JOHNS HOPKINS
BLOOMBERG SCHOOL
*of* PUBLIC HEALTH

# Introduction to Reproducible Research

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

Biostatistics 140.776

How do you know if a data analysis is successful?

When has a data analysis failed?

# Parable

ARTICLES

• Retracted •

nature medicine

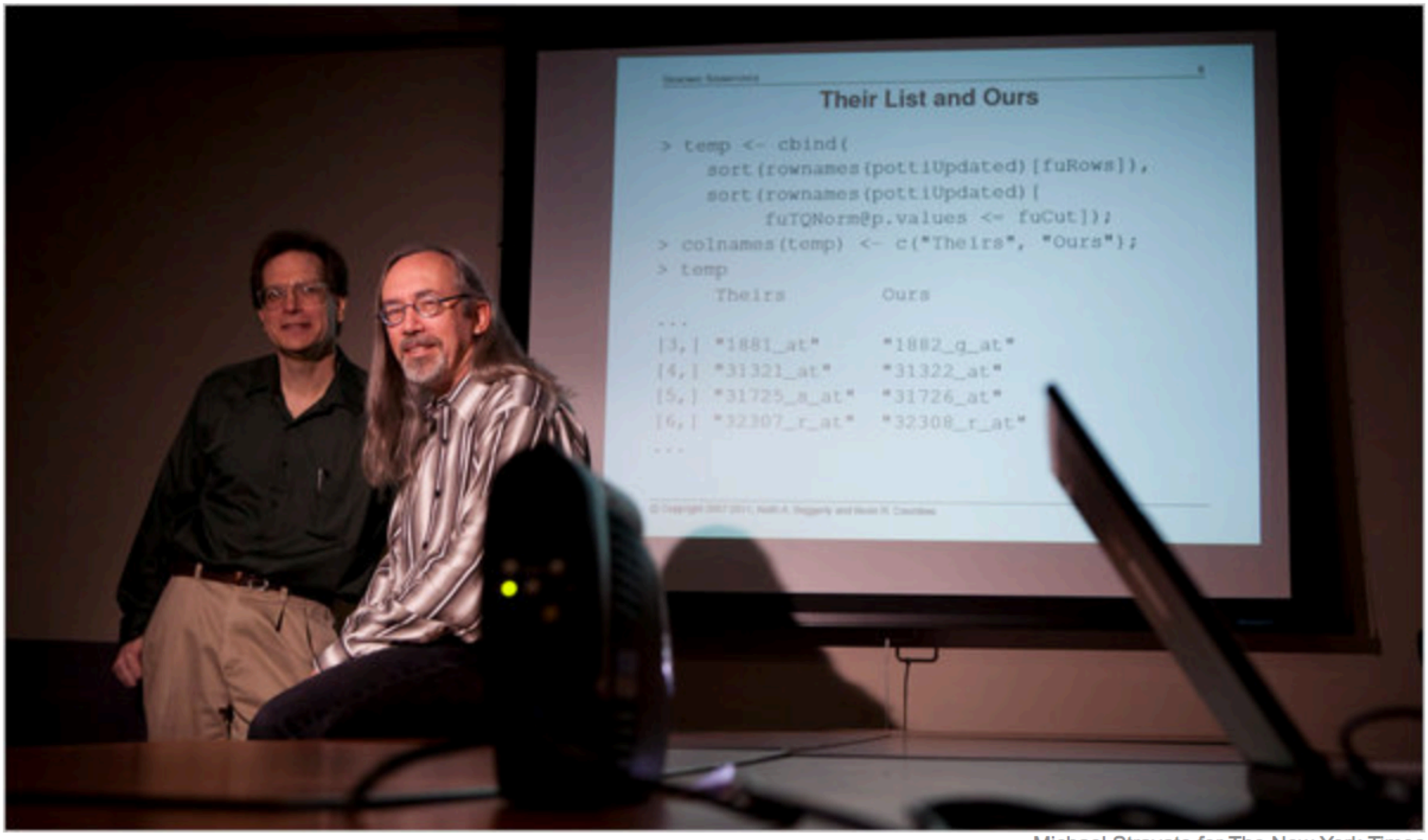
# Genomic signatures to guide the use of chemotherapeutics

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*

# Brief Summary of Problems

- Off-by-one table row labels
- Inadvertent switching of outcome labels
- Duplicated observations
- Genes identified not on microarray used
- Completely arbitrary statistical formulas used

# Lessons?

Clinical medicine (tightly controlled)

+

Genomics (wild west)

=

Wild West

# Institute of Medicine Committee

REPORT BRIEF MARCH 2012

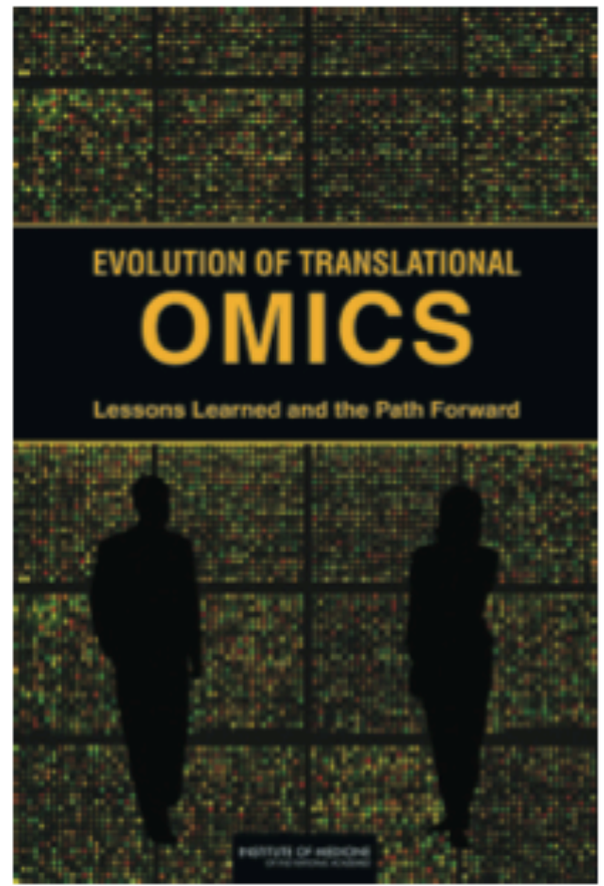
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

