The Data Journalism Handbook 2

Towards a Critical Data Practice

Edited by
Jonathan Gray and Liliana Bounegru.

The Data Journalism Handbook 2

Manual de Jornalismo de Dados

Rumo a uma prática crítica de dados

Editado por
Jonathan Gray e Liliana Bounegru

ABRAJI Insper ESCOLA DE DADOS OPEN KNOWLEDGE BRASIL

Google News Initiative DataJournalism.com European Journalism Centre

Capítulos

Introdução

Jonathan Gray e Liliana Bounegru

Jornalismo de dados em questão

O que é jornalismo de dados? Para que serve? O que pode fazer? Quais limitações e oportunidades apresenta? Quem e o que estão envolvidos em sua criação e compreensão? Este livro é um experimento colaborativo que visa responder estas e outras perguntas. Segue os passos de outro livro já editado, *Manual de Jornalismo de Dados: Como os jornalistas podem usar dados para melhorar suas reportagens* (O'Reilly Media, 2012).[1] Ambas as publicações reúnem pluralidade de vozes e perspectivas no registro do jornalismo de dados, que segue em constante evolução. A primeira edição teve origem em um *book sprint* no MozFest de Londres, em 2011, que reuniu jornalistas, tecnólogos, grupos de apoio e demais interessados, com o objetivo de escrever sobre como o jornalismo de dados é feito. Como dito na introdução, este buscava "documentar a paixão e o entusiasmo, a visão e a energia de um movimento que está nascendo", contar as "histórias por trás das histórias" e permitir a "diferentes vozes e visões brilharem". A edição de 2012 foi traduzida para mais de uma dezena de idiomas — incluindo árabe, chinês, tcheco, georgiano, grego, italiano, macedônio, português, russo, espanhol e ucraniano — e é usada para fins de ensino em diversas universidades de renome, bem como centros de treinamento e aprendizagem espalhados pelo mundo, além de ser frequentemente citado por pesquisadores do campo a nível global.

Por mais que o livro de 2012 continue sendo amplamente utilizado (e esta obra visa complementá-lo, não substituí-lo), muito aconteceu desde o ano de sua publicação. Por um lado, o jornalismo de dados se estabeleceu ainda mais. Em 2011, a área ainda estava sendo construída aos poucos, com apenas um punhado de pessoas usando o termo "jornalismo de dados". De lá para cá, este foi socializado e institucionalizado através de organizações dedicadas, cursos de capacitação, vagas de emprego, equipes profissionais, antologias, artigos em periódicos, relatórios, ferramentas, comunidades online, hashtags, conferências, redes, encontros, listas de email e muito mais. Há uma conscientização mais ampla em torno do termo por conta de eventos visivelmente ligados a dados, em casos como os *Panama Papers*, que o delator Edward Snowden citou como "o maior vazamento na história do jornalismo de dados".

Já pelo outro lado, a prática também passou a ser mais contestada. Os vazamentos de Snowden em 2013 ajudaram a estabelecer a existência de um aparato de vigilância transnacional de estados e empresas como um fato e não mera especulação. Estes vazamentos

[1] Gray et al. (2012).

sugeriam como cidadãos poderiam ser reconhecidos através de *big data*, revelando, assim, um lado sombrio de dispositivos, aplicativos e plataformas geradoras de dados.[2] Nos Estados Unidos, o lançamento do veículo de Nate Silver dedicado ao jornalismo de dados, conhecido como *FiveThirtyEight*, em 2014, foi recebido de maneira hostil pela forma como depositava enorme confiança em determinados métodos quantitativos, acompanhados por desdém ao "jornalismo opinativo".[3] Ao passo que Silver recebia os louros de "senhorio e deus do algoritmo" de Jon Stewart, do *The Daily Show*, após prever com sucesso os resultados da eleição norte-americana de 2012, os métodos estatísticos que defendia acabaram por sofrer críticas e contestações após a eleição de Donald Trump, em 2016. A estas eleições, junto ao Brexit no Reino Unido e à ascensão de líderes populistas de direita pelo mundo, atribuía-se um momento de "pós-verdade",[4] caracterizado pela perda generalizada da confiança em instituições públicas, conhecimento de especialistas e fatos associados a estes, além da mediação da vida pública e política por meio de plataformas online que deixavam seus usuários à mercê de direcionamento, manipulação e desinformação.

Se este momento de "pós-verdade" é considerado prova de fracasso ou um clamor à ação não sabemos, mas algo está claro: não podemos mais tomar os dados e mesmo o jornalismo de dados como certos. Dados não oferecem representações neutras e diretas do mundo, visto que se encontram enredados em meio à política e cultura, ao dinheiro e poder. Instituições e infraestruturas que apoiam a produção de dados — de pesquisas a estatísticas, de ciência climática a redes sociais — passaram a ser questionadas. Logo, cabe perguntar: quais dados, de quem e por quais meios? Dados sobre quais temas e para que finalidade? Quais temas são ricos em dados e quais são pobres? Quem é capaz de se beneficiar destes? Quais tipos de público são agrupados através dos dados, quais competências são suportadas, quais tipos de política pode sancionar e quais os tipos de participação que engendra?

Rumo a uma prática crítica de dados

Em vez de dividir tais perguntas e preocupações, este livro busca "acompanhar o problema", como dito pela proeminente acadêmica feminista Donna Haraway.[5] Em vez de tratar a relevância e importância do jornalismo de dados como assertivas, tratamos como uma questão que pode ser abordada de diversas formas. Os capítulos reunidos nesta obra visam oferecer um panorama mais rico sobre o que é o jornalismo de dados, com o que e para

---

[2] https://zenodo.org/record/1415450.

[3] https://www.politico.com/blogs/media/2014/03/knives-out-for-nate-silver-185394.

[4] https://www.nytimes.com/2016/08/24/opinion/campaign-stops/the-age-of-post-truth-politics.html. Para uma visão crítica do termo, consultar Jasanoff e Simmet (2017).

[5] Consultar Haraway (2018). O capítulo assinado por Helen Verran explora como jornalistas podem acompanhar a problemática em torno de valores e números.

quem. Por meio de nosso trabalho editorial, buscamos encorajar a reflexão em torno do que projetos de jornalismo de dados podem fazer, e as condições sob as quais podem obter sucesso. Isto envolve o cultivo de um tipo diferente de precisão na prestação de contas da prática do jornalismo de dados: é necessário especificar as situações em que este se desenvolve e opera. Tal precisão exige ampliar o escopo do livro para incluir não somente os métodos de análise, criação e uso de dados no contexto do jornalismo, mas também as circunstâncias sociais, culturais, políticas e econômicas nas quais estas práticas estão embutidas.

O subtítulo deste novo livro é "Rumo a uma prática crítica de dados", um reflexo de nossa aspiração enquanto editores de promover o pensamento crítico em torno do jornalismo de dados na prática, bem como refletir a postura cada vez mais combativa de praticantes deste jornalismo. O conceito de "prática crítica de dados" é uma referência à "prática técnica crítica" de Philip E. Agre, descrita como "um pé firme no trabalho habilidoso do design e outro no trabalho reflexivo da crítica".[6] Como já escrevemos em algum outro ponto, nosso interesse com esta obra é compreender como o engajamento crítico com os dados pode mudar a prática, abrindo espaços para a imaginação e intervenções públicas em torno de políticas de dados.[7]

Além da contribuições de jornalistas de dados e demais praticantes no tocante ao que fazem, o livro conta ainda com capítulos assinados por pesquisadores — cujo trabalho pode avançar na reflexão crítica das práticas de jornalismo de dados — de variados campos, como antropologia, estudos de ciência e tecnologia, estudos de (nova) mídia, estudos de internet, estudos de plataforma, sociologia da quantificação, estudos de jornalismo, estudos indígenas, estudos feministas, métodos digitais e sociologia digital. Em vez de operarmos com uma divisão mais tradicional de trabalhos em que pesquisadores oferecem reflexão crítica e praticantes oferecem conselhos e dicas instrumentais, buscamos encorajar pesquisadores para que focassem no lado prático de seu trabalho, dando aos praticantes o espaço necessário para reflexão fora dos prazos do seu cotidiano. Nenhuma destas perspectivas exaure o campo de estudo, e nossa intenção é motivar os leitores a atentarem para os diferentes aspectos de como o jornalismo de dados é feito. Em outras palavras, este livro deve funcionar como o pontapé para uma discussão interdisciplinar e, esperamos, como catalisador para esforços colaborativos.

Não presumimos que o "jornalismo de dados" seja um conjunto unificado de práticas. Trata-se, sim, de uma categoria proeminente ligada a uma série de práticas diversas que podem ser estudadas, definidas e experimentadas de maneira empírica. Como mencionado

---

[6] Agre (1997).

[7] Gray (2018).

em recente análise, precisamos investigar o "quanto da quantificação, assim como o mero fato da mesma", aos quais os efeitos "dependem de intenções e implementação".[8] Nosso propósito não é engessar a prática do jornalismo de dados, e, sim, chamar atenção para seus muitos aspectos, abrindo espaço para fazê-lo de forma diferente.

Um experimento coletivo

É válido, ainda, destacar brevemente o que este livro não é. Não se trata de material didático ou manual no sentido convencional destas palavras, visto que os capítulos não adicionam a um corpo de conhecimento estabelecido, mas visam apontar para direções interessantes rumo a investigações e experimentações mais aprofundadas. Não se trata de um guia prático de tutoriais ou "como fazer" isso ou aquilo, afinal já existem material e cursos disponíveis por aí que abordam os diferentes aspectos da prática de dados, sobretudo análise e visualização de dados. Não é também um livro de "bastidores" de estudos de caso, visto que existem diversos artigos e postagens em blogs que revelam como projetos foram realizados, incluindo entrevistas com seus criadores. Não é, ainda, um livro sobre perspectivas acadêmicas recentes, pois há um corpo crescente de literatura sobre jornalismo de dados espraiado ao longo de múltiplos livros e periódicos.[9]

Este livro, enfim, foi criado como um *experimento coletivo* na prestação de contas de práticas em jornalismo de dados ao longo dos últimos anos e um *convite coletivo* à exploração de como tais práticas podem ser modificadas. Coletivo no sentido de que, assim como na primeira edição, conseguimos reunir um número comparativamente grande de colaboradores (70) para um livro curto. O processo editorial se beneficiou de recomendações de colaboradores ao longo de trocas de email. Uma oficina realizada com vários deles no Festival Internacional de Jornalismo de 2018, em Perugia (Itália), serviu de oportunidade para trocas e reflexões. Uma versão "beta" da obra foi publicada online de forma a possibilitar a prévia pública de uma seleção de capítulos antes que sua versão impressa viesse a ser publicada, gerando comentários e encontros antes que a obra tomasse sua forma final. Através deste processo de curadoria, que poderíamos considerar uma espécie de "editorial bola de neve", buscamos descobrir como o jornalismo de dados é feito por diferentes atores, em diferentes locais, abordando diferentes temas, através de diferentes meios. Ao longo desta jornada, nos debruçamos sobre diversas listas (curtas e longas), veículos e conjuntos de dados de forma a realizar a curadoria de diferentes perspectivas acerca da prática do jornalismo de dados. Por mais que houvesse um sem-fim de colaboradores que gostaríamos de incluir, foi necessário operar dentro das restrições do impresso, considerando ainda o equilíbrio na inclusão de vozes dos mais variados gêneros, localizações e temas.

---

[8] Berman and Hirschman (2018).

[9] https://www.zotero.org/groups/data_journalism_research.

Trata-se de um livro experimental ao considerar que seus capítulos oferecem diferentes perspectivas e provocações em torno do jornalismo de dados, o qual convidamos os leitores a explorarem mais profundamente ao criarem suas próprias combinações de ferramentas, conjuntos de dados, métodos, textos, públicos e problemáticas. Em vez de herdar as formas de observar e compreender já embutidas em elementos como conjuntos de dados oficiais ou dados de mídias sociais, encorajamos os leitores a colocarem estas informações a serviço de suas próprias linhas de investigação. Tudo com base em "analítica crítica" e "métodos criativos" que buscam modificar as questões que são feitas e a maneira como os problemas são apresentados.[10] O jornalismo de dados pode ser encarado não somente em termos de como as coisas são *representadas*, mas também na maneira como este organiza *relações* — tanto que não é apenas uma questão de criar matérias com dados (através da coleta, análise, visualização e narrativa de dados), incluindo quem e o que estas histórias reúnem (diferentes públicos, fontes, métodos, instituições e plataformas de mídias sociais). Cabe, então, perguntar, como Noortje Marres fez há pouco: "Quais são os métodos, materiais, técnicas e arranjos de nossa curadoria para a criação de espaços em que problemas possam ser abordados de maneira diferente?". Os capítulos deste livro mostram como o jornalismo de dados pode ser um ofício criativo, imaginativo e colaborativo, destacando como jornalistas de dados interrogam fontes oficiais, criam e compilam seus próprios dados, experimentam com novos formatos interativos e visuais, refletem sobre os efeitos de seu trabalho e fazem seus métodos responsabilizáveis e códigos reutilizáveis.

Se o futuro do jornalismo de dados é incerto, esperamos que os leitores deste livro juntem-se a nós na avaliação crítica do que foi e é jornalismo, bem como na modelagem de seu futuro.

Uma visão geral do livro

Fiéis à nossa ênfase editorial em especificar o cenário em que este livro se apresenta, cabe esclarecer que a orientação da obra e sua seleção de capítulos levam em consideração nossos interesses e os de nossos amigos, colegas e redes neste momento em especial, o que inclui preocupações crescentes com a mudança climática, destruição do meio ambiente, poluição do ar, evasão fiscal, (neo)colonialismo, racismo, sexismo, desigualdade, extrativismo, autoritarismo, injustiça algorítmica e trabalho de plataforma. Os capítulos exploram como o jornalismo de dados possibilita a compreensão e vivência de questões como estas, bem como os tipos de respostas que pode causar. A seleção destes capítulos reflete nossas próprias oscilações entre os campos de pesquisa acadêmica, jornalismo e defesa de interesses, bem como os diferentes estilos de escrita e prática de dados associada a cada um destes.

---

[10] Ver Rogers (2016) e Lury e Wakeford (2014).

Permanecemos convencidos do potencial generativo de encontros entre colegas destes diferentes campos, e vários dos capítulos atestam o sucesso de colaborações multidisciplinares. Além da exploração de sinergias e semelhanças, cabe notar logo no início (como leitores astutos perceberão) que existem diferenças, tensões e fricções entre as perspectivas apresentadas no decorrer dos vários capítulos que compõem esta obra, incluindo diferentes histórias e histórias de origem; diferentes visões sobre metodologia, dados e tecnologias emergentes; diferentes visões sobre a desejabilidade de convencionalização e experimentação com abordagens diversas; e, por fim, diferentes perspectivas do que é jornalismo de dados, para que serve, suas condições e restrições, como este se organiza e as possibilidades que apresenta.

Após a introdução, o livro inicia com um "menu degustação" sobre como utilizar dados na solução de problemas. Isto inclui uma variedade de formatos para a interpretação de diversos temas em múltiplos locais — caso do monitoramento de mortes de operários na Turquia (Dag), a observação de pessoas e cenas por trás do número de demolição de casas na Jerusalém Oriental ocupada (Haddad), a multiplicação das memórias de árvores em Bogotá (Magaña), o estabelecimento de conexões entre commodities agrícolas, crime, corrupção e colonialismo ao longo de vários países (Sánches e Villagrán), a mobilização para maior segurança nas estradas nas Filipinas (Rey) e o mapeamento da segregação nos EUA (Williams). Os capítulos desta seção ilustram a amplitude de práticas que vão de técnicas de visualização à criação de campanhas para o reaproveitamento de dados oficiais com diferentes prioridades analíticas.

A segunda seção foca em como jornalistas reúnem dados, incluindo projetos voltados a temas como conflitos territoriais (Shrivastava e Paliwal) e crimes envolvendo armas brancas (Barr). Conta, ainda, com relatos sobre como obter e trabalhar com dados em países que não oferecem facilidade de acesso, como Cuba (Reyes, Almeida e Guerra) e China (Ma). Reunir dados também pode ser uma forma de interagir com leitores (Coelho) e concentrar interessados em torno de um assunto, o que por si só pode ser um resultado importante dentro de um projeto. A coleta de dados pode envolver a modificação de outras formas de produção de conhecimento, como sondagens e inquéritos, no contexto do jornalismo (Boros). Um capítulo sobre a soberania de dados indígenas (Kukutai e Walter) explora questões sociais, culturais e políticas em torno de dados oficiais e como trazer à baila outras perspectivas marginalizadas no contexto da organização da vida coletiva com dados. Além do uso de números como material para contar histórias, jornalistas de dados também podem contar histórias sobre como estes números são criados (Verran).

A terceira seção aborda diferentes formas de se trabalhar com dados. Inclui algoritmos (Stray), código (Simon) e métodos computacionais (Borges Rey). Os colaboradores examinam problemas e oportunidades emergentes do trabalho com fontes como dados

textuais (Maseda) e dados de aplicativos, plataformas de mídia social e demais dispositivos online (Weltevrede). Outros tratam sobre como fazer do trabalho jornalístico em dados transparente, responsabilizável e reproduzível (Leon; Mazotte). Bancos de dados também podem abarcar oportunidades de trabalho colaborativo em grandes projetos investigativos (Díaz-Struck, Gallego e Romera) O pensamento e a prática feminista, por sua vez, podem inspirar diversas maneiras de se trabalhar com dados (D'Ignazio).

A quarta seção debruça-se sobre o exame das diferentes formas pelas quais os dados podem ser experimentados, começando com uma observação dos diferentes formatos que o jornalismo de dados pode assumir (Cohen). Muitos dos textos presentes refletem acerca de práticas contemporâneas de visualização (Aisch e Rost), assim como a maneira que leitores reagem e participam na interpretação de dados por meio de visualizações (Kennedy et al.). Outros artigos tratam sobre como dados são mediados e apresentados aos leitores através de bancos de dados (Rahman e Wehrmeyer), interação baseada na web (Bentley e Chalabi), rádio e TV (de Jong), quadrinhos (Amancio) e desenhos com dados (Chalabi).

A quinta seção é dedicada às abordagens emergentes para investigação de dados, plataformas e algoritmos. Projetos de jornalismo recentes consideram o digital não somente um campo que oferece novas técnicas e oportunidades para jornalistas, mas também novos objetos de investigação — caso do trabalho amplamente compartilhado do *Bellingcat* e do *Buzzfeed* sobre conteúdo viral, desinformação e cultura digital.[11] Os capítulos nesta seção esmiuçam as diferentes maneiras de reportar sobre algoritmos (Diakopoulous) e como conduzir colaborações de longo prazo nesta área (Elmer). Outros capítulos dedicam-se a abordar métodos de trabalho com mídias sociais de forma a explorar como plataformas moldam o debate, incluindo abordagens em *storytelling* (Vo). O capítulo final explora as afinidades entre pesquisa de métodos digitais e jornalismo de dados, versando sobre como os dados podem ser usados para contar histórias sobre infraestruturas de monitoramento web (Rogers).

A sexta parte trata da organização do jornalismo de dados, cobrindo variados tipos de trabalho de campo considerados indispensáveis, mas nem sempre prontamente reconhecidos. Entre os temas abordados, temos as mudanças sofridas pelo jornalismo de dados na última década (Rogers); como plataformas e a *gig economy* moldam redes investigativas entre fronteiras (Candea); enredamentos entre jornalismo de dados e movimentos em prol de dados abertos e tecnologia cívica (Baack); práticas de código aberto (Pitts e Muscato); práticas de mensuração de audiência (Petre); arquivamento de jornalismo de dados (Broussard); e o papel da hashtag #ddj na conexão de comunidades de jornalismo de dados no Twitter (Au e Smith).

---

[11] https://www.buzzfeednews.com/topic/fake-news e https://www.bellingcat.com/.

A sétima parte trata do treinamento de jornalistas de dados e do desenvolvimento do jornalismo de dados ao redor do mundo. Esta seção conta com capítulos sobre o ensino do tema em universidades nos EUA (Phillips); o fortalecimento de comunidades marginalizadas para que contem suas próprias histórias (Constantaras, Vaca); e os cuidados com o "universalismo digital" e a subestimação da inovação na "periferia" (Chan).

O jornalismo de dados não ocorre no vácuo e a oitava parte deste livro traz à tona seus diversos cenários sociais, políticos, culturais e econômicos. Um capítulo sobre as genealogias do jornalismo de dados nos EUA serve para estimular a reflexão sobre as diversas ideias e práticas históricas que o fazem ser o que é (Anderson). Os demais capítulos presentes tratam de temas como o jornalismo de dados enquanto resposta aos processos sociais mais amplos de dataficação (Lewis e Radcliffe); como projetos de jornalismo de dados são valorizados através de premiações (Loosen); diferentes abordagens na mensuração do impacto de projetos de jornalismo de dados (Bradshaw; Green-Barber); e questões em torno do jornalismo de dados e colonialismo (Young).

A seção final encerra a obra com reflexões, desafios e possíveis direcionamentos futuros para o setor. Para tanto, temos textos sobre jornalismo de dados e liberalismo digital (Boyer); quais interesses são atendidos por projetos em jornalismo de dados (Young e Callison); e se o jornalismo de dados consegue atender sua antiga aspiração de se transformar um campo de experimentação, interatividade e atividade inspirada (Usher).

Doze desafios para a prática crítica de dados

Inspirados pelo tempo que passamos explorando o campo do jornalismo de dados ao longo do desenvolvimento deste livro, gostaríamos de apresentar doze desafios para uma "prática crítica de dados". Estes desafios consideram o jornalismo de dados em termos de sua capacidade de *moldar relações* entre diferentes atores, bem como a de *criar representações* sobre o mundo.

1. Como projetos de jornalismo de dados podem contar histórias *com e sobre dados* incluindo aí os mais variados atores, processos, instituições, infraestruturas e formas de conhecimento através dos quais dados são gerados?
2. Como projetos de jornalismo de dados podem contar histórias sobre grandes problemas em escala (como mudança climática, desigualdade, taxação de multinacionais, migração) ao passo que *afirmam a provisionalidade e reconhecem os modelos, suposições e incertezas* envolvidos na produção de números?

3. Como projetos de jornalismo de dados levam em conta o *caráter coletivo de dados digitais, plataformas, algoritmos e dispositivos online*, incluindo a relação entre tecnologias e culturas digitais?
4. Como projetos de jornalismo de dados *cultivam suas próprias formas de tornar as coisas inteligíveis, significantes e relacionáveis através dos dados*, sem avançar de forma não crítica sobre os saberes embutidos nos dados advindos de instituições, infraestruturas e práticas dominantes?
5. Como projetos de jornalismo de dados *reconhecem e experimentam com as culturas visuais e estéticas nas quais se baseiam*, incluindo combinações de visualizações de dados e outros materiais visuais?
6. Como projetos de jornalismo de dados podem criar espaços para *participação e intervenção públicas* através do questionamento de fontes de dados já estabelecidas e reinterpretação de quais problemas são explicados através dos dados?
7. Como jornalistas de dados podem cultivar e afirmar conscientemente *seus estilos de trabalho com dados*, que podem se basear em campos como estatística, ciência de dados e analítica de mídias sociais, ainda que distintos?
8. Como o campo de jornalismo de dados pode *desenvolver práticas de memória para arquivamento e preservação de seu trabalho*, bem como situar a produção em relação às práticas e culturas nas quais se baseia?
9. Como projetos de jornalismo de dados podem colaborar em questões transnacionais de forma a *evitar a lógica da plataforma e da colônia, de forma a assegurar inovações na periferia?*
10. Como o jornalismo de dados pode apoiar comunidades marginalizadas no uso de dados para que possam *contar suas próprias histórias do seu jeito*, em vez de contarmos estas histórias em seu lugar?
11. Como projetos de jornalismo de dados podem desenvolver seus próprios *métodos alternativos e criativos para prestação de conta de seus valores e impacto no mundo* para além de métricas de mídias sociais e metodologias de impacto estabelecidas em outras áreas?
12. Como o jornalismo de dados poderia *desenvolver um estilo de objetividade que afirme, não minimize, seu papel interventor no mundo* e na modelagem de relações entre diferentes atores na vida coletiva?

Agradecimentos

Agradecemos à Amsterdam University Press pelo apoio neste projeto experimental, incluindo a publicação de uma edição beta online, e no apoio para uma versão digital de acesso livre deste livro. Talvez esta seja uma escolha adequada, tendo em vista que muitos

dos colaboradores se encontraram em uma conferência internacional sobre jornalismo de dados em Amsterdã, no ano de 2010. O financiamento para livre acesso vem de uma bolsa concedida pela Organização Holandesa para Pesquisa Científica (NWO, 324-98-014).

A visão deste livro nasceu por meio de discussões com amigos e colegas associados ao Public Data Lab. Nos beneficiamos particularmente de conversas sobre a obra com Andreas Birkbak, Erik Borra, Noortje Marres, Richard Rogers, Tommaso Venturi e Esther Weltevrede. Também nos foi cedido espaço para o desenvolvimento deste livro por meio de eventos e visitas às Universidade de Columbia (em conversa com Bruno Latour); Universidade de Utrecht; Universidade da Califórnia em Berkeley; Universidade de Stanford; Universidade de Amsterdã; Universidade de Miami; Universidade Aalborg de Copenhague; Instituto de Estudos Políticos de Paris; Universidade de Cambridge; Escola Londrina de Economia; Universidade de Cardiff; Universidade de Lancaster; e ao Festival Internacional de Jornalismo de Perugia. Estudantes do Mestrado de Artes em Jornalismo de Dados do King's College de Londres nos ajudaram a testar o conceito de "prática crítica de dados" que está no coração deste livro.

Nossa esperança de longa data de lançar outra edição foi nutrida e virou realidade graças a Rina Tsubaki, que ajudou na obtenção de apoio junto ao Centro Europeu de Jornalismo e à Google News Initiative. Somos gratos a Adam Thomas, Bianca Lemmens, Biba Klomp, Letizia Gambini, Arne Grauls e Simon Rogers pela liberdade editorial e pelo apoio duradouro na ampliação de nossos esforços. A assistência editorial de Daniela Demarchi foi inestimável ao nos ajudar a traçar um caminho claro entre a enxurrada de textos, notas de rodapé, referências, emails, documentos compartilhados, versões de artigos, planilhas e demais materiais.

Acima de tudo isso, gostaríamos de agradecer aos praticantes e pesquisadores de jornalismo de dados que se envolveram no projeto (seja por meio de escrita, correspondência ou debate) por nos acompanhar nesta jornada e por apoiar este experimento ao colaborarem com tempo, energia, materiais e ideias, sem os quais este projeto não seria possível. Este livro é, e continua sendo, uma obra coletiva.

Jonathan Gray é Conferencista de Estudos de Infraestrutura Crítica do Departamento de Ciências Humanas Digitais do King's College de Londres. Liliana Bounegru é Conferencista de Métodos Digitais do Departamento de Ciências Humanas Digitais do King's College de Londres. Ambos são cofundadores do Public Data Lab e Pesquisadores Associados da Digital Methods Initiative.

Referências

AGRE, Philip E. *Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI*. In: BOWKER, Geoffrey et al. (ed.). *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*. Mahwah: Erlbaum, 1997, p. 130–57.

BERMAN, Elizabeth P.; HIRSCHMAN, Daniel, *The Sociology of Quantification: Where Are We Now?*. Contemporary Sociology, 2018.

BYERS, Dylan. *Knives out for Nate Silver*. Politico, 2014.

GRAY, Jonathan; BOUNEGRU, Liliana; CHAMBERS, Lucy. *The Data Journalism Handbook: How Journalists Can Use Data to Improve the News*. O'Reilly Media, 2012.

Jonathan GRAY, 'Three Aspects of Data Worlds', *Krisis: Journal for Contemporary Philosophy*, 2018: 1. Disponível em: http://krisis.eu/three-aspects-of-data-worlds/.

GRAY, Jonathan; BOUNEGRU, Liliana. *What a Difference a Dataset Makes? Data Journalism And/As Data Activism*. In: EVANS, J; RUANE S.; SOUTHALL, H. (ed.). *Data in Society: Challenging Statistics in an Age of Globalisation*. Bristol: The Policy Press, 2019.

GRAY, Jonathan, GERLITZ, Carolin; BOUNEGRU Liliana. *Data infrastructure literacy*. Big Data e Society, 5(2), 2018, p. 1–13.

HARAWAY, Donna J. *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press, 2016.

LURY, Celia; WAKEFORD, Nina (ed.). *Inventive Methods: The Happening of the Social*. Londres: Routledge, 2012.

ROGERS, Richard. *Otherwise Engaged: Social Media from Vanity Metrics to Critical Analytics*. International Journal of Communication, 12, 2018, p. 450–72.

Solucionando problemas com dados

Do café ao colonialismo: investigações de dados sobre como pobres alimentam ricos

Raúl Sánchez e Ximena Villagrán

No início de 2016, um pequeno grupo de jornalistas decidiu investigar a jornada de uma barra de chocolate, uma banana ou uma xícara de café, desde o cultivo até a chegada às suas mesas. Nossa investigação foi incitada por relatos que indicavam que todos estes itens eram produzidos em países pobres e consumidos majoritariamente em países ricos.

Com base nestes dados, decidimos fazer algumas perguntas: como são as condições de trabalho nestes cultivos? Há uma concentração da posse destas terras nas mãos de um pequeno grupo? Que tipos de dano ambiental estes produtos causam nestes países? Então, *El Diario* e *El Faro* (dois veículos digitais independentes) juntaram forças para investigar o lado obscuro do modelo de negócios da agroindústria de países em desenvolvimento.[12]

O projeto 'Terra Escrava' (*La Tierra Esclava*, no original) é uma investigação transfronteiriça movida por dados de um ano de duração, com um subtítulo que vai direto ao ponto: "É assim que países pobres são usados para alimentar países ricos".[13] De fato, o colonialismo é a principal questão deste projeto. Enquanto jornalistas, não queríamos contar a história de indígenas pobres sem examinar um panorama mais sistêmico. Queríamos explicar como posse de terras, corrupção, crime organizado, conflitos locais e cadeias de suprimentos de certos produtos ainda fazem parte de um sistema colonialista.

Neste projeto, investigamos cinco tipos de cultivos consumidos amplamente nos EUA e na Europa: açúcar, café, cacau, banana e óleo de palma produzidos na Guatemala, na Colômbia, na Costa do Marfim e em Honduras. Por se tratar de uma investigação baseada em dados, usamos estes dados para partirmos do padrão à história. A escolha destes cultivos e países foi feita com base em análise prévia de dados de 68 milhões de registros do Banco de Dados de Comércio das Nações Unidas (Figura 1).

---

[12] https://www.eldiario.es/, https://elfaro.net/.

[13] https://latierraesclava.eldiario.es/.

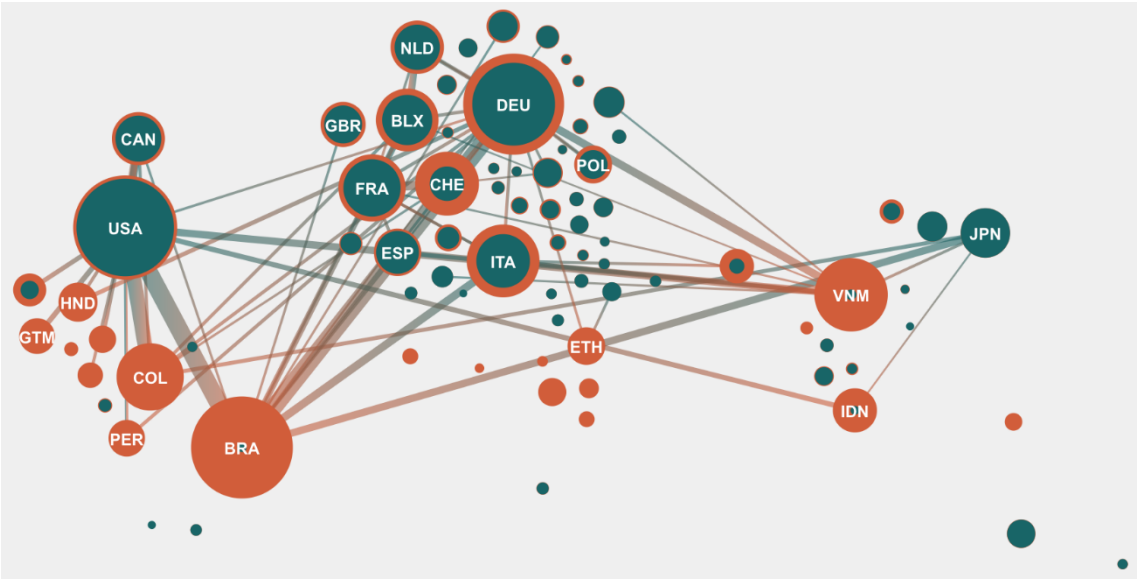


Figura 1: Gráfico de rede detalha importações e exportações de café em 2014.

Esta investigação revelou que a balança de poder entre países ricos e pobres mudou desde o século XV até o presente, provando que estes cultivos são produzidos graças a condições exploratórias, análogas à escravidão, para os trabalhadores envolvidos, práticas comerciais ilegais e danos ambientais continuados.

O foco de nossas histórias foi escolhido com base em dados. Em Honduras, a chave foi utilizar informações geográficas para desenrolar o enredo. Compilamos o atlas de uso da terra do país e comparamos a superfície de cultivos de palmeiras com áreas protegidas. Descobrimos que 7.000 hectares de palmeiras haviam sido plantados ilegalmente em zonas de proteção pelo país. Com base nisso, nosso repórter poderia investigar plantações específicas dentro destas zonas. A história usa casos individuais para destacar e abordar o abuso sistêmico, como 'Monchito', camponês hondurenho que cultiva palmeira-africana dentro do Parque Nacional Jeannette Kawas.

O projeto não foca somente no uso de terras. Na Guatemala, criamos um banco de dados de todos os engenhos de açúcar do país. Mergulhamos fundo no registro local de empresas para conhecermos os proprietários e diretores destes engenhos. Então traçamos as ligações destas pessoas e entidades a empresas offshore com o auxílio de registros públicos do Panamá, das Ilhas Virgens e Bahamas. Para descobrir como esta estrutura offshore era criada e gerida, *El Faro* teve acesso ao banco de dados dos *Panama Papers*, então usamos esta informação para revelar como um dos maiores engenhos do país trabalhava junto ao escritório de advocacia Mossack Fonseca para sonegar impostos.

Uma investigação transnacional cujo objetivo é descobrir corrupção e práticas escusas de negócios em países de terceiro mundo é desafiadora. Tivemos que trabalhar em áreas rurais onde não há presença do governo, correndo riscos na maior parte do tempo. Além disso, tivemos de lidar com países em que há considerável falta de transparência, em que dados não são abertos e, em alguns casos, com uma administração pública que não sabia nem que informação tinha.

Honduras e Guatemala eram apenas uma parte da investigação. Mais de dez pessoas trabalharam juntas na produção deste material. Todo o trabalho foi coordenado a partir dos escritórios do *eldiario.es* na Espanha e do El Faro em El Salvador, em cooperação com jornalistas na Colômbia, na Guatemala, em Honduras e na Costa do Marfim.

Este trabalho foi levado adiante não só por jornalistas, mas também por editores, fotógrafos, designers e desenvolvedores que trabalharam no desenvolvimento e processo produtivo de um produto integrado para a web. Nada disso seria possível sem eles.

Usamos narrativa integrada em sistema de rolagem para cada uma das investigações. Para nós, a forma como os usuários leem e interagem com as histórias é tão importante quanto o próprio processo de investigação. Optamos por combinar imagens de satélite, fotos, visualizações de dados e narrativa pois queríamos que o leitor compreendesse a ligação entre os produtos que consomem e os agricultores, empresas e demais atores envolvidos em sua produção.

Esta estrutura nos permitiu usar um formato narrativo em que histórias pessoais tinham o mesmo peso que a análise de dados. Falamos, por exemplo, de John Pérez — um camponês colombiano cuja terra foi roubada por grupos paramilitares e grandes empresas de banana durante conflito armado — com um mapa em que era possível dar zoom, que levava o leitor da sua plantação até o destino final da produção colombiana de bananas.

Este projeto mostrou que o jornalismo de dados pode enriquecer técnicas tradicionais de reportagem, conectando histórias sobre pessoas a fenômenos sociais, econômicos e políticos mais amplos.

Ele também foi publicado pelos veículos *Plaza Pública*, da Guatemala, *e Ciper*, do Chile, e foi incluído no programa de rádio guatemalteco *ConCriterio*. Este último levou a um pronunciamento da Agência Tributária Guatemalteca, que pedia por recursos para combater a evasão fiscal dos engenhos de açúcar.

Raúl Sanchez é espanhol e jornalista do eldiario.es, onde cobre temas como desigualdade, gênero e corrupção. Ximena Villagrán é jornalista de dados, atuante na América Central e Espanha.

1. Reutilizando dados do censo para medir a segregação nos Estados Unidos

Aaron Williams

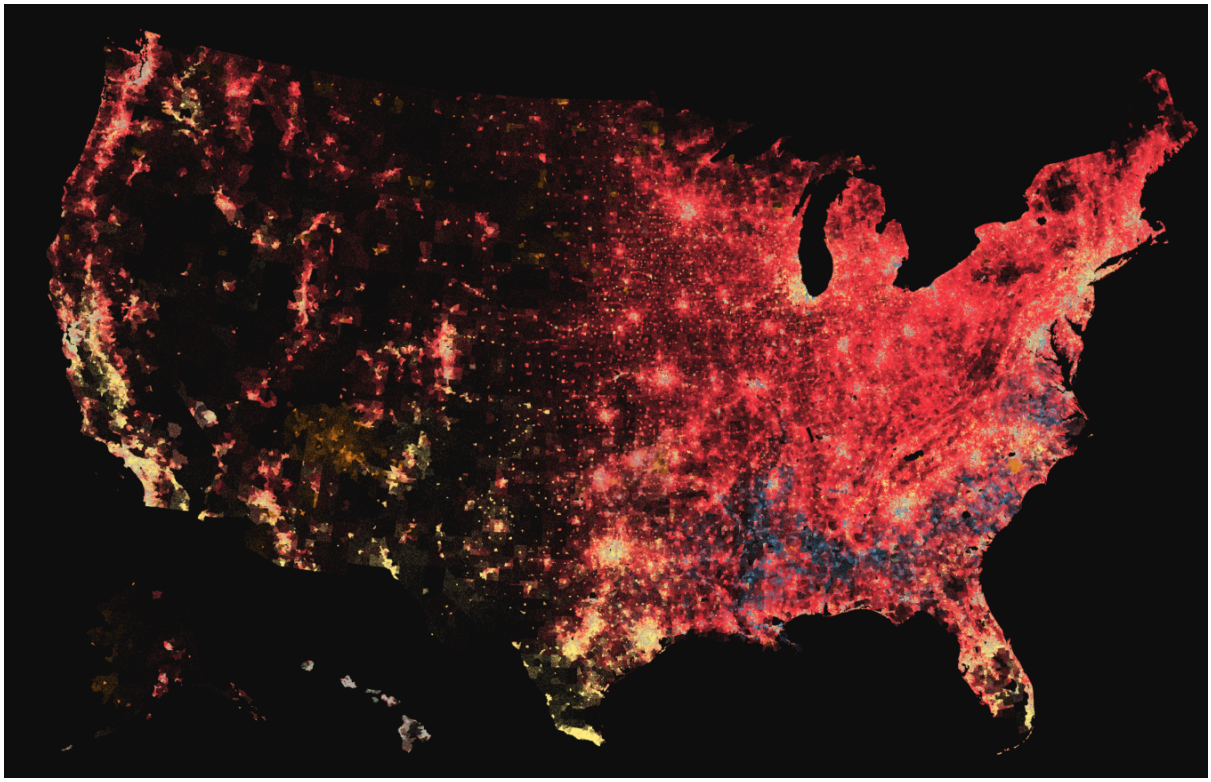


Figura 1: Imagem retirada do artigo *America is more diverse than ever — but still segregated*, publicado no *Washington Post* em 2018.[14]

Como se mede a segregação por raça? Nos EUA, em específico, há um esforço histórico na separação das pessoas desde sua fundação. Ao passo que o país mudava, e leis racistas como a segregação tornaram-se ilegais, novas leis surgiram com o objetivo de manter afro-americanos e outros grupos separados dos norte-americanos brancos. Boa parte da população vem sentindo os efeitos duradouros destas leis, mas o que eu gostaria de saber é se havia como medir o impacto destas medidas de acordo com o local de residência

Ter lido *We Gon' Be Alright: Notes on Race and Resegregation*, de Jeff Chang, me inspirou. A obra é composta por uma série de ensaios em que o autor explora raça e lugar, dois temas interligados. Fiquei impressionado com capítulos que falavam sobre as mudanças demográficas em cidades como São Francisco, Los Angeles e Nova York; aquilo me fez querer trabalhar em um projeto que quantificasse as ideias sobre as quais Chang escrevia.

[14] https://www.washingtonpost.com/graphics/2018/national/segregation-us-cities/.

Muitos dos mapas que mostram a segregação não o fazem de fato. Na maioria das vezes, estes materiais usam um ponto para representar um membro de cada raça ou etnia dentro de um espaço geográfico, baseando sua cor na raça daquele indivíduo. Eles acabam por gerar mapas populacionais fascinantes sobre onde estas pessoas vivem, mas não medem o quão diversas ou segregadas estas regiões são.

Como determinar isso? Bem, segregação e diversidade são dois termos com definições extremamente diversas, dependendo de com quem você está falando. E por mais que muitas pessoas percebam os locais em que vivem como segregados, esta percepção pode mudar com base na forma como se mede esta segregação. Não me interessava agir por conta de relatos esparsos. Assim sendo, busquei por formas de medir a segregação de maneira acadêmica e basear minha reportagem nisso.

Entrevistei Michael Bader, professor adjunto de sociologia da American University em Washington, que me apresentou o Índice de Entropia Multigrupo (também conhecido como Índice Theil), uma medida estatística que determina a distribuição especial de múltiplos grupos raciais simultaneamente. Usamos este índice para medir cada grupo do censo norte-americano em comparação à população racial do condado que habitava.

O projeto levou cerca de um ano para ser completado. A maior parte do tempo foi gasto explorando dados e diversas medidas de segregação. Ao longo de minha pesquisa, descobri que há diversas formas de mensurar esta segregação. O Índice de Entropia Multigrupo, por exemplo, é uma medida de uniformidade, que compara a distribuição populacional ao longo de determinada localização geográfica. Há outros métodos, como o Índice de Exposição, que calcula as chances de dois grupos entrarem em contato na mesma localização. Não há um método único para provar a existência de segregação ou não, mas estas medidas podem ser usadas em conjunto para explicar como uma comunidade é constituída. Li muitas pesquisas a respeito de demografia de censo e tentei basear minhas categorias às presentes na literatura sobre o tema. Logo, escolhi as seis categorias raciais deste projeto com base em pesquisa já existente sobre raça na questão da segregação encomendada pelo órgão responsável pelo censo nos EUA, optando pelo uso do Índice de Entropia Multigrupo pois este me permitia comparar diversos grupos raciais em uma única análise.

Decidi comparar a constituição de cada bloco censitário com a composição racial do condado ao seu redor. Então, meu colega Armand Emamdjomeh e eu passamos meses trabalhando em cima da estrutura que movia a análise destes dados. Já havia visto muita pesquisa demográfica censitária feita em ferramentas como Python, R ou SPSS no passado, mas estava curioso para saber se conseguiria fazer isso com JavaScript. Descobri que o ecossistema node.js e o JavaScript oferecem um conjunto rico de ferramentas para trabalhar

com dados e mostrá-los na web. Um desafio foi ter que criar muitas de minhas funções analíticas na unha, mas em compensação pude entender cada passo de minha análise e usar estas mesmas funções na web. Tanto o Mapbox quanto o d3.js contam com ferramentas poderosas e estabelecidas para lidar com os dados geoespaciais levantados a cada estágio de minha análise.

Dois meses após a publicação da matéria, continuamos trabalhando em cima de seu design e layout. Uma versão mais antiga deste projeto implementou uma abordagem do tipo "narrativa em rolagem", em que o mapa cobria toda a tela e o texto rolava sobre o mapa. Por mais que este seja um formato já bem estabelecido e amplamente utilizado por minha equipe no *Post*, ele impedia a inclusão dos belos gráficos estáticos que havíamos criado de forma holística. No final, optamos por um layout tradicional que explorava a história da segregação e discriminação habitacional nos EUA, com estudos de caso em três cidades, e inclusão do mapa histórico completo, interativo, ao final.[15]

Este é o artigo mais lido de minha carreira jornalística. Acredito que deixar os leitores explorarem os dados ao final do texto conferiu uma camada de personalização que possibilitou a estes se situarem na narrativa. O jornalismo de dados nos permite contar histórias que vão além de palavras e ideias. Podemos colocar o leitor no centro de tudo e deixá-los contarem suas próprias histórias.

Aaron Williams é repórter investigativo de dados especializado em análise demográfica no The Washington Post.

[15] https://www.washingtonpost.com/graphics/2018/national/segregation-us-cities/.

Multiplicando memórias ao descobrir árvores em Bogotá

María Isabel Magaña

Bogotá conta com quase 16% da população colombiana em apenas 1.775 quilômetros quadrados. Uma cidade lotada, de ritmo acelerado. Também uma cidade verde, cercada por montanhas e com diversas árvores. Na maior parte do tempo, estas mesmas árvores passam despercebidas pelos cidadãos em meio à correria do cotidiano. Ou ao menos era o que acontecia com membros da nossa equipe de dados, com exceção de uma das programadoras, que ama árvores e é incapaz de sair na rua sem prestar atenção nelas. Uma conhecedora de todas as espécies e fatos sobre elas. Seu amor pela natureza em meio ao caos urbano nos fez pensar: será que alguém já parou para falar das árvores espalhadas pela cidade?

Um questionamento simples e também um catalisador para tantos outros: o que sabemos sobre estas árvores? Quem é o responsável por cuidar delas? Elas são mesmo úteis na purificação da poluição urbana? Precisamos de mais árvores por aí? É verdade que só bairros ricos têm árvores altas? Existem árvores históricas pela cidade?

Nossa investigação começou com dois objetivos diferentes: primeiro, conectar moradores com as gigantes verdes que veem todos os dias; e segundo, entender a realidade dos planos de conservação e plantio de árvores da cidade.[16]

Para tanto, analisamos os dados do censo urbano de plantio de árvores de Bogotá, publicado pelo Jardim Botânico em 2007, o único conjunto de informações disponível, atualizado mês a mês. O Jardim Botânico se recusou a fornecer os dados completos, mesmo após diversas solicitações de livre acesso à informação, todas feitas com embasamento legal. A posição apresentada era bem simples, afinal, os dados já estavam disponíveis no portal *DataViz*. O problema? Só é possível baixar 10.000 registros e o banco de dados tinha 1,2 milhões destes. São dados públicos, é só nos dar! A resposta: não daremos nada, mas melhoraremos nosso aplicativo para que 50.000 registros possam ser baixados.

Nossa solução? Entrar em contato com outras organizações que haviam colaborado com a coleta de dados do Jardim Botânico. Uma destas entidades era o Ideca, que coleta informações relacionadas ao registro imobiliário da cidade. Nos forneceram os dados rapidinho. Nós, obviamente, optamos por publicá-los, para que qualquer um pudesse ter acesso. Uma pequena vingança contra a falta de transparência. O pessoal do Jardim Botânico

[16] http://especiales.datasketch.co/arboles-bogota/.

percebeu e logo o diálogo foi encerrado. De nossa parte, decidimos não empreender nenhuma batalha legal.

Além disso, incluímos também dados públicos da Prefeitura de Bogotá e do Censo Nacional, cruzando informações para análise em relação às árvores. Por fim, conduzimos entrevistas com especialistas ambientais e engenheiros florestais, o que possibilitou a compreensão dos desafios diante da cidade. Estes especialistas haviam trabalhado muito, investigando não só a realidade dos esquemas de plantio, mas também a história por trás das árvores espalhadas por Bogotá. Trabalho, em sua grande parte, ignorado por autoridades, jornalistas e tantos outros.

O produto final consistiu em um projeto de dados em oito partes que revelava os planos de plantio de árvores pela cidade, mapeando cada espécime — incluindo informações sobre altura, espécie, e benefícios para Bogotá — e desmistificando mitos em torno do plantio, além de desvendar o que havia por trás de algumas árvores históricas. Usamos as plataformas *Leaflet* e *SoundCloud* para interação, com design implementado por nossa talentosa equipe de programadores. Também usamos o *StoryMapJS* para possibilitar aos cidadãos explorarem as árvores históricas da cidade.

Decidimos como e que partes eram importantes para o projeto após pesquisarmos diversos projetos semelhantes, então nos juntamos a um designer para criar uma boa experiência de usuário. Foi nosso primeiro projeto grande com dados, o que envolveu muita tentativa e erro, bem como experimentação e exploração.

Mas mais importante que isso foi o envolvimento dos cidadãos, convidados para a construção de um catálogo colaborativo de árvores e para compartilharem suas histórias sobre as árvores mapeadas. O convite se deu através das redes sociais, onde convidamos todos a adicionarem informações sobre espécies de árvores em uma planilha. Até hoje, os moradores de Bogotá nos ajudam a enriquecer este material. Também disponibilizamos um número de WhatsApp para o qual as pessoas poderiam enviar áudios, contando suas histórias com as árvores. Recebemos quase uma centena de mensagens de voz destas pessoas e seus relatos sobre as árvores onde deram seu primeiro beijo, árvores onde aprenderam a escalar, árvores que os protegeram de assaltos ou que simplesmente deixaram saudade ao serem derrubadas. Decidimos incluir este áudio como um filtro extra no aplicativo de visualização, para que os usuários também pudessem conhecer as árvores da cidade através das histórias das pessoas.

O artigo e sua parte visual foram republicados por um jornal de alcance nacional (em suas versões impressa e online) e compartilhados por autoridades locais e cidadãos que queriam contar suas histórias, de forma a transformar a relação que demais residentes têm

com a cidade. Este mapa tem sido usado para uma compreensão da natureza da cidade e como material de apoio para pesquisa sobre as árvores de Bogotá.

Para nós, um dos projetos mais desafiadores que já fizemos. Mas também um dos mais valorosos, pois mostra como o jornalismo de dados pode ir além dos números, sendo capaz de assumir um papel na criação, na coleta e no compartilhamento de cultura e memórias, ajudando a população a perceber coisas sobre o local onde mora (fora de tabelas e gráficos), e na multiplicação e mudança nas relações entre pessoas, plantas e histórias em espaços urbanos.

María Isabel Magaña é mulher, colombiana, jornalista de dados e professora.

Por trás dos números: demolição de casas na Jerusalém Oriental Ocupada

Mohammed Haddad

Ao observar o gráfico abaixo (Figura 1), você verá uma série de barras laranjas e pretas estáveis, seguida por um grande pico em 2016. Prestando atenção na legenda, você entenderá que este gráfico mostra o número de estruturas destruídas e pessoas afetadas pela política israelense de demolição de casas.

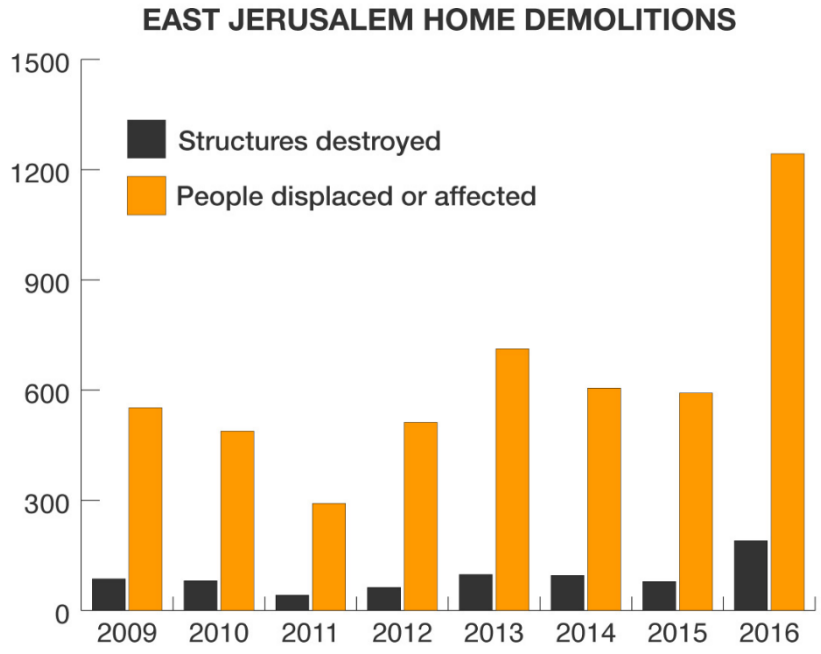


Figura 1: Gráfico da *Al Jazeera* a respeito de demolições em Jerusalém Oriental, de 2009 a 2016.

Como dito por Nathan Yau, autor do livro *Flowing Data* (sem edição em português), "dados são uma abstração da vida". Cada número é uma família e cada número conta uma história.

Broken Homes é o nome do mais abrangente projeto de monitoramento de demolições de casas em Jerusalém Oriental, vizinhança palestina ocupada por Israel há 50 anos.[17]

Em parceria com as Nações Unidas, a *Al Jazeera* monitorou cada demolição doméstica realizada neste território em 2016. Aquele foi um ano recorde, com a marca de 190 estruturas destruídas e mais de 1.200 palestinos afetados.

Decidimos dar início a este projeto após notarmos uma crescente na violência entre israelenses e palestinos ao final de 2015. Tínhamos dois objetivos: entender como as políticas de demolição de Israel seriam afetadas pelo aumento nas tensões e contar aos leitores as histórias humanas por trás dos dados.

O projeto revela o impacto sofrido pelas famílias palestinas por meio de testemunhos em vídeo, fotos em 360 graus e um mapa interativo que destaca a localização, a frequência e o impacto de cada demolição.

Nosso produtor em Doha passou a trabalhar junto com a ONU no final de 2015 para montar o esqueleto do projeto. A ONU coleta, rotineiramente, dados sobre estas demolições e, ao passo que parte destes são disponibilizados online, informações como coordenadas de GPS, dentre outras, são registradas internamente, apenas. Queríamos poder mostrar cada ponto de demolição em um mapa, então passamos a receber dados da ONU mensalmente. Para cada incidente, incluíamos a data da demolição, o número de pessoas e estruturas afetadas, uma breve descrição do ocorrido e um ponto em nosso mapa de Jerusalém Oriental marcando a localização. Cruzamos as informações com notícias e demais informações locais sobre estas demolições. A partir disso, escolhíamos um caso a ser destacado por mês, de forma a mostrar as diferentes facetas da ação israelense — demolições punitivas a administrativas, que afetavam de crianças a idosos.

Nosso repórter de campo percorreu Jerusalém Oriental ao longo de um ano para falar com muitas das famílias atingidas, buscando explorar suas perdas com maior profundidade e fotografar e documentar os locais de demolição.

As reações de cada família variaram grandemente. As entrevistas tinham de ocorrer no local da demolição, o que poderia ser difícil para os atingidos, exigindo paciência e sensibilidade em todas as etapas do projeto, do agendamento dos encontros à gravação do material.

---

[17] https://interactive.aljazeera.com/aje/2017/jerusalem-2016-home-demolitions/index.html.

No geral, a reação das famílias foi positiva. Foram todas muito generosas ao partilharem de seu tempo e suas experiências. Em determinada situação, um homem chegou a escrever uma lista de coisas que gostaria de nos dizer. Em outra, foram necessárias algumas tentativas até que uma das famílias optasse por participar. Uma destas, inclusive, se negou a nos encontrar, então tivemos que entrar em contato com a ONU para encontrar gente disposta a falar da demolição de seu lar.

Figura 2: Foto panorâmica de casa demolida em maio de 2016.

Muitos veículos de notícias, incluindo a *Al Jazeera*, cobriram demolições individuais ao longo dos anos. Um dos principais motivos pelos quais optamos por uma abordagem baseada em dados desta vez foi poder contextualizar, de maneira clara, a escala da história ao enumerar cada uma das demolições. Esta contextualização e uma perspectiva nova se mostram particularmente importantes quando se trata de um tópico continuado, a fim de capturar a atenção do leitor.

Uma dica para aspirantes a jornalista de dados: uma abordagem baseada em dados não precisa ser técnica demais ou custar caro. Por vezes, apenas tomar nota das ocorrências de determinado evento no tempo é o suficiente para descobrir muito sobre a escala de um problema. Contanto que sua metodologia de coletas permaneça consistente, muitas histórias podem ser contadas através de dados que não poderiam ser contadas de outra forma. Por fim, tenha paciência. Coletamos dados durante um ano inteiro para contar esta história. O mais

importante é delinear exatamente quais dados são necessários antes de mandar quaisquer repórteres para o campo. Na maior parte do tempo, não são necessários equipamentos especializados. Usamos um iPhone para todas as fotos em 360 graus e para a captura de coordenadas específicas de GPS.

O projeto, lançado em janeiro de 2017 em inglês, árabe e bósnio, é um alerta sombrio a respeito do que nos espera caso Israel continue a negar licenças para construção a 98% dos solicitantes palestinos, pressionando ainda mais uma população grande e crescente.

Mohammed Haddad é jornalista de dados da Al Jazeera e cofundador do site PalestineRemix.com.

Mapeamento de acidentes rodoviários em prol da segurança nas estradas filipinas

Aika Rey

Dados mostram que acidentes automobilísticos fatais nas Filipinas crescem através dos anos. Ferimentos causados em acidentes são a causa de morte número um entre os jovens filipinos.

Por conta disso, criamos um microsite que compila informações importantes sobre segurança no trânsito. Coletamos e analisamos dados, publicamos artigos e criamos oportunidades de engajamento civil, no mundo real e no digital, em torno de descobertas feitas a partir de dados reunidos de forma a educar o público em relação ao tema segurança nas estradas.[18]

Também começamos uma série em vídeo intitulada "Direito de Passagem" que aborda os problemas enfrentados por motoristas e passageiros na região metropolitana de Manila. Foi assim que a campanha #SaferRoadsPH da *Rappler* surgiu.

Figura 1: Captura de tela do site sobre segurança no trânsito disponível em *rappler.com*

[18] http://www.rappler.com/saferroadsph.

Compilar dados relevantes sobre mortes e ferimentos causados por acidentes de trânsito foi desafiador. Sem um banco de dados nacional abrangente sobre o tema, batemos em muitas portas e coletamos informações junto a autoridades locais e nacionais, incluindo delegacias em diversas cidades e províncias.

Os dados recebidos não eram padronizados. Boa parte do trabalho envolveu a limpeza deste material para que pudesse ser analisado. Um grande desafio foi o mapeamento dos dados quando a informação a respeito da localização estava incompleta ou não havia sido registrada de forma consistente (publicamos uma explicação completa a respeito de nossas fontes). [19]

Ao usar o software de limpeza de dados da Google, Open Refine, obtivemos um banco de dados padronizado, com informações de diversas agências governamentais. Isso nos permitiu determinar locais, datas e número de pessoas afetadas por acidentes. Por mais que ainda estejam incompletas, possivelmente temos a maior compilação de dados sobre acidentes de trânsito nas Filipinas até então.

Mas o que diferenciou essa abordagem é que, além de artigos, análises e visualizações dos dados captados, nos esforçamos para apresentar tudo isso às comunidades interessadas no tema, não só na internet, mas também em atividades de campo. Em meio a este processo, a análise de dados levou a ações de engajamento civil.

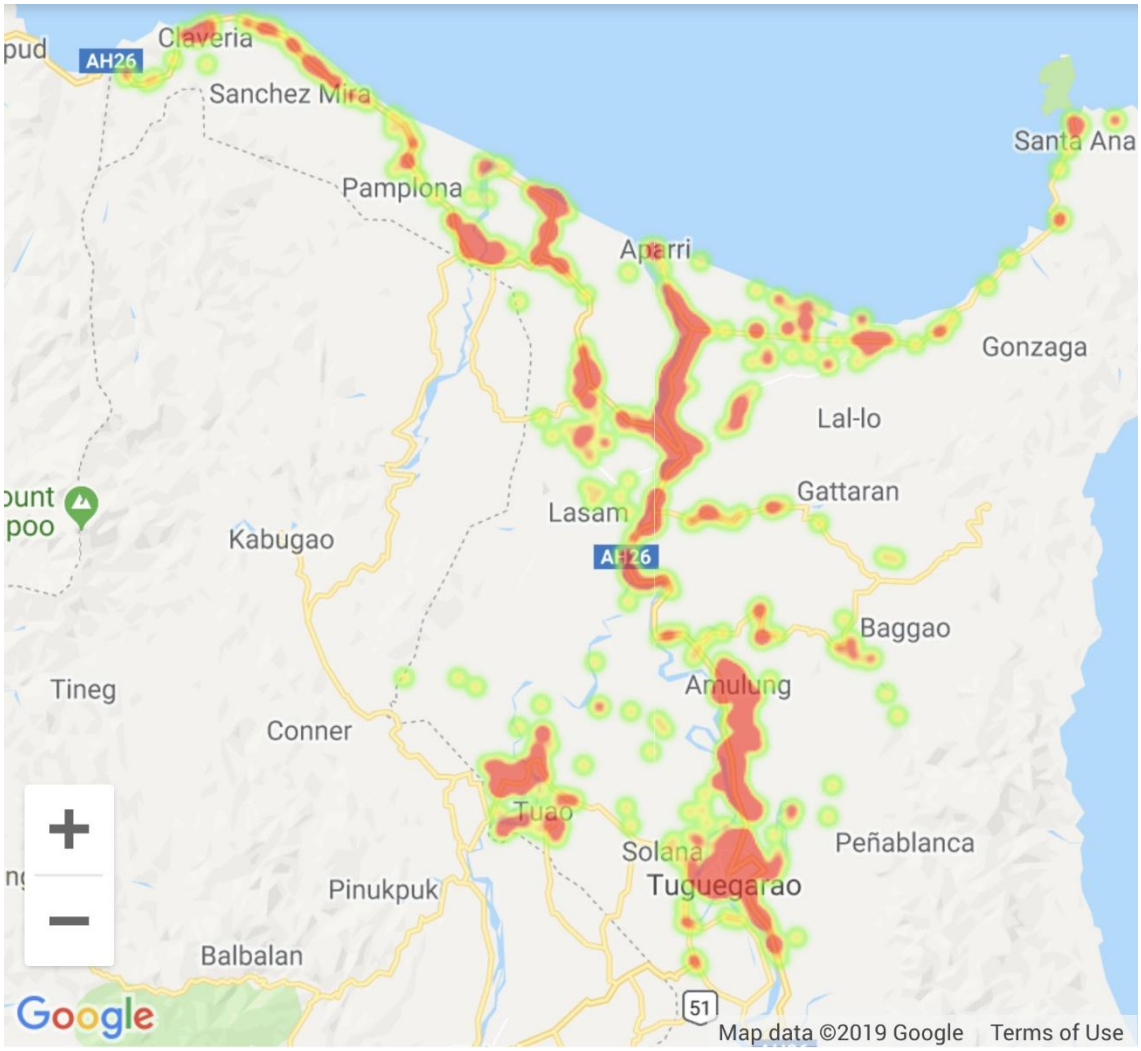
Uma história em particular se destacou em meio à cobertura, um relato detalhado sobre a província de Cagayan, a 600 quilômetros ao norte de Manila, a região mais afetada por mortes ligadas a acidentes de carro. Visitamos locais-chave da província para obtenção de dados relacionados a acidentes, bem como para a condução de entrevistas com vítimas, polícia local e agentes públicos.

Após este exercício, em junho de 2017, a *Rappler* realizou um fórum de conscientização no trânsito na capital da província, Tuguegarao, onde apresentamos nossas descobertas. O objetivo do fórum era educar o público a respeito de questões de segurança no trânsito e influenciar agentes públicos a tratarem da falta de políticas ligadas ao problema.[20]

---

[19] https://www.rappler.com/move-ph/issues/road-safety/171657-road-crash-numbers-data-sources.

[20] https://www.rappler.com/move-ph/issues/road-safety/171778-road-crash-incidents-cagayan-valley.



Além dos gráficos ilustrando os horários de pico em que ocorrem grandes acidentes em um dia, também apresentamos um mapa de calor, ao utilizar o Google Fusion Tables, mostrando os locais com maior número de acidentes em Cagayan.

As autoridades ali presentes atribuíram estes fatos, dentre outros, à ausência (ou falta) de faixas de pedestre. Ao verificar as escolas na cidade, notou-se que não havia faixas de pedestre na frente das mesmas. Após o fórum, foi conduzido um experimento social em que moradores traçaram faixas de pedestre, com giz, diante de uma escola. Agentes da lei queriam ver se os motoristas parariam diante das faixas enquanto estudantes as atravessavam. Posteriormente, a *Rappler* postou um vídeo sobre a experiência no Facebook.[21]

O vídeo atraiu muito interesse. Um leitor da *Rappler* que havia assistido ao vídeo entrou em contato para fornecer tinta, voluntariamente, para pintura de faixas de pedestre pela cidade. Meses depois, através de esforços combinados entre a administração local e voluntários, as escolas finalmente receberam suas faixas. O projeto de pintura foi completado

[21] https://www.rappler.com/move-ph/issues/road-safety/172432-cagayan-police-pedestrian-lane-chalk.

em 30 de setembro de 2017. Passado um ano, a cidade está prestes a aprovar legislação local sobre segurança no trânsito.

Este projeto mostrou que reportagens movidas a dados não acabam quando o editor clica em publicar. É uma prova de que a combinação de jornalismo de dados com engajamento de comunidades dentro e fora da internet pode levar a mudanças sociais e políticas positivas.

Aika Rey é repórter multimídia da Rappler.

Monitoramento de mortes de trabalhadores na Turquia

Pınar Dağ

No rastro do desastre da mina de Soma, na Turquia, em 2014, ficou claro que era extremamente difícil documentar as condições dos trabalhadores. Havia números discrepantes em relação à sindicalização e uma grande escassez de dados em relação à morte destes trabalhadores nas décadas anteriores. O que estava disponível, muitas vezes se encontrava desorganizado e faltavam detalhes. Queríamos abrir estes dados, lançar uma luz sobre as mortes de trabalhadores em outros setores.

Tendo isso em mente, um programador, um editor e eu desenvolvemos o Banco de Dados Aberto de Trabalhadores Falecidos da Turquia; um projeto público que coletou e verificou dados de diversas fontes, disponibilizando-os para acesso e uso de todos.[22] Na Turquia, ao menos 130 trabalhadores morrem todos os meses, por diversas causas. O principal objetivo do projeto era conscientizar o público a respeito destas mortes, sua frequência, e reconhecer publicamente as vítimas e as condições péssimas de trabalho com as quais lidavam. Ele consistia em mapas, gráficos e dados incorporáveis em diferentes formatos.[23] Cobria mortes de trabalhadores em mais de 20 setores diferentes, entre 2011 e 2014. Após a finalização do projeto, continuamos a relatar os incidentes através de monitoramento de mídia comum a cada mês. Um fato crucial: o banco de dados inclui o nome das empresas para as quais trabalhavam.

---

22 http://community.globaleditorsnetwork.org/content/open-database-deceased-workers-turkey-0.

23 http://platform24.org/veri-gazeteciligi/451/turkiyede-isci-olumleri-veritabani-hazirlandi.

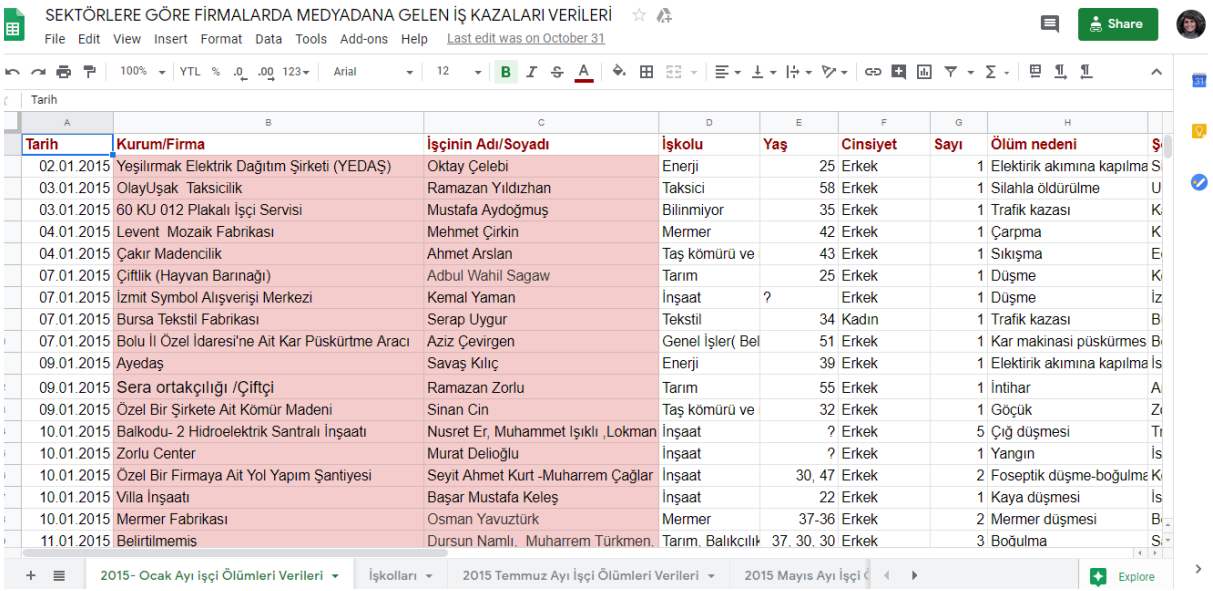


Figura 1: Planilha colaborativa de nomes de empresas baseada em monitoramento de mídia por meio do Google Alerts.

O projeto teve início em 2015. Começamos com a solicitação de pedidos de liberdade de informação e coleta de dados de algumas ONGs de confiança que, por sua vez, extraíam estes dados de diferentes fontes e liberavam para uso público. O primeiro desafio: não foi nada fácil conseguir os dados com base nos pedidos de liberdade de informação. Às vezes esperávamos duas semanas ou até mesmo quatro meses. E, então, uma complicação inesperada. Quando anunciamos o projeto, uma das fontes de nossos dados — İSİG Meclisi (Vigilância em Segurança e Saúde do Trabalho, em tradução livre) — não estava nada contente com a nossa aplicação destas informações.[24] Eles afirmavam que nosso projeto simplesmente republicava os dados que eles haviam reunido. Ao usar seus dados desta forma, pensavam que havíamos negligenciado o seu trabalho. Havíamos pedido permissão, e as informações fornecidas por eles eram abertas a todos, mesmo assim se opuseram ao nosso projeto. Chegaram ao ponto de não mais enviar seu boletim de dados com a mesma regularidade de antes. Sentimos que nosso projeto tinha conseguido disseminar ainda mais os dados fornecidos, de forma visualmente acessível e em formato pronto para download. Nos acusaram de "pornografar" as mortes de trabalhadores com nossas visualizações e tabelas de filtros.

Por mais que as histórias humanas sejam sempre essenciais, sentíamos que dados crus, não estruturados, também eram importantes para a apresentação de uma visão mais sistemática destas injustiças. Infelizmente, não foi possível convencer todos desta lógica e

24 http://guvenlicalisma.org/.

tivemos dificuldades de convencer as pessoas do valor presente em práticas de compartilhamento de dados colaborativos. Acabamos publicando o número mensal de mortes de trabalhadores ao comparar dados oficiais reunidos por meio de pedidos de liberdade de informação com a lista que coletamos através de monitoramento próprio, verificando mês a mês. Depois da realização deste projeto, as instituições às quais solicitamos pedidos de liberdade de informação passaram a compartilhar e apresentar dados de forma mais estruturada. Isso significa que atingimos uma de nossas metas: tornar estes dados mais acessíveis. No final, o projeto foi finalista no Data Journalism Awards de 2015.

Pınar Dag é professor de Jornalismo de Dados na Universidade de Kadir Has

Reunindo Dados

Construindo seu próprio conjunto de dados: crimes com armas brancas no Reino Unido

Caelainn Barr

No começo de 2017, dois de meus colegas, Gary Younge e Damien Gayle, vieram falar comigo na redação do *The Guardian*. Eles queriam examinar as informações em torno de crimes cometidos com armas brancas. Por mais que não faltassem textos detalhados sobre as mortes de vítimas destes crimes, bem como relatos sobre a caçada a suspeitos e matérias sobre julgamentos e condenações destes criminosos, ninguém ainda havia observado os homicídios como um todo.

Meu primeiro questionamento foi o seguinte: quantas crianças e adolescentes haviam sido mortos por armas brancas nos últimos anos? Parecia uma pergunta simples, mas assim que passei a vasculhar os dados em torno disso, ficou claro que ninguém tinha essa resposta. Os dados existiam e estavam em algum lugar, mas não em domínio público. Tinha diante de mim duas opções, desistir ou criar um conjunto de dados do zero, baseado nas informações às quais tinha acesso e poderia reunir e verificar eu mesmo. Foi assim que escolhi montar meu conjunto de dados.

Por que montar seu próprio conjunto de dados?

O jornalismo de dados não deve se basear somente em conjuntos de informações já existentes. Fato é que não faltam motivos para criar seus próprios dados. Há grande riqueza de informações em dados cuja publicação não é rotineira ou que, em certos casos, nem chegam a ser coletados.

Ao criar seu próprio conjunto de dados, cria-se uma série ímpar de informações, uma fonte única, com a qual explorar a sua história. Os dados e histórias subsequentes têm grandes chances de serem exclusivas, uma vantagem perante outros repórteres. Conjuntos de dados únicos também podem ajudar na identificação de tendências que passaram despercebidas por especialistas e autoridades.

Dados são fonte de informação para o jornalismo. A base para uso destes dados no contexto jornalístico é o pensamento estruturado. Para que estas informações sejam usadas com o máximo de seu potencial, o jornalista deve pensar estruturalmente no início do projeto: que história quero poder contar e do que preciso para poder contá-la?

A chave para a criação bem-sucedida de um conjunto de dados para sua história reside em uma abordagem estrutural do que se quer contar e verificar cada fonte de dados com um senso jornalístico de curiosidade. A construção de seu conjunto de dados engloba muitas das habilidades essenciais do próprio jornalismo de dados: pensamento estruturado, narrativa planejada e criatividade na hora de buscar informações. Além do que, pode ser feito com ou sem habilidades em programação. Se você consegue digitar uma planilha e organizar tabelas, já é um passo no desenvolvimento das habilidades básicas do jornalismo de dados.

Não que a prática em si seja sempre simples e direta. Projetos de dados sólidos e minuciosos podem ser complexos e demandar muito tempo, mas com algumas práticas-chave é possível criar uma base forte no uso de dados para contar histórias.

Um passo a passo para a construção do seu conjunto de dados

Planeje o necessário

O primeiro passo na geração ou coleta de dados para sua análise consiste em avaliar o que é necessário e se é possível obtê-lo. No começo de qualquer projeto, vale a pena rascunhar o que se espera e que história tentará ser contada, onde você crê que os dados estejam, quanto tempo levará para consegui-los e onde estão os possíveis gargalos. Este rascunho ajudará a avaliar quanto o processo todo demorará e se o resultado valerá o esforço empreendido. Também pode servir como material de apoio, em um estágio mais avançado.

Considere o foco

No começo de um projeto baseado em dados em que os dados ainda não existem, devemos nos perguntar qual é o foco da história a ser contada. É primordial saber o que os dados devem conter, já que isso define quais perguntas poderão ser feitas a estes. Os dados só podem responder as perguntas baseados na informação que contêm, o que faz desse passo essencial. Logo, para criar um conjunto que atenderá suas necessidades, seja bastante claro com respeito ao que você gostaria de explorar e quais informações são necessárias para isso.

Onde podem estar os dados?

O próximo passo é considerar onde estes dados podem estar, em qualquer formato. Uma forma de fazer isso é refazer seus passos. Como você chegou à conclusão de que há uma história a ser contada aqui? De onde veio essa ideia? Há uma fonte de dados em potencial por trás dela?

A pesquisa também ajudará a esclarecer o que de fato existe sobre o assunto, então é preciso verificar todas as fontes de informação sobre o tópico de interesse, bem como conversar com acadêmicos, pesquisadores e estatísticos que trabalham ou coletam estes

dados. Fazer isso ajudará a identificar deficiências e possíveis gargalos no uso dos dados. Também pode ajudar a fomentar ideias sobre outras possíveis fontes e formas de se obter informações. Toda essa preparação antes da criação do seu conjunto de dados será inestimável caso precise lidar com agências governamentais complicadas ou decida por outra abordagem na coleta das informações.

Questões éticas

Ao planejar e buscar fontes para a história, precisamos pensar nas questões éticas envolvidas. O mesmo vale para dados. Ao criar um conjunto de dados, precisamos levar em conta se a fonte e a metodologia usadas são as mais precisas e completas possíveis.

Essa lógica também se aplica à análise —examinar a informação sob os mais diversos ângulos e não forçar os dados a dizerem algo que não reflita a realidade. Ao apresentar a história, prepare-se para ser transparente sobre fontes, análise e limitação dos dados. Pensar nisso tudo ajudará a criar um material mais robusto e a criar uma relação de confiança com o leitor.

Obtenção de dados

Após a identificação de uma fonte em potencial, o próximo passo é obter estes dados. Isso pode ser feito manualmente, ao inserir dados em uma planilha; transformando informações presas dentro de PDFs em dados estruturados que podem ser analisados; buscando documentos junto a uma fonte humana ou por meio da Lei de Acesso à Informação; com o uso de programação para extração de dados de documentos ou páginas web; ou pela automação para a captura de dados por meio de API.

Seja gentil consigo mesmo! Não sacrifique a simplicidade à toa. Tente buscar a forma mais direta de inserir informações em um conjunto de dados para análise. Se possível, faça de todo esse processo replicável, o que ajudará a verificar seu trabalho e na inclusão de dados em um estágio posterior, se necessário.

Ao obter estes dados, retorne ao rascunho do projeto e faça a seguinte pergunta: eles me permitirão explorar este tópico por inteiro? Contêm as informações que levarão aos pontos que mais me interessam?

Estrutura

A principal diferença entre informações contidas em uma pilha de documentos de papel e um conjunto de dados é a estrutura. Estrutura e repetição são a chave para a criação de um conjunto de dados limpo, pronto para análise.

O primeiro passo é se familiarizar com estes dados. É preciso se perguntar o que está contido ali e o que aquilo te permitirá dizer. O que não poderá ser dito com aqueles dados? Há outro conjunto de informações que poderia ser combinado àquele? É possível dar passos na criação deste conjunto de dados que o permitirá ser combinado com outros?

Pense em como deve parecer esse conjunto ao final do processo. Leve em consideração as colunas ou variáveis que gostaria de analisar. Busque inspiração na metodologia e na estrutura de conjuntos semelhantes.

Seja o mais abrangente possível no começo, considerando todos os dados a serem coletados, e depois reduza, avaliando o que é necessário para a história a ser contada e quanto tempo demorará para a obtenção de todas as informações. Certifique-se de que os dados coletados compararão semelhantes. Escolha um formato e siga nele, te poupará muito tempo ao final! Pense também na dimensão do conjunto de dados que está sendo criado. Horários e datas possibilitarão a análise de informações ao longo do tempo; informações sobre localização possibilitarão mapear os dados em busca de tendências espaciais.

Monitore e verifique seu trabalho no decorrer do processo

Tome nota das fontes usadas para a criação do conjunto de dados e sempre tenha uma cópia dos documentos e dados originais. Crie um dicionário de metodologia e dados para ter sempre à mão suas fontes, como os dados foram processados e o que está contido em cada coluna. Isso ajudará a sinalizar possíveis questionamentos e resolver erros em potencial durante a coleta e a análise de dados.

Não suponha nada e verifique cada descoberta com mais investigações. Não hesite em consultar especialistas e estatísticos a respeito de sua abordagem e descobertas. O ônus de tornar seu trabalho sólido é ainda maior quando seus dados forem consolidados, então é preciso garantir que cada dado, análise e texto estejam corretos.

Estudo de caso: além da lâmina

No começo de 2017, a equipe de projetos de dados se juntou a Gary Younge, Damien Gayle e à equipe de jornalismo comunitário do *The Guardian* em um esforço para documentar a morte de cada criança e adolescente mortos por uma arma branca no Reino Unido.

De forma a compreender a questão e explorar temas-chave em torno de crimes envolvendo armas brancas, sobretudo facas, a equipe precisava de dados. Queríamos saber o seguinte: quem são esses jovens que estão morrendo por conta de esfaqueamentos no Reino Unido? São crianças ou adolescentes? De que gênero ou etnia? Quando e onde esses jovens estavam quando foram mortos?

Após conversarmos com estatísticos, policiais e criminologistas, ficou claro que os dados existiam, mas não estavam disponíveis ao público. Tentar formular uma resposta aos questionamentos consumiria muito de meu trabalho no decorrer do próximo ano.

Os dados que buscava estavam em posse do Ministério do Interior, agrupados em um conjunto conhecido como Índice de Homicídios. As informações chegavam ao Ministério do Interior através das forças policiais da Inglaterra e do País de Gales. Havia duas possíveis rotas para obtenção das informações — enviar uma solicitação de liberdade de informação ao Ministério do Interior ou entrar em contato com as autoridades policiais, uma a uma. De forma a cobrir todas as possibilidades, fiz ambos. Assim sendo, conseguimos dados históricos até o ano de 1977.

Para monitorar as mortes no ano corrente, precisaríamos contabilizá-las quando de sua ocorrência. Como não havia dados públicos ou consolidados de forma centralizada, optamos por monitorar as informações nós mesmos, através de boletins de ocorrência, clippings de notícias, Google Alerts, Facebook e Twitter.

Definimos o que queríamos saber: nome, idade e data do ocorrido certamente eram coisas que queríamos registrar. Mas havia outros aspectos das circunstâncias em torno dessas mortes que não eram tão óbvios. Discutimos o que achávamos saber sobre os crimes com armas brancas — quase sempre envolvia homens, com um número desproporcional de homens negros como vítimas. Para verificarmos estas suposições, incluímos colunas para gênero e etnia. Verificamos, também, todos os números e detalhes relacionados com as autoridades policiais do Reino Unido. Em algumas ocasiões, encontramos casos que haviam passado batido por nós, o que permitiu cruzar as descobertas antes da publicação.

Após termos diversos pedidos de liberdade de informação negados e longos atrasos, os dados acabaram por serem liberados pelo Ministério do Interior. Ali constavam idade, etnia e gênero de todas as pessoas mortas por facas, divididas por distrito policial, ao longo de quase 40 anos. Tudo isso, combinado ao nosso conjunto de dados atual, permitiu que observássemos quem vinha sendo morto e a tendência no tempo.

Os dados revelaram que 39 crianças e adolescentes morreram em crimes envolvendo facas na Inglaterra e no País de Gales em 2017, um dos piores anos (considerando mortes de

jovens) em quase uma década. Os números levaram a preocupações sobre uma crise pública de saúde oculta em meio a anos de cortes no orçamento policial.

Os dados também colocaram em xeque suposições comumente feitas sobre quem este tipo de crime afeta. Descobriu-se que na Inglaterra e no País de Gales, nos dez anos entre 2005 e 2015, um terço das vítimas era negra. Porém, para além da capital, os esfaqueamentos entre jovens não ocorriam predominantemente entre garotos negros, já que no mesmo período menos de uma a cada cinco vítimas fora de Londres era negra.

Por mais que os crimes envolvendo facas sejam tópico constante de debate, os números não eram disponibilizados imediatamente para políticos e legisladores, o que levanta questões sobre como criar políticas eficazes quando detalhes relevantes sobre o tema não são de fácil acesso.

Estes dados foram a base de nosso projeto premiado, que mudou a forma como o debate sobre crimes com facas é realizado. Nada disso seria possível sem a criação de nosso próprio conjunto de dados.

Caelainn Barr é editora de projetos de dados da Guardian News e Media.

Contando histórias por trás de números sem esquecer da questão do valor

Helen Verran

Em 2009, a contribuição ao PIB australiano de transações em que o estado adquiriu intervenções ambientais para melhorar o valor de ecossistemas, compradas de proprietários rurais em Corangamite, uma NRMR, foi calculada em 4,94 milhões de dólares australianos.

O número mencionado aqui veio à tona em uma nota à imprensa emitida pelo governo do estado australiano de Vitória no ano de 2009. A nota anunciava o sucesso de um investimento feito pelo estado em iniciativa de preservação ambiental em uma das cinquenta Regiões de Manejo de Recursos Naturais da Austrália (NRMR, na sigla em inglês) A região ambiental administrativa, constituída por planícies de basalto cobertas por grama que vai de leste a oeste na parte sul de Vitória, se chama Corangamite, um termo aborígene que substituiu o nome dado pelos pastores britânicos que invadiram o país em meados do século XIX, a partir da Tasmânia. Haviam dado à região o nome de "Australia Felix" e começaram a derrubar árvores. Os invasores, que acabaram por virar proprietários de terras, em menos de um século se tornaram uma espécie de pequena nobreza colonial. Em 2008, durante a operação do Programa EcoTender, na região de Corangamite, o governo de Vitória comprou serviços ecossistêmicos dos descendentes daqueles invasores em leilões. Em 2009, calculou-se que estas transações renderam ao PIB australiano algo por volta de 4,94 milhões de dólares australianos. A primeira vez que me deparei com este valor foi na nota à imprensa mencionada anteriormente.

Duvido que qualquer jornalista tenha dado prosseguimento à cobertura deste caso, incluindo o valor citado, que de forma alguma seria considerada uma notícia das mais relevantes. No contexto de uma nota à imprensa, a citação de um número específico é uma espécie de reafirmação. Os registros nacionais são dados reais relevantes, e se esta intervenção regional do governo conta com um valor específico que contribuiu com a economia nacional, então claramente esta intervenção é algo positivo. A especificação do valor puxa para si um quê de verdade em relação às melhorias pelas quais as intervenções do governo vêm passando. A implicação direta é que tal prática leva à boa governança ambiental. É claro que para o que o valor (4,94 milhões de dólares australianos) aponta, o que afirma indexar implicitamente, não interessa tanto assim. Esse número parece apontar para algo que pode ser avaliado de tal forma, o que já é o suficiente para fins de reafirmação.

Minha narrativa ligada a este número oferece um relato absurdamente detalhado dos meios sociotécnicos pelos quais este surgiu. Tal relato tem o perturbador efeito de revelar que este número, tão banal em meio à nota em que surgiu, não passa de uma péssima tentativa de acobertamento. O buraco é mais embaixo. Antes de começar a contar minha história e articular a natureza e a profundidade do buraco em questão — profundo o suficiente para engolir qualquer tentativa de apreciação de valores, mesmo algo tão banal assim —, deixe-me responder preventivamente algumas das perguntas que podem ocorrer aos leitores deste livro.

Primeiro, reconheço que contar o que há por trás de um número em vez de buscar meios para promover a visualização do que tal número significa dentro de um contexto em específico, é uma abordagem bastante inusitada se tratando do jornalismo de dados contemporâneo. Consigo imaginar qualquer jornalista duvidando que uma narrativa como esta funcionaria. Talvez caiba lembrar que não é uma questão de escolher isso ou aquilo e que trabalhar combinando narrativa e visualizações na decodificação e na interpretação de problemas é uma bela maneira de se transmitir ideias. Ao apresentar tal combinação, é dever dos jornalistas lembrarem que há dois pontos de diálogo ao se misturar narrativa e visual. Pode-se proceder como se o visual estivesse embutido dentro da narrativa, no caso você dialoga *com* o visual que representa ou ilustra algo na história. Ou, pode-se seguir como se a narrativa estivesse embutida no visual, no caso dialoga-se *de dentro* do diagrama. Esta é uma estratégia menos comum em jornalismo de dados, mas consigo imaginar que a história que conto aqui possa ser empregada de tal forma. Claro, alternar entre estes dois pontos dentro de um único material talvez seja a estratégia mais eficaz.

Segundo, pode parecer estranho contar a história por trás de um número específico quando o que realmente pode ter alguma relevância em termos de tomada de decisão e formulação de políticas, e que de fato interessa aos jornalistas de dados, é o que pode ser feito com conjuntos de dados na mobilização deste ou aquele algoritmo. Toda essa preocupação pode levar você a questionar as relações entre números e conjuntos de dados. A resposta a esses questionamentos é relativamente direta e pouco interessante. Há diversos números em um conjunto de dados, na relação de um para muitos, embora estes mesmos números estejam reunidos em ordens muito precisas. A pergunta mais interessante trata da relação entre números e algoritmos. Minha resposta seria a seguinte: ao passo que algoritmos mobilizam um protocolo que elabora como lidar com relações embutidas em uma base de dados, os números expressam um protocolo que delineia como lidar com relações de caráter coletivo. Numeração é uma forma de criar algoritmos, e vice-versa.[25] Poderíamos afirmar que os

[25] A noção de que números e algoritmos têm alguma semelhança possivelmente é nova para muitos leitores, acostumados a encararem números como 'abstrações'. Minha visão (incomum) dos números os vê como entidades semióticas compostas por material extremamente comum que habitam o aqui e o agora. Para uma visão de diferentes protocolos no contexto de mobilização de relações dentro de um único momento de caráter coletivo, consultar Watson (2001).

números estão para os algoritmos como as sementes estão para as plantas, algo pode germinar ali. Misturando ainda mais as metáforas, são como ovo e galinha. Por mais que existam características sociotécnicas variadas para a geração de valores enumerados por meios analógicos (através da mistura de recursos cognitivos, linguísticos e gráficos) de enumeração convencional, como ensinados a crianças do ensino fundamental, e criação de valores enumerados através de computação digital, é a semelhança que nos importa aqui: 4,94 milhões de dólares australianos foram gerados através de algoritmo e este número expressa uma série em particular de relações inclusas em um conjunto de dados em específico, ainda assim se apresenta como um número e nada mais.

Agora, de volta à minha história. O relato íntimo da geração de números que conto aqui como história permitirá a um jornalista reconhecer que o conto de fadas que o governo tenta empurrar sorrateiramente com sua nota à imprensa não é uma questão das mais óbvias. Acreditamos que ir fundo na questão política seria mais apropriado. Os detalhes a respeito de como se chegou a tal número revelam que este programa de intervenção ambiental de parceria público-privada consiste em termos o estado pagando proprietários de terra muito ricos para que desempenhem um trabalho que valorizará ainda mais suas propriedades. Uma pergunta que pode ser suscitada com base em minha história é: como um jornalista poderia celebrar ou expor este número, ao agir de boa fé? Ao fim de tudo, sugerirei que esta não é a pergunta certa a ser feita.

A história por trás do número

Qual a série de processos sociotécnicos pelos quais o valor de serviços ecossistêmicos surge no contexto deste programa de parceria público-privada de forma que este seja negociado entre governo, no papel de comprador, e proprietário no papel de fornecedor? E como exatamente o valor econômico dessa negociação contribui para um aumento marginal nos ganhos totais da atividade econômica australiana como um todo, o PIB australiano? Trato destas duas questões com um passo a passo que descrevo o que é necessário para que um proprietário de terras crie um produto, no caso o ‘valor do serviço ecossistêmico’, competitivo o bastante para participar de pregão público em busca de contrato para fornecer ao governo o ‘valor do serviço ecossistêmico’. O trabalho sujo que dá à luz este produto envolve chafurdar na lama, plantar árvores, consertar cercas e, em linhas gerais, buscar reparar o dano causado à terra pelos avôs dos proprietários, digamos, que em ato de ganância desnudaram o país de árvores, cultivando espécies consumidoras de grandes quantidades de água, em busca de mais grãos ou mais lã e maiores fortunas para a família. O valor destes serviços ecossistêmicos é gerado através da intervenção em processos ambientais.

Este mesmo valor, do produto a ser negociado, tem suas raízes no trabalho de servidores públicos dentro de um departamento do governo de Vitória (neste caso, o

Departamento de Sustentabilidade e Ambiente, DSE na sigla em inglês). Coletivamente, estes servidores são responsáveis pela escolha de quais regiões do estado receberão estas licitações. Ao longo deste processo, a *EnSym*, plataforma de modelagem de sistemas ambientais, é essencial. Este sistema é uma maravilha, conhecendo a 'natureza lá fora' como nenhum outro cientista jamais a conheceu. É capaz de gerar representações precisas e focadas, possivelmente de um dia para o outro.

Este software foi desenvolvido pela equipe da EcoMarkets e incorpora ciência, regulamentações, métricas e informações geradas dentro do DSE, bem como diversos modelos científicos de renome internacional e nacional.

A *EnSym* conta com três ferramentas principais: a 'Ferramenta de Avaliação de Locais', para trabalho de campo; a 'Ferramenta de Preferência de Panorama', para priorização de recursos e construção de métricas; e 'BioSim', para planejamento de captação.[26]

A priorização e o mapeamento de áreas em que o estado realizará licitações, a especificação e a quantificação de benefícios ambientais, os valores ecológicos que podem ser aprimorados através de medidas de preservação e reflorestamento em campo, tudo é registrado em formato numérico. Tudo isso representa propriedades do ecossistema lá fora. O software pode fazer mais que isso, pode gerar um roteiro para a intervenção humana. Assim como o roteiro de uma peça clama por produtores, este também. Desta forma, enquanto o roteiro ganha vida, a natureza lá fora parece se aproximar. Deixa de ser uma natureza distante e se torna uma infraestrutura para vidas humanas, uma infraestrutura a receber algumas cutucadas nossas, com o objetivo de consertar o encanamento.

Quando o roteiro de uma produção coreografada de esforço humano coletivo está pronto, o próximo passo dado pelo governo é captar o interesse de proprietários na área do projeto. Em resposta a estas manifestações de interesse, um oficial do governo visita todas as propriedades. Pode-se imaginar este oficial levando o roteiro gerado pela *EnSym* consigo a um lugar de verdade em determinado momento. Ele ou ela terá uma tarefa de tradução formidável pela frente.

O agente de campo avalia possíveis locais que poderão servir de palco para a produção do roteiro. O objetivo é aprimorar a geração dos serviços ecossistêmicos específicos, então o agente precisa avaliar as chances de que as ações especificadas em um lugar em particular levarão a um aumento na oferta de serviços do ecossistema, aumentando o valor do serviço ecossistêmico gerado por aquela propriedade, adicionando os muitos incrementos gerados neste programa de intervenção pelo estado como um todo. Juntos, o

[26] http://www.dse.vic.gov.au/ecomarkets.

proprietário e o agente do governo criam um plano. Na negociação decorrente, um plano formalizado de manejo para lotes específicos é feito. O agente de campo desenvolve este plano em termos passíveis de contrato. Aos proprietários, cabe especificar, em detalhes, como o plano será posto em prática. Logo, um produto que assume a forma de um 'valor de serviço ecossistêmico' em particular é projetado e especificado como uma série de tarefas específicas a serem completadas em um período de tempo determinado: tantas mudas de espécie tal, plantadas de tal forma, neste exato local deste campo em específico, cercado de forma a criar um lote de preservação de tantas dimensões, usando estes materiais.

Os custos destes serviços, especificados pelo estado, são calculados pelos proprietários, certamente considerando um generoso pagamento pela mão-de-obra. Eles inventam o preço a ser pago pelo governo na compra deste produto, um 'valor de serviços ecossistêmicos'. Aqui, eles especificam o montante que estão dispostos a receber para realizarem as tarefas solicitadas, de forma a entregar o valor em serviços ecossistêmicos na data estipulada. Documentos relevantes são enviados ao governo, dentro de envelopes selados.

Então, como funciona o leilão subsequente? Mais uma vez, a *EnSym* entra em cena, para avaliar as ofertas. Não só uma conhecedora da natureza lá fora — este 'lá fora' reimaginado como infraestrutura —, roteirista de intervenções, a *EnSym* é também uma observadora renomada capaz de avaliar os lances feitos para a produção de seu roteiro, não muito diferente do que a Warner Bros. faria na hora de produzir um filme. Os lances são calculados com base em um 'índice de benefícios ambientais', com o preço proposto pelo proprietário. Suponhamos que o governo compre o produto que ofereça o maior 'índice de benefícios ambientais' por custo unitário.

Avaliação de lances. Todos os lances são avaliados objetivamente com base em:

• mudança estimada em resultados ambientais

• valor da mudança em resultados ambientais

• valor dos recursos afetados por estas mudanças (significância)

• custo (preço determinado pelo proprietário).

Fundos, então, são alocados com base na relação entre custo e benefício.

Quando os resultados do pregão são anunciados, os selecionados assinam contrato baseado no planejo e no calendário de trabalhos enviados com definições de organização temporal e espacial. Após assinatura dos documentos, implementam-se sistemas de relatório e o pagamento pode ter início.

“A DSE envia pagamentos para os proprietários inscritos após recebimento de fatura. Os pagamentos estão sujeitos ao progresso satisfatório das ações especificadas no Contrato de Manejo”.

Isso é bom, certo?

Acabo de dar uma explicação precisa sobre como comprar e vender valores de serviços ecossistêmicos. Voltando à nota à imprensa. Uma leitura rápida da declaração talvez deixe o leitor com a impressão de que os 4,94 milhões de dólares australianos se referem ao capital natural adicional gerado por este programa. Em um primeiro momento, estes 4,94 milhões parecem ser um pequeno incremento no valor do capital natural australiano, obtido através deste programa. Isso está errado. 4,94 milhões não se referem ao valor do capital natural. Abaixo explico a que este número se refere. Nesse momento, gostaria de focar no produto que foi negociado neste pregão. Este é o problema que quero manter por perto.

Quero questionar o valor do incremento em “valor de serviços ecossistêmicos” obtido por este complexo e dispendioso programa governamental. Uma leitura cuidadosa a respeito dos detalhes do trabalho pelo qual este valor surge revela que, em momento algum do processo, este número foi citado ou especificado. O produto de tão rigorosa transação *inexiste*. E o pior de tudo é que não haveria como ser de outro jeito. O programa é um exercício elaboradíssimo de contabilidade para distribuição de dinheiro. Quando isso fica claro para quem vê de fora, também fica óbvio que o que esta prática *é* nunca foi de fato escondido. No final das contas, este programa é um meio legítimo de transferir dinheiro dos cofres estatais para as mãos de proprietários de terras privadas.

Reconheço que este é um programa de governança ambiental em uma democracia parlamentarista liberal em que a tecnologia social do partido político é crucial, então deixe-me brincar de analista político um pouco. Corangamite é um campo de disputa eleitoral conhecido por alternar entre esquerda (Partido dos Trabalhadores) e direita (Partido Liberal) na escolha dos representantes da região no Parlamento de Vitória. É claro que o interesse de qualquer governo, à esquerda ou à direita, é apelar ao seu eleitorado. Não há forma melhor de fazer isso senão a descoberta legítima de como transferir recursos do estado para as contas bancárias dos constituintes. Sob esta ótica, o fato de que não há como atrelar um número ao valor do produto negociado entre estado e proprietários aqui pouco importa.

Deixem-me resumir. Do ponto de vista econômico, este programa é justificado por meio da geração de valor em serviços ecossistêmicos. Descrito assim, é material para um bom artigo jornalístico. Dinheiro do contribuinte é usado para melhorar condições ambientais e plantar árvores de forma a compensar a geração excessiva de dióxido de carbono em Vitória. Há uma problemática, porém, visto que o aumento de valor do capital natural do

estado não pode ser numerado, articulado como um número, mesmo se tratando de produto vendido e comprado. Por mais que ainda haja certas complicações técnicas, isso, claramente, *é algo bom*.

Mas, da mesma forma, ao aplicar vieses econômicos diferentes, este programa pode ser descrito como financiamento do plantio de árvores para valorização de terras privadas. É uma forma de intervir a fim de corrigir danos causados por programas anteriores do governo ao subsidiar o trabalho de limpeza de terras através do qual, muito provavelmente, os avós dos proprietários lucraram, criando um benefício a ser desfrutado, mais uma vez, por estes mesmos proprietários. Sob esta ótica, a política governamental aplicada por meio deste pregão não passa de um programa caro para distribuição legítima de dinheiro do contribuinte. *O que é péssimo,* obviamente.

Respeito a números e algoritmos — lidando com questões de valor

O que resta ao jornalista fazer? Como acadêmica e não repórter, consigo responder a esta pergunta de maneira vaga. No começo deste texto, afirmo que o número citado na nota à imprensa é uma tentativa das mais fracas de acobertar a verdadeira questão em jogo. Em minha opinião, valores carregam consigo problemas morais impossíveis de serem deixados de lado por muito tempo. A pergunta certa a se fazer, creio, é "como um jornalista de dados pode lidar com este problema moral?".

Primeiro, cabe esclarecer o que são esses 4,94 milhões. O que este número significa? De onde vem este valor monetário? Ele é descrito da seguinte forma em artigo acadêmico crítico sobre o programa EcoTender:

"Sob este modelo baseado no mercado, o valor econômico dos serviços ecossistêmicos é gerado quando os custos unitários relacionados ao contrato de preservação custam menos que o preço unitário pago aos vencedores do pregão. Ao passo que parte deste valor econômico [para os fornecedores] se perde pela não comercialização de commodities, a restrição na participação de proprietários garante que haverá um aumento líquido no valor [econômico] gerado na condução do pregão".

Sob a modelagem econômica desta prática, supõe-se que os proprietários das terras calcularão os custos envolvidos na produção do roteiro governamental para intervenção na natureza enquanto infraestrutura de forma eficiente — levando a um desempenho mais eficiente no funcionamento da infraestrutura natural. Claro, presume-se que o proprietário lucrará, por mais que também seja possível que o mesmo incorra em prejuízo por conta de algum erro de cálculo, o que pouco interessa ao governo no papel de comprador do valor gerado pelo trabalho do proprietário.

O que interessa ao governo é como esta transação pode ser articulada de maneira correta. Um problema e tanto quando consideramos que o produto negociado existe somente dentro de um pregão. A solução para este formato problemático é a elaborada, complexa e complicada tecnologia do sistema de contas australiano. Ao estabelecer um mercado para o valor de serviços ecossistêmicos, o governo quer mostrar que está fazendo uma diferença no meio ambiente. E as contas nacionais são o lugar conveniente em que isto pode ser mostrado em termos monetários. O 'índice de benefícios ambientais', o valor específico baseado na compra de um produto em especial pelo governo, o valor de um serviço ambiental, é efêmero. Este existe somente dentro do pregão, em um único instante. Apesar da dificuldade envolvendo o formato de sua existência, em um artifício dos mais engenhosos, os meios para se comprar e vender algo tão efêmero são possíveis, e evidências desta atividade econômica podem ser incorporadas às contas nacionais, embora alguns economistas tenham sérias ressalvas quanto à sua precisão.[27]

4,94 milhões de dólares australianos, um valor remotamente ligado à ação do programa EcoTender e à natureza que busca melhorar. Claro, se o governo declara que seus programas melhoraram uma porção da natureza que havia sido danificada, tem que se dar um jeito de indicar a extensão desta melhoria. Qualquer número é melhor que nada em um caso como este, ou ao menos é o que parece. E, certamente, trata-se de um número bom, bacana. Um número ruim, negativo, definitivamente de conhecimento dos contadores do governo, digamos o custo deste programa, não serviria. Por que discutir este número tão deslocado? Não estaríamos indo longe demais? Qual o problema com um truque relativamente fácil de se perceber? Minha preocupação aqui é que o uso errado deste número parece deliberado. É uma falta de respeito com números, uma recusa em reconhecer que números e algoritmos sempre trazem problemas consigo. Joga um protocolo no lixo.

Minha história em torno de um número encontrado em visita a um site governamental acabou por revelar, sem ambiguidade alguma, um programa público que causa bem e mal simultaneamente. Aquele truque ao citar este número (o valor exato de 4,94 milhões de dólares australianos divulgado à imprensa), também descoberto no decorrer da história, acaba por apontar para uma questão que sempre está pairando no ar: valores enquanto campo de tensão moral e problemas.

