

MODULE 1

Introduction



STAT108 - FOUNDATIONS OF DATA SCIENCE – SP23

Room: LGRT 121

Lectures: MWF 11:15AM-12:05PM

Labs: Wednesday afternoon

Instructor: Faith Zhang

TA: Amy Stephen



ABOUT YOUR INSTRUCTORS

- Faith Zhang
- Office hours
 - ❖ MoWe 12:20-1:10PM in LGRT 1344
 - ❖ Fr 2:30- 3:30PM via Zoom 8897994785

- Amy Stephen
- Office hours
 - ❖ Mon 1-2pm, Thu 10-11am
 - ❖ Zoom



DATA SCIENCE



WHAT IS DATA SCIENCE?

DS is the discipline of **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



APPLICATIONS

- Data science is driven by applications
- Data analysis is playing an increasingly important role in many fields including:
 - Biology, Chemistry, Economics, Earth Systems, Education, Environmental Science, Finance, Geography, Geology, Kinesiology, Linguistics, Management, Political Science, Public Health, Psychology, Sociology, ...
- Every data-driven subject brings new challenges



EXAMPLES

In fight against fake news, technology outsmarts humans at detecting misinformation

Researchers have demonstrated an algorithmic solution that is comparable to and sometimes better than humans at correctly identifying fake news stories.

By: IANS | New York | Published: August 22, 2018 11:23 AM

0
SHARES

f SHARE



Now DeepMind's AI can spot eye diseases just as well as your doctor

The AI from Google's DeepMind made correctly identifying fake news stories.

By MATT BURGESS
13 Aug 2018



Inside Facebook: protect the US election from foreign interference

Is it enough?
By Kurt Wagner | Aug 13, 2018

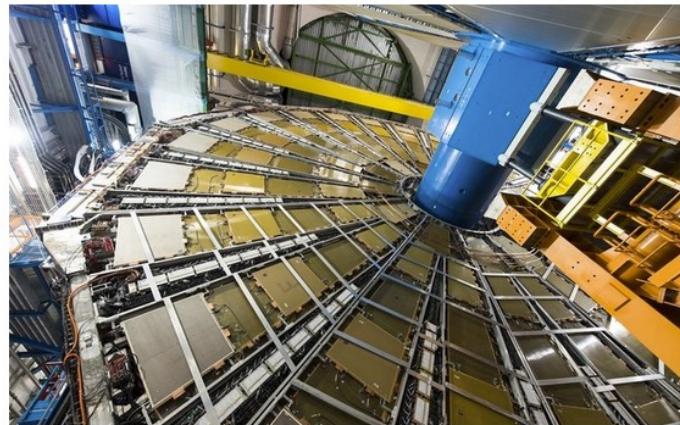


Two weeks ago, on a conference call with executives announced

LHC Physicists Embrace Brute-Force Approach to Particle Hunt

The world's most powerful particle collider has yet to turn up new physics—now some physicists are turning to a different strategy

By Davide Castelvecchi, Nature magazine on August 15, 2018



COURSE STRUCTURE



WHAT DOES THE COURSE COVER?

- An introduction to programming in Python with a focus on manipulating, visualizing, and analyzing data.
- An introduction to statistics that is grounded in computer simulations.
- An introduction to predictive modeling and machine learning.



COURSE COMPONENTS AND GRADING

- Lectures including attendance (often interactive)
- Weekly labs including attendance (pass/fail)
- Weekly graded homework assignments
- Midterm exam & final exam

| | |
|--------------|-----|
| Homework | 35% |
| Labs | 20% |
| Midterm Exam | 20% |
| Final Exam | 25% |

- If attendance drops, HW35% == HW30%+Attendance 5%



COURSE TECHNOLOGY

- **Moodle:** Links to all resource
- **Github.io:** Course website (lecture slides, demos, assignments, labs, etc.)
- **DataHub:** Web-based Python compute environment for completing labs and homework assignments.
- **Gradescope:** Assignments, labs, and exams submission system



COURSE POLICIES

- Late Homework
- Re-grades
- Academic Honesty
- <https://umass-data-science.github.io/CS108website/policies/>



COLLABORATION POLICY

Asking questions is highly encouraged

- You can discuss homework and lab questions with each other
- Do not take notes or pictures out of discussions
- The work you turn in must be your own

The Limits of collaboration

- Don't share solution material of any type with each other
- Copying solutions from any source will be dealt with under UMass' Academic Honesty procedures:

<https://www.umass.edu/honesty/>



GETTING HELP

- The course staff are here to help you be successful in the course!
- When you need help come to office hours.
- The lab sessions are also a good time to ask questions and get help.



ANNOUNCEMENTS



ANNOUNCEMENTS

- Lab 1: Today, 1:25-2:15 or 4:00-4:50.
 - ❖ Check which time is for you
 - ❖ Attendance is needed for a grade
- Syllabus is on Moodle – refer to it often
- TA: Amy Stephen, astephen@umass.edu



ON USING DATAHUB

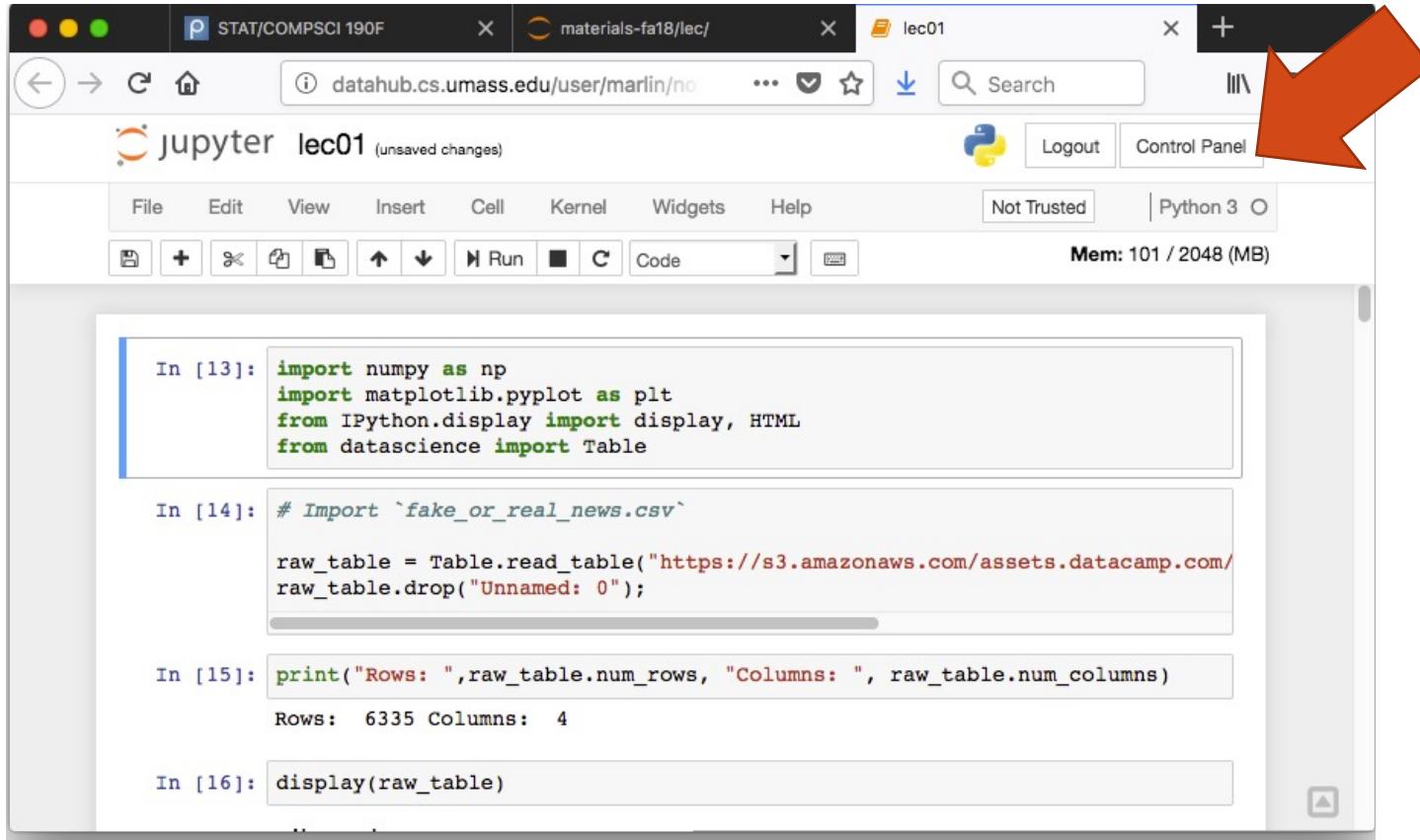


DATAHUB

- If you have an @umass.edu email address, you should now be able to access the course's data hub.
- Datahub uses UMass Google Apps authentication. Use your @umass.edu email address and Spire password to log in. It takes a minute to start up.
- When you're done working with the Datahub, make sure to shut your datahub server down and then log out.



STOPPING YOUR DATA HUB



A screenshot of a Jupyter Notebook interface running in a web browser. The browser tabs show 'STAT/COMPSCI 190F', 'materials-fa18/lec/', and 'lec01'. The main window displays a Jupyter notebook titled 'jupyter lec01 (unsaved changes)'. The notebook contains the following code:

```
In [13]: import numpy as np
import matplotlib.pyplot as plt
from IPython.display import display, HTML
from datascience import Table

In [14]: # Import `fake_or_real_news.csv`
raw_table = Table.read_table("https://s3.amazonaws.com/assets.datacamp.com/
raw_table.drop("Unnamed: 0");

In [15]: print("Rows: ", raw_table.num_rows, "Columns: ", raw_table.num_columns)
Rows: 6335 Columns: 4

In [16]: display(raw_table)
```

The interface includes a toolbar with file operations like File, Edit, View, Insert, Cell, Kernel, Widgets, Help, and a status bar showing 'Mem: 101 / 2048 (MB)'. A large red arrow points to the 'Control Panel' button in the top right corner of the notebook header.



STOPPING YOUR DATA HUB

A screenshot of a web browser window displaying a Jupyter Notebook interface. The browser tabs show the URL `datahub.cs.umass.edu/user/marlin/no`. The notebook title is `lec01`. A confirmation dialog box is overlaid on the page, asking "This page is asking you to confirm that you want to leave - data you have entered may not be saved." It contains two buttons: "Stay on Page" and "Leave Page". A large red arrow points to the "Leave Page" button. The notebook code cells show imports and data loading from a CSV file, followed by a print statement and a display command.

```
In [13]: # Import `fake_or_real_news.csv`

raw_table = Table.read_table("https://s3.amazonaws.com/assets.datacamp.com/
raw_table.drop("Unnamed: 0");

In [14]: # Import `fake_or_real_news.csv`

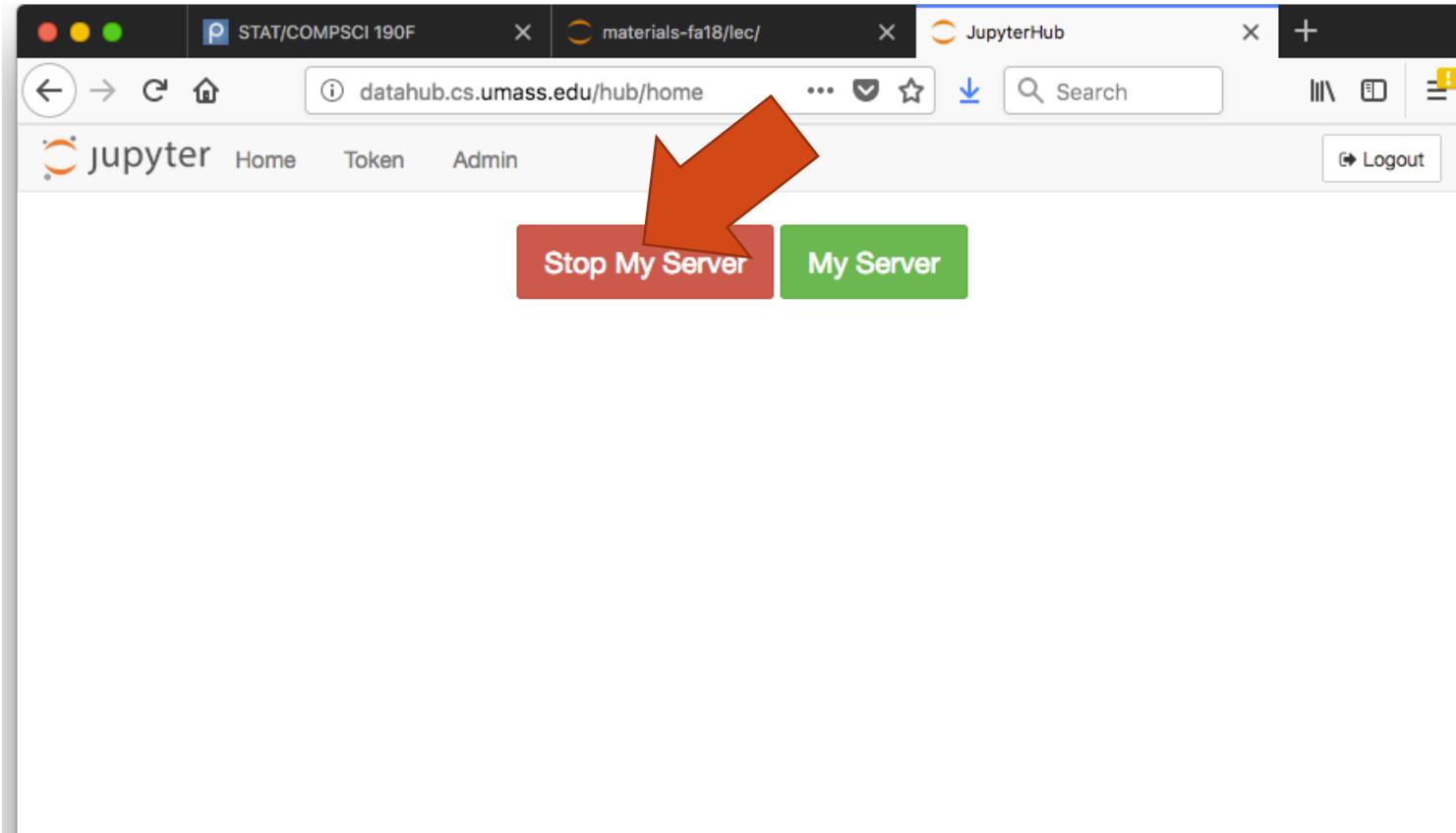
raw_table = Table.read_table("https://s3.amazonaws.com/assets.datacamp.com/
raw_table.drop("Unnamed: 0");

In [15]: print("Rows: ", raw_table.num_rows, "Columns: ", raw_table.num_columns)

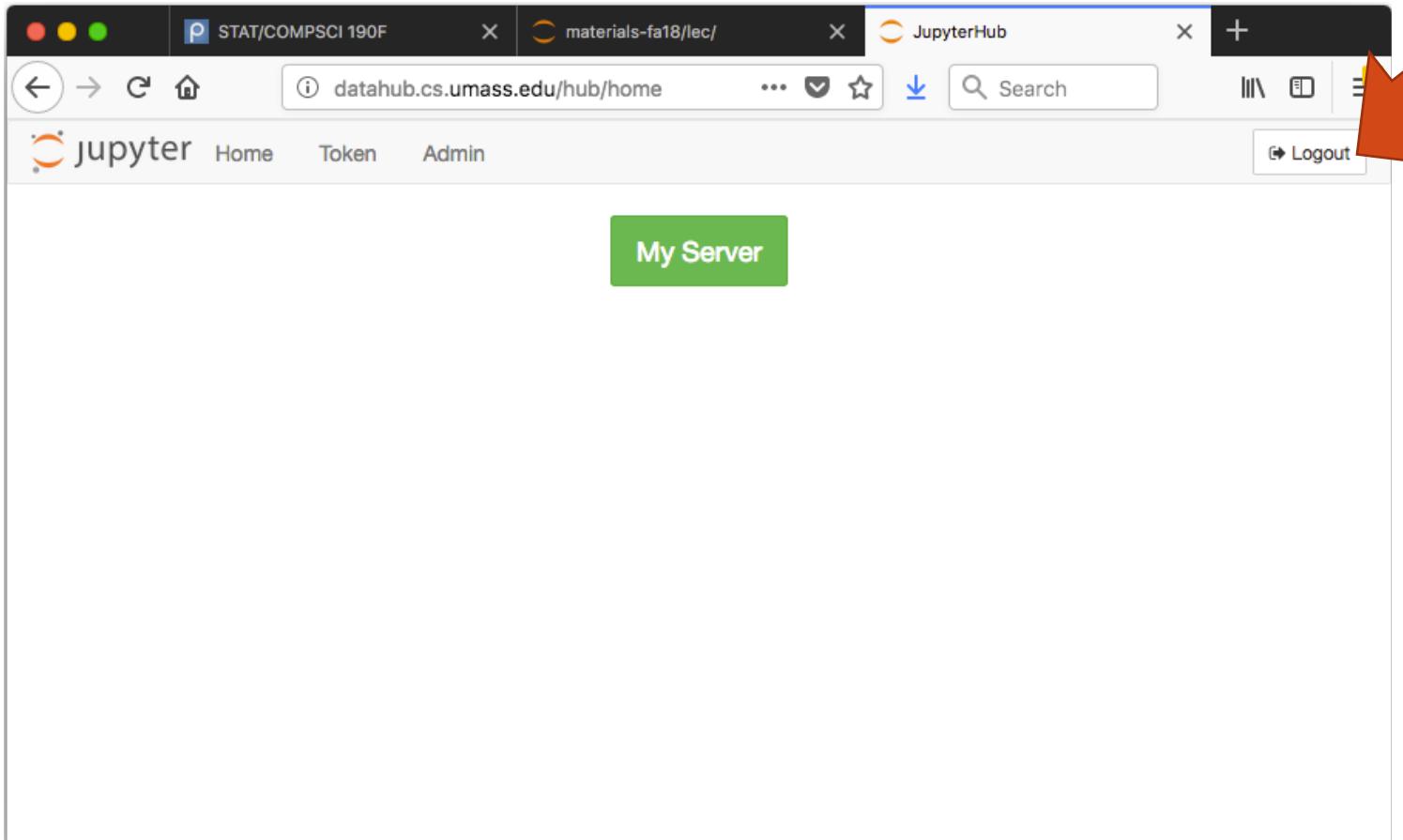
Rows: 6335 Columns: 4

In [16]: display(raw_table)
```

STOPPING YOUR DATA HUB



STOPPING YOUR DATA HUB



PYTHON SYNTAX



CAUSE AND EFFECT



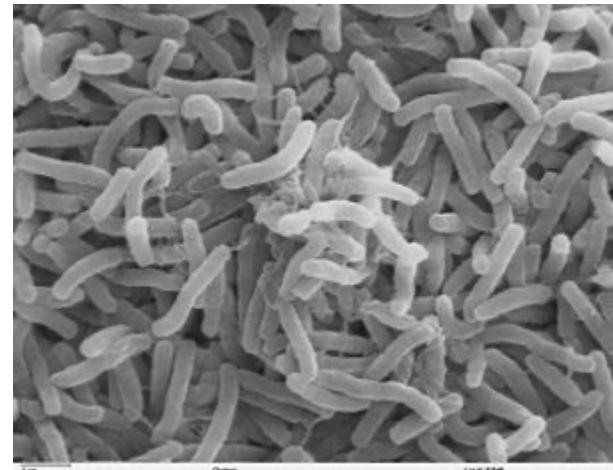
CAUSALITY AND EXPERIMENTS

- Examples of DS questions:
 - Does the death penalty have a deterrent effect?
 - Is chocolate good for you?
 - What causes breast cancer?
- These questions seek to **identify an association between two things**, and for some, **evaluate if that association is causal**
- To identify these associations and evaluate causality, data scientists and researchers conduct experiments.



BROAD STREET CHOLERA OUTBREAK

- The Broad Street cholera outbreak was a severe outbreak of cholera that occurred in 1854 near Broad Street in the Soho district of London, England.
- This outbreak killed 616 people.

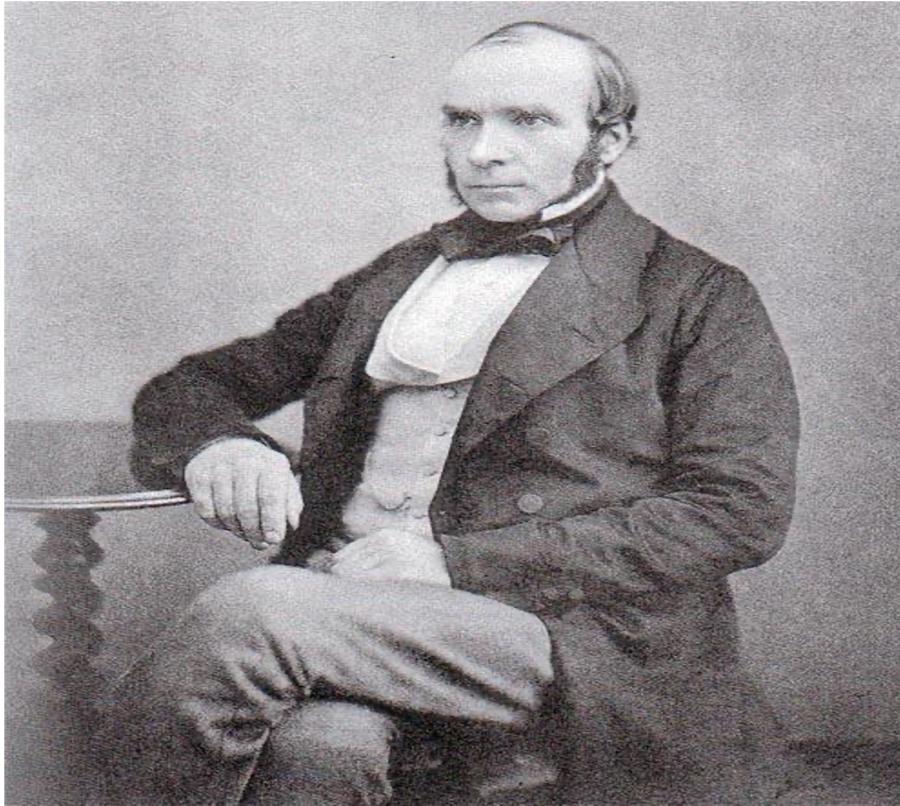


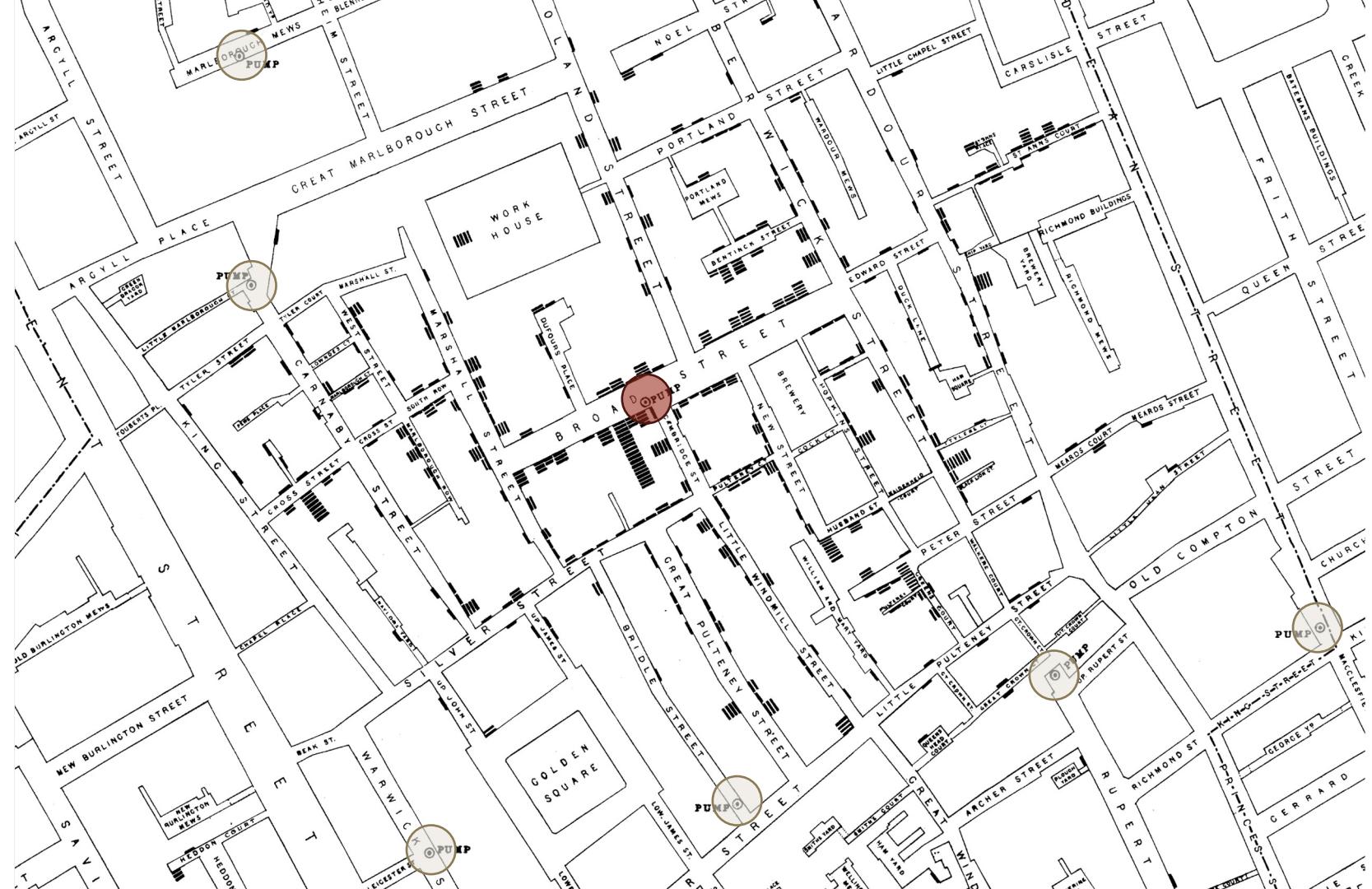
TWO THEORIES OF CHOLERA

- **Miasma theory:** cholera was caused by **particles in the air**, or "miasmata", which arose from decomposing matter or other dirty organic sources.
- **Germ theory:** the principal cause of cholera was a germ cell that had not yet been identified but was **transmitted through food or drink**.



JOHN SNOW, 1813-1858





London Water Supply



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 |



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.*



COMPARISON

- **Treatment group:** Do receive the treatment
- **Control group:** Do not receive the treatment



ASSOCIATION VS CAUSATION



CHOCOLATE AND HEART DISEASE: STUDY

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

June 19, 2015 · 5:03 AM ET

Heard on *Morning Edition*



- **Population** (individuals, study subjects, participants, units): 20K *European adults followed for 12 years.*
- **Treatment:** *chocolate consumption*
- **Outcome:** *heart disease*



CHOCOLATE AND HEART DISEASE: ASSOCIATION

Question 1: Is there **any association** (any relationship) between chocolate consumption and heart disease?

- **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.”
- **Answer:** Yes, this points to an **association**



CHOCOLATE AND HEART DISEASE: CAUSATION

Question 2: Does chocolate consumption lead to a reduction in heart disease?

- **This question asks about 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*



CHOCOLATE AND HEART DISEASE: ALTERNATIVES

Question 3: Is the fact that people ate more chocolate the only possible cause for the observed effect of decreased heart disease risk?

- For example, suppose the people who ate more chocolate tended to live in European countries with better health care?
- What if wealthier people eat more chocolate and can also afford better health care?



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.



RCEs VS OBSERVATIONAL STUDIES

- Observation is a key to good science.
- An *observational study* is one in which scientists make conclusions based on data that they have observed but had no hand in generating.
- *randomized controlled experiment*, RCE, is an experiment where we are able to randomize individuals into the treatment and control groups.
 - also known as a *randomized controlled trial* (RCT)



CONFOUNDING

- 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.



EXAMPLES

No level of alcohol consumption is healthy, scientists say



Fo:

Dairy and meat 'beneficial for heart health and longevity'



Medic

Eating cheese may be associated with a lower risk of death – and scientists are zeroing in on why it might be a better choice than milk



Business Insider • today



RANDOMIZATION AND CONFOUNDING

- 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 Experiments are the gold standard for establishing cause and effect.**



RCEs VS OBSERVATIONAL STUDIES

- **Question:** If randomized controlled experiments can establish causality while observational studies are subject to confounding, why are so many studies observational?

