Data collection

From reality to dataset

Prompt: *Planet Earth being pulled into a computer screen by a robot hand, early 3d CGI*

Schedule

- From objects to data model
- Data collection: concepts and applications
- Webscraper.io workshop

---------- **Break** ----------

- DEDA workshop with Julia Straatman
- Social media as a source
- Data collection for group project

Exercise: objects > data model

Create a data model for the objects you've been given, with columns and data type (30 minutes).

author_lastname (text)

publication_date (date, dd-mm-yy)

From objects to data

- Datasets are a logical representation of data
- A data model indicates the structure of your set: the categories, the interrelationships, the rules
- The data model is represented by means of a diagram

Alexander Severus
A.D. 231-235

Denomination
Denarius (67)

Material
Silver (87)

Iconography
"spear" (62)

Script
Latin (471)

Institution
U.Va. (494)

Department
Roman (448)

Legend
"imperator" (266)

Collection
U.Va. Miscellaneous (48)

Index Terms

Century
A.D. 3rd (363)

Deity
Mars (13)

Dynasty
Severan (14)

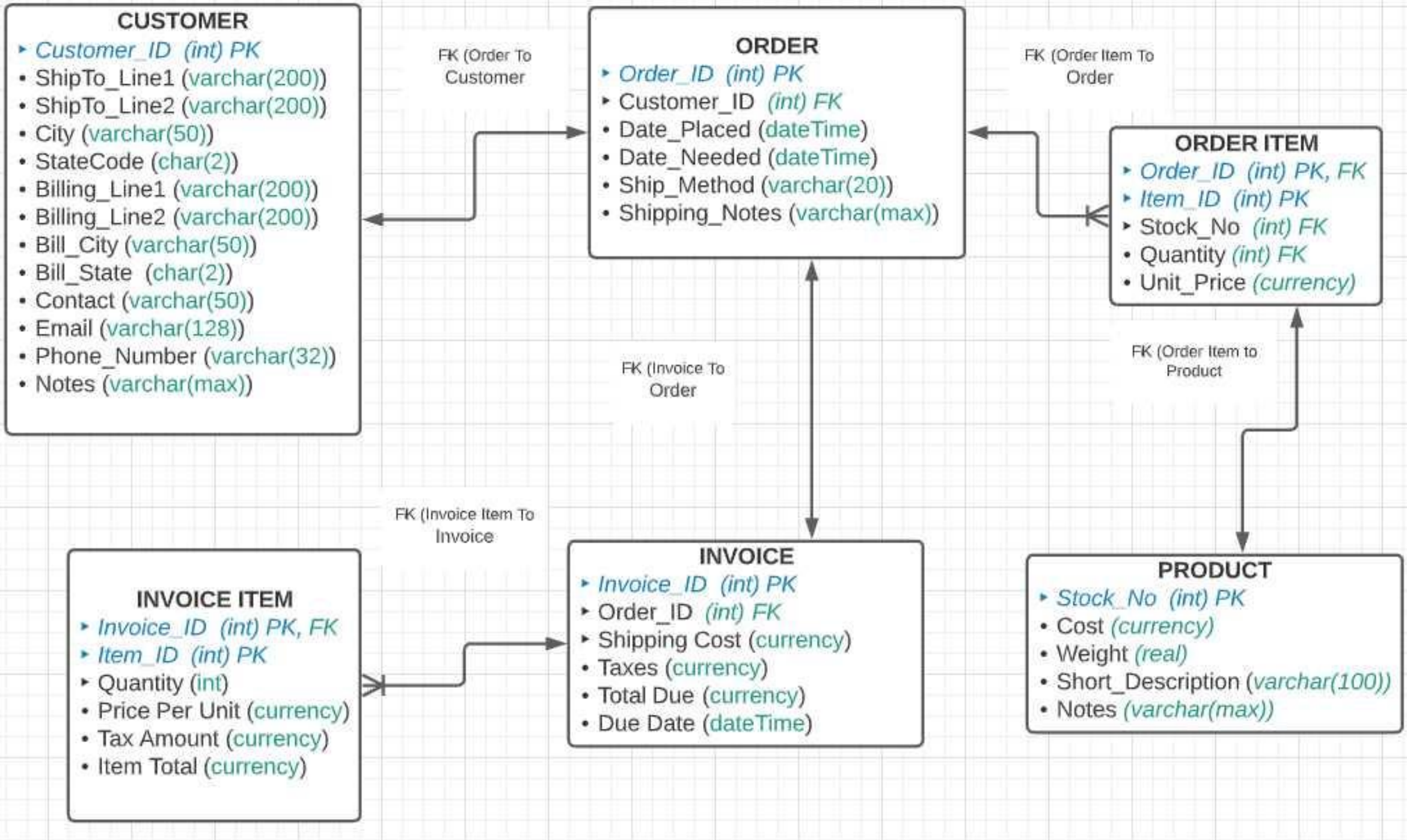
Geographical Locations

Names
Alexander Severus (4)

State
Italy (277)

Region
Lazio (239)

City
Rome (239)


CUSTOMER
Customer_ID (int) PK
ShipTo_Line1 (varchar(200))
ShipTo_Line2 (varchar(200))
City (varchar(50))
StateCode (char(2))
Billing_Line1 (varchar(200))
Billing_Line2 (varchar(200))
Bill_City (varchar(50))
Bill_State (char(2))
Contact (varchar(50))
Email (varchar(128))
Phone_Number (varchar(32))
Notes (varchar(max))
FK (Order To Customer
ORDER
Order_ID (int) PK
Customer_ID (int) FK
Date_Placed (dateTime)
Date_Needed (dateTime)
Ship_Method (varchar(20))
Shipping_Notes (varchar(max))
FK (Order Item To Order
ORDER ITEM
Order_ID (int) PK, FK
Item_ID (int) PK
Stock_No (int) FK
Quantity (int) FK
Unit_Price (currency)
FK (Invoice To Order
FK (Order Item to Product
FK (Invoice Item To Invoice
INVOICE ITEM
Invoice_ID (int) PK, FK
Item_ID (int) PK
Quantity (int)
Price Per Unit (currency)
Tax Amount (currency)
Item Total (currency)
INVOICE
Invoice_ID (int) PK
Order_ID (int) FK
Shipping Cost (currency)
Taxes (currency)
Total Due (currency)
Due Date (dateTime)
PRODUCT
Stock_No (int) PK
Cost (currency)
Weight (real)
Short_Description (varchar(100))
Notes (varchar(max))

Data collection: key questions

- What do we want to study?
- What data can we use for this?
- Where can this data be found?
- How do we get it?
- How do we wrangle the data into the desired format?

Finding data

There are different ways to get to the data you need.

For example:

- Open data
- API Access
- (Paid) data suppliers
- Data breaches
- Automated Web Scraping
- Manual data collection

Open data

- Freely available with open license
- Made available by governments, non-profit organizations, researchers
- Searchable via Google Dataset Search
- There's more available than you might think: just search for "[name of organization] open data“ or “dataset”

API Access

- Application Programming Interface: Building content using a live connection to a database
- Direct access to data from the database
- Sometimes has entry requirements, such as affiliation with an academic institution
- E.g. Spotify, Telegram, YouTube, Facebook

Telegram APIs

We offer two kinds of APIs for developers. The **Bot API** allows you to easily create programs that use Telegram messages for an interface. The **Telegram API and TDLib** allow you to build your own customized Telegram clients. You are welcome to use both APIs free of charge.

You can also add **Telegram Widgets** to your website.

Designers are welcome to create **Animated Stickers** or **Custom Themes** for Telegram.

Bot API

This API allows you to connect bots to our system. **Telegram Bots** are special accounts that do not require an additional phone number to set up. These accounts serve as an interface for code running somewhere on your server.

