XXIII SEMEAD
Seminários em Administração

novembro de 2020
ISSN 2177-3866

O QUE FAZEM CIENTISTAS DE DADOS? Uma Revisão da Literatura Visando Identificar as Principais Habilidades de uma Profissão em Evidência

FABIANO CASTELLO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DA UNIVERSIDADE DE SÃO PAULO - FEA

CESAR ALEXANDRE DE SOUZA
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DA UNIVERSIDADE DE SÃO PAULO - FEA

O QUE FAZEM CIENTISTAS DE DADOS?
Uma Revisão da Literatura Visando Identificar as Principais Habilidades de uma Profissão em Evidência

1. Introdução

O potencial do big data em gerar insights e criar novas formas de valor transformam as organizações e a sociedade. Apesar declaração de que o cientista de dados tem "o trabalho mais sexy do século 21" (Davenport & Patil, 2012), ainda não são abundantes os estudo que analisam esta profissão de forma a, com rigor, defini-la e estudar suas características, principalmente em relação as habilidades que devem possuir.

A demanda de mercado por cientistas de dados aumentou na mesma medida em que aumentou a demanda por big data, e essa demanda surge não apenas de "startups" como também de grandes corporações, sob pressões competitivas globais por inovação e qualidade nas ofertas de serviços e produtos (Chatfield et al., 2014).

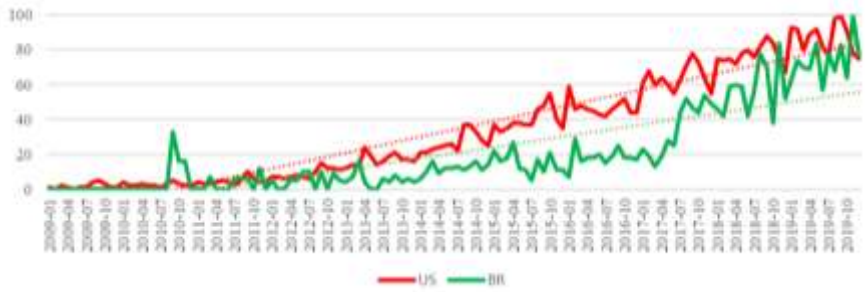
Segundo Yan & Davis (2019) o termo "ciência de dados" foi criado por Jeff C. Wu e utilizado pela primeira vez em 1997, sendo que um ano depois o próprio Jeef C. Wu sugeriu que o termo "ciência de dados" fosse utilizado como um nome moderno para designar a ciência estatística. Poucos anos depois, Cleveland (2001) criou o esboço de um plano para uma nova disciplina, com abrangência mais ampla que a da estatística, e a chamou de "ciência de dados". Do ponto de vista de publicações, o "The International Council for Science: Committee on Data for Science and Technology" começou a publicação do "Data Science Journal" em 2002, enquanto a Universidade de Columbia iniciou a publicação do "The Journal of Data Science" em 2003.

O termo "ciência de dados" tornou-se popular nos anos 2000 por conta do crescimento de companhias baseadas na internet, tais como Yahoo, Google, LinkedIn, Facebook e Amazon, bem como diversas startups, tais como Palantir, Everstring, the Climate Corporation e Stitch Fix. Atualmente o termo "ciência de dados", em conjunto com o termo "big data", tornaram-se termos frequentes nos negócios, nas notícias, na mídia, nas redes sociais e na academia, sendo "cientista de dados" um dos mais populares cargos relacionados (Yan & Davis 2019; Columbus 2018).

Patil (2011) afirma que o termo "ciência de dados" é antigo, e geralmente referia-se à inteligência de negócios. No entanto, "cientista de dados" parecia ser novo, uma vez que, desde que foi cunhado em 2008, não encontrou ninguém que já o usasse antes.

De fato, uma análise do termo cientista de dados pesquisado no Google Trends (Figura 1) identifica que a partir de 2011 inicia-se uma tendência no interesse ao tema, que permanece constante até o final de 2019.

Figura 1 Interesse, segundo Google Trends, pelos termos relacionados à Cientistas de Dados



fonte: autor; dados: Google Trends

Atualmente, a profissão de cientista de dados é considerada uma das melhores opções profissionais segundo relatório do site de empregos Glassdor e a revista Forbes. O relatório define a profissão como a melhor opção, inclusive pelo 4º. ano seguido, baseado num índice de satisfação dos praticantes, no número de posições disponíveis e na mediana da remuneração (Columbus, 2019).

Não há um consenso sobre o termo "cientista de dados" (Harris 2013), particularmente em relação as habilidades que fazem de um profissional um cientista de dados. Este artigo propõe-se a trazer luz especificamente para este aspecto: o das principais habilidades de cientistas de dados, esta relativamente nova e pungente profissão.

