AE 102: Data Analysis and Interpretation

January - April 2016

Authors: Prabhu Ramachandran

General information

- Instructor: Prabhu Ramachandran
- Slot 8: Mon/Thu 2-3:30pm.
- Venue: LC 002
- Office hours: TBD
- TAs:
  - Ajay Vora <15401001@iitb.ac.in>
  - Mohit Rohatgi <143010035@iitb.ac.in>

Preliminaries

- Learn
- Interact
- Be curious

Preliminaries ...

- I don't know everything
- It is harder to teach than you think!
  - Be humble

Data analysis

Data is everywhere!

Data

- Data is everywhere
  - Geography
  - Demographics
  - Wealth
  - Weather
  - Opinions/Polls

Measurement is Key

If you can measure it in some form, you can analyze it.

Data analysis

- Collect the data systematically
- Study it
- Understand and make sense of it

So what?

- Understand correlations and causation
- Understand relationships
- Predict things

Data analysis

- Formally:
  - Visualization
  - Inference
  - Modeling
  - Prediction

Noise, randomness

So what's with all this probability and statistics business?

Noise, randomness

- The data isn’t clean
- Noise is inherent in every measurement

- Consider a simple thing like a coin toss!

Endless examples ...

- The weather
- The stock market
- Interest rates
- The behavior of human beings?

Endless examples ...

- The weather
- The stock market
- Interest rates
- The behavior of human beings?

- Have you crossed the road recently?

So what do we do?

Probability theory

- Study the random and use the same approach
- Quantify uncertainity
- Make statements with levels of certainity

Data analysis

- Collect the data systematically
- Study it
- Understand and make sense of it

So what?

- Understand correlations and causation
- Understand relationships
- Predict

Data analysis

- Formally:
  - Visualization
  - Inference
  - Modeling
  - Prediction

Example 1

- Does vitamin C help fight a cold?

Example 2

- Is Chocolate good for you?

Example 2

- What causes <favorite> cancer?

Data Analysis: How?

- The right tools
    1. Mathematics
    2. Computation

Mathematics

- Statistics
    - Descriptive statistics
        - Gather/describe data
    - Inferential statistics
        - Draw conclusions using the data
    - Probability theory

Computation

- Datasets are large
- Easy to process on the computer
- Simulation!

Some famous examples

- John Graunt (exercise!)
- John Snow
- Abraham Wald
- Target

John Snow story

- Doctor London in the 1850's
- Disease
- Miasamas
- Cholera outbreaks

Interlude: observational study

- Simply observe the data as is
- Nothing is controlled by the scientist
- Q: Does a given "Treatment" have an effect on an "Outcome"?
- Relation between treatment and outcome: "association"
- Association can be "causal"

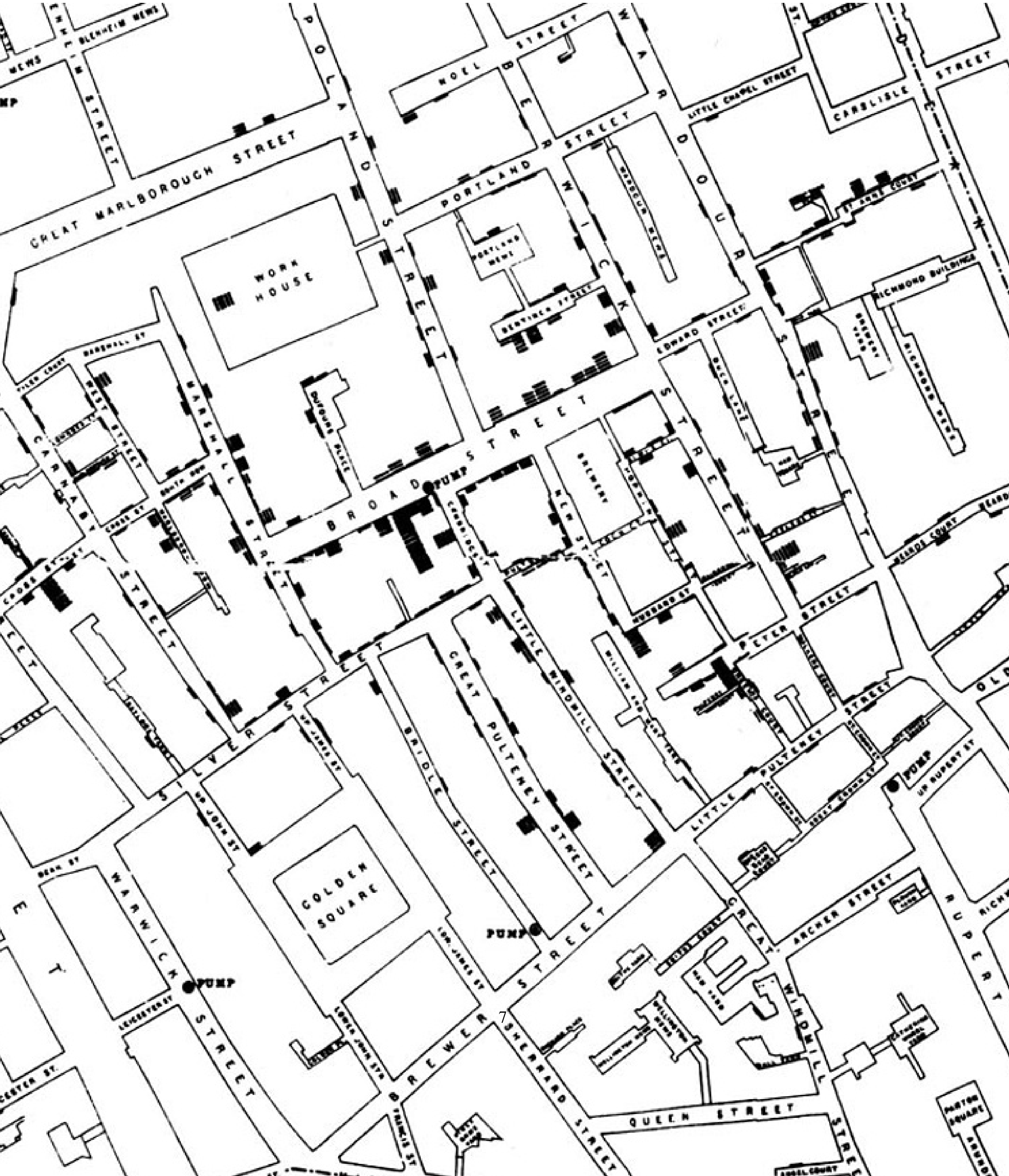
Determining causality

- Causality is key and often takes two steps
    1. Observe to establish an association
    2. More careful analysis/study to determine causality

Back to John Snow

- The importance of visualization!

The map


BLENHEIM MEWS
NEWS
PUMP
GREAT MARLBOROUGH STREET
POLAND STREET
NOEL STREET
BERWICK STREET
PORTLAND STREET
WARDOUR STREET
LITTLE CHAPEL STREET
CARLISLE STREET
DEAN STREET
ST ANNE'S COURT
WORK HOUSE
PORTLAND MEWS
BENTINCK STREET
RICHMOND BUILDINGS
RICHMOND MEWS
EDWARD STREET
MARSHALL STREET
DUFOURS PLACE
CARNABY STREET
WEST STREET
BROAD STREET
PUMP
CAMBRIDGE ST
NEW STREET
BREWERY
HUSBAND ST
CROSS ST
SILVER STREET
LITTLE WINDMILL STREET
GREAT PULTENEY STREET
BRIDLE STREET
PETER STREET
LITTLE PULTENEY STREET
PUMP
UP RUPERT ST
UP JAMES ST
UP JOHN ST
BEAK ST
GOLDEN SQUARE
WARWICK STREET
PUMP
PUMP
LOW. JAMES ST
LOWER JOHN ST
LEICESTER ST
BREWER STREET
SHERRARD STREET
GREAT WINDMILL STREET
ARCHER STREET
RUPERT STREET
QUEEN STREET
ANGEL COURT

Note

- No deaths in brewery
- Some near Rupert street
- Some scattered deaths a bit away
- Strange deaths far away

Lessons

- Established an association
- So what was the cause?

Snow's experiment

- Comparison
- Two **identical** groups with only the water changing
- Eliminate confounding factors

Interlude: confounding factors

- 1960's: studies found coffee drinkers had higher rates of lung cancer than non-coffee drinkers

- Q: Is coffee a "cause" for lung cancer?

Interlude: controlled experiments

- Treatment groups
- Control groups
- Randomized assignment

- Randomized Controlled Experiment
- AKA Randomized Controlled Trial (RCT)

Interlude: the placebo effect

- Placebos
- Malcebo

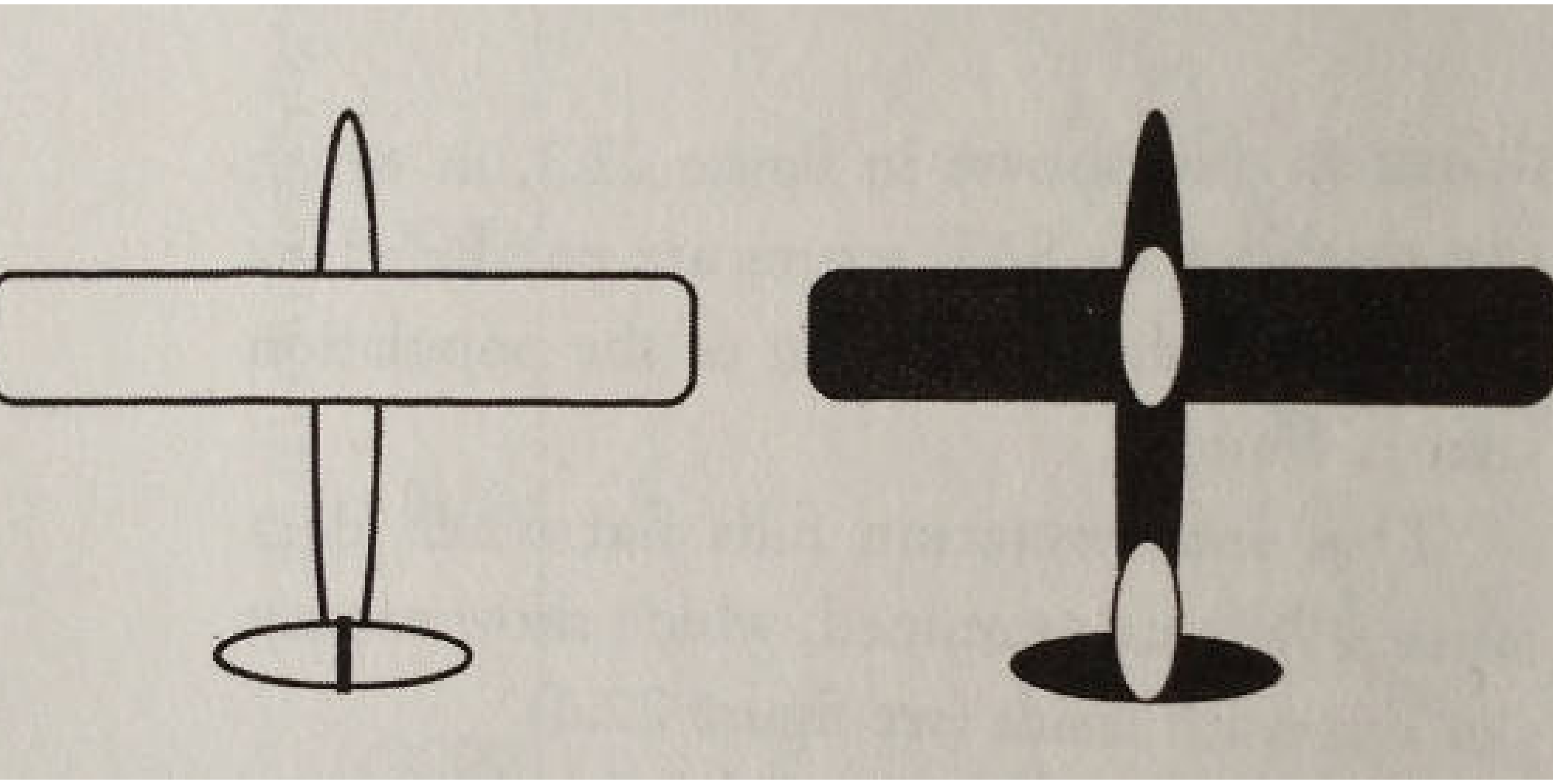
Interlude: blind trials

- Randomized assignment
- Using placebos for the controls

The story of Abraham Wald

- WW2 allied bombers
- Heavy attack by anti-aircraft fire

Visualization to the rescue!



Lesson

- Selection bias!
- Critical thinking
- Reasonable and convicing explanations

The target story

- Recent event

The target story

- Scary possibilities!
- Keep privacy in mind

Estimating chance

- How does one factor chance events?
- How does one measure confidence in a conclusion?

Back to the course

- Data analysis
  - Basic statistical techniques
  - Computational tools to use

Computer setup

- Many assignments will require a computer
- Happy to help make this easier
- Will be using Python

Grading

- 40% of top mark is fail
- Extra tutorial sessions for weak students
- 30% assignments
- 10% Q1
- 10% Q2
- 20% MS
- 30% ES

Resources

- Reference text book:

  Introduction to Probability and Statistics for Engineers and Scientists Sheldon M. Ross, Academic Press.
- Gentle reading:

  The Cartoon guide to statistics by Larry Gonick and Woollcott Smith

Attendance

- Strongly suggest you attend

- 23 out of 84 failed last year!

Plan of Action

- Self-learn chapters 1-3 from the textbook
- Mini-quiz on Thursday 7th.
- Quiz 1 on 18th Jan.
- Meanwhile we learn to use Python for data analysis

Image credits

- John Snow map: http://data8.org/text/assets/images/snow_map.jpg
- Image for Abraham Wald: http://www.fastcodesign.com/1671172/how-a-story-from-world-war-ii-shapes-faceb
- http://h.fastcompany.net/multisite_files/codesign/imagecache/inline-large/inline/2012/11/1671172-inline-inline-wwii-facebook-design.jpg