
DATAPALOOZA
2020

Session 2:

Do It Right from the Start:

Ensuring Reproducibility/Confirmation

Susan Durham: Statistician, Ecology Center

David Bolton: Assistant Professor, Kinesiology & Health Sciences

Richard Cutler: Professor, Mathematics and Statistics Department

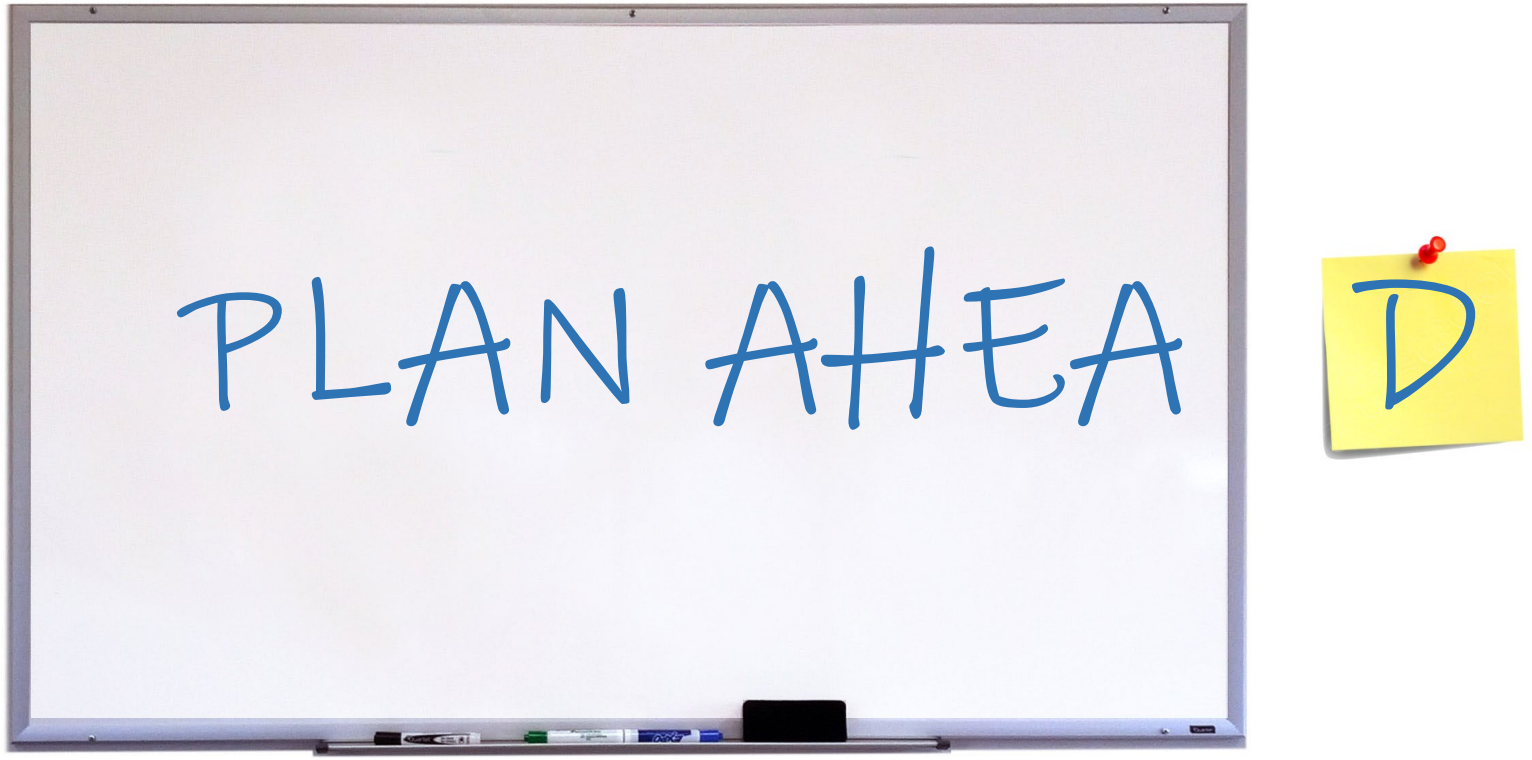
David Rosenberg: Associate Professor, Civil and Environmental Engineering

Doing it right from the start

Susan Durham

Ecology Center

Take-away point #1



Take-away point #2

Do not underestimate the knowledge, time and diligence required in data analysis

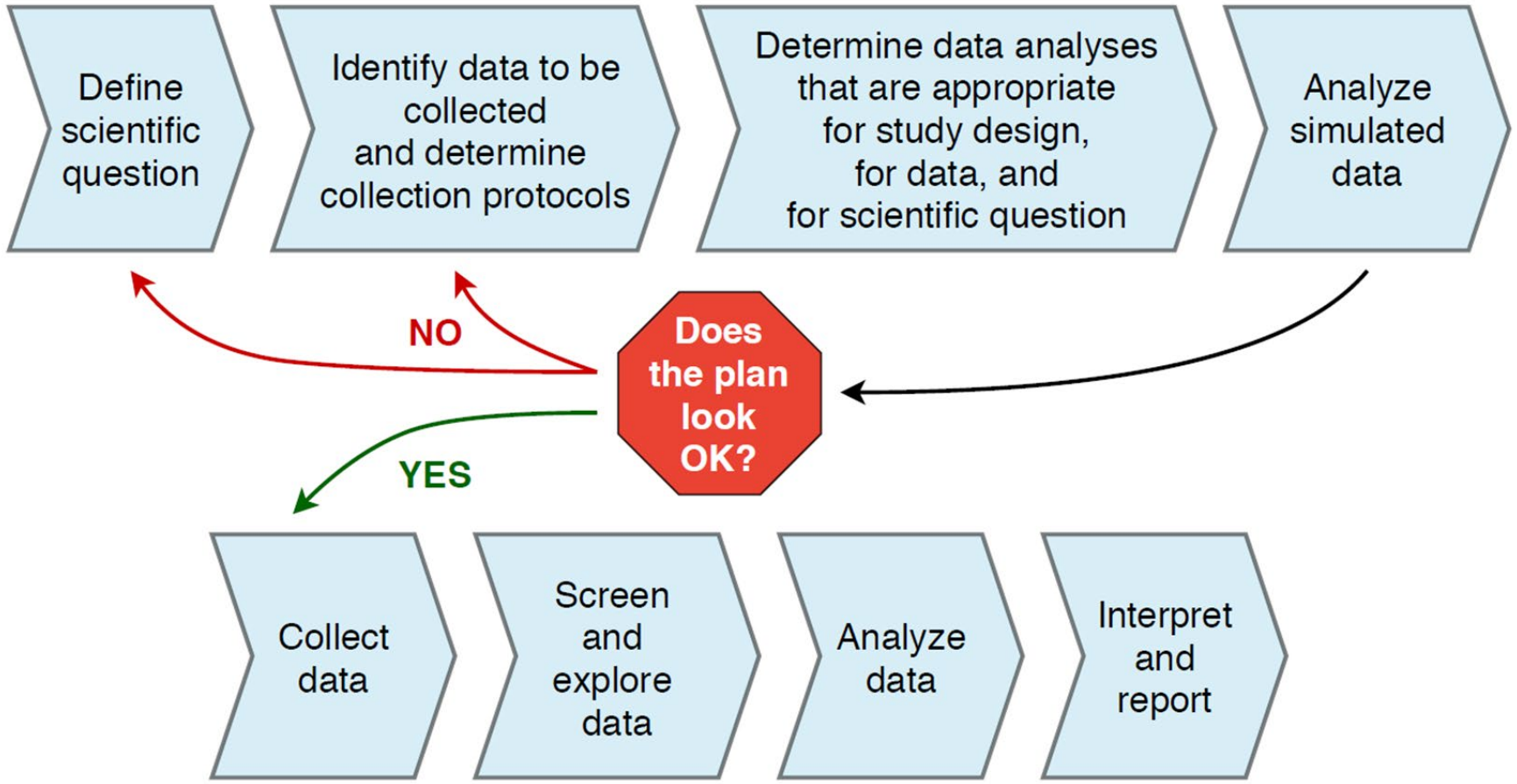
- Educate yourself
  - formal classes, workshops, seminars
  - focused self-study: online tutorials, literature, blogs
- Add the necessary methodology to your toolkit early
  - Coding in a scripting language
  - Data manipulation
  - Data visualization
  - Data analyses

