repec

Revista de Educação e Pesquisa em Contabilidade
Journal of Education and Research in Accounting

Periódico Trimestral, digital e gratuito publicado pela Academia Brasileira de Ciências Contábeis

ABRACICON
ACADEMIA BRASILEIRA DE CIÊNCIAS CONTÁBEIS

ISSN 1981-8610

REPeC, Brasília, v. 11, n. 4, art. 2, p. 396-407, out./dez. 2017
Disponível online em www.repec.org.br
DOI: http://dx.doi.org/10.17524/repec.v11i4.1634

Big Data: Reflexões epistemológicas e impactos nos estudos de finanças e mercado de capitais

Resumo

Objetivo e método: O acesso a séries de dados tem um papel central na área de Finanças. A crescente disponibilidade de grandes volumes de dados, em diferentes formatos e em alta frequência, combinada aos avanços tecnológicos nas ferramentas de armazenamento e proces-samento desses dados, têm criado um novo cenário nas pesquisas acadêmicas em geral, e em Finanças em particular, gerando novas oportunidade e desafios. Entre esses desafios emergem questões metodológicas, vastamente discutidas por pesquisadores de diferentes áreas, mas também questões epistemológicas que merecem maior espaço de discussão. Assim, o objetivo deste ensaio teórico é analisar o aspecto conceitual e epistemológico da utilização de dados intensivos e seus reflexos para a área de Finanças.

Resultados e contribuições: Consideramos que o método hipotético-dedutivo de pesquisas empíricas, que é o mais recorrente, limita a construção do conhecimento na dita 'era Big da-ta', uma vez que tal abordagem parte de uma teoria estabelecida e restringe as pesquisas ao teste à(s) hipótese(s) proposta(s). Defendemos aqui a apropriação de uma abordagem abduti-va, como defendida em Haig (2005), que tem convergência com as ideias da grounded theory e que parece ser a abordagem mais adequada para esse novo contexto, por possibilitar a am-pliação da capacidade de se obter informações de valor dos dados.

Palavras Chave: Big Data, Método Abdutivo, Epistemologia, Finanças.

Talieh Shaikhzadeh Vahdat Ferreira
Mestre em Administração de Empresas pela Universidade Federal da Paraíba (UFPB) e Doutoranda em Administração pela Universidade Federal da Paraíba (UFPB). **Contato:** UFPB, Cidade Universitária, s/n, CCSA/PPGA, Castelo Branco, João Pessoa (PB). CEP.: 58051-900.
E-mail: taliehv@gmail.com

Francisco José da Costa
Doutor em Administração de Empresas pela Fundação Getúlio Vargas (FGV/SP) e Professor do Departamento de Administração na Universidade Federal da Paraíba (UFPB). **Contato:** Cidade Universitária, s/n, CCSA/PPGA, Castelo Branco, João Pessoa (PB). CEP.: 58051-900.
E-mail: franze.mq@gmail.com

Editado em Português e Inglês. Versão original em Português.

Recebido em 24/6/2017. Pedido de Revisão em 19/8/2017. Resubmetido em 9/9/2017. Aceito em 10/9/2017 por Dr. Gerlando Augusto Sampaio Franco de Lima (Editor associado) e por Dr. Orleans Silva Martins (Editor). Publicado em 24/10/2017. Organização responsável pelo periódico: Abracicon.

1. Considerações Iniciais

Os avanços da internet e a ampliação do uso de tecnologias de comunicação móveis trouxeram como consequência um crescimento substancial na quantidade e na armazenagem de dados. Tomando por referência o ano de 2016, de acordo com a *International Business Machines* (IBM), 90% de todos os dados produzidos no mundo foram criados nos dois anos anteriores. Concomitantemente, a mesma empresa projetou que este volume dobraria a cada dois anos adicionais (IBM, 2016). Foi este crescente volume de dados combinado aos avanços tecnológicos em ferramentas de processamento e armazenagem que criaram as bases do que vem sendo denominado de 'Big Data' entre acadêmicos, empresários e governos (Ekbia *et al.*, 2015).

O advento do Big Data estabeleceu uma nova fronteira para o desenvolvimento de pesquisas (Chen & Zhang, 2014), trazendo consigo oportunidades e desafios. As oportunidades surgem à medida que cresce a acessibilidade a dados de forma massiva, o que impulsiona o aprimoramento contínuo de tecnologias e usos nas mais diferentes áreas da vida humana (Demchenko, Grosso, De Laat & Membrey, 2013). Entre essas áreas beneficiadas pelo Big Data, pode-se citar a de Finanças, área em que os dados ocupam um papel central em seu desenvolvimento acadêmico e profissional (Seth & Chaudhary, 2015).

Por outro lado, a 'era do Big Data', por suas particularidades (sobre as quais falaremos mais adiante), traz novos dilemas associados, em primeira instância, a questões tecnológicas e estatísticas e, em um segundo momento, à dimensão epistemológica da pesquisa e da produção de conhecimento (Ekbia *et al.*, 2015). Este ensaio teórico tem como objetivo analisar este último aspecto (epistemológico) e seus reflexos na área de Finanças, discutindo possíveis desdobramentos sobre o mercado de capitais.

Para isso, o trabalho é dividido em cinco seções, incluindo esta introdução. Na segunda seção, buscamos caracterizar o Big Data e apontar as mudanças na construção do conhecimento e nas escolhas de método de pesquisa devido a questões epistemológicas que emergem. Na terceira seção, são explorados alguns dos desafios metodológicos e as alternativas para o processo de produção do conhecimento. Na quarta seção, focamos em algumas questões que a utilização da abordagem de Big Data traz para Finanças. Por fim, algumas considerações são expostas e são apontados outros desafios que entendemos mais relevantes para o novo cenário acadêmico e profissional.

