Da Ciência à eCiência: paradigmas da descoberta do conhecimento

Daniel Cordeiro, Kelly R. Braghetto, Alfredo Goldman e Fabio Kon
Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

1. Introdução

Ao longo dos últimos 3 milênios, o conceito que a sociedade faz de ciência e o modo de se fazer ciência vêm evoluindo, com mudanças significativas acontecendo no decorrer dos séculos. Como tudo o que envolve a evolução da humanidade, a velocidade com que as mudanças ocorriam nos milênios passados era bem menor, foi aumentando incrivelmente no século passado e continua a aumentar no novo milênio. Da evolução do pensamento dos filósofos pré-Socráticos no século VI a.C. até os experimentos científicos de Arquimedes no século III a.C. passaram-se mais de trezentos anos. No último milênio, a velocidade só aumentou. Do trabalho multifacetado artístico-científico-pragmático de Leonardo da Vinci, um homem da renascença, no século XV, aos trabalhos com Matemática, Física e Astronomia de Galileu Galilei, o pai da ciência moderna, no século XVII, até o trabalho especializado de Alan Turing, o pai da ciência da computação, no século XX, observamos grandes mudanças. O ato de fazer ciência passou por significativos aprimoramentos e refinamentos em sua metodologia de trabalho, incluindo novo ferramental lógico-matemático, novos instrumentos de observação do mundo e novos paradigmas de estruturação do pensamento científico (Kuhn, 1962).

Segundo Jim Gray (Hey et al., 2009), a ciência nasceu há milhares de anos de forma **empírica**, descrevendo fenômenos naturais. Nos últimos séculos, ela passou a incorporar uma importante componente **teórica**, utilizando modelos e generalizações. Nas últimas décadas, surgiu uma forte tendência **computacional**, com a possibilidade de realização de sofisticadas simulações de fenômenos complexos. Nos últimos anos, estamos observando o aparecimento de um **quarto paradigma**, a exploração de grandes quantidades de dados, muitas vezes chamado de ***eScience*** (eCiência), que unifica teoria, experimentos e simulação, ao mesmo tempo em que lida com uma quantidade enorme de informação.

Nesta nova forma de se fazer ciência, trilhões de bytes de dados são capturados por instrumentos ou gerados via simulação. O acelerador de partículas *Large Hadron Collider* (LHC) da Organização Europeia para a Pesquisa Nuclear (CERN) captura 25 petabytes (quatrilhões) de dados todos os anos. O sequenciamento do genoma de um único ser humano requer o armazenamento de 4 gigabytes (bilhões) de caracteres. É impossível processar essa enorme quantidade de dados “manualmente”, o processamento precisa ser obrigatoriamente feito por software. A informação precisa ser obrigatoriamente armazenada em grandes bancos

de dados. A análise desses dados precisa obrigatoriamente utilizar um ferramental estatístico avançado, codificado na forma de programas de computador que consomem, filtram, manipulam, transformam e consolidam esses dados com o objetivo de extrair alguma informação relevante.

De fato, a ciência da computação de uma forma geral e, mais especificamente, o software passou a ser um componente central da ciência do século XXI. Salvo raras exceções, não se faz boas ciências exatas e biomédicas hoje em dia sem bons desenvolvedores de software na equipe de trabalho. Até nas ciências humanas, as ferramentas computacionais estão começando a ser utilizadas mais fortemente (Meyer, 2013).

Até meados da década de 1990, a computação de alto desempenho era realizada em supercomputadores que possuíam uma arquitetura especial, memória especializada e barramentos de comunicação de alta velocidade ao custo de milhões de dólares. Na segunda metade daquela década tornou-se frequente a utilização de aglomerados (*clusters*) de dezenas ou centenas de PCs convencionais trabalhando em conjunto para a solução de um único problema científico. Na virada do século, surgiu a ideia de interconectar vários desses aglomerados através da Internet, formando as grades computacionais. Elas podem agregar milhares de computadores compartilhados por cientistas de diversas instituições, permitindo a realização de seus experimentos computacionais e processamento de grandes quantidades de dados.