2. Problema de Pesquisa e Objetivo

A ausência de uma definição clara do que é um cientista de dados, e principalmente quais são suas principais habilidades, traz prejuízo (a) aos cidadãos, principalmente jovens, em busca da profissão de cientista de dados, no sentido que tem dificuldade de definir um caminho de aprendizado, e (b) às empresas buscando cientistas de dados, no sentido que tem dificuldades para definir o perfil do profissional que precisam recrutar.

Dado este problema de pesquisa, que pode ser resumido como a falta de informações abundantes, consensadas e amplamente divulgadas sobre habilidades de cientistas de dados, o objetivo deste artigo é apresentar um recorte específico sobre habilidades de cientista de dados, resultado de uma revisão sistemática de literatura realizada no contexto de pesquisa acadêmica de mestrado.

3. Fundamentação Teórica

Este artigo foi produzido a partir da pesquisa de dissertação de mestrado onde uma ampla revisão sistemática de literatura (RSL) foi planejada e conduzida, visando, através de uma busca consistente e parametrizada, identificar e interpretar todas as informações existentes sobre o tema estudado, de forma completa e imparcial (Kitchenham, 2004).

A revisão foi desdobrada em três eixos para facilitar seu entendimento: (a) qual a definição de cientista de dados? (b) quais as principais áreas de formação? e (c) quais as principais habilidades e ferramentas? Neste artigo o recorte está restrito apenas às habilidades.

A partir de termos como "data literacy", "analytics", "skills", "competencies" e "abilities", sempre utilizados no contexto de ciência de dados ou de cientistas de dados, foram geradas "strings" de buscas que foram pesquisadas nas bases de publicações científicas ACM, EBSCO, Elsevier Science Direct, Google Scholar, IEEE, Proquest, Scopus e Web of Science.

O condução desta pesquisa produziu 2.245 documentos. Adotando o protocolo de Kitchenham (2004), o processo de seleção foi realizado utilizando critérios previamente definidos. Na primeira fase foram excluídos documentos cujo título não está relacionado com a questão de pesquisa, artigos periféricos ou não diretamente relacionadas ao tema, artigos não

publicados em periódicos acadêmicos ou conferências, e artigos cujo conteúdo completo não era possível ser acessado de acordo com os convênios disponibilizados pela universidade. Esta fase da revisão produziu 259 documentos.

Na segunda fase foram analisados os documentos resultantes da primeira fase, com base na leitura de seus sumários e de suas conclusões, atentando para incluir artigos cujo conteúdo contribuiu para definir cientistas de dados ou permitiram auxiliar no processo de identificação de suas áreas de formação e de seu conjunto de habilidades. Uma síntese das bases de dados utilizadas e artigos produzidos está na figura 2.

Figura 2 - Bases de Dados e Número de Documentos

	WoS	Scopus	IEEE	Proquest	Ebsco	Elsevier	ACM	Google Scholar	Total
Resultados Obtidos	780	673	340	95	67	89	149	52	2.245
RSL Fase 1	61	74	20	17	37	8	12	30	259
RSL Fase 2	10	12	5	9	6	2	3	7	54

Fonte: elaborado pelo autor.

O resultado da segunda fase da revisão produziu 54 documentos, e a partir da leitura integral de cada um dos 54 artigos selecionados nas etapas anteriores foi realizado o processo de extração de dados (Tenório et al., 2016), de forma que todas as informações necessárias para abordar os eixos específicos da RSL fossem identificados (Kitchenham, 2004). Conforme informado anteriormente, este artigo faz um recorte específico em relação às habilidades de cientistas de dados. Dos 54 trabalhos analisados na revisão sistemática de literatura, 30 apresentaram conteúdo passível de extração de dados sobre habilidades de cientistas de dados.

4. Discussão

Cientistas de dados dependem de um arcabouço de infraestruturas e aplicações para realizar seu trabalho. Por ser um campo amplo, existem muitas tecnologias e técnicas, não apenas tradicionais como também emergentes, que crescem anualmente, criando um cenário heterogêneo (Salado-Cid et al., 2017).

Dentre as diversas estratégias estudadas para analisar e resumir as informações de forma a melhor apresentá-las para o leitor, a alternativa selecionada foi a segregação de todas as informações em dois grupos distintos: “hard skills” e “soft skills”. Esta segregação está presente na revisão de literatura no trabalho de Pires e Leitão (2018).

