

Introduction to Reproducible Research

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

Biostatistics 140.776

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 Chan⁴, Robyn Sayer⁴,
Janiel Cragun⁴, Hope Cottrill⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵,
Andrew Berchuck^{1,6}, Geoffrey S Ginsburg^{1,2}, Phillip Febbo¹⁻³, Johnathan Lancaster⁴ &
Joseph R Nevins¹⁻³

Deception at Duke



The image shows a screenshot of a 60 Minutes video player. At the top, a red banner features a stopwatch and the text "60 MINUTES". Below this is a navigation bar with links: HOME, UP NEXT, 60 OVERTIME, NEWSMAKERS, POLITICS, SCIENCE, BUSINESS, and ENTERTAINMENT. The main video frame displays a man in a suit standing next to a large open book. The left page of the book has the title "Deception At Duke" and the Duke University logo. The right page has the text "Produced By Kyra Darnton" and a paragraph of text. A "60 MINUTES" logo is in the bottom left corner of the video frame. Below the video frame is a player control bar with a progress bar, a pause button, and the text "0:52 / 13:46". To the right of the progress bar are buttons for "SHARE", a volume icon, and a full screen icon. Below the player control bar is a section for comments and sharing. It includes "23 Comments", a "Share this Video:" prompt, and buttons for "Recommend" (473), "Tweet" (49), and "Like" (363). At the bottom, the video title "Deception at Duke" is displayed, followed by the date and time "February 12, 2012 4:00 PM". Below the title is a short description: "Were some cancer patients at Duke University given experimental treatments based on fabricated data? Scott Pelley reports."

60 MINUTES

HOME UP NEXT 60 OVERTIME NEWSMAKERS POLITICS SCIENCE BUSINESS ENTERTAINMENT

Deception At Duke

Produced By Kyra Darnton

60 MINUTES

0:52 / 13:46

SHARE

23 Comments

Share this Video:

Recommend 473

Tweet 49

Like 363

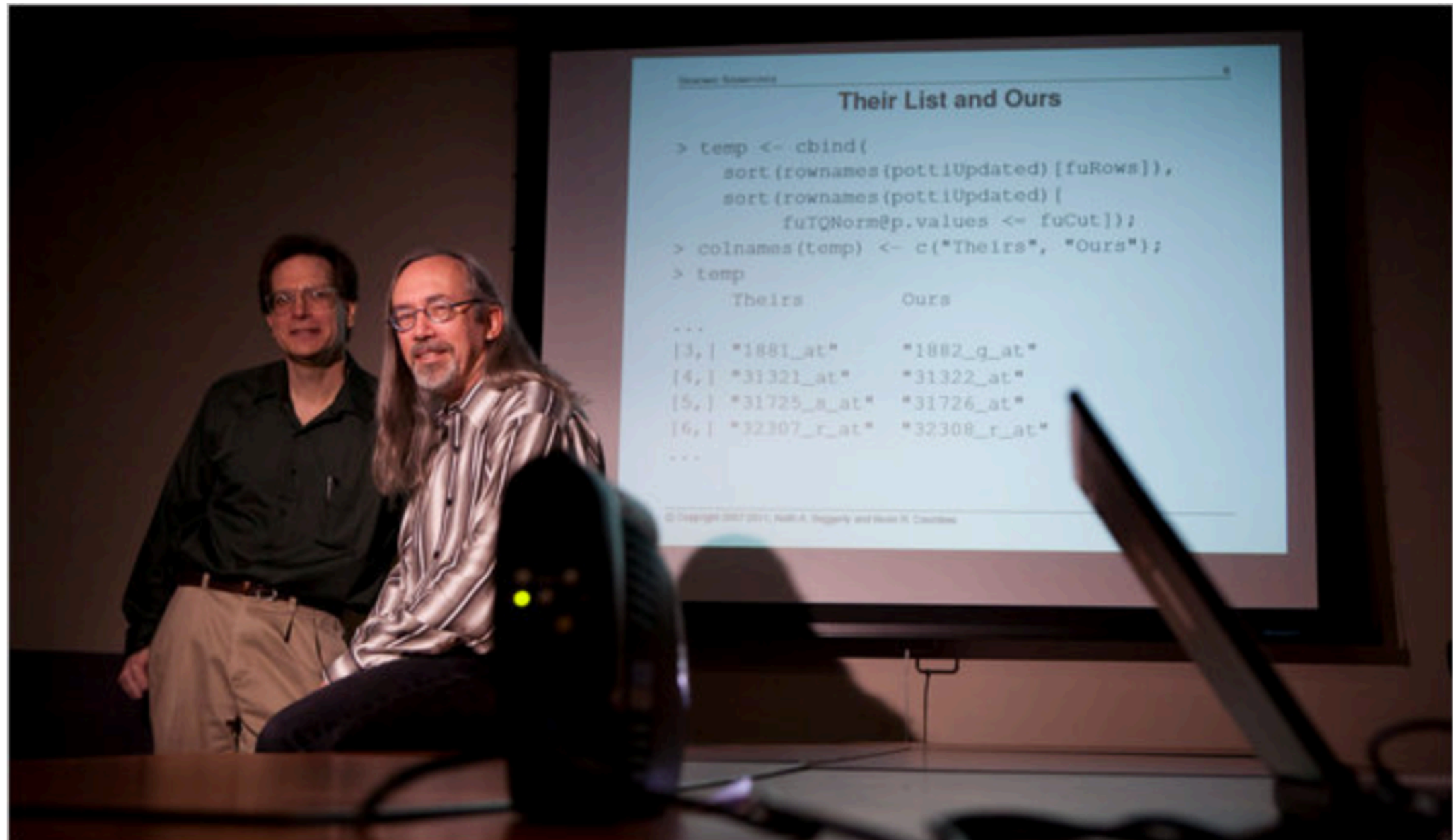
Deception at Duke

February 12, 2012 4:00 PM

Were some cancer patients at Duke University given experimental treatments based on fabricated data? Scott Pelley reports.

