OMICS A Journal of Integrative Biology
Volume 15, Number 4, 2011
© Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2011.0009

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

Fredric Wolf,[1] Russ Hobby,[2] Sonya Lowry,[3] Andrew Bauman,[4] B. Robert Franza,[5] Biaoyang Lin,[6] Sean Rapson,[4] Elizabeth Stewart,[4] and Eugene Kolker[1,4]

Abstract

Data-intensive science will open up new avenues to explore, new questions to ask, and new ways to answer. Yet, this potential cannot be unlocked without new emphasis on education of the researchers gathering data, the analysts analyzing data and the cross-disciplinary participants working together to make it happen. This article is a summary of the education issues and challenges of data-intensive sciences and cloud computing as discussed in the Data-Intensive Science (DIS) workshop in Seattle, September 19–20, 2010.

Introduction

The education working Group (WG) of the Data-Intensive Science Workshop came together to identify issues and needs within the data-intensive sciences with regard to education. Data-intensive science is a new science. As a result, many scientists and lay persons alike do not have an adequate understanding of its application and potential. As data-intensive approaches and methodologies enable scientific discovery in the beginning of the 21st century, it is imperative that all scientists and lay persons be fully aware of the implications of these scientific discoveries and methodologies.

Current State

To explore the factors influencing the current state, this WG created a table with dimensions: (1) different populations of learners based on their area of expertise (rows) and, (2) different functions pertinent to Data-Intensive Science (DIS) (columns). The information in Table 1, although not all-inclusive, allowed the WG to think about the status quo and discuss education in relation to DIS.

The WG generally agreed that the greatest need is further education about cyberinfrastructure (CI) as an entity that must be developed and a field with many facets.

Barriers

CI consists of many components that do not have well-defined relationships that could be used to structure the overall subject. Indeed, each component of CI (data acquisition, management, storage, mining, and visualization to name a few) has been developed independently and patched together. Education is needed to further the development of a cohesive CI and to enable its use.

CI can be looked at as an elephant. Each component can be identified independently but the sum of them all creates a functional elephant (Fig. 1).

In this example, education would be most effective by first looking at the entire elephant and its overall function and then delving into each of the components. Education would consist of teaching the use, operation, or development of CI. Education in all three aspects requires an understanding of the complete animal to varying degrees of detail (Figs. 2 and 3).

To build this infrastructure, the groups listed in Table 1 must educate each other. Scientists must discuss the needs of their fields to receive full benefit from a cohesive CI. Lay persons must be consulted so that the public, which primarily funds these endeavors, is involved and fully enabled. Certainly, developers of both hardware and software must be educated as to the needs of scientists and lay persons. Further, developers must educate those groups to develop an understanding of currently feasible technology and the strengths and limitations of different approaches.

Future State

We need to teach people how to use CI with a new system in which all the components work in a coordinated fashion

[1] University of Washington, Seattle, Washington.
[2] Internet2, Davis, California.
[3] University of Arizona, Tucson, Arizona.
[4] Seattle Children's Research Institute, Seattle, Washington.
[5] Myoonet, Seattle, Washington.
[6] Swedish Medical Center, Seattle, Washington.

Table 1. Current State of DIS Education

	Use of large data applications	*Building of large data applications*	*Use of compute, data, and network infrastructure*	*Infrastructure system design and maintenance*
Lay community (educational pipeline, media, politicians)	Use (Facebook, iPhone apps, g-mail) Exposure through media	Hobbyist	N/A	Unaware
Biological researcher	Beginning to think about it	Not building but providing feature requests, functional requirements	N/A	Unaware but need to know
Bioinformatician	Basic experimentation (Amazon S3, EC2, EBS)	Building tools, applications, mashups	Rare experimentation	Unaware but starting to know
Software systems engineer	Supercomputers, Basic exp. w/cloud computing	Building applications, mashups	Basic knowledge (e.g., knowledge of single systems but not distributed ones)	Starting to sell the idea
Hardware systems engineer	Supercomputers, Basic exp. w/cloud computing	N/A?	Educated users, need to keep current	Specialized hardware for specific applications
Archivists/curators	Increasing awareness, inability to implement	N/A	?	Sense a need
Statisticians/ mathematicians	Application and dissatisfaction with current methods	Modifying current methods to be scalable to large data sets, development of new methods (e.g., map reduce)	N/A? Improvements could be adopted with greater understanding	Don't know don't care

