DOSSIÊ

Big Data e ciência de dados: uma revisão crítica de questões para a pesquisa educacional[1]

Big Data and data science: A critical review of issues for educational research

Daniel, Ben Kei; Maia, Mirtes Dâmares Santos de Almeida; Silva, Danilo Garcia da

Ben Kei Daniel
ben.daniel@otago.ac.nz
Universidade de Saskatchewan (Canadá), Canadá
Universidade de Otago, Nueva Zelanda
Mirtes Dâmares Santos de Almeida Maia
mirtes.almeidam@gmail.com
Universidade Federal de Mato Grosso, Brasil
Danilo Garcia da Silva
danilogsilvas@gmail.com
Univ. Federal de Mato Grosso, Brasil

Percursos
Universidade do Estado de Santa Catarina, Brasil
ISSN-e: 1984-7246
Periodicidade: Cuatrimestral
vol. 21, núm. 45, 2020
revistapercursos.faed@udesc.br

Recepção: 03 Março 2020
Aprovação: 04 Junho 2020

URL: http://portal.amelica.org/ameli/journal/815/8154491005/

DOI: https://doi.org/10.5965/1984724621452020080

Resumo: Big Data refere-se a grandes e díspares volumes de dados gerados por pessoas, aplicativos e máquinas. Vem ganhando atenção crescente em vários domínios, incluindo a educação. Mas quais são os desafios de se empreender a pesquisa em Big Data na área da educação? Este artigo identifica uma ampla gama de questões críticas que os pesquisadores precisam considerar ao trabalhar com Big Data voltado para educação. As questões identificadas incluem diversidade na concepção e significado de Big Data na educação, disparidade ontológica e epistemológica, desafios técnicos, ética e privacidade, barreira digital e dividendo digital, falta de habilidades e oportunidades de desenvolvimento acadêmico que preparem pesquisadores educacionais a alavancar as oportunidades proporcionadas pelo Big Data. O objetivo deste artigo é aumentar a conscientização sobre essas questões e iniciar um diálogo. O artigo foi inspirado em parte por ideias extraídas da bibliografia, mas, principalmente, informado pela experiência em pesquisar Big Data no âmbito da educação.

Palavras-chave: Big Data, Ciência de Dados, Educação.

Abstract: Big Data refers to large and disparate volumes of data generated by people, applications and machines. It is gaining increasing attention from a variety of domains, including education. What are the challenges of engaging with Big Data research in education? This paper identifies a wide range of critical issues that researchers need to consider when working with Big Data in education. The issues identified include diversity in the conception and meaning of Big Data in education, ontological, epistemological disparity, technical challenges, ethics and privacy, digital divide and digital dividend, lack of expertise and academic development opportunities to prepare educational researchers to leverage opportunities afforded by Big Data. The goal of this paper is to raise awareness on these issues and initiate a dialogue. The paper was inspired partly by insights drawn from the literature but mostly informed by experience researching into Big Data in education.

Keywords: Big Data, Data Science, Education.

Introdução[3]

O *Big Data* descreve um fenômeno que envolve o crescimento complexo e dinâmico de dados. Pesquisadores conceituam *Big Data* em dimensões estruturais e funcionais. A dimensão estrutural do *Big Data* abrange elementos de volume, velocidade, veracidade, variedade, verificação e valor (MANYIKA *et al.*, 2011; POULOVASSILIS, 2016). A diversidade estrutural e a complexidade do *Big Data* são atribuídas ao surgimento de novas formas de dados geradas por redes de sensores, aplicativos de mídia social e outros dispositivos móveis e ubíquos (MANYIKA *et al.*, 2011; SNIJDERS, MATZAT e REIPS, 2012; WARD e BARKER, 2013). Além disso, a dimensão funcional descreve o uso de tecnologias inovadoras para capturar, armazenar, distribuir, gerenciar e analisar conjuntos de dados amplos e heterogêneos (DEDE, HO e MITROS, 2016; LAZER, KENNEDY, KING e VESPIGNANI, 2014).

Como um novo paradigma de pesquisa, o *Big Data* voltado para educação estimula novas formas de enquadrar questões de pesquisa, projetar estudos, analisar e visualizar dados (DANIEL, 2015; DEDE *et al.*, 2016). Com a disponibilidade de grandes quantidades de dados na educação, os pesquisadores podem investigar subgrupos em uma população (um grupo específico de pessoas) sem necessariamente depender de métodos probabilísticos sofisticados (MAYER-SCHÖNBERGER e CUKIER, 2013). Além disso, as ferramentas de *Big Data* habilitam os pesquisadores para coletar grandes quantidades de dados de pesquisa a um custo relativamente baixo (MAYER-SCHÖNBERGER, 2015).

O *Big Data* proporciona aos pesquisadores educacionais um conjunto abrangente de ferramentas para manipular e visualizar dados sobre aprendizagem e ensino (BAKER e SIEMENS, 2013; BHAT e AHMED, 2016).

Greer e Mark (2016) propõem o uso de técnicas de visualização para identificar padrões úteis em dados educacionais, que podem não ser óbvios para professores que trabalham com abordagens estatísticas convencionais. A pesquisa também demonstrou que os painéis de visualização podem ajudar professores com conhecimento matemático limitado a navegar e facilmente interpretar os dados dos alunos (BUECKLE, GINDA, RANGA SURI e BÖRNER, 2017; ONG, 2015)[4].

A análise de um grande conjunto de dados educacionais pode informar o desenvolvimento de modelos preditivos para a identificação de oportunidades e para a abordagem de desafios de instituições educacionais (DANIEL e BUTSON, 2013). Também se argumenta que as noções obtidas a partir de modelos preditivos podem ser usadas para explorar as trajetórias de aprendizagem dos alunos de modo a facilitar o desenho de ambientes de aprendizagem adaptáveis e personalizados (MCKENNEY e MOR, 2015).

Embora o *Big Data* utilizado no campo da educação seja um fenômeno novo, com a disponibilidade de grande quantidade de dados educacionais armazenados em bancos de dados institucionais (por exemplo, dados obtidos de mídias sociais e sistemas de gerenciamento de aprendizagem) é provável que a pesquisa educacional se torne um campo com uso intensivo de dados, utilizando métodos e técnicas da Ciência de Dados. A Ciência de Dados se preocupa, principalmente, com o desenvolvimento e o uso de ferramentas, bem como de processos, para extrair e discernir conhecimentos valiosos de dados complexos (LEEK, 2013; WALLER e FAWCETT, 2013).

A Ciência de Dados pode fornecer aos pesquisadores educacionais a estrutura e os princípios necessários para enfrentar dados educacionais complexos. Oferece um conjunto de princípios fundamentais que apoiam a extração de informações e conhecimentos dos dados (PROVOST e FAWCETT, 2013, p. 52). A aplicação dos princípios e técnicas da Ciência de Dados na educação pode gerar benefícios de alta qualidade (KLAŠNJA-MILIĆEVIĆ, IVANOVIĆ e BUDIMAC, 2017).

