Unraveling our internet-mediated lives

Juhi Kulshrestha

Assistant Professor,

Department of Computer Science

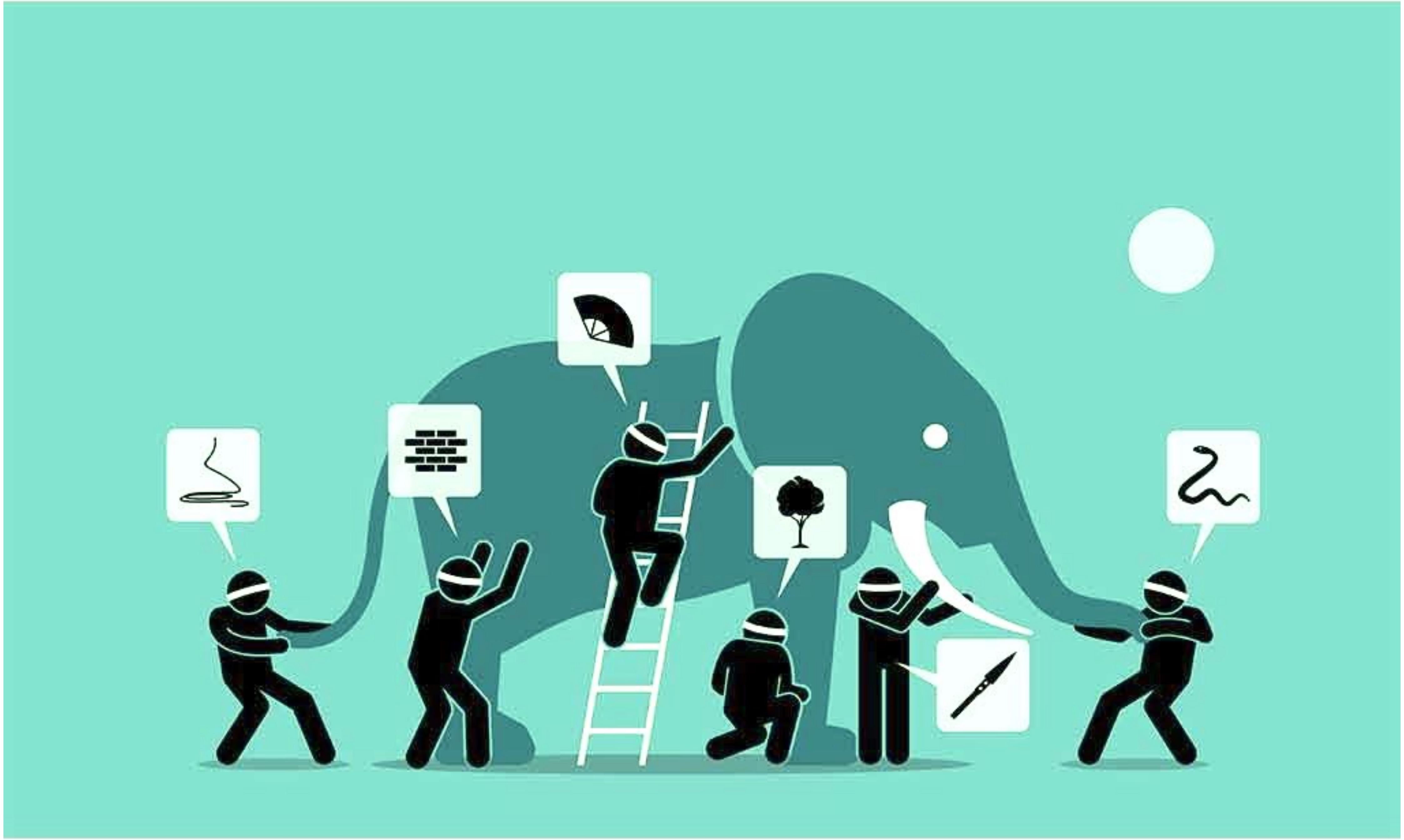
Aalto University, Finland

Keynote @ SICSS Amsterdam

10th June 2024

What is CSS?

What is CSS?



[Image credit: firstcry.com]

What is CSS (to me)?

