UNLOCKING THE POWER OF BIG DATA AT THE NATIONAL INSTITUTES OF HEALTH

Meghan F. Coakley, Maarten R. Leerkes, Jason Barnett, Andrei E. Gabrielian, Karlynn Noble, M. Nick Weber, and Yentram Huyen

Bioinformatics and Computational Biosciences Branch, Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland

Abstract

The era of "big data" presents immense opportunities for scientific discovery and technological progress, with the potential to have enormous impact on research and development in the public sector. In order to capitalize on these benefits, there are significant challenges to overcome in data analytics. The National Institute of Allergy and Infectious Diseases held a symposium entitled "Data Science: Unlocking the Power of Big Data" to create a forum for big data experts to present and share some of the creative and innovative methods to gleaning valuable knowledge from an overwhelming flood of biological data. A significant investment in infrastructure and tool development, along with more and better-trained data scientists, may facilitate methods for assimilation of data and machine learning, to overcome obstacles such as data security, data cleaning, and data integration.

Introduction

The exponential growth of digital data presents both opportunities and obstacles in the domains of science, economics, business, healthcare, media, and virtually all aspects of our lives. Generating vast volumes of data in a multitude of formats is no longer a significant challenge, as an estimated 2.5 quintillion bytes of data are created every day, and 90% of the world's data has been created in the last 2 years alone.[1] The obstacle to making big data truly useful lies in the general shortage of tools and infrastructure to make valuable discoveries, and some federal government agencies have been slower to adapt to the new realities of the era of big data, relative to the private sector. According to predictive analytics expert Nate Silver, "In some cases, the government has the best data in the world, but not always the ability to use it."[2]

To address this shortfall, the Federal "Big Data" Initiative was launched in March 2012 by the White House Office of Science and Technology Policy, with an announcement of a $200 million investment in new research and development projects for big data analytics.[3] Several federal departments and agencies have committed to improving tools and techniques needed to access, organize, and make discoveries from huge volumes of digital data. Based on the recent emphasis on harnessing the power of big data, the Office of Cyber Infrastructure and Computational Biology at the National Institute of Allergy and Infectious Diseases (NIAID) used its third annual Bioinformatics Festival to address the current challenges in data science and big data analytics.

Held at the National Institutes of Health (NIH) in Bethesda, Maryland, on February 6, 2013, "Data Science: Unlocking the

Power of Big Data" featured a diverse group of experts from academia and the public and private sector. Presentation topics included analytics management and governance of big data generated from areas such as astronomy, protein mass spectrometry analysis, and clinical data mining. The speakers presented caveats to leveraging innovative strategies such as crowdsourcing and Twitter analytics. The symposium also featured a noteworthy presentation consisting solely of tweets, delivered by Dr. Michael Rappa of the Institute for Advanced Analytics at North Carolina State University; his talk sparked a flurry of tweets from the audience, a real-time demonstration of the accumulation of big data in the social media sphere.

Opportunities and Obstacles in Big Data Analytics

Biotechnology is advancing at a staggering rate, and the amount of available biological data is overwhelming; the online Molecular Biology Database Collection increased from under 300 entries in 2001[4] to 1,512 online databases in 2013.[5] Dr. Kirk Borne of George Mason University pointed out that with extremely large datasets, researchers can achieve significant statistical analysis of "typical" events while at the same time, "rare" phenomena—outliers that might otherwise be attributed to sampling error—may gain their own statistical significance, equating to new discoveries. Mr. Jonathan Epstein of the National Institute for Child Health and Development echoed these sentiments: "All data can tell you something, even if you think it's worthless." In this respect, Dr. Rappa described big data as "more opportunistic than scientific." In fact, many datasets or data analysis tools are not used for their intended purpose or by their intended audience, according to Dr. Rick Stevens of the University of Chicago and Argonne National Laboratories.

New approaches using machine-learning algorithms such as Bayesian and neural networks and hidden Markov models are a step in the right direction toward mining large datasets for valuable discoveries, but issues exist surrounding a fundamental lack of structure for data integration. In the biological sciences but also in other scientific fields, there is now a well-established tradition of depositing scientific data into various, often specialized, public repositories with different methods for curating and analyzing the data. The high degree of complexity and size of these datasets raises the urgent need for robust and scalable algorithms that link related datasets to maximize the potential for discovery.

Dr. Rappa also brought up the subject of data "cleaning" and the burden it imposes on data scientists. Dr. Stevens referred to a study by Dr. Adam Sadilek and colleagues at the University of Rochester, where artificial intelligence was used to analyze tweets from real-time Twitter feeds to predict the number of flu cases in New York City.[6] By selectively eliminating tweets by healthy people versus tweets from people actually sick with the flu, the group estimated flu cases with 90% accuracy. The ability to select relevant information from large datasets will add further value to data already being collected.

Much of the challenge of big data analysis also boils down to a lack of experienced staff that are cross-trained in informatics and life sciences and that are capable of fully capitalizing on the potential benefits of big data analytics. The McKinsey Report on Big Data estimated that there could be a shortage of 1.5 million skilled data scientists by 2017.[7] Harvard Business Review described data science as the "sexiest job of the 21st century,"[8] and demand could reach even 100 jobs per applicant.[9]

Life scientists must also be able to interpret the data, but selectively sharing knowledge across the disparate disciplines of life science and data science is no easy task. Presentation and interpretation of results by nontechnical domain experts is critical to extracting actionable knowledge. According to Dr. Matthew Clark of BioFortis, Inc., part of the challenge in exploring data is that "the hands-on analytics time to write […] code and specify clearly what you want is very time consuming." BioFortis is looking to bridge this gap with Qiagram, software that uses a drag-and-drop graphical interface to connect visually numerous data sources—combining elements like drugs, reactions, and outcomes—to simplify the process of multimetric clinical data queries.

> "THE MCKINSEY REPORT ON BIG DATA ESTIMATED THAT THERE COULD BE A SHORTAGE OF 1.5 MILLION SKILLED DATA SCIENTISTS BY 2017."

Overall, adoption of big data analytics has been slow in healthcare and life sciences research relative to the physical sciences and the enterprise sector. Commercial entities are incentivized to collect and analyze vast amounts of data, since knowing customer tendencies and desires enhances service and improves sales.[7] Analytics can reveal opportunities in the short-term to decrease operational costs and increase efficiency, while in the long-term it allows businesses to anticipate trends in the market and adapt accordingly. However, the biomedical research sector is largely constrained by informed consent and privacy concerns for health records and patient-specific data—issues that may be overcome in time as patient perceptions evolve along with data policies. Likewise, biomedical research entities may experience challenges collecting or combining large sets of sensitive data based on the need for security for unpublished results or information related to potential biohazards such as influenza or anthrax. In

the federal government, the defense and intelligence communities have been highly regarded for their use of big data,[10] and their solutions to guard against security breaches may present opportunities for NIH and other research institutes to model their own secure protocols and infrastructure.

Finally, it is impossible to ignore the cost burden to implement and maintain high-performance computing capabilities. Dr. Vasant Honavar of the National Science Foundation pointed out that affordable sequencing is irrelevant if a "$1000 genome […] requires a $1 million interpretation." This is particularly true in academic research environments, where expensive computing power is not always accessible. In his introductory talk, Dr. J.J. McGowan, head of NIAID's Office of Science Management and Operations, said that the "instability, competitiveness, and mechanisms and duration of research funding" discourages investigators from dedicating significant capital to computing infrastructure. Researchers are thus reliant on shared resources, which vary based on their institution's investment, or they must sacrifice speed and efficiency by operating on less costly or less advanced computing infrastructure. Developments in cloud computing are addressing this problem by providing remote and on-demand computing power, but computational analysis of data in a cloud environment has yet to become the standard in the scientific community.

"FINALLY, IT IS IMPOSSIBLE TO IGNORE THE COST BURDEN TO IMPLEMENT AND MAINTAIN HIGH-PERFORMANCE COMPUTING CAPABILITIES."

Advancing Data Science at the National Institutes of Health

As part of the Core Techniques and Technologies for Advancing Big Data Science and Engineering Initiative, the Data and Informatics Working Group of the NIH Advisory Committee to the Director recently recommended an NIH-wide strategic plan to create an adaptive and highly collaborative computing environment. The aim is to enable optimal use of big data within NIH intramural and extramural research communities, along with "a governance structure that aligns scientific leadership with resource management and oversight."[11] The Big Data to Knowledge (BD2K) and InfrastructurePlus programs are aimed at enabling the biomedical research enterprise to maximize the value of biomedical data and create a shared computational environment for the NIH community, respectively.