Embora a bibliografia sobre *Big Data* na educação ofereça aos pesquisadores educacionais inúmeras oportunidades, várias questões precisam ser abordadas. Este artigo examina a promessa emergente do *Big Data* na educação e identifica uma ampla gama de questões que, provavelmente, afetarão a utilização futura do *Big Data* no campo educacional.

Notas do Autor

O que já se sabe sobre esse tema?

• Estão crescendo os potenciais de *Big Data* na transformação da educação;

• Pesquisadores em Tecnologia Educacional e Ciências da Computação, em particular, produziram um volume razoável de bibliografia sobre a promessa do Big Data e inteligência analítica no ensino superior, influenciando o ensino, a aprendizagem e a pesquisa;

• Nos últimos 6 anos, vários projetos de pesquisa institucionais se concentraram no desenvolvimento de ferramentas, sistemas e estratégias para o desenvolvimento bem-sucedido da analítica da aprendizagem [em inglês, *Learning Analytics*];

• Também existem pesquisas sobre os desafios da implementação de sistemas de analítica de aprendizagem entre instituições.

O que este artigo acrescenta

• Identifica uma ampla gama de questões que os pesquisadores educacionais precisam considerar ao trabalhar com *Big Data* na pesquisa educacional;

• Apresenta o quadro dos discursos sobre o desenvolvimento do projeto de pesquisa educacional, com as ferramentas e abordagens teóricas e epistêmicas da Ciência de Dados na pesquisa educacional;

• Introduz a Ciência de Dados como a quarta tradição de metodologia de pesquisa em pesquisa educacional.

Implicações práticas e políticas

• Este artigo tem como objetivo conscientizar sobre as questões fundamentais que pesquisadores educacionais enfrentam ao buscarem alavancar, em sua totalidade, a promessa do *Big Data*;

• O artigo apela para uma reconceitualização imediata do valor e da relevância do Big Data na pesquisa educacional;

• O trabalho apresentado neste artigo fornecerá às instituições uma reflexão sobre a criação de programas de pesquisa educacional em Ciência de Dados que apoiem a implementação bem-sucedida de *Big Data* na educação.

Pesquisas relacionadas

A análise de dados estudantis só se tornou um fenômeno importante na educação na última década (LODGE e CORRIN, 2017). No entanto, o uso de dados para auxiliar no processo de aprendizagem dos alunos pode ser rastreado até a pesquisa em Sistemas de Tutoria Inteligente (STI) e Inteligência Artificial na Educação (IAED) (ver Figura 1). Hoje, o principal objetivo do uso de dados na educação é identificar estratégias para projetar melhores ambientes de aprendizagem (MOR, FERGUSON e WASSON, 2015).

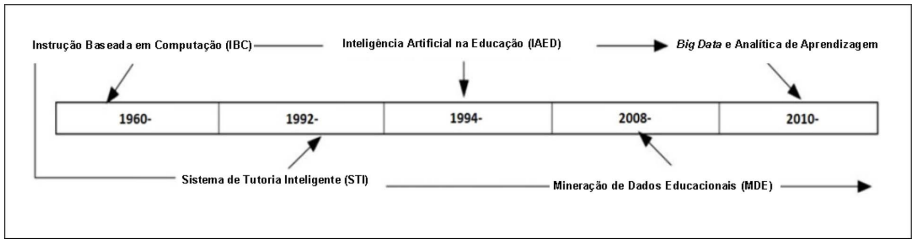

FIGURA 1
Uma breve visão geral dos campos de tecnologia educacional relacionados ao Big Data

Os STI, como mostrado na Figura 1, utilizam abordagens computacionais para rastrear as atividades de aprendizagem dos alunos e construir modelos de diagnóstico para eles (ANDERSON, BOYLE e REISER, 1985; BRUSILOVSKY, SCHWARZ e WEBER, 1996; NWANA, 1990). À medida que aumenta a necessidade de auxiliar alunos diversos e complexos em diferentes formas de ambientes de aprendizagem, novos métodos de coleta e análise de dados foram desenvolvidos. Pesquisadores em IAED e Mineração de Dados Educacionais (MDE), em particular, propuseram várias técnicas de modelagem de [dados de] alunos

(por exemplo, Redes Bayesianas[5], Modelos de Regressão, Modelos Cognitivos etc.) e mecanismos para a análise e visualização de dados (ver, por exemplo, SLATER, JOKSIMOVIĆ, KOVANOVIC, BAKER e GASEVIC, 2016).

A Analítica da Aprendizagem (AA) descreve um conjunto de várias ferramentas e abordagens para lidar com amplos e complexos dados estudantis, bem como os contextos em que a aprendizagem ocorre (GREER e MARK, 2016). Embora a MDE tenha precedido a AA, essas duas comunidades de pesquisa compartilham o objetivo comum de apoiar a educação. A Mineração de Dados Educacionais (MDE) ocupa-se, principalmente, com a descoberta automatizada de conhecimento e oferece uma coleção de ferramentas automatizadas de coleta e visualização de dados destinadas a apoiar a aprendizagem adaptativa (ver, por exemplo, BAKER, 2010; JONES e JO, 2004; LUAN, 2002; ROMERO, VENTURA E GARCÍA, 2008). A pesquisa em Analítica da Aprendizagem (AA), por outro lado, visa fornecer aos alunos e professores ferramentas acionáveis para apoiar a educação (MOR *et al.*, 2015; SIEMENS e BAKER, 2012).

Reconceitualização de *Big Data* na educação

O *Big Data* voltado para a educação é um fenômeno novo (PICCIANO, 2012), com a maioria dos discursos de pesquisa centrados no uso de dados para informar a qualidade da instrução e da pesquisa (EYNON, 2013). Por exemplo, Kalota (2015) sugeriu que a utilização das técnicas de *Big Data* na educação permite às instituições acadêmicas entender os desafios que os alunos enfrentam e identificar estratégias para encará-los. A disponibilidade de grande volume de dados educacionais, em particular, proporciona aos pesquisadores da educação as oportunidades de usar ferramentas e técnicas automatizadas para explorar fenômenos educacionais complexos em grande escala. Daniel (2015) propôs três usos do *Big Data* no âmbito educacional, a saber: apoio à aprendizagem, ensino e administração (ver Figura 2).

