CONHECIMENTO CIENTÍFICO NO CONTEXTO BIG DATA: reflexões a partir da epistemologia de Popper

Scientific Knowledge In The Big Data Context: Reflections from Popper's epistemology

Eugênio Monteiro da Silva Júnior (1), Cezar Karpinski (2), Moisés Lima Dutra (3)

(1) Universidade Federal de Santa Catarina, Brasil, eugeniomonteiro@hotmail.com. (2) Brasil, cezar.karpinski@gmail.com (3) Brasil, moises.dutra@ufsc.br

Resumo

Apresenta o contexto *big data* e sua relação com o conhecimento científico a partir da seguinte pergunta: A extração de informações de grandes volumes de dados representa uma mudança epistemológica para a ciência de forma geral? O objetivo geral é o de refletir sobre as implicações epistemológicas do contexto *big data* a partir das proposições do epistemólogo Karl Popper. Especificamente, os objetivos são: a) discutir as implicações teóricas da ciência no campo da epistemologia; b) discorrer sobre os conceitos do que se convencionou chamar de "*big data*"; c) analisar os possíveis impactos do contexto *big data* ao conhecimento científico. Metodologicamente, trata-se de um estudo exploratório e descritivo de caráter teórico para auxiliar a reflexão crítica acerca do fenômeno *big data* e suas consequências para o fazer científico de qualquer área do conhecimento. Como resultado apresenta-se uma análise crítica acerca de um fenômeno pouco estudado no aspecto epistemológico. Conclui-se que o contexto *big data* tem revolucionado os processos de tomada de decisão nas empresas, mas não é possível afirmar que a mesma revolução ocorre no aspecto epistemológico.

Palavras-chave: Big data; Ciência; Epistemologia; Karl Popper

Abstract

It presents the big data context and its relationship with scientific knowledge from the following question: Does the extraction of information from large volumes of data represent an epistemological change for science in general? The general objective is to reflect on the epistemological implications of the big data context based on the propositions of Karl Popper. Specifically, the objectives are: a) to discuss the theoretical implications of science in the field of epistemology; b) to discourse about the concepts of what was conventionally called "big data"; c) to analyze the possible impacts of the big data context on scientific knowledge. Methodologically, it is an exploratory and descriptive research of the theoretical character to support the critical reflection about the big data phenomenon and its consequences for the scientific practice of any area of knowledge. The result is a critical analysis of a phenomenon little studied in the epistemological aspect. It is concluded that the big data context has revolutionized the decision-making processes in companies, but it is not possible to say that the same revolution occurs in the epistemological field.

Keywords: Big data; Science; Epistemology; Karl Popper

Silva Júnior, Eugênio Monteiro, Karpinski, Cezar, and Dutra, Moisés Lima. Conhecimento científico no context big data: reflexões a partir da epistemologia de Popper. *Brazilian Journal of Information Science: Research trends*, vol. 14, no. 4, set.-dez. 2020, e020017 https://doi.org/10.36311/1940-1640.2020.v14n4.10936

1 Introdução

A partir do Século XXI, a informação passou a ser representada sob a forma digital em grande escala. A tecnologia da informação está cada vez mais presente nas diversas atividades humanas. Isso fica evidente nas transações bancárias, comércio eletrônico, educação à distância, entretenimento, etc. No ambiente científico, essa tecnologia também tem ganhado espaço. Isso significa que a coleta e processamento de dados realizados por computador ganhou importância não apenas social como também científica. Mas, segundo Sayão e Sales (2019), esse protagonismo dos dados não é um fenômeno atual, pois existe um contexto anterior de utilização dos dados pelos governos, empresas, pesquisadores e outros segmentos da sociedade com o objetivo de direcionar a tomada de decisões e fundamentar descobertas. O que ocorreu nas últimas duas décadas foi um aumento na disponibilidade dos dados, passando da escassez para a extrema abundância (Victorino, et al., 2019; Falsarella et al. 2017)

Essa grande disponibilidade de dados, aliada a poderosos métodos de análise computacionais e hardware de alta capacidade de processamento, certamente representam um forte aliado para a ciência de forma geral. Entretanto, a questão que surge é o quão revolucionário é esse avanço. A extração de informações de grandes volumes de dados representa uma mudança epistemológica para a ciência de forma geral?

Para tentar compreender o nível dessa mudança é necessário entender o conceito de epistemologia e ciência, para, então discutir o impacto da ciência orientada por dados. Nesse sentido, este artigo tem como objetivo geral refletir sobre as implicações epistemológicas do contexto big data a partir das proposições do epistemólogo Karl Popper. Especificamente, os objetivos são: a) discutir as implicações teóricas da ciência no campo da epistemologia; b) discorrer sobre os conceitos do que se convencionou chamar de “big data”; c) analisar os possíveis impactos do contexto big data ao conhecimento científico.

A partir de um estudo exploratório e descritivo, a pesquisa se justificou pela necessidade de levantar referenciais teóricos capazes de auxiliar a reflexão crítica acerca do big data e suas consequências para o fazer científico de qualquer área do conhecimento. À Ciência da Informação, este estudo pode servir de apoio às pesquisas desenvolvidas em todos os níveis de pesquisa que problematizam o contexto big data na interface da produção do conhecimento científico da área.

Como conclusão, aponta-se que o contexto big data tem auxiliado os processos de tomada de decisão em organizações que dispõem de recursos financeiros para investimento nesse nicho tecnológico. Entretanto, não é possível afirmar que a mesma revolução ocorre no aspecto epistemológico. Apesar da constatação de que o big data pode fazer o indutivismo prevalecer sobre o método científico atual, ainda não existem argumentos plausíveis para sustentar essa perspectiva no campo epistemológico.

2 Aspectos Metodológicos

A pesquisa caracteriza-se como bibliográfica, pois tem como base livros e artigos que abordam os temas epistemologia e big data. Para obter as fontes, foram realizadas consultas no Google Acadêmico, Web of Science, Scopus, Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (Brapci) e Library, Information Science & Technology Abstracts with Full Text (LISTA). As três primeiras bases de dados foram escolhidas por concentrar publicações nacionais e internacionais no campo multidisciplinar. Já a quarta e a quinta bases foram escolhidas por concentrar publicação de artigos científicos específicos da área de Ciência da Informação no Brasil e no exterior.

