ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 1A: What is Statistics?

William Marsh

Remember to Start Recording

Outline

- Aims of statistics
- *Making the most of online study*
- Comparing Machine Learning and statistics
- Topical examples

Module Introduction

- Aim and teaching style is practical
  - Lectures introduce concepts
  - Apply concepts in practical exercises

- Coursework only: no exam
  - See QMPlus page
  - *More later*

What is the Aim of Statistics?

Reliable Understanding from Data

- Let's collect data on eating tomatoes and bananas

Likes Eating	Count	Percent
Tomatoes and Bananas	5	16.7%
Tomatoes only	7	23.3%
Bananas only	10	33.3%
Neither	8	26.7%
Total	30	

- Understanding (or knowledge)
  - *Bananas are more popular than tomatoes*
  - *Bananas are less popular among tomato likers*

Are these statements reliable?

The Problem of Chance

- Maybe the preferences (bananas & tomatoes) arose by chance

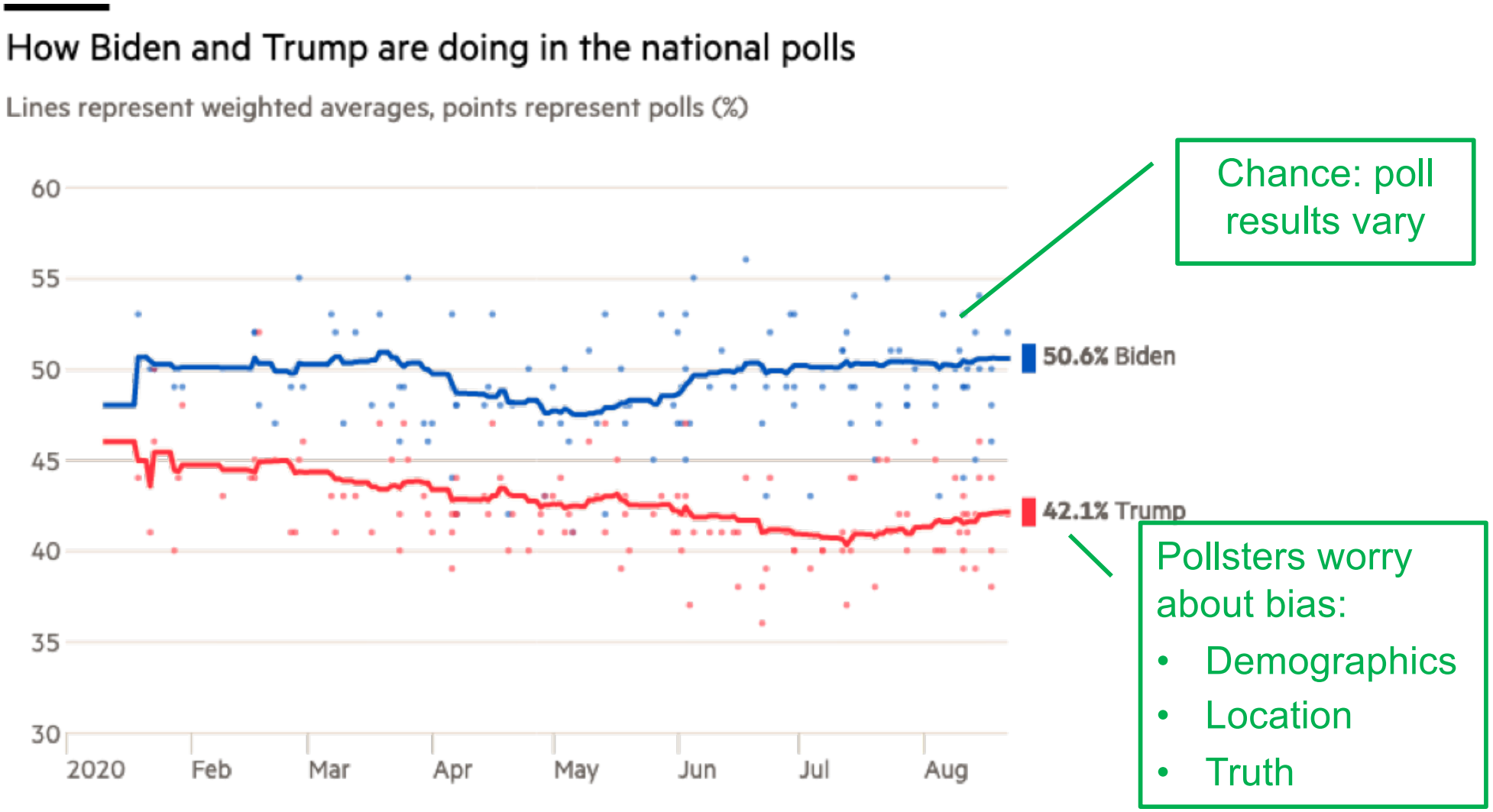
- Statistics looks at way to distinguish chance variation from real variation

The Problem of Bias

- Maybe we gathered data when ‘The Tomato Association’ was running free holidays (to pick tomatoes of course) for members
  - Our data collection was biased
  - Common problem with surveys

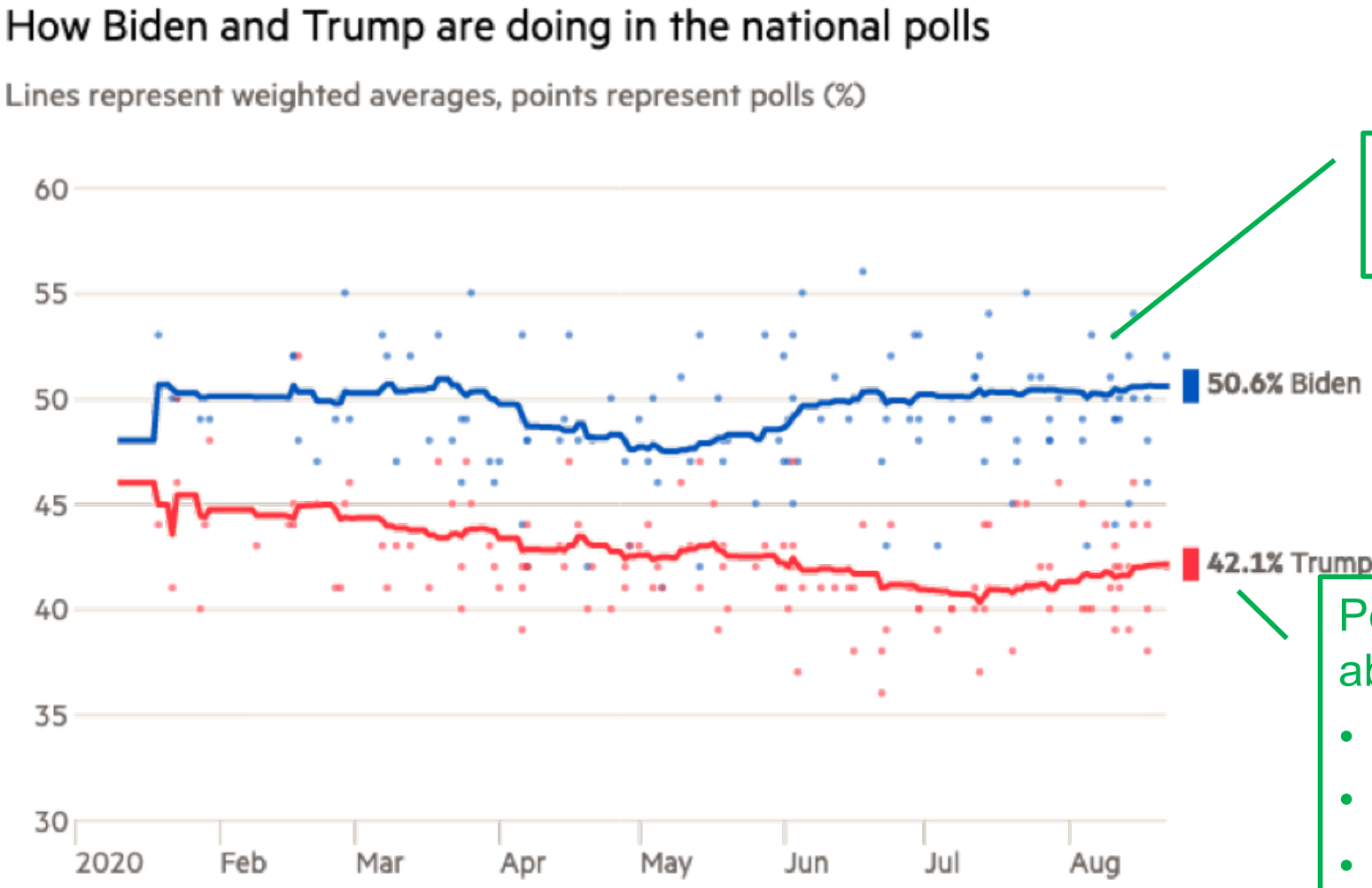
- Statistics considers how to gather data that is not biased (or to correct for possible bias)

Chance and Bias (August 2020)



Source (26/8) https://ig.ft.com/us-election-2020/

Chance and Bias (August 2020)



Chance: poll results vary

Pollsters worry about bias:

- Demographics
- Location
- Truth

Final: 51.3% versus 46.9%
(Wikipedia)

Source (26/8) https://ig.ft.com/us-election-2020/

How to Survive Online Study

Based on my experiences last term,
adapted to 'mixed-mode' education

Every lecture will have a 'learning reflection' slide

Online* Study: Making the Best of It

**Mixed mode: some face-to-face for those able*

The Problems

- Isolation and lack of support
  - We may not meet (e.g. in labs)
  - You may not meet your fellow students ('peer support')
- I cannot tell when you need help
  - Some of you may be asleep

Possible Solutions

- Working together
  - Be willing to meet and talk online
- You have to say if you need help
  - Vital to ask questions (yes, its hard)
  - I am paid (by you) to answer your questions
  - Believe that you can succeed even if you do not get it at first

Data Analysis, Statistics, and Machine Learning

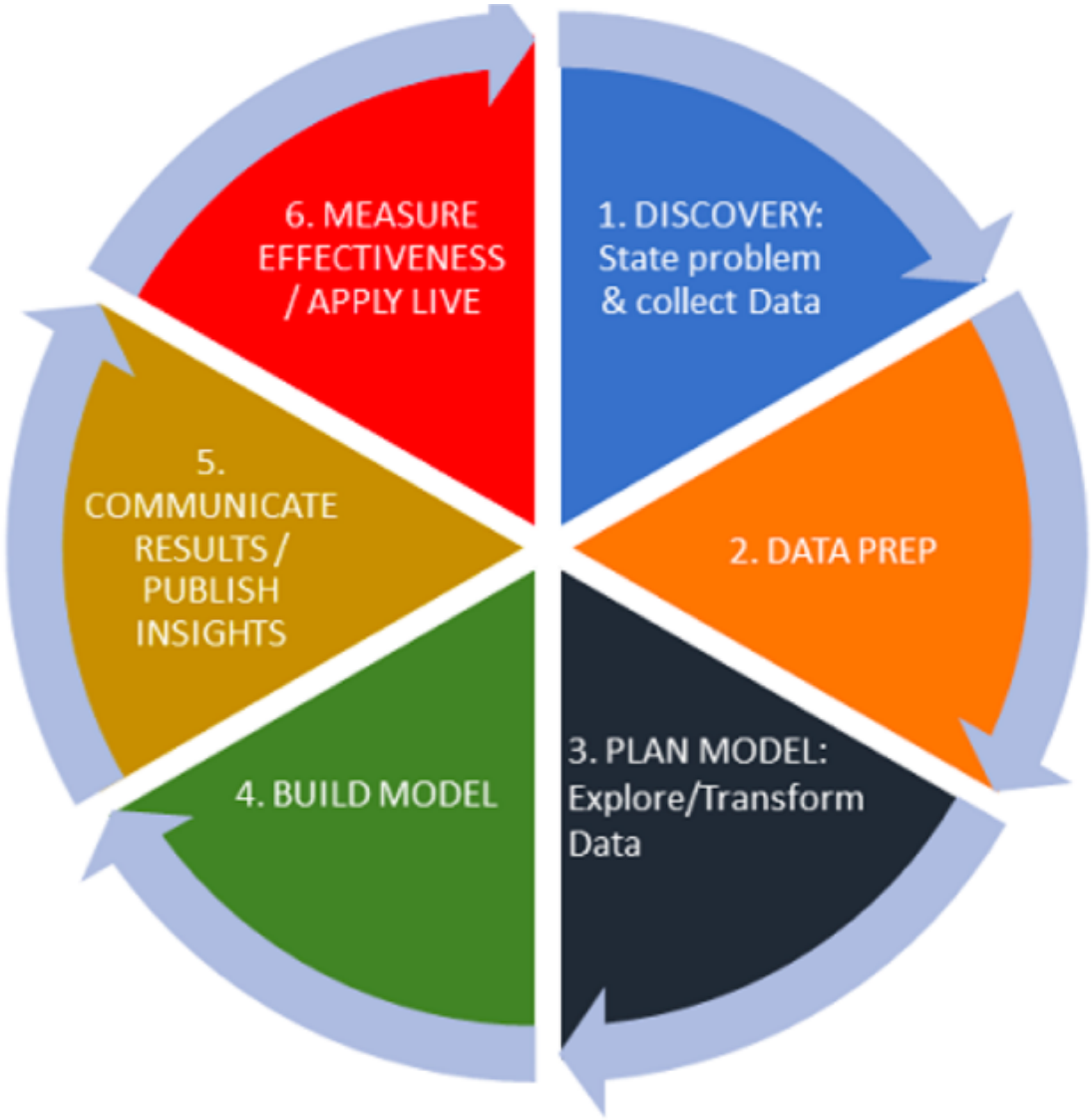
What's the Relationship? (It's changing)

Statistics → ML?

- Both involve data analysis
  - Raw data needs ‘analysis’ - transformation
  - Excel: ok for small datasets
- Computational tools are changing statistics
  - Analysis of large datasets
  - Less emphasis on ‘mathematical solutions’
- Statistics and ML have different aims

Data Analysis Process

- Starts with data collection
- Tidy up data
- Explore data
  - Understand distributions
- Data analysis can stop here
- Both ML and statistics use models and require evaluation



What? versus why?

Comparing Aims

Statistics

- Is there a real difference?
- Explaining differences
  - Why do some students fail?

- Often a group focus

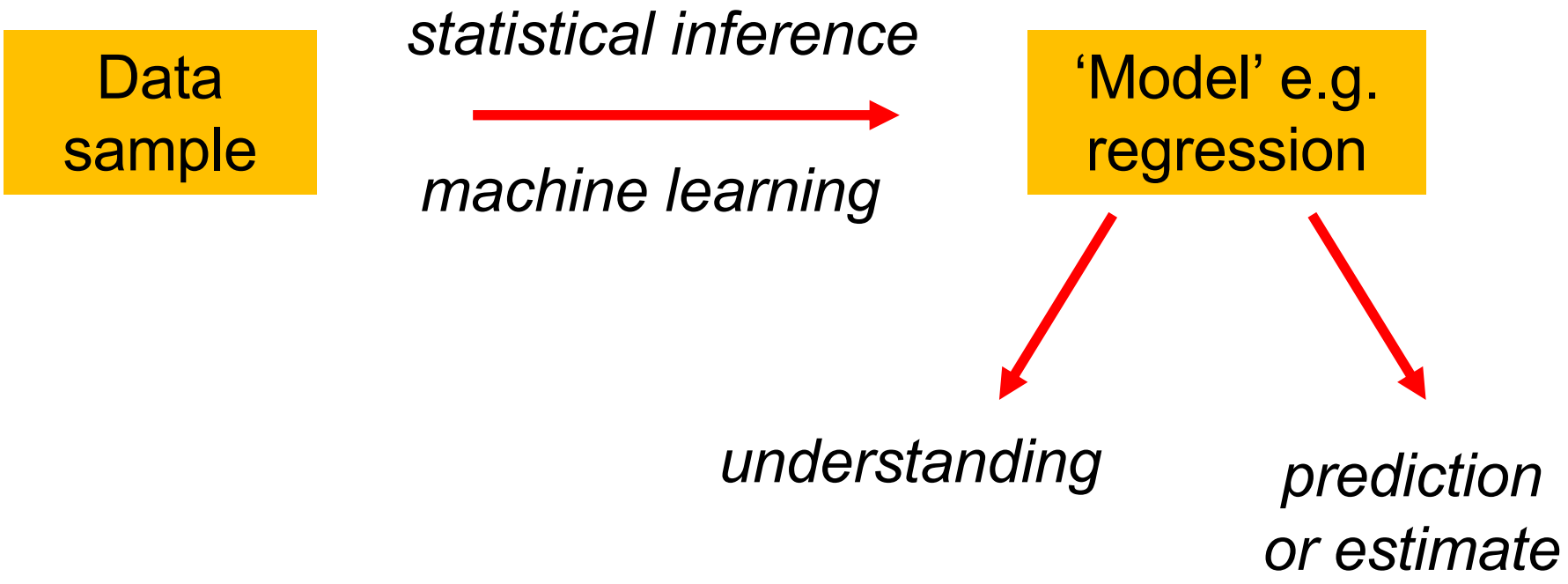
AI / Machine Learning

- Prediction
  - Which students will fail
- Classification
  - What's in a photo?

- Often an individual focus

Overlap: Both Statistics and ML

- Learning (generalising) from data



Lots of Statistics in the News

Challenge of communication

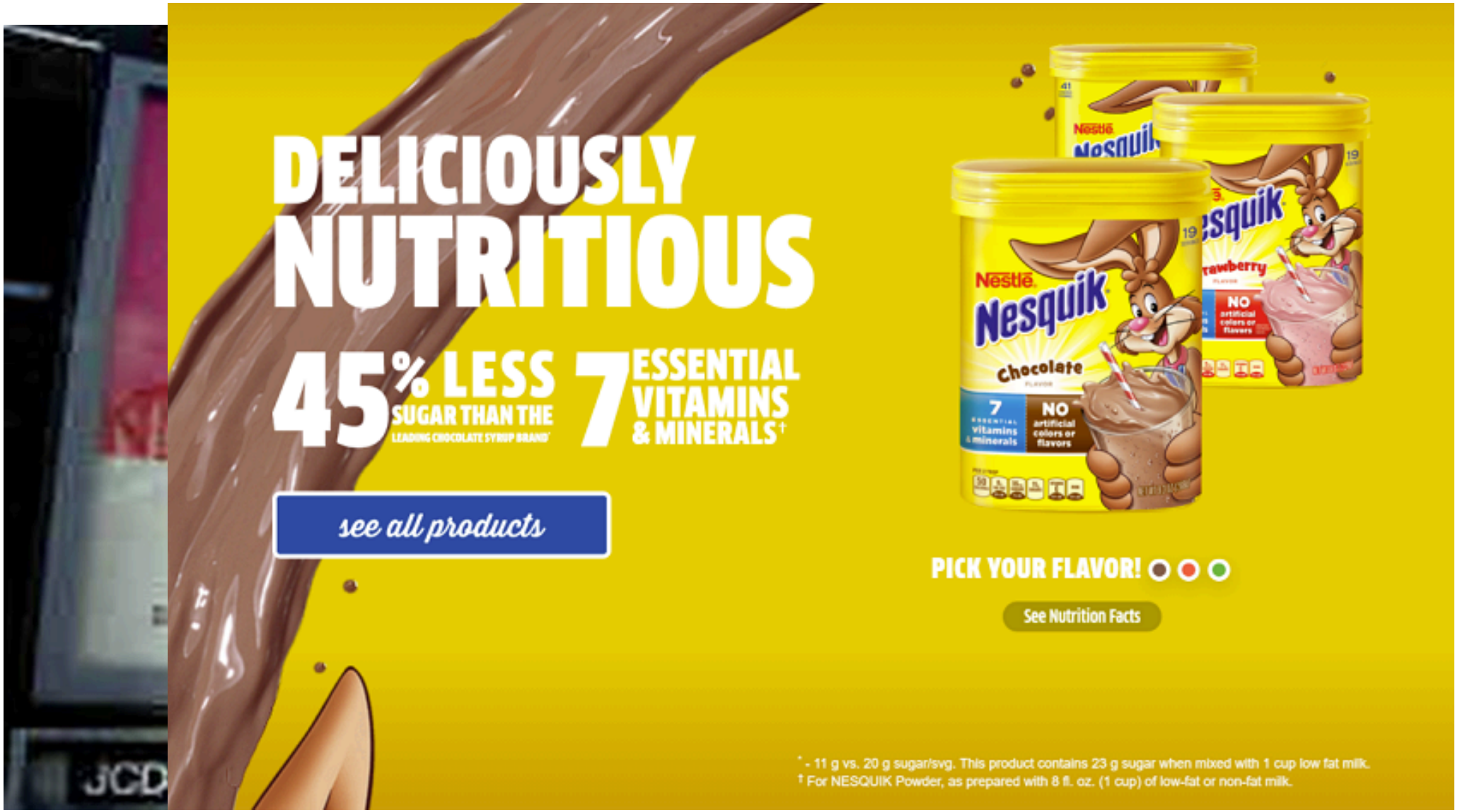
Challenge of uncertainty

Statistical Claims

- What makes a claim statistical?

Statistical Claims

- What makes a claim statistical?



Public Health England data on COVID infection rates

- Rate appears higher in vaccinated patients at some age groups

COVID-19 vaccine surveillance report – week 36

Table 4. COVID-19 cases by vaccination status betw

Cases reported by week of specimen date between week 32 and week 35 2021	Total	Unlinked*	Rates among persons vaccinated with 2 doses (per 100,000)	Rates among persons not vaccinated (per 100,000)
Under 18	167,832	15,901	476.0	1,192.9
18-29	176,392	19,529	711.1	1,520.8
30-39	113,373	12,452	782.2	1,143.9
40-49	97,881	8,930	1,116.2	880.4
50-59	84,488	6,868	962.0	729.7
60-69	45,252	3,657	672.3	487.5
70-79	25,499	2,034	480.5	367.5
80+	12,011	1,124	391.1	427.4

Public Health England data on COVID infection rates

- Rate appears higher in vaccinated patients at some age groups

COVID-19 vaccine surveillance report – week 36

Table 4. COVID-19 cases by vaccination status betw

Cases reported by week of specimen date between week 32 and week 35 2021	Total	Unli		
Under 18	167,832	15,901	476.0	1,192.9
18-29	176,392	19,529	711.1	1,520.8
30-39	113,373	12,452	782.2	1,143.9
40-49	97,881	8,930	1,116.2	880.4
50-59	84,488	6,868	962.0	729.7
60-69	45,252	3,657	672.3	487.5
70-79	25,499	2,034	480.5	367.5
80+	12,011	1,124	391.1	427.4

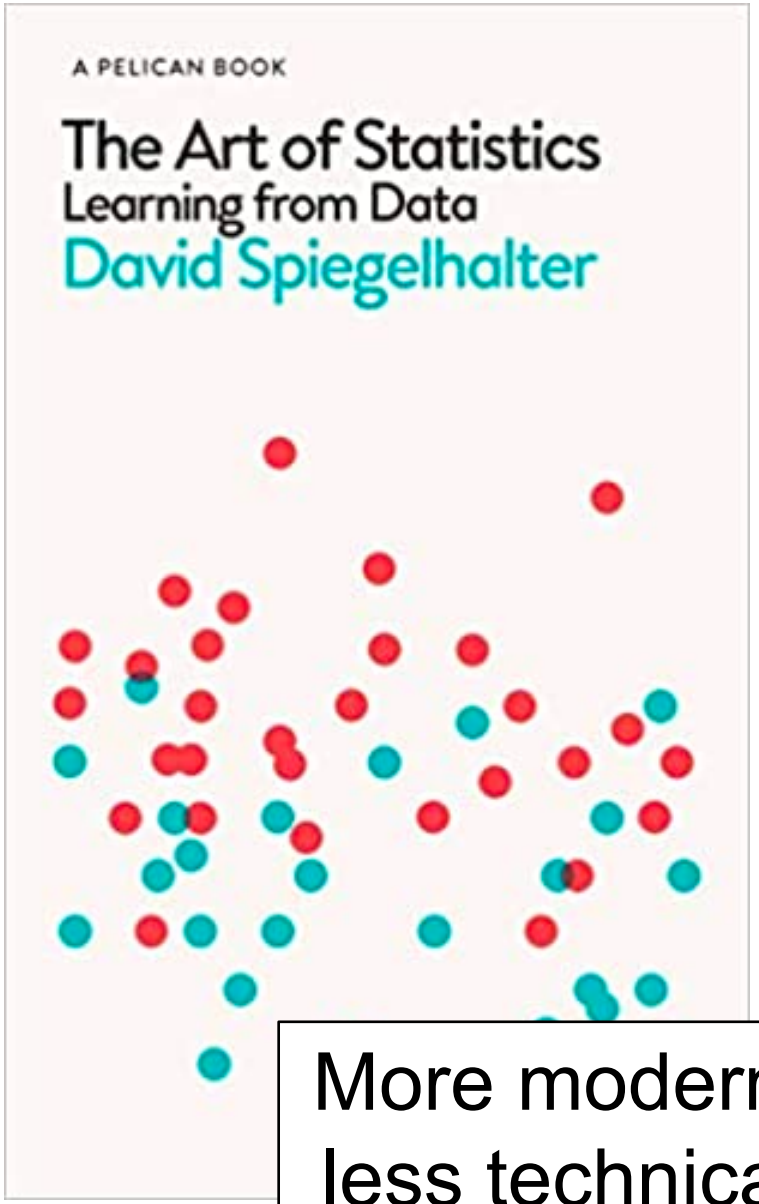
See (listen to): 'More or Less' 15th September 2021, https://www.bbc.co.uk/sounds/play/p09vyy6h

General Resources

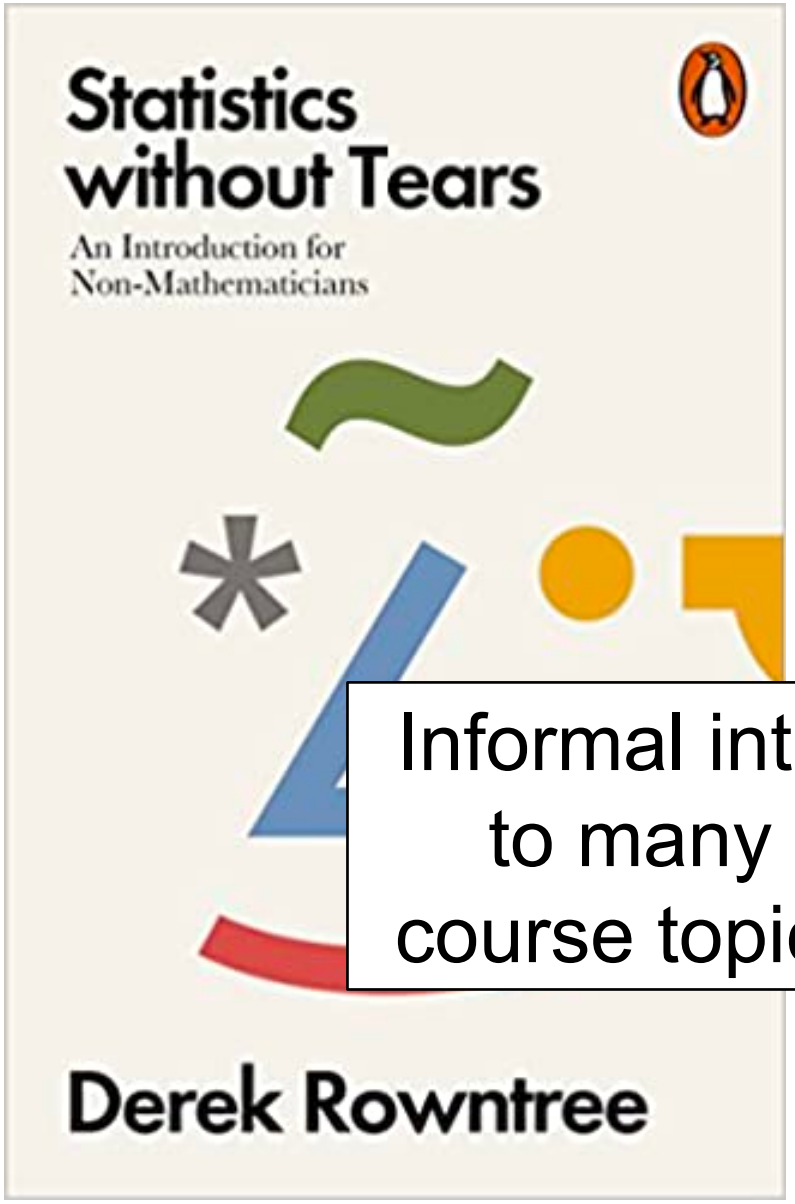
- Radio program on understanding ‘numbers’
  - New series in Jan 2021
  - Outside UK: may need VPN
- Poor communication (or poor statistics) can mislead

https://www.bbc.co.uk/programmes/b006qshd

General Resources



More modern, less technical



Informal intro to many course topics

Summary

- Statistics is about the reliable interpretation of data
  - Is a variation real?
  - Can we explain the variation?
- Data analysis is transforming and understanding data
- Good data analysts need some statistical knowledge (even if it is not often used)

5 Min Breakout

Introduce yourself to other students

Collaborate on any questions