Reproducible and Collaborative Research

Liz Bageant

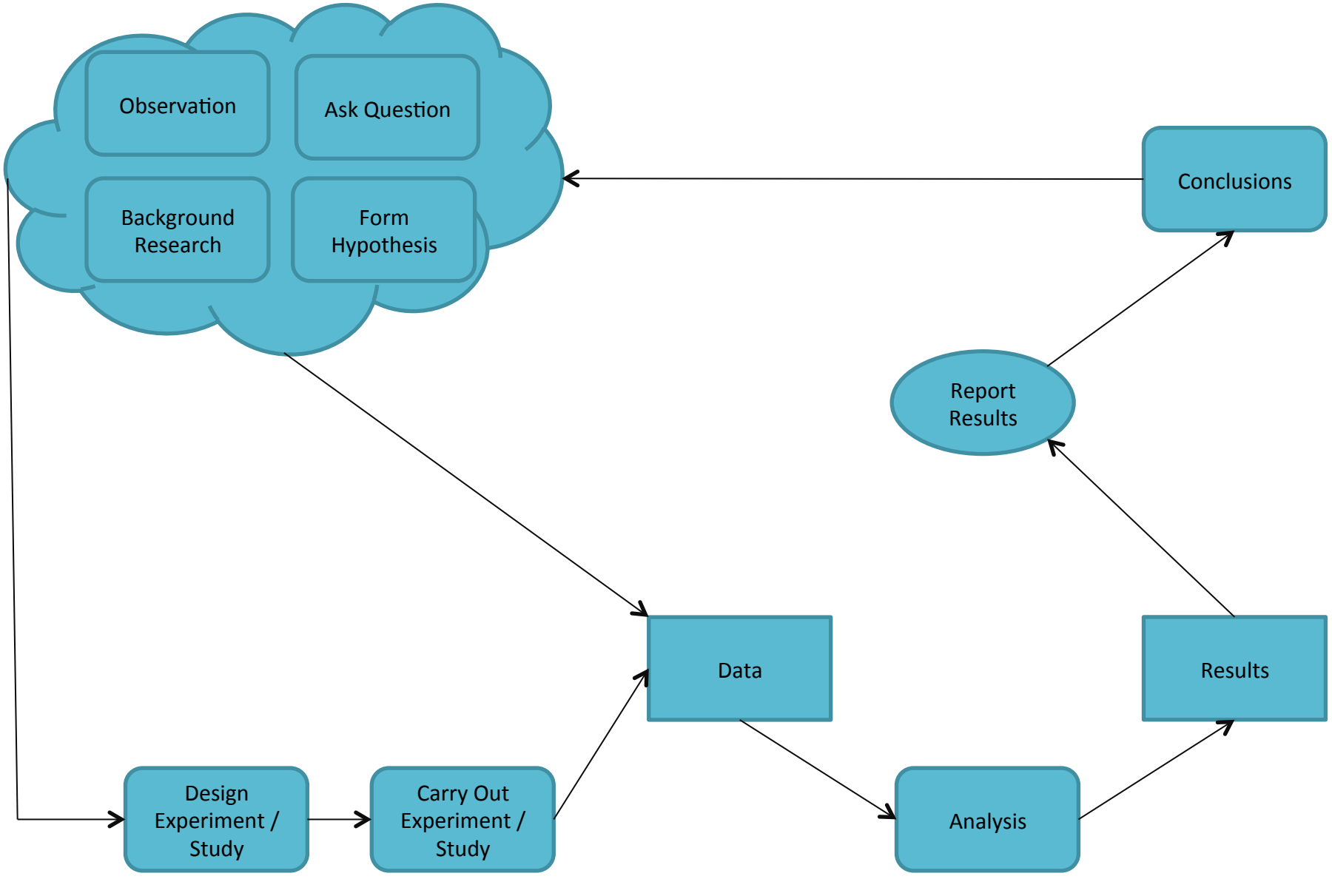
erb32@cornell.edu

Cornell University

Outline

1. Scientific method and research failures
2. Defining reproducible research
3. Strategies for reproducibility

The scientific method



