A epistemologia interdisciplinar do Big Data: principais discussões e dilemas[1]

Allan CANCIAN Marquez[2]
Fábio Luiz MALINI de Lima[3]
Universidade Federal do Espírito Santo, Vitória ES

Resumo

Nossa sociedade recebe a cada minuto uma enxurrada de informações e dados diversos. Essa evolução natural da comunicação entre pessoas legitimou novas formas de pesquisa que tentam descobrir as peculiaridades da vida em sociedade, seja para a obtenção de dinheiro ou para o conhecimento acadêmico. A esses grandes dados, conhecidos simplesmente como "Big Data", estão atreladas todas as ligações digitais dos mais diversos tipos, o que torna essencialmente importante a discussão teórica sobre o termo. O objetivo desta revisão é apresentar, conceituar e analisar este fenômeno, sinalizando os principais recursos tecnológicos envolvidos, as principais discussões e enquadramentos epistemológicos sobre o tema, bem como o foco em discutir os conceitos de Ciência de Dados, seus dilemas e elementos constitutivos.

Palavras-chave: big data; internet; ciência de dados; epistemologia; métodos digitais.

Introdução

Vivemos em um mundo conectado, impulsionado pela expansão da internet e permeado com as inovações sociais trazidas por ela, o que permitiu ao mundo vivenciar uma globalização não apenas mercadológica, como também intercambiada de cultura, experiência e sociabilidade, possibilitando assim a troca de qualquer tipo de informação. Na passagem do modelo dos portais para a atual sociedade dos perfis de redes, o povoamento da internet participativa é menos estabelecido pela intensa criação de sites, fóruns e chats do que da aluvião de aplicações baseadas no *login* de um avatar cuja principal característica é de construção de uma timeline com gostos, atos, *checkins* e opiniões, estabelecendo uma narrativa contínua do cotidiano a partir dessa multiplicidade de rastros pessoais. Nessa sociedade de perfis, um perfil existe porque

[1] Trabalho apresentado no GP Conteúdos digitais e convergências tecnológicas do XVII Encontro dos Grupos de Pesquisa em Comunicação, evento componente do 40º Congresso Brasileiro de Ciências da Comunicação.

[2] Mestrando no programa de Comunicação e Territorialidades na Universidade Federal do Espírito Santo (Ufes) e Pesquisador do Labic (Laboratório de Estudos Sobre Imagem e Cibercultura), E-mail: allancancian@gmail.com.

[3] Prof. Dr. Fábio Malini, Professor Adjunto na Universidade Federal do Espírito Santo (Ufes), onde também coordena o Labic, E-mail: fabiomalini@gmail.com.

está em relação com um outro – seja ele seguidor, amigo, inscrito, etc – e graças a esse entrelaçamento de redes podemos interpretá-los como perfis associados que afirmam conceitos e estão em constante atração e repulsão com outros atores e pontos de vista (MALINI, 2016).

O acúmulo de informações fabricadas pelos perfis de redes sociais fez emergir um quadro desafio para a pesquisa social, uma vez que os padrões de comportamentos e relações digitais passavam ser um problema de "Big Data" (megadados). A partir de agora torna-se imprescindível a presença de laboratórios de pesquisa para extrair e entender esses megadados, permitindo assim estudar diferentes fenômenos de qualquer disciplina, incluindo as do campo da Comunicação.

Hoje esse termo é usado frequentemente em meios de comunicação, de jornais populares a revistas científicas, que usam da disponibilidade desses dados para construir infográficos, matérias densamente analíticas ou até mesmo para legitimar a produção de algumas pautas, o que fez com que o termo adquirisse durante os últimos 50 anos o status de "grande ciência". Borgman (2015) vê isso como algo a ser superado, pois em seu entendimento o que o Big Data quer não é decifrar todos os mistérios do universo, mas descobrir as narrativas de um pedaço da vida humana e a partir delas procurar dar sentido, levar a ações estratégicas e até mesmo preditivas.

O autor, procurando uma definição completa sobre a ciência de dados, afirma que a definição de um dado depende bastante do olhar do pesquisador e qual viés ele quer dar para sua pesquisa. Segundo o escritor, os dados não são objetos puros ou naturais com uma essência própria, pois existem em um contexto e são dotados de significado. Seguindo essa lógica, a interpretação dos dados depende muito das peculiaridades na qual eles estão imersos e do viés em que serão analisados.