Seria a conclusão, aqui, que valores são problemas morais que não podem ser acobertados por muito tempo? A teoria de valor é um tópico extenso com raízes em todas as tradições filosóficas, e este é um buraco fundo o suficiente para que eu me negue a adentrá-lo. Apenas digo que afirmações, ouvidas muitas vezes ao longo dos últimos trinta anos, de que a mão invisível do mercado doma a questão moral que acompanha valores são exageros

---

[27] Stoneham et al. (2012).

dos mais perigosos. Os mercados podem encontrar uma maneira de domar o valor, de forma temporária e efêmera, como meu artigo deixa claro. Mas o problema em torno do valor sempre volta. Lidar com *esse problema* é dever do jornalista de dados.

Aqui deixo algumas sugestões sobre como um jornalista pode tratar, respeitosamente, números e algoritmos como protocolos. Quando você se vê diante de algo que parece não ter problemas na superfície, nenhum tipo de tensão moral, mas sente que há algo ali, é bom ficar de olho. Familiarize-se com os números envolvidos, aprenda a pensar com um número que chame a sua atenção. Encontre maneiras de alargar os buracos que estes números tentam cobrir. Cultive formas respeitosas de lidar com números e algoritmos, botando a curiosidade em prática de forma disciplinada. Reconheça que números têm naturezas pré-estabelecidas e habilidades especiais que vêm à tona no encontro, que as circunstâncias das séries de práticas que os geraram importam. Tenha certeza de que, ao dominar estas técnicas, surpresas te aguardam. Há muito de interessante dentro dos números no momento em que surgem.

Helen Verran é professora da Universidade Charles Darwin.

Referências

Departamento de Sustentabilidade e Meio Ambiente, Governo de Vitória. EcoMarkets. EcoTender e BushTender. 2008. Disponível em: http://www.dse.vic.gov.au/ecomarkets

ROFFE, Jon. *Abstract Market Theory*. Palgrave Macmillan, 2015.

STONEHAM, Gary et al. *Creating physical environmental asset accounts from markets for ecosystem conservation*. Ecological Economics 82, 2012, p. 114–122 e 118.

VERRAN, Helen. *Two Consistent Logics of Numbering', Science and an African Logic*. Chicago University Press, 2001.

VERRAN, Helen. *Enumerated Entities in Public Policy and Governance in Mathematics, Substance and Surmise*', Ernest Davis ePhilip Davis (editores), (Springer International Publishing Switzerland, 2015, DOI 10.1007/978-3-319-21473-3_18.

VERRAN, Helen; WINTHEREIK Brit R. *Innovation with Words and Visuals. A Baroque Sensibility*. In: LAW, John; RUPPERT, Evelyn (ed.). *Modes of Knowing*. Mattering Press: 2016.

WATSON, Helen. *Investigating the Social Foundations of Mathematics: Natural Number in Culturally Diverse Forms of Life*. Social Studies of Science 20, 1990, p. 283-312.

Documentando conflitos por terra em toda a Índia

Kumar Sambhav Shrivastava e Ankur Paliwal

Terra é um recurso escasso na Índia. O país detém somente 2,4% das áreas terrestres do planeta, apesar de conter 17% da população mundial. Uma das economias de mais rápida expansão do mundo, o país exige muitas terras para levar adiante sua ambiciosa agenda de crescimento industrial e infraestrutural. Pelo menos 11 milhões de hectares são necessários para os projetos de desenvolvimento dos próximos 15 anos. Mas grande parte da população indiana, sobretudo a marginalizada, depende de terra para subsistência. Mais de 200 milhões de pessoas dependem das florestas, ao passo que 118,9 milhões dependem de terras para cultivo.

Demandas cuja concorrência causa conflitos. Em muitos casos, as terras são adquiridas à força ou de forma fraudulenta pelo estado ou por interesses privados; dissidentes acabam sendo fichados por agências estatais sob falsas acusações; compensações são pagas apenas parcialmente; comunidades são deslocadas; casas são queimadas; e pessoas morrem. As disparidades sociais entre casta, classe e gênero também motivam estes conflitos. Calamidades relacionadas a mudanças climáticas acabam por tornar comunidades dependentes da terra ainda mais vulneráveis a deslocamentos. Tudo isso se reflete em verdadeiras batalhas por toda a Índia.

Quando começamos a escrever sobre problemas de desenvolvimento no país, nos deparamos com muitos destes conflitos. Porém, logo percebemos que não seria fácil emplacar estas histórias, que se passam em locais remotos, aos editores em Nova Déli. A mídia convencional não tratava da questão dos conflitos de terra, exceto em casos de violência e fatalidade ou que envolvessem a corte nacional. Reportagens esporádicas, escritas por poucos jornalistas, tinham pouco impacto. A voz dos afetados seguia sem ser ouvida. Suas preocupações, deixadas de lado.

O motivo, acreditávamos, era o fato de que jornalistas e editores viam os conflitos como incidentes isolados. Sabíamos que estas brigas por terra continham algumas das mais relevantes histórias sobre a economia indiana. Mas como poderíamos convencer editores e leitores de sua importância? Pensamos que se os jornalistas pudessem aumentar a escala de sua cobertura de conflitos individuais de forma a examinar tendências mais amplas, estas reportagens não só teriam um alcance maior como também poderiam revelar a intensidade destes embates e seu impacto nas pessoas, na economia e no meio ambiente. O maior desafio para a realização disso era a falta de um banco de dados que pudesse ser explorado pelos

jornalistas para detecção de tendências emergentes ligadas a tipos de conflito específicos, envolvendo estradas, municípios, mineração ou zonas de proteção de vida selvagem. Não havia banco de dados com informações sobre conflitos em andamento na Índia. O jeito foi criar um.

Em novembro de 2016, criamos o Land Conflict Watch, projeto de jornalismo de dados baseado em pesquisa, que visa mapear e documentar as disputas territoriais no país. Desenvolvemos uma metodologia de documentação junto a acadêmicos do setor de governança territorial. Reunimos uma rede de pesquisadores e jornalistas, espalhados pelo país, para que registrassem os conflitos em suas regiões seguindo esta metodologia.

Para fins deste projeto, definimos como conflito de terra qualquer situação que envolva demandas ou alegações discordantes a respeito do uso ou propriedade de terras, em que a comunidade seja uma das partes contestadoras. Conflitos em progresso onde tais demandas ou alegações foram registradas em formato escrito ou audiovisual em qualquer lugar, do nível local ao nacional, são incluídos. Estes registros podem ser notícias, resoluções de assembleias destes vilarejos, registros de consultas públicas para projetos de desenvolvimento, reclamações enviadas pelo povo a autoridades do governo, registros policias ou documentos de tribunais. Situações como disputas de posse entre entes privados ou entre ente privado e governo são excluídas, a não ser quando envolvem diretamente públicos mais amplos.

Pesquisadores e jornalistas monitoram a cobertura local e nacional a respeito de suas regiões e interagem com ativistas, organizações comunitárias e advogados para encontrarem estes conflitos. Eles, então, coletam e verificam as informações a partir de documentos do governo disponibilizados ao público em geral, estudos independentes e ao conversar com as partes afetadas. Dados como localização do conflito, motivações, número de pessoas afetadas, área afetada, tipo de propriedade — privada, comum ou floresta —. nomes das agências governamentais e corporativas envolvidas e um resumo da situação, tudo isso é documentado.

Estes dados são inseridos pelos pesquisadores em um software de relatório e revisão integrado ao site do Land Conflict Watch. Revisores dedicados examinam e verificam os dados. O software possibilita o livre fluxo de trabalho entre pesquisadores e revisores, antes da publicação destes dados. O painel, dentro do portal, apresenta não só uma imagem macro dos conflitos atuais a nível nacional, mas também é possível “dar um zoom” e explorar detalhes de cada um destes conflitos, acompanhados de documentos de apoio aos dados, a nível micro. Há, ainda, um mapa interativo que fornece a localização aproximada da disputa em questão.

No momento, são cerca de 35 colaboradores, entre pesquisadores e jornalistas. Até setembro de 2018, mais de 640 casos foram documentados no projeto. Estas disputas por terra afetam algo próximo de 7,3 milhões de pessoas ao longo de 2,4 milhões de hectares de terra. Investimentos estimados em 186 bilhões de dólares estão ligados a projetos e esquemas afetados por estes conflitos.

Cada um destes é delineado no portal e nas redes sociais do projeto, de forma a chamar a atenção de pesquisadores e jornalistas do país. Nossa equipe, então, trabalha junto com estes jornalistas na criação de matérias investigativas, profundas, cobrindo a intersecção entre disputas por terra, política, economia, classe, gênero e meio ambiente, com base nestes dados. Também colaboramos com a mídia nacional e internacional para a publicação deste material. Muitos destes artigos foram republicados por outros veículos da grande mídia. Além disso, fizemos treinamentos com jornalistas a respeito do uso de banco de dados para pesquisa e escalonamento destes artigos em torno da governança territorial.

O projeto Land Conflict Watch é contínuo. Além do desenvolvimento de narrativas, trabalhamos com acadêmicos, pesquisadores e estudantes para fomento do debate público. Os dados do projeto já foram citados por *think tanks* voltados à discussão de políticas em seus relatórios. Especialistas em governança territorial já escreveram editoriais em jornais nacionais usando estes dados. Com frequência, recebemos solicitações de estudantes de universidades indianas e estrangeiras, que querem usar nossos dados em suas pesquisas. Organizações sem fins lucrativos usam estes dados de conflitos territoriais, bem como documentos e relatos de caso, no fortalecimento de sua campanha na luta por direito às terras em comunidades afetadas pelas disputas. Estas narrativas informam o público e ajudam a dar forma ao debate em torno do direito à terra e de questões relacionadas de governança na Índia.

Kumar Sambhav Shrivastava é um jornalista indiano cujo trabalho se dá na intersecção entre políticas públicas, negócios e justiça social. Ankur Paliwal é um jornalista escrevendo na intersecção entre ciência e condição humana.

Práticas alternativas de dados na China

Yolanda Jinxin Ma

Há alguns anos, fiz uma espécie de introdução ao jornalismo de dados da China durante o Google News Summit, evento organizado pelo Google News Lab. Era um belo dia de inverno no coração do Vale do Silício, o público uma centena de profissionais veteranos de comunicação, a maioria destes vindo de países ocidentais. Comecei pedindo primeiro para que levantassem as mãos aqueles que não acreditavam haver bons dados na China e, então, aqueles que não acreditavam haver jornalismo de verdade no país. Bastante gente levantou as mãos nos dois casos, o que foi acompanhado por algumas risadas.

São dois comentários, quem sabe até mesmo vieses, com os quais me deparo quando participo ou falo em conferências internacionais de jornalismo. Com base em minhas observações nos últimos seis anos, ao invés de não haver dados, há enormes quantidades destes sendo gerados e acumulados diariamente na China, e sua qualidade vem aumentando. E, muito pelo contrário, existem muitos jornalistas criando artigos impressionantes todos os dias, por mais que no final das contas nem tudo seja publicado.

Geração de dados com base em problemas

Antes mesmo do termo "jornalismo de dados" ser introduzido na China, já existia material jornalístico baseado em dados. Por mais que, atualmente, usemos o termo "histórias baseadas em dados" no país, houve uma época em que a ótica era invertida: testemunhávamos acontecimentos ou problemas em específico que moviam a geração de dados, em vez de termos dados movendo estas histórias. São sempre questões que ressoam com o cidadão comum, como a poluição do ar.

Desde 2010, o Ministério do Meio Ambiente divulga um índice de poluição do ar em tempo real, mas com a ausência de um dado importante.[28] A informação sobre o volume de PM2.5 — poluentes que medem menos que 2,5 micrômetros de diâmetro, causadores de danos irreparáveis ao corpo — não era divulgada.

Considerando a seriedade da poluição atmosférica e a falta de informações oficiais sobre o volume de PM2.5, uma campanha nacional teve início em novembro de 2011, chamada "Eu testo o ar pela Pátria Mãe", sugerindo que todo cidadão contribuísse com o

[28] http://www.chinanews.com/life/2010/11-26/2682313.shtml.

monitoramento da qualidade do ar, publicando as informações em redes sociais.[29] A campanha foi iniciada por uma organização ambiental sem fins lucrativos, cujo equipamento havia sido financiado coletivamente pelos cidadãos. Além disso, a organização treinava voluntários interessados. O movimento ganhou força após alguns influenciadores digitais se juntarem, incluindo Pan Shiyi, executivo conhecido no país com mais de sete milhões de seguidores na Sina Weibo, uma das redes sociais mais usadas na China.[30]

Após dois anos de campanha, iniciada em janeiro de 2012, os dados a respeito dos volumes de PM2.5 finalmente passaram a ser incluídos nas informações fornecidas pelo governo. Era um bom começo, mas havia outros desafios. Imediatamente observaram-se discrepâncias nos dados divulgados pela Embaixada dos EUA na China, o que gerou dúvidas sobre a precisão e a transparência dos dados.[31]

Em termos de funcionalidade, nada era amigável ao jornalista também. As informações eram fornecidas em atualizações hora a hora, cobrindo mais de 100 cidades, mas só podiam ser visualizadas na página, sem a possibilidade de baixar o conjunto de dados em qualquer outro formato. Por mais que os dados houvessem sido centralizados, não há uma série histórica acessível ao público. Ou seja: sem um código capaz de coletar os dados hora a hora e salvá-los em um arquivo local, é impossível analisar tendências ao longo do tempo ou comparar informações entre cidades.

A história não para por aí. Seguimos gerando dados com base em problemas como estes. Quando os dados não estão bem estruturados ou em formato amigável ao usuário, quando jornalistas se veem às voltas com limitações técnicas, a sociedade civil ou os aficionados por tecnologia podem dar suporte.

Um exemplo de 2011: o site PM5.in, que coleta dados de poluição atmosférica e divulga em um formato limpo. De acordo com o próprio site, mais de 1 bilhão de consultas foram feitas desde o início de sua operação.[32] Outro exemplo é o da Qing-Yue, organização não governamental que coleta e limpa dados ambientais de sites do governo em todos os níveis, divulgando para o público em formatos amigáveis ao usuário. No final das contas, os dados processados são amplamente utilizados não só por equipes especializadas em grandes

---

29 http://www.bjep.org.cn/pages/Index/40-1699?rid=2782.

30 https://blogs.wsj.com/chinarealtime/2011/11/08/internet-puts-pressure-on-beijing-to-improve-air-pollution-monitoring/.

31 https://blogs.wsj.com/chinarealtime/2012/01/23/comparing-pollution-data-beijing-vs-u-s-embassy-on-pm2-5/.

32 http://pm25.in/about.

veículos de mídia, mas também por agências do próprio governo, para melhor desenvolvimento de políticas.

A geração de dados e uma conscientização cada vez maior sobre certos temas caminham de mãos dadas. Em 2015, um documentário abordando a questão da poluição atmosférica tomou o país de assalto. O filme, feito com recursos próprios e intitulado *Under the Dome* (*Sob o Domo*, em tradução livre) expôs a crise ambiental causada pela poluição em todo o país, traçando um caminho até as raízes da questão e os diversos envolvidos.[33] A obra foi comparada a *Uma Verdade Inconveniente*, de Al Gore, por seu estilo e impacto. No desenrolar de sua narrativa, eram apresentados muitos dados científicos, gráficos demonstrando análises e explicando tendências ao longo dos anos, bem como visualizações de redes sociais que mostravam corrupção dentro dos setores de energia e meio ambiente. Logo que foi lançado na rede, o filme viralizou e alcançou 200 milhões de visualizações em três dias, antes de sua censura e remoção, dentro de uma semana. Mesmo assim, o documentário chamou atenção do público e fomentou debate nacional sobre o tema, incluindo a acessibilidade e a qualidade de dados da poluição atmosférica, conscientizando a liderança do país quanto à significância do tema.

Duas semanas após o lançamento do documentário, em coletiva no Congresso Nacional do Povo, ao tratar de uma questão sobre poluição atmosférica fazendo referência ao filme, o premiê Li Keqiang admitiu falha do governo no atendimento a demandas públicas para redução da poluição e reconheceu alguns dos problemas levantados pelo documentário, incluindo fraca fiscalização de restrições à poluição, enfatizando ainda que o governo imporia sanções maiores para reduzir a fumaça tóxica.[34] No final de agosto de 2015, foi lançada a nova Lei de Prevenção e Controle da Poluição do Ar, implementada em janeiro de 2016.[35]

A poluição atmosférica é só um exemplo de que mesmo quando a disponibilidade ou acessibilidade a dados é desafiadora, a preocupação pública com estas questões pode levar a contribuições de cidadãos à geração de dados, bem como a mudanças nas atitudes do governo e na disponibilidade de dados do setor público. Em ecossistemas mais bem estabelecidos, estes dados são disponibilizados com maior agilidade e facilidade de uso, simplificando o trabalho do jornalista: contar histórias com base em dados. Na China, este processo pode ser menos linear, e a dinâmica entre cidadão, governo, sociedade civil e mídia é muito mais interativa. Os dados, ao invés de servirem apenas como o pontapé inicial de uma narrativa,

---

[33] https://www.youtube.com/watch?v=T6X2uwlQGQM.

[34] https://www.nytimes.com/2015/03/16/world/asia/chinese-premier-li-keqiang-vows-tougher-regulation-on-air-pollution.html.

[35] https://www.chinadialogue.net/article/show/single/en/8512-How-China-s-new-air-law-aims-to-curb-pollution.

também podem vir à tona em um estágio mais avançado, possibilitando novas relações entre jornalistas e público.

Evoluindo a cultura de dados

O ambiente de dados na China vem mudando bastante na última década, em parte por conta da dinâmica descrita acima e em parte por conta de outros fatores, como o movimento global pela livre circulação de dados, empresas de internet com crescimento acelerado, índices de penetração mobile surpreendentemente altos etc. A cultura de dados evolui junto a estas tendências.

A legislação implementada pelo governo serve de espinha dorsal política para a disponibilidade de dados. Para a surpresa de muitos, a China tem, sim, leis de liberdade da informação. A regulamentação do conselho de estado sobre a divulgação de informações governamentais foi adotada em 2007 e implementada em 1º de maio de 2008, garantindo a divulgação de dados e firmando compromisso com a transparência governamental. Com base na regulamentação, todas as agências do governo (de todos os níveis) criaram páginas na internet para divulgação de informações, incluindo dados. Contudo, por mais que a nova regulamentação desse aos jornalistas o direito de solicitar certos dados ou informações das autoridades, nos primeiros três anos de aplicação da lei não se soube de nenhum pedido de informação por parte de nenhum veículo ou jornalista, de acordo com estudo de 2011 publicado pelo grupo de mídia chinês *Caixin*.[36] O estudo revelou que, em 2010, o *Southern Weekly*, um dos maiores jornais do país, fez uma solicitação-teste a 29 agências ambientais pedindo a divulgação de certas informações, com uma taxa de resposta de 44%. Cabe notar que dentro destas empresas de comunicação geralmente não há um sistema de apoio, como um setor legal, por exemplo, que ajudaria os jornalistas a darem prosseguimento às suas demandas. Um jornalista, por conta própria, chegou a processar o governo pela não divulgação de informações e acabou perdendo seu emprego. Jornalistas chineses se veem as voltas com riscos e dificuldades muito maiores que seus colegas ocidentais quando se valem de ferramentas legais.

Na esteira do movimento pela livre circulação de dados e do interesse crescente em torno de big data, a China reagiu. Em 2012, Xangai e Pequim lançaram seus portais de dados, com centenas de conjuntos de dados cada, cobrindo temas como uso de terras, transporte, educação, monitoramento de poluição etc. Nos anos seguintes, dezenas de portais foram criados, não só nas cidades maiores, mas também em distritos menores e províncias menos desenvolvidas. O desenvolvimento foi bagunçado, sem modelos padronizados ou estrutura de divulgação de dados a nível local, o que no final das contas não facilitava tanto a coleta de

[36] http://finance.ifeng.com/leadership/gdsp/20110901/4512444.shtml.

dados por parte do usuário. Em 2015, o Conselho de Estado divulgou o Plano de Ação de Desenvolvimento em Big Data, onde a livre circulação de dados era oficialmente reconhecida como um dos dez principais projetos nacionais, com uma linha do tempo definida para a abertura dos dados governamentais, incluindo uma data exata para a apresentação.[37] Porém, os dados oficiais nem sempre são o pontapé inicial para os jornalistas, e nem sempre estas informações estão alinhadas com os interesses públicos.

Por outro lado, o setor privado, em especial gigantes da tecnologia como Alibaba ou Tencent, acumulou enormes quantidades de dados ao longo dos anos. De acordo com dados oficiais recentes, o número de consumidores ativos do Alibaba chegou a 601 milhões em 30 de setembro de 2018.[38] Os dados de e-commerce de uma base de usuários tão forte, equivalente a toda a população do sudeste asiático, pode revelar muitas tendências comerciais, mudanças demográficas, direcionamento de migrações urbanas, mudanças de hábitos de consumo etc. Há, ainda, verticais em que dados mais específicos estão disponíveis, caso de sites como Dianping, o equivalente chinês do Yelp. Mesmo com algumas preocupações em relação à privacidade e à segurança, se utilizadas adequadamente, estas plataformas são fontes ricas de dados para jornalistas.

Um exemplo extraordinário do uso de big data é a atuação do Rising Lab, uma equipe do Shanghai Media Group, especializada em narrativas sobre vida urbana baseadas em dados.[39] A equipe foi criada em resposta a uma urbanização emergente: existem mais de 600 cidades na China atualmente — em 1978 eram apenas 193 — e 56% da população mora nestas áreas urbanas, de acordo com relatório do governo publicado em 2016.[40] Junto a essa rápida urbanização, cresce também o uso de internet e dispositivos móveis, além de mudanças de estilo de vida, como a rápida adoção de modelos de economia compartilhada. Todas tendências com grande impacto na agregação de dados.

Por meio de parcerias e do suporte técnico de empresas de tecnologia, o Rising Lab vem coletando dados de aplicativos e sites frequentemente usados pelos cidadãos, informações como preços de propriedades, número de cafeterias e bares, número de espaços de coworking ou facilidade do transporte público, dentre outras que refletem diversos aspectos da vida urbana. Aliando isso à sua metodologia própria, a equipe criou uma série de rankings da cidade abordando diferentes aspectos, como atratividade comercial, nível de

---

[37] http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

[38] https://www.businesswire.com/news/home/20181102005230/en/Alibaba-Group-Announces-September-Quarter-2018-Results.

[39] https://www.yicaiglobal.com/category/The%20Rising%20Lab.

[40] http://english.gov.cn/news/video/2016/04/20/content_281475331447793.htm.

inovação, diversidade de vida etc. Os rankings e histórias são atualizados ano a ano com base em novos dados, mas sempre seguindo a mesma metodologia, para fins de consistência. O conceito e as histórias publicadas tiveram recepção positiva e começaram até mesmo a influenciar políticas de planejamento urbano e tomada de decisão de empresas, de acordo com Shen Congle, diretora do Rising Lab, cujo sucesso ilustra bem as novas dinâmicas entre provedores de dados, jornalistas e cidadãos que vêm emergindo.

Este mesmo sucesso mostra como assuntos mais leves podem se tornar um celeiro fértil para o jornalismo de dados, junto a outras questões mais urgentes, como crises ambientais, corrupção, injustiça judicial, saúde pública e lavagem de dinheiro. Também explora novos modelos de negócios em potenciais, bem como a forma pela qual produtos baseados em dados podem agregar valor a governos e empresas.

O consumo de notícias por parte de leitores impactou o desenvolvimento do jornalismo de dados, um mais visual e o outro mais móvel. Desde 2011, infográficos se popularizaram graças aos esforços de alguns grandes portais de notícias na construção de vértices dedicados a infográficos, em grande parte baseados em dados. Em 2014, a história da derrocada do ex-chefe de segurança Zhou Yongkang, um dos nove mais antigos políticos chineses, foi a notícia mais importante do ano. Junto do texto, a *Caixin* criou uma visualização interativa para ilustrar a complexa rede em torno de Zhou, incluindo 37 pessoas e 105 empresas ou projetos ligados a ele, e a relação entre todas essas entidades, tudo com base no artigo investigativo de 60.000 palavras escrito por seus repórteres. A parte interativa do projeto recebeu quatro milhões de visitas em uma semana, e mais 20 milhões de visualizações nas redes sociais, de acordo com a *Caixin*.[41] Com a atenção obtida, o projeto apresentou ao público novas formas de narrativa baseadas em dados, gerando um interesse que não existia antes.

Quase que simultaneamente, a indústria midiática começava a abraçar a era dos dispositivos móveis. Como no caso do Rising Lab, mais e mais histórias criadas com dados, bem como qualquer conteúdo online na China, vêm sendo disseminadas através de dispositivos móveis, em sua maioria. De acordo com o Centro de Informações de Rede de Internet da China, mais de 95% dos usuários de internet do país usaram um dispositivo móvel para acessar a rede em 2016.[42] WeChat, o popular serviço chinês de troca de mensagens e rede social, chegou a um bilhão de usuários em março de 2018.[43] A dominância das plataformas móveis implica que narrativas baseadas em dados na China não só são feitas

---

41 http://vislab.caixin.com/?p=50.

42 https://www.emarketer.com/Article/More-than-95-of-Internet-Users-China-Use-Mobile-Devices-Go-Online/1015155.

43 https://technode.com/2018/03/05/wechat-1-billion-users/.

pensando primeiramente nestes dispositivos, como, em muitos casos, acabam sendo exclusivas do meio. Tamanha demanda levou ao desenvolvimento de histórias interativas enxutas, simples, e amigáveis à plataforma móvel.

Em suma, a cultura de dados chinesa vem evoluindo, motivada por diversos fatores, de movimentos globais à legislação, demanda pública a pedidos da mídia, novas gerações de provedores de dados a novas gerações de consumidores de notícias. As relações interdependentes entre os atuantes no setor criaram dinâmicas bastante complexas, em que restrições e oportunidades coexistem. Na China, o jornalismo de dados floresceu e forjou seu próprio caminho.

Dicas práticas

Esta seção é voltada para aqueles que desejam trabalhar em material relacionado à China e não sabem por onde começar. Não será fácil. De cara, temos a barreira do idioma, já que a maioria das fontes estão disponíveis somente em chinês. A partir daí, temos os outros problemas, comuns a qualquer lugar: precisão dos dados, completude dos dados, inconsistências, e por aí vai. Vamos supor que você tenha todas as habilidades necessárias para detectar estes problemas e resolvê-los.

Primeiro, quem são os grandes nomes? Muitos grandes veículos de mídia criaram suas próprias equipes de dados, logo é uma boa ideia acompanhar o que esse pessoal anda fazendo e pedir dicas aos seus repórteres. Deixo aqui uma lista com alguns nomes que você deveria conhecer:

- *Caixin Media*, Data Visualisation Lab[44]
- *The Paper*, Beautiful Data Channel[45]
- *Shanghai Media Group*, The Rising Lab[46]
- *DT Finance*[47]

Segundo, onde encontrar os dados? Fazer uma lista abrangente demandaria um manual à parte, então deixo aqui algumas sugestões iniciais:

---

[44] http://vislab.caixin.com/.

[45] https://www.thepaper.cn/list_25635.

[46] https://www.cbnweek.com/topics/10.

[47] https://www.dtcj.com/.

1. Comece visitando sites governamentais, incluindo ministérios e agências locais. É preciso saber qual ou quais departamentos possuem os dados que você busca. Recomendo que verifique ministérios específicos (por exemplo, o Ministério de Proteção Ambiental) e, também, o site dedicado a dados de nível local, caso exista.

2. Você encontrará dados inesperados, a não ser que espere se deparar, por exemplo, com milhões de decisões judiciais completas disponibilizadas pelo governo chinês após 2014. Documentos legais são relativamente transparentes nos EUA, não na China. Mas a Suprema Corte do Povo criou um banco de dados chamado Julgamentos da China Online, divulgando estes processos.

3. Assim que encontrar dados potencialmente úteis na rede, certifique-se de baixar uma cópia.

4. Por vezes, os dados buscados não estão disponíveis online. Isso ainda é bem comum. Eles podem surgem na forma de um relatório anual do governo que foi publicado e pode ser solicitado pela internet, e, em outras ocasiões, existem somente nos arquivos físicos, em papel, de certos escritórios. Certas agências, por exemplo, têm os registros de empresas privadas, mas nem tudo está na rede.

5. Se os dados não são divulgados pelo governo, vale a pena checar se não há conteúdo gerado pelo usuário por aí. Dados sobre saúde pública são bastante limitados, mas existem sites dedicados ao registro de hospitais ou centros para idosos, dentre outros. Se você conseguir coletar e limpar os dados, terá informação valiosa que possibilitará uma boa visão geral do assunto.

6. Use bancos de dados em Hong Kong, oficiais, como o Registro de Empresas de Hong Kong, ou independentes, como o Webb-site Reports. Com a aproximação da China Continental e de Hong Kong tanto política quanto financeiramente, há mais informação sendo disponibilizada graças ao ambiente mais transparente em Hong Kong e à maior fiscalização, o que pode ajudar a determinar o caminho percorrido pelo dinheiro, por exemplo.

7. Há dados sobre a China que não estão, necessariamente, no país. Existem organizações internacionais ou instituições acadêmicas que têm dados riquíssimos relacionados à China. O jornal *The Paper*, por exemplo, usou dados da NASA e de Harvard em um de seus artigos mais recentes.

Por fim, ao passo que alguns dos desafios e a experiência como um todos sejam únicas à China, estes têm o potencial de ensinar algumas lições úteis para outros países em

que os arranjos sociais, culturais e políticos possuem forma diferente, mas restrições semelhantes.

Yolanda Jinxin Ma é uma entusiasta (aposentada) do jornalismo de dados que apresentou a prática à China e atualmente explora a intersecção entre jornalismo de dados, revolução tecnológica e investimentos de impacto.

Remontagem de dados públicos em Cuba: como jornalistas, pesquisadores e estudantes colaboram quando as informações são inexistentes, desatualizadas ou escassas

Saimi Reyes Carmona, Yudivián Almeida e Ernesto Guerra

A equipe do *Postdata.club* é pequena. No começo, éramos quatro jornalistas e um especialista em matemática/ciência da computação, e havíamos decidido, juntos, nos aventurarmos na prática do jornalismo de dados em Cuba. Também queríamos investigar os problemas relacionados a isso. No país, até então, não existia veículo jornalístico dedicado explicitamente ao jornalismo de dados. Fomos os primeiros.

No momento, somos dois jornalistas e um cientista de dados trabalhando em cima do *Postdata.club*, no nosso tempo livre. Em nossos empregos nós não lidamos com jornalismo de dados, já que Saimi Reyes é editora de um site voltado à cultura, Yudivián Almeida é professor da Escola de Matemática e Ciência da Computação da Universidade de Havana e Ernesto Guerra é jornalista de uma revista voltada à tecnologia e ciência popular. Nosso objetivo é ser mais do que uma organização de mídia, sendo também um espaço experimental em que possamos explorar e aprender mais sobre o país em que vivemos com e através dos dados.

Buscamos usar dados de acesso livre e públicos, no desejo de compartilhar ambos: nossa pesquisa e como a fazemos. Por isso, começamos a usar o GitHub, plataforma que abriga o *Postdata.club*. Com base nos requisitos exigidos pelas histórias que queremos contar, decidimos a extensão dos textos e recursos usados, sejam eles gráficos, imagens, vídeos ou áudios. Focamos no jornalismo de impacto social, em formatos longos e curtos. Nosso interesse reside em todos os assuntos que podemos abordar com dados, mas, acima de tudo, temas que estejam relacionados a Cuba ou seu povo.

Conduzimos as investigações de duas maneiras, a depender dos dados. Por vezes, temos acesso a dados públicos ou de livre acesso. Nestes casos, analisamos as informações para determinar se há uma história a ser contada ali. Por vezes, temos nossas próprias perguntas e vamos direto aos dados para encontrar as respostas e contarmos uma história. Em outras situações, exploramos os dados e encontramos elementos que acreditamos serem interessantes ou surgem questões cujas respostas podem ser relevantes e podem ser respondidas por aquela fonte de dados ou outra fonte.

Se o que recolhemos destas bases de dados parece interessante, a complementamos com outras fontes, como entrevistas, e comparamos com outras informações. A partir disso, pensamos em como se desenrolará a narrativa em torno da pesquisa, através de um ou mais textos sobre o tema, acompanhado por visualizações cujo objetivo é apresentar percepções derivadas dos dados.

Há, ainda, situações que ocorrem com alguma frequência, em que precisamos criar bancos de dados nós mesmos, com base em informações públicas, mas sem a estrutura apropriada. Usamos estas informações como fundamento de nossa análise e questionamento. Por exemplo, para discutirmos as eleições cubanas, tivemos que criar bancos de dados a partir de diferentes fontes de informação. Para tanto, começamos com os dados publicados no site do Parlamento cubano. Porém, estas informações estavam incompletas, e, assim, nossos bancos de dados foram preenchidos com o que encontramos na imprensa e em sites ligados ao Partido Comunista de Cuba. Depois, para tratar da questão do recém-designado Conselho de Ministros, precisamos criar outro banco de dados. Neste caso, as informações fornecidas pela Assembleia Nacional não estavam completas e usamos também material da imprensa, como da *Gazeta Oficial* e de sites informativos, para termos uma visão mais completa de tudo. Em ambas as circunstâncias, criamos bancos de dados em formato JSON que foram, então, processados e utilizados na maioria dos artigos que escrevemos sobre as eleições e os poderes executivo e legislativo em Cuba.

Na maior parte do tempo, compartilhamos estes bancos de dados em nosso site, acompanhados de uma explicação sobre nossos métodos. Dito isso, nosso trabalho na ilha às vezes sofre com a falta de dados que deveriam ser públicos e acessíveis. Muitas das informações que usamos é fornecida por entidades governamentais, mas, em nosso país, várias destas instituições não têm representação adequada na internet ou não divulgam tudo que deveriam. Em alguns casos, fomos bater na porta destas instituições para solicitar acesso a algum dado ou informação em específico, um processo muitas vezes trabalhoso, mas importante.

Para nós, um dos maiores problemas na obtenção de dados em Cuba é sua desatualização. Quando finalmente temos acesso à informação que buscamos, muitas vezes ela está incompleta ou muito desatualizada. Ou seja, os dados podem estar disponíveis para download e consulta em um site, mas a última atualização é de cinco anos atrás. Em alguns casos, nós mesmos completamos as informações ao consultar outros sites confiáveis. Em outros, nos resta buscar documentos impressos, imagens ou indivíduos que possam nos ajudar a trabalhar com informações mais recentes. Tudo isso afeta nosso método de trabalho, baseado em cada investigação e nos dados disponíveis. Estas são as particularidades do nosso meio e o ponto do qual partimos para oferecer aos nossos leitores jornalismo de qualidade

com impacto social. Caso a informação compartilhada por nós seja útil para ao menos uma pessoa, sentimos que valeu todo o esforço.

Além de manter o endereço *Postdata.club* no ar, onde postamos todos os artigos e histórias resultantes de nossa pesquisa, também queremos estender esta forma de fazer jornalismo de dados para outros espaços. Para tanto, desde 2017, damos um curso de Jornalismo de Dados para estudantes de jornalismo da Escola de Comunicação da Universidade de Havana. O tema mal foi ensinado em nosso país, o que exige aprendizagem e preparo contínuos — observando, também, o retorno de estudantes e colegas.

Através destas trocas com estes futuros jornalistas e profissionais da comunicação, aprendemos diversas formas de trabalhar e, surpreendentemente, descobrimos novas formas de acessar informação. Uma das coisas que fazemos nestas aulas é envolver os estudantes na construção de um banco de dados. Não havia uma única fonte em Cuba onde poderíamos obter o nome das pessoas que receberam premiações nacionais, com base em sua obra em diferentes áreas e atividades. Junto a todos os estudantes e professores, coletamos informações e estruturamos um banco de dados de mais de 27 prêmios, desde que começaram a ser distribuídos até então. Esta informação nos permitiu revelar que havia uma lacuna na distribuição destes prêmios em questões de gênero. Mulheres eram premiadas em apenas 25% das ocasiões. Com esta descoberta, pudemos escrever juntos uma história que encorajava a reflexão sobre questões de gênero em relação ao reconhecimento nacional de diferentes tipos de trabalho.

Em 2017, também passamos por outra experiência reveladora que nos ajudou a entender que, em muitos casos, não devemos nos contentar com bancos de dados já disponibilizados e não devemos supor tanto sobre o que é possível ou não. Como parte do trabalho final do curso, pedimos aos estudantes que formassem pequenas equipes na realização de sua tarefa. Estas equipes eram compostas por um dos quatro integrantes do *Postdata.club*, dois estudantes de jornalismo e um estudante de ciências da computação, que participavam do curso de forma a obtermos uma dinâmica interdisciplinar. Uma das equipes propôs abordar as novas iniciativas de trabalho autônomo em Cuba. Aqui, estas pessoas eram conhecidas como “cuentapropistas”. O que há poucos anos era prática extremamente limitada, caminha a passos largos para a aceitação como nova forma de trabalho na sociedade.

Queríamos investigar o fenômeno do trabalho autônomo em Cuba. Por mais que o tema fosse abordado com frequência, não havia quase nada quanto às especificidades deste tipo de trabalho por província, o número de licenças concedidos por área de atividade, ou tendências ao longo do tempo. Junto dos estudantes, discutimos quais questões seriam abordadas e chegamos à conclusão de que faltavam fontes com dados utilizáveis. Em locais

onde estes dados deveriam ser divulgados publicamente, não havia rastro algum. Nem mesmo informação na imprensa nacional que contivesse uma quantia significativa de dados. Com a exceção de algumas entrevistas e números isolados, nada era amplamente divulgado.

Pensamos, então, que estes seriam dados de difícil obtenção. De qualquer forma, estudantes de jornalismo que faziam parte de nosso curso contataram o Ministério do Trabalho e Segurança Social para requisitar informações sobre trabalhadores autônomos em Cuba. Lá, foram informados de que o banco de dados em questão poderia, sim, ser repassado e, em alguns poucos dias, os estudantes tinham as informações em mãos. De repente, tínhamos acesso a dados que interessavam a muitos cubanos, e podíamos compartilhá-los, afinal, aquilo deveria ser público. O ministério não tinha um portal atualizado e nós havíamos presumido, erroneamente, que os dados eram inacessíveis.

Juntos, os estudantes, o futuro cientista da computação e o jornalista do *Postdata.club* escreveram sua história sobre trabalho autônomo no país. A partir dos dados, descreveram de maneira detalhada a situação deste tipo de trabalho em Cuba. Coincidentemente, a informação chegou a nós em uma época de intensa atividade em torno do tema. Naqueles meses, o Ministério do Trabalho e Segurança Social decidiu limitar a concessão de licenças para 28 atividades entre as autorizadas para trabalho não estatal. Desta forma, pudemos usar os dados rapidamente de forma a analisar como estas novas medidas afetariam a economia do país e as vidas dos autônomos.

Grande parte de nossos leitores se surpreendeu com o fato de que havíamos obtido estes dados com relativa facilidade. No final das contas, o acesso a estas informações se deu porque nossos estudantes fizeram uma solicitação ao ministério e, até hoje, o *Postdata.club* foi o único lugar em que esta informação foi tornada pública, de forma que qualquer um pode consultá-la e analisá-la.

O jornalismo de dados em Cuba continua sendo um desafio. Dentre outras coisas, a dinâmica entre criar e acessar dados, bem como as culturas políticas e institucionais, diferem de outros países em que este tipo de informação é disponibilizado com maior prontidão. Logo, devemos sempre ser criativos na busca por novos métodos de acessar informação, e, a partir disso, contar histórias sobre questões relevantes. Isso só é possível se continuarmos tentando, e nós do *Postdata.club* seguiremos empreendendo esforços para sermos um exemplo de como o jornalismo de dados é possível mesmo em regiões em que o acesso à informação passa por maiores dificuldades.

Ernesto Guerra é jornalista do Postdata.club, o primeiro site de jornalismo de dados de Cuba. Saimi Reyes é editora e jornalista do Postdata.club e professora de jornalismo de

dados na Universidade de Havana. Yudivián Almeida é editor de dados do Postdata.club e professora da Escola de Matemática e Ciência da Computação da Universidade de Havana.

Geração de dados com os leitores do *La Nación*

Flor Coelho

No *La Nación*, trabalhamos na produção de grandes investigações baseadas em dados com a colaboração de nossos leitores. Este capítulo apresenta uma visão de "bastidores" sobre como organizamos a participação do público em torno de alguns destes projetos, incluindo o estabelecimento de objetivos, apoio a comunidades investigativas e estímulo a colaborações de longo prazo com nossos leitores, organizações externas e parceiros.

Em projetos como estes, muitas vezes nossa meta é abordar o "impossível" por meio da tecnologia como facilitadora de colaborações em larga escala, permitindo o engajamento de usuários com o jornalismo investigativo e o processo de tornar dados oficiais públicos.

Por exemplo, passamos cerca de cinco anos transcrevendo 10.000 PDFs de despesas do Senado, dois anos ouvindo 40.000 ligações telefônicas interceptadas e mais alguns meses digitalizando mais de 20.000 formulários eleitorais manuscritos.[48]

Para esse tipo de iniciativa colaborativa, usamos a plataforma Vozdata — inspirada no "MP's Expenses", do *The Guardian,* e no "Free the Files", do *Propublica* —, desenvolvida com apoio do Knight Mozilla Open News e da Civicus Alliance. O software por trás da Vozdata foi aberto ao público como Crowdata.[49]

Organizando participações

Para estes projetos, nossos colaboradores eram em grande parte estudantes de jornalismo, voluntários, ONGs voltadas à transparência e aposentados. Cada um tem suas próprias motivações, a depender do projeto, como contribuir com iniciativas de interesse público, trabalhar com nossa equipe de dados e conhecer gente nova em nossos encontros.

O Vozdata tem funcionalidades de equipe e rankings em tempo real. Viemos explorando como estas funções podem melhorar a participação por meio de "gamificação".

---

48 http://blogs.lanacion.com.ar/projects/data/argentina%C2%B4s-senate-expenses-2004-2013/, http://blogs.lanacion.com.ar/projects/data/prosecutor-nisman-phone-interceptions-mapped-in-playlists/, http://blogs.lanacion.com.ar/projects/data/vozdata-2015-unveiling-argentina%c2%b4s-elections-system-failures-with-impact/.

49 http://blogs.lanacion.com.ar/projects/data/vozdata-ii-civic-participation-for-investigative-reporting-and-educational-platform/, http://www.theguardian.com/news/datablog/2009/jun/18/mps-expenses-houseofcommons, https://www.propublica.org/series/free-the-files. https://github.com/crowdata/crowdata.

Obtivemos resultados excelentes quando fomentamos a participação da sociedade civil em torno de feriados nacionais da Argentina. O grosso da participação na construção de um banco de dados colaborativo se dá remotamente (online). Mas também encorajamos usuários a participarem em eventos offline no *La Nación* ou durante maratonas hacker em Datafests ou eventos como o Hacks/Hackers Media Party, em Buenos Aires. Em algumas ocasiões, divulgamos dados durante aulas de jornalismo em universidades parceiras.

Ao passo que estas maratonas hacker duram um ou dois dias, nossas maratonas online podem seguir por meses. Uma barra de progresso mostra quantos documentos foram completados e a porcentagem restante.

Definindo grandes objetivos

O papel principal dos colaboradores nos projetos "Despesas do Senado" e "Telegramas Eleitorais" era coletar dados estruturados específicos dos documentos fornecidos. Essa ação envolveu mais de 1.000 usuários. Além da extração de informações, a estes leitores foi dada a oportunidade de marcar dados como suspeitos ou inaceitáveis e deixar um comentário com informações adicionais (funcionalidade raramente usada). Quando se tem um prazo para finalizar um projeto colaborativo, talvez não seja possível bater a meta. Isso aconteceu conosco no projeto dos telegramas. As eleições estavam se aproximando e precisávamos publicar algumas conclusões. Por mais que alguns locais tivessem chegado a 100%, muitos ficaram entre 10% e 15% dos arquivos, fato que reconhecemos na publicação.

Apoio a comunidades investigativas

No caso do procurador Nisman, cuja investigação envolveu 40.000 arquivos, trabalhamos com uma rede de confiança composta por 100 colaboradores. Muitos arquivos de áudio estavam relacionados a conversas de teor privado (diálogos familiares, por exemplo) que o agente iraniano teve enquanto seu telefone estava grampeado por conta de ordem judicial. Um grupo de seis voluntários mergulhou fundo nesta investigação. Criamos um grupo no WhatsApp onde todos poderiam sugerir pistas e curiosidades.

Um de nossos voluntários resolveu um mistério que havia tirado nosso sono por alguns meses. Havíamos destacado várias ligações em que duas pessoas conversavam utilizando apelidos e números (como "Sr. Dragão, 2000"). Muitos de nossos voluntários haviam ouvido e transcrito tais gravações. Até mesmo pensamos em criar um banco de dados separado para analisar o código usado ali. Até que, um dia, um voluntário descobriu que se tratavam de apostas em corridas de cavalo! Uma rápida pesquisa no Google confirmou, muitos dos nomes utilizados eram de cavalos de corrida.

Sempre temos usuários que consideramos como avançados. Mas, a depender da escala do projeto, muitos voluntários trabalhando em cima de alguns documentos cada geralmente superam estes outros super usuários.

Estímulo a colaborações

Nossa dica para jornalistas e organizações que querem envolver seus leitores em investigações de dados é ter um gestor de comunidades dedicado que organiza e envia comunicados por meio de planilhas colaborativas (como o Google Sheets), listas de email e redes sociais.

Grandes coleções de documentos podem ser um bom lugar para se começar: a curva de aprendizado é rápida e os envolvidos sentem-se como parte de algo maior. Também é bom dar apoio a colaboradores por meio de tutoriais em vídeo ou introduções contextuais dentro da sua organização ou em eventos específicos.

Quando ganhamos um prêmio ligado a estes projetos colaborativos, organizamos um café da manhã para dividir a premiação com os voluntários. Estas relações com os leitores são de longo prazo, então certificamo-nos de dedicar o tempo e a energia necessários para nos encontrarmos em eventos, em visitas a universidades, ao conceder entrevistas para estudantes, e por aí vai.

Tratando-se de parcerias com universidades, professores geralmente servem como pontos de contato. A cada ano eles têm novas turmas ansiosas por colaborarem conosco nestes projetos (planeje com antecedência!).

ONGs ligadas à causa da transparência também podem demonstrar os benefícios destes projetos. Em nossa plataforma, cada dado pode ser registrado, o que significa que projetos e reconhecimentos na imprensa podem ser facilmente mostrados aos seus apoiadores.

Ao publicar resultados e artigos, recomendamos reconhecer o processo colaborativo e organizações envolvidas em cada plataforma (impressa, online e redes sociais) e mala direta. Enfatizar o caráter coletivo de tais projetos pode passar uma mensagem mais forte àqueles que queremos que sejam responsabilizados.

Conclusão

Para criar dados com leitores é essencial alocar tempo e recursos para engajar sua comunidade, lidar com solicitações, analisar produções, aproveitar interações e participar de eventos.

Voluntários se envolvem na classificação destes documentos porque acreditam que estes projetos têm importância. Já para os governos e aqueles que são alvo de cobertura, estes projetos deixam claro que não é apenas uma preocupação da imprensa, mas da sociedade também. Através destas iniciativas, os leitores podem se tornar defensores apaixonados e distribuidores online do conteúdo.

Flor Coelho é gerente de pesquisa e treinamento em novas mídias do La Nación, de Buenos Aires, na Argentina.

Soberania de dados indígenas: implicações para o jornalismo de dados

Tahu Kukutai and Maggie Walter

Tecnologias digitais, dentre elas tecnologias de informação, monitoramento e inteligência artificial (IA) vêm cada vez mais fazendo parte da vida de indígenas, especialmente aqueles que vivem no contexto de economias desenvolvidas e de transição. Dito isso, ao passo que tecnologias movidas por dados podem estimular a inovação e melhorar o bem-estar humano em geral, indígenas dificilmente compartilharão igualmente destes benefícios, considerando sua posição quase universal de marginalização socioeconômica, cultural e política. O uso crescente de big data conectada e integrada por governos e empresas também traz consigo riscos significantes para indígenas. Dentre estes riscos temos a apropriação de conhecimento cultural e propriedade intelectual, a exploração de terras e outros recursos naturais, além da perpetuação da discriminação, estigmatização e marginalização contínua. Tais riscos são ampliados por práticas de narrativa jornalística que reciclam tropos muito utilizados sobre disfunção indígena. Neste capítulo discutiremos alguns dos possíveis danos atrelados à digitalização e como a soberania de dados indígenas (ID-SOV, na sigla em inglês), foco emergente de ciência e ativismo, pode lidar com estas questões ao mesmo tempo que oferece caminhos para o benefício destas pessoas. Concluímos com a sugestão de que a pesquisa e as redes voltadas à ID-SOV também são fontes valiosas de dados e conhecimento em dados que podem levar a abordagens mais igualitárias, críticas e justas do jornalismo voltado a povos indígenas e suas questões.

Povos indígenas e dados

Estima-se que existam 370 milhões de indígenas espalhados pelo mundo, cobrindo todos os continentes e falando milhares de línguas distintas (ONU, 2009) É impossível determinar o número global exato, já que a maioria dos países que abarcam estes povos indígenas não os identifica em seus censos nacionais (Mullane-Ronaki, 2017). Não bastassem estes "desertos de dados" indígenas e a variação significante em termos de autonomia política e condições de vida destes povos pelo mundo, há farta evidência de que estes indígenas estão entre as camadas mais pobres de seus países, às voltas com doenças, encarceramento excessivo e desigualdade de amplo espectro (Anderson et al., 2016; Stephens et al., 2005). Esta marginalização não é coincidência e está diretamente relacionada à história destes povos enquanto colonizados e destituídos. Porém, as consequências devastadoras do colonialismo e seus consortes, supremacia branca e racismo, raramente são reconhecidas, menos ainda criticadas, em material jornalístico dos indígenas e suas comunidades na grande mídia.

Os povos indígenas sempre foram ativos em relação àquilo que agora conhecemos como dados, com suas antigas tradições de registro e proteção de informações e conhecimento por meio de arte, gravuras, totens, músicas, cânticos, dança e oração, por exemplo. Esforços deliberados para expurgo destas práticas e sistemas de conhecimento são integrais a processos de colonização. Ao mesmo tempo, os indígenas passaram a ser conhecidos através dos escritos de viajantes, exploradores e cientistas europeus, considerados "conhecedores" mais objetivos, científicos e credíveis destes povos e suas culturas. Ao longo do tempo, hierarquias raciais que justificavam e sustentavam o colonialismo foram naturalizadas e incutidas através de estruturas ideológicas, arranjos institucionais (como escravatura e segregação) e práticas de classificação estatais. Tomemos por exemplo o caso dos aborígenes e ilhéus do Estreito de Torres, na Austrália, que foram excluídos do censo nacional até 1971. Isso estava ligado à exclusão de direitos civis básicos, como aposentadoria por idade (Chesterman e Galligan, 1997). Em tempos modernos, o poder de decisão sobre se e como indígenas são contados, classificados, analisados e quais ações serão tomadas ainda depende de governos e não destes povos. Mudar o foco de poder sobre os dados indígenas do estado para os povos é o que move a ID-SOV.

Definição de ID-SOV

A terminologia em torno da questão da ID-SOV é relativamente recente, com a primeira publicação relevante sobre o tópico surgindo em 2015 (Kukutai e Taylor, 2015). A ID-SOV volta-se para os direitos dos indígenas de posse, controle e acesso a dados derivados destes, relacionado aos seus membros, sistemas de conhecimento, costumes ou territórios. (FNIGC, 2016; Kukutai e Taylor, 2016; Snipp, 2016).[50] Apoia-se nos direitos inerentes aos povos indígenas de autonomia e governança sobre seus povos, nações (incluindo terras, águas e céu) e recursos como descritos na Declaração Sobre os Direitos dos Povos Indígenas das Nações Unidas (UNDRIP).[51] Implícito na ID-SOV está o desejo de utilizar dados de forma a apoiar e aprimorar o bem-estar coletivo e autonomia de povos indígenas, sentimento enfatizado por ONGs indígenas, comunidades e aldeias (FNIGC, 2016; Hudson, 2016). Na prática, isso significa que estes povos devem ter o poder de decisão sobre o uso e implementação de dados a seu respeito. Assim sendo, a ID-SOV levanta questões como: a quem pertencem os dados? Quem tem o poder de decisão sobre o acesso e circunstâncias de acesso a estas informações? Quem são os beneficiários desejados destes dados e suas aplicações?

---

[50] Em Aotearoa, Nova Zelândia, a rede ID-SOV Te Mana Raraunga define dados maori como "informação digital ou digitalizável de conhecimento sobre ou de pessoas maori, nossa língua, culturas, recursos ou ambientes" (Te Mana Raraunga, 2018a).

[51] https://www.un.org/esa/socdev/unpfii/documents/DRIPS_en.pdf.

Há, ainda, preocupação em torno de temas espinhosos, como o equilíbrio de direitos individuais (como o direito à privacidade), riscos e benefícios ligados aos grupos dos quais estes indivíduos fazem parte. O foco em direitos e interesses coletivos é relevante pois transcende a visão estreita de proteção e controle de dados pessoais que permeia abordagens político-regulatórias como a Regulamento Geral Sobre a Proteção de Dados da União Europeia (RGPD). Conceitos legais anglo-europeus de privacidade individual e propriedade não se aplicam muito bem a contextos indígenas em que indivíduos integram um grupo maior definido, por exemplo, por genealogia ou genes compartilhados. Em tais contextos, o compartilhamento de dados que codifica informação sobre outros membros do grupo não pode ser baseado somente em consentimento pessoal, devendo levar em consideração direitos e interesses coletivos (Te Mana Raraunga, 2018). Ligado intimamente à ID-SOV está o conceito de governança de dados indígenas, que pode ser definido como os princípios, estruturas, mecanismos de responsabilização, instrumentos legais e políticas pelas quais os povos indígenas controlam seus dados (Te Mana Raraunga, 2018). Em sua essência, é uma forma de operacionalizar a ID-SOV (Carroll Rainie, Rodriguez-Lonebear e Rodriguez, 2017). É através de uma governança de dados indígena que os interesses e direitos destes povos em relação a dados podem ser reivindicados (Walter, 2018).

Vigilância estatística e povos indígenas

Delinear perfis de povos indígenas e segmentar serviços não é novidade — a vigilância por parte do estado, suas instituições e agentes há muito se apresenta como característica longeva do colonialismo (Berda, 2013). Mesmo com a exclusão oficial dos aborígenes e ilhéus do Estreito de Torres do censo nacional na Austrália, a vigilância em torno da população aborígene era um processo constante (Briscoe, 2003). A novidade na arena de política social são os processos opacos, complexos e cada vez mais automatizados responsáveis por traçar estes perfis e segmentações (Henman, 2018). No papel de “assuntos de dados” (Van Alsenoy et al., 2009), os povos indígenas se incluem em uma série de agregações de dados, de grupos sociais e políticos autoidentificados (tribos, grupos étnicos ou raciais), a bolsões de interesse definidos por analistas de dados com base em características, comportamentos e/ou circunstâncias.

A posição dos povos indígenas dentro deste processo não é benigna. Veja bem, ao passo que as fontes de dados sobre estes povos evoluem rapidamente, as características destes dados enquanto um relato impiedoso das desigualdades socioeconômicas e sanitárias em que estas pessoas vivem permanecem as mesmas. Walter (2016) chamou isso de dados 5D — dados que focam em diferença, disparidade, desvantagem, disfunção e desamparo. Evidências em apoio a esta afirmação são facilmente encontradas no Google ao buscar por “estatísticas indígenas” ou ao procurar por um povo específico, como nativos americanos, aborígenes e ilhéus do Estreito de Torres, maori, nativo havaiano, primeiras nações e nativos

do Alasca. Invariavelmente, estes resultados geram uma lista detalhando uma representação excessiva de indígenas em dados negativos ligados a saúde, educação, pobreza e encarceramento.

O impacto dos dados 5D na vida destas pessoas também não é positivo. Sendo esta a representação principal dos indígenas na narrativa nacional, tais dados moldam a forma como a população dominante, não indígena, compreende estes povos. As informações que influenciam estas narrativas são disseminadas frequentemente pela mídia. Stoneham (2014) discute um estudo sobre todos os artigos relacionados à saúde aborígene em quatro proeminentes veículos online e impressos da Austrália. Três quartos das matérias publicadas eram negativas, com foco em temas como álcool, abuso infantil, drogas, violência, suicídio e crimes, com apenas 15% dos textos considerados positivos (e 11% neutros); uma proporção de sete artigos negativos para um positivo. Tais narrativas são, em grande parte, retiradas de seu contexto social e cultural, analisadas de maneira simplista em que a população indígena é sistematicamente comparada à (não declarada) norma não indígena (Walter e Andersen, 2013). Isso resulta em um retrato negativo de povos indígenas aos olhos do país, pois são representados como um problema e não pessoas que lidam com o fardo da desigualdade histórica e contemporânea.

Há evidências crescentes de que o viés racial incorporado à big data e os algoritmos criados para sua análise acabarão por ampliar, e não reduzir, o impacto dos dados 5D em povos indígenas (ver Henman, 2015). Desta forma, em estados colonizados altamente desenvolvidos — como Aotearoa (Nova Zelândia) e Austrália —, os resultados prejudiciais de políticas discriminatórias se desenredaram até certo ponto, através de ativismo indígena e movimentos de justiça social. Ao longo de muitos anos, estas práticas de dados emergentes podem, sem querer, entrincheirar desigualdades pré-existentes e reativar antigos padrões. Com a detecção (e melhoria) de problemas sociais cada vez mais relegados a algoritmos, a probabilidade da injustiça retornar ao sistema de forma a afetar povos indígenas aumenta exponencialmente. Cabe, aqui, relembrar e retrabalhar o velho adágio em torno de dados, caso o algoritmo defina como problemas essenciais questões em que os indígenas estão super-representados, então o ‘indígena problemático’ será o alvo.

ID-SOV na prática

Há movimentos de ID-SOV ativos em países como Canadá, Austrália, Nova Zelândia e Estados Unidos, de crescente influência. Os pioneiros na prática foram as Primeiras Nações, do Canadá. Cansados de usuários não indígenas assumindo o manto como “especialistas”, sem qualquer viés ao tratar de questões da população nativa, ativistas da comunidade criaram um novo modelo que deu ao coletivo o controle sobre seus próprios dados. Os princípios OCAP® registrados reivindicam seu direito de reter propriedade,

controle, acesso e posse coletiva de dados sobre si mesmos e, após vinte anos, tornou-se o padrão adotado para condução de pesquisas ligadas às Primeiras Nações (First Nations Indigenous Governance Centre, 2016). Em Aotearoa (Nova Zelândia), a Rede de Soberania de Dados Maori Te Mana Raraunga foi fundada em 2015, reunindo mais de 100 pesquisadores, profissionais de saúde tradicional e empreendedores maori em setores como pesquisa, TI, comunidade e ONGs.[52] Bastante ativa na conscientização da necessidade da soberania e governança de dados sobre os maori ao longo do setor público, a iniciativa abordou a agência estatística nacional em 2018 para questionar seu manejo do Censo Neozelandês (Te Mana Raraunga, 2018b), fato amplamente coberto pela grande mídia e pela mídia indígena. A TMR também abordou questões ligadas à "licença social" de dados no contexto dos dados maori (Te Mana Raraunga, 2017) e criou seus próprios princípios de soberania de dados de forma a guiar o uso ético destas informações (Te Mana Raraunga, 2018a). Para defensores da soberania de dados dos maori, incluindo a TMR, o objetivo não é proteger apenas indivíduos e comunidades de danos e estigmatização futuros, mas proteger o conhecimento e os direitos de propriedade intelectual dos maori e garantir investimentos públicos em dados que possam gerar benefícios e valor de maneira justa e igualitária, de forma que este povo possa desfrutar completamente.

Já na Austrália, o Coletivo de Soberania de Dados Indígenas Maiam nayri Wingara surgiu em 2016 e, em 2018, numa parceria com o Instituto Australiano de Governança Indígena, lançou um comunicado em um encontro nacional com líderes aborígenes e ilhéus do Estreito de Torres. O comunicado delineava a demanda por controle e poder de decisão indígena sobre o ecossistema de dados, o que incluía criação, desenvolvimento, administração, análise, disseminação e infraestrutura destes. Junto a outras entidades indígenas, Maiam nayri Wingara vem ativamente buscando promover mudanças na forma como dados ligados a estes povos são conceitualizados, propostos, implementados, construídos, analisados e interpretados na Austrália. Sua aspiração é colocar em prática a contribuição que os dados podem dar para o bem-estar dos aborígenes e ilhéus do Estreito de Torres. Para que isso aconteça, é necessário reinventar a relação entre geradores/detentores de dados indígenas e os povos ligados a estas informações, de forma a criar uma relação em torno de uma governança indígena.

Rumo a um papel mais relevante de iniciativas em ID-SOV em jornalismo de dados

O jornalismo de dados está em uma posição privilegiada de contestar ao invés de reforçar os dados 5D na questão indígena. Jornalistas de dados têm amplas oportunidades para repensar como usam informações para representar povos indígenas e suas histórias, e

[52] https://www.temanararaunga.maori.nz/tutohinga/.

expor a complexidade das formas pelas quais os dados indígenas são produzidos, controlados, disseminados e aplicados por governo e indústria. Ao fazê-lo, cabe aos profissionais não dependerem somente de produtores/usuários de dados não indígenas. A ascensão das redes de ID-SOV representa um número crescente de especialistas indígenas com os quais se pode contar. Muitos daqueles envolvidos com ID-SOV mantêm laços estreitos com suas comunidades, movidos por um forte compromisso com justiça de dados e encontrar meios para que "bons dados" levem a "bons resultados". As questões levantadas pela ID-SOV, especialmente sobre propriedade, controle, danos e benefício coletivo de dados, têm implicações mais amplas para além das comunidades indígenas. Ao engajar com abordagens e princípios de ID-SOV, jornalistas de dados podem abrir espaço relevante para perspectivas e preocupações indígenas, possibilitando o enquadramento de suas narrativas ao mesmo tempo que preparam o terreno para responsabilizar aqueles no poder.

Tahu Kukutai é professora do Instituto Nacional de Análise Demográfica e Econômica da Universidade de Waikato. Maggie Walter (PhD) é de etnia palawa (aborígene tasmaniana), socióloga e uma das fundadoras do Coletivo Australiano de Soberania de Dados Maiam nayri Wingara.

Referências:

ANDERSON, I. et. al. *Indigenous and tribal people's health (The Lancet and Lowitja institute Global Collaboration): a population study*. The Lancet, 388(10040), 2016, p. 131-157.

BERDA, Y. *Managing dangerous populations: colonial legacies of security and surveillance*. Sociological Forum, 28 (3), 2013, p. 627.

BRISCOE, G. *Counting, health and identity.: A history of Aboriginal health and demography in Western Australia and Queensland 1900-1940*. Canberra: Aboriginal Studies Press, 2003.

CARROLL RAINIE, S.; RODRIGUEZ-LONEBEAR, D.; MARTINEZ, A. *Policy Brief (Version 2): Data Governance for Native Nation Rebuilding*. Tucson: Native Nations Institute, 2017. Disponível em: usindigenousdata.arizona.edu.

CHESTERMAN, J.; GALLIGAN, B. *Citizens without rights, Aborigines and Australian citizenship.* Cambridge: Cambridge University Press, 1997.

Comunicado apresentado durante Congresso de Soberania de Dados Indígenas desenvolvido pelo Coletivo de Soberania de Dados Indígenas Maiam nayri Wingara e

Instituto de Governança Indígena Australiano. Junho de 2018. Disponível em: http://www.aigi.com.au/wp-content/uploads/2018/07/Communique-Indigenous-Data-Sovereignty-Summit.pdf.

First Nations Information Governance Centre. *Pathways to First Nations' data and information sovereignty*. In: Kukutai, T; Taylor, J. (ed.). *Indigenous Data Sovereignty: Toward an Agenda*. Canberra: Australian National University Press, 2016, p. 139-155.

HUDSON, M.; FARRAR, D.; MCLEAN, L. *Tribal data sovereignty: Whakatōhea rights and interests*. In: KUKUTAI, T.; TAYLOR, J. (ed.). *Indigenous Data Sovereignty: Toward an Agenda*. Canberra: Australian National University Press, 2016, p. 157–178.

MULLANE-RONAKI, M. *Indigenising the national census? A global study of the enumeration of indigenous peoples, 1985-2014*. Tese de mestrado não publicada. Hamilton: Universidade de Waikato, 1999.

SMITH, L. *Decolonizing methodologies: Research and indigenous peoples.* Londres: Zed Books, 1999.

STEPHENS, C. et al. *Disappearing, displaced, and undervalued: a call to action for Indigenous health worldwide*. The Lancet, 367: 2019–2028, 2006.

STONEHAM, M. *Bad news: negative Indigenous health coverage reinforces stigma*. The Conversation, 2 de abril de 2014. Disponível em: http://theconversation.com/bad-news-negative-Indigenous-health-coverage-reinforces-stigma-24851.

TE MANA RARAUNGA. *Declaração sobre licença social*. 2017. Disponível em: https://www.temanararaunga.maori.nz/panui/.

_____. *Princípios de soberania de dados maori*. 2018ª. Disponível em: https://www.temanararaunga.maori.nz/new-page-2/.

_____ *Declaração do Te Mana Raraunga sobre o censo neozelandês de 2018: Uma chamada em prol do censo de dados maori*. 2018b. Disponível em: https://www.temanararaunga.maori.nz/panui/.

Organização das Nações Unidas. *State of the World's Indigenous Peoples*. Nova York: Organização das Nações Unidas, 2009. Disponível em: https://www.un.org/esa/socdev/unpfii/documents/SOWIP/en/SOWIP_web.pdf.

VAN ALSENOY, B. et al. *Social networks and web 2.0: are users also bound by data protection regulations?* Identity in the Information Society, *2*(1), 2009, p. 65-79.

WALTER, M. *Data politics and Indigenous representation in Australian statistics*. In: KUKUTAI; T.; TAYLOR, J. (ed.). *Indigenous Data Sovereignty: Toward an Agenda*. Canberra: Australian National University Press, 2016, p. 79-98.

WALTER, M. *The voice of Indigenous data: Beyond the markers of disadvantage*. Griffith Review, 60, 2018.

Processos de pesquisa em investigações jornalísticas

Crina Boros

Será que dado problema é anedótico ou sistemático? Essa é uma questão que você se faz quando percebe que não há dados tabulares sobre o tema em questão, um termo chique que quer dizer que as informações não estão disponibilizadas em linhas e colunas. O que fazer?

O que são dados, afinal? Há muitas definições meio de nerd por aí, algumas delas até intimidadoras.[53] Substituamos estas definições por um conceito simples: informação. Ao passo que você coleta estas informações, em qualquer modelo ou formato, é necessário identificar padrões e pontos fora da curva. Isso significa que você precisa ter um volume considerável de material cru reunido sistematicamente que documente um problema seguindo um método específico (pense em formulários). Não importa se você usa uma planilha, um ambiente de programação, um aplicativo ou lápis e papel.

Por vezes, pensamentos, sentimentos ou experiências íntimas do passado presas dentro dos corações e mentes das pessoas podem ser articulados como dados. Um método para coleta desta informação tão preciosa é uma pesquisa que reuniria e ordenaria tais sentimentos e experiências em tabelas, arquivos ou bancos de dados aos quais ninguém mais teria acesso, além de você.

Por exemplo, a Fundação Thomson Reuters (TRF, na sigla em inglês) desenvolveu um projeto sobre como mulheres nas maiores capitais do mundo percebiam o efeito da violência sexual no transporte público sobre elas.[54] Isso envolveu todo um esforço de pesquisa para uma maior conscientização sobre o tema, mas este mesmo esforço também envolvia comparar e contrastar informações, o que normalmente se faz em estatística.

Para entregar este material, passamos pelos mais variados círculos do inferno, já que há convenções estritas exigidas por métodos científico-sociais como a pesquisa, mesmo quando adotados por jornalistas em seu ofício.

Eis aqui algumas regras cujo conhecimento prévio beneficiaria jornalistas, que raramente recebem treinamento nesse sentido.

---

[53] Consultar as 130 definições de dados, informação e conhecimento em Zins (2007), p. 58, 479–493.

[54] http://news.trust.org/spotlight/most-dangerous-transport-systems-for-women/?tab=stories.

Não se escolhem participantes a dedo. De forma a ser considerado "representativo", o grupo de participantes normalmente incluiria pessoas de todos os grupos sociais, etários, educacionais e localizações sobre os quais queremos falar. Seguindo a metodologia estabelecida, amostras da população estudada devem ser representativas.

A seleção de participantes precisa ser randomizada. Ou seja, todos devem ter a mesma chance em um sorteio. Caso esteja fazendo uma pesquisa e conversando com quem for mais conveniente, sem qualquer critério ou método, há o risco de se produzir dados que induzam ao erro, especialmente caso busque fazer afirmações mais gerais.

O número de participantes em uma pesquisa deve superar certo limiar para ser considerado representativo. Há calculadoras online, como as fornecidas por Raosoft, Survey Monkey e Survey Systems, que podem ajudar.[55] Regra de ouro: manter o nível de confiança em 95% e a margem de erro não deve superar 5.

Deve-se permitir aos participantes não saberem ou não terem certeza de algo. Ao seguir estas regras básicas, as descobertas feitas pelos jornalistas serão quase que indestrutíveis. À época da pesquisa sobre segurança no transporte público da TRF, nossa metodologia seguiu à risca as regras conservadoras das ciências sociais. Abordamos uma experiência humana comum que diz muito sobre como as sociedades funcionam, tanto é que uma agência das Nações Unidas se ofereceu para trabalhar junto conosco. Uma honra e tanto, mas que tivemos que negar, no papel de jornalistas.

Se isso soou bem aos seus ouvidos, é hora de fazer um curso de estatística.

Às vezes, fazer uma pesquisa rigorosa não é nada realista. O que não quer dizer que estas pesquisas não devam ser feitas.

Por mais que existam métodos estabelecidos para tanto, este corpo metodológico não exaure possibilidades, legitimidades ou interesses. Pode haver outras formas de pesquisar, a depender de suas preocupações, restrições e recursos.

Por exemplo, quando o *openDemocracy* quis entrevistar repórteres dos 47 estados integrantes do Conselho Europeu a respeito de pressões comerciais dentro das redações, havia poucas chances de haver significância estatística.

Aí você pode perguntar: por quê?

---

[55] http://www.raosoft.com/samplesize.html, https://www.surveymonkey.com/mp/sample-size-calculator/, https://www.surveysystem.com/sscalc.htm.

Todos os participantes se tornaram delatores. Delatores precisam de proteção, o que inclui não divulgar dados demográficos relevantes, como idade ou gênero. Esperávamos algumas contribuições vindas de países em que o exercício da liberdade de expressão pode levar a sérias consequências pessoais. Decidimos que o fornecimento de dados pessoais não seria obrigatório e, caso fossem fornecidos, não seriam armazenados em um servidor de uma empresa que atuaria como coproprietária de nossas informações.

Os dados da União Europeia variam e se mostram incompletos, tratando-se de jornalistas na região. Ou seja, estabelecer uma amostra representativa por país seria complicado.

Não era possível reunir todos os sindicatos e associações de imprensa para então randomizar os participantes, pois as listas de membros são privadas. Além do que, estas listas não incluem todos os membros, ainda que fosse aceitável tomá-las como base, contanto que fôssemos honestos quanto às nossas limitações. Em alguns países, projetos de transparência acabam por levar à repressão e recebemos dicas de gente experiente sobre em quais países não poderíamos contar com o apoio de sindicatos sem atrair algum tipo de vigilância ou punição.

Nestes casos, não era necessário jogar o bebê junto com a água do banho. Nós não jogamos.

Ao invés disso, identificamos o que era relevante para nossa reportagem e como os métodos de pesquisa poderiam ser ajustados para entregarmos estas histórias.

Decidimos que nosso principal foco seriam exemplos de pressão comercial dentro de redações nacionais; se havia um *padrão como estas ocorriam* e se estes padrões *se repetiam ao longo da região investigada*. Também nos interessava os tipos de entidades acusadas de lavagem de imagem junto à imprensa.

Seguimos em frente e desenvolvemos uma pesquisa, baseada em entrevistas com jornalistas, relatórios sobre liberdade na mídia e feedback obtido junto a grupos focais. Incluímos seções com respostas em aberto.

Levamos essa pesquisa a todos os canais de organizações jornalísticas previamente avaliados. Em essência, não estávamos randomizando nada, mas tampouco tínhamos controle sobre quem estava participando. Também tínhamos parceiros como a Repórteres Sem Fronteiras, o Sindicato Nacional dos Jornalistas do Reino Unido e a Federação Europeia de Jornalistas, que ajudaram a divulgar o questionário.

O retorno da pesquisa foi adicionado a um banco de dados único, com pontuações atribuídas às respostas e contabilização de participantes por país, sendo possível comparar evidências anedóticas (questões relatadas esporadicamente) e problemas sistêmicos (relatados com maior frequência e abrangência).

Campos de texto em aberto se mostraram particularmente úteis, já que os participantes os usavam para nos dar dicas. Pesquisamos as informações obtidas, com um olhar atento a padrões de censura econômica e tipos de supostos malfeitores. Com base nisso, escrevemos nossa reportagem sobre liberdade de imprensa.[56]

Por mais que tenhamos publicado um apanhado das descobertas, nunca divulgamos uma análise aprofundada dos dados pelo simples fato de que eles não foram randomizados e em alguns casos não havia amostragem suficiente por país. [57] Mas pudemos criar um sólido entendimento de como é a imprensa livre de acordo com quem a faz, como ocorre a corrupção midiática, como esta evolui e, tristemente, o quão vulneráveis são tantos repórteres quanto à própria verdade.[58] (Inclusive, caso queira relatar algo, ainda é possível participar da pesquisa.[59])

Então existem regras para quebrar as regras?

Algumas. Sempre descreva seu trabalho com precisão. Caso tenha feito uma pergunta do tipo sim ou não a três grandes conselheiros econômicos do governo, deixe isso claro. Caso tenha entrevistado dez vítimas de bullying, descreva como as escolheu e por que as escolheu. Não trate entrevistas como pesquisas tão facilmente.

Caso faça um estudo com estatísticas relevantes, faça o favor de divulgar sua metodologia.[60] Isso oferece o nível de escrutínio necessário para que o público e especialistas confiem no seu material. Sem metodologia, sem confiança.

Não seja o próximo grande criador de fake news. Se o editor está te forçando a fazer correlações baseadas em inferências e não na coleta de dados precisos, use a linguagem de forma a não sugerir causalidade ou rigor científico. Nosso trabalho é falar da verdade, não apenas soltar fatos. Não use fatos para acobertar uma possível incerteza sobre a verdade.

---

[56] https://www.opendemocracy.net/author/crina-boros.

[57] https://www.opendemocracy.net/openmedia/mary-fitzgerald/welcome-to-openmedia.

[58] https://www.opendemocracy.net/openmedia.

[59] https://www.surveymonkey.co.uk/r/MediaFreedom2017_English.

[60] http://news.trust.org/spotlight/most-dangerous-transport-systems-for-women/?tab=methodology.

Onde está a sua história, em um padrão? Em um ponto fora da curva? Decida que dados precisam ser coletados com base na resposta a estas perguntas. Descubra como e quando estas informações podem ser obtidas antes de definir os métodos mais adequados para tanto. O método nunca é um fim em si, mas sim a narrativa.

Ao realizar uma pesquisa, ponha suas descobertas à prova e proteja sua reportagem contra declarações problemáticas em potencial. Por exemplo, digamos que uma pesquisa sugira que a parte da cidade em que você mora tem o maior índice de crimes, mas você se sente seguro e testemunhou violência quase que semanalmente em outra vizinhança onde morou por um ano, então talvez não dê para confiar nos dados ainda. Antes disso, visite os lugares a serem comparados e contrastados; converse com as pessoas na rua, em comércios, bancos, bares e escolas; observe os dados coletados; os moradores desta região estão mais propensos a fazerem queixas que os de outra área? De quais tipos de crime estamos falando? Considere os tipos de crime inclusos nesta análise, ou teria um furto o mesmo peso que um homicídio? Tais esforços de fundamentação permitirão uma melhor avaliação dos dados e decidir até que ponto se pode confiar nos resultados de análises mais aprofundadas.

Crina Boros é repórter investigava e educadora em jornalismo de dados especializada em investigações de interesse público.

Trabalhando com dados

Jornalismo de dados: o que o feminismo tem a ver com isso?

Catherine D'Ignazio

Por conta de avanços na tecnologia ao longo dos últimos 70 anos, as pessoas podem armazenar e processar mais informações do que nunca. As mais bem-sucedidas empresas do setor no mundo — Google, Facebook, Amazon, Microsof, Apple — ganham dinheiro com a agregação de dados. Nos setores público e privado, cada vez mais se dá valor às decisões tomadas com "base em dados". Dados são poderosos — pois são lucrativos e valorizados pelos poderosos —, mas não são distribuídos igualitariamente, nem as habilidades necessárias para lidar com estes, muito menos os recursos tecnológicos usados em sua armazenagem e processamento. Quem trabalha com dados não representa a população em geral. São, normalmente, homens brancos, localizados ao norte e com educação superior.

Precisamente por conta destas desigualdades do ecossistema de dados, adotar uma abordagem feminista no jornalismo feito em torno dos mesmos pode ajudar a desvendar vieses ocultos na produção de informação. Uma definição simples de feminismo é que se trata da crença na igualdade social, política e econômica dos gêneros cuja atividade organizada se dá em torno desta crença. Conceitos e ferramentas feministas podem ser úteis no questionamento de dinâmicas de poder social, tomando o gênero como dimensão central (não única) da análise. Uma das características principais do feminismo contemporâneo é sua insistência na *interseccionalidade* — a ideia de que devemos considerar não apenas o sexismo, mas também o racismo, o classismo, o capacitismo e outras forças estruturais ao pensarmos como desequilíbrios na balança do poder podem obscurecer a verdade.[61] Para jornalistas que se identificam com aquilo que a profissão convencionou como "confrontar o poder com a verdade", uma abordagem feminista pode parecer bastante familiar.

Este ensaio cobre diversos estágios do processamento de dados — coleta, contexto, comunicação — e aponta gargalos na questão dos vieses e oportunidades para aplicação de uma visão feminista. Cabe notar que uma abordagem feminista não é útil somente em dados ligados a mulheres ou questões de gênero, mas para quaisquer projetos envolvendo seres ou

[61] De fato, o feminismo que não considera como demais fatores identitários se relacionam com questões de gênero deveria ser encarado como "feminismo branco". O conceito de interseccionalidade foi criado pela estudiosa norte-americana Kimberlé Crenshaw, fruto de um legado intelectual de feministas negras que postularam que a desigualdade de gênero não pode ser levada em conta separadamente da desigualdade de raça e classe.

instituições humanas (leia-se: praticamente qualquer projeto), porque onde temos indivíduos, temos desigualdade social.

Coleta de dados

Examinar o poder — como funciona e a quem beneficia — sempre esteve no coração de projetos feministas. A socióloga Patricia Hill Collins, com seu conceito de *matriz de dominação* nos ajuda a compreender que o poder é complicado, e que "raras são as vítimas e opressores puros". Ao passo que pensamos na injustiça no campo interpessoal (caso de um comentário machista), há forças sistêmicas que precisamos entender e expor (caso de machismo em instituições que coletam dados) para causarmos mudanças.

Há duas formas pelas quais a desigualdade se revela na coleta de dados. Primeiro, entidades específicas são *superestimadas* no processo de coleta de informações. Essa super contagem tipicamente se relaciona à vigilância praticada por aqueles no poder versus aqueles em posições menos favorecidas. Por exemplo, a polícia de Boston liberou informações sobre seu programa de abordagens em 2015. Os dados revelam que a polícia patrulha, desproporcionalmente, vizinhanças negras, latinas e de imigrantes, abordando, de forma igualmente desproporcional, jovens negros do sexo masculino. Em casos como estes é importante saber quais grupos estão no poder e quais grupos são alvos prováveis de vigilância. O papel do jornalista de dados é reconhecer e quantificar a disparidade, bem como nomear as forças estruturais em ação — neste caso, o racismo.

A segunda forma pela qual a desigualdade estrutural aparece na coleta de dados é na *subcontagem* ou ausência total de contagem. Por exemplo, por que o mais abrangente banco de dados de feminicídios (homicídios motivados por gênero) do México está sendo alimentado por uma mulher sob o pseudônimo Princesa?[62] Mesmo com o número de mortes de mulheres em Ciudad Juárez e no resto do país crescendo cada vez mais, mesmo com a criação de uma comissão especial de feminicídio em 2006, mesmo com a decisão em 2009 contra o estado mexicano pela Corte Interamericana de Direitos Humanos, o país não monitora feminicídios de forma ampla. A subcontagem ocorre quando muitas questões ligadas às mulheres e pessoas de cor são negligenciadas, sistematicamente, pelas instituições que não consideram os danos pelos quais elas mesmas são responsáveis, ao não realizarem a contagem correta. Ou seja: o ambiente de coleta de dados está comprometido. Em casos de subcontagem, jornalistas podem fazer exatamente como Princesa: contarem eles mesmos, da melhor forma possível. Geralmente, isso envolve um tanto de colaboração coletiva, coleta de informações e inferência estatística. No contexto dos EUA, outros exemplos de subcontagem

[62] https://feminicidios.mx.

incluem homicídios cometidos pela polícia e mortalidade maternal, ambos abordados em projetos de coleta de dados por jornalistas.

Contexto de dados

Por mais que o movimento em torno da livre circulação de dados e a proliferação de APIs possam parecer algo bom para os jornalistas de dados, as informações obtidas "em campo" trazem consigo suas próprias questões, ainda mais quando estamos falando de fenômenos humanos e sociais. A filosofa feminista Donna Haraway afirma que todo conhecimento é "situado", ou seja, sempre está situado em um contexto social, cultural, histórico e material. Desenredar e investigar como conjuntos de dados são produtos destes contextos pode nos ajudar a entender as maneiras pelas quais poder e privilégios podem mascarar a verdade.

Um exemplo: meus alunos da aula de Visualização de Dados queriam fazer seu projeto final a respeito de assédio sexual cometido em câmpus.[63] Faculdades e universidades nos EUA são obrigadas a relatarem assédios sexuais e outros crimes em seus câmpus anualmente por conta da Lei Clery, o que indicava que poderia haver um banco de dados nacional extenso sobre o tema. Mas os dados ligados a esta lei deram uma dorzinha de cabeça para os alunos — o Williams College, por exemplo, apresentava números extremamente altos em comparação a outras faculdades urbanas. Ao investigar o contexto e questionar a estruturação do ambiente de coleta, os estudantes descobriram que os números contavam uma história possivelmente *oposta* à verdade. Assédio sexual é um tema estigmatizado cujas vítimas temem ser culpadas pelo abuso e por possíveis retaliações, o que faz com que não denunciem os casos. Logo, as instituições com maiores índices eram locais que haviam dedicado mais recursos para a criação de um ambiente em que as vítimas se sentiam seguras para se pronunciarem. Por outro lado, aquelas com índices menores de assédio sexual possuíam um clima hostil, em que vítimas não tinham apoio para romper o silêncio.

Aqui, temos um gargalo e uma oportunidade. O gargalo é que jornalistas pegam números na internet e os aceitam como são, sem compreensão das nuances do ambiente em que foram coletados, o que inclui relações de poder, estigmas sociais, e normas culturais de visibilidade perante instituições (grupos como mulheres, imigrantes e pessoas de cor geralmente desconfiam de instituições censitárias, e por um bom motivo). Já a oportunidade reside no fato de que há muitas histórias a serem contadas considerando o contexto dos dados. Antes de usar números na ânsia por novas análises, jornalistas de dados podem usar estes mesmos números para questionar o ambiente de coleta, apontar práticas defeituosas e

---

[63] O artigo final de autoria de Patrick Torphy, Michaele Gagnon e Jillian Meehan está disponível em: https://cleryactfallsshort.atavist.com/reporting-sexual-assault-what-the-clery-act-doesnt-tell-us.

desequilíbrios de poder, bem como mudar práticas de contagem de forma que as instituições registrem aquilo que realmente importa.

Comunicação de dados

O pensamento contemporâneo ocidental em torno dos dados evoluiu a partir de um "estereótipo mestre" em que aquilo que é encarado como racional e objetivo acaba por ser mais valorizado que o que é encarado como emocional e subjetivo (cabe, aqui, pensar qual dos gêneros é considerado "racional" e qual é considerado "emocional"). Este mesmo estereótipo afirma que as emoções nublam a capacidade de julgamento, e a distância potencializa a objetividade. Agora, uma perspectiva feminista põe em xeque tudo a respeito deste estereótipo. Emoções não nublam a capacidade de julgar algo — elas geram curiosidade, engajamento e incentivo para aprender mais. Patricia Hill Collins descreve como uma situação de conhecimento ideal aquela em que "nem emoção nem ética estão subordinadas à razão".

Figura 1: Gráfico por Nigel Holmes. De *Designer's Guide to Creating Charts and Diagrams*, 1984.

O que isso significa para a comunicação de dados? Por mais que práticas anteriores em visualização de dados favorecessem tabelas e gráficos minimalistas como mais racionais, tanto pesquisadores quanto jornalistas estão descobrindo que, ao empregarem as

características únicas da visualização como uma forma de retórica criativa, levam a visualizações mais memoráveis, portanto, compartilháveis. Tomemos como exemplo o gráfico "Monstrous Costs", criado por Nigel Holmes, em 1984, para ilustrar o custo crescente das campanhas políticas — antes considerado como exemplo de "gráfico lixo". Agora, pesquisadores provaram aquilo que a maioria de nós já sabia intuitivamente: alguns leitores gostam mais de monstros do que de gráficos chatos.

Como é o caso com qualquer meio de comunicação, valer-se de dados para causar emoções tem suas implicações éticas. Pesquisadores também demonstraram, recentemente, a importância do título na forma como as pessoas interpretam uma visualização. Títulos num geral tendem à racionalização, ou seja, colocam os dados como algo neutro e objetivo, se apresentado como "Relatos de assédio sexual em câmpus universitários entre 2012 e 2014" ou algo do tipo. Mas há diversos casos — mais uma vez, geralmente ligados a mulheres e outros grupos marginalizados — em que um título mais neutro acaba por prejudicar o grupo representado pelos dados. No caso de assédio e abuso sexual, um título neutro comunica, de maneira implícita, que os dados em questão são verdadeiros e completos, quando sabemos que não é este o caso. Em outras situações, um título neutro como "Mulheres portadoras de doenças mentais mortas em confrontos com a polícia entre 2012 e 2014" serve de oportunidade para a perpetuação de estereótipos danosos, precisamente por não citar as forças estruturais operantes, dentre as quais o capacitismo e o sexismo, que fazem destas mulheres vítimas desproporcionais da violência policial nos EUA.

Conclusão

Optar por uma abordagem feminista ao jornalismo de dados significa prestar atenção às formas com que instituições e práticas preexistentes favorecem um status quo em que homens da elite ficam no topo e os demais se espalham por diversas intersecções da matriz de dominação de Collins. Patriarcado, supremacia branca e colonialismo são forças estruturais e, sendo assim, se dão muito bem com investigações e visualizações movidas por dados. Precisamos questionar as informações recebidas pelo jornalismo de dados de forma a garantir que não estamos perpetuando, inadvertidamente, o status quo e, ao mesmo tempo, devemos usar as ferramentas à nossa disposição para expor e desmantelar injustiças. Quem estiver interessado em seguir por este caminho, pode consultar nosso livro *Data Feminism* (D'Ignazio, 2020), escrito junto com Lauren F. Klein, que discute com maiores detalhes como conceitos feministas podem ser aplicados à ciência e à comunicação de dados.

Catherine D'Ignazio é professora assistente de Visualização de Dados e Mídia Cívica do Emerson College, integrante sênior do Laboratório de Engajamento e docente visitante do Centro de Mídia Cívica do MIT.

Como o ICIJ lida com grandes volumes de dados como *Panama* e *Paradise Papers*

Emilia Díaz-Struck, Cécile C. Gallego e Pierre Romera

O Consórcio Internacional de Jornalistas Investigativos (ICIJ, na sigla em inglês) é uma rede internacional fundada em 1997. Os profissionais que participam de grandes esforços colaborativos do ICIJ têm os mais diferentes perfis e históricos. Há uma grande variedade de repórteres com diversas habilidades, alguns entendem muito de dados e programação, outros contam com fontes excelentes e são ótimos na apuração de campo. Todos unidos por um interesse comum em jornalismo, colaboração e dados.

Quando o diretor do ICIJ, Gerard Ryle, recebeu um disco rígido na Austrália recheado com informações corporativas ligadas a paraísos fiscais e gente de todo o mundo, resultante de sua investigação de três anos do escândalo da empresa Firepower, da Austrália, ele não fazia ideia de como aquilo transformaria a forma de se colaborar no jornalismo. Ele chegou na sede do ICIJ com mais de 260 gigabytes de dados úteis, cerca de 2,5 milhões de arquivos, que acabaram envolvendo mais de 86 jornalistas de 46 países em um enorme esforço colaborativo que ficou conhecido como *Offshore Leaks* (publicado em 2013).[64]

Após o caso dos *Offshore Leaks*, vieram mais projetos investigativos envolvendo enormes conjuntos de dados e milhões de arquivos, mais tecnologias desenvolvidas para lidar com estes projetos em específico e mais redes de jornalistas para cobrirem estes temas. Por exemplo, há pouco compartilhamos com nossos parceiros mais de 1,2 milhões de documentos vazados da mesma empresa envolvida no âmago do caso dos *Panama Papers*, Mossack Fonseca.[65] Essa pilha de informação se juntava aos outros 11,5 milhões de arquivos trazidos até nós em 2015 pelo jornal alemão *Süddeutsche Zeitung* e outros 13,6 milhões de documentos que serviram de base para a investigação seguinte, a dos *Paradise Papers*.[66]

Se um único jornalista parasse para ler, por um minuto, cada um dos arquivos dos *Paradise Papers*, ele levaria 26 anos para dar uma olhadinha em todos eles. Claro que essa não é uma expectativa das mais realistas. Então, nos perguntamos: como achar um atalho para isso? Como tornar a pesquisa mais eficaz e ágil? Como a tecnologia pode nos ajudar a

---

64 https://www.icij.org/investigations/offshore/how-icijs-project-team-analyzed-offshore-files/.

65 https://www.icij.org/investigations/panama-papers/.

66 https://www.icij.org/investigations/paradise-papers/.

encontrar novas pistas em meio a esta pilha gigantesca de documentos, ao mesmo tempo que apoia nosso modelo colaborativo?

Neste capítulo mostramos como nós lidamos com um sem-fim de arquivos vazados não somente através do uso de tecnologias voltadas à "big data", mas também com um aparato analítico caso a caso composto por (1) redes colaborativas internacionais; (2) práticas e infraestruturas de comunicação segura; (3) processos e métodos de criação de dados estruturados a partir de documentos não estruturados; e (4) uso de bancos de dados de gráficos e visualizações que possibilitam explorar conexões juntos.

1. Engajamento com parceiros

O modelo do ICIJ visa investigar o sistema tributário global através de uma rede internacional de jornalistas. Recrutamos repórteres de destaque em cinco continentes para aprimorar esforços de pesquisa e ligar os pontos (de dados) entre um país e outro.[67]

Artigos em torno de questões tributárias são como quebra-cabeças onde faltam peças: um jornalista na Estônia pode entender uma parte da história, enquanto um repórter brasileiro pode se deparar com outra. Reúna-as e tenha, assim, uma imagem melhor do todo. O trabalho do ICIJ é conectar estes repórteres e garantir que compartilhem tudo que encontrarem nos dados.

"Compartilhamento radical", é assim que chamamos nossa filosofia. Descobertas são compartilhadas entre os parceiros ICIJ em meio ao trabalho, não apenas com seus colegas de trabalho diretos, mas com jornalistas do outro lado do mundo.

De forma a promover esta colaboração, o ICIJ oferece uma plataforma de comunicação chamada I-Hub Global, baseada em software de código livre.[68] Ela já foi descrita por seus usuários como um tipo de "Facebook privado" e permite o compartilhamento direto de informações, semelhante ao que ocorreria em uma redação física. Cada jornalista integra um grupo que cobre temas específicos — países, esportes, artes, processos ou quaisquer outros tópicos de seu interesse. Dentro destes grupos, os participantes podem fazer postagens ainda mais específicas, como algo sobre um político encontrado em meio aos dados ou transação na qual estão trabalhando. É aqui que ocorre a maior parte da discussão, onde informações são cruzadas, e anotações e documentos interessantes são compartilhados entre repórteres.

---

[67] http://www.icij.org/journalists.

[68] https://www.icij.org/blog/2014/07/icij-build-global-i-hub-new-secure-collaboration-tool/.

Foram necessários vários projetos do ICIJ até que os jornalistas se sentissem à vontade com o I-Hub. Para melhor acomodar novos usuários e lidar com questões técnicas, os coordenadores regionais do consórcio oferecem suporte. Isso é essencial para garantir que os jornalistas atendam aos requisitos de segurança exigidos.

2. Criptografe tudo

Quando se conduz uma investigação envolvendo 396 jornalistas, é preciso ser realista no tocante à segurança: cada indivíduo é um alvo em potencial para gente mal-intencionada, e o risco de um vazamento é alto. Para reduzir tal risco, o ICIJ emprega diversas linhas de defesa.

Ao se juntar a qualquer investigação da ICIJ, é necessário criar um par de chaves PGP para criptografia de emails. O princípio é simples.[69] São duas chaves: uma delas pública e repassada a correspondentes em potencial, que podem usá-la para enviarem emails criptografados a você. A segunda chave é de uso privativo e nunca deve sair do seu computador. Esta chave privada tem um único propósito: a decodificação dos emails criptografados com a chave pública.

Pense no PGP como uma espécie de cofre em que outras pessoas podem guardar mensagens para você. Você é o único que tem a chave para abri-lo e acessar estas mensagens. Como toda medida de segurança, o PGP tem suas vulnerabilidades. Por exemplo, caso haja algum spyware no seu computador, registrando o que você digita ou analisando cada arquivo no seu HD, sua chave pode ser comprometida. Esta situação destaca a importância de empregar diversas camadas de defesa. Caso uma destas camadas seja rompida, esperamos que as demais diminuam o impacto de um vazamento ou uma invasão.

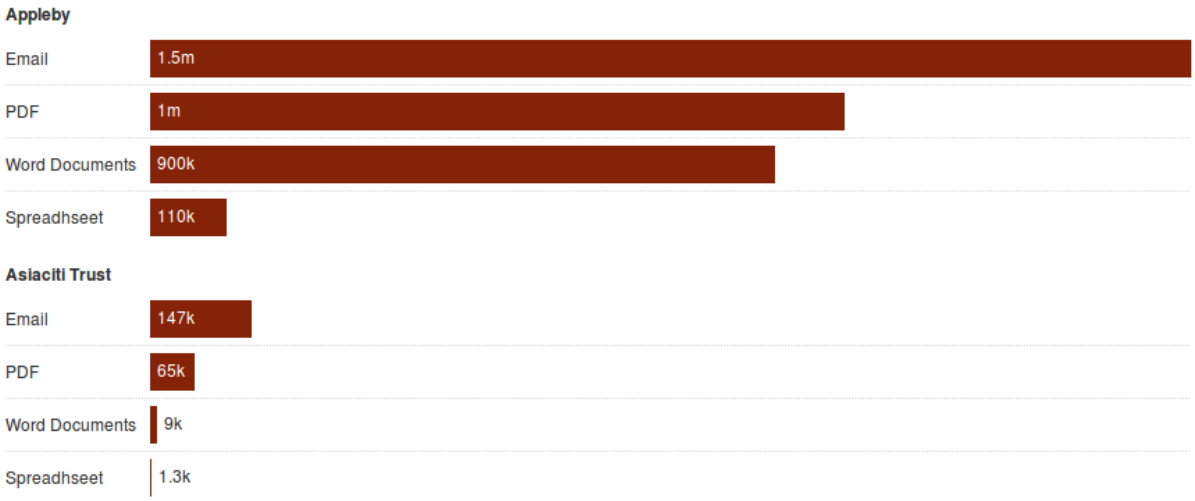
Para assegurar a identidade de seus parceiros, o ICIJ usa autenticação de dois fatores em todas as suas plataformas. Esta técnica é bastante popular em grandes sites como Google, Twitter e Facebook. Ela fornece ao usuário um segundo código, temporário, exigido em seu login, normalmente gerado em outro dispositivo (como um celular), que logo desaparece. Em algumas plataformas mais sensíveis, adicionamos um terceiro fator: o certificado do cliente. Basicamente, um pequeno arquivo armazenado e configurado por jornalistas em seus notebooks. Nossa rede negará acesso a qualquer aparelho que não tenha este certificado. Outro mecanismo digno de menção utilizado pelo ICIJ é o Ciphermail. Este software opera entre nossas plataformas e as caixas de entrada de nossos usuários, assegurando que toda comunicação recebida do ICIJ seja criptografada.

---

[69] https://www.gnupg.org/.

3. Lidando com dados não estruturados

Os *Paradise Papers* eram um grande arquivo secreto de 13,6 milhões de documentos. Um dos principais desafios na exploração destes residia no fato de que o vazamento veio de diversas fontes: Appleby, Asiaciti Trust e 19 outros registros nacionais de empresas.[70] Ao observar os arquivos mais de perto, logo se percebe seu conteúdo e caráter diversos, bem como uma grande presença de formatos "ilegíveis por máquina" como emails, PDFs e documentos no Word, que não podem ser processados diretamente por software de análise de dados estruturados. Estes documentos refletem as atividades internas de dois escritórios de advocacia de offshore investigados pelo ICIJ.



Com isso tudo em mãos, os engenheiros de software do ICIJ criaram um complexo e poderoso sistema que permitiria que os repórteres vasculhassem estes documentos. Utilizando a capacidade expansível da computação em nuvem, os documentos foram armazenados em um disco rígido criptografado que foi enviado a uma "linha de extração de dados", uma série de sistemas de software que extrai os textos dos documentos e os converte em dados que podem ser utilizados por nosso motor de busca.

A maior parte dos arquivos consistia em PDFs, imagens, emails, faturas e afins, o que dificultava a pesquisa. Ao usar tecnologias como Apache Tika (na extração de metadados e texto), Apache Solr (na criação de motores de busca) ou Tesseract (na conversão de imagens

---

70 https://www.icij.org/investigations/paradise-papers/paradise-papers-exposes-donald-trump-russia-links-and-piggy-banks-of-the-wealthiest-1-percent/, https://www.icij.org/investigations/paradise-papers/appleby-offshore-magic-circle-law-firm-record-of-compliance-failures-icij/, https://www.icij.org/investigations/paradise-papers/roll-roll-asiacitis-u-s-marketing-tour./.

em texto), a equipe criou um software de código aberto chamado Extract com o objetivo único de tornar estes documentos legíveis e pesquisáveis por máquinas.[71] Tal ferramenta foi especialmente útil na distribuição destes dados, agora acessíveis, em até trinta servidores.

O ICIJ também criou uma interface de usuário que permite aos jornalistas explorarem as informações refinadas extraídas destes "dados não estruturados", essa verdadeira confusão de diferentes tipos de documentos vindos de várias fontes. Novamente, optamos por reutilizar uma ferramenta de código livre chamada Blacklight, que oferece uma espécie de portal web amigável ao usuário, onde jornalistas podem pesquisar documentos e usar funções avançadas de pesquisa (como a comparação de *strings* aproximados) de forma a identificar pistas escondidas em meio ao vazamento.[72]

4. Uso de gráficos para descoberta de verdadeiras joias raras

A primeira edição do banco de dados Offshore Leaks foi publicada pelo ICIJ em 2013, com a utilização de bancos de dados de gráficos que permitiam aos leitores explorarem as ligações entre oficiais e mais de 100.000 entidades offshore. Até o momento são mais de 785.000 entidades offshore, incluindo aquelas que surgiram em outros vazamentos, como os *Panama* e *Paradise Papers*.

O ICIJ tentou, primeiramente, usar este tipo de banco de dados à época do *Swiss Leaks*, mas foi somente com os *Panama Papers* que eles passaram a desempenhar um papel fundamental na fase de pesquisa e reportagem. Explorar 11,5 milhões de registros financeiros e legais complexos em 2,6 terabytes de dados não foi fácil. Através de ferramentas de gráficos de rede como Neo4J e Linkurious, o ICIJ pôde fornecer aos seus parceiros uma maneira rápida de explorar conexões entre indivíduos e entidades offshore.

Nossas equipes de dados e pesquisa extraíram informações dos arquivos, estruturaram e tornaram os dados pesquisáveis com o Linkurious. De repente, nossos parceiros podiam pesquisar pelos nomes de pessoas de interesse público e descobrir, por exemplo, que o Primeiro-Ministro da Islândia, Sigmundur Gunnlaugsson, era acionista de uma empresa chamada Wintris. A visualização desta descoberta poderia ser salva e compartilhada com outros colegas que trabalhavam nesta investigação em outras partes do mundo.

Um deles poderia, então, retornar à plataforma de documentos Blacklight para fazer pesquisas mais avançadas e explorar registros relacionados à Wintris. Posteriormente, a Blacklight se tornou a Base de Conhecimento dos *Paradise Papers*. Descobertas relevantes

---

[71] https://github.com/ICIJ/extract/.

[72] https://github.com/projectblacklight/blacklight, https://en.wikipedia.org/wiki/Approximate_string_matching.

advindas da exploração de dados e documentos foram compartilhadas através do I-Hub Global, bem como descobertas feitas por meio de apuração em campo.

Bancos de dados de gráficos e tecnologias relacionadas movem o modelo de compartilhamento radical do ICIJ. “Parece mágica!”, disseram muitos de nossos parceiros. Navegar pelos dados não exigia nenhum conhecimento de programação. Fizemos treinamentos sobre o uso de nossas tecnologias para fins de pesquisa e segurança, e de repente mais de 380 jornalistas passaram a vasculhar milhões de documentos, usando bancos de dados de gráficos, fazendo pesquisas avançadas (inclusive em lote), compartilhando não apenas descobertas e resultados de reportagem, mas também dicas úteis de estratégias de pesquisa.

Para o projeto dos *Panama Papers*, bancos de dados de gráficos e outras tecnologias desenvolvidas sob demanda, como a Base de Conhecimento e o I-Hub Global, conectaram jornalistas de quase 80 países, trabalhando em 25 línguas diferentes, em uma redação virtual global.

O fato de que dados estruturados ligados a um grande número de documentos foram compartilhados com o público por meio do banco de dados do Offshore Leaks permitiu que novos jornalistas seguissem novas pistas e trabalhassem em novos projetos colaborativos como *Alma Mater* e *West Africa Leaks*. Também permitiu que cidadãos e instituições públicas usassem estes dados de maneira independente para suas próprias pesquisas e investigações. Até abril de 2019, governos pelo mundo recuperaram mais de 1,2 bilhões de dólares em multas e tributos atrasados por conta da investigação dos *Panama Papers*.

Desde a primeira publicação destes, em 2016, o grupo de jornalistas usando as tecnologias do ICIJ cresceu e mais de 500 repórteres puderam explorar os documentos financeiros vazados e escrever artigos de interesse público ligados a estes milhões de registros.

Emilia Díaz-Struck é Editora de Pesquisa e Coordenadora Latino-Americana do ICIJ. Cécile Gallego é jornalista de dados do ICIJ. Pierre Romera é Chefe de Tecnologia do ICIJ.

