Integrating Replication Tools with Data Repositories

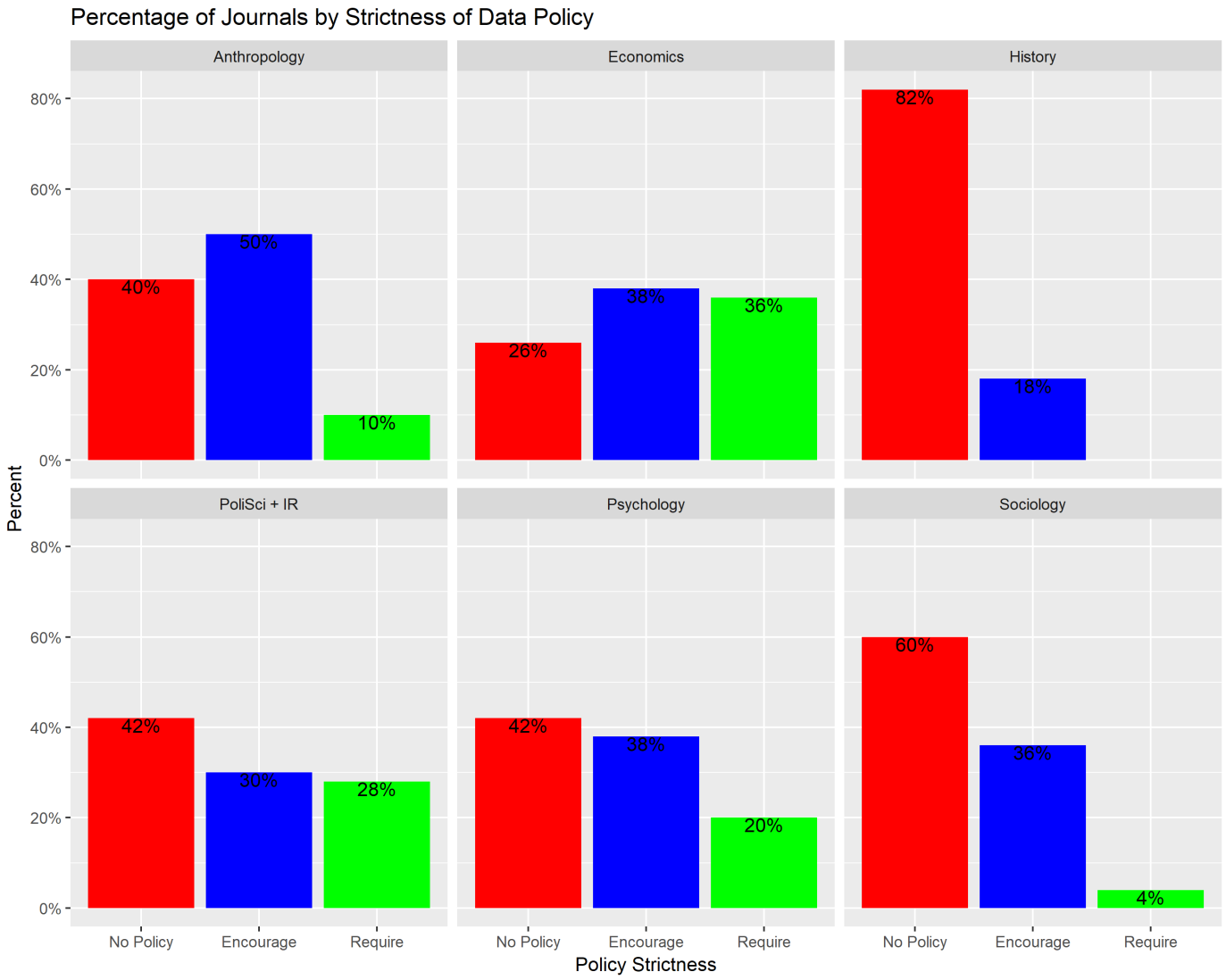
Matthew K Lau and Mercè Crosas
Institute for Quantitative Social Science (IQSS)
Harvard Forest
Dataverse
Harvard University

Increase in Data Sharing

In the last decade, data sharing has become increasingly common in the social sciences:

- More journals are adopting data policies
- More data repositories are available to share research data

