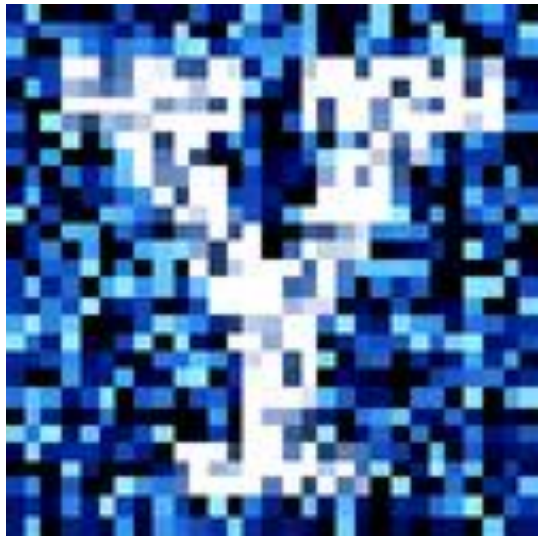
Ydata: Introduction to Data Science



Lecture 02: Cause and Effect

Overview

Brief demonstration on Jupyter Notebooks

The history of Data Science continued…

Experiments and observational studies

- Some motivating examples
- Association vs. causation
- Next class: John Snow and the Broad Street Pump

Assignment 0

Please try assignment 0 to test that you have a working environment to run Jupyter Notebooks

- This assignment does not need to be turned in
- Requires installing Anaconda and the datascience package, or using Google colab
- Instructions are on Canvas and assignment 0 is on the class calendar page

Let's take a quick look at using Jupyter Notebooks…


Questions

Continuation of the history of Data Science…

The history of Data Science

(a very incomplete list)

Ishango bone

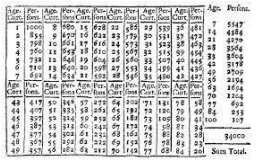
(20,000 BCE)

Cuneiform tablets

(4,000 BCE)

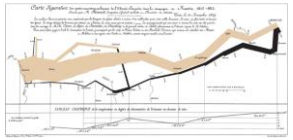
Quipus in South America

(1100-1500)



Demographics

(1600's)



Golden age of data visualization

(1850-1900)

Big data

(now)

Data

Probability

Computers

Key Take Away

Probability models dominated data analysis prior to using computational methods

Initial development

(1600's)

"Small data"

Probability in Statistics

(1820's – 1950's)

Math Stats dominates

(1900-1960's)

"Big data"

Abacus

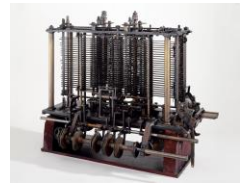
(2400 BCE)

Antikythera mechanism

(100 BCE)

Analytical Engine

(1800's)



Hollerith Tabulating Machine

(1890)

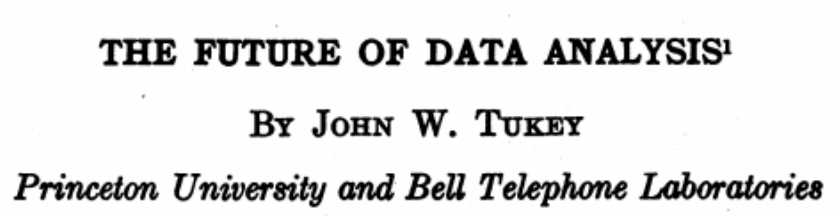
Mainframes, PCs, Internet, etc.

(1950-present)

Brief history of Data Science: the rise of Data Science

Probability models initially dominated data analysis but the rise of powerful computers and plentiful data has given rise to new approaches to analyzing data.

- John Tukey (1962) looks for a broadening of data analysis beyond mathematics
- Breiman (2001) describes a mathematical modeling culture and algorithmic culture
- The term "Data Science" starts being used in the 2000's to describe computational approaches to analyzing data
  - Donoho (2017). 50 Years of Data Science.



THE FUTURE OF DATA ANALYSIS[1]

