UNIVERSIDADE ESTADUAL DE CAMPINAS – UNICAMP
INSTITUTO DE ECONOMIA – IE
GRADUAÇÃO EM CIÊNCIAS ECONÔMICAS

A TECNOLOGIA DO BIG DATA E SEUS IMPACTOS NO PROCESSO DE TOMADA DE DECISÃO

Bárbara Pereira Zidan

Campinas - 2020

BÁRBARA PEREIRA ZIDAN

A TECNOLOGIA DO BIG DATA E SEUS IMPACTOS NO PROCESSO DE TOMADA DE DECISÃO

Trabalho de conclusão de curso a ser apresentada ao Instituto de Economia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de bacharel em Ciências Econômicas.

Orientador: Prof. Dr.
Este trabalho corresponde à versão parcial do trabalho de conclusão de curso a ser defendida pelo aluno Bárbara Pereira Zidan orientada pelo Prof. Dr. Marcio Wohlers de Almeida

Assinatura do Orientador

CAMPINAS

2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Economia
Luana Araujo de Lima - CRB 8/9706

Z621t

Zidan, Bárbara Pereira, 1995-
A tecnologia do Big Data e seus impactos no processo de tomada de decisão / Bárbara Pereira Zidan. – Campinas, SP : [s.n.], 2021.

Orientador: Marcio Wohlers de Almeida.
Trabalho de Conclusão de Curso (graduação) – Universidade Estadual de Campinas, Instituto de Economia.

1. Big data. 2. Processo decisório. 3. Ciência de dados. 4. Mineração de dados (Computação). I. Almeida, Marcio Wohlers de,1949--. II. Universidade Estadual de Campinas. Instituto de Economia. III. Título.

Informações adicionais, complementares

Palavras-chave em inglês:
Big data
Decision making
Data science
Data mining
Titulação: Bacharel em Ciências Econômicas
Data de entrega do trabalho definitivo: 26-01-2021

SUMÁRIO

Resumo

O Big Data faz referência ao grande volume, variedade, velocidade com que os dados estão disponíveis hoje em dia. Por isso, há uma demanda cada vez maior por formas inovadoras para o processamento de dados com o objetivo de transformar toda a informação compilada em recurso para a tomada de decisão. Essa mudança no processo de tomada de decisão é recente, pouco explorada e com acelerado ritmo de mudança. O trabalho pretende elucidar algumas questões envolvendo essa mudança de paradigma. Como podemos entender o Big Data? Qual o valor de entendê-lo? Como os dados são processados e qual o papel do usuário? E, finalmente, qual são os limites da exploração desse novo ativo econômico?

Palavras-chave: Big Data, tomada de decisão, tomada de decisão baseada em dados.

Abstract

Big Data refers to the large volume, variety and speed in which data is available today. There is an increasing demand for innovative ways for data processing with the aim of transforming all the information compiled into a resource for decision making. This change in the decision-making process is recent, little explored and with an accelerated pace of change. The work intends to elucidate some questions involving this paradigm shift. How can we understand Big Data? What is the value of understanding it? How is the data processed and what is the user's role? And finally, what are the limits to the exploitation of this new economic asset?

Keywords: Big Data, decision making, data driven decision making.

1. Introdução

A análise preditiva pode ser definida como o processo de coletar, organizar e analisar grandes conjuntos de dados que orientem padrões geradores de informações úteis. Para isto, são usadas técnicas estatísticas de mineração de dados, modelagem preditiva e aprendizado de máquina (“machine learning”) de forma a analisar dados atuais ou históricos que orientem previsões sobre eventos futuros ou desconhecidos.

Com o advento e progresso da internet aumentou-se de forma exponencial a disponibilidade de dados. Esse volume e variedade de dados extrapolou a capacidade de análise manual e tornou obsoleto o uso de base de dados tradicionais. Essa nova base de dados com volume exponencial é comumente conhecida atualmente como Big Data.

Da mesma forma que as redes se tornam mais complexas, as tecnologias se desenvolvem de forma cada vez mais poderosa e possibilitam análises mais profundas e com maior alcance. A convergência desses fenômenos acabou por criar e impulsionar uma nova área no campo da tecnologia, a ciência de dados.

Essa área é responsável por estruturar, organizar e tratar dados das mais diversas fontes usando habilidades em Matemática, Estatística e Machine Learning com o objetivo último de gerar entendimento e conhecimento profundo no campo que se escolheu analisar. Muitos definem essa ampliação de conhecimento como inteligência, daí surgem definições como inteligência de mercado e inteligência competitiva. Essa maior capacidade de captar e organizar informações relevantes sobre o comportamento de determinado objeto de análise, permite um melhor processo de tomada de decisão para os mais diversos setores da economia ao prever e determinar tendências e cenários no curto e longo prazo.

A aplicabilidade para o setor privado é mais intuitiva e tem seu uso mais prolongado. As empresas como forma de adquirirem vantagens competitivas frente ao mercado cada vez mais consolidado e diversificado procuram se adaptar, mapear e antever o comportamento da concorrência. Assim, os dados garantem identificar novas oportunidades, ameaças e necessidades para novas abordagens. Os setores de empresas que fazem o uso de tal tecnologia não são restritos. Servem de apoio a

empresas de bens primários, bens de consumo duráveis e não duráveis, indústria farmacêutica, entre outros. Também não se restringem internamente na estrutura empresarial nas áreas de tecnologia da informação, uma vez que serve como elo de integração entre as mais diversificadas áreas: Marketing, Vendas, Serviços, Logística.

Mais recentemente a discussão em torno da análise preditiva e ciência de dados tem extrapolado o setor privado e mostrado grande importância no setor público seja no apoio a políticas governamentais, campanhas eleitorais e sistema de saúde pública.

Por se tratar de um fenômeno novo cujas origens remetem a não mais que uma década, ainda existe certo nível de desinformação quanto a aplicabilidade, conceitos, importância e riscos do uso de dados na economia. O presente trabalho tem por objetivo orientar e explicitar a tecnologia vinculada ao uso de dados e como ela tem servido de suporte e mudança ao processo de tomada de decisão. A hipótese que o presente trabalho busca desenvolver e esclarecer é de que a estruturação de dados a partir do Big Data significa uma mudança sem precedentes no processo de tomada de decisão dos agentes econômicos. Isto porque o Big Data associado ao uso aumentado das redes sociais possibilitou estudo e monitoramento do comportamento das pessoas, tanto enquanto consumidores, quanto enquanto cidadãos. Essa mudança de paradigma afeta tanto o setor público quanto o setor privado, influenciando direta ou indiretamente na economia como um todo. Além disso, o seu advento foi de fundamental importância durante a pandemia do coronavírus, seja como mecanismo de combate a sua propagação, seja na mudança de consumo para o ambiente online tendo em vista que a quarentena impossibilitou a compra em ambientes físicos.

Primeiro, o trabalho buscará definir o conceito e relevância da ciência de dados, mais especificamente relativa ao Big Data bem como a evolução do processo de tomada de decisão. Seguido do desenvolvimento de uma análise sobre a aplicabilidade de dados nos mais diversos setores da economia, seja na esfera pública quanto privada. Para isso, o desenvolvimento será dividido em duas partes, a primeira focará em exemplos vinculados ao mercado: na mudança estrutural do modelo de negócio e como ativo estratégico. A segunda parte busca exemplificar a aplicação no setor público mais especificamente no suporte ao sistema de saúde público. Por fim, os riscos do uso dessa tecnologia devem ficar claros uma vez que por se tratar de

uma inovação, pouco foi discutido sobre as consequências no longo prazo que o compartilhamento e armazenamento de dados desregulado pode trazer. Para isso, será feita uma análise acerca dos mecanismos de controle tanto na esfera técnica quanto legal para a garantia de privacidade de dados sensíveis.

Capítulo 1

1.1 Definição do Big Data

O termo Big Data não é novo, ele remete ao ano de 1997, quando foi utilizado no contexto de visualizar grandes bases de dados (Cox e Elisworth 1997). No geral, não há uma definição unificada para Big Data, a mais convencional entende como o conjunto de dados de larga escala que podem ser analisados computacionalmente para revelar padrões, tendências e associações especialmente relacionadas ao comportamento e interações humanas. Ele foi originado pelo rápido crescimento do volume e diversidade de dados digitais gerados em tempo real de acordo com a relevância cada vez maior do papel das tecnologias nas atividades diárias humanas.

A partir desse entendimento, é importante diferenciar o Big Data das bases de dados tradicionais. Dados podem ser entendidos como fatos não processados e não organizados sobre eventos, que servem como pré-requisito à informação. A informação, por sua vez, são as sugestões de entendimento e propósito dada por uma agregação de dados. A informação pode levar a conhecimento (entendimento humano do objeto analisado) a partir de teoria e evidência. Apesar de ambas, tanto base de dados tradicionais quanto o Big Data levarem a um conhecimento específico do objeto, se convencionou diferenciá-los pela complexidade e tamanho dos dados a serem analisados. Logo, o atributo “Big” tem relação direta com as inovações no campo da informática uma vez que faz referência a capacidade computacional, de processamento e armazenamento de dados.

Existem 5Vs que de acordo com avanços e pesquisas, atualmente é comumente utilizado para definir o Big Data: velocidade, volume, variedade, veracidade e valor. Alguns autores como Gandomi e Haider (2015) procuram enfatizar que características além do tamanho devem ser consideradas (Figura 1). Para isso, usaram a definição proposta por Laney (2001) na qual integra volume, variedade e velocidade a noção de veracidade e valor. Para além desses usualmente utilizados, D’Orazio (2017) acredita ainda que mais três outras características devem ser adicionadas: exaustividade,

relacionalidade e escalabilidade:

> “Os três Vs [volume, variedade e velocidade] se referem, respectivamente, à magnitude dos dados; a heterogeneidade estrutural do conjunto de dados e a taxa na qual dados são gerados. Os Vs adicionais [veracidade e valor] se referem à confiabilidade dos dados, devido à sua incerteza e imprecisão (...) e seu valor (...) Para essas características, acreditamos outras três devem ser adicionadas: exaustividade (ou seja, até que ponto o Big Data se esforça para capturar toda a população ou sistemas), relacionalidade (porque o Big Data contém campos comuns que permitem a conexão de diferentes conjuntos de dados) e escalabilidade (porque o BD pode expandir rapidamente em tamanho)” (Traduzido de: D’ORAZIO, P., 2017)

Figura 1 – Os 5 Vs do Big Data

Fonte: International Journal of Social Science and Economic Research (2019) – Formulação própria.

