Ciência de Dados

Ivan Leoni Vilas Boas

[1]Universidade Federal de Itajuba - UNIFEI
Av. BPS, 1303 - Pinheirinho, Itajubá - MG, 37500-903 - Brasil

[2]Sistemas de Informação

ivanleoni18@hotmail.com

Abstract. *The most varied digital technologies allow the production, storage, consultation and circulation of data on a large scale, however the data alone does not have added value and as a result of the challenges imposed by the massive creation of data, the best way of collecting was rethought, analyze and transform the data into useful information in order to bring solutions and support the right strategic decisions. It is necessary, then, to add value and allow new and better ways of using the data. In order to respond to the existing demands for analysis, exploration, understanding of data and as a result of transformations in science and business, an interdisciplinary and intensively computational study area called data science emerged.*

Resumo. *As mais variadas tecnologias digitais permitem a produção, armazenamento, consulta e circulação dos dados em grande escala, porém os dados por si só não apresentam valor agregado e como consequência dos desafios impostos pela criação massiva de dados foi repensado a melhor maneira de como coletar, analisar e transformar os dados em informações útil de modo a trazer soluções e a subsidiar decisões estratégicas acertadas. É preciso, então, agregar valor e permitir novas e melhores formas de uso dos dados. Para responder às demandas existentes de análise, exploração, compreensão dos dados e como resultado da transformações na ciência e nos negócios surgiu a área de estudo interdisciplinar e intensivamente computacional denominada ciência de dados.*

1. Introdução

Antes a preocupação da ciência tecnológica era, entre outros aspectos, em como armazenar um grande volume de dados de forma consistente e de sua recuperação quando necessário, e atualmente a atenção está voltada em como coletar e organizar os dados de forma eficiente e inteligente, e para além disso, em como transforma-los em informação útil de modo a oferecer uma nova interpretação que tenha significado de dimensão social, econômica, comunicacional e informacional. Com a ascensão da grande quantidade massiva de dados gerada diariamente por pessoas e corporações em sistemas privados, redes sociais e plataformas digitais surgiu o termo big data, um conjunto de ferramentas com capacidade de receber, identificar e interpretar com agilidade um grande volume de dados de diferentes tipos. De acordo com o relatório da seagate em parceria com a IDC [REINSEL et al.] os dados serão produzidos globalmente de 33 zettabytes em 2018 para 175 zettabytes até 2025, onde se espera que 30% dos dados do mundo necessite de processamento em tempo real.

Aliado ao big data, a expansão da internet das coisas, onde o aumento cada vez mais significativo de dispositivos inteligentes que estão interconectados via internet e apresentam capacidade para processar, criar e transmitir dados, está revolucionando as formas de como os dados são gerados, uma vez que o grande número de sensores de inovadoras formas criam e armazenam dados diariamente nas nuvens. Foi estimado [EGHAM] que 20,4 bilhões de dispositivos equipados com sensores poderosos já estariam conectados à internet neste ano.

Nesta conjuntura de dados massivos sendo gerados diariamente surgiu a ciência de dados (DS) , que segundo [CURTY and CERVANTES] a ciência orientada a dados faz uso de uma robustas ciberinfraestruturas de comunicação e informação, incluindo as tecnologias de grids, e também de padrões que possibilitam a interoperabilidade e a interligação dos dados. A DS surgiu com finalidade de identificar padrões e extrair respostas a partir de grandes conjuntos de dados com recursos, técnicas de análise e com métodos científicos. Portanto, a DS tem capacidade de resolver problemas, trazer um alto retorno as empresas e aos setores da economia, e ainda de interferir no consumo e por consequência no modo de vida das pessoas, o que a torna atualmente um universo de possibilidades a ser explorada.

O emprego do DS é muitas vezes imperceptíveis aos usuários comuns e atualmente está cada vez mais presentes no cotidiano sendo utilizado nas mais variadas possíveis situações, como por exemplo, segundo [GRUS] na própria internet que possui bases de dados específicos de domínio como músicas, filmes, resultados de esportes e nas estatísticas do governo. Se faz presentes na internet das coisas onde os carros inteligentes que coletam hábitos de direção e as casas inteligentes colhem hábitos de moradia. E também nos smartphones que fazem o registro da localização, nos aplicativos que absorvem os dados de saúde como hábitos de movimentos, batidas do coração, dieta e padrões do sono, e ainda nos comerciais que armazenam os hábitos de seus clientes e dos websites que fazem o rastreamento de todos os cliques dos usuários.

2. O que vem a ser Ciência de Dados

Por ser considerada um ramo recente no Brasil e no mundo não existe uma definição especifica e única do que vem a ser DS, porém nenhuma das definição dadas pelos especialistas e defensores da área se contradizem, apenas se assimilam e complementam. Assim para poder solucionar questões complexas e que envolvem uma vasta quantidade de dados surgiu a DS que conforme [BELL et al. 2009a] tem como objetivo extrair conhecimento, padrões, tendências e insights a partir de conjuntos de dados de vários formatos, estruturados ou não-estruturados através dos processos científicos e computacionais e segundo [AMARAL 2016] a DS é formada pelos processos, modelos e tecnologias que estudam os dados durante todo o seu ciclo de vida, da produção ao descarte e [GRUS] resume a definição como a ciência direcionada para extração de conhecimento a partir de dados desorganizados. Assim para solucionar problemas analíticos complexos a DS utiliza-se de um conjunto multidisciplinar de inferência de dados, do desenvolvimento de algoritmos e da tecnologia. Quando os dados extraídas de um repositório de informações brutas são utilizados de forma criativa tem competência para gerar valor aos negócios. Pontando pode-se dizer que DS é conjunto de técnicas cientificas-computacionais capaz de coletar dados de diversas fontes para processar e analisar grandes quantidades de dados afim de encontrar informações relevantes sobre a base coletada e assim poder descobrir soluções

e agregar valor as organizações.

Para [VANDERPLAS 2016] a DS envolve um conjunto de habilidades interdisciplinares que estão se tornando cada vez mais importantes em muitas aplicações, tanto nas indústrias quanto nas academias. A DS então é considerado uma área nova, mas formada por outras já existentes, que se relacionam como a ciência da computação para armazenar, obter e tratar os dados; a especialização científica para saber fazer a perguntar certa, do que buscar e do porquê buscar; a matemática e estatística para através da ciência realizar a filtragem e mineração dos dados, e também pode-se incluir neste contesto o design gráfico para melhor visualizar e saber refinar as informações geradas. Nesta perspectiva, [CONWAY] criou o diagrama de Venn para especificar as habilidades que concerne a DS (Figure 1) abarcando as principais disciplinas que a compõem DS.