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 quality 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

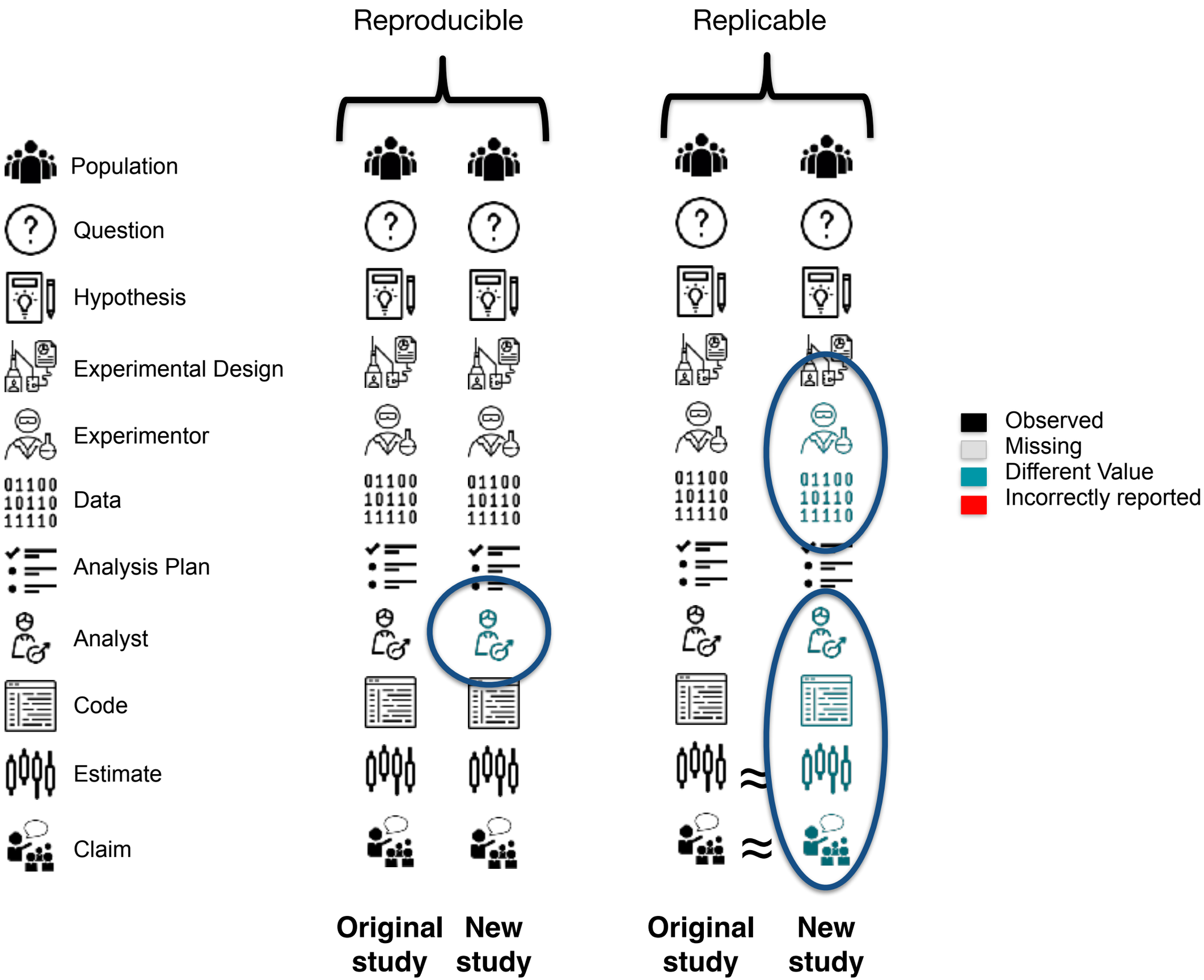
Minimum ⟷ Maximum

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”

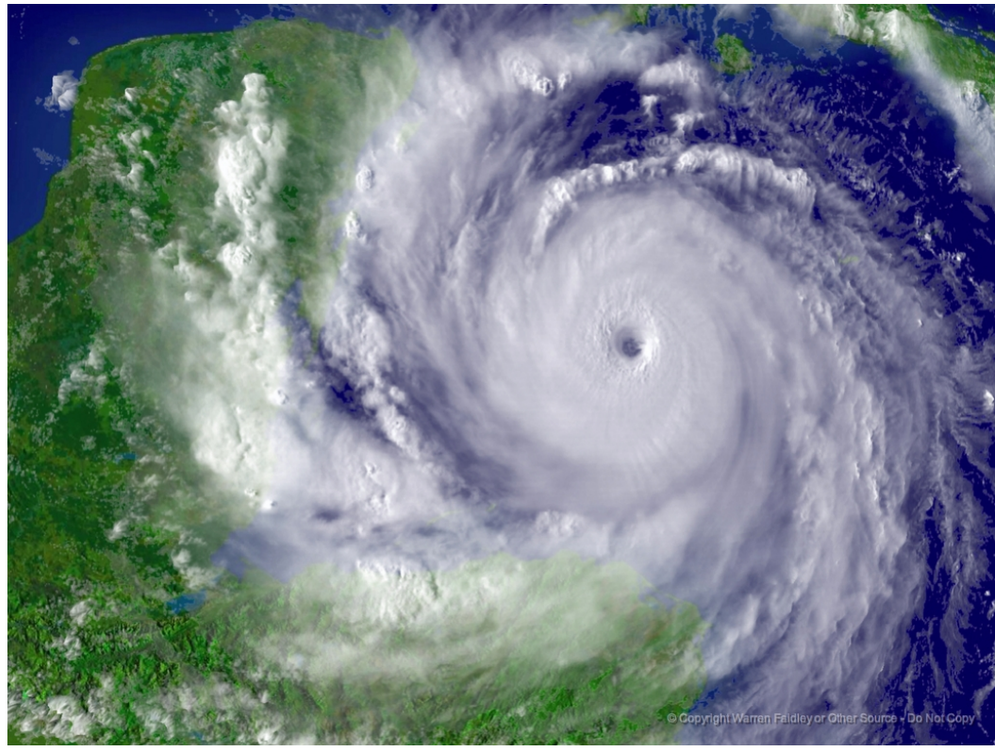
# Air Pollution and Health: A Perfect Storm?

