Foundation of Data Science

Lecture 1

Introduction

Adopted from the data science class at UCB and UCSD

Part 1 - Introduction

Welcome to Data Science and ML

- A course partially adopted from the Data 8 class at Berkeley and DSC 10 UCSD

Fudan Elearning site: elearning.fudan.edu.cn

- Paul Cao— UC-San Diego
- Office hours: TBD
- yic242@eng.ucsd.edu
- TAs: Siyue Han, Ruian He

Course Structure

Syllabus

- Lectures 3 times a week for 4 weeks – T/W/Th
- Lab component for most days
- Final exam on Thursday on 8/1/2019
  - Lecture participation points: 5%
  - Labs: 60%
  - Final: 35%
  - Optional programming assignment

Pair Programming

- You will work in pairs for labs and projects
- Pair programming
  - Driver and navigator
  - Have to work together on the same machine
  - You should have a computer (any platform is ok)

Data Science and Machine Learning

What is Data Science?

Drawing useful conclusions from data using computation

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

Why Data Science?

Demo

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](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”

An answer

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*

- Yes, this points to an association (in my opinion)

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.”

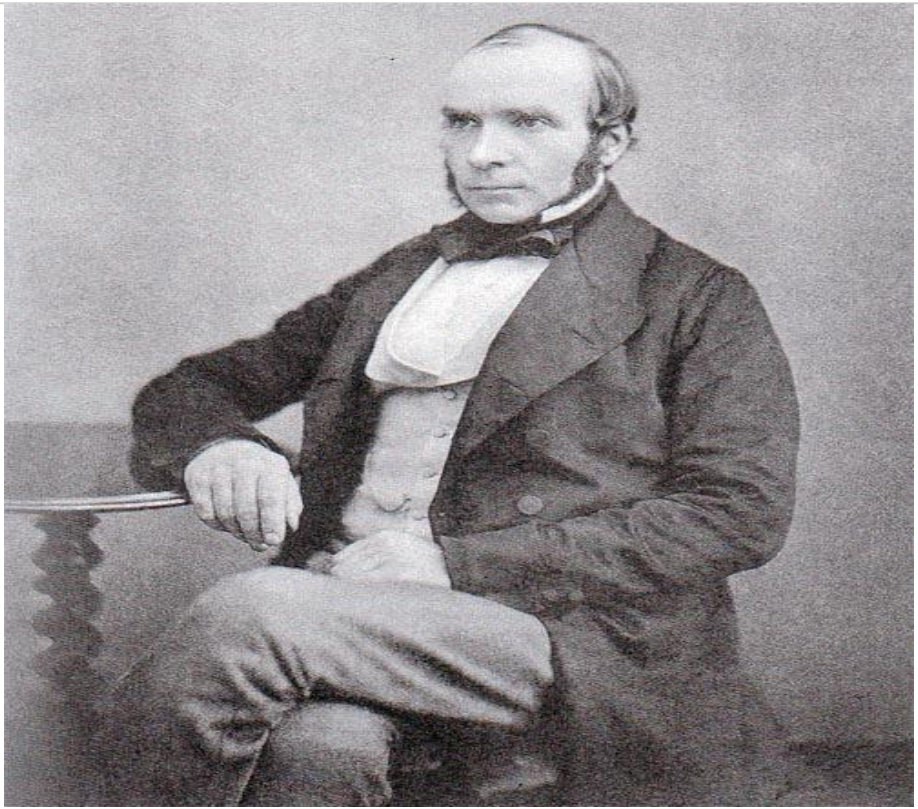
- *JoAnn Manson, chief of Preventive Medicine at Brigham and Women’s Hospital, Boston*

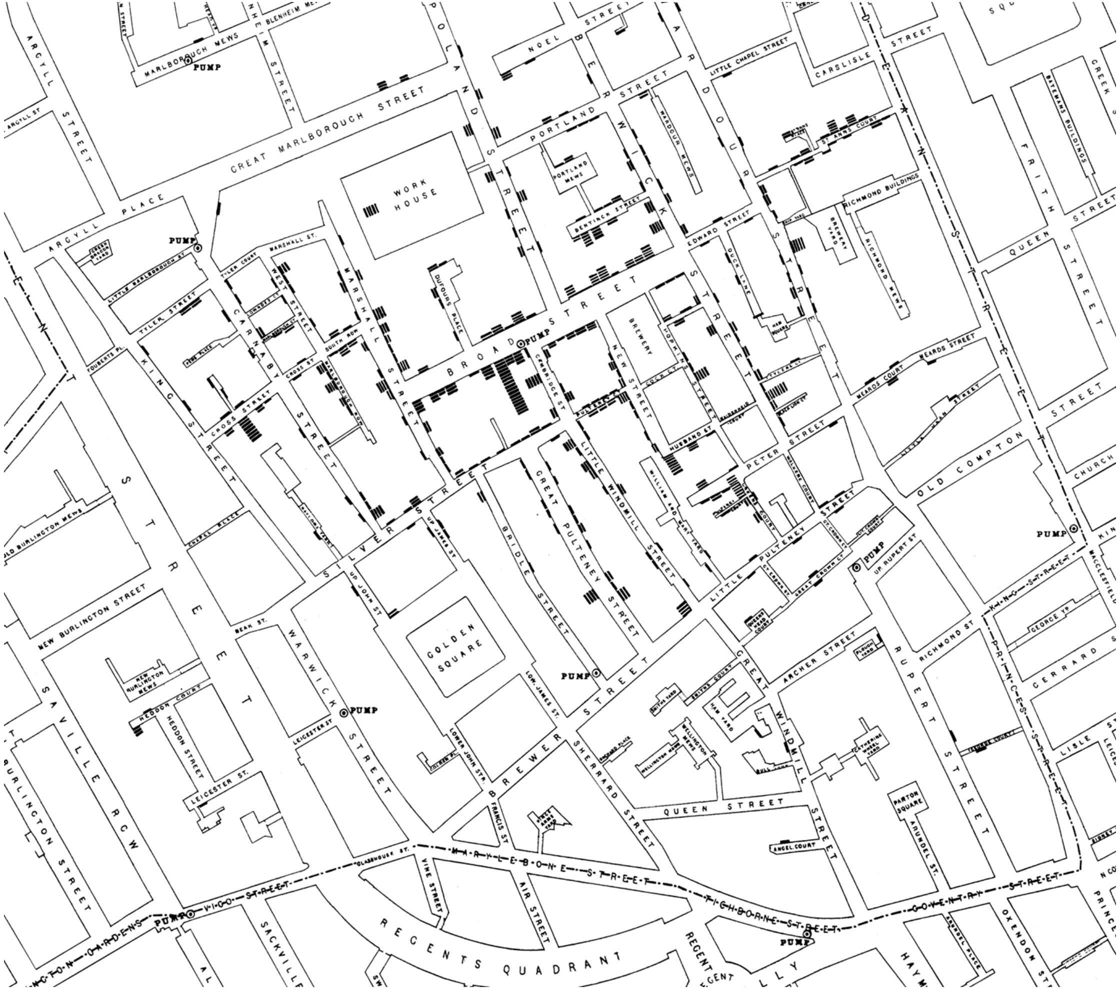
London, 1854

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




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

John Snow

John Snow

3.7 ★★★½☆ 193 reviews

Directions

Pub

SAVE | NEARBY | SEND TO YOUR PHONE | SHARE

Dark-wood saloon bar serving Yorkshire ales, named after doctor who traced London cholera outbreak. - Google

39 Broadwick St, Carnaby, London W1F 9QJ, UK

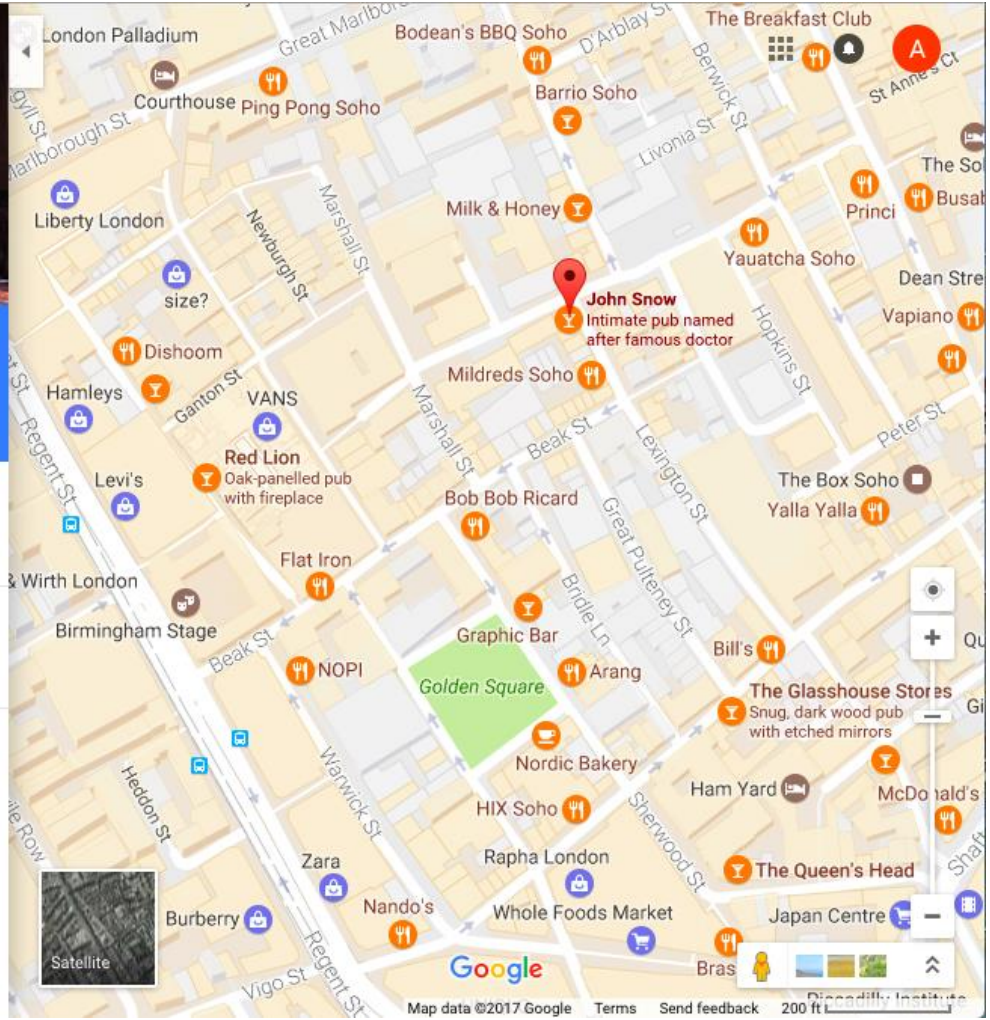
+44 20 7437 1344

Closed. Opens at 12:00 PM

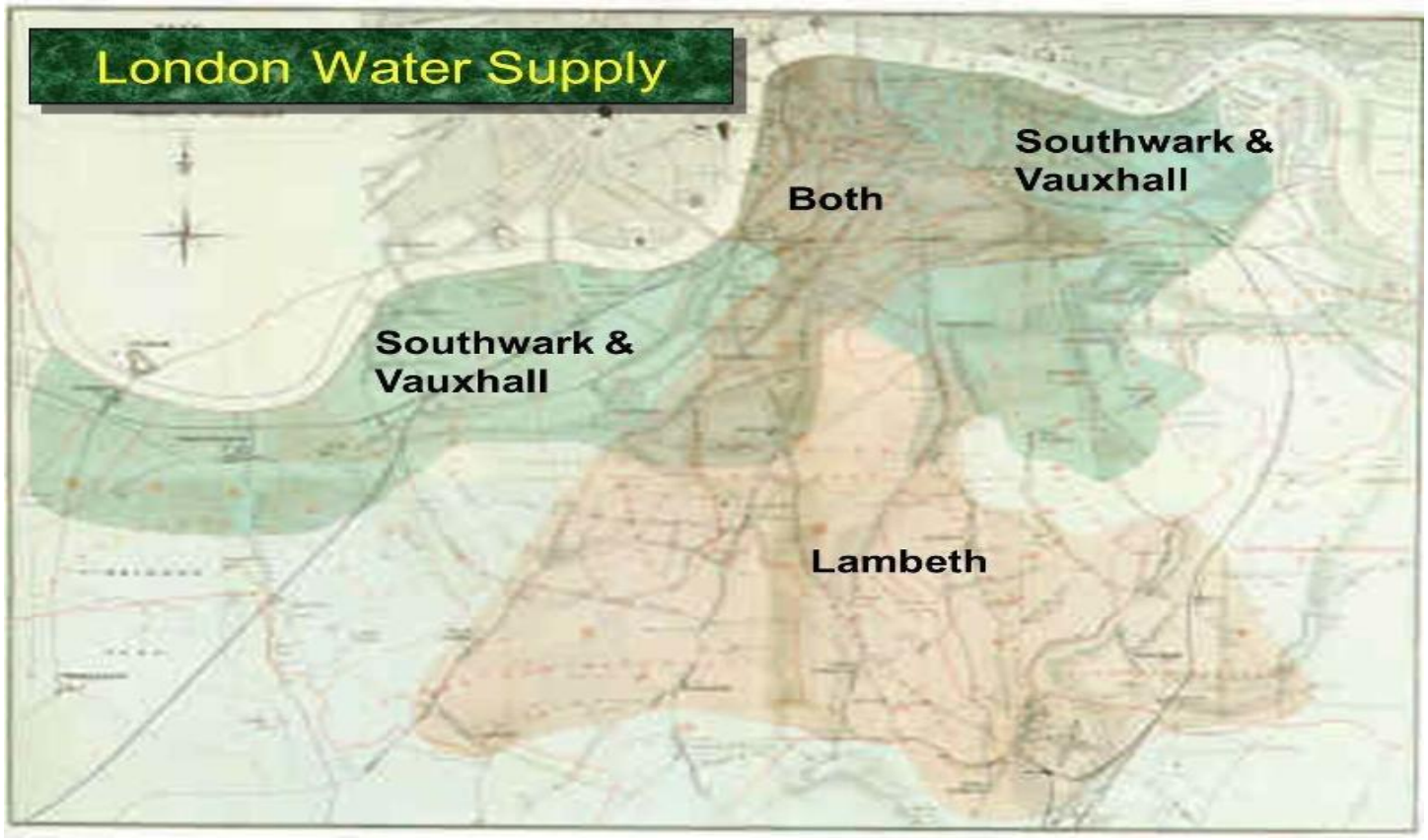
Claim this business

Suggest an edit

Add a label




JOHN SNOW
JOHN SNOW


London Water Supply
Southwark &
Vauxhall
Both
Southwark &
Vauxhall
Lambeth

Comparison

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

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	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

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.

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!

- 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**

Careful ...

Regardless of what the dictionary says,
in probability theory

Random ≠ Haphazard