JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Introduction to Reproducible Research

Roger D. Peng
@rdpeng, @simplystats, simplystatistics.org

UW SISBID
July 2016

How Do You Know if a Data Analysis is Successful?

Parable

ARTICLES

nature medicine

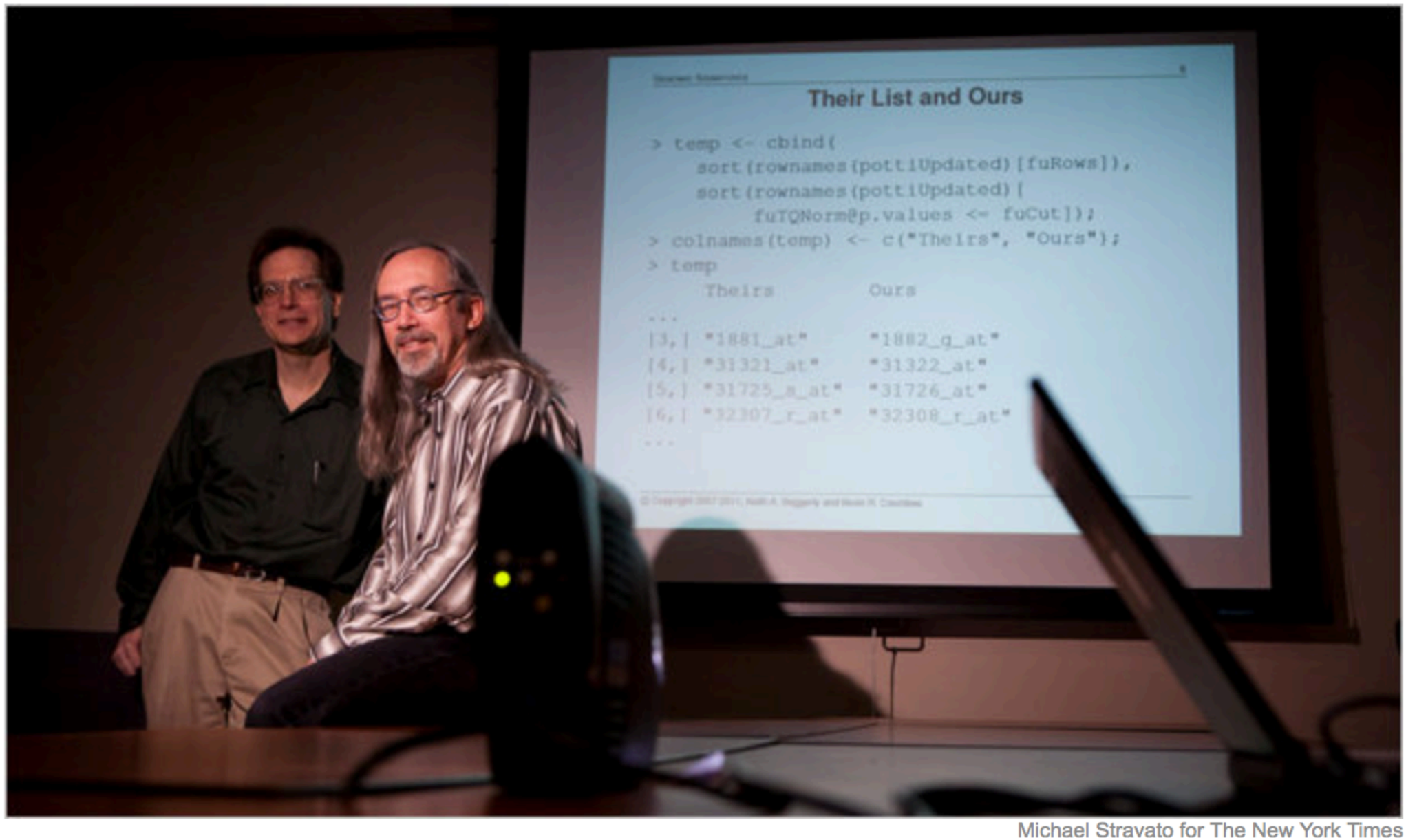
Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

Deception at Duke

“Rock Star” Statisticians

How Bright Promise in Cancer Testing Fell Apart



Michael Stravato for The New York Times

New York Times

“Deception” at MDACC

Roger D. Peng <rpeng@jhsph.edu>
to Keith

Keith, I just had a chance to watch the 60 Minutes segment. Congratulations! I thought it was a very well-done piece. I'm still marveling at how clean your desk is!

Best,
-roger

...

February 2012

Follow-up Discussion

Steve Goodman sgoodman@jhmi.edu via googlegroups.com Unsubscribe 2/15/12

to reproducible-r.