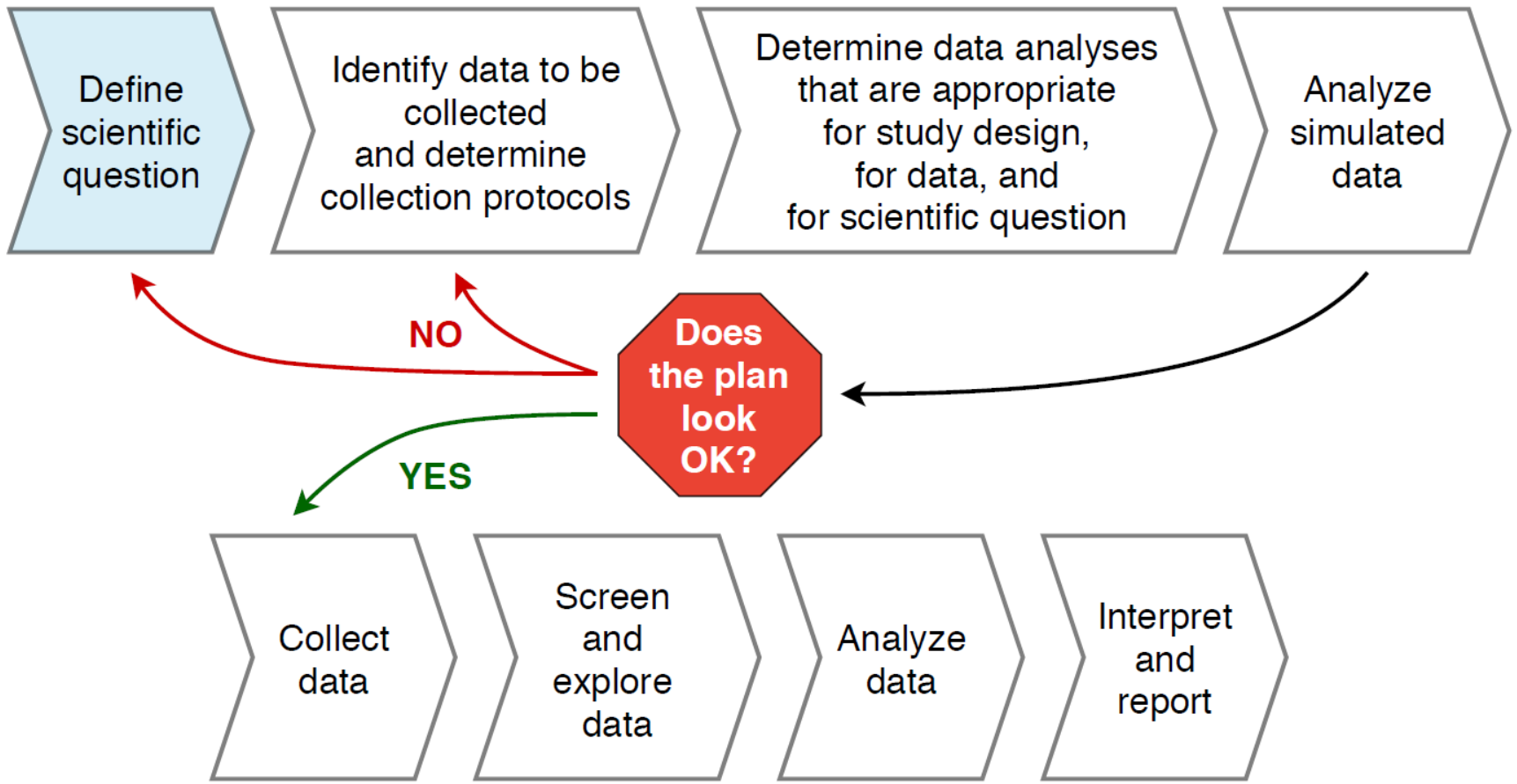
Take-away point #3

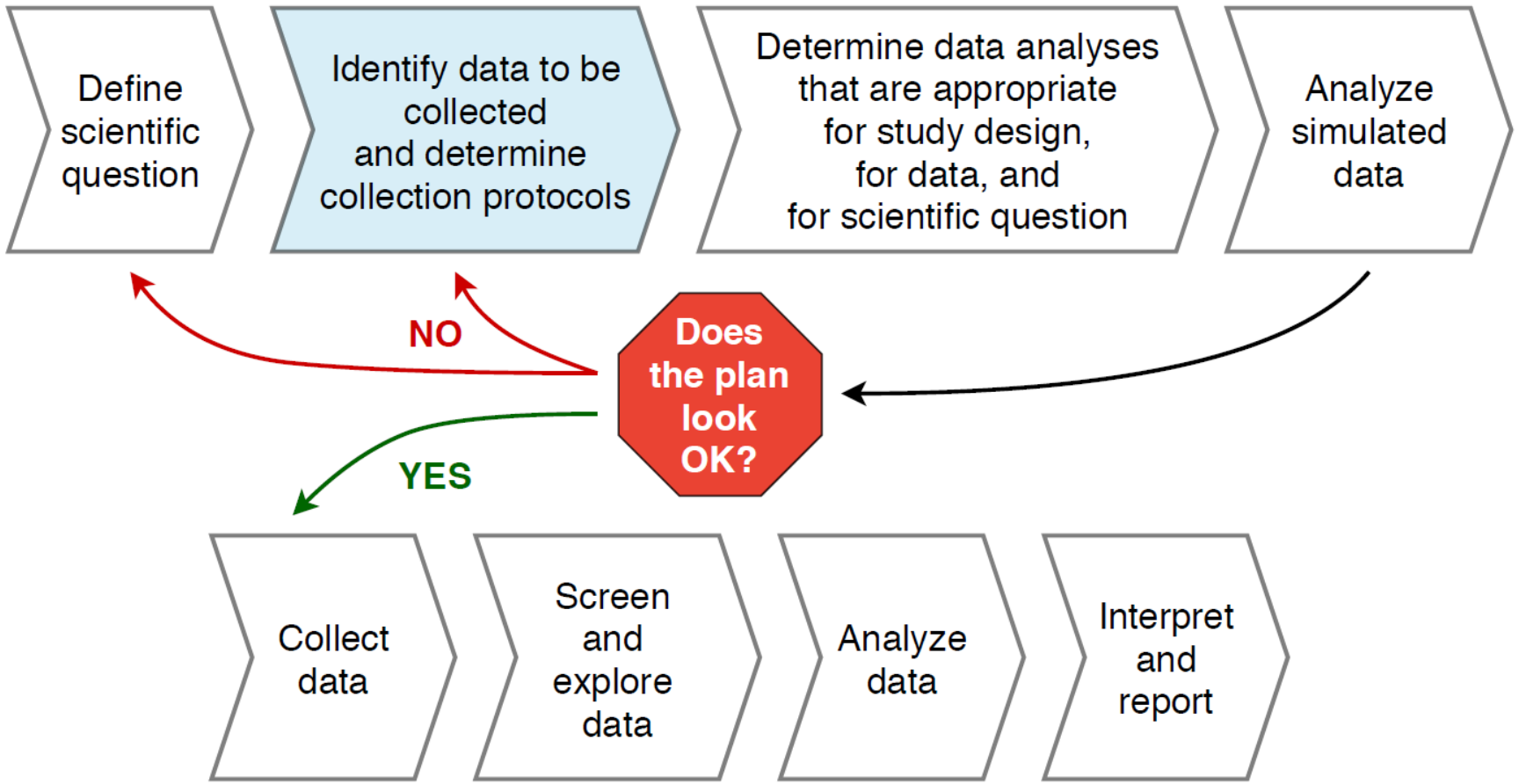
Draw upon the expertise of others

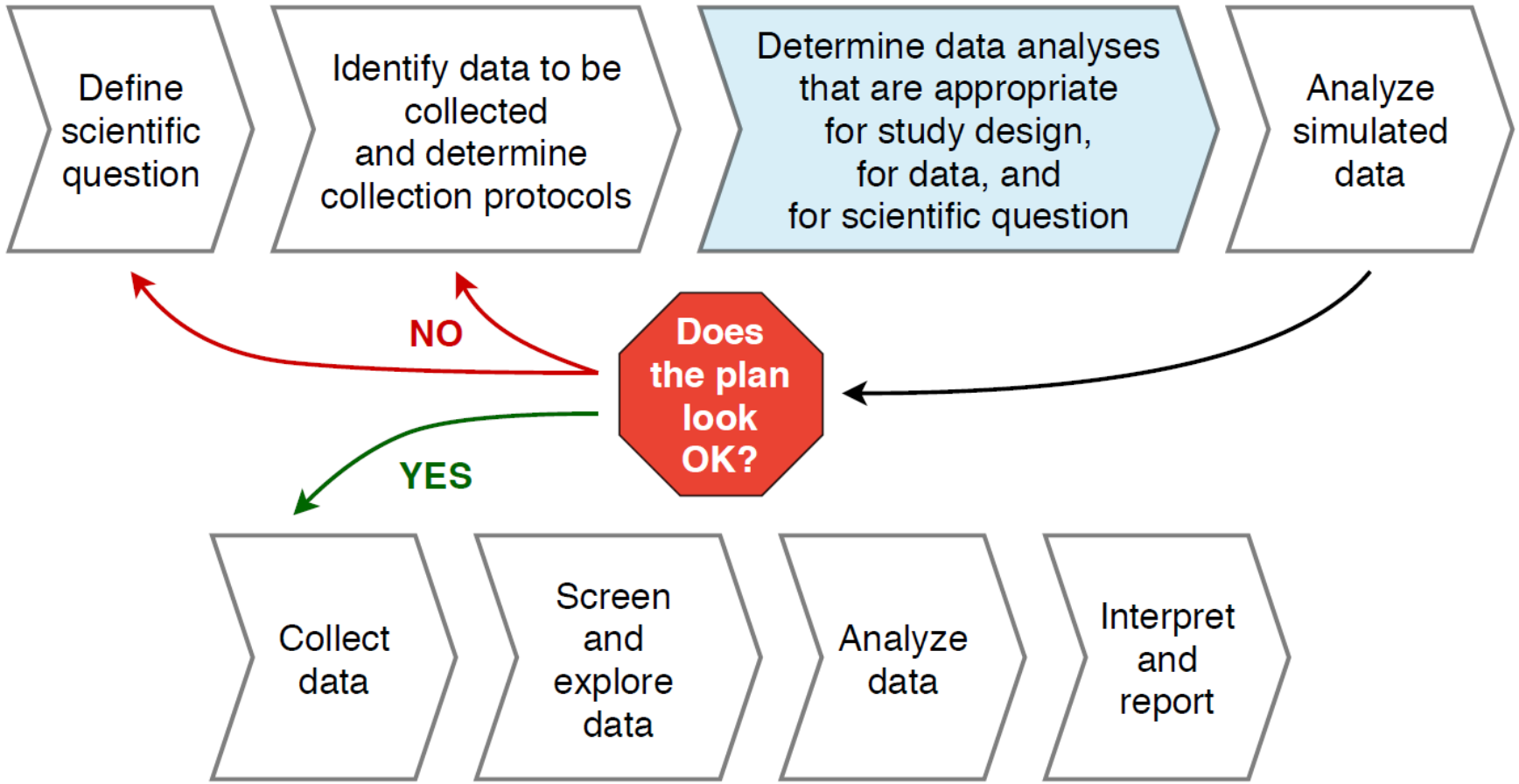
- Your advisor, committee, and faculty in your field
- Statisticians and quantitative colleagues
- Merrill-Cazier Library Research Data Management
- Writing Center