Schematic courtesy of Erika Mudrak

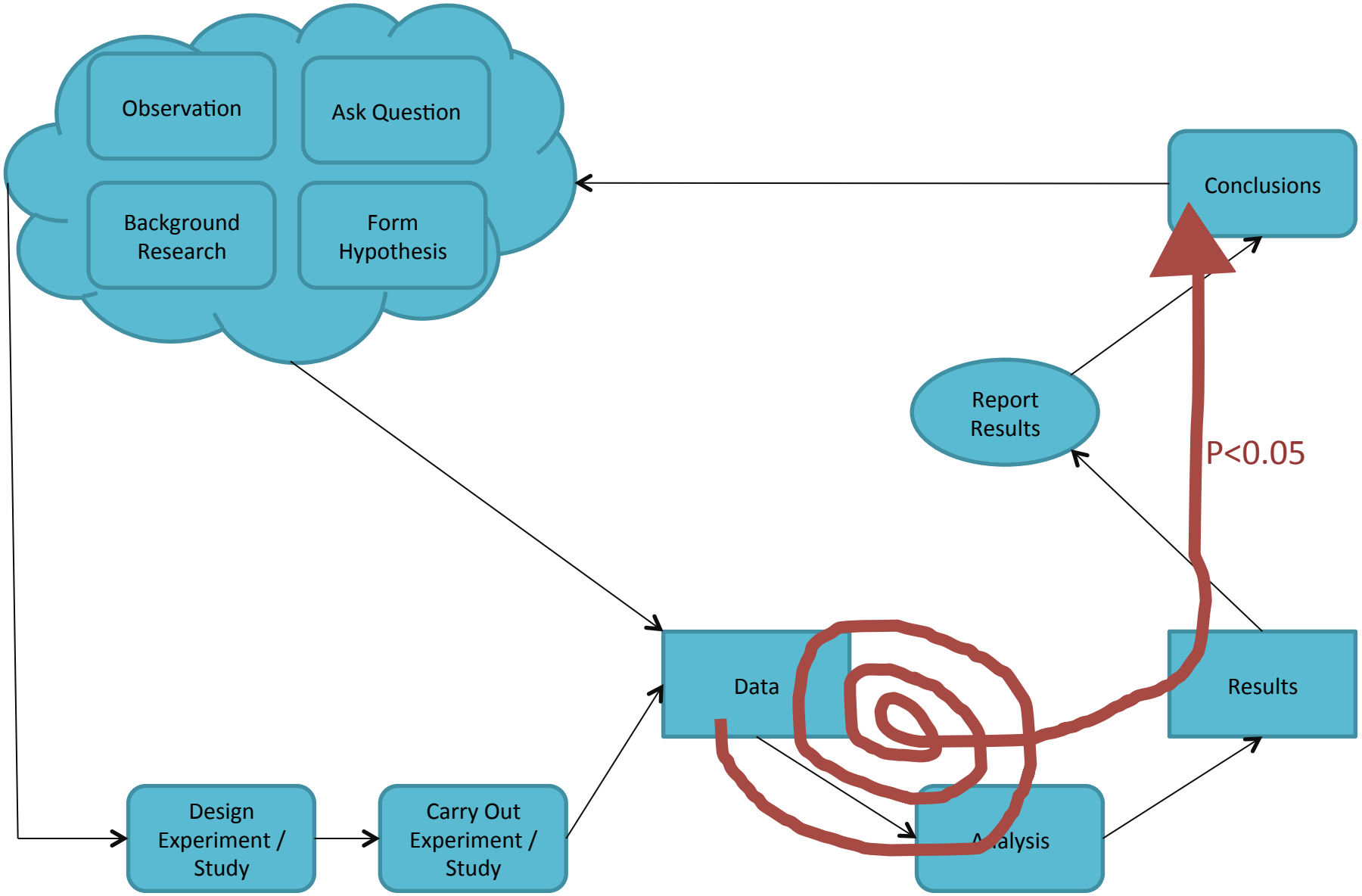
Continuum of research failure

Failure of process **Failure of integrity**

Disorganization **Egregious behavior**

Deliberate manipulation of data to get results
-P-hacking
-"Fishing expeditions"