[Image credit: cambridgeblog.org/]

Combining varied forms of digital behavioural data with diverse research methods to study and solve societally relevant problems.

My journey to CSS

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

Ph.D. in Computer Science

LEIBNIZ INSTITUTE FOR MEDIA RESEARCH HANS-BREDOW-INSTITUT

Visiting postdoctoral fellow
“Algorithmed Public Spheres”

gesis Leibniz Institute for the Social Sciences

Postdoctoral researcher
Computational Social Science (CSS) Dept.

Universität Konstanz

Assistant Professor for CSS
Dept. of Politics & Public Administration

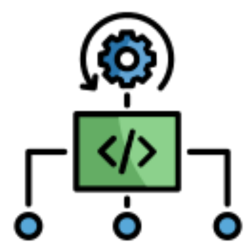
A? Aalto University

Assistant Professor
Computer Science Dept.

Studying our internet mediated lives

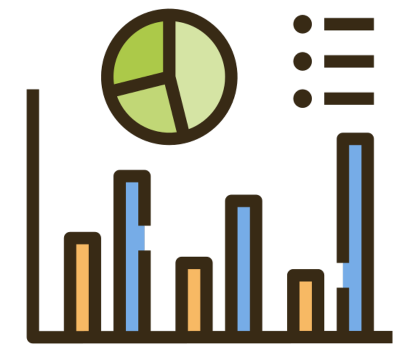
Studying our internet mediated lives

Digital behavioural data



Algorithm audit frameworks

Online surveys & experiments



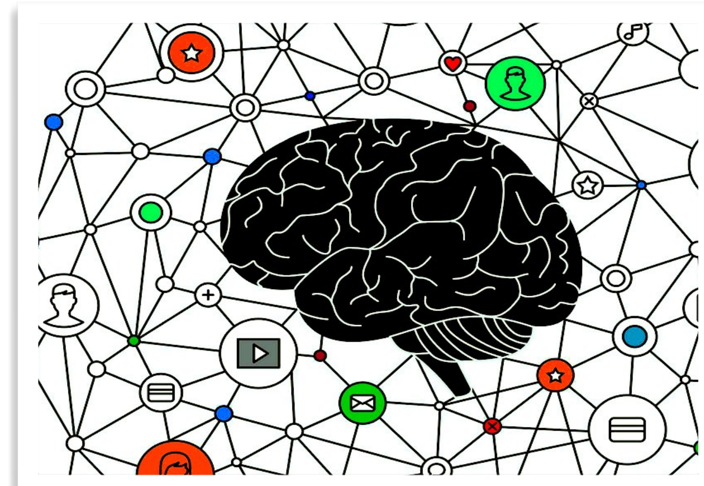
Data science
Machine learning
Statistical analysis
Natural language processing…

Study human behaviour on the web and the impact it has on individuals' opinions, attitudes and behaviours both online and offline

Human behaviour on the web

[Image credit: cyberbullying.org]

Influence of online behaviour



[Image credit: VectorStock.com/26073249]

[Image credit: nbjournal.ru/media-literacy]

Bias & inaccuracies in online news use

[Image credit: digicol.de/news]

Impact of online info retrieval systems

Human behaviour on the web

Influence of online behaviour

credit: VectorStock.com/26073249]

[Image credit: nbjournal.ru/media-literacy]

[Image credit: digicol.de/news]

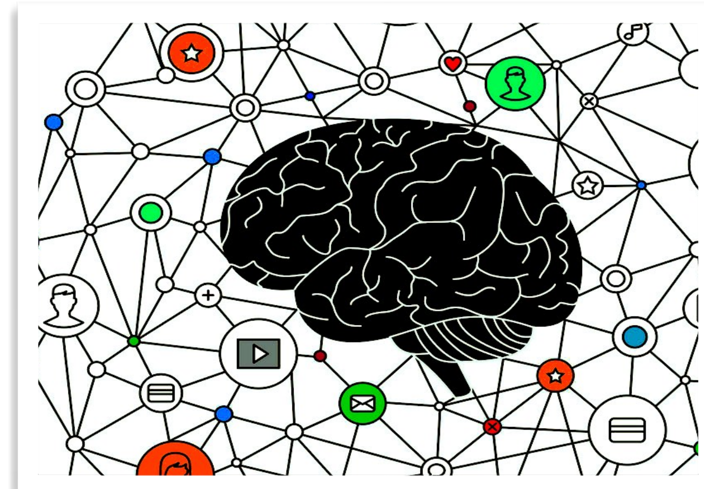
Bias & inaccuracies in online news use

Impact of online info retrieval systems

Human behaviour on the web

[Image credit: cyberbullying.org]