2. O Big Data: características e desafios da construção de conhecimento

Neste item apresentamos e discutimos de forma breve os elementos conceituais e as implicações do Big Data. Inicialmente, expomos o seu conceito, seguindo os caminhos definidos na literatura especializada, que não tem exatamente definido 'Big Data', mas apontado características associadas aos dados captados e armazenados. Na segunda parte, e considerando essa primeira construção, apresentamos a discussão sobre as consequências que uma abordagem de Big Data traz ao processo de produção de conhecimento acadêmico e profissional.

2.1. Caracterização do Big Data

Com as oportunidades de pesquisa e aprimoramento decisório que surgem com os enormes volumes de dados, as possibilidades de uso desses dados têm sido debatidas amplamente tanto dentro como fora da academia, o que contribui para constantes avanços em sua compreensão e utilização (Ekbia *et al.*, 2015). Contudo, a própria definição de Big Data ainda não apresenta um consenso e vem evoluindo e amadurecendo ao longo do tempo. Inicialmente, o termo era caracterizado por sua associação aos chamados "3 V's": volume, velocidade e variedade. Os principais caracterizadores eram as empresas de tecnologia, como IBM e Oracle. Em estudos mais recentes, essa forma de definição foi refinada, adicionando-se dois atributos - valor e veracidade - para caracterizar o Big Data, formando os "5 V's" (Demchenko *et al.*, 2013; Lau, Zhao, Chen, & Guo, 2016).

A primeira característica crucial do Big Data está relacionada, portanto, ao **volume** de dados. Inicialmente, essa discussão pode se limitar a quantificar *terabytes, pentabytes* ou *zettabytes* (Kitchin, 2014), porém é esperado que a capacidade de geração de dados siga se ampliando, o que pode limitar a utilidade desta definição inicial. Lau *et al.* (2016) propõem uma discussão mais ampla para este quesito ao afirmar que o volume de um Big Data deve atingir uma extensão tal que as tecnologias atuais encontrem dificuldade em sua armazenagem, recuperação, análises e utilização.

A **velocidade** está relacionada não só à capacidade de produção dos dados, mas também de processamento e análise dos sistemas informatizados envolvidos. Em verdade, a produção de dados massivos, como um censo nacional, por exemplo, não é uma novidade, contudo, os custos de seu processamento e o tempo necessário para coleta de dados e sua elaboração analítica têm sido elevados (Miller, 2010). Desta forma, o Big Data é caracterizado pela contínua produção de dados, por vezes associados a múltiplos eventos, em nível detalhado, com grande flexibilidade para análise de seu escopo, por meio de ferramentas com crescente agilidade de processamento em tempo real (Kitchin, 2014).

Considerando a possibilidade de versatilidade que as bases de Big Data podem apresentar, a terceira caraterística está associada à **variedade** de formatos e fontes de dados. Com o desenvolvimento da *Web 2.0* e *Web 3.0*, os dados podem ser captados não só em seu formato convencional, como tabelas ou planilhas, mas também em formatos semiestruturados ou não estruturados, ou de forma mista, como imagens, sons, entre outras possibilidades (Demchenko *et al.,* 2013). Essa heterogeneidade de formatos demanda uma nova geração de técnicas e métodos para processamento e armazenamento que estão constantemente sendo aprimoradas.

O **valor** é uma característica-chave quando se está lidando com dados em geral, já que todo o esforço de processamento e investimento somente são justificáveis quando esses dados acrescentam valor à atividade ou à análise em questão (Demchenko *et al.,* 2013). Assim, na medida em que as ferramentas de processamento conseguem transformar os dados coletados de forma a produzir conhecimento, os benefícios obtidos com este tipo de tecnologia poderão atingir níveis mais altos de valor (Lau *et al.*, 2016).

E, por fim, a **veracidade** está relacionada à qualidade e à validade dos dados. Isso implica tanto a consistência que os dados devem possuir, sendo confiáveis em termos de sua mensuração e validade, quanto a qualidade dos dados em si em termos de integridade, que depende de toda uma cadeia, desde sua coleta até os métodos de processamento e armazenagem (Demchenko *et al.,* 2013). Nesse contexto, de acordo com Lau *et al.* (2016), novos desafios tecnológicos emergem na busca por manter a qualidade e consistência dos dados em bases enormes e que estão sendo atualizadas em tempo real. Em conexão próxima com esse desafio da qualidade e da consistência de dados, emerge a questão da produção de conhecimento com base em Big Data, assunto que discutiremos no item seguinte.

2.2. Construção do conhecimento baseado em Big Data

O crescimento da aplicação de Big Data em pesquisas tem alterado as formas convencionais de construção de conhecimento (Demchenko *et al.,* 2013; Chen & Zhang, 2014; Kitchin, 2014). Para esse ensaio, tomamos por referência a visão de Chen e Zhang (2014), que afirmam que, historicamente, a construção do conhecimento científico esteve baseada em três grandes paradigmas: a Ciência Empírica, a Ciência Teórica e a Ciência Computacional. Centenas de anos atrás, a ciência era construída com base em experimentos empíricos que visavam testar possíveis intuições e comprovar sua veracidade, formando o primeiro paradigma. À medida que essas evidências foram se ampliando, foi possível elaborar teorias, chegando ao segundo paradigma, com a construção teórica do conhecimento.