Nos últimos anos, no entanto, a tecnologia das grades, associada a mecanismos de virtualização[1], evoluiu para o modelo de computação em nuvem (Zhang et al., 2010). Uma nuvem oferece uma interface simples por meio da qual usuários podem obter máquinas virtuais para a execução de suas tarefas computacionais. As nuvens são, atualmente, amplamente usadas por empresas, para comércio eletrônico, por governos, para a execução de suas tarefas corriqueiras, e por uma infinidade de serviços disponíveis na Internet. Essa tecnologia permite o compartilhamento mais racional do hardware, baixando os custos, simplificando os processos e diminuindo o impacto ambiental (e.g., consumo de energia). Por exemplo, em vez de 50 grupos de pesquisa da Universidade de São Paulo comprarem, instalarem, configurarem e manterem 50 aglomerados distintos, que possivelmente ficariam ociosos boa parte do tempo, agora torna-se possível que todos compartilhem uma única infraestrutura de nuvem, oferecendo muito mais poder computacional a um menor custo e impacto.

No entanto, as peculiaridades do tipo de computação que a ciência contemporânea exige ainda fazem com que as nuvens sejam pouco utilizadas pelos cientistas. As nuvens atuais não foram projetadas para lidar com o processamento de enormes quantidades de dados da forma como aplicações científicas fazem. O acesso compartilhado a esses dados por centenas de cientistas espalhados em diferentes partes do globo também não é fácil; com o recente movimento no sentido da Ciência Aberta[2], esse tipo de acesso universal e aberto se torna cada vez mais desejado. As interfaces de programação e de administração das máquinas utilizadas nas nuvens ainda são desconhecidas pela grande maioria dos grupos de pesquisa

[1] Virtualização é uma técnica computacional que permite que uma grande quantidade de "máquinas virtuais" sejam criadas dentro de um único computador. Dessa forma, os usuários finais (cientistas) podem multiplicar a quantidade de máquinas utilizadas para o processamento de seus experimentos e compartilhar de forma segura o mesmo conjunto de máquinas físicas.

[2] Veja o verbete *Open Science* na Wikipédia: http://en.wikipedia.org/wiki/Open_science

que, se utilizassem a nuvem, passariam a depender de especialistas em TI para executar tarefas corriqueiras do dia a dia.

Portanto, ainda são necessárias pesquisas em ciência da computação para tornar a tecnologia de nuvem mais apropriada à natureza específica da computação científica e mais facilmente acessível a grupos de pesquisa de outras áreas da ciência. É necessário também a educação da nova geração de biólogos, físicos, químicos, médicos, cientistas sociais, linguistas, etc., no ferramental básico de trabalho dentro desse novo modelo.

Este texto apresentará uma breve descrição da evolução dos paradigmas do modo de se fazer ciência (do empirismo ao panorama atual da *eScience*) e abordará o potencial da computação em nuvem como ferramenta catalisadora de pesquisa transformativa.

2. Evolução da Ciência

Mudanças significativas no processo de criação de conhecimento são raras, mas a história mostra que elas ocorrem tanto quando há a criação de uma nova ferramenta conceitual que se mostra fundamental (por exemplo, o Cálculo Diferencial e Integral de Newton), como quando uma nova ferramenta tecnológica (por exemplo, o microscópio eletrônico) permite a criação de novos "tipos" de ciência.

Computadores desempenham um papel cada vez mais importante no processo científico. Nos últimos 50 anos, o uso de ferramentas computacionais mudou a forma como a ciência é feita em áreas diversas como meteorologia e climatologia, mecânica dos fluidos, astrofísica, química, etc. Não só o uso de computadores permitiu o desenvolvimento de novos tipos de pesquisas nessas áreas, como os novos desafios trazidos por elas promoveram grandes avanços na teoria e prática da ciência da computação. As novas necessidades impostas pela ciência estão promovendo uma mudança importante no papel que a ciência da computação exerce sobre as outras áreas da ciência. Gradualmente, a computação está deixando de ser apenas uma "ferramenta de apoio" a novas pesquisas para se tornar uma parte fundamental das outras áreas com que interage e de seus métodos científicos.

Jim Gray, vencedor do Prêmio Turing de 1998, o "Nobel da ciência da computação", e um dos pioneiros em aplicações de técnicas computacionais para o tratamento de grandes quantidades de dados gerados por cientistas de outras áreas, via nessa mudança uma verdadeira transformação no fazer ciência. Em sua opinião, estaríamos vivendo o início de um quarto paradigma que redefinirá a metodologia científica de diversas áreas do conhecimento (Hey et al., 2009).