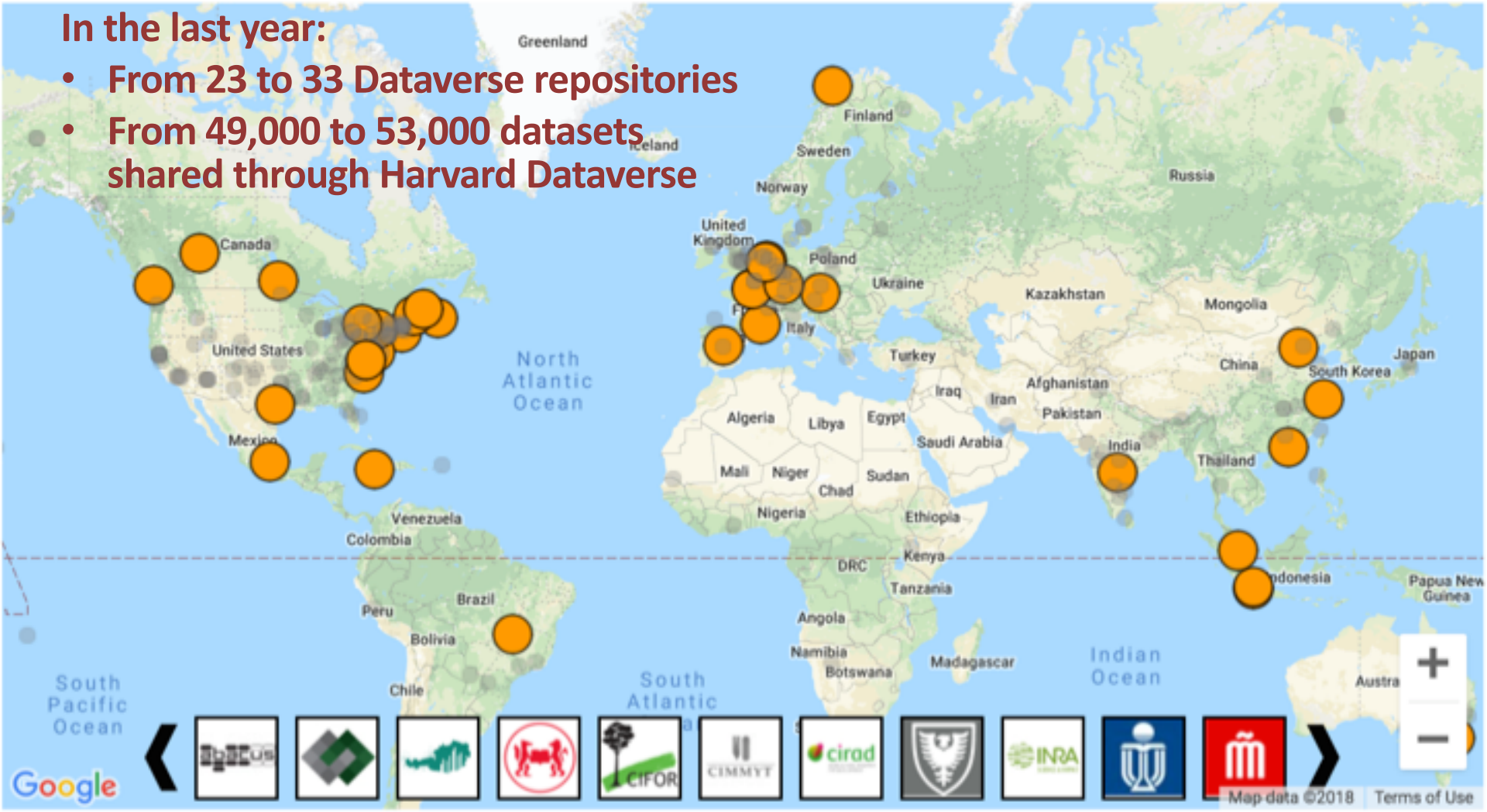
Data Policies in Social Science Journals



More than 50% of the top 50 journals in anthropology, economics, psychology, and political sciences have **data policies that either encourage or require to share the data** associated with the article.

Crosas, Gautier, Karcher, Kirilova, Otalora, Schwartz, 2018. Data Policies of highly-ranked social science journals

Data repositories around the world powered by Dataverse



https://dataverse.org/

But, are these shared data reusable?

- Shared data should be reusable to:
  - Replicate previous studies
  - Conduct new studies using existing data
- For this, we need:
  - Well-documented, well-organized data and code
  - Tools that facilitate replication and reuse

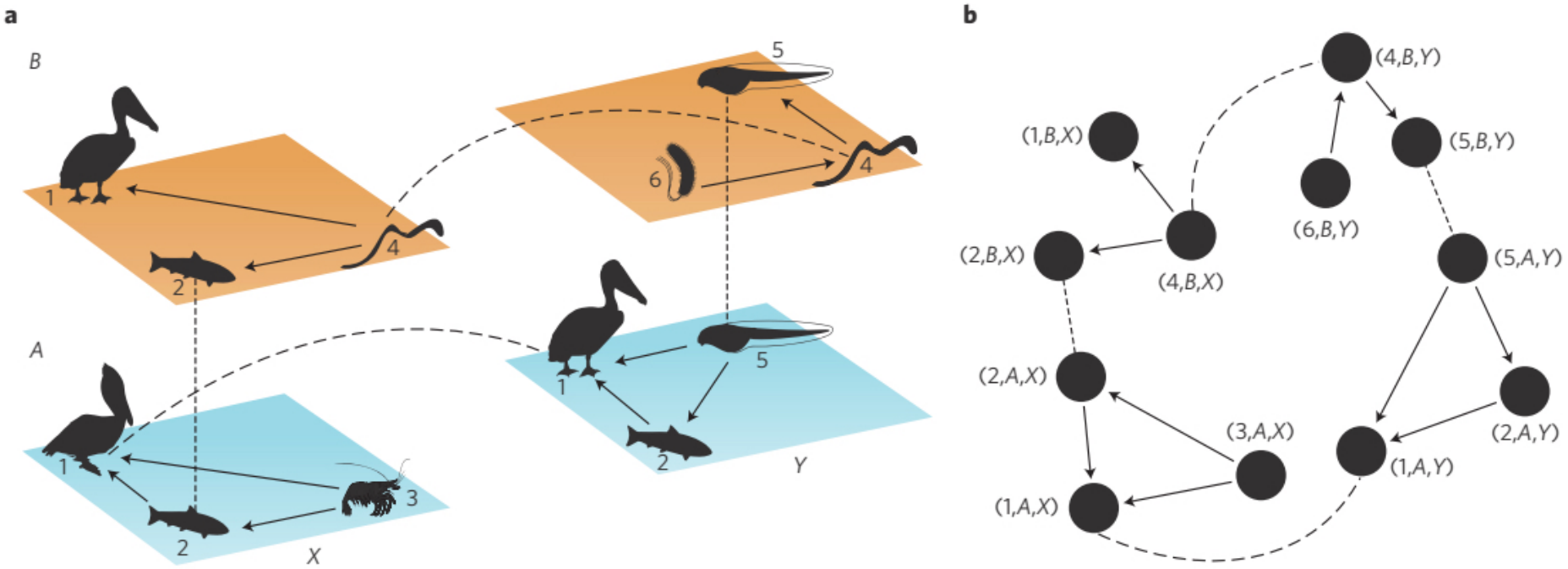
"Increasing Scientific Dataset Quality Through Reproducibility and Curation Tools and Targeted Services in Dataverse Repositories"