"You can't fix by analysis what you bungled by design."
–Light, Singer, and Willett (1990)



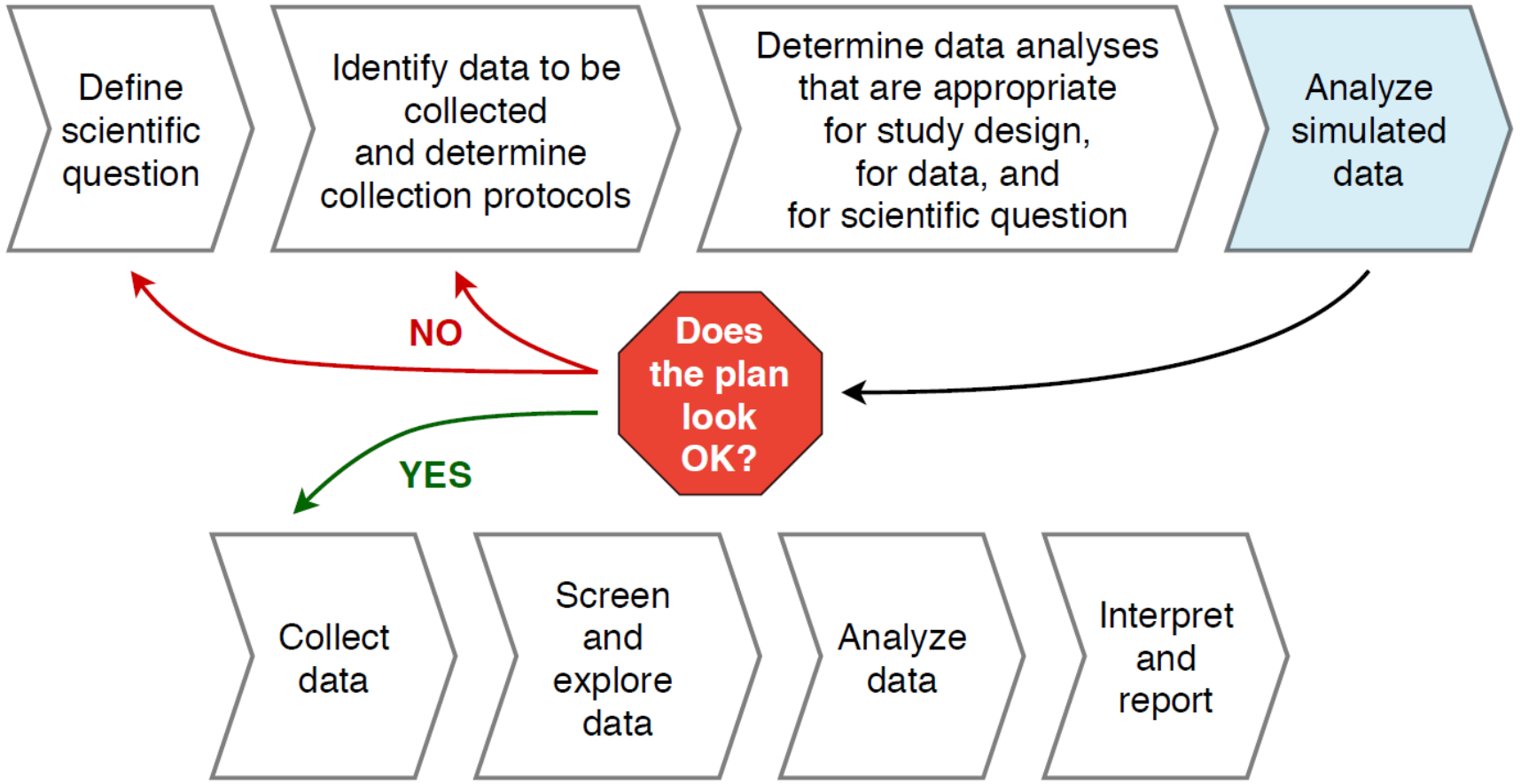

Define scientific question
Identify data to be collected and determine collection protocols
Determine data analyses that are appropriate for study design, for data, and for scientific question
Analyze simulated data
NO
Does the plan look OK?
YES
Collect data
Screen and explore data
Analyze data
Interpret and report

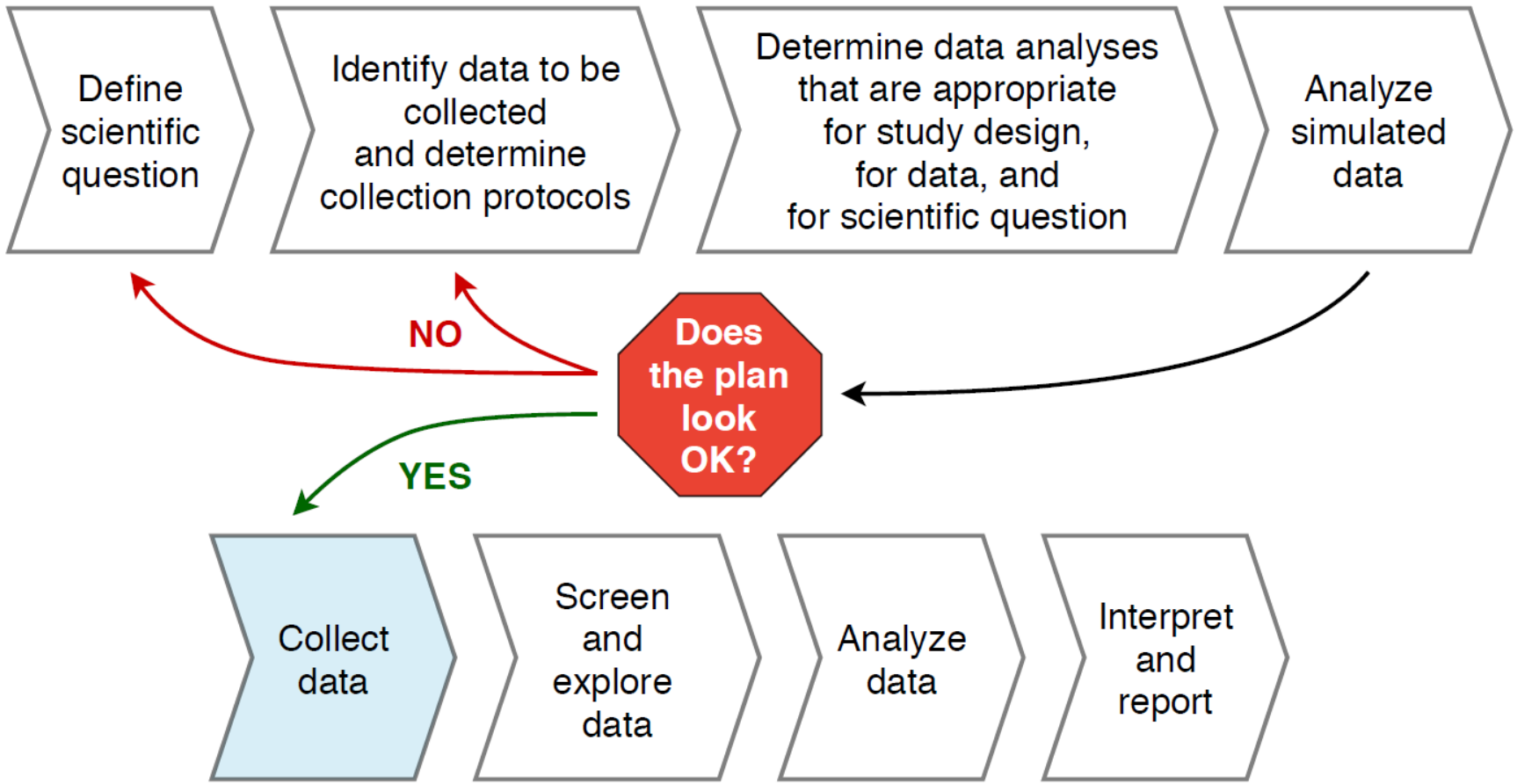

Define scientific question
Identify data to be collected and determine collection protocols
Determine data analyses that are appropriate for study design, for data, and for scientific question
Analyze simulated data
NO
Does the plan look OK?
YES
Collect data
Screen and explore data
Analyze data
Interpret and report