No Google Acadêmico a busca se deu pelos termos epistemology and big data (sem aspas), com recorte temporal de 2000 e 2019, classificando os resultados por relevância, em qualquer idioma e sem a inclusão de patentes e citação. O buscador recuperou um total aproximado de 114 mil documentos e, como artigos que mais se aproximavam dos objetivos desta pesquisa, foram selecionados para a leitura os trabalhos de Kitchin (2014) e Frické (2015). O primeiro apresenta discussões específicas sobre as possibilidades de mudanças epistemológicas promovidas pelo big data. O segundo estabelece uma posição crítica sobre a produção de conhecimento científico por meio do Big Data, uma vez que a ciência, em seus pressupostos epistemológicos, não pode se basear, de forma passiva, em simples análise de dados. O estudo de Kitchin (2014) utiliza como base epistemológica o trabalho de Kuhn (1962). Já Frické (2015), apoia o seu estudo em filósofos e epistemólogos como Floridi (2012), Lakatos (1970; 1974a.; 1974b) e Popper (1959; 1963). Considerando os resultados de Kitchin (2014) e Frické (2015), optou-se por adotar como ponto de partida para as reflexões deste artigo a contribuição do primeiro, mas com um questionamento suscitado pela leitura do segundo: existem artigos que relacionam big data e a epistemologia de Popper na área da

Ciência da Informação? Para tanto, optou-se por recorrer, primeiramente, à produção nacional indexada na BRAPCI.

Na BRAPCI, na data de 19/02/2020, a pesquisa com os termos "epistem*", "big data" e "Popper" não retornou nenhum artigo. Nessa mesma data e diminuindo o recorte para os termos "epistem*" e "big data", foi recuperado nesta base um total de três artigos (Furlan and Laurindo 2017; Sayão and Sale, 2019; Noiret 2015).

Com este resultado, optou-se em ampliar a pesquisa para as bases Web of Science, Scopus e LISTA, a partir dos termos "epistemology", "big data" e "Popper". Na primeira, a pesquisa pelos termos supracitados se deu no dia 18 de novembro de 2020, em pesquisa básica, no campo tópico (pelo seu caráter abrangente) e sem recorte temporal. Na segunda, a pesquisa se deu no mesmo dia, na aba documents, no campo Article titles, Abstract, Keywords. Da mesma forma, a pesquisa na LISTA se deu no mesmo dia e em busca básica. Nas três bases foi recuperado o mesmo artigo de Van Poucke, S. et al. (2016), da área médica e que usa o método do falsificacionismo de Popper aplicado aos estudos de ensaios clínicos randomizados a partir de análise de big data. Por isso foi recuperado pelos termos, mas por não se referir aos objetivos da pesquisa foi descartado.

Esse fato é um indício de que artigos de Ciência da Informação que relacionam a epistemologia de Popper com big data ainda são raros. Desse modo, esse trabalho tem como objetivo preencher essa lacuna tendo como base o trabalho de Kitchin (2014) e suas referências promovendo um diálogo com o livro "A lógica da pesquisa científica" de Karl Popper. Para as discussões sobre epistemologia foram utilizados como referências Japiassu (1977) e Dutra (2010). Já para os conceitos e aplicações da categoria "ciência" foram fundamentais as constatações Chalmers (1999).

3 Epistemologia e Ciência

Para saber se *big data* e novas técnicas de análise de dados são revolucionárias do ponto de vista epistemológico, é necessário saber, antes, o que é epistemologia. Essa é uma palavra que não tem um conceito simples. Segundo Japiassu (1977), etimologicamente, "epistemologia" significa discurso (*logos*) sobre a ciência (*episteme*). Apesar de parecer um termo antigo, sua criação é relativamente recente, pois surgiu no século XIX no vocabulário

filosófico. “Epistemologia é o estudo crítico dos princípios, das hipóteses e dos resultados das diversas ciências” (Japiassu 1977 p. 25).

O conceito de epistemologia é empregado de modo bastante flexível. Entretanto, ela não pode e nem pretende impor dogmas aos cientistas. Não pretende ditar de forma autoritária o que deveria ser o conhecimento científico. Seu papel é estudar a gênese e a estrutura dos conhecimentos científicos. Mais precisamente, o de tentar pesquisar as leis reais de produção desses conhecimentos. A epistemologia procura estudar a produção dos conhecimentos, tanto do ponto de vista lógico, quanto dos pontos de vista linguístico, sociológico, ideológico, etc. Daí seu caráter interdisciplinar (Japiassu 1977).

Segundo Dutra (2010), o termo epistemologia é geralmente empregado para fazer referência à “teoria do conhecimento” e a melhor maneira de saber o que é a epistemologia consiste em examinar o que os estudiosos dedicados a essa disciplina fizeram e fazem. Nesse aspecto, a epistemologia agrega em seu campo, a identificação e problematização dos princípios, da história e da filosofia da ciência em seus aspectos gerais e específicos.

Portanto, epistemologia refere-se a como o conhecimento científico é construído. Dessa maneira, dizer que o *big data* representa uma mudança epistemológica significa dizer que promoverá revolução na forma como o conhecimento científico é produzido. Nas próximas seções, essa proposição será analisada criticamente sob o ponto de vista das teorias do epistemólogo Karl Popper.

Já “ciência” é um termo que parece ter uma definição óbvia, mas ao tentar elaborar uma definição formal para essa palavra, qualquer pesquisador se depara com inúmeras possibilidades e questionamentos. O conceito de ciência geralmente é empregado como sinônimo de saber ou conhecimento (do latim *scientia*). Porém, num sentido mais restrito, ciência refere-se a uma forma específica de conhecimento: o conhecimento científico (Barbieri 1990).