P-hacking / fishing expedition



Continuum of research failure

Failure of process **Failure of integrity**

Disorganization **Egregious behavior**

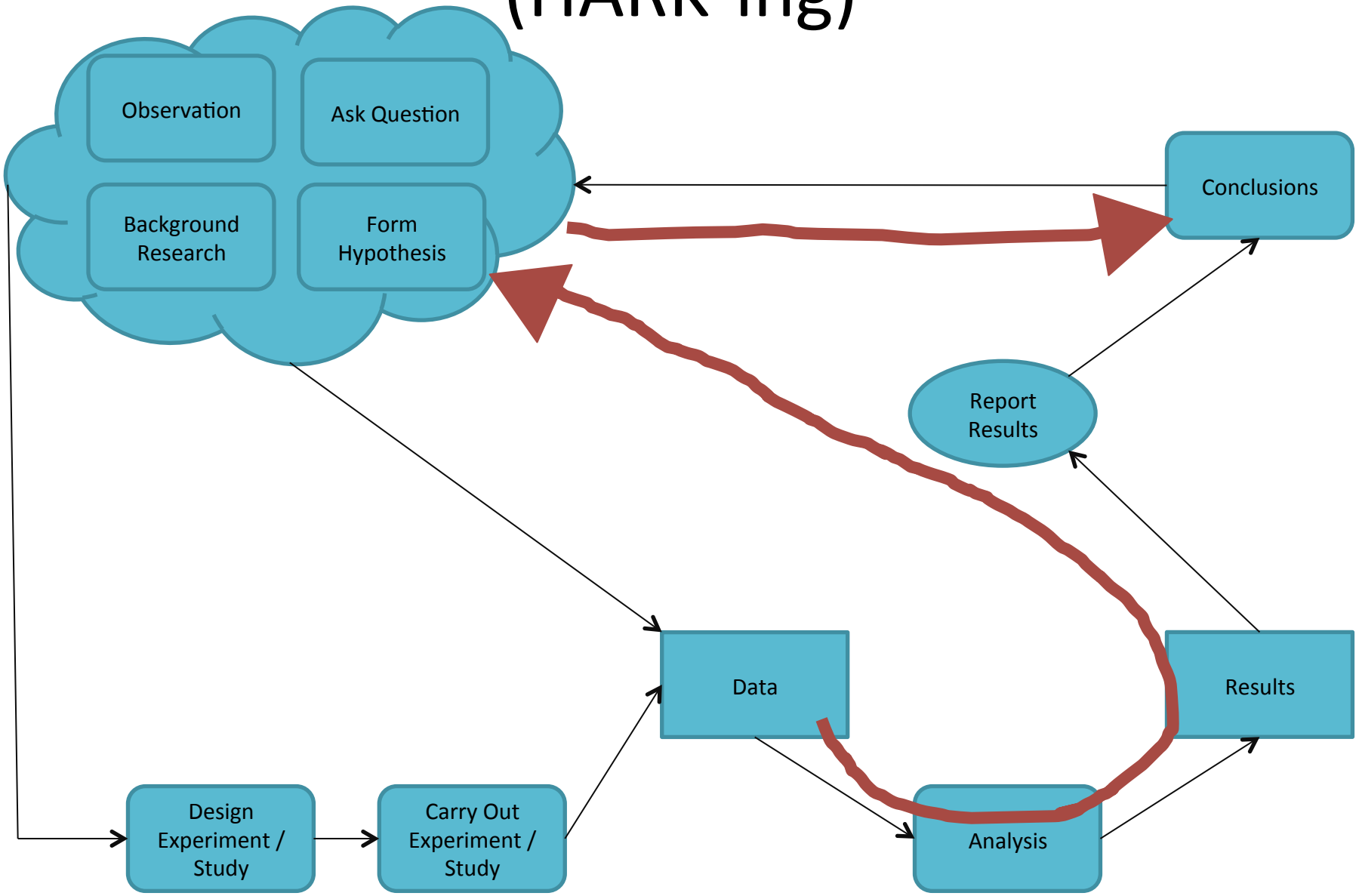
HARK-ing

Deliberate manipulation of data to get results
-P-hacking
-"Fishing expeditions"

P-hack your way to scientific glory

https://projects.fivethirtyeight.com/p-hacking/

Hypothesizing **A**fter **R**esults are **K**nown (HARK-ing)



Is HARK-ing ever okay?