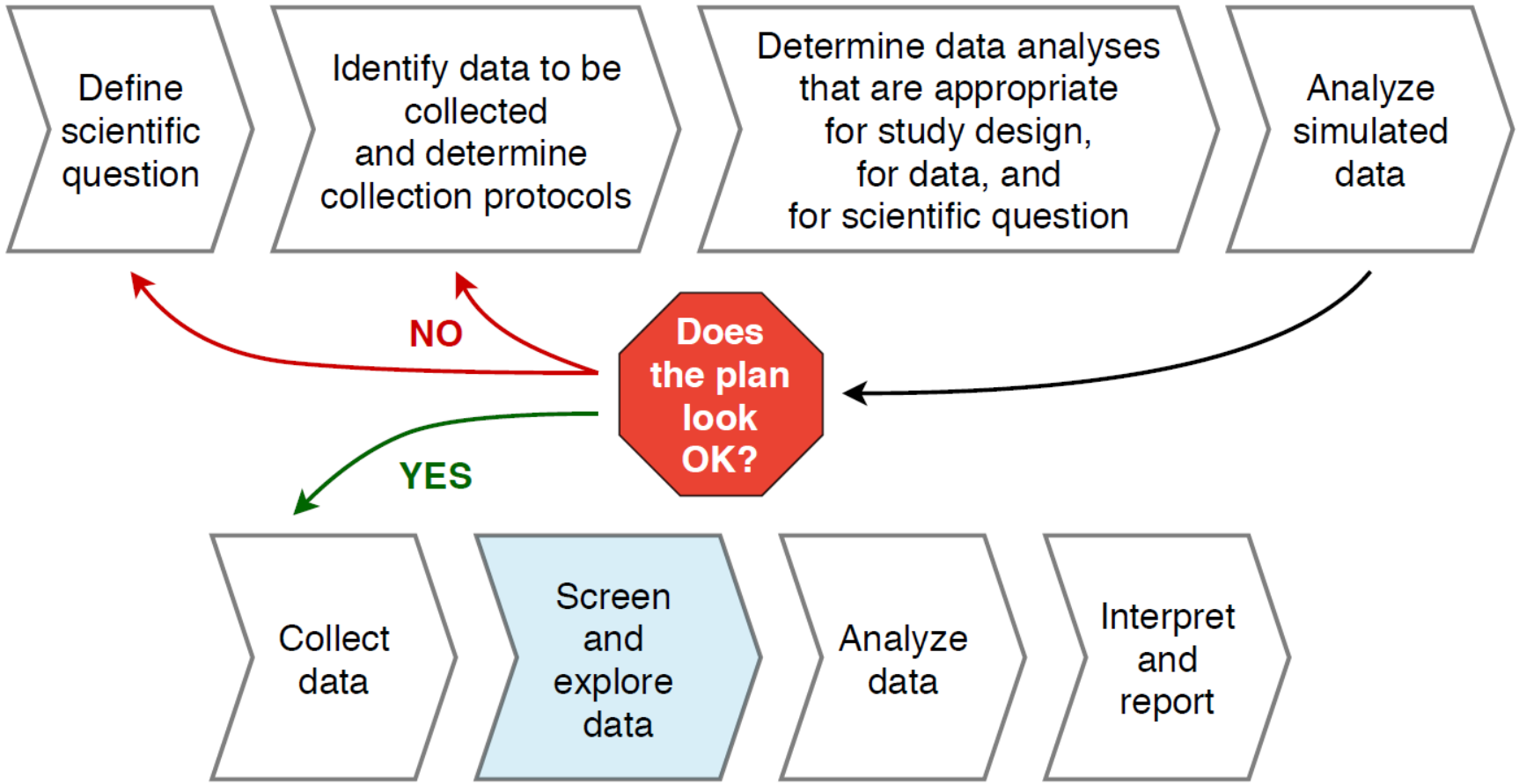

Define scientific question
Identify data to be collected and determine collection protocols
Determine data analyses that are appropriate for study design, for data, and for scientific question
Analyze simulated data
NO
Does the plan look OK?
YES
Collect data
Screen and explore data
Analyze data
Interpret and report

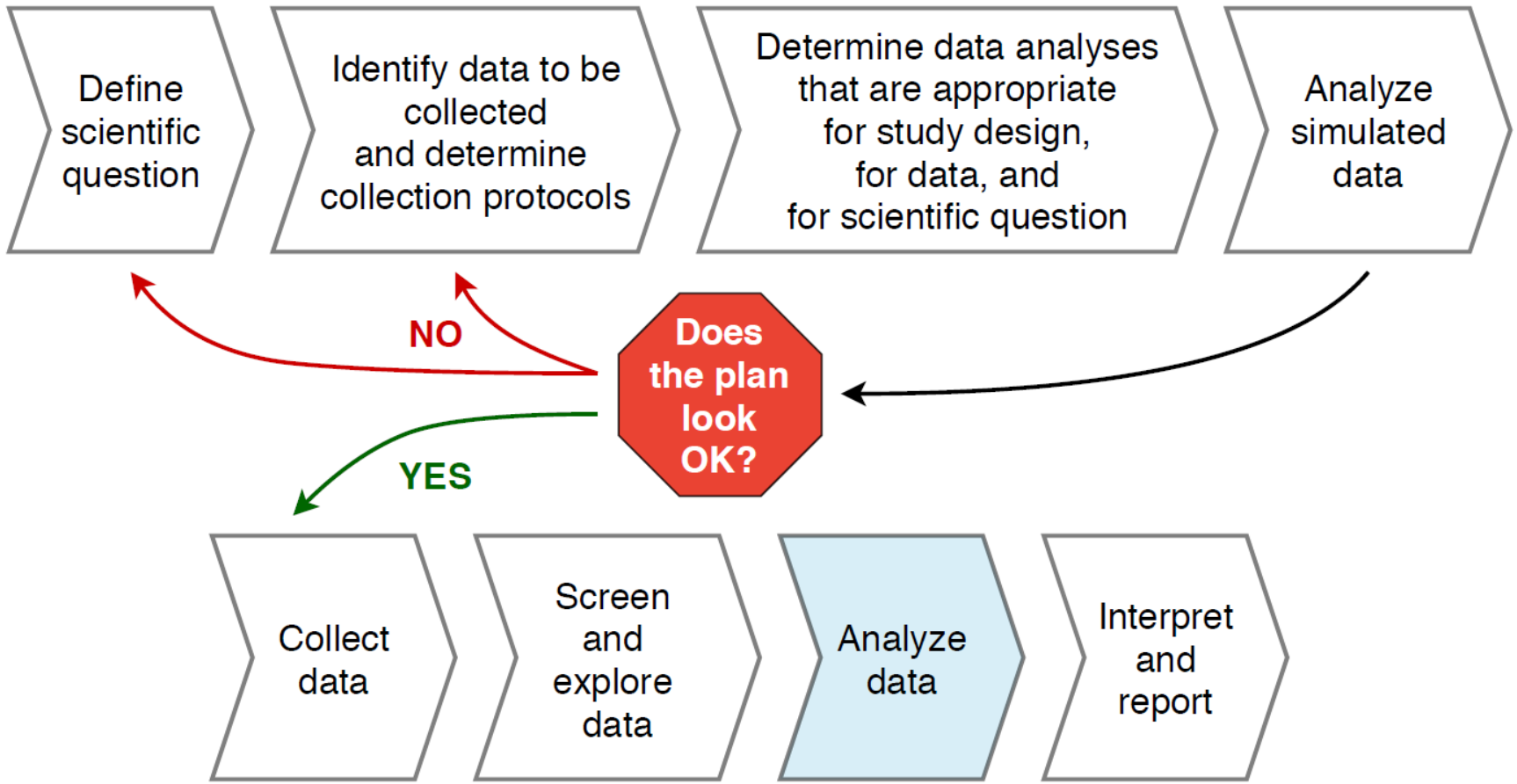
“No one should ever do an experiment without analyzing it first.”

