Research Data Reproducibility

Karen Dobos, PhD
Research Integrity and Compliance Review Office
Colorado State University

Acknowledgements

Tobin Magle, PhD
Bioinformationist
Health Science Library
University of Colorado Anschutz Medical Campus

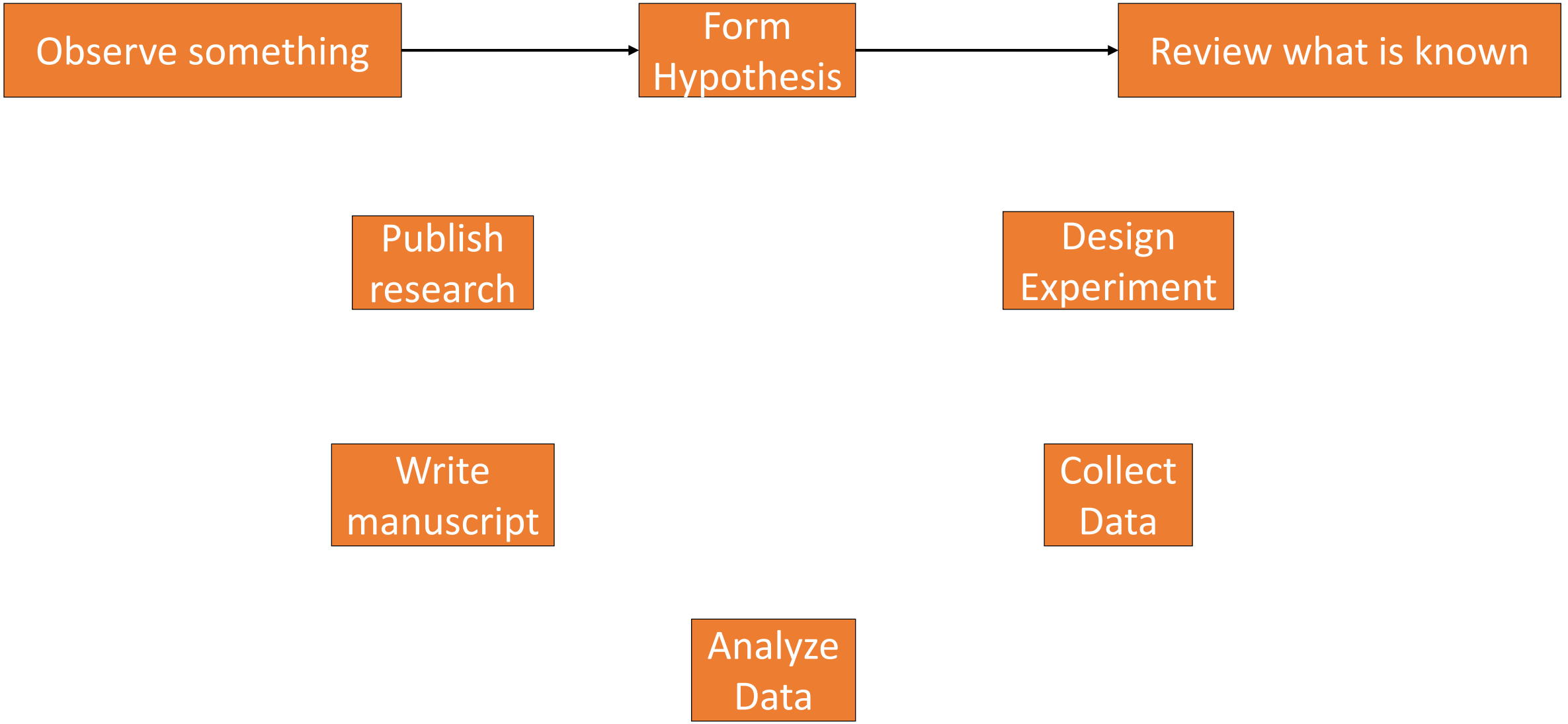
Cat Bens, MS
QA Manager
Research Integrity and Compliance Review Office
Colorado State University

Disclaimer

Unless otherwise stated, the information provided in this talk was gathered from other persons and sources; references are cited in the lower right corner of each slide. I have no conflict of interest with the information, data, and/or resources cited in this presentation.