BTW, I felt that Keith and Kevin's 45 seconds was akin to listening to "Ride of the Valkyries"in a TV commercial instead of hearing the whole of Die Walkure. There ain't nothin' better than the full Die Baggerly, as long as Keith is singing!

Lessons?

Clinical medicine (tightly controlled)

+

Genomics (wild west)

=

Wild West

Institute of Medicine Committee

REPORT BRIEF MARCH 2012

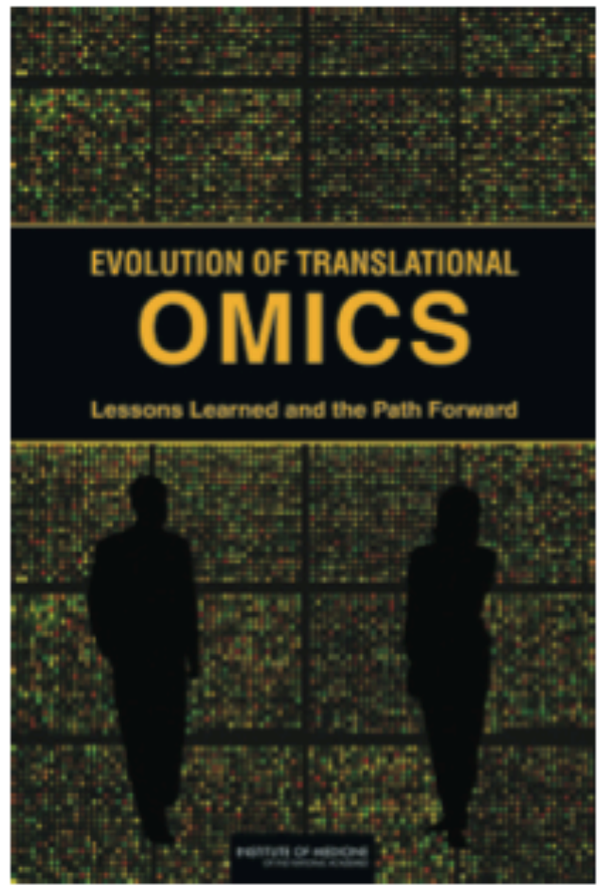
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

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

- **Data/metadata** used to develop test should be made publicly available

- The **computer code** and fully specified computational procedures used or development of the omics-based test should be made available

- Ideally, the computer code that is released will **encompass all of the steps** of computational analysis, including all data preprocessing steps

Replication and Reproducibility

- **Replication**
  - Focuses on the validity of the *scientific claim*
  - "Is this claim true?"
  - Ultimate standard for scientific evidence
  - New investigators, data, analytic methods, labs, instruments, etc.
  - Important in studies that can impact policy or regulation

- **Reproducibility**
  - Focuses on the validity of the *data analysis*
  - "Can we trust this analysis?"
  - A minimum standard
  - New investigators, same data, same methods
  - Important when replication is impossible

What's Wrong with Replication?