- Exploratory research = hypothesis generation
- Confirmatory research = hypothesis testing

Continuum of research failure

Failure of process **Failure of integrity**

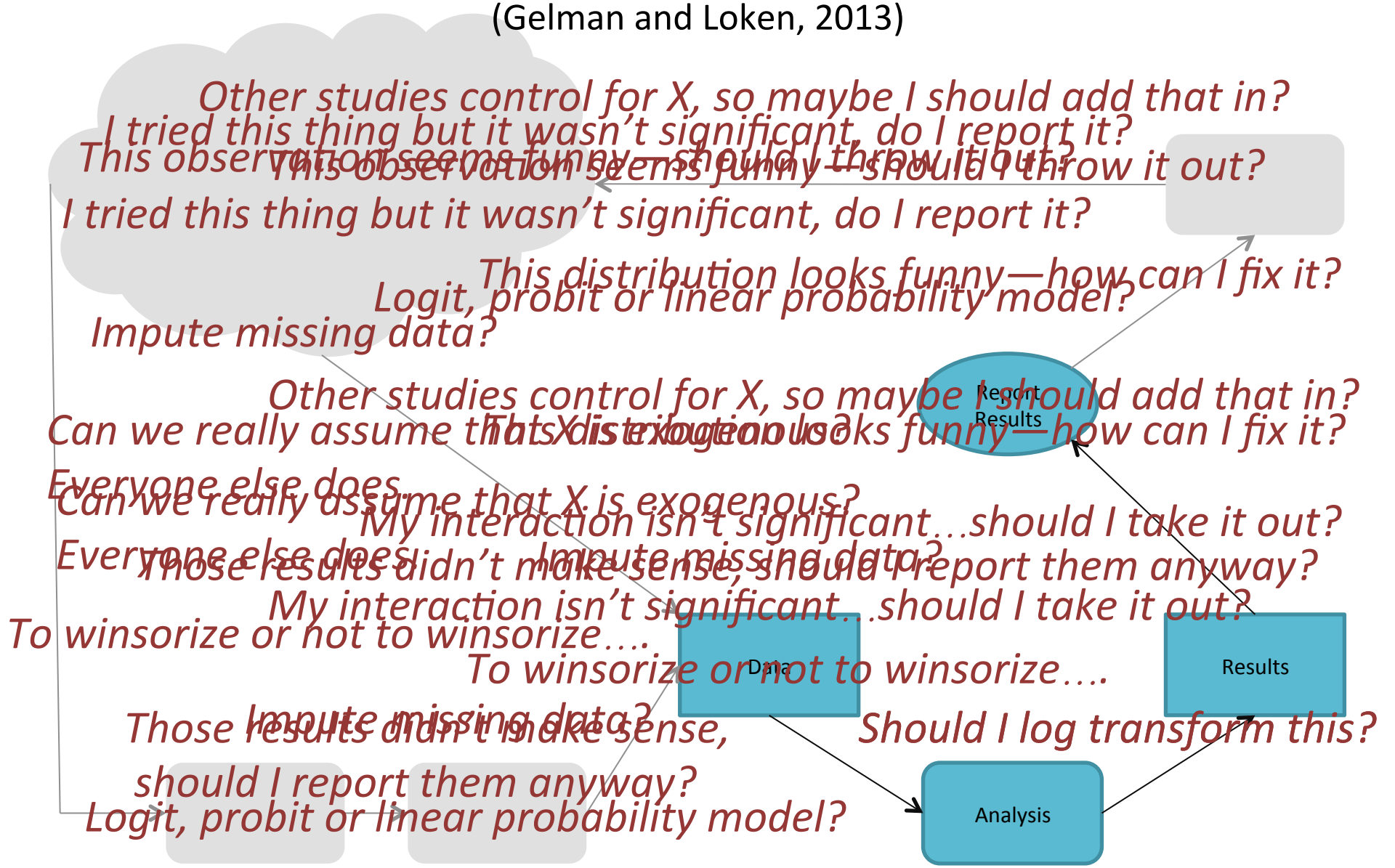
Disorganization **Egregious behavior**

HARK-ing
"Garden of forking paths"

