Reproducible Research: Concepts and Ideas

Reproducible Research

Roger D. Peng, Associate Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

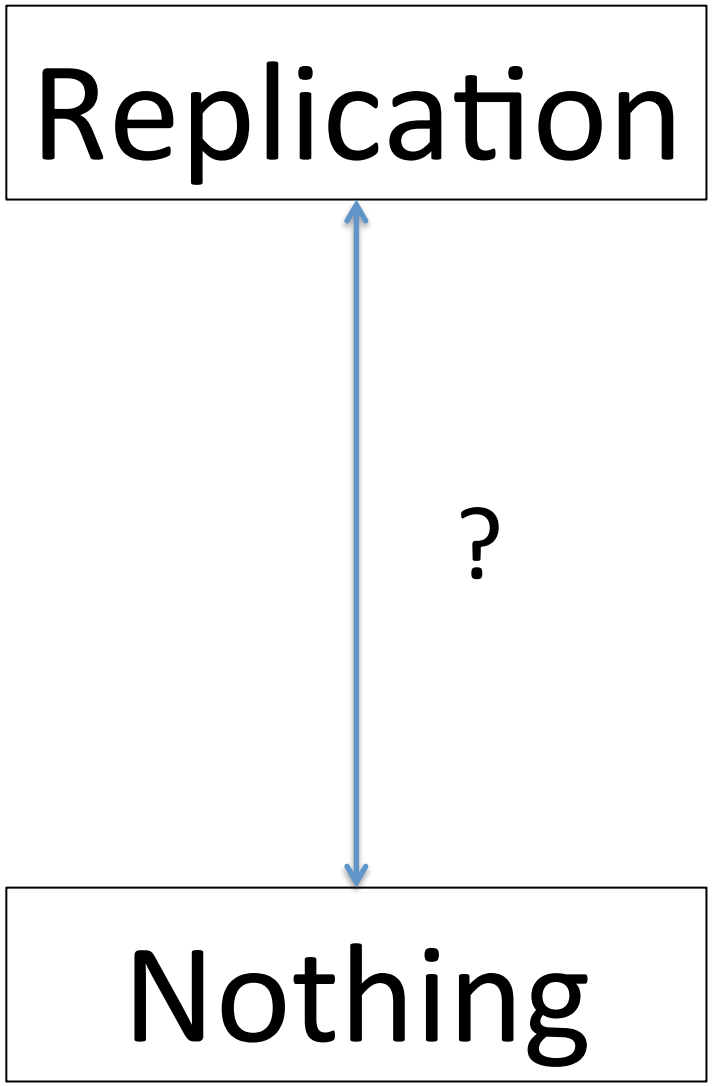
Replication

- The ultimate standard for strengthening scientific evidence is replication of findings and conducting studies with independent
  - Investigators
  - Data
  - Analytical methods
  - Laboratories
  - Instruments
- Replication is particularly important in studies that can impact broad policy or regulatory decisions

