

O perfil e o papel do cientista de dados

the profile and the role of the data scientist

Jorge Sandes*

* Economista do BNDES. Este artigo é de exclusiva responsabilidade do autor, não refletindo, necessariamente, a opinião do BNDES.

Economist at BNDES. the views expressed in this article are the views of the author and do not necessarily reflect the opinion of BNDES.

Resumo

A tomada de decisões pelos gestores é um assunto recorrente nas organizações e empresas nos dias atuais. Com a quantidade massiva de dados disponível nas organizações e, também, provindos de Fontes externas, as organizações têm buscado novas tecnologias e novos métodos entre os sistemas de inFormação para obtenção de inFormações com mais qualidade. *Data science* (DS), ou ciência de dados, um campo emergente no tema dos sistemas de informação, carrega as características de transformação e análise de dados de forma a ajudar a organização também no processo decisório. Este trabalho busca contextualizar e conceituar o perfil e o papel do cientista de dados, bem como o ambiente *big data*. O estudo tem como objetivo apresentar a importância desse profissional e as contribuições que ele pode trazer ao ambiente de grande volume de dados e à ciência de dados.

Falavras-chave: Sistemas de informação. *Big data*. Ciência de dados. Cientista de dados.

Abstract

Decision making by managers is a recurrent issue in organizations and companies today. With the massive amount of data available within organizations and from external sources, organizations have been looking for new technologies and methods among the information systems to obtain better quality information. *Data science* (DS), which is an emerging field in the area of information science, carries the characteristics of data transformation and data analysis in a way that helps the organization in decision making as well. This work seeks to contextualize and conceptualize the profile and role of a data scientist as well as the big data environment. The study aims to present the importance of the information professional, as well as the contributions that they can bring to the environment of large amount of data in data science.

Keywords: Information systems. *Big data*. *Data science*. *Data scientist*.

Introdução

A grande evolução tecnológica ocorrida nos últimos anos, e que se mantém constante, vem impactando diretamente diversos segmentos da sociedade moderna, e com a ciência da inFormação (CI) não seria diferente. Um dos efeitos mais perceptíveis dessa constatação é como o cotidiano dos indivíduos passou a estar repleto de dados e inFormações de variadas origens. Atividades do dia a dia das pessoas, que antes não eram monitoradas por causa da limitação tecnológica, passaram a ser fontes importantíssimas para a obtenção de dados e, conseqüentemente, de informação. Registros do cotidiano como o desempenho da educação, questões de saúde, bens e serviços, fatores relacionados ao Estado, estatísticas sobre a economia, dados sobre o consumismo etc., “passam a nos ajudar a tomar decisões e gerar conhecimento” (RI8EIRO, 2014, p. 98).

Toda organização necessita de um processo de tomada de decisão eficiente. Em sua maioria, os objetivos das organizações são definidos pelos processos de tomada de decisão, auxiliando os funcionários e os negócios a atingir seus propósitos e finalidades (3TAIR; REYWOLD3, 2010). A tomada de decisões eficiente é um dos grandes desafios dos gestores e das empresas atualmente (MORA8ITO WETO; FURE\A, 2012). A partir do conceito de tomada de decisão, é possível constatar a importância desse processo decisório nas organizações.

As companhias são levadas atualmente a estabelecer novas ferramentas de análise de seus dados para melhorar suas tomadas de decisões, aumentando assim a eficiência e eficácia em seus processos produtivos e de negócios (ELATH; 3TEIW, 2017). As Ferramentas devem sempre atender aos tomadores de decisão, bem como dispor de processamento correto de dados para a atividade de tomar a decisão.

A massiva disponibilidade de dados em uma empresa tem aumentado o interesse em métodos para obter inFormações e conhecimentos pertinentes à tomada de decisão. Com a grande oferta de recursos tecnológicos, o processo de tomada de decisões, que era baseado em experiência ou em modelos restritos da realidade, hoje é baseado em produtos de dados. Em outras palavras, uma organização pode reunir mais dados que anteriormente e analisá-los, para melhorar cientificamente suas previsões, decisões e, por fim, a eficácia e produtividade (AMIRIAW; LOGGEREW8ERG; LAWG, 2017). Com o grande volume de dados existentes nas organizações, são necessárias Ferramentas capazes de processá-los de modo eficiente, para transformá-los em inFormações capazes de agregar valor às organizações.

Um sistema de inFormação (3I) eficaz deve entregar inFormações importantes a seus usuários no tempo correto e livres de erros. InFormações Fornecidas no tempo certo implicam tomada de decisões mais eficientes (LAUDOW; LAUDOW, 2016). Por causa da grande quantidade de dados disponíveis, os 3Is necessitam ser ágeis e eficientes, para processar esses dados e fornecer de modo fácil tais inFormações sem erros e que agreguem valor aos gestores das empresas.

De fato, muitos negócios são invadidos por muitos dados, e muitas organizações estão sempre tentando capitalizar os dados por meio de análises para obter vantagem competitiva. A ciência de dados, assim como outras formas de análise de dados, faz parte dessas atividades emergentes de competição envolvendo dados entre as empresas (HAEW *et al.*, 2014). As empresas necessitam a cada dia de mais atualizações na área de análise de dados para se posicionarem de forma competitiva no mercado.

O campo da ciência de dados é uma extensão das estatísticas que são capazes de lidar com uma grande quantidade de dados que são

produzidos nos dias de hoje, e inclui conceitos de ciência da computação, de *business intelligence* e capacidade de trabalhar com algoritmos e outras Ferramentas computacionais (CIELEW; MEY3MAW; ALI, 2016). Como os negócios necessitam cada vez mais de análise de dados, combinar ferramentas pode ser uma forma de agregar valor à organização, auxiliando os tomadores de decisão com as informações obtidas pelo processamento de dados.

A ciência de dados é utilizada em quase todas as organizações, com ou sem fins lucrativos, a fim de obter conhecimento e proporcionar a clientes e usuários uma melhor experiência, armazenar essas informações internamente ou disponibilizá-las ao público (CIELEW; MEY3MAW; ALI, 2016).

Com esse cenário, acabam surgindo indagações de como áreas como a CI são afetadas e sobre o que se espera do profissional que lida diariamente com os processos de gerar, selecionar, representar, armazenar, recuperar, distribuir e usar a informação.

A ciência de dados é essencial e indispensável para empresas que desejam um resultado assertivo de sua estratégia de negócios como: perfil de seus clientes, porcentagem de lucros, novos negócios ou quadro de prejuízos. Dessa forma, a tomada de decisão para a resolução de um problema fica mais objetiva e com menos chances de erros.

A profissão de cientista de dados se resume em cinco tipos de tarefas: filtragem de dados; realização de perguntas objetivas e precisas; análise com base em dados estatísticos e desenvolvimento de *machine learning* (aprendizagem de máquina); visualização de dados e o aperfeiçoamento de modelos e algoritmos para melhores rendimentos; produção de resultados e execução. Apesar da evolução rápida dos computadores com *machine learning*, ainda é importante a presença de um cientista de dados com experiência e domínio de

toda essa tecnologia para identificar pontos em comum por meio de desafios diversos. Fara um resultado preciso e satisfatório, homem e máquina tornam-se uma equipe unida e formam uma dupla essencial. Fica claro que os resultados satisfatórios de qualquer negócio não dependem exclusivamente da quantidade de dados que uma empresa tem, mas sim da forma como serão usadas as informações decorrentes do processamento adequado desses dados, e é esse o ponto de interesse da ciência de dados e do cientista de dados (MEWE\ßEREITA3; FARFIWELLI, 2016).

Wesse sentido, o objetivo central deste trabalho é analisar o perfil e o papel do cientista de dados nas organizações.

A ciência da informação

Partindo da observação sobre a necessidade de compreensão dos processos nos quais a informação se encontra envolvida, e sua real importância para o desenvolvimento sociocultural dos indivíduos, surgiu, com isso, um novo campo do saber, que tem como missão se debruçar sobre os fluxos percorridos pela inFormação, buscando seu entendimento e otimização. Essa nova disciplina chama-se ciência da inFormação (CI).

Žaracevic (1996) discorre sobre a CI apontando as características gerais que constituem sua razão de existir e sua evolução, sendo a CI: uma área interdisciplinar, uma vez que dialoga diretamente com outros campos do conhecimento; está conectada de maneira profunda com a tecnologia da informação, posto que o “imperativo tecnológico” define a CI, assim como em outros ramos do saber, colaborando com o surgimento da chamada *sociedade da inFormação*; e, por fim, sendo a ciência de dados Fortemente vinculada com essa

sociedade da informação. Portanto, para compreender o passado, o presente e o Futuro desse campo, a CI, assim como os desafios enfrentados, é imprescindível entender essas três características Formadoras dessa área.

Os pilares da disciplina de CI foram construídos em meados da década de 1940, ao fim da Segunda Guerra Mundial. Apesar de alguns pesquisadores argumentarem que o termo “ciência da informação” só foi mencionado pela primeira vez por volta de 1960 (FIWHEIRO; LOUREIRO, 1995), é com o desfecho dos conflitos entre as nações, resultando grande desenvolvimento científico e tecnológico, que ocorre profunda propagação da informação no seio social, a chamada *explosão informacional*, o que, por sua vez, serve como ponto de partida para o surgimento da CI. A grande profusão de informação mostrou-se ser, ao mesmo tempo, positiva, porquanto servia de insumo para o desenvolvimento cada vez maior da ciência e da tecnologia, assim como negativa, pela dificuldade de se recuperar a informação produzida.

