Adoption and Adaptation of Data Science in Oceanography

Katie Kuksenok
Computer Science & Engineering
University of Washington
Seattle, WA, USA
kuksenok@cs.washington.edu

Abstract

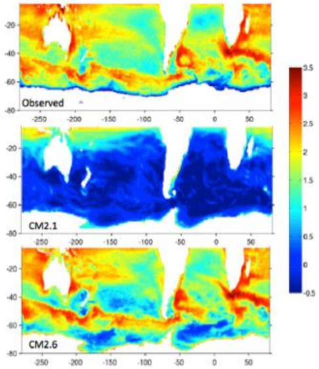
Ocean sciences in the US have had a cultural distinction between *modeling* and *fieldwork*: a researcher either wrote MATLAB code, or went on data collection cruises. Large-scale multi-institution collaborations, and adoption of data science tools and skills, are blurring this distinction. CSCW and STS often study *data*: its production, maintenance, management, and use. In my dissertation, I focus not on the *data* but oceanographer *groups* incorporating data science practice into their work. By studying *challenges* faced by collective actors, this ethnographic research will then lead to developing design and organization implications for supporting data science practice in scientific academic collaborations.

Motivation

New data collection and analysis technologies – belonging to an amorphous and context-dependent umbrella of *data science* – have precipitated radical shifts in the process of physical and life sciences. Oceanography in particular has begun to be flooded by exciting data full of possibilities: covering more of the oceans around the world than ever before, with unprecedented resolution. However, this comes not only with challenges of storage, maintenance, and manipulation, but also the social dynamics around creating and deploying new data collection and analysis technologies. This includes educational initiatives to introduce domain scientists to data science practice; interdisciplinary collaboration between domain scientists and methods scientists, such as statisticians and computer scientists; and integration of data science staff persons.

What is a large-scale, data driven oceanography project? An example:

Southern Ocean Carbon and Climate Observations and Modeling is a large-scale collaboration poised to develop a new observing system for carbon, nutrients, and oxygen; deploy ~200 profiling floats in the Southern Ocean; and provide hi-res simulations and analyses.



Shown: eddy kinetic energy from observations (top), a coarse-res model (middle), and a hi-res model (bottom). An *eddy* is a swirl of fluid; in the ocean, it can transport nutrients and biomass. Eddies are chaotic, so they are modeled differently than tides. Research in ocean sciences spans chemistry, biology, physics - and computational modeling.

In the case of ocean sciences, data gathering through time-consuming cruises is competing for grant funding and researcher-hours against infrastructure-oriented, large-scale collaborations that aim to produce not just scientific results, but standardized data and analytic tools. Such collaborations are positioned to produce increasing amounts of data to be analyzed by the scientific community at large. As an example of a collaboration, the sidebar on the left provides a brief peek at the SOCCOM project, and at one of the research areas (modeling eddies), in oceanography[1].

Computational methods have long been used within the physical and life sciences. Edwards' popular book, *A Vast Machine*, uses the history and politics behind climate science to explain the inseparability of "raw data" from "models" in global climate change research [4]. Computer scientists, and adjacent communities including CHI and CSCW, have long studied scientific technology use, including parallel computing, database systems, analysis platforms, visualization, applications from robotics to remote and/or automated data collection, and communication systems to support distributed collaboration.

Research about scientific work practice is both abundant and largely data-centric. The rationale behind this is articulated by Borgman *et al*: "Data are the 'glue' of a collaboration, hence one lens through which to study the effectiveness of such collaborations is to assess how they produce and use data" [1]. Lee's work on boundary objects used for negotiation around the chaos in work [7] helps to construct a more nuanced account of the role of data-driven artifacts.

The popular use of the word "workflow" exemplifies this data-centric stance. A glance through 'eScience workflow' search results yields scores of work on *data* workflows, with diagrams where the boxes are databases and programs, and the arrows are human actions. Guo's use of the term still uses a data-centric conceptualization, but the boxes of his diagram – reproduced below - correspond to actions [5]:

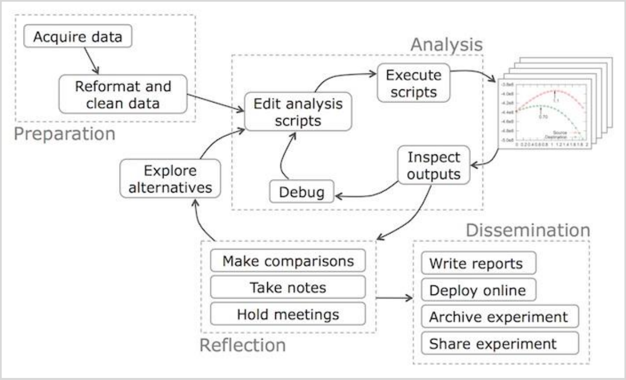


Figure 1: The "Data Science Workflow" diagram [5].

Distinct from "what happens to data" and "what a person does to data" is the question of "what *people* do

[1] The image in this sidebar is taken from the SOCCOM project page, http://soccom.princeton.edu, and the explanatory text was written by me in an effort to simplify. SOCCOM is not involved in the study I am proposing.

before, during, and after data-specific activities." The hypothetical actor performing the tasks in Guo's diagram suggests either an individual or a group that is in agreement. Groups often have to negotiate around uncertainty [7]. Other forms of work, such as anticipation work on developing standards and creating social structures [6], are both vital and require more open-ended inquiry than about data alone.

