Data Rigor and Reproducibly

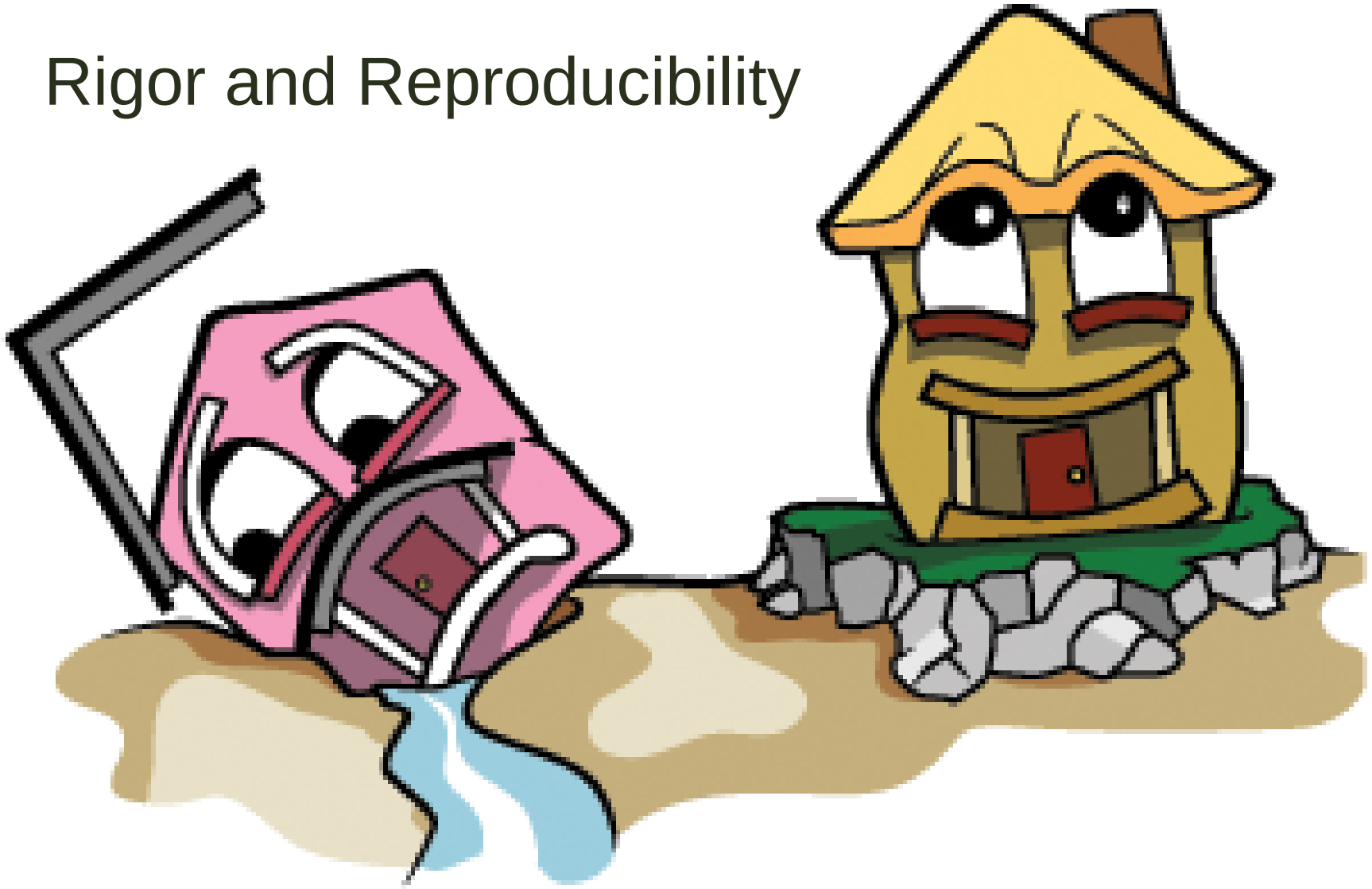
And what do I do with my gene list

Mary Allen

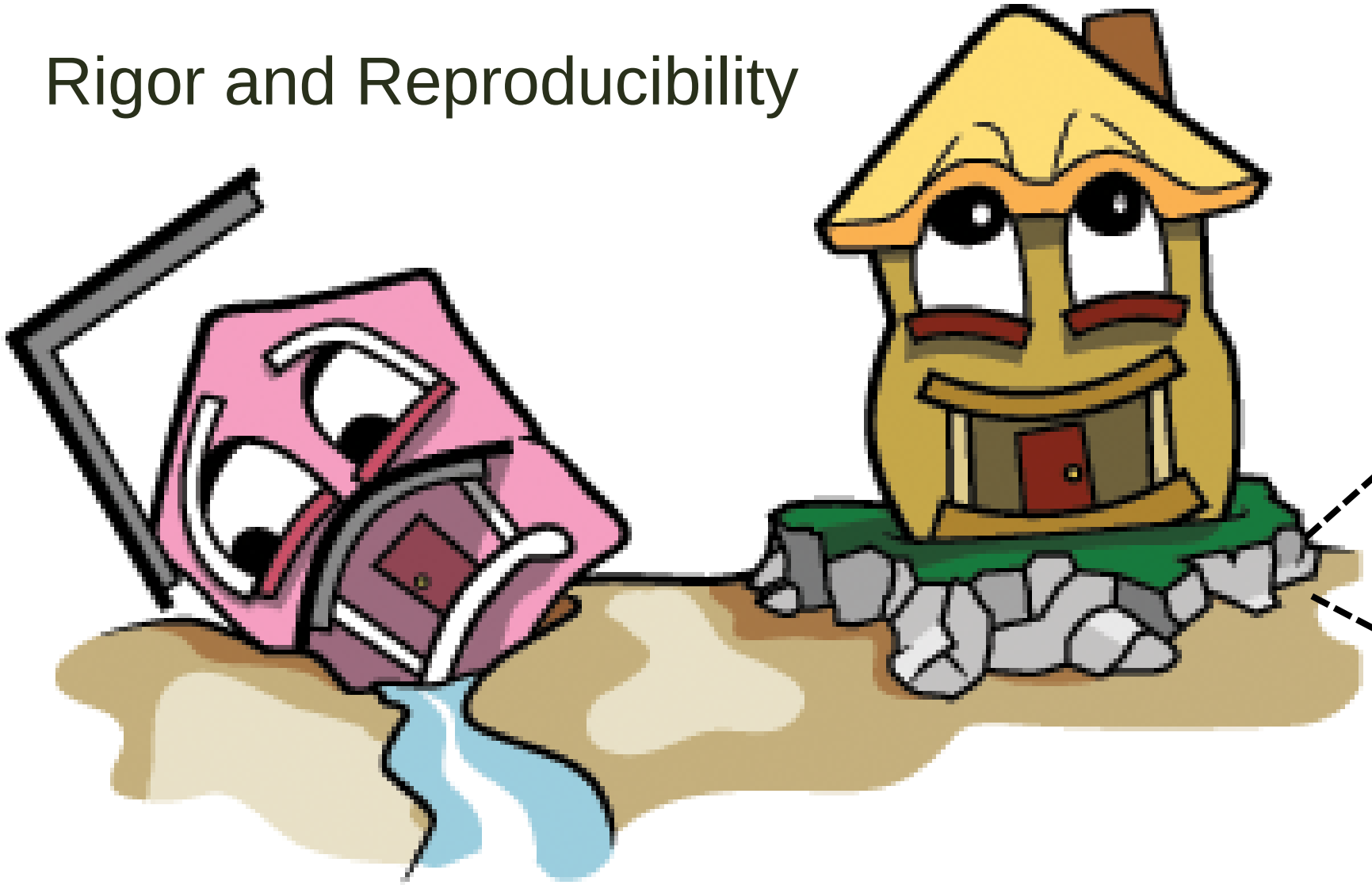
Data Data Data

- Genetic
- Omic
- Evolution
- System
- Images

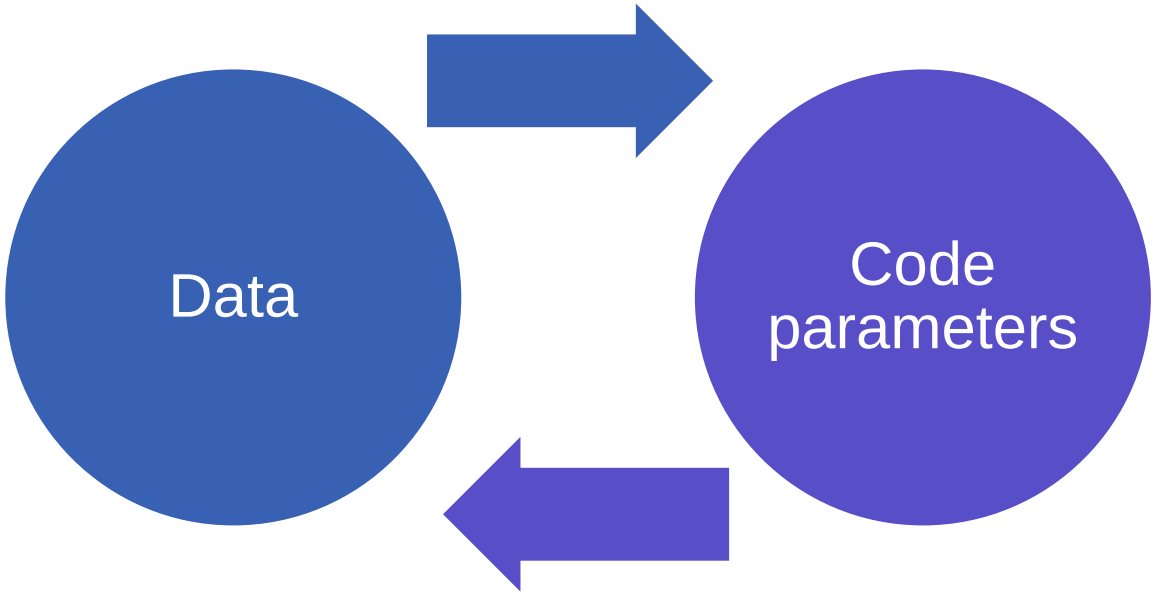
Rigor and Reproducibility

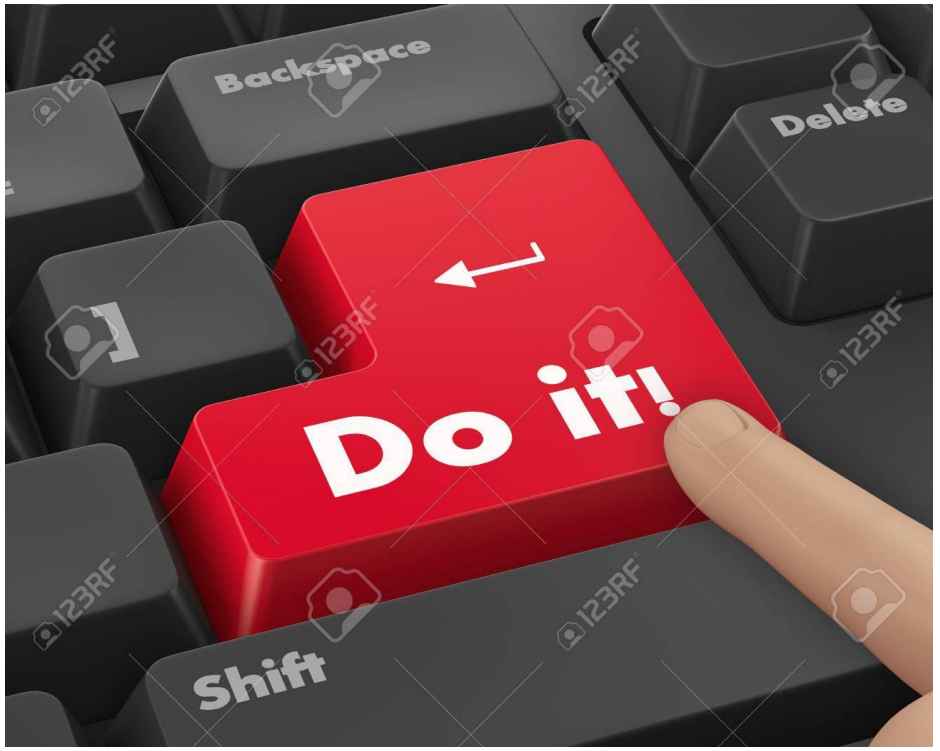


Rigor and Reproducibility



One step at a time

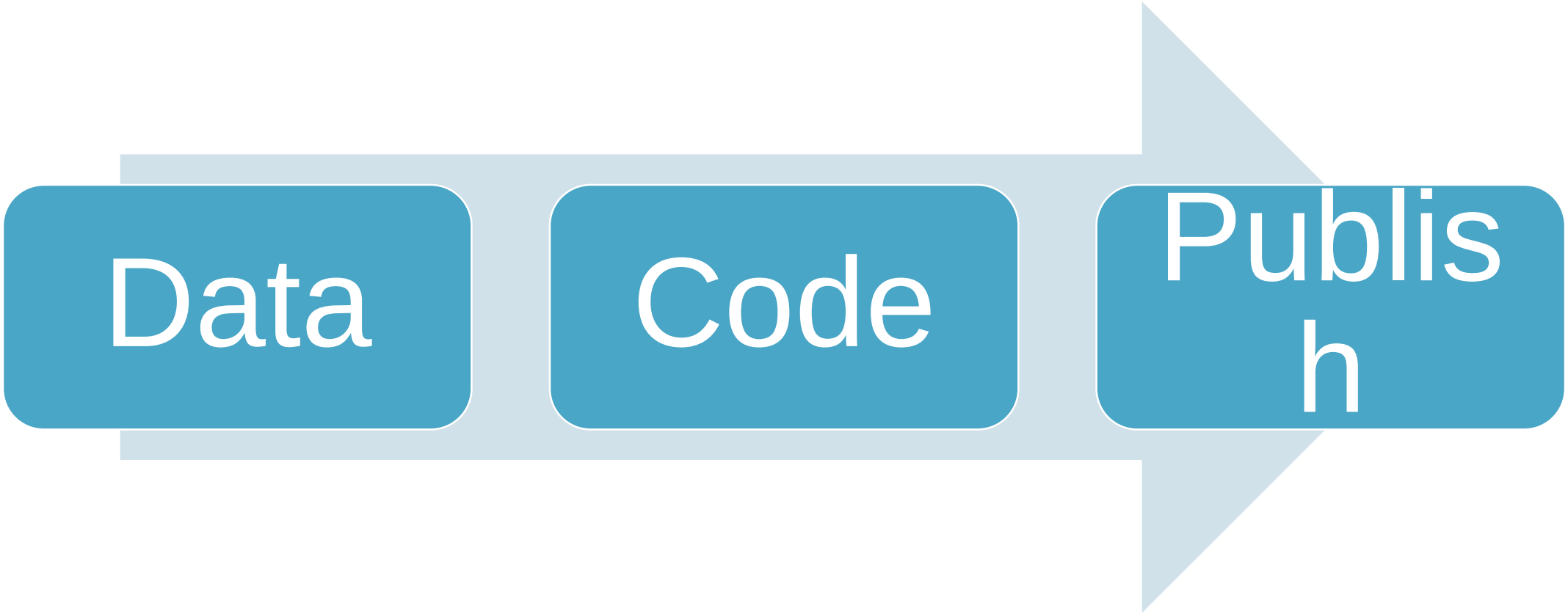




Pipelines

Nextflow
CWL

How we think it works….



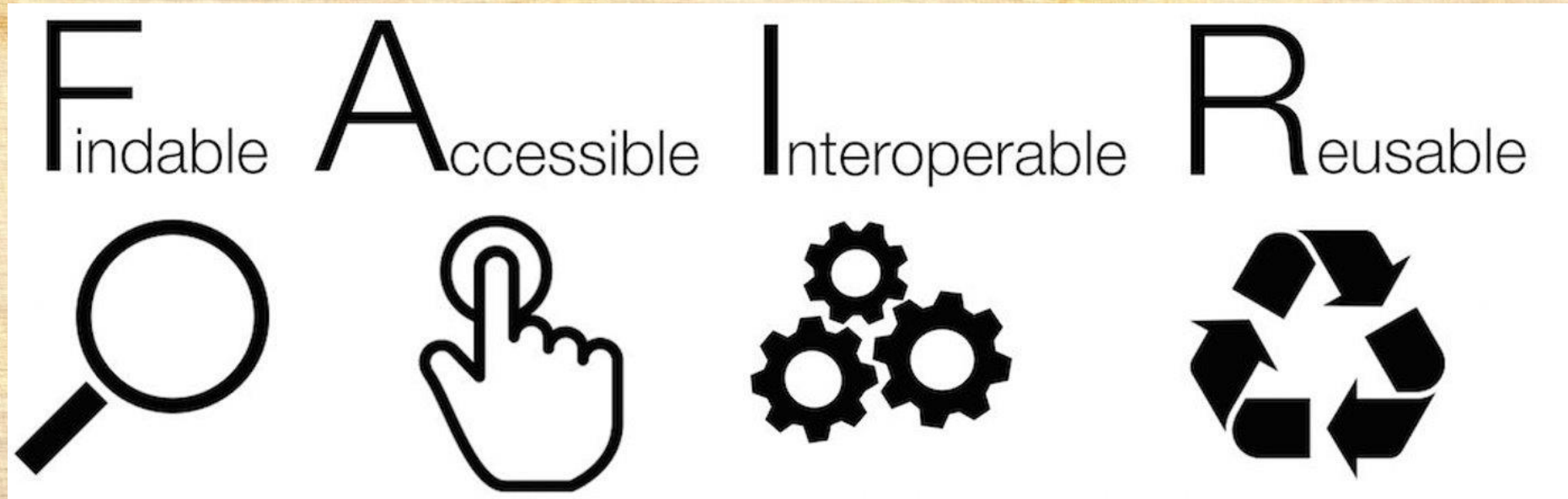
Before you start your foundation

Think about your data before you use it

- Data provenance
- Consent and Purpose
- Trustworthiness
- Privacy and Confidentiality

HAS DATA BEEN USED FOR GOOD? HAS DATA BEEN USED FOR EVIL?

/GERALDKUEHN


Findable
Accessible
Interoperable
Reusable

TRANSFER FIDELITY

sanity checking

When I transfer data

Copy with error check

Checksum algorithms before and after copy


```
-bash-4.2$ md5sum dis_from_mu_to_motif.bins1500.png
b59b8a9128346024b526e4a80024d743  dis_from_mu_to_motif.bins1500.png
```

```
BIOF-DOWELL-MA21:onesamplep53 maryallen$ md5 -r dis_from_mu_to_motif.bins1500.png
b59b8a9128346024b526e4a80024d743 dis_from_mu_to_motif.bins1500.png
```

How do you ensure data transfer quality?

WHAT IS CLEAN DATA? WHAT IS DIRTY DATA?

