Prospectus

Sebastian Benthall

School of Information, UC Berkeley

sb@ischool.berkeley.edu

1. OVERVIEW

The Berkeley Institute of Data Science (BIDS) is a new kind of academic institution, a data science environment (DSE). As part of its structure and mandate, part of its purpose is to monitor, understand, and evaluate itself as a research facility. Its ethnography and evaluation working group (E&E) serves this purpose.

This raises a critical question: how does one quantitatively evaluate a DSE? Evaluation of other research environments has depended on easy to measure quantities such as number of produced papers or grants awarded. That these kinds of metrics can lead to perverse incentives is widely known if not readily acknowledged. The current disruption of the academic publishing industry and emphasis of DSE research on software production further complicate traditional metrics of success. At the same time, E&E has expressed a readiness to instrument the DSE in ways that have not been possible with other research labs. The DSE will itself be the source of a massive, noisy, complex data set about its own inner workings and collaborations. This calls for a data scientific response to the problem of DSE evaluation.

The circularity here–that in order to evaluate the success of the data science environment we need to define and implement perhaps novel data scientific techniques–demands that we approach this problem with rigor and humility. I propose approaching this problem as one of iterative algorithm design. My dissertation will aim to contribute to this effort by taking on an iteration of this work. In doing so, I hope to address not only the immediate needs of BIDS, I hope for this work to address broader themes. For example, this work may contribute to the answer to the still mysterious question, "What is data science?"

1.1 Scientific Python

BIDS is still an emerging and amorphous organization. It has twelve co-PIs from disciplines as diverse as statistics, physics, biology, public policy, computer science, and history. Though it will have a physical space in Doe Library, its work will necessarily draw on resources beyond it. This complicates the question of what the "site" under study is for this research.

Pending the creation of a more concrete instantiation of BIDS, I will focus my attention on the Scientific Python community and especially the I Python project. Fernando Perez, who leads the team developing I Python, is one of the BIDS co-PIs. Arguably, one of the purposes of BIDS is to institutionalize in an academic setting the practices that have already guided the open source scientific programming community for years: building a sustainable and accessible ecosystem of tools for working with data and computing scientific results from it.

Scientific Python consists of a number of loosely connected projects, including NumPy, SymPy, Matplotlib, and others which jointly provide functionality similar to other commercial scientific computing projects like MATLAB and Stata. These projects provide a foundation for further scientific computing libraries that encapsulate specific methods–such as machine learning, natural language processing, and network analysis–as well as domain specific models–as in physics, biology, or ecology. Meanwhile, the I Python Notebook project provides a way of publishing scientific results that includes the software it depends on. This is one promising response to critiques of computational science publishing practices (cite Stodden) and to some of the stated goals of BIDS.

If the committee accepts the premise that the Scientific Python community is an important historical precursor of or ingredient to robust academic data science practices, that justifies focused study on it as a way of gaining purchase on broader questions I'm hoping to address. The danger of this approach is that the specificity of the Scientific Python community culture will not generalize to scientific practice or even research engineering efforts in other programming languages.

1.2 Research Questions

As stated, I conceive this research as an iteration of algorithm design. I see this work logically structured into three phases: interpretive, technical, and reflexive.

1.2.1 "Iteration of algorithm design"?

I mean "algorithm" here in the technical sense common to computer scientists: the specification of programmitic steps to be executed over data by a digital computer. I hope to implement such a thing on data collected from activity logs from scientific infrastructure.

By "design", I am referring to the process of coming to a specification of a process or object through an interpretation of a particular problem or social need and consideration of physical constraints. In this view, design depends importantly on an understanding of the needs and perspectives of those to be impacted by the new process. Hence, my technical work will be logically preceded by qualitative, interpretive work to determine the values that will guide the

design.

By "iteration", I am acknowledging the inevitable imperfections limitations of the design and implementation work. As important as coming up with a "solution" is the eliciting of criticism. The algorithm designed in the preceding steps is intended to impact a community of researchers (either BIDS broadly or the Scientific Python community more narrowly, depending on feasibility). Gathering the feedback of that community as well as other stakeholders (such as software uses and funding agencies) will be important for challenging faulty assumptions that went into the design and guiding future work.

