Evidence-based discovery

Catherine Blake, University of Illinois at Urbana-Champaign

Abstract
Both data-driven and human-centric methods have been used to better understand the scientific process. We describe a new framework called evidence-based discovery, to reconcile the gulf between the data-driven and human-centered approaches. Our goal is to provide a vision statement for how these (and other) approaches can be unified in order to better understand the complex-decision making that occurs when creating new knowledge. Despite the inevitable challenges, the combination of data and human-centric methods are required to understand, characterize, and ultimately accelerate science.

Keywords: scientific discovery, big data


Copyright: Copyright is held by the author(s).

Acknowledgements: This research is made possible in part by grants from the U.S. Institute of Museum and Library Services, (#RE-05-12-0054-12, Socio-Technical Data Analytics) and the National Science Foundation (#0812522, Towards Evidence-Based Discovery). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Contact: ciblake@illinois.edu

1 Introduction
Data-intensive science (Newman, Ellisman, & Orcutt, 2003), in which information and insight emerges from the data, as opposed to hypothesis testing strategy has dominated conversations throughout the sciences for at least the last decade. The “data deluge” (Hey & Trefethen, 2003) is now a fundamental characteristic of e-science and “big science,” especially in disciplines such as in cancer (e.g. National Center for Biotechnology Information), astronomy (e.g., the Sloan Sky Survey), and atmospheric science (e.g., climate models). In the US, much of the funding for data-intensive science was funneled towards the hardware, software and networking infrastructure to support what was called “forth generation science” (Hey & Trefethen, 2003); however, funding was has also been provided to non-computational activities in the information lifecycle. For example, the “DataNet” program was created to integrate “library and archival sciences, cyber-infrastructure, computer and information sciences, and domain science expertise to provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline” (NSF, 2007). Lastly, investments have been made in training a knowledgeable workforce throughout the information lifecycle in particular to support from data curation (Palmer, Renear, & Cragin, 2008; Fearon et al. 2010) to analysis and synthesis (Blake & Pratt, 2006a, 2006b) has also been recognized as we transition towards data-intensive science.

This trend towards data-driven science has not been without controversy, with much of the resistance directed at a magazine article entitled “The End of Theory: The data deluge makes the scientific method obsolete” (Anderson, 2008). The article argues that rather than following the long tradition in science that recognizes correlation is not causation, we should instead consider correlation enough. That we can “throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.” (ibid).

Anderson fails to recognize the importance of knowing where “the numbers” come from and the context in which they were collected. Consider for example “the numbers” that measure how many times an article has been cited. These numbers are particularly important in science where they play a major role in a scientist’s personal (i.e. promotions) and professional (a scientist’s reputation) life. Now let’s place those numbers in the context of how they are obtained. Figure 1 shows the number of citations made to Anderson’s article from web of science, which has historically been the primary source for bibliometric data, but google scholar has started to provide a similar service. Figure 1 shows that without reconciling the different title, volume, and name “the number” could be 1,2,4,7,11 or 20, which differ from the 320 citations reported in google scholar.

This need to contextualize how the data is collected has not overlooked by good statisticians. Box who is infamous for his insight that “All models are wrong, but some are useful” strongly advocates that “the statistician should be involved not merely in the analysis of data, but also in the design of the experiments which generate the data” (Box & Liu, 1999). The paper demonstrated this idea using a series of
experiments that tightly coupled the analytical method (in this case statistics) with domain knowledge (aerodynamics) in order to solve a particular problem (creating a stable helicopter). In contrast to Anderson’s premise that data is just lying around and can be picked up and re-used without knowing where it came from, Box’s paper shows that the actual scientific process is an iterative activity where the data from one experiment leads directly to the next.

Similarly to statistics, the Knowledge Discovery in Databases (KDD) community is “concerned with the development of methods and techniques for making sense of data” (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The KDD process is described in five stages where data is selected, preprocessed, transformed, mined and interpreted. Although much of the work involved in KDD occurs during early stages of the KDD process, where one author claimed that 60% of the effort (Cabena, 1998) goes into data preparation, early stages of the KDD process remain largely undescribed in the literature. One exception was the opening presentation for the KDD challenge in 2000 that provided high-level details of pre-processing (even if the 800 hours spent on pre-processing (versus 1000 hours of computation) was only included in the “extra slide” section). If 40-60% of the effort required in KDD is in the first few stages, you might expect that those steps would be reported and the focus of targeted research efforts, but of the 125 articles published in KDD 2013, only a handful mention cleaning or pre-processing and a level of detail that could be replicated.