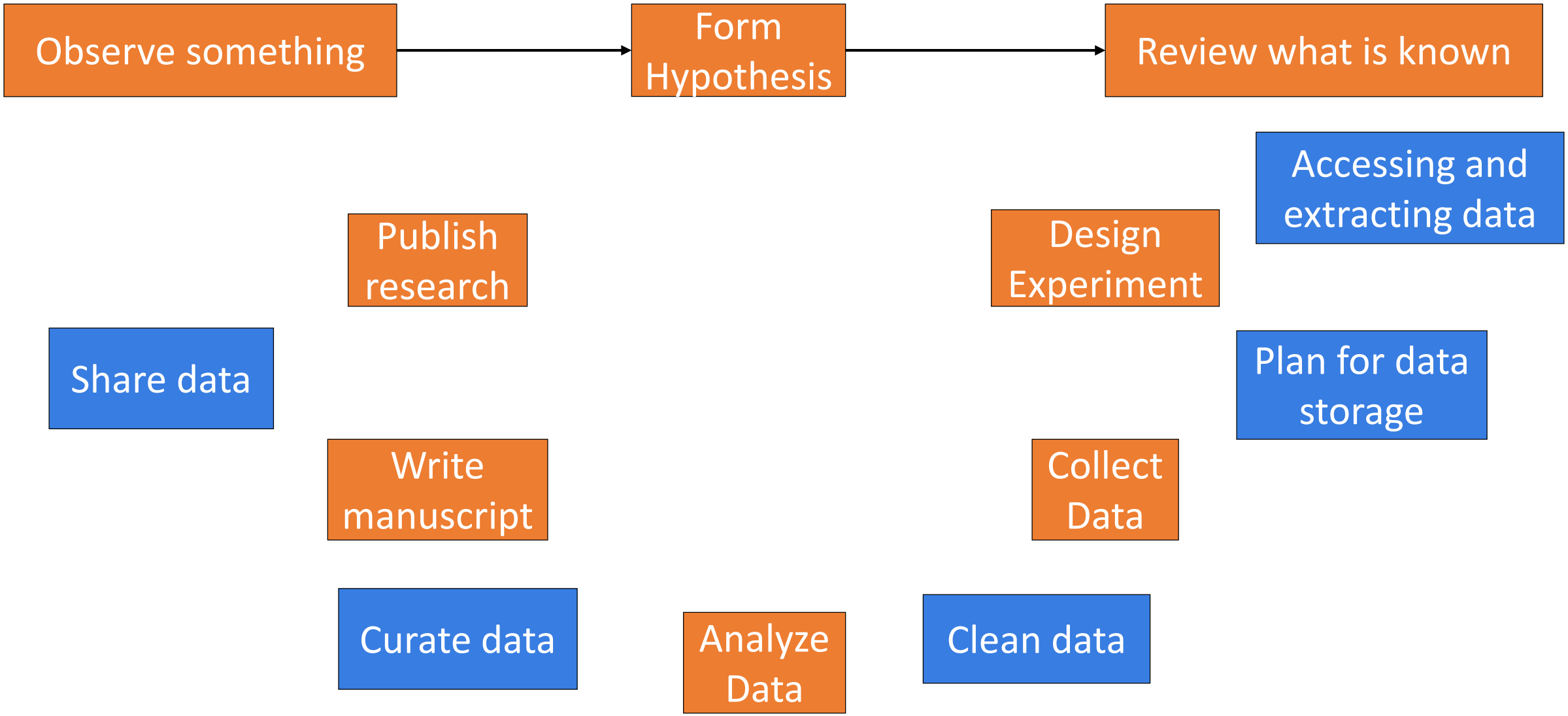
The opinions expressed in this talk are mine.

The Research Lifecycle as taught:



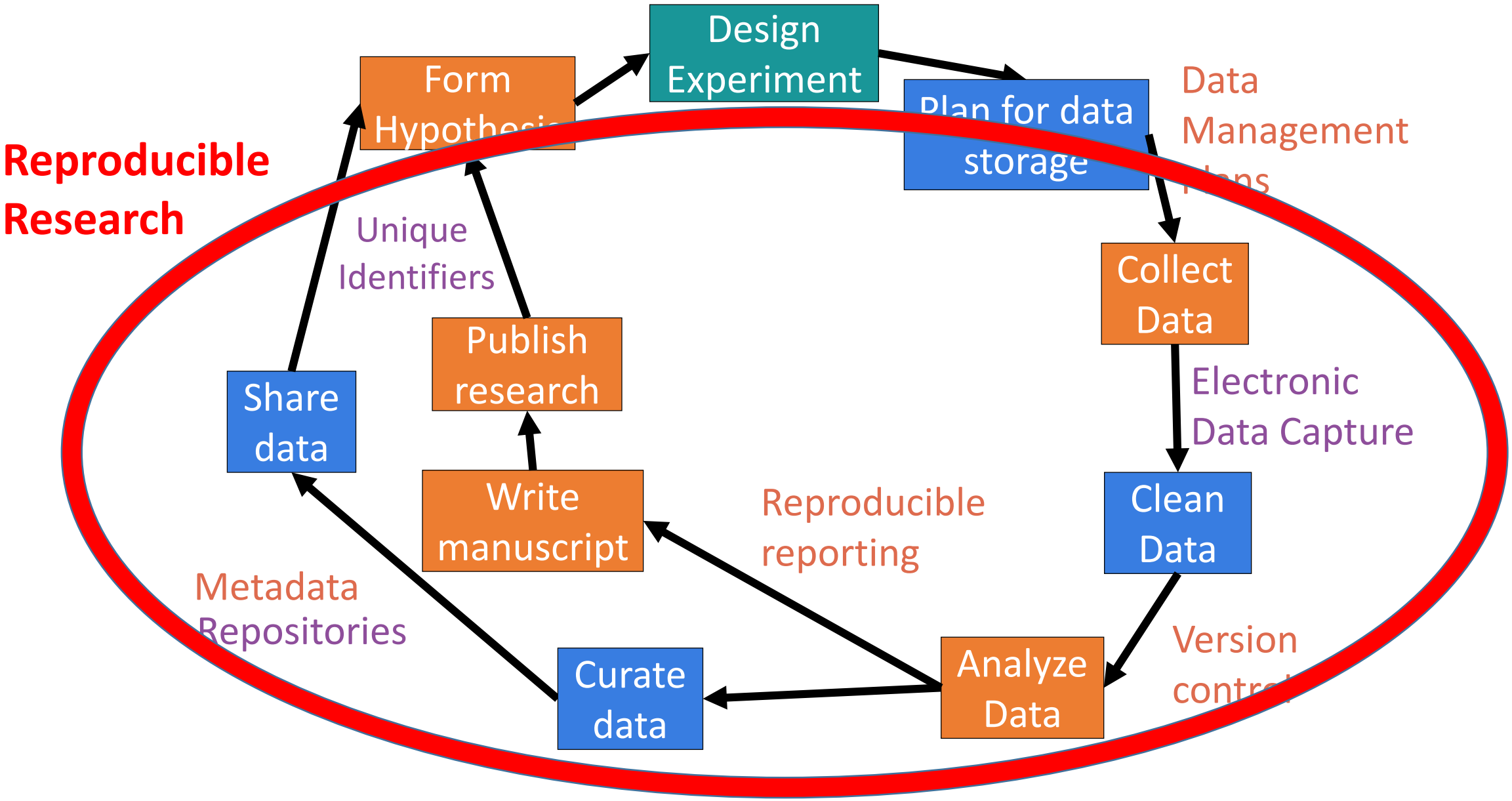
Adapted from: Tobin Magle, PhD, “Reproducible Research Theory”

The Research Lifecycle as taught:



Adapted from: Tobin Magle, PhD, “Reproducible Research Theory”

Requires new **expertise** and **infrastructure**



Reproducibility

is the practice of distributing all data, software source code, and tools required to reproduce the results discussed in a research publication.

https://www.ctspedia.org/do/view/CTSpedia/ReproducibleResearchStandards

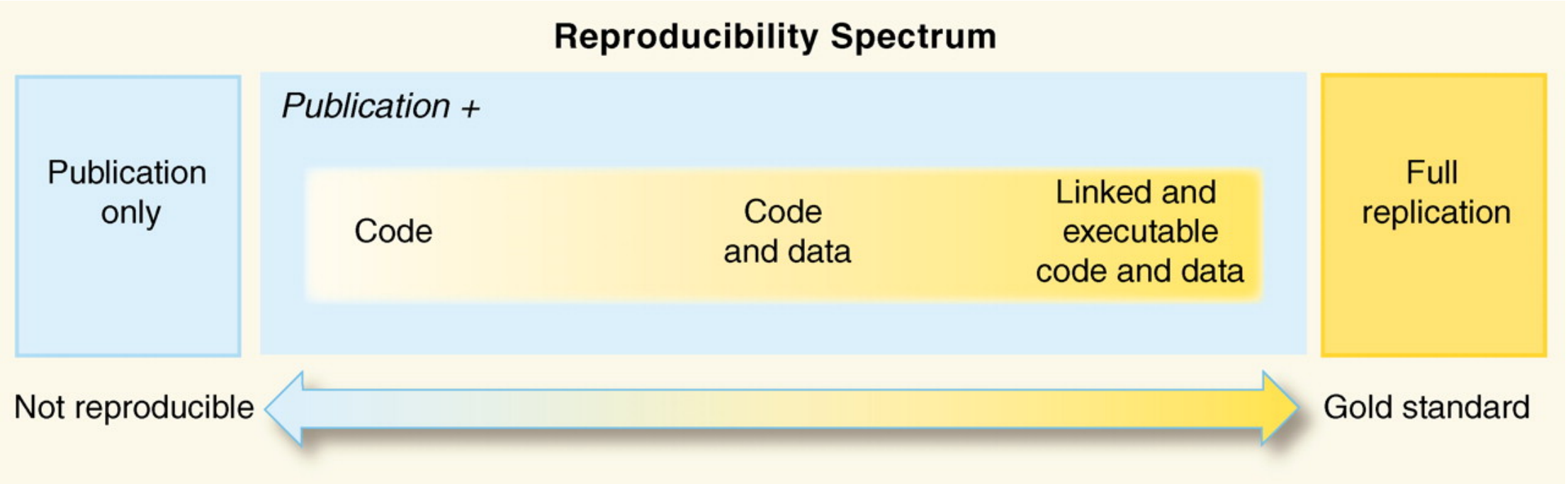
Replication vs. Reproducibility

- **Replication**: The confirmation of results and conclusions from one study obtained independently in another is considered the scientific gold standard.
- "Again, and Again, and Again …" **BR Jasny et. al.** Science, 2011. 334(6060) pp. 1225 DOI: 10.1126/science.334.6060.1225

