Data-Drive Discovery: A New Era of Exploiting the Literature and Data

Ying Ding, Kyle Stirling, Indiana University, USA

Abstract: In the current data-intensive era, the traditional hands-on method of conducting scientific research by exploring related publications to generate a testable hypothesis is well on its way of becoming obsolete within just a year or two. Analyzing the literature and data to automatically generate a hypothesis might become the de facto approach to inform the core research efforts of those trying to master the exponentially rapid expansion of publications and datasets. Here, viewpoints are provided and discussed to help the understanding and challenges of data-driven discovery.

The Panama Canal, the 77-kilometer waterway connecting the Atlantic and Pacific oceans, has played a crucial role in international trade for more than a century. However, digging the Panama Canal was an exceedingly challenging process. A French effort in the late 19th century was abandoned because of equipment issues and a significant loss of labor due to tropical diseases transmitted by mosquitoes. The United States officially took control of the project in 1902. The United States replaced the unusable French equipment with new construction equipment that was designed for a much larger and faster scale of work. Colonel William C. Gorgas was appointed as the chief sanitation officer and charged with eliminating mosquito-spread illnesses. After overcoming these and additional trials and tribulations, the canal successfully opened on August 15, 1914. The triumphant completion of the Panama Canal demonstrates that using the right tools and eliminating significant threats are critical steps in any project.

More than 100 years later, a paradigm shift is occurring, as we move into a data-centered era. Today, data are extremely rich but overwhelming, and extracting information out of data requires not only the right tools and methods but also awareness of major threats. In this data-intensive era, the traditional method of exploring the related publications and available datasets from previous experiments to arrive at a testable hypothesis is becoming obsolete. Consider the fact that a new article is published every 30 seconds (Jinha, 2010). In fact, for the common disease of diabetes, there have been roughly 500,000 articles published to date; even if a scientist reads 20 papers per day, he will need 68 years to wade through all the material. The standard method simply cannot sufficiently deal with the large volume of documents or the exponential growth of datasets. A major threat is that the canon of domain knowledge cannot be consumed and held in human memory. Without efficient methods to process information and without a way to eliminate the fundamental threat of limited memory and time to handle the data deluge, we may find ourselves facing failure as the French did on the Isthmus of Panama more than a century ago.

Scouring the literature and data to generate a hypothesis might become the de facto approach to inform the core research efforts of those trying to master the exponentially rapid expansion of publications and datasets (Evans & Foster, 2011). In reality, most scholars have never been able to keep completely up-to-date with publications and datasets considering the unending increase in quantity and diversity of research within their own areas of focus, let alone in related conceptual areas in which knowledge may be segregated by syntactically impenetrable keyword barriers or an entirely different research corpus.

Research communities in many disciplines are finally recognizing that with advances in information technology there needs to be new ways to extract entities from increasingly data-intensive publications and to integrate and analyze large scale datasets. This provides a compelling opportunity to improve the process of knowledge discovery from the literature and datasets through use of knowledge graphs and an associated framework that integrates scholars, domain knowledge, datasets, workflows, and machines on a scale previously beyond our reach (Ding et al., 2013).

Scientific Discovery

Scientific discovery revolves around the process of problem solving. It either uses existing well-established methods to explore a new area or invents new methods to solve existing problems. Either way, it is a journey into unknown terrain. Trial-and-error remains the most common approach to testing new ideas, learning from failures, and, eventually, finding success. The problem-solving process can be viewed as a search for a path connecting the initial state and the goal state (Klahr, 2000). In cognitive science, a problem space contains the set of states, operators, goals, and constraints, and this problem space can be huge or small depending on whether you are on the right path to the final goal. The time to reach the final goal can be significantly shortened if the right tools are used.

How challenging the problem-solving process is also depends on the basic components in a problem space. The vagueness of some of these components can easily make scientific discovery purposeless. For example, one can have a task with a well-defined goal state (e.g., proving a mathematical equation) but a vague initial state, a task with a clear initial state (e.g., finding potential binding drugs for a given target) but an unclear goal state, or even a task with an ill-defined initial state and goal state (e.g., inventing a cool tool). More knowledge available to the problem-solver can significantly reduce the vagueness of basic components and set clear boundaries on the problem space. It is important to understand the problem space and foresee next steps.

Knowledge Discovery

Hypotheses can be generated from different sources. The dominant approach of developing a hypothesis in biology and medicine, for example, is through first-hand observation, which includes experimental data, electronic medical records, gene sequence data, and lab test results. The alternative method of generating a hypothesis from literature is viewed as a serendipitous process with great uncertainty—even more so now because the vast amount of published research contains a diversity of knowledge beyond what domain experts can humanly reason. Especially for researchers in transdisciplinary domains, it is no longer possible for domain experts in one domain to fully master the knowledge in another domain.

