TO BE SUBMITTED TO APJL: DRAFT - MAY 9, 2013
Preprint typeset using LaTeX style emulateapj v. 08/13/06

WHAT TO KEEP AND HOW TO ANALYZE IT: DATA CURATION AND DATA ANALYSIS WITH MULTIPLE PHASES

ALYSSA GOODMAN, ALBERTO PEPE, VINAY KASHYAP, ASHISH MAHABAL, XIAO-LI MENG, ALEKSANDRA SLAVKOVIC, ANETA SIEMIGINOWSKA, ROSANNE DI STEFANO, CHRISTINE L. BORGMAN, PAUL GROTH, YOLANDA GIL, DAVID W. HOGG, AND KYLE CRANMER

Subject headings:

0.1. *Overview*

This open document is being used to describe and record the events at the Radcliffe Exploratory Seminar on Data Curation and Analysis, to be held at the Radcliffe Institute for Advanced Study, May 9-10 2013.

This Google Drive Directory should be used to deposit all files contributed by participants before and during the meeting. (Click “Open in Drive” on your browser to make a new folder, e.g. with your name as its name.)

This Google Doc is used for collaborative real-time note-taking.

ABSTRACT: Rapid advances in technology have allowed us to collect vast amounts of data in myriad fields and forms, but our ability to manage and analyze these data has not kept pace. As a result, the amount of data collected far exceeds what can be analyzed and, often, what can be archived. These issues only become more pressing as data collection accelerates. Astronomers and astrophysicists, for example, collect terabytes of data per night; the phrase “drowning in a data tsunami” is increasingly used to describe this situation. The issues of what to keep and what to distribute are surprisingly complex, even when we put aside technological issues such as long-term storage and retrieval. A central challenge is the fundamental conflict between reducing the size of data and preserving information for future scientific inquires and statistical analyses. Complicating matters further, the parties/teams involved in the entire data collection, curation, and analysis process often have only limited communication with each other owing to the sequential nature of this process. This seminar brings together a core group of leading experts and emerging scholars in information and natural sciences to discuss, debate, and design principles and strategies to address this grand challenge, which increasingly affects almost every aspect of science and society.

GOAL: By gathering experts from information and natural sciences, we aim to start building a set of principles and methods that will allow us to understand such problems and to provide better preprocessing, analyses, and data preservation, especially in the context of the natural sciences. The ultimate goals of this research include providing methods for assessing the validity of such collaborative analyses, guidance on statistically-principled preprocessing, and a rich new theory of statistical learning and inference with multiple parties. We believe that this collaboration will simultaneously sow the seeds for innovative mathematical theory and shed light on directly usable guidelines for the construction and curation of scientific databases.

0.2. *Draft Schedule of Events, May 9-10, 2013*

Location: Room 112, Radcliffe Gymnasium, Radcliffe Yard, 18 Mason Street, Cambridge, MA (Red pin on this map marks the front door of the Radcliffe Gymnasium–zoom in!)

0.2.1. *Day 1 (Thursday, May 9)*

8:30 AM - 9:00 AM Continental Breakfast

9:00 AM Introductory remarks and welcome address

SESSION I 9:15 AM – 12:30 PM *Quantitative and qualitative perspectives on multiphase science – Beginning a dialogue*

9:15-11:45 Introductions: each of 16 participants will answer the following questions (5 min/person, including short discussions & coffee break, total of 2.5 hours.)

> [1.]What about your background gives you an interest in data curation? What do you think is the most important opportunity good data curation offers? (Please just one!) What do you think is the biggest danger facing scientific research today if we don’t improve data curation? ((Please just one!)

Coffee Break at appropriate stopping point during the above, at roughly at 10:30.

