Scientific Discovery in the Era of Big Data: More than the Scientific Method

A RENCI WHITE PAPER
Vol. 3, No. 6, November 2015

renci

Scientific Discovery in the Era of Big Data: More than the Scientific Method

Authors

Charles P. Schmitt, Director of Informatics and Chief Technical Officer

Steven Cox, Cyberinfrastructure Engagement Lead

Karamarie Fecho, Medical and Scientific Writer

Ray Idaszak, Director of Collaborative Environments

Howard Lander, Senior Research Software Developer

Arcot Rajasekar, Chief Domain Scientist for Data Grid Technologies

Sidharth Thakur, Senior Research Data Software Developer

Renaissance Computing Institute
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
919-445-9640

RESEARCH ENGAGEMENT INNOVATION

AT A GLANCE

- Scientific discovery has long been guided by the scientific method, which is considered to be the "gold standard" in science.
- The era of "big data" is increasingly driving the adoption of approaches to scientific discovery that either do not conform to or radically differ from the scientific method. Examples include the exploratory analysis of unstructured data sets, data mining, computer modeling, interactive simulation and virtual reality, scientific workflows, and widespread digital dissemination and adjudication of findings through means that are not restricted to traditional scientific publication and presentation.
- While the scientific method remains an important approach to knowledge discovery in science, a holistic approach that encompasses new data-driven approaches is needed, and this will necessitate greater attention to the development of methods and infrastructure to integrate approaches.
- New approaches to knowledge discovery will bring new challenges, however, including the risk of data deluge, loss of historical information, propagation of "false" knowledge, reliance on automation and analysis over inquiry and inference, and outdated scientific training models.
- Nonetheless, the time is right for increased focus on the construction of *Collaborative Knowledge Networks for Scientific Discovery* designed to leverage existing data sources and integrate traditional and emerging scientific methods and thereby drive scientific discovery and application.

Introduction

Knowledge discovery in science refers to the systematic process whereby scientists draw logical conclusions regarding the world around us, generate new theories based on those conclusions, and share findings with other scientists and the lay public, thus enabling critical review and consensus before new findings are added to the collective body of knowledge. Historically, scientific discovery has been guided by the scientific method, which dates to ancient times and involves both a philosophical and practical approach to science. Indeed, the renowned philosopher Aristotle (384-322 BC) was one of the first to approach knowledge discovery through rigorous, systematic observation, although it wasn't until millennia later that the scientific method was actually formalized and implemented, largely through the work of Copernicus (1473-1543), Tycho Brahe (1546-1601), Johannes Kepler (1571-1630), Galileo Galilei (1564-1642), Rene Descartes (1596-1765), and Isaac Newton (1643-1727) (Gower 1997; Betz 2011).

At first glance, the scientific method appears to be relatively simple and straightforward. In sum, the method involves a repeating cycle of standardized steps: the process begins with careful observation of the natural world, the framing of a question on the basis of one's observations, and a review of the existing body of knowledge to determine if a reasonable explanation already exists. Assuming that the question remains open ended, a scientist will then formulate a hypothesis, design and implement an experiment to test the hypothesis, and analyze the experimental results. The analytical results are used to formally accept or reject the hypothesis. The hypothesis may then be modified, with the experiment repeated or a new experiment designed. Importantly, the results are then disseminated via presentation to peers and publication, which allow for adjudication[1] or peer consensus regarding the validity of the scientific findings. This, in short, is scientific discovery via the scientific method.

Revisiting Charles Darwin

As any school child will attest, the work of Charles Darwin provides an exemplar of the scientific method and the many challenges involved in scientific discovery (Burkhardt 1996; McKie 2008; Montgomery 2009). Darwin's genius perhaps lies in his keen ability to **observe** the natural world. One of his earliest observations was that similar species are found across the globe and that individuals within a species are not identical but have local variation. He **questioned** why multiple species exist instead of just one. Darwin continuously **researched** and reviewed the existing scientific literature (i.e., the established scientific body of knowledge).

Darwin's work was influenced by the research and writings of Thomas Malthus, who found that humans produce more offspring than are needed to replace themselves and speculated that population size would soon exceed the available resources required for survival. Darwin also observed that populations of plants and animals stay about the same size because of limited resources and competition for those resources. Darwin's thinking also was influenced by the research and writings of Charles Lyell, who found that small, gradual geological processes can produce large changes over time. Darwin made several brilliant inferences on the basis of his scientific observations and the work of Malthus, Lyell, and others that led him to **hypothesize** that species change slowly in a process of evolution from a common ancestor. He then spent decades in observational **experimentation**, ongoing **analysis** of his scientific findings, and refinement of his hypothesis until he eventually reached the now famous **conclusion** that the origin of the species lies in natural selection, or the process whereby individual variation, coupled with competition among individuals for natural resources and social cooperation among kin to increase individual fitness, determines differences among related species.

Coincidentally, while Darwin was refining his hypothesis and drawing a conclusion, Alfred R. Wallace, a much junior researcher who was familiar with Darwin's work, reached a similar conclusion, which he planned to publish and **disseminate** to scientific peers. Aware that his work might go unrecognized,[2] Darwin reached an agreement with Wallace, brokered by Lyell and Joseph D. Hooker, and reluctantly **published** a joint scientific manuscript in 1858. *On the Origin of Species by Natural Selection* wasn't published until November 1859, but only as a book chapter, not the full book that Darwin had intended. Interestingly, the scientific community reacted negatively to Darwin's work (e.g., Gray 1860); many of Darwin's peers felt that he did not have sufficient evidence to put forth his hypothesis, while others were disturbed by the theological, political, and social implications. The **scientific debate** continued for nearly a century until the scientific community reached **adjudication** on Darwin's hypothesis and scientific findings and established the theory of evolution by natural selection. Of note, the **lay debate** continues today, largely on theological and political grounds.

[1] "Adjudication" is a legal term that refers to decision making in the presence of a neutral third party who has the authority to determine a binding resolution through some form of judgment or award. In science, adjudication generally refers to decision making and consensus building among a group of widely regarded scientific experts on the topic under discussion (Spangler 2003).

[2] "I never saw a more striking coincidence, if Wallace had my M.S. sketch written out in 1842 he could not have made a better short abstract! Even his terms now stand as Heads of my Chapters…So all my originality, whatever it may amount to, will be smashed. Though my Book, if it will ever have any value, will not be deteriorated; as all the labour consists in the application of the theory"—Charles Darwin to Charles Lyell, June 18, 1858 (In *Charles Darwin's Letters. A Selection*, F. Burkhardt (Ed.), 1996).