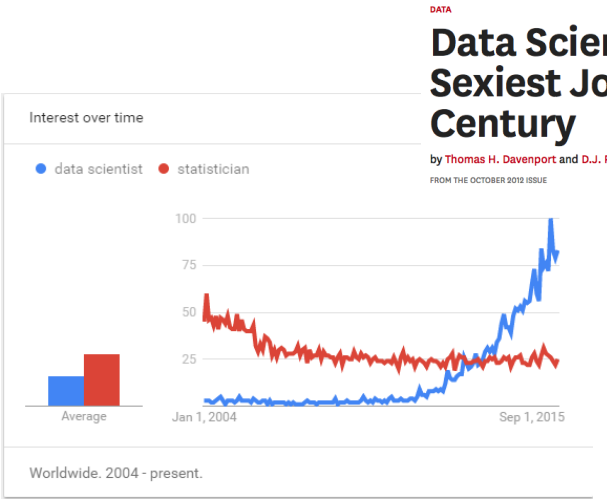
By John W. Tukey

Princeton University and Bell Telephone Laboratories

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

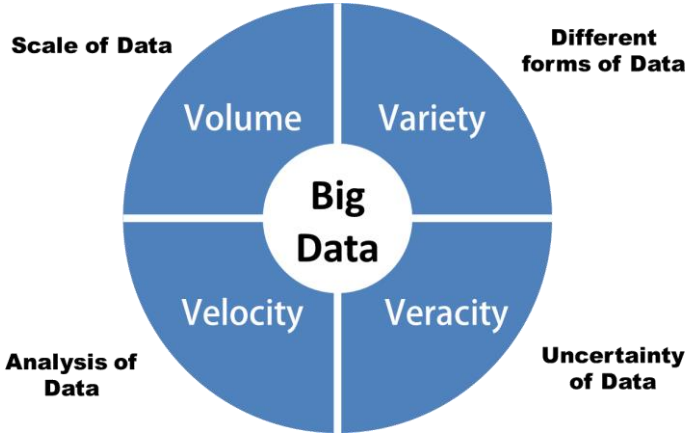
Big Data

New insights:

- Lots of new data from Internet, sensors etc., can be mined to transform our understanding in a range of fields
  - E.g., health, cosmology, social sciences, etc.

New approaches and tools:

- Hypothesis test pick up on very small (meaningless) effects with very large samples
- Data manipulation, programming, computational infrastructure are needed to extract insights

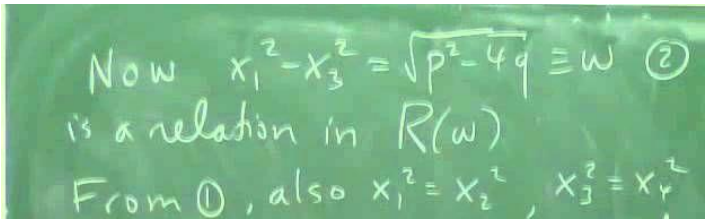


New ways to choose the best methods

Statistics focuses on mathematical models (probability distributions) to analyze data

- Best methods are the ones that have mathematical guarantees (proofs)

The proof is in the math



Data Science empirically evaluates data analysis methods

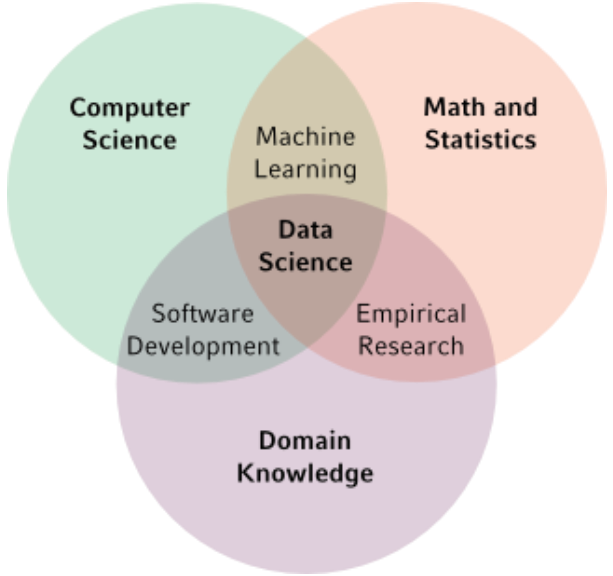
- Best methods are the one that gives the most insight *in practice*

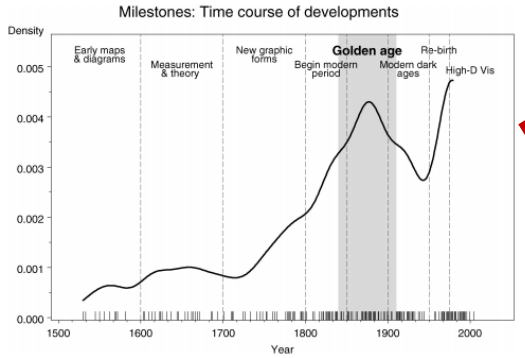
The proof is in the pudding

Data Science vs. Statistician video

So what is Data Science?

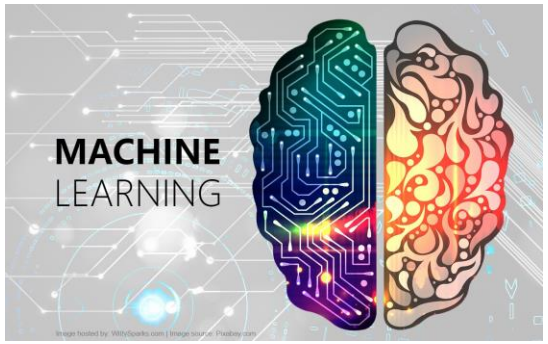
Data Science is a broadening of data analyses beyond what traditional Statistical mathematical/inferential analyses to use more computation