- Nothing, but…

- 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

Upon Seeing Your Work…

Information Required

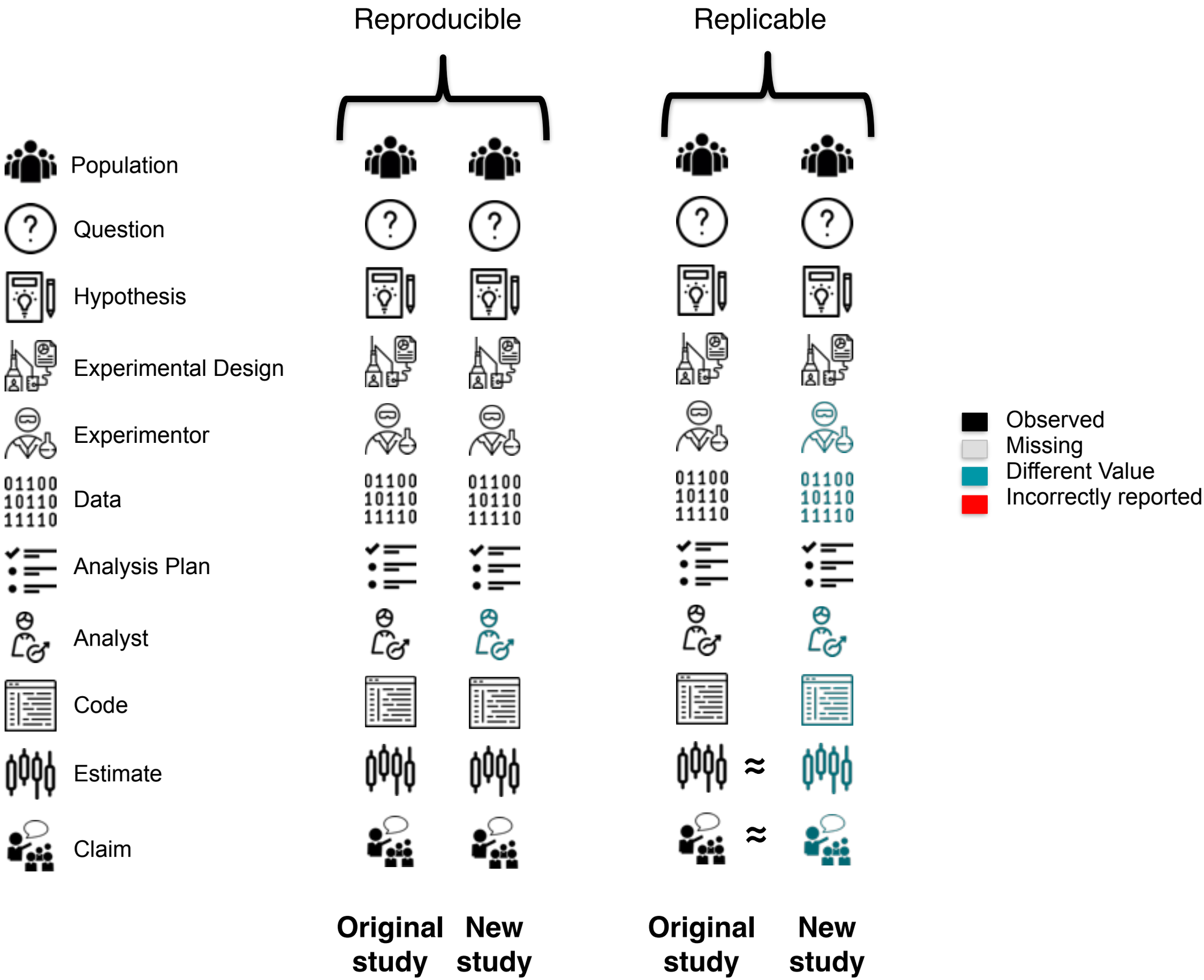
Minimum ⟷ Maximum

Do Nothing

Full Replication

Reproducibility

Most common activity


Reproducible
Replicable
Population
Question
Hypothesis
Experimental Design
Experimentor
Data
Analysis Plan
Analyst
Code
Estimate
Claim
Observed
Missing
Different Value
Incorrectly reported
≈
≈
Original study
New study
Original study
New study

Why Do We Need Reproducible Research?

- New technologies increasing data collection throughput

- Data are more complex and high dimensional

- Existing databases can be merged into new and bigger databases

- Computing power is greatly increased, allowing more sophisticated/complicated analyses

- For every field “X” there is a field “Computational X”

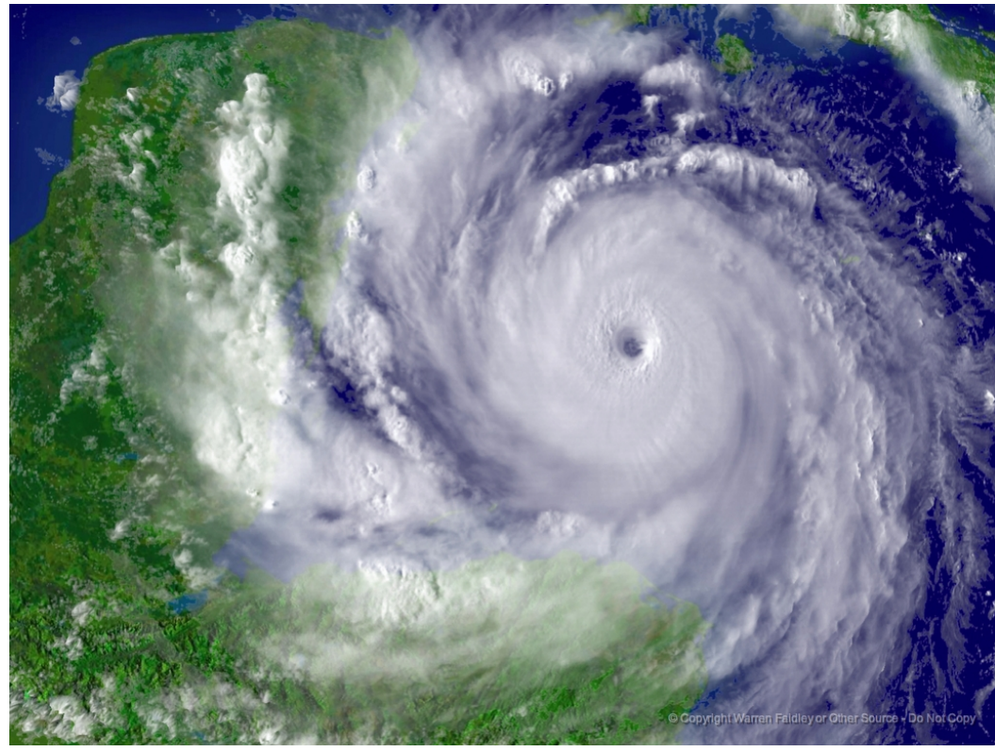
Air Pollution and Health: A Perfect Storm?