No entanto, a complexidade dos fenômenos em análise foi se ampliando e os pesquisadores precisaram utilizar novas ferramentas de simulação científica para validar seus resultados. Assim, emergiu o conceito do 'Terceiro Paradigma', que são as ciências computacionais as quais possibilitaram simulações científicas em larga escala, criando resultados 'suficientemente robustos' (Hey, Tansley & Tolle, 2009). Ocorre que a natureza do 'Big Data' para a elaboração de simulações e testes demanda técnicas e tecnologias totalmente distintas dos outros três paradigmas. Isso derivou a ideia do 'Quarto Paradigma', que seria a ciência de dados intensivos (Miller, 2010; Hey, Tansley & Tolle, 2009). Nesse contexto, cada vez mais pesquisadores e agentes de mercado necessitam combinar os avanços da tecnologia, com o processamento eficiente de grandes volumes de dados e a utilização de métodos científicos mais convencionais para se beneficiar da coleta de dados desejáveis (Demchenko *et al.,* 2013), isso porque, nesse cenário, há abundância de dados disponíveis. Mas para que esses agreguem valor, devem ser processados, transformados em bases para viabilizar a identificação de padrões e, finalmente, serem interpretados de forma a ampliar o conhecimento sobre o que se está buscando compreender (Seth & Chaudhary, 2015). Em outra perspectiva, Haig (2005) aponta uma organização da evolução da pesquisa científica em dois métodos ou paradigmas preponderantes, que são o indutivo e o hipotético-dedutivo. De acordo com Lakatos e Musgrave (1979), a distinção entre o conhecimento e a especulação se baseia na premissa de que o primeiro foi provado pela força do conhecimento ou dos sentidos, e o outro não. Para os indutivistas, essa prova parte de uma observação que fornece a base segura sobre a qual o conhecimento pode ser construído, o que faz com que proposições possam ser feitas por indução, do particular para o todo. Por outro lado, o método hipotético-dedutivo parte de uma teoria estabelecida, gera hipótese e, em seguida, busca testá-las para ampliar seus conhecimentos ou identificar possíveis inconsistências na teoria. Em geral, a perspectiva indutivista está mais associada a procedimentos empíricos qualitativos; e a hipotético-dedutiva está mais associada a métodos quantitativos, principalmente métodos estatísticos que envolvem técnicas inferenciais de estimação e testes de hipóteses.

No entanto, a utilização de grandes volumes de dados cria uma nova complexidade que não parece ser abarcada totalmente por esses métodos de análise. Essa afirmação se baseia na possibilidade de a análise ser feita a partir dos próprios dados, ou seja, não partindo necessariamente de uma hipótese ou teoria específica. Considerando a operacionalização de grandes volumes de dados, necessariamente serão necessárias técnicas quantitativas para procedimentos que são característicos de pesquisas qualitativas. Tem-se então uma aproximação da visão indutiva de procedimentos quantitativos tipicamente aplicados na visão hipotético-dedutiva.

Dessa forma, entendemos que o Big Data traz uma efetiva mudança com relação às abordagens tradicionais, o que faz requerer uma abordagem epistemológica distinta no desenvolvimento da ciência em comparação com o método hipotético-dedutivo corrente (Kitchin, 2014). Em nossa visão, o melhor alinhamento vem do resgate de um debate que já vem sendo feito há algumas décadas em torno da *grounded theory* e do dito 'método abdutivo'. No item seguinte, esta discussão será detalhada.

3. Mudanças epistemológicas e metodológicas da era do Big Data

Nesse item discutimos as consequências metodológicas e epistemológicas de uma realidade de pesquisa baseada em abundância de dados. Inicialmente, descrevemos a perspectiva abdutiva como referência adequada de produção de conhecimento nesse cenário. Em seguida, apresentamos os desafios metodológicos e de ferramentas mais recorrentemente citados na literatura especializada.

3.1. Método abdutivo e o Big Data

Dado o advento do Big Data, parece ser o momento apropriado para se repensarem as questões epistemológicas e os métodos de pesquisa em unidades maiores que possibilitem a ampliação do conhecimento por novas perspectivas a partir dos dados e, não, por uma hipótese específica. Isso se justifica pelo fato de o método indutivo e do hipotético-dedutivo – apesar de apresentarem lógicas distintas – terem uma característica fundamental em comum: focam em uma parte limitada dos dados para realizar suas análises (Haig, 2005), de modo que podem deixar dados importantes do fenômeno ou objeto de pesquisa fora da análise de resultados.

É importante destacar que a tentativa bem estruturada de construir o conhecimento a partir dos dados não é um conceito novo na academia. Em 1967, Barney Glaser e Anselm Strauss iniciaram as discussões da *grounded theory*, que é considerada uma das formas mais puras de se fazer pesquisa qualitativa (Glaser & Strauss, 1967, Bianchi & Ikeda, 2008). Nessa metodologia, o pesquisador não busca fazer verificações nem parte de uma teoria específica para a elaboração de sua pesquisa. Pelo contrário, o pesquisador é convidado a se libertar de seus pressupostos e 'amarras' teóricas já estabelecidas para buscar entender os dados de forma que as perguntas e o conhecimento sejam elaborados a partir deles (Glaser & Strauss, 1967).

A principal finalidade da *grounded theory* é a construção de teorias e, para isso, os pesquisadores assumem dois pressupostos centrais: a) pesquisador, realidade e teoria são entidades contínuas e intrínsecas e assim podem interagir continuamente; b) a teoria evolui ao longo do processo de pesquisa, sendo os resultados de diversas interpolações de dados e análises (Bianchi & Ikeda, 2008). Dessa forma, a ideia é garantir que não só o pesquisador investigue a realidade a partir dos dados, mas que esse é um processo contínuo e que pode se alterar a qualquer momento, já que os dados estão continuamente sendo reacessados e analisados.

