Teoria e Prática em Administração, , Jul.-Dez. 2020
De Moura Jr.
Seção: Dossiê
DOI: http://dx.doi.org/10.21714/2238-104X2020v10i2-54385

Ciência dos Dados e Implicações Teóricas e Práticas em Administração

Data Science and Theoretical and Managerial Implications in Administration

Pedro Jácome de Moura Jr
Universidade Federal da Paraíba – UFPB – Brasil
pjacome@sti.ufpb.br
ORCID: 0000-0001-6548-4614

Submetido em 01/08/2020; **Aprovado** em 01/08/2020.

Resumo
O termo específico "ciência de dados" tem sido amplamente citado, tanto no contexto acadêmico quanto no setor produtivo, o que é resultado de sua alta aplicabilidade e do efeito *buzzword* criado em torno do termo. No entanto, uma "ciência dos dados" ainda carece de identidade distinta e de uma abordagem guiada por teoria, tanto para seu benefício quanto para que possa ser aplicada adequadamente e em conjunto com a Administração. Esta seção temática (dossiê) reflete uma iniciativa que fornece à emergente área uma oportunidade de expressão da sua aplicabilidade ao campo da Administração. Os cinco artigos do dossiê apresentam um conjunto de teorias e/ou conceitos capazes de abordar explicações e apoiar as capacidades de predição/prescrição no campo, principalmente quando aplicadas em conjunto com as ciências sociais.
Palavras-chave: ciência de dados; epistemologia da ciência; teoria.

Abstract
The specific term "data science" has been widely cited, both in the academic context and in the industry, which is the result of its high applicability and also the *buzzword* effect created around the term. However, a "science of data" still lacks a distinct identity and a theory-driven approach, both for its benefit and so that it can be applied properly and in conjunction with Administration. This thematic section (dossier) reflects an initiative that provides the emerging area with an opportunity to express its applicability to the field of Administration. The five articles in the dossier present a set of theories and/or concepts able to address explanations and supporting the abilities of prediction/prescription in the field, especially when applied in conjunction with the social sciences.
Keywords: data science; epistemology of science; theory.

1. Ciência de dados e ciência dos dados

O termo específico "ciência de dados" (*data science*) tem sido amplamente citado, tanto pela academia quanto pelo setor produtivo, o que é resultado de sua alta aplicabilidade e do "alarde" (Cao, 2016, p. 66) criado em torno do termo. Parte desse alarido decorre do uso do termo enquanto clichê (*buzzword*) apenas, o que explica, ao menos em parte, a "confusão" sobre o seu significado (Provost & Fawcett, 2013, p. 51). Alguns autores têm envidado esforços para definir ciência de dados adequadamente e têm, para tanto, adotado abordagens distintas. Por exemplo, a definição para ciência de dados pode adotar uma abordagem computacional, quando associada a "bancos de dados, esquemas e ontologias" (Bell et al., 2009, p. 1298); ou uma abordagem fundamentada em princípios como em "um conjunto de princípios fundamentais que apóiam e orientam a extração por princípios [*principled extraction*] de informações e conhecimentos dos dados" (Provost & Fawcett, 2013, p. 52); ou uma abordagem integradora multidisciplinar, fundamentada em estatística, computação, comunicação, gestão e sociologia (Cao, 2016); ou ainda um "paradigma" (não uma ciência), quando dispensa inclusive o uso de teorias (Hey et al., 2009, p. xi).

Apesar desses esforços, permanece em aberto uma definição suficientemente discriminante para ciência de dados. Ou seja, "ciência de dados" permanece um termo abrangente e "guarda-chuvas", compreendendo um conjunto de conceitos, tecnologias e técnicas pouco articuladas entre si. Por ora, Cao (2016) parece um dos poucos autores a ter tangenciado o que seria a essência desta ciência ao fazer

inferências sobre a natureza dos dados, considerando o conceito de propriedades. Essas inferências incipientes assemelham-se àquilo que Pitágoras – ao investigar as propriedades de certos números, a relação entre eles e os padrões identificáveis que eles formavam – fez e daí favoreceu o desenvolvimento da matemática como uma ciência (Singh, 1997).

