



UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE TECNOLOGIA



Bruno Eduardo Sumeira

O mercado de trabalho da Ciência de Dados

Limeira 2020

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE TECNOLOGIA

Bruno Eduardo Sumeira

O mercado de trabalho da Ciência de Dados

Trabalho de Conclusão de Curso
apresentado como requisito parcial para a
obtenção do título de Bacharel em Sistemas
de Informação à Faculdade de Tecnologia da
Universidade Estadual de Campinas.

Orientador: Prof. Dr. Marcos Augusto Francisco Borges

Limeira 2020

UNIVERSIDADE ESTADUAL DE CAMPINAS

FACULDADE DE TECNOLOGIA

Folha de Aprovação

Trabalho de Monografia do aluno Bruno Eduardo Sumeira apresentado na Faculdade de Tecnologia – FT, da Universidade Estadual de Campinas – UNICAMP, em janeiro de 2021 submetido à aprovação pela Banca Examinadora composta pelos professores:

Prof. Dr. Marcos Augusto Francisco Borges

FT/UNICAMP

Prof. Dr. Plínio Roberto Souza Vilela

FT/UNICAMP

Prof. Gabriel Coutinho Natucci

FT/UNICAMP

AGRADECIMENTOS

Agradeço, primeiramente, à minha mãe, Claudete, e ao meu irmão Thiago por terem incentivado e assegurado minha permanência na vida acadêmica. Se hoje consigo chegar neste momento, muito se deve a vocês.

Agradeço profundamente ao meu orientador, Prof. Dr. Marcos Augusto Francisco Borges, por sua disposição em auxiliar na construção deste trabalho. Suas intervenções foram de grande valia para o desenvolvimento e êxito desta pesquisa.

Agradeço a todos os meus amigos e colegas de curso que ao longo desses cinco anos contribuíram direta e indiretamente com a minha formação profissional. Agradeço, de maneira especial, os amigos Bruno Thomé, Lucas Romani, Victor Suicava e Kaique Siqueira pelos trabalhos produzidos em conjunto, dando sempre mostras de grande comprometimento e responsabilidade. Agradeço à empresa onde desenvolvi atividades de estágio (Honda South América), por ter acrescido competências pessoais e profissionais que hoje mostram-se significativas para a elaboração deste trabalho.

“O sucesso é ir de fracasso em fracasso sem perder entusiasmo.” (Winston Churchill)

RESUMO

A ciência de dados utiliza diferentes tecnologias, metodologias e modelos para capturar, armazenar e processar as informações. Por ser uma área interdisciplinar e em rápida e constante evolução, existem dificuldades em definir qual o conjunto de habilidades que um profissional de ciência de dados (PCD) deve dominar. O objetivo deste trabalho foi apresentar as habilidades essenciais que compõe o perfil do PCD, tal e qual, as contribuições que este pode trazer ao universo corporativo atuando na solução de problemas de informação cada vez mais complexos. A metodologia utilizada foi baseada em pesquisas práticas em âmbito acadêmico (revisão de literatura) e entrevista com público selecionado. Foi desenvolvido um questionário com 9 perguntas que foi respondido por 48 profissionais brasileiros. Conforme a literatura pesquisada, cinco grandes áreas foram consideradas essenciais para um profissional de dados: negócios, matemática, *Big Data*, programação e estatística. Nas respostas dos entrevistados da pesquisa prática, foi possível identificar que um PCD precisa ter um domínio técnico principalmente em Inteligência Artificial, *Big Data*, Estatística, lógica e programação. Desta forma, foi possível encontrar similaridades das habilidades ditas como essenciais na pesquisa bibliográfica. Além de fornecer subsídios e diretrizes gerais a respeito da formação de cientistas de dados para uma reflexão quanto ao público-alvo interessado em ingressar na área de *Data Science*, esta pesquisa poderá contribuir na elaboração de trabalhos futuros que desejem aprofundar mais quanto ao conjunto de habilidades necessária para o PCD, além de incentivar aspirantes a ingressar nessa importante área da tecnologia que transforma a Tecnologia da Informação em uma área estratégica no mercado.

Palavras-chave: Profissional de Ciência de Dados, Habilidades essenciais do cientista de dados, Profissões em alta na Tecnologia da Informação (TI).

ABSTRACT

Data science uses different technologies, methodologies and models to capture, store and process information. It is an interdisciplinary area in fast and constant evolution, so there are difficulties in defining which set of skills a data scientist must master. The objective of this work was to present the essential skills that make up the profile of a data scientist, such as, to present the contributions that this can bring to the corporate universe acting in the solution of increasingly complex information problems. The methodology used was based on practical research available in the academic field (literature review) and interview with selected audience. A questionnaire with 9 questions was developed and 48 Brazilian professionals answered it. It was found in the researched literature= five major areas considered essential for a data scientist: business, mathematics, big data, programming and statistics. In the answers of the interviewees of the practical research, it was possible to identify that a data scientist needs to have a technical domain mainly in Artificial Intelligence, Big Data, Statistics, logic and programming. Thus, it was possible to find similarities in the skills said to be essential in bibliographic research and in interviews. In addition to providing subsidies and general guidelines regarding the training of data scientists for reflection on the target audience interested in entering the field of Data Science, this work aims to contribute in the elaboration of future works about the skill set necessary for the data scientist, in addition to encouraging aspirants to enter this important area of technology that makes Information Technology a strategic area in the market.

Keywords: Data Science Professional, Essential Skills of the Data Scientist, High professions in Information Technology (IT)

LISTA DE FIGURAS

Figura 1 - Analisando os insights comerciais específicos das empresas.....	16
Figura 2 - Popularidade das buscas por "Data Science"	17
Figura 3 - Diagrama de Veen da ciência de dados de Drew Conway(2010)	21
Figura 4 - Proeficiência em habilidades de ciência de dados.....	34
Figura 5 - Área de atuação dos participantes.....	38
Figura 6 - Habilidades técnicas essenciais para o PCD.....	39
Figura 7- Habilidades interpessoais essenciais para o PCD.....	39

LISTA DE QUADROS

Quadro 1. Habilidades do PCD.....	22
Quadro 2.Habilidades em sistemas computacionais.....	36
Quadro 3. Habilidades Técnicas computacional.....	36
Quadro 4.Habilidades de Negócio.....	37

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

BI – *Business Intelligence*

CD – *Ciência de Dados*

IA – *Inteligência Artificial*

JISCr - *Joint Information Systems Committee*

MGI – *McKinsey Global Institute*

NoSQL – *Not Only Structured Query Language*

PCD – *Profissional de Ciência de Dados*

SAD – *Sistema de apoio a decisão*

SQL – *Structured Query Language*

TI – *Tecnologia da Informação*

SUMÁRIO

1.	INTRODUÇÃO	11
1.1	Estrutura desta monografia	12
2.	FUNDAMENTAÇÃO TEÓRICA	13
2.1	A informação	13
2.2	Ciência dos dados	15
2.3	Big Data	19
2.4	Inteligência Artificial – IA	22
2.5	O Profissional de Ciência de Dados (PCD) ou Cientista de dados	25
2.5.1	Habilidades	26
3.	METODOLOGIA	29
3.1.	Pesquisa bibliográfica	30
3.2.	Entrevista com público selecionado	30
3.3.	Discussão	31
4.	ANÁLISE DOS RESULTADOS	32
4.1	Pesquisa bibliográfica	32
4.2	Entrevista com público selecionado	36
4.3	Discussão	40
5.	CONCLUSÕES	42
6.	REFERÊNCIAS	44
	APÊNDICE 1. Questionário referente pesquisa com especialistas sobre o conjunto de habilidades do PCD	50

1. INTRODUÇÃO

A ciência de dados (CD) é uma ciência que utiliza várias tecnologias, metodologias e modelos com a finalidade de capturar, armazenar e processar as informações. Desta forma, um profissional desta área poderá solucionar problemas de informações cada vez mais complexos e dinâmicos (Fundação Instituto de Administração, 2019).

Por ser uma área interdisciplinar e em rápida e constante evolução, ainda restam dúvidas e indagações sobre quais são as habilidades essenciais que um profissional deve ter. O perfil do profissional desta área ainda não é bem definido. Devido ao caráter multidisciplinar da área de dados, são necessários recursos humanos com diferentes habilidades para compor as equipes, tais como engenheiros de dados, cientistas de dados, estatísticos e engenheiros de software.

O mercado, na prática, caminha para uma convergência do papel de cada perfil de profissional no âmbito de suas atividades. O estatístico atua diretamente criando métodos e aplicando conceitos para viabilizar a extração de dados com correção e significância. O engenheiro de dados atua na criação dos pipelines de produção, ou seja, organizando e estruturando os dados de forma que torne o processamento mais rápido, com enfoque na camada de arquitetura. O cientista de dados, que é o principal objeto deste estudo, deve atuar diretamente nos dados, analisando-os e obtendo insights.

Atualmente, é crescente a oferta de cursos para tal fim, com a qual, tanto a academia, quanto o mercado, começam a definir o que realmente seria a formação adequada de um PCD (Lopes, 2019).

A comunidade europeia e da América do Norte tem avançado significativamente na capacitação e qualificação dos cientistas de dados em nível de pós-graduação, por meio de treinamento em informática e estatística e de conteúdo direcionado a essa especialidade de formação (Swan e Brown, 2008; Finzer, 2013). Além disso, diversos autores passaram a utilizar a expressão Moderno Profissional da Informação para se referir ao novo profissional de dados que emerge em um ambiente altamente dinâmico, em meio ao constante avanço das tecnologias de informação e comunicação (Santos, 1996; Pereira e Campos, 2014; Santa Anna, 2017).

O objetivo deste trabalho foi compreender quais conceitos técnicos e quais as principais habilidades são esperadas para o profissional de ciência de dados (PCD). Para

tal, uma abordagem analítica de carácter exploratório foi realizada, através de pesquisas de mercado e revisão da literatura. Com os resultados, foi possível identificar os conceitos essenciais para atuar na área. Essa informação é relevante, visto que é uma profissão extremamente promissora para o mercado de trabalho.

Além disso, foi dado um enfoque especial aos conceitos relacionados a ciência dos dados encontrados na literatura: Informação, *Data Science*, *Big Data* e Inteligência Artificial (IA).

1.1 Estrutura desta monografia

O Capítulo 2 apresenta a fundamentação teórica com o objetivo de apresentar os principais conceitos que abrangem a CD utilizados durante o desenvolvimento deste trabalho. O Capítulo 3 apresenta a metodologia utilizada nesta monografia, com intuito de estabelecer as habilidades e competências de um PCD, tanto na visão da bibliografia, como na de profissionais de mercado. No Capítulo 4 são apresentados os resultados obtidos através da análise teórica e métricas levantadas na pesquisa, resultando na intersecção entre a teoria e os resultados práticos com o objetivo de elencar os principais conceitos para o PCD. Por fim, no Capítulo 5, apresentam-se as conclusões do trabalho destacando os principais resultados obtidos.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os principais conceitos e temas pesquisados visando embasar teoricamente este trabalho. O capítulo apresenta o conceito de informação, de CD, bem como, os conceitos computacionais que permeiam e impulsionam a CD: *Big Data* e IA. Por fim, a seção apresenta o PCD e seu conjunto de habilidades esperadas no mercado de trabalho.

2.1 A informação

Antes de definir o conceito de informação ou citar para qual finalidade a mesma é utilizada, é necessário saber o conceito de dados e dados relacionados.

Diversas definições de dados podem ser encontradas na literatura. De acordo com Rezende (2015), “*O termo dados está associado com fatos. Cada dado é um pequeno fato que não significa muito por si só, ou seja, dados são os registros soltos, aleatórios, sem quaisquer análises*”. Já Pinheiro (2003) define que “*Dados são observações documentadas ou resultados da medição. Os dados podem ser obtidos pela percepção através dos sentidos (por exemplo observação) ou pela execução de um processo de medição*”.