Influence of online behaviour



[Image credit: VectorStock.com/26073249]

[Image credit: nbjournal.ru/media-literacy]

Bias & inaccuracies in online news use

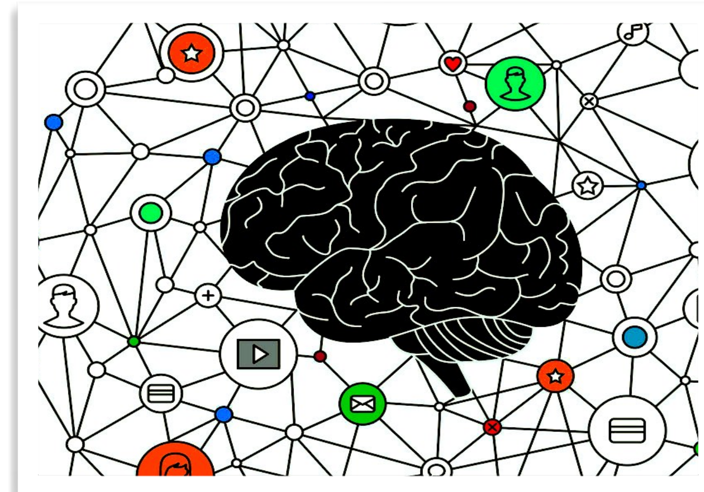
[Image credit: digicol.de/news]

Impact of online info retrieval systems

Human behaviour on the web

[Image credit: cyberbullying.org]

Influence of online behaviour



[Image credit: VectorStock.com/26073249]

[Image credit: nbjournal.ru/media-literacy]

Bias & inaccuracies in online news use

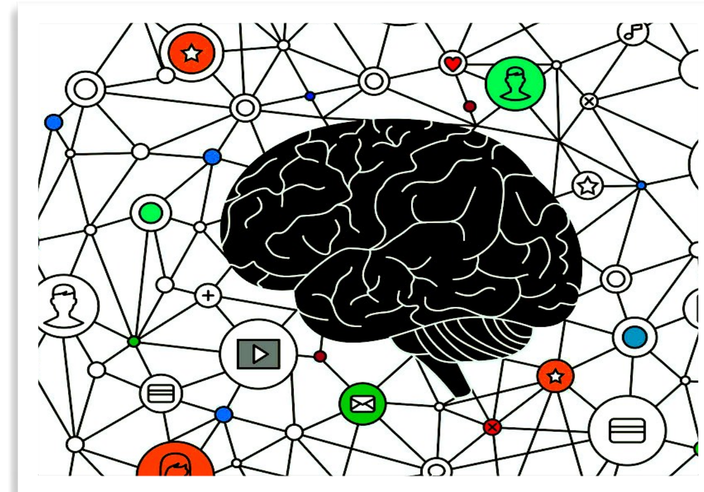
[Image credit: digicol.de/news]

Impact of online info retrieval systems

Human behaviour on the web

[Image credit: cyberbullying.org]

Influence of online behaviour



[Image credit: VectorStock.com/26073249]

[Image credit: nbjournal.ru/media-literacy]