Wersig e Weveling (1975) argumentam que a CI não teve origem em outro campo de estudo, como a psicologia, nem pela junção de duas outras áreas, como a bioquímica, mas a partir da necessidade de um campo de trabalho prático, chamado *documentação* ou *recuperação da informação*. Esses autores ainda afirmam que, apesar de essa disciplina ter sido determinada em grande parte pelo surgimento de novas tecnologias, sua origem, todavia, se encontra na intersecção entre diversas outras disciplinas, na união de uma série de interesses distintos, oriundos da ciência da computação, da biblioteconomia, filosofia e taxonomia, teoria da informação etc.

Pautada na problemática acerca da recuperação da informação, considerando-se o contexto da explosão informacional vivenciada

a partir da década de 1940, surge a preocupação de estimular o debate sobre as melhores e mais adequadas soluções para se garantir a recuperação de inFormações pertinentes. Cabe salientar que o termo “recuperação da inFormação” Foi cunhado inicialmente em 1951, segundo 3aracevic (1996, p. 44), para o qual “engloba os aspectos intelectuais da descrição de inFormações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregadas para o desempenho da operação”.

Em seu artigo, 3aracevic (1996) deFende justamente isto, a importância de se pensar em modelos de recuperação da inFormação como mecanismos fundamentais para o desenvolvimento da CI. O autor arrazoa que, apesar de a recuperação da informação não ter sido a única responsável pelo avanço dessa disciplina, pode-se considerar sua principal motivadora. Esse processo de pensar a recuperação da inFormação Foi, sem dúvida, essencial para a emergência da indústria informacional.

Wessa mesma linha, 8orko (1968) busca delinear uma definição que abarque a amplitude da CI. Dessa forma, a CI seria uma disciplina preocupada com o estudo das propriedades e do comportamento inFormacional, da natureza que demarca o fluxo inFormacional, assim como o processamento da inFormação com o fim de permitir acessibilidade e otimização de uso (8OREO, 1968). Ou seja, a CI está pautada pela preocupação de compreender todo o arcabouço de conhecimentos relacionados ao fluxo inFormacional: a origem da informação, sua coleta, organização, armazenamento, recuperação, significação, diFusão, transFormação e reutilização como nova Fonte informacional.

Convém ressaltar, ainda, as características de ciência social da CI, posto que, as pesquisas desenvolvidas nessa área foram orienta-

das pela procura informacional do ser social, assim como pela necessidade de solucionar um problema social, o da informação. Le Coadic (1996) destaca que a pesquisa em CI, na busca por responder uma necessidade social, acabou prosperando em função dessa necessidade, sendo dirigida e financiada por ela. O autor também defende que, movida pela influência da TI e das máquinas de comunicar, os pesquisadores desse campo emergente tiveram como preocupação maior a utilidade, a eficácia, o prático, negligenciando, de certa forma, a teoria que fundamenta a área.

Todavia, em determinado momento, a CI transcende essa característica quase exclusiva de prática de organização, tornando-se uma ciência social rigorosa, sob efeito tanto de uma crescente demanda social quanto de grandes avanços econômicos (LE COADIC, 1996). Portanto, é possível observar que a CI é forjada a partir da percepção de ausência de uma ciência destinada exclusivamente para analisar e compreender como se dá a fluidez informacional no seio da sociedade, assim como pela busca de aprimoramento do processo de recuperação da informação, em uma época em que o incentivo à pesquisa técnico-científica, bem como o avanço das tecnologias, servia de catalisador para a propagação de fontes informacionais.

O profissional da informação

Com base no exposto, surge a indagação sobre a forma com que o chamado profissional da informação está sendo treinado para lidar com o atual cenário de extensa multiplicidade tecnológica. Além disso, cabe também o questionamento a respeito de quais seriam as características exigidas desses profissionais para suprir a nova demanda existente dos usuários de informação, sejam indivíduos ou organizações.

Realizando um levantamento acerca de trabalhos desenvolvidos com base nas novas características esperadas dos profissionais da inFormação, é possível identificar algumas conclusões alcançadas pelos autores. Cunha (2000), por exemplo, levanta a possibilidade de diversos profissionais poderem atuar nesse ambiente de suporte informacional, como os comunicadores, cientistas da computação, cientistas da informação e gestores da informação, uma vez que, a combinação dessas áreas permite a oferta de serviços de informação gerenciados, estruturados tecnologicamente, analisados e disseminados de maneira eficaz. Cabe indagar qual dessas quatro áreas conseguirá dominar esse setor de consumo de informação que, para Cunha (2000), serão aqueles que possuírem características híbridas, abarcando um pouco de cada disciplina. A autora segue afirmando que a

atividade de informação é muito vasta, envolve muitos aspectos para que seja coberta por um único profissional com uma Formação única; esta abertura e esta troca com profissionais de várias áreas proporciona [...] possibilidades de um trabalho mais diversificado e mais rico (CUNHA, 2000, p. 3).

For sua vez, Targino (2000) buscou sintetizar os requisitos básicos esperados de um profissional que lida com inFormação. Esses requisitos vão desde ter *visão gerencial* – o que permite a esses profissionais tomar decisões de maneira racional e eficiente, como questões relacionadas ao custo da inFormação e seu caráter estratégico –; a *capacidade de análise* – servindo como aporte no momento da tomada de decisão, diante dos diversos tipos de suportes, a variedade de uso da inFormação e as distintas demandas informacionais –; a *criatividade* – característica que permite agir de forma original diante de situações atuais, permitindo buscar novas soluções para proble-

mas antigos –; e, finalmente, a *análise* – processo diretamente ligado à educação continuada, no qual se espera do profissional da informação a constante busca de novas tecnologias e técnicas para auxiliar o usuário/cliente no momento da busca inFormacional.

Valentim (2000), em seu trabalho a respeito das competências que caracterizam o profissional da inFormação tido como “moderno”, esboça quatro conjuntos de habilidades consideradas imprescindíveis, a saber: competências de *comunicação, técnico-científicas, gerenciais e sociais e políticas*. As competências de comunicação dizem respeito aos produtos que possibilitam o processo de comunicação entre o usuário e a Fonte inFormacional (bibliografias, catálogos, índices etc.). Por sua vez, as competências técnico-científicas são aquelas associadas ao desenvolvimento e à execução de atividades relacionadas com o tratamento de fontes de informação, nos diferentes suportes, unidades e serviços de informação. Seguindo, as competências gerenciais abordam as atividades de Formulação, administração, organização e coordenação de unidades, sistemas, projetos e serviços de inFormação. Em conclusão, as competências sociais e políticas estão relacionadas com as ações do profissional da inFormação no âmbito da sociedade, buscando viabilizar seu desenvolvimento pessoal, institucional e social (VALEWTIM, 2000).

Já Eerreira (2003) apresenta um conjunto de habilidades consideradas essenciais pelas organizações na prática de gestão do conhecimento. Segundo a autora, essas Funções estão relacionadas com a execução de atividades na área de classificação das Fontes inFormacionais, acesso, recuperação e análise da informação, desenvolvimento de produtos e serviços a partir da informação, união do conhecimento com a experiência das pessoas dentro da organização e trabalhar com a proteção do conhecimento (EERREIRA, 2003).

Além disso, mediante a confecção de um *raþbiþg* de habilidades mais demandadas pelo mercado, a autora consegue identificar 15 habilidades, sendo que as cinco primeiras são: conhecer o ambiente de negócios da inFormação; ter Facilidade de trabalhar em grupo; ter discernimento sobre inFormações relevantes e a relevância das inFormações; ter capacidade de utilizar equipamentos eletrônicos e operar *soþwares* específicos; e ter conhecimento sobre bases de dados (EERREIRA, 2003). Wota-se, com base na análise da autora, a busca por um profissional que detenha conhecimentos relacionados a gestão de pessoas, liderança e ambiente organizacional (administração), assim como princípios voltados para a busca e análise de Fontes inFormacionais (CI), juntamente com habilidades relativas a *hardware* e *soþware* (ciência da computação).

For seu turno, 8ELLU\\O (2011) aborda cinco grupos de habilidades consideradas Fundamentais para um profissional da inFormação. O grupo *Iþformação* (I) compreende as competências que todos os profissionais devem possuir, em maior ou menor nível (as habilidades essenciais). O grupo *reþologias* (T) vai além das competências essenciais, e diz respeito à utilização de instrumentos mediados pelas tecnologias emergentes. Por sua vez, o grupo *Comnþicação* (C) aborda a interdependência entre as noções de inFormação e comunicação, assim como a complementaridade de seus meios. Já o grupo *Gesião* (M) relaciona-se com a necessidade da gestão da inFormação, bem como a administração das consequências que virão com a qualidade da inFormação que é gerida. For fim, o grupo *Oniros saberes* (3) remete aos conhecimentos que permitem ao profissional da inFormação ser versado em áreas conexas. Dessa Forma, percebe-se, mais uma vez, a exigência de que a CI e, consequentemente, o profissional da inFormação possua conhecimento holístico em campos relevantes para seu desenvolvimento, confirmando seu caráter interdisciplinar.

3anta Anna, Fereira e Campos (2014) utilizam a terminologia cunhada pela Eederação Internacional de Documentação e InFormação (EID) para se reFerir a esses novos profissionais, os modernos profissionais da inFormação (MIF). Além disso, os autores também retomam as quatro habilidades propostas por Valentim (2000), deFendendo, contudo, que novas competências necessitam ser incorporadas, buscando inserir “novas competências de cunho tecnológico, educacional e cultural” (3AWTA AWWA; FEREIRA; CAMFO3, 2014, p. 81).

