

# I Don't Know and I Don't Care

Michael L. Nelson  
Old Dominion University  
Norfolk, VA 23529  
<http://www.cs.odu.edu/~mln/>

When tasked to consider if “data-driven science is becoming a new scientific paradigm – ranking with theory, experimentation, and computational science”, I considered both positions carefully. Ultimately, I arrived at the old adage concerning ignorance vs. apathy: I don't know and I don't care.

It may in fact be a legitimate, new, fourth scientific paradigm. Or it could be an extension to and radical transformation of the existing paradigms of theory, experimentation and computation. Classifying the impact of the vast quantities of data on the scientific process makes for an interesting epistemological exercise, but in practice I suspect it just doesn't matter. The reality is that science is becoming data-driven at a scale previously unimagined. The ubiquity of access and volume of data will fundamentally transform the scientific process. Rather than debate the classification of this phenomenon, I think it is more profitable to focus on the challenges that it presents.

## Rudolphine Tables

Data-driven science is not necessarily new – a compelling argument can be made that Tycho Brahe and his assistant Johannes Kepler were doing data-driven science, at least by the scale of their time. Kepler published the Rudolphine Tables in 1627, some 26 years after Brahe's death. The tables were a catalog of stars and planets and were largely based on Brahe's observations, which were considered to be the most accurate and detailed of the time. The Rudolphine Tables formed the core of the data that Kepler used to derive his laws of planetary motion. That the Rudolphine Tables were published at all is amazing: significant infrastructure costs (in the form of purpose-built observatories), professional jealousy, intellectual property restrictions, and political and religious instability dominate its story. Given the historical context, the cost, scale and legal and social concerns involved in the Rudolphine Tables would seem to place it on par with, say, the Google Books project today.

## What Drives Advancement?

I spent 11 years at NASA Langley Research Center. During this time, I learned two important things concerning aeronautics: the pointy end goes first<sup>1</sup>, and the somewhat counter-intuitive idea that advances in air frames are largely driven by advances in propulsion systems (what you and I would call “engines”). The idea is that air frames evolve to a point where engines are the limiting factor. It is not until the engines become

---

<sup>1</sup> As it turns out, experimental aircraft such as the “Oblique Flying Wing” (<http://www.obliqueflyingwing.com/>) invalidate the universality of this lesson.

appreciably more {efficient|powerful|compact} than previous engines that the entire aircraft and its deployment profile can advance.

Data is the engine that drives all scientific paradigms. The scientific paradigms can be differentiated by the amount of data they produce and consume:

- Theory: the primary scientific paradigm; requires little in the way of resources or data to construct models.
- Experimentation: the use of apparatus, artifacts and observation to test theories and construct models.
- Computation: arguably a specialization of experimentation, with the tools focused around the unique opportunities provided by numerical techniques afforded by computers.

At each level, increasing amounts of data are required. It could be argued that more data makes each successive level possible (e.g. from theory to experimentation), or it could be argued that a significant enough change in the volume and kind of data warrants its own description (as computation can be seen as a form of experimentation). The existence of volumes of data alone does not constitute science, and although I cannot imagine a use of this data that does not fit into one of the three categories that does not mean that a new use does not exist.

## Challenges

There has probably always been an implicit demand in the scientific endeavor for the scale of data we currently enjoy. We may lack the historical context to fully judge the transformation and advances available from this unprecedented level of access, but I will assume that the benefits are real and numerable. Instead, I would like to focus on how the challenges created by this new scale of data-driven science.

The definition and dynamics of the scientific artifact are changing.

The scholarly communication process is optimized for information artifacts of a certain size and description. Books, journals, proceedings and reports have a self-contained nature that facilitates publishing, distribution and long-term preservation. The Rudolphine Tables may have been an early example of data-driven science, but observations of some 1000 objects were easily expressible in book form. Software, data sets, multimedia and the like do not neatly fit into the existing practices and are treated as second-class artifacts. Books with long missing CD-ROMs, papers riddled with 404 URLs and quaint phrases like “contact the authors for the complete data set” capture the difficulty the current processes have with the increasing volume and types of data supporting their endeavor. While the scientific process is becoming more data-driven, the scholarly communication process continues much as it has been for hundreds of years,

the process largely automated but still functionally unchanged<sup>2</sup>. We must account for the increasing amount of scientific data and associated artifacts that go uncollected by the current communication process.