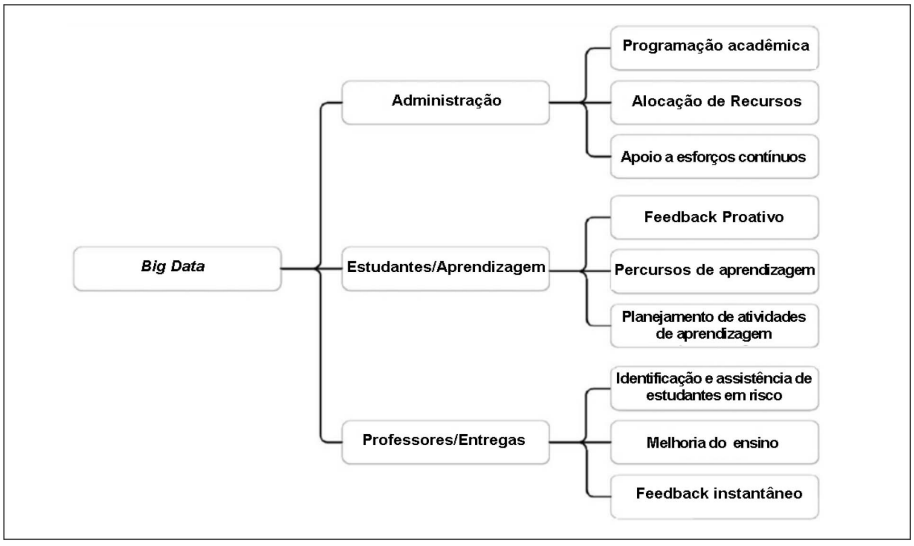

FIGURA 2
Cenários de casos de uso de Big Data na educação
Fonte: Daniel, 2015.

Várias fontes de *Big Da*ta na educação são observadas em outras partes da bibliografia. Por exemplo, Poulovassilis (2016) descreve várias fontes de *Big Data* no campo educacional, incluindo dados gerados e armazenados em ambientes virtuais de aprendizagem, dados de avaliação, registros pessoais de estudantes, modelos de aluno, dados de vídeo e dados fisiológicos (por exemplo, frequência cardíaca, pressão arterial etc.). A Analítica da Aprendizagem (AA) também permite que os professores identifiquem fatores de risco associados ao envolvimento dos alunos na aprendizagem e otimizem o *design* dos ambientes de aprendizagem (LODGE e CORRIN, 2017; MOR *et al.*, 2015). Professores podem usar o painel de AA para visualizar os percursos de aprendizagem dos alunos e identificar as áreas em que os alunos mais têm dificuldade, para que

possam elaborar melhores estratégias de intervenção. De forma similar, fornecer aos alunos o acesso ao painel personalizado [de dados] alimenta maior senso de autoconsciência e promove disposições de aprendizagem autodirecionadas (TAN, KOH, JONATHAN e YANG, 2017).

Embora *Big Data* ofereça várias oportunidades para a educação, *Big Data* voltado para educação e pesquisa educacional são duas áreas distintas de investigação, que exigem diferentes conjuntos de habilidades e conhecimentos (ver Tabela 1). Enquanto a pesquisa educacional está amplamente preocupada com a investigação de vários aspectos da educação, tais como a aprendizagem do aluno, métodos de ensino, aprendizagem aprimorada pela tecnologia, *Big Data* utilizado no campo da educação lida com a análise de dados amplos e complexos, usando técnicas de Ciência de Dados. Trabalhar com *Big Data* no âmbito da educação, portanto, requer um conhecimento adequado de Ciência de Dados e a habilidade de trabalhar com técnicas automatizadas (por exemplo, aprendizagem de máquina[6]) e sistemas de banco de dados de alto desempenho, a exemplo de Hadoop e MapReduce.

Pesquisa educacional	Pesquisa em *Big Data* em educação
• Contexto de dados conhecidos pelos pesquisadores.	• Contexto de dados pode ser desconhecido para pesquisadores.
• O pesquisador pode estar envolvido na coleta de dados.	• O pesquisador pode usar dados já coletados.
• Epistemologia e ontologia focadas.	• Epistemologia e ontologia emergentes.
• Protocolos éticos claros e responsabilidade.	• A responsabilidade ética pode ser desconhecida.
• Requer conhecimento em métodos de educação e pesquisa.	• Precisa de conhecimentos adicionais de Ciência de Dados.
• Amostra de dados pura e geralmente de tamanho pequeno/gerenciável (medido em Megabyte, Gigabyte).	• Estruturas de dados amplas e complexas (medidas em Terabyte, Petabytes, Exabyte).
• Não requer análise em tempo real.	• Pode empregar análise em tempo real.
• Os dados são armazenados dentro dos limites/possibilidades dos mecanismos de armazenamento disponíveis.	• Sistemas de Arquivos Altamente Distribuídos (SAAD) ou NoSQL.
• A análise é manual ou utiliza sistemas de *software* autônomos, tais como SPSS, NVivo, STATA.	• Utiliza Hadoop, sistemas de MapReduce, aplicativos de mineração da Web, redes de sensores, monitoramento de tráfego.

TABELA 1
Características da pesquisa educacional e da pesquisa de Big Data em educação

Embora, atualmente, não haja muitos cientistas de dados trabalhando na educação (BUCKINGHAM *et al.*, 2013; KOPRINSKA, STRETTON e YACEF, 2015), algumas universidades começaram a oferecer diplomas em Analítica da Aprendizagem (AA), um subconjunto de *Big Data*, voltado para educação (a exemplo de Teachers College Columbia University, University of Queensland e Northeastern University), abrindo futuras oportunidades para estender a Ciência de Dados ao domínio educacional.

É evidente que os pesquisadores educacionais geralmente trabalham com dados relativamente pequenos. Entretanto, quantidades massivas de dados educacionais agora podem ser facilmente coletadas, armazenadas, analisadas e compartilhadas entre indivíduos e instituições. A disponibilidade de uso de *Big Data* na educação, no entanto, requer uma compreensão das diferenças fundamentais entre pesquisa educacional e o uso de *Big Data* na área de educação (ver Tabela 1), assim como tratar de possíveis desafios (ver Figura 3) que podem ocorrer durante a transição dos pesquisadores da pesquisa educacional para a pesquisa com *Big Data* em educação.

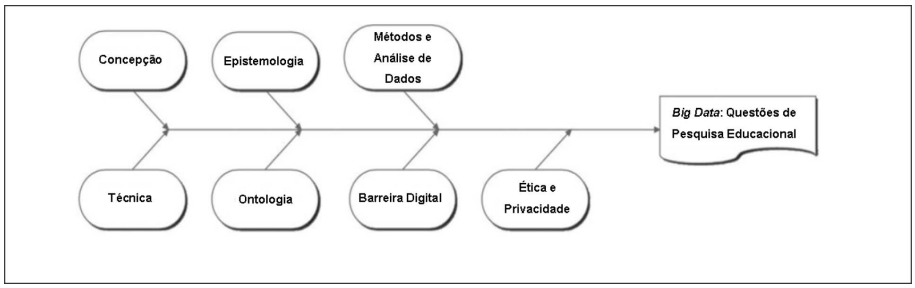

FIGURA 3
Big Data em educação e questões críticas para pesquisa educacional
Fonte: Daniel, 2015

Big Data e pesquisa educacional: questões de concepção

A rápida geração de dados por diferentes dispositivos e pessoas atingiu grandes proporções, excedendo a capacidade de *hardware* ou as habilidades humanas de processar e manipular (VAITSIS, HERVATIS e ZARY, 2016). Subsequentemente, há uma tendência geral na bibliografia de conceituar *Big Data* em paralelo à magnitude dos dados. Por sua vez, isso levou à crença de que os dados educacionais não são grandes o suficiente e, portanto, não podem ser considerados *Big Data*. No entanto, há pouco acordo entre pesquisadores de diversas áreas fora da educação sobre o que constitui o *Big Data* em termos de magnitude. Por exemplo, um Terabyte de dados pode ser qualificado como *Big Data*? Outros argumentaram que a caracterização do *Big Data* em termos de tamanho está relacionada ao domínio (BAKER, 2015).