Research Questions and Methods

The goal is to understand group strategies for overcoming challenges in integrating data science practice into scientific oceanography work. This goal covers several research questions, all in context of oceanography groups studied:

RQ1: What is the division between "data science" and "scientific work"? What constitutes data science? (Ethnographic inquiry into *participants' meanings and values*)

RQ2: What are the challenges in integrating "data science" practice into various oceanography work processes? (CSCW design research inquiry about *work practice*)

RQ3: What are the strategies that *groups* develop for addressing these challenges? (Ethnographic CSCW inquiry about *group dynamics, boundary objects, and norms*)

RQ4: What are the implications of the strategies on the design of technology and educational initiatives? (User-Centered Design, Participatory Design)

I will conduct iterative qualitative study combining ethnographic observation and interviews, and semi-structured interviews [8]. Participant-observation will include *ethnomethdological* inquiry into the minutiae of collective decision-making, and interviews will help ground findings in informants' own values, and meanings. Interpretation during analysis will take an *actor network theory* stance, and incorporate themes grounded in the data [2] with scholarship on (1) affect in the workplace, (2) boundary objects, such as data or software artifacts, and (3) oceanographic practice.

Having thus answered the first three questions, I will draft the recommendations. The nature of these depend on the findings; for example, Chen et al build a large-display visualization based on the findings of ethnographic work [3], but the appropriate design need and solution were not known in advance of the work. The recommendations developed from synthesis of data and literature will then go through iteration in participatory design exercises in order avoid imposing software engineering values in a scientific community with its own norms and priorities.

Novel Data Science in Ocean Science Groups

The challenges of data exist in an arena that places staff programmers, computer scientists, and domain scientists into a decision-making collective. Data is central: *how do we collect the right data? how do we build the software to analyze it? how do we do collection and analysis in a way that allows us to make it public, but still publish? how do we get the next grant?* This research concerns strategies of oceanographers for answering these questions in discussions about incorporating new tools and skills.

Scientific groups. Over summer 2014, I interviewed 15 Software Carpentry Bootcamp[2] attendees from a

[2] software-carpentry.org/workshops/

variety of sciences about integrating new data science skills into their scientific practice. This study was about how individuals incorporate the new skills and tools they learned at a workshop-style educational intervention. In multiple semi-structured interviews, I asked each participant about their progress, revealing that the social context of data science (from group dynamics to inter-institutional collaborations) can inhibit an individual's capacity to learn and integrate new approaches. This exploration of different scientists sets the stage for the more in-depth, targeted inquiry into oceanographers' experience in particular.

<u>Ocean science groups.</u> The research sites include one "modeling" and one "data collection" oriented research group, mirroring the historical split in oceanography. University of Washington is one of the four largest oceanography centers with modeling, data collection, biological, chemical, and physical science arms of oceanography having considerable representation in faculty, graduate students, and undergraduates. Understanding the dynamics of a group may require going beyond that group to interview other actors and stakeholders that influence its decisions, potentially beyond UW.

<u>Novelty.</u> The rhetoric around data science stresses methodological novelty, and from the exploratory interviews and fieldwork so far, the sentiment of wanting to learn something but not having gotten around to it yet ("I've been meaning to learn X...") is pervasive. Educational interventions like SWC are springing up in universities around the US – including, but not limited to, the Moore/Sloan Data Science Environments Initiative, which this research intends to benefit directly in the design recommendation stage.

Proposed Thesis Timeline

At each stage, the hypotheses for and answers to the four core research questions are refined.

SUM14 Exploratory interview study (analysis in-progress)

FALL14 Negotiate field-sites and ocean science teams to study

WIN2015 Participant observation of 2 comparative fieldsites: a "modeling" group and a "data collection" group

SPR2015 Analysis of field-notes, synthesis of findings and literature

SUM2015 More participant-observation; focus on conducting interviews with people important to mapping out collective problem-solving dynamics in the groups studied

FALL2015 Synthesis of data into grounded theory, development of design and organizational implications; complete most of corresponding writing

WIN2015 Design-research study, involving participatory design exercises as appropriate to the emergent implications.

SPR2016 Finish writing, defend

CSCW Doctoral Consortium

The Doctoral Consortium would be an excellent opportunity to refine this study, form a methodological and theoretical perspective, and gather additional related readings, both about scientific work; oceanography; and collaborative decision-making. The timing would be particularly beneficial, as research will be both well underway, but with more than half of the study remaining to be completed. As a participant, I would contribute my own expertise on performing and critiquing mixed methods research, and qualitative open-ended research in HCI.

References

[1] Borgman, C.L., Wallis, J.C., Mayernik, M.S. Who's got the data? Interdependencies in Science and Technology Studies. JCSCW 2012.

[2] Charmaz, K. <u>Constructing Grounded Theory.</u> 2014.

[3] Chen, Y. C., Lee, S., Hur, H., Leigh, J., Johnson, A., & Renambot, L. Case Study: Designing An Advanced Visualization System for Geological Core Drilling Expeditions. CHI Extended Abstracts 2010.

[4] Edward, P.N. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. The MIT Press 2010.

[5] Guo, P. "Data Science Workflow." Communications of the ACM, October 2013.

[6] Steinhardt, S.B., Jackson, S.B. Anticipation Work: Cultivating Vision in Collective Practice. CSCW 2015.

[7] Lee, C.P. "Boundary Negotiating Artifacts: Unbinding the Routine of Boundary Objects and Embracing Chaos in Collaborative Work". CSCW 2007.

[8] Maxwell, J.A. <u>Qualitative Research Design: an Interactive Approach</u>. 2012.