“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?



Institute of Medicine Committee

REPORT BRIEF  MARCH 2012

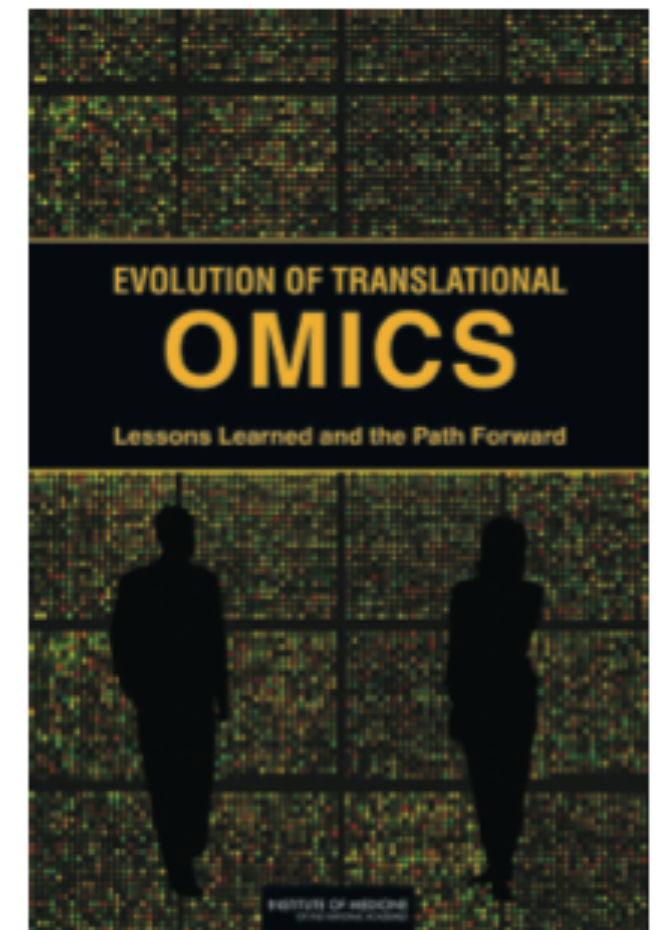
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