Embora existam visões diferentes sobre o que constitui o *Big Data* na bibliografia, Daniel e Butson (2013) propõem uma estrutura teórica para descrever *Big Data* no ensino superior junto com a Analítica Institucional (AI) [em inglês, *Institutional Analytics*], a Analítica da Tecnologia da Informação (ATI) [em inglês, *Information Technology Analytics*], a Analítica Acadêmica (AAc) [em inglês, *Academic Analytics*] e a Analítica da Aprendizagem (AA) [em inglês, *Learning Analytics*].

A Analítica Institucional (AI) se preocupa com a análise de dados administrativos para melhorar a qualidade do processo de tomada de decisão. A Analítica da Tecnologia da Informação (ATI) refere-se à coleta e análise de dados associados ao uso, tanto administrativo, quanto por estudantes, de serviços de tecnologia (por exemplo, repositório de dados, padrões de dados, ferramentas e políticas).

A Analítica Acadêmica (AAc) se refere à análise de dados sobre atividades e desempenho de programas acadêmicos (medidos em termos de índices de conclusão e graduação, índices de aprovação e reprovação etc.). O resultado da AAc informa decisões estratégicas relacionadas a aspectos de administração, como alocação de recursos e retenção de alunos (CHARLTON, MAVRIKIS e KATSIFLI, 2013; SIEMENS, 2013).

A Analítica da Aprendizagem (AA) é a medição, coleta, análise e relatório de dados sobre os alunos e o contexto em que a aprendizagem ocorre (JONES, 2012; SIEMENS e LONG, 2011). Pesquisadores usam os resultados da AA para entender e otimizar o processo de aprendizagem.

É importante observar que a variedade de concepções do que constitui *Big Data* no âmbito da educação levanta questões de diferentes interpretações, as quais, provavelmente, impactarão na implementação de projetos de *Big Data* na educação.

Big Data e pesquisa educacional: questões técnicas

Trabalhar com sistemas de *Big Data* requer acesso a uma infraestrutura computacional de alta velocidade, capaz de lidar com uma quantidade massiva de dados que, em grande parte, pode incorrer em um custo significativo associado a captura, armazenamento, análise e visualização de dados (CHEN & ZHANG, 2014). Embora muitas instituições acadêmicas estejam atualmente coletando várias formas de dados, esses são mantidos em bancos de dados díspares, dificultando a análise. Além disso, a falta de interoperabilidade dos sistemas de dados institucionais torna trabalhosa a agregação de dados para análise de sistemas díspares (DANIEL, 2015). Além disso, a ausência de acordos de compartilhamento de dados e modelos de governança

de dados pode constituir um gargalo adicional para integração e comparação de dados entre instituições (MIYARES e CATALANO, 2016).

Preocupações em proteger a privacidade individual e institucional por meio de autenticação e segurança são outras questões importantes nos sistemas de *Big Data*. Por exemplo, sistemas de *Big Data* como o Hadoop, projetados para gerenciar dados públicos, têm apenas um nível de proteção de dados, dificultando sua implementação no domínio educacional. Construir uma camada adicional de controle e criptografia para proteger os dados na educação pode incorrer em recursos significativos.

Além disso, um dos valores do *Big Data* na educação é o uso extensivo da modelagem preditiva. No entanto, o excesso de confiança na modelagem preditiva pode ser limitante, pois os temas educacionais são muito complexos e difíceis de lidar em um único modelo. Por exemplo, a construção de modelos que possam identificar precisamente alunos sob o risco de fracassar em seu programa de estudos exige uma análise minuciosa dos possíveis gatilhos de problemas computacionalmente não tratáveis. Os problemas computacionais não tratáveis na educação incluem desafios socioeconômicos dos alunos (como histórico familiar, estado de saúde, recursos disponíveis e condições de vida em casa). Geralmente esses problemas estão além do ambiente de aprendizagem; assim sendo, podem ser difíceis de capturar e incorporar em um modelo preditivo. Ademais, a interpretação precisa de modelos preditivos requer conhecimento técnico de Ciência de Dados. No entanto, esse conhecimento pode não estar acessível a muitos pesquisadores educacionais.

Big Data e pesquisa educacional: questões ontológicas

A ontologia constitui uma visão particular da realidade. Os pesquisadores usam uma ontologia própria para situar suas compreensões, dentro de uma perspectiva teórica. Na pesquisa educacional, especialmente na pesquisa qualitativa, o envolvimento com o processo de coleta de dados é uma parte crítica da integridade da pesquisa localizada em um pensamento ontológico ou visão de mundo particular, porque os pesquisadores inferem o significado de um fenômeno com base no contexto em que os dados são coletados e analisados. No entanto, os pesquisadores que trabalham com *Big Data* raramente estão envolvidos na coleta de dados ou no desenho do estudo (DEDE *et al.*, 2016), levantando a questão de como os pesquisadores educacionais podem envolver e analisar significativamente os dados gerados para diferentes intenções e contextos.

A relação entre o pesquisador e o que está sendo pesquisado, referido como reflexividade, é um componente essencial do processo de pesquisa educacional. Uma orientação ontológica facilita a compreensão do fenômeno pesquisado e do próprio processo de pesquisa (WATT, 2007). A falta de envolvimento durante a coleta de dados na pesquisa em *Big Data* voltado para educação pode diminuir drasticamente o valor da reflexividade, comprometendo, possivelmente, o rigor dos resultados da pesquisa.

Como os pesquisadores que trabalham com *Big Data* podem fazer uso de dados já coletados, os elementos essenciais da pesquisa experimental, tais como randomização, contramedidas de ameaças a várias formas de validação e controle manipulativo são difíceis de alcançar (BOYD e CRAWFORD, 2012; CRAWFORD, GRAY e MILTNER, 2014).

Big data e pesquisa educacional: questões epistemológicas

O uso de uma epistemologia específica (por exemplo, positivismo x interpretativismo) informa a escolha de qualquer metodologia de pesquisa e, por sua vez, molda o desenho de um estudo. Trabalhar com *Big Data* em educação requer uma compreensão das teorias científicas universais para inferências indutivas (FRICKÉ, 2015). Isso implica abraçar novas formas de empirismo (KITCHIN, 2014), que transcendem tradições quantitativas e qualitativas.