Textos enquanto dados: encontrando histórias em corpora

Barbara Maseda

Observando a produção em jornalismo de dados nos últimos anos, você pode ter percebido que materiais baseados em dados não estruturados (como texto) são muito menos comuns do que produtos baseados em dados estruturados.

Por exemplo, a análise de mais de 200 indicações ao *Data Journalism Awards*, entre 2012 e 2016, revelou que os trabalhos disputando a premiação dependiam, em grande parte, de dados geográficos e financeiros, seguidos por outras fontes de uso frequentes, como informações de sensores, sociodemográficas ou dados pessoais, metadados e pesquisas (Loosen, Reimer and De Silva-Schmidt, 2017).[73]

Ao passo que as redações têm de lidar com um volume crescente de postagens nas redes sociais, discursos, emails e extensos relatórios oficiais, abordagens computacionais ao processamento e análise destas fontes vêm se tornando cada vez mais relevantes. É possível que você se depare com artigos feitos desta forma, pense nas sínteses estatísticas dos tweets publicados pelo então presidente Donald Trump. Pode-se considerar, ainda, visualizações dos principais tópicos tratados em comunicações públicas ou durante debates de candidatos presidenciais nas eleições dos EUA.

Tratar texto como dado não é pouca coisa. Documentos tendem a ter os mais variados formatos, disposições e conteúdo, o que complica o trabalho de soluções tudo-em-um ou tentativas de replicar uma investigação com um conjunto diferente de arquivos. Limpeza, preparação e análise de dados são processos que podem variar consideravelmente entre uma série e outra de arquivos, e alguns dos passos envolvidos exigirão ação humana antes de qualquer declaração ou descoberta digna de publicação, de forma que algo realmente significativo seja revelado não só para pesquisadores, mas para públicos mais abrangentes.[74]

---

[73] Para mais informações sobre o Data Journalism Awards, ver o capítulo de Loosen publicado neste volume.

[74] O processo de limpeza e preparação de dados pode incluir um ou mais dos seguintes passos: separação do texto em unidades ou tokens (processo conhecido como "tokenização"); "agrupamento" de palavras que compartilham a mesma família ou raiz (stemização e lematização); eliminação de elementos supérfluos, tais como termos irrelevantes e pontuação; alteração de caixa alta/baixa no texto; focar em palavras e ignorar sua ordem (um modelo conhecido como "saco de palavras"), e transformação de texto em representação vetorial.

Neste capítulo, discorro sobre cinco maneiras pelas quais jornalistas podem usar análise de textos para contarem histórias, tudo ilustrado com referências a projetos exemplares de jornalismo de dados.

1. Extensão: o quanto falaram/escreveram

Contar frases ou palavras é a abordagem quantitativa mais simples quando se lida com documentos. Em termos computacionais, esta é uma tarefa feita há muito tempo e pode ser desempenhada pela grande maioria dos processadores de texto. Caso você seja um estudante ou jornalista que teve que enviar material considerando um limite máximo de palavras, não precisa de nenhum treinamento específico em dados para entender isso.

O problema com a contagem de palavras está na interpretação dos resultados em relação a uma linha de base significativa. Tais medidas não são de amplo conhecimento, como temperatura ou velocidade, logo, derivar algum significado a partir do fato de que um discurso contém 2.000 palavras não é um processo simples. Na prática, muitas vezes, a única opção é criar estas linhas de base ou referências para comparação nós mesmos, o que significa ter ainda mais trabalho.

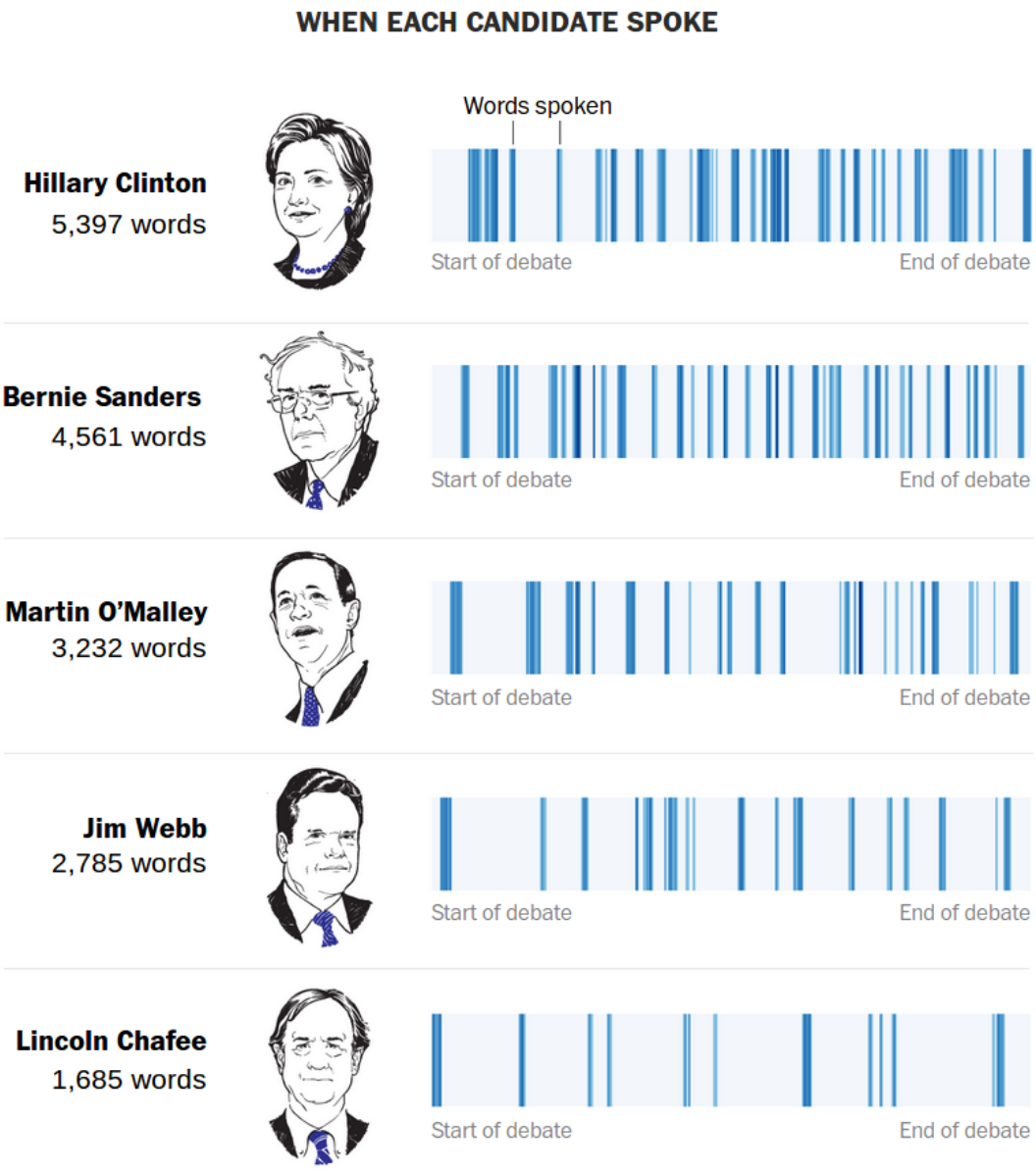
Em alguns casos, é possível encontrar contexto na história do evento ou do orador observado. Por exemplo, na cobertura do Discurso sobre o Estado da União de 2016, a *Vox* calculou a duração de todos os discursos em uma série histórica para determinar que "o presidente Obama foi um dos oradores com discurso mais longo de todos os tempos".[75]

Tratando-se de eventos com mais de um orador, é possível explorar quanto e quando cada indivíduo falou em relação ao número total de palavras proferidas. Para um exemplo disso, sugiro a matéria *Deconstructing the #demdebate: Clinton, Sanders control conversation*, parte da cobertura feita pelo *The Washington Post* em 2015 sobre o debate do Partido Democrata.[76]

---

[75] https://www.vox.com/2016/1/11/10736570/obama-wordy-state-of-the-union.

[76] https://www.washingtonpost.com/graphics/politics/2016-election/debates/oct-13-speakers/.



2. Menções: quem disse o que, quando e quantas vezes

Contar o número de vezes que um termo ou conceito foi usado em um discurso ou texto é outra tarefa simples que fornece uma visão estatística útil dos dados. Para tanto, é importante se certificar de contar os elementos mais apropriados.

A depender das perguntas que você quer fazer aos dados, pode considerar as repetições de cada palavra ou uma série de palavras com raiz em comum por meio de procedimentos de normalização como stemização ou lematização.[77] Outra possível

[77] Tf-idf é uma medida usada por algoritmos para entender o peso de uma palavra dentro de uma série. Peso de tf-idf(p, d) = FreqTermo(p, d) · log (N/FreqDoc(p)), em que (p, d) representa a frequência da palavra no documento (d), N é o número total de documentos, e DocFreq(p) é o número de documentos onde consta a palavra (Feldman e Sanger, 2007).

abordagem é focar nos termos mais relevantes de cada documento, empregando uma medida ponderada chamada "frequência de termo/inverso da frequência no documento", ou tf-idf, na sigla em inglês.[78] Abaixo, alguns exemplos.

Termos e tópicos frequentes

Para sua cobertura das eleições para prefeito de Londres em 2016, o *The Guardian* analisou o número de vezes em que os candidatos falaram sobre diversos temas de campanha — como crime, poluição, moradia e transporte — no Parlamento britânico ao longo dos seis anos antes da corrida eleitoral.[79] Os tópicos a serem analisados podem ser decididos com antecedência, como foi feito neste caso, e explorados através de uma série de palavras-chave relevantes (ou grupos de palavras-chave ligados a um tópico) em coleções de textos comparáveis ou análogos. Termos de pesquisa também podem ser análogos e não necessariamente os mesmos, tomemos por exemplo a análise do *FiveThirtyEight* sobre como os mesmos veículos cobriram diferentes furacões em 2017 (Harvey, Irma e Maria). Outra possibilidade é observar as palavras mais comumente utilizadas em um texto como estratégia para detecção de tópicos.

Discurso ao longo do tempo

Observar o discurso ao longo do tempo também pode ser uma maneira de apontar para tópicos que nunca foram mencionados antes ou não eram comentados há muito tempo. Este foi o método escolhido pelo *The Washington Post* na cobertura do Discurso do Estado da União de 2018, em um artigo que destacava qual presidente havia usado quais palavras primeiro na história do evento.[80] O fato de que os leitores podem saber logo de cara que Trump foi o primeiro presidente a falar do Walmart (em 2017) ou vagabundagem (em 2019) sem ter que ler centenas de páginas de discursos mostra o quão eficazes resumos e visualizações feitos com base em dados de textos podem ser.

Omissões

Um número baixo ou total ausência de menções pode ser notícia. Estas omissões podem ser analisadas ao logo do tempo, mas também podem ser baseadas na expectativa de que alguém ou alguma organização mencionará algo em determinado contexto. Durante a

---

[78] Stemização e lematização são operações usadas para reduzir palavras derivadas à raiz, para que ocorrências de termos como "repórter", "reportando" e "reportado" possam ser contados sob a raiz "report". Estes diferem na forma que o algoritmo determina a raiz da palavra. Diferentemente dos lematizadores, stemizadores removem os sufixos das palavras sem considerar em que parte da fala se encontram.

[79] https://www.theguardian.com/politics/datablog/2016/may/03/london-mayor-data-indicates-candidates-differing-focus-on-issues.

[80] https://www.washingtonpost.com/graphics/2018/politics/trump-state-of-the-union/.

campanha presidencial dos EUA em 2016, o *FiveThirty Eight* escreveu sobre o fato de que o candidato Donald Trump havia parado de falar sobre pesquisas em seu Twitter ao encontrarem um número relativamente baixo de menções a palavras-chaves ligadas ao tema em suas postagens.[81] Tais omissões podem ser detectadas através do monitoramento do mesmo orador no tempo, como neste caso, em que meses antes, o *FiveThirtyEight* descobriu que Trump tuitava bastante sobre pesquisas que o faziam parecer o vencedor da disputa.[82] Este é um bom exemplo de como reportagens baseadas em análise textual podem servir de contexto para outros artigos, de forma a abordar o problema mencionado anteriormente da contextualização de estatísticas textuais. A ausência de certo tópico pode ser medida com base na expectativa de que alguém ou alguma organização falará sobre ele em determinado contexto.

Pessoas, lugares, substantivos, verbos

Ferramentas de Processamento de Linguagem Natural (PLN) possibilitam a extração de nomes próprios, nomes de lugares, empresas e demais elementos (por meio de uma tarefa chamada Reconhecimento de Entidades Nomeadas, REN), bem como a identificação de substantivos, adjetivos e outros tipos de palavras (em um procedimento conhecido como etiquetagem morfossintática). Na matéria do *The Washington Post* mencionada acima, a visualização inclui filtros para destacar empresas, termos ou verbos religiosos.

3. Comparações

Determinar o quão semelhantes dois ou mais documentos são pode ser o início de diversos tipos de matérias. Podemos usar a correspondência de frases aproximadas (também conhecido como correspondência difusa) para expor casos de plágio, revelar as semelhanças entre figuras públicas ou descrever como foi alterada parte da legislação. Em 2012, o site *ProPublica* fez isto para monitorar mudanças em emails enviados a eleitores pelas campanhas, mostrando versões sucessivas das mesmas mensagens lado a lado, comparando o que havia sido apagado, inserido e inalterado.[83]

4. Classificação

Textos podem ser categorizados de acordo com certas funcionalidades predefinidas através do uso de algoritmos de aprendizagem de máquina. No geral, o processo consiste em

---

[81] https://fivethirtyeight.com/features/trump-isnt-tweeting-about-the-polls-anymore/.

[82] https://fivethirtyeight.com/features/shocker-trump-tweets-the-polls-that-make-him-look-most-like-a-winner/.

[83] https://projects.propublica.org/emails/.

treinar um modelo para que classifique entradas com base em determinada funcionalidade, então usá-lo para que categorize novos dados.

Em 2015, o *Los Angeles Times* analisou mais de 400.000 relatórios da polícia obtidos após pedidos de acesso à informação, revelando que cerca de 14.000 agressões graves foram classificadas erroneamente pela polícia de Los Angeles como delitos menores.[84] Ao invés de usar o MySQL para procurar por palavras-chave ("facada", "faca") que levariam a crimes violentos — como havia sido feito em uma investigação anterior em torno de um volume menor de dados —, os repórteres usaram classificadores de aprendizagem de máquina (SVM e MaxEnt) para reclassificar e revisar oito anos de dados em metade do tempo exigido pela primeira investigação, referente a apenas um ano.[85] Este exemplo mostra como a utilização de aprendizagem de máquina pode economizar tempo e ampliar nosso poder de investigação.

5. Sentimento

Muitos jornalistas reconhecem o valor de classificar frases ou documentos como positivos, negativos ou neutros (lembrando que outras gradações são possíveis) de acordo com a atitude do orador perante o tema em questão. Possíveis aplicações incluem a análise de uma postagem feita por um usuário no Twitter, um tópico ou hashtag para avaliação do sentimento em torno de uma questão, ou aplicar a mesma lógica em comentários de usuários em um site ou nota à imprensa e por aí vai. Tomemos por exemplo a comparação feita pelo *The Economist* a respeito do tom dos discursos de Hillary Clinton e Donald Trump durante as convenções de seus partidos.[86] Ao analisar a polaridade das palavras usadas por estes e outros candidatos antigos, puderam mostrar que Trump "fez o discurso mais negativo da memória recente", já Clinton apresentou "um dos discursos mais ponderados das últimas quatro décadas".

Como se tornar um jornalista minerador de textos

Softwares de mineração de textos amplamente disponíveis podem ser um bom começo para se familiarizar com o básico em procedimentos de análise textual e seus resultados (contagem de palavras, extração de entidades, conexões entre documentos etc.). Há plataformas projetadas para jornalistas, como DocumentCloud e Overview, que oferecem funcionalidades como estas.[87] A API Natural Language do Google Cloud é capaz de várias

---

[84] https://www.latimes.com/local/cityhall/la-me-crime-stats-20151015-story.html.

[85] https://www.latimes.com/local/la-me-crimestats-lapd-20140810-story.html.

[86] https://www.economist.com/graphic-detail/2016/07/29/how-clintons-and-trumps-convention-speeches-compared-to-those-of-their-predecessors.

[87] https://www.documentcloud.org/, https://www.overviewdocs.com/.

tarefas, incluindo análise de sentimento, análise de entidade, classificação de conteúdo e análise sintática.[88]

Existem, ainda, ferramentas gratuitas e de código aberto para aqueles que desejam aprender mais sobre mineração de textos. Tais instrumentos permitem análises personalizadas, com software em Python (NLTK, spaCy, gensim, textblob, scikit-learn) e R (tm, tidytext e tantos mais), uma alternativa mais conveniente para jornalistas familiarizados com estas linguagens. Um bom conhecimento de expressões comuns, bem como ferramentas e técnicas necessárias para coleta de texto (raspagem de dados web, consulta de API, solicitações de livre acesso à informação) e processamento destes (reconhecimento óptico de caracteres, ou OCR, em inglês, conversão de formato de arquivo etc.), também são essenciais.[89] E, claro, não faz mal ter uma noção da teoria e dos princípios por trás do trabalho que envolve dados textuais, incluindo coleta de informações, modelos e algoritmos relevantes, e visualização de dados textuais.[90]

Conclusões

A possibilidade de revelar novas percepções ao público a respeito de documentos ou multiplicar nossa capacidade de análise de textos que tomariam meses ou anos para serem lidos são bons motivos para levar a sério o desenvolvimento da análise textual como ferramenta útil ao jornalismo. Ainda restam muitos desafios, como questões de ambiguidade, já que computadores têm maiores dificuldades na "compreensão" do contexto da linguagem em relação a nós, humanos, e problemas específicos de linguagem que podem ser mais facilmente solucionados em inglês do que em alemão, por exemplo, ou que foram abordados mais em um idioma do que outro. Nosso trabalho como jornalistas pode contribuir para o avanço deste campo. Muitos projetos podem ser pensados, como maneiras de expandir o número de conjuntos de dados com anotações, identificar desafios ou, até mesmo, chegar a novas ideias de aplicações. A julgar pelo número de artigos produzidos com esta abordagem recentemente, a mineração de textos parece uma área empolgante e promissora em crescimento dentro do jornalismo de dados.

Barbara Maseda é fundadora e editora da Inventario, iniciativa de dados abertos para Cuba, autora do blog Text Data Stories.

---

[88] https://cloud.google.com/natural-language/.

[89] http://regex.bastardsbook.com/.

[90] Para mais informações, ver *Speech and Language Processing,* de Daniel Jurafsky e James H. Martin; e *The Text Mining Handbook,* de Ronen Feldman e James Sanger. Há, ainda, diversos cursos online gratuitos sobre estes e outros temas relacionados.

Referências

BARR, Caelainn. *London mayor: Commons speeches reveal candidates' differing issue focus*. The Guardian, 3 de maio de 2016. Disponível em: https://www.theguardian.com/politics/datablog/2016/may/03/london-mayor-data-indicates-candidates-differing-focus-on-issues.

CHANG, Alvin. *President Obama is among the wordiest State of the Union speakers ever*. Vox, 11 de janeiro de 2016. Disponível em: https://www.vox.com/2016/1/11/10736570/obama-wordy-state-of-the-union.

Daily Chart: 'How Clinton's and Trump's convention speeches compared to those of their predecessors'. The Economist, 29 de julho de 2016. Disponível em: https://www.economist.com/graphic-detail/2016/07/29/how-clintons-and-trumps-convention-speeches-compared-to-those-of-their-predecessors.

FISCHER-BAUM, Reuben; MELLNIK, Ted; SCHAUL, Kevin. *The words Trump used in his State of the Union that had never been used before*. The Washington Post, 30 de janeiro de 2018. Disponível em: https://www.washingtonpost.com/graphics/2018/politics/trump-state-of-the-union.

LARSON, Jeff; SHAW, Al. *Message Machine: Reverse Engineering the 2012 Campaign*. ProPublica, 17 de julho de 2012. Disponível em: https://projects.propublica.org/emails.

LOOSEN, Wiebke; REIMER, Julius; DE SILVA-SCHMIDT, Fenja. *Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the Data Journalism Awards 2013–2016*. Journalism, 12 de outubro de 2017. Disponível em: https://doi.org/10.1177/1464884917735691.

MEHTA, Dhrumil. *The Media Really Has Neglected Puerto Rico*. FiveThirtyEight, 28 de setembro de 2017. Disponível em: https://fivethirtyeight.com/features/the-media-really-has-neglected-puerto-rico.

MEHTA, Dhrumil; ENTEN, Harry. *Trump Isn't Tweeting About The Polls Anymore*. FiveThirtyEight, 19 de agosto de 2016. Disponível em: https://fivethirtyeight.com/features/trump-isnt-tweeting-about-the-polls-anymore.

MERRILL, Jeremy B. *Chamber of Secrets: Teaching a Machine What Congress Cares About*. ProPublica, 4 de outubro de 2017. Disponível em: https://www.propublica.org/nerds/teaching-a-machine-what-congress-cares-about.

POSTON, Ben; RUBIN, Joel; PESCE, Anthony. *LAPD underreported serious assaults, skewing crime stats for 8 years*. Los Angeles Times, 15 de outubro de 2015. Disponível em: http://www.latimes.com/local/cityhall/la-me-crime-stats-20151015-story.html.

POSTON, Ben; RUBIN, Joel. *LAPD misclassified nearly 1,200 violent crimes as minor offenses*. Los Angeles Times, 9 de agosto de 2014. Disponível em: http://www.latimes.com/local/la-me-crimestats-lapd-20140810-story.html.

STRAY, Jonathan, *Overview: a tool for exploring large document sets*. JonathanStray.com, 1º de dezembro de 2010. Disponível em: http://jonathanstray.com/overview-a-tool-for-exploring-large-document-sets.

Programação com dados dentro da redação

Basile Simon

Inevitavelmente, há um ponto em que dados e linguagem de programação se tornam parceiros. Talvez quando o Google Planilhas trava por conta do tamanho de um conjunto de dados; ou quando fórmulas do Excel ficam complicadas demais; ou quando, simplesmente, fica impossível entender dados ao logo de centenas de colunas. Programação pode simplificar o trabalho com dados, torná-lo mais elegante, menos repetitivo e mais reproduzível. Isso não significa que as planilhas serão aposentadas, mas, sim, que elas estarão entre as muitas opções disponíveis. Jornalistas de dados geralmente usam diversas técnicas assim que estas se fazem necessárias: coleta de dados com ferramentas em Python, cujos resultados vão em uma planilha, depois são copiados para uma limpeza no Refine, antes de retornarem à planilha.

Pessoas diferentes aprendem diferentes linguagens e técnicas de programação, assim como redações diferentes produzem seu trabalho em idiomas diferentes. Isso acontece, em partes, baseado no "stack" de escolha da organização, nome dado ao conjunto de tecnologias usadas internamente (a maior parte do trabalho com dados, visualizações e desenvolvimento dentro do *New York Times* se dá com R, JavaScript e React; no Reino Unido, a *ProPublica* usa Ruby em muitos de seus aplicativos web). Por mais que muitas vezes a escolha das ferramentas seja feita pelo indivíduo, as práticas e culturas de cada empresa de comunicação podem afetar bastante estas escolhas. Por exemplo, a *BBC* vem passando seu fluxo de trabalho de visualização de dados para a plataforma R; o *The Economist* passou seu famosíssimo Índice Big Mac baseado em cálculos de Excel para R e um painel React/d3.js.[91] Há muitas opções e nenhuma resposta correta definitiva. A boa nova para quem está começando é que muitos conceitos centrais se aplicam às mais diversas linguagens de programação. Assim que você aprende como armazenar pontos de dados em uma lista (semelhante ao que se faz com as linhas e colunas de uma planilha) e como fazer diversas operações em Python, fazer o mesmo com Javascript, R ou Ruby passa a ser uma questão de aprender como funciona a sintaxe de cada.

Para os fins deste capítulo, pensemos o jornalismo de dados como programação subdividida em três áreas principais: trabalho com dados — que inclui coleta, limpeza, estatística (coisas que podem ser feitas em planilhas); back-end — o mundo esotérico de bancos de dados, servidores e APIs; front-end — a maior parte do que acontece em um

[91] https://warwick.ac.uk/fac/cross_fac/cim/news/bbc-r-interview/, https://source.opennews.org/articles/how-we-made-new-big-mac-index-interactive/.

navegador, incluindo visualizações de dados interativas. Este capítulo aborda como estas diferentes áreas são moldadas por diversas restrições enfrentadas rotineiramente por jornalistas de dados que se veem às voltas com linguagem de programação dentro da redação, incluindo aí (1) o tempo para aprender, (2) prazos e (3) revisão de código de programação.

Tempo para aprender

Uma das maravilhosas características que unem a comunidade jornalística baseada em dados é a vontade de aprender. Independente se você é um jornalista disposto a entender como funcionam os processos, um estudante atrás de emprego na área ou um praticante do ofício já estabelecido, há muito a se aprender. Como a tecnologia evolui muito rápido, o que faz com que algumas ferramentas caiam em desuso enquanto outras são desenvolvidas por gente talentosa e generosa, há sempre algo de novo a ser feito ou aprendido. Frequentemente, temos diversas iterações e versões de ferramentas para realização de uma tarefa (bibliotecas para obtenção de dados da API do Twitter, por exemplo). Estas ferramentas, muitas vezes, tomam as anteriores por base e expandem seu escopo (extensões e plugins para a biblioteca D3 de visualização de dados). Sendo assim, a programação no contexto do jornalismo de dados é um processo de aprendizagem contínua que demanda tempo e energia, além de um investimento inicial de tempo para começar a aprender.

Uma questão ligada a este aprendizado é a redução inicial de eficiência e agilidade que acompanha a luta que é lidar com conceitos desconhecidos. *Bootcamps* de programação podem te ajudar a entender o que está fazendo em questão de semanas, mas eles também podem ser bem caros. Oficinas realizadas durante conferências são mais curtas, mais baratas e atendem usuários iniciantes e avançados. Ter tempo livre para aprender, como parte do seu trabalho, é essencial. Lá, você lidará com questões práticas, problemas reais, e com sorte terá colegas dispostos a ajudar. Há um jeitinho para encontrar soluções para seus problemas: fazer consultas diversas vezes até desenvolver um certo "faro" para descobrir o que os está causando.

O investimento em tempo e recursos pode compensar — a programação abre novas possibilidades e traz consigo muitas recompensas. Há, porém, um problema que acompanha todos os estágios desta experiência: é difícil estimar o tempo que uma tarefa levará. Um desafio e tanto, já que o trabalho de redação é composto por prazos.

Lidando com prazos

Entregar o que é preciso a tempo é essencial no jornalismo. Imprevisibilidade é uma característica da programação, assim como o ato de reportar. Independentemente da sua experiência, atrasos podem e irão ocorrer, invariavelmente.

Um desafio que afeta os iniciantes é a lentidão atrelada a aprender uma nova forma de trabalho. Quando decidir fazer algo novo, especialmente no começo do processo, certifique-se de ter tempo o suficiente para completar a tarefa em questão com o uso de uma ferramenta já conhecida, como uma planilha. Caso esteja começando a aprender e o tempo seja escasso, faz sentido lidar com uma ferramenta familiar e esperar até ter tempo para experimentações.

Ao lidar com projetos maiores, empresas de tecnologia empregam diversos métodos para dividir projetos em tarefas e subtarefas (até que estas atinjam uma proporção que permita calcular o tempo necessário para sua realização). Além disso, listam e priorizam tarefas por grau de importância.

Estes métodos devem servir de inspiração para o jornalista de dados. Por exemplo, em um projeto do *Sunday Times* sobre a proporção de crimes que a polícia britânica foi capaz de solucionar, priorizamos a exibição de números de acordo com a região do leitor. Depois de termos feito isso e com um tempo extra em mãos, passamos para a segunda parte: uma visualização comparativa da área do leitor e das demais regiões, bem como a média nacional. Graças à maneira como o trabalho foi feito, este projeto poderia ser publicado a qualquer momento. Este fluxo de trabalho iterativo ajuda a focar e gerenciar expectativas ao mesmo tempo.

Revisão de código

Na maior parte do tempo, as redações contam com mecanismos de padronização de seus produtos. Um repórter não apenas entrega seu texto e ele é publicado, não sem ser avaliado por editores e subeditores.

Desenvolvedores de software também têm seus próprios mecanismos para garantir a qualidade dos projetos colaborativos sem a introdução de bugs. Isso inclui "revisões de código", em que um programador apresenta seu trabalho e os demais testam e revisam. Além disso, empregam-se testes automatizados.

De acordo com a Pesquisa Global de Jornalismo de Dados de 2017, 40% das equipes de dados participantes contavam com três a cinco integrantes e 30% era composta por apenas um ou dois integrantes.[92] Estes números, tão baixos, criam um desafio à implementação de práticas de revisão de código internas. Desta forma, jornalistas de dados trabalham sozinhos na maior parte do tempo, por não terem colegas disponíveis, porque não há um mecanismo de revisão de pares implementado ou, simplesmente, não há ninguém por perto com as habilidades necessárias para revisar seu código.

---

[92] https://medium.com/@Bahareh/state-of-data-journalism-globally-cb2f4696ad3d.

Mecanismos internos de controle de qualidade podem ser um luxo que apenas algumas poucas equipes podem bancar, lembrando que não existe subeditor de programação! O preço de não ter tal controle são bugs que podem ser deixados de lado, desempenho abaixo do esperado ou pior: erros que passaram batido. Estas restrições talvez destaquem o porquê ser tão importante para tantos jornalistas buscarem opiniões e colaborações fora de seus veículos, buscando auxílio em comunidades online de programação, por exemplo.[93]

Basile Simon é Editor de Gráficos da Reuters Graphics e professor de Jornalismo Interativo da City University.

[93] Mais informações sobre transparência de código e práticas de revisão podem ser encontradas nos capítulos deste volume assinados por Leon e Mazotte.

Trabalhando de forma aberta no jornalismo de dados

Natália Mazotte

Muitos projetos de software e web relevantes, como Linux, Android, Wikipédia, Wordpress e TensorFlow, foram desenvolvidos de forma colaborativa, com base no fluxo livre de conhecimento. Stallman, notório hacker e fundador do Projeto GNU e da Free Software Foundation, relata que quando começou a trabalhar no MIT em 1971, o compartilhamento de código-fonte de softwares era tão comum quanto trocar receitas.

Por muitos anos, esta abordagem era impensável dentro do jornalismo. No começo de minha carreira como jornalista, trabalhei com comunidades de código aberto no Brasil e comecei a perceber que esta abordagem era o único caminho viável para a prática jornalística. Mas a transparência não tem sido colocada como prioridade ou valor essencial para jornalistas e organizações midiáticas. Na maior parte de sua história moderna, o jornalismo se viu às voltas com um paradigma de competição em torno de informações escassas.

Quando o acesso à informação é privilégio de poucos e descobertas relevantes estão disponíveis somente por testemunhas oculares ou informantes, as formas de garantir qualquer tipo de responsabilização ou prestação de contas ficam limitadas. Citar um documento ou mencionar a fonte de uma entrevista pode não exigir mecanismos elaborados de transparência. Em alguns casos, sigilo significa garantir a segurança da fonte, sendo até mesmo desejável. Mas quando a informação é abundante, não compartilhar *como se chegou até ela* pode privar o leitor dos meios para compreender uma história.

Como jornalistas cobrem e também dependem de dados e algoritmos, seria possível adotar um *ethos* semelhante ao das comunidades de código aberto? Quais as vantagens para jornalistas que adotam práticas digitais emergentes e valores associados a estas comunidades? Este capítulo discorre sobre alguns exemplos e benefícios de uma abordagem aberta no jornalismo de dados, bem como algumas formas de começar a trabalhar assim.

Exemplos e benefícios de uma abordagem aberta

O Washington Post forneceu uma visão única da epidemia de prescrição de opioides nos Estados Unidos ao vasculhar um banco de dados com informações sobre as vendas de milhões de analgésicos. O veículo também disponibilizou publicamente o conjunto de dados e metodologia utilizados, permitindo que jornalistas de outros 30 estados publicassem mais de 90 artigos sobre o impacto desta crise em suas comunidades.

Dois jornalistas computacionais analisaram o algoritmo de aumento de preços do Uber e descobriram que a empresa aparentemente oferece serviço melhor em áreas com maior população branca. O artigo foi publicado pelo Washington Post, a coleta de dados e o código da análise foram disponibilizados abertamente no GitHub, plataforma online que ajuda desenvolvedores a armazenarem e administrarem seus códigos. Isto significa que um leitor, ao encontrar um erro no banco de dados, poderia relatá-lo aos autores do artigo que, por sua vez, seriam capazes de ajustar o erro e corrigir o artigo.

A Gênero e Número, uma revista digital brasileira cofundada por mim, desenvolveu um projeto de classificação de mais 800 mil nomes de ruas para entender a falta de representação feminina em espaços públicos no país. Fizemos isso com a ajuda de um script em Python que fazia referências cruzadas entre nomes de ruas e um banco de dados de nomes do Instituto Brasileiro de Geografia e Estatística - IBGE. O mesmo script foi usado depois por outras iniciativas na classificação de conjuntos de dados sem informações de gênero – caso de listas de candidatos eleitorais e magistrados.

Trabalhar abertamente, com transparência e disponibilização de diversos conjuntos de dados, ferramentas, códigos, métodos e processos, pode ajudar jornalistas a:

1. **Melhorar a qualidade de seu trabalho.** Documentar processos pode estimular jornalistas a serem mais organizados, mais precisos e mais atentos a erros. Pode também reduzir a carga de trabalho na edição e revisão de artigos complexos, ao permitir que mais colaboradores relatem problemas.

2. **Ampliar alcance e impacto.** Uma história pode servir de base para outras matérias e ganhar diferentes perspectivas e atender a diferentes comunidades. Desta forma, cada projeto pode ganhar vida própria, sem as limitações do escopo inicial e demais restrições de seus criadores.

3. **Fomentar o letramento em dados entre jornalistas e públicos mais amplos.** Ter o passo-a-passo de seu trabalho significa que outros podem acompanhá-lo e aprender com isso, possibilitando o enriquecimento e diversificação de ecossistemas, práticas e comunidades de dados.

Na era da “pós-verdade” há também potencial para aumentar a confiança do público na imprensa, que atingiu um novo patamar negativo de acordo com o Barômetro de Confiança Edelman 2018. Trabalhar de maneira transparente pode ajudar a desacelerar ou mesmo reverter esta tendência. Para tanto, jornalistas podem falar mais abertamente sobre como chegaram às suas conclusões e também fornecer “guias” mais detalhados, agindo de forma honesta em relação aos seus vieses e incertezas, abrindo o diálogo com seus públicos.

Há uma ressalva, porém, já que práticas e culturas mais abertas dentro do jornalismo de dados ainda estão em desenvolvimento, num processo contínuo de exploração e experimentação. Ao passo em que avançamos nossa compreensão sobre benefícios em potencial, é preciso entender quando a transparência oferece valor, quando não é prioridade ou, até mesmo, quando é prejudicial. Por exemplo, há situações em que manter os dados e técnicas utilizadas em sigilo por um período é necessário para proteger a integridade da investigação, como ocorrido com os Panama Papers.

Maneiras de se trabalhar abertamente

Caso não haja impedimentos (a serem avaliados caso a caso), uma abordagem comum se dá pela seção de metodologia, ou o que chamamos de "nerd box". Há diversos formatos e tamanhos para apresentar a metodologia usada, a depender da complexidade do processo e do público-alvo.

Caso você queira atingir um público mais amplo, uma caixa de texto dentro do artigo ou mesmo uma nota de rodapé contendo uma explicação sucinta dos métodos empregados já pode ser o bastante. Algumas publicações optam pela publicação de artigos separados explicando como a história original foi escrita. Em ambos os casos, é importante evitar o uso de jargões, explicar como os dados foram obtidos e utilizados, garantir que os leitores entendam as ressalvas envolvidas e também explicar da forma mais clara e direta possível como se chegou àquela conclusão.

Muitos veículos reconhecidos por seu trabalho com jornalismo dados – caso de FiveThirtyEight, ProPublica, New York Times e Los Angeles Times – contam com repositórios em plataformas de compartilhamento de código como o GitHub. A equipe do Buzzfeed News tem até mesmo um índice com todos os seus dados, análises, bibliotecas, ferramentas e guias de código aberto. É compartilhada não só a metodologia por trás da reportagem, como também os scripts usados na extração, limpeza, análise e apresentação de dados. Esta prática torna o trabalho reprodutível (como discutido no capítulo de Sam Leon deste manual), possibilitando ainda aos leitores interessados explorarem os dados eles mesmos. Como os cientistas já fazem há séculos, estes jornalistas convidam seus pares a checarem seu trabalho e ver se conseguem chegar às mesmas conclusões ao seguirem os passos documentados.

Incorporar estes níveis de documentação e colaboração ao trabalho não é fácil para muitas redações. Com recursos cada vez mais escassos e equipes cada vez mais reduzidas, jornalistas dispostos a documentar suas investigações podem ser desencorajados por seus veículos. Isto nos leva às restrições com as quais estes profissionais têm de lidar: muitos veículos jornalísticos lutam pela sobrevivência, enquanto seu papel no mundo e modelos de

negócios mudam. Apesar destes desafios, adotar algumas das práticas das comunidades de código aberto e livre pode ser uma maneira de se destacar como inovadores, e ainda gerar confiança e criar relações com públicos em um mundo cada vez mais complexo e acelerado em suas mudanças.

Natália Mazotte é uma jornalista de dados brasileira, ex-diretora executiva do Open Knowledge Brasil e co-fundadora da Escola de Dados no Brasil e da Gênero e Número, iniciativa independente de jornalismo de dados focada em questões de gênero.

Como prestar contas dos métodos em jornalismo de dados: planilhas, códigos e interfaces de programação

Sam Leon

Com a ascensão do jornalismo de dados, a noção do que é uma fonte jornalística vem mudando. Formatos são muitos: conjuntos públicos de dados, e-mails vazados aos borbotões, documentos escaneados, imagens de satélite e dados de sensores. Junto destes, novos métodos para encontrar histórias em meio a estas fontes. Aprendizagem de máquina, análise de texto e algumas outras técnicas exploradas em outros pontos deste livro vêm sendo cada vez mais utilizadas a serviço da informação.

Mas os dados, apesar de sua aura de verdade objetiva estrita, podem ser distorcidos e representados erroneamente. Há diversas maneiras pelas quais jornalistas podem introduzir erros em sua interpretação de conjuntos de dados, publicando assim uma matéria que induz ao erro. Pode haver problemas na coleta de dados, o que nos impede de fazer inferências generalizando para uma população mais ampla. Isso pode ser resultado, por exemplo, de um viés de autosseleção na forma como uma amostra foi escolhida, problema comum na era de enquetes e pesquisas online. Erros também podem ser introduzidos no estágio de processamento de dados. O processamento, ou limpeza de dados, pode envolver geocodificação, correção de nomes com erros de ortografia, consolidação de categorias ou exclusão integral de certos dados caso sejam considerados pontos fora da curva. Um bom exemplo deste tipo erro é a geocodificação imprecisa de endereços de IP em uma pesquisa, que recebeu ampla cobertura, que supostamente mostrava a correlação entre persuasão política e consumo de pornografia. Isso sem contar o grosso do trabalho do jornalista de dados: a análise. Qualquer falácia estatística pode afetar esta parte do trabalho, como a confusão entre correlação e causalidade, ou a escolha de uma estatística inapropriada para resumir um determinado conjunto de dados.

Considerando as formas como a coleta, tratamento e análise de dados podem mudar uma narrativa, como o jornalista pode assegurar ao leitor que as fontes usadas são confiáveis e o trabalho que levou às suas conclusões também?

No caso do jornalista estar apenas relatando dados e descobertas de terceiros, não é preciso desviar dos padrões editoriais comumente adotados pelos grandes veículos. Uma referência à instituição que coletou e analisou os dados geralmente é o bastante. Por exemplo, uma tabela recente no Financial Times sobre expectativa de vida no Reino Unido é acompanhada da seguinte nota: "Fonte: cálculos do Club Vita com base em dados do Eurostat". Em tese, o leitor pode avaliar a credibilidade da instituição citada. Enquanto um jornalista responsável apenas cobrirá estudos que acredite ser confiáveis, uma instituição terceira é responsável pelos métodos que a levaram até as conclusões apresentadas. Em um contexto acadêmico, isso provavelmente incluiria a revisão de pares e, em caso de publicação científica, contaria ainda com algum grau de transparência metodológica.

No caso cada vez mais comum em que a organização jornalística produz sua própria pesquisa com base em dados, então eles são responsáveis pela confiabilidade dos resultados apresentados na reportagem. Os jornalistas têm lidado com o desafio desta prestação de contas metodológica de diversas maneiras. Uma abordagem recorrente é descrever, em linhas gerais, a metodologia utilizada para chegar às conclusões apresentadas

em um artigo. Estas descrições devem utilizar a linguagem mais distante do jargão técnico possível, buscando a compreensão de maior parte do público. Um bom exemplo desta abordagem é a forma *The Guardian* e *Global Witness* explicaram a contabilização de mortes de ativistas ambientais na sua série *Environmental Defenders*.

Mas existem limitações, como em qualquer prestação de contas. O maior dos problemas é que geralmente não se especificam os procedimentos exatos empregados na produção da análise ou preparação dos dados. Isso dificulta, e por vezes até mesmo impossibilita, reproduzir com exatidão os passos dados pelos repórteres para que chegassem àquelas conclusões. Ou seja, uma prestação de contas escrita muitas vezes não é reprodutível. No exemplo dado acima, em que a aquisição, processamento e análise de dados são relativamente simples, é possível que não haja valor agregado além de uma descrição geral por escrito. Porém, quando são empregadas técnicas mais complicadas, pode haver boas razões para se utilizar abordagens reprodutíveis.

Jornalismo de dados reprodutível

A reprodutibilidade é tida como um dos pilares do método científico moderno. Ela ajuda no processo de corroborar resultados, além de contribuir para identificar e solucionar conclusões problemáticas ou teorias questionáveis. Em princípio, os mesmos mecanismos podem ajudar a eliminar usos errôneos ou enganosos de dados no contexto jornalístico.

Observar um dos mais comentados erros metodológicos da história acadêmica recente pode ser bastante instrutivo. Em um artigo de 2010, Carmen Reihart e Kenneth Rogoff, de Harvard, demonstraram que o crescimento econômico real desacelera (uma queda de 0,1%) quando a dívida de um país sobe para valores acima de 90% de seu produto interno bruto (PIB). Estes números foram usados como munição por políticos que apoiam medidas de austeridade.

No final das contas, esta regressão estava ligada a um erro no Excel. Em vez de incluir a média de uma linha inteira de países, Reinhart e Rogoff cometeram um erro em sua fórmula, fazendo com que apenas 15 dos 20 países observados fossem incorporados. Com a correção, a "queda" de 0,1% revelava uma média de 2,2% de crescimento econômico. O erro só foi percebido quando o candidato a PhD Thomas Herndon e os professores Michael Ash e Robert Pollin observaram a planilha original em que Reinhard e Rogoff trabalharam. Este caso mostra a importância de não apenas ter a metodologia descrita em linguagem simples, como também acesso aos dados e tecnologia utilizados para análise. Mas o erro de Reinhart e Rogoff talvez aponte para outra direção também - softwares de planilhas em geral, incluindo o Microsoft Excel, podem não ser a melhor tecnologia para criar uma análise reprodutível.

O Excel oculta boa parte dos processos envolvendo dados por princípio. Fórmulas, responsáveis pelo trabalho analítico em uma planilha, são visíveis apenas quando se clica em uma célula. Ou seja, revisar os passos dados até determinada conclusão é uma tarefa mais complicada. Por mais que nunca tenhamos como saber, imagina-se que caso o trabalho de Reinhart e Rogoff tivesse sido feito em uma linguagem cujos passos devem ser declarados explicitamente (como em uma linguagem de programação), o erro poderia ter sido notado antes da publicação.

Fluxos de trabalho com base no Excel, na maioria das vezes, encorajam a remoção de passos que levaram a uma conclusão. Valores, e não fórmulas, são frequentemente copiados em outras planilhas ou colunas, sendo o botão de "desfazer" a única rota para entender como um determinado número foi realmente gerado. O histórico da função "desfazer", em linhas gerais, é apagado quando se fecha um programa, logo, não é uma boa ideia armazenar informação metodológica importante ali.

A ascensão do ambiente de programação letrado: Jupyter Notebooks na redação

Uma abordagem emergente para transparência metodológica é usar os chamados ambientes "letrados" de programação. Organizações como Buzzfeed, The New York Times e Corrective utilizam estes ambientes para a criação de documentos legíveis para humanos que também podem ser executados por máquinas, de forma a reproduzir os passos exatos de determinada análise.

Articulada inicialmente por Donald Knuth nos anos 90, a programação letrada é uma abordagem à criação de códigos de computador em que o autor intercala código com linguagem humana comum, explicando os passos dados. Os dois principais ambientes de programação letrada em uso atualmente são Jupyter Notebooks e R Markdown. Ambos geram documentos legíveis por humanos que misturam linguagem comum, visualizações e código em um único documento passível de ser exportado em HTML e publicado na web. Desta forma, os dados originais podem ser linkados explicitamente e demais dependências técnicas, tais como bibliotecas de terceiros, serão identificadas com clareza.

Há uma ênfase em um formato explicativo legível para pessoas e o código é ordenado de forma a refletir a lógica humana também. Documentos criados com este paradigma podem ser interpretados como uma série de passos em um argumento ou uma série de respostas para um conjunto de perguntas de pesquisa.

> O praticante de programação letrada pode ser encarado como um ensaísta, cujas maiores preocupações são clareza e excelência de estilo. Tal autor, com um tesouro em mãos, escolhe cuidadosamente o nome de variáveis, explicando ainda o que cada uma delas significa. Ele ou ela buscam criar um programa compreensível, pois seus conceitos foram introduzidos em uma ordem mais adequada ao entendimento humano, combinando métodos formais e informais que reforçam uns aos outros.

Um bom exemplo do formato está no Jupyter Notebook do Buzzfeed News, que detalha como o veículo analisou tendências nos incêndios florestais da Califórnia. A interface conta com todo o código e dados exigidos para a reprodução da análise, mas o principal do documento é a narrativa ou diálogo com os dados que serviram como fonte. As explicações surgem abaixo de títulos que seguem uma linha lógica de investigação. Já visualizações e tabelas são usados para destacar temas-chave.

Um aspecto da abordagem "letrada" de programação é que os documentos gerados (arquivos do Jupyter Notebook ou R Markdown) podem funcionar para dar segurança até mesmo a leitores que não conseguem compreender o código, mas entendem que os passos dados para produção daquelas conclusões fazem sentido. A ideia é similar à noção de

"testemunho virtual" de Steven Shapin e Simon Schaffer enquanto método para estabelecer questões de fato nos primórdios da ciência moderna. Usando o programa experimental de Robert Boyle como exemplo, Shapin e Schaffer definiram o papel do "testemunho virtual":

> A tecnologia do testemunho virtual envolve a produção, na mente do leitor, da imagem de uma cena experimental que neutraliza a necessidade de uma testemunha direta ou replicação. Por meio do testemunho virtual, a multiplicação de testemunhas poderia ser, em princípio, ilimitada. Assim sendo, tratava-se da mais poderosa tecnologia para a constituição de questões de fato. A validação dos experimentos, e o reconhecimento de seus resultados como questões de fato, necessariamente implicava sua compreensão no laboratório e olho da mente. O que se exigia era uma tecnologia de confiança e garantias de que algo havia sido feito e feito da forma alegada. (Shaping & Schaffer, 'Leviathan and The Air-Pump',1985)

Os documentos gerados em ambientes de programação letrada como o Jupyter Notebook, quando publicados junto de artigos, podem ter efeito similar ao permitir que o leitor não-programador visualize os passos dados na produção das descobertas de determinada história. Por mais que leitor sem conhecimento de programação talvez não seja capaz de compreender ou rodar o código, os comentários e explicações ao longo do documento podem lhe assegurar as medidas apropriadas foram tomadas para mitigar possíveis erros.

Peguemos como exemplo um artigo recente no Buzzfeed News sobre inspeções em casas com crianças no Reino Unido. O Jupyter Notebook conta com etapas específicas para verificar a filtragem correta dos dados (Figura 1), impedindo que erros simples, mas sérios, como aquele cometido por Reinhart e Rogoff se repitam. Por mais que o conteúdo exato do código não seja entendido pelo leitor sem conhecimento técnico, a presença destes testes e checagens, acompanhados de explicações em linguagem simples, servem para mostrar que o trabalho envolvido nas descobertas do jornalista foi feito adequadamente.

```
In [11]: # Make sure that we've identified all private-sector owners
         assert (
             as_at_data_filtered
             .loc[lambda df: df["Sector"] == "Private"]
             ["Owner"].isnull()
             .sum()
         ) == 0
```

Figure 5: Célula do arquivo do Jupyter Notebook com uma explicação para leitores ou comentário explicando que seu propósito é verificar se a filtragem dos dados brutos foi feita corretamente

Mais do que apenas reprodutibilidade

O uso de ambientes de programação letrada ajuda na reprodutibilidade de artigos de dados.

Mas não para por aí - a publicação do código pode fomentar a colaboração entre organizações. Em 2016, a Global Witness publicou um raspador de dados web que extraía detalhes de empresas e acionistas do registro de empresas de Papua-Nova Guiné. A parte inicial da pesquisa buscava identificar os principais beneficiários do comércio madeireiro tropical, sujeito à corrupção e com impacto devastador nas comunidades locais. Por mais que a Global Witness não tivesse planos imediatos de reutilizar o raspador que havia desenvolvido, seu código básico foi disponibilizado no GitHub - o popular site de compartilhamento de códigos de programação.

Pouco tempo depois, um grupo de defesa de interesses comunitários chamado ACT NOW! baixou o código do raspador, aprimorando-o e incorporando-o ao seu projeto iPNG, que permitia ao público fazer verificações cruzadas de nomes de acionistas e diretores de empresas com outras fontes de interesse público. O coletor agora integra a infraestrutura central de dados do site, coletando informações do serviço de registro de empresas de Papua-Nova Guiné duas vezes ao ano.

Criar códigos dentro de um ambiente de programação letrada pode ajudar a simplificar certos processos internos em que outras pessoas dentro de uma organização necessitem compreender e verificar uma análise antes de sua publicação. Na Global Witness, o Jupyter Notebook tem sido usado para dinamizar o processo de revisão legal. À medida que os notebooks definem as etapas seguidas para obter uma determinada conclusão em uma ordem lógica, os advogados podem fazer uma avaliação mais precisa dos riscos legais associados a uma determinada alegação.

No contexto do jornalismo investigativo, esta prática é particularmente importante quando suposições são feitas em torno da identidade de indivíduos específicos citados em um conjunto de dados. Como parte de um recente trabalho sobre o estado da transparência corporativa no Reino Unido, queríamos estabelecer quem eram os indivíduos por trás de um enorme número de empresas. Trata-se de indicativo (não prova) de que estes seriam “nomeados”, o que em certos contextos tais como quando um indivíduo é listado como Pessoa de Controle Significativo (PSC, no inglês) - é ilegal. Ao publicar a lista dos nomes destes controladores responsáveis por um maior número de empresas, a equipe legal gostaria de saber como nós sabíamos que uma pessoa em específico, digamos um tal John Barry Smith, era a mesma pessoa que John B. Smith. O Jupyter Notebook foi capaz de capturar claramente como havíamos realizado este tipo de deduplicação ao apresentar uma tabela na etapa relevante que definia as funcionalidades usadas para confirmar a identidade dos indivíduos (ver abaixo). Estes mesmos processos foram utilizados na Global Witness para fins de checagem de fatos.

Potential nominee PSCs

Which PSCs currently control the most number of companies?

In [44]:
```
temp_df = active_psc_records.groupby(['name_elements.forename','name_elements.surname','month_year
_birth','address.postal_code'])[['company_number']]\
    .agg(unique_company_count).sort_values(by='company_number',ascending=False)
temp_df_for_viz = temp_df.copy()
temp_df_for_viz = temp_df_for_viz.reset_index()
temp_df_for_viz['Name'] = temp_df_for_viz['name_elements.forename'] + ' ' + temp_df_for_viz['name_
elements.surname']
temp_df_for_viz[['Name','company_number']].head(10).to_csv('data/viz/top_10_pscs.csv',index=False)
temp_df.head(10)
```

Out[44]:

				company_number
name_elements.forename	name_elements.surname	month_year_birth	address.postal_code	
Michael	Gleissner	1969-04-01	CT20 2RD	1193
Peter	Valaitis	1950-11-01	BS9 3BY	997
Waris	Khan	1979-02-01	W1G 9QR	639

Figura 2: Seção do arquivo do Global Witness no Jupyter Notebook em que é apresentada tabela com indivíduos e demais dados relacionados ao fato de terem o mesmo primeiro nome, sobrenome, mês/ano de nascimento e CEP.

Dentro da Global Witness, o Jupyter Notebook também se mostrou especialmente útil quando há a necessidade de monitorar um conjunto de dados específicos ao longo do tempo. Por exemplo, em 2018, a Global Witness quis estabelecer a mudança no risco de corrupção no mercado imobiliário londrino ao longo de dois anos. Para tanto, foi obtido um registro das terras de propriedade de empresas estrangeiras, então reutilizaram e publicaram um arquivo que havíamos desenvolvido para o mesmo fim dois anos antes (Figura 2). Obtivemos resultados comparáveis, com o mínimo de ruído. Trabalhar com este tipo de arquivo oferece ainda outra vantagem neste contexto, pois permitia à Global Witness mostrar sua metodologia de trabalho sem republicar os dados de origem, que na época da análise, sofriam certas restrições de licenciamento. Isso tudo é difícil de se fazer com um fluxo de trabalho em planilhas. Claro, a forma mais eficaz de prestar contas do seu método sempre será publicar os dados brutos utilizados. Porém, jornalistas geralmente usam dados que não podem ser republicados por questões de direitos autorais, privacidade ou proteção a fontes.

Ainda que estes ambientes de programação letrada possam claramente aprimorar a prestação de contas e reprodutibilidade do trabalho de um jornalista de dados, além de outros benefícios, há algumas limitações importantes.

Uma destas é que para reproduzir (em vez de apenas acompanhar ou "testemunhar virtualmente") uma abordagem com documentos gerados no Jupyter Notebook ou R Markdown, você precisa saber como escrever, ou ao menos rodar, códigos. O estado ainda relativamente imaturo do jornalismo de dados indica que há ainda um grupo pequeno de jornalistas e suas audiências capazes de programar. Ou seja, é improvável que os repositórios do GitHub de jornais recebam o mesmo escrutínio que os códigos revisados por pares referenciados em periódicos acadêmicos onde partes maiores da comunidade podem analisar e questionar o código de fato. Consequentemente, o jornalismo de dados pode estar mais suscetível a erros ocultos no código quando comparado a pesquisas realizadas por um público

de maior conhecimento técnico. Como apontado por Jeff Harris, não deve demorar muito para que vejamos correções de programação publicadas por veículos jornalísticos, assim como é comum com erros factuais. Cabe notar que, neste contexto, ferramentas como o Workbench (citado também no capítulo deste livro assinado por Jonathan Stray) estão começando a ser desenvolvidas para jornalistas, prometendo algumas das funcionalidades presentes em ambientes de programação letrada sem a necessidade de criar ou compreender códigos.

A esta altura, cabe também pensar se os novos mecanismos de prestação de contas em jornalismo não são apenas novos meios pelos quais um "público" pré-existente pode analisar minuciosamente métodos, mas também maneiras de desempenhar um papel na formação de novos tipos de "públicos". Este é um argumento apresentado por Andrew Barry em seu ensaio, *Transparency as a political device (Transparência enquanto dispositivo político, em tradução livre):*

> A transparência implica não só a publicação de informações específicas, mas também a formação de uma sociedade capaz de reconhecer e avaliar o valor da - e se necessário modificar - informação que é tornada pública. A operação da transparência é endereçada às testemunhas locais, ainda que se espere que estas testemunhas sejam reunidas apropriadamente, e suas presenças validadas. Há, então, uma relação circular entre a constituição de assembleias políticas e prestação de contas da economia do petróleo – uma traz a outra à existência. À transparência não cabe apenas tornar informação pública, mas formar um público interessado em ser informado. (Barry, 'Transparency as a Political Device', 2010)

Os métodos de prestação de contas para trabalho jornalístico envolvendo dados discutidos acima podem desempenhar um papel na emergência de públicos com maior conhecimento técnico que desejem analisar as minúcias destas análises e também conferir maior responsabilização aos jornalistas envolvidos de formas que talvez não fossem possíveis antes do advento e uso de tecnologias como ambientes de programação letrada dentro de um contexto jornalístico.

Esta ideia vai ao encontro do que a Global Witness faz em termos de letramento em dados, de forma a aprimorar a prestação do setor extrativista. A legislação de marcos territoriais da União Europeia força empresas extrativistas a tornarem públicos registros de pagamentos de projetos para governos nos setores de óleo, gás e mineração, altamente suscetíveis à corrupção, o que abriu o caminho para uma observação muito mais cuidadosa de onde este dinheiro acaba se acumulando. Porém, Global Witness e demais grupos associados à coalizão Publish What You Pay (*Divulgue o que Paga, em tradução livre*) já perceberam há tempos que não há um "público" pré-existente que poderia desempenhar este papel imediatamente. Sendo assim, junto de outras organizações, Global Witness desenvolveu ferramentas e programas de treinamento com a intenção de reunir jornalistas e grupos da sociedade civil em países ricos em recursos, capazes de oferecer suporte ao desenvolvimento de habilidades de forma a facilitar o uso deste dados para responsabilização destas empresas. Um elemento deste esforço tem sido o desenvolvimento e publicação de metodologias específicas para sinalizar relatórios de pagamentos suspeitos, passíveis de corrupção.

Atualmente, os ambientes de programação letrada são meios promissores pelos quais os jornalistas de dados vêm tornando suas metodologias mais transparentes e confiáveis. Por mais que dados estejam sempre abertos a múltiplas interpretações, tecnologias que tornam as suposições de um repórter explícitas e seus métodos reprodutíveis possuem valor. Elas ajudam na colaboração e possibilitam maior escrutínio em torno de uma disciplina cada vez mais técnica, por diversos públicos. Tendo em mente a atual crise de confiança do jornalismo, uma adoção maior de abordagens reprodutíveis pode ser um caminho para a manutenção da credibilidade de equipes de dados.

Sam Leon é o Chefe de Investigações de dados da organização anticorrupção Global Witness.

Obras citadas

Ben Leather e Billy Kyte, 'Defenders: Methodology', *Global Witness*, 13 de julho de 2017. Disponível em: https://www.globalwitness.org/en/campaigns/environmental-activists/defendersmethodology/

Donald Knuth, 'Literate Programming', Computer Science Department, Stanford University, Stanford, CA 94305, USA, 1984. Disponível em: http://www.literateprogramming.com/knuthweb.pdf

Andrew Barry, 'Transparency as a political device Em: Débordements: Mélanges offerts à Michel Callon', Paris: Presses des Mines, 2010.

Carmen M. Reinhart e Kenneth S. Rogoff, 'Growth in a Time of Debt', The *National Bureau of Economic Research*, Dezembro de 2011. Disponível em: https://www.nber.org/papers/w15639

Donald E. Knuth, 'Literate Programming', Stanford, California: *Center for the Study of Language and Information*, 1992. Disponível em: https://www-cs-faculty.stanford.edu/~knuth/lp.html

Steven Shapin and Simon Schaffer, 'Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life', Princeton University Press, 1985.

Algoritmos a serviço do jornalismo

Jonathan Stray

O segredinho sujo do jornalismo computacional é que a parte "algorítmica" de um artigo não é aquilo que exige mais tempo e esforço dentro de um projeto.

Não me leve a mal, algoritmos sofisticados podem ser extremamente úteis para reportagem, especialmente se tratando de trabalho investigativo. Treinar computadores para que encontrem padrões, o processo conhecido como aprendizagem de máquina, tem sido utilizado para encontrar documentos importantes em meio a grandes volumes de dados. Já o processamento de linguagem natural, que consiste em treinar computadores para que compreendam linguagem, pode extrair os nomes de pessoas e empresas de documentos, dando a repórteres um atalho para que possam entender quem está envolvido em determinada história. Diversas modalidades de análise estatística vêm sendo empregadas por jornalistas de forma a detectar vieses ou irregularidades.

Mas botar um algoritmo para rodar é a parte fácil disso tudo. Obter dados, limpá-los e seguir pistas baseadas em algoritmos que é dureza.

De forma a ilustrar esta afirmação, cito aqui um caso de sucesso de aprendizagem de máquina no contexto do jornalismo investigativo: o artigo sobre abuso sexual cometido por médicos intitulado *License to Betray* ("Licença para Trair", em tradução livre), publicado no *The Atlanta-Journal Constitution*. Mais de 100.000 registros disciplinares de médicos de cada estado norte-americano foram analisados pelos repórteres do veículo, e nestes foram encontrados 2.400 casos de médicos que haviam abusado sexualmente de suas pacientes, mas que mesmo assim continuavam atuando. Ao invés de lerem cada relato, o que os jornalistas fizeram foi reduzir essa pilha de documentos ao aplicarem aprendizagem de máquina para encontrar os que estivessem relacionados de alguma forma a abuso sexual. Com isso, reduziram o volume em mais de dez vezes, chegando a 6.000 documentos que foram, então, lidos e revisados manualmente.

Não teria como fazer deste projeto um artigo nacional se não fosse pela aprendizagem de máquina, de acordo com o repórter Jeff Ernsthausen: "Talvez houvesse uma chance de criarmos uma matéria de alcance regional".

Uma vitória para algoritmos aplicados ao jornalismo como os conhecemos até então, o tipo de técnica que poderia ser usada de forma mais ampla. Dito isso, a aprendizagem de máquina não é a parte difícil. O método usado por Ernsthausen envolvia "regressão

logística", uma abordagem estatística padrão para a classificação de documentos com base nas palavras que estes contêm. É um processo simples de ser implementado, com algumas dezenas de linhas de código em Python; há diversos bons guias disponíveis online.

Gasta-se a maior parte do tempo organizando o material e explorando seus resultados, para a maioria das histórias. Dados devem ser coletados, limpados, formatados, carregados, checados e corrigidos, em constante preparação. Já os resultados de análise algorítmica são, muitas vezes, pistas ou dicas, que só se transformam em uma história após muito trabalho de reportagem típico, muitas vezes envolvendo equipes que precisam mais de ferramentas colaborativas do que de análise. Essa é a porção nada glamourosa do trabalho com dados, então não é algo que ensinamos muito bem ou discutamos tanto assim. Ainda assim, é toda essa preparação e seguimento que demanda tempo e esforço em uma narrativa baseada em dados.

Em *License to Betray*, a obtenção dos dados já foi um desafio gigantesco. Não existe um banco de dados nacional com relatórios disciplinares de médicos, apenas uma série de bancos de dados estaduais. Muitos destes não contam com um campo específico indicando porque este ou aquele médico sofreu ação disciplinar. Quando há, normalmente não há uma indicação clara de abuso sexual. Em um primeiro momento, a equipe tentou conseguir estes documentos com base em pedidos de liberdade de informação. Tal processo se provou proibitivamente dispendioso, com alguns estados pedindo milhares de dólares em troca do fornecimento destes dados. Logo, a equipe optou por coletar documentos dos sites de conselhos estaduais de medicina. Estes documentos tiveram que passar por programas de OCR (convertidos em texto) e ser carregados em uma aplicação baseada na web para revisões e marcações colaborativas.

Depois disso, os jornalistas tiveram que marcar, manualmente, centenas de documentos para geração de dados de treinamento. Após a categorização dos 100.000 documentos restantes ter sido feita por aprendizagem de máquina, foram necessários muitos meses de leitura manual de 6.000 documentos que se supunha tratarem de abuso sexual, além de milhares de outros arquivos contendo palavras-chave escolhidas a dedo. E é claro que havia o restante do processo de reportagem, o que incluía a investigação de centenas de casos específicos de forma a contar uma história com maiores detalhes. Este processo também se baseava em outras fontes, como reportagens antigas e entrevistas com os envolvidos.

O uso de um algoritmo, de aprendizagem de máquina, foi parte essencial, crítica, para esta investigação. Mas, no final das contas, representava apenas um pequeno volume de todo tempo e esforço envolvidos. Pesquisas feitas por cientistas de dados mostram, consistentemente, que a maior parte do seu trabalho, cerca de 80% do tempo, se dá "lutando" com os dados e limpando-os, e com o jornalismo não é diferente.

Muitas vezes, os algoritmos são encarados como uma espécie de ingrediente mágico. Podem parecer complexos ou nada transparentes, mas são poderosíssimos, sem sombra de dúvidas. É muito mais divertido falar dessa mágica do que de todo o trabalho mundano da preparação de dados ou de dar seguimento a uma longa lista de pistas. Tecnólogos gostam de criar expectativa em torno de sua tecnologia, não em torno do trabalho essencial que se dá em torno dela, e isso afeta como novas e sofisticadas tecnologias adentram o jornalismo. Devemos, sim, ensinar e nos aproveitar de avanços tecnológicos, isso é óbvio, mas nossa responsabilidade maior é fazer jornalismo, o que significa nos atravancar com o resto do processo que envolve dados.

No geral, subestimamos as ferramentas usadas na preparação destes dados. Temos o Open Refine, um herói de longa data quando se trata de tarefas de limpeza. Há, ainda, o Dedupe.io, que aplica aprendizagem de máquina ao problema de fusão de nomes quase duplicados em um banco de dados. Métodos clássicos de trabalho com textos, envolvendo expressões de uso comum, por exemplo, devem fazer parte da educação de todo jornalista de dados. Neste sentido, meu projeto atual, o Workbench, é voltado à demorada e quase invisível tarefa de preparar dados para reportagem, aquilo tudo que acontece antes do algoritmo. Logo, seu objetivo é fazer deste processo mais colaborativo, para que repórteres possam trabalhar juntos em grandes projetos de dados, aprendendo mais sobre o trabalho do outro, o que inclui máquinas.

Algoritmos são importantes para o jornalismo, mas para que funcionem, precisamos falar sobre todas as outras engrenagens do jornalismo baseado em dados. Precisamos criar formas de possibilitar todo esse fluxo de trabalho, não só a parte tecnológica.

Jonathan Stray é jornalista computacional na Universidade de Columbia, onde dá aula no mestrado-sanduíche em ciências da computação e jornalismo, além de liderar o desenvolvimento do Workbench, uma ferramenta integrada de jornalismo de dados.

Referências

DIAKOPOULOS, N. *Automating the News*. Harvard University Press, 2019.

ERNSTHAUSEN, Jeff. *Doctors and Sex Abuse*. Apresentação durante a NICAR 2017. Disponível em: https://docs.google.com/presentation/d/1keGeDk_wpBPQgUOOhbRarPPFbyCculTObGLeAhOMmEM/edit#slide=id.p. Acesso em: 5 de agosto de 2018.

TEEGARDIN, C. et al. *License to Betray*. Atlanta-Journal Constitution, 5 de julho de 2016.

Jornalismo com máquinas? Do pensamento computacional à cognição distribuída

Eddy Borges-Rey

Imagine que você é um jornalista em um futuro não tão distante. Você está trabalhando em cima de um artigo, e para obter o insight necessário, pede ao seu agente conversacional (a quem você carinhosamente chama de Twiki) para juntar quinze bancos de dados anonimizados. Considerando a magnitude e complexidade dos conjuntos de dados fundidos, softwares de visualização são rudimentares demais para isolar as anomalias de seu interesse. Sendo assim, ao usar o implante em seu cérebro, você se conecta ao sistema e navega com facilidade pela abstração destes conjuntos. Por mais que, individualmente, cada conjunto editado funcione para proteger a identidade e os dados pessoais das pessoas listadas, ao combinar tudo você consegue inferir a identidade de alguns indivíduos de alto nível e contextualizar seus dados pessoais. Ao perceber as possíveis implicações legais de revelar os nomes e dados associados, você pede a Twiki para rodar uma rede neural de forma a determinar se a divulgação destas informações levaria a problemas éticos ou legais. A rede executa uma série "n+" de simulações de jornalistas virtuais tomando decisões baseadas em códigos de ética e marcos regulatórios. Com esses processos rodando ao fundo, você é capaz de isolar alguns pontos fora da curva e identificar tendências interessantes. Já que a ideia é certificar-se de que as anomalias têm algo a adicionar à história, e não são apenas erros, você pede para que Twiki verifique os arquivos para ver se os tais pontos coincidem com outros eventos históricos. Pede, também, para que Twiki rode um modelo preditivo para calcular a probabilidade de que as tendências identificadas persistam em um futuro próximo, levando a implicações preocupantes.

Esta breve e fictícia introdução baseia-se em uma fascinante conversa que tive com a ex-jornalista de dados do *Times* Nicola Hughes, há alguns anos. Por mais que a cena descrita pudesse muito bem ter saído de *Minority Report*, escrito por Philip K. Dick, na verdade ela se refere a ferramentas e técnicas já amplamente disponíveis e utilizadas, ou que estão em rápido desenvolvimento. Mais importante que isso, se refere também a um processo de trabalho e mentalidade jornalística despontando nas redações, num mundo em que jornalistas cada vez mais se envolvem com dados, e a computação se torna indispensável.

Estas mudanças recentes refletem como, historicamente, toda vez que uma inovação tecnológica relevante é introduzida ao fluxo de trabalho da produção de notícias, o processo de reportagem passa por uma disrupção e consequente transformação, com o raciocínio e o ideal dos profissionais envolvidos invariavelmente modificados.

Hoje, saímos da era da Big Data rumo à da Inteligência Artificial (IA) e automação, com a prevalência de princípios e práticas computacionais e ciência de dados cada vez mais penetrando no jornalismo. Como dito por Bell:

> Toda empresa em todos os campos de atuação, e toda organização, privada ou pública, terá que repensar a forma como lidam com IA da mesma forma que há 20 anos tiveram que pensar como lidariam com as tecnologias web (Bell, 2017).

Dentro deste contexto, este capítulo faz uma reflexão a respeito das formas como jornalistas que trabalham com dados e processos automatizados internalizam diversos princípios computacionais que, por um lado, melhoram suas capacidades jornalísticas e, por outro, começam a modificar a pedra basilar de suas abordagens e ideais jornalísticos.

Desta forma, este capítulo explora uma série de conceitos teóricos que podem servir de base para antever a cognição jornalística em um ambiente onde a computação se faz cada vez mais presente. Trabalho com a ideia de cognição estendida para estimular discussões mais aprofundadas sobre como a cognição jornalística hoje depende (e se distribui ao longo) das máquinas usadas no ato de dar notícias. Ao longo desta discussão espero encorajar trabalhos futuros na investigação do papel da computação em situações jornalísticas, incluindo trabalho empírico voltado a testar e especificar, ainda mais, o conceito de cognição jornalística distribuída.

Pensamento computacional

Numa tentativa de delinear o significado histórico do conceito de computação, Denning e Martell sugerem que "computação eram os passos mecânicos seguidos para avaliar funções matemáticas [e] computadores eram as pessoas que faziam computações" (2015, p. 1). Nos anos 1980, porém, o conceito passou a ser associado com maior frequência a uma nova forma de fazer ciência, mudando sua ênfase de máquinas para processos de informação (Denning e Martell, 2015).

Esta mudança é essencial para o meu argumento, pois se alinha aos objetivos finais da reportagem e da computação: jornalismo também tem a ver com gerenciar processos informacionais; em linhas gerais, o trabalho do jornalista consiste em dinamizar o fluxo de informação, fazendo sua curadoria e apresentando-a em um formato palatável para o público. Neste ponto, argumentaria que a penetração de uma mentalidade computacional no jornalismo se dá, em partes, por conta das semelhanças entre ambas as práticas profissionais.

Computação e jornalismo são formulaicos, têm a ver com a solução de problemas e exigem maestria sintática. Wing comenta que "em termos operacionais, a computação volta-

se à pergunta ‘como poderia fazer um computador resolver este problema?’” (2008, p. 3.719), e isso já demanda um nível relativamente alto de pensamento computacional. Ao passo que a computação vira parte obrigatória da redação, o raciocínio voltado a ela é empregado por um número cada vez maior de jornalistas ao abordarem narrativas de dados. Bradshaw, por exemplo, argumenta que o pensamento computacional “está no coração do trabalho do jornalista de dados”, possibilitando estes “resolverem problemas tão caros ao jornalismo moderno, com a agilidade e precisão que o processo noticioso exige” (2017).

O pensamento computacional é o processo reflexivo pelo qual uma série de passos programáticos são dados para solucionar um problema (Wing, 2006; Wing, 2008; Bradshaw, 2017). Wing explica que “a essência do pensamento computacional é a abstração” (2008, p. 3.717). Ela fala ainda que, se tratando de computação, cientistas de dados e de computação, desenvolvedores ou programadores abstraem noções além das dimensões físicas do tempo e espaço (2008, p. 3.717) para solucionarem problemas, criarem sistemas e compreenderem o comportamento humano (Wing, 2006). A autora firma que, de forma a responder a pergunta ‘Como poderia fazer um computador resolver este problema?’, profissionais de computação precisam identificar abstrações apropriadas (2008, p. 3.717) que possam ser usadas na criação e implementação de um plano programático que resolva o problema em questão.

Desde que tecnologias de automação foram inseridas nas redações, jornalistas que lidam com profissionais de computação veem-se às voltas com a seguinte pergunta: ‘Como poderia fazer um computador investigar ou escrever um artigo voltado a padrões humanos?’ Gynnild propõe que a infusão do pensamento computacional dentro do jornalismo profissional desafie “o sistema fundamental de raciocínio jornalístico, da narrativa descritiva ao raciocínio abstrato, pesquisa autônoma e visualização de fatos quantitativos” que confere aos jornalistas “habilidades, atitudes e valores complementares voltados à lógica e algoritmos” (2013, p. 13).

Claro, não afirmo que a ideia de abstração computacional seja novidade para os jornalistas. De fato, profissionais cobrindo temas como finanças, negócios, imóveis, ou educação lidam com abstrações diariamente para entender dinâmicas complexas como desempenho de mercado, dividendos, patrimônio líquido doméstico etc. É interessante notar, como dito por Myles, que ao contrário do que se esperava da automação (que esta pouparia jornalistas de tarefas onerosas), ela acabou por introduzir uma nova série de atividades editoriais que antes não cabiam a estes profissionais. Por exemplo, ele explica que a introdução do reconhecimento de imagens ao fluxo de trabalho da *AP* acabou colocando jornalistas e fotógrafos para desempenharem tarefas tradicionalmente associadas à aprendizagem de máquina, como rotulagem de dados de treinamento, avaliação de resultados de teste, correção de metadados ou geração de definição para conceitos (Myles, 2019).

Projeção cognitiva e criatividade estendida

Até então, argumentei a respeito de jornalistas que, como parte de seu ofício, têm de lidar com problemas computacionais introduzidos por processos de automação que transformaram seus fluxos de trabalho e responsabilidades editoriais. O *Wall Street Journal*, por exemplo, há pouco anunciou vagas como Jornalista de Aprendizagem de Máquinas, Editor de Automação e Editor de Processos Emergentes, todas associadas à expansão de sua IA e automação. Como resultado deste tipo de expansão infraestrutural e subsequente diversificação de responsabilidades editoriais, jornalistas muitas vezes se pegam fazendo perguntas que os projetam ao papel de uma máquina que precisa pensar e agir como um jornalista. Um paradoxo interessante, que carrega consigo desafios igualmente interessantes.

Esta ideia de projeção, creio, está se tornando comum na automação de notícias. Tomemos, por exemplo, o maior dos empreendimentos jornalísticos: escrever um artigo. Se desconstruirmos o processo, em linhas gerais, o jornalista tem que usar sua criatividade para organizar uma série de eventos que engaje e/ou informe o público. A questão é: como fazer para uma máquina escrever algo que soa como se tivesse sido feito por um repórter humano? Jornalistas e tecnólogos vêm colaborando juntos nestes últimos cinco anos de forma a responder esta pergunta. Um bom exemplo nesta seara é a implementação de tecnologias de Geração de Linguagem Natural (GLN) na automação de matérias. Mas, ao contrário do que esperamos, o processo ainda envolve repórteres humanos criando modelos de artigos, com espaços em branco a serem preenchidos por software de automação através do uso de um banco de dados. Este processo, bastante bem-sucedido em organizações como a *AP*, ou a colaboração entre PA e Urbs Media, RADAR, que busca aumentar a velocidade e escala da produção de notícias em áreas como esportes, ganhos de empresas e noticiário local.

Neste segmento, a criatividade ganha nova forma, em que jornalistas-programadores têm de repensar a narrativa como uma máquina que decodifica e recodifica o processo de se fazer notícia. Ao invés de pensar qual entrevista melhor apoiaria um argumento ou quais palavras criariam uma manchete mais forte, o objetivo mudou para que a configuração de sentenças condicionais fosse mais eficiente para que o sistema automatizado decidisse qual manchete teria maior apelo ao público da organização em que opera. Ao seguir os princípios da Interação Humano-Computador (IHC) e Experiência do Usuário (UX), jornalistas-programadores precisam antever as formas pelas quais usuários engajarão com experiências informacionais automatizadas, as possíveis maneiras pelas quais navegarão através das diferentes camadas de informação e limites da matéria publicada. Wheeler, ao conceitualizar a ideia de criatividade estendida, explicou que há casos de criação intelectual em que "os veículos materiais responsáveis pelo raciocínio e pensamentos em questão estão distribuídos no espaço ao longo do cérebro, corpo e mundo" (Wheeler, 2018). O conceito de criatividade estendida funciona bem como base para explicar a ideia de que a mente de um jornalista

trabalhando com dados e automação agora funciona bem próxima a uma série de automações, que se espalha em bibliotecas Python, notebooks Jupyter, conjuntos de dados, ferramentas analíticas e plataformas online. Esta dinâmica, por consequência, adiciona mais e mais desafios dignos de atenção.

Mevan Babakar, Chefe de Checagem Automatizada de Fatos da Full Fact, explica que um dos desafios enfrentados por seu checador de fatos automatizado é o contexto. Ela cita como exemplo uma fala de Theresa May em que esta afirma que seu governo alocou mais recursos para o Sistema Nacional de Saúde britânico do que o que o Partido Trabalhista havia prometido em seu manifesto. E por mais que a declaração tenha sido checada e declarada como precisa, para que esta tenha significado e utilidade para o público, precisa ser compreendida dentro de contexto mais amplo: os recursos atuais não são o bastante para que o sistema opere de maneira eficiente (Babakar, 2018). Logo, a partir do momento que sistemas automatizados são incapazes de fazer tais ligações contextuais entre fontes de informação, Babakar e sua equipe tiveram que fazer perguntas como ‘como fazer com que um checador de fatos automatizado entenda as nuances do contexto?”

Cognição jornalística distribuída

Para concluir, gostaria de explorar um pouco mais a ideia de cognição jornalística distribuída e as questões levantadas por esta. Anderson, Wheeler e Sprevak argumentam que ao passo que computadores se inserem na atividade humana, a cognição “se espalha pelo cérebro, pelo corpo não neural e [...] um ambiente consistido por objetos, ferramentas, outros artefatos, textos, indivíduos, grupos e/ou infraestruturas sociais/institucionais” (2018). No contexto do jornalismo, isto significa que, atualmente, quando jornalistas usam software e hardware interligados para aprimorarem sua capacidade de produção de notícias em termos de escala e velocidade, sua cognição é distribuída ao longo das plataformas e ferramentas utilizadas. Isto, é claro, lhes dá acesso ilimitado à maior parte do conhecimento humano disponibilizado online.

A ideia de conhecimento portátil e cognição distribuída, porém, faz com que nos perguntemos quem detém e gerencia o acesso dos jornalistas ao filão de conhecimento e poder analítico “gratuito”? Quem possibilita a cognição jornalística distribuída? Esta questão, digna de uma discussão mais aprofundada, é bastante espinhosa, basta lembrar o que aconteceu quando a Google encerrou sua ferramenta online de visualização de dados, o Google Fusion Tables. Após o encerramento da plataforma, dezenas de projetos de jornalismo de dados que haviam sido desenvolvidos com ajuda da ferramenta ficaram indisponíveis, sem suporte da empresa.

Neste contexto, jornalistas lidam com dinâmicas computacionais diariamente, seu pensamento computacional é normalizado e facilita a projeção de sua cognição nas máquinas que usam ao longo de sua rotina de trabalho. Com a distribuição do conhecimento jornalístico, o mesmo acontece com sua autoridade e controle? Inexoravelmente, a distribuição muda os limites que dão aos jornalistas controle sobre suas rotinas e culturas profissionais, impactando sua autoridade epistemológica. Pensando mais adiante, como feito na introdução fictícia deste capítulo, a distribuição também pode criar uma série de riscos associados, assim que os jornalistas passarem a delegar questões e decisões éticas relevantes às máquinas. É importante, então, que a infraestrutura utilizada para a distribuição de sua cognição seja aberta e disponível ao escrutínio público, se os pilares do jornalismo serão preservados na era dos dados e automação.

Eddy Borges-Rey é estudioso de jornalismo e mídia, Decano Associado de Pesquisa na Escola de Artes e Ciências Humanas da Universidade de Stirling, ex-jornalista de radiodifusão.

Referências

ANDERSON, Miranda; WHEELER, Michael; SPREVAK, Mark. *Series Introduction: Distributed Cognition and the Humanities*. Volumes 1-4 da série *History of Distributed Cognition Series*. Edinburgh: Edinburgh University Press, 2018.

BRADSHAW, Paul. *Computational thinking and the next wave of data journalism*. Online Journalism Blog, 2017. Disponível em: https://onlinejournalismblog.com/2017/08/03/computational-thinking-data-journalism/.

DENNING, Peter; MARTELL, Craig. *Great principles of computing*. Massachusetts: MIT Press, 2015.

GYNNILD, Astrid. *Journalism innovation leads to innovation journalism: The impact of computational exploration on changing mindsets*. Journalism 15.6: 713-730, 2014.

MYLES, Stuart. *Photomation or Fauxtomation? Automation in the Newsroom and the Impact on Editorial Labour — A Case Study*. Computation + Journalism Symposium University of Miami, 1 e 2 de fevereiro de 2019.

WHEELER, Michael. *Talking about more than Heads: The Embodied, Embedded and Extended Creative Mind.* In: GAUT, Berys; KIERAN, Matthew (ed.). Creativity and Philosophy. Londres: Routledge, 2018.

WING, Jeannette. *Computational thinking*. Commun. ACM 49, 2006, p. 33–35.

WING, Jeannette. *Computational thinking and thinking about computing*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 366.1881: 3717-3725, 2008.

Vivência com dados

Formas de fazer jornalismo de dados

Sarah Cohen

dados (*da-dos)*: corpo de fatos ou informação; dados individuais, estatísticas ou itens de informação.

Jornalismo: ofício que consiste na reportagem, escrita, edição, fotografia ou radiodifusão de notícias ou operação de quaisquer organizações de notícias enquanto negócio

Caso esteja lendo esse livro, você escolheu aprender um pouco mais sobre o ofício que veio a ser conhecido como jornalismo de dados. Mas o que isso quer dizer em uma era de portais de dados de livre acesso, visualizações incríveis e disputas por liberdade de informação travadas por todo o mundo?

Uma definição tirada do dicionário das duas palavras ali em cima não ajuda muito, pois sugere que o jornalismo de dados é a prática de produzir notícias baseadas em fatos ou informações. Atualmente, jornalismo de dados é virtualmente qualquer ato de jornalismo que tenha alguma ligação com registros ou estatísticas em formato eletrônico; ou seja, todo tipo de jornalismo.

Por isso, muita gente atuando na área não se considera um jornalista de dados, na maior parte do tempo se vendo como escritores que se dispõem a explicar coisas, jornalistas gráficos ou visuais, repórteres, analistas de público ou desenvolvedores de aplicações de notícias, todos nomes muito mais precisos que descrevem as muitas tribos que compõem este campo em expansão. Obviamente, isso não é o suficiente, então pode adicionar qualquer outra função na redação que exija o uso de números ou programação. O que antes era uma bandinha de garagem, agora tinha gente o bastante para virar uma orquestra.

Jornalismo de dados não é exatamente novidade. Se você encarar "dados" como qualquer tipo de coleção sistemática, então alguns dos primeiros trabalhos em jornalismo de dados nos EUA ocorreram em meados de 1800. Naquela época, Leslie Frank contratou detetives para seguirem carrinhos de laticínios por Nova York para documentar ocorrências de leite contaminado ou rotulados erroneamente. Scott Klein, editor-chefe do site investigativo sem fins lucrativos *ProPublica*, documentou uma fascinante história sobre jornalismo de dados também ocorrida lá pelos idos de 1800, em que jornais ensinaram seus

leitores a interpretarem gráficos de barra. Chris Anderson discorre sobre diferentes genealogias do jornalismo de dados nos anos 1910, 1960 e 2000 em seu capítulo neste livro.[94]

Com estas histórias, taxonomias de diferentes ramos do jornalismo de dados podem ajudar estudantes e profissionais a decidirem o que querem de suas carreiras e as habilidades necessárias para que sejam bem-sucedidos. Estas diferentes formas de fazer jornalismo de dados são apresentadas aqui em uma cronologia aproximada do desenvolvimento da área.

Jornalismo empírico ou dados a serviço de matérias

Maurice Tamman, da *Reuters*, cunhou o termo "jornalismo empírico" de forma a combinar duas tradições em jornalismo de dados. O jornalismo de precisão, desenvolvido nos anos 1960 por Philip Meyer, visava usar metodologia de ciências sociais em suas matérias. Seu trabalho ia desde pesquisa com manifestantes em Detroit ao direcionamento da coleta de dados e análise de uma investigação sobre vieses raciais nos tribunais da Filadélfia. Meyer cimentou as bases do jornalismo investigativo por toda uma geração. Já o jornalismo empírico pode englobar aquilo que se chamou de reportagem assistida por computador nos anos 1990, gênero liderado por Eliot Jaspin em Providence, Rhode Island. Neste ramo jornalístico, repórteres buscam evidências documentais em formato eletrônico — ou as criam, quando necessário — para investigar alguma dica ou pauta.

Há pouco, estes repórteres passaram a utilizar inteligência artificial e aprendizagem de máquina para auxiliar nesta busca ou simplificar o desenvolvimento de uma matéria. Tais práticas podem ser usadas para responder a perguntas simples, como o gênero de um paciente ferido por dispositivos médicos em casos em que o governo tentou ocultar esta informação. Ou podem ser usadas, também, para identificação de padrões mais complexos, como a análise de Peter Aldhous sobre aviões espiões para o *Buzzfeed*.[95]

Estes jornalistas são quase que coletores de notícias puros, seu objetivo não é gerar uma visualização ou contar uma história através de dados. Em vez disso, eles usam registros para explorar matérias em potencial. Seu trabalho é parte integral do projeto de reportagem, muitas vezes guiando o desenvolvimento de uma investigação. Normalmente, não se envolvem tanto na apresentação desta história.

Discutivelmente, o que há de mais novo nesse mundo do jornalismo de dados pode muito bem ser o impacto crescente de investigações visuais e de código aberto pelo mundo. O gênero, derivado de pesquisas em inteligência e direitos humanos, expande nossa noção de

---

[94] https://www.propublica.org/nerds/infographics-in-the-time-of-cholera.

[95] https://www.icij.org/blog/2019/10/using-the-power-of-machines-to-complete-impossible-reporting-tasks/, https://www.buzzfeednews.com/article/peteraldhous/hidden-spy-planes.

"dados" para outros formatos, como vídeos, mídias sociais colaborativas e demais artefatos digitais. Menos dependentes de programação, encaixam bem na tradição do jornalismo de dados ao revelarem, por meio de pesquisa, o que outros gostariam de manter em segredo.

Um dos exemplos mais famosos, *The Anatomy of A Killing*, desenvolvido pelo *Africa Eye* da *BBC*, revela o local exato do assassinato de uma família em Camarões, bem como quando aconteceu, ajudando na identificação dos envolvidos — após o governo camaronês ter tratado tudo como "fake news".[96] A equipe usou ferramentas que iam do Google Earth, para identificação do delineado de uma montanha, ao Facebook, para documentar as roupas usadas pelos assassinatos.

Visualização de dados

Ao observar os vencedores do Data Journalism Awards, um leitor poderia pensar que a visualização é essencial para qualquer material em jornalismo de dados.[97] Se estatísticas são moeda de troca, a visualização é o preço do ingresso.

Visualizações podem ser ferramentas muito importantes do arsenal do jornalista de dados. Elas exigem habilidades ligadas ao mundo do design e da arte, bem como do mundo dos dados, estatísticas e reportagem. Alberto Cairo, um dos mais famosos jornalistas visuais em atuação na academia atualmente, veio dos infográficos de revistas e jornais. Seu trabalho é voltado a contar histórias através de visualização, num processo que envolve tanto noticiar quanto contar algo.

Aplicações de notícias

Na *ProPublica*, a maior parte das investigações começa ou encerra com uma aplicação de notícias — um site ou funcionalidade que fornece acesso a dados locais ou individuais por meio de uma interface interessante e intuitiva. Se a *ProPublica* ficou conhecida por suas aplicações voltadas a notícias, engenheiros de software que começaram suas carreiras em programação acabaram por evoluir ao ponto de se tornarem jornalistas que usam código, e não palavras, em suas matérias.

Ken Schwenke, da *ProPublica*, desenvolvedor por formação que trabalhou em redações como a dos *Los Angeles Times* e *The New York Times*, veio a ser um dos principais jornalistas cobrindo crimes de ódio nos EUA no projeto Documenting Hate, realizado em torno de colaborações vindas da aplicação de notícias da *ProPublica*.

---

[96] https://www.youtube.com/watch?v=4G9S-eoLgX4.

[97] Ver o capítulo assinado por Loosen.

Narrativas de dados

O termo "jornalismo de dados" amadureceu ao passo que repórteres, estatísticos e demais especialistas começaram a escrever sobre dados como uma espécie de jornalismo. Simon Rogers, criador do "Data Blog" do *The Guardian*, popularizou o gênero. *FiveThirtyEight.com*, *vox.com* e, posteriormente, o *Upshot* do *The New York Times*, se tornaram estandartes do setor. Cada um tem uma visão própria de seu papel, mas convergem na ideia de que estatísticas e análises são dignas de nota por si só.

Alguns ficaram conhecidos por conta de suas previsões políticas, discutindo probabilidades na corrida presidencial dos EUA. Outros ganharam fama ao encontrarem conjuntos de dados únicos que permitem um vislumbre da psique pública, caso do mapa de preferências de beisebol nos EUA de 2014 feito a partir de informações colhidas no Facebook. Conjuntos de dados são a força-motriz deste tipo de jornalismo, pareados com conhecimento especializado sobre determinado assunto; é com a combinação destes que os praticantes do ofício se destacam do restante. De fato, Nate Field e outros que vieram a definir este gênero jornalístico não tinham histórico como jornalistas, vindo de campos como estatística e ciência política.

Amanda Cox, editora do *Upshot* do *The New York Times*, disse a estudantes interessados em sua perspectiva do tema que seu o papel do site é ocupar o espaço entre fatos conhecidos, incontestáveis, e o desconhecido, um jornalismo que oferece percepções baseadas em análise especializada de dados disponibilizados entre fato e opinião.

Investigando algoritmos

Uma área emergente dentro do jornalismo de dados é, no final, jornalismo de tecnologia — a área da "responsabilização algorítmica", termo cunhado por Nick Diakapolis, da Northwestern University. Julia Angwin e Jeff Larson, dois jornalistas, deixaram a *ProPublica* para seguirem neste campo com a criação do site *The Markup*, que Angwin afirma ter como objetivo responsabilizar empresas de tecnologia pelos resultados criados por algoritmos de aprendizagem de máquina e inteligência artificial em nossa sociedade, de condenações à prisão aos preços cobrados com base no CEP do consumidor.

A iniciativa já fez com que o YouTube revesse seu sistema de recomendações para que fosse reduzida a tendência de levar espectadores a vídeos cada vez mais violentos. Também responsabilizou o Facebook por seus anúncios de moradia potencialmente

discriminadores, além de ter identificado diferenças nos preços em lojas online com base na localização do usuário.[98]

Sarah Cohen é vencedora do Pulitzer, ex-jornalista e editora do The Washington Post e The New York Times, especializada em jornalismo de dados, atua na Cátedra Knight em Jornalismo de Dados da Escola Cronkite de Jornalismo, Universidade Estadual do Arizona.

98 https://www.washingtonpost.com/technology/2019/01/25/youtube-is-changing-its-algorithms-stop-recommending-conspiracies.

Visualizações de dados: tendências de redação e engajamento cotidiano

Helen Kennedy, William Allen, Martin Engebretsen, Rosemary Lucy Hill, Andy Kirk, Wibke Weber

Este capítulo fala sobre a produção de visualização de dados ("datavis", daqui em diante) em redações e o engajamento cotidiano do público com datavis, com base em dois projetos de pesquisa distintos. *Seeing Data* é o nome do primeiro projeto, uma exploração de como as pessoas interpretam visualizações de dados, o segundo é o *INDVIL*, uma exploração de datavis enquanto recurso semiótico, estético e discursivo na sociedade.[99] O capítulo começa com um resumo das principais descobertas do subprojeto *INDVIL*, focado em datavis no noticiário, onde descobrimos que essas informações são percebidas de diferentes maneiras e implementadas com diferentes propósitos. Então, apresentamos um sumário das principais descobertas feitas no *Seeing Data,*[100] onde também encontramos grande diversidade, desta vez na forma como públicos interpretam datavis. Tal diversidade é importante para o trabalho futuro de pesquisadores e praticantes de datavis.

Visualizações de dados em redações: tendências e desafios

Como a visualização de dados vem sendo incluída na prática de redação? Que tendências e desafios vêm surgindo? Para responder a esta pergunta, em 2016 e 2017 conduzimos 60 entrevistas em 26 redações espalhadas por seis países europeus: Noruega (NO), Suécia (SE), Dinamarca (DK), Alemanha (DE), Suíça (CH) e Reino Unido (GB). Dentre os entrevistados, membros de grandes organizações online de notícias, incluindo aí editores-chefe, chefes de equipes especializadas em visualização de dados, jornalistas visuais, designers gráficos/de visualização de dados e desenvolvedores (por mais que alguns não tivessem cargos específicos, indicador de que este é um campo de rápido desenvolvimento). Apresentamos a seguir alguns dos destaques de nossa pesquisa.

Mudanças no storytelling jornalístico

O uso crescente de visualizações de dados dentro do jornalismo representa uma mudança da escrita enquanto método semiótico principal para o uso de dados e visualizações, agora elementos centrais, no storytelling jornalístico. Muitos dos entrevistados citaram a

[99] http://seeingdata.org/, https://indvil.org/.

[100] A segunda parte do capítulo é um resumo de outro artigo mais longo, disponível online. Kennedy et al. (2016).

visualização de dados como a principal força em um artigo ou matéria, mesmo quando se trata de um gráfico ou diagrama simples.

> "As estatísticas de leitores indicam que quando inserimos uma visualização de dados simples em uma matéria, estes leitores passam mais tempo na página" (SE).

Datavis integram uma ampla gama de intenções comunicativas, dentre as quais: "oferecer percepções" (GB), "explicar com maior facilidade" (SE), "comunicar de forma mais clara que palavras poderiam" (GB), "falar detalhadamente sobre diversas facetas de um assunto, o que só é possível em texto de forma agregada" (DE), para tornar artigos "mais acessíveis" (DK), "para revelar o estado deplorável das coisas" (CH), "para ajudar pessoas a entenderem o mundo" (GB). A visualização de dados é usada para enfatizar um argumento, adicionar evidência empírica, permitir aos usuários explorarem séries de dados, como atração estética para estimular interesse e oferecer um ponto de entrada a histórias não vistas antes.

Estas mudanças vêm acompanhadas de grupos especializados com diversas habilidades dentro da redação, com a priorização de habilidades voltadas a dados e datavis para novos recrutas. Dito isso, não existe um padrão na organização da produção de datavis dentro da redação. Em algumas situações, tudo ocorre dentro do contexto de equipes de dados, em outros, equipes visuais (um dos designers de datavis com quem conversamos também trabalhava em um projeto de realidade virtual à época da entrevista) ou outras equipes completamente diferentes. E, assim como novas estruturas vêm surgindo para acomodar este novo formato visual emergente, a equipe de redação também precisa se adaptar ao aprendizado de novas ferramentas, internas e comerciais, ao desenvolvimento de novas habilidades e ao entendimento de como se comunicar ao longo de diversas equipes e áreas de conhecimento de forma a produzirem materiais eficazes com base em dados.

O mantra 'mobile em primeiro lugar' e suas consequências

O reconhecimento amplo de que o público cada vez mais consome notícias em telas pequenas de dispositivos móveis levou à adoção igualmente ampla do mantra 'mobile em primeiro lugar' quando se trata da produção de datavis em redações. Isso representa um distanciamento das visualizações elaboradas e interativas, características dos primórdios de datavis no contexto de notícias, com um maior foco em simplicidade e linearidade, ou formas visuais simples com baixos níveis de interação. Isso levou a uma predominância de certos tipos de gráficos, como os de barra ou linhas, bem como ao advento do *scrollytelling*, narrativas que se desenrolam com o movimento causado pelo usuário na barra de rolagem, em que as visualizações embutidas no artigo aparecem na hora certa. O ato da rolagem também causa mudanças nas visualizações em si, afastando-se da imagem, por exemplo.

> "Em nossos artigos, é comum usarmos a técnica da rolagem. Não é necessário clicar, mas ao rolar a tela para baixo, algo acontecerá no artigo" (DE).

Ferramentas de automação de datavis dão a chance a jornalistas que não são especialistas no assunto de produzirem gráficos simplificados. De qualquer forma, alguns dos entrevistados se dispõem a educar seus leitores ao apresentarem materiais mais incomuns (um gráfico de dispersão, por exemplo) acompanhados de informações sobre como interpretá-los. Alguns creem que imagens também pode apresentar dados de maneira eficaz — um tablóide escandinavo de alcance nacional representou o tamanho de um avião de carga ao enchê-lo com 427.000 pizzas. Outros reconhecem o valor no uso de animações, por exemplo, para demonstrar mudanças ao longo do tempo, ou experimentar com zoom em visualizações.

O papel social do jornalismo

Ligar datavis a uma fonte de dados, fornecendo acesso aos dados crus e explicando as metodologias, é encarado por alguns dos participantes como boa prática ética de forma a promover transparência e contrabalançar a subjetividade da seleção e interpretação que, para alguns, é inevitável na visualização de dados. Para outros, fazer esta ligação significa dar ao público 'todos os dados', o que entra em conflito com a norma jornalística de identificar e, então, contar uma história. Para alguns, este conflito é abordado através dos processos complexos de compartilhamento de diferentes elementos de dados e processamento em várias plataformas (Twitter, Pinterest, GitHub).

Isso leva jornalistas de dados e designers de visualização a refletirem o quanto de dados compartilhar, seu papel como provedores de fatos e papel social, em linhas gerais. Paul Bradshaw, jornalista de dados, resume em seu blog:

> Quanto de responsabilidade temos em relação às histórias que contamos às pessoas com nossas informações? E quanta responsabilidade temos por entregarmos o tanto de informação que alguém precisa? (Bradshaw, 2013). [101]

O ex-editor do *Guardian* Alan Rusbridger (2009) levantou questão semelhante sobre o papel social do jornalismo ao apontar a gama de atores que fazem o que o jornalismo fez historicamente — ou seja, atuar como guardião de dados e informações oficiais (caso dos sites FixMyStreet e TheyWorkForYou, no Reino Unido). Ele concluiu o seguinte: "Não sei se isso é jornalismo ou não. Não sei se isso importa"(Baack, 2018). Alguns de nossos entrevistados trabalham em grandes projetos semelhantes aos discutidos por Rusbridger — por exemplo, um destes agrupava todos os dados relacionados a escolas no Reino Unido e os

[101] https://onlinejournalismblog.com/2017/09/14/narrative-storytelling-data-journalism-alberto-cairo/.

tornava exploráveis por CEP para ajudar na tomada de decisão quanto a estas escolas. Sendo assim, a questão do que vale o jornalismo no contexto de dados amplamente difundidos e datavis não é nada fácil de responder.

Além do que, o ato de compartilhar conjuntos de dados pressupõe que o público interagirá com estes, por mais que estudos indiquem que a interatividade online é, ao mesmo tempo, mito e realidade, com a imagem idealizada de um explorador ativo e motivado de uma série de dados visualizados contrastando com a do leitor que faz uma leitura rápida das notícias (Burmester et al., 2010). Similarmente, um estudo sobre os projetos de jornalismo de dados enviados ao Nordic Data Journalism Awards conclui que elementos interativos muitas vezes oferecem apenas uma *ilusão* de interatividade, já que a maioria das escolhas foi feita ou predefinida pelos jornalistas (Appelgren, 2017). Isso mais uma vez nos faz questionar a prática de compartilhar 'todos os dados' e o papel social em constante mutação do jornalismo.

Confiança, verdade e visualizações 'em livre circulação'

Outros elementos do processo de visualização de dados levantam questões de confiança e verdade, relacionando-se também com a forma como jornalistas encaram o papel social do ofício. Um aspecto do trabalho com datavis que aponta para estas questões é como os profissionais lidam com dados e suas representações visuais. Alguns veem a prática como uma forma de revelar a verdade, outros como um processo de seleção e interpretação, e há ainda aqueles que acreditam que moldar a visualização de dados através de escolhas é um método de se revelar uma história e é exatamente o que jornalistas deveriam fazer. Estas reflexões destacam a relação entre (des)confiança e apresentação, perspectiva e (in)verdade.

No atual contexto da tal pós-verdade, em que dizem que o público já se cansou de fatos, dados e especialistas e em que notícias falsas circulam ampla e agilmente, nossos participantes estavam de olho nas possíveis formas pelas quais o público poderia reagir às suas visualizações de dados, que poderiam incluir a aceitação ingênua do que foi apresentado, a refutação cética, a descontextualização por meio de compartilhamentos sociais e até mesmo alterações e falsificações. Sentiam que os jornalistas cada vez mais precisam de 'conhecimento leve de cultura da internet', como mencionado por um entrevistado do Reino Unido. Isso inclui a compreensão de como o conteúdo online pode estar mais aberto a questionamentos que suas contrapartes offline, e como visualizações de dados estão mais suscetíveis a circulação na rede do que textos, livres de seus contextos originais, na forma de combinações de números e imagens 'em circulação livre' (Espeland and Sauder, 2007). Consequentemente, é necessária a compreensão de estratégias que possam lidar com estes perigos, como a inclusão de texto explicativo em um arquivo de visualização, de forma que a imagem não possa circular sem este. Tais questões, aliadas a preocupações sobre o letramento

do público em torno de dados e visualizações destes, educam e moldam o pensamento do jornalista em torno de sua audiência.

Como as pessoas interagem com visualizações de dados?

Nesta seção, observamos datavis no contexto de notícias sob a perspectiva do público. Como a audiência interage e interpreta visualizações encontradas no noticiário? Jornalistas de dados geralmente estão ocupados demais para lidar com isso. Já os pesquisadores em visualização de dados não têm essa desculpa, mas raramente voltam sua atenção ao que o usuário final pensa a respeito das visualizações com as quais interagem.