Mining literature to generate hypotheses is not confined to biology or medicine but can be done in almost any science. Publications are no longer just an output of research but rather a vital part of the scientific process. A significant number of associations between different biological entities (e.g., disease, gene, drug, side effect, and pathway) are scattered across millions of biomedical articles. Mining these documented associations can infer innovative associations and generate novel hypotheses, especially in the translational research.

Sciences are being conducted in a totally different way than 20 years ago. For example, biology is shifting from conventional biology to conceptual biology (Blagosklonny & Pardee, 2002) and moving further to systems biology (Oprea, Tropsha, Faulon, & Rintoul, 2007, Kell, 2006), in part because of a strong opinion that the conceptual review and systems thinking of available published knowledge should take its place as an essential component of scientific research. The world of ideas (i.e., published knowledge) interplaying with high-throughput experiments, computational modelling, and technology can generate intelligent hypotheses that will end the aimless fishing adventures in the conventional biology. New knowledge, derived from tens of thousands of publications and manually curated datasets, can be linked back to published knowledge to form a self-evolving ecological knowledge base (Mons et al., 2011). Predictions and experiments that were carried out for other reasons can be reused or revealed in a new context that fully embraces the holistic view of knowledge processing.

New ways of conducting research are in high demand, and examples of new methods can be found in many disciplines (Ding et al. 2013). Don Swanson's (1986) work about undiscovered public knowledge has had a wide impact on association discovery and demonstrated that new knowledge can be discovered from sets of disjointed scientific articles. Swanson's vision of the hidden value of the literature of science in biomedical digital databases is remarkably innovative for information scientists, biologists, and physicians (Swanson et al., 2001, Bekhuis, 2006). Literature-related discovery that mines knowledge in two disparate sets of literature have identified several non-drug approaches that can be used to halt or reverse the symptoms of multiple sclerosis, cataracts, and other chronic diseases (Kostoff, 2012). By combining PubMed literature and public datasets, Chen et al. (2012) can predict potential drug and target pairs based on publications and open datasets. The method performs extremely well in correctly identifying known drug-target pairs in the data and compares favorably with the established Similarity Ensemble Approach, or SEA, method (Keiser et al., 2009) for predicting new drug-target interactions as well as with the Connectivity Map, or CMAP, (Lamb et al., 2006) for associating drugs with changes in gene expression levels.

Scott Spangler and colleagues (2014) mined information contained in published articles to identify new protein kinases that phosphorylate the protein tumor suppressor p53. They successfully demonstrated that it is possible to automatically generate hypothesis for domain experts based on existing published scholarly articles. Even in humanity, Franco Moretti's distance reading solution tackles literary problems by applying computational methods to aggregate and analyze massive amounts of data and generate hypotheses. He advocates that distance reading is needed because nobody is able to read the 60,000 novels published in 19th-century England to understand Victorian fiction (Schulz, 2011). All of these example show that generating hypothesis by mining existing literature and open datasets can advance science and generate huge societal impact.

And while these examples highlight that human brains feature a great capacity for integrating information and recognizing patterns, computers are catching up. IBM Watson, a supercomputer, can process millions of articles, patents, Wikipedia pages, and datasets to facilitate research and diagnostic decision making in lung cancer treatment (Upbin, 2013). It also famously defeated two of the best human *Jeopardy!* players, Ken Jennings and Brad Rutter, in 2011, by parsing keywords in a large set of data to search for related terms as responses. While it is fast, it bears the disadvantage of a misunderstanding of the context of keywords. As well, the recent success of image recognition powered by deep learning outperforms humans (Thomsen, 2015). Project Adam, an initiative by Microsoft, can accurately identify a dog's breed based on a single photo. Soon, it will be possible for computers to provide nutritional information about a meal or help diagnose skin diseases (Chansanchai, 2014).

Translational Thinking

What Hal Varian called "combinatorial innovation" combines or recombines different component parts of previous innovations or ideas to generate new innovations (Mckinsey, 2009). Polymerase chain reaction, which earned Kary Banks Mullis the 1993 Nobel Prize in Chemistry, is the result of recombination of well-understood techniques in biochemistry (Brynjolfsson & McAfee, 2014). Dozens and dozens of publications that documented previous research outputs can be used to rigger translational thinking. These publications can be analyzed and mapped to show the scholarly landscape of unfamiliar fields to a researcher and suggest high-impact works to study and potential collaborators with whom to work. Other examples of combinatorial innovation include medical scientists who mine literature and open data to facilitate diagnostic decision making in cancer treatment, and healthcare professionals who study literature to generate practical guidelines for wound care (Flanagan, 2004).

More and more scientists are thinking about the translational value of their work. Sociologists apply the social concept of structural hole to understand scientific collaboration, and educators utilize literature as a scaffolding technique to enhance active learning. The transdisciplinary collaboration among material scientists, immunologists, and bioengineers has identified an implantable vaccine depot built from a polymer matrix that can kill cancer cells resulting in longer survival, which generates significant impacts on the well-being of society (Ali, Emerich, Dranoff, & Mooney, 2009). Publications and open datasets are ideal instruments to study the success of translational endeavors to further advance scientific innovation.

Transparent Analytics

The process of scientific endeavors, from data curation and analysis to discovery, should be transparent and easily accessible to every researcher so that replication can be easily done and the derived knowledge can be clearly interpreted (Editorial, 2009). Promoting transparency in science is crucial to ensure the reusability of knowledge, avoid reinventing the wheel, and make scientific discovery dedicated. Research, both quantitative and qualitative, is experiencing a methodological revolution (Moravcsik, 2014). Every researcher should make their work completely transparent to fellow scholars, and the process from data to conclusions should be interpretable and reproducible.

In recent years, the American Political Science Association (APSA; 2012) formally established transparency standards for qualitative and quantitative research by reinforcing the ethical obligation of researchers to facilitate the evaluation of their evidence-based knowledge claims through data access, production transparency, and analytic transparency. APSA proposed a new way of citing references called "active citation," which suggests that any citation in a scholarly publication should be annotated with an explanation on how the citation supports the knowledge claim and should include the hyperlink to an excerpt (c.a., 50–100 words) from the original source. These active citations can be located in a "transparent appendix" at the end of the document so that transparent data to conclusions for researchers are only one click away. This can generate a healthy scholarship by actively engaging researchers to establish rigorous research ethics to criticize, evaluate, and extend fellow scholars' research. Provenance has been introduced to data and workflows in scientific research to provide detailed documentation to enable scientific reproducibility. The World Wide Web Consortium has recommended a standard representation for provenance in a human readable and machine understandable way (Groth & Moreau, 2013). Transparency must be considered essential and achieved through active citation and provenance to further advance transparent sciences.

Connecting Intelligence

Machines taking their full place at the table of data-driven discovery is a significant step; these new participants make possible what was unimaginable 20 years ago. With machines, it is now possible to systematically collect, interdigitate, analyze, and disseminate publications and data in ways that will greatly impact the tradition of conducting research while providing powerful new resources that significantly advance the progress of both theoretical and applied research. Further, machines can be used to discover new knowledge and afford breakthroughs in current vexing research questions that can only be answered through transdisciplinary innovations.

The ever-increasing success in the application of full text indexing, taxonomies, and ontologies all dramatically improve the categorization and discovery of related content (Song, Han, Kim, Ding, & Chambers, 2013). The movie *The Imitation Game* has rekindled the memory of Alan Turing's success of machine intelligence (You, 2015). In the current data-enriched era, it may be the right time to revisit machine intelligence and connect machine intelligence with human intelligence. The next generation of artificial intelligence researchers is proposing a new Turing Championship to develop machines with a deeper understanding of the world (e.g., machine comprehension of grammatically ambiguous sentences, machine storytelling from pictures, and machine "humanness" that enables non-disruptive communication between machine and human).

The teamwork of machines and humans can make machines smarter and humans more efficient. The industrial revolution (mainly steam engine) bent the curve of human history and freed the physical muscle labor in the 19th century to allow for modern massive production. Now, the so-called Second Machine Age will bend the curve of human history again pretty soon by freeing the mental labor of humans. This will trigger massive innovation to bring scientific fiction into reality as these innovations are not only generated by human but also machines.

The combination of human and machine power can bring about new capabilities to compile, review, and mash-up related research entities and receive alerts on their activities and interactions, perhaps reaching a scale that was unimaginable 15 years ago. Much like the recent debut of driverless cars, distant scientific dreams could be realized in just a few years, demonstrating the power of the current data and machine progress (Brynjolfsson & McAfee, 2014). In the new world of scholarly analytics, attention and extraction of deeply covered content and findings are the pathways to golden discoveries. Gradually, advances in information technologies, such as the advent of open access, Linked Open Data, semantic publishing, and open science, will make it possible to gather, annotate, and acquire related publications and other data sources and from those discover related content, findings, and conclusions. This could lead to sudden discovery of unanticipated correlations and connections within an incredibly large and expanding research corpus. We are working on one of the oldest and toughest challenges associated with the combination of computer and human intelligence. The combinatorial innovation of human and machine intelligence will allow us to connect the dots for things that have been disconnected and accomplish through research what has been unimaginable, allowing us to dig the canal to connect data with knowledge.