With funding from the Sloan Foundation, we plan to address data reuse and reproducibility by:

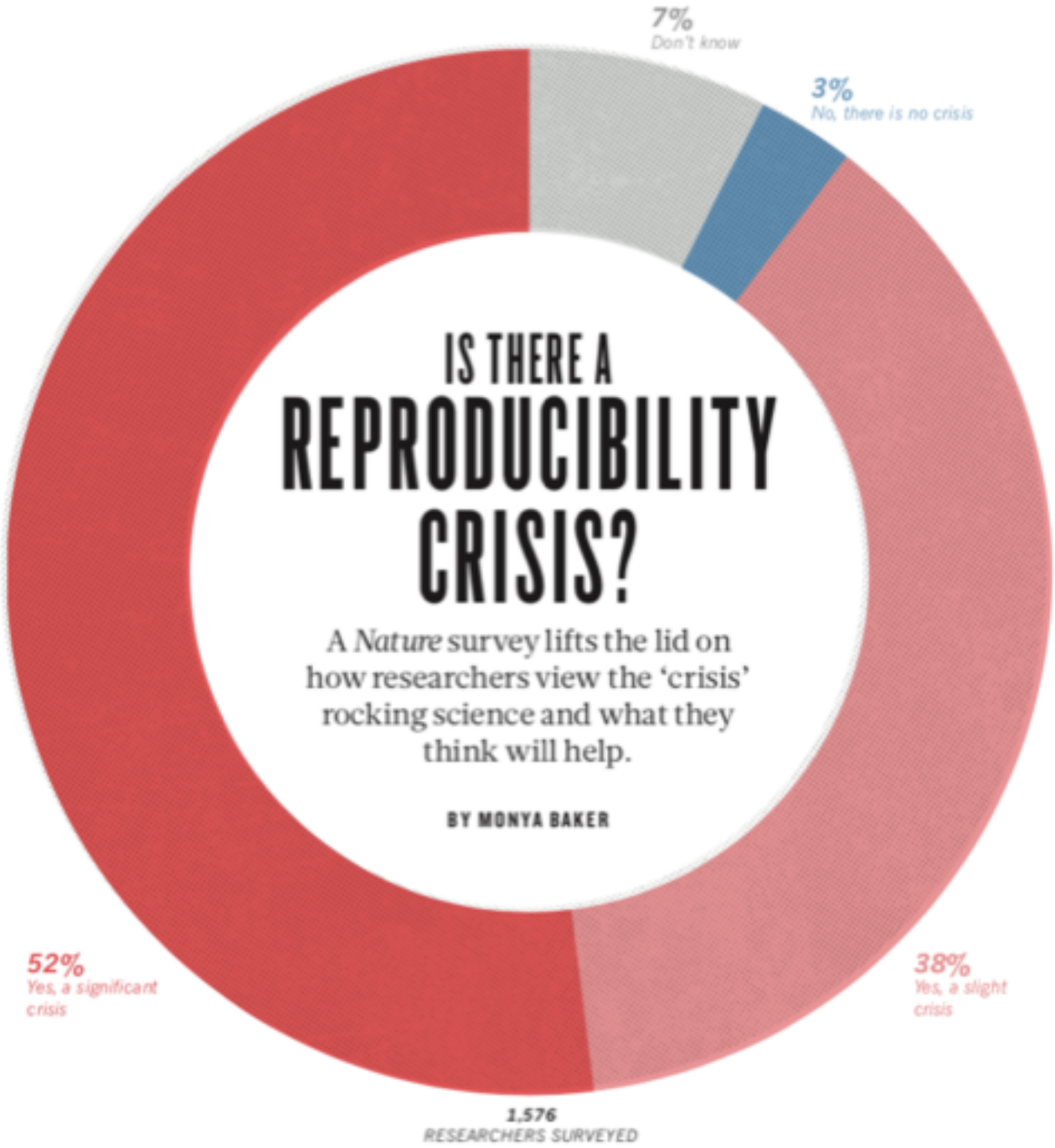
- **Improving curation** through educational materials, friendly user-interface, and services
- **Integrating replication tools** with Dataverse repositories:
  - **Encapsulator** to pack your data and code in a self-contained, documented capsule ***(this talk)***
  - **Code Ocean** to easily run scientific code online
  - **CoRe2** to connect systems in order to streamline the verification workflow (ODUM Institute)

