RICE | D2K

Data to Knowledge Lab

Welcome to DSCI 101

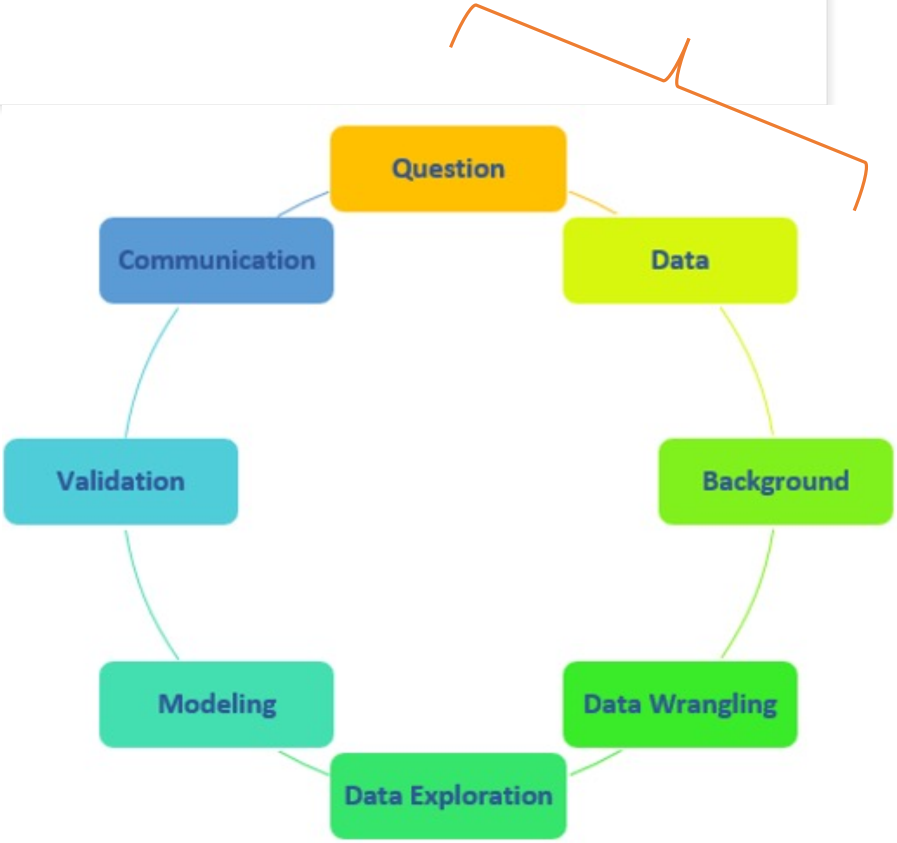
Introduction to Data Science

Week 1 Recap

- What is Data Science?
  - definition of data and data science
  - brief history and modern development

- Data Science pipeline overview
  - end-to-end process
  - a real data example

- Tools you will need for this course
  - Python and Jupyter notebook



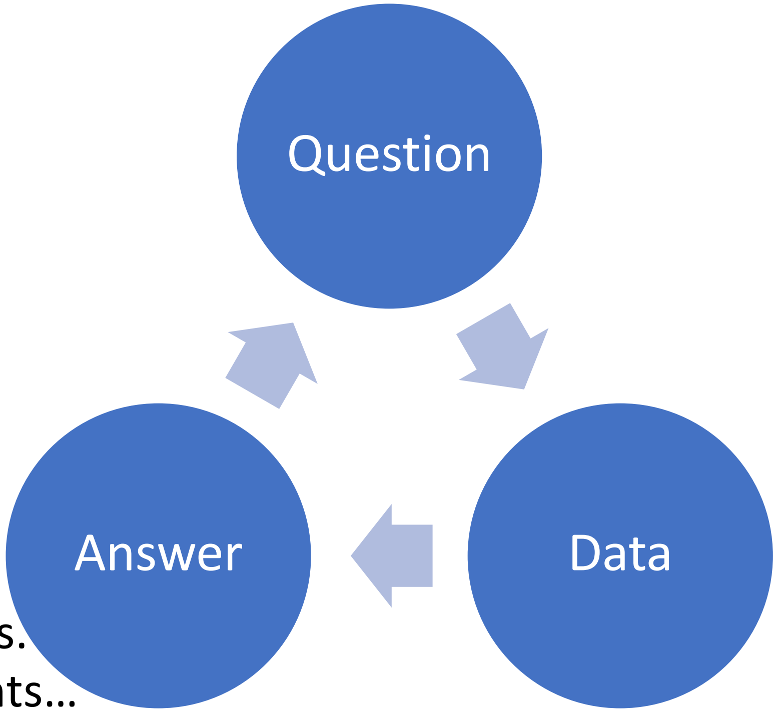
Week 2 Preview

- What type of questions can be answered using data science?
  - Question formulation – 3 types of questions

- Where do data come from?
  - Observational studies vs. randomized experiments

- Data basic concepts
  - data structure and variable types
  - population and random sample
  - sampling bias in surveys

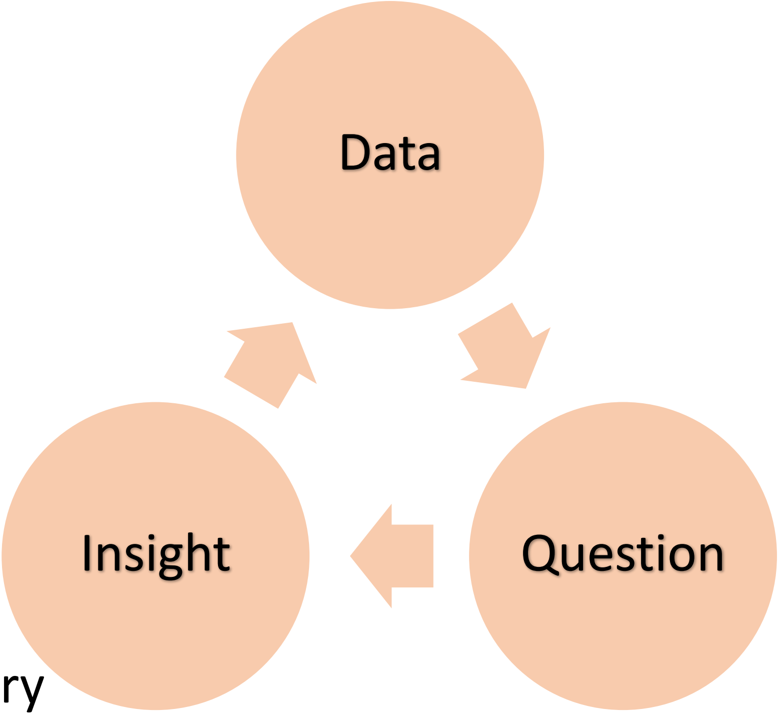
The science of Data Science



- Traditional research model
  - Data-intensive science and engineering
  - Scientists seek to answer questions, using rigorous methods & careful observations. collected from field notes, surveys, experiments…

- Data Science in historical content:
  - Kepler’s laws of planetary motion
  - Tycho Brahe, a Danish astronomer who spent his life collecting data

The science of Data Science



- Motivated by Data-intensive science
    - Data-driven inquiry
    - Tools designed and used by human experts
- Powered by the computer and Internet industry
    - automated data collection --> Big Data
    - Data mining, Artificial Intelligence, and Machine learning
- Data Science today:
    - [How we teach computers to understand pictures]

