Artigo de pesquisa original

Big Data & Society
abril–junho 2014: 1–
12 ! O(s) autor(es) 2014
DOI: 10.1177/2053951714528481
bds.sagepub.com

Big Data, novas epistemologias e mudanças de paradigma

Rob Kitchin

Resumo

Este artigo examina como a disponibilidade de Big Data, juntamente com a nova análise de dados, desafia epistemologias estabelecidas nas ciências, ciências sociais e humanidades, e avalia até que ponto elas estão gerando mudanças de paradigma em várias disciplinas. Em particular, explora criticamente novas formas de empirismo que declaram "o fim da teoria", a criação de ciência orientada por dados em vez de ciência orientada por conhecimento, e o desenvolvimento de humanidades digitais e ciências sociais computacionais que propõem maneiras radicalmente diferentes de fazer sentido da cultura, da história, da economia e da sociedade. Argumenta-se que: (1) Big Data e novas análises de dados são inovações disruptivas que estão reconfigurando em muitos casos como a pesquisa é conduzida; e (2) há uma necessidade urgente de uma reflexão crítica mais ampla dentro da academia sobre as implicações epistemológicas da revolução dos dados em desenvolvimento, uma tarefa que mal começou a ser abordada apesar das rápidas mudanças nas práticas de pesquisa que estão ocorrendo atualmente. Depois de revisar criticamente as posições epistemológicas emergentes, sustenta-se que uma abordagem potencialmente frutífera seria o desenvolvimento de uma epistemologia situada, reflexiva e contextualmente nuançada.

Palavras-chave

Big Data, data analytics, epistemologia, paradigmas, fim da teoria, data-driven science, humanidades digitais, ciências sociais computacionais

Introdução

> As revoluções na ciência muitas vezes foram precedidas por revoluções na medição. Sinan Aral (citado em Cukier, 2010)

> Big Data cria uma mudança radical na forma como pensamos sobre pesquisa... [Ele oferece] uma mudança profunda nos níveis de epistemologia e ética. Big Data reformula questões-chave sobre a constituição do conhecimento, os processos de pesquisa, como devemos nos envolver com a informação e a natureza e a categorização da realidade... Big Data demarca novos terrenos de objetos, métodos de conhecimento e definições da vida social. (boyd e Crawford, 2012)

Tal como acontece com muitos conceitos emergentes rapidamente, o Big Data foi definido e operacionalizado de várias formas, variando de proclamações banais de que o Big Data consiste em conjuntos de dados muito grandes para caber em uma planilha do Excel ou ser armazenado em uma única máquina (Strom, 2012) para mais avaliações ontológicas sofisticadas que provocam suas características inerentes (boyd e Crawford, 2012; Mayer-Schonberger e Cukier, 2013). Com base em um amplo envolvimento com a literatura, Kitchin (2013) detalha que Big Data é:

. enorme em volume, consistindo de terabytes ou petabytes
De dados;
. alta velocidade, sendo criado em ou quase em tempo real; . diversos
em variedade, sendo estruturados e não estruturados por natureza; .
exaustivo
em escopo, esforçando-se para capturar todo o popu
ções ou sistemas (n ¼ todos);

Instituto Nacional de Análise Regional e Espacial, Universidade Nacional de Irlanda Maynooth, County Kildare, Irlanda

Autor correspondente: Rob
Kitchin, Instituto Nacional de Análise Regional e Espacial, Universidade Nacional da Irlanda Maynooth, County Kildare, Irlanda.
E-mail: Rob.Kitchin@nuim.ie

. refinada na resolução e exclusivamente indexical na identificação; . de natureza
relacional, contendo campos comuns que permitem a junção de diferentes conjuntos de dados; . flexível, mantendo as
características de extensionalidade (pode adicionar novos campos facilmente) e escalabilidade (pode expandir em tamanho rapidamente). (ver boyd e Crawford, 2012; Dodge e Kitchin, 2005; Laney, 2001; Marz e Warren, 2012; Mayer-Schonberger e Cukier, 2013; Zikopoulos et al., 2012).

Em outras palavras, Big Data não é simplesmente denotado por volume. De fato, a indústria, o governo e a academia há muito produzem conjuntos de dados massivos – por exemplo, censos nacionais. No entanto, dados os custos e as dificuldades de gerar, processar, analisar e armazenar tais conjuntos de dados, esses dados têm sido produzidos de forma rigidamente controlada, usando técnicas de amostragem que limitam seu escopo, temporalidade e tamanho (Miller, 2010). Para tornar o exercício de compilação de dados do censo gerenciável, eles foram produzidos uma vez a cada cinco ou 10 anos, fazendo apenas 30 a 40 perguntas, e seus resultados são geralmente bastante grosseiros em resolução (por exemplo, áreas locais ou condados em vez de indivíduos e famílias).

Além disso, os métodos usados para gerá-los são bastante inflexíveis (por exemplo, uma vez que um censo é definido e administrado, é impossível ajustar ou adicionar/remover perguntas). Enquanto o censo busca ser exaustivo, enumerando todas as pessoas que vivem em um país, a maioria das pesquisas e outras formas de geração de dados são amostras, buscando ser representativas de uma população.

Já o Big Data se caracteriza por ser gerado continuamente, buscando ser exaustivo e minucioso em seu escopo, e flexível e escalável em sua produção. Exemplos da produção de tais dados incluem: CCTV digital; o registro de compras no varejo; dispositivos digitais que registram e comunicam o histórico de seu próprio uso (por exemplo, telefones celulares); o registro de transações e interações em redes digitais (por exemplo, e-mail ou banco online); dados de clickstream que registram a navegação por meio de um site ou aplicativo; medições de sensores embutidos em objetos ou ambientes; a digitalização de objetos legíveis por máquina, como passes de viagem ou códigos de barras; e postagens em mídias sociais (Kitchin, 2014). Estes estão produzindo fluxos maciços e dinâmicos de dados relacionais diversos e refinados. Por exemplo, em 2012, o Wal-Mart estava gerando mais de 2,5 petabytes (250 bytes) de dados relacionados a mais de 1 milhão de transações de clientes a cada hora (Open Data Center Alliance, 2012) e o Facebook relatou que estava processando 2,5 bilhões de partes de conteúdo (links, comentários, etc.), 2,7 bilhões de ações 'Curtir' e 300 milhões de uploads de fotos por dia (Constine, 2012). Manusear e analisar esses dados é uma proposta muito diferente de lidar com com um censo a cada 10 anos ou uma pesquisa com algumas centenas de entrevistados.

Embora a produção de tal Big Data tenha existido em alguns domínios, como sensoriamento remoto, previsão do tempo e mercados financeiros, por algum tempo, vários desenvolvimentos tecnológicos, como computação onipresente, trabalho generalizado na Internet e novos projetos de banco de dados e soluções de armazenamento, criaram um ponto de inflexão para sua geração e análise de rotina, não menos do que novas formas de análise de dados projetadas para lidar com a abundância de dados (Kitchin, 2014).

Tradicionalmente, as técnicas de análise de dados foram projetadas para extrair insights de conjuntos de dados escassos, estáticos, limpos e pouco relacionais, cientificamente amostrados e aderindo a suposições rígidas (como independência, estacionaridade e normalidade) e gerados e analisados com uma pergunta específica em mente (Miller, 2010).