- 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

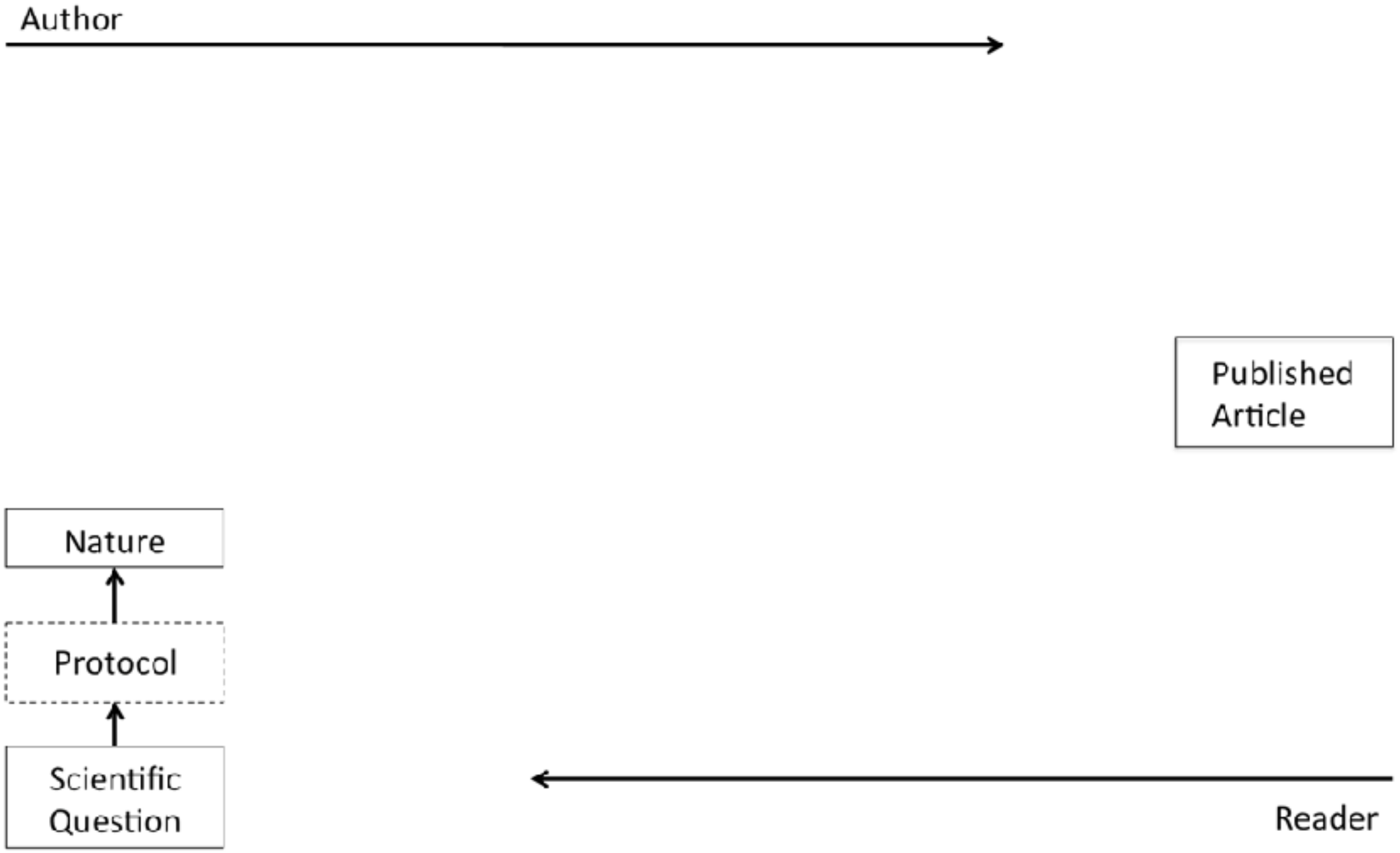


# The End Result

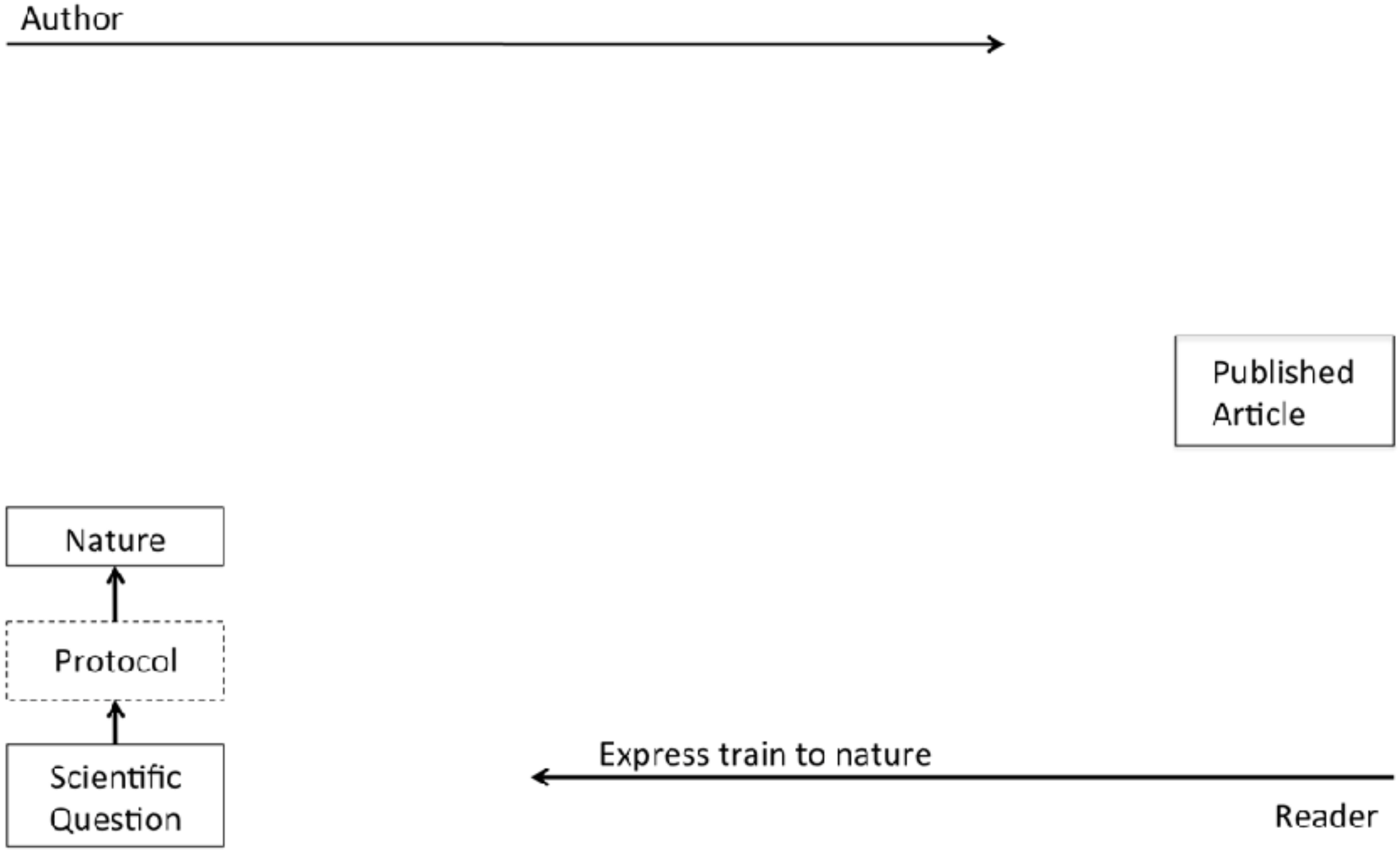
- Basic analyses can be difficult to describe
- Heavy computational requirements are thrust upon people without adequate training in statistics and computing
- Errors are more easily introduced into long and complex analysis pipelines
- Knowledge transfer is limited
- Complicated analyses cannot be trusted