Deliberate manipulation of data to get results
-P-hacking
-"Fishing expeditions"

The garden of forking paths

(Gelman and Loken, 2013)



Continuum of research failure

Failure of process **Failure of integrity**

Disorganization **Egregious behavior**

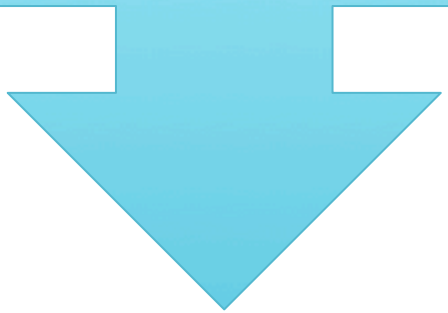
Coding errors
Poor documentation

HARK-ing
"Garden of forking paths"

Deliberate manipulation of data to get results
-P-hacking
-"Fishing expeditions"

To avoid the perils of the garden, HARK-ing, P-hacking, and silly mistakes…

- Integrity! --> Be honest with yourself.
- Transparency! --> Be honest with your readers.
- Do you feel good enough about your decision-making processes to write them down for all to see?

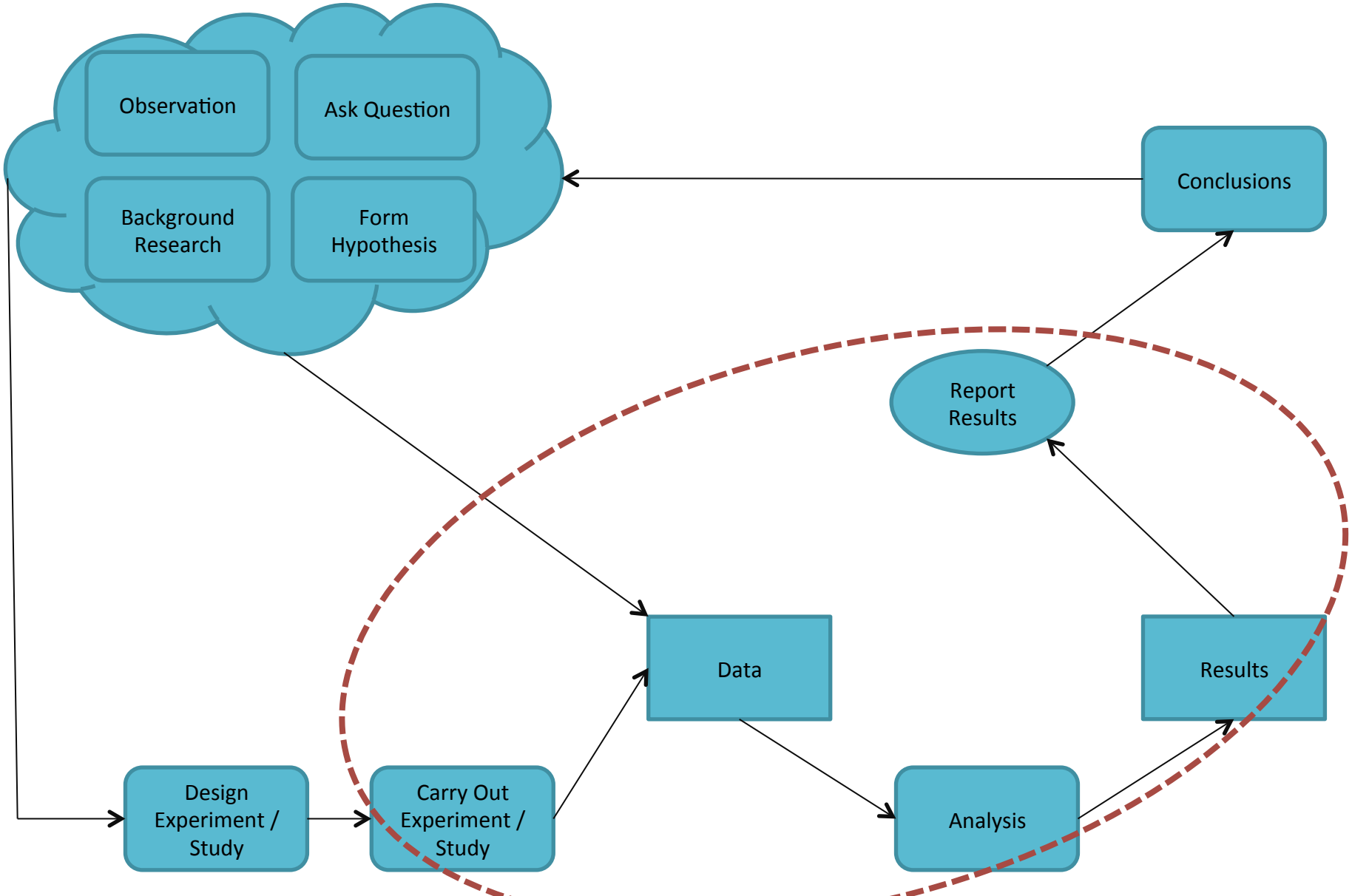


Reproducible research!

Replicability vs reproducibility

- Replicability
  - Essential to the scientific method
  - repeating a study from scratch using new data, analyst and code
  - if a given relationship between X and Y is true, it should show up in multiple studies