- Estimating small health effects in the presence of much stronger signals

- Results inform substantial policy decisions and affect many stakeholders

- EPA regulations can cost billions of dollars

- Complex statistical methods are needed and subjected to intense scrutiny

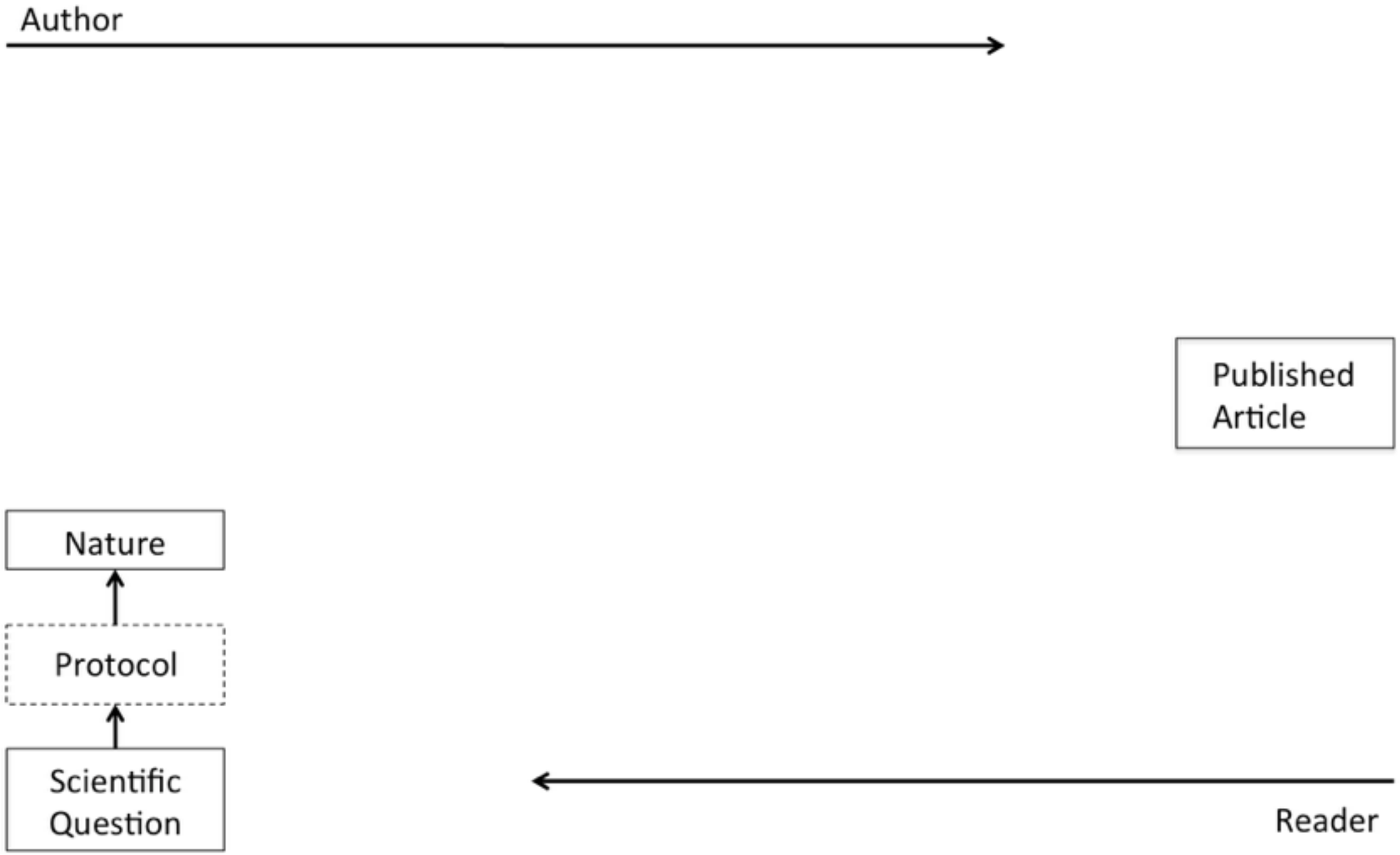


The End Result

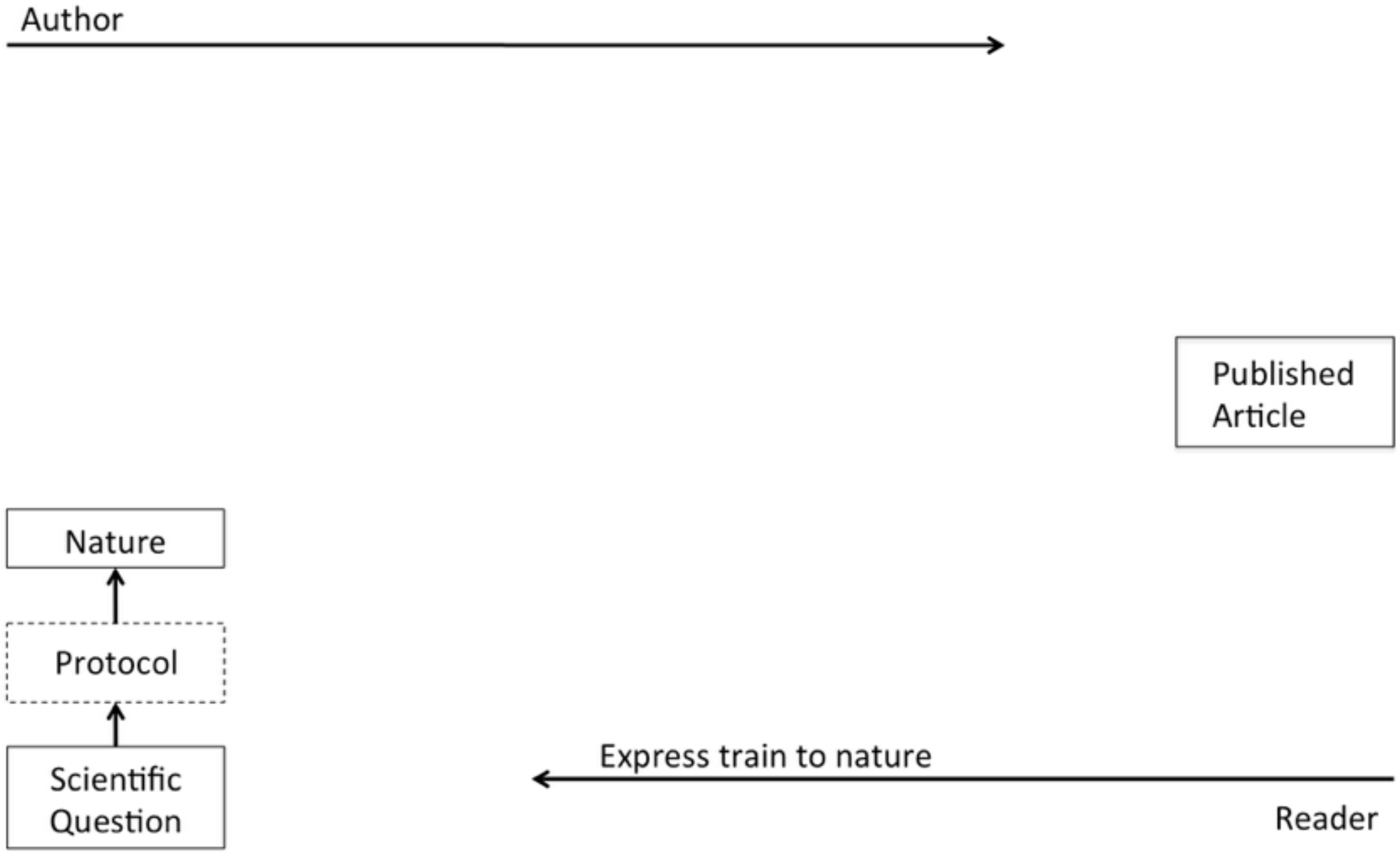
- Basic analyses can be difficult to describe
- Heavy computational requirements are thrust upon people without adequate training in statistics and computing
- Errors are more easily introduced into long and complex analysis pipelines
- Knowledge transfer is limited
- Complicated analyses cannot be trusted

What is Reproducible Research?