As tecnologias mostram-se como as principais variáveis modificadoras do ambiente de atuação do profissional da inFormação. Coneglian, Gonçalves e 3antarém 3egundo (2017) afirmam que cabe ao profissional, que tem a inFormação como insumo de trabalho, possuir conhecimento e domínio no uso da TI. Além disso, desses profissionais espera-se o desenvolvimento de produtos de inFormação, visando o uso interno e externo de suas organizações, como a criação de bases de dados, páginas virtuais, arquivo de texto etc. Já entre as competências esperadas desses profissionais da inFormação, podem-se elencar “visão globalizada; buscar desafios; investir em novas oportunidades; comunicar-se com eficácia; criar parcerias e alianças e construir um ambiente tendo como base o respeito e a confiança” (COWEGLIAW; GOWÇALVE\; 3AWTARŹM 3EGUWDO, 2017, p. 132).

A partir do cenário apresentado, verifica-se que o contexto de atuação do profissional da inFormação vem mudando incessantemente, assim como as atribuições que se espera deles. O Fator que mais corrobora essa mudança vem sendo o surgimento de novas tecnologias de comunicação e informação. Dessa maneira, o mercado passa a buscar profissionais que estejam em constante processo de aprendizagem, evoluindo profissionalmente com o avanço das novas tecnologias e modelos. Um dos desafios surgidos recentemente

para o profissional da inFormação é o *big data*, que, nas palavras de Coneglian, Gonçalves e Zantarelli Segundo (2017, p. 132):

é caracterizado por volumes de dados extremamente densos e que necessitam de competências, habilidades e ferramentas para que essa informação possa ser encontrada; para que isso seja possível, ela necessita ser tratada, analisada e disponibilizada em tempo hábil.

E, nesse caso, pode o profissional da inFormação desempenhar papel fundamental no decorrer desse processo.

Big data

Atualmente deparamo-nos com o fenômeno de produção de dados em larga escala. Tendo como parâmetro o ano de 2012, cerca de 2,5 *exabytes* (1 *exabyte* equivale a 1 bilhão de bytes) de dados foram criados diariamente, e esse número segue dobrando a cada quarenta meses. Mais dados cruzam a internet a cada segundo do que o que foi armazenado em toda a internet há apenas vinte anos. A título de exemplo, estima-se que o Walmart colete mais de 2,5 *petabytes* de dados a cada hora de suas transações com clientes, sendo que 1 *petabyte* equivale a 1 quatrilhão de *bytes*, e 1 *exabyte* é mil vezes esse valor, ou seja, 1 bilhão de *gigabytes* (MCAEEE; 8RYWJOLE33OW, 2012).

Esse fenômeno é motivado, principalmente, pela:

drástica redução de preços para o armazenamento das inFormações; a explosão de aplicações disponíveis na internet (*e-commerce*); a popularização de sensores conectados – internet das coisas, pesquisas científicas – ao projeto genoma; e, as redes sociais (VICTORIWO *et al.*, 2017, p. 230).

Com esse cenário, surge também a necessidade de se pensar em soluções que possibilitem melhorar o tratamento e uso dos dados que são produzidos, objetivando beneficiar a tomada de decisões.

O termo “*big data*” emerge, então, como um modelo de representação das características observadas no contexto de grande profusão de dados. A partir da inserção dos computadores no seio social, meio século atrás, os dados começam a ser acumulados, permitindo o surgimento de algo novo. O mundo passa a não estar apenas repleto de informação, mas a informação começa a ser acumulada com mais rapidez. Dessa forma, o *big data* surge na esteira do avanço de ciências como a astronomia e a genômica, embora o termo, atualmente, esteja migrando para as mais diversas áreas do conhecimento (MAYER-3CHOW8ERGER; CUEIER, 2013).

Davenport (2014) afirma que *big data* nada mais é que um conjunto de dados grande o suficiente para não caber em repositórios usuais, ou seja, um volume de dados grande demais para ser guardado em servidores comuns. Além disso, ainda segundo o autor, esses dados não são estruturados o suficiente para serem alocados em bancos de dados tradicionais – organizados em linhas e colunas –, ou muito demais para serem acomodados em estruturas estáticas de armazenagem.

Wão obstante, a definição mais difundida entre os estudiosos da área advém dos chamados três Vs. Laney (2001) é o primeiro a analisar o fenômeno de grande produção de dados à luz de seu *volume*, *velocidade* e *variedade* (os três Vs). O autor observa que, com o avanço do comércio eletrônico, o que se vê é uma produção de dados em escala cada vez maior (volume) – corroborado pelo progresso crescente da capacidade de armazenamento dos bancos de dados –, com elevada rapidez de produção (velocidade) – em grande parte pelo constante avanço das tecnologias de processamento –, assim como

o diversificado conjunto de Formatos de dados que estão disponíveis (variedade) – textos, imagens, vídeos etc.

Com o passar do tempo, no entanto, novos trabalhos acerca da temática *big data* foram surgindo, e, conseqüentemente, novas proposições foram sendo firmadas a partir dos três Vs iniciais de Laney. Gandomi e Haider (2015, p. 139), por exemplo, acrescentam três dimensões: *veracidade*, *variabilidade* e *valor*. Segundo esses autores, a veracidade está relacionada com a insegurança inerente a algumas fontes de dados – ou seja, lidar com dados imprecisos e incertos pode ser outra faceta do *big data* –; por sua vez, a variabilidade está relacionada com a variação nas taxas de fluxo dos dados – essa dimensão relaciona-se com a velocidade de produção de dados, podendo haver alta e baixa velocidade (variabilidade) –; por último, valor é considerado um atributo definidor do *big data* – os dados recebidos em sua forma original geralmente têm baixo valor em relação a seu volume, podendo ser gerado alto valor a partir da análise de grandes volumes desses dados.

Embora possam existir diversas dimensões quando se trata de *big data*, de acordo com a visão específica de cada autor, é importante destacar que cada dimensão não é independente. A partir do momento que uma característica muda, existe grande probabilidade de isso se refletir nas demais (GAWDOMI; HAIDER, 2015), demonstrando forte ligação entre as dimensões relacionadas ao ambiente *big data*.

Ainda no tocante às definições existentes sobre a matéria *big data*, Rodrigues, Wóbreiga e Dias (2017, p. [5]), na busca de maior entendimento sobre o assunto, elaboram um quadro em que trazem definições e contextos que estão relacionados com o ambiente em questão, auxiliando, assim, “na compreensão do fenômeno e das concepções que ele adquire em campos distintos como Computação, Economia, Ciência da Informação”.

Quadro 1 • Definições e contextos de *big data*

Autor	Definição
Di Martino <i>et al.</i> (2010, p. 5)	“É um campo emergente onde inovadora tecnologia oferece alternativas para resolver os problemas inerentes que surgem quando se trabalha com grandes quantidades de dados, fornecendo novas maneiras de reutilizar e extrair valor a partir de informação”.
Manika <i>et al.</i> (2011, p. 1)	“Refere-se a um banco de dados cujo tamanho vai além da capacidade do software de banco de dados e ferramentas típicas para capturar, armazenar, gerenciar e analisar”.
Boyd e Crawford (2012, p. 663)	“Um fenômeno cultural, tecnológico, acadêmico e que repousa sobre a interação de tecnologia, análise e mitologia”.
Dumbill (2012, <i>on-line</i>)	“São dados que excedem a capacidade de processamento dos sistemas de banco de dados convencionais”.
Mayer-Schonberger e Cukier (2013, p. 4)	“Refere-se a trabalhos em grande escala que não podem ser feitos em escala menor, para extrair novas ideias e criar novas formas de valor de maneira que alterem os mercados, as organizações, a relação entre cidadãos e governos, etc.”.
Moura e Amorim (2015, p. 2)	“Expõe uma nova geração de tecnologia e arquitetura, destinada a extrair valor de uma imensa variedade de dados permitindo alta velocidade de captura, descoberta e análise, transformando dados em informações valiosas”.
Goularte, Zilber e Pedron (2015, p. 3)	“Não se trata apenas de uma ferramenta, mas é, em verdade, uma geração de novas tecnologias e arquiteturas projetadas para extrair valor econômico de grandes volumes de dados”.
Menezes, Freitas e Parpinelli (2016, p. 1)	“Inúmeras bases de dados estão tendendo a possuir grande volume, alta velocidade de crescimento e grande variedade. Esse fenômeno é conhecido como <i>Big Data</i> e corresponde a novos desafios para tecnologias clássicas como Sistema de Gestão de Banco de Dados Relacional”.

Fonte: Rodrigues, Wóbre e Dias (2017, p. [5-6]).

A partir das definições apresentadas no Quadro 1, alguns pensamentos pertinentes acerca do *big data* podem ser extraídos. Muitos autores destacam como sendo o objetivo maior da análise de grandes volumes de dados a geração de “inFormações valiosas” e, consequentemente, “novas ideias” que irão auxiliar as organizações

no momento da tomada de decisão. Outros evidenciam os desafios oriundos dessa nova realidade de proliferação de dados, repletando, assim, na superação da “capacidade de processamento” das tecnologias tradicionais. Isso, por sua vez, irá repercutir no desenvolvimento de “uma geração de novas tecnologias e arquiteturas”, destinadas a otimizar o processo de “captura, descoberta e análise” desse grande volume de dados. A seguir serão descritos os mais importantes princípios e tecnologias criadas para contribuir com o tratamento e análise do ambiente *big data*.

Princípios e tecnologias em *big data*

Os problemas gerados para analisar a enorme quantidade de dados, já mencionada, podem se apresentar de várias maneiras. Em certo momento, as técnicas tradicionais usadas para trabalhar com dados não conseguem mais acompanhar o ritmo de produção desses dados, disponíveis em diversos formatos.