Sobre o uso desses dados digitais em pesquisas mercadológicas e acadêmicas, Rogers (2016) afirma que a internet deve ser usada para buscar uma realidade que vai além da cultura on-line, como diagnóstico de mudanças culturais e condições sociais existentes. O autor acredita que a internet é um espaço de pesquisa e que deve ser repensada como uma fonte de dados sobre a sociedade e seu comportamento. Sendo assim, é a partir dos dados nela inseridos que se tornará possível entender e extrair significado sobre nossos passos, ou seja, pensar uma nova relação entre a web e o que a fundamenta é o maior objetivo teórico desses métodos digitais. "Coletar e analisar esses dados para a pesquisa social e cultural requer não apenas uma nova perspectiva sobre a

internet, mas também novos métodos para fundamentar as descobertas" (ROGERS, 2016, p. 30). Na verdade, Rogers (2016) defende a tese de que o paradigma do virtual está ultrapassado, isto é, os estudos ciberculturais definidos pela transposição de fatos offline para o online são datados até a emergência dos métodos digitais, marcados pela extração contínua da ação dos perfis em diferentes plataformas digitais e da análise posterior das ligações entre esses perfis.

Ao pensarmos sobre a utilização do Big Data em pesquisas científicas, vemos o quão pouco foi produzido até então. Para se ter uma ideia, são escassos os trabalhos em português que abordam o Big Data em diferentes vieses, sejam eles acadêmicos ou voltados para estudos mercadológicos. Durante os últimos 5 anos, apenas 28 trabalhos disponibilizados no Periódico Capes abordaram o tema, sendo que 20 desses estudos foram publicados como artigos e os outros 8 em artigos de jornal. Dessa forma, a carência de produções na área foi um forte impulsionador para a realização desse trabalho.

É com base nessa forma de extrair as mais variadas informações e produzir análises de alto valor informacional que este artigo se debruça. Nossa sociedade recebe a cada minuto uma enxurrada de informações e dados diversos, o que torna essencialmente importante a discussão teórica sobre Big Data e as Ciências Sociais Aplicadas, envolvendo pesquisadores de várias áreas graças a característica interdisciplinar desse objeto. Sendo assim, o objetivo desta revisão é apresentar, conceituar e analisar este fenômeno, sinalizando os principais recursos tecnológicos envolvidos, as principais discussões e enquadramentos epistemológicos sobre o tema, bem como o foco em discutir os conceitos de Ciência de Dados, seus dilemas e elementos constitutivos.

Em busca de uma definição interdisciplinar sobre Big Data

Embora a ciência dos megadados seja relativamente nova, o Big Data tem gerado a curiosidade de alguns teóricos de diferentes áreas acadêmicas. Alguns o veem como um fenômeno sem precedentes, que trabalha com uma imensa quantidade de informações a fim de resolver os mais diversos problemas (FREDERIKSEN, 2005; DI MARTINO, 2010; LAGOZE, 2014; MCAFEE & BRYNJOLFSSON, 2012), já outros o entendem como dados que ultrapassam a capacidade de processamento dos sistemas

convencionais de análise (DUMBILL, 2012; MANOVICH, 2012; MANYKA et al, 2011). Há também os pesquisadores que visualizam o Big Data como trabalhos em grande escala que não devem ser analisados como os de escala menor, produzindo *insights* que ajudam várias áreas da sociedade, como os mercados, a academia, organizações políticas, entre outras (SCHONBERGER-MAYER e CUKIER, 2013; BOYD e CRAWFORD, 2012).

De acordo com Kitchin (2014), "Big Data é caracterizado por ser continuamente gerado, buscando ser exaustivo, refinado no escopo, flexível e escalável em sua produção[4]" (KITCHIN, 2014, p. 2, tradução nossa). Para o autor, embora a produção de conhecimento a partir dos grandes dados já exista há algum tempo nas áreas de sensoriamento remoto, previsão meteorológica ou em mercados financeiros, graças a popularização da internet e sua característica social, outros formatos de extração e trabalho com o Big Data começaram a ganhar espaço, o que o autor vem a definir como "novas formas de análise de dados projetados para lidar com a abundância[5]" (KITCHIN, 2014, p. 2, tradução nossa).

Os estudos mercadológicos e da academia na esfera do Big Data, embora diferentes em alguns aspectos pertinentes a cada área, buscam resultados a partir da análise de dados, como entender as peculiaridades de um público-alvo específico ou a resposta sobre um complexo problema, por exemplo. De acordo com um levantamento epistemológico de artigos publicados em diversas línguas sobre *big data analytcs*, produzido pelos pesquisadores Patricia K. Furlan e Fernando J. B. Laurindo (2016), foram identificados cinco campos de estudo correlatos sobre o tema, sendo eles: a mineração de dados, a análise de dados e geração de conhecimento; estratégia e gestão de negócios; influência da tecnologia no comportamento humano e nas alterações sociais; e discussões a respeito da Internet das Coisas (FURLAN & LAURINDO, 2016).