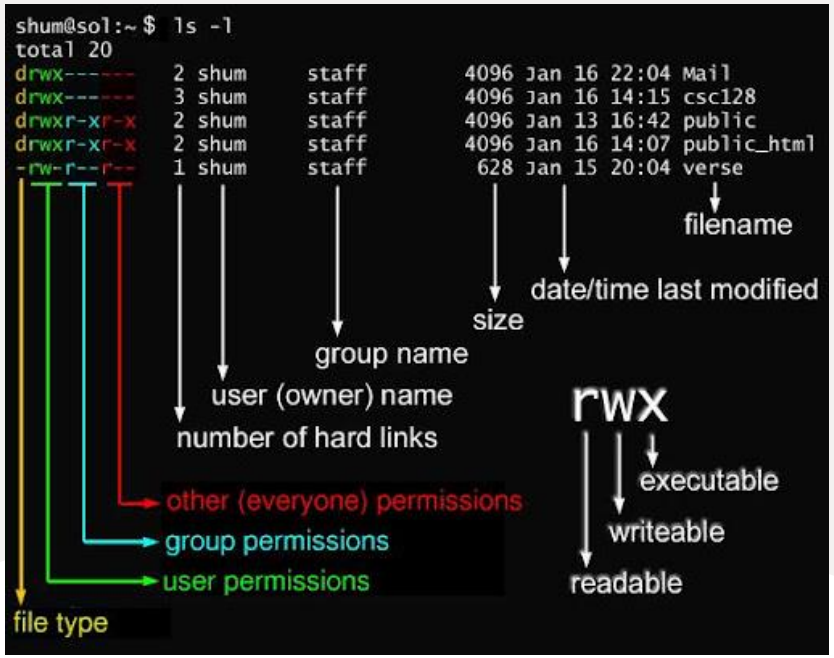
Ethical cleaning of Data

Never touch raw data and have backups

When you clean data you must develop rules for dropping data

All rules for dropped data need to be included in the methods to the publication

chmod 444 rawdata.txt



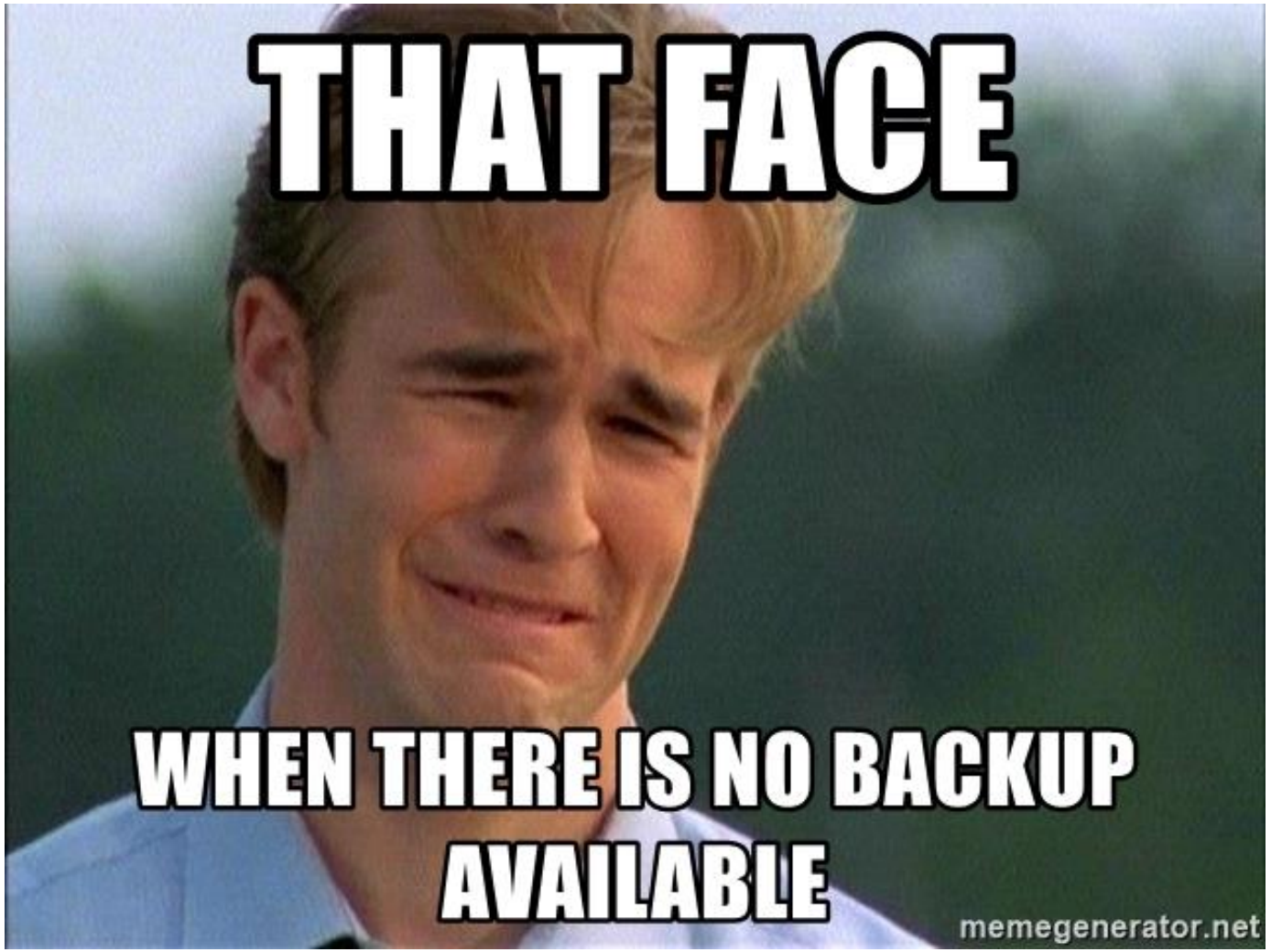
General Data repositories for open data (AND BACKUPS)

Back up that the government pays for?

European Nucleotide Archive - ENA Browser

GEO NIH (Attaches to NIH SRA, where the reads are kept.)

Putting stuff on GEO


THAT FACE
WHEN THERE IS NO BACKUP AVAILABLE
memegenerator.net

How safe is your raw data?

Is my data complete crap, or just dirty?

Quality control matters

How do you asses the quality of your data?

OpenRefine *A power tool for working with messy data.*

- Create Project
- Open Project
- Import Project
- Language Settings

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all sup

Get data from

- **This Computer**
- Web Addresses (URLs)
- Clipboard
- Database
- Google Data

Locate one or more files on your computer to upload:

Browse... No files selected.

Next »

Version 3.4.1 [437dc4d]

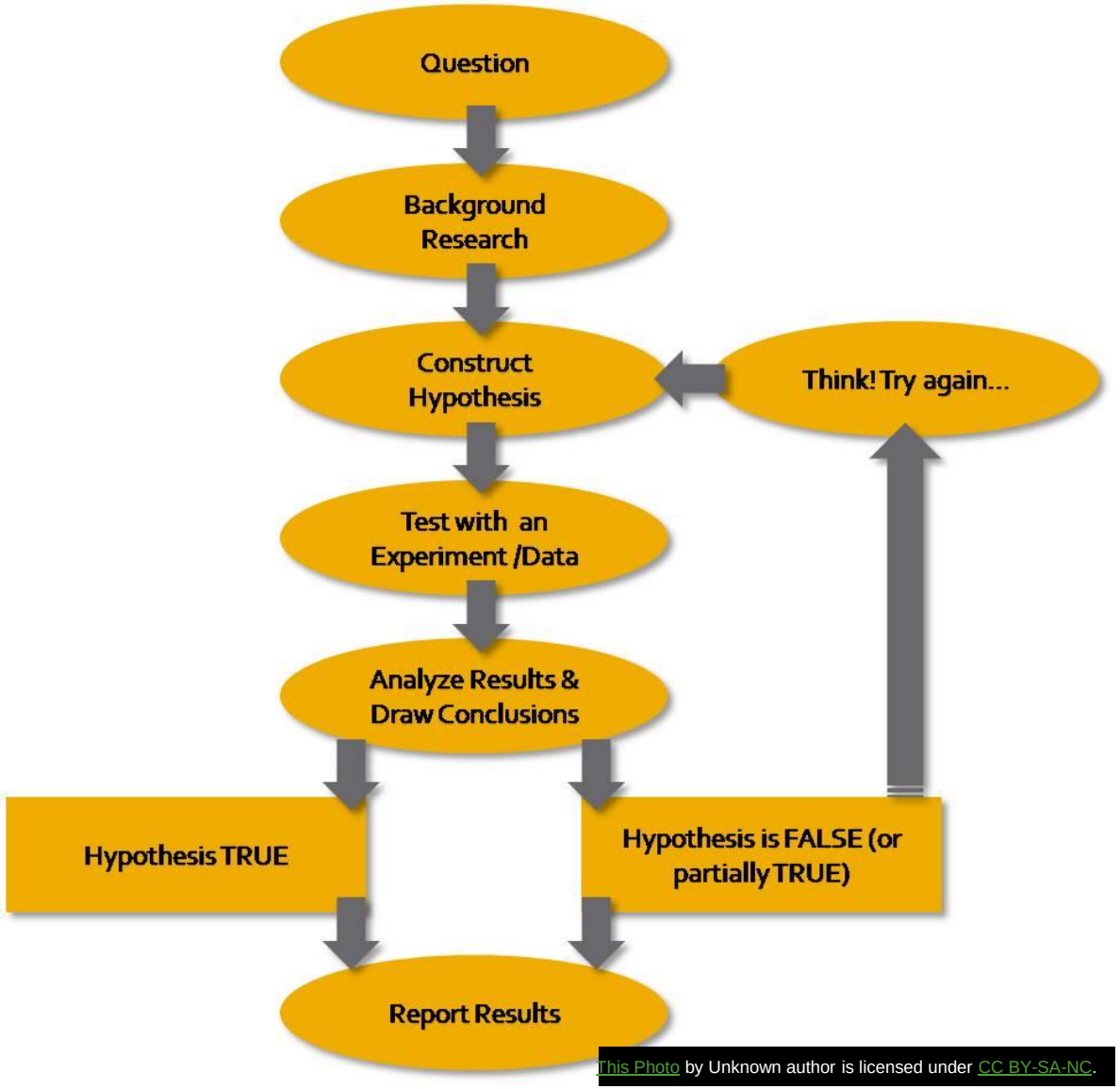
Preferences
Help
About

What will you use to clean dirty data?

Top

Scientific rigor

The strict application of the scientific method to ensure unbiased and well-controlled experimental design, methodology, analysis, interpretation and reporting.



How well did you follow the scientific method in the last month?

Why didn't you follow the scientific method?

“ Date weather ”

“ uneffciencit ”

“ - Dae ”

“ everything ”

“ Takes a long time ”

“ I didn't know what I was doing ”

“ didn't have a hypothesis before starting ”

“ Rushing a little bit. ”

“ also doing what my grad student told me to do ”

“ Time ”

“ Time ”

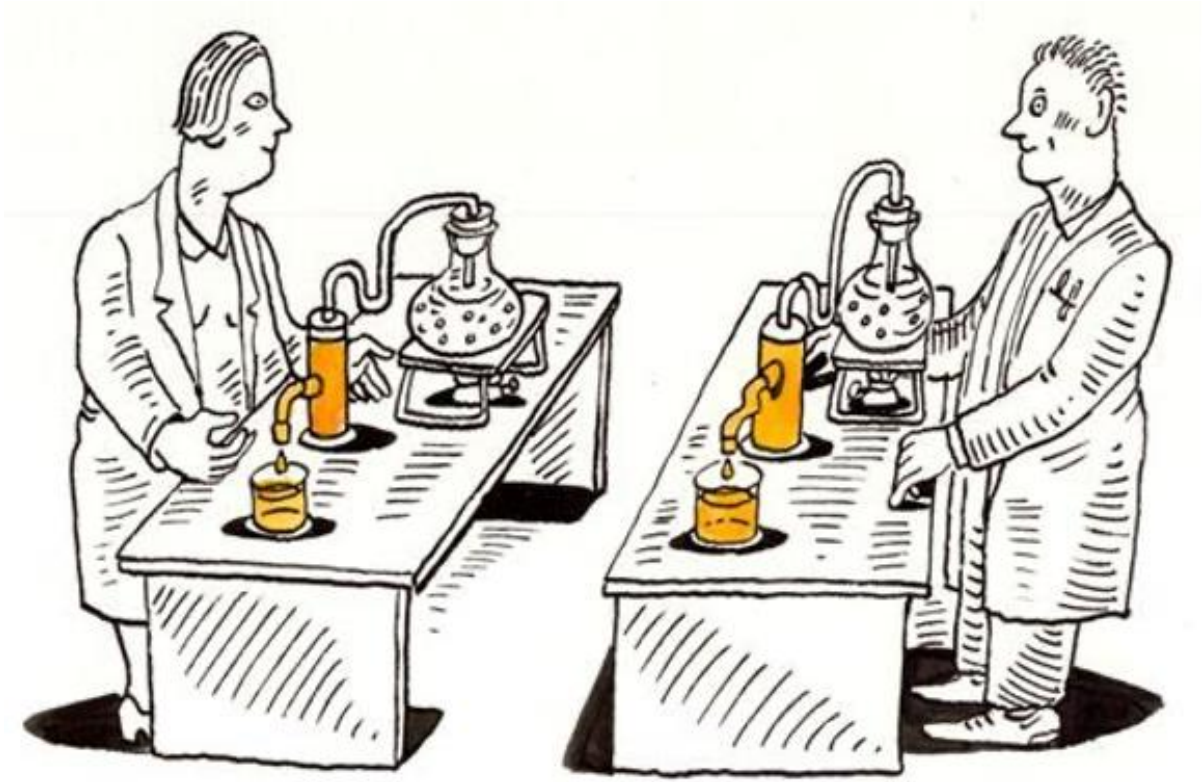
“ time ”

“ Liquid scintillation counter is temperamental and makes me cry ”

“ capitalism ”

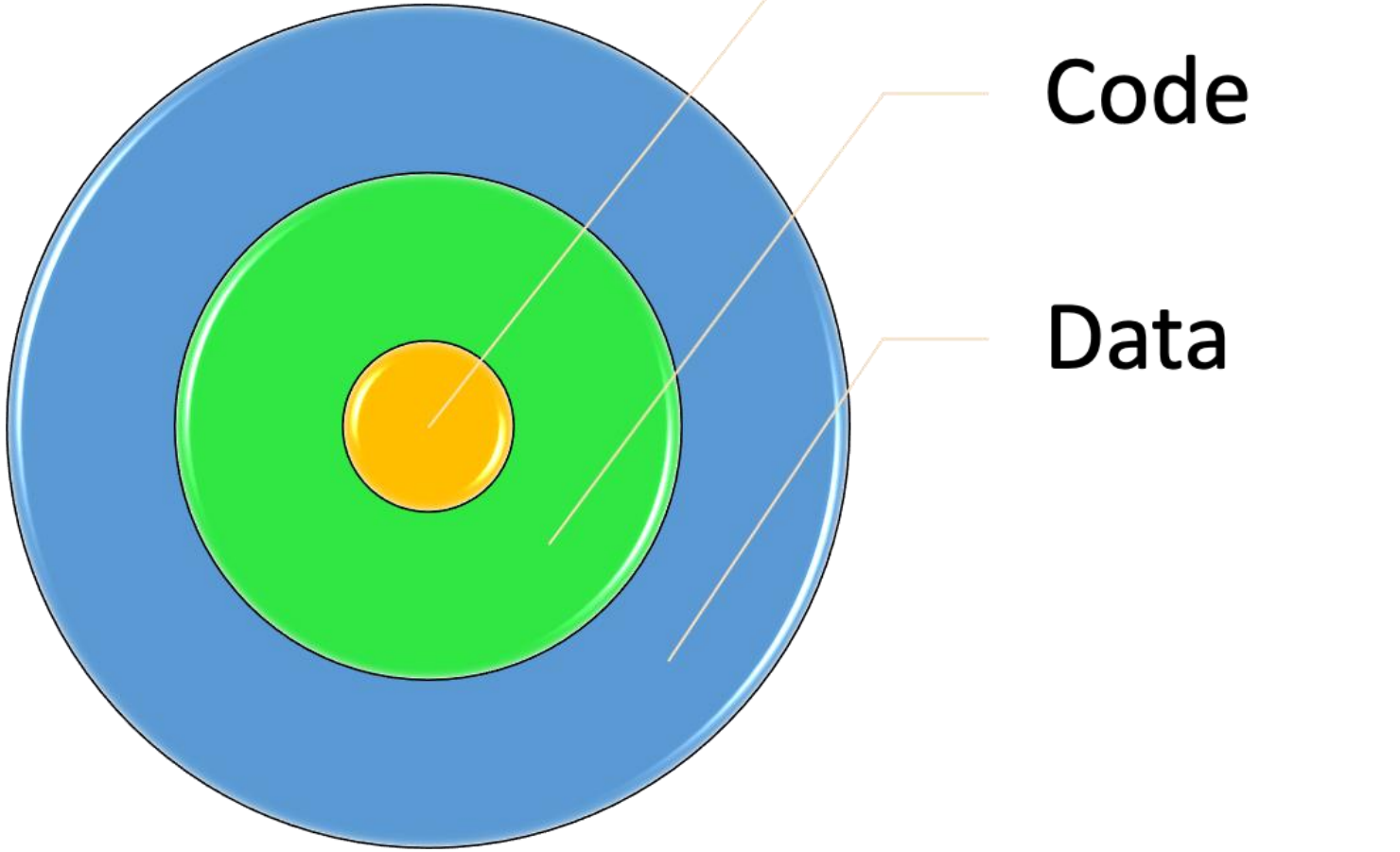
Scientific reproducibility

If someone else read my publication, could they use the same methods as me in another time and place to get the same results?



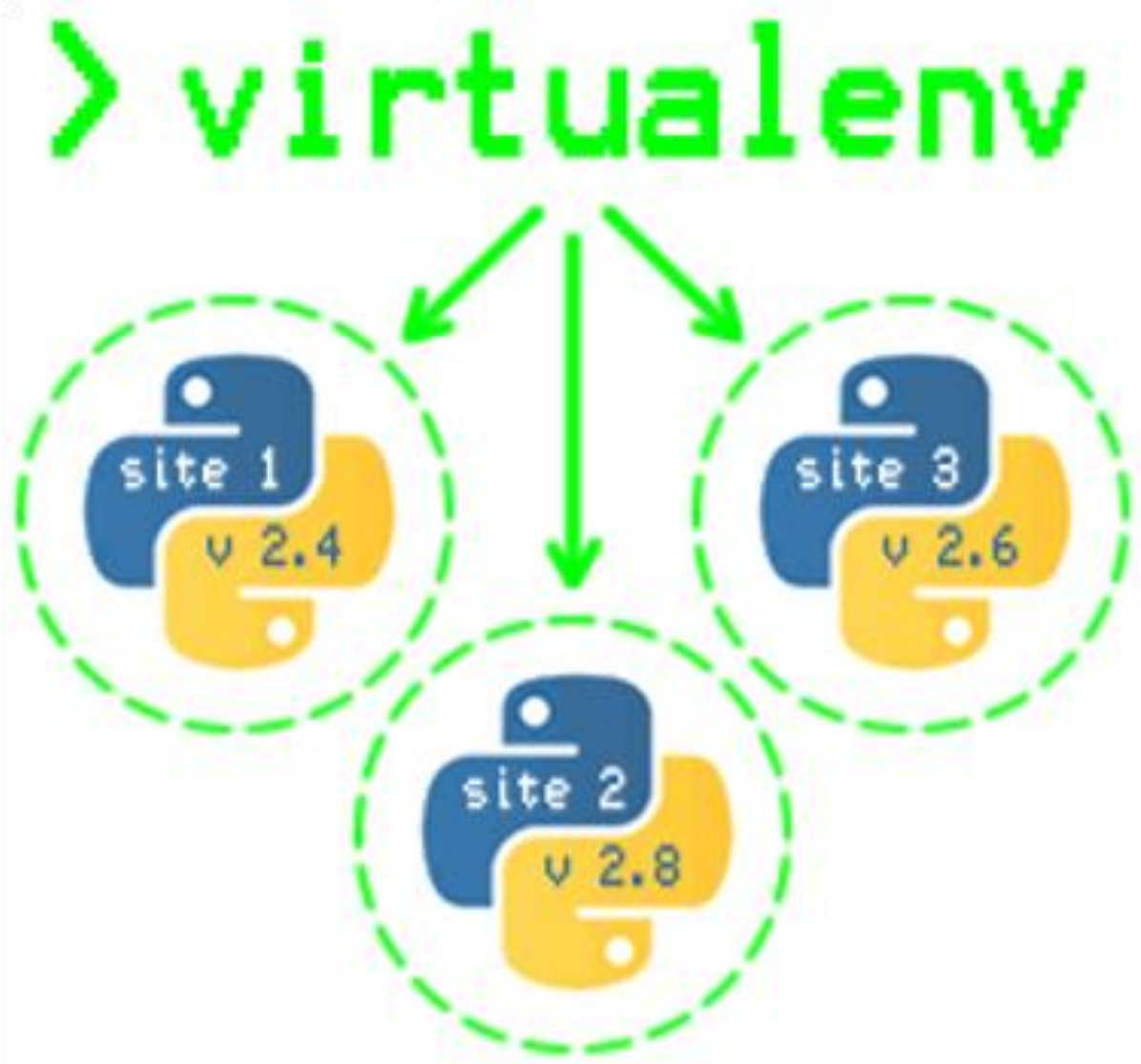
SAVING YOURSELF TIME AND MONEY VIA REPRODUCIBILITY

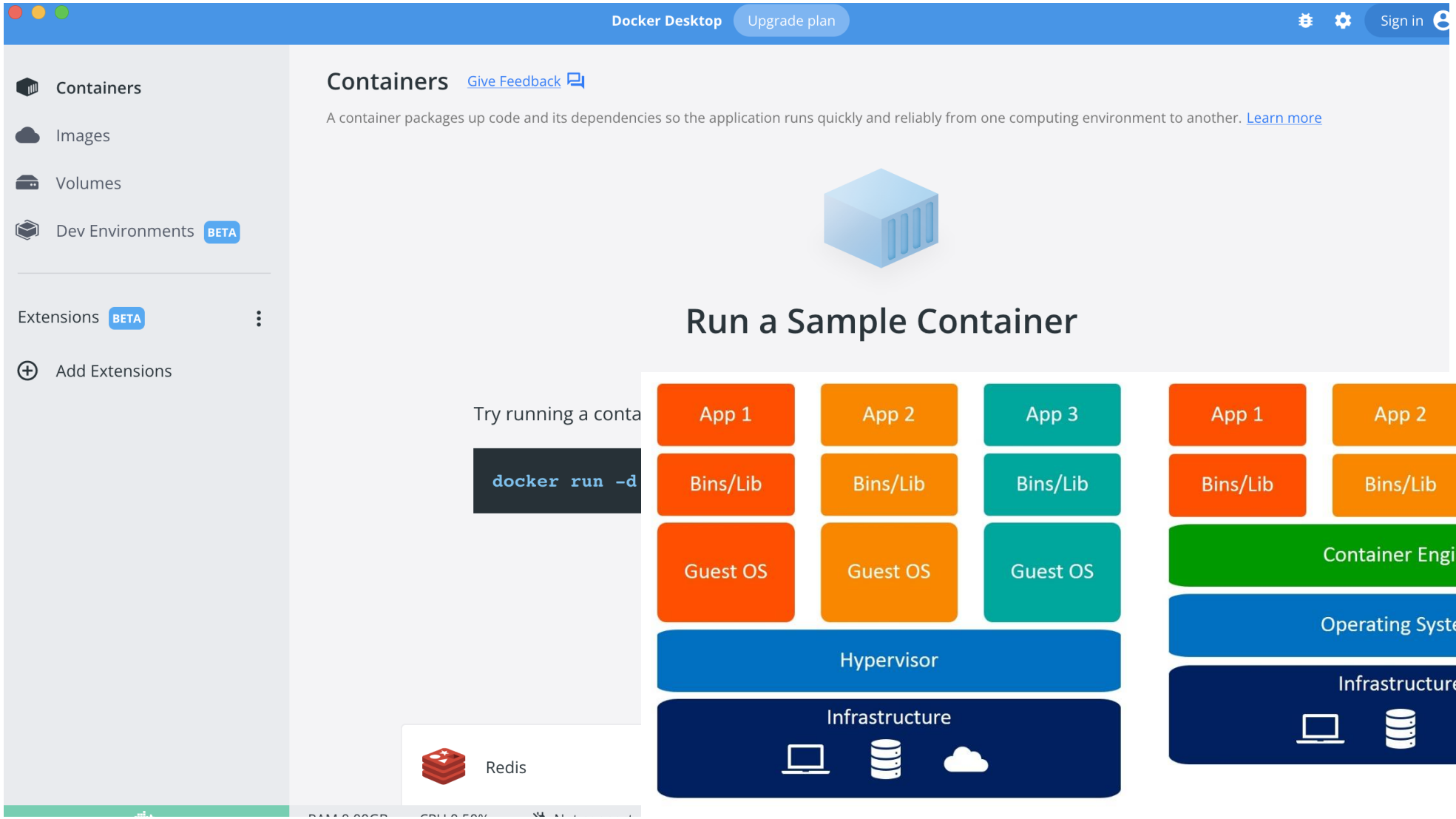
What needs to be reproducible?



“Machine” reproducibility

- Virtual machines
- Docker and other container tools

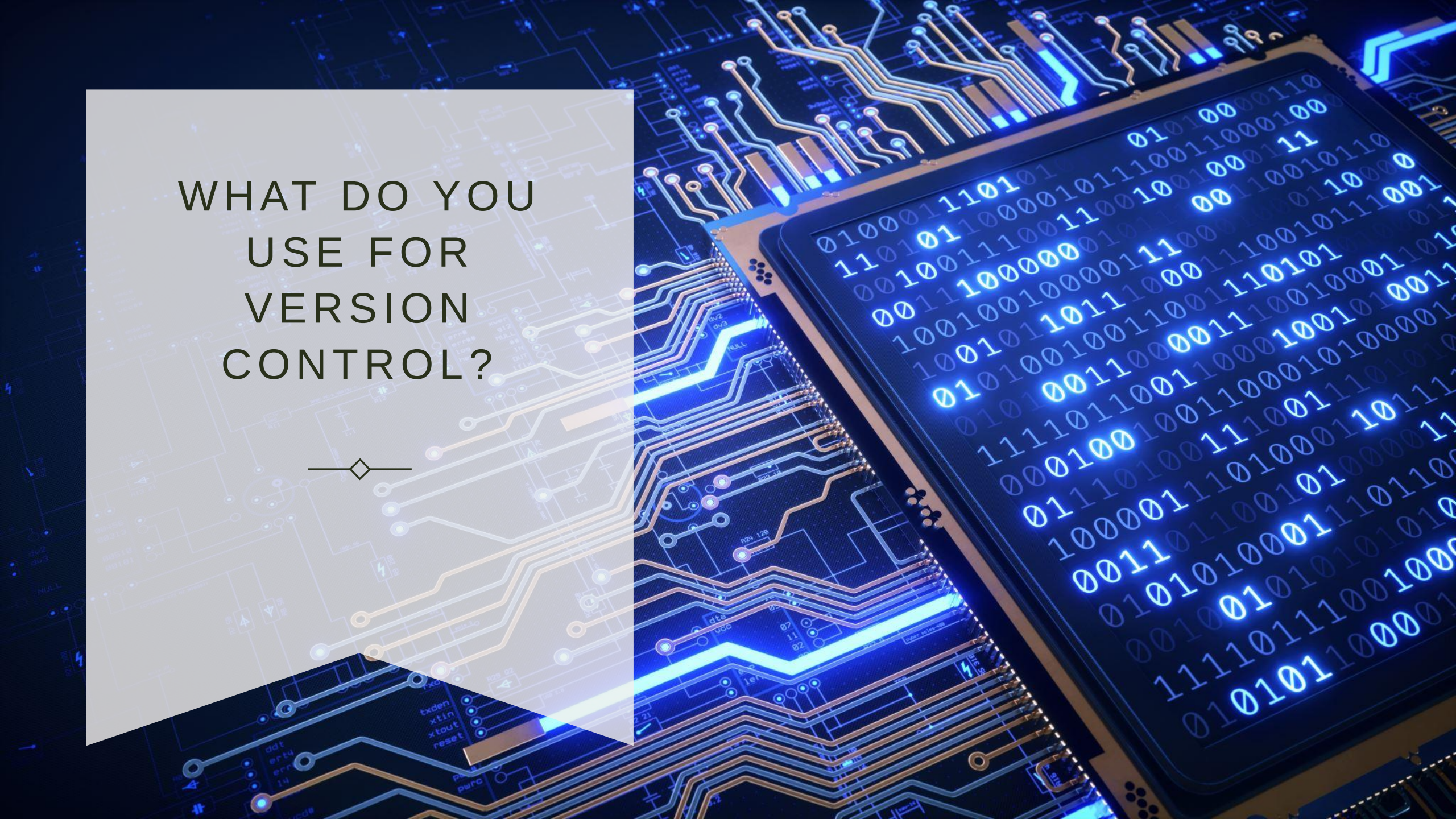




Virtual Machines

Containers

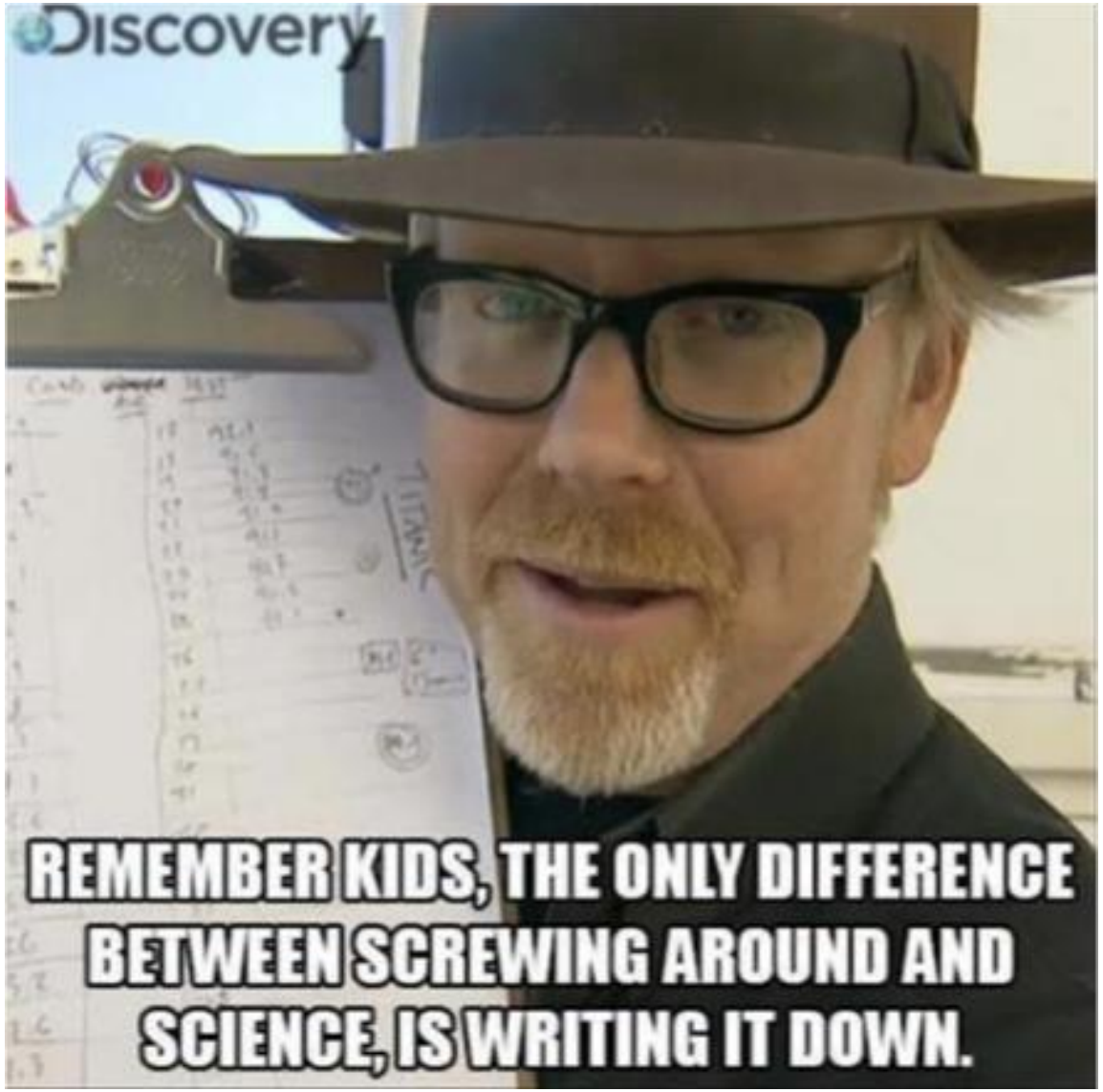
WHAT DO YOU USE FOR VERSION CONTROL?



Code reproducibility

git + GitHub

Version control/history



WRITE IT DOWN

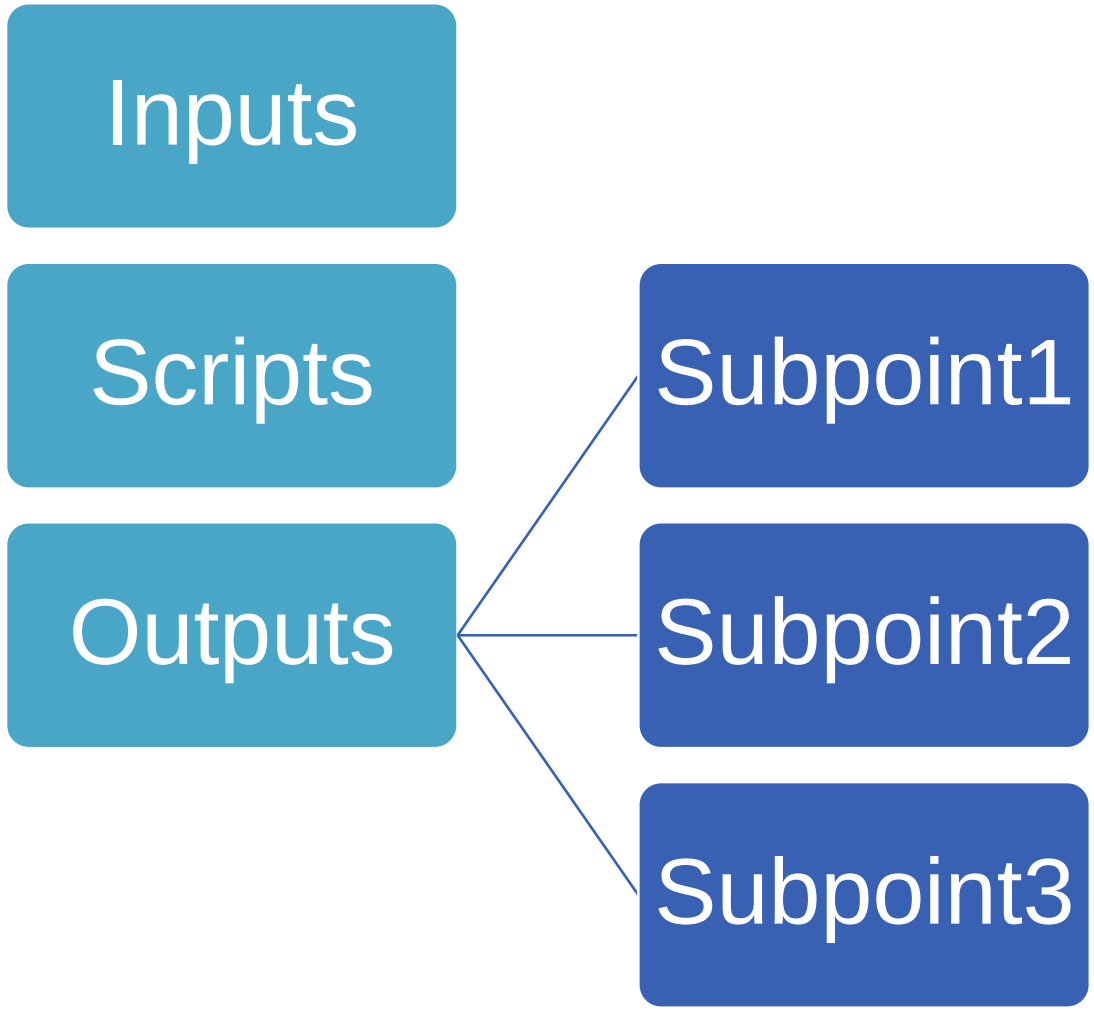
It's easy to lose a tube in a freezer

When biologists say we "lost our shit" at work,

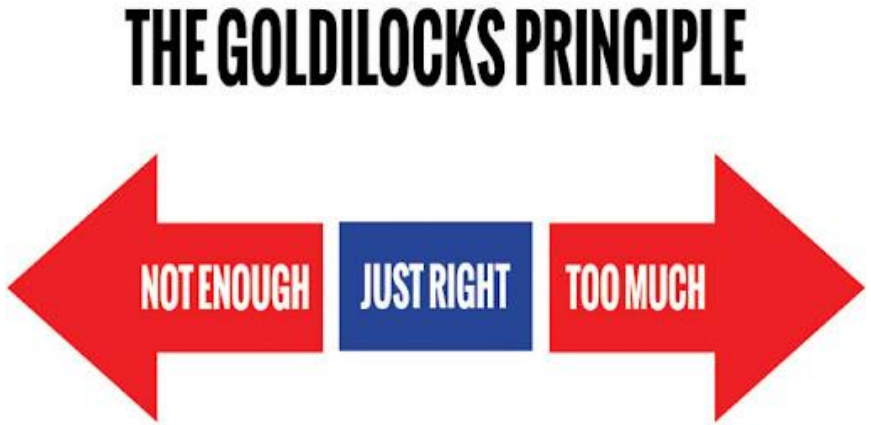
it's not a metaphor.

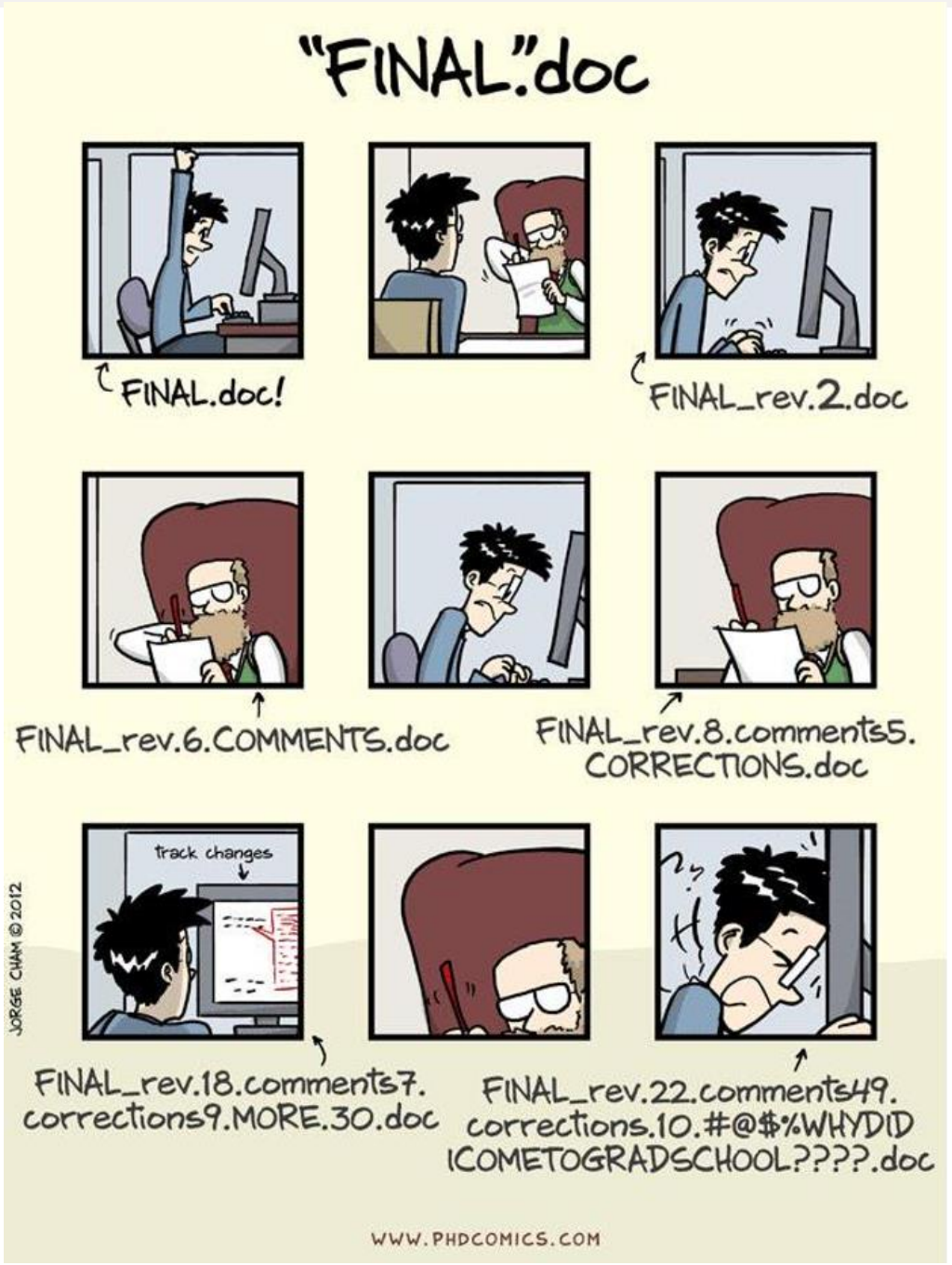
IT'S EASIER TO LOSE A FILE ON A SERVER

My suggested data analysis directory structure



Name of a file/directory



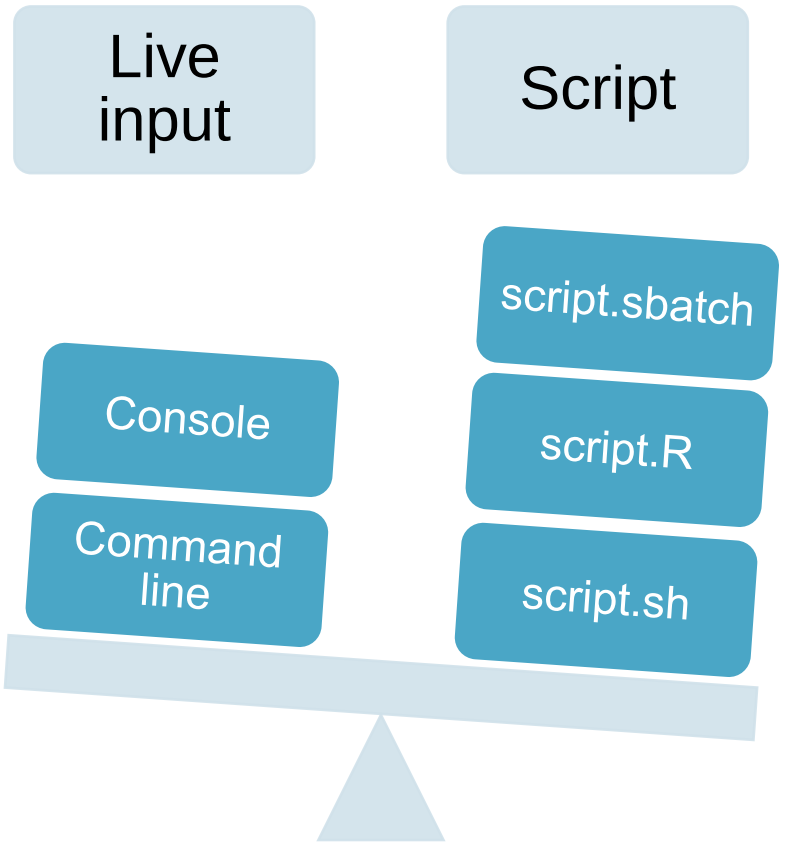


Version Control

GitHub

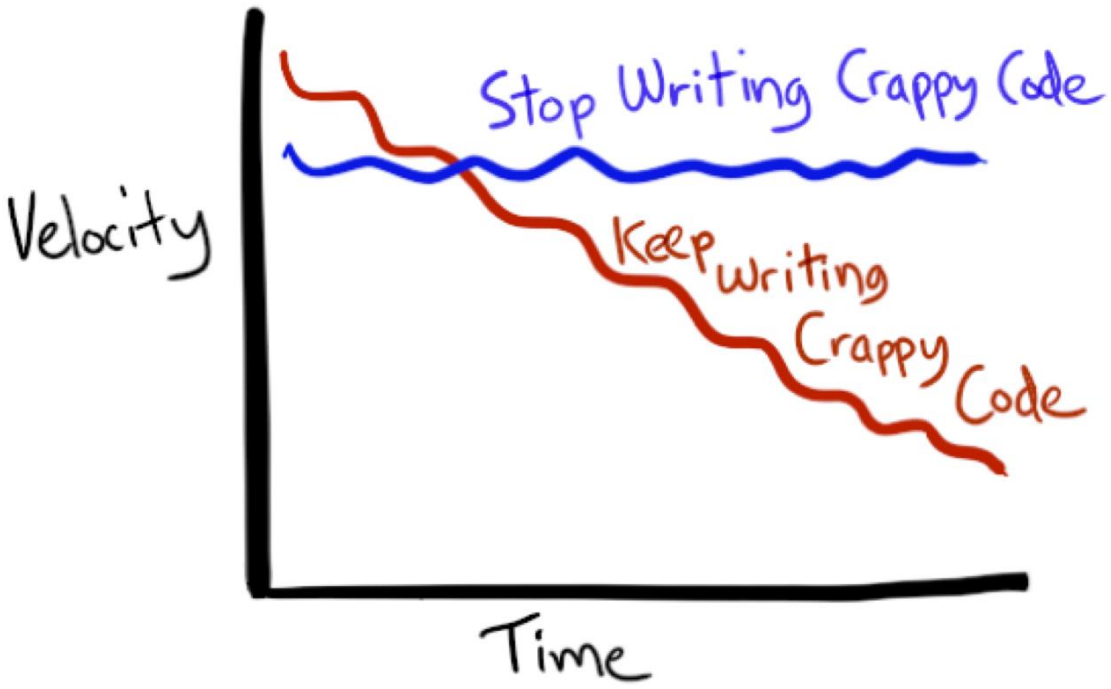
HOW DO YOU KEEP RECORDS OF WHAT YOU HAVE DONE WITH DATA?


WHAT DID I DO


Live input
Script
script.sbatch
Console
script.R
Command line
script.sh

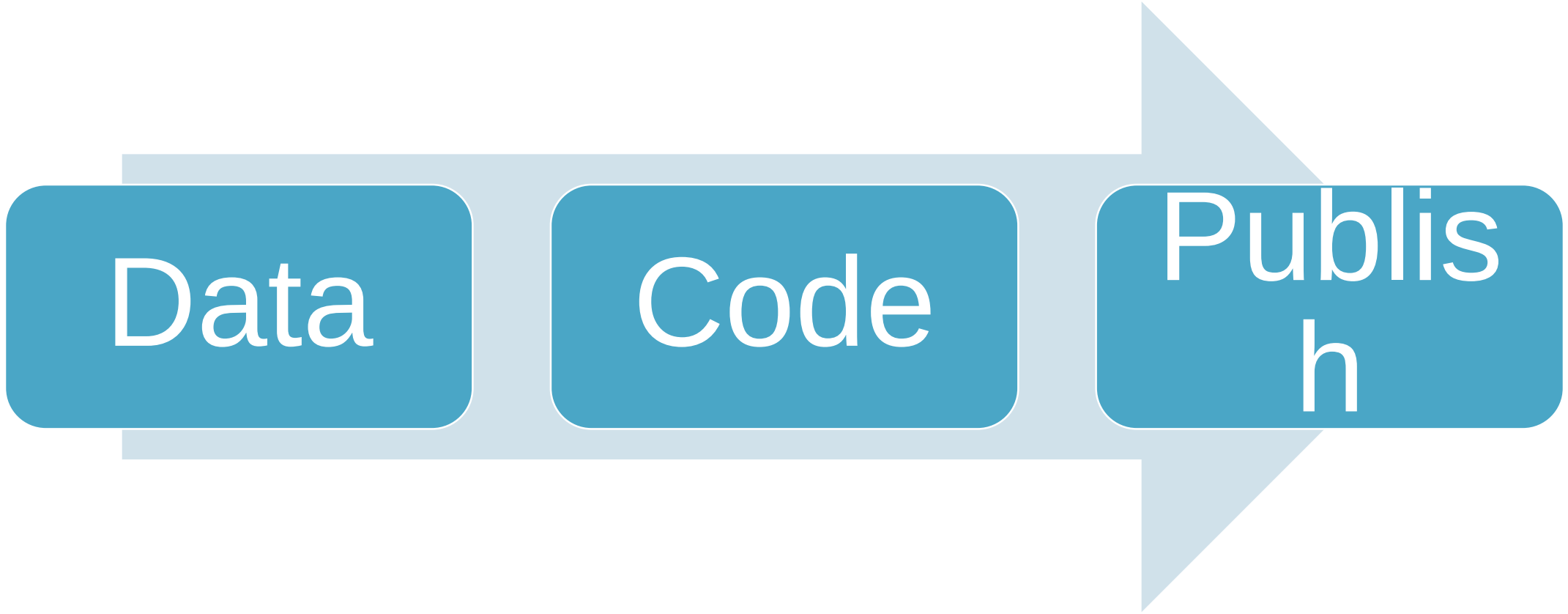
Digital Lab notebooks

- Directory structures
- Readmes
- Jupyter notebooks/R markdown
- Version control



MAKE A CHOICE

What is good code?



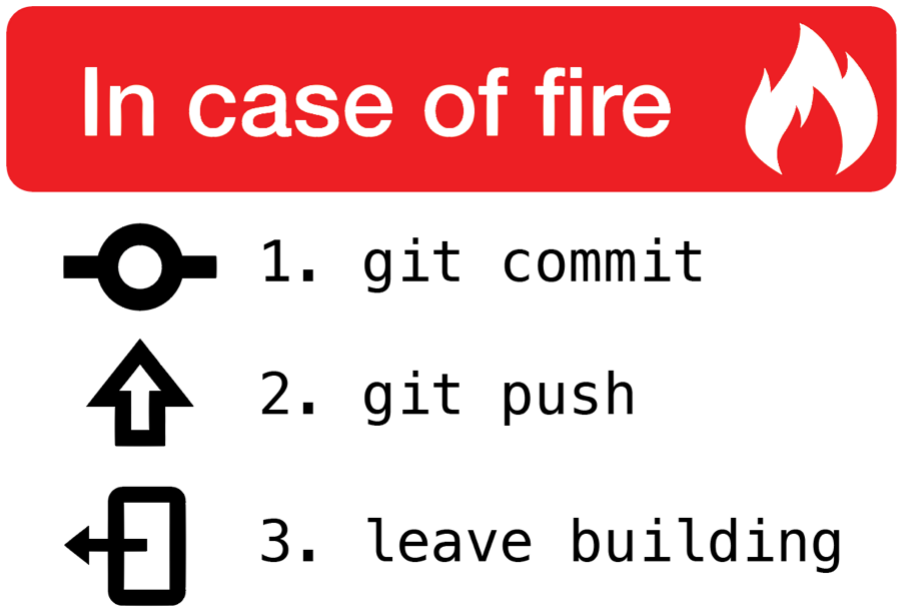
Warns you when you put in garbage!

Garbage *in* Garbage **OUT**

{UX Research}

Has comments that explain code

Frequent use of github with good comments so you can go back to a previous version



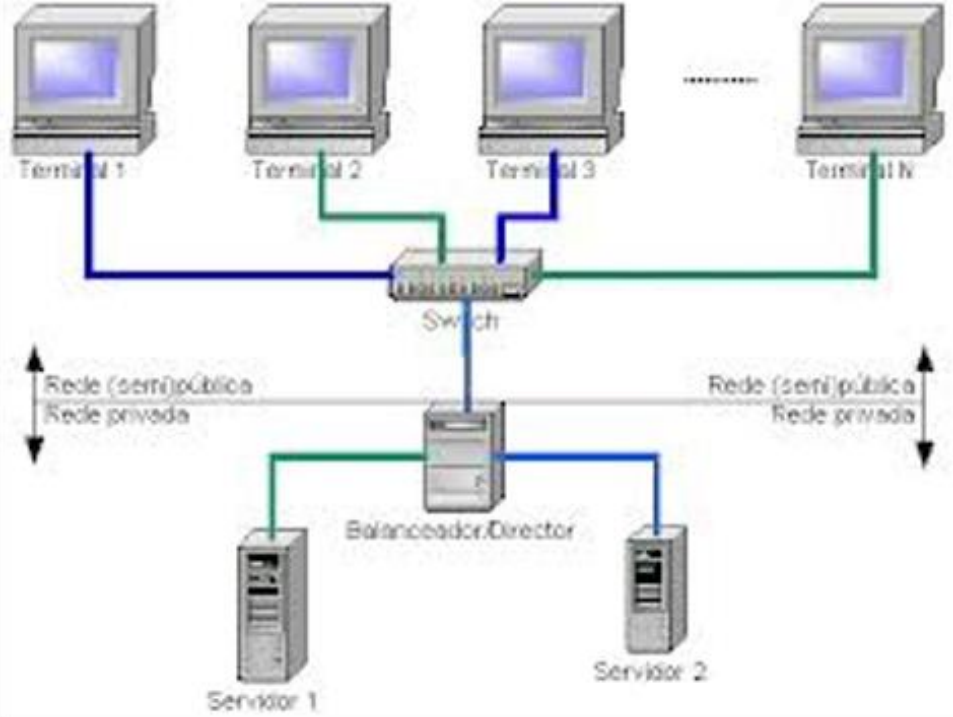
How would you rate your notes and comments on your code???

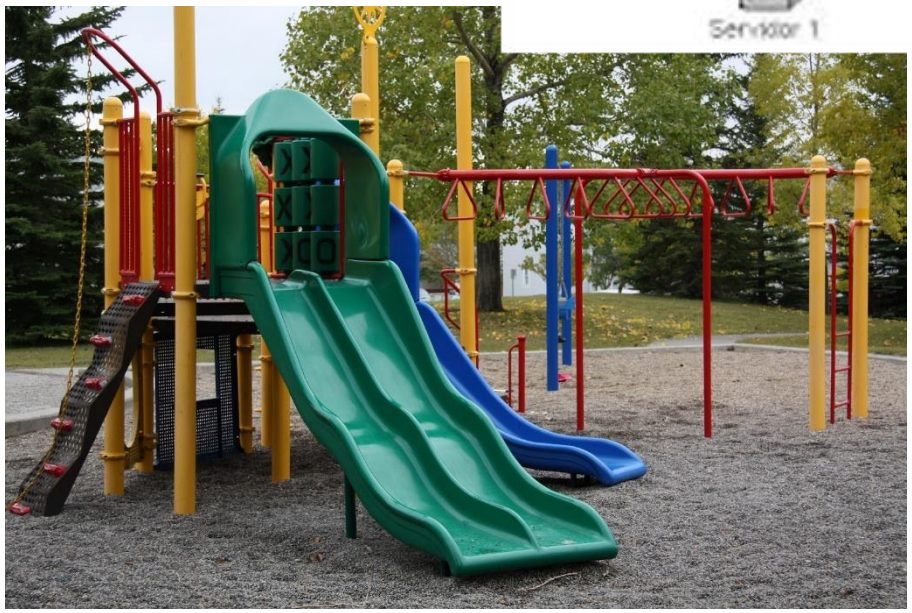
Tests itself frequently

Have you ever tested that your code is giving you the expected result?

THE END

Playing nice on a super computer

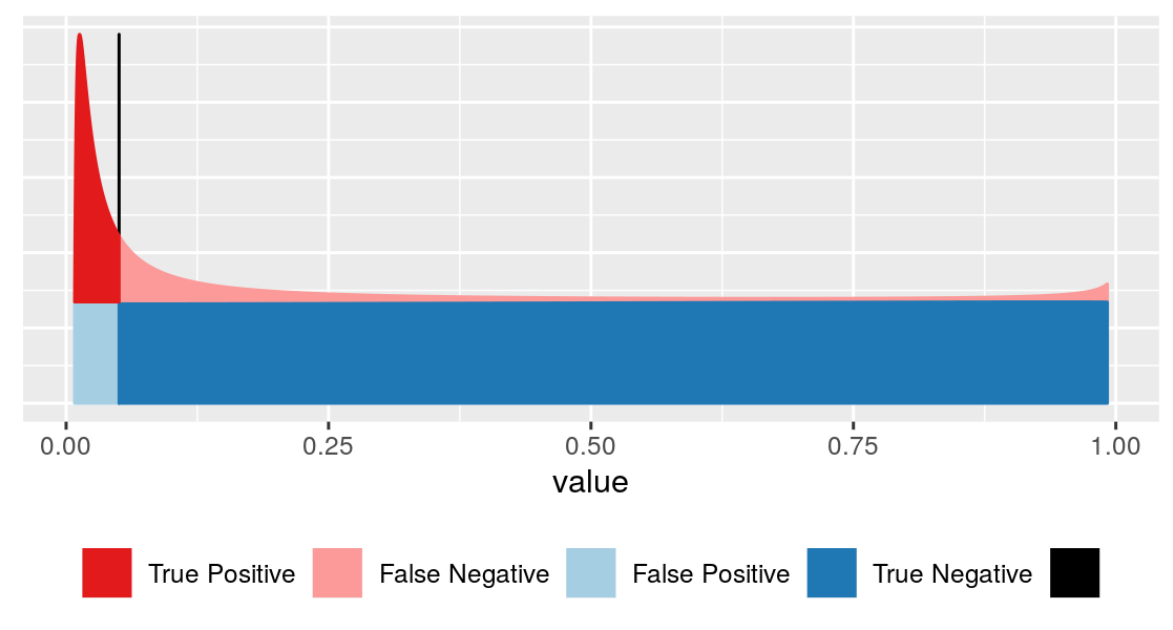




THE END

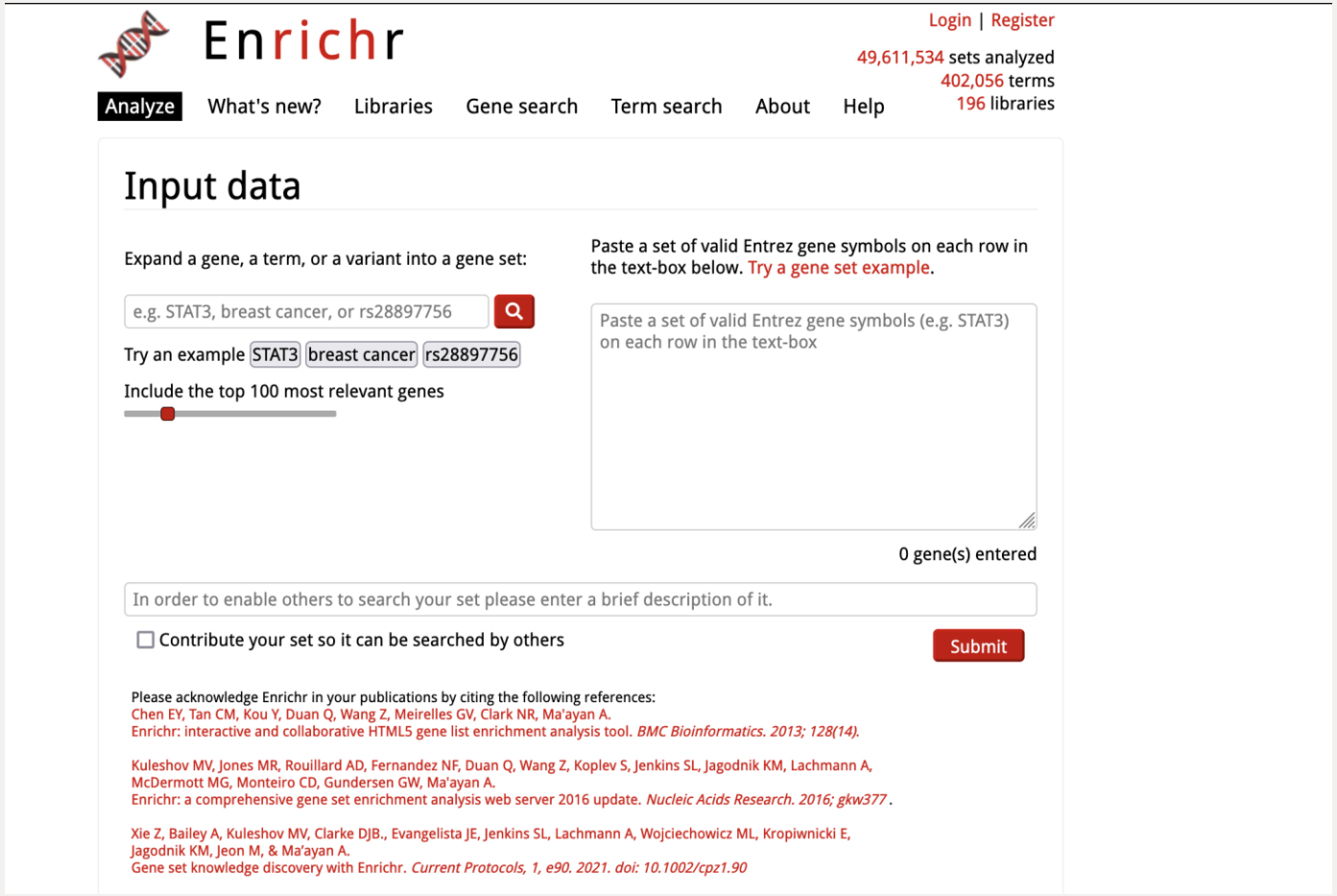
I GOT A GENE LIST! YEAH!
NOW WHAT?

DO I REALLY HAVE SIGNIFICANT GENES?



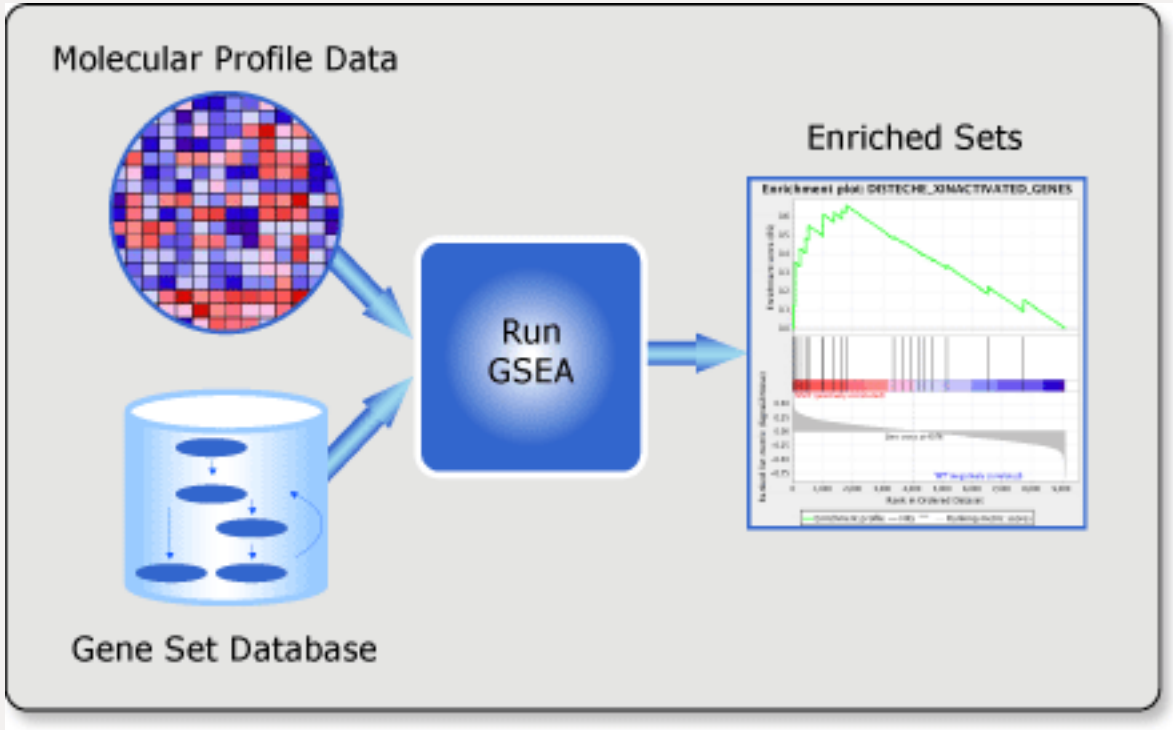
https://www.huber.embl.de/msmb/Chap-Testing.html

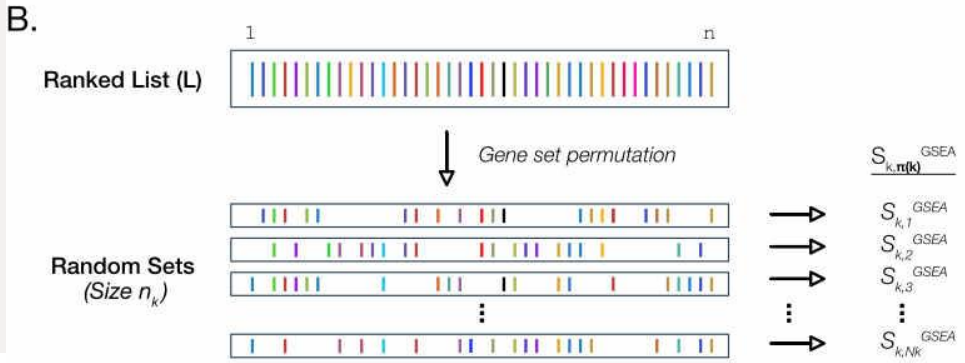
ENRICHR



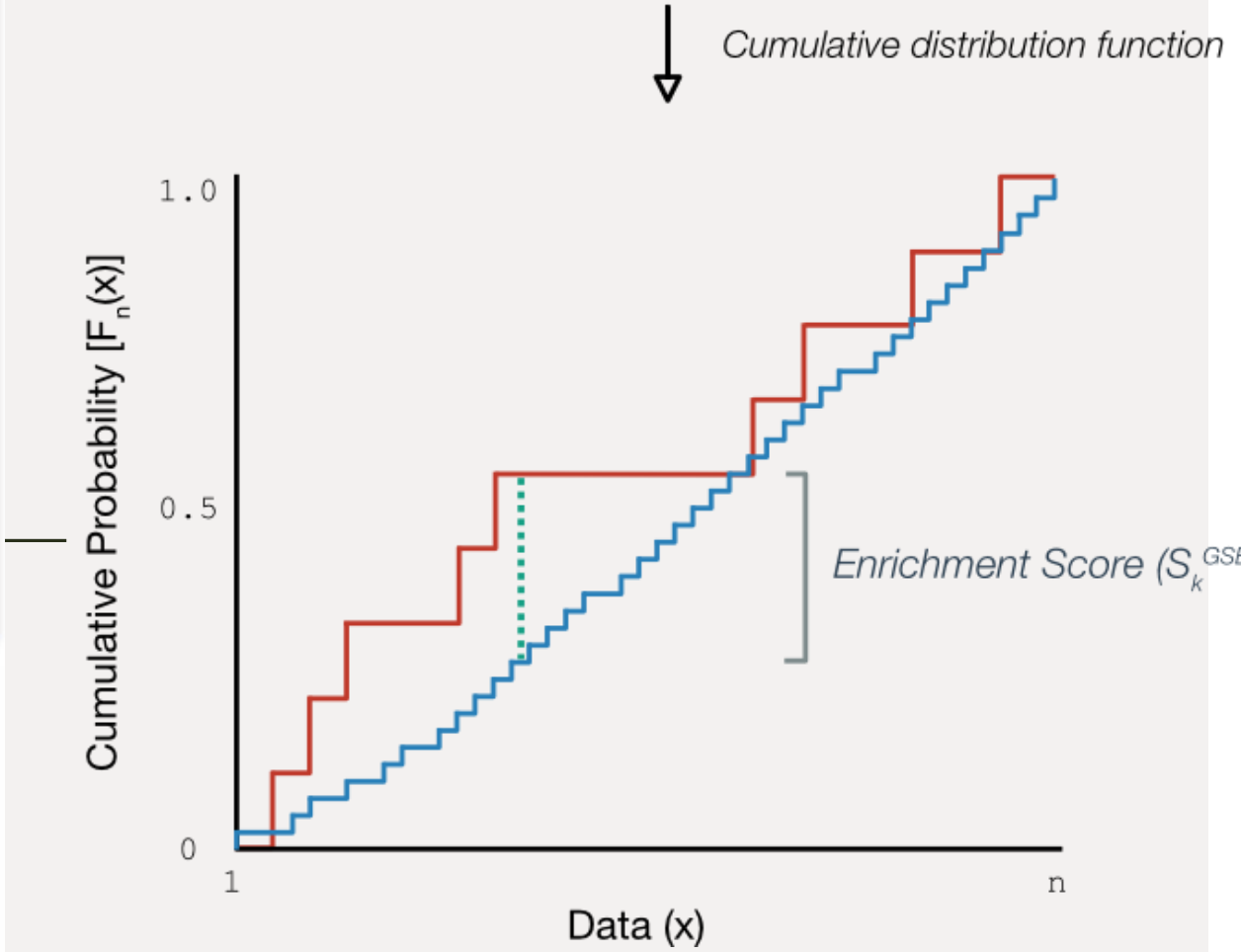
https://maayanlab.cloud/Enrichr/

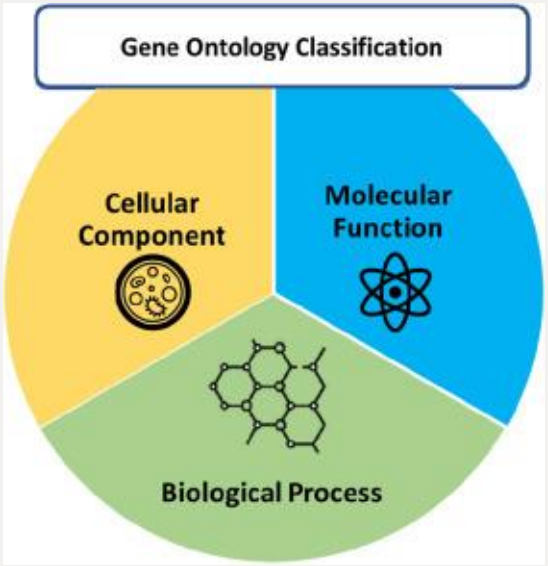
GSEA

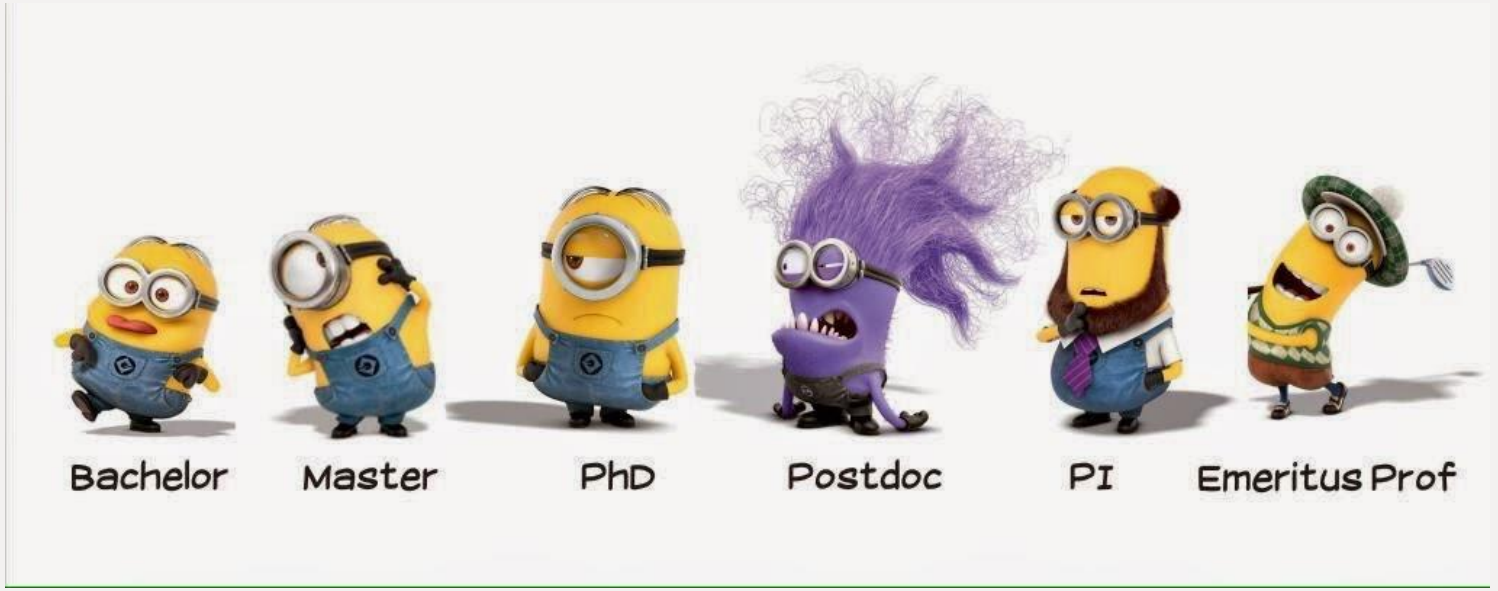

Molecular Profile Data
Run GSEA
Enriched Sets
Gene Set Database


B.
1
n
Ranked List (L)
Gene set permutation
Random Sets
(Size n_k)

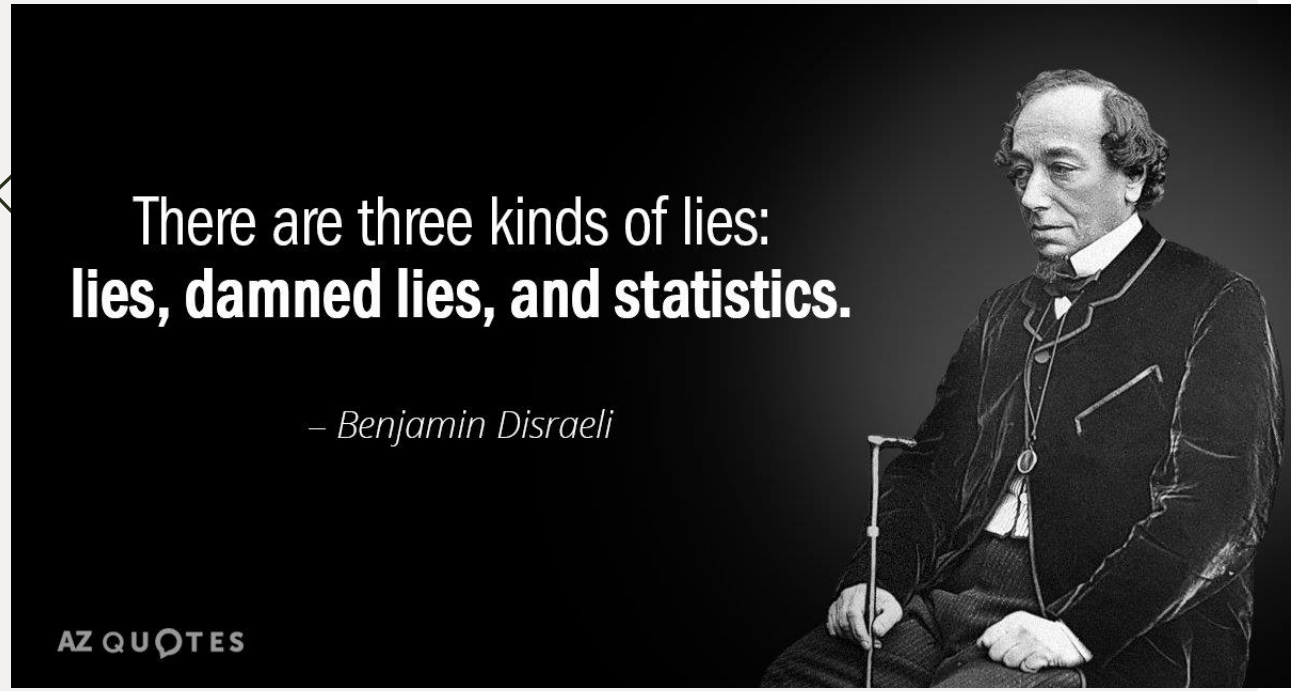

1
n
Ranked Gene List (L)


Cumulative distribution function
Cumulative Probability [$F_n(x)$]
1.0
0.5
0
1
n
Data (x)
Enrichment Score (S_k





GO
GENE ONTOLOGY



THE END

More Best practices for omics


git
+
GitHub

General directory suggestion

Scripts

Input_type1

Input_type2

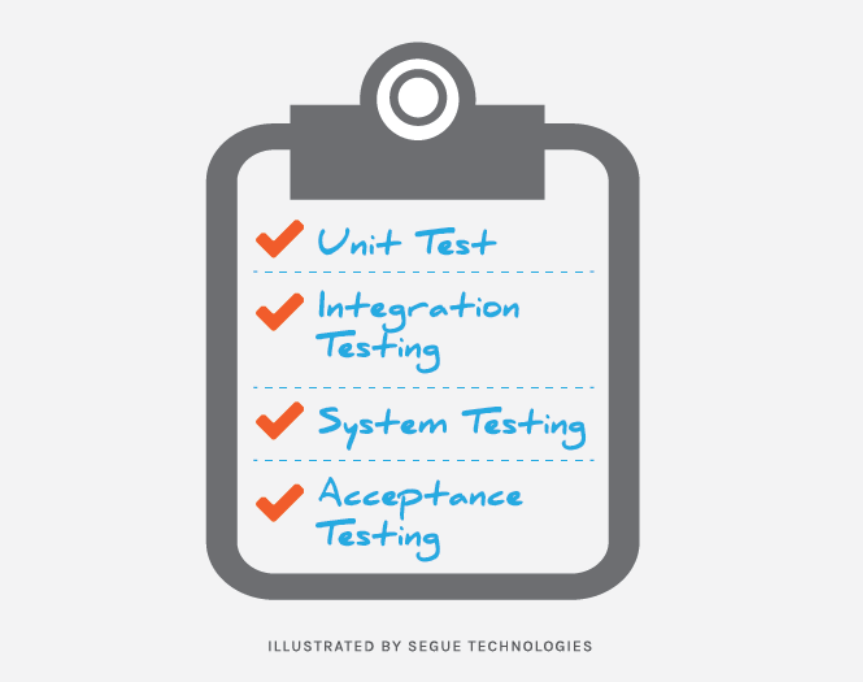
Output_type1

Output_type1

Output_type1

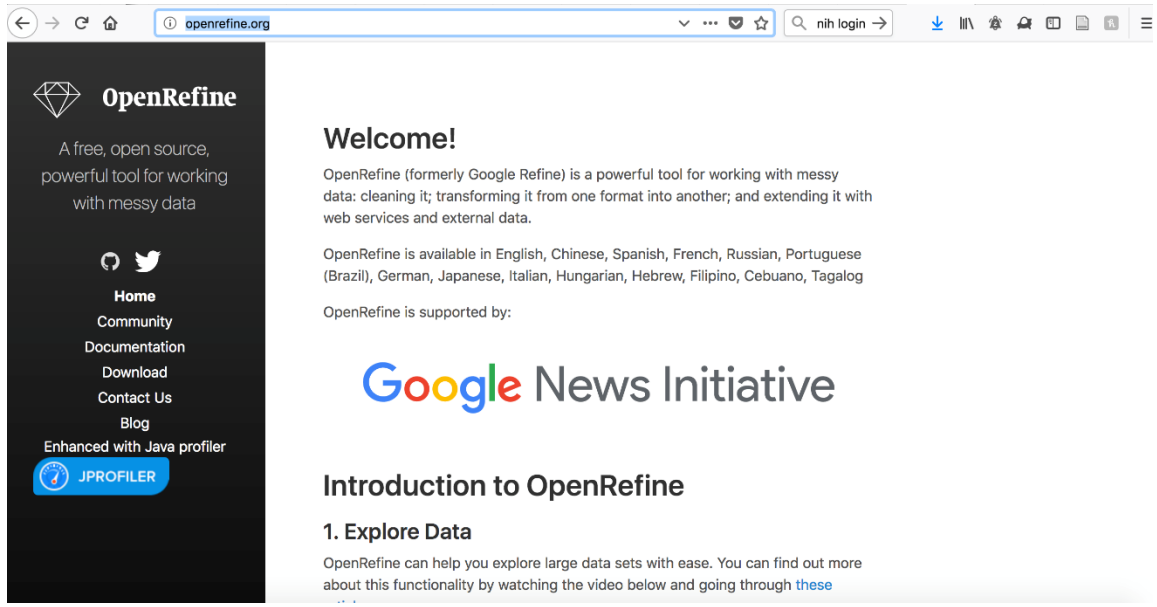
I have been changed for good

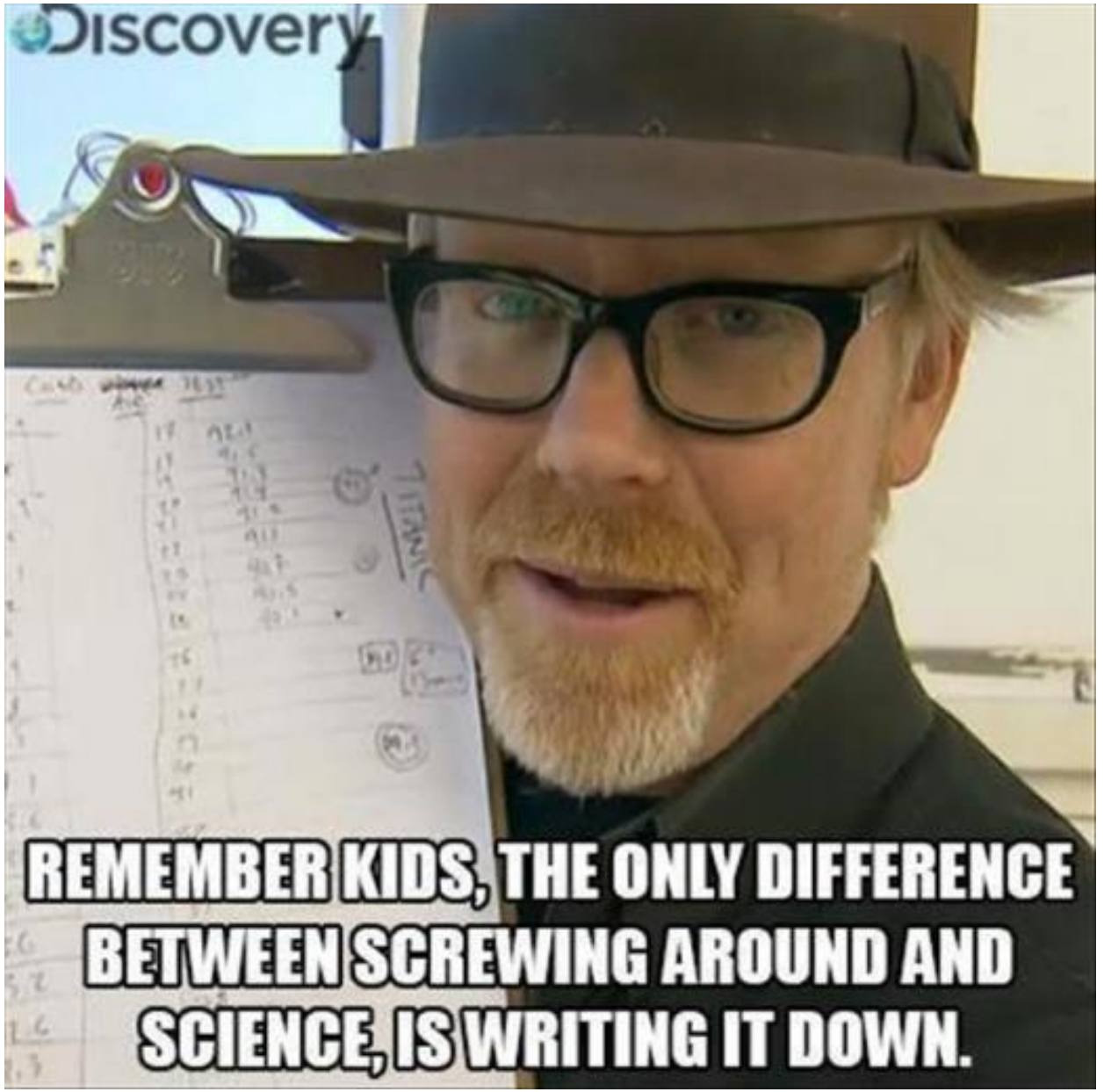
How to write a production-level code in Data Science?



Choosing slurm options

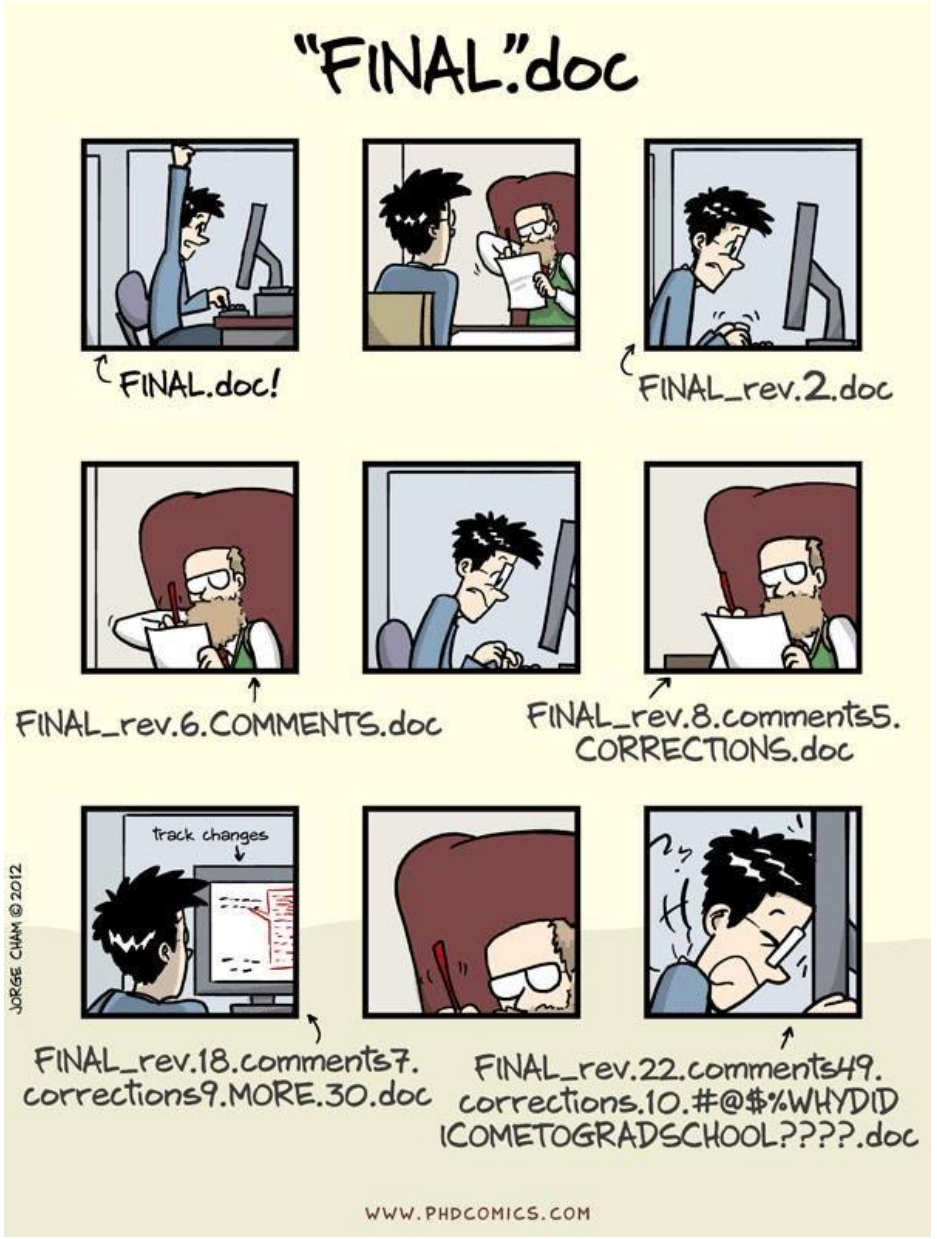
Clean responsibly




Discovery
REMEMBER KIDS, THE ONLY DIFFERENCE BETWEEN SCREWING AROUND AND SCIENCE, IS WRITING IT DOWN.

How understandable is your laboratory notebook?

Report the experiment (at every step!)

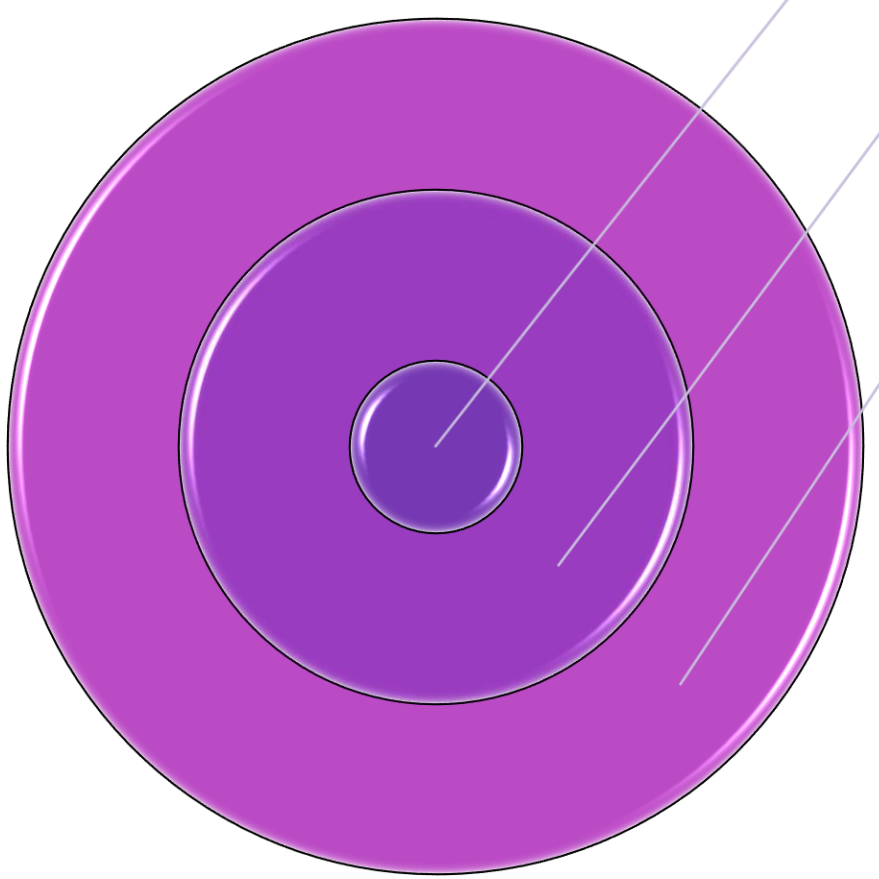


GitHub

CONDA

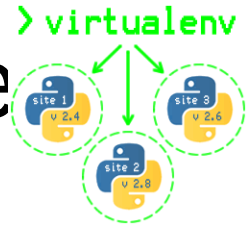
docker

The data, the code and the machine all need to be reproducible.



Machine

Code



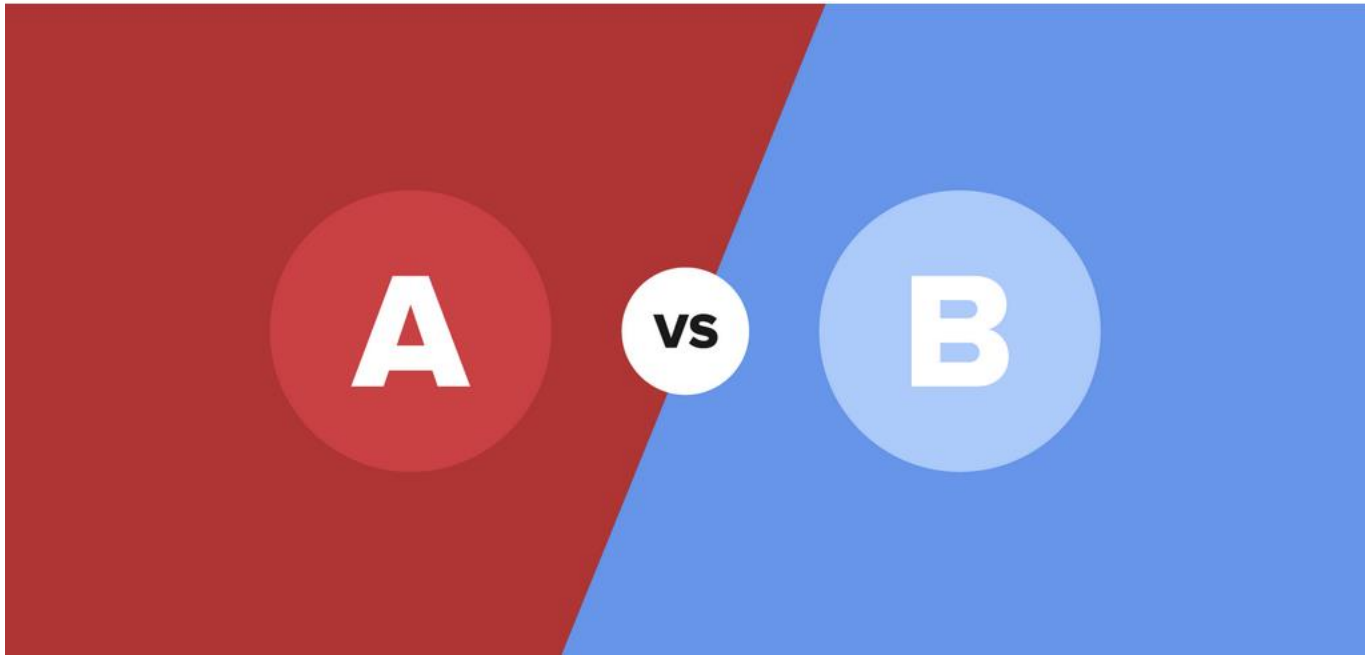
Data

Presenting data ethically

How do you know when your hypothesis may be correct?

"Data don't make any sense,
we will have to resort to statistics."

Is X different between A and B?



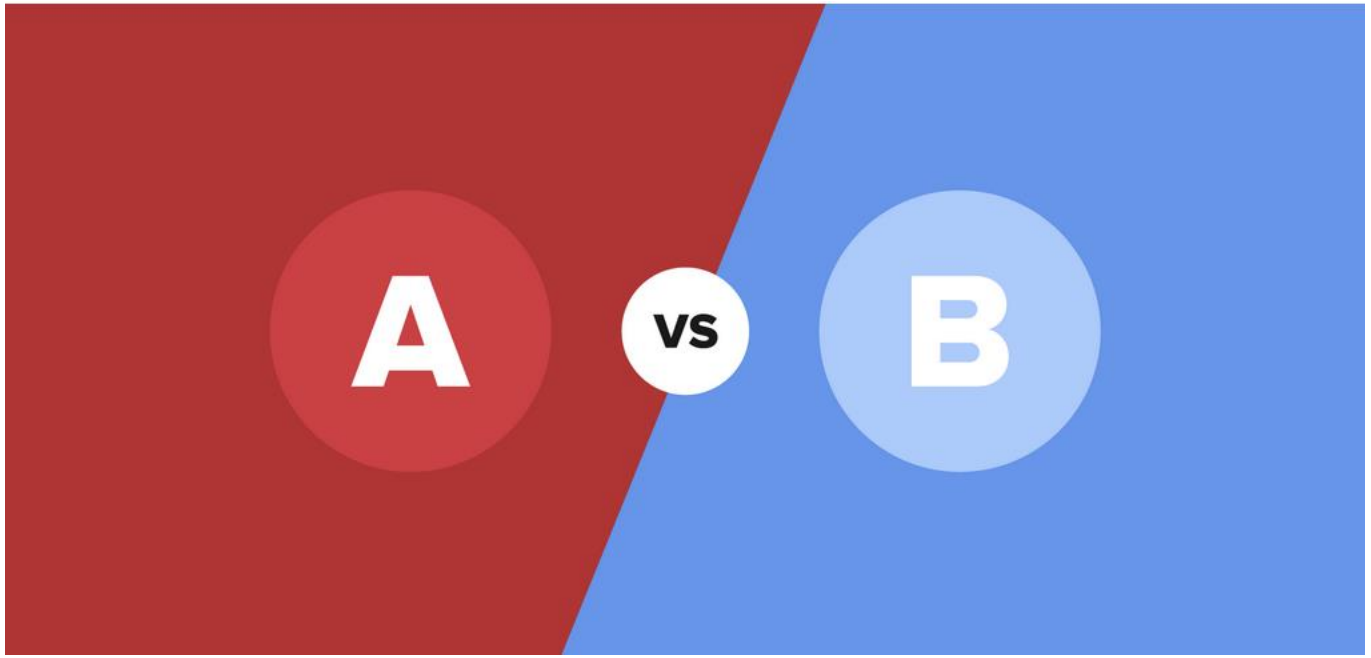
X	30	40

Is X different in A vs. B?

Yes

No

Is X different between A and B?



X	30	40

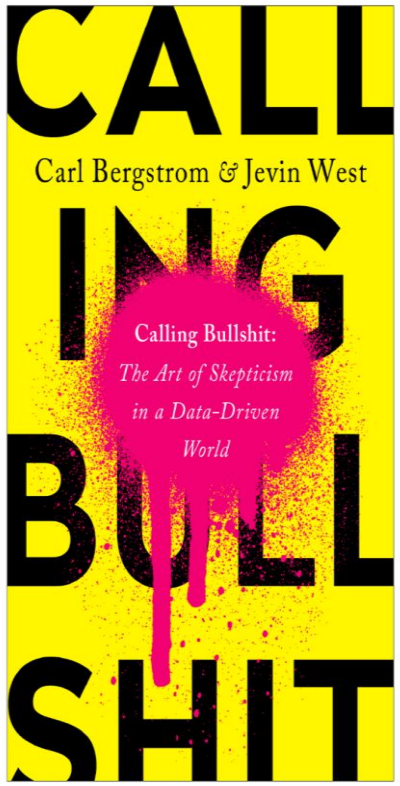
P-value is 0.04

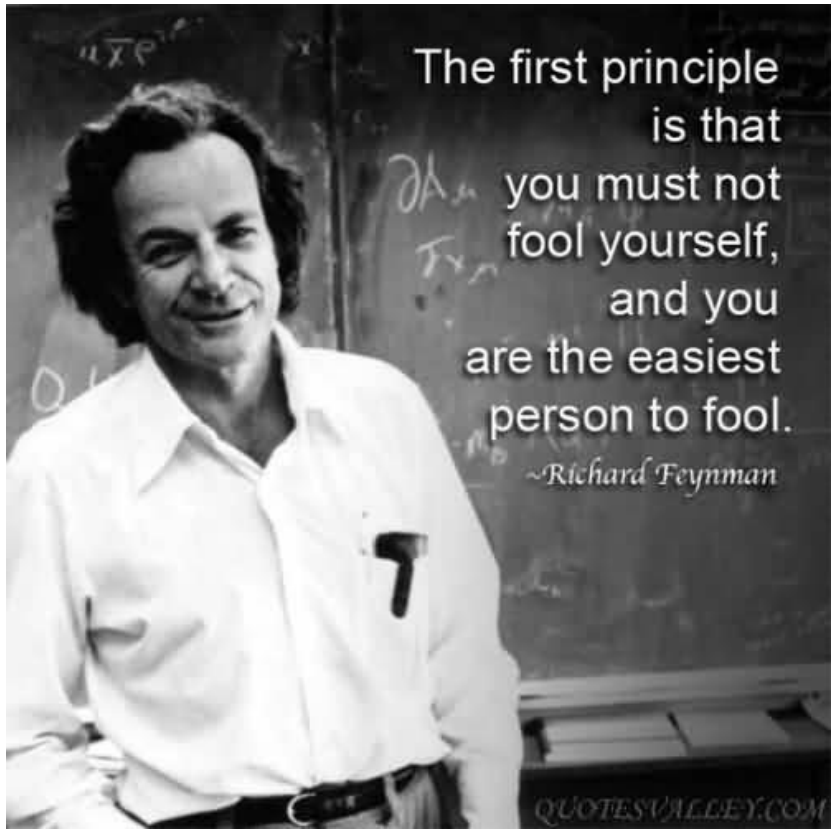
Is X different in A vs. B?

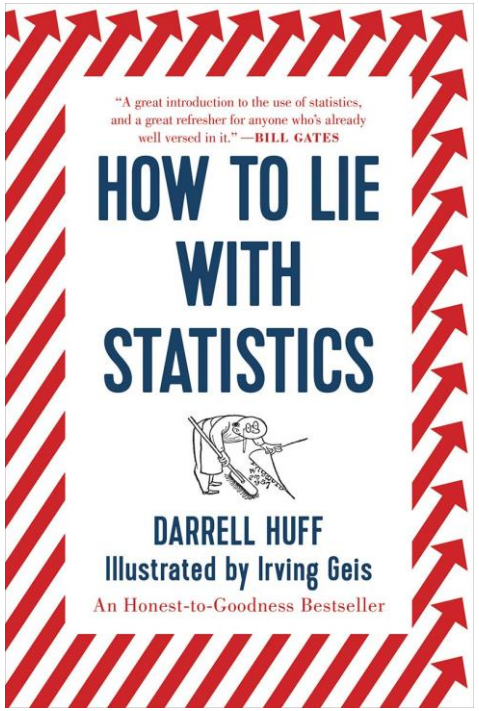
Yes

No

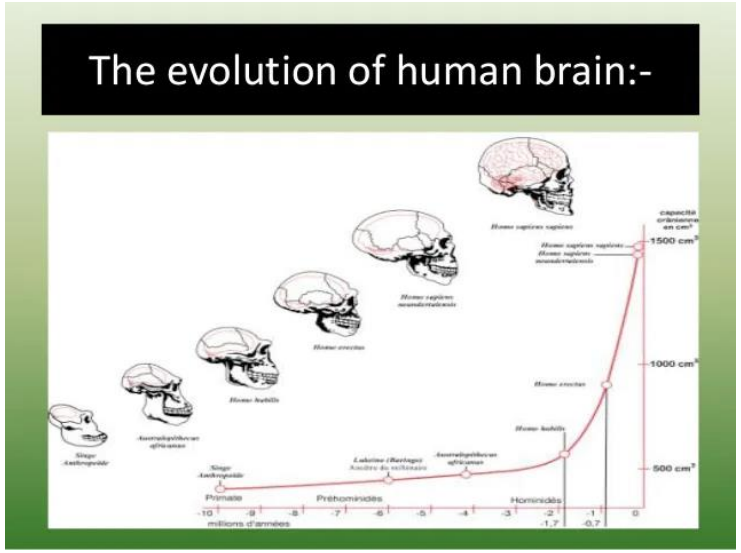
It is easy to fool people with data







Your brain did not evolve to do statistics

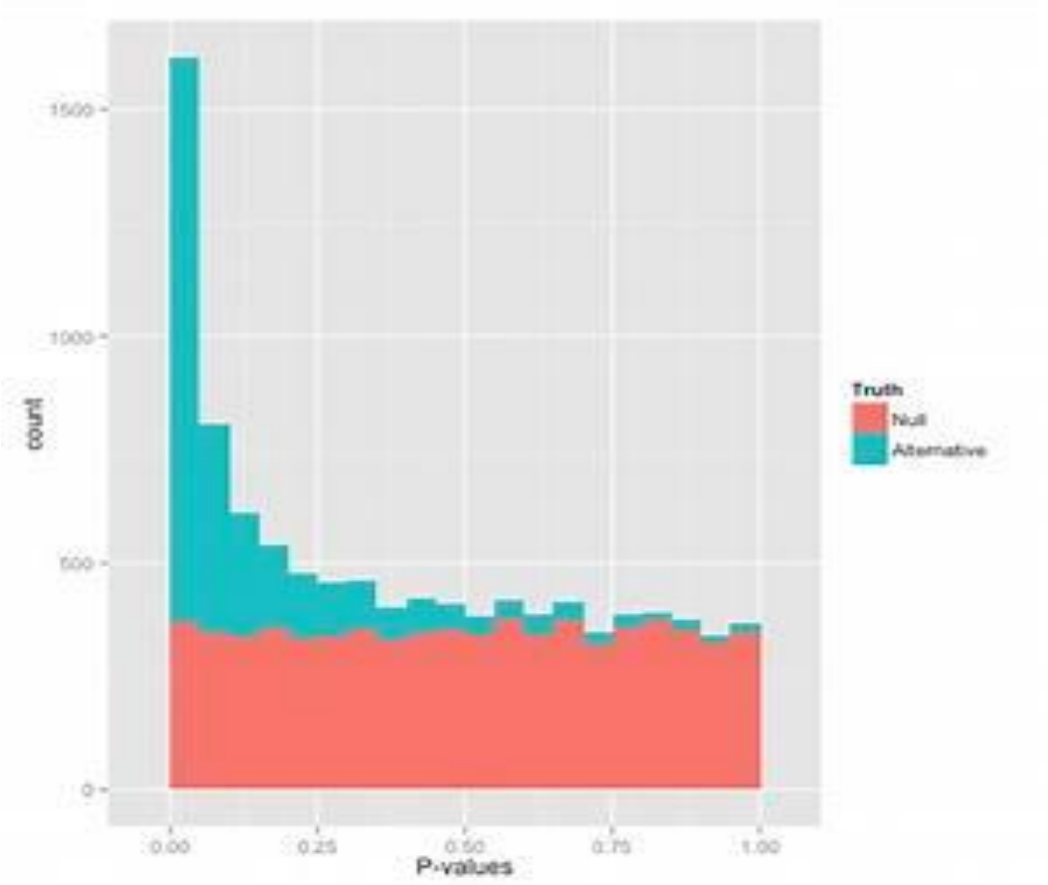


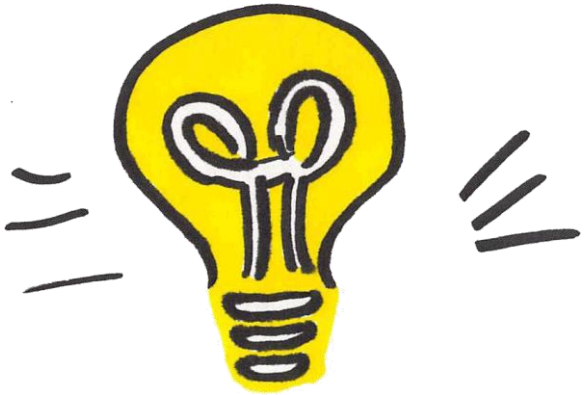
- We jump to conclusions
- We tend to be over-confident
- We see patterns in random data
- We don't realize coincidences are common
- We avoid thinking about ambiguous situation
- We are fooled by multiple comparisons
- We ignore alternative explanations
- We are fooled by regression to the mean
- We have incorrect intuitions about probabilit

Motulsky _Intuitive Biostatistics_ Oxford Univ Press 2010

What is a p-value?

What is a p-value?

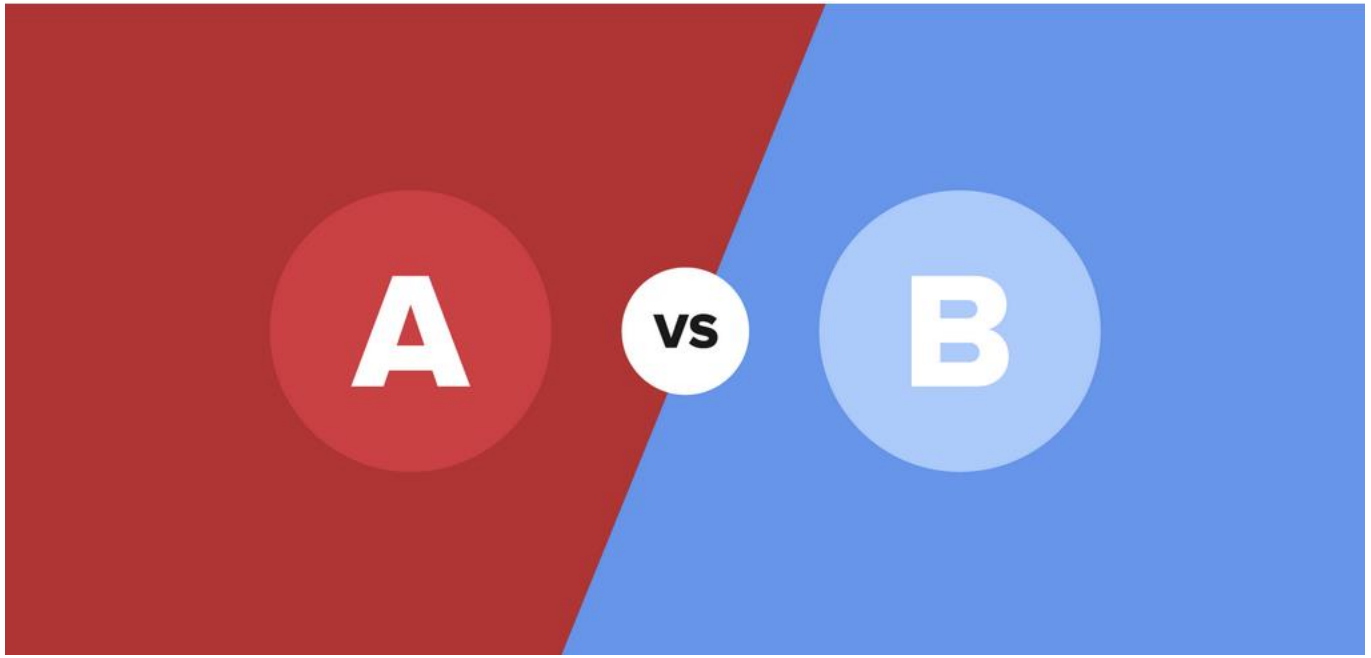




One hypothesis

Multiple hypothesis

Is X different between A and B?



I tested
X1
X2
X3
….
X200,000

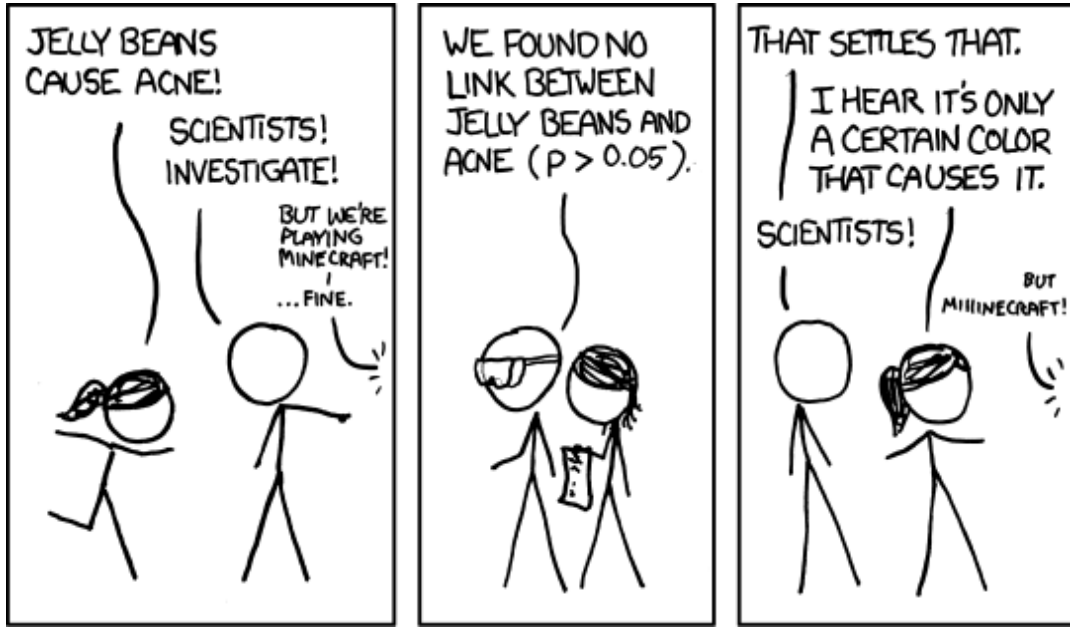
X1 30 40

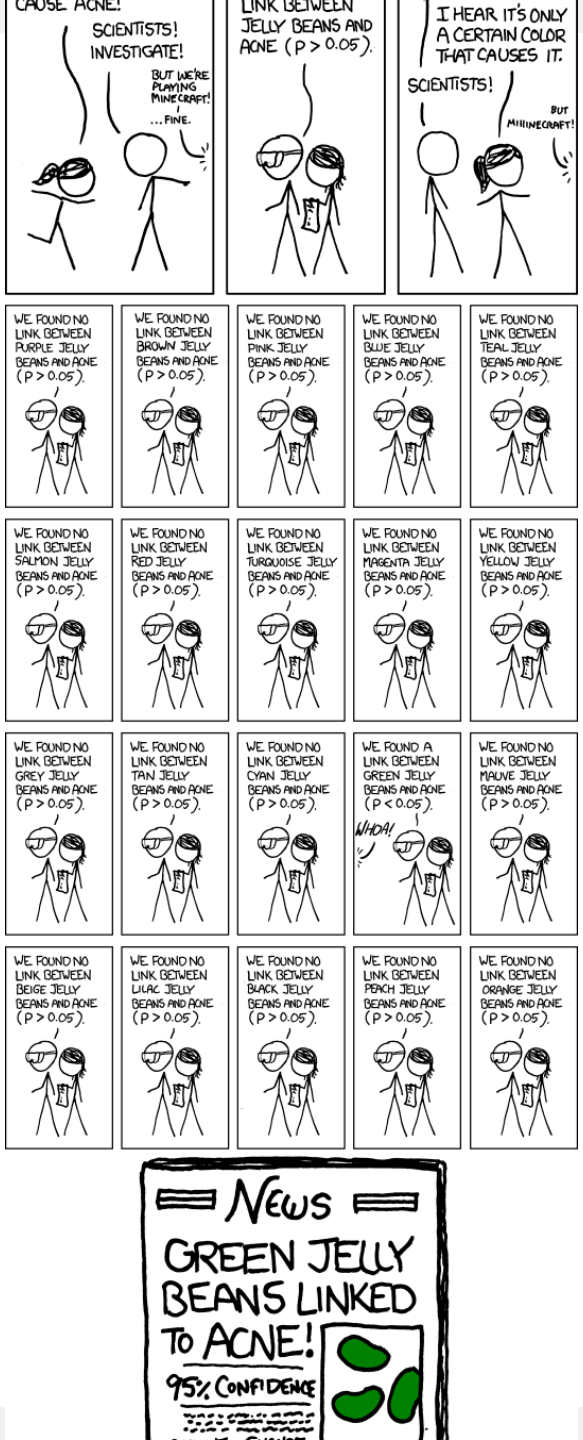
P-value is 0.04

Is X different in A vs. B?

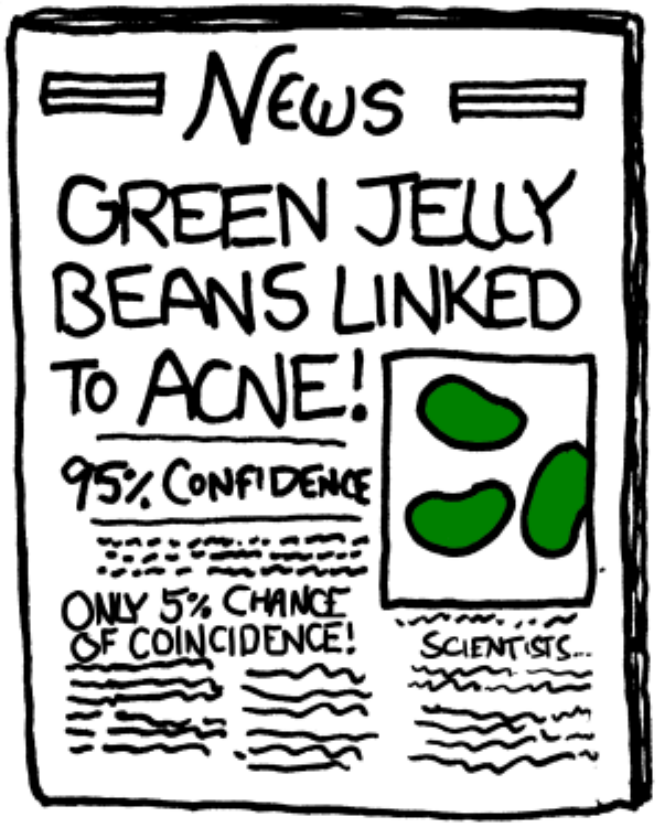
Yes

No


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MILLINECRAFT!


SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MILLINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

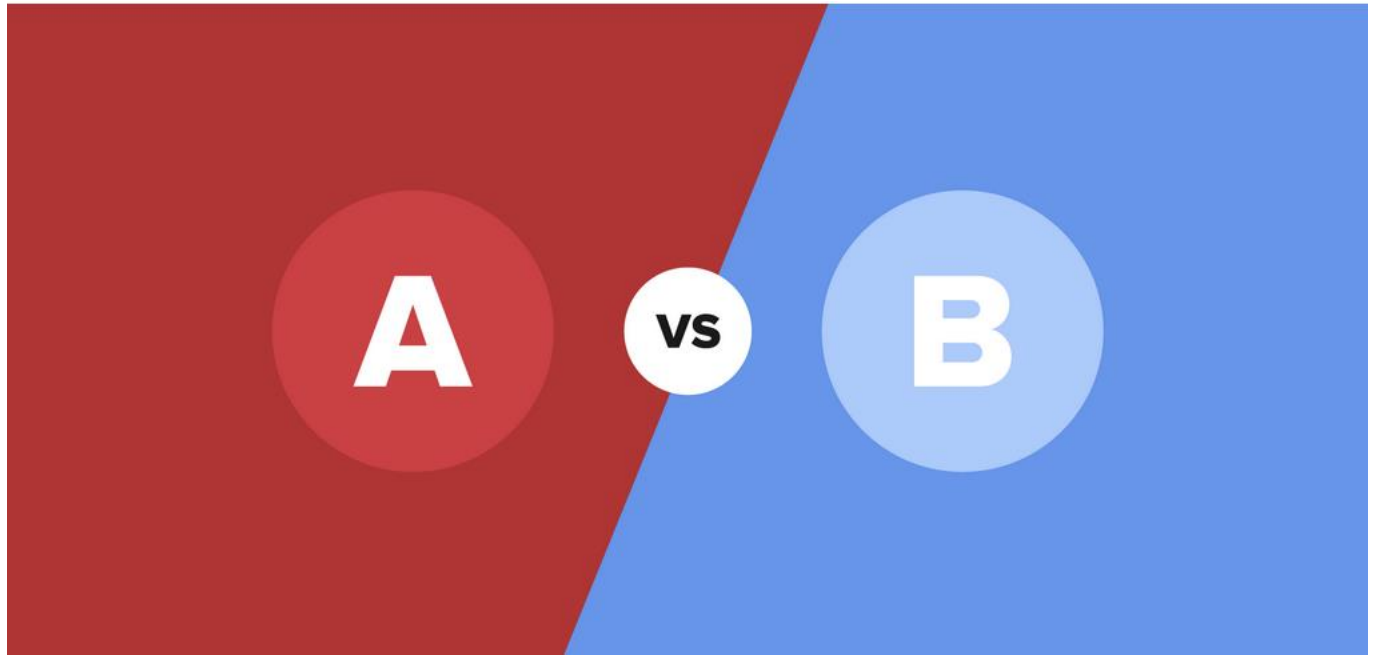
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...


MISTAKES
are proof that you are
TRYING

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Is X different between A and B?



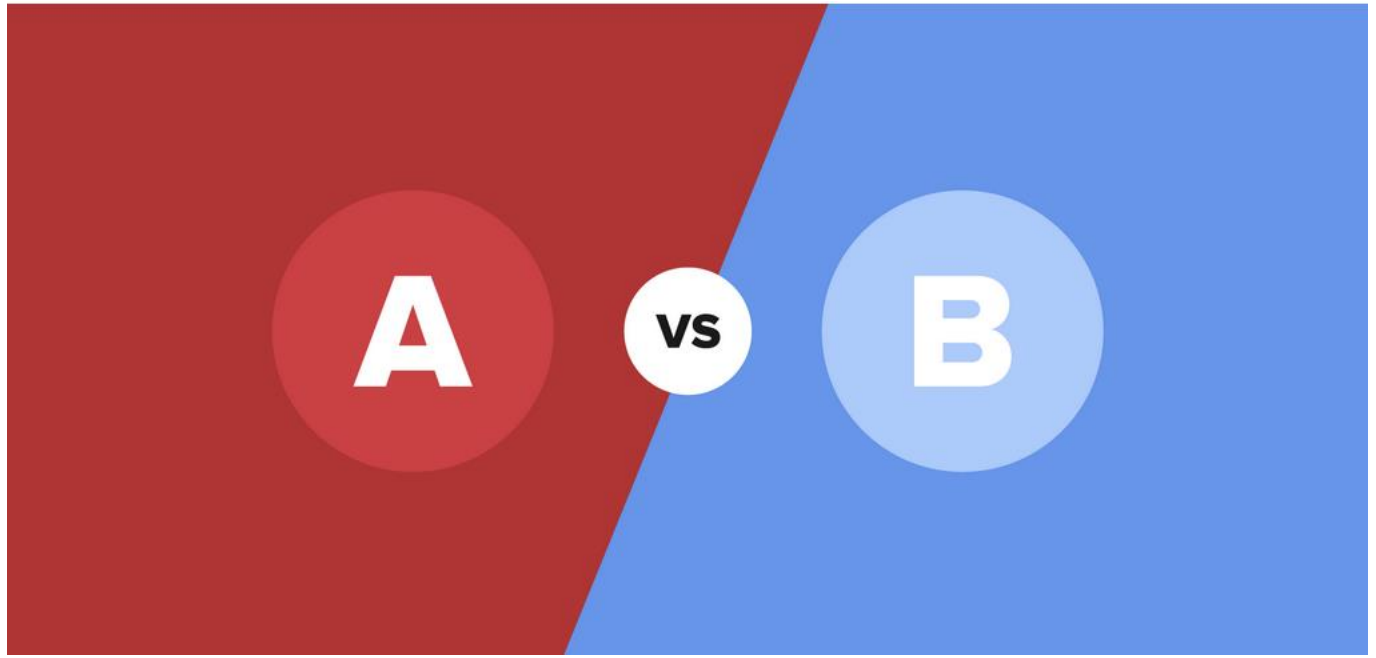
I tested	p-value
X1	0.01
X2	1
X3	.05
….	
X200,000	1e^1

X1 30 40

P-value is 0.04

What is an adjusted p-value?

Is X different between A and B?



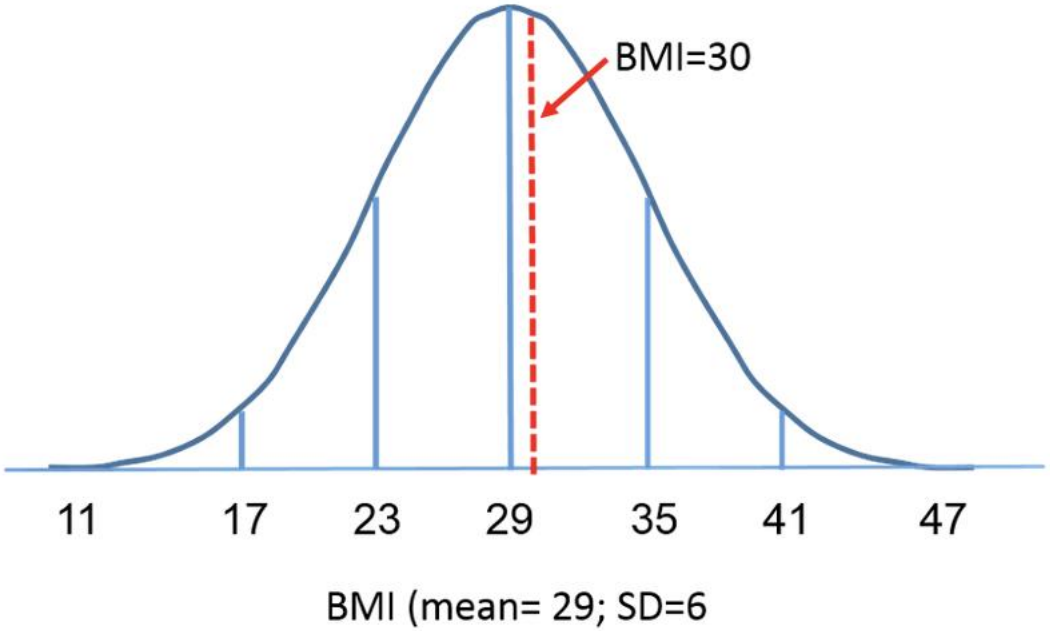
I tested	p-value
X1	0.01
X2	1
X3	.05
….	
X200,000	1e^1

X1 30 40

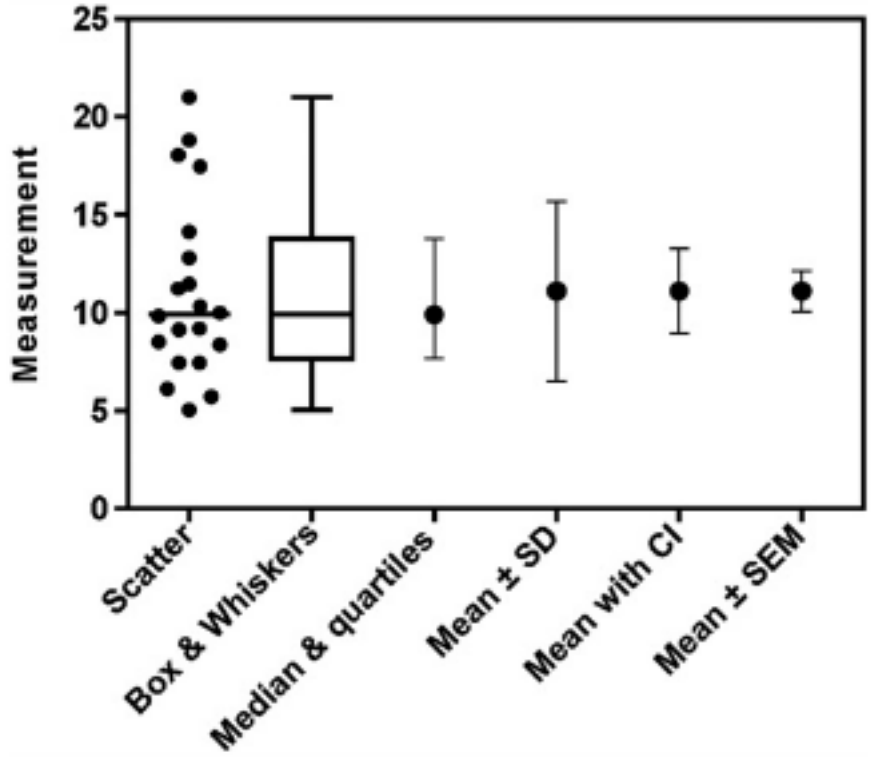
P-value is 0.04

Is X different between A and B?

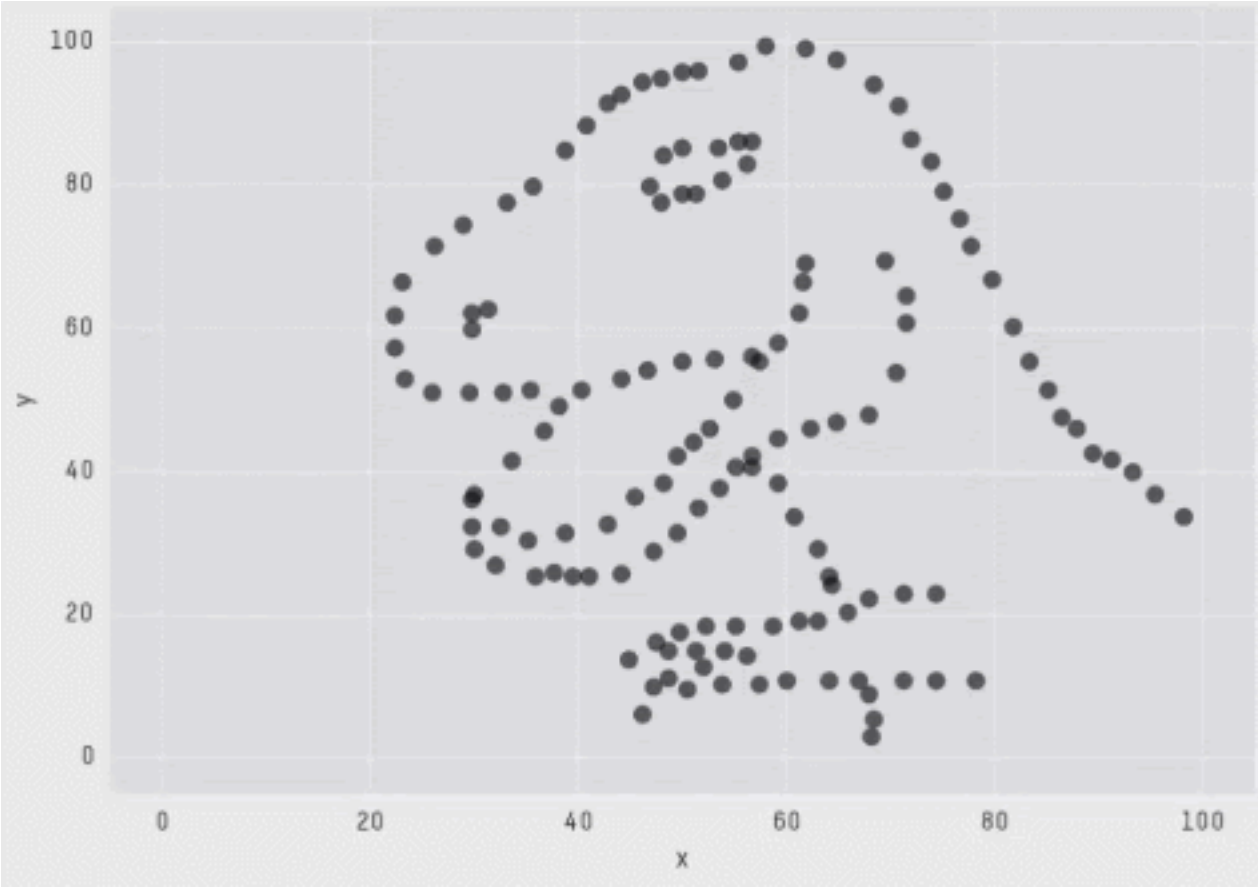
Remember: A value is always comes from a distribution.



Averages can be very misleading!



Maybe you should look at your data?

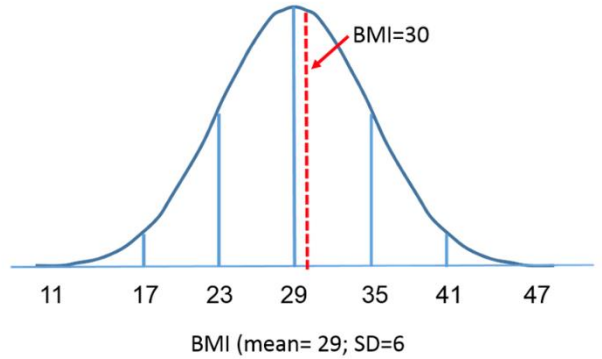


X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

https://www.autodeskresearch.com/publications/samestats

- https://www.youtube.com/watch?v=Hz1fyhVOjr4

How many replicates do I need?



POWER ANALYSIS

Start

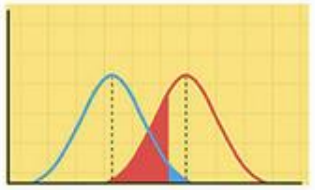
$H_0: \mu = \mu_0$
$H_0: \mu \leq \mu_0$

Determine if your hypothesis test is one-tailed or two-tailed

Determine the research design for the hypothesis

α (.01, .05)
β (.01, .05)

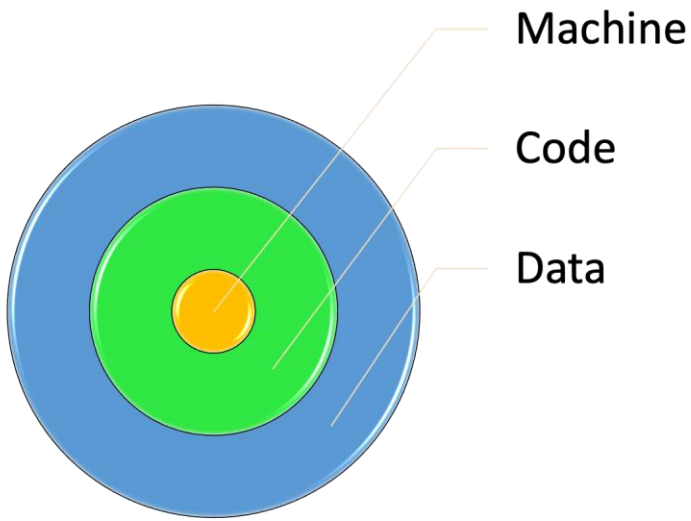
Determine the probability of finding a statistically significant result (i.e., power)



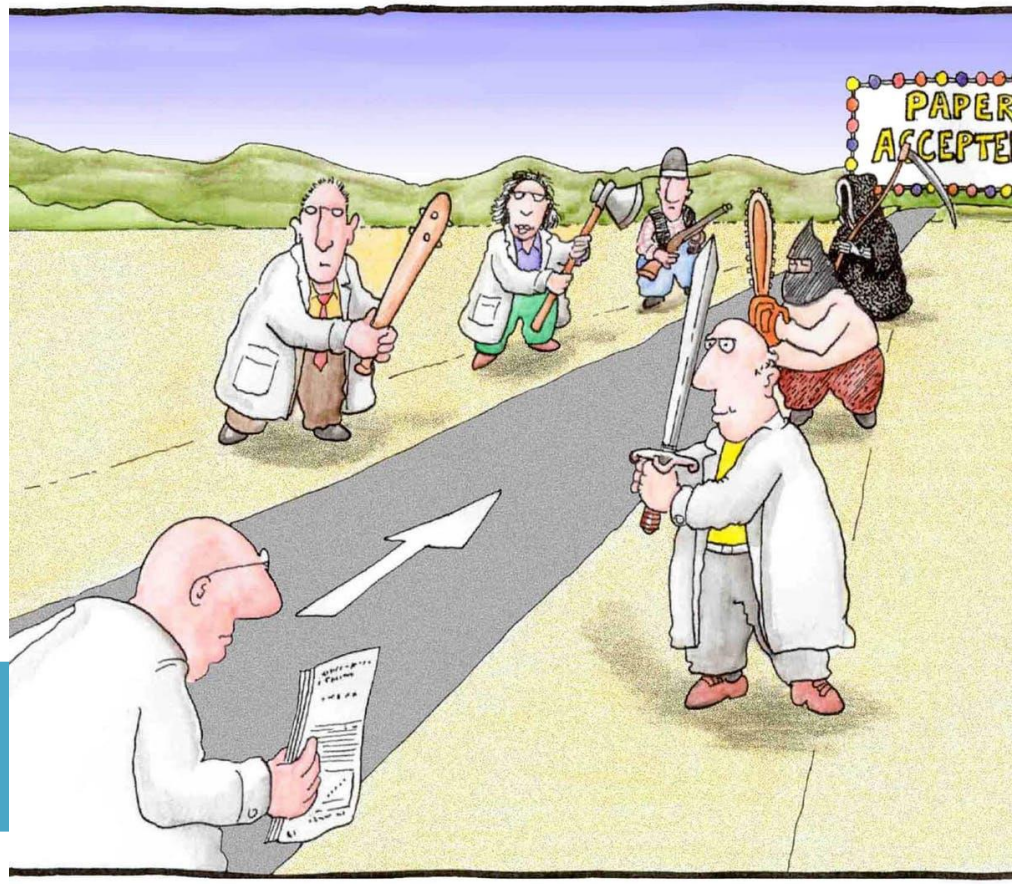
Determine the threshold for finding a meaningful result (i.e., effect size)

Using those inputs, calculate the minimum sample size

End



Data → Code → Publish

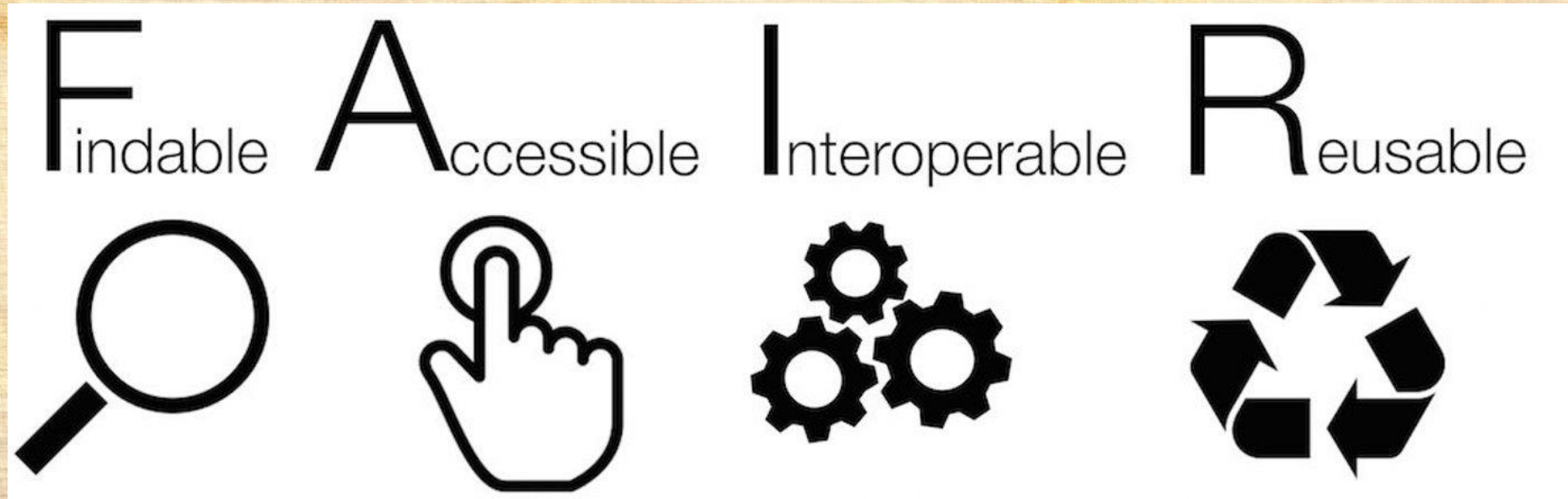


Most scientists regarded the new streamlined peer-review process as "quite an improvement."

Error in one line of code sinks cancer study

Authors of a 2016 cancer paper have retracted it after finding an error in one line of code in the program used to calculate some of the results.

Sarah Darby, last author of the now-retracted paper from the University of Oxford, UK, told *Retraction Watch* that the mistake was made by a doctoral student. When the error was realized,


Findable
Accessible
Interoperable
Reusable

PROS AND CONS OF PIPELINES

nextflow

NLM vs. clean Metadata