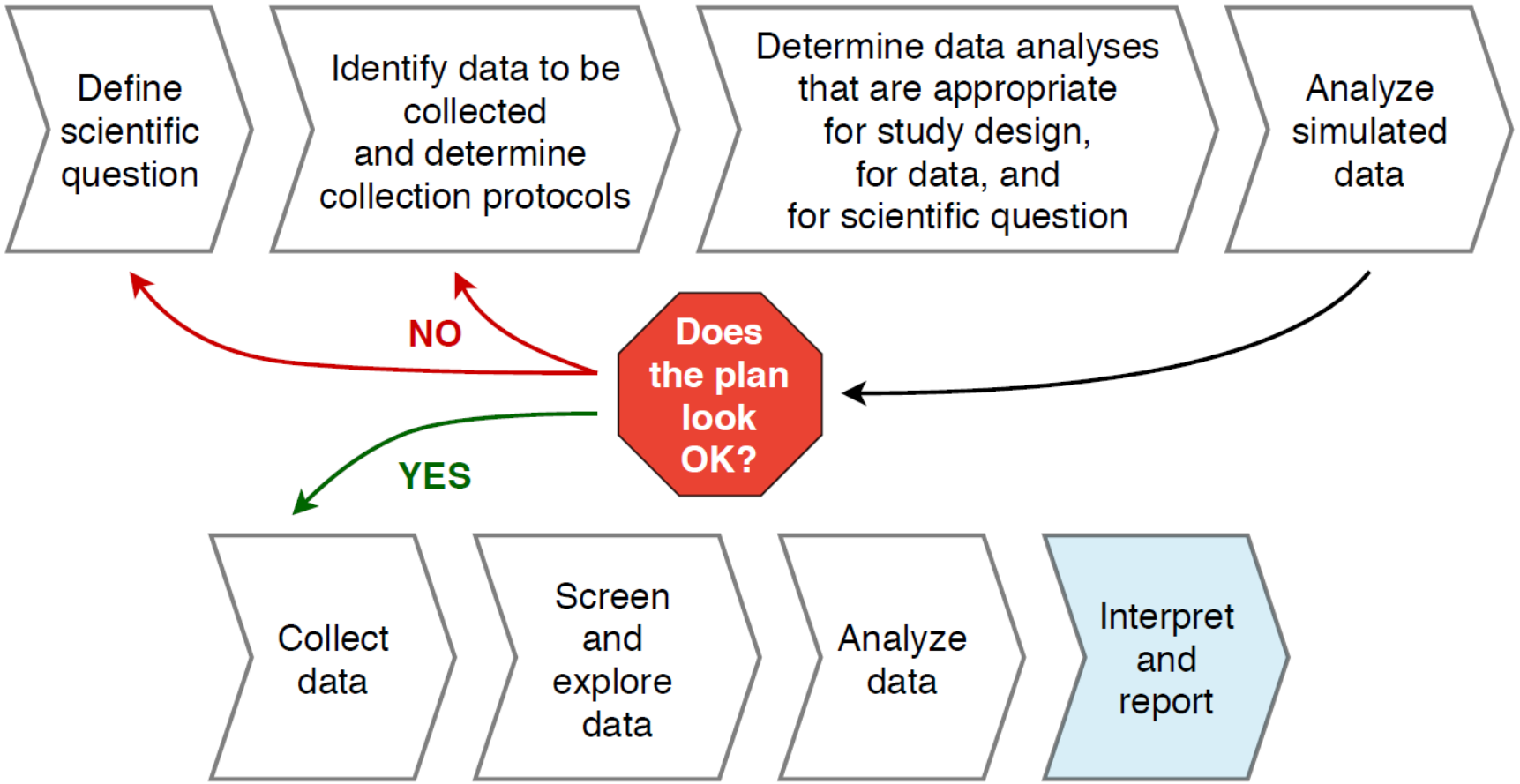
-- Ronald Crosier




Define scientific question
Identify data to be collected and determine collection protocols
Determine data analyses that are appropriate for study design, for data, and for scientific question
Analyze simulated data
NO
Does the plan look OK?
YES
Collect data
Screen and explore data
Analyze data
Interpret and report

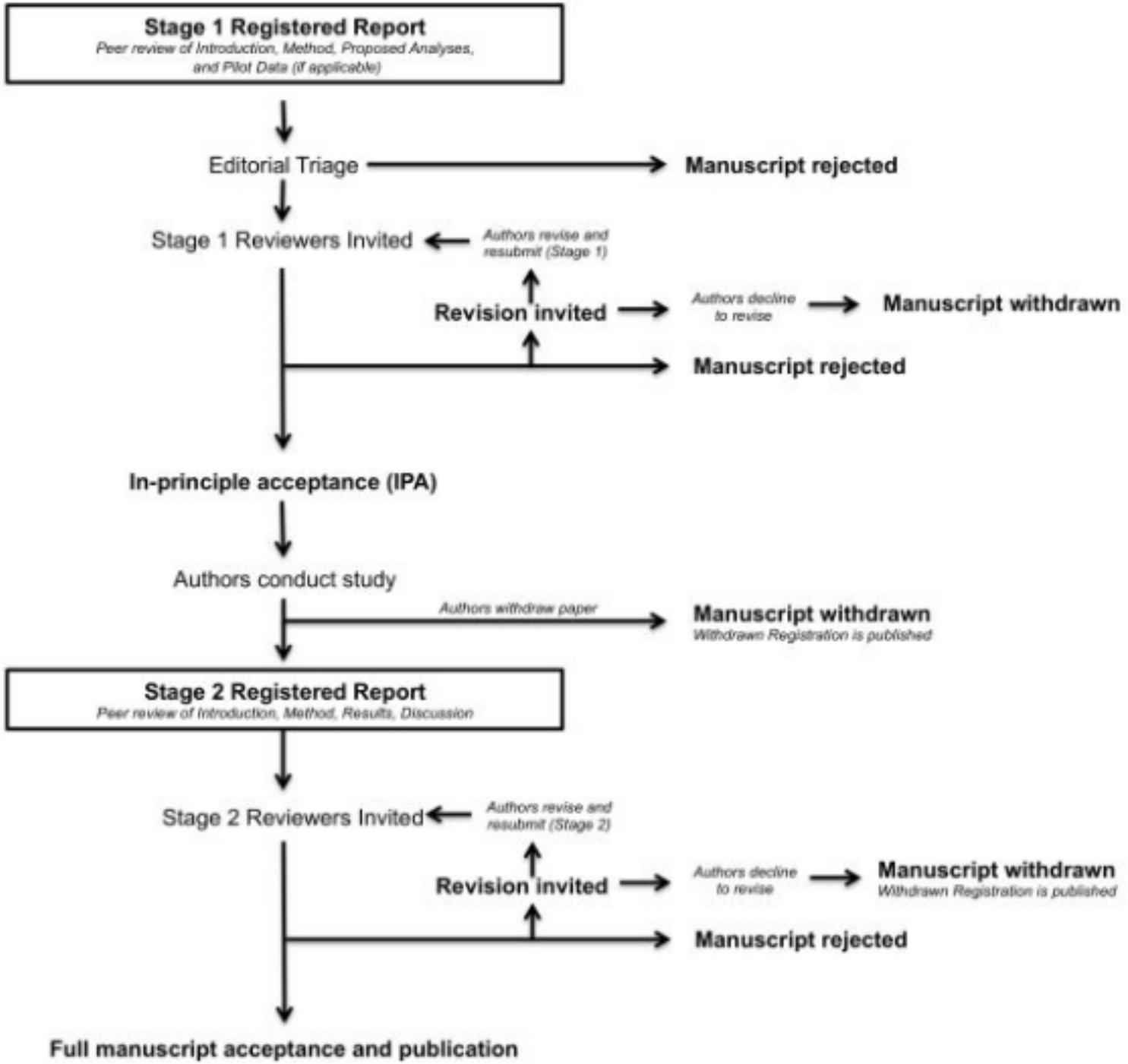

Define scientific question
Identify data to be collected and determine collection protocols
Determine data analyses that are appropriate for study design, for data, and for scientific question
Analyze simulated data
NO
Does the plan look OK?
YES
Collect data
Screen and explore data
Analyze data
Interpret and report


Define scientific question
Identify data to be collected and determine collection protocols
Determine data analyses that are appropriate for study design, for data, and for scientific question
Analyze simulated data
NO
Does the plan look OK?
YES
Collect data
Screen and explore data
Analyze data
Interpret and report


Define scientific question
Identify data to be collected and determine collection protocols
Determine data analyses that are appropriate for study design, for data, and for scientific question
Analyze simulated data
NO
Does the plan look OK?
YES
Collect data
Screen and explore data
Analyze data
Interpret and report

David Bolton: Assistant Professor, Kinesiology & Health Sciences

Editorial pipeline for a Registered Report



CORTEX 117 (2019) 135–146

Available online at www.sciencedirect.com

ScienceDirect

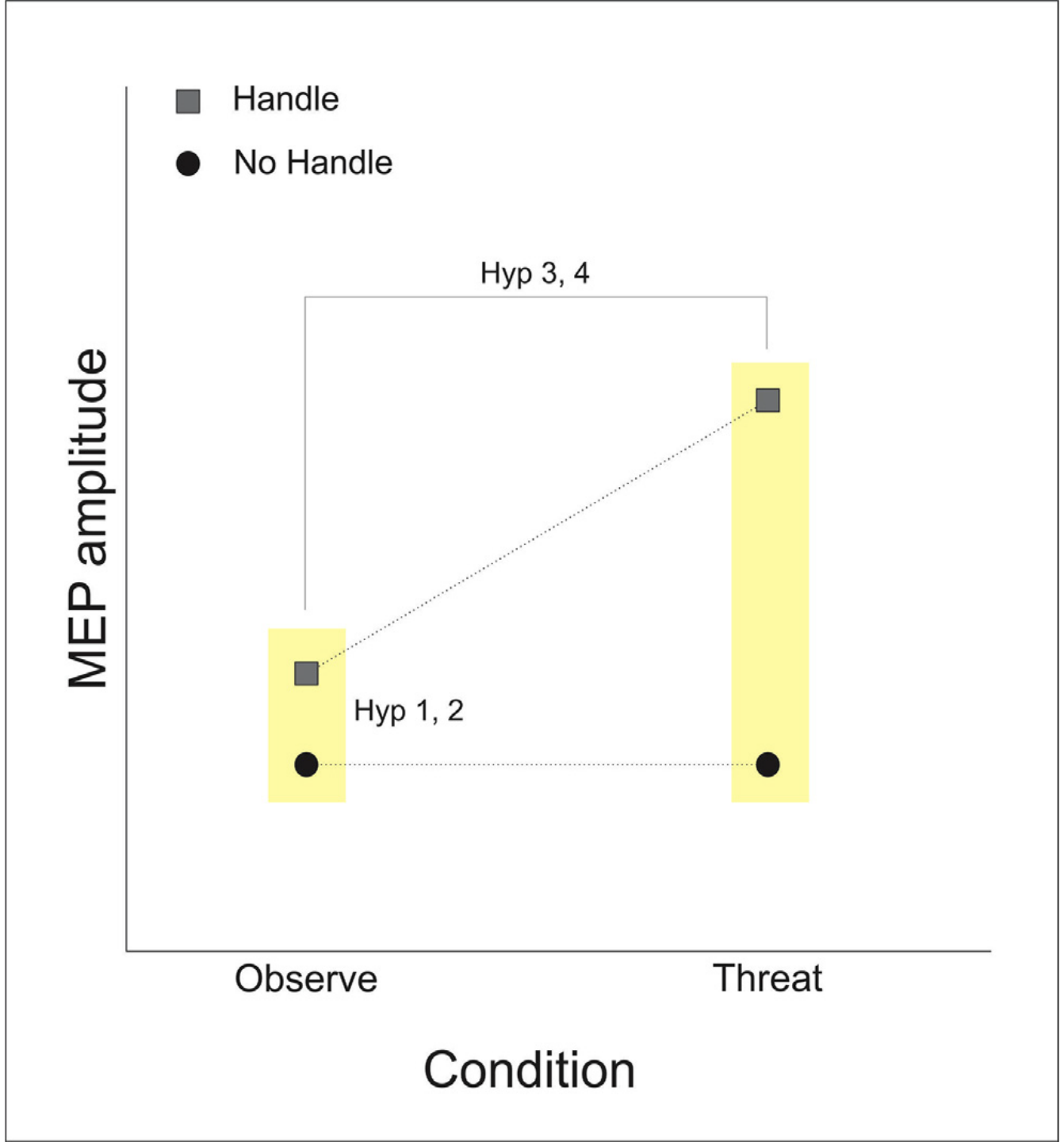
Journal homepage: www.elsevier.com/locate/cortex