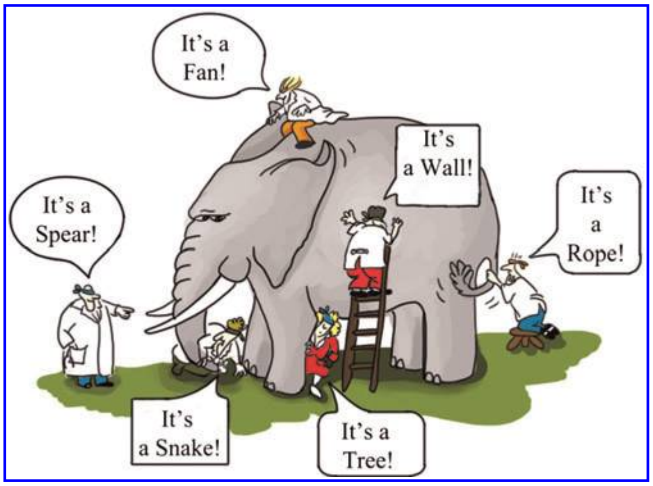

FIG. 1. A functioning CI elephant.

with security, reliability, and good performance. A nationwide integrated system with a common user interface (UI) structure would expedite learning of new applications through familiarity. This future system must: (1) allow access regardless of geographic location; (2) be flexible to allow innovation, change, and growth as needed; (3) foster integration and collaboration across agencies, offices, and resources; and (4) encompass software as well as user/developer communities.

The future state CI will provide a number of significant benefits:

- Scientists will be able to both educate and be educated about available high performance computing resources.
- Lay persons will be able to use educational portals to learn about and use resources in a secure and understandable environment.
- Researchers from all fields will be able to take full advantage of data repositories, applications, and high-speed computing resources without geographic restriction, and then additional scientific findings to the current data repositories without difficulty.
- Bioinformaticians will be able to develop standardized tools and make them widely accessible, adding significantly to the rate of scientific progress.
- Software and hardware engineers will be able to develop, implement, and improve a standardized system that is easy to understand and manipulate.

FIG. 2. We have parts, but what parts?

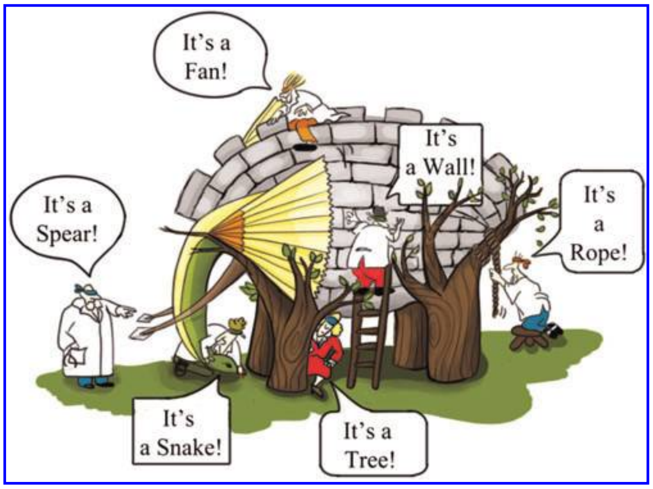

FIG. 3. Parts can be made to look like an elephant. Making them function that way is the next step.

- Archivists and curators will be able to easily access data sets and mine data sources.
- Mathematicians and statisticians will be able to capitalize on increased computer power and larger data sets.

Representation and Engagement with Society

The changes needed for a full realization of DIS impact many sciences. Many different, ongoing organizations tackling these problems could have a bigger impact if they worked together. Thus, a consortium is needed to realize this vision. As research disciplines truly start to depend on each other, they need to learn to share data and skills while working cooperatively, regardless of organization membership or geographic locations. National Science Foundation (NSF) and other funding agencies can recognize the need for cross-disciplinary education and provide leadership and funding to implement the vision.

Conclusion

Change, growth, progress, and innovation are all integral to the scientific endeavor. Without them, the "Why?" and "How?" so often asked would not be answered. To take full advantage of the opportunities change can bring, the groups implementing the change must educate themselves on how to best effect the change and how to best enable the use of the change. A change made due to ignorance can easily lead to wasted time and money. We do not have either to waste.

Acknowledgments

The authors would like to extend their appreciation to Monica Lent and Kathleen Kennedy of the iPlant Collaborative, University of Arizona, for contributing the images and for editorial review.

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.

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