The scientific method has certain desirable characteristics that have enabled it to withstand the passage of time, even as new scientific tools and techniques have been introduced to the scientific process. For example, the method is objective and removed from all personal and cultural biases. All hypotheses are developed to be consistent with accepted scientific truths at the time that the hypothesis is generated. All measurements must be observable and pertinent to the hypothesis. The hypothesis and all conclusions must follow the principle of Occam's Razor in terms of parsimony, or simplicity with few assumptions. The hypothesis also must be falsifiable (and capable of being disproven). Finally, the scientific findings must be reproducible by other scientists.

Without dispute, the scientific method remains the most common and only validated approach to scientific discovery and knowledge extraction. In fact, drug development in the U.S. relies exclusively on the randomized placebo-controlled clinical trial—the exemplar of the scientific method.

However, today's advancements in digital computing and storage capabilities, coupled with new methods for scientific communication, including social media, are introducing new approaches to scientific discovery, each of which brings challenges and opportunities. In this white paper, we consider how the advent of the digital age and today's world of "big data" are changing scientific discovery processes. We close with a framework for *Collaborative Knowledge Networks for Scientific Discovery* designed to leverage existing data sources and integrate traditional and new data-driven scientific methods, allowing for unprecedented advances in scientific discovery and application.

Exploratory data sets and exploratory analysis

Today's scientist has access to numerous large data sets of relevance to multiple scientific domains. For example, the National Oceanic and Atmospheric Administration maintains several databases containing data on climate patterns, earthquakes, ozone levels, and ocean temperatures; these data are useful to scientists in many fields, including environmental science, energy, public health, and medicine. Scientists are increasingly accessing these data sets for exploratory analysis, which is an approach used to determine if a general hypothesis bears any merit and/or if an experimental design is feasible. Exploratory analysis typically begins with a general hypothesis or experimental design that isn't well fleshed out and the application of a variety of statistical approaches and visualization techniques to identify and validate data elements and/or determine if a hypothesis is testable using the data (Behrens 1997; Gelman 2004; Diaconis 2011).

For example, consider an economics researcher who is interested in learning about the loaning practices of a large credit union, in terms of the breakdown of mortgage loans across socioeconomic sectors. S/he requests access to the credit union's loan database in order to determine how the data are structured and how fine-grained the available socioeconomic data elements are. The researcher determines that the database contains data elements on the age, sex, and gross income of loan holders, as well as the loan amount and the location of the property, but it does not contain information on the ethnicity of the loan holders. The researcher uses this information to modify his/her hypothesis for subsequent testing using the same data.

Exploratory analysis has always been part of scientific discovery and, historically, has been used to generate the driving question underlying a scientific hypothesis. However, in the past, the process was informal, slow, and observation-based (e.g., detailed notes on the types of foliage identified at different altitudes within a given region); whereas today, it is fast, large-scale, data-driven and often involves extensive use of advanced statistical methods and visualization techniques. Furthermore, large data sets are being generated strictly for exploratory analysis, and often such data sets incur a considerable expense with questionable cost-benefit.[3] A recent McKinsey report (Manyika, et al. 2015), for example, estimates that less than 1% of data captured and stored from an offshore oil rig equipped with 30,000 sensors is actually used to guide operations. The value of the rest of the data remains to be determined.

Today, exploratory analysis is fast, large-scale, data-driven and often involves extensive use of advanced statistical methods and visualization techniques.

In terms of benefits, exploratory analysis enables a scientist to quickly develop a more refined, testable hypothesis and rigorous experimental design for subsequent hypothesis testing (Behrens 1997; Gelman 2004; Diaconis 2011). Challenges include the costs involved with the generation and long-term storage of exploratory data sets. In addition, drawing conclusions from an analysis of exploratory data sets can introduce error if the data elements are not described with sufficient metadata to fully understand data structure and meaning or if the data elements have inherent biases due to how they were collected. Additional challenges include issues of privacy and the inadvertent or intentional leakage of “sensitive” data, particularly if a scientist does not obtain the requisite authorization to access a data set that contains sensitive data or if sensitive data are accidentally provided to an unauthorized or third-party user (Behrens 1997).

Data mining

Data mining often begins without a hypothesis and involves the application of tools and techniques from statistics, mathematics, and visualization to identify previously unknown patterns and trends in a data set derived from one or more existing databases (Fayyad, et al. 1996; Holzinger, et al. 2014). While data mining is sometimes considered a form of exploratory data analysis (Behrens 1997; Diaconis 2011), we argue for a distinction between *exploratory data analysis*, which enables a scientist to explore a data set in terms of structure and types of data elements, including the relevancy of external data sources (e.g., literature citations) to

[3] The costs and benefits of these exploratory data sets are an open topic of discussion; for example, see Wilhelmsen, et al. 2013 and Evans, et al. 2015 for discussions on the pros and cons of capturing and storing whole genome versus targeted genome sequencing data.

data elements, and *data mining*, which enables a scientist to identify previously unknown patterns in a data set in terms of relationships between data elements. We note that new statistical approaches are being developed to overcome some of the limitations of both types of scientific discovery. For example, the maximal information coefficient, or MIC, overcomes the limitations of Pearson's correlation coefficient, *r*, by allowing for the identification of complex, nonlinear relationships (e.g., exponential, periodic) between data elements in data sets of any size (Reshef, et al. 2011).[4] Suppose a microbiologist generates a gene expression data set to identify genes involved in the regulation of the cell cycle. Using traditional statistical approaches, the scientist is able to identify genes with strong deterministic or linear associations with different aspects of the cell cycle (e.g., a gene that is required for chromatin assembly). Using approaches such as the MIC, the scientist is able to identify genes with periodic relationships with the cell cycle (e.g., a gene associated with an established cyclical event or an event that occurs at a previously unknown frequency during the cell cycle).

Data mining bears little resemblance to the scientific method. In particular, data mining is not: (1) constrained to be objective; (2) consistent with accepted scientific truths; (3) parsimonious; or (4) falsifiable. In addition, all measurements are observable only after the fact. Nonetheless, data mining offers benefits, including the ability to relatively quickly discover previously unknown relationships and thereby generate a new hypothesis that can then be tested using the scientific method (Fayyad, et al. 1996; Holzinger, et al. 2014). Although often cited as a criticism of data mining, a key benefit is the ability to generate multiple associations within a single data set, which may yield powerful new information, especially when combined with associations identified in other data sets. In this regard, data mining can be viewed as akin to meta-analysis of clinical trial data.

