DS 100 – Intro to Data Science

Lecture 1 – Course Introduction

01/21/2025

Adam Poliak


Data Science is the study of extracting value from data
JEANNETTE WING

What is Data Science?

*"Data science is the study of extracting **value** from data"*

- Jeannette Wing

Value

- Requires domain expertise to determine what value is
- *Value from data* is different based on the domain and the needs

What is Data Science?

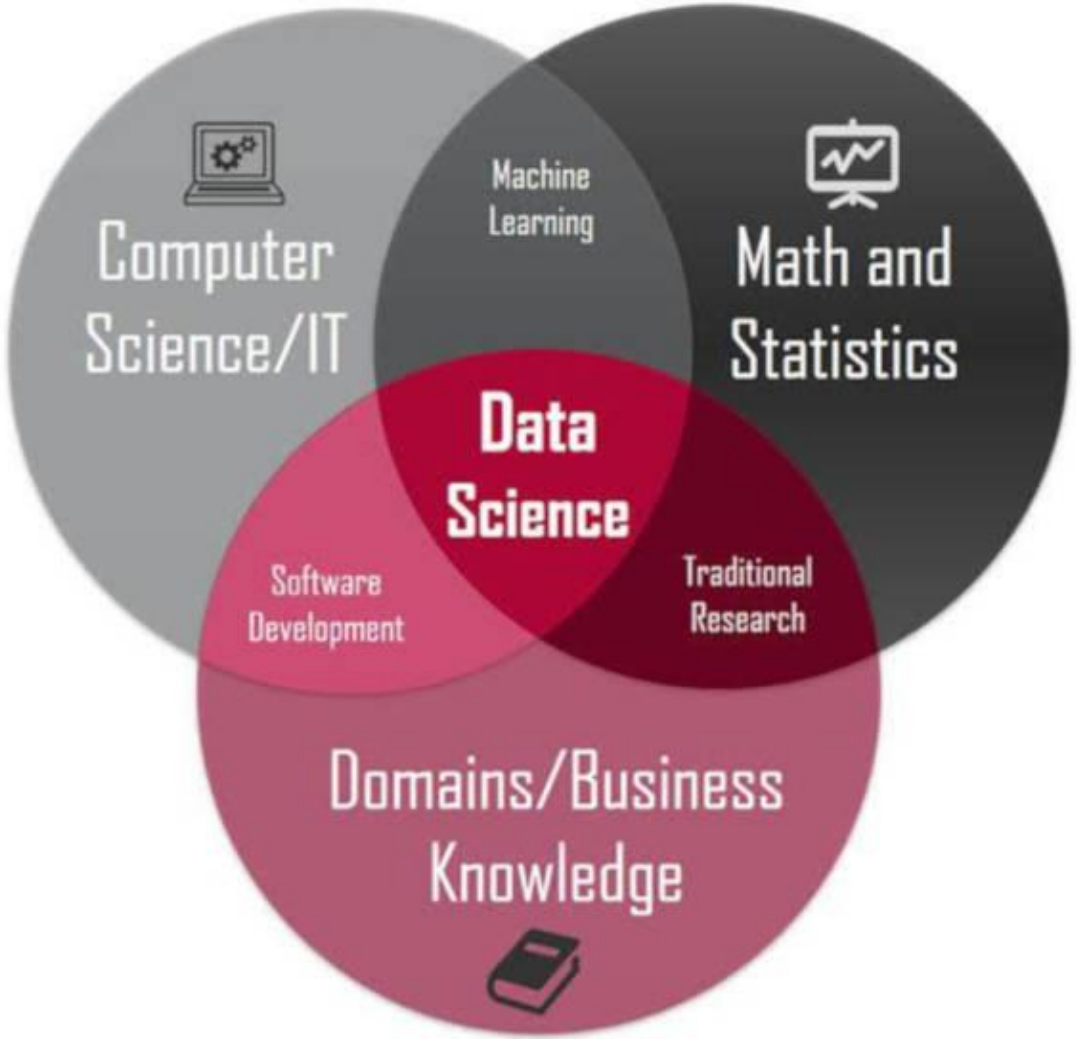
*"Data science is the study of **extracting** value from data"*

