UCSD

DSC 10

Winter 2021

Lecture 1

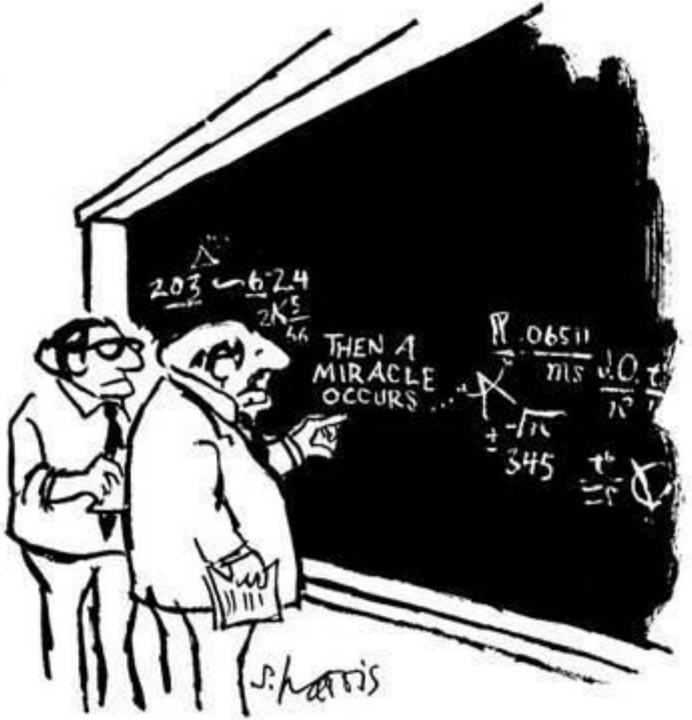
Cause & Effect

Welcome to DSC 10

- A guided tour of data science
- A course developed by UC Berkeley and adapted by UC San Diego
- Learn just enough programming and statistics to do data science
- Statistics without too much math, mostly simulation

About Me

- 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
    - Fifth time teaching DSC 10
  - Outside interests: baking, paper crafts, board games, reading

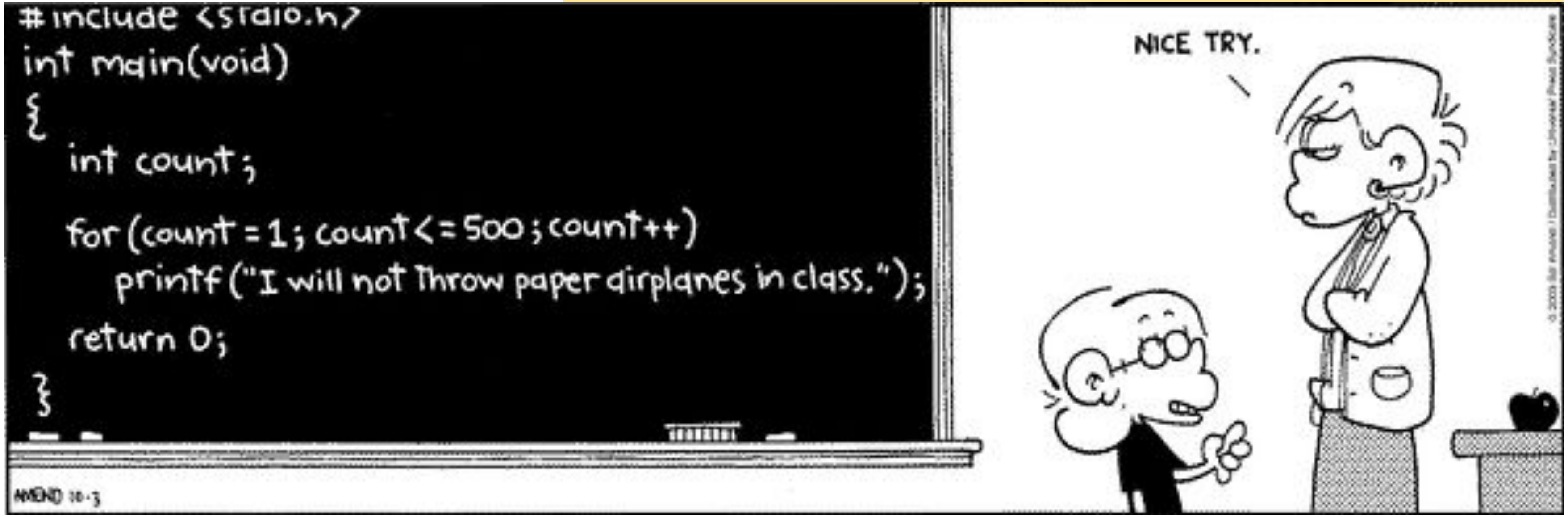


"I think you should be more explicit here in step two."

About you

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!



```
#include <stdio.h>
int main(void)
{
   int count;
   for(count=1;count<=500;count++)
      printf("I will not Throw paper airplanes in class.");
   return 0;
}
```

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 (from same **team**) 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 (from same **team**) 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 often result in failing the course

First Assignment

- Lab 1

  Deadline: Tuesday 11:59pm

- **Start early**.

- **Submit often.**

Data Science

What is Data Science?

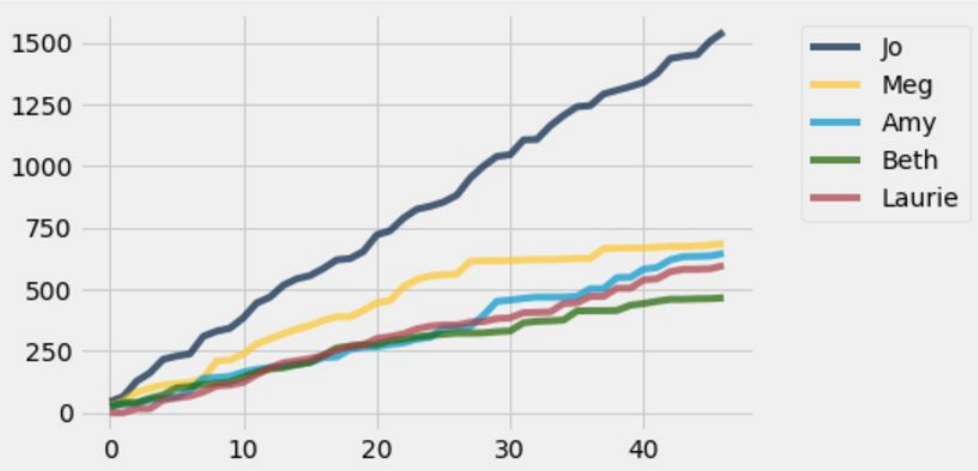
Drawing useful conclusions from data using computation

- **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)

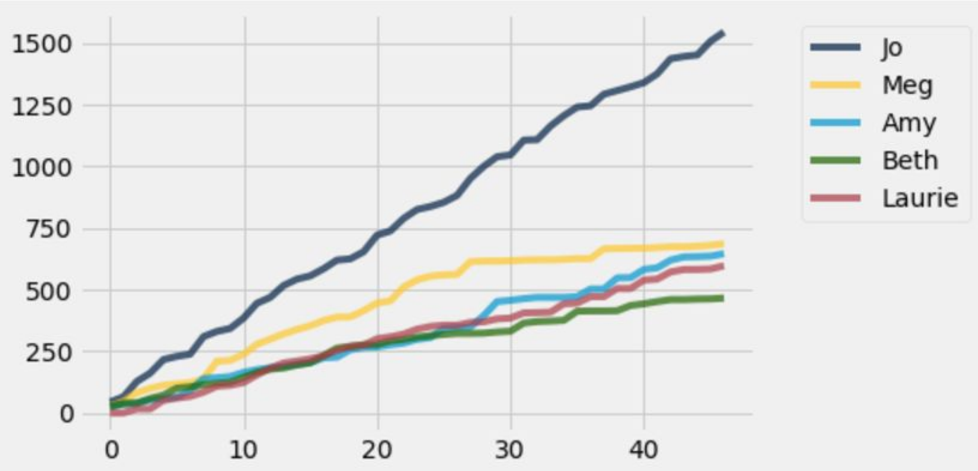
Literature



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

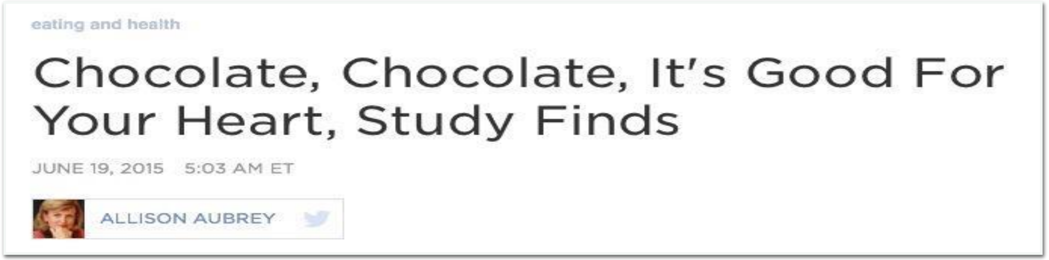


Laurie is a man who marries one of the sisters at the end. Which one?

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

Association and Causality

Really?



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”
  “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*

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."

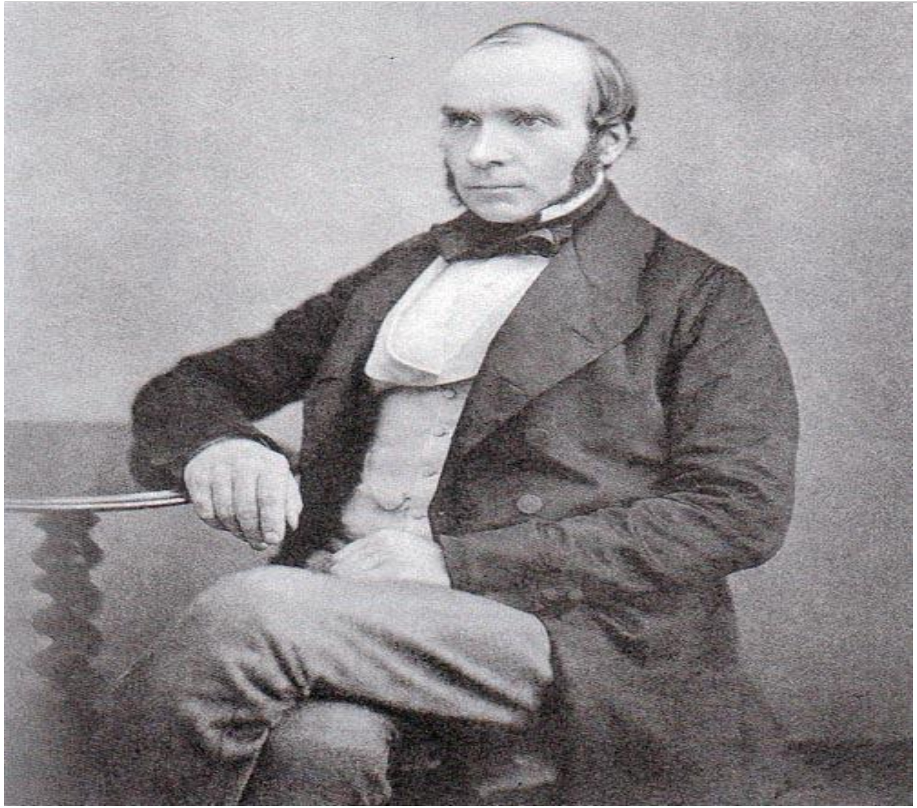
- *JoAnn Manson, chief of Preventive Medicine at Brigham and Women's Hospital, Boston*

London, 1854

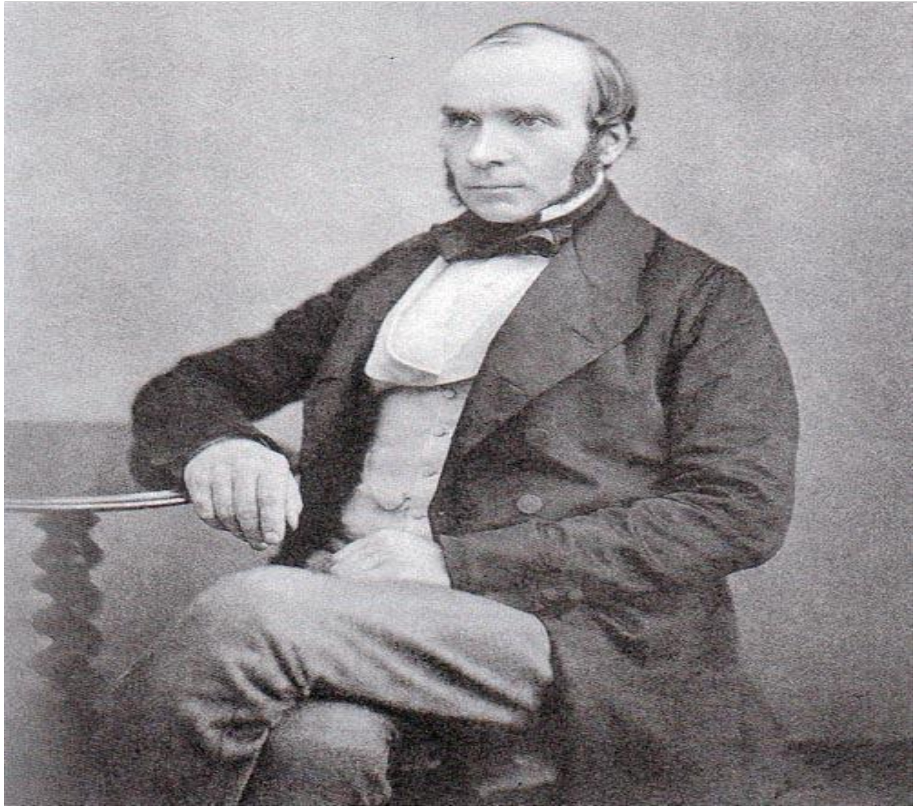
Miasmas, miasmatism, miasmatists

- **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

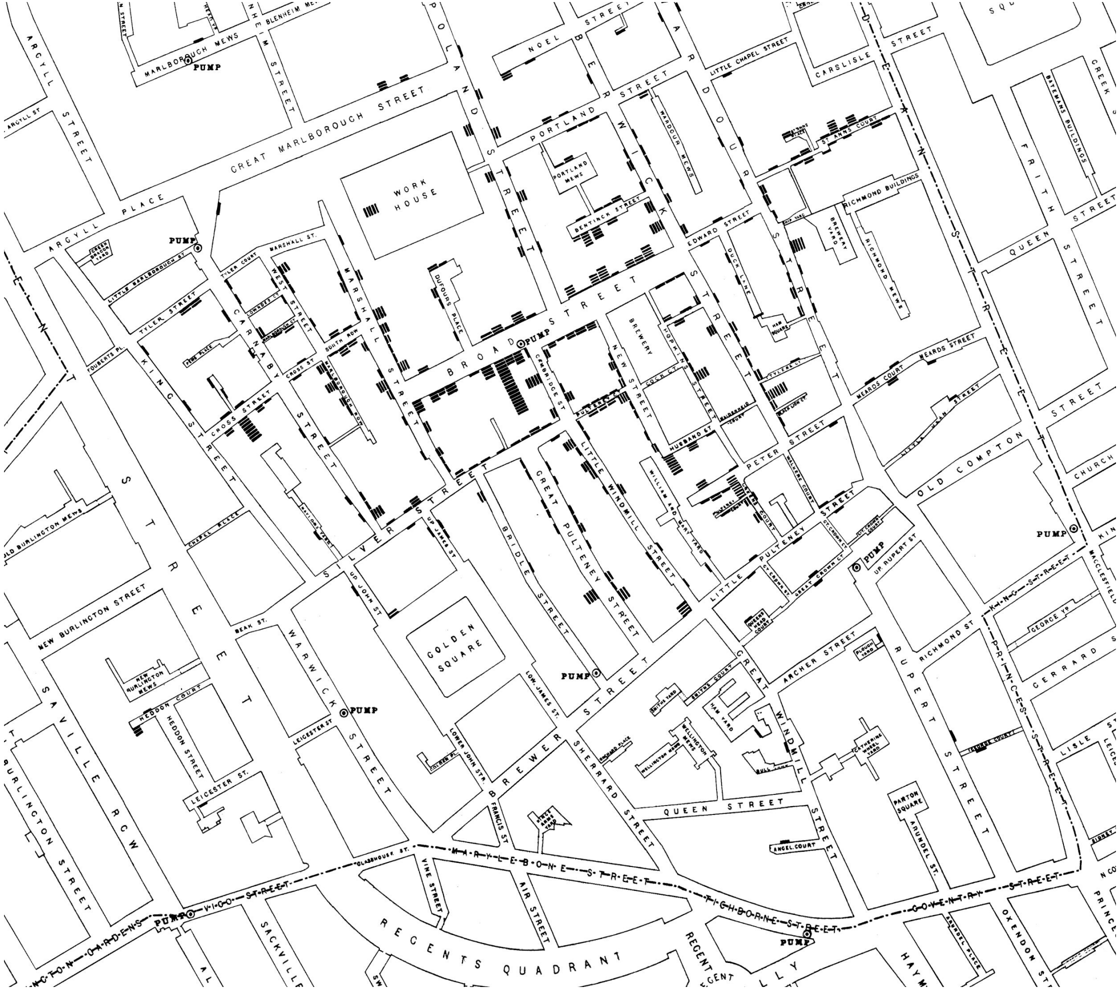
John Snow, 1813-1858



John Snow, 1813-1858



Not this Jon Snow

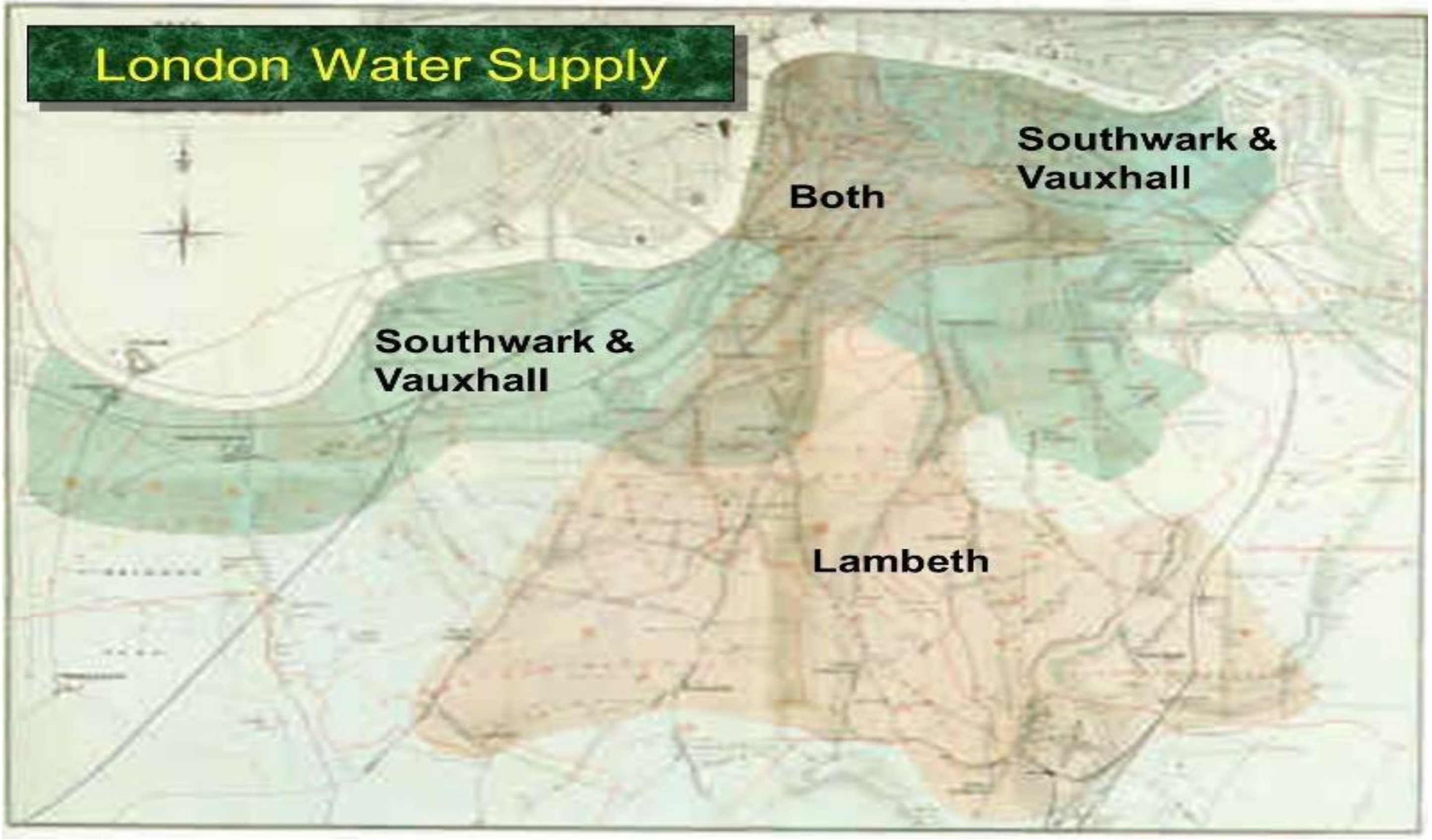


MARLBOROUGH MEWS
PUMP
BLENHEIM MEWS
ARGYLL STREET
ARGYLL ST
GREAT MARLBOROUGH STREET
POLAND STREET
NOEL STREET
BERWICK STREET
WARDOUR STREET
LITTLE CHAPEL STREET
CARLISLE STREET
DEAN STREET
SQU
CREEK ST
BATEMANS BUILDINGS
FRITH STREET
PORTLAND STREET
PORTLAND MEWS
WARDOUR MEWS
ST ANNS COURT
RICHMOND BUILDINGS
RICHMOND MEWS
BREWERY YARD
WORK HOUSE
ARGYLL PLACE
GREEN DRAGON YARD
LITTLE MARLBOROUGH ST
TYLER COURT
MARSHALL ST
WEST STREET
BENTINCK STREET
EDWARD STREET
DUCK LANE
HAM SQUARE
QUEEN STREET
DUFOURS PLACE
BROAD STREET
TYLER STREET
CARNABY STREET
MARSHALL STREET
SOUTH ROW
FOUBERTS PL
KING STREET
CROSS STREET
CAMBRIDGE ST
BREWERY
NEW STREET
HOPKINS STREET
HUSBAND ST
MEARDS COURT
MEARDS STREET
PETER STREET
LITTLE WINDMILL STREET
GREAT PULTENEY STREET
WILLIAM AND MARY YARD
LITTLE PULTENEY STREET
OLD COMPTON STREET
CHURCH STREET
DEAN STREET
SILVER STREET
UP. JAMES ST
UP. JOHN ST
BRIDLE STREET
GOLDEN SQUARE
CHAPEL PLACE
OLD BURLINGTON MEWS
NEW BURLINGTON STREET
NEW BURLINGTON MEWS
REGENT STREET
HEDDON COURT
HEDDON STREET
LEICESTER ST
BEAK ST
WARWICK STREET
LOW. JAMES ST
LOWER JOHN STR.
BREWER STREET
SHERRARD STREET
GREAT WINDMILL STREET
ARCHER STREET
RUPERT STREET
RICHMOND ST
UP. RUPERT ST
GREAT CROWN CT
GEORGE YD
GERRARD STREET
KING STREET
MACCLESFIELD
PRINCES STREET
LISLE STREET
HAM YARD
WELLINGTON MEWS
CATHERINE WHEEL YARD
PLOUGH YARD
RUPERT COURT
NELSONS COURT
QUEEN STREET
ANGEL COURT
PARTON SQUARE
ARUNDEL ST.
COVENTRY STREET
TICHBORNE STREET
MARYLEBONE STREET
GLASSHOUSE ST.
VIGO STREET
BURLINGTON GARDENS
SAVILLE ROW
BURLINGTON STREET
SACKVILLE
VINE STREET
AIR STREET
FRANCIS ST
KINGS ARMS YARD
RECENTS QUADRANT
REGENT
HAYMARKET
ARUNDEL PLACE
OXENDON ST
PICCADILLY

LEXINGTON STREET W1
JOHN SNOW
JOHN SNOW
SALOON
SALOON
JOHN SNOW

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**

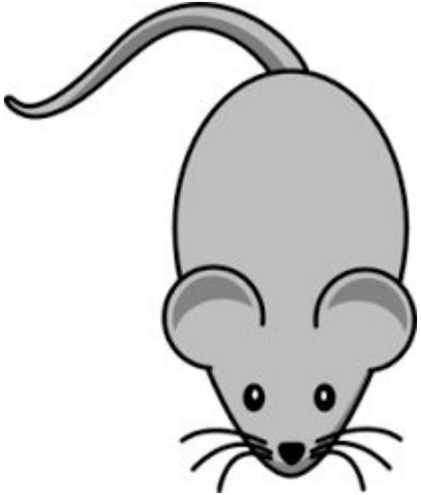
Which of these questions cannot be answered by running a randomized controlled experiment?

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 ≠ 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