Manipulado da forma correta, o Big Data pode ser usado na geração de informação e conhecimento em tempo real, o que significa que existe tempo hábil para que uma nova informação seja usada para mudar decisões antes de serem irreversíveis. De forma resumida, as tecnologias relacionadas ao Big Data descrevem uma nova geração de inovações desenvolvidas para extrair valor econômico de grande volume de uma vasta variedade de dados de forma a possibilitar uma alta velocidade de captura, descoberta e análise.

A razão pela qual se possibilita o uso de tal tecnologia está em três pontos centrais: (1) o aumento na largura da banda de telecomunicações que interligam as redes (4G e 5G, por exemplo) possibilitando avanços significativos na velocidade de

conexão e carregamento de dados; (2) sistemas de armazenamento de dados capazes de guardar exponencialmente o fluxo de informações; e (3) capacidade de processamento de dados adquiridas pelos sistemas computacionais mais recentes (Hilbert, M., 2016).

Existem tanto dados que resultam da atividade online quanto aqueles que são gerados de forma offline. A grande maioria daqueles provenientes da atividade online surgem como resultado da interação em redes sociais, mas também podem surgir de compras em e-commerce, pesquisa em ferramentas de busca e aplicativos. Já como exemplo daqueles que surgem de maneira offline podemos citar compras em lojas físicas, transações bancárias e ferramentas de geolocalização (GPS).

É importante notar que determinar a fonte de dado como offline não a exclui de ser uma informação digital. Essa diferenciação é especialmente importante já que qualquer informação digital é passível de ser analisada no momento que é registrada, bem como em período posterior, apesar de não precisar necessariamente do suporte das redes de conexão. Fontes de dados tradicionais coletam dados para um propósito específico, no Big Data, por outro lado, os dados são reutilizados para propósitos diferentes daquele pretendidos quando foram gerados. Assim, o conceito de reutilização é fundamental para essa tecnologia.

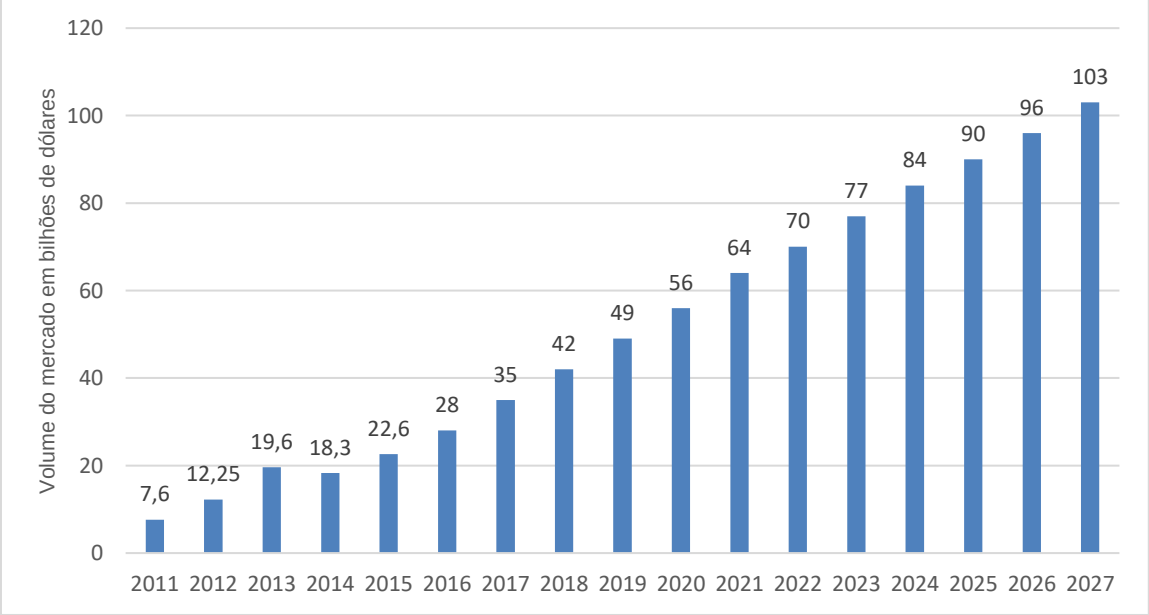
As fontes de dados que alimentam o Big Data particularmente incluem: (1) internet e mídias sociais; (2) registro de transações bancárias e financeiras; (3) informações geradas por pessoas, que incluem telefonemas, e-mail e registros médicos; (4) biometria que incluem reconhecimento facial, escaneamento de retina e impressões digitais; e (4) máquina a máquina, quando algum dispositivo captura e transmite algum evento como, por exemplo, medidores de velocidade, pressão e meteorologia.

1.2 Relevância do Big Data e modelos preditivos

Para além de suas definições e características, a real relevância do Big Data é como ele pode impactar e contribuir para objetivos econômicos, sociais e ambientais. Nos negócios, podem impactar a eficiência e produtividade; na economia, o desenvolvimento econômico; e na sustentabilidade, o monitoramento e reação a determinado dano ambiental.

Assim, o Big Data é visto como uma tecnologia promissora, com capacidade de provocar enormes impactos nas indústrias, empresas e sociabilidade como um todo. Por isso, é esperado um aumento das receitas mundiais relativas a mercado de softwares para Big Data de US$ 42 bilhões em 2018 para US$ 103 bilhões em 2027, segundo reportagem desenvolvida pela Forbes (2018). Esse crescimento na relevância pode ser percebido pelo Gráfico 1.

Gráfico 1 – Evolução do faturamento de mercados de Big Data de 2011 a 2027 (bilhões de dólares).



Fonte: Forbes (2018) – Formulação própria.

O Big Data tem revolucionado muitas premissas do funcionamento do mercado e esse impacto já pode ser percebido no crescimento do faturamento de grandes empresas. Mercados podem ser segmentados de formas mais precisas e produtos personalizados podem ser oferecidos pelas empresas. Essa melhor segmentação possibilita produtividade e lucratividade, com efeitos positivos no crescimento do negócio.

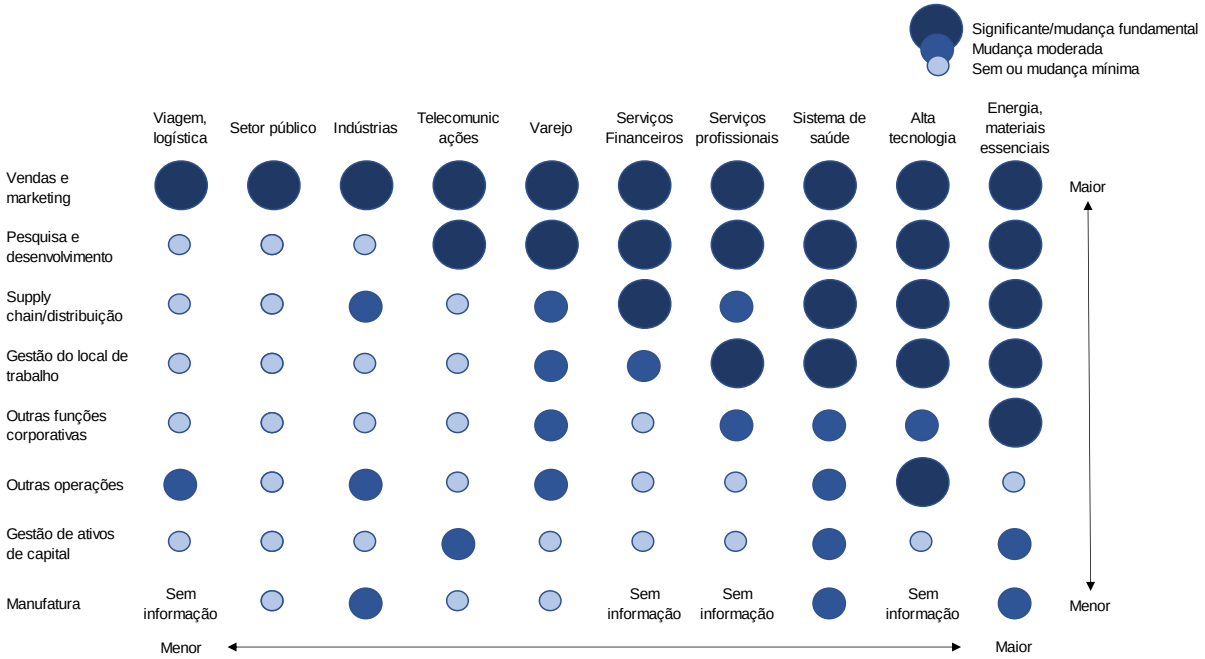
Nas últimas duas décadas de digitalização, a capacidade mundial de trocar informações a partir das telecomunicações cresceu do equivalente de dois jornais impressos por pessoa por dia em 1986 para seis jornais em 2007 (Hilbert e López, 2011). Ao mesmo tempo, a proporção de buscas no Google, compartilhamento de usuários no Facebook e conteúdo publicado no Twitter cresceu de maneira exponencial na última década.

Comparativamente, em 1986, usando toda a capacidade tecnológica de armazenamento disponível na época, seria possível armazenar apenas 1% de toda a informação disponível em escala mundial. Em pouco mais de 20 anos, em 2007, essa participação já tinha aumentado para 16% (Hilbert, M., 2016). Considerando ainda todos os novos adventos tecnológicos das últimas décadas, não é equivocado sugerir que esse percentual tenha aumentado em proporção ainda maior.

Esse crescimento no compartilhamento de informações afetou tanto países desenvolvidos quanto países subdesenvolvidos e em desenvolvimento. Como exemplo, os cinco principais países em termos de usuários do Facebook incluem Índia, Brasil, Indonésia e México. O Brasil, por exemplo, possui mais usuários per capita no Twitter do que a França ou Alemanha (Hilbert, M., 2016).

De acordo com a natureza disruptiva do Big Data, as empresas buscam se adaptar e revisar seus modelos de negócios. Isto porque a habilidade de lidar com a incerteza causada pelo rápido ritmo de mudança no ambiente econômico, institucional e tecnológico se tornou um objetivo fundamental das mudanças organizacionais nessa era da informação (HILBERT, M., 2016).

Gráfico 2 – Mudança nas práticas da indústria trazidas a partir de dados e análises por função, últimos 3 anos



Fonte: Forbes (2018) – Formulação própria.