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

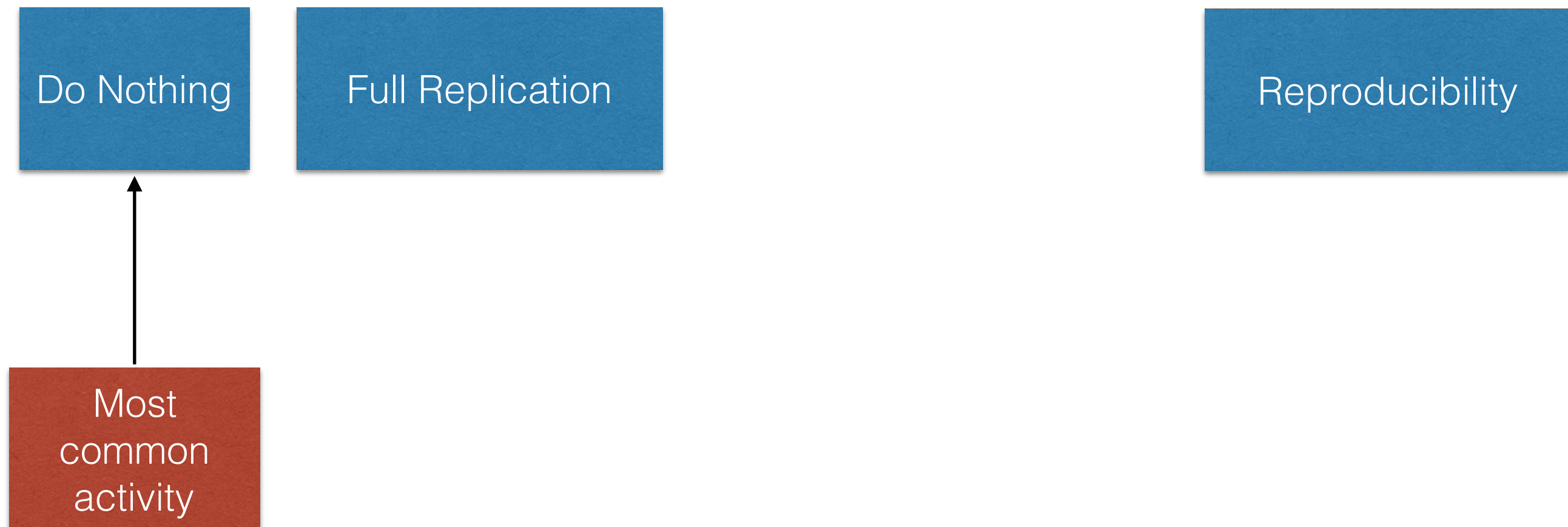
Minimum ←————→ Maximum

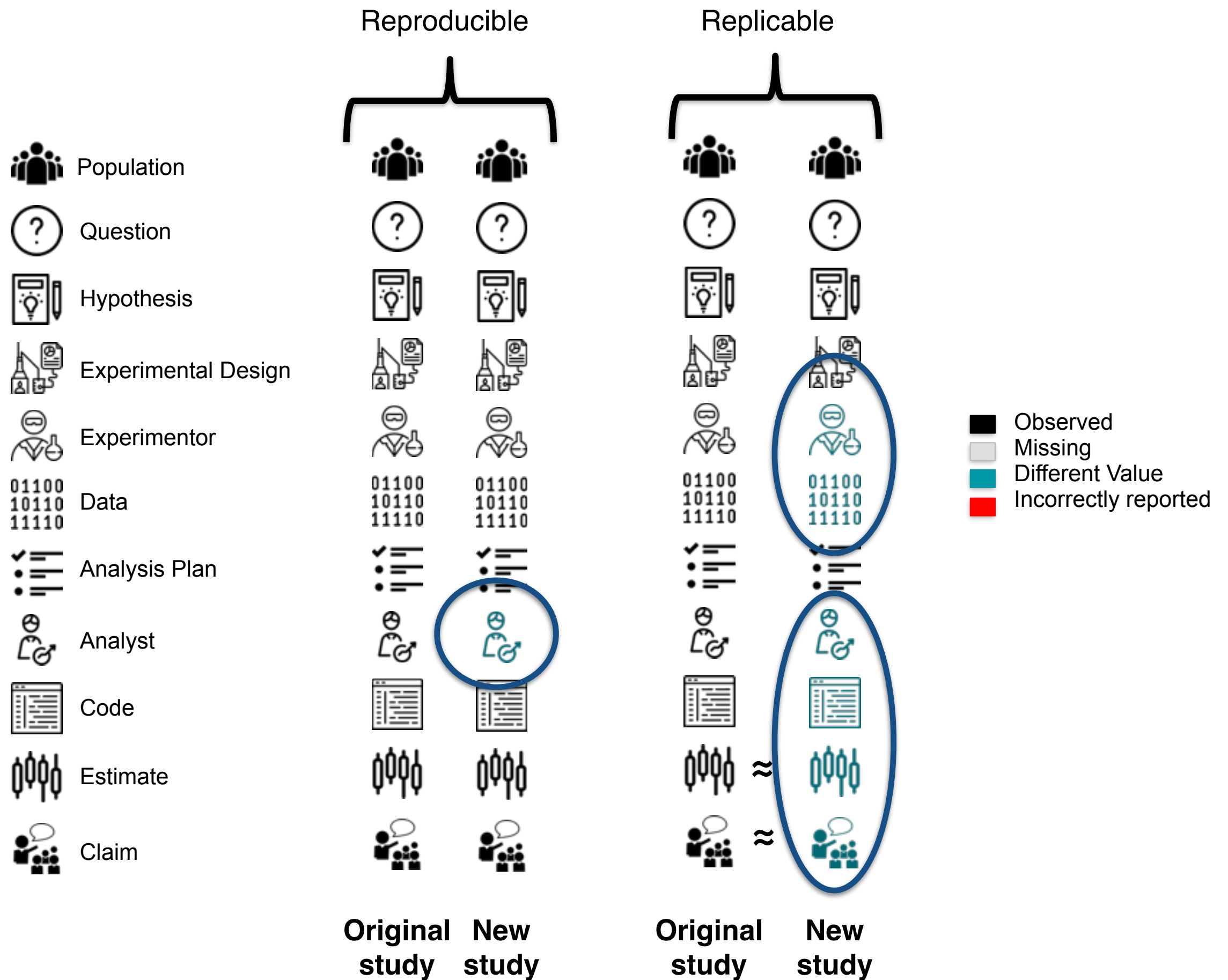
Do Nothing

Full Replication

Reproducibility

Most
common
activity



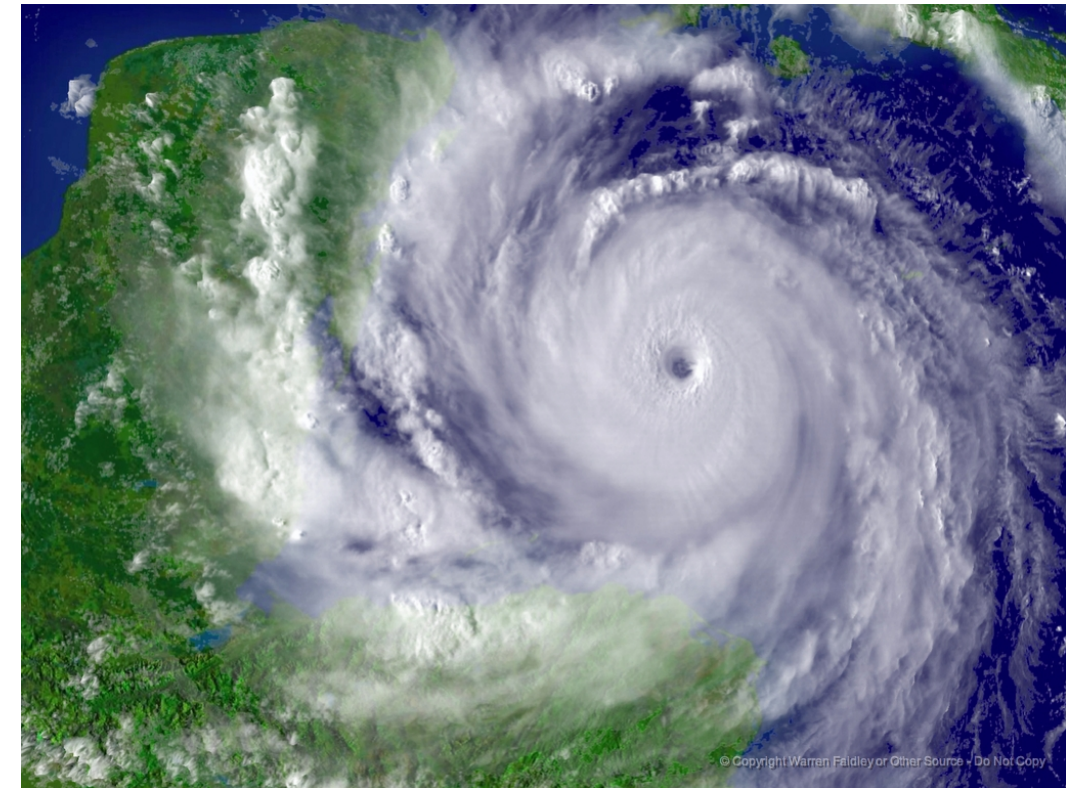


Why Do We Need Reproducible Research?

- 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?