Os dados podem ser classificados em dois grandes grupos. O primeiro, denominado de dados estruturados, são altamente organizados e formatados de forma a ser facilmente pesquisáveis em bancos de dados relacionais, formato de banco que modela os dados de uma forma que eles sejam percebidos pelo usuário finais como tabelas. Sistemas tradicionais de empresas, tais como cadastros contendo nomes, datas, endereços, números de cartão de crédito, informações sobre ações e geolocalização, são excelentes exemplos de dados estruturados (Monteiro, 2019). O segundo grupo é o de dados não estruturados, que não possuem um formato padronizado para leitura. Eles normalmente são classificados como dados qualitativos, ou seja, que não podem ser processados e analisados usando ferramentas e métodos convencionais. Exemplos de dados não estruturados incluem vídeos, textos, áudios, atividades móveis, atividades de mídia social, imagens de satélite (Monteiro, 2019).

O conceito de dado relacionado se refere a um conjunto de dados que está diretamente associado a um tópico específico, como, por exemplo, o grupo universidade,

cujos dados relacionados podem ser a quantidade de alunos, quantidade de alunos por sala, quantidade de salas, etc. Portanto, *“dado relacionado ainda é um conjunto de dado solto, sem quaisquer análises”* (Rezende, 2015), mas com potencial para ser analisado ou correlacionado.

A Informação é um aglomerado de dados relacionados, cuja definição e conceito são extremamente abrangentes na literatura, com muita variação entre autores, por ser objeto de estudo das diferentes áreas da ciência. *“O conceito pode ser utilizado para determinar um padrão de comunicação entre um emissor e um receptor, uma medida de organização física; até mesmo pode ser definida como um objeto ou coisa”* (Rolim, 2018).

Capurro e Hjørland (2007), descrevem a informação como uma ideia subjetiva, ou seja, dependendo da interpretação de um agente cognitivo. Para Setzer (2015), a informação é uma abstração, representando algo significativo para alguém através de textos, imagens, sons ou animação. Buckland (1991) descreveu a informação por meio de três principais conceitos: a informação como processo, como conhecimento e como coisa. A informação como processo é vista a partir da perspectiva de quando um indivíduo é informado sobre algo, aquilo conhecido por ele é modificado: nesse sentido, a informação é identificada como o ato de informar. A informação como conhecimento é caracterizada pela comunicação do conhecimento, como o detalhamento da informação, como uma notícia, de forma a reduzir incertezas. Finalmente, a informação como coisa relaciona-se com os objetos e documentos considerados como informação, relacionados como sendo informativos, dotados de conhecimento comunicável (Buckland, 1991).

Tudo que é dotado de capacidade de transmitir uma ideia, um conceito pode ser considerado informação, estando ela em algum tipo de registro ou sendo de fonte oral. A informação pode ser utilizada em diferentes contextos, tanto em um cenário de imprescindível tomada de decisão, em ambiente gerencial, empresarial e até mesmo governamental, como em um simples comunicado informado no dia a dia das pessoas. A informação pode ser apresentada de forma impressa ou escrita, armazenada sistemicamente; transmitida pelo correio ou por meios eletrônicos (por exemplo: e-mails, redes sociais), apresentada em emissoras de televisão, rádios, falada em conversas (Rolim, 2018).

No atual cenário mundial, grandes volumes de dados uma extensa quantidade de informação é disponibilizada diariamente na Internet. Essas informações, se analisadas corretamente, se tornam uma vantagem competitiva, que as empresas podem utilizar para

traçar estratégias e moldar cenários promissores em benefícios da empresa.

2.2 Ciência dos dados

Diante da evolução e acessibilidade tecnológica, o mundo tem produzido uma infinidade de dados. Devido ao aumento no número de pessoas conectadas constantemente, seja por um aparelho *smartphone*, *smartwatch* ou computador, a quantidade de dados que surgem a cada minuto no mundo inteiro é gigantesca.

Todos estão rodeados de informações de variadas origens. Atividades do cotidiano das pessoas, que antes não poderiam ser monitoradas devido a limitação tecnológica, passaram a ser fonte importante para a obtenção de dados e, consequentemente, de informação. O volume de dados criado nos últimos anos é maior do que a quantidade produzida em toda a história. Hoje, mais da metade da população mundial utiliza a Internet, totalizando aproximadamente 4,4 bilhões de pessoas (Kemp, 2019).

Diante de um cenário cada vez mais dinâmico e competitivo, as empresas também tem motivos para se adaptar, tanto para melhorar a eficácia produtiva quanto na captação de novos clientes. Isso tem exigido que as empresas busquem ser mais produtivas para garantir entregas rápidas, frequentes e com índices elevado de qualidade (Marino, 2006).

Os processos de negócios convivem com um alto dinamismo no novo cenário empresarial, incentivando as empresas a aplicarem e executarem novas práticas. No presente cenário, a Tecnologia da Informação (TI) passa a ser uma área estratégica, estando presente dentro das empresas tanto no processo de desenvolvimento de um produto, quanto na gestão de negócios (Scheer et al., 2004).

Nesse contexto, surge uma grande demanda para a TI que consiste na transformação dessa grande quantidade de dados estruturados e não estruturados, em informação, de forma a possibilitar o uso eficiente desses dados para gerar vantagens competitivas e melhores tomadas de decisões por parte da gestão das companhias.

Por se tratar de um tema comercialmente impactante, essa informação tem potencial de gerar uma revolução no mundo da inteligência do negócio, despertando o interesse da comunidade da tecnologia. Assim, o conceito de CD tem ganhado espaço significativo. De acordo com Cervantes (2016), essa ciência orientada a uma grande quantidade de dados se vale do potencial de robustas *cibeeinfraestruturas* de informação

e comunicação que preveem o uso de tecnologias de *grids*, que consistem em um modelo computacional capaz de alcançar uma alta taxa de processamento dividindo as tarefas entre diversas máquinas, além de padrões que possibilitam a interoperabilidade e a interligação de dados (*linked data*). Cervantes (2016) ainda conclui que esse novo patamar atingido no âmbito tecnológico suportado pelas robustas ciberinfraestruturas possibilita sustentar as diferentes fases do ciclo de vida dos dados disponíveis na web e em repositórios de dados digitais. O principal objetivo da CD, consiste em atribuir sentido e extrair *insights* (momento em que repentinamente se tem uma ideia) de dados aplicáveis a diferentes domínios e contextos, para a resolução de problemas práticos e reais, além de ter um importante e direto papel na tomada de decisões dos gestores das companhias.

De acordo com Singh (2019, p. 01) a CD pode ser definida como:

Um conjunto multidisciplinar de inferência de dados, desenvolvimento de algoritmos e tecnologia para resolver problemas analiticamente complexos. No centro estão os dados: um grande número de informações brutas, transmitidas e armazenadas em data warehouses(ou armazém de dados, ou depósito de dados) corporativos. [...] A ciência de dados é basicamente o uso desses dados de maneiras criativas para gerar valor aos negócios”.

Provosto e Fawcett (2016) definem a CD como: “*Data Science* envolve princípios, processos e técnicas para compreender fenômenos por meio da análise (automatizada) de dados”.

Além disso, a CD pode ser definida como inevitável e é uma base fundamental de concorrência e crescimento, no qual tomar decisões baseadas em dados é de valor crítico nas empresas de todos os setores (McKinsey Global Institute, 2019).

A CD, portanto, abrange o processo de ponta a ponta, desde a coleta inicial do dado até a transformação em inteligência para o negócio. Nesse sentido, Kdnuggets (2017) a caracteriza como uma disciplina multifacetada:

A ciência de dados é uma disciplina multifacetada, que abrange aprendizado de máquina e outros processos analíticos, estatísticas e ramos relacionados da matemática. Cada vez mais se utiliza da computação científica de alto desempenho, tudo isso para extrair informações dos dados e usar essas informações encontradas para contar histórias.

A evolução da tecnologia visa aprimorar as formas para coletar, armazenar, recuperar e transformar os dados, de forma a realizar o conceito de tratativa dos dados, expressão que consiste em coletar e relacionar dados. Atualmente, o volume de dados supera em muito a capacidade humana de lidar com estas novas perspectivas de

quantidade, volume, variedade e principalmente velocidade de acesso a dados, cujo principal benefício é criar informação relevante para a gestão. Na **Figura 1** podemos observar o processo desde a captação do dado até os seus benefícios para o negócio (ISACA, 2013).

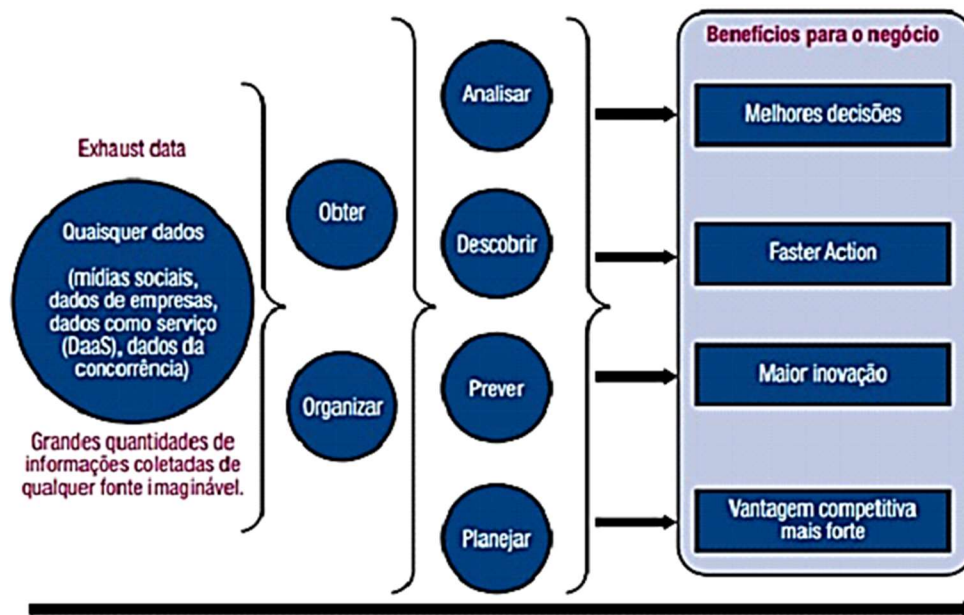


Figura 1. Fluxograma do processamento de dados e seus benefícios para os negócios.

Fonte: ISACA (2013).

Devido as diversas vantagens que a ciência dos dados pode trazer para diferentes empresas, a evolução das buscas pelo termo “*Data Science*” no Brasil e no mundo, de outubro de 2012 até julho de 2018 teve um aumento significativo (Figura 2).

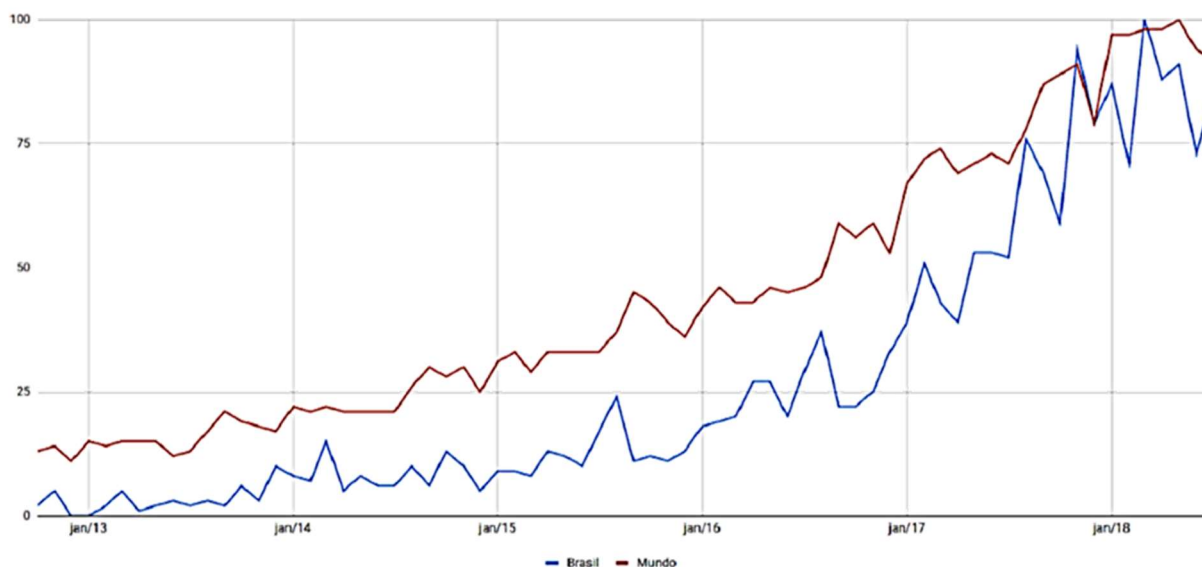


Figura 2. Popularidade das buscas por "*Data Science*" no período de 2012 a 2018.
Fonte: Google Imagem (2018).