Para além do impacto nos negócios (Gráfico 2), o Big Data e a análise preditiva surgem como suporte ao próprio funcionamento do governo e execução de políticas públicas mais assertivas e customizadas. Atualmente a informação é um ativo estratégico e o cabe ao Estado proteger, alavancar e analisar tais informações para melhor atender os requisitos da sociedade.

1.3 A geração de valor a partir do Big Data

A noção de Big Data vai além do volume e qualidade do crescente número de dados disponíveis e tem sua importância extrapolada para o mecanismo de tomada de decisão. Ele possibilita previsões mais inteligentes que acabam por influenciar o comportamento dos diversos atores sociais.

Do ponto de vista teórico, toda a decisão é uma aposta incerta e probabilística com base em algum tipo de informação prévia (Tversky e Khahneman, 1981). Se existir um aprimoramento da estrutura das informações prévias nas quais baseamos nossas estimativas, consequentemente, o nível de incerteza será reduzido e a tomada de decisão será mais assertiva. Nesse sentido, a análise preditiva proporcionada pela tecnologia de Big Data garante uma vasta variedade de informações que servem de instrumento para qualquer tipo de tomada de decisão.

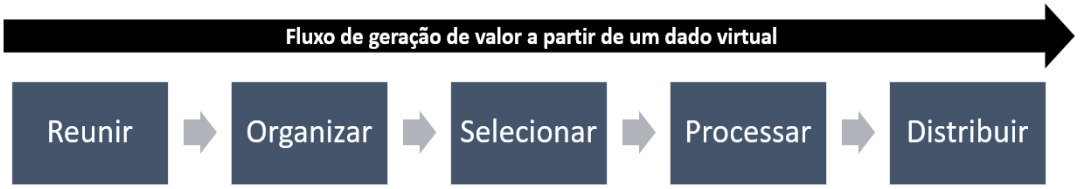
Existem autores que sugerem que o Big Data pode proporcionar um novo passo na escala da evolução da informação. Ele possibilitaria uma nova etapa do que conhecemos como “Era da informação”. Sua capacidade de transformar crescentes volumes de dados de maneira a traduzir uma informação digitalizada em efetivamente conhecimento, que proporciona tomadas de decisões inteligentes e assertivas, significa que o BG proporciona uma nova etapa natural na evolução da “Era da Informação” para a “Era do conhecimento”.

Conhecimento assim como informação não deixa de ser um valor, por isso, é importante ter clara a estrutura de geração de valor que, primeiro, um dado virtual pode proporcionar e, depois, estender a discussão para o Big Data.

Um dos primeiros modelos de estrutura de criação de um valor virtual (VVC) foi apresentado por Rayport e Sviokla (1995). Essa estrutura descreve cinco etapas necessárias para criar valor a partir dos dados: reunir, organizar, selecionar, processar e distribuir. As duas primeiras estão relacionadas ao dado em si, sua

disponibilidade e recursos utilizados para acessá-los. As outras três dependem de um processo analítico, a capacidade de processamento, os sistemas correlacionados e ao manuseamento da informação que o dado em si pode trazer. Ou seja, não há valor em um dado sem que ele perpasse pelo processo de análise.

Figura 2 – Os 5 Vs do Big Data



Fonte: SVIOKA, J. e RAYPORT, J. (1995) - Formulação própria.

Dados virtuais existem desde o surgimento do primeiro computador enquanto a capacidade de processá-los remonta ao desenvolvimento da ciência da computação, mais especificamente, ao surgimento de algoritmos[1]. O processamento de dados, por sua vez, corresponde à fusão do Hardware com o Software. O primeiro, se relaciona a parte física do computador, ou seja, seus componentes. O segundo, é a parte lógica dele, ou seja, a manipulação e execução da máquina, aqui se encaixam os algoritmos. Ou seja, a existência de dados virtuais não indica a capacidade de processá-los. É necessária uma junção de um Hardware com capacidade suficiente de armazenamento e um Software capaz de organizar e processar a complexidade dos mais diversos tipos de dados para se criar o valor de um dado disponível.

É justamente neste ponto que o Big Data se diferencia das bases de dados virtuais tradicionais. A capacidade estrutural do computador atual associada a evolução dos seus sistemas operacionais capacitou um maior nível de processamento de dados que gera mais valor e, logo, mais conhecimento.

1.4 A tomada de decisão com o advento do Big Data

A ciência de dados pode ser entendida como o conjunto de princípios, processos e técnicas que envolvem o entendimento de um fenômeno a partir da análise de dados. Seu objetivo último é servir como suporte para o processo de

[1] Algoritmos são uma sequência de ações executáveis que tem por objetivo obter a solução para determinado tipo de problema.

tomada de decisão.

O também conhecido como DDD (Data-driven decision making) se refere a prática de basear as decisões a partir da análise de dados ao invés de tomar a simples intuição como referência. Como exemplo, a área de marketing de uma empresa pode selecionar uma propaganda baseada puramente pela experiência de mercado, mas também pode selecionar seu público-alvo a partir de dados que refletem a reação de seus consumidores foco. Esse exemplo é especialmente relevante tendo em vista o aumento do tempo que consumidores têm despendido online e a sua capacidade de fazer decisões rapidamente nesse ambiente virtual. Nesse sentido, há uma espécie de revolução na publicidade, utilizando de estratégias de marketing digital cada vez mais frequentes e recorrentes para que se possa aproveitar do tempo online do consumidor e promover uma rápida tomada de decisão (já que uma compra online normalmente leva muito menos tempo do que uma física, e, logo, o processo de tomada de decisão acompanha essa tendência). A disponibilidade de seus dados, que expressam seus interesses de consumo, facilita esse processo ao indicar o produto mais conveniente, a loja com um preço de venda mais adequado e a localização que melhor e mais rapidamente atenda o desejo do consumidor. Esse processo, em escala mais ampla, acaba promovendo uma melhor experiência de compra, com sugestões mais assertivas e com menor risco ao erro.

Essas estratégias de marketing digital que utilizam o Big Data como ferramenta, transformaram a maneira como as empresas avaliam o comportamento de seus clientes no mundo todo. Além de prever tendências passadas, ele permite avaliar o comportamento futuro ampliando a gama de informações disponíveis. A análise preditiva fornece insights significativos sobre a base de consumidores finais, o que ajuda a acompanhar o fluxo da demanda. Esses insights englobam principalmente três setores: a segmentação personalizada, aprimoramento do público-alvo e aumento do volume de consumidores. O primeiro setor, permite criar campanhas direcionadas e personalizadas para um público-alvo específico. Esses públicos-alvo são agrupados de uma forma mais característica e assertiva, uma vez o Big Data possibilita a utilização de um nível muito maior de informações coletadas que permitem uma análise de mercado e definição de público-alvo mais completa e aprimorada. O terceiro e último setor tem por objetivo identificar oportunidades e

atingir clientes potenciais de diferentes maneiras como, por exemplo, utilizando histórico de compras, abordagem de cross-selling (venda cruzada) e abordagem correta de preços.

Tanto no âmbito do consumidor quanto pela perspectiva empresarial, já ficam claros os benefícios dessa tomada de decisão derivada de dados. Eles já foram, inclusive, comprovados por pesquisa conduzida pelo MIT e a Penn's Wharton School. Nelas, foram propostas uma unidade de medida do DDD que classifica as empresas de acordo com seu nível de uso de dados para tomada de decisão. Eles mostraram estatisticamente que quanto mais orientada por dados uma empresa é, mais produtiva também é considerada. (Provost e Fawcett, 2013).

Para além da perspectiva das empresas e consumidores, o DDD pode ser aplicado também no âmbito das decisões macroeconômicas. D'Orazio (2017) defende que em considerando a metodologia macroeconômica a partir de paradigmas[2] que foram se desdobrando e desenvolvendo ao longo do tempo, a difusão de uma quantidade maior de dados complexos com desenvolvimento de tecnologias computacionais com alto poder de armazenamento, possibilitou o surgimento de um novo paradigma. Ele se desdobrou sobretudo a partir da recente crise financeira de 2008 que expôs algumas "anomalias" das teorias macroeconômicas mainstream – hipótese das expectativas racionais, de eficiência de mercado etc. Essa otimização dos agentes propostas por essas teorias não ajudaram a entender a crise de 2008 e, logo, pareciam inconsistentes para entender crises futuras. D'Orazio (2017) argumenta que a teoria macroeconômica deve levar em consideração as interações dos agentes e suas heterogeneidades, para além dos modelos dos agentes representativos neoclássicos. Ela enfatiza que essa mudança pode ser observada com o avanço da computação e utilização do Big Data nas últimas décadas, permitindo uma maior rastreabilidade analítica em contraste ao uso de dados agregados.

Atualmente existe de fato um amplo cenário digital à disposição dos pesquisadores sociais, pois é possível coletar dados sobre opiniões dos usuários, questionários on-line sobre preferências dos consumidores por meio da compra on-

[2] Paradigma é entendido aqui por sua abordagem kuhniana, no qual um paradigma é "[...] o que os membros de uma comunidade científica compartilham e, inversamente, uma comunidade científica consiste em homens que compartilham um paradigma" (Kuhn, 1962, p. 176).

line, redes sociais etc. que permitem utilização de métodos de pesquisas mais ricos e realistas. A maior disponibilidade de dados é de extrema importância porque permite que os pesquisadores levem em conta o contexto e a contingência específicos, além de poderem ser utilizados para refinar a compreensão de alguns agentes específicos ou estruturas de rede. Nesse sentido, a interação constante entre o modelo teórico e os dados disponíveis desempenha um papel crucial.

A tomada de decisão na perspectiva macroeconômica se relaciona ao Big Data em dois pilares que se correlacionam: o primeiro por dar suporte aos modelos metodológicos macroeconômicos de formar a definir análises preditivas econômicas com maior rastreabilidade e o segundo por sua perspectiva política, que reflete na economia real, a partir das decisões governamentais que são diretamente influenciadas pelos modelos metodológicos cada vez mais previsíveis, heterogêneos e granulares.

Geralmente, quando tomamos algum tipo de decisão, este processo é tendencioso e limitado pela incapacidade de processar a sobrecarga de informações. Dados e análises de Big Data podem ajudar decompondo as assimetrias de informações, novas fontes de dados e rapidez com algoritmos automatizados. Quanto mais as fontes de dados se tornem ricas e diversificadas, mais rápida, consistente, precisa e transparente serão as tomadas de decisões.

