Data, Politics and Society

W5 – Data and the Environment

UCL

Where we at?

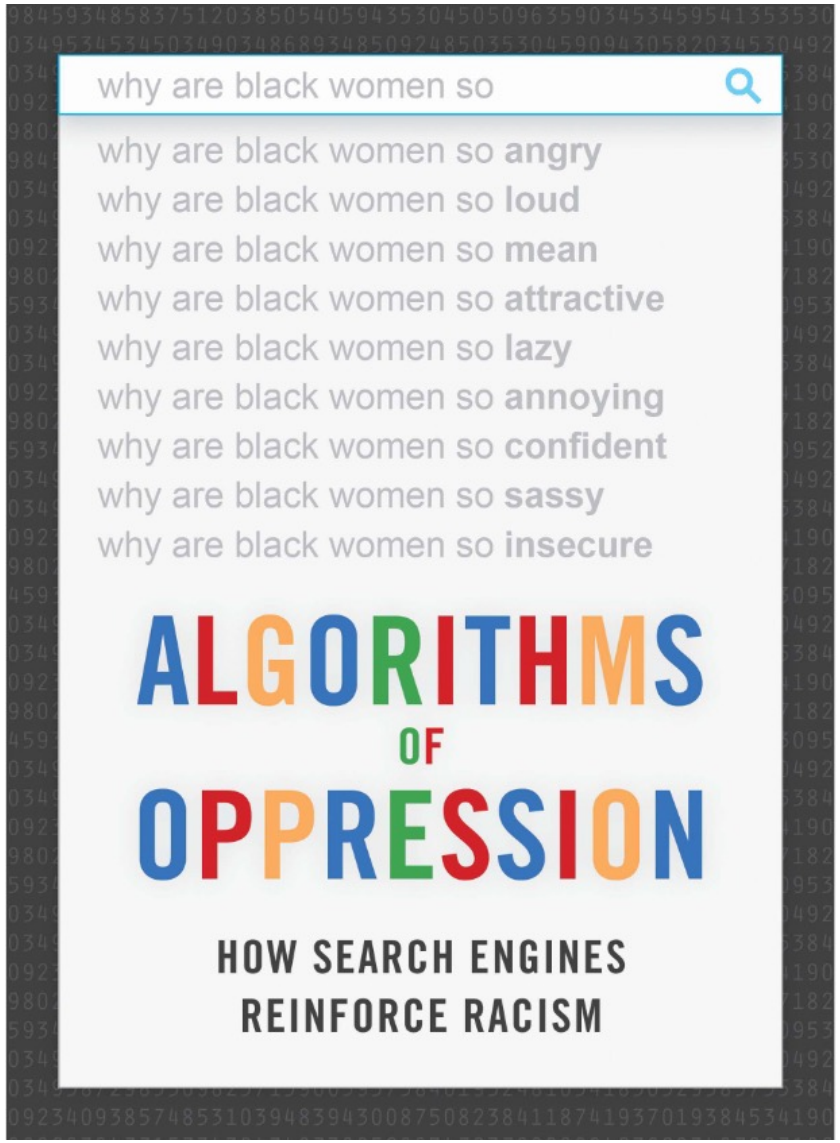
W1
W2
W3

Data: The Good, The Bad, The Ugly

W4
W5

Societal and environmental impacts of data and technology

Algorithms of oppression

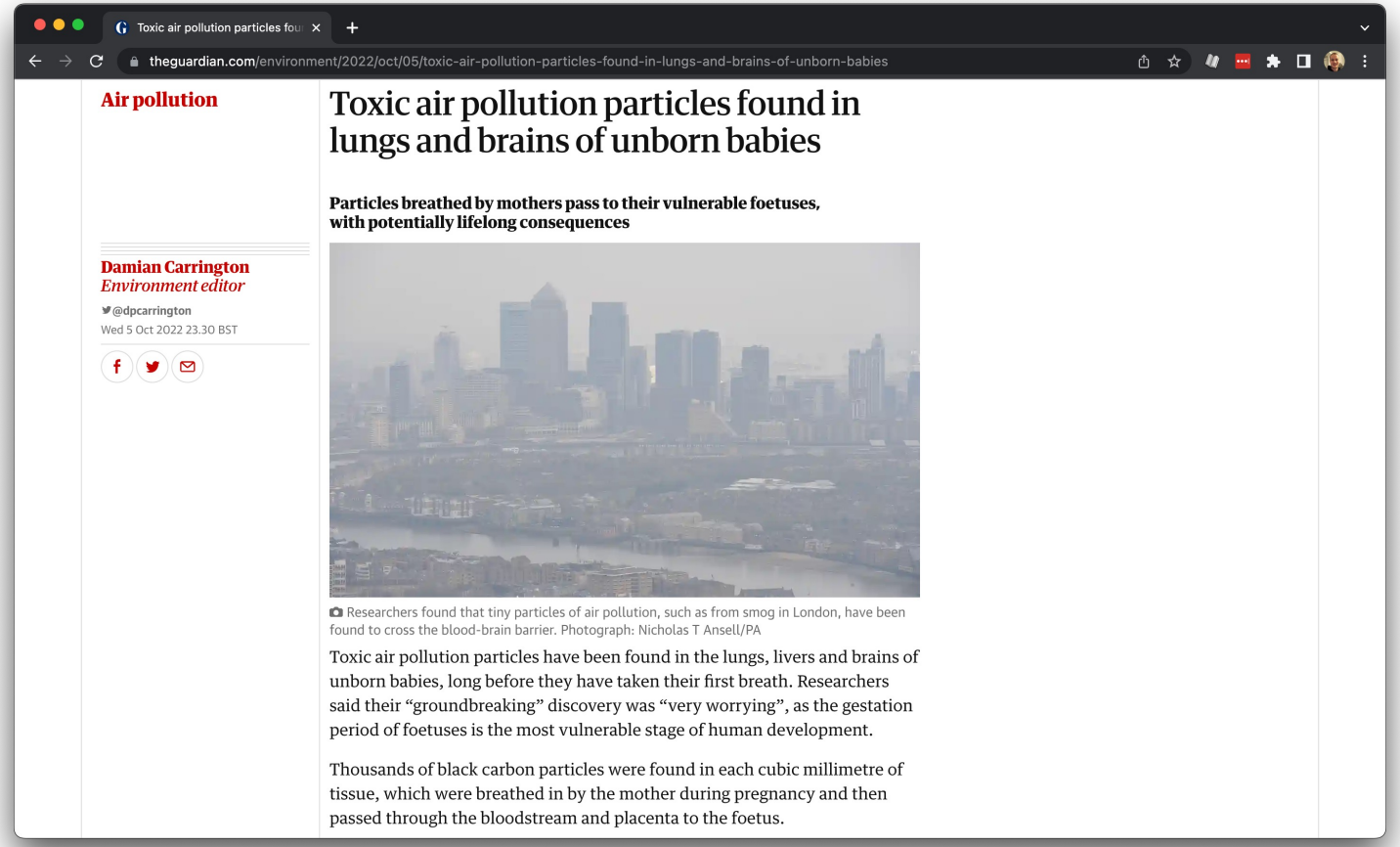


Safiya Noble 2020