Seguindo os mesmos fundamentos que a *grounded theory* traz como método qualitativo, mais recentemente Haig (2005) propôs para a pesquisa quantitativa o método (ou paradigma, como o Haig chamou) dito abdutivo. Esse método segue a mesma lógica epistemológica da *grounded theory*, ou seja, busca o conhecimento a partir dos dados, mas seguindo um protocolo de execução metodológica diferente. O método abdutivo pode ser considerado uma combinação entre os métodos indutivo e hipotético-dedutivo no contexto da pesquisa quantitativa, já que possibilita ao pesquisador descobrir os fatos empíricos (por indução) para construir teorias que explicam esses fatos (por dedução) (Haig, 2005). A Figura 1 ilustra a dinâmica de pesquisa do método abdutivo:

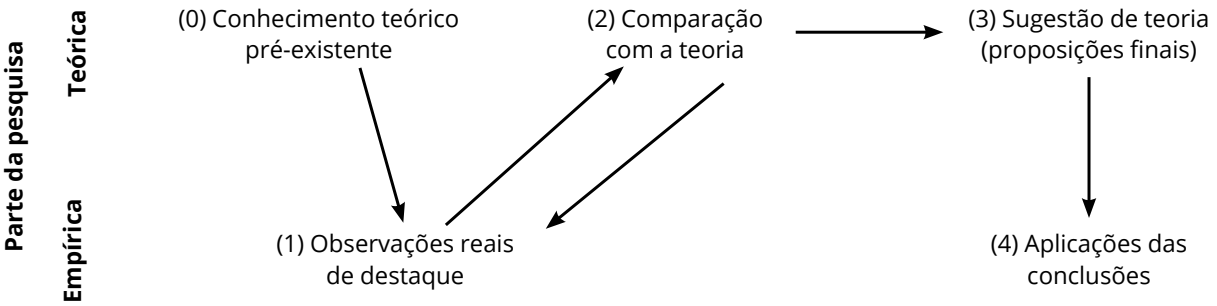

Figura 1. A visão da pesquisa pelo processo abdutivo.

Fonte: Spens e Kovács (2006).

Nessa visão, os fenômenos existem para serem explicados, de modo que não são entendidos simplesmente como objetos de predição ou confirmação de teorias e hipóteses. Em termos práticos, isso significa que, uma vez identificado um fenômeno a partir da exploração quantitativa e exaustiva dos dados, usam-se inferências abdutivas, ou seja, busca-se a melhor explicação para a sua ocorrência, que é entendida como a provável verdadeira explicação. Dessa forma, podem-se utilizar diferentes métodos ou combinações deles para selecionar a 'melhor explicação' que, quando identificada, poderá servir como base para a elaboração de uma teoria, caso não se encaixe em nenhuma existente (Haig, 2005).

Assim, o método abdutivo pode ser uma alternativa interessante enquanto método de estudo que viabiliza a construção do conhecimento a partir de grandes volumes de dados. Com a utilização da abordagem abdutiva, a exploração exaustiva de grandes bases de dados pode impulsionar a geração de conhecimento que até então não estava ao alcance de metodologias baseadas na verificação. Como citado anteriormente, metodologias de verificação têm a limitação que a própria hipótese de pesquisa impõe, de modo que, potencialmente, restringem-se a verificar os dados necessários para a análise da hipótese e não se beneficiam de outras informações que poderiam ser obtidas dessa mesma base de dados para viabilizar o entendimento de temas mais complexos da ciência contemporânea (Miller, 2010). Adicionalmente, o método abdutivo possibilita que em uma mesma amostra possam ser identificados mais de um fenômeno e que teorias possam ser construídas de forma a melhor explicar a realidade, utilizando analogias para buscar respostas a temáticas que, por vezes, são subjetivas ou não se possuem dados censitários (Ekbia *et al.* 2015).

A utilização dessa abordagem para a análise de Big Data também soluciona uma das armadilhas que podem surgir quando se está lidando com a 'ciência de dados intensivos', que significa dar importância exagerada aos dados e suas relações causais (exemplos: *data brokers* e provedores de análise de dados). Essa armadilha pode levar à falsa impressão de que a simples identificação da relação causal seria suficiente para a produção do conhecimento, o que é, em geral, uma simplificação exagerada do fenômeno em análise (Kitchin, 2014). É importante manter em mente que a construção do conhecimento é guiada por inquietações ou preocupações, dando forma ao 'problema de pesquisa', que permitem teorizar sobre o porquê de essas relações acontecerem e quais as suas implicações teóricas e práticas.

Uma questão conexa dessa discussão e que também levanta reflexões relevantes para pesquisadores acadêmicos e profissionais concerne às decisões metodológicas de pesquisa, sobre as quais discorremos no item seguinte.

3.2. Desafios metodológicos para utilização de Big Data

Abordagens de pesquisa baseadas em grandes volumes de dados trazem desafios metodológicos de duas naturezas: a primeira, relativa à capacidade tecnológica de acumulação e processamento dos dados; e a segunda, relativa à análise estatística desses dados.

Relativamente ao primeiro desafio (processamento), novas técnicas de processamento têm sido desenvolvidas, tais como *data mining*, inteligência artificial, *statistical learning, machine learning, predictive analytics.* Essas técnicas permitem, inclusive, detectar padrões sem que uma pergunta específica precise ser fornecida ao sistema, o que possibilita a identificação de informações que talvez não fossem facilmente observáveis. As técnicas referidas têm ainda a capacidade gerada por programações computacionais que permitem 'aprender' em tempo real com os dados na medida em que estes são coletados. A partir desses 'novos' aprendizados, são construídos então modelos de predição que podem se reformular constantemente, de forma a aprimorar continuamente seus resultados e melhorar decisões (Seth & Chaudhary, 2015).

