O fim da teoria: o confronto entre a pesquisa orientada por dados e a pesquisa orientada por hipóteses

The end of theory: the confrontation between data-driven research and hypothesis-driven research

Luís Fernando Sayão [*]

Luana Farias Sales [**]

RESUMO

A ciência contemporânea e seus fundamentos metodológicos têm sido impactados pelo fenômeno do big data, que proclama que na era dos dados medidos em petabytes, de supercomputadores e sofisticados algoritmos, o método científico está obsoleto e que as hipóteses e modelos estão superados. As estratégias do big data científico confia em estratégias de análises computacionais de massivas quantidades de dados para revelar correlações, padrões e regras que vão gerar novos conhecimentos, que vão das ciências exatas até as ciências sociais, humanidade e cultura, delineando um arquétipo de ciência orientada por dados. O presente ensaio coloca em pauta as controvérsias em torno da ciência orientada por dados em contraposição à ciência orientada por hipóteses, e analisa alguns dos desdobramentos desse embate epistemológico. Para tal, tomo como metodologia os escritos de alguns autores mais proximamente envolvidos nessa questão.

Palavras-chave: Big Data; Método Cientifico; Ciência Orientada por Dados;

ABSTRACT

Contemporary science and its methodological foundations have been impacted by the big data phenomenon that proclaims that in the age of data measured in petabytes, supercomputers and sophisticated algorithms the scientific method is obsolete and that the hypotheses and models are outdated.The strategies of the scientific big data rely on computational analysis strategies of massive amounts of data to reveal correlations, patterns and rules that will generate new knowledge, ranging from the exact sciences to the social sciences, humanity and culture, outlining an archetype of data-driven science. The present essay addresses the debates around data-driven science as opposed to hypothesis-oriented science and analyzes some of the ramifications of this epistemological confrontation. For this, the writings of some authors who are more closely involved in this question are taken as methodology.

Keywords: Big Data; Scientific Method; Data-Driven Science; Hypothesis-Driven Science.

[*] Doutor em Ciência da Informação pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict). Tecnologista em Ciência e Tecnologia da Comissão Nacional de Energia Nuclear (CNEN). Professor do Programa de Pós-Graduação em Biblioteconomia da Universidade Federal do Estado do Rio de Janeiro (Unirio). Endereço: Rua General Severiano, nº 90, Botafogo, CEP 22290-040, Rio de Janeiro, RJ. Telefone: (21) 2173-2028. E-mail: lsayao@cnen.gov.br.

[**] Doutora em Ciência da Informação pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Professora do Programa de Pós-Graduação em Ciência da Informação do Ibict. Endereço: Rua Lauro Muller, 455, 4° andar, sala 408, CEP: 22.290-160. Telefone: (21) 97112-7411 / (21) 3873-9450. E-mail: luanasales@ibict.br.

Ciência Orientada por Hipóteses.

INTRODUÇÃO

Em 2008 o ex-editor da revista *Wired,*[1] Chris Anderson, publicou um artigo cujo título já antecipava as tensões e as discussões contundentes que ele desencadearia no mundo da pesquisa científica: "The end of theory: the data deluge makes the scientific method obsolete", que pode ser traduzido para português como: "O fim da teoria: o dilúvio de dados torna o método científico obsoleto". O que Anderson argumenta de forma provocativa no seu texto é que na era da informação medida em *petabytes*,[2] de supercomputadores, algoritmos e ferramentas estatísticas sofisticadas, o que conta realmente é a massiva quantidade de dados, que abre a possibilidade de encontrar informações, que podem ser transformadas em novos conhecimentos. Análises de massivas quantidades de dados vão trazer à tona correlações, conexões, padrões e regras inéditas e muitas vezes surpreendentes, que só se revelam diante do complexo aparato computacional. "Não mais teorias e hipóteses, não mais discussões se os resultados experimentais refutam ou confirmam as hipóteses originais", resume Mazzochi (2015, p.1).

O ensaio de Anderson (2008) tem o mérito único de colocar em pauta debates mais precisos e fundamentados sobre como a disponibilidade massiva de dados associados aos novos métodos computacionais de análise desafia o percurso secular da metodologia cientifica, e a partir desse ponto provoca um deslocamento paradigmático – cuja extensão se desconhece – que se observa atualmente em muitos domínios disciplinares que vão das ciências exatas às artes.