To use this, you don't need to know anything about how our MTProto encryption protocol works — our intermediary server will handle all encryption and communication with the Telegram API for you. You communicate with this server via a simple HTTPS-interface that offers a simplified version of the Telegram API.

Learn more about the Bot API here »

Bot developers can also make use of our **Payments API** to accept **payments** from Telegram users around the world.

TDLib – build your own Telegram

Even if you're looking for maximum customization, you don't have to create your app from scratch. Try our **Telegram Database Library** (or simply TDLib), a tool for third-party developers that makes it easy to build fast, secure and feature-rich Telegram apps.

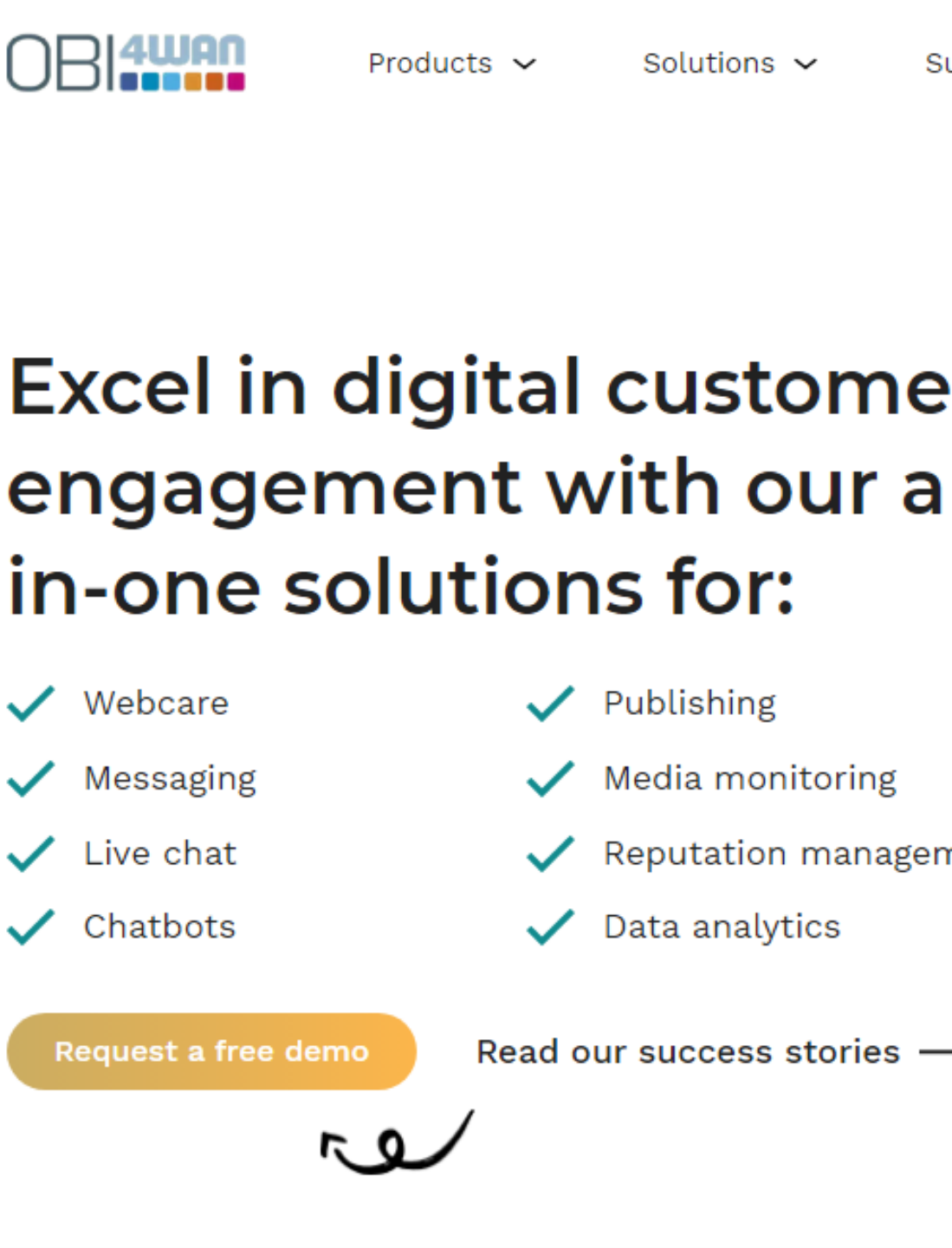
TDLib takes care of all **network implementation** details, **encryption** and **local data storage**, so that you can dedicate more time to design, responsive interfaces and beautiful animations.

TDLib supports all Telegram features and makes developing Telegram apps a breeze on any platform. It can be used on Android, iOS, Windows, macOS, Linux and virtually any other system. The library is open source and compatible with virtually **any programming language**.

Learn more about TDLib here »

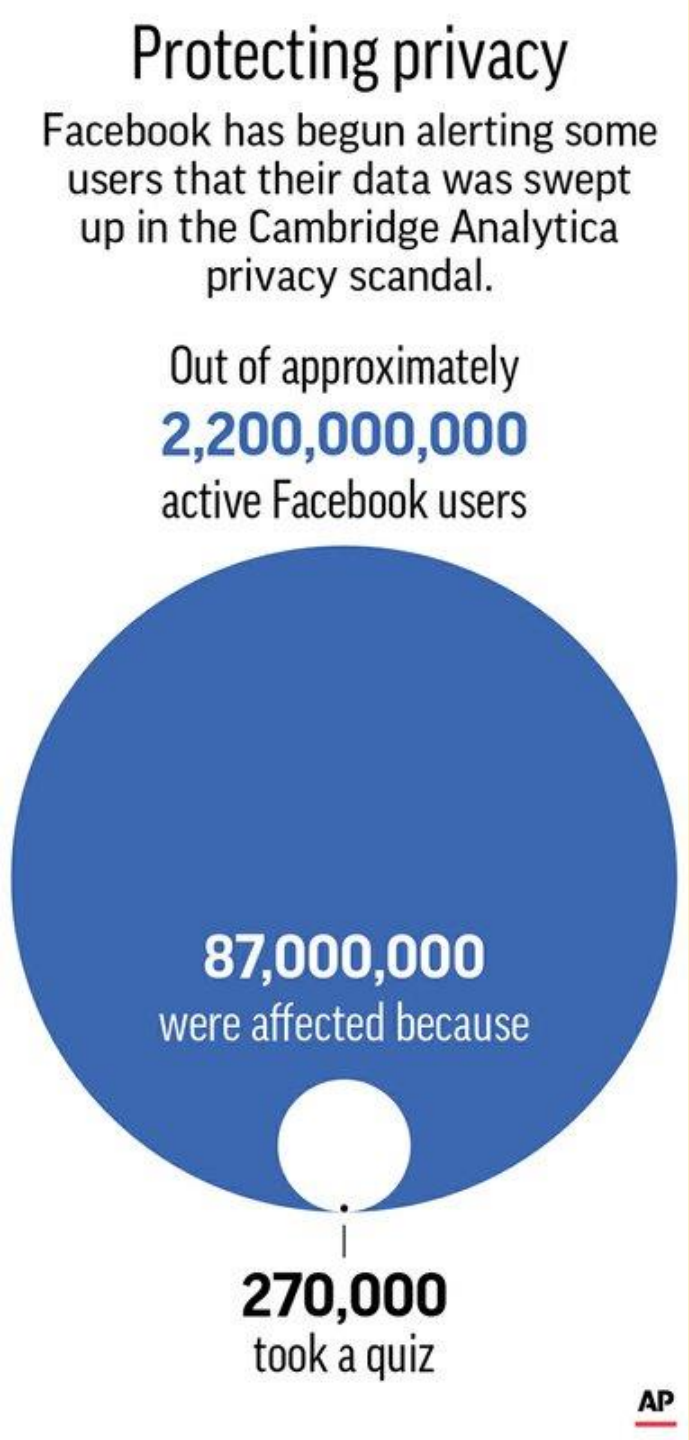
Paid data providers

- Also called ‘data brokers’
- Paying for access to data
- Often combine lots of datasets
- Many focused on personal information (contact details, income level, consumer behavior)
- Examples: Focum, LexisNexis, Spotlr