Adicionalmente, entre 80 e 90% dos dados são disponibilizados de forma não padronizada e não estruturada, fato que cria dificuldades de integração e *gaps* de qualidade que ampliam os desafios para o processamento e a realização de análises com agilidade necessária para a agregação de valor (Oracle, 2012). Há, portanto, a necessidade de estruturas robustas de bancos de dados que permitam a acumulação de dados e possibilidades de cruzamento para pesquisas e colaborações interdisciplinares (Demchenko *et al.*, 2013).

A visão de Big Data como acumulação acelerada de dados por si não tem necessariamente valor. É preciso que ferramentas de **análise** também avancem na mesma velocidade. O volume de informação e seus diferentes formatos têm afetado as pesquisas em múltiplas dimensões, já que as tecnologias e os modelos tradicionais falham ou apresentam grandes inconsistências quando se tem acesso a um volume tão grande de dados (Einav & Levin, 2014).

É nesse contexto que surge a necessidade de se desenvolverem técnicas robustas e adaptadas a esse tipo de situação. Ou seja, novas metodologias devem emergir, inclusive com pressupostos estatísticos que considerem a utilização de amostras que são muito próximas de abranger a população estudada (Einav & Levin, 2014), ou que apresentem maior consistência em termos informativos que as metodologias convencionais de estimação e teste de hipóteses. Por exemplo, em procedimentos de estimação por intervalo que envolvem erro-padrão, esse erro-padrão é função inversa do tamanho da amostra, de modo que em amostras muito grandes o erro padrão tende a zero, o que torna a estimação por intervalo sem sentido. Uma consequência dessa constatação está na possibilidade de um efetivo retorno para (ou na revalorização das) análises de dados baseadas em 'magnitude do efeito' e menos em análises de significância (ou p--valores, conforme Gigerenzer & Marewski, 2015).

Ainda concernente às análises, outra questão emerge das restrições que virão em análises por modelagem, que, em pesquisas tradicionais, utilizam pressupostos rígidos, como, por exemplo, normalidade ou homocedasticidade de erros em modelos de regressão. Com a introdução de grandes volumes de dados, esses pressupostos são potencialmente quebrados, invalidando testes de confirmação de confiança dos resultados, inclusive porque os procedimentos convencionais de verificação desses pressupostos são baseados em testes estatísticos, que têm problemas de utilidade quando as amostras são muito grandes, dada a tendência de sempre se rejeitar a hipótese nula.

Adicionalmente, se não em termos globais, em certas segmentações de dados de algumas variáveis, sempre será possível a verificação de correlações espúrias, dada a grande variedade de dados. Isso gera aqui a necessidade de uma capacidade interpretativa diferenciada para os usuários tanto da técnica de correlação quanto de outras técnicas baseadas em correlação, como análise fatorial, por exemplo.

Em suma, tendo a abordagem de Big Data como uma direcionadora de tecnologias e métodos estatísticos, um grande número de componentes de infraestrutura tecnológica passa a ser repensado para superar desafios relacionados, principalmente, o crescente volume de dados coletados de diferentes fontes e formatos. Há, ainda, a necessidade de um esforço conjunto multidisciplinar entre áreas, tais como Administração, Economia, Matemática, Estatística e Ciência da Computação, no intuito de redefinir os modelos utilizados para tais análises.

A perspectiva de impactos acadêmicos e profissionais decorrentes dessa perspectiva é ampla e vai encontrando repercussões em várias áreas de estudo e pesquisa. Para este artigo, e como já indicado na Introdução, tomamos por desafio analisar as implicações no contexto específico de Finanças e Mercado de Capitais, o que fazemos no item a seguir.

4. Big Data e sua contribuição para estudos de Finanças e o Mercado de capitais

Os dados de transações e atividades do universo das Finanças, seu processamento e seu acesso têm um papel central na área de Finanças (Seth & Chaudhary, 2015). Os mercados de capitais, principal fonte de dados para análises nesta área, transformaram-se significativamente desde os anos 2000 ao atingir níveis cada vez mais altos de geração de Dados Massivos de Alta Frequência (HFD), a exemplo do mercado estadunidense, que possui cerca de 70% de todos os dados de negociação em HDF (Zervoudakis, Lawrence, Gontikas & Al Merey, 2017). Nesse contexto, pode-se observar uma típica configuração de Big Data, dado o registro quase contínuo de atividades, as diferentes formas de manifestação de tais atividades e o elevado volume de registros.

Todas as áreas de Finanças (análise de investimentos, econometria, detecção de fraude, Finanças comportamentais, entre outras) podem se beneficiar desse cenário, já que a análise de dados intensivos amplia a capacidade de se medirem tanto os riscos-sistemáticos, que abrangem o mercado como um todo - como os não sistemáticos - relacionados a cada uma das empresas (Fan, Han & Liu, 2014). Neste ensaio, optamos por restringir os comentários aos potenciais impactos da visão do Big Data nas análises, envolvendo três áreas das Finanças que possuem relação direta com o mercado de capitais: volatilidade, elaboração de carteiras e análise de risco e transparência de mercado. O caráter de ensaio e a nossa intenção de fixar uma referência mais ilustrativa que exaustiva nos conduziram a uma exposição breve, limitada a apontar implicações genéricas e potenciais desdobramentos.

