This paper is SAB Advisory only and does not reflect the opinions or policies of NOAA

D. Wright 17 July 2016

Kareiva Topic Category = 3
(i.e., "explore now, change direction and include in the research portfolio immediately")

1.0 Background and Opportunities

The multidimensional structuring and scaling data, with integrative and innovative approaches to analyzing, modeling, and developing extensive and spatial data from selected places on land and at sea, have revealed how theory and application are in no way mutually exclusive. In fact, it may often be *application* that advances *theory*, rather than vice versa. As such **data science** has emerged from strong collaborations among computer scientists, information scientists, and domain scientists to solve complex scientific questions. And now that we are squarely in an era of regional- to global-scale observation and simulation of the Earth, producing data that are too BIG, move too fast, and do not fit the structures and processing capacity of conventional database systems, NOAA as an agency is well positioned to meet these challenges. Indeed, within NOAA, data science is focused on extracting useful information and providing expedient intelligence from complex and massive collections of data that are often characterized by large degrees of heterogeneity, and rapid rates of collection.

Presently, NOAA manages a diverse portfolio of data that is growing exponentially in size, and is under higher demand than ever before. Therefore, NOAA, in accordance with the Federal Open Data Policy and its increasing dependence on data science, is looking for ways to improve its data science strategies to further promote openness and accessibility and enhance its abilities to harness intelligence from its ever expanding ensemble of data.

1.1 Relevance of Data Science to the NOAA Mission

- **Environmental intelligence**: NOAA is the only federal agency with the operational responsibility to provide weather, water, ocean, climate, and ecosystem forecasts. To continue meeting this charge, and the need for improved precision and accuracy of

these forecasts, enhanced data management practices are a requisite for NOAA moving forward (*see* Data Science within the NOAA Strategic Research Memorandum Guidance or SRGM, NOAA Research Council, 2015), including strategies leading to improved data curation, harmonization, interoperability, and integration.

- **Value-added products:** Big data is the currency of the 'intelligent' economy. As the primary producer of environmental based data, and observations, NOAA has a role to play in improving accessibility of its data and also working with partners to leverage the development of this data into value-added products.

1.2 Examples of NOAA Data Science Projects/Initiatives **(note active hyperlinks)**

- **NOAA Big Data Project**: An innovative approach to publishing NOAA's vast data resources and positioning them near cost-efficient high performance computing, analytic, and storage services provided by the private sector.
- **NOAA Environmental Data Management Framework**: Defines, categorizes, and coordinates the policies, requirements, activities, and technical considerations relevant to the management of observational data and products derived by NOAA.
- **NOAA Public Access to Research Results:** Comprehensive plan to increase public accessibility of publications and digital data produced by NOAA scientists, and extramurally funded researchers.

1.3 NOAA Entities Providing Agency Leadership in Data Science **(note active hyperlinks)**

- Office of the Chief Information Officer (CIO)
- CIO Council and Council Committees
- NOAA Observing Systems Council (NOSC)
- National Centers for Environmental Information (NCEI)
- NOAA Data Management Integration Team (DMIT)
- NOAA Line Office Data Stewards

2.0 Key Challenges

Capitalizing on the above progress to move forward, several challenges should be addressed in a series of workshops or webinar in partnership with selected universities, NOAA cooperative institutes or the private sector. Below are four examples, with more that should be brainstormed upon.

2.1 (i) How can NOAA ensure maximum social return on its data (e.g., public-private partnerships; novel platforms, etc.)?

Indeed, an entire **ocean data industry** has arisen in the last 5-10 years, part of the new Blue Economy as described recently by Spinrad (2016). NOAA's Big Data project is already an excellent example of a partnership ecosystem to help ensure maximum return on its data investments. NOAA must not stop there, especially with the *data acquisition* market estimated

at $80 billion from ships, buoys, satellites, AUVS, ocean communication (Rainer Sternfeld, PlanetOS, pers. comm., 2013). The ocean *data management* market will be $5 billion within a broader geospatial services market of $376 billion, including software and associated costs (PlanetOS, https://planetos.com/). These industry partners are all working on different aspects of data science. In particular, the Research Data Alliance is continually at work fostering public-private partnerships focusing on data use, data quality (**Figure 1**).

Figure 1. An example ecosystem of existing and potential ocean data science partners for NOAA.

2.2 (ii) Looking to the future, what data management policies and practices are needed to ensure long term sustainability and operability of NOAA platforms and databases?

Culling the expertise of DAARWG, CWG, and EISWG and through the aegis of the NOAA Environmental Data Management Committee should be helpful in this regard.

2.3 (iii) What are the workforce requirements of the "data age," and does NOAA need to evolve its workforce to meet these?

