

# Lecture 1

Introduction

Cause and Effect

# Welcome to DSC 10

- A crash course in data science.
- A course developed by UC Berkeley faculty and students and adapted by UCSD.

# Course Structure

# Programming experience

**Do you have any programming experience?**

- A. Yes, I'm a pro (Java, Python etc). Or at least I think I am :)
- B. I have some experience
- C. I know a few basic concepts
- D. No experience whatsoever! Yay!
- E. Why do you ask? Is it a programming class?

# Data Science

# What is Data Science?

Drawing useful conclusions from data in a principled way.

- **Exploration**

- Identifying patterns in information
- Uses visualizations

- **Prediction**

- Making informed guesses
- Uses machine learning and optimization

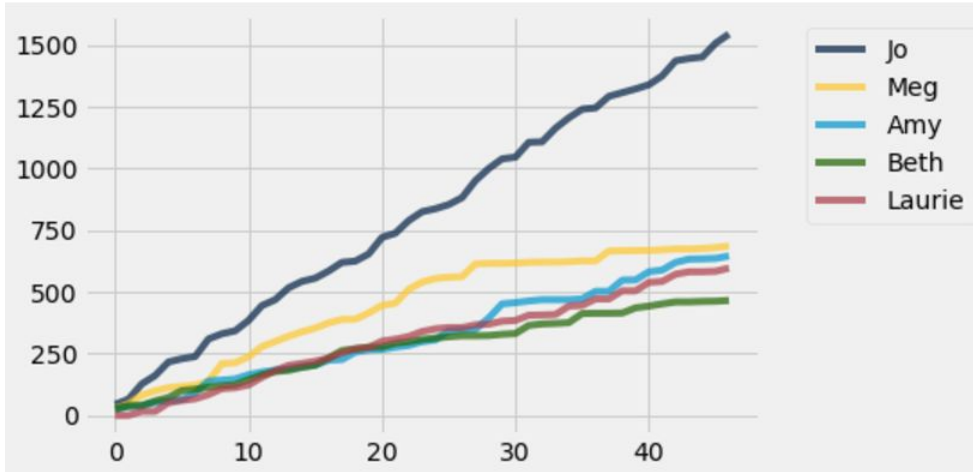
- **Inference**

- Quantifying whether those patterns are reliable
- Uses randomization

# Literature

(Demo)

# Literature

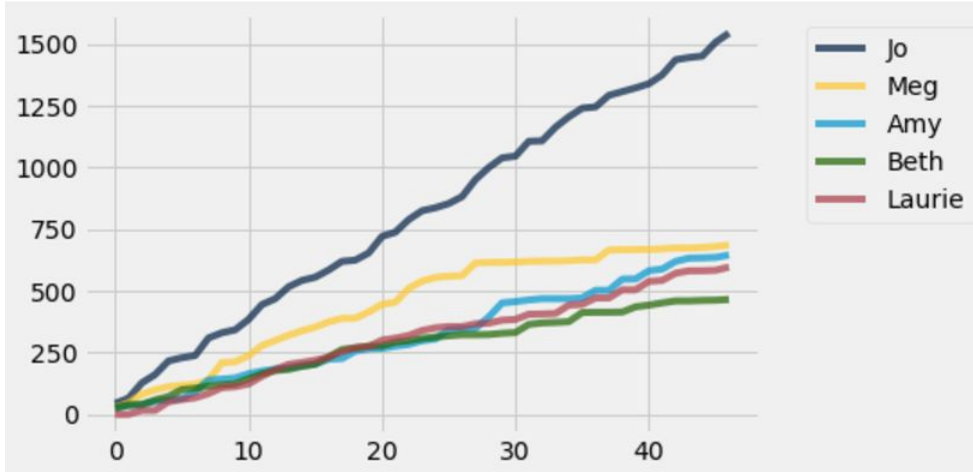


In chapter 27, Jo moves to New York alone. Her relationship with which sister suffers the most from this faraway move?

- A. Amy
- B. Beth
- C. Meg



# Literature



Laurie is a man who marries one of the sisters at the end. Which one?

- A. Amy
- B. Beth
- C. Jo
- D. Meg

[www.dsc10.com](http://www.dsc10.com)

# First week assignments

- Lab 1: out today
- Deadline: Sunday, January 12 @ 11:59 p.m.
  
- Homework 1: out today
- Deadline: next Wednesday, January 15 @ 11:59 p.m.
  
- **Tonight: make sure you can access lab/HW by clicking on link in the schedule**
  
- Start early!

# **Lecture 01 :**

# **Association and Causality**

# Really?



[npr.org](http://npr.org) (report on a study in [heart.bmj.com](http://heart.bmj.com))

# Definitions

- **individuals**, study subjects, participants, units
  - *European adults*
- **treatment**
  - *chocolate consumption*
- **outcome**
  - *heart disease*

# The first question

Is there **any relation** between chocolate consumption and heart disease?

- **Association: any relation**
- **Not necessarily causal! (shark bites and ice cream)**

# 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.”

-Howard LeWine of Harvard Health Blog, reported by [npr.org](https://www.npr.org)

Is there an association (any relation) between chocolate consumption and heart disease?

- A. Yes, I think so
- B. No, I don't think so
- C. Maybe, I can't tell





# London in the 1800s



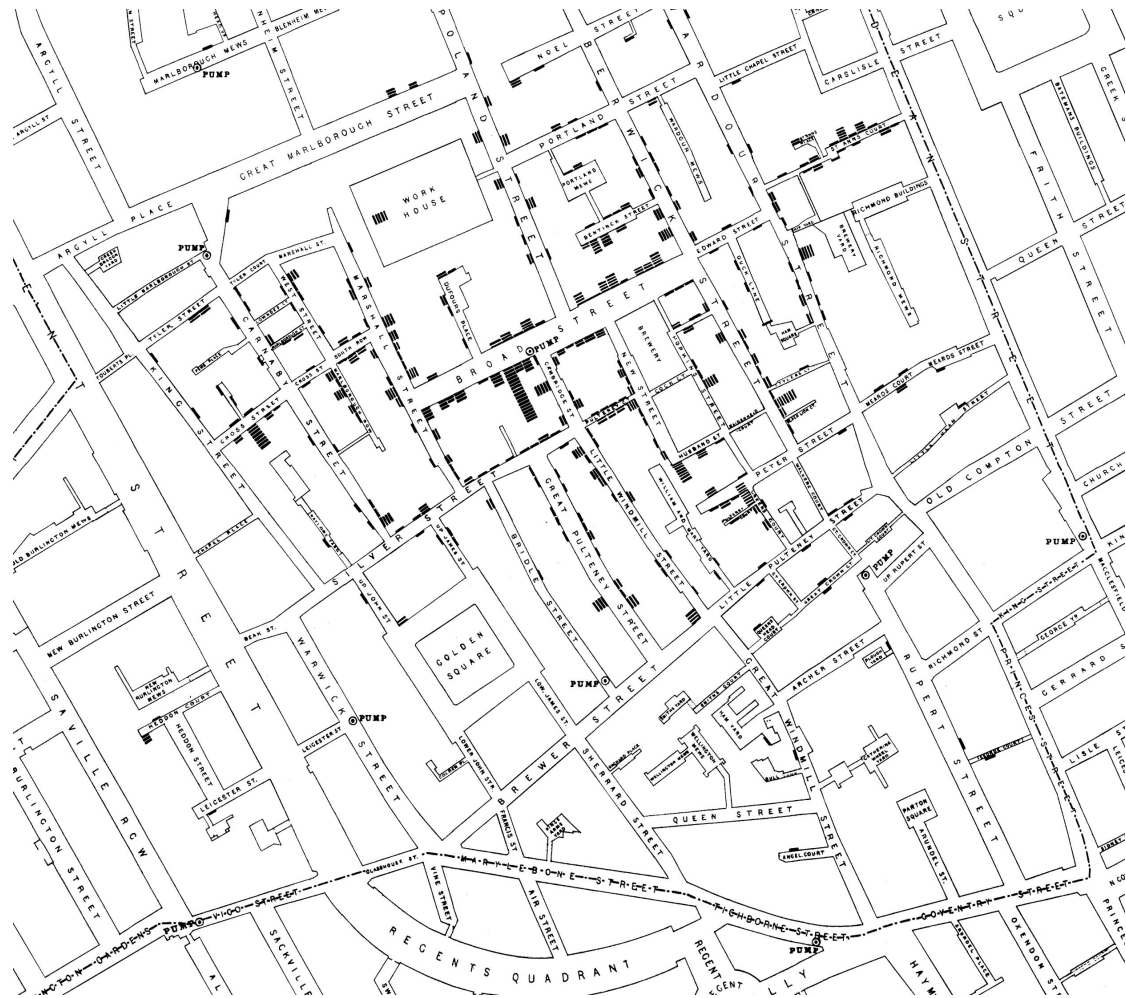
FUTURESKETCHES OF LONDON.—WHITSCAPPEL.

# Miasmas, miasmatism, miasmatists

- **Bad smells** given off by waste and rotting matter
- **Believed to be the main source of disease**
- Suggested remedies:
  - “fly to clene air”
  - “a pocket full o’posies”
  - “fire off barrels of gunpowder”
- Staunch believers:
  - Florence Nightingale
  - Edwin Chadwick, Commissioner of the General Board of Health

