Data Science as a Science: Methods and Tools at the Intersection of Data Science and Reproducibility

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

Berkeley, CA
March 23, 2018

Agenda

1. Framing Data Science as a Science

2. Cyberinfrastructure Tools Supporting Data Science

3. Lifecycle of Data Science as a Framing Tool

Merton's Scientific Norms (1942)

Communalism: scientific results are the common property of the community.

Universalism: all scientists can contribute to science regardless of race, nationality, culture, or gender.

Disinterestedness: act for the benefit of a common scientific enterprise, rather than for personal gain.

Skepticism: scientific claims must be exposed to critical scrutiny before being accepted.

Skepticism and Boyle's Idea for Scientific Communication

Skepticism interpreted to mean claims can be **independently verified**, which requires **transparency** of the research process in publications.

Standards established by Transactions of the Royal Society in the 1660's (Robert Boyle).

Today: Technology drives a re-assessment of transparency

- Big Data / Data Driven Discovery: e.g. high dimensional data,
- Computational Power: simulation of the complete evolution of a physical system, systematically varying parameters,
- Deep intellectual contributions now encoded only in software.

The software contains "ideas that enable biology…"

CSHL Keynote; Dr. Lior Pachter, UC Berkeley

"Stories from the Supplement" from the Genome Informatics meeting 11/1/2013
https://youtu.be/5NiFibnbE8o

Converging Trends

Two (competing?) conjectures:

1. Scientific research will become massively more computational,
2. Scientific computing will become dramatically more transparent.

These trends need to be addressed simultaneously:

Better transparency will **allow people to run much more** ambitious computational experiments.

And **better** computational experiment **infrastructure** will allow **researchers** to be **more transparent**.

Looking ahead…

We imagine a major effort to develop infrastructure that promotes good scientific practice downstream like transparency and reproducibility.

But plan for people to use it not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work.

This infrastructure is used because it enables **efficiency** and **productivity**, and **discovery**.

Infrastructure Innovations

Research Environments

Verifiable Computational Research	SHARE	Code Ocean	Jupyter
knitR	Sweave	Cyverse	NanoHUB
Collage Authoring Environment	SOLE	Open Science Framework	Vistrails
Sumatra	GenePattern	IPOL	Popper
Galaxy	torch.ch	Whole Tale	flywheel.io

Workflow Systems

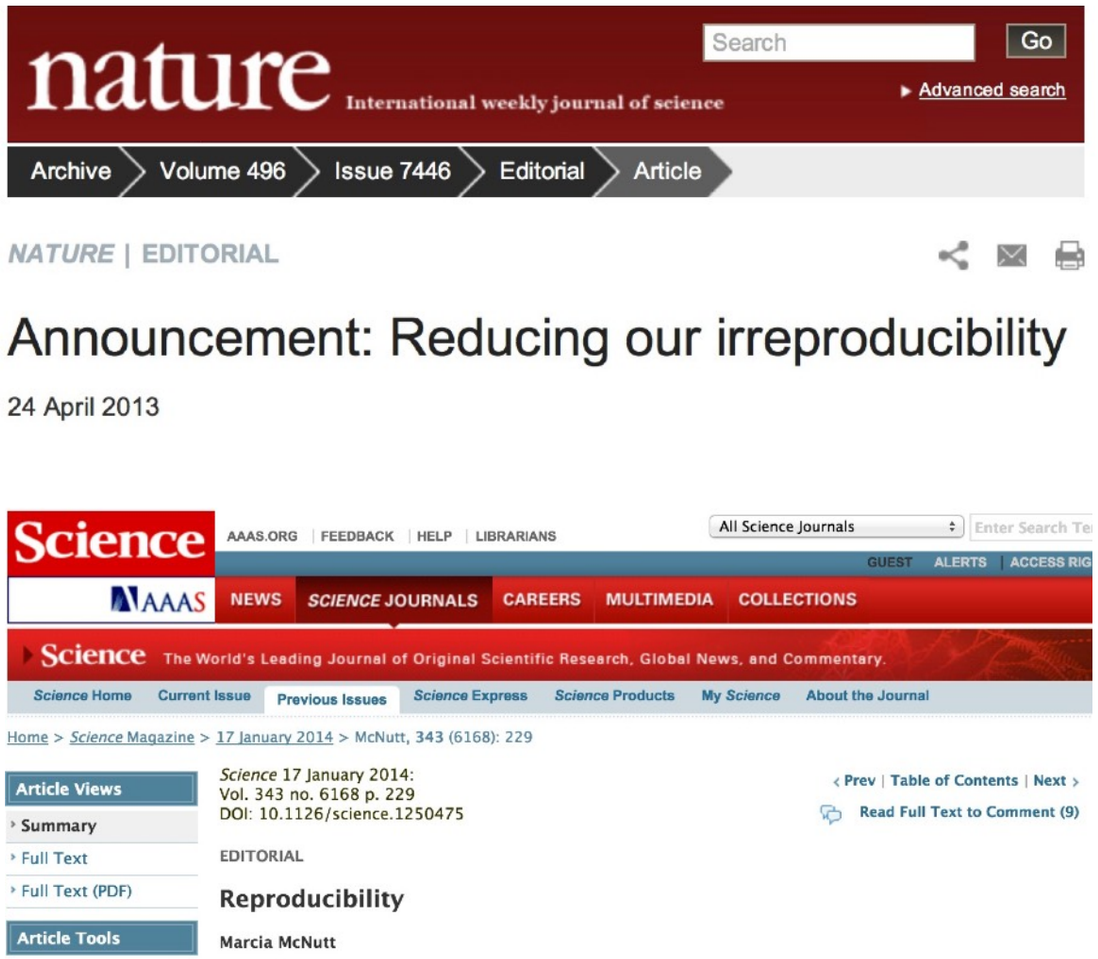
Taverna	Wings	Pegasus	CDE	binder.org
Kurator	Kepler	Everware	Reprozip	