Scientists deeply embedded in the data-driven paradigm are starting to recognize the important role that data collection plays in the process. For example, one co-author of the highly cited article in which the KDD process was introduced stated that “careful consideration of the basic factors that govern how data is generated (and collected) can lead to significantly more accurate predictions.” (Smyth & Elkan, 2010). This realization came more quickly in industry (as opposed to academe) where limited amounts of relevant data and complex transformations leave “a gap between the potential value of analytics and the actual value achieved” (Kohavi, et al, 2004). The influx of data cleaning company’s starting in Silicon Valley suggests that activities that occur before the data selection phase of the KDD process should be considered.

Surely someone must be working on better ways to organize and represent data so that less effort will be required in pre-processing, so why aren’t those efforts reported in the KDD literature? The key challenge is that scientists who work on pre-KDD methods engage with a different scientific community. Of course there are a small number of people who overlap, but for the most part, work on databases is reported in...
different set of journals and presented at a different set of conferences than work on data mining. This is not surprising as scientists must demonstrate that they are making progress and taking a reductionist approach can help with this goal (see (de Solla Price, 1986) for a discussion on little versus big science). This micro-level view of reductionism is mirrored at the macro (institutional) level where the majority of scientists receive tenure from an individual unit and interdisciplinary research is conducted within a project, centers or institute.

Given that work on data driven research is distributed amongst different scientific communities within the computational realm, it is not surprising that research on the broader scientific process also spans multiple communities. A conversation on the way in which science is conducted must also include fields where data is not the focus, but rather the human activity and social contexts in which science takes place is emphasized. Studies conducted in the human-centered paradigm employ a different set of methods from those used in data-intensive science. Consider Latour’s primary research questions posed in his book Laboratory Life that seeks to answer “How are facts constructed in a laboratory, and how can a sociologist account for this construction?” (Latour & Woolgar, 1986, p40). Setting aside the distinction between “fact” and “data”, the answers to such a research question should inform the data-driven paradigm, but connecting social practice to the computational models has yet to be achieved. Consider a similar longitudinal study of two individual scientists that took two years to complete and contributed to our understanding of relevance as a process (Anderson, 2005). Again the results of this work which should inform the metrics used to evaluate information retrieval systems, but major differences in methodological approaches in research efforts make that transition extremely difficult.

In some studies data-driven scientists themselves become the object of study and thick descriptions of the “human cyberinfrastructure” (Lee, Dourish, & Mark, 2006) emerge. Other work from social science researchers has focused on how to understand data sharing practices, where questions such as “1) What are the data management, curation, and sharing practices of astronomers and astronomy data centers, and how have they developed? 2) Who uses what data when, with whom, and why? 3) What data are most important to curate, how, for whom, and for what purposes?” (Fearon et al, 2010) are asked. Despite this tight coupling within a project, analyses of how results from one methodological framework can be seamlessly applied to another is typically out of scope. We need a conceptual framework that allows us to consider all research that contributes to the scientific endeavor, regardless of epidemiological commitments.

In this paper we describe evidence-based discovery to reconcile the gulf between data-driven and human-centered approaches to study the scientific process. Our primary motivation is to provide a vision statement for how these (and other) approaches can be unified in order to better understand the complex decision making that occurs when creating new knowledge and to inform the next generation of computational tools that would better support the knowledge discovery processes.

2 Fundamental characteristics of evidence-based discovery

Evidence-based discovery (EBD) provides a terminological shift from “data” to “evidence” in order to better reflect the terminology used by scientists who employ reflexive methods. However, EBD is not simply a change in nomenclature; evidence-based discovery is characterized by having an instrumented scientific process and support for meta-science, where a meta-science approach “conducts research and develops theory around the documentary products of other disciplines and activities” (Bates, 1999).

2.1 Evidence-based practice

Before defining the key elements of evidence-based discovery, we must first provide context with respect to evidence-based practice. Linking research with practice is a growing trend across disciplines from software engineering (John, 2005), to management (Rousseau, 2012) and librarianship (Canadian Library Association. Evidence-Based Librarianship Interest Group., 2006), and health care, where the phrase evidence-based medicine (Cook, Jaeschke, & Guyatt, 1992) is used. One of the cornerstones of evidence-based practice is the process of unifying, sometimes conflicting results that are reported in the literature which often takes the form of a systematic review(Alderson, Green, & Higgins, 2004) or a meta-analysis, which is a systematic review that combines findings using quantitative methods(Davies & Crombie, 1998; Hunter & Schmidt, 1990). Although a systematic review in medicine might include many
different study designs, the results are only unified within a study design. For example the unified results for all the randomized clinical trials will be presented in one section and the unified results of all the cohort studies will be presented in another.

