

# Van realiteit naar dataset

**Datum:** 5 april 2024

**Docent:** Jeroen Bakker  
([i.j.bakker@uu.nl](mailto:i.j.bakker@uu.nl))

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



# Programma

- **Oefening:** objecten > datamodel
- **Lezing:** van realiteit naar dataset
- Bespreken voorbereidende opdracht

----- **Pauze** -----

- **Lezing:** sociale media als bron
- **Opdracht:** dataverzameling en studie

# Oefening: objecten > datamodel

Maak een datamodel voor de objecten die je hebt gekregen, met kolommen en datatype (30 minuten).

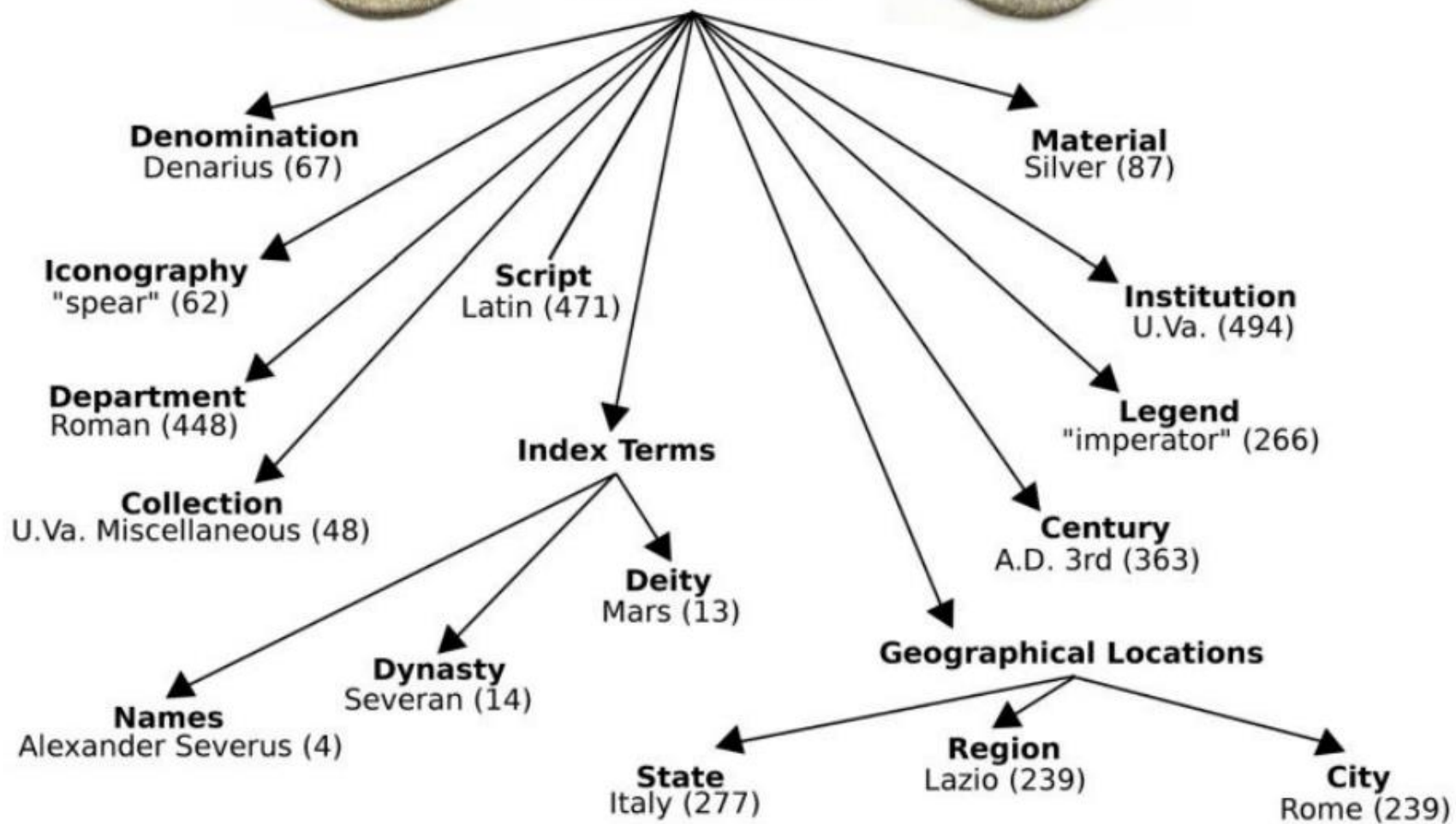
- auteur\_achternaamnaam (*tekst*)
- publicatie\_datum (*datum, dd-mm-jjjj*)

# Van objecten naar data

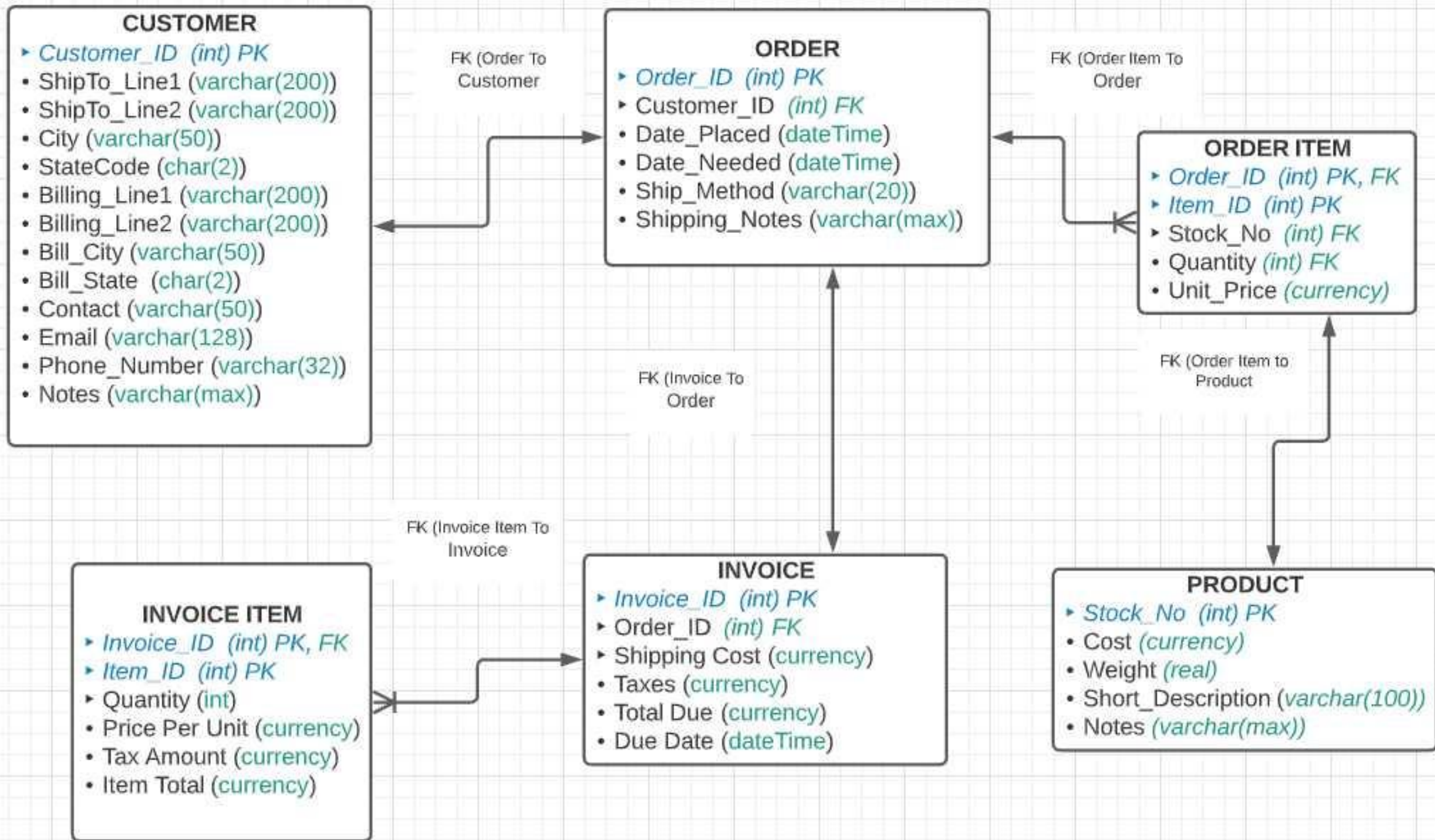
- Datasets zijn een logische weergave van gegevens
- Een datamodel geeft de structuur van je set aan: de categorieën, de onderlinge relaties, de regels
- Het datamodel wordt weergegeven middels een diagram