# John Snow, 1813-1858

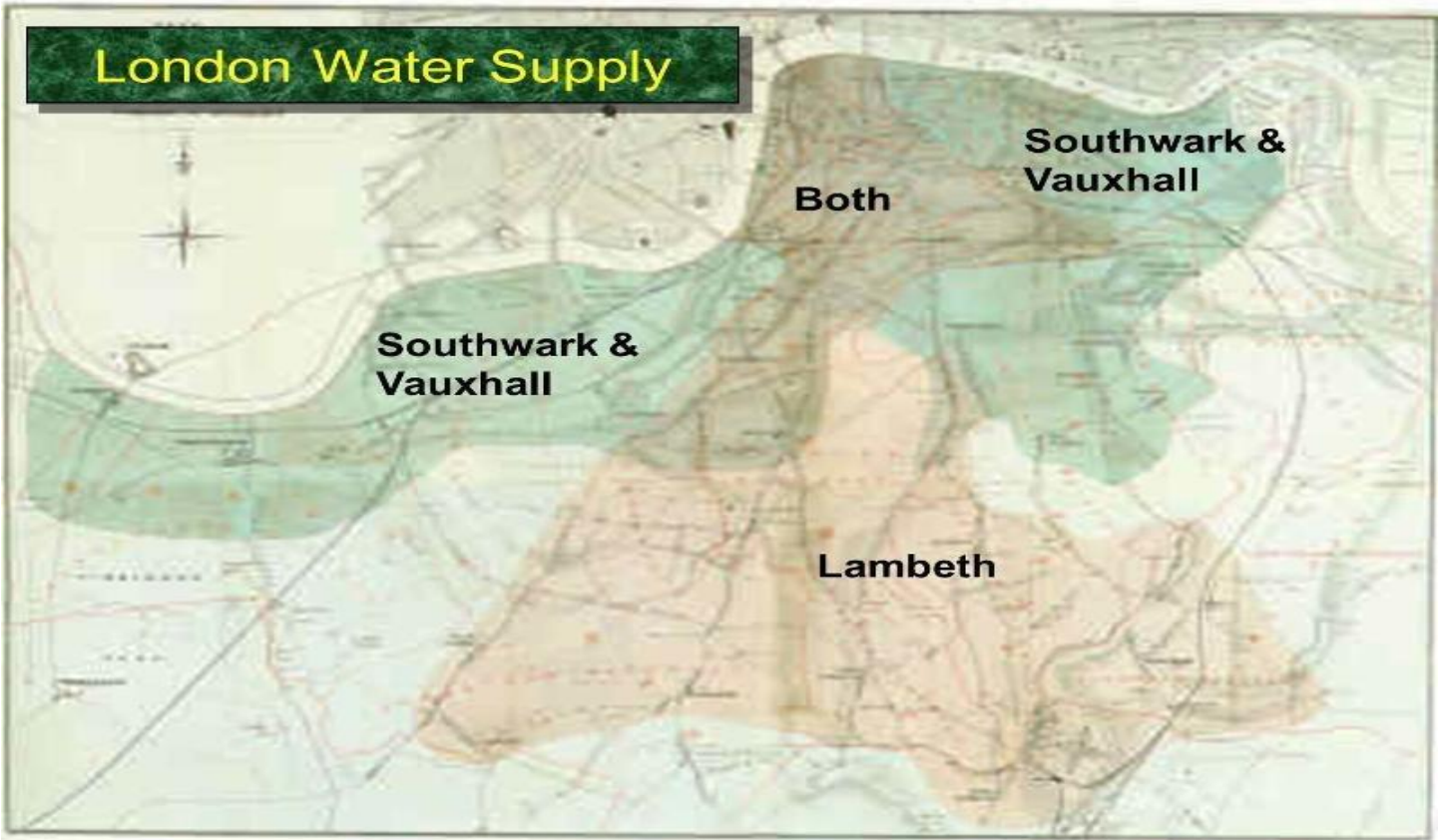








## London Water Supply



# Comparison

- **treatment group**
- **control group**
  - does not receive the treatment

Which houses were part of the treatment group?

- A. All houses in the region of overlap
- B. Houses served by S&V (dirty water) in the region of overlap
- C. Houses served by Lambeth (clean water) in the region of overlap

# Snow's “Grand Experiment”

“... there is no difference whatever in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded ...”

- The two groups were *similar except for the treatment*.



# Snow's table

Supply Area	Number of houses	Cholera deaths	Deaths per 10,000 houses
S&V (dirty water)	40,046	1,263	315
Lambeth (clean water)	26,107	98	37
Rest of London	256,423	1,422	59

Does dirty water cause cholera?

- A. Yes, I think so
- B. No, I don't think so
- C. Maybe, I can't tell

# Key to establishing causality

If the treatment and control groups are *similar apart from the treatment*, then differences between the outcomes in the two groups can be ascribed to the treatment.

# Trouble

If the treatment and control groups have **systematic differences other than the treatment**, then it might be difficult to identify causality.

Such differences are often present in **observational studies**.

When they lead researchers astray, they are called **confounding factors**.



# Randomize!

- If you assign individuals to treatment and control **at random**, then the two groups are likely to be similar apart from the treatment.
- You can account – mathematically – for variability in the assignment.
- **Randomized Controlled Experiment**

# Randomized Controlled Experiments

- Assign individuals to treatment and control **at random**

**Which of these questions cannot be answered by running a randomized controlled experiment?**

- A. Does daily meditation reduce anxiety?
- B. Does playing video games increase aggressive behavior?
- C. Does smoking cigarettes cause weight loss?
- D. Does early exposure to classical music cause higher IQ?
- E. All the above can be answered

# Careful ...

Regardless of what the dictionary says,  
in probability theory

**Random  $\neq$  Haphazard**

# Remember ...

- Get a clicker and register it this week!
- First lab is due Sunday.
- **Tonight: make sure you can access lab/HW by clicking on link in the schedule**