Ciência de Dados

Fábio Porto e Artur Ziviani

Laboratório Nacional de Computação Científica (LNCC)
Av. Getúlio Vargas, 333 – 25651-075 – Petrópolis, RJ

{fporto,ziviani}@lncc.br

Resumo. *Discutimos ciência de dados como um desafio da computação para os próximos anos. Esta proposta está relacionada primordialmente ao desafio "Gestão da Informação em grandes volumes de dados multimídia distribuídos" com aplicação amplamente interdisciplinar, cobrindo múltiplos domínios do eixo ciência-indústria-governo transversalmente.*

1. Introdução

O tratamento do dilúvio de dados sendo produzido pelas ciências e por bilhões de usuários de serviços de Internet globais se apresenta como um dos grandes desafios para a atual sociedade do conhecimento [Bell et al., 2009]. Apresentado de forma geral como um vetor de múltiplas facetas, o fenômeno ainda está sendo interpretado pelos cientistas e vem impulsionando iniciativas em diversas áreas. Nas ciências, o dilúvio apareceu como a expressão de uma nova maneira de investigação [Wright, 2014], incitando biólogos, astrônomos, físicos, e demais pesquisadores em diversas áreas científicas, a enfrentarem problemas computacionais na chamada e-ciência, que se tornam barreiras para as suas descobertas. Na indústria, o dilúvio de dados aparece fortemente como análise preditiva [Dhar, 2013] em sintonia com o ambiente de computação em nuvem, provendo escalabilidade e tolerância a falhas, em ambientes computacionais cada vez mais complexos e de tamanho proporcional ao desafio. Na setor governamental, há oportunidades de se debruçar sobre imensas bases de dados do setor público com vistas a gerar planejamento mais eficiente bem como novos serviços que possam melhorar o atendimento ao cidadão. Novas profissões especializadas no trato e, principalmente, na análise e interpretação de grandes volumes de dados, surgiram, trazendo o método científico para o setor empresarial.

Neste contexto, constitui-se um desafio técnico-científico em computação o estudo metódico para a extração generalizada e em escala de conhecimento relevante a partir de uma imensa massa de dados, em geral dinâmicos [Jagadish et al., 2014]. A abordagem a esse desafio com aplicações em diversas áreas no eixo ciência-indústria-governo emerge como uma nova espécie de ciência. A chamada *Ciência de Dados* incorpora elementos variados e se baseia em técnicas e teorias oriundas de muitos campos básicos em engenharia e ciências básicas, sendo assim intimamente ligada com muitas das disciplinas tradicionais bem estabelecidas, porém viabilizando uma nova área altamente interdisciplinar. Dessa forma, associado a este espírito de aplicação interdisciplinar, a ciência de dados emerge como componente cada vez mais importante nas mais diversas áreas, tais como saúde, petróleo, energia, financeira, esporte, astronomia, bioinformática, Internet, mobilidade urbana, defesa cibernética, comunicação móvel e biodiversidade, apenas para mencionar algumas.

Em ambiente altamente interdisciplinar com aplicações em áreas tão distintas, emerge o *grande desafio* comum às aplicações nessas tão diversas áreas de se identificar os princípios, métodos e técnicas fundamentais para o gerenciamento e análise de grandes volumes de dados, suplantando as dificuldades inerentes ao grande volume de dados em análise [Jacobs, 2009, Lazer et al., 2014]. Especificamente, identificamos três linhas de pesquisa principais cujo amadurecimento acreditamos deverá conduzir rumo à consolidação da área de ciência de dados em um horizonte de alguns anos de pesquisa e desenvolvimento: (i) gerência de dados; (ii) análise de dados; e (iii) análise de redes complexas; todos essas linhas considerando a larga-escala dos dados a serem analisados bem como seu dinamismo. A partir da pesquisa básica nesses aspectos fundamentais de análise de dados em larga-escala, há também um grande potencial tecnológico na pesquisa aplicada em ciência de dados com impacto em diferentes áreas do conhecimento e de setores de atuação ao longo do eixo ciência-indústria-governo.

Um desafio correlato se torna a formação de recursos humanos altamente qualificados no desenvolvimento de pesquisa básica e aplicada na fronteira do conhecimento em ciência de dados. Esse *cientista de dados* possui demanda crescente no eixo ciência-indústria-governo [Davenport e Patil, 2012]. Esse profissional tem uma expectativa de formação tipicamente sólida em ciência da computação e aplicações, modelagem, estatística, analítica e matemática, além do conhecimento mínimo do domínio de aplicação.

Em suma, enfrentar de forma fundamental o grande desafio da ciência de dados permite contribuir a melhor posicionar o Brasil na direção da nova ciência baseada em dados, preparando recursos humanos altamente qualificados, e desenvolvendo o alicerce para sua projeção de forma relevante na sociedade do conhecimento.

O restante deste documento está organizado como segue. Na Seção 2 apresentamos a motivação desta proposta de grande desafio em ciência de dados, tendo por base diferentes cenários de aplicação. Delineamos os desafios de pesquisa básica em ciência de dados nas três principais linhas de pesquisa identificadas ao longo da Seção 3. Os desafios relacionados à formação de recursos humanos na área de ciência de dados são apresentados na Seção 4. Finalmente, a Seção 5 traz algumas considerações finais.

2. Motivação para pesquisa básica e aplicada em ciência de dados

A grande motivação para nossa proposta de desafio relacionado à ciência de dados emerge da experiência anterior em realizar atividades de pesquisa e desenvolvimento em gestão e análise de dados, bem como análise de redes complexas, em cenários de aplicação das áreas mais diversas. Exemplos são astronomia, biodiversidade, Internet, petróleo & gás, saúde e comunicação móvel. Nesta seção, apresentamos uma descrição sucinta da relevância de ciência de dados nesses cenários de aplicação, já em investigação pelos autores.

Essa experiência anterior, portanto, permitiu a identificação de um clamor por pesquisa básica nos aspectos fundamentais de análise de dados em larga-escala, o principal ponto de motivação para a nossa proposta de ciência de dados como grande desafio à computação nos próximos anos. Isso também traz consigo a qualificação e justificativa da relevância deste desafio dado o espectro amplo de impacto e aplicação de avanços de ciência de dados ao longo das linhas de pesquisa delineadas nestes cenários de aplicação, bem como em outros oriundos do eixo ciência-indústria-governo.

