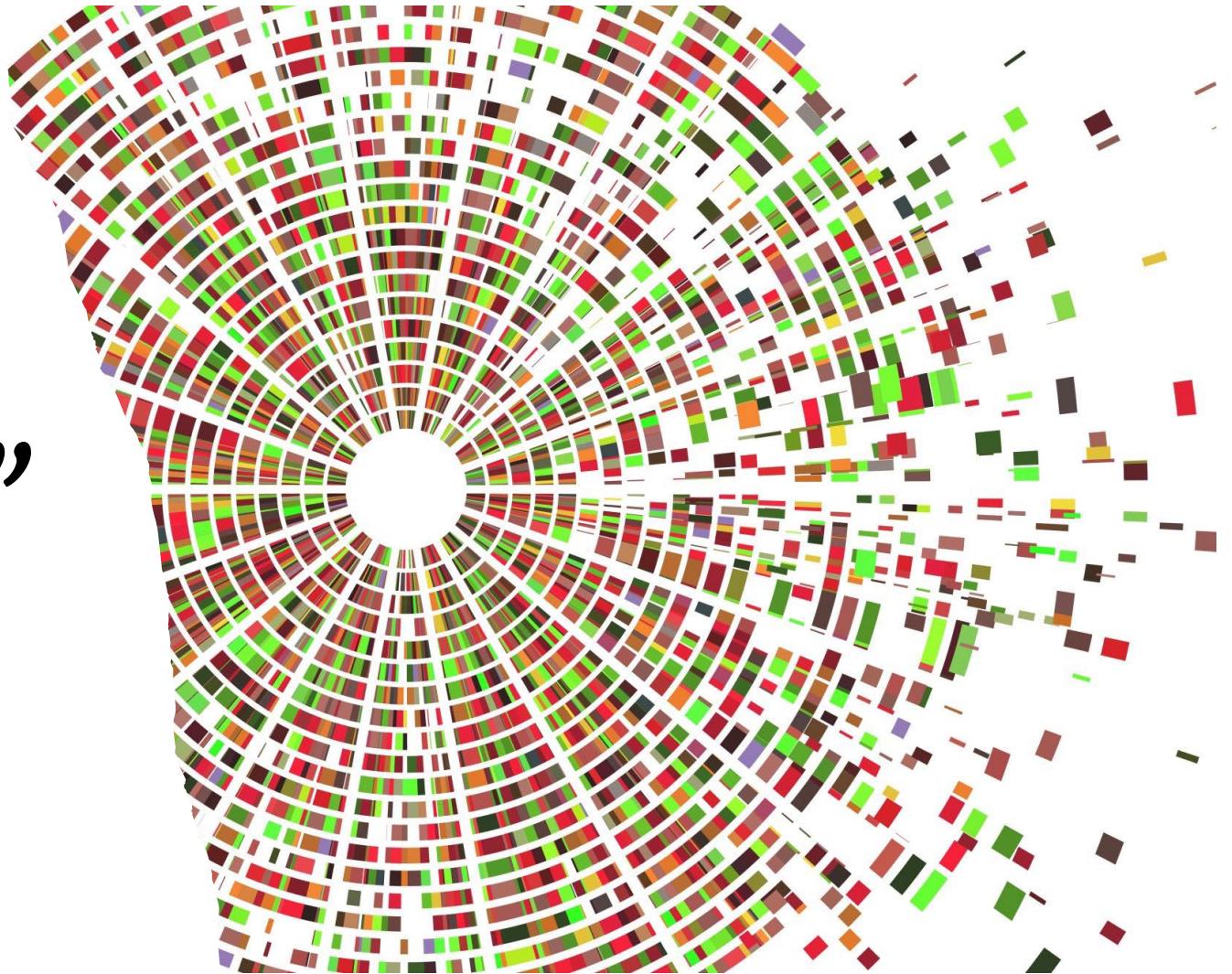


“Thinking in Data”



Dr. Dennis Nguyen

Assistant Professor Digital Literacy & Digital Methods

A complex, abstract network graph composed of numerous small, semi-transparent black dots connected by thin gray lines. The graph forms various small clusters and paths across the left side of the slide.

Agenda

- What are data?
- Digital Humanities & the Computational Turn
- Use Cases
- Discussion

What are data?

Dr. Dennis Nguyen

Assistant Professor Digital Literacy & Digital Methods

“Data are commonly understood to be **the raw material** produced by **abstracting the world into categories, measures and other representational forms** – numbers, characters, symbols, images, sounds, electromagnetic waves, bits – that **constitute the building blocks** from which **information** and **knowledge** are created”

(Kitchin 2022).

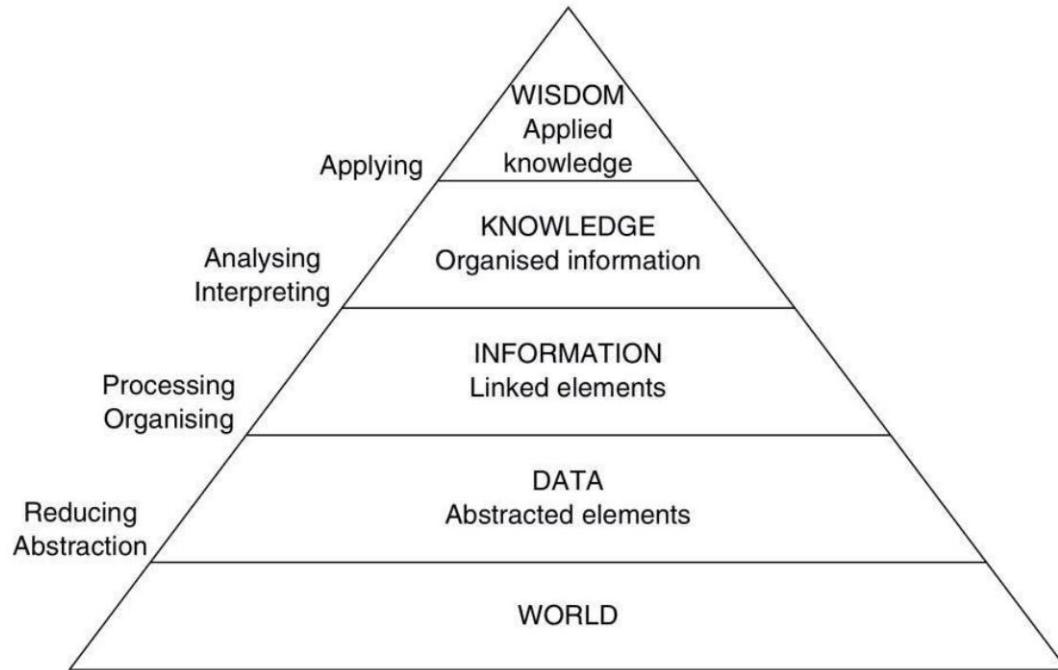


Figure 1.1 Knowledge pyramid (adapted from Adler 1986 and McCandless 2010)

Qualitative vs. Quantitative data

QUANTITATIVE

	A	B	C	D	E	F	G
1	Country	Region	Expenditures (US Dollars)	Total Libraries	Total Librarians	Total Volumes	Total Users
2	Afghanistan	Asia		150		577300	48133
3	Albania	Europe	\$134.036,00	2107	32	12398795	581298
4	Algeria	Africa	\$641.176,00	760	56	4363718	1430
5	American Samoa	Oceania		32		170009	
6	Andorra	Europe	\$5.777.651,00	39	6	222869	16596
7	Angola	Africa		62		392870	
8	Anguilla	Latin America		1		14000	
9	Antigua and Barbuda	Latin America			8	10000	
10	Argentina	Latin America		2537	658	28770630	787788
11	Armenia	Asia	\$133.148,00	2555	1430	30935507	13615227
12	Aruba	Latin America		34		270000	
13	Australia	Oceania	\$1.927.384.739,00	11052	9021	1,28E+08	14477916
14	Austria	Europe	\$182.404.250,00	2695	3486	71909504	1559109
15	Azerbaijan	Asia	\$1.490.047,00	4432	267	59502341	180131
16	Bahamas	Latin America		139	4	201000	3733

QUALITATIVE

474	JS: How would you describe your first reaction to the idea?
475	PR17: Well, hmm [pause], I'm going to have to think back here.
476	[pause] Ahh, I guess I was really worried about how much work it
477	would mean for everybody. The whole program meant doing things
478	in a new way – kind of starting over on some real basics.
479	JS: What's an example here? [Knock on the door; someone enters
480	and leaves] Uhh – I was asking about an example.
481	PR17: Well, the curriculum [unclear], for one. All at once, we were
482	going to have to write out these justifications using new criteria.

Data comes in different forms

- **Structured:** easily organized, stored and transferred
- **Semi-Structured:** loosely structure, no pre-defined model/schema
- **Unstructured:** no data model or identifiable order/structuring

DATE	USER	POSTS, N = 10000
2021-12-05	booly_tier1	5 Japanese Stocks Into Condominiums https://t.co/aNDK44RQiD #Japan #Property #Stocks #Investment #Condo #RealEstate #Housing #Construction #LuxuryEstates #Resorts #Dividends #Trusts #Income #IoT #Stockmarket #Trading #BigData
2021-12-05	Lala35539674	buy fud crypter https://t.co/qzJlaxUbRJ #DataAnalytics #MachineLearning #BigData #DataScience #PyTorch #Python #TensorFlow #JavaScript #ReactJS #Flutter #CloudComputing #DataScientist #WomenWhoCode #Programming #Coding #100DaysofCode #DEVCommunity #opensource #AI #AR
2021-12-05	BrainRoaring	How to Become a #DataScientist #ML #DataScience #BigData #Analytics #AI #IoT #PyTorch #Python #RStats #TensorFlow #JavaScript #ReactJS #CloudComputing #Serverless #DataScientist #Linux #Programming #Coding #100DaysofCode #Flutter #Jupyter #DL #SQL #HTML #CyberSecurity #NLProc #R https://t.co/eadpmqZtFb

With computational methods we can turn unstructured to structured data (e.g., texts, images)!

Levels of Measurement

Nominal

- No intrinsic order
- Hair color – blond/brown/red

Ordinal

- The values are categories with intrinsic order
- ‘Distance’ between the categories is unknown
- Economic status – low/ medium /high

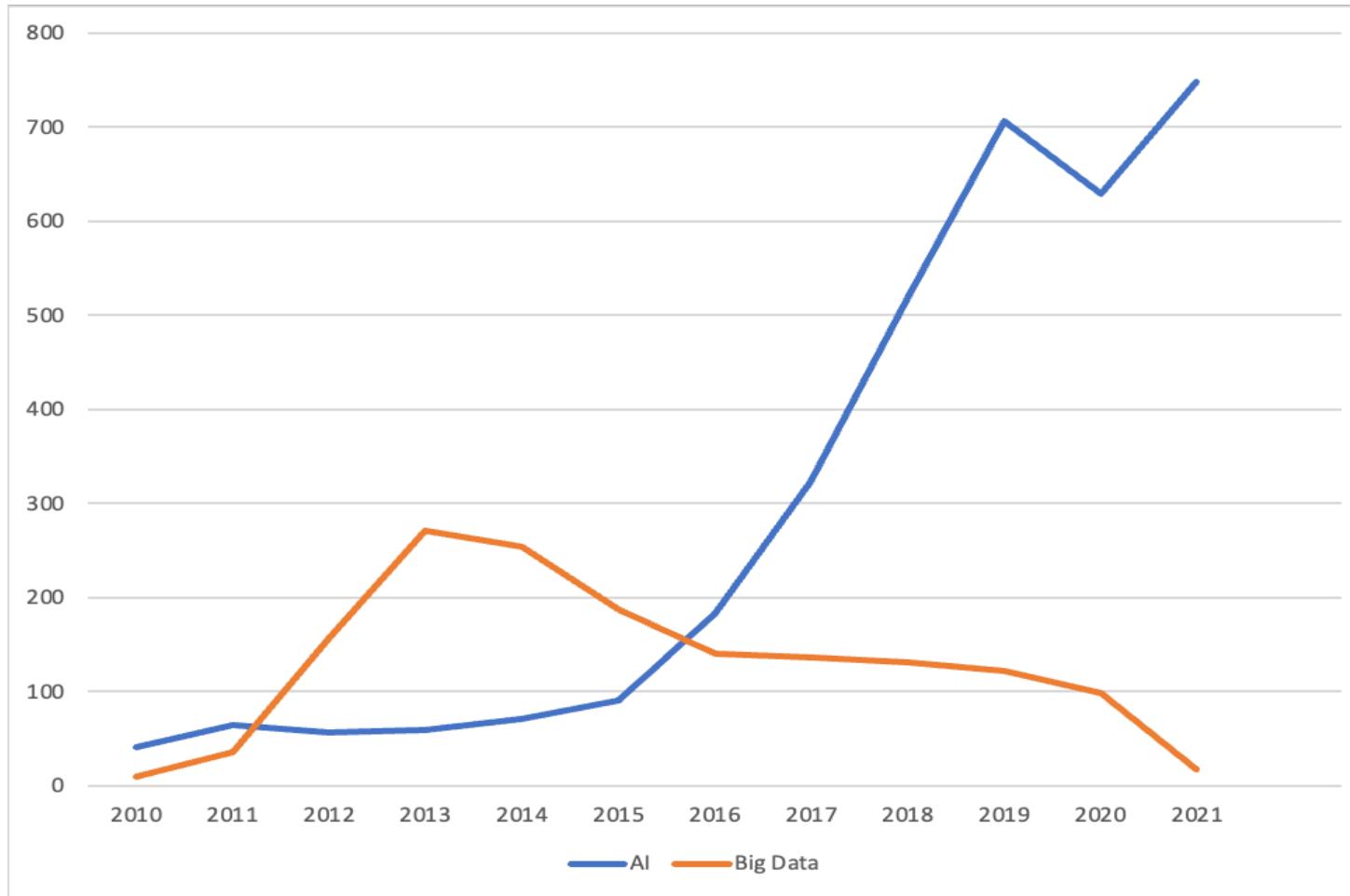
Scale

- The values represent some measured quantities
- ‘Distance’ is known
- Arithmetic operations are possible
- Height – 192cm, 175cm, 169cm, 187cm, 185cm