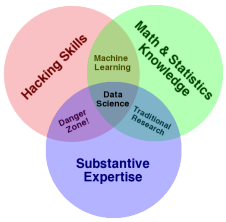

Figura 1. Diagrama de Venn desenvolvido por Conway

É sabido por [CONWAY] que a compreensão da área de DS abrange outras áreas do saber e no diagrama criado pelo autor, a DS aparece no centro e se correlaciona com outras capacidades como conhecimento exato, computacional e especialista, além de aprendizagem de máquinas e analises tradicionais. Conforme [CONWAY] as habilidades de hackers abrangem a capacidade de manipulação de arquivos de texto na linha de comando, a compreensão de operações vetorizadas, a elaboração de algoritmos, a captura e limpeza eficiente dos dados. Porém, os hackers da Zona de Perigo são capazes de extrair e estruturar dados, mas não necessariamente conseguem compreeder o que os resultados significam. A matemática e estatística serão essenciais para a extração das informações dos dados selecionados, através da aplicação de métodos adequados e com a utilização da computação para realizar o aprendizado de máquina, garantindo assim a definição dos modelos e dos algoritmos corretos, e ainda o entendimento das implicações e dos resultados.

O conhecimento especializado é oriundo das experiências do profissional, quanto mais áreas do saber tiver vivência mais apto estará para fazer as perguntas certas, validar as possibilidade e realizar a segregação das variáveis principais de um conjunto de dados e assim mais próximo da soluções estará. Como a base da ciência é descobrir e construir conhecimento requer possuir questões motivadoras sobre o mundo e originar hipóteses, que podem ser trazidas aos dados e testadas com métodos estatísticos. O conhecimento, portanto, acerca do negócio será essencial para a tradução dos resultados em informações validas, que auxiliem a tomada de decisão. Conforme complementam [ZIVIANI and PORTO 2014] a análise de dados em larga-escala possui grande potencial tecnológico com impacto em diferentes áreas do conhecimento e de setores de atuação. Assim a DS possui como característica a interdisciplinaridade conforme [EMMERT-STREIB et al. 2016] e multidisciplinariedade segundo [TIERNEY]. Portanto a DS pode ser descrita como a ciência responsável pela análise e utilização de dados

que incorpora técnicas e teorias de diversas áreas, como lógica, matemática, estatística, programação, computação, mineração de dados (MD), aprendizado de máquina (ML), engenharia, economia e negócios.

Segundo [PROVOST and FAWCETT 2016] o objetivo da DS é aprimorar através da análise automatizada de dados a tomada de decisão, a fim de entender os fenômenos. Essa ciência pode ser utilizada por qualquer setor que gere dados e que dependa de análises para um melhor entendimento de fenômenos, desde a compreensão de eventos passados até a previsão de tendências para o futuro. A computação tem evoluído cada vez mais em prol de uma abordagem mais cognitiva, o que permite aos softwares serem inteligentes o bastante para aprenderem sozinhos, abragendo assim tecnologias como ML e Deep Learning (DL). Cada vez mais se utiliza da computação científica de alto desempenho, para extrair informações dos dados. Neste sentido podemos inferir que DS, é o atual termo para a ciência que analisa dados, combinando a estatística com ML, MD e as tecnologias de base dos dados para responder aos desafios que o big data na atualidade apresenta.

A informação é o resultado do processamento dos dados, portanto, estes somente terão valor para os negócios se quando coletados puderem ser tratados e analisados por meio de algoritmos de previsão e solução. Para haver valor nas tomada de decisões é preciso gerar de forma preditiva soluções e insights, que será feito pela DS através do processo computacional para facilitar a descoberta de padrões. Padrão é uma forma com uma configuração específica e facilmente reconhecível, que se caracteriza por uma regularidade, acumulação de elementos e repetição de partes. [CONWAY], defende que para além da matemática , sãos as tecnologias e os hackers os grandes responsáveis pela descoberta de micro padrões. [BREIMAN 2001] complementa que há duas culturas na modelação de dados. A primeira sendo a cultura dos micro padrões, onde há procura por pequena percentagem e possui como métricas o suporte e a confiança das regras associativas com origem no ML e nos hackers. E a segunda se refere a cultura dos macro padrões, onde utiliza a totalidade dos dados e com origens na matemática e na estatística.

Para que a DS se torne a alternativa mais viável de soluções é necessário que o profissional esteja muito bem preparado para realizar suas funções. Dentre as principais tarefas do profissional que atua com DS, o cientista de dados, de acordo com [ROLIM] estão: recolher grandes quantidades de dados de diversos tipos e fontes para trata-los, afim de transforma-los em informação útil; resolver os problemas relacionados aos negócio ou contextos bem definidos utilizando recursos e técnicas orientadas a dados; trabalhar com uma variedade de linguagens de programação que atenda aos requisitos DS; dominar conceitos estatísticos, incluindo distribuições e testes estatísticos; dominar e acompanhar técnicas analíticas como ML e DL; saber se comunicar com as equipes técnicas e de gestão; ter capacidade de descobrir critérios e ordem em padrões de dados e identificar tendências que venham contribuir para eficácia do negócio ou de domínio especifico.

3. Surgimento da Ciência de Dados

É de conhecimento que a matemática cientifica teve berço no século V e VI A.C. na Grécia. A estatística nasceu no XVII e as teorias de regressão linear teve origem em 1875 por Francis Galton. A computação começou seus passos em meados de 1940 com Turing e as redes neurais já começaram a ser idealizadas a partir 1943 com o neurofisiologista

Warren McCulloch e o matemático Walter Pitts que fizeram analogia entre as células nervosas e o processo eletrônico no artigo "A Logical Calculus of the Ideas Immanent in Nervous Activity". Em 1970 Edgar Frank da IBM apresentou o modelo relacional no artigo "A Relational Model of Data for Large Shared Data Banks". Em 1997, o termo "computação em nuvem" foi utilizado pela primeira vez na academia em uma palestra pelo professor de sistemas de informação, Ramnath Chellappa. Em 2000 a business intelligence atingiu seu ápice com a vasta disponibilidade de ferramentas e recursos. Enfim, da antiguidade a atualidade todo o descobrimento, melhoramento e evolução no que tange as áreas que formam a DS, seja em criação ou atualização de ferramentas, técnicas e métodos propriciam o aperfeiçoamento da própria DS. Não se sabe ao certo quando de fato a DS surgiu, no entanto de acordo com [DONOHO 2017] o termo originou-se em meados da década de 60 quando John Tukey defendia uma reforma das estatísticas acadêmicas no artigo "The Future of Data Analysis", publicado pela Princeton University e Bell Telephone Laboratories. Porém muitos especialistas da área aceitam que a expressão surgiu no início do século XXI por [CLEVELAND 2001]. Ambos cientistas atentavam para necessidade da realização de análises via estatística a partir dos dados e atribuindo, assim, características de ciência. Posteriormente a Cleveland, em abril de 2002, o Conselho Internacional para a Ciência começou a publicar o Data Science Journal junto ao seu Comitê de Dados para Ciência e Tecnologia (CODATA). Depois, em janeiro de 2003, a Columbia University começou a publicar também no Journal of Data Science. E em 2012 o título de cientista de dados teve sua terminologia cunhada por [DAVENPORT] e a partir daí a área de DS rapidamente vem se destacando e buscando aperfeiçoamentos.

