Data, Politics and Society

W5 – Data and the Environment

UCL

Where we at?

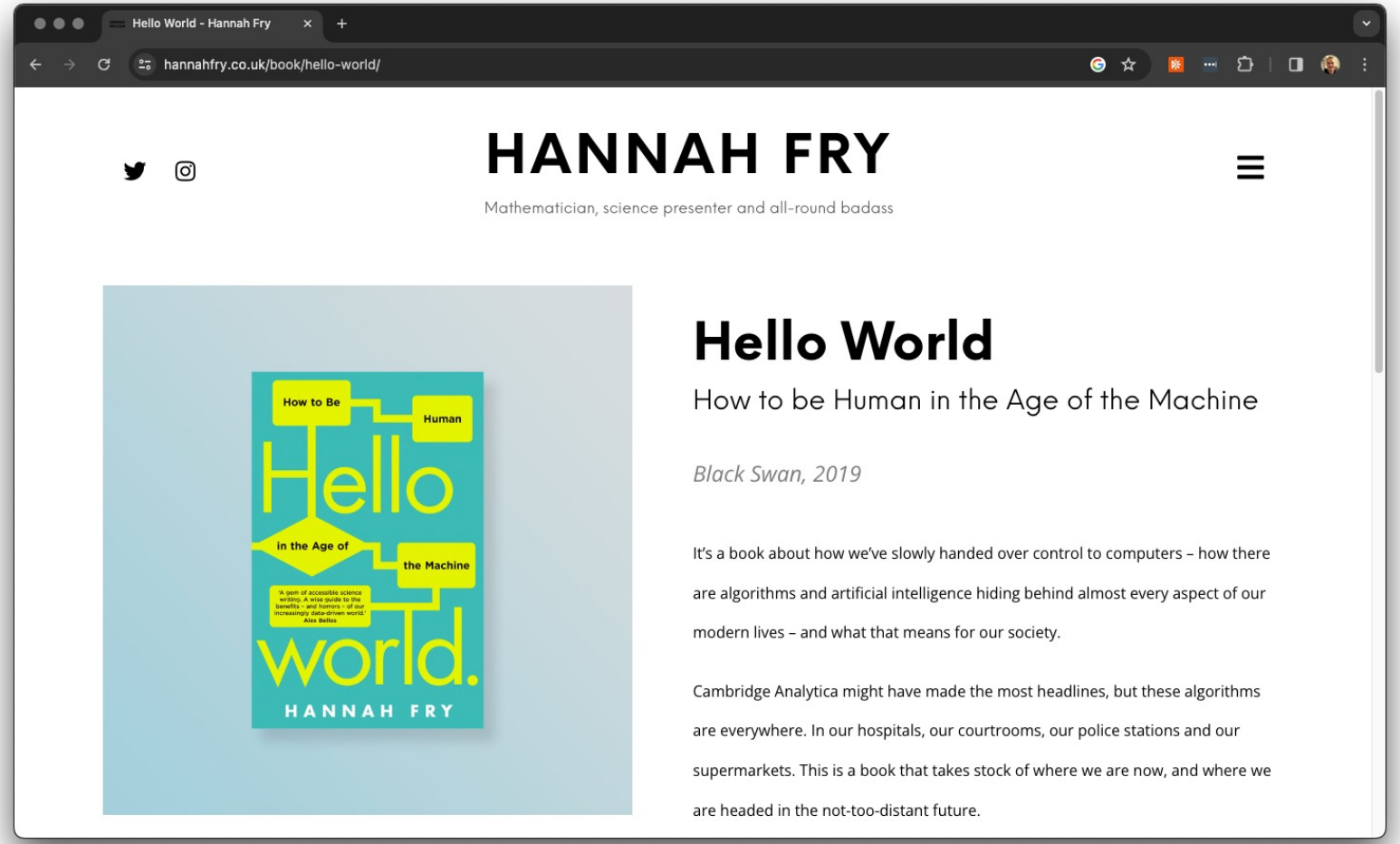
W1
W2
W3

Data: The Good, The Bad, The Ugly

W4
W5

Societal and environmental impacts of data and technology

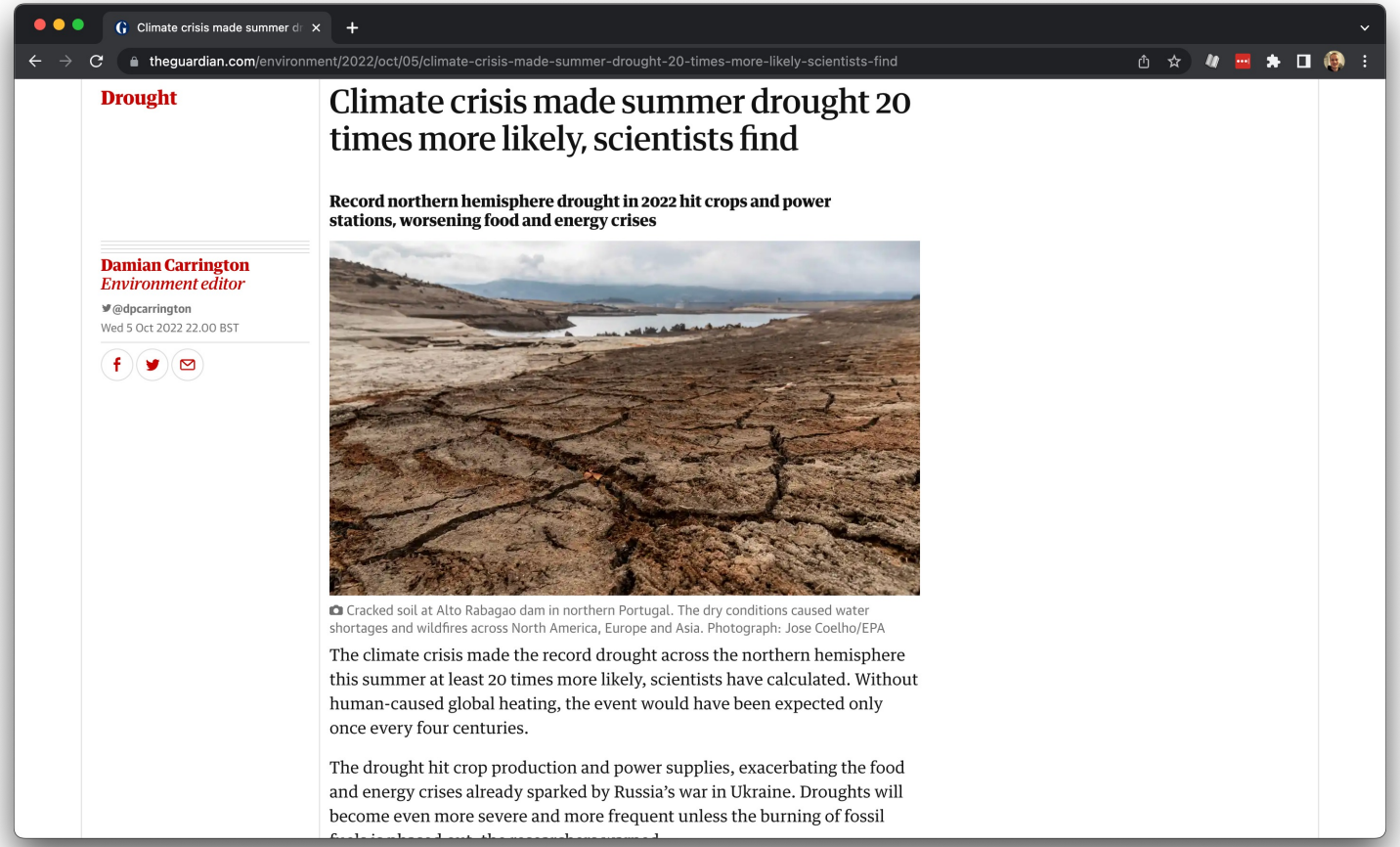
Hello World


Hello World - Hannah Fry
hannahfry.co.uk/book/hello-world/
HANNAH FRY
Mathematician, science presenter and all-round badass
How to Be
Human
Hello
in the Age of
the Machine
world.
HANNAH FRY
Hello World
How to be Human in the Age of the Machine
Black Swan, 2019
It's a book about how we've slowly handed over control to computers – how there are algorithms and artificial intelligence hiding behind almost every aspect of our modern lives – and what that means for our society.
Cambridge Analytica might have made the most headlines, but these algorithms are everywhere. In our hospitals, our courtrooms, our police stations and our supermarkets. This is a book that takes stock of where we are now, and where we are headed in the not-too-distant future.

