

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Changing the Culture of Data Management and Sharing A Workshop

The National Academy of Sciences Engineering Medicine
Rafael A Irizarry
Dana-Farber Cancer Institute and Harvard

 **@rafalab**
<http://rafalab.github.io/>

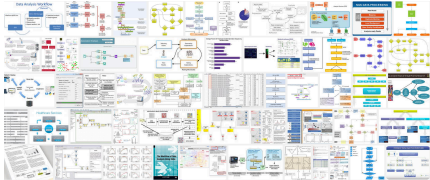
1

Ways public data advances science

- Correcting the scientific literature
- Discoveries from data re-use, for example by meta-analysis studies
- Enables development of better data analysis methods

2

Workflows Everywhere



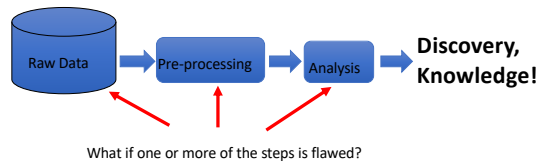
3

Typical workflow



4

Workflows always spit-out an answer



5

And that answer might be completely wrong



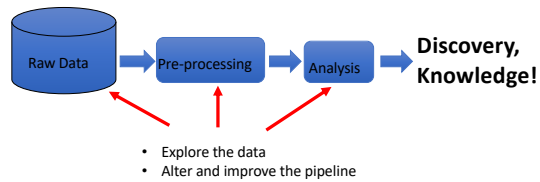
6

Pro tips

- When you find an **unexpected results**, be skeptical and check for systematic errors.
- Data exploration and visualization must be part of every workflow. Always, look at the data!

7

Access to raw data permits EDA



8

The questions

- What challenges exist with data re-use regarding data quality?
- How can researchers improve the quality of the data they are sharing so that it can be more effectively re-used to advance science?

9

My recommendations

- Top down rarely works
- Provide raw data, not dashboards or browsers
- Involve data analysts in decisions about how to share data
- Provide annotations in computer readable: a table for each study is often enough

10

How to share data with data analysts

- Ellis and Leek (2018) *American Statistician*
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7518408/>
- Broman and Woo (2017) *American Statistician*
<https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989>

11