ALFRED P. SLOAN FOUNDATION

Easier reproducibility for scientists with encapsulation

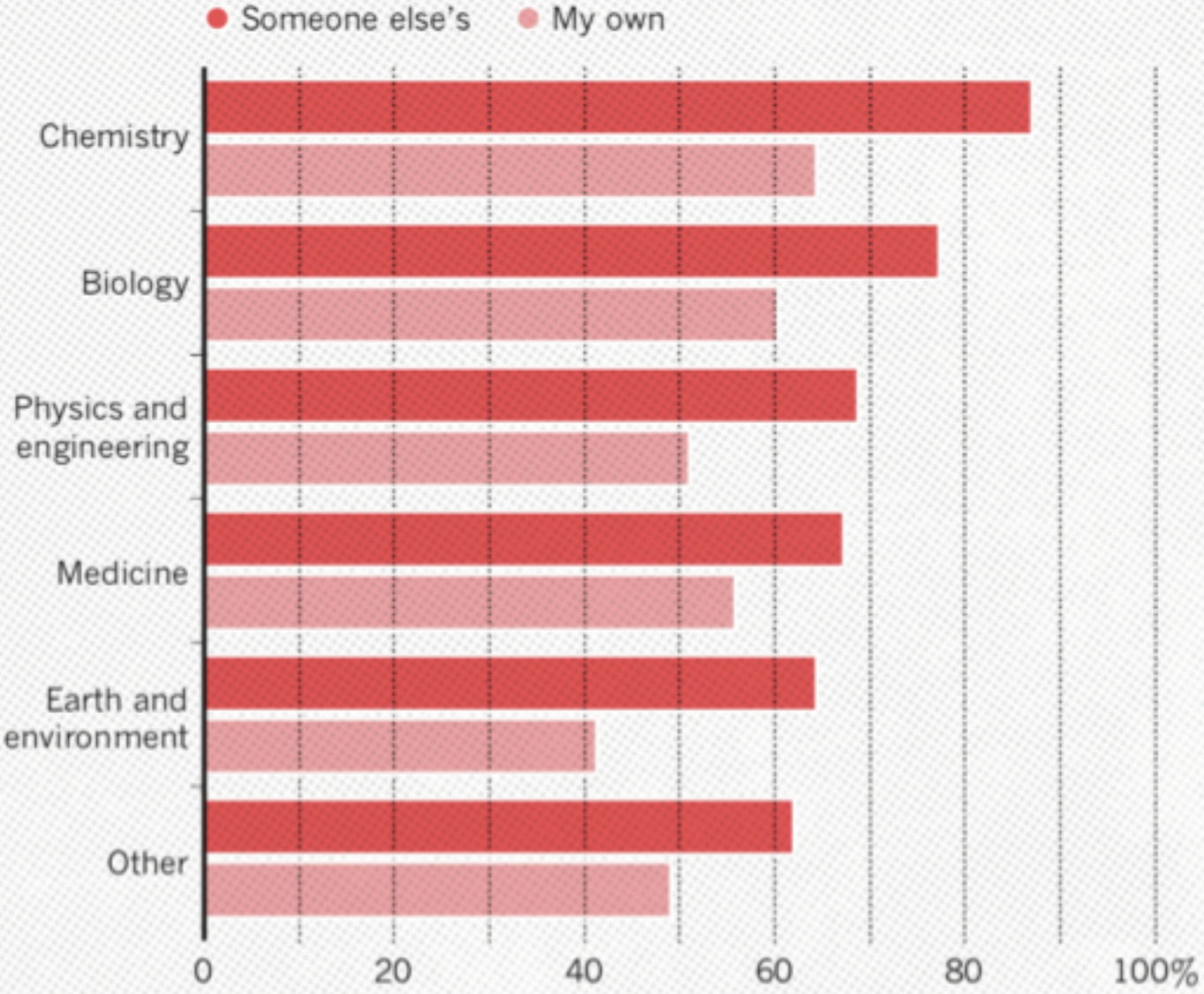

a
B
A
X
Y
1
2
3
4
5
6
b
(4,B,Y)
(1,B,X)
(5,B,Y)
(6,B,Y)
(2,B,X)
(4,B,X)
(5,A,Y)
(2,A,X)
(3,A,X)
(2,A,Y)
(1,A,Y)
(1,A,X)

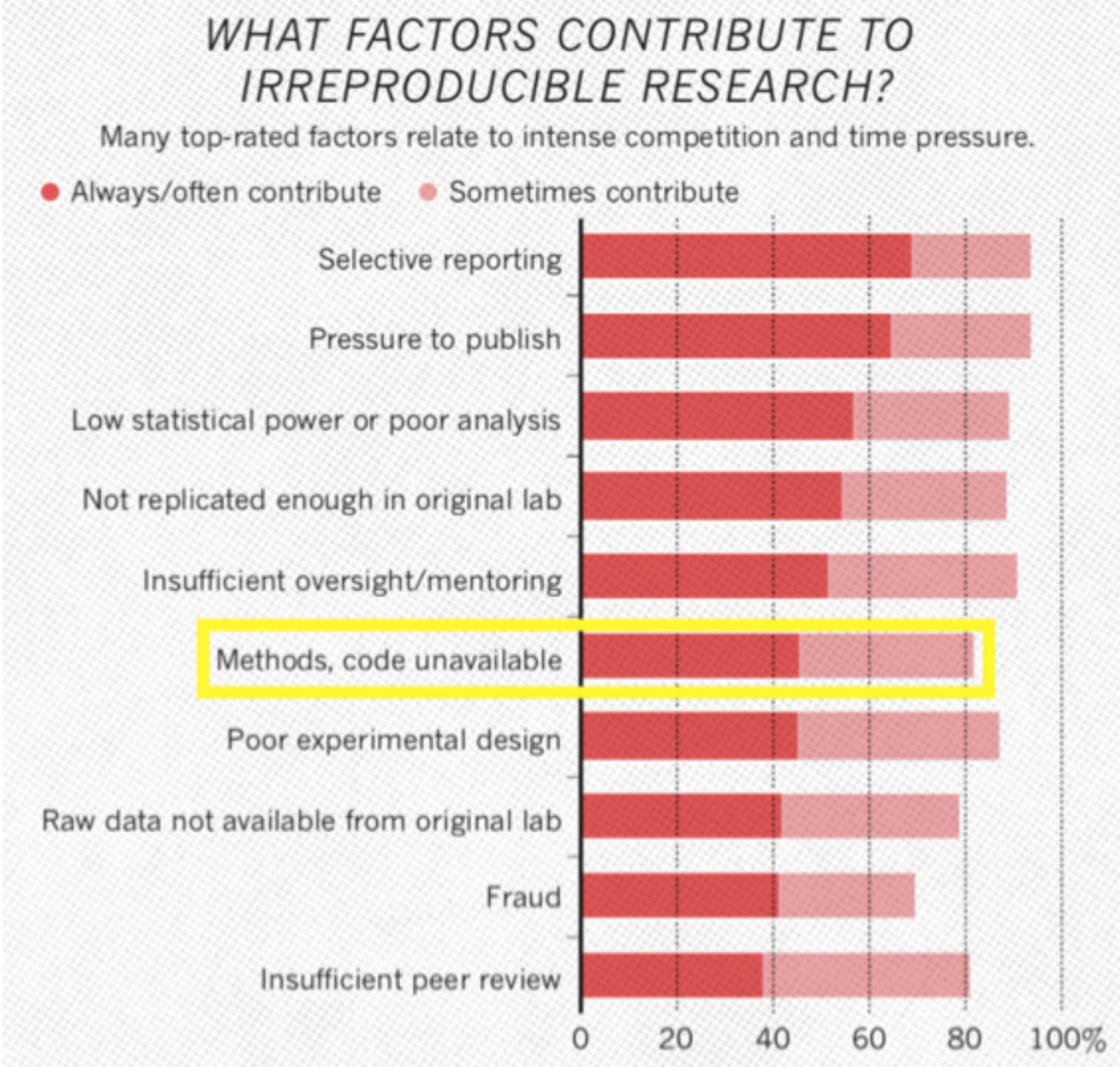
Software should not limit science.