Nesse campo povoado de controvérsias, no qual a forma de fazer pesquisa científica, que atravessa séculos, é substituída por computadores, estatísticas e algoritmos, Fulvio Mazzocchi (2015) coloca algumas questões importantes que merecem ser investigadas do ponto de vista epistemológico. São elas: a pesquisa orientada por dados é um modo genuíno de produção de conhecimento, ou é, simplesmente, uma ferramenta sofisticada para identificar informações potencialmente úteis? Dada a quantidade de dados científicos disponíveis atualmente, é possível desconsiderar o papel das hipóteses e das suposições teóricas? Esse novo modo de coletar informações suplanta o antigo modo de fazer ciência?

Em outro plano, a disponibilidade de massiva quantidade de dados resultantes da atuação pervasiva das empresas que comandam as mídias sociais, somado à oferta de dados governamentais – inflados pelas legislações globais de acesso aberto –, além de outras fontes, desperta o desejo implícito da quantificação das ciências sociais e humanidades, expresso ironicamente por Bruno Latour (2009, p.147) quando discute as ideias de quantificação de Gabriel Tarde. "Números, números, números. A sociologia tem estado obcecada pelo objetivo de se tornar uma ciência quantitativa".

O presenta ensaio tem como objetivo colocar em pauta as discussões em torno das novas metodologias de pesquisa trazidas pelo fenômeno do *big data* científico e o embate entre a ciência orientada por dados e a ciência orientada por hipóteses, e analisar brevemente os problema e limitações do *big data* no âmbito das chamadas

[1] Disponível em: <https://www.wired.com/>. Acesso em: 28 maio 2019.

[2] *Petabytes* é um múltiplo da unidade de informação byte cujo símbolo é PB. Corresponde a 10 [15] *bytes*, 1.000 *terabytes*, 100.000 *gigabytes*.

humanidades digitais. Para tal, toma como recurso metodológico as ideias de alguns autores que se envolveram mais proximamente na discussão sobre o fim da teoria preconizado por Anderson (2008).

ANTECEDENTES E DEFINIÇÕES

Configurando o cenário para fundamentar os seus pressupostos em torno da "Era do *Petabyte*", Anderson (2008) relembra que há 60 anos os computadores digitais tornaram a informação legível por máquina; há 20 anos a internet com seus tentáculos a torna alcançável; há 10 anos a primeira máquina de busca a integra em uma única base de dados. Hoje o Google e empresas similares estão tratando esse massivo corpo de informação como um laboratório da condição humana. Este fato requer uma perspectiva na qual não haja limites para as possibilidades da imensidão de dados, e exige que eles possam ser analisados e visualizados na sua totalidade. Isto nos leva a considerá-los, em primeira instância, matematicamente, e seus possíveis contextos estabelecidos posteriormente

Este é um mundo em que a imensidão dos dados, aliada aos poderosos algoritmos matemáticos, substitui qualquer outra ferramenta. Fora com todas as teorias do comportamento humano, da linguística à sociologia. "Esqueça taxonomia, ontologia e psicologia. [...] Com dados suficientes, os números falam por si. A era do *petabytes* é diferente porque mais é diferente", resume Anderson (2008, p.3) na sua apologia ao fim das teorias científicas.

Mas antes de tudo, é preciso considerar que o protagonismo dos dados não é um fenômeno unicamente do nosso tempo. O governo, as empresas, a pesquisa científica, bem como vários outros segmentos da sociedade sempre lançaram mão de dados e informações para tomar decisões, redirecionar seus empreendimentos, fundamentar suas descobertas. Porém, nas últimas décadas, toda a sociedade experimenta um fenômeno inédito que tem como ponto de inflexão uma mudança na curva de disponibilidade de informação: da escassez à extrema abundância de dados. Isso muda tudo, é "uma revolução que irá transformar como nós vivemos, trabalhamos e pensamos", anunciam Vikor Mayer-Schönberger e Kenneth Cukier (2013) logo no subtítulo do seu livro.

Esses autores justificam o fenômeno argumentando que por séculos nós sempre coletamos, analisamos e processamos somente informações preciosas e de alta qualidade. Isto acontecia pelo alto custo e pela complexidade, e também pela inexistência de dispositivos tecnológicos e metodologias para gerar e processar grandes quantidades de dados. Nesse cenário rarefeito de dados, a confiança era depositada nos dados mais "limpos" tratados quase individualmente, como eram – e ainda são – as bases de dados factuais e numéricas, posto que os dados somente podiam ser pinçados em pequenas porções.

