

Big Data Ethics

Data Rigor and Reproducibility
And what do I do with my gene list

Mary Allen





This Photo by Unknown author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/).

Data Data Data

Genetic

Omic

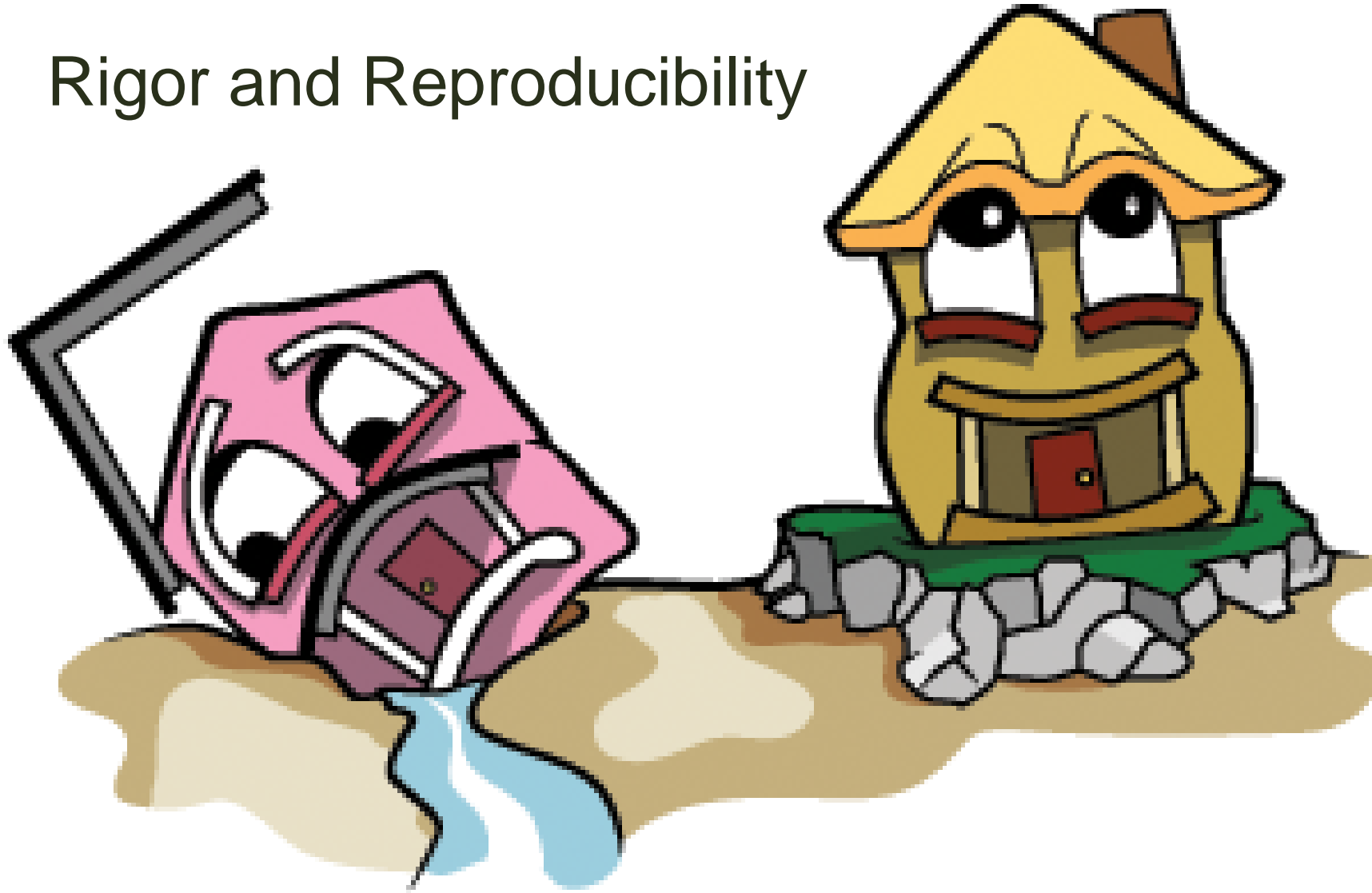
Evolution

System

Images

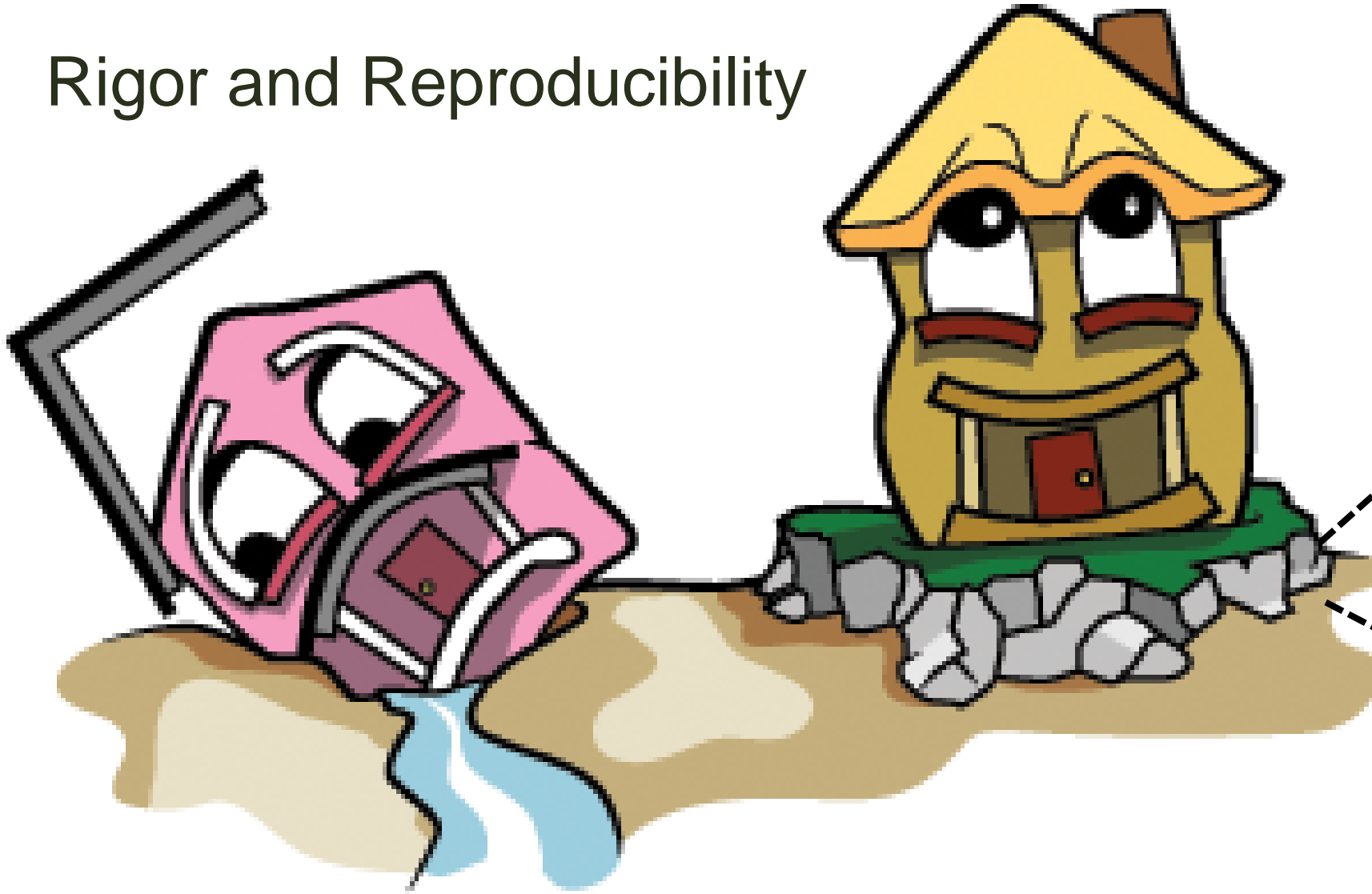


Rigor and Reproducibility



[This Photo](#) by Unknown Author is licensed under [CC BY-NC-ND](#)

Rigor and Reproducibility

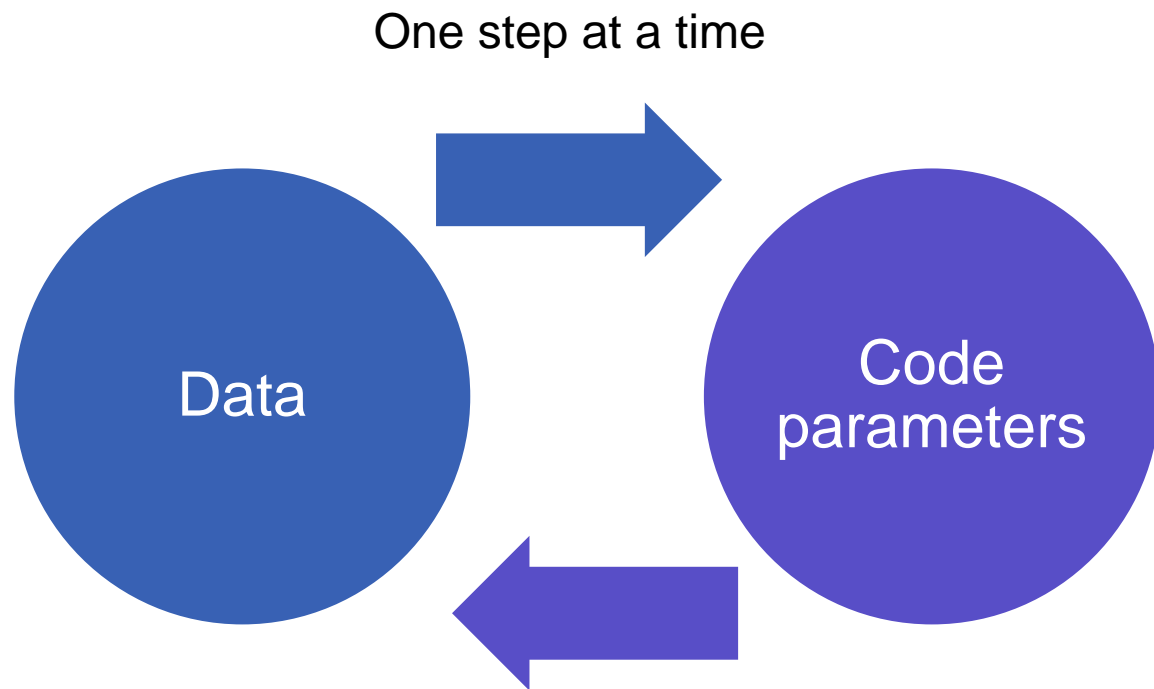


This Photo by Unknown Author is licensed under [CC BY-SA](#)



This Photo by Unknown Author is licensed under [CC BY-NC-ND](#)

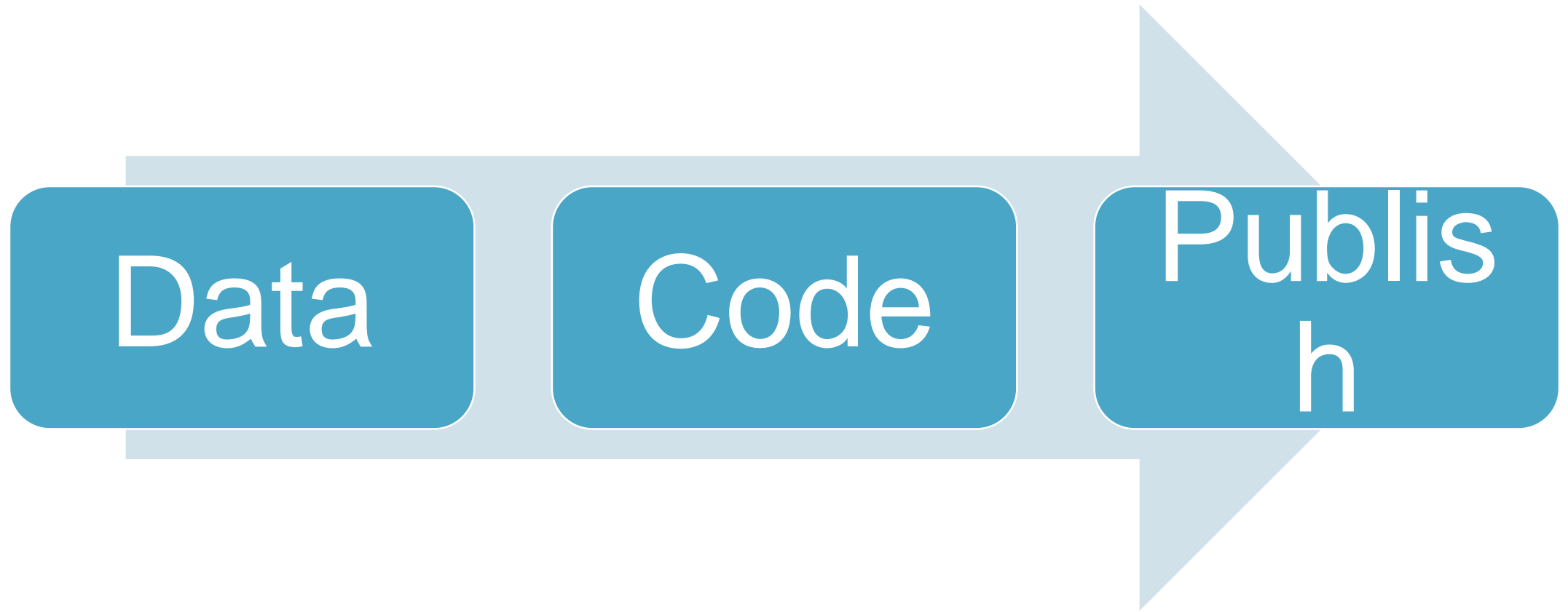
This Photo by Unknown Author is licensed under [CC BY-NC-ND](#)



Pipelines

Nextflow
CWL

How we think it works....



Before you start your foundation



[This Photo](#) by Unknown Author is licensed under [CC BY-SA-NC](#)



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?



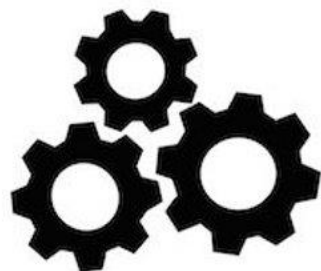
F
Findable



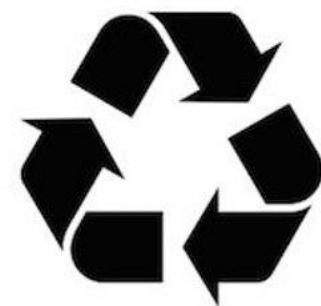
A
Accessible



I
Interoperable



R
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?



[This Photo](#) by Unknown author is licensed under [CC BY-NC](#).

[This Photo](#) by Unknown author is licensed under [CC BY](#).