- **Some studies can't be replicated**: too big, too costly, too time consuming, one time event, rare samples

- **Reproducibility**: minimum standard for assessing the value of scientific claims, particularly when full independent replication of a study is not feasible
- "Reproducible Research in Computational Science". **RD Peng** Science, 2011. 334 (6060) pp. 1226-1227 DOI: 10.1126/science.1213847

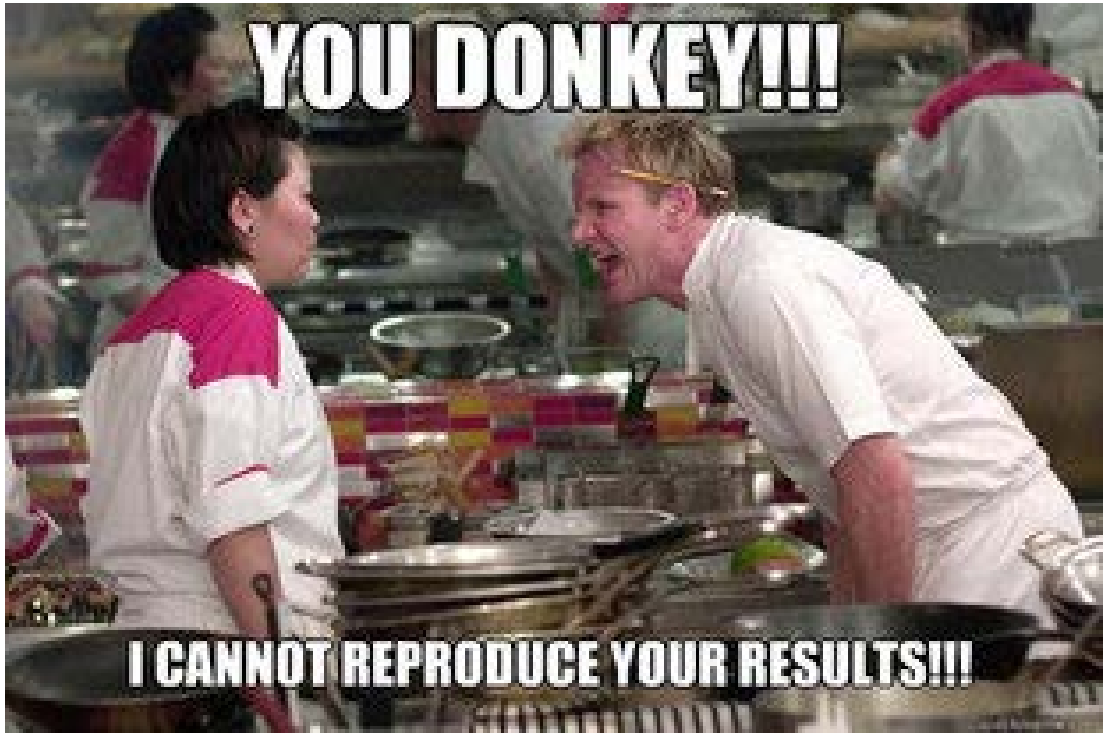
Reproducibility spectrum



"Reproducible Research in Computational Science". **RD Peng** Science, 2011. 334 (6060) pp. 1226-1227 DOI: 10.1126/science.1213847

Why do reproducible research?

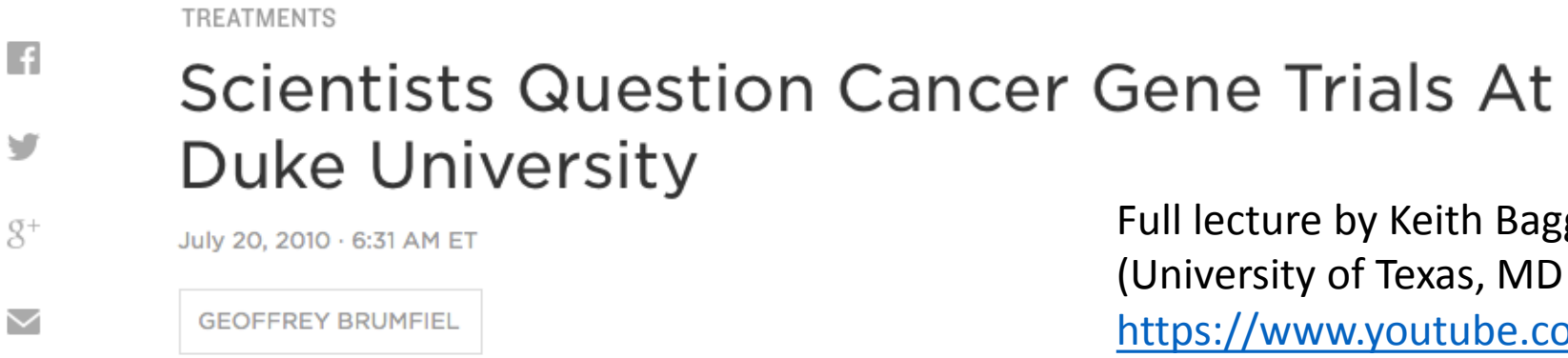
- **Public Good**: transparency
- **Good for YOU:** You are the future user of your data
- Starting to be **recommended** (Journal of Biostatistics, FAIR Principles)
- Will probably be **required** soon