Especificamente em relação ao uso do termo “habilidade” é importante citar que, de uma forma geral, utiliza-se corriqueiramente o termo “competência” de forma intercambiável com o termo “habilidade”. O dicionário Webster (1981) define competência, na língua inglesa, como: “qualidade ou estado de ser funcionalmente adequado ou ter suficiente conhecimento, julgamento, habilidades ou força para uma determinada tarefa”. O dicionário de língua portuguesa Aurélio enfatiza, em sua definição, aspectos semelhantes: capacidade para resolver qualquer assunto, aptidão e idoneidade, bem como introduz outro: capacidade legal para julgar pleito. Nos últimos anos, o tema “competência” entrou para a pauta das discussões acadêmicas e empresariais, associado à diferentes instancias de compreensão: no nível da pessoa, das organizações e dos países (Fleury & Fleury, 1981). Nota-se, desta forma, que o termo “competência” está aberto a um conjunto amplo de interpretações. Já o termo “habilidade” é bem mais restrito e, segundo o dicionário Michaelis da Língua Portuguesa, significa “conjunto

de qualificações para o exercício de uma atividade ou cargo; suficiência." Adicionalmente, o termo "skills" é prevalente na literatura. Dentre as traduções possíveis da palavra "skill" para a Língua Portuguesa, segundo o Google Tradutor, as mais frequentes são "habilidade", "proficiência", "perícia" e "destreza", sendo "habilidade", dentre elas, a mais frequente. Em função das razões apresentadas, neste estudo optou-se por utilizar, de forma exclusiva, o termo "habilidade", com uma única exceção: para evitar a tradução literal de "hard skill" e "soft skill" como habilidades "duras" ou "moles", exclusivamente esses termos foram mantidos em inglês.

Conforme mencionado anteriormente, na revisão de literatura as definições de "*hard skills*" e "*soft skills*" são encontradas no trabalho de Pires & Leitão (2018), sendo que o primeiro representa conhecimentos e capacidades que podem ser adquiridos através de programas de educação e treinamento, enquanto o segundo, também chamado de "people skills", é a capacidade de se relacionar com outras pessoas para expressar ideias e o desejo de realizar o trabalho. Para a apresentação das habilidades de cientistas de dados as definições de Pires & Leitão (2018) foram confirmadas e expandidas através da análise do trabalho de Laker & Powel (2011), que diferencia "*soft skills*" e "*hard skills*" de forma que o primeiro está relacionado a habilidades intrapessoais e interpessoais, enquanto que o segundo são habilidades essencialmente técnicas. As definições acima são a base para segregação e apresentação das habilidades de cientistas de dados encontradas na literatura.

O resultado da consolidação de habilidades encontra-se a seguir, nas tabelas 1 a 4, segregadas da seguinte forma: (a) "*Hard Skills*", segregados em "Conhecimentos Gerais em Computação e Desenvolvimento de Sistemas", "Conhecimentos Gerais em Negócios" e "Inteligência Artificial"; e (b) "*Soft Skills*".

A coluna "frequência" apresenta em quantos artigos analisados a habilidade está presente, e as tabela estão ordenadas por essa informação.

Para algumas habilidades, em geral aquelas mais frequentes, então apresentadas considerações após a tabela. Essas habilidades possuem uma letra-índice para facilitar a localizações das considerações.

Tabela 1: "*Hard Skills*" - Conhecimentos Gerais em Computação e Desenvolvimento de Sistemas

	Frequência	Chatfield et al. (2014)	Debortoli et al. (2014)	Manieri et al. (2015)	Waller & Fawcett (2013)	Wiktorski et al. (2017)	Abidin et al. (2017)	Attwood et al. (2017)	Bašk. & Koronios (2017)	Baumer (2017)	Cao (2017)	Costa & Santos (2017)	De Mauro et al. (2017)	Dichev & Dicheva (2017)	Ecleo & Galido (2017)	Kotzé (2017)	Markow (2017)	Gardiner et al. (2018)	Gibert et al. (2018)	Hu et al. (2018)	Knorr et al. (2018)	Luna-Reyes (2018)	Meyer (2019)	Mikalef (2018)	Pires & Leitão (2018)	Demchenko et al. (2019)	Kross & Guo (2019)	Mikalef (2019)	Miller (2019)	Verma et al. (2019)	Yan & Davis (2019)
Engenharia de Software e Análise de Sistemas (a)	18	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓					✓			✓	✓	✓			
Data Management (b)	15	✓	✓	✓		✓	✓	✓				✓		✓	✓			✓				✓	✓			✓	✓	✓			
Trabalhar com Big Data (c)	14	✓		✓		✓			✓		✓	✓	✓					✓	✓				✓	✓	✓	✓		✓			
Infraestrutura TI (d)	10			✓		✓					✓	✓								✓		✓	✓			✓	✓	✓			
Mineração de Dados	9			✓	✓	✓	✓	✓												✓	✓					✓		✓			
Computação em Nuvem	8			✓		✓		✓	✓		✓											✓				✓		✓			
Data Warehouse	6		✓	✓		✓												✓								✓			✓		
Privacidade e Segurança	6			✓		✓	✓					✓									✓					✓					
Data Quality	5			✓		✓			✓			✓														✓					
Computação de Alta Performance	4			✓		✓		✓																		✓					
Computação Distribuída	4			✓		✓					✓															✓					

Fonte: elaborada pelo autor.