1.2.2 Three questions

The three research questions I will address are:

- **RQ1 (Interpretive):** What core values of BIDS distinguish it from other academic institutions in ways that will enable it to meet challenges to academic legitimacy?
- **RQ2 (Technical):** Taking the goals identified in RQ1 as a starting point, how can they be operationalized as instrumentation of the DSE and algorithmic processing of the data?
- **RQ3 (Reflexive):** When the impacted research communities are made aware of the instrumentation and algorithms developed in RQ2, what is their reaction (both as measured by the RQ2 algorithm and qualitatively)?

Depending on feasibility considerations, I may shift these questions to equivalent ones focused on the Scientific Python ecosystem instead of the intersecting DSE.

As each of these questions build on the previous one, and the outcomes of the first (and even its site) are unknown, the remainder of this document will be provisional. Nevertheless, in order to clarify this research agend and demonstrate its viability, I will elaborate on this project's context, motivations, theoretical and methodological considerations, and my own related prior work as it relates to the three research questions.

2. CONTEXT

In 2013, the Moore and Sloan foundations partnered to fund three pilot Data Science Environments (DSE) in academic institutions–UC Berkeley, NYU, and University of Washington. The project has three stated goals (quoted from Moore foundation public statements):

> - Develop meaningful and sustained interactions and collaborations between researchers with backgrounds in specific subjects (such as astrophysics, genetics, economics), and in the methodology fields (such as computer science, statistics and applied mathematics), with the specific aim of recognizing what it takes to move each of the sciences forward;
> - Establish career paths that are long-term and sustainable, using alternative metrics and reward structures to retain a new generation of scientists whose research focuses on the multi-disciplinary analysis of massive, noisy, and complex scientific data and the development of the tools and techniques that enable this analysis; and
> - Build on current academic and industrial efforts to work towards an ecosystem of analytical tools and research practices that is sustainable, reusable, extensible, learnable, easy to translate across research areas and enables researchers to spend more time focusing on their science.

The foundations are funding the first three experimental DSE's explicitly as institutional experiments. Each DSE is comprised of working groups that will address these goals as research problems in their own right, while at the same time becoming a hub for advances in the specific subjects sciences. The working groups will cooperate cross-institution.

An early and interesting result of DSE planning is the enthusiasm the affiliated faculty have for instrumenting the environments themselves. It has been suggested that everything from GitHub pull requests to data from instrumented lab seats be made available for researchers studying doing ethnographic and evaluative work. This presents an interesting opportunity for the Ethnography and Evaluation Working Groups (E&E), whose responsibility is to both to identify data scientists' bottlenecks and to report back to the foundations about their goals.

The puzzle posed by the task of evaluating the DSE is that the more instrumented the environment, the more a E&E must deal with the same problem the DSE's are designed to solve: "the multi-disciplinary analysis of massive, noisy, and complex scientific data." In theory, any and all data necessary to evaluate the effectiveness of the DSE's in meeting their goals is available to the researchers involved. That puts the burden of evaluation on defining success, operationalizing it as metrics, and implementing those metrics as instruments on the DSE.

2.1 Subtext

While the official Moore and Sloan documents frame the DSE experiments uncontentiously, it is worth noting that the Moore and Sloan foundations are vocal about instigating institutional change, and this is politically challenging. In publically available statements (cite 'Launch of the Berkeley Institute for Data Science' video), Saul Perlmutter, the current faculty director, points to several research questions BIDS raises:

- Given the critique that the peer-reviewed journal system winds up with a system where new generations forget the research that was done 20 years ago, how can scientific publication in a data-rich envirionment make scientific progress more cumulative?
- How can scientific practice keep up with the demand for scarce programming talent?
- How does the success of open source software development change what we think of as ideal scientific practice?

Some of the subtext around BIDS has to do with the attractiveness of scientific careers to talented programmers relative to industry jobs. Perez (cite) has argued that the incentives in academia are toxic to genuine collaboration and that this drives many people away. Naturally, this criticism is not

universally embraced by academics who have succeeded or anticipate success in the institutional status quo.

Just as academia-as-usual is being challenged by the idea of a new kind of "data science", so too is "data science" under academic and popular critique. One widely retweeted tweet (citation needed) jokes:

> "A data scientist is a statistician who lives in San Francisco. Data science is statistics on a Mac. A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."

This raises the question of whether data science is worth considering scientific field or whether it is simply the application of automated statistical techniques to data. The emphasis on software development is discouraging to many academics who have not been trained in computing and see software development as orthogonal or unrelated to academic progress. Nevertheless, the availability of data and the success of organizations that have been able to leverage it in industry and politics are strongly suggestive that working intelligently with massive quantities of data has great potential for academic science.

Various BIDS co-PI's, in conversation, have openly admitted that they do not know what data science is or what BIDS ought to look like as an institution. We can be confident at least that there is not yet a consensus regarding the expanding availability of data and use of software in scientific practice. This research aims to contribute to that broad conversation.

3. RQ1 - WHICH SCIENTIFC VALUES?

A first question facing the ethnography and evaluation working group is: what is valuable? BIDS stakeholders–the foundation funders, the faculty co-PIs, university administrators–have all expressed that BIDS has tremendous potential to face new challenges and introduce important new institutional changes. But what that means concretely is still unknown. As an indication of this, one of the BIDS working groups is dedicated to the task of coming up with 'alternative metrics' of academic success–this is a research problem in its own right. One wonders how the success of *that* research will be evaluated.

While cynics would (and have, in private conversation) said that BIDS lacks coherence because it is mainly an excuse to chase foundation funding, it is premature to doubt the sincerity of the involved faculty. This will only reveal itself over time, perhaps in light of ethnographic work.

All this complicates the initial research of this dissertation. To approach the first research question, "What core values of BIDS distinguish it from other academic institutions in ways that will enable it to meet challenges to academic legitimacy?" I will break the work down into three distinct tasks. First, I will through participation in the E&E working group do qualitative research working with stakeholders to identify a rough consensus of scientific values. Second, I will do a literature review in philosophy of science, science and technology studies, and the methodological literature to identify scientific values that have the blessing of prior scholarship. Lastly, I will do a qualitative investigation of the existing Scientific Python community as an already existing data science culture with alternative values, as a fallback if the BIDS leadership cannot identify other goals. I hope that these three lines of inquiry will triangulate an answer to RQ1.

3.1 BIDS values

The BIDS E&E working group will be responsible for reporting back to the foundations funding the Institute as well as identifying bottlenecks to the scientific progress. Via my participation in this working group, I will attempt to get access to these stakeholders and perform semi-structured interviews about their goals for BIDS. Taking these interviews as data, I will attempt to identify a rough consensus of what they find important and take this as *what BIDS values.*

I am aware that there is a certain naivete to this approach. It is possible that there is no rough consensus among the BIDS stakeholders. On the other hand, the co-PIs and foundations have already gone through a long process to identify common goals and articulate them. So far in my preliminery qualitative exploration of the issues, I have found that these stakeholders are more cynical or undecided about what they are saying in private than in their public statements, and are apprehensive at being held to account for claims that might be merely performative. It is possible that BIDS leadership will be resistant to the idea of a graduate student interpreting their values for them. I have been told that this phase of my research will need to wait until BIDS has systematized its own leadership and come up with a way to make official statements of purpose, and that this will happen within the next six months. If that happens, my work will be so much easier. I see my proposing of this very task as a research question as itself a kind of performative argument or action research that, if successful, will prod BIDS stakeholders to consensus.

Complicating matters is that BIDS is drawing inspiration not just from its internal leadership but also from interested faculty, students, and staff from the rest of UC Berkeley. I have learned that part of E&E's task is to keep BIDS accountable to stakeholders in the larger university context as well. I am currenly involved in reviewing the roughly 120 responses to the BIDS call to participation, which will be the starting point of further ethnographic work about the role of BIDS in its institutional context.

3.2 Scientific values literature review

Since it is uncertain whether the BIDS leadership will ever be able to authoritatively express what it is it is trying to accomplish, I am in parallel setting about the task of literature review in philosophy of science, as well as science and technology studies, to identify normative claims about what science is meant to accomplish that have been identified by scholars. While this kind of inquiry could expand into a dissertation in its own right, I hope to be able to draw from this literature, albiet shallowly, to add theoretical depth and robustness to what is otherwise empirical work.