A critical aspect of both programs is to improve large-scale computing to facilitate data analysis. NIH has multiple scientific clusters to address computation needs at NIH. The challenge for NIH, NIAID, and healthcare and biomedical research in general is not only to make big data accessible but also to provide the tools and infrastructure to facilitate data mining across multiple data sources, for example, by creating or supporting data standards and ontologies. Centralization of data catalogs will promote data sharing, and virtualized cloud computing environments present a major step forward for democratizing access to the necessary cyber infrastructure.

Conclusion

NIAID's Data Science symposium emphasizes that many technical challenges of big data exist, such as volume and scalability as well as privacy issues and the heterogeneity and disparities of differently specialized datasets. Moving forward, skilled data scientists and bioinformaticians will be necessary in the life sciences, regardless of the field of research. Access to computing infrastructure and training will be crucial to allow nontechnical domain experts to interpret results and extract actionable knowledge. In the words of Dr. Borne, "Computational literacy and data literacy are critical for all." Collaborative efforts will be the key to adopting standards and ontologies and learning best practices for maximizing the exciting potential of big data.

Acknowledgments

We thank the speakers and attendees of "Data Science: Unlocking the Power of Big Data" for participating, Dr. J.J. McGowan and Mr. Mike Tartakovsky for sponsoring the event, and Dr. Darrell Hurt and Dr. Sandhya Xirasagar for comments on the manuscript.

The archived webcast, including roundtable discussion, can be found at http://videocast.nih.gov/launch.asp?17798

More information on federal government programs related to big data can be found in the White House OSTP Fact Sheet, "Big Data Across the Federal Government," which can be accessed online at www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf

Reference to any specific persons, commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply endorsement, recommendation, or favoring by the U.S. government. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. government and shall not be used for advertising or product endorsement purposes.

Author Disclosure Statement

No competing financial interests exist.

References

1. IBM. What is big data? – Bringing big data to the enterprise. Available online at www-01.ibm.com/software/data/bigdata/ (Accessed Feb. 2, 2013).
2. Konkel F. February 25, 2013. Nate Silver on big data's future: It's about attitude. *FCW.com*. Available online at http://fcw.com/articles/2013/02/25/nate-silver-data-insights.aspx (Accessed Feb. 27, 2013).
3. Office of Science and Technology Policy, Executive Office of the President. Press Release: Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments. March 29, 2012. Available online at www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf (Accessed March 1, 2013).
4. Baxevanis A. The Molecular Biology Database Collection: an updated compilation of biological database resources. Nucleic Acids Res 2001; 29:1–10.
5. Galperin MY, Fernández-Suárez XM. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucleic Acids Res 2012; 40:D1–D8.
6. Sailek A, Kautz H, Silenzio V. Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. Twenty-Sixth AAAI Conference on Artificial Intelligence. Toronto, Canada, July 22–26, 2012.
7. McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. June 2011. (Accessed March 25, 2013)
8. Davenport TH, Patil DJ. Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review. October 2012.
9. Indeed.com search for "data scientist" and "big data" job trends (accessed online March 2, 2013).
10. SAP/TechAmerica Foundation Quantitative Online Survey August–September 2012. Available online at www.techamericafoundation.org/content/wp-content/uploads/2013/02/SAP-Public-Sector-Big-Data-Report_FINAL-2.pdf (Accessed March 1, 2013).
11. Data and Informatics Implementation, Advisory Committee to the Director Meeting, December 7, 2012; Lawrence A. Tabak, DDS, PHD, Deputy Directory, NIH, DHHS. Available online at http://acd.od.nih.gov/Data-and-Informatics-Implementation-Plan.pdf (Accessed March 1, 2013).

Address correspondence to:

Yentram Huyen
Bioinformatics and Computational Biosciences Branch
Office of Cyber Infrastructure and Computational Biology
OSMO/OD/NIAID/NIH
10401 Fernwood Road, Room 2A11
Mail Stop Code 4821
Bethesda, MD 20892

E-mail: huyeny@niaid.nih.gov