Na Figura 2 podemos observar que o gráfico corrobora com o significativo aumento de buscas pelo tema, sendo 100 o momento de maior popularidade, 50 média popularidade e 0 nenhuma popularidade, evidenciando o crescimento cada vez maior da popularidade da área de CD, estando atualmente em seu pico comparado aos últimos anos. É possível constatar que no Brasil há um certo atraso em relação às buscas gerais em todo o mundo nos anos de 2013 a 2017, mas a partir de 2018 o país se equiparou ao restante do mundo. Além disso, é possível observar que as quedas sazonais no mundo são mais marcantes em dezembro, mas no Brasil não há um padrão bem definido da demanda.

Fica evidente o avanço do interesse por *Data Science* em todo o globo, visto que a aplicabilidade é ampla o suficiente para conseguir amparar a maioria dos ramos do mercado.

Para utilizar os conceitos de CD, basta ter uma quantidade de dados sendo gerada e coletada. Quanto maior for essa quantidade de dados, mais precisas poderão ser as análises. Devido a isso, a Associação Brasileira de Ciência de Dados (ABRACD, 2019) elencou diferentes áreas na qual a CD pode ser aplicada:

- Indústria/Fábrica: por meio da CD é possível otimizar os recursos da

produção, gerar maior eficiência produtiva, proporcionar visão analítica do processo além de controlar a qualidade do processo (ABRACD, 2019).

- Indústria Financeira: auxilia diretamente na detecção de risco de fraudes, na análise de crédito dos clientes, no gerenciamento de riscos, análise de novas oportunidades de mercado e previsões de inadimplência (ABRACD, 2019).
- TI: a CD permite o monitoramento de informações praticamente em tempo real, aumento da segurança da informação, detecção de invasões e maior previsibilidade de erros sistêmicos (ABRACD, 2019).
- Marketing: permite a detecção de padrões comportamentais em nichos de mercado, previsões de novos cenários e tendências de mercado (ABRACD, 2019).
- Saúde: possibilita a análise de imagem e detecção de tumores. Auxilia também, detectando nível avançado de personalização do tratamento por meio de pesquisas em genética e genômica. O objetivo é entender o impacto do DNA em nossa saúde e encontrar conexões biológicas individuais entre genética, doenças e resposta a medicamentos (ABRACD, 2019).

Portanto, fica evidenciado o impacto positivo que a CD está gerando no mercado, alterando e otimizando processos. Neste contexto, a próxima seção (2.3), apresentará o que a literatura relata sobre a relação da CD com o conceito de grandes massas de dados (Big Data).

2.3 Big Data

Big Data é um conceito que está diretamente associado à CD, mas não é nem uma área do conhecimento e nem uma profissão. Segundo Gantz (2012), "*Big Data não é uma coisa, mas uma atividade dinâmica que atravessa muitas fronteiras de TI*".

Mougalas (2005) usou o termo *Big Data* pela primeira vez para referir-se a um grande conjunto de dados que é quase impossível de gerenciar e processar usando as ferramentas tradicionais de *business intelligence (BI)*.

De acordo com o Gartner (2014), *Big Data* é uma expressão que pode ser definida como.

Grande volume de dados, gerados em alta velocidade e variedade, no qual necessitam de formas inovadoras e econômicas para processá-los, organizá-los e armazená-los, a fim de se permitir melhor compreensão para a tomada de decisão e automação de processos.

Big Data está relacionado a uma enorme quantidade de registros gerada pelo aumento significativo no uso e avanço da tecnologia, gerando assim uma grande quantidade de matéria prima para potenciais análises (Gartner, 2014).

O conceito prático de *Big Data* funciona como um conjunto de metodologias utilizadas para a captura, armazenamento e processamento de todo e qualquer tipo de dado, podendo ser processados dados estruturados e não estruturados em um sistema escalável, isto é, em que o número de máquinas aumenta conforme a necessidade (Crawford, 2014).

Big Data está diretamente relacionado com informação rápida. Nesse sentido, os modernos sistemas de computação fornecem a velocidade, potência e flexibilidade necessárias para acessar rapidamente grandes quantidades e tipos de dados. Juntamente com o acesso confiável, as empresas também precisam de métodos para integrar, garantir a qualidade, fornecer controle e armazenamento dos dados, preparando-os para análise. Alguns dados podem ser armazenados localmente em um *data warehouse*, que pode ser definido como um repositório de dados voltado à tomada de decisões gerenciais que armazena grandes quantidades de dados centralizados, integrados, organizados por assuntos, históricos, resumidos e não voláteis (Kimball, 1996).

No modelo convencional de armazenamento de dados e tomada de decisão, a organização filtra dados dos seus vários sistemas e, após criar um *data warehouse*, constrói consultas (*queries*) de forma exploratória afim de tomar decisões. Na prática faz-se garimpagem, forma de expressar o processo de descoberta de conhecimento em banco de dados, em uma base de dados estática, que não reflete o momento, mas sim o contexto de horas, dias ou mesmo semanas atrás. Mas também existem opções flexíveis e de baixo custo para armazenar e manipular *Big Data* por meio de corrente contínua de dados (*stream computing*): essa mineração ocorre em tempo real, com uma corrente contínua de dados (*streaming data*) atravessando um conjunto de consultas, o que pode ser considerado um novo paradigma que pode ser usado em soluções em nuvem, lagos de dados e *Hadoop* (Cetax, 2016).

O *Hadoop* é uma plataforma de software de código aberto, diretamente relacionada com o processamento distribuído de grande quantidade de dados, que faz uso

de *clusters* de computadores com hardware *commodity*. O *Hadoop* vem sendo uma ferramenta fundamental para tornar ágil o processo de análise de dados (Cetax, 2016).

O *Hadoop*, possibilita processar grandes quantidades de dados para disponibilizá-los para análise rapidamente agregando valor competitivo para o negócio, diferente da análise de dados tradicional, que em um contexto de grande quantidade de dados como nos dias atuais, acaba sendo diretamente impactada por morosos e persistentes processos de *turn-around* (consiste no tempo total transcorrido desde o momento em que o software começa a análise e o instante em que termina sua execução) devido à capacidade de organização e processamento dessa grande massa desorganizada de dados, ou seja, para gerar o resultado de uma análise no viés tradicional pode levar horas/dias até que seja finalizado.

Por fim, para que seja possível transformar o dado em informação existem tecnologias de alto desempenho, como computação em grade ou análise na memória. As organizações podem optar por usar todos os seus grandes dados para análises. Outra abordagem é determinar antecipadamente quais dados são relevantes antes de analisá-los. De qualquer forma, a análise de *Big Data* é uma forma das empresas obter valor e *insights* a partir dos dados.

Embora *Big Data* envolva um grande e variado conjunto de dados, por si só não significa nada para o negócio, pois não atua na interpretação dos resultados do processamento de dados para apoiar a tomada de decisões.

O jornal *The Economist* entrevistou executivos de multinacionais em sua primeira pesquisa sobre o tema *Big Data* (EIU, 2011). Na entrevista, metade dos executivos acreditavam que os dados eram um importante ativo para suas organizações e 10% afirmaram que a disponibilidade de dados havia alterado completamente a forma como suas organizações planejavam e operavam. Além disso, a pesquisa apurou também certas dificuldades das organizações em alguns aspectos básicos relacionados à administração de dados e sua efetiva exploração, devido à complexidade de se organizar, analisar e compreender essa grande quantidade de dados.

Por fim, conclui-se que o objeto gerador de interesse para as empresas não é a quantidade de dados em si, mas sim a exploração e transformação do dado em *insights*, por meio do PCD. O objetivo empresarial é coletar dados de qualquer fonte e analisá-los para encontrar respostas que possibilitem reduzir custos e tempo empreendido, desenvolver novos produtos e gerar ofertas otimizadas, além de apoiar na tomada de decisão inteligente.

Em um nível estratégico empresarial, a associação do PCD com *Big Data* tem como finalidade construir uma visão capaz de supervisionar e melhorar a maneira de armazenar, gerenciar, compartilhar e usar dados dentro e fora da organização para garantir vantagens competitivas e maior lucratividade.

Conclui-se então que *Big Data* fomenta e ampara o avanço da TI. Dados confiáveis e bem gerenciados levam a análises e decisões confiáveis. Para se manter competitivas, as empresas precisam aproveitar todo o valor da *Big Data* e operar de maneira orientada a dados, tomando decisões com base nas evidências apresentadas pela *Big Data*, e não em suposições.

Os benefícios de ser orientado a dados são evidentes. As organizações orientadas a dados têm melhor desempenho, são operacionalmente mais previsíveis e mais lucrativas. Um estudo realizado pelo Google em parceria com o Boston Consulting Group (BCG) em 2018, comprovou que organizações orientadas a dados obtiveram 20% de aumento em receita comparada com as organizações tradicionais.

2.4 Inteligência Artificial – IA

O conceito de IA surgiu na década de 50, mesma época em que Turing propôs seu teste que verificava se uma máquina tinha capacidade de exibir comportamento inteligente equivalente a um ser humano. O nome de IA surgiu a primeira vez em 1956 durante uma conferência de verão em Hanover, no estado de New Hampshire nos Estados Unidos da América (Russel, 2004).

As primeiras pesquisas de IA nos anos 1950 exploraram temas como a resolução de problemas e métodos simbólicos. Na década de 1960, o Departamento de Defesa dos EUA se interessou por este tipo de tecnologia e começou a treinar computadores para imitar o raciocínio humano básico. Por exemplo, a *Defense Advanced Research Projects Agency* (DARPA) completou um projeto de mapeamento de ruas nos anos 1970. E a DARPA criou assistentes pessoais inteligentes em 2003, muito tempo antes de Siri, Alexa ou Cortana serem conhecidos (SAS Institute, 2019).

Devido a complexidade de se obter uma definição formal, surgiram algumas propostas de definições operacionais para a IA, dentre as quais se destaca a de Michie e Meltzer (1969): “...uma máquina é inteligente se ela é capaz de solucionar uma classe de

problemas que requerem inteligência para serem solucionados por seres humanos". Nesse sentido, a IA pode ser associada com a capacidade de um dispositivo computacional replicar algumas habilidades cognitivas e assim, automatizar atividades que associamos ao pensamento humano, como tomada de decisão, resolução de problemas e aprendizagem (Russell e Norvig, 1995). A IA integra várias áreas da ciência, como da computação, estatística, filosofia, psicologia e matemática.

Com os esforços analíticos avançando com a CD, alimentados pela gama de dados fornecidos pelo Big Data, a IA ganha notoriedade no mercado. Esse contexto vem ocasionando mudanças tanto na tecnologia quanto na maneira como nos relacionamos com a mesma.

De forma geral, a IA provê recursos tecnológicos capazes de resolver problemas complexos, pelo fato de possibilitar que as máquinas aprendam com experiências e se ajustem a novas entradas de dados. A IA atua diretamente nos sistemas de apoio à decisão e sistemas inteligentes de pesquisa que podem ser projetados para complementar e expandir as capacidades humanas.

Os sistemas inteligentes, por meio da IA, são capazes de tarefas frequentes, volumosas e programadas de modo confiável e sem fadiga. Para este tipo de automação, a interferência humana ainda é parcialmente necessária, principalmente na configuração do sistema e para fazer os questionamentos certos.

Muitos pesquisadores de tecnologia defendem que a IA é uma tecnologia chave para o software do futuro. As pesquisas em IA estão relacionadas com áreas de aplicação que envolvem o raciocínio humano, com o intuito de chegar de imitar o mais próximo possível esse raciocínio. As áreas de aplicação que geralmente são incluídas e ou associadas nas definições de IA incluem, entre outras:

- **Sistemas Especialistas ou Sistemas Baseados em Conhecimento:** segundo FEIGENBAUM (1977), são sistemas que solucionam problemas que, normalmente, apenas pessoas especialistas conseguem resolver;
- **Sistemas Inteligentes/Aprendizagem:** “... são plataformas tecnológicas que executam diversas funções que aproximam a capacidade racional do ser humano para solucionar problemas.” (Instituto Avaliação, s.d.);
- **Compreensão/Tradução de Linguagem Natural:** “Consiste no desenvolvimento de modelos computacionais para a realização de tarefas que dependem de informações expressas em alguma linguagem natural”(Pereira, 2019);
- **Machining Learning (Aprendizado de Máquina):** “... é um campo de estudo que dá a computadores a capacidade de aprender sem ser programado de forma explícita.” (Samuel, 1959);

- **Deep Learning(Aprendizagem Profunda):** *é uma subárea do Aprendizado de Máquina, que utiliza um conjunto de algoritmos buscando aprender diferentes níveis de abstração, utilizando várias camadas de processamentos não lineares* (Deng; Yu, 2014; Hochuli, 2016).

O campo de IA tem como objetivo o contínuo aumento da "inteligência" do computador, pesquisando, para isto, também os fenômenos da inteligência natural. Por esse motivo, a inteligência artificial não é comercializada como uma aplicação individual, os produtos que são aprimorados com funcionalidades inteligentes. Para este fim, IA é definida aqui como sendo uma coleção de técnicas suportadas por computador emulando algumas capacidades dos seres humanos. Esta coleção inclui:

- Resolução de problemas;
- Compreensão de Linguagem Natural;
- Visão e Robótica;
- Sistemas Especialistas e Aquisição de Conhecimento;
- Metodologias de Representação de Conhecimento.

A Automação, plataformas de conversa (Siri, Alexa), robôs e aparelhos inteligentes podem ser combinados com grandes quantidades de dados para aprimorar muitas tecnologias para o cotidiano das pessoas, tais como sistemas inteligentes de segurança, cidades inteligentes à análise de investimentos.

Por meio de algoritmos de aprendizagem progressiva torna-se possível que se encontrem estruturas e regularidades nos dados para que o algoritmo por si só se torne inteligente, podendo classificar os dados. Então, assim como o algoritmo pode ensinar a si mesmo a jogar xadrez, ele pode ensinar a si mesmo quais produtos recomendar em seguida. E os modelos se adaptam quando recebem mais dados.

Existem diversos motivos para aplicar a inteligência artificial no mercado de trabalho. Entre as vantagens, estão a maior eficiência, a possibilidade de automatizar processos e, conseqüentemente, o aumento na velocidade de conclusão de tarefas. Indiretamente, esse uso pode estar associado à redução de custos operacionais e a eficiência do trabalho.

A IA pode estar associada aos conceitos de *Data Science* e *Big Data*, pois é capaz de armazenar e processar as informações que recebem a partir da experiência humana (leitura, preferência por calçados, viagens, situações de crise etc.) e, com base nesse processamento, pode sugerir soluções por conta própria. A IA, assim como a CD, é alimentada pelos dados gerados pelo *Big Data* de forma a processar esses dados e gerar

inteligência por meio deles

2.5 O Profissional de Ciência de Dados (PCD) ou Cientista de dados

Nos últimos anos surgiu a necessidade de um profissional responsável por trazer soluções para os problemas oriundos da enorme produção de dados, o PCD. O PCD é responsável por desenvolver produtos e serviços a partir desses dados, trata-se de um perfil de profissional abrangente e multidisciplinar, na prática envolve uma diversidade de cargos, desde o cientista de dados, engenheiros de dados, estatístico ao engenheiro de software.

O PCD é um profissional que por meio de treinamentos, curiosidade e pesquisa consegue efetuar análises que geram valor. O trabalho do PCD é realizar descobertas enquanto analisa uma grande massa de dados. Além disso, o profissional também pode estruturar grandes quantidades de dados que não possuem forma, tornando possível analisá-los (Davenport e Patil, 2012).

Os PCDs são capazes de encontrar ricas fontes de dados, conectando-as a outras fontes que estão incompletas, trabalhando em conjunto para alcançar os objetivos previamente traçados (Davenport e Patil, 2012). A função do profissional que trabalha com dados começou a se difundir nas organizações no final dos anos 2000, principalmente na região da Baía de São Francisco, nos Estados Unidos, onde as empresas se voltaram para a exploração de dados provenientes da Internet e das mídias sociais. As empresas acumulavam grandes volumes de dados, para os quais começou-se a desenvolver uma enorme variedade de novas tecnologias para armazená-los, processá-los e analisa-los. Muitas dessas tecnologias foram inventadas por cientistas de dados (Davenport, 2014).

Miller (2013) corrobora com a explicação de Davenport e Patil (2012) sobre o perfil do profissional, afirmando que “*esses cientistas são os mágicos da era da big data*”... Para o autor, 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 (Miller, 2013).

Nesse contexto, o PCD emerge como sendo um profissional altamente requisitado pelas grandes companhias que tem como principal missão e objetivo trazer clareza, de forma a tirar *insights* que irão gerar vantagens competitivas ao negócio, sobre o cenário

de grande produção e acúmulo de dados. Além disso, espera-se desse especialista um conjunto de habilidades que são essenciais no desempenho do papel para o qual é destinado, conforme será detalhado no tópico das habilidades esperada para o PCD.

2.5.1 Habilidades

O mercado de trabalho para o PCD tem como objetivo encontrar um especialista que esteja munido de um conjunto de habilidades tidas como fundamentais. Para atuar na área é exigida uma multidisciplinaridade, devido a isso, existe um elevado grau de dificuldade para chegar a um consenso quanto à definição deste profissional (Davenport, 2014).

O PCD pode ser definido como aquele que trabalha in loco onde pesquisas sobre dados são realizadas, ou em estreita colaboração com os pesquisadores ou grupos de cientistas em centros de dados, e que está envolvido na investigação criativa e de análise de dados, oferecendo soluções tecnológicas para o manuseio e uso de dados digitais (Swan e Brown, 2008).

A literatura tem desenvolvido diversos estudos para buscar encontrar o perfil desse especialista. Nos artigos consultados, há uma convergência ao enfatizar que os PCDs devem apresentar domínio estatístico e computacional para a programação e uso de sistemas capazes de processar grandes volumes de dados (Chatfield et al., 2014); e ter capacidade de explorar a inteligibilidade em dados que a princípio são desestruturados e sem sentido (Van Der Aalst, 2014).

Conaway (2010) propõe que a CD deve constituir-se como a congruência entre domínios computacionais, habilidades de matemática e estatística, e especialidade na área de aplicação de dados, conforme evidência o diagrama de Venn apresentado na Figura 3.

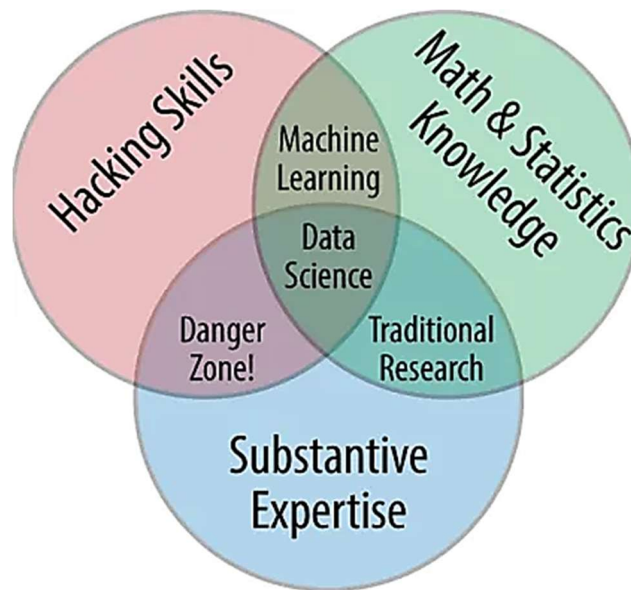


Figura 3. Diagrama de Veen da ciência de dados . Fonte: Drew Conway (2010).

Stanton et al. (2012) propõe que a formação do PCD é fortemente interdisciplinar. Nesse aspecto, os autores complementam que os PCD são responsáveis pela identificação, coleta, tratamento, transformação, análise, visualização e curadoria de grandes conjuntos de dados heterogêneos estruturados ou não estruturados.

Com o avanço nas pesquisas, passa a ser possível estabelecer um conjunto bastante abrangente e ainda pouco detalhado de disciplinas essenciais para o PCD. Davenport (2014) segmentou e detalhou quais seriam as habilidades necessárias para esse profissional, sem entrar na especialização do conhecimento. O profissional de dados é “*um hacker, um cientista, um analista quantitativo, um conselheiro de confiança e um expert em negócios*”. No quadro 1 é possível identificar que um PCD deve possuir um conjunto de características e competências que abarcam desde o entendimento computacional (programação e arquiteturas), princípios analíticos da estatística, essencial para as atividades de coleta, ou seja, mineração e tratamento dos dados e, bem como, conhecer os conceitos fundamentais acerca da gestão de negócios e liderança. a CD envolve muito mais do que apenas algoritmos de mineração de dados. Os cientistas de dados bem-sucedidos devem ser capazes de visualizar os problemas de negócios de uma perspectiva de dados (Provost e Fawcett, 2016).

Quadro 1. Habilidades do PCD.

Hacker
• Capacidade de codificar; • Domínio de arquiteturas tecnológicas de <i>big data</i>.
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: Rolim, Mesaque Vidal (2018)

Fica evidente que um PCD deve possuir conhecimentos interdisciplinares de diversas áreas da ciência (Davenport, 2014). Davenport (2014) relata que, apesar de muitos cientistas de dados possuírem formação acadêmica e títulos, a academia nem sempre é o melhor local para desenvolver as habilidades necessárias na atuação como PCD. Como o avanço dessa área tem sido notável, há uma crescente oferta de cursos em massa online, em uma solução educacional de cursos abertos e disponíveis na internet (os chamados *Massive Open Online Courses - MOOCs*). *MOOCs* adotam o estilo da educação a distância, com a vantagem de ser ainda mais acessível e global. Esse tipo de modalidade de curso costuma oferecer opções gratuitas de aprendizagem e não exigem pré-requisitos para se matricular.

Nas seções seguintes (2.2 e 2.3), serão descritos alguns conceitos fundamentais para os cientistas de dados.

3. METODOLOGIA

A metodologia empregada neste estudo possui abordagem qualitativa e exploratória, com o principal objetivo de elencar o conjunto de habilidades essenciais para um PCD. Para tal, foi utilizado como método a pesquisa bibliográfica, pesquisa de resultados práticos e entrevista com público selecionado.

A abordagem do estudo é qualitativa, pois, por meio de pesquisa foi estabelecida a fundamentação teórica, focada no caráter subjetivo do objeto analisado. Prodanov e Freitas (2013) definem a análise qualitativa como *"uma relação dinâmica entre o mundo real e o sujeito, isto é, um vínculo indissociável entre o mundo objetivo e a subjetividade do sujeito que não pode ser traduzido em números"*. Em outras palavras, a metodologia visa buscar compreender o comportamento do crescente cenário da tecnologia da informação, estudando as suas particularidades e coletando as experiências individuais dos autores, entre outros aspectos. Nesse método qualitativo de coleta de dados, as respostas costumam não ser objetivas, ou seja, os resultados obtidos não são necessariamente contabilizados em números exatos. Normalmente, a amostra é pequena e os entrevistados são estimulados a se sentirem à vontade para dar sua opinião sobre assuntos relacionados com o objeto de estudo.

Este trabalho constitui-se, também, como uma pesquisa exploratória, uma vez que, uns dos objetivos foi estabelecer a existência de vínculo entre os conceitos analisados nas pesquisas bibliográficas, neste caso, o crescimento da demanda por PCD e o conjunto de habilidades esperado pelo mercado de trabalho para esse profissional. Além disso, por meio da intersecção entre os conceitos encontrados na literatura e na pesquisa com profissionais do mercado de trabalho, foi possível identificar as habilidades e competências que o profissional de dados necessita dominar. Com objetivo de corroborar com o enfoque dado a esta pesquisa, Gil (2002) propõe que o principal objetivo da pesquisa exploratória seja “proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a constituir hipóteses”. O autor considera que as técnicas de coleta de dados mais utilizadas na análise qualitativa são a pesquisa bibliográfica e o estudo de caso.

Nas subseções seguintes são descritas em detalhes cada uma das metodologias utilizadas neste trabalho.

3.1. Pesquisa bibliográfica

Marconi e Lakatos (2003) sinalizam que uma pesquisa bibliográfica, também conhecida como de fontes secundárias, tende a analisar os materiais que já foram produzidos sobre determinado tema de estudo, abrangendo diversos tipos de publicações, tais como livros, revistas, monografias, jornais, etc.

Para este trabalho foram considerados como fontes bibliográficas artigos científicos e livros. Durante a pesquisa bibliográfica, a seleção das fontes utilizadas para embasar este trabalho foi feita a partir da técnica de leitura dinâmica do título e dos principais capítulos aplicáveis ao tema pesquisado.

A pesquisa bibliográfica, realizada no Google Acadêmico, levantou dados sobre os termos “boom da ciência de dados”, “ TI e grande quantidade de informação”, “especialização em *Data Science*”, “perfil do profissional da informação”, “*information professional*”, “*Big Data*”, “ciência de dados”, “*Data Science*”, “cientista de dados” e “*data scientist*”. Nessa pesquisa foram encontrados aproximadamente dez trabalhos relacionados a este tema. O critério utilizado para seleção dos trabalhos foi aquele cujo o foco da pesquisa do autor estava no levantamento de habilidades dos cientistas de dados em âmbito geral, com o objetivo de correlacionar a tendência do agrupamento das áreas mais importantes para o perfil desse profissional.

Nessa pesquisa foram encontrados aproximadamente dez trabalhos relacionados a este tema. O critério utilizado para seleção dos trabalhos foi aquele cujo o foco da pesquisa do autor estava no levantamento de habilidades dos cientistas de dados em âmbito geral, com o objetivo de correlacionar a tendência do agrupamento das áreas mais importantes para o perfil desse profissional.

3.2. Entrevista com público selecionado

Um formulário, utilizando a ferramenta Google Formulários foi construído com o objetivo de compreender a necessidade do mercado quanto as habilidades esperadas do profissional de dados.

O formulário inicialmente foi direcionado para um público selecionado da região

de São Paulo e, posteriormente, para as demais regiões do país. O objetivo era questionar esse público sobre as habilidades essenciais para o profissional que deseja ingressar na área de dados e como deve ser sua preparação.

O formulário contém um questionário com 09 questões e é apresentado integralmente no Apêndice I. As três primeiras questões buscam identificar o participante, sua área de atuação e idade, de forma a permitir compreender o público que respondeu e garantir a correlação dos participantes com o objetivo da pesquisa. A quarta questão busca identificar quais conceitos técnicos são essenciais para um PCD no dia a dia, visando coletar dados das habilidades técnicas essenciais na visão do mercado no Brasil. A quinta questão almeja questionar se existe um perfil desejado de habilidades interpessoais para o PCD. Essa questão permitirá analisar se existem habilidades interpessoais que tornam um PCD mais apto a exercer essa atividade. A sexta questão, opcional, busca identificar, de forma dissertativa, qual habilidade os ingressantes na área de dados mais carecem, buscando analisar possíveis gargalos na formação convencional. A sétima questão, também dissertativa e opcional, tem como objetivo identificar se existe uma ferramenta ou um framework que podem ser considerados como essenciais para ingressar na área: o resultado dessa questão poderá servir de guia para profissionais que busquem atuar, no futuro, na área de dados. As últimas duas questões visam indagar se a grade dos cursos de tecnologia - que devem seguir a grade convencional estabelecida pelo Ministério da Educação (MEC) - estão abrangentes o suficiente para preparar um aluno para área da tecnologia; a análise dessas questões poderá servir de base para futuras pesquisas, que buscarão aprofundar a discussão acerca da formação tradicional, em comparação com o dinamismo do mercado.

3.3. Discussão

A partir do referencial teórico identificado através da pesquisa bibliográfica, e da pesquisa prática, realizada com formulário disponibilizado, foi feita uma análise do resultado - de ambas as pesquisas - de forma a avaliar o alinhamento dessas duas visões acerca do conjunto de habilidades do PCD. No Capítulo 4 é apresentado o detalhamento dos resultados obtidos em cada pesquisa e são analisadas as correlações entre os estudos.

4. ANÁLISE DOS RESULTADOS

Este capítulo visa abordar os resultados obtidos com os levantamentos descritos no capítulo anterior. Na Seção 4.1 é apresentado o resultado da análise de trabalhos similares no âmbito acadêmico e na Seção 4.2 é apresentado e analisado o resultado da pesquisa feita com pessoas que atualmente atuam no mercado de dados e ou áreas correlatas. Por fim, na Seção 4.3, é apresentada uma discussão sobre a correlação entre os resultados obtidos na bibliografia e na pesquisa com profissionais, de forma a traçar o perfil esperado pelo mercado a respeito desse profissional

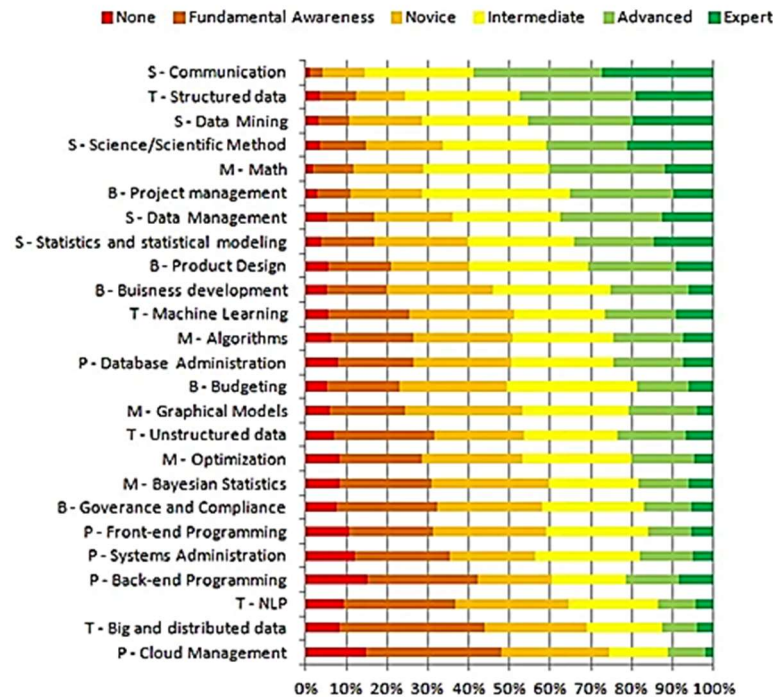
4.1 Pesquisa bibliográfica

Esta seção tem por objetivo apresentar os resultados e os *insights* obtidos por meio do levantamento bibliográfico em bases acadêmicas acerca da PCD, conforme método descrito na Seção 3.2.

Uma pesquisa com teor prático que abrange, em detalhes, as habilidades dos cientistas de dados, foi realizada por Harris et al. (2015). Nela, os pesquisadores entrevistaram várias centenas de profissionais, buscando identificar suas proficiências em 22 habilidades de dados diferentes. Esses pesquisadores descobriram que as habilidades de dados se enquadravam em cinco grandes áreas: Negócios, Matemática, *Big Data*, Programação e Estatística.

Os entrevistados, nessa ocasião, foram membros da comunidade *AnalyticsWeek* e de mídias sociais (por exemplo o *LinkedIn*). Eles responderam um questionário sobre sua proficiência em diferentes habilidades relacionados a dados, educação, funções, membros da equipe e satisfação com seus resultados de trabalho. No total, foram coletadas 490 respostas completas da pesquisa. A maioria dos entrevistados era da América do Norte (representavam 68%) e trabalhavam para empresas B2B (79%) com menos de 1000 funcionários (53%). As áreas de Tecnologia da Informação, Serviços

Financeiros, Educação / Ciência, Consultoria, Saúde e Medicina representaram 68% das respostas da pesquisa. Os homens representaram a maioria das respostas coletadas, com 75% da amostra. A maioria dos entrevistados possuía graduação (30%), mestrado (49%) ou doutorado (18%). A Figura 4 apresenta os resultados obtidos na pesquisa.



Note: Data are based on responses from 490 data professionals. Data professionals were asked to rate their proficiency across 25 skills using a scale from 0 (don't know) to 100 (expert).

Figura 4. Proficiência em habilidades de ciência de dados.

Fonte: Harris et al. (2015).

A Figura 4 mostra que os níveis de proficiência variaram amplamente entre as diferentes habilidades. Os entrevistados relataram um alto grau de competência em áreas como comunicações, dados estruturados, mineração de dados, método científico e matemática. Além disso, na Figura 4 é possível observar que os entrevistados indicaram um baixo grau de competência em áreas como administração de sistemas, programação de *front-end* e *back-end*, dados grandes e distribuídos e gerenciamento de nuvem. Os resultados da pesquisa mostraram que os profissionais de dados tendem a trabalhar juntos para resolver problemas, dado que 76% disseram que trabalham com pelo menos uma outra pessoa em projetos que envolvem análises.

Segundo o Ex Vice-Presidente de *Big Data* da IBM e atual Vice-Presidente da Adobe, Anjul Bhambhri, os PCD possuem uma base sólida normalmente em ciência da

computação, aplicações, modelagem, estatísticas, análises e matemática.

Os pesquisadores Kim e Lee (2016) realizaram a análise de cerca de mil ofertas de emprego para cargos de cientistas de dados de 736 empresas registradas em três diferentes plataformas online de recrutamento dos EUA (*Monster.com*, *Indeed.com* e *CareerBuilder.com*). Nesse trabalho, os autores puderam identificar três abrangentes conjuntos de habilidades e conhecimentos essenciais para exercer a profissão do ponto de vista do mercado: sistemas, negócios e técnicas. Para a classe definida como Sistema, os autores subdividiram a mesma em dois grupos menores que consistem em Desenvolvimento e Solução de problemas, para então listar as áreas de conhecimento associadas, conforme apresentado no

Quadro 2. Quanto a classe de negócios, os autores dividiram em social, de negócio propriamente e gerencial, conforme o Quadro 4. Por fim, a categoria técnica abrange as subclasses de software, arquitetura e redes e hardware, conforme o Quadro 3.

Quadro 2.Habilidades em sistemas computacionais essenciais para o PCD

Desenvolvimento	Solução de problemas
Análise de sistemas	Modelagem de dados
Implementação e teste	Análise
Gestão de dados	Estatística
Conhecimento variado em diferentes tecnologias	Pensamento analítico
Desenvolvimento de metodologias de software	Criatividade/Inovação
Programação	Habilidade de solucionar problemas
Operação e manutenção	Adaptabilidade e flexibilidade
Integração	Capacidade estratégica.
Documentação.	-

Fonte: Adaptado de Kim e Lee (2016).