Dissemination Platforms

ResearchCompendia.org	DataCenterHub	RunMyCode.org	ChameleonCloud
Occam	RCloud	TheDataHub.org	Madagascar
Wavelab	Sparselab		

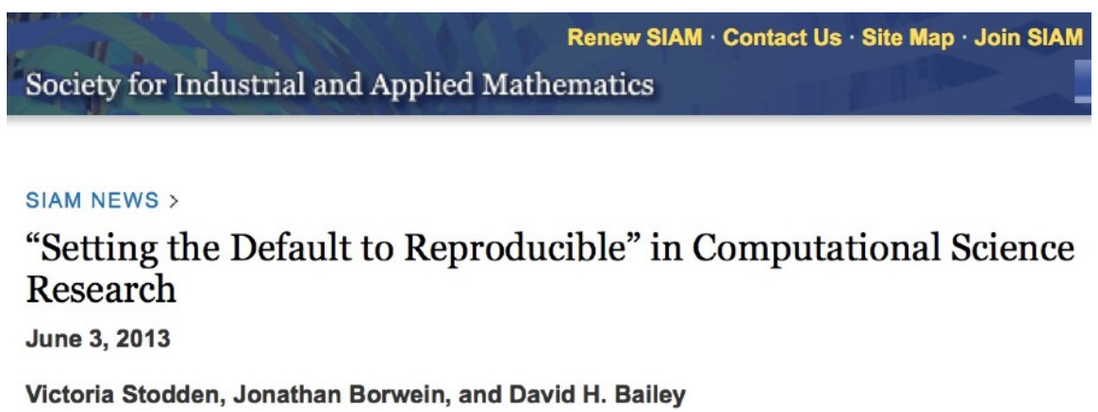
Parsing Reproducibility

“Empirical Reproducibility”



“Statistical Reproducibility”

“Computational Reproducibility”



V. Stodden, IMS Bulletin (2013)

Training for DataScience on New Tools and Practices

- Information extraction from data is a *science*, and so the processes are expected to be transparent,
- Therefore training must extend to include computational methods and tools used in scientific research, as well as theory and computational techniques.