Todavia, cabe destacar que os problemas que o *big data* traz consigo não são uma percepção recente, embora isso tenha ganhado mais espaço para discussão nas últimas décadas. A questão da “armazenagem” e “compreensão” de grandes quantidades de dados já era identificada na década de 1960, quando a empresa americana RAWD trabalhava em um projeto de *relational data file* (arquivo de dados relacionais) – sistema projetado para analisar de maneira lógica uma grande coleção de dados Factuais. Dessa forma, em 1967, dois cientistas da computação encontravam dificuldades em trabalhar com grandes conjuntos de dados, pois notaram que, com o vasto volume de dados, vinha também uma variedade de problemas de caráter lógico e linguístico, de *hardware* e *software*, práticos e teóricos, trazendo prejuízo a seus empreendimentos (CRAWFORD; MILTWER; GRAY, 2014).

Assim sendo, como discorrem Victorino e outros (2017), várias pesquisas despontam motivadas pela busca por desenvolver novas tecnologias para lidar com os problemas de armazenamento e processamento desse vasto volume de dados, produzidos em grande velocidade e de forma variada.

Analytics em big data

A seguir, são apresentados alguns princípios fundamentais no ambiente *big data*, assim como tecnologias úteis para facilitar a ação de análise e interpretação desses dados. Os tópicos escolhidos para breve descrição se baseiam em Davenport (2014) e Victorino e outros (2017).

Para falar sobre o uso de *analytics* no ambiente *big data* é preciso primeiro identificar as raízes e conceituar o termo. A definição de *analytics* está diretamente associada com a expressão *business intelligence* (BI), que, por sua vez, surge por volta da década de 1950 (DAVEWFORT, 2014), quando pesquisadores de inteligência artificial passam a utilizá-la. *Analytics* é um campo abrangente e multi-dimensional que se utiliza de técnicas matemáticas, estatísticas, de modelagem preditiva e *machine learning* para encontrar padrões e conhecimento significativos em dados.

Todavia, o BI tornou-se um termo popular nas comunidades de negócios e de TI apenas nos anos de 1990. No fim dos anos 2000, o conceito de *business analytics* foi introduzido para representar o componente analítico no BI. Mais recentemente, o termo “*big data analytics*” é usado para descrever os conjuntos de dados e técnicas analíticas em aplicações que são tão grandes (de *terabytes* a *exabytes*) e complexas (de dados vindos de sensores a mídias sociais), que exigem avançadas e exclusivas tecnologias de armazenamento, ge-

renciamento, análise e visualização de dados (CHEW; CHIAWG; 3TOREY, 2012).

Dessa forma, embora o termo *analytics* em BI tenha se difundido mais na atualidade, suas raízes datam de muito antes. O que se vê hoje é uma adaptação, para o ambiente *big data*, daquilo que já era feito com dados comuns. Davenport (2014, p. 4) destaca as diferenças básicas entre o *analytics* tradicional, utilizado com *business intelligence*, e o *analytics* utilizado em grandes volumes de dados. No primeiro, de acordo com o autor, os dados são formatados em linhas e colunas – podendo ser armazenados em bancos de dados convencionais –; o volume dos dados está na casa dos *terabytes* ou menos; o fluxo de dados é *pool* estático; os métodos de análises são baseados em hipóteses; e o objetivo principal é dar suporte ao processo decisório da organização. Já o *analytics* em *big data* tem seus dados em formatos não estruturados – o que exige bancos de dados especiais para armazená-los –; o volume dos dados gira em torno de 100 *terabytes* a *petabytes*; o fluxo de dados é constante; o método de análise é por meio de *machine learning*; e o objetivo principal é gerar produtos baseados em dados.

Logo, o que se vê é um princípio que já era utilizado anteriormente, só que desta vez em um contexto distinto. *Analytics* em *big data*, chamado por Davenport (2014) de *analytics 3.0*, está associado a grandes volumes de dados. Esses conjuntos de dados, ao contrário do *analytics* tradicional, exigem novos dispositivos tecnológicos capazes de resistir ao fluxo constante de dados, em sua mais variada forma.

Qualidade de dados em *big data*

Desde o surgimento do fenômeno da explosão informacional, a informação, relacionando-se com o contexto da tomada de decisão

e do desenvolvimento científico, passa a ter maior destaque. Wão obstante, surge o debate a respeito da qualidade dessa informação que é produzida e quais seriam os parâmetros necessários para avaliar a qualidade de determinado conjunto de dados e inFormações. Esse debate sobre a qualidade das inFormações na CI ganha Força a partir do seminário promovido pela Wordic Council For 3cientific InFormation and Research Libraries, ocorrido no ano de 1989. O encontro foi visto como um importante esforço na busca da teorização sobre o assunto e do desenvolvimento de critérios e atributos que pudessem trazer maior qualificação para os dados e inFormações (EAGUWDE3; MACEDO; EREUWD, 2018, p. 197).

A partir disso, alguns autores começam a propor pesquisas empíricas tentando identificar as dimensões que conseguiriam trazer maior entendimento à qualidade dos dados. Wang e 3trong (1996 *apnd* EAGUWDE3; MACEDO; EREUWD, 2018), por exemplo, enxergavam o conceito de qualidade de dados como sendo multidimensional. Os autores então propuseram um quadro conceitual em que a qualidade dos dados era vista a partir de quatro aspectos: a acessibilidade dos dados, a facilidade de compreensão da sintaxe e semântica dos dados; a utilidade dos dados; e a credibilidade dos dados para os usuários. Dando andamento aos estudos, os autores conseguiram definir quatro grupos de categorias, incluindo o total de 15 dimensões (atributos): *itirítseca* (precisão, objetividade, credibilidade e fidedignidade); *cotrexinal* (relevância, valor agregado, atualização, completeza e valor apropriado); *represeñaciotal* (interpretável, fácil de entender, representação concisa e representação consistente); e *acessibilidade* (acessível e seguro).

Por sua vez, a maior produção, processamento e variedade de dados em ambiente virtual (*big data*) traz consigo novas abordagens

e procedimentos para a geração, seleção e manipulação dos dados (EAGUWDE3; MACEDO; DUTRA, 2017), o que, por conseguinte, influencia as discussões relacionadas com a temática da qualidade de dados, que passa a ser tratada como qualidade de dados em *big data*.

Eagundes, Macedo e Dutra (2017), por exemplo, buscam criar um paralelo entre os aspectos utilizados na avaliação de qualidade das inFormações e as dimensões que caracterizam o ambiente *big data*. Os autores utilizam o critério de qualidade das inFormações denominado Methodology For InFormation Quality Assessment (AIMQ), que também utiliza 15 critérios para definir essa qualidade, sendo estes: acessibilidade, suficiência, credibilidade, completeza, representação concisa, representação consistente, facilidade de operação, exatidão, interpretabilidade, objetividade, relevância, reputação, segurança, atualidade e compreensibilidade. Já os aspectos selecionados para representar o ambiente *big data* foram: volume, velocidade, variedade, valor, veracidade, variabilidade e visualização. A partir da análise feita, os autores conseguiram identificar a existência de relações entre todos os critérios de qualidade da inFormação, propostos pela metodologia AIMQ, e os sete Vs usados na representação do *big data*. No entanto, não foi possível propor um modelo de qualidade inFormacional eficiente apenas com os critérios utilizados em sua análise (EAGUWDE3; MACEDO; DUTRA, 2017, p. [14]).

Também tratando da qualidade de dados e inFormações em ambientes de *big data*, Eirmani e outros (2016) abordam a dificuldade de propor uma definição única sobre qualidade dos dados no *big data*. Dessa forma, defendem os autores, existem várias noções de qualidade, aplicadas nos diferentes tipos de dados, que devem ser

cuidadosamente consideradas quando se lida com grandes volumes de dados e suas análises.

Assim como nos estudos anteriores, Eirmani e outros (2016) discorrem sobre um conjunto de dimensões capazes de capturar aspectos importantes da qualidade dos dados e informações, sendo que essas dimensões podem ser divididas em oito conjuntos: acurácia, completude, consistência, redundância, legibilidade, acessibilidade, confiança e utilidade. Wota-se grande semelhança entre os modelos propostos por Wang e 3trong (1996 *apnd* EAGUWDE3; MACEDO; EREUWD, 2018), Eagundes, Macedo e Dutra (2017) e Eirmani e outros (2016), indicando que provavelmente eles têm uma origem comum.

Apoiando-se nas pesquisas realizadas sobre o tema da qualidade dos dados em ambientes *big data*, é possível perceber a complexidade existente para criar modelos de análise que sirvam para todos os tipos de cenários de dados. Quando a proposta é direcionada a um contexto comum de dados, as dificuldades já se mostram reais, quando o foco é direcionado para o *big data*, os obstáculos podem ser ainda maiores em virtude de suas características singulares. Por exemplo, em comparação com um cenário tradicional de dados, o ambiente *big data* traz consigo duas complexidades adicionais: variadas Fontes para os dados (Fontes de origem humana, Fontes mediadas por processos e fontes geradas por máquinas) e o fato de ser altamente desestruturado e desprovido de esquemas (EIRMAWI *et al.*, 2016). O debate a respeito do tema pode ser desenvolvido a partir de várias áreas, inclusive de campos de estudo que surgem para tratar de Forma específica a problemática *big data*, como a ciência de dados, assunto da próxima seção.

Ciência de dados – *data science*

Fara Elath e 3tein (2017), com a onipresença de dados em todos os setores pessoais e empresariais, o desejo de se ter conhecimento sobre os negócios e o valor dos dados está aumentando. Portanto, a ideia de “análise de dados” descreve a ciência de dados no contexto de uma organização, e essa ideia tem se expandido rapidamente nelas nos últimos anos.