A **volatilidade** é comumente definida como a dispersão do preço de um ativo durante certo período de tempo e é uma área que tem, historicamente, significativa relevância nas pesquisas em Finanças, dado que a medição e a projeção de volatilidade dos ativos são cruciais para atividades como alocações de ativos, precificação de derivativos e opções de análise e gerenciamento de investimentos (Seth & Chaudhary, 2015). Com efeito, os estudos que buscam medir a volatilidade de mercado no contexto de Big Data têm a possibilidade de fazer avançar a precisão de seus modelos de predição a partir da manipulação de grandes volumes de dados (Louzis, Xanthopoulos-Sisinis & Refenes, 2013), de modo a gerar análises mais confiáveis que proporcionem ganhos para seus investidores.

Avaliações financeiras mais precisas podem minimizar a probabilidade de falhas, especialmente em períodos de crise, quando a volatilidade pode se ampliar de forma significativa. Adicionalmente, a volatilidade apresenta alto poder explicativo de resultados devido à alta persistência e dependência condicional, sendo não estacionária e previsível (Louzis, Xanthopoulos-Sisinis & Refenes, 2013). Dada a sua relevância, muitos modelos foram desenvolvidos para estimar volatilidade, sendo o modelo autorregressivo de heterocedasticidade condicional (ARCH) de Engel e suas variações o mais destacado como um dos mais populares no campo (Seth & Chaudhary, 2015). No entanto, esses modelos foram concebidos para utilizar dados de baixa frequência, com retornos diários, perdendo informações intradiárias que podem ser valiosas para ampliar a precisão dessas estimações.

O acesso a dados em alta frequência, o que é típico em Big Data, tem a potencialidade de alterar esse cenário de forma significativa na medida em que modelos estatísticos e teorias sejam desenvolvidos, considerando as adequações necessárias para a utilização desses dados (Cartea & Karyampas, 2011). Em suma, no que concerne à volatilidade, a abordagem de Big Data influencia no aprimoramento da medição, dos processos de predição e dos modelos, a partir tanto do grande volume de dados quanto das características de continuidade da geração de dados.

Com relação à **elaboração de carteiras de investimento e análise de risco** para a seleção de ativos, a perspectiva do Big Data também tem ampliado os horizontes. Até o início dos anos 2010, os modelos de avaliação de ativos têm se baseado fortemente em dados específicos da companhia sob análise ou da economia a qual está exposta, proporcionando basicamente uma análise fundamentalista (Damodaran, 2012). Com o crescimento da disponibilidade de dados somado à possibilidade de processamento e quantificação, os modelos de avaliação de empresas passaram a considerar uma gama muito maior de fontes de informação do que há dez anos, ampliando sua capacidade de identificação das melhores oportunidades (GSAM, 2016). Essas inovações incluem a análise de textos, imagens, áudios de reuniões com os acionistas, apresentações, entre outras informações em formatos não estruturados.

O desenvolvimento das técnicas de *machine learning* e de ferramentas de *statistical learning* (James, Witten, Hastie & Tibshirani, 2013) também tem contribuído para a criação de modelos de avaliação mais dinâmicos, que se adaptam e aprendem com o constante processamento de grandes volumes dados, acelerando o tempo em que a informação fica disponível para a tomada de decisão dos investidores ou analistas (GSAM, 2016). Adicionalmente, em mercados emergentes, como o Brasil, nos quais a assimetria de informação é mais acentuada, a deficiência de informações pode levar à ampliação da incerteza e ao erro de precificação dos ativos (Martins & Paulo, 2014). Assim, os modelos baseados em dados intensivos têm a potencialidade de criar uma vantagem adicional para os investidores, reduzindo a assimetria e direcionando mais claramente o capital para os melhores investimentos.

Além dos benefícios monetários e de eficiência financeira que esses avanços podem proporcionar para os diferentes agentes de mercado, há uma área das Finanças que deve ganhar especial relevância nesse contexto, que é a área de **transparência de mercado**. A utilização de ferramentas Big Data, como as de *business intelligence* ou de *business analytics*, traz uma nova perspectiva para os estudos e esforços relacionados à ampliação da transparência e, consequentemente, possibilita o aumento da liquidez, eficiência de mercado (Ye, 2010), e um ambiente mais estável e confiável para o seu desenvolvimento.

As demandas do mercado de capitais por maiores níveis de transparência ganharam força especialmente após os escândalos financeiros do início dos anos 2000 que envolveram grandes corporações, como a Enron e a WorldCom, o que motivou diferentes discussões sobre seu conteúdo e regulação (Chong & Lopez-de-Silanes, 2007). Esses eventos desencadearam uma intervenção governamental sem precedentes no mercado de capitais dos Estados Unidos (EUA), que foi seguida pela maior parte dos países, e teve como ponto de destaque o estabelecimento da Lei Sarbanes-Oxley (Aksu & Kosedag, 2006). Mais recentemente, a crise de crédito imobiliário nos EUA em 2008 revelou que ainda havia significativas limitações e inadequações do sistema de divulgação de informações dentro do sistema financeiro. A falta de procedimentos padrão, *gaps* de qualidade na divulgação de informação e falta de capacidade de processamento de dados em um tempo adequado levaram os órgãos reguladores a não terem a capacidade de processar tais informações, o que impossibilitou medidas proativas ou a identificação mais precisa de quais informações estavam faltando (Seth & Chaudhary, 2015).

Assim, nesse cenário de crises, a demanda dos órgãos reguladores tem desempenhado um papel-chave na aceleração do desenvolvimento de soluções para processamento de Big Data na busca por graus mais elevados de transparência (Oracle, 2012). Esse movimento, obviamente, não se restringiu ao mercado de capitais, pois todas as companhias que nele operam têm sido forçadas a ampliar o volume de dados divulgados e a redesenhar sua infraestrutura tecnológica para suportar tal demanda.

Esses avanços possibilitam o monitoramento e a identificação de fraudes e outras atuações que possam colocar a transparência e a confiabilidade dos mercados em risco com mais agilidade e precisão (Louzis, Xanthopoulos-Sisinis & Refenes, 2013), o que preservaria e ampliaria a liquidez e a eficiência de mercado. Desta forma, a visão do Big Data traz novas possibilidades para inovação e crescimento da área que, por ter uma característica de ampliação da acessibilidade, possibilita que órgãos reguladores, companhias e pesquisadores possam colaborar para a obtenção de soluções conjuntamente, extraindo o máximo de valor dos dados.

Nesse cenário, as discussões relacionadas à transparência têm alterado seu foco do requerimento de volume de dados, já que são mais abundantes, para refletir sobre como superar os desafios de seu processamento (Seth & Chaudhary, 2015). Alguns exemplos de tais desafios poderiam ser: *Como processar em tempo hábil um volume crescente de dados? Como superar a ausência de protocolos padrões de divulgação? Como definir programações que identifiquem com agilidade fraudes?* A busca por respostas para essas e para tantas outras questões relacionadas a transformar dados do mercado em conhecimento pode gerar ferramentas de regulação cada vez mais efetivas para que os mercados atinjam um novo nível de transparência e eficiência.

Esses três aspectos que comentamos de forma breve, além de diversas outras questões que podem surgir nas diferentes áreas de pesquisa em Finanças, podem se beneficiar não só da ampliação de dados, proporcionada por acesso a Big Data, mas, também, de uma nova percepção epistemológica de como esses dados podem ser analisados, principalmente em nível acadêmico. É aqui que a utilização dos métodos de inspiração mais abdutiva possibilitaria a ampliação do escopo de informações sobre as quais o pesquisador se debruça para conhecer a realidade, estimulando a observação de fenômenos não restritos a um conjunto de hipóteses propostas para serem confirmadas ou refutadas. Este novo olhar, que considera em um primeiro momento um movimento indutivo, tem a potencialidade de criar uma revolução em uma área que historicamente tem se baseado em métodos puramente hipotético-dedutivos para a construção do conhecimento.

5. Considerações Finais

A incorporação da perspectiva de Big Data, tanto na academia como no universo profissional, parece ser um movimento inevitável e irreversível. A possibilidade de basear a análise em dados com magnitudes quase censitárias amplia as possibilidades de minimizar assimetria de informação em todos os campos das ciências e dos negócios, ao alterar a preocupação da obtenção de dados para a busca de ferramentas de processamento e análise. Uma 'nova linguagem multidisciplinar' deverá emergir para abarcar essa crescente complexidade, redesenhando e adequando metodologias e protocolos para a obtenção do máximo de valor dos dados.

Este novo panorama gera questões ainda mais profundas relacionadas à construção do conhecimento como tem sido tradicionalmente desenvolvido. Com o estabelecimento das ciências baseadas em dados (*data-drive science*; conforme Kitchin, 2014), a epistemologia mais tradicional, baseada fortemente no método hipotético-dedutivo, parece ser limitada para alcançar todo o valor dos dados que a perspectiva do Big Data disponibiliza.

As discussões relacionadas ao Big Data sobre o aprimoramento do método, do processamento e protocolos metodológicos para a extração do maior valor possível de dados é importante para as diversas dimensões das Finanças, que é uma área em que os dados desempenham um papel central para a produção de conhecimento e para decisões gerenciais e regulatórias. Destacamos aqui os estudos relacionados à volatilidade, à elaboração de carteiras e análise de risco e à transparência de mercado, que são áreas que já iniciaram o desenvolvimento das ferramentas necessárias para se beneficiar de dados intensivos, mas ainda apresentam muitos desafios de processamento para que se possa extrair o máximo de valor.

Buscamos analisar como as características de Big Data, em particular as de volume, velocidade e variedade, alcançam as práticas de pesquisas e gestão nesses campos. Mesmo para uma visualização genérica e restrita a esses três temas, já foi possível observar o quanto essas áreas deverão se aprimorar na medida em que se apropriam de mais dados, com acesso que se aproxima da continuidade, e com uso ou desenvolvimento de melhores ferramentas para manipular esses crescentes volumes e maior diversidade de tipos e formatos de dados. Deixamos como desafio para outros estudos tanto aprofundar análise de implicações de Big Data nesses campos quanto em outras especialidades da área de Finanças que não abordamos.

Esse cenário não requer ainda a mudança na perspectiva paradigmática em termos metodológicos e epistemológicos. Por isso, sugerimos ao longo do texto e enfatizamos novamente que a incorporação de uma abordagem abdutiva nas pesquisas da área de Finanças poderá ampliar a capacidade de extração de informações de valor e possibilitará a construção do conhecimento de forma mais completa, não limitando as análises ao que se pergunta como hipóteses previamente definidas, mas acessando todas as possíveis informações disponíveis nos dados. A busca da solução desses desafios tem demandado não só uma visão multidisciplinar, mas também uma aproximação de diferentes atores da sociedade que, por vezes, ficam distantes na resolução de desafios, em especial governos (órgãos reguladores), companhias e acadêmicos. Sugerimos, portanto, que estudos empíricos analisem visões, desafios e experiências concretas envolvendo esses grupos de interesse no campo profissional de Finanças, de modo a gerar uma visão cada vez mais completa e consistente das melhores práticas e melhores decisões no campo de Finanças no Brasil.

Referências

Aksu, M., & Kosedag, A. (2006). Transparency and disclosure scores and their determinants in the Istanbul Stock Exchange. *Corporate Governance - an International Review*, *14*(4), pp. 277-296. DOI: https://doi.org/10.1111/j.1467-8683.2006.00507.x/abstract

Bianchi, E. M. P. G., & Ikeda, A. A. (2008). Usos e aplicações da *grounded theory* em administração. *Revista Eletrônica de Gestão Organizacional*, *6*(2), pp.231-248.

