BIG DATA E CIÊNCIA DE DADOS: COMPLEMENTARIEDADE CONCEITUAL NO PROCESSO DE TOMADA DE DECISÃO

Big Data and Data Science: conceptual complementarity in the decision-making process

Sandro Rautenberg (1), Paulo Ricardo Viviurka do Carmo (2)

(1) Universidade Estadual do Centro-Oeste, srautenberg@unicentro.br. (2) pauloviviurka4@gmail.com

Resumo

Atualmente, produz-se volumosas bases de dados para os mais variados domínios, devido à evolução e ao uso massivo das Tecnologias de Informação e Comunicação. Neste contexto, desenvolver instrumentos voltados para extrair informações a partir do *Big Data*, fomentando o conhecimento no Processo de Tomada de Decisão, desperta a atenção das organizações. Na evolução dados → informação → conhecimento, requer-se a sinergia de competências de especialistas ao fomentar uma nova área de atuação, a Ciência de Dados. Como atribuições da área estão: obter dados originados de fontes heterogêneas e distribuídas na web; formalizar o tratamento dos dados e metadados; e arquitetar a exploração dos dados e metadados para produzir informação relevante no Processo de Tomada de Decisão. Considerando essas assertivas, o objetivo deste artigo é evidenciar a diferença e a complementariedade dos conceitos *Big Data* e Ciência de Dados. Como resultado, pontua-se que o *Big Data* é um termo intrínseco à infraestrutura do hardware e de serviços computacionais em nuvem necessários para o armazenamento, processamento e a distribuição de recursos. Em relação à Ciência de Dados, este é conexo à camada dos softwares para transformação dados em informação, subsidiando os tomadores de decisão em suas Tarefas Intensivas em Conhecimento.

Palavras-chave: *Big Data*; Ciência de Dados; Processo de Tomada de Decisão; Tarefas Intensivas em Conhecimento.

Abstract

Nowadays, huge databases are produced in a wide range of domains due to the evolution and the massive use of the Information and Communication Technologies. Therefore, the development of instruments for extracting information from Big Data and fostering actionable knowledge in Decision-Making Processes arise interest by several organizations. In this context, the evolution of data → information → knowledge requires the synergy of competences in a new domain, the Data Science. Among the activities of that domain, can be cited: obtaining data from various sources distributed on the web; creating models for handling data and metadata; and planning the exploration of data and metadata to produce relevant information in Decision-Making Processes. Considering these statements, this paper aims to discuss the difference and the complementarity between the Big Data and Data Science concepts. As a result, it is pointed out that Big Data delineates the cloud computing services for storing, processing and distributing data resources. Regarding to Data Science, it is a concept related to the use of software for transforming data into information, supporting the decision makers when dealing with the Knowledge-Intensive Tasks.

Keywords: Big Data; Data Science; Decision-Making Process; Knowledge-Intensive Tasks.

1 Introdução

É notório que os avanços e o uso massivo das Tecnologias da Informação e Comunicação influenciam o comportamento de um coletivo social. Tal fato instiga amplos debates no campo da Ciência da Informação no que tange, principalmente, a utilização salutar de dados, informação e conhecimento gerados a partir dos rastros digitais produzidos por artefatos computacionais (câmeras, celulares, cartões de credito, sensores de vários tipos, etc.). Por isso, no âmbito da Ciência da Informação, admite-se a necessidade de um espaço interdisciplinar para discutir questões polêmicas sobre informação, conhecimento e ação autônoma, relacionando-as com o fenômeno tecnológico denominado *Big Data* (Eiica, 2019).

Corroborando essa visão de mundo, mediante o avanço da Internet, a humanidade vem produzindo cada vez mais dados nas mais variadas plataformas digitais (Figura 1 no apêndice). Vários dispositivos interconectados (sensores, computadores, câmeras, dentre outros) e aplicativos relacionam uma miríade de eventos na web (van der Aalst, 2014), coletando e armazenando enormes quantidades de registros, sinais, imagens, vídeos e *posts*. Bugnion, Manivannan e Nicolas (2017) pontuam que cerca de 90% dos dados produzidos são resultado do uso intenso das Tecnologias de Informação e Comunicação nos últimos tempos. Por conseguinte, os dados são abundantemente e velozmente produzidos, servindo de matéria-prima para tomada de decisão em grandes corporações (Economist, 2017).

Neste contexto, o desenvolvimento de soluções computacionais que obtém insumos de conhecimento de imensas bases de dados é foco de investimento em grandes organizações. Isso introduz o conceito de *Big Data*, referindo-se aos conjuntos de dados cujo tama-

nho é fator impeditivo de captura, armazenamento, gerenciamento e análise por parte de ferramentas computacionais tradicionais (Manyika *et al*., 2011). Ou seja, o *Big Data* requer formas inovadoras de processamento de grandes volumes de dados heterogêneos, amparando o Processo de Tomada de Decisão guiado por Dados (Gartner, 2018a; Provost e Fawcett, 2013). Por isso, atualmente, enfrenta-se desafios tecnológicos para coletar, guardar e disponibilizar volumosos conjuntos de dados e produzir informação relevante.

Neste sentido, o *Big Data* também requer que seus profissionais detenham competências diversas na organização, representação de dados para, em um segundo momento, desenvolver ações de recuperação e visualização de informação nos processos decisórios. Por isso, pressupõe-se que salvaguardar volumosas coleções de dados (*Big Data*) distingue-se da produção de informação a partir dessas coleções.

Essa distinção conceitual para com o *Big Data*, complementarmente, enseja a Ciência de Dados. Em suma, a Ciência de Dados é devotada à extração de informação útil a partir de imensas, complexas e dinâmicas bases de dados (Bugnion, Manivannan e Nicolas, 2017). Entende-se que a Ciência de Dados é um conceito conexo à camada dos métodos, na qual os softwares são empregados para transformar dados em informação, resultando no apoio à tomada de decisão.

Ao considerar a evolução dados → informação → conhecimento, disserta-se sobre os conceitos *Big Data* e Ciência de Dados, apresentando o *locus* interdisciplinar de competências das Ciências da Informação e da Computação.