O surgimento da ciência de dados é recente e decorre de questões pragmáticas, dada a necessidade das empresas e organizações não governamentais de conhecer seus clientes e aumentar a própria eficiência. Muitos aspectos de diversas organizações são agora potencialmente abertos à coleta de dados. A disponibilidade desses dados tem levado a um crescente interesse em métodos para extrair inFormações e conhecimento dos dados para auxílio na tomada de decisão (AMIRIAW; LOGGEREW8ERG; LAWG, 2017).

Em 1962, John W. Tukey, um estatístico americano, já defendia a necessidade da realização de análises a partir de dados. O autor de *me fninre of daia atalysis* explicava que, durante muito tempo, ele acreditou estar interessado apenas nas inFerências Feitas do particular para o geral, proporcionadas pelos métodos oriundos da estatística clássica. Mas, a partir de determinado momento, percebeu que seu interesse estava de fato na área de análise de dados (FRE33, 2013). Aliás, ele pensava que a estatística deveria passar a Fazer a análise de dados, assumindo, assim, características de ciência aplicada aos negócios, em vez de ser vista apenas como um ramo da matemática pura.

For meio de uma linha temporal, é possível perceber como a ciência de dados evoluiu ao longo das últimas décadas (mesmo que o uso

do termo só tenha se tornado mais comum a partir das décadas de 1990-2000). E fundamentado nessa cronologia, nota-se que os primeiros princípios norteadores da atual ciência de dados começam a se desenvolver na década de 1960, inicialmente com os trabalhos de Tukey. Desde então, o que se vê é o surgimento de livros, publicações seriadas, artigos, *workshops*, entidades e encontros de especialistas que abordam, inicialmente, os temas de processamento e análise de dados e, posteriormente, *big data* e ciência de dados, sendo alguns exemplos: os livros *Exploratory data analysis*, em 1977, de John W. Tukey; e *From data mining to knowledge discovery in databases*, em 1996, de Usama Fayyad, Gregory Elbert H. Shapiro e Padhraic Smyth; os periódicos *The Journal of Data Mining and Knowledge Discovery*, lançados em 1997; e *Data Science Journal*, lançados em 2002; o livro *Competing analytics*, em 2005, de Thomas H. Davenport, Don Cohen e Al Jacobson; o artigo *Rise of the data scientist*, em 2009, de Wathan Yau; o artigo *What Is Data Science?*, em 2010, de Mike Loukides; o artigo *Data scientist: the sexiest job of the 21st century*, em 2012, de Thomas H. Davenport e D. J. Patil; entre outros. Assim sendo, a busca progressiva da ligação entre métodos estatísticos tradicionais com as tecnologias computacionais, que estão em constante evolução, mostra ser algo bastante característico dessa genealogia da ciência de dados.

Nos dias de hoje, há diversas funcionalidades que são focadas em dados. A internet utilizada é formada por serviços de dados e por bases de dados. O papel de um cientista dos dados nunca foi mais importante, gerando oportunidades de criações de produtos de dados que têm seu valor extraído dos próprios dados e gerando mais dados como resultado, aumentando ainda mais seu valor (COSTA; SANTOS, 2017).

O *business intelligence* pode se integrar com a ciência de dados nos negócios para criar impactos positivos na análise de uma grande quantidade de dados, criando *onipnis* na visualização e na análise interativa, e a alta gerência pode entender esses *onipnis* sem um conhecimento técnico (WEWMAW *et al.*, 2016).

De acordo com Costa e Santos (2017), a *data science*, nos últimos anos, tem atraído grande atenção, surgindo como um casamento entre a estatística e a ciência da computação. Porém, se assume que as habilidades de um cientista de dados vão além das dessas duas áreas.

Ciência de dados é uma atividade interdisciplinar com foco em extrair conhecimento de dados de diversas maneiras, sendo a nova fonte de conceitos como *data mining* e análise de dados. A partir da *ciência de dados*, a análise de dados passou a ser auxiliada por estatísticas e algoritmos de aprendizado de máquina (*machine learning algorithms*), para produzir produtos de dados ou modelos que funcionam como análises descritivas, preditivas e prescritivas.

Além disso, por ser também multidisciplinar, a ciência de dados possibilita que diversos especialistas de múltiplas áreas trabalhem e estudem em conjunto. A principal razão de ser tão atraente para as organizações é sua associação ao processo, análise, interpretação dos dados (WEWMAW *et al.*, 2016). A área em evolução da ciência de dados combina campos como matemática, estatística, ciência da computação, ciência do comportamento e análise preditiva.

A Figura 1 mostra os três pilares da ciência de dados: dados, tecnologia e pessoas. “Dados” são todos os dados utilizados, estruturados ou não; “tecnologia” engloba as utilizadas para processar os dados; e “pessoas” podem incluir os cientistas de computação, estatísticos, cientistas de dados e analistas de negócios envolvidos com a *ciência de dados* (3OWG; \HU, 2016).

Figura 1 • Os três pilares da *data science*



Fonte: Song e \hu (2016).

Fara 3chutt e O³Weil (2014), *data science* é uma série de ações de melhoria utilizadas nas empresas para lidar com uma ampla gama de problemas que podem ser solucionados com dados, e por isso pode merecer o nome de ciência, apesar de muitas vezes ser tratada como a “solução para todos os problemas”, o que deve ser evitado. Os resultados de um trabalho de ciência de dados podem ser, entre outros:

- análise exploratória de dados;
- visualizações (relatórios);
- *dashboards* e métricas;
- resultados de conhecimento da organização;

- tomada de decisão baseada nos dados;
- engenharia de dados e *big data*;
- pesquisas e investigações;
- otimizações diversas;
- encontro de correlações com dados.

Segundo Frovost e Eawcett (2013), a ciência de dados abrange uma série de princípios e técnicas que auxiliam a análise automatizada de dados, com o objetivo principal de melhorar o processo decisório. A tomada de decisão baseada em dados se refere à prática de basear as decisões na análise dos dados, e não apenas na intuição.

Fara Amirian, Lojerenberg e Lang (2017), *ciência de dados* é a ciência que usa de métodos computacionais para identificar e encontrar padrões nas séries de dados. Seu objetivo principal se dá no conhecimento adquirido que afeta as decisões, tornando estas mais consistentes e eficientes, ajudando os tomadores de decisão.

Ainda para Amirian, Lojerenberg e Lang (2017), os dados são necessariamente uma medida de informação histórica e, por definição, a ciência de dados analisa dados históricos. No entanto, os dados utilizados na ciência de dados podem ter sido coletados há alguns anos ou há alguns milissegundos, com um processo contínuo. Portanto, seu processo pode ser em tempo real ou próximo do tempo real.

Em relação aos principais desafios de pesquisa em ciência de dados, Maneth e Foulouvassilis (2017) incluem: o desenvolvimento de técnicas computacionais capazes de escalonar os volumes e as variedades de dados que são gerados por meio de tecnologias baseadas em *web*,

móveis e difusas; a proporção de dados que estão sendo produzidos por empresas de grande porte; as aplicações científicas e de mídias sociais; o desenvolvimento de Ferramentas de limpeza, transformação, modelagem, análise, integração e visualização de dados, permitindo aos cientistas de dados entender e melhorar a veracidade do *big data* e extrair valor com maior rapidez, Facilidade e confiabilidade; e, por fim, a garantia de segurança, privacidade e propriedade de dados das organizações e dos usuários.

Wo que diz respeito à aplicação da ciência de dados em áreas tradicionais da sociedade, é cada vez mais perceptível o surgimento gradual de novas iniciativas que buscam tirar maior proveito desse campo. No setor governamental, por exemplo, como asseveram Viviani, Forto e Ogasawara (2015), há uma grande profusão de bases de dados que podem possibilitar a oportunidade para análise das atividades desenvolvidas pelo setor público, objetivando tornar o planejamento mais eficiente, além de criar novos serviços que melhorem o relacionamento com o cidadão.

Por sua vez, existem empreendimentos na iniciativa privada que buscam trazer um conjunto de vantagens de negócio pela análise de dados. Isso ocorre não apenas com os dados produzidos pelas próprias entidades, mas também com dados que são comercializados com terceiros, visando agregar maior valor aos serviços e produtos que serão desenvolvidos, como apontado por Loukides (2010, p. 2, tradução nossa), quando afirma que

a questão enfrentada por toda empresa hoje, *startups*, organizações sem fins lucrativos [...] que desejam atrair uma comunidade é como usar os dados de maneira eficaz, não apenas seus próprios dados, mas todos os dados que estão disponíveis e são relevantes.

Um exemplo disso é o comércio eletrônico de dados, *e-commerce*, pelo qual diversas empresas compram dados de navegação de usuários da internet com o objetivo de desenvolver *marketing* direcionado, baseado no histórico de navegação desses usuários.

De acordo com Wewman *et al.* (2016), a ciência de dados nos negócios inclui os conceitos dos cinco Vs do *big data*: volume, velocidade, variedade, veracidade e valor. Volume se refere à quantidade e ao volume propriamente dito de dados que são processados. Velocidade significa a rapidez com que os dados são gerados para análise nos negócios e variedade, pois existem inúmeros tipos de dados que os negócios trabalham. Veracidade parte do pressuposto de que a análise correta desses dados resulte em dados que reflitam a realidade, aumentando o desenvolvimento do negócio. Por fim, valor trata da agregação de valor às organizações com a utilização de um modelo de ciência de dados.

A ciência de dados tornou-se relevante para as organizações. Primeiramente, proporcionou a elas adquirir e analisar os dados de suas operações, assim como seu desempenho e sua estratégia. Em segundo lugar, permitiu a melhoria de suas operações e serviços, com base nos resultados das análises. Em terceiro lugar, o negócio pode ampliar a qualidade de suas previsões, para os tomadores de decisão planejarem suas estratégias (WEWMAW *et al.*, 2016).