**Alexander Severus**  
A.D. 231-235







# **Dataverzameling: hoofdvragen**

Wat willen we bestuderen?

Welke data kunnen we hiervoor gebruiken?

Waar is deze data te vinden?

Hoe komen we eraan?

Hoe krijgen we de data in onze gewenste vorm?

# Data vinden

Er zijn veel verschillende manieren om bij de gegevens te komen die je nodig hebt. Bijvoorbeeld:

- Open data
- API-toegang
- (Betaalde) dataleveranciers
- Datalekken
- Geautomatiseerde web scraping
- Handmatige gegevensverzameling



# Open data

- Vrij beschikbaar met open licentie
- Beschikbaar gesteld door overheden, non-profitorganisaties, onderzoekers
- Doorzoekbaar via o.a. [Google Dataset Search](#)
- Er is meer beschikbaar dan je denkt: zoek eens op "[naam van de organisatie] open data"

## Climate Change Data

The indicators in this category examine carbon dioxide atmospheric concentrations, as well as trends in global warming, such as rising sea levels, rising temperatures and frequency of natural disasters which are key indicators to monitor climate change and its impacts on populations.

### Annual Surface Temperature Change

This indicator presents the mean surface temperature change during the 1961-2021, using temperatures between 1951 and 1980 as a baseline. Use the drop-down menus to search for temperature changes by country.

This data is provided by the Food and Agriculture Organization Corporate Statistical Database (FAOSTAT) and is based on publicly available GISTEMP data from the National Aeronautics and Space Administration Goddard Institute for Space Studies (NASA GISS).

# API-toegang

- Application Programming Interface: het bouwen van content met behulp van een live-verbinding met een database
- Rechtstreeks toegang tot gegevens uit de database
- Heeft soms toegangsvereisten, zoals aansluiting bij een academische instelling
- 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 »](#)

# Betaalde dataproviders

- Betalen voor toegang tot gegevens
- Gespecialiseerde gegevens, gearchiveerde gegevens, auteursrechtelijk beschermde gegevens
- Voorbeelden: LexisNexis, OBI4wan, veel sociale platforms

## Excel in digital customer engagement with our all-in-one solutions for:

✓ Webcare

✓ Messaging

✓ Live chat

✓ Chatbots

✓ Publishing

✓ Media monitoring

✓ Reputation management

✓ Data analytics

[Request a free demo](#)

[Read our success stories](#)





# Datalekken

- Geheime informatie gelect door klokkenluiders
- Vaak opgemaakt in een gemakkelijk toegankelijk formaat om onderzoek aan te moedigen
- [Offshore Leaks Database](#) (Panama Papers, Paradise Papers, Pandora Papers)
- [WikiLeaks](#)
- Niet zonder risico's



INTERNATIONAL CONSORTIUM  
OF INVESTIGATIVE JOURNALISTS

Answer our user survey to help shape the future of the Offshore Leaks Database.

TAKE OUT  
THE SURVEY >

THIS DATA  
**SHOULD BE  
PUBLIC**

We need your support to keep it that way.

DONATE

## OFFSHORE LEAKS DATABASE

Find out who's behind more than **810,000** offshore companies, foundations and trusts from the **Pandora Papers, Paradise Papers, Bahamas Leaks, Panama Papers** and **Offshore Leaks** investigations.

Search the full Offshore Leaks database

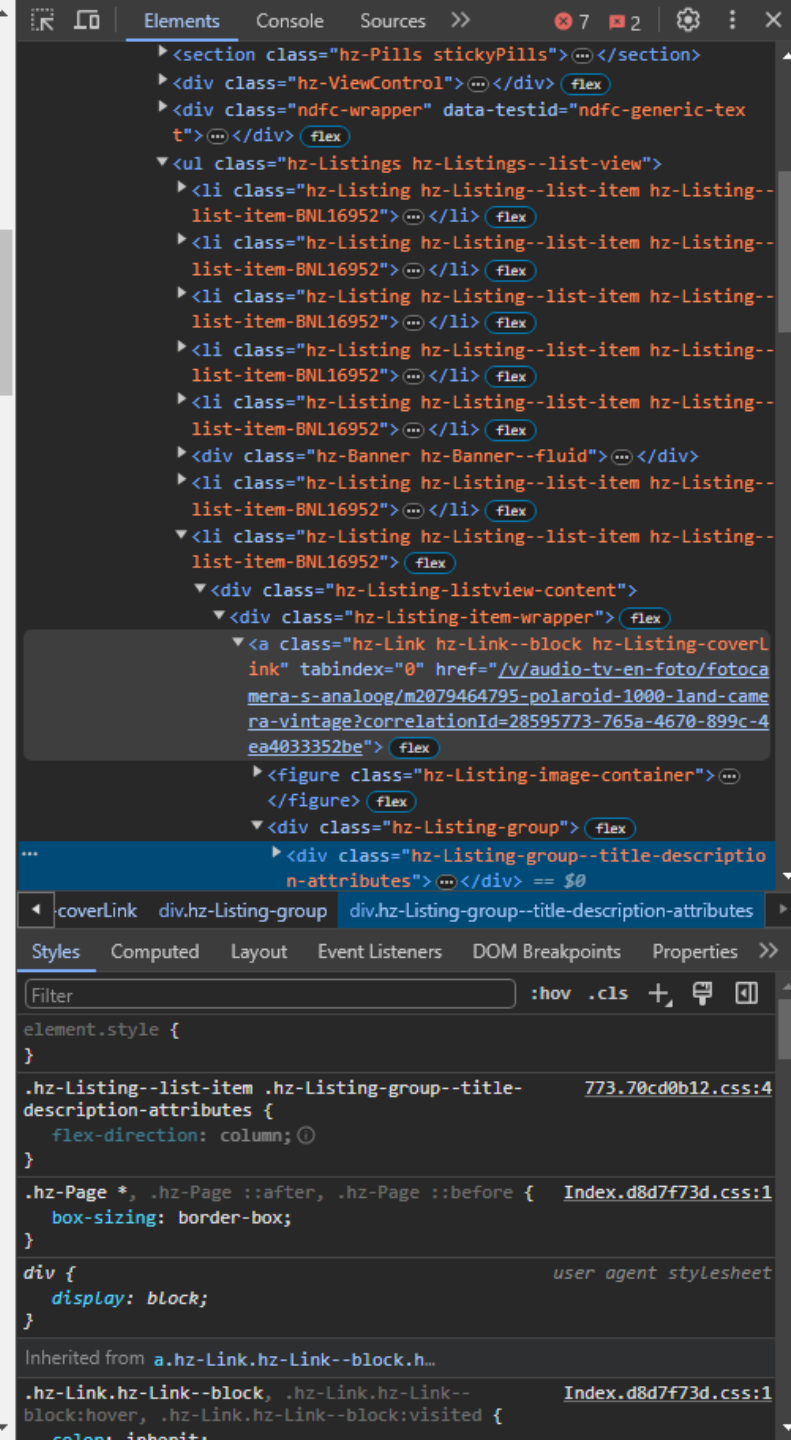
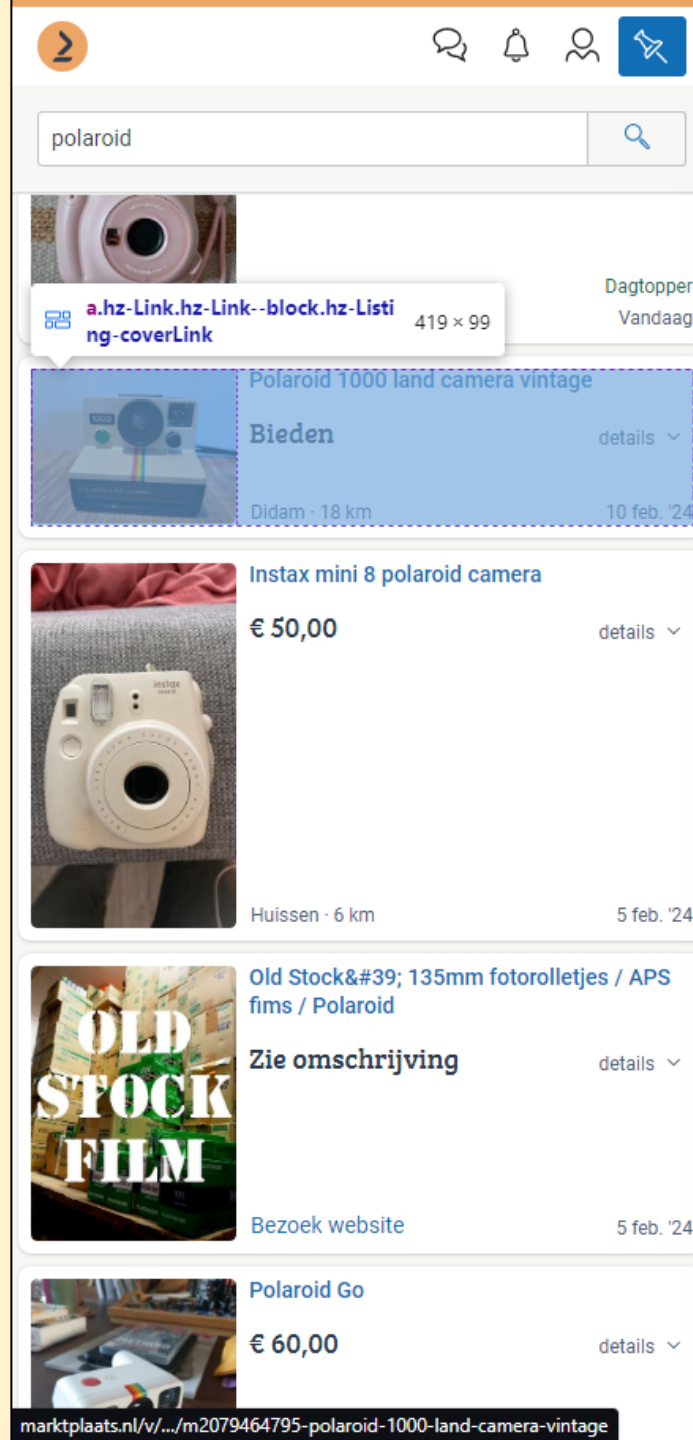
SEARCH

Explore the investigations

[Pandora Papers >](#) [Paradise Papers >](#) [Panama Papers >](#) [Bahamas Leaks >](#) [Offshore Le](#)

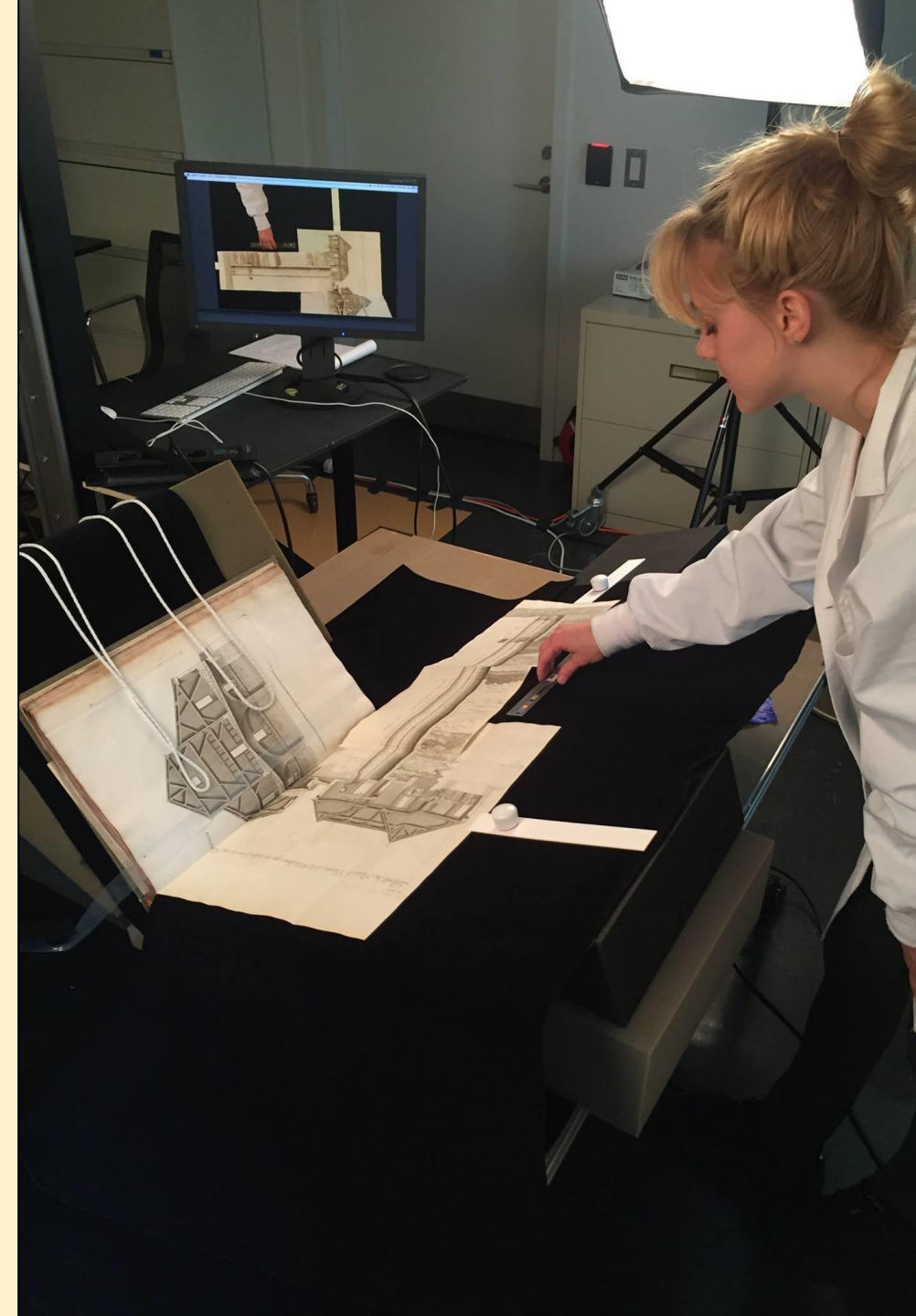
# Web scraping

- Automatisch verzamelen van webdata op basis van patronen in webpagina's
- Kan worden gebruikt om elk type gestructureerde pagina te verzamelen
- [Zeeschuimer](#) (voor sociale media)
- [Web Scraper](#) plugin



# Handmatige dataverzameling

- Old-school
- Soms nodig, bijvoorbeeld bij het werken met niet-digitaal archiefmateriaal
- Doel: een gestructureerde gegevenstabel maken op basis van ongestructureerde bronnen





# Oefening: welke data?

**Doel: onderzoeken hoe de Olympische Zomerspelen van 2024 in Parijs de lokale toeristenindustrie beïnvloeden (accommodatie, vervoer, restaurants enz.).**

*Welke gegevens zou je gebruiken? Bespreek in groepjes van 2 en bereid een pitch van 2 minuten voor, waarin je uitlegt:*

*... welke soorten data je zou gebruiken*

*... welke perspectieven op het onderwerp deze data kunnen bieden*

*... waar je denkt dat je deze data kunt krijgen*

**Tijd: 10 minuten**



# Discussie

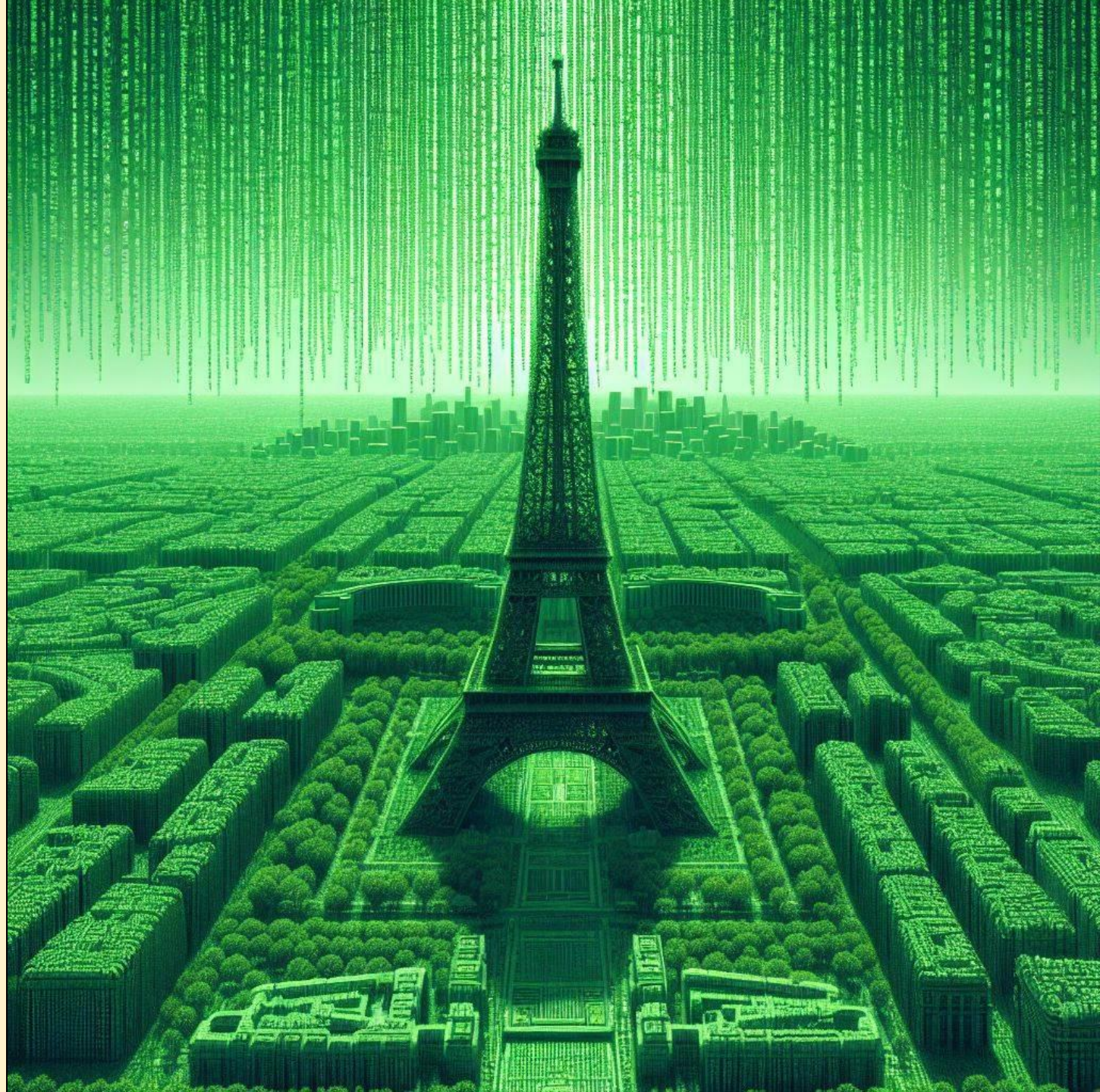
Leg uit...

... welke soorten gegevens je zou gebruiken

... welke perspectieven op het onderwerp deze gegevens kunnen bieden

... waar je denkt dat je deze gegevens kunt krijgen

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





# Data verrijken

- Meer informatie toevoegen aan je data
- Methoden:
  - Datasets combineren
  - Samenvoegen met andere datasets, vaak op basis van dezelfde kolom (d.w.z. naam, productcode)
  - Annotatie
  - Handmatig meer informatie toevoegen op basis van annotatieschema

# Kwaliteitscontrole

- Alle gegevens worden direct of indirect door mensen gecreëerd
  - Motieven
  - Fouten kunnen gebeuren!
- Gegevens zijn geen gegeven!
- Hoe zou je de kwaliteit van een dataset controleren?

# Checklist voor kwaliteitscontrole

- ✓ Wat is de bron?
- ✓ Wat zijn hun motieven om deze data te delen?
- ✓ Wat wordt er wel en niet weergegeven in de data?
- ✓ Welke methoden zijn gebruikt om deze data te verzamelen?

# Samenvattend

- Gegevens zijn geen gegeven: data zijn er nooit zomaar
- Data hebben altijd een vertaalslag gemaakt: realiteit > data
  - Gevolg: abstrahering, verlies van details
- Er zijn duizend manieren om aan data te komen, maar een kritische blik is altijd nodig
  - Doe de checklist!



# **Dataverzameling: hoofdvragen**

Wat willen we bestuderen?

Welke data kunnen we hiervoor gebruiken?

Waar is deze data te vinden?

Hoe komen we eraan?

Hoe krijgen we de data in onze gewenste vorm?

# Vorbereidende opdracht

Marktplaats

Help en info Voorwaarden Veiligheidscentrum

Berichten Meldingen J. Bakker

Plaats advertentie

polaroid camera

Fotocamera's Analoo

6813KZ


Alle afstanden...

Zoek

Audio, Tv en Foto

Fotocamera's Analoo

Verwijder filters




Collectors item! Een schitterende mamiya rb67 met veel toebehoren. Alles in zeldzaam uitstekende...

Zo goed als nieuw · Ophalen

Vandaag Rotterdam 99 km

Dagtopper



Polaroid Land Zip Camera.  
Polaroid land zip camera.

Gebruikt · Ophalen

Bieden Jeronimerel  
1 apr. '24 Westervoort 8 km

Sitemaps Sitemap Polaroid\_MP Create new sitemap

7 2

\_root / page\_next / advertentie [Data preview](#)

ID	Selector	type	Multiple	Parent selectors	Actions
titel	h3	SelectorText	no	advertentie	<a href="#">Element preview</a> <a href="#">Data preview</a> <a href="#">Edit</a> <a href="#">Delete</a>
conditie	span.hz-Attribute--default.nth-of-type(1)	SelectorText	no	advertentie	<a href="#">Element preview</a> <a href="#">Data preview</a> <a href="#">Edit</a> <a href="#">Delete</a>
prijs	span.hz-Listing-price	SelectorText	no	advertentie	<a href="#">Element preview</a> <a href="#">Data preview</a> <a href="#">Edit</a> <a href="#">Delete</a>
aangeboden_sinds	span.hz-Listing-date--desktop	SelectorText	no	advertentie	<a href="#">Element preview</a> <a href="#">Data preview</a> <a href="#">Edit</a> <a href="#">Delete</a>
naam_verkoper	span.hz-Listing-seller-name	SelectorText	no	advertentie	<a href="#">Element preview</a> <a href="#">Data preview</a> <a href="#">Edit</a> <a href="#">Delete</a>
locatie_verkoper	span.hz-Listing-distance-label	SelectorText	no	advertentie	<a href="#">Element preview</a> <a href="#">Data preview</a> <a href="#">Edit</a> <a href="#">Delete</a>

# Pauze

***Prompt:** A group of students enjoying a cup of coffee in the sun, goofy claymation style*



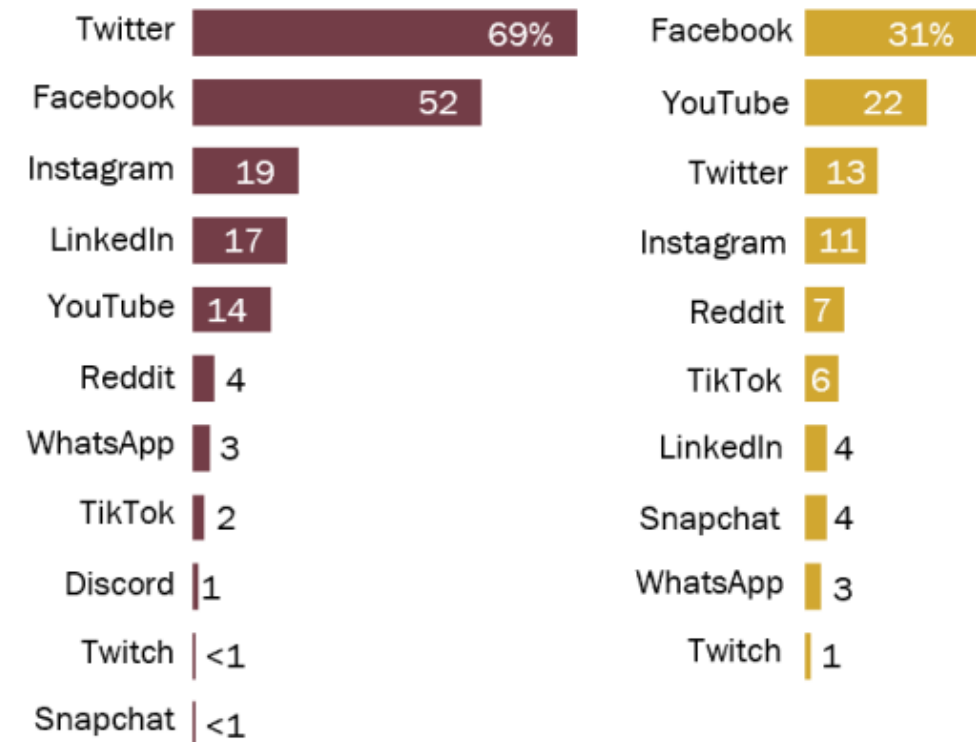
# Sociale media als een bron

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/](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/)

## 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

% of *U.S. journalists* who say —  
is the social media site they use  
most or second most in their job

% of *U.S. adults* who say they  
regularly get news on ...



Note: Discord was not asked about in the survey of U.S. adults.

Source: Survey of U.S. journalists conducted Feb. 16-March 17, 2022. Survey of U.S. adults conducted July 26-Aug. 8, 2021.

PEW RESEARCH CENTER



# Voors en tegens

- Voordelen
  - Extreem snelle manier om nieuwsgebeurtenissen te volgen
    - *Citizen reporting*
  - Ongemedieerde communicatie
  - (Bijna) iedereen heeft een stem
- Nadelen
  - De rest van deze lezing



Geert Wilders  
@geertwilderspvv

Follow

Never trust the press.



RETWEETS

2,635

LIKES

3,537

12:16 PM - 10 Dec 2016

198

2.6K

3.5K

# Ongemedieerde communicatie?

- Ja, iedereen kan posten en delen...
- ... Maar alle platformen worden gedreven door algoritmes
  - Geautomatiseerde mediatie
  - Zichtbaarheid wordt bepaald door een ondoorzichtig systeem
- Het platform *is* een medium
  - Eigen gebruiken, mogelijkheden en limitaties



# Filterbubbles

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



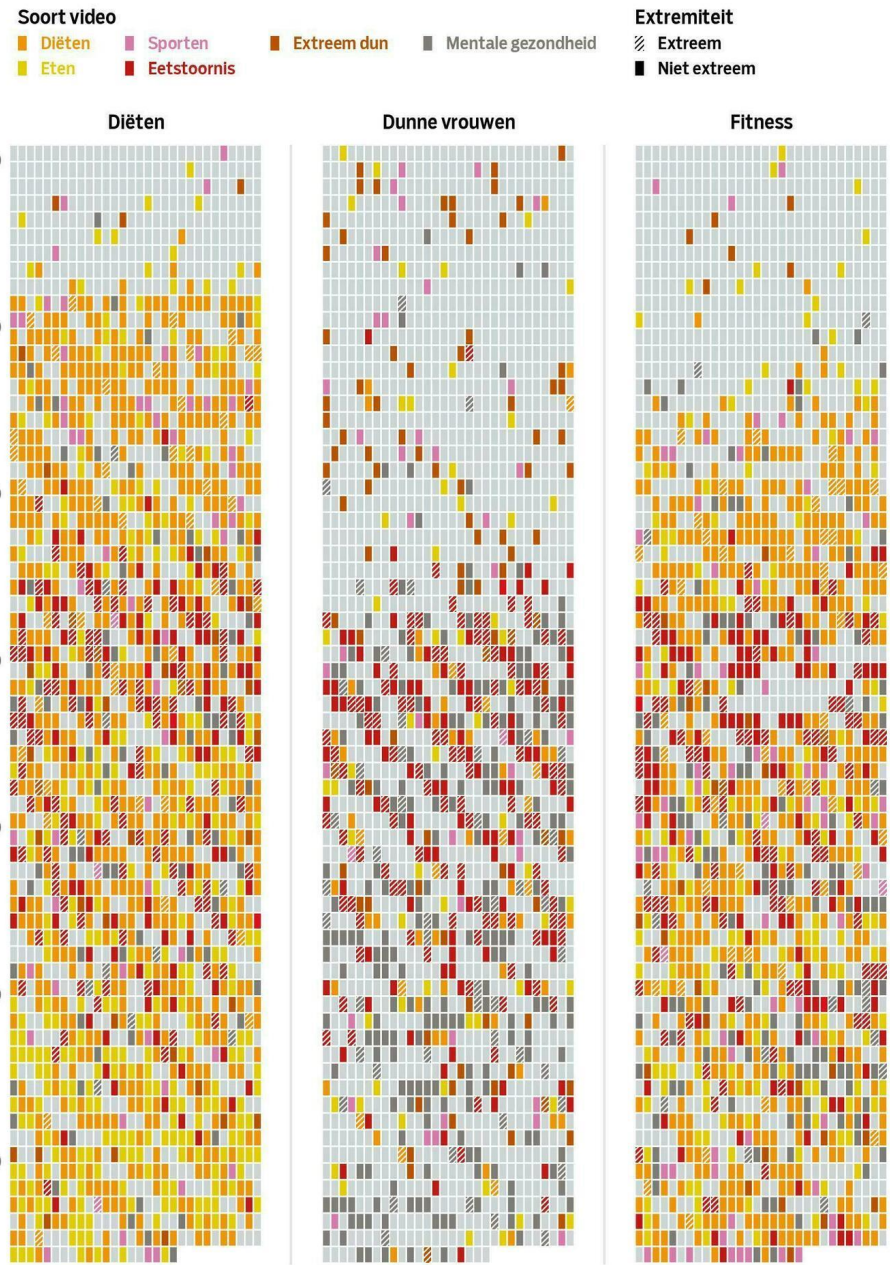
# Filterbubbels

- Websites richten content op gebruikers
  - *Recommendation*-algoritmes
  - Gepersonaliseerde feeds
- Filterbubbels (Pariser)
  - Informatie is gepersonaliseerd geworden
  - Gebruikers worden ingesloten door een gefilterd ecosysteem
  - Effect: mensen worden bevestigd in hun eigen vooroordelen



# 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

## Soort video

Diëten

Sporten

Extreem dun

Mentale gezondheid

Eten

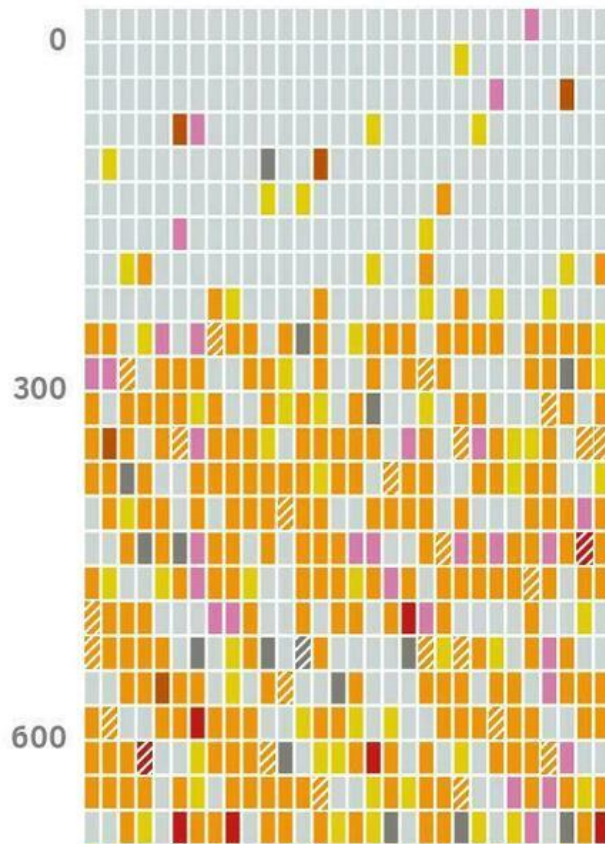
Eetstoornis

## Extremiteit

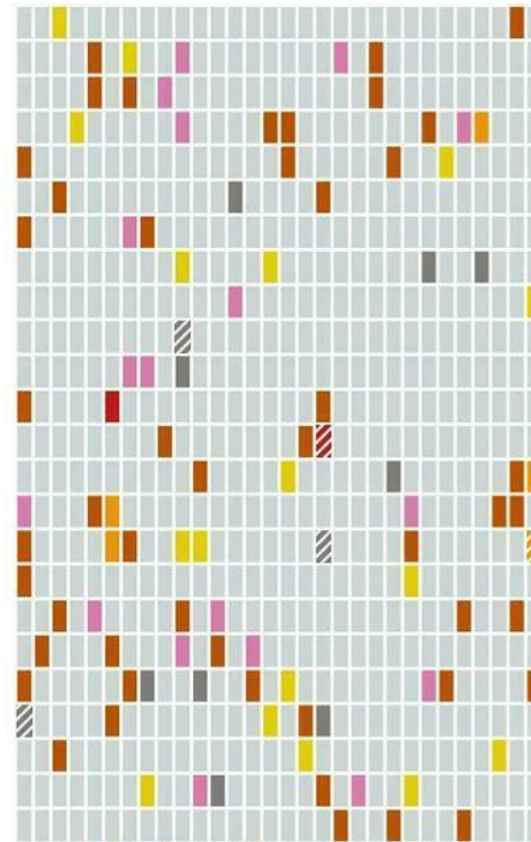
Extreem

Niet extreem

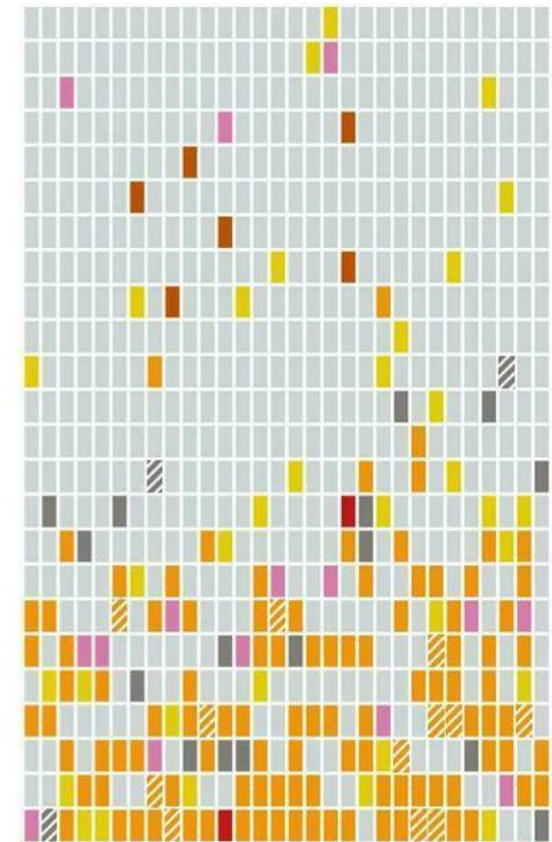
Diëten



Dunne vrouwen



Fitness



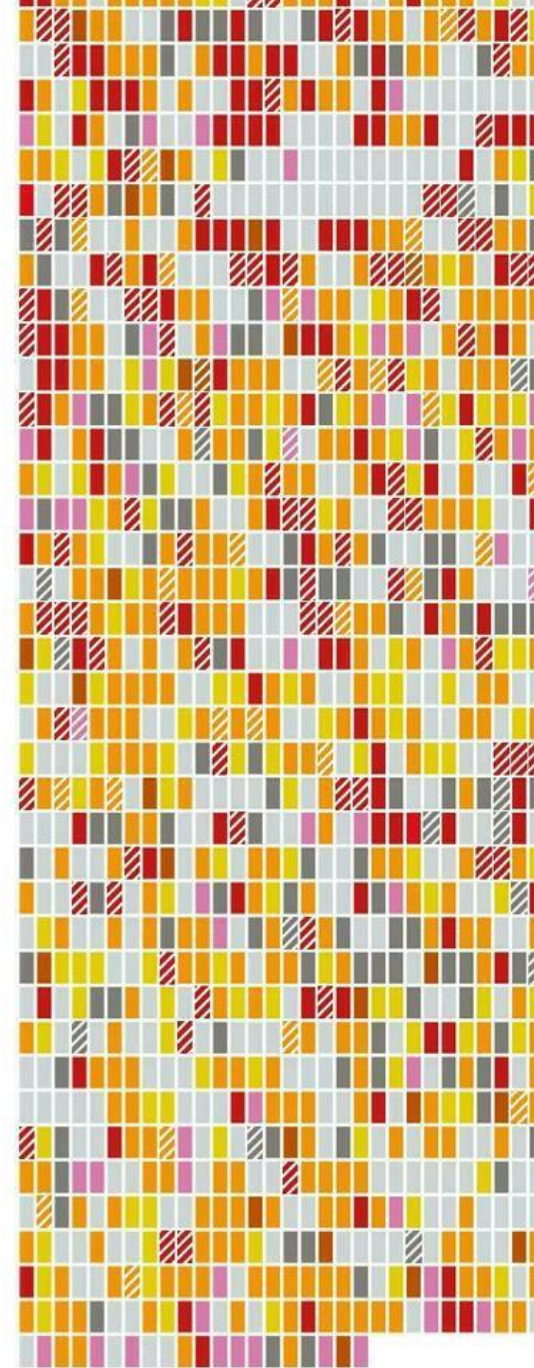
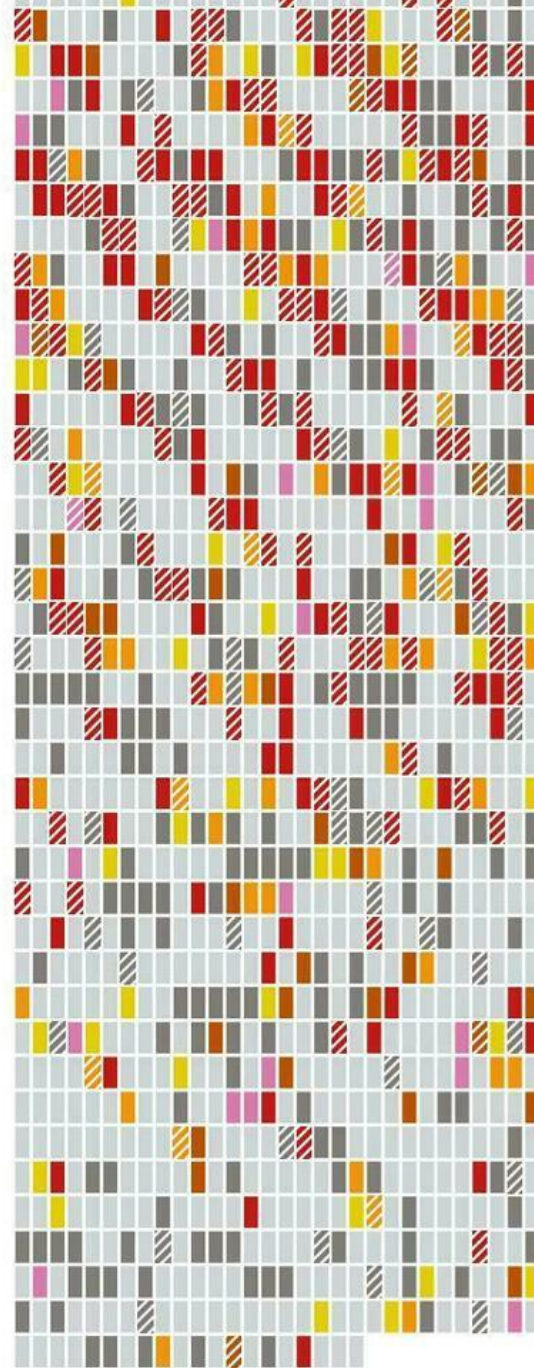
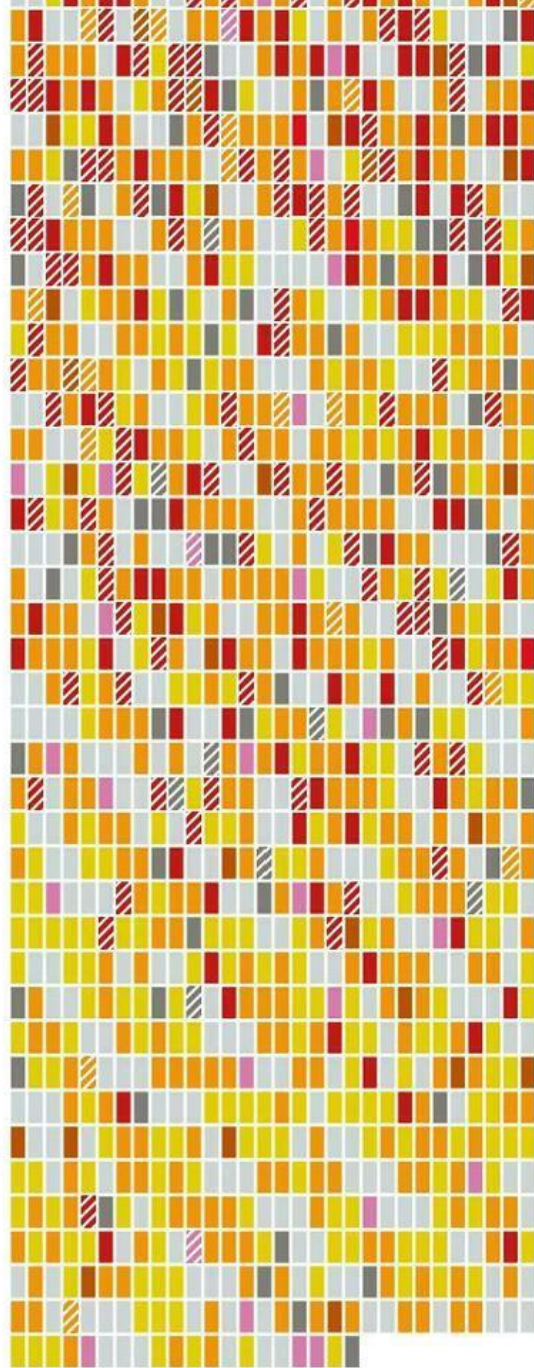


900

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

## Reacties

 **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.

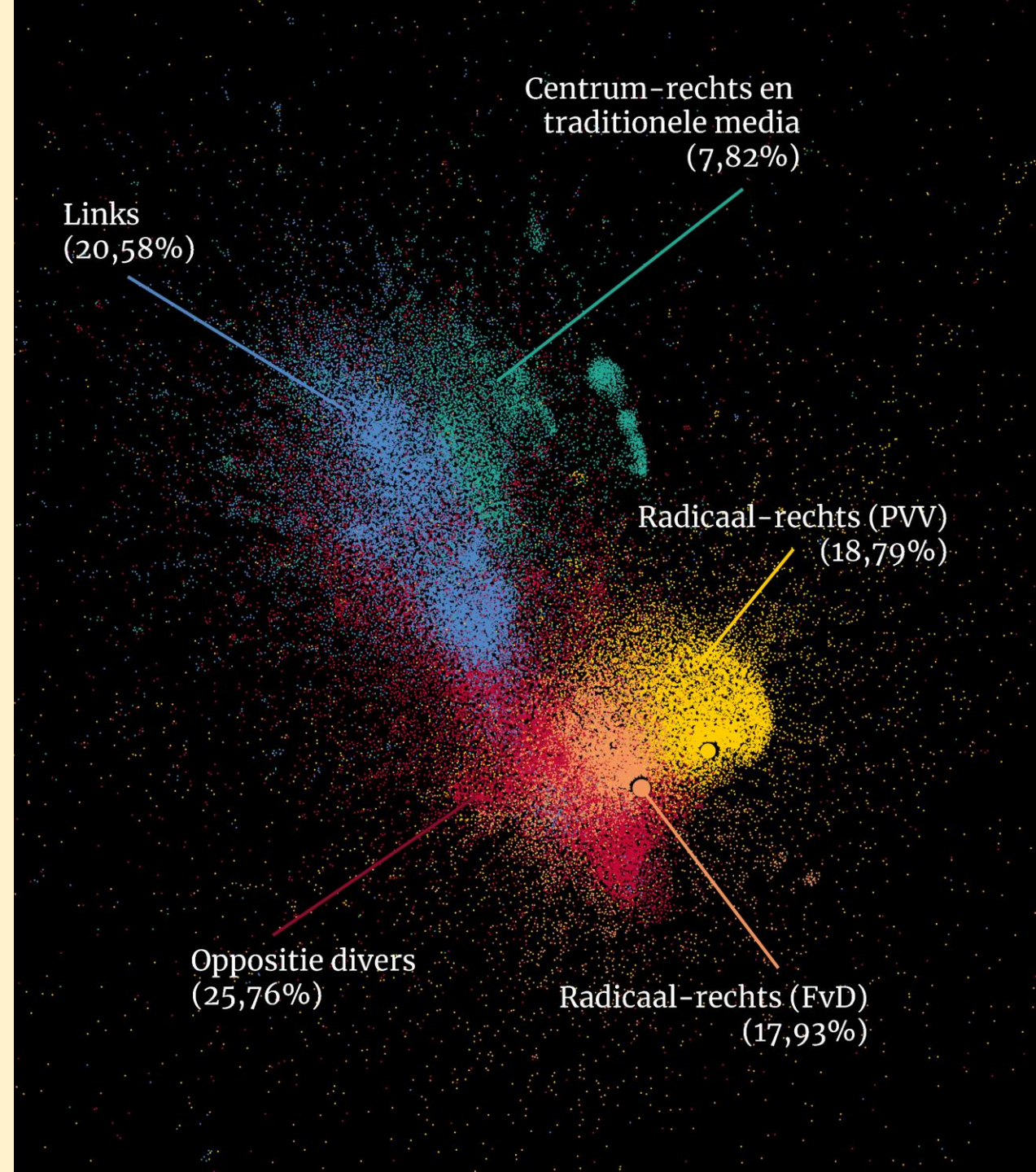
Stop prioritizing foreign countries over the welfare of Americans citizens.

Put America First.

8 7 98 1.6K

# Onderzoekers zijn niet immuun

- We worden allemaal beïnvloed door algoritmen
- De feed van een onderzoeker is net zo bevooroordeeld als elke andere
- Het publiek van sociale media is geen afspiegeling van de echte wereld
- Doelgroepen verschillen



# Alternatief: API's

- Directe toegang tot gegevens door middel van zoeken op trefwoorden
- Zoekresultaten worden niet beïnvloed door algoritme
- Echter...
  - Technische hindernissen: vereist vaak programmeervaardigheden
  - Velen zijn stilgelegd

**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](#)

# Tools voor dataverzameling

- YouTube: [YouTube Data Tools](#)
- Tumblr: [TumblrTool](#)
- TikTok, Instagram, LinkedIn, Imgur, Twitter: [Zeeschuimer](#) (wat technische kennis vereist)
- Of bouw je eigen scraper met [Web Scraper](#)



# Opdracht:



1. Kies met je team welke bronnen jullie willen verzamelen en onderzoek hoe deze data te bereiken zijn. Verzamel vervolgens per persoon een dataset. Minstens één persoon moet hiervoor gebruik maken van Web Scraper. Wees voorbereid om je proces en keuzes uit te leggen.
2. Onderzoek jullie data middels de tidyverse in R. Bereid een presentatie voor waarbij jullie de hoofdinzichten uit jullie studie presenteren. Focus hierbij op de verhoudingen tussen de datasets: wat kunnen we leren door datasets te combineren?



# Wensen voor volgende sessie

- Voorstel van Chiel: SQL
  - Programmeertaal om met relationele databases te werken