Entra aí o *Seeing Data*, projeto de pesquisa que visava explorar como as pessoas interagem com visualizações de dados com as quais se deparam no cotidiano, muitas vezes na mídia. Este projeto explorava os fatores que afetam engajamento e o que isso representa para a forma como encaramos aquilo que acreditamos fazer de uma visualização eficaz. No *Seeing Data* usamos grupos focais e entrevistas para explorar estas questões, de forma a compreender as atitudes, sentimentos e crenças subjacentes à interação das pessoas com datavis. Foram 46 participantes, incluindo pessoas que, presume-se, teriam interesse em dados, visual ou migração (tema de muitas das visualizações mostradas). Ou seja, gente já "engajada" em uma das questões que compunham o cerne do projeto, e pessoas das quais não poderíamos fazer tais suposições a respeito.

Nos grupos focais, pedíamos aos participantes que avaliassem oito visualizações, escolhidas (após muita discussão) porque representavam uma variedade de temas, tipos de gráficos, fontes de mídia, formatos e visavam explicar ou estimular a exploração. Metade das visualizações vinham de fontes jornalísticas (*BBC*; *The New York Times*; *The Metro*, jornal britânico de distribuição gratuita; e a revista *Scientific American*). Já as demais eram de organizações que criam estas visualizações e compartilham dados como parte de seu trabalho (Observatório de Migração da Universidade de Oxford; Escritório de Estatística Nacional do Reino Unido; e Organização para a Cooperação e Desenvolvimento Econômico – OCDE).

Após os grupos focais, sete participantes escreveram um diário ao longo de um mês, para que obtivéssemos mais informações sobre suas interações com visualizações 'em circulação livre', não escolhidas por nós

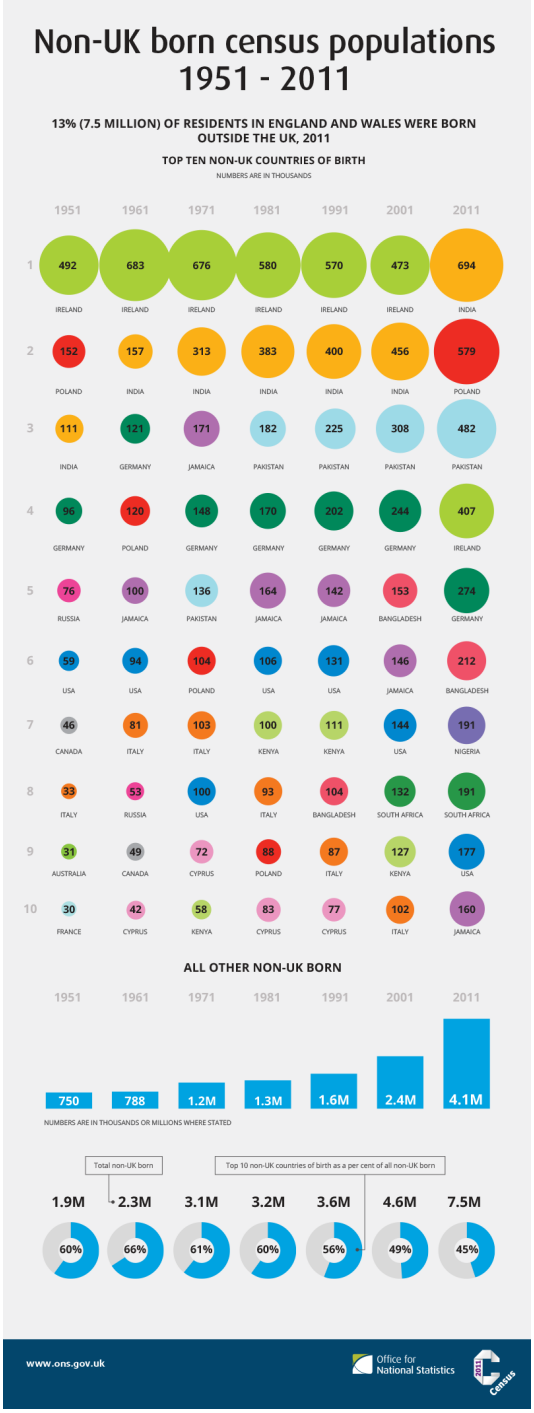


Figura 1: Censo populacional de não nascidos no Reino Unido, 1951-2011, Escritório de Estatística Nacional do Reino Unido.

Figura 2: Migração no censo, produzida para o Observatório de Migração da Universidade de Oxford.[102]

Fatores que afetam o engajamento com datavis

Tema

Visualizações não estão isoladas do tema que representam. Quando o tema era de interesse dos participantes, estes se mostravam engajados — no caso de profissionais da sociedade civil interessados em questões relacionadas à migração, havia interesse natural em visualizações ligadas a isto. Em contraste, um participante (homem, 38 anos, branco, britânico, agricultor) não tinha interesse algum nas visualizações mostradas durante o grupo focal, nem mesmo confiança para buscar mais informações sobre o tema. Porém, sua falta de interesse e confiança e mesmo suspeita em relação à mídia (afirmava que "eles tentam te confundir") o impediu de observar estas visualizações por inteiro; ele comentou que ao se deparar com esse tipo de material no *The Farmer's Guide*, publicação que lê regularmente, já que trata de temas de seu interesse, gastava um tempo nelas.

Fonte ou localização da mídia

A fonte das visualizações importa, pois existem implicações quanto à confiança dos usuários. Preocupações a respeito da mídia querer confundir o povo foram mencionadas por muitos participantes e levou alguns destes a encararem visualizações de certos veículos como suspeitas. Em contraponto a isso, alguns participantes demonstraram confiar em visualizações sobre migrações com o logo da Universidade de Oxford, pois sentiam que a "marca" da universidade invoca um ar de qualidade e autoridade. Mas, durante o período em que utilizaram o diário, as coisas mudaram de figura. Os participantes tendiam a interagir

[102] http://www.compas.ox.ac.uk/migrationinthecensus/, http://migrationobservatory.ox.ac.uk/.

com estas visualizações nos veículos de sua escolha, nos quais confiavam, então havia maior probabilidade de confiarem no que era visto ali. Um participante (homem, 24 anos, agricultor, branco, britânico), leitor do *The Daily Mail*, demonstrou isso quando disse durante entrevista que "você vê mais coisas erradas no *The Sun*, eu acho". Levando em conta as semelhanças ideológicas entre ambas as publicações, este comentário destaca a importância da localização da mídia no engajamento de datavis.

Crenças e opiniões

Os participantes confiavam nos jornais que liam regularmente, e, sendo assim, confiavam nas visualizações apresentadas nestes jornais, visto que o conteúdo de ambos muitas vezes batia com suas visões de mundo. Isso aponta para a relevância de crenças e opiniões e sua influência em como as pessoas interagem ou não com determinadas visualizações. Alguns participantes afirmaram gostar de visualizações que validavam suas crenças e opiniões. Mas essas crenças não importam somente quando visualizações as validam. Um participante (homem, 34 anos, branco, britânico, trabalha em TI) foi surpreendido com dados sobre migração em visualização do Escritório Nacional de Estatística do Reino Unido, apresentados na Figura 1. Ele disse não saber que tanta gente no Reino Unido havia nascido na Irlanda. Essas informações questionavam suas crenças e foi uma experiência da qual ele gostou. Algumas pessoas gostam, ou ao menos se interessam, por visualizações de dados que questionam crenças pré-existentes, pois provocam e desafiam percepções. Ou seja, crenças e opiniões importam nesse sentido também.

Tempo

Interagir com visualizações é encarado como trabalho por pessoas que não têm facilidade com a prática. Ter tempo disponível é essencial para determinar se as pessoas estão dispostas ou não a 'trabalhar'. A maioria dos participantes que afirmaram não ter tempo para visualizações era do sexo feminino e tal falta de tempo foi atribuída ao trabalho, à família e aos afazeres domésticos. Uma mãe trabalhadora falou que seu trabalho, tanto doméstico quanto remunerado, era tão cansativo que, ao final do dia, ela não queria nem ver notícias, o que incluía visualizações de dados. Tais atividades eram como 'trabalho' aos seus olhos, e ela estava cansada demais para isso ao fim de um dia cheio de afazeres. Um agricultor nos disse por email que seu horário de trabalho era tão longo a ponto de impactar sua capacidade de manter o diário mensal de interações com visualizações de dados após a pesquisa com grupos focais.

Confiança e habilidades

O público precisa sentir que tem as habilidades necessárias para decodificar visualizações, e muitos participantes demonstraram falta de confiança nesse sentido. Um dos

participantes, conselheiro vocacional em meio período, disse o seguinte a respeito de uma visualização em especial: “Tinha esse monte de cores e círculos, olhei aquilo e pensei que parecia trabalhoso demais; não sabia se havia entendido”. Muitas das pessoas que participaram da pesquisa demonstraram preocupação quanto à sua falta de habilidades, ou mesmo chegaram a demonstrar a falta destas, fosse uma questão de letramento visual, linguagem, matemática e estatística (saber como interpretar certos tipos de gráficos, por exemplo) e até mesmo pensamento crítico.

Emoções

Apesar de vir em último nesta lista, uma grande descoberta de nossa pesquisa foi o papel relevante das emoções quando se trata da interação das pessoas com visualizações de dados. Uma ampla gama de emoções ocorreu em relação a interações com datavis, incluindo prazer, raiva, tristeza, culpa, vergonha, alívio, preocupação, amor, empatia, empolgação e ofensa. Diversas respostas emocionais a visualizações, em geral, foram relatadas pelos participantes, estendendo-se aos dados representados, estilo visual adotado, tema, fonte ou localização original destas visualizações, e à própria habilidade dos participantes em interpretarem estas visualizações.

Por exemplo, dois profissionais da sociedade civil usaram linguagem forte para descrever como se sentiam em relação às visualizações de migração no Reino Unido mostradas na Figura 2. Tais dados os fizeram refletir como deve ser a experiência do imigrante que vem ao país e se depara com manchetes xenofóbicas na mídia. Descreveram-se como ‘culpados’ e ‘envergonhados’ por serem britânicos.

Outros tiveram forte reposta emocional ao estilo de algumas visualizações. Uma representação de recibos de cinema do *The New York Times* dividiu os participantes, alguns atraídos pela estética e outros dissuadidos:

> ‘Foi um prazer ver esta apresentação por conta da coordenação entre imagem e mensagem.
>
> Frustrado. Era uma apresentação feia, para começo de conversa, difícil de entender com clareza, sem informação, uma bagunça (Bloch, Carter e Cox, 2008).[103]

[103] http://archive.nytimes.com/www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html?_r=1.

O que isso tudo significa na hora de criar visualizações eficazes?

Uma série de visões sobre o que torna uma visualização eficaz despontou com base em nossa pesquisa. Representações na mídia voltadas a leigos podem ter o objetivo de persuadir, por exemplo. Todas precisam atrair pessoas, para que disponham de seu tempo para saber mais sobre os dados nos quais a visualização se baseia. Estas visualizações podem estimular emoções em particular, que inspirarão as pessoas a passarem mais tempo analisando o que foi apresentado ou mesmo a irem mais fundo e além. Podem provocar interesse ou o oposto disso. Uma visualização eficaz pode:

- Provocar questões/desejo de debater com outros
- Criar empatia em relação aos outros seres humanos apresentados naqueles dados
- Gerar curiosidade o bastante para prender o usuário
- Reforçar ou dar suporte a conhecimento prévio
- Provocar surpresa
- Persuadir ou mudar ideias
- Apresentar algo novo
- Gerar confiança na interpretação de datavis
- Apresentar dados úteis para fins daquele indivíduo
- Possibilitar engajamento informado ou crítico de determinado tópico
- Ser uma experiência agradável
- Provocar uma forte resposta emocional

Não há uma receita para o que torna uma visualização eficaz, não há uma definição única que se aplique no espectro como um todo. Por exemplo, se entreter com uma visualização é relevante em alguns contextos, mas não em outros. Estas representações têm diversos objetivos, dentre os quais comunicar novos dados, informar o público, influenciar tomadas de decisão, possibilitar exploração e análises de dados, surpreender e afetar comportamentos. Já os fatores que afetam o engajamento identificados em nossa pesquisa devem ser encarados como *dimensões* de eficácia, com pesos diferentes em relação a representações, contextos e propósitos diferentes. Muitos destes fatores estão fora do alcance

de profissionais de visualização, já que estão relacionados ao consumo e não à produção destas visualizações. Trocando em miúdos, a eficácia ou não de uma visualização depende em grande parte como, por quem, quando e onde é acessada. Infelizmente, nossa pesquisa não sugere um checklist simples que garanta a produção de representações de eficácia universal. Porém, se queremos criar visualizações acessíveis e eficazes, é importante que os jornalistas se envolvam com estas descobertas.

A pesquisa de Helen Kennedy cobriu diversos caminhos pela mídia digital; seu foco atual é a experiência vivida e visualizada de dataficação e fenômenos relacionados (algoritmos, IA, aprendizagem de máquina), desigualdades e perspectivas cotidianas a respeito de práticas de dados 'justas'. William Allen é Pesquisador do Centro de Migração, Políticas e Sociedade (COMPAS, na sigla em inglês), cujo trabalho é voltado a mídia, migração, políticas de dados e atitudes públicas. Martin Engebretsen é professsor de linguagem e comunicação da Universidade de Agder, especializado nos campos da análise de discurso multimodal e estudos de jornalismo. Rosemary Lucy Hill pesquisa gênero, música popular e políticas de visualização de dados, é autora de "Gender, Metal and the Media: Women Fans and the Gendered Experience of Music" (Palgrave). Andy Kirk é especialista em visualização de dados. Wibke Weber é professora de linguística de mídia e estuda visualização de dados, gráficos de informação, semiótica visual, jornalismo em quadrinhos, realidade virtual e multimodalidade.

Referências:

BAACK, Stefan. *Knowing What Counts: how data activists and data journalists appropriate and advance datafication*. Tese de PhD. Universidade de Groningen, 238 p., 2018. Disponível em: https://www.rug.nl/research/portal/files/56718534/Complete_thesis.pdf.

ESPELAND, Wendy Nelson; SAUDER, Michael. *Rankings and reactivity: how public measures recreate social worlds*. American Journal of Sociology, 113(1), 2007, p. 1-40.

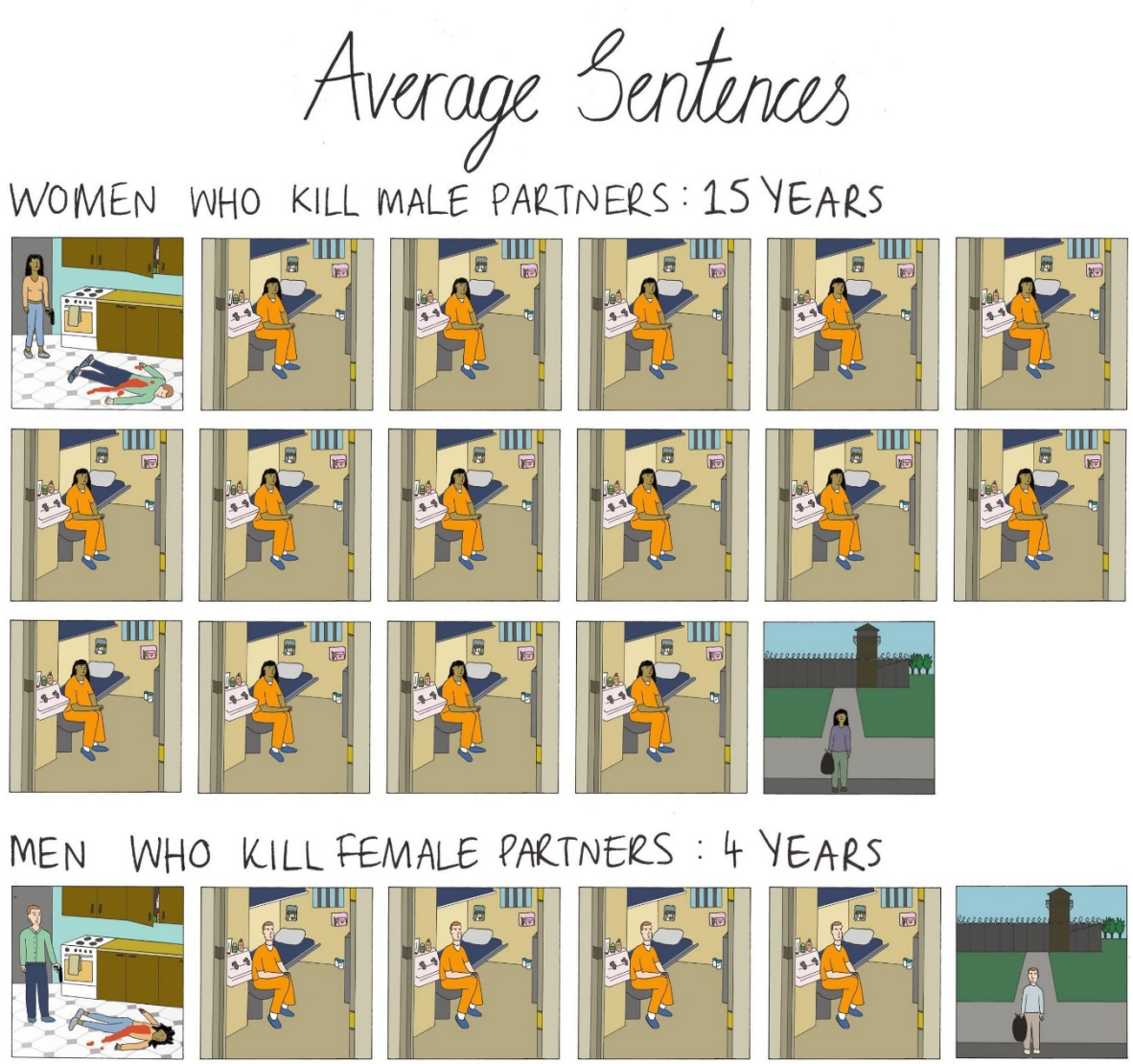
KENNEDY, Helen et al. *Engaging with (big) data visualisations: Factors that affect engagement and resulting new definitions of effectiveness*. First Monday 21:11, 2016.

KENNEDY, Helen; HILL, Rosemary. *The Feeling of Numbers: Emotions in Everyday Engagements with Data and Their Visualisation*. Sociology 52:4, 2018, p. 830-848.

RUSBRIDGER, Alan. *Why Journalism Matters*. Media Standards Trust Series, agosto de 2010. Disponível em: http://mediastandardstrust.org/wp-content/uploads/downloads/2010/08/Why-Journalism-Matters-Alan-Rusbridger.pdf.

Esboços com dados

Mona Chalabi[104]



Como você começou a criar esboços com dados?

Quando trabalhava no *FiveThirtyEight* sentia que não estava escrevendo para leitores como eu. Seu material interativo complexo atendia a um tipo de leitor um pouco diferente. Nessa época, comecei a fazer esboços, coisa que eu podia fazer ali da minha mesa mesmo. Quando comecei a fazê-los, percebi que poderiam ser uma maneira bastante eficaz de comunicar a incerteza de projetos de dados. Poderiam servir de lembretes às pessoas de que um ser humano foi responsável por tomar todas aquelas decisões de design. Poderiam ser bastante democratizadores também, comunicando dados de uma forma que qualquer um

[104] Este capítulo é baseado em uma entrevista com Mona Chalabi conduzida pelos editores.

poderia fazer. Eu costumava escrever uma coluna estilo faça-você-mesmo no *The Guardian* que guiava o público durante cada passo do processo. Era divertido, enquanto jornalista, guiar as pessoas não só por onde você encontrou seus dados, mas também ao longo de seu processamento e do que foi feito, dando ao público a chance de replicar tudo isso, derrubando a barreira entre eles e você, com sorte criando novos tipos de acessibilidade, participação e relações com os leitores.

No livro falamos sobre como jornalistas de dados não devem apenas refletir e reforçar tipos de conhecimento já estabelecidos (como ciência de dados e metodologia estatística avançada), podendo também promover outros tipos de práticas e culturas de dados. Você diria que seu trabalho, em parte, se relaciona à busca por outras formas de trabalhar e se relacionar com dados?

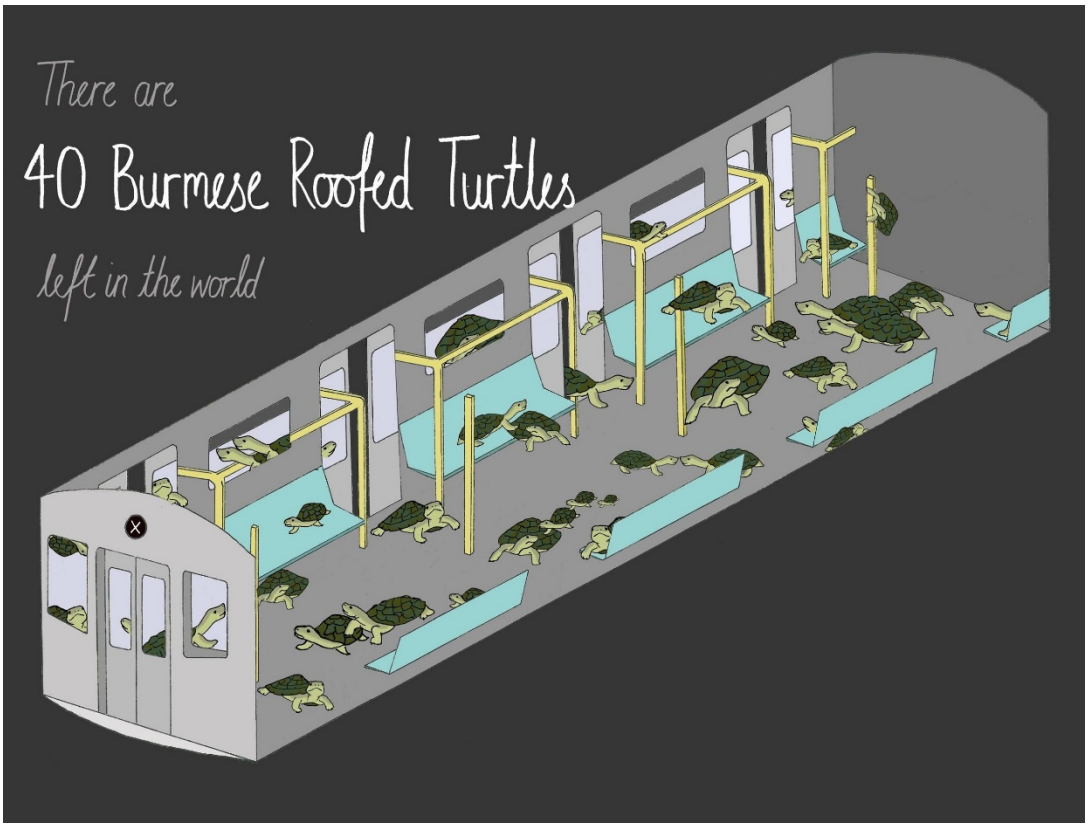
Eu não tenho conhecimento avançado em estatística. A forma como começo a analisar dados, na maioria das vezes, se dá através de cálculos simples que podem ser replicados por outras pessoas. De certa forma, isso torna os dados que uso muito mais confiáveis. Em determinado momento, ao utilizar abordagens estatísticas mais avançadas, você dá aos leitores um ultimato: há de se confiar na análise do jornalista ou não. Isto difere da noção de confiar em dados do governo e multiplicação básica ou não. Há certo benefício em fazer as coisas através de cálculos simples. E isso importa muito para o que faço e como faço.

Dados podem servir como oportunidade para fazer duas coisas: aproximar e distanciar. Por um lado, meu dever enquanto jornalista de dados é me distanciar de um incidente específico e dar aos leitores o contexto do ocorrido através de dados. Digamos que no advento de determinado incidente ou atentado, possamos mostrar como estes ocorrem, onde ocorrem, se sua prevalência aumenta ao longo do tempo e se há grupos mais afetados que outros. Eis uma oportunidade para leitores terem melhor compreensão de tendências mais amplas, o que pode ser extremamente informativo para eles; talvez os ajude a não surtarem ou surtar com razão em reação às notícias.

Por outro lado, podemos também fazer o oposto e nos aproximar. Digamos que a Secretaria de Estatísticas Trabalhistas (BLS, na sigla em inglês) dos EUA divulga dados sobre desemprego e a maioria dos veículos jornalísticos apenas divulga o índice de desemprego. Nós, jornalistas de dados, podemos nos aproximar dessa questão: podemos dizer aos leitores que aqui está o índice de desemprego nacional, mas veja só como se aplica às mulheres, a homens e diferentes faixas etárias, aqui como se aplica a diferentes grupos raciais e étnicos. Isso permite aos leitores explorarem os dados mais de perto.

Meu trabalho alterna entre estes dois modos. Creio que uma de minhas maiores críticas a veículos como o *FiveThirtyEight* é que o trabalho desempenhado pode, às vezes,

soar como bravata intelectual: “olha aqui o que conseguimos fazer”. Eu não gosto disso. Meu propósito é atender ao leitor e, em específico, à mais ampla comunidade de leitores possível, não só homens brancos que se dizem geeks. Os leitores do *FiveThirtyEight* se apresentam como geeks e os jornalistas do site também. Não foi para isso que entrei no jornalismo.



Pegando um exemplo recente do seu trabalho, poderia nos falar um pouco mais sobre o artigo *Endangered Species on a Train*, publicado no *The Guardian*? Como você se aproximou deste tema, como o projeto surgiu e como foi sua abordagem?[105]

Foi tudo bem estranho, na verdade. Não foi algo inspirado pelas notícias, tendo mais a ver com minha prática destas ilustrações e o desejo de fazer algo um pouco mais ambicioso. Parte do motivo pelo qual comecei a criar estas ilustrações é sua eficácia: as reações chegam rápido e elas podem ser completadas em horas, se necessário. Queria criar algo maior, que demandasse mais tempo. Comecei abordando um tema muito maior com os quais as pessoas já estão acostumadas — espécies em risco de extinção —, mas cuja linguagem visual preexistente talvez fosse pouco inspirada. Os dados vieram da “Lista Vermelha” da União Internacional para a Conservação da Natureza (UICN).[106] Muitos dos números a respeito de

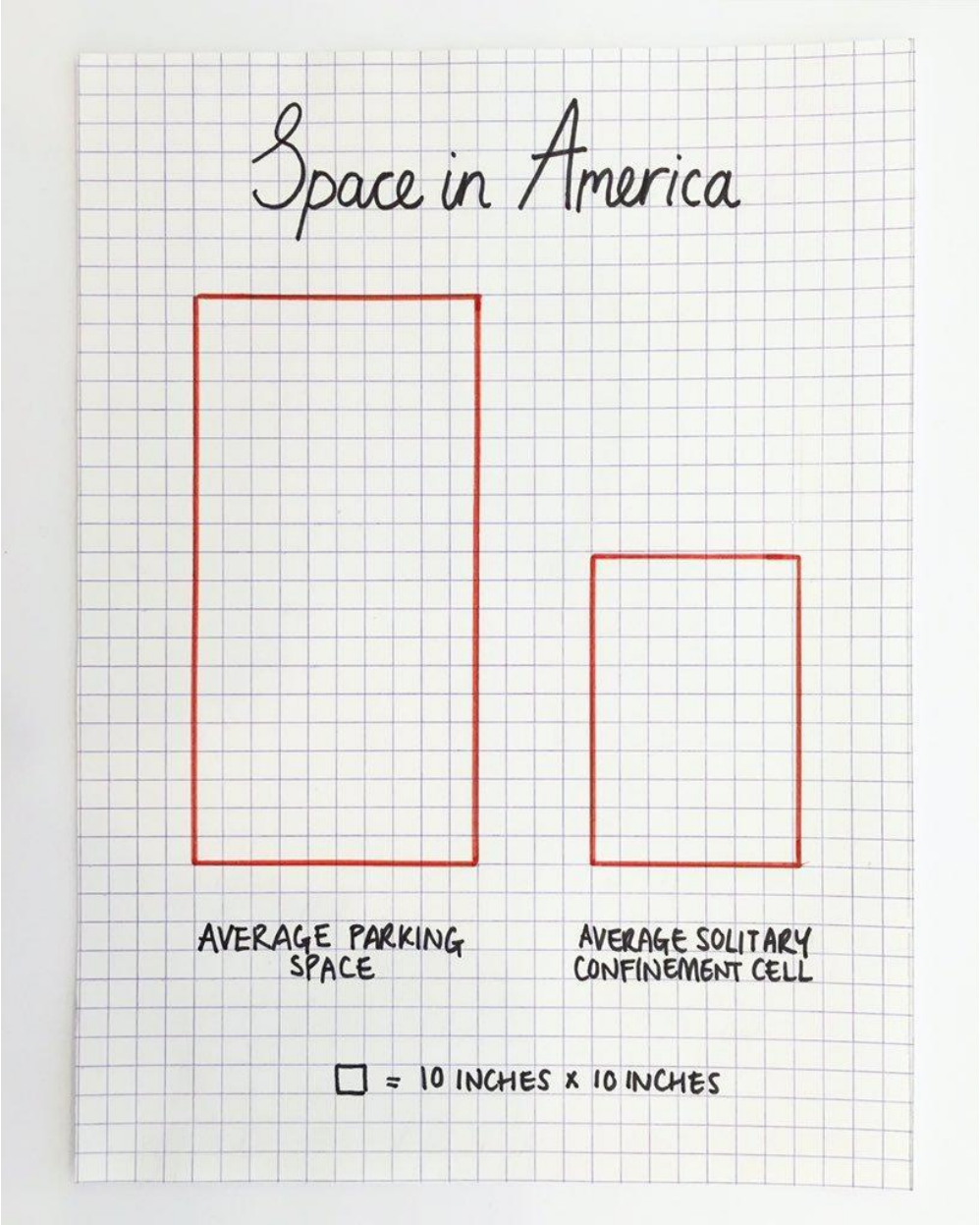
---

105 https://www.theguardian.com/environment/gallery/2018/sep/17/endangered-species-on-a-train.

106 https://www.iucnredlist.org/.

espécies ameaçadas estavam divididos em categorias, e escolhi um ponto médio para cada uma destas.

Dando um passo para trás, você poderia ver minhas ilustrações como gráficos. Escala é a única coisa que transforma estas representações em gráficos. Cada ilustração publicada tem um senso de escala e é exatamente assim que um gráfico se comporta. Um dos problemas disso é que diversos países e lugares usam escalas diversas, como milímetros no Reino Unido e polegadas nos EUA. Escalas significam coisas diferentes para pessoas diferentes. Muitos jornalistas de dados esquecem disso. O que “1 milhão” significa para alguém? O que “1” significa para alguém? Tudo isso depende de contexto. Quando se lida com números baixos, pode ser mais fácil lidar com essa questão, afinal, você sabe o que ou quanto 27 representa. Mas o que isso significa de fato?



Parte da beleza na visualização de dados é que ela pode fazer com que tudo pareça mais visceral. Uma ilustração que me deixou bastante orgulhosa e teve boa resposta foi uma em que comparei a área média de uma vaga de estacionamento com a de uma cela solitária. Esta é uma prática comum ao lidar com números em jornalismo: você não fala "banqueiros em Londres ganham X", mas sim "banqueiros em Londres ganham 7.000 vezes mais que um assistente social". Esse tipo de analogia ajuda as pessoas.

Pelo jeito, parte da sua prática tem a ver com a justaposição de diferentes elementos (como o perturbador e o familiar). Há um elemento de curadoria aí?

Em meu trabalho, o humor também tem um papel importante. Não que meu material seja engraçado, mas muitas vezes há algo de irônico no estilo. A melhor comédia consiste em dizer, basicamente, "isso aqui é uma merda". Sempre há algum tipo de comentário social. Se você consegue incluir um pouco disso no jornalismo de dados, tudo fica mais impactante.

Voltando ao exemplo de Endangered Species como um caso de transformar números em material com o qual o público se identifica através do humor e o uso de diferentes espaços visuais de comparação, você começou pela carruagem em vez do gráfico?

Primeiro desenhei a carruagem, então sete ou oito de cada espécie. Usei Photoshop para separar as camadas, colorir e contabilizar. Para me certificar de que fiz tudo certo, cada animal era uma camada diferente. Minha primeira ideia consistia em desenhar espécies ameaçadas dentro de diferentes coisas de identificação universal. O metrô de Nova York não é perfeito (é maior ou menor que o londrino?), mas serve para dar uma noção de escala. Comecei com uma planilha contendo diversas possibilidades de combinação entre animais ameaçados e espaços de fácil identificação. Pensei em colocar um tubarão dentro de uma piscina. Mas, com tantos espaços, seria difícil ter uma compreensão daquilo tudo e assim que comecei a desenhar, percebi que o processo demoraria bastante. Em vez de desenhá-los em locais diversos, todos estariam no mesmo lugar, o que funciona melhor.

Não é perfeito, claro, colocar todos os rinocerontes ali, em escala, é meio questionável (muitos teriam que ser bebês e não adultos!). Mas o resultado te faz pensar naqueles números todos. Além da transparência quanto às restrições do meio. Quando você se depara com o gráfico feito pelo *FiveThirtyEight*, como você, especialmente se for leigo, teria condições de entender até que o ponto as informações apresentadas são precisas? Aos leitores, é feito um ultimato: confie na gente ou não. Quando eles veem as ilustrações destes animais ameaçados, podem muito bem olhar para aqueles rinocerontes e pensarem "está meio esquisito, mas entendi". Aos leitores é dado acesso à crítica de uma forma que não ocorre com gráficos gerados por computador.

Antes você comentou que gostaria que seu trabalho democratizasse a forma como pessoas interagem com dados. Poderia falar mais sobre isso?

Sem que os leitores possam participar da interpretação dos dados e desenvolverem suas próprias opiniões, qual a diferença entre jornalistas e políticos? Temos jornais de esquerda e direita dizendo "ou você confia na gente ou não". Nós deveríamos estar dando autonomia às pessoas para que possam tomar decisões bem-informadas sobre suas vidas. Não é só uma questão de dizer "eis os fatos e agora você claramente tem que fazer isso aqui". Se trata de falar o seguinte: "eis os fatos, e foi assim que chegamos a esse ponto". Não é uma

questão só de jornalismo; acho que há muito a ser feito no campo da medicina também. Gostaria de trabalhar mais em cima de mudanças quanto a embalagens de medicamentos, por exemplo. No lugar de caixas com dizeres como "você precisa fazer isso", um bom médico deveria ser capaz de dizer ao paciente "eis os riscos desta medicação, os riscos de não tomá-la, os riscos de seguir outro tratamento, os riscos de não segui-lo", para que as pessoas possam tomar as decisões por conta própria, considerando que ninguém é igual.

Acredito que uma boa visualização de dados deve comunicar incertezas.[107] Incertezas são integrais àquela conversa toda sobre tomar decisões bem-informadas em sua vida. Poucos jornalistas de dados gastam tempo comunicando estas incertezas. Poucos jornalistas de dados gastam tempo tentando se aproximar de comunidades compostas por não geeks. Só porque você não tem esse repertório de habilidades estatísticas ou computacionais significa que você não é inteligente? Que você não merece entender estas informações? Claro que não. Mesmo assim, alguns jornalistas referem-se a termos como se não fossem nada demais, em uma atitude do tipo "não vou explicar isso toda vez, ou você entende ou não". É idiota. Minha abordagem dentro do jornalismo de dados se baseia na ideia de que você não precisa de repertório ou conhecimento específicos para ser inteligente.

Há também um elemento de participação popular em decidir o que importa?

Um dos motivos pelos quais comecei a coluna *Dear Mona* foi para que as pessoas pudessem me enviar suas perguntas. Recebo mensagens constantemente no Instagram de pessoas falando sobre coisas que lhes importam, muitas delas que eu não teria necessariamente pensado sobre. Há alguns caminhos que não gostaria de percorrer, como a relação entre saúde mental e controle de armas de fogo, que pode estigmatizar pessoas com problemas mentais e levar a um buraco fundo demais. Mas se eu recebesse muitas mensagens de gente querendo saber mais sobre o assunto, então seria hora de refletir se não daria para deixar de lado a nuance por conta de ser algo tão complicado ou, ao invés disso, ir para cima com tudo. Sendo assim, sempre busco saber o que importa para os leitores. Não acho que se trate de uma abdicação da responsabilidade jornalística. É parte do papel democrático do jornalismo, e as pessoas percebem que têm voz no produto final, do começo ao fim, na criação e na compreensão — não foi algo que foi feito e entregue para eles desse jeito meio "ame-o ou deixe-o".

Poderia falar um pouco mais sobre as reações ao seu trabalho? Houve reações inesperadas ou dignas de nota?

---

[107] Consultar também o capítulo de C. W. Anderson presente neste volume e seu livro *Apostles of Certainty: Data Journalism and the Politics of Doubt* (2018).

Vejo as mais variadas reações ao meu trabalho. Algumas pessoas focam no tema abordado. Então, sempre que faço algo voltado a diferenças salariais, por exemplo, um monte de homens brancos chega dizendo coisas como "mulheres negras ganham menos porque trabalham menos" e você tem que explicar a eles como as ilustrações são baseadas em comparações de igual para igual entre trabalhadores de tempo integral, e se há variação entre os níveis de seus cargos (gerente sênior, por exemplo), isso também é parte da problemática. Sempre estou disposta a avaliar a crítica em primeiro lugar.

Mas, no geral, recebo mais apoio. Por vezes, as pessoas respondem a estas críticas nos comentários antes mesmo de eu vê-las. Pessoas cujas vidas estão representadas nas ilustrações acabam por intervir, falando coisas como "minha experiência pessoal reflete isso". Há ocasiões em que o público pede por mais dados. Muitos estudantes escrevem dizendo que gostariam de fazer aquilo (o interessante é que recebo mais mensagens de estudantes do sexo feminino do que masculino). Muitas ONGs e instituições de caridade entram em contato pois buscam sentir algo a respeito dos seus dados e não pensar sobre estas informações — algo que meu trabalho consegue fazer às vezes. Um de meus artigos foi citado em um projeto de lei dos EUA.

Meu trabalho foi visto e compartilhado por muita gente em redes sociais, pessoas que não necessariamente se interessam por jornalismo de dados por si só, o que leva a prática para um novo público. Bernie Sanders compartilhou minha ilustração sobre violência armada, Miley Cyrus compartilhou outra, assim como a modelo Iman, e Shaun King, ativista de direitos civis. São pessoas que não conheço e que não necessariamente acompanham meu trabalho, mas que veem outros compartilhando o que faço e acaba entrando no seu radar. É incrível ver gente interagindo assim. Assim que uma pessoa proeminente compartilha o material, ele pode acabar ganhando vida própria.

Alguns dos trabalhos citados neste capítulo podem ser encontrados no *sitemonachalabi.com* e no Instagram, no perfil @monachalabi.

Mona Chalabi vem tentando deixar números mais emocionantes, acumulando um monte de informação no processo.

2. A web como meio de visualização de dados

Elliot Bentley

Nem toda mídia é igual. Uma série de televisão com 20 episódios é capaz de contar uma história de uma forma diferente quando comparada a um longa-metragem de duas horas, por exemplo. Da mesma forma, uma humilde página na internet oferece possibilidades únicas para visualização de dados.

A web foi projetada para lidar com documentos hiperlinkados simples, constituídos em grande parte por texto e imagens estáticas. JavaScript, novas funcionalidades e ferramentas expandiram o arsenal do que pode ser feito.[108]

Ainda que teorias e técnicas clássicas de visualização de dados (caso de Tufte, Bertin) continuem aplicáveis a gráficos na rede, as funcionalidades únicas da web oferecem vasto potencial para novas formas de jornalismo de dados. Estes trabalhos muitas vezes são chamados de "interativos", uma palavra esquisita que acaba por nublar alguns dos pontos fortes intrínsecos à web.

Abaixo, uma lista com alguns exemplos de como gráficos na internet podem se aproveitar do meio que os abriga.

Conjuntos de dados enormes e exploráveis

Um uso clássico da interatividade consiste em apresentar ao leitor um conjunto de dados gigantesco e permitir que ele 'mergulhe' e explore o quanto quiser. Às vezes, uma enorme tabela ou um grande mapa interativo.

Este formato muitas vezes é subestimado atualmente, ao jogar no leitor a responsabilidade de encontrar o que interessa por conta própria, mas pode valer a pena se os dados são bons o bastante. Em minha experiência, as versões mais bem-sucedidas lidam bem com o fato de que se tratam de ferramentas e nada mais (ao invés de serem artigos), como os apps extremamente populares de serviço público com informações sobre rankings de universidades da *ProPublica.*[109]

---

[108] Outras plataformas "multimídia", em especial o Flash, adicionaram grande riqueza de opções antes da web aberta. Por bem ou mal, estas tecnologias caíram em desuso. E por mais que tais funcionalidades, até mais que estas, estejam disponíveis em aplicações nativas, a web é muito mais fácil de se trabalhar e sua distribuição é praticamente gratuita.

[109] https://www.propublica.org/newsapps/.

Um guia para o leitor ao longo de gráficos complexos

Bastante comum atualmente é o seguinte formato: inicia com um único gráfico que, então, é manipulado por zooms, viagens ao longo do tempo, dados trocando de lugar, meios para explorar aquela série de dados inteiramente. Tudo isso funciona muito bem quando pareado com scrollytelling e é especialmente valoroso quando falamos de dispositivos móveis, que podem não ter espaço de tela o suficiente para exibir todos os dados de uma vez.[110]

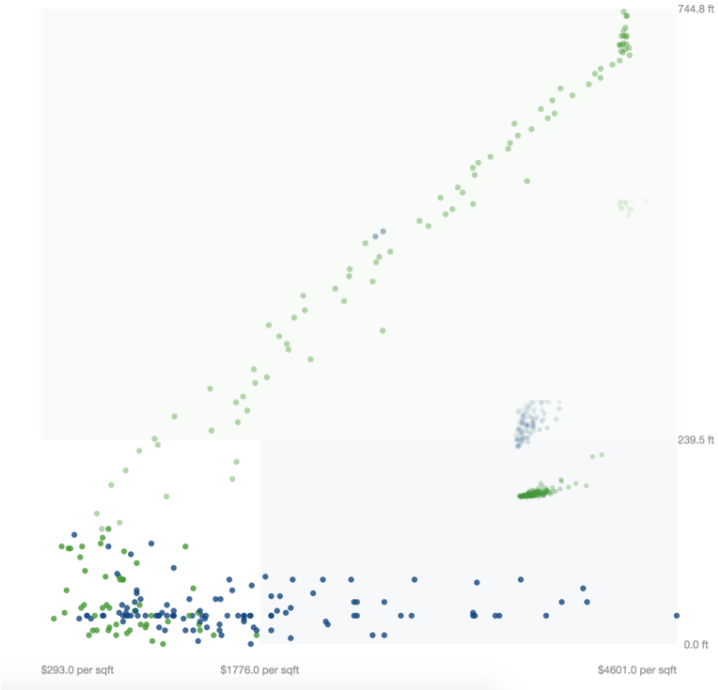


Figura 1: *A visual introduction to machine learning*. Fonte: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/.

Na agora clássica matéria *A visual introduction to machine learning* (acima), os mesmos dados transitam entre múltiplos formatos, ajudando os leitores a acompanharem como os algoritmos de aprendizagem de máquina estão organizando-os.[111] Outro bom exemplo é o trabalho da Vox em *100 years of tax brackets, in one chart*, que aproxima e afasta o zoom em um conjunto de dados que seria intimidador caso fosse apresentado de outro jeito.[112]

---

110 https://pudding.cool/process/how-to-implement-scrollytelling/.

111 http://www.r2d3.us/visual-intro-to-machine-learning-part-1/.

112 https://www.vox.com/2015/10/26/9469793/tax-brackets.

Dados atualizados em tempo real ou quase

Por que escolher um conjunto de dados estático quando pode ter os números atualizados sobre seja lá o que for? Eleições, esportes, clima e finanças são boas fontes de dados em tempo real, interessantes o bastante para serem exibidos desta forma. Mais legal ainda é fornecer o contexto destas informações de formas inusitadas, por exemplo, ao mostrar quais países são beneficiados pelo preço atual do petróleo (abaixo).[113]

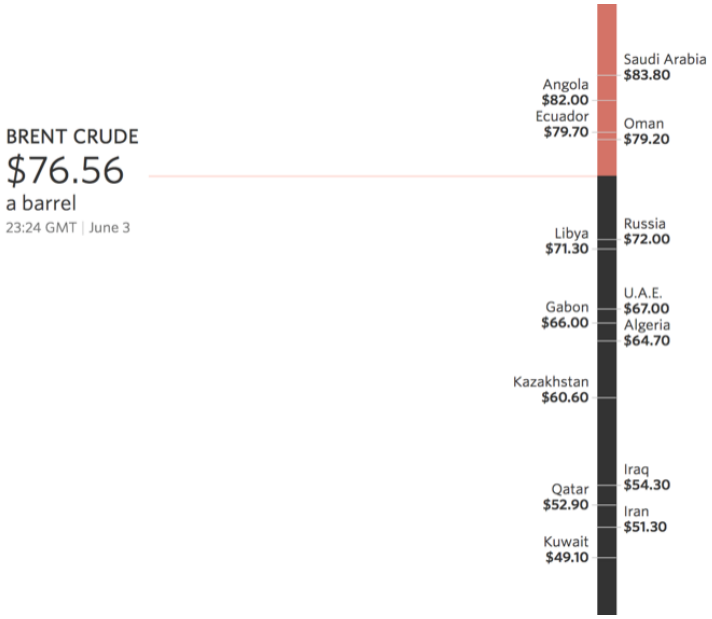


Figura 2: Países beneficiados pelo preço atual do petróleo. Fonte: http://graphics.wsj.com/oil-producers-break-even-prices/.

Ampp3d, um veículo experimental de curta existência voltado ao jornalismo pop baseado em dados, usava contadores em tempo real para dar vida a números de maneiras únicas, como o número de imigrantes entrando no Reino Unido e o faturamento do jogador de futebol Wayne Rooney.[114] Infelizmente, estes projetos já não estão mais disponíveis para visualização.

Colocar o leitor dentro do conjunto de dados

Outra maneira de lidar com a ideia de grandes conjuntos de dados, que acredito ser bastante atraente para os leitores, é mostrar ao público onde ele se encaixa dentro dos dados, geralmente ao pedir por algumas informações pessoais. Publicado em 2013, o questionário

---

[113] http://graphics.wsj.com/oil-producers-break-even-prices/.

[114] https://onlinejournalismblog.com/2015/05/13/the-legacy-of-ampp3d-usvsth3m-and-row-zed/.

sobre dialetos em forma de mapa do *The New York Times* (abaixo) foi o artigo mais popular do veículo naquele ano — mesmo com sua publicação no dia 20 de dezembro.[115]

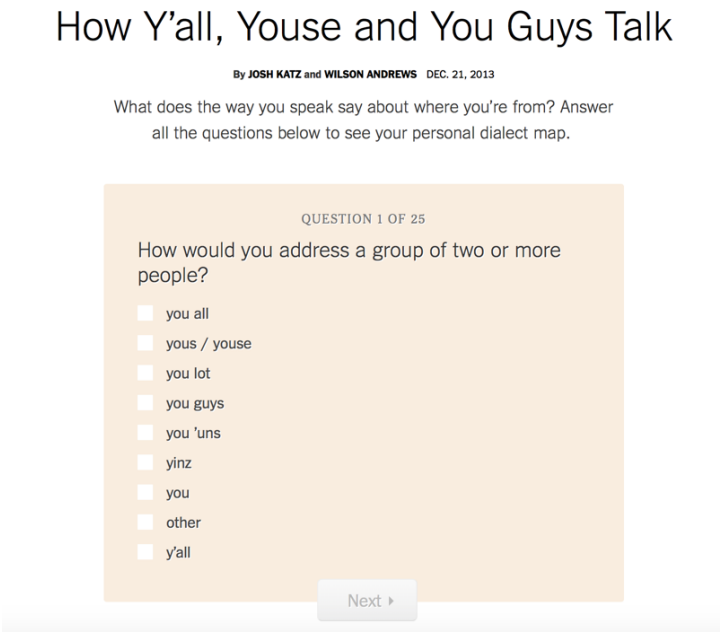


Figura 3: Questionário sobre dialetos do *The New York Times*, 2013. Fonte: https://www.nytimes.com/interactive/2014/upshot/dialect-quiz-map.html.

A *BBC* parece sempre estar fazendo coisas do tipo, muitas vezes como uma ferramenta de serviços públicos, caso da "UK fat scale calculator".[116] Gosto também de um projeto publicado pela *Quartz* sobre como pessoas de diferentes culturas desenham círculos, que começa pedindo ao leitor para desenhar um círculo, uma introdução atrativa para o que poderia ter sido um artigo chato.[117]

Coleta de conjuntos de dados originais

Outros projetos vão um passo além da categoria anterior ao usar os dados enviados por leitores não somente para dar uma resposta imediata, mas também compilar uma nova série de informações para análise posterior.

---

115 https://www.nytimes.com/interactive/2014/upshot/dialect-quiz-map.html, https://knightlab.northwestern.edu/2014/01/20/behind-the-dialect-map-interactive-how-an-intern-created-the-new-york-times-most-popular-piece-of-content-in-2013/.

116 https://www.bbc.com/news/health-43697948.

117 https://qz.com/994486/the-way-you-draw-circles-says-a-lot-about-you/.

A Australian Broadcasting Corporation trabalhou junto a cientistas políticos no 'Vote Compass' para ajudar os leitores a entenderem seu lugar no panorama político — e publicou uma série de artigos com base nestes dados.[118]

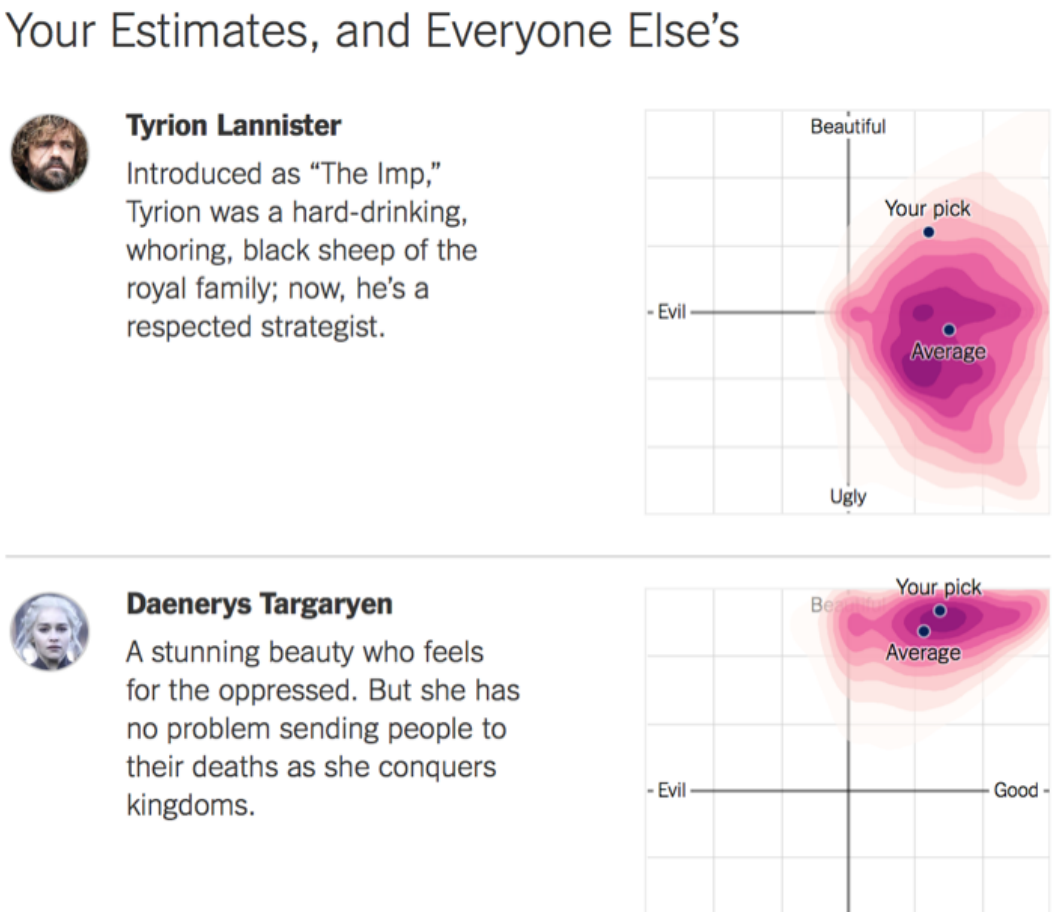


Figura 4: Gráfico de classificação de personagens em *Game of Thrones*. Fonte: https://www.nytimes.com/interactive/2017/08/09/upshot/game-of-thrones-chart.html.

Recentemente, o *The New York Times* usou esta mesma ideia para tratar de um assunto mais leve, pedindo para que os leitores classificassem os personagens de *Game of Thrones*, com os resultados publicados em gráficos (Figura 4).[119]

Tela infinita

A web é infinita em escopo e capacidade, mas, mais especificamente, páginas podem ter a altura ou largura que bem entenderem, uma espécie de "tela infinita" para se trabalhar. Peguei este termo do artista Scott McCloud, que argumenta "não haver motivo para

---

118 https://www.abc.net.au/news/nsw-election-2015/vote-compass/, https://www.abc.net.au/news/nsw-election-2015/vote-compass/results/.

119 https://www.nytimes.com/interactive/2017/08/09/upshot/game-of-thrones-chart.html.

quadrinhos longos serem divididos em páginas quando se publica na internet".[120] E, de fato, por que nossos gráficos deveriam se restringir aos limites do papel também?

No artigo *The depth of the problem*, do *The Washington Post*, um gráfico de 16.000 pixels de altura é usado para ilustrar a profundeza da área do oceano investigada em busca do voo MH370 (trecho abaixo).[121] Claro que esta informação poderia ter sido enfiada em uma única tela, mas faltariam detalhes e impacto emocional, que acompanham este gráfico extremamente alto.

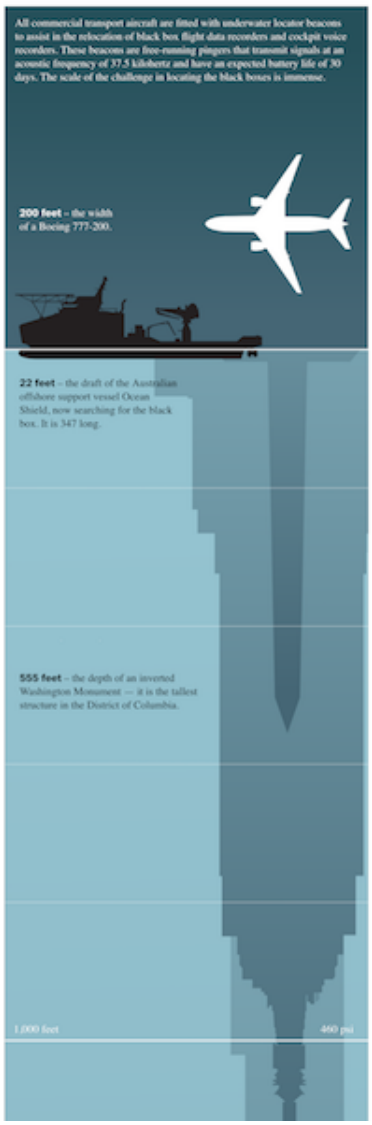


Figura 5: Gráfico ilustrando a profundeza da área no oceano das buscas do voo MH370. Fonte: http://apps.washingtonpost.com/g/page/world/the-depth-of-the-problem/931/.

120 http://scottmccloud.com/4-inventions/canvas/index.html.

121 http://apps.washingtonpost.com/g/page/world/the-depth-of-the-problem/931/.

No artigo *How the List tallies Europe's migrant bodycount,* publicado no *The Guardian*, dezenas de milhares de mortes de imigrantes são representadas, de forma poderosa, como pontinhos surgindo um a um enquanto o leitor rola a página.[122]

Jogos baseados em dados

'Newsgames', experiências interativas que tomam para si mecânicas de videogames para explorar notícias, já existem há algum tempo, com graus variados de sucesso.

A série *You Draw It*, do *The Upshot* (abaixo), desafia as premissas dos leitores ao pedir que preencham um gráfico em branco, antes de revelar a resposta correta e partir para uma exploração mais aprofundada do assunto.[123]

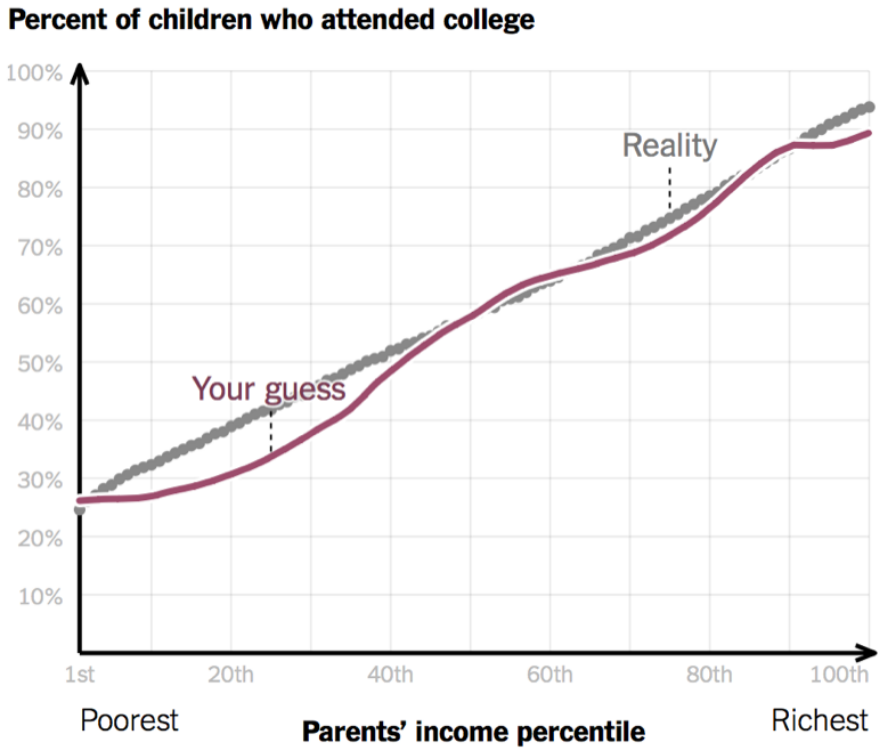


Figura 6: Gráfico da série 'You Draw It' do *The Upshot.* Fonte: https://www.nytimes.com/interactive/2015/05/28/upshot/you-draw-it-how-family-income-affects-childrens-college-chances.html.

122 https://www.theguardian.com/world/2018/jun/20/the-list-europe-migrant-bodycount.

123 https://www.nytimes.com/interactive/2015/05/28/upshot/you-draw-it-how-family-income-affects-childrens-college-chances.html.

Alguns jogos são mais elaborados e pedem ao leitor para resolver uma versão simplificada de um problema do mundo real — como financiar a operação da *BBC*, por exemplo — para provar quão complicado ele realmente é.[124]

Poderíamos considerar estes artifícios meros brinquedos que oferecem aos leitores informação superficial, mas, feitos do jeito certo, podem oferecer uma perspectiva nova a assuntos batidos. 'How To Win A Trade War', do *FiveThirtyEight*, é um jogo em que o leitor escolhe uma estratégia comercial e compete contra outro visitante daquela página, dando vida à possivelmente enfadonha cobertura de teoria econômica[125]

Experimentos randomizados em tempo real

Um formato relacionado a isso é permitir ao leitor que rode uma simulação em tempo real no seu próprio navegador. Mais que uma simples explicação animada, a técnica adiciona um grau de aleatoriedade capaz de levar a resultados únicos a cada vez que é empregada, excelente para demonstrar probabilidades estatísticas.

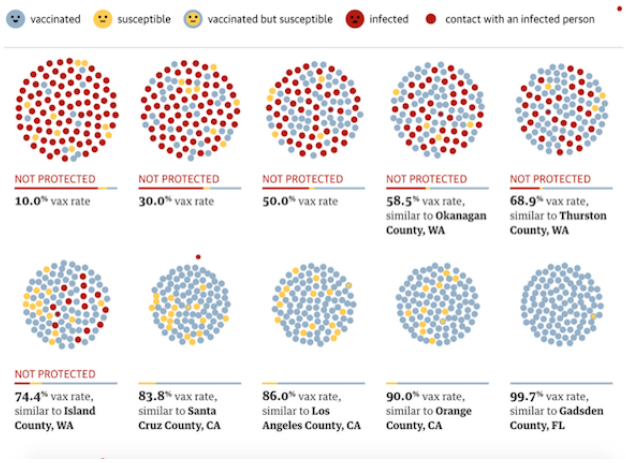


Figura 7: Simulação de um surto de sarampo em dez populações únicas com índices variados de vacinação.

O material do *The Guardian* citado acima simula um surto de sarampo em dez populações únicas com índices variados de vacinação.[126] Ao utilizar gráficos web, ficam claros os resultados de uma forma que não seria possível com uso apenas de porcentagens. Em 'Years You Have Left to Live, Probably', de Nathan Yau, um gráfico simples de linha

---

124 https://ig.ft.com/sites/2015/bbc/.

125 https://fivethirtyeight.com/features/how-to-win-a-trade-war/.

126 https://www.theguardian.com/society/ng-interactive/2015/feb/05/-sp-watch-how-measles-outbreak-spreads-when-kids-get-vaccinated.

(indicando a probabilidade de viver até o próximo ano) ganha mais força com 'vidas' que se perdem aleatoriamente e vão se acumulando.[127]

Não é necessário usar dados fictícios para esse tipo de simulação. 'The Birthday Paradox', por exemplo, verifica a probabilidade de aniversários compartilhados com base em quem visitou a página anteriormente.[128]

3D, realidade virtual e realidade aumentada

Gráficos tridimensionais e realidades virtuais são ferramentas difíceis de serem utilizadas a serviço do jornalismo de dados para além da visualização de mapas.

Dois experimentos notáveis, ambos de 2015, voltados a dados financeiros ('Is the Nasdaq in Another Bubble?' e 'A 3-D View of a Chart That Predicts The Economic Future: The Yield Curve') se mostram novidades interessantes, mas que não chegaram a gerar uma onda de gráficos em três dimensões.[129] Talvez seja melhor assim.

Já o potencial da realidade aumentada, em que a imagem de uma câmera do mundo real é sobreposta por gráficos, precisa ser posto à prova.

Conclusão: como surgem os novos formatos

Alguns dos gráficos web citados acima representam novos formatos cuja emergência se deu ao longo dos últimos anos. Alguns chegaram para ficar, caso do guia de gráficos complexos (geralmente por meio de interação em 'scrollytelling'). Outros, como os gráficos tridimensionais, foram fogo de palha.

Ainda assim, não são apenas gostos pessoais que determinam quais tipos de gráficos estão em alta na rede, há de se considerar a tecnologia disponível e os hábitos de consumo de leitores como influenciadores de tendências.

Peguemos o exemplo do mapa interativo, amplamente utilizado. Além de ser um formato já consagrado, visualmente atrativo e de fácil compreensão, seu amplo uso e sua familiaridade certamente receberam um empurrão de ferramentas que facilitam sua criação e manipulação, citando aqui Google Maps e Leaflet como as mais comuns.

---

127 https://flowingdata.com/2015/09/23/years-you-have-left-to-live-probably/.

128 https://pudding.cool/2018/04/birthday-paradox/.

129 http://graphics.wsj.com/3d-nasdaq/, https://www.nytimes.com/interactive/2015/03/19/upshot/3d-yield-curve-economic-growth.html.

Sem muitos dados relevantes a serem apresentados, a impressão é de que menos mapas interativos são publicados atualmente. Ainda que fosse fácil atribuir esta tendência a uma percepção crescente entre jornalistas de que tal interatividade (ou o mapa em si) pode ser supérflua, é possível que novas tecnologias também tenham colaborado com esta queda.

Uma grande proporção dos leitores acessa a internet pelo celular, e mapas interativos oferecem uma experiência péssima em telas de toque pequenas. Além disso, há uma nova solução tecnológica superior em diversos aspectos: ai2html, um script de programação de código aberto do *The New York Times* que gera um snippet responsivo em HTML a partir de arquivos do Adobe Illustrator.[130] Mapas criados com ai2html podem se aproveitar das habilidades de um cartógrafo tradicional e, ainda assim, apresentar textos afiados, legíveis por máquinas. A falta de interatividade nestes mapas acaba por ser uma bênção, mesmo que ofereça suas limitações.

Este é só mais um exemplo de como jornalistas de dados precisam pensar bem a respeito do uso de funcionalidades únicas da web. Com tantas possibilidades, é importante avaliá-las e usá-las somente quando há necessidade real.

Elliot Bentley trabalha no departamento de gráficos do Wall Street Journal desde 2014, além de ter criado o app de transcrição de código aberto Transcribe.

130 https://github.com/newsdev/ai2html.

Quatro desdobramentos recentes em gráficos jornalísticos

Gregor Aisch e Lisa Charlotte Rost

O campo de gráficos jornalísticos ainda é jovem e tenta responder perguntas como: como representar vieses e incertezas em dados (de pesquisas)?[131] Como trabalhar junto aos repórteres? Como comunicar dados complexos em redes sociais de ritmo ágil?[132] Aqui, tentamos cobrir quatro desdobramentos que consideramos fundamentais para os próximos anos.

O conceito 'mobile em primeiro lugar' começa a ser levado a sério

A ideia de 'mobile em primeiro lugar' é amplamente utilizada, mas no mundo veloz dos gráficos de notícias, a experiência com dispositivos móveis continuou em segundo lugar. Só agora estamos vendo estes dispositivos móveis avançarem na lista de prioridades. Fato que, por sua vez, traz duas consequências.

Primeiro, se pensa mais em como fazer gráficos funcionarem nestes dispositivos. Um aviso do tipo "esta experiência funciona melhor no desktop" acaba por se tornar uma gafe. Gráficos precisam ser responsivos, não afastar metade dos usuários. Mas pensar dentro da caixinha mobile e seus tão poucos pixels pode ser bem frustrante para jornalistas gráficos, muitos acostumados ao "luxo" de poder preencher páginas inteiras no impresso e telas em experiências voltadas ao desktop. Sob as melhores circunstâncias, estas limitações motivam os jornalistas a irem além e serem mais criativos. Já vemos isso acontecendo: o *Financial Times*, por exemplo, virou em 90 graus seu gráfico a respeito do Parlamento, essencialmente criando um novo tipo de gráfico.[133]

A segunda consequência da visualização de dados que coloca dispositivos móveis em primeiro lugar é que desenvolvedores e jornalistas encararão não só aparelhos com telas pequenas, mas também dispositivos cheios de sensores. Como consequência, novas experiências com dados podem surgir. O *Guardian* criou um aplicativo em que é possível fazer um tour virtual em áudio pelo Rio de Janeiro com a mesma extensão da maratona ocorrida lá em 2016.[134] "Nosso desafio para você: completar todas as 26,2 milhas — ou 42,2 km — da rota ao longo de três semanas", escreveu. Realidades aumentada e virtual fazem uso semelhante de nossos smartphones, e vemos essas técnicas chegando ao jornalismo também.

---

131 https://www.nytimes.com/interactive/2019/08/29/opinion/hurricane-dorian-forecast-map.html.

132 https://medium.com/severe-contest/lessons-for-showcasing-data-journalism-on-social-media-17e6ed03a868.

133 https://ig.ft.com/italy-poll-tracker/.

134 https://www.theguardian.com/sport/2016/aug/06/rio-running-app-marathon-course-riorun.

Está morta a interatividade, exceto quando dá sinais de vida

Cada vez menos vemos interatividade em gráficos simples, uma tendência intensificada ao longo dos últimos anos. Agora, interação é um recurso reservado para grandes projetos desenvolvidos anualmente nas redações. Cabe notar que interatividade já não é determinante para o sucesso. Redações de veículos como *Financial Times*, *FiveThirtyEight* e *National Geographic* vêm publicando gráficos que viralizaram repetidas vezes mesmo sem permitir interação dos usuários.

Acreditamos haver duas razões principais para o declínio dos gráficos interativos. Para começo de conversa, menos pessoas do que o esperado interagem com gráficos.[135] Gente curiosa, experiente com internet — caso dos jornalistas gráficos —, sempre tentará passar o mouse ou o que for sobre uma visualização em busca de mais detalhes. Os jornalistas, por sua vez, querem dar mais vida aos seus artigos. Mas estamos criando para um público que prefere o consumo passivo, especialmente em dispositivos móveis. Grande parte das pessoas não acessará conteúdo escondido em camadas interativas, o que fez muitos jornalistas optarem por não esconder nada.

Fora isso, os gráficos agora fazem parte do ciclo de notícias urgentes. Jornalistas envolvidos neste processo foram ficando cada vez mais ágeis na criação destas visualizações e uma notícia urgente logo conta com mapas demarcando onde algum evento ocorreu, por exemplo. Interatividade bem-feita demanda tempo, porém. Muitas vezes, é deixada de lado para entrar no artigo em algum outro momento.

Ainda vemos gráficos interativos em notícias, mas sua relevância mudou. No lugar de adicionar algo a uma história, a interatividade torna-se aquela história. Já vimos grandes exemplos de materiais exploráveis em que os leitores podem inserir dados pessoais como localização, renda ou mesmo opinião e ver como estão inseridos dentro de um contexto mais amplo. Caso de publicações como *You Draw It: How Family Income Predicts Children's College Chances* e *Is It Better to Rent or Buy*, publicados no *The New York Times.*[136] Ambos não adiantam de nada aos leitores caso não insiram suas informações: só se gera valor *através* da interação.

---

135 https://vimeo.com/182590214. https://medium.com/@dominikus/the-end-of-interactive-visualizations-52c585dcafcb.

136 https://www.nytimes.com/interactive/2015/05/28/upshot/you-draw-it-how-family-income-affects-childrens-college-chances.html, https://www.nytimes.com/interactive/2014/upshot/buy-rent-calculator.html.

Redações passam a usar mais ferramentas gráficas (internas)

Mais do que nunca, jornalistas são pressionados para que seus artigos se destaquem. Adicionar gráficos a uma reportagem pode ser uma solução para isso, mas a equipe responsável não consegue dar conta de todos os pedidos que chegam. Por isso, cada vez mais vemos redações usarem ferramentas que facilitam a criação de gráficos, mapas e tabelas em apenas alguns poucos cliques. Redações têm duas opções quando se fala de ferramentas para criação de gráficos:

- soluções externas como Datawrapper e Infogram;
- desenvolver internamente um programa ajustado às necessidades da redação, integrado ao seu sistema de gerenciamento de conteúdo.

Por mais que a segunda alternativa soe excelente, há de se encarar que é algo que demanda mais recursos que o esperado para muitas redações. Já as ferramentas externas são criadas por equipes dedicadas que oferecem suporte e treinamento. Dentro da redação, isso geralmente será feito pela equipe de dados ou interação, o que toma seu tempo para lidar com projetos noticiosos de fato. O desenvolvimento de uma solução própria só tem como dar certo se for considerada uma prioridade: há uma equipe de três desenvolvedores dedicados exclusivamente à criação e manutenção da ferramenta Q, utilizada pelo *Neue Zürcher Zeitung*.

Publicações baseadas em dados fomentam a inovação e o letramento visual

Há alguns anos, uma abordagem de dados era considerada útil apenas em algumas ocasiões; hoje temos publicações (bem-sucedidas!) inteiramente dedicadas a esta ideia. Geralmente, estes sites usam dados como forma de comunicar temas específicos de uma publicação, caso do *FiveThirtyEight* em sua cobertura de política e esportes, *The Pudding* com cultura pop e *Our World in Data* em sua tarefa de falar sobre o desenvolvimento da humanidade a longo prazo. Talvez a principal diferença entre estas publicações e outras sobre os mesmos temas é o público: curioso e interessado em dados, um público que não tem medo de gráficos e tabelas. Por conseguinte, publicações voltadas a dados podem mostrar aos seus leitores materiais mais complicados de serem interpretados, como gráficos de dispersão interligados. Com a aplicação correta, estes recursos oferecem uma visão mais complexa e menos agregada do mundo, traçando comparações de maneira que um simples gráfico de barras não faria.

Gregor Aisch foi editor de gráficos do The New York Times, cofundador e atual chefe de tecnologia do Datawrapper. Lisa Charlotte Rost é designer e já criou visualizações para diversas redações (SPIEGEL, NPR, Bloomberg, ZEIT Online) antes de integrar a equipe do Datawrapper.

Bancos de dados pesquisáveis enquanto produto jornalístico

Zara Rahman e Stefan Wehrmeyer

Um formato jornalístico que aos poucos vem surgindo é o banco de dados pesquisável online, tipo de interface web que dá acesso a uma série de informações, disponibilizado por redações. Não é novo, mas sua presença em projetos de jornalismo de dados ainda é relativamente escassa.[137]

Neste artigo, discutiremos vários tipos de bancos de dados, daqueles que cobrem assuntos diretamente ligados à vida dos leitores até interfaces criadas a serviço de trabalho investigativo mais aprofundado. Nosso trabalho se baseia no envolvimento de um dos coautores no projeto "Euros für Ärzte" (Euros para Médicos), descrito abaixo como estudo de caso ilustrativo.[138] Cabe mencionar que, por mais que o compartilhamento de dados puros tenha se tornado uma boa prática a ser seguida após a conclusão de uma investigação, criar uma interface pesquisável de dados é bem menos comum.

Levamos em consideração os recursos específicos envolvidos na criação de bancos de dados em jornalismo, mas também destacamos como a prática dá vazão a uma série de problemáticas relacionadas à ética e à privacidade no tocante a como dados são usados, acessados, modificados e interpretados. Então, examinamos que reflexões acerca do uso responsável de dados surgem como consequência do uso das informações desta forma, levando em conta as dinâmicas inerentes de poder, bem como as consequências de disponibilizar tais informações online. Concluímos ao fornecer uma lista de melhores práticas que, certamente, evoluirão no futuro.

Exemplos de bancos de dados jornalísticos

Bancos de dados podem integrar a face pública do jornalismo investigativo de várias formas.

Um exemplo com fortes elementos de personalização é o "Dollars for Docs", da *ProPublica*, que compilou dados de pagamentos a médicos e hospitais de ensino feitos por

---

137 http://www.holovaty.com/writing/fundamental-change/.

138 https://correctiv.org/recherchen/euros-fuer-aerzte/.

empresas farmacêuticas e fabricantes de dispositivos médicos.[139] Estes tema e abordagem foram tomados como base pelo *Correctiv* e pelo *Spiegel Online* na criação do "Euros für Ärzte", através do desenvolvimento de um banco de dados pesquisável de receptores de pagamentos de empresas farmacêuticas, como explicado abaixo em maiores detalhes. Ambos os projetos envolveram a compilação de dados já disponíveis em outras fontes, com o objetivo de aumentar a acessibilidade destas informações de forma que leitores pudessem pesquisá-las por conta própria e, presumidamente, checar se seus médicos haviam recebido tais pagamentos. Tudo acompanhado de reportagens e investigações.

Seguindo o mesmo caminho, o *Berliner Morgenpost* criou o "Schul Finder", para ajudar pais a encontrarem escolas em sua região. Neste caso, a interface é o produto final.[140]

Em contraste com este tipo de banco de dados em que os dados são coletados e preparados pela redação, há outra prática em que os leitores contribuem com os dados, sendo estes conhecidos como dados 'gerados por cidadãos', ou colaborativos. Isso se mostra particularmente eficaz quando os dados desejados não são coletados por meio de fontes oficiais, caso do banco de dados colaborativo The Counted, do *The Guardian*, que reunia informações sobre pessoas mortas pela polícia nos EUA nos anos de 2015 e 2016.[141] O banco de dados em questão baseava-se em diversas reportagens online, bem como contribuições de leitores.

Qualquer outro tipo de banco de dados envolve uma série de informações já existentes e a criação de uma interface em que o leitor possa gerar um relatório baseado em um conjunto de critérios definidos pelos mesmos — os Nauru Files, por exemplo, permitem aos leitores visualizarem um resumo das reportagens sobre os incidentes ocorridos na prisão de Nauru entre 2013 e 2015, assinados pela equipe na Austrália. O *Bureau de Jornalismo Investigativo*, com base no Reino Unido, compila dados de várias fontes reunidas ao longo de suas investigações, dentro de um banco de dados conhecido como Drone Warfare.[142] Este banco de dados possibilita aos leitores escolherem os países nos quais o *Bureau* atua e o período de tempo desejado, gerando um relatório com visualizações que resume os dados.

Por fim, bancos de dados também podem ser criados a serviço de mais jornalismo, como ferramenta para auxiliar pesquisas. O Consórcio Internacional de Jornalistas Investigativos criou e mantem o Offshore Leaks Database, baseado em dados obtidos junto

---

139 https://projects.propublica.org/docdollars/.

140 https://interaktiv.morgenpost.de/schul-finder-berlin/#/.

141 https://www.theguardian.com/us-news/ng-interactive/2015/jun/01/the-counted-police-killings-us-database.

142 https://www.thebureauinvestigates.com/projects/drone-war/.

aos *Panama Papers*, *Paradise Papers* e outras investigações.[143] De maneira semelhante, a OCCRP mantém e atualiza o OCCRP Data, que dá aos usuários a oportunidade de pesquisar mais de 19 milhões de registros públicos.[144] Em ambos estes exemplos, considera-se jornalistas e pesquisadores como principais usuários destas plataformas, não leitores, e, sim, profissionais em busca de mais informações através do uso destas ferramentas.

Abaixo, algumas considerações a respeito da criação de bancos de dados enquanto produto jornalístico:

- **Público:** voltado diretamente a leitores ou como fonte de pesquisa para outros jornalistas
- **Temporalidade:** atualizado com frequência ou como publicação única
- **Contexto:** parte de um artigo ou investigação ou tendo o banco de dados em si como produto principal
- **Interatividade:** leitores são encorajados a contribuir para melhoramento do banco de dados ou leitores são considerados como visualizadores das informações apresentadas.
- **Fontes:** baseado em dados públicos ou tornando novas informações de conhecimento público através do banco de dados

Estudo de caso: "Euros für Ärzte" (Euros para Médicos)

A Federação Europeia de Indústrias e Associações Farmacêuticas (EFPIA, na sigla em inglês) é uma associação comercial que reúne outras 33 associações nacionais e 40 empresas farmacêuticas. Em 2013, foi decidido que as empresas integrantes deveriam divulgar pagamentos realizados a profissionais e organizações de saúde nos países de sua operação a partir de julho de 2016.[145] Inspirado no projeto da *ProPublica* intitulado "Dollars for Docs", a redação alemã sem fins lucrativos *Correctiv* decidiu coletar estas informações nos sites das empresas farmacêuticas do país e criar um banco de dados centralizado pesquisável sobre quem recebeu estes pagamentos para visualização pública.[146] O projeto foi batizado como "Euros für Ärzte" ("Euros para Médicos").

Em colaboração com o veículo alemão de cobertura nacional *Spiegel Online*, documentos e dados foram coletados em 50 sites e convertidos a partir de variados formatos em dados tabulares consistentes. Posteriormente, os dados passaram por uma limpeza e foram

[143] https://offshoreleaks.icij.org/.

[144] https://data.occrp.org/.

[145] https://efpia.eu/media/25046/efpia_about_disclosure_code_march-2016.pdf.

[146] Ver 'Dollars for Docs' de Tigas et al. (2018) e Correctiv.org

encontrados diversos recebedores destes pagamentos feitos por várias empresas. O processo de limpeza destes dados levou cerca de dez dias e envolveu cinco pessoas. Uma interface personalizada de banco de dados, pesquisável, com endereços individuais por recebedor foi projetada e publicada pela *Correctiv*.[147] Em 2017, o banco de dados foi atualizado através de processo semelhante. A *Correctiv* também empregou a mesma metodologia e interface web para publicar dados da Áustria, em colaboração com o site *derstandart.at*, ORF e dados da Suíça obtidos junto ao site *Beobachter.ch*.

O objetivo jornalístico da iniciativa era destacar a influência sistêmica da indústria farmacêutica sobre profissionais de saúde, por meio de eventos, organizações e conflitos de interesse relacionados. Desejava-se que este banco de dados pesquisável encorajasse leitores a iniciarem um diálogo com seus médicos sobre o tema, além de chamar atenção para o ocorrido.

Além disso, a iniciativa destacava a inadequação de regras de declaração voluntária. Como a exigência da publicação era uma iniciativa da indústria e não uma exigência legal, o banco de dados estava incompleto, o que dificilmente mudaria de figura sem a obrigação legal da divulgação destas informações.

Como descrito acima, os dados estavam incompletos, o que significa que bastante gente que recebeu estes pagamentos de farmacêuticas não era mencionada ali. Por tabela, quando os usuários buscavam o nome de seus médicos, a falta de resultados poderia significar que o profissional em questão não recebeu nada ou não divulgou nada, duas conclusões a mundos de distância uma da outra. Críticos notaram que isso lançava os holofotes sobre indivíduos transparentes e cooperativos, o que deixava transferências de dinheiro mais escandalosas no escuro. De forma a contrabalançar isso, a *Correctiv* criou uma funcionalidade para médicos que não haviam recebido pagamentos e queriam aparecer no banco de dados, oferecendo contexto importante à narrativa, mas, ainda assim, gerando incertezas nos resultados de pesquisa.

Após a publicação, tanto a *Correctiv* quanto o *Spiegel Online* receberam dezenas de queixas e ameaças legais de médicos que constavam no banco de dados. Como os dados vinham de fontes públicas, ainda que difíceis de serem encontrados, a equipe jurídica do *Spiegel Online* optou por deferir as queixas às farmacêuticas e ajustar o banco de dados somente em caso de mudanças na fonte.

---

147 https://correctiv.org/thema/aktuelles/euros-fuer-aerzte/.

Considerações técnicas sobre a criação de bancos de dados

Para uma redação pensando em como disponibilizar e tornar acessível um conjunto de dados aos seus leitores, há diversos critérios a serem levados em consideração, como tamanho e complexidade dos dados, capacidade técnica interna da redação e como os leitores interagirão com estes dados.

Quando se opta por criar um banco de dados passível de ser um produto apropriado para uma investigação, a criação deste demanda desenvolvimento e implementação sob medida, ou seja, uma quantidade significativa de recursos. Tornar estes dados acessíveis através de um serviço externo geralmente é mais simples e demanda menos.

Por exemplo, no caso da *Correctiv*, a necessidade de busca e listagem de cerca de 20.000 receptores e suas ligações financeiras com farmacêuticas demandou software personalizado. O programa do banco de dados foi desenvolvido em repositório à parte do site principal, ainda assim, de forma que poderia ser conectado ao seu sistema de gerenciamento de conteúdo. Tal decisão foi tomada para possibilitar a integração visual e conceitual ao site principal e à seção dedicada à investigação. As informações foram armazenadas em um banco de dados relacional separado do banco de dados de conteúdo para evitar que se misturassem. Neste caso, a presença de processo e interface para ajuste de informações no banco de dados em uso era crucial, já que dezenas de correções chegaram após a publicação.

Porém, séries de dados menores e de estruturas simples podem se tornar acessíveis sem projetos de desenvolvimento de software dispendiosos. Algumas ferramentas de planilhas de terceiros (Planilhas Google, digamos) permitem o embutimento de tabelas. Há, também, diversas bibliotecas frontend em JavaScript que possibilitam a inclusão de funções em tabelas HTML como pesquisa, filtragem e categorização, muitas vezes o suficiente para que algumas centenas de colunas se tornem acessíveis aos leitores.

Um meio-termo atraente são aplicações web baseadas em JavaScript que acessam dados por meio de uma API. Funciona bem com interfaces embutíveis i-frame sem a necessidade de uma aplicação web completa. Esta mesma API pode ser operada através de serviços de terceiros, retendo o controle total do estilo do frontend.

Recursos oferecidos por bancos de dados

Bancos de dados integrados a um artigo podem oferecer diversos novos recursos para leitores e redações.

Tratando-se do público, disponibilizar um banco de dados online permite aos leitores buscarem por suas cidades, políticos ou médicos, ligando aquela matéria às suas próprias

vidas. Oferece, ainda, um novo canal para engajamento com materiais jornalísticos a nível pessoal. Considerando que há analítica envolvida nestas solicitações de pesquisa, a redação acaba por ter acesso a mais dados sobre os interesses de seus leitores, possivelmente guiando esforços futuros.

Falando da redação, se o banco de dados é considerado um investimento investigativo a longo prazo, pode ser usado para cruzar entidades com outros bancos de dados ou documentos para geração de novas pautas. Similarmente, se ou quando outras redações decidirem disponibilizar bancos de dados semelhantes, facilita-se a colaboração e a cobertura através da reutilização de infraestruturas e metodologias já existentes.

Bancos de dados também oferecem melhor otimização para motores de busca, gerando mais tráfego para o site do veículo jornalístico. Quando o banco de dados fornece endereços individuais para as entidades contidas nele, motores de busca detectarão estas páginas e as colocarão em posições altas nas buscas por termos-chave infrequentes relacionados a estas entidades, a tal "cauda longa" das pesquisas na internet, desta forma gerando mais tráfico ao site de quem o publicou.

Otimização para motores de busca pode ser uma prática não muito bem-vista dentro do jornalismo. Porém, oferecer aos leitores informações jornalísticas enquanto buscam por questões específicas também pode ser encarado como engajamento de público bem-sucedido. Por mais que o objetivo do banco de dados público não seja competir com termos-chaves de busca, certamente há um benefício no tráfego orgânico gerado, que pode atrair novos leitores.

Considerações acerca de dados e responsabilidade

Inspirados na abordagem da comunidade de dados responsáveis, que trabalha na criação de boas práticas, visando os desafios éticos e de privacidade ligados ao uso de dados de novas e diferentes formas, podemos pensar riscos em potencial de várias maneiras.[148]

Primeiro: a distribuição de poder, em que uma redação decide publicar um banco de dados com informações sobre as pessoas. Geralmente, estas pessoas não têm qualquer envolvimento nem oportunidade de corrigir ou vetar dados antes de sua publicação. O poder nas mãos destes indivíduos depende basicamente de quem são. Uma pessoa politicamente exposta inclusa neste banco de dados esperaria por este desdobramento, bem como teria recursos para agir com base neste, diferente de um profissional de saúde, que dificilmente esperaria ser envolvido em uma investigação. Assim que um banco de dados é publicado, a visibilidade dos envolvidos ali pode mudar rapidamente. Os médicos citados no projeto

[148] https://responsibledata.io/what-is-responsible-data/.

“Euros für Ärzte” nos disseram que um dos primeiros resultados quando buscavam por seus nomes era sua página neste projeto.

Dinâmicas de poder relacionadas ao leitor ou espectador também devem ser levadas em consideração. Para quem estas informações podem ser mais úteis? Estas pessoas têm acesso às ferramentas e ao conhecimento necessários para usarem este banco de dados, ou esta informação será utilizada por aqueles em posições de poder para levarem adiante seus interesses? Isso pode significar uma ampliação do escopo da testagem de usuário antes da publicação, para garantir que há contexto o suficiente para explicar o funcionamento do banco de dados ao público-alvo, ou mesmo a inclusão de certas funcionalidades que fariam da interface mais acessível a este grupo.

A suposição de que mais dados levam a decisões melhores para a sociedade foi questionada nos mais variados níveis nestes últimos anos. Clare Fontaine, estudiosa de educação, expande este ponto ao mencionar que as escolas nos EUA vêm ficando cada vez mais segregadas, apesar (ou talvez por conta) do aumento das informações disponíveis sobre ‘desempenho escolar’.[149] “Uma relação causal entre escolha de escolas e segregação galopante ainda não foi estabelecida”, comentou, mas ela e outros estudiosos estão trabalhando para melhor entender esta relação, questionando a talvez muito simplificada relação de que mais informação leva a melhores decisões, pondo em xeque possíveis significados de ‘melhor’.

Em seguida, temos o banco de dados em si. Um banco de dados contém muitas decisões humanas, como o que foi coletado e o que foi deixado de lado e como foi feita a categorização, a organização ou a análise. Dados não são objetivos, mas letramento e compreensão das limitações destes dados são pouco prevalentes, o que significa que leitores podem muito bem interpretar erroneamente as conclusões apresentadas.

Tomemos por exemplo a ausência de determinada organização dentro de um banco de dados de organizações políticas envolvidas no crime organizado, que pode não necessariamente representar a falta de envolvimento com o crime organizado em si; significa apenas que não havia dados disponíveis sobre sua atuação. Michael Golebiewski e Danah Boyd chamam este fenômeno de “vácuo de dados”, comentando ainda que, em alguns casos, este vácuo pode “refletir vieses ou preconceitos em uma sociedade, de forma passiva”.[150] Esta ausência no contexto de um espaço saturado de informações se aproxima do que a artista

---

[149] https://points.datasociety.net/driving-school-choice-16f014d8d4df.

[150] Golebiewski e Boyd (2018). Disponível em: https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf.

e pesquisadora Mimi Onuoha, do Brooklyn, chama de "série de dados faltante", e destaca as escolhas feitas pela sociedade no ato da coleta de dados.[151]

Em terceiro lugar, o direcionamento da atenção. Bancos de dados podem mudar o foco do interesse público de um problema sistêmico maior para as ações de indivíduos e vice-versa. Transações financeiras entre farmacêuticas e profissionais de saúde claramente são de interesse público, mas, a nível individual, estes médicos podem não se considerar pessoas de interesse amplo. Fato é que para apresentar um problema como sistêmico e generalizado (um padrão e não uma ocorrência isolada) são necessários dados de múltiplos indivíduos. Alguns bancos de dados, como o do "Euros für Ärzte" mencionado acima, também alteram as fronteiras de quem ou o que é de interesse público.

Mesmo quando os indivíduos concordam com a publicação de seus dados, cabe aos jornalistas a decisão de por quanto tempo estas informações serão de interesse público e quando deverão ser removidas. O Regulamento Geral sobre a Proteção de Dados da União Europeia certamente afetará a forma como jornalistas lidam com estes dados pessoais e os tipos de mecanismos disponibilizados àqueles que queiram não consentir com a inclusão de seus dados.

Com todos estes desafios, nossa abordagem consiste em ponderar como os direitos das pessoas são afetados tanto pelo processo quanto pelo resultado final de uma investigação ou produto. No âmago de tudo está o entendimento de que práticas responsáveis em dados são contínuas e não apenas uma lista de coisas a serem consideradas em pontos específicos. Sugerimos que estas abordagens devam priorizar os direitos dos indivíduos refletidos nos dados ao longo de toda a investigação, da coleta de informações à publicação, ponto central da otimização do jornalismo (de dados) em termos de confiança.[152]

Boas práticas

Para os jornalistas que cogitam criar um banco de dados para compartilhar sua investigação com o público, deixamos aqui algumas recomendações e dicas de boas práticas. Acreditamos que estas medidas evoluirão com o tempo e sugestões são bem-vindas.

Antes da publicação, desenvolva um projeto para fazer correções no banco de dados. Boas práticas ligadas à procedência de dados podem ajudar a encontrar fontes de erros.

---

[151] https://github.com/MimiOnuoha/missing-datasets#on-missing-data-sets.

[152] https://medium.com/de-correspondent/optimizing-journalism-for-trust-1c67e81c123.

Crie um canal de comunicação: quando pessoas não esperam ser mencionadas em uma investigação, certamente haverá algum retorno (ou queixa). Oferecer uma boa experiência de usuário para que essa queixa possa ser feita pode ajudar com a experiência.

Mantenha o banco de dados atualizado ou deixe claro que este não é mais atualizado: em um contexto jornalístico, publicar um banco de dados exige nível de manutenção mais alto do que um artigo. O nível de interatividade possibilitado por um banco de dados significa que há expectativas diferentes em relação a quão atualizado ele está em comparação a um artigo.

Disponha de recursos suficientes para manutenção ao longo do tempo: manter dados e software associado de banco de dados atualizados exige recursos significativos. Por exemplo, adicionar dados do ano seguinte a um banco de dados exige fundir informações novas e antigas, adicionado uma dimensão extra de tempo à interface do usuário.

Observe como os leitores estão usando o banco de dados: tendências de pesquisa ou uso podem indicar caminhos para futuros artigos e apurações.

Seja transparente: um banco de dados 100% completo é raridade e cada banco traz consigo uma série de escolhas embutidas. No lugar de tentar ocultar tais escolhas, torne-as visíveis para que os leitores saibam com o que estão lidando.

Zara Rahman é pesquisadora e escritora, cujo trabalho volta-se à intersecção entre poder, tecnologia e dados. Stefan Wehrmeyer é jornalista de dados e trabalha com tecnologia em liberdade de informação.

27. Conflitos sobre água narrados através de dados e quadrinhos interativos

Nelly Luna Amacio

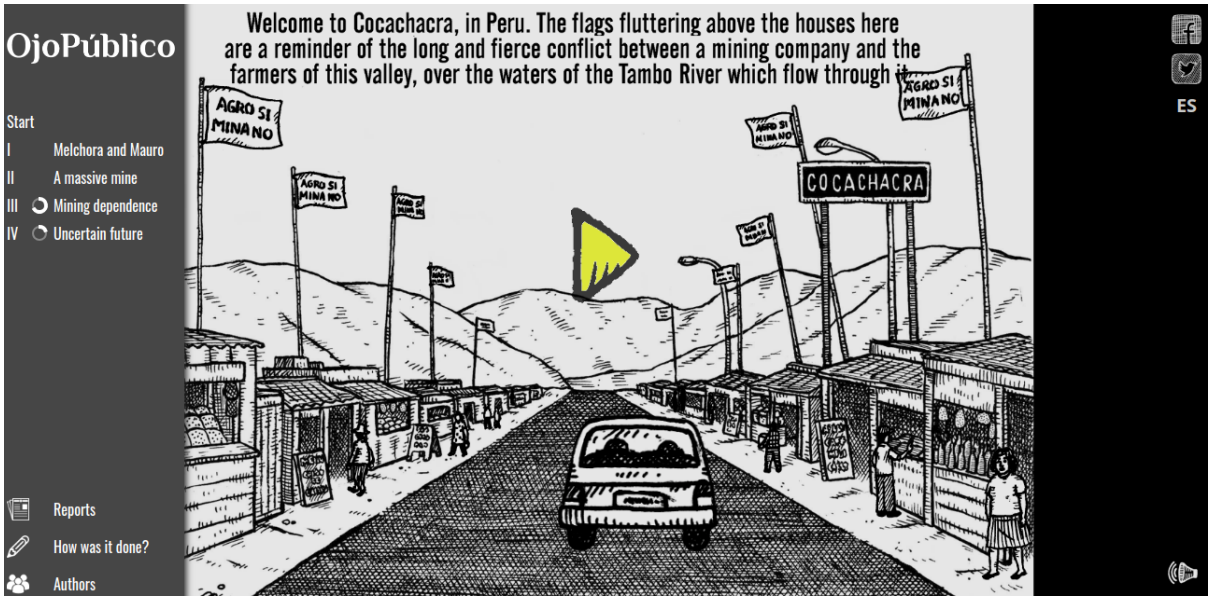


Figura 1: Página principal de "A Guerra Pela Água" (*La Guerra por el Agua*), do *Ojo Publico*.

Tudo apresentado no quadrinho "A Guerra Pela Água" (*La Guerra por el Agua*) é real. Seus personagens principais, Mauro Apaza e Melchora Tacure, existem, assim como seus medos e incertezas. Conheci ambos em um dia quente de setembro de 2016. Era meio-dia e não havia sombras ou vento. Ela removia ervas daninhas do solo com as mãos, ele fazia sulcos no solo acidentado. Por mais de 70 anos eles vêm cultivando comida em um pequeno lote no Vale de Tambo, zona rural no sul do Peru que há tempos recebe propostas para um projeto de mineração. A história deste casal, como a de milhares de agricultores e comunidades indígenas, tem a ver com as disputas entre estes e as poderosas indústrias extrativistas que atuam em torno de um dos recursos mais estratégicos do mundo: a água.

Como falar deste confronto em um país como o Peru, onde há mais de duzentos conflitos ambientais em andamento e o orçamento nacional depende do dinheiro movimentado pelo setor? Como abordar uma história de tensões entre fazendeiros humildes, interesses de multinacionais e um governo que precisa aumentar suas fontes de arrecadação? Que narrativa pode ajudar a compreender esta situação? Como mobilizar as pessoas em torno de tema tão urgente? Perguntas como estas levaram à criação de "A Guerra Pela Água",

primeira HQ interativa do Peru, criada pelo site *Ojo Publico*, reunindo dados e visualizações em uma narrativa sobre este conflito.[153]

Por que um quadrinho interativo?

O projeto teve início em julho de 2016. Optamos por narrar o conflito a partir de uma perspectiva econômica, mas abordando o leitor dentro da visão de dois agricultores em uma rota que imita uma viagem intimista a uma das regiões mais emblemáticas de toda a disputa. A interatividade do formato possibilita ao público descobrir os sons e diálogos do conflito, ao longo e além das páginas.

Escolhemos contar a história do projeto de mineração Tía Maria, da Southern Copper Corporation, uma das maiores mineradoras do mundo, de propriedade de German Larrea, um dos homens mais ricos do México e do planeta. Houve oposição local a este projeto, o que levou a violenta repressão policial e morte de seis cidadãos.

A equipe que produziu este quadrinho era composta por uma jornalista (eu mesma), o desenhista Jesus Cossio e o desenvolvedor web Jason Martinez. Nós três fomos até o Vale de Tambo, em Arequipa, coração do conflito, para conversar com líderes comunitários, agricultores e autoridades, documentando todo esse processo em anotações, fotos e desenhos, que se tornariam os primeiros rascunhos da HQ. Após voltarmos a Lima, estruturamos aquilo que seria o seu primeiro protótipo. Com base nele, escrevemos o roteiro final, definimos as funcionalidades interativas e começamos a desenvolver o projeto.

Honestidade em quadrinhos

Escolhemos a HQ como meio porque acreditamos que jornalistas não deveriam, nas palavras do cartunista Joe Sacco, "neutralizar a verdade em nome da imparcialidade". Sacco participou conosco de uma apresentação do primeiro capítulo do projeto, e cabe mencionar que foi uma de suas obras que nos inspirou: *Srebrenica*, uma HQ para web sobre o massacre em que mais de 8.000 muçulmanos bósnios morreram em 1995.

Foram necessários oito meses para que "A Guerra Pela Água" chegasse ao fim. Um quadrinho baseado em fatos, com uma estrutura narrativa que permite ao público ver como é a rotina dos personagens e trazer à tona um dos maiores dilemas da economia peruana: agricultura ou mineração? Há água o suficiente para ambos?

---

153 https://laguerraporelagua.ojo-publico.com/en/.

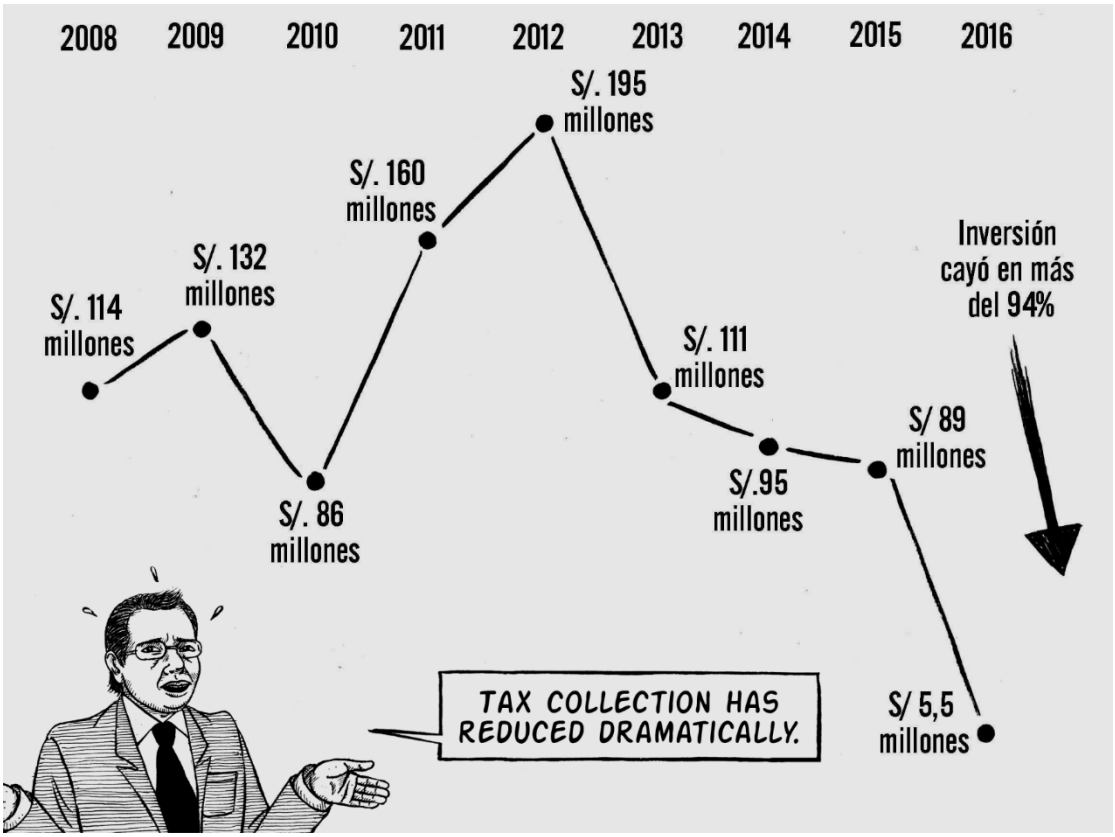


Figura 2: Visualização ilustrando a queda em arrecadação desde 2008.

Contamos a história deste conflito através dos olhares de Mauro e Melchora. Esta mesma história é acompanhada por representações de dados que revelam a dependência econômica da região e os incentivos fiscais recebidos pelas mineradoras. Todas as cenas e diálogos da HQ são reais, produtos de nossa apuração local, entrevistas com autoridades, moradores e investigações das finanças da Southern Copper. Buscamos criar cenas a partir destes diálogos, números, entrevistas e ambientes, com honestidade e precisão.

Do papel à internet

Para o cartunista Jesus Cossio, o desafio foi repensar como trabalhar com a questão do tempo em uma HQ interativa: "Quando se trata de material impresso ou digital estático, a ideia é fazer o leitor ser impacto pelas imagens e parar por um instante, já uma HQ interativa a composição e as imagens precisam ser adaptadas para um fluxo mais ágil e dinâmico de leitura".

De um ponto de vista tecnológico, o projeto foi um desafio para a equipe do *Ojo Publico*, já que nunca havíamos desenvolvido um quadrinho interativo antes. Usamos a Plataforma de Animação GreenSock (GSAP, na sigla em inglês), uma biblioteca que nos

permitiu fazer animações e transições, bem como a padronização de cenas e linha do tempo. Este processo foi complementado com uso de JavaScript, CSS e HTML5.

O saldo final da HQ: 42 cenas e mais de 120 ilustrações. Jesus Cossio desenhou cada um dos personagens, cenas e cenários do roteiro à base de lápis e nanquim. Estas imagens foram digitalizadas posteriormente e separadas em camadas — fundos, cenários, personagens e elementos que precisavam interagir uns com os outros.