While there is a strong demand for "data scientists" as highlighted in Manyika et al. (2011; aka as the now famous McKinsey Report") and new data science programs are emerging in industry and academia, true consensus as to what a data scientist does or should do (e.g., Dhar, 2013; Murata et al., 2013; Stocks et al., 2015). The traditional professions of researcher, computer programmer, systems engineer, lab or field technician do not fully address the complex,

mediating requirements of making data work, nor the various computational approaches to analyzing the entire ecosystem around scientific discovery itself (Goth, 2012). And Berman and Bourne (2015) point out that while data science has the potential to narrow the gender gap and set a new bar for inclusion there exist the persistent challenges of increasing the number of women going into the field, and evolving professional cultures so as to retain such diversity.

2.4 (iv) How could NOAA enable transdisciplinary use of its data, particularly leveraging social science insights (e.g., monitoring social media platforms to update real-time intelligence during weather events)?

As advised by the National Research Council (2011): "It would be beneficial for federal agencies to periodically examine and adopt data management practices that come from *beyond* the ocean sciences, as well as approaches to grow access to and use of community-wide facilities." Proven efforts from beyond the ocean sciences can be very informative and helpful. As illustrated in **Figure 1**, community-specific organizations that focus on data use and data quality will also be valuable to the ocean sciences and they include the participation and expertise of social scientists (e.g., NSF EarthCube, AGU Earth and Space Informatics, Research Data Alliance).

3.0 Leading Thinkers to Consider Inviting to SAB Strategy Sessions or Workshops

(with numbers indicating which challenge(s) they might best be able to address)

Fran Berman (ii, iii), bermaf@rpi.edu**,** is the Edward P. Hamilton Distinguished Professor in Computer Science at Rensselaer Polytechnic Institute. She is the U.S. lead of the Research Data Alliance (RDA), "a community-driven international organization created to accelerate research data sharing world-wide, through the development and adoption of technical, organizational and social infrastructure needed to support data-driven innovation." Previously, Fran served as Director of the San Diego Supercomputer Center from 2001 to 2009 and as Vice President for Research at Rensselaer Polytechnic Institute (RPI) from 2009-2012. In 2015, Fran was nominated by President Obama and confirmed by the U.S. Senate to become a member of the National Council on the Humanities.

Ruth Duerr (i, ii, iii), rduerr@nsidc.org, is the manager of the data stewardship program at the National Snow and Ice Data Center (NSDIC) within the NOAA Cooperative Institute for Research in Environmental Science (CIRES) in Boulder. In her role as a data scientist for a number of collaborative projects, Ruth participates in workshops to advise on characteristics of certain sea ice data, metadata, and documentation. Ruth is also the primary liaison between NSDIC and the Data Conservancy.

Peter Fox (i-iv), pfox@cs.rpi.edu, is one of the most well-respected thought leaders in Data Science today. He is currently the Tetherless World Constellation Chair and Professor of Earth and Environmental Science, Computer Science and Cognitive Science at RPI. Previously, he was Chief Computational Scientist at the High Altitude Observatory of the National Center for Atmospheric Research and before that a research scientist at Yale University. In 2015, the was

elected as the first Earth and Space Science Informatics (aka Data Science) fellow to the American Geophysical Union. He has developed one of the first higher education courses in data science at RPI. Peter is also a member of **NOAA SAB DAARWG**.

Vincent Granville (i, iv), vincentg@datashaping.com, Executive Data Scientist and co-founder of DataScienceCentral, one of the world's largest online resources for big data practitioners in companies ranging from startups to Fortune 100, across multiple industries (finance, Internet, media, IT, security) and domains (data science, operations research, machine learning, computer science, business intelligence, statistics, applied mathematics, growth hacking, IoT). Vincent is a visionary data science executive with a broad spectrum of domain expertise, technical knowledge, and industry connections (and with over 143,000 followers on Twitter).

Steve Kempler (ii, iv), steven.j.kempler@nasa.gov, is Manager of the NASA Goddard Earth Sciences Data and Information Services Center (DISC) with an interest in mentoring early career professionals in a data science career. **Steve is also a member of DAARWG.**

Kerstin Lehnert (i, ii), lehnert@ldeo.columbia.edu, is Senior Research Scientist at the Lamont-Doherty Earth Observatory, Columbia University and Director of the Intedisciplinary Earth Data Alliance (IEDA)**.** A marine geochemist by training, Kerstin has become a global citizen in the data science community with a service record including Science Officer and Sub-programme Co-convener for Earth and Space Science Informatics focus groups of both the American Geophysical Union (AGU) and the European Geosciences Union (EGU), member of the Data Science Credentialing Editorial Board for the AGU, Member of the NSF Advisory Committee for Cyberinfrastructure, and Leadership Council of the NSF-funded EarthCube Initiative.