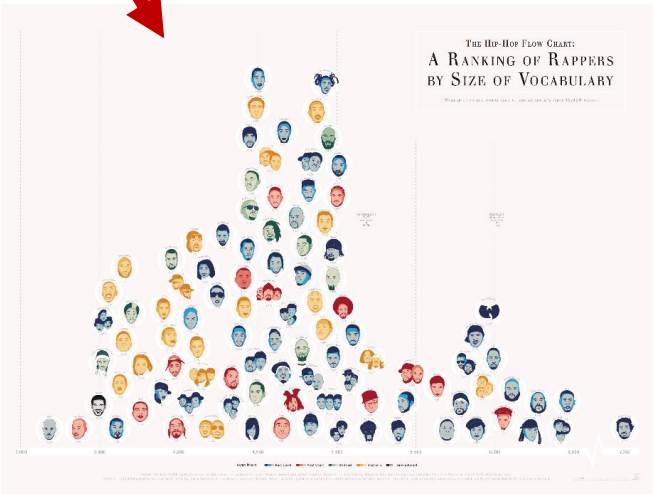
Many other fields impacted by 'Data Science

- Making business decisions
- Predictive medicine
- Fraud detection
- Etc.



Examples:

- NYC city bike visualization
- Wind map visualization



Ethical concerns around privacy, fairness and other issues


ANY
QUESTIONS?

Observational and Experimental Studies

Is it best to run in the afternoon?

The New York Times

PHYS ED

The Best Time of Day to Exercise

Men at risk for diabetes had greater blood sugar control and lost more belly fat when they exercised in the afternoon than in the morning.

370

Getty Images

Should we drink a lot of coffee?

Three coffees a day linked to a range of health benefits

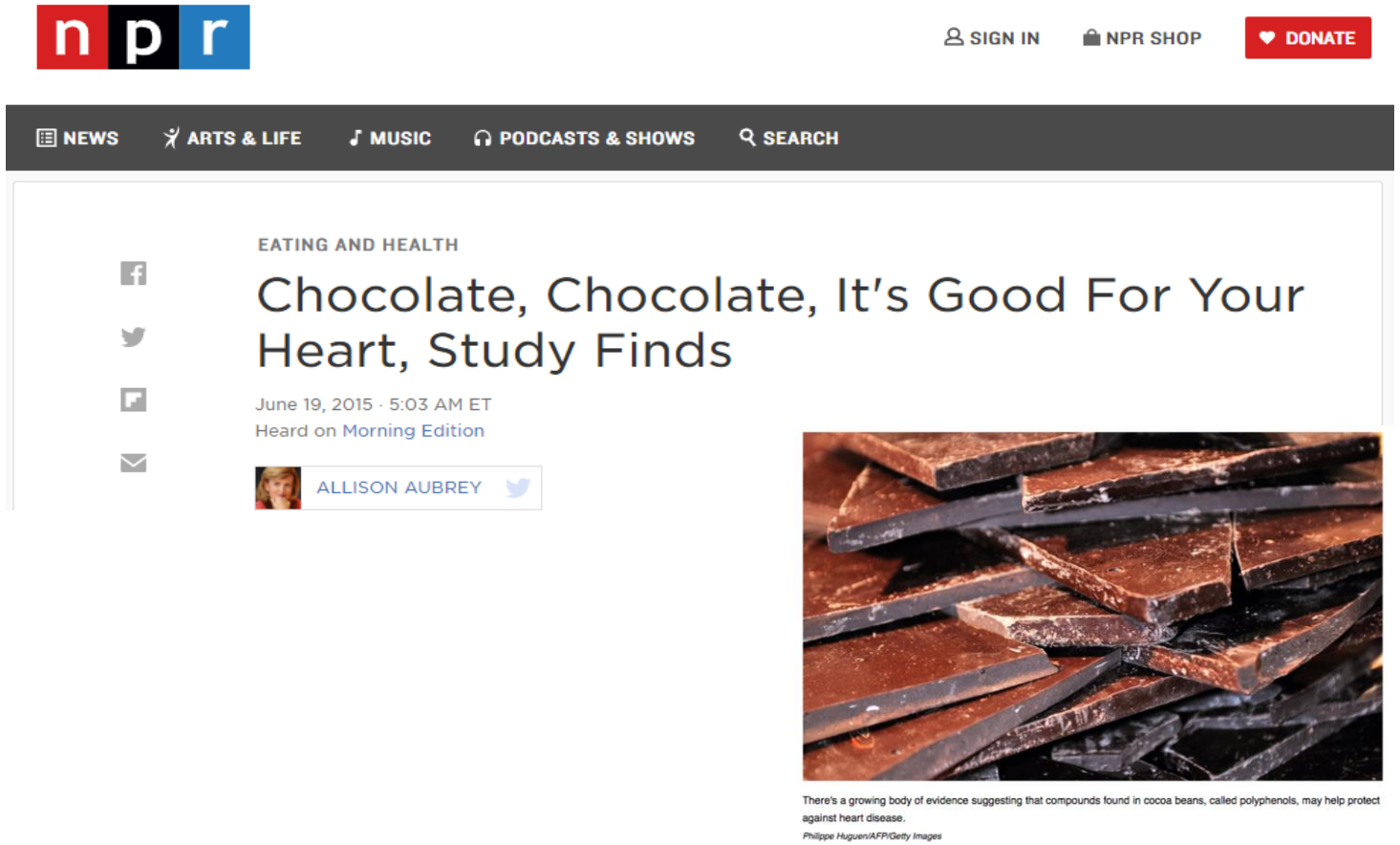
Research based on 200 previous studies worldwide says frequent drinkers less likely to get diabetes, heart disease, dementia and some cancers

Staff and agencies

Wed 22 Nov 2017
19.54 EST

The findings supported other studies showing the health benefits of drinking coffee. Photograph: Wu Hong/EPA

Is chocolate good for your heart?



