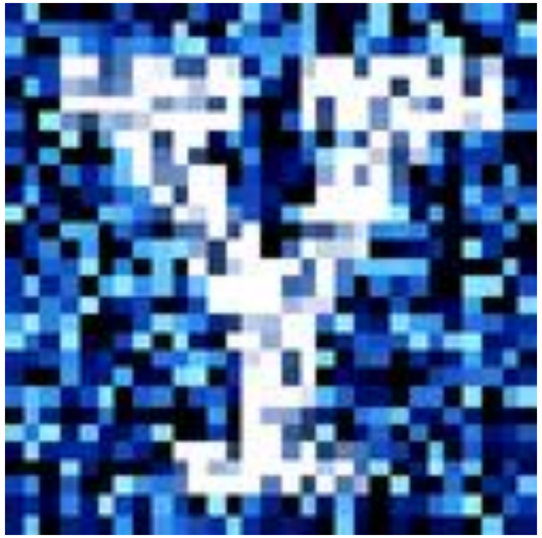
Ydata: Introduction to Data Science



Lecture 03: Intro to Python

Overview

Review and continuation on association vs. causation

- Examples of John Snow and cholera

Start on the basics of Python

- Basic expressions
- Assigning values to names
- Calling function
- If there is time: Operations on Tables

Announcements

Homework 1 has been posted, It is due on Sunday February 6th at 11pm

Practice exercises have also been posted

- These are not turned in but will be useful to complete to gain more Python practice

Any questions about anything?