<table>
<thead>
<tr>
<th>Synthesis</th>
<th>Discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong agreement about factors</td>
<td>Little agreement about factors</td>
</tr>
<tr>
<td>Shared vision of experimental methods</td>
<td>Little agreement on experimental methods</td>
</tr>
<tr>
<td>Standard data collection (and sharing)</td>
<td>High level of change in data collection methods</td>
</tr>
<tr>
<td>Many “replicated” experiments</td>
<td>Few repeated experiments</td>
</tr>
</tbody>
</table>

Figure 2. Spectrum between synthesis and discovery

Some might argue that evidence-based practice does not apply to discovery. Figure 2 shows the continuum between the synthesis activities required to support evidence-based practice and the information available in the discovery process. The left-hand side involves synthesis work that can only occur when a shared understanding of key factors exists and when multiple studies explore the same phenomena. The right-hand side shows the other end of the spectrum, discovery, where there is little agreement on the experimental methods that should be used and where there are few “replicated” studies.

2.2 Instrument the scientific process

Motivated by the evidence-based practice movement, the first tenant of evidence-based discovery, is that the scientific process itself should be informed by research about the scientific process. To achieve this goal will require better instrumentation of the scientific process. The scientific endeavor is highly non-linear, complex and dynamic which suggests that the instrumentation necessary will take a variety of forms in order to capture interactions amongst people, information, technology, and policy.

Some scientists have argued that the scientific process is already well documented in the methods section of a journal article; however, this position was scrutinized in the UK report on climate change data that recommended “scientists should take steps to make available all the data used to generate their published work, including raw data” (The disclosure of climate data from the Climatic Research Unit at the University of East Anglia, 2010).

Even if the methods section of an article did adequately describe the experimental design, publication bias, where factors other than the quality of a study influence the likelihood that the article will be accepted for publication (Thornton & Lee, 2000), mean that the set of published articles does not reflect the entire scientific endeavor. One consequence of publication bias is that negative findings take longer to publish, for example clinical trials reporting a significant positive finding take 4.5 years to publish compared with trials that report a null or negative findings that take 6-8 years (Hopewell, Clarke, Stewart, & Tierney, 2003). In one study, articles with identical study designs, but different levels of statistical significance were sent to several different journals. The articles that showed statistical significance were three times more likely to be published (Atkinson, Furlong, & Wampold, 1982). Rosenthal aptly stated that “For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the “file drawer problem” is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show non-significant results.” (Rosenthal, 1979). Clearly publication bias has an important impact on the validity of bibliometric approaches, but our point here is that publications alone are insufficient to instrument the processes used to generate new knowledge.

Consider a study that asked experienced scientists to articulate how they arrived at their research question (Blake & Rendall, 2006). The qualitative analysis revealed that discussions with colleagues, previous projects, combining expertise, and reading literature informed the question explored. The methods section of an article may cite prior work, but it is rare that authors document informal discussions. If the scientific process were fully instrumented, we would be able to learn more about the type of environments that foster different types of scientific progress and such studies would be more accurate than relying on human memory of situated examples. The number of invited speaker series, conferences and workshops held also suggests that human-to-human interactions are a part of the
scientific process, but we do not currently have a way to measure the amount or impact of those activities. To instrument this activity (i.e. discussions) will require extensions in both the human-centered paradigm (consider the practicalities of recording every conversation and the IRB required to ensure privacy) and in the data-driven (consider the background noise while recording a discussion during a conference) paradigms.

Workflow systems such as MyExperiment (www.myexperiment.org) and Kepler (kepler-project.org) are examples of tools that both instrument parts of the scientific process and enable scientists to provide an abstract representation of data processing. Such interfaces (that are also appearing in commercial products such as oracle) have enormous potential with respect to supporting evidence-based discovery and also allude to some of the advantages that instrumenting the scientific process might bring, where scientists are able to share data flows and benefit from being able to use data flows that optimize activities. It is however, unlikely that any single technology will support all of the activities involved in the scientific endeavor; nor that an individual tool would support the range of methods employed to understand science. Thus, we envision that evidence-based discovery will be supported by multiple information ecologies, where an ecology is defined as "a system of people, practices, values and technologies in a particular local environment." (Nardi & O'Day, 1999)

2.3 **Support meta-science**

The complexity and scope of the scientific process transcends multiple analytical methods and philosophical commitments, from the data-driven and human-centered paradigms mentioned earlier to purely theoretical approaches and new analytical methods and epistemological constructs as they are realized. The second fundamental tenant is that evidence-based discovery must not be tightly coupled to an individual discipline, but rather must span disciplines in order to support meta-science.

