Data science for everyone

Prof. Jones-Rooy & Prof. Policastro

Feb. 3, 2020

2.1: Causality

ANNOUNCEMENTS

1. Homework 1 out today, Feb. 3; due Tue., Feb. 18
2. Lab 0 out Wed., Feb. 5; due Wed., Feb. 12
   - Content: practice formatting and submitting assignments for this semester
   - Also: counts for *Academic Engagement*
3. Wed., Feb. 5, Lecture 2.2: Everything you need to know to submit assignments
   - Sections Feb. 6/7: Practice, Lab 0
4. Data Science major/minor open house, Feb. 20, 12p, CDS

JupyterHub

https://dsua-111.rcnyu.org

Where to find all assignments for this course, including Homework 1 tonight!

jupyter

Logout | Control Panel

Files | Running | Clusters | Assignments

Released, downloaded, and submitted assignments for course: dsua-111

Released assignments

Homework1	dsua-111	Fetch

Downloaded assignments

There are no downloaded assignments.

Submitted assignments

There are no submitted assignments.

jupyter

Logout | Control Panel

Files | Running | Clusters | Assignments

Released, downloaded, and submitted assignments for course: dsua-111

Released assignments

There are no assignments to fetch.

Downloaded assignments

Homework1 ▸ dsua-111 Submit

Submitted assignments

There are no submitted assignments.

jupyter

Logout | Control Panel

Files | Running | Clusters | Assignments

Released, downloaded, and submitted assignments for course: dsua-111

Released assignments

There are no assignments to fetch.

Downloaded assignments

Homework1 dsua-111 Submit

Homework1 Validate

Submitted assignments

There are no submitted assignments.

DS-UA 111: Homework One

This homework is due Tuesday, Feb. 18 by 8:00p. Late homework will be graded down, no exceptions. Improperly formatted homeworks also count as late. Note that the course academic honesty policy applies to every homework, including this one. Some of the questions refer to articles, which you can find by clicking on the links provided. This homework is worth 41 points (one point per sub-question).

Instructions

jupyter

Logout | Control Panel

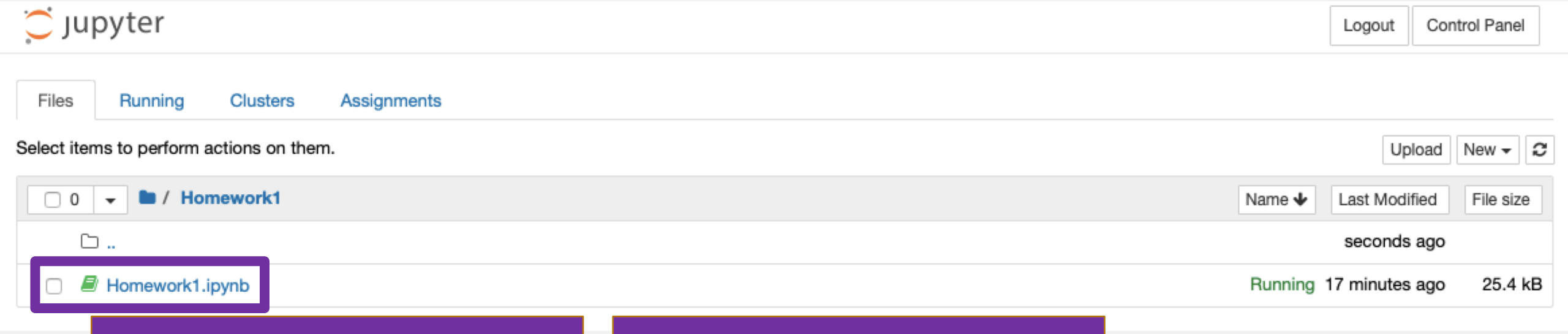
Files | Running | Clusters | Assignments

Select items to perform actions on them.

Upload | New | ↻

0 / | Name ↓ | Last Modified | File size

class_materials	an hour ago	
Homework1	a minute ago	
lost+found	3 hours ago	
shared	10 days ago	
your_materials	3 hours ago	



Recommendation: Make a copy if you want to have "scratch paper" to try things out

But: Ultimately you'll need to submit the original assignment

jupyter Homework1 (autosaved)

Logout | Control Panel

File Edit View Insert Cell Kernel Widgets Help | Not Trusted | Python 3

Run | Markdown | Validate

DS-UA 111: Homework One

This homework is due Tuesday, Feb. 18 by 8:00p. Late homework will be graded down, no exceptions. Improperly formatted homeworks also count as late. Note that the course academic honesty policy applies to every homework, including this one. Some of the questions refer to articles, which you can find by clicking on the links provided. This homework is worth 41 points (one point per sub-question).

Instructions

In Lecture 2.2 we will go over how to **submit** this homework
You will be able to practice this in sections & Lab 0

Outline

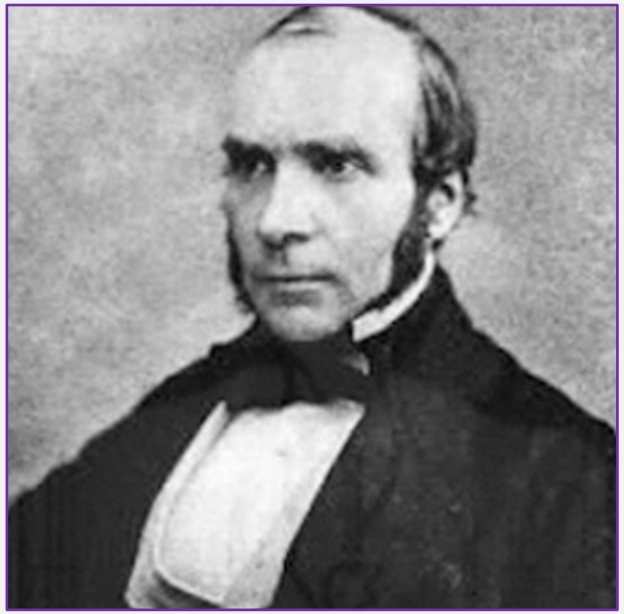
1. John Snow & thinking like a scientist
2. John Snow & the search for causality
3. Key concepts in causality

John Snow and Thinking Like a Scientist

John Snow and the Search for Causality

JOHN SNOW & THE BROAD ST. PUMP

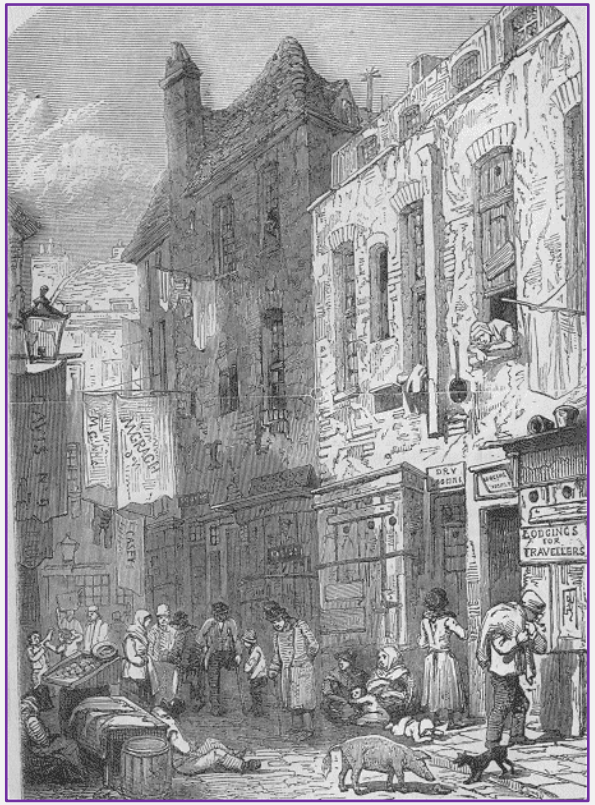
Inferential Thinking, Ch. 2



- 1850s London
- Waves of cholera killing 10K people each
- Dr. Snow is treating patients during a big outbreak in 1854
- He wants to know:

What is causing this cholera outbreak?

(Scientific method pre-step!)



SNOW THINKS LIKE A SCIENTIST

SCIENTIFIC METHOD

DATA

1. Observation
2. Question
3. Theory
4. Hypothesis

DATA

5. Test
6. Update theory

DATA

7. Repeat as desired

Snow's initial observations

1. People die within days of contracting
2. Often people within one house would all die
3. But neighbors would not necessarily become infected
4. Symptoms: digestive problems

Snow's question

Why do some people become infected with cholera and not others?

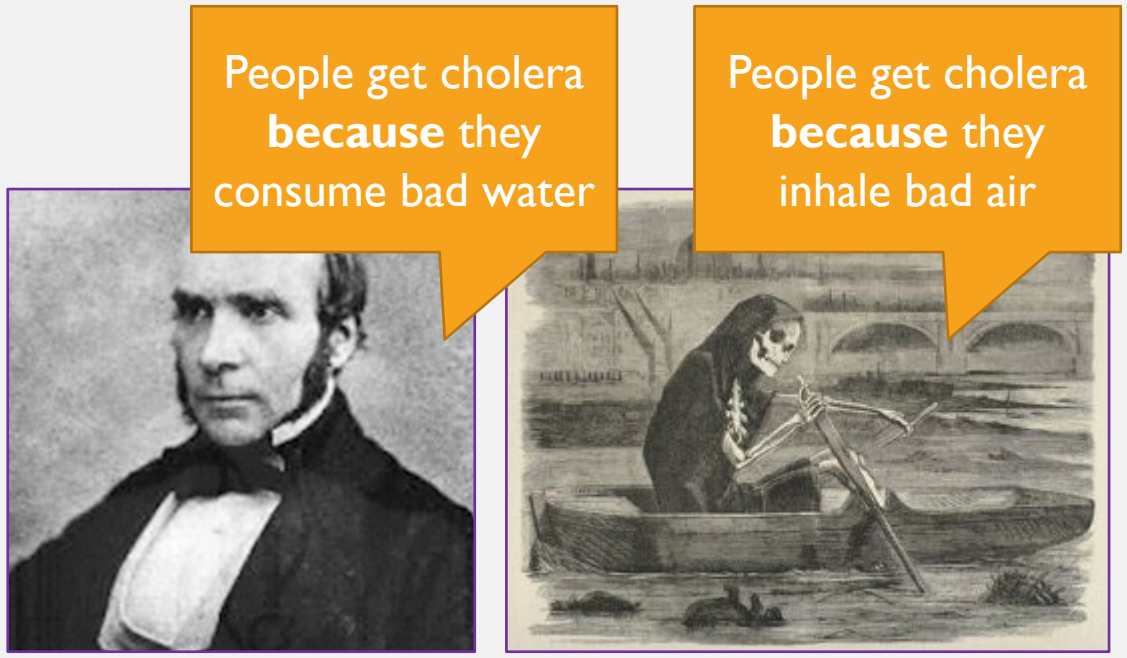
Note: He could have asked something else (e.g., why do people die from cholera?)

FORMING A THEORY

- Snow suspects the Broad Street Pump might be to blame
  - Drink the water → get infected
- This is counter to the prevailing theory
  - *Miasma* – it's something in the air
- How can Snow find out which is the better theory?
  - What do we mean by "better"?
  - Note: **Not** how can he find out which one is right?

Theory

A conjecture about the causes of some phenomenon or outcome of interest



Read more on what makes a good theory here!

WHAT MAKES A GOOD THEORY?

Observable implications

Predicts something about the world that we can check with evidence

Falsifiable

We can imagine evidence that would be incompatible with our theory being true

Useful

It explains (a lot of) the variation we see in the outcome we care about

How much is a lot? Stay tuned!
(Hint: it involves art & science)

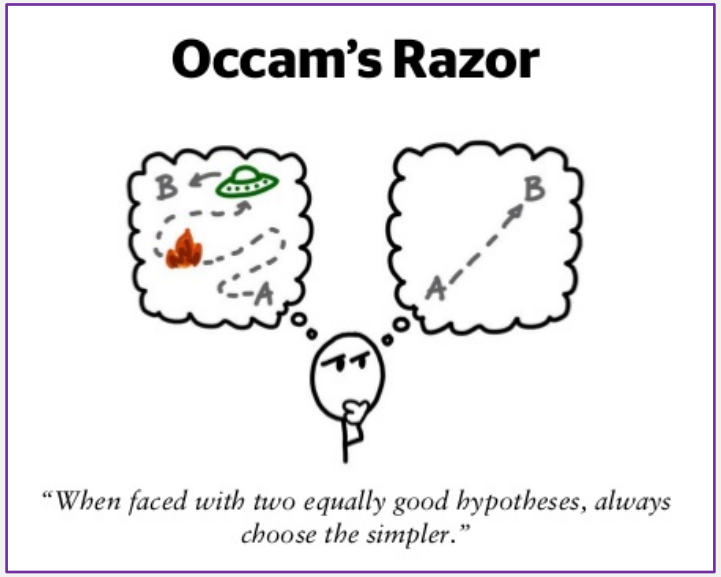
Parsimonious

We prefer a theory that explains equivalently with fewer moving parts

Quick read on Occam's razor and machine learning here!

EVALUATING THEORIES

- A theory cannot be proven correct, but it can be "better" than another theory if it explains more variation in the outcome of interest (e.g., in infection rates)
  - Or if it explains equivalently but with fewer moving parts (parsimonious)
- Over time, as we gather evidence compatible with a theory, it can be shown to be "less wrong" than others
- All theories are wrong, but some are useful
  - This will apply to statistical models as well later in the term
- How can we disprove or falsify a theory? Test it!



TESTING THEORIES

1. Think about what **observable implications** follow from the theory

2. Generate **testable hypotheses** that we can check with evidence

- Snow's theory:
  - Drinking water from the pump is causing people to become infected with cholera
- Observable implications of this theory:
  - People who drink from the pump should be more likely to become infected than people who do not drink from the pump
- Testable hypothesis:
  - Hypothesis (H1 or H_A): Drinking from the pump is associated with higher rates of infection
  - Null hypothesis (H_0): Drinking from the pump is associated with no difference in rates of infection

SNOW THINKS LIKE A SCIENTIST

SCIENTIFIC METHOD

DATA

1. Observation
2. Question
3. Theory
4. Hypothesis

DATA

5. Test
6. Update theory

DATA

7. Repeat as desired

Snow's falsifiable theory

Drinking from the Broad Street Pump causes people to become infected with cholera

Snow's testable hypothesis

People who drink from the Broad Street Pump are more likely to get cholera than those who don't

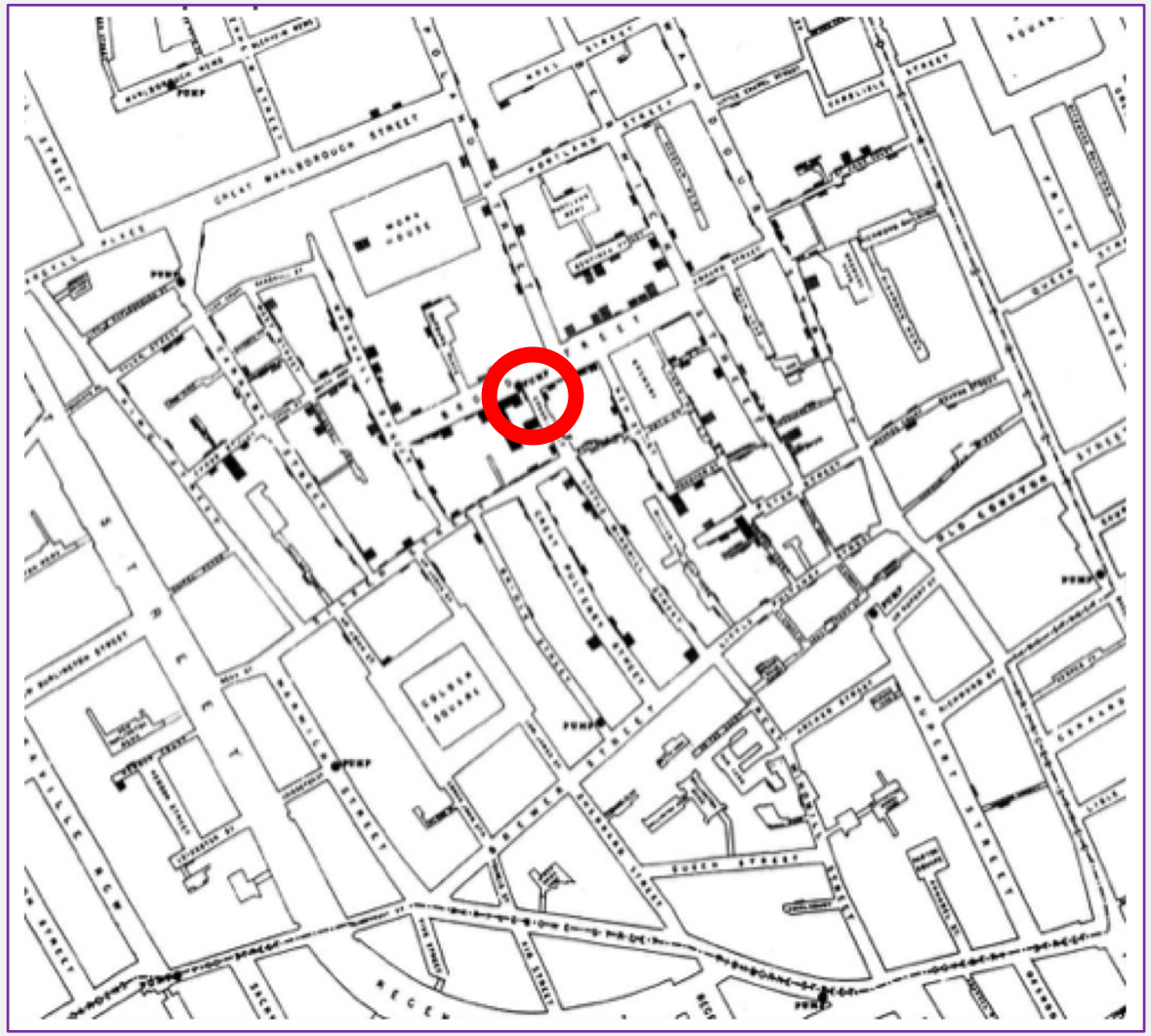
Snow's test

Collect data on cholera infections relative to the pump and see if the patterns follow the observable implication of this theory

TESTING SNOW'S HYPOTHESIS

Snow records the locations of patients on a map

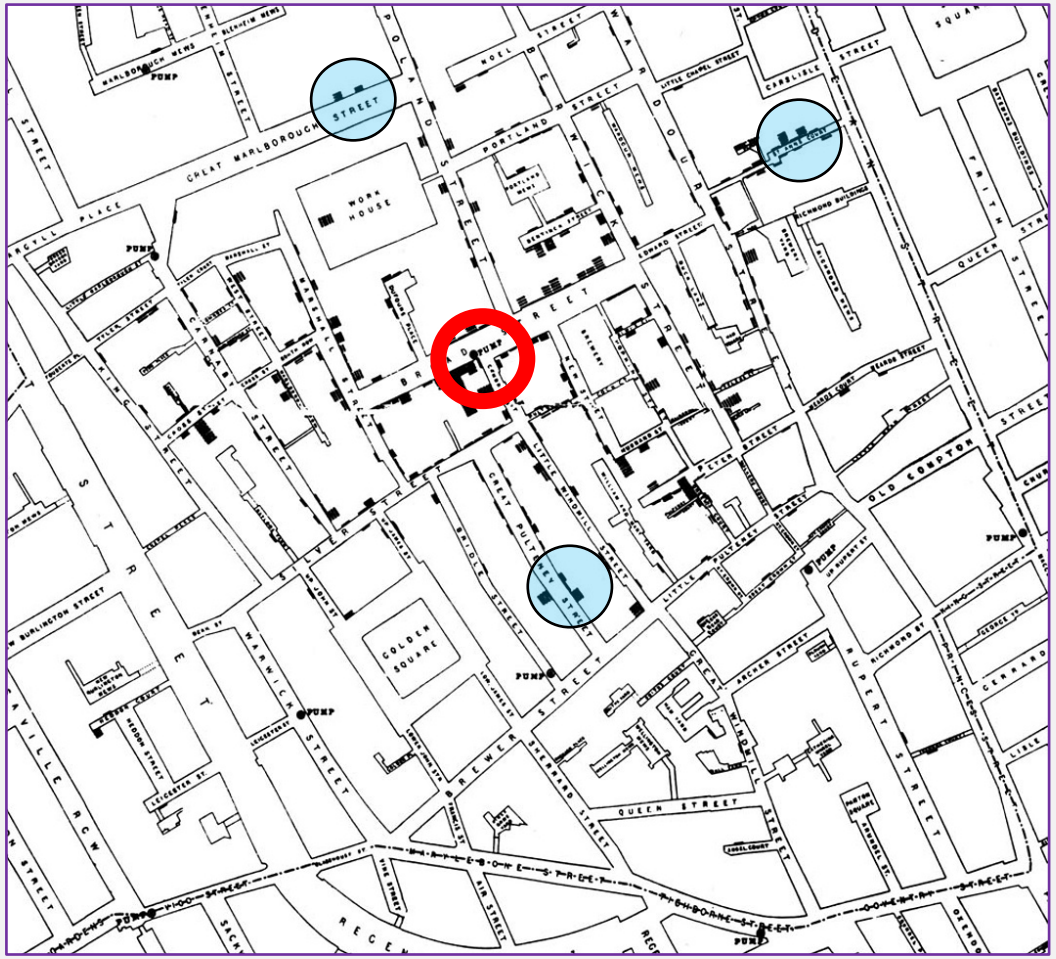
Indeed, outbreaks seem to be concentrated around the Broad Street Pump!



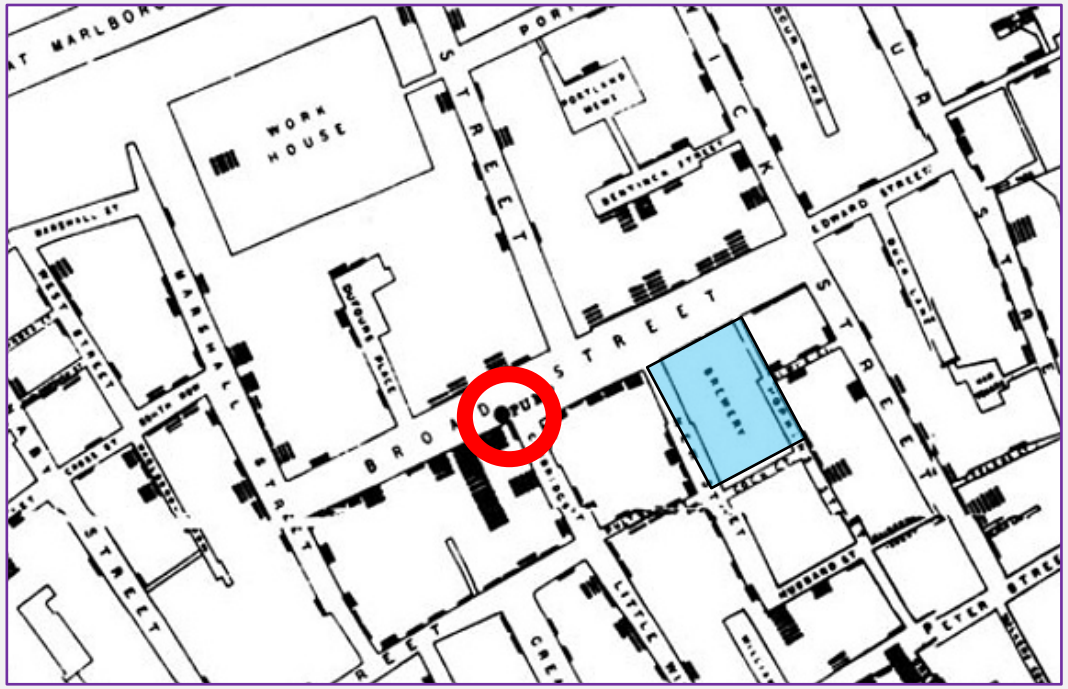
Does this mean it's definitely the pump?

Not so fast…

There are a whole bunch of infections far from the pump!



…and no infections at the brewery!



Snow's testable hypothesis

People who drink from the Broad Street Pump are more likely to be infected with cholera than those who don't

EVALUATING SNOW'S HYPOTHESIS

- Generally speaking, cholera infections do take place near the Broad Street Pump
- But there are some exceptions to this pattern
- Do we reject Snow's hypothesis and conclude his theory been falsified?
- Not necessarily: the original theory is about drinking from the pump. Geography is a **proxy**
- Let's look deeper

- Houses close to the pump were more likely to be infected
- Houses further from the pump, but less convenient to other pumps, more likely to be infected
- Near the pump, workers from the brewery were not infected because they had their own well
- Other scattered deaths were of children who walked past the pump on the way to school
- Woman far away who died had pump water delivered!

SNOW THINKS LIKE A SCIENTIST

SCIENTIFIC METHOD

DATA

1. Observation
2. Question
3. Theory
4. Hypothesis

DATA

5. Test

6. Update theory

DATA

7. Repeat as desired

Snow's testable hypothesis

People who drink from the Broad Street Pump are more likely to get cholera than those who don't

Update theory

We cannot rule out the Broad Street Pump, but we can still go deeper to better understand **why**

Notice this language: We haven't "proven" anything; we have only failed to rule it out, or disprove

Repeat & extend

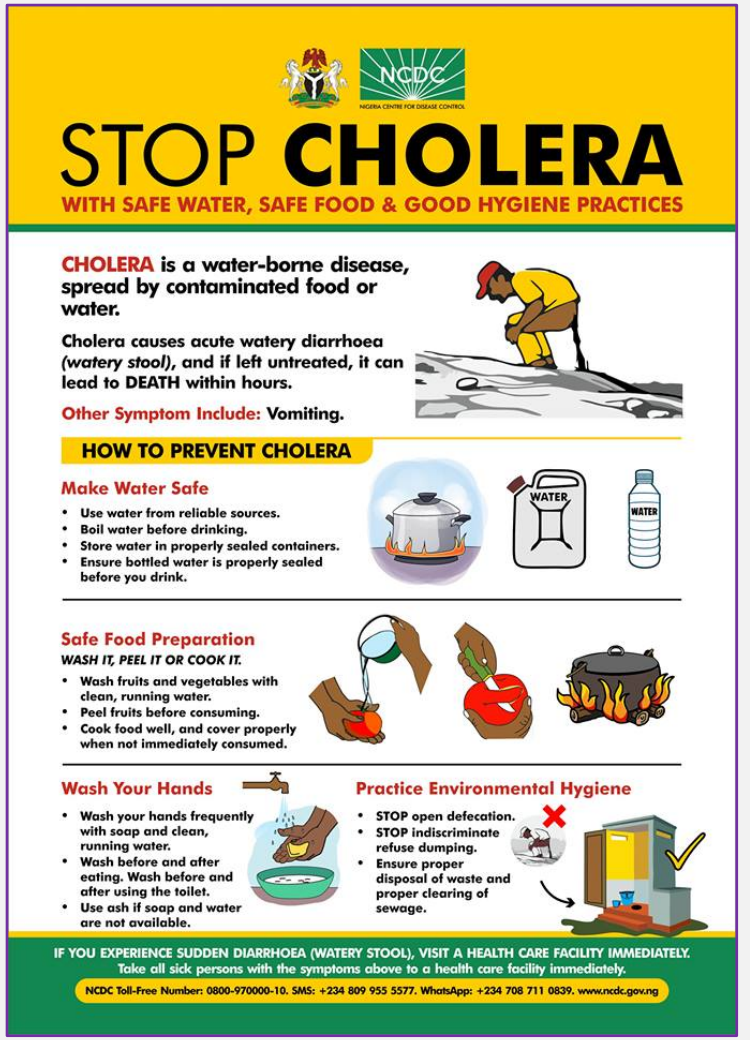
We still haven't established causality. Need more rigor!

Outline

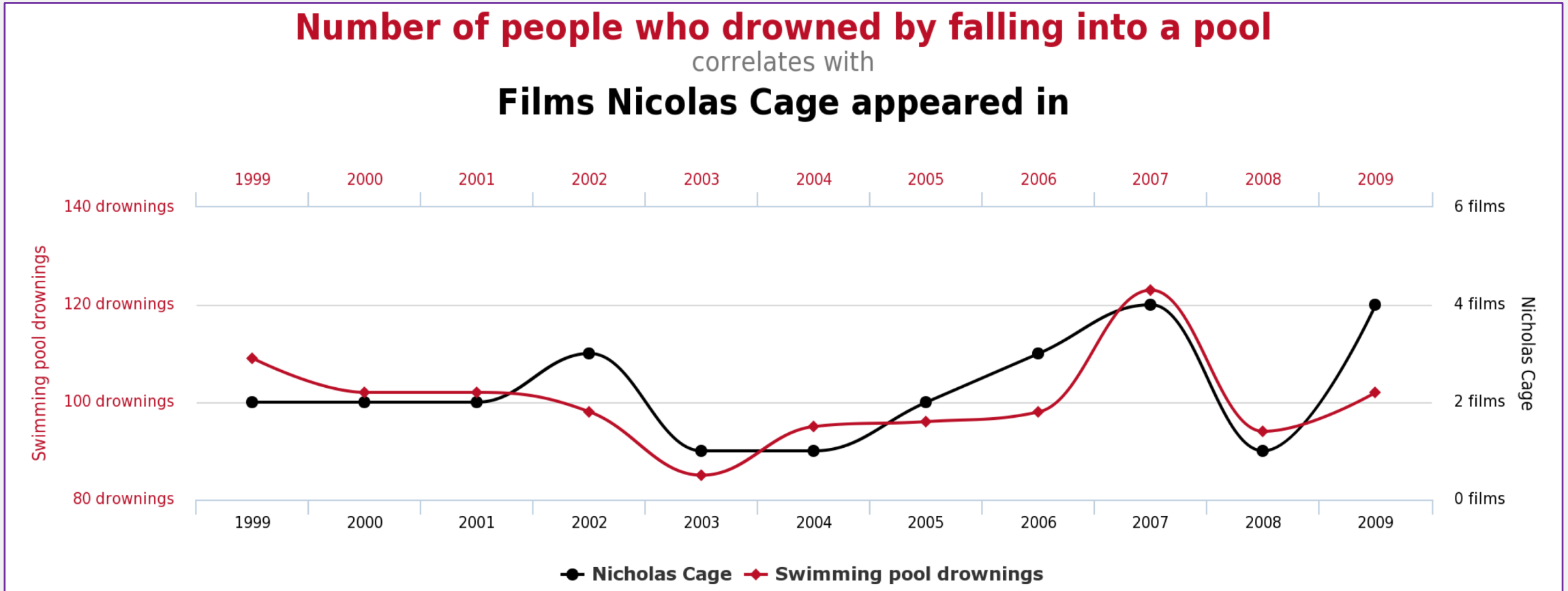
1. John Snow & thinking like a scientist
2. John Snow & the search for causality
3. Key concepts in causality

ASSOCIATION VS. CAUSATION

- Snow has reason to believe the pump water is making people sick – there's an association between drinking from the pump and cholera
- But, he wants to know **why**
  - This is what epidemiology does: works out why disease spreads and how to contain it.
- Snow's can make his theory more specific – what about the pump is making people sick?
  - He thinks the water supplied to pump was contaminated by sewage
- How can he draw firmer conclusions? The problem is that finding associations is easy, but showing causality is hard



ESTABLISHING ASSOCIATION IS EASY



Theory: Nic Cage films cause people to fatally fall into pools

H1: Rate of Cage films is associated with rate of pool-falling-deaths

Test: We see association!

We cannot rule out that Cage causes fatal pool-falling

But we have not (yet) established that Cage films **cause** this, or **why** they do

Read more about counterfactuals here and here!

ESTABLISHING CAUSALITY IS HARD

It relies on counterfactual reasoning, which means:

The causal effect of some factor X is the difference between what **actually** happened and what **would have happened** if X had been different in some way

The causal effect of treatment X on an outcome for a specific case at a specific time is the difference between the actual outcome and the hypothetical outcome that would have occurred in the same case, at the same time, had X not been present

Counterfactual =
"Counter" to the "Facts"

Imagining what would have happened if something other than what happened, happened

"What if" scenarios

What if 9/11 never happened?
What if you didn't come to NYU?
What was the effect of Obama on the US economy?

EXPERIMENTS CAN HELP!

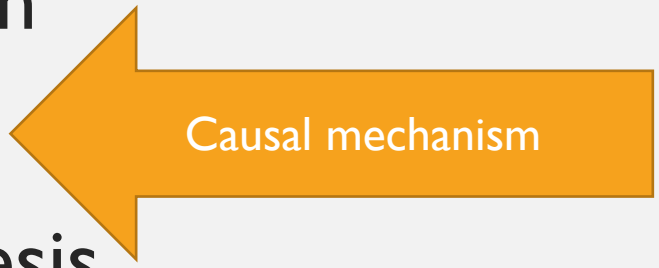
treatment

control

1. **Treatment and control group with a placebo**
   - Treatment: people receiving X, the treatment, water, drug, etc. 
   - Control: people not receiving X (but think they do)
   - Placebo: So it's not obvious who is in the control group 
2. **Random assignment of treatment/control to subjects**
   - To avoid problems like selection ← more on this to come
3. **Double-blind design**
   - Neither participants nor scientists know who is in treatment vs. control
   - Participants may behave differently if they know, scientists may behave, record, or infer differently if they know

SNOW'S EXPERIMENT

1. Observation
2. Question
3. Theory
4. Hypothesis
5. Test
6. Update theory
7. Repeat as desired



Snow's revised theory:

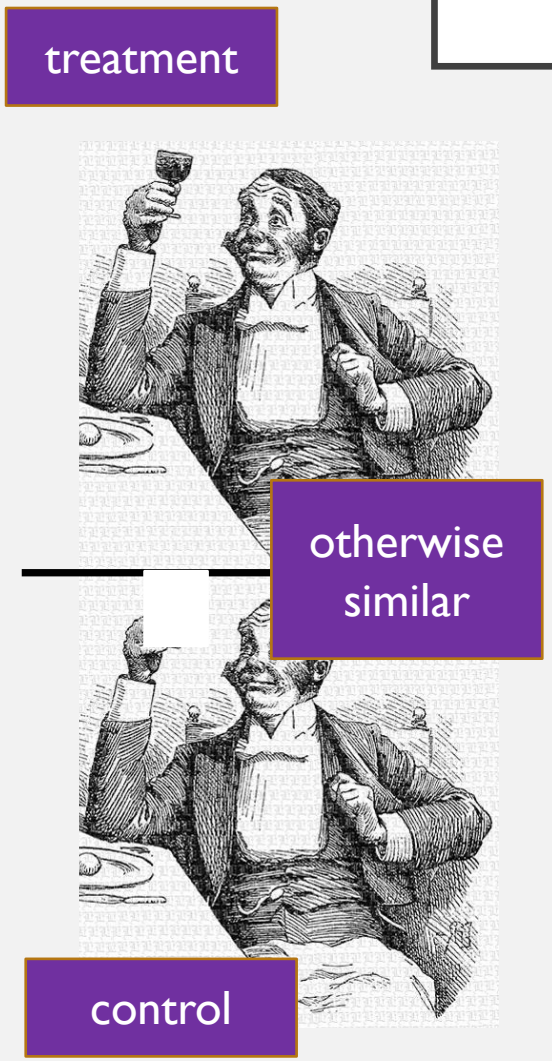
The pump is causing people to become infected **because** the water supplied to pump was contaminated by sewage

Causal mechanism: How and why a hypothesized cause, in a given context, contributes to a particular outcome

ESTABLISHING CAUSALITY IS HARD FOR SNOW!

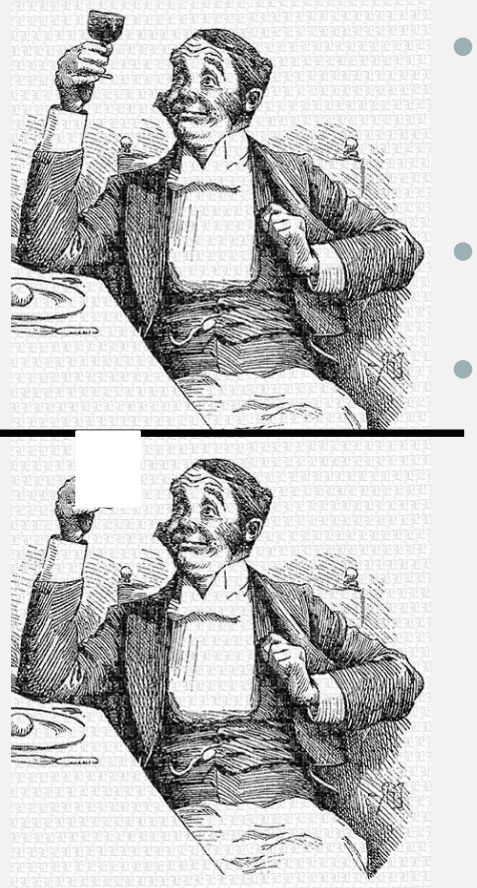
- The **causal** effect of drinking 'bad' water on the infection status for a particular person on a given day is:
  - The difference between their infection status after drinking the water
  - And the infection status they would have had on the same day had they not drunk 'bad' water
- Fundamental problem of causal inference: we never get to see both scenarios for the same unit (person) at the same time
  - This means we can never know the causal effect with certainty!
- What we can do is find **good** comparison cases and assign them to treatment and control groups
  - What makes for good comparison cases?

RANDOM ASSIGNMENT



- We prefer that subjects **not select into** (choose) treatment or control: Why?
  - What types of people do you think would volunteer to try (or not) a new drug?
  - Would we expect different outcomes for them?
- We can achieve this by **randomizing** subjects to treatment and control
  - This makes treatment and control groups **similar**, on average, with respect to everything except the treatment of interest

CONTROL GROUPS ARE KEY



- Compare the treatment group to the control group in terms of their outcomes (net of placebo effect)
- Without a control group, you cannot make inferences
- Snow needed at least some people who did not drink the pump water (why?)
  - If we only observe people who drink the pump water and they all die, we still don't know what would have happened if they didn't
  - Example: All humans who die all drank water during their lives

CAN JOHN SNOW RANDOMIZE?

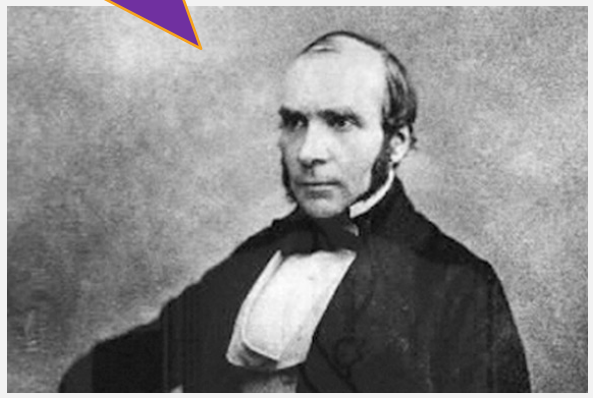
- Are locals randomized into treatment and control?
  - No, it appears proximity or access to the pump matters
- Can he randomize them himself? I.e., can he do a randomized control trial (in lab, or "field experiment")?
  - No, and there are many issues with this:
  - **Practical**: it's hard to get people to drink from different pumps
  - **Compliance**: People prefer to drink from the pump nearest them so may not follow instructions
  - **Ethical**: We suspect the pump is deadly. We can't assign half the neighborhood to drink from it!
  - A (lesser) consideration: It would be a single-blind study not double-blind unless Snow got another researcher to conduct it.
    - Even then it would only be double-blind if residents don't know about the Pump theory or where the water they are drinking is from

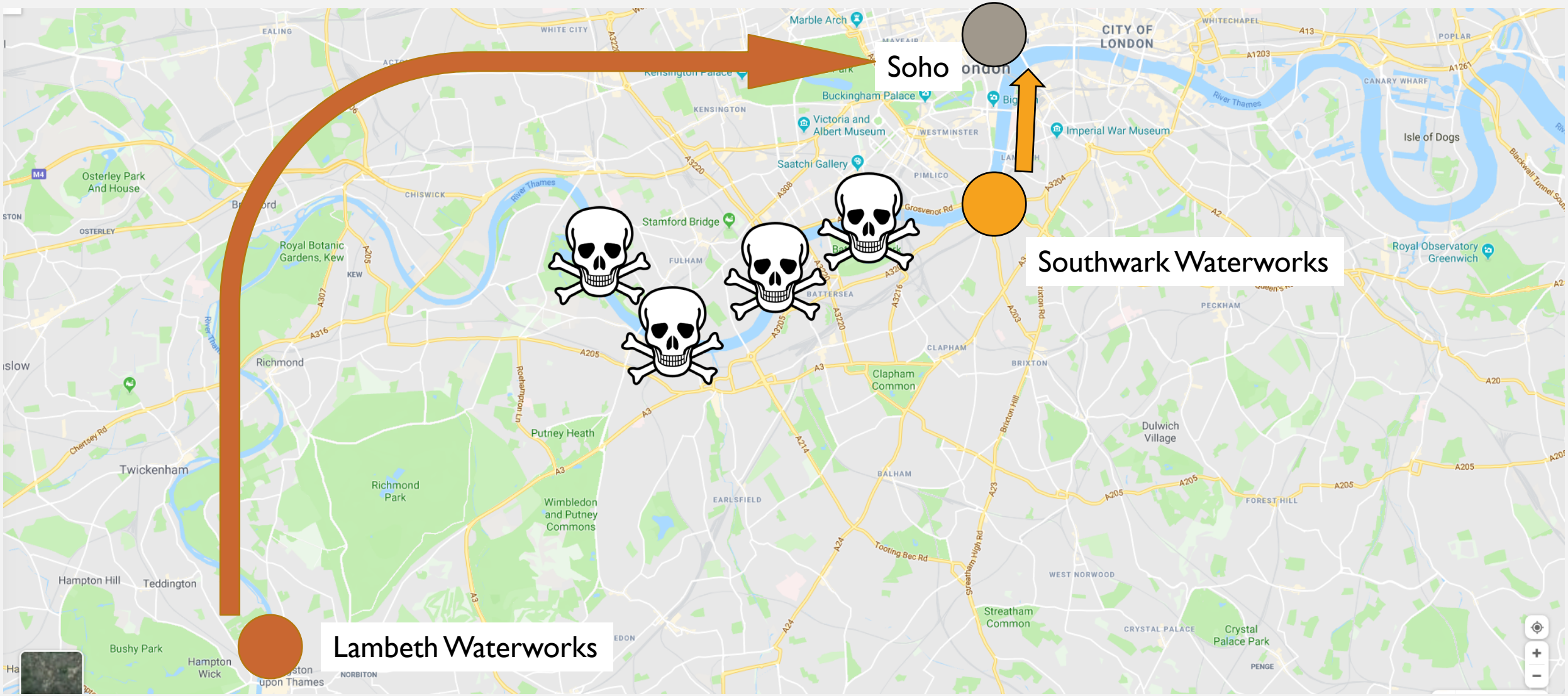
WHAT CAN SNOW DO?

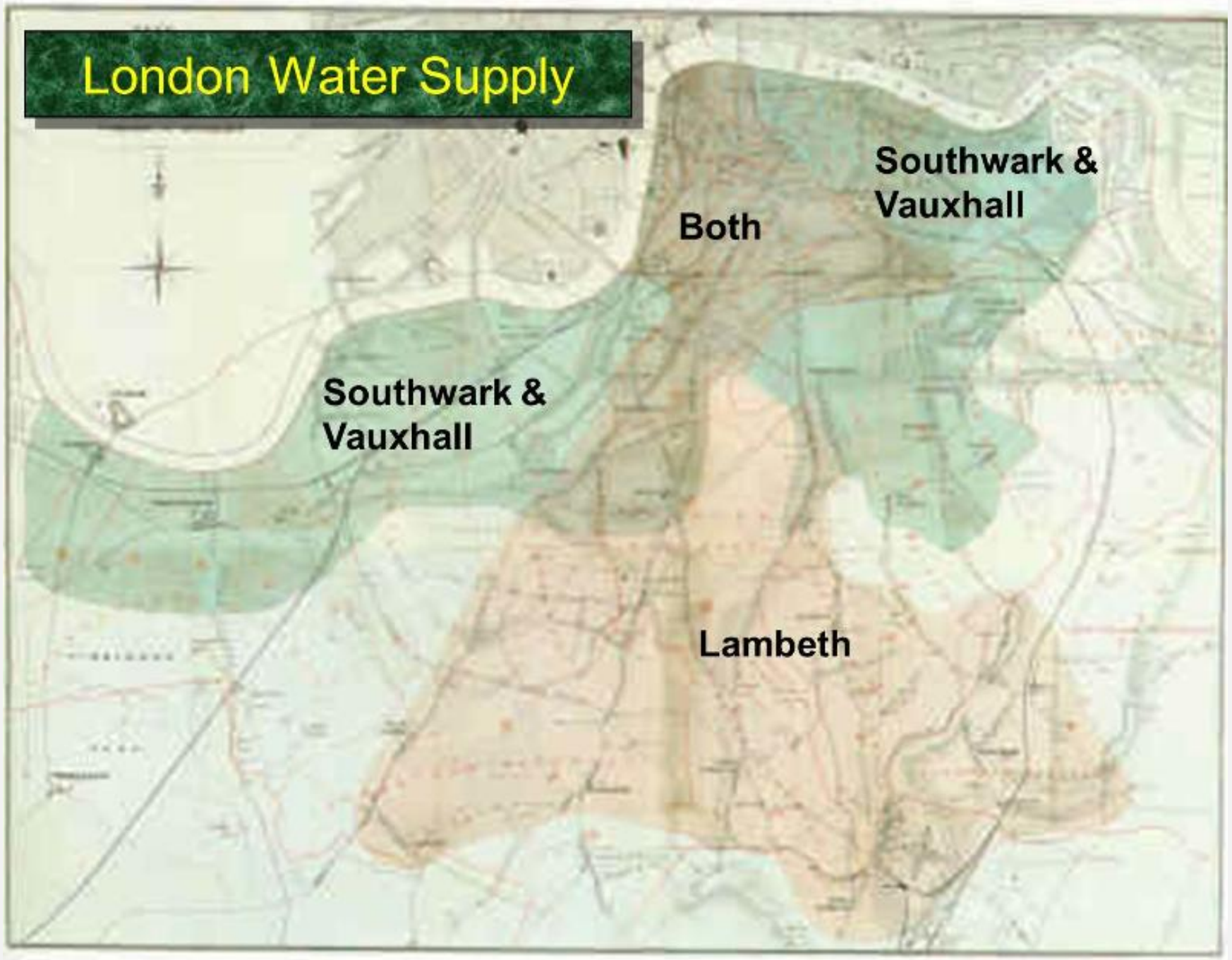
He uses a mix of assumptions and observational data:

1. Assume/show populations are similar in all relevant respects except for which water company supplies them
   - What counts as a relevant respect?
   - We use our best judgment
2. Some areas of London (including the pump) are supplied by Southwark and Vauxhall (S&V) Company
   - S&V drew water downstream of sewage
3. Some areas of are supplied by Lambeth Company
   - Drew water upstream of sewage

"Rich and poor, both large houses and small, there is **no difference** either in the condition or occupation of the persons receiving the water of the different companies."




Soho
Southwark Waterworks
Lambeth Waterworks
Marble Arch
Buckingham Palace
Victoria and Albert Museum
Saatchi Gallery
Imperial War Museum
Stamford Bridge
Royal Botanic Gardens, Kew
Richmond
Putney Heath
Richmond Park
Wimbledon and Putney Commons
Clapham Common
Dulwich Village
Streatham Common
Crystal Palace Park
Twickenham
Hampton Hill
Teddington
Bushy Park
Hampton Wick
Osterley Park And House
Royal Observatory Greenwich
Isle of Dogs
CITY OF LONDON
WESTMINSTER
KENSINGTON
CHISWICK
FULHAM
BATTERSEA
CLAPHAM
BRIXTON
PECKHAM
BALHAM
EARLSFIELD
WEST NORWOOD
CRYSTAL PALACE
FOREST HILL
River Thames



NATURAL EXPERIMENT

- Snow couldn't control the assignment of the treatment himself (diseased water)
- But he could plausibly argue that assignment to treatment was either random, or not related to other factors that affect the outcome
- Today, we call this setup a "natural experiment"
- Subjects receive treatment or control in a way that is **as if random**, and unrelated to **confounders**

RESULTS: LOOKS PRETTY BAD FOR S&V

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

In this case: S&V is the **treatment** case and Lambeth is the **control**

With this experiment he was able to go deeper than just the pump as the problem

The experiment fails to disprove the theory, it still doesn't prove

Also doesn't address what it is in the water specifically that's so deadly

Outline

1. John Snow & thinking like a scientist
2. John Snow & the search for causality
3. Key concepts in causality

KEY CONCEPTS FROM TODAY

- Theory
- Good theory
- Observable implication
- Hypothesis
- Causal mechanism
- Association
- Causation
- Experiment
- Treatment
- Control
- Randomization
- Natural experiment

Coming up!

- Independent variable
- Dependent variable
- Confounders
- Deterministic vs. Probabilistic causality
- Selecting on the dependent variable
- Endogeneity
- Necessary vs. sufficient conditions

Outline

1. John Snow & thinking like a scientist
2. John Snow & the search for causality
3. Key concepts in causality