Information may want to be free, but data is not as driven.

As the type and scale of the data increases, the difficulty in preserving and understanding it increases: data sets masquerading as books and source code frozen in appendices of journals are insufficient to support data-driven science as it is today. In the Library of Congress sponsored archive ingest and handling test (AIHT), we were one of four teams tasked with “preserving” a medium-sized web site and exchanging our archive with another project participant after one year. The sobering reality was once processed for “archiving”, the exchange of the content was very difficult and required significant manual intervention, despite the level of coordination between project members, the short duration of the project, and the fact that three of the four participants used the same XML encoding scheme (METS).

One of the lessons I learned from the AIHT project was that digital preservation is the software engineering problem writ large. Unsurprisingly, digital preservation as a nascent field has been dominated by the adaptation of traditional archiving practices. I believe it would do well to embrace the software engineering discipline. The division between code and data is somewhat artificial (not unlike the data vs. metadata distinction made in web based information retrieval) and to focus solely upon one without the other is myopic.

Who will capture this data and where will it live?

Not only are the nature and the size of the science artifacts changing, but the manner in which they are acquired and stored changes too. Figure 1 shows an approximation of the cost functions required to acquire and curate scientific artifacts with different data dependencies: the blue represent line represents the cost of the theoretical model: small initial costs, small runout costs; the red line represents the experimental model: high initial costs and small runout costs; the black line represents data-driven model: some initial costs, but the curve is dominated by nearly flat runout costs.

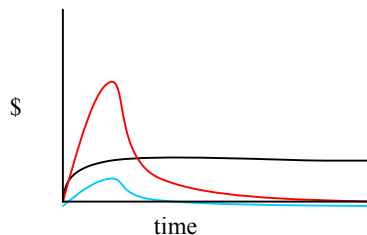


Figure 1: Cost functions for Theory (blue), Experimentation (red) and Data-Driven (black).

---

<sup>2</sup> See the NSF-funded Cornell/LANL “Pathways” Project for a proposed implementation of a modern scholarly communications infrastructure (<http://www.infosci.cornell.edu/pathways/>).

The runout costs are dominated not just by maintaining the repositories, but also in the effort required for acquisition and long-term preservation: refreshing, migrating to new formats, support continued access, tracking changes, verifying provenance – all the things that are currently handled in small scale by the book and journal publisher / library arrangement. The acquisition component should not be overlooked; there are currently many woefully under-filled “institutional repositories” (IRs). The reasons for slow accession rates of these IRs are manifold, but it is probably just as much due to improper incentives and as a lack of technology. In 1994, the National Research Council released a report recommending changes in the academic process to support academicians doing experimental computer science and engineering<sup>3</sup> and their findings should be revisited to determine what can be done to support a data-driven emphasis in all scientific disciplines. Similarly, we should look to the NSF Cyberinfrastructure Report<sup>4</sup> and see where data-driven science fits within the vision they describe.

## Summary

I have consulted with many organizations desiring to set up repositories with an OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) interface. Nearly all organizations ask questions similar to “why should we focus on adopting OAI-PMH when we still need to improve our local search interface?” I always tell them that if they can create a large enough collection of data, someone else will build a search interface for them. Just as new engines cause advances in air frames, large data collections cause advances in interfaces and related systems.

So while I don’t know if data-driven science is a new paradigm, I do care about the data itself: where it will come from, how it will be stored and preserved. Web-scale collections of data will drive new innovations in science. Perhaps all science wants to be data-driven but until now, could not. The story of Brahe, Kepler and the Rudolphine Tables indicates that being data-driven is expensive and difficult, but in the end it is worth it.

---

<sup>3</sup> “Academic Careers for Experimental Computer Scientists and Engineers”, National Research Council, 1994. (<http://books.nap.edu/html/acesc/>).

<sup>4</sup> “Revolutionizing Science and Engineering Through Cyberinfrastructure: report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure; Daniel E. Atkins (Chair), 2003. (<http://www.nsf.gov/cise/sci/reports/atkins.pdf>).