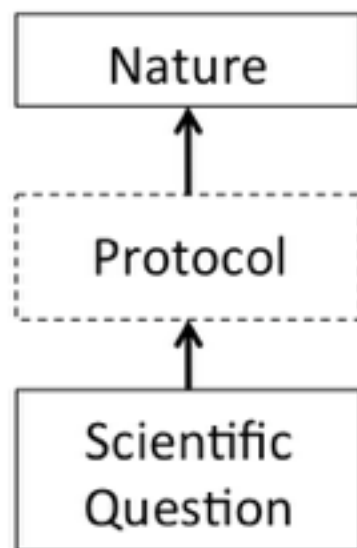
- Estimating small health effects in the presence of much stronger signals
- Results inform substantial policy decisions and affect many stakeholders
- EPA regulations can cost billions of dollars
- Complex statistical methods are needed and subjected to intense scrutiny



The End Result

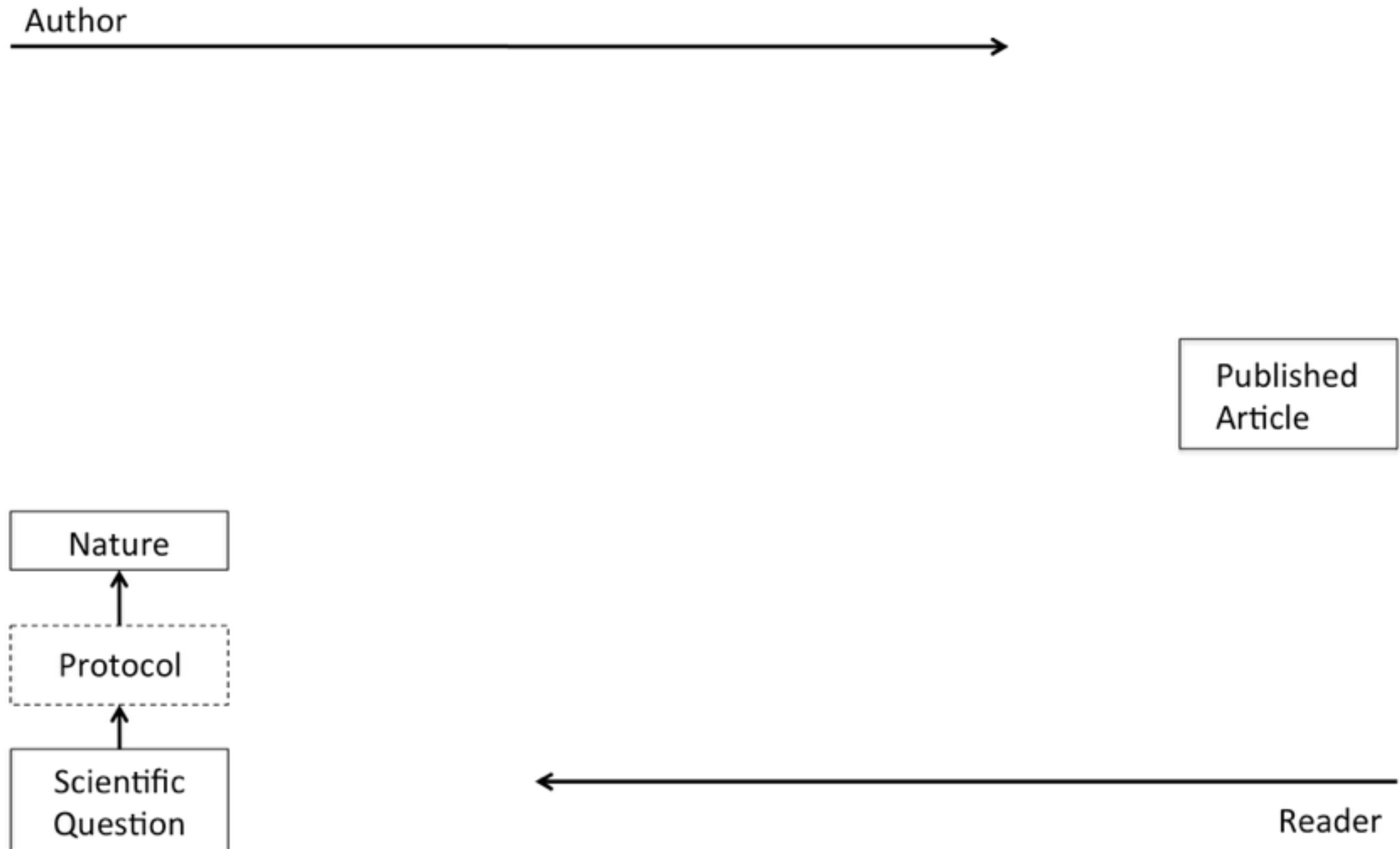
- Basic analyses can be difficult to describe
- Heavy computational requirements are thrust upon people without adequate training in statistics and computing
- Errors are more easily introduced into long and complex analysis pipelines
- Knowledge transfer is limited
- Complicated analyses cannot be trusted