Assim, entre os aspectos que a análise de dados fornecida pelo Big Data proporciona na melhoria no processo de tomada de decisão estão: a rapidez/adaptabilidade, precisão, consistência e transparência. Dados aplicados a algoritmos podem fornecer respostas em tempo real garantindo rapidez e adaptabilidade. Este é o caso dos carros autônomos, que garantem a segurança da direção a partir da resposta rápida à obstáculos, perigos e curvas nas vias por exemplo. A precisão possibilita um uso mais acertado dos recursos. Podemos citar a aplicabilidade no sistema de saúde que garante a melhor escolha de medicação e dosagem para pacientes em tratamento como exemplo. Além disso, a mesma informação não gera conclusões variadas o que garante a minimização dos erros. Modelos digitais de projetos ajudam arquitetos e engenheiros a tomar decisões assegurando integridade estrutural e diminuir o retrabalho. Por fim, a transparência permite que decisões possam ser revisadas e melhoradas no futuro. Mais transparência no sistema jurídico, por exemplo, poderiam tornar as decisões mais

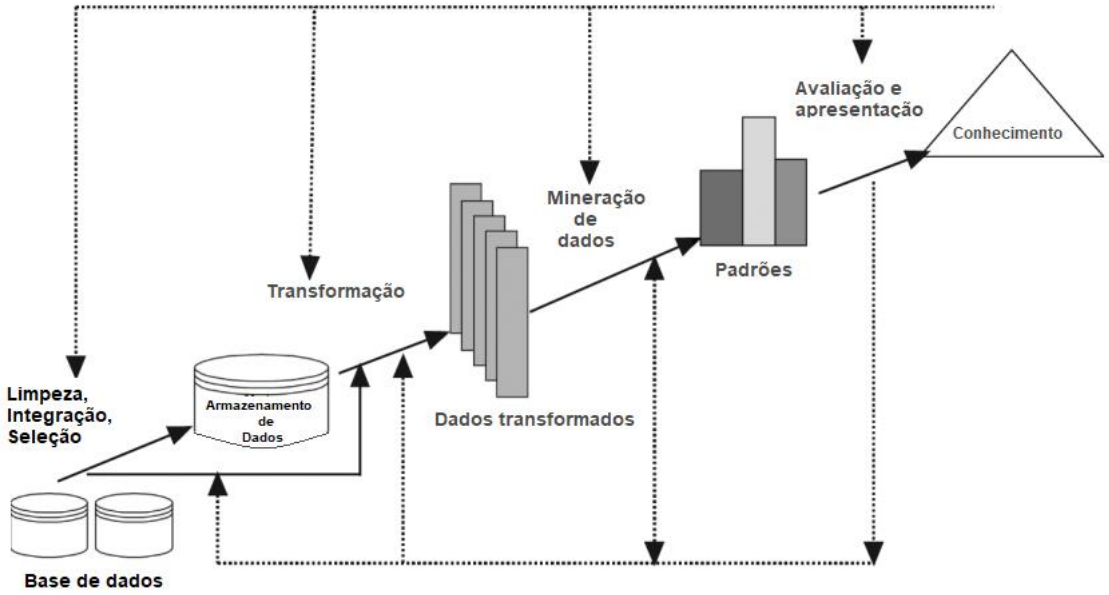
imparciais quando alimentadas pelo maior número de dados possível.

1.5 O processo de tomada de decisão baseada em dados.

O processo que correlaciona o Big Data, tomada de decisão e os usuários é comumente entendido como mineração de dados. Esse termo também é tratado por seu sinônimo, o *Knowledge Discovery from Data* (KDD). Esta sigla se traduz como “descoberta de conhecimento a partir de dados” e foi utilizada pela primeira vez em 1996 em artigo publicado pelo MIT e desenvolvido pelos autores Fayyad, Piatsky-Shapiro e Smyth. O processo de descoberta de conhecimento útil a partir de dados ou KDD pode ser descrito pelas as etapas mostradas na Figura 3.

As etapas evoluem de forma que o destino final seja o conhecimento/informação. Nesse sentido, a Etapa 1 engloba o pré-processamento de dados cujas operação inclui: seleção, limpeza e integração de dados. A Etapa 2 é o estágio de transformação de dados e tem como objetivo encontrar recursos úteis (aplicativos, programas etc.) para representar os dados. Já na terceira etapa, métodos são aplicados para identificação de padrões de dados (por exemplo, clusters, classificações, etc). A quarta e última etapa, é a etapa de avaliação e apresentação onde são identificadas informações úteis aos conhecimentos e efetivamente compartilhadas de forma mais simples.

Figura 3 – Visão geral do processo KDD

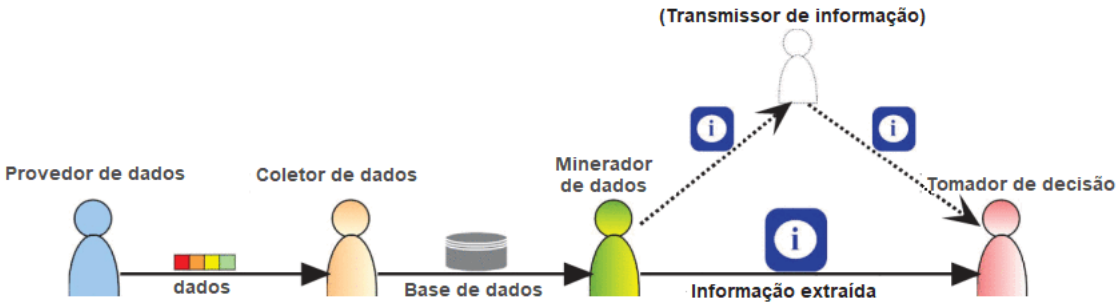


Fonte:

Com base na divisão dos estágios de processamento do KDD, podemos

identificar quatro tipos diferentes de usuários que garantem como objetivo último do Big Data, o suporte à tomada de decisão. Podemos perceber essas etapas a partir do demonstrado na Figura 4.

Figura 4 – Dos dados à tomada de decisão



Fonte: L. Xu, C. Jiang, J. Wang, J. Yuan and Y. Ren (2014)

As funções desses usuários se dividem em 4. O Provedor de dados é aquele usuário que possui a capacidade de compartilhar informações desejadas através de dados. O coletor de dados, recolhe os dados providos e os dispõe para a mineração de dados. Já o minerador de dados tem como função a descoberta de conhecimento a partir dos dados dispostos (KDD). O tomador de decisão representa aquele que toma decisões com base nos resultados da mineração de dados para atingir determinados objetivos. Neste cenário, um usuário é a representação de uma pessoa ou organização. Além disso, um único usuário pode desempenhar várias funções ao mesmo tempo.

Essas funções são importantes não só para compreender a arquitetura por trás da mineração de dados, obtenção de conhecimento e, logo, tomada de decisão, como também para compreender a questão da segurança da informação (será discutida posteriormente no Capítulo 3). O desempenho de cada um dos usuários pode comprometer ou não o resultado esperado e, assim, a tomada de decisão está diretamente relacionada com o nível de confiança de todo o processo de mineração.

Capítulo 2

Nesta seção, começamos descrevendo os recursos de dados disponíveis para o governo bem como os dados podem ser usados para acompanhar e prever melhor a atividade econômica, como pode ser usado em decisões políticas ou como suporte para melhoria dos serviços governamentais.

2.1 O uso do Big Data para o setor público

2.1.1 Big Data no setor público.

Conforme a economia inteligente vai se moldando nos nossos meios profissionais e sociais, a empresa inteligente emerge, possibilitando crescimento, solução de problemas e aceleração da inovação.

Os avanços da tecnologia e a crescente quantidade de informações disponíveis também transformou a condução de políticas governamentais. Essa informação disponível é um ativo estratégico e que o governo precisa proteger, alavancar e analisar para melhor atender seus objetivos de gestão. Dessa forma, podem se adequar ao que se chama de “economia inteligente”.

O termo economia inteligente é o nome dado ao próximo estágio da economia tal como conhecemos hoje. Esse novo paradigma será guiado prioritariamente pelas inteligências dos computadores nas tomadas de decisão, nos seus bancos de dados e nas suas análises preditivas. Dessa forma, soluções governamentais serão criadas a partir das mais diversificadas tecnologias disruptivas como: Big Data, serviços em nuvem e redes sociais.

Pesquisas mostram que tecnologias vinculadas ao Big Data estão no cerne da economia inteligente. A partir de novas ferramentas de captura, pesquisa, descoberta e análise, as organizações do setor público conseguem obter insights de seus dados não estruturados (que representam mais de 90% do universo digital).

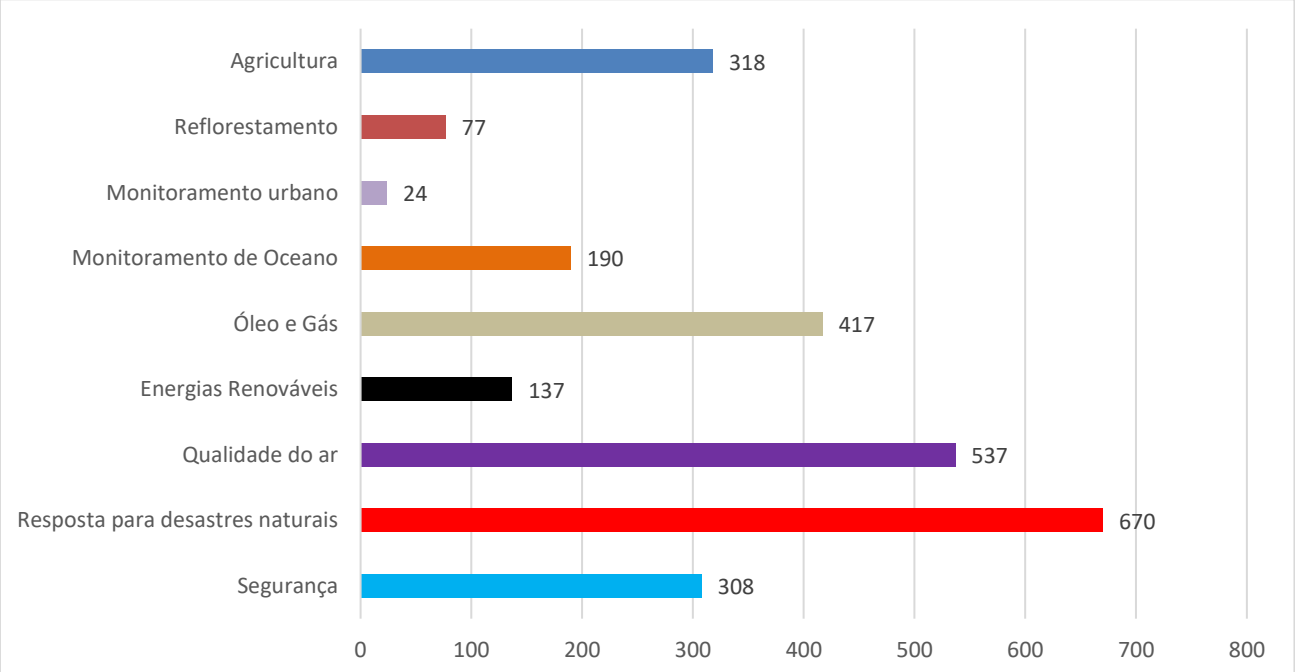
Como exemplo, o governo de alguns países já usa o Big Data para analisar conjuntos de dados massivos em pesquisas científicas, controle ambiental (green data), bem como na tentativa de controlar a violência e alguns tipos de fraudes fiscais.