Today

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

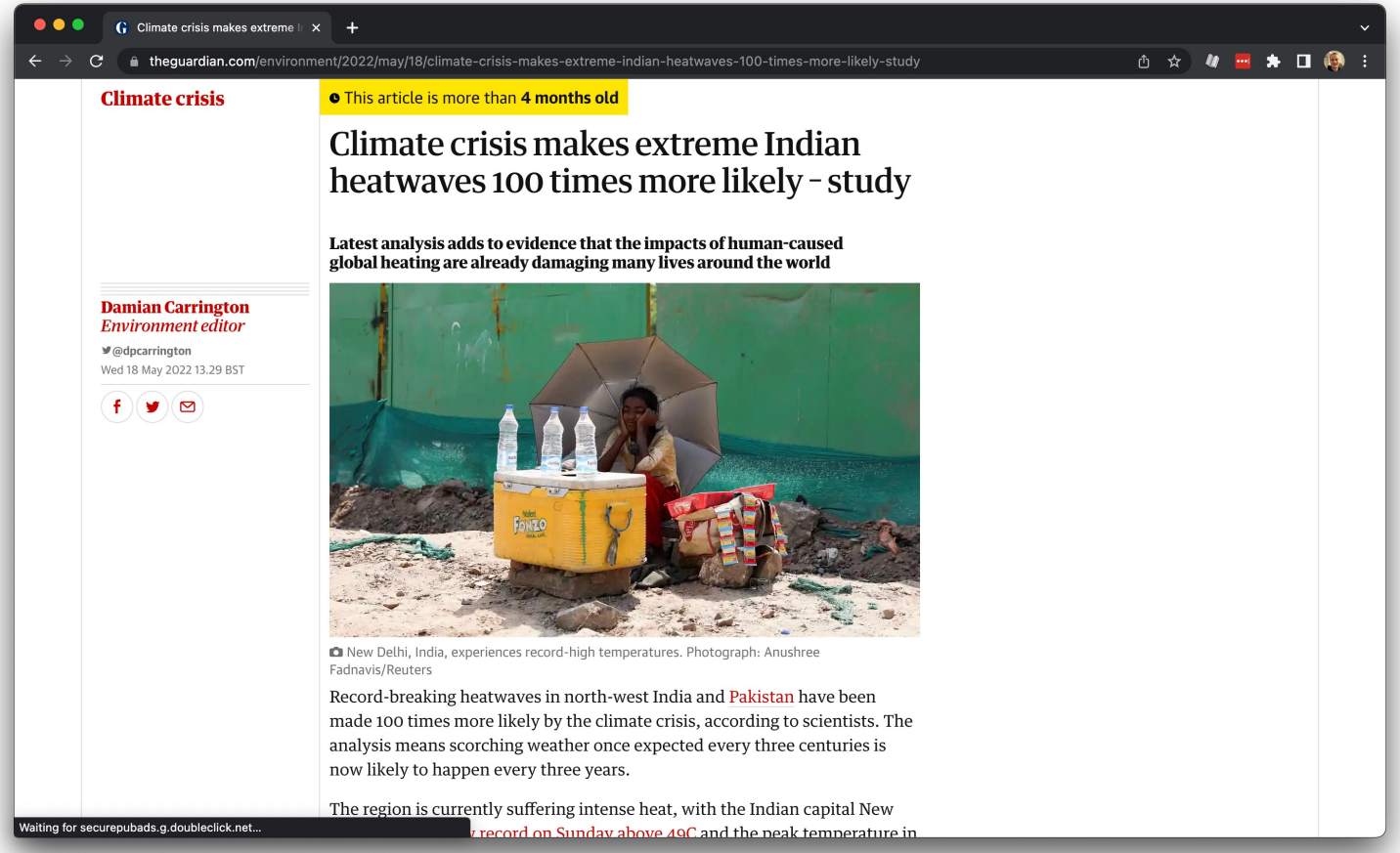
Climate crisis



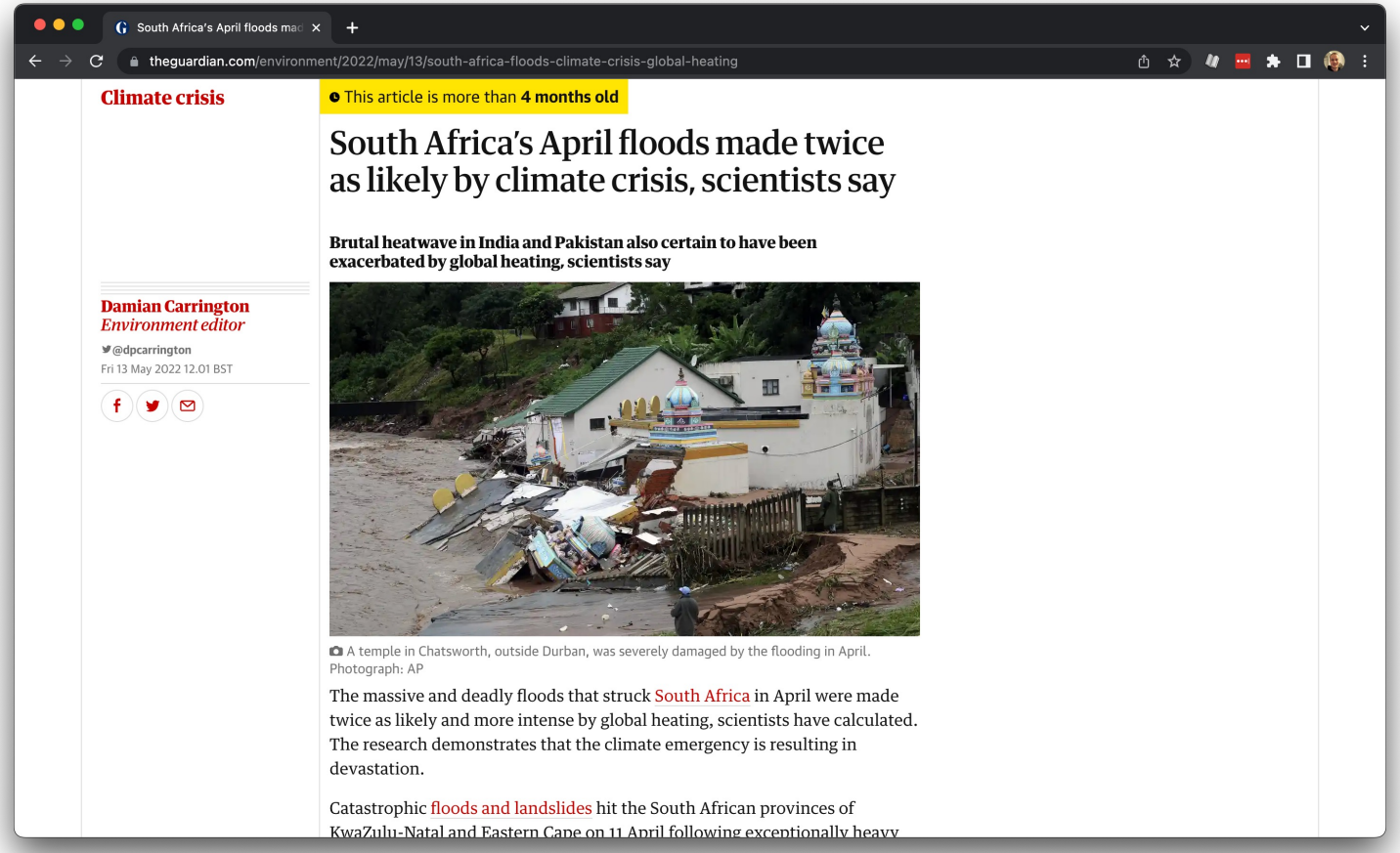
Climate crisis

Climate crisis

Climate crisis

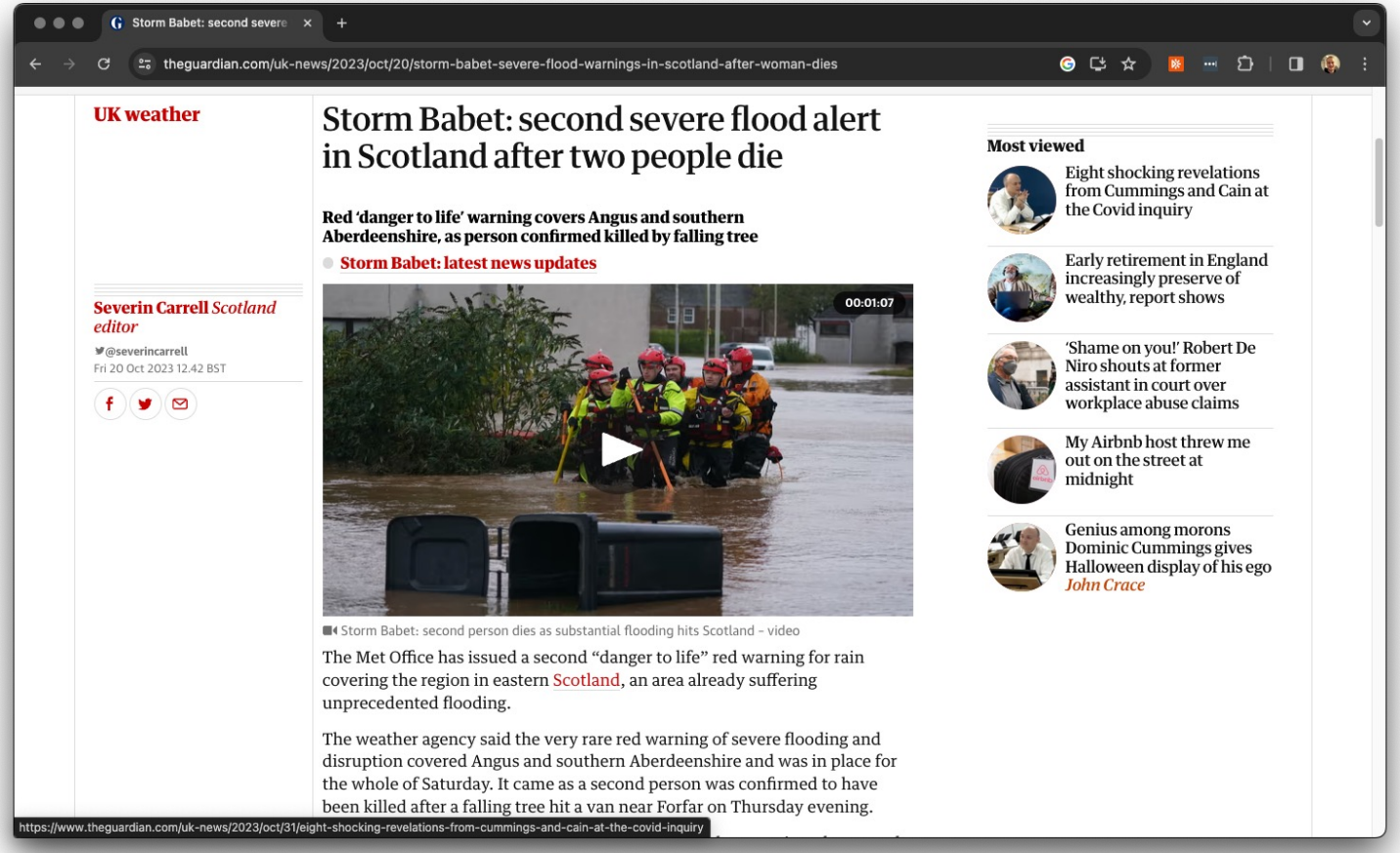


Climate crisis



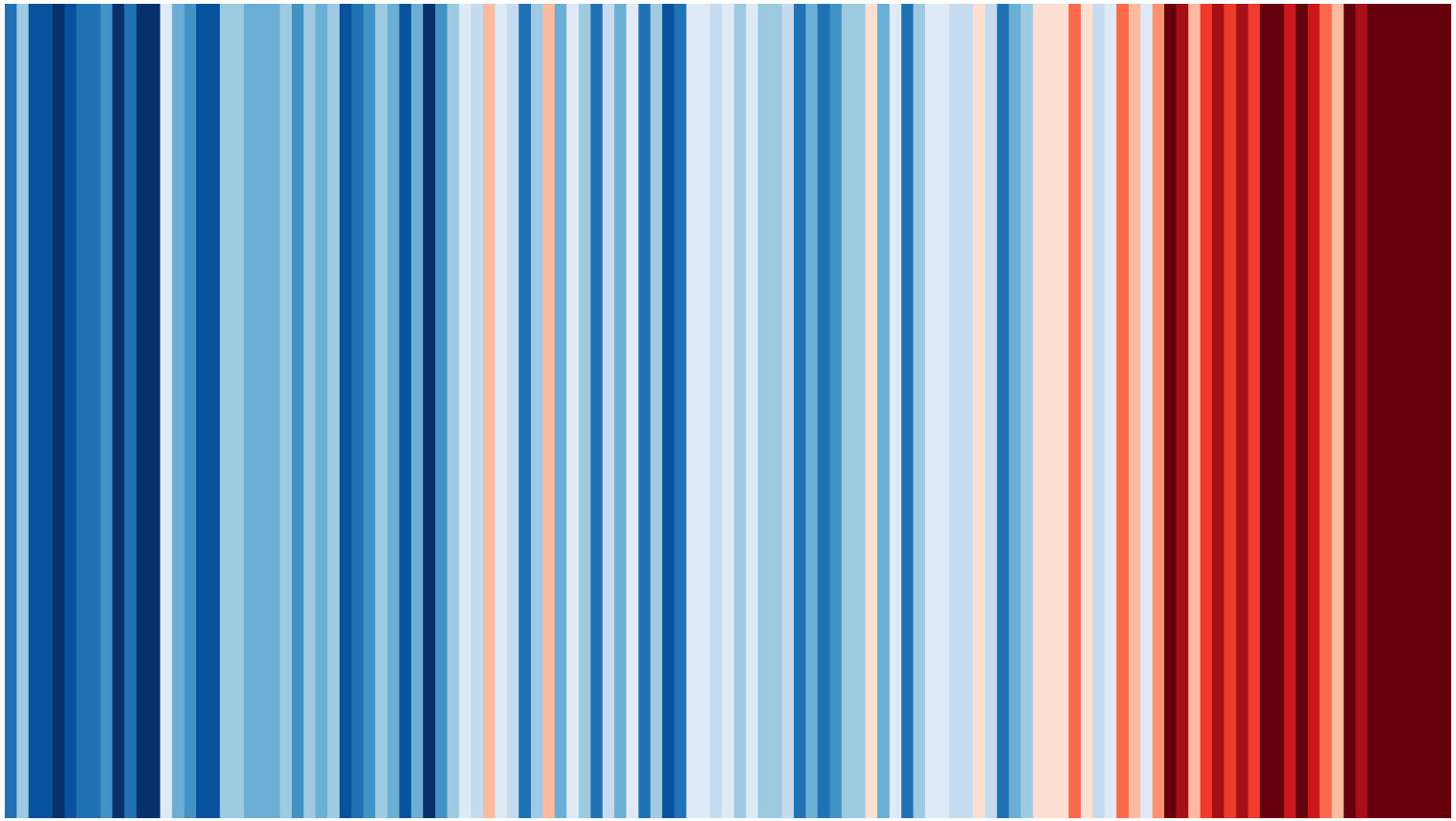
Climate crisis

Climate crisis



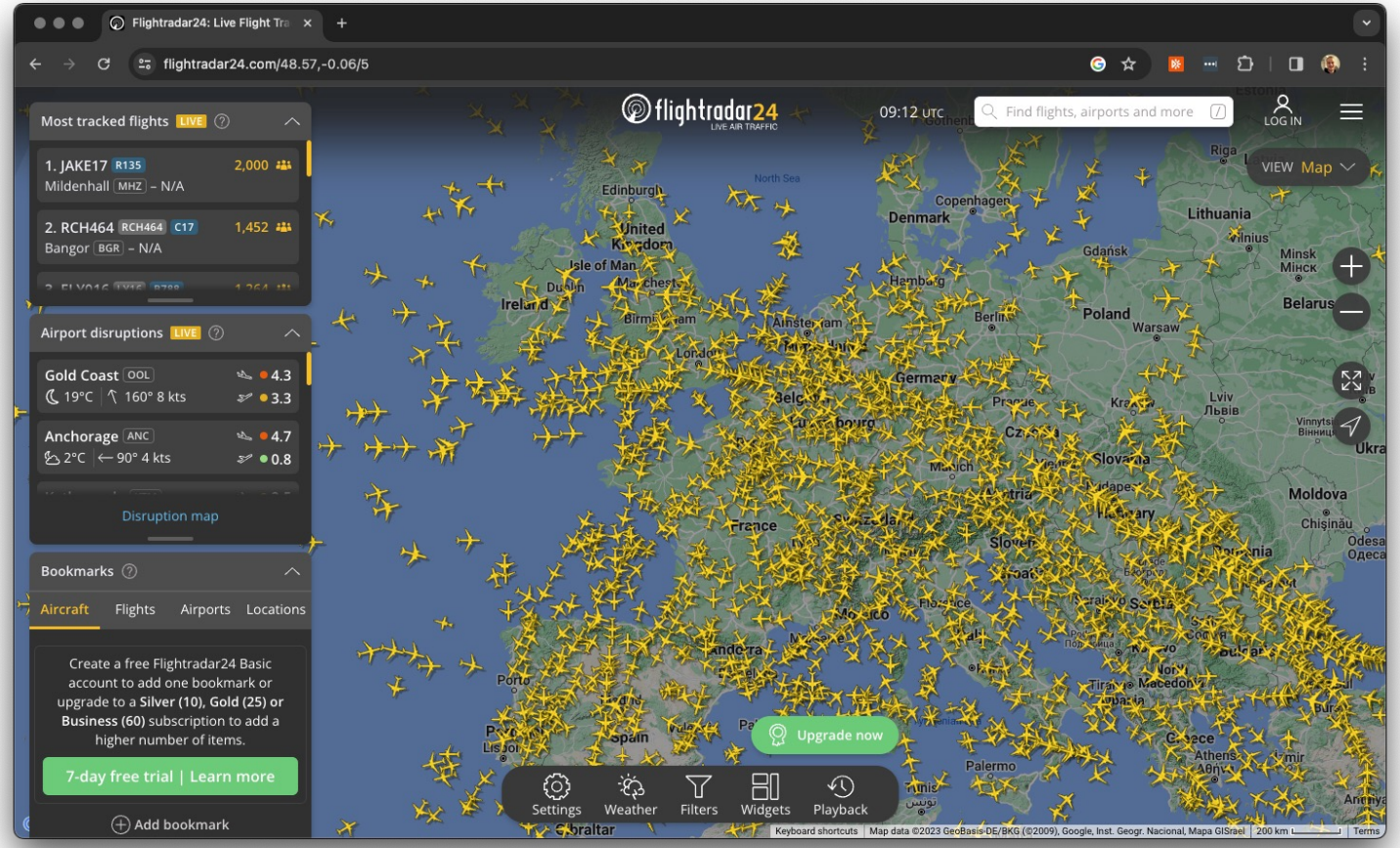
Climate crisis

Climate crisis

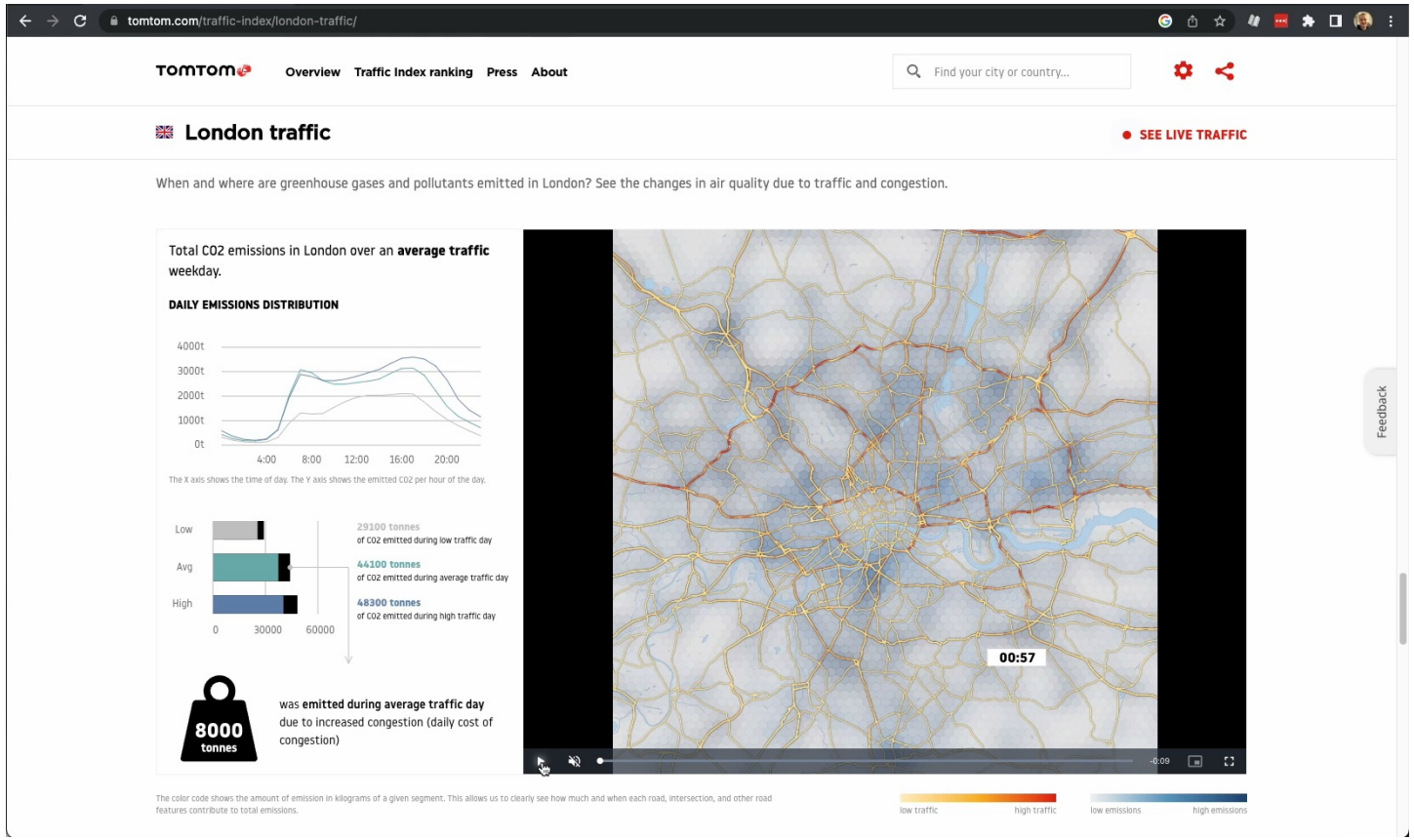


showyourstripes.info, created by Ed Hawkins

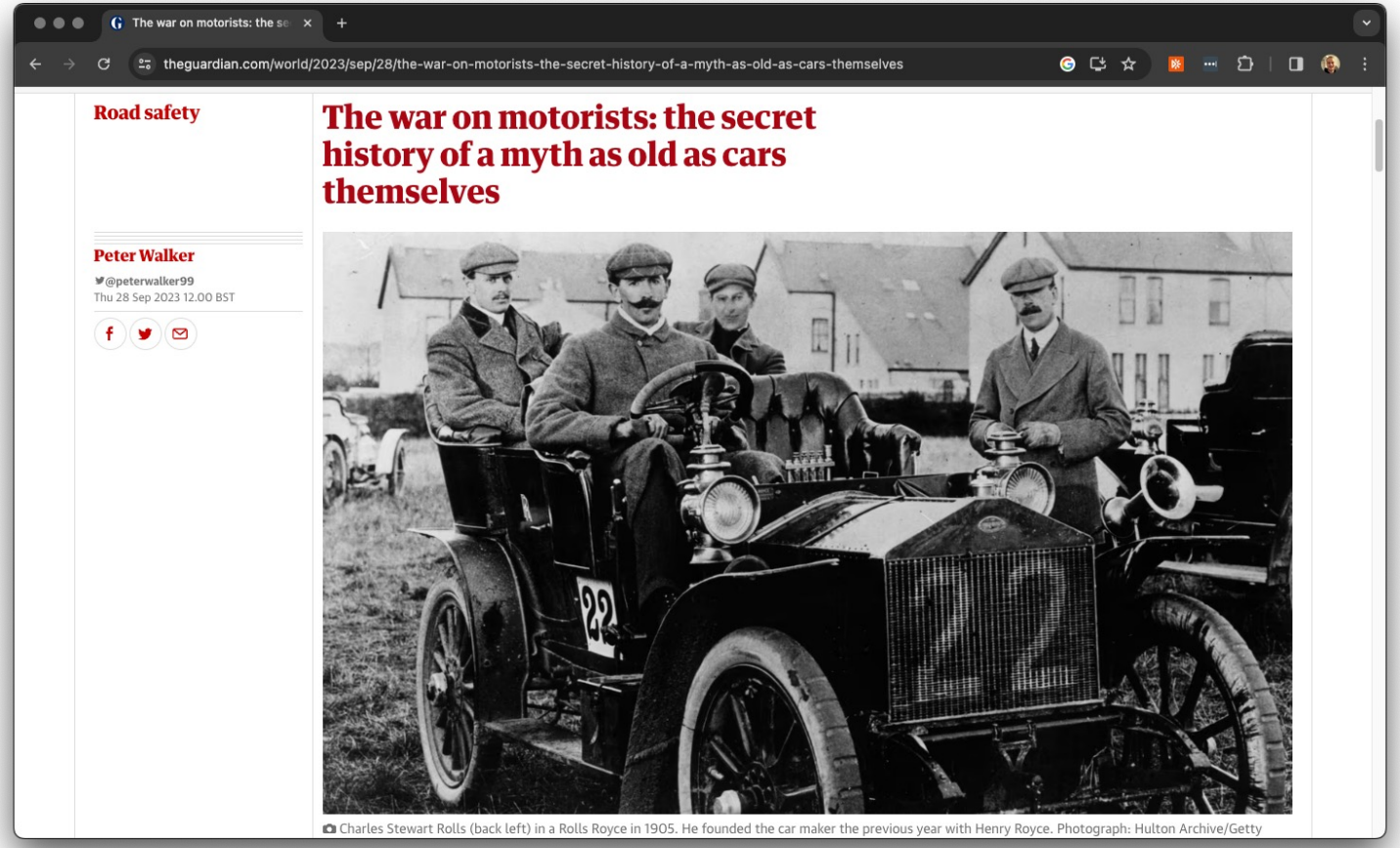
Climate crisis?



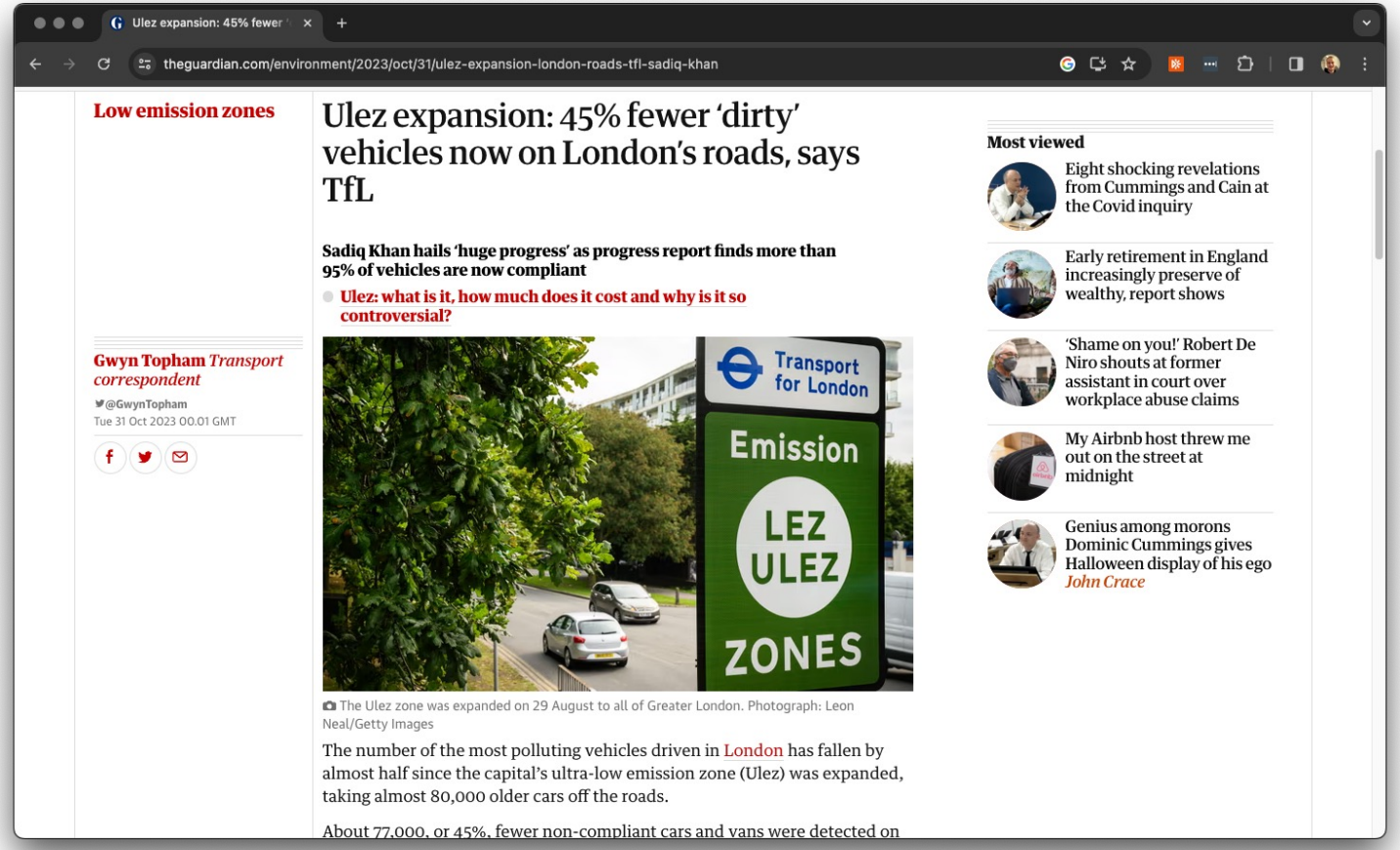
Climate crisis?



Climate crisis?

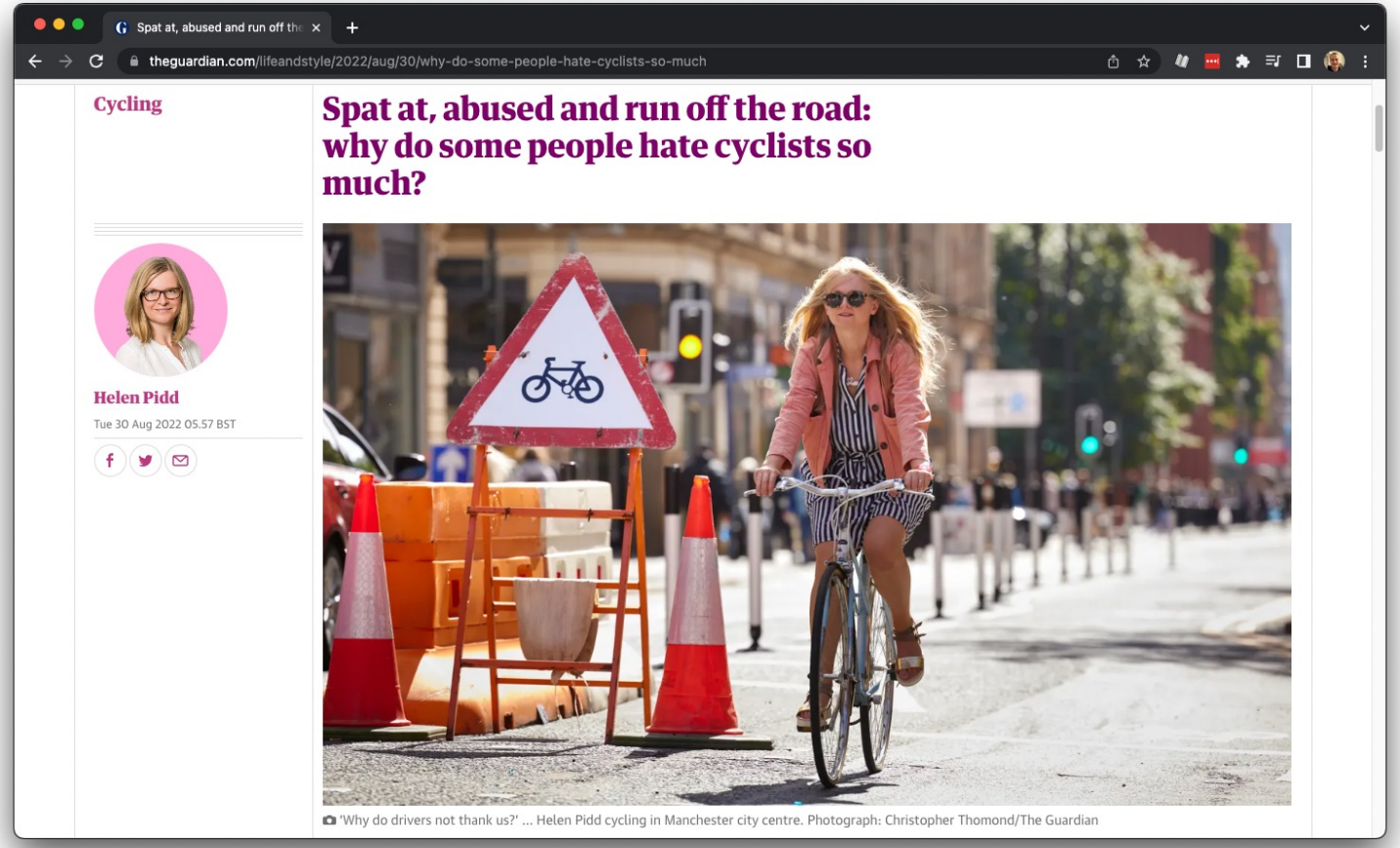


Local resistance



Local resistance

Local resistance



Data and technology

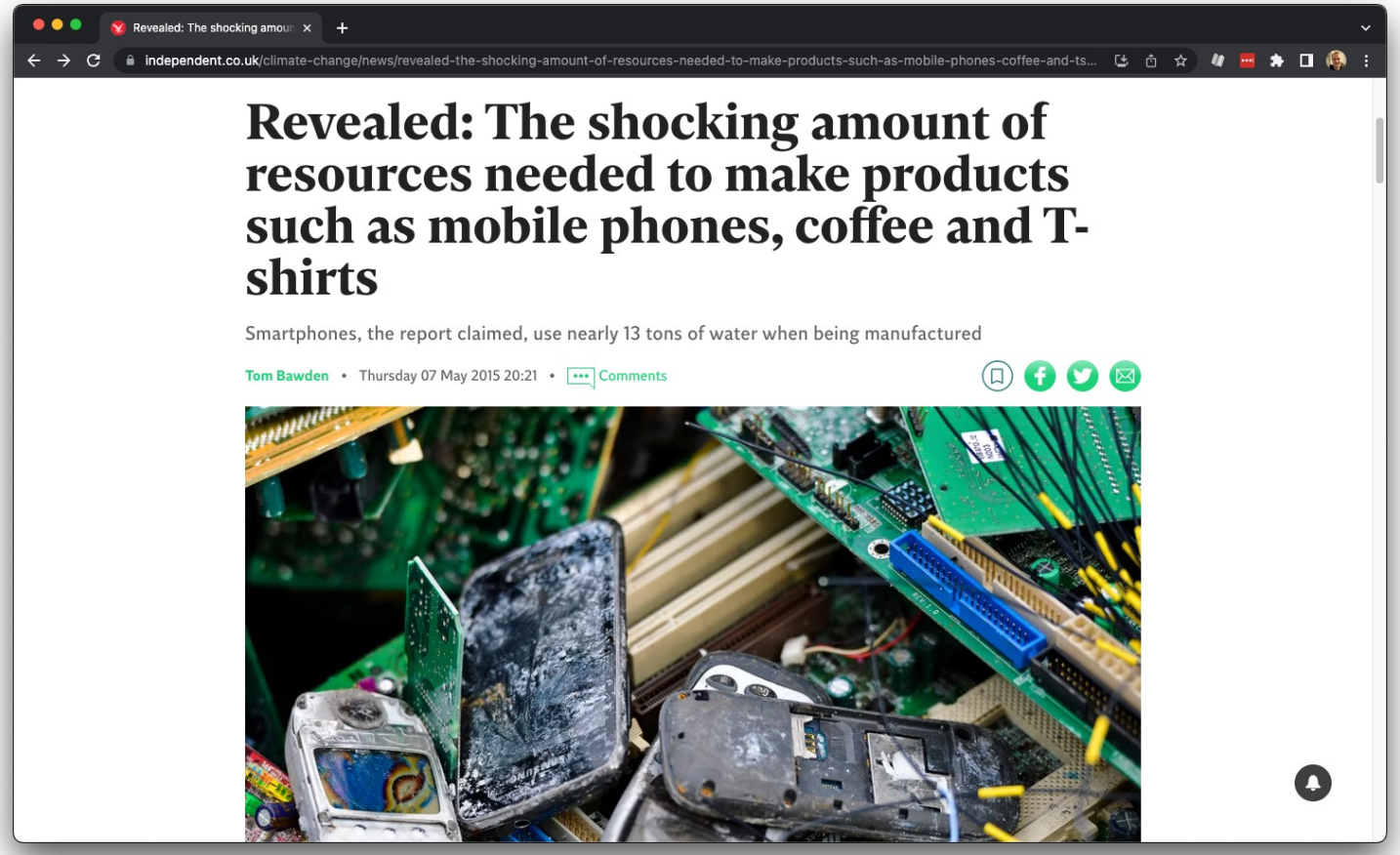
Although the transport sector is still one of the larger contributors to greenhouse gas emissions, the climate impact of information and communication technologies has been rapidly increasing.

Data and technology

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

Data and technology

Natural resources



Data and technology

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

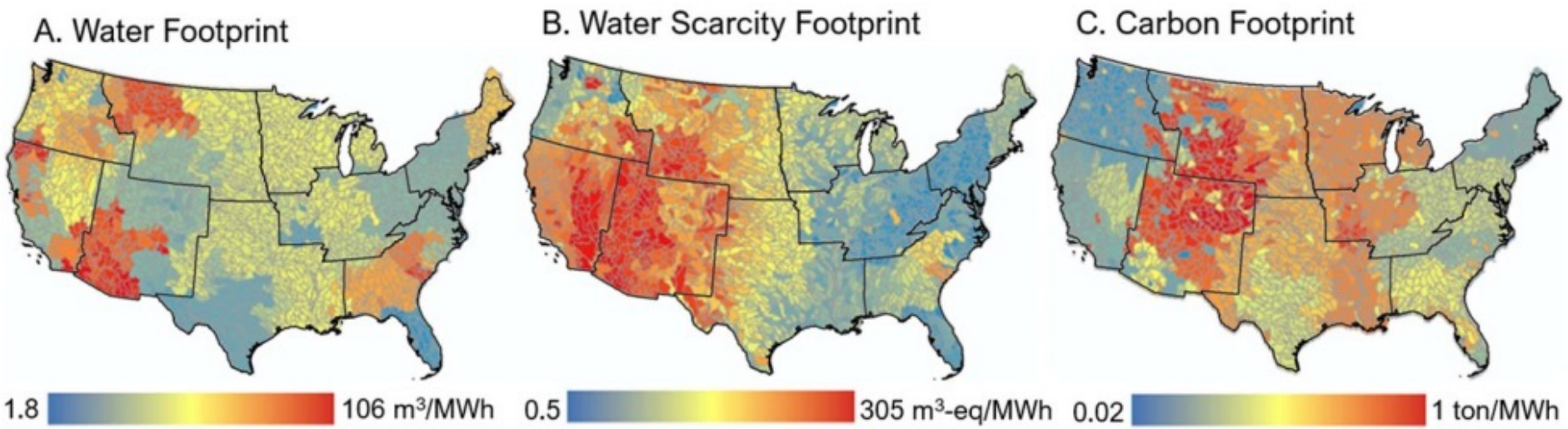


Figure 4. A data center's environmental footprint is highly contingent on where it is located. The (A) water intensity (m^3 MWh^{-1}), (B) water scarcity intensity (m^3 US-eq MWh^{-1}), and (C) GHG emissions intensity (tons CO_2-eq MWh^{-1}) of a hypothetical 1 MW data center placed in each of the 2110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

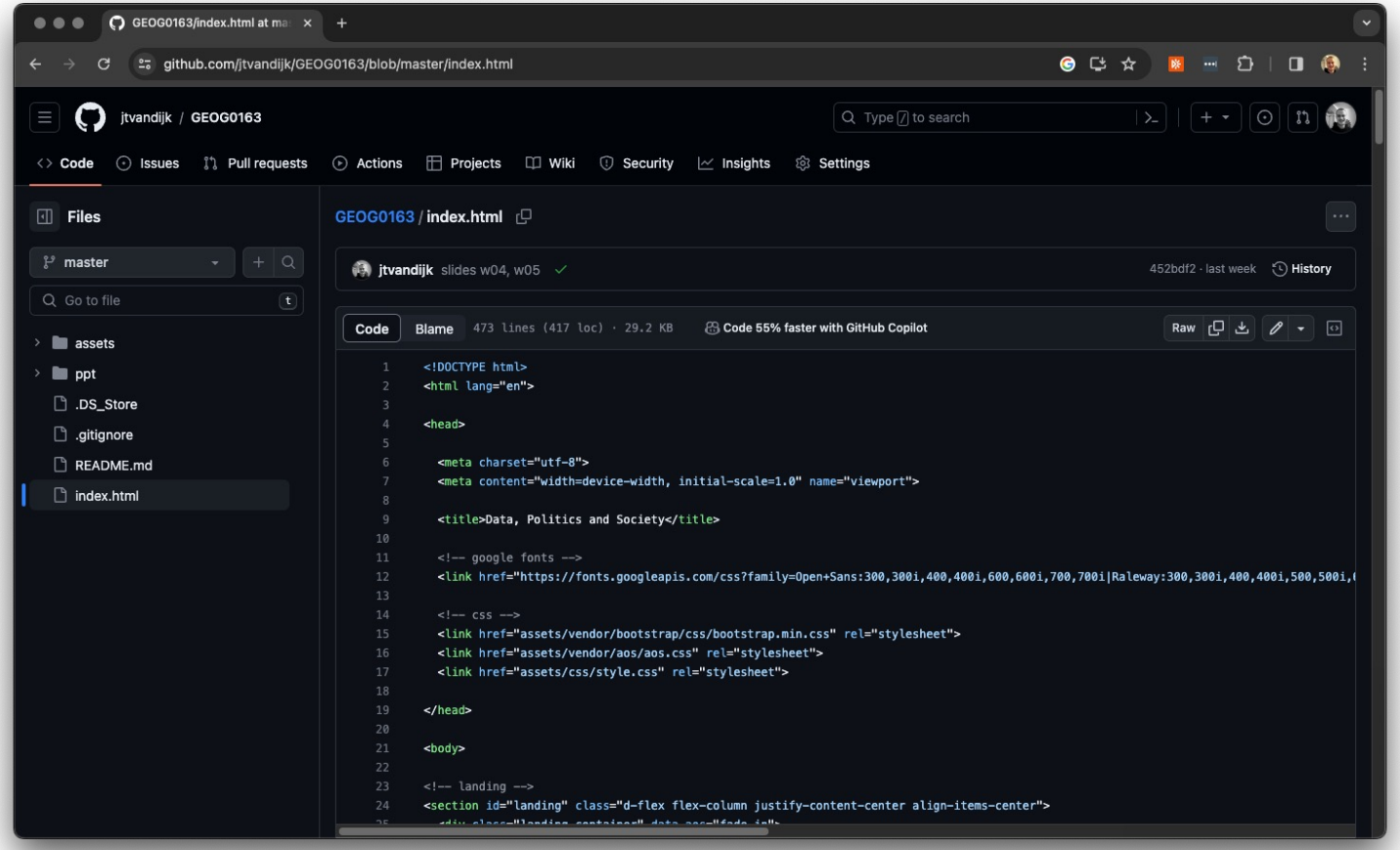
Consumer	Renew.	Gas	Coal	Nuc.
China	22%	3%	65%	4%
Germany	40%	7%	38%	13%
United States	17%	35%	27%	19%
Amazon-AWS	17%	24%	30%	26%
Google	56%	14%	15%	10%
Microsoft	32%	23%	31%	10%

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).

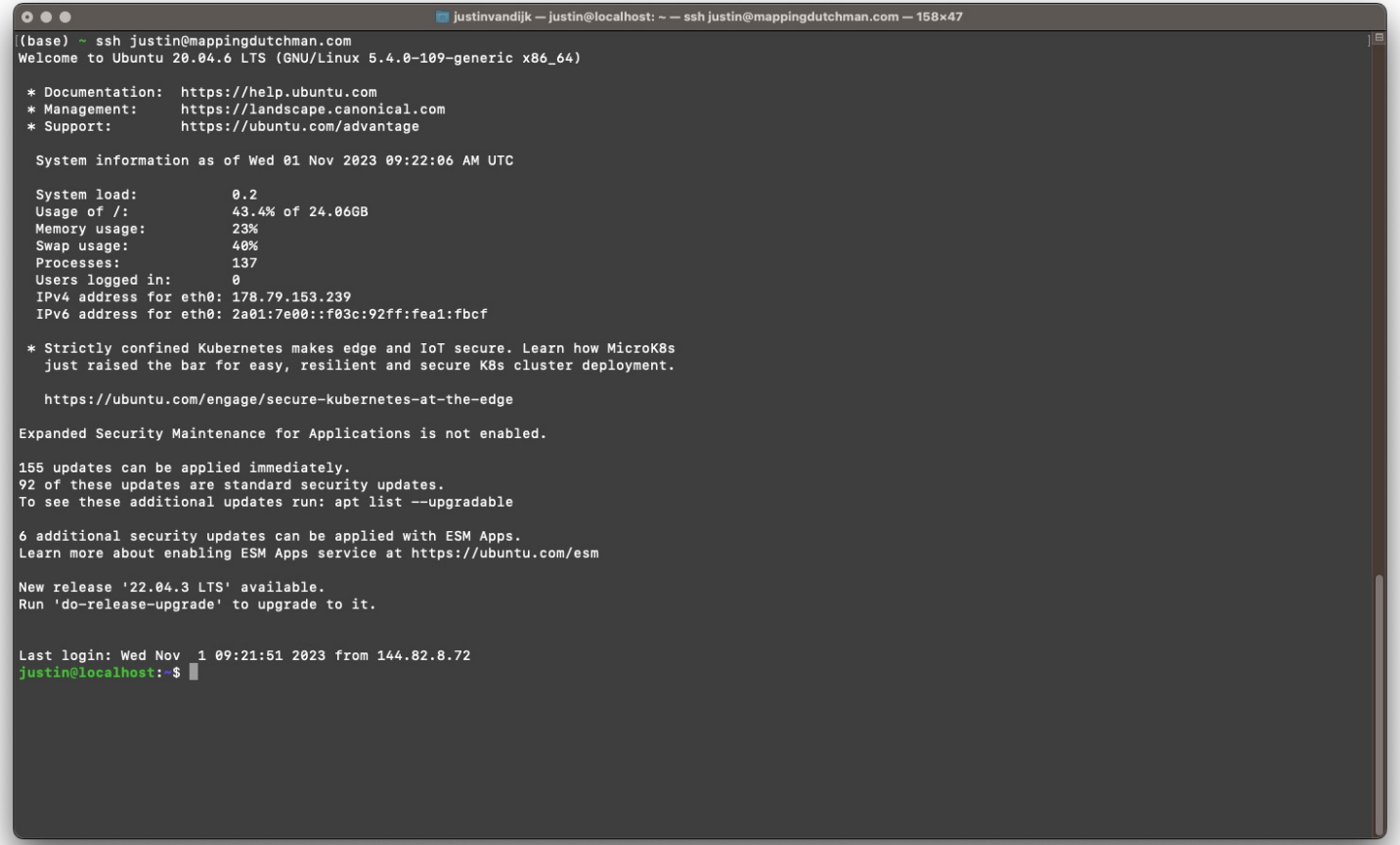
Data processing tools

- 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.

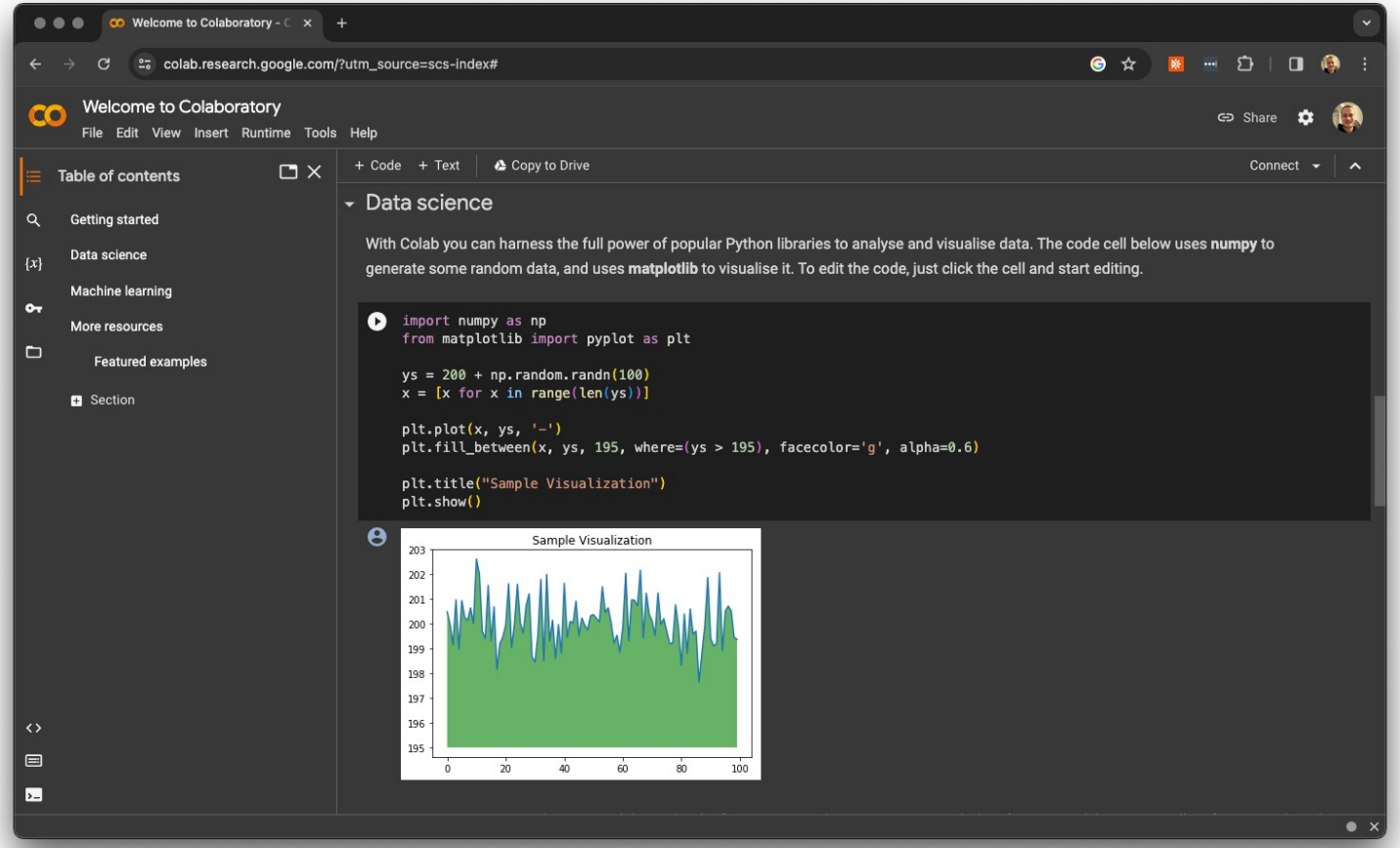
Data processing tools



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?

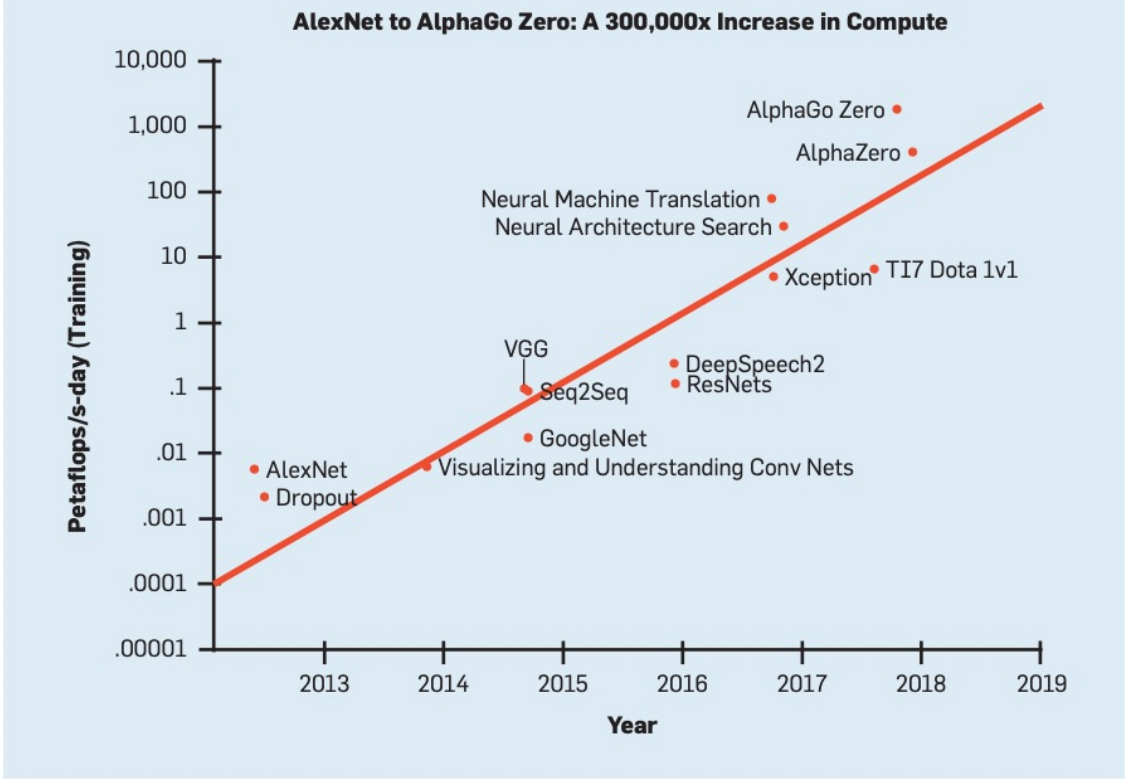
- Strübel *et al.* 2019
- Authors should report training time and sensitivity to hyperparameters.
- Academic researchers need equitable access to computation resources.
- Researchers should prioritize computationally efficient hardware and algorithms.

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?

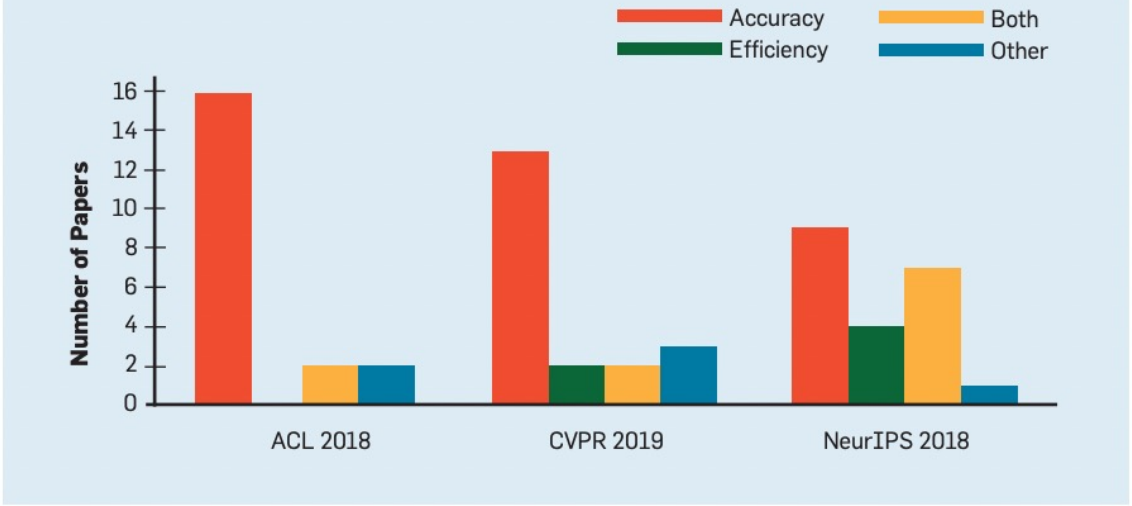
Figure 1. The amount of compute used to train deep learning models has increased 300,000x in six years. Figure taken from Amodei et al.[2]



Schwartz *et al.* 2020

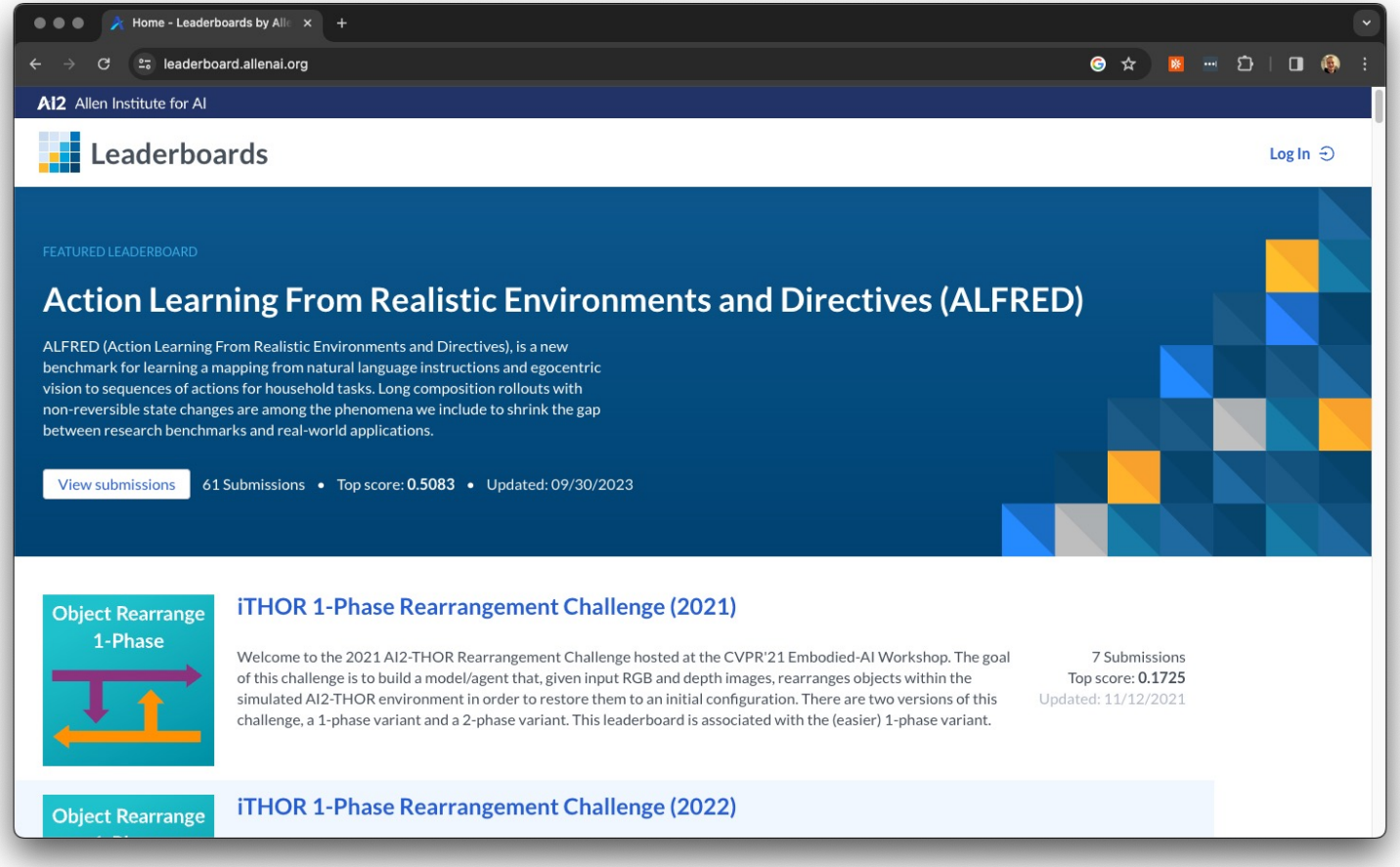
What can we do?

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?

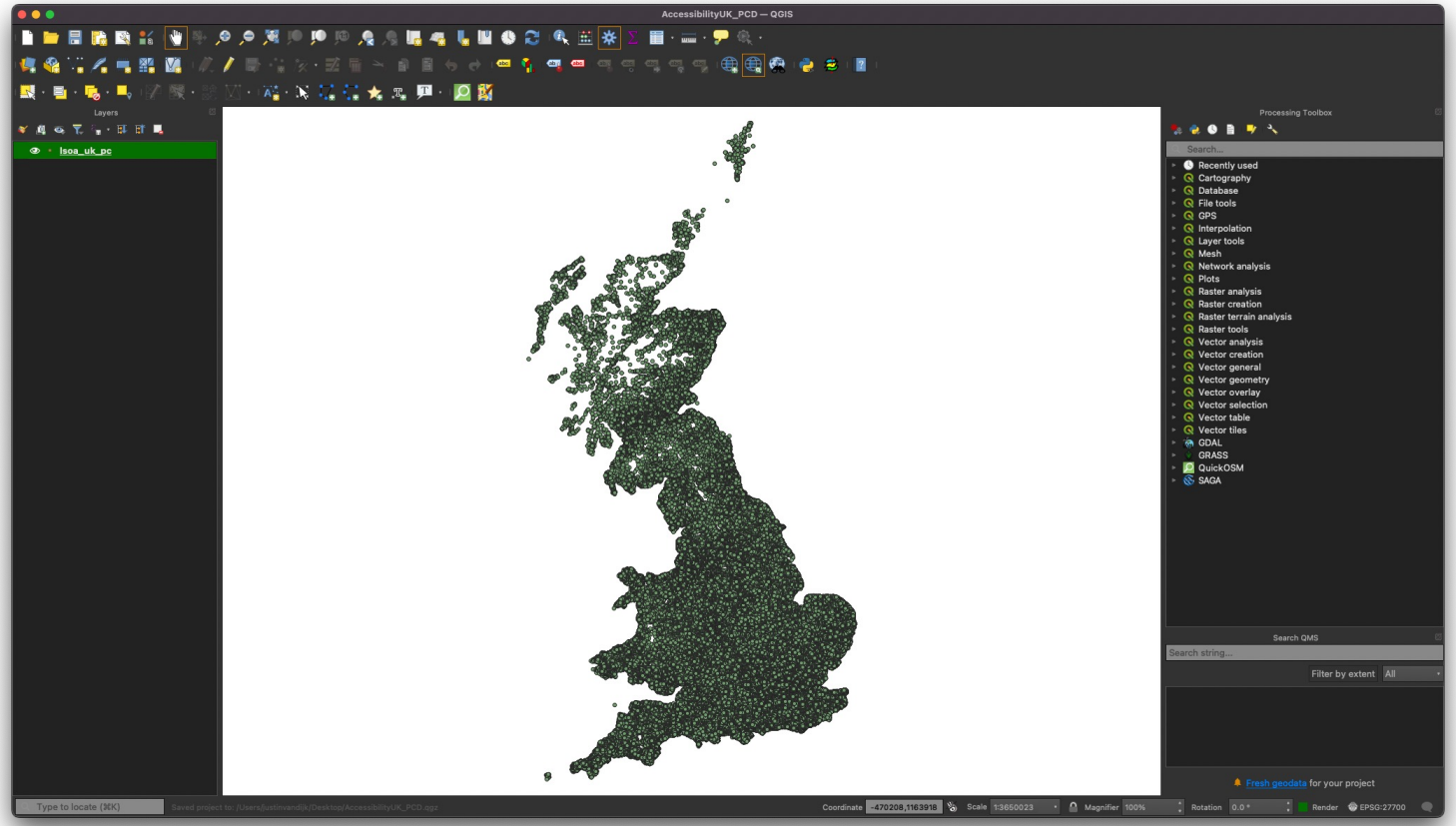
Task

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

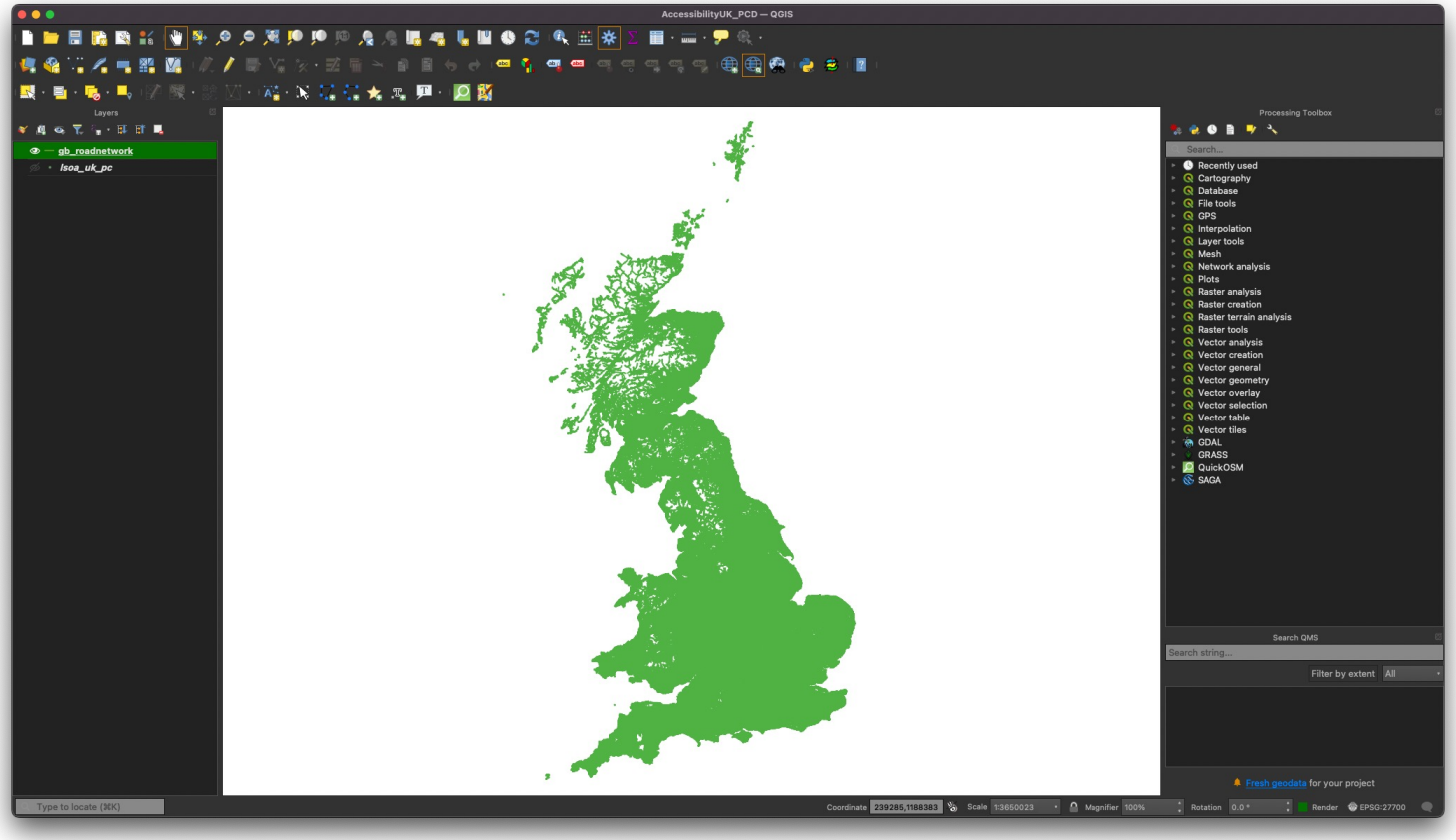
Data

- 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 questions: what are data and how can they be used?

- 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 paying attention to the *essay outline* part of your first coursework assignment through a peer-feedback assignment. For this Friday this means that everyone should bring their draft essay topic outline (either printed or on a digital device). This should include the main research question or aim and some bullet points on how you intend to address the question or aim.

This is an excellent opportunity to get some feedback – the better your draft outline, the more useful your feedback will be.

Questions

Justin van Dijk

j.t.vandijk@ucl.ac.uk

Enjoy reading week