Kwok, et al (2015). Heart.

Terminology

Cases

- Units that we take measurements from
  - E.g., European adults

Treatment (explanatory variable)

- A property that differs between groups
  - E.g., chocolate consumption

Outcome (response variable)

- A measurement of interest
- We want to see if the outcome differs depending on the treatment
  - E.g., heart disease differ between groups

Cases

Treatment (explanatory variable)	Outcome (response variable)
Chocolate	Cardiovascular disease
No Chocolate	Cardiovascular disease
Chocolate	No cardiovascular disease
Chocolate	No cardiovascular disease

Association

An association is the presence of a reliable relationship between the treatments an outcome

E.g., Do people who eat chocolate have lower rates of heart disease?

Some data:

- "Among those in the top tier of chocolate consumption, 12 percent developed or died of cardiovascular disease during the study, compared to 17.4 percent of those who didn't eat chocolate."

This suggests there may be an association!

Causation

A causal relationship is when changing the value of a treatment variable influences the value outcome variable

E.g., Consuming chocolate leads to a reduction in heart disease

Association does not ≠ causation!

Incorrectly implying causation is one of the most common mistakes in Statistics and Data Science

THE SCIENCE NEWS CYCLE

Start Here

Your Research

Conclusion: **A is correlated with B (ρ=0.56)**, given C, assuming D and under E conditions.

...is translated by...

UNIVERSITY PR OFFICE
(YES, YOU HAVE ONE)

FOR IMMEDIATE RELEASE: SCIENTISTS FIND POTENTIAL LINK BETWEEN A AND B (UNDER CERTAIN CONDITIONS).

...which is then picked up by...

NEWS WIRE ORGANIZATIONS

A CAUSES B, SAY SCIENTISTS.

...who are read by ...

THE INTERNETS

...then noticed by...

Scientists out to kill us again
POSTED BY RANDOM DUDE
Comments (377)
OMG! i kneeew it!!!
WTH??????

We saw it on a Blog!
A causes B all the time
What will this mean for Obama?
BREAKING NEWS BREAKING NEWS BREA

...and caught on ...

CNC Cable NEWS

LOCAL EYEWITLESS NEWS

...eventually making it to...

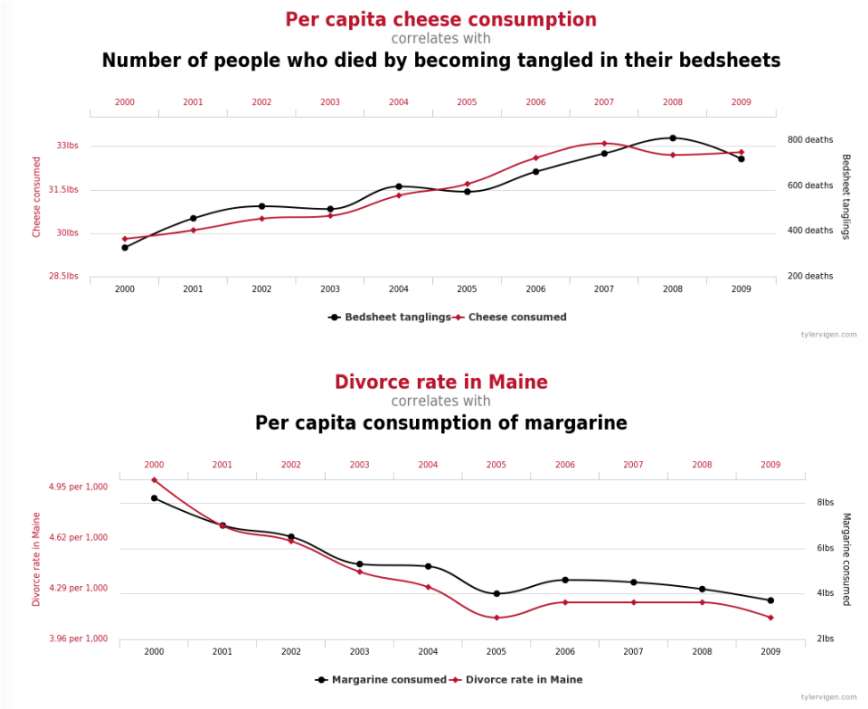
WHAT YOU DON'T KNOW ABOUT "A"... CAN KILL YOU! MORE AT 11...