I should not here that the research design of this dissertation is based heavily on Habermas' philosophy of science, as articulated in *Knowledge and Human Interests* (1968). In it, Habermas argues that there are three distinct kinds of scientific inquiry, each important but irreconcilable.

- *Hermeneutic* inquiry, whose method is interpretation of texts and speech and whose purpose is shared understanding of values to guide collective action.
- *Technical* inquiry, whose method is positivist experi-

mentation and whose purpose is power through prediction and control of the natural environment.

- *Reflective* inquiry, whose method is critical self-examination and whose purpose is emancipation from the self-imposed limits of ideology.

If Habermas was the first to suggest this trinary breakdown of modes of inquiry, he was not the last. Star (cite 'The Ethnography of Infrastructure', 1999) identifies a similar set of ways to study information infrastructure: as a collection of veridical representations of the world to be interpreted faithfully; as a material artifact with pragmatic properties and material impact; and as trace or record of activities that can be critiqued as evidence for cultural values or conflicts that may not be expressed explicitly in the records themselves.

I have a hunch that data science, if it is to be successful, will need to incorporate elements of all three of these lines of inquiry, which map roughly to stages of iterative software development in e.g. industry practice. (citation needed). As a test and performative argument for this approach, I have structured my dissertation into research questions enacting and combining each of these three modes.

It is worth noting that Habermas has been heavily critiqued and that these critiques anticipate two possible failures of this research question. One critique of Habermasian thought, articulated by Fraser (cite 'Rethinking the Public Sphere: A contribution to the critique of actually existing democracy', 1990), problematizes Habermas's account of the possibility of legitimate consensus through communicative rationality. (cite, theory of communicative action) Echoing the concerns of feminist epistemologists (cite Haraway here?), the anticipation of rational consensus is seen as either bourgeios masculinist impulse or else the proposed mechanism of consensus (in Habermas' early view, the public sphere) is critiqued as exclusionary to marginalized groups, thereby necessitating multiple publics. If the process of deriving consensus values through E&E investigation winds up excluding groups who would otherwise be engaged, that would threaten the legitimacy of the answer to RQ1.

A second critique, articulated by Rash (cite 'Theories of Complexty, Complexities of Theory: Havermas, Luhmann, and the study of Social Systems) drawing on Luhmann (cite), is that social systems, and especially systems *of social scientists*, cannot be understood according to a simple generative logic such as Habermasian rationality or universal pragmatics because they upon self-examination either evolve further complexity or cease to exist. Hence, an evaluation metric, even if ideally constructed and well-intended, will upon consideration by the community being measured become unstable or result in the community's disolution.

As far as I know, these critiques of Habermas have not been put to an empirical test. RQ3 is designed in anticipation of these possible failures of RQ1.

3.3 The values of the Scientific Python community

As noted earlier, the Scientific Python community, through Fernando Perez, is one of the influences on the vision of BIDS. Though I have only prelimenary observations to work with so far, this community appears to be comprised of academic scientists who have adopted the open source software practices and in particular the community-based development model common to Python projects. While there is no reason to believe a priori that these practices will get adopted by BIDS as a whole, it is true that Scientific Python represents an alternative set of practices to science as usual with claims to solve some of the problems faced by normal academic science.

If BIDS leadership and stakeholders are unable to provide a satisfactory answer to RQ1, as a fallback I will work with the more culturally homogenous Scientific Python community to identify *it's* scientific values. I am comfortable with this step because I have been a part of similar communities of practice before and have done academic work on open source communities at the I School.

In particular, the value of *efficient collaboration* is one prized by open source communities. Perez has been outspoken about how this value is lacking from normal academic science, and it resonates well with the foundations' stated goals for DSE's. This also resonates well with the Habermasian framing of scientific inquiry described above. Interestingly, community-based collaboration on software involves interpretive/hermeneutic work (discourse through text to decide on collective action) to achieve a technical goal (working software that recieves some purpose).

Clearly my professional and theoretical background biases my answer to RQ1. I will need to be reflexively explicit about this in my work, drawing on Kelty (cite, Two Bits) (and Coleman – who I still need to read!) and the practicioner literature (Raymond, Fogel) to explain how these communities function and socialize new members. I expect (though need to verify with qualitative research) that the Scientific Python communities are similar to the other documented open source communities and the ones I've experienced myself.