Tabela 1: "*Hard Skills*" - Conhecimentos Gerais em Computação e Desenvolvimento de Sistemas (continuação)

	Frequência	Chatfield et al. (2014)	Debortoli et al. (2014)	Manieri et al. (2015)	Waller & Fawcett (2013)	Wiktorski et al. (2017)	Abidin et al. (2017)	Attwood et al. (2017)	Bašk. & Koronios (2017)	Baumer (2017)	Cao (2017)	Costa & Santos (2017)	De Mauro et al. (2017)	Dichev & Dicheva (2017)	Ecleo & Galido (2017)	Kotzé (2017)	Markow (2017)	Gardiner et al. (2018)	Gibert et al. (2018)	Hu et al. (2018)	Knorr et al. (2018)	Luna-Reyes (2018)	Meyer (2019)	Mikalef (2018)	Pires & Leitão (2018)	Demchenko et al. (2019)	Kross & Guo (2019)	Mikalef (2019)	Miller (2019)	Verma et al. (2019)	Yan & Davis (2019)
Pré-processamento de dados, ETL	4			✓		✓				✓																✓					
Computação em Tempo Real ("Streaming")	3			✓		✓																				✓					
Controle de Versionamento ("GIT")	3			✓		✓																				✓					
Armazenamento de dados	2						✓																					✓			
Hardware	1														✓																
Integração de Dados	1																											✓			
Limpeza e preparação de dados	1																											✓			

Fonte: elaborada pelo autor.

Tabela 2: "*Hard Skills*" - Conhecimentos Gerais em Conhecimentos Gerais em Negócios

	Frequência	Chatfield et al. (2014)	Debortoli et al. (2014)	Manieri et al. (2015)	Waller & Fawcett (2013)	Wiktorski et al. (2017)	Abidin et al. (2017)	Attwood et al. (2017)	Bašk. & Koronios (2017)	Baumer (2017)	Cao (2017)	Costa & Santos (2017)	De Mauro et al. (2017)	Dichev & Dicheva (2017)	Ecleo & Galido (2017)	Kotzé (2017)	Markowet al. (2017)	Gardiner et al. (2018)	Gibert et al. (2018)	Hu et al. (2018)	Knorr et al. (2018)	Luna-Reyes (2018)	Meyer (2019)	Mikalef (2018)	Pires & Leitão (2018)	Demchenko et al. (2019)	Kross & Guo (2019)	Mikalef (2019)	Miller (2019)	Verma et al. (2019)	Yan & Davis (2019)
Negócios (e)	19	✓		✓		✓	✓		✓	✓	✓	✓	✓		✓					✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Visualização de Dados (f)	17	✓		✓		✓	✓	✓			✓	✓	✓	✓	✓	✓						✓	✓	✓		✓		✓			✓
Comunicação (f)	14	✓					✓		✓	✓	✓	✓	✓		✓	✓		✓					✓		✓		✓			✓	
Gestão de Projetos	10			✓		✓					✓	✓	✓		✓			✓								✓	✓	✓			
Trabalho em time	7	✓										✓						✓					✓		✓			✓		✓	
Liderança	6								✓						✓			✓							✓			✓		✓	
Metodologias (g)	6			✓		✓					✓				✓									✓		✓					
Storytelling (f)	6			✓		✓	✓									✓							✓			✓					
Ética	5						✓					✓	✓	✓							✓										
Resolução de Problemas	5										✓	✓			✓										✓						✓
Conformidade	3			✓		✓																				✓					
Gestão de Processos de Negócios	3			✓		✓																				✓					
Conhecimento Multidisciplinar	3	✓									✓	✓																			
Gestão	2														✓															✓	

Fonte: elaborada pelo autor.

Tabela 2: "*Hard Skills*" - Conhecimentos Gerais em Conhecimentos Gerais em Negócios (continuação)

	Frequência	Chatfield et al. (2014)	Debortoli et al. (2014)	Manieri et al. (2015)	Waller & Fawcett (2013)	Wiktorski et al. (2017)	Abidin et al. (2017)	Attwood et al. (2017)	Bašk. & Koronios (2017)	Baumer (2017)	Cao (2017)	Costa & Santos (2017)	De Mauro et al. (2017)	Dichev & Dicheva (2017)	Ecleo & Galido (2017)	Kotzé (2017)	Markow (2017)	Gardiner et al. (2018)	Gibert et al. (2018)	Hu et al. (2018)	Knorr et al. (2018)	Luna-Reyes (2018)	Meyer (2019)	Mikalef (2018)	Pires & Leitão (2018)	Demchenko et al. (2019)	Kross & Guo (2019)	Mikalef (2019)	Miller (2019)	Verma et al. (2019)	Yan & Davis (2019)
Gestão de Conhecimento	2										✓													✓							
Visão Estratégica	2						✓								✓																
Auditoria	1												✓																		
Desenvolvimento de Produtos	1																						✓								
Gestão de mudança	1														✓																
Fluência em inglês	1																							✓							
Orientação ao cliente	1														✓																

Fonte: elaborada pelo autor.