Para fomentar a discussão, além desta seção introdutória, este artigo aborda: i) o conceito *Big Data*, estabelecendo seu relacionamento com a Curadoria Digital; ii) a Ciência de Dados como método de transformação de dados em informação; iii) o Processo de Tomada de Decisão, amparando-se nas Tarefas Intensivas em Conhecimento; iv) a discussão da complementariedade do *Big Data* e Ciência de Dados em processos decisórios; e v) as considerações finais.

2 *Big Data*: a camada dos materiais e da Curadoria Digital

O *Big Data* é um termo derivado dos avanços recentes relativos à massificação da utilização de recursos tecnológicos e da farta produção de dados. Em suma, é um conceito que caracteriza volumosos conjuntos de dados heterogêneos, os quais não são passíveis de processamento por soluções computacionais tradicionais, considerando seu dinamismo e sua complexidade. Originalmente, o *Big Data* preconizava três características essenciais dos dados, denominadas por Laney (2001) como 3Vs:

- **Volume**. Grandes volumes de dados são gerados mediante o uso de recursos computacionais abundantes. Com a evolução das mídias sociais e outros recursos e serviços da Internet, as pessoas produzem mais e mais conteúdo, vídeos, fotos, *tweets*, entre outros tipos de dados.
- **Velocidade**. Os dados são gerados em grande velocidade, à medida que os recursos computacionais têm sua capacidade de produção, captura e processamento de dados aumentada.
- **Variedade**. Os dados advêm de variadas fontes (sistemas legados, *e-mails*, *posts* em mídias sociais, arquivos de vídeo/áudio, gráficos, dispositivos ou sensores), as quais implementam tecnologias distintas para representação e armazenamento de recursos digitais.

Ao considerar o atual estágio da utilização de Tecnologias de Comunicação e Informação, outros Vs são adicionados aos 3Vs originais, conforme a visão de especialistas ou o domínio de aplicação. Neste sentido, Akhtar (2018) pontua a existência de 6Vs (Figura 2), incrementando as características com:

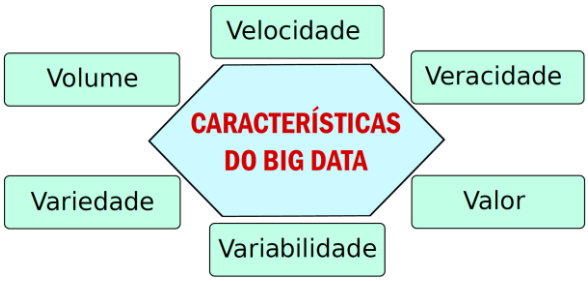

Figura 2. Representação dos 6Vs do *Big Data* (Akhtar, 2018) [tradução dos autores]

- **Veracidade**. Refere-se à integridade e à precisão dos dados, contrapondo o fenômeno GIGO (*garbage-in, garbage-out* – lixo entra, lixo sai) na recuperação da informação. Neste sentido, deve-se evitar ruídos e incertezas no armazenamento dos dados de modo a não interferir, consequentemente, na análise da informação e no Processo de Tomada de Decisão.
- **Variabilidade**. Relaciona-se à compreensão e ao tratamento dos fenômenos subliminares e temporariamente presentes nos dados. Por exemplo, sazonalmente, alguns eventos específicos (virais nas mídias sociais, como a estreia de um filme a muito aguardado ou o acontecimento de um fato midiático) podem refletir em padrões de comportamento que não se sustentam ao longo do tempo.
- **Valor**. É característica mais importante em termos dos dados, independente das demais dimensões (volume, velocidade, variedade, variabilidade e veracidade). O valor em *Big Data* é, principalmente, percebido mediante a análise com dados precisos e,

por conseguinte, a aquisição de informação e *insights* úteis para o Processo de Tomada de Decisão.

Dadas as características do *Big Data*, algumas questões importantes afloram. Por exemplo:

- Como armazenar os dados e metadados em ecossistemas de *Big Data*?
- Como organizar e catalogar os dados e metadados armazenados nesses ecossistemas?
- Como garantir que os dados críticos estejam disponíveis no *Big Data* para o Processo de Tomada de Decisão?

Interdisciplinarmente, essas questões ensejam algumas competências da Ciência da Informação, introduzindo a Curadoria Digital (Figura 3 no apêndice) como elemento importante na definição do ecossistema de *Big Data*.

A Curadoria Digital é um conceito vinculado à veracidade e à proveniência, bem como à garantia da qualidade dos dados (Roy, Underwood e Chang, 2015). Em suma, a Curadoria Digital é envolta por boas práticas de planejamento e de gestão de dados. No contexto de ecossistemas de *Big Data*, conforme seu ciclo de vida (Digital Curation Centre, 2018), a Curadoria Digital pode auxiliar em:

- **Conceituar**. É a formalização de documentos que definem as orientações, as políticas, os requisitos legais e ações de criação, representação, captura, limpeza, avaliação e guarda dos dados e metadados.
- **Criar ou Receber**. São as ações para criar dados em um ecossistema de *Big Data*. Os metadados decorrentes dessas ações (metadados administrativos, descritivos, estruturais, técnicos e de preservação) também devem ser considerados/mantidos. Na criação ou recebimento de dados, deve-se proceder em consonância às políticas de coleta documentadas na fase conceituar.
- **Avaliar e Selecionar**. Antes de inserir novos dados no ecossistema de *Big Data*, deve-se avaliar os dados quanto aos requisitos de qualidade estabelecidos (as orientações, as políticas e os requisitos legais de criação, captura e guarda de dados). Uma vez avaliados, seleciona-se o conjunto íntegro de dados para ser custodiado e preservado.
- **Inserir**. Definido o conjunto íntegro de dados, o próximo passo é armazenar os dados no ecossistema do *Big Data*, de acordo os documentos previamente formalizados.
- **Ação de preservação**. Realiza-se as ações para garantir a preservação dos dados ao longo do tempo. As ações de preservação são previamente definidas e devem ser orquestradas de modo que os dados permaneçam autênticos, confiáveis e usáveis, mantendo perenemente sua integridade.
- **Armazenar**. Ao custodiar os dados e metadados, deve-se garantir que estes sejam mantidos seguramente, utilizando tecnologias apropriadas para armazenamento e representação num ecossistema de *Big Data*.
- **Acesso, uso e reutilização**. Possibilitar que os dados sejam facilmente acessíveis pelos usuários. Os controles de acesso/autenticação devem ser implementados, de acordo com as políticas previamente definidas.
- **Transformar**. Em algumas circunstâncias, existe a possibilidade de sumarizar ou derivar novos dados a partir dos dados armazenados.
- **Descarte**. Ocasionalmente, pode ocorrer a remoção de dados (desatualizados, invalidados, ou por orientação legal) conforme as políticas documentadas. Normalmente, os dados são retirados de um ambiente de produção, sendo transferidos para um arquivo morto passível de recuperação. Em outros casos, os dados são definitivamente destruídos, por razões legais que sustentam a destruição segura.
- **Reavaliar**. Quando necessário, pode-se reavaliar uma versão mais recente dos dados que anteriormente não foram validados de acordo com os procedimentos formalizados na fase conceituar.
- **Migrar**. Em virtude de avanços tecnológicos, deve-se executar ações de migração dos dados para um formato mais atual. Desta forma, preserva-se os dados e metadados a longo prazo, mesmo ocorrendo a obsolescência de hardware ou de software em ecossistemas de *Big Data*.