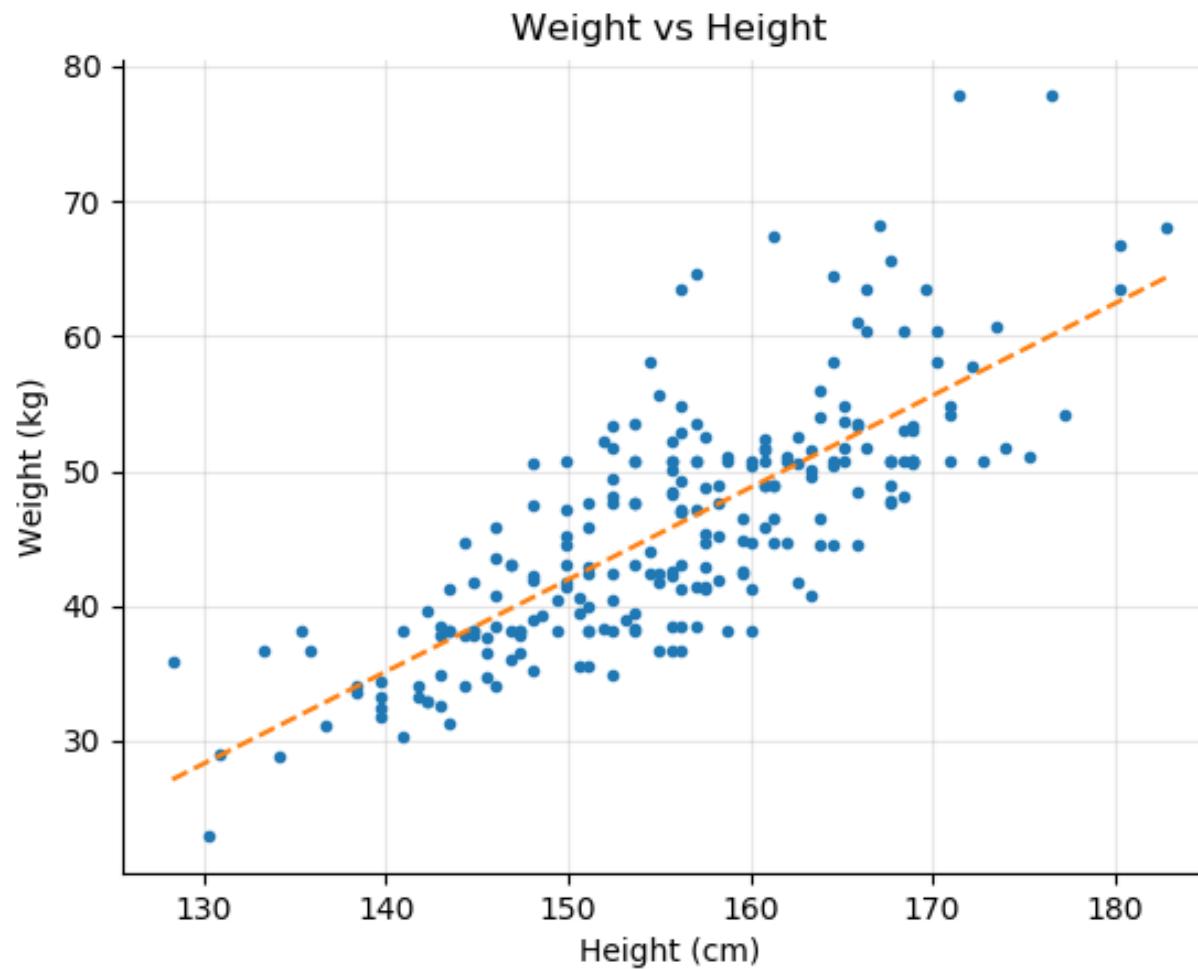
It matters how we count stuff!

The levels of measurement determine what analyses you can do with your data!

Example descriptive statistics: Item Frequencies over Time



Example inferential statistics: relationship height and weight in a study population



As researchers and practitioners, we need to critically engage with data:

- **Technically:** representativeness; uncertainty; reliability; error; bias; meta-data;
- **Ethically:** justice; equality; fairness; honesty; respect; rights; entitlements; care;
- **Politically & Economically:** norms; usage; regulation; legislation; funding; decision-making;
- **Temporally & Spatially:** histories; geographies/locations; transformations;
- **Philosophically:** ontology; epistemology; objectivity; subjectivity; normativity;



"The point is, **data are never simply just data**; how data are conceived and used varies between those who capture, analyse and draw conclusions from them"

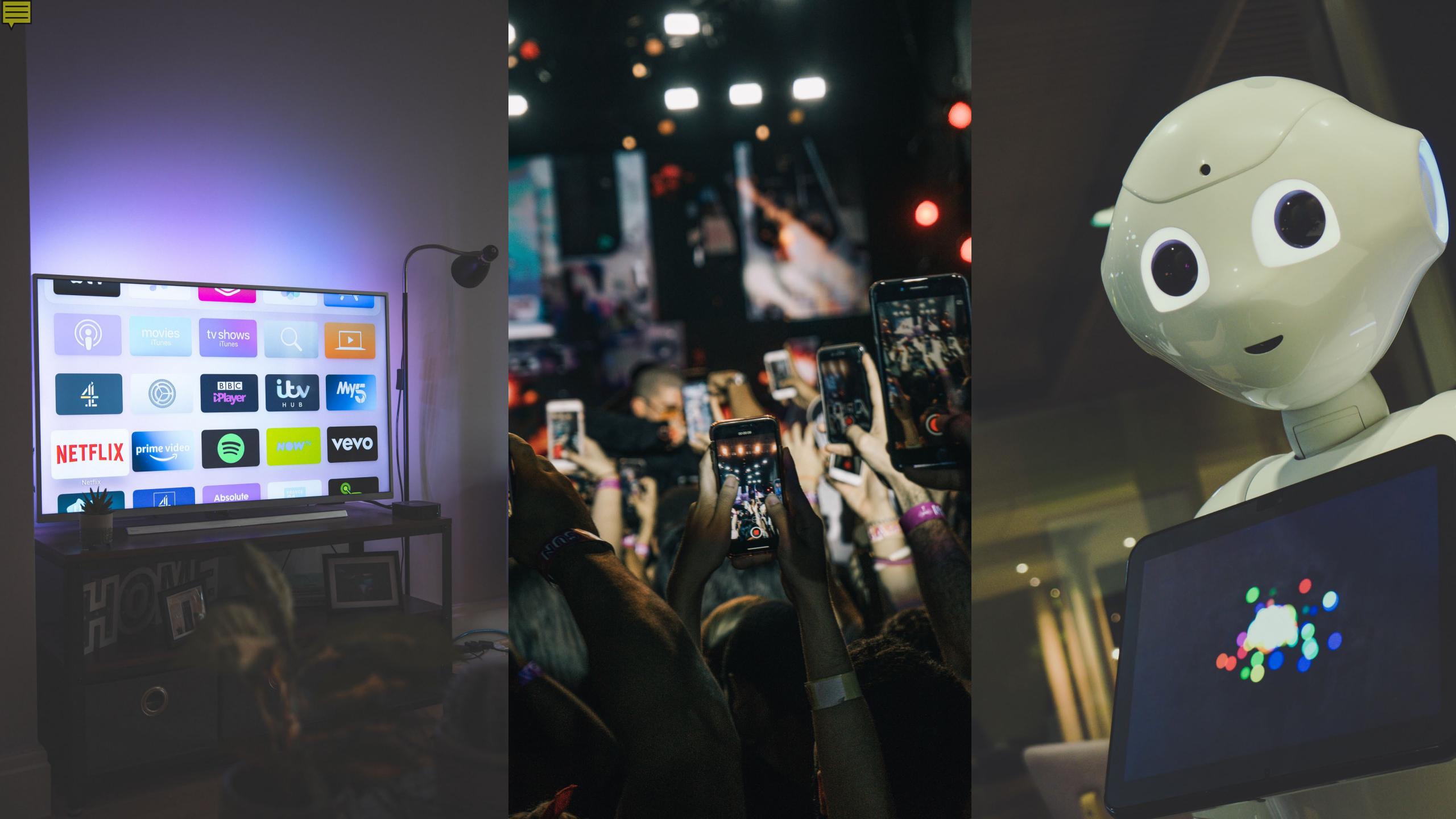
(Kitchin 2022)

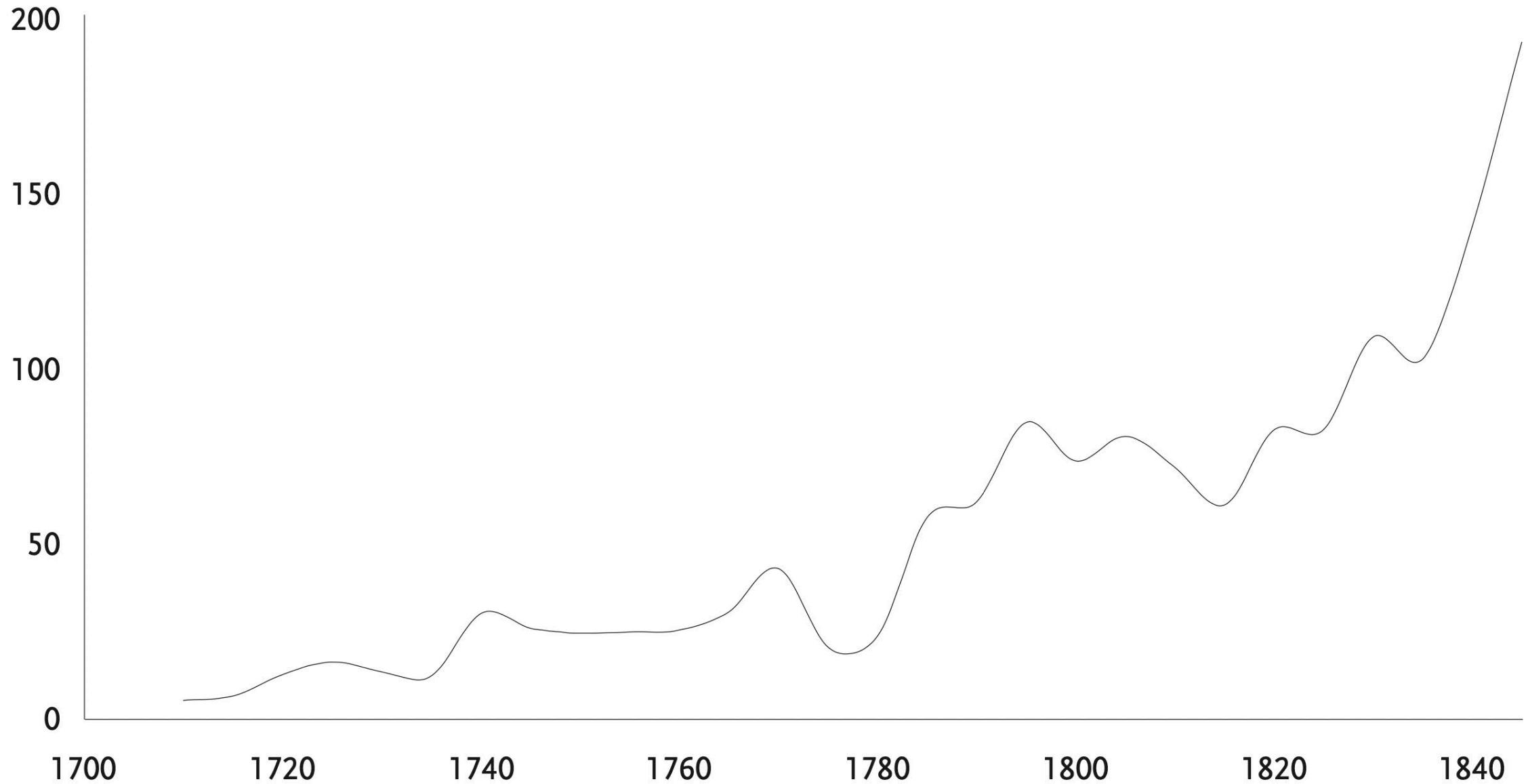
Digital Humanities & the Computational Turn

Dr. Dennis Nguyen

Assistant Professor Digital Literacy & Digital Methods







New novels per year, by 5-year average. Sources: McBurney, *Check List of English Prose Fiction, 1700–39*;

Humanities research...