A: KILLER AMONG US??
LOCAL EYEWITLESS NEWS

YOUR GRANDMA

WWW.PHDCOMICS.COM

Causal relationships?

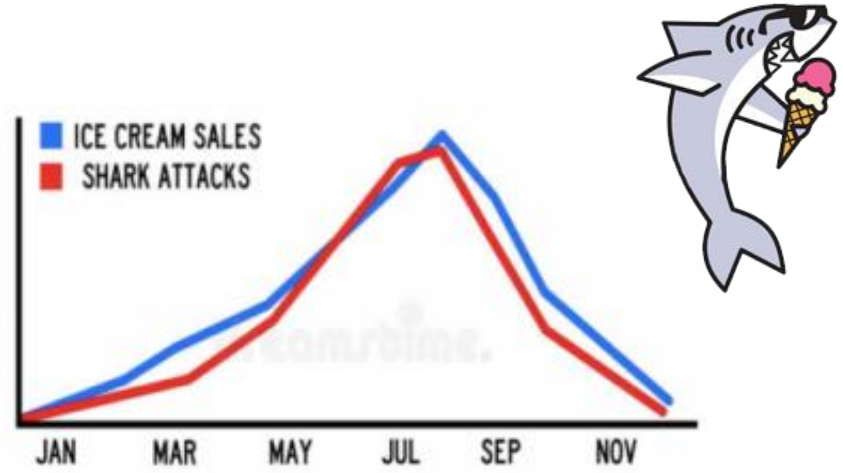


Since the 1950's both the atmospheric $C0_2$ level and the obesity levels have increased sharply. $C0_2$ causes obesity?

- Spurious correlation

As ice cream sales increase, the rate of shark attacks increases. Does eating ice cream cause shark attacks?