Dessa maneira, e nesse contexto, uma definição pontual da ciência de dados é a feita por Loukides (2010, p. 1, tradução nossa, grifo nosso), ao afirmar que

apenas o uso de dados não é exatamente o que se quer dizer com “ciência de dados”. Uma aplicação de dados adquire seu valor dos dados em si, criando mais dados como resultado. Não é apenas uma aplicação com dados; é um produto de dados. *A ciência de dados permite a criação de produtos de dados.*

Ou seja, a finalidade máxima da ciência de dados é permitir que sejam desenvolvidas aplicações que utilizem dados como insumos na geração de produtos e serviços para as organizações. Assim sendo, quais seriam as características e os requisitos de um profissional apto a trabalhar na construção desses produtos e serviços?

O cientista de dados

Como exposto anteriormente, o cenário atual é de uma vasta produção de dados, em grande parte decorrente do maior desenvolvimento das tecnologias de computadores, redes e sensores, bem como do barateamento e da rapidez do processamento e armazenamento que é feito pelas máquinas. Em 2013, por exemplo, de acordo com a International Data Corporation (EMC/IDC),¹ havia cerca de 4 *zeiabyies* de inFormações armazenadas no mundo, e esse montante vem dobrando a cada dois anos. Para se ter uma ideia, 1 *zeiabyie* corresponde a 1.020 *byies*, o que equivaleria ao poder de armazenamento de centenas de CD-ROMs distribuídos para cada ser humano (ERICEZ, 2015).

Davenport e Fatil (2012) afirmam que a carreira de cientista de dados é a mais “*sexy*” do século XXI. Tal afirmação decorre, em grande parte, da demanda por profissionais que consigam lidar com o cenário *big data* e a consequente necessidade de obtenção de resultados a partir desses dados, gerando vantagens competitivas para as organizações. Corroborando esse contexto de procura por especialistas em

¹ INTERNATIONAL Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. Disponível em: www.idc.com/about. Acesso em: 25 de abril de 2019.

ciência de dados, uma análise do Google Trends, em 2012, mostrou haver uma crescente busca, de usuários de diversos países, por inFormações relacionadas com os termos “*data science*” e “*data science*”, buscas essas quase sempre combinadas com termos sobre formação profissional, cursos, salário, habilidades necessárias e certificação profissional (CURTY; 3ERAEIM, 2016). Dessa Forma, percebe-se a maior compreensão, tanto das organizações quanto dos profissionais, acerca das potencialidades e oportunidades proporcionadas por essa nova atividade.

O ambiente do *big data*, mais evidente nas últimas duas décadas, assim como o entendimento tanto corporativo – o viés empresarial –, quanto acadêmico – o viés do conhecimento – sobre a necessidade de uma área que fosse responsável por trazer soluções para os problemas oriundos da enorme produção de dados, propiciou o surgimento de um novo tipo de profissional responsável por desenvolver produtos e serviços a partir desses dados, designado *cientista de dados*. Segundo Miller (2013), esses cientistas são os mágicos da era do *big data*. Eles analisam os dados utilizando modelos matemáticos e criam narrativas ou visualizações que consigam explicá-los, e depois sugerem como usar as inFormações para tomar decisões.

For sua vez, Davenport e Fatil (2012) afirmam que o cientista de dados é um profissional de alto nível, com treinamento e curiosidade suficientes para conseguir efetuar descobertas no mundo do *big data*. Segundo os autores, o termo foi cunhado em 2008 por D. J. Fatil e JeΦ Hammerbacher, respectivos líderes das iniciativas sobre análise de dados no LinkedIn e no Facebook.

Ainda de acordo com Davenport e Fatil (2012), o que os cientistas de dados fazem é realizar descobertas enquanto “nadam em dados”. Estando à vontade no mundo digital, eles conseguem estruturar

grandes quantidades de dados que não têm Forma, tornando possível analisá-los. Esses profissionais são capazes de encontrar ricas fontes de dados, conectando-as a outras fontes que estão incompletas (DAVEWFORT; FATIL, 2012), propiciando, assim, que trabalhem em conjunto para alcançar os objetivos previamente traçados.

O termo “cientista de dados” Foi cunhado para definir os profissionais que trabalhavam com aplicações de dados e que tinham um impacto nas organizações descobrindo e entendendo melhor questões de negócios a partir dos dados, explorando-os de uma maneira científica (CO3TA; 3AWTO3, 2017).

ConForme 3chutt e O³Weil (2014), o cientista de dados é um cientista que pode provir de diversos campos, das ciências sociais à biologia, que trabalham com uma grande quantidade de dados e necessitam trabalhá-los computacionalmente, com seus respectivos problemas, estruturas, tamanhos e complexidades, solucionando problemas do mundo real. Deve estar alinhado à estratégia da organização em que está inserido, resolvendo questões desde a engenharia e inFraestrutura de dados da empresa até preocupações acerca de tomada de decisões dos gestores.

Segundo Costa e 3antos (2017), para melhor distinguir “cientista de dados” de “analista de dados”, considera-se que o papel do cientista de dados seja uma evolução do papel de analista de dados, pois ele dispõe de habilidades em negócios e comunicações para superar desafios nas organizações, agregando valor a elas. Além disso, o cientista de dados combina conhecimentos da estatística e da ciência da computação, porém se encaixa melhor sob o leque dos sistemas de inFormação.

Os principais papéis do cientista de dados são: extrair conhecimento dos dados para resolver problemas nas organizações; realizar as

perguntas corretas e necessárias para alinhar seus resultados com os objetivos do negócio em questão; identificar os dados corretos a usar ou reutilizar; selecionar as melhores tecnologias e Ferramentas; analisar, avaliar e visualizar os dados; e, por fim, ajudar a tomada de decisão relacionada aos dados, entre outras Funções menores.

Habilidades e competências de um cientista de dados

O mercado de trabalho para o cientista de dados, assim como para um profissional de qualquer área, busca um especialista que esteja munido de um conjunto de habilidades tidas como fundamentais. Inicialmente, vislumbrava-se que todas essas competências almejadas estivessem em um único indivíduo, como explica Davenport (2014). Todavia, a partir de determinado momento, tendo em conta a dificuldade para encontrar um profissional considerado completo e que estivesse disponível no mercado, passa a ser admitido um modelo mais realista quanto ao saber exigido desses indivíduos. Isso permite que diversos especialistas, das mais variadas áreas, unam seus conhecimentos no momento de apresentar soluções para analisar, tratar e interpretar a grande quantidade de dados existentes.

Os cientistas de dados são distinguidos em dois tipos: os verticais e os horizontais. Os verticais são aqueles especialistas que possuem um profundo conhecimento em algum campo específico (cientistas da computação, estatísticos, engenheiros de *software* etc.), cada qual, sendo um *expert* de sua área, podendo agregar valor a processos específicos da análise, tratamento e interpretação dos dados. Por sua vez, os horizontais são os cientistas de dados que têm um pouco de conhecimento em cada uma das áreas, que podem contribuir com a dinâmica na qual os dados estão inseridos. Em virtude do abrangente entendimento que possuem, esses últimos são os profissio-

nais mais almejados pelo mercado, conseguindo, muitas vezes, lidar individualmente com todo o processo de análise, interpretação e obtenção de resultados a partir dos dados. No entanto, como observado anteriormente, esses profissionais são escassos, em grande parte pela Falta de cursos de Formação específicos para cientistas de dados nas instituições de ensino.

Wo Brasil, por exemplo, até o ano de 2016, existiam poucas iniciativas de cursos na modalidade *laio seṡsn*, como asseveram Curty e 3erafim (2016), o que gera prejuízo no desenvolvimento das habilidades buscadas. Muitas vezes, as próprias empresas contratantes oferecem cursos que possibilitam o aperfeiçoamento de seus cientistas de dados (DAVEWFORT; FATIL, 2012).

Entre as habilidades buscadas em um cientista de dados, nas palavras de Davenport (2014, p. 85), estão: “[ser] um *hacber*, um cientista, um analista quantitativo, um conselheiro de confiança e um *experi* em negócios”. O autor apresenta um quadro, exposto a seguir, em que melhor especifica cada uma dessas habilidades:

A partir do Quadro 2 é possível depreender que um cientista de dados ideal, de acordo com Davenport (2014), deve ter um conjunto de características e competências que abarque desde o entendimento razoável de programação e arquiteturas desenvolvidas especificamente para o ambiente *big data*, perpassando os princípios básicos da estatística, de extrema importância na ocasião da mineração e tratamento dos dados, chegando aos fundamentos acerca da gestão de negócios, liderança e proatividade, ou seja, conceitos advindos da administração. Isso corrobora o entendimento da ciência de dados como um campo extremamente interdisciplinar.

Quadro 2 • Habilidades do cientista de dados

Cientista
Tomada de decisões baseada em evidências Improvisação Impaciência e inclinação à ação
Conselheiro de confiança
Grandes habilidades de comunicação e relacionamento Capacidade de elaborar decisões e entender os processos decisórios
Analista quantitativo
Análise estatística Visual <i>analytics</i> Aprendizado de máquina Análise de dados não estruturados, como texto, vídeo ou imagens
Expert em negócios
Compreensão de como o negócio funciona e lucra Boa noção de onde aplicar o <i>analytics</i> e o <i>big data</i>

Fonte: Davenport (2014, p. 86).