Quadro 3. Habilidades Técnicas computacional essenciais para o PCD

Software	Arquitetura e redes	Hardware
Programação	Internet	Dispositivo de armazenamento
Banco de dados	Dispositivo de redes	Impressoras
<i>Softwares Open Source</i>	Computação Cloud	Desktop
Visualização de dados	Arquitetura e segurança da rede.	Conhecimentos gerais de hardware
-	-	Servidores.

Fonte: Adaptado de Kim e Lee (2016).

Quadro 4. Habilidades de Negócio essenciais para o PCD

Social	Negócio	Gerencial
Boa comunicação	Conhecimento especializado da área de negócios	Administração geral
Automotivação	Habilidade de análise macro	Organização e liderança
Pensamento Crítico.	Negócios online/e-commerce.	Capacidade de monitoramento e controle
		Planejamento e treinamento

Fonte: Adaptado de Kim e Lee (2016)

Por fim, é possível constatar uma certa convergência dos estudos para concluir que o PCD é um profissional multifacetado, com aptidões matemáticas, conhecimento técnico em computação e perfil de solucionador de problemas. Diante da atual necessidade do mercado, com uma alta demanda por essas aptidões, 36% dos líderes de grandes e médias empresas estariam dispostos a pagar mais para profissionais que apresentem habilidades com dados (Portal Saúde Business, 2019).

Embora ainda seja uma carreira recente em território nacional e em fase de consolidação, é possível constatar um número significativo de anúncios ativos pelo *LinkedIn* no Brasil. Em uma busca feita pelo autor na plataforma *LinkedIn* pelo termo

Data Science no mês de março de 2020, foram identificadas em torno de 600 vagas, evidenciando a crescente da área no país. Diante disso, para preparar os profissionais a crescente demanda deste tipo de profissional para as diversas áreas, existem algumas iniciativas no Brasil na modalidade *lato sensu* de cursos voltados para a CD oferecidos por algumas universidades, como Universidade Estadual de Campinas (UNICAMP), Universidade Presbiteriana Mackenzie, Fundação Getúlio Vargas, Universidade Federal do Rio Grande do Sul (UFRGS). Em outros casos, como o curso de graduação em Ciência da Informação da Universidade Federal de Santa Catarina (UFSC) e a Faculdade de Informática e Administração Paulista (FIAP), também se observa uma tendência em preparar profissionais com habilidades e competências semelhantes às comuns aos cientistas de dados. Muito embora a profissão ainda esteja em fase de consolidação, a ciência já caminha para um consenso acerca do conjunto de disciplinas essenciais para o PCD.

4.2 Entrevista com público selecionado

Esta seção apresenta a pesquisa realizada neste trabalho, cujo objetivo foi compreender as principais habilidades esperadas para um PCD no âmbito do mercado nacional.

Foram convidados, inicialmente, quinze profissionais que atuam diretamente na área de CD no estado de São Paulo. Além disso, o questionário foi divulgado em grupos especializados do *LinkedIn*. A lista completa das questões que compõe o questionário está localizada no Apêndice 1 deste trabalho, tal e qual, o detalhamento das questões dissertativas estão localizadas nos Apêndices de 2 a 5. Ao todo, o questionário foi respondido por quarenta e oito profissionais brasileiros. As respostas obtidas foram tabuladas e analisadas, todas as respostas foram com base na interpretação de cada participante sobre os termos disponibilidades nas opções. As questões que foram abordadas de maneira diferente por algum membro, foram consideradas: nas questões de 1 a 5 foram listadas como “outras” e nas questões de 6 a 9 foram listadas nas tabelas e agrupadas, para fins estatísticos. Para as questões de 6 a 9, a resposta final considerada para resultado da análise, foi o que a maioria dos membros apontou. Os homens representaram a maioria das respostas coletadas, com 87% da amostra. Em relação a faixa etária, a maioria dos entrevistados tinham de 20 a 30 anos (48%), seguidos de 30 a 40 anos (31%), e a pessoas de 40 a 50 anos representaram 21% dos entrevistados.

Por meio da Questão 3 (Qual área de atuação?) dissertativa da pesquisa, foi possível coletar a área de atuação dos participantes, apresentado na **Figura 5**. Fica claro que o maior subgrupo entre as pessoas pesquisadas atua diretamente na área de *Data Science* (22%) seguido das demais áreas relacionadas indiretamente com CD. É possível constatar que há uma distribuição significativa de áreas.

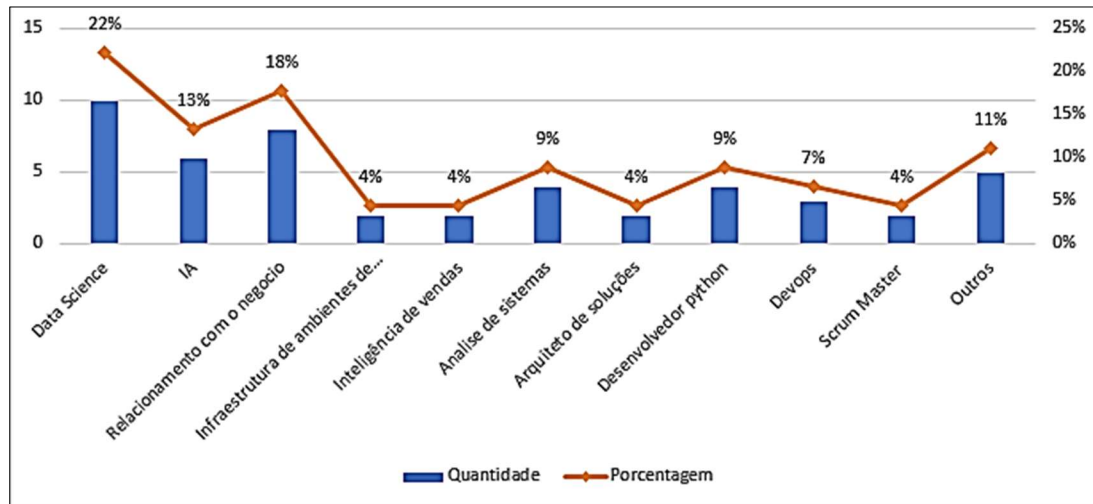


Figura 5. Área de atuação dos participantes da pesquisa realizada neste trabalho.

Visto que o objetivo da pesquisa foi elencar as principais habilidades do PCD, por meio da análise da Questão 4 (Quais são as habilidades técnicas essenciais para o PCD?), deve-se usar com cautela o que foi levantado neste trabalho, dado que apenas 22% dos respondentes indicaram atuar com *Data Science*, atuação que mais se alinha com o que se define como PCD. Com base na pesquisa, foi possível identificar que, entre os entrevistados, um PCD deve possuir um domínio técnico abrangente das áreas da tecnologia. Destaca-se o conhecimento na área de IA, *Big Data* e Estatística, bem como espera-se um bom entendimento lógico e de programação, em específico de *Python* (Figura 6).

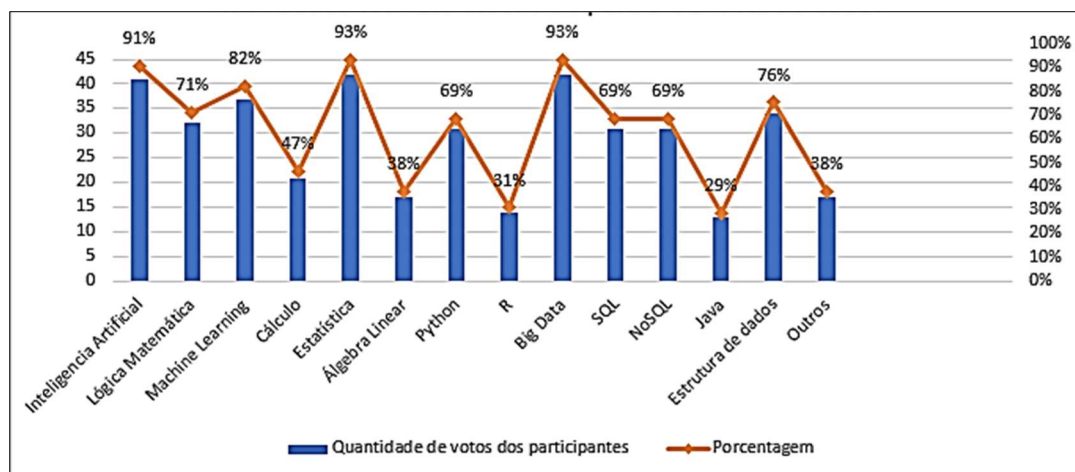


Figura 6. Habilidades técnicas essenciais para o PCD definidas pelas respostas obtidas pelos entrevistados deste trabalho. (No campo descrito como outros, foram considerados os conceitos de técnicas de apresentação e visualização da informação, algoritmo para organização de dados, expressões regulares e deep learning.)

Por meio da Questão 5 (Quais são as habilidades interpessoais essenciais para o PCD?), foi possível mapear as qualidades interpessoais que se destacam como essenciais para o PCD. Com ilustrado na Figura 7, é possível constatar que o PCD tem um perfil analítico (92%), seguido de uma capacidade de senso crítico (63%) e visão estratégica de negócio (63%).

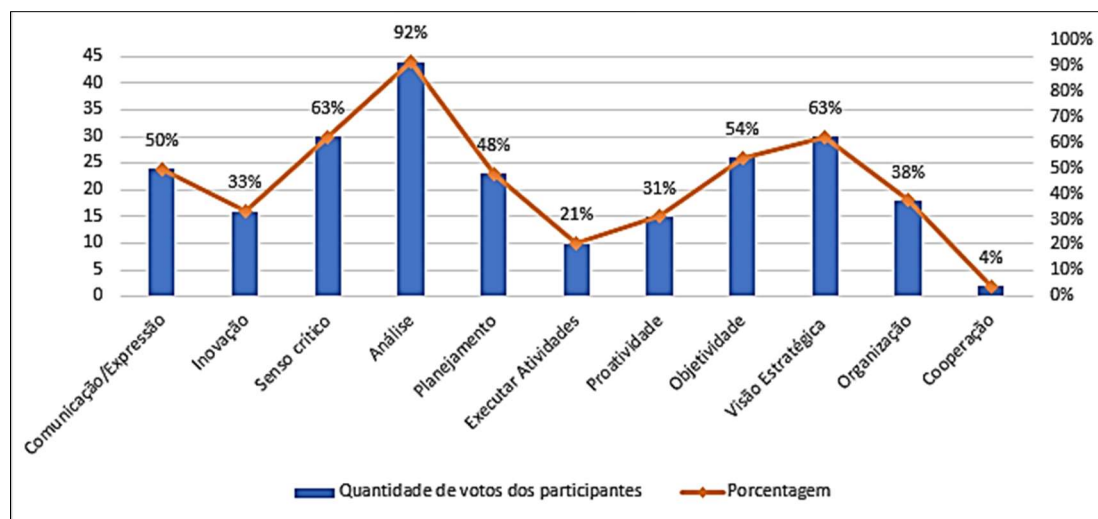


Figura 7. Habilidades interpessoais essenciais para o PCD definidas com base nas respostas dos entrevistados por este trabalho.

O intuito da Questão 6 (Quais técnicas/conceitos o mercado identifica que o ingressante mais carece na área de dados?) do formulário, cujo formato é dissertativo, foi

identificar se o mercado observa carências de habilidades ao procurar um profissional iniciante. Nesse sentido, os comentários foram avaliados e agrupados quando aplicável. O consolidado das respostas pode ser consultado no Apêndice 2, no qual foi possível identificar que os conceitos que o mercado procura para ingressantes na área, estão associados a sólidos conceitos nas novas tecnologias, principalmente em *Big Data*, bem como conceitos lógicos e conhecimento de sistemas de apoio a decisão. Além disso, teve bastante destaque a necessidade da prática propriamente de conduzir a solução de um problema complexo com uma gama variada de dados.