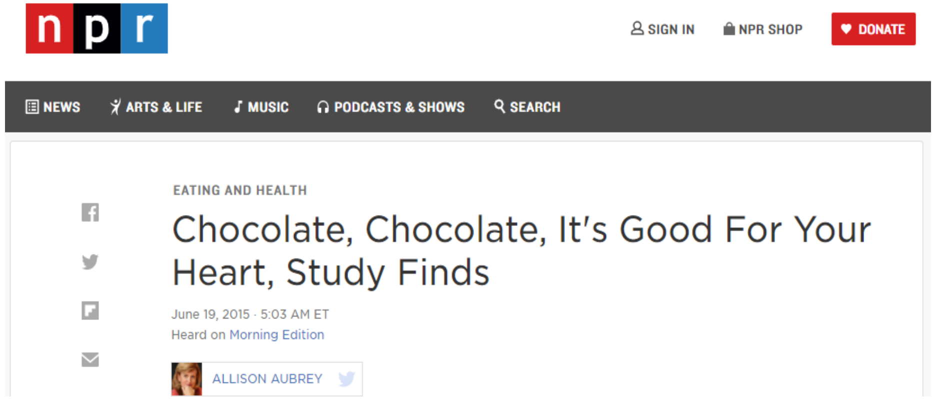
Review association vs. causation

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

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

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



EATING AND HEALTH

Chocolate, Chocolate, It's Good For Your Heart, Study Finds

June 19, 2015 · 5:03 AM ET

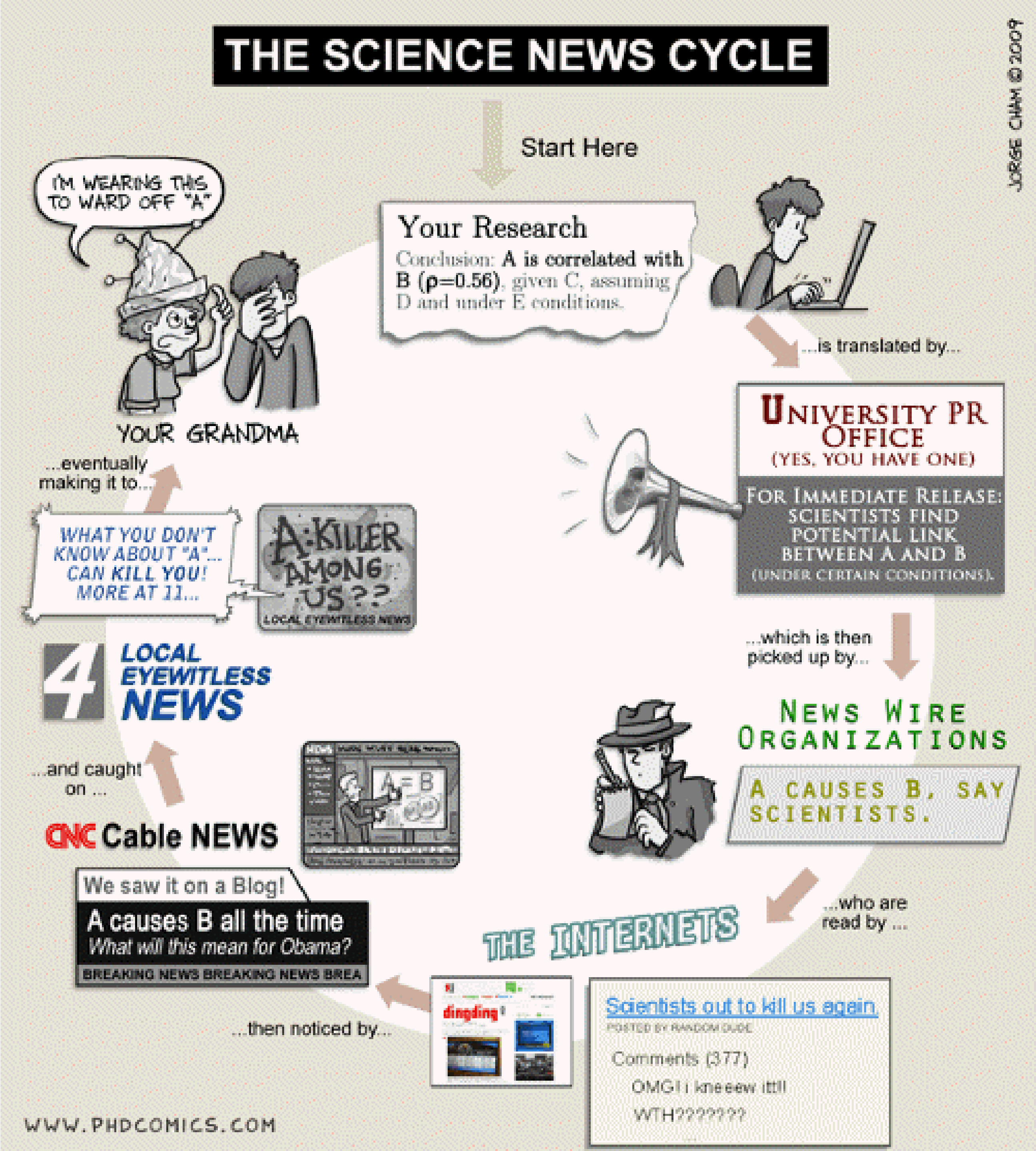
Heard on Morning Edition

ALLISON AUBREY

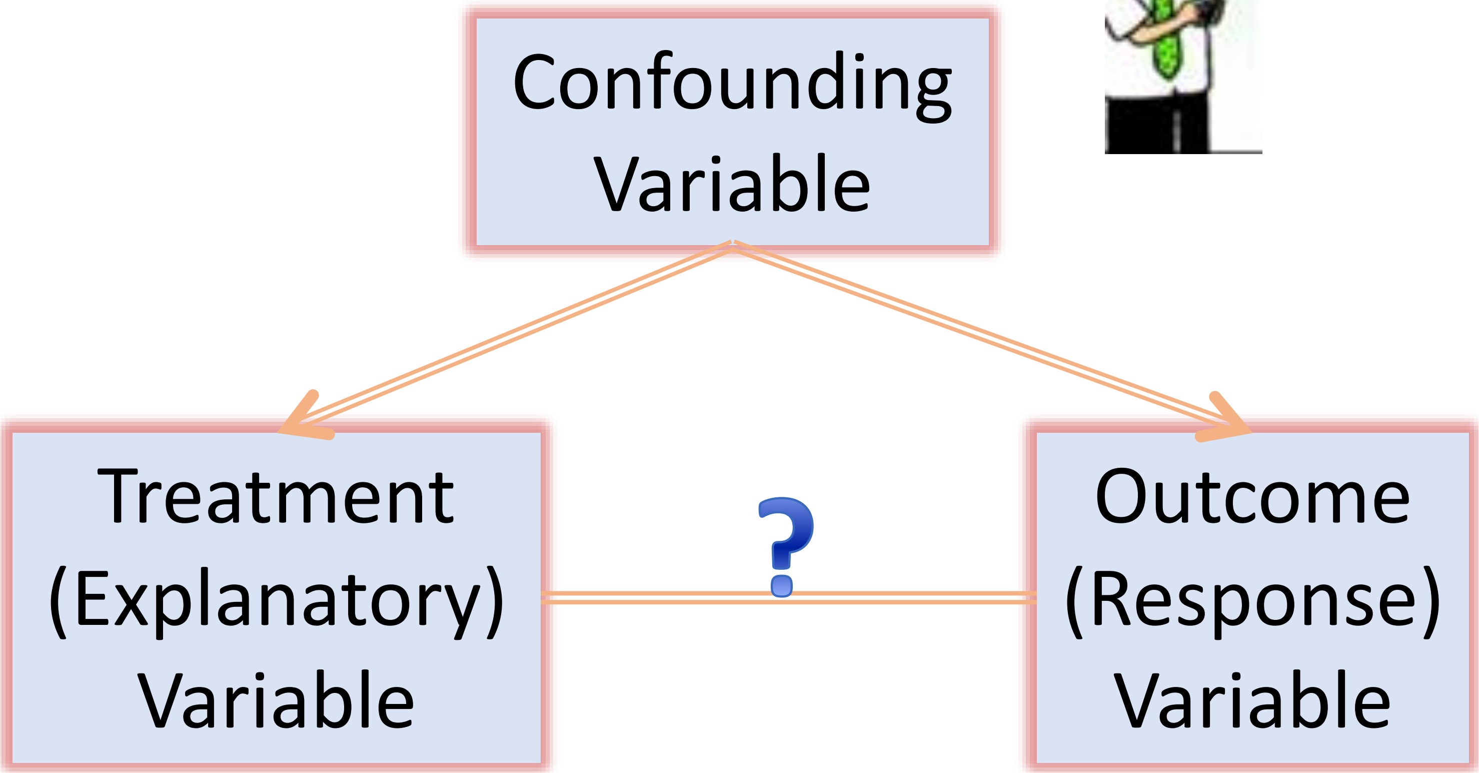
There's a growing body of evidence suggesting that compounds found in cocoa beans, called polyphenols, may help protect against heart disease.

Philippe Huguen/AFP/Getty Images

Association does not ≠ causation!


THE SCIENCE NEWS CYCLE
JORGE CHAM ©2009
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
Scientists out to kill us again
POSTED BY RANDOM DUDE
Comments (377)
OMG! i kneeew it!!
WTH???????
dingding
...then noticed by...
We saw it on a Blog!
A causes B all the time
What will this mean for Obama?
BREAKING NEWS BREAKING NEWS BREA
CNC Cable NEWS
A = B
...and caught on ...
4 LOCAL EYEWITLESS NEWS
WHAT YOU DON'T KNOW ABOUT "A"... CAN KILL YOU! MORE AT 11...
A: KILLER AMONG US??
LOCAL EYEWITLESS NEWS
...eventually making it to...
YOUR GRANDMA
I'M WEARING THIS TO WARD OFF "A"
WWW.PHDCOMICS.COM