Percebendo essa interdisciplinaridade do objeto, Vis (2013) apresenta duas concepções para Big Data: uma direcionada ao campo empresarial e outra responsável por desenvolver projetos acadêmicos. De acordo com o autor, enquanto a perspectiva empresarial produz sua pesquisa utilizando o Big Data como tomada de decisão para

[4] No original: "Big Data is characterized by being generated continuously, seeking to be exhaustive and fine-grained in scope, and flexible and scalable in its production".
[5] No original: "new forms of data analytics designed to cope with data abundance".

aumentar seus lucros, a perspectiva da academia está preocupada em discutir a criação de conhecimento para o mundo científico. Ao focar sua discussão para as pesquisas científicas que usam dos grandes dados a fim de construir seus estudos, Vis apresenta três questões sensíveis, sendo elas: o fortalecimento da precisão e força dos cálculos de coleta, análises e comparações dos dados (questão tecnológica); a descoberta e definição de padrões sociais, econômicos, legais ou técnicos no Big Data (questão da análise); e a crença de que grandes conjuntos de dados possibilitam uma inteligência e conhecimento nunca antes vistos nas ciências, com a aura da verdade, objetividade e precisão (questão mitológica) (VIS, 2013).

Também debatendo sobre os diferentes aspectos desse fenômeno, Gray (2009) irá propor quatro paradigmas na ciência para compreender seu percurso para com os dados provenientes do Big Data. O primeiro paradigma seria o empírico, em que as práticas científicas buscam sempre relatar os fenômenos naturais; o segundo seria o teórico, em que há esforços para determinar os modelos e considerações generalistas sobre os fenômenos estudados; o terceiro elemento seria o computacional, paradigma essencial para o tratamento de fenômenos mais complexos, além de suas visualizações; por fim, o quarto paradigma seria o da exploração de grandes volumes de dados (*data exploration* ou *e-Science*), por criar um ambiente repleto de informações que podem ser gerenciadas e categorizadas de diferentes tipos (GRAY, 2009).

A Ciência de Dados para pesquisas acadêmicas

A Ciência de Dados (*Data Science*) é o ramo científico que nasceu como forma de estudar os fenômenos provenientes dos megadados. O termo apareceu na literatura da ciência da computação ao longo das últimas décadas do século passado, porém, foi no final dos anos 1990 que o campo começou a ganhar suas primeiras comunidades de adeptos à extração e análise dos dados, alcançando pela primeira vez em 2001 o status de disciplina independente (HERMAN et al, 2013). De acordo com Manovich (2012), foi o avanço das ferramentas da internet quem contribuiu para instituir o *data science* como uma ciência dos dias atuais.

> "O surgimento de mídias sociais no meio da década de 2000 criou oportunidades para estudar processos sociais e culturais e dinâmicas de novas formas. Pela primeira vez, podemos seguir a imaginação, opiniões, ideias e

> sentimentos de centenas de milhões de pessoas. Podemos ver as imagens e os vídeos que eles criam e comentar, monitorar as conversas que estão envolvidos, ler seus posts e tweets, navegar em seus mapas, ouvir suas listas de faixas, e seguir suas trajetórias no espaço físico[6]". (MANOVICH, 2012, p. 2, tradução nossa)

Para autores como Rodrigues (2017), é essa quantidade de possibilidades e laços em rede que contribui para a análise dos dados. A pesquisadora observa que "a imensa quantidade de perfis e informações oriundas desses dados criam perspectivas diversas sobre determinados assuntos e áreas de interesse, permitindo a visualização de novos padrões e de pesquisa interdisciplinar" (RODRIGUES, 2017, p. 37). Também de acordo com a autora, é com fundamentação em teorias de disciplinas tradicionais e já consolidadas, bem como na utilização em perspectivas interdisciplinares e que não precisam estar necessariamente ligadas a um território fixo, que o Big Data – assim como a Ciência de Dados – acaba ganhando novos valores e angariando outras possibilidades. "Dessa maneira, as metodologias em torno de Ciência de Dados ainda estão em fase de consolidação para acercar os fenômenos relacionados a megadados" (RODRIGUES, 2017, p. 36).

Por ser uma ciência nova e carregada de conteúdo, muitos entusiastas têm ajudado a enumerar as melhores formas de se trabalhar com os dados. Rodrigues ainda cita os autores Porto e Ziviani (2014) para discutir os aspectos centrais da Ciência de Dados com base em sua expansão teórica e conceitual com relação a outras disciplinas. Para eles, tais aspectos seriam a gerência de dados, a análise desse conteúdo e a análise de redes complexas. "A ciência de dados emerge como componente cada vez mais importante nas mais diversas áreas" (PORTO E ZIVIANI, 2014, p. 2 apud RODRIGUES, 2017, p. 37).

