OMICS A Journal of Integrative Biology
Volume 15, Number 4, 2011

DOI: 10.1089/omi.2011.0008

Communication and Data-Intensive Science in the Beginning of the 21st Century

Jack Faris,[1] Evelyne Kolker,[2] Alex Szalay,[3] Leon Bradlow,[4] Ewa Deelman,[5] Wu Feng,[6] Judy Qiu,[7] Donna Russell,[8] Elizabeth Stewart,[8] and Eugene Kolker[2,8]

Abstract

The advent of data-intensive science has sharpened our need for better communication within and between the fields of science and technology, to name a few. No one mind can encompass all that is necessary to be successful in controlling and analyzing the data deluge we are experiencing. Therefore, we must bring together diverse fields, communicate clearly, and build crossdisciplinary methods and tools to realize its potential. This article is a summary of the communication issues and challenges as discussed in the Data-Intensive Science (DIS) workshop in Seattle, September 19–20, 2010.

Introduction

Since the beginning of mankind, the human species has constantly struggled to answer questions about the world around us. As time goes on, the way we have approached answering questions has evolved along with us. Pioneering computer scientist Jim Gray saw the evolution of science as beginning with experimentation, moving to theory, and simulation, and then characterized the fourth stage as data-intensive science. In this fourth paradigm, how we answer questions is beginning to shift from creating hypotheses that are tested over and over, to analyzing large data sets to provide answers and conclusions. Even the very means of communication is changing as paper gives way to digital publishing.

As Voltaire once said, "If you wish to converse with me, define your terms." This idea may sound simple, but it is complex to enact, particularly when communication is occurring across disciplines. The reality now is that with data-intensive science, communication between biologists, computer scientists, statisticians, and many other disciplines is necessary and crucial to success. Our working group outlined four main constituencies that must communicate, both within and between themselves.

Five Major Constituencies of the Fourth Paradigm

1. Scientists and Researchers
2. Funders and Policymakers
3. Students
4. General public
5. Industry

In this report, we will be discussing the current state of communication in these groups, current barriers to communication, and our hopes and goals for the future state. We will also discuss the importance of communication and representation to society at large. With this report, we hope to convey the importance of communication in DIS and what we believe to be imperative changes that must be made in order to achieve success in this fourth paradigm of science.

Scientists and Researchers

The first key constituency includes scientists and researchers. The current state of communication between scientists and researchers, these being broad umbrella terms for all the myriad of disciplines involved with data-intensive science, is one where communication is not clear, standardized,

[1]Pacific Northwest Diabetes Research Institute, Seattle, Wahsington.
[2]University of Washington, Seattle, Washington.
[3]The Johns Hopkins University, Baltimore, Maryland.
[4]Hackensack University Medical Center, Hackensack, New Jersey.
[5]University of Southern California, Los Angeles, California.
[6]Virginia Tech, Blacksburg, Virigina.
[7]Indiana University, Bloomington, Indiana.
[8]Seattle Children's Research Institute, Seattle, Washington.

or completely open. As people from different disciplines come together to discuss their common projects, they come from separate backgrounds, experiences, and educational training. There is a lack of understanding because the terminology used is not standardized, not only between different disciplines, but also within disciplines. There is often a lack of: (1) comprehension of one another's fields, (2) capabilities within each other's fields, and (3) trust. Furthermore, collaboration among different universities, labs, groups, etc., is rare due to funding policies that do not encourage collaboration nor reward it. This is what we see as the current state of communication among researchers and scientists. The desired future we envision is one where there is a break down of silos and, instead, researchers communicate and collaborate across disciplines. We would like to see the establishment of frameworks that increase productivity of creative science, intellectual property laws that support collaboration, and common data repositories that are open source and can be accessed by anyone.

Although it is not yet clear how to achieve this vision, it is clear that there are barriers standing in the way of achieving it. Although some of the barriers are technological, such as creating easily accessible data repositories through cloud computing, it is evident that to create a truly nonredundant and successful data-intensive science existence we must begin with communication and the changing of perspectives and practices. First, funding policies need to be adapted to support major collaborative efforts. Funding needs to be awarded, at least in part, based on the merit of proposed joint efforts rather than the number of citations a principal investigator may have. Crossdisciplinary research needs a clearly established home, an innovation that may require some institutional restructuring. Academic promotion guidelines must also be examined to make sure collaborative efforts are fully recognized. It is absolutely vital for scientists and researchers to overcome the communication barrier and begin to establish standards for data formats, data sharing, and data storage and collection, including standards of high quality data and citation procedures for researchers using other researchers' data. Communication is also needed to overcome technology barriers that we currently face, and will be instrumental in the education of students and future researchers to properly prepare them for being citizens of fourth paradigm science (Fig. 1).