- Confounding

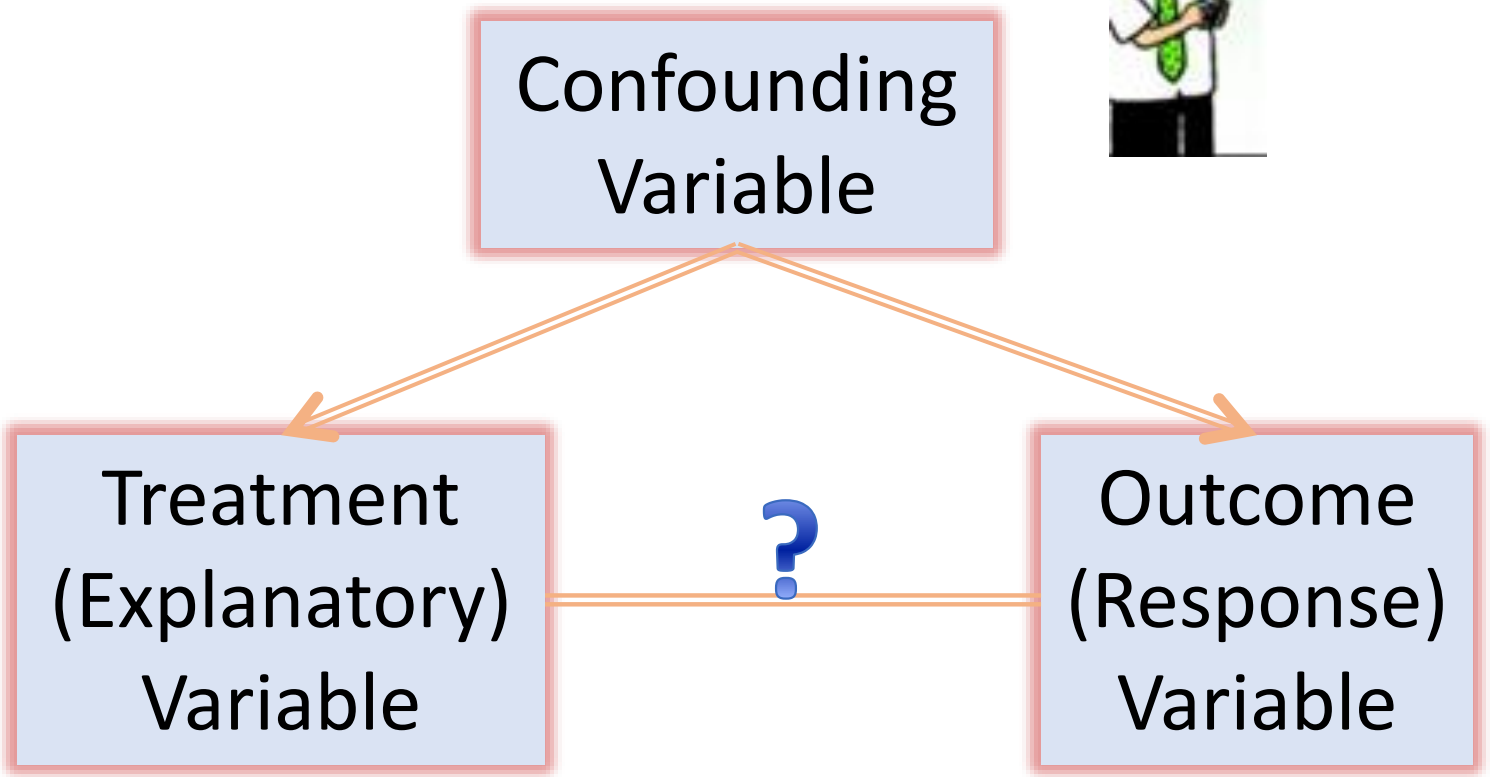


Confounding

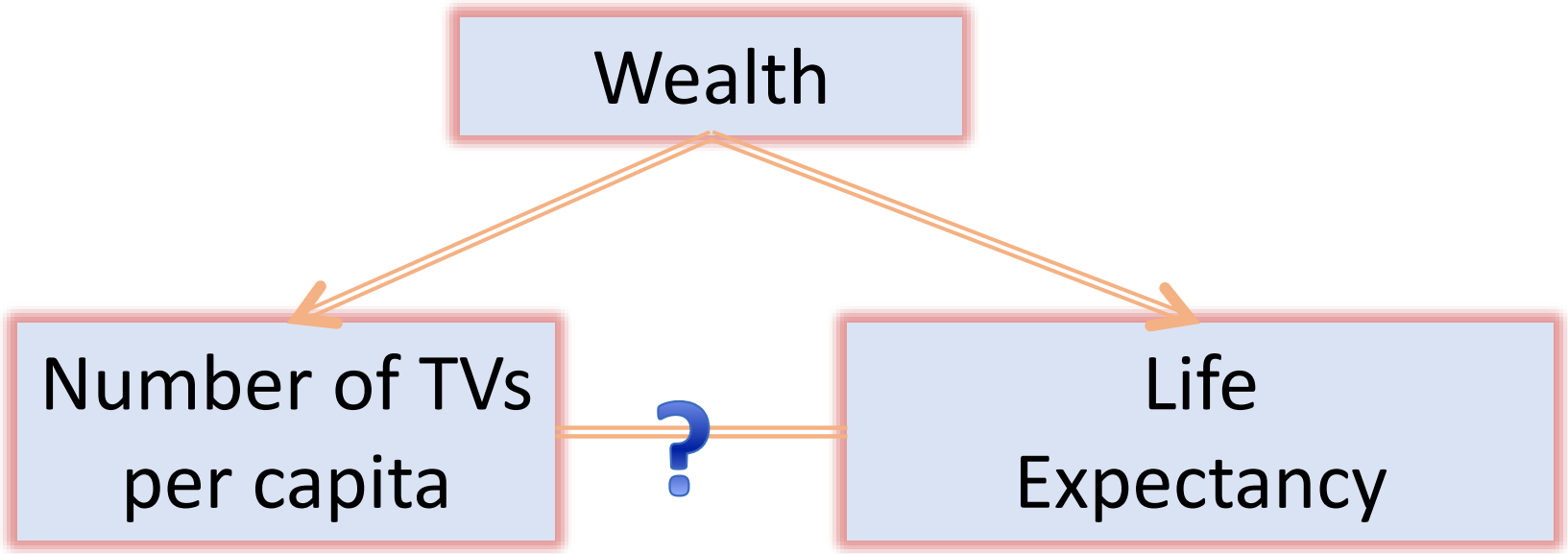
A **confounding variable** (also known as a **lurking variable**) is a third variable that is associated with both the treatment (explanatory) variable and the outcome (response) variable

A confounding variable can offer a plausible explanation for an association between the other two variables of interest

Confounding



Do TVs increase life expectancy?



Observational and experimental studies

An **observational study** is a study in which the researcher does not actively control the value of any treatment variable but simply observes the values as they naturally exist

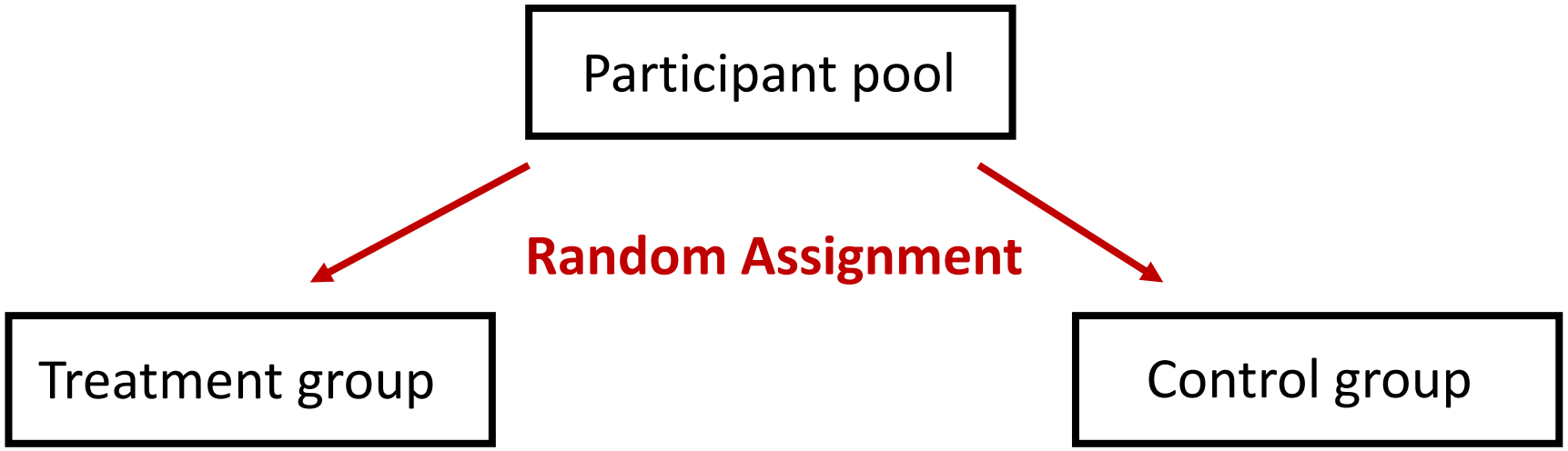
An **experiment** is a study in which the researcher actively controls one or more of the treatment variables