O primeiro paradigma da metodologia científica teria sido o empirismo. Há milhares de anos, o processo de descoberta era feito somente a partir de experimentos. Todo o conhecimento acerca dos fenômenos naturais era baseado e adquirido unicamente por meio do que se podia apreender pelos sentidos.

A primeira quebra de paradigma teria ocorrido há algumas centenas de anos, com o surgimento das primeiras tentativas de se explicar fenômenos por meio de modelos teóricos. Modelos como as Leis de Kepler, as Leis de Netwon ou a Lei de Boyle-Mariotte (dentre tantas

outras) permitiram não só um melhor entendimento dos fenômenos observados empiricamente, como também a realização de previsões sobre o comportamento de novos fenômenos.

Com o passar do tempo, cientistas criaram modelos grandes e complexos demais para serem resolvidos de forma puramente analítica. Os cientistas passaram a utilizar simulações que, tipicamente, avaliam a evolução de um fenômeno em função do tempo utilizando o modelo desenvolvido. A complexidade das simulações impulsionou o desenvolvimento da computação científica.

O ENIAC, um dos primeiros computadores digitais eletrônicos de uso geral, começou a ser desenvolvido em 1943 (durante a Segunda Guerra Mundial) para a realização de simulações de modelos balísticos. A partir de 1946, quando a existência do ENIAC foi anunciada ao mundo, ele passou a ser utilizado para a realização de simulações em diferentes áreas do conhecimento. Foi utilizado, por exemplo, no desenvolvimento de túneis de vento, na análise de números aleatórios, em cálculos de energia atômica e em aplicações de previsão meteorológica. Desde então, muitas pesquisas científicas foram validadas com base não apenas em dados obtidos por meio de observações experimentais, mas também com base em resultados de simulações numéricas.

O desenvolvimento de computadores cada vez mais rápidos, juntamente com os avanços feitos pela ciência da computação nas áreas de computação paralela e distribuída, permitiram que diversas ciências começassem a realizar simulações numéricas em escalas cada vez maiores. Isso possibilitou a simulação de modelos cada vez mais complexos, que analisam e produzem uma grande quantidade de dados e que requerem muito poder computacional para sua execução. Ao mesmo tempo em que a computação evoluiu, as ciências experimentais também evoluíram e passaram a ser capazes de coletar uma quantidade maior de dados. Jim Gray dizia que, atualmente, os astrônomos não olham mais através de seus telescópios. Ao invés disso, eles “olham” através de instrumentos complexos que estão conectados a centrais de processamento de dados e, só então, utilizam seus computadores para visualizar as informações coletadas.

Em muitas áreas da ciência, principalmente nas ciências naturais, as novas tecnologias criaram novas possibilidades (ou “tipos”) de pesquisa. Criou-se uma nova metodologia de pesquisa em que dados experimentais, coletados por meio de instrumentos ou gerados por simulação, são processados por sistemas de software complexos e só então a informação (ou o conhecimento) resultante é armazenada em computadores. Os cientistas só analisam os dados no final do processo. Trata-se, de fato, de uma mudança importante no processo de pensamento científico, que está substituindo o processo de “formulação de hipótese → experimentação → análise de resultados” por “formulação de hipótese → busca da resposta no banco de dados” (Emmott et al., 2006). Esse novo processo científico, baseado no processamento e análise de grandes quantidades de dados, requer tecnologias e metodologias tão distintas que diversos cientistas, ao lado de Gray, dizem estarmos presenciando o início de um novo (o quarto) paradigma de exploração científica.

A ciência da computação exerce um papel fundamental nesse novo processo científico. As técnicas desenvolvidas nas últimas décadas permitem que essas grandes quantidades de dados sejam processadas por algoritmos eficientes, capazes de explorar o grande poder computacional fornecido por soluções modernas como as plataformas de computação em nuvem. Mas mais do que uma ferramenta operacional para as outras ciências, os conceitos e

teorias da ciência da computação já são parte intrínseca de pesquisas em outras áreas do conhecimento.

A primeira prova da importância dos conceitos da Computação como parte integrante de uma outra ciência foi mostrada no Projeto Genoma Humano[3]. A escolha apropriada das abstrações matemáticas que representam os elementos da pesquisa – tais como a representação das sequências de DNA como um *string* (ou seja, uma sequência finita de símbolos) ou a representação da estrutura tridimensional das proteínas como um grafo rotulado – permitiu o uso de teorias sofisticadas que garantiram o processamento eficiente e a disponibilização de grandes bancos de dados de sequências de DNA digitalizados para cientistas de diversas partes do mundo. A codificação de conhecimento científico cria uma nova metodologia científica, na qual o conhecimento pode ser analisado computacionalmente no mundo virtual mesmo antes que qualquer experimento seja realizado no mundo real. Mais ainda, a codificação promove um processo de pesquisa ainda mais colaborativo, no qual novos modelos podem ser facilmente avaliados e testados usando outros modelos e dados disponibilizados por outros cientistas, interconectados pela Internet.

Além das abstrações, muitas outras ferramentas e teorias da computação podem ser aplicadas a outras ciências. A computação estuda há vários anos diversas noções de complexidade. A complexidade de Kolmogorov, por exemplo, avalia qual o menor programa que pode produzir um determinado *string* e poderia ser usado no estudo de árvores filogenéticas na biologia (Emmott et al., 2006, p. 27). Outro exemplo interessante é o uso de modelos algébricos de computação paralela para descrever conceitos como concorrência, indeterminismo, comunicação, sincronização, troca de mensagens, etc. Esses modelos estão sendo usados para entender melhor os processos biológicos inter e intracelulares (Cardelli, 2005).

3. *eScience* hoje

A ciência é hoje, mais do que nunca, uma atividade colaborativa. Um exemplo disso é o *Worldwide LHC Computing Grid* (WLCG)[4], projeto que integra grades computacionais de mais de 200 centros em 36 países, com o objetivo de prover os recursos computacionais necessários para armazenar, distribuir e analisar os dados gerados pelo *Large Hadron Collider* (LHC). O LHC é o maior acelerador de partículas existente no mundo e gera todos os anos cerca de 25 petabytes (aproximadamente 25 quatrilhões de bytes) de dados. Para dimensionar esse volume de dados, considere que, se estivessem armazenados em DVDs comuns, os dados ocupariam mais de 220 mil DVDs. A manutenção desse volume gigantesco de dados não seria possível não fosse a rede de colaboração mantida pelo WLCG. Os dados resultantes dos experimentos conduzidos no LHC são distribuídos a centenas de centros computacionais de instituições de pesquisa espalhadas pelo mundo. Esses centros, por sua vez, processam os dados e os disponibilizam a uma comunidade de mais de 8 mil físicos.

---

[3] *The Human Genome Project*: http://www.genome.gov/10001772.

[4] Projeto *Worldwide LHC Computing Grid*: http://wlcg.web.cern.ch/.

Há poucos anos, era difícil de se imaginar que ambientes para a colaboração científica em uma escala global (como o provido pelo WLCG) seriam exequíveis. Mas a rápida evolução das redes de computadores de abrangência local e global, o aumento da capacidade de armazenamento, processamento e transmissão de dados, e o barateamento dos equipamentos eletrônicos impulsionaram a criação de plataformas computacionais de alto desempenho que hoje são usadas para amparar o desenvolvimento da ciência.

De nada adiantaria toda essa infraestrutura computacional para amparar os processos científicos se não houvesse programas de computador que escondessem dos cientistas a complexidade envolvida no uso desses ambientes. George Johnson, em um artigo escrito para o *New York Times* em 2001, constatou que hoje "*toda ciência é ciência da computação*" (Johnson, 2001). Porém, nem todo cientista precisa (ou quer!) ser um cientista da computação. Essa é a principal razão para que sistemas de software como o *Taverna*[5] e o *Pegasus*[6], gerenciadores de fluxos de trabalho (*workflows*) científicos, tenham se tornado muito populares entre físicos, químicos, biólogos e astrônomos.

Os sistemas gerenciadores de *workflows* permitem que um cientista descreva um experimento científico como um conjunto de tarefas a serem realizadas pelo computador. Esse conjunto de tarefas é o *workflow*. As tarefas comumente realizadas em um experimento se relacionam a coleta, homogeneização, filtragem e análise de dados. Um cientista define o seu *workflow* usando um modelo gráfico, de compreensão bastante intuitiva. A partir desse modelo, o sistema gerenciador de *workflows* é capaz de executar o experimento de forma automática, com pouca ou nenhuma intervenção do cientista, utilizando, para isso, a infraestrutura computacional disponível. O próprio sistema gerenciador se encarrega de traçar estratégias para o bom uso dos recursos computacionais, garantindo que os experimentos sejam executados de forma eficiente e segura. Portais da Web como o *MyExperiments* [7] complementam as funcionalidades dos sistemas gerenciadores, atuando como canais para o compartilhamento de modelos de *workflows*, estabelecendo novos meios de comunicação entre cientistas e promovendo a colaboração científica.

Além dos sistemas gerenciadores de *workflow*, que auxiliam o projeto e a execução de experimentos, existem outros programas de computador que desempenham um papel fundamental nas descobertas científicas. São programas que implementam algoritmos complexos de análise de dados, como os que realizam reconhecimento de padrões ou os que extraem modelos de predição a partir de dados históricos. Entre os programas desse tipo que são desenvolvidos na USP, tem-se, por exemplo, os que fazem a detecção automática de anomalias em imagens médicas (Rimkus et al., 2011) ou identificam a correlação entre mutações do HIV e a resistência à medicação retroviral (Cintho et al., 2012).

A computação faz mais do que amparar o desenvolvimento da ciência; ela aproxima ciência e sociedade. Não apenas pelo fato de redes de abrangência global (como a Internet) facilitarem a divulgação dos resultados científicos, mas também por possibilitar que o cidadão comum colabore na análise de dados e na realização de experimentos científicos. Exemplos disso são a *computação voluntária* e a *ciência cidadã*.

---

[5] *Taverna Workflow Management System*: http://www.taverna.org.uk/.

[6] Pegasus *Workflow Management System*: http://pegasus.isi.edu/.

[7] Projeto *MyExperiments*: http://www.myexperiment.org/.

Na computação voluntária, pessoas comuns "doam" a grandes projetos científicos a capacidade de processamento ociosa de seus computadores de uso pessoal. Um dos maiores projetos de computação distribuída da história da computação é um projeto de computação voluntária, o *SETI@home*[8], que analisa dados de radiotelescópios em busca de vida inteligente fora da Terra. O *SETI@home* foi lançado em 1999 e nos seus primeiros 10 anos de funcionamento processou mais de 160 terabytes (aproximadamente 160 trilhões de bytes) de dados com o auxílio de mais de 6 milhões de computadores voluntários (Korpela et al., 2011).

Embora o conceito moderno de ciência cidadã não tenha sido cunhado recentemente (ele existe desde o século XIX), a computação facilitou o seu uso e potencializou seus benefícios. Diferentemente do que ocorre na computação voluntária, em que as pessoas participam de forma passiva, na ciência cidadã, um voluntário colabora com projetos científicos ativamente, usando o seu próprio cérebro. Existem diversas atividades relacionadas a coleta e análise de dados científicos que não podem ser completamente automatizadas. São nessas atividades que a ajuda de cidadãos não-especialistas pode ser bem-vinda. Além disso, a inteligência e o conhecimento coletivo tem grande valia em vários domínios da ciência. A *Citizen Science Alliance* (CSA)[9] apoia, desenvolve e gerencia projetos de ciência cidadã que se amparam na Internet. Um dos projetos mantidos pela CSA é o *Ancient Lives*[10], da Universidade de Oxford, cujo objetivo é decifrar um importante conjunto de manuscritos greco-romanos encontrados próximo à cidade de Oxirrinco, no Egito, entre 1897 e 1907. Os voluntários consultam, no sítio Web do projeto, imagens de fragmentos dos manuscritos e fazem a transcrição dos caracteres identificáveis. Essas transcrições são combinadas ao conhecimento de especialistas e aos resultados de análises computacionais das imagens, agilizando assim o processo de identificação dos documentos.

Exemplos de projetos como o *Ancient Lives*, que associam a computação às ciências sociais e humanas, eram (até pouco tempo atrás) relativamente infrequentes. Entretanto, hoje, dados em grande escala (*Big Data*) estão sendo coletados por meio de dispositivos eletrônicos (como aparelhos celulares, computadores de mão, GPSs, etc.) que estão cada dia mais integrados ao cotidiano das pessoas. Esses dados são o objeto de estudo da *ciência social computacional*, que investiga fenômenos sociais por intermédio da computação e, em particular, de tecnologias avançadas de processamento de informações.

Segundo o Prof. Alex Pentland, "*a habilidade de ver os detalhes do mercado, das revoluções políticas, e ser capaz de predizê-las e controlá-las é, definitivamente, um caso de fogo de Prometeu - ela pode ser usada para o bem ou para o mal; e, assim, 'Big Data' nos conduz a tempos interessantes. Terminaremos por reinventar o que significa ter uma sociedade humana*" (Pentland, 2012). Pentland dirige o laboratório de *Human Dynamics* do MIT e é considerado um pioneiro da ciência social computacional e um dos maiores cientistas de dados do mundo (O'Reilly, 2011). Em seus projetos de pesquisa mais recentes, Pentland tem usado dados coletados a partir de equipamentos como telefones celulares para fazer o que ele chama de "mineração da realidade": a identificação de padrões humanos de comportamento individual ou coletivo. Esses padrões podem se relacionar a diferentes aspectos humanos, como a

[8] Projeto *SETI@home*: http://setiathome.berkeley.edu/.

[9] *Citizen Science Alliance:* http://www.citizensciencealliance.org/.

[10] Projeto *Ancient Lives*: http://ancientlives.org/.

comunicação e a movimentação. Padrões como esses podem ser usados, por exemplo, no rastreamento de ações terroristas ou no monitoramento preventivo do tráfego de uma cidade. As aplicações práticas desse tipo de estudo são inúmeras.

Apesar dos vários cenários bem-sucedidos que ilustram este texto, a *eScience* ainda está longe da realidade de muitas instituições de pesquisa do mundo todo. A justificativa mais frequente para esse fato é a falta de recursos físicos, humanos e financeiros para criar e manter o ambiente computacional mínimo que é um requisito para a prática da *eScience*. Entretanto, o surgimento de uma nova tecnologia renovou as esperanças da comunidade científica em tornar a *eScience* mais "acessível". Essa tecnologia é a computação em nuvem.

4. Ciência da Nuvem

Plataformas de computação em nuvem revolucionaram a indústria de tecnologia da informação ao permitir que, pela primeira vez, grandes quantidades de recursos computacionais (por exemplo, armazenamento, processamento ou aplicações) fossem oferecidos aos usuários como um serviço sob demanda. Uma nuvem abstrai uma estrutura computacional complexa, tornando-a disponível aos usuários através de interfaces simples e acessíveis por uma rede (como a Internet). Outras características essenciais de uma nuvem são: compartilhamento de recursos, elasticidade (serviços podem ser alocados e liberados rapidamente, conforme a demanda) e serviço mensurado (o que permite um uso mais eficiente dos recursos, por parte tanto dos usuários quanto dos provedores de serviços de nuvem).

Atualmente, um dos grandes desafios científicos em computação é conseguir efetivamente fazer *eScience* em nuvens computacionais. Antes, aplicações que demandavam muito processamento eram executadas em aglomerados ou em super computadores paralelos (HPC, de *high-performance computing*). Hoje, uma tendência clara é o uso da nuvem para o processamento de alto desempenho. Tanto que a Amazon EC2[11], um dos maiores provedores de serviços de nuvem, atualmente oferece uma plataforma computacional específica para HPC. Nela, é possível comprar não apenas máquinas isoladas, mas instâncias de aglomerados com as características desejadas, como, por exemplo, a presença de placas aceleradoras gráficas (GPUs) ou redes de alto desempenho.

Como uma forma de mostrar o poder computacional que uma nuvem pode prover, em 2011 foi realizado um teste de desempenho, do mesmo tipo que o usado para criar a lista dos 500 supercomputadores mais rápidos do mundo (projeto *Top500*)[12]. Para o teste, foi criada uma instância de um aglomerado com 1.064 máquinas (17.024 núcleos) do tipo *Eight Extra Large* (que é o modelo de máquina mais caro e poderoso disponível na Amazon EC2). O desempenho dessa instância foi equivalente ao de uma máquina com 240,09 TeraFLOPS[13], que ficaria na posição 42 entre os maiores supercomputadores do mundo segundo a lista

[11] *Amazon Elastic Compute Cloud* (Amazon EC2): http://aws.amazon.com/ec2/.

[12] Projeto *Top500*: http://www.top500.org/.

[13] Um TeraFLOP corresponde a um trilhão de operações de ponto flutuante por segundo.

TOP500 de novembro de 2011. Essa mesma máquina ficaria na posição 102 da lista *TOP500* de novembro de 2012.

Um dos principais interesses da *eScience* relacionado às plataformas de computação em nuvem é a possibilidade de se executar *workflows* científicos nessas plataformas. Apesar de todas as relativas facilidades que existem hoje para a alocação de aglomerados de máquinas, executar de forma eficiente um *workflow* científico em um ambiente de nuvem ainda é um desafio. A execução de um *workflow* na nuvem envolve atribuir para cada tarefa uma máquina e garantir as transferências dos dados para as tarefas sempre que necessário.

Mesmo considerando as diversas garantias oferecidas por meio de contratos de qualidade de serviço (SLAs, de *service-level agreements*), o ambiente fornecido pelas nuvens não é completamente conhecido e controlado. Por exemplo, ao se solicitar recursos computacionais na nuvem, não há garantia de que os recursos fornecidos estarão alocados em uma mesma máquina física ou em máquinas independentes. O desempenho de um *workflow* depende de como essa alocação é feita. Além disso, quando um *workflow* possui tarefas que trocam grandes quantidades de dados entre si, é desejável que essas tarefas sejam executadas em máquinas adjacentes ou próximas (para se diminuir o tempo necessário para a transmissão dos dados entre as tarefas). Nos modelos tradicionais de nuvem, não há garantia de que isso aconteça.

A execução de *workflows* em nuvens pode se beneficiar do que é hoje conhecido como federação de nuvens, ou seja, nuvens formadas por diferentes provedores interconectados entre si. Entretanto, a criação de métodos eficientes para promover essa interconexão, que permitirá explorar as vantagens oferecidas por diferentes provedores, ainda é um desafio. Para permitir a execução de um *workflow* em diferentes nuvens simultaneamente, é importante que exista a interoperabilidade entre as nuvens envolvidas. Outros desafios estão ligados a segurança e privacidade (Hashizume, 2013), já que certos *workflows* podem manipular dados sensíveis (com restrições de acesso) ou mesmo dados que só podem ser armazenados em localidades geográficas pré-definidas.

Recentemente, o projeto *Magellan*[14] investigou o potencial da computação em nuvem. Uma infraestrutura distribuída foi preparada e foram analisadas diversas aplicações científicas em áreas variadas, como a metagenômica e física nuclear e ótica. Entre as principais constatações do estudo, é possível citar as seguintes:

- Iniciativas de computação em nuvem trazem várias vantagens, como: ambientes personalizados sem um grande custo adicional de administração (compra e manutenção), habilidade de se conseguir mais recursos rapidamente para problemas maiores e economia de escala.
- A adaptação de aplicações já existentes para a execução em nuvem pode exigir um esforço considerável, que não pode ser negligenciado antes de se decidir pelo uso de nuvens.
- Ainda existem vários desafios nas áreas de gerenciamento de ambientes virtuais, *workflows*, dados, segurança, entre outros. É necessário o desenvolvimento de ferramentas que simplifiquem o uso de computação em nuvem.

[14] Projeto *Magellan: Cloud Computing for Science*: http://www.alcf.anl.gov/magellan.

- Aplicações científicas com pouca movimentação de dados são as que melhor se adaptam aos ambientes de nuvens. Para outros tipos de aplicações, a perda de desempenho em relação a um ambiente dedicado pode ser grande.

No início do século XX, o fornecimento de energia elétrica era feito por centrais elétricas locais, sem conexão umas com as outras e com características de tensão e frequência diferentes. Pensar em uma rede elétrica conectada, em que os consumidores também pudessem ser fornecedores, poderia parecer ficção científica naquela época. A computação em nuvem, que proporciona a obtenção de processamento e armazenamento sob demanda, pode chegar a ser em breve um ambiente completamente integrado, de forma que o usuário nem saiba se está usando recursos locais ou externos. As possibilidades são inúmeras, não só para *eScience*, mas para a computação em geral.

5. Democratização da *eScience*

A ciência da computação deixou de ser uma ferramenta de apoio para se tornar um verdadeiro alicerce do processo de criação de conhecimento em diversas ciências. As mudanças trazidas pelo uso de seus conceitos, teoremas, técnicas e métodos provocaram o surgimento de um novo paradigma de metodologia científica – o que hoje se conhece por *eScience* (ou eCiência). A coleta e análise de uma grande quantidade (antes inimaginável) de dados agora é possível com o uso de computação. Experimentos com novos modelos podem ser realizados de forma virtual, por meio de técnicas como simulação. Isso possibilitou novas maneiras de se fazer ciência. A Ciência está se transformando progressivamente em eCiência.

O uso de novas técnicas de computação paralela e distribuída como a computação em nuvem promove a democratização do acesso ao poder computacional. Oportunidades de pesquisa que antes eram restritas ao seleto grupo dos que tinham acesso a supercomputadores agora podem ser exploradas por milhares de pesquisadores. Espera-se que a evolução dessas novas tecnologias, aliada ao aumento da integração da ciência da computação às outras ciências, permita que todo pesquisador tenha condições de fazer pesquisa transformativa em qualquer área do conhecimento, promovendo novas mudanças de paradigma na ciência.

Referências

CARDELLI, Luca. *Abstract machines of systems biology*. Transactions on Computational Systems Biology III, p. 145-168, 2005.

CINTHO, Mina et al. *Data-intensive analysis of HIV mutations*. In: IEEE 8th International Conference on E-Science, 2012, Chicago, p. 1-7

EMMOTT, Stephen et al. (Orgs.). *Towards 2020 Science*, 2006. Disponível em: http://research.microsoft.com/towards2020science/. Acesso em: 12 mar. 2013.

HASHIZUME, Keiko. *An analysis of security issues for cloud computing.* Journal of Internet Services and Applications. 4:5, 2013.

HEY, Tony et al. (Eds.). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009. Disponível em: http://research.microsoft.com/collaboration/fourthparadigm. Acesso em: 12 mar. 2013.

JOHNSON, George. *The World: In Silica Fertilization; All Science Is Computer Science*. The New York Times, 25 mar. 2001. Disponível em: http://www.nytimes.com/2001/03/25/weekinreview/the-world-in-silica-fertilization-all-science-is-computer-science.html. Acesso em: 12 mar. 2013.

KORPELA, Eric J. et al. *Status of the UC-Berkeley SETI efforts*. Proc. SPIE 8152, Instruments, Methods, and Missions for Astrobiology XIV, 815212, 23 set. 2011.

KUHN, Thomas S. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

MEYER, Eric. *Accessing and Using Big Data to Advance Social Science Knowledge*. Oxford Internet Institute, 2013. Página Web do projeto de pesquisa: http://www.oii.ox.ac.uk/research/projects/?id=98. Acesso em: 12 mar. 2013.

O'REILLY, Tim. *The World's 7 Most Powerful Data Scientists*. Forbes, 2 nov. 2011. Disponível em: http://www.forbes.com/sites/nicoleperlroth/2011/11/02/tim-oreilly-the-worlds-7-most-powerful-data-scientists/. Acesso em: 12 mar. 2013.

PENTLAND, Alex. *Reinventing society in the wake of Big Data - A Conversation with Alex (Sandy) Pentland*. Edge, 30 ago. 2012. Disponível em: http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data. Acesso em: 12 mar. 2013.

RIMKUS, Carolina M. et al. *Corpus Callosum Microstructural Changes Correlate with Cognitive Dysfunction in Early Stages of Relapsing-Remitting Multiple Sclerosis: Axial and Radial Diffusivities Approach*. Multiple Sclerosis International. v. 2011, p. 1-7, 2011.

ZHANG, Qi, et al. *Cloud computing: state-of-the-art and research challenges.* Journal of Internet Services and Applications, 1(1), pp. 7-18, 2010.