Diante o exposto, percebe-se que a definição de um ecossistema de *Big Data* perpassa por várias competências. Neste sentido, a National Science Foundation (2005) pontua que os cientistas da informação e da computação são agentes cruciais e devem cooperar na guarda perene dos dados digitais.

No contexto deste artigo, entende-se que o *Big Data* se reserva à infraestrutura de manutenção grandes coleções de dados, atuando como a camada de suporte para extrair informações relevantes dessas coleções. Neste sentido, adicionalmente, pontua-se que a extração de informação é uma atividade desafiadora, considerada complementar ao *Big Data*, ensejando o conceito Ciência de Dados. Como discutido a seguir, admite-se que Ciência de Dados circunscreve as soluções computacionais que, a partir dos dados, abstraem insumos úteis no Processo de Tomada de Decisão (Grady e Chang, 2015).

3 Ciência de Dados: a camada dos métodos de transformação dos dados em informação

Aliado ao surgimento do *Big Data*, tem-se o advento da Ciência de Dados como um campo de atuação de

competências interdisciplinares em ascensão. Atribui-se à Ciência de Dados a extração de informação útil a partir de imensas bases de dados complexas, dinâmicas, heterogêneas e distribuídas (Bugnion; Manivannan; Nicolas, 2017). Conforme a Figura 4, para se atuar na Ciência de Dados, três domínios de conhecimento se inter-relacionam: Programação de Computadores; Estatística e Matemática; e Domínio do Conhecimento. Neste sentido, existem três pressupostos:

- Dentre as habilidades necessárias na Ciência de Dados, seus especialistas devem apresentar habilidades na área da Ciência da Computação, visto que basilarmente os dados são armazenados, manipulados e transmitidos por computadores. Neste contexto, os ambientes computacionais para o Desenvolvimento de Software são ferramentas essenciais para promover a Curadoria Digital e a implementar os algoritmos de Aprendizado de Máquina e das interfaces de Visualização da Informação. É imperativo saber utilizar essas tecnologias de modo a acessar e transformar os dados para abstrair e representar informação útil.

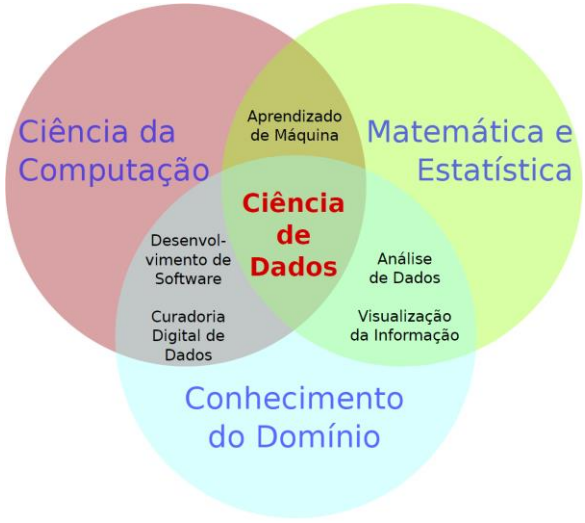

Figura 4. Interdisciplinaridade da Ciência de Dados (baseado em Conaway, 2010) [tradução dos autores]

- O conhecimento sobre Matemática e Estatística também é necessário para a realização de atividades de Análise de Dados. Ou seja, os profissionais da Ciência de Dados devem entender o funcionamento dos algoritmos de Aprendizado de Máquina, bem como, saber interpretar os resultados, estatisticamente. Interdisciplinarmente, a atividade de interpretação é facilitada pela Visualização da Informação, a qual privilegia a utilização elementos de representação gráfica da informação.
- Para o efetivo sucesso das soluções de Ciência de Dados, o Conhecimento do Domínio do problema deve ser disponível e amplamente utilizado no Processo de Tomada de Decisão. Neste sentido, as soluções de Ciência de Dados são voltadas para a formulação de hipóteses e a aquisição de informação aderente como insumo no processo decisório.

Ressalta-se que em ecossistemas de *Big Data*, o Processo de Tomada de Decisão é guiado por dados (Provost; Fawcett, 2013). Como pode ser percebido nos pressupostos relatados, geralmente, tal processo emprega soluções computacionais baseadas em algoritmos de Aprendizado de Máquina à aquisição de informação relevante. Conceitualmente, o Aprendizado de Máquina é uma subárea da Inteligência Artificial que investiga a captura automatizada de modelos de abstração de informação a partir registros contidos em (volumosas) bases de dados (Blum; Hopcroft; Kannan, 2018). Em outras palavras, o Aprendizado de Máquina aplica métodos computacionais e/ou estatísticos para a extração automatizada de informação útil a partir de dados históricos. Neste contexto, como métodos computacionais, pode-se citar:

- **Redes Neurais Artificias.** São modelos computacionais que imitam o funcionamento mais básico do cérebro humano. Em poucas palavras, similarmente ao cérebro quando acionado em relação a um evento, uma Rede Neural Artificial recebe estímulos (sinais de entradas), processa sinais e produz uma saída (Munakata, 2008). Como soluções de Aprendizado de Máquina aplicadas ao Processo de Tomada de Decisão, as Redes Neurais Artificiais são empregadas nas Tarefas Intensivas em Conhecimento (ver *Seção 4*) de: Associação; Avaliação; Diagnóstico; Monitoramento; e Predição.
- **Algoritmos Genéticos**. Resumidamente, são modelos computacionais baseados na teoria da evolução das espécies (Munakata, 2008). São fundamentados na premissa de que somente os seres mais adaptados ao ambiente têm maior chance de gerar descendentes. Computacionalmente, os Algoritmos Genéticos implementam: a seleção dos melhores indivíduos (soluções) baseada na aptidão à resposta de um problema; a reprodução das melhores soluções; e a ocorrência ocasional de mutação sobre as soluções. Com estas metáforas da Evolução das Espécies, um Algoritmo Genético otimiza a busca de uma solução ótima dentre várias soluções possíveis dado um problema. Os Algoritmos Genéticos, geralmente, são empregados em tarefas de: Associação; Avaliação; Diagnóstico; e Predição.
- **Inteligência Coletiva**. Foi originalmente introduzida no contexto de sistemas autônomos baseados na coletividade e auto-organização de simples agentes (Tarasewich; Mcmullen, 2002). Seus algoritmos são inspirados pela observação do comportamento de indivíduos que cooperam coletivamente na resolução de problemas globais, como por exemplo, o comportamento de uma colônia de formigas na busca por alimento. Em poucas palavras, Inteligência Coletiva é um paradigma de Aprendizado de

Máquina bio-inspirado, baseado na distribuição e no comportamento coletivo (enxame, cardume, revoada ou colônia) de elementos biológicos (formigas, cupins, abelhas, entre outros) para resolver problemas de otimização. Segundo Abraham; Guo e Liu (2006), os algoritmos de Inteligência Coletiva são utilizados em atividades de Mineração de Dados ou Descoberta de Conhecimento em Bases de Dados, sendo adequados às tarefas de: Associação; Avaliação; Diagnóstico; Monitoramento; e Predição.

Para Bugnion, Manivannan e Nicolas (2017), independentemente de método computacional de Aprendizado de Máquina utilizado, sete passos podem ser executados iterativamente em soluções de Ciência de Dados (Figura 5):

Figura 5. Ciclo de Vida da Ciência de Dados (Bugnion, Manivannan e Nicolas, 2017) [tradução dos autores]

- **Obtenção de Dados**. Preconiza a realização de tarefas de avaliação e seleção de dados primários e seus metadados a partir: do processamento de arquivos de texto; do monitoramento de uma rede de sensores; de consultas a bases de dados de sistemas legados; de dados oriundos da web; dentre outros.
- **Ingestão de Dados**. Trata da transformação e carga dos dados primários advindos de fontes diferentes e formatos diversificados em uma base de dados centralizada. Esta etapa implica em organizar e representar os dados de modo a inserir os recursos pré-processados em um repositório de dados principal, mitigando os esforços futuros da geração de informação relevante.
- **Exploração de Dados**. Privilegia a execução de estudos preliminares para estabelecer as conjecturas iniciais acerca dos dados disponibilizados em relação à informação requisitada. Neste sentido, esta atividade é importante para o estabelecimento do fluxo de trabalho (*workflow*), definindo o roteiro de como relacionar os dados primários à informação relevante.
- **Definição dos Parâmetros**. Passo intimamente ligado as escolhas necessárias para o emprego do(s) algoritmo(s) de Aprendizado de Máquina. Nesta atividade, por exemplo: i) converte-se os dados de entrada conforme os requisitos de manipulação do algoritmo de aprendizado; ii) transforma-se os dados de saída de modo a refletir uma saída legível aos seres humanos; iii) estabelece-se os intervalos dos parâmetros de entrada a serem considerados; iv) define-se os critérios de parada do algoritmo de aprendizado; v) o nível de confiabilidade exigido da resposta gerada; dentre outros.
- **Implementação do Modelo**. Prima-se pela utilização dos algoritmos de Aprendizado de Máquina para estabelecer modelos a partir dos dados de entrada e saída. Iterativamente, isso envolve o emprego de estratégias de treinamento e de testes dos algoritmos para a definição dos parâmetros mais adequados dentre aqueles avaliados. Como resultado, deve-se abstrair um modelo que estatisticamente melhor represente as características dos dados utilizados.
- **Utilização do Modelo**. Uma vez estabelecido um modelo, pode-se utilizá-lo para inferir informações sobre dados em um ambiente de produção. Isso confirmará o poder de generalização do modelo em gerar informação relevante perante situações do mundo real. Uma vez confirmado o poder de generalização, o modelo poderá ser empregado em Tarefas Intensivas em Conhecimento.
- **Tomada de Decisão**. Nas Tarefas Intensivas em Conhecimento, mediante a combinação do resultado gerado pelo modelo na análise dos dados com seu conhecimento especializado, o gestor ampara suas decisões tomadas. Uma parte fundamental nesta etapa envolve a customização da apresentação de dados e da visualização da informação através de relatórios e gráficos, respectivamente. Isso torna os *insights* mais claros e convincentes, auxiliando as atividades cognitivas dos tomadores de decisão.

Considerando as atividades relatadas, assume-se que a geração de informação útil a partir de dados brutos normalmente é um processo interativo. Neste sentido, os atores envolvidos podem formular premissas iniciais a respeito do problema e, gradualmente, refiná-las ao adicionar novas dimensões de dados ou testar outros algoritmos de Aprendizado de Máquina. Em outras palavras, diante um volumoso conjunto de dados de baixo nível, iterativamente, encontra-se outras formas de representação mais abstratas e úteis acerca dos dados para a Tomada de Decisão. Subliminarmente, isso evidencia o processo de evolução dados → informação → conhecimento. Ou seja, os dados são transformados em informações, que por sua vez, são agrupadas em

padrões apresentados ao usuário para avaliação, descoberta de novos conhecimentos e suporte à Tomada de Decisão (conhecimento em ação).

4 Processo de Tomada de Decisão: a camada das Tarefas Intensivas em Conhecimento

Considerando que o *Big Data* aporta grande volume de dados estruturados ou desestruturados para o processo decisório, a curadoria desse aporte deve permitir às organizações as condições para realizar análises, *insights* e/ou julgamentos baseados em dados precisos. Neste sentido, conforme a Figura 6, a Ciência de Dados configura-se como um suporte metodológico ao Processo de Tomada de Decisão, facilitando: a obtenção de informação contextualizada; a explicitação de fenômenos subliminares contidos nos dados; ou a refutação/confirmação de hipóteses previamente estabelecidas. Esse processo é denominado por Provost e Fawcett (2013) como Tomada de Decisão Guiada por Dados.