A União Europeia por exemplo conta com diferentes modelos geradores de Big Data ambiental (green data), entre eles o Copernicus. Ele é um programa de observação da Terra por satélite capaz de calcular a influência do aumento das temperaturas os rios. Esse mecanismo se torna fundamental na medida que proporciona informações relevantes para a otimização de recursos hídricos, biodiversidade, qualidade do ar e agricultura.

Os dados e informações processados e divulgados colocam o Copernicus na vanguarda da mudança de paradigma tecnológico. Nesse sentido, ele se relaciona aos 5Vs que definem o Big Data: volume, velocidade, variedade, veracidade e valor. Quanto ao volume, mais de cem mil usuários fizeram o download de dados desde o início da operacionalização do sistema; a velocidade se relaciona ao tempo de processamento e entrega dos dados, mais de 100 TB de dados são divulgados todos os dias; a variedade vinculada ao projeto estão nos diferentes tipos de sensores que os satélites possuem e que podem gerar diferentes conjuntos de dados; essa diversidade de dados obtidos necessitam de uma fonte confiável, por isso, a veracidade é fator essencial na medida em que avalia a qualidade dos dados; por último o valor, já que a extração de informações do Copernicus trazem benefícios econômicos diretos para a Europa já que possui o potencial de impactar na criação de empregos, inovação e crescimento como um todo.

É estimado que os benefícios gerais para usuários finais que se utilizam dos dados obtidos pelo Copernicus podem chegar a 206 milhões de euros em 2020. Eles mostram grande disposição de acessar as informações personalizadas, que podem gerar valor ao seu negócio como mostra o Gráfico 3.

Gráfico 3 – Visão geral de benefícios para usuário final dos dados gerados pelo Copernicus (em Milhões de Euros)



Fonte: PwC Market Report (2019) – Edição 2 – Página 7.

Um fator importante e que merece devida atenção são os custos vinculados a esse tipo de tecnologia. O investimento estimado no Copernicus, por exemplo, está em 7,4 bilhões de euros. Esse patamar de custo pode significar uma barreira na introdução desse tipo de tecnologia em países subdesenvolvidos. Esse fator fica ainda mais agravante quando percebemos que o acesso à tecnologia parece ser chave para o crescimento econômico na atualidade.

Para além de fatores socioeconômicos, barreiras técnicas também se aplicam uma vez que a capacidade de processamento de dados cada vez mais volumosos depende diretamente da capacidade de armazená-los. É necessário o investimento em infraestrutura de servidores para a sua implementação. Assim, o armazenamento está se tornando um ativo estratégico na medida que o volume de informações disponíveis gera preocupações quanto a recuperação e arquivamento de dados governamentais.

Além disso, questões vinculadas a segurança das informações também se tornam cada vez mais complexas. Devido à natureza sensível dos dados governamentais, se faz necessário evitar ataques cibernéticos, proteger informações críticas e garantir proteção da identidade dos cidadãos.

2.1.2 Big Data no Governo brasileiro

Em uma sociedade que muda tão rapidamente, a análise em tempo real é vital. Permite que os governos tomem decisões mais rápidas e permite que monitorem essas decisões e implementem mudanças rapidamente, se necessário.

O governo federal brasileira coleta grande volume de dados dos mais diversos setores: saúde, educação, economia, cultura e segurança pública. Ao longo dos últimos governos podemos perceber um avanço na estruturação e acesso a dados como por exemplo pelo “Portal Brasileiro de Dados Abertos” e o “GovData”. Segundo o próprio site do “GovData”, ele disponibiliza dados necessários para monitorar informações e orientar a tomada de decisão baseada em evidências.

Como suporte nas políticas sociais, permitiu por exemplo zerar a fila do Bolsa Família a partir do cruzamento de dados. Isto porque possibilitou identificar de forma mais precisa os cidadãos que cumprem os requisitos do programa. Foram

usadas as informações do SISOBI e SIM (dados relativos a óbitos), RAIS e CAGED (dados relativos a empregabilidade), SIAPE (dados relativos a recursos humanos), NIS, INSS e Grupo Familiar do BCP.

O uso do Big Data e Inteligência Artificial na segurança pública permitiu a partir de parceria com a Universidade Federal do Ceará (UFC) acompanhar as ocorrências de todo o país, buscar informações e ficha criminal de suspeitos, monitorar veículos roubados, atuar no combate ao tráfico nas regiões de fronteiras, além de agir de prontidão na prevenção de assaltos e homicídios. O projeto prevê a implantação de uma plataforma de Big Data e inteligência artificial (IA) para o Sistema Nacional de Informações de Segurança Pública, do Ministério da Justiça. O sistema integra os dados das secretarias de segurança pública de todo o País.

Órgãos como a Receita Federal e as Secretarias Estaduais utilizam essas ferramentas para cruzar dados e detectar sonegações de impostos e evasões ficais. Essa pode ser considerado uma forma de gerir com maior eficácia a arrecadação de verbas públicas. Como exemplo prático do Big Data Fiscal, ele permite regularizar os parcelamentos de débitos de acordo com o Programa de Regularização Tributária do Governo Federal. Isto porque os registros desses débitos em parcelamento ficam cadastrados em grandes Big Datas do Governo. Isso permite aos fiscos identificarem rapidamente quais contribuintes possuem impedimentos ficais com maior materialidade.

Fora da área de segurança pública, projetos relativos a Big Data e Inteligência Artificial também vem sendo desenvolvidos pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). O tema foco de pesquisa da FAPESP é saúde pública e entre os tópicos a serem desenvolvidos com o suporte do Big Data estão: medicina de precisão; visão computacional e realidade aumentada em saúde humana; sistemas de apoio à decisão em medicina; telemedicina e eHealth; segurança, privacidade, confiabilidade e visualização de dados para a saúde humana; gestão, administração e administração de medicamentos; e sistemas de monitoramento de pacientes

2.2 O uso do Big Data, saúde pública e coronavírus

A pandemia do coronavírus também conhecida como pandemia de COVID-19, é uma pandemia em curso declarada pela Organização Mundial da Saúde em 11

de Março de 2020, causada pelo vírus SARS-CoV-2, que provoca síndrome respiratória aguda grave e foi identificada pela primeira vez em Wuhan, província de Hubei, na China.

Conforme o avanço do número de infectados pela doença, muitos sistemas hospitalares ao redor do mundo entraram em colapso. De acordo com a mobilização cada vez maior por parte da sociedade em combater os avanços e os prejuízos causados pelo vírus, o Big Data permitiu uma resposta tecnológica aos conflitos vivenciados pela comunidade médica.

Graças aos mais recentes avanços no campo das técnicas computacionais e das tecnologias de informação e comunicação (TICs), a inteligência artificial (IA) e o Big Data podem ajudar no manuseio da enorme e sem precedentes quantidade de dados derivados da vigilância da saúde pública em tempo real, monitoramento de surtos epidêmicos, previsão de tendências, atualizações e informações sobre a utilização de unidades de saúde.

Com o Big Data está se tornando possível customizar tratamentos médicos, isto porque os dados que descrevem a saúde dos pacientes estão se tornando cada vez mais granulares de acordo com o decrescente custo de sequenciamento genético, o advento da análise de proteínas, o aumento no número de sensores, monitores e possibilidade de diagnósticos que possibilitam um fluxo constante de dados. Existe um esforço nos dias de hoje para integrar esses novos recursos e seus vastos conjuntos de dados dos meios eletrônicos para efetivamente o tratamento médico. Mecanismos de análise de dados avançados poderiam transformar os tratamentos padrões em tratamentos personalizados de acordo com variantes como: avaliação de risco, diagnóstico, tratamento e monitoramento.

Para Filho (2015), existem três novas áreas para análise do Big Data em saúde: medicina de precisão, prontuários eletrônicos do paciente e internet das coisas.

A medicina de precisão garantiria com o suporte da análise de dados avançados um tratamento customizado para pacientes de forma a garantir a eficácia assertiva do tratamento. Ele utiliza como exemplo um estudo acerca de um medicamento anticoagulante que garantiria a redução de 19% de acidentes vasculares cerebrais (AVC). O problema colocado é que somente alguns pacientes no estudo tiveram o risco diminuído em 100% (não tiveram AVC) e outras 0%

(tiveram AVC). Ou seja, é sabido que o uso desses anticoagulantes reduz a presença desse evento (ter AVC) na população como um todo, mas não exatamente para quem. O objetivo é que a medicina de precisão garanta tratamento individualizado apenas para aqueles que o medicamento funcione com eficácia.

Para que isso seja possível, é fundamental que se aumente o tamanho das amostras das pesquisas, por isso, a importância dos prontuários eletrônicos. Esta ferramenta ainda é defasada no Brasil pois não existe um registro único para os pacientes e sim prontuários (muitas vezes em papel) específicos para cada unidade de saúde. Para existir essa universalização da digitalização dos prontuários no Brasil, seria fundamental o uso integrado do prontuário eletrônico do paciente (PEP), que permitiria o uso remoto individual por todas as unidades de saúde, tanto públicas quanto privadas. O SUS poderia acompanhar a tendência de outros países e tomar uma posição de liderança para garantir o uso integrado das PEPs.

A internet das coisas prevê que todos os objetos de uso diário se conectem de alguma forma à Internet. Na saúde, por exemplo, isto poderia ser traduzido pelo uso de sensores conectados à Internet que detectam uma situação iminente de perigo e que podem garantir socorro prontamente. Ou seja, a própria internet das coisas facilitaria a coleta de dados que poderiam ser úteis para diagnóstico, chamadas de urgência e tratamentos, por exemplo.