Portanto, para definir adequadamente ciência de dados, é preciso pensar no significado de uma "ciência dos dados" (*science of data*), conforme referido por Cao (2016, p. 73), dando aos dados uma natureza específica e distinguível. Não é nosso objetivo encampar tal empreitada neste espaço, embora alguns pressupostos iniciais possam ser elaborados, de modo que possamos contribuir para uma definição apropriada de ciência dos dados (como referenciaremos daqui em diante). Assumimos, inicialmente, que a natureza exclusiva de "dados" depende essencialmente do tratamento de "dados" como entidades sujeitas a – e caracterizados por – propriedades e propensos a estabelecer relacionamentos (entre si e também com entidades da mesma natureza). Só assim, dados podem tornar-se um fenômeno *per se*, dignos de investigação por meio de uma ciência específica.

É provável que uma das dificuldades centrais de tal definição resida na (1) abrangência de aplicações da ciência dos dados; e (2) na emergência desta ciência em meio à emergência de recursos, demandas e conceitos correlatos. Por abrangência, entende-se a variedade de contextos em que "dados" podem ser o fenômeno investigado (*small data* ou *big data*, estruturados ou não-estruturados, centralizados ou distribuídos, locais ou em *streaming*, analisados em *batch* ou em tempo real, produzidos por humanos ou não-humanos, contínuos ou discretos, para suporte à decisão ou para processamento transacional etc). Por emergência, entende-se o caráter incipiente do olhar específico para "dados" enquanto fenômeno, o que ocorre simultaneamente à emergência de dados abundantes, recursos de processamento, demandas analíticas e conceitos emprestados de todas as áreas do conhecimento, inclusive do senso comum. Uma definição que supere essas dificuldades e, parcimoniosa e, compreensivamente, contemple a essência da nova ciência, pode ser fruto de um exercício coletivo de uma comunidade em formação que consiga se articular e refletir sobre a sua área, para além da aplicação imediata.

2. Teoria e ciência dos dados

A falta de um conjunto adequado de teorias para orientar os estudos da ciência dos dados também é causa de confusão em torno do campo. Mesmo considerando-se uma natureza potencialmente abdutiva da pesquisa em ciência dos dados – como tem sido sugerido por Hey et al. (2009) e Simsek et al. (2019), por exemplo –, e a eventual apologia a petabytes de dados como suficientes para aquisição de conhecimento – como é o caso de Anderson (2008) –, as ciências sociais, incluindo o campo da Administração, reivindicam explicação como uma contribuição científica indispensável. Modelos, diagramas, dados, referências ou construtos em si não são teoria (Sutton & Staw, 1995), de modo que previsões e prescrições fundamentadas apenas em tais recursos tendem a gerar desconfiança (Simsek et al., 2019). Nesse sentido, cabe crítica ao reducionismo adotado por Anderson (2008), considerando-se que: (1) análises de dados *per se* não são capazes de oferecer explicação fundamentada para fenômenos investigados e (2) ciência não se refere apenas à identificação ou ao reconhecimento de padrões (Pigliucci, 2009).

As evidências da necessidade de modelos e teoria têm sido abundantes durante o episódio de SARS-CoV-2 (também conhecido como Covid-19), quando mesmo um dilúvio de dados (*data deluge*, termo frequentemente usado em referência e suporte a *big data*) mostra-se insuficiente para explicar e, consequentemente, prever o comportamento pandêmico. Por exemplo, apesar dos esforços de Wu & McGoogan (2020) para descrição de 72.314 casos suspeitos e confirmados de Covid-19, os modelos adotados não consideraram características sociodemográficas suficientemente detalhadas de pacientes que receberam alta; ou, mesmo dispondo de milhares de casos, Xie et al. (2020) queixam-se, pois "seria muito útil desenvolver modelos matemáticos que prevejam o número esperado de pacientes e os recursos necessários (equipamento e pessoal)". Na verdade, ciência dos dados (em algumas de suas vestes mais *fashion*, como *big data* e *data mining*) tem sido muito eficaz na descrição do uso de cartões de crédito, mobilidade urbana e variações nos preços de ações de mercado, quase em tempo real (e.g. Baker et al., 2020; Pavlyshenko, 2020; Warren & Skillman, 2020). No entanto, tem cometido erros grosseiros na predição de taxas de mortalidade, propagação do contágio ou número de leitos de UTI

