There and Back Again
a Data Scientist's Tale

by Lucy D'Agostino McGowan


lucymcgowan.com/talk


lucymcgowan.com/talk
@LucyStats

da·ta sci·en·tist

n. *a person who effectively extracts, manipulates, draws inference from, & communicates about data*

We’re all mad here

Carroll, Lewis. *Alice's adventures in wonderland*. Macmillon, 1865

data scientists

We're all ^~~mad~~

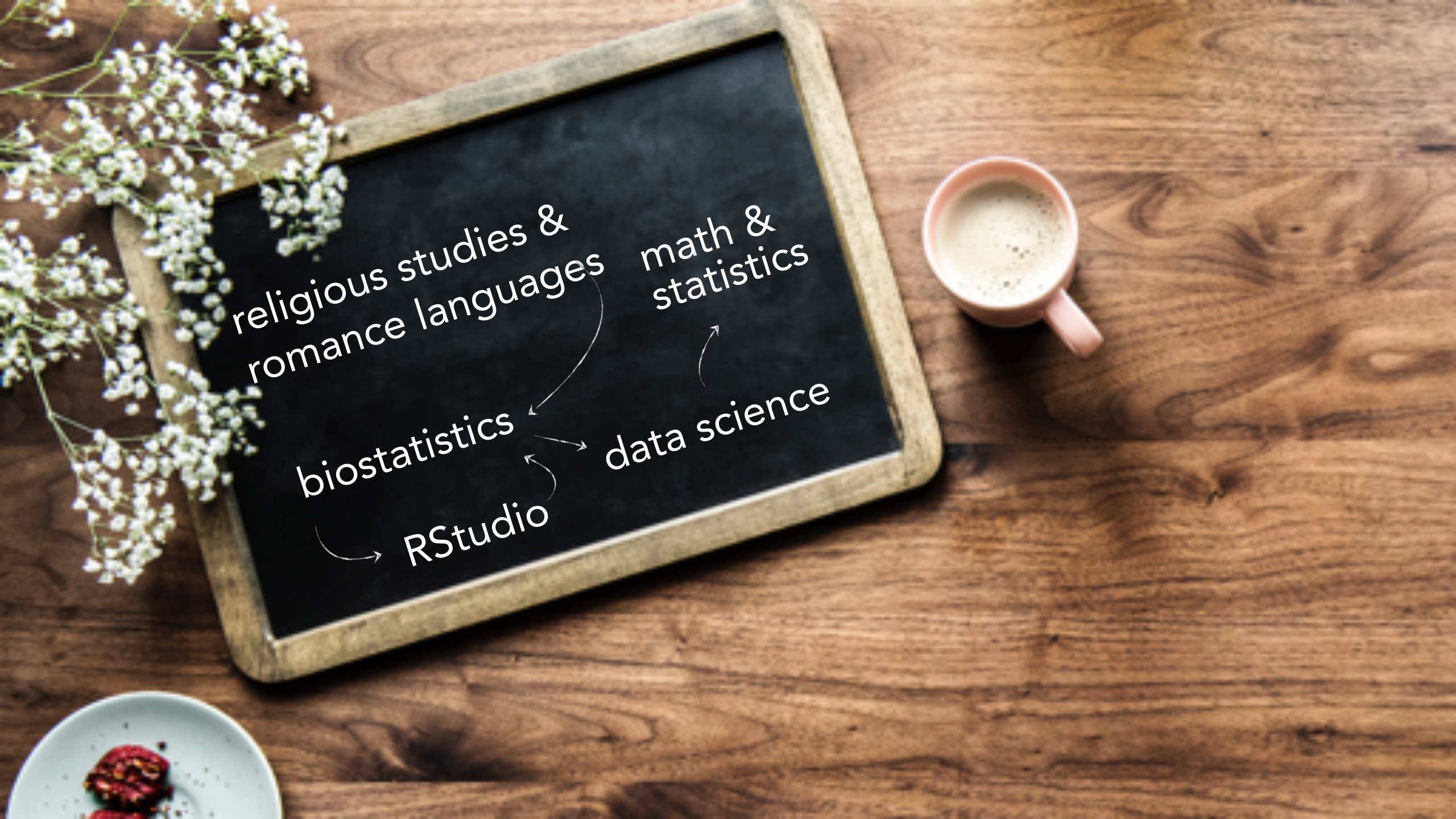
here

Carroll, Lewis. *Alice's adventures in wonderland*. Macmillon, 1865


I'm going on an adventure!


Tips / lessons
learned


background


religious studies &
romance languages
math &
statistics
biostatistics
data science
RStudio

religious studies &
romance languages

math &
statistics

biostatistics

data science

RStudio

religious studies &
romance languages

math &
statistics

biostatistics

data science

RStudio


religious studies &
romance languages
math &
statistics
biostatistics
data science
RStudio
google
drive

religious studies &
romance languages

math &
statistics

biostatistics

data science

RStudio

religious studies &
romance languages

math &
statistics

biostatistics

data science

RStudio

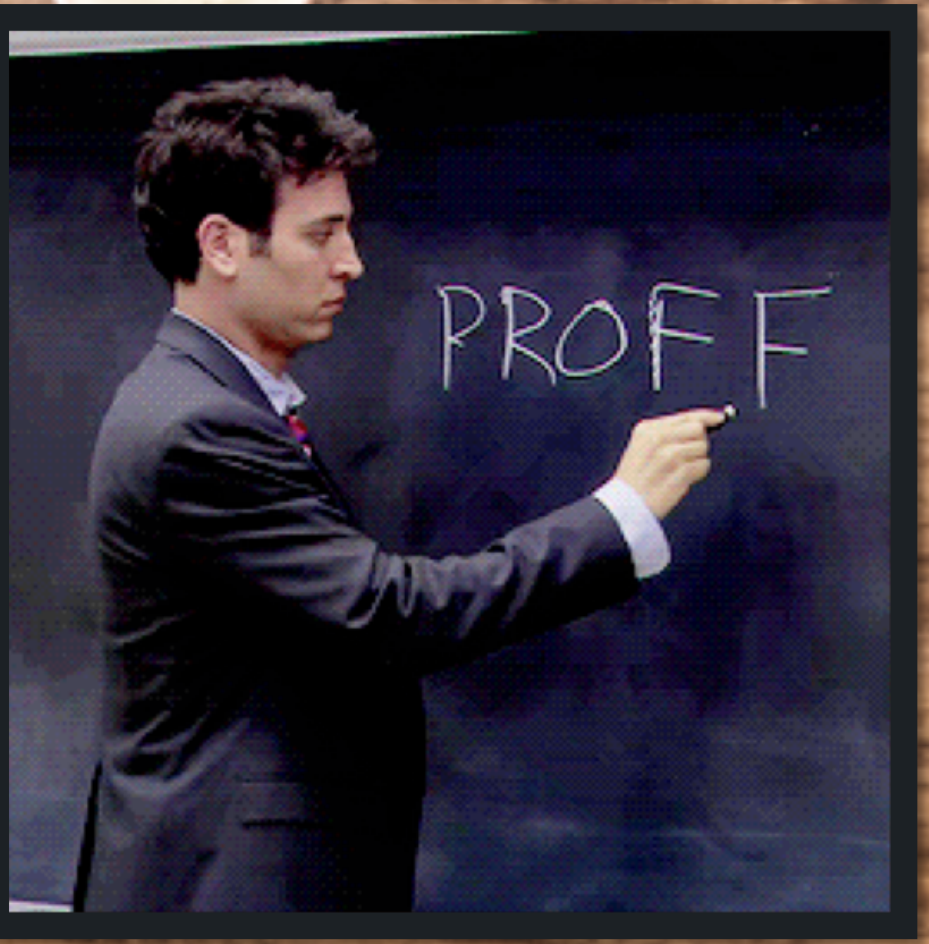
religious studies &
romance languages

math &
statistics

biostatistics

data science

RStudio




Embrace
non-linearity


religious studies &
romance languages
math &
statistics
biostatistics
data science
RStudio

religious studies &
romance languages

math &
statistics

biostatistics

data science

RStudio

religious studies &
romance languages

math &
statistics

biostatistics

data

RStudio

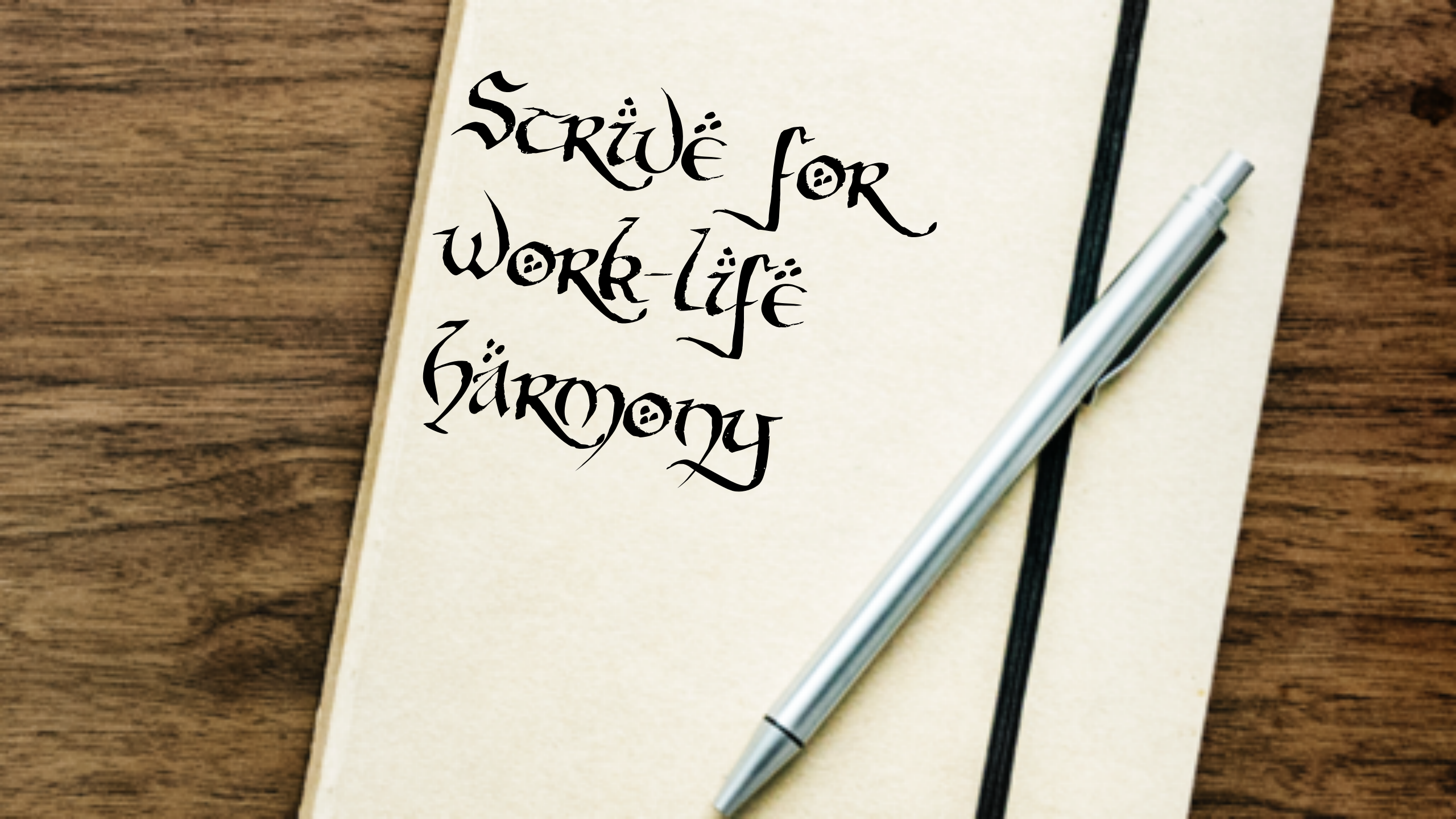

religious studies &
romance languages
math &
statistics
biostatistics
RStudio
data scienc
#rchickenladies
www.rstudio.com
R
Ladies

religious studies &
romance languages

math &
statistics

biostatistics → data science

RStudio


Scribe for
work-life
harmony


thoughts on
data science

da·ta sci·en·tist

n. *a person who effectively extracts, manipulates, draws inference from, & communicates about data*

da·ta sci·en·tist

n. *a person who effectively extracts, manipulates, draws inference from, & communicates about data*

de·gree of free·dom

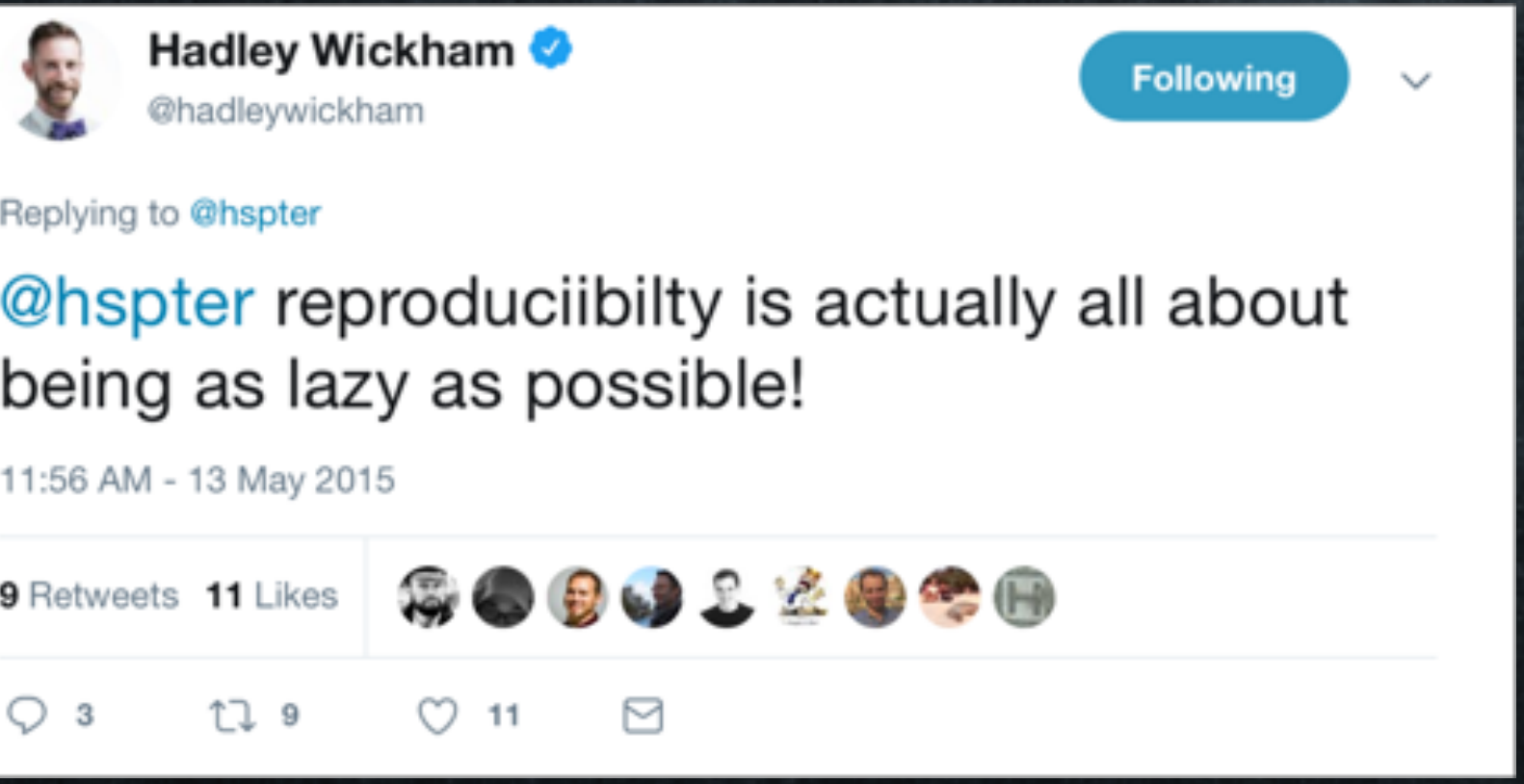
n. *the number of independent values or quantities that can vary in an analysis*

researcher

^**de·gree of free·dom**

n. *the number of* decisions *independent ~~values or quantities~~ that can vary in an analysis*


Reproducibility


Hadley Wickham
@hadleywickham
Following
Replying to @hspter
@hspter reproduciibilty is actually all about being as lazy as possible!
11:56 AM - 13 May 2015
9 Retweets 11 Likes
3
9
11

Names Matter

Bryan, Jenny. Naming things.
https://speakerdeck.com/jennybc/how-to-name-files

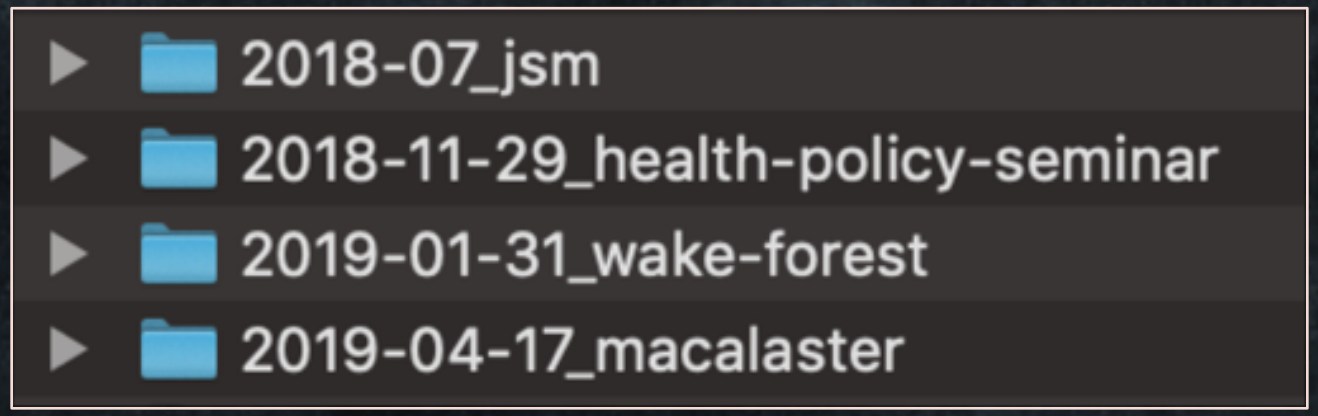
Names Matter

- machine readable
- human readable
- play well with default ordering

Bryan, Jenny. Naming things.
https://speakerdeck.com/jennybc/how-to-name-files

Names Matter

- machine readable
- human readable
- play well with default ordering



Bryan, Jenny. Naming things.
https://speakerdeck.com/jennybc/how-to-name-files

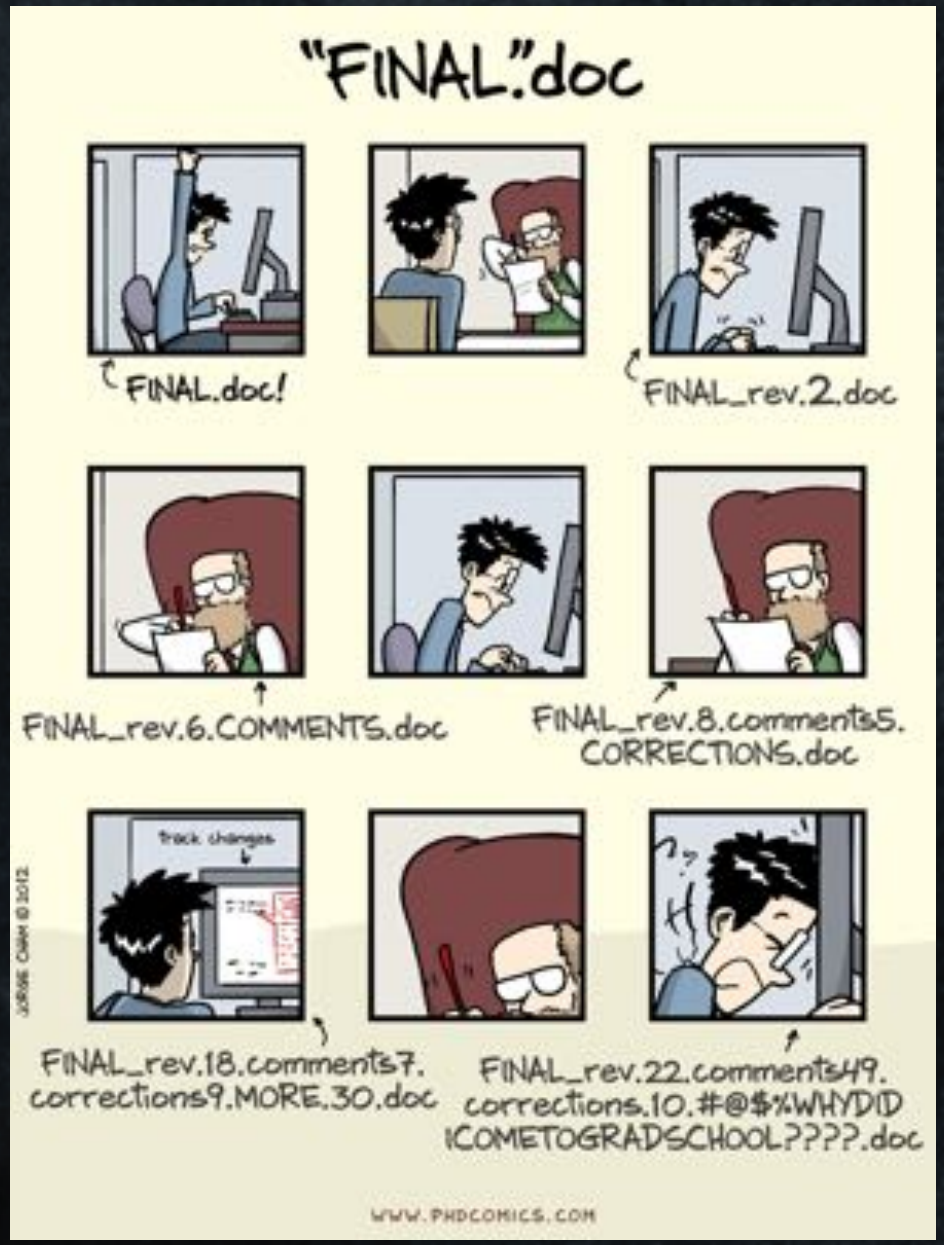
Version control

Cham, Jorge. notfinal.doc.
http://phdcomics.com/comics/archive_print.php?comicid=1531

Version control

Dropbox

Cham, Jorge. notfinal.doc.
http://phdcomics.com/comics/archive_print.php?comicid=1531



Cham, Jorge. notfinal.doc.
http://phdcomics.com/comics/archive_print.php?comicid=1531

Version control

Dropbox

googledrive

Write a function

Gannon, Chris. Many Me!
https://codepen.io/chrisgannon/details/OmpPPv

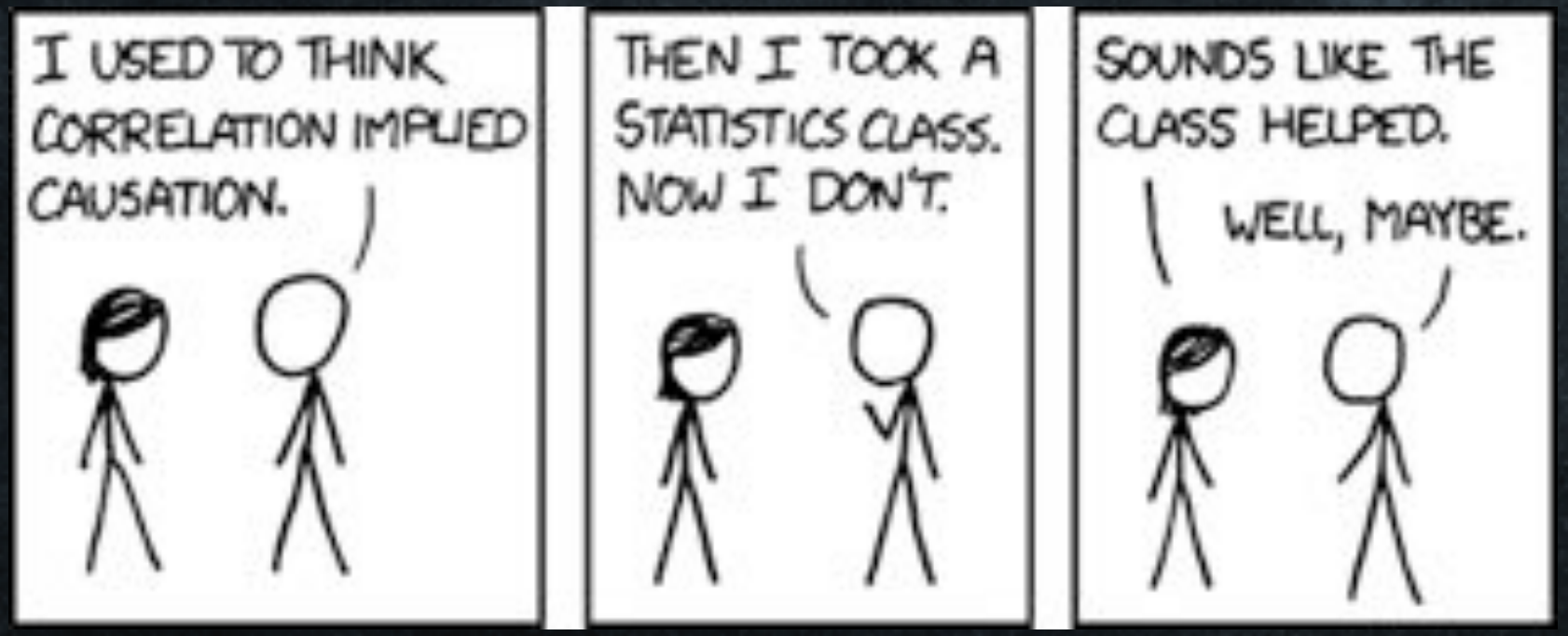

Be kind to
future
collaborators


Be kind to
future
~~collaborators~~

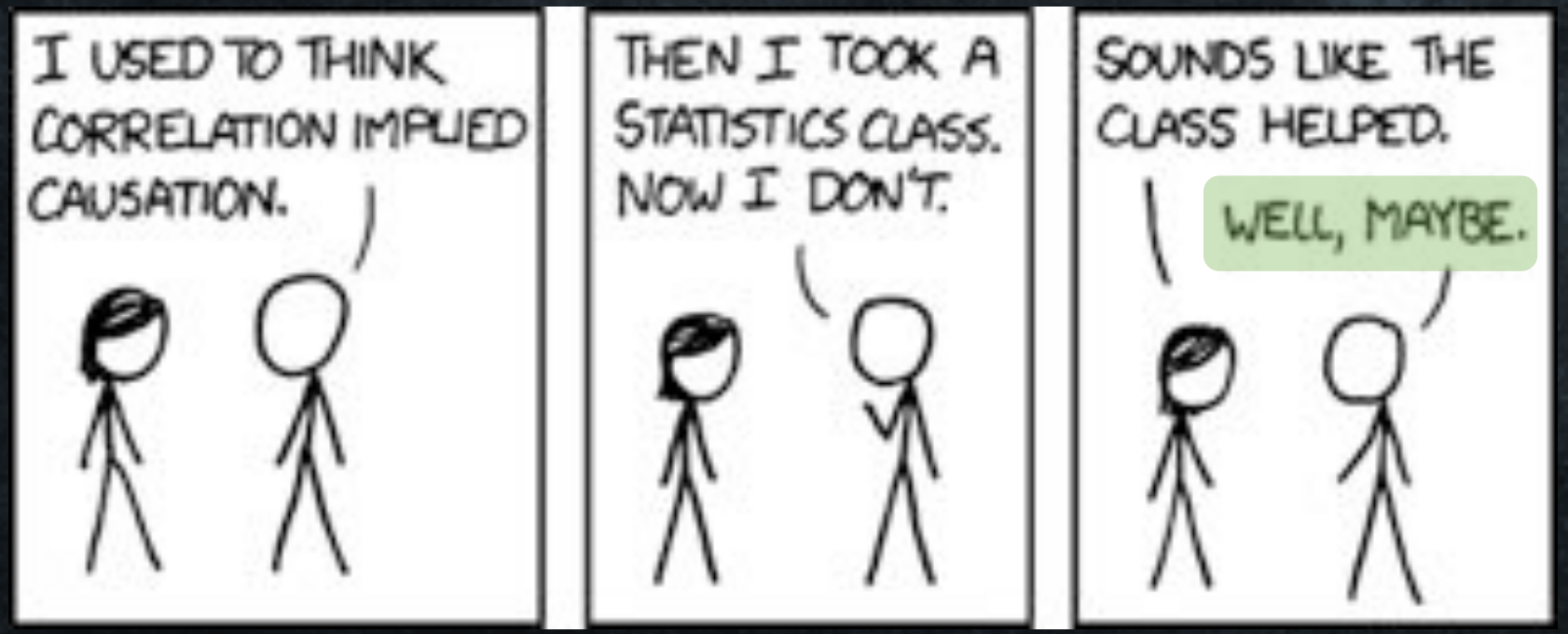

Be kind to
future you
~~collaborators~~

da·ta sci·en·tist

n. *a person who effectively extracts, manipulates, draws inference from, & communicates about data*



https://xkcd.com/552/



https://xkcd.com/552/

Lucy 🌻 @LucyStats · 11 Oct 2017

Kaplan claims: As statisticians we are too focused on "abstinence only" method, we need to teach **safe** causality #SSI2017 #ASASymposium2017

Lucy 🌻 @LucyStats

Daniel Kaplan makes an exciting claim: correlation *is* causation! (cc: @theeffortreport) #SSI2017 #ASASymposium2017

Show this thread

13 45

The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data

Causal inference is a core task of science. However, authors and editors often refrain from explicitly acknowledging the causal goal of research projects; they refer to causal effect estimates as associational estimates.

This commentary argues that using the term "causal" is necessary to improve the quality of observational research.

Miguel A. Hernán, MD, DrPH

See also Galea and Vaughan, p. 602; Begg and March, p. 620; Ahern, p. 621; Chiolero, p. 622; Glymour and Hamad, p. 623; Jones and Schooling, p. 624; and Hernán, p. 625.

You know the story:

> Dear author: Your observational study cannot prove causation. Please replace all references to causal effects by references to associations.

Many journal editors request

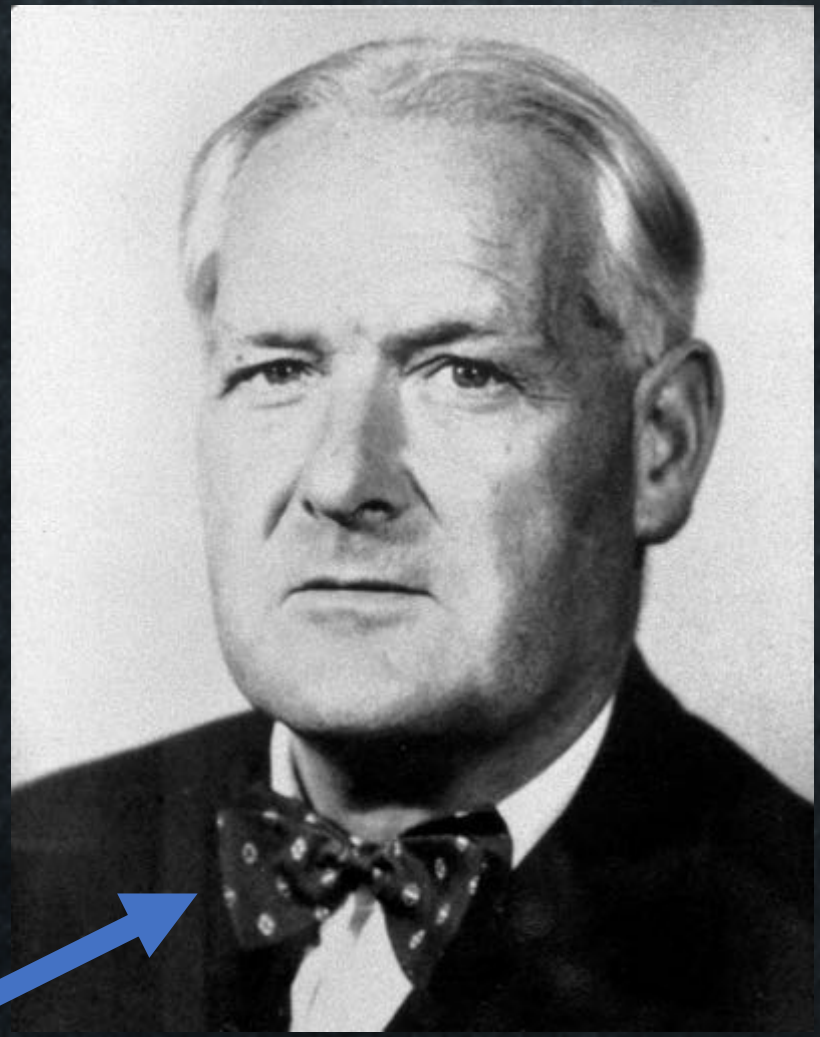
Confusion then ensues at the most basic levels of the scientific process and, inevitably, errors are made.

We need to stop treating "causal" as a dirty word that respectable investigators do not

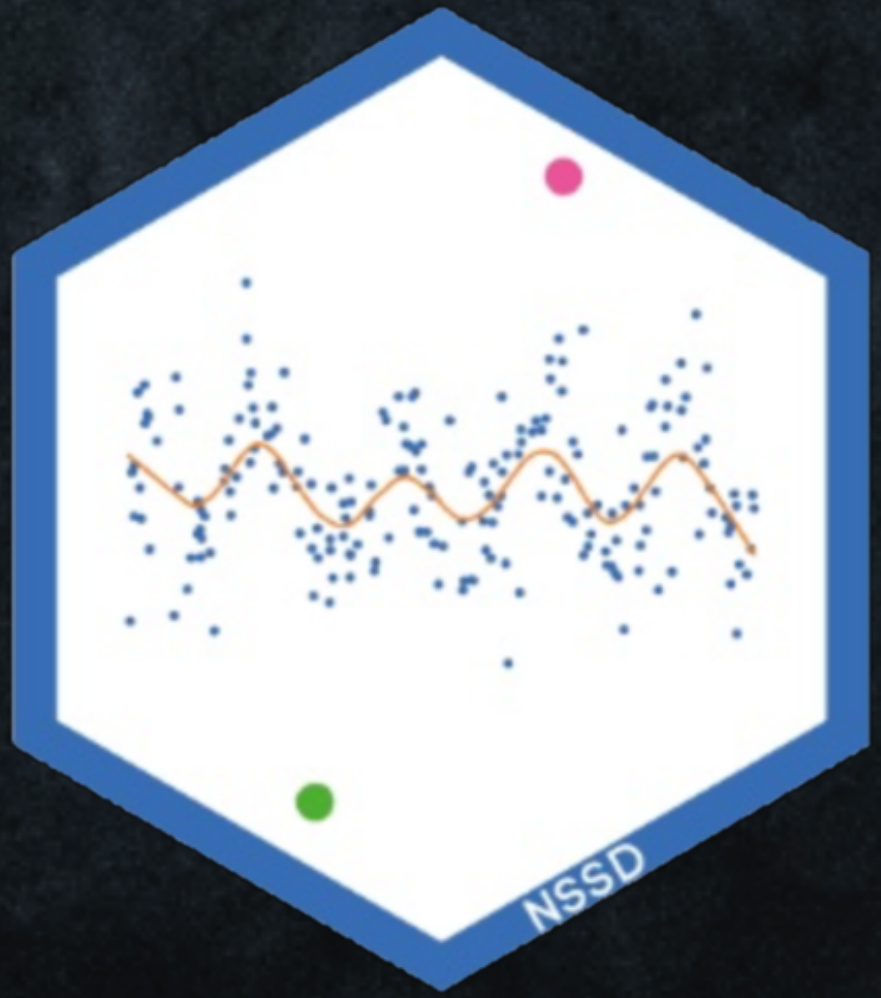
glass of red wine per day versus no alcohol drinking. For simplicity, disregard measurement error and random variability—that is, suppose the 0.8 comes from a very large population so that the 95% confidence interval around it


causal inference

Hill's Criteria

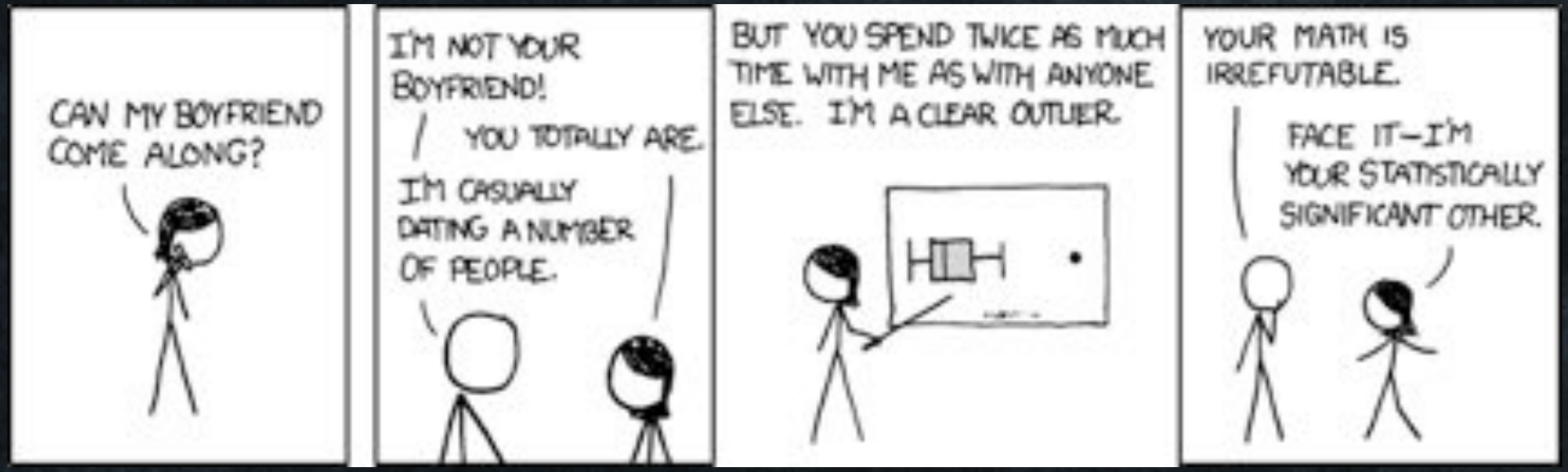


https://commons.wikimedia.org/wiki/File:Austin_Bradford_Hill.jpg



http://nssdeviations.com

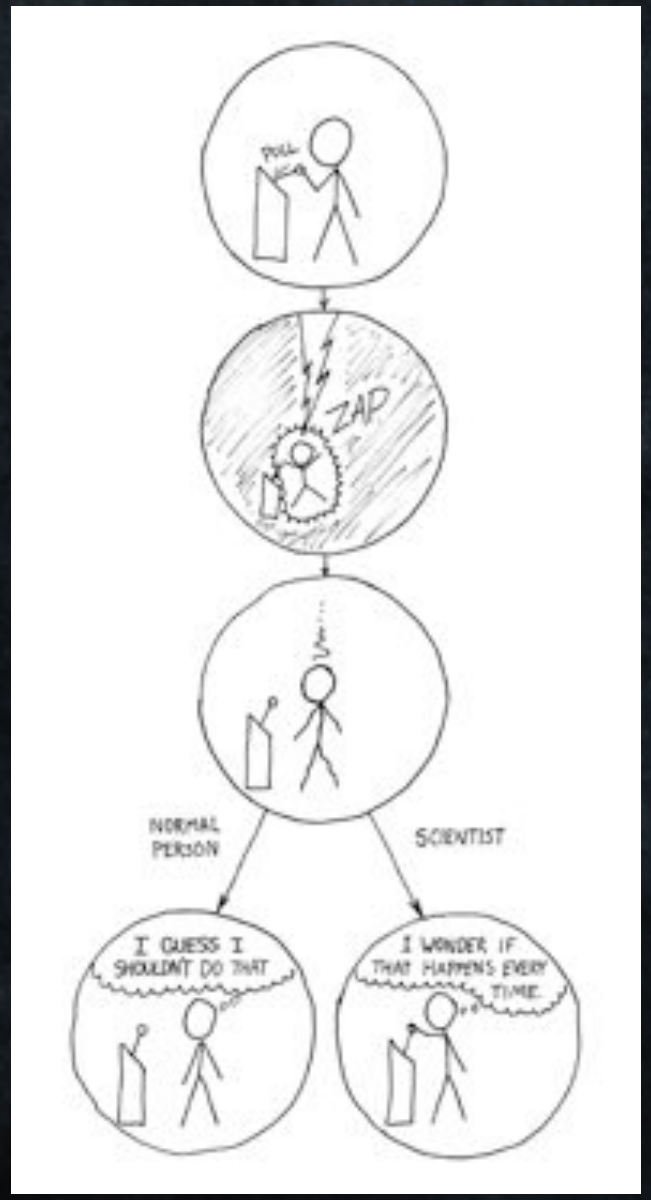
http://livefreeordichotomize.com/2016/12/15/hill-for-the-data-scientist-an-xkcd-story/



https://xkcd.com/539/

Strength

Consistency

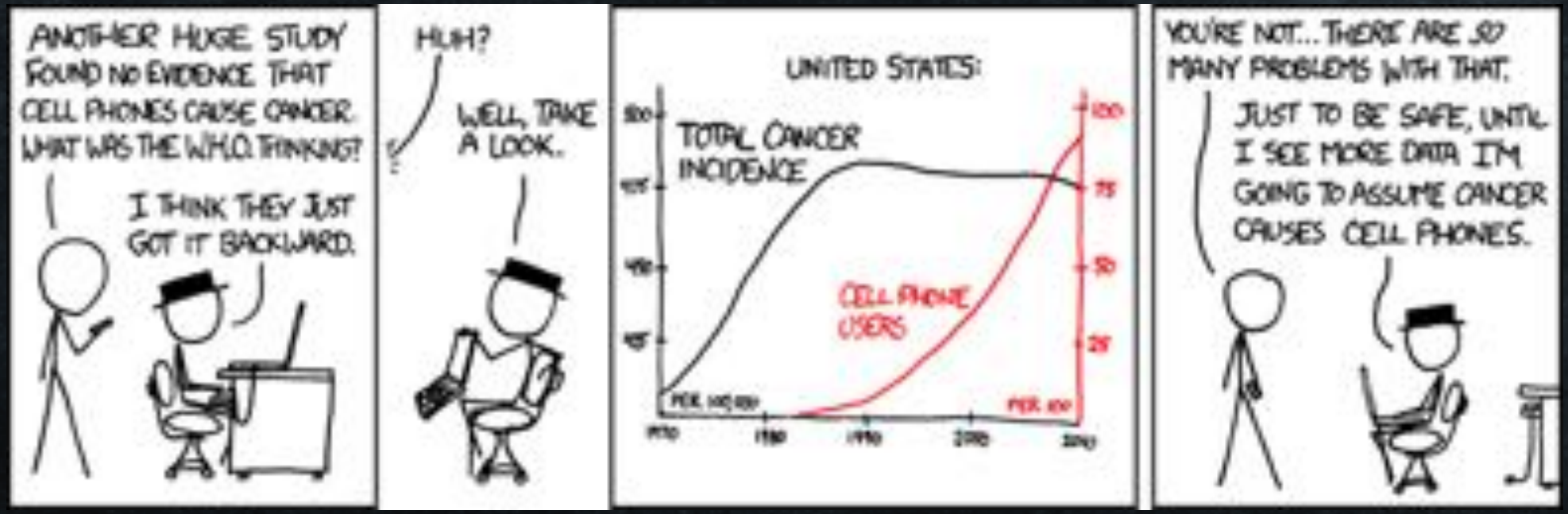


https://xkcd.com/242/

Specificity

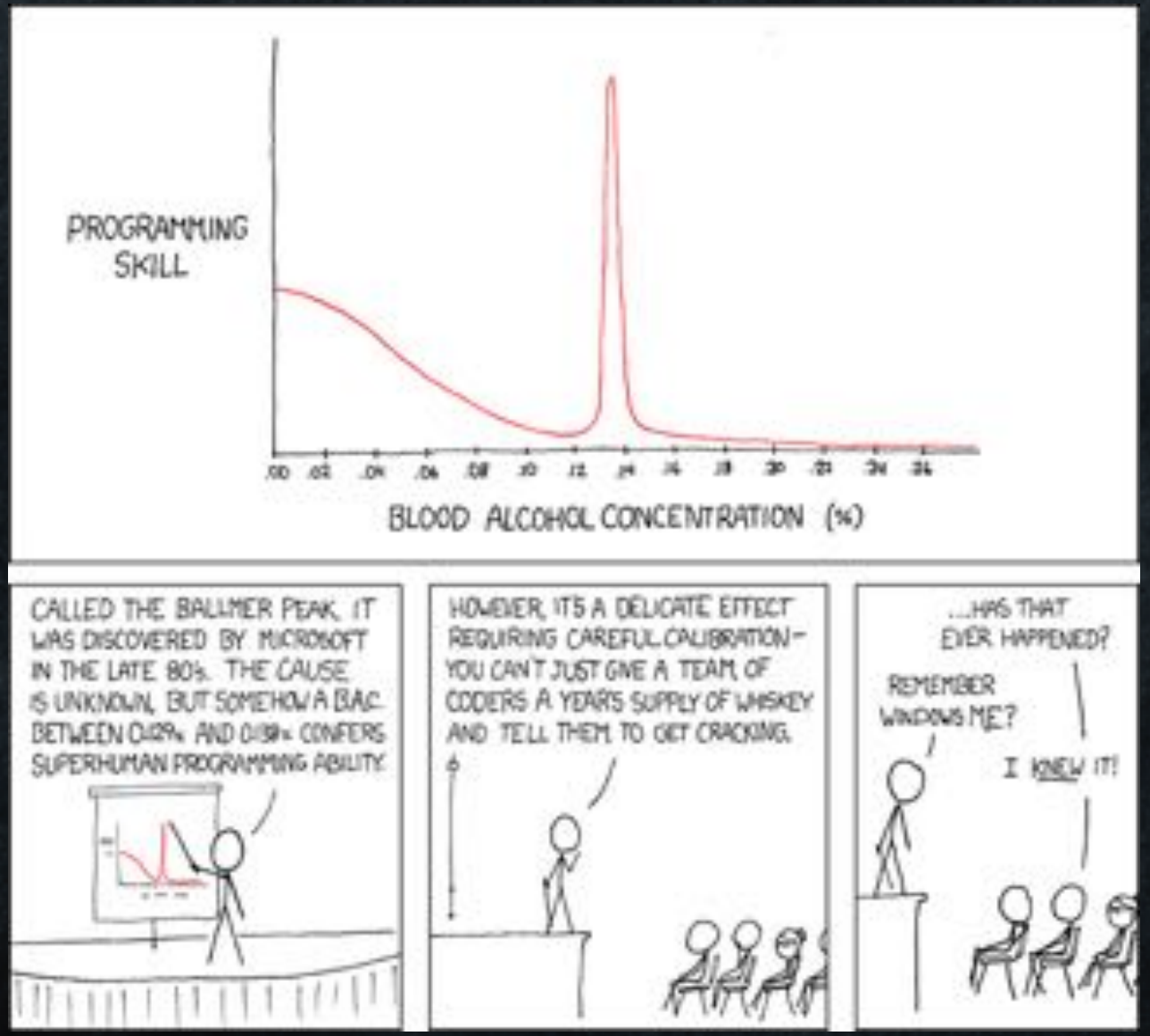
https://xkcd.com/1217/

Temporality



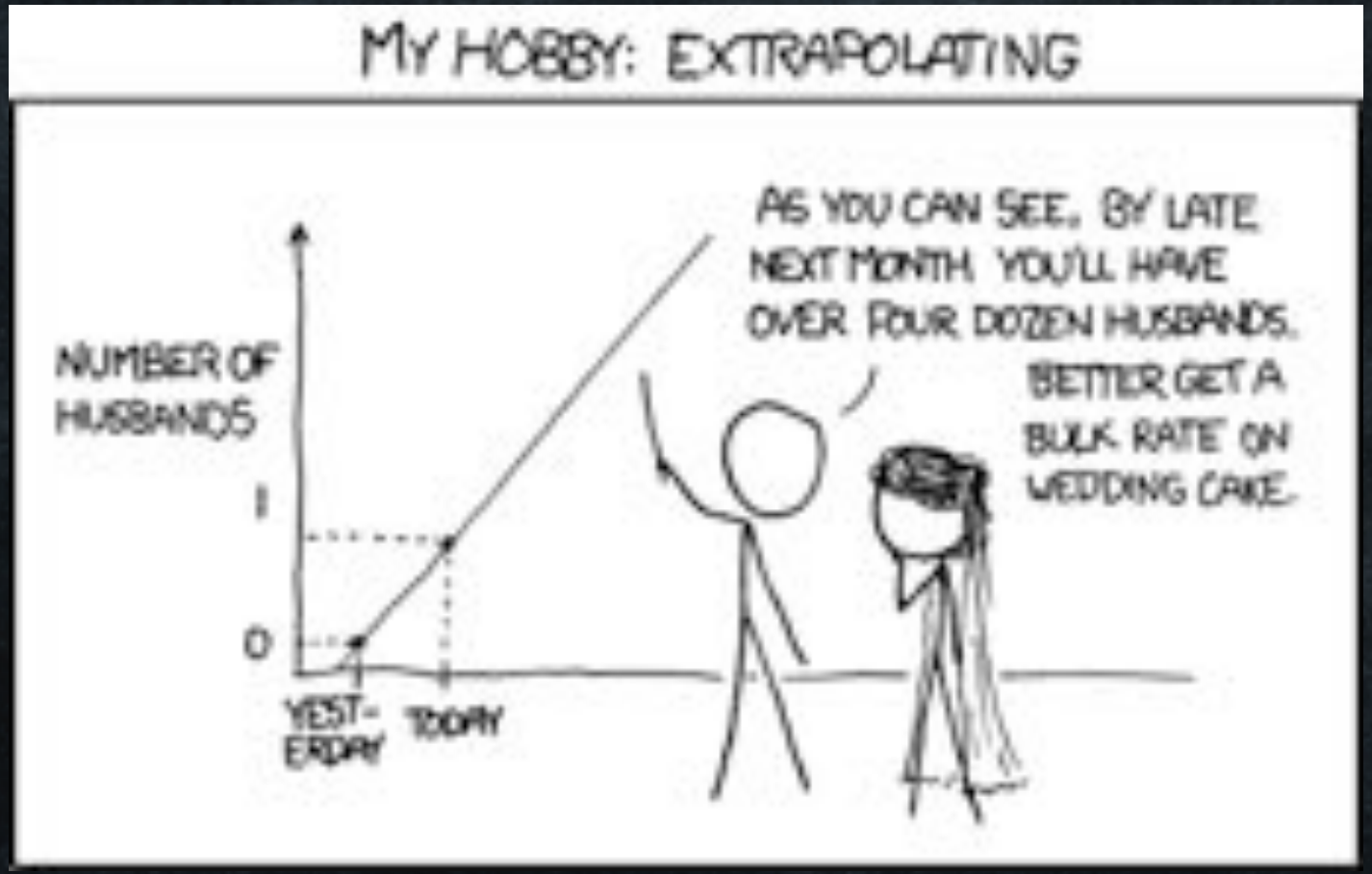
https://xkcd.com/925/

Biological gradient

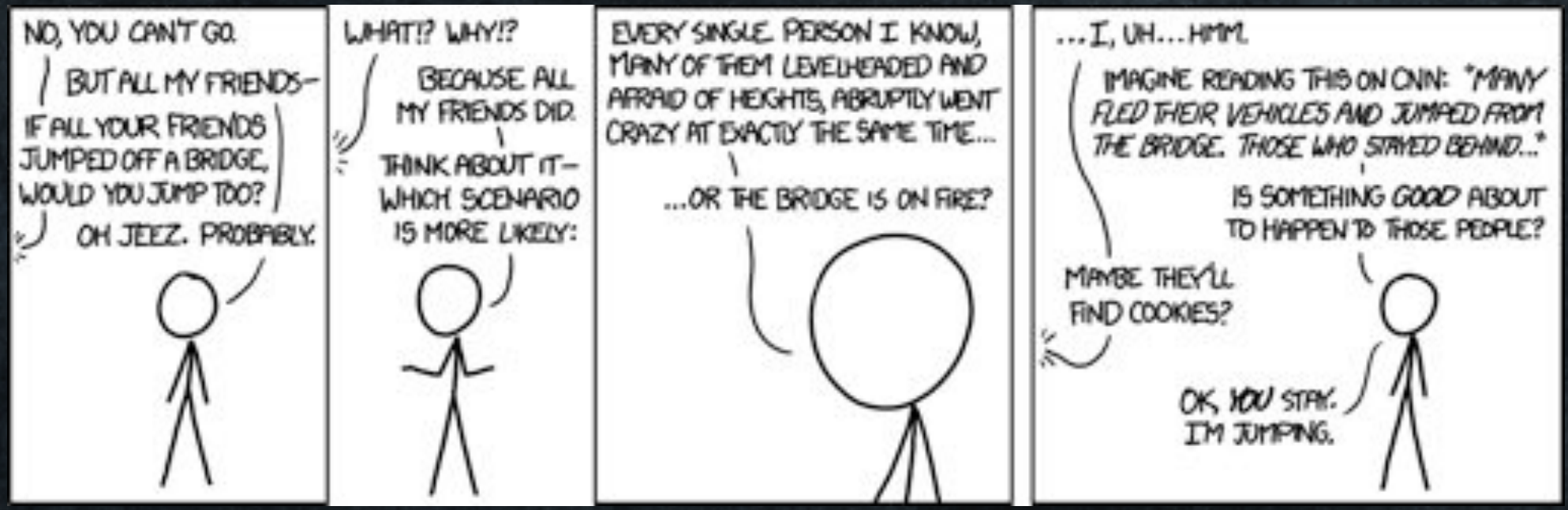


https://xkcd.com/323/

Plausibility



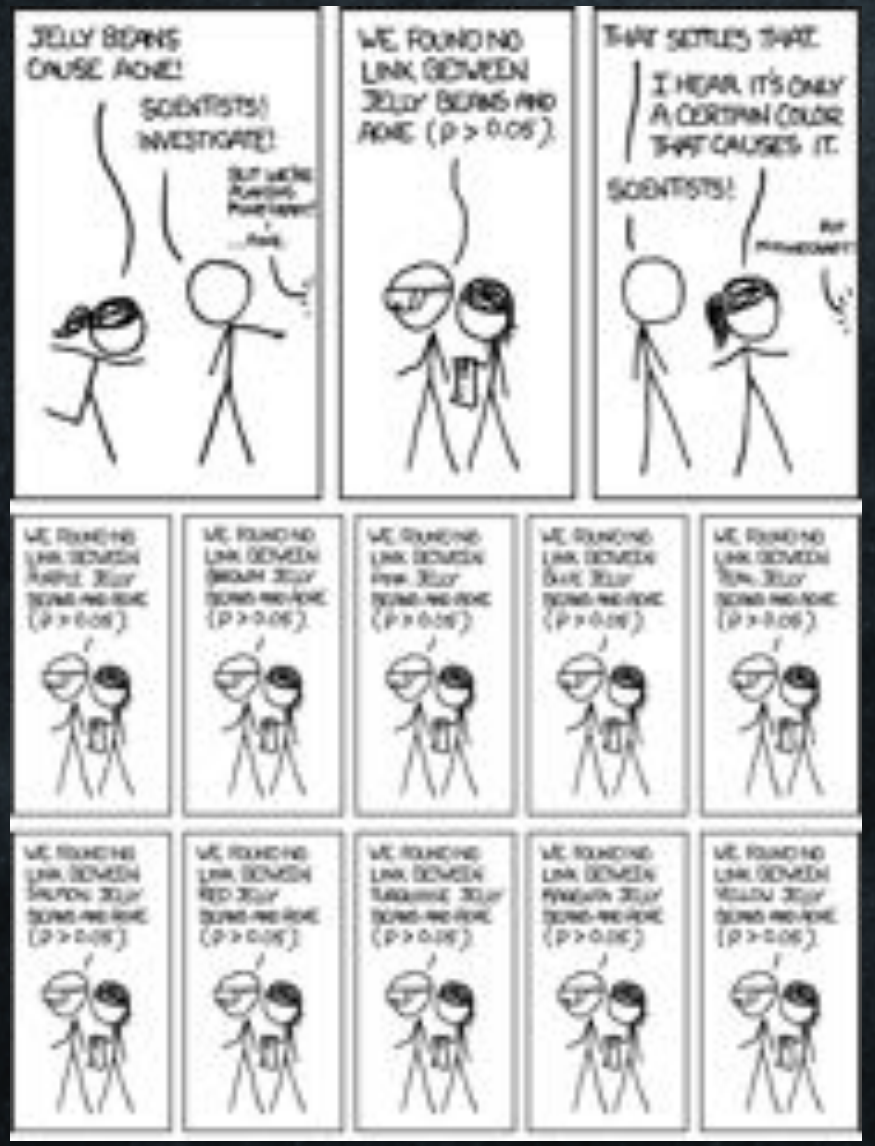
https://xkcd.com/605/

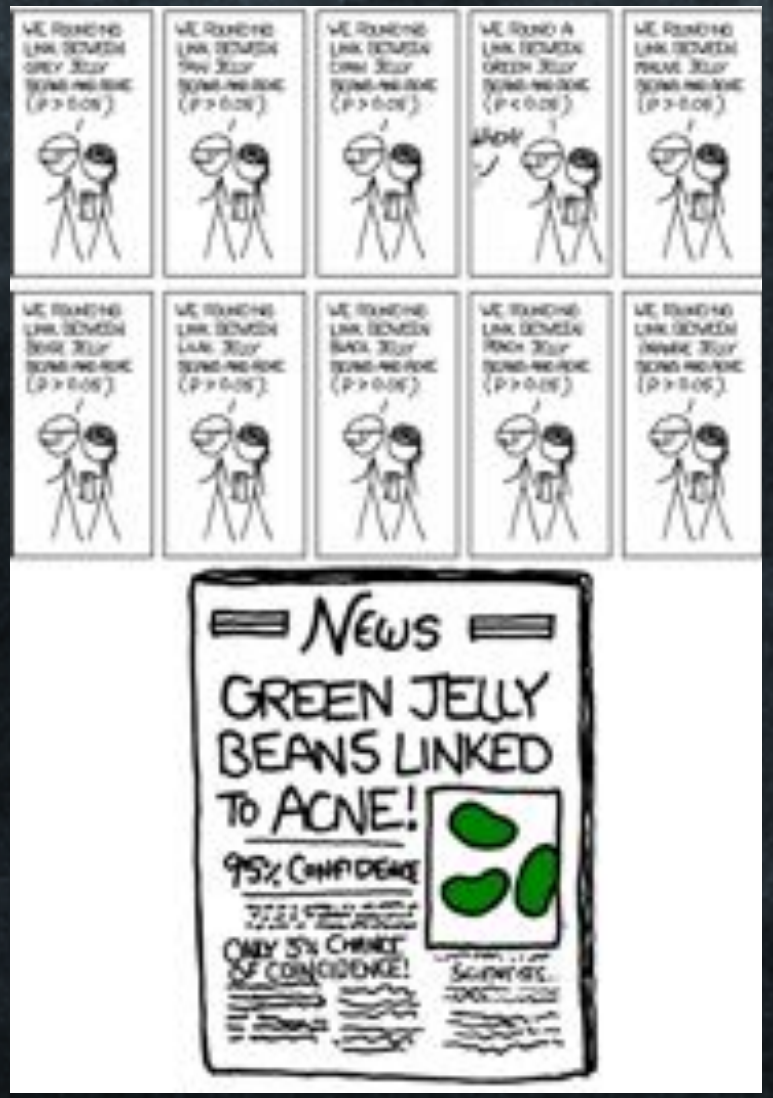


https://xkcd.com/1170/

Coherence

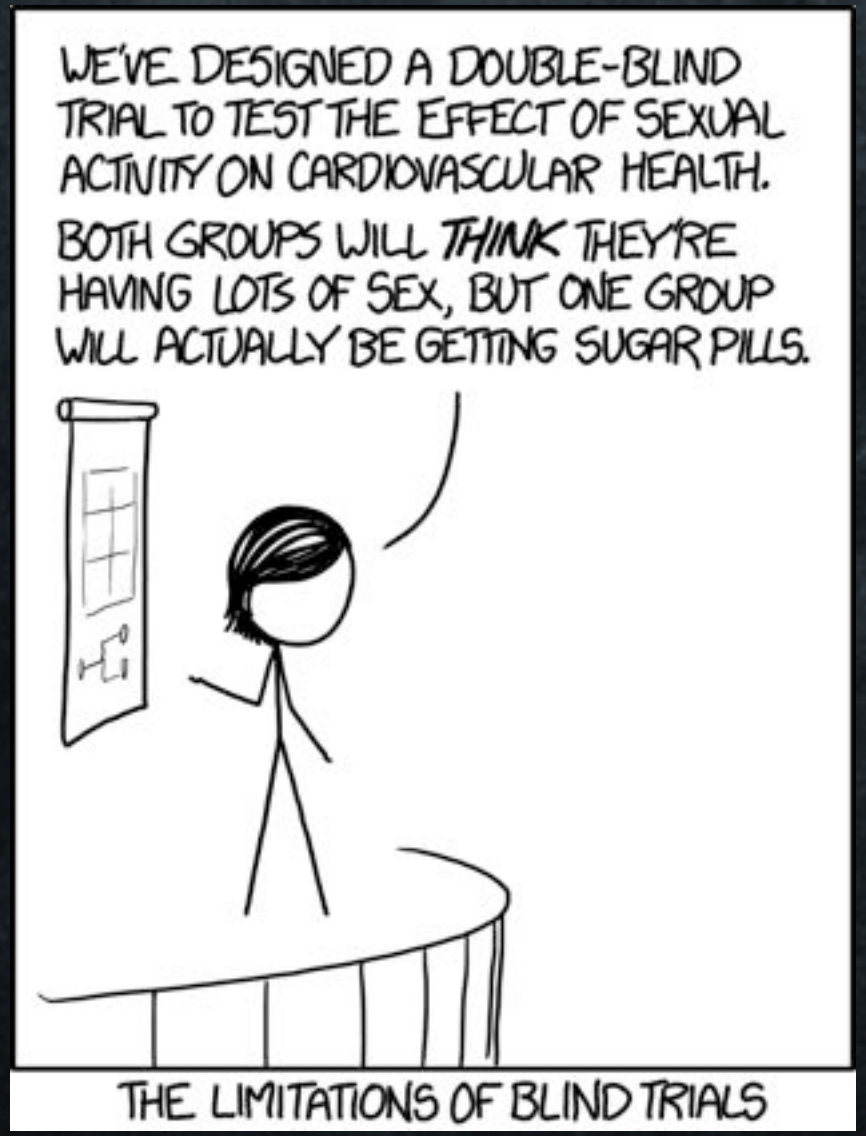
Analogy



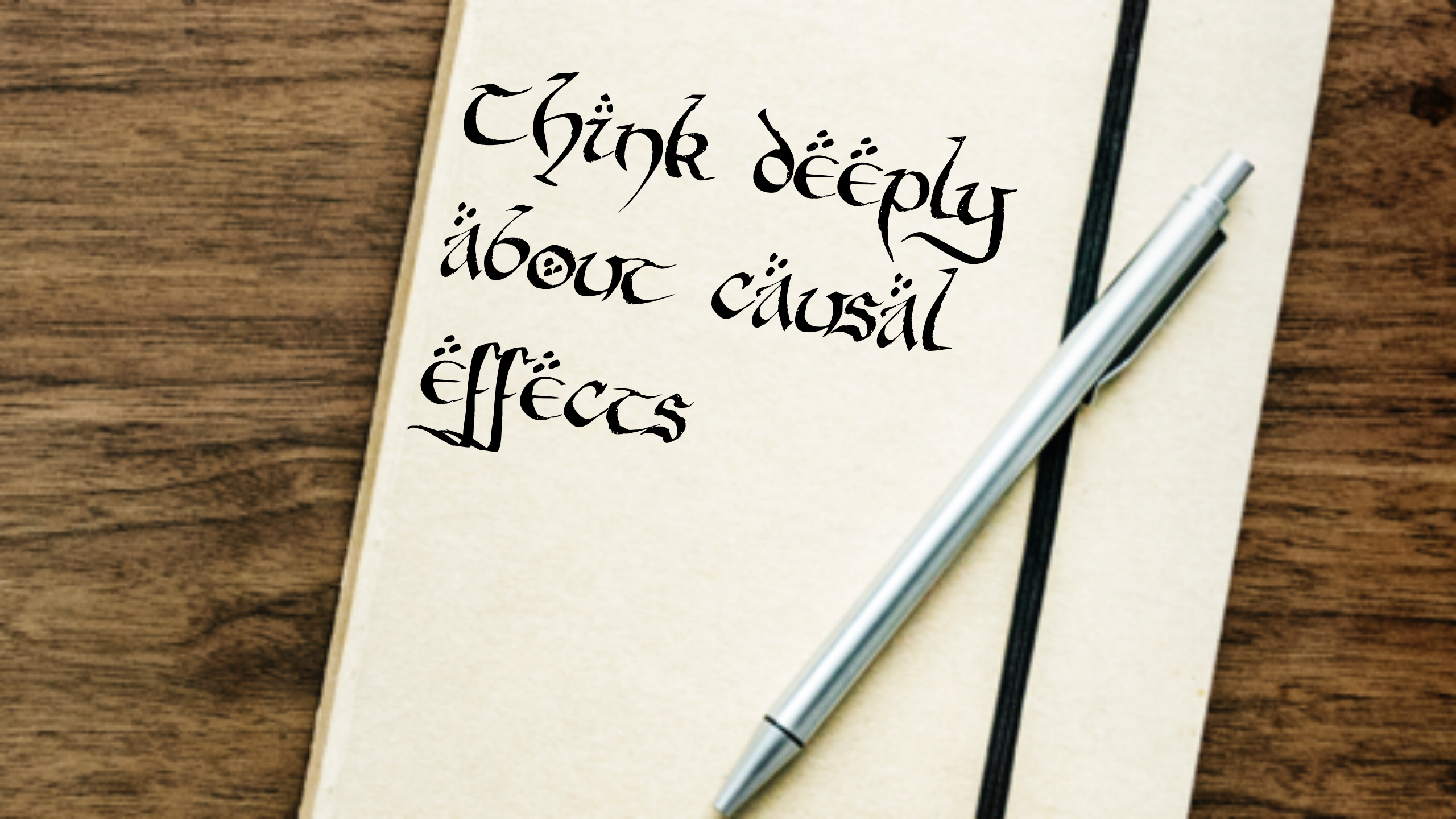


https://xkcd.com/882/

Experiment



https://xkcd.com/1462/


Think deeply
about causal
effects

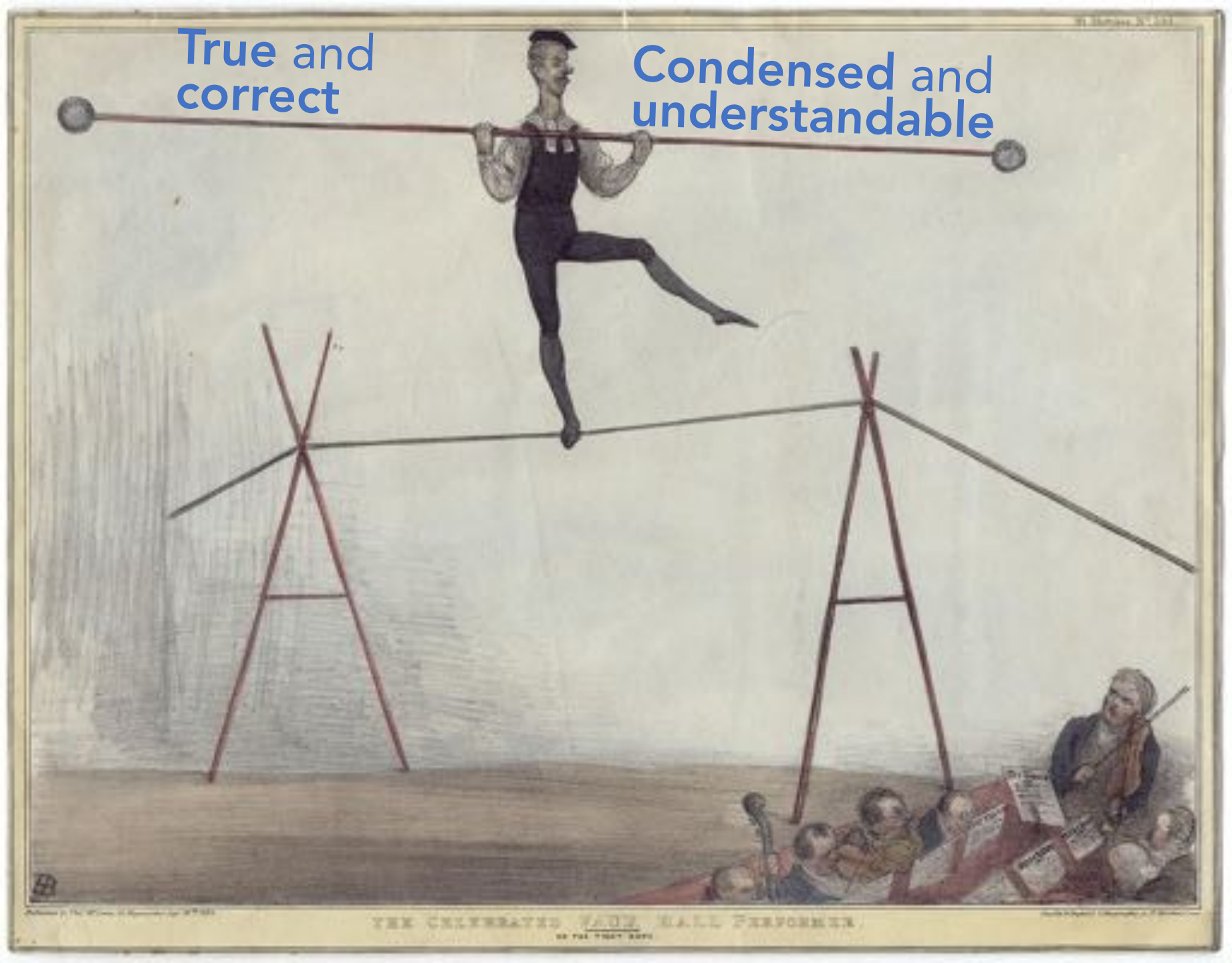
da·ta sci·en·tist