The New York Times

Pfizer's I
Vaccine I

Pfizer announced pos
vaccine trial, cementir
has unfold

NETFLIX RESEAR

RESEARCH AREAS

Recomm

Figuring out how to

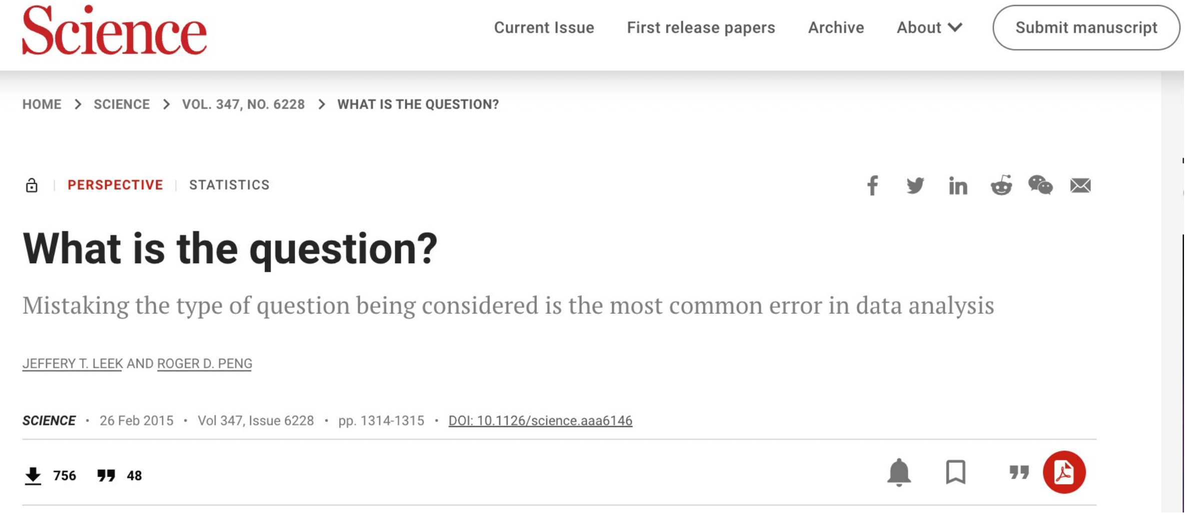
al polls?

polls, accounting for each poll's
ncy

Question Formulation

- How to ask the “right” data questions?

- In general, these 3 types of questions!
  - Descriptive / exploratory
    - “what” about observed data
  - Predictive / decision-making
    - “what” about future data
  - Inferential / causal
    - “how” and “why”
    - about larger population



What is the question?

Mistaking the type of question being considered is the most common error in data analysis

JEFFERY T. LEEK AND ROGER D. PENG

SCIENCE • 26 Feb 2015 • Vol 347, Issue 6228 • pp. 1314-1315 • DOI: 10.1126/science.aaa6146

Descriptive summary

- The “facts” summarized from data:
  - as numbers that are easy to interpret
  - plots and graphs for visual comparison

- The most straightforward type of questions
  - do your data have it or not?

source

- This looks simple, but can be tricky sometimes:
  - data can be deceptive depends on the way you present it
  - “facts” and “opinions” sometimes can intertwine

Predictive models

Sad to report that the inventor of predictive text has passed away
His funfair will be held next Monkey
Joko Jokes

- This is usually what the buzz is all about
  - but…

- The three important aspects before building a model:
  - Is the question really predictive or is it inferential…often not a clear cut!
  - Do I have data to train a predictive model?
  - Can I validate and assess my model?

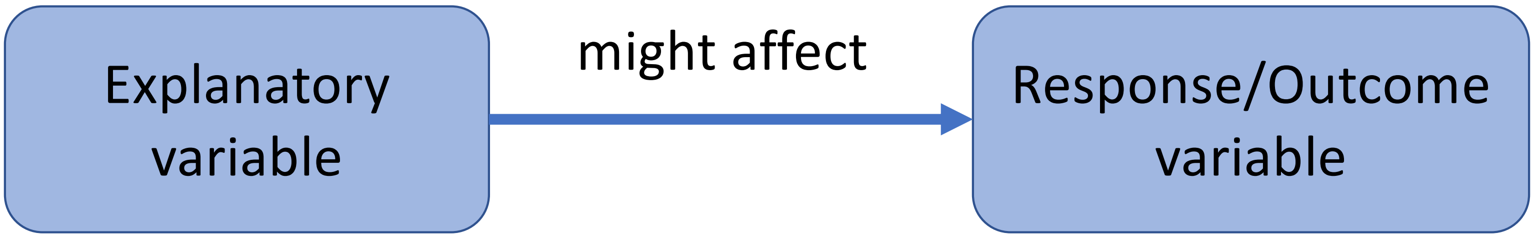
- Besides the question of what, we often are obsessed with "why”
  - Why does the model make certain prediction?

Scientific inference

- Statistical inference:
  - **the process of drawing conclusions about populations or scientific truths from data -** data collected on a random sample from the population
  - quantify the uncertainty of the conclusions made – uncertainty due to lack of data on the entire population

- Causal inference:
  - identify and quantify causal relationship between variables of interest
  - association and causation – why are they not the same?

Cause and effect



- Questions:
  - Does the death penalty have a deterrent effect?
  - Is chocolate good for you?
  - What causes breast cancer?

- These questions attempt to assign a cause to an effect
  - for intervention - control outcome by manipulating the cause
  - a careful examination of data can help shed light on questions like these

Association is not causation, not always!

- Association is when two variables observed to happen at the same time

- Association does not necessarily mean one variable caused the other!
  - common cause
  - common effect
  - spurious correlation

- **Causation can be inferred from:**
  - **randomized experiments + proper statistical inference**
  - **observational studies + proper causal inference**

Observational studies vs. experiments

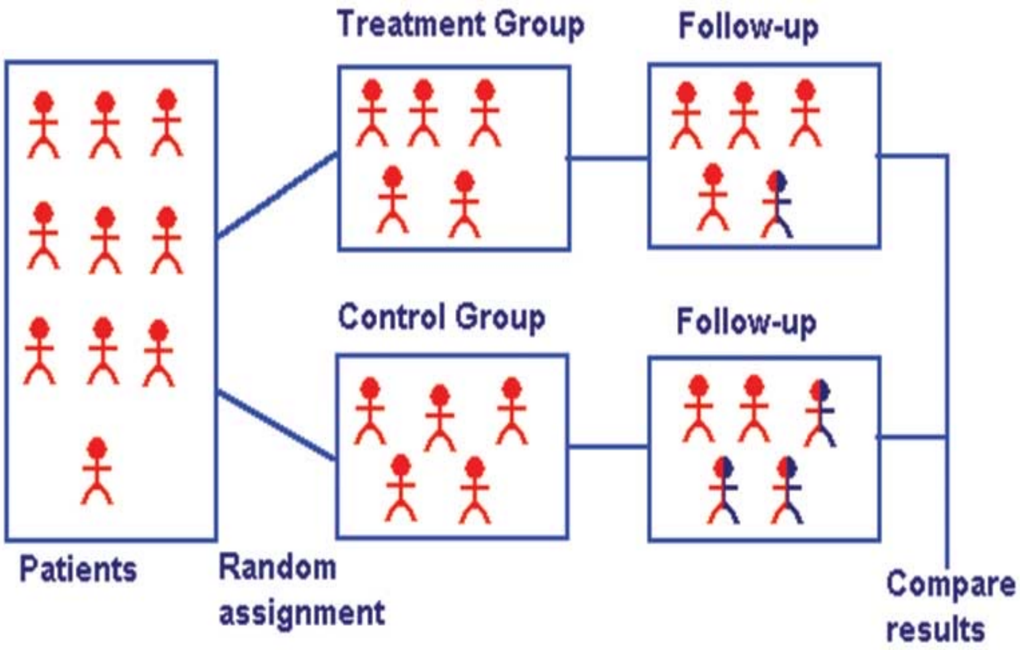
- Two primary types of data collection

- Observational study:
  - researchers make conclusions based on data that they have observed but had no hand in generating
  - In general can provide evidence of naturally occurring association between variables

