DSC 10: Lecture 1

Introduction
Cause and Effect

Credit: Anindita Adhikari and John DeNero

Welcome to DSC 10

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

Welcome to DSC 10

- A guided tour of data science.
- Learn just enough programming, statistics to do data science.
- Statistics done without (much) math. Instead: simulation.

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?

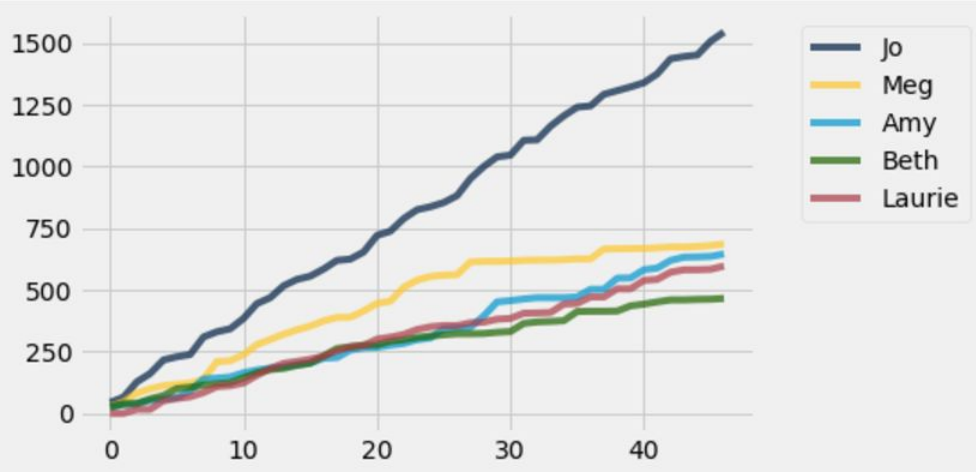
Drawing useful conclusions from data in a principled way.

- **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)

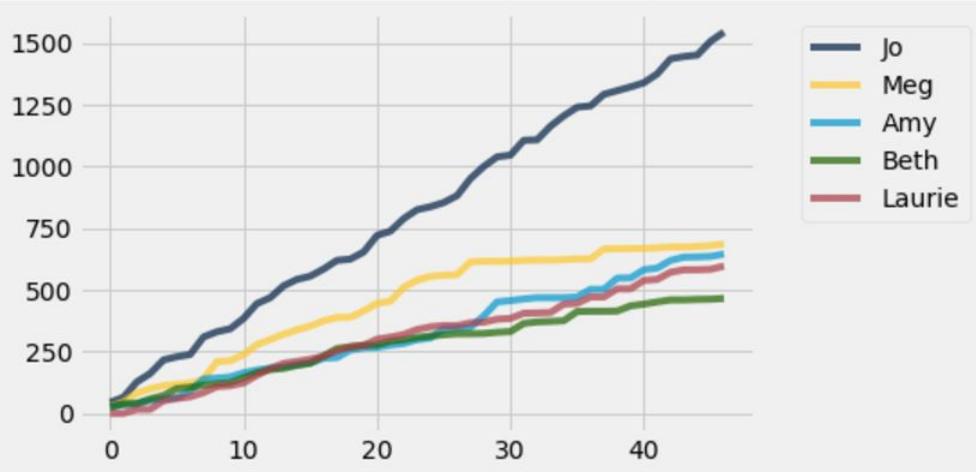
Literature



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

Course Page:

www.dsc10.com

Lecture 01 :
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)

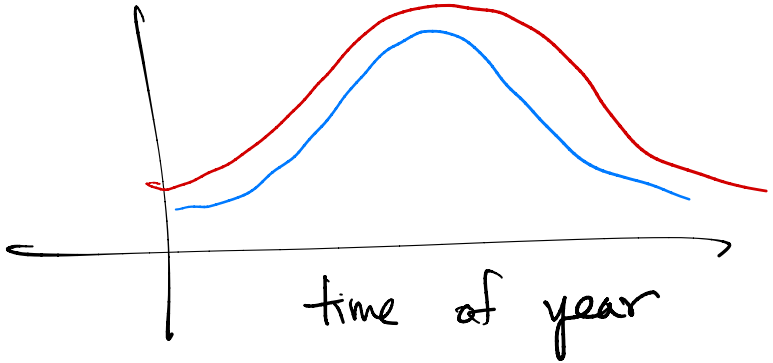
Definitions

- **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*

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

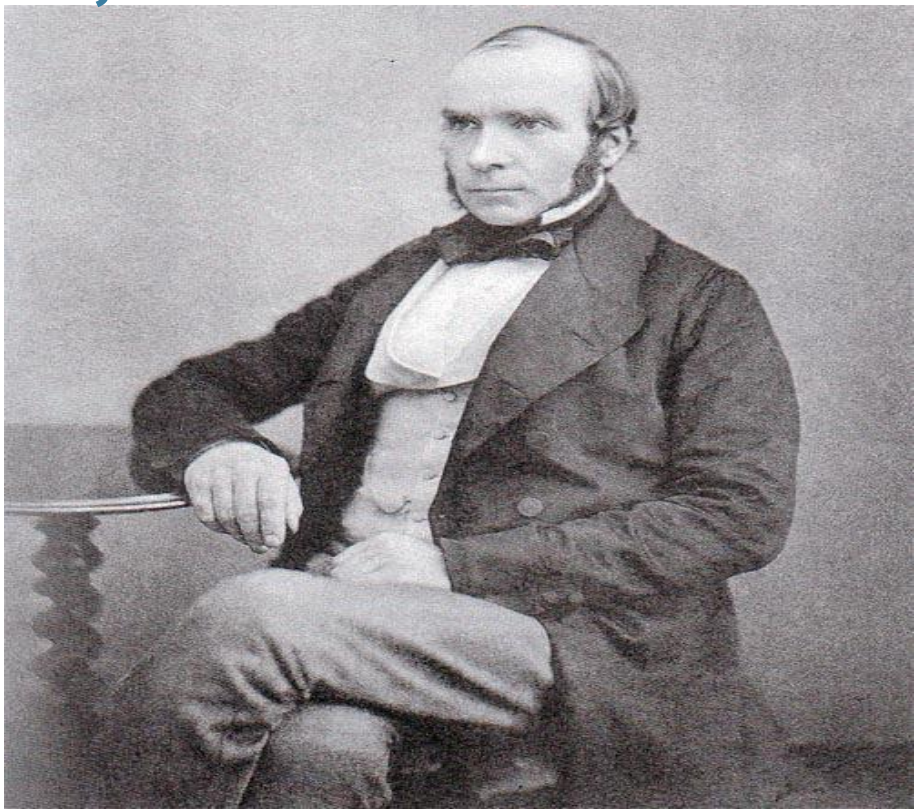
A. Yes, I think so
B. No, I don't think so
C. Maybe, I can't tell

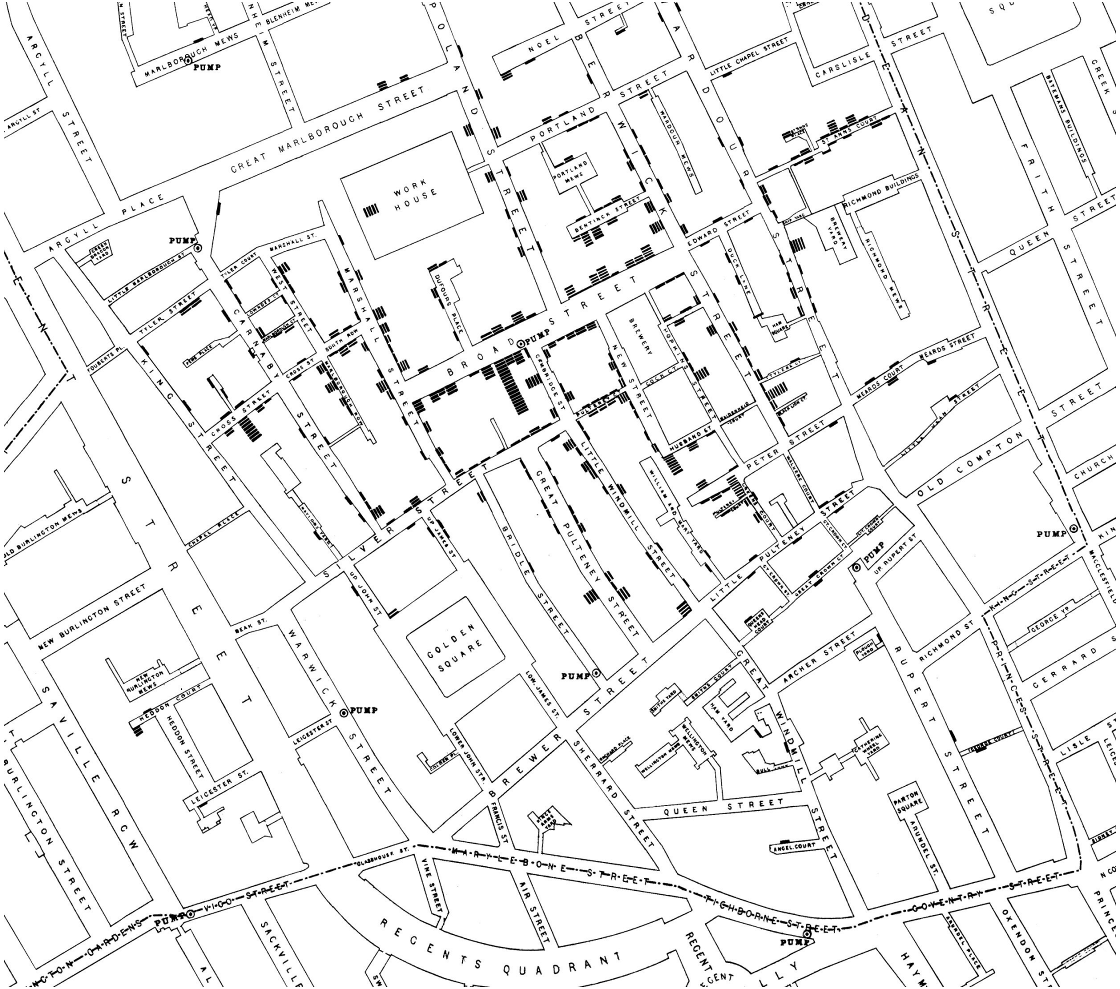
London in the 1800s

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

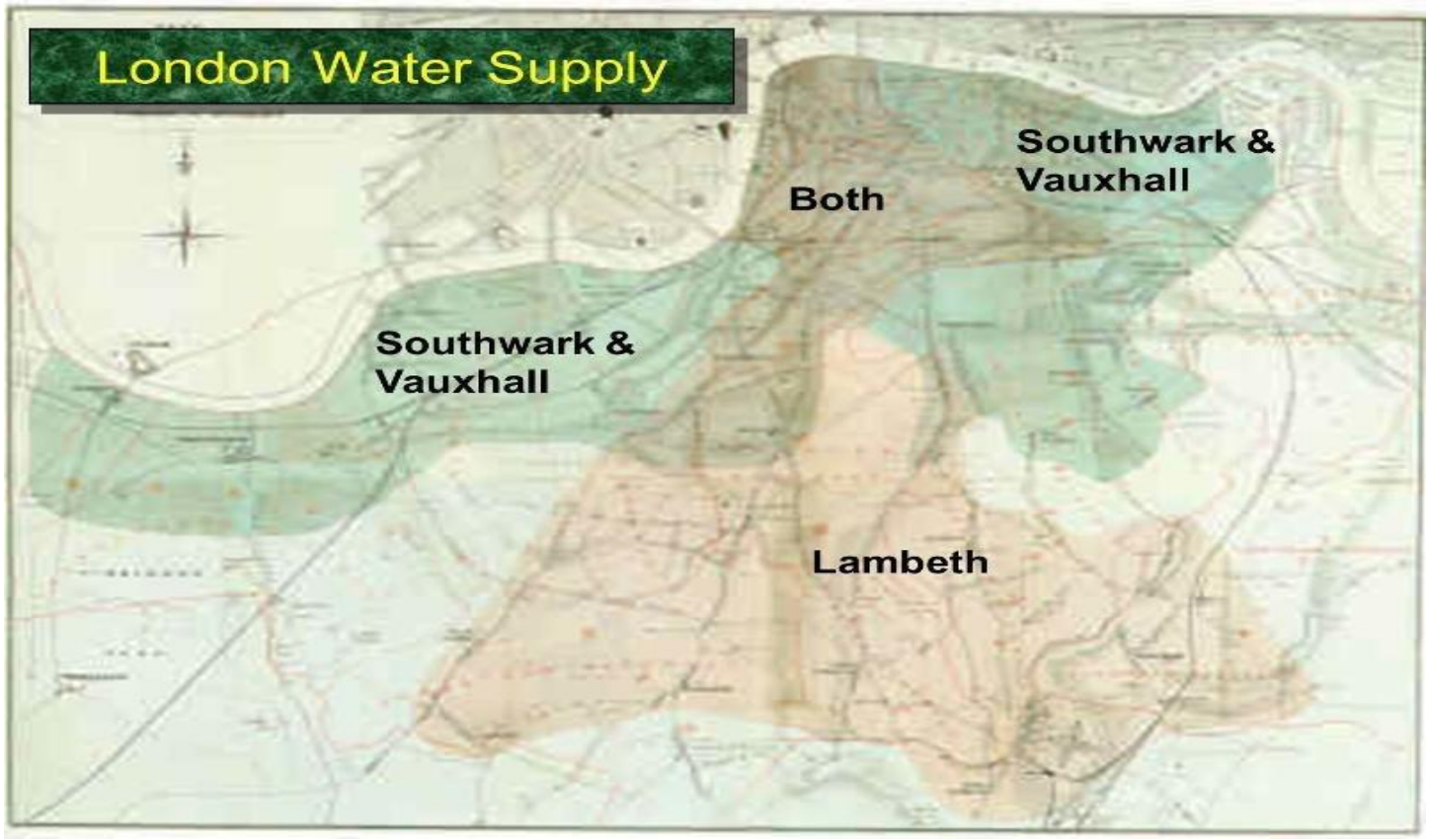
John Snow, 1813-1858





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


JOHN SNOW
JOHN SNOW


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