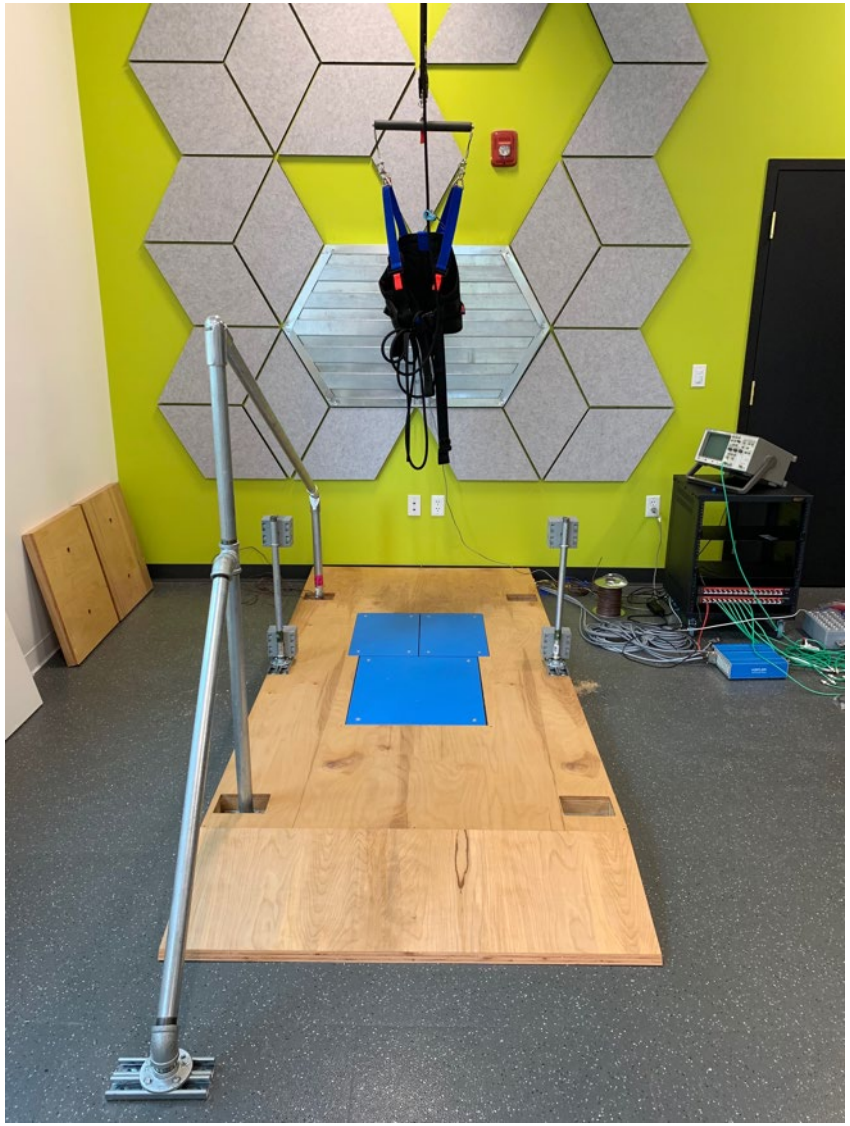
Research Report

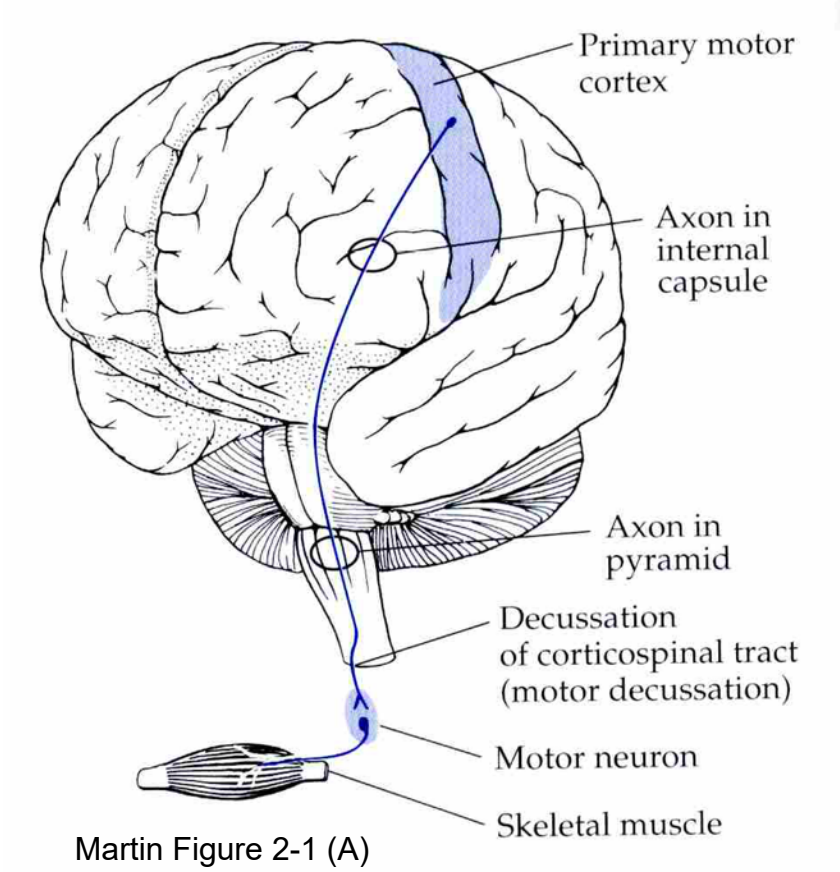
Motor preparation for compensatory reach-to-grasp responses when viewing a wall-mounted safety handle

David A.E. Bolton [a,*], David M. Cole [a,b], Blake Butler [a], Mahmoud Mansour [c], Garrett Rydalch [d], Douglas W. McDannald [a] and Sarah E. Schwartz [b]

[a] Department of Kinesiology & Health Science, Utah State University, United States
[b] Department of Psychology, Utah State University, United States
[c] Department of Electrical & Computer Engineering, Utah State University, United States
[d] Department of Biology, Utah State University, United States