Assim como Rodrigues e os demais autores que trabalham a inserção do conhecimento proveniente do Big Data nas pesquisas acadêmicas, Margetts, John, Hale e Yasseri (2016) defendem que os cientistas sociais devam trabalhar conjuntamente com os pesquisadores e profissionais de outras áreas, a fim de realizar totalmente a extração e análise de Big Data, algo que do ponto de vista dos autores não é aplicado nas escolas de ciências sociais tradicionais (MARGETTS, JOHN, HALE, REISSFELDER, 2016).

[6] No original: "The emergence of social media in the middle of 2000s created opportunities to study social and cultural processes and dynamics in new ways. For the first time, we can follow imaginations, opinions, ideas, and feelings of hundreds of millions of people. We can see the images and the videos they create and comment on, monitor the conversations they are engaged in, read their blog posts and tweets, navigate their maps, listen to their track lists, and follow their trajectories in physical space.

Também na mesma linha de raciocínio interdisciplinar, Malini (2016) apresenta a Ciência de Dados como "um campo em formação - derivada da mistura de Ciências Humanas, Estatísticas, Física e Ciências da Computação, predominantemente - que nos permite testar novas possibilidades" (MALINI, *online*).

É essa capacidade de testar em variados formatos que fez com que pesquisadores como Caldas e Silva (2016) buscassem uma interpretação mais próxima da realidade do Big Data. Eles afirmam que, apesar de seu conceito não ser tão claro e gerar controvérsias em alguns momentos, "seu crescimento é volumoso, decorrente do enorme número de informações não estruturadas, além de ser necessário processamento em grande escala para que seja possível a extração de informações" (CALDAS & SILVA, 2016, p. 79). Essa extração, de acordo com os autores, busca estabelecer valores que impactem a economia, o governo, organizações e relações interpessoais.

Sobre de onde prover esses dados, seguindo a lógica conceitual de Manovich (2012), os pesquisadores afirmam que os grandes dados possibilitaram dar uma atenção a informações antes consideradas sem valor, como um comentário em uma rede social, por exemplo. Para eles, "em vez de analisar apenas um percentual de dados, como uma amostragem, por exemplo, seriam analisados além da amostragem, dados de diversas fontes nunca utilizadas antes" (CALDAS & SILVA, 2016, p 74).

Outros autores mais preocupados com a utilização e o estado dos dados adotaram pesquisas direcionadas a estudar as características multidimensionais do Big Data. Um desses pesquisadores foi Laney (2001 apud LAGOZE, 2014) que, tentando compreender a complexidade dos dados, defini-os como sendo volume, velocidade e variedade (os chamados 3V's do Big Data). O *volume* refere-se a seu tamanho. A *velocidade* está ligada a seu caráter dinâmico e sua capacidade de processamento em uma escala necessária para torná-lo útil e mantê-lo em funcionamento. Já a *variedade* significa a mistura de tipos de dados heterogêneos em um mesmo *dataset*[7], criando uma necessidade em resolver essas diferenças para tornar os dados úteis.

Autores como Marr (2015, *online*) acrescentam ainda a *veracidade* e o *valor* como características inerentes aos grandes dados. O primeiro refere-se a sua desordem ou confiabilidade, ou seja, devem ser de qualidade e com informações precisas sobre o

[7] *DataSet* é um conjunto de dados que consiste em uma série de registros tabulados (em formato de tabelas). Cada coluna representa uma variável particular e cada linha corresponde a um determinado elemento do conjunto de dados em questão.

tema tratado. Já o segundo significa a capacidade de transformar os dados em valor, isto é, a habilidade de quem está pesquisando os dados em compreender a real importância do mesmo, seu recorte e elementos essenciais para sua compreensão. Dessa forma os autores que buscam uma definição mais global do Big Data se utilizam desses 5V's como conceitos iniciais de suas pesquisas e análises.

Como vimos até então, a Ciência de Dados está focada na obtenção de informações a partir dos dados, ao contrário de outras ciências que buscam os relatos em fatos históricos. Entretanto, algo deve ser entendido quando falamos de Big Data: a quantidade de dados não significa dizer que necessariamente encontraremos melhores pesquisas ou informações completamente destrinchadas em seu entendimento, já que quantidade não significa qualidade. É muito importante também estabelecer estratégias de ir do *big* ao *small* data, a fim de tornar mais qualitativa qualquer tipo de análise social.

Antes de confiar totalmente em um *dataset* é necessário verificar sua genuinidade, o que, de acordo com a empresa de consultoria e marketing Booz Allen Hamilton (HERMAN et al, 2013), isso não é compreendido pela maioria das pessoas. Segundo a empresa, enquanto a maior parte das pessoas associa volume, velocidade e variedade com a veracidade dos dados, pesquisas incorretas e de metodologia incerta são produzidas mundo afora. "Você deve avaliar a veracidade e a exatidão dos dados, bem como identificar informações ausentes ou incompletas[8]" (HERMAN et al, 2013, p. 82, tradução nossa).

O caráter interdisciplinar do Big Data faz com que ele esteja envolvido em questões de áreas completamente distintas entre si. Tufekci (2014) considera que os grandes dados devem ser compreendidos como um processo político que está inserido em discussões sobre transparência, poder e vigilância. A autora cita seis motivações que fizeram a quantidade de estudos a partir do Big Data aumentar exponencialmente durante os últimos anos: a primeira seria o crescimento digital das comunicações sociais, políticas e econômicas; a segunda, o refinamento do conceito de "público-alvo" para um alvo muito mais individual e assertivo a partir das redes sociais; já a terceira seria a obtenção de informações e repostas a perguntas que nem sequer foram feitas pelos e aos usuários a partir das técnicas de coleta atuais (cruzamento de dados, inserção

[8] No original: "You must assess the truthfulness and accuracy of the data as well as identify missing or incomplete information".

de técnicas interdisciplinares); a quarta motivação é o avanço dos estudos sobre o comportamento humano, cada vez mais precisos e inteligentes; a quinta motivação é referente a resposta em tempo real nas pesquisas de opinião graças às conexões das redes digitais; e a sexta são as ferramentas e técnicas disponíveis que permitem a coleta de dados, boa parte delas disponibilizadas publicamente e com acesso fácil (TUFEKCI, 2014).

Extração, mineração e visualização do Big Data como metodologia digital

É essa sexta grande questão, a extração de dados, que permite transformar inúmeras informações complexas em megadados "tangíveis" e analisáveis. As redes sociais, por exemplo, são uma das ferramentas mais dinâmicas da internet e novas ligações entre seus atores podem surgir a todo instante, daí a relevância em se utilizar das técnicas de mineração de dados para descobrir mais sobre a vida desses indivíduos, suas preferências, interesses políticos, etc. Alves (2016), ao dar o exemplo do Facebook, explica que "sempre que curtimos uma página, criamos um laço na rede social, um canal para receber informações" (ALVES, 2016, p. 77). O pesquisador também joga luz às características de um banco de dados dessa rede social, onde cada nó representa um ator (página, grupo ou usuário, por exemplo) e as arestas significam as ações desses atores (curtidas, comentários e compartilhamentos), um conhecimento importante na hora de visualizar e trabalhar com um *dataset*.

Dias (2002), ao refletir sobre os motivos de se utilizar da coleta de dados para a realização de pesquisas, sustenta os elementos característicos do método digital. Para ela, "os principais objetivos da mineração de dados são descobrir relacionamentos entre dados e fornecer subsídios para que possa ser feita uma previsão de tendências futuras baseadas no passado" (DIAS, 2002, p. 1716). Já segundo Caldas e Silva (2016), muito mais que apenas garimpar em busca de informações quantitativas, o *data mining* permite descobrir padrões e avaliar pressupostos preditivos, controle de processos, tomada de decisão entre outras aplicações.

> "O termo DN, utilizado nessa pesquisa para designar Data Mining, é conhecido também como mineração de dados, consiste em um processo analítico projetado para explorar grandes quantidades de dados (tipicamente relacionados a negócios, mercado ou pesquisas científicas), na busca de padrões consistentes e/ou relacionamentos sistemático entre variáveis para, então, validá-los

> aplicando os padrões detectados a novos subconjuntos de dados" (CALDAS E SILVA, 2016, p. 69).

Todavia, para a mineração dos dados conseguir coletar todas as informações que uma pesquisa busca encontrar, deve existir uma conexão entre o banco de dados do site e o computador do pesquisador na qual a mineração será realizada. Essa conexão é possibilitada pela API de cada site (sigla para Interface de Programação de Aplicativos, "*Application Programming Interface*" em inglês), um conjunto de comandos que podem ser usados por um programa de usuário para recuperar os conteúdos armazenados em bancos de dados de sites como Facebook, Twitter, Flickr, YouTube entre outros.

É graças a extração possibilitada pela API e pelo programa coletor que conseguimos descobrir quantos usuários deram *like* em determinados comentários do Facebook ou quantas pessoas utilizaram uma *hashtag* no Twitter e quais as postagens mais compartilhadas, por exemplo. Manovich (2012), todavia, alerta para um ponto sensível da mineração de conteúdo desses sites: algumas informações específicas ficam bloqueadas pelas empresas, o que impossibilitaria alguns avanços em pesquisas. Na API do Twitter, por exemplo, não há como extrair todos os *replies* de um post, mas, apenas aqueles comentários com o termo de busca requerido[9]. Segundo o autor, "as API's públicas fornecidas pela mídia social e pelas empresas de redes sociais não dão todos os dados que essas próprias empresas estão capturando sobre os usuários[10]" (MANOVICH, 2012, p. 5, tradução nossa).

Voltando nossos olhares teóricos para o *data mining* e a fim de deixar mais claro o processo de obtenção e análise dos megadados, a Booz Allen Hamilton, junto com seus especialistas, aprofundou-se em definir quatro procedimentos que esses dados precisam passar para serem legíveis e modeláveis. De acordo com a empresa, o primeiro desses procedimentos é a *coleta*, uma atividade que se concentra na obtenção dos dados necessários. Os detalhes dessa atividade dependem de qual será a pesquisa e quais informações o pesquisador irá coletar; O segundo procedimento é o *preparo*, responsável por manipular os dados para encontrar suas necessidades analíticas, seja visualizando previamente alguns dados para encontrar um objeto de pesquisa ou

[9] É ao partir dessa realidade que muitos pesquisadores têm apoiado a ideia do *Open Data*, uma linha de pensamento sobre a transparência do Big Data que julga imprescindível a abertura das informações dos sites para todos na internet, sem restrições de direitos autorais e patentes ou outros mecanismos de controle (RODRIGUES, 2017; BORGMAN, 2015; MANOVICH, 2012).

[10] No original: "The public APIs provided by social media and social network companies do not give all data that these companies themselves are capturing about the users".

transformando esses dados em informações legíveis para o pesquisador; A *análise* é o terceiro procedimento e refere-se aos programas utilizados para lapidar os dados, bem como os pesquisadores que os utilizam para dar sentido e forma a tais informações. O tipo de análise permite a compreensão em tempo real dos riscos e oportunidades, avaliando os dados situacionais, operacionais e comportamentais. Finalmente, o quarto procedimento definido pela empresa é o *ato*, conceituado assim como a habilidade de fazer uso da análise crítica, trabalhando com resultados que tenham sentido e que mostrem a veracidade dos dados. (HERMAN et al, 2013, p. 25). Talvez falte a essa definição uma quinta etapa: a visualização de dados, imprescindível para fornecer novos modos de visibilidade de temas complexos inscritos nos megadados.

Como o Big Data é por sua natureza repleto de dados precisos e densamente conectados, é necessário que o pesquisador recorra a técnicas de modelagem para facilitar sua análise. Além disso, é preciso ter programas que facilitem a visualização dos dados e a análise das relações entre os atores e seus rastros nas redes sociais, com o intuito de correlacionar os dados e levar a respostas (TUFEKCI, 2014). Visualizar o Big Data é observar o social se desenvolver, conseguindo dessa forma perceber o caos antes da estabilização. É necessário entrar e estudar os laços sociais que formam a rede, entendendo as reviravoltas de uma narrativa e como certos atores conseguiram moldar tal emaranhado de conexões (LEMOS, 2013).

Os principais dilemas do Big Data

Já descrevemos aqui que a ciência por trás do Big Data é responsável por evoluir consideravelmente o nosso entendimento sobre inúmeras indagações. De simples padrões em redes de conversação no Twitter até imensos bancos de dados sobre o comportamento de um consumidor em diferentes mercados, a era dos grandes dados veio para ficar. Por ser uma ciência recém estruturada e sempre em construção, alguns dilemas e incertezas acabam por aparecer em textos científicos e discussões entre autores da área, o que fez com que enumerássemos as duas principais questões sensíveis a respeito do Big Data.

O primeiro grande dilema sobre os megadados diz respeito à privacidade dos usuários. Autores como Rodrigues (2017) e Breternitz & Silva (2013) reportam o propósito do Big Data como método de vigilância e dizem que isso precisa ser evitado.

De acordo com Rodrigues, esse propósito "confere novas potencialidades e implicações econômicas, sociais, culturais e tecnológicas, inclusive de vigilância com a mineração dos dados e o cruzamento com bases diversas" (RODRIGUES, 2017, p; 70). Como forma de otimizar os resultados sem revelar a identidade dos atores, uma possível solução seria tornar anônimo os dados sensíveis de pessoas e tratá-las como um grupo de usuários, evitando a completa identificação dos mesmos.

Já a segunda fonte de questões a respeito do Big Data refere-se à forma como os pesquisadores e analistas utilizam e interpretam os dados. Kitchin (2014) é um dos autores que alegam esse dilema, já que critica a forma como algumas pesquisas que utilizam o Big Data como metodologia são criadas. Para ele, essas pesquisas parecem ser geradas sem problemas (objetos), sendo o foco impulsionado pela aplicação de um método ou o conteúdo do conjunto de dados em vez de uma questão específica, o que para o autor acaba por produzir pesquisas fracas e que não geram grandes resultados.

Todavia, o dilema da interpretação dos dados para uso em pesquisas científicas também rende discussões a respeito da incorporação da metodologia da Ciência de Dados em teorias clássicas. Se para Frické (2015) os dados não teriam significado sem a teoria, para Kitchin (2014) os grandes dados não necessitam de um enquadramento teórico, modelos ou hipótese a priori, já que através da aplicação da análise os dados mostrarão informações íntegras e livres de viés humano ou enquadramento, possuindo significado verdadeiro. Para o pesquisador, o significado transcende um contexto ou domínio de conhecimento específico, o que possibilitaria os dados serem interpretados por qualquer pessoa.

Bruns (2013), todavia, acrescenta nessa discussão o desafio em documentar e publicar os métodos de extração e análise do Big Data. Por ser uma nova área, muitos pesquisadores têm trabalhado em seus próprios métodos e precisam explica-los com todos os detalhes, o que acaba deixando as pesquisas com pouco espaço para a apresentação dos resultados e suas especificidades. Por perceber o problema e buscar uma solução, o autor defende que os métodos sejam detalhados em publicações separadas dos objetos, para assim o pesquisador ter espaço para escrever seu *corpus*.

Considerações finais

O Big Data exige uma perfeita realização dos processos de mineração, preparo, visualização e análise dos megadados, para que os resultados possam ser altamente

assertivos e confiáveis. Apresentamos neste trabalho considerações preliminares de autores das mais diversas áreas do conhecimento, a fim de construir uma pesquisa que possa futuramente ajudar outros pesquisadores a entenderem um pouco mais sobre as discussões da Ciência de Dados nas ciências sociais. O simples fato de existir um trabalho que aborde e explique o tema foi uma das nossas metas, pois ainda são poucos os artigos em português que trazem uma visão mais acadêmica do Big Data. Esse trabalho ainda está a ser complementado a partir do levantamento de artigos científicos sobre o tema na base do Periódico Capes.

Sobre suas características, vimos que o Big Data é menos sobre volume de dados e mais sobre a capacidade de interpretação, pesquisa e uso crítico desses dados (Boyd e Crowford, 2012). Essa análise dos megadados permite novas abordagens metodológicas que possibilitam fazer e responder as perguntas de novas maneiras, uma afirmação compartilhada pelos autores que estudam o tema.

Encontramos escritores que apostam em um futuro em que esse fenômeno se tornará ainda mais presente em nossa sociedade, seja criando um ambiente de inovações diversificadas em relação às pesquisas tradicionais, seja desafiando a atual logística da gestão de dados não digitais (Rodrigues, 2017). Essa visão sobre o futuro da sociedade com os megadados é ampliada por Manovich (2012), que imagina uma população dividida em três diferentes categorias, o que o autor vem chamar de "classes de dados": (1) aqueles que criam dados (todos inseridos em redes sociais ou sites que permitam a inserção de informações), (2) aqueles que têm os meios para coletá-lo e (3) aqueles que têm experiência para analisá-lo.

Vimos, todavia, que a questão do Big Data envolve dilemas importantes, principalmente no que diz respeito à privacidade dos dados e ao uso metodológico que alguns pesquisadores utilizam para expor seus objetos à luz dos grandes dados. Trabalhar com Big Data não é apenas quantificar dados e apresentar respostas superficiais, mas principalmente encontrar informações que antes não se tinha acesso e criar *insights* a partir dos resultados identificados.

Por fim, os pesquisadores dessa nova área não descartam os resultados e métricas criados por profissionais de outras áreas. Ao pensar nesse caráter multidisciplinar do Big Data, autores como Borgman (2015) afirmam que os “dados de pesquisas acadêmicas podem ter valor comercial e dados comerciais podem servir aos interesses

acadêmicos, levando a novas parcerias e novas tensões[11]" (BORGMAN, 2015, p. 8, tradução nossa). Em resumo, é essa troca de informações que permite ao Big Data continuar crescendo em relevância e poder de investigar as mais variadas nuances da vida em sociedade.

REFERÊNCIAS

ALVES, M. **Abordagens da coleta de dados nas mídias sociais**. In Monitoramento e pesquisa em mídias sociais: metodologias, aplicações e inovações. São Paulo: Uva Limão, 2016.

BORGMAN, C. **Big data, little data, no data**. Scholarship in the networked world. Cambridge-London: The MIT Press, 2015.

BOYD, D.; CRAWFORD, K. **Critical questions for big data:** Provocations for a cultural, technological, and scholarly phenomenon. In Information Communication & Society, v. 15, n. 5, p. 662-679, 2012.

BRETERNITZ, V. J.; SILVA, L. A. **Big data: Um novo conceito gerando oportunidades e desafios.** Revista Eletrônica de Tecnologia e Cultura, vol. 2, no. 2, pp. 1–8, 2013.

BRUNS, A. **Faster than the speed of print:** Reconciling 'big data' social media analysis and academic scholarship. First Monday, oct. 2013. Disponível em <http://firstmonday.org/ojs/index.php/fm/article/view/4879>. Acesso em 29 de junho de 2017.

CALDAS, M. S.; SILVA, E. C. C. **Fundamentos e aplicação do Big Data: como tratar informações em uma sociedade de yottabytes**. Bibliotecas Universitárias: pesquisas, experiências e perspectivas, Belo Horizonte, v. 3, n.1, p. 65-83, jan. /jun. 2016.

DI MARTINO, B. et al. **Big data (lost) in the cloud.** In International Journal of Big Data Intelligence, Vol.1, No.1/2, pp.3 – 17, 2014.

DIAS, M. **Parâmetros na escolha de técnicas e ferramentas de mineração de dados**. Acta Scientiarum, Maringá, Vol. 24, Nº 6, p. 1715-1725

DUMBILL, E. **What is big data? An introduction to the big data landscape.** O'Reilly Media, Inc., 2012. Disponível em: < https://goo.gl/W2b7F >. Acesso em 24 de junho de 2017.

FRICKÉ, M. **Big Data and its Epistemology**. Journal of the American Society for Information Science and Technology, 66 (4): 651-661, 2015.

FURLAN, P.K.; LAURINDO, F.B. **Agrupamentos epistemológicos de artigos publicados sobre big data analytics**. Transinformação, Campinas , v. 29, n. 1, p. 91-100, Jan/Abr, 2017. Disponível em < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862017000100091&lng=en&nrm=iso >. Acesso em 20 de junho de 2017.

[11] No original: "Data form academic research can have commercial value and commercial data can serve academic inquiry, leading to new partnerships and new tensions".

GRAY, J. **Jim Gray on science: a transformed scientific method**. In HEY, T.; TANSLEY, S.; TOLLE, K. (Ed.). The fourth paradigm: data-intensive scientific discovery. Washington: Microsoft Research, 2009.

HERMAN, M. et al. **The Field Guide to Data Science**. Booz Allen Hamilton Inc, 2013.

KITCHIN, R**. Big Data, new epistemologies and paradigm shifts**. In Big Data & Society. April–June 2014.

LAGOZE, C. **Big Data, data integrity, and the fracturing of the zone control**. In Original Research Article. Big Data & Society. July-December, 2014.

MALINI, F. **Um método perspectivista de análise de redes sociais**: cartografando topologias e temporalidade em rede. XXV Encontro Anual da Compós, Universidade Federal de Goiás, Goiânia, 2016.

_____. **A ciência de dados e o marketing político:** inclusão experimental nas páginas de Casagrande e Hartung. Disponível em <https://goo.gl/n62uRA> Acesso em 28 de junho de 2017.

MANIKA, J. et al. **Big data: The next frontier for innovation, competition, and productivity**. Disponível em: < https://goo.gl/EHAhVV >. Acesso em 25 de junho de 2017.

MANOVICH, L. **Trending: The Promises and the Challenges of Big Social Data**. Disponível em: < http://goo.gl/IqlgGF >. Acesso em 24 de junho de 2017.

MARGETTS, H., JOHN, P., HALE, S., & YASSERI, T. **Political turbulence:** How social media shape collective action. Princeton University Press, 2015.

MARR, B. **Big Data: The 5 Vs everyone must to know**. Disponível em: <https://goo.gl/i9TLRh>. Acesso em 16 de junho de 2017.

McAFEE, A.; BRYNJOLFSSON, E. **Big data: The management revolution**. Harvard Business Review, v. 90, n. 10, p. 60, 2012.

RODRIGUES, A. Proposta de Qualificação de Doutorado apresentada ao Programa de Pós-graduação em Ciência da Informação do Centro de Ciências Sociais Aplicadas da Universidade Federal da Paraíba. **Visualização de dados científicos no cenário da data science**: Produção de formatos inovadores disruptivos. UFP: Paraíba, 2017.

ROGERS, R. **O fim do virtual: os métodos digitais**. Juiz de Fora: Revista Lumina, 2016.

SCHÖNBERGER-MAYER, V.; CUKIER K. **Big data: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana**. Rio de Janeiro: Elsevier, 2013.

TUFEKCI, Z. **Engineering the public: Big data, surveillance and computational politics**. First Monday. Julho, 2014.

VIS, F. **A critical reflection on Big Data:** Considering APIs, researchers and tools as data makers. First Monday, [S.l.], oct. 2013. Disponível em <http://firstmonday.org/ojs/index.php/fm/article/view/4878>. Acesso em 20 de junho de 2017.