A análise das respostas apresentadas para a Questão 7 (É necessário ter conhecimento prévio de algum software em específico?), no qual o detalhamento completo está no Apêndice 3, permitiu identificar que 60,9% dos que responderam essa questão acreditam que conhecimento de software é essencial para quem visa ingressar na área de dados. O conhecimento de softwares de inteligência do negócio (BI) apareceu com 21%, distribuídos em software como *Tableau* e *Power B.I*, assim como o conhecimento em *Apache* (26%). Nessa questão também foi possível identificar a importância do conhecimento da linguagem de programação *Python*. É também possível constatar a pluralidade e abrangência de conteúdos elencados pelos participantes.

Por fim, as Questões 8 (O que a universidade precisa adicionar/alterar na grade para embasar esse futuro profissional?) e 9 (O que a universidade precisa manter que auxilia na preparação desse perfil de profissional?), no qual o detalhamento completo está nos Apêndices 4 e 5, tiveram como finalidade indagar os especialistas e propor uma reflexão acerca da abrangência da grade curricular das universidades. Foi questionado o que poderia ser incluído na formação dos novos profissionais, e o que deve ser mantido, que já funciona bem, dado o dinamismo do mercado de tecnologia atual.

Por meio da análise das respostas, é possível identificar que a universidade tem uma abrangência inicial de quase todos os importantes conceitos que permeiam a CD, tais como: estatísticos, lógica e programação. Entretanto, poderia incluir disciplinas específicas para as novas tecnologias, de forma a contemplar softwares direcionados, ambientes com grande massa de dados (*Big Data*), detalhar aprendizado de máquina e *deep learning*. Constata-se, por meio dessas questões, que seria importante uma interação de disciplinas com teor mais prático.

4.3 Discussão

A partir da análise feita a respeito das habilidades almejadas em um PCD é possível traçar uma intersecção entre o que se constatou na literatura e na pesquisa prática. Pode-se concluir que é necessário um viés bem amplo e multifacetado para ser um bom PCD. Essa diversidade de conhecimentos se deve a abrangência das etapas relacionadas com a análise dos dados, que incluem identificação das fontes, captura e armazenamento, acesso, análise e, por fim, exposição dos resultados dos dados analisados (Coneglian e Gonçalves, 2017).

É possível constatar que, apesar da pluralidade de conhecimentos e abrangência de formações, há uma maior exigência de conhecimento sobre Computação e Estatística para o PCD. Esse fato corrobora com a ideia de Coneglian, Gonçalves e Santarém Segundo (2017), que defendem que a atividade de análise dos dados, no ambiente *Big Data*, é de atribuição do cientista da computação, responsável pelo desenvolvimento de algoritmos e a implantação de princípios estatísticos. Além disso, o fato também confirma a percepção de alguns autores de que a profissão de PCD não diz respeito a um único indivíduo munido de todo o conhecimento necessário para realizar a análise e interpretação dos dados, mas sim, de diversos profissionais contribuindo mutuamente (Davenport, 2014; Coneglian e Gonçalves, 2017).

De certa forma, é possível identificar que existe uma similaridade entre as etapas do processo de transformação do dado em informação (Captura, Processamento, Geração de interpretação e Construção da informação) com as etapas do método científico (Definição do problema, Medição. Análise, Resultados e Discussão) (PROVOST, F., FAWCETT, T. (2016) Além disso nota-se uma similaridade nos conceitos Estatístico com o conceito do Negócio que permeia a CD, nos casos de formas de resumir e visualizar dados; testes de hipóteses, técnicas de análise preditiva, modelagem de dados

É possível destacar a expectativa de domínio de linguagens de programação passíveis de utilização em ambientes *Big Data*, como por exemplo Python, além do domínio sobre a matéria de bases de dados (Ferreira, 2003). Foi possível constatar, também, por meio da pesquisa com público selecionado, que garantir uma boa base de dados é um dos maiores desafios do PCD. Além disso, outro conceito bastante pertinente quanto se trata de habilidades para o PCD é o de banco de dados não convencionais, como

no caso do NoSQL, muito utilizado para atuar junto a grandes volumes de dados (Victorino et al., 2017; Loukides, 2010).

5. CONCLUSÕES

Por meio deste trabalho foi possível verificar que com o rápido avanço da TI, e seu uso no cotidiano por boa parte da população, a CD tornou-se uma das principais áreas para as empresas gerarem vantagens competitivas, pois utiliza diferentes tecnologias, metodologias e modelos para capturar, armazenar e processar as informações. Por ser uma área interdisciplinar e em rápida e constante evolução, existem dificuldades em definir qual o conjunto de habilidades que um PCD deve dominar. O objetivo deste trabalho foi apresentar as habilidades essenciais que compõe o perfil do PCD, enumerando contribuições que este pode trazer ao universo corporativo atuando na solução de problemas de informação cada vez mais complexos.

Em ambas as pesquisas realizadas neste trabalho, tanto a bibliográfica, quanto a prática foi possível constatar que um PCD deve possuir domínio em diversas áreas de conhecimento. No levantamento bibliográfico, cinco grandes áreas foram consideradas essenciais para um PCD: negócios, matemática, *Big Data*, programação e estatística. De acordo com as respostas referentes ao questionário aplicado neste trabalho, foi possível encontrar similaridades das habilidades ditas como essenciais na pesquisa bibliográfica. Com base nas respostas dos entrevistados da pesquisa prática, um PCD precisa ter um domínio técnico principalmente em IA, *Big Data*, Estatística, lógica e programação. Além disso, dentre as qualidades interpessoais importantes, destaca-se o senso crítico, analítico e visão estratégica de negócios. Em relação aos iniciantes da área, é preciso ter conceitos bem sólidos referentes às novas tecnologias, principalmente no que se refere a conceitos lógicos, *Big Data*, além de conhecimentos de sistemas de apoio a decisão e softwares.

Foi possível identificar que o PCD é um profissional com aptidões no âmbito interpessoal relacionadas a um perfil analítico associada a um senso crítico e visão estratégica de negócio, habilidades essas que permitem uma melhor compreensão das possibilidades diante de uma grande massa de dados.

Os resultados deste trabalho podem fornecer subsídios e diretrizes gerais a respeito da formação de cientistas de dados para uma reflexão quanto ao público-alvo interessado em ingressar na área de CD sendo possível compreender tanto os desafios do PCD como profissional que obtém *insights* em um cenário com grandes e voláteis quantidades de dados, bem como as principais habilidades, tidas como essenciais para ser um profissional preparado para atuar na área. Desta forma, esta monografia poderá

contribuir para futuros trabalhos mais aprofundados quanto ao conjunto de habilidades necessárias para o CD, além de incentivar outros profissionais a ingressar nessa importante área da tecnologia que transforma a TI em uma área estratégica no mercado.

6. REFERÊNCIAS

Associação Brasileira de Ciência de dados (ABRACD). **O que é Data Science: a famosa Ciência de Dados**, 2019. Disponível em: <https://abracd.org/o-que-e-data-science-a-famosa-ciencia-de-dados/>. Acesso em: 20 de nov 2020.

BUCKLAND, M. K. **Information as thing**, 1991. Journal of the American Society for Information Science, [S.l.], v. 45, n. 5, p. 351-360. Disponível em: <https://pdfs.semanticscholar.org/b3d4/d7980d6a628b503003ef4e7763a93544508e.pdf> >. Acesso em: 15 de jan. 2020

CAPURRO, R.; HJORLAND, B. **O conceito de informação. Perspectivas em Ciência da Informação, Belo Horizonte**, 2007, v. 12, p. 148-207. Disponível em: http://www.capurro.de/conceito_informacao.pdf . Acesso em: 19 nov. 2020.

CERVANTES, R. G. C. N. **DATA SCIENCE: CIÊNCIA ORIENTADA A DADOS**, 2016. Disponível em: <https://brapci.inf.br/index.php/res/download/45442>. Acesso em: 20 de nov. 2020.

CETAX. **Hadoop: O que é, conceito e definição**, 2016. Disponível em: <https://www.cetax.com.br/blog/apache-hadoop/#:~:text=Hadoop%20%C3%A9%20uma%20plataforma%20de,seguran%C3%A7a%20e%20opera%C3%A7%C3%B5es%20de%20Dados>>. Acesso em: 19 de nov. 2020

CHATFIELD, A. T. et al. **Data Scientists as a game changers in big data environments**, 2014. PROCEEDINGS OF THE 25TH AUSTRALASIAN CONFERENCE ON INFORMATION SYSTEMS (ACIS), Auckland University of Technology, p.1-11.

CONAWAY, D. **The data science venn diagram**, 2010. Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>>. Acesso em: 15 de dez. 2019.

CONEGLIAN, C. S.; GONÇALVES, P. R. V. A. SANTARÉM SEGUNDO, J. E. **O profissional da informação na era da big data**, 2017. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 22, n. 50, p. 128-143, set./dez. 2017. Disponível em: <<https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2017v22n50p128>>. Acesso em: 10 fev. 2020.

CRAWFORD, K.; MILTNER, K.; GRAY, M. L. **Critiquing big data: politics, ethics, epistemology**. International Journal of Communication, [S.l.], v. 8, p. 1.663-1.672, 2014. Disponível em: <<http://ijoc.org/index.php/ijoc/article/view/2167/1164>>. Acesso em: 10 fev. 2020.

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

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

DENG, L.; YU, D. Deep Learning: Methods and Applications. 2014. Disponível em <<https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/>>.

FEIGENBAUM, E in HARMON & KING - EXPERT SYSTEMS: Artificial Intelligence in Business - 1985 - New York, Ed Wiley

FILHO, D. L. Ba. **Experiências com desenvolvimento ágil**. Instituto de Matemática e Estatística da Universidade de São Paulo (Dissertação de Mestrado). São Paulo: USP, 2008.

FINZER, W. **The data science education dilemma**. Technology Innovations in Statistics Education, California, v. 7, n. 2, p. 1-9, 2013. Disponível em: <<http://escholarship.org/uc/item/7gv0q9dc>>. Acesso em: 15 fev. 2020.

FUNDAÇÃO INSTITUTO ADMINISTRAÇÃO (FIA). **Ciência de Dados ou Data Science: O que é, Aplicações e Perfil Profissional**, 2019. Disponível em: <https://fia.com.br/blog/ciencia-de-dados-data-science/>. Acesso em: 20 nov. 2020

GANTZ, J.; REINSEL, D. **Extracting Value from Chaos**, 2011. Disponível em <https://pt.scribd.com/document/87162672/IDC-IView-Extracting-Value-From-Chaos-2011-Data-Storage-Etc>. Acesso em 19 nov. 2020.

GARTNER, **Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s**, 2014. Disponível em: <https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/?sh=5dda5ff942f6>. Acesso em: 20 de nov. 2020.

HARRIS, D.; MURPHY, S.P.; VAISMAN, M. **Analyzing the Analyzers An Introspective Survey of Data Scientists and Their Wor**, 2015. Disponível em: http://www.stat.wvu.edu/~jharner/courses/dsci503/docs/Analyzing_the_Analyzers.pdf. Acesso em: 19 nov. 2020.

HOLUCHI, ANDRÉ GUSTAVO. **Abordagens livres de segmentação para reconhecimento automático de cadeiras numéricas manuscritas utilizando aprendizado profundo**, 2016. Disponível em: <https://acervodigital.ufpr.br/bitstream/handle/1884/59972/R%20-%20T%20-%20ANDRE%20GUSTAVO%20HOCHULI.pdf?sequence=1&isAllowed=y>
Acesso em: 24 jan. 2021

INSTITUTO AVALIAÇÃO, **O que são Sistemas Inteligentes?** s.d. Disponível: <http://www.avaliacao.org.br/pages/areas-de-atuacao/sistemas-inteligentes.aspx#:~:text=Sistemas%20Inteligentes%20s%C3%A3o%20plataformas%20tecnol%C3%B3gicas,ser%20humano%20para%20solucionar%20problemas>. Aceso em: 20 de nov 2020.