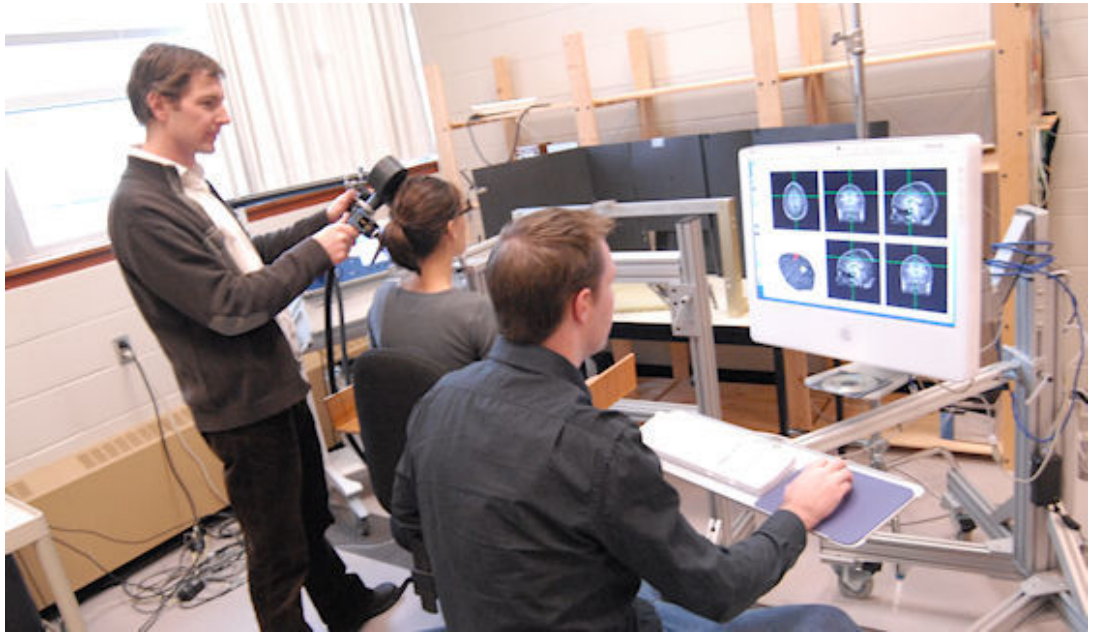

Handle
No Handle
Hyp 3, 4
Hyp 1, 2
MEP amplitude
Observe
Threat
Condition

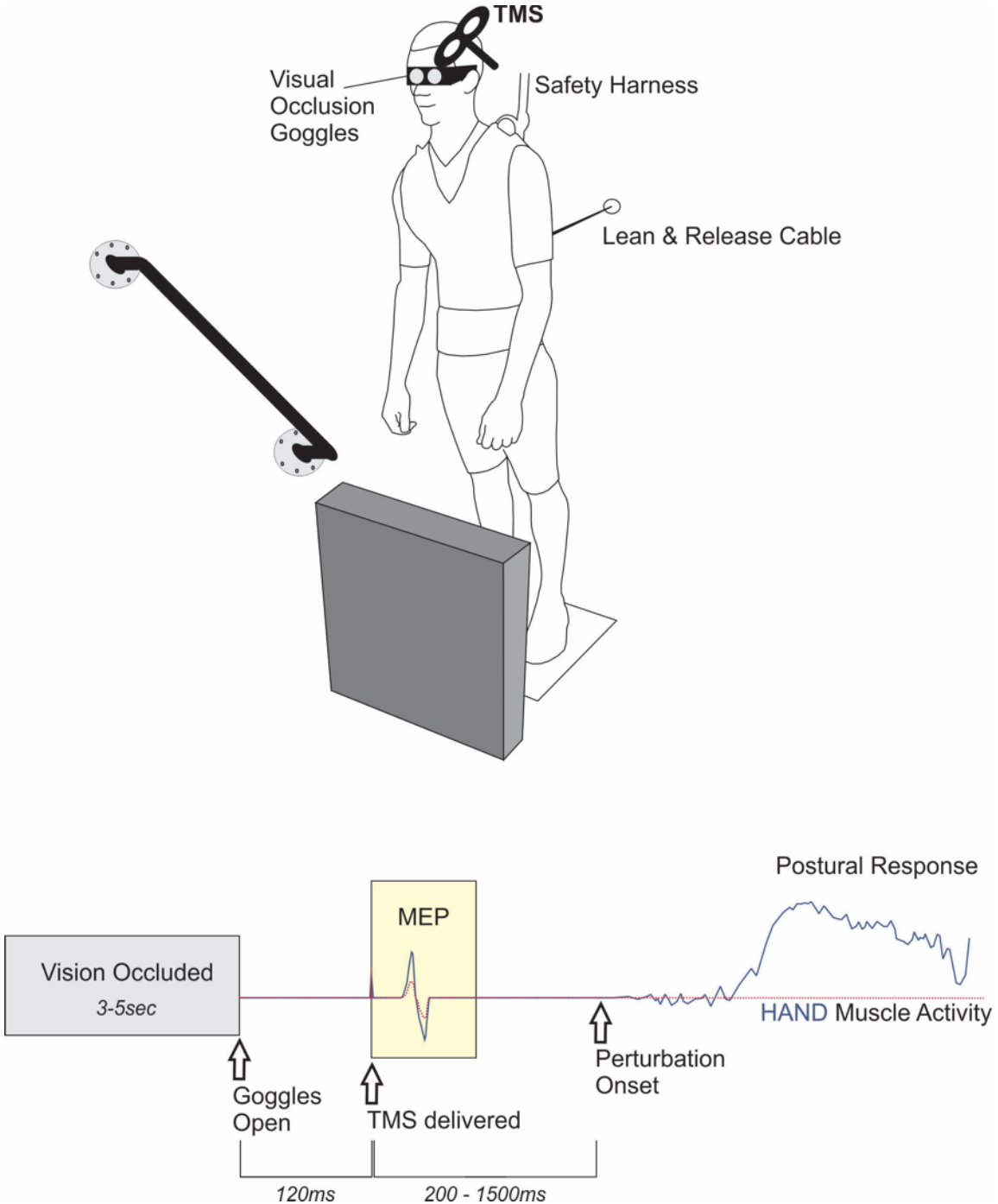


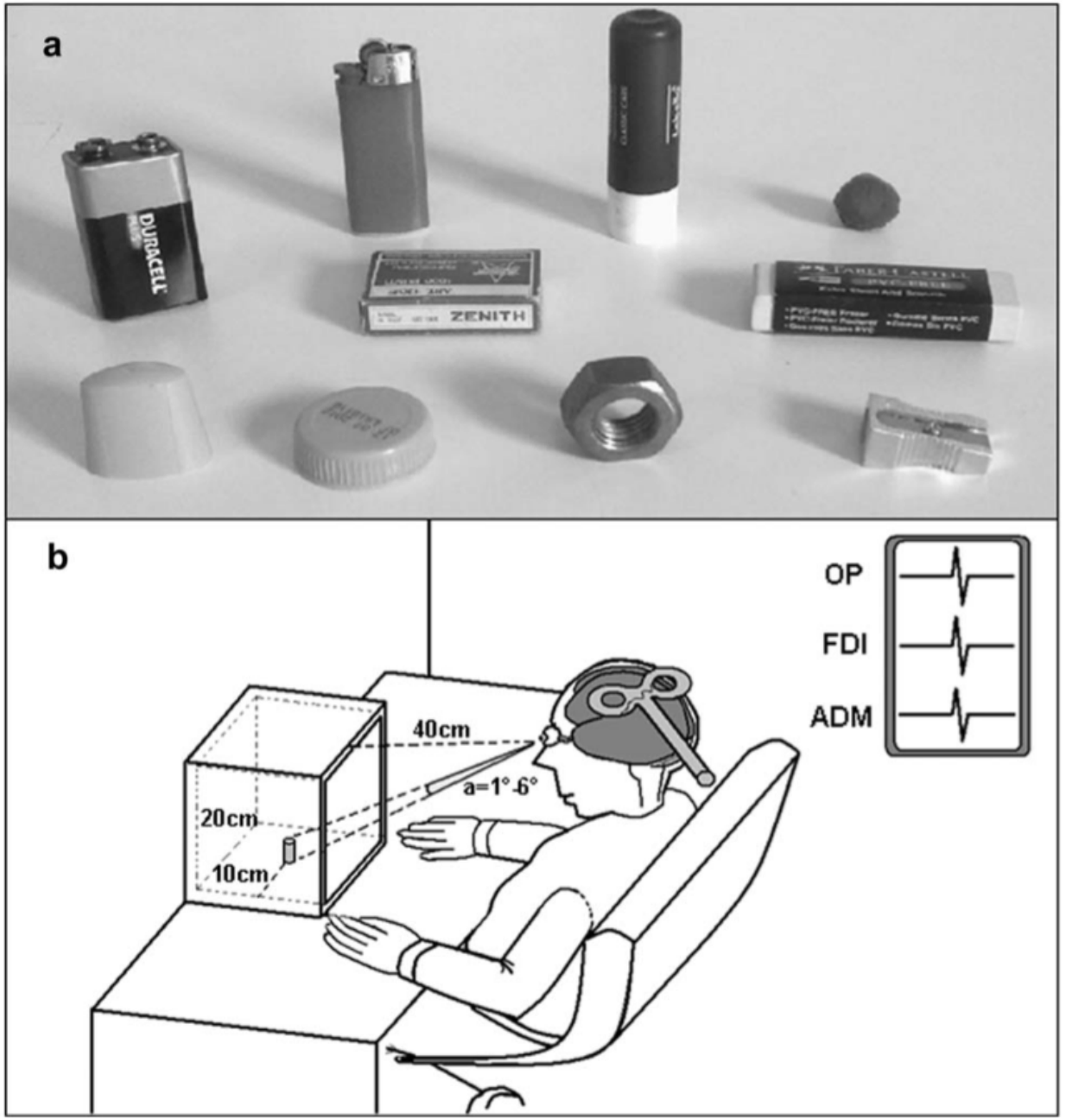


Martin Figure 2-1 (A)

Transcranial Magnetic Stimulation TMS




TMS
Visual
Occlusion
Goggles
Safety Harness
Lean & Release Cable
Postural Response
MEP
Vision Occluded
3-5sec
HAND Muscle Activity
Goggles
Open
TMS delivered
Perturbation
Onset
120ms
200 - 1500ms



Franca et al, 2013. Corticospinal Facilitation during Observation of Graspable Objects: A Transcranial Magnetic Stimulation Study

Datapalooza

Doing it right from the start

Richard Cutler
Department of Mathematics and Statistics

Datapalooza

About me …

- Member of the statistics faculty
- Ran the Statistical Consulting Center for 3 years
  - Worked on over 250 problems from every college at Utah State University
  - Mainly worked with graduate students
- I work on application of statistical methods in many areas but particularly ecology and the environment, and in the design and analysis of experiments
- My main area of expertise is statistical learning (machine learning) methods
  - *Data Science* methodology

Datapalooza

A Cautionary Tale

- 12 soil cores are taken
- Each core split into top (2—5cm), middle (7—10cm), and bottom (12—15cm)
- The 12 portions at each depth are combined and thoroughly mixed
- 36 samples from each depth are made and randomly assigned to each treatment
- At each Depth and for each treatment, for each of 6 different times 3 samples are randomly selected and destructively tested.

Datapalooza

A Cautionary Tale

- This is a sophisticated experimental design
- A great deal of thought was put into designing the study
- A HUGE amount of effort went into implementing the design and collecting data
- The students advisor and committee signed off on it
- But …

Datapalooza

A Cautionary Tale

- There is no replication at the “whole plot” level
- Traditional ANOVA techniques are not valid for some of the analyses of interest
- The analysis did not fully meet the student’s objectives
- Two years later, using some extant data we were able to improve the analysis but we made heroic assumptions in the process and it was still not completely satisfactory
- The problem with the design could have been averted by talking with a statistician knowledgeable in the design and analysis of experiments

Datapalooza

Some thoughts about data analysis

- It can be really hard, much harder than you think
- It's important to know the details, not just generalities
  - Even down to the level of generating "fake" data
- There are usually multiple methods, perhaps multiple approaches, for analyzing a given dataset
- Learn as much as you can about the techniques you are going to use
- Talk to people who are experts in the areas of analysis
  - Including statistical consultants whenever possible

Building reproducibility from the start

David E. Rosenberg

Datapalooza

February 25, 2020

@WaterModeler
david.rosenberg@usu.edu

Utah State University
UTAH WATER RESEARCH LABORATORY

Civil & Environmental ENGINEERING

Drought management with reconstructed monthly paleo flows

Water conservation

Reproducibility is a continuum

1. Data and artifacts available?

- Input data
- Models/Code
- Results
- Instructions
- Other digital artifacts

2. Results reproducible?

- Figures
- Tables
- Other results

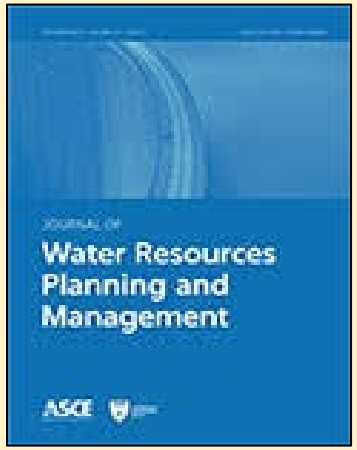
3. Findings replicable?