2.3 O uso do Big Data para o setor privado

No setor privado, as empresas e organizações usam do conhecimento para otimizar os processos de tomada de decisão corporativa, prever tendências futuras e são considerados um ativo valioso na economia. Com o apoio do Big Data fica cada vez mais evidente que a análise de dados significa uma base de competição entre as empresas da qual os empresários preveem grandes vantagens. Acima de tudo, dados e análises podem permitir uma tomada de decisão mais rápida e baseada em evidências.

Essa adaptação, no entanto, enfrenta certa dificuldade organizacionais, principalmente no que diz a respeito de encontrar colaboradores capacitados e reorganizar os processos do negócio para esse novo modelo. Isto porque os dados e análises estão mudando todo princípio de competição, propondo uma mudança não só operacional, mas sim do próprio modelo de negócio.

Várias companhias já se aproveitam desses recursos para melhorar sua operação ou lançar modelos de negócio totalmente novos. Líderes em análises de dados como Apple, Alphabet/Google, Amazon, Facebook, Microsoft, GE, Baidu e Alibaba já se estabelecem como as empresas mais valiosas do mundo. Essas empresas têm se diferenciado sobretudo devido ao investimento em três fatores: investimento em fornecimento de dados, ou seja, empresas ou áreas especializadas em coletar, agregar e fornecer os dados; investimento em recursos humanos tecnicamente capacitados para tarefa analítica; e investimento em estrutura de dados, que podem englobar toda a estrutura física ou digital capaz de armazenar, analisar e fornecer esse grande volume dados. Em relatório da Mckinsey divulgado em 2016, são apresentados exemplos dessa nova realidade:

> “A UPS alimenta dados em sua plataforma ORION para determinar as rotas mais eficientes para seus motoristas de forma dinâmica. Só nos Estados Unidos, a empresa estima que o sistema irá reduzir o número de milhas que seus veículos percorrem a cada ano em 100 milhões, economizando mais de US$ 300 milhões anualmente. O Google está executando experimentos para induzir consultas de pesquisa mais rápidas, já que alguns milissegundos podem se traduzir em milhões de dólares adicionais em suas receitas. (...) A Netflix refinou seu mecanismo de recomendação e o lançou para clientes que estima gerar cerca de US$ 1 bilhão em receita anual.” (Traduzido de Mckinsey Global, 2016.)

Assim, os dados passam a representar uma nova classe de ativos coorporativos, providenciando informações para marketing, vendas, desenvolvimento de produtos, otimização das operações e produtividade. O valor dos dados e análises alterou toda a relação tradicional entre os consumidores e produtores. Hoje, toda e qualquer interação das empresas com o consumidor podem gerar informações valiosas. As vezes, os dados são tão valiosos que empresas estão dispostas a oferecer serviços gratuitos a fim de obtê-lo. Essa situação coloca a prova a segurança e privacidade do usuário pois muitas vezes não é de seu conhecimento que essa troca está sendo feita. Esses limites serão explorados no próximo capítulo.

É importante ressaltar que essa tendência tecnológica ainda é restrita a um pequeno grupo de empresas, que muitas vezes já são pertencentes a setores ligados à tecnologia. Este fator acaba criando disparidades entre esse pequeno

grupo de empresas líderes e as demais. Além disso, considerando que essas empresas de tecnologia estão sobretudo localizadas em países desenvolvidos, é fato que percebemos certo atraso em países que são dependentes da tecnologia vinda de fora.

Capítulo 3

3.1 Os limites e segurança dos dados

3.1.1 Preocupações do usuário.

Apesar da grande variedade de aplicação e o retorno informativo que o Big Data pode oferecer, existe uma preocupação crescente dos indivíduos com relação às ameaças à privacidade que podem surgir. A sua privacidade pode ser violada devido ao acesso não autorizado a dados pessoais, à descoberta indesejada de informações privadas e ao uso inadequado dos dados para fins diferentes dos quais os dados foram inicialmente coletados. Assim, podemos afirmar que existe um conflito entre segurança da privacidade e Big Data.

A Cloud Security Alliance (CSA) é uma organização sem fins lucrativos de escopo internacional que possui a missão de promover o uso das melhores práticas de forma a garantir segurança em todas as formas de computação em nuvem. Ela oferece conteúdos didáticos, pesquisa científica e eventos centrados especificamente em segurança. Entre seus conteúdos, apresentou em novembro de 2012 o chamado “Top Ten Big Data Security and Privacy Challenges” (ou “Os Top Dez Desafios de Segurança e Privacidade no Big Data”). Esses desafios contam com problemas de segurança e privacidade prioritários como: armazenamento seguro de dados; controle de acesso criptografados; auditorias; fonte de dados; monitoramento; e controle de acesso granular.

Para lidar com essas questões de privacidade, surgiu um subcampo conhecido como Privacy Preserving Data Mining (PPDM) que teve grande desenvolvimento ao longo dos últimos anos. O objetivo do PPDM é proteger as informações confidenciais da divulgação não autorizada enquanto preserva a utilidade dos dados. Os algoritmos atuais propostos para a PPDM se concentram principalmente em como ocultar essas informações confidenciais de certas operações de mineração.

Retomando o processo de mineração, já explorado inicialmente no Capítulo 1, no qual dividimos a função do usuário em quatro (Figura 4) podemos explorar os mais diversos problemas de privacidade no processo de mineração de dados. A diferenciação da função de cada usuário justifica preocupações diferentes acerca da privacidade dos dados, isto porque observam a questão da segurança a partir da sua própria perspectiva.

Os autores XU e col. (2014), diferenciam estas perspectivas das preocupações sobre segurança de dados de acordo com cada usuário. Segundo eles, a principal preocupação de um provedor de dados é a confidencialidade de dados privados. Os dados uma vez coletados pelo coletor de dados, devem passar por modificações, a preocupação nesse processo está em garantir que os dados modificados não contenham informações confidenciais, mas que ainda preservem alta utilidade. Assim, segundo essas funções, o coletor de dados deve assumir a responsabilidade principal ao proteger os dados confidenciais, enquanto o terceiro usuário (minerador de dados) deve se concentrar em como ocultar os resultados confidenciais. O tomador de decisões, por sua vez, concentra suas preocupações na confiabilidade do resultado da mineração. As abordagens para lidar com essas preocupações quanto a segurança da informação também são diferenciáveis de acordo com a função de cada usuário.

3.1.2 Abordagens quanto a questão da segurança de dados.

Não há garantia quanto a confidencialidade dos dados uma vez que são entregues à terceiros, por isso, é importante que o provedor se certifique que seus dados confidenciais estejam fora do alcance desde o início da cadeia. Para isso, podem ser usadas tecnologias anti-rastreamento nos navegadores de Internet chamadas DNT (*Do Not Track* ou Não Rastrear). Esta parece ser uma solução viável opara os problemas de privacidade, visto que permite o controle sobre quem pode ter acesso a sua atividade online. Porém, não há garantias de que o servidor honrará com a solicitação do usuário. A proteção perfeita seria não revelar dados confidenciais a terceiros, mas isso deve obstruir os benefícios trazidos pela mineração. Mais atualmente foram criados mecanismos que permitem atribuir maior valor à privacidade, de forma que os fornecedores de dados ganhem mais benefício com a divulgação de informações sensíveis. São conhecidos como leilões

de dados sensíveis, no qual se busca consentimento e transparência no fornecimento de dados confidenciais.

Um ponto importante a ser destacado no contexto do processo de mineração de dados é a falta de conhecimento do provedor de dados quanto a como seus dados estão sendo usados. Na falta de maneiras de monitorar o comportamento do coletor e minerador de dados, os provedores de dados acabam por ter conhecimento sobre o uso indiscriminado dos seus dados mais por divulgação da mídia do que efetivamente por um conhecimento técnico. Assim, podemos afirmar que existe certo nível de obscuridade quanto ao processo que seus dados são expostos. Em relatório divulgado em 2013 pela empresa de telecomunicações dos Estados Unidos, Verizon Communications Inc., cerca de 70% dos incidentes relativos a violação de dados são descobertos por alguém que não seja o proprietários dos dados.

Como acima exposto, o coletor de dados deve garantir a proteção dos dados fornecidos pelo provedor de dados. A abordagem para garantia dessa preocupação envolve estudos atuais sobre o que chamam de PPDP (Privacy Preserving Data Publishing ou Preservação de Privacidade dos Dados Publicados), que é o processo adotado pelo coletor de dados no qual o objetivo principal é a preservação da privacidade e utilidade de dados simultaneamente. A maioria desses estudos têm seu enfoque em processos estatísticos, isto porque generalizam o grau de preservação de dados de uma amostra de acordo com a preferência da maior parte dos usuários. Na realidade, porém, a opinião quanto a sensibilidade de cada tipo de dado muitas vezes variam de pessoa para pessoa. Como exemplo, dados quanto a orientação sexual ou retorno salarial podem ser considerados sensíveis para uma pessoa, enquanto para outras, não. Estudos em desenvolvimento mais recentemente já citam a questão da preservação de privacidade personalizada chamada de anonimato personalizado, no qual provedores de dados precisam declarar seus próprios requisitos de privacidade ao coletor. Apesar de surgir como uma tendência, ainda é raro que um provedor de dados defina sua preferência de privacidade de uma forma tão formal e muitas vezes vistas como dificultosas.

A preocupação do minerador de dados deve ser ocultar informações sensíveis que possivelmente venham a seu conhecimento. Por isso, sua abordagem deve englobar a modificação dos dados originais por meio de randomização, bloqueio,

transformação ou reconstrução. É preciso garantir que qualquer modificação não afete negativamente a utilidade dos dados de forma equilibrada com possíveis implicações sobre a sensibilidade dos dados. A PPDM (Privacy Preserving Data Mining ou Preservação de Privacidade dos Dados Minerados) visa desenvolver algoritmos que possam preservar a privacidade de modo a não trazer efeitos negativos aos resultados da mineração. A PPDM pode ser classificada em duas categorias: mineração de dados centralizados e mineração de dados distribuídos.

Figura 5 – Categorias da PPDM

Distribuição de dados. (a) dados centralizados. (b) dados horizontais. (c) dados verticais.