Bias & inaccuracies in online news use

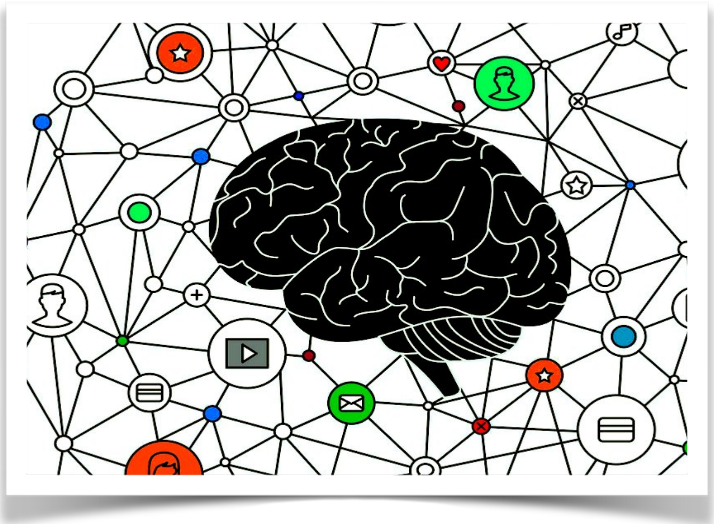
[Image credit: digicol.de/news]

Impact of online info retrieval systems

Human behaviour on the web

[Image credit: cyberbullying.org]

Influence of online behaviour



[Image credit: VectorStock.com/26073249]

[Image credit: nbjournal.ru/media-literacy]

Bias & inaccuracies in online news use

[Image credit: digicol.de/news]

Impact of online info retrieval systems

Human behaviour on the web

[Image credit: cyberbullying.org]

Human behaviour on the web

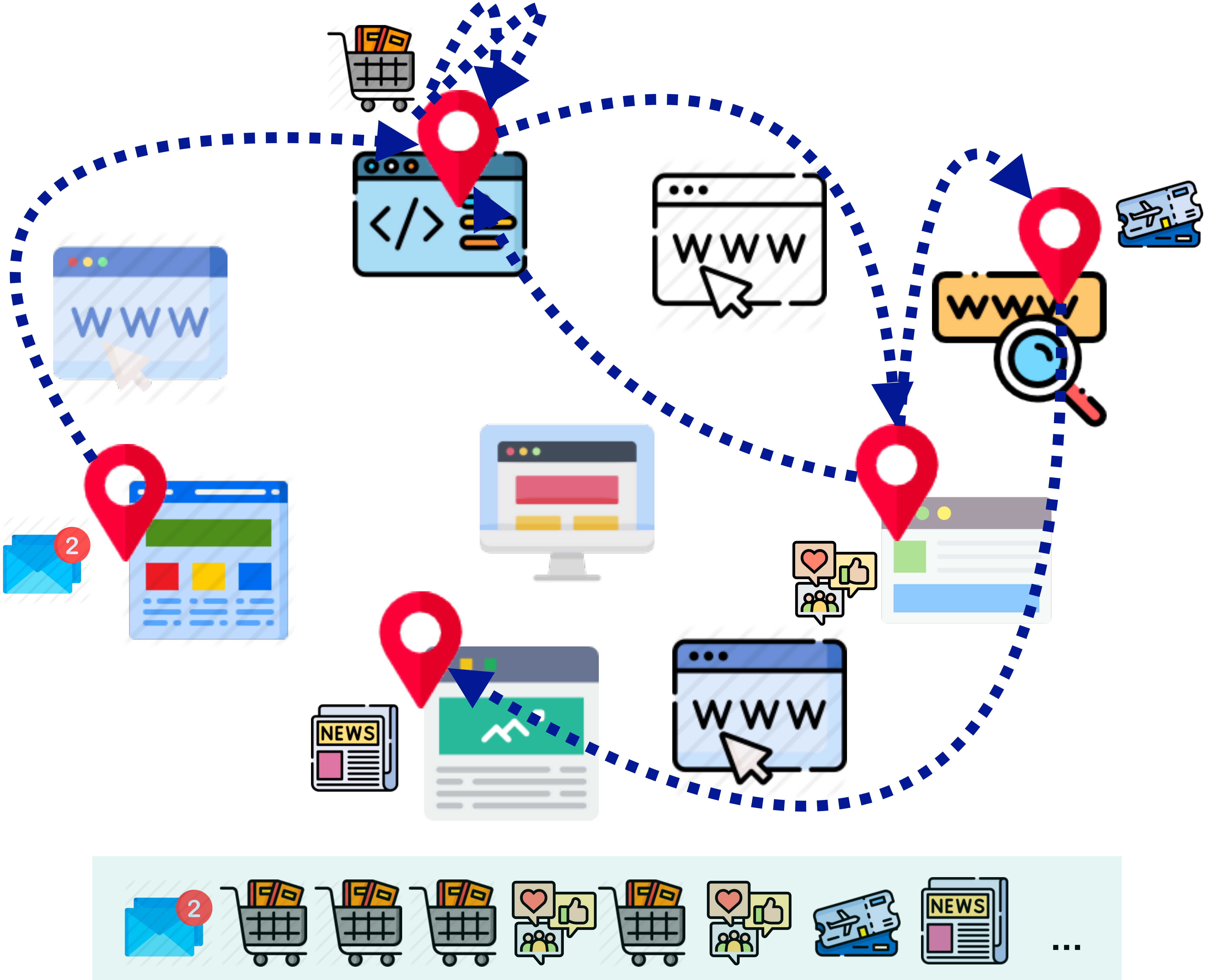
[Image credit: cyberbullying.org]