4. RQ2 - HOW CAN THE GOALS OF RQ1 BE OPERATIONALIZED AS INSTRUMENTATION OF THE DSE AND ALGORITHMIC PROCESSING OF THE DATA?

A critically important part of this dissertation is the technical implementation of the values identified in RQ1. This question breaks down into subtasks.

- A *modeling* problem. Given the goals identified in RQ1, what would constitute a viable and internally valid model of the problem BIDS is trying to solve? I anticipate drawing on such mathematical tools such as statistical learning theory, models of information diffusion, and self-excited Poisson processes. [1]

- An *instrumentation* problem. Taking the model identified above, I will need to work with E&E to develop a feasible plan for instrumentation of the DSE that can collect observational data relevant to model verification. This will no doubt involve political challenges and an IRB application.

The intention of this research question is to attempt a data scientific evaluation of data science. In more ways than one, this will be a test of data science itself.

4.1 Massive, noisy, and complex

Anticipating the challenges of evaluating DSE's, note that the evaluation problem is a data scientific problem in the

sense implied by the Moore/Sloan public statements: the evaluation will performed on data that is massive, noisy, and complex. This raises corresponding methodological problems:

- **Massive**. Since the data collected in instrumented environments will be of much larger scale than academic social scientists are used to, DSE data will challenge methods that depend on p-values as a measure of statistical significance. This means that evaluative statistics may require mathematical rigor that goes beyond the normal toolkits of disciplinary social sciences. Taking past work in data science as a cue, DSE evaluation may involve translating concepts from sociology of science used in RQ1 into the operational languages of signal detection and statistical learning theory. It's unclear how this interdisciplinary move can be made convincingly given the current lack of overlap between these academic communities. This social problem is yet another one that DSE's must be expected to overcome in order to competently evaluate their own success as a site of meaningful collaboration between specific subjects and methodology fields.
- **Noisy**. Data drawn from instrumented environments does not have the same affordances as data collected from a cleanly designed experiment. When data scientists work with such data, often they must begin with a preprocessing phase in which they clean the data of measurement errors and try to separate the salient features from the noisy ones. In this stage, the data scientist filters the available data through a model before testing a second model that is the substance of their work. [citation needed] This prior modeling and filtering step creates opportunity for statistical bias that may pollute later results. Unless one is very careful, a researcher can find anything they are looking for in a large and noisy enough data set. Evaluating DSE's effectively will require separating the scientific signal from the noise.
- **Complex**. If an underlying system is complex in certain ways, then researchers studying it are faced with modeling problems that contemporary machine learning techniques might not be able to adequately address. For example, if one of the properties of a complex system is that it can contain positive feedback loops between variables, then techniques for inferring causal structure from data based on causal graphical models will be of limited use because these methods assume acyclic causality. [is this true for frequentist techniques as well?] [SW notes I'm not being consistent about whether I'm testing a robust theory here or using inductive algorithms. Need to specify that induction might be part of any iterative approach. Where is it appropriate to do that?] Meanwhile, many simulated and theoretical complex systems are unpredictable in the sense that they are mathematically chaotic. Hence, the more faithful our models are to the mechanisms of a complex system, the less useful they may be for prediction and control. If evaluation metrics for DSE's are designed to provide a generalizable pattern for replication of future DSE's, then chaotic complexity might be a significant challenge. (c.f. Rasch)

The upshot of the above is that if DSE data is indeed massive, noisy, and complex, the task of using it to evaluating the effectiveness of DSE's may require novel methodological innovation or at least data scientific sophistication. But that means the task of evaluation raises the very problems of technique that might make data science, and hence DSE's, ineffective at scientific progress. So there is the danger from the start that statistical fallacies will make their way into the evaluation process, thereby guaranteeing the perceived success (or failure) or the DSE itself, or having that success depend on noise. Beauty may be in the eye of the beholder. Since whatever metrics that are devised to evaluate DSE's and the data scientists within them are likely to be replicated in future DSE's built around the same model, errors like this at the early stage can have compounded negative effects.

