Lecture 1

Introduction
Cause and Effect

About me

- Justin Eldridge

- I'm new to UCSD!

- Before UCSD:
    - Taught at Ohio State
    - Ph.D. in machine learning from Ohio State
    - M.S. in computer science from Ohio State
    - B.S. in physics, B.S. in applied math from Ohio State

Welcome to DSC 10

- A course developed by UC Berkeley faculty and students and adapted by UCSD.
- All information for the course is searchable from the course website:

https://sites.google.com/eng.ucsd.edu/dsc-10-fall-2018/

Course Structure

Components

- In-class participation through iClickers

To change your remote frequency:

1. Press and hold power button until flashing
2. Enter two-letter frequency code
3. Checkmark / green light indicates success

What year are you at UCSD?

A.   First-year
B.   Second-year
C.   Third-year
D.   Fourth-year
E.   Other

Components

- In-class participation through iClickers **5%**

- Weekly lab **15%**

- Weekly homework assignments **25%**

- Projects **15%**

- Discussions led by a head tutor / TA

- Exams
  - Nov 1 (Thurs): Midterm in class **10%**
  - Dec 13 (Thurs): Final exam 3:00pm-6:00pm **30%**

Collaboration

Asking questions is highly encouraged

- Discuss all questions with each other (except exams)
- Submit lab assignments **individually**, but you can work with others in the same lab room

- Submit homework and a project individually or in pairs, but feel free to discuss with others

Collaboration

Asking questions is highly encouraged
- Discuss all questions with each other (except exams)
- Submit lab assignments individually, but you can work with others in the same lab room
- Submit homework and a project individually or in pairs, but feel free to discuss with others

The limits of collaboration
- Don't share solutions with each other or look at someone's code
- Partners should work together and be physically in the same place
- Academic integrity violations will result in failing the course

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?

First week assignments

- Lab 1: out today.
- Deadline: Wednesday @ 11:59 p.m.

- Homework 1: out today
- Deadline: Sunday, October 7 @ 11:59 p.m.

- **Start early**. Every week, make sure to finish most (if not all) by Friday.

Announcements

- There will be no discussion tomorrow, September 28!

Data Science

What is Data Science?

Drawing useful conclusions from data using computation

- **Exploration**
    - Identifying patterns in information
    - Uses visualizations
- **Prediction**
    - Making informed guesses
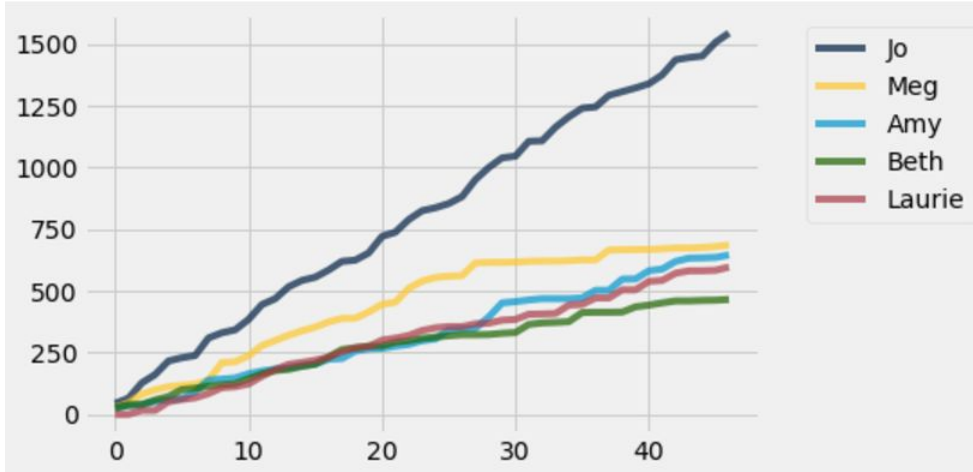   - Uses machine learning and optimization
- **Inference**
    - Quantifying whether those patterns are reliable
    - Uses randomization

Literature

[(Demo)](#)

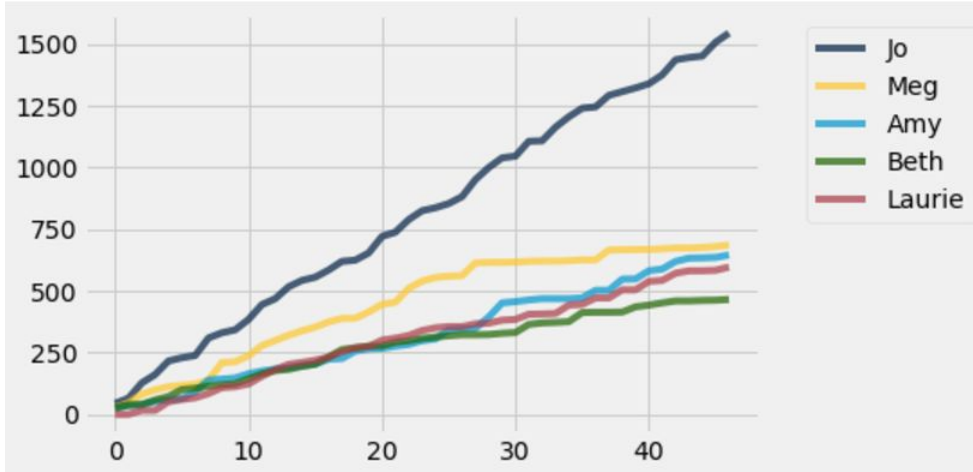# Literature

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

Part 2. Association and Causality

Really?

eating and health

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

JUNE 19, 2015   5:03 AM ET

ALLISON AUBREY

npr.org (report on a study in heart.bmj.com)

Observation

- **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"

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

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

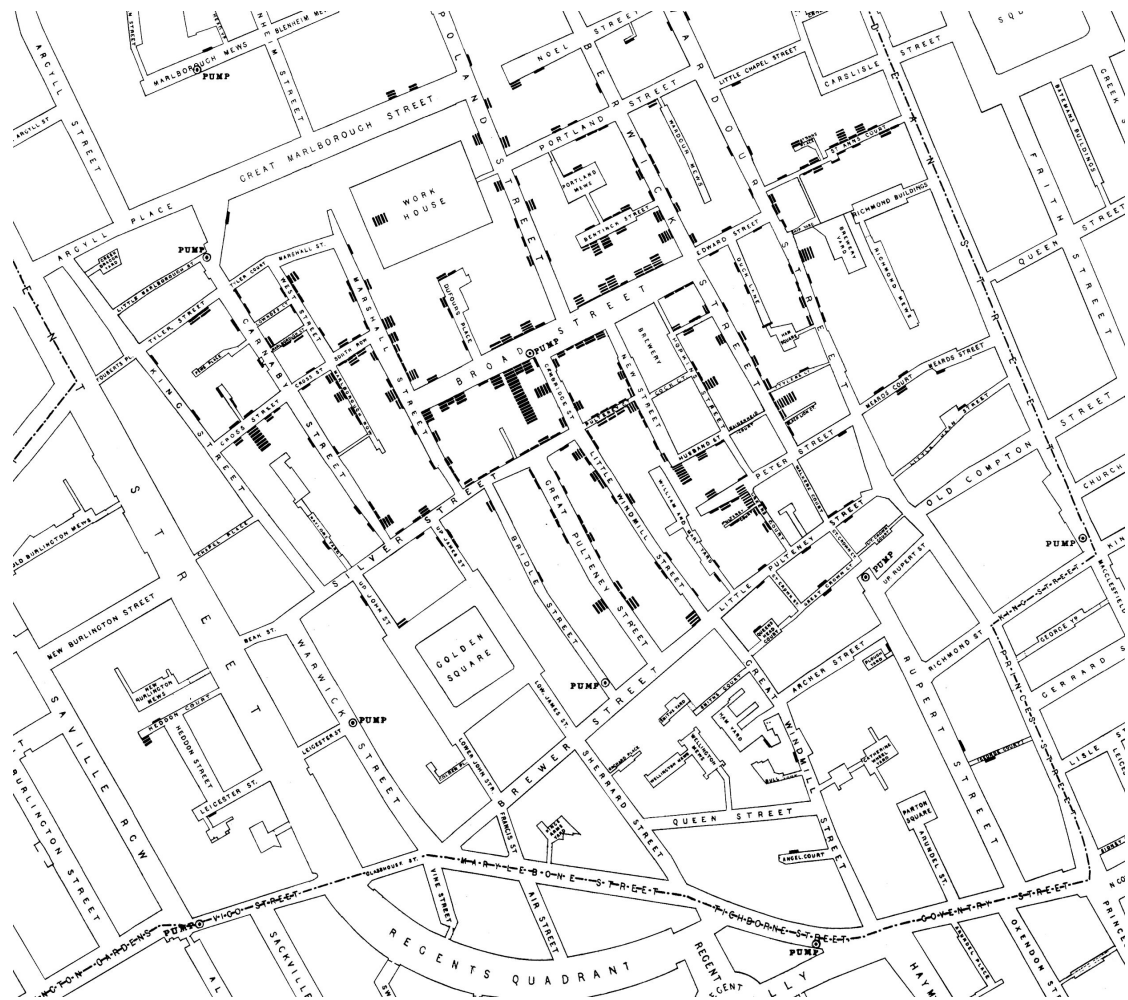FICTURESQUE SKETCHES OF LONDON.—WHITECHAPEL.

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
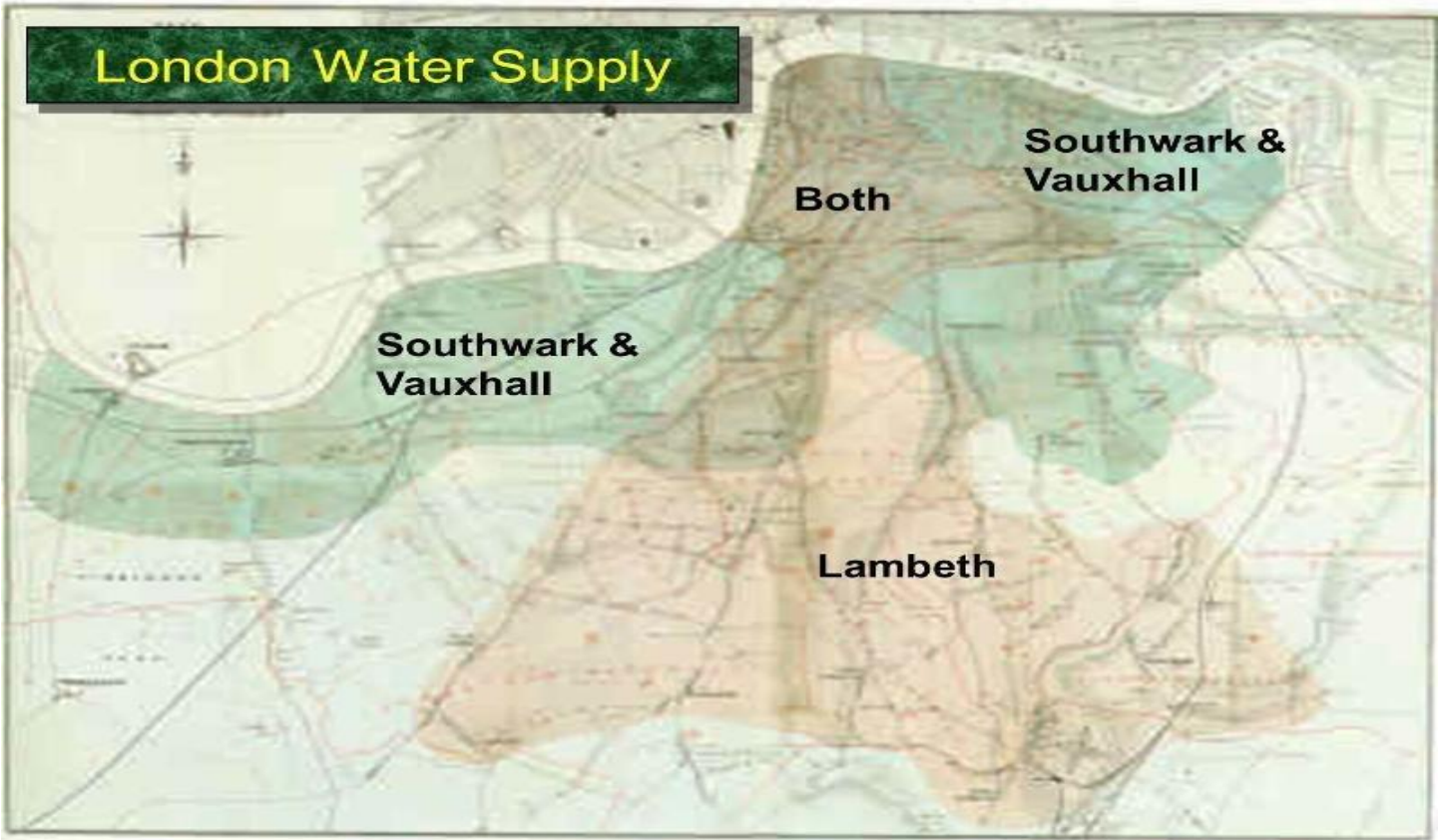 - Edwin Chadwick, Commissioner of the General Board of Health

John Snow, 1813-1858

London Water Supply

Southwark & Vauxhall

Both

Southwark & Vauxhall

Lambeth

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 ≠ Haphazard

Remember ...

- There is no discussion tomorrow.
- Bring clickers for Tuesday class to get attendance grade.