7%
Don't know
3%
No, there is no crisis
IS THERE A
REPRODUCIBILITY
CRISIS?
A Nature survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.
BY MONYA BAKER
52%
Yes, a significant crisis
38%
Yes, a slight crisis
1,576
RESEARCHERS SURVEYED

HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.




WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?
Many top-rated factors relate to intense competition and time pressure.
Always/often contribute
Sometimes contribute
Selective reporting
Pressure to publish
Low statistical power or poor analysis
Not replicated enough in original lab
Insufficient oversight/mentoring
Methods, code unavailable
Poor experimental design
Raw data not available from original lab
Fraud
Insufficient peer review
0
20
40
60
80
100%


BLAZING
SADDLES

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

Comment: If these data could talk

Thomas Pasquier[1], Matthew K. Lau[2], Ana Trisovic[3,4], Emery Boose[2], Ben Couturier[3], Mercè Crosas[5], Aaron M. Ellison[2], Valerie Gibson[4], Chris Jones[4] & Margo Seltzer[1]

Received: 12 April 2017
Accepted: 24 July 2017
Published: xx xxx 2017

In the last few decades, data-driven methods have come to dominate many fields of scientific inquiry. Open data and open-source software have enabled the rapid implementation of novel methods to manage and analyze the growing flood of data. However, it has become apparent that many scientific fields exhibit distressingly low rates of repeatability and reproducibility. Although there are many dimensions to this issue, we believe that there is a lack of formalism used when describing end-to-end published results, from the data source to the analysis to the final published results. Even when authors do their best to make their research and data accessible, this lack of formalism reduces the clarity and efficiency of reporting, which contributes to issues of reproducibility. Data provenance aids both repeatability and reproducibility through *systematic* and *formal* records of the relationships among data sources, processes, datasets, publications and researchers.


The
Dataverse®
Project

Sharing and Preserving Computational Analyses for Posterity with *encapsulator*

Thomas Pasquier
University of Cambridge

Matthew K. Lau and Xueyuan Han
Harvard University

Elizabeth Fong and Barbara S. Lerner
Mount Holyoke College

Emery R. Boose, Mercè Crosas, Aaron M. Ellison, and Margo Seltzer
Harvard University

Editors: Lorena A. Barba, labarba@gwu.edu; George K. Thiruvathukal, gkt@cs.luc.edu

Reproducibility has become a recurring topic of discussion in many scientific disciplines.[1] Although it might be expected that some studies will be difficult to reproduce, recent conversations highlight important aspects of the scientific endeavor that could be improved to facilitate reproducibility. Open data and open source software are two important parts of a concerted effort to achieve reproducibility.[2] However, multiple publications point out these approaches' shortcomings,[3,4] such as the identification of dependencies, poor documentation of the installation processes, "code rot," failure to capture dynamic inputs, and technical barriers.

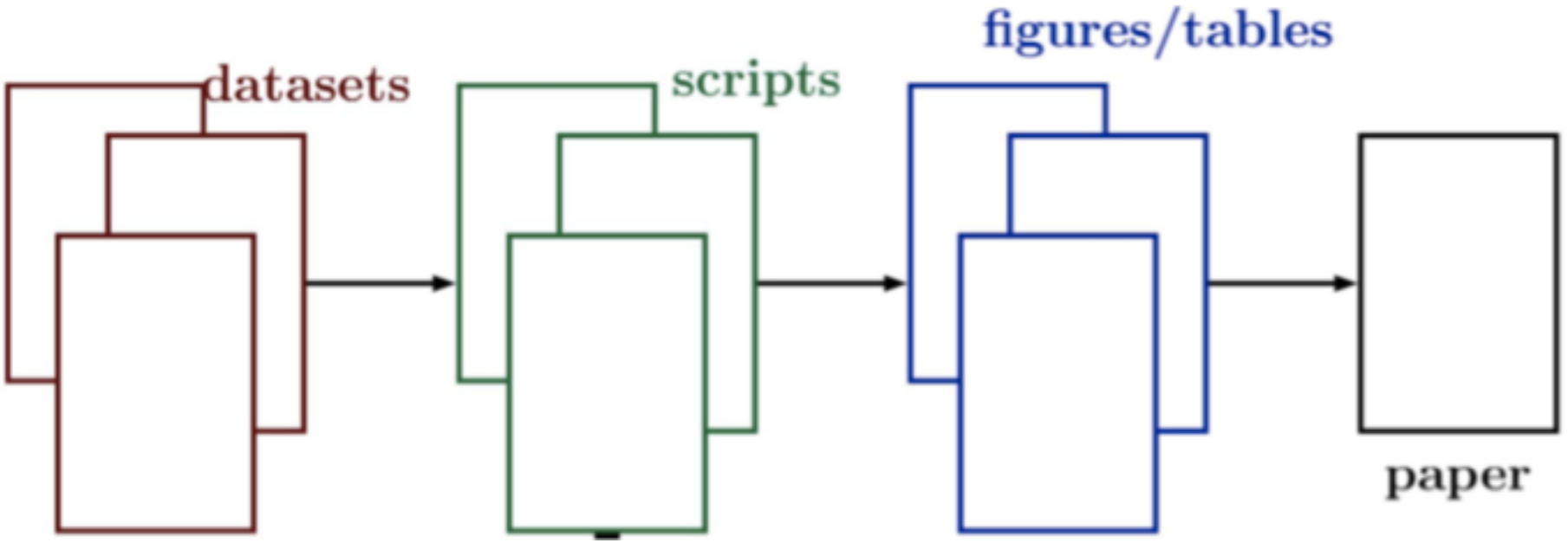
In prior work,[5] we pointed out that open data and open source software alone are insufficient to ensure reproducibility, as they do not capture information about the computational execution, that is, the "process" and context that produced the results using the data and code. In keeping with the "open" culture, we defined open process as the practice of both sharing the source and the input data and providing a description of the entire computational