O desafio de analisar Big Data é lidar com abundância, exaustividade e variedade, pontualidade e dinamismo, confusão e incerteza, alta relacionalidade e o fato de que muito do que é gerado não tem uma questão específica em mente ou é subproduto de outra atividade. Tal desafio era até recentemente muito complexo e difícil de implementar, mas tornou-se possível devido à computação de alta potência e novas técnicas analíticas. Essas novas técnicas estão enraizadas na pesquisa sobre inteligência artificial e sistemas especialistas que buscaram produzir aprendizado de máquina que pode minerar e detectar padrões de forma computacional e automática, construir modelos preditivos e otimizar resultados (Han et al., 2011; Hastie et al., 2009). Além disso, uma vez que diferentes modelos têm seus pontos fortes e fracos, e muitas vezes é difícil prejulgar qual tipo de modelo e suas várias versões terão melhor desempenho em qualquer conjunto de dados, uma abordagem conjunta pode ser empregada para construir várias soluções (Seni e Velho, 2010). Aqui, literalmente centenas de algoritmos diferentes podem ser aplicados a um conjunto de dados para determinar o melhor ou um modelo composto ou explicação (Siegel, 2013), uma abordagem radicalmente diferente daquela tradicionalmente usada em que o analista seleciona um método apropriado com base em seu conhecimento de técnicas e os dados. Em outras palavras, a análise de Big Data permite uma abordagem epistemológica totalmente nova para dar sentido ao mundo; em vez de testar uma teoria analisando dados relevantes, a nova análise de dados busca obter insights "nascidos dos dados".

A explosão na produção de Big Data, juntamente com o desenvolvimento de novas epistemologias, está levando muitos a argumentar que uma revolução de dados está em andamento e tem consequências de longo alcance sobre como o conhecimento é produzido, os negócios conduzidos e a governança promulgada (Anderson, 2008; Bollier, 2010; Floridi, 2012; Mayer Schonberger e Cukier, 2013). No que diz respeito à produção de conhecimento, afirma-se que o Big Data apresenta

Tabela 1. Quatro paradigmas da ciência.

Paradigma	Natureza	Forma	Quando
Primeiro	ciência experimental	Empirismo; descrevendo fenômenos naturais	pré-renascentista
Segundo	ciência teórica	Modelagem e generalização	pré-computadores
Terceiro	ciência computacional	Simulação de fenômenos complexos	pré-Big Data
Quarto	ciência exploratória	Intensivo em dados; exploração estatística e mineração de dados	Agora

Compilado de Hey et al. (2009).

a possibilidade de um novo paradigma de pesquisa em múltiplas disciplinas. Conforme definido por Kuhn (1962), um paradigma constitui uma forma aceita de interrogar o mundo e sintetizar o conhecimento comum a uma proporção substancial de pesquisadores em uma disciplina em qualquer momento no tempo. Periodicamente, argumenta Kuhn, surge uma nova forma de pensar que desafia as teorias e abordagens aceitas. Por exemplo, a teoria da evolução de Darwin alterou radicalmente o pensamento conceitual dentro das ciências biológicas, bem como desafiou a doutrina religiosa do criacionismo. Jim Gray (conforme detalhado em Hey et al., 2009) mapeia a evolução da ciência por meio de quatro paradigmas amplos (consulte a Tabela 1). Ao contrário da proposição de Kuhn de que as mudanças de paradigma ocorrem porque o modo dominante de ciência não pode dar conta de fenômenos particulares ou responder a questões-chave, exigindo assim a formulação de novas ideias, as transições de Gray são baseadas em avanços em formas de dados e no desenvolvimento de novas técnicas analíticas. métodos. Assim, ele propõe que a ciência está entrando em um quarto paradigma baseado na crescente disponibilidade de Big Data e novas análises.

O argumento de Kuhn tem sido objeto de muitas críticas, até porque dentro de alguns domínios acadêmicos há pouca evidência de paradigmas operando, notavelmente em algumas ciências sociais onde há um conjunto diversificado de abordagens filosóficas empregadas (por exemplo, geografia humana, sociologia), embora em Em outros domínios, como as ciências, tem havido maior unidade epistemológica em torno de como a ciência é conduzida, usando um método científico bem definido, sustentado por testes de hipóteses para verificar ou falsificar teorias. Além disso, relatos paradigmáticos produzem histórias excessivamente higiênicas e lineares de como as disciplinas evoluem, suavizando as formas confusas, contestadas e plurais nas quais a ciência se desenvolve na prática. No entanto, embora a noção de paradigmas seja problemática, ela é útil para enquadrar os debates atuais sobre o desenvolvimento do Big Data e suas consequências, porque muitas das reivindicações feitas com relação à produção de conhecimento afirmam que uma epistemologia fundamentalmente diferente está sendo criada. ; que está em curso uma transição para um novo paradigma. No entanto, a forma que essa nova epistemologia está tomando é contestada. O restante deste artigo examina criticamente o desenvolvimento de um quarto paradigma emergente na ciência e sua forma, e explora até que ponto os dados A revolução está levando a epistemologias alternativas nas ciências humanas e sociais e mudando as práticas de pesquisa.

Um quarto paradigma na ciência?

Enquanto Jim Gray prevê que o quarto paradigma da ciência seja intensivo em dados e uma extensão radicalmente nova do método científico estabelecido, outros sugerem que Big Data inaugura uma nova era de empirismo, em que o volume de dados, acompanhado por técnicas que podem revelar sua verdade inerente, permite que os dados falem por si mesmos, livres de teoria. A visão empirista ganhou credibilidade fora da academia, especialmente nos círculos de negócios, mas suas ideias também se enraizaram no novo campo da ciência de dados e outras ciências. Em contraste, um novo modo de ciência orientada por dados está surgindo nas disciplinas tradicionais da academia. Nesta seção, as reivindicações epistemológicas de ambas as abordagens são examinadas criticamente, atentas aos diferentes impulsionadores e aspirações dos negócios e da academia, com a primeira preocupada em empregar análise de dados para identificar novos produtos, mercados e oportunidades, em vez de avançar o conhecimento per se , e o último enfocou a melhor forma de dar sentido ao mundo e determinar explicações sobre fenômenos e processos.

O fim da teoria: o empirismo renasce

Para comentaristas como Chris Anderson, ex-editor-chefe da revista Wired, Big Data, novas análises de dados e abordagens de conjunto sinalizam uma nova era de produção de conhecimento caracterizada pelo 'fim da teoria'. Em uma peça provocativa, Anderson (2008) argumenta que 'o dilúvio de dados torna o método científico obsoleto'; que os padrões e relacionamentos contidos no Big Data produzem inerentemente conhecimento significativo e perspicaz sobre fenômenos complexos. Essencialmente argumentando que o Big Data permite um modo empirista de produção de conhecimento, ele afirma:

> Agora existe uma maneira melhor. Os petabytes nos permitem dizer: 'A correlação é suficiente.' ... Podemos analisar os dados sem hipóteses sobre o que eles podem mostrar. Podemos jogar os números no maior computador

> clusters que o mundo já viu e permitir que algoritmos estatísticos encontrem padrões onde a ciência não pode... A correlação substitui a causação, e a ciência pode avançar mesmo sem modelos coerentes, teorias unificadas ou realmente qualquer explicação mecanicista.
> Não há razão para nos apegarmos aos nossos velhos hábitos.

Da mesma forma, Prensky (2009) argumenta:

> os cientistas não precisam mais fazer suposições fundamentadas, construir hipóteses e modelos e testá-los com experimentos e exemplos baseados em dados. Em vez disso, eles podem explorar o conjunto completo de dados em busca de padrões que revelem efeitos, produzindo conclusões científicas sem mais experimentos.

Dyche (2012) assim argumenta que 'a mineração de Big Data revela relacionamentos e padrões que nem sabíamos procurar'. Da mesma forma, Steadman (2013) argumenta:

> A abordagem de Big Data para coleta de inteligência permite que um analista obtenha a resolução completa dos assuntos mundiais. Nada se perde ao examinar muito de perto uma seção específica de dados; nada se perde ao tentar obter uma perspectiva muito ampla de uma situação em que os detalhes finos são perdidos... . O analista nem precisa mais se preocupar em propor uma hipótese.