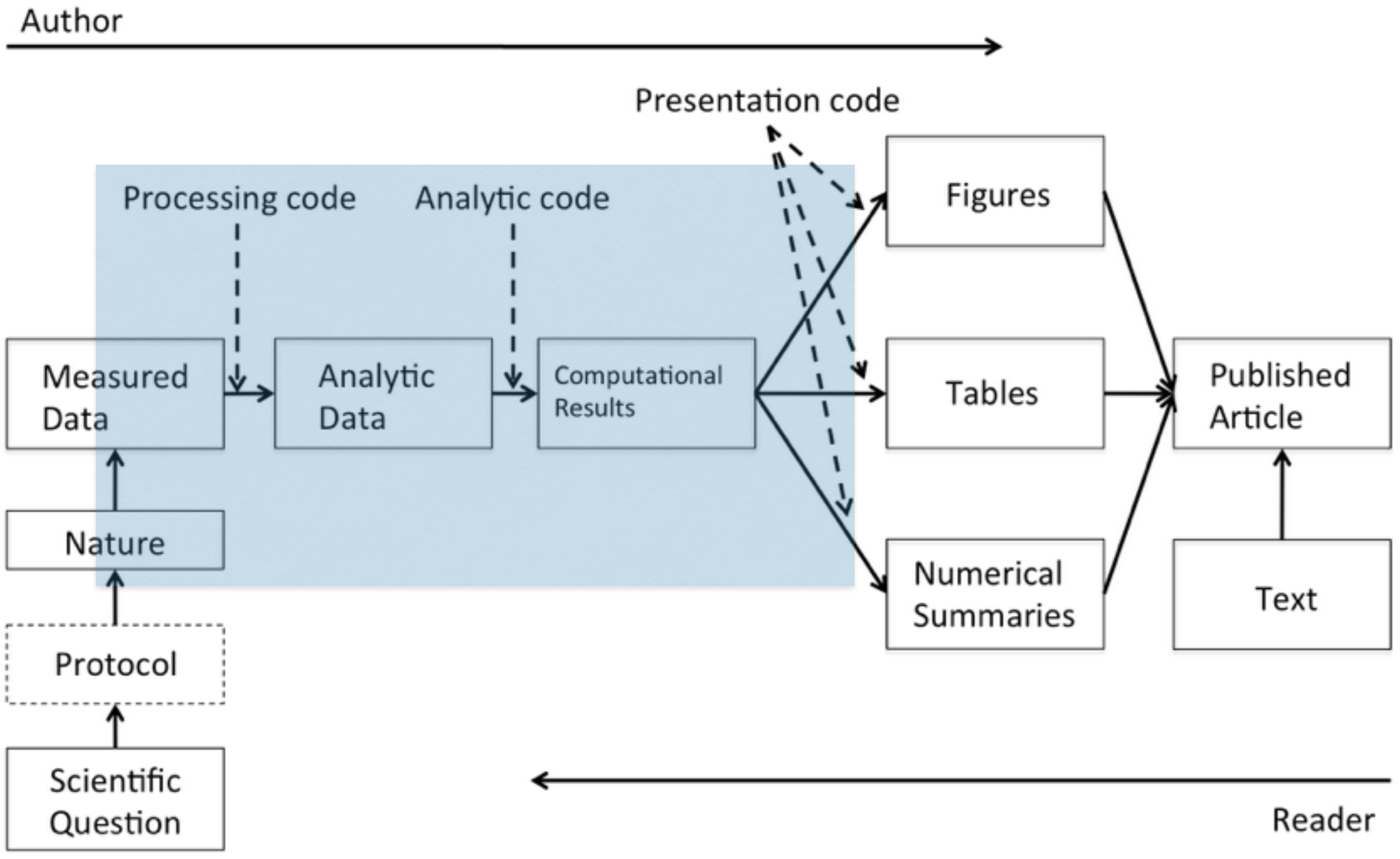
What is Reproducible Research?



What is Reproducible Research?



What is Reproducible Research?



What is Reproducible Research?

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

What is Reproducible Research?

- 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 and 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