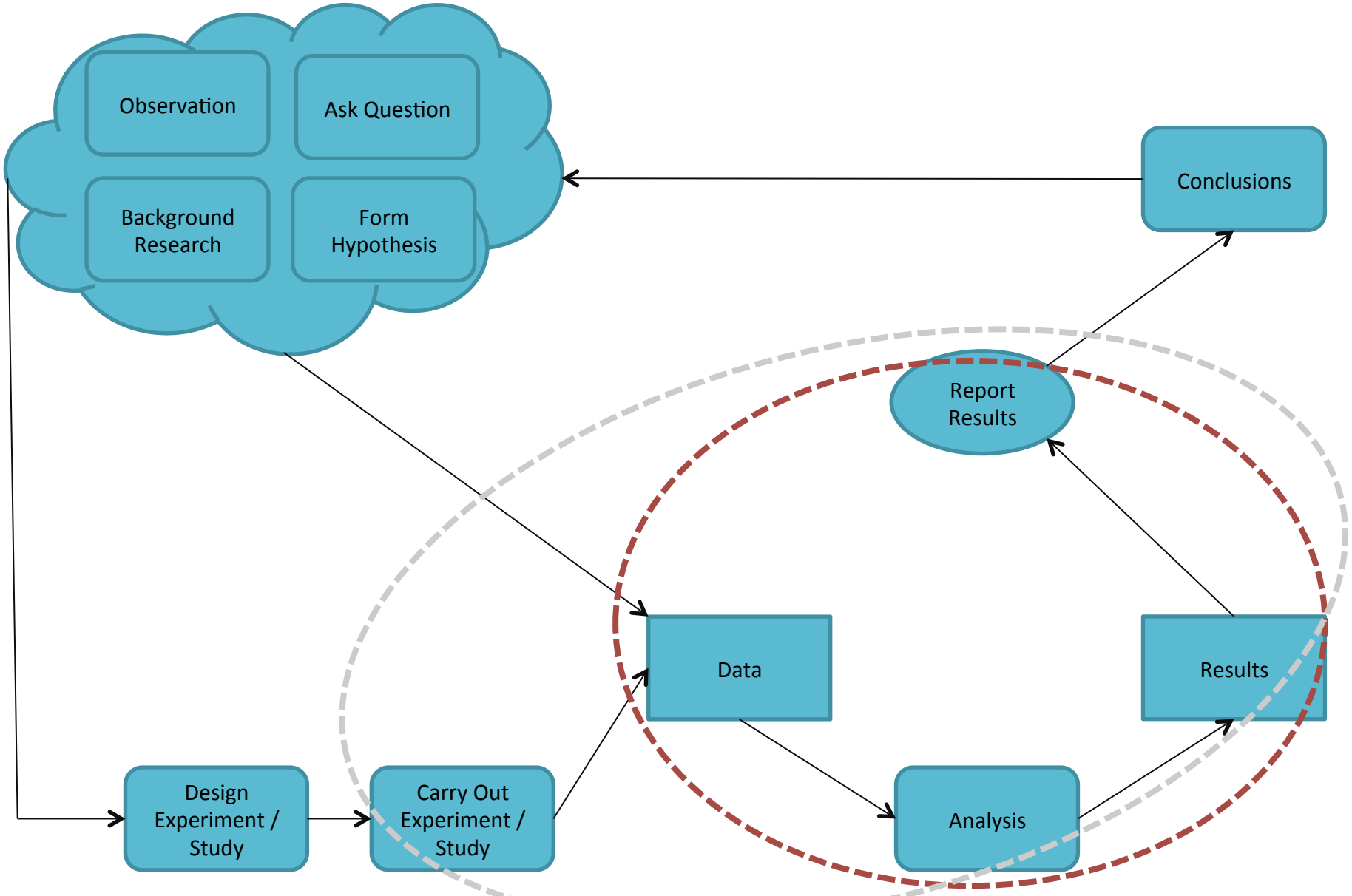
Replicability



Replicability vs reproducibility

- Reproducibility
  - Getting the exact same result as an existing study using new analyst, but same data and code
  - Recently tractable due to computing and software advances