Confounding

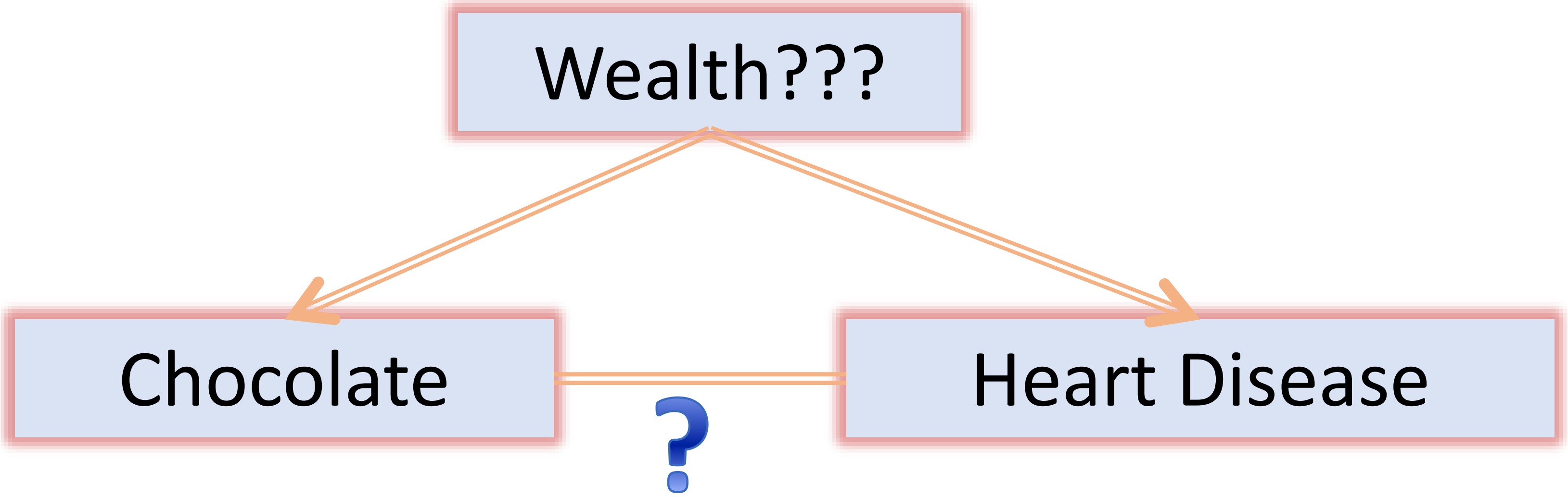

Confounding Variable

Treatment (Explanatory) Variable

?

Outcome (Response) Variable

Does chocolate reduce heart disease?



Wealth???

Chocolate

Heart Disease

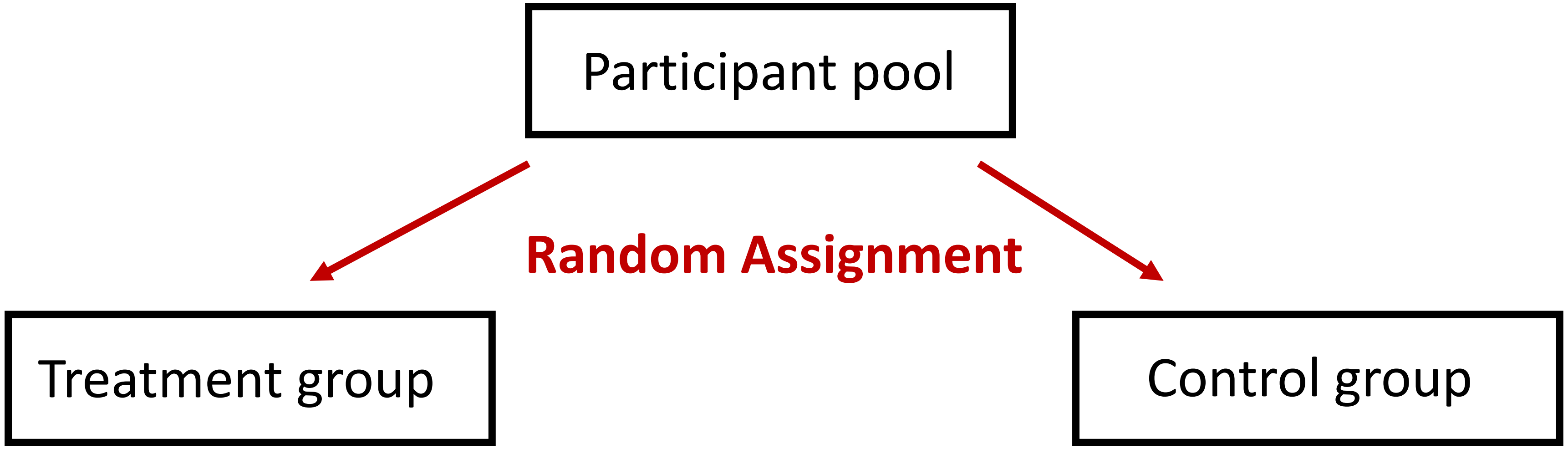
?

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**!



Participant pool

Random Assignment

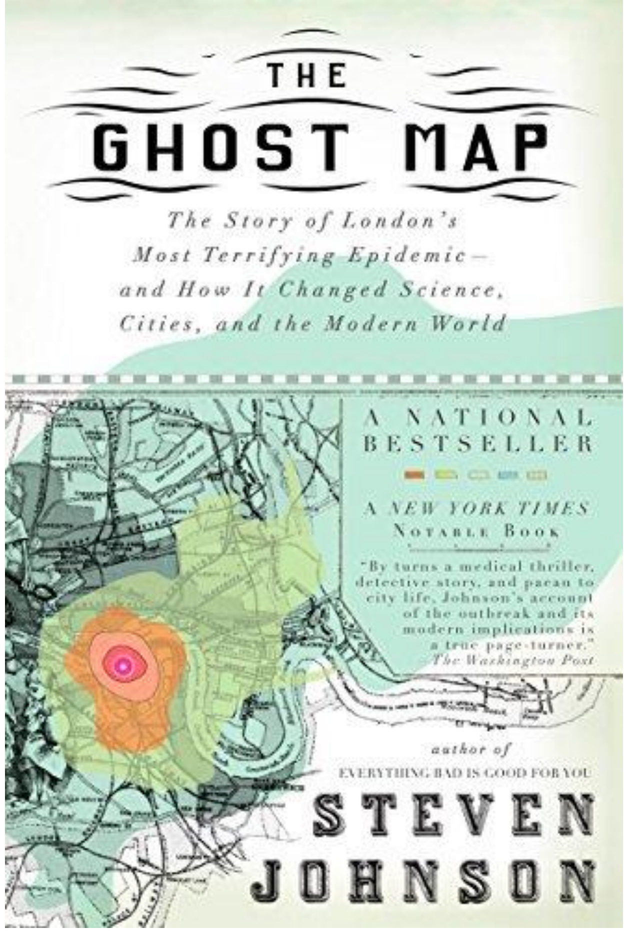
Treatment group

Control group

Determining the causes of cholera

London Pub by [illegible]

