Data-driven Geoscience: Key Issues and Recommendations from the 2015 Unidata Users Workshop

by:

Michael Baldwin, Purdue University

Steven Lazarus, Florida Institute of Technology

Kevin Tyle, University at Albany

Joshua Young, Unidata

Sen Chiao, San Jose State University

Ibrahim Demir, University of Iowa

Kevin Goebbert, Valparaiso University

Gretchen Mullendore, University of North Dakota

Sam Ng, Metropolitan State University of Denver

Kimberly Hoogewind, Purdue University

Mohan Ramamurthy, Unidata

Submitted to the Bulletin of the American Meteorological Society

August 17, 2017

Corresponding author address: Michael Baldwin, Dept. of Earth, Atmospheric, and Planetary Sciences, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907

Email: baldwin@purdue.edu

What: The 2015 Unidata Users Workshop, organized by the Unidata Users Committee and the Unidata Program Committee, with support from the National Science Foundation, was entitled: Data-Driven Geoscience: Applications, Opportunities, Trends, and Challenges. Session topics involved four main themes associated with the rapidly-changing landscape of managing and analyzing geoscience data: Python, cloud computing, big data, and data management. Based on the presentations and discussions at the workshop, the Unidata Users Committee identified several key issues and recommendations, which are presented in this summary.

When: 22-25 June 2015

Where: Boulder, Colorado

The 2015 Unidata Users Workshop, organized by the Unidata Users Committee and the Unidata Program Committee (UPC), with support from the National Science Foundation (NSF), took place in Boulder, Colorado, from 22 to 25 June 2015. Around 75 participants attended the event and represented numerous organizations, mainly educational institutions from the United States. Similar workshops have been held on a triennial basis over the past three decades to provide a forum for addressing issues related to enhancing education in the atmospheric and related sciences and have been an important venue for the Unidata community to come together to share ideas, techniques, and course materials.

Atmospheric science, along with numerous other disciplines, finds itself in the middle of major changes regarding data and computing. Data volumes have been increasing and are expected to explode in the near future as new observational capacities are implemented and numerical model output grows in both size and scope. Demands from funding agencies, publishers, and scientific societies to make research more open by sharing data, publishing code, and providing results that are easily replicated are additional challenges for our field. Educators, researchers, and students face daunting challenges associated with an ever-shifting landscape of computing paradigms, software tools, and programming languages. Cloud computing is supplanting traditional data management techniques with new mechanisms for processing and sharing data. To help address these challenges, this workshop focused on helping members of the community become more aware of emerging technologies and software in the age of data-driven science.

The workshop consisted of a combination of keynote speakers, in-depth presentations, student posters, and hands-on working sessions that generally covered one (or more) of four main topics: Python, big data, cloud computing, and data management. Many of the workshop logistics were inspired by Software Carpentry workshops (http://software-carpentry.org/workshops/operations/) which are notable in how they help to establish an environment that is supportive of learning new programming skills, a process that can often be quite intimidating. For instance, a pre-workshop survey of the participants was implemented to determine the level of experience with Python and other programming environments, allowing the presenters to better understand their audience prior to the start of the workshop. The workshop utilized a Github site (https://github.com/Unidata/unidata-users-workshop) to help manage the various Python environments that presenters needed to use for their hands-on sessions. The participants installed custom “environments” that could be activated as needed, preventing conflicts between specialized environments among different presenters. Time was allotted at the start of the workshop to allow participants to install these custom environments on their laptops, with assistance from Unidata staff if needed. The room was configured with multiple “pods” of seating to allow participants to view the main presentation screen as well as facilitating working in smaller groups. Since the majority of the presenters included hands-on demonstrations, we attempted to collect feedback from the participants in real-time by using “sticky notes” of different colors. Participants were asked to place a red sticky note on the front of their open laptop if they were having a problem or had a question, and a green sticky-note to indicate that they were following along and things were working smoothly. Feedback was also collected continuously during the workshop by accumulating comments written on the sticky notes as well as those submitted through a web-based form (http://www.unidata.ucar.edu/community/surveys/2015users_workshop/survey_results.html).