n. *a person who effectively extracts, manipulates, draws inference from, & communicates about data*


communication

Opper, Frederick Burr. *The fin de siècle newspaper proprietor.* 1984

Doyle, John. The Celebrated Vaux Hall Performer on the Tight Rope



Doyle, John. The Celebrated Vaux Hall Performer on the Tight Rope


there and back again
a data scientist's tale

"Macalester was founded in 1874 on a firm belief in the transformational power of the **liberal arts** education."

https://www.macalester.edu/about/


Communication
is key

da·ta sci·en·tist

n. *a person who effectively extracts, manipulates, draws inference from, & communicates about data*

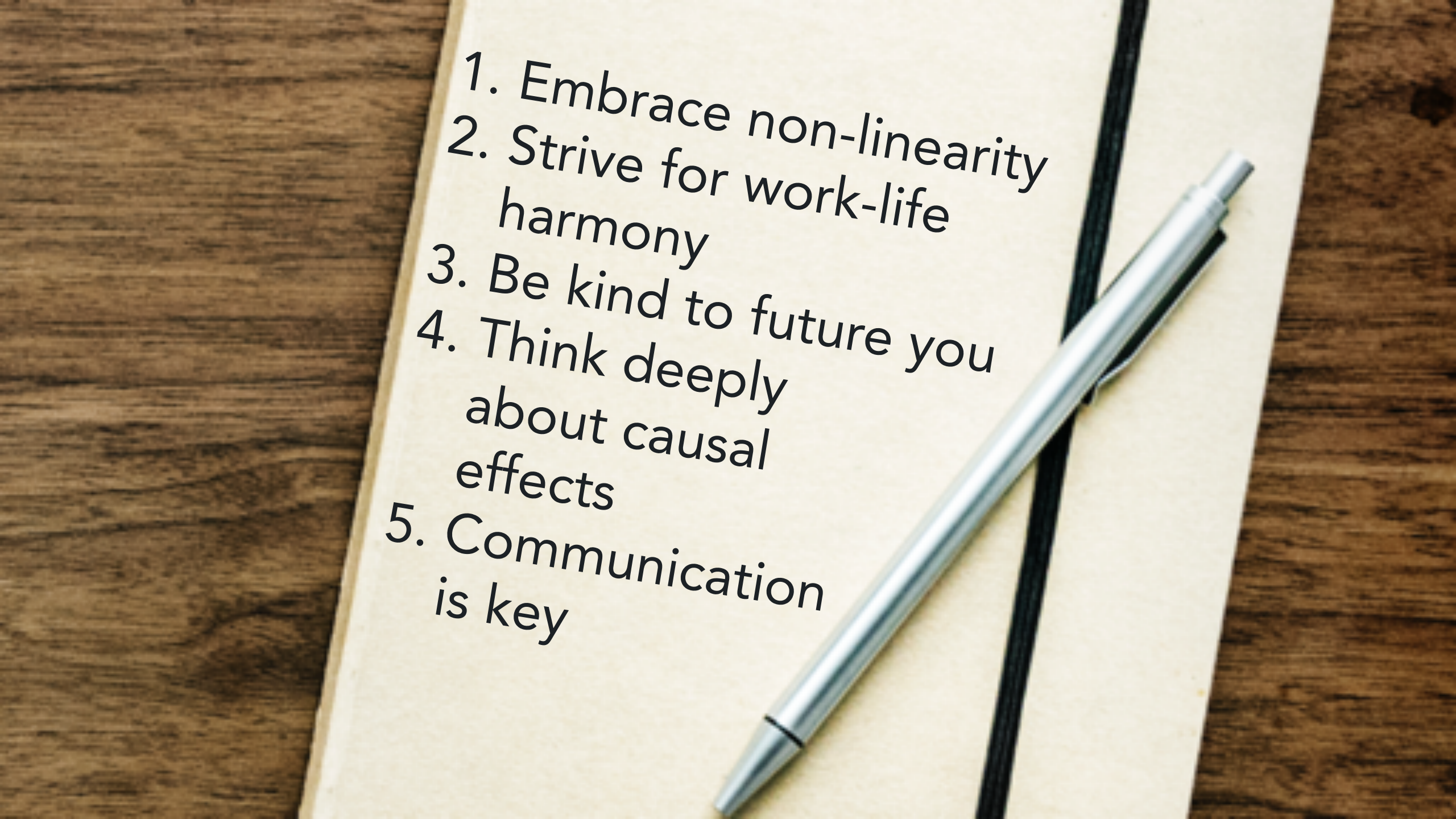

1. Embrace non-linearity
2. Strive for work-life harmony
3. Be kind to future you
4. Think deeply about causal effects
5. Communication is key


1. Embrace non-linearity
2. Strive for work-life harmony
3. Be kind to future you
4. Think deeply about causal effects
5. Communication is key


1. Embrace non-linearity
2. Strive for work-life harmony
3. Be kind to future you
4. Think deeply about causal effects
5. Communication is key


1. Embrace non-linearity
2. Strive for work-life harmony
3. Be kind to future you
4. Think deeply about causal effects
5. Communication is key


1. Embrace non-linearity
2. Strive for work-life harmony
3. Be kind to future you
4. Think deeply about causal effects
5. Communication is key


1. Embrace non-linearity
2. Strive for work-life harmony
3. Be kind to future you
4. Think deeply about causal effects
5. Communication is key


lucymcgowan.com/talk
@LucyStats