# What is Reproducible Research?

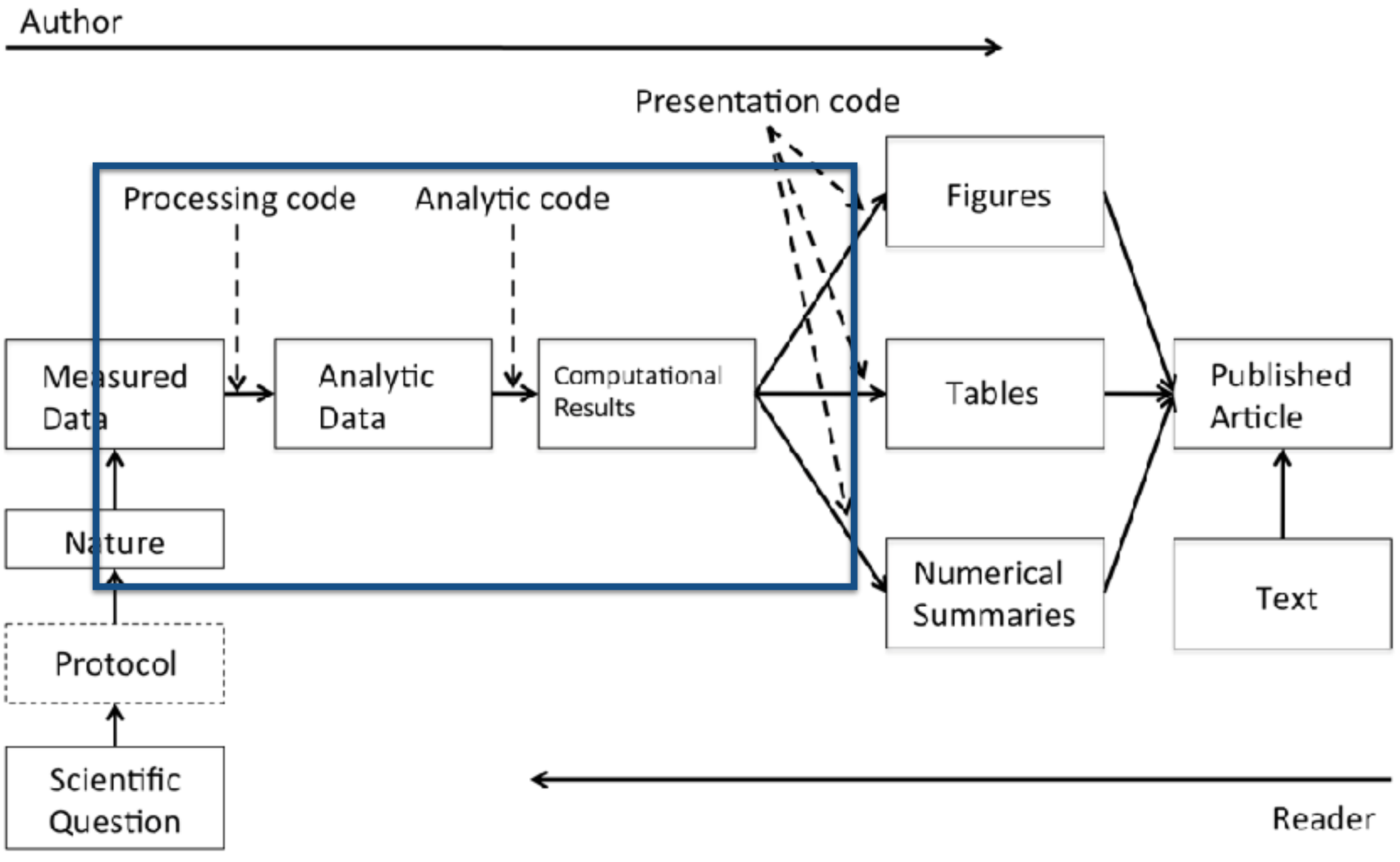
# What is Reproducible Research?



