The DOI for this manuscript is doi: 10.1175/BAMS-D-17-0304.1

The final published version of this manuscript will replace the preliminary version at the above DOI once it is available.

If you would like to cite this EOR in a separate work, please use the following full citation:

Sellars, S., 2018: “Grand Challenges” in Big Data and the Earth Sciences. Bull. Amer. Meteor. Soc. doi:10.1175/BAMS-D-17-0304.1, in press.

"Grand Challenges" in Big Data and the Earth Sciences

S. L. Sellars

[1] Center for Western, Weather, and Water Extremes, Scripps Institute of Oceanography, La Jolla, California.

Corresponding author: Scott L. Sellars (scottsellars@ucsd.edu), orcid.org/0000-0003-0778-8964

Title: Big Data and the Earth Sciences: Grand Challenges Workshop

What: Over 100 participants interested in Big Data and the Earth Sciences from industry, academia, government, and research organizations met to discuss advanced cyberinfrastructure and technologies as well as Big Data approaches that are emerging in the Earth sciences.

When: May 31 to June 2, 2017

Where: La Jolla, California

Introduction:

The Big Data and the Earth Sciences: Grand Challenges Workshop[1] held in late spring 2017 in California, was assembled so researchers in the Earth sciences, computer sciences, and information technology could learn, network together, collaborate, and focus on the challenges they all face in using Big Data capture and "data sciences" approaches. It was attended by 127 participants, including 60 undergraduate/graduate students from the *Machine Learning for physical applications* class taught by Scripps

[1] The Big Data and the Earth Sciences: Grand Challenges Workshop was hosted by the Pacific Research Platform (PRP) and the Center for Western Weather and Water Extremes (CW3E) of UC San Diego's Scripps Institution of Oceanography.

Institution of Oceanography. workshop The Grand Challenges aspect of the workshop was to focus on bringing together thought leaders on how to bridge the disciplines needed for the Earth science community to take full advantage of data science tools provided by advanced cyberinfrastructure.

The three main topics of discussion of Earth sciences research included:

- Cyberinfrastructure technological advancements: Big Data acquisition, collection, management, storage, access, and collaboration.
- Computational Science: statistical sampling, modeling and methods for Earth sciences data exploration, analysis, understanding, and interpretation.
- Challenges: those faced in Big Data approaches for Earth science investigation.

Each day had at least one Grand Challenges lecture, laying the foundation for the sessions during that day. The four lectures, summarized in this report, included distinguished researchers and experts who have engaged in these areas:

- **Dr. Larry Smarr**, Founding Director of the California Institute for Telecommunications and Information Technology (Calit2), a UC San Diego/UC Irvine partnership, holds the Harry E. Gruber professorship in Computer Science and Engineering (CSE) at UC San Diego's Jacobs School.
- **Dr. Michael Wehner**, Senior Staff Scientists, Computational Research Division at the Lawrence Berkeley National Laboratory.

- **Dr. Vipin Kumar**, Regents Professor at the University of Minnesota, holds the William Norris Endowed Chair in the Department of Computer Science and Engineering, University of Minnesota.
- **Dr. Padhraic Smyth**, Professor, Director, UCI Data Science Initiative and Associate Director, Center for Machine Learning and Intelligent Systems, UC Irvine.

Workshop Highlights:

A noticeable theme throughout the workshop was that technological advances in hardware and software have allowed data driven approaches to emerge as powerful tools that can be used in the era of Big Data and “deep analysis.” In addition, many of these technologies allow for massive data transfers, storage, and analysis approaches—necessary features to process enormous and often complex datasets. The first series of sessions discussed many technologies emerging from projects like the NSF-funded Pacific Research Platform (PRP, such as the Flash I/O Network Appliance (FIONA) and an end-to-end 10-100Gbps network backbone for data transfers), Globus data transfer service, and workflow technologies, which are transforming how science is performed. A Senior Engineer at UC San Diego’s Qualcomm Institute/Calit2, [Mr. John Graham], stated early in his talk that “we can't even keep up *[referring to technology]*, and that is a good thing.” His statement emphasizes the fast pace of innovation in the field of Big Data, technology, and data science, and that even the top centers and experts struggle to keep up with it.