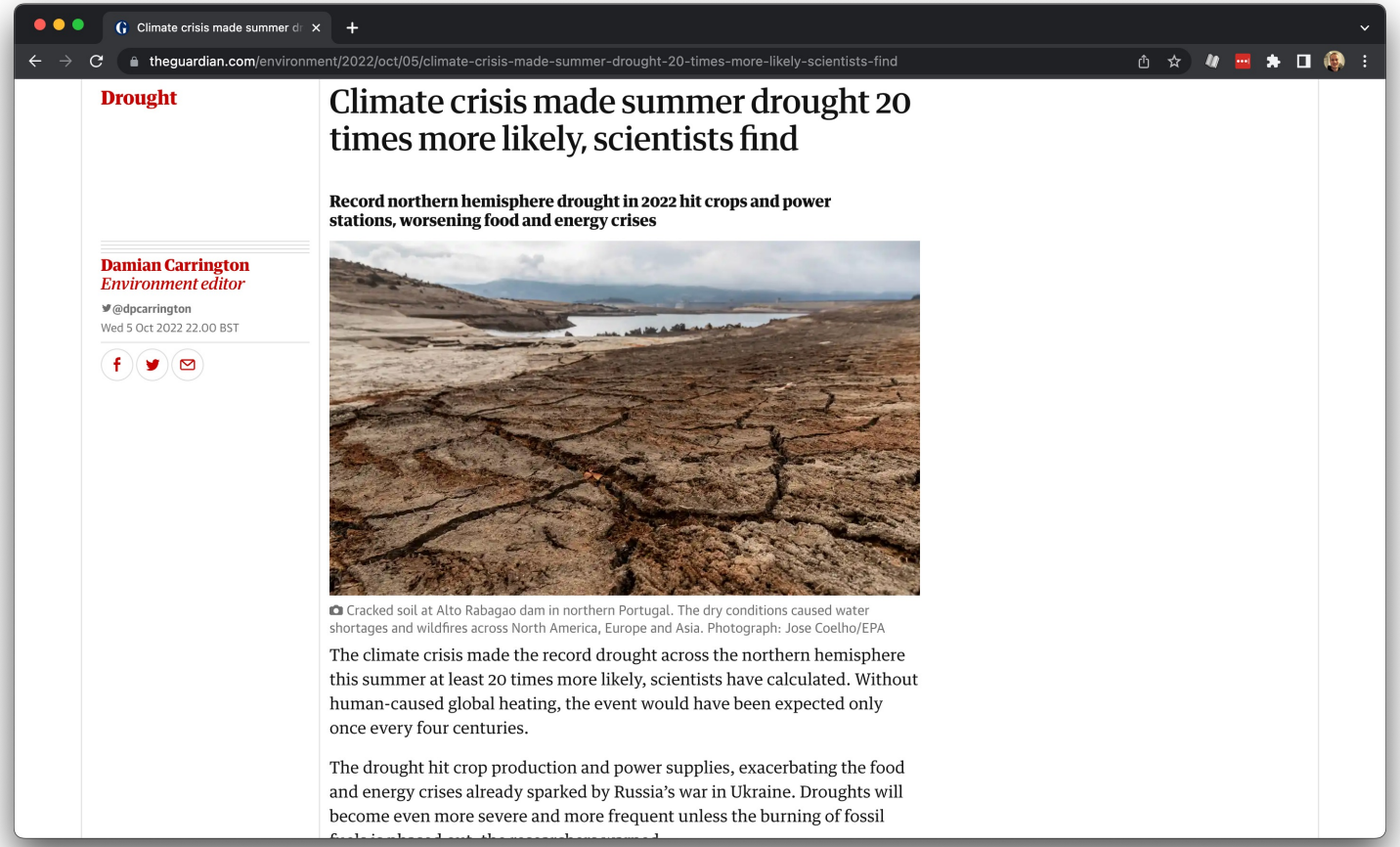
Today

- Climate crisis
- Role of data and technology
- Responsible data science

Climate crisis



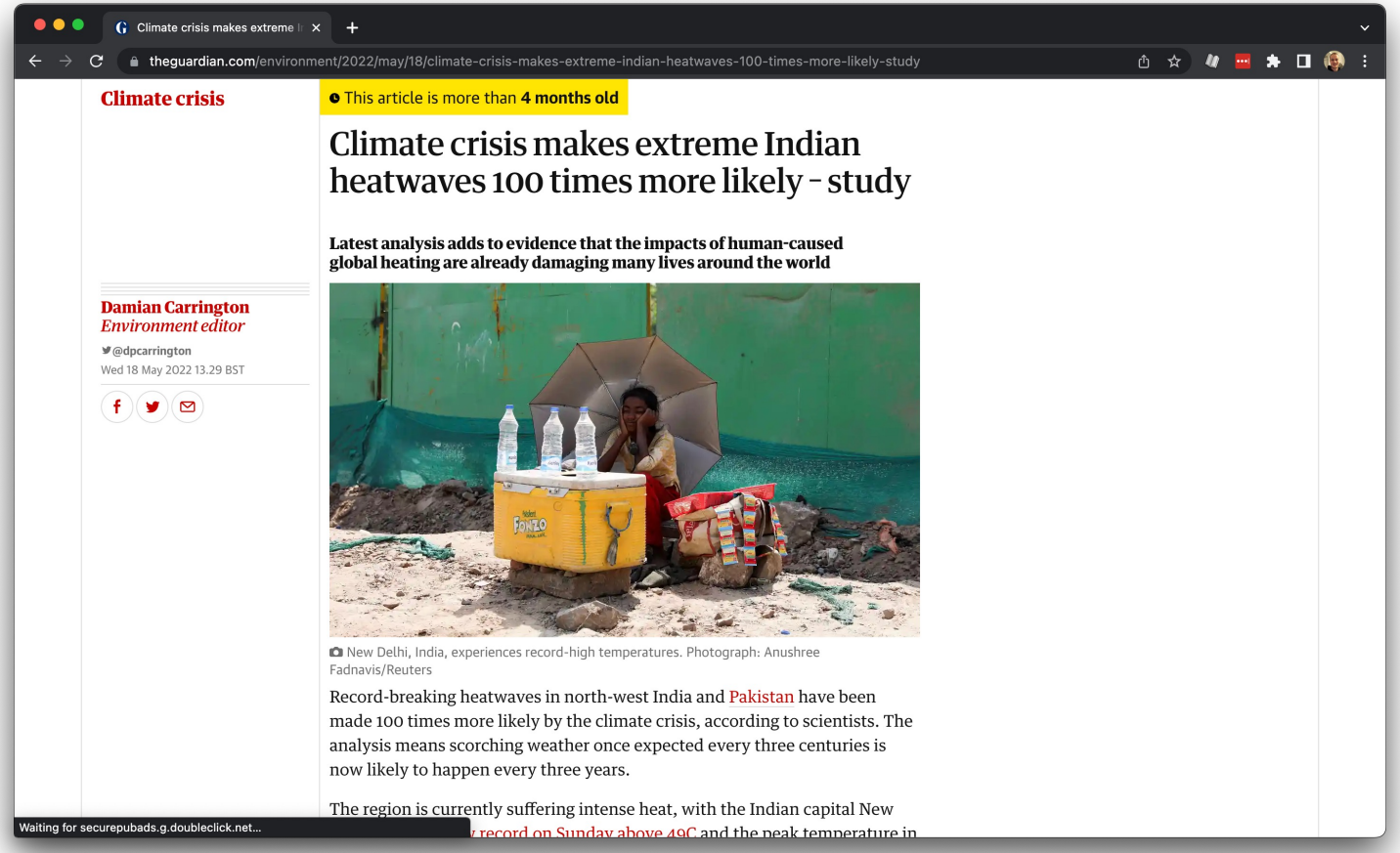
Climate crisis



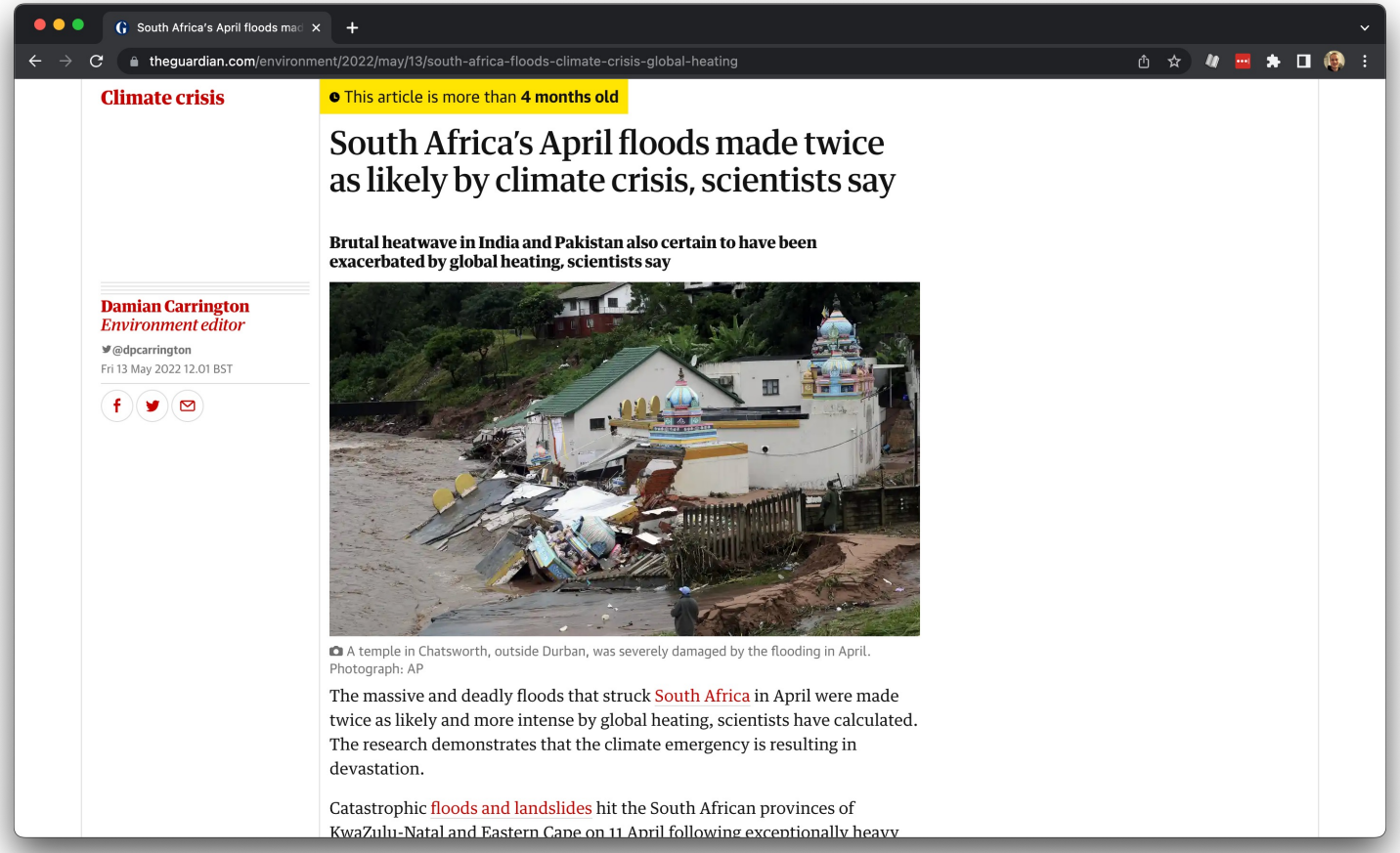
Climate crisis

Climate crisis

Climate crisis