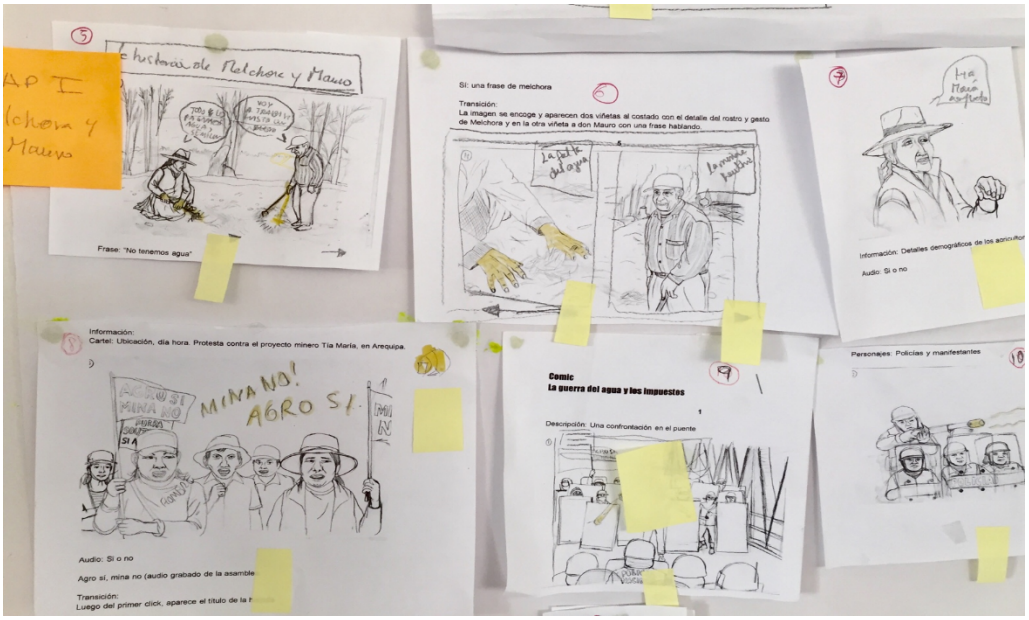


Figura 3: Desenvolvimento do storyboard de “A Guerra Pela Água”.

Da internet de volta ao papel

“A Guerra Pela Água” é uma experiência transmídia. Também publicamos uma versão impressa. Através destas duas plataformas, a HQ visa atingir públicos diferentes. Um dos principais interesses da equipe do *Ojo Publico* é a exploração de narrativas e formatos para relatar histórias (muitas vezes complexas) de interesse público. Já fomos premiados por nossas investigações com dados.

Em outros projetos do site usamos o formato de HQ para abordar temas como violência. No “Projeto Memória” (*Proyecto Memoria*), as imagens usadas narram o horror do conflito doméstico ocorrido no Peru entre 1980 e 2000. Quadrinhos são uma linguagem poderosa para contar histórias com dados. Acreditamos que jornalistas investigativos devem testar todas a linguagens possíveis para contar histórias a diferentes públicos. Mas, acima de tudo, queremos denunciar os desequilíbrios na balança do poder — neste caso, falando especificamente da administração de recursos naturais no Peru.

Nelly Luna Amancio é fundadora e editora do Ojo Publico, e jornalista investigativa especializada em meio ambiente, povos indígenas e direitos humanos.

Jornalismo de dados deve focar em pessoas e histórias[154]

Winny de Jong

Há mais em comum entre o jornalismo de dados e o jornalismo do que diferenças, como ocorre com as pessoas. Por mais que a apuração movida por dados utilize tipos diferentes de fontes, que demandam outras habilidades para questionamento, o raciocínio é meio que o mesmo. Na verdade, com a distância adequada, percebemos que os processos são quase que indistinguíveis.

Desconhecidos conhecidos

Em seu âmago, jornalismo é o ofício de transformar coisas sabidamente desconhecidas em conhecidas de fato. O conceito de conhecidos e desconhecidos foi popularizado pelo Secretário de Defesa dos EUA, Donald Rumsfeld, em 2002. Na época, faltavam provas de que o governo iraquiano havia fornecido armas de destruição em massa a grupos terroristas. Durante coletiva de imprensa sobre o tema, Rumsfeld disse o seguinte:

> Relatos que dizem que algo não aconteceu sempre me interessam pois, como sabemos, há fatos conhecidos que são de fato conhecidos; coisas que sabemos que sabemos. Também sabemos de fatos desconhecidos conhecidos; ou seja, sabemos que há coisas que não conhecemos. Mas há também os desconhecidos desconhecidos, aqueles que não sabemos que não sabemos. E, se observarmos a história de nosso país e outros países livres, é esta última categoria que é difícil de se lidar.

[154] Já que ideias são novas combinações de antigos elementos, este ensaio baseia-se na previsão da Fundação Nieman para o jornalismo escrita por Winny em 2019, em uma palestra durante a Smart News Design Conference, em Amsterdã, e no site holandês *alshetongeveermaarklopt.nl*, que ensina matemática para jornalistas.

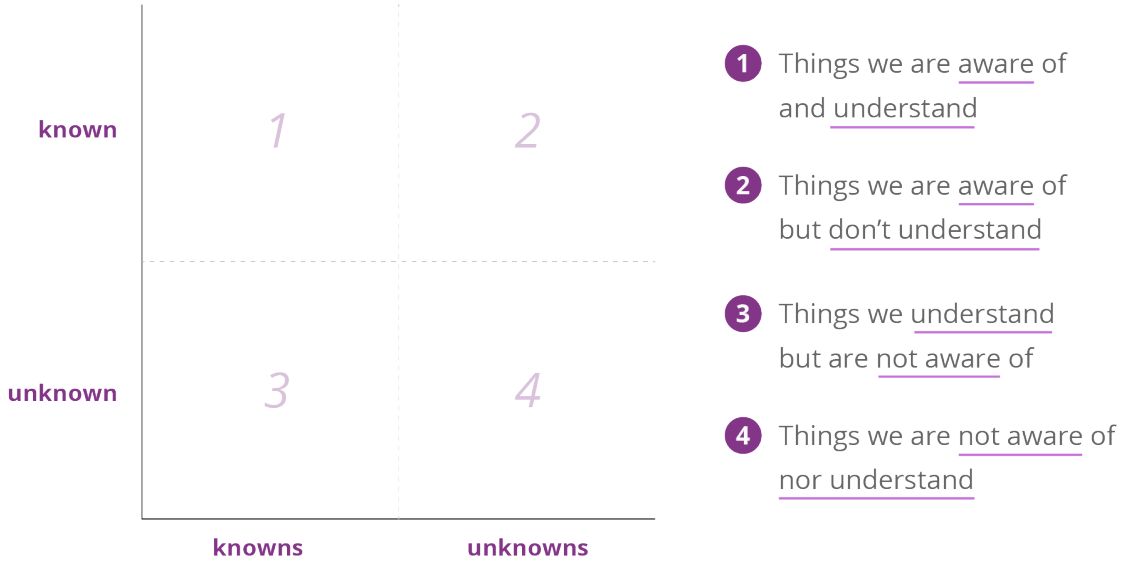


Figura 1: Matriz de conhecidos e desconhecidos.[155] Gráfico por Lars Boogaard.

Todo processo jornalístico consiste em mover peões ao longo da matriz de combinações do que se sabe e não se sabe. Todo jornalismo começa com uma pergunta, ou, no caso desta matriz, um desconhecido conhecido (você sabe que há algo sobre o qual você não sabe, logo, existe o questionamento). Ao se preparar para sair da pergunta ou palpite para um material pronto para publicação, idealmente todos os peões são movidos para a categoria de conhecidos conhecidos. Mas, como qualquer jornalista te diria, a realidade geralmente não é bem assim. Durante a apuração, através de entrevistas com pessoas ou verificação de documentos ou conjuntos de dados, provavelmente você irá se deparar com coisas que você não fazia ideia que não sabia (desconhecidos desconhecidos) e que também exigem respostas. Com sorte, pode acabar esbarrando com algumas coisas com as quais já era familiarizado e não sabia (desconhecidos conhecidos). Trabalhando em direção ao seu prazo final, você está às voltas com três categorias de conhecimento — desconhecidos conhecidos, os questionamentos que deram início a tudo; desconhecidos desconhecidos, questionamentos que você não fazia ideia que deveria ter feito; e desconhecidos conhecidos, as respostas que você não sabia que tinha — e precisa transformar tudo isso em fatos realmente conhecidos. Diferentemente de como agem os governos, jornalistas só podem agir, ou publicar, o que se sabe de fato.

[155] Esta matriz é um pouco diferente da Janela de Johari, por vezes utilizada em psicologia cognitiva para ajudar as pessoas a melhor compreenderem sua relação consigo mesmas e com terceiros.

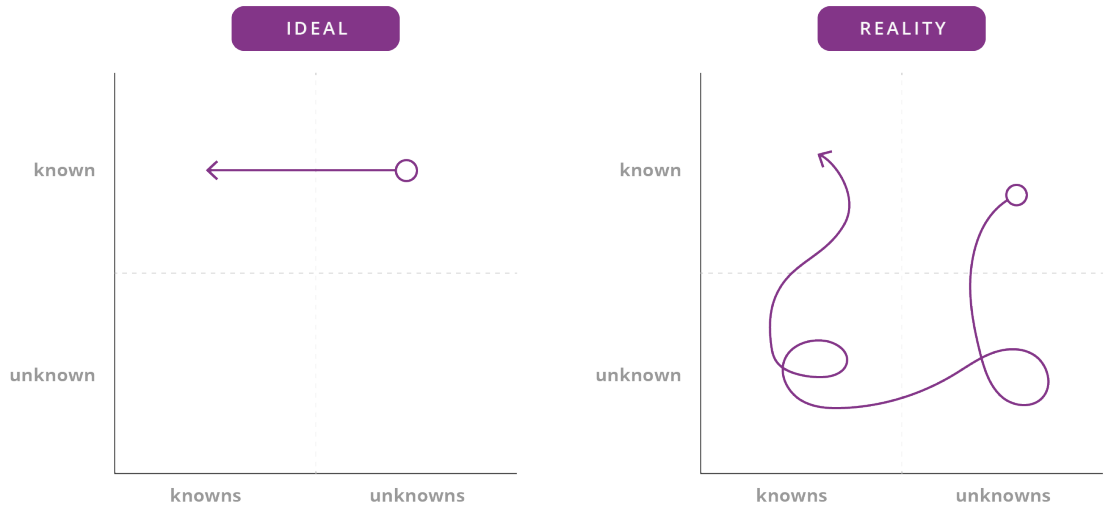


Figura 2: Navegando entre o que se sabe e o que não se sabe no jornalismo. Gráfico por Lars Boogaard.

Jornalismo sólido

Sendo quase indistinguíveis, o jornalismo tradicional e o de dados certamente seguem os mesmos padrões. Ambos devem ser verdadeiros, independentes e livres de vieses. Como todos os outros fatos, dados também precisam ser checados. Então, antes de mais nada, é preciso se perguntar se os dados são reais. O que cada número de fato significa? Qual a fonte? Por que os dados foram coletados? Quem criou o conjunto destes dados? Como a tabela de dados foi criada? Há pontos fora da curva em meio a estas informações? Estas informações fazem sentido? E, muitas vezes esquecemos, mas isso vale para qualquer entrevista: o que a fonte não diz? Ao passo que as exigências e, por conseguinte, as perguntas são as mesmas, as ações resultantes destas variam um pouquinho.

Como Bill Kovach e Tom Rosenstiel descrevem em *The Elements of Journalism*, a primeira tarefa do jornalista é "checar quais informações são confiáveis e ordená-las de forma que as pessoas as compreendam de maneira eficaz". No caso dos jornalistas de dados, especialmente aqueles trabalhando em rádio ou TV, isso significa que aqueles números que aprenderam a amar talvez não tenham espaço no produto final.

Nerdice restrita

Precisão é necessária se tratando de análise de dados, óbvio. Mas, para que as pessoas compreendam sua história de "maneira eficaz", há um limite para essa precisão, no caso, o

número de casas decimais usadas no produto final. Ou seja, usar uma medida do tipo “4 a cada 10 pessoas” certamente é melhor do que algo como ‘41,8612%”. Com base em minha experiência, diria que você pode ser tão preciso quanto normalmente seria ao discutir o tema com seus amigos não nerds-de-dados em uma tarde de sábado.

A não ser que o seu público precise saber dos métodos e ferramentas utilizados para compreender o que está sendo dito, é melhor deixar a nerdice para a parte de metodologia. Porque quando as pessoas estão lendo, ouvindo ou assistindo ao seu produto, eles precisam pensar no que está sendo relatado, não nos dados, na análise, ou na tecnologia que apoia aquela história. No final, isso quer dizer que o que há de melhor em jornalismo de dados muitas vezes nem é encarado assim, o que faz deste um ofício invisível. Contanto que tal invisibilidade sirva ao propósito da história, fazendo o seu jornalismo mais fácil de ser compreendido, melhor. Afinal, a prática jornalística cria diferentes mapas com os quais os cidadãos podem navegar pela sociedade, então devemos nos certificar que estes mapas sejam legíveis para todos e lidos por muitos.

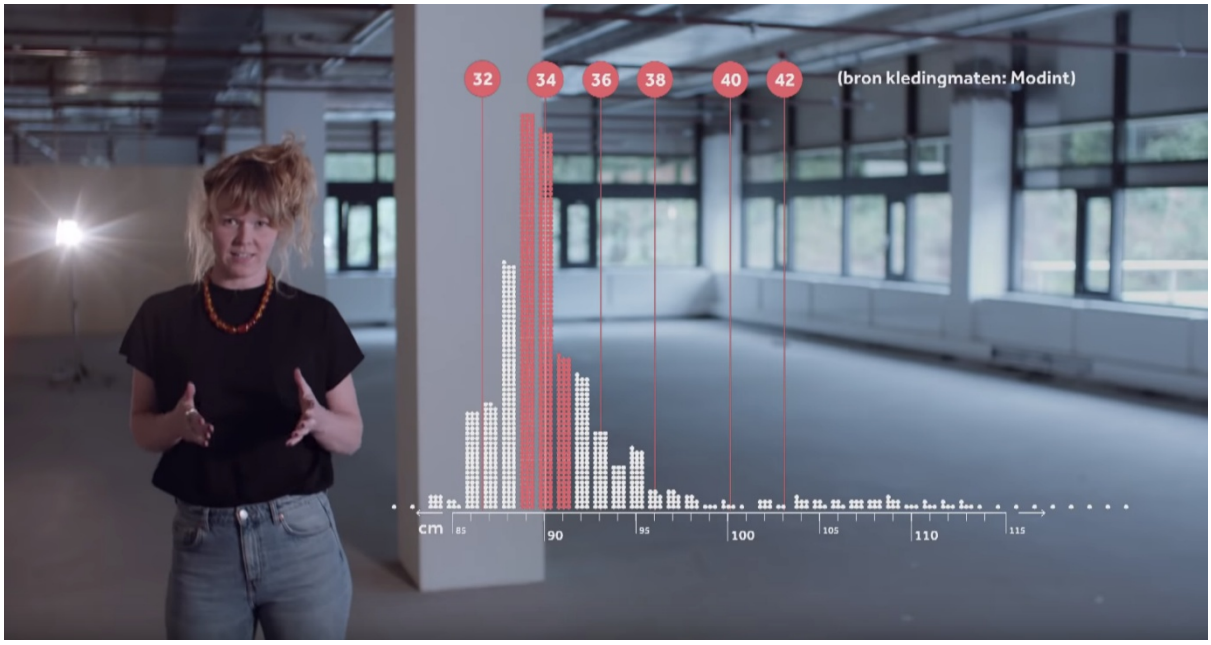


Figura 3: Quão magra você precisa ser para se tornar modelo. Captura de tela de vídeo da NOS.

Rádio e televisão

Em rádio e televisão, ao divulgar produtos de jornalismo de dados, menos é mais. Na redação da *NOS*, repórteres discutem o número de segundos disponíveis para contarem suas histórias. Isso significa que não há tempo hábil para falar sobre como a reportagem foi feita

ou porque se optou por usar uma fonte e não outra se isso não contribui para o produto final ou a compreensão do público sobre este. Em um vídeo publicado na internet sobre o quão magra se precisa ser para trabalhar como modelo, gastamos 20 segundos explicando nossos métodos.[156] Quando se tem 90 segundos para contar uma história em rede nacional, 20 segundos é muito. Neste caso, 'menos é mais' significa que não há tempo para explicar como foi feita a apuração. Sob condições de tempo e espaço limitados, a história prevalece.

Visual modesto

Claro que o adágio 'menos é mais' vale também para visualizações de dados. Jornalismo de dados se assemelha a sexo na adolescência: todo mundo fala sobre, mas quase ninguém faz. Quando uma redação finalmente inclui dados em seu arsenal, a tendência é agir de forma espalhafatosa, criando representações visuais para tudo. Olha, eu amo essas representações, especialmente as mais inovadoras, de alto nível, mas só se contribuem para a história sendo contada. (Visualizações podem agregar valor ao jornalismo das mais diversas formas. Dentre elas, o aprofundamento da compreensão do público a respeito do tema em questão e a ampliação deste entendimento ao oferecer percepções adicionais a nível regional, por exemplo.) A dica é: seja discreto. Restrinja-se a criar representações quando estas podem adicionar algo à história. Hoje em dia, muita gente ouve rádio ou assiste TV enquanto faz outras coisas. Isso limita sua capacidade de absorver informações; ao dirigir, ouvir as notícias é uma tarefa secundária, o mesmo vale para assistir TV enquanto cozinha. Então, é bom ter cuidado e não exigir demais do público. Tudo isso pode fazer o nosso ofício invisível, mas estamos aqui para dar notícias e contar histórias, não para exibir nossas habilidades em visualização de dados.

156 https://www.youtube.com/watch?v=DWRGqmywNY.

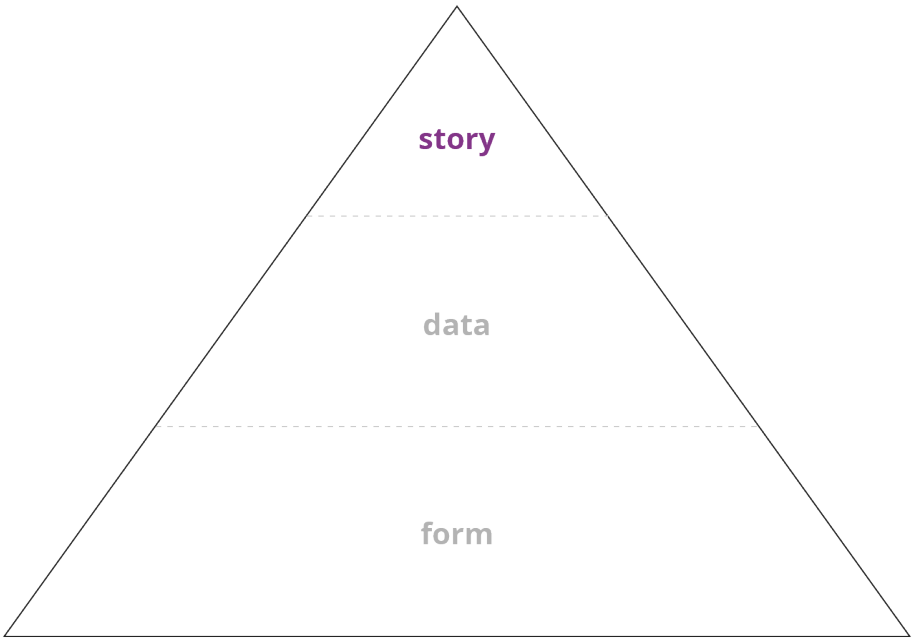


Figura 4: Gráfico por Lars Boogaard.

Foco nas pessoas

Tudo isso para dizer que o que realmente importa, no seu artigo, no jornalismo, na vida como um todo, não cabe em um conjunto de dados. Nunca coube, nunca caberá. No final das contas, o que importa são as pessoas. Então, independentemente do que fizer ou onde publicar, fale de pessoas e não de dados. Quando se sentir tentado a usar mais dados, tecnologia ou qualquer outra nerdice jornalística além do necessário, lembre-se que você é um dos poucos habilitados neste campo. Isso por si só é incrível, não precisa destacar o óbvio. Siga o que todo bom jornalismo de dados faz: o formato facilita os dados que facilitam a história. Tudo e todos precisam orbitar em torno desta história, ela é nosso sol. A história manda. Não há espaço para egos aqui.

Winny de Jong é jornalista de dados na emissora nacional holandesa NOS.

Investigação de dados, plataformas e algoritmos

Ciência forense digital: reutilização de IDs do Google Analytics[157]

Richard Rogers

Quando um jornalista investigativo descobriu uma rede secreta de sites russos que espalhava desinformação sobre a Ucrânia em julho de 2015, esta revelação não foi apenas um presságio da campanha de influência financiada pelo estado antes das eleições presidenciais dos EUA de 2016. Ela também popularizou uma técnica de descoberta de redes para jornalistas de dados e pesquisadores sociais (Alexander, 2015). Quais sites compartilham o mesmo ID do Google Analytics (ver Figura 1)? Se os sites usam a mesma ID, por tabela são operados pela mesma entidade que os registrou, seja um indivíduo, organização ou grupo de mídia. Lawrence Alexander, o jornalista em questão, se viu motivado em sua pesquisa ao não encontrar uma fonte por trás do endereço emaidan.com.ua, site que aparentemente fornecia informações sobre os protestos Euromaidan de 2013 a 2014, que derrubaram o presidente ucraniano pró-Rússia em prol de uma figura favorável ao ocidente. Em busca da fonte e "intrigado pelo anonimato", Alexander mergulhou no código do site (2015).

157 O autor gostaria de prestar o devido reconhecimento aos fundamentos do tema desenvolvidos por Mischa Szpirt. Para maiores informações, checar Rogers (2019), capítulo 11, e Bounegru et al. (2017), capítulo 3.

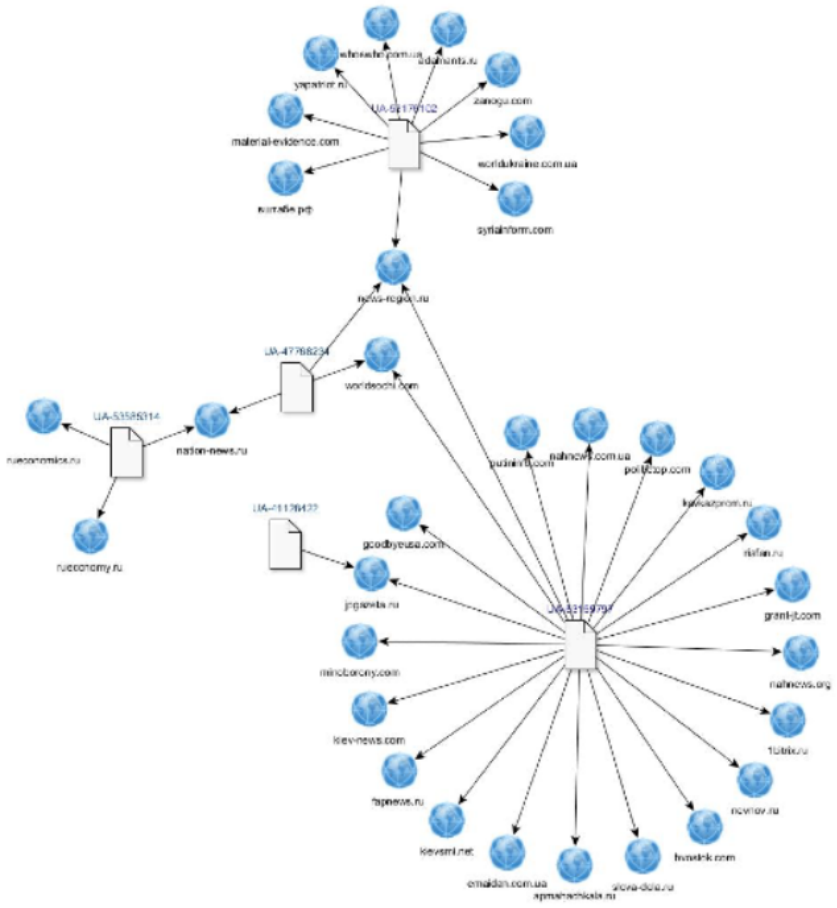


Figura 1: Rede de sites descoberta através de IDs compartilhadas do Google Analytics. Fonte: Alexander (2015).

Em meio ao código-fonte da página, ele encontrou um ID do Google Analytics e inseriu-a em um software de pesquisa reverso, que lhe deu uma lista de outros sites com o mesmo ID.[158]

Alexander, então, encontrou uma rede (em forma de estrela) que ligava um ID do Google Analytics a outros oito websites (Figura 1, no topo do diagrama). Todos adotavam a mesma narrativa antiucraniana. Um destes sites também usava um ID adicional, que levou a outra rede de sites relacionados (Figura 1, no canto inferior direito), seguindo a mesma linha política. Ao examinar os registros whois destes sites, Alexander chegou a um endereço de email associado, perfil e foto na rede social russa VKontakte (atual VK). O nome deste indivíduo estava em uma lista vazada de funcionários da Agência de Pesquisa da Internet em São Petersburgo, local de trabalho do "exército troll" patrocinado pelo governo russo (Chen, 2015; Toler, 2015). Traçando elos entre pontos de dados, Alexander conseguiu dar rosto e

[158] A pesquisa também pode revelar o endereço IP de cada site, ID do Google AdSense, registro de domínio whois e demais informações relevantes.

nome ao tal troll russo. Também humanizou este troll, de certa forma, ao apontar sua página no Pinterest, onde postava fotos de conquistas russas relacionadas ao espaço. Descobrimos, ali, que o troll era fã de cosmonautas.

Figura 2: ID do Google Analytics.

O uso de ferramentas de 'inteligência open source' (OSINT, na sigla em inglês) como técnicas de descoberta (e métodos digitais, no sentido de reutilização de Google Analytics e softwares de busca reversa) permite a Alexander e outros jornalistas que criem e deem seguimento a links em código, registros públicos, bancos de dados e vazamentos, possibilitando a criação de um dossiê para saber quem está por trás de operações específicas (Bazzell, 2016). Este processo de 'descoberta' é uma abordagem investigativa, ou mesmo de ciência forense, digital para mineração e exposição jornalística, em que se busca identificar e contatar o indivíduo, organização ou grupo de mídia envolvido para entrevistá-lo, dando a oportunidade para que se responsabilize por seu trabalho.[159] Ambos os relatos, a descoberta do jornalista e a explicação do descoberto, constituem a história a ser contada. O objetivo é tornar tudo público, tirar da bagunça do código dos sites o trabalho político secreto sendo feito, junto do reconhecimento desta prova em particular (Latour, 2005).

[159] A ciência forense digital tem origens na investigação de fraudes corporativas através de técnicas como *data carving*, que possibilita a recuperação de arquivos excluídos.

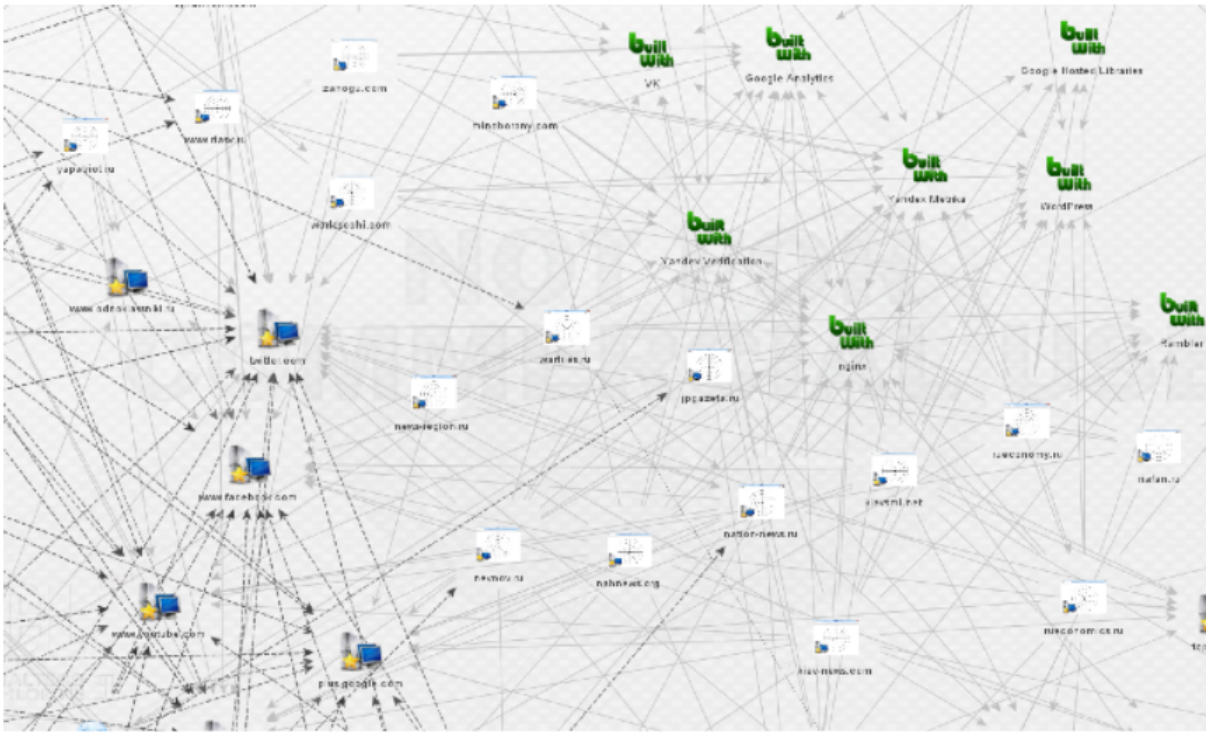


Figura 3: Objetos digitais embutidos em sites, apresentados como diagrama de rede. Fonte: Alexander (2015).

Investigações com base no ID do Google Analytics fazem parte de uma linhagem de práticas que visam desmascarar anônimos na internet por meio de brechas ou pontos de acesso a dados pessoais que não foram previstos por seus criadores. Minerar IDs do Google Analytics para descoberta e mapeamento de redes é, também, um exercício de reutilização, empregando o software de formas imprevistas para fins de pesquisa social.

O criador desta técnica, Andy Baio, jornalista da revista *Wired*, conta a história de um blogueiro anônimo que postava material altamente ofensivo e cobria seus rastros com os métodos de sempre: "ocultando informações pessoais no registro de domínio, usando um IP diferente de seus outros sites, e retirando quaisquer recursos compartilhados de sua instalação do WordPress" (2011). Baio descobriu de quem se tratava por conta do ID do Google Analytics, deixado à vista junto de outros sites do mesmo dono. Esta historinha a respeito desta técnica de descoberta e identificação se conclui com Baio fornecendo um guia para outros blogueiros anônimos *que lutam por uma causa justa*, como aqueles que monitoram cartéis de tráfico mexicanos, cuja descoberta poderia levar a perigo e até mesmo à morte. Desta forma, poderiam testar a robustez do anonimato e informar aos jornalistas trabalhando online sobre quaisquer vulnerabilidades ou brechas em potencial.

Protocolo de pesquisa para descoberta de rede através de IDs do Google Analytics

1. Reúna uma lista de sites que não informam sua fonte.

2. Encontre seus IDs do Google Analytics e Adsense.

3. Verifique a lista de endereços em softwares de busca reversa, como o disponível em dnslytics.com.

4. Procure por sites que compartilham os mesmos IDs.

5. Agrupe por tema e caracterize os endereços que compartilham estes IDs.

6. Considere utilizar softwares como Gephi para visualização de redes.

Richard Rogers é professor de Novas Mídias e Cultura Digital na Universidade de Amsterdã e Diretor da Digital Methods Initiative e da Escola de Pesquisa Holandesa de Estudos de Mídia.

Referências

ALEXANDER, Lawrence. *Open-Source Information Reveals Pro-Kremlin Web Campaign*. GlobalVoices, 13 de julho de 2015. Disponível em: https://globalvoices.org/2015/07/13/open-source-information-reveals-pro-kremlin-web-campaign/.

BAIO, Andy. *Think you can hide, anonymous blogger? Two words: Google Analytics*. Wired, 15 de novembro de 2011.

BAZZELL, Michael. *Open Source Intelligence Techniques: Resources for Searching and Analyzing Online Information.* North Charleston: CreateSpace Independent Publishing Platform, 2016.

BOUNEGRU, Liliana et al. *A Field Guide to Fake News*. Amsterdã: Public Data Lab, 2017.

CHEN, Adrian. *The Agency*. New York Times, 2 de junho de 2015.

COOKSON, Robert. *Jihadi website with beheadings profited from Google ad platform*. Financial Times, 17 de maio de 2016.

CUSH, Andy. *Who's Behind This Shady, Propagandistic Russian Photo Exhibition?* Gawker, 10 de outubro de 2014.

LATOUR, Bruno. *From Realpolitik to Dingpolitik or How to Make Things Public*. In: LATOUR, Bruno; WEIBEL, Peter (eds). *Making Things Public: Atmospheres of Democracy*. Cambridge: MIT Press, 2015, p. 14-41.

ROGERS, R. *Doing Digital Methods*. Londres: SAGE Publications, 2019.

TOLER, Aric. *Inside the Kremlin Troll Army Machine: Templates, Guidelines, and Paid Posts*. GlobalVoices, 14 de março de 2015. Disponível em: https://globalvoices.org/2015/03/14/russia-kremlin-troll-army-examples/.

Contando histórias com as redes sociais[160]

Lam Thuy Vo

Nos tornamos os maiores produtores de dados da história. Quase todo clique online, deslizada em nossos tablets ou toque em nossos celulares produz um ponto de dados em um repositório virtual. Somente o Facebook gera dados sobre as vidas de mais de 2 bilhões de pessoas. Já o Twitter registra as atividades de mais de 330 milhões de usuários mensais. Um estudo do MIT descobriu que o trabalhador de escritório norte-americano médio gerava 5 gigabytes de dados por dia.[161] Isso aconteceu em 2013 e não vai parar. Cada vez mais e mais pessoas vivem online, e a penetração dos smartphones vêm crescendo em locais antes desconectados pelo mundo, o que significa que esse volume segue aumentando.

Muitos pesquisadores tendem a tratar cada usuário de redes sociais como um tópico individual, nada além de evidência anedótica e pontos únicos de contato. Mas fazer isso com um punhado de usuários e suas postagens individuais é ignorar o potencial de centenas de milhões de outros e suas interações entre si. Há muito que poderia ser contado com base nos enormes volumes de dados produzidos por usuários de redes sociais e suas plataformas, considerando que pesquisadores e jornalistas só agora começaram a se especializar em técnicas para retirarem dados relevantes de dados de larga escala como estes.

Eventos recentes mostram que é cada vez mais crucial para repórteres entenderem melhor esta web voltada ao social. A interferência russa nas eleições presidenciais dos EUA em 2016; o Brexit; a disseminação perigosa de discurso de ódio contra muçulmanos através do Facebook na Europa e em Mianmar; o uso pesado do Twitter por líderes globais — todos estes desdobramentos mostram que há uma necessidade crescente de um entendimento maior sobre a utilidade e os gargalos nas redes sociais como um todo.

Como jornalistas podem usar dados de redes sociais?

Por mais que existam diversas maneiras pelas quais as redes sociais podem ser úteis em reportagens, pode valer a pena examinar os dados que podem ser coletados das plataformas sob duas óticas.

---

[160] https://source.opennews.org/articles/what-buzzfeed-news-learned-after-year-mining-data-/, http://www.niemanlab.org/2016/12/the-primary-source-in-the-age-of-mechanical-multiplication/.

[161] https://www.technologyreview.com/s/514351/has-big-data-made-anonymity-impossible/.

Primeiro, as redes sociais podem ser usadas como meio para melhor entender indivíduos e suas ações. Sejam declarações públicas ou interações privadas entre dois indivíduos — muitas das ações das pessoas, mediadas e disseminadas através da tecnologia, atualmente deixam rastros online que podem ser analisados para oferecem novas percepções. Esta técnica é especialmente útil quando se observam políticos e demais figuras relevantes, cujas opiniões públicas podem indicar o direcionamento de suas políticas e ações ou ter consequências reais, como queda de ações ou a demissão de indivíduos de alto escalão.

Segundo, a web pode ser encarada como um ecossistema próprio em que as histórias se dão em plataformas sociais (ainda que movidas por ações humanas e automatizadas). Campanhas de desinformação, universos de informação enviesados por algoritmos e ataques de trolls são alguns dos fenômenos únicos da rede.

Como dados sociais são usados em materiais jornalísticos

Em vez de discutir estes tipos de narrativa de maneira abstrata, talvez seja melhor compreender dados de redes sociais no contexto de como podem ser usados para contar histórias em particular. As seções a seguir discutem diversos projetos jornalísticos que utilizaram este tipo de informação.

Entendendo figuras públicas: dados de redes sociais voltados à responsabilização na imprensa

Para pessoas públicas e comuns, as redes sociais se tornaram uma forma de dialogar publicamente de maneira direta. Atualizações de status, tweets e postagens podem servir como maneiras de burlar antigos mecanismos de projeção, como entrevistas com grandes veículos, notas à imprensa ou coletivas.

Para políticos, porém, estes anúncios públicos, verdadeiras projeções do seu eu, podem se tornar declarações vinculativas e, no caso de gente poderosa, podem preceder políticas que ainda não foram postas em prática.

Como parte do trabalho de um político consiste em lidar com o público, pesquisar suas redes sociais pode nos ajudar a compreender melhor sua ideologia. Para um projeto em específico, meu colega Charlie Warzel e eu coletamos e analisamos mais de 20.000 tweets de Donald Trump para responder à seguinte pergunta: que tipo de informação ele espalha e como esta informação pode indicar o tipo de informação que ele consome?

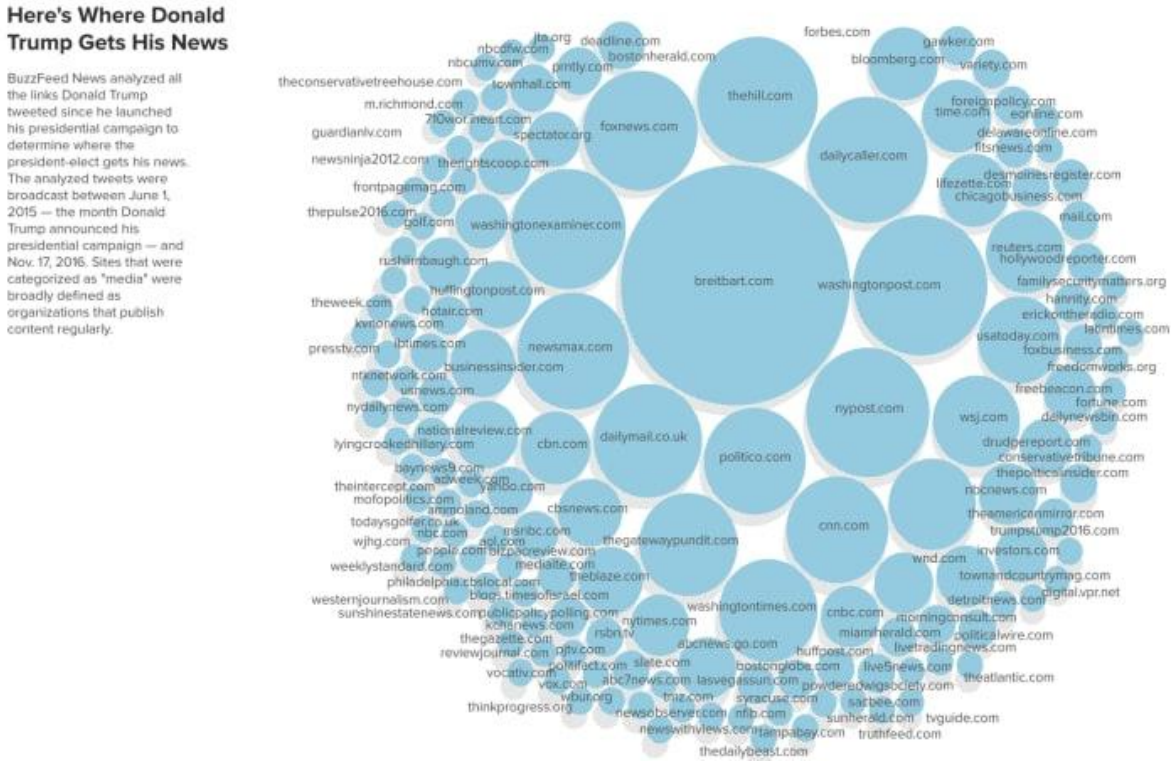


Figura 1: Uma amostra dos links compartilhados por Trump no Twitter durante sua campanha presidencial.

Pontos de dados de redes sociais não oferecem uma imagem completa de quem somos, em parte por conta de sua natureza performativa e em parte porque estes conjuntos de dados são incompletos e sujeitos a interpretações individuais. Podem funcionar de maneira complementar, porém: a ligação entre o presidente Trump e o site *Breitbart*, mostrada acima, era um indicativo precoce de sua relação com Steve Bannon na vida real. Seus constantes retweets de blogs conservadores como *The Conservative Tree House* e *News Ninja 2012* poderiam ser considerados indícios de sua desconfiança em relação à "grande mídia".[162]

Traçando ações humanas ao ponto de origem

Comunicações públicas e semipúblicas, como tweets e postagens abertas no Facebook, podem nos dar uma noção de como as pessoas se apresentam para terceiros, mas há também o tipo de dado que vive dentro destas plataformas sociais por trás de portas fechadas, caso de mensagens privadas, pesquisas no Google ou dados de geolocalização.

Christian Rudder, cofundador do OKCupid e autor de *Dataclisma: Quem somos quando achamos que ninguém está vendo,* apresentou uma descrição bastante adequada para

[162] https://theconservativetreehouse.com/, http://newsninja2012.com/.

este tipo de informação: estatísticas de nosso comportamento registradas quando "achamos que ninguém está vendo".

Ao usar uma plataforma de redes sociais, pessoas criam dados longitudinais de seu próprio comportamento. Por mais que seja difícil extrapolar estes montes de dados para além de quem os gerou, estas informações podem ser extremamente poderosas ao tentar se contar a história de uma única pessoa. Gosto de chamar esta abordagem de Selfie Quantificada, termo criado por Maureen O'Connor para falar de meu trabalho durante uma conversa que tivemos.

Tomemos a história de Jeffrey Ngo como exemplo. Quando tiveram início os protestos pró-democracia em Hong Kong, sua cidade natal, no começo de setembro de 2014, Ngo, estudante da Universidade de Nova York à época, sentiu-se impelido a agir. Ele começou a entrar em contato com outros expatriados de Hong Kong em Nova York e Washington; acabou organizando protestos em 86 cidades pelo mundo e forjando uma história emblemática em meio a diversos movimentos que se originam a partir de uma revolta global sobre determinado assunto.

Para esta matéria da *Al Jazeera America*, Ngo nos permitiu analisar seu histórico do Facebook, um arquivo que pode ser baixado por qualquer usuário da plataforma.[163] Coletamos as mensagens trocadas com outro organizador-chave de Hong Kong e nos deparamos com 10 salas de chat diferentes em que estes e outros dois organizadores discutiam suas atividades políticas.

O gráfico abaixo (Figura 2) documenta a ida e vinda de mensagens. Em um primeiro momento, há um pico de mensagens quando certa notícia casou revolta pública: a polícia de Hong Kong lançou bombas de gás lacrimogêneo sobre manifestantes pacíficos. Surge, então, uma sala de chat, a de cor bege, que se tornou a sala usada pelos organizadores para planejarem atividades políticas para além das primeiras notícias.

[163] http://projects.aljazeera.com/2015/04/loving-long-distance/hong-kong-umbrella-protest.html.

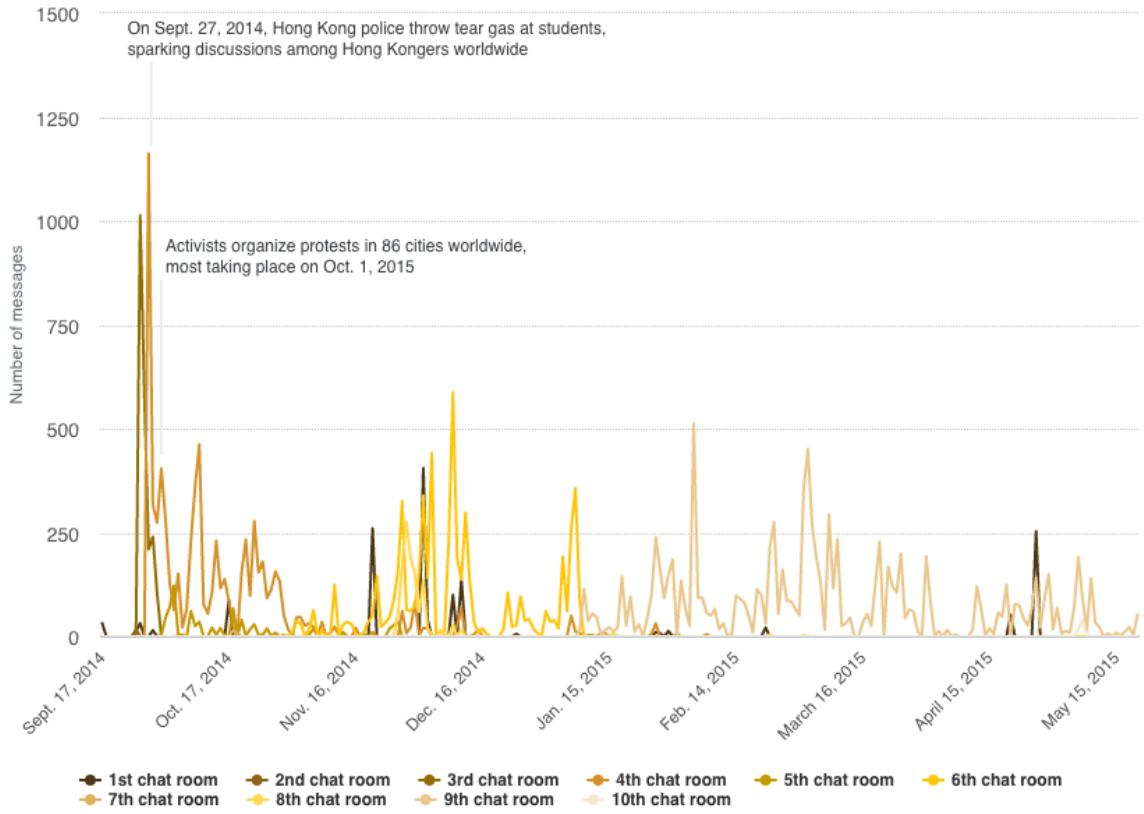


Figura 2: Grupo no Facebook *United for Democracy: Global Solidarity with Hong Kong*. Fonte: dados do Facebook cedidos por Jeffrey Ngo.

Considerando que a maior parte das discussões e do planejamento se deu dentro destas salas de chat, também pudemos recuperar o momento em que Ngo conheceu Angel Yau, coorganizadora dos protestos. Ngo não lembrava de suas primeiras conversas, mas graças ao arquivo do Facebook, pudemos reconstruir o primeiro contato entre os dois.

Está claro que a evolução de Ngo enquanto agitador político é a de um único indivíduo e em que nada reflete a situação de todos que participaram deste movimento, mas, mesmo assim, é representativa do *tipo* de caminho que pode ser percorrido por um manifestante na era digital.

Fenômenos específicos de ecossistemas online

Muitas de nossas interações agora se dão exclusivamente em plataformas online.

Muito de nosso comportamento social on e offline se mistura, mas os ambientes online seguem bastante únicos, pois, na rede, pessoas têm o auxílio de ferramentas poderosas.

A prática de bullying, por exemplo. Tão antiga quanto a própria humanidade. A diferença é que, agora, valentões contam com a ajuda de milhares de outros valentões, que podem ser acionados em um piscar de olhos. Estes têm acesso a motores de busca e rastros digitais da vida de alguém, por vezes até informações sobre as personas online dos indivíduos em questão. Além disso, há a questão da amplificação — um valentão gritando do outro lado do corredor é uma coisa, milhares indo para cima de você ao mesmo tempo é outra completamente diferente. Essa é a natureza do que hoje chamamos de trollagem.

A editora do *Washington Post* Doris Truong se viu em meio à controvérsia política na internet. No decorrer de alguns dias, trolls (e um monte de gente defendendo-a) direcionaram 24.731 tweets a ela. Ataques online tão virulentos acabam cobrando seu preço.

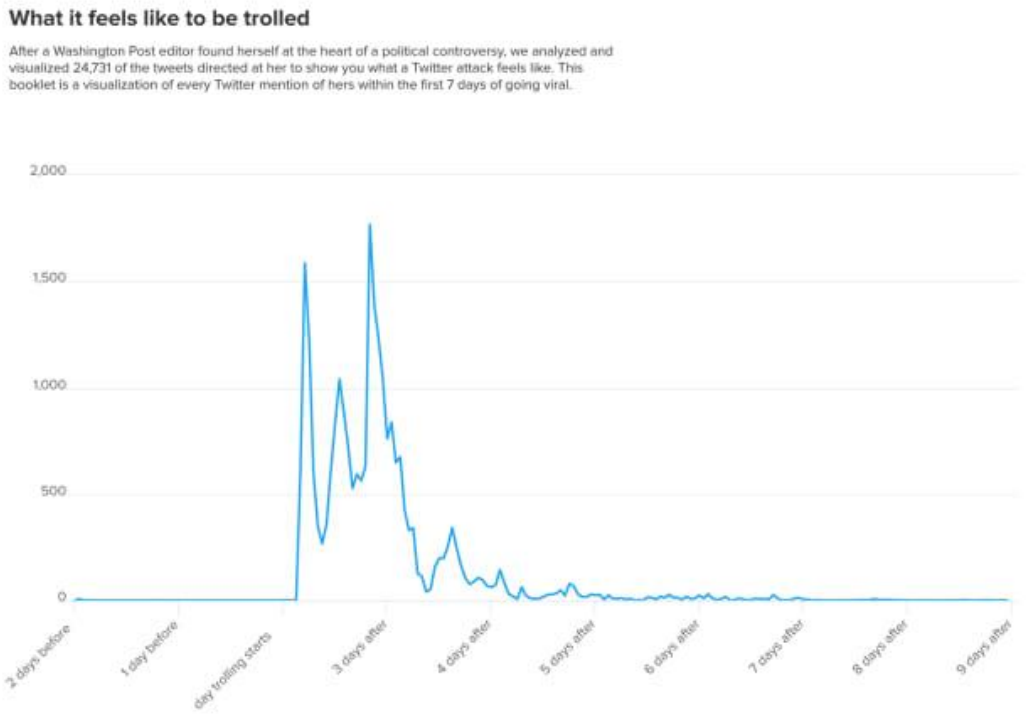


Figura 3: Gráfico com as menções a Doris Truong no Twitter desde o início do ataque.[164]

Trollagem, não muito diferente de outros tipos de ataques online, virou um problema que pode atingir qualquer um agora, famoso ou não. De avaliações no Yelp que viralizam — caso da confeitaria que se recusou a fazer um bolo de casamento para um casal gay — às maneiras com que a viralização causou a demissão e o constrangimento público de Justine Sacco, profissional de relações-públicas que fez uma piada péssima sobre HIV e sul-africanos

[164] https://www.buzzfeednews.com/article/lamvo/heres-what-it-feels-like-to-be-trolled-in-trumps-america.

pouco antes de uma viagem intercontinental — muito do que afeta nosso cotidiano hoje ocorre na internet.

Guerra de informação

A ascensão e onipresença das redes sociais trouxe consigo um novo fenômeno para nossas vidas: a viralidade.

Compartilhamentos em redes sociais possibilitaram a qualquer conteúdo ser visto não por centenas, mas milhões de pessoas sem grandes e dispendiosas campanhas de marketing, sem comprar tempo na TV.

Junto a isso, muita gente também descobriu como passar a perna em algoritmos com seguidores falsos ou comprados, bem como contas (semi)automatizadas, comandadas por bots e ciborgues.[165]

Bots não são maus, em princípio: muitos nos entretêm com sua poesia excêntrica ou dicas de autocuidado. Mas, como dito pelo pesquisador do Atlantic Council, Ben Nimmo, que estuda exércitos de bots há anos, em entrevista ao *BuzzFeed*: "[Bots] têm o potencial de distorcer seriamente qualquer debate [...] Eles podem fazer com que um grupo de seis pessoas pareça ter 46.000 membros".

As próprias plataformas de redes sociais estão em um ponto de virada na sua existência, em que devem reconhecer sua responsabilidade na definição e repressão do que podem entender como "bots problemáticos". Enquanto isso, cabe aos jornalistas reconhecerem também a presença cada vez maior de não humanos e sua força online.

Para uma matéria explicativa sobre o tema, queríamos comparar tweets feitos por um humano com tweets feitos por um bot.[166] Por mais que não haja um jeito 100% preciso de determinar se uma conta é operada com base em linhas de código, ou seja, não por um ser humano, há formas de analisar as características de um usuário para determinar se seu comportamento é suspeito. Uma das características analisadas foi a atividade da conta.

Para tanto, comparamos a atividade de uma pessoa real com a de um bot. No dia e na hora de maior atividade, o bot analisado fez mais de 200 tweets. Já o humano, apenas 21.

---

165 https://www.buzzfeednews.com/article/lamvo/twitter-bots-v-human#.kpLQjnEKa.

166 https://www.buzzfeed.com/lamvo/heres-what-we-learned-from-staring-at-social-media-data-for.

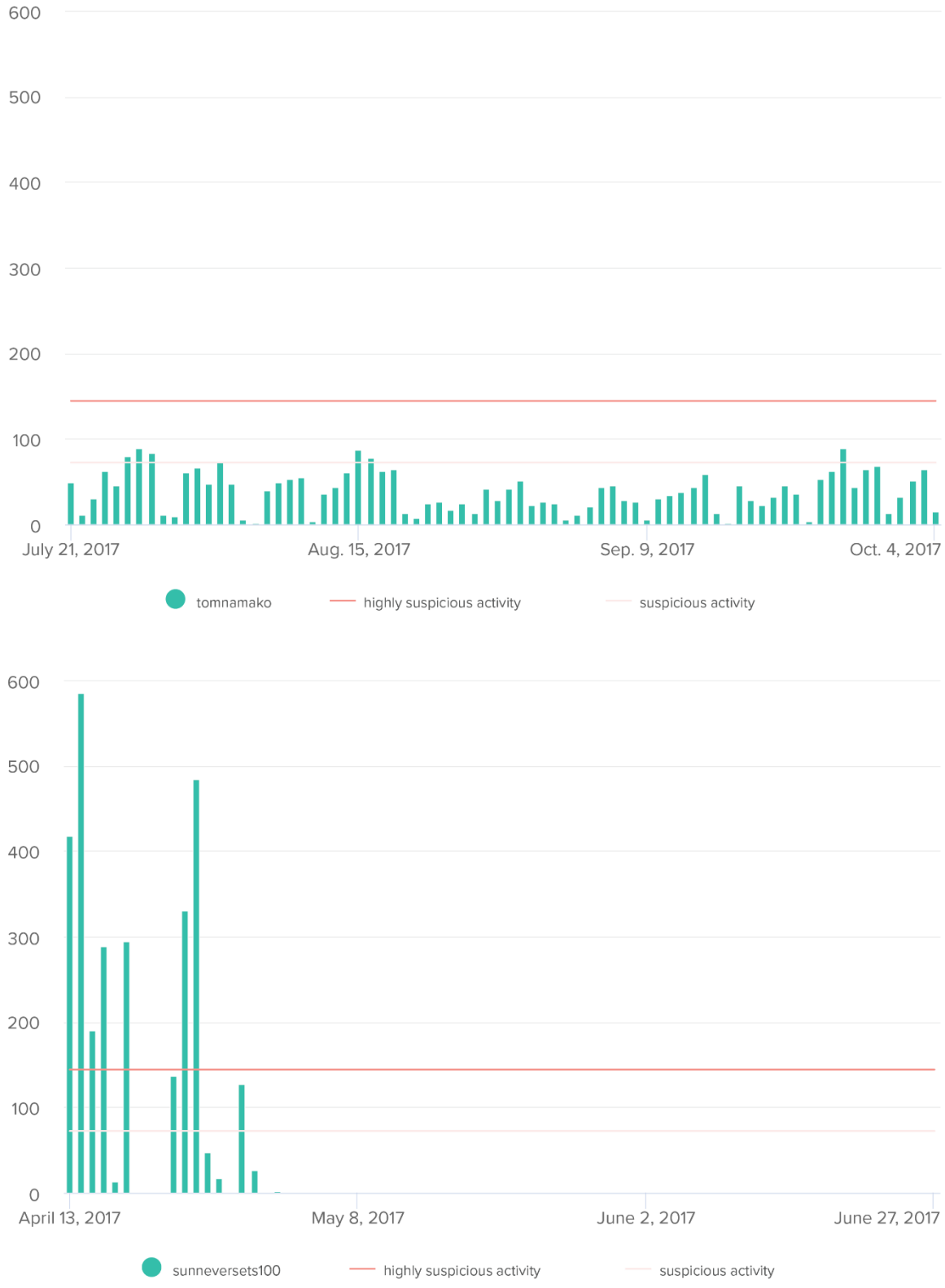


Figura 4: Comparação do *BuzzFeed News* de dados de um de seus editores no Twitter, @tomnamako, com os dados de diversas contas que apresentaram atividades semelhantes às de bots para destacar suas diferenças em personas e comportamento. O primeiro gráfico mostra que os últimos 2.955 tweets do jornalista se distribuem ao longo de vários meses. Sua contagem de tweets diários mal ultrapassa 72 publicações, que o Laboratório de Pesquisa em Ciência Forense Digital considera nível de atividade

suspeito. O segundo gráfico mostra os últimos 2.955 tweets do bot. Com um número suspeito e contínuo, o robô chegou a disparar 584 postagens em um dia. E, então, parou abruptamente.

Como coletar dados sociais

Há três maneiras bem diferentes de coletar dados de redes sociais: APIs, arquivos pessoais e raspagem.

Os dados fornecidos por canais oficiais como APIs são bastante limitados. Por mais que coletem dados em volumes gigantescos, as empresas de redes sociais oferecem só um pouquinho destes através de suas APIs (no caso do Facebook, pesquisadores já conseguiram informações de páginas e grupos públicos, mas isso não é mais possível desde que a empresa implementou restrições na disponibilidade deste tipo de informação em reposta à Cambridge Analytica. No caso do Twitter, este acesso geralmente se restringe a número predefinido de tweets da linha do tempo de um usuário ou período predeterminado por busca).

Além disso, existem limitações quanto ao tipo de dados que usuários podem solicitar sobre sua persona e comportamento online. Alguns serviços, como Facebook e Twitter, permitem aos usuários que baixem um histórico dos dados que constituem seu eu online — postagens, mensagens ou fotos de perfil —, mas este arquivo nem sempre incluirá todos os dados que as empresas têm sobre eles.

Por exemplo, usuários só podem visualizar os anúncios nos quais clicaram até três meses atrás, complicando a tarefa de determinar se clicaram ou não em uma postagem patrocinada pela Rússia.

Por fim, a extração de dados de redes sociais das plataformas por meio de raspagem, muitas vezes, vai contra seus termos de serviço. A raspagem de dados de uma rede social pode fazer com que o usuário seja banido e pode até mesmo levar a um processo legal.[167]

No caso das plataformas, pode fazer sentido processar seus usuários por raspagem, do ponto de vista financeiro. Grande parte das informações reunidas por estas plataformas a respeito de seus usuários está à venda, não diretamente, mas outras empresas e anunciantes podem lucrar com isso por meio de anúncios e marketing. Rivais poderiam raspar informações do Facebook para criar uma plataforma semelhante, por exemplo. Tais processos podem não apenas impedir gente em busca de ganhos financeiros, mas também acadêmicos e jornalistas que desejam obter informações de plataformas para fins de pesquisa.

---

[167] https://caselaw.findlaw.com/summary/opinion/us-9th-circuit/2016/07/12/276979.html.

Isso significa que jornalistas precisam ser mais criativos em seu trabalho e em como contam histórias, talvez queiram comprar bots para entender como funcionam na rede ou comprar anúncios no Facebook para entender como a plataforma funciona.[168]

Independente do meio, operar dentro e fora das restrições definidas por empresas de redes sociais será um grande desafio para os jornalistas circulando por este ambiente cibernético em constante mutação.

Para o que *não* servem os dados de redes sociais

É imperativo compreender melhor o universo de dados extraídos de redes sociais a partir de suas ressalvas.

Entendendo quem *usa* e quem *não usa* redes sociais

Um dos grandes problemas dos dados de redes sociais é que não podemos presumir que as pessoas que vemos no Twitter ou no Facebook são amostras representativas de grandes públicos fora da internet.

Muita gente tem conta no Twitter ou no Facebook, mas jornalistas não deveriam crer que as opiniões expressas online refletem a população em geral. Como uma pesquisa da Pew de 2018 ilustra bem, a utilização das redes sociais varia com cada plataforma.[169] Mais de dois terços dos usuários adultos de internet dos EUA estão no YouTube e no Facebook, mas menos de um quarto usa o Twitter. Este tipo de informação pode ser muito mais eficaz no contexto de uma história específica, seja uma observação do discurso de ódio disseminado por políticos em Mianmar ou o tipo de cobertura jornalística oferecida pelo veículo conspiracionista *Infowars* ao longo do tempo.

Nem todo usuário representa um ser humano de verdade

Além do que, nem todo usuário corresponde a uma pessoa. Há contas automatizadas (bots) e contas semiautomatizadas, bem como contas semicontroladas por pessoas (ciborgues) e usuários que atuam em diversas contas.

Mais uma vez, entender que há um mundo de atores lá fora manipulando o fluxo de informação para fins econômicos ou políticos é um aspecto importante a se ter em mente ao observar dados de redes sociais em grandes volumes (por mais que este tema, de manipulação da mídia e informação, tenha se tornado um grande tema por si só, que jornalistas tentam desvendar de maneiras cada vez mais sofisticadas).

---

[168] https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters.

[169] Smith e Anderson (2018).

A tirania daqueles que falam mais alto

Não menos importante do que tudo que já foi discutido é o ato de reconhecer que nem sempre o comportamento de tudo ou todos é mensurado. Há quem prefira ficar em silêncio. E com o desaparecimento de vozes moderadas, apenas as reações extremas são computadas e alimentam algoritmos que amplificam de maneira desproporcional a prominência daqueles que falam mais alto.

Ou seja, o conteúdo que vêm à tona em plataformas como Facebook, Twitter e outras mais em nossos feeds, se baseia nas curtidas, retweets e comentários daqueles que optaram por participar da conversa. Quem não se manifestou acaba sumindo no processo. Logo, precisamos levar em conta o que não é mensurado, assim como fazemos com o que é mensurado, e como a informação é categorizada e emerge como resultado destes dados mensurados e não mensurados.

Lam Thuy Vo é uma jornalista que conta histórias baseadas em dados no BuzzFeed News e ensina outros jornalistas a usarem tecnologia a favor da coleta de informações e narrativa na Escola de Graduação em Jornalismo Craig Newmark, na Universidade da Cidade de Nova York.

Referências

ANDERSON, Monica; SMITH, Aaron. *Social Media Use in 2018*. Pew Research Centre, 1º de março de 2018.

Aplicativos e suas *affordances* para investigações com dados

Esther Weltevrede

Há pouco, a Netvizz, ferramenta usada para extração de dados do Facebook, deixou de ter acesso à funcionalidade de Acesso a Conteúdo Público de Página da plataforma, fato que aparentemente encerrou a já precária relação que o desenvolvedor do software, o pesquisador de métodos digitais Bernhard Rieder, mantinha com a API do Facebook nos últimos nove anos.[170] O fim da Netvizz é sintomático de uma mudança mais significativa em pesquisa digital e investigações onde plataformas cada vez mais restringem coleta de dados por meio de suas APIs (Interfaces de Programação de Aplicações) e Políticas de Desenvolvedor. Por mais que a eficácia dos métodos utilizados pela Cambridge Analytica sejam questionáveis (Smout e Busvine, 2018; Lomas, 2018), o escândalo em torno da empresa levou ao debate em torno da privacidade e proteção de dados em redes sociais, e a reação do Facebook foi restringir ainda mais o acesso a dados em suas plataformas.

Desde seu anúncio inicial em março de 2018,[171] a implementação vacilante de restrições no acesso a dados por parte do Facebook em sua linha de aplicativos acabou por tornar visível a vasta rede de terceiros que haviam passado a depender da plataforma para os mais variados fins. Aplicativos deixaram de funcionar, houve restrição em anúncios direcionados, por exemplo, mas os mais afetados foram os pesquisadores digitais, já que aplicativos cuja principal função é coletar dados não são mais permitidos. Resistindo a estas mudanças (Bruns, 2018), pesquisadores argumentavam que tudo isso representaria um prejuízo à pesquisa para fins de interesse público. Entre as referências ao artigo sobre a Netvizz (Rieder, 2013) constavam mais de 450 publicações, número que na realidade era bem maior — basta considerar o volume de projetos de estudantes que utilizavam a ferramenta. De maneira semelhante, um inventário de Bechmann a respeito de estudos que "não teriam como existir sem acesso a dados de API"[172] inclui uma lista impressionante de publicações nas áreas de jornalismo, ciências sociais e pesquisa digital.

Em uma reflexão sobre o impacto das restrições de acesso a dados dentro da pesquisa digital, diversos autores contextualizaram estes desdobramentos e classificaram a última

---

[170] http://thepoliticsofsystems.net/?s=netvizz.

[171] https://about.fb.com/news/2018/03/cracking-down-on-platform-abuse/.

[172] https://docs.google.com/document/d/15YKeZFSUc1j03b4lW9YXxGmhYEnFx3TSy68qCrX9BEI/edit.

década como "Pesquisa baseada em API" (Venturini e Rogers, 2019) ou "Pesquisa Relacionada a API" (Perriam et al., 2019), definindo-a pela abordagem em pesquisa digital baseada na extração de dados disponibilizados por plataformas online através de suas APIs. Certamente estas APIs, com informações já prontas para pesquisa social, facilitaram estudos feitos com base em dados obtidos em plataformas de redes sociais, além de dar a chance a uma geração de estudantes de conduzir experimentos em pesquisa digital. Não há a necessidade de conhecimento técnico aprofundado e, tratando-se de dados web, estes apresentam-se relativamente limpos, prontos para uso. Pesquisas com base em APIs também foram criticadas desde o princípio, mais notavelmente por conta das *affordances* destas plataformas, movidas por sua própria conveniência, afetando a atuação do pesquisador no desenvolvimento de questões de pesquisa relevantes (Marres, 2017).

Este capítulo aborda o clamor recente por "pesquisa pós-API" feito por Venturini e Rogers (2019) e a Digital Methods Initiative,[173] com foco nas oportunidades que surgem em resposta aos desdobramentos atuais ocorridos dentro do ecossistema das redes sociais. A pesquisa digital, no sentido utilizado ao longo deste texto, é definida pelo princípio metodológico de "seguir o meio", reações e métodos de interação com o que acontece no ambiente digital. Nos parágrafos seguintes, discuto as restrições de API recentes ao argumentar em prol da necessidade renovada e potencial de explorações criativas, inovadoras de diferentes tipos de dados sociotécnicos essenciais para modelagem dos atuais ambientes de plataforma. A seguir, comento as oportunidades já identificadas por pesquisadores digitais, somando a isso a proposta de uma perspectiva metodológica para o estudo de relações app-plataforma. Com isso, espero dar aos jornalistas de dados interessados no potencial de dados sociais para storytelling (recomendo consultar o capítulo de Thuy Vo neste volume) alguns pontos de partida para abordagem de investigações com e sobre plataformas e seus dados no momento pós-API que vivemos.

O que os métodos digitais têm em comum é o fato de usarem uma série de técnicas de coleta e análise de dados que otimiza o emprego de formatos nativos de dados digitais emergentes da introdução das mídias digitais à vida social. Pesquisadores de métodos digitais desenvolvem ferramentas inspiradas nestas mídias para poderem lidar com estes formatos de dados de maneiras inovadoras, dentro de uma perspectiva metodológica. Desta forma, a história dos métodos digitais pode ser encarada como a narrativa de formatos e estruturas de dados essenciais da internet; adaptam-se às mudanças da mídia e incluem estas em suas análises. A seguir, gostaria de contribuir com as abordagens de pesquisa pós-API ao propor o estudo de plataformas como infraestruturas de dados a partir de uma perspectiva app-plataforma. O impacto das restrições de acesso a dados no ecossistema mais amplo da mídia atesta o fato de que o conhecimento avançado, detalhado, sobre infraestruturas destas

[173] https://wiki.digitalmethods.net/Dmi/WinterSchool2020.

plataformas e sua relação com apps de terceiros é de extrema necessidade. Tais acontecimentos demonstram a necessidade de maior letramento a respeito de infraestrutura de dados (Gray et al., 2018), além de conhecimento sobre como empresas e apps de terceiros operam em ambientes de redes sociais.

Aplicativos e plataformas enquanto infraestrutura

As restrições de dados de plataformas são parte integral dos desdobramentos das redes sociais rumo ao formato de plataforma enquanto infraestrutura, o que destaca a evolução do foco destes ecossistemas digitais em termos de parcerias corporativas (Helmond et al., 2019). Após um ano de cobertura negativa na esteira do papel da plataforma nas eleições norte-americanas, Zuckerberg postou uma nota onde delineava a mudança de perspectiva do Facebook de "conectar pessoas" para a construção de "uma infraestrutura social" (Hoffmann et al., 2018; Helmond et al., 2019).[174] O conceito de infraestrutura social destaca tanto as atividades sociais enquanto principal produto da plataforma na conexão e geração de valor para o mercado de diversos ângulos como a mudança da empresa, antes uma rede social, agora uma infraestrutura de dados para além da plataforma, que inclui seus sites e outros 70 apps (Nieborg e Helmond, 2018).[175] Esta mudança marca um novo passo do Facebook e sua capacidade de extensão de infraestrutura de dados a apps, plataformas e websites de terceiros, bem como uma maior facilidade em integrações internas.

Por mais que o conceito de plataformas enquanto infraestrutura venha recebendo cada vez mais atenção (Plantin et al., 2018), como ocorre com aplicativos individuais, a forma como estes apps operam em e entre infraestruturas de dados é um campo pouco estudado e muitas vezes deixado de lado. Ainda, estes apps continuam a transformar e valorizar práticas cotidianas dentro do ambiente destas plataformas. Levo em consideração uma definição relacional do que são apps ao focar em terceiros, definidos como aplicações construídas sobre uma plataforma por desenvolvedores externos, não operadores ou de posse da mesma. Quando um app se conecta a uma plataforma, ele recebe acesso a suas funções e dados, a depender das permissões concedidas. Apps também permitem que as partes interessadas, como lojas de aplicativos, anunciantes e usuários, integrem e agreguem valor a estes de maneiras simultâneas e múltiplas. Em outras palavras, apps possuem esta tendência natural de se relacionarem com, e serem relacionados a, diferentes infraestruturas operativas de dados. A posição específica de apps de terceiros os torna especialmente apropriados para nossos estudos voltados aos ambientes de plataforma enquanto infraestrutura.

---

[174] https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/.

[175] https://www.appannie.com/company/1000200000000034/.

Plataformas de redes sociais oferecem desafios metodológicos, visto que, como mencionado anteriormente, o acesso aos dados gerados por usuários é cada vez mais limitado, colocando pesquisadores em situação de considerar o que há de novo em "dados sociais" e criar espaço para perspectivas alternativas. Diferente da forma como plataformas de redes sociais fornecem acesso a dados gerados por usuários para pesquisa digital, de maneira estruturada via APIs, as fontes de dados de apps caracterizam-se cada vez mais por serem fechadas ou de natureza proprietária. Por mais que ofuscação seja uma técnica largamente utilizada na engenharia de software (Matviyenko et al., 2015), esforços para tornar códigos e dados ilegíveis ou inacessíveis têm impacto significativo em pesquisa digital. Desafios crescentes impostos por ambientes de plataformas e apps que dificultam pesquisa empírica, aquilo que colegas e eu chamamos de "resistência infraestrutural" (Dieter et al., 2018). No lugar disso, os dados disponíveis para pesquisa digital hoje caracterizam-se por formatos heterogêneos, originados em dispositivos (como GPS), bibliotecas de software (kits de desenvolvimento de software, SDKs na sigla em inglês) e conexões de rede (redes de anúncios). Apps podem coletar dados gerados por usuários, mas normalmente não oferecem acesso via APIs abertas, logo, faltam informações imediatamente disponíveis para maiores investigações.

A seguir, apresento três diferentes explorações de dados feitas de baixo para cima onde pesquisadores e jornalistas puderam empregar diferentes "*affordances* de pesquisa" (Weltevrrede, 2013), usando-as no avanço ou no início de uma linha de inquérito. Estas *affordances* de pesquisa estão em sintonia com possíveis ações dentro do software da perspectiva dos interesses do pesquisador e alinham com estes. Esta abordagem permite o desenvolvimento de métodos digitais inovadores (Rogers, 2013; Lury e Wakeford, 2012), que exigem repensar formatos e formas técnicas das relações entre apps e plataformas ao explorar suas oportunidades analíticas. Estas explorações tomam por base pesquisas recentes nos quais eu e outros colegas trabalhamos, servindo como inspiração e um lembrete dos desafios para a pesquisa, e indicando caminhos quanto ao tipo de solicitações que estas fontes podem atender para que tenhamos melhor compreensão do ambiente de plataforma enquanto infraestrutura.

Infraestruturas sociais falsas

A primeira exploração tratou da questão de seguidores falsos e sua relação com a infraestrutura social do Facebook. Há maior atenção por parte das plataformas e pesquisadores digitais quanto a seguidores falsos em ambientes de redes sociais. Do ponto de vista das plataformas, o mercado de seguidores falsos geralmente é excluído de discussões que tratam estas plataformas como mercados multilaterais; o mercado de seguidores falsos não é encarado como um aspecto destes ambientes, muito menos como parte da "família". Seguidores falsos estabelecem uma infraestrutura não oficial de relações, reconhecida pelas plataformas como uso incorreto e indesejável. Não é algo que as plataformas querem, mas

que funciona junto e em sintonia com os mecanismos da plataforma, por virtude destes. Além do que, tais práticas diminuem o valor do produto principal aqui, a atividade social.

Outros colegas e eu analisamos os preparativos para o Brexit no Twitter ao focar nos apps mais usados naquele conjunto de dados (Gerlitz e Weltevred, 2019; Figura 1). Uma análise sistemática daqueles aplicativos e suas funcionalidades oferece informação sobre os mecanismos de engajamento automatizado e falsos dentro da estrutura de governança da plataforma em questão. Em um projeto contínuo junto a Johan Lindquist, estamos explorando um conjunto de mais de 1.200 plataformas que possibilitam a compra e venda de engajamentos nas mais diversas redes sociais. Estas apurações iniciais mostram como seguidores falsos se relacionam a estas plataformas de um ponto de vista técnico, com apps oficiais e de terceiros conectando-se via API e de uma infraestrutura de plataformas que se conectam de maneira não oficial a estas redes sociais. O que descobrimos com isto é que a pesquisa terá que se desdobrar para acomodar dados de fontes legítimas e falsas. Contas automatizadas e falsas não podem ser tratadas (somente) como atuantes neste processo, mas também como *prática*, situada e emergente em relação às *affordances* do meio. Como mostrado no caso do Twitter, uma conta não necessariamente representa um usuário humano, visto que existe de maneiras distribuídas e situadas, assim como um tweet não é apenas um tweet, geralmente compreendido como uma postagem criada unicamente (Gerlitz e Weltevrede, 2019).

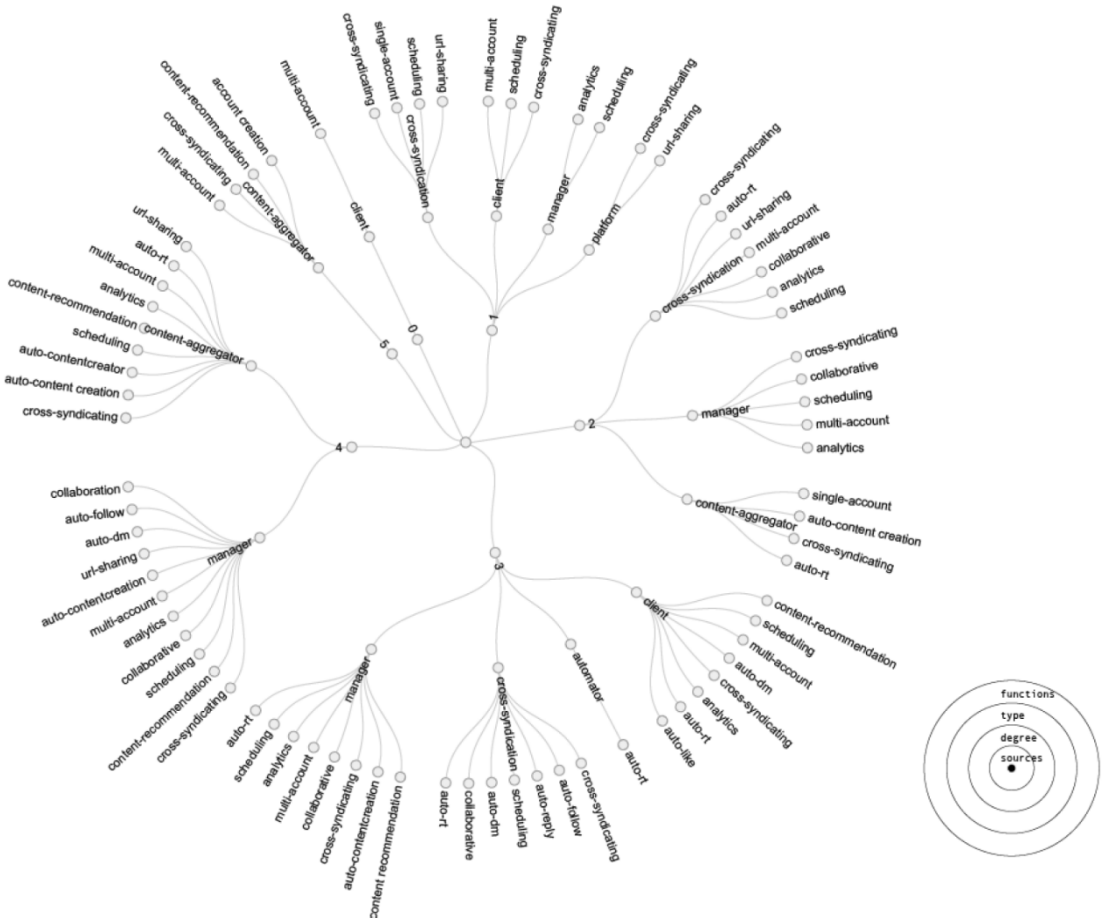


Figura 1: Funções de automação (Gerlitz e Weltevrede, 2019). O dendograma ilustra a hierarquia de fontes, graus de automação, tipos de fontes e suas funções na série de dados sobre o Brexit, de 17 a 23 de junho de 2016.

Relações app-plataforma

A segunda exploração considera lojas de apps como infraestruturas de dados para aplicativos. Hoje, o principal ponto de entrada para apps — considerando desenvolvedores e usuários — se dá pelas lojas de aplicativos, onde usuários podem buscar por apps específicos ou demarcarem coleções e gêneros. Com base em métodos de estudos algorítmicos (Sandvig et al., 2014; Rogers, 2013), um indivíduo pode engajar com a tecnicidade da "cultura de classificação" (Rieder et al., 2018) encontrada no Google Play e na App Store, por exemplo. Tal empreitada envolve poder econômico e algorítmico, bem como suas consequências sociais. Com base nisso, pode-se ganhar conhecimento sobre mecanismos de categorização e compreensão do porquê da importância destes na circulação de conteúdo cultural.

As lojas de aplicativos também podem ser usadas na demarcação de coleções ou gêneros de apps, possibilitando o estudo das relações app-plataforma da perspectiva dos próprios aplicativos. Em *Regramming the platforms* (Gerlitz et al., 2019), alguns colegas e eu

investigamos mais de 18.500 apps e as diferentes formas pelas quais os aplicativos se relacionam com funcionalidades e características de plataforma. Uma das principais descobertas deste estudo foi a de que desenvolvedores encontram maneiras criativas de lidar com as APIs oficiais das plataformas, por conseguinte o mesmo acontece com sistemas de governança oficiais destas plataformas (Tabela 1). A abordagem centrada em apps para a questão das plataformas enquanto infraestrutura permite um vislumbre sobre os apps de terceiros desenvolvidos na periferia das plataformas de redes sociais, práticas e funcionalidades suportadas e estendidas por estes aplicativos, e relações bagunçadas e contingentes entre estes softwares e as próprias redes sociais (Gerlitz et al., 2019).

Relation	[Facebook]	[Instagram]	[Snapchat]	[Twitter]
Brand (mentions)	1,449 (34.96%)	2,945 (80.03%)	614 (12.17%)	1,107 (21.80%)
Legal (mentions)	302 (7.29%)	318 (8.65%)	268 (5.31%)	305 (6.01%)
Technical (mentions)	61 (1.47%)	62 (1.68%)	70 (1.39%)	114 (2.24%)
Technical (libraries, SDKs)*	83 (33.20%)	0 (0%)	0 (0%)	40 (16.13%)
Technical (HTTP requests)*	156 (62.40%)	89 (35.60%)	12 (4.80%)	102 (41.13%)

* Only for Google Play search results (N = 998).

Tabela 1: Relações app-plataforma detectadas por conjunto de fontes (Gerlitz et al., 2019). Os indicadores técnico (bibliotecas, SDKs) e de inteligência (solicitações de HTTP) indicam que aplicativos mantêm uma relação "oficial" com as redes sociais através de suas APIs.

Conexões de dados de terceiros

A terceira exploração considera aplicativos e como estes se relacionam com as infraestruturas de dados de interessados externos. Através deste tipo de investigação é possível mapear como o app enquanto objeto de software embute infraestruturas externas de dados e fluxos dinâmicos de dados, entrando e saindo (Weltevrede e Jansen, 2019). Apps chegam até nós como objetos discretos e limitados, ao mesmo tempo que por definição são objetos de dados infraestruturais, ligados a plataformas para extensão e integração com a infraestrutura de dados destas.

De forma a ativar e explorar o fluxo de dados entrando e saindo, usamos uma variação do "método passo a passo" (Light et al., 2016). Com foco nas conexões de dados, a visualização resultante mostra quais dados são canalizados para dentro de apps a partir das plataformas de redes sociais e da plataforma mobile (Figura 2). Em outro momento,

mapeamos as redes de anúncios, serviços de nuvem, analítica e demais redes de terceiros usadas pelos apps para monetização de dados, melhoria de funcionalidades ou distribuição de hospedagem para entidades externas, dentre outras funções (Figura 3). Mapear o fluxo que entra e sai destes apps oferece uma visão essencial da economia política da circulação e recombinação de dados: as conexões estabelecidas, como são acionadas e que tipos de dados são transferidos a quem.

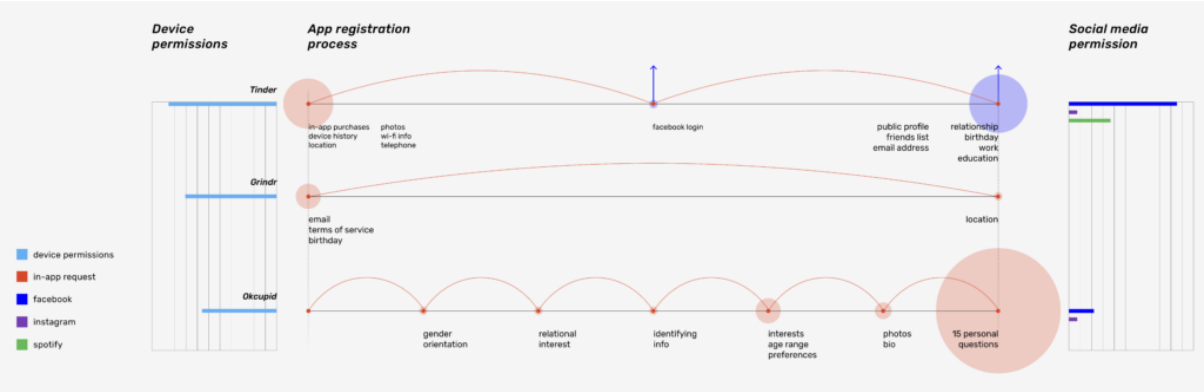


Figura 2: Passo a passo de fluxos de dados durante processo de registro (Weltevrede e Jansen, 2019).

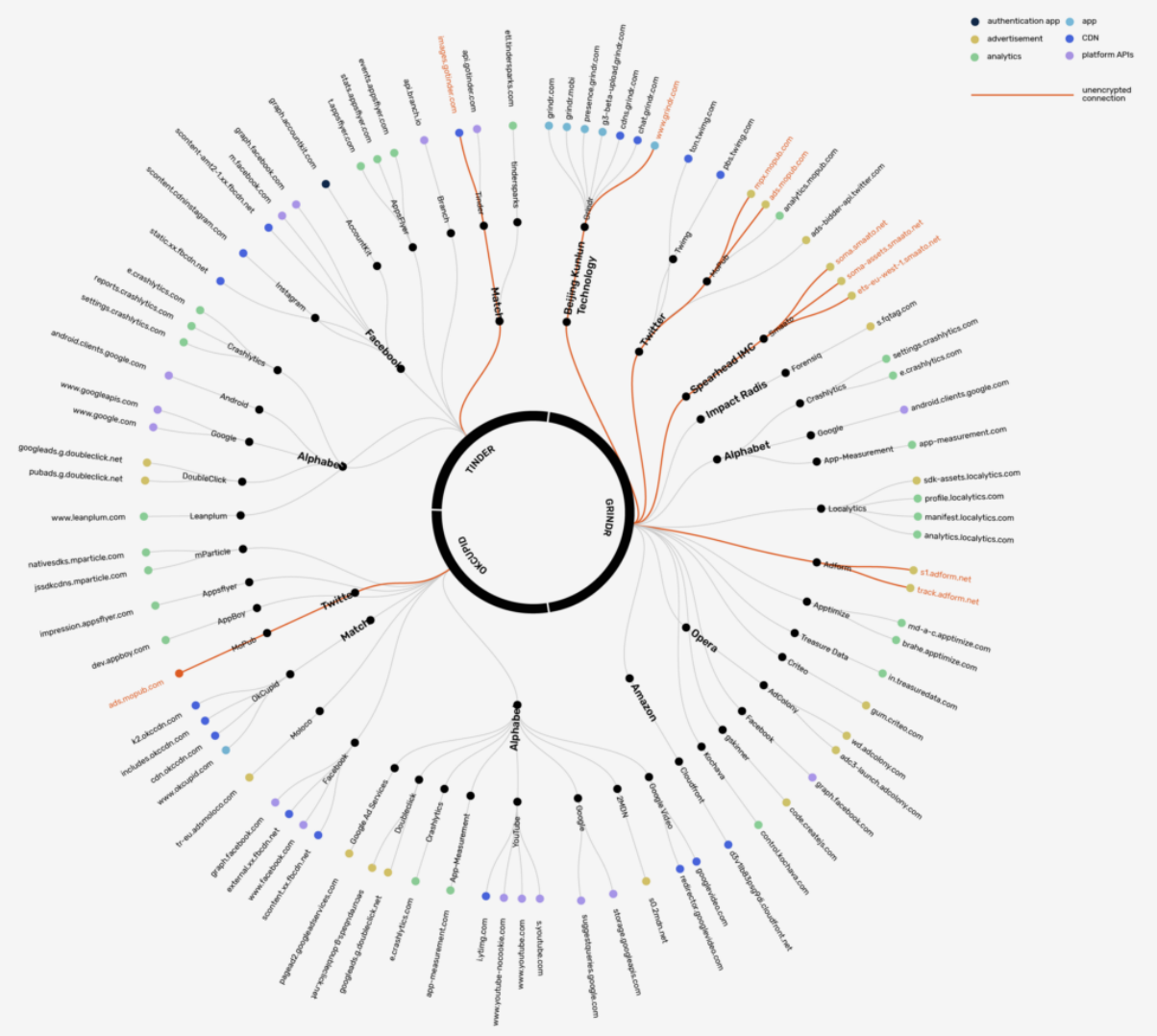


Figura 3: Conexões de rede estabelecidas entre os aplicativos de relacionamento Tinder, Grindr e OKCupid e terceiros (Weltevrede e Jansen, 2019).

Conclusão

Plataformas e aplicativos fazem parte do nosso cotidiano a ponto de ninguém parar para pensar nestes. Esta tendência de ficar nos bastidores é o motivo pelo qual pesquisadores digitais, jornalistas de dados e ativistas deveriam explorar a forma como eles operam e as condições que sustentam sua criação e uso. É importante aprimorar o entendimento quanto à infraestrutura de dados para que possamos entender como isso se relaciona a diferentes plataformas e redes, como operam entre si e envolvem outras partes interessadas desconhecidas.

Após o escândalo envolvendo a Cambridge Analytica, cada vez mais as plataformas restringem o acesso via API a dados já prontos para investigações de cunho social em

resposta à pressão pública. Neste capítulo, sugeri que pesquisadores, jornalistas e grupos da sociedade civil poderiam reagir através de criativas e inovadoras formas de exploração de dados em termos de *affordances* para investigações de dados. Explorei três tipos de dados para investigar o funcionamento de apps. Além disso, há diversas oportunidades para expandir este tema. Deve-se reforçar que, ao longo deste processo, abordei principalmente os aplicativos, mas esta pesquisa pode servir de inspiração para investigação de outros ambientes ricos em dados, como cidades inteligentes e a internet das coisas. Um entendimento mais profundo das infraestruturas de dados que cada vez mais moldam nosso cotidiano segue como um projeto contínuo.

Esther Weltevrede é professora assistente de Novas Mídias e Cultura Digital e coordenadora da Digital Methods Initiative na Universidade de Amsterdã, onde explora affordances de pesquisa em mídias digitais, com interesse específico em inovações metodológicas no estudo de infraestruturas de plataformas de redes sociais e terceiros.

Referências

BRUNS, Axel. *Facebook Shuts the Gate after the Horse Has Bolted, and Hurts Real Research in the Process*. Internet Policy Review, 25 de abril de 2018. Disponível em: https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786.

DIETER, Michael et al. *Multi-situated app studies: Methods and propositions*. Social Media + Society, 2019.

GERLITZ Carolin et al. *Regramming the Platform: Infrastructural Relations between Apps and Social Media*. Computational Culture 7, 21 de outubro de 2019. Disponível em: http://computationalculture.net/regramming-the-platform/.

GERLITZ, Carolin; WELTEVREDE, Esther. *What Happens to ANT, and Its Socio-Material Grounding of the Social, in Digital Sociology?*. In: BLOK, Anders; FARÍAS, Ignacio; ROBERTS, Celia (ed.). *Companion to Actor-Network Theory*. Londres e Nova York: Routledge, 2019.

GRAY, Jonathan; GERLITZ, Carolin; BOUNEGRU, Liliana. *Data infrastructure literacy*. Big Data e Society 5.2, 2018.

HELMOND, Anne; NIEBORG, David B; VAN DER VLIST, Fernando N. *Facebook's evolution: development of a platform-as-infrastructure*. Internet Histories, 3:2, 2019, p. 123-146.

HOFFMANN, A. L.; PROFERES, N.; ZIMMER, M. *Making the world more open and connected: Mark Zuckerberg and the discursive construction of Facebook and its users*. New Media e Society, 20(1), 2018, p. 199–218.

LIGHT, Ben; BURGESS, Jean; DUGUAY, Stefanie. *The Walkthrough Method: An Approach to the Study of Apps*. New Media e Society, 20(3), 2016, p. 881–900.

LOMAS, Natasha. *Kogan: 'I Don't Think Facebook Has a Developer Policy That Is Valid'*. TechCrunch, 2018. Disponível em: https://techcrunch.com/2018/04/24/kogan-i-dont-think-facebook-has-a-developer-policy-that-is-valid/.

LURY, Celia; WAKEFORD, Nina. *Inventive Methods: The Happening of the Social*. Londres e Nova York: Routledge, 2012.

MARRES, Noortje. *Digital Sociology: The Reinvention of Social Research*. Maden: Polity Press, 2017.

MATVIYENKO, Svitlana; CLOUGH, Patricia T. *On Governance, Blackboxing, Measure, Body, Affect and Apps: A conversation with Patricia Ticineto Clough and Alexander R. Galloway*. The Fibreculture Journal, (25), 2015, p. 10–29.

NIEBORG, David; HELMOND Anne. *The political economy of Facebook's platformisation in the mobile ecosystem: Facebook Messenger as a platform instance*. Media, Culture e Society, 41(2), 2019, p. 196–218.

PERRIAM, Jessamy; BIRKBAK, Andreas; FREEMAN, Andrew. *Digital methods in a post-API environment*. Artigo preliminar, 2019.

PLANTIN, Jean-Christophe et al. *Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook*. New Media e Society 20, nº 1, 2018, p. 293–310.

RIEDER, Bernhard. *Studying Facebook via data extraction: the Netvizz application*. In: *Proceedings of the 5th Annual ACM Web Science Conference on 'WebSci'13*. Paris: ACM Press, 2013, p. 346–355.

RIEDER, Bernhard; MATAMOROS-FERNÁNDEZ, Ariadna; COROMINA, Òscar. *From Ranking Algorithms to 'Ranking Cultures:' Investigating the Modulation of Visibility in YouTube search Results*. Convergence 24.1, 2018, p. 50-68.

ROGERs, Richard. *Digital Methods*. Cambridge: The MIT Press, 2013.

SANDVIG, Christian et al. *Auditing algorithms: Research Methods for Detecting Discrimination on Internet Platforms*. Data and Discrimination: Converting critical concerns into productive inquiry, 2014, p. 1-23.

SMOUT, Alistair; BUSVINE Douglas. *Researcher in Facebook Scandal Says: My Work Was Worthless to CA*. 2018. Disponível em: https://www.reuters.com/article/us-facebook-privacy-cambridge-analytica/researcher-in-facebook-scandal-says-my-work-was-worthless-to-cambridge-analytica-idUSKBN1HV17M.

WELTEVREDE, Esther; JANSEN, Fieke. *Infrastructures of Intimate Data: Mapping the Inbound and Outbound Data Flows of Dating Apps*. Computational Culture 7, 21 de outubro de 2019. Disponível em: http://computationalculture.net/infrastructures-of-intimate-data-mapping-the-inbound-and-outbound-data-flows-of-dating-apps/.

Jornalismo aplicado a algoritmos: métodos e pontos de vista investigativos

Nicholas Diakopoulos

A série *Machine Bias* da *ProPublica* teve início em maio de 2016, um esforço para investigar algoritmos e sua atuação na sociedade.[176] Talvez o que mais tenha chamado atenção ao longo da sua publicação tenha sido a investigação e análise expondo o viés racial de algoritmos de avaliação de risco de recidivas utilizadas em decisões criminais.[177] Estes algoritmos conferem pontuações a pessoas para determinar os riscos de estas cometerem crimes novamente. Estados e municípios usam estas pontuações para definir a necessidade de detenção antes do julgamento, liberdade vigiada, condicional, e, por vezes, até mesmo sentenças. Jornalistas da *ProPublica* pediram acesso às pontuações do Condado de Broward, na Flórida, e compararam estas informações com históricos criminais para saber se um indivíduo havia reincidido dentro do período de dois anos. A análise dos dados mostrou que réus negros tendiam a receber pontuações mais altas que réus brancos, com maior chance de serem categorizados como de alto risco, mesmo não sendo presos novamente após dois anos.[178]

Este sistema de pontuação da justiça penal é só mais um domínio em que algoritmos estão sendo implementados na sociedade. Os artigos da série *Machine Bias* vêm, desde então, tratando de assuntos como os anúncios direcionados do Facebook, taxas de seguro automotivo discriminatórias com base em localização geográfica e práticas injustas de precificação na Amazon. Tomadas de decisões baseadas em algoritmos cada vez mais integram os setores público e privado. Vemos isso em áreas como avaliação de risco em seguros e crédito, sistemas de emprego, gestão de benefícios, categorização educacional e de professores, curadoria de mídia online, entre outras.[179] Operando em larga escala e, muitas vezes, impactando grandes volumes de pessoas, estes algoritmos podem tomar decisões de cálculo, classificação, categorização, associação e filtragem consequenciais e, por vezes, contestáveis. Algoritmos, movidos por pilhas e mais pilhas de dados, são uma nova forma de exercer poder na sociedade.

---

176 https://www.propublica.org/series/machine-bias.

177 https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

178 https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm/.

179 O'Neil (2016), Pasquale (2015) e Eubanks (2018).

Como atestado pela série, um novo tipo de jornalismo computacional e de dados surge para investigar e responsabilizar a forma como o poder é exercido através de algoritmos. Chamo isso de "reportagem de prestação de contas algorítmica", uma reorientação da tradicional função de sentinela do jornalismo, considerando exercícios de poder com algoritmos.[180] Apesar de sua objetividade ostensiva, algoritmos não estão livres de erros e vieses que exigem maior escrutínio. Aos poucos, aprende-se a lidar com eles como algo inerente à prática jornalística, em conjunto com habilidades técnicas, de forma a prover o escrutínio mencionado acima.

Há, é claro, diversos tipos de transparência ou prestação de contas algorítmicas que podem se dar em múltiplos fóruns além do jornalismo, caso de contextos políticos, legais, acadêmicos, ativistas ou artísticos.[181] Mas meu foco neste capítulo é a reportagem de prestação de contas algorítmica como empreitada jornalística independente que contribui para prestação de contas e transparência ao mobilizar a pressão pública. Pode-se encarar isso como complementar a outras medidas que, no final, podem contribuir para questões de transparência e prestação de contas, como desenvolvimento de regulamentação e normas legais, criação de instituições de auditoria na sociedade civil, elaboração de políticas eficazes de transparência, exibição de mostras de arte que promovam reflexão, e publicação de crítica acadêmica.

Ao decidir o que constitui um bom procedimento jornalístico, ajuda primeiro definir o que há nos algoritmos que é digno de nota. Em termos técnicos, algoritmos são sequências de ações feitas em ordem para solucionar determinado problema ou chegar a determinado resultado. Tratando-se de processos de informação, os resultados dos algoritmos são, tipicamente, decisões. O ponto crucial do poder algorítmico, muitas vezes, se resume à capacidade de computadores tomarem decisões rapidamente e em larga escala, possivelmente afetando um grande número de pessoas. Na prática, transparência e prestação de contas com algoritmos não se resumem ao lado técnico destes. Algoritmos devem ser compreendidos como combinações de tecnologia reunidas através do trabalho de pessoas, incluindo designers, operadores, proprietários e mantenedores em complexos sistemas sociotécnicos.[182] Prestação de contas, neste contexto, tem a ver com entender como estas pessoas exercem poder dentro e ao longo de um sistema, sendo elas as responsáveis por suas decisões. Muitas vezes, um algoritmo chega ao noticiário quando toma uma decisão "ruim". Pode ser que ele

---

[180] http://www.slate.com/articles/technology/future_tense/2013/08/words_banned_from_bing_and_google_s_autocomplete_algorithms.html, https://www.theatlantic.com/technology/archive/2013/10/rage-against-the-algorithms/280255/.

[181] https://samatt.github.io/algorithmic-disobedience/, https://socialmediacollective.org/reading-lists/critical-algorithm-studies/.

[182] Seaver (2017) e Ananny (2015).

tenha feito algo que não deveria ou não fez algo que deveria. Para fins jornalísticos, a significância e as consequências públicas de uma decisão ruim são essenciais. Quais os possíveis danos causados a um indivíduo ou à sociedade? Decisões ruins podem impactar indivíduos diretamente, ou, quando agregadas, podem reforçar vieses estruturais, dentre outros problemas. Podem, ainda, custar caro. Vamos ver como várias decisões ruins podem levar a notícias.

Perspectivas sobre algoritmos

Ao observar os desdobramentos da forma de se trabalhar com algoritmos em jornalismo ao longo dos anos e através de minhas próprias investigações, identifiquei pelo menos quatro pontos que parecem estar atrelados a muitos materiais jornalísticos sobre prestação de contas algorítmica: (1) discriminação e injustiça, (2) erros ou enganos em previsões ou classificações, (3) violação de normas legais e sociais, e (4) mau uso de algoritmos por parte de pessoas, intencionalmente ou não. Darei exemplos de cada um nas próximas subseções.

Discriminação e injustiça

Revelar casos de discriminação e injustiça é um tema bastante explorado na reportagem de prestação de contas algorítmica. O artigo da *ProPublica* que abre este capítulo é um exemplo impressionante de como um algoritmo pode levar a disparidades sistêmicas no tratamento de diferentes grupos de pessoas. Northpoint, a empresa que criou as pontuações de avaliação de risco (atual Equivant), argumentou que as pontuações eram precisas ao longo de todas as raças, logo, eram justas. Mas sua definição de justiça não levava em consideração o volume desproporcional de enganos que afetava pessoas negras. Histórias de discriminação e injustiça dependem da aplicação da definição de justiça, que pode refletir diferentes visões políticas.[183]

Também trabalhei em matérias que desvelam a injustiça causada por sistemas algorítmicos, em especial a dinâmica de preços do Uber em diferentes vizinhanças de Washington.[184] Com base em observações iniciais de diferentes tempos de espera e como estes mudavam com base no algoritmo de preço dinâmico do Uber, consideramos a hipótese de que diferentes vizinhanças teriam diferentes índices de qualidade de serviço (no caso, tempo de espera). Ao recolher sistematicamente amostras dos tempos de espera em diversos locais ao longo do tempo, percebemos que aqueles com maior população negra/parda tendiam a esperas mais longas, mesmo tendo em conta outros fatores de controle, como

[183] Lepri et al. (2017).

[184] https://www.washingtonpost.com/news/wonk/wp/2016/03/10/uber-seems-to-offer-better-service-in-areas-with-more-white-people-that-raises-some-tough-questions/.

renda, índice de pobreza e densidade populacional do bairro. É difícil atribuir esta injustiça diretamente ao algoritmo do Uber, já que outros fatores humanos afetam o sistema, como o comportamento e possíveis vieses de motoristas do aplicativo. Mas os resultados sugerem que, no todo, o sistema apresenta disparidades associadas à demografia.

Erros e enganos

Algoritmos também podem ganhar as manchetes ao cometerem erros ou enganos específicos em suas decisões de classificação, previsão ou filtragem. Tomemos como exemplo plataformas como Facebook e Google, que empregam filtros algorítmicos para reduzir a exposição a conteúdo prejudicial, como discurso de ódio, violência e pornografia. Isso pode ser relevante para a proteção de populações particularmente vulneráveis, caso de crianças, especialmente em produtos como o YouTube Kids, da Google, alardeado como seguro para o público infantil. Erros no algoritmo de filtragem do app têm valor noticioso porque estes significam que as crianças podem se deparar com conteúdo inapropriado ou violento.[185] Via de regra, algoritmos cometem dois tipos de engano: falsos positivos e falsos negativos. No contexto do YouTube Kids, um falso positivo seria se deparar com um vídeo classificado, por engano, como inapropriado, quando na verdade é adequado para o público infantil. Um falso negativo ocorre quando um vídeo classificado como adequado não é algo que você gostaria que crianças vissem.

Decisões de classificação impactam pessoas quando aumentam ou diminuem o tratamento negativo ou positivo que um indivíduo recebe. Quando um algoritmo escolhe alguém para ganhar sorvete por engano (tratamento positivo mais intenso), você não verá nenhuma reclamação dessa pessoa (mas quando outros descobrirem o ocorrido, podem falar que é uma situação injusta). Erros geralmente ganham manchetes quando levam a um tratamento negativo mais intenso para uma pessoa, tal como expor uma criança a conteúdo inapropriado em vídeo. Erros também são dignos de nota quando levam a um decréscimo no tratamento positivo recebido por um indivíduo, como quando alguém perde uma oportunidade. Tomemos como exemplo um consumidor elegível a diversos descontos e que nunca os recebe porque o algoritmo o excluiu. Por fim, erros também podem gerar interesse jornalístico quando diminuem atenção negativa justificada. Considere a ocasião em que um algoritmo de avaliação de risco categoriza um indivíduo de alto risco como de baixo risco, um caso de falso negativo. Por mais que isso seja ótimo para o indivíduo em questão, há maior risco para a segurança pública ao permitir que um criminoso siga livre.

---

[185] https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html?_r=0.