Eim e Lee (2016) traçam sua própria representação das habilidades e conhecimentos imprescindíveis para a atuação de um cientista de dados. Os autores chegaram a esse resultado realizando um levantamento de ofertas de emprego na área de ciência de dados em três sites americanos (Indeed.com, Monster.com e Career8uilder.com) e identificando os requisitos exigidos pelos empregadores. Com base nesse levantamento, Eim e Lee (2016) chegaram a um conjunto de habilidades que compreendem três classes: sistemas, negócios e técnicas. Por sua vez, essas três classes se subdividem em subclasses que apontam para as competências demandadas. É possível perceber, *grasso modo*, que as habilidades assinaladas por esses autores têm uma “espinha dorsal” bastante similar às aquelas identificadas por Davenport (2014), mostrando tendência às áreas de computação e estatística, além de alguns princípios de administração.

Quadro 3 • Habilidades e conhecimentos de um cientista de dados

Sistemas	Negócios	Técnicas
Desenvolvimento	Social	Software
Análise; Implementação/teste; Gestão de dados; Conhecimento de diferentes tecnologias; Desenvolvimento de metodologias; Programação; Operação/manutenção; Integração; Documentação.	Habilidades interpessoais; Comunicação; Automotivação.	Linguagem de programação; Banco de dados/ <i>data warehouse</i> ; Plataformas <i>open source</i> ; Domínio de diferentes pacotes de <i>software</i> ; Visualização de dados.
Negócios		
Conhecimento específico do setor/negócio; Habilidade de análise macro; Negócios <i>on-line/e-commerce</i> .		
Solução de problemas	Gerencial	Arquitetura de redes
Modelagem de dados; Análise quantitativa/estatística; Pensamento analítico/lógico; Criatividade/inovação; Capacidade para solução de problemas; Adaptabilidade/flexibilidade; Capacidade estratégica.	Administração geral; Organização/Liderança; Capacidade de monitoramento e controle; Planejamento; Treinamento; Gestão de mudança; Gerenciamento de projetos.	Internet; Dispositivos de rede; Computação em nuvem; Arquitetura e segurança de rede.
Hardware		
Dispositivos de armazenamento; Impressoras; <i>Desktop/PC</i> ; Servidores/Estações de trabalho.		

Fonte: Adaptado de Eim e Lee (2016, p. 166).

For seu turno, Rodrigues, Wóbreiga e Dias (2017) também esboçam um cenário em que são apresentadas algumas competências exigidas de um cientista de dados. Percebe-se que, ao contrário de Davenport (2014) e Eim e Lee (2016), esses autores deram maior ênfase às habilidades relacionadas com os campos da ciência da

computação e da estatística, tidas, para muitos, como aquelas imprescindíveis para todo profissional que decida trabalhar com grandes volumes de dados.

Quadro 4 • Competências esperadas do cientista de dados

- t Capacidade de estruturar grandes volumes de dados amorfos
- t Tornar os dados possíveis para análise
- t Identificar fontes de grandes volumes de dados e cruzar com outras fontes
- t Criar ferramentas e analisar grande quantidade de dados
- t Domínio de ferramentas que deem conta do volume de dados (Hadoop, por exemplo)
- t Formação em qualquer área, desde que tenha foco em dados e na computação

Fonte: Adaptado de Rodrigues, Wóbrega e Dias (2017, p. [10-11]).

O último ponto, anotado por Rodrigues, Wóbrega e Dias (2017), diz respeito à Formação que se espera de um profissional que decida trabalhar com dados. Essa matéria sustentada pelos autores, assim como por Davenport (2014), indica que muitos profissionais que atuam na área de dados não têm Formação necessariamente em ciência de dados, mas em campos correlatos, como a ciência da computação e a estatística, até pela ainda escassa oferta de cursos que sejam específicos, como já exposto anteriormente.

Assim sendo, o cientista de dados emerge como sendo um profissional que tem como principal missão trazer clareza sobre o cenário de grande produção e acúmulo de dados. Não obstante, espera-se desse especialista um conjunto de saberes e habilidades que são essenciais ao desempenho do papel para o qual é destinado.

Recentemente, diversas *startups*, como Wubank, Weon e Quinto Andar, contrataram milhares de cientistas de dados, dada a importância desse profissional no contexto atual. Produtos já são criados especificamente para seus clientes a partir da gama de dados que tais corporações possuem. A tendência é que essas empresas contra-

tem não mais economistas ou contadores ou administradores, mas sim profissionais com essas habilidades que tenham também conhecimento de ciência de dados, ou seja, os profissionais horizontais.

Cabe destacar a imensa lacuna que há no setor público com relação a esses profissionais, e desconhece-se, até hoje, a realização de qualquer concurso público para a carreira de cientista de dados. Como o governo é detentor de inúmeros dados sobre a população, há aí uma enorme oportunidade a ser abarcada, com grandes impactos sobre diversas políticas públicas, o que é de suma importância em cenário de restrição fiscal. Fara citar um exemplo, um relatório da Controladoria-Geral da União (CGU) de abril de 2017 (8RA3IL, 2018) mostra que 11 estados e o Distrito Federal jogaram remédios Fora em 2014 e 2015 em razão da validade vencida e do armazenamento incorreto, causando um prejuízo de R\$ 16 milhões.

Considerações finais

Weste trabalho, verificou-se que a área de CI surge da constatação da necessidade de compreender os processos nos quais a informação se encontra envolvida e sua real importância para o desenvolvimento sociocultural dos indivíduos. Além disso, a explosão informacional também contribuiu com o processo de nascimento dessa disciplina, iniciada em meados da década de 1940, assim como a necessidade de recuperação da informação, questão que ganha destaque por volta da década de 1960, em grande parte pelo avanço das novas tecnologias. Dessa forma, a CI surge com a missão de se debruçar sobre o fluxo percorrido pela inFormação, buscando seu entendimento e sua otimização.

Essa evolução tecnológica também trouxe grandes avanços para o meio social, desde *hardware* (equipamentos) até *software* (lógica responsável por possibilitar o funcionamento dos equipamentos), influenciando, de maneira ímpar, a Forma como os indivíduos interagem com o mundo que os circunda, com os outros e com eles mesmos.

Provocado justamente por essa evolução tecnológica, a partir do barateamento do armazenamento das inFormações e do advento da internet e das mídias sociais, surge o conceito de *big data*. Esse fenômeno se constitui como grandes “volumes” de dados, produzidos em rápida “velocidade” e com grande “variedade” de formatos. Partindo-se desse contexto, acaba emergindo, em conjunto, a necessidade de um profissional que consiga apresentar soluções, realizar análises e extrair valor dessa grande quantidade de dados, o que, por sua vez, colabora com o surgimento de um novo campo de estudo, a chamada ciência de dados. Os cientistas de dados podem ser considerados mágicos da era do *big data*. Eles conseguem analisar os dados por meio de modelos matemáticos, criando narrativas e visualizações que possibilitam explicá-los.

Uma das características da profissão de cientista de dados é seu viés multifacetado, sendo que diversos especialistas podem exercer essa atividade, desde cientistas da computação, estatísticos, administradores, profissionais da inFormação, entre outros. Isso se deve à abrangência das etapas relacionadas com a análise dos dados, sendo: identificação das Fontes, captura e armazenamento, acesso, análise e, por fim, exposição dos resultados dos dados analisados. Todavia, destaca-se que essa profissão é marcada por sua natureza mais voltada a conhecimentos relacionados com a computação – linguagens algorítmicas, tecnologias para *big data*, banco de dados, aprendiza-

do de máquina etc. – e com a estatística – análise quantitativa, mineração de dados, visualização de dados etc., – posto que a ciência de dados surgiu pautada nessas duas disciplinas.

O profissional da inFormação pode contribuir com a aplicação de conceitos sobre dados e inFormações que vão além do paradigma computacional, como gestão e fluxo dos dados.

Há um grande espaço para o cientista de dados no setor público brasileiro, detectando padrões na grande massa de dados que o governo possui, podendo, assim, otimizar diversas políticas públicas, o que acarretaria um grande aumento de eficiência.

Portanto, apesar de a temática sobre *big data*, ciência de dados e cientista de dados ainda ser vista quase que exclusivamente com o olhar das tecnologias, da ciência da computação e/ou da estatística, áreas, de fato, extremamente importantes para o assunto, todavia, existe, sim, espaço para debate e colaboração em outros campos, inclusive nas organizações, com o intuito de ajudar nas tomadas de decisão.

Referências

AMIRIAW F.; LOGGEREW8ERG E.; LAWG T. *Big data in healthcare* – extracting knowledge From point-of-care machines. OE, University of Oxford:3pringer, 2017.

8ELLU\O, R. C. 8. As competências do profissional da inFormação nas organizações contemporâneas. *R88D. Revisia Brasileira de Bibliorecoçomia e Docnmeçriação*, São Paulo, v. 7, n. 1, p. 58-73, jan./jun. 2011.

8OREO, H. InFormation science: what is it? *Americaç Docnmeçriaçoiç*, Baltimore, v. 19, n. 1, p. 3-5, jan. 1968. Disponível em: <https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoEerneda/k---artigo-01.pdf>. Acesso em: 29 ago. 2019.

8RA3IL. Ministério da Transparência e Controladoria-Geral da União. Secretaria Executiva. *Relatório de Gestão - Exercício 201Z*. Brasília, 2018.

Disponível em: <https://www.cgu.gov.br/sobre/auditorias/arquivos/2017/cgu-se-relatorio-gestao-2017.pdf/view>. Acesso em: 10 jul. 2019.

CHEW, H.; CHIAWG, R. H. L.; STOREY, V. C. Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, [s.l.], v. 36, n. 4, p. 1.165-1.188, Dec. 2012.

CIELEW, D.; MEYMAW A. D. 8.; ALI M. *Introducing data science*. New York: Manning, 2016. 300p.