http://campus.murraystate.edu/academic/faculty/cmecklin/RWebpage.html

Public good

- Current lack of transparency:
  - Only 1 of 441 papers provided a “full protocol”, 0 had data
  - Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JPA (2016) Reproducible Research Practices and Transparency across the Biomedical Literature. PLoS Biol 14(1): e1002333. doi:10.1371/journal.pbio.1002333
- Stopped potentially harmful clinical trials



Full lecture by Keith Baggerly, Bioinformatician (University of Texas, MD Anderson Cancer Center)
https://www.youtube.com/watch?v=7gYIs7uYbMo

http://www.npr.org/sections/health-shots/2010/07/19/128630549/scientists-question-cancer-gene-trials-at-duke-university

Good for you:

- You are the future user of your data
  - Version control – revert to older versions
  - Save time when writing your methods sections
  - The you of 2 years ago is bad at answering emails

Starting to be recommended

SCIENTIFIC DATA | COMMENT OPEN

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier […] Barend Mons

Affiliations | Contributions | Corresponding author

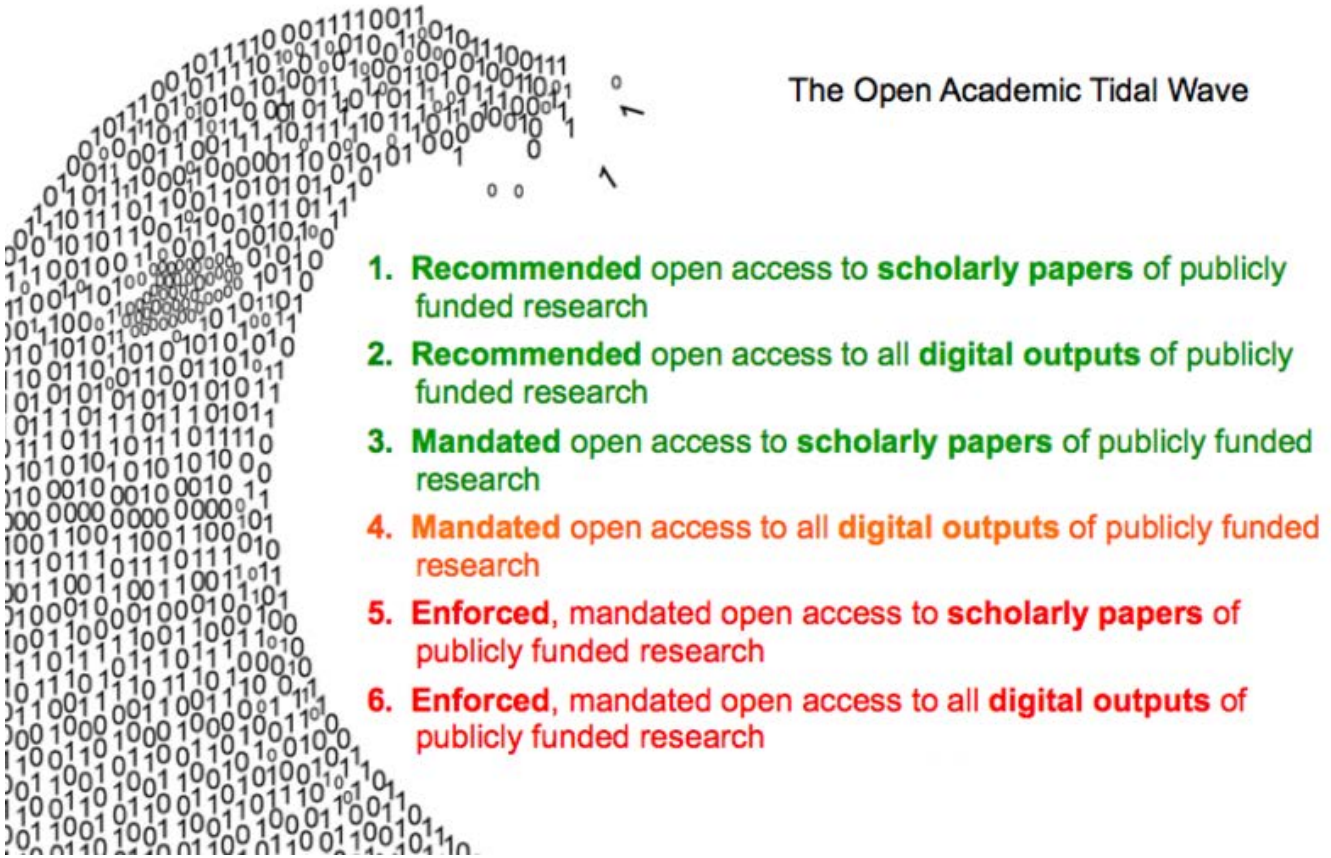
Scientific Data **3**, Article number: 160018 (2016) | doi:10.1038/sdata.2016.18
Received 10 December 2015 | Accepted 12 February 2016 | Published online 15 March 2016