> A key benefit of data mining is the ability to generate multiple associations within a single data set, which may yield powerful new information, especially when combined with associations identified in other data sets.

Challenges include the fact that data mining requires extensive training beyond the skills of a typical domain scientist; without such training, a scientist may identify patterns or associations that are not valid or reproducible (Fayyad, et al. 1996; Behrens 1997; Diaconis 2011; Holzinger, et al. 2014). Further, there is no commonly accepted approach or method for data mining, which makes the field somewhat more of an art than a science (Fayyad, et al. 1996; Diaconis 2011). There is also a tendency to treat all findings as conclusive when they may be chance

[4] As an aside, we note that the theoretical foundation of the MIC lies in the concept of mutual information (MI) in pairs of random variables, which was developed by Claude Shannon, the founder of information theory, more than 50 years ago (Speed 2011). The implementation of MI into practice as MIC cannot be achieved through manual or semi-manual computation and instead requires significant digital computational power—something not available until recently.

findings (Fayyad, et al. 1996; Diaconis 2011). In addition, while the power of data mining increases when multiple heterogeneous data sets are integrated before the data are mined, so too does the complexity of the process (Fayyad, et al. 1996; Holzinger, et al. 2014). With high-dimensional data, multiple approaches must be used to analyze the data; for example, sophisticated visualization approaches may need to be combined with traditional statistical approaches to data mining (Fayyad, et al. 1996; Diaconis 2011; Holzinger, et al. 2014).

Computer modeling

Computer modeling involves conceptual, mathematical, computer-generated, or physical representations of real-world objects or phenomenon (Bowers 2012; Buytaert, et al. 2012; Perra et al., 2012; Berman, et al. 2015). It is used to test a hypothesis or observe and manipulate an object or phenomenon that is otherwise difficult (or unethical) to observe and manipulate.

As an example, consider a scientist who wishes to determine how the beta-amyloid protein influences the development of Alzheimers Disease. S/he develops a software program using established principles of protein folding to visually explore in 3-D different scenarios whereby beta-amyloid may fold in the brain to influence cognitive function and lead to the development of Alzheimers Disease.

Modeling represents a fundamental aspect of scientific discovery and has been used throughout the history of modern science (e.g., anatomical models). The use of computer modeling is relatively new, however, and as such, computer-driven modeling tends to be a highly specialized tool as opposed to a general tool. Benefits of computer modeling include the addition of a potentially powerful tool to the formerly manual process, particularly when models incorporate the vast streams of data that are available from technologies such as crystallography and other sophisticated sources of data (Perra, et al. 2012; Berman, et al. 2015). Computer models become even more powerful when they are generated using data derived from multidisciplinary science (Bowers, et al. 2012; Buytaert, et al. 2012).

Challenges, however, include the fact that computer-generated models are only as good as the underlying data and software programs used to create them (Joppa, et al. 2013; Berman, et al. 2015). In addition, the risk of introducing error increases significantly when models are created for poorly understood objects or phenomenon, when the data are qualitative or otherwise described in a non-standardized format, or when models developed in one field are applied (without validation) to another field or even shared between different research groups within the same field (Bowers, et al. 2012; Buytaert, et al. 2012; Perra, et al. 2012; Joppa, et al. 2013; Bergman, et al. 2015). Once introduced, errors tend to propagate and may even amplify (Buytaert, et al. 2012). Moreoever, many models are not designed to work in a dynamic, flexible, user-driven, web environment (Buytaert, et al. 2012). Finally, modeling costs can be quite high, sometimes higher than the costs of the instruments used to generate the data that are used in the models (Berman, et al. 2015).

Interactive simulation and virtual reality

Interactive simulation and virtual reality are similar to modeling in that a computer is used to create realistic scenarios for testing a hypothesis or manipulating an object or phenomenon that is otherwise difficult (or unethical) to test or manipulate. However, the difference, as defined herein, is that with interactive simulation and virtual reality, human behavior, not the underlying software program, guides the outcome of the simulation or virtual experience (Zyda 2005; Kunkler 2006; Diemer, et al. 2015).

Suppose a medical device company has been developing a new anesthesia machine. In order to obtain approval from the U.S. Food and Drug Administration, the device has to be tested for safety in Phase I and II clinical trials. To achieve this, the company creates a "dummy" patient that is programmed to mimic physiological responses during particular types of surgeries in a simulated operating room. A Phase I clinical trial is then conducted in the simulated environment, using a team of actual surgeons, anesthesiologists, and nurses as study subjects.

Interactive simulation and virtual reality can be used for hypothesis testing as part of the scientific method, but they also can be used for exploratory analysis. The benefits include the ability to investigate human behavior in the context of technology or computer-assisted scenarios that resemble the real world (Zyda 2005; Kunkler 2006; Diemer, et al. 2015). Thus, the approach facilitates both team-based science and science on human teams. Several challenges exist, however. First, a team with expertise in both software development and human behavior must create the software programs that drive the simulations and virtual scenarios, with careful attention to every detail of the scenario (Diemer, et al. 2015). Also, human behavior in a simulated environment might not be identical to human behavior in the real world; at present, software programs cannot fully capture the true human experience, including behaviors such as emotional reactions (Kunkler 2006). Finally, technological challenges grow exponentially with the complexity of the scenario, and upfront costs are high (Zyda 2005).

Scientific workflows

Scientific workflows refer to the abstract steps required to complete a specific scientific task. Today, these are typically designed as specialized workflow management systems that are capable of orchestrating and automating many of the steps, including data flow and, in some cases, personnel (e.g., scheduling, access rights, alerts or reminders), required to complete a specific task and/or test a hypothesis(es) (Curcin & Ghanem 2008; Perraud, et al. 2010; Achilleos, et al. 2012; Bowers 2012; Guo 2013). The concept of the automated scientific workflow appears to have originated with an undated publication by Singh and Vouk (see references). Scientific workflows are used to automate tedious, time-consuming, or highly complex steps in experimental testing and/or analysis.

For instance, imagine that a biomedical research company relies on flow cytometry, a technique that enables the visualization and quantification of the protein composition of live cells, as a critical part of its drug discovery efforts in diabetes. The company decides to automate much of the process and hires a software engineering team to design a scientific workflow system to automate and coordinate the technologies and research teams involved in each step of the process, including isolation of blood cells from patients with diabetes, incubation of cells with conjugated antibody(ies), multiple washes, immunofluorescence visualization, and analysis to determine cellular subpopulations.

The design of most scientific workflow systems models the actual scientific method, *except* that the measurements are automated and not always observable by the scientist. Moreover, scientific workflows are increasingly being used for automated deduction and inference, which are steps that historically relied on the scientist. In the example above, for instance, the use of flow cytometry to identify cellular subpopulations on the basis of their fluorescent profile can be more of an art than a science, especially when multiple fluorescent conjugates are used; as such, automating the process can yield erroneous or inconsistent results. That said, scientific workflows enable a scientist to conduct experiments more quickly, efficiently, and with greater statistical power and reproducibility due to the large data volumes that a well-designed scientific workflow can handle and the ability to integrate heterogeneous data sources, often in real time (Perraud, et al. 2010; Achilleos, et al. 2012; Bowers 2012; Guo 2013).

> Scientific workflows are increasingly being used for automated deduction and inference, which are steps that historically relied on the scientist.

While potentially quite powerful, the scientist in the realm of the workflow is often dependent on the technology and removed from critical decision-making steps (e.g., calculations, incubation times, etc.), which, in the absence of careful system design and implementation, threatens the validity and reproducibility of workflow findings. Furthermore, unless the workflow steps are clearly annotated and openly accessible, a peer reviewer will be unable to fully evaluate and reproduce the scientific findings (Bowers 2012). Moreover, workflow design can be quite complex, involving hundreds of individual analysis steps, large sets of heterogeneous data, and multiple existing workflows that often were not designed to work together; the complexity presents difficulties in user-friendliness, design, and capture and record of provenance[5] (Perraud, et al. 2010; Achilleos, et al. 2012; Bowers 2012; Gup 2013). There is also a practical limit to the degree of granularity in workflow tasks; highly granular activities are typically not feasible due to a negative impact on overall system performance (Perraud, et al. 2010). An additional concern is that scientific workflows can introduce systematic bias in that errors (e.g., an incorrect calculation) may be introduced and propagated indefinitely because automation makes it more difficult to detect and correct errors. Finally, workflows tend to be very specialized and often cannot be realistically adapted for different domains (Curcin & Ghanem 2008).

[5] "Provenance" refers to the capture of metadata that describes the origins of the data, each step of data transformation and analysis, and a record of versioning to identify how up-to-date the data are (Guo 2013).

Dissemination and adjudication

Dissemination and adjudication of scientific discoveries are critical components of scientific discovery, as these are the processes through which new scientific "truths" are added to the collective scientific body of knowledge (Smith 2006; ACS publications 2013; Almeida 2013; Cold Spring Harbor Laboratory 2014). Dissemination has historically been achieved through presentation to scientific peers and publication in scientific journals in order to obtain critical peer review and validation (or rejection) of scientific findings and inform the scientific and lay communities. Adjudication is the method by which a consensus is reached among scientific experts regarding the validity of the scientific findings (also see footnote 1).

As an example of the dissemination and adjudication processes, assume a scientist delivers a PowerPoint presentation on new technologies for hydrology at a scientific conference on water safety. S/he receives immediate feedback on the presentation from colleagues in attendance at the lecture. On the basis of that feedback, the scientist decides that an additional experiment is required before his/her work is ready for publication in a peer-reviewed journal. To provide another example, imagine a scientist intends to publish his/her new ceramics model in a scientific journal focused on materials science. The journal requires critical peer review.[6] The scientist submits the manuscript to the journal for potential publication, and it is reviewed by anonymous peers. On the basis of the peer review, the editor decides that before the manuscript can be published, the text needs to be revised to better frame the need for a new ceramics model in the context of existing, similar models.

Historically, dissemination and adjudication have been key components of the scientific method and the final screen before new scientific truths are added to the scientific body of knowledge. The benefits are tremendous because the process enables scientific findings to be rigorously evaluated by expert scientists, thus ensuring the robustness and integrity of scientific findings (Smith 2006; ACS publications 2013; Almeida 2013; Cold Spring Harbor Laboratory 2014). The challenges, however, are growing in the era of big data. Specifically, the digital world is changing the way that scientific findings are generated, presented, published, and catalogued. For example, traditional, paper-based, peer-reviewed scientific journals are competing with new, electronic, open access scientific journals[7] that may not require a slow, rigorous *peer* review, but only a quick, minimal *editorial* review before publication (Almeida 2013; Cold Spring

[6] "Peer review" refers to the process whereby scientific journals or scientific funding organizations enlist (for free) the help of colleagues within the same field as the manuscript or grant proposal's investigative team to evaluate the scientific merit and integrity of a document (Smith 2006).

[7] "Open access" scientific journals are online journals that permit access to all journal content without a subscription fee. (Cold Spring Harbor Laboratory 2014). The premise is to allow ordinary citizens to access all scientific findings, particularly those that are generated with federal money, as a "public good" (e.g., http://publicaccess.nih.gov/), much the same way that federal salaries are released to the public for evaluation. "Public goods", as defined by economists, are those items (commodities or services) that are both *non-excludable* and *non-rivalrous*; examples include public parks, sewer systems, highways, police services, etc. (Samuelson 1954). Ironically, many journals require the author to pay a fee for open access publication.

Harbor Laboratory 2014). In addition, many scientists are now publishing non–peer-reviewed scientific findings in electronic form, via websites, digital white papers and technical reports, blogs, webcasts, YouTube videos, etc. (ACS publications 2013; Almeida 2013). This approach allows for rapid dissemination of scientific findings, but it bypasses both the peer review process for publication and the peer selection process for speaker presentation at a scientific meeting. That said, the current peer review process is not without flaw, in terms of bias and error (Smith 2006), so perhaps the shift to digital self-publication will reform the peer review process or kindle an entirely new form of peer review. New hybrid models such as eGEMs combine peer review with free publication and open access permissions; whether these models are fiscally sustainable has yet to be seen. Another challenge is the wealth of scientific information available today. This dissemination deluge makes it challenging to adjudicate existing findings and identify important new findings that may become lost in the wealth of available data (i.e., the "long tail of science") (Trader 2012).

Knowledgebases

Traditionally, the dissemination and adjudication of scientific knowledge involve social processes; however the storage, integration, search, and inference of knowledge is rapidly changing to a hybrid model that relies equally upon humans *and* large, complex, collections of digital information that include data, relationships between data elements, human assertions on the data, and, in some instances, capabilities for automated cognitive processing: knowledgebases. The term "knowledgebase" was introduced in the 1970s (Jarke, et al. 1978) to distinguish it from the existing database of the time, which essentially stored data in tabular form for user access via query. The term remains loosely defined, but a knowledgebase can be considered to be a computing system in which assertions are represented and persisted within an overarching ontological framework that provides the semantic context and that allows for querying and computing across assertions. The human user is integral to the knowledge derived from and contained within the knowledgebase; and, in some cases, the system can be structured to derive new, automated assertions based on machine learning or new data. While traditional databases have evolved to become transaction-based and relational and can be part of a knowledgebase system, knowledgebases remain differentiated in that the human user's ability to dynamically access, add to, and use the knowledge is an essential part of the design and function of the system (e.g., Wikipedia) (Pugh & Prusak 2013).

Completely automated knowledgebases are under development, but automated curation of human assertions is not yet possible and may never be. Nonetheless, emerging knowledgebases are incorporating sophisticated algorithms for automated and semi-automated structuring, parsing, summarization, retrieval, and visualization of information. Leading-edge production knowledgebases currently contain 10^9 to 10^{11} assertions mined from a variety of unstructured and semi-structured data sources, including Wikipedia and PubMed (Bordes, et al. 2015; Southern 2015). Prominent commercial examples of knowledgebases include the IBM Watson system and Elsevier Pathway Studio; proprietary examples include WalMart's Social Genome and Google's Knowledge Graph; open source systems include OpenBEL, Open PHACTS, and DARPA's DeepDive. Other examples include the Semantic Web and the Linked Data movement.

The use of knowledgebases in scientific research is growing. For instance, the GenBank knowledgebase allows one to search for specific genes and then identify a wealth of curated information about that gene and related genes and external data sources. The GeneOntology (GO) knowledgebase provides annotation on genes, biological processes, cellular components, and molecular functions and has been prominent in the application of knowledgebases for new statistical approaches such as enrichment analysis (Mi, et al. 2013). As scientists become more aware of the benefits of knowledgebases, and as the quality and quantity of their assertions increase, their application in scientific discovery is expected to grow.

> Emerging knowledgebases are incorporating sophisticated algorithms for automated and semi-automated structuring, parsing, summarization, retrieval, and visualization of information.

Major challenges in the development of knowledgebases for scientific discovery include the costs of human curation, in terms of the amount of time and experience required to contribute to a knowledgebase in a meaningful way, and the lack of incentives for scientists to contribute to knowledgebases, especially those perceived to be outside of a scientist's area of expertise. As exemplified in the knowledgebase systems listed above, recent research has advanced our understanding of how to construct knowledgebases, including the incorporation of probabilistic models, structured representation of unstructured data, deep-learning systems, question-answer schemas, and natural language processing algorithms. Recent research also has advanced our understanding of how to integrate knowledge from different sources with differing quality and differing semantics (Southern, 2015; Nickel 2015). For example, DeepDive is built upon a probabilistic graph model that allows for the encoding of assertions and relationships between data elements that have inaccuracies, such as those that are derived from predictive models and data-driven approaches, along with assertions that have higher validity. The system uses inference engines to mine the relationships, as well as the strength of the assertions, in order to construct different "world models" upon which new inferences can be made. These advances show promise, but the human user, *the scientist*, remains the roadblock in the further development and widespread application of knowledgebases for scientific discovery.

The Big Picture: The Future of Knowledge Discovery in Science

We recognize that there have always been scientific approaches that do not rely on the scientific method, as others have argued previously (e.g., Cleland 2001). Yet, the scientific method remains the "gold standard" in terms of scientific discovery. The wealth of data available today offers the unprecedented opportunity to conduct science in ways never before envisioned: real-time science, real-world data, and new methods of scientific discovery. In embracing new approaches to scientific discovery that are not necessarily aligned with the scientific method, or maybe even contradict it, we must consider how to best employ and unify the new approaches with each other and with the traditional scientific method.

John Tukey, considered by many to be the most influential statistician in modern times, once stated: “The best part of being a statistician is that you get to play in everyone’s backyard.” Among Tukey’s many accomplishments are the introduction of the terms “bit” (in 1946) and “software” (in 1958) and the concepts and methods of exploratory data analysis, data mining, the Tukey Fast Fourier Transformation, and a variety of other statistical and mathematical approaches for scientific discovery, many of them also named after him (Bittrich 2000; Leonhardt 2000; Brillinger 2002). Tukey also introduced the term “uncomfortable science”, which has been described as lying somewhere between classical mathematical statistics and “magical thinking” (Diaconis 2011). The concept is that in many instances, scientific inference can and must be made from intuition and exploration, rather than controlled experimentation, using a finite amount of potentially rich data that are flawed and often nonreplicable.[8] Tukey died in 2000, before the introduction of the terms “big data” and “Internet of Things”, but one can bet that he’d be the first to take advantage of the many opportunities that the “datafication” of society has provided (Bertolucci 2013).

The wealth of data available today offers the unprecedented opportunity to conduct science in ways never before envisioned: real-time science, real-world data, and new methods of scientific discovery.

Another forward-thinking scholar is the late James (Jim) Gray, whose last position was as a Technical Fellow and Manager of the Bay Area Research Center at Microsoft. In addition to his pioneering work on large databases and transaction processing systems, Gray introduced the concept of the *Fourth Paradigm* for science (Hey, et al. 2009).[9] While Gray’s initial focus was on exploratory analysis and data-intensive scientific computation, the Fourth Paradigm essentially represents the radical transformation of the scientific method from hypothesis-driven to hypothesis-generating science, as described herein.[10] Gray’s paradigm was developed in the late 1990s and early 2000s, but his genius lies in his vision for today’s world, just a decade or

[8] “Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.”—John Tukey, 1962, The future of data analysis. *Annals of Mathematical Statistics,* 33, 1–67 (quoted on pp. 13-14).

[9] After Gray’s death, Alex Szalay and colleagues formally introduced the concept of the “Fourth Paradigm”, with his co-authored publication in the journal *Science* (Bell, et al. 2009).

[10] During Gray’s last lecture on January 11, 2007 [http://research.microsoft.com/en-us/um/people/gray/jimgraytalks.htm], he is quoted as stating the following: “The world of science has changed, and there is no question about this. The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, *fourth paradigm* for scientific exploration.” (Hey, et al., 2009).

two later, where nearly every scientific domain is moving toward data-intensive and data-enabled scientific discovery. This shift involves traditional fields such as physics and astronomy, which have always had access to rich data sources but are now facing unprecedented computational and analytical challenges, and emergent fields such as environmental science, which has only recently had access to the wealth of distributed data available from environmental sensors and satellites. Even the "soft" sciences such as social science and political science are recognizing the power of data derived from social media data, mobile devices, and "smart cities". Moreover, applied fields such as medicine are embracing the Fourth Paradigm and these myriad new data sources. For example, in 2011, the National Research Council organized an *ad hoc* committee to develop a framework for a unified taxonomy of human disease that, when implemented, will accelerate progress toward precision medicine, which is a new area of medicine that aims to personalize medicine through the integration of data on individual variability in genes, environment, and lifestyle in order to identify the most appropriate medical monitoring and treatment plan for any given patient (National Research Council 2011). The committee's work largely motivated President Obama's January 2015 announcement during the State of the Union Address on a new National Initiative in Precision Medicine (Collins & Varmus 2015).

A Framework for Scientific Discovery in the Era of Big Data: The Collaborative Knowledge Network

The time is right to whole-heartedly embrace the flexibility and power that new approaches to scientific discovery offer, while maintaining the power that the traditional scientific method holds.

The scientific approaches presented in this white paper are increasingly being adopted by scientists; however, there remains hesitancy about their roles and legitimacy in the scientific process, a lack of training in and incentives for their use, and the need for additional research to tailor and improve these approaches for scientific discovery. *We argue that a primary challenge in moving towards a combination of hypothesis- and data-driven science is the integration of approaches at a community level.*

The National Research Council's framework to create a path toward personalized medicine includes the concept of a *Knowledge Network of Disease* as a federated discovery network organized around an *Information Commons* of structured patient-centric data based largely on genomics and including patient medical history, current signs and symptoms, etc. (National Research Council 2011). We suggest a similar, but broader framework for a *Collaborative Knowledge Network for Scientific Discovery* (**Figure 1**). The envisioned network would build upon the knowledgebases that are already being developed, but through an open, community-based effort designed to link these knowledgebases and thereby create a knowledge network for scientific discovery. The effort would involve not only scientific experts, but all stakeholders, including policy makers, industry representatives, and even ordinary citizens, thereby enabling nontraditional data sources and data types to drive scientific discovery. Indeed, scientists today have access to an abundance of new data sources and data types, which we've classified as: (1)

archetypal or visible big data such as the large, well-curated, public data sets available from NASA and other large-scale research initiatives; (2) crowd-sourced or supernova big data such as the moment-to-moment data sets available from Twitter and the Internet of Things; and (3) long-tail or dark big data such as those hard-to-find data sets held by small scientific teams and amateur or citizen scientists. Efforts such as Wikipedia provide a model for open, collaborative, community-governed knowledge accumulation (Cohen 2014). A community-driven and community-governed knowledge network could revolutionize scientific discovery and move science in directions not imaginable today.

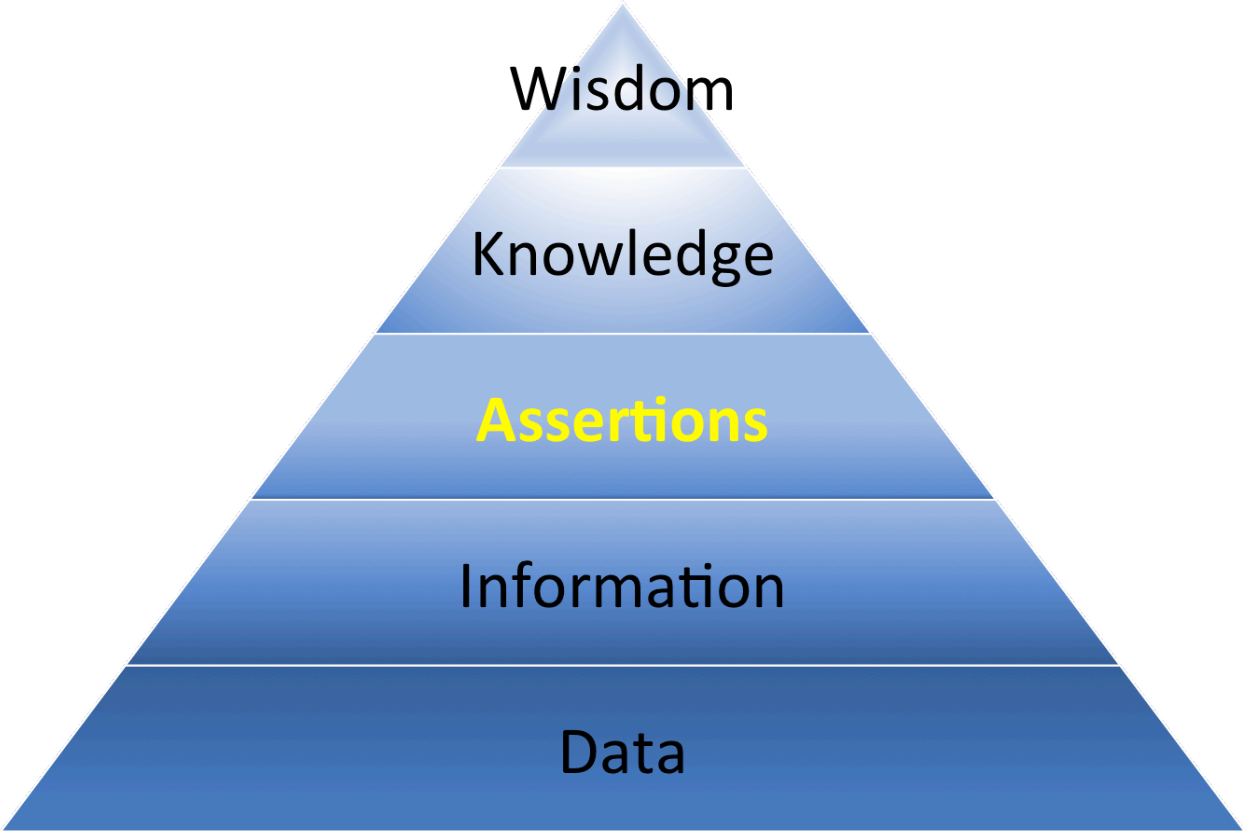

Figure 1. Conceptual framework for the proposed Collaborative Knowledge Network for Scientific Discovery.

Closing Considerations

The Collaborative Knowledge Network for Scientific Discovery, as proposed and envisioned herein, will require new scientific methods, approaches, and policies in order to fully realize its potential. For example, issues related to data provenance will need to be addressed, as will issues related to funding for ongoing development and long-term sustainability. The integration of knowledge assertions generated through different scientific approaches that have differing levels of certainty and expressiveness remains a fundamental challenge, but one where progress is being made. These are not trivial issues, as the creation of new knowledge from collective knowledge, particularly when automated, yields complicated issues regarding intellectual property and ownership. With multiple stakeholders involved in the envisioned knowledge network (e.g., academics, industry, government, the public), ownership issues will need to be resolved. Ongoing development and long-term sustainability bring related challenges that likely will require significant up-front community buy-in. Complex models for provenance and sustainability may need to be developed to account for conflicting interests. Furthermore, today's students and workers aren't being trained for the Fourth Paradigm of science and often lack the critical thinking skills required to objectively navigate through the abundance of data available today and make meaningful inferences about the world around them. This gap in training and workforce development will need to be addressed in order to ensure that a knowledge network is not misused.

Perhaps the most important consideration, however, is the development of new approaches to enable the critical evaluation and adjudication of scientific findings in the most traditional sense. For, in the absence of valid automated methods for drawing conclusions regarding scientific "truths," scientists will continue to face a data deluge without a rational path toward peer consensus, erroneous knowledge may be introduced and propagated, and the lay public will lose trust in science. Indeed, the public already has a tendency to mistrust science, especially when scientific findings go against intuition or personal belief (Achenbach 2015). Without a facile method to determine the legitimacy of scientific findings, any public mistrust in science will only increase. In the absence of public trust, science itself will suffer.

How to reference this paper:

Schmitt, C., Cox, S., Fecho, K., Idaszak, R., Lander, H., Rajasekar, A. and Thakur, S. (2015): Scientific Discovery in the Era of Big Data: More than the Scientific Method. RENCI, University of North Carolina at Chapel Hill. Text. http://dx.doi.org/10.7921/G0C82763

References*

Achenbach, J. (2015). Why do many reasonable people doubt science? *National Geographic*. http://ngm.nationalgeographic.com/2015/03/science-doubters/achenbach-text.

Achilleos, K. G., Kannas, C. C., Nicolaou, C. A., Pattichis, C. S., & Promponas, V. J. (2012). Open source workflow systems in life sciences informatics. *Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE).* Larnaca, Cyprus.

Alemida, P. (2013). The origins and purpose of scientific publications. *JEPS Bulletin*. http://blog.efpsa.org/2013/04/30/the-origins-of-scientific-publishing/.

American Chemical Society Publications. (2013). Be found or perish: writing scientific manuscripts for the digital age. *BIO JOURNALS.* American Chemical Society. http://pubs.acs.org/bio/ACS-Guide-Writing-Manuscripts-for-the-Digital-Age.pdf.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychol Methods.,* 2(2), 131–160.

Bell, G., Hey, T, & Szalay, A. (2009). Beyond the Data Deluge, *Science*, 323(5919), 1297-1298.

Berman, H. M., Gabanyi, M. J., Groom, C. R., Johnson, J. E., Murshudov, G. N., Nicholls, R. A., Reddy, V., Schwede, T., Zimmerman, M. D., Westbrook, J., Minor, W. (2015). Data to knowledge: how to get meaning from your result. *IUCrJ.,* 2(Part 1), 45–58.

Bertolucci, J. (2013). Big data's new buzzword: datafication. *Information Week.* http://www.informationweek.com/big-data/big-data-analytics/big-datas-new-buzzword-datafication/d/d-id/1108797?.

Betz, F. (2011), Origin of Scientific Method. In *Managing Science: Methodology and Organization of Research (*Innovation, Technology, and Knowledge Management Series, No. 9), by F. Betz. Springer Science+Business Media.

Bittrich, M. E. (2000). Biography of John Wilder Tukey. AT&T Bell Laboratories Communication. http://cm.belllabs.com/cm/stat/tukey/bio.html.

Bordes, A., Weston, J., Collobert, R., Bengio, Y. (2009). Learning structured embeddings of knowledge bases. *Association for the Advancement of Artificial Intelligence*, 301-306. http://www.thespermwhale.com/jaseweston/papers/AAAI11.pdf.

Bowers, S. (2012). Scientific workflow, provenance, and data modeling challenges and approaches. *J. Data Semant.,* 1 (1), 19–30.

Brillinger, D. R. (2002). John W. Tukey: his life and professional contributions. *Annals Statist.*, 30 (6), 1535–1575.

Burkhardt, F. (1996). Charles Darwin's Letters. A Selection. Cambridge University Press.
Buytaert, W., Baez, S., Bustamante, M., Dewulf, A. (2012). Web-based environmental simulation: bridging the gap between scientific modeling and decision-making. *Environ Science Technol.,* 46, 1971–1976.

Cohen N. Wikipedia vs. the Small Screen. *The New York Times.* February 9, 2014. http://www.nytimes.com/2014/02/10/technology/wikipedia-vs-the-small-screen.html?_r=2m.
Cold Spring Harbor Laboratory. (2014). Guide to open access. http://cshl.libguides.com/content.php?pid=222607&sid=1847688.

Collins, F. S., Varmus, H. (2015). A new initiative on precision medicine. *NEJM*, 372(9), 793–795. http://www.nejm.org/doi/pdf/10.1056/NEJMp1500523.

Curcin, V., Ghanem, M. (2008) Scientific workflow systems – can one size fit all? *Proceedings of the 2008 IEEE, CIBEC.* IEEE.

Diaconis, P. (2011). Theories of Data Analysis: from magical thinking through classical statistics. In: Exploring Data Tables, Trends, and Shapes, by D. C. Hoaglin, F. Mosteller, and J. W. Tukey. John Wiley & Sons, 2011.

Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., Mühlberger, A. (2015). Perception and presence on emotional reactions: a review of research in virtual reality. *Front Psychol.,* 6, Article 26.

Evans, J., Wilhelmsen, K., Berg, J., Schmitt, C., Krishnamurthy, A., Fecho, K., & Ahalt, S. (2015). Clinical genomics: how much data is enough?. RENCI, University of North Carolina at Chapel Hill. Text. doi: 10.7921/G0F769G9. http://renci.org/wp-content/uploads/2015/02/0215White-Paper-CLinicalGenomics-highres.pdf.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, Fall 1996, 37–54.

Gelman, A. (2004). Exploratory data analysis for complex models. *J. Computational Graphic Statist.*, 13(4), 755–779.

Gower, B. (1997). Scientific method: an historical and philosophical introduction. London, England: Routledge.

Guo, P. (2013). Data science workflow: overview and challenges. *Communications of the ACM blog.* http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext.

Hey T, Tansley S, Tolle K (Eds.) The Fourth Paradigm. Data-Intensive Scientific Discovery. Seattle, Washington: Microsoft Corporation; 2009.

Holzinger, A., Dehmer, M., Jurisica, I. (2014). Knowledge discovery and interactive data mining in bioinformatics – state-of-the-art, future challenges and research directions. *BMC Bioinformatics*, 15(Suppl 6), 1.

Jarke, M., Neumann, B., Vassiliou, Y., Wahlster, W. (1978). KBMS requirements of knowledge-based systems. In: Schmidt, J. W., & Thanos, C. (eds.) Foundations of Knowledge Base Management. Contributions from Logic, Databases, and Artificial Intelligence. Berlin, Germany: Springer, pp. 391-395. http://www.dfki.de/wwdata/Publications/KBMS_Requirements_of_Knowledge-Based_Systems.pdf.

Joppa, L. N., McInerny, G., Harper, R., Salido, L., Takeda, K., O'Hara, K., Gavaghan, D., Emmo, S. (2013). Troubling trends in scientific software use. *Science*, 340(6134), 814–815.
Kunkler, K. (2006). The role of medical simulation: an overview. The International Journal of Medical Robotics and Computer Assisted Surgery, 2, 203-210. http://onlinelibrary.wiley.com/doi/10.1002/rcs.101/pdf.

Leonhardt, D. (2000). John Tukey, 85, statistician; coined the word 'software'. *The New York Times*. http://www.nytimes.com/2000/07/28/us/john-tukey-85-statistician-coined-the-word-software.html.

Manyika, J., Chu, M., Bisson, P., Woetzel, J., Dobbs, R., Bughin, J., Aharon, D. (2015). The Internet of Things: mapping the value beyond the hype. McKinsey & Company. http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights/Business%20Technology/Unlocking%20the%20potential%20of%20the%20Internet%20of%20Things/Unlocking_the_potential_of_the_Internet_of_Things_Executive_summary.ashx.

McKie, T. How Darwin won the evolution race. *The Guardian*. June 21, 2008. http://www.theguardian.com/science/2008/jun/22/darwinbicentenary.evolution.

Mi H, Muruganujan A, Casagrande JT, Thomas PD. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc*., 8(8), 1551-1566
Montgomery, S. (2009). Charles Darwin & evolution, 1809 ~ 2009. Natural selection. Cambridge, UK: Christ's College, University of Cambridge. http://darwin200.christs.cam.ac.uk/pages/index.php?page_id=d3.

National Research Council. (2011). Toward Precision Medicine. Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, D. C.: The National Aademies Press. Contract No. N01-0D-4-2139. http://www.nap.edu/catalog/13284/toward-precision-medicine-building-a-knowledge-network-for-biomedical-research.

Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. From multi-relational link prediction to automated knowledge graph construction. Cornell University Library: arXiv:1503.00759. http://arxiv.org/abs/1503.00759.

Perra, N., Goçalves, B., Pastor-Satorras, R., Vespignani, A. (2012). Activity driven modeling of time varying networks. *Scientific Reports*, 2, 469, 1.

Perraud, J. M., Bai, Q., Hebir, D. (2010). On the appropriate granularity of activities in a scientific workflow applied to an optimization problem. *International Environmental Modelling and Software Society (iEMSs) 2010 International Congress on Environmental Modelling and Software Modelling for Environment's Sake, Fifth Biennial Meeting*. Ottawa, Canada.

Pugh, K., Prusak, L. Designing effective knowledge networks. *MIT Sloan Management Review.* September 12. 2013. http://sloanreview.mit.edu/article/designing-effective-knowledge-networks/.

Reshef, D. N, Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., Sabeti, P. C. (2011). Detecting novel associations in large datasets. *Science*, 334(6062), 1518-1524.

Samuelson, P. A. (1954). The pure theory of public expenditure. *The Review of Economics and Statistics,* 36(4), 387–389. http://www.ses.unam.mx/docencia/2007II/Lecturas/Mod3_Samuelson.pdf.

Singh, M. P., Vouk, M. A. (undated). Scientific workflows: scientific computing meets transactional workflows. http://www.csc.ncsu.edu/faculty/mpsingh/papers/databases/workflows/sciworkflows.html.
Smith, D. F. (2014). A brief history of gamification. *EDTECH, Focus on Higher Education*. http://www.edtechmagazine.com/higher/article/2014/07/brief-history-gamification-infographic.
Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *J Royal Society Med.*, 99(4), 178–182.

Southern, M. Google is looking into ways to rank sites based on accuracy of information. *Search Engine Journal.* March 4, 2015. http://www.searchenginejournal.com/google-is-looking-into-ways-to-rank-sites-based-on-accuracy-of-information/127394/#.

Spangler, B. (2003). Adjudication. *Beyond Intractability.* http://www.beyondintractability.org/essay/adjudication.

Speed, T. (2011). Mathematics. A correlation for the 21st century. *Science,* 334(6062), 1502-1503.

Trader, T. (2012). Taming the long tail of science. *HPC Wire*. http://www.hpcwire.com/2012/10/15/taming_the_long_tail_of_science/.

Wilhelmsen, K., Schmitt, C. & Fecho, K. (2013). Factors influencing data archival of large-scale genomic data sets: a mathematical formalism to comprehensively evaluate the costs-benefits of archiving large data sets. RENCI Technical Report Series, TR-13-03. University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. doi:10.7921/G0MW2F25. http://renci.org/technical-reports/factors-influencing-data-archival-of-large-scale-genomic-data-sets/.

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer*, 38(9), 25–32.

*All hyperlinks were last accessed on September 4, 2015.