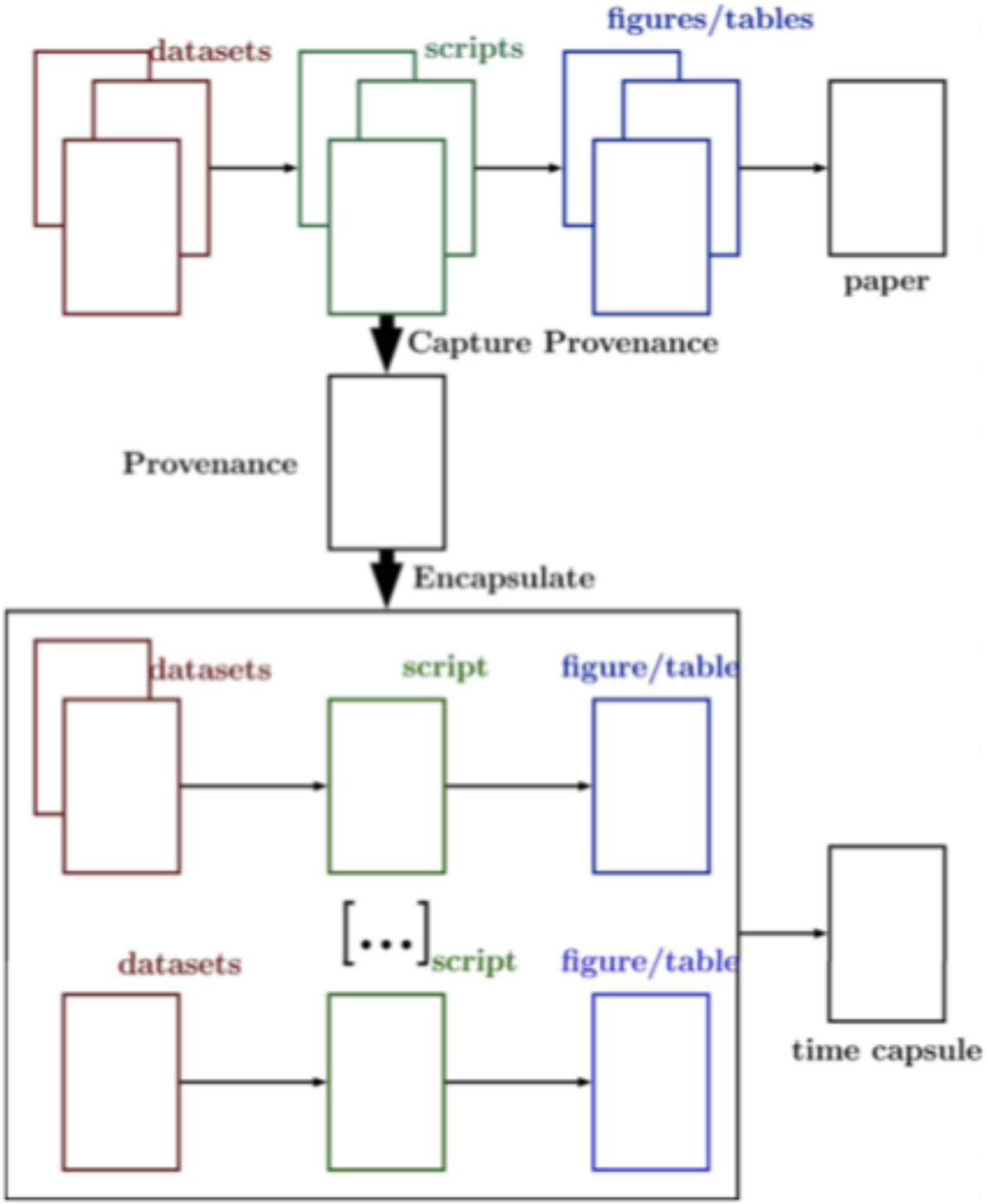
IEEE: Computing in Science & Engineering 2018