Funders and Policymakers

Another principal constituency in which communication changes need to occur is funding agencies (public and private) and policymakers. Currently, the system rewards researchers and their labs for the amount of work they have done. If they see the research initiatives that are proposed as safe promises on return, they will reward these projects time and time again rather than collaborative efforts that seek to achieve higher goals and, thus, are riskier. The optimal future state mindset for funding agencies and policymakers is one in which a cultural shift has occurred and data-centric research is the norm rather than the exception. As such, we would like to see a policy of significant rewards and widespread recognition for those who are conducting data-intensive research. In addition to this, we would like funding agencies and policymakers to support high-quality data collection and curation as well as commit to funding and supporting long-term data preservation through accessible repositories. As funding agencies begin to support collaborative, data-intensive science, there will be a shift in many agendas and project proposals to these kinds of initiatives.

Currently, we see barriers to success coming from the cultural paradigm of one PI, one grant. This blocks high-risk, data-intensive projects, which can reap the greatest benefits and advancement. As the amount of available funding decreases while research initiatives increase, the competition for available money increases dramatically. This leads to more time spent writing grants and applications than the actual experimentation and research that leads to solutions and answers. Policymakers also create roadblocks to cross-disciplinary collaboration by creating and encouraging the single discipline departments in science, rather than supporting the coming together of individuals from different fields to approach common inquiries. Additionally, if academics feel they will not achieve tenure nor be promoted unless they specialize in one field, very few academics will pursue a double-pronged approach in their studies (Fig. 2).

Students

The third constituency that needs improved communication is the student group. Clearer communication is needed by the individuals and groups training students as well as those

FIG. 1. Communicating DIS.

FIG. 2. Collaborative science.

groups that provide and create opportunities for students. Students will be the next generation to be involved in data-intensive science and, as such, must be prepared for a new approach to science. This will require changes to the curriculum to train students to be pi-shaped, meaning they have two disciplines they know a lot about and pieces of knowledge across a variety of disciplines. The creation of scholarships, fellowships, and opportunities in various disciplines in data-intensive science will encourage students to become experts in bridging the gaps between disciplines. Students need to be told that there is a need for fourth paradigm trained researchers, and that there will be rewards for such training. Incentives such as increased funding, higher pay, and more career choices will lead to more students following this new path. This, in turn, will result in more interested, trained, and better equipped professionals to address the problems faced by current professionals in data-intensive science.

General Public

Communication with the general public is key to creating a positive representation of science in society. This begins with creating a greater understanding of what scientists and researches do, as well as how their success can lead to improvements in society, enhancing people's daily lives, and progress in terms of people's health. Such a cultural shift can be achieved by creating ad campaigns that target the general public, as well as increasing the media presence of scientists and researchers. Everyone should be encouraged to be a citizen scientist, from collecting samples for researchers out in the wilderness to lab experiments in a child's classroom. This involvement will be a crucial part of changing how science is viewed in society. The hope is that such communication will alleviate any antiscience feelings that stem from low science literacy and lead to broad-based support for big science, including data-intensive science.

Industry

Industry has had to confront the challenges of large datasets head on. More customers mean more information, and the recent trend has been to gather extra information for more specialized marketing. Amazon exists because of its incredible ability to track information and products. Wal-Mart has an added layer of complexity with brick and mortar stores. Google is well aware of the power of data and the need for security with miles of security fencing outside of its data warehouses. In all cases, customers expect their information to stay secure, and any efforts to breach that security must be taken extremely seriously. However, these companies do not talk much about their systems and the amount of data they have stored. Citing the security issues mentioned above, they only allude to the massive data centers and terabytes of data hidden within them.

The success of these companies rest in part on the trust a customer gives them, and communication aids trust. However, sometimes their data-gathering efforts can back fire and trust is eroded. The FCC has investigated data gathering by darling Google as it pursued its street-mapping project. Apple and Facebook, two other darlings of the computer/social networking era, have had to deal with skittish users as Apple introduces Ping, and Facebook wrestles with making its privacy updates as simple as possible.

Conclusion

In conclusion, communication is extremely vital to any discipline and even more so when disciplines come together. No matter the technological advancements that occur in science, communication among individuals is imperative to progress in science. By changing the paradigm from research driving data to data driving research, as well as encouraging collaborative, crossdisciplinary data-centric, and open projects with many groups coming together, we can begin to fully live Jim Gray's vision of the fourth paradigm. Data-intensive science can lead to great gains and advances in science and society. Yet, we must be mindful of the challenges these large data sets bring and balance the need for reproducibility of experiments with the need to avoid wasteful, time-consuming duplication, as discussed in the report of the National Science Foundation Office of Cyberinfrastructure Task Force on Data and Visualization (NSF_OCI_TFDV, in press). Clear communication between all constituencies, be it among scientists and researchers, funding agencies and policymakers, students and the general public, will succeed in making data-intensive science effective, eye-opening, and ultimately, provide answers for our many queries and questions.

Acknowledgments

This policy report and DIS workshop were supported by SCRI and NSF Grant DBI-0969929 to E. Kolker (Principal investigator). The views expressed in this article are entirely personal opinions of the authors and do not necessarily represent positions of their affiliated institutions or NSF.

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

Reference

National Science Foundation, NSF (2011). Office of Cyberinfrastructure, Task Force on Data and Visualization (in press).

Address correspondence to:
Eugene Kolker, Ph.D.
Seattle Children's Research Institute
1900 Ninth Avenue
C9S-9
Seattle, WA 98101

E-mail: eugene.kolker@seattlechildrens.org