- always dealt with the “digital” (esp. media studies)
- experimented with quantitative approaches to studying culture (see Moretti, 2005)

*However, with datafication we see a “**paradigm shift**”*





Epistemology

'[...]how questions are asked and how
they are answered[...]'

(Kitchin, 2021: 113)

What is a “paradigm”?

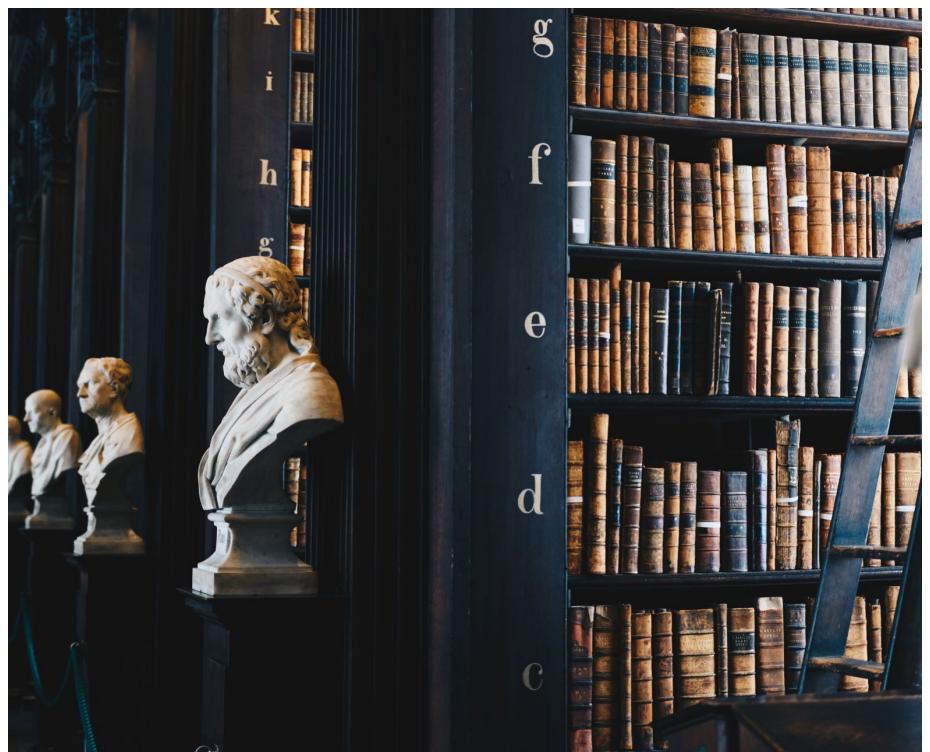
[...]a paradigm constitutes an accepted way of **interrogating the world** and **synthesising knowledge** common to a substantial proportion of researchers in a common discipline[...]

(Kitchin, 2021:113, paraphrasing Kuhn, 1962, my emphasis)

What is a “paradigm”?

Paradigms are neither static nor permanent.
They change over time with new insights, new arguments, and novel methods for observing the world.

However, they become at least **temporarily hegemonic**.





What is a “paradigm”?

Social Sciences & Humanities

- Positivism
- Phenomenology
- Structuralism
- Poststructuralism
- ...

(Natural) Sciences

- Scientific method
- Hypothesis testing
- Verification/falsification

There is an ongoing debate over how data and computational methods change the “nature” of humanities research.

For some a paradigm shift has already taken place, while others are still sceptical about “digital humanities” as a new methodological proposition.



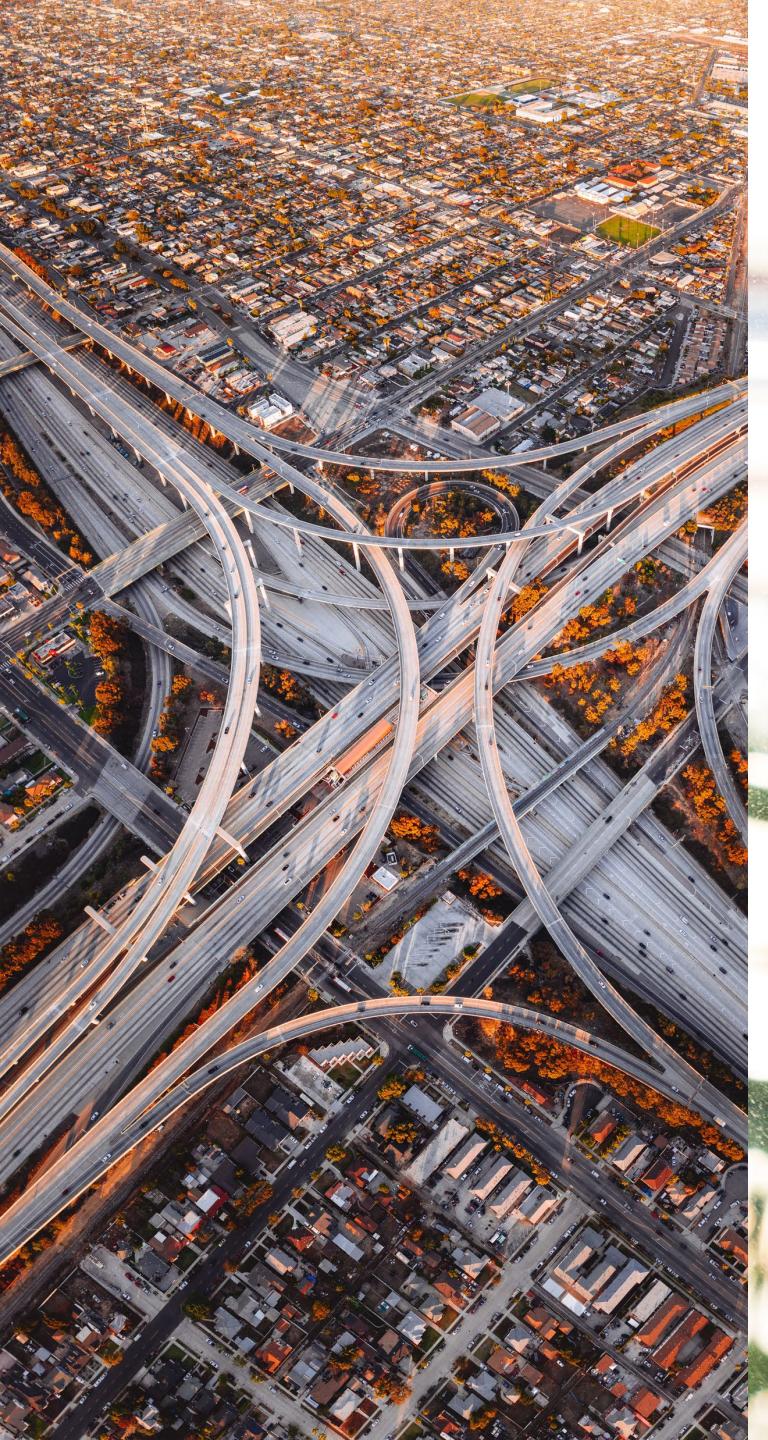
Big Data = A New Empiricism?

```
0100011101  
11010110000101110011000100  
00100111001100001010010100  
0011100000  
10010010011100011001000101  
10010110111000111001000101  
010100001100010010000101  
011101001110011001000101  
000100  
0111010011100011001000101  
10000110010110100101  
0011  
01010100011001000101  
001001010110100010000101  
1111011001110011001000101  
010101100010000101000101  
1111011001110011001000101  
010101100010000101000101
```

A New Era of “Empiricism”?

Simply put, the main idea of empiricism (from Greek *empeiria* for ‘experience’) is that observations of the world through data drive the creation of knowledge.

It predates big data but has gained in popularity with the rise of datafication (esp. in the private sector).



A New Era of “Empiricism”?

Claim: data '**can reveal their inherent truth**' and '**speak for themselves**' (Kitchin, 2021: 115). Correlation beats causation. Induction trumps deduction. No need to ask questions, theory is dead.

The Fallacies of Big Data Empiricism



- Big data are never “complete” and always sampled
- Big data do not materialise in a vacuum
- Big data are not free from human bias
- Big data are meaningless without context and domain knowledge

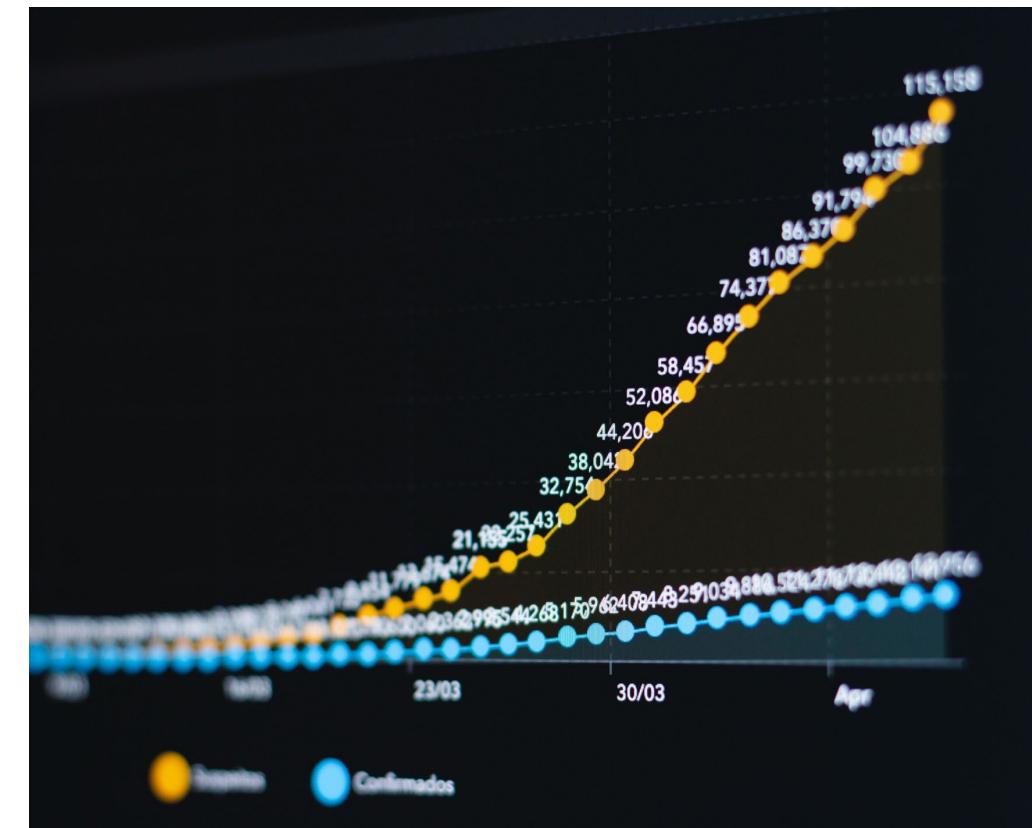
*'Indeed, all data provide **oligoptic views** on the world: views from certain vantage points, using particular tools, rather than an all-seeing, infallible god's -eye view'*