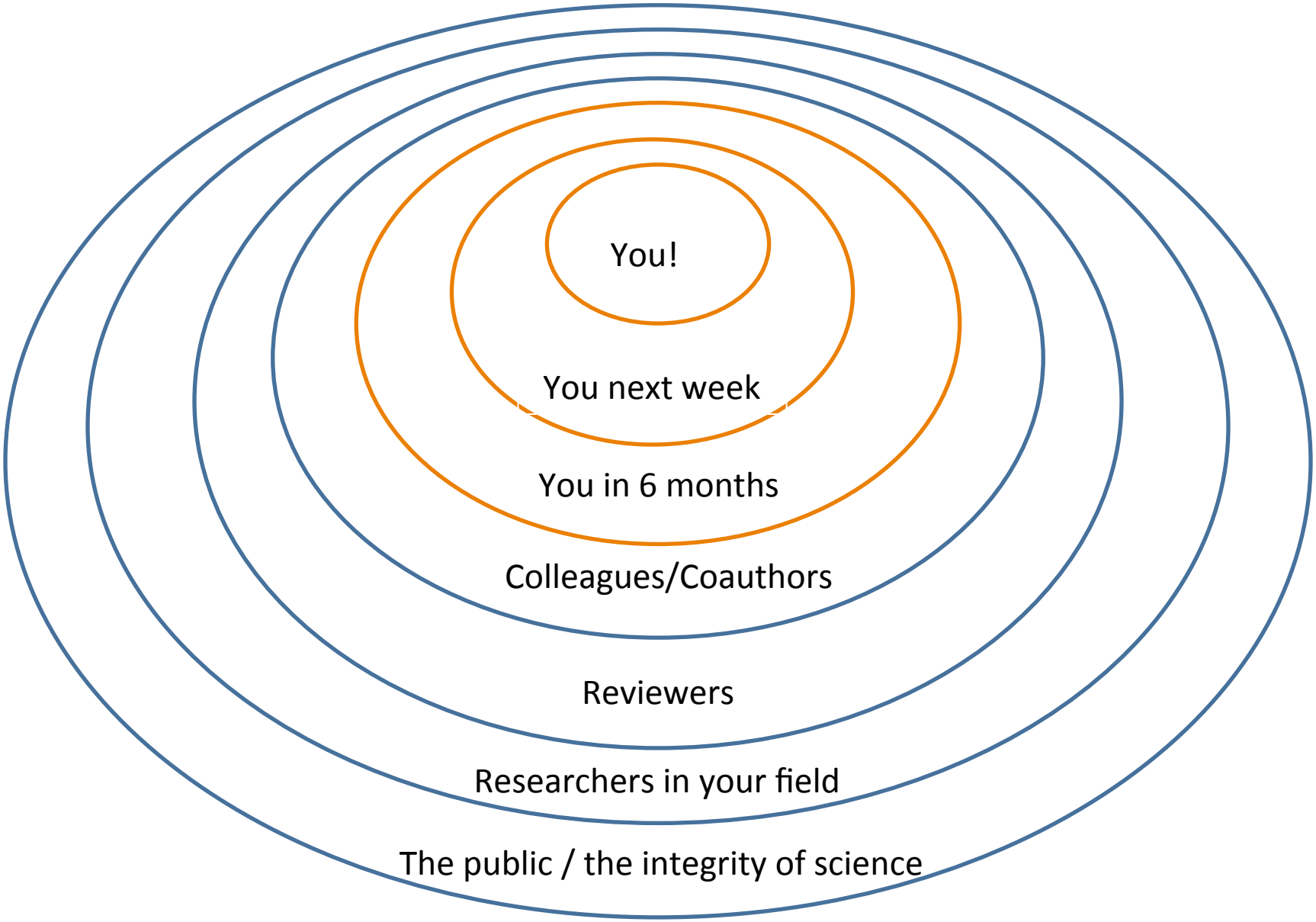
Reproducibility



Reproducibility

- Facilitate transparency by communicating procedures easily
- Identify inadvertent errors
- Avoid embarrassment
- Facilitate collaboration
- Save time
- Greater potential for extension of work --> higher impact over time

Who are you accountable to?



What are we aiming for?

- Sufficient documentation to bring an unfamiliar user up to speed
  - Codebook
  - Readme file
  - Variable and value labels in analysis data set
  - Effective comments in code
- A single click executes your project from start to finish.
  - Downloading
  - Reformatting
  - Cleaning and variable construction
  - Analysis
  - Output tables, graphs, figures
  - Reproducible report

How do we get there?

- Separate the phases of data work
- Systematic file and naming structures
- Effective and organized scripting
- Reproducible reports

Separate phases of data work

1. Data conversion/cleaning/variable construction
2. Analysis
3. Report generation

Naming conventions

- Agree with your collaborators on naming conventions.
- Human readable
  - Short, useful names
  - Information on content
- Machine readable
  - Avoid special characters, spaces, etc
    - CamelCase, ALLCAPS, lowercase, alloneword, underscore_between
  - Consistent naming to facilitate searching
- Default ordering
  - Date format YYYYMMDD
  - Other numbers—add leading zeros
- Never call something “final”. It probably isn’t.

Systematic file structure

- Must be common to all users!
- Choose a file structure and stick to it.
- Make skeleton of folders when you start a project.

- **/dta**
  - /original → Copy of read-only original files exactly as obtained.
  - /stata raw → Data after conversion to format of choice
  - /clean → Variable- or module-specific clean files
  - /analysis → Data set(s) you will use for analysis
- **/documentation**
  - /metadata → Any/all codebooks or metadata related to data