11:45-12:30 Introduction to solutions proposed in the literature (Part I) Presented by: Meng, Borgman, Crosas, Pepe et al. (TBD)

12:30 PM – 1:30 PM Lunch

1:30-2:00 Introduction to solutions proposed in the literature (Part II) Presented by: Meng, Borgman, Crosas, Pepe et al. (TBD)

SESSION II 2:00 PM – 5:00 PM Specific challenges in data curation, provenance, and multiphase analysis

2:00 PM–4:00 PM

Roughly 40 minutes for each of the topics below (as amended at the Workshop). *Suggested* discussion leaders indicated, but changes can and will(!) be made to respond to participant suggestions. Each workshop attendee will each “sign up” (at lunchtime) for 3 discussions total, to be held within groups of roughly 5 or 6 people each. Multiple rooms will be available, and a schedule of which discussions will take place in which room will be made on-the-fly, and posted here. There will be three “blocks” of 40 minutes, with two or three topics to choose from within each block.

- 1 group collaboration challenges (Cranmer/Hogg)
- provenance, what’s realistic? (Hedstrom/Pepe)
- storage, ideas on what to keep, sociological & algorithmic approaches (Groth/Blocker)
- can statistics help? (Slavkovic/Siemiginowska)
- the divide between theory and practice: what we should do, versus what we do do (Goodman/Borgman)
- what has & has not worked in Astronomy? (DiStefano/Kashyap/Mahabal)

- working with & educating the community of data producers (Gil/Crosas)

4:00-4:20 Coffee Break

4:30-5:30 Group discussion of smaller group's discussions, used to refine plans for Day 2.

6:30 PM Group Dinner at NuBar, Cambridge (in the Sheraton Commander)

0.2.2. *Day 2 (Friday, May 10)*

8:30 AM - 9:00 AM Continental Breakfast

SESSION III 9:00 AM – 12:30 PM Where can we connect? Addressing foundational issues from interdisciplinary perspectives

12:30 PM – 1:30 PM Lunch

SESSION IV 1:30 PM – 5:00 PM What can we do together? Identifying opportunities for collaboration

6:30 PM Group Dinner (social event, location TBD)

0.3. *Participants*

Alexander Blocker, Statistics –Bio– Email ablocker@gmail.com

Christine L. Borgman, Information Science –Bio– Email: borgman@gseis.ucla.edu

Kyle Cranmer, Particle Physics –Bio– Email Kyle.Cranmer@nyu.edu

Merce Crosas, Data Science –Bio– Email mcrosas@iq.harvard.edu

Rosanne DiStefano, Astrophysics –Bio– Email distefano.rosanne@gmail.com

Yolanda Gil, Information Science –Bio– Email gil@isi.edu; anava@isi.edu

Alyssa Goodman, Astrophysics, Visualization –Bio– Email agoodman@cfa.harvard.edu

Paul Groth, Computer Science –Bio– Email p.t.groth@vu.nl

David Hogg, Astrophysics, Data Science –Bio– Email david.hogg@nyu.edu

Vinay Kashyap, Astrophysics, Statistics –Bio– Email vlk.astro@gmail.com

Margaret Hedstrom Information Science –Bio– Email hedstrom@umich.edu

Ashish Mahabal, Astrophysics –Bio– Email aam@astro.caltech.edu

Xiao-Li Meng, Statistics–Bio– Email mengharvard@gmail.com

Alberto Pepe –Bio– Email apepe@cfa.harvard.edu

Aneta Siemiginowska, Astrophysics, Statistics –Bio– Email asiemiginowska@cfa.harvard.edu

Aleksandra B. Slavkovic, Statistics –Bio– Email sesa@stat.psu.edu

Click here to email all Workshop Participants at once.