Devido as necessidades do mundo dos negócios a tecnologia evolui para auxiliar as decisões empresariais. No entanto antes da DS algumas ferramentas e técnicas foram implementadas para auxiliar na tomada de decisão empresarial e também tiveram seu grau de contribuição para surgimento da DS. Assim como defende [PRIMAK 2008] nos anos de 1970 os Decision Support Systems (DSS) designavam um modelo genérico para tomada de decisão analisando um grande número de variáveis para tornar possível o posicionamento a cerca de uma determinada pergunta. Nos anos 80 os Executive Information Systems (EIS) davam suporte aos executivos para tomada de decisão fornecendo indicadores de desempenho e com fácil acesso a informações internas e externas. Nos anos 90 os Online Analytical Processing (OLAP) permitiram aos analistas de negócios, executivos e gerentes a análise e visualização dos dados corporativos de forma rápida, consistente e interativa, tanto nas atividades analíticas como nas navegacionais devido ao dinamismo e multidimensionalidade dos dados. Nos anos 2000 um conjunto de técnicas e ferramentas denominada de Business Intelligence (BI) se cosolidou para dar suporte a gestão de negócios através da transformação dos dados brutos em informações úteis por meio de processos de coleta, organização, análise, compartilhamento e monitoramento de um grande volume de dados que capacitava a interpretação dos dados e consequentemente a identificação, desenvolvimento e criação de novas oportunidades de estratégias. No entanto a BI e DS não devem ser compreendidas como iguais e nem como uma única ferramenta, mas sim como um conjunto de métodos e técnicas. Conforme [MATOS] enquanto o BI tem como objetivo converter dados brutos em insights de negócio para que os líderes empresariais possam tomar decisões, através dos analistas de negócios com a utilização de ferramentas para criar produtos de apoio à gestão, como relatórios e dashboards.A DS, por outro lado, emprega o método científico para a exploração dos dados,

com a formação e testes de hipóteses, por meio de simulação e modelagem estatística e com aplicação de ferramenta como o ML para automatizar a transformação dos dados em informação. A principal diferença entre os dois diz respeito a utilização dos dados. Apesar de BI poder utilizar métodos para previsão de futuro, esses métodos são gerados para fazer inferências simples a partir de dados históricos e atuais. Desta forma, BI extrapola o passado e o presente para inferir previsões sobre o futuro ao analisar os fatos que já tenham ocorrido em um determinado momento e se fundamenta em dados estruturados e exatos que já existem e assim apresentam informações relevantes do histórico e da momento atual para contribuir no monitoramento das operações de negócio e para auxiliar os gestores na tomada de decisões mais pontuais de curto a médio prazo, sem atentar para importância de predições em prazos longínquos e sem usufruir da base cientifica, mas objetivando obter painéis (dashboards), elaborar Key Performance Indicador e Key Risk Indicators ao utilizar ferramentas como Pentaho, Qlikview, Qliksense, Microstrategy, SAS Business Intelligence, Dundas, TIBCO Spotfire, Microsoft Power BI, Tableau e Oracle BI. Em contraste, os cientistas de dados fazem valer a ciência via análise exploratória de grandes quantidades de dados estruturados e não estruturados com métodos científicos, estatísticos e matemáticos avançados e algoritmos de ML para a extração de conhecimentos e insights preditivos relevantes ao negócio a longo prazo, sendo assim, capaz de identificar as tendências, realizar as previsões mais assertivas e utilizando ferramentas como R, Python, Octave, Matlab, Julia, Spark ML, Weka, Scala, Google ML, Amazon ML, Azure ML.

Desde os anos 1970 a cada dois anos a capacidade de processamento praticamente dobra e com grande aumento na quantidade de dados sendo gerada e manipulada aliada ao barateamento de armazenamento e da necessidade de análise e extração do grande volume de dados para obter informações úteis para as corporações impulsionaram a utilização da DS. Podemos destacar que entre os fatores que culminaram e alavancaram a DS então o big data, pois com grande volume de dados gerados com variedade e em alta velocidade percebeu-se a necessidade de criar formas inovadoras e econômicas para processar, organizar e armazenar os diferentes tipos de dados, a fim de compreender os dados e tomar as melhores decisões, e ainda de automatizar os processos. Outro fator importante conforme [MEYER 2017] foi o avanço na alta capacidade de processamento em nuvem aliada ao seu barateamento por meio de processamento horizontal com clusters, assim como também menciona [RAMOS 2018]. Sem esse aumento de processamento a DS certamente não existiria. Isso ocorre porque o processamento vertical tradicional é caro e ineficiente para grandes quantidades de dados. A especialização da capacidade computacional disponibilizada por fornecedores com enormes plataformas de computação na nuvem distribuídas paralelamente como Amazon (AWS), Google (GCP) e Microsoft (Azure) possibilitaram a locação de hardware sob demanda e a sua redistribuição para atingir máxima eficiência, contribuindo assim diretamente para o maior armazenamento dos dados. As tecnologias de nuvem além de permitirem que os dados fossem salvos em locais remotos e acessados por diversos dispositivos, também possibilitou aos usuários a execução de ferramentas para análise de primeira linha nos dados gravados nos mais diversos locais e ao mesmo tempo. A combinação das maiores capacidades de processamento nos computadores atuais com softwares inventivos fornecem ferramentas de ponta e a alta capacidade de armazenamento e com o barateamento tornaram os dados recursos renováveis, podendo ser combinados com outros conjuntos de dados e usados várias

vezes com diferentes perguntas, afim de gerar novas respostas e obter insights valiosos.

4. Desafios da Ciência de Dados

O big data, a internet das coisas, a inteligência artificial com ML e DL, a computação em nuvem, o aumento da capacidade de processamento, o barateamento de armazenamento, os algoritmos de ponta, as novas ferramentas e técnicas de análise e exploração contribuíram e continuam a contribuir para sucesso da DS, porém os desafios sempre existem, conforme [PROVOST and FAWCETT 2016] o fato do contexto tecnológico marcado por grandes quantidades de dados disponíveis e da impossibilidade da análise de forma manual excedia a capacidade de bases de dados computacionais mais usuais. A evolução tecnológica permitiu que computadores se tornassem cada vem mais poderosos, a conexão em rede mais ubíqua, o processamento mais rápido e os algoritmos com aprimoramentos possibilitaram a conexão de bases de dados e a realização de análises mais profundas. Se antes sofríamos com a falta de dados, hoje há abundancia nas mais diversas áreas de conhecimento e conforme [BELL et al. 2009b] o desafio aparente em comum se concentra justamente em como lidar de forma eficiente com a quantidade massiva de dados que cada área de conhecimento potencialmente produz diariamente e, sobretudo, em extrair o conhecimento relevante em benefício da própria área de conhecimento.