- Randomly assigns treatments to cases
- Allows one to get at questions of **causation**!

Randomized Controlled Experiment

Take a group of participant and ***randomly assign***:

- Half to a *treatment group* where they get chocolate
- Half in a *control group* where they get a fake chocolate (placebo)
- See if there is more improvement in the treatment group compared to the control group

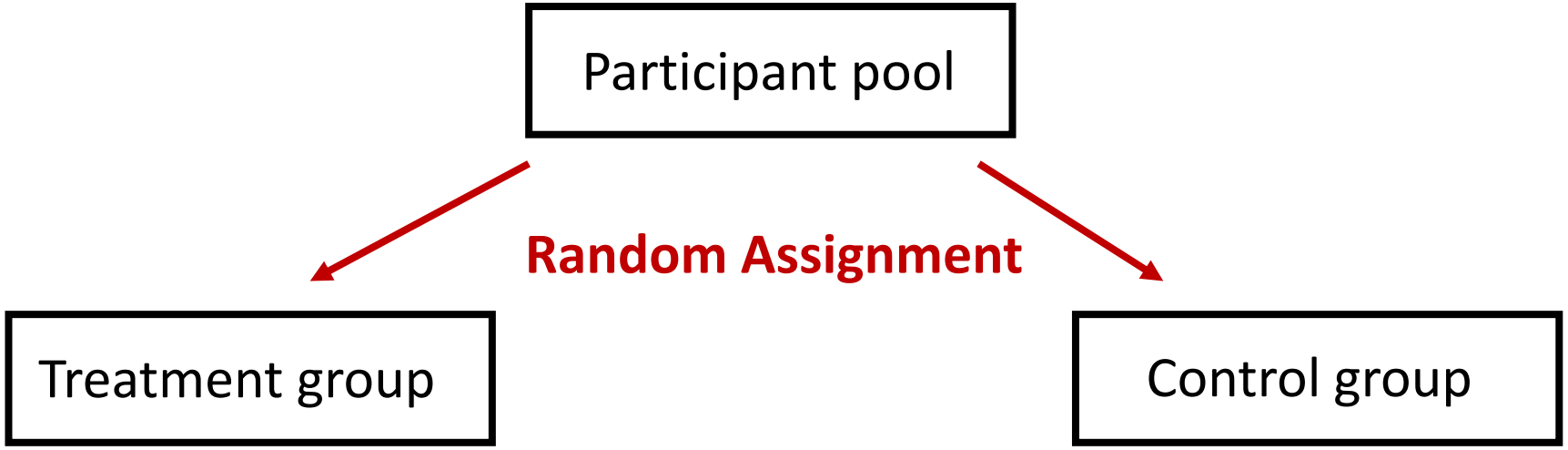


Randomized Controlled Experiment

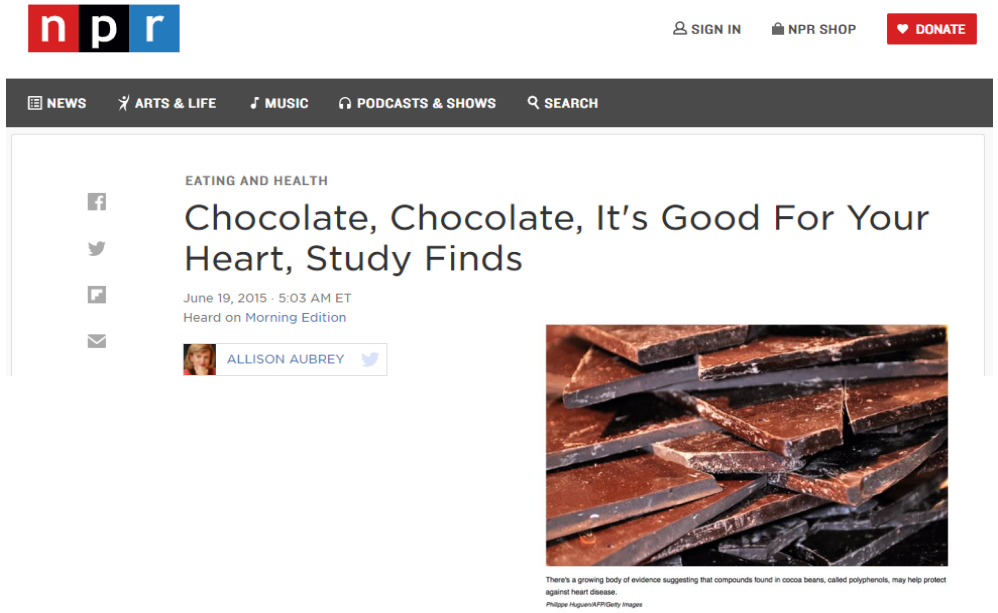
If chocolate was not effective at reducing heart disease, would we expect a difference in the amount (proportion) of heart disease between the treatment and control groups?