As novas formas de empirismo são caracterizadas por projetos de pesquisa emergentes, moldados pelo ambiente tecnológico, por dados complexos e dinâmicos. Esse novo tipo de empirismo constitui a quarta tradição da metodologia de pesquisa (Ciência de Dados). A primeira tradição da metodologia de pesquisa baseia-se nos métodos científicos (quantitativos), caracterizados pela epistemologia positivista, enquanto que a segunda tradição consiste em práticas de pesquisa que estão teoricamente situadas no interpretativismo (métodos qualitativos). Métodos mistos com sua epistemologia abrangente do pragmatismo formam a

terceira tradição. As quatro tradições, com suas formas de dados respectivamente associadas, são mostradas na Figura 4.

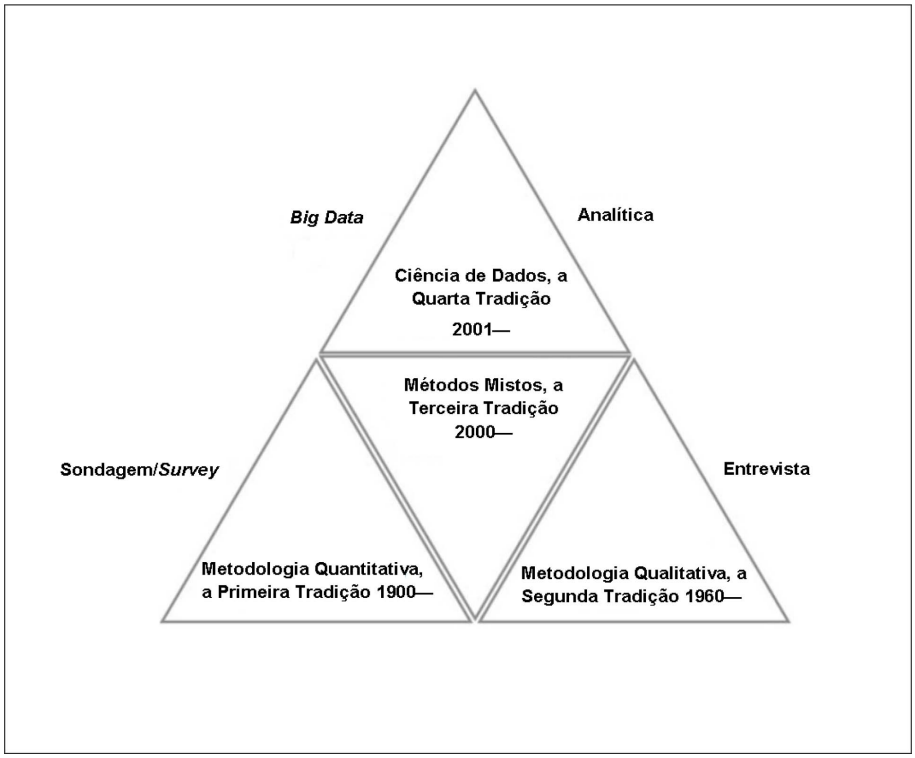

FIGURA 4
As quatro tradições da metodologia de pesquisa

A epistemologia associada à Ciência de Dados difere dos métodos convencionais (HARFORD, 2014) porque o processo de pesquisa em Ciência de Dados não depende de paradigmas pré-determinísticos e hegemônicos, mas requer uma negociação contínua de significados, restringidos pelo ambiente em que a pesquisa é realizada. Diferentemente das três tradições metodológicas, trabalhar efetivamente com a Ciência de Dados requer habilidade por parte dos pesquisadores para lidar com dados complexos e heterogêneos (FAN, HAN e LIU, 2014). A quarta tradição de metodologia de pesquisa proposta no artigo substancia as visões de Tansley e Tolle (2009) sobre o quarto paradigma de pesquisa, que apela por novas abordagens e procedimentos para a realização de pesquisas científicas à luz de novas formas de dados gerados publicamente, que podem ser reajustados e supervisionados segundo certas restrições regulatórias (TOLLE, TANSLEY e HEY, 2011).

Big Data e pesquisa educacional: métodos e análise de dados

Romero e Ventura (2010) observaram que os pesquisadores do ensino superior trabalharam com quantidades relativamente pequenas de dados, com poder interpretativo, latência e validade limitados. *Big Data* voltado para educação oferece aos pesquisadores abordagens robustas para descobrir padrões populacionais sutis que dificilmente serão alcançados com dados em pequena escala (FAN *et al.*, 2014). No entanto, o resultado da pesquisa em *Big Data* é, em grande medida, limitado a modelos correlacionais e análises preditivas, deixando a causalidade dos resultados da pesquisa educacional desejável, mas até certo ponto inatingível.

Muitos métodos na pesquisa de *Big Data* preocupam-se em fazer as perguntas "o quê" mais que "por quê". No entanto, o resultado da pesquisa educacional é frequentemente necessário para tratar de problemas específicos de aprendizagem. Portanto, é necessário identificar as causas dos problemas em vez de

simplesmente descrevê-los, a fim de desenvolver melhores estratégias para alcançar resultados educacionais desejáveis.

Pesquisas contestam o uso de correlação *versus* causalidade (ver, por exemplo, BOLLIER, 2010; MAYER-SCHÖNBERGER, 2015). Mayer-Schönberger (2015) argumenta que a análise correlacional da pesquisa em *Big Data* pode frequentemente produzir conexões úteis para o desenvolvimento de intervenções, mesmo na ausência de causalidade. No entanto, tratar equivocadamente a correlação como causalidade pode resultar na escolha de intervenções ineficazes, mesmo que esse resultado seja baseado na análise de um grande conjunto de dados.

Big Data e pesquisa educacional: questões de barreira digital e dividendo digital

Big Data é uma fonte de competição para algumas instituições, porque os pesquisadores podem extrair indicações úteis dos dados e usá-los para melhorar a produtividade (GURSTEIN, 2011; MCGUIRE, MANYIKA e CHUI, 2012). Junqué de Fortuny, Martens e Provost (2013) afirmaram que instituições com maiores recursos de dados poderiam tirar proveito do *Big Data* para obter vantagem competitiva sobre outras instituições (dividendo digital).

A pesquisa em *Big Data* em educação requer habilidades especializadas que faltam em muitos pesquisadores educacionais. O uso da visualização de dados, em particular, requer conhecimento de estatística e visualização de informações, limitando a acessibilidade a muitos pesquisadores educacionais.

Além disso, muitos pesquisadores educacionais não estão familiarizados com as tecnologias associadas à pesquisa em *Big Data* (por exemplo, Hadoop, NoSQL e MapReduce). Também faltam oportunidades de desenvolvimento profissional para pesquisadores educacionais interessados em trabalhar com *Big Data* (barreira digital). Trabalhar com *Big Data* requer o envolvimento de um cientista de dados, conhecedor das questões corretas de pesquisa educacional. No entanto, há um número limitado de cientistas de dados que estão familiarizados ou interessados em trabalhar no domínio da educação.