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

```
shum@sol1:~$ ls -l
total 20
drwx----- 2 shum  staff  4096 Jan 16 22:04 Mail
drwx----- 3 shum  staff  4096 Jan 16 14:15 csc128
drwxr-xr-x  2 shum  staff  4096 Jan 13 16:42 public
drwxr-xr-x  2 shum  staff  4096 Jan 16 14:07 public_html
-rw-r--r--  1 shum  staff   628 Jan 15 20:04 verse
```

Annotations for the terminal output:

- file type (points to the first character of the permissions)
- user (owner) name (points to the user name)
- group name (points to the group name)
- size (points to the file size)
- date/time last modified (points to the date and time)
- filename (points to the file name)
- number of hard links (points to the link count)
- other (everyone) permissions (points to the last three characters of the permissions)
- group permissions (points to the middle three characters of the permissions)
- user permissions (points to the first three characters of the permissions)
- rwx breakdown:
 - rwx (points to the permissions string)
 - executable (points to the 'x' character)
 - writable (points to the 'w' character)
 - readable (points to the 'r' character)



General Data repositories for open data (AND BACKUPS)

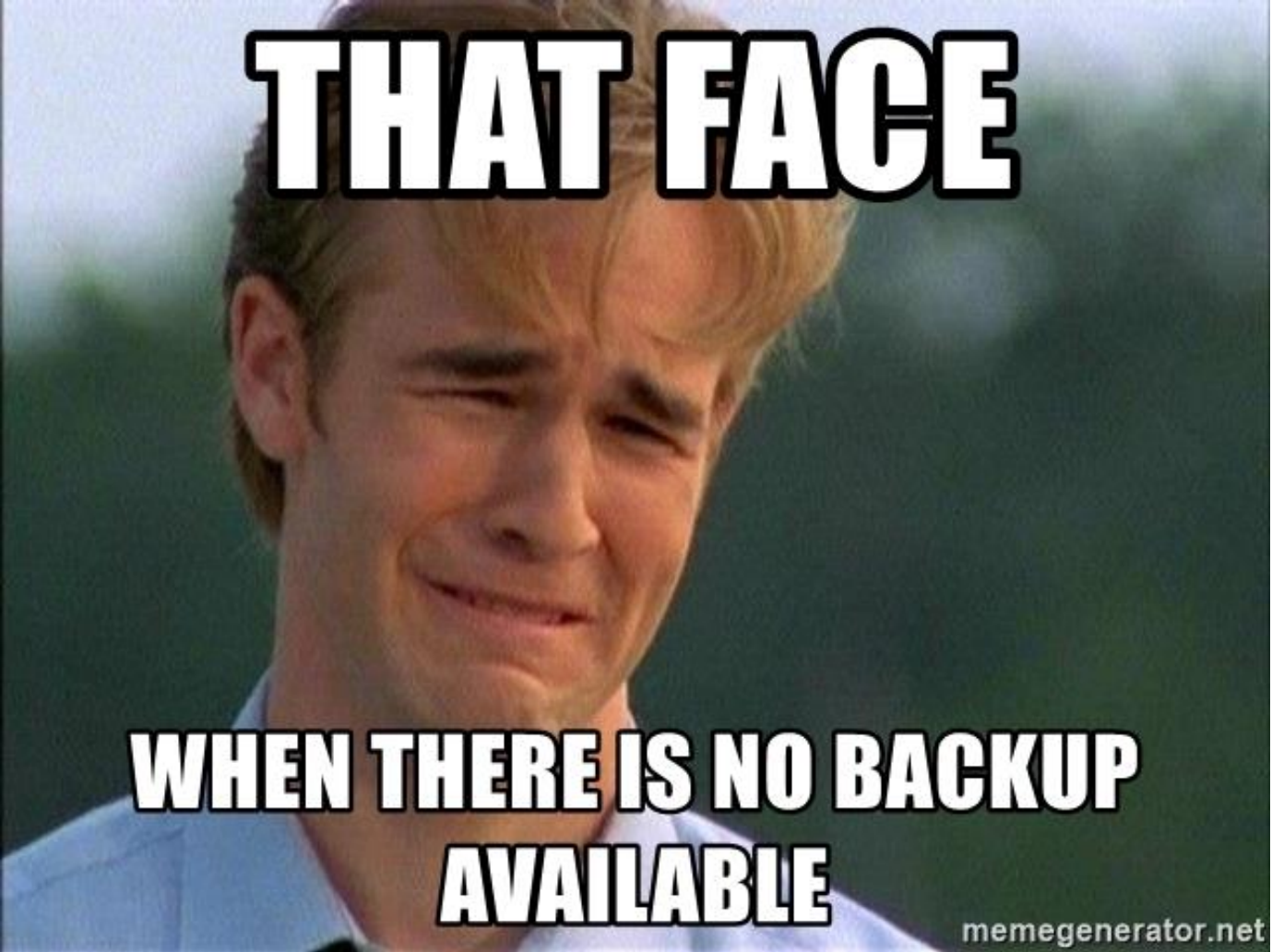


Back up that the government pays for?

[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 assess the quality of your 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 supported.

Get data from

Locate one or more files on your computer to upload:

This Computer

Browse...

No files selected.

Web Addresses (URLs)

Next »

Clipboard

Database

Google Data



Version 3.4.1 [437dc4d]

Preferences

Help

About



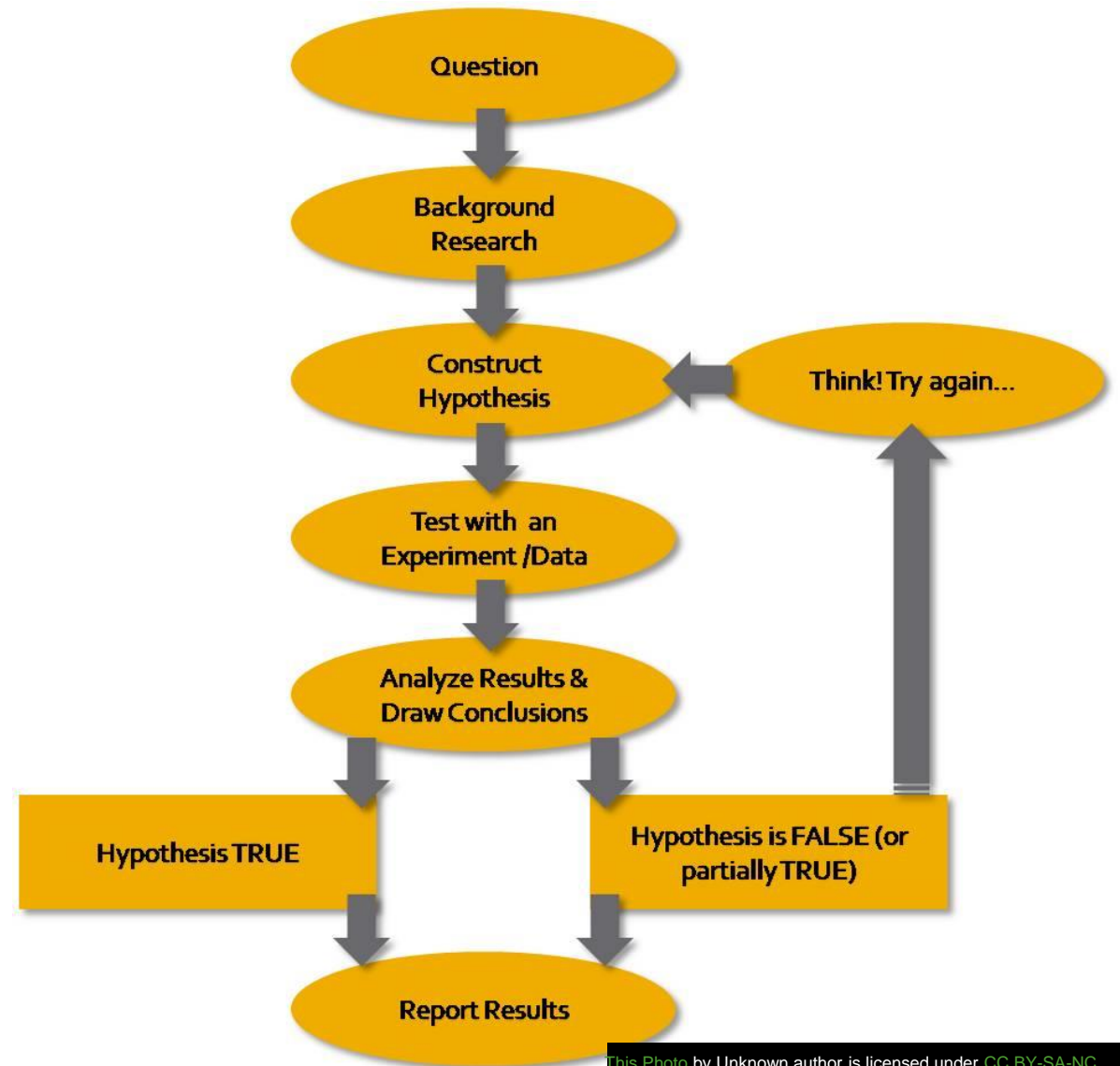
What will you use to clean dirty data?

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 ”

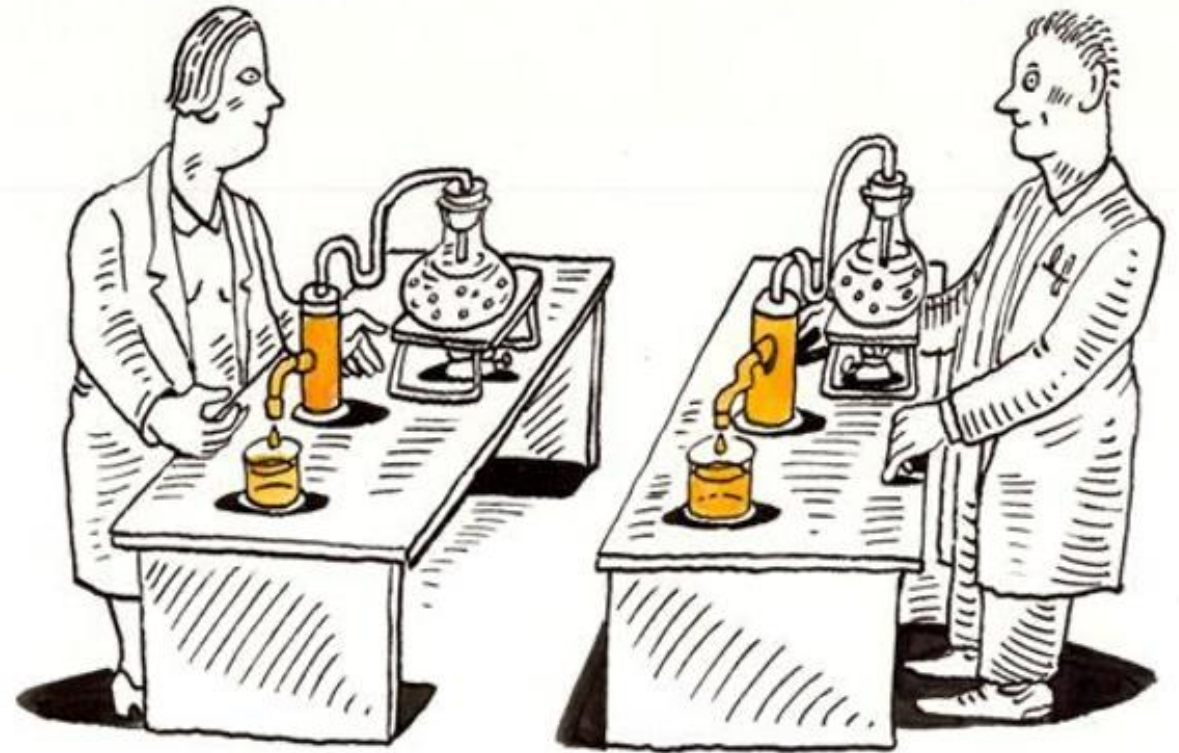
“ Liquid scintillation counter is temperamental and makes me cry ”

“ capitalism ”

Scientific reproducibility



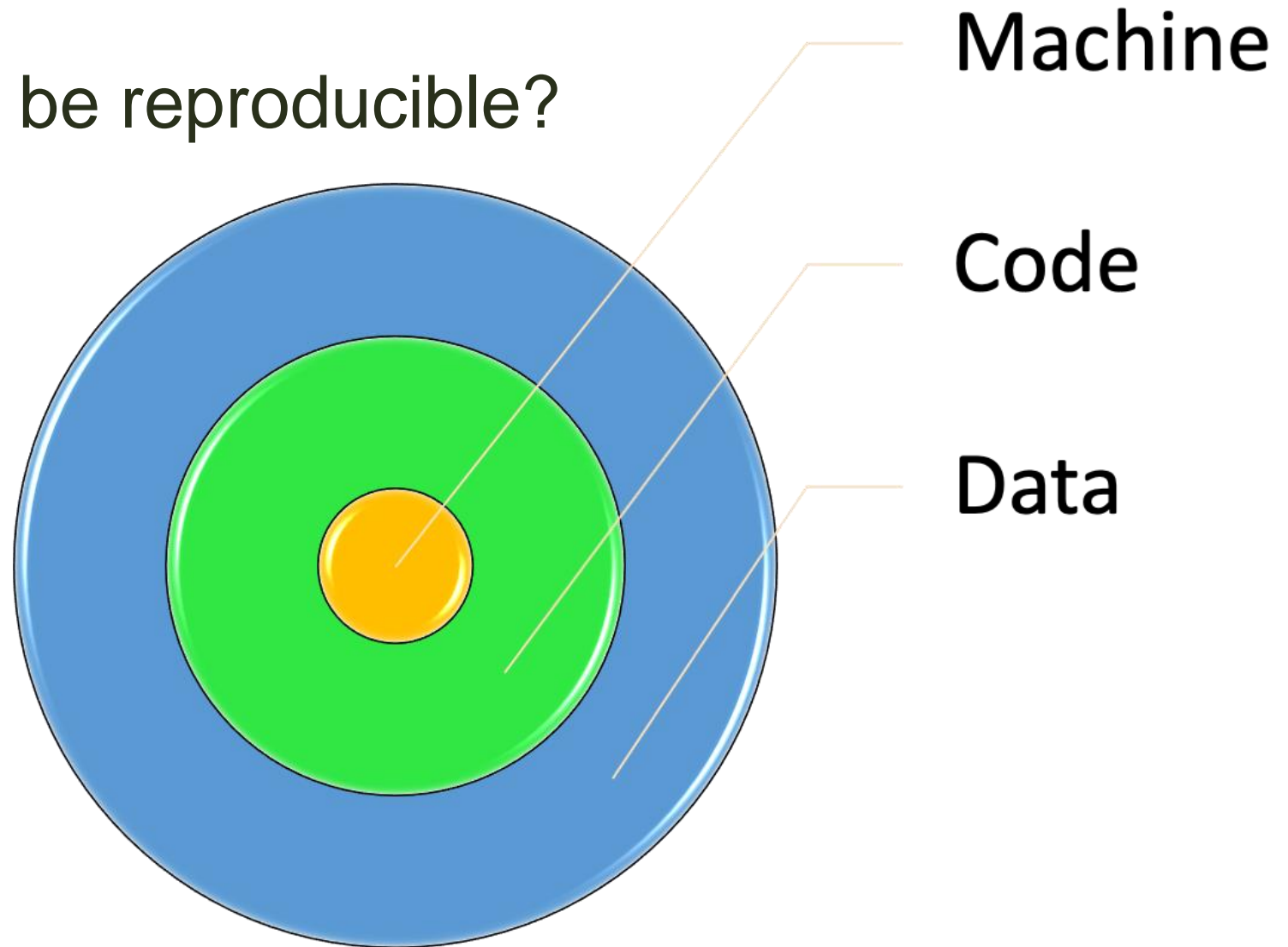
If someone else read my
publication, could they use the
same methods as me in another
time and place to get the same
results?



SAVING YOURSELF
TIME AND MONEY
VIA
REPRODUCIBILITY



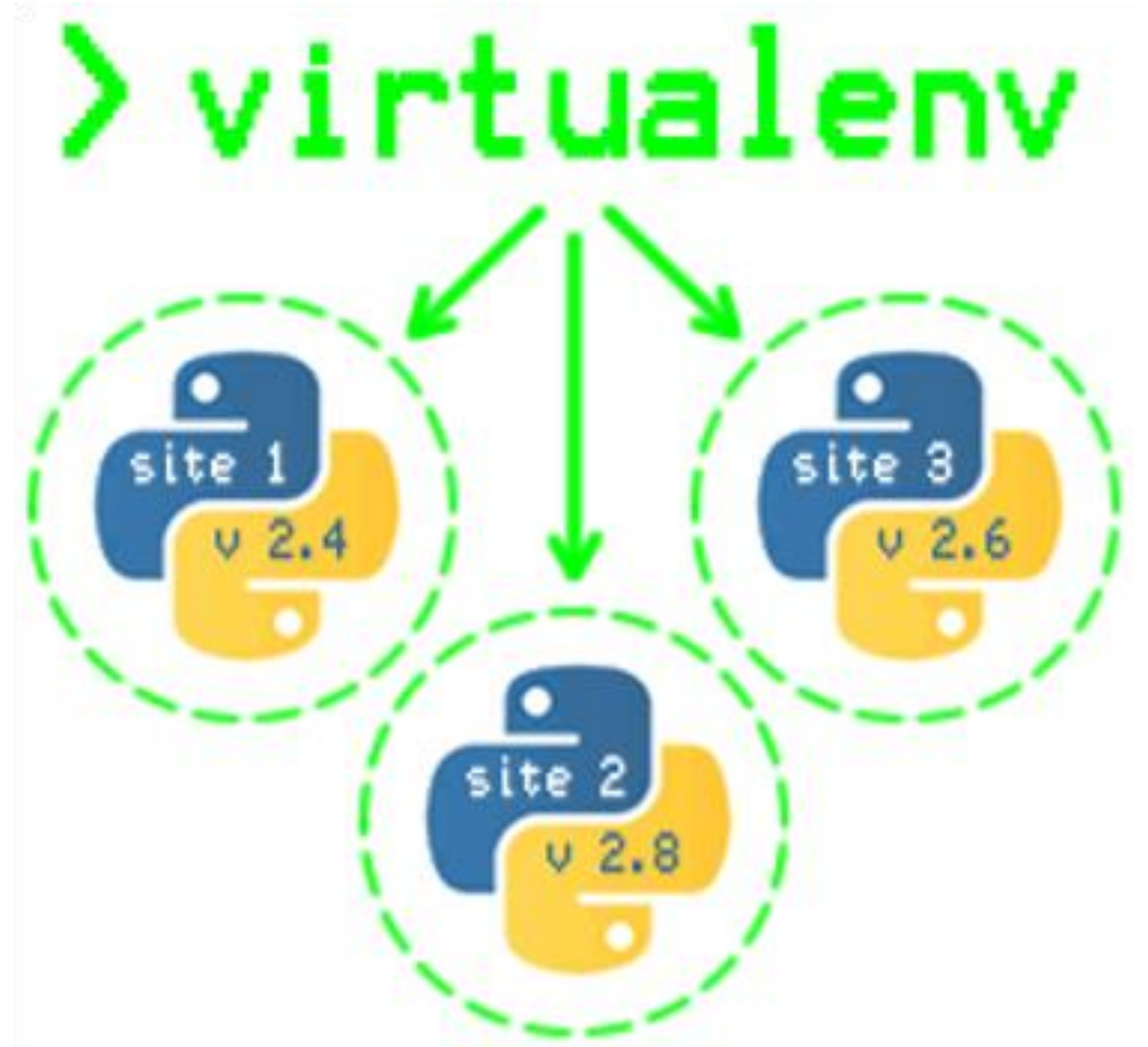
What needs to be reproducible?