Ultrapassado o século XX, os avanços das tecnologias digitais, dos computadores pessoais e da internet tornaram possíveis para um amplo espectro de pessoas – incluindo pesquisadores, profissionais de *marketing*, agências governamentais, instituições de ensino e indivíduos motivados – produzir, compartilhar, interagir e organizar grandes quantidades de dados, usando ferramentas de *software* padronizadas e *notebooks*. "Conjunto de dados que estavam obscuros e difíceis de gerenciar [...] estão agora sendo agregados e tornados facilmente acessíveis para qualquer um que tenha interesse, independentemente de sua qualificação", resumem Boyd e Crawford (2012, p. 604).

Hoje o termo "*big data*" é usado com bastante frequência nas mídias populares, negócios, na ciência da computação e na indústria de computadores, referindo-se às transformações desencadeadas pela o aumento da capacidade tecnológica para capturar, armazenar, processar e compreender quantidades massivas de dados. Partindo do fato que as coleções de dados crescem indefinidamente, a capacidade para encontrar respostas fundamentais está mudando a ciência, medicina, tecnologia, educação e negócios. Como um fenômeno ainda considerado novo e em constante mutação, as definições ainda estão em fase de consolidação e vão variando conforme o domínio de aplicação e as intenções de seus praticantes.

Boyd e Crawford (2012) analisam *big data* de uma forma compreensiva que se ajusta aos objetivos das discussões do presente estudo. Eles o caracterizam como um fenômeno tecnossocial, que se realiza no espaço de interação entre tecnologia, análise e mitologia: a *tecnologia*, maximizando o poder computacional e a precisão dos algoritmos para coletar, analisar, conectar e comparar grandes conjuntos de dados; a *análise*, identificando, em grandes conjuntos de dados, padrões, relações, conexões e regras que permitem fazer afirmações; e a *mitologia*, espalhando a crença de que grandes conjuntos de dados oferecem uma alta forma de inteligência e conhecimento que podem gerar *insights* que eram anteriormente impossíveis, tudo isso com uma aura de verdade, objetividade e acurácia.

O que fica evidente sob todos os pontos de observação é que o volume de dados gerados pelos sistemas tecnossociais é excepcional. Entretanto, há bases para se argumentar que isso não é o que define a característica básica dessa nova ecologia informacional. O *big data* não se refere unicamente a grandes coleções de dados e a ferramentas e procedimentos para manipulá-los e analisá-los, mas é também uma mudança fundamentada na computação de pensar e pesquisar cientificamente. Uma possível evidência desse fato é que algumas das coleções de dados tipicamente enquadradas como *big data*, como os dados do Twitter, não são maiores que os *data sets* gerados ou coletados por sistemas anteriores – não considerados *big data* –, como os dados dos censos, exemplificam Boyd e Crawford (2012, p. 663), que concluem: "*Big data* tem menos a ver com o grande volume de dados e mais com a capacidade de pesquisar, agregar e cruzar grandes conjuntos de dados".

Quando consideramos esse fenômeno pelos vieses das ciências sociais, cultura e humanidades, fica claro que o volume de dados gerados, processados e consumidos pelas ciências exatas é extraordinariamente maior do que nesses domínios. Nas disciplinas caracterizadas como *big science* – fortemente conectado aos padrões do *big data* científico –, o regime de distribuição, compartilhamento e acumulação de coleções de dados é determinado pelas idiossincrasias de cada domínio disciplinar (SAYÃO; SALES, 2019). Por exemplo, nas ciências exatas, existem disciplinas híbridas formadas essencialmente por bases de dados, como bioinformática e astroinformática, cuja sofisticação das análises computacionais desempenha um papel crucial na geração de novos conhecimentos (THE ROYAL SOCIETY, 2012); o Grande Colisor de Hádrons do Cern[3] – na busca incansável pelo bóson de Higgs – gera mais de 600 milhões de colisões por segundo e distribui, no instante que são gerados, 15 *petabytes* (15 milhões de *gigabytes*) por ano, por meio de computação em grade, para centenas de centros de processamento de dados espalhados pelo mundo, incluindo o nosso vizinho, o Centro Brasileiro de Pesquisas Físicas. Mesmo encontrando novos padrões que podem gerar hipóteses inéditas para a física de

[3] Cern: Conseil Européen pour la Recherche Nucléaire. Disponível em: <https://home.cern/>.

partículas, "a descoberta do bóson de Higgs não é uma pesquisa orientada por dados. Os experimentos de colisão são na maioria orientados por predições teóricas", alerta Mazzocchi (2015, p.7)

Porém, o redimensionamento das mídias sociais por volta dos anos 2000 alavancou oportunidades para estudar os processos e as dinâmicas sociais e culturais de formas absolutamente novas. "Pela primeira vez nós podemos seguir imaginação, opiniões, ideias e sentimentos de milhões de pessoas" (MANOVICH, 2011, p.2), extrapolando os limites artificias da amostragem. O fenômeno do *big data* reconfigura o panorama dessas áreas, posto que os pesquisadores podem trabalhar sobre conteúdos nascidos digitais gerados pelas próprias fontes de estudos, como os bilhões de fotos, imagens, vídeos, comentários conversas, *blogs*, *tweets* e trajetórias e rastros nos espaços físicos e virtuais.