Violações de normas legais e sociais

Algoritmos de previsão podem testar os limites de normas legais ou sociais estabelecidas, levando a outras oportunidades e perspectivas para cobertura. Pense, por um momento, na possibilidade de difamação algorítmica.[186] Difamação pode ser definida como "uma declaração falsa de um fato que expõe um indivíduo ao ódio, constrangimento ou desonra, rebaixamento em relação a seus pares, causando isolamento ou prejuízo na condução de seus negócios ou comércio".[187] Ao longo de muitos anos vimos diversas histórias e batalhas legais se desenrolarem em torno de pessoas que acreditavam ter sido difamadas pelo algoritmo de preenchimento automático do Google. Este mesmo preenchimento automático pode ligar uma pessoa ou o nome de uma empresa a qualquer coisa: de crime e fraude à falência ou conduta sexual, com consequências à reputação. Algoritmos também podem ganhar as notícias quando infringem normas sociais como o direito à privacidade. O site *Gizmodo*, por exemplo, fez uma cobertura extensa do algoritmo "Pessoas que Você Talvez Conheça" (PVTC, daqui em diante) do Facebook, que sugere "amigos" em potencial dentro da plataforma, muitas vezes gente inapropriada ou indesejada.[188] Em um dos artigos, os repórteres chegaram a identificar uma situação em que o algoritmo revelou a identidade real de uma trabalhadora do sexo para seus clientes.[189] Isso é problemático não só por conta do estigma possivelmente associado ao trabalho sexual, pois envolve também o medo de alguns de seus clientes se tornarem *stalkers*.

Difamação e violações de privacidade são apenas duas possíveis perspectivas para se criar uma matéria jornalística aqui. Cabe aos próprios jornalistas detectarem outras violações legais ou sociais causadas por algoritmos nos mais variados contextos sociais. Já que estes algoritmos operam em uma versão quantificada da realidade que incorpora somente o mensurável por dados, podem acabar deixando de fora boa parte do contexto legal e social que seria essencial para tomada de decisões. Ao compreender o que determinado algoritmo quantifica sobre o mundo, como ele "vê" as coisas, é possível fazer uma crítica ao lançar luz sobre as partes faltantes que apoiariam uma decisão dentro do contexto integral.

Mau uso por humanos

Decisões algorítmicas muitas vezes integram processos mais amplos de tomada de decisão que envolvem todo um conjunto de pessoas e algoritmos trabalhando juntos em um

---

186 https://towcenter.org/algorithmic-defamation-the-case-of-the-shameless-autocomplete.

187 http://www.dmlp.org/legal-guide/defamation.

188 https://gizmodo.com/how-facebook-figures-out-everyone-youve-ever-met-1819822691.

189 https://gizmodo.com/how-facebook-outs-sex-workers-1818861596.

sistema sociotécnico. Apesar da inacessibilidade de alguns de seus componentes técnicos mais sensíveis, a natureza sociotécnica dos algoritmos abre novas oportunidades para investigar a relação que usuários, designers, proprietários e demais partes interessadas possam vir a ter com o sistema como um todo.[190] Se algoritmos são utilizados incorretamente pelas pessoas envolvidas no sistema sociotécnico, isso também é digno de nota. Quem cria algoritmos é capaz de, por vezes, antecipar e articular diretrizes para uma série razoável de contextos de uso de um sistema. Caso as pessoas ignorem isso tudo na prática, podemos ter um caso de negligência ou mau uso. O caso de avaliação de risco publicado pela *ProPublica* é um exemplo que salta aos olhos. A Northpointe, desenvolvedora responsável, havia criado duas versões e calibragens da ferramenta, uma para homens e outra para mulheres. Modelos estatísticos precisam ser treinados com dados que reflitam a população de onde serão utilizados, gênero sendo um fator relevante em previsão de recidivas. Mas o Condado de Broward estava utilizando o algoritmo do jeito errado: a pontuação de risco projetada e calibrada para homens também era utilizada em mulheres.[191]

Como investigar um algoritmo

Há vários caminhos a serem seguidos na investigação de uma força algorítmica, não há uma receita de bolo. Há, contudo, um arsenal crescente de ferramentas e técnicas à disposição, desde engenharia reversa e inspeção de código de programação, ambas altamente técnicas, até auditoria com base em dados coletados automaticamente ou de forma colaborativa, sem contar abordagens simples que consistem em fuçar e encarar de forma crítica reações geradas por algoritmos.[192] Cada projeto, artigo ou história pode exigir uma abordagem diferente a depender da perspectiva e do contexto, considerando fatores como grau de acesso ao algoritmo, seus dados, e se seu código foi disponibilizado. Por exemplo, um artigo revelador sobre discriminação sistemática pode se apoiar em um método de auditoria que usa dados coletados online, enquanto uma revisão de código pode se fazer necessária para verificar a implementação correta da política desejada.[193] O trabalho jornalístico tradicional envolve buscar fontes junto a empresas e profissionais como designers, desenvolvedores e cientistas de dados, bem como a solicitação de registros públicos e contato com indivíduos afetados pelo ocorrido. Tudo isso é mais importante do que nunca. Não tenho como me aprofundar em todos estes métodos neste capítulo tão curto, mas gostaria de falar um pouco mais sobre como jornalistas podem investigar algoritmos por meio de auditoria.

---

[190] https://www.cjr.org/tow_center/algorithms-reporting-algorithmtips.php.

[191] http://datastori.es/85-machine-bias-with-jeff-larson/.

[192] Para uma discussão mais completa sobre opções metodológicas, consulte Diakopoulos (2017).

[193] https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

Técnicas de auditorias vêm sendo empregadas há décadas no estudo de vieses sociais em setores como habitação, tendo sido adaptados recentemente para estudo de algoritmos.[194] A ideia básica consiste no seguinte: se os dados recebidos pelo algoritmo variam o suficiente e os resultados são monitorados, então pode-se fazer uma correlação entre o que entra e sai para que se teorize o funcionamento do algoritmo.[195] Se nos deparamos com violação dos resultados esperados após inserção de determinado dado ou informação, isso pode nos ajudar a tabular erros e determinar se há um viés sistemático atrelado a eles. No caso de algoritmos que fornecem acesso via APIs ou páginas na internet, resultados podem ser coletados automaticamente.[196] No caso de algoritmos personalizados, técnicas de auditoria se aliam à colaboração coletiva para coletar dados de uma série de pessoas que pode oferecer uma visão única do algoritmo. O AlgorithmWatch vem usando esta técnica na Alemanha de maneira eficaz em seu estudo da personalização de resultados de busca no Google, coletando quase 6 milhões de resultados de mais de 4.000 usuários, cujos dados foram compartilhados através de um plugin em seus navegadores (Christina Elmer discute esta iniciativa em seu capítulo neste livro).[197] Já o *Gizmodo* usou uma variação desta mesma técnica para investigar o PVTC, do Facebook. Neste caso, os usuários baixavam um software em seu computador que monitora, periodicamente, os resultados de PVTC, armazenando tudo localmente, protegendo, assim, a privacidade dos usuários. Os jornalistas poderiam, então, entrar em contato com usuários que acharam os resultados preocupantes ou surpreendentes.[198]

Auditoria de algoritmos, porém, não é para os fracos de coração. Déficits de informação limitam a capacidade do auditor até mesmo de saber por onde começar, o que perguntar, como interpretar resultados e como explicar os padrões observados no comportamento de um algoritmo. Há, ainda, o desafio em saber e definir o que se espera de um algoritmo, e como estas expectativas variam ao longo de contextos variados e de acordo com padrões e normas morais, sociais, culturais e legais pelo mundo. Esperam-se coisas diferentes quando falamos de justiça no contexto de um algoritmo de avaliação de risco criminal em comparação a um outro que cobra preços diferentes para pessoas diferentes por uma passagem de avião. Para identificar um erro ou viés digno de publicação, é preciso definir o que deveria ser o normal, sem quaisquer vieses. Por vezes, a definição vem de uma

---

[194] Gaddis (2014). Apresentação feita durante a pré-conferência da International Communication Association intitulada *Data and Discrimination Converting Critical Concerns into Productive Inquiry*.

[195] Diakopoulos (2015).

[196] https://www.wsj.com/articles/SB10001424127887323777204578189391813881534.

[197] https://datenspende.algorithmwatch.org/en/index.html.

[198] https://gizmodo.com/keep-track-of-who-facebook-thinks-you-know-with-this-ni-1819422352.

linha de base determinada por dados, caso de nossas auditorias de fontes de notícias em resultados de busca do Google durante as eleições dos EUA de 2016.[199]

A questão do acesso legal à informação sobre algoritmos também surge e, claro, depende bastante da jurisdição.[200] Leis de Acesso à Informação, nos EUA, definem o acesso do público a documentos do governo, mas a resposta quanto a algoritmos varia de agência para agência, sendo irregular, no melhor dos casos.[201] Talvez reformas legais pudessem facilitar o acesso à informação a respeito de algoritmos por parte do público. Caso a falta de informações, expectativas difíceis de serem articuladas e acesso legal incerto não sejam desafiadores o bastante, cabe lembrar que os próprios algoritmos costumam ser bastante caprichosos. Um algoritmo hoje pode ser diferente do algoritmo de ontem, o Google, por exemplo, costuma mudar seu algoritmo de busca entre 500 e 600 vezes por ano. Dependendo do que estas mudanças envolvem, pode ser necessário monitorar algoritmos ao longo do tempo para entender sua alteração e evolução.

Recomendações para seguir adiante

Para começar e aproveitar ao máximo a prática de reportagem de prestação de contas algorítmica, recomendaria três coisas. Primeiro, desenvolvemos uma fonte de informações chamada Algorithm Tips, que faz curadoria de metodologias relevantes, exemplos e demais recursos educacionais, além de abrigar um banco de dados de algoritmos para possíveis investigações (primeiramente cobrindo algoritmos de uso pelo governo dos EUA e, então, expandindo para outras jurisdições globais).[202] Caso busque fontes confiáveis de informação para aprender mais e para ajudar um projeto a sair do papel, pode ser um bom começo.[203] Segundo, foque nos resultados e impactos causados por algoritmos no lugar de tentar explicar o mecanismo exato utilizado na tomada de decisão. Identificar discriminação algorítmica (um exemplo de resultado) muitas vezes vale mais para a sociedade como passo inicial do que explicar exatamente como essa discriminação se deu. Ao focar em resultados, jornalistas podem dar um diagnóstico imediato e soar um alarme para outros interessados, que, por sua vez, poderão atuar junto a outras searas para fins de prestação de contas e responsabilização. Por fim, muito do que já vi feito na área e citei ao longo deste capítulo foi feito por equipes e

---

[199] Diakopoulos et al. (2018).

[200] https://www.aclu.org/other/data-journalism-and-computer-fraud-and-abuse-act-tips-moving-forward-uncertain-landscape.

[201] https://theconversation.com/we-need-to-know-the-algorithms-the-government-uses-to-make-important-decisions-about-us-57869.

[202] http://algorithmtips.org/.

[203] Trielli, Stark e Diakopoulos (2017).

há um bom motivo para isso. A prática eficaz da reportagem para fins de prestação de contas algorítmica exige todas as habilidades clássicas que um jornalista deve ter em sua prática, incluindo reportar e entrevistar, conhecimento dos procedimentos, solicitações de registros públicos e análise dos documentos recebidos, bem como escrita clara e cativante a respeito dos resultados. Tudo isso com apoio de novas técnicas, como raspagem e limpeza de dados, desenho de estudos de auditoria e uso de técnica estatística avançada. O conhecimento destas áreas pode ser distribuído entre uma equipe, ou colaboradores externos, contanto que exista comunicação clara, noção do assunto tratado e liderança. Desta forma, especialistas neste ou naquele método podem se juntar a outros especialistas para entender como o algoritmo exerce seu poder sobre uma maior variedade de campos sociais.

Nicholas Diakopoulos é professor assistente da Escola de Comunicação da Northwestern University, onde atua também como diretor do Laboratório de Jornalismo Computacional, e autor de "Automating the News: How Algorithms are Rewriting the Media".

Referências

ANGWIN, Julia et al. *Machine Bias*. ProPublica, maio de 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

ANANNY, Mike. *Toward an Ethics of Algorithms*. Science, Technology e Human Values 41 (1), 2015.

BHANDARI, Esha; GOODMAN, Rachel. *Data Journalism and the Computer Fraud and Abuse Act: Tips for Moving Forward in an Uncertain Landscape*. Computation + Journalism Symposium, 2017. Disponível em: https://www.aclu.org/other/data-journalism-and-computer-fraud-and-abuse-act-tips-moving-forward-uncertain-landscape.

BRAIN, Tega; MATTU, Surya. *Algorithmic Disobedience*._s/d. Disponível em: https://samatt.github.io/algorithmic-disobedience/#/.

BRAUNEIS, Robert; GOODMAN, Ellen. *Algorithmic Transparency for the Smart City*. 20 Yale Journal of Law e Technology, 103, 2018.

BUCHER, Taina. *If... Then: Algorithmic Power and Politics.* Oxford University Press, 2018.

DIAKOPOULOS, Nicholas et al. *I Vote For — How Search Informs Our Choice of Candidate*. In: MOORE, M; TAMBINI, D. (ed.). *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple.* Junho de 2018.

DIAKOPOULOS, Nicholas. *We need to know the algorithms the government uses to make important decisions about us*. The Conversation, maio de 2016. Disponível em: https://theconversation.com/we-need-to-know-the-algorithms-the-government-uses-to-make-important-decisions-about-us-57869.

DIAKOPOULOS, Nicholas. *Algorithmic Accountability: Journalistic Investigation of Computational Power Structures*. Digital Journalism 3 (3), 2015.

DIAKOPOULOS, Nicholas. *Sex, Violence, and Autocomplete Algorithms*. Slate, agosto de 2013. Disponível em: http://www.slate.com/articles/technology/future_tense/2013/08/words_banned_from_bing_and_google_s_autocomplete_algorithms.html.

DIAKOPOULOS, Nicholas. *Rage Against the Algorithms*. The Atlantic, outubro de 2013. Disponível em: https://www.theatlantic.com/technology/archive/2013/10/rage-against-the-algorithms/280255/.

DIAKOPOULOS, Nicholas. *Algorithmic Defamation: The Case of the Shameless Autocomplete*. Tow Center, agosto de 2013. Disponível em: https://towcenter.org/algorithmic-defamation-the-case-of-the-shameless-autocomplete.

DIAKOPOULOS, Nicholas. *Automating the News: How Algorithms are Rewriting the Media*. Harvard University Press, 2019.

DIAKOPOULOS, Nicholas. *Enabling Accountability of Algorithmic Media: Transparency as a Constructive and Critical Lens*. In: CERQUITELLI, Tania; QUERCIA, Daniele; PASQUALE, Frank (ed.). *Towards glass-box data mining for Big and Small Data*. Springer, junho de 2017, p. 25-44.

DIAKOPOULOS, Nicholas. *Algorithmic Accountability: Journalistic Investigation of Computational Power Structures*. Digital Journalism 3 (3), 2015.

EUBANKS, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

FINK, Katherine. *Opening the government's black boxes: freedom of information and algorithmic accountability*. Digital Journalism 17(1), 2017.

GADDIS, Steven M. *An Introduction to Audit Studies in the Social Sciences*. In: GADDIS, Michael (ed.). *Audit Studies Behind the Scenes with Theory, Method, and Nuance.* Springer, 2017, p. 3-44.

HILL, Kashmir. *How Facebook Figures Out Everyone You've Ever Met*. Gizmodo, novembro de 2017. Disponível em: https://gizmodo.com/how-facebook-figures-out-everyone-youve-ever-met-1819822691.

HILL, Kashmir. *How Facebook Outs Sex Workers*. Gizmodo, outubro de 2017. Disponível em: https://gizmodo.com/how-facebook-outs-sex-workers-1818861596.

HILL, Kashmir; MATTU, Surya. *Keep Track Of Who Facebook Thinks You Know With This Nifty Tool*. Gizmodo, janeiro de 2018. Disponível em: https://www.wsj.com/articles/SB10001424127887323777204578189391813881534.

LARSON, Jeff et al. *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica, maio de 2016. Disponível em: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm/.

LARSON, Jeff. *Machine Bias with Jeff Larson*. Podcast Data Stories, outubro de 2016. Disponível em: http://datastori.es/85-machine-bias-with-jeff-larson/.

LECHER, Colin. *What Happens When An Algorithm Cuts Your Health Care*. The Verge, março de 2018. Disponível em: https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

LEPRI, Bruno et al. *Fair, Transparent, and Accountable Algorithmic Decision-making Processes*. Philosophy e Technology, 84(3), 2017.

LEWIS, Seth C.; SANDERS, Kristin; CARMODY, Casey. *Libel by Algorithm? Automated Journalism and the Threat of Legal Liability*. Journalism e Mass Communication Quarterly 80(1), 2018.

MAHESHWARI, Sapna. *On YouTube Kids, Startling Videos Slip Past Filters*. New York Times, novembro de 2017. Disponível em: https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html?_r=0.

O'NEIL, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2016.

PASQUALE, Frank. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.

SANDVIG, Christian et al. *Auditing algorithms: Research methods for detecting discrimination on Internet platforms. Presented at International Communication*

Association preconference on Data and Discrimination Converting Critical Concerns into Productive Inquiry. 2014.

SEAVER, Nick. *Algorithms as culture: Some tactics for the ethnography of algorithmic systems*. Big Data e Society, 4(2), 2017.

STARK, Jennifer; DIAKOPOULOS, Nicholas. *Uber seems to offer better service in areas with more white people. That raises some tough questions*. Washington Post, março de 2016. Disponível em: https://www.washingtonpost.com/news/wonk/wp/2016/03/10/uber-seems-to-offer-better-service-in-areas-with-more-white-people-that-raises-some-tough-questions/.

TRIELLI, Daniel; DIAKOPOULOS, Nicholas. *How To Report on Algorithms Even If You're Not a Data Whiz*. Columbia Journalism Review, maio de 2017. Disponível em: https://www.cjr.org/tow_center/algorithms-reporting-algorithmtips.php.

TRIELLI, Daniel; STARK, Jennifer; DIAKOPOULOS, Nicholas. *Algorithm Tips: A Resource for Algorithmic Accountability in Government*. Computation + Journalism Symposium, outubro de 2017.

VALENTINO-DEVRIES, Jennifer; SINGER-VINE, Jeremy; SOLTANI, Ashkan. *Websites Vary Prices, Deals Based on Users' Information*'. Wall Street Journal, 24 de dezembro de 2012.

Algoritmos em destaque: investigações colaborativas no Spiegel Online

Christina Elmer

A demanda por transparência na questão dos algoritmos não é novidade na Alemanha. Já em 2012, o colunista Sascha Lobo, do *Spiegel Online*, reivindicou que a mecânica por trás do algoritmo de busca do Google fosse divulgada,[204] mesmo que isso causasse problemas à empresa. O motivo? Nossa visão de mundo pode ser moldada pelo Google através da função de preenchimento automático, por exemplo, como demonstrado em um caso famoso no país. Nesta ocasião, a esposa de um ex-presidente resolveu ir à corte contra o Google por conta da sugestão de termos problemáticos feitos pela função de preenchimento automático quando se buscava por seu nome. Dois anos depois, o Ministro de Justiça alemão repetiu este apelo, estendido novamente pela chanceler em 2016. De acordo com Angela Merkel, algoritmos deveriam ser mais transparentes.[205]

Nos últimos anos, a questão da transparência e prestação de contas de algoritmos vem sendo bastante discutida no *Spiegel Online*, inicialmente como oportunidade jornalística, não como um projeto próprio de pesquisa ou análise. Pode haver duas razões principais para a mídia alemã ter começado a experimentar nesta área após seus colegas nos EUA: por um lado, jornalistas alemães não têm os mesmos direitos e instrumentos de acesso à informação; por outro, a prática de jornalismo de dados não possui uma tradição de longa data como nos Estados Unidos. O *Spiegel Online* só foi ter um departamento dedicado ao setor em 2016, aos poucos expandindo-o. Claro que é possível para redações com menos recursos atuarem no campo, através de colaborações com outras organizações e freelancers, por exemplo. Em nosso caso, todos os projetos realizados no campo da reportagem de prestação de contas algorítmica foram feitos assim. Por isso, este capítulo se concentrará nestas colaborações e nas lições que aprendemos com elas.

---

204 http://www.spiegel.de/netzwelt/netzpolitik/google-suchvorschlaege-was-bettina-wulff-mit-mettigeln-verbindet-a-855097.html.

205 http://www.spiegel.de/wirtschaft/unternehmen/google-heiko-maas-fordert-offenlegung-von-algorithmus-a-991799.html.
http://www.spiegel.de/netzwelt/netzpolitik/angela-merkel-warum-die-kanzlerin-an-die-algorithmen-von-facebook-will-a-1118365.html.

Google, Facebook, Schufa — uma rápida observação destes três projetos

Nossa equipe editorial baseia-se principalmente em colaboração quando se trata da investigação de algoritmos. Na prévia das eleições federais de 2017, nos juntamos à ONG AlgorithmWatch para aprendermos mais sobre a personalização dos resultados de busca do Google.[206] Foi solicitado aos usuários que instalassem um plugin cuja função seria fazer pesquisas predefinidas em seus computadores regularmente. Cerca de 4.400 participantes doaram quase seis milhões de resultados de busca, fornecendo dados para uma análise que colocaria em xeque a tese da bolha, ao menos se tratando do Google e da área investigada.

Neste projeto, o pessoal do AlgorithWatch veio até nós do *Spiegel Online*, pois buscava um veículo parceiro de grande alcance para que pudesse fazer uma ação colaborativa e obter os dados necessários. Ao passo que toda a reportagem foi de responsabilidade de nosso departamento de internet e temas correlatos, a equipe de jornalismo de dados deu apoio no planejamento e à avaliação metodológica da operação toda. Além do que, o apoio de nosso departamento jurídico foi essencial para a implementação do projeto de maneira que fosse à prova de balas, em termos legais. Por exemplo, havia de se esclarecer questões de proteção de dados em meio à reportagem, de maneira que todos os participantes do projeto pudessem compreender.

Quase que simultaneamente, o *Spiegel Online* colaborou com a *ProPublica* no lançamento de seu AdCollector na Alemanha, nos meses que precederam as eleições.[207] Este projeto visava tornar transparente o direcionamento de anúncios de Facebook feitos pelos partidos alemães. Para tanto, um plugin coletava os anúncios políticos visualizados por um usuário em sua linha do tempo, revelando as propagandas que esta pessoa não via. Nesta ocasião, nos juntamos a outros veículos alemães, como *Süddeutsche Zeitung* e *Tagesscha*u — em uma colaboração incomum envolvendo rivais, mas que parecia necessária levando em conta o interesse público, de forma a atingir a maior quantidade possível de pessoas. Os resultados obtidos poderiam ser publicados em produtos jornalísticos, mas o foco mesmo era a transparência. Passadas duas semanas, cerca de 600 anúncios de cunho político haviam sido coletados e disponibilizados ao público.

Julia Angwin e Jeff Larson, da *ProPublica*, vieram com a ideia da colaboração durante a conferência anual da associação sem fins lucrativos Netzwerk Recherche em Hamburgo, onde realizaram um debate sobre prestação de contas algorítmica e a prática jornalística em torno do tema. Desde o princípio, a ideia foi desenvolvida com o envolvimento de especialistas técnicos e metodológicos de diversos departamentos da

---

206 https://datenspende.algorithmwatch.org/en/index.html.

207 https://www.propublica.org/article/help-us-monitor-political-ads-online.

redação do *Spiegel Online*. Nossa colaboração prévia com a ONG AlgorithmWatch também foi bastante valorosa para que pudéssemos aprender mais sobre as questões legais envolvidas e incluí-las em nossa pesquisa. Após o evento, demos prosseguimento à ideia e a expandimos por meio de conferências telefônicas regulares. Posteriormente, parceiros de outros veículos se envolveram.

Em 2018, o *Spiegel Online* passou a apoiar um grande projeto voltado à investigação de um poderoso algoritmo em uso na Alemanha, o relatório de crédito Schufa, usado para avaliação de crédito de indivíduos. Este relatório teria como objetivo mostrar a probabilidade de alguém pagar suas contas, aluguel ou conseguir um empréstimo. Desta forma, pode ter grandes implicações na vida privada de uma pessoa e um impacto negativo na sociedade em geral. Por exemplo, é possível que essa pontuação aumente a discriminação social ou trate pessoas de forma desigual, com base na quantidade de dados disponíveis sobre elas. Além disso, dados incorretos vindos de fontes integradas ou enganos de qualquer tipo podem ter consequências fatais.

A pontuação não é transparente, não se sabe quais informações são consideradas e com qual peso. Além do que, nem sempre os afetados sabem de qualquer coisa sobre este processo. Todos estes fatores fazem da Schufa uma instituição controversa na Alemanha — projetos como OpenSCHUFA acabam sendo vitais para o debate público em torno de transparência e prestação de contas de algoritmos, em nossa opinião.[208]

Tal projeto é operado principalmente pelas ONGs Open Knowledge Foundation (OKFN) e AlgorithmWatch, com o *Spiegel Online* sendo um de seus parceiros, junto ao *Bayerischer Rundfunk* (Sistema de Radiodifusão Bávaro). A ideia deste projeto surgiu meio que simultaneamente, envolvendo diversas partes. Após algumas empreitadas bem-sucedidas em associação com as ONGs citadas acima, com a participação da equipe de jornalismo de dados do *Bayerischer Rundfunk*, o *Spiegel Onlin*e passou a integrar as discussões.

Cabe notar alguns desafios específicos que ali havia. Para as duas equipes de mídia, era importante trabalhar separadamente das ONGs para garantir sua independência em relação ao processo de financiamento coletivo, em especial. Sendo assim, por mais que houvesse discussões envolvendo os atuantes no caso, não era possível firmar uma parceria oficial ou uma avaliação conjunta de dados. Este exemplo enfatiza quão importante é para jornalistas refletirem sobre sua autonomia, especialmente em assuntos de grande destaque.

Tornar a iniciativa OpenSCHUFA conhecida foi um dos fatores centrais para o sucesso deste projeto. O primeiro passo foi valer-se de financiamento coletivo de forma a criar a infraestrutura necessária para coleta de dados, que viria a ser realizada em 2018, por

---

[208] http://www.openschufa.de/.

meio de colaboração voluntária. Posteriormente, os resultados seriam avaliados em conjuntos pelos parceiros, de maneira anonimizada. As questões centrais: o algoritmo da Schufa discrimina certos grupos populacionais? Ele aumenta a desigualdade social?

Em março de 2018, a campanha se provou bem-sucedida. Foi possível bancar o software por meio do sistema de financiamento coletivo.[209] Fora isso, mais de 16.000 pessoas solicitaram seus dados pessoais à Schufa. Com base nestes relatórios, será feita a análise do algoritmo e seus efeitos.

Indicadores de sucesso e ressonância

Em questão de resultados, ambos os projetos envolvendo Facebook e Google não foram nada espetaculares nem revelaram os efeitos esperados. Aparentemente, partidos políticos mal usavam as opções de direcionamento do Facebook e a tal bolha do Google provou ser imensurável, com base nos esforços colaborativos realizados na Alemanha. De qualquer forma, para nós mais valia aumentar o letramento em torno de algoritmos em meio ao nosso público e ilustrar suas funcionalidades e riscos.

A suposição de que conseguimos tornar o assunto mais conhecido se apoia no alcance de artigos publicados. 335.000 leitores foi o alcance do texto introdutório sobre o projeto Schufa, a maioria destas pessoas chegando até ele por canais internos, como nossa página principal. No caso de uma história pessoal de um de nossos jornalistas, publicada como relatório de campo, chegamos a 220.000 leitores. Um quinto destas pessoas chegou ao artigo por meio de redes sociais, um número bem acima da média. Ou seja, pelo jeito é possível atingir novos grupos com este tema. O tempo de leitura também estava claramente acima do normal, batendo uma média de quase três minutos. Além do que, o tema foi amplamente debatido pelo público e em outros veículos, incluindo diversas conferências.

Mas e o impacto no cotidiano? Em um primeiro momento, consideramos relevante ancorar o tema na consciência pública. Até então, não percebemos nenhuma grande diferença na forma como atores políticos lidam com algoritmos publicamente eficazes. Esperamos, porém, que projetos como estes aumentem a pressão para a criação de legislação e normas ligadas à transparência no setor.

De qualquer forma, mais esforços fazem-se necessários. Por conta destes projetos discutidos anteriormente, pudemos trabalhar em cima de aspectos relevantes de algoritmos específicos, mas claro que seria no mínimo recomendável dedicar mais recursos ao tema. É ótimo que o trabalho pioneiro de Julia Angwin e Jeff Larson venha a se desdobrar através de uma nova organização de mídia voltada ao impacto social da tecnologia, capaz de dar mais

---

[209] https://www.startnext.com/openschufa.

atenção ao tópico em questão. Maior experimentação faz-se necessária, em parte porque ainda há o que ser feito para a regulamentação de algoritmos. Esta área jornalística, voltada à transparência e prestação de contas de algoritmos, se desenvolveu nos últimos anos. Para enfrentar os desafios impostos por um mundo cada vez mais digitalizado, está claro que precisa crescer ainda mais.

Sobre organização de investigações colaborativas

Trabalhar junto com diversas cabeças não só facilita o compartilhamento de competências e recursos, também permite definir papéis claramente. Como parceiro de mídia, o *Spiegel Online* pode atuar como um comentarista neutro sem se envolver muito profundamente com o projeto em questão. Os editores seguem independentes, justificando a confiança de seus leitores. Claro, estes mesmos editores aplicam seus critérios de qualidade à reportagem feita dentro de iniciativas como estas, dando oportunidade de resposta para qualquer um dos envolvidos nas acusações, por exemplo. Em comparação às ONGs envolvidas, mecanismos como estes podem desacelerar o trabalho de parceiros da mídia mais do que eles gostariam, mas, ao mesmo tempo, há a garantia de que os leitores terão toda a informação necessária, o que enriquece o debate a longo prazo.

Definir estes papéis logo de cara provou ser um importante fator para o sucesso de colaborações no campo de transparência e prestação de contas de algoritmos. Uma linha do tempo comum também deve ser implementada no início do projeto, bem como regras de linguagem para apresentação do mesmo em diferentes canais. Afinal, uma divisão clara de funções só pode funcionar se comunicada de maneira consistente. Isso inclui, por exemplo, uma terminologia clara sobre os papéis assumidos pelos parceiros dentro do projeto e o emprego de avisos legais no caso de conflitos de interesse.

Devem-se empregar métodos de gestão de projetos de maneira prudente nos bastidores, objetivos devem estar definidos claramente e é necessário discutir os recursos disponíveis. Cabe aos coordenadores auxiliar na comunicação geral e dar o espaço necessário aos editores participantes durante suas investigações. Para manter todos atualizados, canais de informações devem ser os mais simples possíveis, especialmente próximo ao lançamento de etapas relevantes do projeto.

Quanto ao planejamento editorial, as três áreas tinham seus desafios. Por mais que a relevância e o valor noticioso dos temas nunca tenham sido questionados, foram necessários artigos específicos para alcançar uma base de leitores mais ampla. Muitas vezes, estes artigos focavam nos efeitos pessoais causados pelos algoritmos em questão. Dados da Schufa associados incorretamente dificultaram um colega da equipe editorial do *Spiegel Online* na conclusão de um contrato de internet, por exemplo O relato de sua experiência mostrou, de

forma impressionante, os possíveis efeitos do algoritmo a nível pessoal, criando uma conexão com a realidade de nosso público.[210]

Desta forma, moldamos o escopo de nossa reportagem aos interesses de nosso público o máximo possível. Claro que os jornalistas de dados envolvidos também têm grande interesse no funcionamento dos algoritmos investigados, interesse de utilidade extrema para fins de pesquisa. Porém, tais detalhes só podem se tornar o foco da reportagem caso possuem influência relevante sobre os algoritmos, e apenas se estes forem discutidos de uma forma acessível ao leitor.

Internamente, houve bastante apoio por parte da editoria para os três projetos. No entanto, não foi fácil liberar cursos no cotidiano de uma redação focada em notícias, especialmente quando os resultados não eram sempre espetaculares.

Mesmo assim, temos o tema da transparência e prestação de contas de algoritmos como muito caro a nós. Veja, na Europa ainda temos condições de discutir o assunto em sociedade e moldar nossa relação com ele. É parte de nossa função enquanto jornalistas oferecer o conhecimento necessário à população para que possa compreender e dar forma a este escopo. Até onde for possível, também levamos adiante o papel de sentinela ao tentar tornar algoritmos e seus efeitos transparentes, identificando riscos e confrontando os responsáveis. Para tanto, é preciso estabelecer colaborações que em outras circunstâncias seriam consideradas fora do comum, incluindo rivais e atores de outros setores.

O que aprendemos com estes projetos

Colabore sempre que possível. Apenas equipes diversas conseguem delinear um bom projeto de pesquisa e juntar forças ao abordar estes tópicos, um argumento importante considerando a escassez de recursos e as restrições legais com as quais grande parte dos jornalistas precisa lidar. Como estes projetos reúnem gente de diferentes sistemas, é crucial discutir critérios de relevância, requisitos e capacidades logo de cara.

Defina seus objetivos de maneira clara. Conscientizar as pessoas em torno dos princípios operacionais dos algoritmos pode ser um grande objetivo neste tipo de projetos. Claro que cabe também a estas iniciativas o máximo de transparência possível. Podemos verificar se tais algoritmos têm ação discriminatória, no melhor dos cenários, mas os parceiros envolvidos devem ter em mente que um objetivo como esse demanda a disponibilidade de conjuntos extensos de dados.

---

210 http://www.spiegel.de/wirtschaft/service/schufa-wie-ich-zum-deutlich-erhoehten-risiko-wurde-a-1193506.html.

Implemente projetos com cuidado. A depender da carga de trabalho e pressão cotidiana sofrida pelos jornalistas envolvidos, talvez seja necessário um gestor de projetos. Tenha em mente que a linha do tempo da operação pode conflitar com as demandas cotidianas da redação. Leve isso em consideração na comunicação com demais parceiros e, se possível, tenha alternativas prontas para lidar com estas situações.

Dedique-se ao desenho da pesquisa. Um desenho relevante que produza dados úteis pode demandar a ajuda de especialistas. Manter relações próximas com cientistas nas áreas de computação, matemática e outras disciplinas relacionadas é bastante útil ao debruçar-se sobre os aspectos mais técnicos dos algoritmos. Além do que, pode valer a pena cooperar com pesquisadores sociais e culturais para maior compreensão de classificações e normas implementadas nestes algoritmos.

Proteja bem os dados de seus usuários. Ao investigar algoritmos, podemos receber dados fornecidos voluntariamente por usuários, de forma a considerar o maior número de cenários possível. Em projetos colaborativos, especialmente, apoio jurídico é indispensável para garantir a proteção dos dados considerando os requisitos de normas e leis nacionais. Se a sua empresa tem alguém responsável pela proteção de dados, envolva-o no projeto o quanto antes.

Christina Elmer criou a editoria de jornalismo de dados do Spiegel Online, fomenta projetos de inovação e integra o conselho da Netzwerk Recherche.

Referências

AlgorithmWatch: Datenspende BTW17. Setembro de 2017. Disponível em: https://datenspende.algorithmwatch.org/en/index.html.

ANGWIN, Julia; LARSON, Jeff. *Help Us Monitor Political Ads Online*. ProPublica, setembro de 2017. Disponível em: https://www.propublica.org/article/help-us-monitor-political-ads-online.

KOLLENBROICH, Philipp. *Wie ich bei der Schufa zum "deutlich erhöhten Risiko" wurde*. Spiegel Online, março de 2018. Disponível em: http://www.spiegel.de/wirtschaft/service/schufa-wie-ich-zum-deutlich-erhoehten-risiko-wurde-a-1193506.html.

LOBO, Sascha. *Was Bettina Wulff mit Mettigeln verbindet*. Spiegel Online, setembro de 2012. Disponível em: http://www.spiegel.de/netzwelt/netzpolitik/google-suchvorschlaege-was-bettina-wulff-mit-mettigeln-verbindet-a-855097.html.

REINBOLD, Fabian. *Warum Merkel na die Algorithmen will*. Spiegel Online, outubro de 2016. Disponível em: http://www.spiegel.de/netzwelt/netzpolitik/angela-merkel-warum-die-kanzlerin-an-die-algorithmen-von-facebook-will-a-1118365.html.

Site do projeto OpenSCHUFA. Disponível em: http://www.openschufa.de/.

Sistema de financiamento coletivo do OpenSCHUFA. Disponível em: https://www.startnext.com/openschufa.

VOR KARTELL, Sorge. *Maas hätte gerne, dass Google geheime Suchformel offenlegt*. Setembro de 2014. Disponível em: http://www.spiegel.de/wirtschaft/unternehmen/google-heiko-maas-fordert-offenlegung-von-algorithmus-a-991799.html.

Organização do jornalismo de dados

A hashtag #ddj no Twitter

Eunice Au e Marc Smith

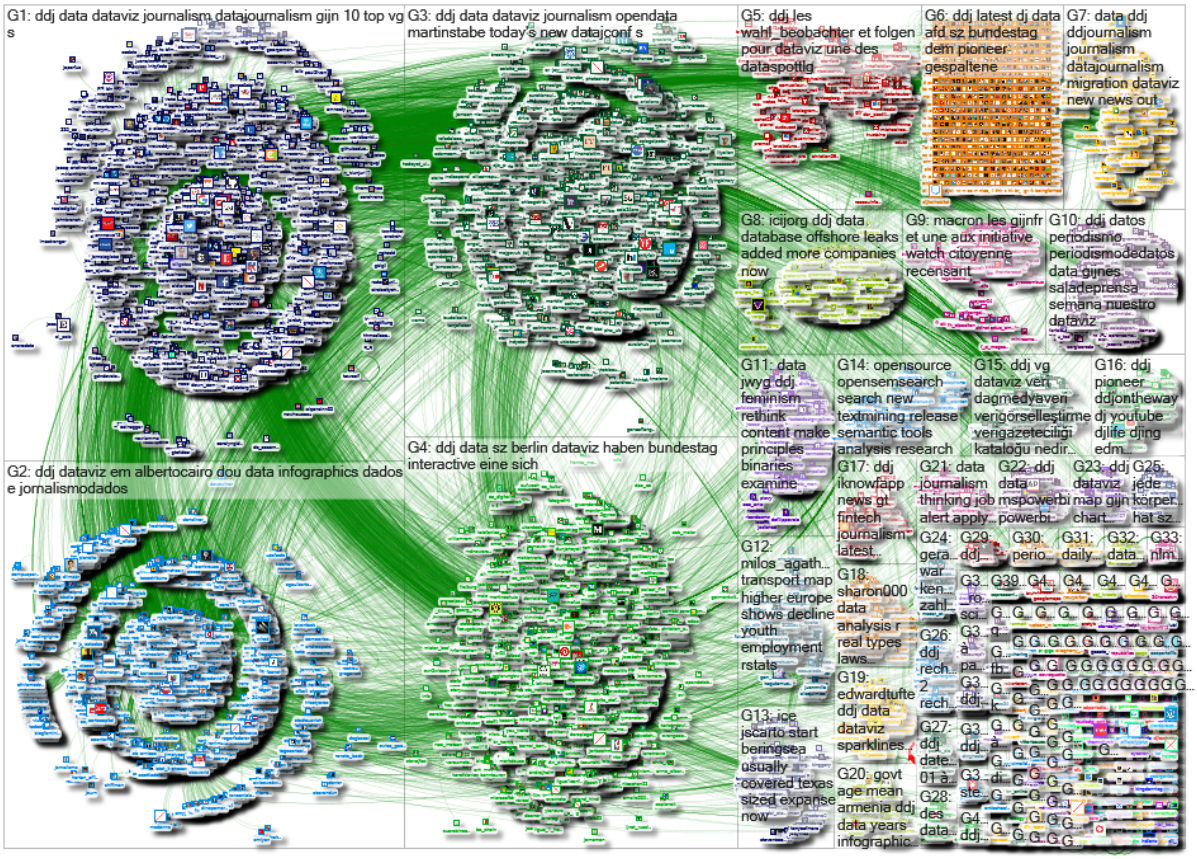


Figura 1: Mapeamento da hashtag #ddj no Twitter de 1º de janeiro de 2018 a 13 de agosto de 2018.

Escolher um único termo para monitorar o campo de jornalismo de dados não é fácil, já que seus jornalistas usam uma miríade de hashtags associadas ao seu trabalho, como #datajournalism, #ddj #dataviz, #infographics e #data. Quando a Rede Global de Jornalismo Investigativo (GIJN, na sigla original), uma associação internacional de organizações jornalísticas que apoia o treinamento e compartilhamento de informações entre jornalistas investigativos e de dados, começou a cobrir conversas em torno do jornalismo de dados no Twitter há cinco anos, a hashtag mais popular parecia ser #ddj (*data-driven journalism;* "jornalismo baseado em dados" em português).

O termo jornalismo baseado em dados já é controverso por si só, já que se pode argumentar que a prática não é baseada ou movida por dados; eles apenas informam, são uma ferramenta a serviço do jornalismo. Dados são fatos e estatísticas estruturadas que demandam

filtragem, análise e descoberta de padrões por parte dos jornalistas para que se crie uma narrativa. Assim como ninguém chamaria um perfil de "jornalismo baseado em entrevistas" ou artigo que se apoia em documentos de "jornalismo baseado em documentos", boas matérias em jornalismo de dados usam estas informações apenas como uma parte do todo.

O papel da #ddj

Apesar deste argumento, a aceitação ampla da hashtag #ddj entre comunidades jornalísticas de dados fez dela uma fonte de destaque para o compartilhamento de projetos e atividades do setor pelo mundo. Profissionais usam a hashtag para promover seu trabalho e divulgá-lo para um público internacional mais amplo.

O uso da hashtag também facilita discussões sobre o tema nas redes sociais, em que membros da comunidade jornalística de dados podem pesquisar, descobrir e compartilhar conteúdo relacionado pelo uso da mesma. Discussões em torno da #ddj vão de previsões eleitorais à interpretação errônea de gráficos de probabilidade, percorrendo campos como ética em dados e responsabilização de atos de inteligência artificial.

Surgimento do Top 10 #ddj

A série semanal Top 10 #ddj da GIJN teve início em janeiro de 2014, quando o sociólogo Marc Smith, coautor deste capítulo, tweetou um gráfico de rede acompanhado da hashtag #ddj. O gráfico, que mapeava tweets mencionando a hashtag, incluindo respostas a estas postagens, foi criado com NodeXL, um pacote de análise e visualização de redes sociais baseado no software de planilhas Excel. Estes gráficos revelam a forma da multidão que emerge a partir dos padrões de interconexão criados por quaisquer pessoas que respondam, mencionem ou retweetem outros usuários. Tais padrões destacam as principais pessoas, grupos e tópicos sendo discutidos.

No papel de uma organização internacional voltada ao jornalismo investigativo, a GIJN sempre busca por formas de chamar atenção para o que está acontecendo nos campos do jornalismo investigativo e de dados. Quando o diretor-executivo da GIJN, David Kaplan, viu o gráfico de Smith, aproveitou a ideia do mapa ali incluso para criar uma espécie de ranking semanal com o objetivo de dar destaque a exemplos interessantes e populares de jornalismo de dados, criando assim o Top 10 #ddj. (David e Smith também tentaram fazer o mesmo com pautas de jornalismo investigativo, mas não havia hashtag que chegasse nem perto do caráter agregador da #ddj para o jornalismo de dados). Por mais que a GIJN siga as sugestões do gráfico ao máximo, curadoria humana é necessária para remoção de duplicatas e para destacar o que há de mais interessante.

Desde o surgimento da série, já reunimos mais de 250 instantâneos das discussões da comunidade de jornalismo de dados em torno da hashtag #ddj durante estes (quase) seis anos. Atualmente, ela serve como um bom resumo rápido para os interessados que não conseguem acompanhar cada tweet feito com a hashtag.

O uso da palavra "instantâneo" também não é uma simples metáfora. Esta análise nos oferece uma espécie de retrato da comunidade jornalística de dados no Twitter, da mesma forma que o fotojornalismo retrata multidões de verdade nas páginas principais dos grandes veículos.

Organização da comunicação em redes sociais

Para ter uma noção de como o tráfego no Twitter com a hashtag #ddj evoluiu, fizemos uma análise bem básica dos dados coletados de 2014 a 2019. Escolhemos um período de amostragem curto, compreendendo oito semanas em fevereiro e março de cada um dos seis anos de operação, totalizando 48 semanas. Havia grande variedade de conteúdo sendo compartilhado e com o qual as pessoas interagiam. Os mais populares incluíam análises e artigos opinativos, prêmios, bolsas, eventos, cursos, vagas, ferramentas, recursos e investigações. O tipo de conteúdo se provou consistente ao longo dos anos.

Em 2014, vimos artigos que discutiam um setor de jornalismo de dados em franca expansão. Isso incluía artigos que argumentavam a relevância deste tipo de jornalismo pois fomentava transparência e percepções, ainda com previsões que indicavam a análise de dados como o futuro dos jornalistas. Em anos posteriores, vimos novos temas sendo discutidos, como inteligência artificial, grandes vazamentos de dados e investigações colaborativas. Havia guias aprofundados, nos quais jornalistas de dados ofereciam um vislumbre de seus processos de trabalho, compartilhando como melhor utilizar bancos de dados no lugar de debater se a indústria deveria ou não incorporar a prática em suas redações. Também notamos a grande presença de temas como eleições, imigração, poluição, clima e futebol.

O apanhado semanal da GIJN através da #ddj não só destaca os tweets e URLs mais populares, como também lista os principais participantes desta discussão. Entre os nomes citados costumeiramente em meio à discussão de #ddj temos especialistas em jornalismo de dados como Edward Tufte, Alberto Cairo, Martin Stabe, Nate Silver e Nathan Yau, e equipes de dados da Europa e da América do Norte, incluindo veículos como *Le Telegramme*, *Tagesanzeiger*, *Berliner Morgenpost*, *FiveThirtyEight*, *Financial Times* e *UpshotNYT*. Seus trabalhos, de alta qualidade, por vezes se mostram educativos, inspiradores e servem como trampolim para debates mais aprofundados. A comunidade também pode aproveitar e criar laços com estes influenciadores.

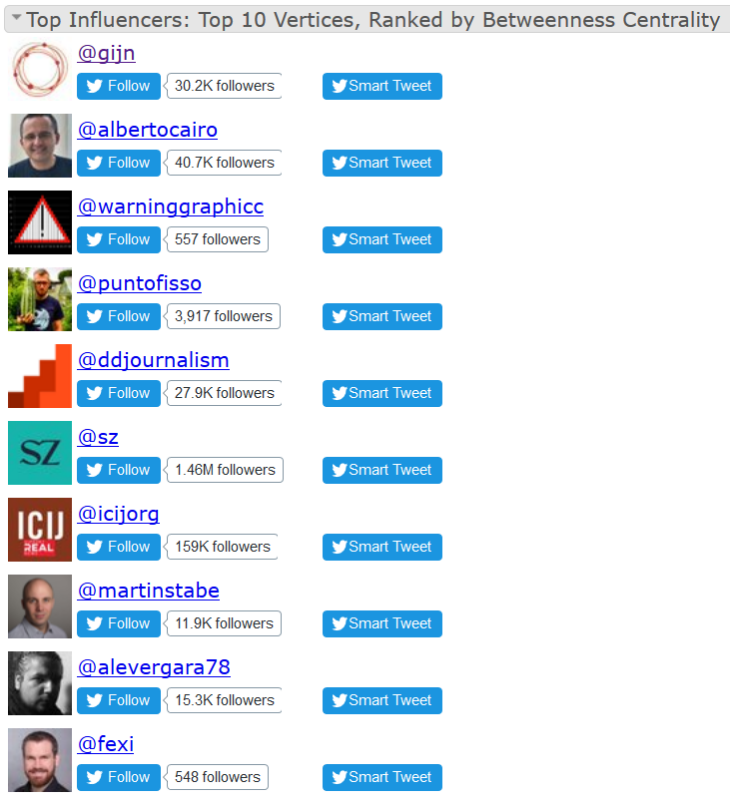


Figura 2: Exemplos de principais influenciadores (de 1º de janeiro de 2018 a 13 de agosto de 2018).

O mapeamento da hashtag #ddj feito pela Connected Action também revela outras tags que costumam aparecer junto da #ddj, possibilitando aos membros da comunidade buscarem conteúdo semelhante.

Hashtags mais usadas ao longo de todo o gráfico:

[22540] ddj

[6765] dataviz

[1783] datajournalism

[1578] opendata

[1517] vg

[1253] data

[1080] infographics

[589] opensource

[541] datajournalismawards

[534] journalism

Figura 3: Exemplos de principais hashtags relacionadas (de 1º de janeiro de 2018 a 13 de agosto de 2018).

De longe, as hashtags mais utilizadas junto à #ddj foram #dataviz, #visualization, #datajournalism, #opendata, #data e #infographics. Isso sinaliza para nós que há gente neste campo que se preocupa não só com a disponibilidade de dados públicos, mas também com a forma como estes dados se apresentam criativamente para os leitores e como são criadas suas visualizações.

Representação

Dito isso, o mapeamento da #ddj feito pelo NodeXL de forma alguma representa o campo como um todo, pois analisa somente pessoas que tweetam. Além do que, geralmente, aqueles que têm mais seguidores no Twitter e conseguem mais retweets costumam aparecer com maior destaque em nossa seleção.

Notamos ainda que grande parte dos tweets mais relevantes costuma vir da Europa ou das Américas, especialmente Alemanha e Estados Unidos, com alguns vindos de Ásia e África. Isso pode estar relacionado à base de usuários da plataforma ou, talvez, tenha a ver com uma presença relativamente menos robusta de comunidades voltadas ao jornalismo de dados em outras regiões.

Ao longo do último ano, percebemos que parte do trabalho realizado por organizações relevantes do setor, de ampla circulação no Twitter, não aparecia em nosso gráfico de rede. Podemos relacionar este acontecimento com o não uso de #ddj ao tweetar sobre o material, ou o uso de outras hashtags, ou mesmo de hashtag nenhuma. Suspeitamos que o aumento no limite de caracteres do Twitter de 140 para 280 em novembro de 2017 possa ter contribuído para que as pessoas optassem por hashtags mais extensas, como #datajournalism.

Descobertas divertidas com #ddj

O material jornalístico impressiona e as visualizações criadas, muitas vezes, são de tirar o fôlego, mas há situações em que encontramos coisas que são simplesmente divertidas. Listamos aqui algumas das descobertas mais engraçadas do último ano junto à #ddj:

Xaquìn G.V fez um ensaio visual adorável e inteligente onde mostra o que pessoas em diferentes países buscam quando querem consertar algo. Em países mais quentes,

normalmente são geladeiras; na América do Norte e no Leste Asiático, vasos sanitários; já no norte e leste da Europa, todos parecem mais preocupados com lâmpadas.

Uma tabela, encontrada em meio à publicação *Collection of Doughnut Ephemera*, de Sally L Steinberg, do *Smithsonan*, apresenta a teoria de que o tamanho do buraco de uma rosquinha vem reduzindo gradualmente ao longo dos anos.

De corrida de lesmas a carregamento de esposa, o designer gráfico Nigel Holmes ilustrou e falou sobre competições bizarramente maravilhosas pelo mundo em um livro chamado *Crazy Competitions*.

Mulheres pelo mundo já sabem que os bolsos em calças jeans femininas são pequenos a ponto de serem inúteis, e o *Puddingviz* nos deu os dados e análise que provam isso.

Existe uma época certa para fazer bebês? Uma análise dos dados de nascimentos das Nações Unidas indica que sim. O pessoal do *Visme* descobriu uma correlação entre três variáveis diferentes: meses com mais nascimentos, estações do ano e latitude do país (sua distância do equador), apontando que os dois últimos podem ter influência no ritmo da concepção em diferentes países.

Eunice Au é coordenadora de programa da Rede Global de Jornalismo Investigativo, e responsável por coletar os 10 destaques da semana em #ddj com tweets mais populares sobre jornalismo de dados. Marc Smith é sociólogo, atua no desenvolvimento e na aplicação de ferramentas para estudo de redes sociais.

Referências

ARTHUR, Charles. Analysing Data is the Future For Journalists, Says Tim Berners-Lee. The Guardian, 22 de novembro de 2010. Disponível em: https://www.theguardian.com/media/2010/nov/22/data-analysis-tim-berners-lee.

CHIBANA, Nayomi. Do Humans Have Mating Seasons?. Visme, 2017. Disponível em: https://visme.co/blog/most-common-birthday/.

DIEHM, Jan; THOMAS, Amber. Pockets. The Pudding, agosto de 2018. Disponível em: https://pudding.cool/2018/08/pockets/.

EDWARDS, Phil. Have donut holes gotten smaller? This compelling vintage chart says yes. Vox, 1º de junho de 2018. Disponível em: https://www.vox.com/2015/9/20/9353957/donut-hole-size-chart.

GALLEGO, Cecile S. How to Investigate Companies Found in the Offshore Leaks Database. International Consortium of Investigative Journalists, 23 de janeiro de 2018. Disponível em: https://www.icij.org/blog/2018/01/investigate-companies-found-offshore-leaks-database/.

GIJN. Top 10 #ddj series, *2014 a 2019*. Disponível em: https://gijn.org/series/top-10-data-journalism-links/.

GROSSENBACHER, Timo. (Big) Data Jouranlism with Spark and R. 8 de março de 2019. Disponível em: https://timogrossenbacher.ch/2019/03/big-data-journalism-with-spark-and-r/.

G.V., Xaquín. *How to Fix A Toilet*. 1º de setembro de 2017. Disponível em: http://how-to-fix-a-toilet.com/.

HOWARD, Alex. Data-driven Journalism Fuels Accountability and Insight in the 21st Century. *TechRepublic*. 3 de março de 2014. Disponível em: https://www.techrepublic.com/article/data-driven-journalism-fuels-accountability-and-insight-in-the-21st-century/.

Rede Global de Jornalismo Investigativo (GIJN). Disponível em: https://gijn.org/.

SMITH, Marc. *#ddj mapping on Twitter from Jan 1, 2018 to Aug 13, 2018*. NodeXL, 13 de agosto de 2018. Disponível em: https://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=163145.

SMITH, Marc. *first NodeXL #ddj network graph*. Twitter, 22 de janeiro de 2014. Disponível em: https://twitter.com/marc_smith/status/425801408873385984.

YAU, Nathan. *Nigel Holmes new illustrated book on Crazy Competitions*. Flowing Data, 21 de maio de 2018. Disponível em: https://flowingdata.com/2018/05/21/nigel-holmes-new-illustrated-book-on-crazy-competitions/.

Preservação em jornalismo de dados

Meredith Broussard

Na primeira edição do "Manual de Jornalismo de Dados", publicada em 2012, o pioneiro do jornalismo de dados Steve Doig escreveu que um de seus materiais favoritos do campo era o projeto *Murder Mysteries* de Tom Hargrove[211]. Publicado pela *Scripps Howard News Service*, tratava-se de um estudo de Hargrove sobre dados demográficos detalhados de 185.000 homicídios não solucionados, incluindo a criação de um algoritmo que sugeria quais dos casos poderiam estar relacionados. Estas ligações poderiam indicar a atuação de um *serial killer*. "Esse projeto tem de tudo", disse Doig na época: "Trabalho duro, um banco de dados superior ao do governo, análise inteligente com uso de técnicas de ciências sociais, e uma apresentação interativa dos dados disponível online para que os leitores possam explorá-los por conta própria".

Quando a segunda edição do livro saiu, seis anos depois, o endereço do projeto já não funcionava mais — <projects.scrippsnews.com/magazine/murder-mysteries>. Sumiu da internet, junto com quem o havia publicado; a *Scripps Howard* não existia mais, após diversas fusões e restruturações, incluindo sua fusão com a *Gannett*, editora da rede local de notícias *USA Today*.

Sabemos bem que as pessoas trocam de emprego, empresas de mídia vêm e vão. Essa dinâmica, porém, tem consequências desastrosas para projetos de jornalismo de dados.[212] Trabalhos em jornalismo de dados são mais frágeis do que outros produtos jornalísticos "comuns" baseados em textos e imagens, publicados nas edições impressas de jornais ou revistas.

Normalmente, links que deixam de funcionar não são um grande problema para arquivistas, tendo em vista que é possível usar soluções como Lexis-Nexis ou ProQuest, ou qualquer outro provedor de banco de dados, para encontrar uma cópia de tudo que já foi publicado, digamos, na versão impressa do *The New York Times*, em qualquer dia do século XXI. Mas, tratando-se de narrativas com dados, a morte destes links indica um problema mais profundo. Estes produtos não são preservados em arquivos tradicionais. Logo, estão sumindo da internet. A não ser que as organizações de notícias e bibliotecas ajam, os

---

[211] http://www.murderdata.org/.

[212] Para uma discussão mais ampla do tema, ver Broussard, *Preserving news apps present huge challenges*; Boss e Broussard, *Challenges of archiving and preserving born-digital news applications*; Broussard, *Future-Proofing News Apps*; e ProPublica, *A Conceptual Model for Interactive Database Projects in News*.

historiadores do futuro não terão como ler tudo que o *Boston Globe* publicou em qualquer dia de 2017, por exemplo. Isso tem sérias implicações para estudiosos e para a memória coletiva do campo. Muitas vezes, o jornalismo é chamado de "o primeiro rascunho da história". Se esse primeiro rascunho está incompleto, como os estudiosos do futuro entenderão o presente? Ou, com o desaparecimento dessas histórias da internet, como jornalistas manterão seus portfólios pessoais?

Uma questão humana, não apenas computacional. Para entender por que o jornalismo de dados não vem sendo preservado para a posteridade, ajuda começar com o entendimento de como as notícias "comuns" são preservadas. Todas as organizações e veículos de notícia usam sistemas de gerenciamento de conteúdo (*Content Management Systems,* CMS na sigla em inglês), que permitem a estas planejarem e gerenciarem centenas de conteúdos criados todos os dias, fazendo com que o material publicado se mantenha coeso. Historicamente, veículos e organizações mais antigas usavam um CMS diferente para as versões impressa e web. O CMS para a web permitia ao veículo embutir anúncios em cada página, uma das formas pelas quais este ganha dinheiro. Já o CMS impresso possibilita aos editores gerenciarem diferentes versões do layout para impressão, que serão enviados posteriormente à gráfica. No caso do vídeo, geralmente se usa um terceiro CMS. Para postagens em redes sociais, pode-se usar ou não soluções dedicadas como SocialFlow ou Hootsuite. Fluxos de dados para preservação tendem a estar atrelados ao CMS impresso, em grandes provedores como Lexis-Nexis ou outros. A não ser que alguém nestas organizações de comunicação se lembre de colocar o CMS web na jogada, o conteúdo publicado em meios digitais não é incluído no fluxo de informações recebido por bibliotecas e serviços de arquivamento. Este é um lembrete de que preservação e arquivamento não são neutros e se baseiam em decisões humanas a respeito do que importa (e o que não importa) para o futuro.

"Mas e o Internet Archive?" é a pergunta que a maioria faria a essa altura. Que conste que o Internet Archive é um achado e tanto, e o grupo faz um trabalho admirável na captura de instantâneos de sites de notícias. Sua tecnologia está entre o que há de melhor em software de arquivamento digital. Dito isso, sua abordagem não captura tudo. O Internet Archive só coleta páginas disponíveis publicamente. Veículos que exigem logins ou que usam paywalls como parte de sua estratégia financeira não têm como serem preservados no Internet Archive automaticamente. Páginas de conteúdo estático ou HTML simples são as mais fáceis de se preservar. Estas, sim, são captadas pelo Internet Archive. Já conteúdo dinâmico, como JavaScript, uma visualização de dados ou qualquer outra coisa que em algum momento já foi chamada de "Web 2.0" é muito mais complicado de lidar e acaba não sendo armazenado pelo Internet Archive, na maior parte do tempo. "Há diversos tipos de páginas dinâmicas, algumas delas facilmente armazenadas em um arquivo e outras que acabam se deteriorando por completo", de acordo com a seção de Perguntas Frequentes do próprio Internet Archive:

“Quando uma página dinâmica gera um HTML padrão, o processo de arquivamento funciona bem. Quando uma página dinâmica contém formulários, JavaScript, ou demais elementos que demandam interação com a hospedagem de origem, a página arquivada não manterá a funcionalidade do site original”.

Visualizações dinâmicas de dados e apps de notícias, atualmente o que há de mais inovador em jornalismo de dados, não têm como serem capturados por tecnologias de preservação atuais. Além do que, por conta de uma série de razões institucionais, estes produtos jornalísticos são desenvolvidos fora de um CMS. Então, mesmo que fosse possível preservá-los (não é o caso, na maior parte do tempo), nenhum processo automatizado os captaria pois não estão dentro do CMS.

É uma questão complicada. Não há respostas simples. Trabalho com uma equipe de jornalistas de dados, bibliotecários e cientistas da computação que estão tentando criar tecnologia para solucionar este problema espinhoso. Para tanto, emprestamos métodos de pesquisa científica reproduzível para nos certificar de que as pessoas poderão ler as notícias de hoje nos computadores do amanhã. Estamos adaptando uma ferramenta chamada ReproZip, que capta código, dados e ambiente de servidor usados em experimentos de ciência computacional. Acreditamos que a ReproZip pode ser integrada com outra ferramenta, o Webrecorder.io, de forma a coletar e preservar apps de notícias, que são ao mesmo tempo notícia e software. Como projetos de jornalismo de dados baseados na web e em mobile dependem e existem em relação a uma série de outros ambientes de mídia, bibliotecas, funcionalidades de navegador e entidades de rede (sujeitas à mudança contínua), esperamos poder usar a ReproZip de forma a coletar e preservar bibliotecas e códigos remotos que possibilitam objetos complexos de jornalismo de dados a funcionarem na rede. Precisaremos de um ano ou dois para provar nossa hipótese.

Enquanto isso, há alguns passos concretos que toda equipe de dados pode dar para certificar-se de que seu trabalho será preservado para o futuro.

Registre em vídeo. Essa é uma estratégia emprestada da preservação de videogames. Mesmo quando um console não existe mais, uma partida em vídeo pode mostrar o jogo rodando em seu ambiente original. O mesmo vale para produtos de jornalismo de dados. Armazene o vídeo em localização central com metadados em texto simples descrevendo o conteúdo apresentado. Sempre que surgir um novo formato de vídeo (como ocorreu na mudança do VHS para o DVD, ou quando este foi substituído pelo streaming), atualize todo o material para este formato.

Crie uma versão simplificada para posteridade. Bibliotecas como a Django-bakery permitem que páginas dinâmicas sejam apresentadas como estáticas. Este processo também é

conhecido como “bake-out”. Mesmo em um banco de dados com milhares de registros, cada registro dinâmico pode ser reduzido a uma página estática que demanda pouca manutenção. Em tese, todas estas páginas estáticas poderiam ser importadas para o sistema de gerenciamento de conteúdo da empresa. Além disso, não é preciso fazer isso logo no lançamento do conteúdo. Um projeto de dados pode ser lançado como um site dinâmico e transformado em um site estático quando o tráfego cair, passados alguns meses. A ideia principal é adaptar seu trabalho para sistemas de arquivamento e preservação ao criar a versão mais simples possível deste, certificando-se que esta versão está na mesma localização digital que todos os outros conteúdos publicados na mesma época.

Pense no futuro. Jornalistas tendem a planejar suas publicações e, então, seguir adiante. Em vez disso, tente planejar o encerramento de seu projeto de dados ao mesmo tempo que planeja seu lançamento. *Kill All Your Darlings*, de Matt Waite, publicado no *Source, The Open News Blog* é um bom guia para como pensar o ciclo de vida de um produto de jornalismo de dados. Com o tempo, você será promovido ou mudará de emprego. Certamente gostaria que seu trabalho sobrevivesse a essa partida.

Trabalhe junto a bibliotecas, institutos de preservação e empresas de arquivamento. Enquanto jornalista, é imprescindível manter cópias de seu trabalho. Ninguém vai procurar uma caixa no seu armário ou vasculhar seu HD, muito menos seu site pessoal, quando forem buscar por jornalismo no futuro, pense nisso. O que essas pessoas farão é consultar grandes repositórios comerciais como Lexis-Nexis e ProQuest. Para saber mais sobre preservação comercial e arquivamento digital, o livro de Kathleen Hansen e Nora Paul *Future-proofing the News: Preserving the First Draft of History* serve como guia canônico para a compreensão do panorama de preservação de notícias e os desafios tecnológicos, legais, e organizacionais relacionados ao tema.

Meredith Broussard é professora de jornalismo de dados no Instituto Arthur L. Carter de Journalismo da Universidade de Nova York, e autora de “Artificial Unintelligence: How Computers Misunderstand the World”.

Referências

FISHER, *Tyler;* LAB, Knight; KLEIN, Scott. *A Conceptual Model for Interactive Database Projects in News*. ProPublica, 2016. Disponível em: https://github.com/propublica/newsappmodel.

BOSS, Katherine; BROUSSARD, Meredith. *Challenges of archiving and preserving born-digital news applications*. IFLA Journal 42:3, 2017, p. 150-157. Disponível em: http://journals.sagepub.com/doi/abs/10.1177/0340035216686355.

BROUSSARD, Meredith. *Preserving news apps present huge challenges*. Newspaper Research Journal 36:3, 2015, p. 299-313. Disponível em: http://journals.sagepub.com/doi/abs/10.1177/0739532915600742.

BROUSSARD, Meredith. *The Irony of Writing Online About Digital Preservation*. The Atlantic, 20 de novembro de 2015. Disponível em: https://www.theatlantic.com/technology/archive/2015/11/the-irony-of-writing-about-digital-preservation/416184/.

BROUSSARD, Meredith. *Future-Proofing News Apps*. Media Shift, 23 de abril de 2014. Disponível em: http://mediashift.org/2014/04/future-proofing-news-apps/.

Do *Guardian* ao Google News Lab: uma década trabalhando com jornalismo de dados

Simon Rogers

Quando decidi que queria ser jornalista, em algum momento entre o primeiro e o segundo ano do ensino fundamental, nunca imaginei que isso envolveria dados de alguma forma. Agora, trabalhando com dados todo dia, percebi o quão sortudo fui. Se estou onde estou hoje, certamente não foi por conta de um planejamento detalhado de carreira. Apenas estava no lugar certo na hora certa. O *jeito* como tudo aconteceu, porém, diz muito sobre o estado do jornalismo de dados em 2009. Creio que também nos fale bastante sobre o mesmo tema, mas em 2019.

Adrian Holovaty, desenvolvedor de Chicago que trabalhou no *Washington Post* e fundou o Everyblock, veio falar à redação no antigo centro educacional do *Guardian* na Farringdon Road, em Londres. Naquela época, eu trabalhava como editor de notícias no impresso (o centro gravitacional do veículo então), com experiência na porção online e tendo atuado como editor de uma seção de ciências. Quanto mais Holovaty falava sobre o uso de dados para contar histórias e ajudar as pessoas a entenderem o mundo, mais sentia algo despertar em mim. *Eu* não só poderia fazer aquilo, como aquilo refletia o que eu *fazia* cada vez mais. Talvez eu pudesse ser um jornalista que trabalha com dados. Um "Jornalista de Dados".

Como editor de notícias responsável pela parte gráfica, tive a oportunidade de trabalhar com designers que mudaram a forma como vejo o mundo, parte da equipe talentosa de Michael Robinson. Ao passo que o número de gráficos aumentava, acabou que eu também havia acumulado um monte de números: Matt McAlister, que estava lançando a API aberta do *Guardian,* descreveu aquilo como uma "mina de ouro". Tínhamos informações sobre o PIB, emissões de carbono, gastos governamentais e muito mais, tudo limpinho em planilhas no Google e pronto para uso sempre que precisássemos.

E se publicássemos essas informações em formato de dados abertos? Nada de PDFs, apenas dados acessíveis, interessantes e prontos para uso, para quem quiser usar. E foi isso que fizemos com o *Datablog* do *Guardian*, primeiro com 200 conjuntos de dados diferentes, com temas como índices de criminalidade, indicadores econômicos, detalhes de zonas de guerra, semanas de moda e até mesmo vilões de *Doctor Who*. Começamos a perceber que dados poderiam ser aplicados a qualquer coisa. Ainda era algo esquisito de se estar fazendo, veja bem. "Editor de dados" não era um cargo lá muito conhecido, tendo em vista que

pouquíssimas redações tinham uma equipe de dados. Inclusive, o uso da palavra "dados" em uma reunião de pauta gerava algumas caras feias. Isso não era jornalismo "de verdade", certo?

Mas 2009 marcava o início da revolução dos dados abertos: em maio daquele ano, o governo dos EUA havia lançado seu portal de dados, data.gov, com apenas 47 conjuntos de dados. Portais do tipo estavam sendo lançados por países e cidades pelo mundo inteiro, acompanhados por campanhas que exigiam acesso a mais informações. Ao longo de um ano, contamos com a colaboração de nossos leitores para reunir milhares de informações sobre os gastos do Parlamento inglês; e o governo britânico havia divulgado seu conjunto de dados mais completo sobre o tema: COINS, sigla em inglês para Sistema de Informações Combinadas Online. A equipe do *Guardian* havia desenvolvido uma espécie de explorador para encorajar os leitores a explorarem ambos os bancos de dados.[213] Assim que produzimos artigos com aqueles dados, porém, nos perguntamos "como extrair mais disso aqui?"

A resposta não demorou a vir, na forma de uma então nova organização com base na Suécia que poderia muito bem ser descrita como uma empreitada radical de transparência: Wikileaks. Seja lá o que for que você pensa sobre o Wikileaks hoje, seu impacto na história recente do jornalismo de dados não tem como ser exagerado. Ali estava uma coleção gigantesca de milhares de registros detalhados das zonas de conflito no Afeganistão, em primeiro lugar, depois vieram dados sobre o Iraque. Disponibilizados em uma enorme planilha, grande demais para que a equipe de investigações do *Guardian* pudesse trabalhar inicialmente.

Era um volume maior do que os *Pentagon Papers*, arquivos divulgados durante a Guerra do Vietnã que lançavam luz sobre a quantas andava o conflito, de fato. Era tudo bastante detalhado, incluindo uma lista de incidentes com número de vítimas, geolocalização, pormenores e categorias. Com base nisso, pudemos observar o aumento no número de atentados com explosivos improvisados no Iraque, por exemplo, assim como o quão perigosas as estradas pelo país haviam se tornado. A combinação daqueles dados com as habilidades clássicas de jornalistas veteranos de guerra mudou a forma como o mundo via as guerras.

Não era difícil produzir conteúdo de aparente impacto mundial. Os dados geográficos nas planilhas eram ótimos para mapeamento e havia uma nova ferramenta gratuita que poderia ajudar com isso: Google Fusion Tables. Sendo assim, criamos rapidamente um mapa com todos os incidentes no Iraque envolvendo no mínimo uma vítima fatal. Dentro de 24 horas, este conteúdo que havia levado uma hora para ser feito estava rodando o mundo, os

---

213 https://www.theguardian.com/politics/coins-combined-online-information-system.

usuários agora capazes de explorarem a zona de conflito por conta própria, de forma que tudo aquilo parecia mais real. Como os dados eram estruturados, a equipe gráfica conseguiu criar representações visuais sofisticadas, fornecendo um tipo de reportagem mais aprofundada.

Ao final de 2011, ano anterior à publicação do volume, o projeto *Reading the Riots* havia aplicado as técnicas de reportagem assistida por computador de Phil Meyer, surgidas nos anos 1960, a um surto de violência que tomou a Inglaterra.[214] Meyer havia usado estas técnicas de ciências sociais ao jornalismo em torno dos protestos ocorridos em Detroit no final da década de 1960. Uma equipe do *Guardian* liderada por Paul Lewis fez o mesmo com a onda de agitação que varreu o país naquele ano, dados sendo uma parte essencial desse trabalho. Eram matérias de capa, feitas com base em dados.

Mas havia outra mudança ocorrendo na forma como consumimos informação, e acontecia rápido. Não lembro de me deparar com o termo "viral" em qualquer coisa que não fosse um texto sobre saúde antes de 2010. Claramente essa noção não se aplica mais aos dias de hoje, considerando que a ascensão do jornalismo de dados veio junto da ascensão das redes sociais. Usávamos tweets para promoção de matérias e artigos para usuários espalhados pelo mundo e o tráfego resultante disso levou a mais gente buscando por mais material baseado em dados como estes. Uma visualização ou um número poderia ser visto, em segundos, por milhares de pessoas. As redes sociais transformaram o jornalismo, mas a amplificação em torno do jornalismo de dados foi a mudança de paradigma que o levou do nicho para o mainstream.

Para começo de conversa, a dinâmica com o consumidor mudou. No passado, as palavras de um repórter eram consideradas sacrossantas; agora, você é só mais uma voz entre milhões. Ao cometer um erro com uma série de dados, 500 pessoas aparecerão prontas para te avisar do ocorrido. Lembro de ter conversas longas (e profundas) no Twitter com designers sobre esquemas de cores em mapas que afetaram a forma como eu trabalhava. O ato de compartilhar fez meu trabalho melhor.

De fato, esse espírito colaborativo ainda persiste no jornalismo de dados. A primeira edição deste livro, afinal, foi desenvolvida inicialmente por um grupo de pessoas que se encontrou no MozFest em Londres. Com a realização de cada vez mais eventos sobre dados, surgiram mais oportunidades para que jornalistas da área trabalhassem juntos e compartilhassem suas habilidades. Se os dados divulgados do Iraque e o Wikileaks foram grandes exemplos iniciais de cooperação transatlântica, veja só como estes se tornaram enormes reportagens globais envolvendo centenas de jornalistas. Os vazamentos de Snowden e os *Panama Papers* foram notáveis nesse sentido, levando em conta como repórteres pelo

---

214 https://www.theguardian.com/uk/video/2011/dec/09/reading-the-riots-detroit-meyer-video.

mundo trabalharam de forma coordenada no compartilhamento de material e construção de narrativas a partir do trabalho do outro.[215]

Tomemos como exemplo um exercício como o *Electionland*, que usava técnicas colaborativas de reportagem para monitorar questões ligadas à votação no dia da eleição. Também tomamos parte nisso, fornecendo dados do Google e ajudando na visualização destas questões, ambos em tempo real. Até o momento, *Electionland* foi o maior exercício de reportagem de um único dia da história, com mais de 1.000 jornalistas atuantes. É possível traçar uma linha direta deste projeto ao que fazíamos naqueles primeiros anos.

Meu ponto aqui não é listar projetos, mas destacar o contexto mais amplo daqueles primeiros anos, não só no *Guardian*, mas em redações pelo mundo. *New York Times*, *LA Times*, *La Nacion,* na Argentina: jornalistas pelo mundo descobriam novas formas de trabalhar ao contar histórias com base em dados de maneiras inovadoras. Este foi o pano de fundo para a primeira edição deste livro. O caso do *La Nacion* é um bom exemplo disso. Uma equipe pequena de repórteres entusiasmados aprendeu por conta própria a como criar visualizações com Tableau (uma nova ferramenta à luz da época), juntando isso a pedidos de acesso à informação que serviram de trampolim para um mundo de jornalismo de dados na América do Sul e América Latina.

Não mais uma província de uns poucos solitários, agora esta nova prática jornalística integrava muitas grandes redações. Com isso em mente, uma tendência ficou clara mesmo naquele tempo: sempre que uma nova técnica é introduzida ao arsenal de reportagem, não só dados seriam parte essencial daquilo como os jornalistas especializados nestes estariam no centro de tudo. Em menos de três anos, jornalistas encontraram dados, publicaram conjuntos de dados, colaboração voluntária passou a ser uma ferramenta essencial às redações, jornalistas usaram bancos de dados para lidar com volumes gigantescos de documentos e técnicas analíticas baseadas em dados foram aplicadas a produtos jornalísticos complexos.

Isso não deve ser encarado como um desdobramento isolado dentro do jornalismo, afinal, estes são efeitos diretos de outros desdobramentos em questões de transparência internacional que vão bem além do surgimento de portais de dados. Incluem-se aí campanhas como as realizadas pelo movimento *Free Our Data*, a *Open Knowledge Foundation* e demais grupos civis colados à tecnologia, cujo objetivo era pressionar o governo britânico para que abrisse outros dados para consulta e uso públicos, bem como disponibilizasse APIs que todos pudessem explorar. Além disso, houve um acesso maior a ferramentas de visualização e limpeza de dados gratuitas e poderosas, como Open Refine, Google Fusion Tables, Many

---

[215]Para mais informações sobre colaborações em larga escala realizadas em torno dos *Panama Papers*, consultar o capítulo de Díaz-Struck, Gallego e Romera neste volume

Eyes, Datawrapper, Tableau Public. Estas ferramentas, combinadas ao acesso a grande volume de dados públicos, facilitaram a produção de mais projetos e representações visuais de dados voltados ao público. Redações como a do *Texas Tribune* e *ProPublica* começaram suas operações em torno destas informações.

Consegue entender como isso tudo funciona? Um ciclo virtuoso de dados, processamento fácil, criação de visualizações, mais dados e por aí vai. Quanto mais dados circulam por aí, mais trabalho se faz com dados e maior a pressão para que mais informações sejam divulgadas. Quando escrevi o artigo *Data Journalism is the New Punk*, o argumento era o seguinte: estávamos em um ponto onde a criatividade poderia correr alta,[216] mas também era um ponto em que todo esse trabalho cairia no mainstream.

Dados não podem fazer tudo. Como dito por Jonathan Gray: "A atual onda de empolgação em relação a dados, tecnologia de dados e tudo que é coisa relacionada a dados pode levar alguém a pensar que todo esse negócio estruturado, legível por máquina, tem algo de especial."[217] No final, é só mais uma peça que o jornalista tem que usar em seu quebra-cabeça. Mas com mais e mais dados disponíveis, esse papel muda e fica ainda mais importante. Poder acessar e analisar grandes conjuntos de informações foi o principal atrativo de minha próxima mudança de carreira.

Em 2013, tive a oportunidade de me mudar para a Califórnia e trabalhar no Twitter como seu primeiro editor de dados, e ali havia ficado claro que o termo havia penetrado no vocabulário das grandes redações, especialmente nos EUA e na Europa. Diversos sites especializados em jornalismo de dados surgiram com poucas semanas de diferença entre si, caso do *Upshot*, do pessoal do *New York Times,* e do *FiveThirtyEight*, de Nate Silver. O público ao redor do mundo se mostrava mais letrado e valorizava visualizações sofisticadas de temas complexos. Você pode me perguntar que provas tenho de que o mundo está à vontade com estas visualizações, aí eu digo que não tenho muito além de minha experiência em produções visuais que chamem atenção na rede, um processo mais complicado do que já foi no passado. Antigamente, todos nos surpreendíamos com qualquer coisa visual, agora é difícil conseguir mais que um simples dar de ombros.

Quando entrei no Google News Lab para trabalhar com jornalismo de dados, em 2015, já estava claro que havia acesso a conjuntos de dados mais relevantes e maiores do que nunca. Diariamente, há bilhões de buscas, uma boa parte das quais nunca vimos antes.

---

216 https://www.theguardian.com/news/datablog/2012/may/24/data-journalism-punk.

217 https://www.theguardian.com/news/datablog/2012/may/31/data-journalism-focused-critical.

Jornalistas pegam estas informações e as analisam, junto de tweets e curtidas no Facebook.[218] É o escape da vida moderna, adaptado para que possamos extrair percepções sobre a forma como vivemos hoje.

O jornalismo de dados está mais forte do que nunca. Em 2016, o Data Journalism Awards recebeu 471 inscrições, já a edição de 2018 recebeu quase 700, mais da metade vindas de pequenas redações e muitas de todo o mundo — cada vez mais inovadoras, cabe comentar. Inteligência artificial, ou aprendizagem de máquina, tornou-se uma ferramenta para o jornalismo de dados, como demonstrado pelo trabalho de Peter Aldhous no *Buzzfeed*.[219] Enquanto isso, o acesso a novas tecnologias, como realidade virtual e realidade aumentada, abre possibilidades para que possamos contar histórias com dados de novas maneiras. No papel de alguém cujo trabalho é imaginar como o jornalismo de dados pode mudar e o que podemos fazer para apoiá-lo, observo como tecnologias emergentes podem ser simplificadas de forma que mais repórteres possam integrá-las ao seu trabalho. Por exemplo, há pouco trabalhamos com o estúdio de design Datavized na criação do TwoTone, ferramenta visual que traduz dados em sons.[220]

O que um jornalista de dados da Google faz? Tenho a chance de contar histórias com a ajuda de uma vasta coleção de conjuntos de dados, além de poder trabalhar com designers talentosos para imaginar o futuro das visualizações de dados no contexto de notícias e o papel de novas tecnologias no jornalismo. Parte do meu trabalho consiste em ajudar na exploração de novas tecnologias, de maneira que estas sejam usadas em contexto adequado para que sejam úteis. Isso também envolve a exploração de como jornalistas vêm usando dados e tecnologias digitais de formas novas em seu trabalho. Um projeto recente, *El Universal*, da *Zones of Silence*, demonstrou o uso de IA em jornalismo, usando processamento de linguagem para analisar a cobertura jornalística de homicídios ligados a cartéis de drogas e comparação com dados oficiais, já que há um silêncio entre as duas áreas em termos de reportagem; nós os ajudamos com isso através de acesso a APIs de inteligência artificial e recursos de design.

Os desafios são grandes, para todos nós. Todos consumimos cada vez mais informações em dispositivos móveis, o que traz consigo suas próprias dificuldades. Visualizações complexas que ocupam uma tela inteira deram de cara com o fato de que mais da metade de nós lemos as notícias em nossos celulares ou outros aparelhos móveis (um terço

---

[218] Para perspectivas mais aprofundadas sobre o tema, consulte a seção “Investigação de dados, plataformas e algoritmos”.

[219] https://www.buzzfeednews.com/article/peteraldhous/hidden-spy-planes.

[220] https://twotone.io/.

de nós consome notícias no banheiro, de acordo com estudo da *Reuters*).[221] Isso significa que os designers da redação têm que pensar em telas pequenas e ter uma atenção mais dispersa sempre que vão criar algo.

Também temos um novo problema que pode nos impedir de aprender sobre o passado. Códigos morrem, bibliotecas deixam de funcionar e, com o tempo, muito do que há de mais ambicioso no jornalismo simplesmente desaparece. Projetos como MPs Expenses e EveryBlock, dentre outros, acabaram sucumbindo a uma memória institucional que simplesmente não existe mais. Esta questão já vem sendo abordada de forma inovadora (como podemos ver no capítulo assinado por Meredith Broussard). No longo prazo, são necessários investimentos adequados e nos resta esperar para ver se a comunidade tem motivação o suficiente para fazer isso acontecer.

Além do que, temos um problema maior e mais alarmante: confiança. Análise de dados é uma prática sempre sujeita a interpretação e discordância, mas jornalismo bem-feito pode superar tais questões. Numa época em que a crença nas notícias e em um conjunto compartilhado de fatos é questionada todos os dias, o jornalismo de dados pode ser um novo caminho ao reunir fatos e evidências, apresentando-os de forma acessível.

No final, apesar de todas as mudanças, algumas coisas seguem constantes neste campo. O jornalismo de dados tem uma história longa,[222] mas em 2009 parecia uma maneira importante de chegar a uma verdade comum, algo que todos podemos defender. Isso se faz ainda mais necessário nos dias de hoje.

Simon Rogers é fundador do Datablog do Guardian, editor de dados na Google, diretor do Data Journalism Awards e autor do livro "Facts are Sacred" e de mais um monte de infográficos para livros infantis da Candlewick.

---

221 https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf.

222 Consulte, por exemplo, os capítulos assinados por Anderson e Cohen neste volume.

Emaranhados entre jornalismo de dados e tecnologia cívica

Stefan Baack

Por mais que a reportagem assistida por computador fosse considerada uma prática exclusiva de jornalistas (investigativos), o jornalismo de dados se caracteriza pela mistura com o setor de tecnologia e outras formas de trabalho e cultura com/de dados. Em uma comparação direta com a prática jornalística assistida por computador, a ascensão do jornalismo de dados nos EUA e na Europa se deu em meio de vários desdobramentos dentro e fora das redações: dados cada vez mais disponíveis na internet, não só por conta de iniciativas ligadas ao acesso a estas informações e vazamentos; redações passaram a contratar desenvolvedores e integrá-los à equipe editorial para melhor manejo de dados e criação de aplicações web interativas; surgimento de vários movimentos do tipo "tecnologia a serviço do bem", atraídos pelo jornalismo como forma de usar suas habilidades tecnológicas para um "bem comum". Isso tudo contribuiu para a entrada de tecnólogos na redação desde que o jornalismo de dados surgiu e se popularizou nos anos 2000, no ocidente e além. Porém, os emaranhados resultantes da atuação de jornalistas de dados e outras formas de trabalho com dados distinguem-se ao longo de diferentes regiões. Além do que, o jornalismo de dados está ligado a novas e empreendedoras formas de se fazer jornalismo, surgidas em resposta ao esforço constante das organizações de mídia para o desenvolvimento de um modelo de negócios sustentável. Estas novas organizações, redações sem fins lucrativos como a *ProPublica* ou startups bancadas por investidores como *BuzzFeed,* tendem a questionar as limitações tradicionais do jornalismo em sua aspiração de "reviver" ou "melhorar" a prática, com dados e tecnologia tendo papéis essenciais em sua atuação.[223]

Estes emaranhados criam novas dependências, mas não só isso, surgem também energias que permitem novas formas de colaborações entre os setores relacionados. Gostaria de usar a relação próxima entre jornalismo de dados e tecnologia cívica como exemplo, pois em muitos lugares ambos os fenômenos surgiram na mesma época e moldaram um ao outro, em seus estágios iniciais. Tecnologia cívica trata do desenvolvimento de ferramentas que visam conferir autonomia a cidadãos ao facilitar sua interação com governos (ou para cobrá-los). Alguns casos de projetos de tecnologia cívica: OpenParliament, site de monitoramento parlamentar que, entre outras coisas, torna discursos parlamentares mais acessíveis;

[223] Recomendo a leitura dos estudos de Wagemans et al., *Impact as driving force of journalistic and social change*, e Usher, *Venture-backed News Startups and the Field of Journalism*.

WhatDoTheyKnow, um site que ajuda usuários a enviarem e encontrarem pedidos de acesso à informação; FixMyStreet, site que simplifica o relato de problemas às autoridades locais.[224]

Tecnólogos cívicos e jornalistas de dados compartilham algumas características importantes. Primeiro, muitos dos profissionais de ambos os grupos se comprometem com princípios de cultura de código aberto e promovem compartilhamento, uso de ferramentas de código aberto e padrões de dados. Segundo, jornalistas de dados e tecnólogos cívicos dependem de dados, sejam obtidos junto a instituições oficiais, através de esforços colaborativos voluntários ou quaisquer outras fontes. Terceiro, por mais que os métodos sejam diferentes, ambos aspiram oferecer um serviço público que confira autonomia aos cidadãos e responsabilize as autoridades. Por conta do conjunto de habilidades voltadas a dados compartilhados por estes, ambições semelhantes e compromisso em compartilhar, tecnólogos cívicos e jornalistas de dados entendem seus papéis como complementares. Além disso, o apoio de organizações de mídia, fundações como a Knight Foundation e iniciativas de base como Hacks/Hackers criaram intercâmbio e colaboração contínuos entre os dois grupos.

A tensão entre expandir e reforçar o "núcleo" jornalístico

Com base em um estudo de caso desenvolvido na Alemanha e no Reino Unido que observou como jornalistas de dados e tecnólogos cívicos se completam, podemos descrever suas relações como algo que se dá em torno de duas práticas principais: facilitação e mediação.[225] Facilitar, neste caso, significa possibilitar a outros que ajam por conta própria, já a mediação se dá ao papel jornalístico tradicional de mediar informações relevantes ao público. Para ilustrar a diferença entre estes, sites de monitoramento parlamentar desenvolvidos por tecnólogos cívicos têm o objetivo de possibilitar aos seus usuários que *se informem eles mesmos*, através de pesquisas de discursos parlamentares (facilitação), mas sem *levar informação a eles de maneira proativa*, o tipo de informação considerada relevante por profissionais (mediação). Facilitação se relaciona com a autonomia de cada um, já a mediação tem a ver com direcionamento do debate público e causar impacto.

O que caracteriza o emaranhado entre jornalistas de dados e tecnólogos cívicos é o fato de que as práticas de facilitação e mediação são complementares, capazes de reforçarem umas às outras. Aplicações de tecnologia cívica não facilitam as coisas só para cidadãos comuns, jornalistas de dados podem usá-las em suas próprias apurações. O trabalho jornalístico, por sua vez, pode chamar a atenção para assuntos em específico e encorajar o público a utilizar estes serviços facilitadores. Além disso, o direito à informação é essencial

---

[224] Ver Openparliament.ca; WhatDoTheyKnow.com e FixMyStreet.com.

[225] https://www.tandfonline.com/doi/full/10.1080/21670811.2017.1375382.

para práticas de facilitação e mediação, e acaba por criar sinergias adicionais. Mais um exemplo: jornalistas podem usar seus direitos de acesso exclusivo a dados para compartilhá-los com tecnólogos cívicos, ao passo que estes mesmos jornalistas podem se beneficiar do ativismo por parte dos tecnólogos em prol de maior acesso à informação e à política aberta de dados.

Novas formas empreendedoras de jornalismo desempenham um papel específico na relação entre jornalismo de dados e tecnologia cívica, já que se mostram mais abertas para a expansão da noção tradicional de mediação aliada ao conceito de facilitação atrelado à tecnologia cívica. Podemos citar o caso da *ProPublica*, que criou diversos bancos de dados enormes, pesquisáveis, com o objetivo de facilitar não a interação entre cidadãos e seus governos, mas, sim, apurações jornalísticas por parte de redações locais às quais faltam recursos e conhecimento especializado para coletar, limpar e analisar dados por conta própria. Outra redação sem fins lucrativos, a alemã *Correctiv*, veio com uma abordagem semelhante e até mesmo integrou o site dedicado à liberdade de informação *Open Knowledge Foundation*, em sua seção alemã, a alguns de seus aplicativos para que os usuários possam solicitar mais informações diretamente, adicionadas automaticamente ao banco de dados da *Correctiv*.[226]

Estes casos ilustram o número crescente de organizações que ampliam a noção tradicional do que é jornalismo através da incorporação de práticas e valores associados a diversas culturas de dados, mas vemos também o oposto: jornalistas de dados que reagem às semelhanças na prática e nas aspirações vindas de outros campos do trabalho com dados e abraçam sua identidade profissional enquanto mediadores e contadores de histórias. Não que estes jornalistas rejeitem a tecnologia cívica, mas sua resposta é uma especialização ainda maior do ofício, se aproximando do conceito de jornalismo investigativo tradicional.

Oportunidades oferecidas por limites pouco definidos

Em suma, o emaranhar do jornalismo de dados com outros campos e culturas ligados a dados contribui para maior diversificação de como o jornalismo é entendido e praticado, seja rumo à expansão ou ao reforço de valores e identidades tradicionais. Tanto jornalistas quanto pesquisadores podem considerar o jornalismo de dados como um fenômeno embutido em transformações tecnológicas, culturais e econômicas mais amplas. Voltei-me ao emaranhar entre jornalistas de dados e tecnólogos cívicos ao longo deste artigo, mas gostaria de apontar duas lições essenciais para profissionais, relevantes para além deste caso em específico.

---

226 Correctiv.org.

Benefícios de limites pouco definidos: jornalistas tendem a descrever falta de limites profissionais em relação a outras áreas como problemática, mas a sinergia entre tecnólogos cívicos e jornalistas de dados mostra que isso pode ser uma vantagem. No lugar de encarar esta falta de definição como problemática, cabe aos jornalistas perguntarem se existem sinergias semelhantes com outros campos ligados a dados e como se beneficiar destas. Mais importante ainda, isso não significa que estes profissionais tenham que adotar práticas de facilitação para si mesmos. Por mais que existam exemplos disso, jornalistas que rejeitam esta ideia podem tentar encontrar novas formas de se beneficiar sem abrir mão de sua identidade profissional.

Adoção da diversidade no contexto do jornalismo profissional: as descobertas de minha pesquisa refletem como o jornalismo é realizado cada vez mais por variada série de atores, ainda mais especializados. Esta diversificação preocupa alguns dos jornalistas com quem conversei. Para eles, organizações de mídia que adotam práticas de facilitação podem enfraquecer sua noção de jornalismo investigativo "linha-dura". Porém, aos jornalistas é necessário reconhecer a improbabilidade de haver uma única forma definida de se fazer jornalismo no futuro.

Resumindo, uma maior conscientização dos laços históricos e contemporâneos a outros tipos de trabalho e cultura com/de dados pode ajudar jornalistas a refletirem sobre o seu próprio papel em meio a isso, além de gerar maior conhecimento a respeito de novas dependências e sinergias a serem usadas no apoio e na expansão em potencial de sua missão.

Stefan Baack concluiu seu doutorado no Centro de Estudos de Mídia e Jornalismo da Universidade de Groningen e estuda como uma dependência cada vez maior de dados ao longo de diversos setores da sociedade atravessa visões e práticas democráticas.

Referências

BAACK, Stefan. *Practically Engaged. The entanglements between data journalism and civic tech*. Digital Journalism, 6(6), 2018, p. 673–692. Disponível em: https://doi.org/10.1080/21670811.2017.1375382

USHER, Nikki. *Venture-backed News Startups and the Field of Journalism*. Digital Journalism, 5(9), 2017, p. 1116–1133. Disponível em: https://www.tandfonline.com/doi/full/10.1080/21670811.2016.1272064.

WAGEMANS, Andrea; WITSCHGE, Tamara; HARBERS, Frank. *Impact as driving force of journalistic and social change*. SAGE journals, 2018. Disponível em: https://doi.org/10.1177/1464884918770538.

Práticas de código aberto no contexto do jornalismo de dados

Ryan Pitts e Lindsay Muscato

Imagine o seguinte: alguns jornalistas trabalham juntos para coletar os registros de sites governamentais, transformar estes documentos em dados, analisar estes dados em busca de padrões e publicar uma visualização destas informações que conte uma história para os leitores. Versões deste processo ocorrem em redações pelo mundo todos os dias. Em muitas destas, cada passo depende, parcialmente ao menos, de algum software de código aberto, reunindo ferramentas testadas pela comunidade em um fluxo de trabalho mais ágil do que nunca.

Mas não foi só o software de código aberto que passou a fazer parte do trabalho atual do jornalismo de dados, há também a *filosofia* em torno do código aberto. Compartilhamos conhecimento e habilidades uns com os outros, em eventos e através de canais da própria comunidade e redes sociais. Publicamos metodologias e dados, em um convite para que colegas corrijam o que presumimos e dando motivo para que nossos leitores confiem nos resultados aos quais chegamos. Tais abordagens abertas e colaborativas podem fazer nosso jornalismo ainda melhor. Sempre que buscamos algum retorno ou contribuições externas, tornamos nosso trabalho mais resiliente. Outra pessoa pode notar um problema na forma como usamos dados em determinado produto, ou contribuir com uma nova funcionalidade que aprimora nosso software.

Tais práticas podem ter benefícios ainda maiores, para além de nossos projetos e organizações. Grande parte de nós nunca mergulhará fundo em um projeto sem usar nada além de ferramentas e técnicas desenvolvidas por conta própria. Em vez disso, nos baseamos nos trabalhos de outros, aprendemos com mentores, prestando atenção no que é dito em conferências e aprendendo sobre como projetos que gostamos foram realizados.

Na *OpenNews*, trabalhamos com jornalistas em projetos abertos, apoiamos colaborações entre desenvolvedores e até mesmo escrevemos o *Field Guide to Open Source in the Newsroom* (“Guia de Código Livre na Redação”, em tradução livre).[227] Ao longo deste capítulo refletiremos sobre algumas das coisas que aprendemos sobre o papel de práticas de código livre no contexto do jornalismo de dados, incluindo desafios comuns e características de projetos bem-sucedidos.

---

227 https://opennews.org/, http://fieldguide.opennews.org/.

Desafios comuns

Trabalhar de forma aberta pode ser recompensador, divertido e pode-se aprender mais no decorrer do processo, mas nem sempre é simples! Planejamento com sucesso em mente consiste em ter noção clara dos desafios comumente enfrentados por projetos deste tipo.

Vendendo a ideia

Pode ser difícil convencer editores, equipe legal e demais envolvidos de que "ceder" seu trabalho é uma boa. Pode haver preocupações envolvendo questões legais, de negócios, reputação e sustentabilidade. Em resposta a isso, viemos trabalhando com jornalistas na documentação dos benefícios em uma abordagem de código livre às ferramentas e processos usados, que incluem programação mais robusta, boa reputação junto à comunidade e maior credibilidade.[228]

As pessoas seguem em frente, a tecnologia também

Quando um membro fundamental de uma equipe aceita outro emprego, o tempo que ele teve para manter e atuar em prol de um projeto de código aberto vai junto. Por exemplo, há alguns anos, o *The New York Times* lançou a Pourover, uma estrutura em JavaScript que alimentava um serviço de filtragem rápida de conjuntos de dados gigantescos, de dentro do navegador. Amplamente compartilhada, deu início a uma comunidade própria. Mas um de seus principais desenvolvedores aceitou uma vaga em outro lugar e a equipe passou a buscar ferramentas novas para a solução de problemas semelhantes. Isso não é uma cutucada na forma como a Pourover foi criada ou planejada, é só que, às vezes, o tempo de vida de um projeto sai diferente do imaginado.

Pressões do sucesso

Pode soar contraintuitivo, mas descobrir que as pessoas estão empolgadas com algo que você criou pode gerar trabalho para o qual você não está preparado. Popularidade súbita e explosiva vem acompanhada de pressão extra para continuar trabalhando em cima de um projeto, consertar bugs e responder às contribuições da comunidade. Elliot Bentley se viu às voltas com tudo isso após lançar o Transcribe, aplicativo para web criado por ele para solucionar um problema de seu trabalho: a transcrição de entrevistas em áudio. Passados alguns meses, Elliot contava com dezenas de milhares de usuários ativos e perguntas sobre o futuro do projeto.

---

228 http://fieldguide.opennews.org/en/latest/Chapter01-Choosing-Open-Source/.

Características de projetos de sucesso

Há muitos exemplos de abordagens de código aberto aplicadas ao jornalismo, de projetos feitos por uma redação e adotados por muitas outras àqueles que eram colaborativos desde o início. Os esforços de maior sucesso que já vi compartilham uma ou mais das seguintes qualidades:

Resolvem um problema cotidiano

As chances de que alguém esteja empacando no mesmo lugar que você ou fazendo as mesmas tarefas repetitivas são bem prováveis. Cobrindo a justiça penal por todo o país, *The Marshall Project* monitora centenas de sites em busca de mudanças ou anúncios novos. Visitar uma lista de endereços repetidamente não é um exemplo do bom uso do tempo de um jornalista, mas é um excelente uso de servidores em nuvem. O software Klaxon fica de olho nestes sites e envia um alerta sempre que ocorre uma mudança, num processo tão rápido que a redação tem acesso às novas informações antes mesmo de serem anunciadas oficialmente.[229] Esse tipo de monitoramento é útil para qualquer área, e quando *The Marshall Project* solucionou um problema para os seus repórteres, resolveu os problemas para repórteres de outros veículos e organizações também. Ao lançar o Klaxon como código aberto, seus desenvolvedores ajudaram o trabalho de reportagem de dezenas de redações e passaram a receber colaborações que aprimoram sua ferramenta.

Resolvem problemas que não são divertidos de se lidar

A equipe de dados e visualizações da NPR precisava encontrar um jeito de fazer com que gráficos mudassem suas dimensões junto com as páginas nas quais estavam integrados. Trata-se de uma funcionalidade crítica, considerando que cada vez mais leitores usam dispositivos móveis para o consumo de notícias, mas não é o tipo de coisa que é divertida de se resolver. Quando a NPR lançou a Pym.js, biblioteca em código aberto que resolveu o problema, não demorou para que ela fosse adotada pela comunidade jornalística.

A documentação é ótima

Há uma enorme diferença entre só jogar o código na internet e explicar exatamente para que serve um projeto e como usá-lo. Prazos tendem a tornar a criação destas documentações uma das prioridades mais baixas dentro de um projeto, mas estes não podem prosperar sem elas. Novos usuários precisam saber por onde começar e você mesmo agradecerá ao revisitar seu trabalho, mais adiante. Existe um serviço chamado Wherewolf, em JavaScript, que pode ser usado para descobrir onde um endereço está localizado dentro de

[229] https://newsklaxon.org/.

uma série de limites (distritos escolares ou fronteiras municipais, por exemplo). O código em si não é atualizado há tempos, mas a comunidade de usuários segue em expansão, em parte porque a documentação relacionada é bastante minuciosa e cheia de exemplos.

Contribuições são bem-vindas

A California Civic Data Coalition oferece uma série de ferramentas de código aberto que ajuda repórteres a utilizarem os dados financeiros de campanhas estaduais. Teve início como uma colaboração entre desenvolvedores de duas redações, mas cresceu graças às contribuições de estudantes, estagiários, pessoal do setor de dados cívicos, cidadãos interessados e até mesmo jornalistas sem qualquer experiência em programação. Não foi à toa: a iniciativa contava com um planejamento de funcionalidades a serem implementadas e bugs a serem consertados, gerando tarefas para diferentes níveis de especialização, além de participar de conferências e sprints de planejamento abertos ao público.

Ryan Pitts é jornalista experiente e chefe de programas em tecnologia da OpenNews, organização sem fins lucrativos que ajuda desenvolvedores, designers e analistas de dados de redações a colaborarem em projetos de tecnologia e a contarem histórias através de linguagem de programação. Lindsay Muscato é editora da Source, plataforma de comunidade integrante da OpenNews.

Feudalismo de dados: como plataformas moldam redes de investigação transfronteiriças

Ştefan Cândea

A plataformização do jornalismo investigativo transfronteiriço é um fenômeno em expansão, endossado pelo mesmo positivismo tecnológico como tendência atual da plataformização da sociedade (Van Dijck, Poell, e De Waal, 2018). Plataformas de hospedagem de dados para investigações transfronteiriças começaram a ganhar proeminência por volta de 2019, no contexto de apurações envolvendo vazamento de dados. Talvez o exemplo mais notável de colaboração jornalística de grande porte baseada em plataforma sejam os *Panama Papers*, vencedores do prêmio Pulitzer.

De forma a organizar o processo de investigação e reportagem feito por 500 jornalistas envolvidos na empreitada, o Consórcio Internacional de Jornalistas Investigativos (ICIJ, na sigla em inglês) criou uma plataforma chamada "Global I-Hub" (Wilson-Chapman, 2017). Ryle (2017) descreve esta plataforma como "tecnologia desenvolvida especialmente [...] usada para interrogar e distribuir informação, conectar jornalistas em uma redação online e garantir que a equipe trabalhe como uma única equipe global", também chamada de "escritório virtual do ICIJ [...] um Facebook para jornalistas" pela equipe editorial e de pesquisa do próprio ICIJ (Hare, 2016; Raab, 2016).

Espera-se que dados e investigações transfronteiriças sejam uma espécie de par sem igual e uma forma de possibilitar colaborações jornalísticas independentes (Coronel, 2016; Houston, 2016). Organizações como ICIJ e Projeto de Reportagem sobre Crime Organizado e Corrupção (OCCRP, na sigla em inglês), dentre outras, oferecem a um grupo selecionado de centenas de jornalistas pelo mundo acesso gratuito (ou melhor, subsidiado) a conjuntos de dados exclusivos para pesquisa em uma plataforma eletrônica privada, inacessível para o mundo exterior. Oferecem, ainda, uma plataforma para publicação e anúncio de materiais produzidos por estes jornalistas.