Os exemplos usados para ilustrar tal posição geralmente vêm do marketing e do varejo. Por exemplo, Dyche (2012) detalha o caso de uma rede de varejo que analisou 12 anos de transações de compra para possíveis relações despercebidas entre produtos que acabaram nas cestas dos compradores. A descoberta de correlações entre determinados itens levou à colocação de novos produtos e a um aumento de 16% na receita por carrinho de compras no primeiro mês de teste. Não havia hipótese de que o Produto A fosse frequentemente comprado com o Produto H que era então testado. Os dados foram simplesmente consultados para descobrir quais relações existiam que poderiam ter passado despercebidas anteriormente. Da mesma forma, o sistema de recomendação da Amazon produz sugestões de outros itens nos quais um comprador pode estar interessado sem saber nada sobre a cultura e as convenções dos livros e da leitura; ele simplesmente identifica padrões de compra entre os clientes para determinar se a Pessoa A gosta do Livro X, ela também provavelmente gostará do Livro Y, dados seus próprios padrões de consumo e os de outros. Embora possa ser desejável explicar por que existem associações nos dados e por que elas podem ser significativas, tal explicação é considerada desnecessária. Siegel (2013: 90) argumenta assim com relação à análise preditiva: 'Geralmente não sabemos sobre causalidade e muitas vezes não nos importamos necessariamente ... o objetivo é mais prever do que entender o mundo...Só precisa funcionar; a previsão supera a explicação'.

Alguns softwares de análise de dados são vendidos precisamente com essa noção. Por exemplo, o software de mineração e visualização de dados Ayasdi afirma ser capaz de

> descubra insights automaticamente – independentemente da complexidade – sem fazer perguntas. Os clientes da Ayasdi podem finalmente aprender as respostas para perguntas que eles não sabiam fazer em primeiro lugar. Simplificando, Ayasdi é 'serendipidade digital'. (Clark, 2013)

Além disso, pretende ter removido totalmente

> o elemento humano que entra na mineração de dados – e, como tal, todo o viés humano que o acompanha. Em vez de esperar para fazer uma pergunta ou ser direcionado para links de dados existentes específicos, o sistema fornecerá - sem direcionamento - padrões que um controlador humano pode não ter pensado em procurar. (Clark, 2013)

Há um poderoso e atraente conjunto de ideias em ação na epistemologia empirista que vai contra a abordagem dedutiva que é hegemônica dentro da ciência moderna:

. Big Data pode capturar um domínio inteiro e fornecer resolução total; . não há necessidade
de teoria, modelos ou hipóteses a priori; . por meio da aplicação de análise de
dados agnóstica, os dados podem falar por si mesmos, livres de preconceito ou enquadramento humano, e quaisquer padrões e relacionamentos dentro do Big Data são inerentemente significativos e verdadeiros;

. o significado transcende o contexto ou o conhecimento específico do domínio, portanto, pode ser interpretado por qualquer pessoa que possa decodificar uma estatística ou visualização de dados.

Estes trabalham juntos para sugerir que um novo modo de ciência está sendo criado, no qual o modus oper andi é puramente indutivo por natureza.

Embora essa epistemologia empirista seja atraente, ela se baseia em um pensamento falacioso com relação às quatro ideias que sustentam sua formulação. Em primeiro lugar, embora o Big Data possa procurar ser exaustivo, capturando todo um domínio e fornecendo resolução completa, é tanto uma representação quanto uma amostra, moldada pela tecnologia e plataforma utilizada, a ontologia de dados empregada e o ambiente regulatório, e é sujeito a viés de amostragem (Crawford, 2013; Kitchin, 2013). De fato, todos os dados fornecem visões oligópticas do mundo: visões de certos pontos de vista, usando ferramentas específicas, em vez de uma visão infalível e onisciente do olho de Deus (Amin e Thrift, 2002; Haraway, 1991). Assim, os dados não são simplesmente

elementos naturais e essenciais que são abstraídos do mundo de maneira neutra e objetiva e podem ser aceitos pelo valor de face; os dados são criados dentro de um conjunto complexo que molda ativamente sua constituição (Ribes e Jackson, 2013).

Em segundo lugar, o Big Data não surge do nada, livre da 'força reguladora da filosofia' (Berry, 2011: 8). Por outro lado, os sistemas são projetados para capturar certos tipos de dados e as análises e algoritmos usados são baseados em raciocínio científico e foram refinados por meio de testes científicos. Como tal, uma estratégia indutiva de identificação de padrões nos dados não ocorre em um vácuo científico e é discursivamente enquadrada por descobertas, teorias e treinamento anteriores; pela especulação alicerçada na experiência e no conhecimento (Leonelli, 2012). Novas análises podem dar a ilusão de descobrir insights automaticamente sem fazer perguntas, mas os algoritmos usados com certeza surgiram e foram testados cientificamente quanto à validade e veracidade.

Em terceiro lugar, assim como os dados não são gerados livres da teoria, eles também não podem simplesmente falar por si mesmos, livres de preconceitos ou enquadramentos humanos. Como Gould (1981: 166) observa, “dados inanimados nunca podem falar por si mesmos, e nós sempre trazemos algum quadro conceitual, seja intuitivo e mal formado, ou rígido e formalmente estruturado, para a tarefa de investigação, análise , e interpretação'. Dar sentido aos dados é sempre enquadrado – os dados são examinados através de uma lente específica que influencia a forma como são interpretados. Mesmo que o processo seja automatizado, os algoritmos usados para processar os dados são imbuídos de valores particulares e contextualizados dentro de uma abordagem científica particular.

Além disso, os padrões encontrados em um conjunto de dados não são inerentemente significativos. Correlações entre variáveis dentro de um conjunto de dados podem ser de natureza aleatória e ter nenhuma ou pouca associação causal, e interpretá-las como tal pode produzir sérias falácias ecológicas. Isso pode ser exacerbado no caso do Big Data, pois a posição empirista parece promover a prática da dragagem de dados – caçar cada associação ou modelo.

Em quarto lugar, a ideia de que os dados podem falar por si sugere que qualquer pessoa com um conhecimento razoável de estatística deve ser capaz de interpretá-los sem contexto ou conhecimento específico do domínio. Este é um conceito expresso por alguns cientistas de dados e da computação e outros cientistas, como físicos, todos os quais se tornaram ativos na prática de pesquisas em ciências sociais e humanidades. Por exemplo, vários físicos voltaram sua atenção para as cidades, empregando a análise de Big Data para modelar processos sociais e espaciais e para identificar supostas leis que sustentam sua formação e funções (Bettencourt et al., 2007; Lehrer, 2010).

Esses estudos geralmente ignoram deliberadamente alguns séculos de estudos em ciências sociais, incluindo quase um século de análise quantitativa e construção de modelos. O resultado é uma análise das cidades reducionista, funcionalista e que ignora os efeitos da cultura, da política, da política, da governança e do capital (reproduzindo os mesmos tipos de limitações geradas pelas ciências sociais quantitativas/positivistas em meados do século XX). Um conjunto semelhante de preocupações é compartilhado por aqueles nas ciências. Strasser (2012), por exemplo, observa que dentro das ciências biológicas, os bioinformáticos que têm uma forma muito estreita e particular de entender a biologia estão reivindicando um espaço outrora ocupado pelo clínico e pelo biólogo experimental e molecular. Esses cientistas estão, sem dúvida, ignorando as observações de Porway (2013):

> Sem especialistas no assunto disponíveis para articular os problemas com antecedência, você obtém resultados [ruins]... . Especialistas no assunto são duplamente necessários para avaliar os resultados do trabalho, especialmente quando você está lidando com dados confidenciais sobre o comportamento humano. Como cientistas de dados, estamos bem equipados para explicar o “o quê” dos dados, mas raramente devemos tocar na questão do “porquê” em assuntos nos quais não somos especialistas.

Simplificando, embora os dados possam ser interpretados sem contexto e conhecimento específico do domínio, essa interpretação epistemológica provavelmente será anêmica ou inútil, pois carece de incorporação em debates e conhecimentos mais amplos.