Encapsulator

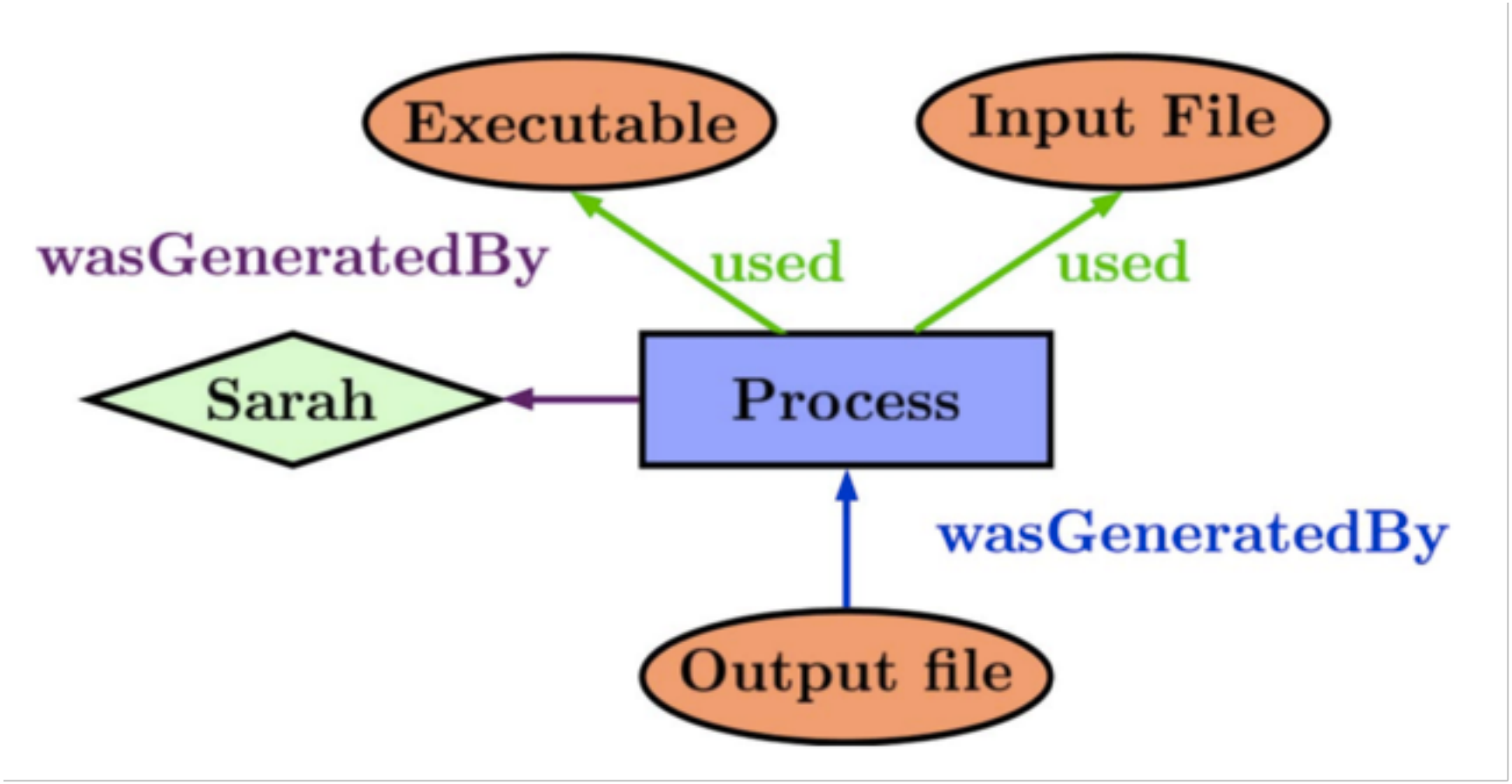
Purpose: Simplify computational reproducibility