References

American Political Science Association (APSA). (2012). *A guide to professional ethics in political science* (2nd ed., rev.). Washington, DC: Author. www.apsanet.org/Portals/54/APSA%20Files/publications/ethicsguideweb.pdf.

Ali, O. A., Emerich, D., Dranoff, G., & Mooney, D. J. (2009). In situ regulation of DC subsets and T cell mediates tumor regression in mice. *Science Translational Medicine*, *1*(8), 8ra19.

Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Library, 3*(2).

Blagosklonny, M.V., & Pardee, A.B. (2002). Conceptual biology: unearthing the gems. *Nature*, *416*(6879): 373.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. New York: W.W. Norton & Company Inc.

Chansanchai, A. (2014). Microsoft research shows off advances in artificial intelligence with Project Adam. Microsoft Blog, July 14. blogs.microsoft.com/next/2014/07/14/microsoft-research-shows-advances-artificial-intelligence-project-adam.

Chen, B., Ding, Y., & Wild, D. (2012). Assessing Drug Target Association using Semantic Linked Data. *PLoS Computational Biology*, 8(7): e1002574. doi:10.1371/journal.pcbi.1002574

Editorial (2009). Data's shameful neglect [editorial]. *Nature, 461*, 145.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. *PLoS One*, *8*(8): 1–14.

Evans, J. A., & Foster, J. G. (2011). Metaknowledge. *Science*, *332*(6018), 721–725.

Flanagan, M. (2004). Barriers to the implementation of best practice in wound care. *Wounds UK*, 74–84. www.woundsinternational.com/pdf/content_87.pdf.

Groth, P., & Moreau, L. (2013). PROV-Overview: An overview of the PROV family of documents. www.w3.org/TR/prov-overview.

Jinha, A.E. (2010). Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing, 23*(3), 258–263.

Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S. J., … Roth, B. L. (2009). Predicting new molecular targets for known drugs. *Nature*, 462(7270): 175–181.

Kell, D. B. (2006). Metabolomics, modelling and machine learning in systems biology: Towards an understanding of the languages of cells. *FEBS Journal*, *273*(5), 873–894.

Klahr, D. (2000). Exploring science: The cognition and development of discovery processes. Cambridge, MA: MIT Press.

Kostoff, R. N. (2012). Literature-related discovery and innovation—update. *Technological Forecasting & Social Change*, *79*(4), 789–800.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., … Golub, T. R. (2006). The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, *313*(5795), 1929–1935.

Mckinsey (2009). Hal Varian on how the web challenges managers [commentary]. www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers.

Mons, B., Van Haagen, H., Chichester, C., Hoen, P. B. T., Den Dunnen, J. T., … Schultes, E. (2011). The value of data. *Nature Genetics*, *43*(4), 281–283.

Moravcsik, A. (2014). Transparency: The revolution in qualitative research. *Political Science & Politics*, 47(1), 48–53.

Oprea, T. I., Tropsha, A., Faulon, J. & Rintoul, M. D. (2007). Systems chemical biology. *Nature Chemical Biology*, 3, 447-450.

Schulz, K. (2011). What is distance reading. New York Times, Jan 24. www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html?pagewanted=all&_r=0.

Song, M., Han, N., Kim, Y., Ding, Y., & Chambers, T. (2013). Discovering implicit entity relation with the gene-citation-gene network. *PLoS One*, *8*(12), e84639.

Spangler, S., Wilkins, A.D., Bachman, B. J., Nagarajan, M., Dayaram, T., Haas, P., … Lichtarge, O. (2014). Automated hypothesis generation based on mining scientific literature. Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 24–27, 2014, New York, USA.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1), 7–18.

Swanson, D. R., Smalheiser, N. R., & Bookstein, A. (2001). Information discovery from complementary literatures: categorizing viruses as potential weapons. Journal of the American Society for Information Science and Technology, 52(10), 797–812.

Thomsen, M. (2015). Microsoft's deep learning project outperforms humans in image recognition. Forbes, February 19.

www.forbes.com/sites/michaelthomsen/2015/02/19/microsofts-deep-learning-project-outperforms-humans-in-image-recognition.

Upbin, B. (2013). IBM’s Watson gets its first piece of business in healthcare. Forbes, February 8. www.forbes.com/sites/bruceupbin/2013/02/08/ibms-watson-gets-its-first-piece-of-business-in-healthcare.

You, J. (2015). Beyond the Turing test. *Science*, *347*(6218), 116.