Nos tempos modernos, uma concepção que se tornou aceitável foi a de que o conhecimento científico é apenas o conhecimento provado. Nessa concepção, as teorias científicas são oriundas de uma maneira rigorosa de obtenção de dados da experiência adquiridos por observação e experimento. A ciência moderna, de acordo com a crítica feita por Chalmers (1999), foi concebida como o conhecimento objetivo, onde não poderia haver espaço para opiniões pessoais ou suposições. Essa parece ser a visão definitiva do significado de ciência, entretanto, trata-se de apenas uma das vertentes possíveis.

Ao refletir sobre o campo epistemológico da Ciência da Informação, Rendón-Rojas (2008) estabelece as principais características do conhecimento científico da seguinte forma:

a) É um sistema de conhecimentos a respeito de uma parte da realidade.

b) Tem um objeto de estudo determinado.

c) Seus conhecimentos são verdadeiros.

d) Seus conhecimentos são justificados (metodologia).

e) Possui um corpo teórico.

Assim, é premissa da ciência ser um sistema de conhecimentos, mas não de todo o tipo de conhecimento, apenas daquele que pode ser entendido como uma crença verdadeira e justificada. Assim, a ciência sempre se diferencia da opinião, pois é verdadeira e apresenta razões para justificar essa crença. Além disso, a ciência investiga uma parte da realidade e essa parte constitui seu objeto de estudo. Cada ciência possui um objeto de estudo e o analisa de uma perspectiva particular. Dizer que o conhecimento científico possui uma estrutura teórica definida significa que ele possui conceitos, enunciados gerais e uma inter-relação entre eles. A base teórica de uma ciência constitui uma rede de conceitos e proposições interdependentes. (Rendón-Rojas 2008; 2012)

As funções epistemológicas de uma ciência são a explicação e a predição científica, assim como a compreensão. Hempel (1979) afirma que a explicação científica consiste em deduzir a partir de leis gerais e condições iniciais um determinado fenômeno. Já a predição científica é o mesmo processo da explicação, porém, quando o fenômeno que é deduzido não está presente. A predição é que permite manipular a realidade para obter ou evitar os fenômenos previstos de acordo com a teoria. A compreensão tem lugar nas ciências sociais e humanas nas quais não é possível descobrir leis gerais que permitam a explicação e a predição científica, já que o objeto de estudo é um sujeito que tem variáveis difíceis de controlar como a liberdade, a imaginação e as emoções. Entretanto, o conhecimento deste sujeito não é impossível, mas é dado com base no sentido de suas ações.

A outra característica inerente ao conhecimento científico é a sua justificação, ou seja, dar razões para o que se afirma. É por meio do método que se alcança a justificação e é a sua utilização que possibilita a replicação dos resultados de um experimento por parte da comunidade científica (Redón-Rojas 2008). A metodologia é, portanto, um integrante essencial na construção do conhecimento científico, porém, toda metodologia depende de

uma epistemologia. De acordo com os pressupostos epistemológicos de que se parta, serão as exigências metodológicas que se terá (Redón-Rojas 2008). Desse modo, a metodologia merece uma investigação mais profunda, pois não há um consenso em relação a esse tema.

Redón-Rojas (2008) defende a posição de que é necessária a existência de um método para justificar o conhecimento e a existência da comunidade científica, mas neste caso a palavra "um" é um artigo indeterminado e não um numeral. Desse modo, o referido método não é único e absoluto, emprega-se aquele que esteja de acordo com o objeto de estudo. Esse autor cita como exemplo, que estudar a Lua como objeto da astronomia ou como objeto de culto em uma cultura, não é algo que possa ser feito utilizando a mesma metodologia.

4 Big Data

Esta seção se dedica a fornecer uma visão geral de Big data, um tema relativamente novo e ainda sem uma conceituação definitiva. A ausência de uma definição conceitual e da compreensão social de seus princípios faz com que várias perguntas ainda não possam ser respondidas sobre o Big data. Nesse sentido, serve como parâmetro das discussões conceituais acerca da temática o trabalho de Mazieri e Soares (2016) que, por meio de uma pesquisa bibliométrica, forneceram subsídios relevantes para a conceituação do termo Big data.

No campo da Ciência da Informação, segundo Costa et al (2020), é possível considerar que big data também é objeto de estudo, uma vez que o objetivo dessa ciência e concentrar esforços para resolver problemas informacionais.

O desafio fundamental para as aplicações de big data é explorar os grandes volumes de dados para extrair informações ou conhecimentos úteis para ações futuras (Rajaraman and Ullman 2011). Entretanto, o termo big data não deve ser entendido exclusivamente como um grande volume de dados. Assim como em muitos conceitos que emergem rapidamente, big data foi definido de várias formas, desde definições triviais de que big data consiste em um conjunto de dados muito grande para caber em uma planilha ou ser armazenada em uma única máquina até definições mais detalhadas. Em busca de uma definição mais completa, Kitchin (2013) realizou uma pesquisa na literatura e elaborou uma lista com as seguintes características que normalmente são utilizadas para definir big data:

Enorme em volume, composto por terabytes ou petabytes de dados;

Alta velocidade, sendo gerados praticamente em tempo real;

Diverso, sendo de natureza estruturada ou não estruturada;

Exaustivo em escopo, pois busca capturar populações ou sistemas inteiros;

Refinado em resolução, ou seja, busca ser o mais detalhado possível;

De natureza relacional, contendo campos comuns que permitem a junção de diferentes conjuntos de dados.

Flexível, sendo possível a adição de novos campos e escalável, permitindo a expansão do seu tamanho de maneira rápida.

Alguns autores adotam três letras "v" como uma forma de descrever big data representando as iniciais das palavras volume, velocidade e variedade. Apesar de não haver consenso sobre a utilização destes três "v" na literatura científica, percebe-se que o mercado passou a adotar essa concepção como constituinte do contexto de Big data. Grandes empresas de tecnologia, por exemplo, adicionaram mais três letras "v" para complementar o conceito de big data. International Business Machines Corporation (IBM) adicionou o v de veracidade, pois lidar com a incerteza e a imprecisão é outra faceta do big data. A Statistical Analysis System (SAS) introduziu o v de variabilidade (e complexidade). Variabilidade refere-se à variação nas taxas de fluxo de dados. Frequentemente, a velocidade de geração do big data não é constante e tem altos e baixos periódicos. Complexidade refere-se ao fato de que grandes volumes de dados são gerados por uma infinidade de fontes. Isso impõe um desafio crítico: a necessidade de conectar, combinar, limpar e transformar dados recebidos de diferentes fontes. O último v foi proposto pela Oracle e é referente a valor. Com base na definição dada por essa empresa, o big data geralmente é caracterizado pela "baixa densidade de valor", ou seja, na sua forma original, os dados têm pouco valor em relação ao volume. No entanto, é possível conseguir um alto valor analisando grandes volumes desses dados (Gandomi and Haider 2015).

Wu et al. (2013) utiliza a sigla "hace" para listar as características do big data: grande volume de dados, heterogêneos, com fontes autônomas distribuídas e de controle descentralizado e tem o objetivo de explorar relações complexas e em evolução entre os dados.

É importante ressaltar que dois conjuntos de dados do mesmo tamanho podem exigir diferentes tecnologias de gerenciamento de dados dependendo do seu tipo, por exemplo,

dados tabulares e dados de vídeo. Assim, as definições de big data dependem da indústria de tecnologia. É impraticável estabelecer um limiar específico para o que é ou não big data (Gandomi and Haider 2015).

Qual a utilidade do big data? Como esses dados podem ser operacionalizados de modo a fornecerem informações potencialmente úteis? Pelo que se percebe na literatura e na prática, o big data é inútil no vácuo. Seu valor potencial é desbloqueado apenas quando é utilizado para impulsionar a tomada de decisão. Para permitir essa tomada de decisão baseada em evidências, as organizações precisam de processos eficientes para transformar grandes volumes de dados velozes e diversos em informações precisas. De acordo com Gandomi e Haider (2015), todo o processo de extração de informações de big data pode ser dividido em cinco etapas: 1) aquisição e gravação; 2) extração, limpeza e anotação; 3) integração, agregação e representação; 4) modelagem e análise; 5) interpretação. Essas cinco etapas podem ser agrupadas em dois subprocessos principais: gerenciamento e análise de dados. O gerenciamento de dados envolve os processos e as tecnologias de suporte para coletar, armazenar e preparar os dados para serem analisados. A análise dos dados, por outro lado, compreende as técnicas utilizadas para obter informações úteis a partir do big data.

Para trabalhar com esses grandes volumes de dados para obter informações de valor, são necessárias técnicas próprias. Essas técnicas são diferentes das utilizadas nas análises de dados desenvolvidas no passado. Tradicionalmente, as técnicas de análise de dados forma projetadas para extrair informações de conjuntos de dados escassos, estáticos, limpos e pouco relacionados, amostrados cientificamente e respeitando premissas estritas (independência e normalidade), gerados e analisados com uma questão específica em mente (Miller 2010).

Não faz parte do escopo desse artigo, detalhar cada uma dessas técnicas de análise de dados, mas é possível citar algumas. Gandomi e Haider (2014) fazem uma breve descrição de várias técnicas de análise de dados estruturados e não estruturados. Entre elas, é possível destacar a mineração de texto, análise de áudio e de vídeo e análise de mídias sociais.

O big data não se refere exclusivamente a grandes coleções de dados e às ferramentas e procedimentos para manipulá-los e analisá-los, mas é também uma mudança fundamentada no modo de pensar e pesquisar cientificamente. Uma possível evidência desse fato é que coleções de dados que geralmente são consideradas big data, como os dados do Twitter, não são maiores que os conjuntos de dados gerados ou coletados por sistemas anteriores que não eram considerados big data, como os dados do censo (Sayão and Sales 2019).

6 Questões Epistemológicas

O aumento da geração de grandes volumes de dados digitais e o desenvolvimento de novas técnicas computacionais para analisá-los tem proporcionado enormes avanços no meio corporativo. É muito comum encontrar aplicações desse tipo no comércio eletrônico, por exemplo. No meio científico, também se observa o aumento de pesquisas que utilizam análise de big data para alcançar seus objetivos. No entanto, isso permite dizer que, do ponto de vista epistemológico, existe de fato uma revolução em curso?

É importante esclarecer que este artigo se refere somente ao grau do impacto que o big data pode provocar na ciência em sentido estrito. Não está entre os objetivos deste artigo questionar a utilidade do big data nos processos de tomada e decisão em empresas, por exemplo. Entretanto, as aplicações meramente práticas ou comerciais, ainda que tenham seu próprio mérito e algum embasamento científico, não serão consideradas para efeito dessa análise.

Um bom ponto de partida para a discussão epistemológica é o ensaio publicado por Chris Anderson na Wired Magazine (Anderson 2008) com o título que pode ser traduzido como "O fim da teoria: o dilúvio de dados torna obsoleto o método científico". Esse texto gerou uma série de críticas na literatura científica justamente por sugerir que o big data representaria a fim da teoria na ciência. O título aparenta trazer uma ideia inovadora e radical. Entretanto, segundo Mazzocchi (2015), essa proposta de relegar as hipóteses a um papel secundário é ainda mais antiga. O próprio Francis Bacon, considerado o "pai do método científico", argumentou na sua obra Novum Organum, de 1620, que o conhecimento científico não deveria se basear em noções preconcebidas, mas em dados experimentais. O raciocínio dedutivo é eventualmente limitado, pois estabelecer uma premissa antes de um experimento restringiria o raciocínio de modo a coincidir com essa premissa.

No contexto atual do big data, observa-se que alguns autores consideram que pode estar ocorrendo uma mudança no método científico, uma vez que, literalmente, centenas de algoritmos diferentes podem ser aplicados a um conjunto de dados para determinar o melhor ou um modelo ou explicação composta (Siegel 2013), uma abordagem radicalmente diferente daquela utilizada tradicionalmente, na qual o analista seleciona um método apropriado com base em seu conhecimento das técnicas e dos dados. Em outras palavras, a análise de big data permite uma abordagem epistemológica inteiramente nova para dar sentido ao mundo. Em

vez de testar uma teoria analisando dados relevantes, as novas análises de dados buscam obter intuições "nascidas dos dados" (Kitchin 2014).

Com base nesses textos, é possível identificar ao menos duas vertentes quando o assunto é a revolução epistemológica provocada pelo big data: o renascimento do empirismo e o desenvolvimento de uma metodologia inteiramente nova. Nas próximas subseções, são apresentados os argumentos de cada uma dessas vertentes para, então, serem discutidos conforme a proposta epistemológica de Karl Popper.

6.1 Epistemologia Empirista

Alguns autores sugerem que o *big data* inaugura uma nova era do empirismo, em que o volume de dados, acompanhado por técnicas que podem revelar sua verdade inerente, permite que os dados falem por si próprios, sem a necessidade de teoria. A visão empirista ganhou credibilidade fora da academia, especialmente nos círculos empresariais, mas suas ideias também se enraizaram no novo campo denominado Ciência de Dados e de outras ciências relacionadas à temática (Kitchin 2014). Na peça provocativa de Anderson (2008), na qual ele argumenta que "o dilúvio de dados torna o método científico obsoleto"; que os padrões e relacionamentos contidos no *big data* produzem conhecimento significativo a respeito de fenômenos complexos. Argumentando essencialmente que o *big data* permite um modo empirista de produção de conhecimento, ele afirma: os *petabytes* permitem dizer que correlação é suficiente, é possível analisar os dados sem hipóteses a respeito do que podem mostrar (Kitchin 2014). Segundo Anderson (2008), diante dos grandes volumes de dados a abordagem científica tradicional – hipótese, modelo e teste – está se tornando obsoleta.

Essa ideia geralmente vem das análises de dados utilizadas no *marketing* e no comércio nas quais não existe hipótese para explicar porque o produto A é comprado geralmente junto com o produto B (Kitchin 2014). Nessas situações, muitas vezes, encontrar correlações é suficiente, ou seja, não há necessidade de entender o "porquê". No entanto, quando o assunto é ciência, não se pode adotar um método tão simplório e abandonar de maneira definitiva as teorias. De acordo com Popper (2001), a missão do cientista é a de buscar leis que o habilitem a deduzir previsões. Essa missão compreende duas partes: De um lado, ele deve tentar descobrir leis que lhe deem condição para deduzir previsões isoladas (leis "causais" ou "deterministas" ou "enunciados de precisão"); De outro lado, deve tentar formular hipóteses acerca de frequências, ou seja, leis que assegurem probabilidades, a fim de deduzir previsões de frequência. Nada há nessas duas tarefas que as torne incompatíveis.

De acordo com Kitchin (2014), existe um conjunto poderoso e atraente de ideias presentes na epistemologia empirista que vai contra a abordagem dedutiva que é hegemônica na ciência moderna:

1) O *big data* pode capturar um domínio inteiro e fornecer resolução total;
2) Não há necessidade de teorias, modelos ou hipóteses *a priori*;
3) Por meio da aplicação de análise de dados agnóstica, os dados podem falar por si próprios, sem ideias preconcebidas ou enquadramentos humanos, e que quaisquer padrões e relacionamentos dentro *big bata* são inerentemente significativos e verdadeiros;
4) 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 (Kitchin 2014).

Ainda segundo Kitchin (2014), embora os argumentos dessa epistemologia empirista pareçam atrativos, eles estão fundamentados em ideias falaciosas. A partir dos contrapontos apresentados por Amin e Thrifit (2002) e de Haraway (1991), pode-se inferir que o big data não contempla todo o domínio, os dados sempre proporcionam um ponto de vista específico dependendo das ferramentas utilizadas. Dizer que a análise de um grande volume de dados permite encontrar padrões gerais a respeito de um assunto nada mais é do que evocar o método indutivo. Segundo Popper (2001), são chamadas de "indutivas" as inferências que partem de "enunciados singulares" ou "particulares", como descrições de resultados de observações ou experimentos, para "enunciados universais", tais como hipóteses ou teorias.

A ideia de que o big data é suficientemente numeroso de modo a permitir a extração de teorias, ou seja, enunciados universais, também é errônea a partir do ponto de vista de Popper (2001). Segundo ele, ao inferir enunciados universais a partir de enunciados singulares, independentemente de quão numerosos sejam estes, qualquer conclusão obtida dessa maneira sempre pode se revelar falsa. Independentemente de quantos cisnes brancos forem observados, não será possível concluir que todos os cisnes são brancos. Popper (2001) rejeita inclusive a ideia de probabilidade. Segundo ele, não há nenhuma vantagem em afirmar que o princípio da indução em vez de ser "verdadeiro" é "provavelmente" verdadeiro. "Nunca suponho que, por força de conclusões 'verificadas', seja possível ter por verdadeiras ou mesmo por meramente 'prováveis' quaisquer teorias" (Popper 2001 p. 34).

Para Popper (2001) é inadmissível a existência da indução, ou seja, produzir teorias a partir de enunciados singulares "verificados por experiência". Ele afirma categoricamente que teorias nunca são empiricamente verificáveis. Ele faz uma ressalva apenas em relação à indução matemática, que não deve ser contestada. Popper (2001) não exige que um sistema científico seja suscetível de ser dado como válido, de uma vez por todas, em sentido positivo. Porém, exige que sua forma lógica seja tal que se torne possível validá-lo por meio de provas empíricas, em sentido negativo: deve ser possível refutar, pela experiência, um sistema científico empírico.

O segundo ponto é que o Big data surge de uma necessidade muito bem delimitada no tempo e no espaço. Isso porque, os sistemas são projetados para capturar certos tipos de dados e as análises e algoritmos usados são baseados em raciocínio científico e foram aprimorados por meio de testes científicos. Como tal, uma estratégia indutiva de identificação de padrões nos dados não ocorre no vácuo científico e é discursivamente enquadrada por descobertas, teorias e treinamentos anteriores. Além disso, resultam de especulações baseadas em experiência e conhecimento (Leonelli 2012). A tese da "não teoria" desconhece o fato de que a ciência não coleta dados aleatoriamente e que os experimentos científicos são projetados dentro de escopos teóricos, metodológicos e instrumentais bem definidos (Sayão and Sales 2019). Por que coletar certas informações em vez de outras? Por que usar determinadas palavras-chave para organizar a pesquisa e não outras? Toda escolha que se faz a esse respeito é um reflexo de um conjunto, muitas vezes não declarado, de suposições e hipóteses sobre o que se quer e se espera dos dados. Sem modelos, matemático ou conceitual, os dados são apenas ruídos (Pigliucci 2009).

Um exemplo muito conhecido é a descoberta do bóson de Higgs. Só foi possível provar a existência dessa partícula por meio do Large Hadron Collider (LHC), o mais poderoso acelerador de partículas do mundo e a maior máquina de propósito específico já construída pela humanidade. O LHC gera mais de 600 milhões de colisões por segundo e produz 15 petabytes de dados por ano. Para encontrar os traços das partículas elementares é necessário examinar essa vasta quantidade de dados procurando por padrões específicos. Essa tarefa fica a cargo do Worldwide LHC Computing Grid (WLCG) que conecta centenas de centros de processamento de dados em todo o mundo. O desempenho do WLCG é essencial para apoiar os experimentos do LHC e fornecer resultados de maneira rápida. Big data, computação distribuída e sofisticadas análises de dados foram empregadas em conjunto para a descoberta do bóson de Higgs – e talvez na descoberta de novos "padrões" que também

podem gerar novas hipóteses nesse campo. Apesar disso, a descoberta do bóson de Higgs não foi orientada por dados. Os experimentos conduzidos no LHC obedeciam às previsões teóricas que indicavam a existência da última partícula que faltava para completar o Modelo Padrão de partículas elementares (Mazzocchi 2015). Esse exemplo demonstra que, nem sempre, uma pesquisa conduzida envolvendo grandes volumes de dados pode abandonar a teoria.

O terceiro ponto é que assim como os dados não são gerados livres de teoria, eles também não podem simplesmente falar por si mesmos livres de ideias preconcebidas. Os dados são examinados por meio de uma lente particular que influencia em como eles são interpretados. Além disso, correlações entre variáveis podem ser aleatórias na natureza e não apresentarem qualquer relação causa e efeito (Kitchin 2014). Para a ciência, não basta encontrar correlações, é necessário estabelecer uma relação causal.

Por último, a ideia de que os dados falam por si mesmos sugere que qualquer pessoa com razoável entendimento de estatística seja capaz de interpretá-los sem conhecimento do contexto ou do domínio em que os dados foram gerados (Kitchin 2014). Entretanto, na maioria dos casos, será necessário ter conhecimento do domínio em que os dados foram coletados ou gerados, por exemplo, uma pessoa sem conhecimento suficiente de química não conseguirá interpretar qualquer insight a partir de dados desse contexto.

Anderson corajosamente fecha seu pedaço de bravata epistêmica afirmando que:

> [...] nova disponibilidade de grandes quantidades de dados [...] oferece uma maneira totalmente nova de entender o mundo. A correlação substitui a causalidade, e a ciência pode avançar mesmo sem modelos coerentes, teorias unificadas ou realmente nenhuma explicação mecanicista (Anderson 2008, p.4, tradução nossa).

No entanto, de acordo com Pigliucci (2009), a ciência só avança se puder fornecer explicações, sem isso, torna-se uma atividade mais parecida com a coleta de selos. Para esse autor, agora, há uma área em que *petabytes* de informação podem ser usados por eles mesmos, mas não deve ser chamada de ciência. Segundo Sayão e Sales (2019), o ensaio de Anderson (2008) tem o único mérito de colocar em pauta debates mais precisos e fundamentados a respeito de como a disponibilidade massiva de dados associada aos novos métodos computacionais de análise desafia o percurso secular da metodologia científica.

6.2 Ciência orientada por dados

Segundo Kitchin (2014), em contraste com as novas formas de empirismo, a ciência orientada por dados procura se apegar aos princípios do método científico, mas é mais aberta ao uso de uma combinação de abordagens abdutivas, indutivas e dedutivas para avançar na compreensão de um fenômeno. Cabe destacar que, de acordo com Morin (1996), foi Charles Sanders Peirce (1839-1914) que introduziu a palavra abdução para se referir à invenção de hipóteses.

Em outras palavras, procura incorporar um modo de indução ao desenho da 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 está situado e contextualizado dentro de um domínio teórico altamente evoluído. Como tal, a estratégia epistemológica adotada na ciência orientada por dados é usar técnicas guiadas de descoberta de conhecimento para identificar possíveis questões (hipóteses) dignas de mais exames e testes (Kitchin 2014).

De fato, muitos supostos relacionamentos dentro de conjuntos de dados podem ser rapidamente descartados como triviais ou absurdos por especialistas do domínio, com outros sinalizados como merecedores de mais atenção (Miller 2010). Mazzocchi (2015) cita uma frase do livro de Mayer-Schonberger e Cukier (2013) que afirma que as correlações podem não dizer precisamente porque algo está acontecendo, mas elas alertam que algo está acontecendo. E, em muitas situações, isso é suficiente. Entretanto, na maioria dos casos, entender o porquê é crucial para atingir um nível de conhecimento que possa ser utilizado para produzir previsões confiáveis. Desse modo, as análises de big data, podem servir para o levantamento de hipóteses que, posteriormente, podem ser estudadas com maior rigor.

De acordo com Kitchin (2014), 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. Mas quais a características as hipóteses levantadas a partir do big data devem ter? Essas hipóteses precisam ser objetivas, passíveis de serem reproduzidas por outros pesquisadores e, principalmente, falseáveis.

Popper (2001) utiliza os termos “objetivo” e “subjetivo” da mesma maneira que Kant. Ele utiliza “objetivo” para indicar que o conhecimento científico deve ser justificável, independentemente de capricho pessoal; uma justificação será “objetiva” se puder, em

princípio, ser submetida à prova e compreendida por todos. O fato de ser falseável é o principal requisito da epistemologia de popperiana. Como exemplo, Popper (2001) apresenta o enunciado "choverá ou não choverá aqui amanhã" como não empírico, uma vez que não admite refutação, ao passo que o enunciado "choverá aqui, amanhã" é empírico. O segundo enunciado é falseável, pois basta "não chover aqui, amanhã" para contestá-lo.

As teorias da Ciência Natural e, em particular, as que são denominadas de leis naturais, tem a forma lógica de enunciados estritamente universais. Assim, podem ser expressas sob forma de negações de enunciados estritamente existenciais ou, caberia dizer, sob a forma de enunciados de não-existência (ou enunciados-não-há). A lei da conservação da energia, por exemplo, admite ser expressa sob a forma: "não existe máquina de movimento perpétuo". Esses enunciados são como proibições e, por agirem nesse sentido, é que são falseáveis. Por outro lado, os enunciados existenciais não podem ser falseados. Por exemplo, para contradizer o enunciado existencial "há corvos brancos", seria necessário um enunciado universal (Popper 2001). Em outras palavras, haveria a necessidade de observar todos os corvos do planeta, em todas as épocas e não encontrar nenhum corvo branco, somente assim o enunciado estaria refutado.

Uma teoria é dita falseada somente quando existem enunciados básicos aceitos que a contradigam. Essa condição é necessária, porém não é suficiente. O falseamento só é aceito se uma hipótese empírica de baixo nível, que descreve esse efeito, for proposta e corroborada. Para falsear o enunciado "todos os corvos são negros", bastaria um enunciado suscetível de teste como "no zoológico de Nova Iorque existe uma família de corvos brancos" (Popper 2001).

Enunciados básicos têm a forma de enunciados existenciais singulares. Enunciados básicos indicam que um evento observável está ocorrendo em certa região individual do tempo e do espaço (Popper 2001). Esse mesmo autor ressalta que as teorias não são verificáveis, mas podem ser "corroboradas". Uma teoria está "corroborada" enquanto ela resistir aos testes.

Desse modo, toda teoria é temporária. Enquanto não existir um enunciado básico capaz de refutar determinada teoria, ela permanece válida. Além do caráter temporário das teorias na ciência, não se deve abandonar a busca pela causalidade. No entanto, as correlações podem ser úteis. É necessário, porém, estar atento às correlações espúrias. As correlações desempenham um papel importante como dispositivos heurísticos e precisam ser

melhor analisadas - usando modelos e experimentos – com o objetivo de distinguir entre correlações significativas e espúrias. De acordo com Mazzocchi (2015), um exemplo de correlação espúria vem de técnicas de mineração de dados financeiros que apontaram forte associação estatística entre eventos que não apresentam qualquer relação. Apesar de todo o auxílio que as informações extraídas do big data podem dar ao pesquisador, deve-se ter em mente que as teorias não podem ser excluídas do processo.

7 Considerações Finais

O *big data* tem revolucionado os processos de tomada de decisão nas empresas, especialmente aquelas que possuem recursos financeiros para desenvolvimento de tecnologias e contratação de recursos humanos qualificados. Entretanto, não é possível afirmar que a mesma revolução ocorre no aspecto epistemológico. Apesar de algumas afirmações de que o *big data* pode fazer o indutivismo prevalecer sobre o método científico atual, ainda não existem argumentos plausíveis para sustentar essa ideia. Nota-se que há certo exagero no discurso dos que alegam que correlação é suficiente no contexto *big data* e que não há necessidade de estabelecer a relação 'causa e efeito'. Isso pode até ser válido para o contexto de recomendação de produtos no comércio eletrônico, por exemplo. Mas, os que tentam extrapolar esse pensamento para a ciência demonstram desconhecimento dos seus fundamentos.

Conforme o que é defendido por Popper, a ciência tem como objetivo estabelecer leis, ainda que essas leis sejam temporárias. No entanto, é fundamental que se encontre relação causal nos fenômenos estudados. Além disso, não é possível realizar uma pesquisa livre de teorias. A própria escolha dos dados que serão coletados ou dos termos de uma busca obedecem, ainda que de maneira informal, alguma teoria.

O que se percebe é que o *big data* pode auxiliar no levantamento de hipóteses para serem posteriormente avaliadas. É importante que essas hipóteses possam ser avaliadas por outros pesquisadores de maneira objetiva e que sejam falseáveis. Portanto, pelo menos por enquanto, a essa pesquisa demonstra que é um equívoco acreditar em uma 'revolução epistemológica' promovida pelo *big data*. Concorda-se que esse contexto promove maior agilidade nos processos de levantamento e avaliação de hipóteses. No entanto, o conhecimento científico permanece sendo construído com base em teorias e buscando relação causal entre os eventos, agora dispondo de novas ferramentas para auxiliar nessa busca.

Referências

Amin, A. and Thrift, N. *Cities: Reimagining the Urban*, Polity. 2002.

Anderson, C. "The end of theory: the data deluge makes the scientific method obsolete". *Wired Magazine*, 2008, https://www.wired.com/2008/06/pb-theory/. Acessado 15 fev. 2020.

Barbieri, J. C. *Produção e transferência de tecnologia*. Ática, 1990.

Chalmers, A. F. *O que ciência afinal?*. Brasiliense, 1993.

Costa et al. *Dados científicos: estudos práticos, teóricos e epistêmicos*. Ideia, 2020.

Dutra, L. H. D. *Introdução à Epistemologia*. Editora Unesp, 2010.

Falsarella, O. M., et al."Gestão estratégica empresarial: proposição de um modelo de monitoramento informacional na era do big data". *Revista Digital de Biblioteconomia & Ciência da Informação*, v. 15, n. 2, p. 420-441, 2017. DOI: 10.20396/rdbci.v15i2.8647124. Acesso em: 18 nov. 2020.

Floridi, L. "Big Data and their epistemological challenge". *Philosophy of Tecnology*, v.24, n.4, p.435-437, 2012.

Frické, Martin. "Big Data and Its Epistemology". *Journal of the Association for Information Science and Technology*, v.66, n.4, p.651-661, 2015. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/asi.23212. Acesso em: 18 nov. 2020.

Furlan, P. K.; Laurindo, F. J. B. "Agrupamentos epistemológicos de artigos publicados sobre big data analytics". *Transinformação*, v. 29, n. 1, p. 91-100, 2017. DOI: 10.1590/2318-08892017000100009. Acesso em: 18 nov. 2020.

Gandomi, A., and Haider, A. "Beyond the hype: Big data concepts, methods and analytics." *International Journal of Information Management*, v. 35, o. 2, 2015. DOI: https://doi.org/10.1016/j.ijinfomgt.2014.10.007. Acesso em: 15 mar. 2020.

Haraway, D. *Simians, Cyborgs and Women: The Reinvention of Nature.* New York: Routledge. 1991.

Hempel, C. G. La explicación científica: estudios sobre filosofía de la ciencia, Buenos Aires, Paidós. 1979.

Japiassu, H. *Introdução ao pensamento epistemológico*. Livraria Francisco Alves Editora. 1977.

Kitchin, R. "Big Data and Human Geography: Oportunities, Challanges and Risks." *Dialogues in Human Geography*, vol. 3 no. 3, 2013, pp. 262-267, doi: 10.1177/2043820613513388. Acessado 01 fev. 2020.

Kitchin, R. "Big Data, new epistemologies and paradigms shift." *Big Data and Society*, vol. 1, 2014, doi: 10.1177/2053951714528481. Acessado 10 dez. 2019.

Kuhn, T. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1962.

Lakatos, I. “Popper on demarcation and induction”. In Schilpp, P.A. (Ed.). *The philosophy of Karl Popper*. La Salle, IL: Open Court, 1974a.

Lakatos, I. “The role of crucial experiments in science”. *Studies in History and Philosophy of Science*, v. 4, n. 4, p. 309–325, 1974b.

Lakatos, I. Falsification and the methodology of scientific research programs. In. Lakatos, I ; Musgrave, A. E. (Eds.). *Criticism and the growth of knowledge*. Cambridge, England: Cambridge University Press, 1970. p. 91–195.

Leonelli, S. “Introduction: Making sense of data-driven research in the biological and biomedical sciences”. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 2012, doi: 10.1016/j.shpsc.2011.10.001. Acessado 25 jan. 2020.

Mayer-Schonberger, V. and Cukier, K., Big Data: a revolution that will transform how we live, work and think. New York: Houghton Mifflin Harcourt. 2013.

Mazieri, M. and Soares, E. D. “Conceptualization and theorization of the big data”. *International Journal of Innovation (IJI Journal)*, São Paulo, v. 4, n. 2, p. 23-41, 2016. Disponível em: https://periodicos.uninove.br/innovation/article/view/10171/4844. Acesso em: 18 nov. 2020.

Mazzocchi, R. “Could big data be the end of theory in science?” *EMBO Reports*, vol. 16, no. 10, 2015, doi: 10.15252/embr.201541001. Acessado 13 fev. 2020.

Miller H. J., “The data avalanche is here. Shouldn’t we be digging?”. *Journal of Regional Science*, vol. 50, 2010, doi: 10.1111/j.1467-9787.2009.00641.x. Acessado 21 jan. 2020.

Morin, E. O problema epistemológico da complexidade. Publicações Europa-América. 1996.

Noiret, S. História pública digital. *Liinc em revista*, v. 11, n. 1, 2015. DOI: 10.18617/liinc.v11i1.797. Acesso em: 18 nov. 2020.

Pigliucci, M. “The end of theory in science?” *EMBO Reports*, vol. 10, no. 6, 2009, doi: 10.1038/embor.2009.111. Acessado 01 mar. 2020.

Popper, K. R. *A lógica da pesquisa científica*. Edusp e Cultrix. 2001.

Popper, K. R. *Conjectures and refutations*. London: Routledge and Kegan Paul, 1963.

Popper, K. R. *The logic of scientific discovery*. London: Hutchinson, 1959.

Rajaraman, A. e Ullman, J. Mining of Massive Datasets, Cambridge University Press, 2011.

Rendón-Rojas, M. A. “Ciencia bibliotecológica y de la información en el contexto de las ciencias sociales y humanas. Epistemologia, metodologia e interdisciplina”, *Investigación Bibliotecológica*, 2008, http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-358X2008000100004&lng=es&nrm=iso. Acessado 20 dez. 2019.

Rendón-Rojas, M. A. “Epistemologia da Ciência da Informação: objeto de estudo e principais categorias”, *InCID: Revista de Ciência da Informação e Documentação*, v. 3, n. 1, p. 3-14, jan./jun. 2012, doi: https://doi.org/10.11606/issn.2178-2075.v3i1p3-14 Acessado 31 ago.2020.

Sayão, L. F., Sales L. F. “O fim da teoria: o confronto entre a pesquisa orientada por dados e a pesquisa orientada por hipóteses”, *Liinc em Revista*, vol. 15, no. 1, 2019, doi: https://doi.org/10.18617/liinc.v15i1.4688. Acessado 20 jan. 2020.

Siegel, E. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die.Wiley. 2013.

Van Poucke, S. et al. “Are randomized controlled trials the (g)old standard? From clinical intelligence to prescriptive analytics”. *J Med Internet Res*, v.18, n.7, e185, 2016. DOI: https://www.jmir.org/2016/7/e185/. Acesso em: 18 nov. 2020.

Victorino, M. C. et al. Uma proposta de ecossistema de big data para a análise de dados abertos governamentais conectados. *Informação & Sociedade: Estudos*, v. 27, n. 1, 2017. DOI: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/29299. Acesso em: 18 nov. 2020.

Wu, X. et al. “Data Mining with big data”. *IEEE Transactions on Knowledge and Data Engineering*, vol.26, no. 1, 2013, doi: 10.1109/TKDE.2013.109. Acessado em: 30 dez. 2019.

Recived: 10/09/2020 Acepted: 03/12/2020