What is Reproducible Research?

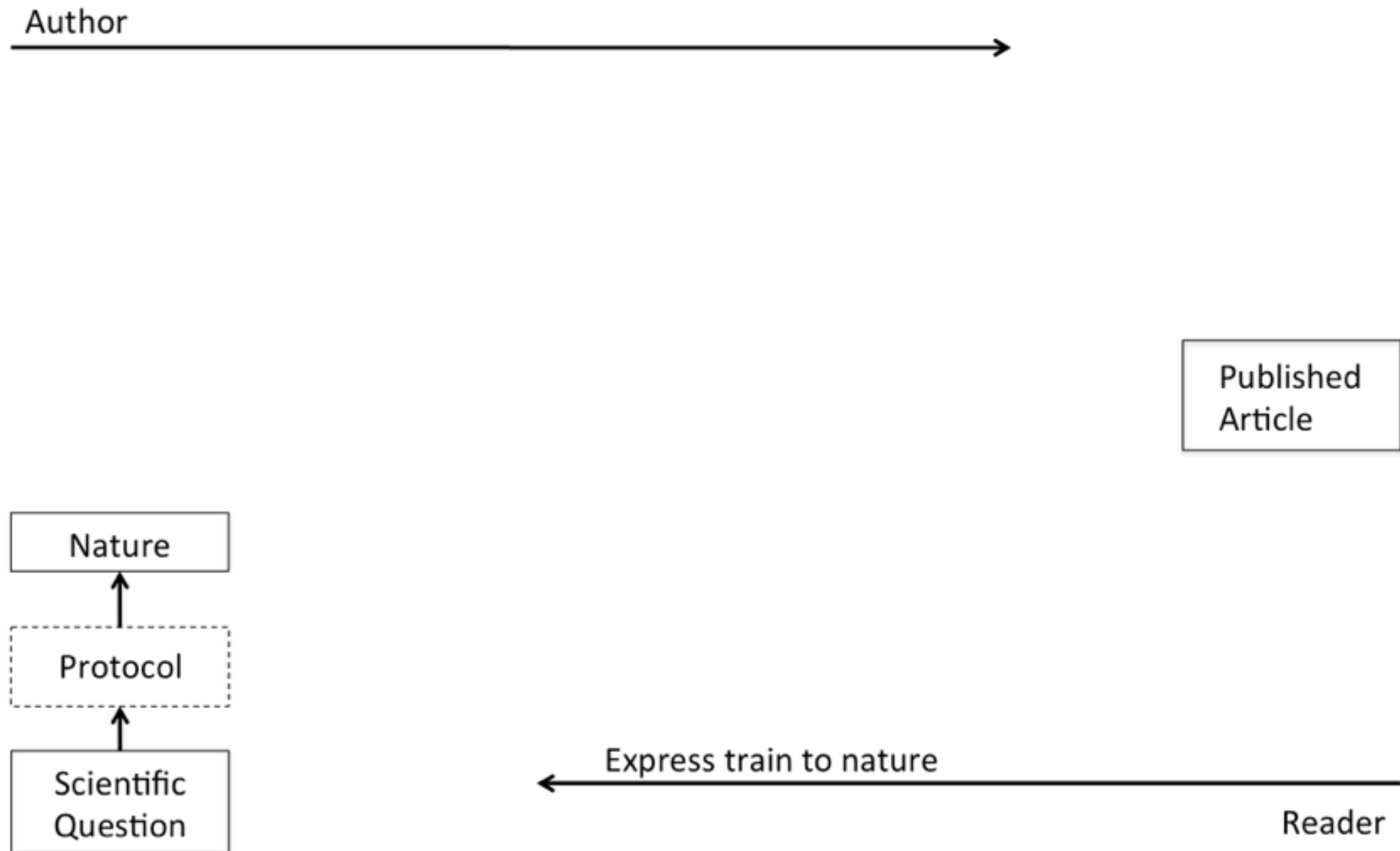


Published
Article

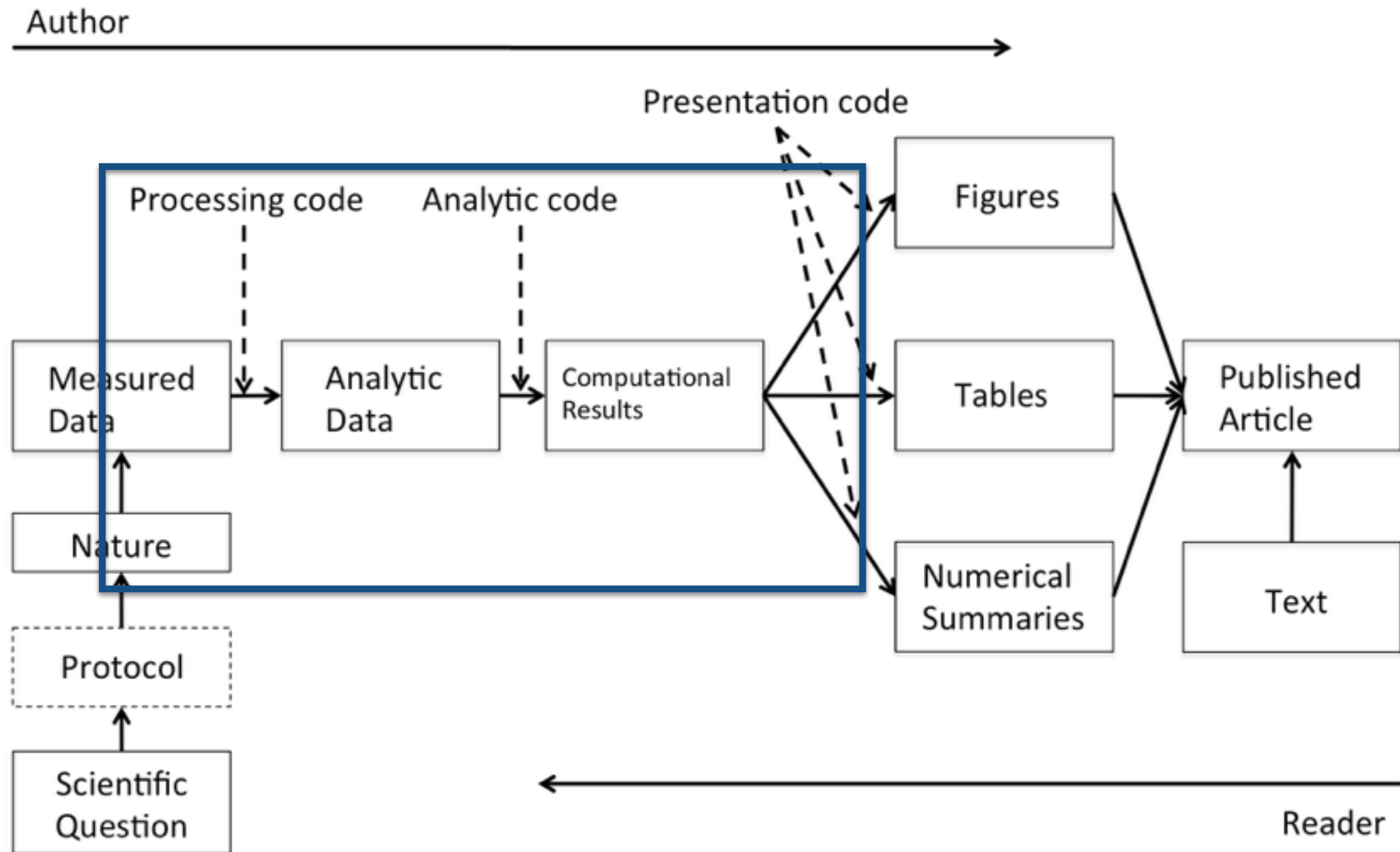
What is Reproducible Research?