1. Create a data “capsule” (code + data + environment)
2. Increase transparency with “cleaned” code

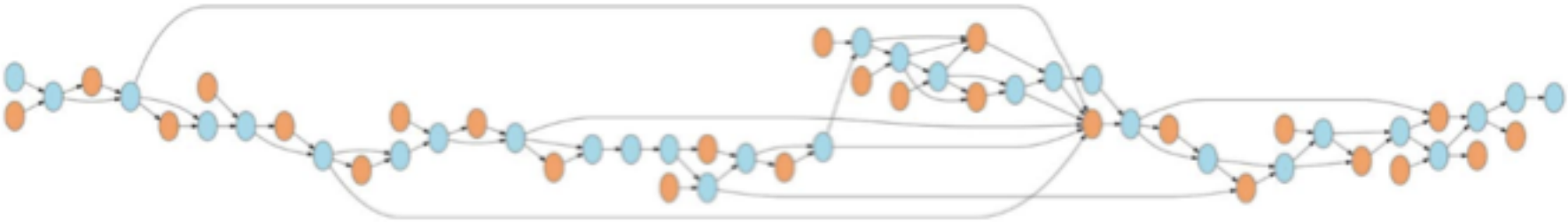

datasets
scripts
figures/tables
paper

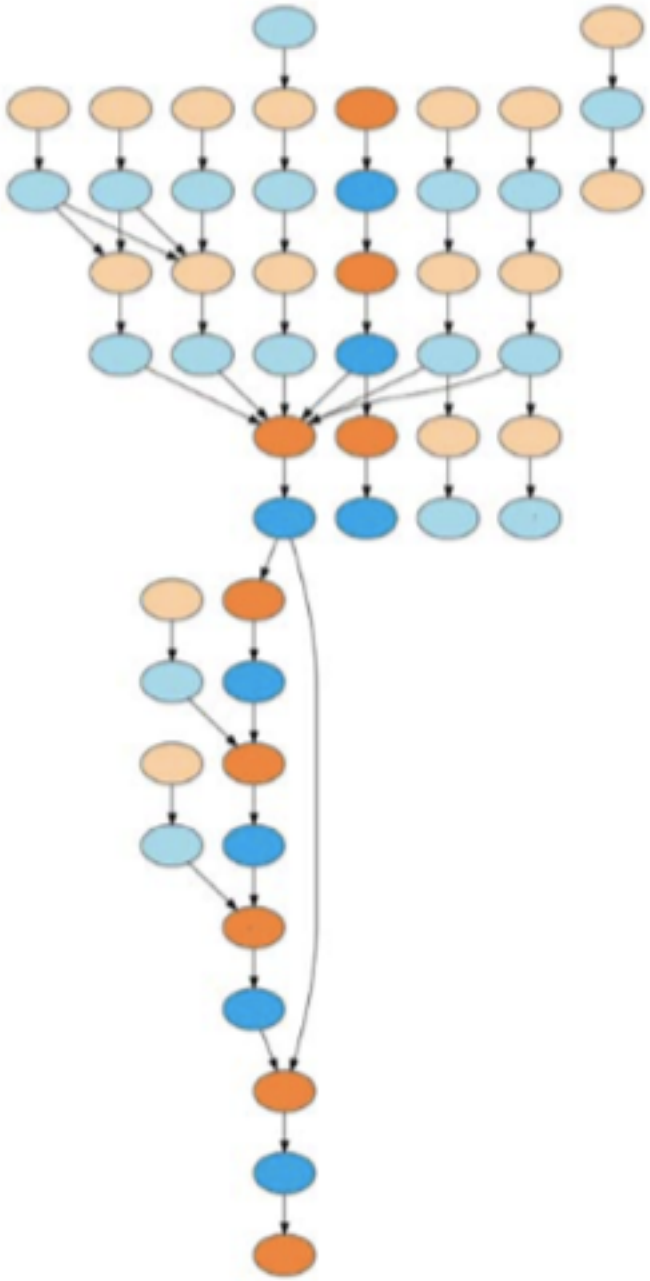

datasets
scripts
figures/tables
paper
Capture Provenance
Provenance
Encapsulate
datasets
script
figure/table
[...]
datasets
script
figure/table
time capsule

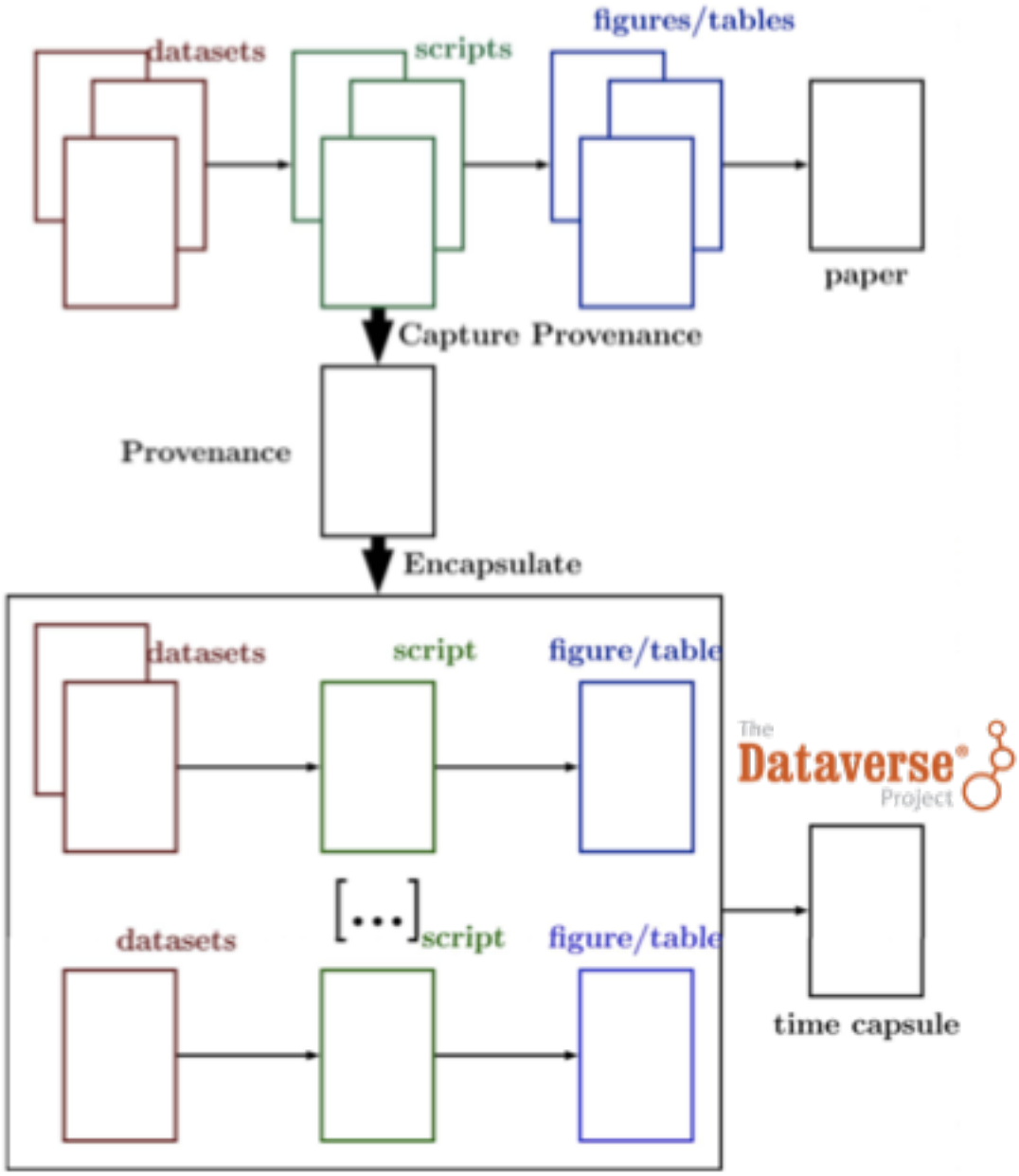
What is data provenance?



Prov. Huh. What is it good for?






figures/tables
datasets
scripts
paper
Capture Provenance
Provenance
Encapsulate
datasets
script
figure/table
[...]
datasets
script
figure/table
The Dataverse Project
time capsule

Future Work

- **Challenges**: domain-specific environments, improved code cleaning (*Rclean*), Capsule OS (*Linux*), nondeterminism
- Testing Dataverse products
- Integration with IDEs (*RStudio*)
- Container Support (e.g. *Docker*)

Email: *matthewklau@fas.harvard.edu*

Github: MKLau

This work was supported by NSF grants SSI-1450277 (End-to-End Provenance) and ACI-1448123 (Citation++).

More details about those projects are available at https://projects.iq.harvard.edu/provenance-at-harvard.