To illustrate our vision of what support for a meta-science might look like, so we have developed a scenario from the health sciences, but a similar level of complexity would emerge if you were to consider any of the grand challenges that face society in the 21st century such as improving education, or minimizing our impact on the environment. The goal in this scenario is to identify a set of actions that would have the greatest impact in lowering the prevalence of breast cancer.

![Image of a systems perspective to identify strategies that prevent breast cancer.](image-url)
Of the people who have breast cancer approximately half have none of the known risk factors (Vincent, Samuel et al. 2001). If you were to try and identify risk factors for this disease you would need consult the epidemiology literature where you would find the items listed in the top left corner of Figure 3. However many of the factors in epidemiology also manifest in the biological literature. Consider family history from epidemiology that is described as genetic mutations in the BRCA1 and BRCA2 genes in the biology literature. The biological literature would also reveal that breast cancer is a hormone related disease that manifests in the epidemiology literature as obesity because hormones accumulate in fat cells. The family history from epidemiology also manifests in the medical literature as clinical trials because family history is such a strong indicator, that drug used to treat breast cancer are sometimes prescribed before any actual symptoms emerge, despite the known treatment side effects.

A medical solution is certainly one way to prevent the prevalence of breast cancer, but it is certainly not the only way. Consider again the relationship between obesity and breast cancer. One way to lower cancer rates would be to design our urban environments such that public transportation and parks encourage more physical activity. Perhaps a more effective strategy might be explore the connection between estrogens reported in the chemistry literature and food reported in the agriculture literature (certain types of mushrooms lower breast cancer risk). Certain mushrooms only have health benefits when fresh, so such a strategy may also have implications to the food distribution channels.

Both medical and non-medical approaches are being explored within the current domain centric organizations of faculty, methods, data, funding and publications, but it is very difficult to explore and analyze alternative strategies that draw from heterogeneous academic traditions. Moreover, it is very difficult to develop systematic ways of identifying non-medical solutions health challenges within the current organizational structures. We are not suggesting that disciplines be removed as we are in agreement with P.W. Anderson’s assessment that “The reductionist hypothesis may still be a topic for controversy among philosophers, but among the great majority of active scientists I think it is accepted without question” (Anderson, 1972). However, an evidence-based discovery approach that supported meta-science, would provide the same types of support for a scientist either within or between disciplines and perhaps lower the barriers between disciplines. Similar ideas were proposed in literature-based discovery (Swanson, 1986), but in evidence-based discovery the goal is not to uncover new treatments, but rather to better understand the processes used to generate new knowledge across disciplines. Consider for example, an experiment where different groups of scientists are provided with different types of meta-science support in order to understand how different support structures influence the amount or nature of the new knowledge produced. Consider also the co-dependency between the way in which new knowledge is generated and the mechanisms used to disseminate that knowledge (which is also being explored in conferences such as Beyond the PDF).

3 Closing comments

We are under no illusions that instrumenting the scientific process and supporting meta-science in order to achieve evidence-based discovery will be easy. Key challenges in both the data-centered and human-centered paradigms will need to be addressed in order to realize this vision. Moreover some philosophical positions will be difficult to reconcile, such as the following epistemological stance where data is seen as “the product of the reflexive relationship between researcher and researched, constrained and informed by biographical, historical, political, theoretical and epistemological contingencies, data cannot be treated as discrete entities” (Mauthner, & Backett-Milburn, 1998).

Despite the inevitable challenges, the evidence-based discovery framework presented here provides a way to unify research efforts from different traditions, which each provide a glimpse into the underlying cognitive processes used in the scientific process. It is only by combining of these methods that we will be able to understand, characterize, and ultimately accelerate the scientific process.
References

Table of Figures
Figure 1. Context information necessary to accurately count the number of citations to an article (from Web of Science, accessed Sept, 5, 2014)
Figure 2. Spectrum between synthesis and discovery
Figure 3. A systems perspective to identify strategies to prevent breast cancer.

Table of Tables
No table of figures entries found.

de Solla Price, D. J. (1986). Little Science, Big Science ... and Beyond. New York.
Great Britain, House of Commons, Science and Technology Committee Publication, (2010). The disclosure of climate data from the Climatic Research Unit at the University of East Anglia, Eighth report of session 2009-10, HC387-I.