THE APPEARANCE AFTER DEATH OF A VICTIM TO THE INDIAN CHOLERA



THE GHOST MAP

The Story of London's Most Terrifying Epidemic — and How It Changed Science, Cities, and the Modern World

A NATIONAL BESTSELLER

A *NEW YORK TIMES* NOTABLE BOOK

"By turns a medical thriller, detective story, and paean to city life, Johnson's account of the outbreak and its modern implications is a true page-turner." *The Washington Post*

author of
EVERYTHING BAD IS GOOD FOR YOU

STEVEN JOHNSON

Cholera in London in the 19th century

Cholera reached London in early 1830s

It was greatly feared as it was often deadly

- An outbreak in 1849 killed over 14,000 people in London

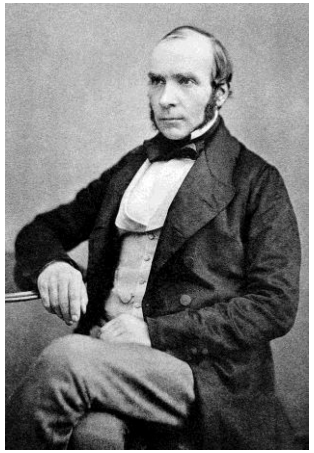
Cause of cholera was unknown. Several theories:

1. Miasmas theory: caused by bad air/smells

- Florence Nightingale, Edwin Chadwick (board of health)

2. Water born disease

- John Snow (anesthesiologist)

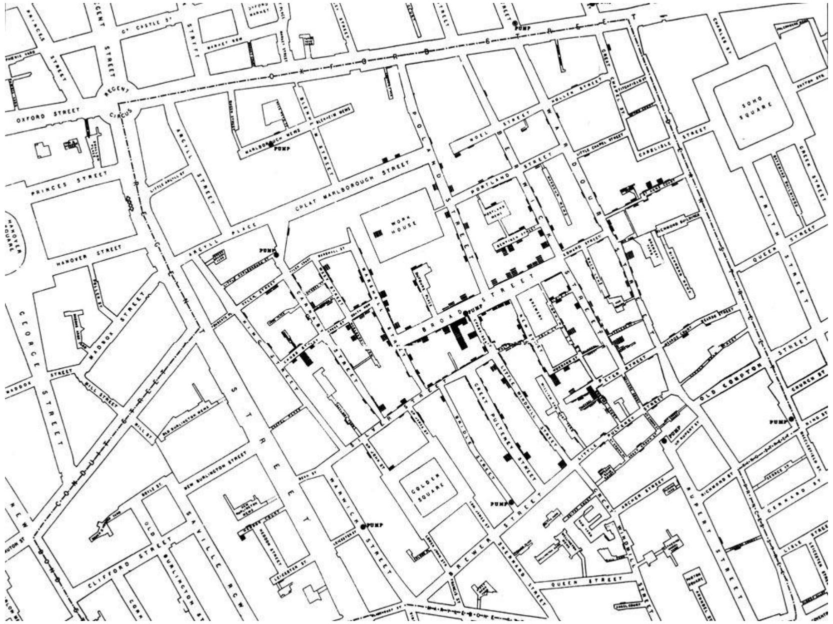


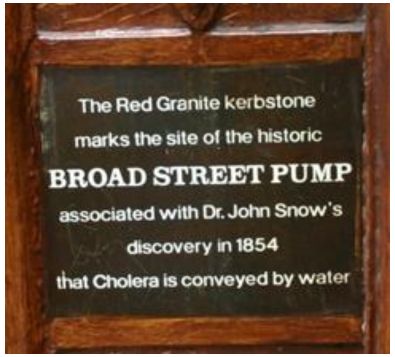
John Snow and spatial mapping

To try to understand the cause of the cholera outbreak of 1854, John Snow plotted a map of cholera deaths

Based on this map and interviews, he concluded that the source of cholera was the Broad Steet well

- He famously removed the handle of the well to prevent the spread of disease
- Now he is considered the founder of epidemiology





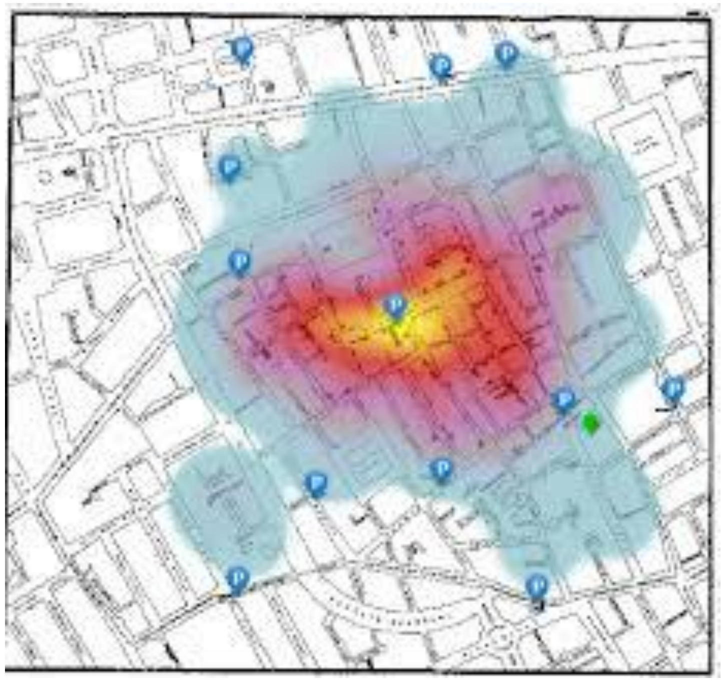
The Red Granite kerbstone
marks the site of the historic
BROAD STREET PUMP
associated with Dr. John Snow's
discovery in 1854
that Cholera is conveyed by water

John Snow and causation

Q_1: Did John Snow show there was a causal link between drinking water from the Broad Street well and cholera?

Q_2: What is an indicator that no causal link was shown?

Q_3: Is observational data useless of finding causal links?