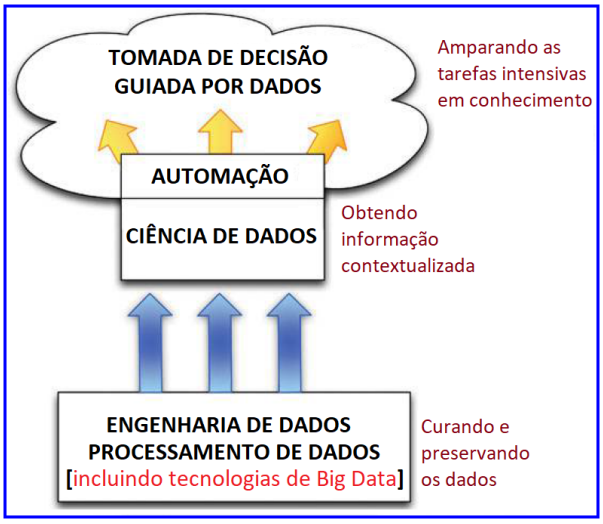

Figura 6. Evolução Dados → Informação → Conhecimento na utilização do *Big Data* como base no Processo de Tomada de Decisão (adaptado de Provost e Fawcett, 2013) [tradução dos autores]

Em face disso, tecnologicamente, a Tomada de Decisão Guiada por Dados auxilia os gestores em suas atividades cognitivas. E, por conseguinte, contribui à qualidade das decisões deliberadas e à produtividade laboral. Neste enredo, as soluções computacionais de Ciência de Dados auxiliam os gestores em suas Tarefas Intensivas em Conhecimento (Schreiber *et al*., 2000). Dentre as referidas tarefas, são listadas:

- **Associação**. Tarefa em que o conhecimento de um mapeamento entre dois conjuntos de objetos é utilizado. Por exemplo, considere as tarefas em que a relação entre causa e efeito é evidente. Por exemplo, pode-se citar a falta de óleo em um motor leva ao desgaste prematuro das peças mecânicas, comprometendo a vida útil). Outro exemplo do cotidiano de tarefa de Associação é sugestão de um vinho adequado em uma refeição. Dadas as características do prato principal, o *sommelier* recomenda o vinho ideal a ser consumido. Neste caso, em um ecossistema de *Big Data*, as regras de associação entre vinhos e pratos principais podem ser abstraídas de bases de dados das vinícolas e das opiniões de experiências dos consumidores, mediante os processos de Mineração de Dados.
- **Avaliação**. O objetivo principal em tarefas de avaliação é caracterizar um caso em conformidade às opções de decisão. Para exemplificar uma Tarefa Intensiva em Conhecimento, considere um sistema de avaliação de crédito. Baseando-se nos históricos de empréstimos realizados e armazenados no ecossistema de *Big Data*, para determinado caso, um gestor pode melhor discernir sobre sua decisão ao compará-lo com casos anteriores. Neste tocante, a decisão a ser tomada pode ser: conceder o crédito; recusar o crédito; sugerir uma contraproposta, ou requisitar mais informações do contratante.
- **Diagnóstico**. Dado um conjunto de atributos, resume-se na tarefa de inferir o estado de um objeto (caso em investigação) em contraste ao conhecimento que rege um domínio. Neste sentido, dispõe-se de uma heurística (geralmente expressa por regras) que caracteriza a compatibilidade ou a discrepância de um caso em análise em relação ao comportamento esperado. Em um ecossistema de *Big Data*, o conhecimento do domínio (as regras) pode ser abstraído dos dados, mediante o uso de soluções computacionais de Ciência de Dados.
- **Monitoramento**. Em suma, é um processo de diagnóstico iterativo, no qual o estado de um objeto é aferido ciclicamente ao longo do tempo. Ou seja, periodicamente, dados sensitivos são capturados e criticados por regras que definem a dinâmica da normalidade do objeto em monitoramento. Nas situações em que anormalidades são detectadas, alertas podem ser disparados para a execução de atividades de correção. Assim como na Tarefa Intensiva em Conhecimento de Diagnóstico, em um ecossistema de *Big Data*, as regras podem ser abstraídas a partir dos dados com o uso de soluções computacionais de Ciência de Dados.
- **Predição**. Considerando um conjunto de dados de históricos e os dados correntes, na Tarefa Intensiva em Conhecimento de Predição, estima-se o(s) evento(s) vindouro(s) para algum ponto futuro no tempo. Exemplos de aplicações voltadas à Predição são a estimativa de vendas, a previsão de safras de *commodities*, dentre outros.

Em suma, em ecossistemas de *Big Data*, quando as Tarefas Intensivas em Conhecimento são tecnologicamente suportadas, estas permeiam os processos de transformação dos dados primários em informação, apoiando o discernimento dos tomadores de decisão.

5 *Big Data* e Ciência de Dados: sua Complementariedade na Tomada de Decisão

O objetivo deste artigo é evidenciar a diferença e a complementariedade dos conceitos *Big Data* e Ciência de Dados no Processo de Tomada de Decisão. Neste sentido, a Figura 7 (no apêndice) ilustra o alinhamento conceitual a ser pontuado.

O *Big Data* se caracteriza principalmente nos volume, variedade, velocidade, veracidade, variabilidade e valor de imensas bases de dados, requerendo estruturas computacionais escaláveis para tratamento dos recursos armazenados (Grady e Chang, 2015). Neste contexto, o *Big Data* atua como a primeira camada de suporte (camada basilar dos materiais) de ambientes computacionais voltados à tomada de decisão.

Em face disso, a infraestrutura de *Big Data* deve suportar o gerenciamento, a proveniência, a curadoria e o arquivamento dos dados e seus metadados (Mishra e Chang, 2015). Nessa dinâmica, interdisciplinarmente, a Ciência da Informação contribui no fomento das competências de organização e representação de dados e informação, privilegiando os serviços de coleta, registro, filtragem, classificação e entrega de dados e seus metadados às atividades reservadas à camada da Ciência de Dados.