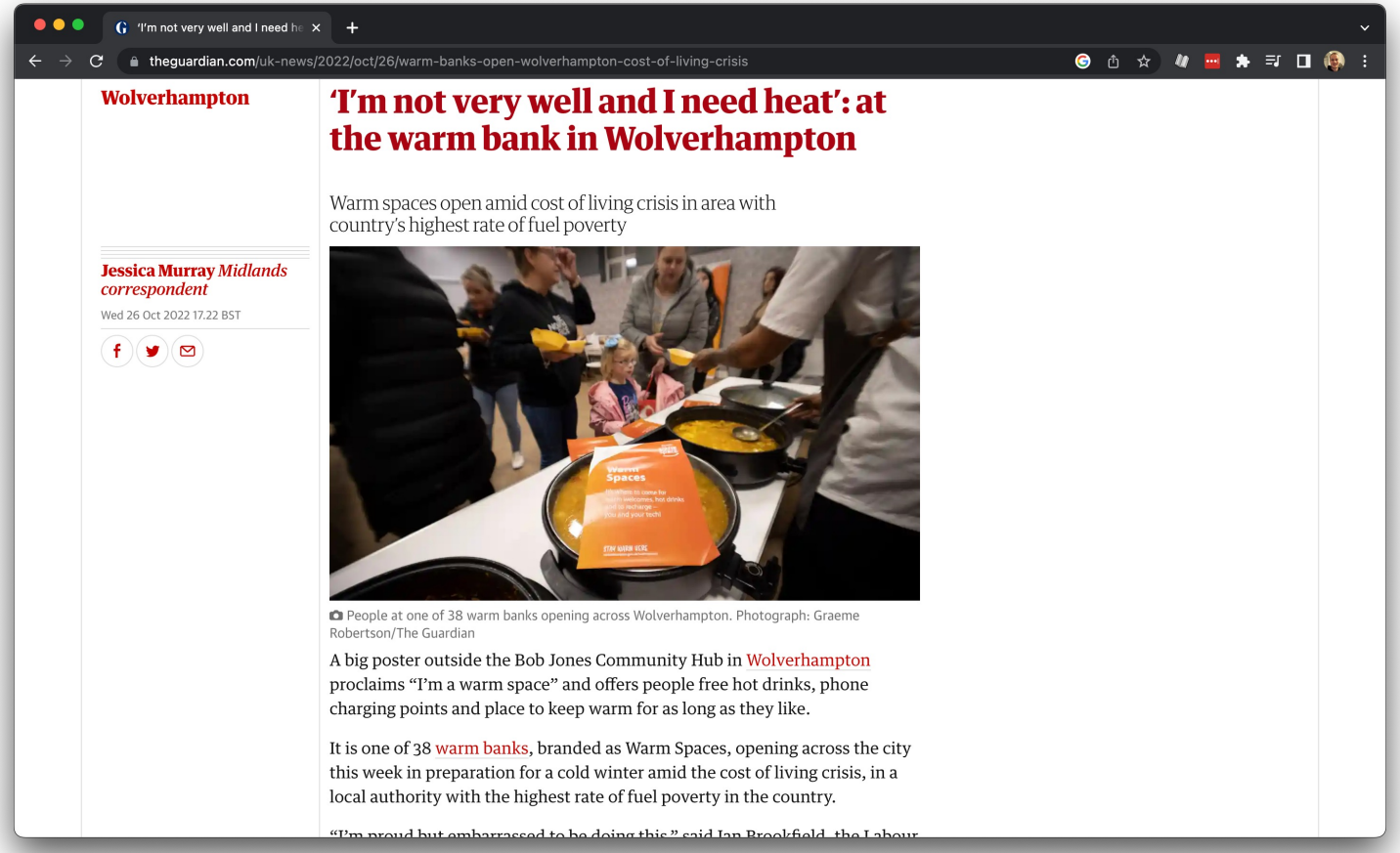


Climate crisis



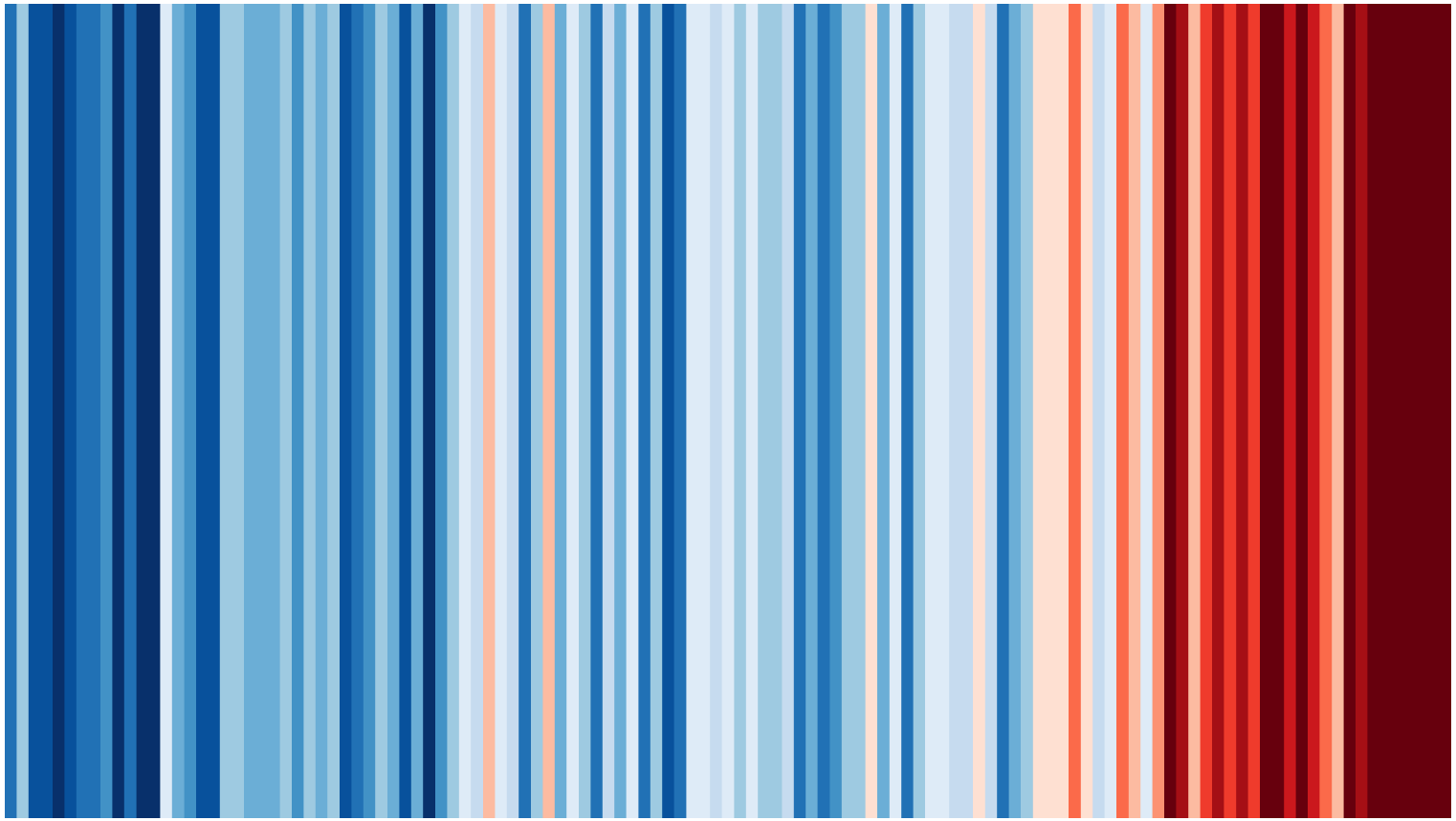
Climate crisis

Climate crisis



Climate crisis

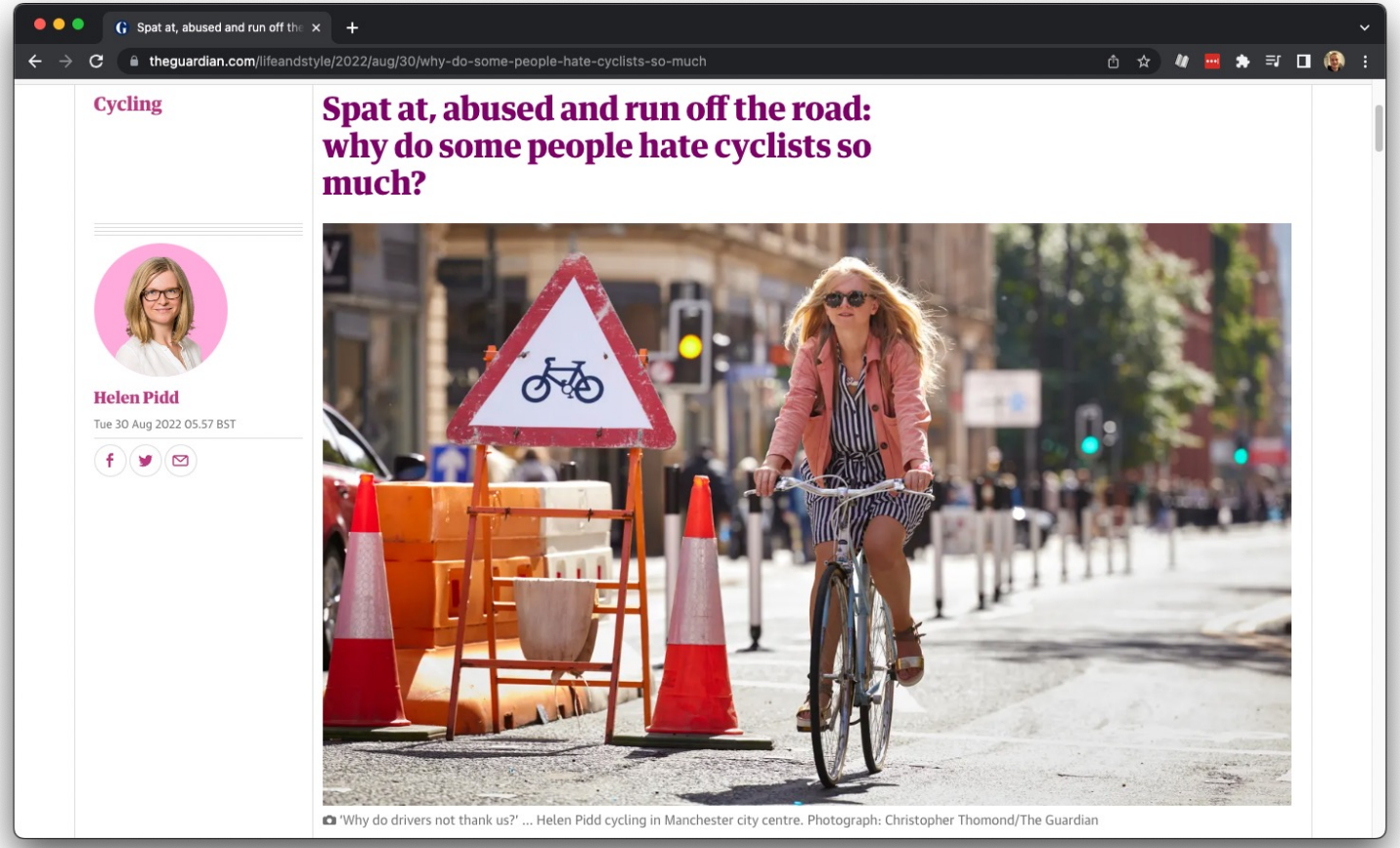
Climate crisis



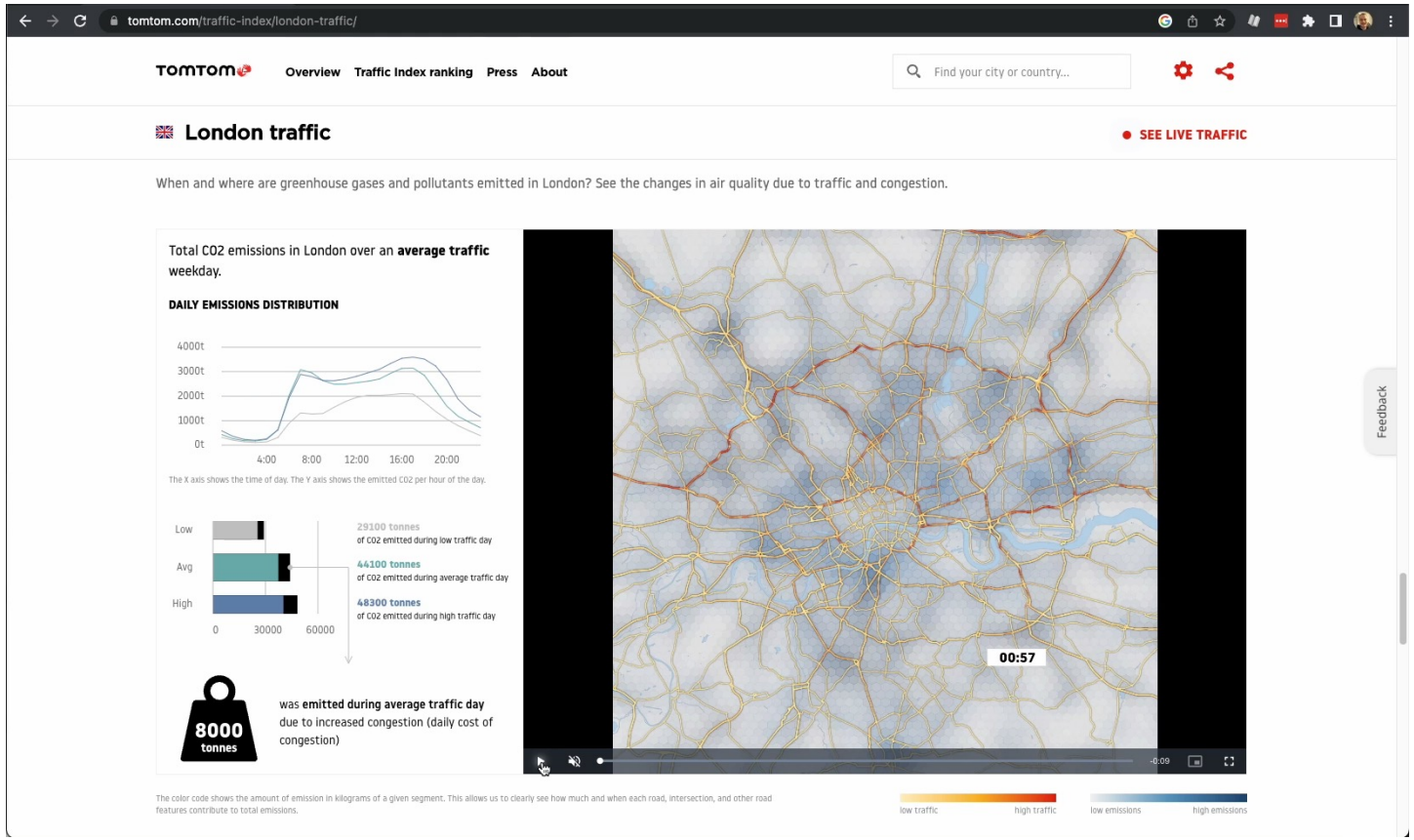
showyourstripes.info, created by Ed Hawkins

Local resistance