# 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

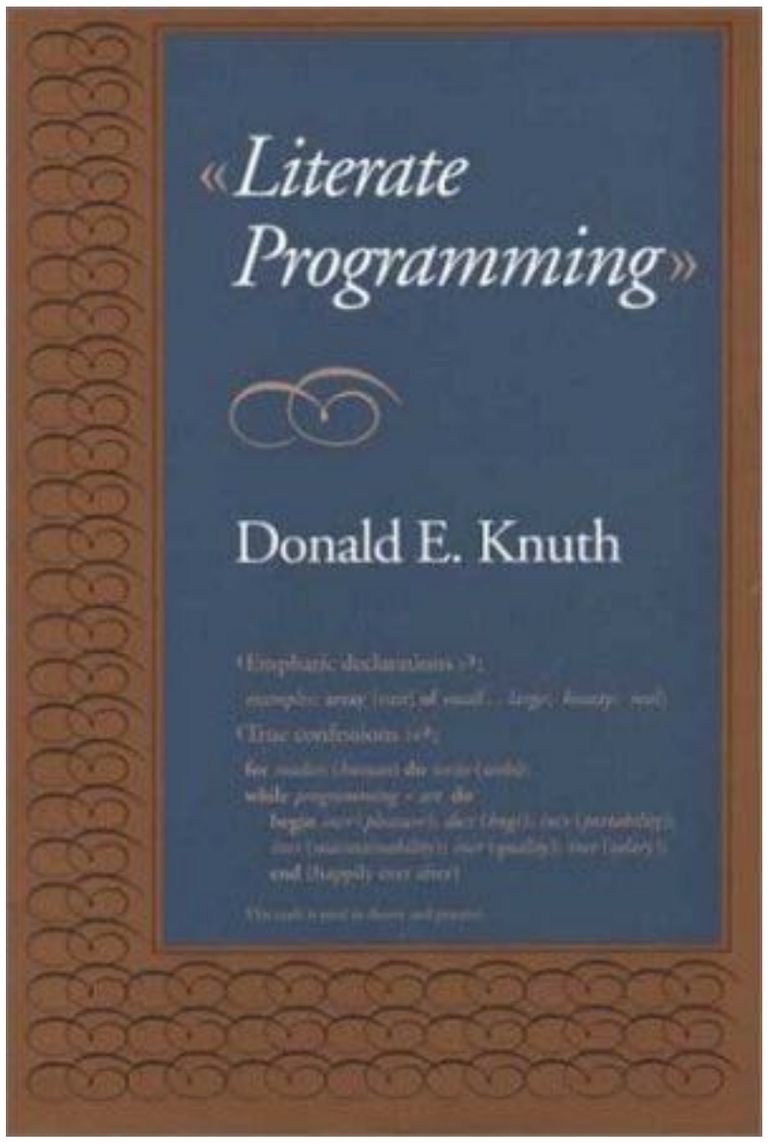
- **Software**: Jupyter Notebooks, knitr, markdown, LONI, Galaxy

- **Repositories**: GitHub, NCBI, ICPSR, Dataverse, Open Science Framework, *Google Dataset Search*

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

  - LaTeX, Markdown, HTML

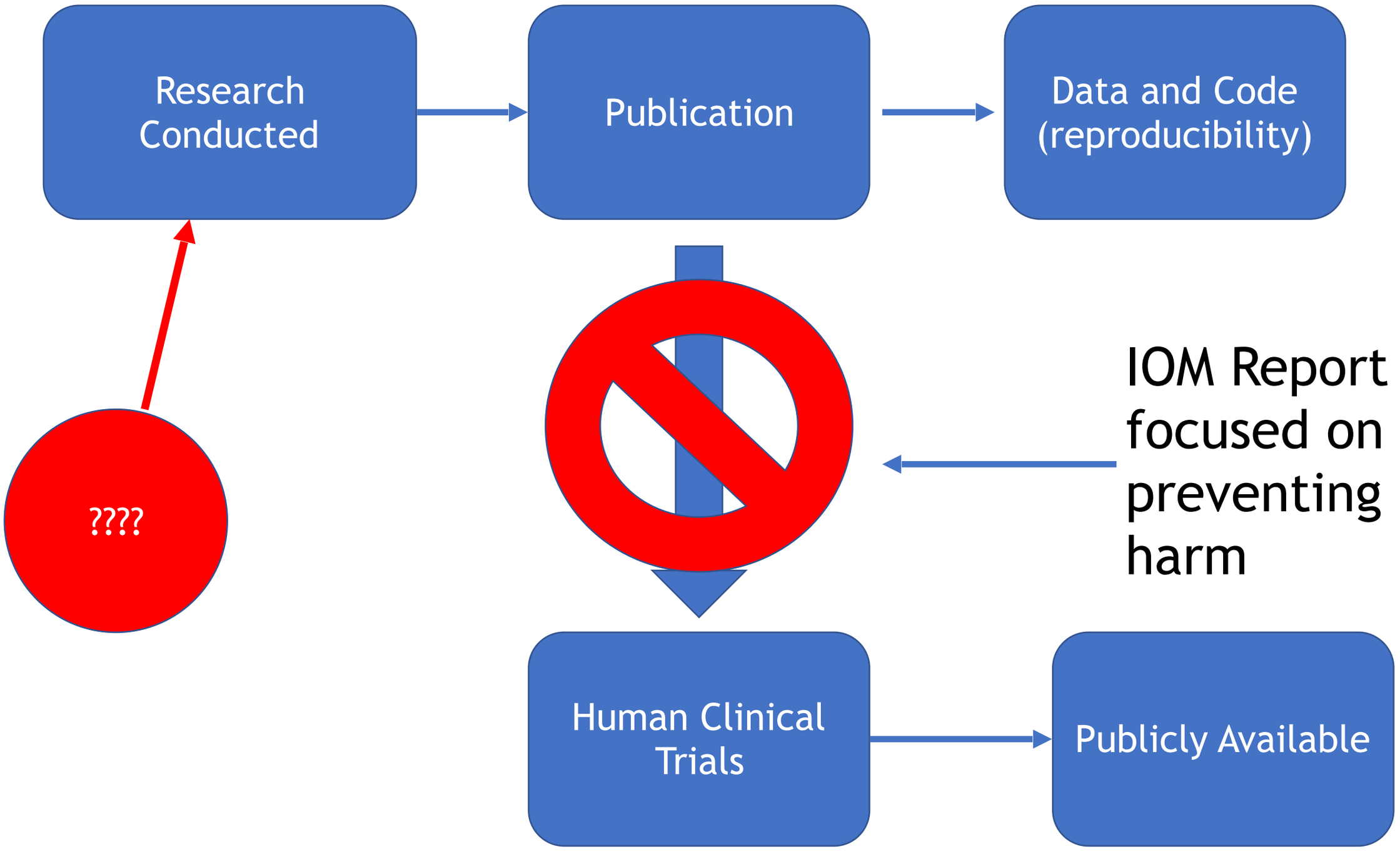
- Built into RStudio pipeline

- 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

# Where to Intervene?



# Lessons?

- Reproducibility
- Expertise and training
- Publication pressure; glamour journals
- Funding, conflicts of interest