Cartea, A., & Karyampas, D. (2011). Volatility and covariation of financial assets: a high-frequency analysis. *Journal of Banking and Finance*, *35*(12), pp.3319-3334. DOI: https://doi.org/10.1016/j.jbankfin.2011.05.012

Chen, C. L. P., & Zhang, C.Y. (2014). Data-intensive applications, challenges, techniques, and technologies: a survey on Big Data. *Information Science*, *27*(5), pp. 314-374. DOI: https://doi.org/10.1016/j.ins.2014.01.015

Chong, A., Lopez-de-Silanes, F. (2007). *Investor protection and corporate governance:* intra-firm evidence across Latin-America. Palo Alto: Stanford University Press.

Damodaran, A. (2012). *Investment valuation*: tools and techniques for determining the value of any assets. 3rd edition. New Jersey: Wiley&Sons.

Demchenko Y., Grosso, P., De Laat C., & Membrey, P. (2013). Addressing Big Data issues in scientific data infrastructure. *Collaboration Technologies and Systems* (CTS), International Conference on 2013.

Einav L. & Levin J. (2014). Economics in the age of big data. *Science*, *346* (6210), pp. 12430891–12430896. DOI: https://doi.org/10.1126/science.1243089

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., Tsou, A., Weingart, S., & Sugimoto, C. R. (2015). Big Data, bigger dillemas: a critical review. *Journal of the Association for Information Science and Technology*, *66*(8), pp. 1523-1545.

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data Analysis. *National Science Review*, *1*(2), pp. 293–314. DOI: https://doi.org/10.1093/nsr/nwt032

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: the idol of a universal method for scientific inference. *Journal of Management*, *41*(2), pp.421-440. DOI: https://doi.org/10.1177/0149206314547522

Glaser, B, G., & Strauss, A. L. (1967). *The discovery of grounded theory: strategies for qualitative research*. New York: Aldine de Gruyter.

GSAM – Goldman Sachs Asset Management (2016). Perspectives: the role of Big Data in investing. Recuperado em 26 de abril, 2017, de: https://www.gsam.com/.

Haig, B. D. (2005). An abdutive theory of scientific method. *Psychological Methods*, *10*(4), pp. 371-388. DOI: https://doi.org/10.1037/1082-989X.10.4.371

Hey, T., Tansley, S., & Tolle K. (2009). *Jim Grey on eScience: A transformed scientific method.* In: Hey T, Tansley S and Tolle K (eds) The fourth paradigm: data-intensive scientific discovery. Redmond: Microsoft Research, pp. xvii–xxxi.

Big Data, New Epistemologies and Paradigm Shift (PDF Download Available). Recuperado em 15 de setembro, 2017, de: https://www.researchgate.net/publication/271525133_Big_Data_New_Epistemologies_and_Paradigm_Shift.

IBM, 2016. Recuperado em 08 de abril, 2107, de: https://www-01.ibm.com/software/data/bigdata/what-is-bigdata.html.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: springer. DOI: https://doi.org/10.1007/978-1-4614-7138-7

Kitchin, R. (2014). Big Data, new epistemologies and paradigms shift. *Big Data & Society*, April-June, pp. 1-12. DOI: https://doi.org/10.1177/2053951714528481

Lakatos, I., & Musgrave, A. (1979). *A crítica e o desenvolvimento do conhecimento*. São Paulo: Editora da Universidade de São Paulo. DOI: http://dx.doi.org/10.11606/issn.2447-9799.cienciaefilosofi.1980.107354

Lau, R. Y.K., Zhao, J.L., Chen, G., & Guo, X. (2016). Big Data commerce. *Information & Management*, *53*, pp. 929-933. DOI: https://doi.org/10.1016/j.im.2016.07.008

Louzis, D.P., Xanthopoulos-Sisinis, S., & Refenes, A.P., (2013). The role of high-frequency intra-daily data, daily range and implied volatility in multi-period value-at-risk forecasting. *Journal of Forecasting*, *32*(6), pp. 561–576.

Martins,O. S. & Paulo E. (2014). Assimetria de informação na negociação de ações, características econômico financeiras e governança corporativa no mercado acionário brasileiro. *Revista Contabilidade & Finanças (Online)*, *25*(64), pp. 33-45. jan./fev./mar./abr. 2014

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, *50*(1), pp. 181-201. DOI: https://doi.org/10.1111/j.1467-9787.2009.00641.x

Oracle (2012). Financial services data management: Big Data technology in financial services. Oracle Financial Services, *An Oracle White paper.*

Seth, T., & Chaudhary, V. (2015). "Big Data in Finance". In: Li, K.; Jiang, H.; Yang, L. T.; Cuzzorea, A. (Eds) "Big Data: algorithms, analytics and applications". Chapman and Hall/ CRC, pp. 329-356.

Spens, K. M., & Kovács, G. (2006). A content analysis of research approaches in logistics research, *International Journal of Physical Distribution & Logistics Management*, *36*(5), pp. 374-390. DOI: https://doi.org/10.1108/09600030610676259

Ye, G. (2010). *High frequency trading models*. NJ: John Wiley & Sons Inc.

Zervoudakis, F., Lawrence, D., Gontikas, G., & Al Merey, M. (2017). *Perspectives on high-frequency trading*. Recuperado em 30 de março, 2017, de: http://www0.cs.ucl.ac.uk/staff/f.zervoudakis/docs/hft.pdf.