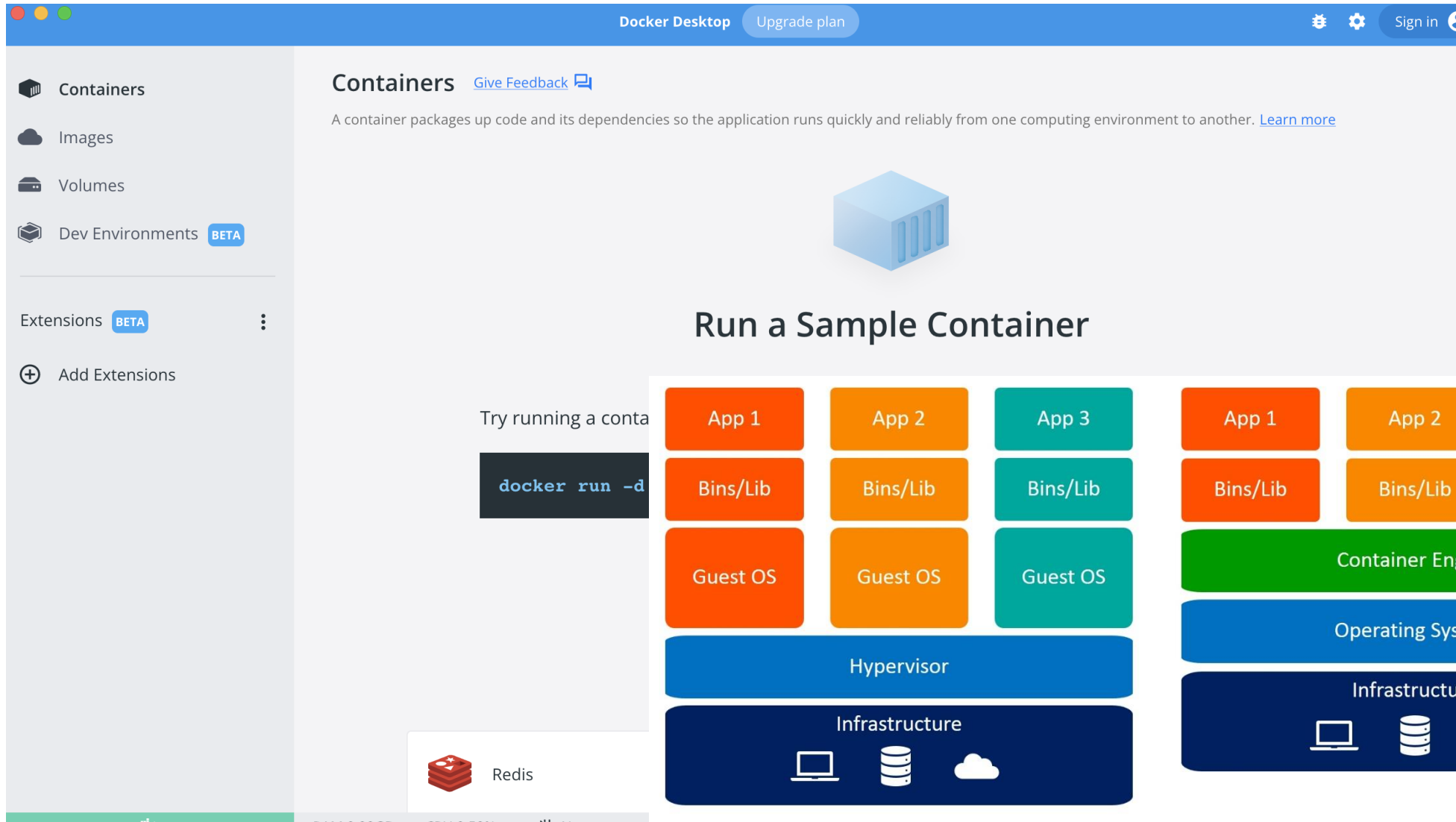


[This Photo](#) by Unknown Author is licensed under [CC BY-NC-ND](#)

“Machine” reproducibility

- Virtual machines
- Docker and other container tools

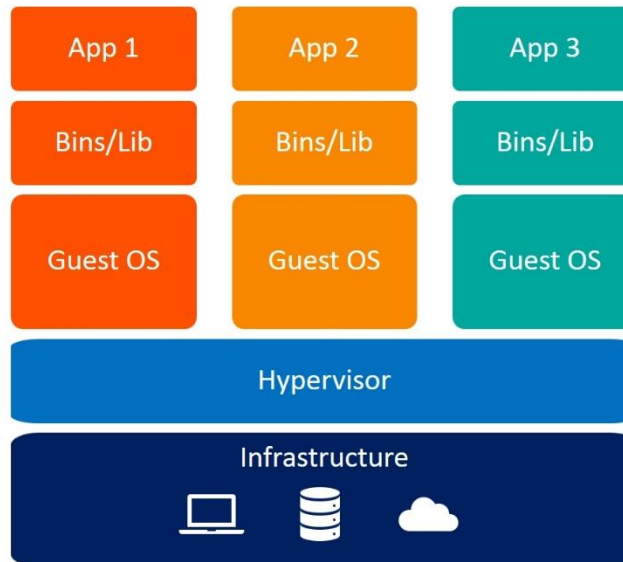




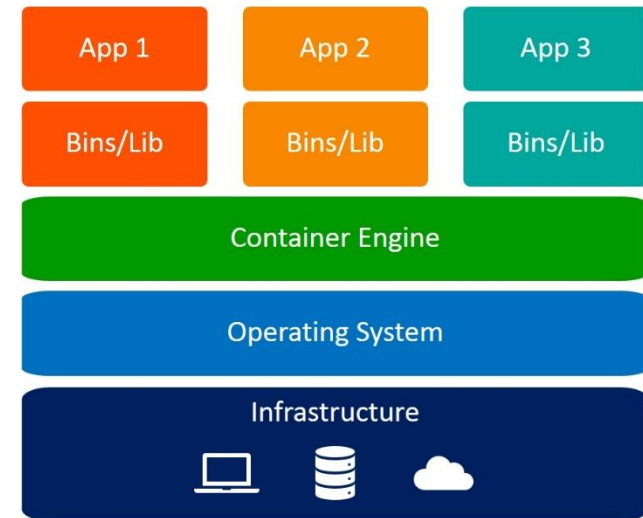
Run a Sample Container

Try running a container

```
docker run -d
```

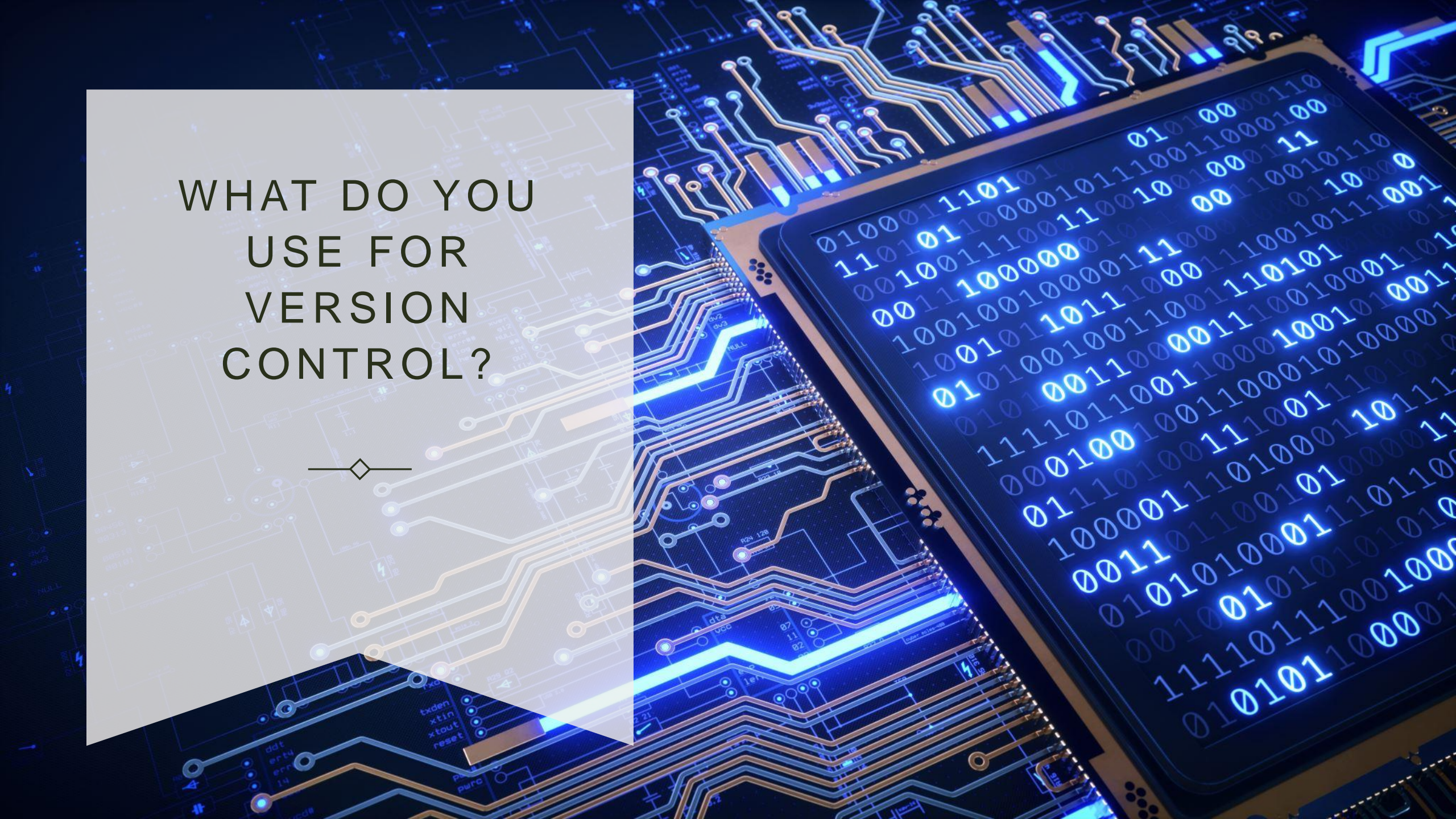


Virtual Machines



Containers

WHAT DO YOU
USE FOR
VERSION
CONTROL?