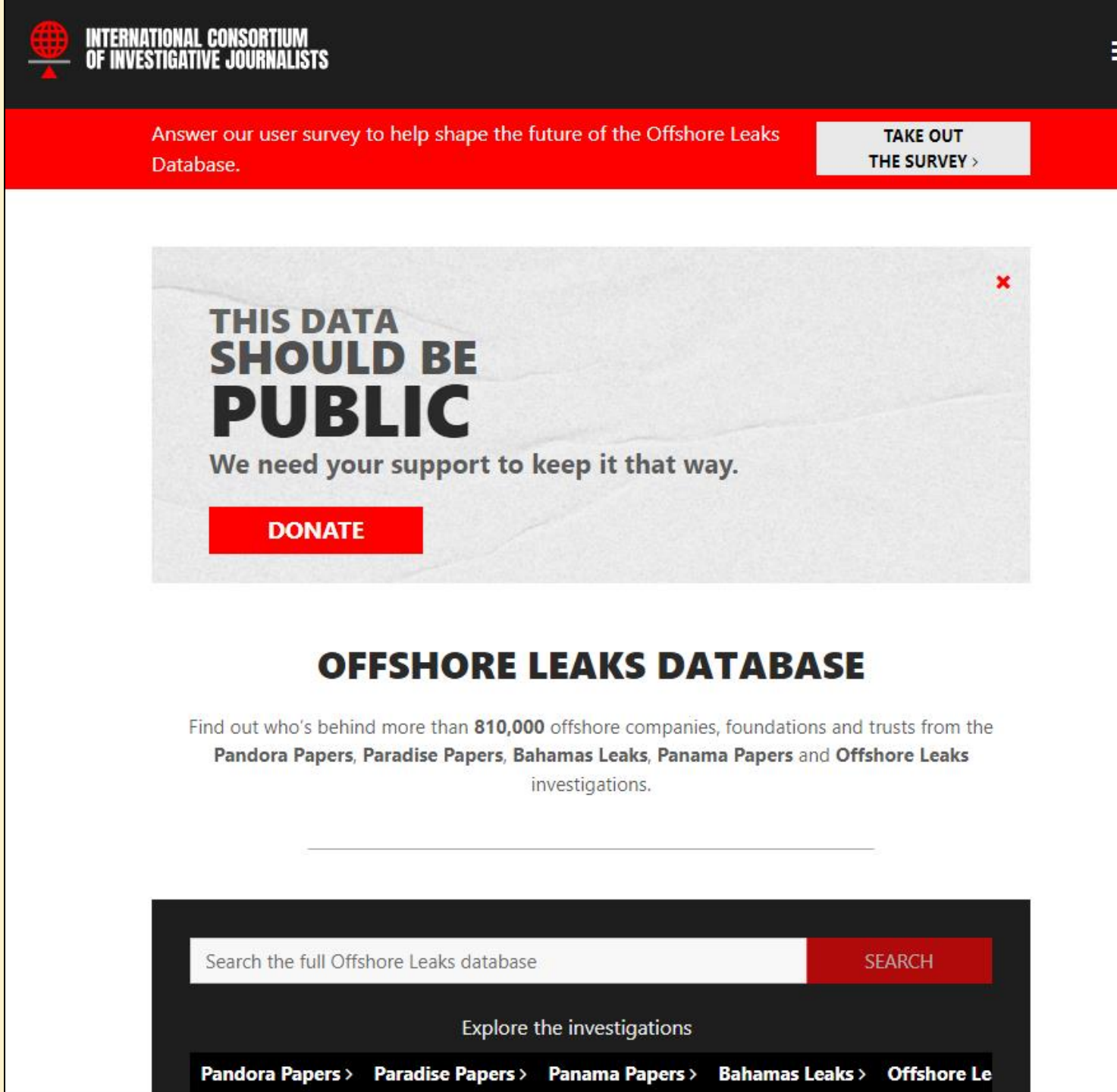
Paid data providers

- Can be questionable: think Cambridge Analytica
  - Commercial companies that aim to get as detailed information as possible about people, sometimes using dubious methods
- Sometimes a last-ditch effort
  - Some keep archive of social media data that isn't available otherwise



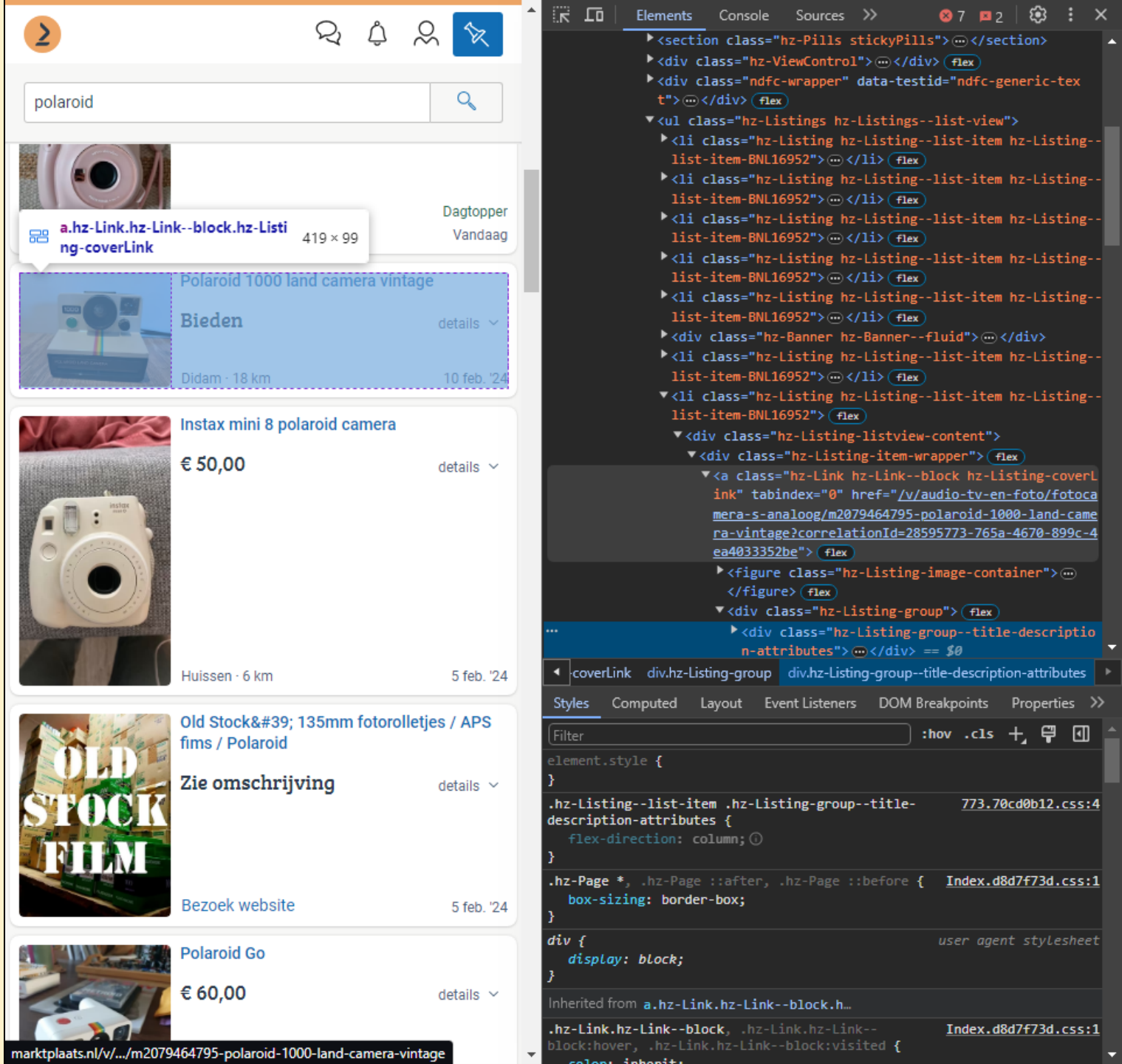
Data leaks

- Secret information leaked by whistleblowers
- Often formatted in an easily accessible format to encourage research
- Offshore Leaks Database (Panama Papers, Paradise Papers, Pandora Papers)
- WikiLeaks
- Not without risks