Tabela 3: "*Hard Skills*" - Inteligência Artificial

	Frequência	Chatfield et al. (2014)	Debortoli et al. (2014)	Manieri et al. (2015)	Waller & Fawcett (2013)	Wiktorski et al. (2017)	Abidin et al. (2017)	Attwood et al. (2017)	Bašk. & Koronios (2017)	Baumer (2017)	Cao (2017)	Costa & Santos (2017)	De Mauro et al. (2017)	Dichev & Dicheva (2017)	Ecleo & Galido (2017)	Kotzé (2017)	Markowet al. (2017)	Gardiner et al. (2018)	Gibert et al. (2018)	Hu et al. (2018)	Knorr et al. (2018)	Luna-Reyes (2018)	Meyer (2019)	Mikalef (2018)	Pires & Leitão (2018)	Demchenko et al. (2019)	Kross & Guo (2019)	Mikalef (2019)	Miller (2019)	Verma et al. (2019)	Yan & Davis (2019)
Métodos Quantitativos (h)	25	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓		✓		✓	✓	✓	✓
“Machine Learning” (i)	19	✓	✓	✓		✓					✓	✓		✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		
Modelagem de Dados (j)	7	✓					✓							✓	✓					✓									✓		✓	
Processamento de Linguagem Natural (NLP)	5			✓		✓	✓																			✓		✓			
Redes Neurais	3												✓													✓			✓			
Sistemas de Recomendação ou Classificação.	3			✓		✓																				✓					
Deep Learninig	2	✓																							✓						
Regressão	2																												✓	✓	
Agrupamento ("clustering")	1																													✓	
Análise de Redes Sociais	1						✓																								
Classificação	1												✓																		
Otimização	1				✓																										
Previsões ("forecast")	1				✓																										

Fonte: elaborada pelo autor.

Tabela 4: "*Soft Skills*"

	Frequência	Chatfield et al. (2014)	Debortoli et al. (2014)	Manieri et al. (2015)	Waller & Fawcett (2013)	Wiktorski et al. (2017)	Abidin et al. (2017)	Attwood et al. (2017)	Bašk. & Koronios (2017)	Baumer (2017)	Cao (2017)	Costa & Santos (2017)	De Mauro et al. (2017)	Dichev & Dicheva (2017)	Ecleo & Galido (2017)	Kotzé (2017)	Markowet al. (2017)	Gardiner et al. (2018)	Gibert et al. (2018)	Hu et al. (2018)	Knorr et al. (2018)	Luna-Reyes (2018)	Meyer (2018)	Mikalef (2018)	Pires & Leitão (2018)	Demchenko et al. (2019)	Kross & Guo (2019)	Mikalef (2019)	Miller (2019)	Verma et al. (2019)	Yan & Davis (2019)
Pensamento Analítico	3										✓	✓			✓																
Criatividade	2										✓				✓																
Empreendedorismo	2	✓																										✓			
Pensamento Crítico	2						✓				✓																				
Aprendizado contínuo	1									✓																					
Curiosidade	1	✓																													
Paixão	1																							✓							
Proatividade	1																							✓							
Relacionamento interpessoal	1															✓															
Senso de Responsabilidade	1																							✓							

Fonte: elaborada pelo autor.

A seguir são apresentadas as considerações sobre determinados itens presentes nas tabelas:

a) Considera-se “Engenharia de Software e Análise de Sistemas”, de uma forma geral, como aquelas habilidades citadas na literatura que estão relacionadas com desenvolvimento de sistemas e atividades afins. Na literatura aparecem em diversas formas, tais como engenharia de software, análise de sistemas, programação, desenvolvimento de scripts, desenvolvimento de algoritmos, extração, limpeza, enriquecimento e transformação de dados, “habilidades, técnicas, metodologias e conhecimentos relacionados ao design, construção e implantação de sistemas de informação”, “suporte ao desenvolvimento de aplicações”, “conhecimentos gerais sobre tecnologia, gestão de dados, projeto, análise, implementação e teste, conhecimento de metodologias de desenvolvimento, integração, programação, documentação, operações e manutenção, desenvolvimento para aplicativos móveis, algoritmos”, aplicativo baseado na Web, programação estruturada múltipla, linguagens de consulta (“query”), princípios de engenharia de sistemas e software para o design e desenvolvimento de sistemas de informação das organizações, incluindo o design de requisitos, uso integrado com sistemas corporativos e colaborativos, algoritmos eficientes para acessar e analisar grandes quantidades de dados, incluindo APIs para diferentes bancos de dados, gestão de relacionamento com cliente (CRM), requisitos de experiência do usuário (UX) e design; metodologia e plataforma de desenvolvimento e gerenciamento de software ágil Scrum.

