What's a finding?

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

Social Science Foo Camp
February 3, 2018
Massive Scale in Data and Computation

Passive data collection, massive compute infrastructure providing ever more research opportunities

• How does discovery change with (data/compute) scale?

• Does this impact reliability of findings?

Can we learn something about these questions by studying the scholarly record?
A Story from the Biomedical Sciences

• Elsevier publishes ~1,000 medical journals with ~1 million articles a year, mostly clinical findings

• Typically single-center studies with a small number of patients (e.g. n = 20)

• **Meta Analysis**: aggregate across many studies

Meta-analysis of the association between TP53 status and the risk of death at 2 years

Kyzas et al., “Selective Reporting Biases in Cancer Prognostic Factor Studies,” *JNCI*, 97(14), 2005
What Does Meta-Analysis Tell Us?

- Most published findings do not replicate
- Most published effects are inflated
- Incorrect findings have more impact than true ones, e.g. negative results

Suggests an important approach: *study the scholarly record as a body of evidence*
“Candidate Gene” Research (RIP)

• Late 1990’s: microarray and sequencing technology provided gene expression data for statistical analysis

• Goal was to find “candidate genes” that were related to a phenomena of interest:
  - small n studies
  - risk factors chosen from “diverse considerations”
  - use of conventional statistical tests and thresholding ($p < 0.05$)
  - studies subject to confounding and selective reporting

• Entirely replaced by Genome-Wide Association Studies (GWAS)
Recall: False Positives and False Negatives

Statistical Inference

True Underlying Relationship

- +
- -

+ -

- +

- -

N₁

N₂

True Positive
False Positive
False Negative
True Negative
Efforts to Replicate “Candidate Gene” Association Studies Fail

• “at least 20 false-positive findings for every one true-positive result”
• “approximately 1000 early gene loci-phenotype associations for the conditions listed in were false positives from the candidate-gene approach.”
• “There are no documented false-negative results arising from candidate-gene studies. Therefore, for the phenotypes listed in Table 1, the numerator of the FP:FN ratio is over 1000, while the denominator is apparently 0”

Ioannidis et al. “The false-positive to false-negative ratio in epidemiologic studies,” Epidemiology, 22(4), 2011
Studying the Scholarly Record

We now see the scholarly record as a body of numerical data, and we find:

- False Positives can overwhelm fields
- Entire fields are systemically failing
- Publications unstructured for analysis

Why?

- Overuse of underpowered studies
- Editorial preference for positive results
- Researcher degrees of freedom

“Only when the tide goes out do you discover who’s been swimming naked.”

--Warren Buffett
REPRODUCIBILITY

Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

By Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P.A. Ioannidis, Michela Taufer

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency, the lack of metadata, and the difficulty in verifying results. To address these challenges, we offer a framework for understanding how computational results were derived and to reconciling any differences that might arise between independent replications. We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard, which includes workflow information that explains what raw data and intermediate results are input to which computations. Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include the computational steps taken to process data and generate findings.

Access to the computational steps taken to process data and generate findings is as important as access to data themselves.


Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://...
Reproducibility Enhancement Principles

1: To facilitate reproducibility, share the data, software, workflows, and details of the computational environment in open repositories.

2: To enable discoverability, persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

3: To enable credit for shared digital scholarly objects, citation should be standard practice.

4: To facilitate reuse, adequately document digital scholarly artifacts.

5: Journals should conduct a Reproducibility Check as part of the publication process and enact the TOP Standards at level 2 or 3.

6: Use Open Licensing when publishing digital scholarly objects.

7: Funding agencies should instigate new research programs and pilot studies.
Does this Generalize?

- Carp evaluated methods and reporting for 241 recent fMRI articles.

- Many studies did not report critical methodological details with regard to experimental design, data acquisition, and analysis.

- Many studies were underpowered to detect any but the largest statistical effects.

- Data collection and analysis methods were highly flexible across studies, with nearly as many unique analysis pipelines as there were studies in the sample.

- Since the rate of false positive results is thought to increase with the flexibility of experimental design, functional neuroimaging may be especially susceptible to false positives.

Carp, “The secret lives of experiments: methods reporting in the fMRI literature,” Neuroimage, 63(1) 2012