Beyond the technological capabilities, presentations on computational research in predictive modeling in the Earth sciences focused on the advancing capabilities of data science approaches to Big Data. Prominent researchers and graduate students discussed state-of-the-art machine learning methods, such as the Extreme Learning Machines, Generative Adversarial Networks, and Recurrent Neural Networks that are being successfully applied to pressing Earth science prediction problems such as precipitation, cloud, and river streamflow forecasting. These methods are often available from open source software packages.

In the Earth sciences, numerical models have also advanced, including data assimilation, higher space and time resolution, advanced physics and optimization, and coupling of Earth systems. Many participants who have worked in modeling physical-based systems continue to raise caution about the lack of physical understanding of machine learning methods that rely on data-driven approaches.

Dr. Bruce Cornuelle, Senior Researcher and Oceanographer at Scripps Institution of Oceanography, led his talk with the question: “How can we merge machine learning with data assimilation?” He then focused on a discussion about how physical models and data-driven models are competing in real-world prediction problems and how we need to bring these two closer together. He suggested that our efforts should be improved optimization for physical models and better diagnostics for data-driven models. In the end, he posed a powerful question that turned out to be more of a challenge to the computer science community, “Could a data-driven model infer the equations of motion from a sparse, incomplete, and noisy ocean dataset?” A grand question indeed that highlights the need for multi-disciplinary collaboration and inclusion of discipline specific knowledge to address these problems.

Summary of the Grand Challenges Lectures:

Dr. Larry Smarr kicked off the workshop presenting the progress made over the last decade in science data networking and architecture by Universities. He also laid out his vision for a National Research Platform, the next iteration of the PRP that was originally envisioned in 2009, that would “link together Universities across the country on a national scale”. Throughout the first day, terms like ESnet, CENIC, Internet2, XSEDE, Globus, Kubernetes, non-von Neumann processors, Rook, and Kepler Workflows were used. The use of these terms sent many in the audience “Googling” and seeking definitions of the tool names, ideas, and processes that were discussed. Although, the overarching session relied on discipline specific jargon, the benefits of the use of these technologies for handling Big Data were made clear by examples after example of science being enhanced (e.g., improved scientific workflow, data sharing, and collaboration). Many participants were very interested to not only learn about the state-of-the-art in Big Data technologies and data sciences but also how to start the process of engagement with a technologist.

Dr. Michael Werner’s Grand Challenges Lecture that afternoon emphasized the challenges that large-scale climate modeling projects present with the ability to transfer and analyze the “copious” amounts of data that the numerical climate models produce. His talk discussed how we do large scale weather and climate science, including international climate modeling intercomparison projects. He suggested that in the era of Big Data, these projects may not be able to succeed without a strategic plan to deal with storing and distributing these massive datasets for research teams to access. Beyond access to data, he highlighted the serious challenges scientists face in analyzing the many model realizations, runs, and variables.

Dr. Vipin Kumars presented the third lecture and showed how he and his colleagues are utilizing machine learning approaches to provide a new ability for scientists to understand land use and land cover change dynamics on a global scale. He cautioned about the challenges that traditional data science approaches face when applied to Earth science data as well. His concerns include the “unstructured” nature of the data, the quality and/or scope of the data, and the source of the data that includes many different sensors and different space and time modalities. Although these cautions do exist, he saw these as exciting opportunities for the computer science arena. He showed examples of research on labeling and describing complex and unstructured data [*Mithal et al.,* 2017]”, and using known physical properties of the data to guided labeling and describing it when the quality is poor [*Jia et al.,* 2016, 2017, *Khandelwal et al*., 2015].

Dr. Padhraic Smyth in the final Grand Challenges Lecture cautioned the participants that with these promising results and discoveries these methods and approaches are not always easy to apply directly to Earth science problems. He identified, for instance, that simply training a predictive model on data from one region, in general, will not transfer to other regions. Dr. Smyth shared another example of the challenges by reporting results from a study in a state-of-the-art pattern recognition algorithm trained to detect either guitars or penguins [*Nguyen et al.,* 2015] and showed enormous accuracy when presented with pictures of one or the other (upwards of 98.90% accuracy for Guitars and 99.99% accuracy for penguins). The issue was that it was also extremely confident (99.99% certainty) that a picture of an abstract pattern with similar colors to a penguin/guitar was a penguin/guitar. To a human observer, it is obvious that none of these patterns resemble a penguin or guitar. These and other issues exist with these powerful algorithms and highlight Dr. Cornuelle's point about the importance of domain knowledge.