b) O termo “Data Management” agrupa habilidades que estão relacionadas, de uma forma geral, a capacidade de gerenciar dados fora do contexto de projetos de inteligência artificial. Na literatura tais habilidades são citadas como criação e gerenciamento da arquitetura de informação da organização e gerenciamento de ativos de dados; gerenciamento de dados (curadoria, preparação, análise exploratória, integração de diversas fontes de dados); especificação, desenvolvimento e implementação do gerenciamento de dados corporativos, bem como definir estratégia para arquitetura de governança de dados; desenvolvimento e implementação de arquitetura de dados, tipos e formatos de dados, modelagem e design de dados, incluindo tecnologias relacionadas (ETL, OLAP, OLTP, etc).

c) “Trabalhar com Big Data” são habilidades referentes a dados não estruturados, como por exemplo utilização de bancos de dados NoSQL e mineração de texto; tecnologias baseadas na nuvem para sistemas e aplicativos de processamento de grandes conjuntos de dados; desenvolvimento de políticas de governança de big data.

d) “Infraestrutura TI” refere-se a habilidades que aparecerem na literatura principalmente como implementação de modelos em produção, mas também como infraestrutura de reprodutibilidade (por exemplo utilização de Docker e Virtual Machines); orquestração de negócios e TI; gerenciamento e operação de infraestrutura e serviços; utilização de Unix ou de suas variações, como por exemplo Linux; arquitetura geral de redes, computação em nuvem, cliente-servidor e redes distribuídas, internet, LAN e WAN, devices de rede (firewall, routers).

e) Negócios ou “Domain Knowledge” é uma categoria de habilidades que é prevalente na literatura, e está diretamente relacionada ao conhecimento do cientista de dados sobre negócios, ou conhecimento sobre domínio ou, ainda, entendimento do contexto dos dados que são base para algum tipo de análise que está sendo realizada. Exemplos citados são marketing, finanças, saúde e cadeia de suprimento. Estão presentes também o entendimento dos desafios de negócios, dos principais motivadores da competitividade digital, do conhecimento de casos de sucesso, do desenvolvimento de cultura orientada a dados e do desenvolvimento de planos de negócios.

f) “Visualização de Dados” contempla a habilidade de utilizar softwares de visualização

de dados. Aqueles citados na literatura são Datawrapper, API de visualização do Google, Google Charts, Flare, D3.js, Tableau, Raphael, Gephi e Qlik. Esta habilidade em geral está presente explorando o fato de que cientistas de dados devem atentar para aspectos de comunicação em todas as fases de um projeto, mas, sobretudo, em relação aos resultados, que são melhor comunicados de maneira visual para pessoas não envolvidas com dados. Na literatura "comunicação" aparece em várias nuances: oral, escrita e não técnica, bem como a habilidade de falar com vários públicos e, ainda, a comunicação interpessoal. "Storeytelling" é uma habilidade que, quando presente, está associada a comunicação e a visualização de dados.

g) Habilidades relacionadas à metodologia aparecem na literatura como métodos de pesquisa e validação empírica; métodos de pesquisa, tanto no contexto de pesquisa acadêmica como no contexto de negócios.; tecnologias de desenvolvimento ágil, como DevOps e ciclo de aprimoramento contínuo, para aplicativos orientados a dados; conhecimento de pacotes e "*framework*s".

h) Habilidades em métodos quantitativos foram agrupadas quando aparecem na literatura de forma genérica, como por exemplo análises quantitativas em geral, matemática, estatística, testes de hipótese, simulação de eventos discretos e probabilidade aplicada. Aparecem também técnicas gerais de análise estatística e análise descritiva, preditiva e prescritiva, métricas de desempenho para avaliação e validação de análise de dados, pesquisa operacional, otimização e simulação.

i) As habilidades agrupadas na categoria "machine learning", ou aprendizado de máquina, incluem aprendizado supervisionado, não supervisionado ou de aprendizado por reforço, bem como modelagem preditiva. Esta categoria existe para agrupar a habilidade de trabalhar com "machine learning" de forma genérica. Quando o artigo revisado menciona um algoritmo específico, como por exemplo "basket analysis", este é computado na tabela de forma também específica.

j) Habilidades relacionadas à modelagem de dados poderiam ter sido classificadas dentro de conhecimento gerais de TI. A razão por estar agrupada dentro de "Inteligência Artificial" é que é uma habilidade essencial na criação de "datasets", ou conjunto de dados, que constituem a base para treinamentos de modelos de inteligência artificial.

5. Conclusão / Contribuição

O objetivo deste artigo é apresentar um recorte específico sobre habilidades de cientistas de dados, resultado de uma revisão sistemática de literatura realizada no contexto de pesquisa acadêmica de mestrado, que pode contribuir para auxiliar (a) cidadãos, particularmente jovens no momento de definição de seus primeiros passos profissionais, que desejam tornar-se cientista de dados, e este artigo pode ajuda-los a melhor entender do se trata a profissão, da mesma forma que pode contribuir com (b) profissionais já maduros atualmente desocupados, ou sem oportunidades de trabalho no seu ofício corrente, que percebem-se vislumbrando outra profissão; e (c) para empresas publicarem vagas com maior clareza e, consequentemente, atraírem candidatos mais assertivos para posições de cientistas de dados e, com isso, melhorarem a produtividade e diminuírem a rotatividade de colaboradores.