Code reproducibility



git

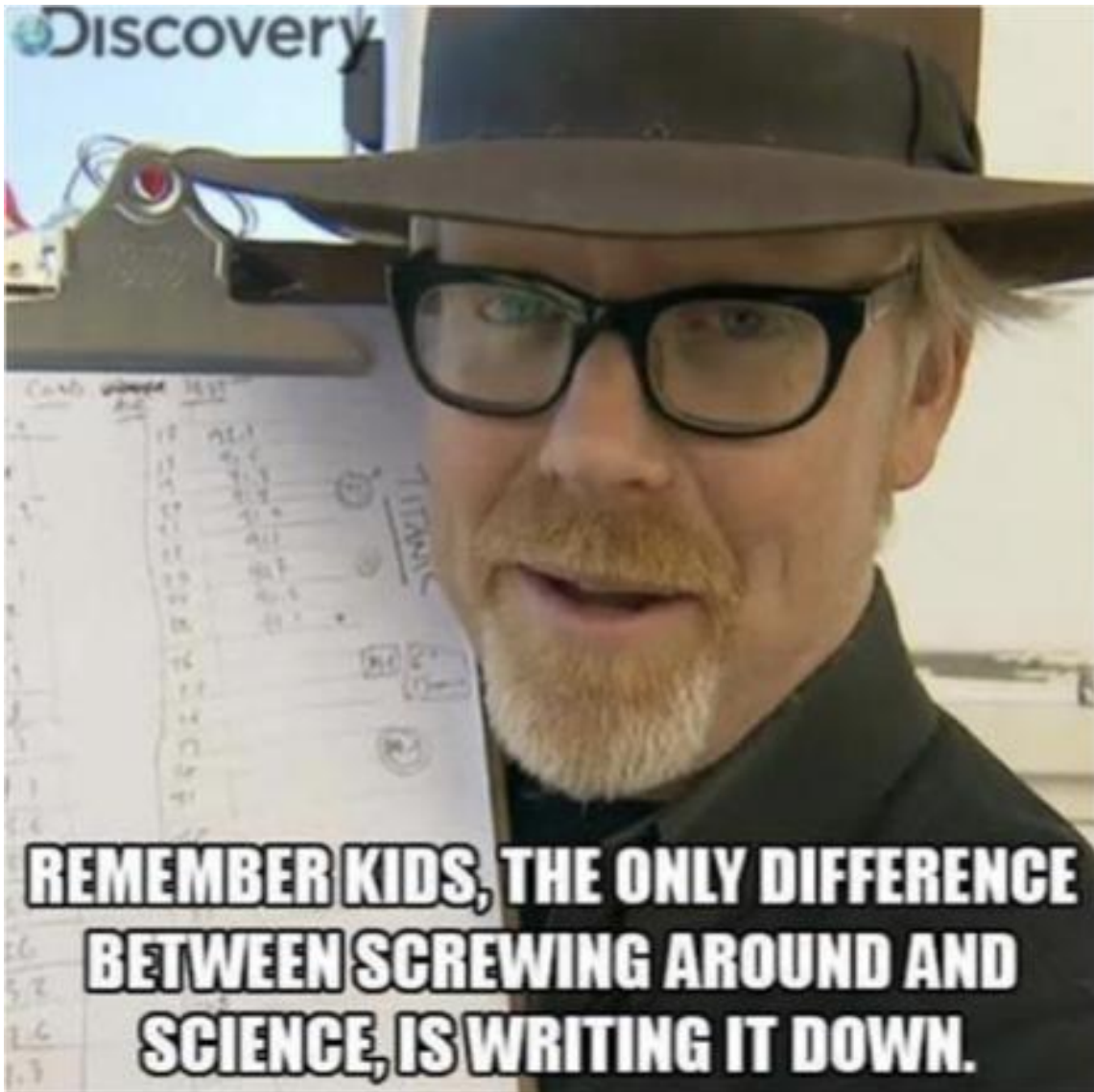


GitHub



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,

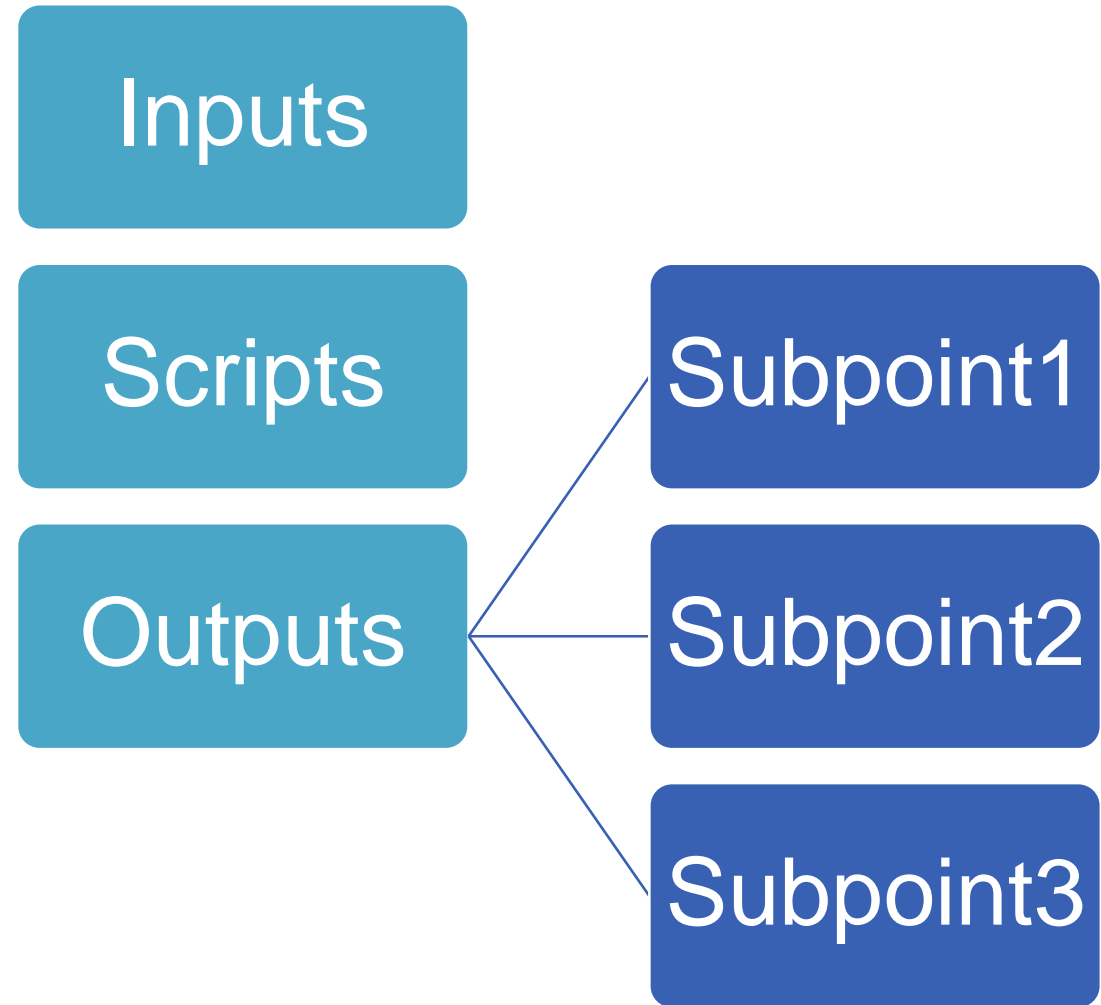


it's not a metaphor.

IT'S EASIER TO
LOSE A FILE ON
A SERVER



My suggested
data analysis
directory
structure



Name of a file/directory

THE GOLDILOCKS PRINCIPLE



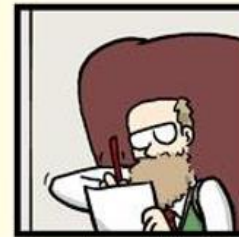
"FINAL".doc



FINAL.doc!



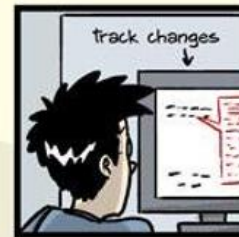
FINAL_rev.2.doc



FINAL_rev.6.COMMENTS.doc



FINAL_rev.8.comments5.
CORRECTIONS.doc



FINAL_rev.18.comments7.
corrections9.MORE.30.doc



FINAL_rev.22.comments49.
corrections.10.##\$%WHYDID
ICOMETOGRADSCHOOL?????.doc



JORGE CHAM © 2012



1955

1963

1966

1970

1971



1994

1996

1997

1997

2000



Version Control



2002

current

summer '09 & '10

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



WHAT DID I DO



Live
input

Script

Console

Command
line

script.sbatch

script.R

script.sh



Digital Lab notebooks

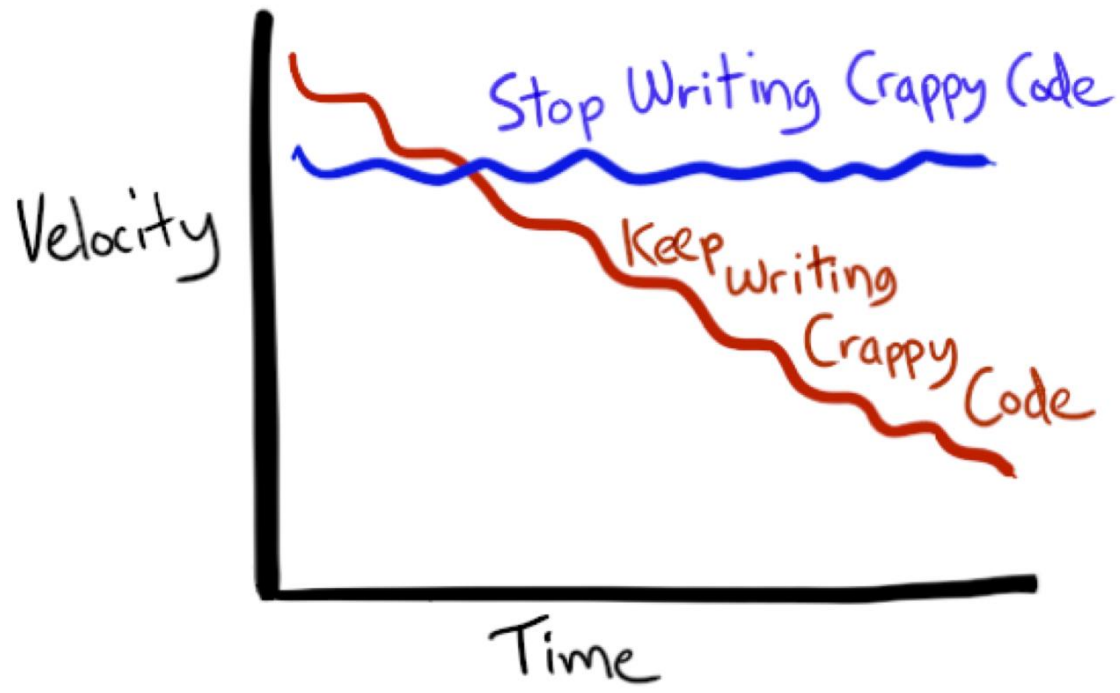


Directory structures

Readmes

Jupyter notebooks/R
markdown

Version control



MAKE A CHOICE



What is good code?



Data

Code

Publis
h

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

In case of fire



1. git commit



2. git push



3. leave building



Developers Swearing

61.2K Tweets



Follow

Developers Swearing

@gitlost

Unfiltered commit messages containing profanity from GitHub's API. Picture is of a burning NeXT Cube. Developed by @uiri00

gitlost.net Joined March 2013

5 Following 39.8K Followers

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



Tests itself frequently



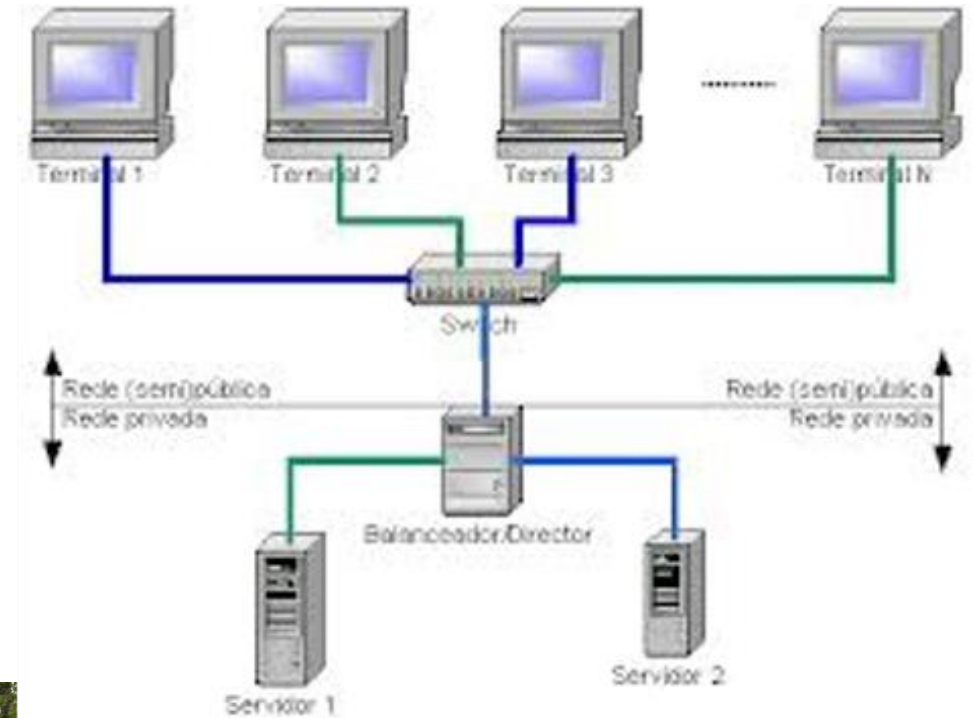
Have you ever tested that your code is giving you the expected result?



THE END



Playing nice on a super computer



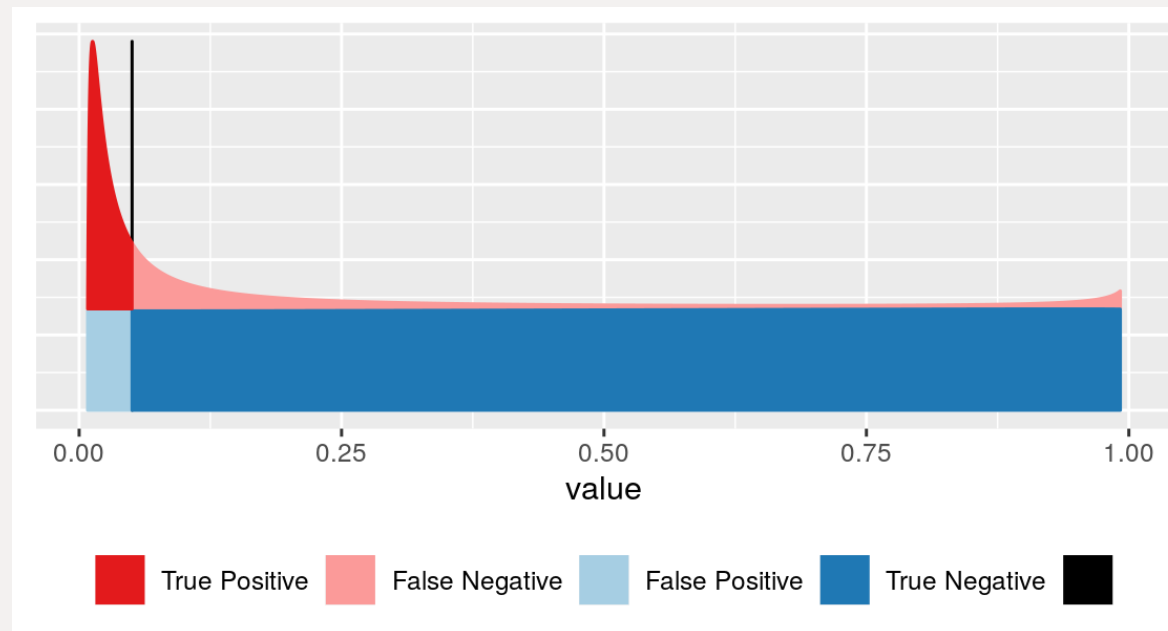
THE END



I GOT A GENE LIST! YEAH!
NOW WHAT?



DO I REALLY HAVE SIGNIFICANT GENES?



ENRICHR



Enrichr

[Login](#) | [Register](#)

49,611,534 sets analyzed

402,056 terms

196 libraries

Analyze

[What's new?](#)

[Libraries](#)

[Gene search](#)

[Term search](#)

[About](#)

[Help](#)

Input data

Expand a gene, a term, or a variant into a gene set:

e.g. STAT3, breast cancer, or rs28897756



Try an example

Include the top 100 most relevant genes



Paste a set of valid Entrez gene symbols on each row in the text-box below. [Try a gene set example.](#)

Paste a set of valid Entrez gene symbols (e.g. STAT3) on each row in the text-box

0 gene(s) entered

In order to enable others to search your set please enter a brief description of it.

Contribute your set so it can be searched by others

Please acknowledge Enrichr in your publications by citing the following references:

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A.

Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013; 128(14).

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A.

Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*. 2016; gkw377.

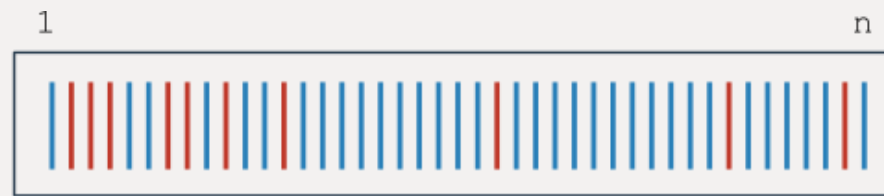
Xie Z, Bailey A, Kuleshov MV, Clarke DJB., Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M, & Ma'ayan A.

Gene set knowledge discovery with Enrichr. *Current Protocols*, 1, e90. 2021. doi: 10.1002/cpz1.90

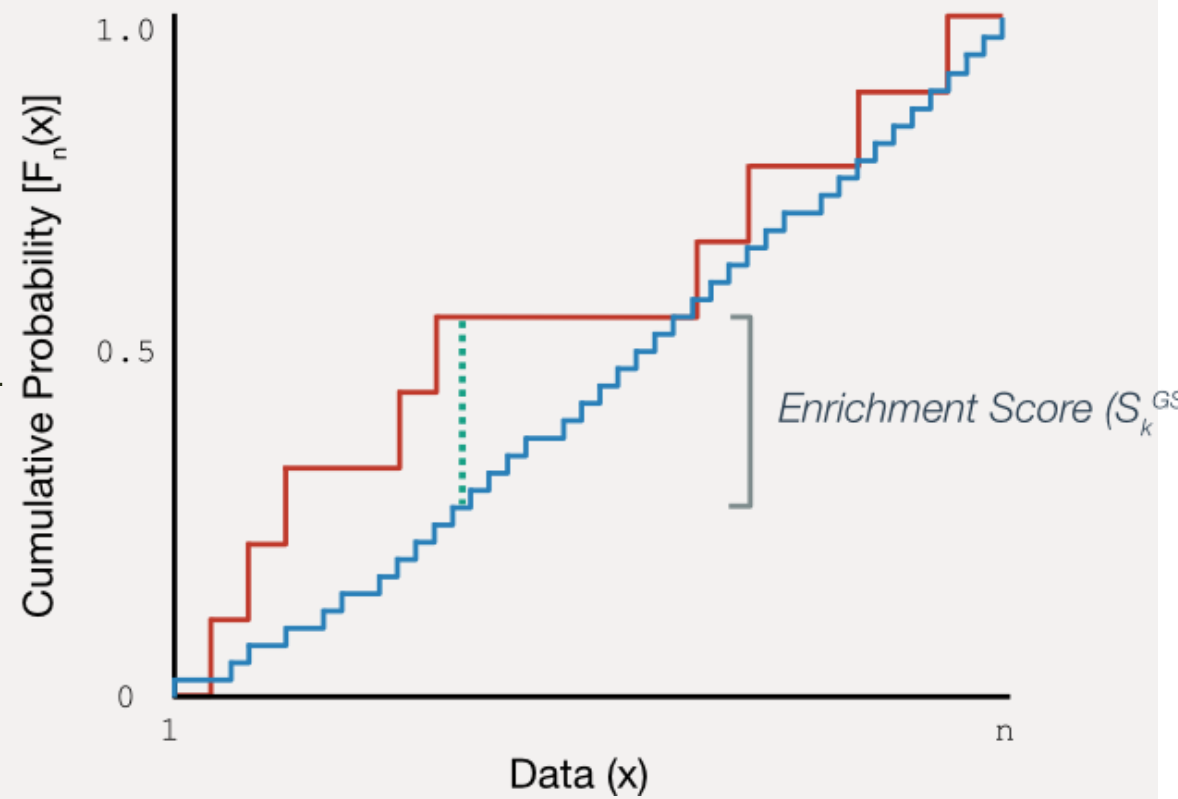
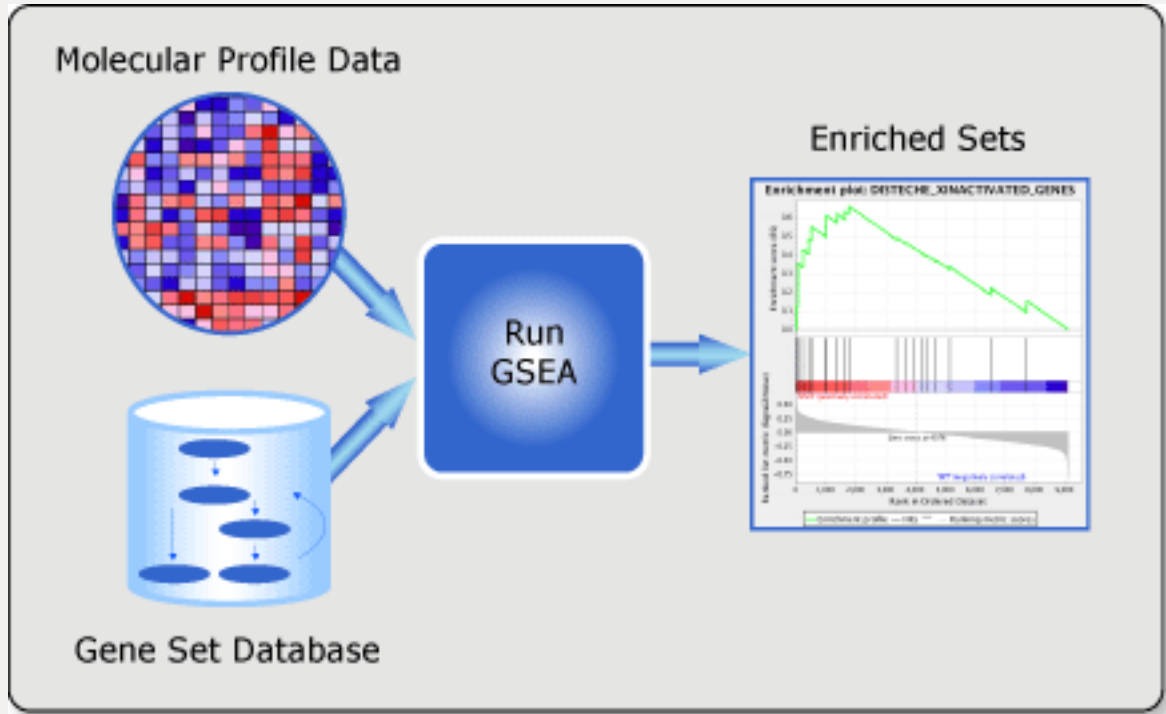
<https://maayanlab.cloud/Enrichr/>