Em relação à Ciência de Dados, esta é considerada a segunda camada de suporte (camada dos métodos) em ambientes computacionais voltados à tomada de decisão. Caracterizada como uma camada de transformação dados → informação, a Ciência de Dados visa agregar valor aos dados armazenados na camada de *Big Data*. Para tanto, as organizações que queiram extrair informações a partir do *Big Data* necessitam combinar habilidades diversas, geralmente, atendidas por equipes multidisciplinares (Gartner, 2018b). Neste sentido, Manyika *et al*. (2011) identifica três oportunidades de atuação:

- **Suporte tecnológico**. Oportunidade reservada aos profissionais com competência em computação que desenvolvem, configuram e mantém, por exemplo: os programas para a aquisição de dados a partir do *Big Data*; as interfaces para realização de análise de dados; a implementação de algoritmos de Aprendizado de Máquina; dentre outros.
- **Análise de dados**. Envolve os profissionais com habilidades técnicas em Estatística e Aprendizado de Máquina para explorar os grandes volumes de dados na obtenção de *insights* de negócios nas Tarefas Intensivas em Conhecimento.
- **Tomada de Decisão**. Oportunidade ligada aos gestores com conhecimento do domínio e que tenham as habilidades para formular questões pertinentes a serem investigadas. Mediante as Tarefas Intensivas em Conhecimento, tais atores realizam análises, interpretações e resolvem problemas, apoiando-se em Processos de Tomada de Decisão guiada por Dados.

Ressalta-se que as oportunidades anteriormente relacionadas ensejam habilidades multidisciplinares de novos profissionais, os cientistas de dados. Em poucas palavras, um cientista de dados lida com conhecimento sobre tecnologias, formas de comunicação, habilidades analíticas e domínio aplicados no ciclo evolutivo dados → informação → conhecimento. Na perspectiva deste artigo, interdisciplinarmente, as competências dos profissionais das Ciências da Computação e da Informação (engenheiros e programadores de software, analistas de banco de dados, curadores, bibliotecários, arquivistas, entre outros) e dos gestores são necessárias na produção e na utilização do conhecimento advindo a partir do *Big Data*. Com o uso inovador de métodos e tecnologias, os cientistas da computação e da informação devem municiar os gestores com ferramental propício à resolução de problemas nos ambientes coorporativos (Swan e Sheridan, 2008). Ou seja, com base em suas habilidades, os cientistas da computação e da informação desenvolvem as interfaces dos ecossistemas de *Big Data* que auxiliam os gestores no Processo de Tomada de Decisão guiada por Dados. Neste sentido, tais profissionais cooperam no(a):

- obtenção de dados de fontes primárias heterogêneas internas à organização ou distribuídas na web;
- definição dos procedimentos de ingestão de dados no ecossistema de *Big Data*;
- pré-processamento, estruturação e formalização dos dados e seus metadados para o uso;
- modelagem dos processos de transformação de dados e seus metadados de modo a gerar informações relevantes;
- utilização de métodos computacionais ou estatísticos de Aprendizado de Máquina para automatizar os processos de sumarização e visualização de informações a partir dos dados disponíveis; e
- instrumentalização dos meios de exploração da informação com intuito de subsidiar os entendimentos dos tomadores de decisão no desempenho das Atividades Intensivas em Conhecimento.

6 Considerações Finais

Com o advento da Internet, tem-se produzido imensas bases de dados para os mais variados domínios. Este fato é acelerado em função do uso massivo e da evolução das Tecnologias de Informação e Comunicação. Este é o ensejo do *Big Data* como conceito contemporâneo para processamento de dados complexos e dinâmicos perante às exigentes demandas de informação da atual Sociedade do Conhecimento.

Notadamente, o desenvolvimento de instrumentos voltados à extração automatizada de informação a partir do *Big Data* têm despertado atenção das organizações. Principalmente, para subsidiar os gestores na execução das Tarefas Intensivas em Conhecimento, facilitando o Processo de Tomada de Decisão guiada por Dados.

Neste contexto, a evolução dados → informação → conhecimento em ecossistemas de *Big Data* requer a sinergia de competências de profissionais (cientistas da informação, cientistas da computação, estatísticos, gestores, dentre outros).

Considerando a interdisciplinaridade supracitada, neste trabalho dissertou-se sobre a diferença tecnológica e a complementariedade dos conceitos *Big Data* e Ciência de Dados.

Como resultado, aponta-se que o *Big Data* é um termo intrinsecamente ligado à infraestrutura do hardware e de serviços de computação na nuvem, necessários para o armazenamento, o processamento e a distribuição de recursos. Em outras palavras, considerando a evolução dados → informação → conhecimento, o conceito *Big Data* é relacionado à camada basilar de materiais, privilegiando os 6Vs atribuídos aos dados (Velocidade, Variedade, Variabilidade, Veracidade, Volume e Valor). Neste sentido, advoga-se que a Ciência da Informação tem papel fundamental na consolidação dos ecossistemas de *Big Data*. Principalmente, no tocante às competências de organização/representação de dados e metadados e da Curadoria Digital dos recursos mantidos nesses ecossistemas.

Em relação à Ciência de Dados, entende-se que este conceito é conexo à camada dos softwares, a qual metodologicamente transforma os dados em informação para o Apoio à Tomada de Decisão. Neste sentido, as competências dos cientistas da computação e da informação são necessárias na concepção de modelos de representação, interfaces de comunicação e informações relevantes. Em ecossistemas de *Big Data*, tais competências são úteis para customizar o ferramental utilizado pelo gestor na Tomada de Decisão guiada por Dados.

Notas

Os autores agradecem à Fundação Araucária pelas bolsas de Iniciação Científica concedida (PIBIC-2018/UNICENTO - Programa Institucional de Iniciação Científica) e de Produtividade (FA - Convênio 046/2019).

Referências

Abraham, Ajith; Guo, He; Liu, Hongbo (2006). Swarm Intelligence: Foundations, Perspectives and Applications. // Nedjah, Nadia, Mourelle, Luiza de M. (eds). Swarm Intelligent Systems. Heidelberg: Springer, 2006. 3-25.

Akhtat, Syed Muhammad Fahad (2018). Big Data Architect's Handbook. Birmingham: Pack Publishing, 2018.

Blum, Avrim; Hopcroft, John; Kannan, Ravi (2018). Foundations of Data Science (2018). https://www.cs.cornell.edu/jeh/book.pdf (2018-07-26).

Bugnion, Pascal; Manivannan, Arun; Nicolas, Patrick R. (2017). Scala: Guide for Data Science Professionals. Birmingham: Packt Publishing, 2017.