- With new datasets
- In new locations

(Follows NSF and U.S. Dept. of Education definitions)

Push work up the continuum!!

How can I make my results more reproducible?



David E. Rosenberg
Yves Filion
Rebecca Teasley
Samuel Sandoval-Solis
Jory S. Hecht
Jakobus E. van Zyl
George F. McMahon
Jeffrey S. Horsburgh
Joseph R. Kasprzyk
David G. Tarboton

1. **Build reproducibility into the project from start – budget time, money, storage, IRB, and tools**
2. **Put all materials in a repository**
3. **Make all inputs to and outputs of proprietary, private, & computationally intensive steps available**
4. **Ask someone to verify your results are reproducible**
5. **Train students and employees in reproducible practices**

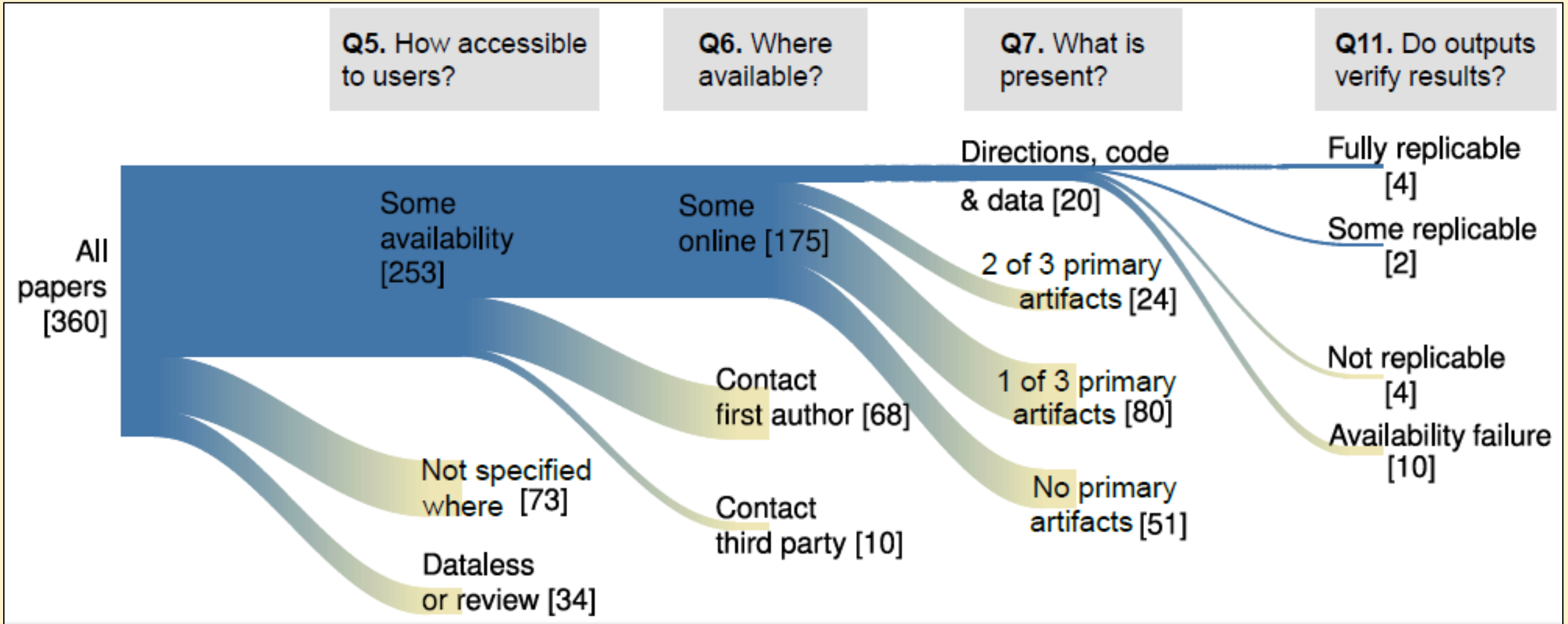
+ 42 other tips

ASCE
AMERICAN SOCIETY OF CIVIL ENGINEERS

You are not alone

Low availability and reproducibility in 360 water papers in 2017

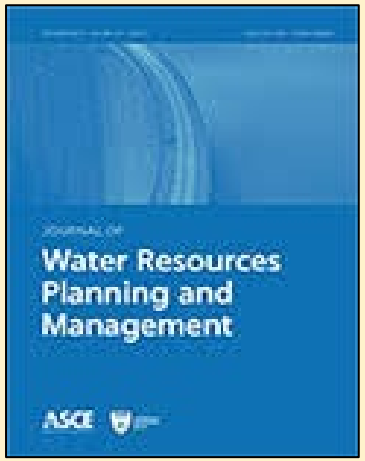
(Stagge et al, 2019 in Nature-*Scientific Data*)



Environmental Modeling & Software	Water Resources Research	J. American Water Resources Association
Hydrology and Earth System Sciences	Journal of Hydrology	J. Water Resources Planning & Management

We must create a culture where we …

Authors



Journals

Institutions

Funders

… reproduce results

What prevents me to make my results more reproducible?

1. More author effort
2. Must learn new skills
3. Fear being scooped
4. Can not share proprietary data
5. Nor share high performance computing, big data, or methods with long run times
6. Takes time and expertise to reproduce others' results.
7. We value journal articles
8. Unintentionally encourage researchers to pursue easily reproduced methods

Why make my results more reproducible?

1. Increase access, use, and extend work
2. Improve trust
3. Enable benchmark studies
4. Organize materials in perpetuity
5. Reduce effort to respond to email requests
6. Learn by doing
7. Narrow gap between academics and professionals
8. **Funders often require you to!!**

Authors, journals, funders, institutions must create a culture where we reproduce results

How can I make proprietary data and model results more reproducible?

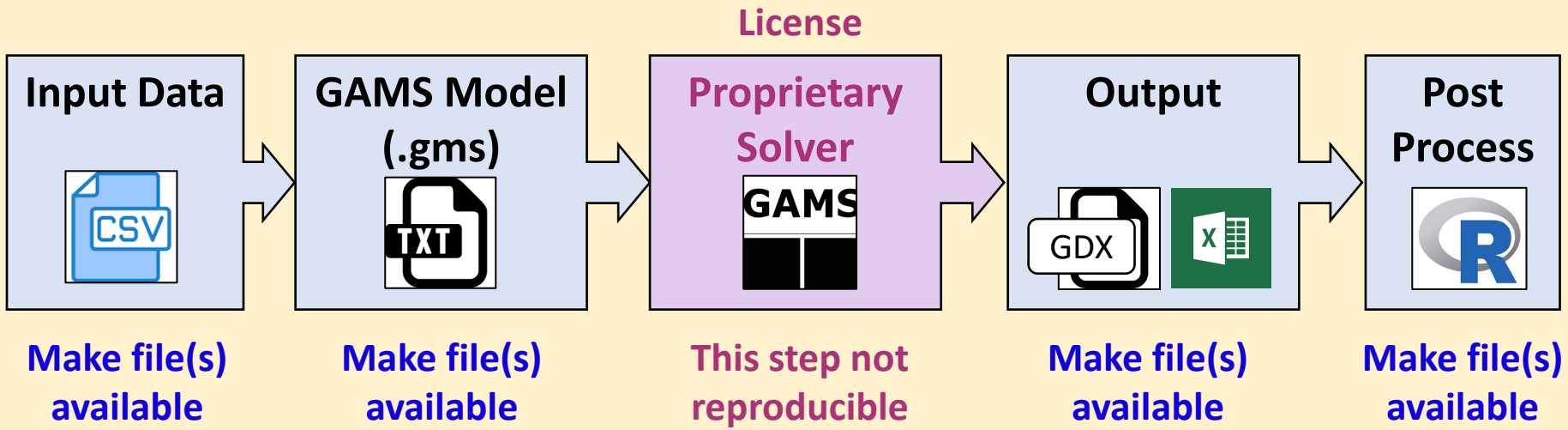


Figure 2. Make a proprietary work flow more reproducible. Example with General Algebraic Modeling System (GAMS) optimization model

What is next?

- Go to https://tinyurl.com/WhatNext2Reproduce
- Pick a project. Then answer:
  1. What data, models, code, and other electronic materials will the project generate?
  2. Current plan to make materials available to others?
  3. How to make materials more available and reproducible?

Additional Resources

- Rosenberg et al (in press). "The Next Frontier: Making Research More Reproducible." *Journal of Water Resources Planning and Management*.
- Stagge et al (2019). "Assessing data availability and research reproducibility in hydrology and water resources." *Nature-Scientific Data*, 6, 190030.
- Reproducibility survey tool (2019)