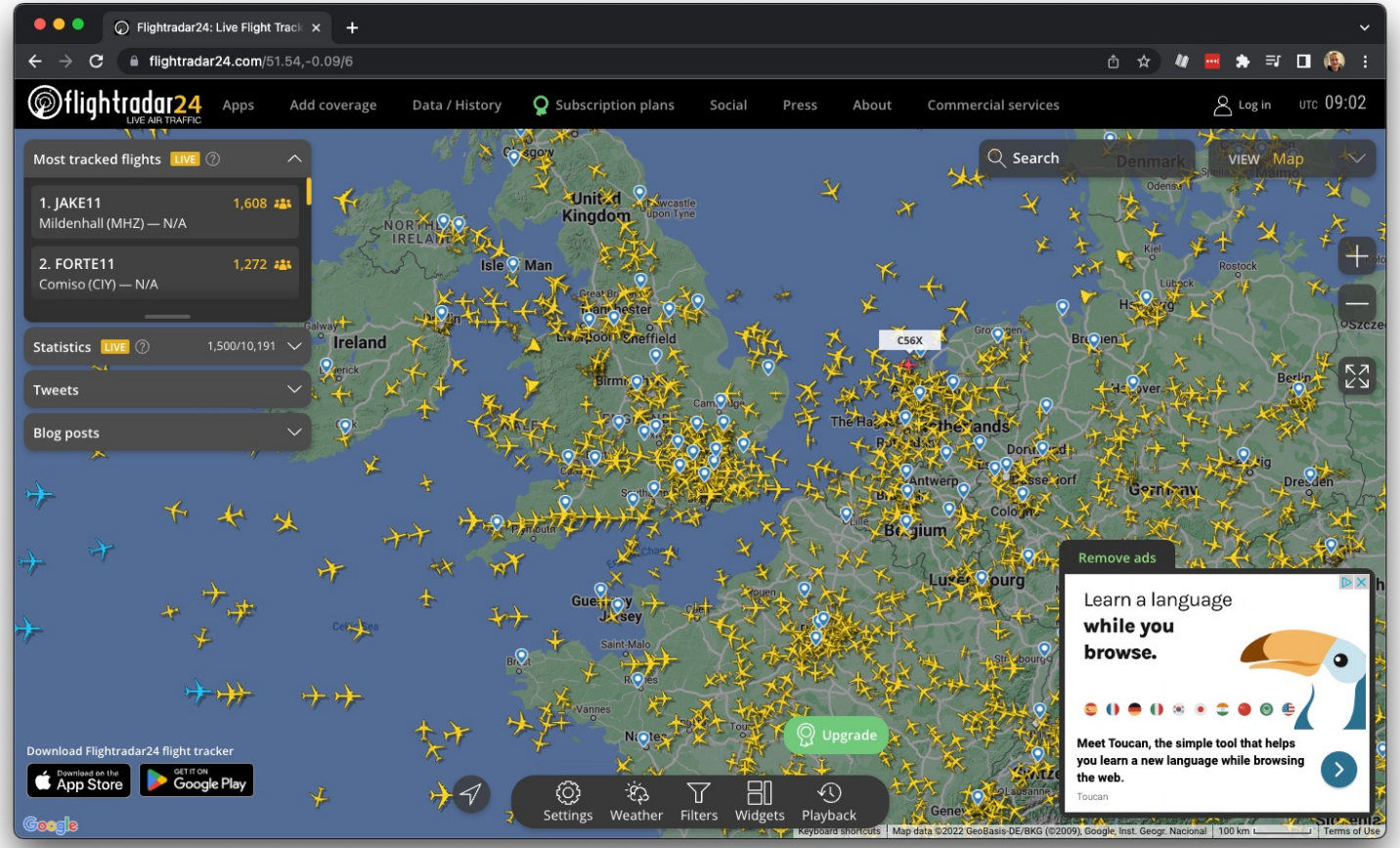
Local resistance



Climate crisis?



Climate crisis?



Data and technology

Although the transport sector is still one of the larger contributors to greenhouse gas emissions, the climate impact of information and communication technologies has been rapidly increasing.

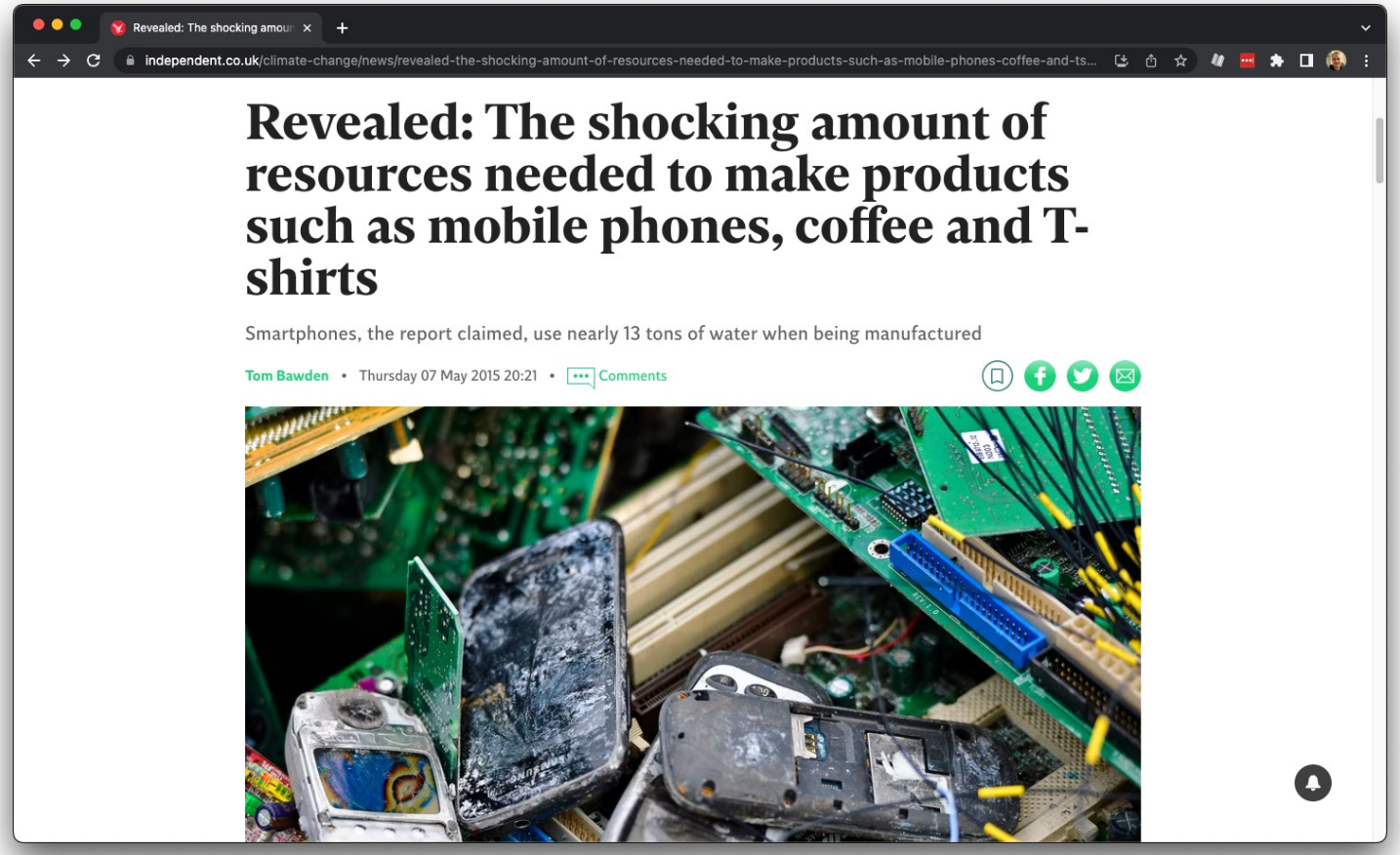
Data and technology

Data and technology

Data and technology

- Natural resources
- Data centres responsible for data storage and processing

Natural resources



Data centres

Data centres

- Siddik *et al.* 2021
- Data are stored in large data centers.
- Large amounts of energy required: 1.8% of electricity use in the US.
- Large amounts of greenhouse gas emissions: 0.5% of total emissions in US.
- Energy demand is between 15–100 times as large as those of typical commercial buildings.

Data centres

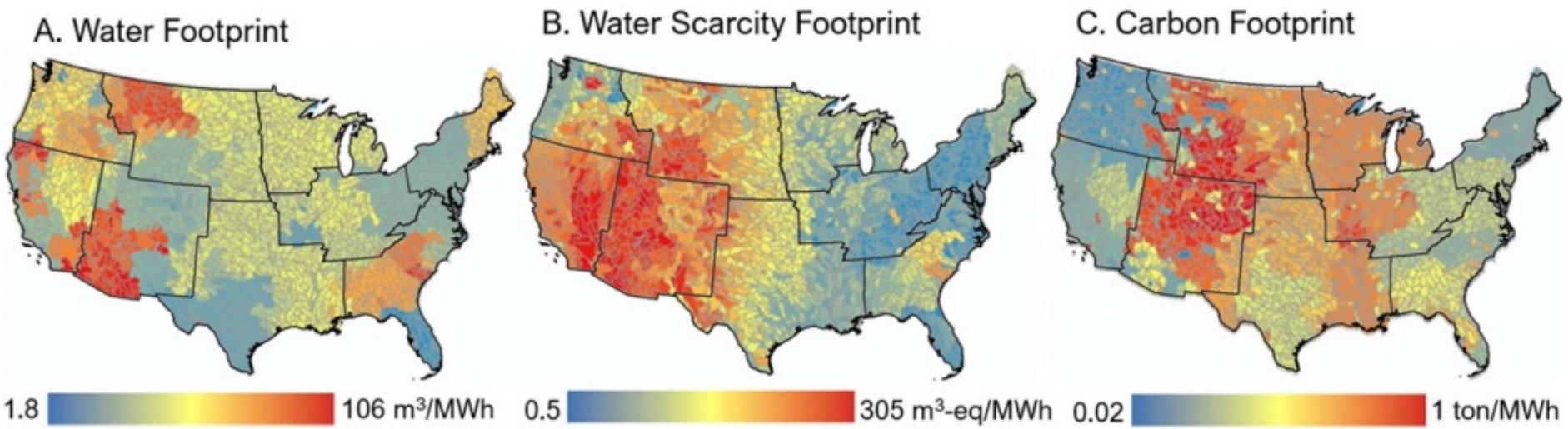


Figure 4. A data center's environmental footprint is highly contingent on where it is located. The (A) water intensity (m^3 MWh^{-1}), (B) water scarcity intensity (m^3 US-eq MWh^{-1}), and (C) GHG emissions intensity (tons CO_2-eq MWh^{-1}) of a hypothetical 1 MW data center placed in each of the 2110 subbasins of the continental United States.

Data centres

Siddik *et al*. 2021:

“Though the amount of data center computing workloads has increased nearly 550% between 2010 and 2018, data center electricity consumption has only risen by 6% due to dramatic improvements in energy efficiency and storage-drive density across the industry. However, it is unclear whether energy efficiency improvements can continue to offset the energy demand of data centers as the industry is expected to continue its rapid expansion over the next decade.”

Data centres

- Current estimate that the tech sector contribute around 3% of global greenhouse, of which around 45% can be attributed to data centers.
- At the same time: lack of transparency on the share of green and dirty energy being used.
- Some estimates suggest that the wider tech industry's carbon footprint could increase to 14% of global emissions by 2040; think crypto currency, energy intensive 5G networks, autonomous vehicles, Internet of Things.

Data centres

But it is not only the storage and access of data; data processing with data hungry and computationally intensive models is having a growing influence.

Data centres

- Strübel *et al.* 2019
- Training a state-of-the-art model now requires substantial computational resources which demand considerable energy.
- Research and development of new models multiplies these costs by thousands of times by requiring retraining to experiment with model architectures and hyperparameters.
- Estimated carbon footprint from training a large Natural Language Processing model: 300,000 kilograms.

Data centres

Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Data centres

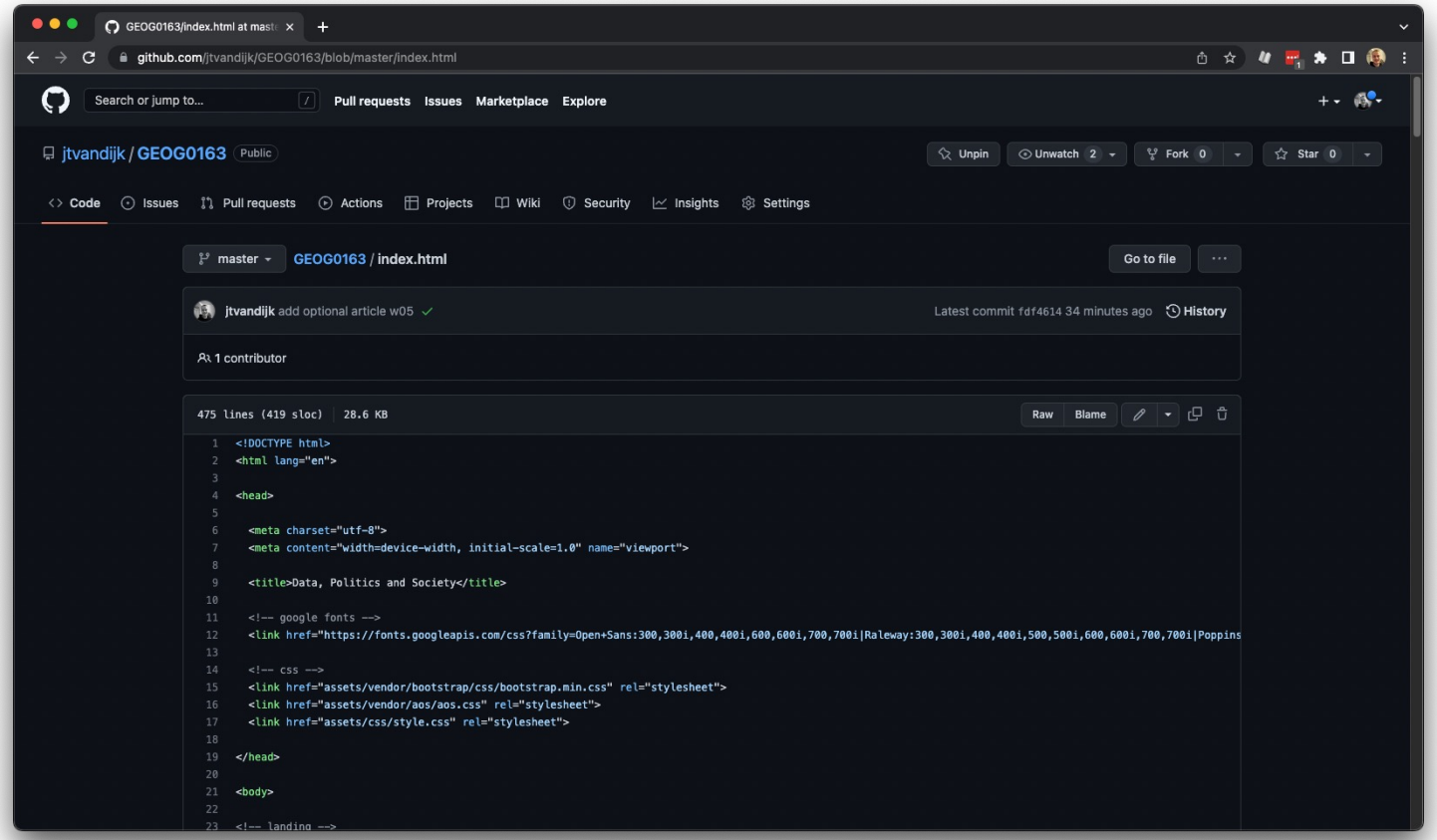
Consumer	Renew.	Gas	Coal	Nuc.
China	22%	3%	65%	4%
Germany	40%	7%	38%	13%
United States	17%	35%	27%	19%
Amazon-AWS	17%	24%	30%	26%
Google	56%	14%	15%	10%
Microsoft	32%	23%	31%	10%

Table 2: Percent energy sourced from: Renewable (e.g. hydro, solar, wind), natural gas, coal and nuclear for the top 3 cloud compute providers (Cook et al., 2017), compared to the United States,[4] China[5] and Germany (Burger, 2019).

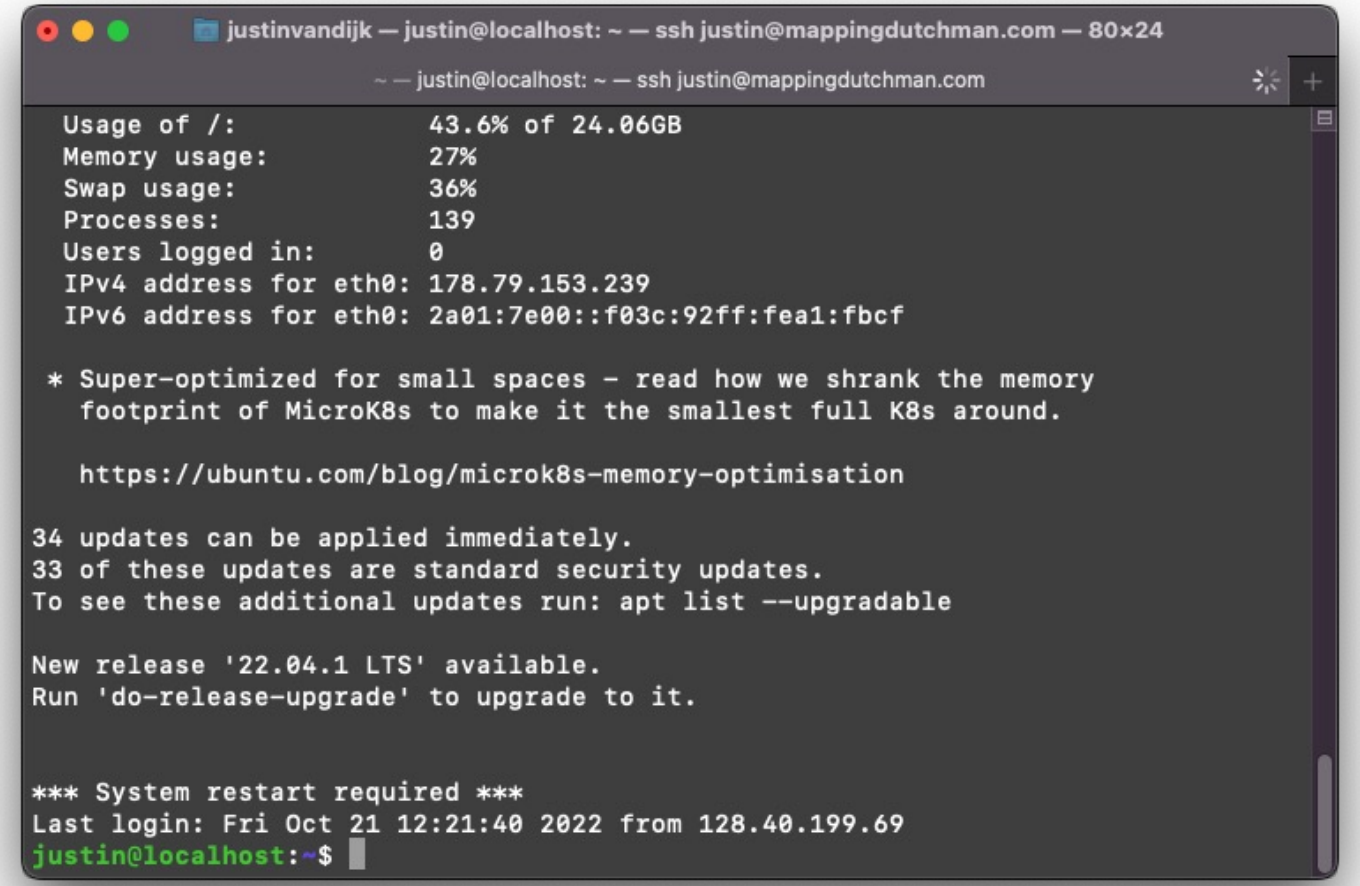
Data processing tools

- Increased accessibility in terms of using cloud-distributed computing and move away from laptop/desktop-based calculations.
- Lots of tools are part of a typical data science workflow.

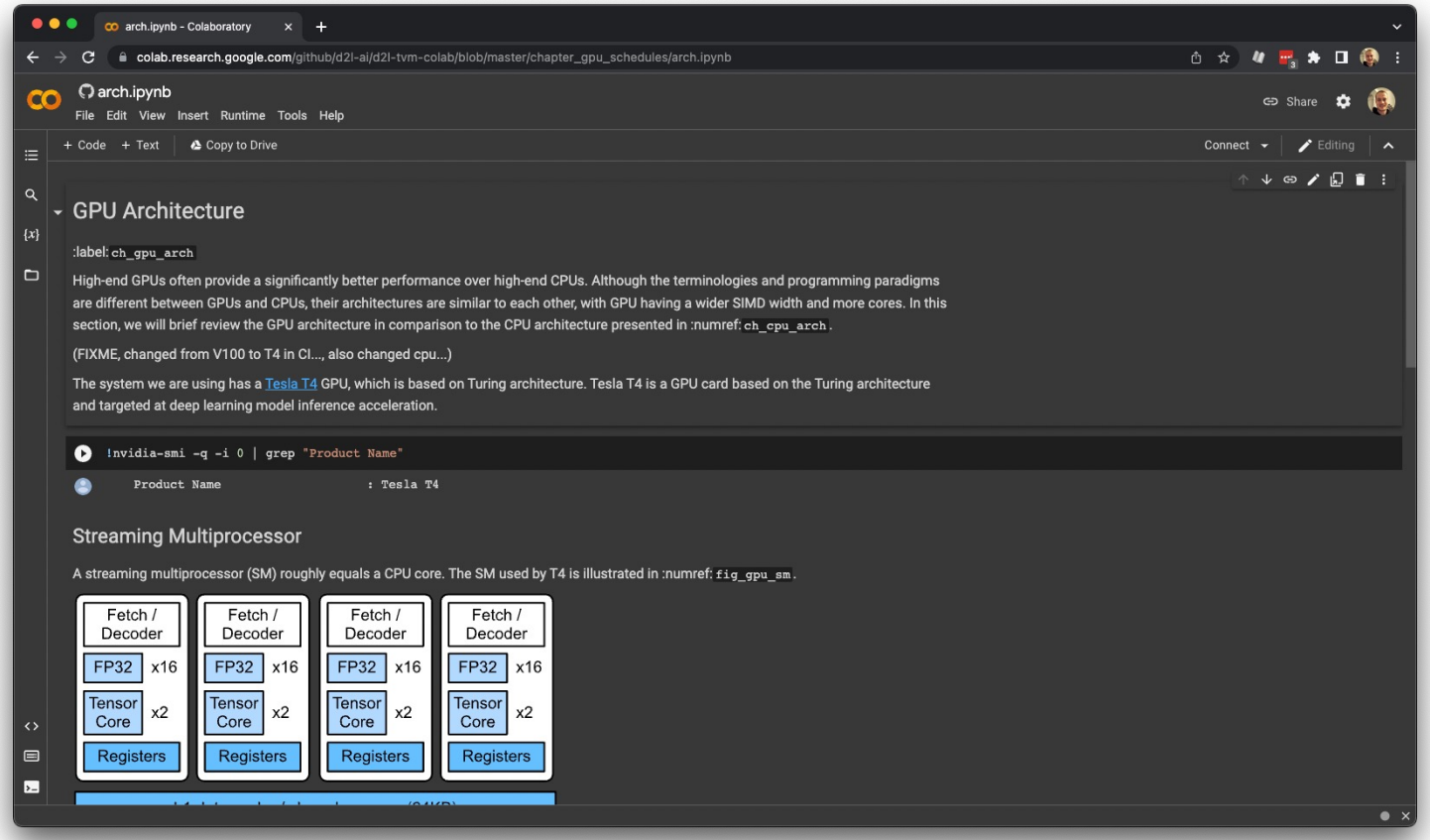
Data processing tools



Data processing tools



Data processing tools



What can we do?

What else can we do in a research context to reduce our environmental impact?

What can we do?

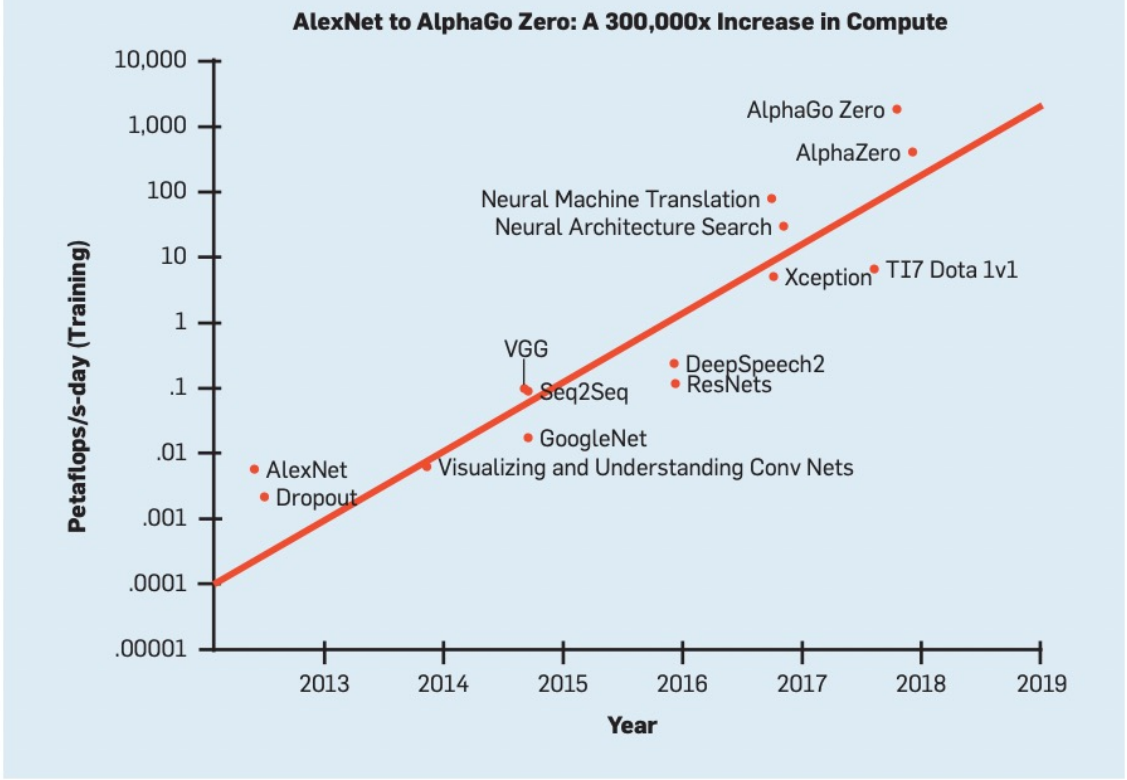
- Strübel *et al.* 2019
- Authors should report training time and sensitivity to hyperparameters.
- Academic researchers need equitable access to computation resources.
- Researchers should prioritize computationally efficient hardware and algorithms.

What can we do?

- Improvements in the field of AI: object recognition, game playing, speech recognition, and machine translation.
- The computational costs of state-of-the art AI research has increased 300,000x recent years: prioritising accuracy over efficiency. Red AI.

What can we do?

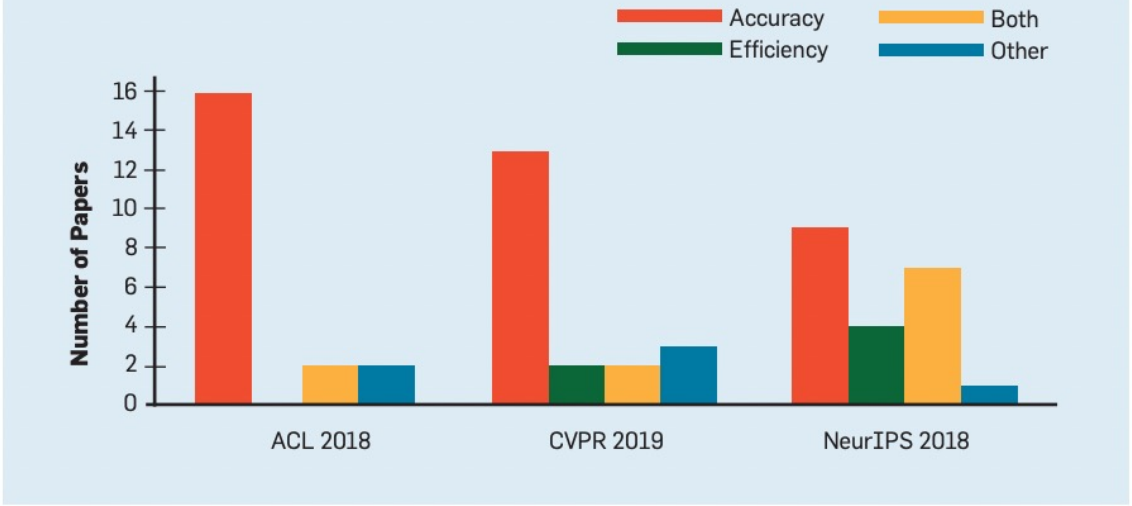
Figure 1. The amount of compute used to train deep learning models has increased 300,000x in six years. Figure taken from Amodei et al.[2]



Schwartz *et al.* 2020

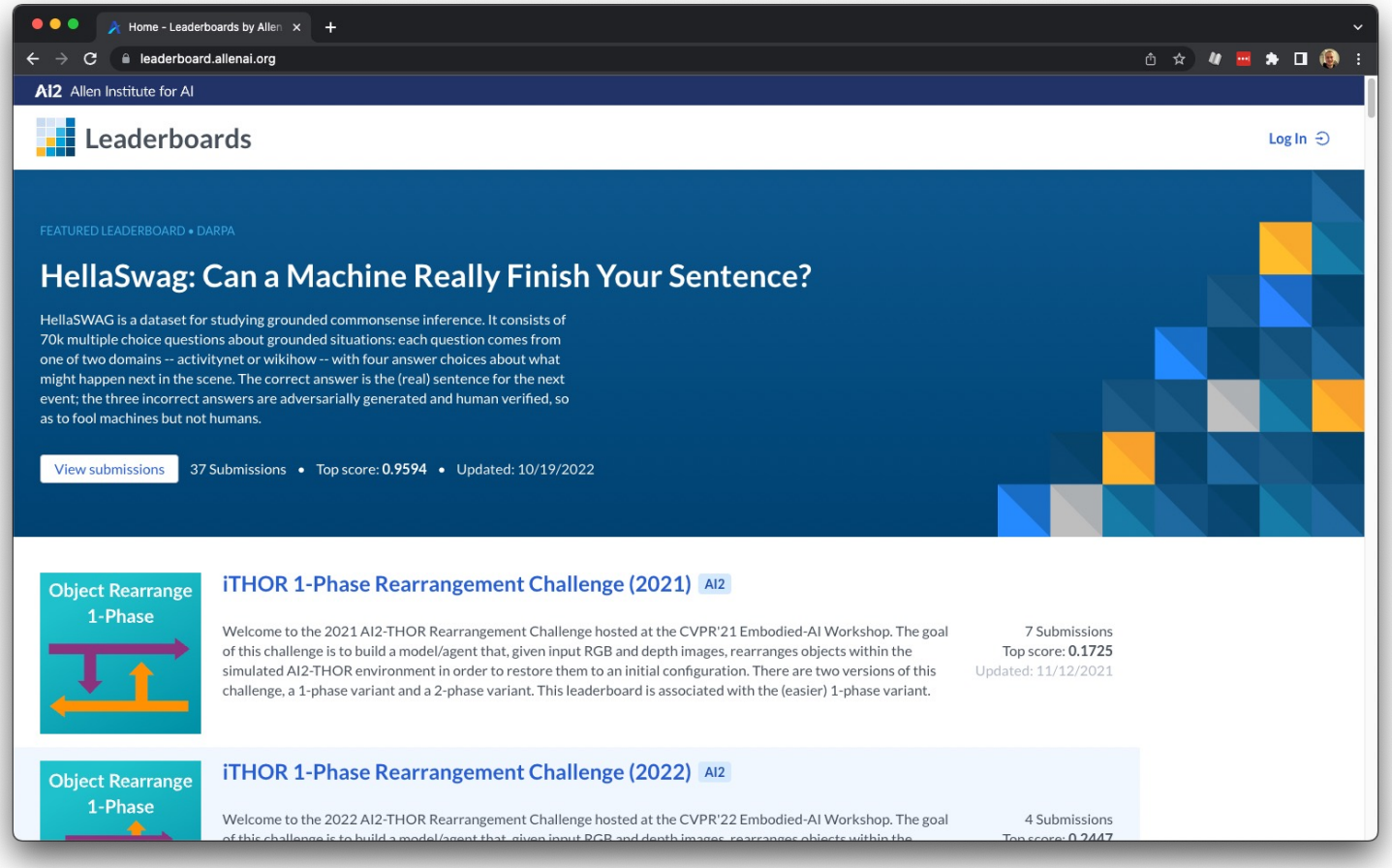
What can we do?

Figure 2. AI papers tend to target accuracy rather than efficiency. The figure shows the proportion of papers that target accuracy, efficiency, both or other from a random sample of 60 papers from top AI conferences.



Schwartz *et al.* 2020

What can we do?



What can we do?

- Transfer learning
- Efficiency of coding language
- Be aware of diminishing returns
- Together: Green AI?

How to measure efficiency?

- Carbon emissions
- Electricity usage
- Elapsed real time
- Number of parameters
- Reporting on number of floating-point operations
- Workflow?

What can we do?

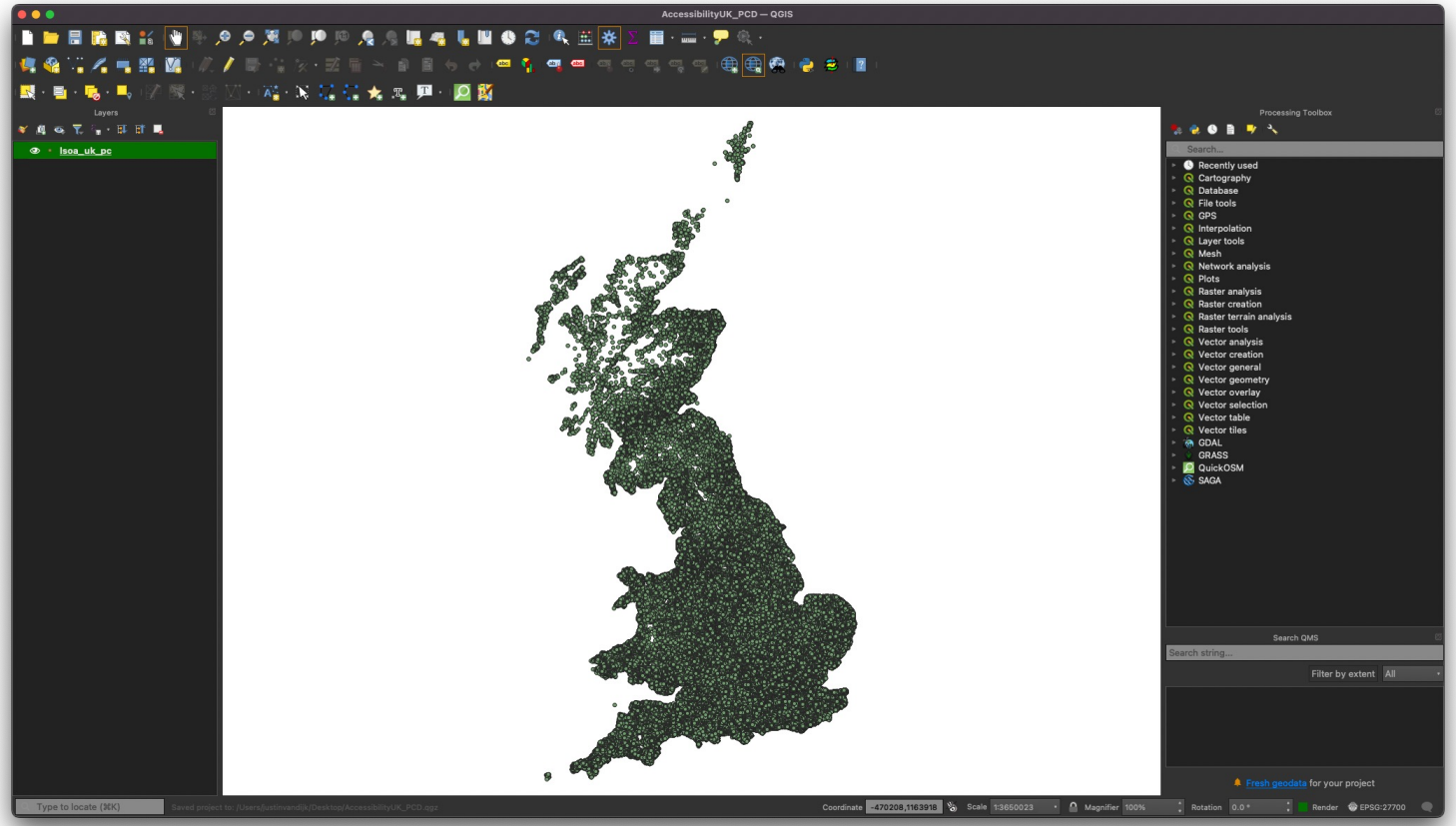
Task

- For each Output Area in Great Britain, calculate the average accessibility by unit postcode to the nearest supermarkets or vegetable shops with fresh foods.

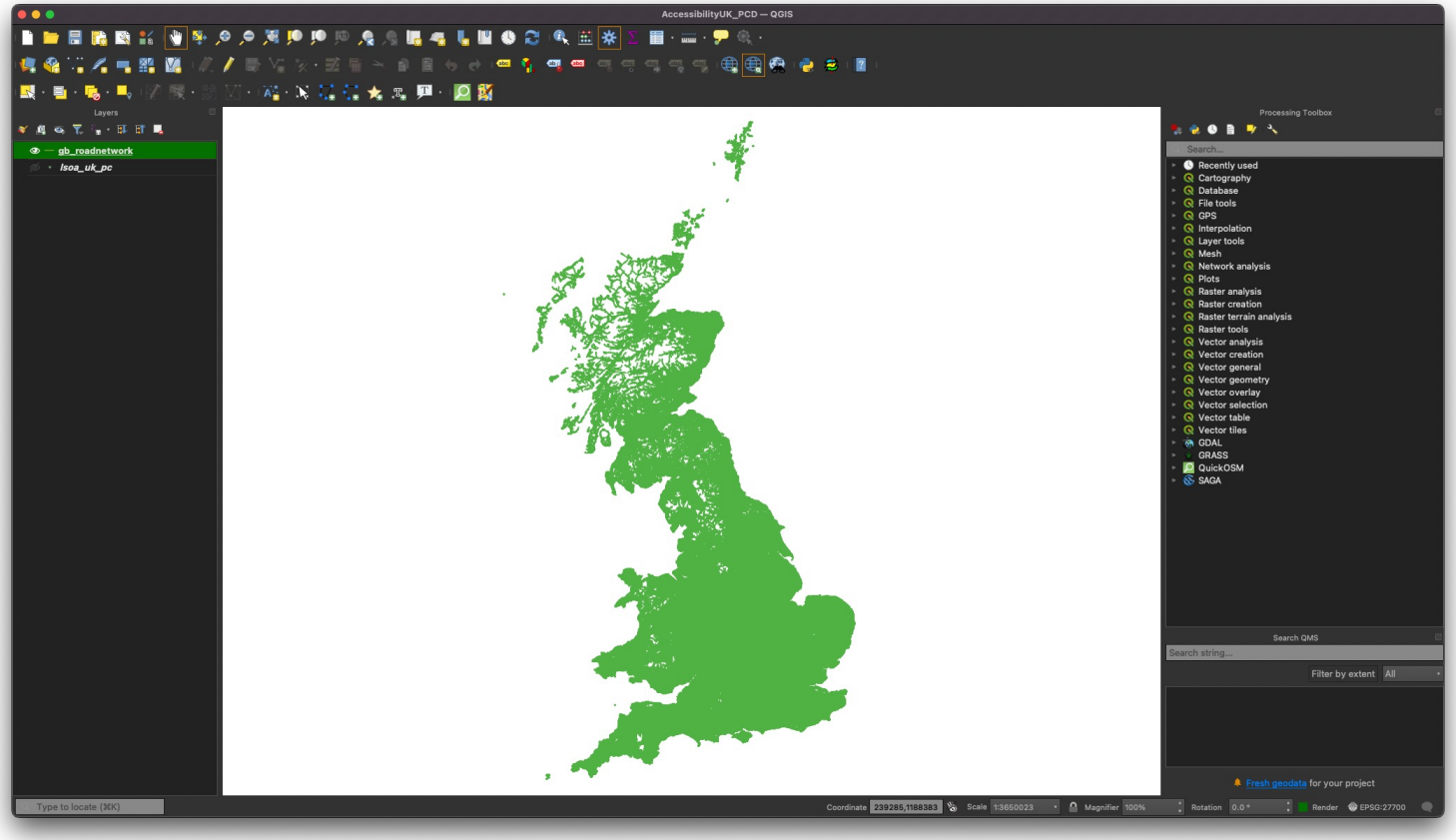
Data

- 2.5 million postcodes in Great Britain
- 3.7 million individual road segments in Great Britain

What can we do?



What can we do?



What can we do?

Take 3 minutes to discuss with your neighbour how you would approach this with the idea of efficiency in mind.

What can we do?

To start with:

- Confirm aims and objectives (e.g. single transport mode or multimodal analysis).
- Identify a data source that can be used to get an idea of the number of shops?
- Define how to measure accessibility. Travel time? If so, what transport mode?

What can we do?

A possible workflow:

- Identify small case study area #1.
- Develop and test your code, including sensitivity analysis and areal aggregation.
- Randomly select case study area #2.
- Test your code to see if no unexpected errors show up.
- Discuss case study results with colleagues / supervisor / manager.
- Run code over your entire dataset.

Conclusion

- The climate crisis is real and it is here.
- Massive impact of the ”information society” not only in terms of natural resources needed for chips and infrastructure, but also in terms of electricity and water required for data storage and processing – heavy carbon footprint.
- Responsibility as a data scientist / researcher to maximise efficiency of analytical work; even at small scale simple strategies can be employed.

Recap

- Part I: Data and its role in Society.
- Mainly: raised issues and concerns, “taking stock”.
- Fundamental question: what are data?

- Part II: Mitigating the risks of working with large-scale datasets.
- Mainly: what are “we” currently doing to address some of these issues and concerns.
- Fundamental question: is what ”we are doing” enough?

Seminar preparation

- This week we will be holding a formal debate. A formal debate involves two sides: one supporting a resolution and one opposing it.
- This week you will be debating the resolution: There is no such thing as Green Data.
- One half of the Seminar group will be in support of this resolution, the other half will be opposing this. Each of you has been assigned to either one of the sides; Excel spreadsheet available on GitHub.

Seminar preparation

- You are asked, as a group, to prepare arguments in advance.
- Divide and conquer with your team mates - e.g. divide different domains and have individuals focus on different parts of the resolution.
- Think about statistics, case studies, and examples that you can use to support your argument or that you can use to counteract the other team's arguments.

How are we going to do it?

- Both teams start with 10 minutes each to finalise their arguments and plan for the debate.
- The first speaker on the supporting team presents arguments in support of the resolution (3 mins). *Opening statements, definitions.*
- The first speaker on the opposing team presents arguments opposing the resolution (3 mins). *Opening statements.*

How are we going to do it?

- The second speaker on the supporting team presents further arguments in support of the resolution, identifies areas of conflict, and answers questions that may have been raised by the opposition speaker (3 mins). *Arguments and evidence.*
- The second speaker on the opposing team presents further arguments against the resolution, identifies further areas. of conflict, and answers questions that may have been raised by the previous supporting speaker (3 mins). *Arguments and evidence.*

How are we going to do it?

- Both team have 5 minutes to prepare for the rebuttal.
- The opposing team begins with the rebuttal, attempting to defend the opposing arguments and to defeat the supporting arguments without adding any new information (3 mins).
- The supporting team follows with their rebuttal, attempting to defend the supporting arguments and to defeat the opposing arguments without adding any new information (3 mins).
- Each team gets a second rebuttal for closing statements with the supporting team having the last opportunity to speak (2 mins each).

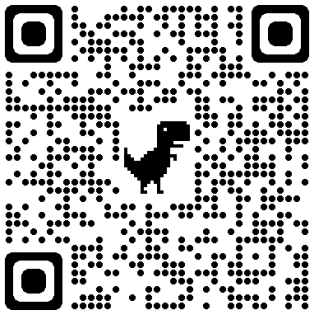
How are we going to do it?

- All speakers are timed by the Seminar leader, you are not allowed to run over time and will be cut off if you do.
- The first two rounds is when you lay out your arguments and introduce your evidence in favour or against the resolution; no new information or evidence should be introduced in the final two rounds.
- There cannot be any interruptions. Speakers must wait their turns.
- Every speaker may only speak once.
- Have a little fun at the last class before reading week

Questions

Justin van Dijk

j.t.vandijk@ucl.ac.uk



Enjoy reading week