# New Details Emerge (Jan 2015)

*"In raising these concerns, I have nothing to gain and much to lose."*
*— Bradford Perez*

# The Perez Memo (cont'd)

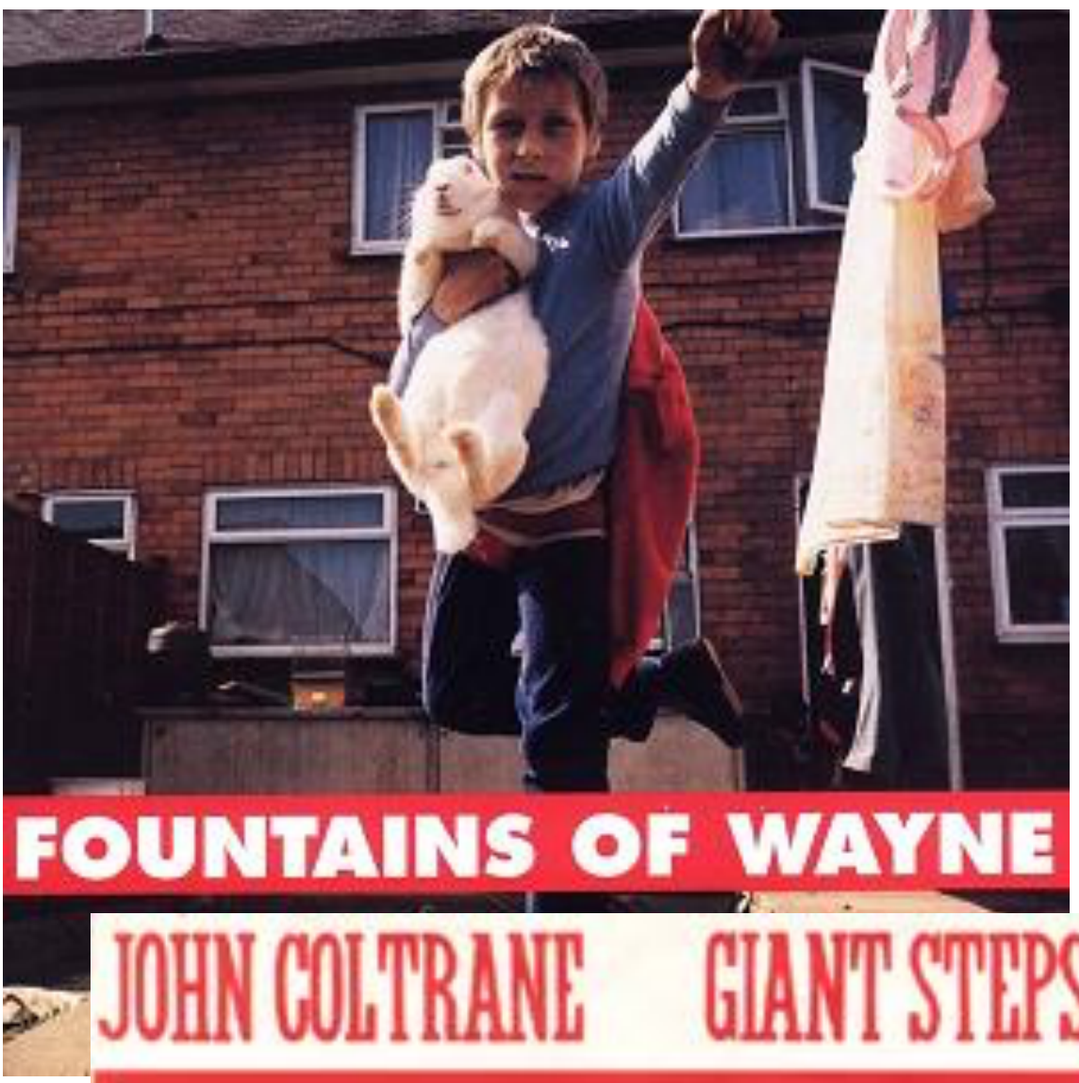
"At this point, I believe that the situation is serious enough that **all further analysis should be stopped** to evaluate what is known about each predictor and it should be reconsidered which are appropriate to continue using and under what circumstances…. I would argue that at this point nothing…should be taken for granted**. All claims of predictor validations should be independently and blindly performed.**" [emphasis added]

*-Memo from Bradford Perez, April 2008*

# Lessons Learned

- Not all problems have a technical solution
- Analyses were not “too complicated” in that there was insufficient expertise; problems were readily recognized
- Lab/institute cultural problems lead to unwillingness to communicate obvious problems
- From the analyst perspective, a breakdown in communication is an early warning sign of potential data analytic problems
- Making analyses more reproducible likely would not have made much difference

# Communicating a Data Analysis?


FOUNTAINS OF WAYNE


JOHN COLTRANE
GIANT STEPS

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 askes 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
It's you and me on a beach/ In nineteen-n
Bb Dm
Leaning into the breeze/ From the willows
Am Bb F
Are reborn in this place/ I'm assured the
F Bb
The taxicab with no brakes/ around the mo
Bb Dm
Keep your head in your hands/ if anybody
Am Bb F
You were picking a fight/ you were only c

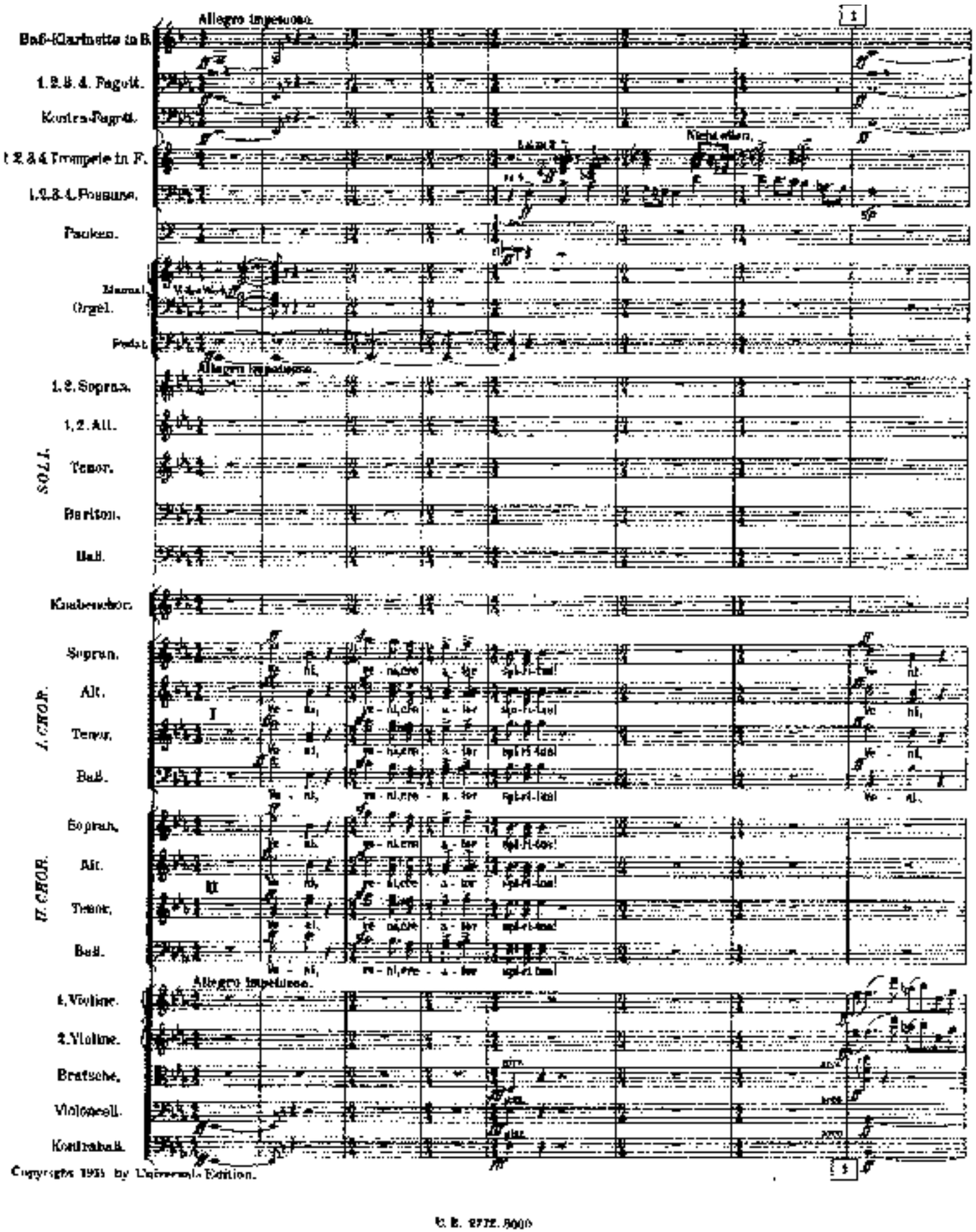
Chorus 1:

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

(Repeat Intro)

Verse 2:

F Bb
The waterfront is alight/ with Citronella
Bb Dm
Tourists flash in the night/ from the gro
Bb F Am Bb
Gathering now on the heel-worn planks for
C
Form another, a mumble (?)
F Bb F
And lovers paddle a boat/ on the molten b
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



CD

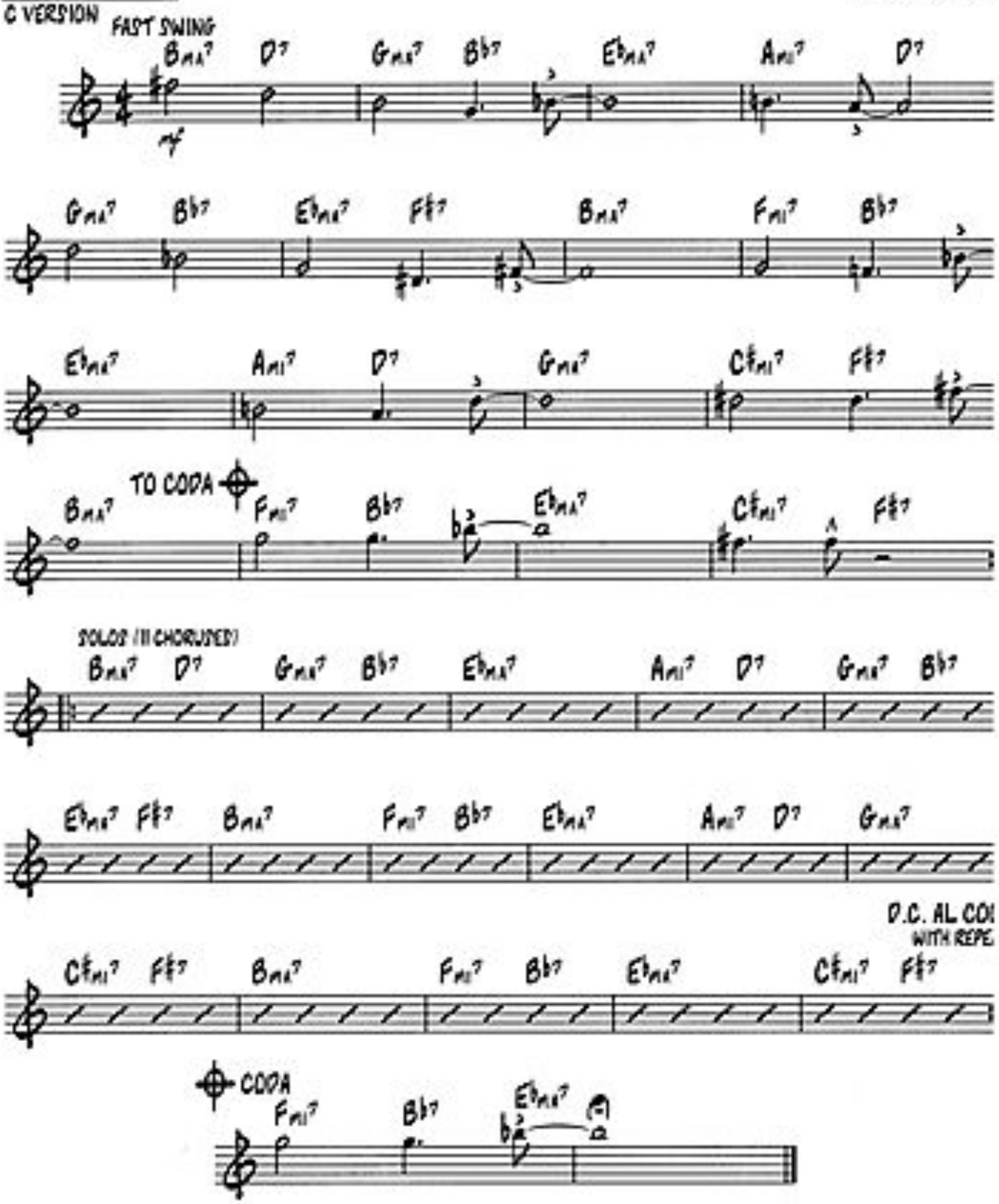
◆ : SPLIT TRACK/MELODY
◆ : FULL STEREO TRACK