(Kitchin, 2021: 117, paraphrasing Haraway, 1991, my emphasis)

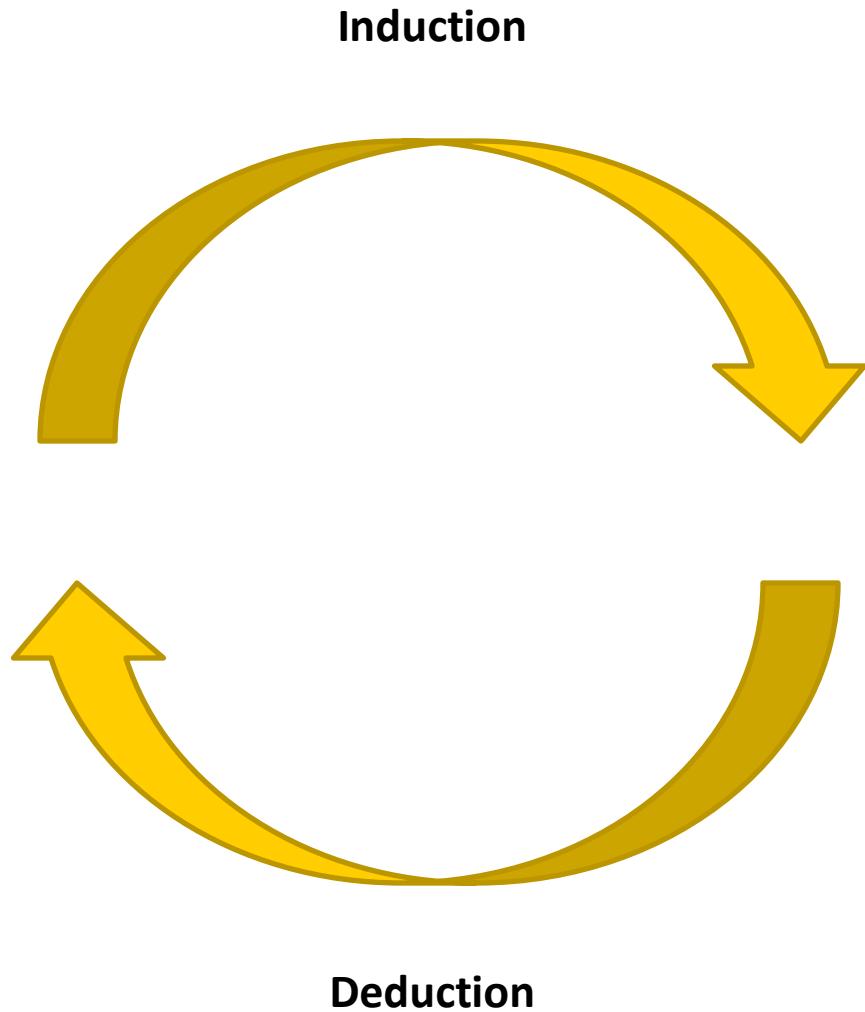


Data-Driven Science

- Avoids the fallacies of (naive) empiricism
- Maintains tenets of scientific method but reconfigures them
- Hybrid of abductive, inductive, and deductive
- Generates hypotheses and insights from data/incorporates induction in research designs



Data-Driven Science



- Links induction and deduction
- Anchored in domain knowledge
- *Guided knowledge* discovery & careful planning of data collection/analysis

Abductive Reasoning (proposed by C. S. Peirce)

- Making logical inferences and reasoning without making definite claims.
- Puts induction into context and connects it to deductive reasoning. Neither is isolated but embedded within one framework for generating knowledge.

Data-Driven Science

Data-Driven Sciences is proposed as a new paradigm that overcomes the limitations of *knowledge-driven science* while avoiding the fallacies of *big data empiricism*.

It aims to make use of data and novel analytical methods to explore, extract value from, and make sense of massive datasets.

The goal is to develop **holistic approaches** for doing science (Kitchin, 2021: 120).



Digital Humanities

- Focus on disciplines of the humanities
- History, art history, literature, culture studies, media studies etc.
- Understanding (digital) culture through digital data
- Introducing quantitative methods to humanities (not for the first time)



BANGKOK



BERLIN



59.4%



38.5%



23.7

26.3

EST. AGE ————— 20 ————— 30 —————>

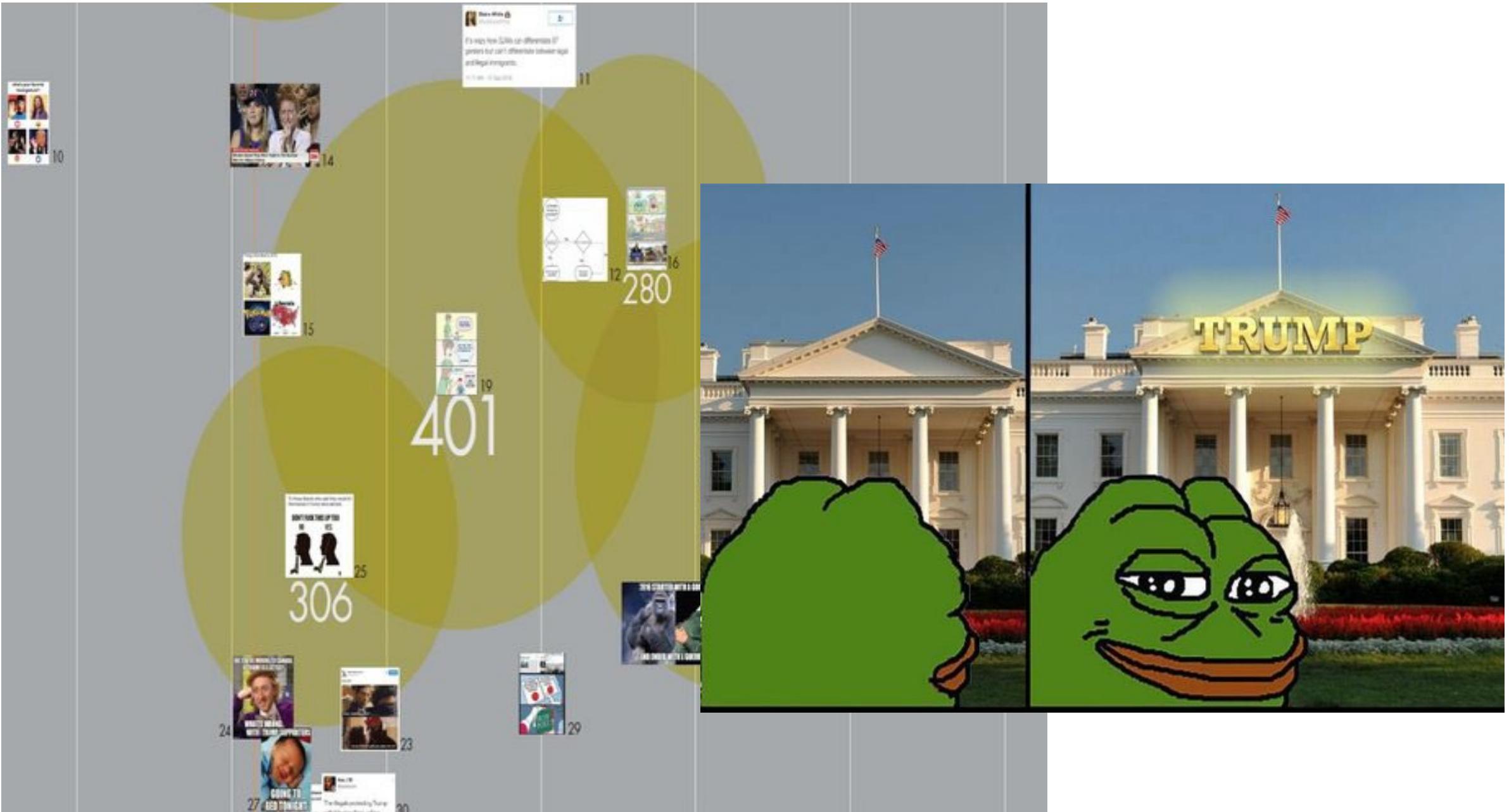
MOSCOW

82.0%

17.3%

23.3

25.7



```

#header - authorization
def create_headers(bearer_token):
    headers = {"Authorization": "Bearer {}".format(bearer_token)}
    return headers

# define url and parameters - set for historical search in full archive
# erik: the 500 max is 500 results per API call.. not just 500 and it will iterate as far as I can
def create_url(keyword, start_date, end_date, max_results = 500):
    search_url = "https://api.twitter.com/2/tweets/search/all" #Change to the endpoint you want to use
    #change params based on the endpoint you are using
    query_params = {'query': keyword,
                    'start_time': start_date,
                    'end_time': end_date,
                    'max_results': max_results,
                    'expansions': 'author_id,in_reply_to_user_id,geo.place_id',
                    'tweet.fields': 'id,text,author_id,in_reply_to_user_id,geo,conversation_id,created_at,public_metrics,retweet_count,referenced_tweets',
                    'user.fields': 'id,name,username,created_at,description,public_metrics,verified',
                    'place.fields': 'full_name,id,country,country_code,geo,name,place_type',
                    'next_token': {}}
    return (search_url, query_params)

#connect to endpoint
def connect_to_endpoint(url, headers, params, next_token = None):
    params['next_token'] = next_token #params object received from create_url function
    response = requests.request("GET", url, headers = headers, params = params)
    print("Endpoint Response Code: " + str(response.status_code))
    if response.status_code != 200:
        raise Exception(response.status_code, response.text)

    # returns the JSON
    return response.json()

```

Dashboard

Projects

Public Perception of AI

For non-commercial use only ⓘ ACADEMIC

MONTHLY TWEET CAP USAGE ⓘ

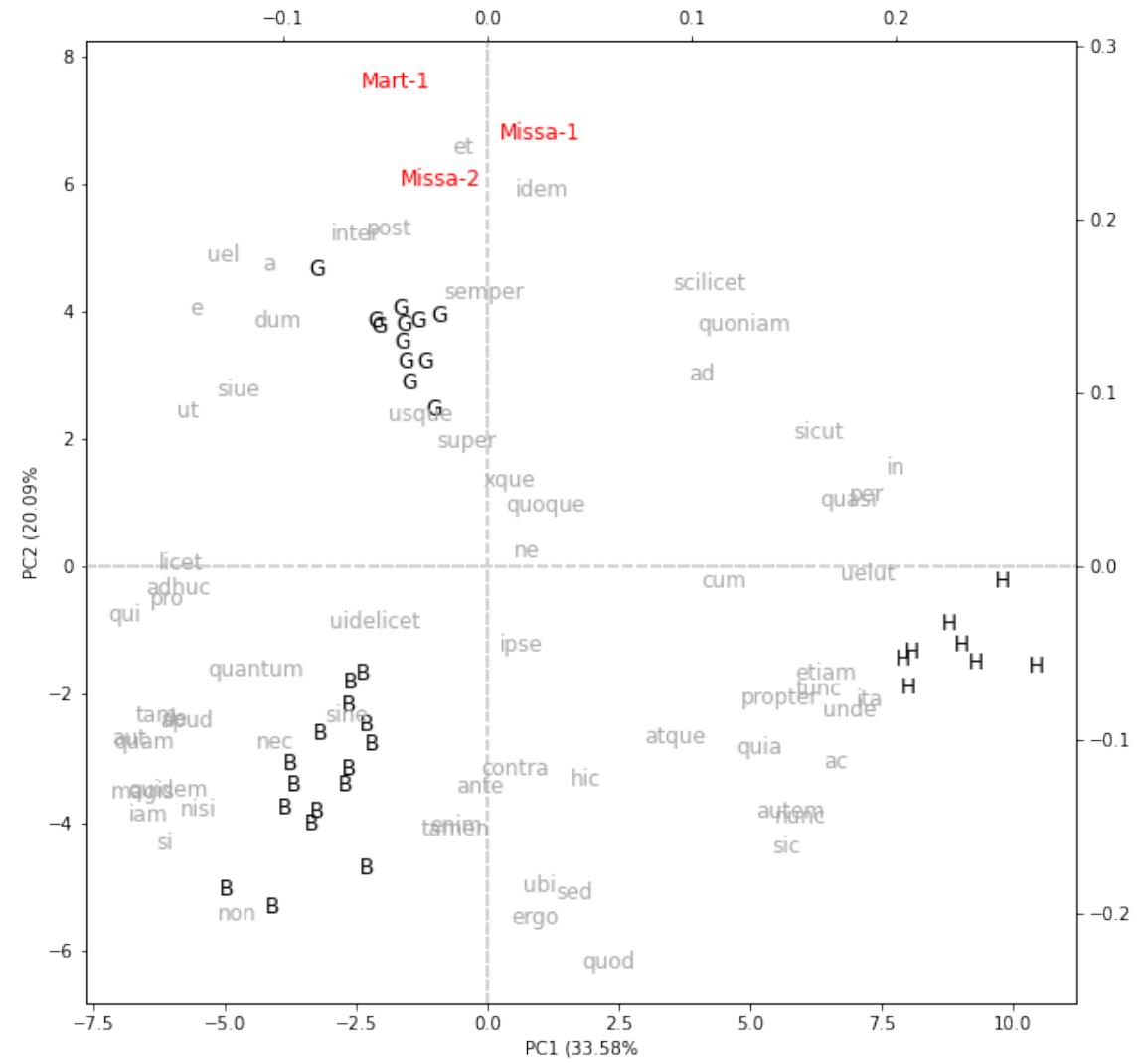
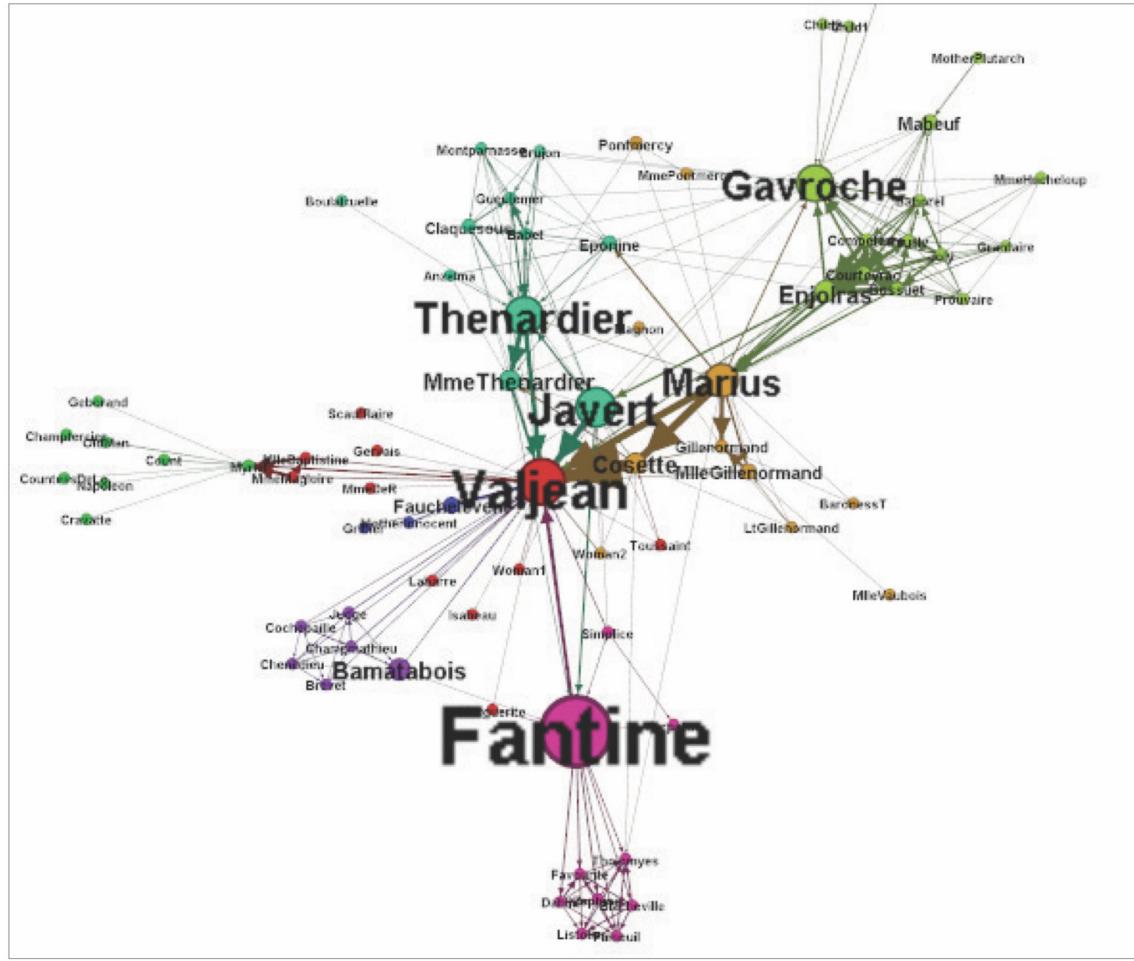
70%

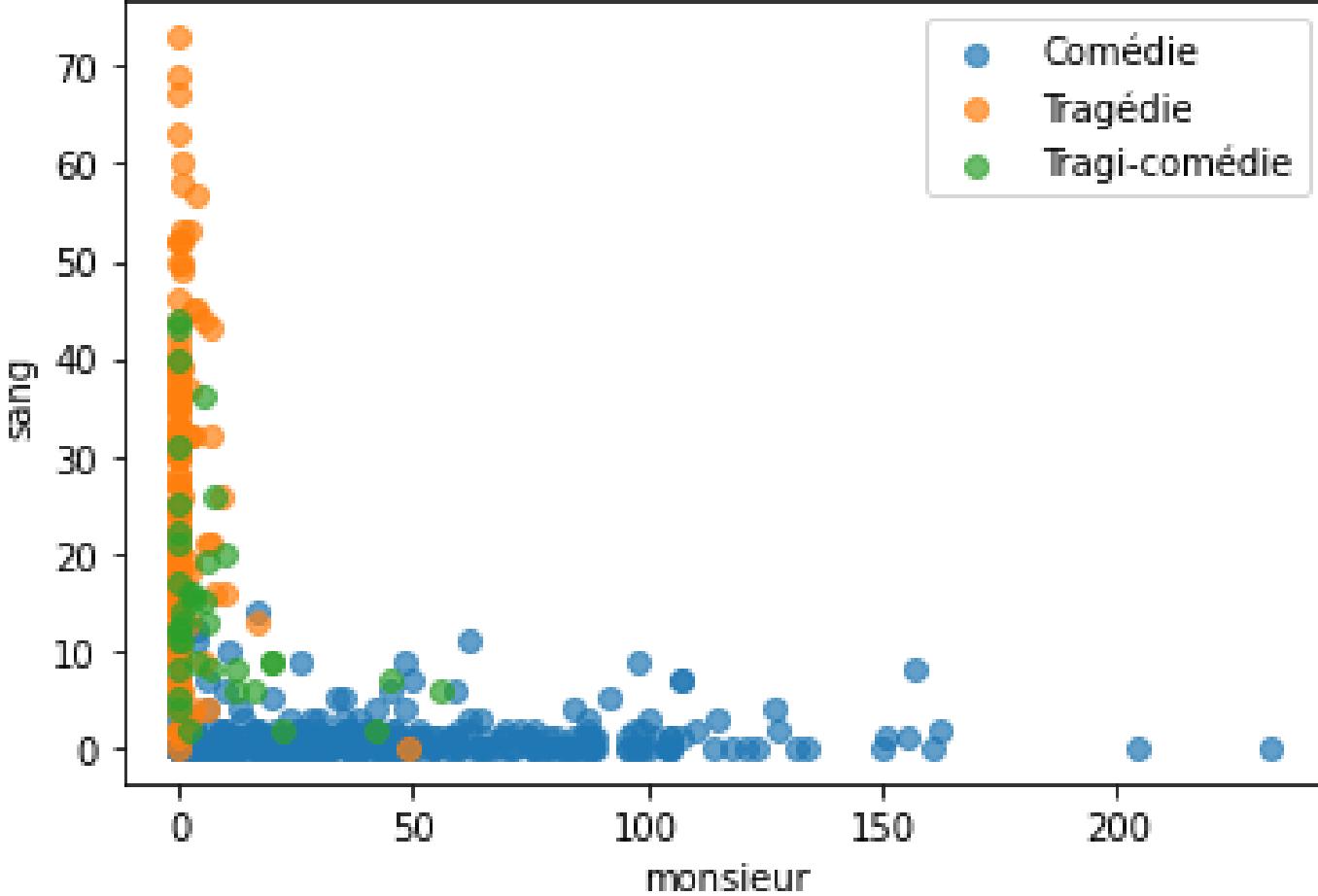
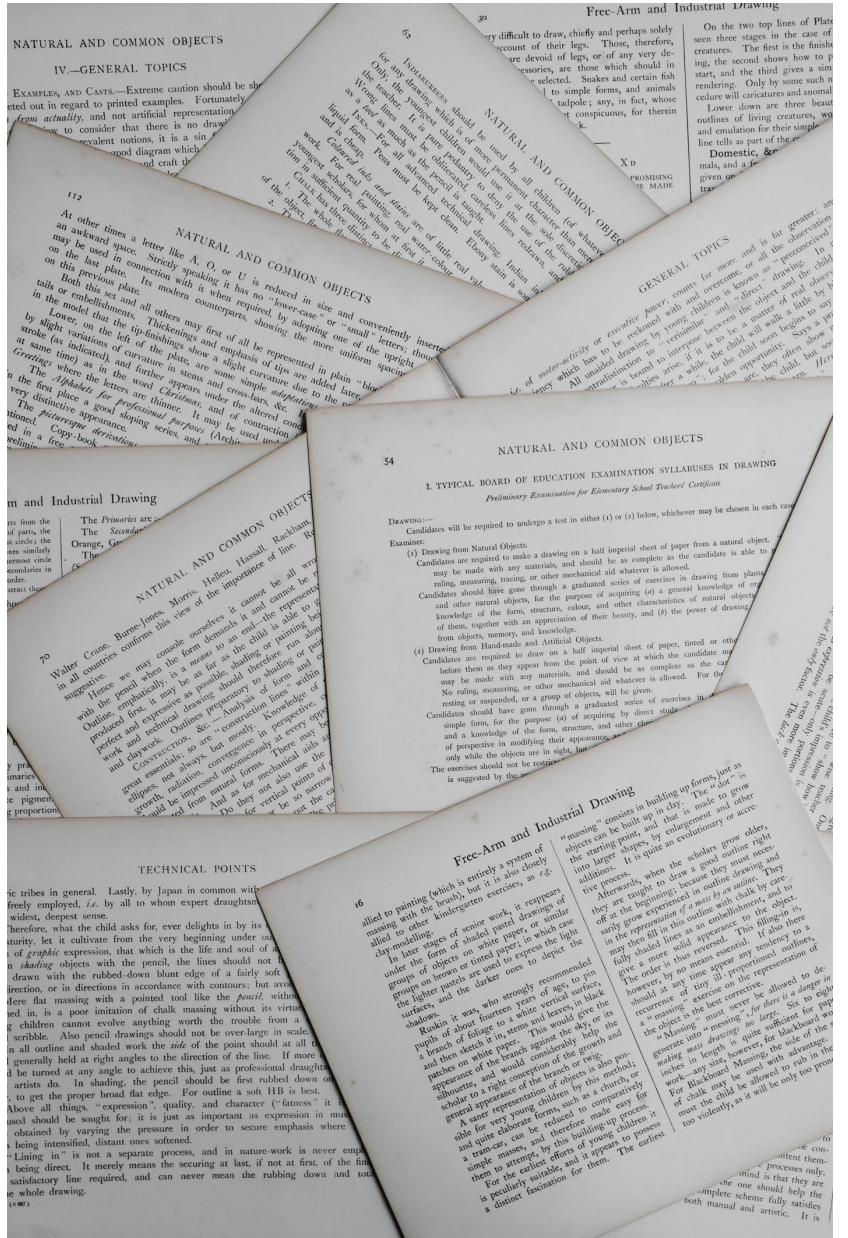
6,999,124 Tweets pulled of 10,000,000

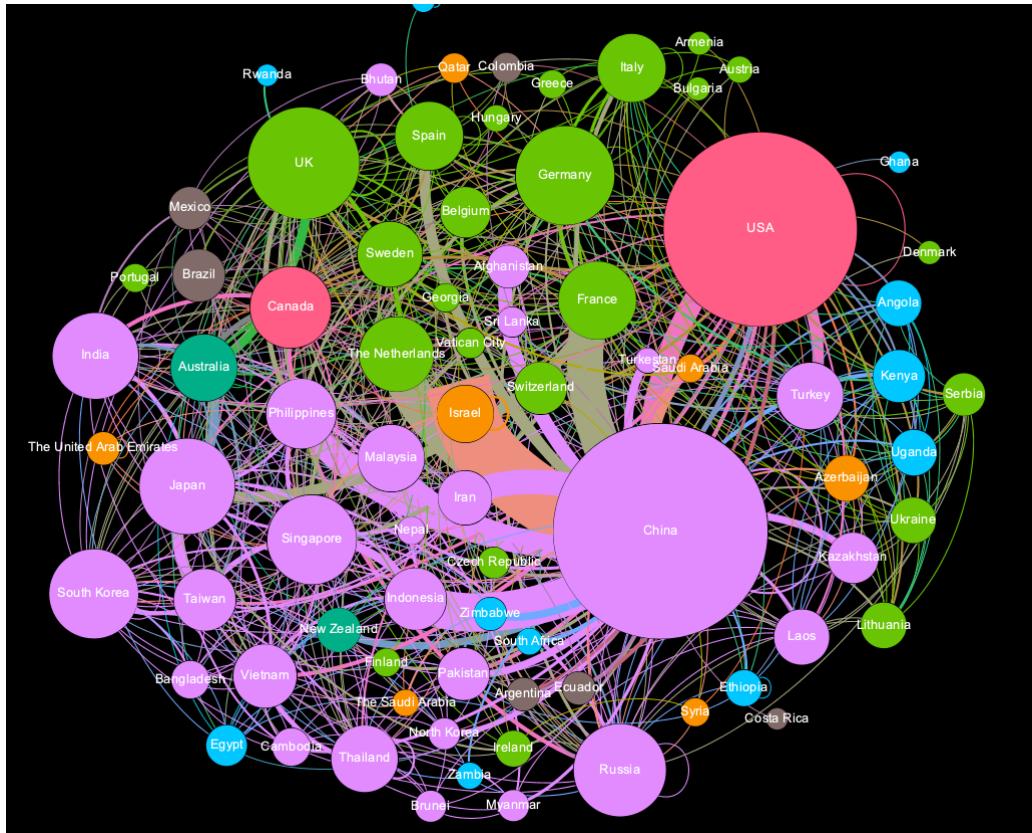
Resets on February 14 at 00:00 UTC

PROJECT APP

Technology Discourse Observer







- Shift from *data-scarce* to *data rich*
- Dynamic instead of static
- High resolution instead of coarse aggregations
- Complex not “simple”



Distant Reading

vs.

Close Reading



Digital Humanities

“Digital Humanities scholars use computers to analyse mostly historical artefacts created by professionals, such as writers, artists and musicians”

(Manovich, 2017: 56)



Digital Humanities (DM)

How did genres change over time?

How did language change over time?

Who is the author of this text (stylometry)?

How do people give meaning to cultural output?



Criticism on Digital Humanities

- Distant reading is no reading at all
- Ignorant of context and domain knowledge
- Cultural artifacts are not just data
- Digital humanities approaches are superficial
- Does not explain causes (e.g., concerning social injustice)





Digital Humanities

“We cannot have a computational digital humanities without **critically examining** the hardware, software, and social infrastructures and the history of methods that make computation as such possible.”

(Dobson, 2019: 30, my emphasis)



Computational Social Sciences

- Behaviour
- Relationships
- Social processes and communication
- Creating of (social) meaning
- Discourses
- Descriptive and inferential

Digital Humanities/Cultural Analytics

- Cultural activity
- Creating of (cultural) meaning
- Focus on arts, literature, film, music, history etc.
- Media formats and content
- Creativity also in digital networks
- Discourses
- Mostly descriptive



Counter-Arguments to Criticism

- Not to replace human experts but offer novel tools & techniques
- Explore massive datasets and then zone-in on individual cases
- Link distant and close reading or *surface* and *depth* studies
- Overcome limitations of conventional sampling
- Increase consistency and reliability
- Data can reveal patterns and support research on causation
- Acknowledges that quantitative approaches alone do not suffice

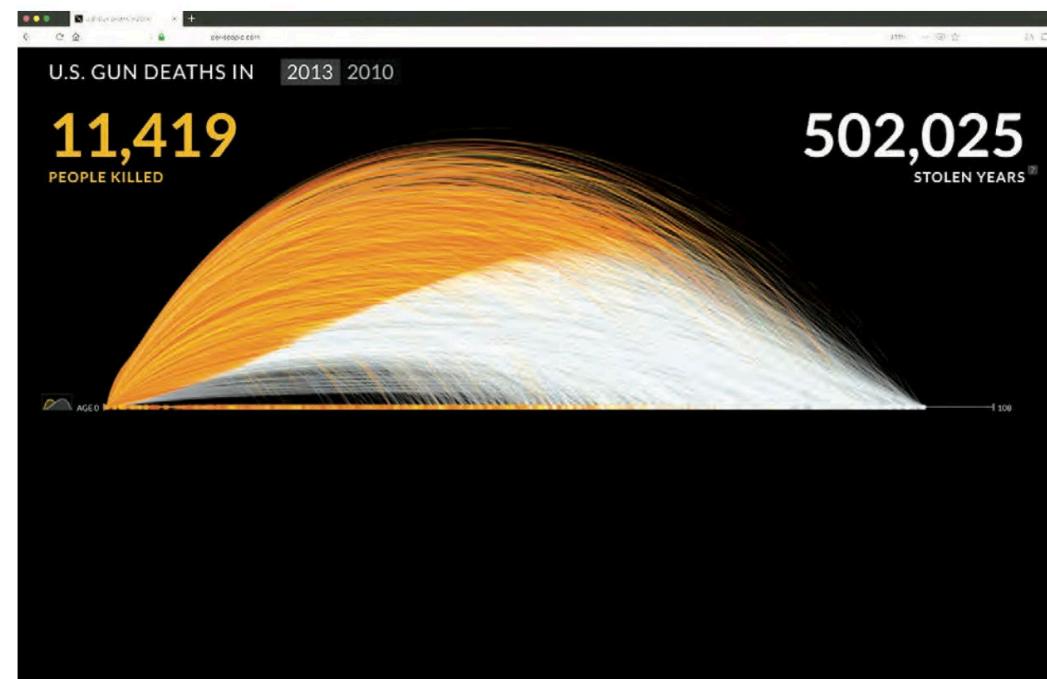
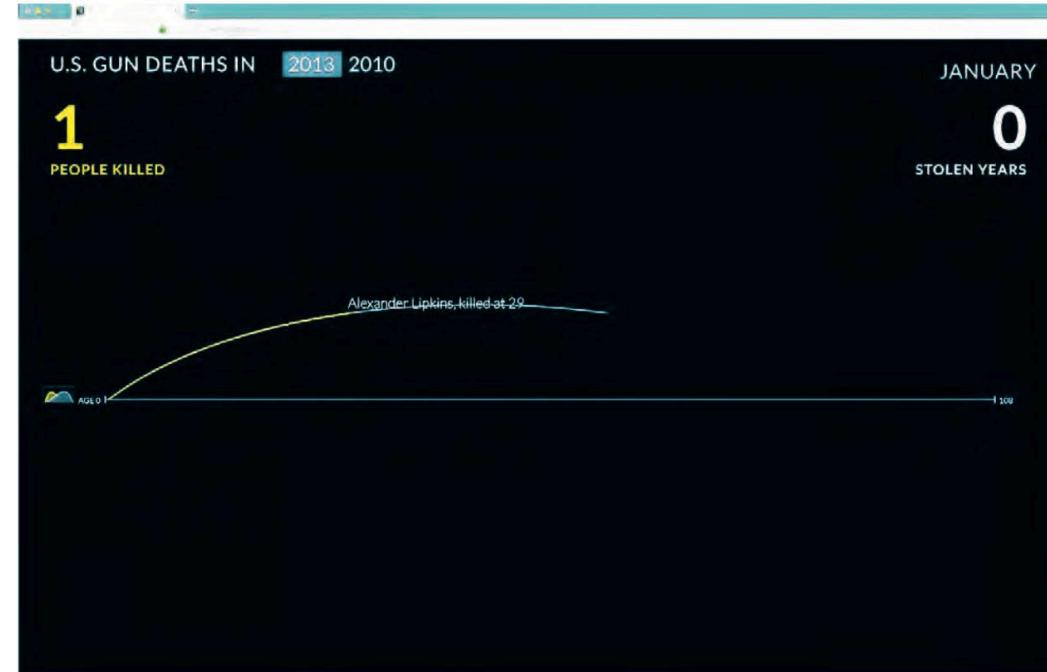
Scientific method is reframed and not abandoned!

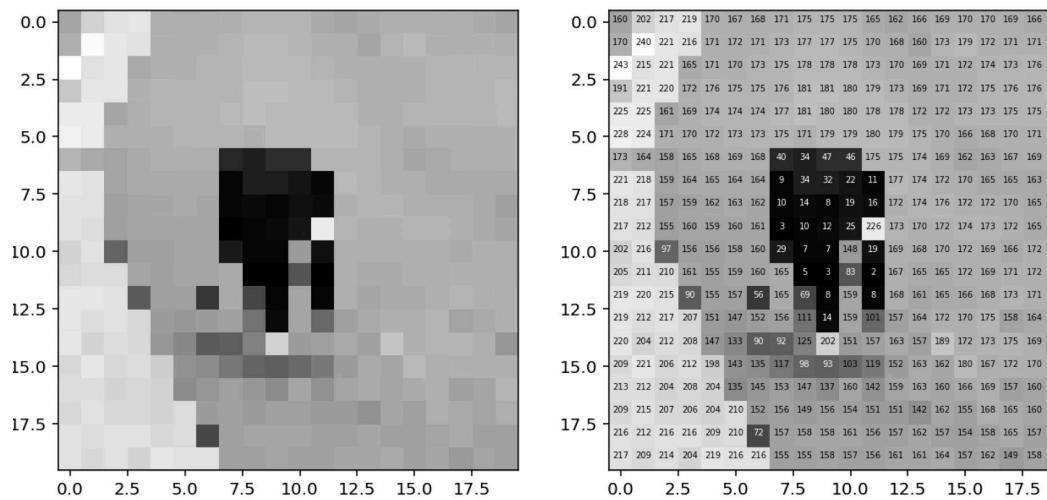


What does it now mean to “think in data”?

“Media and communications researchers need to acknowledge that media are **consumed and used on a large scale**, implying a need for quantitative research, but eventually become a very **individual, subjective experience**, which indicates a need for qualitative research”

(Nguyen, 2020: 26, paraphrasing Baumann & Scherer, 2012)





```
array([[[160.],
       [202.],
       [217.],
       [219.],
       [170.],
       [167.],
       [168.],
       [165.],
       [162.],
       [166.],
       [169.],
       [170.],
       [170.],
       [171.],
       [171.],
       [170.],
       [170.],
       [170.],
       [170.]]])
```



ChatGPT

If you're encountering ages that are implausibly young for your dataset, such as 2 and 5 years old, it's reasonable to apply additional filtering criteria to your data. This process involves setting a minimum age threshold that is considered plausible for your dataset's context and excluding any data points below this threshold.

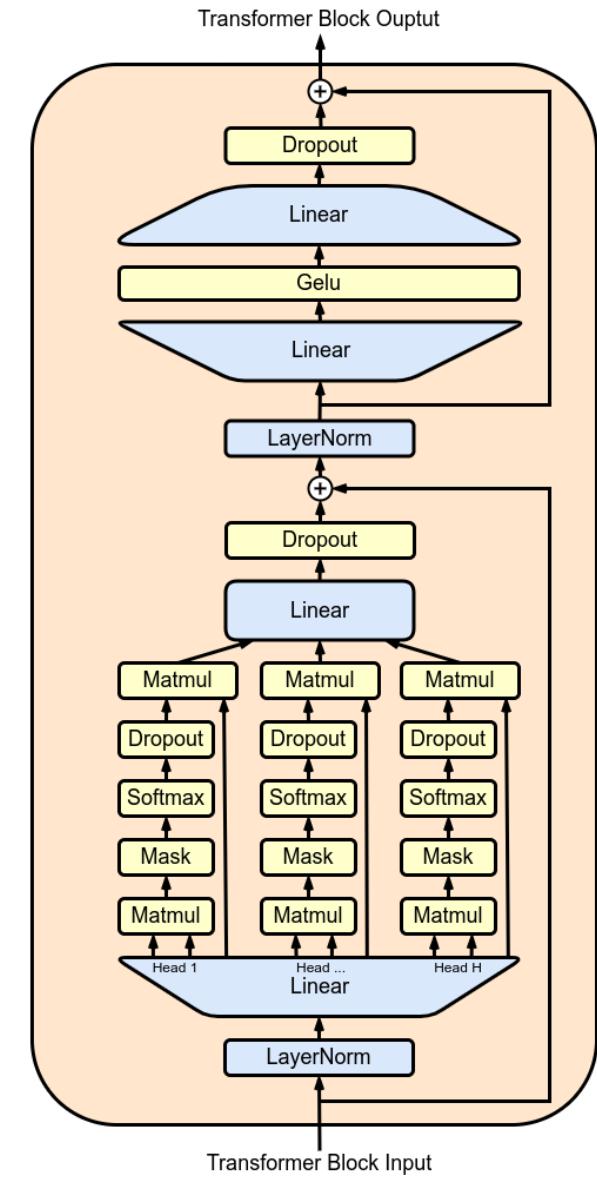
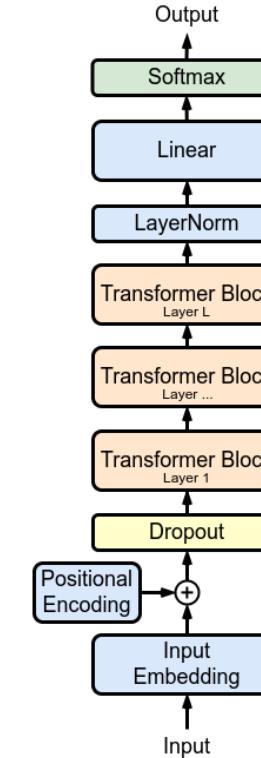
Let's apply a filter to exclude ages that are below a certain threshold, for example, ages less than 18, in addition to filtering out extreme outliers as previously described. You can adjust this minimum age threshold based on the specific requirements of your dataset and analysis:

python

Copy code

```
import pandas as pd  
import matplotlib.pyplot as plt
```

Generative pre-trained transformers (GPT)



Data Ethics

- Privacy
- Consent
- Anonymity/Identifiability
- Harm & Vulnerability



Questions pertaining to Tool Criticism:

- Where do your data come from? What do they actually represent?
- How does the source of data determine the shape and meaning of data?
- What pre-processing steps were applied and why? What about post-processing?
- What methods were applied to the data to what analytical ends? What parameters were chosen and why?
- Can the data be shared along with the scripts?

Use Cases

Dr. Dennis Nguyen

Assistant Professor Digital Literacy & Digital Methods



Utrecht
University

DATA SCHOOL

"How well did Dutch journalism cover the COVID-19 pandemic?" -> empirical research into journalistic practices and journalistic quality for professional self-reflection, combining data science and media studies



0,45

0,4

0,35

0,3

0,25

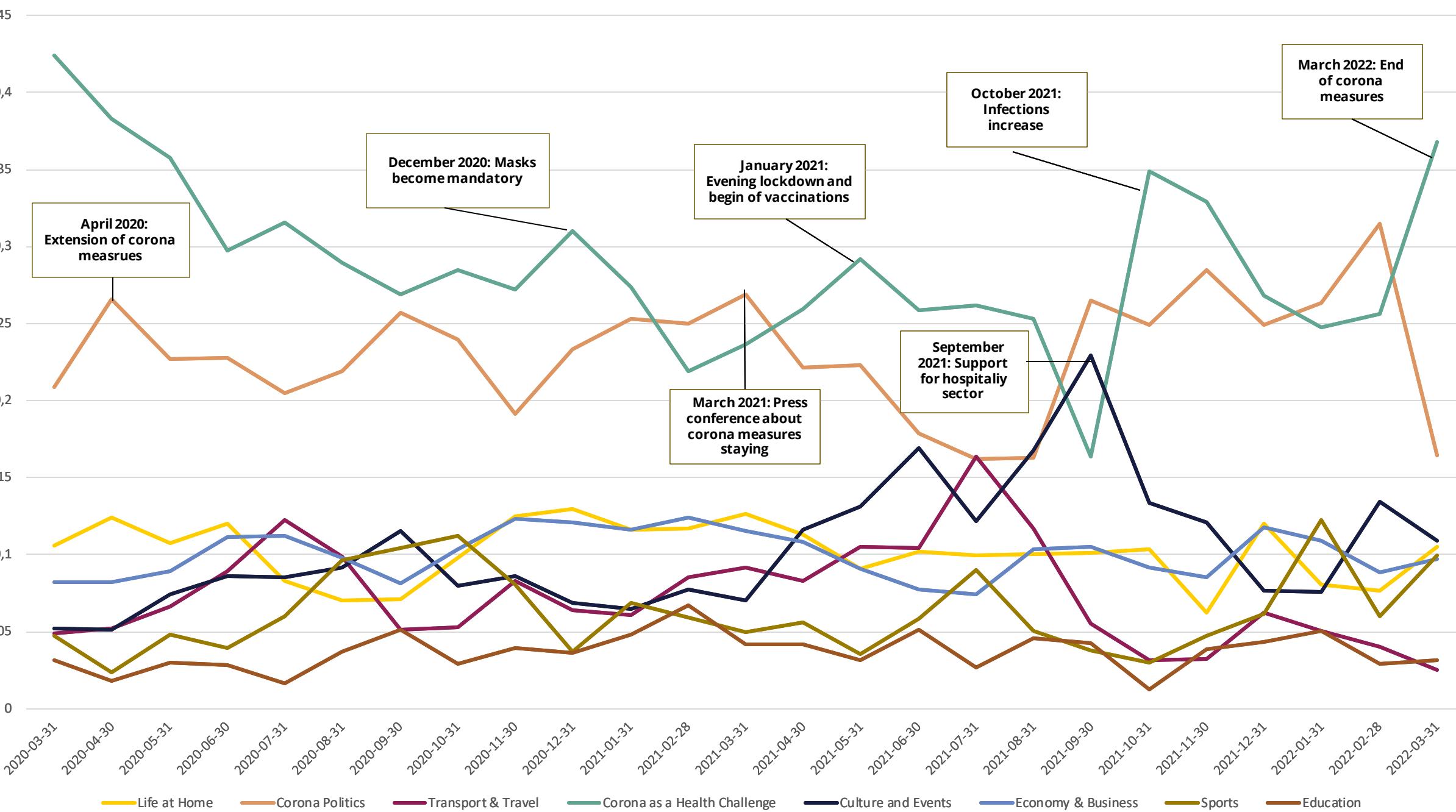
0,2

0,15

0,1

0,05

0



ORIGINAL ARTICLE



The news framing of artificial intelligence: a critical exploration of how media discourses make sense of automation

Dennis Nguyen¹ · Erik Hekman²



Volume 12, Issue 2

How news media frame data risks in their coverage of big data and AI

Dennis Nguyen, Utrecht University, The Netherlands, d.nguyen1@uu.nl

PUBLISHED ON: 23 May 2023 DOI: [10.14763/2023.2.1708](https://doi.org/10.14763/2023.2.1708)

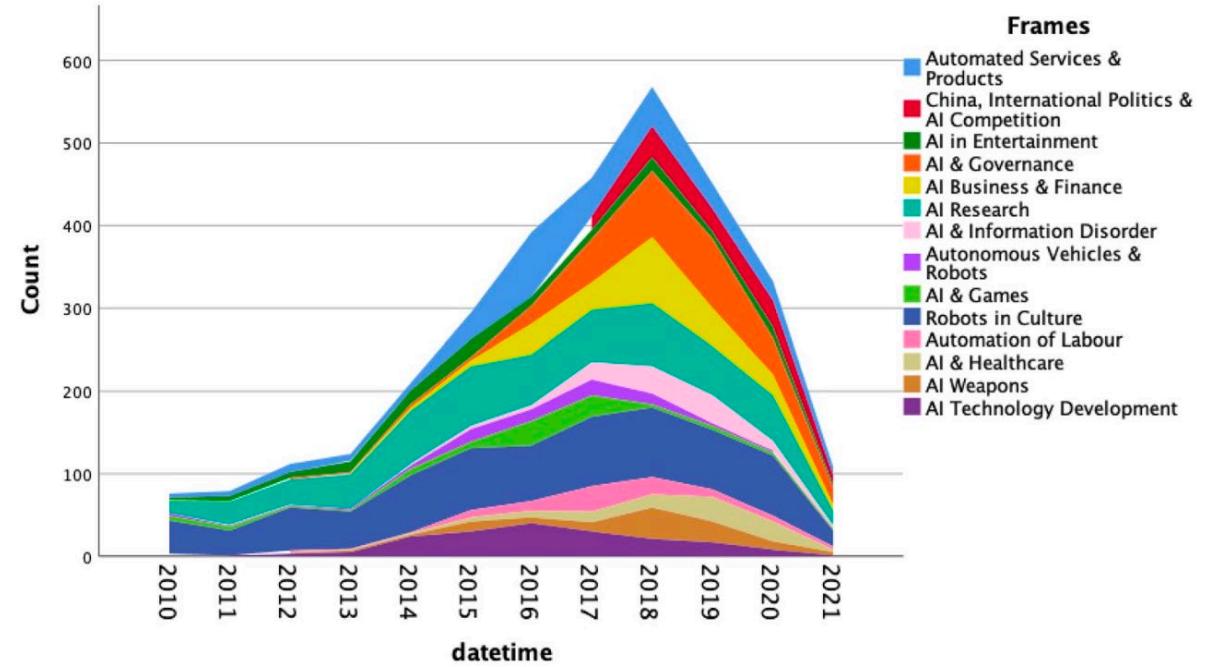
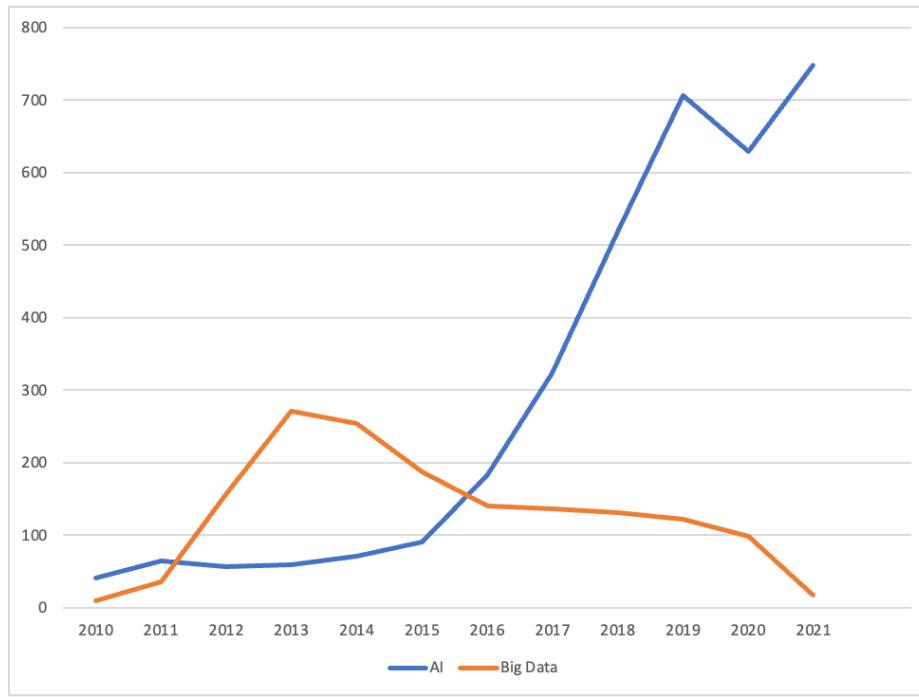


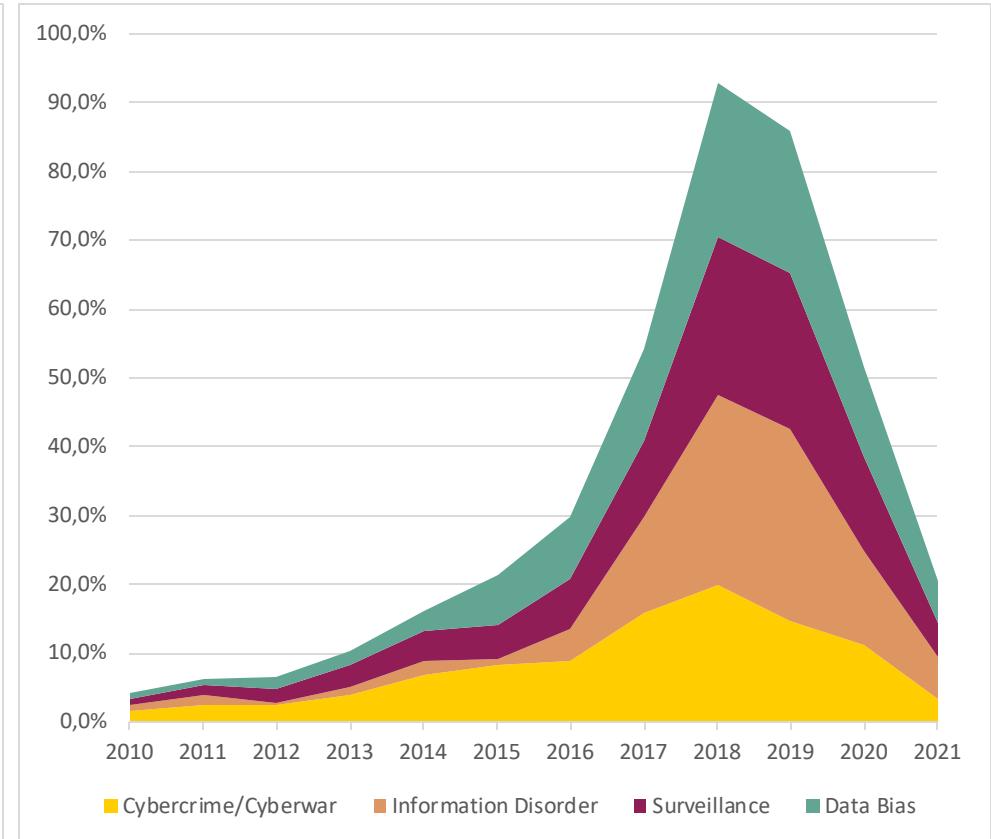
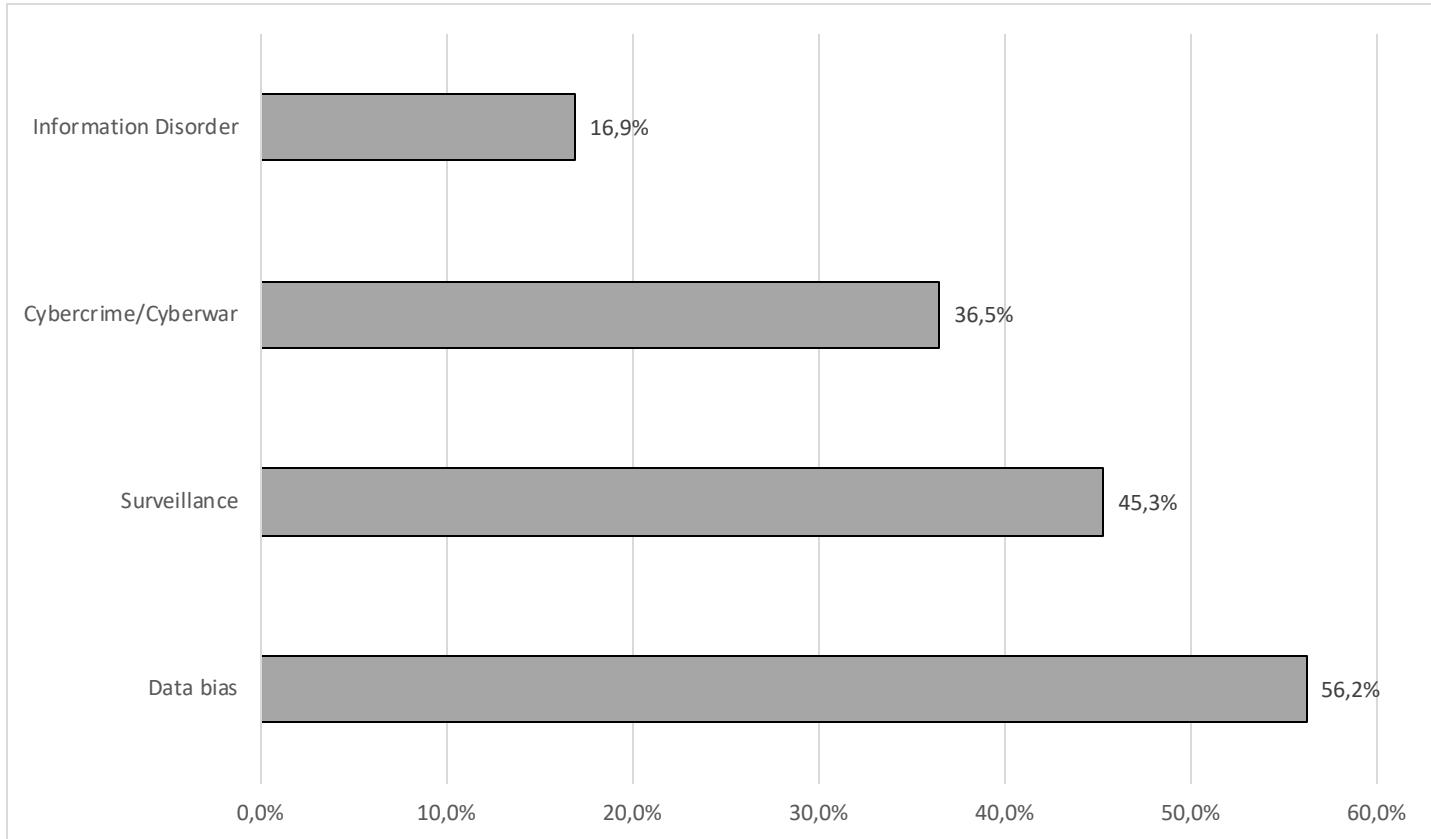
Original Article

A ‘New Arms Race’? Framing China and the U.S.A. in A.I. News Reporting: A Comparative Analysis of the Washington Post and South China Morning Post

Global Media and China
2022, Vol. 7(1) 58–77
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: [10.1177/20594364221078626](https://doi.org/10.1177/20594364221078626)
journals.sagepub.com/home/gch









```

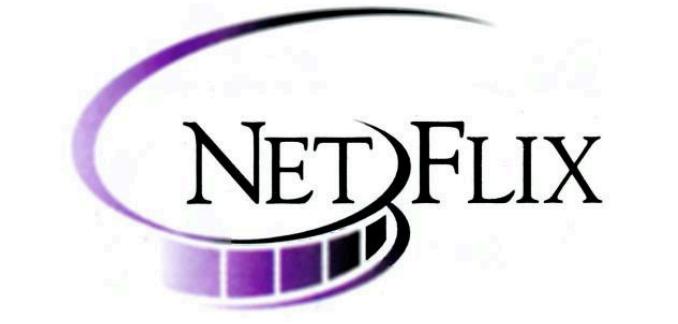
1 #inspect
2 df_viewings = pd.concat([df1, df2], ignore_index=True)
3 df_viewings

```

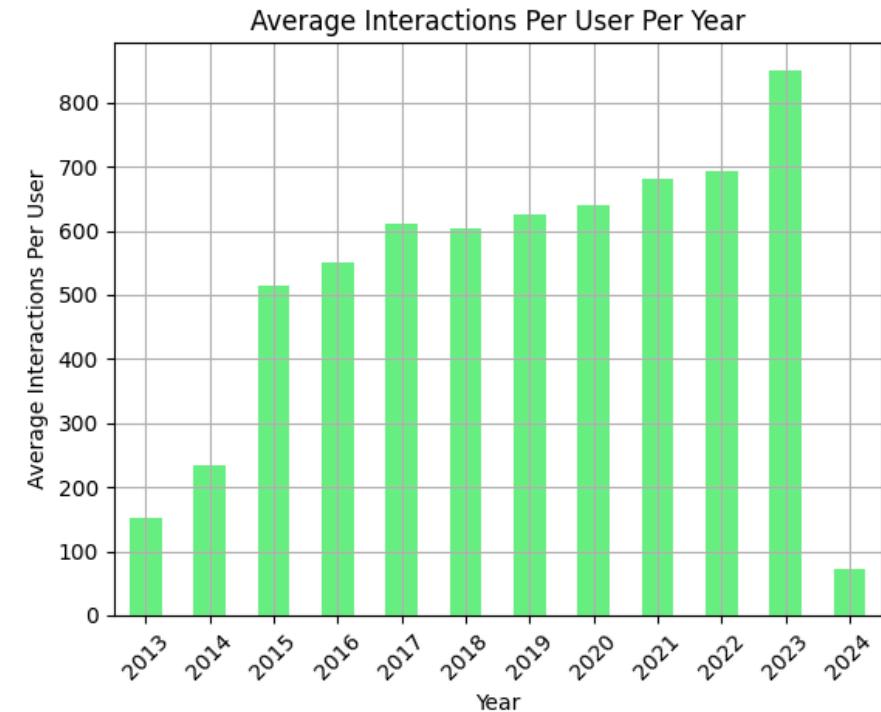
[3]

	ID	Start tijd	Aantal uur gekeken	Titel	Apparaat
0	1706049637869	2023-12-31 22:02:01	1.977	A Man Called Otto	Chrome PC (Cadmium)
1	1706049637869	2023-12-31 22:01:55	0.001	A Man Called Otto_hook_01_16x9	Chrome PC (Cadmium)
2	1706049637869	2023-12-31 21:25:32	0.044	Good Will Hunting	Chrome PC (Cadmium)
3	1706049637869	2023-12-31 21:25:24	0.002	Good Will Hunting_hook_primary_16x9	Chrome PC (Cadmium)
4	1706049637869	2023-12-31 21:25:04	0.002	Season 1 Trailer 1: Onimusha	Chrome PC (Cadmium)
...
10515311	1706785016375	2020-04-19 14:52:57	0.102	Cheer: Seizoen 1: Full Out (Aflevering 5)	Android DefaultWidevineL3Tablet Android Tablet
10515312	1706785016375	2020-04-19 13:52:50	1.000	Cheer: Seizoen 1: Hit Zero (Aflevering 4)	Android DefaultWidevineL3Tablet Android Tablet
10515313	1706785016375	2020-04-18 17:42:34	1.735	Central Intelligence	Android DefaultWidevineL3Tablet Android Tablet
10515314	1706785016375	2020-04-17 15:59:33	0.932	Cheer: Seizoen 1: Blood, Sweat and Cheers (Afle...)	Android DefaultWidevineL3Tablet Android Tablet
10515315	1706785016375	2020-04-16 15:22:16	0.910	Cheer: Seizoen 1: Making Mat (Aflevering 2)	Android DefaultWidevineL3Tablet Android Tablet

10515316 rows x 5 columns



Titel	
Miniserie (Clip): Fool Me Once	196
The Crown: Seizoen 1: Wolferton Splash (Aflevering 1)	195
The Vampire Diaries: Seizoen 4: We'll Always Have Bourbon Street (Aflevering 8)	193
The Vampire Diaries: Seizoen 4: My Brother's Keeper (Aflevering 7)	188
Seizoen 1 (Trailer): The Night Agent	180
How to Lose a Guy in 10 Days	176
The Vampire Diaries: Seizoen 3: Heart of Darkness (Aflevering 19)	174
Stranger Things: Seizoen 1: Hoofdstuk 1: The Vanishing Of Will Byers (Aflevering 1)	163
Glass Onion: A Knives Out Mystery	162
Undercover: Seizoen 1: Camping Zonnendauw (Aflevering 1)	160
To All the Boys I've Loved Before	160
Seizoen 1 (Clip): The Recruit	157
Seizoen 1 (Clip): The Night Agent	153
Miniserie (Clip): Treason	150
The Lord of the Rings: The Fellowship of the Ring	148
The Vampire Diaries: Seizoen 4: Graduation (Aflevering 23)	144
Lupin: Deel 1: Hoofdstuk 1 (Aflevering 1)	142
Miniserie (Clip): Liebes Kind	142
The Kissing Booth	141
Miniserie (Clip): En helt vanlig familj	141
The Crown: Seizoen 1: Hyde Park Corner (Aflevering 2)	141
Seizoen 1 (Clip): The Watcher	141
La casa de papel: Deel 1: Aflevering 1 (Aflevering 1)	140
Extraction	138
The Vampire Diaries: Seizoen 5: Rescue Me (Aflevering 17)	136
Name: count, dtype: int64	



References

- Kitchin, R. (2022): *The Data Revolution. A Critical Analysis og Big Data, Open Data & Data Infrastructures*. Sage.
- Nguyen, D. (2020): 'Media and Communication Studies in the Age of Digitalization and Datafication: How Practical Factors and Research Interests Determine Methodological Choices', in Nguyen D., Dekker, I. & Nguyen, S. (eds): *Understanding Media and Society in the Age of Digitalisation*. London: Palgrave Macmillan.



Utrecht
University

Sharing science,
shaping tomorrow