- Jeannette Wing

Extracting

- Emphasizes action on data
- Mining information

Math + Computer Science + Domain Knowledge



Data Science in this course

Exploration

- Discover patterns in data
- Articulate insights (visualizations)

Inference

- Make reliable conclusions about the world
- Statistics is useful

Prediction

- Informed guesses about unseen data

Course outline

Exploration

- Introduction to Python
- Working with data

Inference

- Probability
- Statistics

Prediction

- Machine Learning
- Regression & Classification

Logistics

brynmawr.edu

Communication

Course webpage

https://bmc-ds-100.github.io/

Piazza:

Course communication

Ask & Answer Questions

Can post anonymously

Course Meetings: 1:10 – 2:30 T/Th

Live class

- primarily lectures
- Discussions & exercises
- Q/A
- Recorded

Pre-class readings:

- Expected to read assigned readings before class
- Distributed on course schedule

Assignments

LEARNING BY DOING

Assignments

Labs
Homework
Projects

Labs

Tuesdays after class (Carpenter Library Computer Room)

Submit on Gradescope

Individual submission, but can discuss among each other

Due following Monday

Homeworks

12 through out semester

Completely individually

Late day policy

8 late days (no questions ask), can use at most 2 on a hw

Due Wednesdays

HW00 – available, due Wed 01/29

Projects

Like homeworks, but longer

2 weeks to complete

Can be done in pairs

3 in the semester

Assessments

Midterm

- After spring break
- Date tbd

Final

- Schedule (not self-scheduled)

Rubric

Participation	5%
Weekly Lab	5%
Weekly HW	30%
Projects	20%
Midterm	15%
Final	25%

Participation

During class:

- Discussion
- Asking questions

Asynchronous

- Active on Pizza
- Daily quizzes

Assignment Logistics

Distribution

Course website – schedule page

JupyterHub

https://dsci-b100-hub.brynmawr.edu/

Gradescope

Course Staff

Adam Poliak

apoliak@brynmawr.edu

3rd year at BMC

3rd time teaching this class (previous 2 at Barnard)

Research:

Natural Language Processing

Data Science applied to text data

Our job is to help you succeed!

Course Policies

brynmawr.edu

Collaboration

Encouraged to discuss problems

Do not share solutions

Late days & Dropped Assignments

8 Late Days for homeworks and projects

- Can only use 2 per assignment

Drop 2 lowest homeworks & 2 lowest labs

Learn By Doing

brynmawr.edu

Cause & Effect

A link (Coffee and Health)

Three coffees a day linked to a range of health benefits

Research based on 200 previous studies worldwide says frequent drinkers less likely to get diabetes, heart disease, dementia and some cancers

▲ The findings supported other studies showing the health benefits of drinking coffee. Photograph: Wu Hong/EPA

A link (Chocolate and Health)

EATING AND HEALTH

Chocolate, Chocolate, It's Good For Your Heart, Study Finds

June 19, 2015 · 5:03 AM ET

Heard on Morning Edition

ALLISON AUBREY

Observation

individuals, study subjects, participants, units

- European adults

Treatment

- *Chocolate through out the day*

outcome

- *heart disease*

Question 1: Association

Is there **any relation** between consuming chocolate and heart disease?

- **association**
  - any relation
  - Three coffees a day **linked** to improve health

An answer

Look at 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

Does this point to an association?

Question 2: Causality

- Does eating chocolate **lead to** reduced heart disease?
  - **Causality**

- Causality 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

Association

brynmawr.edu

King Cholera – London 1850's

Illustration from *Punch* (1852)

Miasma, miasmatism, miasmatists

Bad smells given off by waste & rotting matter

Potential remedies:

- “fly to clean air”
- “a pocket full of posies”
- “fire off barrels of gunpowder”

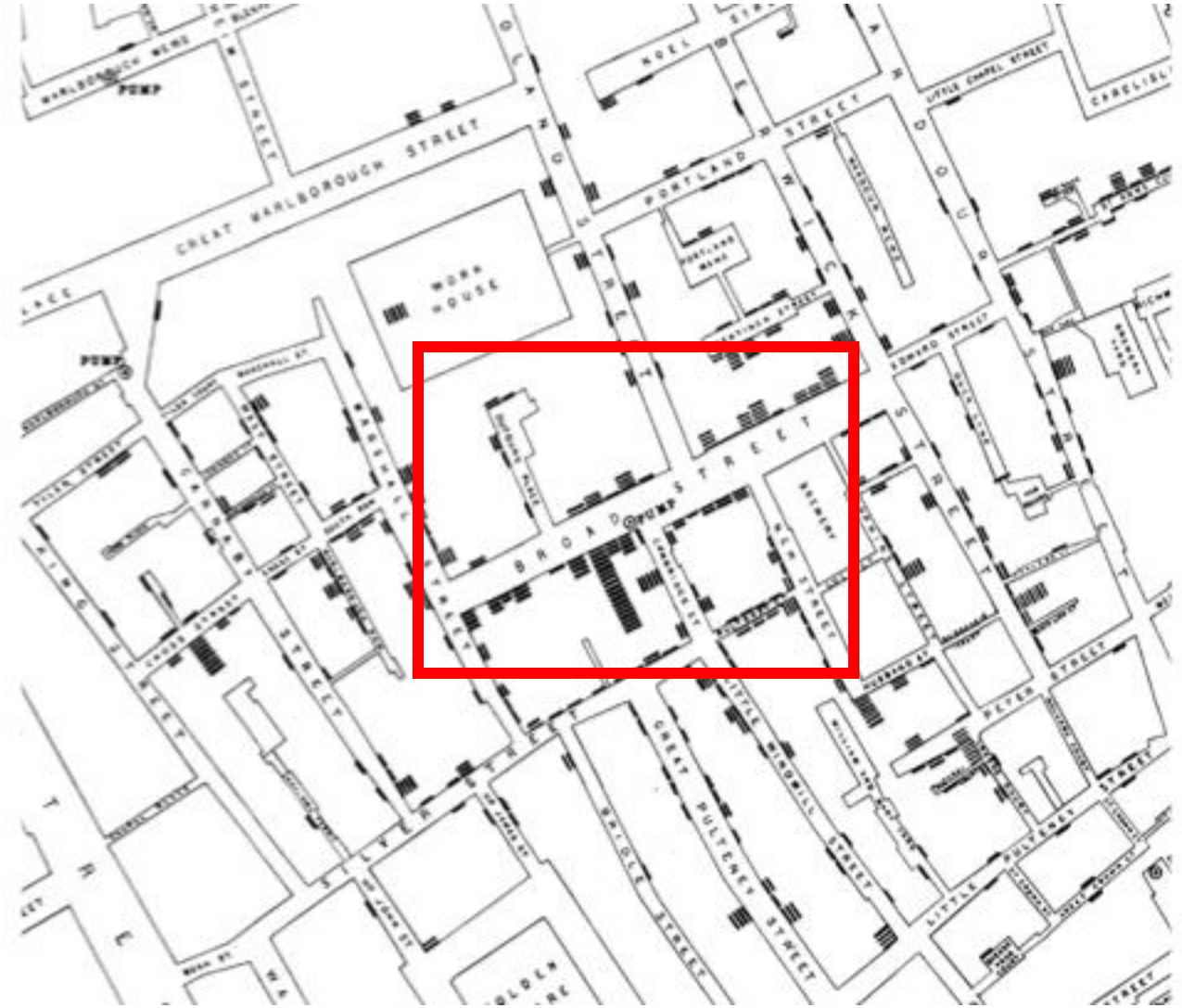
Popular miasmatists

- Florence Nightingale (founder of modern nursing)
- Edwin Chadwick (Commissioner of the Board of Health)

John Snow, 1813 - 1858

Mapping the disease

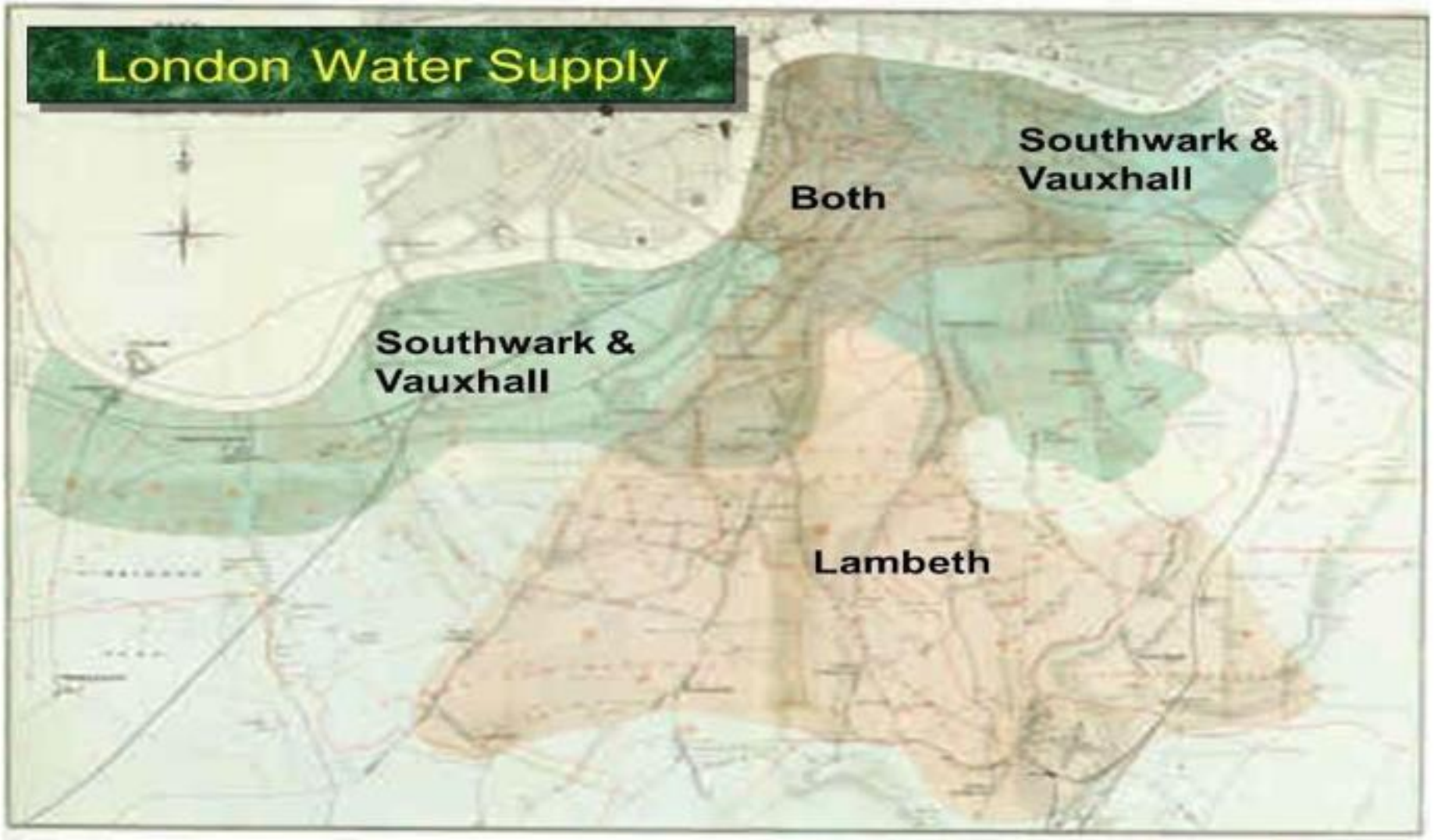
Bond Street



BR
COLLEGE

Causation

brynmawr.edu


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 ..."

Two groups different **only in the treatment**

Snow's Table

Supply Area	Number of houses	Cholera deaths	Deaths per 10,000 houses
S&V	40,046	1,263	315
Lamberth	26,107	98	37
Rest of London	256,423	1,422	59

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 Factors

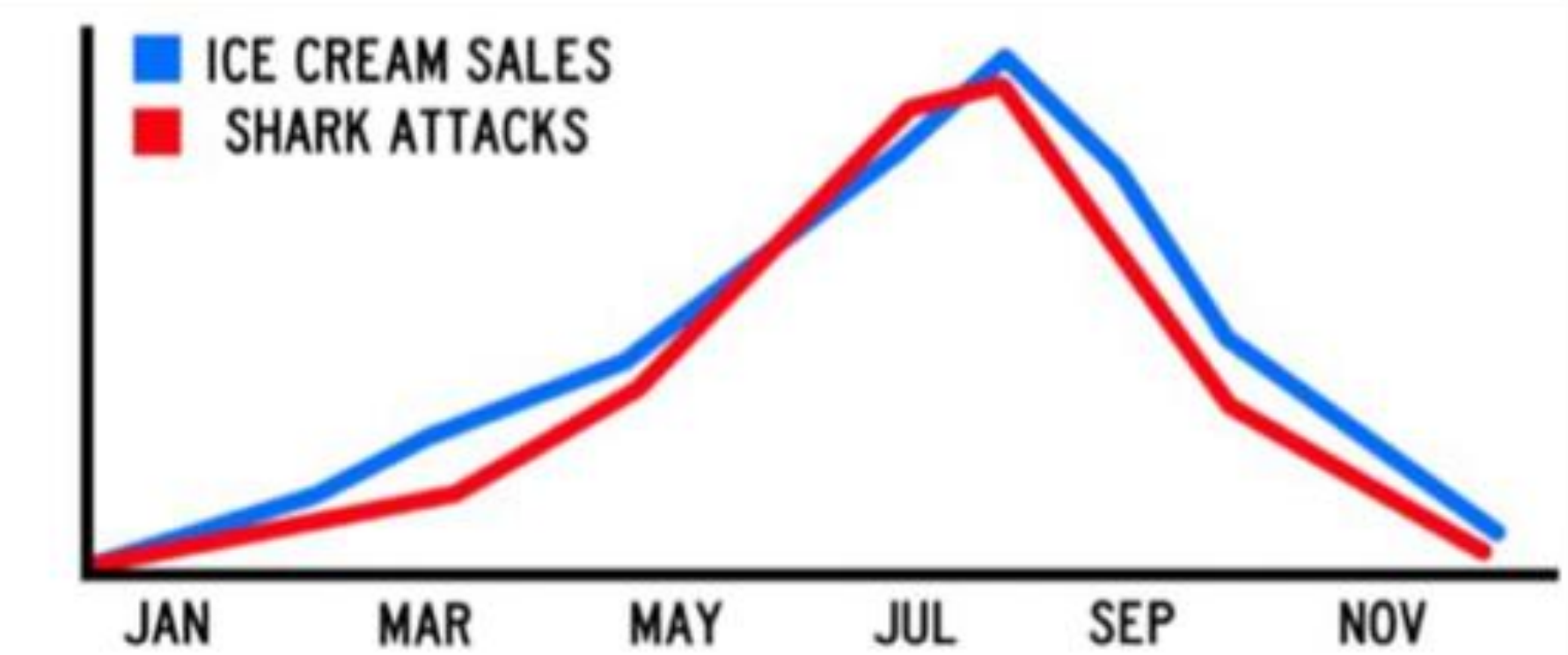
Issue

If the treatment and control groups have **systemic differences other than the treatment,** then it might be difficult to identify causality

Such differences are often present in **observational studies**

When these differences lead researchers astray, they are called **confounding factors**

Confounding Factor Example



Solution: Randomize!

If you assign individuals to treatment and control groups **at random,** the two groups are likely to be similar apart from the treatment

You can account (mathematically) for variability in the assignment

Randomized Controlled Experiment