Neste momento, há bases de dados volumosas de materiais digitais que podem ser usadas por acadêmicos em humanidades e ciências sociais. Esses imensos estoques digitais vão de livros, jornais e música até dados transacionais como buscas na *web*, dados de sensores ou registros de chamadas por celulares. "Como o mundo se torna crescentemente digital, novas técnicas serão necessárias para pesquisar, analisar e compreender esses materiais cotidianos", conclui Manovich (2011, p.1).

Além de tudo, *big data,* com a sua visão avassaladora e totalizante, oferece para as ciências sociais e para as disciplinas humanísticas um argumento forte para a sua reivindicação de *status* de ciências quantitativa e detentora de métodos objetivos. Isso acontece no momento em que torna muitos de seus espaços de pesquisa perfeitamente quantificáveis, mesmo quando o que é quantificável não possua necessariamente o valor de uma verdade objetiva.

Entre um extremo e outro – das ciências duras às ciências sociais e humanidades –, hoje muito mais cientistas da computação estão trabalhando com grandes *data sets*, configurando uma nova área de estudo por eles denominada de "computação social" (do inglês *social computing*), que visa à facilitação por meio de computação de estudos sociais e de dinâmicas sociais humanas bem como o planejamento e uso das tecnologias de informação e comunicação que considerem o contexto social (KITCHIN, 2014).

A CIÊNCIA ORIENTADA POR DADOS *VERSUS* A CIÊNCIA ORIENTADA POR HIPÓTESES: UM CONFRONTO EPISTEMOLÓGICO

"Historicamente a descoberta científica tem sido guiada pelo método científico que remonta a tempos ancestrais e envolve uma abordagem filosófica e prática da ciência" (SCHIMITT et al., 2015, p.1). O método científico compreende um processo contínuo de formular hipóteses testáveis por meio de experimentos e análises, cujos resultados fundamentam a rejeição, aceitação a reformulação da hipótese. Essa forma de procedimento tem sido empregada por séculos e é aceita nas sociedades ocidentais como a mais confiável forma de produzir conhecimento científico robusto.

Porém, nestes últimos tempos, as tecnologias digitais com seu poder avassalador de transformação reconfiguram a trajetória secular de dois pilares monumentais dos processos científicos: a comunicação e o método científico. O poder dos modernos computadores permite que relacionamentos e padrões altamente complexos e imperceptíveis a "olho nu" sejam identificados em ambientes povoados por grandes bases de dados, e a partir daí sejam formuladas novas hipóteses. Nesses novos ambientes de pesquisa, cada vez mais se reconhece que a computação não precisa se

limitar meramente a apoiar as formas tradicionais de conduzir uma investigação científica, mas pode mudar fundamentalmente o desenvolvimento e a forma de gerar conhecimento de determinadas disciplinas (THE ROYAL SOCIETY, 2012).

A contemporaneidade e o centro de ruptura desencadeado pelo fenômeno estão localizados no protagonismo dos dados e de sua prevalência sobre o fluxo tradicional de pesquisa. Os dados vêm em primeiro lugar: as hipóteses, modelos e contextualização vêm depois, retroalimentadas pelos *insights* reconhecidos na imensidão de dados. No âmbito da eScience, esse ponto de inflexão é catalisado por novos métodos, instrumentos, ferramentas e escalas, e está apoiado pelo vertiginoso progresso alcançado pelas tecnologias digitais: redes, processamento, armazenamento e, sobretudo, pela simulação por meio de *software* (RODRIGUES; SARAIVA, 2010). O fato de o mundo real, objeto de pesquisa, ser substituído por programas de computadores, consumidores e geradores de dados, muda muita coisa.

Nesse contexto de mudanças vertiginosas, as metodologias computacionais de análise de dados estão rapidamente criando condições para a adoção de novos enfoques para a investigação científica que não estão em perfeita harmonia com o método científico tradicional ou diferem dele radicalmente. São exemplos dessas novas abordagens as análises exploratórias de coleções de dados não estruturados, mineração de dados, modelagem por computador, simulação interativa e realidade virtual, entre muitas outras estratégias (SCHMITT et al., 2015; GUO, 2013). As abordagens metodológicas e a geração de novos conhecimentos delineiam um verdadeiro *big data* científico, baseado em três ações críticas sobre grandes conjuntos de dados: captura, curadoria e análise.

Mas o conceito de big data caminha para permear toda a sociedade. Os defensores do *big data* científico consideram que as suas abordagens na procura por novos conhecimentos irão mudar não somente os fluxos de trabalho dos laboratórios, mas também o modo como conduzimos nossos empreendimentos e o cotidiano de trabalho, lazer e os padrões comportamentais. Partindo desses pressupostos, Mayer-Schönbergerand e Cukier (2013) apontam três inovações fundamentais do *big data* que se aplicam de forma abarcante em toda a sociedade. Porém, se transpostas para o mundo da pesquisa, essas inovações criam um ponto de ruptura na metodologia científica tradicional. São elas: abundância, diversidade e correlações. Abundância de dados irá garantir que as múltiplas faces dos problemas complexos poderão ser investigadas oferecendo uma visão abrangente, em vez de uma visão focada em uma porção aleatória do problema, reduzindo a importância da amostragem e de seus limites reducionistas. A segunda inovação, a diversidade de dados reduz a ansiedade por exatidão. Em vez de buscar resultados precisos sob condições controladas e simplificadas – como são os laboratórios e modelos –, os cientistas estão tendendo a encontrar na imensidão de dados um espelhamento da complexidade da natureza. A terceira inovação, considerada a mais importante, é a forte ênfase na correlação entre fenômenos ou coisas ou ainda variáveis matemáticas e estatísticas. É a primazia das correlações sobre a compreensão das causas, dos mecanismos internos.

A ideia corrente de que nenhuma conclusão pode ser tirada da correlação entre dois elementos, e que é necessário identificar com confiança os mecanismos internos que os conectam por meio, por exemplo, de modelos teóricos (SAYÃO, 2001), está superada.

Anderson (2008, p.3) defende que, em face da massiva quantidade de dados, o enfoque tradicional da ciência baseado em hipótese, modelo, experimento, está se

tornando obsoleto. "*Petabytes* nos permitem dizer: correlação é suficiente. Nós podemos parar de procurar modelos. Nós podemos analisar os dados sem hipóteses sobre o que eles podem revelar". Correlação supera, nessa perspectiva, a causa. Desse modo, a ciência pode avançar mesmo sem modelos coerentes, teorias unificadas e explicações mecanicistas. O processo de construção de modelos – de forma inversa aos procedimentos tradicionais – é reorientado pela massiva quantidade de dados e menos dependente de proposições teóricas e hipóteses.

O que se observa é que as técnicas de análise de dados ajudam o pesquisador a lidar com a assustadora complexidade dos domínios científicos atuais, como genômica, ecologia, astronomia e interações sociais, entre outros. Isso acontece especialmente quando grandes escalas temporais e espaciais são envolvidas. A mineração de dados, por exemplo, está continuamente aumentando a capacidade dos pesquisadores de visualizarem e identificarem padrões relevantes. O objetivo é descobrir coisas desconhecidas e inesperadas, e observar relações e conexões esperadas previamente ou não entre os diversos elementos da base de dados. Esses procedimentos não são baseados em hipóteses, e devem ser, o máximo possível, independentes de modelos (MAZZOCCHI, 2015).

Mayer-Schönberger e Cukier (2013 p.14) enfatizam também a correlação como essencial nas novas análises. Eles ilustram suas ideias por meio do seguinte exemplo hipotético: se o exame de milhões de prontuários médicos de pacientes de câncer que tomam uma combinação de aspirina com suco de laranja mostram que a doença regride, a causa exata da recuperação da saúde, argumentam os autores, pode ser menos importante do que o fato de os pacientes sobreviverem. "As correlações podem não manifestarem precisamente *porque* alguma coisa está acontecendo, mas elas nos alertam sobre *o que* está acontecendo. Em muitos casos isso é suficiente", defendem os autores. "Entretanto, em muitos casos, compreender o *porquê* é crucial para se alcançar o nível de confiança que permita aplicações práticas e para se fazer previsões confiáveis", rebate Mazzocchi (2015, p.5), com aval de diversos outros pesquisadores como Pigliucci (2009, p.1): "Mas, se nos pararmos de construir modelos e hipóteses, estamos realmente fazendo ciência?". Ciência, diferentemente de propaganda, não objetiva descobrir padrões – embora isso certamente faça parte do processo –, objetiva a descobrir explicações para esses padrões. E completa: "sem modelos matemáticos ou conceituais, dados são apenas ruídos". Mazzocchi (2011) complementa: "Eu não acredito na presumida neutralidade dos números ou na tese de que as correlações se tornam mais importante do que as causas". Os números não falam por si próprios.