É marcante, analisando os resultados da revisão, um aspecto bastante desafiador para cientistas de dados: apenas de um forte "background" técnico, com foco em computação e métodos quantitativos, exige-se um conhecimento significativo em negócios e habilidades interpessoais, como por exemplo comunicação, explorada sempre de uma forma bastante abrangente.

Outras duas características relevantes que florescem a partir da revisão são a quantidade e a heterogeneidade das habilidades encontradas. Estas duas características, em conjunto, sugerem que cientistas de dados são profissionais super-qualificados, ou seja, que para desempenhar suas atividades são necessárias muitas habilidades. Dois autores analisados, Baškarada & Koronios (2017), utilizam o termo "unicórnio" em sua pesquisa para designar cientistas de dados, exatamente no mesmo contexto: um conjunto muito grande e muito diferente de habilidades. Mais próximo da realidade, onde unicórneos ainda não se provaram factuais, De Mauro et al. (2017) sugerem que, na verdade, existem grupos ou "famílias" de cientistas de dados, que podem ser organizados de acordo com a homogeneidade de suas habilidades. A análise proposta por De mauro et al., à luz das habilidades encontradas na revisão da literatura, está além do escopo deste artigo.

Finalmente, outra característica que pode ser depreendida da análise de habilidades de cientistas de dados é o fato de que apenas seis artigos - Manieri et al. (2015), Wiktorski et al. (2017), Cao (2017), Ecleo & Galido (2017), Mikalef (2018 e 2019) – mencionam aspectos que estão relacionados à metodologias. Esta análise é importante porque um cientista de dados, sendo um cientista, deveria seguir os protocolos científicos. Da forma como o resultado da análise sugere, o termo cientista de dados está mais próximo de um termo genérico adotado pelo mercado do que de um profissional que, efetivamente, segue o método científico.

6. Referências Bibliográficas

Abidin et al., W. Z., Ismail, N. A., Maarop, N., & Alias, R. A. (2017, August). Skills Sets Towards Becoming Effective Data Scientists. In International Conference on Knowledge Management in Organizations (pp. 97-106). Springer, Cham.

Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., & Schneider, M. V. (2019). A global perspective on evolving bioinformatics and data science training needs. Briefings in Bioinformatics, 20(2), 398-404.

Baškarada, S., & Koronios, A. (2017). Unicorn data scientist: the rarest of breeds. Program.

Baumer, B. S. (2018). Lessons from between the white lines for isolated data scientists. The American Statistician, 72(1), 66-71.

Cao, L. (2017). Data science: a comprehensive overview. ACM Computing Surveys (CSUR), 50(3), 1-42.

Chatfield, A. T., Shlemoon, V. N., Redublado, W., & Rahman, F. (2014). Data scientists as game changers in big data environments.

Choudhury, G. S. (2008). Case study in data curation at Johns Hopkins University. Library Trends, 57(2), 211-220.

Columbus, L. (2018), "Data Scientist Is the Best Job in America According to Glassdoor's 2018 Rankings," available at https://www.forbes.com.

Columbus, L. (2019), "Data Scientist Leads 50 Best Jobs In America For 2019 According To Glassdoor," available at https://www.forbes.com.

Costa, C., & Santos, M. Y. (2017). 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, 37(6), 726-734.

Cleveland, W. S. (2001), “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics,” International Statistical Review, 69, 21–26. doi:10.1111/j.1751-5823.2001.tb00477.x.

Debortoli, S., Müller, O., & vom Brocke, J. (2014). Comparing business intelligence and big data skills. Business & Information Systems Engineering, 6(5), 289-300.

De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2017). Human resources for Big Data professions: A systematic classification of job roles and required skill sets. Information Processing & Management, 54(5), 807-817.

Demchenko, Y., Comminiello, L., & Reali, G. (2019, March). Designing Customisable Data Science Curriculum Using Ontology for Data Science Competences and Body of Knowledge. In Proceedings of the 2019 International Conference on Big Data and Education (pp. 124-128).

Dichev, C., & Dicheva, D. (2017, January). Towards Data Science Literacy. In ICCS (pp. 2151-2160).

Davenport, H & Patil, D. J. (2012) "Data Scientist: The Sexiest Job of the 21st Century." Harvard Business Review 90, no. 10 (October 2012): 70–76.

Ecleo J., Galido A, (2017) Surveying LinkedIn Profiles of Data Scientists: The Case of the Philippines, Procedia Computer Science, Volume 124, Pages 53-60, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2017.12.129.

Fleury, M.T. L. & Fleury, Af (2001). Construindo o conceito de competência. Revista de Administração Contemporânea, 5(spe), 183-196. https://dx.doi.org/10. 1590/ S1415 - 65552001000500010

Gardiner, A., Aasheim, C., Rutner, P., & Williams, S. (2018). Skill requirements in big data: A content analysis of job advertisements. Journal of Computer Information Systems, 58(4), 374-384.

Gibert, K., Horsburgh, J. S., Athanasiadis, I. N., & Holmes, G. (2018). Environmental data science. Environmental Modelling & Software, 106, 4-12.

Harris D. (2013). Kaggle now has 100K data scientists, but what’s a data scientist? GigaOM.com. https://gigaom.com/2013/07/11/kaggle-now-has-100k-data-scientists-but-whats-a-data-scientist/ Recuperado da internet em 31/3/2020.

Hu, H., Luo, Y., Wen, Y., Ong, Y. S., & Zhang, X. (2018). How to find a perfect data scientist: A distance-metric learning approach. IEEE Access, 6, 60380-60395.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele, UK, Keele University, 33(TR/SE-0401), 28.

Knorr, E. M., Riva, G. V. D., & Vakarelov, O. (2018, May). Anatomy of a New Data Science Course in Privacy, Ethics, and Security. In Proceedings of the 23rd Western Canadian Conference on Computing Education (pp. 1-5).

Kotzé, E. (2017, July). A survey of data scientists in South Africa. In Annual Conference of the Southern African Computer Lecturers' Association (pp. 175-191). Springer, Cham.

Kross, S., & Guo, P. J. (2019, May). Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

Laker, D. R., & Powell, J. L. (2011). The differences between hard and soft skills and their relative impact on training transfer. Human Resource Development Quarterly, 22(1), 111–122.

Luna-Reyes, L. F. (2018). The search for the data scientist: creating value from data. ACM SIGCAS Computers and Society, 47(4), 12-16.

Manieri, A., Brewer, S., Riestra, R., Demchenko, Y., Hemmje, M., Wiktorski, T., ... & Frey, J. (2015, November). Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists. In 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom) (pp. 588-593). IEEE.

Markow, W., Braganza, S., Hughes, D., & Miller, S. (2017). The Quant Crunch. How Demand for Data Science Skills Is Disrupting the Job Market. Boston: Burning Glass Technologies.

Marr, B. (2020). Coronavirus: How Artificial Intelligence, Data Science And Technology Is Used To Fight The Pandemic. https://www.forbes.com/sites/bernardmarr/2020/03/13/coronavirus-how-artificial-intelligence-data-science-and-technology-is-used-to-fight-the-pandemic. Recuperado da Internet em 20/04/2020.

Mikalef, P., Giannakos, M. N., Pappas, I. O., & Krogstie, J. (2018). The Human Side of Big Data: Understanding the skills of the data scientist in education and industry. In 2018 IEEE Global Engineering Education Conference (EDUCON) (pp. 503-512). IEEE.

Mikalef, P., & Krogstie, J. (2019, April). Investigating the Data Science Skill Gap: An Empirical Analysis. In 2019 IEEE Global Engineering Education Conference (EDUCON) (pp. 1275-1284). IEEE.

Miller, G. J. (2019, June). The influence of big data competencies, team structures, and data scientists on project success. In 2019 IEEE Technology & Engineering Management Conference (TEMSCON) (pp. 1-8). IEEE.

Meyer, M. A. (2019). Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings. Journal of the American Medical Informatics Association, 26(5), 383-391.

Patil, D.J. 2011. “Building Data Science Teams: Data Science Teams Need People with the Skills and Curiosity. http://radar.oreilly.com/2011/09/building-data-science-teams.html. Recuperado da internet em 18/11/2019.

Pires, F., Barbosa, J., & Leitão, P. (2018, July). Data scientist under the Da. Re perspective: analysis of training offers, skills and challenges. In 2018 IEEE 16th International Conference on Industrial Informatics (INDIN) (pp. 523-528). IEEE.

Salado-Cid, R., Ramírez, A., & Romero, J. R. (2018). On the need of opening the big data landscape to everyone: challenges and new trends. In Digital Marketplaces Unleashed (pp. 675-687). Springer, Berlin, Heidelberg.

Tenório, T., Bittencourt, I. I., Isotani, S., & Silva, A. P. (2016). Does peer assessment in on-line learning environments work? A systematic review of the literature. Computers in Human Behavior, 64, 94–107. https://doi.org/10.1016/j.chb. 2016.06.020

Verma, A., Yurov, K. M., Lane, P. L., & Yurova, Y. V. (2019). An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements. Journal of Education for Business, 94(4), 243-250.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics, 34(2), 77-84.

Wiktorski, T., Demchenko, Y., & Belloum, A. (2017). Model Curricula for Data Science EDISON Data Science Framework. In 2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom) (pp. 369-374). IEEE.

Yan, D., & Davis, G. E. (2019). A First Course in Data Science. Journal of Statistics Education, 27(2), 99-109.