[AAAI ICWSM 2021]

Web Routineness and Limits of Predictability

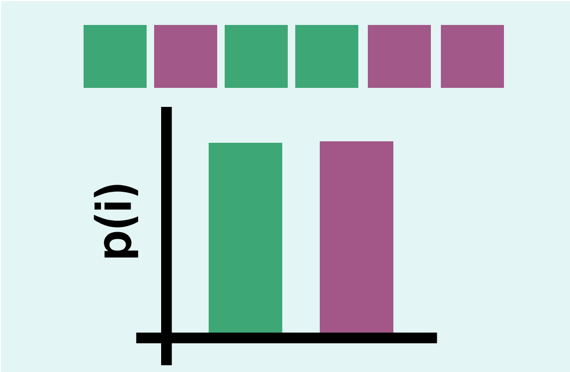
Investigating Demographic and Behavioral Differences Using Web Tracking Data

Juhi Kulshrestha, Marcos Oliveira, Orkut Karaçalik, Denis Bonnay, Claudia Wagner

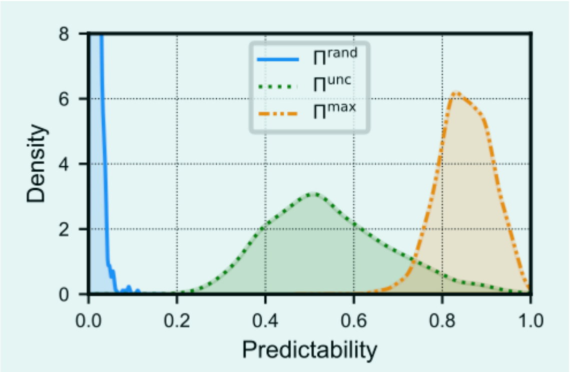


WWW
WWW
WWW
WWW
2
NEWS
2
NEWS
...

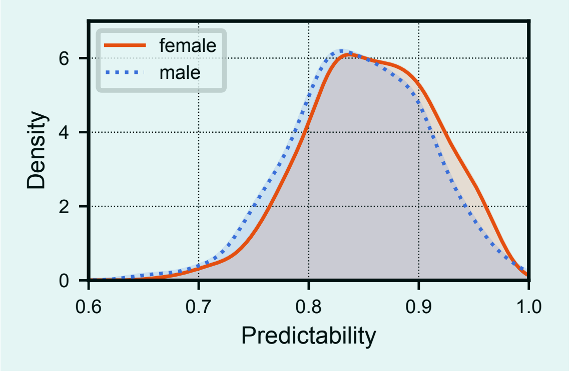
Web routineness & limits of predictability