0.4. *Contributed links*

[1.]5-minute data/code sharing survey from the Harvard-Smithsonian Center for Astrophysics, April 2013 http://projects.iq.harvard.edu/seamlessastronomy/book/three-highlighted-graphs-spring-2013-cfa-data-code-sharing-survey

1. 0.5. *Appendix 1: Original Workshop Justification*

With the dramatic increases in the size, diversity, and complexity of data available for scientific discoveries, medical advances, education reforms and evidence-based policy making, the entire enterprise of scientific quantitative inquiry has been presented with unprecedented challenges and opportunities. In particular, the vast majority of current quantitative inquires are not made by a single individual or even a single team. The final scientific inference and, more generally, quantitative learning is a result of a multi-party effort, with teams/parties entering the process sequentially over several phases (e.g. data collection, processing, curation, and analysis). Due to practical constraints such as resource limitations and confidentiality, each team involved in a given phase may not have full knowledge of the assumptions made by, and resources available to, those coming before or after it. This fact compels all of us involved in the production and preservation of scientific data to rethink the traditional paradigms of statistical analysis and data preservation. These have been built around two ideas: (1) the academic paper as the primary repository of scientific knowledge and information, and (2) the analysis of data beginning (and ending) with a single team, who has essentially full knowledge of the data's origins and all assumptions made in its genesis.

Shifts in the scientific landscape call for revision of both of these ideas. Projects in astronomy, biology, ecology, and social sciences (to name a small sampling) are increasingly focused on building databases for future analyses as a primary objective. These projects must decide what levels of preprocessing to apply to their data and what additional information to provide to their users. Clearly, providing all of the original data allows the most flexibility in subsequent analyses. In practice, the journey from raw data to a complete analysis is typically too intricate and problematic for the majority of users, who instead choose to use preprocessed output. Unfortunately, decisions made at this stage can be quite treacherous from a statistical perspective because of the potential for serious information loss and/or information distortion.

Scientific data released to end-users almost always undergo editing, imputation, and other forms of preprocessing before they are analyzed. When such steps are taken, the data analysis becomes a collaborative endeavor by all parties involved in data collection, preprocessing, and analysis. Such settings are rife with subtleties and pitfalls. Teams subsequently handling those data do not and often cannot have a perfect understanding of the entire phenomenon at hand; the final results will inevitably contain some combination of their judgments, and some preprocessing can irreversibly destroy information from the raw data. By gathering experts from information and natural sciences, we aim to start building a set of principles and methods that will allow us to understand such problems and to provide better preprocessing, analyses, and data preservation, especially in the context of the natural sciences. The ultimate goals of this research include providing methods for assessing the validity of such collaborative analyses, guidance on statistically-principled preprocessing, and a rich new theory of statistical learning and inference with multiple parties. We believe that this collaboration will simultaneously sow the seeds for innovative mathematical theory and shed light on directly usable guidelines for the construction and curation of scientific databases.

Defects incurred by earlier parties may cause more damage than those in subsequent analyses, just as problems in the data collection stages are usually harder to address than problems in the analysis stage. This is especially true when some of those steps are "irreversible". An example of great current interest in astronomy and astrophysics concerns the use of data from Chandra X-ray Observatory. As described in the Chandra documentation (http://cxc.harvard.edu/ciao/dictionary/sdp.html), the "Chandra data" come with different level of processing, from Level 0 "raw data", which are not recommended for analysis, to Level 3 "higher lever information" available to public, where the Level 2 data processing is considered to be irreversible, which was defined as "By 'irreversible' we mean that information that has been lost cannot be regained from the L2 products alone." Evidently judgments have been made regards what to retain and what to discard, and as such assessing their impact on the subsequent analyses is of great importance for the so called V&V (Verification and Validation) process. Indeed, the question of "what to keep" has been a much debated and discussed topic in the rapidly growing literature on data curation, yet currently there is few collaboration between fields with overlapping interests in this area. For example, statisticians have been largely absent such discussion and debates.

Such collaborations would appear quite natural given the complementary strengths of the participants. Literature in the field of data curation has largely focused on describing how scientists use data, their motivations for data sharing, and the organizational and cultural issues involved in implementing better data curation practices. Simultaneously, computer scientists are developing technical solutions to enable tracking of data provenance and easier access to scientific resources, to name only a few directions. Statisticians are interested in developing principled statistical methods for these situations. These lines of research are distinct, but they provide necessary complements for each other and could benefit immensely from greater communication and collaboration.