Life Cycle of Data Science

Example:

Recommendation:
Structure curricula around the skills represented in the Life Cycle of Data Science (note. the lifecycle is not uniquely defined)

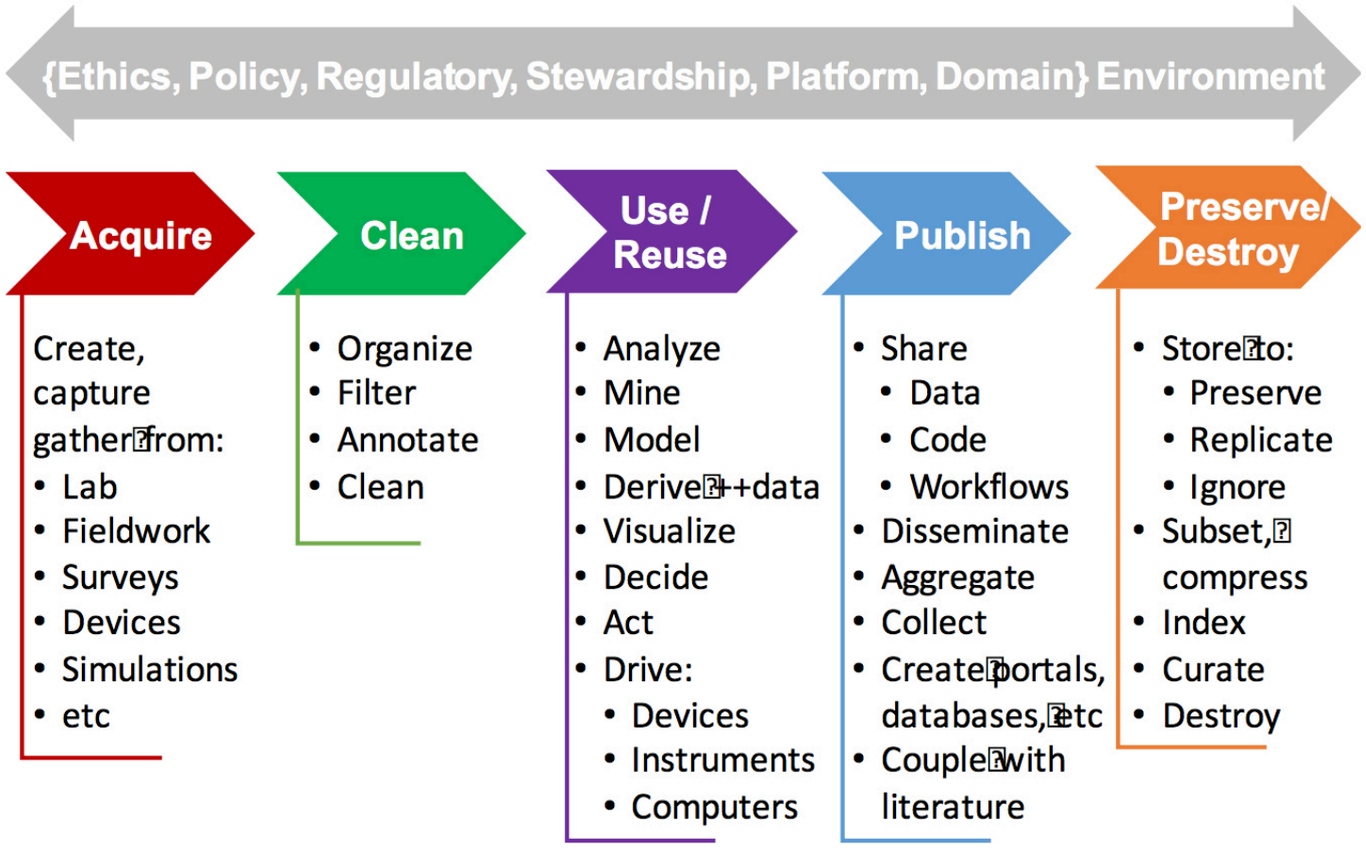


FIGURE 1: The Data Life Cycle and Surrounding Data Ecosystem

The *Data Life Cycle* is critical to understanding the opportunities and challenges of making the most of digital data. **Figure 1** shows a simplified cartoon with essential components of the data life cycle. Data is *acquired* from some source (measured, observed, generated), *cleaned* and edited to remove the outliers inevitable in real-world measurement scenarios and render it suitable for subsequent analysis; *used* (or reused) via some analysis leading to insight, action, or decision; *published* or disseminated in some way so the community at large is made aware of the data and its outcome(s); *preserved* (or not) so that others can revisit and reuse this data now or in the future. Surrounding this overall pipeline is a broader *environment* of concerns: *stewardship* to maximize the quality of the data and promote effective use, *ethics* issues that touch on proper or improper actions with these data; *policy* and *regulatory* constraints that impose legal limitations on these data; *platform* and infrastructure issues that affect technically how we can work with data; and *domain* and disciplinary needs specific to the application communities that create, operate, and use the data from these pipelines.