Predictability measurement framework

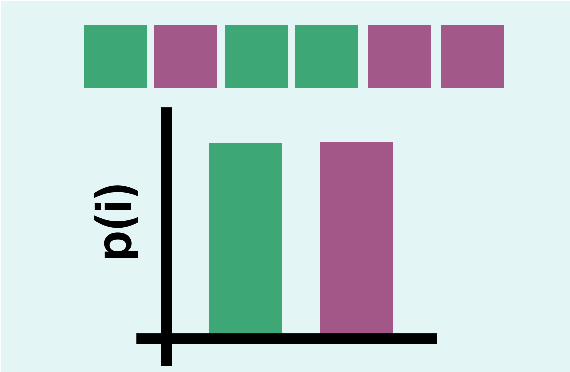


Web routineness and predictability

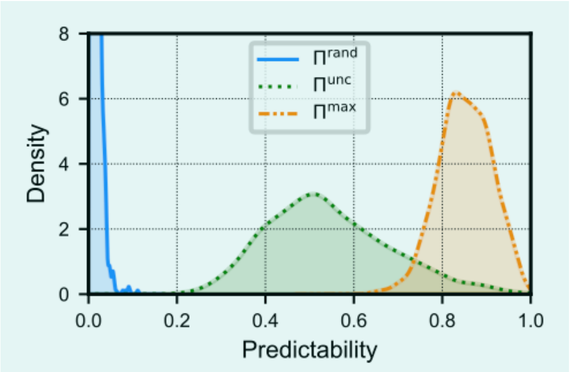


Demographic & behavioral differences

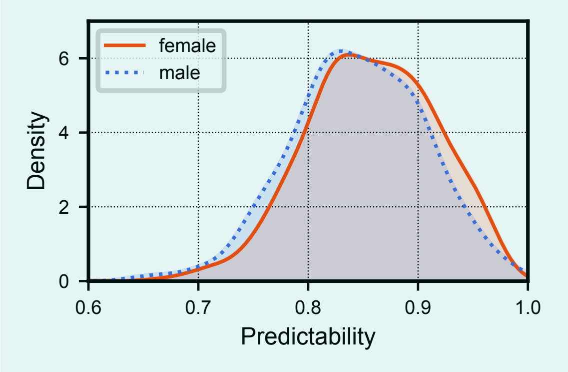
Web routineness & limits of predictability



Predictability measurement framework



Web routineness and predictability



Demographic & behavioral differences

Creating trajectories

- From Web tracking data to trajectories

Time	Duration (s)	URL
2013-05-24 08:19:06	70	https://mail.google.com/mail/u/1/#inbox
2013-05-24 09:15:02	30	https://www.youtube.com/watch?v=I33IGc8Vqy4
2013-05-24 08:35:26	26	https://www.youtube.com/watch?v=EJ6UvCkInRk
2013-05-24 09:15:56	186	https://twitter.com/home

⋮

Creating trajectories

- From Web tracking data to trajectories

Time	Duration (s)	URL
2013-05-24 08:19:06	70	https://mail.google.com/mail/u/1/#inbox
2013-05-24 09:15:02	30	https://www.youtube.com/watch?v=I33IGc8Vqy4
2013-05-24 08:35:26	26	https://www.youtube.com/watch?v=EJ6UvCkInRk
2013-05-24 09:15:56	186	https://twitter.com/home

⋮

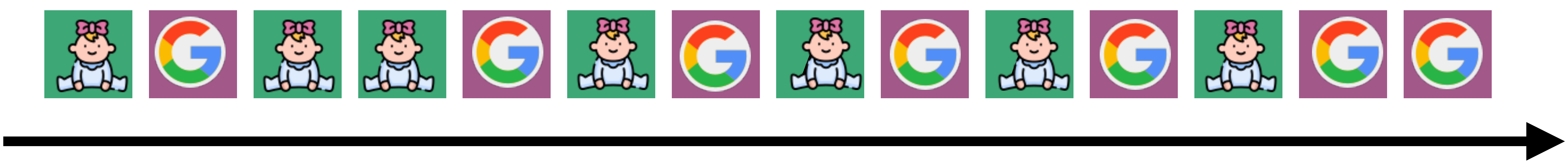
A trajectory is a discrete sequence of locations visited by a user, where a location can be a website, a domain, or a category.