As a specific example of the fundamental restructuring needed to address the aforementioned grand challenge, consider the current paradigm for conducting and evaluating statistical inferences. Statisticians are trained to regard their mathematical models as approximations to a true underlying reality. Consequently, these models are typically not designed to capture the journey from data collection to data analysis. This is very problematic because such journeys necessarily involve judgments and data preprocessing from other teams. If the assumptions made and procedures used in this preprocessing phase are incompatible with those used in the final analysis (so-called "uncongeniality" in the literature of statistical analysis), then the current statistical framework is ineffective, or, at worst, entirely inapplicable. In particular, standard notions such as estimation consistency and unbiasedness become misguided mathematical idealizations. They are misguided because they do not take into account the fact that even if every team in this sequence has reached the perfect answer given their available information and resources, the lack of mutual knowledge can still make the final output significantly inferior to that possible using all the information available to every team. Yet it is clear that we still can and should have a theoretical foundation for comparing different methods in such environments. In mathematical terms, we need to reformulate our criteria by taking into account additional practical constraints and then seek the most effective methods, instead of comparing methods using a criterion that none can ever satisfy. A general statistical framework for this purpose is now being built. This development can greatly benefit from the input and perspectives of the data curation community, which has a much better understanding of the practical constraints and goals involved in these collaborative research settings.

Conversely, approaches to data curation would benefit greatly from the involvement of statisticians. Scientists and librarians alike often rely on general principles of future utility to base decisions on what to select and on what to keep, rather than on analyses of the actual trends in data or on demonstrated utility. As a concrete example, at the Center for Embedded Networked Sensing, a five-university NSF Science and Technology Center based at UCLA, the involvement of a statistician (Mark Hansen, a suggested Seminar attendee) midway through the Center's lifespan radically changed the course of data collection and data curation. Scientists changed their data collection, storage, and retrieval methods, and involved their information science partners in developing better data curation and management methods.

In a nutshell, Radcliffe Exploratory Seminar provides an ideal forum for intense interdisciplinary exchanges on emerging challenges that truly require collaborations from multiple disciplines in order to make meaningful headways. As far as we are aware, if funded, this would be the first workshop that brings leading computer scientists, information scientists, natural scientists, and statisticians under one roof to address some of most intellectually stimulating and practically challenging problems of the information age.

0.6. *Participant Comments*

Hogg–Astrometry.net–just did it because they could, now they see it's important. Q. What's the difference between provenance and metadata? Very interested in being able to reconstruct the reasons people took data. How did they do what they did? In Astronomy–you never get the photons again–one chance only. Strong relationship between data curation and software curation. Knowledge about data is embedded within the software.

Gil more information about the data, the more we can write software and intelligent systems that can assist scientists; provenance standards–including metadata from people with different expertise; opportunity–discovery informatics–helping scientists with intelligent assistance; how to preserve connections between data & models? (e.g. "Eureqa" system); will post link to workshop on discovery informatics; many problems are social rather than technological

- Crosas mentions system similar to Eureqa in social science

- DOE-funded software innovation project at Michigan mentioned by Hedstrom–looking for "models" of data–do they need to be "scientific" or "physical" or can they just be "statistical" (with no *a priori* knowledge of phenomenology).

- Cranmer: thinking about the "multi-phase" nature of modeling–many steps are taken along the say, some may have more "statistical" nature and some more "scientific"–the distinction between these two is clearer at the higher level, where scientific models are needed to connect disparate data sets/results

- (Hedstrom) Is social science (and maybe life science?) different–in that there isn't necessarily an underlying theory–maybe there it's more statistical/empirical from the start

Groth – offers definition of provenance information– rough quote (fix later): "Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness." (one source) We need "Data Connoisseurs" obsessed with data & provenance. Biggest opportunity: (re-)using other people's data, assuming it has enough provenance to use! Danger: more bad results, due to poor use of others' data–negative impact on science overall