Fonte: L. Xu, C. Jiang, J. Wang, J. Yuan and Y. Ren (2014)

Algumas abordagens que utilizam a Teoria dos Jogos[3] já foram propostas para esse tipo de problema, especialmente para casos de mineração de dados distribuídas. Nesta metodologia da PPDM, é adotado um protocolo chamado de computação segura de múltiplas partes (SMC ou *Secure Multi-Party Computation*) cujo objetivo é garantir o resultado correto da mineração de dados sem revelar quaisquer dados para outros. A aplicação da Teoria dos Jogos se dá pela utilização de um algoritmo baseado em uma técnica de compartilhamento secreto de Shamir[4] para evitar conluio entre as partes envolvidas, ou seja, que algumas das partes acesse informação privada dos usuários. O equilíbrio de Nash seria alcançado quando todas as partes enviam ações e obtêm um comportamento de não conluio.

Pode-se dizer que existe um paralelo que se faz entre mecanismos de proteção de dados e situações desenvolvidas pelo campo econômico como: racionalidade,

[3] A Teoria dos Jogos é o estudo acerca da tomada de decisão de um indivíduo quando o resultado dessa escolha depende de uma relação de interdependência com outras partes.

[4] O compartilhamento secreto de Shamir é conhecido como o método de algoritmo criptográfico proposto pelo criptógrafo israelense Adi Shamir baseado em estimar valores desconhecidos em uma lacuna entre dois pontos de dados conhecidos sem a necessidade de saber nenhuma informação sobre o que está em cada lado desses pontos. Esse algoritmo possibilita que várias partes que não se conheçam armazenem informações privadas. Assim, possibilita, por exemplo, armazenar com segurança os dados dos usuários (como chaves de acesso e informações pessoais).

utilidade e equilíbrio. Nessas abordagens, cada usuário diferente busca altos interesses próprios em termos da preservação de seus dados ou, por outro lado, a utilidade dos dados. Usando metodologia das teorias dos jogos e aplicadas regras algorítmicas, são desenvolvidas implicações úteis quanto ao comportamento do usuário. Dessa maneira, os temas econômicos são estendidos ao campo computacional.

De forma semelhante, é possível questionar os pressupostos que compõe tal metodologia, isto porque existem limitações quanto a simplificação e generalização do modelo. Existe certo desprezo pelo contexto e preferências de cada indivíduo. Para isso, são considerados pressupostos de racionalidade dos atores, maximização da utilidade e existência de equilíbrio. Esses pressupostos muitas vezes não condizem com a realidade e são questionados por teóricos econômicos.

Paralelamente, é importante a preocupação quanto a generalização da preferência do usuário quanto a sua privacidade. Como explorado anteriormente, a opinião quanto a sensibilidade de cada tipo de dado muitas vezes variam de pessoa para pessoa. É relevante que se exista um tipo de preservação de privacidade personalizada de ampla utilização, mesmo que seja necessária uma fase de adaptação do usuário das redes para tal mudança.

Além disso, confiabilidade no resultado da mineração é uma preocupação do tomador de decisão (o último usuário da cadeia da mineração de dados). Quanto mais generalista e simplificado o resultado, menor o nível de representação da realidade. Ademais, na maioria dos casos, a proveniência dos resultados da mineração de dados não está disponível. Existe uma falta de controle e transparência quanto ao resultado.

Para evitar esse tipo de situação e garantir maior credibilidade ao resultado, é importante a implementação de protocolos e normas específicas para o Big Data. A inexistência de tal controle permite resultados de mineração de dados distorcidos e podem gerar uma influência errônea em relação a tomada de decisão. Os protocolos e normas específicas relativos ao Big Data serão explorados no item a seguir.

3.1.3 Controle do Big Data.

Existem dois âmbitos de controle quanto falamos de segurança da informação contida no Big Data, o primeiro é representado por um caráter técnico enquanto o segundo, como um mecanismo legal, que depende diretamente das regras e regulamentações impostas pelo governo. Podemos simplificar o entendimento dividindo as ações que envolvem o caráter técnico como aquelas específicas ao ambiente digital, enquanto as ações de caráter legal, ao ambiente físico.

O controle proporcionado pelo caráter técnico, envolvem o ambiente digital e compreendem questões relacionadas à programação, criptografia, armazenamento, entre outros. Com relação a ele, será utilizado como exemplo, o desafio mais urgente considerado pela CSA no que diz respeito a privacidade e segurança das redes: o controle de acesso. A propriedade de segurança que importa da perspectiva do controle de acesso é o sigilo, ou seja, impedir o acesso aos dados por pessoas que não deveriam ter acesso. Para garantia do sigilo, existem protocolos de segurança[5] que utilizam criptografia ao autenticarem qualquer tipo de acesso.

Esses protocolos de segurança, porém, estão muitas vezes expostos a falhas e podem contribuir para o comprometimento de dados sensíveis dos usuários. A própria autenticação mais comumente usada, a baseada em login e senhas de acesso, podem sofrer violação por hackers. Para essas lacunas, existe um desenvolvimento crescente de tecnologias baseadas em blockchain[6] que permitem minar tais problemas.

Como o Big Data é um sistema distribuído complexo, que é de difícil gerenciamento centralizado, novas abordagens de segurança são necessárias. A autenticação e o controle de acesso aos dados devem ser gerenciados de forma descentralizada, flexível, escalonável e forte que evite qualquer tipo de acesso de usuários não autorizados. Alguns autores sugerem, por exemplo, um tipo de autenticação sem senha, que utilizam tecnologia de blockchain e são baseadas no

[5] Os protocolos de rede englobam um conjunto de normas que permitem que qualquer máquina conectada à Internet possa se comunicar com outra também já conectada à rede. Funciona como se fosse uma “língua universal” entre computadores. Entre os protocolos de rede, existem àqueles destinados especificamente a segurança. A estes, são utilizados métodos de criptografia de forma a garantir confidencialidade, integridade e anonimato das informações.

[6] O blockchain é uma tecnologia de registro que permitiu o funcionamento e transação seguros das chamadas criptomoedas (bitcoin etc.). Ele permite rastrear o envio e recebimento de algum tipo de informação pela internet a partir de códigos complexos gerados, o que garante que os dados não sejam violados.

chamado Número de Identidade Seguro (ou SIN – “Secure Identity Number”). O SIN funcionaria como uma identidade digital. Uma identidade digital é uma credencial validada para o mundo digital, semelhante à identidade de uma pessoa para o mundo real, que poderia ser emitida e regulamentada por um esquema de identificação nacional. Esta tecnologia é fortemente apresentada como uma tendência para os próximos anos e tem grande valor econômico.

No próprio Fórum Econômico Mundial (em Davos 2019), foi apresentado em conferência a discussão acerca do “Valor da Identidade Digital para a Economia Global e a Sociedade”. Isto porque:

> “A identidade digital pode aumentar a inclusão política e social, facilitar a proteção de direitos e promover a transparência. Ela também pode impulsionar o crescimento econômico inclusivo e criar valor substancial para os indivíduos, governos e empresas em todo o mundo (...) os países que implementam um bom uso da identidade digital podem criar valor econômico equivalente a 3-6% do PIB em média até 2030. Para desbloquear este potencial, precisamos criar uma definição e padrões compartilhados que garantam que a identidade digital seja segura e respeite a privacidade e outros direitos fundamentais” (Traduzido de YouTube, 2019).

Já sobre o controle proporcionado pelo aspecto legal, podemos usar o exemplo da Lei Geral de Proteção de Dados promulgada no Brasil em 2018. Ela dispõe sobre:

> “o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais da liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural.” (Ministério da Defesa, 2020)

Ela traz garantias quanto a privacidade, confidencialidade e controle sobre seus dados pessoais:

> “Toda pessoa natural tem assegurada a titularidade de seus dados pessoais e garantidos os direitos fundamentais de liberdade, de intimidade e de privacidade, nos termos da LGPD (artigo 17 da LGPD).
>
> O titular dos dados pessoais tem direito a obter do controlador, em relação aos dados do titular por ele tratados, a qualquer momento e mediante requisição:
>
> I - Confirmação da existência de tratamento;

II - Acesso aos dados mantidos pelo controlador;

III - Correção de dados incompletos, inexatos ou desatualizados;

IV - Anonimização, bloqueio ou eliminação de dados desnecessários, excessivos ou tratados em desconformidade com o disposto na LGPD;

V - Portabilidade dos dados a outro fornecedor de serviço ou produto, mediante requisição expressa;

VI - Eliminação dos dados pessoais tratados quando revogado o consentimento dado pelo titular;

VII - Informação com quem o controlador realizou compartilhamento de seus dados;

VIII - Informação sobre a possibilidade de não fornecer consentimento e sobre as consequências da negativa;

IX - Revogação do consentimento.” (Ministério da Defesa, 2020)

Experiências recentes mostraram a vulnerabilidade de bancos de dados conectados à internet. Um dos casos mais emblemáticos de negligência com informações foi o vazamento de dados de milhões de usuários do Facebook pela empresa Cambridge Analytica. A LGDP, apesar de ainda não estar em vigência, pretende exigir a adequação de qualquer operação que envolva tratamento de dados pessoais no Brasil. Assim, pode-se dizer que os negócios serão impactados profundamente. Empresas e instituições devem se proteger não somente de multas que podem ocorrer, mas também de qualquer exposição midiática negativa que uma vulnerabilidade pode causar.

Com a garantia legal que a LGDP promove, cabe aos cidadãos se conscientizarem sobre seus direitos e deveres quanto aos seus dados disponibilizados na Internet. Acompanhando a rapidez das inovações e modernizações que o Big Data promove, a necessidade de adaptação deve seguir velocidade similar. Essa é a maior barreira que o controle e segurança da informação enfrenta, a percepção rápida e eficaz de todas as partes. Muitas vezes as transformações tecnológicas não são seguidas de um processo uniforme de conscientização social, principalmente devido a barreiras de caráter econômico e social. Afinal, requisições, revogações, portabilidade e correções garantidas por lei só serão atendidas mediante solicitação prévia dos usuários. É preciso garantir que

o grande volume de informações que dispomos nas redes seja de total conhecimento e transparência com os cidadãos.

Podemos concluir que o Big Data proporcionou oportunidades importantes, porém, trouxe também preocupações com relação a segurança e proteção da privacidade das informações. Em resumo, é necessário o desenvolvimento de três pilares principais como forma de evitar maiores prejuízos nesse quesito, são eles: técnico com o suporte de tecnologia de segurança da informação, conscientização social sobre o uso dos dados e suporte legal amparado pelo Estado.