Christopher Lenhardt (i-iv), clenhardt@renci.org, is Domain Scientist, Environmental Data Science and Systems division of the Renaissance Computing Institute (RENCI), University of North Carolina at Chapel Hill, and **current chair of DAARWG**. As such, Chris has a particularly valuable insights into various data science issues for NOAA, combined with 20+ years experience managing the integration of information technology and data stewardship best practices to support the science enterprise; data center operations management, digital archiving, metadata standards, application of leading edge information technology, sociology of science data management.

Mark Parsons (i-iv), parsom3@rpi.edu, is the first Secretary General of the (RDA), an Associate Director of the Rensselaer Institute for Data Exploration and Applications, and is similarly a global spokesperson and thought leader of data science. Before being appointed Secretary General, Mark was the Managing Director of RDA/United States and the RPI Center for the Digital Society and a Senior Associate Scientist at NSIDC within CIRES. With his training in

geography, he focuses on stewarding research data and making them more accessible and useful across different ways of knowing.

Amy Nurnberger (iii, iv), anurnberger@columbia.edu, Research Data Manager at the Center for Digital Research and Scholarship of the Columbia University Libraries who is currently conducting an extensive community survey of data-science-related educational programs so as to better define what a data science professional needs to be today and into the future.

Carole Palmer (iii, iv), clpalmer@uw.edu, Professor in the Information School of the University of Washington, works in the areas of data curation and digital research collections and advancing data services for interdisciplinary inquiry. She has also been a leader in professional workforce development in data curation for nearly a decade, recognized in 2013 with the Information Science Teacher of the Year Award from the Association for Information Science & Technology.

Erin Robinson (i, ii), erinrobinson@esipfed.org, is the Executive Director for the Foundation for Earth Science and the Federation of Earth Science Information Partners (ESIP). ESIP is open, networked data science community, which held its annual summer meeting (July 19-26) on the theme: Frontiers in Earth Sciences Big Data. Erin facilitates virtual and in-person collaboration across many academic, government, and economic sectors to expedite progress toward data interoperability.

Karen Stocks (i,ii, iii), kstocks@ucsd.edu, is Director of the Geological Data Center at Scripps Institution of Oceanography and involved in a bevy of data science projects and initiatives with NSF EarthCube and the Ocean Data Interoperability Platform. Karen is particularly interested in the issues surrounding cementing data science as an accredited profession and in training the next generation of practitioners accordingly. **Karen is also a member of DAARWG**.

Kristin Tolle (i), Kristin.Tolle@microsoft.com or ktolle@microsoft.com, is the Director of the Data Science Initiative in Microsoft Research Outreach. To many in the Earth Science community she is most well known as the co-editor, along with Tony Hey and Stewart Tansley, of one of the earliest books on data science, *The Fourth Paradigm: Data Intensive Scientific Discovery*. Her current focus is develop an outreach program to engage with academics on data science in general and more specifically around using data to create meaningful and useful user experiences across devices platforms.

4.0 References

Berman FD, Bourne PE. 2015. Let's make gender diversity in data science a priority right from the start. *PLoS Biol* **13**(7): e1002206. doi:10.1371/journal.pbio.1002206.

Dhar V. 2013. Data science and prediction. *Communications of the ACM (Association for Computing Machinery)* **56**(12): 64-73. doi:http://dx.doi.org/10.1145/2500499.

Goth G. 2012. The science of better science. *Communications of the ACM (Association for Computing Machinery)* **55**(2): 13-15.

Manyika J, Chui M, Brown B, Bughin J, Dobbs, R, Roxburgh C, Hung Byers, A. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. New York: McKinsey Global Institute, 156 pp. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation, last accessed 17 July 2016.

Murata KT, Watari S, Nagatsuma T, Kunitake M, Watanabe H, et al. 2013. A science cloud for data intensive sciences. *Data Science Journal* WDS139–WDS146. doi: http://doi.org/10.2481/dsj.WDS-024.

National Research Council. 2011. *An Ocean Infrastructure Strategy for U.S. Ocean Research in 2030*. Washington, DC: National Academies Press, 98 pp., doi: 10.17226/13081, http://www.nap.edu/catalog/13081/critical-infrastructure-for-ocean-research-and-societal-needs-in-2030, last accessed 17 July 2016.

NOAA Research Council. 2015. *Strategic Research Guidance Memorandum*, 9 pp., http://nrc.noaa.gov/CouncilProducts/StrategicResearchGuidanceMemorandum.aspx, last accessed 17 July 2016.

Spinrad RW. 2016. The new blue economy: A vast oceanic frontier. *Eos, Trans AGU* **97**: doi:10.1029/2016EO053793.

Stocks K, Yarmey L, Duerr R, Wyborn, L. 2015. Data science careers: A sampling of successful strategic, pitfalls, and persistent challenges, Eos, Trans AGU **96**, Fall Meet. Suppl., Abstract IN33D-1827.