Para estas organizações, o uso de tais plataformas possibilita maior escala e eficiência. Para os jornalistas individuais, acesso exclusivo e seguro, em um único lugar, a montanhas de dados — incluindo vazamentos, registros de empresas, resultados de solicitações de acesso à informação, arquivos, notas de repórteres, matérias antigas, arquivos digitalizados de processos e documentos judiciais — é um paraíso na terra, especialmente para aqueles que trabalham isolados e sem recursos para viajar, armazenar e processar dados.

Por mais que estes benefícios a curto prazo sejam reconhecidos, pesquisas mais detalhadas sobre como tais plataformas investigativas vêm moldando a posição e o trabalho de jornalistas individuais que as usam, e as redes das quais fazem parte, ainda precisam ser conduzidas.

Uma das consequências de termos tão poucos atores operando tais plataformas e grande número de profissionais dependentes destas no campo do jornalismo transfronteiriço poderia ser o que vêm sendo chamado de "um tipo de feudalismo hipermoderno" dentro da seara "big tech" baseada em posse e propriedade de dados (Morozov, 2016). Estes termos são utilizados para chamar atenção ao fato de que controle total dos dados de usuários e interações estão nas mãos de algumas poucas empresas, sem nenhum tipo de concorrência.

Um modelo operacional que traz consigo diversas preocupações, entre elas o controle de acesso. Por muitos bons motivos, o acesso a tais plataformas está atrás de várias camadas de segurança e nem todo jornalista o tem. As questões essenciais aqui são quem decide esse processo de inclusão e exclusão, quais regras regem estes processos e quais as possíveis tensões e conflitos que podem surgir em relação a estas. A participação em tais plataformas é regida, normalmente, por um acordo de confidencialidade básico ou de cooperação, em que os deveres do jornalista ou veículo receptor do acesso são listados detalhadamente, com referências escassas aos seus direitos, na maior parte dos casos. Tais sistemas e seus métodos de governança não são criados com princípios de copropriedade em mente, mas, sim, como estruturas centralizadas, o que inclui vigilância em torno da atividade do usuário e monitoramento de brechas de acordo com as funcionalidades embutidas.

Além do que, a adoção deste modelo, bem como o resto da "economia compartilhada", corre o risco de gerar um precariado dentro da área do próprio jornalismo investigativo. Indicadores deste risco são as descrições das organizações que operam estas plataformas. A OCCRP, por exemplo, se apresenta como "o Airbnb ou Uber dos jornalistas" que desejam realizar "grandes investigações transfronteiriças" (OCCRP, 2017).

Muitas vezes jornalistas trabalham sem receber nada em cima dos dados de posse destas organizações, tendo que pagar pelo acesso às informações com sua produção, com o risco de serem removidos da plataforma a qualquer instante. Apesar das condições nada favoráveis, cada vez mais jornalistas têm de permanecer ativos nestas mesmas plataformas para não serem removidos do jogo.

Por estes motivos, este modelo de negócios no contexto do intermediário de uma grande rede investigativa atual pode ser comparado às plataformas digitais da *gig economy*, ou economia dos bicos. O acesso pode ser revogado a qualquer instante, governança não é algo a ser discutido, o usuário é vigiado pela própria plataforma e "é melhor deixar o dinheiro

de fora da equação" (Buzenberg, 2015). O trabalho não remunerado e o "compartilhamento radical" presente nas interações entre centenas de jornalistas são "vendidos" a doadores, sem o compartilhamento de quaisquer lucros. Os dados vazados e o intercâmbio de informações que enriquece estes mesmos dados não são compartilhados com os usuários. Dados produzidos pela troca de informações entre usuários só são compartilhados novamente na forma de funcionalidades que tornam a plataforma mais eficiente e geram mais interação, mais usuários e, por extensão, mais doadores. O custo real dos serviços é desconhecido para os usuários.

Mas o que fazer para remediar essa tendência atual do mundo do jornalismo investigativo? Um primeiro passo fundamental é reconhecer que o compartilhamento de dados baseado em plataformas no contexto de uma rede de jornalismo investigativo precisa vir acompanhado de discussões em torno de regras de governança e design de tecnologia, copropriedade de dados e ferramentas digitais. Estas redes precisam desenvolver e adotar códigos de conduta públicos e mecanismos de transparência para lidar com abusos de qualquer tipo. A ausência destes pode piorar as condições precárias de trabalho de jornalistas individuais, em vez de romper com práticas do passado. Além disso, o objetivo não deve ser aumentar a escala de redes investigativas transfronteiriças para milhares de pessoas cada. Em seu lugar, deve-se buscar um bom modelo a ser aplicado às mais variadas redes passíveis de colaboração umas com as outras. Em vez de uma única rede de 150 parceiros de mídia, uma abordagem mais desejável seria ter 10 redes com 15 parceiros cada. Esta última seria adequada a um sistema de mídia saudável, incluindo concorrência justa e pluralismo midiático. Sem tais abordagens, o potencial participativo de redes investigativas transfronteiriças fracassará em se materializar e, por efeito de rede, algumas poucas plataformas consolidarão um sistema investigativo feudal de dados global.

Ștefan Cândea é pesquisador de doutorado na CAMRI, Universidade de Westminster, membro do ICIJ e coordenador do EIC.

Referências

BUZENBERG, W. *Anatomy of a Global Investigation: Collaborative, Data-Driven, Without Borders.* Shorenstein Center, 2015. Disponível em: https://web.archive.org/web/20191210111153/https://shorensteincenter.org/anatomy-of-a-global-investigation-william-buzenberg/. Acesso em: 10 de dezembro de 2019.

CORONEL, S. *Coronel: A Golden Age of Global Muckraking at Hand*. Global Investigative Journalism Network, 2016. Disponível em: https://web.archive.org/web/

20191210111301/https://gijn.org/2016/06/20/a-golden-age-of-global-muckraking/. Acesso em: 10 de dezembro de 2019.

DIJCK, J. et al. *The Platform Society: Public Values in a Connective World*. Nova York: Oxford University Press, 2018.

HARE, K. *How ICIJ got hundreds of journalists to collaborate on the Panama Papers*. Poynter, 2016. Disponível em: https://web.archive.org/web/20191210111558/https://www.poynter.org/reporting-editing/2016/how-icij-got-hundreds-of-journalists-to-collaborate-on-the-panama-papers/. Acesso em: 10 de dezembro de 2019.

HOUSTON, B. *Panama Papers Showcase Power of a Global Movement*. Global Investigative Journalism Network, 2016. Disponível em: https://web.archive.org/web/20191210111725/https://gijn.org/2016/04/13/panama-papers-showcase-power-of-a-global-movement/. Acesso em: 10 de dezembro de 2019.

MOROZOV, E. *Tech titans are busy privatising our data*. The Guardian, 2016. Disponível em: https://www.theguardian.com/commentisfree/2016/apr/24/the-new-feudalism-silicon-valley-overlords-advertising-necessary-evil. Acesso em: 10 de dezembro de 2019.

OCCRP. *2016 Annual Report*. Sarajevo: OCCRP, 2017. Disponível em: https://web.archive.org/web/20191210112002/https://www.occrp.org/documents/AnnualReport2017.pdf. Acesso em: 10 de dezembro de 2019.

RAAB, B. *Behind the Panama Papers: A QeA with International Consortium of Investigative Journalists Director Gerard Ryle*. Ford Foundation, 2016. Disponível em: https://web.archive.org/web/20191210112045/https://www.fordfoundation.org/ideas/equals-change-blog/posts/behind-the-panama-papers-a-qa-with-international-consortium-of-investigative-journalists-director-gerard-ryle/. Acesso em: 10 de dezembro de 2019.

RYLE, G. *Paradise Papers: More documents, more reporters, more Revelations. ICIJ*, 2017. Disponível em: https://web.archive.org/web/20191210112134/https://www.icij.org/blog/2017/11/more-documents-more-journalists-and-bigger-revelations/. Acesso em: 10 de dezembro de 2019.

WILSON-CHAPMAN, A. *Panama Papers a 'notable security success'*. ICIJ, 2017. Disponível em: https://web.archive.org/web/20191210112217/https://www.icij.org/blog/2017/08/panama-papers-notable-security-success/. Acesso em: 10 de dezembro de 2019.

Editorial baseado em dados? Considerações acerca de métricas de audiência[230]

Caitlin Petre

Em 23 de agosto de 2013, o site de notícias satíricas *The Onion* publicou um editorial supostamente escrito pela editora digital da *CNN*, Meredith Artley, intitulado "Deixe-me explicar porque a apresentação de Miley Cyrus no VMA foi nossa matéria mais visualizada esta manhã". A resposta, explicava o texto, era "bem simples":

> Foi uma tentativa de fazer com que você fosse até o endereço CNN.com para que aumentássemos nosso tráfego, o que por sua vez nos permitiria aumentar nossos lucros com publicidade. Não havia nada, nada mesmo, nessa matéria que tivesse qualquer relação com as notícias relevantes do dia, nenhuma crônica de eventos significantes ou mesmo a ideia de que o jornalismo pode ser uma força rumo a uma mudança positiva no mundo, mas meu deus, como nos deu audiência.

No decorrer do texto, mencionavam-se métricas específicas como visualizações de página e taxas de rejeição como fatores que motivaram a *CNN* a dar destaque à matéria sobre Miley Cyrus em sua página principal.

Claro, Artley não escreveu nada daquilo, mas alguns círculos sentiram uma pontada mesmo assim, ainda mais se levarmos em conta que a infame apresentação de Cyrus no MTV Video Music Awards havia, de fato, ocupado um espaço proeminente no site *CNN.com*, e a própria Meredith Artley confirmou, posteriormente, que aquela matéria havia, sim, gerado o maior tráfego visto pelo veículo naquele dia. O editorial falso pode ser encarado não só como uma crítica à *CNN*, mas também um comentário mais abrangente sobre o triste estado do noticiário em meio à era das métricas de internet.

Empresas de mídia sempre se esforçaram para coletar dados demográficos e de comportamento de seu público, mas as capacidades de monitoramento da internet, bem como a de armazenar e analisar grandes volumes de dados, significam que métricas de público sofisticaram-se cada vez mais nos últimos anos. Além das visualizações e taxas de rejeição mencionadas anteriormente, ferramentas analíticas monitoram variáveis como taxas de

---

230 Texto adaptado de *The Traffic Factories: Metrics at Chartbeat, Gawker Media, and The New York Times*, publicado originalmente pelo Centro Tow de Jornalismo Digital da Escola de Graduação em Jornalismo da Universidade de Columbia, em 2015. Republicado com permissão.

retorno de visitantes, sites de referência, profundidade de rolagem e tempo gasto em uma página. Muitas destas informações são entregues aos veículos de notícias em tempo real.

Painéis de métricas são quase que onipresentes nas redações contemporâneas, com frequentes debates emocionados sobre como e quando devem ser consultados, tão frequentes quanto o próprio uso destas métricas. Não é de surpreender que o assunto tenha se tornado uma questão controversa dentro do jornalismo. Sua presença traz à tona uma série de tensões sempre presentes no cenário dos meios de comunicação comerciais, dentre as quais: qual a missão fundamental do jornalismo e como as organizações de notícias podem determinar quando cumpriram esta missão? Como empresas de comunicação podem reconciliar seu imperativo de lucros e seu imperativo cívico? Até que ponto a distinção entre jornalista e público ainda é relevante, que relação os jornalistas devem ter com seus leitores? Métricas de audiência fazem parte do cotidiano das empresas de notícias, mas há pouca pesquisa empírica sobre como estes dados são produzidos ou como afetam a cultura da redação e a rotina dos jornalistas.

Com o apoio do Centro Tow de Jornalismo Digital da Universidade de Columbia, dei início a um longo projeto de pesquisa etnográfica para entender como o uso de métricas altera o comportamento dos jornalistas e o que isso significa para o jornalismo em si. Dentre as principais perguntas de minha pesquisa estava a seguinte: como estas métricas são produzidas?

Ou seja, como programadores, cientistas de dados, designers, líderes de produto, marqueteiros e equipes de vendas que criam e comercializam estas ferramentas decidem quais aspectos do comportamento do público serão mensurados e como mensurá-los? Quais ideias — a respeito de quem tem seu comportamento mensurado (consumidores de notícias) e de quem usará a ferramenta (jornalistas) — estão embutidas nestas decisões? Como empresas de análise comunicam o valor destas métricas para empresas de comunicação?

Segundo: como estas métricas são interpretadas? Apesar de suas posições opostas, argumentos em torno de como métricas são boas ou ruins para a prática jornalística têm um ponto em comum, ambos tendem a presumir que o significado destas métricas é claro e direto. Mas um número por si só não significa nada sem uma estrutura conceitual para interpretá-lo. Quem interpreta estas métricas e como fazem isso?

Terceiro, como métricas são utilizadas dentro do jornalismo? Estes dados têm algum papel na forma como redações distribuem, produzem e promovem pautas? De que maneiras, se é que isso ocorre, dados são considerados na hora de tomar decisões relacionadas à equipe, como aumentos, promoções e demissões? Estes dados têm um papel mais relevante no

trabalho diário ou em uma estratégia de longo prazo? E como as respostas a estas perguntas diferem ao longo de contextos organizacionais variados?

Para responder a tudo isso, conduzi um estudo etnográfico do papel das métricas no jornalismo contemporâneo ao examinar três estudos de caso: Chartbeat, *Gawker Media*, e *The New York Times*. Combinando observações e entrevistas com gerentes de produto, cientistas de dados, repórteres, blogueiros, editores e outros, minha intenção era desvendar premissas e valores subjacentes às métricas de audiência, o efeito destas métricas no trabalho cotidiano de jornalistas e as formas pelas quais estas interagem com a cultura organizacional. A seguir, falo de algumas de minhas principais descobertas.

Em primeiro lugar, painéis analíticos têm importantes dimensões emocionais que muitas vezes são deixadas de lados. Métricas e o fenômeno mais abrangente da "big data", do qual faze parte, são descritas como uma força de racionalização, ou seja, permitem que as pessoas tomem decisões com base em informações objetivas e não intuições ou julgamentos nada confiáveis. Não é um retrato incorreto das coisas, mas está incompleto. O poder e apelo das métricas firmam-se em grande parte na capacidade de dados causarem sentimentos em específico, como empolgação, decepção, validação e segurança. Na Chartbeat, sabia-se que esta valência emocional era uma parte poderosa do apelo de um painel, e a empresa incluía funcionalidades para engendrar estas emoções em seus usuários. Por exemplo, o painel foi projeto para transmitir uma sensação de consideração pelo julgamento do jornalista, amaciar o impacto do baixo tráfego e oferecer oportunidades para comemoração dentro das redações.

Além disso, o impacto de uma ferramenta de análise depende da organização utilizando-a. Muitas vezes, presume-se que a presença de uma ferramenta como essa mudará a forma de operação de uma redação de maneiras específicas, porém, descobri que o contexto organizacional influencia grandemente como e se métricas afetam a produção de notícias. Tanto *Gawker Media* e *The New York Times* são clientes da Chartbeat, mas a ferramenta surge em contextos diferentes em cada cenário. No caso da Gawker, métricas tinham altos índices de influência e visibilidade. Já no *Times*, não tanto, sendo usadas para corroborar decisões que os editores já haviam tomado. Isso sugere que é impossível saber como métricas afetam o jornalismo sem examinar como estão sendo utilizadas dentro de redações

Por fim, para quem escreve, uma cultura baseada em métricas pode ser, ao mesmo tempo, uma fonte de estresse e de segurança; surpreendentemente, também é bastante compatível com visões de liberdade editorial. Enquanto os redatores da *Gawker Media* consideravam a pressão relacionada ao tráfego estressante, muitos eram bem mais afetados, de um ponto de vista psicológico, pelo ódio online presente em comentários e redes sociais. Em meio a um clima de hostilidade ou até mesmo de intimidação, redatores usavam métricas como lembrete de sua competência profissional. É interessante notar, porém, que redatores e

editores geralmente não encaravam os sistemas de avaliação da empresa, baseados em tráfego, como obstáculo para sua autonomia editorial. Isso sugere que jornalistas em empresas operando somente em ambientes online, como a *Gawker Media,* possam ter noções diferentes de liberdades e restrições editoriais em relação a outras empresas de comunicação.

Com base nestas descobertas, faço as seguintes recomendações a empresas de comunicação em geral: primeiro, priorizar o pensamento estratégico quando se trata de questões ligadas à análise (o papel apropriado das métricas dentro da empresa e as formas pelas quais dados interagem com os seus objetivos jornalísticos, por exemplo). A maior parte dos jornalistas está ocupada o bastante com suas tarefas diárias para se debruçar sobre o papel das métricas em seu veículo, ou quais métricas melhor complementam seus objetivos jornalísticos. Com base nisso, estes jornalistas tendem a consultar, interpretar e usar métricas de acordo com a necessidade de cada ocasião. Mas há um problema: são dados poderosos demais para serem implementados assim, de supetão. Cabe às redações criarem oportunidades — internas ou através da parceria com pesquisadores externos — para raciocínio reflexivo e deliberado, distante das pressões diárias, a respeito de como melhor usar os frutos de análise.

Segundo, ao escolher um serviço de análise, chefes de redação devem olhar para além das ferramentas e considerar quais objetivos estratégicos, imperativos de negócios e valores de cada fornecedor melhor complementam o de sua redação. Tendemos a ver números, e por tabela painéis de análise, como reflexões racionais e dominantes do mundo empírico. Ao escolher um serviço de análise, porém, é importante lembrar que as empresas que prestam estes serviços têm seus próprios interesses.

Terceiro, ao desenvolver políticas internas para o uso de métricas, chefes de redação devem levar em conta os possíveis efeitos de tráfego de dados não só no conteúdo editorial, mas também em quem trabalha na editoria. Quando rankings começam a ter um lugar de destaque no mural de uma redação ou um site, pode ser difícil limitar sua influência. Rankings baseados em tráfego podem se sobrepor a outras formas de avaliação, mesmo que não seja essa a intenção.

Por fim, esforços para o desenvolvimento de melhores métricas mostram-se necessários e valorosos, mas redações e empresas de análise devem prestar atenção às limitações destas métricas. Ao passo que prioridades e sistemas de avaliação organizacionais cada vez mais tomam métricas por base, há o perigo de confundir o que é quantitativamente mensurável com o que de fato tem valor. Nem tudo pode, ou deve, ser contabilizado. Redações, empresas de análise, financiadores e pesquisadores de mídia devem levar em conta como algumas das características mais atraentes e indispensáveis do jornalismo, caso de sua missão social, não podem ser mensuradas facilmente. Em um momento de constante

valorização de análise de dados, devemos tomar cuidado para não igualar o que é quantificável com o que é valoroso.

Caitlin Petre é socióloga e seu trabalho se dedica a examinar as implicações sociais e materiais de um mundo cada vez mais saturado de dados, com foco em particular na relação entre tecnologias digitais, conhecimento especializado e indústria de comunicação.

Treinamento para jornalistas de dados

Jornalismo de dados, universalismo digital e inovação na periferia mundial

Anita Say Chan

O universalismo digital é a estrutura onipresente, ainda que errônea, moldando o imaginário global em torno do digital que pressupõe que uma única narrativa universal levada adiante por "centros" de inovação seja capaz de representar de forma precisa as formas pelas quais o desenvolvimento digital vem se dando pelo mundo nos dias de hoje. Pressupõe que estes centros de "inovação" e design tecnológico contemporâneos inevitavelmente determinarão o modelo de futuro digital a se espalhar pelo resto do mundo, para a maioria do "restante digital". Um conceito que é ressoado pela suposição, feita casualmente, de que os melhores e mais "legítimos" locais de estudo e observação de transformações tecnológicas, produtividade e prática digitais ou inovação e investigação com base em informações vêm destes centros. De destaque entre estes: laboratórios, escritórios e centros de pesquisa instalados no Vale do Silício e seus equivalentes espalhados por outras capitais da inovação pelo mundo, concentradoras de uma espécie de elite do conhecimento digital. É a partir destes centros que, presume-se, surge a cultura digital, em sua forma e manifestação mais pura, de forma a ser replicada em outros lugares; é ali que as visões de futurismo digital em suas acepções mais precisas ou ideais surgem; e é ali que avanços tecnológicos e, por conseguinte, avanços na cultura digital, surgem no contexto que entendemos em grande parte como o que há de mais dinâmico, vivo e inspirado. Em outras palavras, presume-se que a cultura digital, apesar de suas dimensões globais únicas, vive em lugares mais "autênticos" e produtivos de onde pode-se basear estudos e observar suas dinâmicas.

No papel de uma jovem pesquisadora estudando e escrevendo sobre ativismo cultural digital e política no Peru e na América Latina desde o começo dos anos 2000, isso moldou minha experiência de maneiras fundamentais. Muitas vezes me deparo com a pergunta aparentemente inocente: "Por que ir ao Peru ou à América Latina para estudar cultura digital?". Não havia locais "melhores" onde poderia estudar o tema e meu tempo não seria melhor gasto participando e documentando atividades no Vale do Silício, por exemplo? Para quem faz esse tipo de pergunta, a imagem que se tem do Peru é de um país sul-americano montanhoso que já foi o coração da civilização inca, terra que abriga Machu Picchu, boa parte dos Andes e enormes populações que se comunicam em quéchua e aymara. Este mesmo Peru pode ser conhecido como espaço ideal para observar tradições do passado, cultura nativa ou a abundância da natureza, mas pouco sobre a cultura digital contemporânea, de acordo com esta linha de raciocínio. O mesmo se aplica aos fluxos de tecnologia e aos

desdobramentos futuros associados. Locais como o Peru podem levar a um caminho cheio de relíquias e tesouros de um passado tecnológico que teríamos que nos esforçar para *não* esquecer, em outras palavras. Ao passo que lugares como o Vale do Silício abrigam segredos de um futuro tecnológico cujo caminho ainda precisamos trilhar por inteiro, lugares onde estes segredos seriam desvendados. Esta pergunta esconde uma assertiva: a leve certeza de que futuros digitais imaginados por uma população selecionada de tecnólogos em centros de design de elite podem falar pelo resto do mundo, e que o presente que se desdobra em centros de inovação por aí certamente representa o futuro da periferia mundial.

A força do mito deste universalismo digital se manifesta não só através dos meios pelos quais fixa narrativas e imaginário público em torno de centros já estabelecidos de inovação, mas também pelo modo como desencoraja observar dinâmicas digitais para além destes centros. Desta maneira, estreita a diversidade e a circulação global de narrativas em torno de dinâmicas digitais *reais* ocorrendo em grande variedade de lugares, invisibiliza formas diversas de generatividade digital e amplifica e reforça, artificialmente, a representação destas capitais da "inovação" como locais exclusivos de produtividade digital.

Há uma visão especialmente colonialista da periferia aqui, do tipo que jornalistas e estudiosos de culturas digitais globais devem tomar cuidado para não reproduzir: a de que a "periferia" serve de agente para imitação global ou zona de difusão para um futuro inventado antes em algum outro lugar. Cabe notar que a periferia está longe de ser passiva ou pouco criativa. Feiras livres ou cybercafés animados e dinâmicos, lotados de computadores e componentes usados, reciclados, remontados representam algumas das inovações do Sul Global que estenderam o acesso de baixo custo à internet e aumentaram a escala da circulação de conteúdo de mídia global e local para incluir populações de zonas rurais e urbanas. Estas gambiarras tecnológicas e improvisos locais fazem parte do panorama tecnológico da periferia, cuja vibração é parcialmente captada ao comparar-se com as redes comerciais formais de bens digitais ou fornecedores de computadores e internet. Como observado pelos cientistas sociais Daniel Miller e Don Slater (2001) em seu estudo de Trinidad, "a internet não é um ciberespaço monolítico", existindo como uma tecnologia globalmente expansível com diversas realidades locais, práticas a serem assimiladas e políticas culturais em torno de suas várias localizações. Houve, de fato, várias maneiras de se imaginar como seria a prática e conexão digital.

No Peru, evidências de culturas digitais em plena atividade aproximaram diversos atores e interesses de formas muitas vezes inesperadas e contraditórias que já se mostravam visíveis. Coletivos em defesa do software livre, que haviam ajudado a realizar a primeira conferência sobre uso de software livre na América Latina — com apoio da ONU, um evento divisor de águas, realizado na capital inca de Cuzco em 2003 —, buscavam reestruturar a adoção destas tecnologias abertas. E, ainda, mudar essa visão não só como uma questão de

liberdades individuais e escolhas, como havia sido no caso do ativismo do software livre de código aberto (FLOSS, na sigla em inglês) nos EUA, mas também de diversidade cultural, transparência estatal e soberania política em relação ao poder monopolizante de empresas transnacionais no Sul Global. Salas de aula voltadas à "inovação digital" instaladas em escolas rurais pelo estado seriam convertidas na maior rede de implementação do projeto *One Laptop per Child* (um notebook por criança, OLPC na sigla em inglês) do MIT, alguns muitos anos depois, em nome da inclusão digital. Novas propriedades intelectuais aplicadas de forma agressiva por programas estatais a bens "tradicionais" prometiam transformar produtores rurais e artesões em novas classes de "trabalhadores da informação" prontos para exportação, com base em crescentes iniciativas voltadas à sociedade da informação realizadas pelo país. Defensores do software livre e de código aberto, juntos de ativistas de tecnologia em Cuzco, "aulas de inovação" promovidas pelo estado em escolas rurais e artesões no papel de "trabalhadores da informação" não representam os interesses ou protagonistas convencionais que se espera de um centro de cultura digital. Ver o desdobramento de cada uma dessas histórias foi como ver cada uma de suas particularidades se sobrepondo às estruturas e narrativas desta tal cultura digital. O imaginário global em torno da TI no novo milênio, afinal de contas, fez dos hackers do Vale do Silício um modelo a ser seguido por engenheiros. E as empreitadas estratégicas de empreendedores de tecnologia competitivos serviram de base para grandes filmes de Hollywood e contas com muitos seguidores no Twitter. Eis um elenco de atores, heróis e vilões cada vez mais reconhecíveis. Mas, para capturar as interações dinâmicas e experimentos tensos em cultura digital no Peru, é necessário prestar atenção a muitos outros interesses, agentes e desdobramentos — aqueles que tentaram construir novas ligações e trocas entre rural e urbano no contexto do digital, entre a tecnologia de ponta e o tradicional, com orientações distintas em torno do global com intenso compromisso local.

Atualmente, jornalistas de dados contam com um número cada vez maior de ferramentas e recursos tecnológicos/digitais para testemunharem, capturarem e relembrarem culturas e atividades digitais em diversos locais pelo mundo. Antes mesmo da onda de protestos que tomou o Ocidente Médio no começo de 2011, mídias digitais em rede oferecem novas capacidades de transmissão global para movimentos que adotaram usos estratégicos de redes sociais em contextos tão diversos quanto os de México (Schultz, 2007), Irã (Burns e Eltham, 2009; Grossman, 2009), Filipinas (Uy-Tioco, 2003; Vincente, 2003), e Ucrânia (Morozov, 2010). Na esteira da Primavera Árabe de 2011, movimentos como os Indignados de 15 de Março, na Espanha, e o Occupy dos norte-americanos fizeram uso estratégico de hashtags para organização e ativismo em redes sociais. Mais recentemente, movimentos como #MeToo e #BlackLivesMatter, originados nos EUA, acompanharam mobilizações globais como a #NiUnaMenos, da América Latina; a nigeriana #BringBackOurGirls; #Sosblakaustralia, da Austrália; #Idlenomore, de nativos canadenses; e #UmbrellaRevolution,

de Hong Kong. O crescente fluxo de mídia gerada por usuários destes movimentos multiplica práticas de dados cívicos, descentralizando as aplicações dominantes de "big data" em plataformas de redes sociais com vieses de perfilamento orientado ao marketing. Em vez disso, valem-se de práticas de dados em prol de novos tipos de narrativa que rompem com os centros estabelecidos de comunicação e notícias, ao mesmo tempo que cedem suas informações e evidência online da extensão de seu público a documentaristas, jornalistas e organizadores espalhados geograficamente.

Mas o crescimento no número de fontes de informações digitais e repositórios de dados —de acervos online criados por movimentos sociais em redes sociais a formas paralelas de ativismo criativo com dados — também aumenta os riscos para jornalistas do setor. Destacando-se entre estes riscos está a habilidade sedutora da big data e de plataformas de redes sociais de valerem-se da abundância de dados e informações coletados para convencer o público de que seu monitoramento extensivo compila e cria a melhor maneira de documentar a atividade e experiência humana no momento, além de formas de avaliar e *prever* o futuro de suas ramificações políticas ou econômicas. A *presuntividade* temporal da projeção dos universalistas digitais de que as formas do "presente" digital cultivadas em centros de inovação hoje poderão e representarão com precisão os futuros digitais de periferias globais encontra um novo complemento nas afirmações feitas a respeito das capacidades de previsão do processamento de dados algorítmicos pela indústria de dados. Tais pronunciamentos permanecem, apesar de flagrantes fracassos contemporâneos de grandes analistas de dados políticos, empresas de redes sociais e comentaristas do Ocidente em preverem com precisão grandes mudanças políticas dos últimos anos, como a eleição presidencial dos EUA em 2016, o Brexit, o escândalo da Cambridge Analytica e a "surpresa" da ascensão de movimentos *alt-right* no Ocidente.

Os jornalistas de dados de hoje devem evitar ter acervos e monitoramento de dados, independentemente de quão extensos sejam, como a única ou dominante forma de documentar, falar em prol ou avaliar as diversas realidades sociais nas quais o público depende destes profissionais. Em paralelo aos crescentes pedidos de estudiosos latino-americanos ou pós-coloniais para que se ampliem métodos de pesquisa e documentação para incluir o que e quem representa informação, tecnologia e novas culturas de mídia sob uma estrutura de "computação decolonial" (Chan, 2018), jornalistas de dados críticos da abordagem digital universalista precisam, de maneira consciente, diversificar suas fontes de informações e descentralizar métodos que privilegiam big data como forma exclusiva ou mais legítima de mapear eventos empíricos e realidades sociais. Movimentos rumo a uma "decolonização do conhecimento" sublinham a significância dos diversos modos pelos quais cidadãos e pesquisadores no Sul Global interagem com práticas de dados de baixo para cima. Estas se valem de ênfase em práticas comunitárias e meios centrados em indivíduos de

avaliar e interpretar dados para mudança social, assim como para falar da resistência ao uso de big data, que aumenta opressão, desigualdade e danos sociais.

Jornalistas de dados críticos da expansão do universalismo digital no universalismo de dados devem buscar aliados e preocupações em comum para o desenvolvimento de uma maneira de operar com dados de forma transparente e responsável junto a estudiosos em estudos críticos de dados, algoritmos, software e plataforma, bem como de computação pós-colonial. Isso inclui uma rejeição reforçada do fundamentalismo de dados (Boyd e Crawford, 2012) e determinismo tecnológico em torno de relatos convencionais sobre algoritmos em aplicação, e envolve uma recentralização fundamental do ser humano dentro de mundos dataficados e indústrias de dados — resistente ao impulso de interpretar big data "e algoritmos como objetos fetichizados… resistindo firmemente a colocar a tecnologia no explanatório banco do motorista" (Crawford, 2016). Também envolve tratar infraestruturas de dados e os algoritmos subjacentes que dão vida política a estas como ambíguas e acessíveis, intencionalmente, de forma a desenvolver metodologias que não só exploram novos cenários empíricos e cotidianos para política de dados — seja segurança em aeroportos, score de crédito, monitoramento hospital-paciente ou redes sociais ao longo de diversos pontos globais —, mas também encontrar formas criativas de fazer dados produtivos para análise (Gillespie, 2013; Ziewitz, 2016).

Por fim, talvez valha a pena lembrar que preservar os atuais centros de inovação digital como focos exclusivos de invenção digital ou criação de futuros em dados, claro, negligencia outro detalhe crucial: os centros do presente também já estiveram na periferia. Focar nestes centros como criadores de modelos que passam a ser adotados e copiados por aí se baseia na ideia de que funções e forças replicativas se estendem de maneira perfeita e contínua. Fracassa, porém, em não considerar a possibilidade de mudança dentro de um sistema maior e desestabilizações e realinhamentos de centros anteriores, logo, deixa de considerar os realinhamentos do que antes eram periferias. A "surpresa" da Primavera Árabe de 2011 e sua influência ao longo de diversos centros no Ocidente e no Oriente; a ascensão recente de mercados digitais não ocidentais e concorrentes econômicos em nações categorizadas como "em desenvolvimento" há menos de duas décadas; e a desestabilização de democracias ocidentais robustas de hoje servem como lembrete de que o equilíbrio de poderes estabelecidos e a permanência da relação centro-periferia podem ser questionados. Longe de ficar para trás ou imitar outros centros, as atividades dinâmicas da periferia sugerem como agentes que outrora não tinham grande influência podem emergir como fontes novas de produtividade distinta. Seu desenrolar diverso joga para escanteio a afirmação não dita de que uma única narrativa universal poderia representar adequadamente distintos futuros e imaginários digitais em uma série de centros hoje.

Anita Say Chan é professora adjunta do Departamento de Estudos em Cinema + Mídia e fellow do Centro Nacional para Aplicações de Supercomputação da Universidade de Illinois Urbana-Champaign.

Referências

BOYD, Danah; CRAWFORD, Kate. *Critical Questions for Big Data*. Information, Communication e Society 15, nº. 5, 2012, p. 662–679.

BURNS, Alex; ELTHAM, Ben. *Twitter-Free Iran An Evaluation of Twitter's Role in Public Diplomacy and Information Operations in Iran's 2009 Election Crisis*. In: PAPANDREA, F.; ARMSTRONG, M. (ed.). *Record of the Communications Policy e Research Forum 2009*. Sydney: Network Insight Institute, 2009.

CHAN, Anita. *Decolonial Computing and Networking Beyond Digital Universalism*. Catalyst, 4(2), 2018.

CRAWFORD, Kate. *Can an Algorithm be Agonnaic? Ten Scenes from Life in Calculated Publics*. Science, Technology e Human Values, 41(1), 2016, p. 77-92.

GILLESPIE, Tarleton. *The Relevance of Algorithms*. In: GILLESPIE, Tarleton; BOCZKOWSKI, Pablo J., FOOT, Kirsten A. (ed.). *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge: MIT Press, 2014, p. 167-194.

GROSSMAN, Lev. *Iran's Protests: Why Twitter Is the Medium of the Movement*. Time Magazine, junho de 2017.

MILLER, Daniel; SLATER, Don. *The naternet: An Ethnographic Approach*. Londres: Berg Publishers, 2001.

MOROZOV, Evgeny. *The Net Delusion: The Dark Side of Internet Freedom*. Nova York: Public Affairs, 2010.

SCHULZ, Markus. *The Role of the Internet in Transnational Mobilization: A Case Study of the Zapatista Movement, 1994–2005*. In: HERKENRATH, Mark (ed.). *Civil Society: Local and Regional Responses to Global Challenges*. Piscataway: Transaction Publishers, 2007.

UY-TIOCO, Cecilia. *The Cell Phone and EDSA 2: The Role of a Communication Technology in Ousting a President*. Critical Themes in Media Studies Conference, New School University, 11 de outubro de 2003.

VICENTE, Rafael. *The Cell Phone and the Crowd: Messianic Politics in the Contemporary Philippines*. Public Culture 15 (3), 2003, p. 399–425.

ZIEWITZ, M. *Governing algorithms: Myth, mess, and methods*. Science, Technology e Human Values 41(4), 2016, p. 3–16.

Jornalismo de dados feito por, sobre e para comunidades marginalizadas

Eva Constantaras

Atuo como jornalista de dados em países onde, geralmente, considera-se que as coisas vão muito mal, não por se tratar de um momento difícil ou solavanco político, mas, sim, sistemas políticos e econômicos inteiros em estado de falência. Em tais lugares, vemos notícias sobre como a corrupção paralisou o governo, cidadãos seguem desesperançosos e a sociedade civil vive em estado de sítio. Tudo está péssimo. Produzir jornalismo de dados em alguns dos lugares mais pobres, sem acesso à educação e perigosos do mundo me fez chegar a uma importante conclusão sobre a prática. Injustiça, desigualdade e discriminação são onipresentes, atuam nas sombras e são subestimadas na maioria dos países. Os jornalistas com quem trabalho vêm abraçando, sem hesitar, novas ferramentas que possibilitam, pela primeira vez, medir o quão ruins as coisas estão, quem sofre por conta disso, de quem é a culpa e como melhorar a situação. Dentro destes contextos, jornalistas tomaram para si os dados como meio de influenciar a política, mobilizar cidadãos e combater propaganda. Apesar das restrições à livre imprensa, o jornalismo de dados é encarado como um caminho rumo à autonomia.

O que trago à baila neste texto e gostaria de explorar é o compromisso do jornalismo de dados feito por, sobre e para comunidades marginalizadas. Ao lidar com diversos aspectos da injustiça, desigualdade e discriminação, bem como suas consequências nas vidas de comunidades marginalizadas, tornamos estes problemas visíveis, mensuráveis e, com sorte, solucionáveis. Estas histórias engajam jornalistas cujas raízes profundas se encontram em comunidades marginalizadas. Eles tratam de temas com os quais grupos que sofrem discriminação institucional se importam para mobilizar cidadãos. São materiais disseminados por meios de comunicação de massa locais de forma a atingir o máximo de gente e pressionar governos a tomarem decisões melhores para todo o país. Cito abaixo cinco tipos de produtos jornalísticos envolvendo jornalismo de dados que atendem aos interesses e preocupações de comunidades marginalizadas no Afeganistão, Paquistão, Quênia, Quirguistão e Países Balcãs.

1. Por que nosso povo está passando fome se o país tem recursos para alimentar a todos?

No Quênia, doadores financiavam os programas de alimentação errados. Uma matéria de 12 minutos na televisão, transmitida pela NTV e feita por Mercy Juma, a respeito de Turkana, região isolada e empobrecida no norte do país, revelou que a desnutrição infantil era

um problema crescente por conta da seca e da fome cada vez mais frequentes e intensas. O dinheiro ia parar nas mãos de programas emergenciais de alimentação, não de iniciativas de longo prazo voltadas à seca. O mesmo volume de dinheiro gasto em um ano nestes programas poderia financiar uma iniciativa de sustentabilidade para toda a região e seus quase um milhão de habitantes, de acordo com projetos no Parlamento. Mercy ameaçou não levar a matéria ao ar quando seus editores quiseram cortar os dados ali mencionados, afinal, o trabalho dependia de influenciar doadores, causar ultraje aos cidadãos e envergonhar o governo através, em grande parte, da televisão, mas também em versão impressa e resumida na internet.[231]

Ela convenceu os doadores com a robustez de seus dados. Coletou informações climáticas, agriculturais e de saúde junto a ministérios, pesquisas de saúde pública, agências doadoras e à Cruz Vermelha do Quênia. A missão queniana da Agência dos Estados Unidos para o Desenvolvimento Internacional (USAID, na sigla em inglês) se deparou com uma visualização de dados demonstrando que um ano de sua ajuda alimentar de emergência poderia financiar a estratégia de sustentabilidade alimentar da Cruz Vermelha queniana para a região de Turkana. Mercy demonstrou o impacto na saúde destas crianças por conta de atrasos e o contraste abismal com países cultivando comida em desertos. A jornalista foi convidada a apresentar suas descobertas na sede da USAID em Nairóbi e, em 2015, a estratégia quanto à agricultura e à segurança alimentar da agência mudou de ajuda humanitária para agricultura sustentável.[232]

Ela conseguiu o apoio do público ao documentar de maneira intimista a situação de famílias famintas em Turkana. Foram três dias com as famílias que participam da matéria, junto de tradutor e cinegrafista. O telefone da emissora não parou de tocar antes mesmo da reportagem terminar de ser exibida, com quenianos buscando formas de doar dinheiro para as pessoas apresentadas ali. Por conta da reação gigantesca de indivíduos e organizações à notícia veiculada, em algumas horas a emissora havia criado um fundo de auxílio para o Condado de Turkana. Estas e outras matérias sobre a desesperadora situação da fome na porção norte do Quênia levaram à cobertura diária nos meios de comunicação do país, historicamente desinteressados no sofrimento de regiões pobres e isoladas naquela região. Seu público se relacionou com uma história humana, forte, e não com os dados que sugeriam que as doações poderiam ser melhor aplicadas em desenvolvimento.

---

[231] Versão impressa: http://www.internewskenya.org/summaries/internews52e7747b74fff.pdf. Versão resumida online: https://www.nation.co.ke/lifestyle/dn2/When-will-Kenya-have-enough-to-feed-all-its-citizens-/957860-2163092-p8mgj9z/index.html.

[232] https://www.usaid.gov/kenya/agriculture-and-food-security.

Por fim, o governo cedeu à pressão pública e de doadores. O Comitê de Monitoramento de Secas pediu a Juma que compartilhasse os dados de sua matéria, pois afirmou não estar ciente de que a situação era tão desesperadora, por mais que este mesmo departamento tenha tentado lhe cobrar por acesso a estas informações quando ela deu início à sua apuração. Com base nas informações de falta de água de Juma, o Ministério da Água planeja ir até Turkana para fazer mais poços artesianos. Além disso, o governo, através do Ministério de Planejamento e Devolução, separou 27 milhões de dólares para distribuição de auxílio em Turkana, um desdobramento que a jornalista acompanhou de perto. Por conta da reação absurda à matéria, tanto por parte de pessoas quanto organizações, legislação ligada à sustentabilidade alimentar que redireciona auxílios finalmente foi levada à discussão no Senado em maio daquele ano.[233] Juma seguiu produzindo matérias baseadas em dados sobre a desconexão entre percepção pública, programas de doação e políticas, incluindo a matéria "Mães Adolescentes em Kwale", investigação sobre o impacto do uso de contraceptivos em índices de gravidez na adolescência em uma parte conservadora do país.[234]

2. Como garantir que nossa justiça está protegendo os marginalizados?

No Afeganistão, a equipe do *Pajhwok Afghan News* usou dados para avaliar o impacto de duas políticas alardeadas como fundamentais para o progresso da justiça no país — a Lei de Eliminação da Violência Contra a Mulher, de 2009, e a Estratégia de Controle Nacional de Drogas (2012-2016) — e acabou se deparando com duas vítimas inesperadas destas, mulheres sofredoras de abusos e trabalhadores rurais. Por mais que no Afeganistão não exista legislação para acesso à informação, muitas agências que recebem doações, incluindo os ministérios de questões da mulher e de combate aos narcóticos, são obrigadas, por força de contrato, a fornecerem dados.

Cinco anos após a implementação da lei contra violência doméstica, o *Pajhwok Afghan* quis saber o que tinha acontecido com abusadores e abusadas. A equipe obteve dados de 21.000 casos de abusos junto ao Ministério de Questões da Mulher e diversas agências das Nações Unidas incumbidas de monitorarem tais casos, desde sua denúncia até a fase de veredicto final e mediação. Descobriram que no pior país do mundo para mulheres, a lei amplamente elogiada as havia inserido em um processo de mediação local enraizado no machismo tradicional que as colocava junto de seu abusador.[235] Dois anos depois, a Human Rights Watch publicou estudo que confirmava as descobertas da *Pajhwok Afghan News*:

---

[233] http://kenyalaw.org/kl/fileadmin/pdfdownloads/bills/2014/TheFoodSecurityBill2014.pdf.

[234] NTV Kenya, #*TeenMumsOfKwale*, outubro de 2016.

[235] https://www.pajhwok.com/en/2016/05/11/cases-violence-against-women-mediation-best-option.

legislação e mediação haviam fracassado com as mulheres afegãs.[236] Mesmo se mais mulheres tivessem acesso ao judiciário, que conta com altos índices de condenação para abusadores, fica no ar a questão espinhosa de como lidar com divorciadas em uma sociedade onde mulheres não trabalham.

Desafios práticos semelhantes surgem na implementação de uma estratégia de combate às drogas. O Escritório das Nações Unidas sobre Drogas e Crime teve acesso, o que é raro, aos condenados por crimes relacionados a drogas e cedeu os dados crus da pesquisa para a equipe da *Pajhwok*. Uma análise das descobertas destas pesquisas revelou que a nova legislação acabou por levar mais motoristas e agricultores pobres e iletrados à prisão, enquanto chefões de cartel seguiram em liberdade.[237] Cabe notar que a maioria dos prisioneiros planejava voltar a operar no ramo das drogas após soltura, sendo este o único meio de sustentar suas famílias em áreas rurais isoladas.

Tais histórias atenderam a um propósito triplo para a equipe de dados da *Pajhwok*: checar na prática como funcionam as políticas desenvolvidas a partir de uma visão legal ocidental; destacar as consequências da marginalização econômica por gênero e localização; e entregar conteúdo de interesse público, baseado em dados, em idiomas como dari, pashtu e inglês, a um público variado.

3. Como garantir educação de qualidade para todos?

Acesso à educação, muitas vezes tido com um grande nivelador, permite que comunidades marginalizadas quantifiquem o fracasso de um governo em fornecer serviços públicos básicos e direcionem líderes locais na direção de algum tipo de reforma. Em uma série de artigos, o desenvolvedor e jornalista Abdul Salam Afridi construiu uma narrativa em torno do acesso à educação entre os menos favorecidos, o que o colou entre os indicados do Data Journalism Awards. No primeiro artigo, usou estatísticas oficiais do governo e dados nacionais de pesquisas sobre educação para mostrar que pais na remota região tribal de Passo de Khyber que, em um ato de desespero, enviavam cada vez mais crianças para escolas particulares estavam, na verdade, fazendo um péssimo investimento. O artigo de Abdul mostrava que a maioria dos formandos de escolas públicas e privadas fracassava em testes padronizados básicos.[238] Mais artigos sobre educação pública no Território Federal das Áreas

---

236 https://unama.unmissions.org/sites/default/files/unama_ohchr_evaw_report_2018_injustice_and_impunity_29_may_2018.pdf.

237 https://www.pajhwok.com/en/2016/09/28/most-jailed-drug-offenders-are-poor-illiterate.

238 http://www.newslens.pk/in-kp-parents-still-prefer-private-over-public-schools/.

Tribais, terra natal de Abdul, e no Passo de Khyber exploraram os motivos por trás deste fracasso.[239]

Outro artigo — sobre alunos escalados para participar do programa de treinamento vocacional e de vagas de emprego a nível nacional — mostrava uma enorme lacuna entre as habilidades destes estudantes e as demandas do mercado. A investigação revelou que o país treina especialistas em TI e esteticistas, quando precisa de motoristas e metalúrgicos, deixando metade dos formandos desempregados, em grande parte por conta de quem maneja este projeto. Financiado pelo fundo de desenvolvimento do governo alemão, GiZ, o governo paquistanês fez sua própria análise do tema e chegou à mesma conclusão, agindo rapidamente para alterar o programa e oferecer novos cursos, alinhados com as habilidades exigidas pelas vagas.[240]

Uma vantagem inata à reportagem baseada em dados em meio a comunidades marginalizadas é que o jornalista pode seguir trabalhando na pauta mesmo após o choque inicial do escândalo ter se esvaído. O que histórias como essas têm em comum é que usam dados não apenas para relatar um problema, mas também na busca de soluções para ele. Estes jornalistas reúnem dados para mensurar o problema, seu impacto, causas e soluções. Em termos globais, há uma pressão para que se faça jornalismo de dados de forma acessível por, sobre e para comunidades marginalizadas, de forma a ganhar sua confiança e colocá-las para atuar na vida cívica.

Jornalismo de dados, mas com restrições

Muito da divisão na academia a respeito da viabilidade de longo prazo do jornalismo de dados deve-se à seguinte questão: seu objetivo é produzir produtos interativos de alto nível ou reportagens de interesse público baseadas em fatos? Jornalistas de países em desenvolvimento utilizam dados para responderem a perguntas básicas sobre discriminação de gênero institucionalizada, sistemas judiciários preconceituosos e negligência proposital para com a fome, entregando essas informações para o máximo de pessoas possível. Fazem isso sabendo que se trata de questões complicadas e que qualquer mudança prática é difícil de ser obtida. Já os jornalistas de dados do Ocidente, com acesso a recursos melhores, dados e uma mídia livre, além de um governo mais responsivo, muitas vezes não aproveitam as oportunidades oferecidas para garantir que, em tempos tão turbulentos, estejamos tratando das necessidades de informação de cidadãos marginalizados ou responsabilizando governos por suas ações.

---

[239] Afridi (2017 e 2018).

[240] Salam (2017).

A maioria destes problemas era invisível antes e se farão invisíveis de novo se os jornalistas pararem de falar sobre eles. Em seus melhores momentos, o jornalismo de dados é feito por, sobre e para aqueles que a sociedade decidiu ignorar. Com sorte, a sociedade civil, ativistas, acadêmicos, governos e outros estão trabalhando juntos para fazer um trabalho melhor em incluir aqueles que sempre foram deixados de lado. No meio disso tudo, jornalistas têm um papel essencial, o de garantir que as pessoas discutam e trabalhem para solucionar estes problemas. Tudo era terrível, é terrível e será terrível a não ser que continuemos falando sobre o assunto. Ano após ano precisamos contabilizar os famintos, os abusados, os presos, os não educados e os não ouvidos porque em cada canto do planeta a situação está horrível para alguém.

Eva Constantaras é jornalista investigava de dados especializada na implementação de unidades dedicadas de dados em grandes veículos de países em desenvolvimento para o atendimento de comunidades locais.

Referências

AFRIDI, Abdul Salam. In: KP. *Parents Still Prefer Private Over Public Schools*. News Lens, 18 de fevereiro de 2017.

AFRIDI, Abdul Salam, *Half of FATA Schools Functioning in Dire Straits*. News Lens, 16 de junho de 2017.

AFRIDI, Abdul Salam Afridi. *Despite Huge Investment The Outlook of Education in KP Remains Questionable*. News Lens, 2 de março de 2018.

AFRIDI, Abdul Salam. *TVET Reform Programmes Targeting Wrong Skills*. Data Journalism Pakistan, 16 de setembro de 2017.

JUMA, Mercy. *When will Kenya have enough to feed all its citizens?* Daily Nation, 28 de janeiro de 2014.

BARAKZAI, Navid Ahmad; WARDAK, Ahsanullah. *Most Jailed Drug Offenders Are Poor, Illiterate*. Pajhwok Afghan News, 28 de setembro de 2016.

NTV KENYA, *#TeenMumsOfKwale*. YouTube, 2 de outubro de 2016.

MUNSEF, Abdul Qadir; SALEHAI, Zarghona. *Cases of Violence Against Women*. Pajhwok Afghan News, 11 de maio de 2016.

United Nations Assistance Mission in Afghanistan. *Injustice and Impunity: Mediation of Criminal Offences of Violence Against Women*. Maio de 2018.

Ensino de jornalismo de dados[241]

Cheryl Phillips

Cindy Royal dá aulas de desenvolvimento para a web na Universidade Estadual do Texas. A alguns milhares de quilômetros ao leste, na Universidade da Flórida, Mindy McAdams ocupa a Cátedra Knight de Tecnologias de Jornalismo e Processo Democrático, já o professor adjunto Norman Lewis ensina de programação a jornalismo de dados tradicional e desenvolvimento de apps. Alberto Cairo, que ocupa a Cátedra Knight de Jornalismo Visual na Escola de Comunicação da Universidade de Miami, leciona um curso inteiro voltado a visualizações de dados.

Siga ao norte e estudantes da Universidade de Columbia e da Universidade da Cidade de Nova York (CUNY, na sigla em inglês) têm aulas com jornalistas de dados em atividade que trabalham em veículos como *NBC* e *New York Times*, onde aprendem o básico de jornalismo investigativo e análise de dados. Na Universidade de Maryland, aulas de lei de imprensa regularmente abordam o processo de solicitar informações públicas para projetos jornalísticos. Em Nebraska, Matt Waite ensina alunos a visualizarem dados com uso de lego. Na Universidade de Stanford, ensinamos análise de dados básica, programação em Python e R e visualização de dados básica, mais para fins de compreensão geral do que apresentação.

Professores de jornalismo de dados — muitos tendo começado como jornalistas — lecionam das mais variadas formas pelo mundo (e os exemplos citados acima contemplam apenas os Estados Unidos). Qual dos cursos mencionados trata do verdadeiro jornalismo de dados? Pegadinha: todos. Sendo assim, como lecionar?

Da mesma forma como lecionamos qualquer tipo de aula em jornalismo. Qualquer especialização — do jornalismo esportivo ao de negócios ou científico — tem habilidades e conhecimentos específicos do campo a ser coberto que precisam ser aprendidas. Mas todas dependem de princípios jornalísticos básicos.

O mesmo vale para jornalismo de dados: devemos começar pelos fundamentos. E quando digo fundamentos, não estou me referindo a planilhas, por mais que considere este aprendizado ideal para compreender alguns dos princípios essenciais do jornalismo de dados. Não há nada como compreender a bagunça inerente aos dados ao fazer com que seus alunos embarquem num exercício em sala de aula envolvendo inserir informações em caixinhas na tela de um computador. Também não me refiro a um tipo específico de linguagem de

[241] Créditos a Charles Berrett, coautor de *Teaching Data and Computational Journalism*, publicado com apoio da Universidade de Columbia e da Fundação John S. e James L. Knight.

programação, seja Python ou R, por mais que acredite que ambas têm muitos benefícios. Não há nada como ver um aluno às voltas com uma única linha de código que apresenta resultados que precisariam de quatro ou mais passos em uma planilha.

O aprendizado em jornalismo de dados começa com o entendimento de como pensar de maneira crítica sobre informações e como estas podem ser coletadas, padronizadas e analisadas para fins jornalísticos. Começa com entender a história a ser contada e fazer as perguntas certas para chegar lá.

Tratando-se dos educadores em jornalismo, é bem provável que já saibam as formas que estas perguntas podem assumir.

- **Quem** criou os dados?
- **O que** deve constar nos dados?
- **Quando** os dados foram atualizados pela última vez?
- **Que lugar** do mundo estes dados representam?
- **Por que** precisamos destes dados para contar nossa história?
- **Como** encontrar as respostas às questões que queremos fazer a estes dados?

Então, vá em frente, monte o currículo com planilhas, SQL, Python, R, o que for. Não importa. Assim como não importa o fato de que em algum ponto da vida eu soube usar um programa de DOS chamado Paradox. O que importa mesmo é saber quais passos seguir na hora de coletar e analisar dados. Visualizações são essenciais tanto para análise quanto apresentação, mas se a análise visual para compreensão precede a apresentação, esta última é facilitada.

Este capítulo conta com algumas abordagens diferentes e pontos iniciais, assim como dicas de como ensinar a fazer jornalismo de dados com base em quem você é, o nível do programa curricular em mãos e como criar projetos colaborativos. Após introduzir o método "mala" ao ensino de jornalismo de dados, temos modelos de curso único, de sala de aula invertida e integrados, além de experimentos em coensino em diversas disciplinas.

Um único curso: preparando a mala

Quando vamos acampar, brincamos que levamos de tudo, até a pia da cozinha.

O segredo é saber o que pode ser levado e o que vai ser só peso extra ao ponto da improdutividade. Aquela pia da cozinha, na verdade, é uma tigelinha dobrável de tecido.

Se você está dando uma só aula e é o único educador em jornalismo de dados, não tente enfiar tudo em um curso só, não é preciso incluir análise de dados com planilhas e Linguagem de Consulta Estruturada, processamento de dados com Python, análise com R e visualizações geradas com D3 em um único trimestre ou semestre.

Escolha as ferramentas essenciais. Considere basear ao menos parte das aulas em um projeto. De qualquer forma, siga os passos. Repita-os e mantenha tudo simples. Continue focado no jornalismo produzido a partir das ferramentas escolhidas.

Em 2014 e 2015, Charles Beret, da Universidade de Columbia, e eu, conduzimos uma pesquisa e longas entrevistas com jornalistas de dados e educadores de jornalismo. A maioria daqueles que lecionam jornalismo de dados relatou que, ao começar com uma planilha que introduz o conceito de dados estruturados, os alunos tiveram maior facilidade em entendê-lo.

Outro passo é aumentar a complexidade ao incluir outras técnicas valorosas em jornalismo de dados, indo além de organizações e filtros, e partir para consultas do tipo "agrupar por" ou pela junção de conjuntos de dados díspares em busca de padrões até então não descobertos.

Mas isso não necessariamente significa envolver uma série de ferramentas ou mesmo usar uma entre as mais recentes. É possível apresentar o tema aos alunos usando qualquer tecnologia que funcione para você e o programa curricular de jornalismo de sua instituição. Se é parte do currículo de uma universidade onde estudantes podem usar o MS Access, use-o, mas vá além da interface simples para ter certeza de que todos os alunos entendam a linguagem de consulta estruturada (SQL) por trás de cada consulta feita. Ou use MySQL. Ou use Python dentro de um ambiente Jupyter Notebook. Ou use R e R Studio, que conta com pacotes excelentes para consultas do tipo SQL.

O objetivo maior é fazer jornalismo, entender o que precisa acontecer e que há muitas formas de operação semelhante com dados a serviço de narrativas.

Mais uma vez, mantenha a simplicidade. Não crie obstáculos para os seus alunos no tocante a ferramentas de tecnologia, que devem ser usadas para fazer do jornalismo mais robusto e facilitá-lo. Voltando à analogia do acampamento: leve só o que você precisa para dentro da sala de aula. Não traga uma serra elétrica se tudo que você precisa é uma machadinha ou um canivete.

Mas, assim que você tiver estabelecido como será aquela aula, pense além desse modelo. Pense em formas de desenvolver componentes de jornalismo de dados no departamento ou instituição. Encontre motivações compartilhadas com outras disciplinas. Pergunte-se se há a possibilidade de trabalhar com outros colegas que estão dando uma aula

de reportagem básica, por exemplo, verifique se há interesse de seus alunos aprenderem um pouco mais sobre como integrar dados.

Alguns professores vêm experimentando o formato de "sala invertida" de forma a equilibrar aprendizagem, pensamento crítico e reflexão teórica. Estudantes podem seguir guias de acordo com seu próprio ritmo e focar em solução de problemas junto a instrutores durante a aula e através da aprendizagem de outros métodos para abordagem de desafios variados em jornalismo de dados. A professora McAdams, da Universidade da Flórida, emprega este formato em sua aula de design de apps para web.

Um dos benefícios deste modelo é que ele leva em consideração jornalistas dos mais variados níveis de habilidade. Em algumas ocasiões, uma aula de jornalismo pode atrair o interesse de um estudante que é bom em ciência da computação e, ao mesmo tempo, de outro aluno que nunca usou uma planilha.

Mas o ensino de jornalismo de dados de dados vai além do modelo de sala invertida. Ensinar jornalismo de dados significa pensar em outras formas de passar seus conceitos adiante.

Durante a SRCCON, desconferência realizada regularmente, Sarah Cohen — professora da Cátedra Knight de Jornalismo de Dados da Universidade Estadual do Arizona e vencedora recente do prêmio Pulitzer por seu trabalho no *New York Times* — defendeu o uso de outras atividades análogas para engajar estudantes.

Cohen e Waite, professor de prática profissional da Universidade de Nebraska, introduziram a ideia de um currículo comum, baseado em módulos, possível de ser adotado por educadores de qualquer lugar. O objetivo é desenvolver um sistema em que professores não precisem criar nada do zero. No evento realizado no verão de 2018, os dois lideraram um grupo de participantes na contribuição de possíveis módulos para esta empreitada.

"Estamos tentando não ser religiosos com essas coisas (ferramentas)", disse Cohen ao grupo: "Estamos tentando, sim, catequizar as pessoas em relação aos valores fundamentais do jornalismo e da análise de dados".

Isso acabou levando à criação de um repositório no GitHub em que colaboradores adicionam e ajustam módulos para uso no ensino de jornalismo de dados.[242] O repositório conta, ainda, com links para outros recursos voltados ao tema, incluindo este manual.

---

[242] https://github.com/datajtext/DataJournalismTextbook.

Interpretação de enquetes e estudos são algumas possibilidades de módulos. Numeramento básico é um componente importante de cursos de jornalismo. Encontrar dados online também é sucesso na hora de animar as aulas.

Além do que, significa que você não precisa doar o seu tempo livre inteiro à causa. Desenvolva um módulo ou tutorial uma única vez e poderá ser usado por terceiros, múltiplas vezes. Ou valha-se dos muitos tutoriais grátis disponíveis por aí. As edições anuais das conferências promovidas por organizações como *Investigative Reporters* e *Editors* e do Instituto Nacional de Comunicação Assistida por Computador (NICAR, na sigla em inglês) rendem ainda mais guias e tutoriais para seus associados, cobrindo temas que vão de tabelas dinâmicas a raspagem e mapeamento.

Uma vez por trimestre atuo como professora convidada junto a um colega, lecionando sobre busca de dados online. Dentre os benefícios, cria-se um grupo de estudantes interessados na exploração do jornalismo de dados e estar integrado a uma atmosfera universitária com outros docentes.

Se possível, considere criar módulos que estes colegas possam adotar. Jornalistas ambientais, por exemplo, poderiam fazer um módulo sobre temperaturas médias ao longo do tempo utilizando planilhas.

Há outros benefícios em potencial: você mostra aos colegas o valor do jornalismo de dados, que pode ajudar a justificar um programa curricular que integre, sistematicamente, tais práticas e abordagens.

Mais pensamentos sobre modelos integrados ou lecionar sem fronteiras

Um modelo plenamente integrado significa que há mais de uma pessoa dedicada ao ensino de conceitos de jornalismo de dados — o que, por sua vez, pode indicar o potencial para ir além dos limites do programa curricular de um curso de jornalismo. Em Stanford, lançamos o Projeto de Policiamento Aberto de Stanford e firmamos uma parceria com a Poynter para treinar jornalistas na análise de dados de policiamento. Professores dos departamentos de jornalismo e engenharia trabalharam juntos em aulas que rompem limites, acolhendo alunos de jornalismo, direito e ciência da computação. Isso é importante, porque as melhores equipes colaborativas de redações contam com gente de diversas áreas. Recentemente, instituições acadêmicas não só passaram a adotar estes modelos integrados, como também produziram trabalhos que chegam às redações e ensinam seus alunos, tudo ao mesmo tempo.

Neste mês, a Fundação Scripps-Howard anunciou bolsas de 3 milhões de dólares para a Universidade Estadual do Arizona e Universidade de Maryland, que lançarão seus próprios

centros de jornalismo investigativo.[243] Estes centros servirão tanto para a educação dos alunos quanto para a produção de produtos de jornalismo investigativo, assumindo dois papéis: edição e educação.

Aulas dotadas de uma missão e que vão além da sala de aula são mais atraentes para os alunos e podem levar a uma experiência de aprendizado mais engajadora. Uma das aulas mais bem-sucedidas das quais já participei foi a de Lei, Ordem e Algoritmos durante a primavera de 2018, dada por mim e pelo professor assistente de engenharia Sharad Goel. Quem batizou o curso foi Goel, mas demos uma mexida nisso aí. Minha aula de vigilância, de mesmo nome, acabou combinando com a dele. Considerando as duas turmas, acabamos lecionando para alunos de engenharia e ciência da computação, direito e jornalismo. As equipes de estudantes produziram análises estatísticas avançadas, informativos e jornalismo a partir de seus projetos. Goel e eu ensinamos cada um sobre sua área de especialidade. Gosto de pensar que aprendi algo sobre direito e como algoritmos podem ser usados para o bem e para o mal, e que Goel também aprendeu um pouco sobre o que é necessário para se fazer jornalismo investigativo e de dados.

Já os estudantes, considerando a natureza das aulas voltadas a projetos, puderam aprender o que era preciso para atingir os objetivos do projeto de suas equipes. O que evitamos fazer foi pedir que os alunos aprendessem tanto a respeito de ferramentas e técnicas, ao ponto de só perceberem progresso a nível incremental. Tentamos incluir ali o que era preciso para ser bem-sucedido, como naquelas viagens para acampar.

Cheryl Phillips é jornalista investigativa e de dados de longa data, leciona jornalismo de dados na Universidade de Stanford e é diretora da Big Local News, empreitada que visa coletar, fazer curadoria, arquivar e compartilhar dados locais para fins de jornalismo de transparência pública.

---

[243] Para mais informações sobre estas bolsas e o lançamento destes centros, ver Boehm (2018). Disponível em: https://amp.azcentral.com/amp/902340002.

Organização de projetos de dados com mulheres e minorias na América Latina

Eliana A. Vaca Muñoz

Chicas Poderosas (Garotas Poderosas) é uma rede transnacional de jornalistas, designers e desenvolvedoras que trabalha para desenvolver projetos de mídia digital por e para mulheres e comunidades marginalizadas da América Latina. Como designer integrante do grupo, meu trabalho explora o papel que o design pode assumir como agente de cultura e diversidade, através de pesquisa interdisciplinar e participativa na exploração de patrimônio cultural, identidade, apropriação de territórios e reconhecimento de mulheres e populações vulneráveis.

Este capítulo trata da organização de diversas iniciativas das *Chicas Poderosas* na Colômbia e na América Central. Ao passo que desigualdades sociais e culturais crescem cada vez mais no país sul-americano, é importante que minorias sejam ouvidas e tratadas como iguais e tenham seus conhecimentos compartilhados. Para tanto, a equipe *Chicas Colômbia* conduziu uma série de oficinas colaborativas voltadas ao jornalismo de dados e práticas de mídias digitais associadas. Nas seções a seguir discorro sobre os dois métodos que utilizamos para facilitar a participação nestas oficinas: coleta analógica de dados e visualização analógica de dados. Estas abordagens podem se mostrar relevantes à prática e à cultura de jornalismo de dados em comunidades e regiões em que conectividade, dispositivos e letramento tecnológico não são uma certeza.

Coleta de dados analógicos

Figura 1: “Sou tão criativa(o)?”, atividade de coleta analógica de dados.

Em maio de 2016, *Chicas Colômbia* foi até Villa España, em Quibdó, Chocó, para trabalhar com mulheres e adolescentes integrantes do coletivo AJODENIU (sigla para “Associação de Juventude Deslocada”). Desde 2002, este grupo atuava em prol dos interesses e direitos de crianças removidas de Chocó, Río Sucio, Bojayá e Urabá. Todas estas regiões são de difícil acesso, sem internet e com poucos serviços de apoio disponíveis. Logo, as oficinas começaram com o ensino de técnicas de coleta analógica de dados voltada a dados qualitativos. Com base nestas informações, trabalhamos na construção de narrativas a respeito de temas como deslocamento forçado e gravidez na adolescência, por meio de gravação e registro de histórias orais e escritas.[244]

Com base nestas abordagens, trabalhamos junto ao programa de Desenvolvimento das Nações Unidas em Honduras, no mês de setembro de 2018, no desenvolvimento de uma oficina com os Observatorios Municipales de Convivencia y Seguridad Ciudadana, que lidavam com dados a respeito de mortes violentas de homens e mulheres e buscavam como apresentar estas informações de forma desagregada por gênero. Um dos objetivos era criar caminhos emocionais para iniciar conversas em torno destes tópicos delicados com a

[244] https://chicaspoderosas.org/2016/11/22/the-pacific-counts-chicas-poderosas-quibdo-colombia/.

comunidade por meio de atividades participativas, e usar recursos limitados para o compartilhamento de dados relevantes e sensíveis.

Nestas oficinas, as atividades iniciais são para quebrar o gelo, com perguntas simples e divertidas (Figura 1). Durante sua realização em Honduras, houve dificuldades ao discutir violência com participantes por conta das diferentes normas sociais da região e barreiras de linguagem. Sendo assim, voltamos a oficina a diferentes atividades de coleta exploratória de dados de forma a trazer à tona diferentes concepções e experiências de violência. Usamos desenhos, imagens e fotos para criação de pôsteres em grupo, onde participantes podiam adicionar adesivos como forma de coletar dados — inclusive sobre a forma como se viam, sua compreensão de direitos e como haviam vivido diferentes tipos de violência doméstica (física, psicológica e econômica) em suas próprias vidas.

Visualização analógica de dados

Em um esforço para melhor entender as questões afetando estas comunidades indígenas, planejamos em 2017 oficinas interativas com a tribo Embera, da região de Vigía del Fuerte, de forma a ter um vislumbre de suas vidas, apesar da barreira idiomática. Historicamente, interações entre a tribo e forasteiros envolviam predominantemente homens, então priorizamos o acesso de populações femininas para medir seus níveis educacionais e facilitar discussões sobre empoderamento.

Figura 2: Exemplo de visualização analógica usando contas em que as diferentes cores representam as línguas faladas e a quantidade de contas representa a fluência em cada uma delas.

Na falta de tecnologias modernas, exploramos expressões tradicionais de cultura de forma a acessar de forma mais significativa as vidas de nossas participantes, através de práticas como tecelagem, decoração com contas e artesanato (Figura 2).

No mês de setembro de 2018, desenvolvemos uma oficina em Honduras sobre como "humanizar" dados, através da condução de projetos de resiliência junto a vítimas e populações em risco. Projetamos oficinas de visualização analógica de dados com técnicas de design empático utilizando tesouras, papel, figurinhas, notas adesivas e balões. Estas oficinas serviram tanto para facilitar o compartilhamento de informações sensíveis junto a organizações relevantes de forma a melhor atender estas comunidades, como para ensinar diferentes formas pelas quais estas populações vulneráveis e de baixo letramento poderiam compartilhar dados sobre suas vidas, experiências e problemas. Trabalhamos com participantes para criar visualizações analógicas sobre homicídios e feminicídios por região, tipo e idade, por exemplo.

Em outra oficina, realizada em Belize, exploramos diferentes abordagens colaborativas para a visualização de dados a respeito de crime e violência. Originalmente, queríamos ver como informações do Observatório da Violência de Belize poderiam ser usadas para coordenar diversos tipos de respostas coletivas. Por mais que os participantes tivessem alto letramento, os recursos tecnológicos e a conectividade eram muito precários, o que dificultava o uso de ferramentas básicas de visualização online. Tudo isso levantou questionamentos e trouxe à baila vários desafios sobre práticas de visualização de dados online, muitas vezes tomadas como certas, mas que não funcionariam nas circunstâncias que nos encontrávamos, o que sugere a relevância de abordagens analógicas à visualização com os materiais disponíveis.

Eliana A. Vaca Muñoz é designer voltada ao trabalho com minorias em estado de desfavorecimento e à colaboração em projetos envolvendo o empoderamento feminino através da visualização de dados sobre direitos humanos, dados qualitativos e coletados pela comunidade, e 'humanização' de conjuntos de dados quantitativos.

Situando o jornalismo de dados

Genealogias do jornalismo de dados

C.W. Anderson

Introdução

Por que alguém deveria se importar com a história do jornalismo de dados? Não só "história" é um tema abstrato e um tanto quanto acadêmico para a maior parte das pessoas, como também parece bem distante da realidade de jornalistas de dados em atuação, ocupados com seus próprios trabalhos, com prazos apertados e o objetivo de passar adiante informações complicadas de forma ágil e compreensível para o maior número de leitores possível. Não é de se estranhar que estes não gostem tanto da ideia de perder tempo refletindo sobre o ofício. Na maior parte do tempo, esta relutância em olhar para o próprio umbigo é uma qualidade admirável; já quando falamos de práticas e conceitos em jornalismo de dados e reportagem computacional, a hostilidade ao pensamento histórico pode ser um problema que atravanca a produção de jornalismo de qualidade.

O jornalismo de dados pode muito bem ser a mais poderosa forma de se fazer jornalismo coletivo no mundo hoje. Ou, no mínimo, trata-se da forma mais positiva e positivista de se fazer jornalismo. Este *poder* (a capacidade do jornalismo de dados de criar jornalismo de altíssima qualidade, junto da força retórica de seu modelo), *positivismo* (a maioria dos praticantes deste tipo de jornalismo tem esperanças para o futuro da área, convencida de que está em ascensão) e *positivismo* (estes jornalistas acreditam firmemente na pesquisa guiada em métodos para a captura de fatos reais e prováveis ao redor do mundo) criam o que eu chamaria de uma profissão empiricamente segura de si. Uma consequência desta segurança, poderia argumentar, é que ela pode gerar uma ideia otimista de que o jornalismo de dados está em um processo constante de melhoria e melhorando o mundo. Tal atitude pode, também, levar à arrogância e à falta de autorreflexão crítica, aproximando o ofício das instituições que critica e busca responsabilizar.

Neste capítulo, gostaria de argumentar que mais atenção à história pode, de fato, melhorar a rotina diária do jornalismo de dados. Ao entender que processos e práticas têm uma história, jornalistas podem abrir suas mentes ao fato de que as coisas no presente podem ser feitas de outro jeito, porque talvez já tenha sido assim em algum outro momento. Sendo mais específico, estes jornalistas podem pensar mais a respeito de como representar incerteza de maneira criativa em seu trabalho empírico. Podem, por exemplo, levar em conta técnicas para atrair leitores com diferentes sensibilidades políticas e persuasão que vai além de simples evidência factual. Podem, em suma, se abrir para o que as estudiosas e historiadoras de estudos em ciência e tecnologia Catherine D'Ignazio e Lauren Klein (2018) batizaram de "visualização feminista de dados", um formato que repensa binários, abraça o pluralismo,

examina relações de poder e considera diferentes contextos (ver o capítulo assinado por D'Ignazio neste livro). Para empreender tais mudanças, o jornalismo de dados, mais do que outras práticas jornalísticas, deve inculcar esta sensibilidade histórica por conta da natureza de suas próprias força e segurança em si. Nenhum tipo de história é melhor equipada para levar à autorreflexão do que a abordagem genealógica ao desenvolvimento conceitual, liderada por Michel Foucault e adotada por alguns historiadores nos campos da ciência e de estudos de ciência e tecnologia.

"Genealogia", como definida por Foucault e inspirada em trabalhos mais antigos de Nietzsche, é uma abordagem única ao estudo da evolução de instituições e conceitos no tempo, distinta da história. A análise genealógica não busca um único ponto de origem para práticas ou ideias no passado nem tenta entender como conceitos se desenvolveram seguindo uma linha evolucionária de ontem para hoje. Na verdade, foca-se mais em *descontinuidade* e *mudanças inesperadas* do que na presença do passado no presente. Como dito por Nietzsche em uma passagem de *A Genealogia da Moral*, citada por Michel Foucault:

> O "desenvolvimento" de algo, uma prática ou órgão nada tem em comum com seu progresso em direção a um único objetivo, menos ainda o progresso lógico e mais curto alcançado com o mínimo dispêndio de poder e recursos. Em vez disso, é a sequência de processos mais ou menos profundos, mais ou menos mutuamente independentes de dominação, que ocorrem naquela coisa, junto com a resistência que surge contra essa dominação a cada vez, as mudanças de forma que foram tentadas para o propósito de defesa e reação, e os resultados de contramedidas bem-sucedidas. A forma é fluida; o "significado", porém, é mais ainda (Foucault, 1980).

Uma "genealogia do jornalismo de dados", então, revelaria as formas pelas quais a prática evoluiu de maneiras que seus criadores e praticantes jamais esperaram, ou de maneiras contrárias aos seus desejos. Esta genealogia observaria as formas pelas quais a história nos surpreende e por vezes nos leva a direções inesperadas. Como argumentado anteriormente, uma abordagem como esta seria especialmente útil para os jornalistas de dados de hoje, ajudando-os a entender processos — creio que estes jornalistas *não* estejam trabalhando dentro de uma tradição predefinida e com um passado venerável; muito pelo contrário, estão improvisando tudo pelo meio do caminho de maneiras radicalmente contingentes. Levaria, também, a um tipo útil de autorreflexão crítica, que poderia ajudar a lidar com a autoconfiança (compreensível e muitas vezes merecida) de jornalistas e repórteres de dados em atividade.

Tentei traçar uma genealogia destas em meu livro *Apostles of Certainty: Data Journalism and the Politics of Doubt*. Nas páginas a seguir, tento resumir algumas das suas

descobertas e discutir maneiras pelas quais estas lições podem ser úteis para os dias atuais. Gostaria de concluir ao argumentar que o jornalismo, especialmente o dataficado, pode e deve fazer um trabalho melhor em mostrar aquilo que não sabe, e que tais gestos de incerteza honrariam as origens do jornalismo de dados na crítica de poderes legítimos, em vez de retificá-lo.

Jornalismo de dados através do tempo: anos 1910, 1960 e 2000

Jornalistas podem usar dados com outras formas de informação quantificada — como documentos impressos com números, visualizações de dados, tabelas e gráficos — para produzir um jornalismo melhor? E como esse jornalismo pode ajudar o público a tomar decisões políticas melhores? Estas foram as principais perguntas que guiaram *Apostles of Certainty: Data Journalism and the Politics of Doubt*, que tentava oferecer um panorama mais amplo da história das notícias. Com paradas ao longo dos anos 1910, anos 1960 e o presente, o livro traça genealogias do jornalismo de dados e seus suportes materiais e tecnológicos, argumentando que o uso de dados na notícia está inevitavelmente ligado a políticas nacionais, evolução de bancos de dados computáveis e história de campos científicos profissionais. É impossível entender os usos jornalísticos de dados, digo no livro, sem compreender as relações muitas vezes controversas entre ciências sociais e jornalismo. Também é impossível se desvencilhar de formas empíricas de levar a verdade ao público sem entender a notavelmente persistente crença progressiva de que a publicação de informações empiricamente verificáveis levará a um mundo mais justo e próspero. *Apostles of Certainty* conclui que esta intersecção de tecnologia e profissionalismo levou a um jornalismo melhor, mas não necessariamente a políticas melhores. Para atender às demandas da era digital por completo, o jornalismo precisa se sentir mais à vontade ao expressar a dúvida empírica tanto quanto a certeza. Ironicamente, este "abraçar da dúvida" poderia levar o jornalismo a ser encarado mais como uma ciência, e não menos.

O desafio das ciências sociais

A narrativa de *Apostles of Certainty* alicerça-se em três períodos de tempo diferentes nos EUA, cada um com suas perspectivas em relação ao desenvolvimento do jornalismo de dados. A primeira era, a chamada "Era Progressiva", período de ascensão política liberal acompanhada da crença de que o estado e os cidadãos comuns, informados pelas melhores estatísticas disponíveis, poderiam fazer do mundo um lugar mais humano e justo. O segundo momento ocorre nos anos 1950 e 1960, quando alguns reformistas do jornalismo começam a prestar atenção em ciências sociais quantitativas, especialmente sociologia e ciência política,

como possíveis fontes de novas ideias e métodos para fazer do jornalismo mais empírico e objetivo. Para ajudar nesta jornada, surgiu uma nova série de bancos de dados mais acessíveis e computadores poderosos. O terceiro momento tem início nos primeiros anos da década de 2010, quando a inovação do jornalismo de dados foi suplementada pelo jornalismo "computacional" ou "estruturado". Na atual situação da big data e da "aprendizagem profunda de máquina", estes jornalistas afirmam que a objetividade jornalística depende menos de referências externas, vindo de dentro da estrutura do próprio banco de dados.

Em cada um destes períodos, o jornalismo baseado em dados *reagia,* mas também se definia *em oposição parcial* a correntes maiores em operação dentro das ciências sociais em geral, e esta relação com correntes políticas e sociais maiores ajudou a informar as escolhas dos casos que destaquei neste capítulo. Em outras palavras, busquei pontos de inflexão na história do jornalismo que poderiam ajudar a lançar uma luz sobre estruturas políticas e sociais mais amplas, além do próprio jornalismo. Na Era Progressiva,[245] a reportagem tradicional, em grande parte, rejeitava a atenção emergente da sociologia em estruturas sociais e informação contextual despersonalizada, preferindo manter seu foco individualista em personalidades poderosas e eventos relevantes. Com a profissionalização do jornalismo e da sociologia, os dois ficaram mais à vontade em fazer reivindicações estruturais, mas foi só nos anos 1960 que Philip Meyer e outros reformistas se reuniram em torno da filosofia do Jornalismo de Precisão e começaram a tomar a sociologia quantitativa e a ciência política como modelos para o próximo nível de exatidão e contexto ao qual o jornalismo aspirava. Na virada do século XXI, um modelo amplamente normalizado de jornalismo de dados começou a lidar com questões de replicabilidade e causalidade que cada vez mais assolavam as ciências sociais. Como ocorrido com estas, o jornalismo passou a conduzir experimentos para determinar se "big data" e formas não causais de behaviorismo correlacional poderiam oferecer percepções sobre atividades sociais.

Apostles of Certainty, sendo assim, argumenta implicitamente que formas de conhecimento especializado jornalístico e autoridade nunca surgem isoladas ou apenas internamente, no contexto do jornalismo. O jornalismo de dados não se tornou o jornalismo de dados por razões puramente profissionais, sendo um processo que não pode ser analisado inteiramente ao analisar o próprio discurso jornalístico. De fato, o tipo de conhecimento especializado que nos anos 1960 passou a ser chamado de jornalismo de dados só pode ser entendido *relacionalmente*, ao examinar como jornalistas de dados reagiam e interagiam com seus colegas de ciências sociais, mais poderosos e com maior autoridade. Além do que, este processo não pode ser compreendido somente em termos de ações e embates humanos, em isolamento ou grupos. Conhecimento especializado, de acordo com o modelo que apresentei

---

[245] Nos Estados Unidos, o período conhecido como Era Progressiva foi de 1880 até 1920, considerado o tempo de grandes reformas liberais e uma tentativa de alinhar políticas públicas com a era industrial.

em *Apostles of Certainty,* é um fenômeno de rede em que agrupamentos profissionais sofrem para estabelecer jurisdição sobre ampla gama de artefatos materiais *e* discursivos. Em linhas simples, o jornalismo de dados seria impossível sem a existência do banco de dados, mas o banco de dados mediado por um entendimento profissional específico do que é um banco de dados e como este poderia ser implementado de formas apropriadamente jornalísticas (para uma tentativa mais abrangente ligada a este argumento sobre a natureza interligada do conhecimento especializado, ver Anderson, 2013). É impossível entender autoridade jornalística sem entender a autoridade das ciências sociais (e o mesmo pode ser dito a respeito de ciência de computação, antropologia ou narrativas longas de não ficção). Profissionalismo e conhecimento jornalístico não podem ser entendidos ao se observar somente o campo jornalístico.

A persistência da política

O jornalismo de dados deve ser compreendido de maneira genealógica em relação a campos especializados adjacentes, como sociologia e ciência política. Todas estas áreas, por sua vez, devem ser analisadas através de concepções mais abrangentes de política e como estas lidam com o fato de que os "fatos" que revelam são "políticos", quer gostem disso ou não. O próprio desejo por conhecimento factual é, em si, um ato político. Ao longo da história do jornalismo de dados, como digo em *Apostles of Certainty*, testemunhamos uma tentativa distinta de se escorar na neutralidade das ciências sociais para levar adiante o que só posso descrever como objetivos políticos progressistas. O contexto mais amplo em que esta ligação se dá, porém, mudou drasticamente com o tempo. Estas mudanças maiores devem contrabalançar qualquer entusiasmo ligado à ideia de que o que estamos vendo acontecer no jornalismo é o desdobramento teleológico da certeza jornalística possibilitada por dispositivos digitais cada vez mais sofisticados.

Na Era Progressiva, protojornalistas de dados viram a coleta e o acúmulo de fatos quantitativos como um processo de iluminação social e política, e, independentemente disso, livre de quaisquer outros comprometimentos políticos. Ao coletar fatos granulares sobre os índices sanitários da cidade, distribuição de pobreza através de espaços urbanos, estatísticas sobre presença na igreja e prática religiosa, condições de trabalho e uma série de outros dados fatuais — e ao transmitir estes fatos para o público através da imprensa —, pesquisadores sociais acreditaram que o organismo social ganharia uma compreensão mais robusta de suas condições. Ao obter melhor compreensão de si mesma, a sociedade melhoraria, por conta própria e por fazer com que políticos levassem adiante medidas reformistas. Neste caso, o conhecimento factual sobre o mundo fala por si só. Precisava apenas ser coletado, representado e divulgado, de tal forma que a iluminação seguiria. Podemos chamá-la de uma noção "ingênua e transparente" do que são fatos — não exigem interpretação e seu acúmulo

levará a mudanças sociais positivas. O jornalismo de dados, neste momento, pode ser político sem explicitar sua política.

Nos tempos de Philip Meyer, nos anos 1960, esta congruência entre fatos transparentes e política havia sido arruinada. Meyer e seus correligionários diziam que o jornalismo estava corrompido durante as décadas de 1950 e 1960 porque havia confundido objetividade com o simples registro do que todos os lados envolvidos em uma disputa política consideravam ser verdade, permitindo ao leitor que tirasse sua própria conclusão a respeito do que é verdade ou não. Em tempos de agitação social e desordem política, a objetividade jornalística precisava de um fundamento mais robusto, encontrado no campo das ciências sociais objetivas. O ponto de partida para reportagem de um tema não deve ser as declarações discursivas de políticos voltados a seus interesses e, sim, a verdade, dura e fria, advinda da análise de dados relevantes acompanhada da aplicação de métodos apropriados. Tal análise seria *profissional* e *não política*. Ao agir como um grupo altamente profissionalizado em busca da verdade, jornalistas poderiam ir além do viés político e ajudar a colocar o público no campo da verdade objetiva. As direções às quais esta verdade pode levar, porém, pouco importavam. Diferentemente da geração anterior de jornalistas de dados, alegre e ingenuamente progressiva, as consequências iluminadas destes dados não eram uma conclusão deixada de lado.

Hoje, diria que uma nova geração de jornalistas computacionais reabsorveu, inconscientemente, parte das crenças políticas e epistemológicas de seus antecessores da Era Progressiva. Em termos epistemológicos, há uma crença cada vez mais difundida entre jornalistas computacionais de que fatos digitais "falam por si mesmos" ou, ao menos, farão isso quando forem coletados, organizados e limpos. Em escala, quando ligados a bancos de dados semânticos maiores e consistentes internamente, fatos geram um tipo de *excesso correlacional* em que problemas com significado ou causalidade são apagados por uma enxurrada de dados computacionais. Em termos profissionais, jornalistas de dados entendem objetividade como emergente da estrutura de um banco de dados e não como parte de um processo ocupacional interpretativo mais amplo. E, por fim, em termos políticos, argumentaria que houve uma espécie de retorno ao "cripto-progressismo" por parte de muitos dos jornalistas de dados distintamente neutros, com uma esperança política enraizada de que mais e mais dados, belamente representados e divulgados por uma imprensa robusta, poderão irromper as tendências políticas mais irracionais ou patológicas presentes em democracias ocidentais. Esta era a esperança no ar, ao menos, antes de 2016 e dos choques duplos que foram o Brexit e a eleição de Donald Trump.

Certeza e dúvida

O desenvolvimento do jornalismo de dados nos EUA através do arco mais longo do século XX deve ser encarado como um em que alegações cada vez mais precisas em relação à certeza profissional jornalística coexistem desconfortavelmente com uma conscientização nascente de que todos os fatos, independentemente de suas origens, estavam encardidos de política. Estas crenças, muitas vezes contraditórias, estão evidentes nos mais variados campos ligados a dados, claro, não só no jornalismo. Em um artigo de 2017 do *The Atlantic,* por exemplo*,* o colunista de ciências Ed Yong tratou de como um movimento rumo a uma "ciência aberta" e à crescente crise de replicabilidade poderia ser usado por membros anticientíficos do congresso para difamar e retirar o financiamento de pesquisas científicas. Yong citava Christie Aschwanden, jornalista de ciências do *FiveThirtyEight*: "Parece que o público pensa duas coisas opostas a respeito da ciência". "[Ou é] uma varinha mágica que transforma tudo que toca em verdade, ou é tudo bobagem porque o que pensávamos mudou... A verdade está no meio do caminho. Ciência é um processo de redução de incertezas. Se você não mostra incerteza como parte do processo, permite que outras pessoas peguem incertezas genuínas e usem isso para atrapalhar as coisas", comentou a Yong (2017). Este raciocínio alinha-se com a obra da estudiosa da STS Helga Nowotny, em *The Cunning of Uncertainty*, no qual declara que "a relação entre superar a incerteza e buscar a certeza sustenta o desejo de saber" (Nowotny, 2016). A essência da ciência moderna, ao menos em sua forma ideal, não chega à certeza e sim ao fato de que esta declara tão abertamente a provisionalidade de seu conhecimento. Falando de ciência, nada está escrito em pedra. Ela admite saber pouco, muitas vezes. E é através deste, um dos mais modernos paradoxos, que a ciência se faz digna da confiança pública.

Uma das percepções oferecidas por esta visão geral genealógica do desenvolvimento e da implementação do jornalismo de dados, diria, é que jornalistas que se guiam por estes dados ficaram obcecados com exatidão e certeza, deixando de lado uma compreensão mais simples de provisionalidade e dúvida. Como tentei demonstrar, desde meados do século XX, jornalistas têm empreendido esforços cada vez mais bem-sucedidos para que suas alegações sejam mais acertadas, contextuais e explicativas. Em grande parte, isso se deu pela utilização de diferentes tipos de evidência por parte dos jornalistas, especialmente evidências quantitativas. De qualquer forma, deve ficar claro que este profissionalismo elevado e a confiança cada vez maior dos jornalistas em sua capacidade de fazer alegações contextualizadas nem sempre levaram aos resultados democráticos esperados por estes profissionais. O discurso político norte-americano moderno tenta lidar com a incerteza da modernidade ao engajar com uma série de alegações cada vez mais estridentes de certeza. Em vez de solucionar este dilema, o jornalismo profissional acabou por exacerbá-lo. Para melhor atuar em meio à complexidade do mundo moderno, concluiria que o jornalismo deve

repensar os meios e mecanismos pelos quais transmite suas provisionalidade e incerteza. Feito corretamente, isso poderia aproximar o jornalismo da ciência moderna, em vez de afastar.

C. W. Anderson é autor de Rebuilding the News: Metropolitan Journalism in the Digital Age e Apostles of Certainty: Data Journalism and the Politics of Doubt.

Referências

ANDERSON, C. W. *Towards a Sociology of Computational and Algorithmic Journalism*. New Media e Society, 15:7, 2013, p. 1005–1021.

ANDERSON, C. W. *Apostles of Certainty: Data Journalism and the Politics of Doubt*. Nova York: Oxford University Press, 2018.

FOUCAULT, Michel. *Power/Knowledge: Selected Interviews and Other Writings*. Nova York: Vintage, 1980, p. 1972-1977.

NOWOTNY, Helga. *The Cunning of Uncertainty*. Londres: Polity Press, 2016.

YONG, Ed. *How the GOP Could Use Science's Reform Movement Against It*. The Atlantic, 2017

Padrão ouro em dados: o que o setor valoriza como digno de premiação e como o jornalismo coevolui com a dataficação da sociedade

Wiebke Loosen

Introdução: a resposta do jornalismo à dataficação da sociedade

Talvez possamos entender, melhor do que em seus primórdios, o jornalismo de dados e sua ascensão como uma espécie de resposta jornalística à dataficação da sociedade.[246] Dataficação é um termo que se refere à crescente disponibilidade de dados originada na digitalização de nosso ambiente (de mídia), rastros digitais e big data acumulados ao longo do processo de viver em tal ambiente (Dijck, 2014). Este processo transforma muitos aspectos de nossa vida social em dados computadorizados que, para diversos fins, são agregados e processados por meio de algoritmos. A dataficação leva a uma série de consequências e se manifesta de variadas formas na política, por exemplo, diferentemente do que acontece no mundo financeiro ou no campo da educação. Porém, o que todos os domínios financeiros têm em comum é que podemos presumir que cada vez mais dependerão de uma variada gama e um volume ainda maior de dados em seus processos de (auto)entendimento.

Situar a dataficação do jornalismo em relação à dataficação da sociedade em geral nos ajuda a olhar além do jornalismo de dados, identificando-o como "apenas" uma ocorrência de um processo e, para melhor entender, a transformação do jornalismo em uma prática baseada em dados, operada com algoritmos, movida por métricas e, talvez, até mesmo automatizada.[247] Isso inclui, em específico, os objetos e tópicos que o jornalismo deve cobrir, ou, falando de outra forma, a função do jornalismo como observador da sociedade: quanto mais campos e domínios sociais que deveriam ser cobertos pelo jornalismo tornam-se "dataficados", mais o ofício precisa ser capaz de compreender e produzir dados para cumprir seu papel social. Essa relação se reflete no jornalismo de dados contemporâneo, que depende precisamente desta disponibilidade maior de dados para expandir o repertório de fontes para pesquisa jornalística, identificar e contar histórias.

---

[246] https://www.kofi.uni-bremen.de/fileadmin/user_upload/Arbeitspapiere/CoFi_EWP_No-18_Loosen.pdf.

[247] https://www.kofi.uni-bremen.de/fileadmin/user_upload/Arbeitspapiere/CoFi_EWP_No-18_Loosen.pdf.

Premiações: meios de estudar o que é definido e valorizado como jornalismo de dados

Uma forma de traçar a evolução do jornalismo de dados enquanto estilo de reportagem é observar a sua produção. Ao passo que os primeiros estudos em pesquisa jornalística tendiam a focar mais nos atores envolvidos em sua produção, em grande parte baseados em entrevistas, cada vez mais pesquisas vêm sendo conduzidas através da análise de conteúdo para melhor compreender o jornalismo de dados com base em seus produtos (Ausserhofer et al., 2017). Premiações de jornalismo são um bom ponto de acesso empírico para este fim por diversas razões: primeiro, o material enviado a premiações já se mostrou um objeto útil para análise de gêneros e aspectos narrativos (Wahl-Jorgensen, 2013). Segundo, o jornalismo de dados é um objeto de estudo difuso que não só dificulta, mas torna precondicional a identificação de respectivo material para análise de conteúdo. A amostragem de indicados, por sua vez, impede de se começar com uma definição muito estreita ou muito ampla, uma estratégia que serve como uma observação da auto-observação no jornalismo, já que tais produtos representam o que o próprio campo considera como jornalismo de dados e crê serem exemplos significativos deste estilo de reportagem. Terceiro, indicações a premiações internacionais podem influenciar o desenvolvimento do campo como um todo por conta de seu reconhecimento, uma espécie de padrão ouro dentro da área — e, desta forma, contando com impacto transfronteiriço. Além do que, observar premiações internacionais nos permite investigar uma amostra que cobre grande amplitude geográfica e temporal.

É importante ter em mente, porém, que estudar premiações (jornalísticas) traz consigo seus próprios vieses. O estudo com o qual trabalhamos aqui é baseado na análise de 225 produtos jornalísticos indicados (39 deles vencedores) para o Data Journalism Awards (DJA), premiação anual realizada pela Global Editors Network[248] entre 2013 a 2016 (Loosen et al., 2017). Isso significa que nossa amostragem está sujeita a um viés de dupla seleção. Primeiramente por seleção própria, já que cabe aos jornalistas enviarem seu material para que este possa ser indicado. No segundo passo, um júri de especialistas que muda quase que anualmente decidirá quais inscritos serão, de fato, indicados. Prêmios e premiações representam um tipo específico de "capital cultural", o que explica porque alguns projetos premiados podem ter um efeito de sinal para o campo como um todo, servindo de modelo para empreitadas futuras (English, 2002). Isso significa que premiações não representam somente o setor (de acordo com certos padrões), mas também o constituem. Ou seja, no nosso caso, ao rotular conteúdo como jornalismo de dados, estas premiações têm um papel na reunião de diferentes práticas, atores, convenções e valores. Pode-se considerar, então, que estes eventos não têm apenas uma *função de criação de premiações,* mas também de *criação do próprio setor*. Produtos dignos de premiação sempre são resultado de uma espécie de

[248] https://www.globaleditorsnetwork.org/about-us/, https://www.datajournalismawards.org/.

“coconstrução” por parte de inscritos e jurados e das expectativas moldadas por ambos. Tais efeitos podem ser particularmente influentes no caso do jornalismo de dados, visto que é um estilo de reportagem relativamente novo e todos os atores envolvidos ainda estão em fase de experimentos, em maior ou menor grau.

Evoluindo, mas sem revolucionar: algumas tendências em jornalismo de dados (digno de premiação)

Estudos que analisam matérias baseadas em dados geralmente mostram que a evolução do jornalismo de dados não é necessariamente uma revolução no ofício jornalístico. O resultado disso é contestar a crença generalizada de que este tipo de jornalismo vem revolucionando o setor ao substituir métodos tradicionais de reportagem e descoberta de notícias. Nosso próprio estudo em grande parte concorda com as descobertas de outras análises empíricas de amostras de jornalismo de dados “cotidiano” (Loosen et al., 2017). Estas amostras representam coletas de dados relativamente limitadas, mas aos menos nos permitem traçar alguns desdobramentos e, acima de tudo, algum grau de consistência na produção do jornalismo de dados.

Em termos de quem está produzindo jornalismo de dados a nível digno de premiação, resultados indicam que o “padrão ouro”, aquele digno de reconhecimento para os seus pares, está dominado por jornais e seus departamentos online. Ao longo dos quatro anos analisados, estes representam o maior grupo entre todos os indicados, bem como entre os vencedores (total: 43.1%; premiados no DJA: 37,8%). O outro único agrupamento proeminente é constituído por organizações envolvidas em jornalismo investigativo, caso da *ProPublica* ou do *Consórcio Internacional de Jornalistas Investigativos* (ICIJ), também muitas vezes premiados. Isso pode refletir o viés inerente destas premiações, direcionado a atores já estabelecidos e relevantes, ecoando descobertas de outras pesquisas que o jornalismo de dados acima de certo nível parece empreender em prol de organizações maiores, detentoras de recursos e comprometimento editorial que garantem o investimento em equipes multidisciplinares compostas por redatores, programadores e designers (Young et al., 2017). Tudo isso tem impacto, ainda, no tamanho das equipes: dos 192 projetos em nossa amostragem que contavam com assinatura, em média eram equipes compostas por cinco indivíduos citados como autores ou colaboradores e cerca de um terço dos projetos foi realizado através de colaboração com parceiros externos que contribuíram com a análise ou representações visuais. Tais afirmações se aplicam a projetos premiados, com equipes maiores do que aqueles que foram apenas indicados, de acordo com nossa pesquisa (M = 6,31, Dp = 4.7 x M = 4,75, Dp = 3,8).

Em relação à geografia do jornalismo de dados que é reconhecido nesta corrida, podemos notar a dominância dos Estados Unidos, já que quase metade dos indicados é dos

EUA (47,6%), seguido de longe por Grã-Bretanha (12,9%) e Alemanha (6,2%). Porém, o jornalismo de dados aparenta ser um fenômeno cada vez mais global, com o número de países representados por indicados crescendo a cada ano, totalizando 33 nações de todos os cinco continentes em 2016.

A dependência do ofício de certas fontes influencia também os temas que pode ou não abordar. Como resultado direto disso, o jornalismo de dados pode negligenciar domínios sociais para os quais dados não são produzidos regularmente ou não estão acessíveis. Em termos de temas cobertos, os indicados ao DJA se caracterizam pelo foco em questões políticas, sociais e econômicas, quase metade (48,2%) dos materiais analisados tratando de algum tema político. A pequena parcela de produtos voltados a educação, cultura e esportes – informação alinhada com outros estudos – pode não representar o jornalismo de dados em geral, sendo resultante de um viés voltado a temas "sérios" inerentes a premiações do setor. Porém, isso também pode refletir a disponibilidade ou falta de fontes de dados para diferentes campos e tópicos, ou, no caso de nossa amostragem, os vieses de seleção de inscritos tomando por base o que estes consideram digno de inscrição e que esperam que agradaria os jurados. De forma a obter informações mais confiáveis sobre este ponto de crucial importância, precisaríamos de um estudo comparativo internacional que relacionasse disponibilidade e acessibilidade de dados a temas cobertos pelo jornalismo de dados em diferentes países. Tal estudo ainda está ausente da literatura, mas certamente poderia lançar uma luz sobre quais domínios sociais e tópicos são atendidos através de quais métodos analíticos e fontes de dados. Tal abordagem também ofereceria informações valorosas sobre o outro lado da moeda, os pontos cegos na cobertura baseada em dados por conta da falta de fontes de dados (disponíveis).

Um achado recorrente na pesquisa relacionada a conteúdo em jornalismo de dados é que há uma "dependência de dados públicos previamente processados" advindos de escritórios de estatística e outras instituições governamentais (Tabary et al., 2017; Borges-Rey, 2017). Isso também se aplica a produtos baseados em dados a nível de premiação: observamos uma dependência de informações fornecidas por instituições oficiais (quase 70% das fontes) ou demais organizações não comerciais, como institutos de pesquisa, ONGs e afins, além de dados disponíveis publicamente através de solicitação, ao menos (quase 45%). Por um lado, isso mostra que o jornalismo de dados tem um papel no entendimento da disponibilidade cada vez maior de fontes de informação, por outro, mostra que depende bastante desta mesma disponibilidade, visto que a parcela de informações coletadas, vazadas ou solicitadas por conta própria é substancialmente menor. Independentemente disso, o jornalismo de dados segue ligado à reportagem investigativa, que "levou a uma percepção de que o jornalismo de dados trata apenas de conjuntos enormes de informações, adquiridos por

meio de atos de bravura jornalística".[249] Casos recentes, como os *Panama Papers,* contribuíram para essa visão.[250] Dito isso, o que este caso ainda mostra é que algumas questões complexas de relevância global estão inseridas em meio a dados que demandam cooperação transnacional entre diferentes organizações de mídia. Além do que, é provável que vejamos mais casos como estes logo que rotinas sejam desenvolvidas para monitorar o fluxo internacional de dados — em finanças, por exemplo, não apenas como serviço, mas também na forma de matérias mais profundas e investigativas. Isso poderia estimular um novo tipo de *jornalismo investigativo em tempo real baseado em dados*, que monitora constantemente fluxos de dados financeiros, digamos, e busca anomalias.

Interatividade é tida como critério de qualidade dentro do jornalismo de dados, mas esta é geralmente implementada com um conjunto definido de funcionalidades — nesse tocante, nossos resultados também estão alinhados com outros estudos no que muitas vezes é descrito como "falta de sofisticação" em interatividade relacionada a dados (Young et al., 2017). Mapas com função de zoom e filtros são bastante frequentes, talvez por conta da tendência em utilizar softwares de uso simplificado ou mesmo gratuitos, resultando em visualizações e funcionalidades interativas menos sofisticadas. Porém, projetos premiados em sua maioria fornecem ao menos uma funcionalidade interativa e integram um número maior de representações visuais diferentes. A tendência rumo a opções interativas um tanto quanto limitadas também pode ser um reflexo das experiências dos jornalistas com pouco interesse por parte dos leitores em interatividade mais sofisticada (oportunidades para gamificação ou ferramentas de personalização que possibilitam adequar um produto com dados personalizados, por exemplo). Simultaneamente, porém, funções interativas e visualizações devem apoiar a narrativa e função explicativa de um artigo, nos melhores casos. Isso, por sua vez, demanda soluções adaptadas para cada produto jornalístico baseado em dados.

Um resumo das tendências de desenvolvimento ao longo dos anos mostra um padrão misto ao passo que as parcelas e números médios de categorias sob observação mantiveram-se estáveis ao longo do tempo, ou, em caso de mudança, não foram percebidas mudanças lineares. Muito pelo contrário, foram observados picos e vales erráticos em anos individuais, sugerindo a evolução no esquema tentativa e erro que se espera de um campo em ascensão como o jornalismo de dados. Assim sendo, nos deparamos com alguns desdobramentos consistentes ao longo dos anos, como uma parcela significativamente crescente de material voltado a negócios, bem como um número médio consistente e também crescente de diferentes tipos de visualizações, e um (não significativo do ponto de vista estatístico, mas) constante crescimento no volume de materiais que incluíam algum tipo de crítica (por

---

[249] Ver Parasie (2015), Royal e Blasingame (2015) e Knight (2015).

[250] https://panamapapers.icij.org/.

exemplo, aos métodos de confiscação errôneos empregados pela polícia) ou chamadas à intervenção pública (no caso de emissões de carbono). Estas cresceram de maneira consistente ao longo de quatro anos (2013: 46,4% x 2016: 63%) e se mostraram maiores entre premiados (62,2% x 50%). Podemos interpretar este fato como um indicativo da valorização do potencial investigativo e de sentinela do jornalismo (de dado) e, talvez, como forma de legitimar este campo emergente.