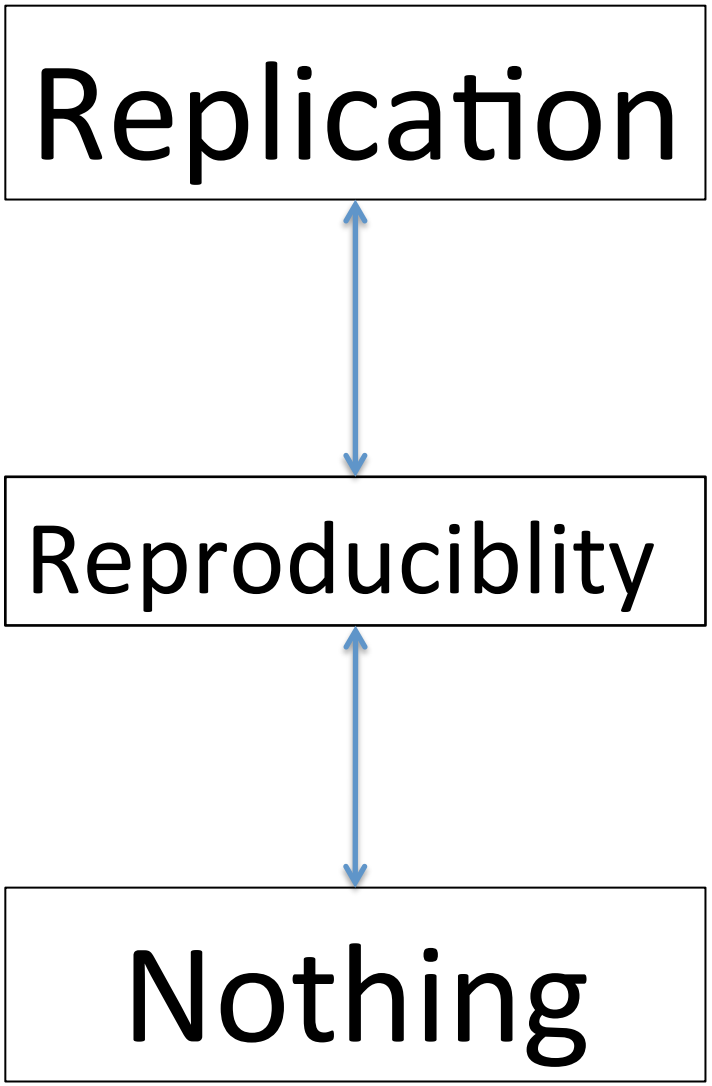
What's Wrong with Replication?

- Some studies cannot be replicated
  - No time, opportunistic
  - No money
  - Unique
- Reproducible Research: Make analytic data and code available so that others may reproduce findings

How Can We Bridge the Gap?



How Can We Bridge the Gap?



Why Do We Need Reproducible Research?

- New technologies increasing data collection throughput; data are more complex and extremely high dimensional
- Existing databases can be merged into new "megadatabases"
- Computing power is greatly increased, allowing more sophisticated analyses
- For every field "X" there is a field "Computational X"

Example: Reproducible Air Pollution and Health Research

- Estimating small (but important) health effects in the presence of much stronger signals
- Results inform substantial policy decisions, affect many stakeholders
  - EPA regulations can cost billions of dollars
- Complex statistical methods are needed and subjected to intense scrutiny

Internet-based Health and Air Pollution Surveillance System (iHAPSS)

iHAPSS
Internet-based Health & Air Pollution Surveillance System

ABOUT iHAPSS

iHAPSS is an internet system for monitoring the effects of air pollution on mortality and morbidity in the United States.

iHAPSS is funded by the Health Effects Institute (HEI). It provides published material, software and data to monitor the association between air pollution and mortality and morbidity.

iHAPSS is developed and maintained by the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health.

PUBLICATIONS

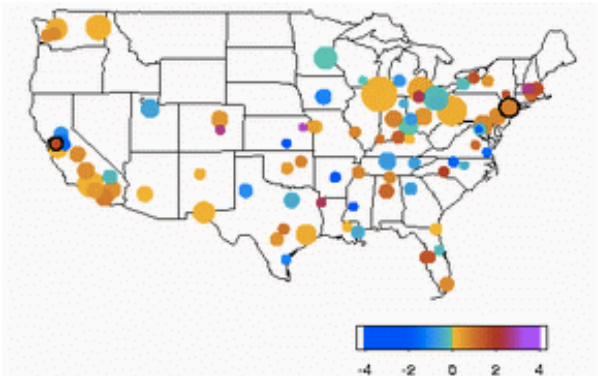
Current and previous publications and reports.

SOFTWARE

Tools for data analysis.

DATA

Air pollution and meteorological data for 108 U.S. cities 1987–2000.