Os empecilhos da DS apresentam-se em diferentes aspectos, dentre eles podemos destacar a heterogeneidade, a fragmentação, o processamento, a disponibilidade, a privacidade e integridade e a falta de mão de obra qualificada. Todavia, estes desafios estão sendo enfrentados pela expertise dos cientistas e pelo próprio avanço tecnológico, que objetivam sempre o aperfeiçoamento e principalmente a eficácia em velocidades para análise e rápidas respostas. A heterogeneidade dos dados permitem a manipulação de diferentes formatos de dados e assim a realização de diferente níveis de acurácia se faz necessária, uma vez que, dificulta a extração e limpeza dos dados. A fragmentação dos dados fornece a possibilidade de armazenamento em múltiplas bases de dados e se utiliza de vários controladores e tomadores de decisão que tolhem o uso transparente e integrado dos dados, fazendo necessários obter a interoperabilidade entre as diferentes fontes de bases de dados. O processamento de dados apresenta desafios específicos na capacitação e viabilização de formas eficientes de realizar o gerenciamento dos dados para assim fornecer o devido acesso, tratar os diferentes níveis de dados, propiciar agilidade nas consultas e permitir o compartilhamento. Segundo [MANETH and POULOVASSILIS 2016] a alta produção de dados por empresas, aplicações científicas e mídias sociais fazem surgir a necessidade do desenvolvimento de técnicas computacionais capazes de escalonar os grandes volumes e as variedades de dados, que são gerados por meio de tecnologias baseadas em Web, móveis e difusas. Também é necessário que haja uma redução de complexidade temporal de quase todos os algoritmos existentes, desde o cálculo da variância em estatística até ao mais complexo problema de sequence mining, para assim permitir a realização mais eficaz de buscas exploratórias e a obtenção de repostas rápidas.

A Qualidade dos dados garante agilidade de extração e da análise de dados. Como os dados estão disponíveis em formatos diferentes, estruturados ou não, e localizados em diferentes fontes, as ferramentas utilizadas para o devido processamento ira combinar dados inconsistentes que acarretarão em duplicidade, conflitos lógicos e informações faltantes. A má qualidade dos dados leva a relatórios e análises defeituosas, que podem prejudicar o processo de tomada de decisões. Ademais, a analise precisa que os dados

sejam previamente preparados, sendo que atualmente um cientista de dados usufruí da maior parte do tempo de trabalho para coletar, limpar e organizar dados de interesse. É estimado conforme [SEPUVIDA et al. 2016] que 50% a 80% do tempo total gasto para realização de um projeto DS é na preparação dos dados digitais desorganizados para que possam ser explorados em busca de conteúdo útil. Nesse sentido, a qualidade dos dados é um aspecto crucial, espera-se que novas ferramentas possam ser desenvolvidas de modo a permitir melhor extração de diferentes dados oriundos de diversas fontes heterogêneas e em distintas escalas espaço-temporais.

É inegável a abundância de dados, porém não necessariamente é permitido o acesso e uso de todo e qualquer dado. Então a disponibilidade de dados se apresenta hoje como desafio, uma vez que, devido à sensibilidade de dados há diversas políticas extremas de proteção, seja por razões comerciais ou culturais, sendo estabelecidas com a justificativa de privacidade. Segundo [BSA] alguns países têm como condição de mercado políticas que restringem o livre fluxo de dados e exigem que os servidores fiquem somente na jurisdições e territorio de domínio do estado. Tais restrições prejudicam as enormes eficiências de escala, pois inibem os benefícios econômicos que podem surgir com a inovação em dados e a capacidade de combinar conjuntos de dados diferentes para fazer descobertas significativas, a partir de uma crescente abundância de dados. Desta forma, também podem vir a prejudicar a segurança ao impedir que os dados valiosos realizem backup em diferentes geolocalizações. Deixando-os, assim, desprotegidos em caso de falhas técnicas ou desastres naturais. Contudo as leis dos países não precisam ser iguais, mas sim se convergirem para serem mais compatíveis. Permitir, portanto, o livre fluxo de dados entre estados é um princípio básico para possibilitar os benefícios gerados pelos dados e as regulamentações que buscam unicamente tratar a privacidade e segurança exigindo que os dados sejam armazenados localmente inibem e limitam os tipos de benefícios sociais que a inovação em dados pode oferecer ao próprio pais e ao mundo. Entretanto garantir a segurança, privacidade e propriedade de dados das organizações e dos usuários é fundamental. As empresas tem se dedicado cada vez mais na atualidade com área de segurança, porém é notório o alto risco de uma grande quantidade de dados confidenciais serem manipulada sem a devida autorização e principalmente sem o conhecimento consciente dos usuários. A realidade se apresenta com vazamento de dados sigilosos, seja por descuido de alguns cientistas ou por invasões propositais assim como previu [CHIAVEGATTO and ALEXANDRE 2015]. No Brasil para a proteção de dados e da privacidade a lei geral de proteção aos dados (LGPD) direciona as empresas para reverem as estratégias existentes ou para criar estratégias de dados alinhadas com a lei e o uso ético dos dados. A LGPD tem como diretriz a prevenção de corrupção e inibição dos vazamentos e a sua utilização ética. No entanto, a solução para o problema está além da conscientização ética dos cientistas e de seus empregadores sobre a real importância da privacidade. Sendo cada vez mais comum a análise de dados exclusivamente dentro de um terminal de acesso restrito, os profissionais devem ser comprometidos com o desenvolvimento de protocolos de segurança e a criação de novas técnicas que garantem o direito e sigilo dos detentores dos dados. Assim, o cientista de dados é o profissional responsável não apenas pelo processo de análise em si, mas por garantir que os dados estejam atendendo a todos os regulamentos cabíveis. Neste sentido, devem ser incorporadas técnicas de criptografias em projetos e pesquisas de DS, porém é necessário levar em conta a disponibilidade dos dados, escândalos e vazamentos, apesar de certamente

prejudiciais para as vítimas, não podem ser utilizados para restringir todos os acessos ao dados, tais crimes devem ser apurados e responsabilizados legalmente e judicialmente, se for o caso, os envolvidos. Afinal não se pode esquecer dos benefícios que a DS traz para a sociedade com ganhos econômicos, financeiros, sociais, além de conseguir salvar vidas.