- question from Cranmer–will the run-of-the-mill physicists/scientists ever use provenance systems properly? (psychology/sociology questions!)
- (Hedstrom) what do connoisseurs need, in comparison with "regular" people
- Hogg-difference between "goods" and "data"–data is easier to spoof, harder to control

Pepe – mention of CDS *manual* curation, enriching documents with hand-done links to astronomical sources; discussion of ADS All-Sky Survey, which uses manual curation to make all-sky heat maps of the sky, showing where it's been studied, when, and why; then discussion of automated image-extraction and solving for the images positions/scale (astroreferencing), and how that led back to manual curation, using Zooniverse to create the "oldAstronomy" platform where citizens will enter the necessary metadata to make images useful. Point about authorea.com is that it will allow for automated provenance in the future.

- question about reliability of classification from Kayshap–answer is 3 people do each image
- Hedstrom mention of ebird as another good citizen science platform, in addition to Zooniverse
- Hogg mention of value of people (connoisseurs) talking about data sets, Pepe mentions "TALK" page at Zooniverse, that looks great & performs this function

Kayshap – work as a calibration scientist for Chandra is relevant to this discussion; perspective from someone who thinks about the "measurement to data" part of this situation; regime of "small data" where every photon matters needs special consideration, because each bit is precious, unlike the case with "big data" sets

Mahabal – transient science (in big surveys)- must make decisions in ∼real-time about what is interesting; archival information is critical – to put the new data into context of old data, immediately, in order to make these decisions; idea of a "portfolio" for each object

Siemiginowska – high-energy astrophysics has an amazing archive, "HEASARC", 50 years worth of data, to use this it is critical to understand the metadata and to have the original software used to reduce the data (but danger is that the old s/w is not usable, so how does one reproduce the analysis?) Danger: what about people who use a hybrid system of computers and "paper" to do their research–even young postdocs–and then their research cannot be reproduced? Shouldn't we **educate** people to work in a different way to make their research reproducible. 2nd issue: people who write their own s/w vs. people who write/use software meant for a group–the latter is typically better-documented & what is value in that?

Borgman – big data is hard, but little data is harder, since there's so much less consistency and regularity in the small data case. Basic premises: invest in metadata on ingest OR "google model" chuck it all in & try to make sense of it "later." (. . . then there's "digital archaeology"..) Opportunity: Grab data early in the life cycle. That's also the DANGER–if you do this early, and the data are "dirty," they won't get cleaned up early. AG silent comment: that's like putting GPS on digital cameras, built-in!

Crosas – her IQSS team builds tools to solve all these problems! trying to add-on functionality to alleviate extra burdens on researchers "at the end". Opportunity: "Sustainable science." Automating the process as much as possible will help this happen.

Hedstrom – started as an archivist, including archiving paper, and "electronic records" Interested now in "general" vs. "(discipline-) specific" tools for archiving. IGERT program trains many varieties of students. Works with different communities of scientists (life sciences where data deposit is required with publications to materials scientists who have no idea, essentially, what data reuse means). Works on SEAD, Sustainable Environment Actionable Data–massive data integration problem. Most important opportunity: re-create an environment where researchers can do research–less time on "data wrangling". Biggest danger: false conclusions from messy data.

Cranmer –described re-analysis of archived high-energy physics data (and software) that could have led to Nobel Prize (if particle hd been found in the data!). Quick discussion of "Collaborative Statistical Modeling" (see link), showed (network) graph of multi-phase analysis of a huge amount of data by a very large amount of people, leading to better statistical limits on Higgs Boson. Also, service (prototype) where theorists can go get data to test their models.

- Groth mentions European LEAD factory from pharmaceutical companies that led people run models over data & get results on how predictive they are, without direct access to all the data.

Di Stefano –opportunities in data sets that can find important phenomena like gravitational lensing (Pan-STARRS, LSST), only a tiny fraction of what's possible have been realized–lack of curation seems to be the problem

Slavkovic –more after lunch