Conclusão

É inegável a nível de transformação que o Big Data proporcionou ao mecanismo de tomada de decisão. Apoiado por técnicas modernas de análises que extrapolaram a análise de dados manuais e de dados tradicionais, essa mudança significa uma completa mudança de paradigma. Muitos autores sugerem que proporcionou uma novo passo na escala da evolução da informação, possibilitando uma nova etapa do que conhecemos como “Era da Informação” para a “Era do conhecimento”. Essa mudança é relativamente nova e muitas vezes as pesquisas acerca do tema são restritas a pesquisadores de países mais avançados tecnologicamente.

O presente trabalho buscou esclarecer como o Big Data permitiu essa mudança de paradigma, o processo pelo qual a tomada de decisão se insere com as novas tecnologias e como nós enquanto usuários participamos desse processo. Visto que toda a decisão é uma aposta incerta e probabilística com base em informações prévias, quanto mais aprimorada a estrutura dessas informações nas quais baseamos nossas estimativas, menor o nível de incerteza e maior a assertividade da tomada de decisão. Entre os aspectos que a análise de dados fornecida pelo Big Data proporciona na melhoria no processo de tomada de decisão estão: a rapidez/adaptabilidade, precisão, consistência e transparência.

Essa habilidade de lidar com um menor nível de incerteza se caracteriza como um ativo econômico importante na atualidade. Lidar com a incerteza proporcionada pelo rápido ritmo de transformação que as mudanças no cenário econômico, institucional e tecnológico estão sujeitas se tornou objetivo fundamental em todos os setores da economia.

Na macroeconomia podem dar suporte nas considerações quanto as interações dos agentes e suas heterogeneidades, para além dos modelos dos agentes representativos neoclássicos. Na saúde garantem maior precisão e possibilita um uso mais acertado dos recursos, com melhor escolha de medicação e dosagem para pacientes em tratamento. Na sustentabilidade possibilita monitoramento em tempo real e decisões mais acertadas quanto a otimização dos recursos naturais. Nos negócios, revolucionou a publicidade ao garantir maior assertividade e rapidez para atingir o público-alvo e converter vendas. Também proporcionou uma mudança em todo o princípio da competitividade, propondo uma mudança não só operacional como também de todo o modelo de negócio de várias empresas.

Essa nova categoria de ativos, porém, alimentou uma discussão quanto a segurança e privacidade dos usuários. Isto porque dados pessoais são tratados como uma verdadeira mercadoria. Serviços aparentemente gratuitos, são pagos em troca de dados do comportamento do consumidor. Informações estão sendo usadas para além do que foram autorizadas inicialmente. Por isso, passa a ser fundamental mecanismos técnicos e legais que defendam o direito do usuário online. Para além desses fatores, cabe também um processo uniforme de conscientização e adaptação quanto aos seus direitos e deveres.

Referências Bibliográficas

NEO WAY. **O que é Big Data e qual a importância de implementá-lo na empresa?.** Jul de 2020. Disponível em: www.neoway.com.br/o-que-e-big-data . Acessado em Maio de 2020.

FORBES. **12 Big Data Definitions: What's Yours?** Set de 2014. Disponível em: www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/. Acessado em Maio de 2020.

FORBES. **10 Charts That Will Change Your Perspective Of Big Data's Growth,** Maio de 2018. Disponível em: www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/ . Acessado em Mai de 2020.

IRON MOUNTAIN. (2012). **The Impact of Big Data on Government**. Disponível em: www.ironmountain.com/resources/whitepapers/t/the-impact-of-big-data-on-government . Acessado em Mai de 2020.

SUMMIT SAÚDE. **Big Data: o uso da informação na luta contra o coronavírus,** Maio de 2020. Disponível em: www.summitsaude.estadao.com.br/big-data-o-uso-da-informacao-na-luta-contra-o-coronavirus/ . Acessado em Maio de 2020.

EL PAÍS. **Os Big Data do coronavírus,** Março de 2020 Disponível em: www.brasil.elpais.com/opiniao/2020-03-09 . Acessado em Mai de 2020.

KHALID, B. **Big Data in Economic Analysis: Advantages and Challanges,** Agosto de 2019 Disponível em: www.researchgate.net/publication/335234998_BIG_DATA_IN_ECONOMIC_ANALYSIS_ADVANTAGES_AND_CHALLENGES . Acessado em Mai de 2020.

HILBERT, M. **Big Data for Development: A Review of Promises and Challanges,** 2016. Disponível em : www.researchgate.net/publication/286907720_Big_Data_for_Development_A_Revie

w_of_Promises_and_Challenges. Acessado em Mai de 2020.

ECLAC. **Big data and open data as sustainability tools: A working paper prepared by the Economic Commission for Latin America and the Caribbean,** Outubro de 2014. Disponível em: www.repositorio.cepal.org/handle/11362/37158/. Acessado em Mai de 2020.

ABREU, G. NICOLAU, M. **Big Data, publicidade e o consumidor dataficado: o caso da série House of Cards.** Disponível em: www.periodicos.ufpb.br/index.php/cm/article/view/35074. Acessado em Mai de 2020.

OMS. **Global research on coronavirus disease (COVID-19).** Disponível em: www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov. Acessado em Mai de 2020.

EINAV, L. LEVIN, J. **The Data Revolution and Economic Analysis,** Maio 2013. Disponível em: www.nber.org/papers/w19035 . Acessado em Mai de 2020.

D'ORAZIO, Paola, 2017. **Big data and complexity: Is macroeconomics heading toward a new paradigm?** Journal of Economic Methodology. 10.1080/1350178X.2017.1362151.

NICKERSON, David W., e ROGERS,T. **Political Campaigns and Big Data**. HKS Faculty Research Working Paper Series RWP13-045, Revisado Fevereiro 2014.

ZHU, Christina., **Big Data as a Governance Mechanism**, Julho de 2018. Disponível em SSRN: www.ssrn.com/abstract=3164624.

YLIJOKI, Ossi & PORRAS, Jari. (2016). **Perspectives to Definition of Big Data: A Mapping Study and Discussion**. Journal of Innovation Management. 4. 69-91. 10.24840/2183-0606_004.001_0006.

KHALID, Balar & RACHID, Chaabita. (2019). **Big data in economic analysis: advantages and challenges**. 04. 5196.

HILBERT, Martin. (2016). **Big Data for Development: A Review of Promises and Challenges**. Development Policy Review. 34. 135-174. 10.1111/dpr.12142.

ECLAC. (2014). **Big data and open data as sustainability tools: A working paper prepared by the Economic Commission for Latin America and the Caribbean**.

COX, Michael & ELLSWORTH, David. (1997). **Managing big data for scientific visualization**.

LANEY, D. **3D Data Management: Controlling Data Volume, Velocity, and Variety**, Gartner, file No. 949. (2001). Disponível em: http://blogs.gartner.com/douglaney/files/2012/01/ad949-3D-Data-Management ControllingData-Volume-Velocity-and-Variety.pdf

HAIDER, Murtaza e GANDOMI, Amir. (2015). **Beyond the hype: Big data concepts, methods, and analytics**

SVIOKA, J. e RAYPORT, J. (1995). **Exploiting the Virtual Value Chain.** Disponível em: hbr.org/1995/11/exploiting-the-virtual-value-chain

PROVOST, Foster & FAAWCETT, Tom. (2013). **Data Science and Its Relationship to Big Data and Data-Driven Decision Making**. Big Data. 1. 10.1089/big.2013.1508.

XCUBE LABS**. How Is Big Data Influencing Digital Marketing Strategy?** Disponível em: https://www.xcubelabs.com/blog/how-is-big-data-influencing-digital-marketing-strategy/#:~:text=Big%20Data%20plays%20a%20key,customer%27s%20behavior%20around%20the%20world. . Nov de 2017. Acessado em Agosto de 2020.

D'ORAZIO, Paola. (2017). **Big data and complexity: Is macroeconomics heading toward a new paradigm?.** Journal of Economic Methodology, 2017 Vol. 24, No. 4, 410–429.

How Big Data and Artificial Intelligence Can Help Better Manage the COVID-19

Pandemic. Int. J. Environ. Res. Public Health, 2020**.** Disponível em: https://www.mdpi.com/1660-4601/17/9/3176/htm

FAPESP. **Proposals to FAPESP Research Program on eScience and Data Science: Digital Human Health.** 2019. Disponível em: https://fapesp.br/en/13653/call-for-proposals-escience-2019.

ESTADAO. **Big Data: o uso da informação na luta contra o coronavírus.** 2020. Disponível em: https://summitsaude.estadao.com.br/tecnologia/big-data-o-uso-da-informacao-na-luta-contra-o-coronavirus/

YouTube. (2019). **Davos 2019 – Press Conference The Value of Digital Identity for the Global Economy and Society.** Disponível em: https://www.youtube.com/watch?v=1-V7lyxrOmw&t=18s

MINISTERIO DA DEFESA. (2020). **Lei Geral de Proteção de Dados – LGPD.** Disponível em: https://www.gov.br/defesa/pt-br/acesso-a-informacao/lei-geral-de-protecao-de-dados-pessoais-lgpd

FILHO, Alexandre. (2015). **Uso do Big Data em saúde no Brasil: perspectivas para um futuro próximo.** Disponível em: https://www.scielo.br/scielo.php?pid=S2237-96222015000200325&script=sci_abstract&tlng=pt

MACKINSEY GLOBAL INSTITUTE. (2016). **The Age of Analytics: Competing in a Data-Driven World.** Disponível em: https://www.mckinsey.com/~/media/McKinsey/Industries/Public%20and%20Social%20Sector/Our%20Insights/The%20age%20of%20analytics%20Competing%20in%20a%20data%20driven%20world/MGI-The-Age-of-Analytics-Full-report.pdf

THALES. **Digital identity trends – 5 forces that are shaping 2020**. Disponível em: https://www.thalesgroup.com/en/markets/digitalidentityandsecurity/government/identity/digital-identity-services/trends

ABDULLAH, N. & HAKSSON, A. & MORADIAN, E. (2017). **Blockchain based**

approach to enhance big data authentication in distributed environment. Disponível em: https://www.researchgate.net/publication/318743976_Blockchain_based_approach_to_enhance_big_data_authentication_in_distributed_environment

CLOUD SECURITY ALLIANCE. (2012). **Ten Big Data Security and Privacy Challenges.** Disponível em: https://s3.amazonaws.com/content-production.cloudsecurityalliance/

XU, L. & JIANG, C. & WANG, J. & YUAN, J. & REN, Y. (2014). **Information Security in Big Data: Privacy and Data Mining**. Disponível em: https://ieeexplore.ieee.org/document/6919256

KHALID, B. & RACHID, C. (2019). **Big data in economic analysis: advantages and challenges - International Journal of Social Science and Economic Research.**