Conamay, Drew (2010). The data science venn diagram (2010). http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram (2018-07-27).

Digital Curation Centre (2018). The DCC Curation Lifecycle Model (2018). http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf (2018-07-25).

Economist, The (2017). The world's most valuable resource is no longer oil, but data (2017). https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (2018-07-28).

Eiica (2019). X Encontro Internacional de Informação, Conhecimento e Ação. Marília, 2018. http://enancib.marilia.unesp.br/index.php/EIICA/XEIICA. (2019-02-27).

Gartner (2018a). What is Big Data? – Gartner IT Glossary – Big Data (2018a). http://www.gartner.com/it-glossary/big-data (2018-07-28).

Gartner (2018b). Data Scientist – Gartner IT Glossary (2018c). https://www.gartner.com/it-glossary/data-scientist (2018-07-28).

Grady, Nancy; Chang, Wo (2015). NIST Big Data Interoperability Framework: Volume 1, Definitions (2015). https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf (2018-07-28).

Laney, Doug (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety (2001). https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (2018-07-25).

Manyika, James; Chui, Michael; Brown, Brad; Bughin, Jacques; Dobbs, Richard; Roxburgh, Charles Byers, Angela Hung (2011). Big data: The next frontier for innovation, competition, and productivity (2011). https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf (2018-07-28).

Mishra, Sanjay; Chang, Wo (2015). NIST Big Data Interoperability Framework: Volume 5, Security and Privacy (2015). https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-5.pdf (2018-07-28).

Munakata, Toshinori (2008). Fundamentals of the New Artificial Intelligence: Neural, Evolutionary, Fuzzy and More. Heidelberg: Springer, 2008.

National Science Foundation (2005). Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century (2005). https://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf (2018-07-28).

Provost, Foster; Fawcett, Tom (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. // Big Data, 1:1 (March 2013) 51-59.

Rautenberg, Sandro; Carmo, Paulo Ricardo Viviurka do (2018). Big Data e Ciência de Dados: Complementariedade Conceitual no Processo de Tomada de Decisão. // Encontro Internacional de Informação, Conhecimento e Ação, Marilia. Caderno de Resumos. Marilia: Unesp, 10, 1, 2018, p. 219-221.

Roy, Arnab; Underwood, Mark; Chang, Wo (2015). NIST Big Data Interoperability Framework: Volume 4, Security and Privacy (2015) https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-4.pdf (2018-07-28).

Schreiber, Guss; Akkermans, Hans; Anjewierden, Anjo; de Hoog, Robert; Shadbolt, Nigel; van der Welde, Walter; Wielinga, Bob (2000). Knowledge Engineering and Management: the CommonKADS Methodology. Cambridge: The MIT Press, 2000.

Swan, Alma; Brown, Sheridan (2008). The Skills, Role and Career Structure of Data Scientists and Curators: an Assessment of Current Practice and Future Needs (2008). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8960&rep=rep1&type=pdf (2008-07-28).

Tarasewich, Peter; Mcmullen, Patrick R (2002). Swarm Intelligence: power in numbers. // Communications of the ACM 45:8 (August 2002) 62-66.

van der Aalst, Wil (2014). Data Scientist: The Engineer of the Future. // Mertins, Kai; Bénaben, Frédérick; Poler, Raul; Bourrières Jean-Paul (eds.) (2014). Proceedings of the Interoperability of Enterprises Systems and Applications Conference (I-ESA'2014): Albi, France. Mar. 24-28, 2014. Heidelberg: Springer.

Received: 2018-10-15. Accepted: 2019-03-22

Apêndice

Figura 1.

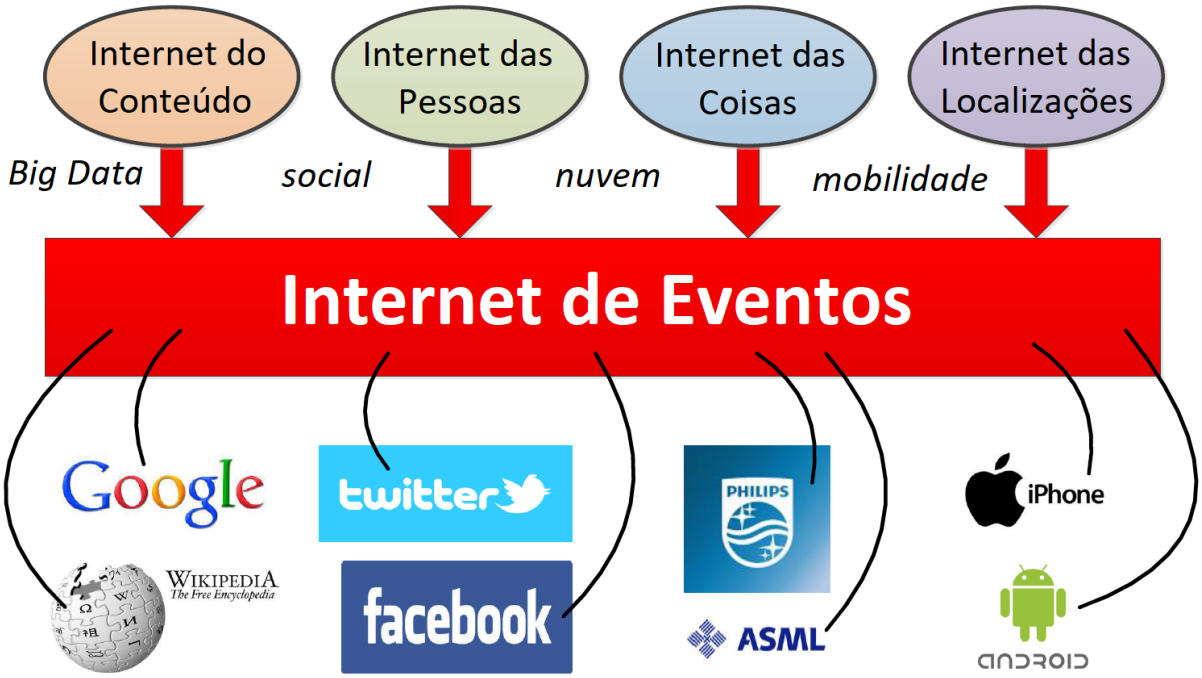

Figura 1. A Internet de Eventos baseada em Conteúdo, Pessoas, Coisas e Localizações (van der Aalst, 2014) [tradução dos autores]

Figura 3.

