Graduate students’ vision for EarthCube

S. de Larquier, N.A. Frissell, A.J. Ribeiro, E.G. Thomas
Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, Virginia

1. Introduction

As graduate students, we envision EarthCube as a tool that could significantly ease the transition from student to researcher. This paper is derived from the experience of senior graduate students (over three years in graduate school), and is meant to reflect the difficulties encountered in dealing with data intensive science using the training and tools available to young researchers. New graduate students face numerous challenges as they begin a career in research, many of which we believe are unnecessarily time consuming and mundane. These tasks include obtaining data from disparate sources, working to understand undocumented data, writing code to visualize the data, and searching for the best techniques for managing and interpreting a large, heterogeneous data set.

It is our opinion that too much of a graduate student’s time is spent dealing with cyberinfrastructure issues rather than working within their elected field. This paper presents ideas designed to improve a graduate student’s transition to efficient research practices, which would also benefit the entire data-driven geoscience community. In the following sections, we identify three important issues to be addressed, and then offer potential solutions to these issues.

2. Issues encountered in a graduate student’s path to data handling

2.1 Issues with data usability

The first major issue that we would like to address is a simple one: the task of accessing and understanding data sets. Only recently have massive amounts of data, ease of data access and advanced computing resources been available to both general public and the research community. However, data access is hindered by the lack of common data access points. If data is posted on the web, it is either spread out among different websites or one must contact individual PI’s for access. Even after access to the data is gained, it is often undocumented and exceedingly difficult to decipher without prior expertise. Once data formats are finally understood by a user, time must then be spent writing code to read the file. These three steps to acquire data are required every time a new data set is studied and are unnecessarily repetitive since they are performed by each new user.

2.2 Issues with accessibility of analysis tools

The second issue that we would like EarthCube to address also focuses on users not repeating similar basic tasks in data processing. We believe that pre-existing tools to visualize and manipulate data sets should be shared throughout and between communities. Individual groups have developed techniques and tools for visualizing and manipulating their data set. Frequently, new data users outside of the groups also have to spend time developing tools which are very similar to those already in existence. This accounts for a large part of a researcher's time, which could be spent analyzing the data rather than plotting it. Additionally, new tools are constantly being developed for various purposes, e.g. visualization, data-mining, and modelling. All too often, these routines are only available to the individual who developed them, instead of being shared amongst researchers with common goals.

2.3 Issues with graduate student training

The third and final issue addressed in this paper is the fact that there is a steep learning curve associated with entering the field of research for new graduate students. Many current educators/researchers were trained in techniques designed to maximize knowledge from minimal amounts of data, i.e., case studies. These sorts of techniques are what we are taught in classes, and although they remain very useful, larger statistical studies are also important to study scale and frequency of phenomena. Increasingly, science is performed by examining large amounts of data for trends and then deducing theory from these trends, yet we feel that graduate students are not adequately trained in this approach. This all means that students end up re-imagining methods and techniques previously developed by other experts.

None of the issues which have been mentioned involve inventing new solutions. Rather, we believe the solutions to the problems lie in collecting sparse solutions which should be shared throughout the geosciences community.

3. Proposed approaches to solving these issues

3.1 Improving data usability

Concerning usability of data, our approach for solving the first problem comes in the form of cyberinfrastructure. First, we imagine EarthCube as a giant clearinghouse of data, similar to the NOAA website or CDAweb, but reaching across several disciplines.

Such a data site should provide at least the following mandatory feature: sufficient documentation to identify data formats including a short description of available fields (no abbreviations). This would make it much easier for users to understand and properly manipulate the available data. Another requirement for data on EarthCube should be that it is kept up-to-date. Many instruments operate continuously, providing new data every second. It would be difficult for EarthCube to be updated in real time, but updates on a scale of weeks or months should be a reasonable requirement. Any revision in file formats should also be documented and readily available to the community along with the corresponding data files. Because this would require extra effort from data providers, there could be an incentive system in place rewarding PIs and groups involved in good data sharing practices.

On a higher level, tools for reading file metadata and browsing its contents would be very useful if provided with the data. For example, a simple program that could give a quick snapshot of a data file's contents, without having to read the entire file. Further down the road, it would be a great achievement if the geosciences community as a whole could move towards common data formats.

3.2 Improving accessibility of analysis tools

When considering the issue of sharing data visualization and manipulation tools throughout the community, we imagine a massive toolkit with routines written by experts on specific data sets. Currently, every individual group develops and maintains their own software libraries which could be included in a global EarthCube toolkit. It would be important for such a toolkit to be expandable, i.e., as new routines are written and old ones improved, it should be possible to add them in. Of course, the process of adding to the existing library should be controlled to prevent a global spreading of errors. First, it is imperative that all code be well-documented. And second, there should be a peer-review process where one or more existing users thoroughly test and evaluate the new code before it is added. Some models for managing this toolkit could be inspired from large open-source projects, such as Firefox and the Linux kernel, as well as existing scientific projects such as the THEMIS software toolkit.

An ideal interface for basic usage would be a web-interface. This would require no installation of software, which can get exceedingly complicated, and would be platform independent. In our perfect scenario, a user would be able to visualize data and run models through a web page, and download open-source code if desired. If that user were then to develop some new feature for handling the data, it could in turn be included into the EarthCube toolkit. A global shared toolkit would also allow for easier reproduction of published results, considerably helping new research and results to be built upon existing ones.

3.3 Improving graduate student training

To address the issue of training new graduate students in handling large data sets, we propose the development and implementation of a curriculum which will provide practical experience in tools that are designed to extract data from large data sets. This curriculum would be taught in one or two semester-long classes involving teachers from multiple disciplines, including areas such as geoscience, informatics, and computer science. Based on the existing model of bioinformatics, classes should be designed to address the issues and demands of geoinformatics. The classes should provide introductions to specific techniques that are designed to allow researchers to manage, understand, and process large, heterogeneous data sets. These classes should be project-based using real data sets and allow for a significant amount of practical student experience. As an alternative to classes, a seminar or webinar series could be developed. Additionally, this would be an excellent opportunity to train new students in the use of the new EarthCube facilities.

4. Conclusion

In conclusion, we have identified three areas in need of improvement from a geoscience graduate student's perspective. In response, we have identified solutions which would allow EarthCube to ease the transition from student to researcher. We believe the ideas proposed here would make research more efficient, more interactive, and increase collaboration within and across specific fields.