Big Data e pesquisa educacional: questões éticas e de privacidade

Big Data voltado para educação apresenta ameaças potenciais à segurança e proteção dos alunos. O uso de AA, onde os alunos estão sendo rastreados e seu desempenho sinalizado, pode levar a resultados indesejados. Por exemplo, o uso de dados de estudantes para tomar uma decisão pode negar o acesso estudantil a programas futuros. Além disso, alguns estudantes podem se opor ao uso de seus dados, mesmo que seja obtido um consentimento adequado. Como Prinsloo, Archer, Barnes, Chetty e Van Zyl (2015) observaram, a coleta de dados sem qualquer objetivo claro ou sem a obtenção de um consentimento apropriado dos alunos levanta questões de ética, privacidade e propriedade dos dados. As preocupações com ética e privacidade em *Big Data* voltado para educação são complexas, exigindo um entendimento das relações de poder entre estudantes e instituições (SLADE e PRINSLOO, 2013). Para abordar questões de ética e privacidade, as instituições precisam considerar a criação de modelos de governança de dados e políticas de proteção de dados, bem como o contexto em que os dados podem ser usados (DIARIES *et al.*, 2014; DYCKHOFF, ZIELKE, BÜLTMANN, CHATTI, e SCHROEDER, 2012; METCALF e CRAWFORD, 2016).

No entanto, os padrões atuais para obter o consentimento dos participantes na pesquisa em *Big Data* são desafiadores, pois a maioria dos dados já existe em bancos de dados institucionais. Outro dilema ético associado ao uso do *Big Data* para pesquisa é manter a integridade da pesquisa ao usar dados acessíveis ao público, pois os possíveis geradores desses dados podem não estar dispostos a consentir com seu uso, ou tais indivíduos não estarem mais acessíveis para pesquisadores.

O direito à propriedade e ao acesso aos dados são questões adicionais a serem consideradas. Por exemplo, um aluno deve ter acesso aos mesmos dados que um professor? Os educadores devem poder ver dados analíticos informatizados de outros cursos? Ainda, seria apropriado que instituições acadêmicas disponibilizassem dados estudantis para terceiros, incluindo empregadores? Há também questões de obrigações morais institucionais associadas ao uso de dados do aluno para modelagem preditiva. Por exemplo,

se ficar evidente que um aluno em particular está enfrentando dificuldades, uma instituição será moralmente obrigada a ajudá-lo, mesmo que a causa da dificuldade possa ter um fundo social e financeiro complexo?

Como os pesquisadores educacionais exploram a análise dos dados armazenados na computação baseada em nuvem, as questões de privacidade e segurança podem vir a se tornar ainda mais complexas, carecendo do estabelecimento de ética global e obrigações morais para o uso de dados educacionais.

Questões de confiança precisam ser tratadas ao compartilhar dados de pesquisa entre instituições. Um número crescente de revistas acadêmicas (por exemplo, *British Journal of Educational Technology*) incentiva os pesquisadores a compartilhar dados com outros pesquisadores. No entanto, o compartilhamento de dados sem diretrizes adequadas pode provocar direitos de propriedade intelectual e preocupações com o consentimento esclarecido, já que aqueles que consentiram com o uso de seus dados podem não ter permitido o compartilhamento com terceiros.

Conclusões e pesquisas futuras

Big Data voltado para educação levou pesquisadores e desenvolvedores a ver possibilidades de como introduzir diferentes tecnologias de processamento e geração de informações que auxiliem na aprendizagem do aluno. Apesar da crescente pesquisa sobre *Big Data* na área da educação e seu evidente valor para a aprendizagem, muitas instituições acadêmicas são lentas na implementação de projetos de *Big Data* (MACFADYEN, 2017). Eynon (2013) nos alertou sobre o uso excessivo de *Big Data* na educação como uma forma de "reparo técnico" em vez de uma forma de empoderar os pesquisadores para realizar uma melhor pesquisa educacional.

Como o uso de *Big Data* na educação começa a se tornar um paradigma de pesquisa dominante, é preciso tratar de questões de conceitualização antes que seja amplamente adotado. É necessária uma nova concepção de *Big Data* no contexto da pesquisa educacional, que leve em consideração a complexidade dos ambientes educacionais e a natureza dos dados que estão sendo coletados. *Big Data* cria oportunidades únicas para pesquisa. No entanto, essas oportunidades não são imediatamente acessíveis a todos os pesquisadores educacionais, a menos que sejam oferecidas oportunidades de desenvolvimento profissional (DANIEL, 2017). Além disso, o estabelecimento de programas de pesquisa educacional usando *Big Data* exigirá enfrentar questões de epistemologia, ontologia, metodologia e desigualdade na alavancagem dos resultados de *Big Data* na educação.

É importante abordar questões de infraestrutura, ferramentas e capacidade humana necessárias para a coleta, limpeza, análise e distribuição eficientes de grandes conjuntos de dados. Além disso, preocupações críticas quanto a privacidade, ética, acesso e governança continuam sendo majoritárias (GASEVIC, DAWSON e JOVANOVIC, 2016). Como as instituições ampliam a necessidade de compartilhar dados educacionais, será imperativo o desenvolvimento de padrões nacionais e internacionais para tratar de questões de segurança e interoperabilidade de dados, bem como privacidade e acesso. Os educadores podem se engajar na coleta de vários tipos de dados para melhorar as aulas, mais que para a pesquisa (HO, 2017); reaproveitar esses dados para pesquisa pode não ser ético.

Pesquisas futuras precisam explorar essas questões e identificar estratégias para auxiliar pesquisadores educacionais. Além disso, a implementação bem sucedida de *Big Data* na educação depende da habilidade dos pesquisadores educacionais em trabalhar com princípios e abordagens da Ciência de Dados, impulsionados por ideias da quarta tradição metodológica de pesquisa.

Referências

ANDERSON, J. R.; BOYLE, C. F.; REISER, B. J. Intelligent tutoring systems. **Science**, Washington, v. 228, n. 4698, p. 456-462, 1985.

BAKER, R. S. Big data and education. 2. ed. New York: Teachers College, Columbia University, 2017. Disponível em: <http://www.columbia.edu/~rsb2162/bigdataeducation.html>; Acesso em: 02 nov. 2017.

BAKER, R. S. J. D. Data mining for education. In: McGRAW, B. *et al.* (Eds.) **International Encyclopedia of Education**, Oxford, Elsevier, 2010, v. 7, p. 112-118.

BAKER, R. S. J. D.; SIEMENS, G. **Educational data mining and learning analytics**. Cambridge: Cambridge University, 2013.

BHAT, A. Z.; AHMED, I. Big data for institutional planning, decision support, and academic excellence. In: **Third MEC International Conference on Big Data and Smart City (ICBDSC), March 2016**. Muscat, Oman: IEEE, 2016, p. 1-5.

BOLLIER, D. The promise and peril of big data. Washington, DC: Aspen Institute, Communications and Society Program, 2010, p. 56. Disponível em: <http://23.66.85.199/collateral/analystreports/10334-ar-promise-peril-of-big-data.pdf>;. Acesso em: 02 nov. 2010.

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

BRUSILOVSKY, P.; SCHWARZ, E.; WEBER, G. ELM-ART: an intelligent tutoring system on World Wide Web. In: **International conference on intelligent tutoring systems**. Berlin/ Heidelberg: Springer, 1996, p. 261-269.