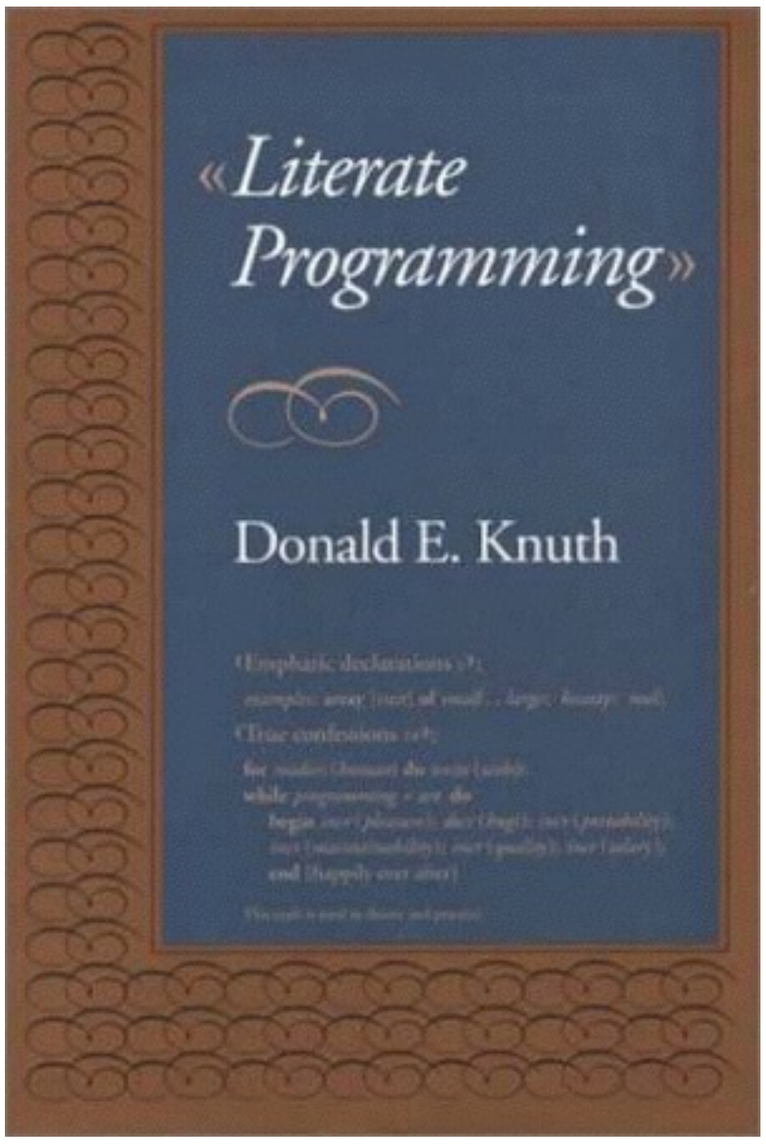
- Few tools to help authors/readers (although toolbox is growing!)

Recent Developments

- **Software**: iPython Notebooks, knitr, markdown, LONI, Galaxy
- **Repositories**: GitHub, NCBI, ICPSR, Dataverse
- **Policy**: *Science*, *Nature*, *PLOS ONE*, OSTP, NIH

Literate Statistical Programming

- An article/report 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
- See *Literate Programming* by Donald Knuth



Literate Statistical Programming

- Literate programming is a general concept that requires
  - A documentation language (human readable)
  - 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

Literate Statistical Programming

- knitr is package that brings together many features added on to Sweave to address limitations

- knitr uses R as the programming language knitr was developed by Yihui Xie (while a graduate student in statistics at Iowa State, now at RStudio)

- knitr uses the R programming language (although others are allowed) and variety of documentation languages

    - LaTeX, Markdown, HTML

- Built into RStudio pipeline

- See http://yihui.name/knitr/

What Problem Does Reproducibility Solve?

- What we get
  - Transparency / Improved knowledge transfer
    - Data availability
    - Software / Methods
- What we do NOT get
  - Validity / Correctness of the analysis

Computational Research has a Communication Problem


Thing
-a-
Week
Three
Jonathan Coulton

Intro:
E (p.m.)

Verse:
E (p.m.)
Code Monkey get up get coffee
E (p.m.)
Code Monkey go to job
E (p.m.)
Code Monkey have boring meeting
E (p.m.)
With boring manager Rob
E B/E A/E
Rob say Code Monkey very dilligent
E B/E A/E
But his output stink
E B/E A/E
His code not "functional" or "elegant"
E B/E A/E
What do Code Monkey think?

Pre-chorus:
N.C. A B7 G# C#m
Code Monkey think maybe manager wanna write god damned login page himself
B A
Code Monkey not say it out loud
B7
Code Monkey not crazy, just proud