  - /reports → Collection of documents where the data was used, cited, described
- **/do**
  - /cleaning → Cleaning, merging, reshaping, variable construction scripts
  - /analysis → Analysis scripts
  - master.do → Script that sets up relative file paths and calls all scripts
- **/output**
  - /figures → Subfolders depend on type of project
  - /tables
  - /old output → Keep for reference, if you choose.
- **/writing**
  - / paper 1 → Separate folders if multiple papers using the same data
  - / paper 2
  - /notes → Optional as needed
  - /old drafts → Keep older versions of paper, but get them out of the way
- **/temp** → Get rid of clutter as you make it

Scripting tips

- Data + Script = Reproducible Output
- Master script: Runs other scripts in correct order
- Modular scripting vs. one big file
  - Separate types of processes (cleaning, analysis)
  - Avoid repeating blocks of code: Separate program for repeated processes
- Notes/comments.
  - Consistent headers
  - Useful comments, not expressions of feeling
- Clarity > efficiency? Consider your collaborators.
- Re-run script from the beginning regularly. It must run!

Reproducible Reports

- Integrate code into the prose of your report
- Single file that executes all steps of data process and outputs a final paper
- Know exactly what data was used for analysis, what code made which figure, etc.
- Disadvantages—learning curve, initial investment.
- Alternative method: Copy and paste.

Avoid research failures by implementing reproducible research techniques to improve organization and transparency

1. Separate phases of research
2. Systematic file naming and structure
3. Effective and organized scripting
4. Reproducible reports

- Prioritize elements that are attainable for you.

Your future self thanks you!

Additional resources

- P-hack your way to scientific glory! https://projects.fivethirtyeight.com/p-hacking/
- Gelman and Loken (2013) Garden of Forking Paths. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf