



# Lecture 1

---

Cause & Effect

# Welcome to DSC 10

- A course developed by UC Berkeley faculty and students and adapted by UC San Diego.
- Instructor: Dr. Janine Tiefenbruck (call me Janine)
  - My path:
    - BS in Math and Computer Science at Loyola MD
    - PhD in Math (Combinatorics) at UCSD
    - Teaching at UCSD: Math, CSE, now DSC
    - Fourth time teaching DSC 10
  - Outside interests: baking, paper crafts, board games, reading

# Course Website

<https://dsc10.com>

# Collaboration

Asking questions is highly encouraged

- Discuss all questions with each other (except exams)
- Submit lab assignments **individually**, but you can work with others
- Submit homework and a project individually or in pairs using **pair programming**, 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
- Submit homework and a project individually or in pairs using **pair programming**, 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

BC

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

**Do you have any programming experience?**

- A. Yes, I'm a pro!
- B. I have some experience.
- C. I know a few basic concepts.
- D. No experience whatsoever!

# First Assignment

- Lab 1  
Deadline: Friday 11:59pm
- **Start early.**
- **Submit often.**

# First Meetings

- Lecture: now!
- Discussion Section:
  - Wednesday 2pm or 3pm, Center 216
- Programming Basics:
  - Thursday 5pm or 6pm, CSE B230
  - Starts next week, not today



# 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

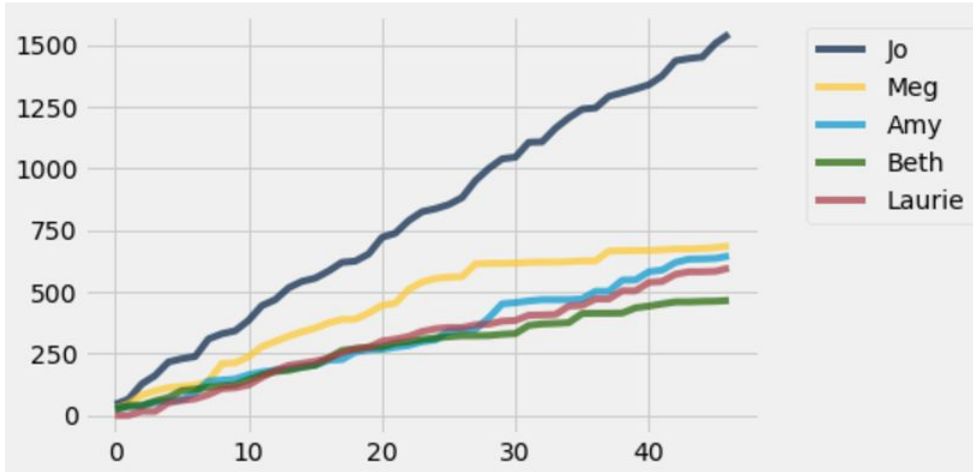
- **Inference**

- Quantifying whether those patterns are reliable
- Uses randomization

# Literature

(Demo)

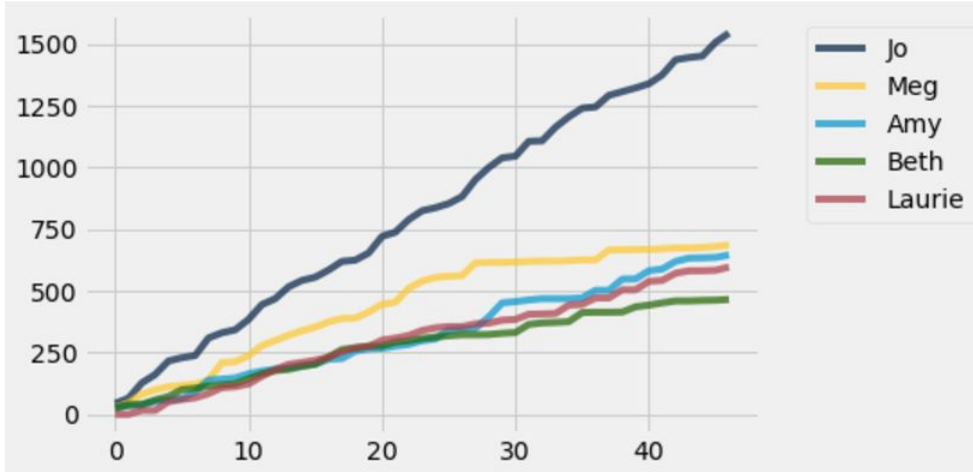
# Literature



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

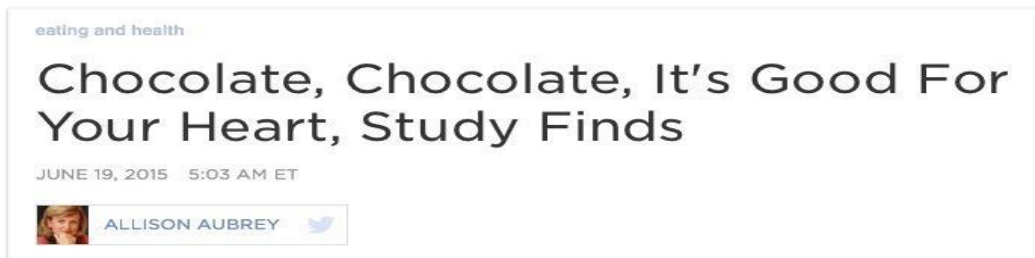


Laurie is a man who marries one of the sisters at the end. Which one?

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

# **Association and Causality**

# Really?



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

“link”

# 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

# The next question

---

Does chocolate consumption **lead to** a reduction in heart disease?

- **causality**

Does chocolate consumption lead to (cause) a reduction in heart disease?

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

# The next question

---

Does chocolate consumption **lead to** a reduction in heart disease?

- **causality**

This question is often harder to answer.

“[The study] doesn’t prove a cause-and-effect relationship between chocolate and reduced risk of heart disease and stroke.”

– JoAnn Manson, chief of Preventive Medicine at Brigham and Women’s Hospital, Boston

---

**London, 1854**

# Miasmas, miasmatism, miasmatists

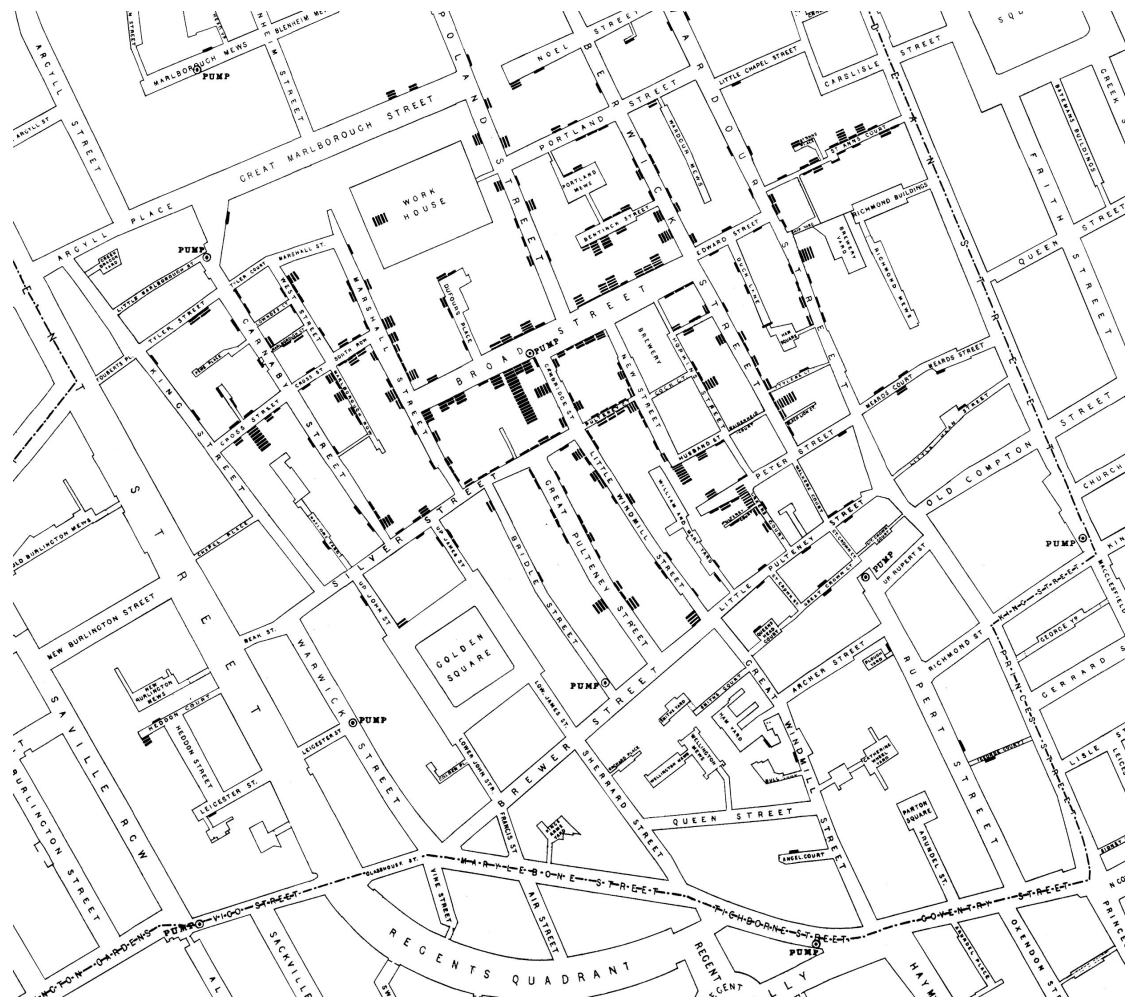
---

- **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

---



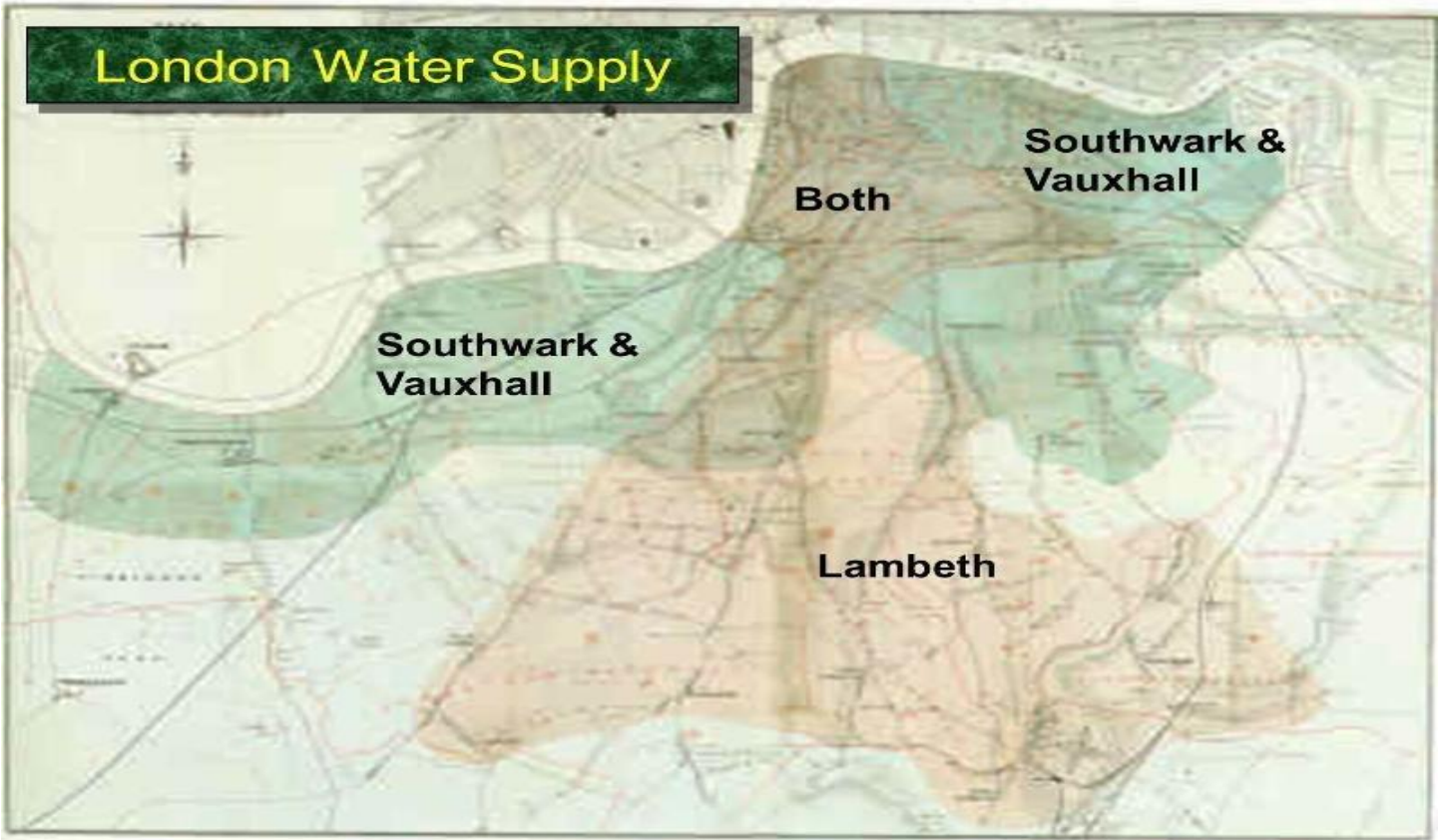






# Establishing Causation

## 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.

---

# Confounding



# 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?

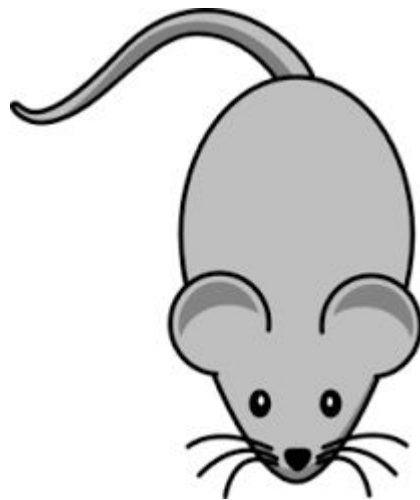
- A. Does daily meditation reduce anxiety?
  - B. Does playing video games increase aggressive behavior?
  - C. Does smoking cigarettes cause weight loss?
  - D. Does early exposure to classical increase a child's IQ?
-

# Careful ...

---

Regardless of what the dictionary says,  
in probability theory

**Random  $\neq$  Haphazard**



# **Summary: Cause & Effect**

# Comparison

---

- Group by some *treatment* and measure some *outcome*
  - Simplest setting: a *treatment group* and a *control group*
  - If the *outcome* differs between these two groups, that's evidence of an *association* (or *relation*)
    - E.g., the top-tier chocolate eaters died of heart disease at a lower rate (12%) than chocolate abstainers (17%)
  - If the two groups are similar in all ways but the *treatment*, a difference in the *outcome* is also evidence of *causality*
-

# Confounding

---

- If the treatment and control groups have systematic differences other than the treatment itself, then it might be difficult to identify a causal link
  - When these systematic differences lead researchers astray, they are called *confounding factors*
  - Such differences are often present in observational studies
    - *Observational study*: the researcher does not choose which subjects receive the treatment
    - *Controlled experiment*: the researcher designs a procedure for selecting the treatment and control groups
-

# Randomize!

---

- When subjects are split up *randomly*, it's unlikely that there will be systematic differences between the groups
  - And it's possible to account for the chance of a difference
  - Therefore, *randomized controlled experiments* are the most reliable way to establish causal relations
-