Essas noções falaciosas ganharam alguma força, especialmente nos círculos de negócios, porque possuem uma narrativa conveniente para as aspirações de negócios orientados ao conhecimento (por exemplo, corretores de dados, provedores de análise de dados, fornecedores de software, consultorias) na venda de seus serviços. Dentro do quadro empirista, a análise de dados oferece a possibilidade de conhecimento perspicaz, objetivo e lucrativo sem ciência ou cientistas e suas despesas gerais associadas de custo, contingências e busca por explicação e verdade. Nesse sentido, embora as técnicas de ciência de dados empregadas possam ter relevância genuína para os profissionais, a articulação de um novo empirismo opera como um dispositivo retórico discursivo projetado para simplificar uma abordagem epistemológica mais complexa e convencer os fornecedores da utilidade e valor do Big Data análise.

Ciência orientada por dados

Em contraste com as novas formas de empirismo, a ciência orientada por dados busca manter os princípios do método científico, mas está mais aberta a usar uma combinação híbrida de abordagens abdutivas, indutivas e dedutivas para avançar na compreensão de um fenômeno. Difere do design dedutivo experimental tradicional porque procura gerar hipóteses e insights

'nascido dos dados' em vez de 'nascido da teoria' (Kelling et al., 2009: 613). Em outras palavras, procura incorporar um modo de indução no projeto de pesquisa, embora a explicação por indução não seja o ponto final pretendido (como nas abordagens empiristas).
Em vez disso, forma um novo modo de geração de hipóteses antes que uma abordagem dedutiva seja empregada. O processo de indução também não surge do nada, mas é situado e contextualizado dentro de um domínio teórico altamente evoluído. Como tal, a estratégia epistemológica adotada na ciência orientada por dados é usar técnicas de descoberta de conhecimento guiada para identificar possíveis questões (hipóteses) dignas de exame e teste mais aprofundados.

O processo é guiado no sentido de que a teoria existente é usada para direcionar o processo de descoberta de conhecimento, em vez de simplesmente esperar identificar todos os relacionamentos dentro de um conjunto de dados e assumir que eles são significativos de alguma forma. Como tal, a forma como os dados são gerados ou reaproveitados é dirigida por certas suposições, sustentadas por conhecimento teórico e prático e experiência sobre se as tecnologias e suas configurações irão capturar ou produzir material de pesquisa apropriado e útil. Os dados não são gerados por todos os meios possíveis, usando todo tipo de tecnologia disponível ou todo tipo de estrutura de amostragem; em vez disso, as estratégias de geração e reaproveitamento de dados são cuidadosamente pensadas, com decisões estratégicas tomadas para colher certos tipos de dados e não outros.

Da mesma forma, a forma como esses dados são processados, gerenciados e analisados é guiada por suposições sobre quais técnicas podem fornecer insights significativos. Os dados não estão sujeitos a todos os enquadramentos ontológicos possíveis, ou a todas as formas de técnicas de mineração de dados na esperança de revelar alguma verdade oculta. Em vez disso, decisões teoricamente informadas são tomadas sobre a melhor forma de lidar com um conjunto de dados de modo que ele revele informações que serão de interesse potencial e dignas de pesquisas adicionais. E, em vez de testar se todas as relações reveladas têm veracidade, a atenção se concentra naquelas – com base em alguns critérios – que aparentemente oferecem o caminho mais provável ou válido a seguir. De fato, muitos relacionamentos supostos dentro de conjuntos de dados podem ser rapidamente descartados como triviais ou absurdos por especialistas de domínio, com outros sinalizados como merecedores de alta atenção (Mejias, 2010).

Essa tomada de decisão com relação aos métodos de geração e análise de dados é baseada em raciocínio abdutivo. Abdução é um modo de inferência lógica e raciocínio encaminhado por CS Peirce (1839-1914)
(Moleiro, 2010). Busca uma conclusão que faça sentido razoável e lógico, mas não é definitiva em sua afirmação.
Por exemplo, não há uma tentativa de deduzir qual é a melhor forma de gerar dados, mas sim de identificar uma abordagem que faça sentido lógico dado o que já se sabe sobre essa produção de dados. Abdução é muito comumente usado na ciência, especialmente na formulação de hipóteses, embora tal uso não seja amplamente reconhecido. Quaisquer relações reveladas nos dados não surgem do nada e nem simplesmente falam por si. O processo de indução – de percepções emergentes dos dados – é enquadrado contextualmente. E esses insights não são o ponto final de uma investigação, arranjados e fundamentados em uma teoria.

Em vez disso, os insights fornecem a base para a formulação de hipóteses e o teste dedutivo de sua validade. Em outras palavras, a ciência orientada por dados é uma versão reconfigurada do método científico tradicional, fornecendo uma nova maneira de construir a teoria. No entanto, a mudança epistemológica é significativa.

Em vez do empirismo e do fim da teoria, alguns argumentam que a ciência baseada em dados se tornará o novo paradigma do método científico em uma era de Big Data porque a epistemologia favorecida é adequada para extrair insights adicionais e valiosos que o conhecimento tradicional "ciência dirigida" não conseguiria gerar (Kelling et al., 2009; Loukides, 2010; Miller, 2010).
A ciência orientada pelo conhecimento, usando uma abordagem dedutiva direta, tem utilidade particular na compreensão e explicação do mundo sob as condições de dados escassos e computação fraca. Continuar a usar essa abordagem, no entanto, quando os avanços tecnológicos e metodológicos significam que é possível realizar análises de dados muito mais ricas – aplicando novas análises de dados e sendo capaz de conectar dados grandes e díspares de maneiras que até então eram impossíveis e que produzem novos dados valiosos e identificam e abordam questões de maneiras novas e empolgantes – faz pouco sentido. Além disso, os defensores da ciência orientada por dados argumentam que ela é muito mais adequada para explorar, extrair valor e dar sentido a conjuntos de dados maciços e interconectados, fomentando pesquisas interdisciplinares que unem expertise de domínio (já que é menos limitada pelo quadro teórico inicial), e que levará a modelos e teorias mais holísticos e extensivos de sistemas complexos inteiros, em vez de elementos deles (Kelling et al., 2009).

Por exemplo, afirma-se que a ciência baseada em dados transformará nossa compreensão dos sistemas ambientais (Bryant et al., 2008; Lehning et al., 2009). Ele permitirá que dados de alta resolução (Mejias, 2010) a partir de uma variedade de fontes, geralmente em tempo real (como estações meteorológicas convencionais e móveis, imagens aéreas e de satélite, radar meteorológico, observações de fluxo e estações de medição, observações de cidadãos, e LIDAR aéreo, amostragem de qualidade da água, medições de gás, núcleos de solo e sensores distribuídos que medem domínios selecionados, como temperatura e umidade do ar) para serem integrados para fornecer modelos muito detalhados de ambientes em fluxo (em oposição a congelamento- pontos no tempo e no espaço) e identificar relações específicas entre

fenômenos e processos que geram novas hipóteses e teorias que podem ser testadas posteriormente para estabelecer sua veracidade. Também ajudará a identificar e entender melhor os pontos de conexão entre diferentes esferas ambientais – como a atmosfera (ar), biosfera (ecossistemas), hidrosfera (sistemas de água), litosfera (casca rochosa da Terra) e pedosfera (solos).) – e ajuda na integração de teorias em um conjunto teórico mais holístico. Isso fornecerá uma melhor compreensão dos diversos processos inter-relacionados no trabalho e as interconexões com sistemas humanos, e pode ser usado para orientar modelos e simulações para prever tendências de longo prazo e possíveis estratégias adaptativas.

Ciências Sociais Computacionais e Humanidades Digitais

Embora as epistemologias do empirismo de Big Data e da ciência orientada por dados pareçam destinadas a transformar a abordagem de pesquisa adotada nas ciências naturais, da vida, físicas e de engenharia, sua trajetória nas ciências humanas e sociais é menos certa. Essas áreas de conhecimento são altamente diversas em seus fundamentos filosóficos, com apenas alguns estudiosos empregando a epistemologia comum nas ciências. Aqueles que usam o método científico para explicar e modelar os fenômenos sociais, em termos gerais, recorrem às ideias do positivismo (embora possam não adotar tal rótulo; Kitchin, 2006). Esse trabalho tende a se concentrar em informações factuais e quantificadas – fenômenos empiricamente observáveis que podem ser medidos de forma robusta (como contagens, distância, custo e tempo), em oposição a aspectos mais intangíveis da vida humana, como crenças ou ideologia – usando testes estatísticos para estabelecer relações causais e construir teorias e modelos preditivos e simulações. Abordagens positivistas estão bem estabelecidas em economia, ciência política, geografia humana e sociologia, mas são raras nas humanidades. No entanto, dentro dessas disciplinas mencionadas, houve um forte movimento ao longo do último meio século em direção a abordagens pós-positivistas, especialmente na geografia humana e na sociologia.

Para os estudiosos positivistas das ciências sociais, o Big Data oferece uma oportunidade significativa para desenvolver modelos mais sofisticados, de escala mais ampla e mais refinados da vida humana. Não obstante as preocupações com o acesso ao Big Data social e econômico (muito do qual é gerado por interesses privados) e questões como a qualidade dos dados, o Big Data oferece a possibilidade de mudar 'de estudos de sociedades com escassez de dados para estudos ricos em dados; de instantâneos estáticos a desdobramentos dinâmicos; de agregações grosseiras a altas resoluções; de modelos relativamente simples a simulações mais complexas e sofisticadas' (Kitchin, 2014: 3). O potencial existe para um novo era da ciência social computacional que produz estudos com amplitude, profundidade, escala e linhas de tempo muito maiores, e que são inerentemente longitudinais, em contraste com a pesquisa existente em ciências sociais (Lazer et al., 2009; Batty et al., 2012) . Além disso, a variedade, exaustividade, resolução e relacionalidade dos dados, além do poder crescente da computação e da nova análise de dados, abordam algumas das críticas dos estudos positivistas até o momento, especialmente as do reducionismo e do universalismo, fornecendo informações mais refinadas, análise sensível e diferenciada que pode levar em conta o contexto e a contingência e pode ser usada para refinar e ampliar os entendimentos teóricos do mundo social e espacial (Kitchin, 2013). Além disso, dada a extensão dos dados, é possível testar a veracidade de tal teoria em uma variedade de configurações e situações. Em tais circunstâncias, argumenta-se que o conhecimento sobre indivíduos, comunidades, sociedades e ambientes se tornará mais perspicaz e útil no que diz respeito à formulação de políticas e ao tratamento das várias questões enfrentadas pela humanidade.

Para estudiosos pós-positivistas, Big Data oferece oportunidades e desafios. As oportunidades são a proliferação, digitalização e interligação de um conjunto diverso de dados analógicos e não estruturados, muitos deles novos (por exemplo, redes sociais) e muitos dos quais até agora têm sido de difícil acesso (por exemplo, milhões de livros, documentos, jornais, fotografias , obras de arte, objetos materiais, etc., de toda a história que foram transformados em formato digital nas últimas duas décadas por uma série de organizações; Cohen, 2008), e também o fornecimento de novas ferramentas de curadoria de dados, gerenciamento e análises que podem lidar com um grande número de objetos de dados. Conseqüentemente, em vez de se concentrar em um punhado de romances ou fotografias, ou em alguns artistas e seus trabalhos, torna-se possível pesquisar e conectar-se a um grande número de trabalhos relacionados; em vez de focar em um punhado de sites, salas de bate-papo, vídeos ou jornais online, torna-se possível examinar centenas de milhares dessas mídias (Manovich, 2011). Essas oportunidades estão sendo amplamente examinadas no campo emergente das humanidades digitais.

Inicialmente, as humanidades digitais consistiam na curadoria e análise de dados que nasceram digitais e nos projetos de digitalização e arquivamento que procuravam transformar textos analógicos e objetos materiais em formas digitais que pudessem ser organizadas, pesquisadas e submetidas a formas básicas de , análise automatizada ou guiada, como visualizações resumidas de conteúdo (Schnapp e Presner, 2009). Posteriormente, seus defensores foram divididos em dois campos. O primeiro grupo acredita que as novas técnicas de humanidades digitais – contagem, representação gráfica, mapeamento e leitura à distância – trazem rigor metodológico e objetividade

a disciplinas que até então eram assistemáticas e aleatórias em seu foco e abordagem (Moretti, 2005; Ramsay, 2010). Em contraste, o segundo grupo argumenta que, em vez de substituir os métodos tradicionais ou fornecer uma abordagem empirista ou positivista para o estudo das humanidades, as novas técnicas complementam e aumentam os métodos existentes das humanidades e facilitam as formas tradicionais de interpretação e construção de teoria, permitindo estudos de escopo muito mais amplo. para responder a perguntas que seriam praticamente irrespondíveis sem computação (Berry, 2011; Manovich, 2011).

As humanidades digitais não foram universalmente bem-vindas, com detratores afirmando que o uso de computadores como 'máquinas de leitura' (Ramsay, 2010) para realizar 'leitura à distância' (Moretti, 2005) contraria e prejudica os métodos tradicionais de leitura atenta. Culler (2010: 22) observa que a leitura atenta envolve prestar “atenção em como o significado é produzido ou transmitido, em que tipos de estratégias e técnicas literárias e retóricas são empregadas para alcançar o que o leitor considera serem os efeitos da obra ou passagem”. – algo que uma leitura distante não é capaz de realizar. Sua preocupação é que uma abordagem de humanidades digitais promova estudos literários que não envolvam leitura real. Da mesma forma, Trumpener (2009: 164) argumenta que um “modelo estatisticamente orientado de história literária... senso astuto e historicizado de como gêneros e instituições literárias funcionam e ferramentas interpretativas incisivas' (pp. 170-171).

Da mesma forma, Marche (2012) afirma que fatos artísticos culturais, como a literatura, não podem ser tratados como meros dados. Uma escrita não é simplesmente uma ordem de letras e palavras; é contextual e transmite significado e tem qualidades inefáveis. Os algoritmos são muito fracos para capturar e decifrar o significado ou o contexto e, argumenta Marche, tratam "toda a literatura como se fosse a mesma". Ele continua:

> [a] análise algorítmica de romances e artigos de jornal está necessariamente no limite do redutivismo. O processo de transformar literatura em dados remove a própria distinção. Ele remove o gosto. Ele remove todo o refinamento da crítica. Retira o histórico da recepção das obras.

Jenkins (2013) assim conclui:

> o valor das artes, a qualidade de uma peça ou de uma pintura, não é mensurável. Você poderia colocar todos os tipos de dados em uma máquina: datas, cores, imagens, recibos de bilheteria, e nada disso poderia explicar o que é a obra de arte, o que significa e por que é poderosa. Isso requer homem [sic], não máquina.

Para muitos, então, as humanidades digitais estão promovendo uma análise fraca e superficial, em vez de uma visão profunda e penetrante. É excessivamente reducionista e grosseiro em suas técnicas, sacrificando complexidade, especificidade, contexto, profundidade e crítica por escala, amplitude, automação, padrões descritivos e a impressão de que a interpretação não requer conhecimento contextual profundo.

Os mesmos tipos de argumento podem ser usados na ciência social computacional. Por exemplo, um mapa da linguagem dos tweets em uma cidade pode revelar padrões de concentração geográfica de diferentes comunidades étnicas (Rogers, 2013), mas as questões importantes são quem constitui tais concentrações, por que elas existem, quais foram os processos de formação e reprodução, e quais são suas conseqüências sociais e econômicas? Uma coisa é identificar padrões; outra é explicá-los. Isso requer teoria social e profundo conhecimento contextual. Como tal, o padrão não é o ponto final, mas sim um ponto de partida para análises adicionais, que quase certamente exigirão outros conjuntos de dados.

Assim como nas críticas anteriores das ciências sociais quantitativas e positivistas, as ciências sociais computacionais são criticadas pelos pós-positivistas como sendo mecanicistas, atomizantes e paroquiais, reduzindo diversos indivíduos e estruturas sociais complexas e multidimensionais a meros pontos de dados (Wyly, no prelo). Além disso, a análise está repleta de pressupostos de determinismo social, como exemplificado por Pentland (2012): 'o tipo de pessoa que você é é amplamente determinado por seu contexto social, então, se eu puder ver alguns de seus comportamentos, posso inferir o descanse, apenas comparando você com as pessoas da sua multidão'. Em contraste, as sociedades humanas, argumenta-se, são muito complexas, contingentes e confusas para serem reduzidas a fórmulas e leis, com modelos quantitativos que fornecem pouca visão sobre fenômenos como guerras, genocídio, violência doméstica e racismo, e apenas uma visão circunscrita sobre outros sistemas humanos, como a economia, inadequadamente contabilizando o papel da política, ideologia, estruturas sociais e cultura (Harvey, 1972). As pessoas não agem de maneira racional e predeterminada, mas vivem vidas cheias de contradições, paradoxos e ocorrências imprevisíveis. A forma como as sociedades são organizadas e operam varia ao longo do tempo e do espaço e não existe uma forma ótima ou ideal, ou características universais. De fato, existe uma incrível diversidade de indivíduos, culturas e modos de vida em todo o planeta. Reduzir essa complexidade aos assuntos abstratos que povoam os modelos universais é uma violência simbólica ao modo como criamos conhecimento. Além disso, as abordagens positivistas ignoram intencionalmente os aspectos metafísicos da vida humana (relacionados com significados, crenças, experiências) e questões normativas (dilemas éticos e morais sobre como as coisas deveriam ser em oposição a como são) (Kitchin, 2006). Em outras palavras, as abordagens positivistas apenas

concentram-se em certos tipos de questões, às quais procuram responder de uma forma reducionista que aparentemente ignora o que significa ser humano e viver em sociedades e lugares ricamente diversos. Isso não quer dizer que as abordagens quantitativas não sejam úteis – elas obviamente são – mas que suas limitações na compreensão da vida humana devem ser reconhecidas e complementadas com outras abordagens.

Brooks (2013) afirma, assim, que a análise de Big Data luta com o social (as pessoas não são racionais e não se comportam de maneira previsível; os sistemas humanos são incrivelmente complexos, tendo relações contraditórias e paradoxais); lutas com o contexto (os dados são amplamente desprovidos do contexto social, político, econômico e histórico); cria palheiros maiores (consistindo de muito mais correlações espúrias, dificultando a identificação de agulhas); tem dificuldade em lidar com grandes problemas (especialmente sociais e econômicos); favorece memes sobre obras-primas (identifica tendências, mas não necessariamente características significativas que podem se tornar uma tendência); e obscurece valores (dos produtores de dados e daqueles que os analisam e seus objetivos). Em outras palavras, embora a análise de Big Data possa fornecer alguns insights, é preciso reconhecer que eles são limitados em escopo, produzem tipos específicos de conhecimento e ainda precisam de contextualização com relação a outras informações, sejam teorias existentes, documentos de políticas , pequenos estudos de dados, ou registros históricos, que podem ajudar a dar sentido aos padrões evidentes (Crampton et al., 2012).

Além da abordagem epistemológica e metodológica, parte da questão é que muitos Big Data e análises parecem ser gerados sem questões específicas em mente, ou o foco é direcionado pela aplicação de um método ou pelo conteúdo do conjunto de dados, em vez de um pergunta específica, ou o conjunto de dados está sendo usado para buscar uma resposta para uma pergunta que nunca foi projetada para responder em primeiro lugar. Com relação a este último, os dados georreferenciados do Twitter não foram produzidos para fornecer respostas com respeito à concentração geográfica de grupos linguísticos em uma cidade e os processos que conduzem tal autocorrelação espacial. Talvez não devêssemos nos surpreender, então, que ele forneça apenas um instantâneo superficial, embora seja um instantâneo interessante, em vez de percepções profundas e penetrantes nas geografias de raça, idioma, aglomeração e segregação em locais específicos.

Enquanto a maioria dos humanistas digitais reconhece o valor das leituras próximas e enfatiza como as leituras distantes as complementam ao fornecer profundidade e contextualização, as formas positivistas da ciência social se opõem às abordagens pós-positivistas. A diferença entre as humanidades e as ciências sociais a esse respeito é porque as estatísticas usadas nas humanidades digitais são amplamente descritivas – identificando e plotando padrões. Em contraste, as ciências sociais computacionais empregam o método científico, complementando a estatística descritiva com a estatística inferencial que busca identificar associações e causalidade. Em outras palavras, eles são sustentados por uma epistemologia cujo objetivo é produzir modelos estatísticos sofisticados que expliquem, simulem e prevejam a vida humana. Isso é muito mais difícil de conciliar com as abordagens pós-positivistas.

A defesa então se baseia na utilidade e no valor do método e dos modelos, não no fornecimento de análises complementares de um conjunto mais amplo de dados.

Há uma alternativa potencialmente frutífera a essa posição que adota e estende as epistemologias empregadas em GIS crítico e estatística radical. Essas abordagens empregam técnicas quantitativas, estatísticas inferenciais, modelagem e simulação, embora sejam conscientes e abertas em relação às suas deficiências epistemológicas, baseando-se na teoria social crítica para enquadrar como a pesquisa é conduzida, como as descobertas fazem sentido e o conhecimento empregado. Aqui, há o reconhecimento de que a pesquisa não é uma atividade neutra e objetiva que produz uma visão do nada, e que há uma política inerente que permeia os conjuntos de dados analisados, a pesquisa conduzida e as interpretações feitas (Haraway, 1991; Rose, 1997). Como tal, reconhece-se que o pesquisador possui uma certa posição (no que diz respeito ao seu conhecimento, experiência, crenças, aspirações, etc.), que a pesquisa está situada (dentro de debates disciplinares, o cenário de financiamento, políticas sociais mais amplas, etc.), os dados refletem a técnica usada para gerá-los e possuem certas características (relativas à amostragem e estruturas ontológicas, limpeza de dados, completude, consistência, veracidade e fidelidade), e os métodos de análise utilizados produzem efeitos particulares em relação a os resultados produzidos e as interpretações feitas. Além disso, reconhece-se que a forma como a pesquisa é empregada não é ideologicamente neutra, mas é enquadrada de maneiras sutis e explícitas pelas aspirações e intenções dos pesquisadores e financiadores/patrocinadores, e aqueles que traduzem essa pesquisa em várias formas de política, instrumentos e ações. Em outras palavras, dentro de tal epistemologia, a pesquisa realizada é reflexiva e aberta em relação ao processo de pesquisa, reconhecendo as contingências e relacionalidades da abordagem empregada, produzindo assim relatos e conclusões nuançadas e contextualizadas. Tal epistemologia também não impede a complementação da ciência social computacional situada com pequenos estudos de dados que fornecem insights adicionais e ampliadores (Crampton et al., 2012). Em outras palavras, é possível pensar em novas epistemologias que não descartem ou rejeitem a análise de Big Data, mas que empreguem a abordagem metodológica da data-driven science dentro de um enquadramento epistemológico diferente que possibilite

cientistas para obter insights valiosos de Big Data que são situados e reflexivos.

Conclusão

Há pouca dúvida de que o desenvolvimento de Big Data e novas análises de dados oferecem a possibilidade de reenquadrar a epistemologia da ciência, ciências sociais e humanidades, e tal reenquadramento já está ocorrendo ativamente em todas as disciplinas. Big Data e novas análises de dados permitem que novas abordagens para geração e análise de dados sejam implementadas, tornando possível fazer e responder perguntas de novas maneiras. Em vez de buscar extrair insights de conjuntos de dados limitados por escopo, temporalidade e tamanho, o Big Data oferece o contra-problema de lidar e analisar conjuntos de dados enormes, dinâmicos e variados. A solução tem sido o desenvolvimento de novas formas de gerenciamento de dados e técnicas analíticas que contam com aprendizado de máquina e novos modos de visualização.