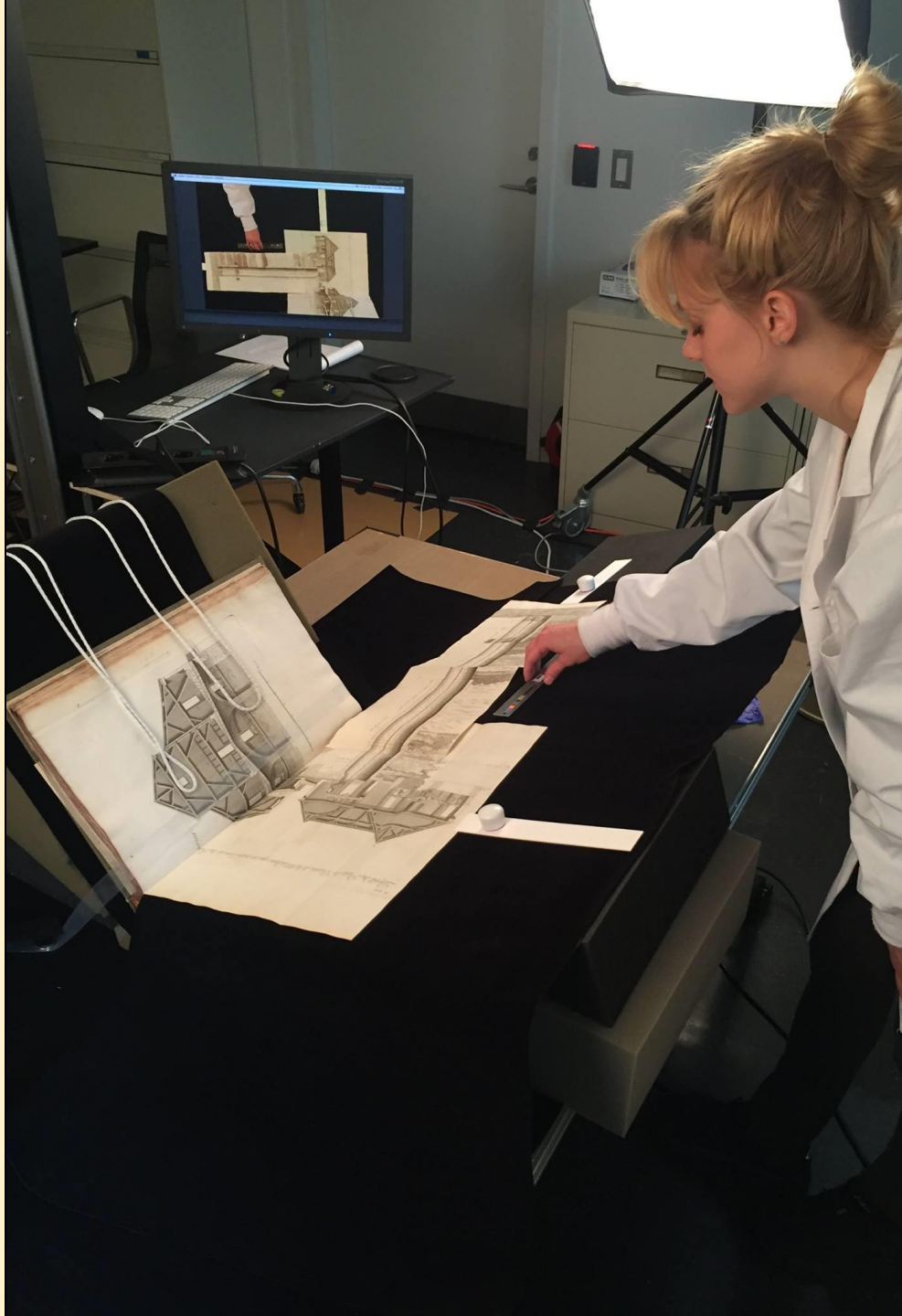
Web scraping

- Automatic collection of web data based on patterns in web page
- Can be used to collect any type of structured page
- Zeeschuimer (for social media)
- Web Scraper plugin
- BeautifulSoup (Python)



Manual data collection

- Old-school
- Sometimes necessary, for example when working with non-digital archive material
- Purpose: Create a structured data table from unstructured sources



Exercise: what data?

Objective: To investigate how the 2024 Summer Olympics in Paris affect the local tourism industry (accommodation, transport, restaurants, etc.).

What data would you use? Discuss with your group and prepare a 2-minute pitch, explaining:

… what types of data you would use

… what perspectives on the topic these data can offer

… where you think you can get this data

Time: 10 minutes

Discussion

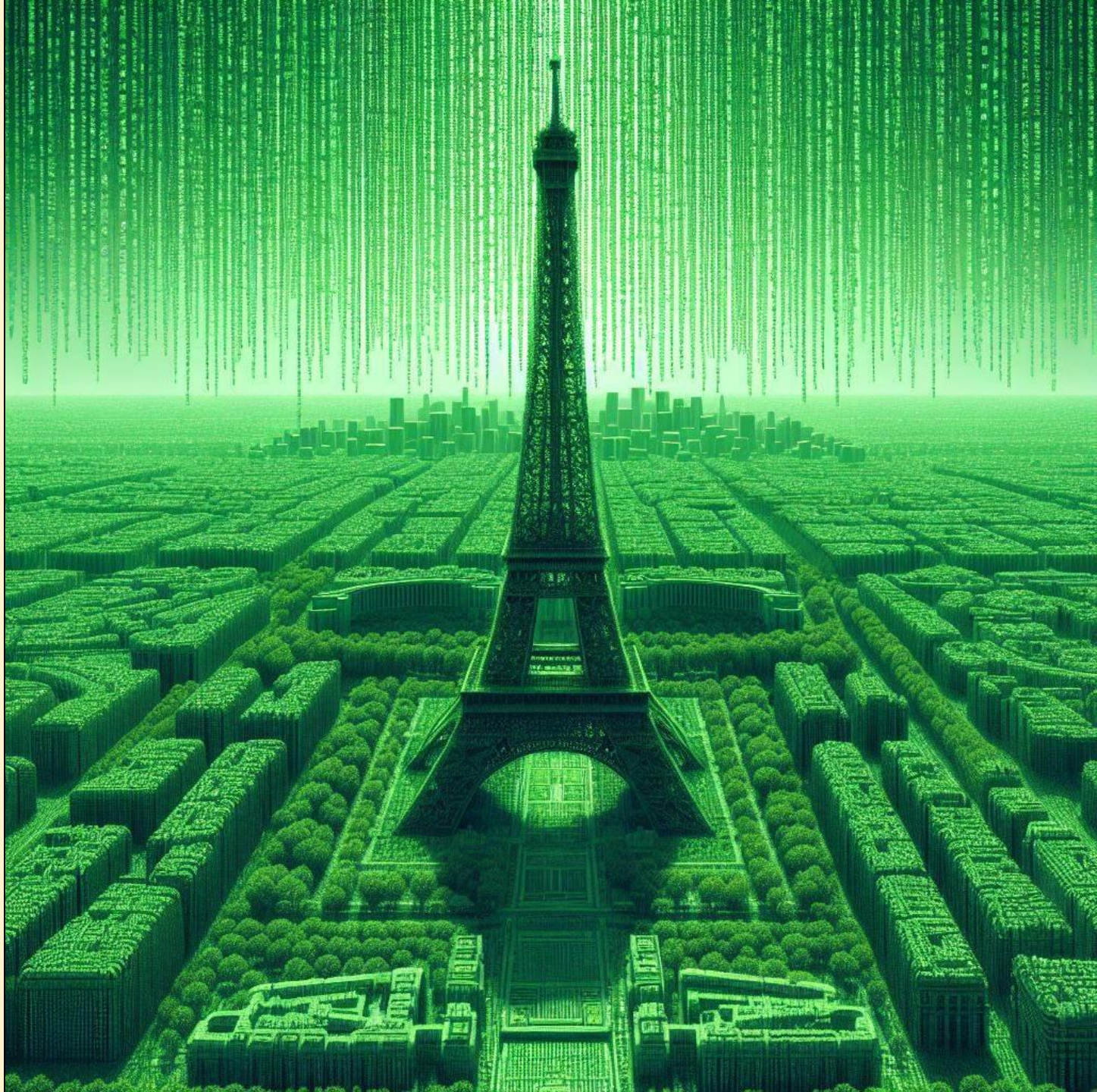
Explain...

...what types of data you would use

... what perspectives on the topic this data can provide

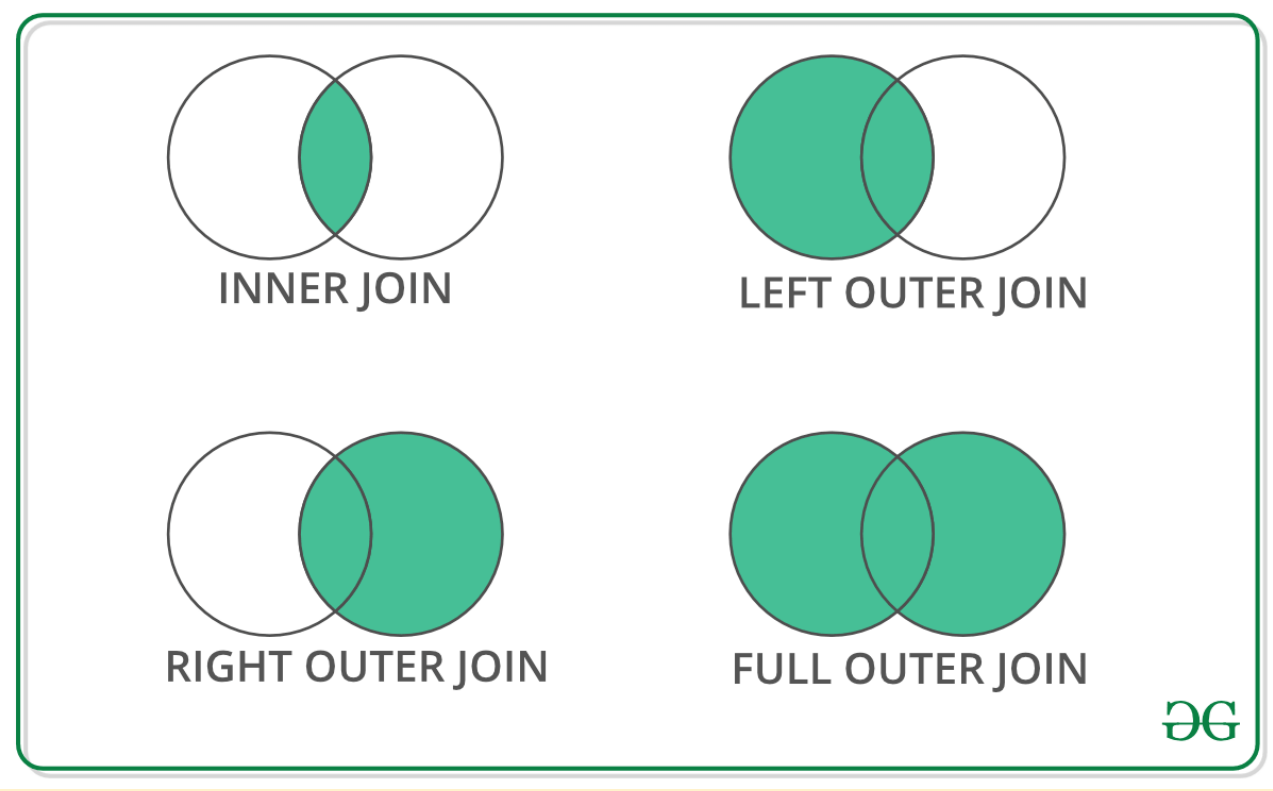
... where you think you can get this data

Prompt: *The Eiffel Tower, but represented like in the Matrix, made of green rows of letters and numbers.*



Enriching data

- Add more information to your data
- Methods:
  - Combining datasets
  - Merge with other datasets, often based on the same column (i.e., name, product code)
  - Annotation
  - Manually add more information based on annotation scheme



Quality control

- All data is created directly or indirectly by humans
  - Motives
  - Mistakes can happen!

- Data are not a given!
- How would you check the quality of a dataset?

Quality Control Checklist

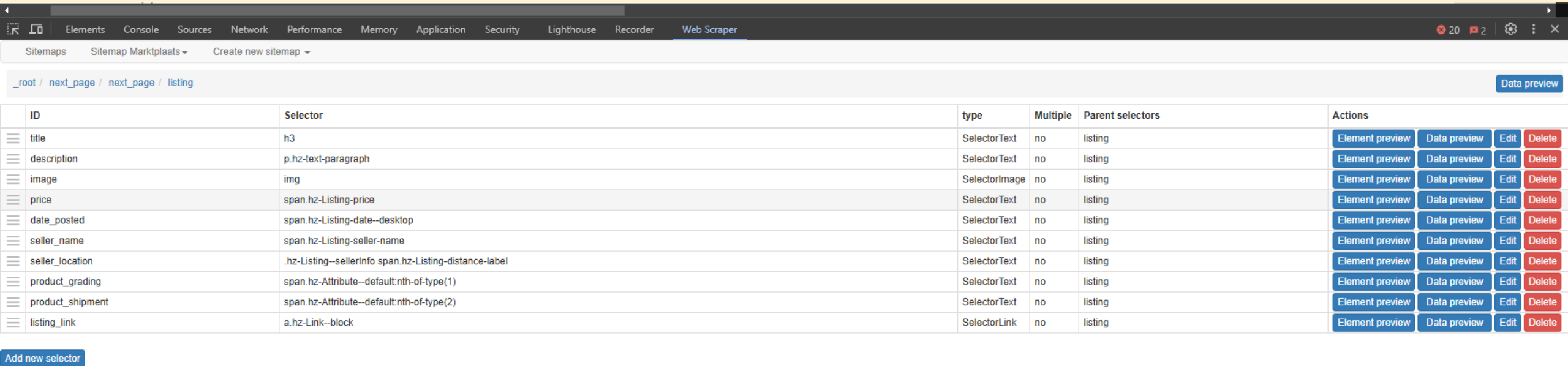
- ✓ What is the source?
- ✓ What are their motives for sharing this data?
- ✓ What is and isn’t included in the data?
- ✓ What methods were used to collect this data?

Summary

- Data is not a given: data are never just there
- Data have always made a translation: reality > data
  - Result: abstraction, loss of details
- There are a thousand ways to get data, but a critical eye is always needed
- Do the checklist!

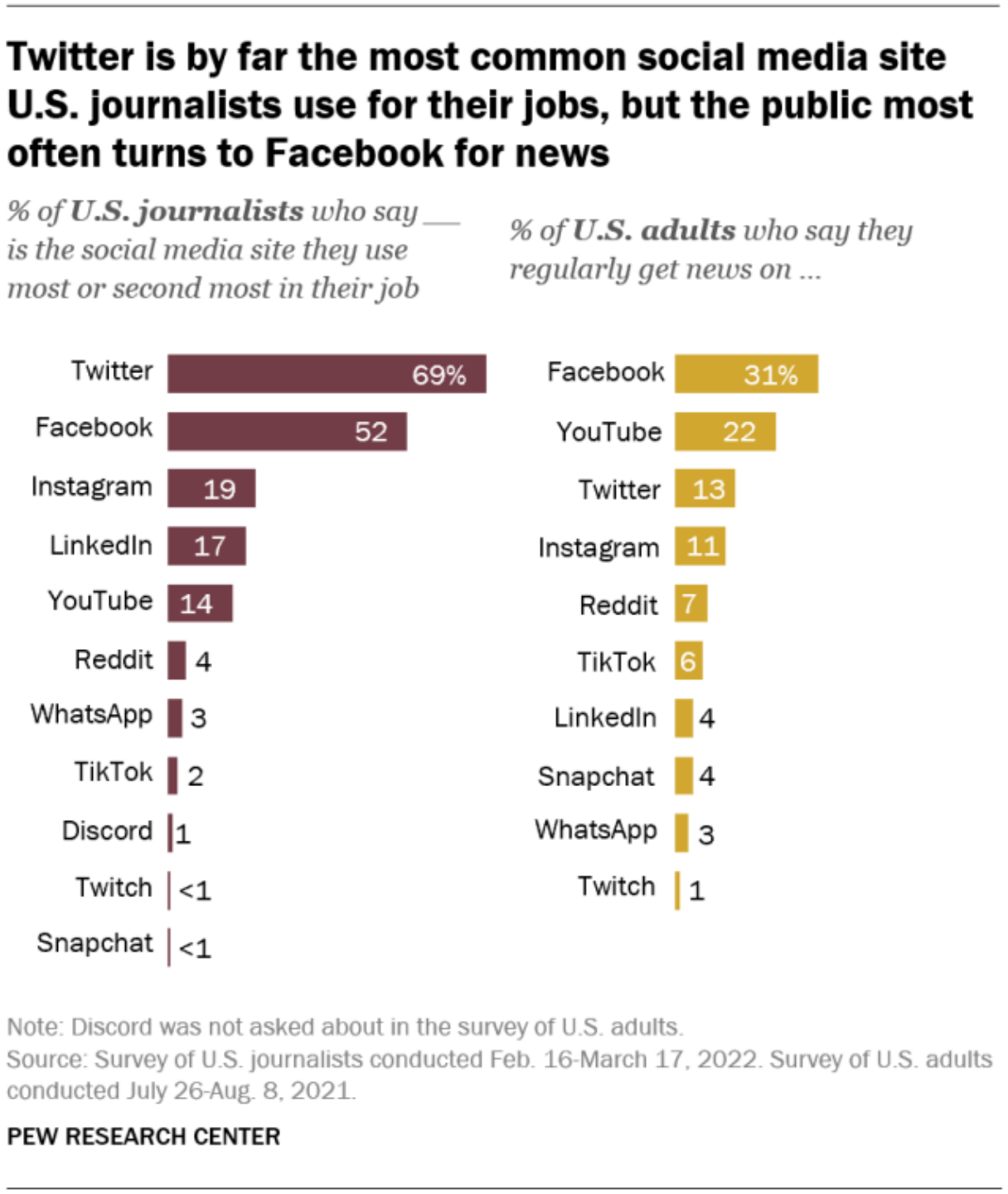
Webscraper.io

- Free plugin
- Graphical user interface for building web scrapers
- No coding needed!



```
for time in day:
    if time == 11:30:
        break
    print('We'll continue at 12:30!')
```

Social media as a source



Pew Research Center. (2022, June 24). *Twitter is by far the most common social media site U.S. journalists use for their jobs, but the public most often turns to Facebook for news | Pew Research Center*. https://www.pewresearch.org/short-reads/2022/06/27/twitter-is-the-go-to-social-media-site-for-u-s-journalists-but-not-for-the-public/ft_2022-06-27_journalistssocialmedia_01/

Pros and sons

- Advantages
- Extremely fast way to follow news events
  - Citizen reporting
  - Unmediated communication
  - (Almost) everyone has a voice
- Disadvantages
- The rest of this lecture

Unmediated communication?

- Yes, anyone can post and share...
  - … But all platforms are driven by algorithms
  - Automated mediation
  - Visibility is determined by an opaque system
- The platform is a medium
- Own customs, possibilities and limitations

Filter bubbles

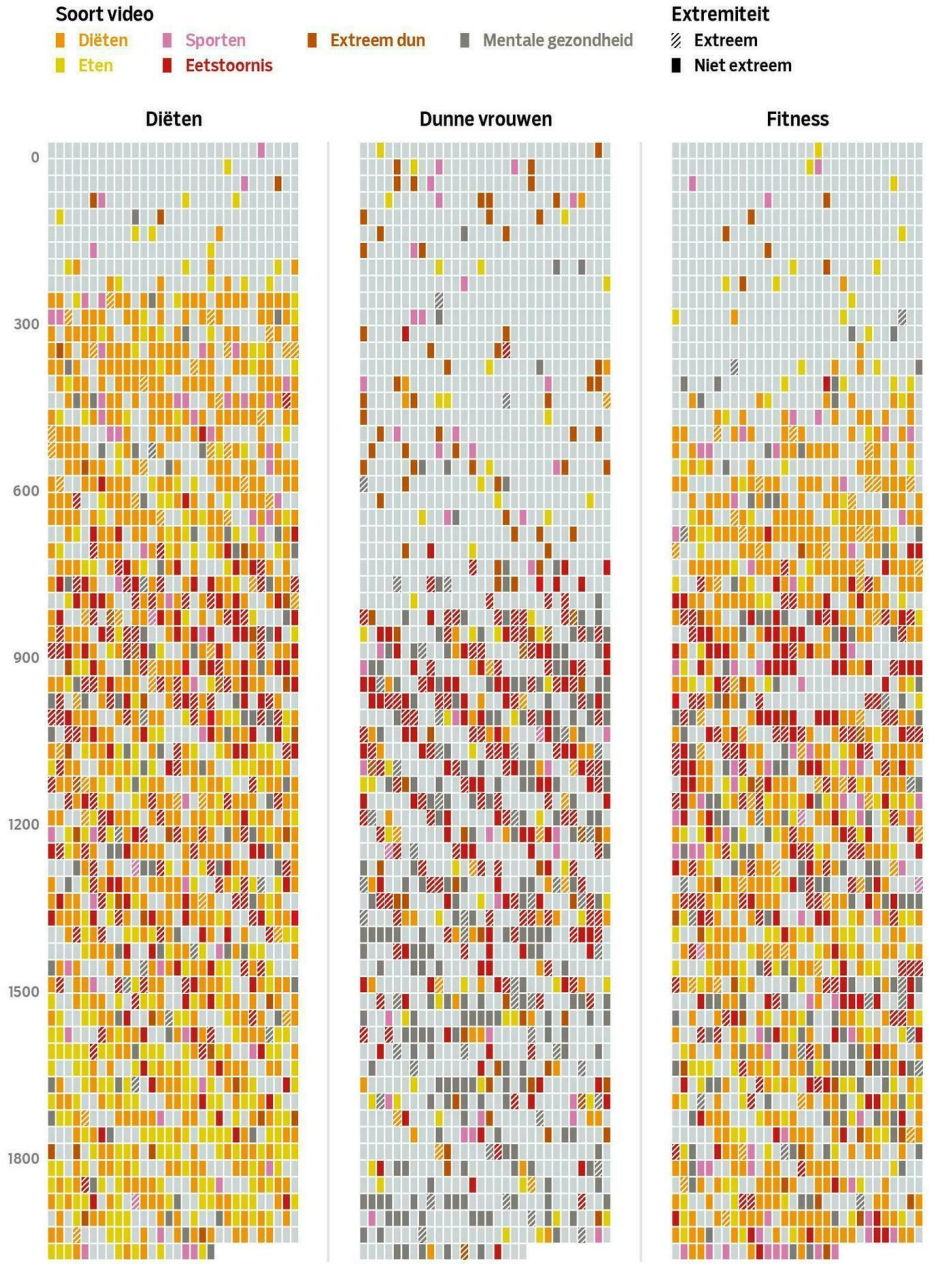
Prompt: *Groups of people floating in bubbles high in the sky, dramatic*

Filter bubbles

- Websites target content to users
- Recommendation algorithms
- Personalized feeds
- Filter bubbles (Pariser)
  - Information has become personalized
  - Users are enclosed in a filtered ecosystem
  - Presumed effect: people are confirmed in their own prejudices

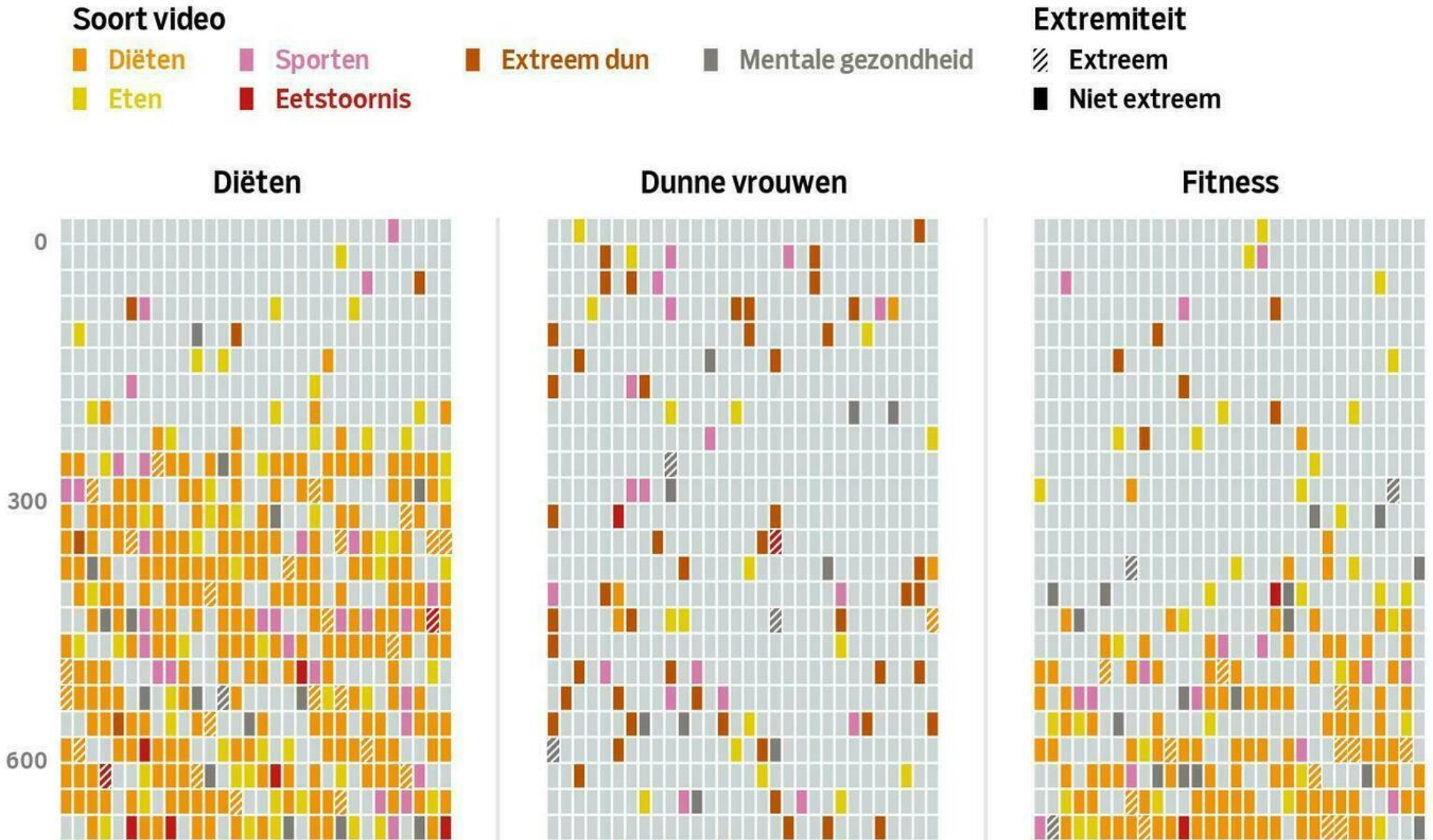
Een constante stroom aan extreme diëten, eetstoornisvideo's en slanke lichamen

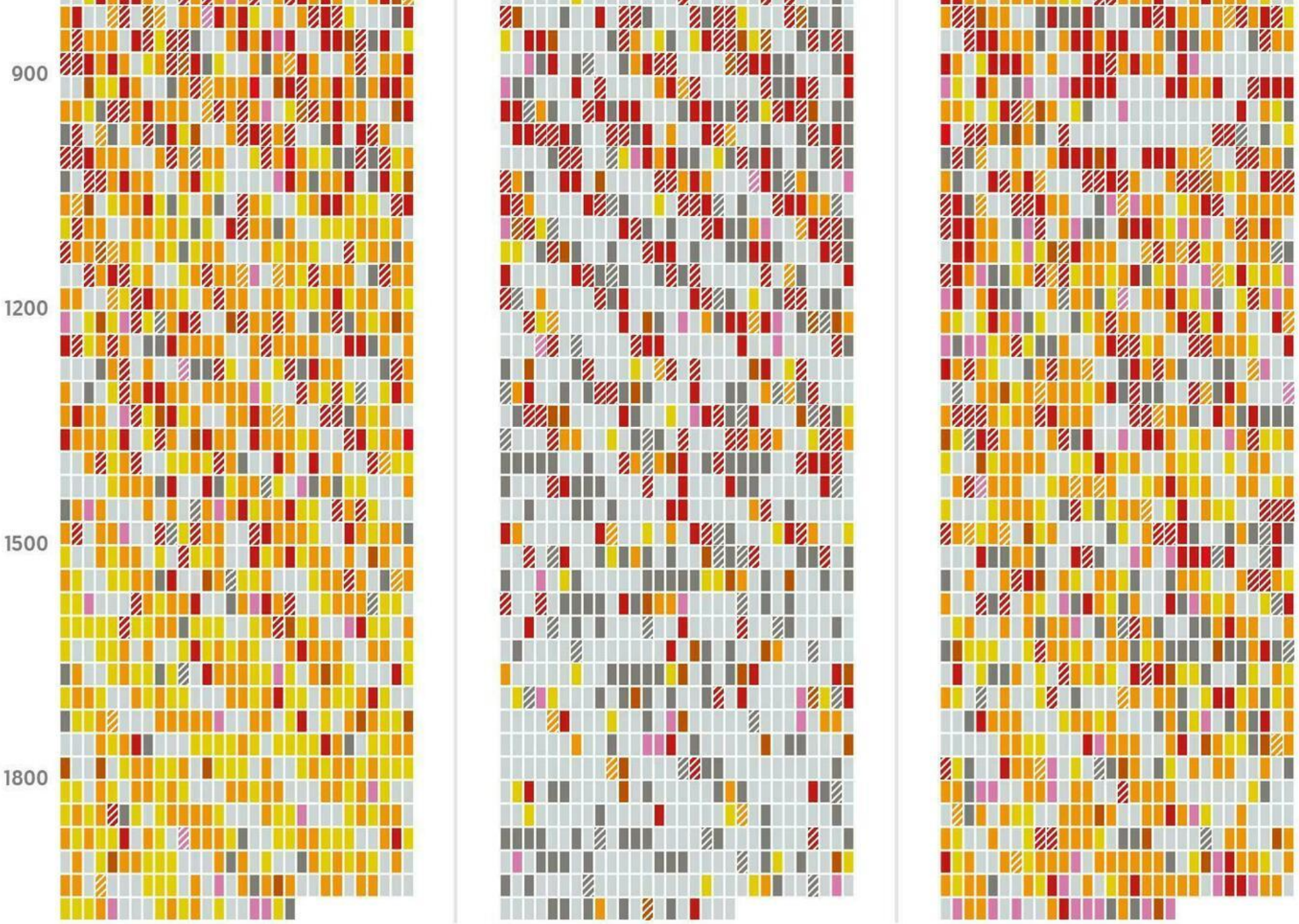
De eerste tweeduizend video's van drie accounts die scrollen vanuit een interesse in diëten, dunne vrouwen of fitness



Een constante stroom aan extreme diëten, eetstoornisvideo's en slanke lichamen

De eerste tweeduizend video's van drie accounts die scrollen vanuit een interesse in diëten, dunne vrouwen of fitness




900
1200
1500
1800

In recent years researchers have however begun to question this explanation. 'The main problem is that we just aren't finding any echo chambers,' says Törnberg. 'In fact, studies suggest that social media is characterized by *more* interaction outside our local network, and more interaction with political opponents than in our offline life.'

Universiteit van Amsterdam. (2022, October 11). *Social media polarize politics for a different reason than you might think*. University Of Amsterdam. https://www.uva.nl/en/shared-content/faculteiten/en/faculteit-der-maatschappij-en-gedragswetenschappen/news/2022/10/social-media-polarize-politics-for-a-different-reason-than-you-might-think.html

Tweet

President Biden ✓
@POTUS

Some great news:

We've come to an agreement with Congressional leaders on a path forward for the remaining full-year funding bills.

The House and Senate are now working to finalize a package that can quickly be brought to the floor, and I will sign it immediately.

6:35 PM · Mar 19, 2024 · **510.5K** Views

1.4K 2K 10K 75

Replies

Freedom @PU28453638 · 3h
The great news is that they are com back!

9 15 659

Shoegal8720 @shoegal8720 · 2h
I have some great news. I just voted for Donald Trump and so did my husband and children.

8 4 62 632

ZNO @therealZNO · 3h
Full-year funding bills for who?

If these bills send money and aid to anyone other than Americans and the United States, it should be DOA.

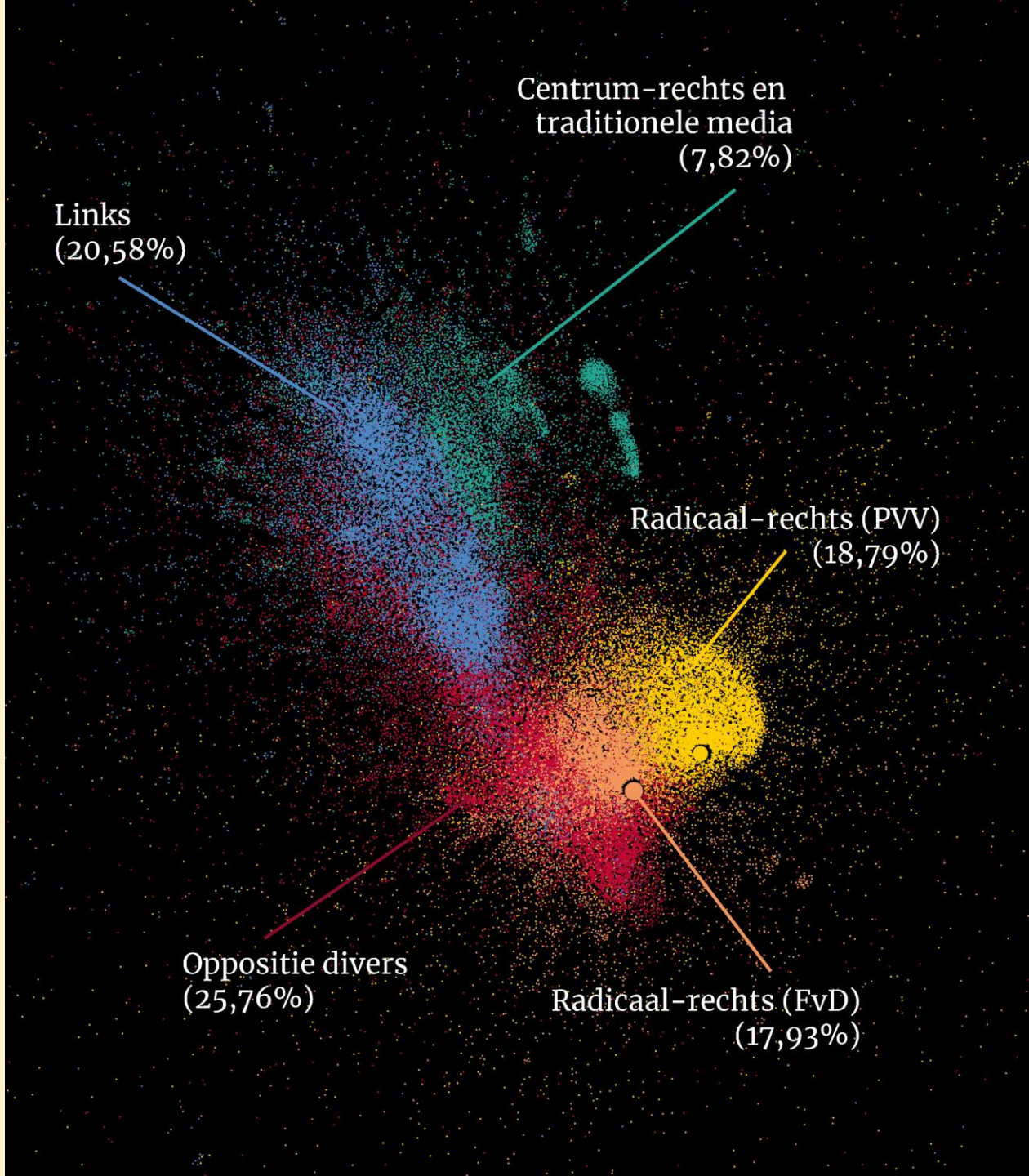
Stop prioritizing foreign countries over the welfare of Americans citizens.

Put America First.

8 7 98 1.6K

Researchers are not immune

- We are all influenced by algorithms
- A researcher's feed is as biased as any other
- Social media audiences are not a reflection of the real world
- Target audiences differ



Alternative: APIs

- Instant access to data through keyword search
- Search results are not influenced by algorithm
- However,...
  - Technical hurdles: often requires programming skills
  - Many have been shut down

nature human behaviour

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature › nature human behaviour › comment › article

Comment | Published: 02 November 2023

Platform-controlled social media APIs threaten open science

Brittany I. Davidson ✉, Darja Wischerath, Daniel Racek, Douglas A. Parry, Emily Godwin, Joanne Hinds, Dirk van der Linden, Jonathan F. Roscoe, Laura Ayravainen & Alicia G. Cork

Nature Human Behaviour (2023) | Cite this article

Social media data collection tools

- YouTube: YouTube Data Tools
- Tumblr: TumblrTool
- TikTok, Instagram, LinkedIn, Imgur, Twitter: Zeeschuimer in combination with Zeehaven

- Or build your own scraper using Web Scraper

Group assignment

- Time to collect!
- Investigate which data sources you will need for your project. Before you start collecting, write down what each source will add.
- Collect at least three datasets that show different perspectives on your topic.