COWEGLIAW, C. 3.; GOWÇALVE, F. R. V. A.; SAWTARZM SEGOWDO, J. E. O profissional da informação na era do *big data*. *Estudos Bibli: revista eletrônica biblioteconomia e ciência da informação*, Florianópolis, v. 22, n. 50, p. 128-143, set./dez. 2017.

CO3TA, C.; SAWTO3, M. Y. The data scientist profile and its representativeness in the European e-competence Framework and the skills Framework For the inFormation age. *International Journal of Information Management*, [s.l.], n. 37, p. 726-734, 2017.

CRAWEORD, E.; MILTWER, E.; GRAY, M. L. Critiquing big data: politics, ethics, epistemology. *International Journal of Communication*, [s.l.], v. 8, p. 1.663-1.672, 2014.

CUNHA, M. V. O profissional da informação e o mercado de trabalho. *Informação & Sociedade: Estudos*, João Pessoa, v. 10, n. 1, p. 1-5, 2000.

CURTY, R. G.; SERAEM, J. 3. A Formação em ciência de dados: uma análise preliminar do panorama estadunidense. *Informação & Formação*, Londrina, v. 21, n. 2, p. 307-328, mai./ago. 2016.

DAVENPORT, T. H. *Big data no trabalho*: derrubando mitos e descobrindo oportunidades. Tradução de Cristina Yamagami. Rio de Janeiro: Elsevier, 2014.

DAVEWFORT, T. H.; FATIL, D. J. Data scientists: the sexiest job of the 21st century. *Harvard Business Review*, [s.l.], v. 90, n. 10, p. 70-76, Oct. 2012.

EAGOWDE3, F. 8.; MACEDO, D. D. J.; EREUWD, G. F. A produção científica sobre qualidade de dados em *big data*: um estudo na base de dados *Web of Science*. *RD8CI: Revista Digital de Biblioteconomia e Ciência da Informação*, Campinas, v. 16, n. 1, p. 194-210, jan./abr. 2018.

EERREIRA, D. T. Profissional da inFormação: perfil de habilidades demandadas pelo mercado de trabalho. *Ciência da Informação*, Brasília, v. 32, n. 1, p. 42-49, jan./abr. 2003.

- FIRMANI, D. *et al.* On the meaningfulness of "big data quality". *Data Science and Engineering*, [s.l.], v. 1, n. 1, p. 6-20, Mar. 2016. Disponível em: <<https://link.springer.com/article/10.1007/s41019-015-0004-7>>. Acesso em: 27 ago. 2019.
- ELATH, C. M.; STEIN, W. Towards a data science toolbox for industrial analytics applications. *Computers in Industry*, Würzburg, Germany: Julius Maximilians University, n. 94, 2017.
- ERICEZ, M. Big data and its epistemology. *Journal of the Association for Information Science and Technology*, [s.l.], v. 66, n. 4, p. 651-661, 2015.
- GAWDOMI, A.; HAIDER, M. Beyond the hype: big data concepts, methods, and analytics. *International Journal of Information Management*, [s.l.], v. 35, n. 2, p. 137-144, Apr. 2015.
- HALEW, S. T. *et al.* Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, [s.l.], n. 154, p. 72-80, 2014.
- EIM, J. Y.; LEE, C. E. An empirical analysis of requirements for data scientists using online job posting. *International Journal of Software Engineering and Its Applications*, [s.l.], v. 10, n. 4, p. 161-172, 2016.
- LANEY D. *3D data management*: controlling data volume, velocity, and variety. META group Inc., 2001. Disponível em: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Acesso em: 14 mai. 2019.
- LAUDOW, J.; LAUDOW, E. *Management information system – managing the digital firm*. 14. ed. Global Edition. United States of America: Pearson, 2016. 675p.
- LE COADIC, Y-F. *A ciência da informação*. Tradução de Maria Yêda E. S. de Eigueiras Gomes. Brasília: Briquet de Lemos, 1996.
- LOUEIDE3, M. What is data science?: the Future belongs to the companies and people that turn data into products. *O'Reilly Radar*, [s.l.], June 2, 2010.
- MAWETH, S.; FOULOVA33ILI3, A. Data science. *The Computer Journal*, [s.l.], v. 60, n. 3, p. 285-286, 2017.
- MAYER-SCHWENGER, V.; CUEIER, E. *Big data*: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Tradução de Paulo Folzono Junior. Rio de Janeiro: Elsevier, 2013.

MCAEEE, A.; 8RYWJOLE33OW, E. Big data: the management revolution. *Harvard Business Review*, [s.l.], v. 90, n. 10, p. 60-68, Oct. 2012.

MEWE\ E3; 3. R. L; EREITA3; R. 3; FARFIWELLI, R. 3. Mineração em grandes massas de dados utilizando hadoop mapreduce e algoritmos bio-inspirados: uma revisão sistemática. *Revista de InFormática Teórica e Aplicada (ot-lite)*. PortoAlegre, 2016.

MILLER, C. C. Data science: the numbers of our lives. *me New Yorb rimes*, Wew York, 23 Apr. 2013.

MORA8ITO WETO, R.; FURE\ A V. Modelagem e simulação. *In*: MIGUEL, P. A. C (Org.). *Meiodologia de pesqnisa em etgepharia da prodnção e gesião de operações*. 2. ed. Rio de Janeiro: Elsevier: Abepro, 2012. p. 169-198.

WEWMAW, R. *et al*. Model and experimental development for business data science. *Interaiotat Jonrtal of Itformaiot Matagemet*, Rio de Janeiro, Elsevier, n. 36, p. 607-617, 2016.

FIWHEIRO, L. V. R; LOUREIRO, J. M. M. Traçados e limites da ciência da informação. *Ciêtcia da Itfformação*, Brasília, v. 24, n. 1, abr. 1995.

FROVO3T, E.; EAWCEW, T. *Data scietce for bnsitess*. United States of America: O³Reilly Media, Inc., 2013. p. 409.

FRE33, G. A very short story oF data science. *Forbes*, [s.l.], May 28, 2013. Disponível em: <http://www.mat.uFrGs.br/~viali/estatistica/mat2274/material/textos/A%20Very%203hort%20History%200F%20Data%203cience%20-%20Eorbes.pdf>. Acesso em: 27 ago. 2019.

RIBEIRO, C. J. S. *8ig daia*: os novos desafios para o profissional da inFormação. *Itfformação & reçtologia (IrEC)*, João Fessoa/Marília, v. 1, n. 1, p. 96-105, jan./jun. 2014.

RODRIGUE3, A. A.; W68REGA, E; DIA3, G. A. Desafios da gestão de dados na era *big daia*: perspectivas profissionais. *In*: ENCONTRO NACIONAL EM FE3QUISA EM CIZWCIA DA IWEORMAÇÃO, 18, 2017, Marília. *Atais...* Marília, 3F: Unesp, 2017. p. [1-19].

3AWTA AWWA, J.; FERREIRA, G.; CAMFO3, 3. O. 3ociedade da inFormação x biblioteconomia: em busca do moderno profissional da inFormação (MIF). *R88D. Revista Brasileira de Bibliorecoçomia e Docnmetração*, São Paulo, v. 10, n. 1, p. 68-85, jan./jun. 2014.

- 3AWT03, C. J. Atuação profissional da inFormação no processo de inteligência competitiva. *Rebecit*, v.3, n. 2, p. 26-50, jul./dez. 2016. Disponível em: <https://goo.gl/dEqb3L>. Acesso em: 24 jul. 2017.
- 3ARACEVIC, T. Ciência da inFormação: origem, evolução e relações. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.
- 30WG I., \HU Y. 8ig data and data science: what should we teach? *Experí Systems*. Philadelphia, Drexel University, Wiley Publishing Ltd., v. 33, n. 4, 2016.
- 3CHUW, R.; O³WEIL, C. *Doitg daia scietce*. United 3tates oF America: O³Reilly Media, Inc., 2014. 405p.
- 3TAIR, R. M.; REYWOLD3, G. W. *Prinípios de sistemas de itfomação*. Tradução da 9. ed. norte-americana. [S.l.], Cengage Learning, 2010.
- TARGIWO, M. G. Quem é o profissional da inFormação? *rra3itfomação*, Campinas, v. 12, n. 2, p. 61-69, jul./dez. 2000.
- TUEEY, J. W. *me collecied worbs of Joh3 W. rnrbey*. London: Chapman Hall, 1991.
- VALEWTIM, M. L. F. O moderno profissional da inFormação: Formação e perspectiva profissional. *Etco3tros Bibli: revisia eleirô3ica de bibliotecotomia eciê3cia da itfomação*, Elorionópolis, v. 5, n. 9, p. 16-28, jan. 2000.
- VICTORINO, M. C. *er al*. Uma proposta de ecossistema de *big daia* para a análise de dados abertos governamentais conectados. *Itfomação & Sociedade: Esindos*, João Fessoa, v. 27, n. 1, p. 225-242, jan./abr. 2017.
- WAWG, R. Y.; 3TROWG, D. M. Beyond accuracy: what data quality means to data consumers. *Jonrtal of Ma3agemet3 Itfomaio3 System*, v.12, n. 4, 1996, p.5-34. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/>. Acesso em: 20 mar. 2019.
- WER3IG, G; WEVELIWG, U. fie phenomena oF interest to inFormation science. *Itfomaio3 Sciet3rist*, [s.l.], v. 9, n. 4, p. 127-140, Dec. 1975.
- \IVIAWI, A.; FORTO, E.; OGA3AWARA, E. Ciência de dados: desafio para a ciência, indústria e governo. *ComCiê3cia: Revisia Eleirô3ica de Jor3alismo Cie33ífico*, Campinas, v. 170, jul. 2015.