Um centro de discussões é a validade da amostragem nos processos do *big data* científico cujos enfoques totalizantes desprezam as técnicas de seleção de amostras representativas. Comentando sobre o livro de Mayer-Schönberger e Cukier (2013), Naimi e Westreich (2014, p.1) registram a seguinte observação: "Em muitas situações os autores se referem à amostragem como um obstáculo desatualizado para as descobertas [científicas]". A tese da "não teoria" desconhece o fato de que a ciência não coleta dados randomicamente, e que os experimentos são projetados e conduzidos dentro de escopos teóricos, metodológicos e instrumentais bem definidos.

Na medida em que o exame dos dados emerge de percursos *bottom-up* baseados em processos indutivos e cálculos estatísticos, nesse cenário aparentemente nenhuma teoria é necessária. Os padrões surgem a partir dos dados e irão fornecer hipóteses de pesquisa sobre os processos subjacentes que produziram a observação. "Neste sentido, o enfoque computacional pode ser visto como um gerador de hipóteses",

em contraste o com o fluxo de formulação de hipótese e teste característico da ciência clássica.

Os defensores do *big data* reacendem as discussões sobre os limites do pensamento dedutivo – no qual o conhecimento científico e os experimentos podem estar baseado em noções preconcebidas, como hipóteses, e não em dados experimentais. O triunfo do pensamento indutivo, representado pelas análises de Kepler sobre os dados observacionais exaustivamente levantados por Tycho Brahe, que resultaram na publicação das leis do movimento planetário, parece representar melhor esse momento.

O motor do enfoque *big data* parece estar no uso de algoritmos indutivos. Nesse sentido, os seus percursos metodológicos renovam a primazia do raciocínio indutivo na forma de empirismo baseado em tecnologia, e inspira uma visão de futuro no qual a mineração de dados automatizados levará diretamente a novas descobertas. De acordo com essa visão, o novo modo de criar conhecimento com "hipóteses neutras" vai substituir a tradicional pesquisa orientada por hipótese.

No domínio específico dos estudos qualitativos sociais, a interação pessoal sempre foi determinante desde o princípio; neste século, entretanto, a capacidade teórica de se poder acessar a totalidade dos dados capturados por tecnologias digitais de traços e registros de atividades humanas abre novas possibilidades. Esse poderio tecnológico pode ser uma ajuda significante na compreensão de muitos fenômenos, mas não uma forma de substituição das metodologias tradicionais, conforme argumentam uma parcela de pesquisadores.

Não importa o quanto sejam boas as fontes de dados, os cientistas de dados, os algoritmos, as ferramentas de análise de dados, eles nunca chegarão aos mesmos *insights* e compreensões da realidade e das interações estudadas. Manovich (2015) defende que haja um equilíbrio de visões e que num cenário hipotético, pesquisadores e cientistas da computação acessam diferente tipos de dados, que os possibilitam dirigir diferentes questões, observar diferentes padrões e chegar a diferentes *insights* (MANOVICH, 2015, p.7). "As questões sobre o que pode ser descoberto e compreendido por meio da análise computacional de dados sociais e culturais em contraste com os métodos qualitativos tradicionais são particularmente importantes para as humanidades digitais" (MANOVICH, 2015, p.8).

ALGUNS PROBLEMAS

As tensões que se instalam entre as utopias e distopias do *big data* povoam um amplo território, ao mesmo tempo mítico e real. Essa dualidade vai do estabelecimento de uma nova forma de inteligência, que se manifesta por métodos inéditos de interrogar a natureza, a sociedade, a história, a arte e a cultura e tudo mais e chegar a novos *insights;* até a concepção de formas de controle corporativo e governamental que são mais velozes do que o desenvolvimento de códigos éticos e legais. Porém, é necessário investigar com um grau a mais de profundidade as limitações econômicas, tecnológicas e políticas de acesso às coleções massivas de dados (in)disponíveis especialmente para as áreas de ciências sociais.