No que diz respeito às ciências, o acesso a Big Data e novas práticas de pesquisa levaram alguns a proclamar o surgimento de um novo quarto paradigma, enraizado na exploração intensiva de dados que desafia a abordagem dedutiva científica estabelecida. Actualmente, embora seja claro que o Big Data é uma inovação disruptiva, apresentando a possibilidade de uma nova abordagem à ciência, a forma desta abordagem não está definida, propondo-se dois caminhos potenciais que têm epistemologias divergentes – o empirismo, em que os dados podem falar por si mesmos, livres de teoria e ciência orientada por dados que modifica radicalmente o método científico existente, combinando aspectos de abdução, indução e dedução. Dadas as fraquezas nos argumentos empiristas, parece provável que a abordagem baseada em dados acabará por vencer e, com o tempo, à medida que o Big Data se tornar mais comum e novas análises de dados forem avançadas, apresentará um grande desafio para a abordagem baseada em conhecimento estabelecida. método científico. Para acompanhar tal transformação, os fundamentos filosóficos da ciência orientada por dados, no que diz respeito aos seus princípios epistemológicos, princípios e metodologia, precisam ser trabalhados e debatidos para fornecer uma estrutura teórica robusta para o novo paradigma.

A situação nas ciências humanas e sociais é um pouco mais complexa dada a diversidade de seus fundamentos filosóficos, sendo improvável que Big Data e novas análises levem ao estabelecimento de novos paradigmas disciplinares. Em vez disso, o Big Data aprimorará o conjunto de dados disponíveis para análise e permitirá novas abordagens e técnicas, mas não substituirá totalmente os estudos tradicionais de pequenos dados. Isso se deve em parte a posições filosóficas, mas também porque é improvável que Big Data adequado seja produzido e possa ser utilizado para responder a perguntas específicas, necessitando de estudos mais direcionados. No entanto, como Kitchin (2013) e Ruppert (2013) argumentam, Big Data apresenta uma série de oportunidades para cientistas sociais e estudiosos de humanidades, não menos do que grandes quantidades de dados sociais, culturais, econômicos, políticos e históricos muito ricos. Também apresenta uma série de desafios, incluindo um déficit de habilidades para analisar e dar sentido a esses dados, e a criação de uma abordagem epistemológica que permita formas pós-positivistas de ciência social computacional. Um caminho potencial a seguir é uma epistemologia que se inspire em GIS crítico e estatísticas radicais nas quais métodos e modelos quantitativos são empregados dentro de uma estrutura que é reflexiva e reconhece a localização, posicionalidade e política da ciência social que está sendo conduzida, em vez de rejeitá-la. uma abordagem fora de mão. Tal epistemologia também tem utilidade potencial nas ciências para reconhecer e explicar o uso da abdução e criar uma ciência mais reflexiva orientada por dados. Como ilustra esta tentativa de discussão, há uma necessidade urgente de uma reflexão crítica mais ampla sobre as implicações epistemológicas do Big Data e da análise de dados, uma tarefa que mal começou apesar da velocidade da mudança no cenário de dados.

Agradecimentos Evelyn

Ruppert e Mark Boyle forneceram alguns comentários úteis sobre o rascunho inicial deste artigo. A pesquisa para este artigo foi financiada por um Prêmio de Investigador Avançado do Conselho Europeu de Pesquisa, 'A Cidade Programável' (ERC-2012-AdG-323636).

Referências

Amin A e Thrift N (2002) Cidades: Reimaginando o Urbano. Londres: Política.

Anderson C (2008) O fim da teoria: O dilúvio de dados torna o método científico obsoleto. Telegrafado, 23 de junho de 2008. Disponível em: http://www.wired.com/science/discoveries/revista/16-07/pb_theory (acessado em 12 de outubro de 2012).

Batty M, Axhausen KW, Giannotti F, et al. (2012) Cidades inteligentes do futuro. European Physical Journal Special Topics 214: 481–518.

Berry D (2011) A virada computacional: pensando nas humanidades digitais. Culture Machine 12. Disponível em: http://www.culturemachine.net/index.php/cm/article/view/440/470 (acessado em 3 de dezembro de 2012).

Bettencourt LMA, Lobo J, Helbing D, et al. (2007) Crescimento, inovação, escala e ritmo de vida nas cidades. Proceedings of the National Academy of Sciences 104(17): 7301–7306.

Bollier D (2010) A Promessa e o Perigo do Big Data. Instituto Aspen. Disponível em: http://www.aspeninstitu te.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf (acessado em 1º de outubro de 2012).

boyd D e Crawford K (2012) Questões críticas para big data. Informação, Comunicação e Sociedade 15(5): 662–679.

Brooks D (2013) O que os dados não podem fazer. New York Times, 18 de fevereiro de 2013. Disponível em: http://www.nytimes.com/ 2013/02/19/ opinion/brooks-what-data-cant-do.html (acessado em 18 de fevereiro de 2013).

Bryant R, Katz RH e Lazowska ED (2008) Big-data com puting: Criando avanços revolucionários no comércio, ciência e sociedade. In: Iniciativas de Pesquisa em Computação para o Século 21, Associação de Pesquisa em Computação, Ver. 8. Disponível em: http://www.cra.org/ccc/docs/init/ Big_Data.pdf (acessado em 12 de outubro de 2012).

Clark L (2013) Sem perguntas: A empresa de big data mapeia soluções sem intervenção humana. Telegrafado, 16 de janeiro de 2013. Disponível em: http://www.wired.co.uk/news/archive/2013-01/16/ayasdi-big-data-launch (acessado em 28 de janeiro de 2013).

Cohen D (2008) Contribuição para: A promessa da história digital (mesa redonda). Journal of American History 95(2): 452–491.

Constine J (2012) Qual é o tamanho dos dados do Facebook? 2,5 bilhões de conteúdos e 500þ terabytes ingeridos todos os dias, 22 de agosto de 2012. Disponível em: http://techcrunch.com/2012/ 22/08/how-big-is-facebooks-data-2-5-billion-pieces -of content-and-500-terabytes-ingested-every-day/ (acessado em 28 de janeiro de 2013).

Crampton J, Graham M, Poorthuis A, et al. (2012) Além do Geotag? Desconstruindo 'Big Data' e Aproveitando o Potencial da Geoweb. Disponível em: http://www. uky.edu/tmute2/geography_methods/ readingPDFs/ 2012-Beyond-the-Geotag-2012.10.01.pdf (acessado em 21 de fevereiro de 2013).

Crawford K (2013) Os vieses ocultos do big data. Blog da Harvard Business Review. 01 de abril. Disponível em: http://blogs. hbr.org/2013/04/the-hidden-biases-in-big-data/ (acessado em 18 de setembro de 2013).

Cukier K (2010) Dados, dados em todos os lugares. O Economista, 25 fevereiro (acessado em 12 de novembro de 2012).

Culler J (2010) A proximidade da leitura atenta. Boletim ADE 149: 20–25.

Dodge M e Kitchin R (2005) Códigos da vida: Códigos de identificação e o mundo legível por máquina. Meio Ambiente e Planning D: Society and Space 23(6): 851–881.

Dyche J (2012) Big data 'Eurekas!' não simplesmente aconteça. Blog da Harvard Business Review. 20 de novembro. Disponível em: http://blogs.hbr.org/cs/2012/11/eureka_doesnt_just_acontecer.html (acessado em 23 de novembro de 2012).

Floridi L (2012) Big data e seu desafio epistemológico. Filosofia e Tecnologia 25(4): 435–437.

Gould P (1981) Deixando os dados falarem por si. Anais da Associação de Geógrafos Americanos 71(2): 166–176.

Han J, Kamber M e Pei (2011) Mineração de Dados: Conceitos e Técnicas, 3ª ed. Waltham: Morgan Kaufmann.

Haraway D (1991) Simians, Cyborgs and Women: The Reinvention of Nature. Nova York: Routledge.

Harvey D (1972) Justiça Social e a Cidade. Oxford: Blackwell.

Hastie T, Tibshirani R e Friedman J (2009) Os Elementos da Aprendizagem Estatística: Mineração de Dados, Inferência e Previsão, 2ª ed. Nova York: Springer.

Hey T, Tansley S e Tolle K (2009) Jim Gray em eScience: Um método científico transformado. In: Hey T, Tansley S e Tolle K (eds) O Quarto Paradigma: Descoberta Científica Intensiva em Dados. Redmond: Microsoft Research, pp. xvii–xxxi.

Jenkins T (2013) Não conte com big data para obter respostas. The Scotsman, 12 de fevereiro de 2013. Disponível em: http://www. scotsman.com/the-scotsman/opinion/comment/tiffany jenkins-don-t-count-on-big-data-for-answers-1-2785890 (acessado em 11 de março de 2013).

Kelling S, Hochachka W, Fink D, et al. (2009) Data-intensive Science: Um novo paradigma para estudos de biodiversidade. BioScience 59(7): 613–620.

Kitchin R (2006) Geografia positivista e ciência espacial. In: Aitken S e Valentine G (eds) Abordagens em Geografia Humana. Londres: Sage, pp. 20–29.

Kitchin R (2013) Big data e geografia humana: oportunidades, desafios e riscos. Diálogos em Geografia Humana 3(3): 262–267.

Kitchin R (2014) A cidade em tempo real? Big data e inteligência urbanismo. GeoJournal 79: 1–14.

Kuhn T (1962) A Estrutura das Revoluções Científicas. Chicago: University of Chicago Press.

Laney D (2001) Gerenciamento de dados 3D: Controlando volume, velocidade e variedade de dados. Metagrupo. Disponível em: http:// blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume Velocity-and-Variety.pdf (acessado em 16 de janeiro de 2013).

Lazer D, Pentland A, Adamic L, et al. (2009) Ciência social computacional. Ciência 323: 721–733.

Lehning M, Dawes N, Bavay M. et al. (2009) Instrumentando a Terra: redes de sensores de próxima geração e ciência ambiental. In: Hey T, Tansley S e Tolle K (eds) O Quarto Paradigma: Descoberta Científica Intensiva em Dados. Redmond: Microsoft Research, pp. 45–51.

Lehrer J (2010) Um físico resolve a cidade. New York Times, 17 de dezembro. Disponível em: http://www.nytimes.com/2010/12/19/magazine/ 19Urban_West-t.html (acessado em 23 de dezembro de 2013).

Leonelli S (2012) Introdução: Dar sentido à pesquisa baseada em dados nas ciências biológicas e biomédicas. Estudos em História e Filosofia das Ciências Biológicas e Biomédicas 43(1): 1–3.

Loukides M (2010) O que é ciência de dados? O'Reilly Radar, 2 de junho de 2010. Disponível em: http://radar.oreilly.com/2010/06/what-is-data-science.html (acessado em 28 de janeiro de 2013).

Manovich L (2011) Tendências: as promessas e os desafios do big data social. Disponível em: http://www. manovich.net/DOCS/ Manovich_trending_paper.pdf (acessado em 9 de novembro de 2012).

Marche S (2012) A literatura não é um dado: Contra as humanidades digitais. Los Angeles Review of Books, 28 de outubro de 2012. Disponível em: http://lareviewofbooks.org/article.php? id¼1040&fulltext¼1 (acessado em 4 de abril de 2013).

Marz N e Warren J (2012) In: MEAP (ed.), Big Data: Princípios e Melhores Práticas de Sistemas de Dados em Tempo Real Escaláveis. Westhampton: Manning.

Mayer-Schonberger V e Cukier K (2013) Big Data: uma revolução que mudará a forma como vivemos, trabalhamos e pensamos. Londres: John Murray.

Miller HJ (2010) A avalanche de dados está aqui. Não deveríamos estar cavando? Journal of Regional Science 50(1): 181–201.

Moretti F (2005) Graphs, Maps, Trees: Abstract Models for a História Literária. Londres: Verso.

Open Data Center Alliance (2012) Guia do Consumidor de Big Data. Open Data Center Alliance. Disponível em: http://www. opendatacenteralliance.org/docs/Big_Data_Consumer_ Guide_Rev1.0.pdf (acessado em 11 de fevereiro de 2013).

Pentland A (2012) Reinventando a sociedade após o big data. Edge, 30 de agosto de 2012. Disponível em: http://www. edge.org/conversation/ reinventing-society-in-the-wake-of big-data (acessado em 28 de janeiro de 2013).

Porway J (2013) Você não pode simplesmente hackear seu caminho para a mudança social. Harvard Business Review Blog, 7 de março de 2013. Disponível em: http://blogs.hbr.org/cs/2013/03/you_cant_ just_hack_your_way_to.html (acessado em 9 de março de 2013).

Prensky M (2009) H. sapiens digital: De imigrantes digitais e nativos digitais à sabedoria digital. Inovar 5(3). Disponível em: http://www.innovateonline.info/index.php? view¼article&id¼705 (acessado em 12 de outubro de 2012).

Ramsay S (2010) Máquinas de leitura: Rumo a uma crítica algorítmica. Campanha: University of Illinois Press.

Ribes D e Jackson SJ (2013) Homem da mordida de dados: O trabalho de sustentar o estudo de longo prazo. In: Gitelman L (ed.) 'Dados brutos' é um oxímoro. Cambridge, MA: MIT Press, pp. 147–166.

Rogers S (2013) Mapeou as línguas do Twitter de Nova York. The Guardian, 21 de fevereiro de 2013. Disponível em: http:// www.guardian.co.uk/news/datablog/interactive/2013/feb/21/twitter-languages-new-york-mapped (acessado em 3 de abril de 2013).

Rose G (1997) Situando conhecimentos: posicionalidade, reflexividades e outras táticas. Progress in Human Geography 21(3): 305–320.

Ruppert E (2013) Repensando as ciências sociais empíricas. Dialogues in Human Geography 3(3): 268–273.

Schnapp J e Presner P (2009) Manifesto de Humanidades Digitais 2.0. Disponível em: http://www.humanitiesblast. com/manifesto/ Manifesto_V2.pdf (acessado em 13 de março de 2013).

Seni G e Elder J (2010) Ensemble Methods in Data Mining: Melhorando a precisão por meio da combinação de previsões. San Rafael: Morgan e Claypool.

Siegel E (2013) Análise preditiva. Hoboken: Wiley.

Steadman I (2013) Big data e a morte do teórico. Wired, 25 de janeiro de 2013. Disponível em: http://www.wired. co.uk/ news/archive/2013-01/25/big-data-end-of-theory (acessado em 30 de janeiro de 2013).

Strasser BJ (2012) Ciências orientadas por dados: de cabines maravilhosas a bancos de dados eletrônicos. Estudos em História e Filosofia das Ciências Biológicas e Biomédicas 43: 85–87.

Strom D (2012) Big data torna as coisas melhores. Slashdot, 3 de agosto. Disponível em: http://slashdot.org/topic/bi/big data-makes-things-better/ (acessado em 24 de outubro de 2013).

Trumpener K (2009) Resposta crítica I. Paratexto e sistema de gênero: Uma resposta a Franco Moretti. Critical Inquiry 36(1): 159–171.

Wyly E (no prelo) (pós)positivismo automatizado. Geografia Urbana.

Zikopoulos PC, Eaton C, DeRoos D, et al. (2012) Entendendo o Big Data. Nova York: McGraw Hill.