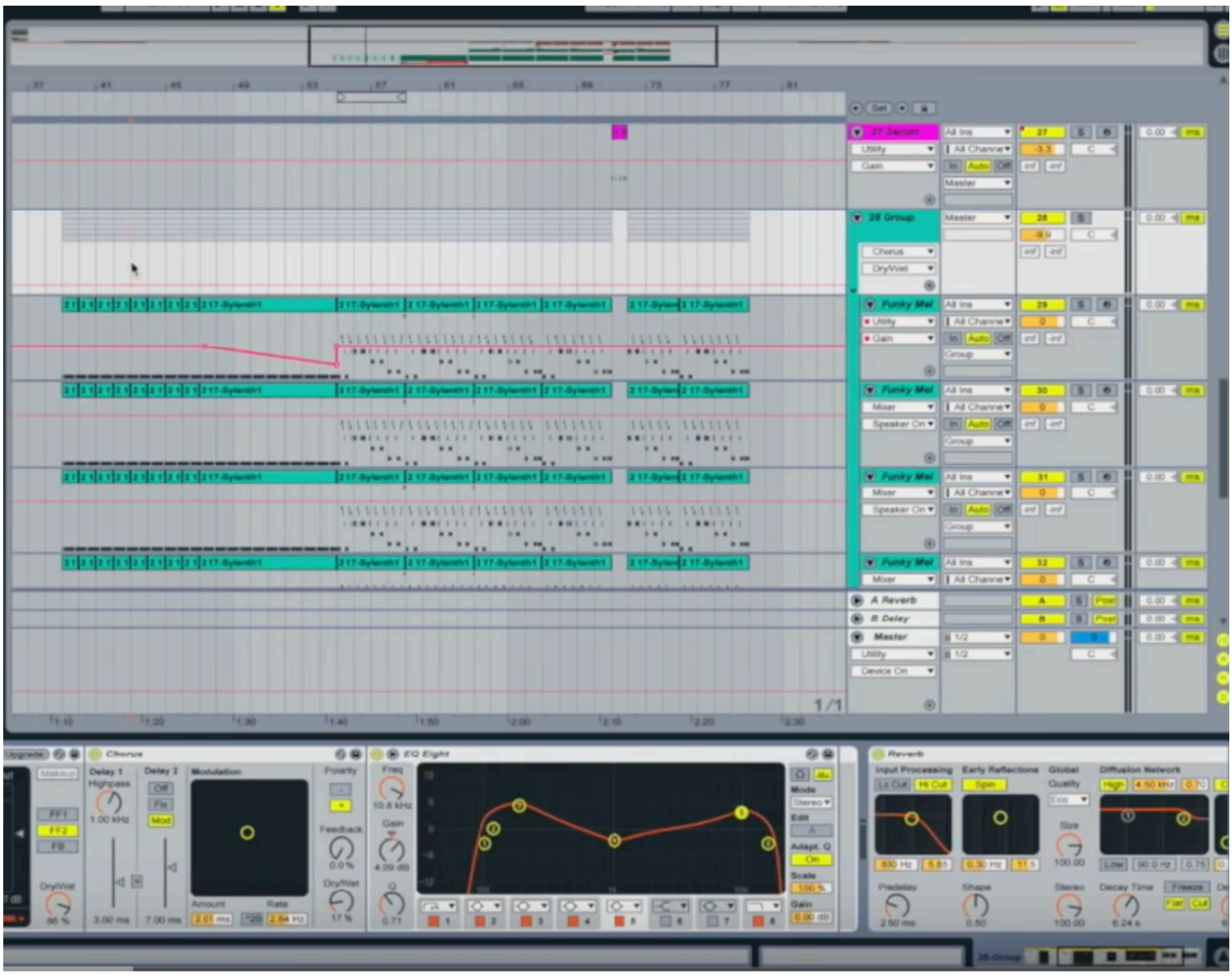
C VERSION

# GIANT STEPS

BY JOHN COLTRANE

FAST SWING




Chorus
Delay 1
Highpass
1.00 kHz
Delay 2
Off
Fix
Mod
Modulation
Amount
Rate
Polarity
Feedback
0.0 %
Dry/Wet
17 %
3.00 ms
7.00 ms
EQ Eight
Freq
10.8 kHz
Gain
4.09 dB
Q
0.71
Mode
Stereo
Edit
Adapt. Q
On
Scale
100 %
Gain
0.00 dB
Reverb
Input Processing
Early Reflections
Global
Diffusion Network
Quality
Size
100.00
Predelay
2.50 ms
Shape
0.50
Stereo
100.00
Decay Time
6.24 s
Freeze
Flat
Cut
Dry/Wet
86 %

# 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