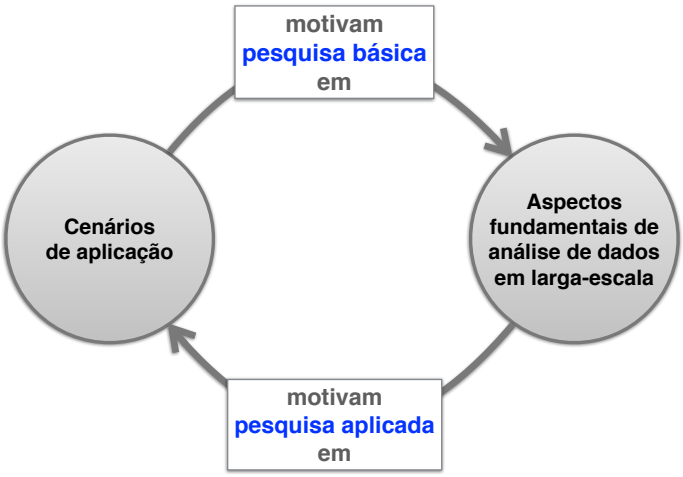

Figura 1. Motivação cíclica para pesquisa básica e aplicada em ciência de dados.

As diferentes aplicações práticas de ciência de dados, tais como os cenários de aplicação ilustrativos descritos nesta seção, ao mesmo tempo em que são alvos para elaboração de novas soluções em pesquisa aplicada, muitas vezes propiciam a oportunidade de elaboração de novos arcabouços teóricos em pesquisa básica, de caráter mais geral, para a solução dos problemas práticos. A Figura 1 ilustra esse ciclo de motivação para a pesquisa básica e aplicada em ciência de dados. Essa abordagem que liga teoria e prática é uma das estratégias gerais adotadas pelos autores, que tem as suas pesquisas centradas na análise de dados em diferentes campos. Exemplos ilustrativos de cenários de aplicação atuais de ciência de dados em áreas diversas, nos quais os autores tem experiência, são:

1. **Astronomia:** O LNCC é membro do Laboratório Inter-institucional de eAstronomia (LIneA),[1] onde tem-se gerenciado e processado dados obtidos de grandes levantamentos astronômicos. Estes levantamentos produzem dados a partir de imagens telescópicas fotografadas por instrumentos terrestres. A partir das imagens, corpos celestes são identificados e suas características anotadas produzindo um conjunto de dados chamado Catálogo Astronômico. Tais catálogos podem abrigar até centenas de bilhões de objetos celestes. Processar tal volume incomum de dados desses catálogos de forma eficiente requer seu particionamento e alocação distribuída em um cluster. Estratégias de particionamento devem atender a requisitos de diferentes dataflows de análise, dificultando a determinação de critérios adequados a distintas aplicações. Encontrar estratégias que coincidam com os critérios de dataflows é um problema em aberto. A integração de diversos catálogos, produzidos por diferentes levantamentos, também traz o problema de resolução de entidades, uma vez que a identificação de objetos estrelares é feita com base em sua posição, cuja medida varia em diferentes telescópios [Freire et al., 2014].
2. **Biodiversidade:** Para monitorar as mudanças na biodiversidade é essencial coletar, documentar, armazenar e analisar indicadores sobre a distribuição espaço-

[1] `http://www.linea.gov.br`

temporal das espécies, além de obter informações sobre como elas interagem entre si e com o ambiente em que vivem [Michener et al., 2012]. Nesse contexto, o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr)[2] visa integrar e disseminar dados coletados e publicados por diversas instituições brasileiras. O SiBBr, cuja infraestrutura computacional está sediada no LNCC, desempenha também o papel de nó brasileiro do Global Biodiversity Information Facility (GBIF).[3] O SiBBr já permite a agregação de dados de espécies e ocorrências disponibilizadas por diversas instituições acadêmicas e de pesquisa bem como de órgãos governamentais. Um primeiro protótipo de workflow científico para modelagem de distribuição de espécies [Gadelha Jr. et al., 2012b] permite uma execução escalável e com registro de informações de proveniência [Gadelha Jr. et al., 2012a].

3. **Internet:** A Internet apresenta grandes desafios para a caracterização de sua estrutura e comportamento [Chen, 2001]. De fato, a Internet mostra-se atualmente como um conjunto de redes complexas interdependentes entre si, abrangendo desde as redes de comunicação que formam a infraestrutura básica de interconexão até redes sociais online envolvendo bilhões de usuários, passando por redes no nível aplicativo de troca de conteúdo. Há, portanto, grandes desafios para a caracterização, análise e modelagem de tais redes na Internet, bem como a rede WWW sobre esta, pois esses estudos devem também preservar a privacidade de usuários, o que impõe desafios adicionais à coleta eficiente e detalhada de informações importantes para condução de pesquisa. Os autores contam com experiências diversas em diferentes aspectos da coleta, análise e modelagem do imenso volume de dados envolvidos na investigação da estrutura e comportamento das atuais redes complexas que são formadas na e pela Internet bem como o impacto em suas aplicações [Gueye et al., 2006, Freire et al., 2008, Ziviani et al., 2012, Las-Casas et al., 2013].
4. **Petróleo e Gás:** A pesquisa de petróleo e gás em áreas profundas é um grande desafio no Brasil, com grandes campos em profundidades de mais de 5 kms na área do pré-sal. A investigação nestes campos envolve a captura de reflexos de ondas sísmicas enviadas a partir da superfície. Ondas enviadas em direção às camadas subaquáticas são refletidas por camadas rochosas no fundo do mar e recapturadas por sensores na superfície. Uma vez capturadas e processadas para limpeza dos dados, os chamados traços sísmicos são combinados em um grande conjunto de dados representando a região investigada. A atividade de analisar os sinais sísmicos para detecção de feições de interesse é chamada de interpretação geofísica e tem valor econômico bastante relevante. Neste sentido, o desenvolvimento de técnicas que possam apoiar a detecção de falhas sobre campos muito grandes, como o pré-sal brasileiro, é um problema cujas soluções estão ainda em sua infância. Além do problema básico da gerência de grande volume de dados, a inferência de feições a partir de sinais em ondas sísmicas é um grande desafio.
5. **Saúde:** A área de saúde lida rotineiramente com enormes quantidades de dados. Esse volume de dados somente aumenta devido à adoção crescente de sistemas de informação em saúde e prontuários eletrônicos do pa-

[2] http://www.sibbr.gov.br
[3] http://www.gbif.org

ciente. O LNCC tem experiência na área de sistemas de informação em saúde [Correa et al., 2011, Gomes et al., 2012], além da instituição ser a atual sede do INCT-MACC (INCT em Medicina Assistida por Computação Científica).[4] Há grandes desafios na gestão e análise de dados ligados à área de saúde, tais como a agregação, manutenção, interoperabilidade, interpretação desses dados, sem mencionar questões de privacidade devido à evidente sensibilidade dos dados [Nambiar et al., 2013]. A tendência é de ainda maior expansão no volume de dados num futuro próximo devido ao uso crescente de sensores ou mesmo dispositivos móveis para coleta de dados individualizados em ambientes residenciais ou pré-hospitalares [Estrin, 2014]. Outra tendência recente é a abordagem de modelagem por redes complexas de problemas relativos à área de saúde, seja relacionando doenças [Barabási et al., 2011], seja relacionando serviços de saúde para melhor coordenação de cuidados e uso de recursos de forma centrada no paciente [Pretz, 2014].