Hoje, há uma escassez global de analistas e gerentes com capacidade de compreensão dos dados, de acordo com [BSA] so nos EUA há uma escassez de pessoas com talento analítico avançado necessário para revelar o potencial oculto nos dados entre 140 e 190 mil e de gerentes e analistas capazes de compreender e tomar decisões com base na análise da economia dos dados em 1,5 milhão. É preciso, então, possuir habilidades técnicas para a gestão de problemas complexos, mas também conhecimentos em inteligência de dados. Muitas vezes, só é possível compreender os insights contidos nos dados por meio da criatividade humana. Conforme [DAVENPORT] existe a demanda crescente por um novo perfil profissional que deve ser treinado nas habilidades imprescindíveis para fazer descobertas a partir de um grande conjunto de dados. Os profissionais especializados devem possuir a capacidade de realizar a seleção de modelos de simulação e estatística e realizar a implemenção das práticas da DS em um negócio entregando o produto final gerado pelos dados. É fundamental que os profissionais possuam o conhecimento sobre o ramo do negócio em questão, dada a necessidade de coletar os dados especificos e importantes para o setor. O profissional em ciência de dados bem qualificado precisa de uma boa base em computação, matemática aplicada, estatística, modelagem, além de possuir conhecimento do domínio alvo de aplicação para, assim, estar apto a desenvolver soluções inovadoras.

Em suma, há diversos desafios envolvendo DS e muitos destes já estão sendo tratados em diferentes níveis para apresentar soluções tecnológicas cada vez mais satisfatórias e que trarão segurança e respostas ainda mais rápidas e menos custosas, economizando tempo e possibilitando, portanto, que grandes volumes de dados complexos sejam extraídos e analisados de forma mais eficaz.

5. Futuro da Ciência de Dados

À medida que os dados passaram a ser cada vez mais abundantes, com o baixo custo de armazenamento, com o aumento da capacidade de processamento em nuvem e do número de sensores que capturam dados, a quantidade de dados que já é, será juntamente com as possibilidades de uso destes dados ainda maiores, uma vez que, novas fontes de dados continuarão surgindo cada vez mais pela internet da coisas (IOT), onde a interconexão entre dispositivos continuará a crescer no futuro, resultando em mais conexões entre diferentes tipos de dispositivos eletrônicos. A DS impulsionará a solução de novos problemas perante os desafios do mundo em que as oportunidades geradas pelos dados surgem para uma maior eficiência dos recursos na saúde, no transporte, na energia, na economia, na segurança , em todas as áreas possíveis. Sendo os dados a matéria-prima dos cientistas de dados e quanto maior a fartura de matéria-prima maior se dará a utilização dos dados. Assim a DS nos direciona para possíveis aplicações e respostas futuras que valem a pena discutir.

A DS com geração automatizada de conteúdo e com agregação e classificação de conteúdo poderá a vir ser utilizada para várias finalidades futuras como, por exemplo, na correção automatizada de ensaios de estudantes para detecção de plágio, na otimização

do tráfego automóvel, na identificação, seleção e manutenção de empregados ideais, nas auditorias automatizadas para evitar litígios dispendiosos e perda de tempo, melhorias no planejamento urbano, no combate ao cibercrime, em julgamentos para fortalecer a evidência ou falta contra um réu em um tribunal, no fortalecimento da agricultura de precisão, onde os agricultores promovem o crescimento eficiente dos alimentos, bem como a sua entrega, reduzindo o desperdício de alimentos e nas organizações sem fins lucrativos que serão capazes de prever as necessidades de captação de recursos e aumentar seus esforços de forma inteligentes. Há, portanto, uma infinidade de aplicações futuras que se limitam apenas à capacidade do penssamento humano. Muitas das áreas já contém inumeros estudos e projetos em adamento com DS corforme [ZIVIANI and PORTO 2014], no entanto ainda requerem aperfeiçoamentos. A saúde pública poderá ser aprimorada com a utilização de DS, por meio de rastreadores vestíveis e IOT, que instigarão as pessoas a adotarem hábitos mais saudáveis e irão avisa-las sobre possíveis problemas de saúde futuros e também novas doenças poderão ser descobertas e evitadas com a apuração e analise do histórico dos pacientes hospitalares, aumentado assim a expectativa de vida. Complementa [MARCOS] o papel da IOT no futuro da saúde trará mais confiança aos médicos e melhor controle das doenças e dos medicamentos, permitirá o atendimento remoto ao paciente e facilitará a manutenção de dispositivos médicos. Na área de recursos humanos a DS poderá ser aplicada tanto na identificação de possíveis demissões quanto na contratação. No recrutamento a identificação dos perfis de candidatos a ocupar alguma vaga, de modo a melhor se enquadrar no perfil solicitado, será realizado com DS via análise dos dados dos candidatos com base no histórico das causas de pedido de demissão e através dos modelos preditivos permitirá obter uma lista com os colaboradores com maior probabilidade de desligamento futuro com a empresa. Compremenda [SANTANAEM] que a DS trará grande beneficios ao RH, um vez que, a seleção de funcionários se dará com o perfil mais alinhado e adequado aos objetivos empresariais com habilidades e fatores que não seriam considerados pela seleção tradicional e aumentará a rapidez na análise e seleção, pois todos os currículos serão analisados juntos.

Atualmente as empresas utilizam principalmente dados de compra, vendas, dados de fluxo de cliques, mas futuramente será necessário que as empresas estejam cada vez mais bem informadas, e além de moldarem suas estratégias de acordo com os insights deveão ser capazes de incluir para analise os dados capturados das mais variadas fontes, como fluxos de fabricação, ambientes de varejo, funcionários e até mesmo, se possível, bases de dados governamentais e os das concorrentes, para assim obter resultados de analises mais satisfatórias e vantagens competitivas.

As melhorias nos algoritmos personalizados se tornarão ainda mais importantes. Conforme [SANTOS et al. 2009] o sucesso de aplicações e a necessidade de soluções melhores, mais rápidas, mais escaláveis e mais precisas para a obtenção de conhecimento a partir da Web motiva investimento de empresas, a participação de cientistas e o envolvimento de professores, estudantes e profissionais que usam a Web para coletar ou fornecer informações. Assim como empregar técnicas de otimização apriori como defende [DE AMO 2004] para melhorar a performance de algoritmos. Os cientistas de dados serão futuramente capazes de criar uma estratégia de dados individualizada e os correspectivos algoritmos de MD visando o sucesso nos negócios com base nos objetivos organizacionais exclusivos de uma empresa. Em decorrencia do desenvolvimento de algoritmos otimizados, funções avançadas poderão ser criadas para fornecer soluções automatizadas

e feedback à medida que os dados já forem sendo coletados. Com algoritmos evolutivos e a mudança do ML, onde o foco se dará na mecânica do ML para estimular a criatividade e a utilização de vários tipos diferentes de modelos. Assim com a nova maneira de praticar o ML os cientistas de dados terão a responsabilidade de traduzir as mudanças em novas soluções que acrescentem valor às empresas. Por fim os algoritmos da DS cada vez mais complexos serão incluídos em pacotes e tecnologias que tornarão suas ordens de magnitude mais fáceis de implementar, uma vez que a experiência de treinamento e implantação poderá ser realizada com um maior grau de qualidade e com menos conhecimento técnico e estatístico.