necessários[1].

O que dizer dos esforços de explicação dessas demandas e dos motivos de tal discrepância na predição? Em princípio, sem um conjunto de teorias e modelos derivados, a ciência dos dados não poderia reivindicar o status de ciência, resignando-se a um conjunto de técnicas para coleta, armazenamento, processamento e análise de dados. Parece-nos que o grande objetivo da ciência dos dados deveria estar bastante próximo daquilo que Johnson et al. (2012, p. 44) denominam "pesquisa informada pela teoria". O problema central de uma "ciência de dados sem teoria" (e sem o decorrente vínculo desta com o respectivo resgate da literatura) é que inexiste noção da efetiva contribuição para o conhecimento. Pode-se permanecer sistematicamente "reiventando a roda". A noção de que "passamos a saber mais sobre" pode se diluir em meio ao deslumbramento técnico/tecnológico promovido pelo "dilúvio de dados" e crescente poderio computacional. Esse é o nosso receio.

3. Esta edição especial TPA

Este dossiê reflete uma iniciativa que fornece à ciência dos dados uma oportunidade de expressão da sua aplicabilidade ao campo da Administração. Os cinco artigos do dossiê apresentam um conjunto de teorias e/ou conceitos capazes de abordar explicações e apoiar as capacidades da ciência dos dados para predição/prescrição no campo, principalmente quando aplicadas conjuntamente às ciências sociais.

O primeiro artigo, de autoria de Hilton Ramalho, Aléssio de Almeida e Alcimar Fraga, aplica a teoria dos jogos à análise de atividades ligadas à auditoria e à fiscalização de licitações do setor público. Em conjunto com técnicas de ciência dos dados, especificamente o algoritmo Apriori (que implementa regras de associação ou padrões de relacionamento entre dois conjuntos de dados), o estudo identifica padrões comportamentais de empresas licitantes e elabora indicador de suspeição por empresa, o que tem grande potencial de aplicação em tribunais de contas e demais órgãos públicos e privados de auditoria e regulação.

O artigo seguinte, de autoria de Jorge Silva, Magnus Emmendoerfer e Nina Cunha, fundamenta-se na abordagem *Design Science*, avança sobre uma abordagem metodológica particular, que dá suporte ao posicionamento teórico de argumentos extraídos de documentos públicos. Isso se dá por meio de uma proposta de sistematização operacional em fases, para extração de informações do crescente volume de documentos produzidos pela administração pública, alinhando-se às emergentes áreas de governança de dados e gestão de dados abertos.

O terceiro artigo, de autoria de Humberto Marques, Larissa Gomes, André Zambalde e André Grützmann, não adota teoria específica, embora produza subsídios que podem ser analisados à luz das "teorias da aprendizagem", com extensões nos domínios construtivista, transformativo, cooperativo etc. Esses autores implementam revisão sistemática da literatura sobre mineração de dados educacionais no ensino a distância. O artigo contribui diretamente para o desenvolvimento de conceito incipiente (*e-learning*) e chega em momento oportuno, quando a demanda por recursos *e-leaning* e o reconhecimento de sua aplicabilidade são elevados em consequência da crise pandêmica SARS-CoV-2 e medidas de distanciamento social.