GSEA

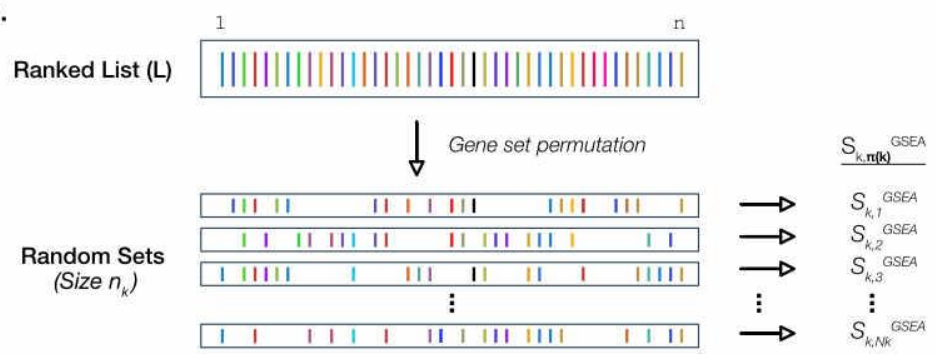
Ranked Gene List (L)

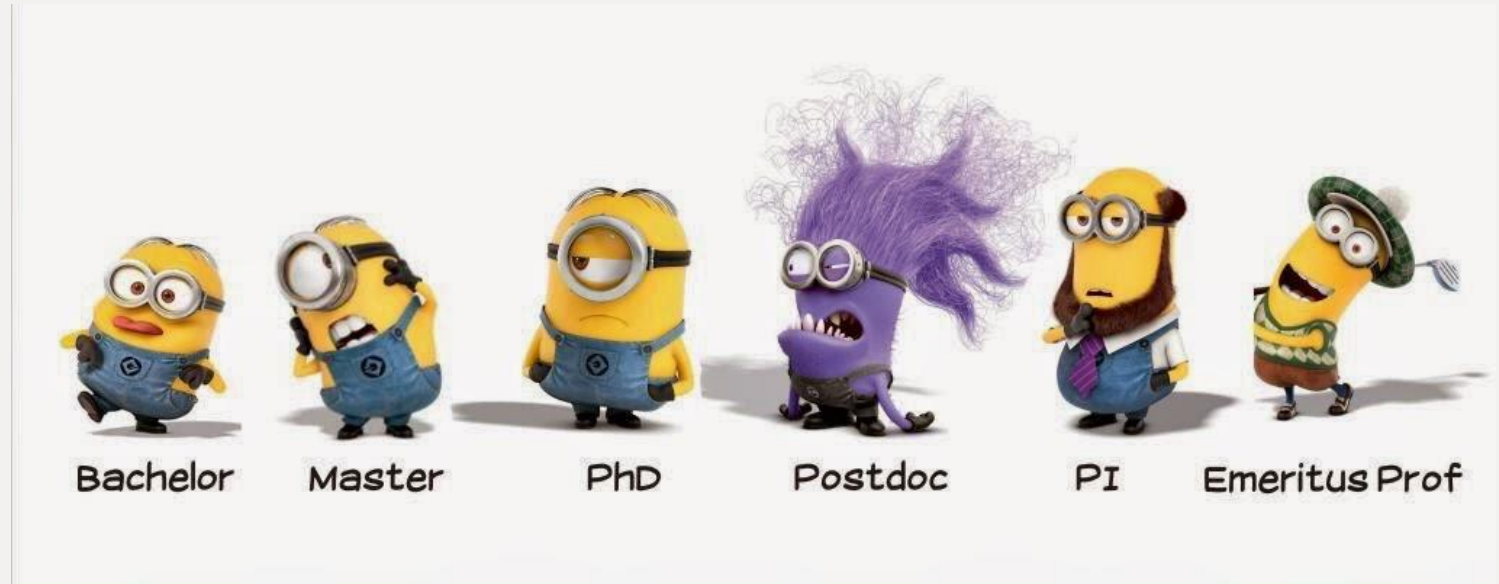
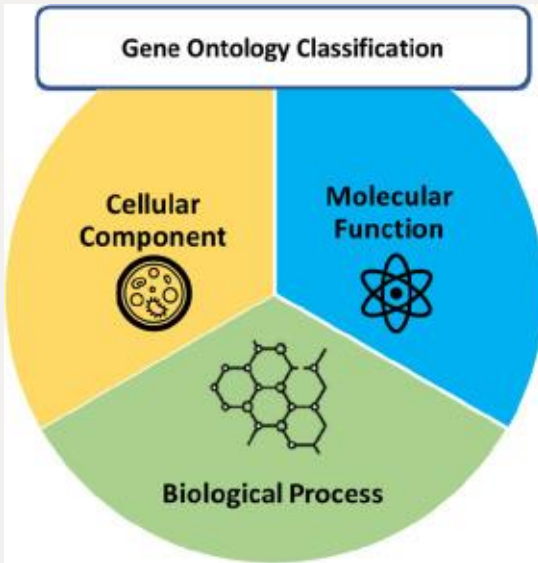


↓ Cumulative distribution function



B.



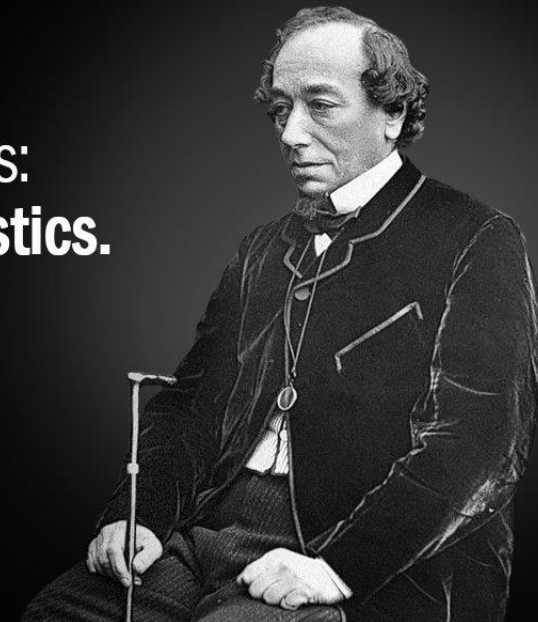


GO
GENE ONTOLOGY

There are three kinds of lies:
lies, damned lies, and statistics.

– Benjamin Disraeli

AZ QUOTES



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?



ILLUSTRATED BY SEGUE TECHNOLOGIES

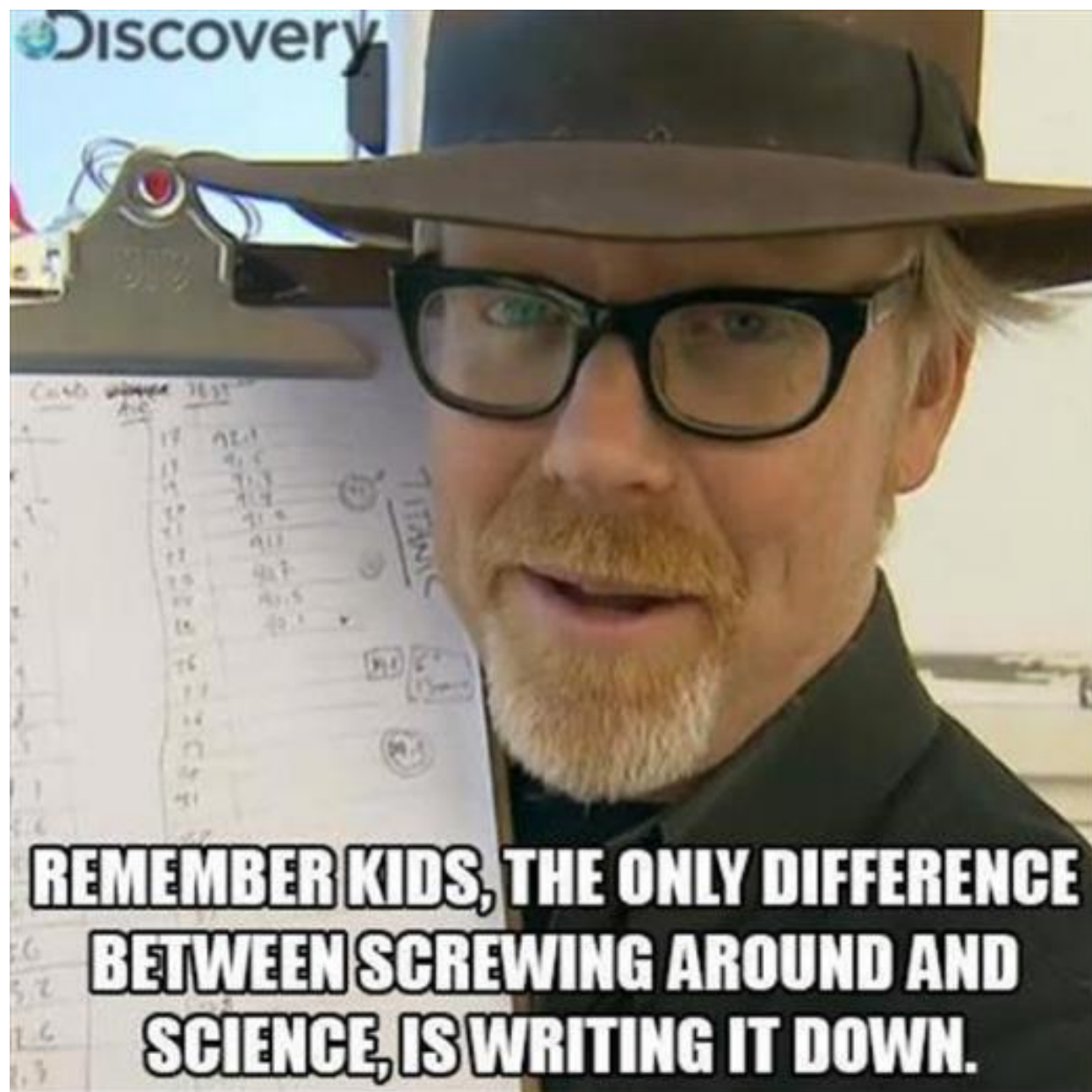
Findable 
Accessible 
Interoperable 
Reusable 

Choosing slurm options

Clean responsibly



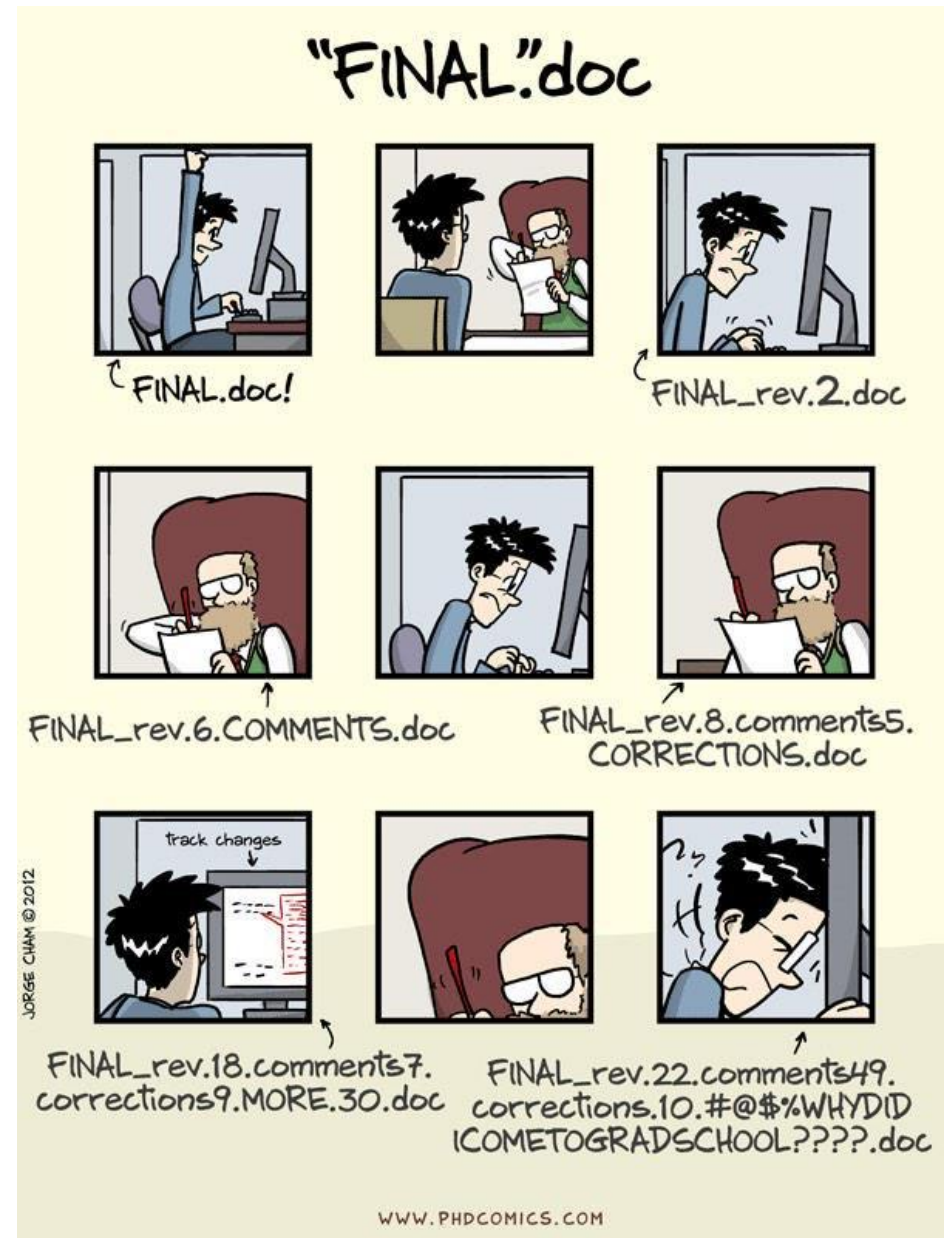
A screenshot of the OpenRefine website homepage. The browser address bar shows 'openrefine.org'. The page has a dark sidebar on the left with the OpenRefine logo and navigation links: Home, Community, Documentation, Download, Contact Us, Blog, and a badge for 'Enhanced with Java profiler' and 'JPROFILER'. The main content area has a 'Welcome!' heading, a paragraph describing OpenRefine as a tool for working with messy data, a list of supported languages, and a 'Google News Initiative' logo. Below that is an 'Introduction to OpenRefine' section with a sub-heading '1. Explore Data' and a paragraph of introductory text.