ISACA. **Big Data: Impactos e Benefícios**, 2013. Disponível em: < <http://www.isaca.org/Knowledge-Center/Research/Documents/Big->

Data_whp_Por_0413.pdf>. Acesso em: 19 nov. 2020.

KDNUGGETS, M. M. **Data Science Primer: Basic Concepts for Beginners**, 2017. Disponível em: <https://www.kdnuggets.com/2017/08/data-science-primer-basic-concepts-for-beginners.html>. Acesso em 19 nov. 2020.

KEMP, S. **DIGITAL 2019: Q4 GLOBAL DIGITAL STATSHOT**, 2019. Disponível em: <https://datareportal.com/reports/digital-2019-q4-global-digital-statshot>. Acesso em: 19 nov. 2020.

KIM, J. Y.; LEE, C. K. **An empirical analysis of requirements for data scientist using online job postings**, 2016. International Journal of Software Engineering and its Applications, Seoul, v. 10, n. 4, p.161-172.

KIMBALL, R. **The Data Warehouse Toolkit: practical techniques for building dimensional data warehouses**, 1996. New York: John Wiley & Sons.

LOUKIDES, M. **What is data science? the future belongs to the companies and people that turn data into products**, 2010. O'Reilly Radar, [S.l.].

MARINO, L. H. F. C. **Gestão da qualidade e gestão do conhecimento: fatores-chave para produtividade e competitividade empresarial**, 2006. Disponível em: https://simpep.feb.unesp.br/anais/anais_13/artigos/598.pdf. Acesso em 20 de nov. 2020

MARCONI, M. A.; LAKATOS, E. M. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2003.

MCAFEE, A.; BRYNJOLFSSON, E. **Big data: the management revolution**, 2012. Harvard Business Review, [S.l.], v. 90, n. 10, p. 60-68. Disponível em: https://eller.arizona.edu/sites/ai/files/mis510/big_data_-_the_management_revolution_0.pdf. Acesso em: 20 mar. 2020

MICHALSKI, R. **Knowledge acquisition through conceptual clustering: A**

theoretical framework and an algorithm for partitioning data into conjunctive concepts, 1980. Disponível em:

https://books.google.com.br/books/about/Knowledge_Acquisition_Through_Conceptual.html?id=r5blXwAACAAJ&redir_esc=y. Acesso em 20 de nov. de 2020.

MICHIE, D. MELTZER. **Machine Intelligence 4**, Pp. viii, 508. Edinburgh University Press, 1969.

MILLER, C. C. **Data science: the numbers of our lives**, 2013. The New York Times, 23. Disponível: <https://www.nytimes.com/2013/04/14/education/edlife/universities-offer-courses-in-a-hot-new-field-data-science.html>. Acesso em: 20 nov. 2020.

McKinsey Global Institute (MGI). **Highlights of our research this year**, 2019. Disponível:<https://www.mckinsey.com/~media/McKinsey/Featured%20Insights/Innovation/Ten%20highlights%20from%20our%202019%20research/MGI-in-2019-A-compendium-of-our-research-this-year-vF.ashx>. Acesso em 19 de nov. de 2020

MONTEIRO, L. P. **Dados Estruturados e Não Estruturados**, 2019. Universidade da Tecnologia. Disponível em <https://universidadedatecnologia.com.br/dados-estruturados-e-nao-estruturados/>. Acesso em: 20 nov. 2020.

PINHEIRO, H. P. **Sistemas inteligentes: fundamentos e aplicações**, 2003. Disponível em: <https://www.ime.unicamp.br/~hildete/dados.pdf>. Acesso em 20 de nov. 2020.

Portal Saúde Business. **Data Literacy Project apresenta certificação em Alfabetização de dados reconhecida mundialmente**, 2019. Disponível em: <https://saudebusiness.com/gestao/data-literacy-project-apresenta-certificacao-em-alfabetizacao-de-dados-reconhecida-mundialmente/>. Acesso em: 19 de nov. 2020.

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**, 2013. 2. ed. Novo Hamburgo: Feevale.

PROVOST, F.;FAWCETT, T. **Data science para negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados**, 2016. Rio de Janeiro: Alta

Books.

REZENDE, E. **Dados, Informação e Conhecimento**, 2015. Disponível em: <https://eliana-rezende.com.br/dados-informacao-e-conhecimento-o-que-sao/>. Acesso em 19. nov. 2020.

ROLIM, MESAQUE VIDAL. **Análise do perfil do profissional da informação para a atuação como cientista de dados em ambientes de big data: uma perspectiva a partir das disciplinas do curso de biblioteconomia da UnB**, 2018. Disponível em: https://bdm.unb.br/bitstream/10483/20898/1/2018_MesaqueVidalRolim_tcc.pdf. Acesso em 24. jan. 2021

RUSSEL, S.; NORVIG, P.; Artificial Intelligence: A Modern Approach. Prentice-Hall, Second Edition, 1995. Disponível em: <https://www.cin.ufpe.br/~tfl2/artificial-intelligence-modern-approach.9780131038059.25368.pdf>. Acesso em 17. Jan. 2021

APÊNDICE 1. Questionário referente pesquisa com especialistas sobre o conjunto de habilidades do PCD

- 1) Qual seu nome? * - Dissertativa
- 2) Qual idade? * Alternativa
- 3) Qual área de atuação? Dissertativa
- 4) Quais são as habilidades técnicas essenciais para o PCD? *Caixa de seleção
- 5) Quais são as habilidades interpessoais essenciais para o PCD? * Caixa de seleção
- 6) Quais técnicas/conceitos o mercado identifica que o ingressante mais carece na área de dados? Dissertativa
- 7) É necessário ter conhecimento prévio de algum software em específico? Dissertativa
- 8) O que a universidade precisa adicionar/alterar na grade para embasar esse futuro profissional? Dissertativa
- 9) O que a universidade precisa manter que auxilia na preparação desse perfil de profissional? Dissertativa

APÊNDICE 2. Principais respostas da questão de número 6 dos participantes da pesquisa realizada neste trabalho.

Quais técnicas/conceitos o mercado identifica que o ingressante mais carece na área de dados?
Classificação e previsão de dados
Estruturação e compreensão dos dados
Ferramentas de modelagem
Base de cálculo e álgebra linear
Desprender dos paradigmas antigos.
Não saber o que fazer com grande quantidade de dados (<i>Big Data</i>)
Big Data e SAD*
Senso Crítico e Análise
Comunicação e senso critico
Habilidade de aplicar conhecimento técnico para resolver os problemas reais do negócio
Habilidades de desenvolvimento; conhecimento mais aprofundado de frameworks, <i>machine learning</i> e <i>deep learning</i> .
Muitos alunos recém formados tem pouca ou nenhuma base em <i>python</i> .
O mercado carece de pessoas da computação com conhecimento avançado em matemática.
Falta de experiência com <i>big data</i>
Estruturação de dados: Desempenho de operações baseados no fluxo de dados
SAD
<i>Machine Learning</i> , <i>Deep Learning</i> e conceitos de mineração de dados
Trabalhar com dados não estruturados
Conhecimento de software de processamento de dados

APÊNDICE 3. Respostas referente a questão 7 dos entrevistados deste trabalho.

Conhecimento de Software é essencial para o PCD?
Sim, Apache
Sim, Apache Spark, Matlab e Tlabeau
Não.
Nao.
Não, python já é suficiente
Sim, SAS e Tableau
Sim, Frameworks de testes e de front end
Sim, SAS, power BI, Alterix, Python, sql
Sim, software de Data Viz para elaborar hipóteses
Não.
Não, minimo de excel ja e o suficiente para analise, e conhecimento logico.
Sim, softwares de BI(Tableaux, power BI)
Não, linguagens de programação (Python e R) são suficientes
Sim, ferramentas estatísticas como R ou the data <i>wrangling</i> como Python/pandas são essenciais
Sim, conhecimento de software de mineração de dados
Não.
Não, se entender um pouco de lógica, terá base para qualquer linguagem de programação.
Não, se tiver conhecimentos de estatística, banco de dados e programação ajuda muito
Sim, Python e as principais bibliotecas para ML/DL . Conhecimentos de ferramentas de desenvolvimento como Git, Jupyter Notebooks.
Sim, Apache Hadoop
Sim, R, Python, Colab, Pentaho.
Sim, Banco de Dados, Power BI
Sim, ferramentas de dashboard são essenciais

APÊNDICE 4. Respostas referente a questão 8 dos entrevistados deste trabalho.

O que a grade curricular dos cursos de tecnologia poderia incluir para melhor prepara o PCD?
Um maior estudo sobre as ferramentas utilizados pelos profissionais de Data Science
Incluir disciplinas de <i>Big Data</i> , Mineração de dados, Python como obrigatórias nos cursos.
Curso específico para essa demanda ou um caminho com disciplinas que guiem para um TCC nessa área.
Atividades interdisciplinares que foquem na análise de dados (Estatística e programação)
Curso práticos em parceria com empresas e institutos
Solidificar mais os conceitos de cálculo e álgebra para alunos de T.I, facilitando o entendimento dos modelos
Incluir <i>Data Analytics</i> como disciplina
Colocar mais atividades práticas que envolvam análise dos dados e aplicação de estratégias
Proposta reais, de empresas e situações que cabem ao dia a dia que atraiam o aluno e mostre a importância desse conceito.
Tentar trazer casos reais para universidade, ensinar diversas informações sobre mineração, análise, interpretação de dados e diferentes softwares e linguagens voltadas a esse fim, por exemplo em python temos a biblioteca pandas
Disciplinas de Big Data, SAD
Fortificar estatística básica e aplicada. Laboratórios de análise de dados também são essenciais
Precisa se aprofundar mais na matemática necessária (probabilidade e álgebra linear) e propor o ensino de técnicas mais diversas, menos clássicas e talvez mais complexas.
Disciplinas que trabalhem a colaboratividade entre os alunos, desenvolva habilidades interpessoais
Aplicar aulas práticas para extração de dados, ensino nas ferramentas de visualização de dados (PowerBI, tabelau), falar sobre <i>storytelling</i> , estudo de caso de startup, método ágil.
Incluir disciplinas de resolução de problemas, big data, python, R, análise de dados
Incluir git/bash/containers e boas práticas de código.
Focar mais em linguagens de programação recentes (Python, R, Angular)
Trabalhos práticos com dados reais
Deve-se incluir na formação os conceitos de BIG Data e conhecimento em Storage
Análise de dados

Geralmente, a Ciência de Dados é abordada de forma superficial, diluída em outras disciplinas e com um viés puramente mercadológico. Seria importante que as universidades contassem com disciplinas específicas para Ciências de Dados e que estas envolvessem sua discussão a partir de uma perspectiva científica.

APÊNDICE 5. Respostas referente a questão 9 dos entrevistados deste trabalho.

O que a grade curricular dos cursos de tecnologia já faz para preparar o PCD
As matérias de estatística, álgebra e banco de dados dão uma boa base inicial
Estudos na área de Estatística e estrutura de dados.
Matérias de IA
Análise de sistema 1 e 2 dão uma boa base
Estrutura de dados é algo bem interessante
Visualização gráfica.
Disciplinas de introdução à programação, logica é a base para todas as demais.
Lógica, conceitos de IA
Aulas de estatística
Precisa manter uma formação generalista e de base (ou "teórica").
Lógica de programação, matemática, probabilidade e estatística, banco de dados, linguagem de programação.
Lógica e estatística
O ensino avançado de disciplinas através das ICs e matérias eletivas. Como é uma área que está numa fronteira da computação e evolui rapidamente, muito do material e do SOTA só é aprendido através de papers.
Manter a perspectiva de fazer o aluno pesquisar além da sala de aula.
Estatística
Base de conhecimento técnico
IA, logica matemática, python
Disciplinas referentes a Banco de Dados e análise estatística devem ser mantidas, pois contribuem para a formação do Cientista de Dados