Editorial

Biostatistics (2009), **10**, 3, *pp.* 405–408
doi:10.1093/biostatistics/kxp014

As coeditors of *Biostatistics*, we wish to encourage the practice of making research published in the journal reproducible by others. The following invited piece by Roger Peng sets out our policy on this; Roger will be assuming the role of Associate Editor for reproducibility as set out in his piece.

Will probably be required soon



http://figshare.com/blog/2015_The_year_of_open_data_mandates/143

Whitehouse's 2013 OSTP

"The Obama Administration is committed to the proposition that **citizens deserve easy access to the results of research their tax dollars** have paid for. That's why, in a policy memorandum released today, OSTP Director John Holdren has directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the **results of federally funded research freely available to the public**—generally within one year of publication."

http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

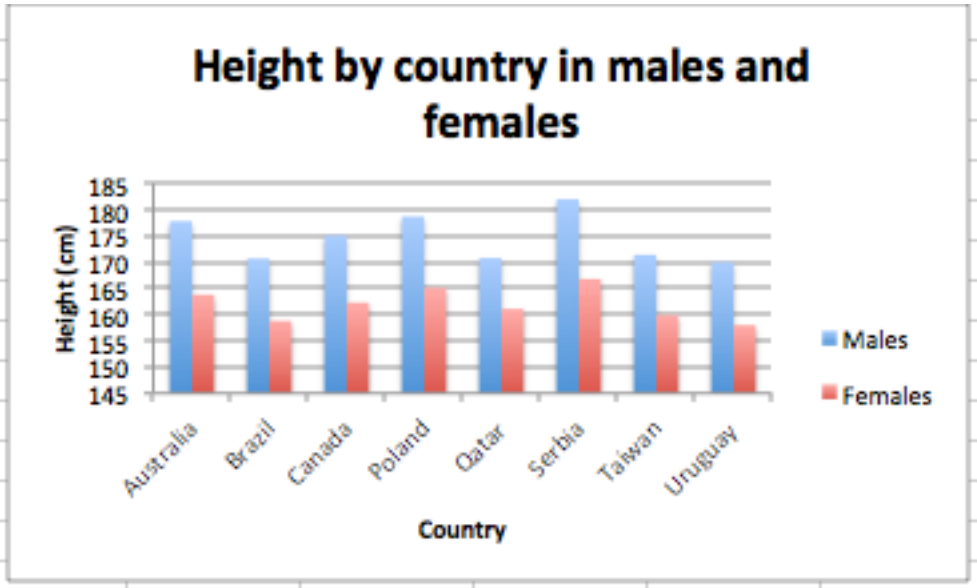
Workflow

- **Optimal**: include the whole workflow
- **Minimum**: Processed data to results
- **Optimal**: the instructions should be an automated script file (ie, “code”)
- **Minimum**: Written instructions that allow for the complete reproduction of your analysis

Exercise 1:
Pt1-Write instructions

Describe how to make a bar graph in excel

	Height in cm	
	Males	Females
Australia	177.8	163.8
Brazil	170.7	158.8
Canada	175.1	162.3
Poland	178.7	165.1
Qatar	170.8	161.1
Serbia	182	166.8
Taiwan	171.4	159.9
Uruguay	170	158



Exercise 1:
Pt2-follow instructions

- Make a graph using ONLY your class mate's instructions
  - What was described well?
  - What details are missing?

Exercise 1:
Pt3- Doing it reproducibly

Run code in R

Reproducibility =

Data

+

Source Code/Tools

What is research data?

White House Office of Management and Budget:

“The recorded factual material commonly accepted in the scientific community as necessary to validate research findings."

OMB Circular A-110: http://www.whitehouse.gov/omb/circulars_a110

Types of Data

- **Raw Data** – what you record
  - Exported from EMRs
  - Readings from machine
  - Results from surveys
- **Processed Data** – cleaned up
  - Ready for analysis
  - Properly formatted

Adheres to FAIR principles

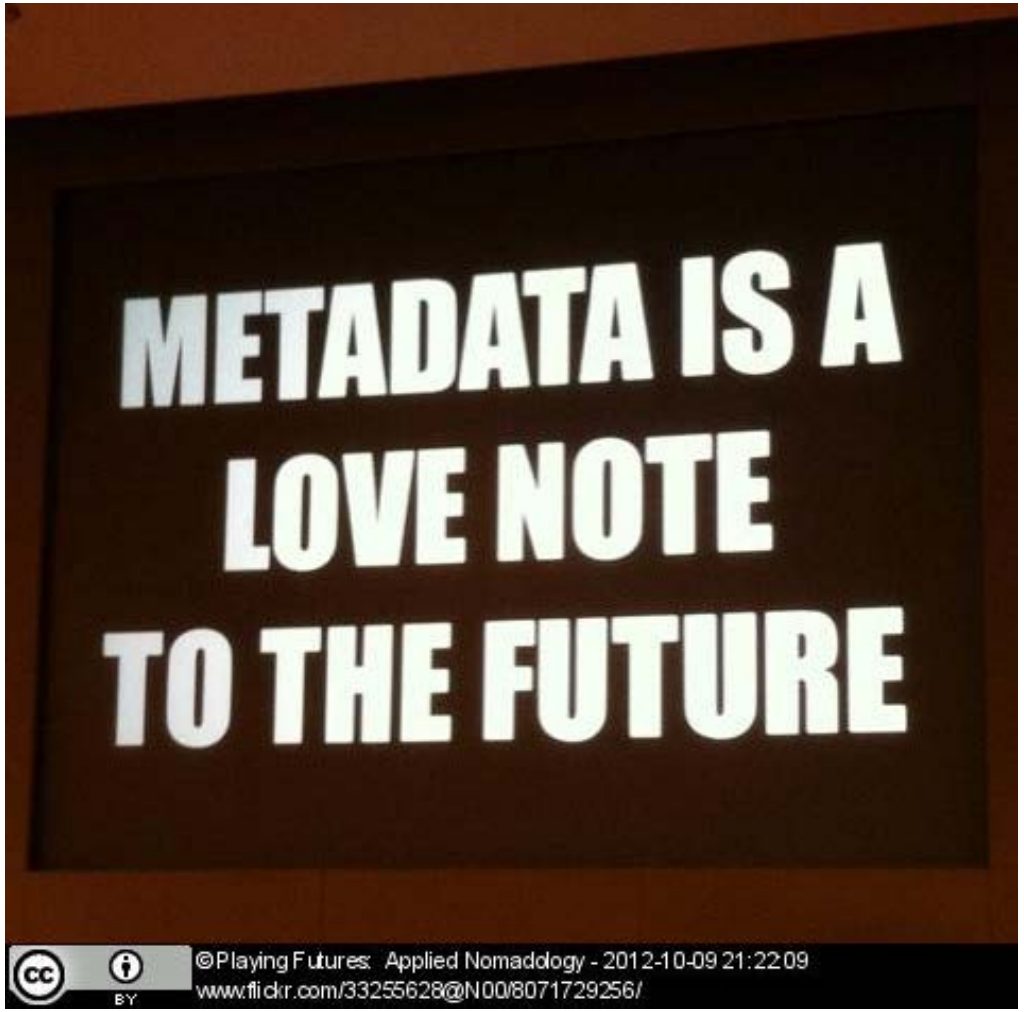
FAIR principles

- **Findable**: UIs, good **metadata**, in a searchable index
- **Accessible**: Quick access to **metadata** (and hopefully data)
- **Interoperable**: Use established **terminologies** where available
- **Reusable**: good **metadata**, usage license, provenance, meets established community standards

http://www.nature.com/articles/sdata201618

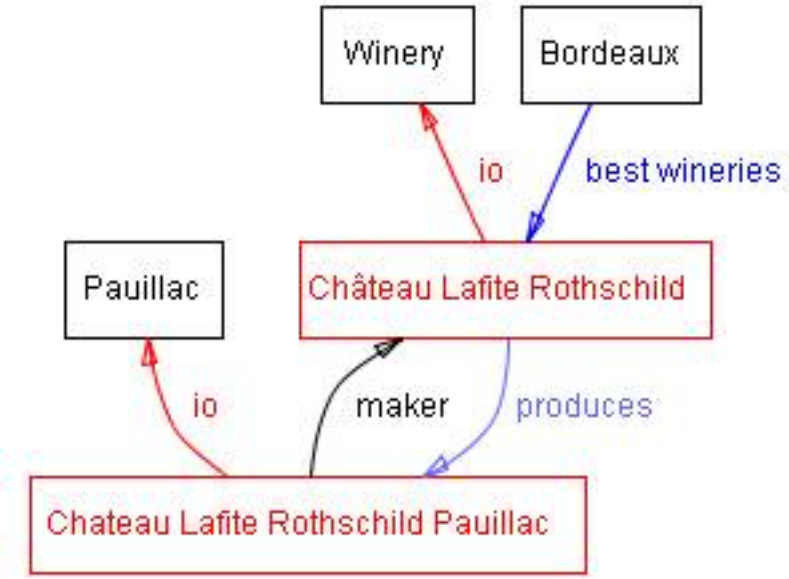
Metadata

- Metadata describes your research data:
  - Format depends on what type of data you have
- Dublin Core: http://dublincore.org/documents/dcmi-terms/
  - Can be applied to anything
- Many discipline specific metadata standards
  - MIAME: http://fged.org/projects/miame/
  - FHIR: https://www.hl7.org/fhir/index.html
  - Search for other standards: https://biosharing.org/standards/



Controlled Terminologies

- The language you use to describe your data
- **Examples**:
  - Airport codes
  - MeSH: http://www.ncbi.nlm.nih.gov/mesh
- **Ontologies**:
  - Gene Ontology: http://geneontology.org
  - SNOMED: http://bioportal.bioontology.org/ontologies/SNOMEDCT
  - Search for relevant ontologies: http://bioportal.bioontology.org



http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

File formats

Proprietary formats

- .doc, docx
- .ppt, pptx
- .xls, .xlsx

Non-proprietary formats*

- .txt
- .jpeg
- .tiff
- .csv

*Non-proprietary file formats are the most appropriate to use to ensure access to the data in the future

Software/Tools

- **What did you do to get from your raw data to the processed data**?
  - What software do you use?
  - Does it support log files/scripts?
  - Include version info and settings
  - What else does the software need to run?
    - Computer architecture
    - OS/Software/tool/add ons (libraries/packages)
    - External databases

- **Automate as much as possible**
  - Write scripts or save log files
  - If you're doing things by hand, record exactly how to did it

Reproducible research coding tools:

R Markdown and knitr

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

sas
StatRep

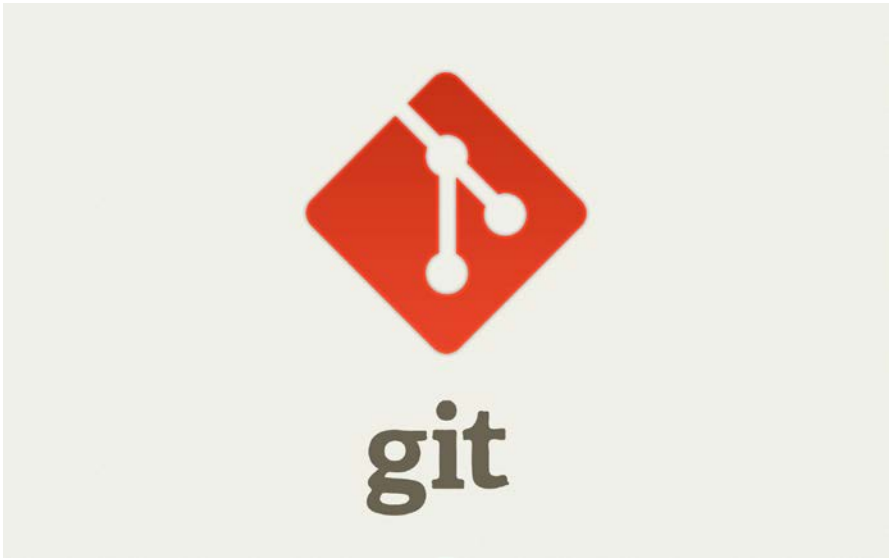
jupyter

GUIs can be reproducible

Version Control

- Save versions of your data/code
  - So you can go back when disaster strikes
- Old school: manuscriptV3.doc
- New school: version control systems
  - Allows for collaboration

GitHub



Reproducible research checklist

- **Think about the entire pipeline**: are all the pieces reproducible?
- **Is your cleaning/analysis process automated?**– guarantees reproducibility
  - Are you doing things "by hand"? editing tables/figures; splitting/reformatting data
  - Does your software support log files or scripts?
  - If no, do you have a detailed description of your process?
- **Are you using version control?**
- **Are you keeping track of your software?**
  - Computer architecture;
  - OS/Software/tool/add ons (libraries/packages)/external databases
  - version numbers for everything (when available)
- **Are you saving the right files?**: if it's not reproducible, it's not worth saving
  - Save the data and the code
  - Data + Code = Output

Adapted from:
https://github.com/DataScienceSpecialization/courses/blob/master/05_ReproducibleResearch/Checklist/Reproducible%20Research%20Checklist.pdf

Exercise 2: Assess your research Pt 1-assess yourself

Fill out the reproducible research checklist with your own work in mind

Which of the FAIR principles do you follow in your own work?

Exercise 2: Assess your research Pt2- Brainstorm

- Explain your research (elevator speech) to your table
- Explain where you're good at reproducible research
- Explain where you're not doing as well
- Brainstorm ways to fix it

Exercise 2: Assess your research Pt 3- share with group

- Pick a representative to give their elevator speech
- Explain the good
- Identify areas or improvement
- Explain how to improve

Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser