

Data, Politics and Society

W3 – Data III: The Ugly



Where we at?

W1

W2

W3

Data: The Good, The Bad, The Ugly

W4

W5

Societal and environmental impacts of data and technology

The Ugly

Today

We will be talking about the ugly side of large human-generated datasets and the datafication of daily life, around three themes:

- Privacy
- Consent
- Neutrality

Privacy

Privacy

Privacy fears as schools use facial recognition to speed up lunch queue

Nine schools in North Ayrshire begin using technology to take payments, with others in UK expected to follow



▲ The company supplying the technology claimed it was more Covid-secure than other systems, as it was cashless and contactless, and sped up the lunch queue. Photograph: Susan Walsh/AP

Privacy campaigners have raised concerns about the use of facial recognition technology on pupils queueing for lunch in school canteens in the UK.

Nine schools in North Ayrshire began taking payments for school lunches this week by scanning the faces of their pupils, according to a [report in the Financial Times](#). More schools are expected to follow.

The company supplying the technology claimed it was more Covid-secure than other systems, as it was cashless and contactless, and sped up the lunch queue, cutting the time spent on each transaction to five seconds.

With [break times shortening](#), schools are under pressure to get large numbers of students through lunch more quickly.

Privacy

Google workers can listen to what people say to its AI home devices

Company admitted that contractors can access recordings made by Assistant, after some of its recordings were leaked



▲ In 2017, Google confirmed a bug in its Home Mini speaker allowed the smart device to record users even when it was not activated by the wake-up word. Photograph: Samuel Gibbs/The Guardian

Google acknowledged its contractors are able to listen to recordings of what people say to the company's artificial-intelligence system, [Google Assistant](#).

The company admitted on Thursday that humans can access recordings made by the Assistant, after some of its Dutch language recordings were leaked. [Google](#) is investigating the breach.

The recordings were obtained by [the Belgian public broadcaster VRT](#), which reviewed [more than 1,000 audio clips](#) and found 153 had been captured accidentally.

Google Assistant begins automatically recording audio when prompted by a user, usually by saying a wake-up word or phrase like, "OK, Google".

Privacy

Some more accidents (?) with voice assistants:

- Echo Dot voice assistant spitting out fragmentary commands, seemingly based on previous interactions with the device.
- Amazon customer in Germany was mistakenly sent about 1,700 audio files from someone else's Echo, providing enough information to name and locate the unfortunate user and his girlfriend.

Future?

Shhh ... Alexa might be listening

Amazon has filed a patent that could allow its Echo devices to one day listen in on conversations to help with user recommendations. A handy feature or more fodder for conspiracy theories?



▲ Amazon's Alexa Echo may become a more proactive assistant. Photograph: Alamy Stock Photo

Should you whisper around your [Amazon](#) Echo, lest it whisper back to you?

That's the future suggested by a patent recently filed by the company, which examined the possibility of eavesdropping on conversations held around its voice-activated devices in order to better suggest products or services to users.

The idea seems to be to turn [Alexa](#), the company's virtual assistant, from a dutiful aide under the user's command to one with a more proactive attitude. For instance, the [patent suggests](#): "If the user mentions how much the user would like to go to a restaurant while on the phone, a recommendation might be sent while the user is still engaged in the conversation that enables the user to make a reservation at the restaurant."

Locational privacy

- Individual spatial data is by nature very revealing and poses significant disclosure risks.
- Disclosure risks are everywhere, but location is very specific to the field of geography.
- Different to an economist's nationally representative dataset on transaction and purchasing behaviour.

Locational privacy

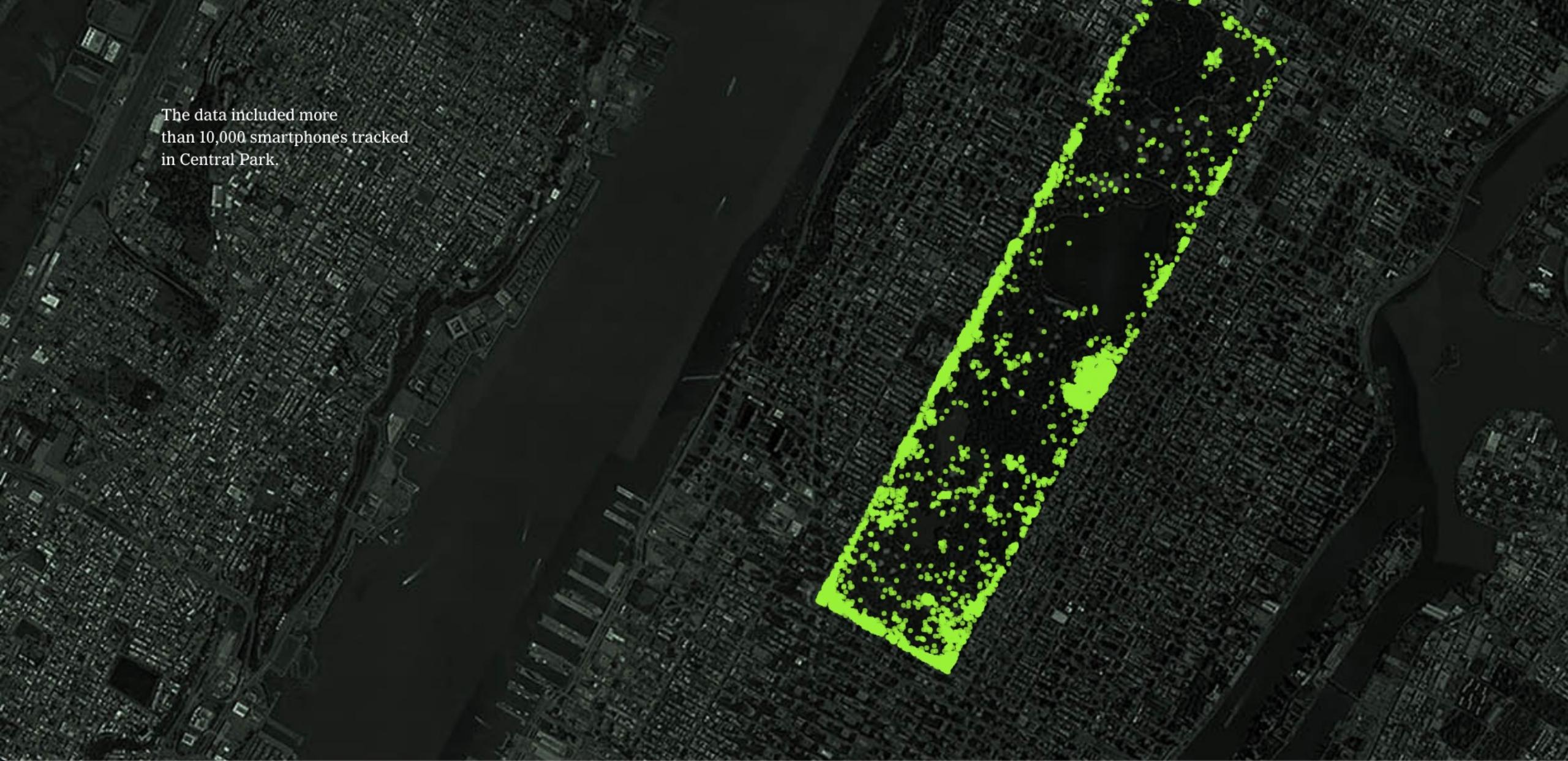
- The Privacy Project by the New York Times (2019).
- NY Times obtained a dataset comprising of more than 50 billion location pings of more than 12 million Americans as they moved through several major cities (e.g. Washington, New York, Los Angeles).
- Each piece of information in this file represents the precise location of a single smartphone over a period of several months in 2016 and 2017, data originated from a location data company (i.e. mobile applications).
- *The data was provided to Times Opinion by sources who asked to remain anonymous because they were not authorised to share it and could face severe penalties for doing so.*

The Privacy Project



New York Times. 2019. The Privacy Project: Twelve Million Phones, One Dataset, Zero Privacy. [Online]
<https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html>

The Privacy Project



The data included more than 10,000 smartphones tracked in Central Park.

The Privacy Project



Here is one smartphone, isolated
from the crowd.

The Privacy Project



Here are all pings from
that smartphone over the period
covered by the data.

The Privacy Project



Connecting those pings reveals a diary of the person's life.

Locational privacy



New York Times. 2019. The Privacy Project: Twelve Million Phones, One Dataset, Zero Privacy. [Online] <https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html>

Consent

Unwilling and unknowing participants

- Not just individual spatial data from location-based services can be very informative: social media data is another source used to generate population insights.
- Analytics as black-boxed trade secrets.
- Classifications could reproduce those “social and geographical categories in people’s material consumptive practices” and in this way “produces the conditions for its own reproduction”.
- Dalton and Thatcher 2015: production of consumer geographies.
- New ‘Big’ Data: from quantified space to quantified individuals.

Unwilling and unknowing participants

- Cambridge Analytica and Facebook scandal 2018.
- The Times reported that in 2014 contractors and employees of Cambridge Analytica, eager to sell psychological profiles of American voters to political campaigns, acquired the private Facebook data of tens of millions of users — the largest known 'data breach' in Facebook history.
- CA used personal information taken without authorisation to build a system that could profile individual US voters, in order to target them with personalised political advertisements.
- Stephen Bannon, former Trump aide, was a board member of the company.

Informed consent

In a research setting:

- purpose of the research
- type of research intervention, e.g. questionnaire, interview, etc.
- participation is voluntary
- benefits and risks of participating
- procedures for withdrawal from the study
- usage of the data during research

Who has access

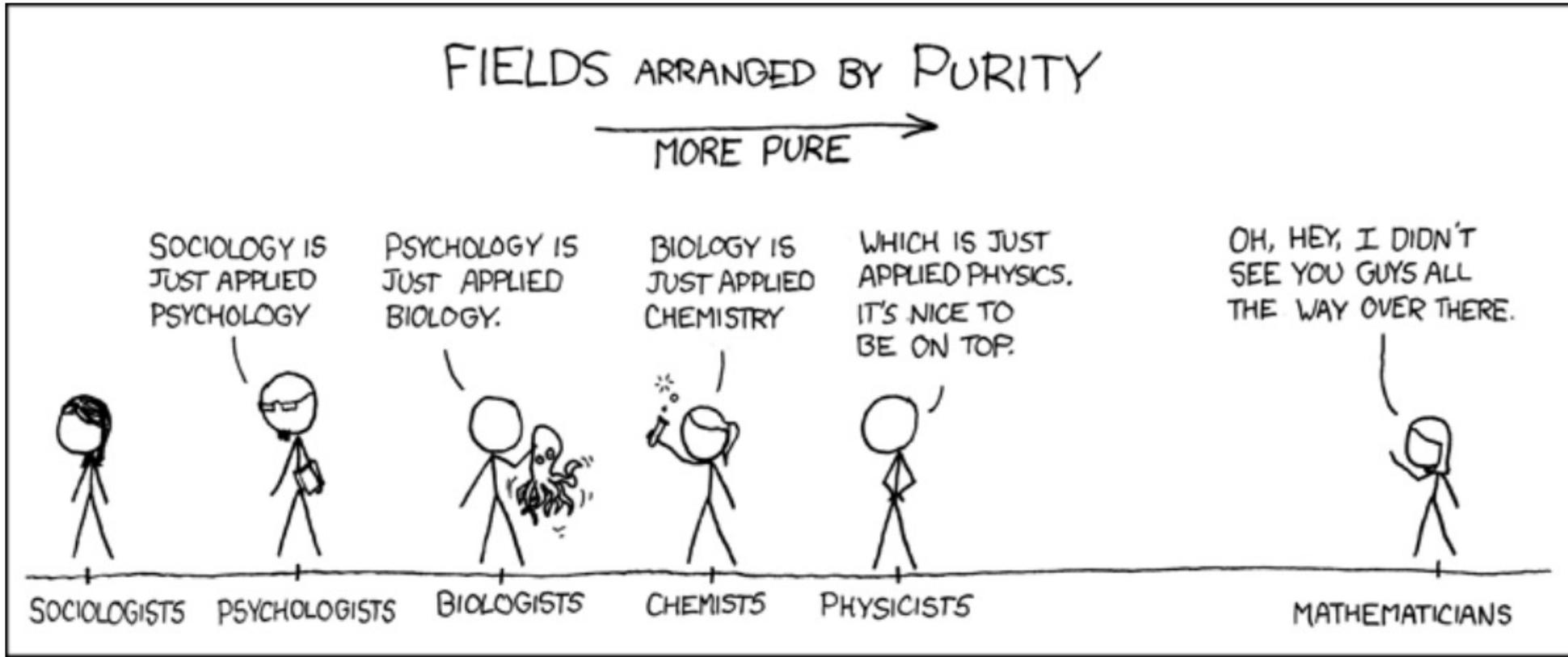
- Privacy concerns arise first and foremost because of the questions: who can access my data and for what purposes.
- Is informed consent enough (more in W06, W07)?

Neutrality

Knowledge is not neutral

There is an uneven spatial politics of geographic knowledge production where theory from the Anglo-American sphere is privileged and universalised.

Knowledge is not neutral



Knowledge is not neutral

More general:

positionality the idea that the identities of the researcher influence the research process and their interactions with research participants (also applies to data!)

reflexivity the process of considering the researcher's positionality and the effects of this positionality on one's research

Knowledge is not neutral

- There is an implicit danger in an attempt to attain objective knowledge, that what appears to one group of researchers at a particular point in time is often treated as a fact (see also the article by Schurr *et al.* 2020).
- When defining social identities, social scientists have often translated prejudices into "objective" categories.

Institutionalisation of prejudice

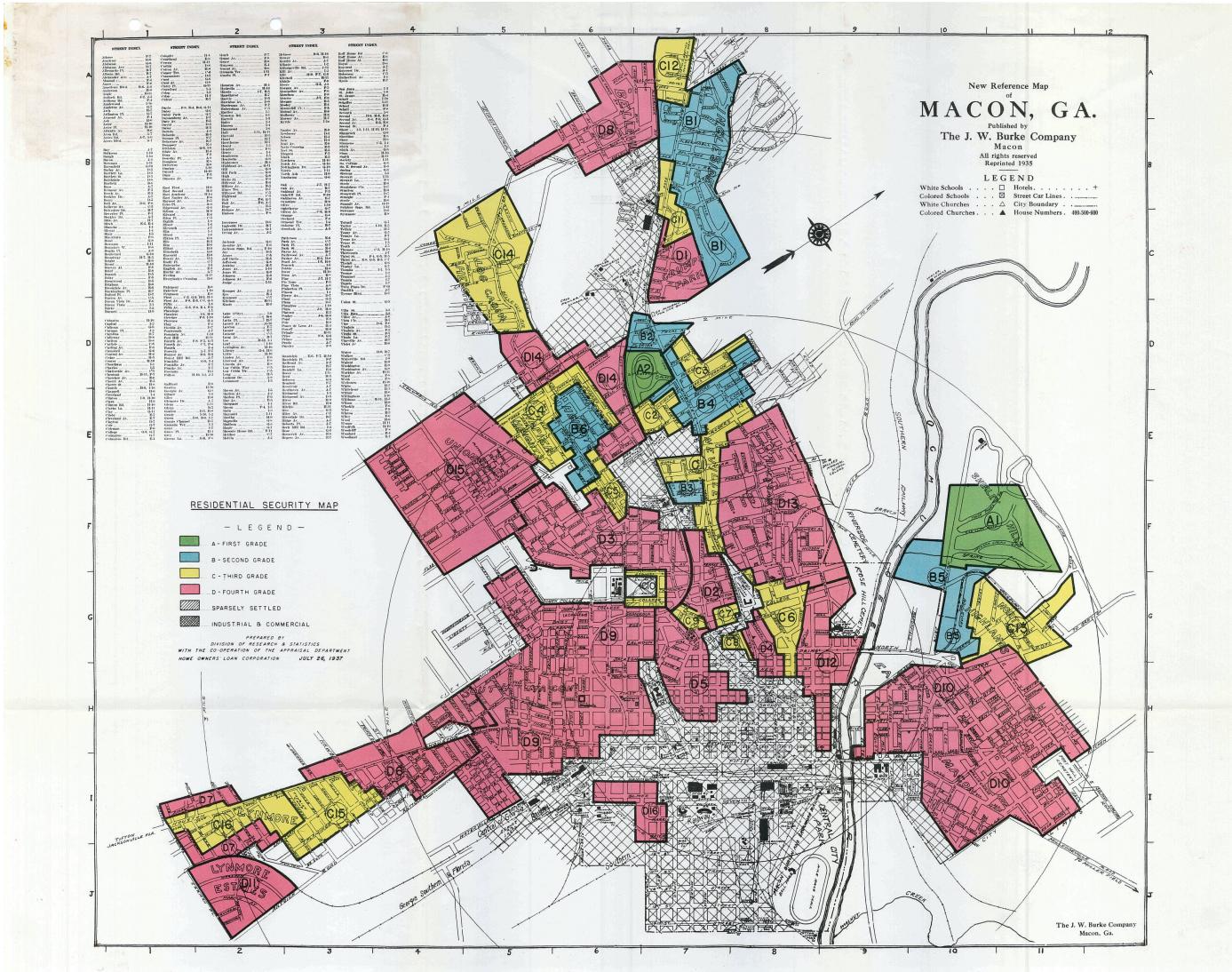
From prejudice to "objective prejudice" in extrema:

- Eugenics: a 'science' of improving the human species through gene selection, which in effect is the advocacy of selective breeding of the population to further racist or discriminatory aims.
- Scientific racism: much of the knowledge was developed through western perspectives, with some using it as a reason to segregate populations (e.g. 'separated development' in Apartheid South Africa)

Institutionalisation of prejudice

- Redlining. US 1930s: government surveyors graded neighborhoods in 239 cities, color-coding them green for “best,” blue for “still desirable,” yellow for “definitely declining” and red for “hazardous.”
- The “redlined” areas were the ones local lenders discounted as credit risks, in large part because of the residents’ racial and ethnic demographics.
- Loans in these neighborhoods were unavailable or very expensive, making it more difficult for low-income minorities to buy homes and setting the stage for the country’s persistent racial wealth gap.

Institutionalisation of prejudice



Washington Post. 2018. Redlining was banned 50 years ago. It's still hurting minorities today. [Online] <https://www.washingtonpost.com/news/wonk/wp/2018/03/28/redlining-was-banned-50-years-ago-its-still-hurting-minorities-today/>

Institutionalisation of prejudice

G How dividing US cities along racial lines led to an air pollution crisis 100 years on

the guardian.com/us-news/2022/mar/09/redlining-air-pollution-us-cities?CMP=Share_IOSApp_Other

Supported by

OPEN SOCIETY FOUNDATIONS

About this content

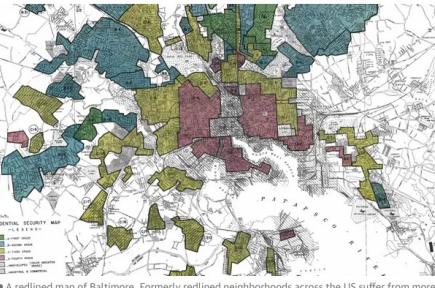
Erin McCormick

Wed 9 Mar 2022 13.00 GMT

[f](#) [t](#) [e](#)

How dividing US cities along racial lines led to an air pollution crisis 100 years on

Study of 200 cities shows dangerous environmental inequality fueled by 20th-century practice of redlining



A redlined map of Baltimore. Formerly redlined neighborhoods across the US suffer from more air pollution than other areas, researchers found. Photograph: US Federal Housing Administration via Richmond University's Mapping Inequality

A new study has found that neighborhoods in which the federal government discouraged investment nearly 100 years ago - via a racist practice known as redlining - face higher levels of air pollution today.

Looking at more than 200 cities across the nation, researchers from the University of California, Berkeley, found that people who live in neighborhoods that were once categorized as "hazardous", based on racist factors such as how many Black or "foreign-born" people lived there, now breathe 56% more of the freeway pollutant nitrogen dioxide than those in top-rated areas.

Those formerly redlined neighborhoods also suffer from higher levels of the sooty particle known as PM 2.5, the study found. And both pollutants are associated with health effects, including higher rates of asthma, cardiovascular disease and even Covid-19.

Historically redlined neighborhoods experienced the highest levels of pollution

How much greater nitrogen dioxide levels were in 2010 in redlined neighborhoods (Grade D) compared to the entire city

Less NO₂ | More NO₂
← pollution →

Grade A

Black people and immigrants were especially targeted in neighborhoods

Most viewed

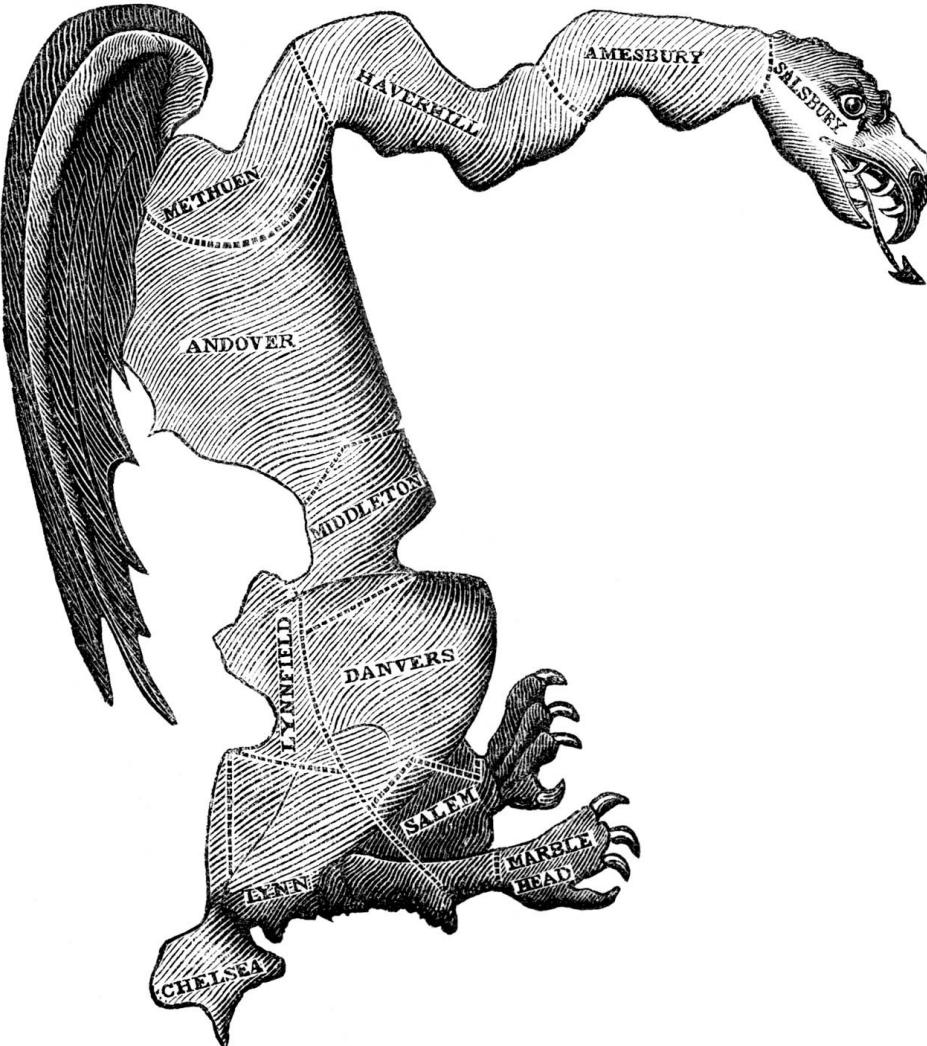
Live Liz Truss defends mini-budget, saying she has to do 'what I believe is right' - UK politics live

Live Russia-Ukraine war live: Vladimir Putin to sign

Institutionalisation of prejudice

- Gerrymandering. The practice of drawing the boundaries of electoral districts in a way that gives one political party **an unfair advantage** over its rivals (political or partisan gerrymandering) or that dilutes the voting power of members of ethnic or linguistic minority groups (racial gerrymandering).
- The term is derived from the name of Gov. Elbridge Gerry of Massachusetts, whose administration enacted a law in 1812 defining new state senatorial districts. The law consolidated the Federalist Party vote in a few districts and thus gave disproportionate representation to Democratic-Republicans. The outline of one of these districts was thought to resemble a salamander: "The Gerry-mander".

Institutionalisation of prejudice



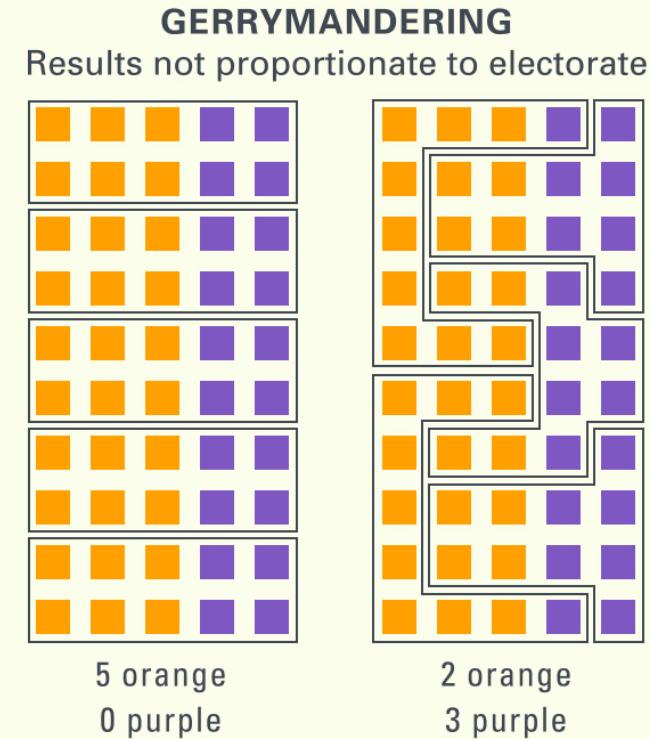
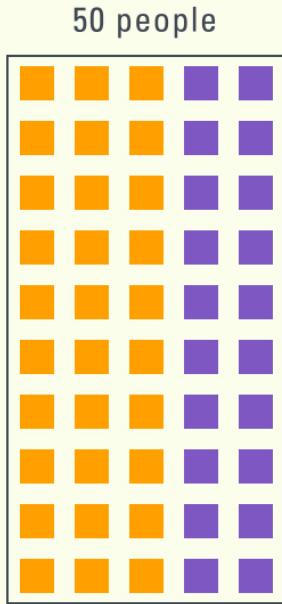
Encyclopaedia Britannica, Inc.. Gerrymandering. [Online]
<https://www.britannica.com/topic/gerrymandering>

Institutionalisation of prejudice

GERRYMANDERING

How differently drawn district maps produce different electoral results

FOUR WAYS TO DIVIDE 50 PEOPLE INTO 5 DISTRICTS:



© Encyclopædia Britannica, Inc.

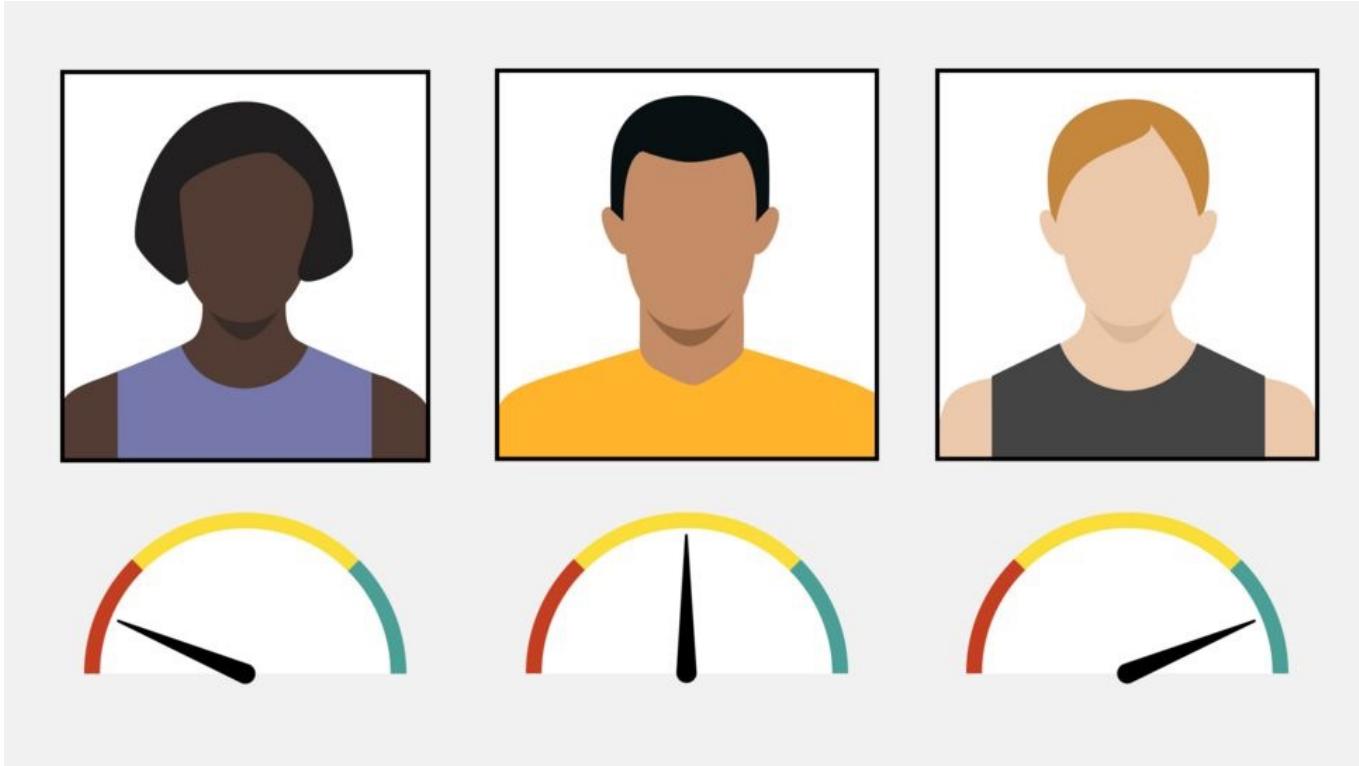
Making the same mistakes of the past?

- Texas: enactment of most restrictive abortion law in the United States.
- Democratic-leaning cities and their suburbs are growing quickly, while Republican-leaning rural areas are not.
- The Guardian 5 September 2021:

"A decade ago, Republicans had complete control over the process of drawing the boundaries for state legislative and congressional districts. It allowed them to distort the lines to help Republicans win elections and guarantee their election in the state legislature over the past 10 years. This year the lines will be redrawn again and Republicans once again will have complete control of the process."

Making the same mistakes of the past?

BBC 2020: Women with darker skin are more than twice as likely to be told their photos fail UK passport rules when they submit them online than lighter-skinned men. How?



Making the same mistakes of the past?

- Netherlands: "Leefbaarheidsbarometer" ("neighbourhood liveability")
- Composite measure, comparable to the Index of Multiple Deprivation in the UK
- 100s of variables such as distance to nearest highway, access to greenery, access to education, access to health.
- Percentage of immigrants ("non-Western"): specifically percentage of people with a Turkish, Surinam, Moroccan, or Eastern Europeans background.
- Producers of the indicator would claim they look at predictive power ('correlation' not 'causation') – what about the consumer of these indicators and maps?
- How to measure 'neighbourhood liveability' to begin with – very subjective

Making the same mistakes of the past?

Algorithms applied on large, unrepresentative datasets may exacerbate bias and prejudice, especially when they are used for decision-making ("e.g. algorithmic governance").

Algorithm

What is an algorithm?

Algorithm

Veggie chilli | Vegetables recip... [+ Not Secure | jamieoliver.com/recipes/vegetables-recipes/veggie-chilli-with-crunchy-tortilla-avocado-salad/](#)

Jamie Oliver

RECIPES DISCOVER FAMILY BUDGET-FRIENDLY ONE SIGN UP / LOG IN

CHILLI & RICE

1 red onion
1 dried smoked chipotle or ancho chilli
½ a fresh red chilli
1 teaspoon sweet smoked paprika
½ teaspoon cumin seeds
1-2 cloves of garlic
1 big bunch of fresh coriander
olive oil
2 mixed-colour peppers
1 x 400 g tin of chickpeas
1 x 400 g tin of black beans
700 g passata
1x 250 g pack of cooked mixed long grain & wild rice

SALAD

4 small corn tortilla wraps
2 ripe avocados
3 heaped tablespoons fat-free natural yoghurt, plus extra to serve
2 limes
1 romaine lettuce
½ a cucumber
1 fresh red chilli
1 handful of ripe cherry tomatoes

Ingredients out • Oven at 200°C/400°F/gas 6 • Food processor (bowl blade) • Lidded casserole pan, high heat • Stick blender

START COOKING

Peel and halve the red onion. Put the chillies, onion, paprika and cumin seeds into the processor, squash in the unpeeled garlic through a garlic crusher, then add the coriander stalks (reserving the leaves) and 2 tablespoons of oil, and whiz until fine.

Tip into the pan. Deseed and roughly chop the peppers, drain the chickpeas and black beans, then add to the pan with a pinch of sea salt and black pepper and the passata, stir well and put the lid on.

Fold the tortillas in half, slice into 0.5cm strips, sprinkle on to a baking tray and pop in the oven until golden and crisp.

Put most of the coriander leaves, a pinch of salt and pepper, half a peeled avocado, the yoghurt and the lime juice into a jug and whiz with a stick blender until silky.

Check and adjust the seasoning of the chilli, then leave the lid off.

Remove the tortillas from the oven into a bowl, cut the lettuce into chunky wedges and add to the bowl. Scoop and dot over curls of avocado.

Peel the cucumber into ribbons and finely slice half a chilli, then scatter both over the top.

Make a well in the middle of the chilli and tip in the rice, then pop the lid on for the last few minutes to warm the rice through.

Pour the dressing over the salad, pick over the remaining coriander leaves, finely slice the remaining chilli and sprinkle over the top along with the halved cherry tomatoes, then toss everything together. Serve with dollops of yoghurt.

Buy ingredients online £4.67 per serving
[Change supermarket](#)

Shop at Ocado

Create a shopping list [View list](#)

ONE PAN, ONE BOOK, ONE QUICK FIX BUY NOW

ASUS

Windows 365

Welcome to your Windows 365 Cloud PC

Enable flexibility for hybrid work.

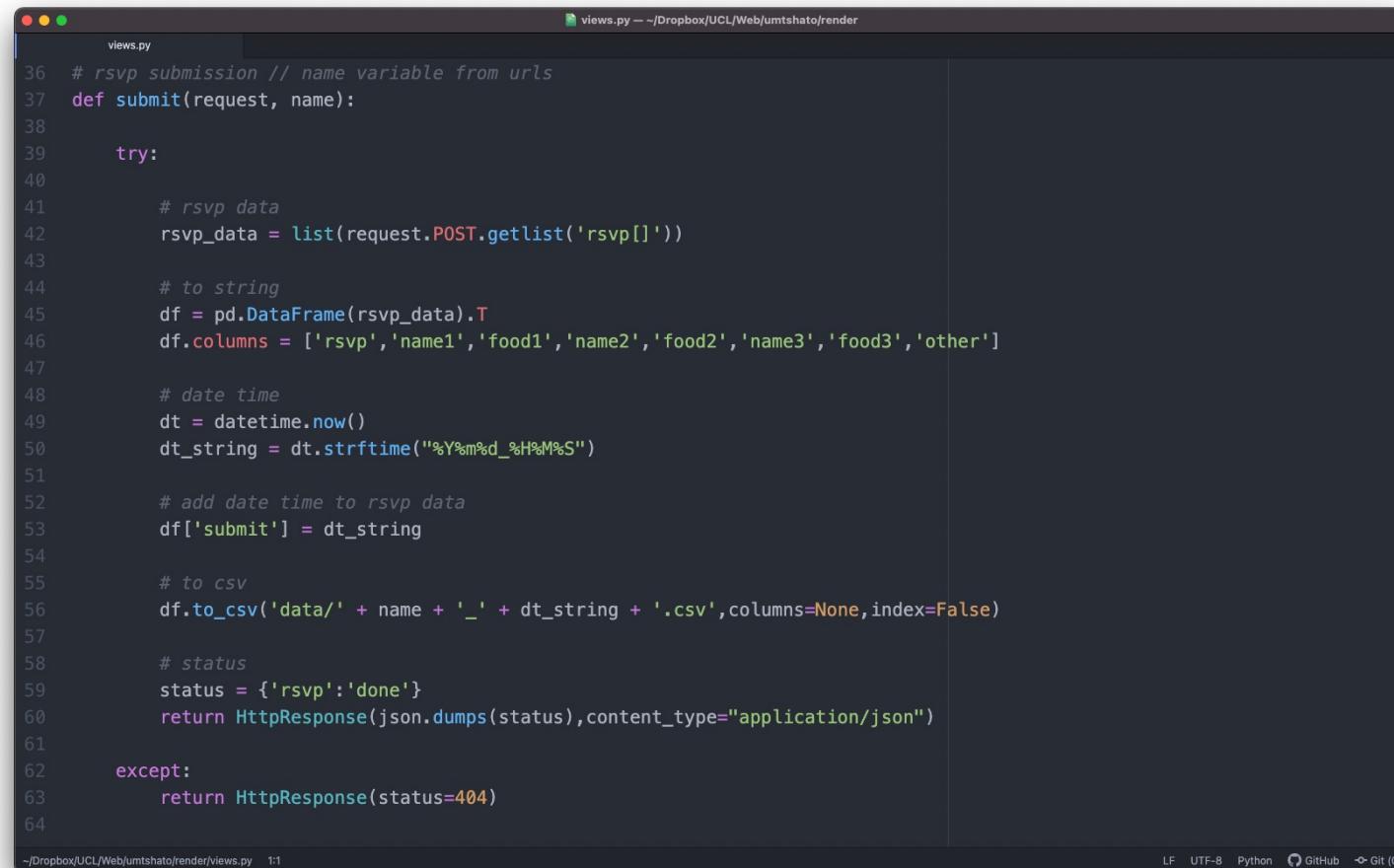
Try it for free

Related video

5 OF YOUR 5 A DAY

Sweet potato & white bean chilli: Jamie Oliver's food team

Algorithm



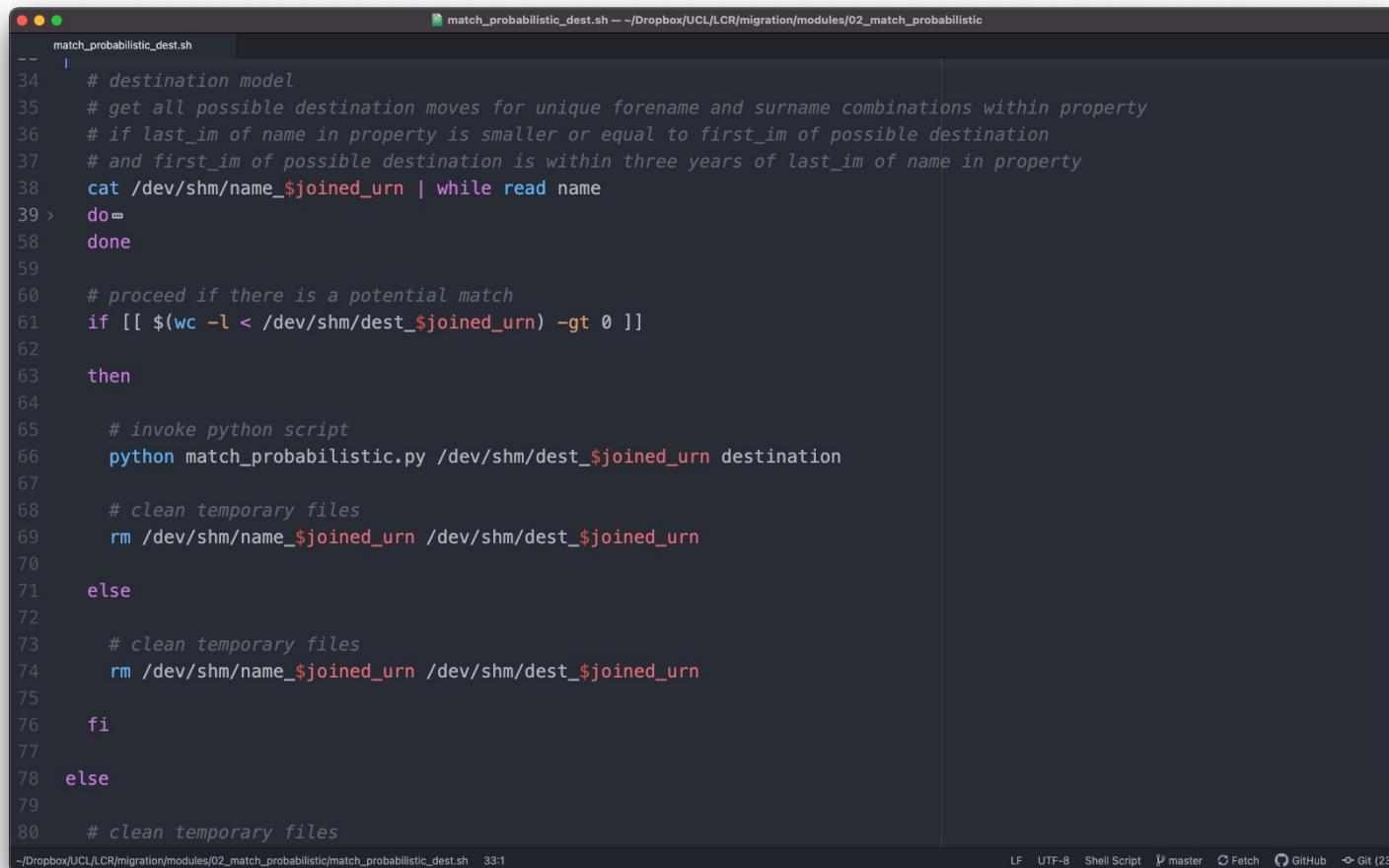
A screenshot of a code editor window titled "views.py". The code is a Python script for handling rsvp submissions. It includes imports for os, json, and HttpResponse. It defines a submit function that processes POST data, converts it to a DataFrame, adds a date/time column, and saves it as a CSV file. It also handles errors by returning a 404 status.

```
views.py
36 # rsvp submission // name variable from urls
37 def submit(request, name):
38
39     try:
40
41         # rsvp data
42         rsvp_data = list(request.POST.getlist('rsvp[]'))
43
44         # to string
45         df = pd.DataFrame(rsvp_data).T
46         df.columns = ['rsvp','name1','food1','name2','food2','name3','food3','other']
47
48         # date time
49         dt = datetime.now()
50         dt_string = dt.strftime("%Y%m%d_%H%M%S")
51
52         # add date time to rsvp data
53         df['submit'] = dt_string
54
55         # to csv
56         df.to_csv('data/' + name + '_' + dt_string + '.csv',columns=None,index=False)
57
58         # status
59         status = {'rsvp':'done'}
60         return HttpResponse(json.dumps(status),content_type="application/json")
61
62     except:
63         return HttpResponse(status=404)
64
```

~/Dropbox/UCL/Web/umtshato/render/views.py 1:1

LF UTF-8 Python GitHub Git (0)

Algorithm

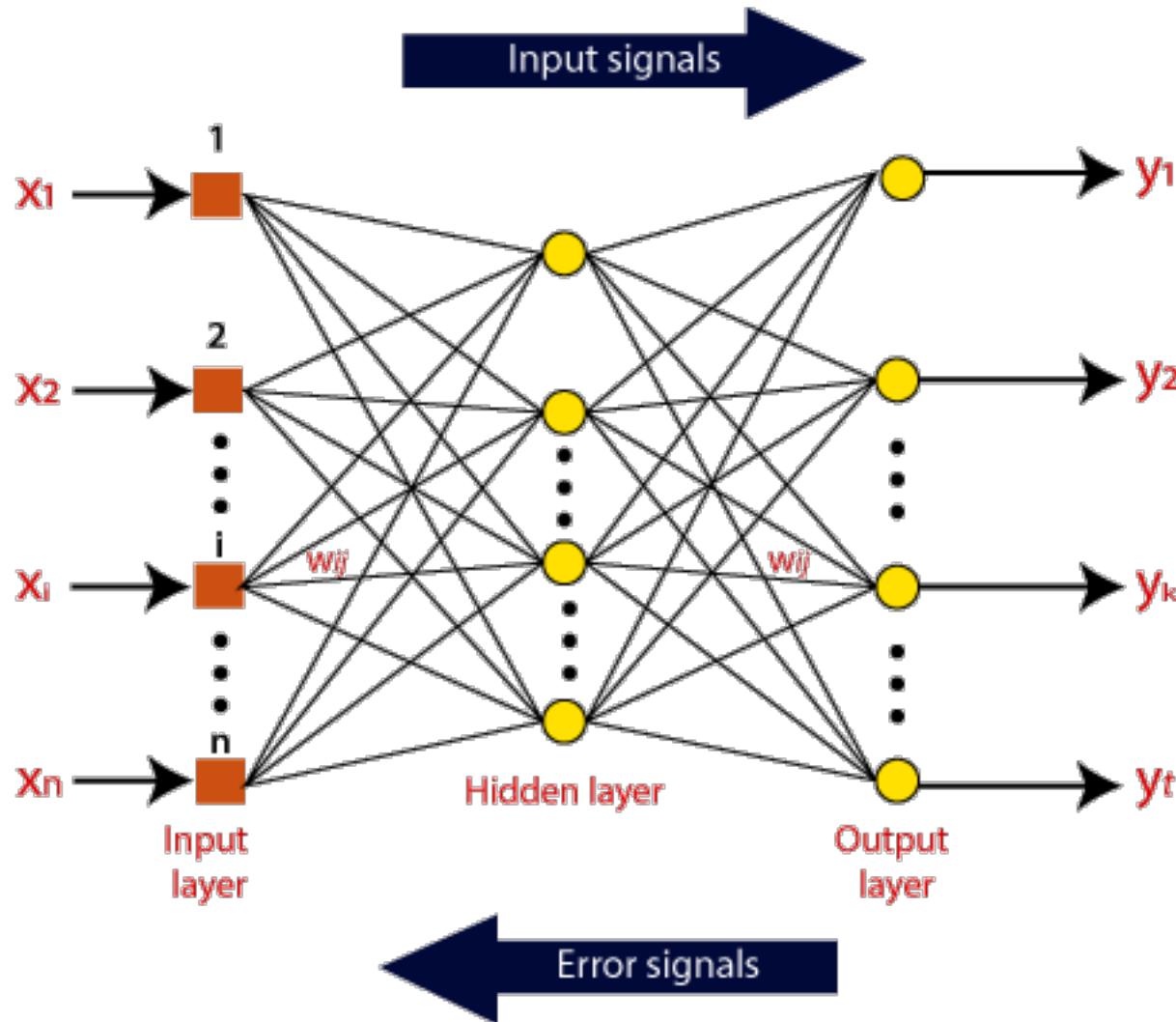


The screenshot shows a terminal window with a dark theme. The title bar reads "match_probabilistic_dest.sh ~/Dropbox/UCL/LCR/migration/modules/02_match_probabilistic". The script content is as follows:

```
match_probabilistic_dest.sh
34  I
35  # destination model
36  # get all possible destination moves for unique forename and surname combinations within property
37  # if last_im of name in property is smaller or equal to first_im of possible destination
38  # and first_im of possible destination is within three years of last_im of name in property
39 > cat /dev/shm/name_${joined_urn} | while read name
40 > do=
41 done
42
43 # proceed if there is a potential match
44 if [[ $(wc -l < /dev/shm/dest_${joined_urn}) -gt 0 ]]
45 then
46     # invoke python script
47     python match_probabilistic.py /dev/shm/dest_${joined_urn} destination
48
49     # clean temporary files
50     rm /dev/shm/name_${joined_urn} /dev/shm/dest_${joined_urn}
51
52 else
53     # clean temporary files
54     rm /dev/shm/name_${joined_urn} /dev/shm/dest_${joined_urn}
55
56 fi
57
58 else
59
60     # clean temporary files
61     rm /dev/shm/name_${joined_urn} /dev/shm/dest_${joined_urn}
```

At the bottom of the terminal window, the status bar displays: ~/Dropbox/UCL/LCR/migration/modules/02_match_probabilistic/match_probabilistic_dest.sh 33:1 LF UTF-8 Shell Script master Fetch GitHub Git (23)

Algorithm



Algorithmic bias

- Machine learning and AI: data-driven algorithms.
- Unsupervised machine learning is trying to recognise patterns in datasets.
- Supervised machine learning is all about learning from examples.
- Classification confidence goes up the more data is fed into such an algorithm.

Algorithmic bias



= CAT



= DOG

Algorithmic bias

So, what will happen?



Algorithmic bias

So, what will happen?



Algorithmic bias

So, what will happen?



Algorithmic bias

Dog or muffin?



Isolated incidents?

- Twitter's saliency algorithm, trained on eye-tracking data, used to crop images
- It was found that the algorithm showed:
 - An 8% difference in favour of women over men
 - A 4% difference favouring white people over black people of both sexes
 - A 7% difference favouring white women over black women
 - A 2% difference in favour of white men over black men

Isolated incidents?

- Software used in self-driving cars.
- 2019 research by Georgia Institute of Technology tested several models used by academic researchers, trained on publicly available datasets.
- Results: if you're a person with dark skin, you may be more likely than your white friends to get hit by a self-driving car because automated vehicles may be better at detecting pedestrians with lighter skin tones.
- *Note that the research did not use actual data from autonomous car manufacturers as they refuse to make these available.*

Isolated incidents?

Examples are far from limited to data science and novel automated procedures in the domain of machine learning and artificial intelligence.

Isolated incidents?

Medical textbook for nurses with a section on “cultural differences in response to pain”.

Some actual quotes:

“Pain is considered a test of faith. Muslim clients must endure pain as a sign of faith in return for forgiveness and mercy.”

“Black people often report higher pain intensity than other cultures because they believe pain must be shared and validated by others.”

Isolated incidents?

London Metropolitan Police's Gang Database:



80%
12-24 year old



78%
Black people



75%
have been victims of
violence themselves



35%
never committed any
serious offences

Isolated incidents?

NHS

From oximeters to AI, where bias in medical devices may lurk

Analysis: issues with some gadgets could contribute to poorer outcomes for women and people of colour

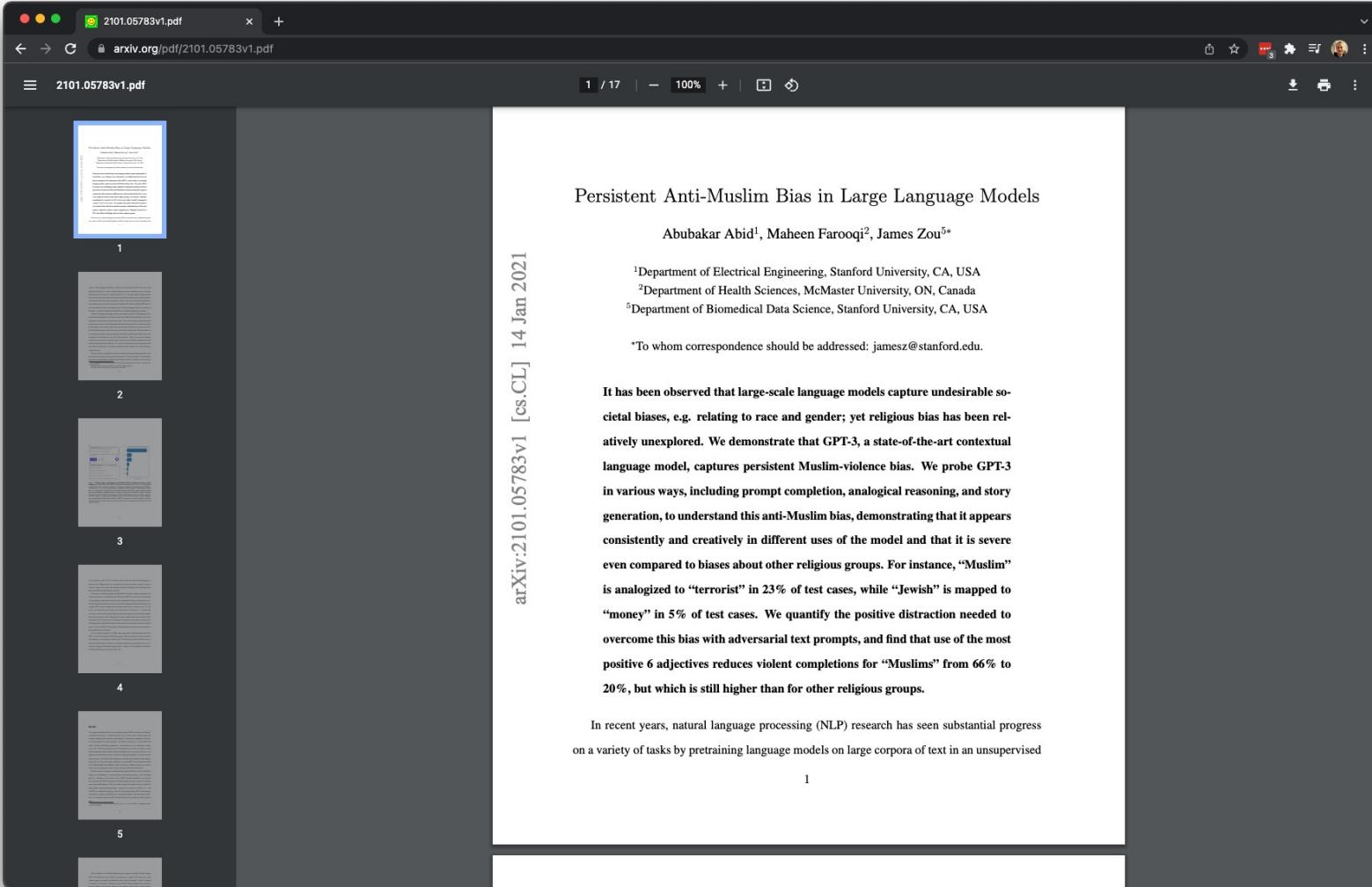
Nicola Davis
*Science
correspondent*

✉ @NicolaKSDavis

Sun 21 Nov 2021 17.56
GMT



Isolated incidents?



Conclusion

- We explored some of the 'ugly' side of (large) human-generated datasets and datafication.
- Neutrality within data science, nor the wider social science, does not exist. This is not new, however, existing biases are perpetuated and exacerbated – and in some cases institutionalised (e.g. Met's Gang matrix).
- Data are created and analysed within socio-political assemblages that inscribe it with particular biases and representations. Gets worse with black-box algorithms. **Ugly!**
- Thomas theorem: "When people define situations as real, they become real in their consequences."

Conclusion

Piketty 2017, p.3:

"Social scientific research is and always will be tentative and imperfect. It does not claim to transform economics, sociology, and history into exact sciences. But by patiently searching for facts and patterns and calmly analyzing the economic, social, and political mechanisms that might explain them, it can inform democratic debate and focus on the right questions."

... we probably should keep both Thomas and Piketty in mind.

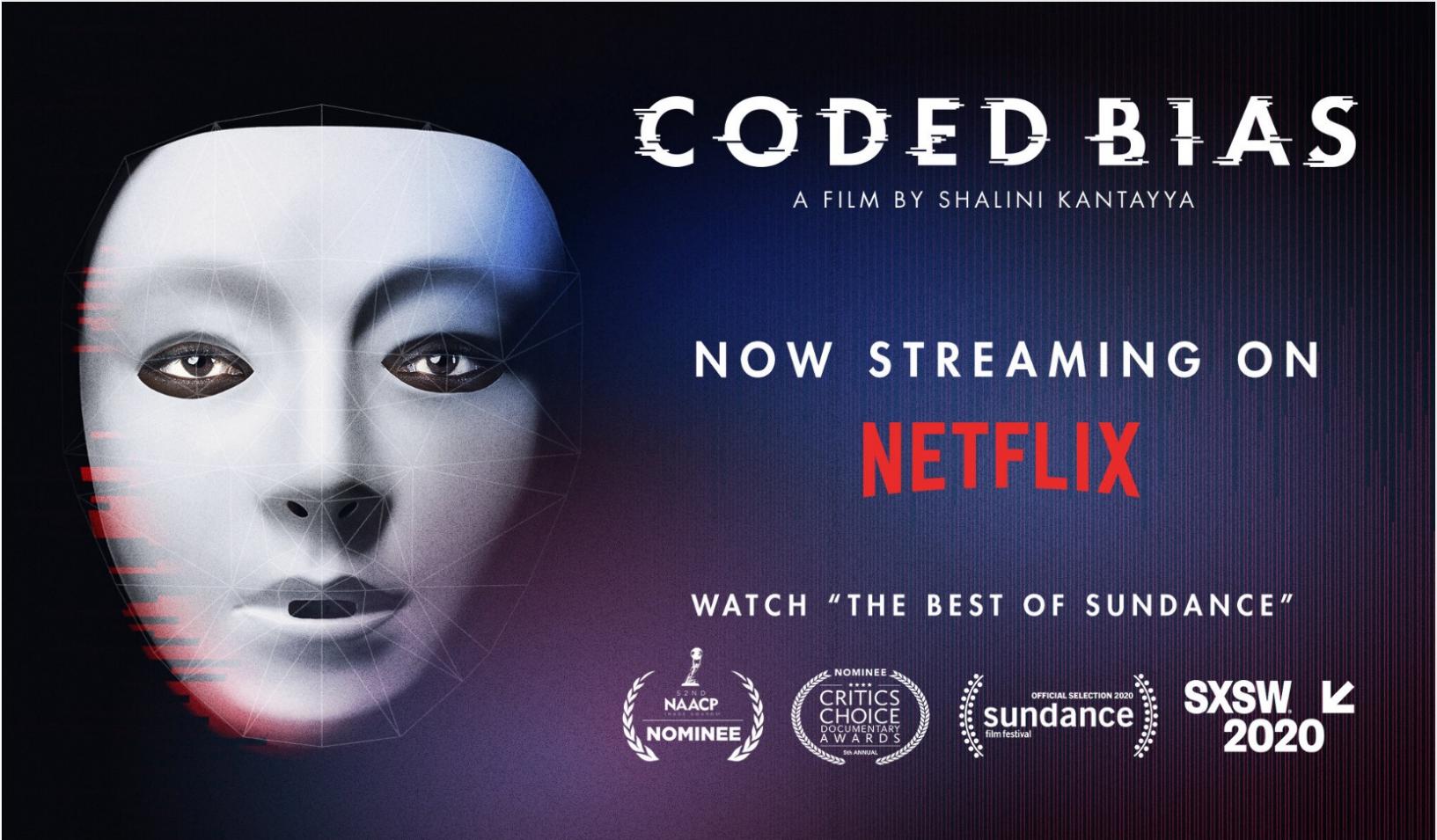
Seminar preparation

In preparation for the next seminar, carefully read the following article by Denton *et al.* 2021 and think about the linkages between this article and the other readings on this week's list:

Denton, E., Hanna, A., Amironesei, R. *et al.* 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*.

You do not have to submit anything, however, do come prepared as this article will be guiding our discussion.

Netflix



Questions

Justin van Dijk

j.t.vandijk@ucl.ac.uk