Do jornalismo de dados ao jornalismo dataficado e seu papel na sociedade de dados

O jornalismo de dados representa a emergência de um novo subcampo jornalístico em coevolução paralela com a dataficação da sociedade, um passo lógico na adaptação do ofício à disponibilidade crescente de dados. Porém, o jornalismo de dados não é mais um fenômeno em franca expansão, integrando firmemente a prática dos grandes veículos de comunicação. Um indicador digno de nota disso pode ser encontrado ao se observar o Data Journalism Awards: em 2018, a premiação apresentou uma nova categoria chamada "inovação no jornalismo de dados". Parece, então, que o jornalismo de dados não é mais encarado como inovador por si só, pois já busca por abordagens inovadoras dentro de sua prática contemporânea.[251]

Podemos esperar que a relevância e proliferação do jornalismo de dados evolua junto à crescente dataficação da sociedade como um todo, uma sociedade em que interpretação, decisões e todo tipo de ações sociais dependem de dados. Neste contexto, não é difícil presumir que o termo "jornalismo de dados" se tornará supérfluo em um futuro não tão distante visto que o jornalismo como um todo, bem como o ambiente que integra, é cada vez mais dataficado. Independentemente deste prognóstico se confirmar ou não, os termos "jornalismo de dados" e "sociedade de dados" ainda nos sensibilizam quanto aos processos de transformação fundamental que se dão no jornalismo e além. Isso inclui como e por quais meios o jornalismo observa e cobre a sociedade (dataficada), como monitora seu próprio desempenho, como controla seu alcance e participação do público, e como produz e distribui conteúdo (automaticamente). Em outras palavras, o jornalismo contemporâneo caracteriza-se por sua transformação em uma prática baseada em dados, algoritmos, métricas ou, até mesmo, automatizada.

Porém, dados não são um tipo de "matéria-prima"; não permitem o acesso direto, objetivo ou até privilegiado ao mundo social (Borgman, 2015). Esta circunstância é ainda mais importante para um jornalismo de dados responsável enquanto o processo de dataficação da sociedade avança. O avanço da dataficação e a relevância cada vez maior do

[251] https://www.datajournalismawards.org/categories/.

jornalismo baseado em dados também podem incentivar outros domínios sociais a produzirem ou disponibilizarem mais dados (a jornalistas) e é provável que vejamos a evolução paralela de uma espécie de "RP de dados", *relações públicas baseadas em dados*, produzida e divulgada para influenciar comunicações públicas para seus próprios fins. Isso significa que rotinas de verificação de qualidade, origem e relevância de dados vêm se tornando cada vez mais importantes para o jornalismo (de dados) e trazem à baila a questão de porque talvez não existam dados disponíveis sobre certos fatos ou desdobramentos.

Trocando em miúdos, consigo organizar nossas descobertas de acordo com sete Cs, sete desafios e capacidades subutilizadas do jornalismo de dados que podem ser úteis para sugerir práticas alternativas ou modificadas no setor:

1. **Coleta**: O jornalismo investigativo e o jornalismo crítico de dados devem superar sua dependência quanto a dados publicamente acessíveis. São necessários mais esforços para obtenção de informações e coleta independente destas.

2. **Colaboração**: Por mais que materiais cotidianos baseados em dados sejam cada vez mais fáceis de serem produzidos, projetos mais complexos exigem maiores recursos e equipes dedicadas. Espera-se que o número de tópicos de relevância global também aumente. Estes exigirão investigações baseadas em dados transfronteiriças, envolvendo diversas organizações e, em alguns casos, colaboração com outros campos, como ciência e ativismo de dados.

3. **Crowdsourcing**: O potencial interativo real do jornalismo de dados não está em funcionalidades interativas complexas, mas, sim, em abordagens de crowdsourcing que envolvem usuários ou cidadãos, colocando-os no papel de coletores, categorizadores e coinvestigadores de dados.[252]

4. **Cocriação**: Abordagens de cocriação, comuns na área de desenvolvimento de software, podem servir de modelo para projetos de longo prazo baseados em dados. Em tais casos, os usuários integram todo o processo, desde a descoberta de um tema até seu desenvolvimento e sua manutenção por um período mais longo.

5. **Competências**: Jornalismo de dados de qualidade exige equipes com amplos conjuntos de habilidades. O papel do jornalista segue relevante, mas estes precisam ter um entendimento mais aprofundado de dados, estruturas de dados e métodos analíticos. Organizações de mídia, por sua vez, precisam de

[252] http://jonathangray.org/2018/08/08/what-can-citizen-generated-data-do/.

recursos para o recrutamento de analistas de dados, também desejáveis em outros setores.

6. **Combinação**: Dados mais complexos exigem análises mais sofisticadas. Métodos que combinam fontes de dados e os interpretam a partir de variados pontos de vista podem ajudar a criar uma imagem mais substancial de fenômenos sociais e fortalecer a capacidade analítica do jornalismo de dados.

7. **Complexidade**: Quando se fala em complexidade, não se pensa somente nos dados em si, mas na sua importância em variados campos sociais, decisões políticas etc.; ao longo destes desdobramentos, o jornalismo de dados cada vez mais será confrontado com RP de dados e informações falsas.

O que isso significa? Considerando o que sabemos a respeito do jornalismo de dados (premiado) em termos de que tipo de material é valorizado, atenção pública generalizada (caso dos *Panama Papers*), e contribuições para uma apreciação geral do jornalismo, que tipo de jornalismo de dados realmente queremos? Nesse tocante, argumentaria que o jornalismo de dados se mostra particularmente relevante em seu papel único como jornalismo responsável integrante da sociedade de dados, ou seja, um jornalismo de dados que:

- Investiga temas relevantes para a sociedade e faz da sociedade de dados algo compreensível e criticável através de seus próprios meios;

- É ciente de seus pontos cegos ao mesmo tempo que questiona a deficiência de dados em certas áreas e se isso é um bom ou mal sinal;

- Tenta ativamente desvendar casos de manipulação e abuso de dados;

- Não esquece, explica e enfatiza que dados são "artefatos humanos", jamais uma série de fatos autoevidentes, muitas vezes coletados em relação a condições e objetivos bastante particulares (Krippendorf, 2016).

Ao mesmo tempo, a particularidade desse tipo de jornalismo, sua dependência de dados, é também seu ponto fraco. Esta limitação tem a ver com a disponibilidade de dados, sua confiabilidade, qualidade e manipulabilidade. Um jornalismo de dados responsável deve refletir sobre sua dependência destes dados, um tema que deveria ser considerado central na discussão ética dentro do contexto deste ofício. Tais condições indicam que o jornalismo de dados não é só um novo jeito de reportar, mas também um meio de intervenção que desafia e questiona a sociedade de dados, marcada por questões epistemológicas centrais que confrontam (não apenas) o que o jornalismo de dados pensa sobre *o que* sabemos (ou podemos saber) e *como* sabemos (através de dados).

Estas questões se fazem mais urgentes quanto maior o volume e a diversidade de dados acabam incorporados ao "circuito da notícia" — como métodos de observação e investigação jornalística, integrando rotinas de produção e distribuição, e como meios de monitorar o consumo por parte do público. É desta forma que o jornalismo dataficado afeta: (1) *a forma como o jornalismo observa o mundo* e constrói notícias a partir de dados; (2) a própria atuação do jornalismo na *facilitação da automação da produção de conteúdo*; (3) a *distribuição e circulação da produção jornalística* dentro de um ambiente moldado por algoritmos e sua lógica subjacente para o processamento de dados; (4) o que se *entende como digno de noticiar* para segmentos de públicos mensurados de maneira cada vez mais granular.

Estes desdobramentos trazem consigo três responsabilidades essenciais para o jornalismo (de dados): observar nosso desenvolvimento rumo a uma sociedade dataficada de maneira crítica, torná-la compreensível através de seus próprios meios, e tornar visíveis as limitações do que pode e deve ser contado e observado pela visão dos dados.

A Profa. Dra. Wiebke Loosen é pesquisadora sênior em jornalismo no Instituto Hans Bredow de Pesquisa de Mídia em Hamburgo e professora da Universidade de Hamburgo.

Referências

AUSSERHOFER, Julian et al. *The datafication of data journalism scholarship: Focal points, methods, and research propositions for the investigation of data-intensive newswork*. Journalism, 2017.

BORGES-RENA, Eddy. *Towards an epistemology of data journalism in the devolved nations of the United Kingdom: Changes and continuities in materiality, performativity and reflexivity*. Journalism, 2017.

BORGMAN, Christine L. *Big data, little data, no data: Scholarship in the networked world*. Cambridge: MIT Press, 2015.

ENGLISH, James F. *Winning the Culture Game: prizes, awards, and the rules of art*. New Literary History 33(1), 2002, p. 109-135.

KNIGHT, Megan. *Data journalism in the UK: A preliminary analysis of form and content*. Journal of Media Practice 16(1), 2015, p. 55–72.

KRIPPENDORF, Klaus. *Data*. In: JENSEN, Klaus B.; CRAIG, Robert T. (ed.). *The International Encyclopedia of Communication Theory and Philosophy.* Volume 1 A-D. Wiley Blackwell: 2016, p. 484-489.

LOOSEN, Wiebke. *Four forms of datafied journalism. Journalism's response to the datafication of society*. Communicative Figurations, artigo preliminar nº 18, 2018. Disponível em: https://www.kofi.uni-bremen.de/fileadmin/user_upload/Arbeitspapiere/CoFi_EWP_No-18_Loosen.pdf.

LOOSEN, Wiebke; REIMER, Julius; DE SILVA-SCHMIDT, Fenja. *Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the data journalism awards 2013-2016*. Journalism, 2017.

PARASIE, Sylvain. *Data-driven revelation? Epistemological tensions in investigative journalism in the age of 'big data'*. Digital Journalism 3(3), 2015, p. 364-380.

ROYAL, Cindy; BLASINGAME, Dale. *Data journalism: An explication*. #ISOJ 5(1), 2015, p. 24-46.

TABARY, Constance; PROVOST, Anne-Marie; TROTTIER, Alexandre. *Data journalism's actors, practices and skills: A case study from Quebec*. Journalism: Theory, Practice, and Criticism 17(1), 2016, p. 66-84.

VAN DIJCK, Jose. *Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology*. Surveillance e Society 12(2), 2014, p. 197–208.

WAHL-JORGENSEN, Karin, *The strategic ritual of emotionality: a case study of Pulitzer Prize-winning articles*. Journalism: Theory, Practice, and Criticism 14(1), 2013, p. 129–145.

YOUNG, Mary Lynn; HERMIDA, Alfred; FULDA, Johanna. *What makes for great data journalism? A content analysis of data journalism awards finalists 2012–2015*. Journalism Practice, 2017, p. 115-135.

Além de cliques e compartilhamentos: como e por que mensurar o impacto de projetos em jornalismo de dados

Lindsay Green-Barber

Jornalismo e impacto

Por mais que muitos jornalistas façam cara feia diante da ideia de impacto jornalístico, a prática contemporânea, enquanto profissão, se constrói sobre uma base de impacto: informar o público para que possamos nos engajar civicamente e cobrar os responsáveis. E por mais que jornalistas se preocupem que pensar sobre, falar sobre, traçar estratégias para e mensurar o impacto positivo (e negativo) de seu trabalho é uma aproximação perigosa do jornalismo com o ativismo, profissionais e comentaristas já gastaram muitos caracteres e pixels discutindo os efeitos negativos das "fake news", desinformação e reportagem partidária a respeito de indivíduos, nossa sociedade e a democracia. Ou seja, enquanto jornalistas evitam discutir o impacto de seu trabalho, reconhecem o impacto cultural, social e político sério causado pelas "fake news".

Além do que, antes de sua profissionalização no final do século XVIII e começo do XIX, o jornalismo tratava-se de trabalho de influência, apoiado por partidos políticos e produzido com o objetivo expresso de apoiá-los e garantir a eleição de seus candidatos.[253] Assim sendo, numa perspectiva histórica, a profissionalização do jornalismo e a adoção (do mito) da imparcialidade são novidade (Hamilton, 2005). A luta do jornalismo em busca da tal "imparcialidade" não foi uma decisão normativa, mas, sim, uma questão atrelada a modelos econômicos em mudança e à necessidade de atingir o maior público possível de forma a gerar receita (Hamilton, 2004).

Dada as crises paralelas e intimamente relacionadas que assolam o modelo de negócios da indústria de notícias e a desconfiança pública na comunicação dentro dos EUA e Europa Ocidental, pode-se argumentar que a negativa do jornalismo em reconhecer seu impacto tenha sido, no melhor dos casos, uma abdicação de suas responsabilidades ou, até mesmo, um fracasso, no pior dos cenários.

Há esperança, porém. Nos últimos anos, algumas organizações de mídia começaram a aceitar o fato de que são influentes na sociedade. A proliferação de veículos de comunicação sem fins lucrativos, muitas vezes com apoio de fundações filantrópicas com uma missão a cumprir ou mesmo de indivíduos, criou uma placa de Petri para experimentos com impacto.

---

253 https://academiccommons.columbia.edu/catalog/ac:jdfn2z34w2.

Muitas empresas de comunicação começaram a aceitar a ideia de que comunicar o impacto positivo de seu trabalho junto ao público funciona como estratégia para gerar confiança e lealdade, que com sorte se traduz em aumento na receita. Em 2017, por exemplo, o *Washington Post* adicionou a frase "A democracia morre na escuridão" em seu expediente, abraçando (e anunciando) seu papel em nosso sistema político. Já a *CNN* criou uma seção intitulada "Impacte Seu Mundo" em seu site, com links para eventos mundiais, reportagens, artigos "impactantes" e caminhos para que o público pudesse agir, de campanhas com hashtags a doações.[254]

Organizações de mídia também passaram a adotar novas estratégias para maximizar o impacto positivo de seu trabalho, bem como o uso de métricas não relacionadas à publicidade e métodos de pesquisa para entender a efetividade destas estratégias. Em alguns casos, métricas digitais podem ser substitutas úteis à mensuração de impacto. Métricas de publicidade como visualizações únicas de página ou dados mais avançados como o tempo gasto em uma página são usados para medir o alcance do conteúdo sem considerar os efeitos deste no indivíduo.

Gostaria de propor uma infraestrutura para o impacto de mídia, uma mudança no status quo como resultado de uma intervenção que inclui quatro tipos de impacto — em indivíduos, em redes, em instituições e no discurso público. Estes impactos estão relacionados entre si. Por exemplo, como muitas vezes supõe um jornalista, reportagens podem aumentar o nível de conhecimento de uma pessoa a respeito de determinado assunto, afetando a escolha de um candidato e, por fim, instituições inteiras. Uma reportagem pode ter, ainda, efeitos imediatos em instituições, levando a demissões ou reestruturações, que, por sua vez, impactam indivíduos. Porém, o impacto catalisado pelo jornalismo geralmente leva tempo e envolve processos sociais complexos.

Diferentes tipos de jornalismo estão melhor equipados para diferentes tipos de impacto. James T. Hamilton mostra que reportagens investigativas podem poupar o dinheiro de instituições ao revelarem casos de prevaricação, corrupção ou irregularidades, estimulando mudanças. Documentários também se mostraram especialmente eficientes na geração ou no fortalecimento de redes de ativismo em prol de mudança.[255]

O restante deste capítulo explora a relação entre jornalismo de dados e impacto, mostrando como a prática pode contribuir para os mais variados tipos de mudanças sociais. Mais adiante, sugere métodos de como mensurar a efetividade do jornalismo de dados e o que jornalistas e organizações de mídia podem fazer com esta informação.

---

[254] https://edition.cnn.com/specials/impact-your-world.

[255] https://www.documentcloud.org/documents/1278731-waves-of-change-the-case-of-rape-in-the-fields.html.

Por que jornalismo de dados?

Por mais que profissionais empreguem técnicas de jornalismo de dados por diversos motivos, dois se destacam: a possibilidade de fornecer provas robustas para as alegações feitas no decorrer de uma narrativa e para apresentar informações ao público na forma de dados, no lugar de uma narrativa baseada em texto. A prática do jornalismo de dados baseia-se no julgamento basilar de que dados são dignos de confiança — e, por extensão, um produto jornalístico que inclui reportagem com dados passa essa mesma credibilidade, possivelmente mais do que teria caso dados não fizessem parte deste.

Reportagens com dados usadas na comunicação de informações como números, dados, tabelas, gráficos ou quaisquer outras representações visuais se assemelham a outros formatos jornalísticos (texto, vídeo, áudio) no sentido de que, em essência, trata-se de uma forma linear de comunicação de informações selecionadas para um público. Já a reportagem apresentada ao público através de interação é um formato único de narrativa no sentido de que supõe que o público interagirá com os dados, fará suas próprias perguntas e buscará por respostas nos dados disponibilizados. Logo, a "narrativa" depende tanto do usuário quanto do trabalho jornalístico.

Até mesmo esta versão tosca do jornalismo de dados implica quatro tipos de impacto.

Indivíduos

O jornalismo de dados tende a focar em membros individuais do público como unidade potencial para mudança, dando informações com credibilidade para que ganhem mais conhecimento sobre determinado tema e, por tabela, tomem decisões mais informadas. Ao passo que esta prática jornalística enquanto base para narrativas tradicionais e lineares aumente a confiança do público em relação ao conteúdo, produtos interativos de notícias ou dados apresentam maior potencial para causar impacto individual, tratando-se de jornalismo de dados.

Com um produto interativo voltado a dados, ou seja, "um grande banco de dados interativo que narra uma notícia", qualquer usuário pode gerar suas próprias perguntas e vasculhar os dados em busca de respostas (Klein, 2012). Empresas de comunicação, muitas vezes, presumem que produtos interativos possibilitarão ao público ir fundo na exploração de dados, encontrando informação relevante e criando narrativas. Em análise de produtos interativos desenvolvidos por uma empresa de comunicação, o autor deste capítulo descobriu que os apps de dados mais bem-sucedidos, aqueles com maior tráfego e índices de exploração, integravam um pacote editorial completo que incluía outros conteúdos, funções

para pesquisar dados locais ou relevantes para a localização geográfica solicitada, com alto grau de interatividade, estética agradável e bem projetada, com carregamento rápido.[256]

Dollars for Docs, da *ProPublica*, é um exemplo clássico de jornalismo de dados no sentido de que acessa volumes significativos de dados, neste caso a respeito de pagamentos feitos por farmacêuticas e fabricantes de dispositivos médicos a profissionais de medicina, estrutura estes dados e apresenta ao público na forma de um banco de dados interativo com o objetivo e inspirar outras pessoas a fazerem suas próprias pesquisas e, possivelmente, agirem.[257] O projeto instrui o público a "usar esta ferramenta" para pesquisar possíveis pagamentos feitos aos seus médicos e, em uma barra lateral, orienta: "Pacientes, ajam. Queremos saber como você usou ou pretende usar essa informação em seu dia a dia. Você chegou a conversar com seu médico? Planeja falar com ele? Nos fale".[258]

Redes

O jornalismo de dados oferece informação com credibilidade que pode ser usada por redes (formais e/ou informais) para fortalecimento de seus posicionamentos e trabalho. Organizações ativistas, por exemplo, muitas vezes valem-se de reportagens baseadas em dados para alavancarem suas reivindicações em chamadas públicas ou durante procedimentos legais, especialmente em casos em que tais informações não estão prontamente disponíveis ao público. A prática jornalística de solicitar acesso a dados não disponíveis ao público, analisar estes dados e publicar as descobertas feitas, absorve custos que seriam intransponíveis para indivíduos ou redes (Hamilton, 2016)..

Instituições

O jornalismo de dados pode gerar o tipo de reportagem que instituições dão duro para manter guardada a sete chaves, visto que revelam casos de corrupção, prevaricação, irregularidades e/ou incompetência. Quando esse tipo de informação vem à tona, há pressão para que as instituições mudem, a partir de ameaças associadas à eleição, para políticos, ou forças de mercado, no caso de empresas de capital aberto.

Os *Panama Papers,* do Consórcio Internacional de Jornalismo Investigativo, são um exemplo, uma investigação colaborativa que analisou mais de 11,5 milhões de documentos para desmascarar "políticos de mais de 50 países ligados a empresas offshore em 21 paraísos

---

256 https://s3-us-west-2.amazonaws.com/revealnews.org/uploads/CIR+News+Interactives+White+Paper.pdf.

257 https://projects.propublica.org/docdollars/.

258 https://propublica.forms.fm/was-the-dollars-for-docs-information-helpful-tell-us-how-you-will-use-it/forms/2249.

fiscais".[259] Este trabalho levou à renúncia de políticos, caso do primeiro-ministro da Islândia, Sigmundur David Gunnlaugsson, e a investigações de outros, como o antigo primeiro-ministro do Paquistão, Nawaz Sharif (condenado a dez anos de prisão em 2018), dentre outras incontáveis reações institucionais.

Discurso público

Como o jornalismo de dados pode ser dividido em partes menores, geograficamente, demograficamente, ou com base em outros fatores, suas informações podem ser usadas para contar histórias diferentes em diferentes meios. Sendo assim, sua prática pode ser localizada de forma a causar mudança no diálogo público sobre diversos temas por meio de localizações geográficas, grupos demográficos ou outras fronteiras sociais.

O *Center for Investigative Reporting* publicou conjuntos interativos de dados dos EUA sobre o Departamento de Assuntos dos Veteranos, um deles sobre o tempo médio de espera de veteranos ao buscarem atendimento médico em hospitais ligados ao departamento, o outro com o número de opiáceos prescritos a veteranos pelos sistemas do mesmo departamento. Em ambos os casos, organizações jornalísticas locais usaram as informações de base para as reportagens sobre estes temas com foco na região em que se encontram.

Mas, então, como jornalistas podem criar uma estratégia voltada ao impacto?

Você fez toda a parte difícil: obteve acesso aos dados, trabalhou em cima dos números, estruturou a informação e tem uma história importante a contar. E agora?

Uma estratégia de alto impacto para produtos jornalísticos pode seguir estes cinco passos:

1. Defina objetivos

O que pode acontecer como resultado de seu projeto? Quem ou o que tem poder e/ou incentivo para tratar de quaisquer irregularidades? Quem deveria ter acesso à informação que você está trazendo? Faça estas perguntas a si mesmo para definir qual ou quais tipos de impacto fazem sentido para o seu projeto.

2. Conteúdo

Com objetivos definidos, é hora de identificar o público-alvo da empreitada. Em quais fontes de notícia e informação estes públicos confiam? Como podem ter o melhor acesso

[259] https://www.icij.org/investigations/panama-papers/.

possível a estas informações? É necessário criar um produto interativo ou uma narrativa linear será mais eficaz?

3. Engajamento

Como você e seu veículo de comunicação engajarão o público e como o público engajará com o seu trabalho? Por exemplo, se você identificou outro veículo como fonte de confiança para determinado público, proponha uma colaboração. Se seu produto interativo tem informações importantes a respeito da comunidade de determinada ONG, promova um webinário explicando como usá-lo.

4. Pesquisa estratégica

Dependendo de seus objetivos, planos de conteúdo e engajamento, escolha os métodos e/ou indicadores apropriados de pesquisa de forma a monitorar o progresso e entender o que está funcionando ou não. Ao passo que veículos tendem a falar sobre "mensurar" o impacto de seu trabalho, prefiro usar o termo "pesquisa estratégica", pois tanto métodos quantitativos quanto qualitativos devem ser levados em consideração. Quanto antes se identificarem métodos e indicadores de pesquisa, melhor a informação a ser obtida (a seção a seguir discute opções de mensuração em maior profundidade).

5. Repetição

Tempo e recursos foram gastos com reportagem, conteúdo, engajamento e mensuração de seu projeto em jornalismo de dados. O que deu certo? O que mudará para a próxima vez? Quais questões ficaram em aberto? Compartilhe este aprendizado com sua equipe e o campo como um todo para levar seu próximo projeto além.

Como "mensurar" o impacto de nosso trabalho?

Em outro momento, aludimos ao fato de que a pesquisa de impacto de mídia tem sido dominada por métricas de publicidade. Porém, métricas como visualizações de página, tempo gasto na página e taxas de rejeição podem representar algum impacto. Seu objetivo é medir a exposição total de conteúdo a indivíduos sem considerar suas opiniões sobre as questões tratadas, se aprenderam ou não novas informações ou sua intenção em agir com base no conteúdo apresentado. Ao considerar o impacto do conteúdo em indivíduos, redes, instituições e discurso público, há outros métodos qualitativos e quantitativos inovadores que podem ser utilizados para um melhor entendimento de como o produto afetou estes mesmos indivíduos, redes, instituições e discurso público. Esta seção explora um punhado de métodos promissores de pesquisa para compreensão do impacto do jornalismo de dados.

Analítica

Métricas de mídia podem ser encaradas como representativas de resultados desejados, como maior conscientização ou maior conhecimento sobre determinado assunto. Porém, empresas de comunicação devem tomar cuidado e agirem intencionalmente ao atribuir mudanças à analítica. Por exemplo, se um projeto de jornalismo de dados tem como objetivo causar mudanças institucionais, visualizações únicas de página não são um indicador apropriado de sucesso; menções de dados por agentes públicos em documentos seriam um indicador melhor.

Pesquisa experimental

A pesquisa experimental cria condições constantes sobre as quais os efeitos de uma intervenção podem ser testados. O Centro de Engajamento de Mídia da Universidade do Texas, em Austin, conduziu pesquisas fascinantes a respeito dos efeitos do layout de páginas iniciais de notícias na lembrança e no sentimento do público, e sobre reportagem voltada a soluções e sua relação com o sentimento do público para organizações de notícias. Empresas de tecnologia frequentemente testam os efeitos de elementos interativos diversos em usuários. Veículos de comunicação podem fazer o mesmo para melhor entender o efeito de produtos interativos em sua base de usuários, em parceria com universidades ou trabalhando diretamente com pesquisadores de áreas como marketing, desenvolvimento de negócios e engajamento de audiência dentro da própria redação.

Pesquisa de campo

Por mais que pesquisas de campo não sejam o que há de mais moderno dentro do campo de pesquisa como um todo, são um método já estabelecido de obtenção de informações junto a indivíduos sobre mudanças de interesse, conhecimento, opiniões e ações. Estas pesquisas podem ser feitas de forma criativa pelas organizações, utilizando tecnologias que permitem, por exemplo, o uso de pop-ups acionados pelo retorno do usuário ou monitoramento de cliques em newsletters para criação de um grupo de possíveis participantes.

Análise de conteúdo

Trata-se de um método de pesquisa usado para determinar mudanças no discurso ao longo do tempo. Este método pode ser aplicado a qualquer corpus baseado em texto, o que o torna extremamente flexível. Um exemplo: quando uma organização produz conteúdo com o objetivo de influenciar o discurso público a nível nacional, é esperado que realize também análise de conteúdo após o projeto considerando os dez principais jornais do país para determinar a influência de suas narrativas. Se o objetivo é influenciar uma legislatura

estadual, a organização pode empregar a análise de conteúdo pós-projeto em cima de agendas legislativas disponíveis ao público.[260] Ou, se o objetivo é disponibilizar informações para redes ativistas, esta análise de conteúdo pode ser aplicada a newsletters de organizações.

Esta análise de conteúdo pode ser feita de, no mínimo, três formas. Em um nível mais básico, o veículo pode procurar pelas citações a um projeto para documentar onde e quando foi citado. Muitos jornalistas criam alertas no Google usando uma palavra-chave de sua reportagem, junto com seu sobrenome, para determinar em que outros veículos seu projeto foi citado. Não é uma metodologia sem falhas, mas pode fornecer informações interessantes e uma noção informal do impacto obtido. Tal processo pode gerar questões adicionais a respeito do impacto de um projeto que valem uma análise mais profunda. Muitos veículos e empresas de comunicação empregam serviços de clipping, como alertas do Google News ou Meltwater, para este fim.

Uma análise de conteúdo rigorosa, porém, identifica termos-chave, dados, e/ou frases em um projeto, analisando sua prevalência antes e depois da publicação em um corpus finito de texto para documentar mudanças. A análise computacional de textos dá um passo além e infere mudanças no discurso através de técnicas avançadas de contagem e análise. Estes métodos mais rigorosos geralmente exigem que um veículo ou empresa de comunicação firme uma parceria com pesquisadores especializados.

Pensando no futuro: por que jornalistas devem se importar com o impacto do jornalismo de dados

Para continuar relevante, o jornalismo precisa aceitar não apenas seu impacto na sociedade, mas abraçar este fato. Ao trabalhar para compreender o ecossistema da mudança no qual o jornalismo opera, bem como seu papel específico dentro deste, a indústria pode trabalhar para maximizar seu impacto positivo e demonstrar seu valor para o público.

Jornalistas de dados, dotados de sua compreensão do valor e da importância de dados quantitativos e qualitativos, estão em uma posição privilegiada para isso. Ao articular os objetivos de projetos de jornalismo de dados, criando formas criativas de engajar com o público e estratégias de distribuição, acompanhadas de métodos sofisticados para a mensuração do sucesso destes projetos, repórteres podem liderar este movimento atuando de dentro.

---

[260] https://s3-us-west-2.amazonaws.com/revealnews.org/uploads/CIR+News+Interactives+White+Paper.pdf.

Lindsay Green-Barber é uma profissional motivada e colaborativa com mais de dez anos de experiência em pesquisa quantitativa e qualitativa, novas tecnologias de informação e comunicação, inovação de mídia, planejamento estratégico e comunicação, mensuração e avaliação de programas, gestão, captação de recursos e administração de negócios.

Referências

GREEN-BARBER, Lindsay. *Changing the conversation: The VA backlog*. The Center for Investigative Reporting, 2015. Disponível em: https://s3.amazonaws.com/uploads-cironline-org/uploaded/uploads/VA+Backlog+White+Paper+11.10.14.pdf.

GREEN-BARBER, Lindsay. *What makes a news interactive successful? Preliminary lessons from The Center for Investigative Reporting*. The Center for Investigative Reporting, 2015. Disponível em: https://s3-us-west-2.amazonaws.com/revealnews.org/uploads/CIR+News+Interactives+White+Paper.pdf.

GREEN-BARBER, Lindsay. *Waves of Change: The Case of Rape in the Fields*. The Center for Investigative Reporting, 2014. Disponível em: https://www.documentcloud.org/documents/1278731-waves-of-change-the-case-of-rape-in-the-fields.html.

GREEN-BARBER, Lindsay; FERGUS, Pitt. *The Case for Media Impact: A Case Study of ICIJ's Radical Collaboration Strategy*. Tow Center for Digital Journalism, 2017. Disponível em: https://academiccommons.columbia.edu/catalog/ac:jdfn2z34w2.

GROSECLOSE, Tim; MILYO, Jeffrey. *A Measure of Media Bias*. The Quarterly Journal of Economics (4), 2005, p. 1191-1237.

HAMILTON, James. *All the News That's Fit to Sell: How the Market Transforms into News*. Princeton University Press, 2004.

HAMILTON, James. *Democracy's Detectives: The Economics of Investigative Journalism*. Cambridge: Harvard University Press, 2016.

KLEIN, Scott. *The Data Journalism Handbook*. O'Reilly Media, 2012.

Jornalismo de dados com impacto

Paul Bradshaw

Se você não assistiu *Spotlight: Segredos Revelados*, o filme sobre a investigação do *Boston Globe* a respeito do silêncio institucional sobre abuso infantil, essa é a hora. Indo direto ao ponto: assista até os intertítulos no final.[261]

Uma lista passa pela tela. Ela detalha as dezenas de lugares onde se descobriram escândalos desde os eventos mostrados no filme — de Akute, na Nigéria, a Wollongong, na Austrália. Estes intertítulos nos fazem deixar de lado qualquer comemoração ao declararem que uma das peças centrais envolvidas no escândalo foi transferia para "uma das maiores igrejas católico-romanas do mundo".

Este é o desafio que envolve a questão do impacto no contexto do jornalismo de dados: chamar atenção para um problema pode ser considerado "impacto"? O desfecho da história precisa levar a penalidade, compensação, ou uma mudança visível de políticas? Quão relevante é o impacto? E para quem?

Estas últimas duas questões merecem ser abordadas primeiro. Tradicionalmente, o impacto é considerado relevante por duas razões principais: comercial e cultural. Em termos comerciais, mensurações de impacto como reconhecimento de marca e altos índices de audiência podem contribuir diretamente para a margem de lucro de uma publicação, através de publicidade (afetando preço e volume) e vendas de assinaturas/cópias.[262] Culturalmente, porém, artigos e matérias de impacto deram a veículos e organizações de notícias, bem como jornalistas, uma espécie de direito a se gabar entre seus pares. Tudo ficou mais complicado, como veremos adiante.

Mensurações de impacto no jornalismo sempre foram, historicamente, limitadas. Vendas agregadas e índices de audiência, uma série limitada de premiações do setor e pesquisas ocasionais com o público: era nisso que editoras se baseavam.

Agora, claro, o desafio não é somente a proliferação de métricas, mas de modelos de negócios, com a expansão de veículos de notícias sem fins lucrativos, em especial, gerando

---

261 https://www.imdb.com/title/tt1895587/quotes/qt3112625?mavIsAdult=false&mavCanonicalUrl=https%3A%2F%2Fwww.imdb.com%2Ftitle%2Ftt1895587%2Fquotes.

262 https://www.theguardian.com/news/2018/aug/31/alan-rusbridger-who-broke-the-news.

uma ênfase crescente em questões de impacto e discussões sobre como este pode ser mensurado.[263]

Além do que, a possibilidade de mensurar impacto a cada matéria ou artigo significa que editores não são mais os únicos responsáveis por este, incluindo também jornalistas.

Mensurando impacto com números

Talvez a medida mais fácil seja *alcance* puro e simples. Produtos interativos baseados em dados — como *7 billion people and you: What's your number?,* da *BBC* — atingiram milhões de leitores com um produto atual, ao passo que em determinado momento de 2012, o jornalismo de dados feito por Nate Silver alcançava um entre cada cinco visitantes do *New York Times*.[264]

Alguns podem fazer cara feia diante de medições tão brutas, mas elas são importantes. Se em algum ponto jornalistas foram criticados ao tentarem impressionar seus colegas à custa do público, no mínimo espera-se do jornalismo moderno que prove que consegue se conectar a este mesmo público. Na maior parte dos casos, tais provas são necessárias para anunciantes, mas até mesmo provedores universais de notícias financiados com dinheiro público como a *BBC* precisam destas informações, de forma a provar que estão atendendo os requisitos necessários para receberem este financiamento.

O *engajamento* é a relação mais sofisticada do alcance e nesse tocante o jornalismo de dados se dá bem. Durante conferência da editora de jornais *Reach*, realizada com editores, foi revelado que adicionar uma visualização de dados a uma página pode aumentar o tempo gasto nela em até um terço. A interatividade fundamentada em dados pode transformar o mais tedioso dos assuntos. Em 2015, David Higgerson, desta mesma empresa, percebeu que mais de 200.000 pessoas inseriram seus CEPs em um widget interativo criado por sua equipe de dados com base em estatísticas de privação — o que David considerou um número bem maior "do que o esperado [para] um artigo mais direto do tipo 'dados nos dizem isso ou aquilo'".[265]

O engajamento é particularmente importante para organizações que dependem de publicidade (alta de preços acompanha altos índices de engajamento), mas para aquelas que

---

263 https://reutersinstitute.politics.ox.ac.uk/our-research/speed-not-everything-how-news-agencies-use-audience-metricsSchlemmer, https://www.tandfonline.com/doi/abs/10.1080/21670811.2018.1445002?journalCode=rdij20.

264 https://www.bbc.com/news/world-15391515, https://www.politico.com/blogs/media/2012/11/20-of-nyt-visitors-read-538-148670.

265https://davidhiggerson.wordpress.com/2015/10/14/how-audience-metrics-dispel-the-myth-that-readers-dont-want-to-get-involved-with-serious-stories/.

têm assinaturas, doações e eventos relevantes para a receita, bom, estas também mostram-se ligadas a engajamento.

Financiamentos voltados a organizações sem fins lucrativos ou bolsas, muitas vezes, são acompanhados de uma demanda explícita por monitoramento ou demonstração de impacto, que vai além de alcance puro e simples. *Mudança* e *ação*, políticas ou legais, são referenciadas. O Consórcio Internacional de Jornalistas Investigativos (ICIJ), por exemplo, destaca o impacto de seu projeto *Panama Papers* por ter resultado em "pelo menos 150 consultas, auditorias ou investigações... Em 79 países" junto de métricas mais comuns, como quase 20 premiações, incluindo um Pulitzer.[266] No Reino Unido, há um espaço dedicado na história do jornalismo de dados para o escândalo dos gastos do Parlamento. Isso não só colocou o *The Telegraph* como guia das pautas do noticiário por semanas, como levou à criação de um novo órgão, o IPSA, sigla em inglês para Autoridade Independente de Normas Parlamentares. O órgão, agora, divulga dados abertos sobre as despesas de políticos, permitindo maior transparência e fomentando o jornalismo de dados.

Mas diretrizes, guias e formulação de políticas podem ir muito além da própria política. As diretrizes de empréstimos de bancos afetam milhões de pessoas e passaram por um notório processo de responsabilização no final da década de 1980 nos EUA, quando Bill Dedman publicou uma série de artigos intitulada *Colour of Money*, premiada pelo Pulitzer. Ao identificar práticas de empréstimo que levavam em conta critérios raciais (uma prática conhecida como "*redlining*", exclusão por razões discriminatórias), a apuração baseada em dados causou mudanças políticas, financeiras e legais, por meio de investigações mais aprofundadas, novos financiamentos, processos e leis aprovadas ao longo da publicação de cada artigo.[267]

Trinta anos depois, vemos uma versão bastante moderna desta mesma abordagem na série *Machine Bias,* da *ProPublica*, que lança uma luz sobre prestação de contas algorítmica, enquanto o *Bureau Local* utilizou sua rede para juntar informações através de crowdsourcing a respeito de '*dark posts*' direcionados por algoritmos nas redes sociais.[268] Ambos os projetos contribuíram para a mudança de políticas do Facebook. Já os métodos da *ProPublica* foram adotados por um grupo voltado aos direitos de moradia para dar base a um processo contra a

---

266 https://www.icij.org/investigations/panama-papers/20161201-global-impact/. https://www.icij.org/about/awards/.

267 http://powerreporting.com/color/.

268 https://www.thebureauinvestigates.com/stories/2017-05-18/campaigners-target-voters-brexit-dark-ads.

rede social.[269] Com a influência cada vez maior dos algoritmos em nossas vidas, afetando fatores como a alocação da polícia e o preço do Uber em áreas povoadas por pessoas de cor, a transparência e responsabilização destes acaba por ser tão importante quanto as cobranças voltadas a formas mais tradicionais de poder.[270]

O que é notável a respeito de alguns destes exemplos é que seu impacto depende e é parcialmente demonstrado através da colaboração com terceiros. Quando o *Bureau Local* fala de impacto, por exemplo, eles estão se referindo aos números de histórias produzidas por membros de sua rede de base, inspirando outros a agirem; enquanto o ICIJ lista a escala crescente de suas redes: "*LuxLeaks* (2014) envolveu mais de 80 repórteres em 26 países.[271] O *Swiss Leaks* (2015) envolveu mais de 140 repórteres em 45 países". O número sobe para mais de 370 repórteres em quase 80 países no caso dos *Panama Papers*: 100 organizações de mídia publicando 4.700 artigos.[272]

Os dados coletados e publicados como resultado destas investigações podem se tornar uma fonte de impacto: o banco de dados Offshore Leaks, de acordo com o ICIJ, "é usado regularmente por acadêmicos, ONGs e agências fiscais".

Há algo de surpreendente nesta mudança de sentir orgulho ao publicar algo em busca de aclamação para a atuação de facilitadores, organizadores e administradores de bancos de dados. Como resultado desse processo, o ato de colaborar virou uma habilidade por si só. Muitas organizações sem fins lucrativos contam com vagas de gerente de comunidades ou projetos dedicados a construir e manter relações com colaboradores e parceiros. A educação jornalística também reflete estas mudanças cada vez mais.

Podemos traçar esse processo à influência da cultura dos primórdios do jornalismo de dados. Ao escrever sobre a prática no Canadá, em 2016, Alfred Hermida e Mary Lynn Young notaram "uma divisão do trabalho em evolução que prioriza relações jornalísticas interorganizacionais em rede".[273] Tal influência foi reconhecida ainda mais em 2018, quando o Instituto Reuters publicou um livro sobre a ascensão do jornalismo colaborativo, onde

---

269 https://www.seattletimes.com/business/facebook-vows-more-transparency-over-political-ads/, https://www.propublica.org/article/facebook-fair-housing-lawsuit-ad-discrimination, https://newsroom.fb.com/news/2017/02/improving-enforcement-and-promoting-diversity-updates-to-ads-policies-and-tools/.

270 https://www.themarshallproject.org/2016/02/03/policing-the-future#.wOF9SrXzh, https://gijn.org/2016/05/02/investigating-uber-surge-pricing-a-data-journalism-case-study/.

271 https://www.icij.org/blog/2017/11/icij-went-no-data-team-tech-driven-media-organization/.

272 https://niemanreports.org/articles/how-some-370-journalists-in-80-countries-made-the-panama-papers-happen/.

273 https://www.tandfonline.com/doi/abs/10.1080/21670811.2016.1162663.

constava que "a colaboração pode virar uma história por si só, contribuindo para o impacto do jornalismo".[274]

Mudanças no que contamos, como contamos e se fizemos certo

Conhecimento técnico avançado não é um requisito obrigatório na criação de uma história de impacto. Um dos mais longevos projetos de jornalismo de dados, *Drone Warfare*, do *Bureau de Jornalismo Investigativo*, vem monitorando ataques com drones feitos pelos EUA há mais de cinco anos.[275] Sua metodologia central pode ser resumida em uma única palavra: persistência.[276] Semanalmente, jornalistas do *Bureau* transformaram "texto corrido" em um conjunto estruturado de dados que pode ser analisado, pesquisado e investigado. Estes dados, complementados por entrevistas com outras fontes, vêm sendo utilizados por ONGs, e o próprio *Bureau* enviou provas textuais ao Comitê de Defesa do Reino Unido.[277]

Enumerar aquilo que não foi contado é uma maneira especialmente importante pela qual o jornalismo de dados pode gerar impacto, de fato. Seria razoável dizer que é o equivalente do setor a "dar voz aos sem voz". O projeto *Migrants Files*, que envolve jornalistas de mais de 15 países, teve início após estes jornalistas de dados notarem "não haver um banco de dados utilizável de pessoas que morreram em suas tentativas de chegar ou ficar na Europa".[278] Seu impacto forçou outras agências a atuarem, entre elas a Organização Internacional para as Migrações, que agora coletam seus próprios dados.

Mesmo quando um governo parece contabilizar algo, pode valer a pena investigar. Ao trabalhar junto com a Unidade de Dados da *BBC* da Inglaterra durante investigação sobre a escala de cortes orçamentários em bibliotecas, vivi um momento de pânico quando vi que uma pergunta sobre dados envolvendo o tema havia sido feita no Parlamento.[279] Será que a resposta sairia na frente do trabalho de meses que vínhamos fazendo? No final das contas, nada disso aconteceu e foi ali que soubemos que o próprio governo sabia menos do que a gente sobre a escala real daqueles cortes, pois não havia analisado a situação com a mesma profundidade que nós.

---

[274] https://reutersinstitute.politics.ox.ac.uk/risj-review/global-teamwork-rise-collaboration-investigative-journalism.

[275] https://www.thebureauinvestigates.com/projects/drone-war.

[276] https://www.thebureauinvestigates.com/explainers/our-methodology.

[277] https://publications.parliament.uk/pa/cm201314/cmselect/cmdfence/772/772vw08.htm.

[278] http://www.themigrantsfiles.com/.

[279] https://www.bbc.com/news/uk-england-35707956.

Por vezes, o impacto está não apenas na existência dos dados, mas em sua representação. Um projeto do jornal mexicano *El Universal*, chamado *Ausências Ignoradas*, deu rosto a mais de 4.500 mulheres que sumiram no país ao longo de uma década.[280] Os dados existiam, mas não haviam sido apresentados a um nível mais "humano". *Meurtres conjugaux, des vies derrière les chiffres*, do *Libération*, fez o mesmo a respeito de homicídios domésticos de mulheres. Já o projeto *Kadin Cinayetleri*, do *Ceyda Ulukaya*, mapeou feminicídios na Turquia.[281]

Quando os dados são ruins: impactos na qualidade da informação

Alguns de meus projetos favoritos enquanto jornalista de dados são aqueles que destacam ou levam à identificação de dados falhos ou à falta destes. Em 2016, a Unidade de Dados da *BBC* da Inglaterra observou quantas escolas acadêmicas seguiam regras de transparência. Para tanto, foi escolhida uma amostragem aleatória de 100 escolas e foi verificado se estas publicavam registros dos interesses comerciais e pecuniários de seus diretores, como exigido pela legislação. Uma em cada cinco academias não cumpriu a exigência, resultando em ação por parte da reguladora Ofcom para com as escolas identificadas por nós.[282] Mas a pergunta a ser feita é: esta fiscalização perduraria? Retomar o assunto anos depois seria importante para determinar se o impacto foi de curto prazo ou mais sistêmico.

Por vezes, o impacto de um projeto de jornalismo de dados é um subproduto, identificado apenas quando o material já está pronto e buscam-se respostas. Quando o *Bureau Local* reportou que 18 conselhos espalhados pela Inglaterra não dispunham de caixa para lidar com quaisquer incertezas financeiras, e buscou informações junto a estes, descobriu-se que os dados estavam errados.[283] Ninguém havia percebido o erro em meio aos dados, disseram à época: "Nem os conselhos que compilaram as informações, nem o Ministério de Habitação, Comunidades e Governo Local, que examinaram e então divulgaram [estes]". A investigação contribuiu com uma campanha crescente para que órgãos locais publicassem dados de maneira mais consistente, aberta e precisa.

---

280 https://onlinejournalismblog.com/2016/06/22/mexico-data-journalism-ausencias-ignoradas/.

281 https://www.liberation.fr/apps/2018/02/meurtres-conjugaux-derriere-les-chiffres/, http://kadincinayetleri.org/, http://datadrivenjournalism.net/featured_projects/kadincinayetleri.org_putting_femicide_on_the_map.

282 https://www.bbc.com/news/uk-england-37620007.

283 https://www.thebureauinvestigates.com/blog/2018-05-02/inaccurate-and-unchecked-problems-with-local-council-spending-data.

Impacto além da inovação

Ao passo que o jornalismo de dados se torna mais rotineiro e integrado a modelos de negócio cada vez mais complexos, seu impacto mudou da esfera da inovação para a da entrega. Como dito pelo editor David Ottewell a respeito do tema, em 2018:

> A inovação leva o jornalismo de dados à primeira página. Entrega é chegar à primeira página dia após dia. Inovação é criar um produto interativo incrível que permite a leitores explorarem e compreenderem questões importantes. Entrega é fazer isso e conseguir com que muita gente de fato use esse produto; então criar outro no dia seguinte, e no dia após esse (Ottewell, 2018).[284]

Entrega, é claro, também tem a ver com o impacto além de nossos pares, além de impressionar com uma visualização de dados ou mapa interativo, tem a ver com afetar o mundo real. O efeito pode ser imediato, óbvio e mensurável, ou pode ser algo que toma tempo, passa despercebido e de forma difusa. Por vezes, podemos sentir que não fazemos diferença alguma, como no caso do padre da matéria do *Boston Globe*, mas a mudança pode levar tempo: o jornalismo pode semear a mudança, com resultados aparecendo anos ou décadas depois. Tanto o *Bureau Local* quanto a *BBC* não sabem se os dados dos conselhos ou das escolas serão mais confiáveis no futuro, mas sabem que o holofote foi lançado sobre estas instituições para que possam melhorar.

Às vezes, lançar luz sob determinado tema e aceitar que é responsabilidade de terceiros agir é tudo que cabe ao jornalismo; em outras, o jornalismo pode agir por conta própria e promover campanhas para maior abertura. Para tanto, o jornalismo de dados conta com a habilidade de forçar uma maior abertura ou a criação de ferramentas que possibilitam que terceiros ajam.

No final das contas, o jornalismo de dados com impacto pode definir o curso de ação. Ele alcança públicos que outros tipos de jornalismo não alcançam e interage com estas pessoas de formas que outros jornalismos não o fazem. Ele dá voz aos sem voz, e joga uma luz sobre informações que, de outra forma, permaneceriam obscurecidas. Ele reforça a transparência e responsabilização em cima de dados, e é verdadeiro quanto à sua força.

Parte desse impacto é quantificável, enquanto o restante pode ser difícil de ser mensurado, o que qualquer tentativa de monitoramento de impacto deve levar em consideração. O que não quer dizer que não devamos tentar.

---

[284] https://towardsdatascience.com/the-evolution-of-data-journalism-1e4c2802bc3d.

Paul Bradshaw é responsável pelos programas de mestrado em Jornalismo de Dados e em Jornalismo Multiplataformas e Móvel da Universidade da Cidade de Birmingham, além de trabalhar como consultor em jornalismo de dados junto à Unidade de Dados da BBC da Inglaterra, escrever livros e treinar jornalistas em diversos veículos de comunicação.

Reflexões

Jornalismo de dados: ao interesse de quem?

Mary Lynn Young e Candis Callison

Uma das primeiras contribuições significativas em jornalismo de dados nos EUA foi o site *chicagocrime.org*, um mapa online de Chicago com estatísticas sobre crimes (Anderson, 2018; Holovaty, 2005 e 2008). De acordo com seu fundador, Adrian Holovaty, o site lançado em 2005 foi um dos "primeiros mapas mistos, combinando informações do Departamento de Polícia de Chicago com o Google Maps. Ele contava com uma página e feed RSS para cada quarteirão de Chicago e múltiplas formas de navegar dados sobre crimes: por tipo, tipo de local (calçada ou apartamento, por exemplo), por CEP, por endereço, data, ou mesmo seguindo uma rota arbitrária" (Holovaty, 2008).[285] Alguns anos depois, o *Los Angeles Times* criou o blog jornalístico *Homicide Report*, que se baseava em dados da polícia para criar postagens a respeito dos mais de 900 homicídios ocorridos na região. Ambos os projetos usavam informações sobre crimes e dados geográficos em grandes centros metropolitanos dos EUA, fornecendo percepções sobre críticas e desafios constantes relacionados aos objetivos e impactos do jornalismo baseado em dados e do jornalismo em geral.

Os motivos para Holovaty lançar o endereço *chicagocrime.org* estavam relacionados aos objetivos do jornalismo de gerar "notícias úteis" junto de sua identidade cada vez mais técnica e foco em "coisas técnicas legais" (Holovaty, 2005). Já os objetivos da fundadora do *Homicide Report*, Jill Leovy, do *LA Times*, eram mais críticos. Leovy queria ter um registro de todos os homicídios em Los Angeles de forma a desconstruir normas e práticas de jornalismo tradicional, que cobria apenas alguns casos de homicídio (Leovy, 2008; Young e Hermida, 2015).[286] Durante entrevista concedida em 2015 ao programa *Fresh Air* da *National Public Radio*, Leovy falou sobre seus motivos para lançar o *Homicide Report* como uma espécie de resposta ao viés estrutural presente no noticiário e sua frustração a respeito de como reportagens sobre o crime "não chegavam nem perto da realidade":

> O papel do jornal é cobrir eventos fora do comum, e quando falamos de homicídios, isso significa que nos prendemos à parte mais baixa de uma curva. Nunca nem chegamos perto do meio, porque ali estão os homicídios de rotina, por mais que, claro, um homicídio nunca seja algo rotineiro. Estes homicídios

[285] Uma iteração prévia do elogiado site de jornalismo de dados comunitários *EveryBlock*, lançado por Holovaty em 2008 e comprado pela *MSNBC.com* em 2009 (Holovaty, 2013).

[286] Mais tarde foi recriado como um blog voltado ao jornalismo algorítmico.

> vêm ocorrendo no mesmo formato, dos mesmos jeitos e há tanto tempo na América, especialmente em cidades americanas, que acabaram se tornando o papel de parede da vida urbana. São parte do cenário e é muito difícil extrair deles uma narrativa, uma história que funcione no contexto de um jornal.

Ao combinar sua experiência como jornalista policial com o espaço infinito e menos hierárquico do jornalismo digital (em comparação à primeira página do jornal) e acesso a dados públicos, Leovy imaginou uma reportagem que representasse informações sobre todos os homicídios na região, "em grande parte de jovens latinos e, de maneira desproporcional, jovens negros" (Leovy, 2008), com o máximo de equivalência possível (Leovy, 2015). De acordo com a jornalista, a reação foi grande: "A resposta do público foi forte. 'Meu Deus', iniciava uma das primeiras postagens de um leitor. 'O volume é chocante', escreveu outro. 'É quase como se fossem pessoas dispensáveis', disse um terceiro" (Leovy, 2008).

Como articulações novas de uma subespecialidade em desenvolvimento, estes exemplos de jornalismo de dados foram aclamados por sua inovação. O site *chicagocrime.org*, de acordo com Holovaty (2008), foi parte de uma exibição no Museu de Arte Moderna de Nova York. Agora, questionamentos sobre a quem interessava ou qual era o público imaginado para estes projetos e outros que alegam compartilhar dados em prol dos interesses de um bem público não foram feitos. Acadêmicos do campo de estudos em ciência e tecnologia demonstraram repetidas vezes o quão danosas as relações entre populações vulneráveis e certos tipos de dados podem ser e persistem nesse sentido mesmo quando a tecnologia é anunciada como nova e transformadora (Nelson, 2016; Reardon, 2009; TallBear, 2013). A orientação positivista do jornalismo de dados (Coddington, 2019) é implicada também, apesar de críticas extensas sobre a construção social de raça e o papel da tecnologia na replicação da supremacia branca (Benjamin, 2019; Noble, 2018; Townsley, 2007). Além do que, estudos sobre representações jornalísticas, normas, práticas, economia e noticiário criminal indicam uma longa história de racialização, controle social, danos e colonialismo contínuo (Anderson e Robertson, 2011; Callison e Young, 2020; Ericson, Baranek e Chan, 1991; Hamilton, 2000; Schudson, 2005; Stabile, 2006). Este capítulo explora, de maneira breve, quais estruturas seguem sendo apoiadas e quais dados estão mais sujeitos à coleta — ou não — ao passo que levanta questões sobre a necessidade dos jornalistas de incorporar uma espécie de 'ética da recusa' ao decidir se e como empregar práticas em jornalismo de dados (Simpson, 2007; Tallbear, 2013). Como argumentado por Judith Butler, "há maneiras de se distribuir vulnerabilidade, formas diferenciais de alocação que tornam algumas populações mais sujeitas à violência arbitrária que outras" (Butler, 2004, p. xii).

Nos baseamos em Coddington (2015 e 2019) para definir jornalismo de dados como jornalismo quantitativo que consiste em três formas: jornalismo de dados, jornalismo computacional e reportagem assistida por computador. Críticas persistentes a práticas

científicas e instituições sociais que unem justificativas para pesquisa e coleta de dados, novas tecnologias e questões sociais mostram-se relevantes para os três tipos de jornalismo de dados, assim como questões de vulnerabilidade e outras relacionadas aos interesses de quem estariam sendo suportados. A acadêmica de estudos em ciência e tecnologia e estudos indígenas Kim TallBear trabalhou com pesquisa genômica entre populações indígenas nos EUA e descobriu que muitos dos estereótipos e narrativas coloniais associados com a ideia de que 'índios estão desaparecendo' faziam parte de justificativas para pesquisa junto a declarações de identidade potencial (ou seja, conhecimento sobre migração e ligações ancestrais) e benefícios de saúde.

> Ao passo que a ideia de conexão genética talvez possa ter substituído a de hierarquia racial no léxico da ciência corrente, relações de poder, diferença e hierarquia continuam integrando nossa cultura em termos mais amplos, nossas instituições e estruturas, e a cultura em que a ciência se dá e que a ciência ajuda a produzir (TallBear, 2007, p. 413).

> O que TallBear argumenta é que reforçar ideias científicas de objetividade maquínica ou baseada em laboratório faz parte de uma série de prerrogativas institucionais, relações históricas e continuadas com comunidades, e estruturas culturais que guiam justificativas para pesquisa e articulação de benefícios desejados. Deve-se sempre questionar a interesse de quem e por que. E, em alguns casos, análises de pesquisa e/ou mineração de dados podem ser recusadas, pois processos de significação são predicados com base em ideias enraizadas de raça, gênero e classe. O que TallBear chama de "suposições e práticas coloniais que continuam a informar a ciência", diríamos que fazem o mesmo com o jornalismo e, por extensão, o jornalismo de dados (TallBear, 2007, p. 423).

Povos indígenas também tiveram de lidar com extensos arquivos antropológicos e governamentais, assim como representações consistentemente erradas e estereotipadas na mídia (Anderson e Robertson, 2011), a serviço de variadas formas e histórias de colonialismo (Wolfe, 2006; Tuck e Yang, 2012). Sendo assim, as implicações para o jornalismo de dados, especificamente, como extensão de ideias de objetividade baseadas em máquina são profundas. Na crítica à antropologia de Audra Simpson (2014), ela sugere que comunidades Mohawk regularmente engajam em formas de recusa quando se trata de lidar com tais arquivos e participação em pesquisas centradas em instituições e estruturas colonialistas. A recusa na estrutura de Simpson é multidirecional: recusa em ser eliminado, recusa em internalizar as representações erradas de sua identidade, cultura e terras, e uma recusa em conformar-se com as expectativas de diferença que o estado ou outra forma de reconhecimento (neste caso, mídia) faz sobre você ou seu grupo.

Tais argumentos feitos por acadêmicos indígenas oferecem desafios diretos às intenções, justificativas e práticas de jornalismo de dados centradas em questões de história e poder. Estas questões estão ligadas não somente ao estado, mas também ao papel do jornalismo na manutenção de ordens sociais que apoiam objetivos do estado, assim como estruturas e ideologias como patriarcado, colonialismo e supremacia branca (Callison e Young, 2020).

Uma complicação ainda mais profunda para comunidades indígenas é que dados e representações midiáticas precisas quase sempre são difíceis de se encontrar, junto do fato de que os dados refletem os contextos institucionais em que estas informações são coletadas, arquivadas e acessadas.[287] A crítica de Ericson às estatísticas da polícia por não refletirem a realidade social do crime e, sim, "construtos culturais, legais e sociais produzidos... para fins organizacionais" (1998, p. 93) é relevante para jornalistas focados somente em processamento de dados. A jornalista do *Laguna Pueblo* Jenni Monet, por exemplo, caracteriza comunidades indígenas nos EUA como "nações asterisco", aquelas para as quais não existem dados (2018a). Especialmente no Alaska, muitos gráficos de dados sociais contam com um asterisco indicando que não há informações para nativos da região. Mídias digitais como o Facebook oferecem plataformas alternativas promissoras que podem ser encaradas como ferramentas para que o jornalismo possa interagir com públicos indígenas e suas questões de forma a criar representações significativas e precisas, que abordem desigualdades estruturais e a própria falta de dados (Monet, 2018b). Mais uma vez, a questão de participar ou não e como participar se dá em torno de quem está se beneficiando, quais os processos utilizados para coleta de dados e quais processos de significação prevalecem.

Para o jornalismo, em termos amplos, processos de significação muitas vezes estão ligados a questões de dissidência, desvio, conflito ou "o comportamento de uma coisa ou pessoa que se afasta da norma" (Ericson, 1998, p. 84) dentro de uma orientação positivista. O papel do jornalismo no ordenamento social teve e continua a ter impactos materiais e efeitos danosos em populações construídas como desviantes (Callison e Young, 2020; Rhodes, 1993; Stabile, 2006). Stabile, em seu estudo histórico do noticiário criminal nos EUA — que inclui jornais, cobertura televisiva do tema e programas de rádio — e sua relação com raça, articula o impacto das normas de desvio e "a implantação de informação cultural sobre índices de crime" (2006, p. 1) em populações estruturalmente vulneráveis dentro de um contexto ideológico de supremacia branca e jornalismo com fins lucrativos. Ela foca em raça e gênero, pois "estão entre os pontos focais de disputa em torno do significado histórico atribuído a desvio" (2006, p. 5), argumentando que a mídia suporta "o processo de criminalização" de homens negros pelo estado e seus agentes, dentre os quais a polícia (2006, p. 3). Um exemplo é como a mídia amplia e reforça práticas de coleta de dados por parte da polícia ao focar em

[287] Para mais informações sobre o assunto, ver o capítulo de Kukutai e Walter neste mesmo volume.

tipos específicos de crime, como roubo de carros, infratores e vítimas. Ela se depara com uma "sociedade branca gananciosa e violenta que prosperou nos EUA, onde ficções de terror branco consistentemente deslocam a materialidade do terrorismo branco" (Stabile, 2006, p. 2). Aqui, a análise de Carey de 1974 do jornalismo como gerador de inimigos e aliados pode ser entendida como igualmente relevante para as relações da profissão com o capitalismo e o estado na América do Norte, o que inclui genocídio estatal e colonialismo continuado. Somando isso à alergia do jornalismo à ideia de que fatos e conhecimentos são construídos socialmente, jornalismo e em especial as notícias tornam-se a fáscia pela qual discursos de ordenamento social têm sido e continuam sendo cogerados, replicados e potencialmente transformados.

Sobre estas questões tão críticas, a literatura nos campos de jornalismo, criminologia, estudos de ciência e tecnologia, dentre outras disciplinas, levanta sérias preocupações que até então foram pouco abordadas no contexto do jornalismo de dados. Estudiosos gastaram um bom tempo com tipologias (Coddington, 2019), o estado do jornalismo de dados (Heravi, 2017) e seus efeitos em epistemologias, culturas, práticas e identidades jornalísticas mais amplas (Anderson, 2018; Borges-Rey, 2017; De Maeyer et al., 2015; Gynnild, 2014; Lewis e Usher, 2014; Parasie e Dagaril, 2013; Young e Hermida, 2015) — mais do que com seus efeitos e consequências mais abrangentes. Poucos acadêmicos discutem questões relacionadas a poder, caso da pesquisa de Borges-Rey (2016 e 2017) que integra uma análise de economia política do crescimento do jornalismo de dados no Reino Unido.

O jornalismo de dados pode levar a alguns impactos, como nesta declaração de Holovaty:

> Muito de bom aconteceu após a criação do *chicagocrime.org*. A nível local, incontáveis cidadãos de Chicago entraram em contato comigo para expressar sua gratidão pelo serviço público prestado. Grupos comunitários levaram páginas impressas do site para suas discussões sobre a polícia, e alguns cidadãos mais emocionados levaram os relatos do site aos seus representantes para apontar cruzamentos problemáticos em que a cidade poderia instalar postes com luzes mais brilhantes (Holovaty, 2008).

Neste caso, grupos comunitários pegaram dados e criaram seus próprios significados e justificativas para ação. Mas a forma como isso funciona em maior escala, em áreas rurais distantes de centros de poder e mídia, em comunidades que já sofrem vigilância desproporcional e em casos nos quais comunidades não estão bem representadas dentro de redações, predominantemente brancas no Canadá e nos EUA, demanda um conjunto maior de

diagnósticos éticos. Considerando estes exemplos e a evidência baseada em literatura crítica fora do campo jornalístico, dano em potencial pode e deve ter prioridade perante normas do setor como o uso de "notícias úteis" e experimentações movidas por tecnologia. A forma como jornalistas cobrem notícias sobre crimes de um ponto de vista de dados exige entendimento profundo das consequências e problemas de se levar em conta somente as intenções internas do jornalismo, evidências de sucesso e justificativas para inovação.[288] Diagnósticos éticos devem melhor embasar a ideia da recusa, longos históricos de representação errônea e serviços prestados ao colonialismo por parte do jornalismo, bem como os processos desiguais pelos quais significação e coleta de dados ocorrem. 'A interesse de quem?' e 'por que?' se tornaram perguntas essenciais para jornalistas ao considerar como, quando, onde e para quem o jornalismo de dados faz qualquer tipo de contribuição.

Mary Lynn Young é professora adjunta da Escola de Jornalismo da Universidade da Colúmbia Britânica e cofundadora da The Conversation Canada, organização jornalística nacional sem fins lucrativos afiliada da rede global The Conversation. Candis Callison é professora adjunta da Escola de Jornalismo da Universidade da Colúmbia Britânica e autora do livro "How Climate Change Comes to Matter: The Communal Life of Facts".

Referências

ANDERSON, Chris. *Apostles of Certainty: Data Journalism and the Politics of Doubt*. Oxford: Oxford University Press, 2018.

ANDERSON, Mark C.; ROBERTSON, Carmen. *Seeing Red: A History of Natives in Canadian Newspapers.* Winnipeg: University of Manitoba Press, 2011.

BARTHEL, Michael. *In the news industry, diversity is lowest at smaller outlets*. Pew Research, 4 de agosto de 2015.

BENJAMIN, Ruha. *Race after technology: Abolitionist Tools for the new Jim Code.* Cambridge: Polity Press, 2019.

BORGES-REY, Eddy. *Towards na epistemology of data journalism in the devolved nations of the United Kingdom: Changes and continuities in materiality, performativity and reflexivity*. Journalism, 2017, p. 1-17.

BORGES-REY, Eddy. *Unravelling data journalism: A study of data journalism practice in British newsrooms*. Journalism Practice 10 (7), 2016, p. 833–843.

[288] Ver o capítulo assinado por Loosen neste volume.

BUTLER, Judith. *Precarious Life: The Powers of Mourning and Violence.* Nova York: Verso, 2004.

CALLISON, Candis; YOUNG, Mary L. *Reckoning: Journalism's Limits and Possibilities*. NovaYork: Oxford University Press, 2020.

CAREY, James W. *The Problem of Journalism History*. Journalism History, 1 (1), 1974, p. 3-27.

CODDINGTON, Mark. *Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting*. Digital Journalism 3 (3), 2015, p. 331-348.

CODDINGTON, Mark. *Defining and mapping data journalism and computational journalism: A review of typologies and themes*. In: ELDRIDGE, Scott II; FRANKLIN, Bob (ed.). *The Routledge Handbook of Developments in Digital Journalism Studies*. Abingdon: Routledge, 2019.

ERICSON, Richard. *How Journalists Visualize Fact*. The Annals of the American Academy of Political and Social Science, Vol. 560. The Future of Fact, novembro de 1998, p. 83-95.

ERICSON, Richard; BARANEK, Patricia; CHAN, Janet. *Representing Order: Crime, Law and Justice in the News Media*. Toronto: University of Toronto Press, 1991.

GERTZ, Matt. *Stagnant American Newsroom Diversity In Charts*. Media Matters for America. 25 de junho de 2013. Disponível em: https://www.mediamatters.org/blog/2013/06/25/stagnant-american-newsroom-diversity-in-charts/194597.

GYNNILD, Astrid. *Journalism innovation leads to innovation journalism: The impact of computational exploration on changing mindsets*. Journalism 15 (6), 2014, p. 713-730.

HAMILTON, James T. *Channeling Violence: The Economic Market for Violent Television Programming*. Princeton: Princeton University Press, 2000.

HERAVI, Bahareh. *State of Data Journalism Globally: First Insights into the Global Data Journalism Survey*. Medium, 1º de agosto de 2017. Disponível em: https://medium.com/@Bahareh/state-of-data-journalism-globally-cb2f4696ad3d.

HOLOVATY, Adrian. *RIP EveryBlock*. Blog de Adrian Holovaty, 7 de fevereiro de 2013. Disponível em: http://www.holovaty.com/writing/rip-everyblock/.

HOLOVATY, Adrian. *In Memory of chicagocrime.org*. Blog de Adrian Holovaty, 31 de janeiro de 2008. Disponível em: http://www.holovaty.com/writing/chicagocrime.org-tribute/.

HOLOVATY, Adrian. *Announcing chicagocrime.org*. Blog de Adrian Holovaty, 18 de maio de 2005. Disponível em: http://www.holovaty.com/writing/chicagocrime.org-launch/.

LEWIS, Seth; USHER, Nikki. *Code, collaboration, and the future of journalism: A case study of the Hacks/Hackers global network*. Digital Journalism 2 (3), 2014, p. 383-393.

LEOVY, Jill. *On the Air interview with Dave Davies*. NPR, 26 de janeiro de 2015. Disponível em: https://www.npr.org/2015/01/26/381589023/ghettoside-explores-why-murders-are-invisible-in-los-angeles.

LEOVY, Jill. *Unlimited Space for Untold Sorrow*. Los Angeles Times, 4 de fevereiro de 2008. Disponível em: http://articles.latimes.com/2008/feb/04/local/me-homicide4.

MONET, Jenni. *A Conversation with Jenni Monet*. Women's Media Centre, 14 de março de 2018. Disponível em: http://www.womensmediacenter.com/news-features/a-conversation-with-jenni-monet.

MONET, Jenni. *#DeleteFacebook? Not in Indian Country*. Yes Magazine, 23 de março de 2018. Disponível em: https://www.yesmagazine.org/peace-justice/deletefacebook-not-in-indian-country-20180323.

NELSON, Alondra. *The Social Life of DNA: Race, Reconciliation, and Reparations after the Genome.* Boston: Beacon Press, 2016.

NOBLE, Safiya. *Algorithms of Oppression: How Search Engines Reinforce Racism* Nova York: NYU Press, 2018.

REARDON, Jenny. *Race to the Finish: Identity and Governance in an Age of Genomics*. Princeton: Princeton University Press, 2009.

RHODES, Jane. *The Visibility of Race and Media History*. Critical Studies in Mass Communication 10 (2), 1993, p. 184-190.

STABILE, Carol. *White Victims, Black Villains: Gender, Race and Crime News in US Culture.* Nova York: Routledge, 2006.

TALLBEAR, Kim. *The Emergence, Politics and Marketplace of Native American DNA*. In: LEE KLEINMAN, Daniel; MOORE, Kelly (ed.). *The Routledge Handbook of Science, Technology, and Society*. Londres: Routledge, 2014, p. 21-37.

TALLBEAR, Kim. *Native American DNA: Tribal Belonging and the False Promise of Genetic Science.* Minneapolis: University of Minnesota Press, 2013.

TALLBEAR, Kim. *Narratives of Race and Indigeneity in the Genographic Project*. Journal of Law, Medicine e Ethics*,* 35(3), 2007, p. 412–424.

TUCK, Eve; YANG K. Wayne. *Decolonization is not a metaphor*. Decolonization: Indigeneity, education e society 1(1), 2012.

YOUNG, Mary L.; HERMIDA Alfred. *From Mr. and Mrs. outlier to central tendencies: Computational journalism and crime reporting at the Los Angeles Times*. Digital Journalism 3 (3), 2015, p. 381-397.

WOLFE, Patrick. *Settler Colonialism and the Elimination of the Native*. Journal of genocide research 8 (4), 2006, p. 387-409.

Para que serve o jornalismo de dados? Dinheiro, cliques, tentativa e erro

Nikki Usher

Ficar dando F5 diariamente nas previsões do mapa eleitoral interativo do *Five Thirty Eight* de 2016 era um ritual entre os pares em Washington, de políticos a jornalistas, de estudantes a servidores públicos e mais. Alguns favoreciam o agregador de pesquisas *Upshot* do *New York Times*; outros; voltados a probabilidades, recorriam ao *Real Clear Politics*; já o pessoal de gostos mais exóticos acompanhava a cobertura feita pelo *The Guardian*. Para estes viciados em F5, tudo estava e continuaria bem no mundo contanto que as chances estivessem favorecendo Hillary Clinton na corrida à presidência dos EUA, uma espécie de *Jogos Vorazes* eleitorais: quanto maior a margem, melhor.

Sabemos como essa história acaba, com o mapa de Nate Silver indicando uma vantagem de 71,4% para Hillary Clinton mesmo no dia da votação. Talvez já seja hora de superarmos a eleição dos EUA de 2016, afinal, a obsessão por mapas eleitorais talvez seja um passatempo particularmente americano, considerando o ciclo regular das eleições no país — além, claro, de um público global que também fica de olho (Lewis e Waters, 2017). Mas, até que o mapa saia do ar, segue ali, assombrando jornalistas e apoiadores de Clinton, e dando munição aos republicanos para lembrar que a "mídia mainstream boboca" não passa de "fake news". Política à parte, a eleição dos EUA em 2016 não deve ser esquecida pelos jornalistas de dados. Mesmo que a quantificação estivesse correta, e todos acreditavam que sim, erros de mapeamento e visualização acabam por ser mais uma ferramenta pelas quais pode-se desmantelar a reinvindicação jornalística da autoridade epistêmica (ou, em termos mais simples, lá se vai o argumento de autoridade).

Sim, é injusto tratar jornalismo de dados como previsão eleitoral — certamente, é bem mais que isso, especialmente de uma posição privilegiada global, mas por vezes parece que essas são as principais contribuições da prática: mapas sem fim, gráficos clicáveis e calculadoras sujeitos a erro por parte do usuário, simplificação exacerbada e marginalização, independentemente do rigor da computação e da habilidade estatística que gerou tais produtos. Com a segunda edição deste guia em suas mãos, podemos declarar que o jornalismo de dados chegou a um ponto de amadurecimento e autorreflexão, e, de tal forma, é importante perguntar "para que serve o jornalismo de dados?".

A prática, em seu estágio atual, apenas sugere um potencial para moldar e reacender a chama do jornalismo. A primeira edição deste guia teve início como um projeto colaborativo,

no contexto de um grande grupo, em 2011, durante o Mozilla Festival, um esforço que observei e duvidei que chegaria a algum resultado tangível (claramente errei). Esta segunda edição agora é publicada pela University of Amsterdam Press e distribuída nos EUA pela University of Chicago Press, contando com a presença de colaboradores convidados, o que sugere que a natureza independente do jornalismo de dados deu lugar, de certa forma, a uma espécie de profissionalismo, ordem e legitimidade. Veja bem, é este mesmo o caso: o jornalismo de dados chegou no mainstream, é ensinado em cursos de jornalismo e foi normalizado na redação (Usher, 2016). Ele também foi padronizado; sendo assim, pouca coisa mudou nos últimos cinco a sete anos. Análises de concursos nacionais do setor mostram pouca inovação na forma e nos temas (política, em grande parte), com mapas e gráficos ainda sendo os recursos mais utilizados (Loosen et al., 2017). A interatividade se limita àquilo que são consideradas "técnicas de entrada" para os envolvidos com visualização de informações (Young, Hermida e Fulda, 2017). Além do que, o jornalismo não avançou o suficiente para visualizar "gráficos dinâmicos, dirigidos e ponderados" (Niederer et al., 2015). Jornalistas de dados ainda lidam com dados pré-processados e não "big data", que são "grandinhos", na melhor das hipóteses, dados governamentais e não dados multinível em termos de profundidade e tamanho, do tipo que um provedor de internet pode coletar.

A crítica que ofereço vem em grande parte de uma perspectiva ocidental, se não norte-americana, mas isso não atrapalha a chamada a qual destaco: jornalistas de dados estão sentados em cima de uma caixa de ferramentas possivelmente revolucionária que ainda não mostrou a que veio. A revolução, porém, se mal executada, só servirá para atrapalhar os esforços voltados à experiência de usuário e à busca de conhecimento de consumidores de notícias e, no pior dos casos, espalhar mais desconfiança em relação às notícias. Se o jornalismo de dados continuar do mesmo jeito destes últimos cinco a dez anos, então pouco ele faz para levar adiante a causa jornalística em uma era digital e de plataformas. Logo, para começar a perguntar esta questão existencial, "para que serve o jornalismo de dados?", proponho que jornalistas do setor — assim como jornalistas menos voltados a dados, mas inseridos no contexto da web e que trabalhem com vídeo, áudio, e programação —, bem como estudiosos do tema, precisam repensar a história de origem do jornalismo de dados, seu jeito de pensar atual e seu futuro.

Jornalismo de dados nos EUA: o início

A história de origem é aquela que nos contamos sobre como e por que viemos a ser assim, muitas vezes vista por lentes que não correspondem bem à realidade, acompanhadas de um tanto de gabolice. Nos EUA, o jornalismo de dados começou mais ou menos assim: no mundo primordial, pré-jornalismo de dados, a prática já existia de outra maneira, conhecida como reportagem assistida por computador, ou ao menos era chamada assim nos EUA, uma oportunidade para levar o rigor das ciências sociais para o jornalismo.

No mito da introdução do jornalismo de dados à web, seus praticantes se tornam versões turbinadas de jornalistas investigativos dotados de habilidades computacionais superiores do século XXI, libertadores de dados (ou documentos) com o objetivo de contar histórias que não seriam contadas de outra forma. Mas, além da investigação de histórias, jornalistas de dados conseguiram salvar o jornalismo com suas novas habilidades web, adicionando um novo nível de transparência, personalização e interatividade a notícias que quem consome notícias apreciaria, aprenderia e, claro, clicaria. A internet e suas notícias de outrora nunca mais seriam as mesmas. O jornalismo de dados corrigiria erros e daria a base objetiva necessária que as avaliações qualitativas do ofício não possuíam, fazendo isso em escala e com uma maestria inimaginável antes do ambiente digital interativo em tempo real repleto de servidores poderosos em nuvem que tiram a pressão computacional de qualquer veículo de comunicação. Sinais iniciais de sucesso guiariam o caminho adiante e tornariam leitores comuns em colaboradores investigativos ou cientistas cidadãos, caso da cobertura do *The Guardian* sobre o escândalo com o Parlamento ou o projeto *Cicada* do *WNYC*, que fez com que um pequeno exército de residentes de Nova York construíssem termômetros de solo para ajudar a mapear a chegada dos terríveis instintos de verão. E esta inspirada orquestração envolvendo jornalismo, computação, multidões, dados e tecnologia seguiria adiante, fazendo justiça à verdade.

O presente: de "jornalista hacker" a mais um funcionário (chato) da redação

Os dias de hoje não estão tão distantes da história de origem que os jornalistas de dados atuais contaram a si mesmos, nem em questão de visão ou realidade. Emergiram dois tipos diferentes de jornalismo de dados: o "investigativo", que leva adiante o nobre manto dos esforços jornalísticos; e o cotidiano, que pode ser otimizado para fins virais, que pode ser qualquer coisa, desde cartografia jornalística feita a toque de caixa ou transformar uma enquete de opinião pública ou pesquisa em um meme de fácil compartilhamento com um verniz jornalístico aplicado. Na melhor das hipóteses, o jornalismo de dados ficou chato e profissional demais. Na pior, tornou-se só mais uma estratégia para geração de receita online.

Não é exagero dizer que a prática poderia ter transformado o jornalismo como o conhecemos, mas não fez isso até então. Em 2011, durante o MozFest, um hack de destaque no festival foi uma espécie de plugin que permitiria a qualquer um colocar seu rosto em destaque numa simulação de página inicial do *Boston Globe*. Tudo muito divertido, mas com certeza o veículo não permitiria que conteúdo gerado pelo usuário, sem qualquer tipo de triagem, fosse parar em sua página. De maneira semelhante, durante o nascimento do primeiro Manual de Jornalismo de Dados, o jornalista de dados era o "jornalista hacker", imaginado como alguém que vinha do setor de tecnologia para o jornalismo ou que, no mínimo, carregasse consigo um espírito de código aberto, hacker, de forma a inspirar projetos que jogavam para escanteio processos convencionais de jornalismo institucional, dando

espaço para experimentação, imperfeição e diversão — uma vontade de fuçar nas coisas em busca de algo que talvez não fosse excelente na forma ou no conteúdo, mas de que de algum jeito "hackeasse" o jornalismo (Lewis e Usher, 2013). Em 2011, se tratava de gente de fora entrando no jornalismo, agora em 2018 temos gente de dentro profissionalizando programação no jornalismo. O espírito de inovação, invenção, assumiu uma forma corporativa, de negócios, certamente bem menos divertida (Lewis e Usher, 2016).

Tudo bem ser chato, já que isso também cumpre um papel. Parte da profissionalização do jornalismo de dados foi justificada através de coisas como a percepção que os próprios jornalistas têm, ideias como a do "jornalista de dados herói", gente que graças a um conjunto diferente de valores (como colaboração e transparência) e habilidades (visualização e habilidades computacionais variadas) poderia fazer a verdade cantar mais alto de novas formas. Os casos dos *Panama* e *Paradise Papers* podem ser encarados como algumas das melhores expressões desta visão. Mas, o jornalismo de dados investigativo demanda tempo, esforço e conhecimento especializado que vão além do processamento de dados, incluindo muitas outras fontes de dados mais tradicionais, como entrevistas, reportagem de campo e documentos. De ocorrência regular, jornalismo investigativo inovador é um conceito controverso, não por falta de dedicação — a Rede Europeia de Jornalismo de Dados, o Instituto de Notícias Sem Fins Lucrativos dos EUA e a Rede Global de Jornalismo Investigativo revelam uma ampla rede de esforços investigativos. A verdade é que uma investigação que seja de fato avassaladora é raridade, por isso conseguimos enumerar com os dedos das mãos casos de enorme sucesso. O escândalo do Parlamento do *The Guardian*, feito com base em colaborações voluntárias em 2010, segue sem paralelo até então.

O passado é preâmbulo quando se fala de jornalismo de dados. *Snow Fall*, projeto narrativo imersivo revolucionário do *New York Times*, vencedor de um Pulitzer em 2012, retornou em dezembro de 2017 na forma dos artigos *Deliverance from 27,000 Feet* e *Everest*. Cinco anos depois, o *New York Times* publicou um texto longo sobre desastre ocorrido em uma montanha nevada, mas outra montanha desta vez (ainda que o mesmo autor, John Branch, assinasse a peça). Nestes cinco anos, "Snowfall" ou "Snowfalled" eram termos usados internamente pelo pessoal do jornal — e fora também — para essa ideia de dar um *tchan* interativo a qualquer matéria. Depois de 2012, teve início um debate não só no *Times*, mas em outras redações dos EUA e do Reino Unido, sobre se jornalistas de dados deveriam gastar seu tempo com ferramentas pré-desenvolvidas que poderiam dar esse toque interativo a qualquer história automaticamente ou se deveriam trabalhar em projetos únicos inovadores (Usher, 2016). Enquanto isso, o projeto *Snow Fall* original, que oferecia o mínimo de interatividade em 2012, seguiu minimamente interativo, na melhor das hipóteses, até 2017.

"Mas espere!", diria o jornalista de dados de outrora, "*Snow Fall* não é um exemplo de *jornalismo de dados,* talvez um truque metido à besta feito por desenvolvedores de apps

de notícias, mas não tem nada de dados ali!" Eis o problema: talvez jornalistas de dados não encarem *Snow Fall* como jornalismo de dados, mas por que não? *O que é jornalismo de dados senão contar histórias de novas formas com novas habilidades que se aproveitam do que há de melhor na web?*

O jornalismo de dados não pode se voltar a mapas e gráficos somente, muito menos o mapeamento de dados ou sua organização em gráficos concede ao jornalismo de dados superioridade intelectual em relação a empreitadas de jornalismo digital imersivo. O que pode ser mapeado é mapeado. Mapeamento de eleições dos EUA à parte, as consequências éticas de quantificar e visualizar os dados mais recentes disponíveis em formato clicável e coerente precisam, sim, de alguma crítica. Em sua forma mais corriqueira, o jornalismo de dados é como a saladinha que acompanha um prato, falando de visualização. Isso se aplica ainda mais considerando o movimento em prol de projetos diários de jornalismo de dados. Talvez novas estatísticas de trabalho, dados sobre bicicletas pela cidade, índices de reciclagem, resultados de um estudo acadêmico, criação de visualizações porque tem como visualizar aquilo ali (com sorte, conseguem-se alguns cliques). Na pior das hipóteses, o jornalismo de dados pode simplificar demais algo a ponto de desumanizar o tema ao qual busca lançar alguma luz. Mapas de migrantes e seus fluxos pela Europa assumem a forma de setas interativas ou ícones sem gênero, como argumentado pelo geógrafo humano Paul Adams. A cartografia em notícias digitais transformou a crise dos refugiados em uma série desencarnada de ações clicáveis, o oposto do que poderia fazer, enquanto jornalismo, para tornar "refugiados" desconhecidos em figuras dignas de empatia, para além dos números (Adams, 2017). Antes de mapear qualquer outro problema ou estudo acadêmico, jornalistas de dados precisam se perguntar: para que estamos mapeando e criando gráficos (e artigos junto a isso tudo)?

O problema está em algum ponto entre *Snow Fall* e mapas de migração: *para que serve o jornalismo de dados*? O presente nos dá muitas provas de profissionalização e isomorfismo, com um toque de incentivo corporativo para que o jornalismo de dados não sirva só para ajudar consumidores de notícias em seu processo de entendimento do mundo, mas também para garantir o sustento de veículos e organizações de comunicação. Claro que isso não é tudo que o jornalismo de dados pode ser.

O futuro: como o jornalismo de dados pode retomar seu valor (e ser divertido também)

Para que serve o jornalismo de dados? A prática precisa voltar às raízes de mudança e revolução, de experimentação e inspiração, de uma visão autodeterminada de renegados em meio a uma indústria cansada e nada inspirada para forçar jornalistas a confrontarem sua suposta autoridade sobre conhecimento, narrativa e distribuição. Jornalistas de dados

precisam lembrar de sua inspiração hacker e hackear a redação, como já prometeram anteriormente; precisam ir além do foco no lucro e no profissionalismo de suas redações. Retomar o status de outsider nos aproximará daquilo que o jornalismo de dados tinha como oferta primordial: uma forma diferente de pensar jornalismo, um jeito diferente de apresentar o jornalismo, e um modo de trazer novos pensadores e atores para dentro das redações. E, além disso tudo, uma maneira de revigorar a prática jornalística.

No futuro, consigo imaginar o jornalismo de dados livre do termo "dados", focado somente na palavra "jornalismo". Presume-se que jornalistas de dados têm habilidades que o resto da redação ou outros jornalistas não têm, como a capacidade de compreender dados complicados (ou guiar um computador para que faça isso por eles) e de visualizar dados de forma significativa e a habilidade em programação. O jornalismo de dados, porém, precisa se tornar aquilo que chamei de jornalismo interativo, precisa deixar de lado seu impulso de trabalhar com mapas e gráficos, bem como seu desdém por tecnologias e outras habilidades que não focam em dados, como vídeos em 360 graus, realidade aumentada e animação. Em minha visão do futuro, teremos muitos mais produtos interativos como *Secret Life of the Cat,* da *BBC*, e *Dialect Quizzes*, do *New York Times*; e mais projetos que combinem vídeos em 360 graus ou realidade virtual com dados, caso da empreitada do Dataverse financiada pela iniciativa de jornalismo imersivo da *Journalism 360*. Teremos muito menos mapeamento de eleições e cartografia para ilustrar notícias cotidianas, daquelas que reduzem fatalidades a fluxos e linhas clicáveis. Com sorte, veremos o fim da tendência de produtos interativos com enquetes em tempo real, um fetiche dos grandes veículos de notícias dos EUA. No lugar disso tudo, teremos muito mais originalidade, diversão e uma ruptura inspirada com o que se espera do jornalismo em forma, ação e conteúdo. Jornalismo de dados trata de transparência, mas também de diversão e imaginação; ganha força não só porque algum parlamentar abriu mão do cargo ou porque deixou uma tendência ainda mais evidente, mas também porque pessoas comuns conseguem perceber o valor de se voltar a veículos e jornalistas porque estes jornalistas atendem a diversas necessidades de informação: orientações, entretenimento, comunidades e mais.

E para poderem falar de conhecimento superior sobre dados, jornalistas de dados precisam transformar estas informações em conhecimento e coletar estes dados por conta própria. Não é uma questão de gerar visualizações a respeito de informações reunidas por terceiros. Na melhor das hipóteses, trabalhar em cima de dados de terceiros dá visibilidade às alegações de quem os forneceu; já na pior, o jornalismo de dados não passa de uma espécie de assessoria de imprensa para o fornecedor das informações. Dito isso, muitos jornalistas continuam não tendo muito interesse em coletar seus próprios dados e creem que isso não é parte de sua função. Como explicado pelo editor de dados do *Washington Post*, Steven Rich, em postagem no Twitter, o *Post* "e outros não deveriam coletar e manter bancos de dados que

o governo poderia dar conta sem grandes esforços. Esse não deveria ser nosso trabalho, porra."[289] Porém, simultaneamente, as estatísticas sobre violência à mão armada que Rich frustra-se tanto em manter são mais poderosas do que ele jamais imaginou: em dados do governo estão embutidos vieses e decisões sobre o que coletar que precisam ser questionados e ponderados. Dados não são inertes, mas, sim, repletos de pensamentos predeterminados a respeito de quais fatos importam. Jornalistas que buscam controlar o domínio da facticidade precisam ser capazes de explicar por que os fatos são como são. E, de fato, a produção sistemática destes foi o meio pelo qual estes mesmos jornalistas conseguiram tomar para si uma autoridade epistêmica ao longo de boa parte da história do jornalismo moderno.

Ou seja, o jornalismo de dados é muito mais amplo e serve para muito mais do que para o que é empregado agora. Pode ser divertido, pode ser experimental, pode ser uma forma de mudar como histórias são contadas e um convite a repensar como esse processo se dá. Mas também tem um papel essencial em recolocar o jornalismo como meio de difusão da verdade e provedor de fatos. Ao criar e ter conhecimento de dados, sendo capaz de explicar processos de observação e coleta de informações que levam a um fato, o jornalismo de dados pode ser uma linha de defesa vital para o ofício jornalístico, capaz de coletar fatos melhor que qualquer outra ocupação, instituição ou pessoa comum jamais poderia.

Nikki Usher estuda produção de notícias na era digital e das plataformas; é autora dos livros "Making News at The New York Times" e "Interactive Journalism: Hackers, data, and code".

Referências

ADAMS, PC. *Migration Maps with the News: Guidelines for ethical zof mobile populations*. Journalism Studies 1:21, 2017, p. 527-547.

LEWIS, Seth C.; USHER, Nikki. *Trading zones, boundary objects, and the pursuit of news innovation: A case study of journalists and programmers*. Convergence 22:5, 2016, p. 543-560.

LEWIS, Norman P.; WATERS, Stephenson. *Data Journalism and the Challenge of Shoe-Leather Epistemologies*. Digital Journalism 1:18, 2017, p. 719-736.

LEWIS, Seth C; USHER, Nikki. *Open source and journalism: Toward new frameworks for imagining news innovation*. Media, Culture e Society 35:5, 2013, p. 602-619.

[289] https://twitter.com/dataeditor/status/964160884754059264.

LOOSEN, Wiebke; REIMER, Julius; DE SILVA-SCHMIDT, Fenja. *Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the data journalism awards 2013-2016*. Journalism, 2017.

NIEDERER, Christina; AIGNER, Wolfgang; RIND, Alexander. *Survey on visualizing dynamic, weighted, and directed graphs in the context of data-driven journalism*. Proceedings of the International Summer School on Visual Computing, 2015, p. 49-58.

USHER, Nikki. *Interactive journalism: Hackers, data, and code*. Urbana-Champaign: University of Illinois Press, 2016.

Jornalismo de dados e liberalismo digital

Dominic Boyer

Os últimos trinta anos foram palco de uma enorme transformação na profissão de jornalista e na cultura organizacional de notícias. As causas e os efeitos de tais transformações são complexas demais para serem abordadas aqui. Basta dizer que o modelo de impressão e transmissão terrestre que parecia bastante robusto até o final da década de 1990 foi quase que inteiramente substituído por um modelo digital de comunicação criado pela ascensão da internet, por motores de busca, pelas redes sociais como sistemas de comunicação e informação dominantes, e pela ampla financeirização e privatização de veículos de comunicação movidas pela filosofia econômica do neoliberalismo. Como argumentado ao longo deste volume, fluxos de dados e símbolos digitais em franca proliferação são a condição padrão da prática jornalística atual. Todo jornalismo é agora, até certo ponto, "jornalismo de dados". Em meu livro *The Life Informatic,* de 2013, descrevi este processo como uma "revolução lateral", sugerindo que testemunhamos uma mudança ecológica da dominância de infraestruturas radiais (em grande parte unidirecionais, do centro para demais pontos) em notícias para um modelo lateral (multidirecional, ponto a ponto). Como observado por Raymond Williams em seu estudo histórico brilhante a respeito da ascensão da televisão (1974), a mídia eletrônica exibe potencialidades laterais e radiais desde o século XVIII. Como estas potencialidades foram descobertas e institucionalizadas sempre foi uma questão de circunstâncias sociais e políticas para além das tecnologias envolvidas. Já existia um protótipo de uma máquina de fax por pelo menos um século antes de existir uma necessidade social óbvia de tal tecnologia, então sua "invenção" formal foi atrasada com base nisso. Sistemas de transmissão do rádio à televisão primeiramente se fizeram necessários, argumenta Williams, até que aquilo que ele chamou de "privatização móvel" da sociedade ocidental avançou ao ponto de dificultar governo e indústria a localizarem e se comunicarem com cidadãos-consumidores de outra forma que não por um sistema de mensagem radial que cobrisse todo um território. A lição que temos aí para nossa situação contemporânea é que não devemos presumir que a atual revolução dos dados no noticiário se dá ou é movida primariamente por novas tecnologias e infraestruturas como a internet. Devemos, sim, prestar atenção à forma como veículos de comunicação evoluíram (e continuam evoluindo) dentro de uma ecologia mais complexa de forças sociais.

A abordagem de Williams deu bases para o conceito de "liberalismo digital" que desenvolvi em *The Life Informatic,* de forma a captar um palpite que elaborei durante meu trabalho de campo com jornalistas ao final dos anos 2000: a sensação de que havia uma relação simbiótica entre práticas jornalísticas de informação digital e a neoliberalização mais

ampla da sociedade e da economia ocorrida desde os anos 1980. Eu estava interessado, por exemplo, na importância crescente do trabalho realizado diante de uma tela por jornalistas. Pode-se considerar que este tipo de trabalho é uma espécie de precondição infraestrutural para o surgimento do jornalismo de dados. O trabalho mediado por tela surgiu como aspecto do trabalho jornalístico nas décadas de 1970 e 1980, movido por iniciativas organizacionais na mídia e em outros setores da cultura corporativa ocidental para utilizar computadores pessoais e sistemas digitais de informação de escritório para gerar novas eficiências produtivas. Na indústria de notícias, a informatização foi vista, originalmente, como forma de melhorar a velocidade do processamento de textos e reduzir custos através da automatização da escrita e de algum trabalho de edição. Mas, no decorrer de seu processo de institucionalização, os computadores logo passaram a integrar todos os aspectos possíveis da produção de notícias, do marketing ao design e arquivamento, criando novas oportunidades para a automação de tarefas previamente feitas por humano e concentrando, assim, demais tarefas relacionadas à produção nas mãos de cada vez menos trabalhadores. Jornalistas veteranos, que lembram de como eram as coisas antes da informatização, frequentemente me falam sobre como a equipe de apoio costumava ser maior, como passam mais tempo em suas mesas agora e como a carga de trabalho individual aumentou.

É fato que o jornalismo sempre teve um lado sedentário. Datilografar, por exemplo, também era um processo feito sentado a uma mesa, assim como o uso de telefone antes da invenção dos aparelhos celulares. A principal diferença entre os formatos antigos de jornalismo sedentário e suas variantes contemporâneas é como o atual trabalho mediado por tela reúne um número sem precedentes de tarefas jornalísticas relevantes (processamento de texto, edição, pesquisa em arquivos, monitoramento de notícias e outros veículos, comunicação e coordenação dentro da redação) em uma única interface, geralmente de localização fixa. A importância combinada do smartphone e das redes sociais ágeis como o Twitter para o jornalismo fez com que o trabalho mediado por telas móveis adquirisse, no mínimo, a mesma relevância que o trabalho feito no computador, com poucas mudanças ao fato de que jornalistas permanecem "grudados em suas telas". Poucos contestariam a afirmação de que a tela se tornou um aspecto central da prática jornalística. Quase tudo que um jornalista faz, quase toda fonte de informação e quase todo aspecto de sua entrega envolve interagir com uma ou mais telas.

Esta organização de tarefas essenciais gera uma espécie de conveniência, mas também distração. Muitos jornalistas afirmam se sentir assoberbados pelo número e pela velocidade de fluxos de informação que precisam administrar. É importante reconhecer que a experiência de fazer jornalismo de dados é, frequentemente, uma experiência de ansiedade. Em minha pesquisa de campo, estes jornalistas de tela relataram depender, muitas vezes, de outras fontes confiáveis de notícia para determinar algo (por exemplo, se havia valor naquela

notícia) por conta de estarem tão sobrecarregados. É fácil perceber como este trabalho mediado por tela contribui com uma muito malvista "mentalidade de rebanho" no contexto do noticiário atual, com jornalistas distraídos e sobrecarregados, muitas vezes dependentes um dos outros para direcionamento, enquanto informações chegam na velocidade da luz.

Ao saber que a dominância deste tipo de trabalho não surgiu no vácuo, uma investigação paralela do neoliberalismo se mostra útil. Surgido nos séculos XVII e XVIII, o liberalismo clássico é uma espécie de desdobramento da cultura intelectual europeia adaptada às realidades da formação de impérios coloniais pelo mundo. O domínio cultural do conservadorismo medieval cristão e, até mesmo, do humanismo renascentista foi cada vez mais deixado de lado em detrimento de filosofias sociais focadas em trabalho, liberdade, propriedade privada e produtividade. Um problema fundamental do liberalismo em seus primórdios foi como fazer da busca e conquista da propriedade privada o caminho virtuoso a se trilhar, visto que esta ameaçava tomar dos pobres sua parcela das dádivas de Deus à humanidade. A solução encontrada foi enfatizar que a capacidade da ciência e da indústria humanas de melhorarem o uso produtivo de recursos combinada à abundância das novas fronteiras coloniais significava que a aquisição de propriedade privada não se oporia aos valores cristãos. Talvez uma consequência não intencional desta nova ética foi a atenção voltada ao indivíduo como sujeito da razão, ação, liberdade e virtude. Ao passo que o liberalismo se desenvolvia junto ao capitalismo e suas formas modernas de vida, o indivíduo foi ganhando destaque na cultura ocidental. Em um primeiro momento, buscava contrabalançar de maneira harmoniosa as forças limitadoras da "sociedade", mas a individualidade cada vez mais era um objetivo por si só, em que todas as relações sociais e econômicas serviam para desenvolver e possibilitar o advento de indivíduos robustos, produtivos e autossustentáveis — imaginados como sendo idealmente livres de determinações sociais, livres para pensar e agir como bem entendessem. Descrevo este modelo de individualidade, com base no trabalho da antropóloga Elizabeth Povinelli, como "autológico", por conta de sua hipótese ideológica de que indivíduos em grande parte são capazes de "sucesso" por conta própria, uma proposta muito benquista no espectro político liberal até os dias de hoje.

Aí você pode se perguntar: o que isso tem a ver com computadores? É verdade que as primeiras grandes empreitadas em computação analógica e digital se deram nos anos 1930 enquanto as social-democracias keynesianas de meados do século XX se preparavam para a guerra. Mas o desenvolvimento da computação pessoal, que serviu de antepassado direto do trabalho mediado por telas atual, ocorreu nas décadas de 1970 e 1980, ao passo que o neoliberalismo passava a dominar o pensamento político e filosófico, desdobramentos acompanhados pelo aparente colapso da social-democracia keynesiana sob múltiplas pressões geopolíticas ao final das décadas de 1960 e 1970 (a guerra do Vietnã, conflitos entre árabes e

israelenses, a formação da OPEP, entre outras crises). Onde o liberalismo há muito acreditava que a melhor forma de atender aos interesses públicos era dar autonomia aos privados, é possível descrever o neoliberalismo como impiedosamente autológico em seu reforço dos interesses privados às custas de investimentos e instituições públicas. Na esfera política, o neoliberalismo teve impacto profundamente negativo no tipo de jornalismo de interesse público que acompanhava normas keynesianas de meados do século XX, mesmo com o fomento a grandes investimentos em infraestruturas de comunicação e informação como internet, transmissão via satélite e telefonia celular pelo mundo. Originalmente, a invenção e a criação de tais infraestruturas pouco tinha a ver com veículos de notícias. A internet, como muitos sabem, surgiu por conta de interesses compartilhados entre militares e pesquisadores. O que poucos sabem, mas é de igual importância, é a questão da utilidade de comunicações ágeis entre nações para práticas financeiras como arbitragem. De qualquer forma, estas novas infraestruturas de informação e comunicação impactaram todas as áreas da comunicação social, o que inclui, claro, notícias. Um de seus feitos foi o fortalecimento radical de capacidades de troca lateral de mensagens ponto a ponto, bem como a pluralização e retemporalização de transmissões radiais de tal forma que estas ainda existem, mas de maneira cada vez mais transfronteiriça e assíncrona. A ideia de um país inteiro parar para ouvir as notícias da noite junto simplesmente não existe mais em quase nenhum lugar do mundo, mesmo na Europa e na Ásia, continentes em que tradições mais fortes relacionadas à transmissão pública perduram.

Nossa ecologia de notícias contemporânea não inclui o individualismo robusto, mesmo que tenha feito do processo de encontrar informações da comunidade e informações confiáveis ainda mais precário. O liberalismo digital é posto em prática na experiência individualizadora do trabalho mediado por tela (e do entretenimento mediado por telas também). A evolução da computação pessoal, da internet e das redes sociais foi grandemente moldada pela importância social de princípios neoliberais/liberais de maximizar capacidades individuais de ação, comunicação e ideação. Ao longo da última década, uma porcentagem crescente da população (mais de 70% nos EUA, por exemplo) carrega consigo um dispositivo portátil de mídia multifuncional, quase como um membro extra do seu corpo. Este membro possibilita acessar diversos fluxos de informação e fazer curadoria destes de forma a refletir interesses e desejos pessoais, uma multitude de formas de encaminhar mensagens com visões e pensamentos pessoais das coisas e constituir micropúblicos autocentrados. Herdeiro de séculos de epistemologia liberal e, também, o aparelho crucial que possibilita a reprodução e intensificação desta epistemologia, naquilo que chamamos de “era digital”. Você já viu fotos por aí de estranhos em um bar ou trem, todos colados na tela. Claro que smartphones não inventaram a alienação social. O que eles inventaram foi uma interface comunicacional que nos permite vivenciar uma individualidade ativa e produtiva, ao mesmo tempo que minimiza

conexões sociais e responsabilização, mesmo quando estamos cercados por estranhos em qualquer canto do mundo. Em outras eras, estes estranhos poderiam ter se valido de oportunidades como essas para serem sociáveis uns com os outros.

Em suma, sigo convencido de que a individualidade autológica segue reforçada pela proliferação e intensificação de telas e suas interfaces, mesmo que a existência destas interfaces tenha muito a ver com tecnologias criadas para materializar visões de mundo e prioridades liberais ao longo dos últimos séculos. Parafraseando Marshall McLuhan, presumimos que trabalhamos em nossas telas, mas devemos reconhecer que elas têm efeitos em nós. Esta conjuntura formada por mídia portátil baseada em telas e percepções liberais de individualidade é aquilo que chamo de "liberalismo digital" e será interessante ver como o liberalismo evoluirá no futuro. E se todos aqueles desconhecidos no trem estivessem usando headsets de realidade virtual que lhes possibilitaria acesso imersivo a mundos virtuais? Como novas interfaces de mídia levariam a novos modos de individualidade e socialidade? Por mais que muitas vezes se suspeite da aproximação entre jornalismo de dados, tecnologias de vigilância e autoritarismo algorítmico, diria que a evolução do liberalismo digital é, na verdade, a história mais profunda do jornalismo de dados.

Dominic Boyer é antropólogo, cineasta e podcaster, diretor do Centro de Pesquisa Energética e Ambiental em Ciências Humanas (CENHS) da Rice University.

Referências

BOYER, Dominic. *The Life Informatic.* Ithaca: Cornell University Press, 2013.

WILLIAMS, Raymond. *Television*. Hanover: Wesleyan University Press, 1974.