A: We would expect some small difference due to chance, but if the difference is very large, we can say the treatment caused the outcome.

- We will examine this more later in the semester



Central question: is chocolate good for your heart?



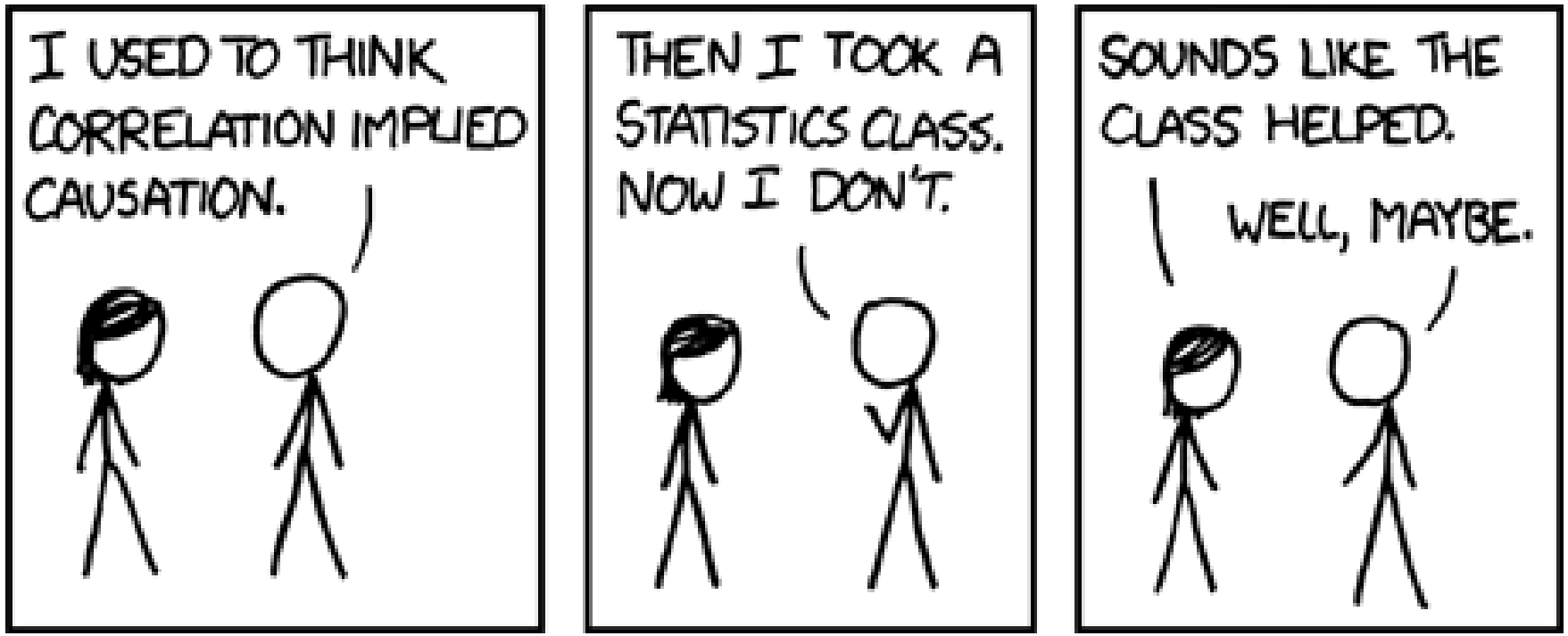
Does the cholate study show there is a causal link?

Why or why not?

"Now, the rub with this kind of study is that the link between chocolate and health is just an association. "It doesn't prove a cause-and-effect relationship between chocolate and reduced risk of heart disease and stroke," says JoAnn Manson, chief of the Division of Preventive Medicine at Brigham and Women's Hospital in Boston."


BUSTED!
DON'T BELIEVE THESE
CHOCOLATE
MYTHS!

Questions?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.


ANY
QUESTIONS
?

Next class: Cholera and Python

Things to do:

- Try to install Anaconda/Python on your computer and/or try out Google colabs
- Try out homework 0
  - Does not need to be turned in