Os avanços da tecnologia cada vez maiores são fortemente impactados pela DS. A computação quântica, por sua vez, conforme [PACHECO and DISCONZI 2019] promete revolucionar a indústria da computação e consequentemente trará nova soluções a DS. Permitirá realizar o treinamento de modelos 1000 vezes mais rápido do que é feito atualmente, em escalas exponenciais. Com os computadores e processadores quântico novos algoritmos serão necessários e resultarão na criação de processos de automação quase que instantâneos da DS, gerando ainda mais valor para as empresas. E, por consequência, novos frameworks irão surgir e permitirão absorver mais insights dos dados. Criando, assim, um círculo de inovação virtuoso.

Para maximizar as oportunidades criadas pela DS de forma a acelerar de forma positiva a produtividade, o crescimento econômico e os benefícios individuais, é necessário dar prioridade à inovação e preparar o caminho para as soluções avançadas e os poderosos impactos que a DS pode ocasionar. Cabe aos cientistas de dados criar futuros modelos para potencializar a produtividade e eficiência de todas as áreas. Não existe restrição sobre nenhuma área para a atuação da DS e como a abundância de dados direciona a novos desafios e oportunidades, basta que o cientistas de dados sejam cada vez mais inovadores, criativos e responsáveis para garantir que a evolução da DS traga consigo as respostas aos problemas e aos maiores desafios que ainda estão por surgir.

6. Aplicações da Ciência de Dados

Com ferramentas avançadas a DS realiza previsões que solucionam grandes problemas, melhoram a vida das pessoas e beneficiam as corporações. Como os dados são a nova chave de poder para o mercado desta nova era de integração e inovação, se faz necessária a implementação com a DS para manter o papel de liderança, uma vez que, a nível empresarial converter dados em ações garante a evolução do próprio negócio. As empresas orientadas a dados como Amazon, Google, Netflix, IBM, Facebook, Youtube e Instagram utilizam e investem fortemente tecnologias da DS para tomar decisões importantes. As empresas de busca na internet como Google, Yahoo, Bing, Ask, AOL, DuckDuckGo fazem uso de algoritmos de DS para entregar o melhor resultado da consulta de forma extremamente rápida. O Google segundo [FRAGAH] em 2008 já processava 20 petabytes de dados diariamente, sem a DS seria incapaz de processar a quantidade de informações que recebe por segundo. Devida a infinidade de atuação da DS será exemplificado apenas os setores de agronegocio, saúde, logistica e também apresentado algumas das diversas aplicações que já obtiveram resultados de grande valia para as organizações e aos usuários de um determinado serviço.

No agronegócio a inovação em dados aumenta a produtividade no campo e nos

processos industriais e ainda reduzir custos. Algumas das aplicações que utilizam DS estão fenômica [GIGLIOTI et al.], genômica [SANTOS et al. 2009], e principalmente a agricultura/pecuária de precisão [DE MIRANDA et al. 2017], que para aumentar a eficiência das técnicas de controle de danos, a produtividade e reduzir os cultos envolvidos faz as recomendações dos melhores uso de sementes e das aplicação de defensivos e fertilizantes em áreas especificas. Com o uso dos dados captados no campo de sementes juntamente com algoritmos avançados, satélites e sensores se torna possível tomar decisões melhores visando em conjunto com o aumento da produtividade e qualidade dos alimentos a redução do desperdício de insumos e o consumo de água. Por exemplo, ao utilizar ferramentas de análise de dados, conforme [BSA] os agricultores reduzem os custos com insumos, o uso de pesticida e produtos químicos e melhoram a produtividade em 5 a 10 vezes e para os produtores de laticínios da Croácia criou-se uma plataforma analítica de software na nuvem que fornece dados em tempo real sobre itens como o impacto da qualidade da ração na produção e as taxas de concepção de cada animal, o que, por sua vez, aumentou a produtividade e a eficiência em até 50%. Com a DS é possível análise de dados para a recomendação assertivas de cultivo. Seria enorme os ganhos obtidos com uma melhor utilização de dados, uma vez que, um quarto da população mundial está no setor da agricultura e produção de alimentos. A DS ajuda a tomar decisões melhores sobre o que cultivar, quando plantar, como monitorar o frescor dos alimentos da fazenda ao prato do consumidor e ainda como adaptar o cultivo as possíveis mudanças climáticas. [RAVI and GOPAL 2017] apontam a importância da DS para o setor produtivo, considerando a adoção sensores para captura, técnicas de integração, robôs e drones para criação do modelo, chamado Smart Farm, onde se torna possível entender todos os fatores que afetam a produtividade e a geração de perdas no âmbito das cadeias produtivas.

Na área da saúde a DS esta na personalização de tratamentos e na detecção antecipada de doenças através de análise genética. Entre alguns exemplos de aplicados DS, conforme [BSA], estão : rastreando e correlacionando mais de 1000 pontos de dados por segundo, pesquisadores canadenses surpreenderam médicos ao descobrir que crianças nascidas prematuramente com sinais vitais estranhamente estáveis relacionavam-se a grave casos de febre no dia posterior, permitindo assim que medidas preventivas fossem tomadas para salvar a vida ao combinar dados em tempo real com o histórico médico de pacientes. Também foi desenvolvido algoritmo de aprendizado de máquina capaz de prever paradas cardíacas com quatro horas de antecedência e com 66% de precisão. Na Angola reportagens de jornais de um período de duas décadas estão sendo usadas para prever quando epidemias de cólera ocorrerão. No Quênia, através dos dados móveis estao identificando padrões de infecção por malária e definindo os locais de contágio para orientarem os esforços de erradicação do governo. Dados de consultas médicas e informações sobre medicamentos receitados mostram que pacientes com doenças autoimunes correm risco maior de epilepsia. Os hospitais estão implementando sistemas de apoio a decisões clínicas, onde são analisados dados de diversas fontes para realizar o diagnósticos mais rápidos e confiáveis em um complexo ambiente de dados, comprovando-se eficácia em mais de 70% dos casos. Segundo [KENT] os cientitas de dados Andrew Satz e Brett Averso, formados no Data Science Institute da Columbia University dos EUA estão usando algoritmos de ML capazes de gerar, rastrear e otimizar computacionalmente centenas de milhões de anticorpos terapêuticos para encontrar rapidamente tratamentos para o Covid-19. Conforme [MANYIKA et al.] caso os dados fossem utilizados com

mais eficácia para aumentar a qualidade e eficiência da saúde é estimado que este setor poderia economizar mais de 300 bilhões de dolares por ano ao reduzir os gastos em 8%. Portanto, os maiores impactos da DS na saúde não são apenas no ambito econômico, mas abrangem a capacidade de salvas vidas humanas e de aumentar suas expectativas.

No que tange a logística, conforme [SARASWAT] as empresas DHL, FedEx, UPS, Kuehn, Nagel melhoraram a sua eficiência operacional quando começaram a utilizar a DS nos dados gerados pelos GPS instalados nos transportes. Com os modelos preditivos criados as transportadoras obtiveram uma maior eficiência de custos, pois foi possível identificar as melhores rotas de envio, o tempo mais adequado de uma entrega e o melhor modo de transportes.

A DS esta empregada no varejo, esportes, educação, entretenimento, biotecnologia, astronomia, telecomunicações, transporte, energia, finanças, economia, industria de produção, turismo, física, entre outras. Mesmo sendo considerada um conceito realtivamente novo a DS já se apresenta em diversas áreas e setores. Algumas das aplicações mais atuais da Ds estão: propaganda, jogos, recomendação, reconhecimento de imagem e voz, planejamento de rotas aéreas, comparação e geolocalização.

Nos anúncios multiplataforma direcionados em realtime e exibidos nos websites, onde o conteúdo dos anúncios são escolhidos por meio de algoritmos, que com base na análise dos dados individuais de cada consumidor seleciona aquele com a maior probabilidade de interesse. Por considerar o comportamento particular de cada visitante, a taxa de cliques é maior do que a de anúncios tradicionais.

Nos Jogos a DS fornece experiência única e personalizada para cada pessoa. As empresas Sony, EA Sports, Activision-Blizzard, Zynga e Nintendo estão criando os jogos com algoritmos de ML que se melhoram e se atualizam enquanto o jogador passa de fases. Em jogos de ação e aventura, o computador (adversário) molda o seu próprio jogo com a análise dos movimentos realizados anteriormente pelo jogador, adaptando assim ao estilo único de cada pessoa. Conforme [CROCOMO and SIMOES 2008] um algoritmo evolutivo de aprendizado cria estratégias inteligentes e adaptativas que não são controladas pelo jogador.

Nos sistemas de recomendação as atuais gigantes da internet utilizam DS para melhorar a experiência do usuário promovendo seus produtos e oferecendo sugestões de acordo com o interesse do visitante e da relevância da informação. As recomendações são feitas com base em resultados de pesquisas anteriores de cada usuário. Entres as gigantes estao corforme [SARASWAT] a Amazon, Twitter, Google Play, Netflix, Linkedin e IMDB. Com a personalização, por exemplo, a rede Social Facebook faz recomendações de novos prováveis contatos, empresa Netflix recomenda filmes a seus clientes de acordo com os gostos e histórico individuais. As Empresas E-commerce fazem sugestões que além do histórico abrangem os hábitos e as associações entre os produtos.

No reconhecimento de imagem a análise automatizada de imagens acelera a detecção de doenças e reduz o tempo de espera hospitalar, auxilia na busca de pessoas perdidas, no reconhecimento de criminosos, facilita a análise do padrão de consumo dos clientes e marcação de ponto nas empresas. Em virtude do rápido processamento de imagens é possível obter soluções na computação cognitiva, na neurobiologia, na robótica e na detecção de câncer. Devido a aparência e forma imprevisíveis, os tumores cerebrais

são difíceis de identificar nas imagens médicas, mas com a computação em nuvem e algoritmos avançados de análise de imagens segundo a [BSA] os cientistase estão buscando identificar tumores cerebrais com mais rapidez e precisão. Outros exemplos onde DS é empregada está no Facebook quando se publica uma foto, e logo em seguida se tem a opção de marcar as pessoas presentes e no Google que permite a realização de pesquisa por imagem. Já No reconhecimento de fala DS se faz presente, por exemplo, nas tecnologia Cortana (Microsoft), Siri (Apple) e Alexa (Amazon) que estão investindo pesado no uso de DL para o reconhecimento de voz. Essas tecnologias conversacionais permitem ao usuário interagir com uma inteligência artificial por meio de comandos de voz revelando de forma bastante compreensiva como funciona a transformação entre dados não estruturados (voz) em informações úteis (comandos computacionais).

No planejamento de rotas aéreas, segundo [SARASWAT] as companhias Southwest Airlines, Alaska Airlines com o objetivo de manter as taxas de ocupação e os lucros operacionais estão entre as empresas que passaram a utilizar DS para identificar áreas estratégicas que requerem melhorias como a prevenção de atrasos em voos, decisões acerca de que aviões comprar, estabelecimento de rotas diretas ao destino ou com conexões, gerenciamento mais eficazes de seus programas de fidelização e informando às equipes de manutenção quais peças precisam de reposição antes de falharem. Correlacionando dados históricos de uma década de voos e padrões climáticos, passageiros de companhias aéreas podem descobrir quais voos têm menor probabilidade de sofrer atrasos. Conforme [BSA] com a capacidade de capturar dados em tempo real para melhorar a eficiência dos motores e definir rotas com mais eficácia, uma economia de combustível de apenas 1% significaria economias de 30 bilhões de dolares ao longo de 15 anos.

Nos Websites de comparação de preço é possível em um só lugar comparar os valores de diversos fornecedores de uma só vez. Alguns exemplos de sites conforme [RAJ 2019] estão PriceGrabber, PriceRunner, DealTime, Junglee e Shopzilla. Os buscadores de preço revolucionaram e aumentaram as compras online quando com DS passaram a analisar e comparar os dados recebidos de diversas fontes como das interfaces de programação de aplicações e dos feeds RSS.

Na geolocalização por meio de mapas a DS pode der utilizada para identificação de padrões geográficos de sinistralidade, doenças, vendas, reclamações. O Facebook, por exemplo, utiliza da cidade natal e a localização atual do perfil do usuário para analisar as localizações e identificar padrões de migração global e encontrar os fã-clubes dos times de futebol conforme [GRUS].

Todas as aplicações trouxeram benefícios seja social, econômico, ecológico, informacional, lazer, melhoria na qualidade e na expectativa de vida. As aplicações exisententes poderão vir a sofrer melhorias conforme o avanço tecnologico e muitas outras irão ainda surgir, assim como defende [ZIVIANI and PORTO 2014]. Por fim, as aplicabilidade de DS se divergem para todas as possíveis áreas e se restringem unicamente ao conhecimento e a inteligência dos cientistas de dados.

7. Conclusão

Os dados são uma fonte essencial de benefícios sociais e econômicos para o mundo e com a DS, mesmo perante alguns desafios, propriciou a obteção de soluções rápidas para atender as perspectivas mais reais das organização sobre os as operações, produtos, serviços e

as suas relações com os clientes, fornecedores e concorrentes. Futuramente com a possibilidade de novas e poderosas ferramentas mais avançadas de análise de dados e profissionais qualificados a DS conseguirá realizar previsões ainda mais assertivas que resolvem grandes problemas e que melhorarão muito a vida cotidiana das pessoas. A inovação em DS permitirá que clientes tomem melhores decisões e que empresas personalizem produtos e serviços para atender e entreter melhor seus consumidores. O futuro da DS nos direciona de uma economia fundamentada em produção em massa para uma baseada na personalização.

Referências

AMARAL, F. (2016). *Aprenda Mineração de Dados: Teoria e prática*. Alta Books Editora, 1 edition.