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

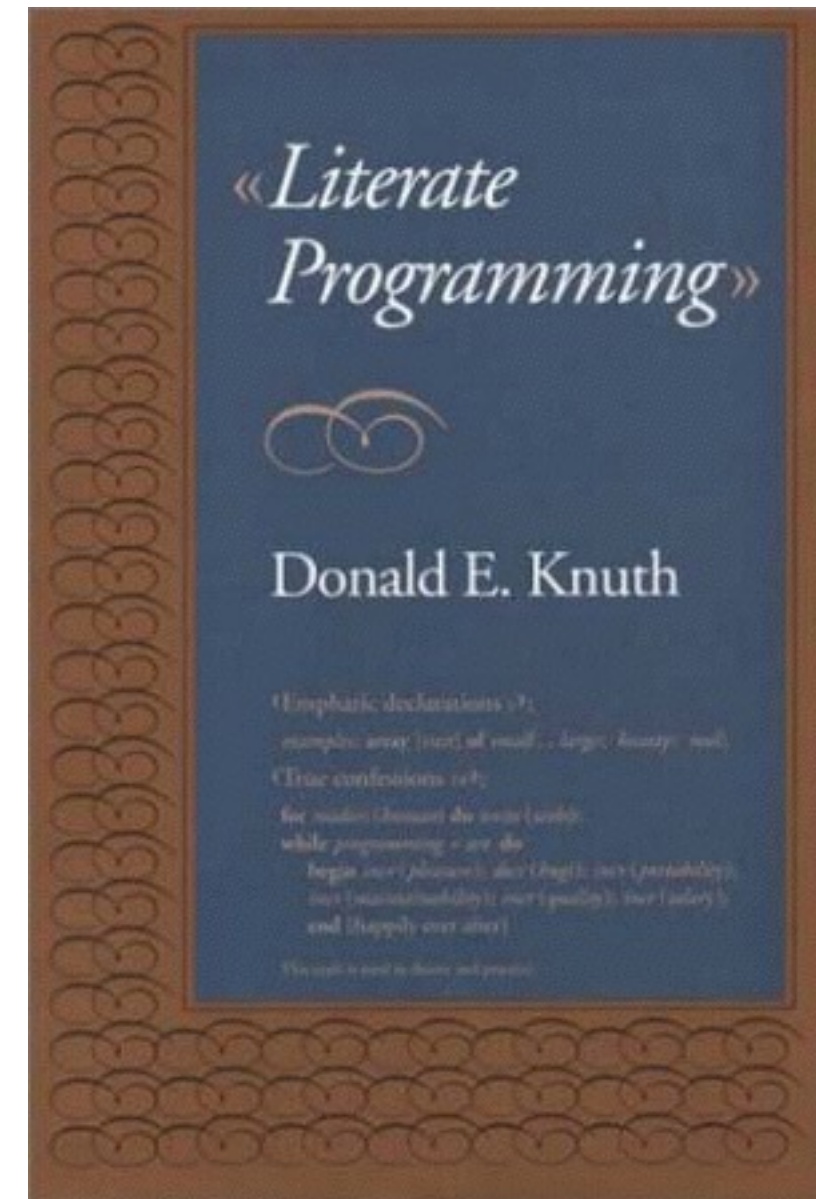
- Authors must undertake considerable effort to put data and results on the web (may not have resources like a web server)
- Readers must download data/results individually and piece together which data go with which code sections, etc.
- Readers may not have the same resources as authors
- Few tools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

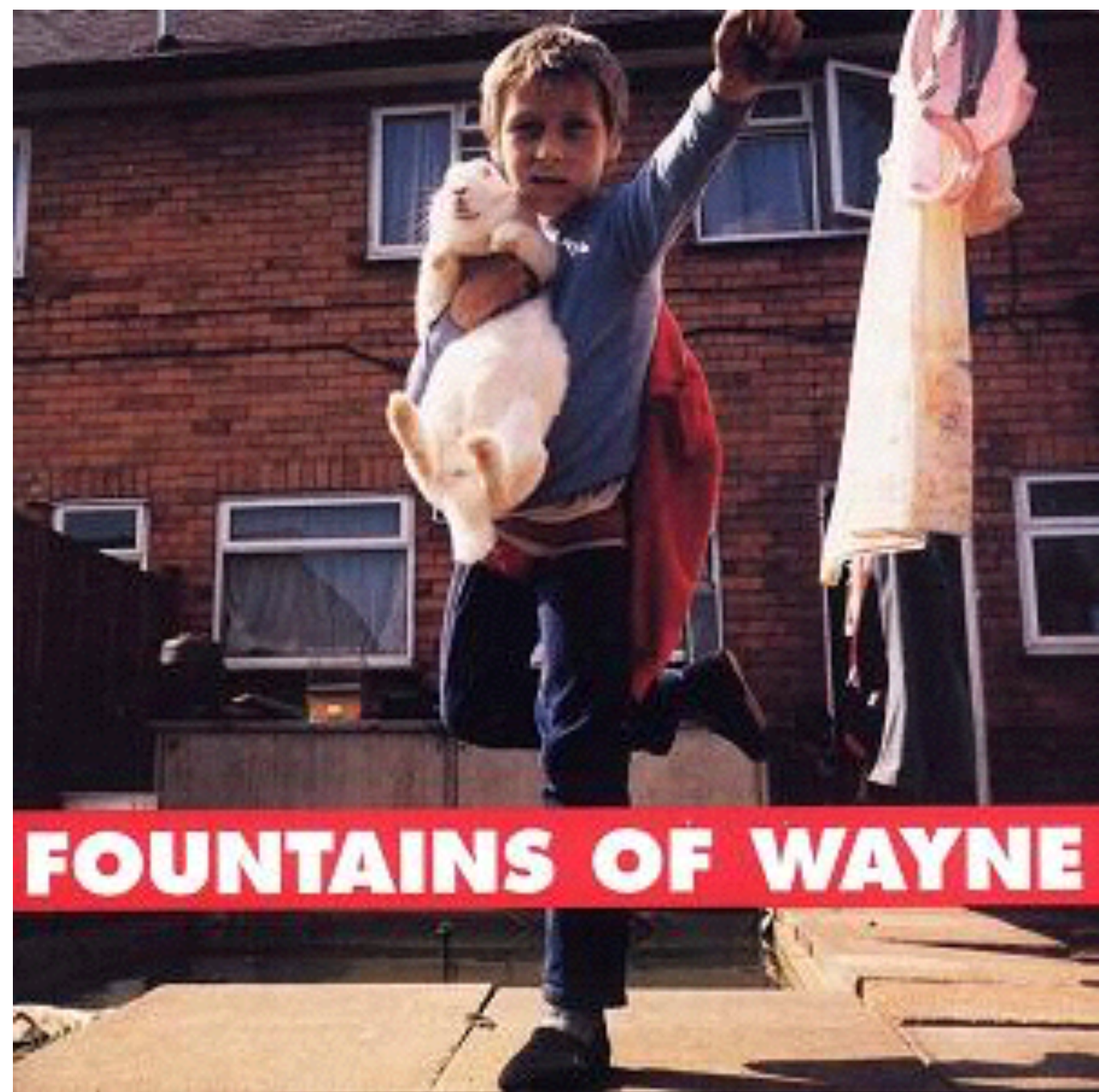
- knitr is package that brings together many features added on to Sweave to address limitations
- knitr uses R as the programming language knitr was developed by Yihui Xie (while a graduate student in statistics at Iowa State, now at RStudio)
- knitr uses the R programming language (although others are 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