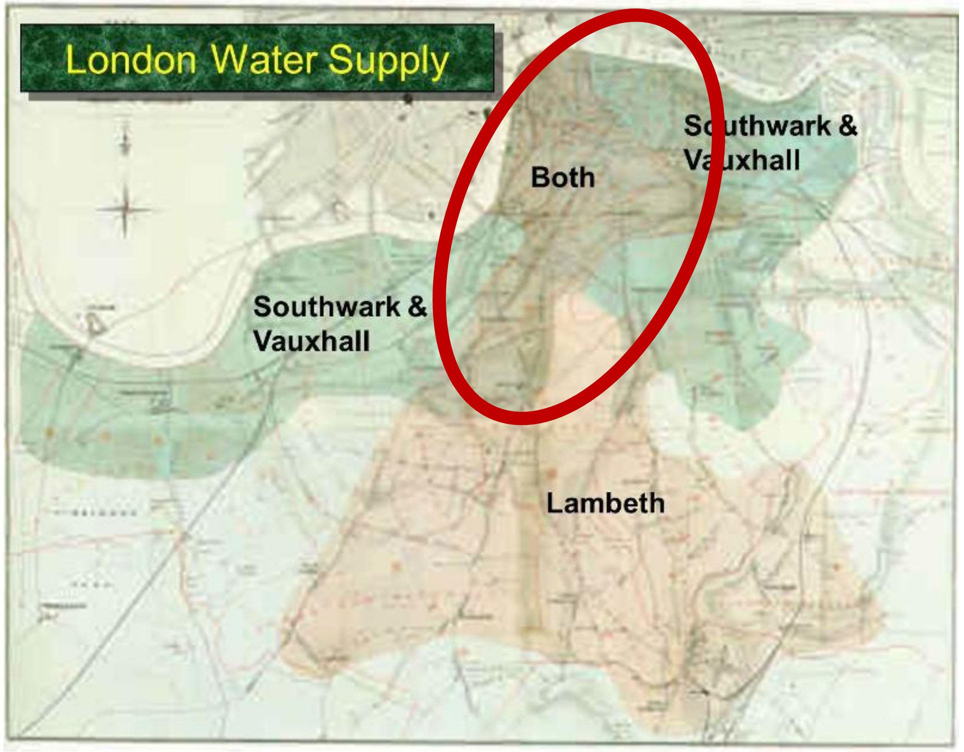


BREWERY
You know everything, John Snow by Kenneth Field

Water supply to London

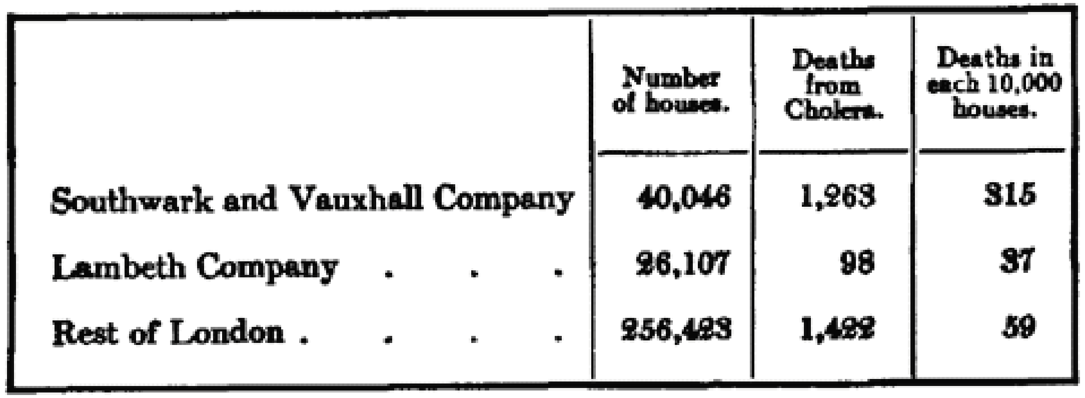
Two companies supplied water to London in the mid 19th century

- Lambeth drew water **upriver** from sewage dump into the River Thames
- Southark & Vauxhall drew from **below** the sewage dump



Snow focused on areas that were served by both companies to see if there were different rates of cholera

- Showed very different death rates between these two companies



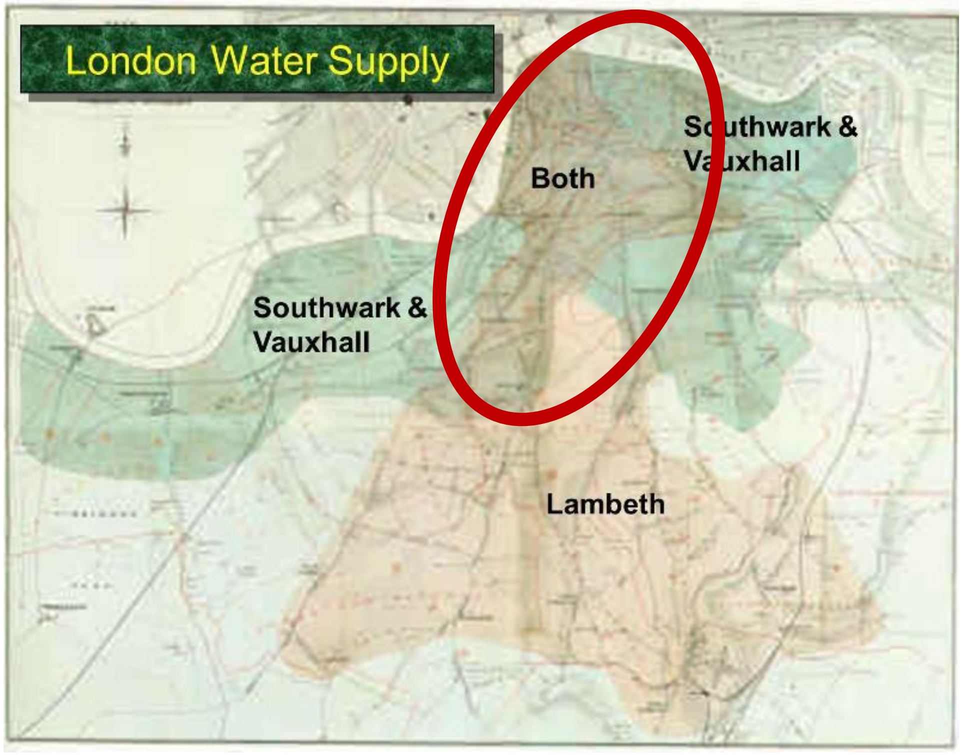
	Number of houses.	Deaths from Cholera.	Deaths in each 10,000 houses.
Southwark and Vauxhall Company	40,046	1,263	315
Lambeth Company . . .	26,107	98	37
Rest of London	256,423	1,422	59