- Experiment:
  - studies where researchers assign treatments to cases and collect data
  - randomized experiment: when assignment includes randomization

Randomized controlled trial

- Investigate **causal** connection between explanatory variable and outcome
- Study subjects are **randomly** assigned to Treatment vs. Control
- Key to establish causality: the two groups were **similar except for the treatment**.
  - Double blinded
  - Gold standard of clinical trials



Confounding factors

- 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.

- 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**, called **confounding factors**.

Experiment- Covid vaccine safety and efficacy

- Pfizer and BioNTech’s trial included nearly 44,000 volunteers, half of whom received the vaccine. The other half received a placebo shot of salt water. Then the researchers waited to see how many in each group developed Covid-19.
- “out of 170 cases of Covid-19, 162 were in the placebo group, and 8 were in the vaccine group. Out of 10 cases of severe Covid-19, 9 had received a placebo.”
  - Vaccine efficacy measure = $\frac{\text{Risk among unvaccinated group} - \text{risk among vaccinated group}}{\text{Risk among unvaccinated group}} = \frac{162-8}{162}$=95%

Observational studies- Covid lockdown policy and impact

- COVID-19: The impact of social distancing policies, cross-country analysis

- Associations Between Governor Political Affiliation and COVID-19 Cases, Deaths, and Testing in the U.S.

- Levels of economic developement and the spread of coronavirus disease 2019 (COVID-19) in 50 U.S. states and territories and 28 European countries: an association analysis of aggregated data

Causal inference examples

- Causation can be inferred from:
  - randomized experiments + proper statistical inference
  - **observational studies + proper causal inference**

- The “Grand Experiment” by John Snow:
  - John Snow and the 1854 Broad Street cholera outbreak in London
  - Case Study: John Snow and the Origin of Epidemiology

- The big debate in the 50s: does smoking cause lung cancer?
  - The study that helped spur the US stop smoking movement
  - Smoking and Lung Cancer: From Association to Causation
  - Why the Father of Modern Statistics Didn’t Believe Smoking Caused Cancer

Case study: London, early 1850's

A COURT FOR KING CHOLERA.

Illustration from *Punch* (1852).

"These problems are, and will probably ever remain, among the inscrutable secrets of nature. They belong to a class of questions radically inaccessible to the human intelligence"

-- The times of London, September 1849,
on how cholera is contracted and spread

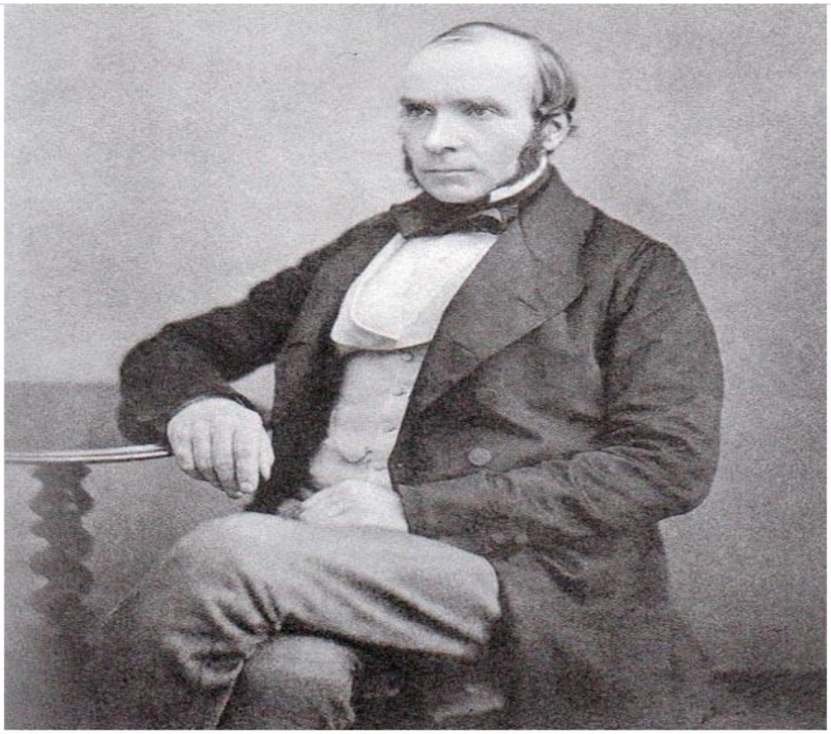
Miasma theory

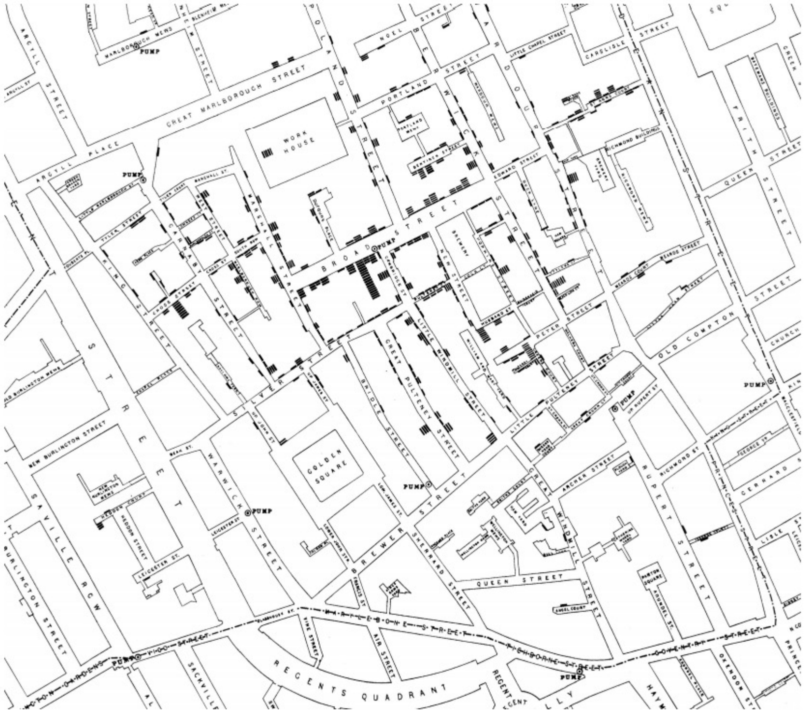
- Bad smells believed to be the main source of disease
- Suggested remedies:
  - ‘fly to clean air”, “a pocket full oposies”, “fire off barrels of gunpowder”
- Celebrity Miasmatists
  - Florence Nightingale, Edwin Chadwick…

