UC BERKELEY
LEARN2LAUNCH
DATA SCIENCE

Image credit patswerk.nl

LECTURE 1
PIONEERS OF DATA SCIENCE

FABIAN PEDREGOSA

COURSE STRUCTURE

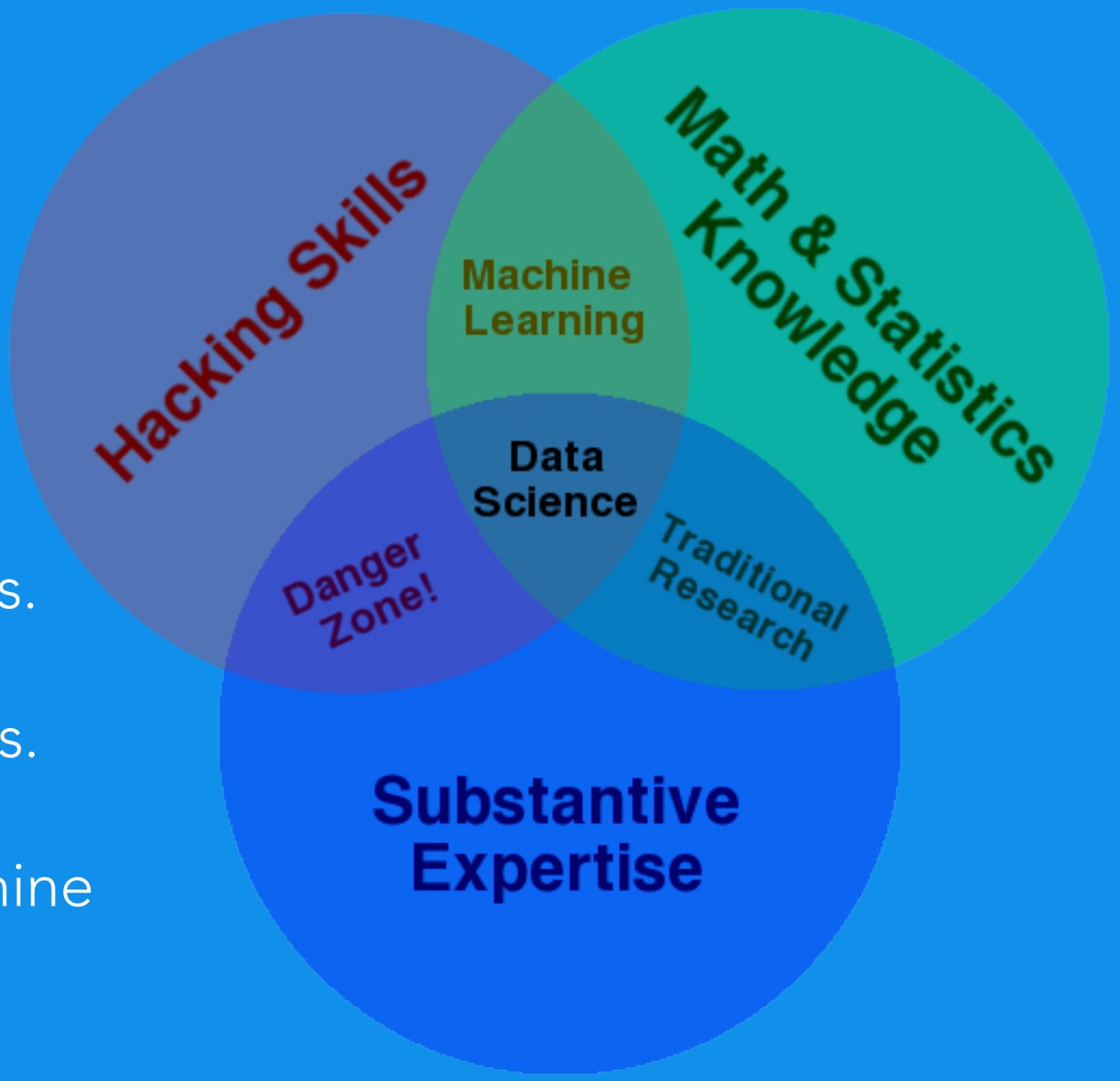
- 3 hour lecture, divided into
    - 1h lecture (slides and whiteboard).
    - 2h lab and exercises jupyter
- Teaching material: https://dsl2l2017.github.io .
- Communication: slack channel

VALIDATION OF THE COURSE

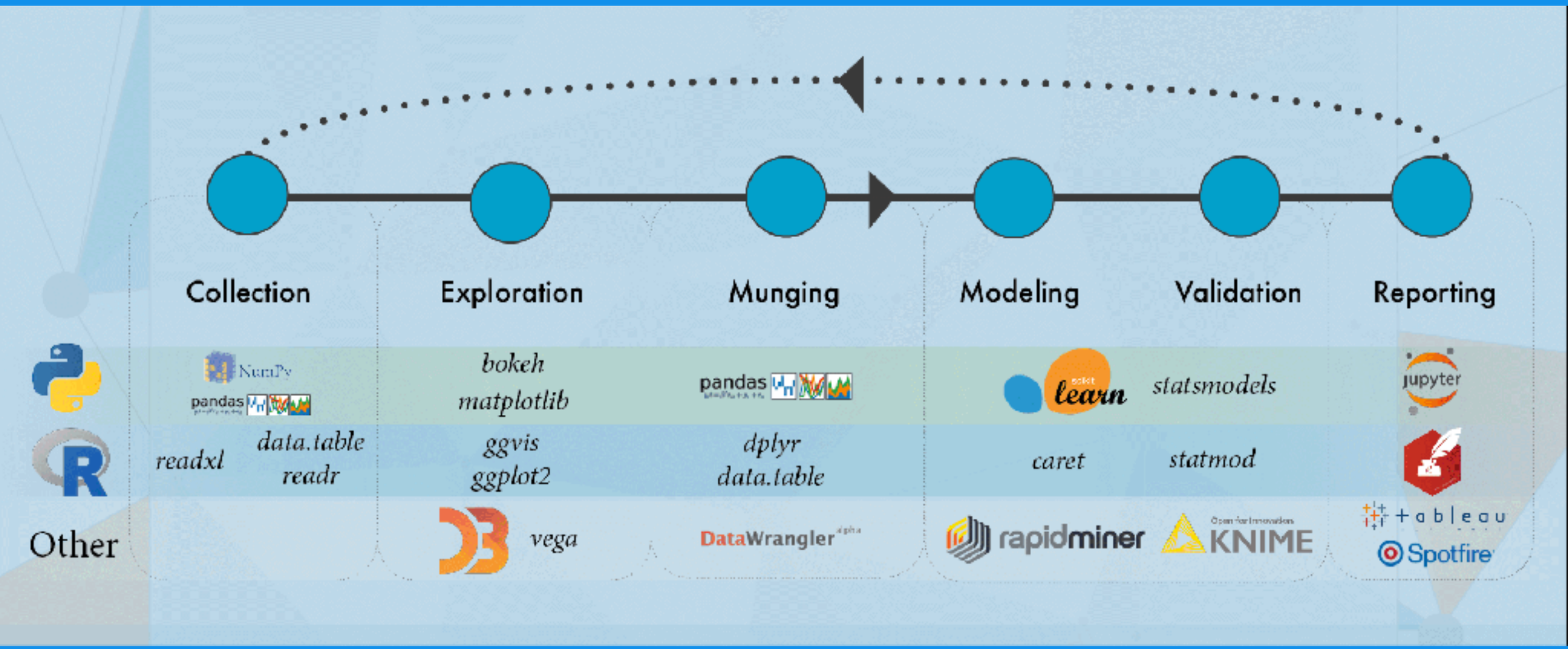
- 2 presentations:
  - October 10th: Dataset analysis and visualization (50%).
  - November 28th: Predictive model implementation (50%).
- List of proposed projects:
  - https://dsl2l2017.github.io/assignments.html

GOALS OF THIS COURSE

- Learn to organize data.
- Communicate critical findings.
- Uncover patterns and insights.
- Make predictions using machine learning.



DATA SCIENCE WORKFLOW

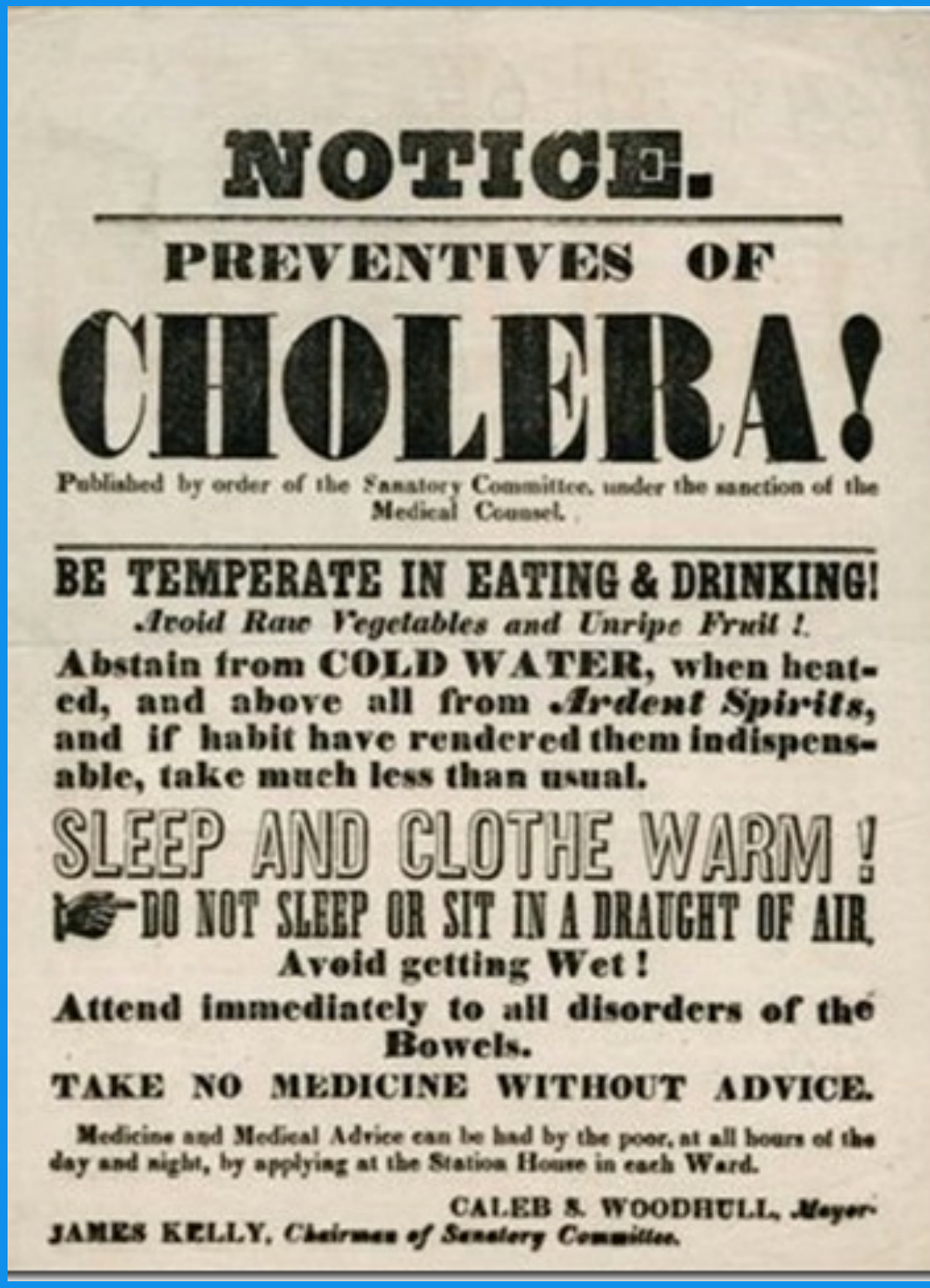


TODAY'S LECTURE

- Pioneers of data science
  - John Snow and the cholera epidemic (1850).
  - R. A. Fisher and the Lady Tasting Tea.

DATA SCIENCE PIONEERS

- London 1850.
- World's wealthiest city but many of its people were desperately poor.
- Cholera epidemic.
- Unclear origin.



NOTICE.

PREVENTIVES OF

CHOLERA!

Published by order of the Sanatory Committee, under the sanction of the Medical Counsel.

BE TEMPERATE IN EATING & DRINKING!

Avoid Raw Vegetables and Unripe Fruit!

Abstain from COLD WATER, when heated, and above all from *Ardent Spirits*, and if habit have rendered them indispensable, take much less than usual.

SLEEP AND CLOTHE WARM!

DO NOT SLEEP OR SIT IN A DRAUGHT OF AIR.

Avoid getting Wet!

Attend immediately to all disorders of the Bowels.

TAKE NO MEDICINE WITHOUT ADVICE.

Medicine and Medical Advice can be had by the poor, at all hours of the day and night, by applying at the Station House in each Ward.

CALEB S. WOODHULL, *Mayor.*

JAMES KELLY, *Chairman of Sanatory Committee.*

DATA SCIENCE PIONEERS

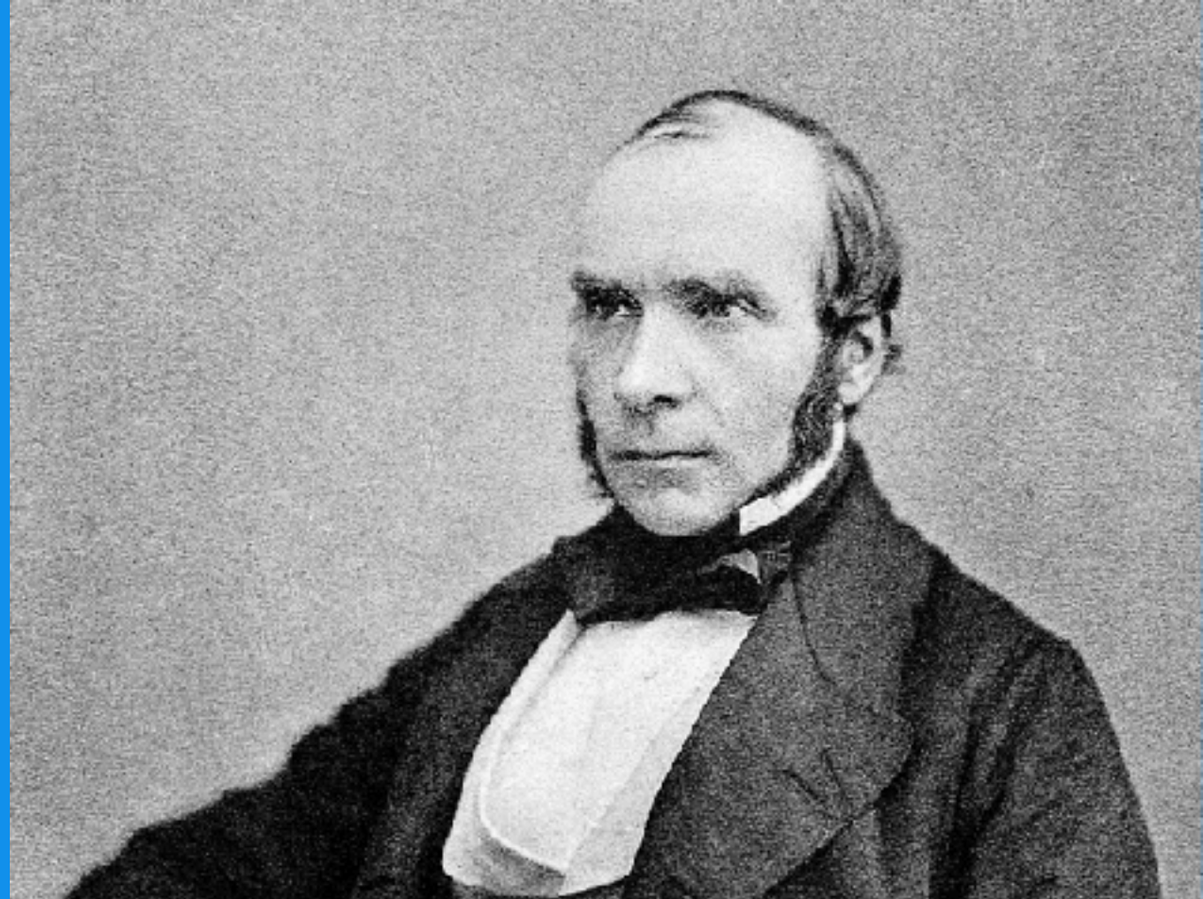
- Jon Snow

- Bastard son of Eddard Stark, Lord of Winterfell.

- Joined the night's Watch

DATA SCIENCE PIONEERS

- Another John Snow (1813-1858).
- Considered the father of modern epidemiology.



CHOLERA EPIDEMIC

- John Snow printed occurrence of cholera-deaths
- Deaths seem to be located around the pump.

TOWARDS CAUSALITY

- Strong indication that water supply was key to controlling cholera.
- Pump removed.
- Cholera epidemic controlled.

ESTABLISHING CAUSALITY

- Can establish causality only if the design is randomized.

LADY TASTING TEA

- England, early 20th century.
- Dr. Muriel Bristol claimed to be able to tell whether the tea or the milk was added first to a cup.

R. A. FISHER

- “Father” of modern statistics.
- Designs an experiment to settle the matter: was her truly able to differentiate both kids of tea?
- Quantify likelihood for this argument.

FISHER'S EXPERIMENT

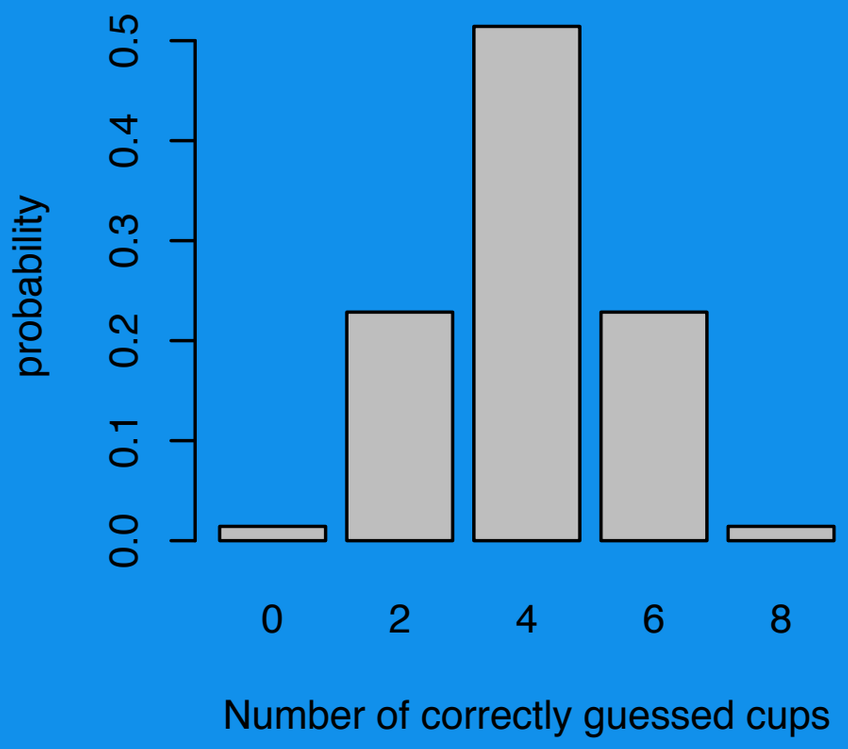
- Present 8 cups of tea (4 with milk first + 4 with tea first) in random order. Have her identify which is which.

- Debate between a skeptic and an advocate.
  - Skeptic: she could get all right purely by chance.
  - Advocate: the chance of getting the answer by chance is very small.

FISHER'S KEY INSIGHT

- If the skeptic's right, the guess of the lady is random.
- Permuting the actual value would also result in random guess.
- Repeat this process to estimate distribution of "getting X cups right".

cups	actual	lady's guess	other scenarios				⋯
1	M	M	T	T	T	T	
2	T	T	T	T	M	M	
3	T	T	T	T	M	M	
4	M	M	T	M	T	M	
5	M	M	M	M	M	M	
6	T	T	M	M	T	T	
7	T	T	M	T	M	T	
8	M	M	M	M	T	T	
correctly guessed		8	4	6	2	4	⋯



FISHER'S MAGIC

- We haven't assumed data comes from a specific statistic.
- The permutation process can be carried by a computer.

"If you can program a computer, you have direct access to the deepest, most fundamental ideas in statistics."

–JOHN RAUSER