- Quem cria, quem coleta e quem analisa os dados
- A explosão de dados e o advento da análise de dados por computador – como pontos-chave nas abordagens informacionais científicas e econômicas das sociedades contemporâneas – criam um novo princípio

de divisão: "aqueles que criam dados (seja conscientemente ou deixando "pegadas digitais"); aqueles que têm os meios de coletá-los; e aqueles que têm a *expertise* de analisá-los" (MANOVICH, 2011, p.10). O primeiro grupo inclui quase todo mundo que usa a *web*, redes sociais e telefones celulares; o segundo grupo é menor; e o terceiro grupo é ainda muito menor. Nós podemos nos referir a essa estratificação como uma nova classe de pessoas e organizações no contexto de uma "sociedade *big data*", conclui o autor. Essa divisão por classes de geração, coleta e análise pode ser um ponto de partida para inúmeras discussões que incluem *expertise* tecnológica, acesso aos dados.

- **Domínio da tecnologia**: É necessário ter clareza sobre quais *expertises* os humanistas digitais necessitam para obter vantagens da nova escala dos dados humanos. Embora as API's públicas disponibilizadas pelas empresas detentoras de bases de dados sociais não sejam algo tecnologicamente complicado, os projetos de pesquisa de larga escala são geralmente desenvolvidos por pesquisadores da área de ciência da computação.
- **Acesso limitado aos dados:** Muito do entusiasmo em relação ao *big data* tem origem na percepção de que é sempre possível ter acesso fácil à massiva quantidade de dados. Isto é parcialmente verdadeiro nas áreas de ciências exatas, porém raro nas áreas de ciências sociais. Somente as companhias de mídias sociais têm acesso às grandes bases de dados desenvolvidas pelos acumuladores de dados sociais. Um pesquisador pode obter algum desses dados por meio de API's oferecidas pela maioria dos serviços das mídias sociais e dos grandes revendedores de mídias *online*. O acesso limitado a quantidades massivas de dados transacionais sociais que estão sendo coletados pelas grandes empresas "é uma das razões pelas quais ciências sociais e humanidades contemporâneas, orientadas por grandes volumes de dados, não são fáceis de serem realizadas na prática" (MANOVICH, 2011, p.12). Como consequência, o que se pode fazer com as ferramentas disponíveis é muito limitado, posto que essas empresas fazem dinheiro analisando padrões sobre os dados que coletam sobre o nosso comportamento *online* e físico e/ou vendendo os dados para outras companhias.
- **Dados governamentais:** Na contramão das grandes empresas acumuladoras de dados provenientes de mídias sociais, observa-se uma abertura gradual nos dados coletados pelas agências governamentais catalisados pela pressão da opinião pública e pela consequente legislação (por exemplo a LAI, no Brasil). Porém, nem sempre os dados estão prontos e tratados por processos de curadoria digital – por exemplo, adição de metadados – que facilitem a recuperação e as possíveis reanálises. Manovich (2011) assinala que os governos estão se tornando fornecedores de dados, porém esses dados são tipicamente sumários estatísticos muito diferentes dos dados que registram o comportamento *online* das pessoas e das mídias coletadas pelas companhias de mídias sociais.
- **Acessível pode não ser ético**: Qualquer dado que envolva seres humanos inevitavelmente coloca em pauta questões de privacidade e o risco de abuso. Esse fato torna necessário questionar sobre que sistemas éticos, legais e econômicos estão ordenando as práticas de dados e quais códigos as estão regulando. Há pouca compreensão sobre as implicações

éticas do fenômeno do *big data*, especialmente nas ciências sociais. Os pesquisadores que trabalham com o *big data* raramente estão alertas de que existe uma considerável diferença entre o dados estarem disponíveis e poderem ser reusados sem limites.

À GUISA DE CONCLUSÃO

O enfoque orientado por dados se consolida como um novo e perfeitamente válido sistema de ferramentas e técnicas para a descoberta científica. É uma nova forma de inteligência científica que dirige novas questões ao enigmático sorriso de pedra da esfinge da natureza, da sociedade, dos homens e de seus artefatos, e que abre a oportunidade de quantificação das ciências sociais e humanas, e da cultura. Os debates deixam claro que suas estratégias de geração de conhecimento não vão substituir os procedimentos cognitivos e metodológicos tradicionais refinados por muitos séculos, assim como o paradigma experimental não foi substituído pelo paradigma teórico, e ambos não foram substituídos pelo paradigma computacional. Pelo contrário, um ponto de observação mais privilegiado revela que o *big data,* na pesquisa contemporânea, é uma nova forma de interrogação que reforça o teste de hipótese, o desenvolvimento de modelos e a experimentação, redefinindo, talvez, uma nova forma de empirismo.

Entretanto, essas inovações disruptivas no mundo acadêmico, que reconfiguram em muitas instâncias como as pesquisas são conduzidas e o conhecimento é gerado, que recriam disciplinas e criam outras inteiramente baseadas em dados, clamam urgentemente por uma ampla reflexão crítica sobre as implicações epistemológicas da pervasiva revolução dos dados. A partir desse ponto é que se avalia o texto de Anderson.

Encarada como uma provocação apocalíptica, a tese do fim da teoria, proclamada por Anderson, criou um pretexto oportuno para estudos, debates e respostas – às vezes contundentes, porém mais precisas e aprofundadas – sobre a complexidade conceitual e epistemológica da busca pelos conhecimentos científicos em cenários povoados por dados e permeados por tecnologias e algoritmos sofisticados. Sem tomar fôlego, essas novas metodologias avançam na direção da datificação de sentimentos e emoções, trazendo mais elementos para essa discussão otimistamente infindável.

Artigo recebido em 31/01/2019 e aprovado em 07/05/2019.

REFERÊNCIAS

ANDERSON, Chris. The end of theory: the data deluge makes the scientific method obsolete. Science. *Wired*, 2008. Disponível em: <https://www.wired.com/2008/06/pb-theory/>. Acesso em: 25 março 2019.

BOYD, Danah; CRAWFORD, Kate. Critical questions for big data**.** *Information, Communication & Society*, v. 15, n. 5, p. 662-679, 2012. Disponível em: <http://dx.doi.org/10.1080/1369118X.2012.678878>. Acesso em: 20 março 2019.

GUO, Philip. *Data science workflow*: overview and challenges. Disponível em: <http://pgbovine.net/CACM-data-science-workflow.htm>. Acesso em: 20 mar. 2019.

KITCHIN, Rob. Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, v. 1, n. 12, 2014. Disponível em: <https://journals.sagepub.com/doi/full/10.1177/2053951714528481>. Acesso em: 20 mar. 2019.

LATOUR, Bruno. *Tarde's ideia of quantification*. In: CANDEA, Matei. *The social after Gabriel Tarde*: debates and assessments. London: Routledge, 2009. Disponível em <https://hal-sciencespo.archives-ouvertes.fr/hal-00973004/document>. Acesso em: 20 mar. 2019.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. *Big data*: a revolution that will transform how we live, work, and think. Boston: Eamon Dolan: Houghton Mifflin Harcourt, 2013.=

MANOVICH, Lev. *The promises and challenges of big social data*. 2011. Disponível em <http://manovich.net/content/old/03-articles/64-article-2011/64-article-2011.pdf>. Acesso em: 30 mar. 2018.

MAZZOCCHI, Fulvio. Could big data be the end of theory in science? *EMBO Reports*, v. 16, n. 10, 2015. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.15252/embr.201541001>. Acesso em: 30 mar. 2018.

NAIMI, Ashley; WESTREICH, Daniel. Big data: a revolution that will transform how we Live, work, and think. *American Journal of Epidemiology*, v.179, n. 9, p. 1.143-1.144, abr. 2014. Disponível em: <https://academic.oup.com/aje/article/179/9/1143/2739247>. Acesso em: 25 mar. 2019.

PIGLIUCCI, Massimo. The end of theory in science? ***EMBO Reports***, v. 10, n. 6, 2009. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1038/embor.2009.111>. Acesso em: 25 mar. 2019.

RODRIGUES, Eloy; SARAIVA, Ricardo. *Os repositórios de dados científicos*: estado da arte. Porto: RCAAP, 2010. Disponível em: <https://repositorio-aberto.up.pt/handle/10216/23806>. Acesso em: 7 dez. 2018.

SAYÃO, Luis Fernando. Modelos teóricos em ciência da informação: abstração e método científico. *Ciência da Informação*, v. 30, n. 1, p. 82-91, jan./abr. 2001 Disponível em: <http://www.scielo.br/pdf/ci/v30n1/a10v30n1>. Acesso em: 25 mar. 2019.

SAYÃO, Luis Fernando; SALES, Luana Farias. A ciência invisível: os dados da cauda longa da pesquisa científica. In: DIAS, G. A; OLIVEIRA, B. M. J. F (Org.). *Dados científicos*: perspectivas e desafio. João Pessoa: Ed. UFPB, 2019. No prelo.

SCHMITT, Charles et al. Scientificdiscovery in the era of big data: more than the scientific method. *RENCI White Paper Series*, v. 3,n. 6, p. 1-22, 2015. Disponível em: <https://renci.org/wp-content/uploads/2015/11/SCi-Discovery-BigData-FINAL-11.23.15.pdf>. Acesso em: 12 dez. 2018.

THE ROYAL SOCIETY. *Science as an open enterprise*. London: The Royal Society Science Policy Centre, 2012. Disponível em: <https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf>. Acesso em: 23 mar. 2019.