BELL, D. C., LEMME, M. C., STERN, L. A., WILLIAMS, J. R., and MARCUS, C. M. (2009a). Precision cutting and patterning of graphene with helium ions. *Nanotechnology*, v.20(n.45).

BELL, G., HEY, T., and SZALAY, A. (2009b). Beyond the data deluge. *Science*, v.323(n.5919):p.1297–1298.

BREIMAN, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, v.16(n.3):p.199–215.

BSA. Qual é o x da questão com relação a dados?, the software alliance, 2015, disponível em (https://data.bsa.org/wp-content/uploads/2015/10/bsadatastudy_br.pdf).

CHIAVEGATTO, F. and ALEXANDRE, D. P. (2015). Uso de big data em sao paulo no brasil: perspectivas para um futuro. *Epidemiologia e Serviços de São Paulo*, v.24:p.325–332.

CLEVELAND, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, v.69(n.1):p.21–26.

CONWAY, D. The data science venn diagram, drewconway.com, drewconway, 2010, disponível em (http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram).

CROCOMO, M. K. and SIMOES, E. (2008). Um algoritmo evolutivo para aprendizado on-line em jogos eletrônicos. *Proceedings of SBGames*.

CURTY, R. G. and CERVANTES, B. M. N. Data science: Ciência orientada a dados. informação informação, uel, 12/2016. disponível em: (http://www.uel.br/revistas/uel/index.php/informacao/article/view/27929/20119). note = Acesso em: 25 jun de 2020.

DAVENPORT, Thomas H.and PATIL, D. J. Data scientists: the sexiest job of the 21st century. harvard business review, [s.l.], v. 90, n. 10, p. 70-76, oct. 2012. disponível em: (https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/).

DE AMO, S. (2004). Técnicas de mineração de dados. *Jornada de Atualização em Informatica*.

DE MIRANDA, A. C. C., VER´SSIMO, A. M., and CEOLIN, A. C. (2017). Agricultura de precisão: Um mapeamento da base da scielo. *Gestao.org*, v.15(n.6):p.129–137.

DONOHO, D. (2017). 50 anos de ciência de dados. *Journal of Computational and Graphical Statistics*, v.26(n.4).

EGHAM. Gartner says 8.4 billion connected things will be in use in 2017, up 31 percent from 2016, gartner, 2017, disponível em (https://www.gartner.com/en/newsroom/press-releases/2017-02-07-gartner-says-8-billion-connected-things-will-be-in-use-in-2017-up-31-percent-from-2016). Acesso em: 25 de jun de 2020.

EMMERT-STREIB, F., MPUTARI, S., and DEHMER, M. (2016). The process of analyzing data is the emergent feature of data science. *Frontiers in Genetics*, v.7:p.12.

FRAGAH, R. Google processa 20,000 terabytes de informação por dia , google discovery , 09/01/2008 disponível em (https://googlediscovery.com/2008/01/09/google-processa-20000-terabytes-de-informacao-por-dia/). Acesso em: 27 de jun de 2020.

GIGLIOTI, A., SUMIDA, C. H., and CANTERI, M. G. Fenômica de doenças.

GRUS, J. *Data Science From Scratch: First Principles with Python*. 1 edition.

KENT, J. Cientistas de dados usam o aprendizado de máquina para descobrir tratamentos covid-19, health itanalytics, 25/03/2020, disponível em (https://healthitanalytics.com/news/data-scientists-use-machine-learning-to-discover-covid-19-treatments). Acesso em: 27 de jun de 2020.

MANETH, S. and POULOVASSILIS, A. (2016). Data Science. *The Computer Journal*, v.60(n.3):p.285–286.

MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C., and ANGELA, H. B. Big data: The next frontier for innovation, competition, and productivity, mckinsey global institute , 01/05/2011, disponível em (https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation). Acesso em: 27 de jun de 2020.

MARCOS. Iot nos hospitais: Impactos e vantagens para a Área da saÚde,biocam, 01/08/2020, disponível em (https://www.biocam.com.br/category/iot/). Acesso em: 27 de jun de 2020.

MATOS, D. Business intelligence x data science. cienciaedados, 2015. disponível em: (http://www.cienciaedados.com/business-intelligence-x-data-science/). Acesso em: 26 de jun de 2020.

MEYER, L. A. V. C. (2017). Uma visão geral dos sistemas distribudos de cluster e grid e suas ferramentas para o processamento paralelo de dados. *IBGE [sd]. Disponvel em (https://ww2. ibge. gov. br/confest_e_confege/pesquisa_trabalhos/CD/palestras/368-1. pdf).*, v.8:p.30.

PACHECO, B. B. M. and DISCONZI, M. S. (2019). Ciência de dados : desafio de abordagem não processo. *Pesquisa, Sociedade e Desenvolvimento*, v.8(n.11).

PRIMAK, F. V. (2008). *Decisões com bi (business intelligence)*. Fabio Vinicius Primak.

PROVOST, F. and FAWCETT, T. (2016,). *Data Science para Negócios,*. Alta Books,, 1 edition.

RAJ, P. (2019). *Novel Practices and Trends in Grid and Cloud Computing*. IGI Global.

RAMOS, H. D. S. (2018). Computação em cluster as vantagens do uso de cluster de alto desempenho em empresas.

RAVI, A. K. and GOPAL, C. R. (2017). Smart technology in farming development. *International Journal of Management (IJM)*, v.8(n.2).

REINSEL, D., GANTZ, J., and RYDNING, J. The digitization of the world, seagate, 2018, disponível em (https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf). Acesso em: 25 de jun de 2020.

ROLIM, M. V. 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, disponível em: (https://bdm.unb.br/handle/10483/20898). Acesso em: 25 de jun de 2020.

SANTANAEM, F. Data science e big data aplicado no rh, 01/29/2020, disponível em (https://minerandodados.com.br/data-science-e-big-data-aplicado-no-rh/). Acesso em: 27 de jun de 2020.

SANTOS, R. et al. (2009). Conceitos de mineração de dados na web. *Anais do XV Simposio Brasileiro de Sistemas Multimidia e Web e VI Simpósio Brasileiro de Sistemas Colaborativos*, pages p.81–124.

SARASWAT, M. 13 applicações práticas de data science hoje , vooo, 21/04/2016, disponível em (https://www.vooo.pro/insights/13-applicacoes-praticas-de-data-science-hoje/). Acesso em: 27 de jun de 2020.

SEPUVIDA, W. R. et al. (2016). Predição de evasão na educação a distância como subsídio a tomada de decisão.

TIERNEY, B. Data science is multidisciplinary, oralytics, 2012, disponível em: (https://com/2012/06/13/data-science-is-multidisciplinary/).

VANDERPLAS, J. (2016,). *Jake. Python Data Science Handbook,*. O'Reilly Media,, 1 edition.

ZIVIANI, A. and PORTO, F. (2014). Ciência de dados, in: 3o. seminario de grandes desafios da computção no brasil, sbc.