

# chéRÉ and back again at a data scientist's cafe

by Lucy D'Agostino McGowan



[lucymcgowan.com/talk](http://lucymcgowan.com/talk)

M  
L





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



data scientists  
We're all <sup>^</sup>mad  
here



***I'm going on an adventure!***

Cips / Lëssons  
Lëärdeed



background



A chalkboard resting on a wooden surface, surrounded by a garland of white flowers. The board contains handwritten text and arrows indicating connections between various fields:

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

Arrows show connections from "religious studies & romance languages" to "math & statistics", "biostatistics", and "RStudio". Arrows also point from "math & statistics" to "data science", and from "RStudio" to "data science".

religious studies &  
romance languages

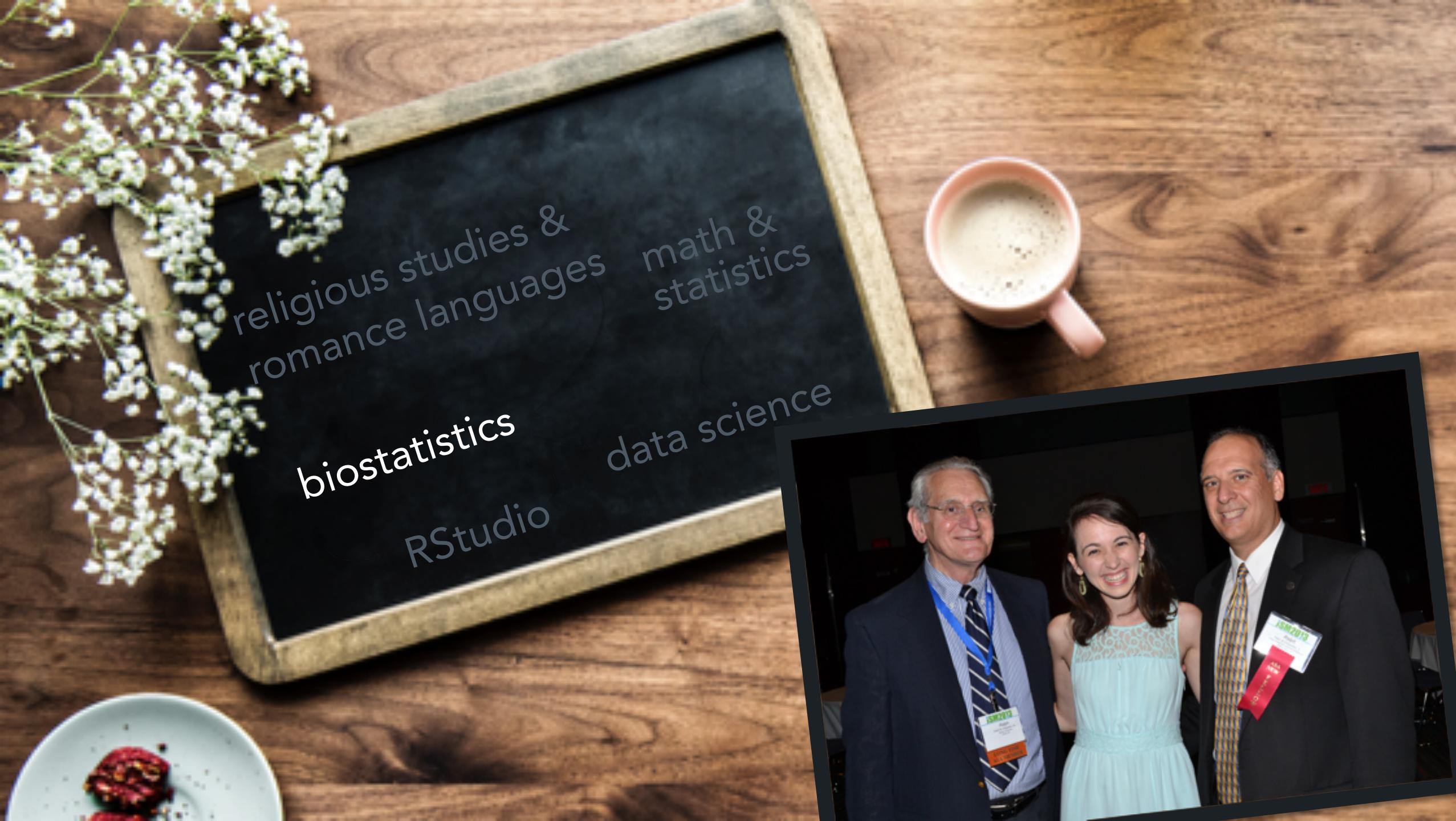
biostatistics

RStudio

data science

math &  
statistics





religious studies &  
romance languages

math &  
statistics

biostatistics

RStudio

data science





religious studies &  
romance languages

biostatistics

RStudio

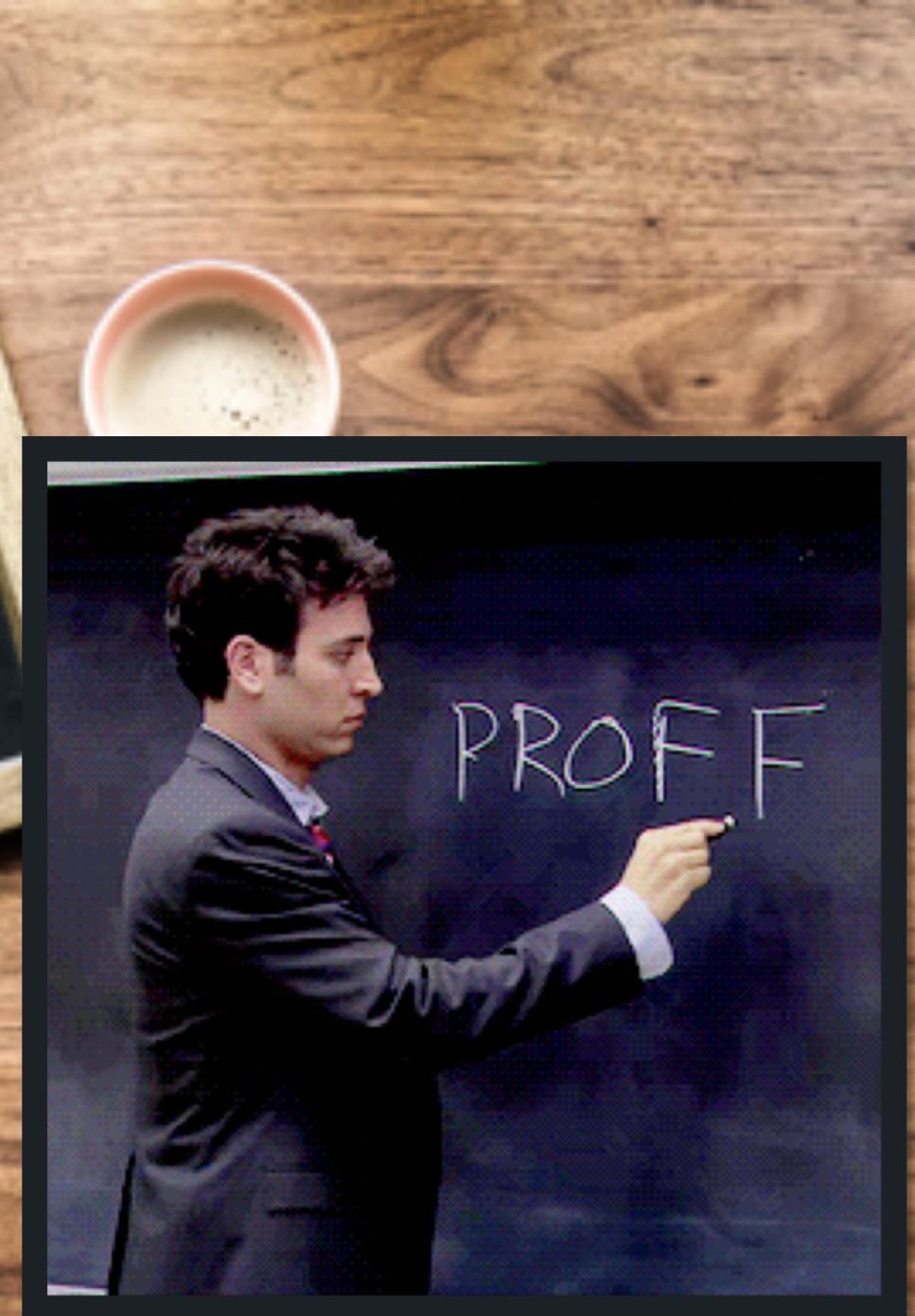
data science

math &  
statistics

google  
drive







Embrace  
non-linearity



A chalkboard resting on a wooden surface, surrounded by white flowers. The board contains handwritten text and arrows:

- "religious studies & romance languages" (with a curved arrow pointing to "biostatistics")
- "math & statistics" (with a curved arrow pointing to "data science")
- "biostatistics" (with a curved arrow pointing to "RStudio")
- "RStudio" (with a curved arrow pointing to "data science")
- "data science"

The text is written in white chalk on a dark board.

religious studies &  
romance languages

math &  
statistics

biostatistics

data science

RStudio





religious studies &  
romance languages

math &  
statistics

biostatistics  
RStudio

data





SCRÜWE for  
WORK-LİFE  
HÄRMONY



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 ^degree of freedom

n. the number of decisions  
*independent values<sup>or</sup>  
quantities* that can vary in  
an analysis



REPRODUCIBILITY



Hadley Wickham



@hadleywickham

Following



Replying to @hspter

@hspter reproducibility 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

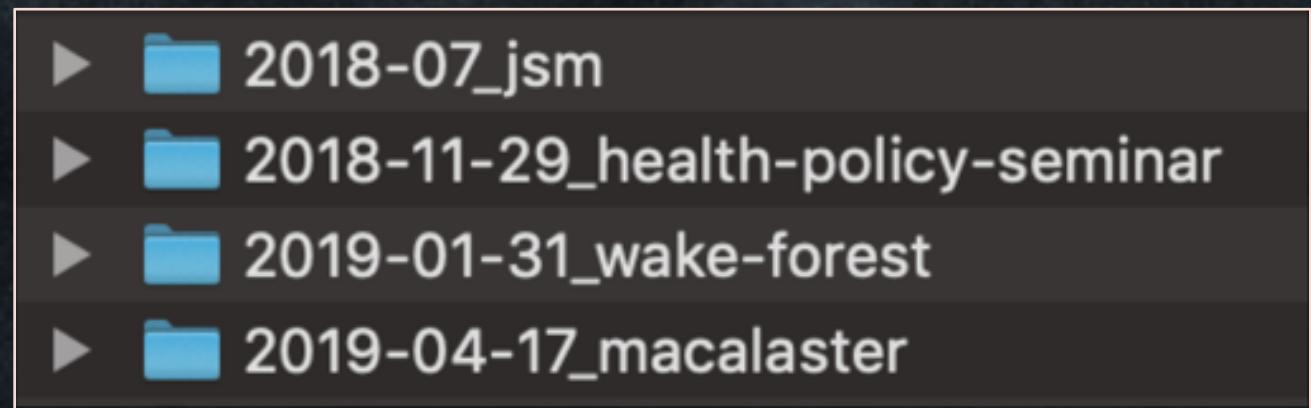
- machine readable
- human readable
- play well with default ordering

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](http://phdcomics.com/comics/archive_print.php?comicid=1531)

# Version control



Cham, Jorge. notfinal.doc.  
[http://phdcomics.com/comics/archive\\_print.php?comicid=1531](http://phdcomics.com/comics/archive_print.php?comicid=1531)

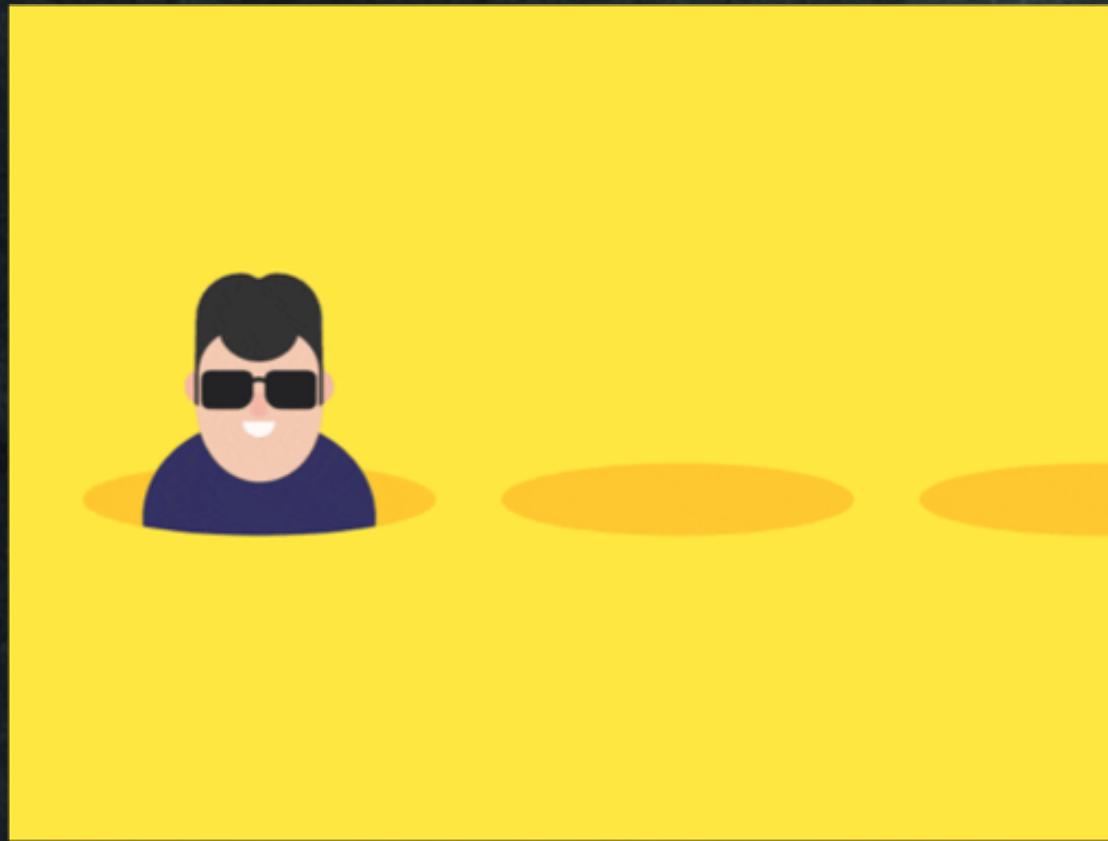
# Version control



Dropbox



Cham, Jorge. notfinal.doc.  
[http://phdcomics.com/comics/archive\\_print.php?comicid=1531](http://phdcomics.com/comics/archive_print.php?comicid=1531)



# 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

I USED TO THINK  
CORRELATION IMPLIED  
CAUSATION.



THEN I TOOK A  
STATISTICS CLASS.  
NOW I DON'T.



SOUNDS LIKE THE  
CLASS HELPED.  
WELL, MAYBE.



<https://xkcd.com/552/>

I USED TO THINK  
CORRELATION IMPLIED  
CAUSATION.



THEN I TOOK A  
STATISTICS CLASS.  
NOW I DON'T.



SOUNDS LIKE THE  
CLASS HELPED.

WELL, MAYBE.



<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; Abern, 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.

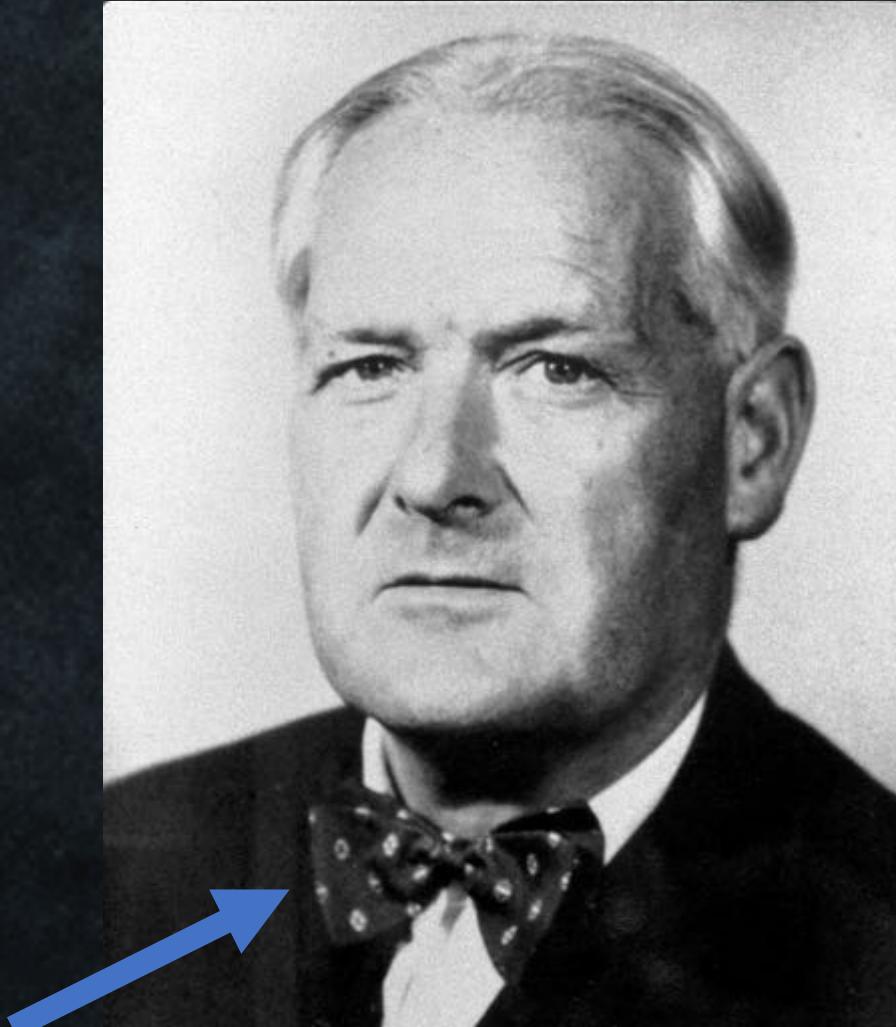
We need to stop treating "causal" as a dirty word that respectable investigators do not

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

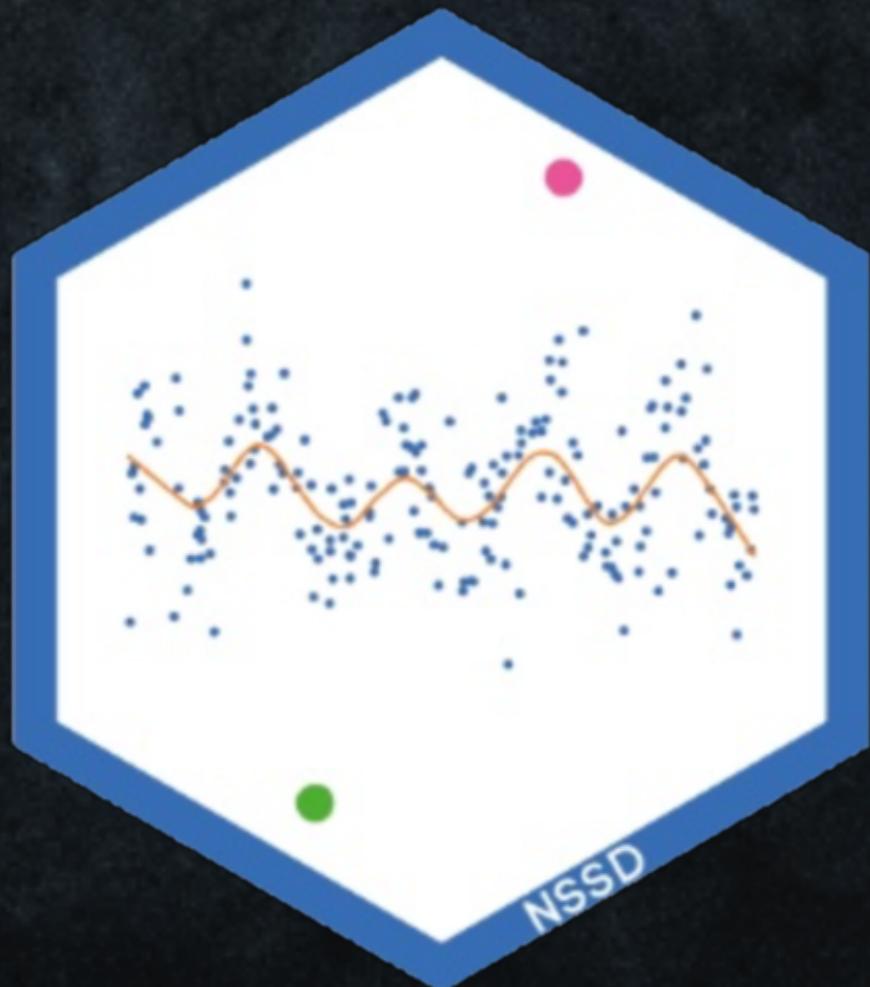


causal inference

# Hill's Criteria

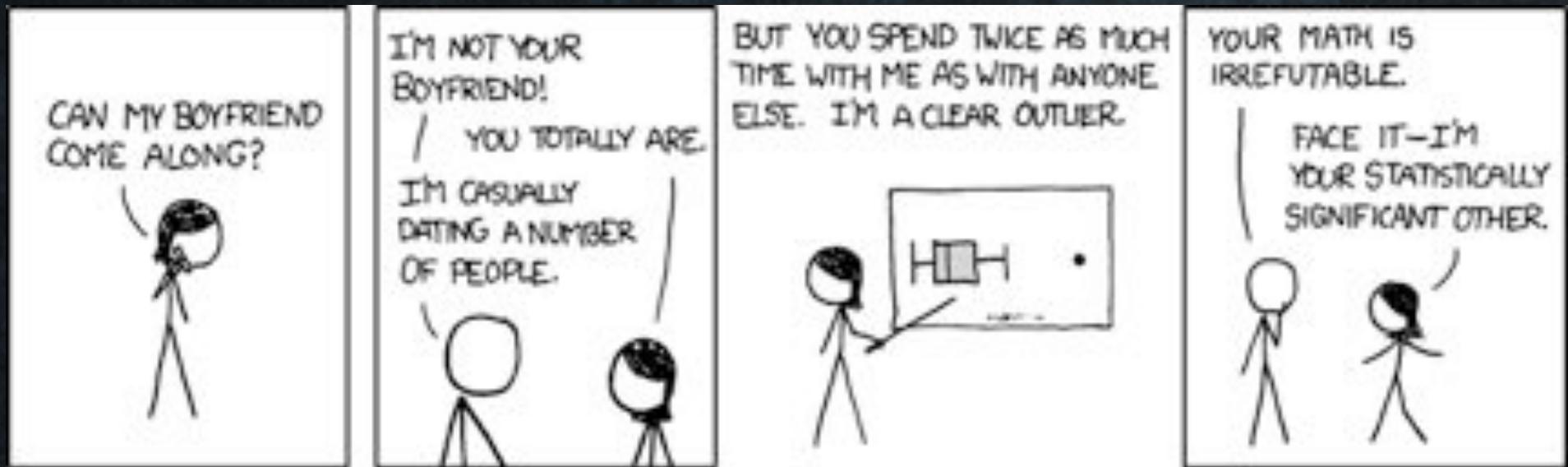


[https://commons.wikimedia.org/wiki/File:Austin\\_Bradford\\_Hill.jpg](https://commons.wikimedia.org/wiki/File:Austin_Bradford_Hill.jpg)



<http://nssdeviations.com>

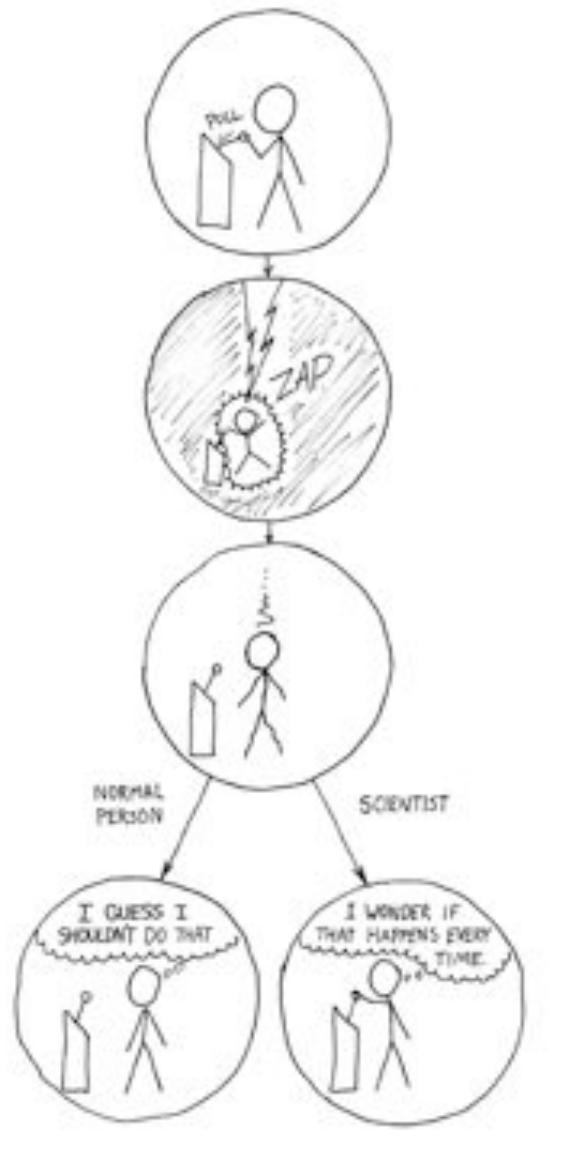
<http://livefreeordichotomize.com/2016/12/15/hill-for-the-data-scientist-an-xkcd-story/>



<https://xkcd.com/539/>

# Strength

# Consistency



<https://xkcd.com/242/>

# Specificity

WHEN YOU SEE A CLAIM THAT A  
COMMON DRUG OR VITAMIN "KILLS  
CANCER CELLS IN A PETRI DISH,"

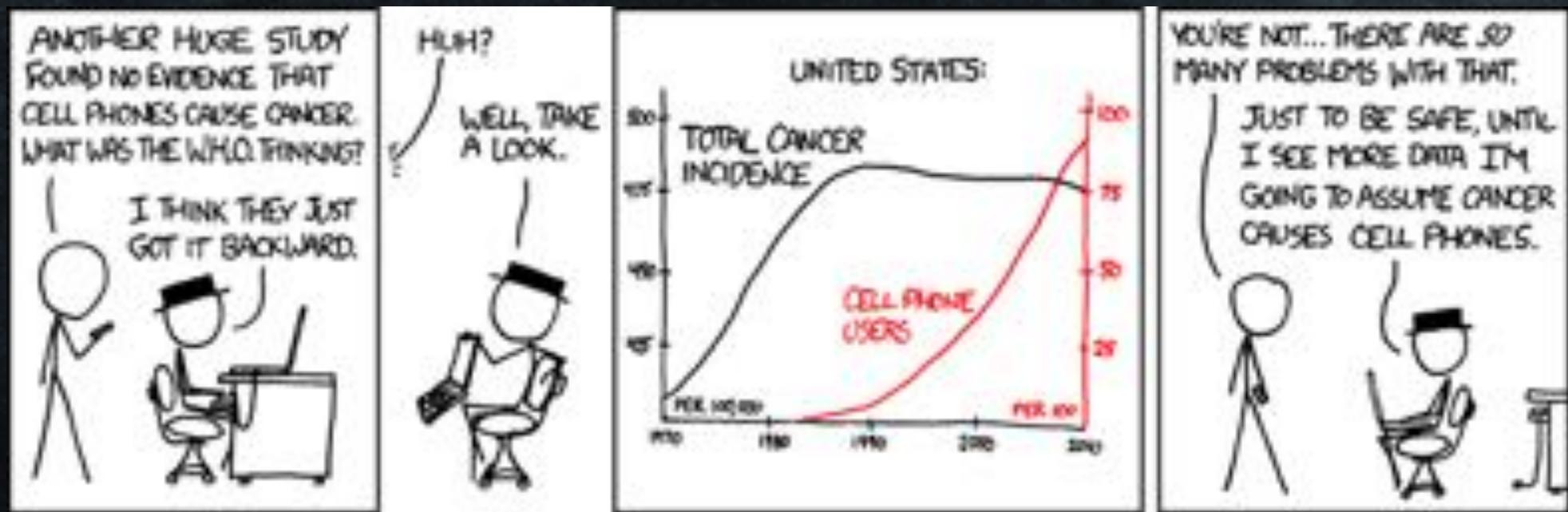
KEEP IN MIND:



SO DOES A HANDGUN.

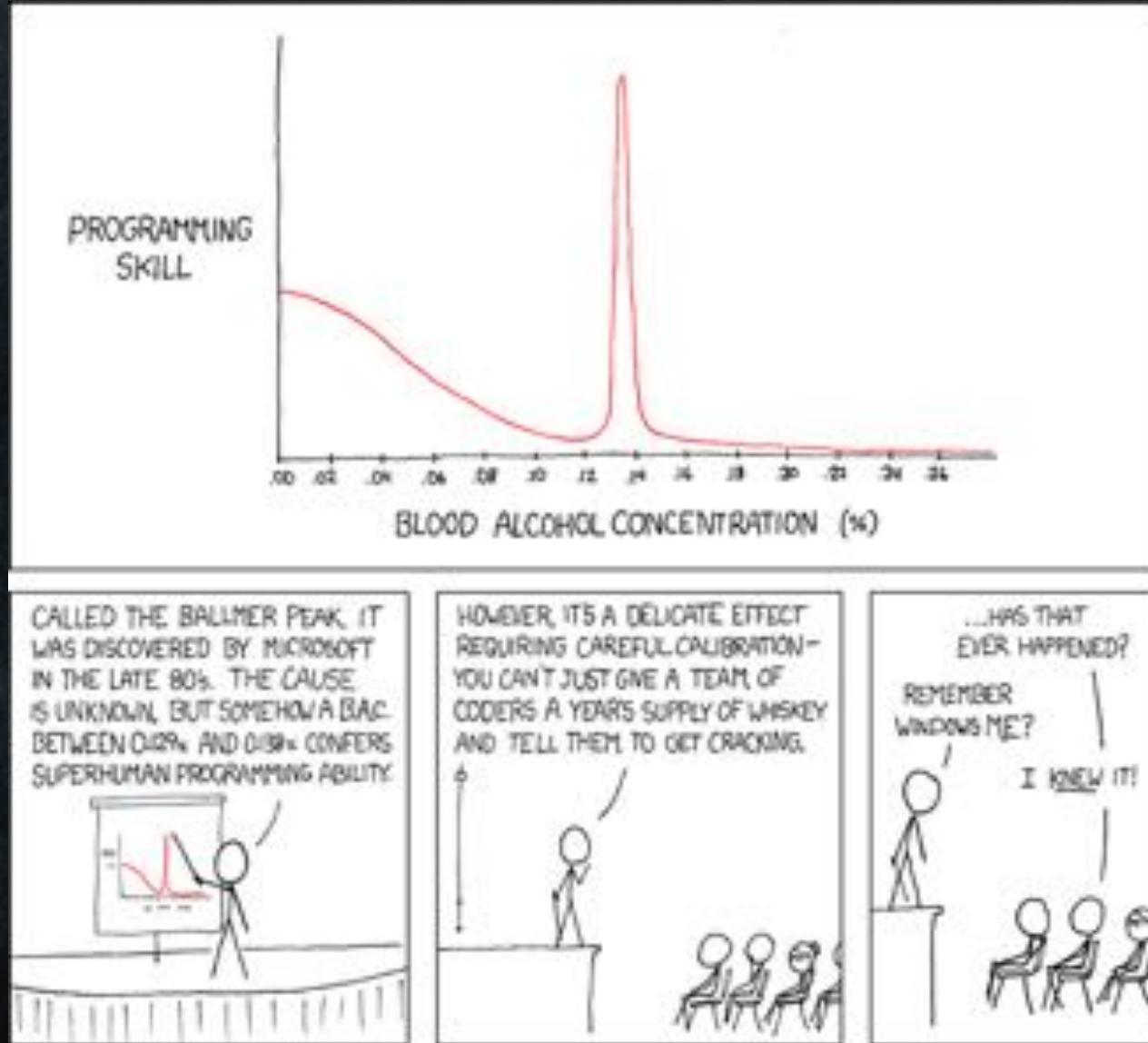
<https://xkcd.com/1217/>

# Temporality



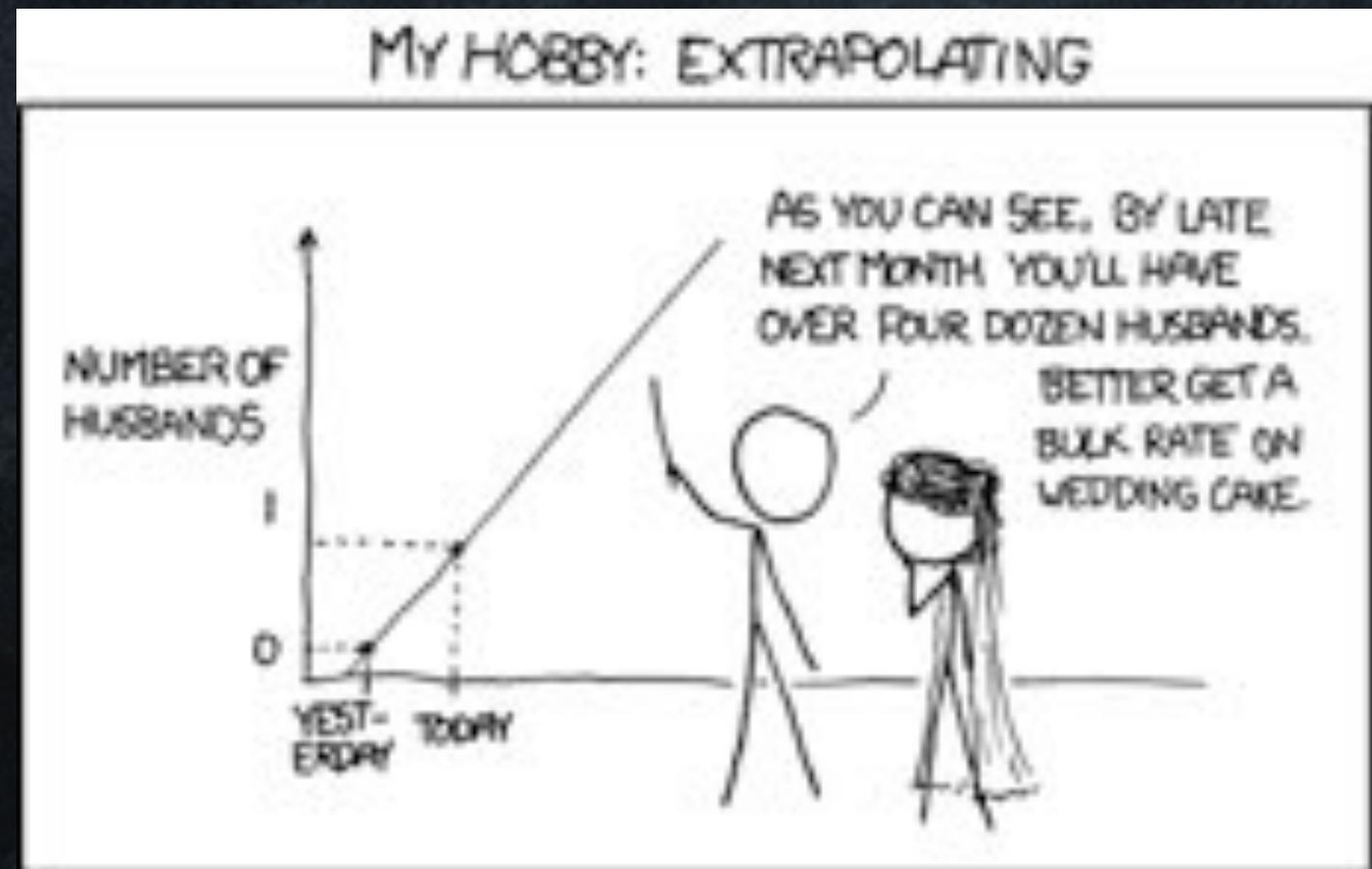
<https://xkcd.com/925/>

# Biological gradient



<https://xkcd.com/323/>

# Plausibility



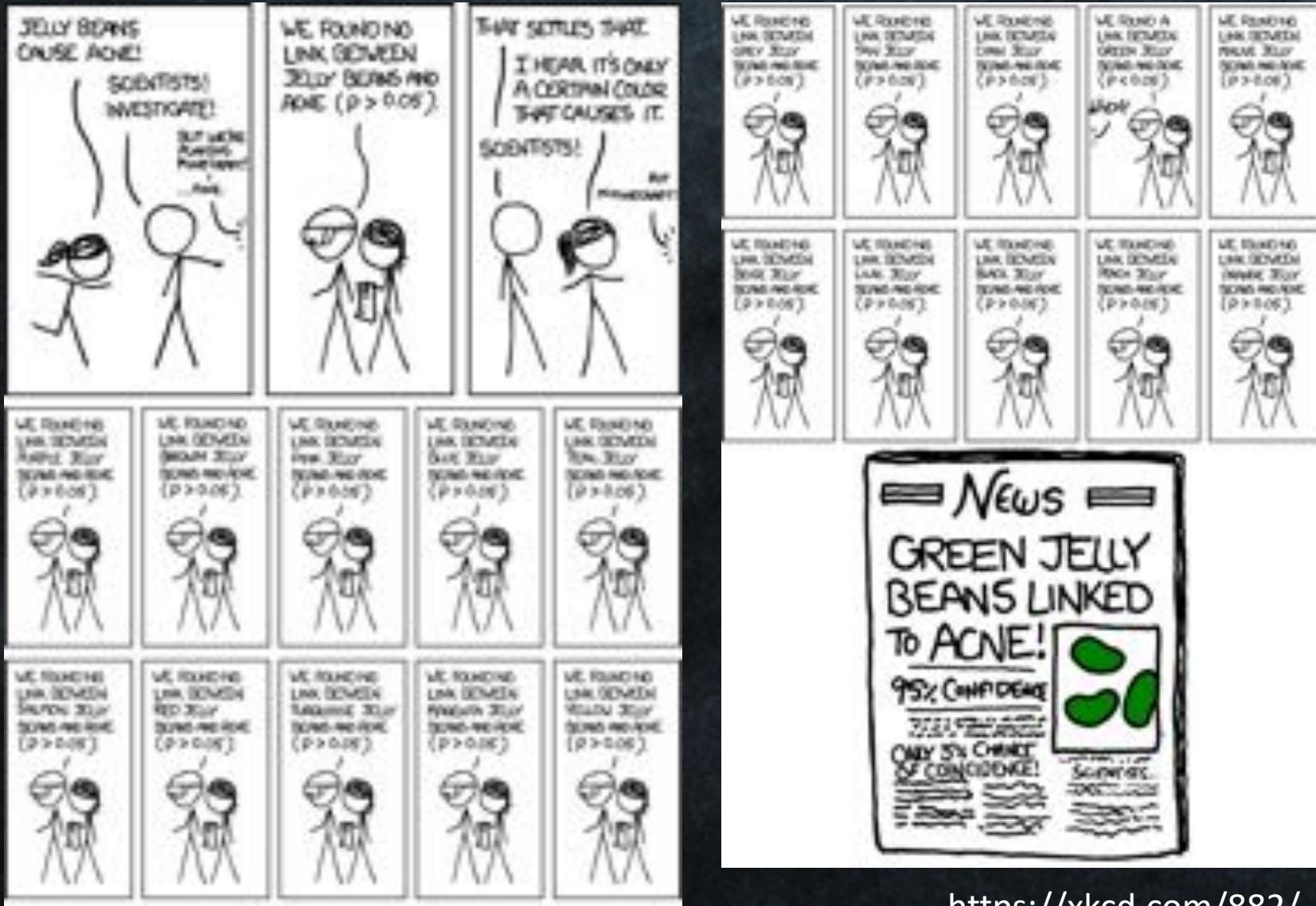
<https://xkcd.com/605/>



<https://xkcd.com/1170/>

# Coherence

# Analogy



<https://xkcd.com/882/>

# Experiment

WE'VE DESIGNED A DOUBLE-BLIND TRIAL TO TEST THE EFFECT OF SEXUAL ACTIVITY ON CARDIOVASCULAR HEALTH. BOTH GROUPS WILL *THINK* THEY'RE HAVING LOTS OF SEX, BUT ONE GROUP WILL ACTUALLY BE GETTING SUGAR PILLS.



THE LIMITATIONS OF BLIND TRIALS

Think deeply  
about causal  
effects



# da·ta sci·en·tist

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



communication



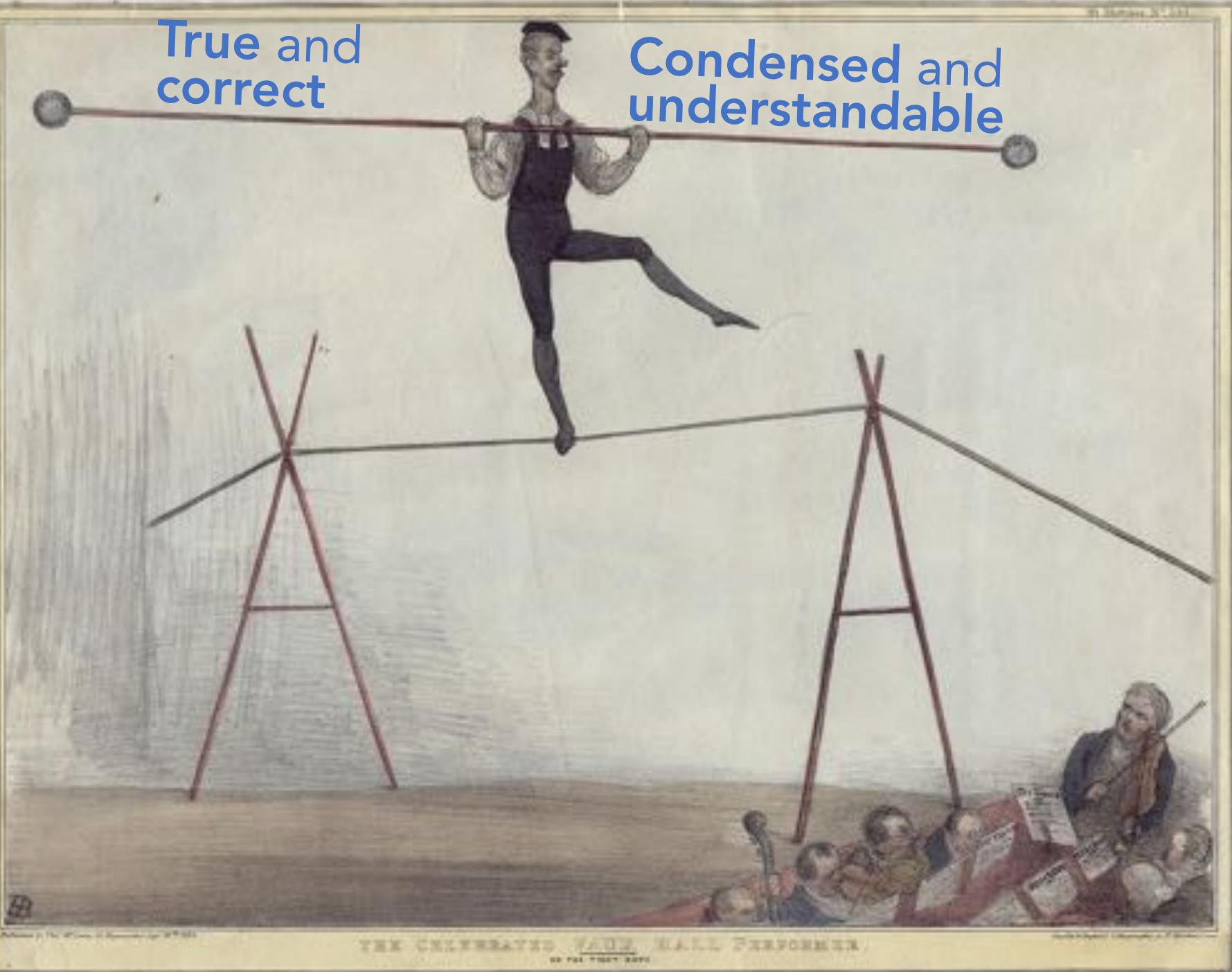
Opper, Frederick Burr. *The fin de siècle newspaper proprietor*. 1984



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

True and  
correct

Condensed and  
understandable



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

therē 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.”**



**MACALESTER**

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

- 
- A photograph of a light-colored wooden desk. On the desk, there is a open notebook with a black cover, a silver pen lying diagonally across it, and a small potted plant in the background.
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

- 
- A photograph of a light-colored wooden desk. On the desk, there is a open notebook with a black cover, a silver pen lying diagonally across it, and a small potted plant in the background.
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



@LucyStats