6. **Comunicação móvel:** Dados coletados de redes de telefonia celular tem um enorme potencial de prover informações valiosas sobre o relacionamento dinâmico de indivíduos [Eagle et al., 2009] ou sobre mobilidade humana [Becker et al., 2013] a um custo relativamente baixo e numa escala sem precedentes. A análise de enormes volumes de dados de redes celulares hoje apresenta impacto em diversas áreas, de melhor planejamento e dimensionamento das próprias redes de telecomunicação até mais indiretamente, por exemplo, planejamento urbano [Iqbal et al., 2014]. O LNCC tem experiência, com colaboradores, no estudo de dados de redes celulares para a investigação da dinâmica da carga da rede e mobilidade humana devido a eventos de larga-escala em ambiente urbano [Xavier et al., 2012, Xavier et al., 2013]. São desafios nesse cenário de aplicação a gestão e análise dos dados em grande volume, assim como a análise das redes complexas de relacionamento que emergem tipicamente desse tipo de dado.

3. Desafios de pesquisa em ciência de dados

Esta proposta de desafio em ciência de dados situa-se diretamente relacionada ao tema estratégico de tecnologias da informação e comunicação (TICs), tendo relação imediata com o primeiro dos programas prioritários para os setores portadores de futuro, tal como definido na Estratégia Nacional de Ciência, Tecnologia e Inovação (ENCTI) 2012-2015,[5] publicada pelo Ministério da Ciência, Tecnologia e Inovação (MCTI). De fato, gestão de informação em grandes volumes de dados já era reconhecido em 2006 com um dos grandes desafios para a computação brasileira para os 10 anos seguintes no primeiro relatório a respeito realizado pela Sociedade Brasileira de Computação (SBC). Esses desafios são hoje ainda maiores vistos os atuais volumes de dados a analisar, bem como seu dinamismo e capilaridade, que surgem como oportunidades de progresso científico e inovação tecnológica em diferentes áreas do eixo ciência-indústria-governo.

Nesta seção, as linhas de pesquisa identificadas como portadoras de desafios em ciência de dados para os próximos anos são detalhadas, sendo essas: (i) gestão de dados; (ii) análise de dados; e (iii) análise de redes complexas; todos essas linhas considerando

[4] `http://macc.lncc.br`

[5] `http://www.mct.gov.br/upd_blob/0218/218981.pdf`

Figura 2. Desafios de pesquisa básica em ciência de dados para diversos cenários de aplicação no eixo ciência-indústria-governo.

a larga-escala dos dados a serem analisados bem como seu dinamismo. Avaliadas de modo integrado, essas linhas de pesquisa encontram-se nos desafios apresentados ao se confrontar com os grandes volumes de dados produzidos atualmente nas ciências, governo e indústria. A Figura 2 ilustra essa visão de ciência de dados como central a diversas áreas do eixo ciência-indústria-governo, tendo as linhas de pesquisa propostas como ponto de ligação entre essas áreas e ciência de dados. Acreditamos não serem essas linhas de pesquisa absolutamente exaustivas quanto aos desafios de ciência de dados. Porém estas cobrem em grande parte os aspectos fundamentais da pesquisa básica e aplicada em ciência de dados em larga-escala de forma transversal a diversos cenários de aplicação no eixo ciência-indústria-governo, o foco principal motivador de nossa proposta de desafio.

3.1. Gestão de dados

3.1.1. Representação de dados

O sucesso de Sistemas de Bancos de Dados Relacionais aplicados a problemas comerciais reduziu a representação de dados a tabelas bidimensionais. Em se tratando de gerência de grandes volumes de dados, a propriedade de sua representação se reflete positivamente no desempenho do acesso aos dados e, por sua vez, espelha necessidades de aplicações mais complexas do que aquelas apoiadas por bancos de dados relacionais. Assim, domínios que têm sido em grande parte negligenciados pelo suporte de banco de dados, tais como simulações numéricas, análises sísmicas e redes de interação gênica, para citar apenas alguns de muitos domínios nesta área, requerem representações de dados em sintonia com representações complexas, tais como: espaço, tempo, grafos e seqüências.

Sistemas como o SciDB,[6] por exemplo, propõem a representação de dados científicos em matrizes multidimensionais, o que de fato se apresenta como uma estrutura interessante, na medida em que generaliza o conceito de sistemas de coordenadas como índice de valores de variáveis calculadas. Vimos, no entanto, em [Costa et al., 2012], que o modelo apresenta problemas na representação de malhas irregulares e que mais esforços

[6] http://www.scidb.org

precisam ser dedicados nessa área. Igualmente, dados do tipo sísmico do tipo *pos-stack* apresentam como estrutura uma lista de seqüências, representando traços sísmicos, e ordenadas no espaço 2D e 3D. Efetuar consultas em dados deste tipo é um grande desafio e pode alavancar pesquisas importantes na área de petróleo, por exemplo. Ainda nesta linha, dados de acompanhamento de atletas [Porto et al., 2012] podem ser modelados por trajetórias virtuais, onde cada ponto reflete o valor de um elemento observável de interesse em múltiplas dimensões, incluindo o estado de treinamento, o atleta sendo observado e a data da observação. O mesmo modelo [Spaccapietra et al., 2008] tem sido usado em soluções de mobilidade urbana para representação de indivíduos em regiões urbanas, a partir de dados de mobilidade obtidos durante o uso de telefones celulares.

Um outro modelo que tem sido bastante utilizado em contextos bastante distintos como redes sociais e biologia de sistemas utiliza a representação em grafos, formando uma rede complexa (ver Seção 3.3). Interações humanas ou as produzidas pela expressão de certos genes podem ser fielmente retratadas em associações entre nós de um grafo, onde estes espelham os objetos do domínio de interesse (i.e. pessoas ou genes) e aqueles referem-se aos tipos de interações ocorrendo entre esses objetos. Características dos objetos ou das interações podem ser retratadas em suas propriedades, de forma semelhante aos atributos de relações. Na medida em que o conjunto de dados que se pretende representar em grafos aumenta, problemas como o desempenho de consultas apontam para estratégias semelhantes às adotadas em sistemas relacionais, como o particionamento. No entanto, a natureza de consultas em grafos baseada na navegação através das associações tornam a definição de critérios de particionamento de dados mais difícil.

Finalmente, de forma mais comum, uma mesma aplicação pode requerer a representação de parte de seus dados em diferentes modelos. Assim, retornando ao cenário de biologia de sistemas, uma aplicação de inferência de redes de interação gênica pode utilizar o modelo de grafos para representar as interações, o modelo relacional para características dos genes e organismos envolvidos e um modelo XML para armazenamento de dados de proveniência do processo de inferência.

Um desafio é a pesquisa de estruturas para representação de dados, como acima discutidas, tendo por exemplo as aplicações discutidas na Seção 2 como cenários motivadores da investigação. O objetivo principal seria identificar as lacunas na adoção de tais representações, principalmente tendo como alvos a expressividade para aplicações e o alcançado desempenho para grandes volumes de dados.

3.1.2. Tratamento de incerteza em dados

A crescente disponibilização de informações, seja na web ou capturadas por instrumentos e sensores, potencialmente subsidia tomada de decisões mais precisas. Isto ocorre em parte pelo maior conhecimento do fenômeno observado, seus estados limiares e casos espúrios. No entanto, conforme discutido em [Srivastava, 2012], informações contraditórias disponibilizadas de forma independente produzem o resultado oposto, confundindo os usuários. Neste aspecto, graus de confiança podem ser atrelados às fontes de informação, caracterizando-as e permitindo que sejam selecionadas a partir de um modelo de custo. No contexto de dados científicos, a acurácia da informação pode depender do instrumento sendo utilizado para sua captura. Em cenários de múltiplos instrumentos,

semelhantes a diferentes fontes de dados, a discrepância entre caracterizações de objetos comuns a várias fontes dificulta uma possível integração e torna complexa respostas a consultas. Finalmente, em modelos de simulação computacional, a incerteza própria dos modelos e dos parâmetros precisa ser identificada e informada ao usuário para tomada de decisão [Gonçalves e Porto, 2014].

Em geral, tais exemplos mostram uma faceta de dados em grandes volumes associados à imprecisão da informação. Enquanto em situações mais controladas os dados se mostram bem comportados, em contextos em que dados são capturados de diversas fontes autônomas, ou que advêm de processos por natureza imprecisos tais como modelos computacionais, é preciso equipá-los com características de incerteza e propagar essa representação nas inferências deles extraídas [Suciu et al., 2011].

Neste tópico, um desafio é se avaliar o tratamento de incerteza em representações como as discutidas na Seção 3.1.1. Em modelos de matrizes multi-dimensionais, por exemplo, a incerteza pode se distribuir ao longo do espaço-tempo e novos mecanismos de inferência probabilística podem ser necessários. Igualmente, no cálculo de inferência de redes de interações, as associações se estabelecem de forma imprecisa e será interessante avaliar a atribuição de incerteza em nós e arestas de grafos.

3.1.3. Particionamento de dados

O processamento de grandes volumes de dados, para que se torne escalável, requer o particionamento de dados entre nós de um cluster de computadores. O problema de particionamento de dados é antigo [Özsu e Valduriez, 2011], no entanto os tipos de aplicações que acessam grandes volumes de dados é de natureza fundamentalmente distinta das aplicações convencionais: (i) aplicações varrem grandes volumes de dados; (ii) estratégias de acesso variam devido à característica exploratória do processo de pesquisa; (iii) dados não sofrem atualizações. Neste contexto, as estratégias de particionamento precisam ser re-avaliadas. No conhecido framework Hadoop,[7] pedaços de tamanho uniforme, sem conotação semântica, estabelecem as fronteiras de cada unidade de alocação (i.e., partição). Fica evidente que estratégias mais próximas das características dos dados devem favorecer tanto o armazenamento quanto o acesso.

Neste sentido, [Curino et al., 2010] propõem uma estratégia de particionamento chamada Schism, que se baseia na análise de grafos em que nós representam os objetos e arestas o acesso conjunto daqueles. Dessa forma, algoritmos como *min-cut* encontram partições que minimizam o acesso em mais de uma partição (i.e. maximizam o acesso local). Apesar de interessante, as estratégias derivadas de Schism não são adequadas para o particionamento de petabytes de dados e, muito menos, para acesso de varredura de boa parte dos dados, como é o caso freqüente em grandes volumes de dados. Por outro lado, o processamento em dataflow requer mudança na estratégia de particionamento, uma vez que deve atender a uma seqüencia de atividades do dataflow. A combinação de paralelismo de tarefas em dataflows com particionamento de dados nas fontes, e aqueles produzidos por etapas intermediárias do dataflow, sugere que novas estratégias para o particionamento de dados devam ser avaliadas.

[7] `http://hadoop.apache.org`

Neste tópico de pesquisa, há o desafio de se investigar estratégias para o armazenamento de grandes volumes de dados, tendo como foco técnicas para particionamento, replicação e indexação de dados.

3.1.4. Processamento de grandes volumes de dados

A necessidade de se utilizar programas ad-hoc para o processamento intensivo de dados norteou o desenvolvimento de modelos de processo baseados em dataflows, cujo maior expoente é o paradigma MapReduce [Dean e Ghemawat, 2008] e sua implementação aberta Hadoop. Processos em dataflows se diferem do processamento de consultas tradicionais em vários aspectos: semântica de transformação de dados desconhecida; programas e dados fora do alcance dos SGBDs; otimização reduzida; estatísticas escassas, apenas para comentar alguns desses aspectos. Ainda assim, alguns trabalhos tentam recuperar para dataflows os benefícios de otimização automática e gerência de proveniência [Ogasawara et al., 2011, Hueske et al., 2012]. Neste contexto, a coincidência entre os critério de interesse de dados expresso pelas atividades e aquele referente ao do particionamento de dados devem dirigir a estratégia de execução. Modelos de execução norteados a processos intensivos de CPU, tal como o modelo de ativação, podem ser integrados ao dirigido a dados formando um modelo de execução heterogêneo apropriado à diferentes etapas do dataflow.

Neste tópico de pesquisa, é um desafio investigar estratégias e algoritmos para o processamento eficiente de grandes volumes de dados por dataflows, assim como novos sistemas tipo NoSQL [Mohan, 2013], privilegiando aspectos específicos de cada aplicação e dos dados nela armazenados.

3.2. Análise de dados

3.2.1. Processo de análise de dados

Em linhas gerais, a análise de dados corresponde a um conjunto de atividades que devem ser desempenhadas, desde a seleção dos dados até a produção do conhecimento, que é o principal produto da análise. A análise de dados envolve o processamento de coleções de objetos em busca de padrões consistentes, de forma a detectar relacionamentos sistemáticos entre variáveis componentes desses objetos e gerar conhecimento não facilmente detectado. Dá-se o nome de processo de análise de dados à especificação do encadeamento desse conjunto de atividades. As atividades que compõem o processo de análise de dados podem ser organizadas em quatro etapas: seleção, pré-processamento, métodos de análise e avaliação [Han et al., 2006].

O processo de análise pode ser compreendido como um caso particular de experimentação científica *in-silico* [Stevens et al., 2007], no qual os dados são volumosos, as estruturas de dados precisam ser bem definidas e os métodos de análise de dados são computacionalmente intensivos. Neste contexto, é apropiado estabelecer um tratamento datacêntrico a esses experimentos, compreendendo a desafios diretamente ligados aos apresentados na Seção 3.1.1. A pesquisa envolve, então, estrutura de dados e algoritmos para apoiar tanto às etapas (seleção de dados, pré-processamento, algoritmos de mineração e análise), quanto ao processo de análise de dados como um todo.

No que tange ao processo de análise de dados, há também uma necessidade preemente de utilizar processamento de alto desempenho (PAD) para se conseguir realizar a análise de dados em larga-escala. Além dos desafios mencionados na Seção 3.1.4 sobre o processamento desses grandes volumes de dados, há outros importantes desafios no estabelecimento desses processos. Esses processos são comumente modelados como workflows [Goderis et al., 2006]. Nestes workflows, as atividades e dados estão direcionados a execução em algum ambiente de PAD (clusters, computação em nuvem) [Oliveira et al., 2010, Ogasawara et al., 2013], onde ocorre a decomposição destes workflows em dataflows e, tem-se a alocação destes dataflows e seus respectivos dados aos recursos computacionais.

Em função da diversidade de plataformas existentes, um dos grandes desafios é estabelecer uma representação deste workflow que seja agnóstica ao meio em que será executado e, ao mesmo tempo, possibilite a otimização de sua execução no ambiente alvo, considerando-se os aspectos mencionados nas Seções 3.1.3 e 3.1.4. Diante deste cenário repleto de desafios para a execução de workflows de análise em larga-escala, tem-se oportunidades para explorar técnicas e métodos de acesso para grande volumes de dados, as quais possam lidar, principalmente, com os problemas relacionados com o particionamento, distribuição, movimentação e sumarização de dados presentes nos experimentos. Novamente, essas técnicas e métodos dependem da plataforma na qual os dados são processados.

Um desafio neste tópico é investigar métodos para lidar com replicas parciais do banco de dados, ao mesmo tempo em que desenvolveremos métodos que tirem vantagem das infraestruturas virtualizadas de processamento em larga-escala para reduzir o tempo de acesso aos dados processados e persistidos em memória principal.

3.2.2. Técnicas de análise em grandes volumes de dados

A análise de dados propriamente dita é apoiada por um conjunto de métodos que incluem tanto os métodos tradicionais de mineração de dados — pré-processamento, classificação, predição, agrupamento, associação e visualização — quanto os métodos de modelagem computacional [Liao et al., 2012]. Esses métodos, por sua vez, são apoiados por técnicas clássicas, dentre as quais incluem-se k-means, Partitioning Around Medoids (PAM), árvores de decisão, redes neurais, Support Vector Machines (SVM) e Apriori.

A partir dessas técnicas de análise clássicas, diversos algoritmos, implementações, adaptações e variações estão presentes em ferramentas de análise de dados consolidadas, como, por exemplo, a linguagem R. O desafio consiste tanto na correta aplicação desses algoritmos, como também na implementação adequada para se atingir a escalabilidade desses algoritmos nos processamento de grandes volumes de dados. Essas aplicações devem ser encapsuladas em wrappers para fazer parte dos experimentos de análise (workflows). Nesse contexto, um aspecto muito importante para a ciência de dados, consiste em como preparar os dados para a aplicação destas técnicas. A correta aplicação das técnicas de normalização, transformação [Ogasawara et al., 2010], remoção de outliers [Gupta et al., 2014], seleção de atributos e definição de amostras, pode significar a diferença entre obter ou não conhecimento e produzir valor agregado. Outro aspecto fundamental consiste em se ter uma infraestrutura que possibilite explorar as di-

ferentes técnicas para selecionar a mais adequada para os dados trabalhados. No processo de condução da ciência de dados, isso consiste em visualizar os resultados parciais e poder ajustar parâmetros nas técnicas de análise durante a execução do experimento [Mattoso et al., 2013].

3.2.3. Análises orientadas a hipóteses

A análise de dados em larga-escala, como advogado nessa proposta de desafio, requer um processo científico no qual o cientista de dados formula hipóteses e utiliza os dados disponíveis, ou desenvolve um experimento *in-silico* para sua produção, como base para validação. Neste contexto, permitir que hipóteses científicas sejam gerenciadas computacionalmente amplia o suporte computacional ao ciclo-de-vida da ciência *in-silico*. De fato, o novo cientista de dados exige dos sistemas computacionais o mesmo aparato encontrado pela ciência tradicional em seus laboratórios. Por um lado, dados observacionais são capturado de forma digital e formam a base para formulação de hipóteses sobre os fenômenos sendo investigados. Por outro lado, sistemas de workflows fornecem o ferramental para levar a cabo a validação das hipóteses seja através de simulações computacionais, cujos resultados são confrontados com as observações, seja na determinação de padrões nos dados que estatisticamente estabeleçam relações causais propostas pela hipótese científica.

Neste sentido, vários trabalhos têm sido propostos para apoio na criação e validação de hipóteses de forma automática [Schmidt e Lipson, 2009]. A ferramenta Eureqa[8] obteve grande sucesso, produzindo através de um algoritmo de programação dinâmica expressões matemáticas que modelam um conjunto de dados fornecido, sem de fato apresentar semântica física correspondente.

No contexto de pesquisas orientadas à formulação de hipóteses que reflitam a interpretação do fenômeno observado, o problema é abordado de maneira top-down. Um modelo teórico inicial é proposto, possivelmente como um conjunto de equações matemáticas. Utilizando-se de métodos numéricos, deriva-se sua representação equivalente em forma computacional. A avaliação do modelo computacional produz os dados a partir dos quais as hipóteses podem ser avaliadas. Em [Haas, 2014], um sistema de banco de dados para apoio ao cálculo de simulações numéricas é apresentado. O sistema permite que atualizações e progressos na análise de hipóteses sejam realizados inteiramente baseadas em dados. A partir de sua geração, dados de simulação computacional passam a estar disponíveis para sua avaliação analítica. Em [Gonçalves e Porto, 2014], estes são carregados em um sistema de bancos de dados probabilísticos onde a característica de incerteza dos modelos associados é inferida. O cálculo da incerteza associada aos dados de simulação permite a adoção de modelos Bayesianos, que através do cálculo de probabilidade condicional permite integrarmos dados observacionais com dados simulados, a priori.

A ciência de dados orientada a hipóteses pode tomar, igualmente, uma abordagem menos teórica. Em sistemas de recomendação, por exemplo, nos quais ferramentas de busca na web fazem sugestões automaticamente a seus usuários, encontra-se um modelo

[8] `http://www.nutonian.com/products/eureqa`

ad-hoc relevante de formulações de hipóteses e de sua avaliação sobre a grande quantidade de dados disponíveis. Neste contexto, pode-se avaliar e valorar milhões de hipóteses simultaneamente, ordenando-as e escolhendo a que melhor reflita a natureza dos dados.

A expressão e validação de hipóteses científicas *in-silico* está na base da realização da ciência em dados. Seja através de expressões formais matemáticas ou de determinação de padrões em dados, ela é fundamental no avanço desta nova disciplina. A compreensão e adoção do métodos baseados em hipóteses *in-silico* está, no entanto, na sua infância. A compreensão das diferentes facetas de sua representação e de modelos mais ou menos formais influencia o processo científico e precisa ser melhor compreendido. Do ponto de vista do suporte computacional, modelos, técnicas e algoritmos devem ser concebidos para abrigar a ciência em dados vista desta forma. Neste contexto, há um claro desafio relacionado a se aprofundar a compreensão sobre essas diversas facetas do processo científico *in-silico*, ampliando o suporte computacional à ciência praticada em dados.

3.3. Análise de redes complexas

Habitamos um mundo extremamente conectado e esta conectividade em permanente crescimento impacta nossas vidas de maneiras que ainda não compreendemos totalmente [Vespignani, 2009]. Simultaneamente, novas ferramentas, métodos e tecnologias nos permitem atualmente um potencial sem precedentes de extrair conhecimento de enormes volumes de dados de alguma maneira interconectados em diversos campos da ciência, de redes sociais a redes biológicas.

No campo bastante ativo de redes complexas, mais recentemente também chamado de *ciência de redes* [Kocarev e In, 2010], o desenvolvimento de um grande conjunto de atividades de pesquisa nos últimos 10-15 anos foi muito estimulado pela disponibilidade crescente de dados empíricos e o aumento correspondente na capacidade computacional para analisar tais dados. Isso permitiu a percepção de similaridades nas estruturas de redes oriundas de áreas bastante distintas, o desenvolvimento de uma série de ferramentas e métodos para caracterizar e modelar tais redes, bem como o entendimento do impacto da estrutura dessas redes nos processos dinâmicos que ocorrem nessas redes. Esse desenvolvimento acelerado da disponibilidade de dados e aplicações imediatas com base nesses colabora para a atual demanda por pesquisa básica nos aspectos fundamentais de análise de redes complexas.

Nesse contexto, diferentes sistemas de grande porte, tanto naturais quanto artificiais, com elementos diversos interconectados podem ser representados por meio de redes complexas de larga-escala [Albert e Barabási, 2002, Newman, 2003] (ver Seção 3.1.1, onde discute-se a possibilidade de representação por grafos). A adoção dessa abordagem baseada em redes complexas para modelagem atualmente impacta, não somente áreas científicas, mas também diversos setores do governo ou da indústria como base para aplicações estratégicas em definição de políticas públicas ou sistemas de recomendação, respectivamente, apenas para nomear algumas aplicações. Podemos identificar duas frentes relacionadas com a caracterização, análise e modelagem de redes complexas, uma é lidar com a dimensão que as mesmas tipicamente tem adquirido em diversas áreas e outra frente é lidar com o dinamismo das redes complexas atuais, sendo que este último pode estar associado a redes complexas dinâmicas com larga-escala. Discutimos essas frentes a seguir.

3.4. Redes complexas de larga-escala

A caracterização, análise e modelagem de redes complexas de larga-escala [Albert e Barabási, 2002, Rosvall e Bergstrom, 2010] é um desafio chave no domínio de ciência de dados, dada a vasta presença maciça e escala das redes complexas com as quais várias áreas do conhecimento lidam atualmente. Um dos desafios nesse tema é investigar algoritmos, métodos ou técnicas que permitam a análise de características globais de redes complexas, muitas vezes muito custosas computacionalmente ou mesmo intratáveis dependendo da escala, por métricas locais, viabilizando a análise dessas redes, mesmo que aproximadamente [Everett e Borgatti, 2005, Wehmuth e Ziviani, 2013]. Esse contexto se aplica a áreas científicas diversas, bem como sistemas naturais ou artificiais modelados por redes complexas presentes em desafios na ciência, indústria e também no setor governamental. Lidar com redes complexas de larga-escala, mesmo estáticas, impõe desafios em todas as subáreas de pesquisa relacionadas à gestão e análise de dados (em interseção com as outras duas linhas de pesquisa consideradas nesta proposta e discutidas nas Seções 3.1 e 3.2).

3.5. Dinamismo em redes complexas de larga-escala

O principal desafio na linha de pesquisa de análise de redes complexas reside no estabelecimento de fundações sólidas para a caracterização, análise e modelagem de redes complexas dinâmicas de larga-escala. Redes complexas dinâmicas podem apresentar um dinamismo espaço-temporal. Esse dinamismo pode ser variante no tempo (i.e., arestas e nós variam ao longo do tempo, podendo ser representados por grafos variantes no tempo [Holme e Saramäki, 2012, Wehmuth et al., 2014b]; variante no espaço, onde múltiplas redes interdependentes podem ser associadas em camadas (podendo ser representadas por redes multicamadas [Kurant e Thiran, 2006, De Domenico et al., 2013]; ou mesmo ambos [Kivelä et al., 2014, Wehmuth et al., 2014a].

Dado esse dinamismo das redes complexas que emergem em diferentes cenários de aplicação, tais como os descritos na Seção 2, esse é um desafio chave para o avanço de aspectos fundamentais na área de ciência de dados ao lidar com sistemas naturais ou artificiais modelados por redes complexas de larga-escala, sobretudo dinâmicas. Esse desafio se projeta tanto na caracterização, análise e modelagem da dinâmica da estrutura dessas redes complexas, mas também na caracterização, análise e modelagem dos processos dinâmicos que ocorram sobre essas redes complexas. A análise se torna ainda mais desafiadora em cenários combinados onde se requer a análise de processos dinâmicos em execução sobre estruturas de redes dinâmicas, requerendo portanto a caracterização, análise de modelagem *de* e *em* redes complexas dinâmicas de larga-escala. Exemplos de processos dinâmicos são a difusão de informação, identificação de comunidades de interesse, partição dos grafos, ou detecção de anomalias.

Associado à investigação dos aspectos fundamentais de análise do dinamismo de e em redes complexas de larga-escala, também torna-se um desafio a pesquisa aplicada para o desenvolvimento de métodos, técnicas e ferramentas que sejam de relevância prática em cenários de aplicação reais, tais como os discutidos na Seção 2.

4. Formação de recursos humanos

A formação de recursos humanos é um dos grandes desafios mais importantes para o avanço da área de ciência de dados no Brasil. A pesquisa básica e aplicada envolve tipica-

mente o desenvolvimento de técnicas, metodologias, modelos, algoritmos e arquiteturas em ciência de dados. O perfil profissional de cientista de dados possui demanda crescente no eixo ciência-indústria-governo [Davenport e Patil, 2012]. Esse profissional tem uma expectativa de formação tipicamente sólida em ciência da computação e aplicações, modelagem, estatística, analítica e matemática, além do conhecimento mínimo do domínio de aplicação, dada sua atuação intrinsecamente interdisciplinar. Assim, o novo profissional em ciência de dados reúne um conjunto de competências interdisciplinares dificilmente encontradas no profissional formado pelos cursos verticais atualmente oferecidos nas universidades. No Brasil, começam a aparecer alguns poucos cursos curtos de especialização, mas sua estruturação em cursos latu-sensu, de graduação e pós-graduação ainda é incipiente, se tanto.

Há, portanto, o grande desafio de formação de recursos humanos altamente qualificados em pesquisa básica e aplicada na fronteira do conhecimento em ciência de dados.

5. Considerações finais

O avanço tecnológico das últimas décadas culminou com a capacidade de obtenção e geração de imensos volumes de dados, tanto de fenômenos naturais quanto de sistemas artificiais. O novo cenário delineado nesse contexto abre em realidade novas necessidades, perspectivas e oportunidades de avanços tecnológicos relacionados ao desenvolvimento de técnicas, metodologias, modelos, algoritmos e arquiteturas para se fazer frente ao desafio de analisar e interpretar esses imensos volumes de dados que emergem em aplicações de diversas áreas do conhecimento. Há, portanto, um grande potencial tecnológico na pesquisa básica e aplicada em ciência de dados, tal como aqui discutido, dado o foco nos aspectos fundamentais da análise de dados em larga-escala com impacto em diferentes áreas de conhecimento básico bem como cenários de aplicação.

Ciência de dados é uma área recente tanto no Brasil quanto no exterior. Há, entretanto, já algumas iniciativas recentes em instituições de ponta no exterior que focam em ciência de dados. Alguns exemplos são o Data Science Institute[9] no Imperial College, o Institute for Data Sciences & Engineering[10] na Columbia University, o Berkeley Institute for Data Science (BIDS)[11] na UC Berkeley, o Center for Data Science (CDS)[12] da New York University ou a iniciativa de recrutamento expressivo em Data Science a partir de junho/2014 na Boston University.[13] É, portanto, relevante concentrar esforços para enfrentar o grande desafio de ciência de dados em nosso país, contribuindo para preencher a lacuna existente atualmente no Brasil nesta área.

Em suma, é necessário posicionar o Brasil na direção da nova ciência baseada em dados, enfrentando os desafios de pesquisa básica e aplicada em ciência de dados, preparando recursos humanos altamente qualificados na área, de forma a desenvolver o alicerce para a projeção do país de forma relevante e em bases sólidas na sociedade do conhecimento.

[9] http://www3.imperial.ac.uk/data-science

[10] http://idse.columbia.edu

[11] http://vcresearch.berkeley.edu/datascience

[12] http://cds.nyu.edu

[13] http://www.bu.edu/provost/2014/06/23/university-provosts-faculty-hiring-initiative-in-data-science

Agradecimentos

Os autores agradecem FAPERJ, CNPq, FINEP e MCTI pelo apoio. Os autores agradecem a Eduardo Ogasawara (CEFET-RJ) por sua contribuição na parte de análise de dados.

Referências

Albert, R. e Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Barabási, A.-L., Gulbahce, N., e Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., e Volinsky, C. (2013). Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82.

Bell, G., Hey, T., e Szalay, A. (2009). Beyond the Data Deluge. *Science*, 323:1298–1298.

Chen, T. M. (2001). Increasing the observability of internet behavior. *Communications of the ACM*, 44(1):93–98.

Correa, B. S., Gonçalves, B., Teixeira, I. M., Gomes, A. T., e Ziviani, A. (2011). Atoms: a ubiquitous teleconsultation system for supporting ami patients with prehospital thrombolysis. *International journal of telemedicine and applications*, 2011:2.

Costa, R. G., Porto, F., e Schulze, B. (2012). Towards analytical data management for numerical simulations. In *Proceedings of the 6th Alberto Mendelzon International Workshop on Foundations of Data Management*, pages 210–214.

Curino, C., Jones, E., Zhang, Y., e Madden, S. (2010). Schism: a workload-driven approach to database replication and partitioning. *Proceedings of the VLDB Endowment*, 3(1-2):48–57.

Davenport, T. H. e Patil, D. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*.

De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., Gómez, S., e Arenas, A. (2013). Mathematical formulation of multilayer networks. *Physical Review X*, 3(4):041022.

Dean, J. e Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.

Eagle, N., Pentland, A. S., e Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278.

Estrin, D. (2014). small data, where n = me. *Communications of the ACM*, 57(4):32–34.

Everett, M. e Borgatti, S. P. (2005). Ego network betweenness. *Social networks*, 27(1):31–38.

Freire, E. P., Ziviani, A., e Salles, R. M. (2008). Detecting voip calls hidden in web traffic. *Network and Service Management, IEEE Transactions on*, 5(4):204–214.

Freire, V. P., Macedo, J. A. F., e Porto, F. (2014). NACluster: A non-supervised clustering algorithm for matching multi catalogs. In *The e-Science Workshop for Work In Progress, IEEE International Conference on e-Science*.

Gadelha Jr., L. M., Wilde, M., Mattoso, M., e Foster, I. (2012a). MTCProv: a practical provenance query framework for many-task scientific computing. *Distributed and Parallel Databases*, 30(5-6):351–370.

Gadelha Jr., L. M. R., Stanzani, S., Correa, P., Dalcin, E., Gomes, C. R. O., Sato, L., e Siqueira, M. (2012b). Scalable and provenance—enabled scientific workflows for predicting distribution of species. In *Proc. 8th International Conference on Ecological Informatics (ISEI 2012)*, Brasília, DF.

Goderis, A., Li, P., e Goble, C. (2006). Workflow discovery: the problem, a case study from e-science and a graph-based solution. In *International Conference on Web Services (ICWS)*, pages 312–319. IEEE.

Gomes, A. T. A., Ziviani, A., Correa, B. S. P. M., Teixeira, I. M., e Moreira, V. M. (2012). SPLiCE: a software product line for healthcare. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 721–726. ACM.

Gonçalves, B. e Porto, F. (2014). Υ-DB: Managing scientific hypotheses as uncertain data. In *Proc. of the Very Large Data Bases (VLDB)*.

Gueye, B., Ziviani, A., Crovella, M., e Fdida, S. (2006). Constraint-based geolocation of internet hosts. *Networking, IEEE/ACM Transactions on*, 14(6):1219–1232.

Gupta, M., Gao, J., Aggarwal, C., e Han, J. (2014). Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250–2267.

Haas, P. J. (2014). Model-data ecosystems: challenges, tools, and trends. In *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 76–87. ACM.

Han, J., Kamber, M., e Pei, J. (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Holme, P. e Saramäki, J. (2012). Temporal networks. *Physics reports*, 519(3):97–125.

Hueske, F., Peters, M., Sax, M. J., Rheinländer, A., Bergmann, R., Krettek, A., e Tzoumas, K. (2012). Opening the black boxes in data flow optimization. *Proceedings of the VLDB Endowment*, 5(11):1256–1267.

Iqbal, M. S., Choudhury, C. F., Wang, P., e González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, 52(8):36.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., e Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., e Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, 2(3):203–271.

Kocarev, L. e In, V. (2010). Network science: A new paradigm shift. *IEEE Network*, 24(6):6–9.

Kurant, M. e Thiran, P. (2006). Layered complex networks. *Physical review letters*, 96(13):138701.

Las-Casas, P. H., Guedes, D., Almeida, J. M., Ziviani, A., e Marques-Neto, H. T. (2013). Spades: Detecting spammers at the source network. *Computer Networks*, 57(2):526–539.

Lazer, D., Kennedy, R., King, G., e Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, 343(6176):1203–5.

Liao, S.-H., Chu, P.-H., e Hsiao, P.-Y. (2012). Data mining techniques and applications—a decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12):11303–11311.

Mattoso, M., Ocaña, K., Horta, F., Dias, J., Ogasawara, E., Silva, V., de Oliveira, D., Costa, F., e Araújo, I. (2013). User-steering of HPC workflows: State-of-the-art and future directions. In *Proceedings of the 2nd ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies*, page 4. ACM.

Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., Kelling, S., Koskela, R., Tenopir, C., e Vieglais, D. A. (2012). Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, 11:5–15.

Mohan, C. (2013). History repeats itself: sensible and nonsensql aspects of the nosql hoopla. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 11–16. ACM.

Nambiar, R., Bhardwaj, R., Sethi, A., e Vargheese, R. (2013). A look at challenges and opportunities of big data analytics in healthcare. In *IEEE International Conference on Big Data*, pages 17–22.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Ogasawara, E., Dias, J., Oliveira, D., Porto, F., Valduriez, P., e Mattoso, M. (2011). An algebraic approach for data-centric scientific workflows. *Proceedings of the VLDB Endowment*, 4(12):1328–1339.

Ogasawara, E., Dias, J., Silva, V., Chirigati, F., Oliveira, D., Porto, F., Valduriez, P., e Mattoso, M. (2013). Chiron: a parallel engine for algebraic scientific workflows. *Concurrency and Computation: Practice and Experience*, 25(16):2327–2341.

Ogasawara, E., Martinez, L. C., de Oliveira, D., Zimbrão, G., Pappa, G. L., e Mattoso, M. (2010). Adaptive normalization: A novel data normalization approach for non-stationary time series. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Oliveira, D., Ogasawara, E., Baião, F., e Mattoso, M. (2010). Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows. In *IEEE International Conference on Cloud Computing*, pages 378–385. IEEE.

Özsu, M. T. e Valduriez, P. (2011). *Principles of distributed database systems*. Springer.

Porto, F., Moura, A. M., Silva, F. C., Bassini, A., Palazzi, D. C., Poltosi, M., Castro, L. E. V., e Cameron, L. (2012). A metaphoric trajectory data warehouse for olympic athlete follow-up. *Concurrency and Computation: Practice and Experience*, 24(13):1497–1512.

Pretz, K. (2014). Better health care through data. Tech Focus, The Institute, IEEE.

Rosvall, M. e Bergstrom, C. T. (2010). Mapping change in large networks. *PloS one*, 5(1):e8694.

Schmidt, M. e Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85.

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., e Vangenot, C. (2008). A conceptual view on trajectories. *Data & knowledge engineering*, 65(1):126–146.

Srivastava, D. (2012). Towards analytical data management for numerical simulations. In *Proceedings of the 6th Alberto Mendelzon International Workshop on Foundations of Data Management*.

Stevens, R., Zhao, J., e Goble, C. (2007). Using provenance to manage knowledge of in silico experiments. *Briefings in bioinformatics*, 8(3):183–194.

Suciu, D., Olteanu, D., Ré, C., e Koch, C. (2011). Probabilistic databases. *Synthesis Lectures on Data Management*, 3(2):1–180.

Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, 325(5939):425.

Wehmuth, K., Fleury, É., e Ziviani, A. (2014a). On multiaspect graphs. *arXiv preprint arXiv:1408.0943*.

Wehmuth, K. e Ziviani, A. (2013). DACCER: Distributed assessment of the closeness centrality ranking in complex networks. *Computer Networks*, 57(13):2536–2548.

Wehmuth, K., Ziviani, A., e Fleury, E. (2014b). A Unifying Model for Representing Time-Varying Graphs. Technical Report RR-8466, INRIA.

Wright, A. (2014). Big data meets big science. *Communications of the ACM*, 57(7):13–15.

Xavier, F. H. Z., Silveira, L., Almeida, J., Malab, C., Ziviani, A., e Marques-Neto, H. T. (2013). Understanding human mobility due to large-scale events. In *3rd Conference on the Analysis of Mobile Phone Datasets (NetMob)*.

Xavier, F. H. Z., Silveira, L. M., Almeida, J. M. d., Ziviani, A., Malab, C. H. S., e Marques-Neto, H. T. (2012). Analyzing the workload dynamics of a mobile phone network in large scale events. In *Proceedings of the first workshop on Urban networking (URBANE), ACM CoNEXT*, pages 37–42. ACM.

Ziviani, A., Cardozo, T. B., e Gomes, A. T. A. (2012). Rapid prototyping of active measurement tools. *Computer Networks*, 56(2):870–883.