A recommendation to consider using Python as a primary programming language appeared as a common theme from the workshop presenters. Python users reported much less frustration when performing complex data workflow in contrast to the older approach of using numerous computing environments (such as: shell scripts, FORTRAN code, visualization software, web server) to process and share data. Python is quickly gaining traction as a platform for scientific computing across a wide variety of disciplines, and many Unidata community members are enthusiastic participants. Recent AMS Short Course events and Unidata training workshops related to Python have been in high demand. Numerous presentations were made in this workshop from scientists reporting on their experiences using Python in their scientific workflows as well as progress in developing Python tools for data analysis and visualization. Python has become a very popular computing language; in June 2015 it was ranked #6 on the TIOBE Programming Community Index (http://www.tiobe.com/tiobe_index, moved to #5 as of August 2017), behind languages such as Java and C, but ahead of Matlab and R. This popularity is seen as an advantage for using Python, since there are numerous packages available for data analysis and visualization, for example. However, some dissenting views regarding the adoption of Python were expressed at the workshop. The importance and challenges of community development and digital curation in the rapidly-changing area of Python computing were discussed. A complete and mature meteorological data analysis and visualization package, such as Unidata's Integrated Data Viewer, is not currently available in Python. Keeping abreast of new developments and modules in Python is a difficult challenge. In addition, issues regarding how computing should be incorporated in the undergraduate curriculum as well as effective strategies for teaching scientific computing were discussed. Development of software tools for data analysis and visualization which are both powerful and easy to use remains a challenge for this community.

Cloud computing appears to be a potential avenue for researchers to gain access to significant computing resources beyond the traditional supercomputing center. However, the "pay as you go" model of cloud computing might prove difficult for many institutions. There is also a steep learning curve for a user to be able to take advantage of these resources in their data workflow. Another recommendation from the workshop is to adopt virtualization/container technologies (e.g., Docker) for cloud computing for easier deployment and improved reproducibility. Software containers wrap code with a complete system that contains everything needed for it to run: system tools, libraries, etc. This approach ensures that the software will always execute in the same way, regardless of the environment in which it is running.

Another important outcome from this workshop was the dissemination of resources, such as Jupyter notebooks (Python) with examples of data analysis and visualization procedures that participants can take back and use in their own teaching and research. To extend the impact of these resources, a website was established for the workshop (http://www.unidata.ucar.edu/events/2015UsersWorkshop/), which is recommended to serve as a hub of information related to geoscience data analysis and management. Links to videos of the workshop presentations as well as the workshop Github and RAMADDA sites are provided, along with other useful information. For example, Unidata has developed the Data Management Resource Center, a "one-stop" collection of materials (http://www.unidata.ucar.edu/data/dmrc/) intended to help scientists keep up with rapidly changing data management requirements and tools. The site includes separate links to data management requirements from different funding agencies (e.g., NSF, NOAA, NASA), a collection of best practices and recommended tools for a variety of common activities in planning how data are gathered, stored, used, and shared. Unidata has a long history of helping to create data management solutions for the community, which will benefit from an improved flow of information regarding these issues. Therefore, it is recommended that Unidata continue to serve as a repository for best practices, training materials, and recommended tools for scientific computing.

The feedback collected from participants both during and after the workshop can be summarized as follows: "I wish there was more time to dig deeper into these topics." We recommend additional training be considered, such as Unidata's Python training sessions and Software Carpentry workshops, to assist the proliferation of data science throughout the community. A post-workshop survey indicated that several participants have incorporated concepts and materials from the workshop into their teaching and research.

Acknowledgement:

Support for the workshop was provided by the National Science Foundation grant 1535378 "2015 Unidata Users Workshop Data-Driven Geoscience: Applications, Opportunities, Trends and Challenges."