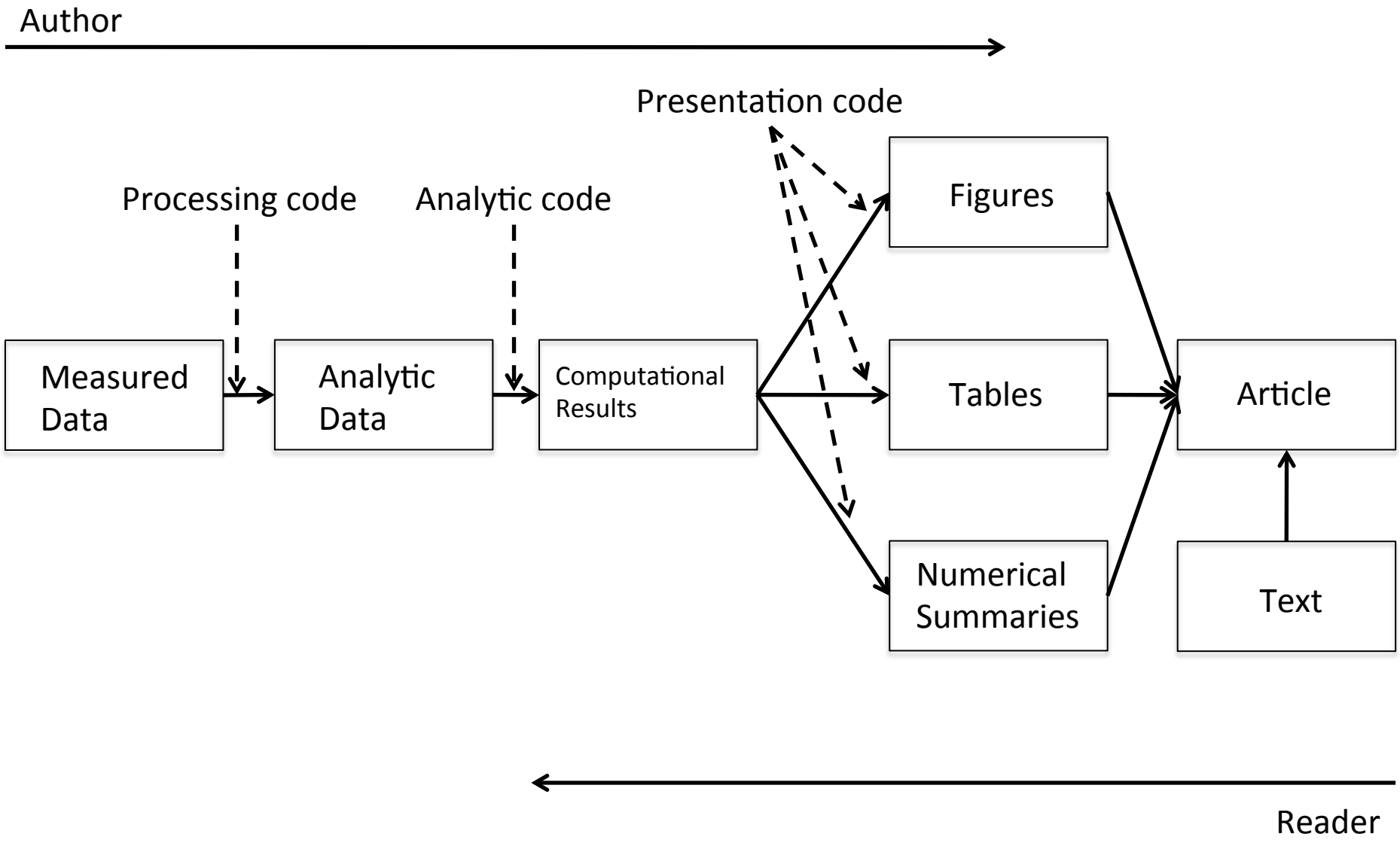


http://www.ihapss.jhsph.edu

Research Pipeline

Reader

Research Pipeline



Recent Developments in Reproducible Research

Data Replication & Reproducibility

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

Recent Developments in Reproducible Research

The Duke Saga

Recent Developments in Reproducible Research

REPORT BRIEF MARCH 2012

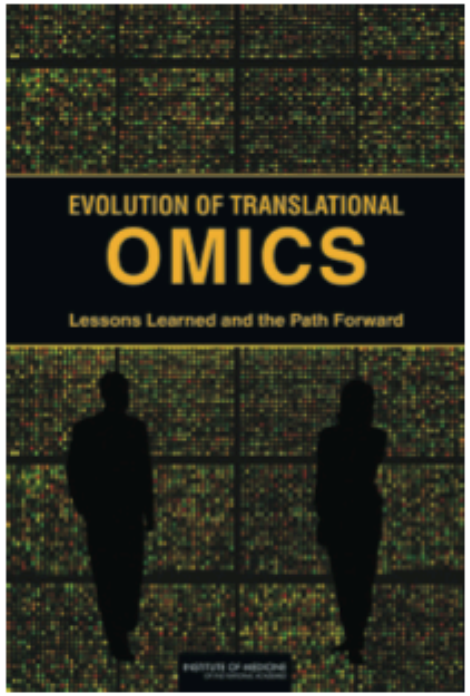
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

Advising the nation • Improving health

For more information visit www.iom.edu/translationalomics

Evolution of Translational Omics

Lessons Learned and the Path Forward



The IOM Report

In the Discovery/Test Validation stage of omics-based tests:

- **Data/metadata** used to develop test should be made publicly available
- The **computer code** and fully specified computational procedures used for development of the candidate omics-based test should be made sustainably available
- "Ideally, the computer code that is released will **encompass all of the steps of computational analysis**, including all data preprocessing steps, that have been described in this chapter. All aspects of the analysis need to be transparently reported."

What do We Need?

- Analytic data are available
- Analytic code are available
- Documentation of code and data
- Standard means of distribution

Who are the Players?

- Authors
  - Want to make their research reproducible
  - Want tools for RR to make their lives easier (or at least not much harder)
- Readers
  - Want to reproduce (and perhaps expand upon) interesting findings
  - Want tools for RR to make their lives easier

Challenges

- Authors must undertake considerable effort to put data/results on the web (may not have resources like a web server)
- Readers must download data/results individually and piece together which data go with which code sections, etc.
- Readers may not have the same resources as authors
- Few tools to help authors/readers (although toolbox is growing!)

In Reality…

- Authors
  - Just put stuff on the web