Os dois últimos artigos não adotam teoria específica, mas produzem subsídios que podem ser analisados à luz da literatura sobre controle social e desenho de políticas públicas. O estudo de Andréa Silva, Aléssio de Almeida e Hilton Ramalho discute problema relevante (evasão e retensão escolares) e identifica meios de antecipação (predição) de reprovações em disciplinas específicas, com vistas à atenuação dos efeitos indesejáveis da evasão e retensão. O estudo de Luciana Militão, Paulo Silveira, Andrea León e Liziane Oliveira contribui diretamente para o desenvolvimento do conceito emergente "*on-line dispute resolution*", com análises de dados da plataforma consumidor.gov, criada em 2014 pelo governo brasileiro para moderação de disputas entre consumidores e empresas.

4. Agradecimentos e sugestão de pauta

Todos os artigos que compõem este dossiê foram avaliados por especialistas nas temáticas e

[1] Preferimos não citar casos específicos, embora amplamente conhecidos, e nos basear nas tentativas de predição que fizemos com variações sobre técnicas de regressão (simples, múltiplas, LASSO, ElasticNet etc.) e implementações *random forest*, todas com precisão razoável no curto prazo, mas incapazes de prever pouco além de 24 horas à frente.

metodologias específicas e, em média, passaram por três rodadas de avaliação e ajustes, afora a revisão de forma final. Grande ênfase foi dada às contribuições práticas, de modo que acreditamos haver contribuições valiosas para a prática gerencial, derivadas de literatura atualizada e rigorosa aplicação metodológica. Somos devedores e gratos aos autores e ao corpo editorial que entusiasticamente abraçaram a ideia desta edição especial. Sabemos que este momento de distanciamento social tem provocado mudanças significativas na rotina dos professores-pesquisadores-avaliadores. Tais dificuldades têm reflexo imediato nas demandas de avaliação de trabalhos para periódicos, já que se trata de atividade altamente dependente de engajamento pessoal/profissional (é não diretamente remunerada, tem pouco ou nenhum reconhecimento para fins de promoção na carreira e é *time-consuming*). Todos esses aspectos têm contribuído para sobrecarga nas editorias, com eventuais postergações de prazo para avaliadores e autores. Entendemos que todos esses aspectos devem ser oportunamente discutidos em forum adequado, pois cremos não serem de caráter transitório.

Referências

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7), 16-07.

Baker, S. R., Farrokhnia, R. A., Meyer, S., Pagel, M., & Yannelis, C. (2020). How does household spending respond to an epidemic? Consumption during the 2020 COVID-19 pandemic (No. w26949). *National Bureau of Economic Research*.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science*, *323*(5919), 1297-1298.

Cao, L. (2016). Data science: Nature and pitfalls. *IEEE Intelligent Systems*, *31*(5), 66-75.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Microsoft research.

Johnson, P., Ekstedt, M., & Jacobson, I. (2012). Where's the theory for software engineering?. *IEEE software*, *29*(5), 96-96.

Pavlyshenko, B. M. (2020). Regression approach for modeling COVID-19 spread and its impact on stock market. *arXiv preprint arXiv:2004.01489*.

Pigliucci, M. (2009). The end of theory in science?. *EMBO reports*, *10*(6), 534-534. Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, *1*(1), 51-59.

Singh, S. (1997). *Fermat's last theorem*: the story of a riddle that confounded the world's greatest minds for 358 years. Fourth Estate.

Simsek, Z., Vaara, E., Paruchuri, S., Nadkarni, S., & Shaw, J. D. (2019). New ways of seeing Big Data. *Academy of Management Journal*, 62(4).

Sutton, R. I., & Staw, B. M. (1995). What theory is not. *Administrative science quarterly*, 371-384.

Warren, M. S., & Skillman, S. W. (2020). Mobility changes in response to COVID-19. *arXiv preprint arXiv:2003.14228*.

Wu, Z., & McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *Jama*.

Xie, J., Tong, Z., Guan, X., Du, B., Qiu, H., & Slutsky, A. S. (2020). Critical care crisis and some recommendations during the COVID-19 epidemic in China. *Intensive care medicine*, 1-4.