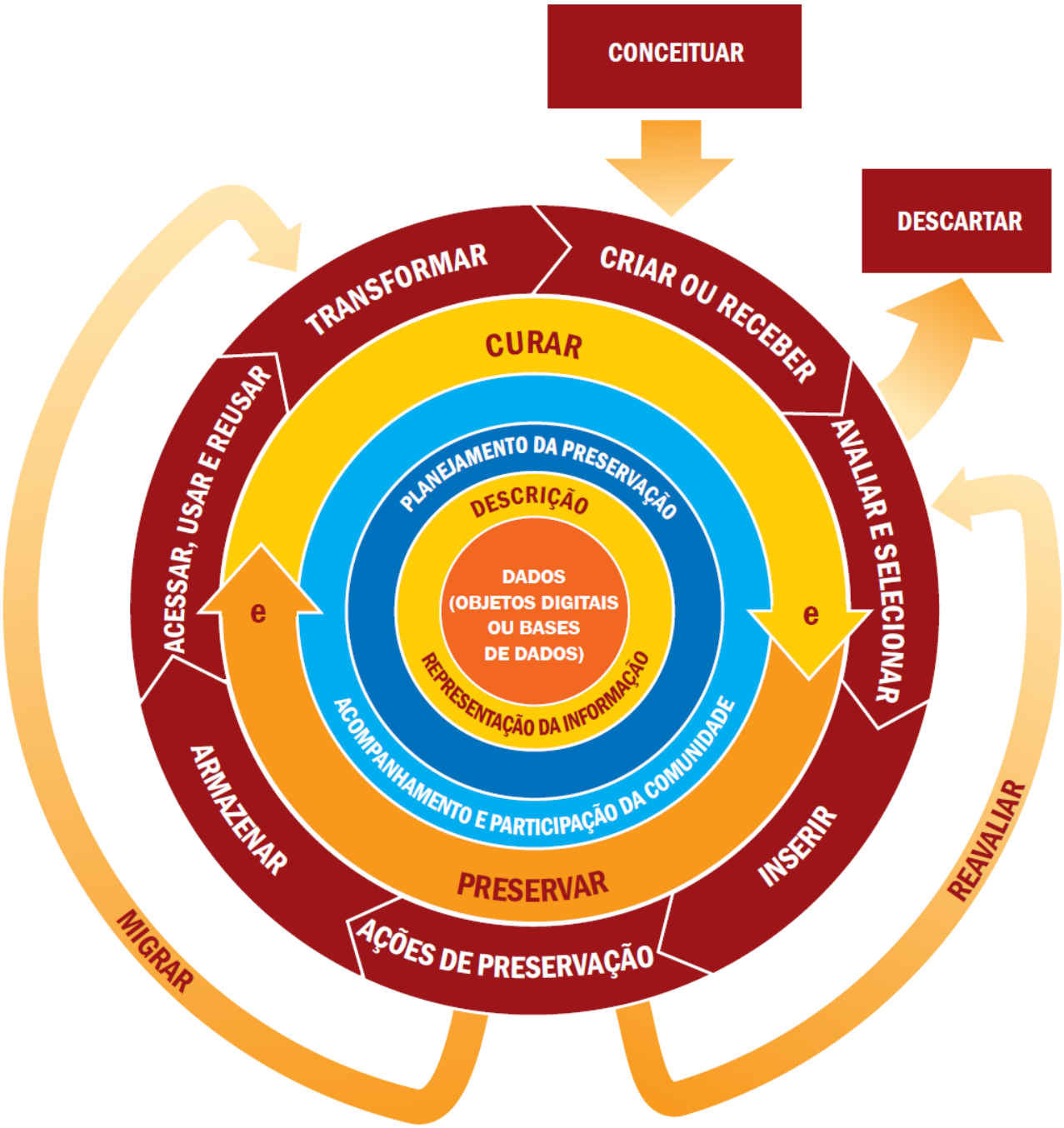

Figura 3. Ciclo de Vida da Curadoria Digital (Digital Curation Centre, 2018) [tradução dos autores]

Figura 7.

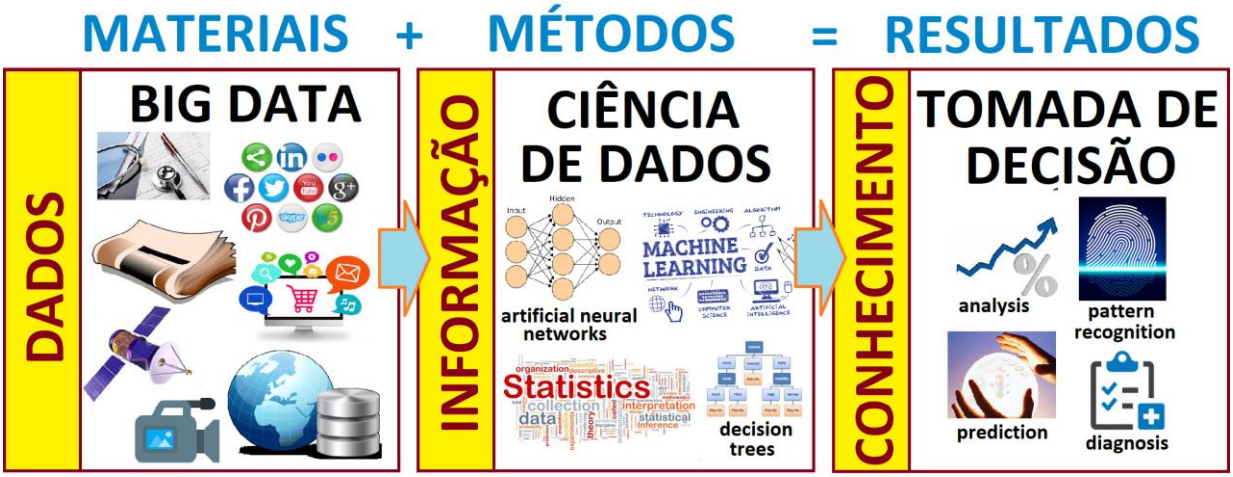

Figura 7. Representação do Alinhamento de *Big Data* e Ciência de Dados no Processo de Tomada de Decisão