How understandable is your laboratory notebook?



Report the
experiment
(at every
step!)



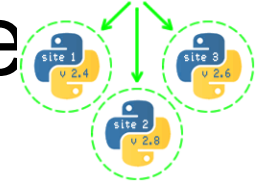


Machine

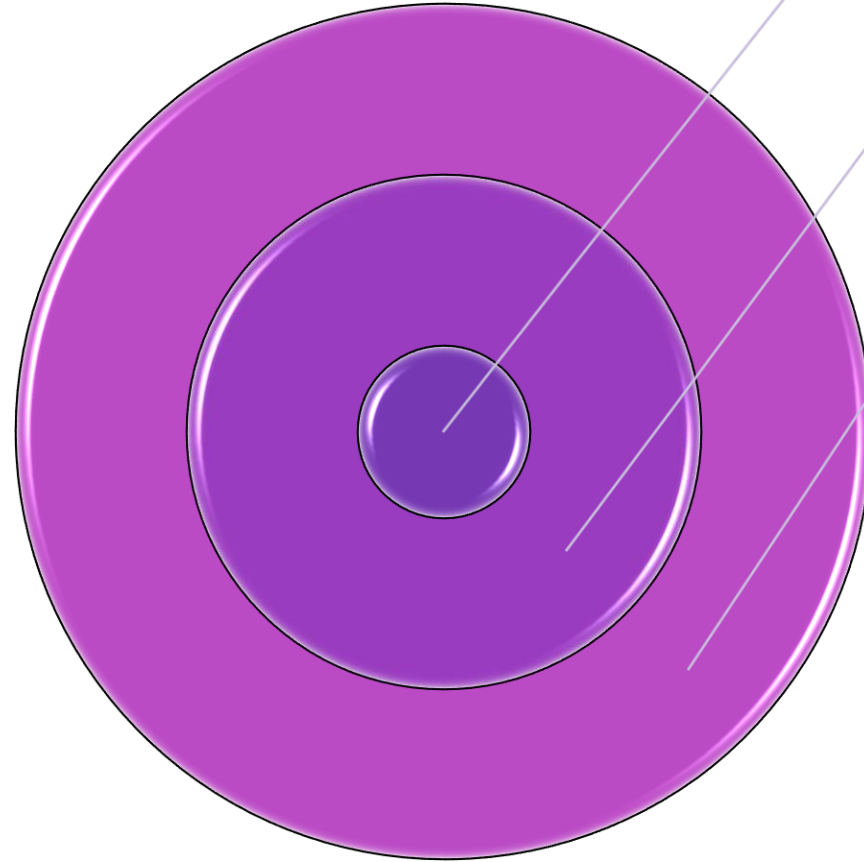
Code

Data

> virtualenv



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



Presenting data ethically

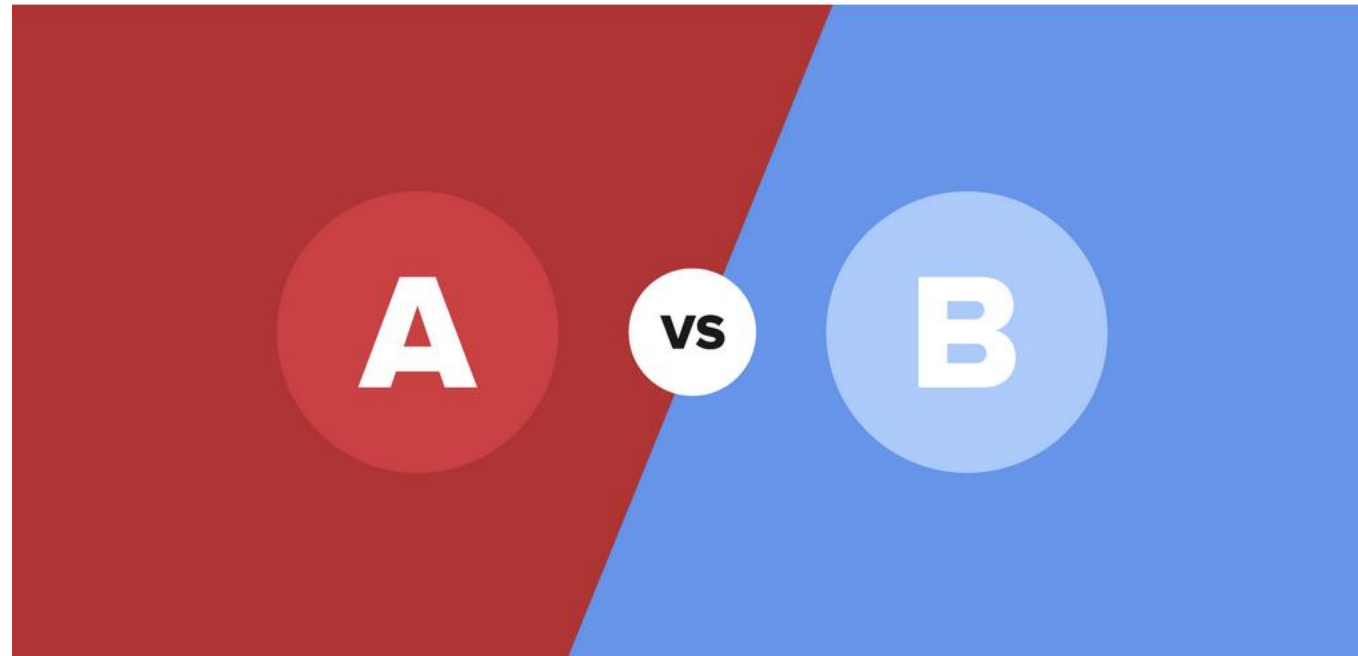


How do you know when your hypothesis may be correct?



“Data don’t make any sense,
we will have to resort to statistics.”

Is X different between A and B?



X

30

40

When poll is active, respond at pollev.com/maryallen084

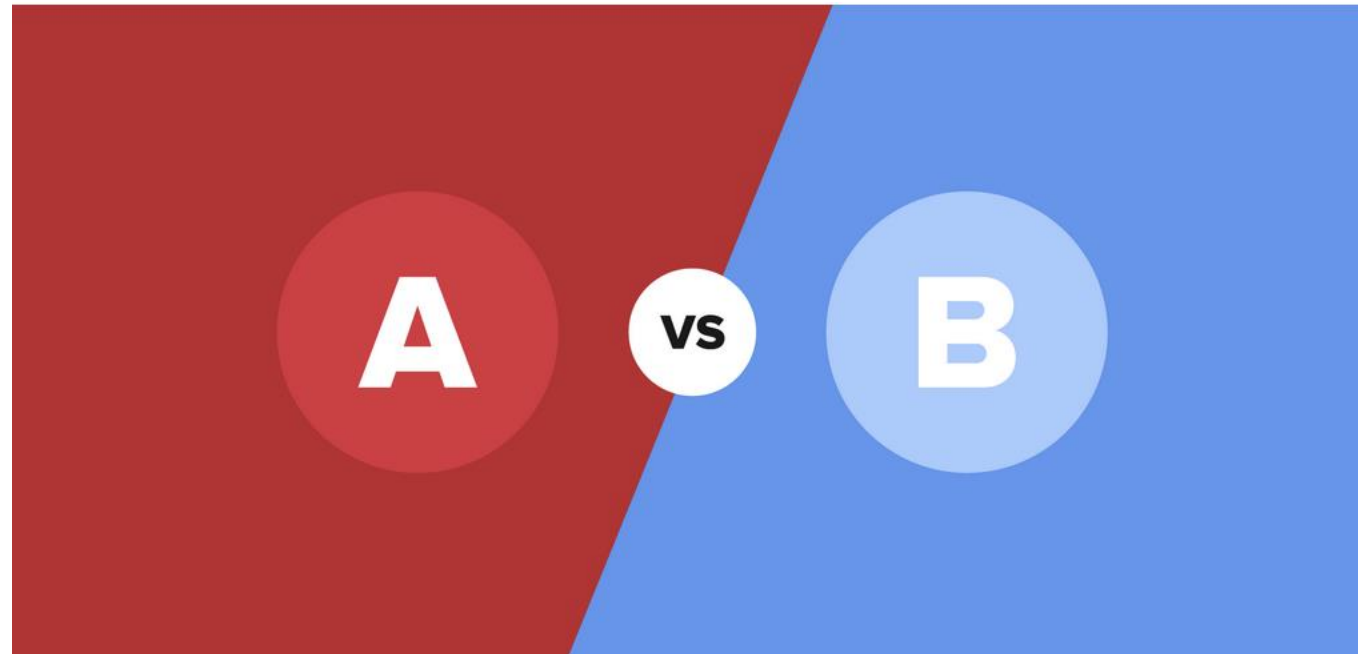
Text **MARYALLEN084** to **37607** once to join

Is X different in A vs. B?

Yes

No

Is X different between A and B?



X

30

40

P-value is 0.04

When poll is active, respond at pollev.com/maryallen084

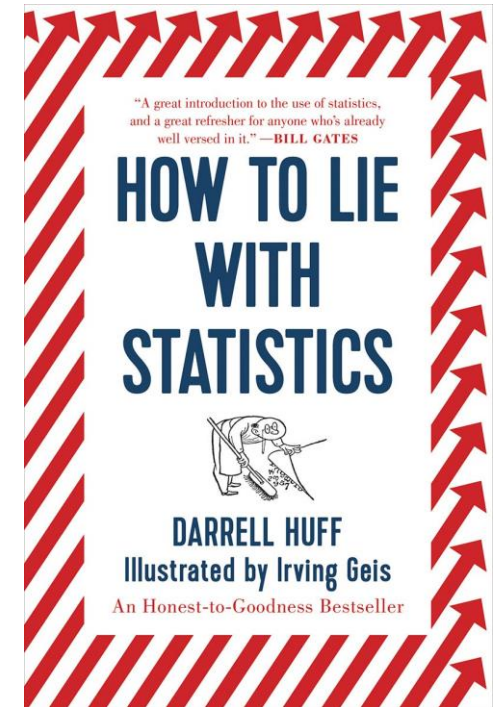
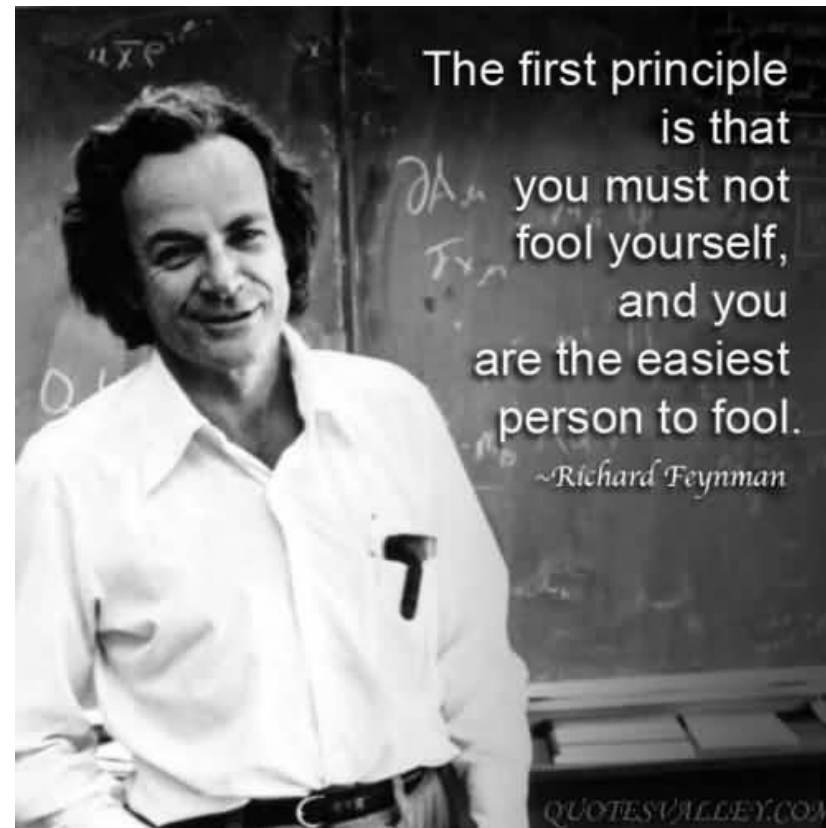
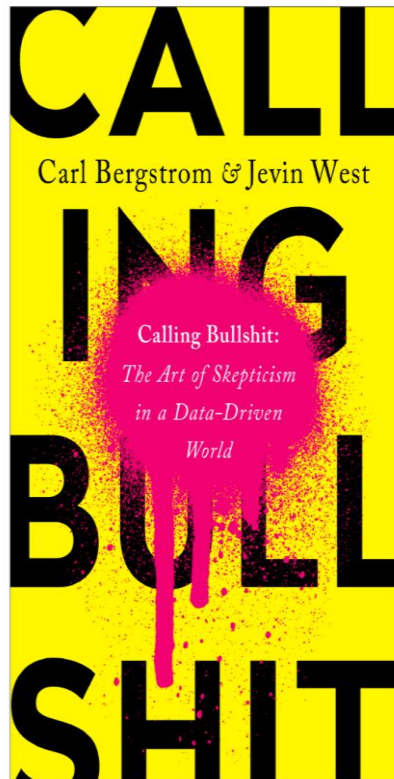
Text **MARYALLEN084** to **37607** once to join

Is X different in A vs. B?

Yes

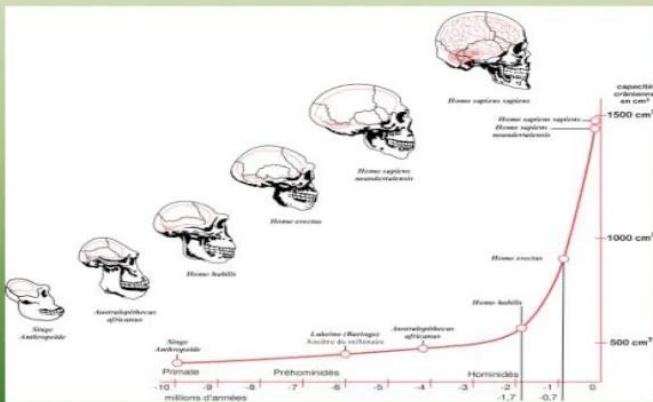
No

It is easy to fool people with data



Your brain did not **evolve** to do statistics

The evolution of human brain:-

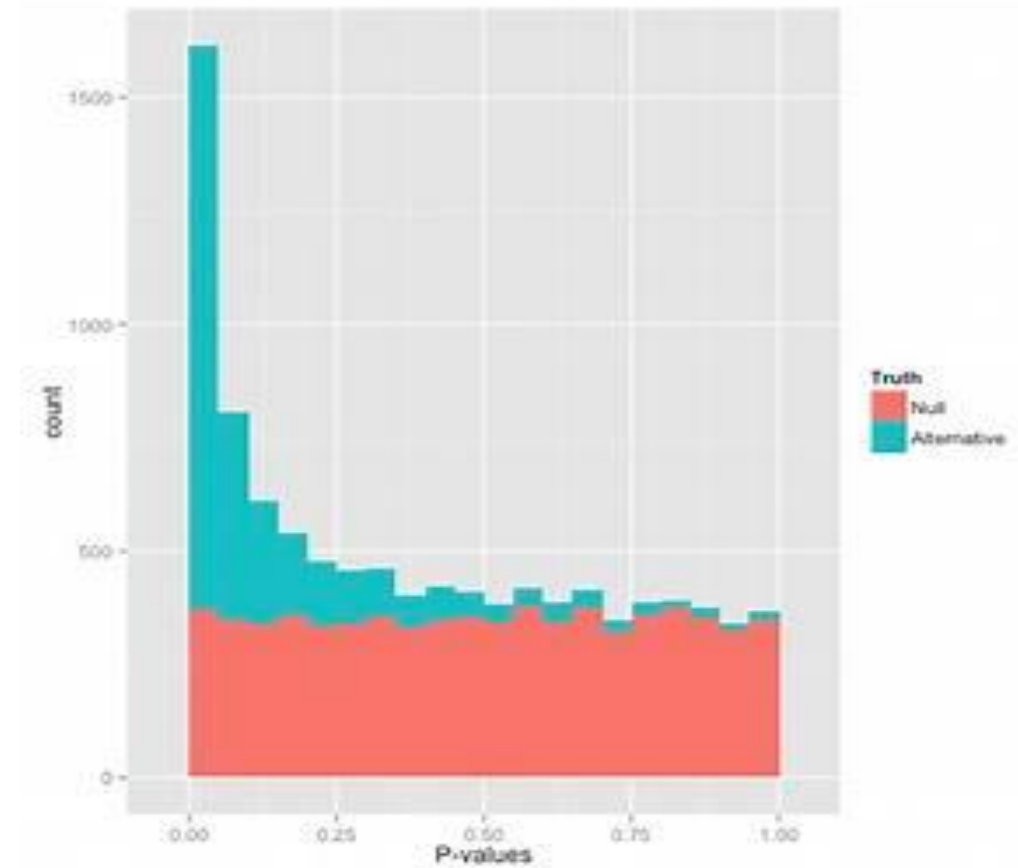


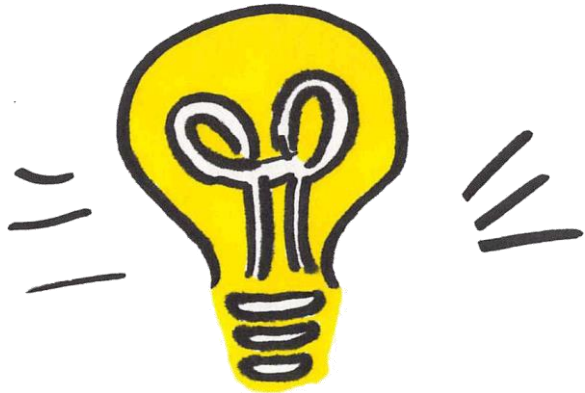
- We jump to conclusions
- We tend to be over-confident
- We see patterns in random data
- We don't realize coincidences are common
- We avoid thinking about ambiguous situations
- We are fooled by multiple comparisons
- We ignore alternative explanations
- We are fooled by regression to the mean
- We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



What is a p-value?

What is a p-value?





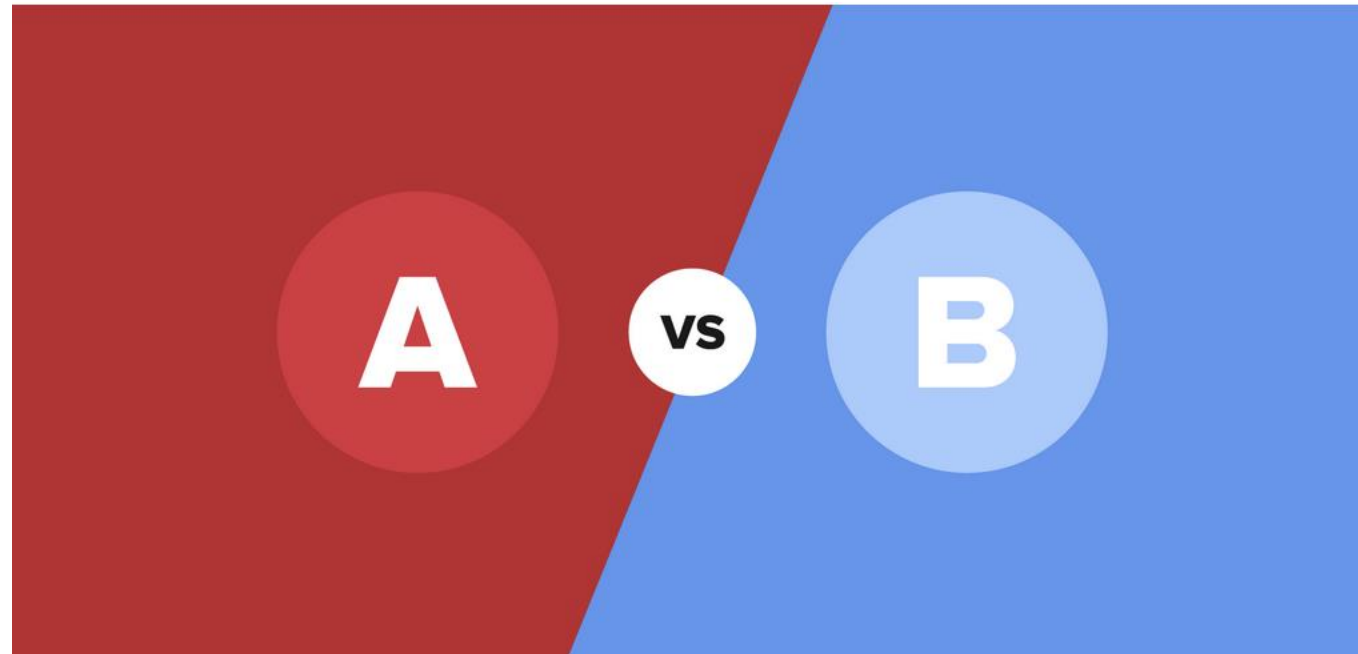
This Photo by Unknown author is licensed under [CC BY-SA](#).

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

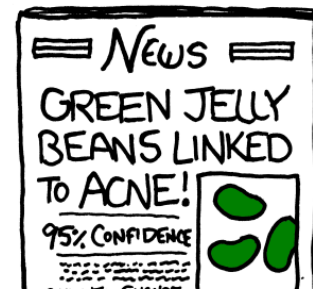
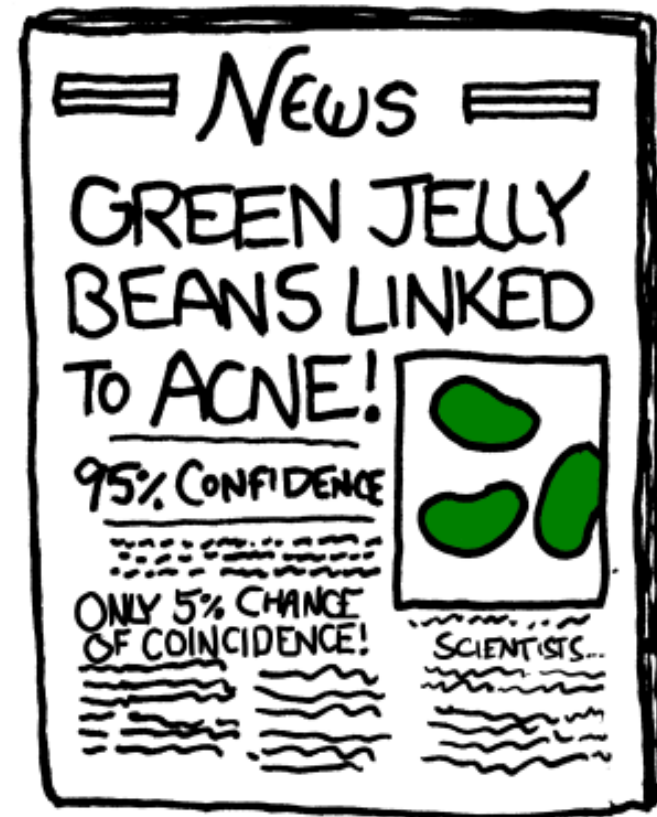
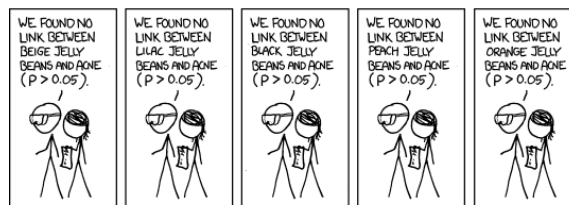
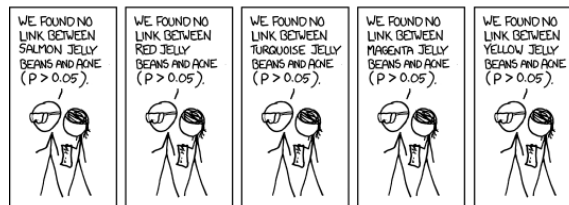
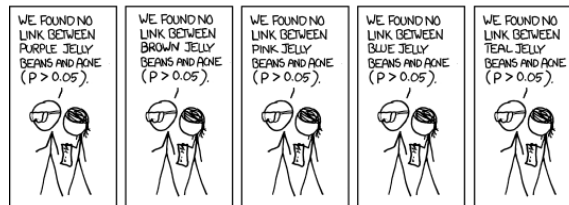
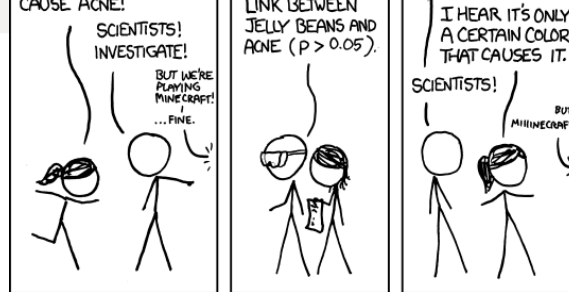
When poll is active, respond at pollev.com/maryallen084

Text **MARYALLEN084** to **37607** once to join

Is X different in A vs. B?

Yes

No





MISTAKES
are proof that you are
TRYING



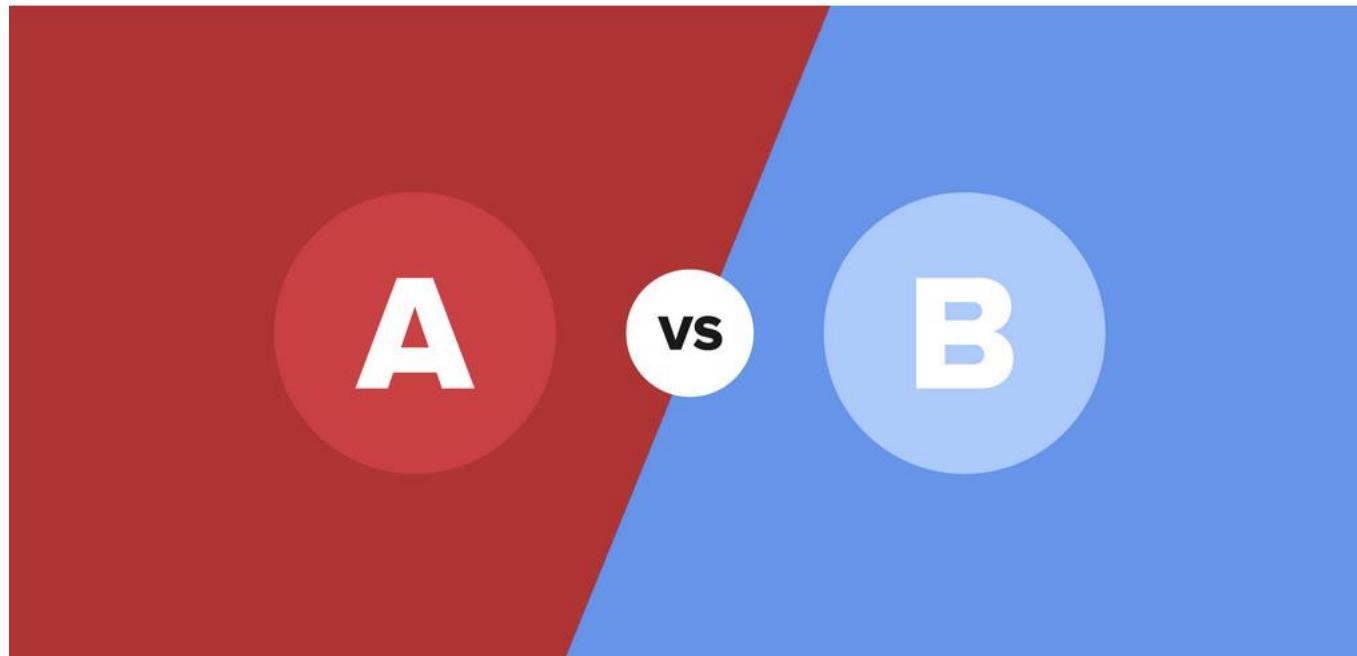
MISTAKES
are proof that you are
TRYING



shutterstock · 1010689966

| | | Predicted Class | | |
|--------------|----------|--|--|--|
| | | Positive | 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?



X1

30

40

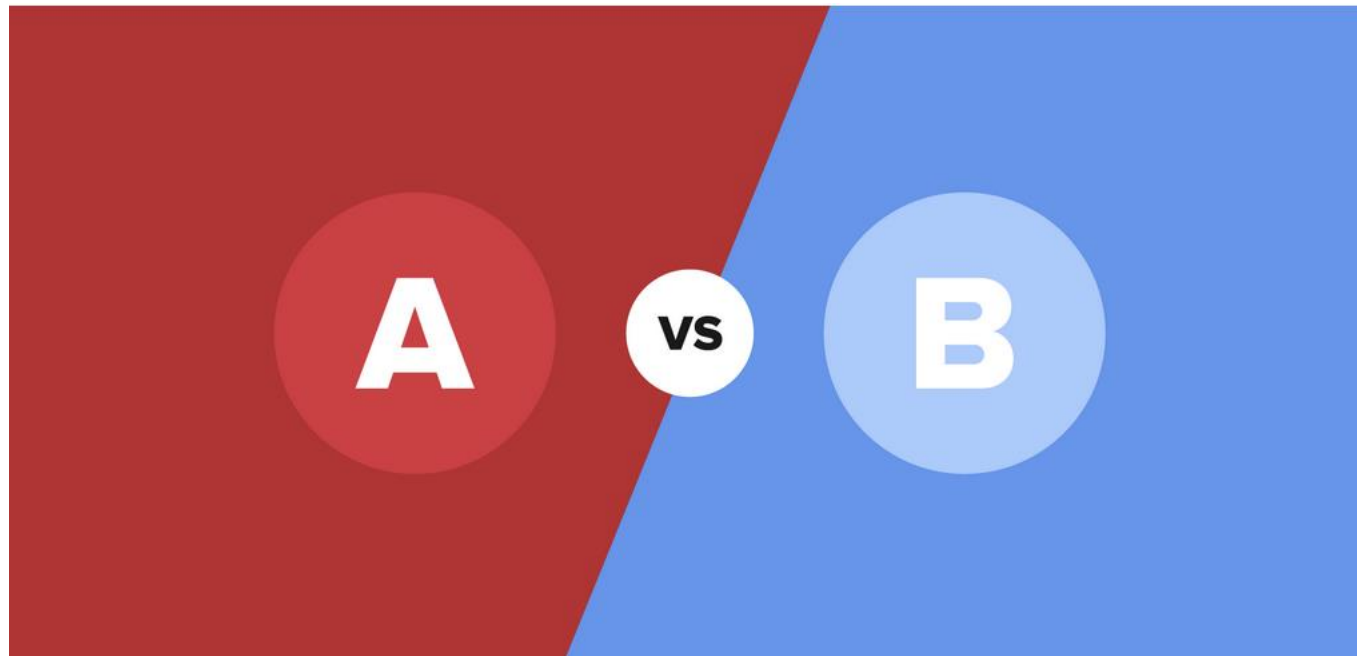
| I tested | p-value |
|----------|---------|
| X1 | 0.01 |
| X2 | 1 |
| X3 | .05 |
| | |
| X200,000 | 1e^1 |

P-value is 0.04

What is an adjusted p-value?



Is X different between A and B?



X1

30

40

I tested

X1

p-value

0.01

X2

1

X3

.05

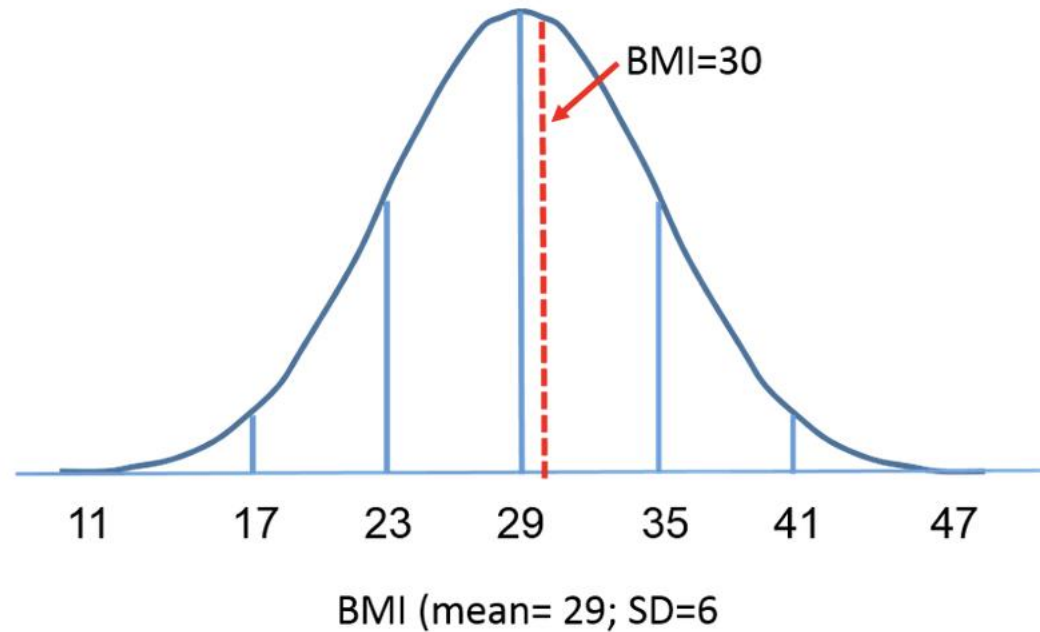
....

X200,000

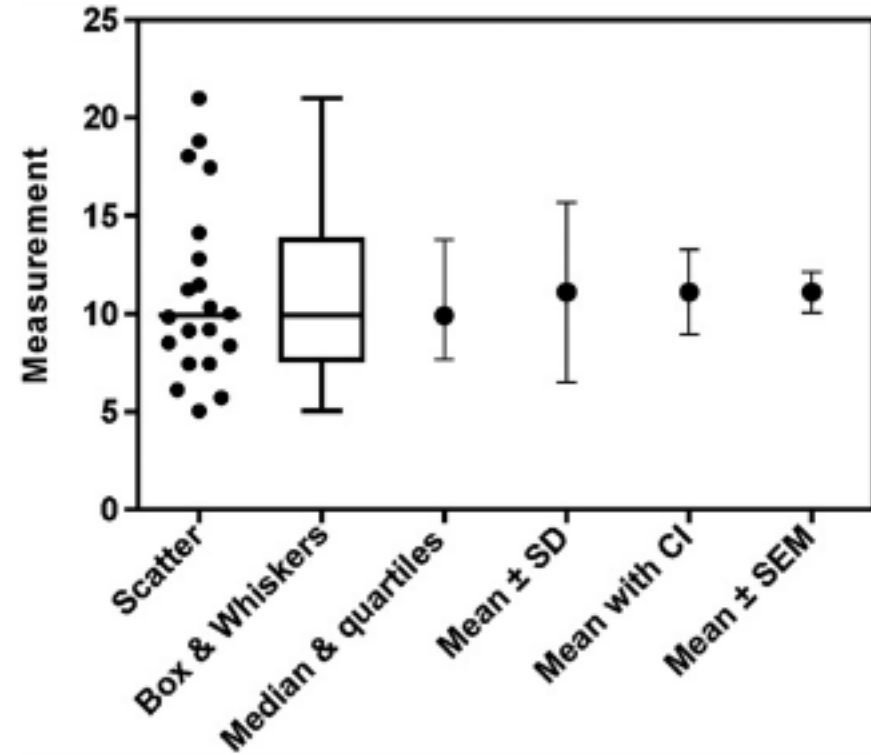
1e^1

P-value is 0.04

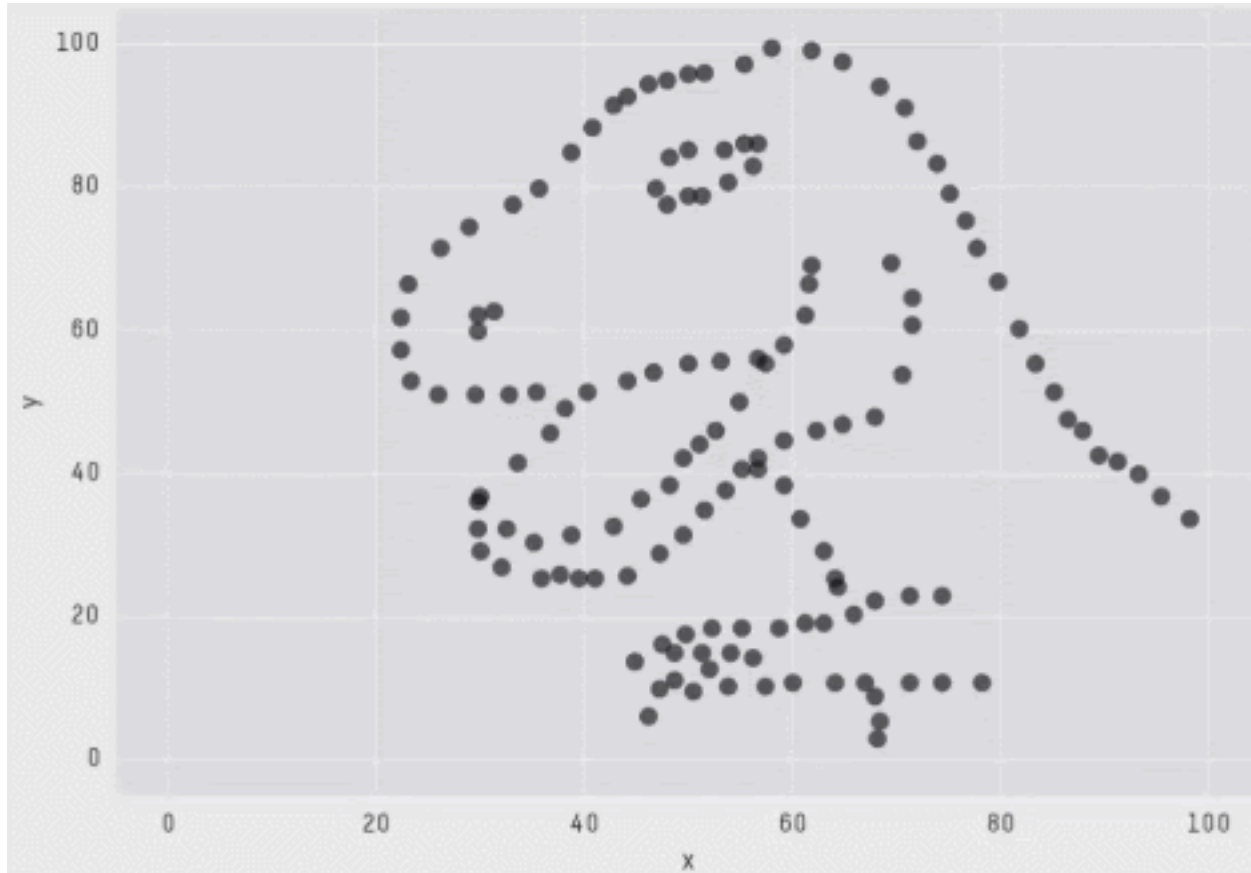
Is X different between A and B?
Remember: A value is always comes from a distribution.



Averages can be very misleading!



Maybe you should look at your data?

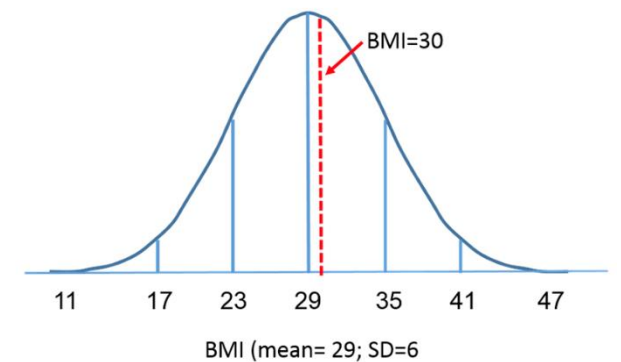


```
X Mean: 54.2659224  
Y Mean: 47.8313999  
X SD   : 16.7649829  
Y SD   : 26.9342120  
Corr.  : -0.0642526
```

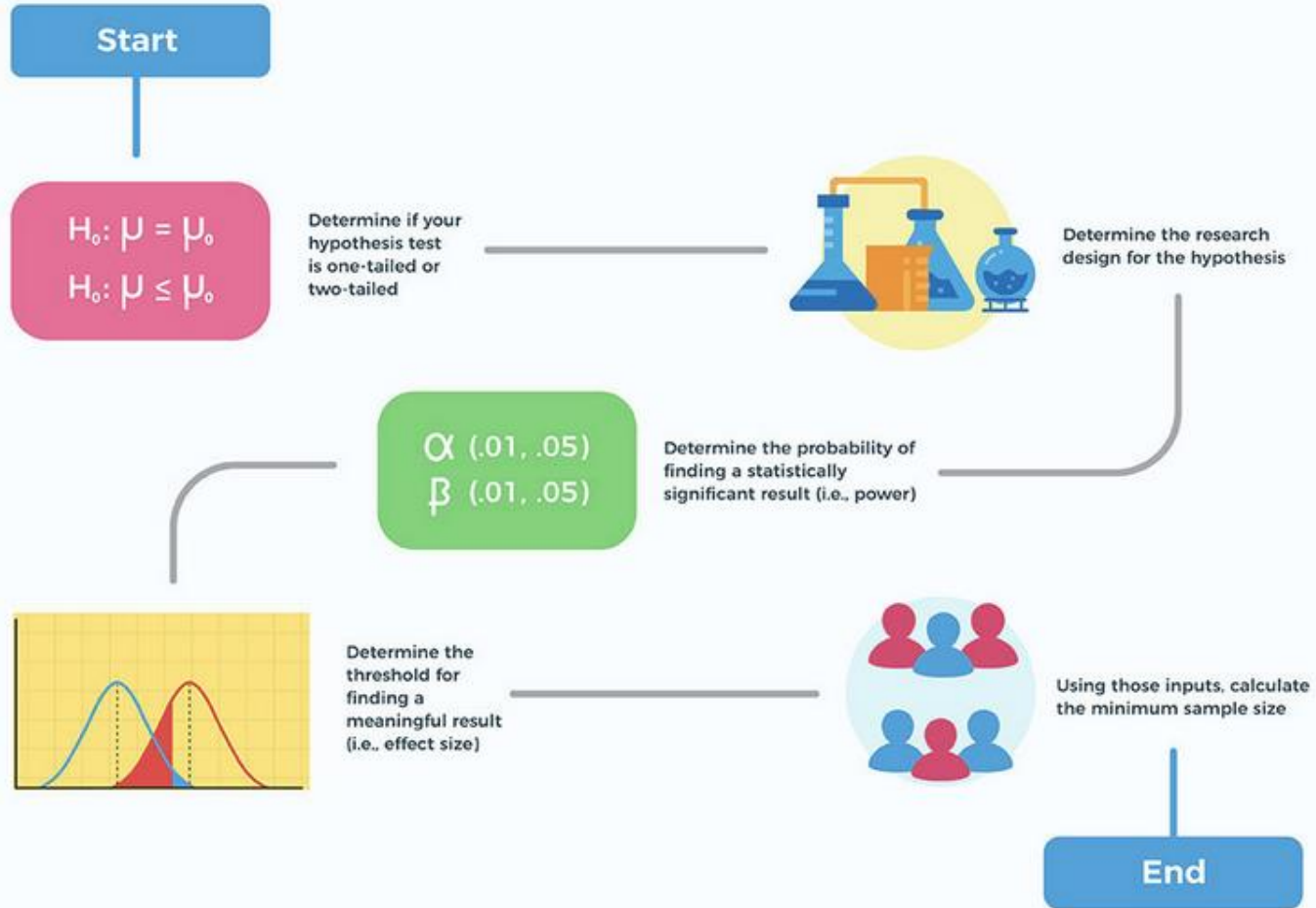
<https://www.autodeskresearch.com/publications/samestats>

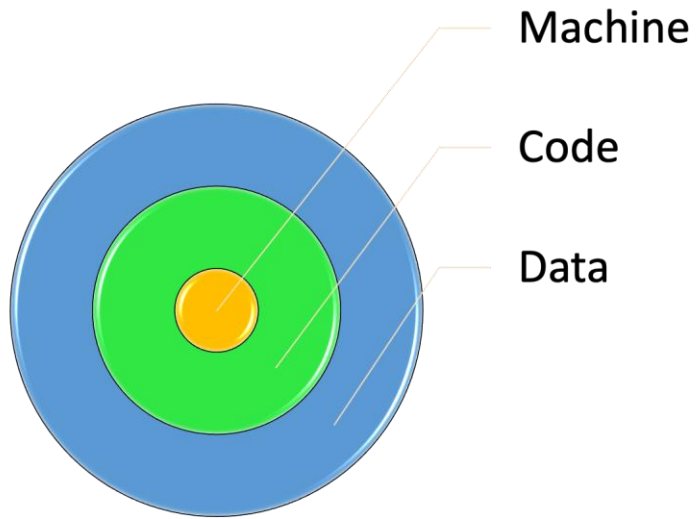
• <https://www.youtube.com/watch?v=Hz1fyhVOjr4>

How many replicates do I need?



POWER ANALYSIS

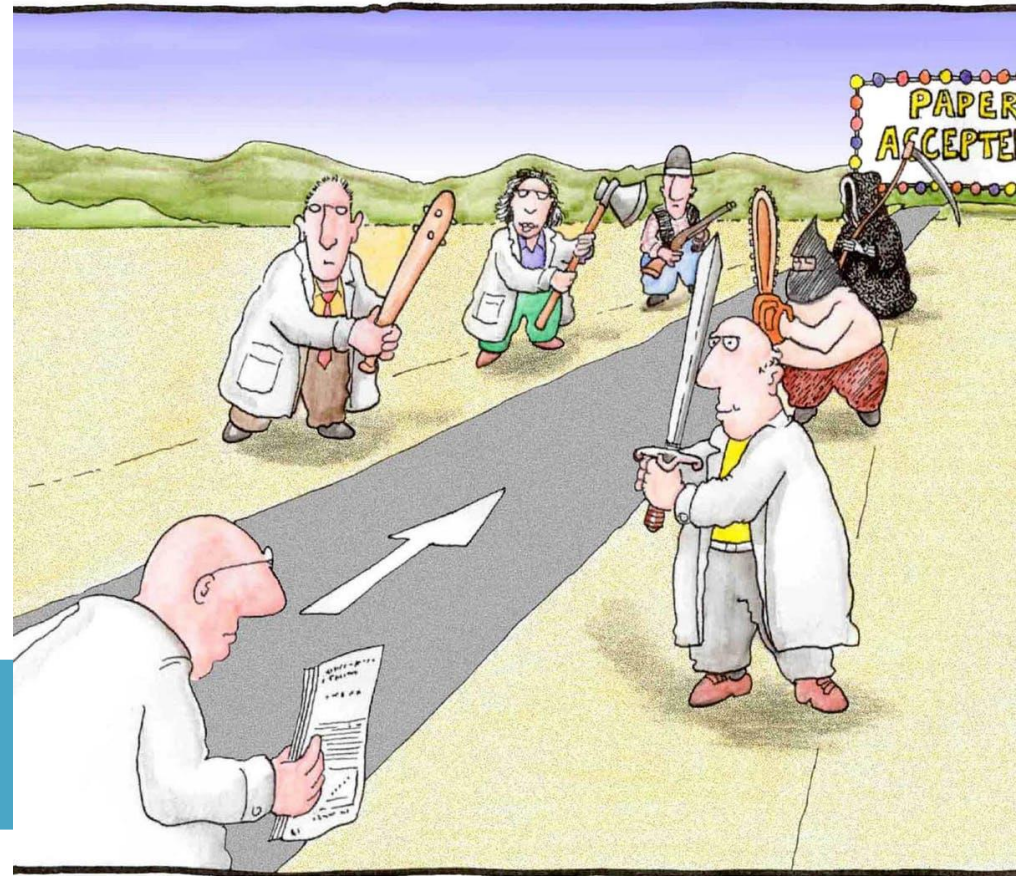




Data

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,



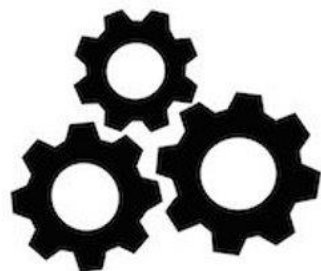
F
Findable



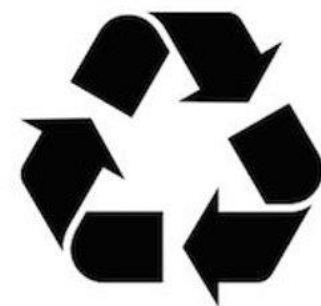
A
Accessible



I
Interoperable



R
Reusable





PROS AND CONS OF PIPELINES



nextflow



NLM vs. clean Metadata