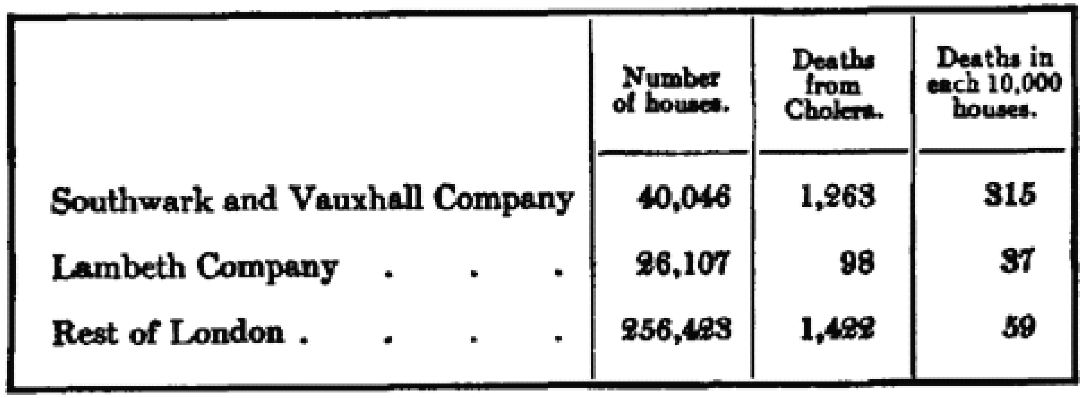
Water supply to London

Q$_1$: Does this analysis provide evidence there is a causal link?

A "natural experiment" results in data that are a lot like randomized controls

- E.g., tries to make sure there are no systematic differences between groups apart from the treatment





	Number of houses.	Deaths from Cholera.	Deaths in each 10,000 houses.
Southwark and Vauxhall Company	40,046	1,263	315
Lambeth Company	26,107	98	37
Rest of London	256,423	1,422	59

Take away

Most data examined by Data Scientists is observational data

Observational data can give real insights

- E.g., did provide evidence that cholera is a water born disease

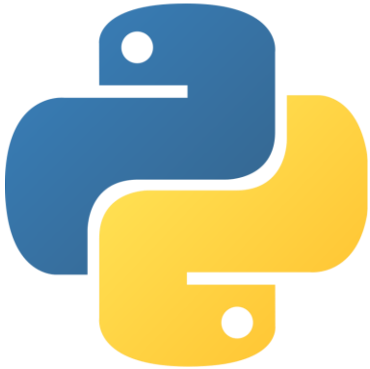
However, we need to be aware of limitations of observational data

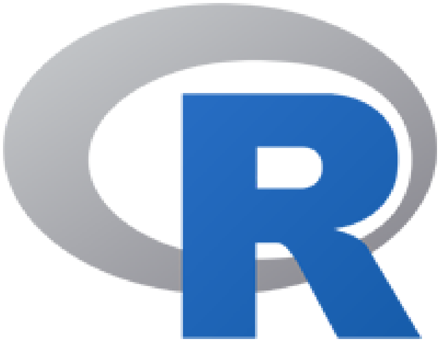
- If possible, we should run a randomized controlled trial to definitively show causal links

ANY
QUESTIONS?

Programming languages for Data Science

The two most popular languages for Data Science are:



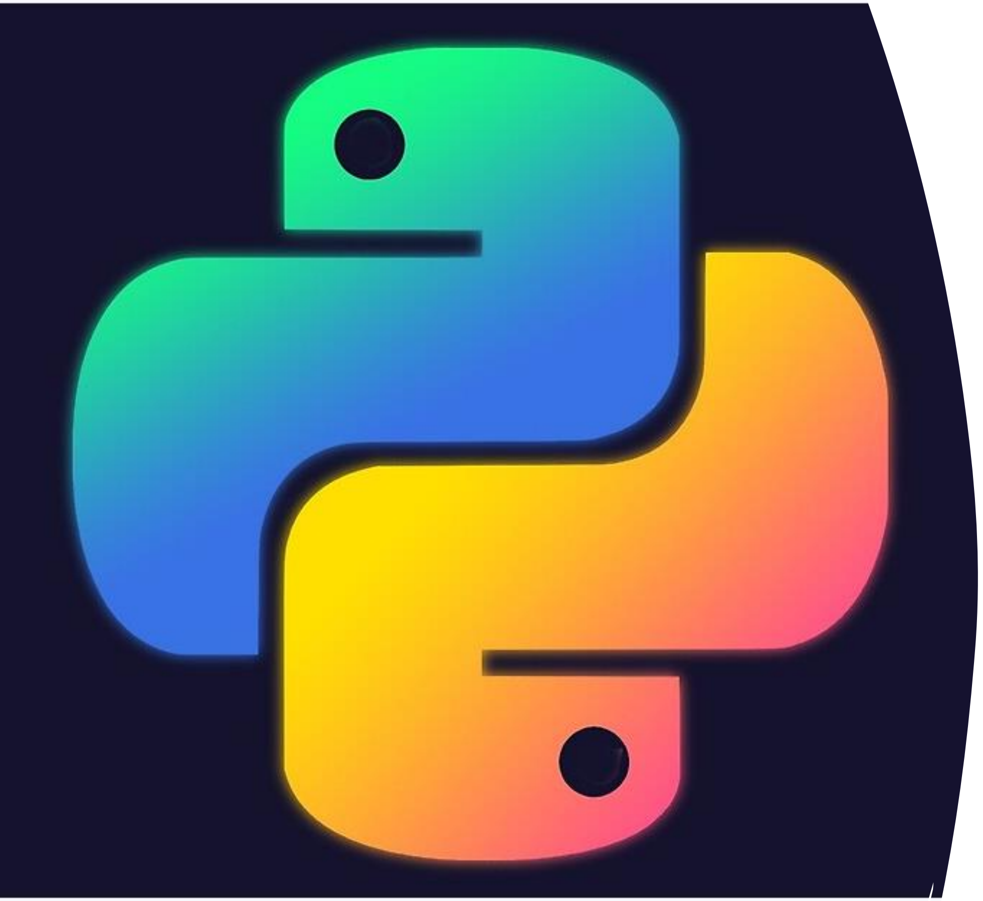


General purpose programming language

- Can do a lot more than data analysis
- Easy to read
- Easy to write larger software packages
- Good machine learning package (scikitlearn)

Focused on data analysis

- Better for creating pdf reports
- Easy to create interactive apps
- RStudio created a great IDE and support



Programming in Python

Understanding the language fundamentals is important

Learn through practice, not by reading or watching but by doing

- Like learning to ride a bike

Follow along with:

- demo/lec03
- binder

Expressions

Expressions

Expressions describe how a computer should combine pieces of data

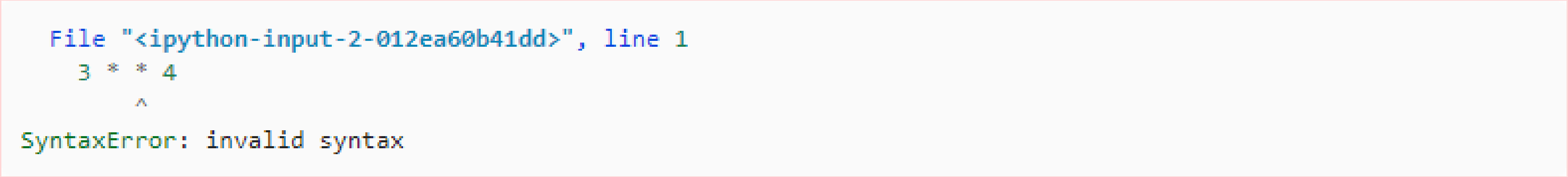
- They are evaluated to by the computer and return a value
- E.g., mathematical expressions
  - Multiplication: 3 * 4
  - Exponentiation: 3**4

Operation	Operation	Example	**Value**
Addition	+	2 + 3	5
Subtraction	-	2 – 3	-1
Multiplication	*	2*3	6
Division	/	7/3	2.667
Remainder	%	7 % 3	1
Exponentiation	**	2**.05	1.414

Syntax

The *Syntax* of a language is its set of grammar rules for how expressions can be written

- *SyntaxError* indicates that an expression structure doesn’t match any of the rules of the language.
- E.g., failed attempt at exponentiation: 3 * * 4

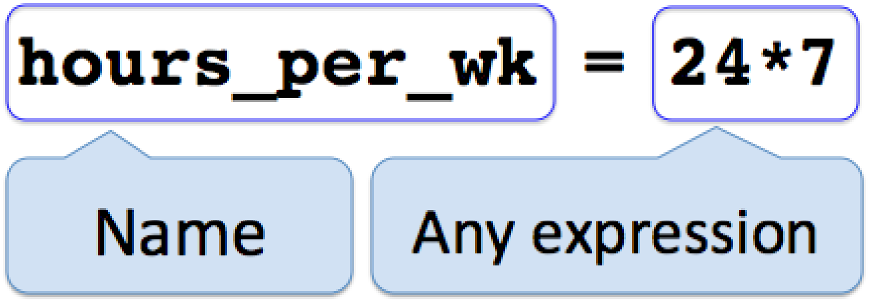


```
  File "<ipython-input-2-012ea60b41dd>", line 1
    3 * * 4
        ^
SyntaxError: invalid syntax
```

Let's explore this in Jupyter!

Names

Assignment statements



```
hours_per_wk = 24*7
```

Names store the values (from an expression)

- i.e., they are like variables in algebra

Names are assigned values using the = symbol

- E.g., my_number = 7

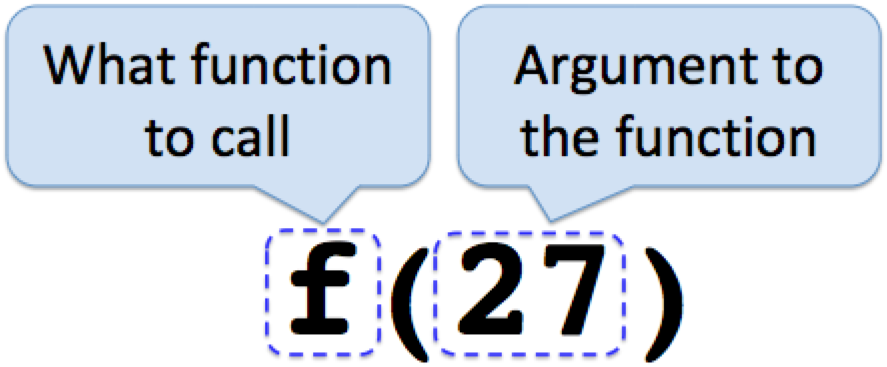
Let's explore this in Jupyter!

Call Expressions

Anatomy of a Call Expression

Call expressions are expressions that call functions

- Functions take in one or more values (arguments) and (usually) return another value

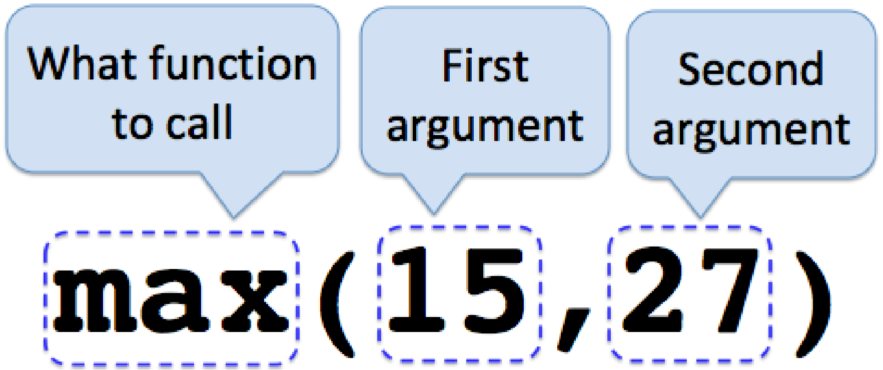


What function to call

Argument to the function

```
f(27)
```

Example: taking the maximum value



What function to call

First argument

Second argument

```
max(15,27)
```

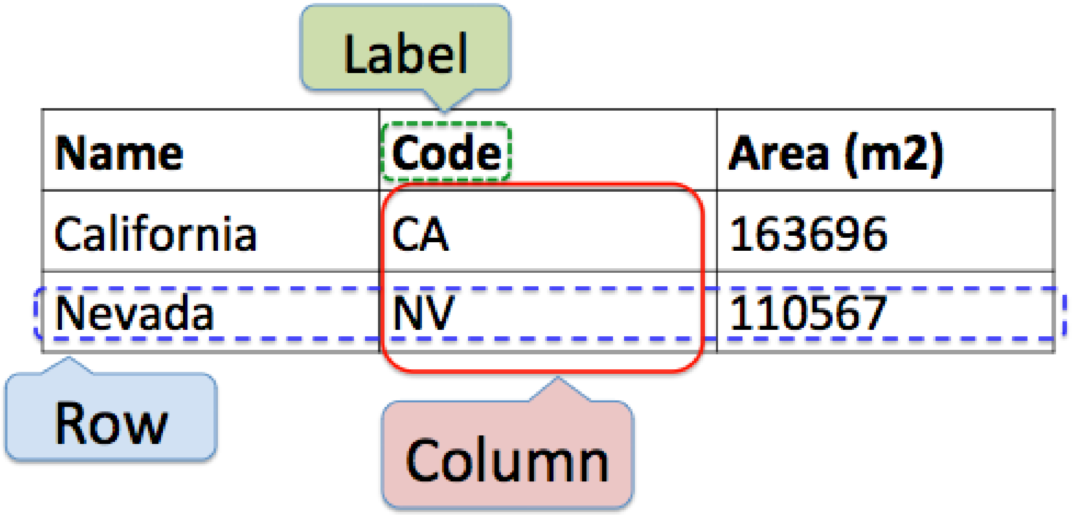
Let's explore this in Jupyter!

Tables

Table structure

A Table is a sequence of labeled columns

- Each row represents one individual case
- Data within a column represents one attribute



Name	Code	Area (m2)
California	CA	163696
Nevada	NV	110567

Some Table Operations

t.select(label) - constructs a new table with just the specified columns

t.drop(label) - constructs a new table in which the specified columns are omitted

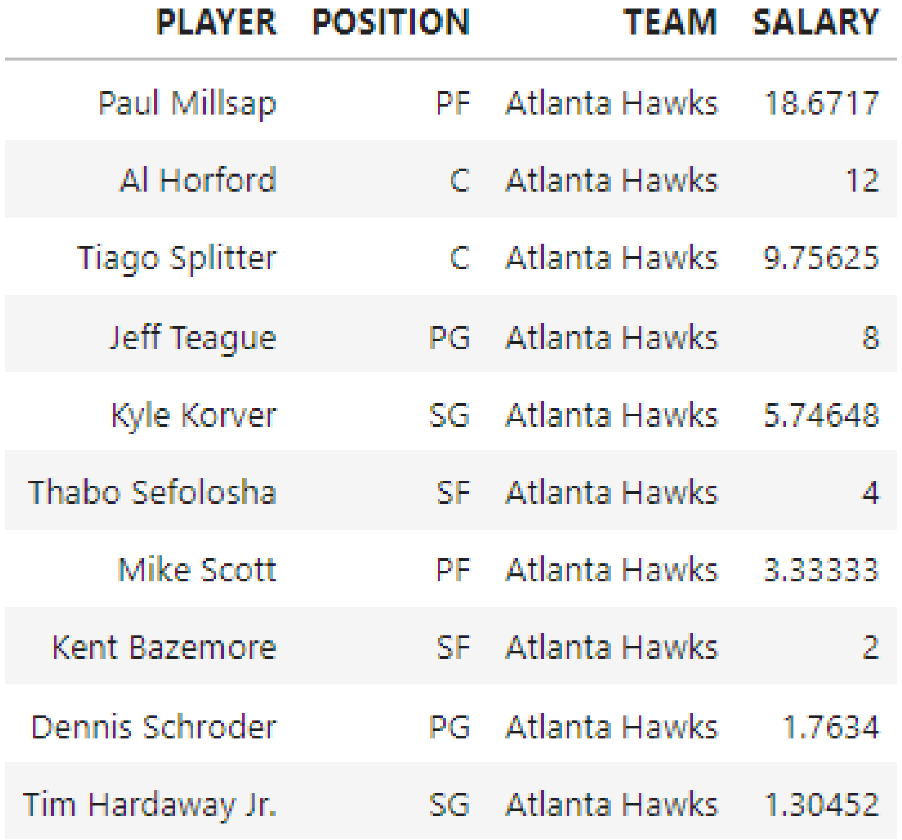
t.sort(label) - constructs a new table with rows sorted by the specied column

t.where(label, condition) - constructs a new table with just the rows that match the condition

Discussion question

How to display just the row corresponding to the player who had the highest salary?

nba table



PLAYER	POSITION	TEAM	SALARY
Paul Millsap	PF	Atlanta Hawks	18.6717
Al Horford	C	Atlanta Hawks	12
Tiago Splitter	C	Atlanta Hawks	9.75625
Jeff Teague	PG	Atlanta Hawks	8
Kyle Korver	SG	Atlanta Hawks	5.74648
Thabo Sefolosha	SF	Atlanta Hawks	4
Mike Scott	PF	Atlanta Hawks	3.33333
Kent Bazemore	SF	Atlanta Hawks	2
Dennis Schroder	PG	Atlanta Hawks	1.7634
Tim Hardaway Jr.	SG	Atlanta Hawks	1.30452

Pandas

FYI: The datascience package is a Berkeley product

It's a light wrapper on top of pandas

Hopefully at the end of the class we'll have time to discuss Pandas

Summary

Today we talked about how to:

- Wrap of up association/causation
- Assign a value to a name
- Call a function
- Operate on Tables