  - (Infamous) Journal supplementary materials
  - There are some central databases for various fields (e.g. biology, ICPSR)
- Readers
  - Just download the data and (try to) figure it out
  - Piece together the software and run it

Literate (Statistical) Programming

- An article is a stream of **text** and **code**
- Analysis code is divided into text and code "chunks"
- Each code chunk loads data and computes results
- Presentation code formats results (tables, figures, etc.)
- Article text explains what is going on
- Literate programs can be **weaved** to produce human-readable documents and **tangled** to produce machine-readable documents

Literate (Statistical) Programming

- Literate programming is a general concept that requires
  1. A documentation language (human readable)
  2. A programming language (machine readable)
- Sweave uses LaTeX and R as the documentation and programming languages
- Sweave was developed by Friedrich Leisch (member of the R Core) and is maintained by R core
- Main web site: `http://www.statistik.lmu.de/~leisch/Sweave`

Sweave Limitations

- Sweave has many limitations
- Focused primarily on LaTeX, a difficult to learn markup language used only by weirdos
- Lacks features like caching, multiple plots per chunk, mixing programming languages and many other technical items
- Not frequently updated or very actively developed

Literate (Statistical) Programming

- knitr is an alternative (more recent) package
- Brings together many features added on to Sweave to address limitations
- knitr uses R as the programming language (although others are allowed) and variety of documentation languages
  - LaTeX, Markdown, HTML
- knitr was developed by Yihui Xie (while a graduate student in statistics at Iowa State)
- See http://yihui.name/knitr/

Summary

- Reproducible research is important as a **minimum standard**, particularly for studies that are difficult to replicate
- Infrastructure is needed for **creating** and **distributing** reproducible documents, beyond what is currently available
- There is a growing number of tools for creating reproducible documents