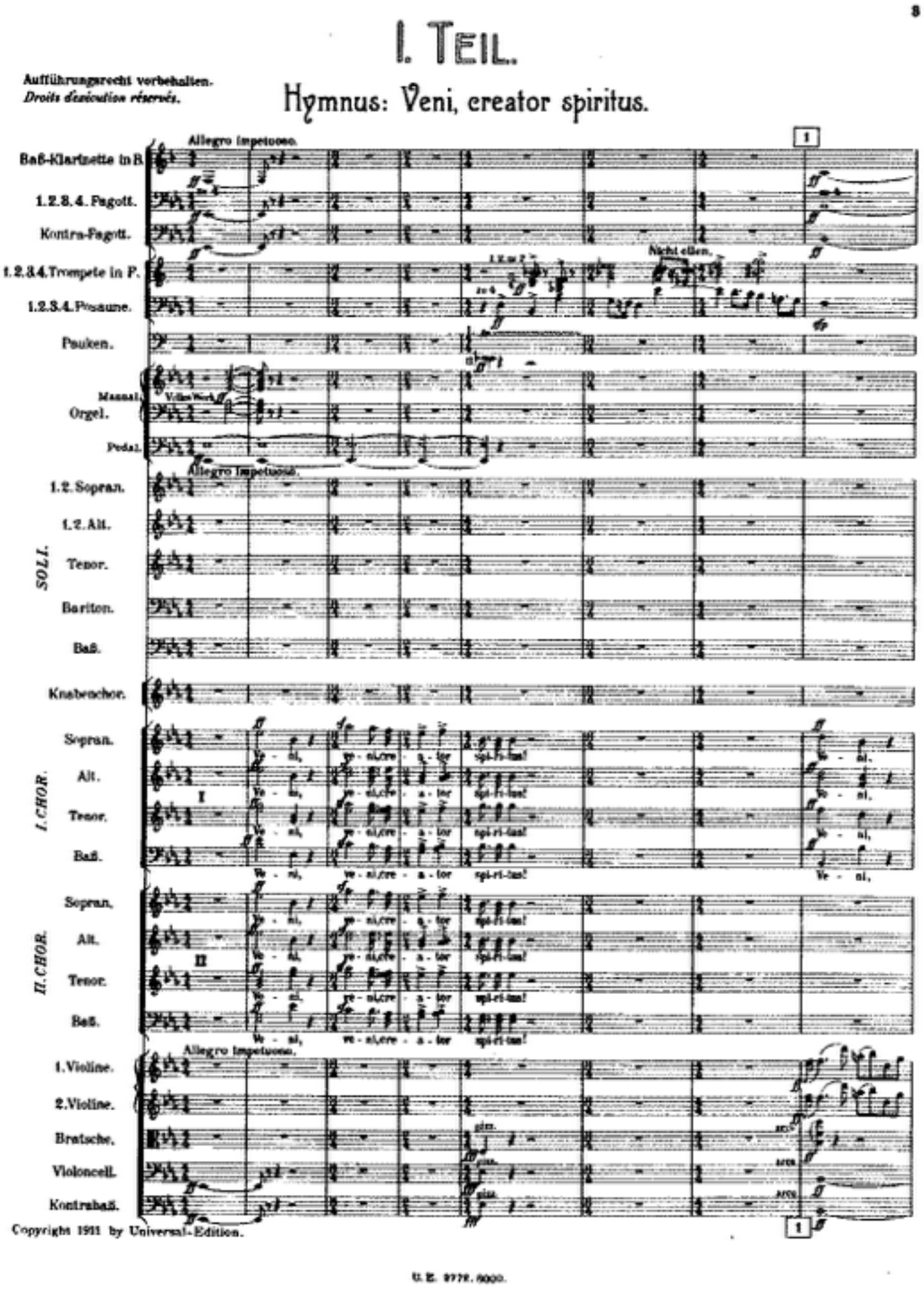
Chorus:
B7 E Emaj7
Code Monkey like Fritos
E6 E
Code Monkey like Tab and Mountain Dew
Amaj7sus2 Amaj7 Amaj7sus2
Code Monkey very simple man
Amaj7 B7
With big warm fuzzy secret heart:
B9
Code Monkey like you

```
Intro:
E (p.m.)

Verse:
E (p.m.)
Code Monkey get up get coffee
E (p.m.)
Code Monkey go to job
E (p.m.)
Code Monkey have boring meeting
E (p.m.)
With boring manager Rob
E    B/E       A/E
Rob say Code Monkey very dilligent
E        B/E     A/E
But his output stink
E    B/E       A/E
His code not "functional" or "elegant"
E             B/E              A/E
What do Code Monkey think?

Pre-chorus:
N.C.                A                B7
Code Monkey think maybe manager wanna writ
      B           A
Code Monkey not say it out loud
                  B7
Code Monkey not crazy, just proud

Chorus:
B7                  E       Emaj7
Code Monkey like Fritos
                    E6                      E
Code Monkey like Tab and Mountain Dew
                    Amaj7sus2  Amaj7 Amaj7sus
Code Monkey very simple man
          Amaj7       B7
With big warm fuzzy secret heart:
B9
Code Monkey like you
```



The Central Problem

Data Analysis = ???

What's Next?

- Reproducibility is critical for *communicating* a data analysis
- One cannot sufficiently describe an analysis in words
- General consensus about its importance
- Infrastructure for making all research reproducible is not there yet, but things are ever improving

How Do You Know if a Data Analysis is Successful?

- Reproducible

- Uses the best available statistical methods of analysis

“There ain’t nothin’ better than the full Die Baggerly, as long as Keith is singing!”

–Steve Goodman