BUCKINGHAM SHUM, *et al.* Educational data scientists: a scarce breed. In: **Proceedings of the Third International Conference on Learning Analytics and Knowledge**, 2013, p. 278-281.

BUECKLE, A. *et al*. Empowering instructors in learning management systems: interactive heat map analytics dashboard. Disponível em: <http://cns.slis.indiana.edu/docs/publications/2016%20ginda%20LAK17%20submit.pdf>;. Acesso em: 02 nov. 2017.

CHARLTON, P.; MAVRIKIS, M.; KATSIFLI, D. The potential of learning analytics and big data. Ariadne, n. 71, p. 1-15, 2013. Disponível em: <http://www.ariadne.ac.uk/issue71/charlton-et-al>;. Acesso em: 04 nov. 2017.

CHEN, C. P.; ZHANG, C. Y. Data-intensive applications, challenges, techniques, and technologies: a survey on Big Data. **Information Sciences**, n. 275, p. 314-347, 2014.

CRAWFORD, K.; GRAY, M. L.; MILTNER, K. Big Data / Critiquing Big Data: politics, ethics, epistemology, special section introduction. **International Journal of Communication**, v. 8, p. 10, 2014.

DANIEL, B. K. Big Data and analytics in higher education: opportunities and challenges. **British Journal of Educational Technology**, n. 46, p. 904-920, 2015.

DANIEL, B. K. **Big data and learning analytics in higher education**: current theory and practice. New York: Springer, 2017.

DANIEL, B. K.; BUTSON, R. Technology enhanced analytics (TEA) in higher education. In: KOMMERS, T. *et al.***.roceedings of the International Conference on Educational Technologies (ICEduTech)**. Lisbon, Portugal: IADIS Press, 2013, p. 89-96.

DEDE, C.; HO, A.; MITROS, P. Big Data analysis in higher education: promises and pitfalls. EDUCAUSE Review, v. 51, n. 5, p. 8-9, 2016. Disponível em: <http://er.educause.edu/articles/2016/8/big-data-analysis-in-higher-education-promisesand-pitfalls>;. Acesso em: 01 set. 2016.

DIARIES, J. P. *et al.* Privacy, anonymity, and big data in the social sciences. C**ommunications of the ACM**, n. 57, p. 56-63, 2014.

DYCKHOFF, A. L. *et al.* Design and implementation of a learning analytics toolkit for teachers. **Educational Technology & Society,** n. 15, p. 58-76, 2012.

EYNON, R. The rise of Big Data: what does it mean for education, technology, and media research? **Learning, Media, and Technology**, n. 38, p. 237-240, 2013.

FAN, J.; HAN, F.; LIU, H. Challenges of big data analysis. **National Science Review**, n. 1, p. 293-314, 2014.

FRICKÉ, M. Big data and its epistemology. **Journal of the Association for Information Science and Technology**, n. 66, p. 651-661, 2015.

GASEVIC, D.; DAWSON, S.; JOVANOVIC, J. Ethics and privacy as enablers of learning analytics. **Journal of Learning Analytics**, n. 3, p. 1-4, 2016.

GREER, J.; MARK, M. Evaluation methods for intelligent tutoring systems revisited. **International Journal of Artificial Intelligence in Education**, n. 26, p. 387-392, 2016.

GURSTEIN, M. B. Open data: empowering the empowered or effective data use for everyone?. First Monday, v. 16, n. 2, 2011. Disponível em: <http://journals.uic.edu/ojs/index.php/fm/article/view/3316/2764>;. Acesso em: 15 set. 2016.

HARFORD, T. Big data: a big mistake?. **Significance**, n. 11, p. 14-19, 2014.

HO, A. **Advancing educational research and student privacy in the "Big Data" era**. Washington, DC: National Academy of Education, 2017.

JONES, S. Technology Review: the possibilities of learning analytics to improve learnercentred decision-making. **Community College Enterprise**, n. 18, p. 89-92, 2012.

JONES, V.; JO, J. H. Ubiquitous learning environment: an adaptive teaching system using ubiquitous technology. In: **Beyond the Comfort Zone: Proceedings of the 21st ASCILITE Conference - Perth, Western Australia, 5-8 December,** 2004, v. 468, p. 474.

JUNQUÉ DE FORTUNY, E.; MARTENS, D.; PROVOST, F. Predictive modeling with big data: is bigger better?. **Big Data**, n. 1, p. 215-226, 2013.

KALOTA, F. Applications of Big Data in education. **International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering**, n. 9, p. 1602- 1607, 2015.

KITCHIN, R. Big Data, new epistemologies, and paradigm shifts. **Big Data & Society**, v. 1, n. 1, p. 1-12, 2014.

KLAŠNJA-MILIĆEVIĆ, A; IVANOVIĆ, M.; BUDIMAC, Z. Data science in education: Big Data and learning analytics. **Computer Applications in Engineering Education**, n. 25, p. 10661078, 2017. Disponível em: <https://doi.org/10.1002/cae.21844>;. Acesso em: 02 nov. 2017.

KOPRINSKA, I.; STRETTON, J.; YACEF, K. Students at risk: detection and remediation. In: **EDM 2015 - The 8th International Educational Data Mining- 22-26 June 2015**. Madrid, Spain, International Educational Data Mining Society, UNED - the National University for Distance Education, 2015, p. 512-515.

LAZER, D.; KENNEDY, R.; KING, G.; VESPIGNANI, A. The parable of Google Flu: traps in big data analysis. **Science**, n. 343, p. 1203-1205, 2014.

LEEK, J. Six types of analyses every data science should know. **Data scientist Insights blog**, 2013. Disponível em: <http://datascientistinsights.com/2013/01/29/six-types-ofanalyses-everydata-scientistshould-know/>;. Acesso em: 02 nov. 2017.

LODGE, J. M.; CORRIN, L. What data and analytics can and do say about effective learning. **NPJ Science of Learning**, v. 2, n. 5, p. 1-2, 2017. Disponível em: <http://www.nature.com/articles/s41539-017-0006-5>;. Acesso em: 02 nov. 2017.

LUAN, J. Data mining and its applications in higher education. In: SERBAN, A.; LUAN, J. (Eds.). **Knowledge management**: building a competitive advantage in higher education. San Francisco, CA: Josey-Bass, 2002, p. 17-36.

MACFADYEN, L. P. Overcoming barriers to educational analytics: how systems thinking and pragmatismo can help. **Educational Technology**, n. 57, p. 31-39, 2017.

MANYIKA, J., M. *et al.***Big Data:** the next frontier for innovation, competition, and productivity. [San Francisco, EUA]: McKinsey Global Institute, 2011. Disponível em: <http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation >. Acesso em: 29 mar. 2016.

MAYER-SCHÖNBERGER, V. Big Data for cardiology: novel discovery?. **European Heart Journal**, n. 37, p. 996-1001, 2015.