The overall message conveyed by all lecturers was that, although each of the Earth sciences' disciplines requires independent knowledge and expertise, future Earth science research would depend upon the successful collaboration and integration of knowledge from a diverse set of domains.

Outcomes: Meeting the Challenge – paths forward for Big Data in Earth Sciences

Throughout the 2.5 days of discussions, there was a wealth of insight into the many ways to move forward in harnessing Big Data approaches in the Earth sciences.

Education

It was obvious that a curriculum that allows for a student to learn computer science, machine learning, systems thinking, as well as Earth sciences (or other disciplines for that matter) is needed, yet it was unclear how to do this, given that most students are rooted in a single domain. It was suggested that we need to "build the paradigm of machine learning that can incorporate the knowledge of these different disciplines." In the end, it was unanimous that there is a dire need for people with skills in both camps, but no clear answer on how best to integrate or coordinate their knowledge.

Discipline Knowledge and Reward Structure for "Renaissance Teams"

"How do we alleviate the challenges faced by multidisciplinary teams?" Cross disciplinary engagement is very challenging and exciting, as viewed by academia. Dr. Smarr described what his colleague, Dr. Donna Cox from the National Center for Supercomputing Applications (NSCA), calls "Renaissance Teams." These multidisciplinary teams learn enough about each other's discipline to be productive. They are still quite rare, but are necessary for innovative approaches to be successful. There must be rewards, venues, journals, and workshops for these interdisciplinary teams, and fortunately more of these types of venues have been developing recently. The reward structure was brought up throughout the workshop, and there was agreement that there are major barriers to what is needed to bring together the disciplines. It seemed clear that if a reward structure was set up to support these types of teams and projects, more students, scientists, and researchers would participate.

Cyberinfrastructure and Big Data Partners in the Earth Sciences

Geosciences are major drivers for cyberinfrastructure investment and use. Yet, with these drivers, and even considering that there has been more standardization over the decades, there still is little national data set conformity. Any graduate student working in the Earth sciences knows this well, as obtaining and organizing data from various research groups and modeling centers takes up a major portion of their time. To alleviate this, from a research perspective, we need to have a national strategy for linking Earth science researchers and data.

It was also highlighted that we really need improvements in "metadata," describing the data to be used in research (i.e., what is measured, what type of devise measured it, and what units are used). The metadata is important and that these types of improvements are necessary for the longevity of the data and to keep a sustained community involved.

More information about the workshop can be found here:

http://prp.ucsd.edu/events/big-data-and-the-earth-science-grand-challenges-workshop

http://prp.ucsd.edu/BigDataEarthScience_Agenda_FINAL.pdf

https://www.youtube.com/playlist?list=PLbbCsk7MUIGfenfd5OV6ggpimI5A91Brg

http://prp.ucsd.edu/workshop-reports/BigDataWorkshop2017_Report_FINAL_082417.pdf

Acknowledgments:

The organizers would like to thank UC San Diego Qualcomm/Calit2 and Pacific Research Platform (#ACI-1541349), The Center for Western Weather and Water Extremes (CA AR Program award 4600010378 and NOAA PSD award NA15OAR4320071), and the Scripps Institution of Oceanography's Directors Office for financial support.

References:

Jia, X., Khandelwal, A., Gerber, J., Carlson, K., West, P., and Kumar, V. Learning Large-scale Plantation Mapping from Imperfect Annotators. In IEEE Big Data (Big Data), 2016.

Jia, X., Khandelwal, A., Gerber, J., Carlson, K., Samberg, L., West, P., and Kumar, V. Automated Plantation Mapping in Southeast Asia Using Remote Sensing Data. In Department of Computer Science and Engineering-Technical Reports.

Khandelwal, A., Mithal, V., and Kumar, V. (2015). Post Classification Label Refinement Using Implicit Ordering Constraint Among Data Instances, Proceedings of the IEEE International Conference on Data Mining.

Mithal, V., Nayak, G., Khandelwal, N., Kumar, V., Oza N., and Nemani, R. (2017). RAPT: Rare Class Prediction in Absence of True Labels. IEEE Transactions on Knowledge and Data Engineering.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Vol. 7-12-NaN-2015, pp. 427–436). http://doi.org/10.1109/CVPR.2015.7298640