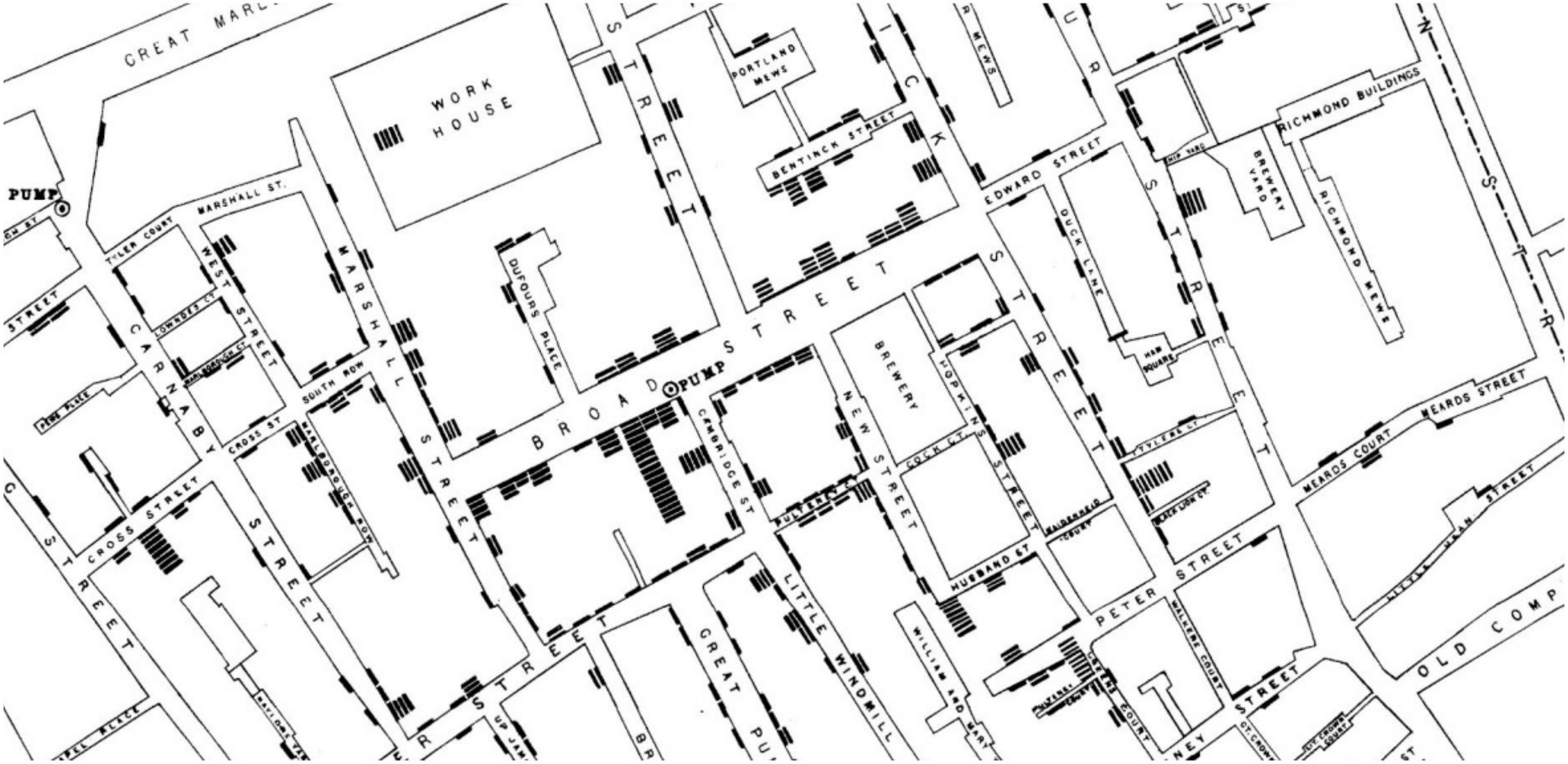
- One person who was a little doubtful

An 1831 color lithograph by Robert Seymour depicts cholera as a robed, skeletal creature emanating a deadly black cloud.

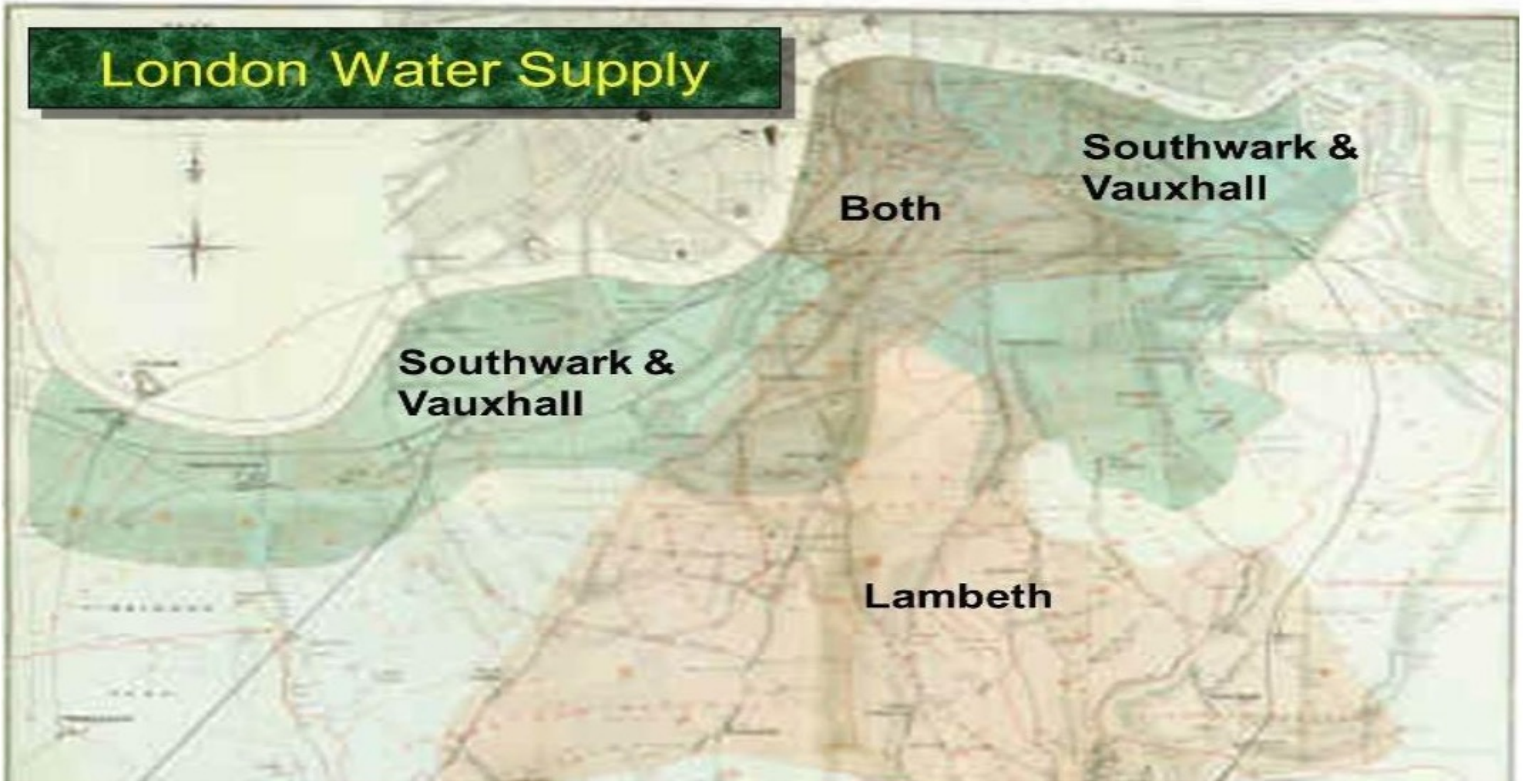
John Snow, 1813 - 1858



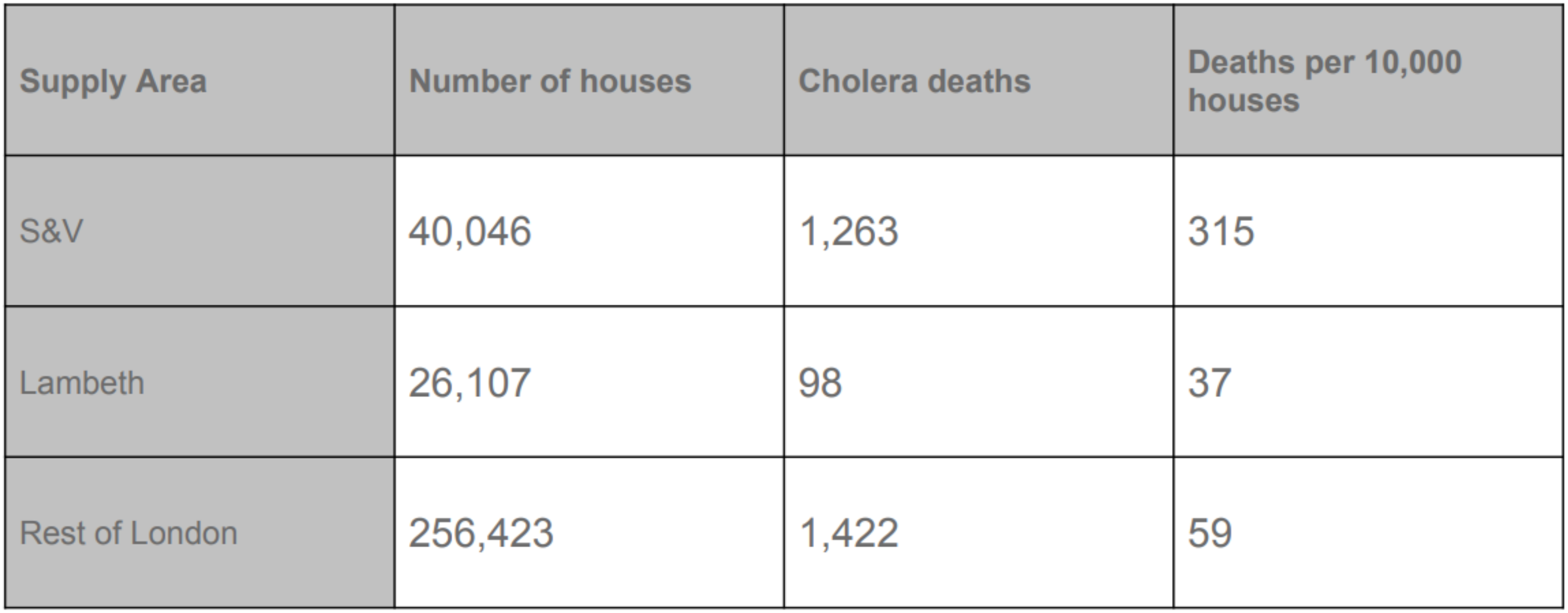



GREAT MARL
WORK HOUSE
PUMP
MARSHALL ST.
TYLER COURT
CARNABY
STREET
WEST STREET
LOWNDES CT.
MARLBOROUGH CT.
SOUTH ROW
CROSS ST.
MARLBOROUGH ROW
MARSHALL STREET
CROSS STREET
DUFOURS PLACE
BROAD STREET
PUMP
CAMBRIDGE ST
PORTLAND MEWS
BENTINCK STREET
EDWARD STREET
DUCK LANE
BREWERY
HOPKINS STREET
NEW STREET
COCK CT.
PULTENEY CT.
HUSBAND ST
MAIDENHEAD COURT
PETER STREET
LITTLE WINDMILL
GREAT PU
WILLIAM AND MARY
RICHMOND BUILDINGS
BREWERY YARD
RICHMOND MEWS
HAM YARD
TYLERS CT.
BLACK LION CT.
MEARDS COURT
MEARDS STREET
WALKERS COURT
OLD COMP
CHAPEL PLACE
UP JAMES
BR

London Water Supply



Snow’s Table

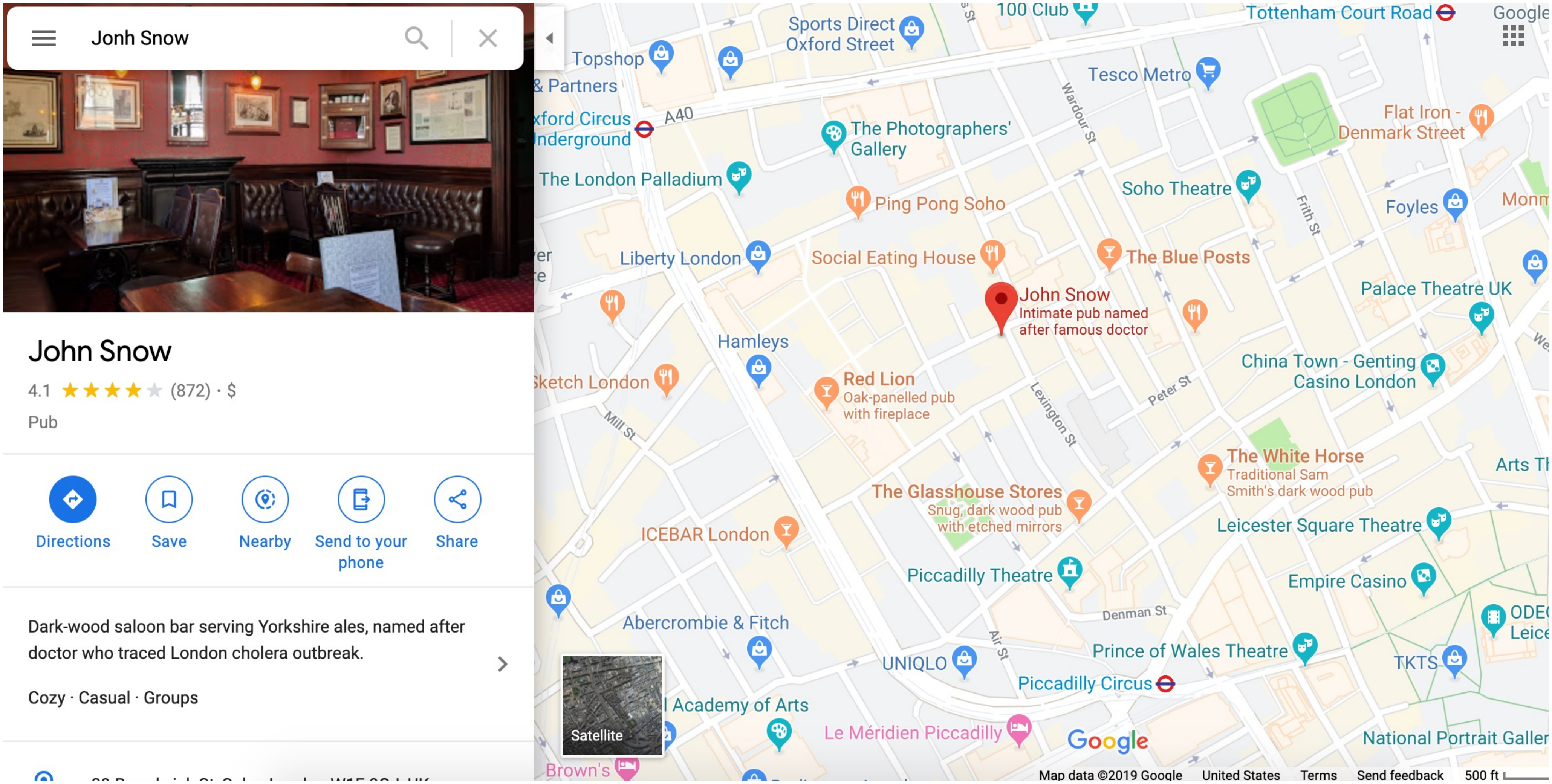


Supply Area	Number of houses	Cholera deaths	Deaths per 10,000 houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

The “Grand Experiment”

“No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentle folks down to the very poor, were divided into two groups without their choices, and, in most cases, without their knowledge; ... 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...”

- **Key to establishing causality**: The two groups were similar except for the treatment.



Jonh Snow

John Snow

4.1 ★★★★☆ (872) · $

Pub

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

Cozy · Casual · Groups

John Snow legacy

- Father of modern epidemiology
- Early data scientist

- Today, epidemiologist still ask the question: “where is the handle to this pump?”

Data structure

- Tabular structure: data tables
    - arrange data values in rows and columns
    - each row is one observation / individual / entity …
    - each column is a variable / attribute / feature …

- A database contains many data tables
    - these data tables are linked to each other
    - relational database management

- Other data structure: hierarchical
- Unstructured: free text, sound, pictures, video, EEG, fMRI…

Data file format

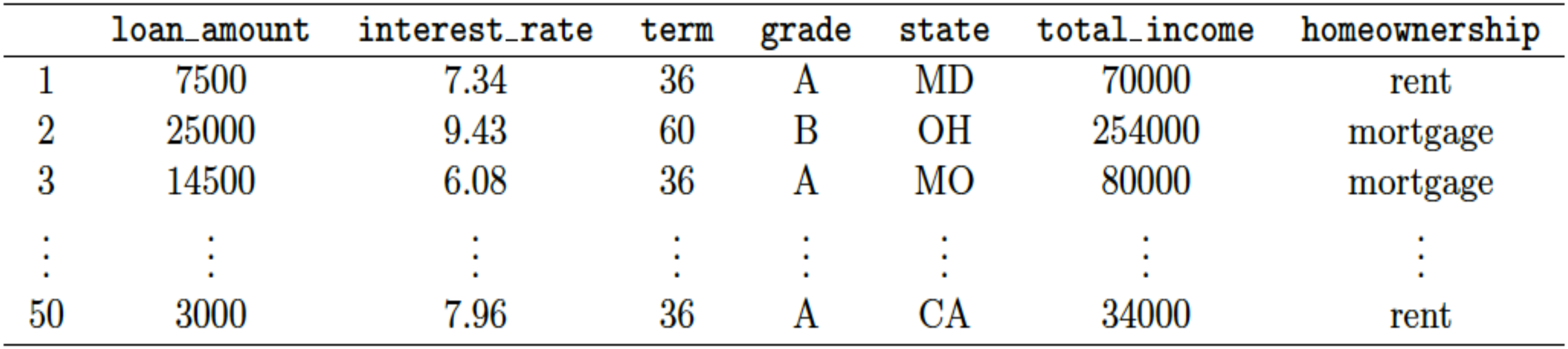
- Do not mistaken this as data structure!

- Comma-Separated Values (CSV) and Tab-Separated Values (TSV)
  - almost always contain tabular data
  - most datasets in this format

- Other formats: hierarchical, key-value format, …
  - JavaScript Object (JSON)
  - eXtensible Markup Language (XML) and HyperText Markup Language(HTML)
  - picture, video, signal: preprocessing needed

We love tabular data!

- Easy to understand and easy to work with
- Who are the data collected on? – the rows!
- What variables / data attributes are collected? – the columns!



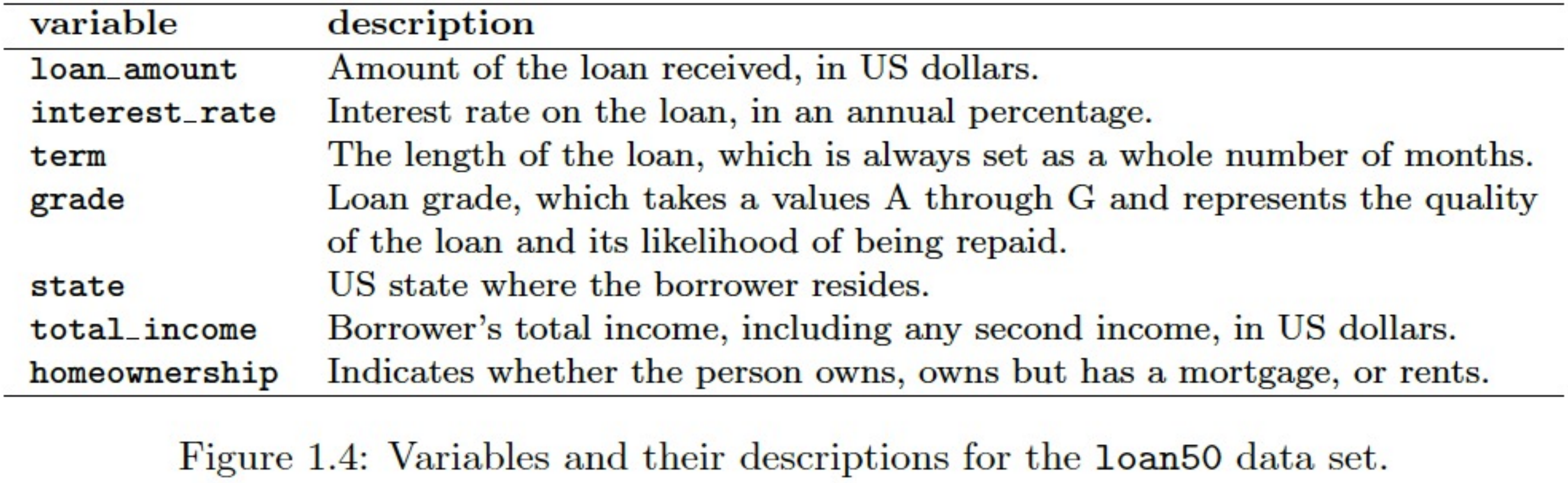
	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Figure 1.3: Four rows from the loan50 data matrix.

OpenIntro Statistics

Understand the columns

- always look for the data keys!
  - what is each variable, possible values and unit of measurement



variable	description
loan_amount	Amount of the loan received, in US dollars.
interest_rate	Interest rate on the loan, in an annual percentage.
term	The length of the loan, which is always set as a whole number of months.
grade	Loan grade, which takes a values A through G and represents the quality of the loan and its likelihood of being repaid.
state	US state where the borrower resides.
total_income	Borrower's total income, including any second income, in US dollars.
homeownership	Indicates whether the person owns, owns but has a mortgage, or rents.

Figure 1.4: Variables and their descriptions for the loan50 data set.

OpenIntro Statistics

Types of variables

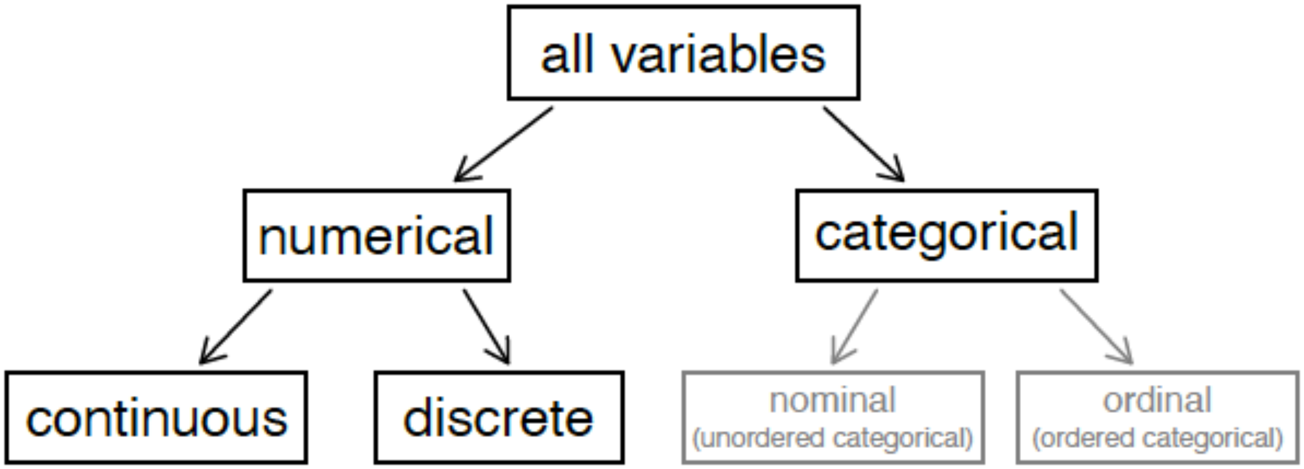
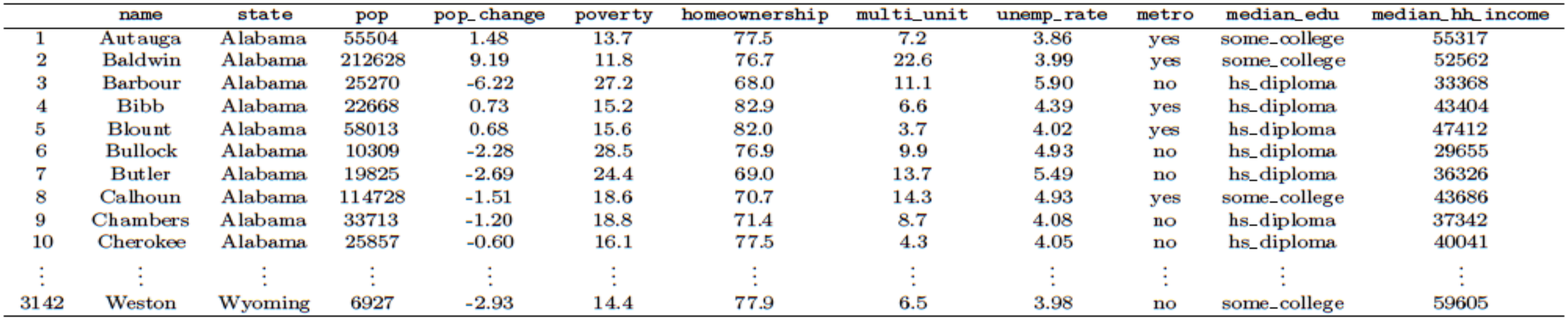


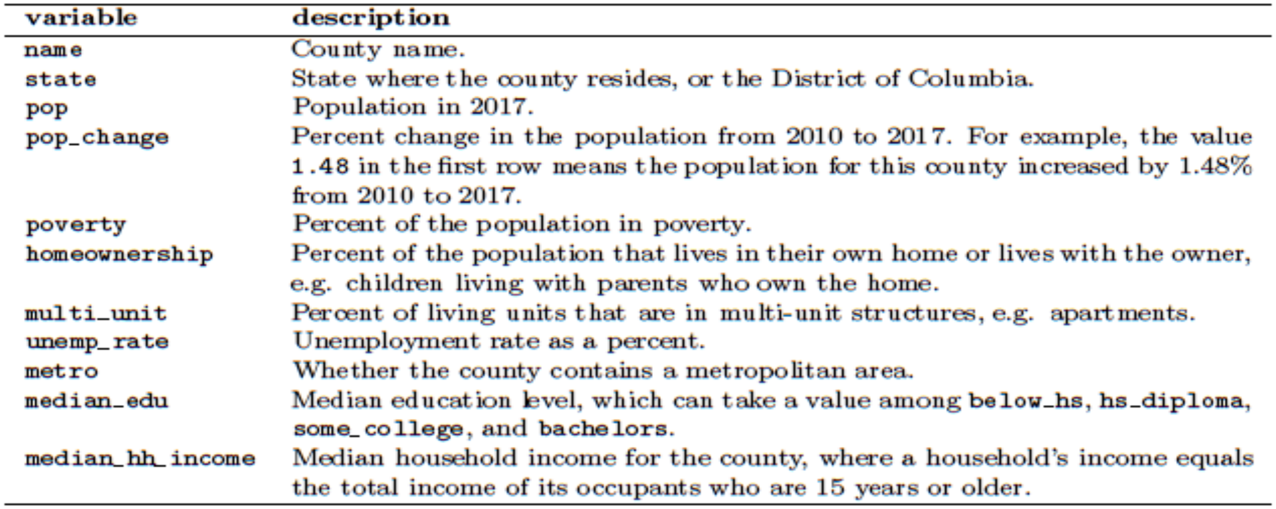
Figure 1.7: Breakdown of variables into their respective types.

OpenIntro Statistics



	name	state	pop	pop_change	poverty	homeownership	multi_unit	unemp_rate	metro	median_edu	median_hh_income
1	Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86	yes	some_college	55317
2	Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99	yes	some_college	52562
3	Barbour	Alabama	25270	-6.22	27.2	68.0	11.1	5.90	no	hs_diploma	33368
4	Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39	yes	hs_diploma	43404
5	Blount	Alabama	58013	0.68	15.6	82.0	3.7	4.02	yes	hs_diploma	47412
6	Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93	no	hs_diploma	29655
7	Butler	Alabama	19825	-2.69	24.4	69.0	13.7	5.49	no	hs_diploma	36326
8	Calhoun	Alabama	114728	-1.51	18.6	70.7	14.3	4.93	yes	some_college	43686
9	Chambers	Alabama	33713	-1.20	18.8	71.4	8.7	4.08	no	hs_diploma	37342
10	Cherokee	Alabama	25857	-0.60	16.1	77.5	4.3	4.05	no	hs_diploma	40041
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Weston	Wyoming	6927	-2.93	14.4	77.9	6.5	3.98	no	some_college	59605

Figure 1.5: Eleven rows from the county data set.



variable	description
name	County name.
state	State where the county resides, or the District of Columbia.
pop	Population in 2017.
pop_change	Percent change in the population from 2010 to 2017. For example, the value 1.48 in the first row means the population for this county increased by 1.48% from 2010 to 2017.
poverty	Percent of the population in poverty.
homeownership	Percent of the population that lives in their own home or lives with the owner, e.g. children living with parents who own the home.
multi_unit	Percent of living units that are in multi-unit structures, e.g. apartments.
unemp_rate	Unemployment rate as a percent.
metro	Whether the county contains a metropolitan area.
median_edu	Median education level, which can take a value among below_hs, hs_diploma, some_college, and bachelors.
median_hh_income	Median household income for the county, where a household's income equals the total income of its occupants who are 15 years or older.

Figure 1.6: Variables and their descriptions for the county data set.

OpenIntro Statistics

US Decennial Census

- Count **every person** living in US
- Mandated by the Constitution
- Participation is required by law
- Important uses:
  - Allocation of federal funds
  - Congressional representation
  - Drawing congressional and state legislative districts

- data.census.gov

Census

- Consider the following questions:
  - How many squirrels living on Rice campus?
  - How many people living in US?
  - Over the last 10 years, how long does it take to graduate from Rice (undergrads)?

- Each question refers to a target population.

- A census is a complete data set you collect for the entire population.

Samples

- A census is great, but expensive and difficult to execute

- A sample is a subset of the population
  - used to make **inferences** about the population
  - how you draw the sample will affect accuracy

- Two common sources of error:
  - chance error: random samples can vary from what is expected in **any direction**
  - bias: a systematic error in **one direction**

Probability sampling

- **A simple random sample is the most common probability sample!**

- Each member has an equal probability of being chosen

- Sample is meant to be an unbiased representation of the population

Why sample at random?

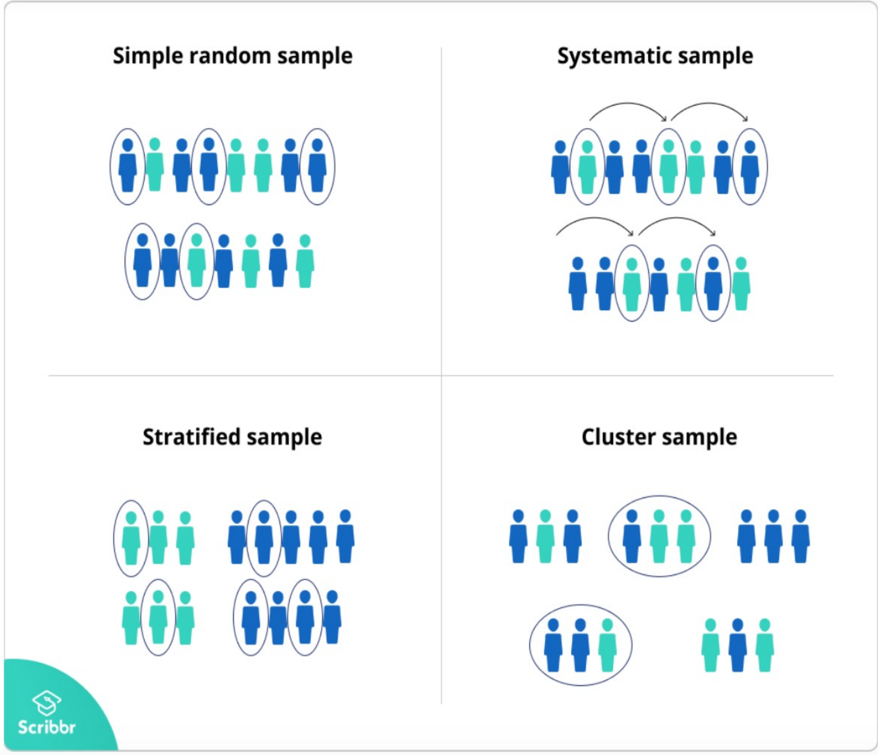
- With random samples we might be able to estimate the chance error and bias
  - remember the two common sources of error?

- We can quantify the uncertainty
  - can measure the errors because you know all the probabilities

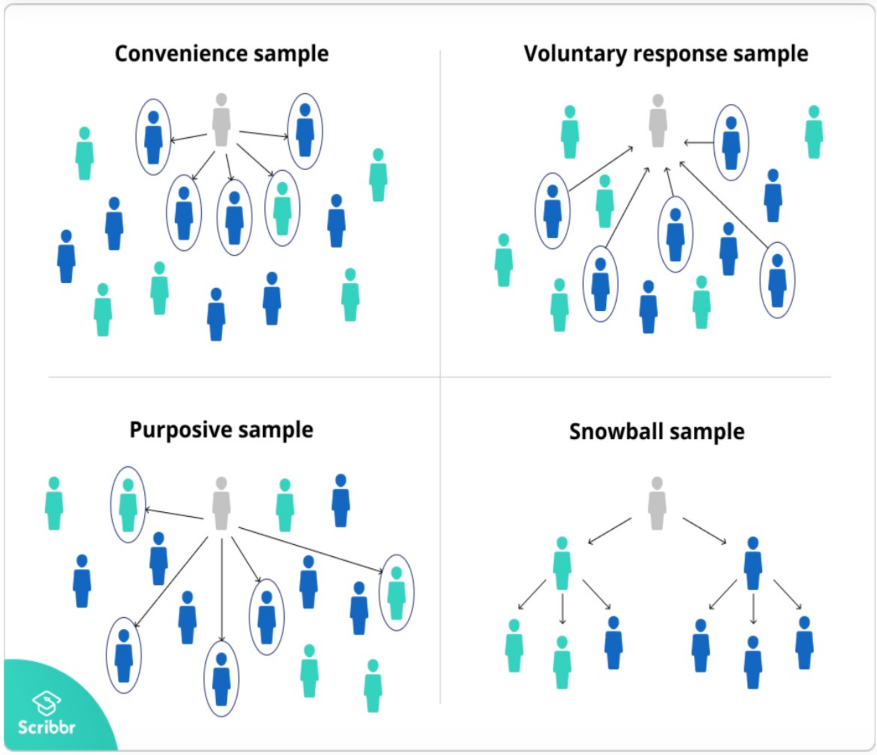
- **Almost all statistical inference rely on random samples!**
  - more on this later in the course

Sampling from a finite population

Good ways to sample



Bad ways to sample



Quality, not quantity

- Design a sample for inference about a population

- Try to ensure that the sample is representative of the population!
  - Don’t just try to get a BIG sample
  - If your method of sampling is BAD, and your sample is BIG
  - What you will have is a BIG BAD sample

- Presidential election polls often have very large sample, but
  - An example of 1936 US Presidential election

Common biases in survey data

- Selection Bias
  - Systematically excluding or favoring particular groups
  - Examine the sampling frame and the method of sampling

- Response Bias
  - People don’t always respond truthfully
  - Examine the nature of questions and the method of surveying

- Non-response Bias
  - People don’t always respond
  - Keep your surveys short and be persistent

A very common approximation

- A common situation in practice:
  - enormous population
  - can only afford to sample a relatively small number

- If the population is huge compared to the sample, then random sampling with and without replacement are pretty much the same

- So it's important to understand random sampling with replacement
  - key to simulating random samples

Demo

Take away messages

- Random sample is very important for studying a population
  - two sources of errors: chance error and bias

- Causal inference / learning is a booming area of data science!

- The Causal Ladder: associations -> interventions -> counterfactuals

- Ask questions about all observational studies that claim causal relationship, are there possible confounding factors?