MAYER-SCHÖNBERGER, V.; CUKIER, K. **Big Data**: a revolution that will transform how we live, work, and think. Boston, MA: Houghton Mifflin Harcourt, 2013.

MCGUIRE, T.; MANYIKA, J.; CHUI, M. Why Big Data is the new competitive advantage. **Ivey Business Journal**, n. 76, p. 1-4, 2012.

MCKENNEY, S.; MOR, Y. Supporting teachers in data-informed educational design. **British Journal of Educational Technology**, n. 46, p. 265-279, 2015.

METCALF, J.; CRAWFORD, K. Where are human subjects in Big Data research?. The Emerging Ethics Divide, **Big Data and Society**, p. 1-14, 2016.

MIYARES, J.; CATALANO, D. Institutional analytics is hard work: a five-year journey. **EDUCAUSE review**, v. 51, n. 5, p. 8-9, 2016. Disponível em: <http://er.educause.edu/~/media/files/articles/2016/8/erm1656.pdf>;. Acesso em: 01 set. 2016.

MOR, Y.; FERGUSON, R.; WASSON, B. Editorial: Learning design, teacher inquiry into student learning and learning analytics: a call for action. **British Journal of Educational Technology**, n. 46, p. 221-229, 2015.

NWANA, H. S. Intelligent tutoring systems: an overview. **Artificial Intelligence Review**, n. 4, p. 251-277, 1990.

ONG, V. K. Big Data and its research implications for higher education: cases from UK higher education institutions. In: **4th International Congress on Advanced Applied Informatics (IIAI-AAI)**, 2015, p. 487-491.

PICCIANO, A. G. The evolution of Big Data and learning analytics in American higher education. **Journal of Asynchronous Learning Networks**, n. 16, p. 9-20, 2012.

POULOVASSILIS, A. **Big Data and education** [Technical Report BBKCS-16-01]. Birkbeck, 2016. Disponível em: <http://www.dcs.bbk.ac.uk/research/techreps/2016/bbkcs-16-01.pdf>;. Acesso em: 02 nov. 2017.

PRINSLOO, P. *et al.* Big(ger) data as better data in open distance learning. **The International Review of Research in Open and Distributed Learning**, n. 16, p. 284-306, 2015.

PROVOST, F.; FAWCETT, T. Data science and its relationship to Big Data and data-driven decision making. **Big Data**, n. 1, p. 51-59, 2013.

ROMERO, C.; VENTURA, S. Educational data mining: a review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews**, n. 40, 2010, p. 601-618.

ROMERO, C.; VENTURA, S.; GARCÍA, E. Data mining in course management systems: Moodle case study and tutorial. **Computers & Education**, n. 51, p. 368-384, 2008.

SIEMENS, G. Learning analytics: the emergence of a discipline. **American Behavioral Scientist**, n. 57, p. 1380-1400, 2013.

SIEMENS, G.; BAKER, R. S. Learning analytics and educational data mining: towards communication and collaboration. In: **Proceedings of the Second International Conference on Learning Analytics and Knowledge**, Vancouver, Canada, 2012, p. 252-254.

SIEMENS, G.; LONG, P. Penetrating the Fog: analytics in learning and education. **EDUCAUSE Review**, n. 46, p. 30-32, 2011.

SLADE, S.; PRINSLOO, P. Learning analytics: ethical issues and dilemmas. **American Behavioral Scientist**, n. 57, p. 1510-1529, 2013.

SLATER, S. *et al.* Tools for educational data mining a review. **Journal of Educational and Behavioral Statistics**, v. 42, n. 1, p. 85-106, 2016.

SNIJDERS, C.; MATZAT, U.; REIPS, U. D. "Big Data": big gaps of knowledge in the field of internet science. **International Journal of Internet Science**, n. 7, p. 1-5, 2012.

TAN, J. P. L. *et al.* Learner dashboards a double-edged sword? Students'sense-making of a collaborative critical reading and learning analytics environment for fostering 21st century literacies. **Journal of Learning Analytics**, n. 4, p. 117-140, 2017.

TANSLEY, S.; TOLLE, K. M., HEY, T. (Ed.). **The Fourth Paradigm:** Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research, 2009.

TOLLE, K. M.; TANSLEY, D. S. W.; HEY, A. J. The Fourth Paradigm: Data-Intensive Scientific Discovery [point of view]. **Proceedings of the IEEE**, v. 99, n. 8, p. p. 1334-1337, 2011.

VAITSIS, C.; HERVATIS, V.; ZARY, N. Introduction to Big Data in education and its contribution to the quality improvement processes. In: VENTURA SOTO, S., LUNA, J,. CANO, A. (Ed.). **Big Data on Real-World Applications**. London: InTechOpen, 2016. Disponível

em: <https://www.intechopen.com/books/big-data-on-real-world-applications/introduction-tobig-data-in-education-and-its-contribution-to-the-qualityimprovement-processes>;. Acesso em: 02 nov. 2017.

WALLER, M. A.; FAWCETT, S. E. Data science, predictive analytics, and Big Data: a revolution that will transform supply chain design and management. **Journal of Business Logistics**, n. 34, p. 77-84, 2013.

WARD, J. S.; BARKER, A. Undefined by data: a survey of Big Data definitions [2013]. Disponível em: https://arxiv.org/pdf/1309.5821.pdf. Acesso em: 02 nov. 2017.

WATT, D. On becoming a qualitative researcher: the value of reflexivity. **The Qualitative Report**, v. 12, n. 1, p. 82-101, 2007. Disponível em: <http://nsuworks.nova.edu/tqr/vol12/iss1/5>;. Acesso em: 02 nov. 2017.

Notas

1 [Nota da editora] Tradução para o português do artigo "Big Data and data science: A critical review of issues for educational research", originalmente publicado em inglês no British Journal of Educational Technology (v.50, n.1, p. 101-113, 2019). A tradução e a publicação na revista PerCursos foram autorizadas pelo autor e por Ross Fulton, Publications Manager da British Journal of Educational Tecnology.

2 [Nota dos tradutores] Palavras-chave, assim como keywords, incluídas na versão traduzida

3 Declarações sobre dados abertos, ética e conflito de interesses: As ideias apresentadas neste artigo são desenvolvidas a partir da revisão da bibliografia publicada. Não apresentam riscos para indivíduos ou instituições. Nenhum potencial conflito de interesses foi relatado pelo autor.

4 [Nota da editora] A visualização de dados envolve sua representação imagética, geralmente em quadro, tabela, gráfico, mapa ou diagrama. Já os painéis de visualização de dados são ferramentas automatizadas que permitem rastrear, reunir e apresentar indicadores-chave a respeito daquilo que se pretende gerenciar ou estudar.

5 [Nota da editora] Redes construídas sobre variáveis cujo grau de probabilidade busca ser aferido ou controlado com base no Teorema de Bayes.

6 [Nota da editora] Procedimento que submete equipamentos computadorizados a operações de reconhecimento de padrões de dados, de modo que "aprendam" a reconhecê-los, selecionando e apresentando os resultados desejados.