Verse 1:

F Bb F
It's you and me on a beach/ In nineteen-ninety eight
Bb Dm Bb F
Leaning into the breeze/ From the willows and rivermen grace
Am Bb F C
Are reborn in this place/ I'm assured the procedure is painless
F Bb F
The taxicab with no brakes/ around the mountain pass
Bb Dm Bb F
Keep your head in your hands/ if anybody asks what you mean when
Am Bb F C
You were picking a fight/ you were only complimenting the waitress

Chorus 1:

Am Bb
Give us a room, with a mountain view:
F C F C Bb
A tiny cabana by the water, yeah by the water and I,
C F Bb
Got a rental for an hour or two, for a ride up the coast
C F
And a dip in the ocean.

(Repeat Intro)

Verse 2:

F Bb F
The waterfront is alight/ with Citronella flame
Bb Dm
Tourists flash in the night/ from the grottoes and
Bb F Am Bb F
Gathering now on the heel-worn planks for a drunken
C
Form another, a mumble (?)
F Bb F
And lovers paddle a boat/ on the molten bay
Bb Dm
Peering into the the reeds/ on a ripple
Bb F Am Bb
And playing it cool/ in a bar by the pool
F C
With a Caribbean Kiss Amaretto

Verse 1:

F Bb F
It's you and me on a beach/ In nineteen-ninety eight
Bb Dm Bb F
Leaning into the breeze/ From the willows and rivermen grace
Am Bb F C
Are reborn in this place/ I'm assured the procedure is painless
F Bb F
The taxicab with no brakes/ around the mountain pass
Bb Dm Bb F
Keep your head in your hands/ if anybody asks what you mean when
Am Bb F C
You were picking a fight/ you were only complimenting the waitress

Chorus 1:

Am Bb
Give us a room, with a mountain view:
F C F C Bb
A tiny cabana by the water, yeah by the water and I,
C F Bb
Got a rental for an hour or two, for a ride up the coast
C F
And a dip in the ocean.

(Repeat Intro)

Verse 2:

F Bb F
The waterfront is alight/ with Citronella flame
Bb Dm
Tourists flash in the night/ from the grottoes and
Bb F Am Bb F
Gathering now on the heel-worn planks for a drunken
C
Form another, a mumble (?)
F Bb F
And lovers paddle a boat/ on the molten bay
Bb Dm
Peering into the the reeds/ on a ripple
Bb F Am Bb
And playing it cool/ in a bar by the pool
F C
With a Caribbean Kiss Amaretto

Verse 1:
It's you and me on a beach/ In nineteen-ninety
Leaning into the breeze/ From the willows and
Are reborn in this place/ I'm assured the proce
The taxicab with no brakes/ around the mountain
Keep your head in your hands/ if anybody asks
You were picking a fight/ you were only compli

Chorus 1:
Give us a room, with a mountain view:
A tiny cabana by the water, yeah by the water
Got a rental for an hour or two, for a ride up
And a dip in the ocean.

(Repeat Intro)

Verse 2:
The waterfront is alight/ with Citronella flame
Tourists flash in the night/ from the grottoes
Gathering now on the heel-worn planks for a dr
Form another, a mumble (?)
And lovers paddle a boat/ on the molten bay
Peering into the the reeds/ on a ripple
And playing it cool/ in a bar by the pool
With a Caribbean Kiss Amaretto

I. TEIL.

Hymnus: Veni, creator spiritus.

Aufführungsrecht vorbehalten.
Droits d'exécution réservés.

Allegro impetuoso.

1. Baß-Klarinette in B.

1. 2. 3. 4. Fagott.

Kontra-Fagott.

1. 2. 3. 4. Trompete in F.

1. 2. 3. 4. Posaune.

Pauken.

Musikl.

Orgel.

Pedal.

1. 2. Sopran.

1. 2. Alt.

SOLI.

Tenor.

Bariton.

Baß.

Knabenchor.

Sopran.

Alt.

Tenor.

Baß.

I. CHOR.

Sopran.

Alt.

Tenor.

Baß.

II. CHOR.

1. Violine.

2. Violine.

Bratsche.

Violoncell.

Kontrabaß.

Allegro impetuoso.

1

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