The E&E groups balance quantitative evaluation with qualitative ethnographic techniques. The latter are intended to surface challenges facing data scientists and report on aspects of the DSE's that would otherwise be missed. While necessary, ethnography does not simply by complementing evaluation resolve the problems stated above. That is because quantitative evaluations aim at generalizable criteria of DSE success, while ethnographic results are not normally considered generalizable.

The above arguments show only that the question of how to properly instrument a DSE in order to properly evaluate its success is a deep and interesting problem. Not only is it a question necessary to properly fulfill the Moore/Sloan mission, it is also a question that bridges between specific science and data science methodology in novel ways.

Without an answer to RQ1 in hand, I can point to prior work on this subject only as familiarity with the concepts underlying data scientific inquiry. In the next section, I will briefly discuss the history of data science methods through cybernetics and artificial intelligence. I suggest that DSEs should operationalize their values using the concepts provided by these rich methodological traditions. I will then describe as an example how collaboration might be modeled and instrumented on data from the Scientific Python community, in lieu of BIDS instrumentation.

4.2 A History of Data Science

[N.B. This section as originally written had a number of factual and terminology errors. I'm eager for corrections and references on this material, as I think taking a historical perspective on data science will be valuable both to demystify it and to demonstrate the robustness of its techniques to less quantitatively inclined audiences. What's currently in the section below is something of a 'folk history' that I've pieced together through courses and independent research. Obviously citations are needed and as I do more research on statistical methods I'm finding a lot of holes and misconceptions so please don't take this as verified.]

One important historical precursor to contemporary data science is cybernetics, a transdisciplinary scientific movement born out of MIT weapons labs in World War II. (cite Turner, 'counterculture to cyberculture') Researchers who were concerned during the war with the mechanics of human controlled gun turrets and homing missiles sought to develop a unifying theory of dynamics systems. These theories were consolidated into papers by Norbert Wiener and others. While Drew Conway has characterized data science

as the application of machine learning–the intersection of 'statistics' and 'hacker skills'–to domain science (cite, show venn diagram?), in the 1940s cybernetics researchers were building 'learning machines' designed to use negative feedback loops in an analog system to converge on intelligent behavior.

While the early cyberneticists were involved in designing intelligent analog systems, Claude Shannon's information theory formalized digital communication. Shannon's theory was embraced by Wiener (cite), who had defined cybernetics as the study of control and communication between the animal and the machine. But cybernetics declined in the United States with the rise of symbolic artificial intelligence as a discipline in 1956.(citation needed–follow up on Wikipedia footnote) (Cybernetics continued to be an active research field in Britain).

Symbolic artificial intelligence drove mathematical theory of computation as researchers sought to apply computing to increasingly complex problems. This lead to the development of algorithmic information theory, which was independently discovered in the 1960's Solomonoff, Kolmogorov, and Chaitin. Extending Shannon's idea of signal entropy, these researchers developed general metrics of a representation's complexity that depended on the length of its shortest description in an algorithmic language. Solomonoff developed this measure further into a general theory of computational inductive inference that is both responsive to new evidence and concordant Occam's razor, the epistemic principle that simpler explanations are more likely–a method now known as Solomonoff induction. [I may be overplaying the role of Solomonoff induction here as it doesn't have much practical application despite being beautiful theory showing the ultimate viability of Bayesian inference in the limiting case. Occam's razor is better represented in practice by the Akaike Information Criterion (AIC) or other automated model selection techniques, and in frequentist statistics in the χ^2 test??]

Over time, the limitations of symbolic AI (now sometimes called "Good Old Fashioned AI" or GOFAI) became apparent [cite] and researchers in the U.S. turned back to techniques that had excited cyberneticists, like artificial neural networks (ANN) and other connectionist paradigms. This was possible partly because of increases in computing power which made it more feasible to run large scale stochastic simulations to perform numerical approximations. [cite] Over time, while ANN's proved to be adaptable learning algorithms, they lacked for clear relationship between design and performance. [cite]

Modern machine learning theory came with the shift from symbolic AI and connectionism to algorithms explicitly designed to approximate Bayesian statistical updating. Again, the viability of Bayesian algorithms was due mainly to increases in computational power, which allowed researchers to numerically approximate integrals and sample from complex distributions. One of the appeals of these methods is their statistical soundness; whereas other inference algorithms can be seen as engineering solutions to the inference problem, arguably Bayesian statistics is *the* mathematical principle of inductive inference, and implementations are only correct insofar as they approximate it.

[My background in computational cognitive science and Michael Jordan's statistical learning theory is showing here. Very recently I have been persuaded by frequentist statisticians of the utility of those techniques, but I'm less aware of the history of these and how they are adapted to contemporary data scientific methods. My best source now says Bayesian modeling with frequentist model checking techniques is the way to go.]

It is of mathematical interest that parameter inference algorithms used in machine learning commonly employ stochastic gradient descent, a computational method that is in many ways simply a digital virtualization of the machines designed by cyberneticists to learn through negative feedback loops. It has also been shown that Solomonoff induction and Bayesian updating are two sides of the same coin: when dealing with computable probability distributions, every Bayesian update is also an instance of Solomonoff induction, given the correct choice of universal computing language. Related, notions of Shannon entropy feature heavily in the statistical machine learning literature, for example in its use of 'maximum entropy distributions' to minimize assumptions in model. Contemporary data science methodology is a convergence of many distinct prior mathematical innovations.

[I am underplaying the problems of model selection and the usefulness of frequentist techniques especially when featurization is used to fit more complex data sets into a linear model.]

Another common feature of modern data science is the use of large graphical data sets. Perhaps most famously, the invention of PageRank, a method of determining the relative importance of nodes in a network, lead to the Google Search engine. [Griffiths et al., cite properly] discovered that the steady state of an artificial neural network has the same mathematical property as the PageRank of a network of linked pages–it is the first eigenvector of the connectivity matrix. The fact that matrix eigenvectors–a concept from linear algebra–represent the result of computational learning process driven by negative feedback is another example of the theoretical coherence of mathematical learning theory and its representational power.

[Other things to include in history of data science:

1. Singular value decomposition (SVD) and principal component analysis (PCA) (ugh linear algebra my Achilles heel)
2. Natural language processing (Chomksy and the symbolic AI tradition, actually...huh)
3. Pearl's *Causality*

]

From multiple and various origins, statistical learning theory has evolved into a powerful, statistically sound, and concretely applicable paradigm for explaining both the substance and form of learning. Several distinct branches of mathematics have been discovered and integrated over decades into what are now conventional tools. Along the way, mathematicians have independently derived general theories of computational learning and have built bridges and conversions between them.

These theories of statistical learning, now the baseline tools of data science methodologists, have both normative and descriptive power. As normative guidelines, they have been both proven mathematically to be accurate and have demonstrated themselves to be invaluable in creating powerful tools. But they have also been used by computational

cognitive scientists (e.g. ...) to model the human mind and brain, which results in hypotheses that can be tested in the lab. [cite]

For my dissertation work, I would like to use these data scientific concepts to build a model of DSE performance. My reasons are twofold: first, a confidence in the rigor with which these concepts model learning, and second, their cultural fit with the expectations of data science methodologists, who are more likely to accept a theory presented in their own terms.

4.3 Analysis of Scientific Python community data

As instrumentation of BIDS depends on the existence of an organization that is currently only on paper, I will at first pursue a parallel analysis of Scientific Python. As an open source project adhering to best practices in the field (cite Fogel), Scientific Python communities have publically available mailing lists, version control systems, and issue trackers. This makes *instrumentation* of these communities remarkably easy, as they have been thoroughly tracking sometimes up to fifteen years of scientific collaboration through their normal practice.

The availability of this data shifts the burden of this analysis to the modeling and featurization of the scientific values chosen in RQ1. I will proceed here with the provisionally chosen value of *effective collaboration*.

We can draw some of our modeling assumptions from the structure of the available data itself. Mailing list data can be modeled as a bipartite graph between users and lists where edges are annoted with text and timestamps...

...

[insert preliminery analysis of scientific collaboration on open source mailing lists here]

...

5. RQ3 REFLEXIVITY

After approximately one year of observation, I will intervene on the DSE by presenting the results of the observational study to the DSE and build an interface which makes these measurements transparent to the researchers involved. I will then begin to measure whether this algorithmically mediated self-awareness on the part of DSE participants affects the DSE's ability to attain the operationalized goal. I will also conduct ethnographic work about the response and interpretation of these measurements. The purpose of this phase will be to suggest refinements to the BIDS values and/or their operationalization.

5.1 Stability and problematization

The preceding discussion has outlined the motivation for operationalizing scientific values in the DSE, the challenges presenting research design in the instrumented environment, and the development of modeling tools that are provided by the history of data science itself.

The last third of the proposed research is what I have called reflexive data science–testing the designed algorithm on the DSE by intervening on it with the derived results. There are several motivation for this second step which depend on the currently unknown quality of the output of the first phase of work.

If the algorithm design is successful at capturing the performance goals of the Moore/Sloan funders and other DSE stakeholders, then it will be of interest to DSE evaluators whether or not performance improves, declines, becomes unstable, or is unaffected by having DSE participants aware of how they are being evaluated. I posit that one general desideratum for a performance metric (one which will influence the algorithm design) is that when it is used to incentivize behavior it robustly increases the performance of the system.

On the other hand, far more likely than a successful operationalization of good data science is a problematic performance metric. While I intend to do my best in the algorithm design phase, realistically I can only attempt a first iteration in a PhD dissertation. So as important as testing the impact of the performance metric on the DSE on its own terms, it will also be necessary to use ethnographic methods to look for ways in which the reflective intervention has incentivized perverse behavior or made DSE stakeholders reconsider their goals.

6. BROAD THEMES AND CONCERNS

I hope that this dissertion work will contribute directly to the operational goals of BIDS and its E&E working group. But as the organization of the research I am proposing is rather unusual, I want to be explicit about the several broader themes and concerns motivating this work. Below are four kinds of questions that have been on my mind as I outline this dissertation.

- While idealized views of science from philosophy of science and epistemology describe science in terms of the soundness of its inference processes, fashionable sociology of science and current meta-scientific critiques of publishing practices have shown that often actually existing science falls far short of these ideals. Arguably, this is due to an institutional emphasis on the product of science–paper publication–instead of its process. With instrumented DSE's and transformed publishing practices, we have the opportunity to evaluate scientists according to their virtues as scientists rather than their formal prolificity. Can we use this process to develop better scientific institutions, or will a process-oriented scientific lab be prone to different pitfalls?

- The social sciences are fraught with methodological disputes about the appropriateness and generality of quantitative extrapolations of social data. Just one of the problems with social prediction is that a social system can be unstably responsive to the insights and incentives proposed by those who study it scientifically. (This is especially true in the case of adversarial systems, such as finance.) These methodological questions are unavoidably implicated in the aspiration that DSE's be sites of interdisciplinary research spanning computer science, statistics, physics, biology, and the social sciences. The research design presented above is designed to face these questions head on.

- While it may be possible to find straightforward statistical proxies with which to operationalize BIDS goals, I anticipate that operationalization will raise questions that are of interest to data science methodologists. I hope these results will be interesting to methodologists not only because they will be applied to the methodologists themselves, but also because they raise challenge

problems and result in opportunities for collaboration. For example, if scientific advance is idealized and modeled according to statistical learning theory, is it possible in theory or practice to build a scientific advance detector? How does one account for the contributions of multiple scientists to a process of distributed cognition? Can an understanding of distributed computational processing shed light on this question?

- This research is being conducted as a dissertation for the School of Information, which has only loosely inherited its legacy as a Library school. As research artifacts are now less books and journals and more PDF's and software code, library science has evolved into information organization and retrieval. At the same time, universities have been outsource their critical research information management infrastructure to external businesses. Hence, the academic research process has become increasingly algorithmically mediated, but by algorithms tailored to the commercial interests of their hosts, not the interests of science itself. If successful, the research described above might speak to questions of scientific infrastructure design: how might DSE's build their own information management systems to improve the quality of scientific work?

If allowed to participate in BIDS as described above, I hope to use the opportunity of my dissertation research to contribute directly to the success of BIDS as an institution. I am not interested in working on a dissertation which amounts only to a long book that nobody reads. I would like to help BIDS work better.

7. REFERENCES

[1] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, October 2008. PMID: 18824681.