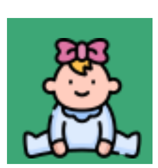
Studying web trajectories

- We would like to learn about:
  - Website preferences
  - Visitation routines

Studying web trajectories

- We would like to learn about:
  - Website preferences
  - Visitation routines



 **= whattoexpect.com/milestones/**

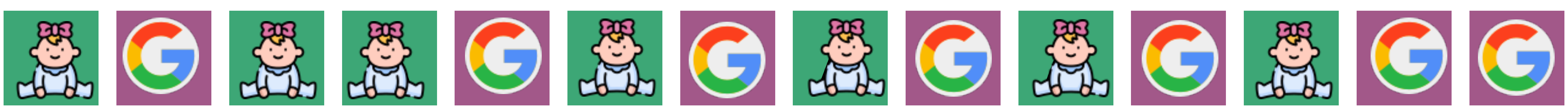
 **= google.com**

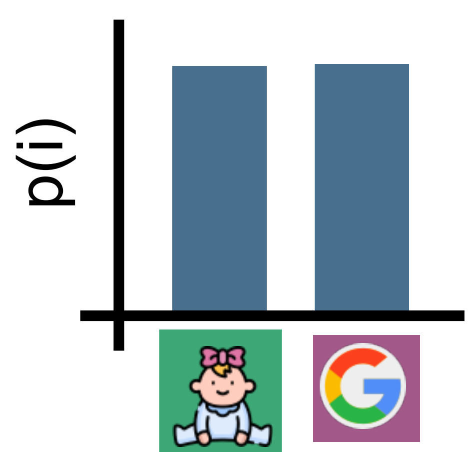
Uncertainty (website preferences)

- Measuring the uncertainty of a random variable.

X

Analyzing the probability distribution of a variable.



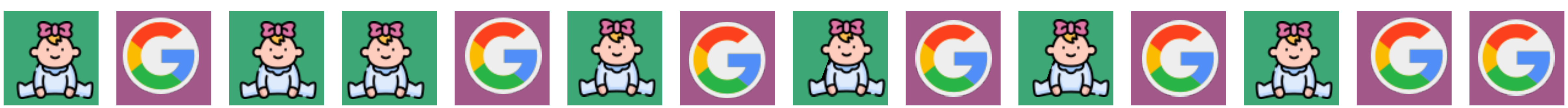


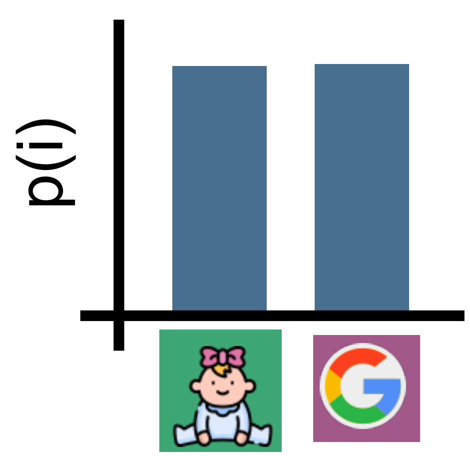
Uncertainty (website preferences)

- Measuring the uncertainty of a random variable.

X

Analyzing the probability distribution of a variable.





Time-*un*correlated entropy (Shannon entropy)

Uncertainty (visitation routines)

- Examining a trajectory as a result of a process

$$X^L$$

Analyzing the probability distribution of joint variables.

$$X^2 = X_1 X_2$$

$$X^3 = X_1 X_2 X_3$$

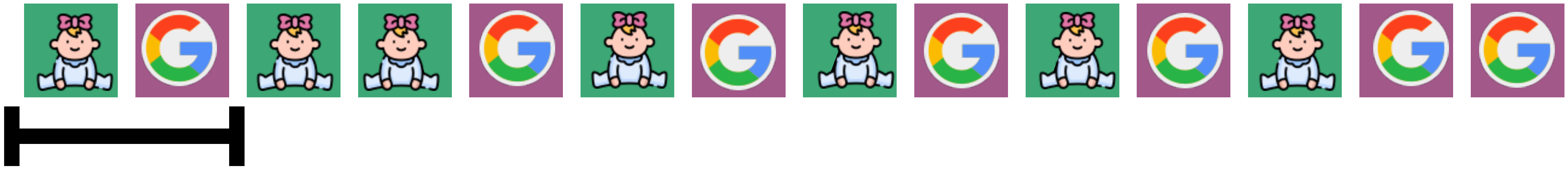
$$X^L = X_1 \dots X_L$$

Uncertainty (visitation routines)

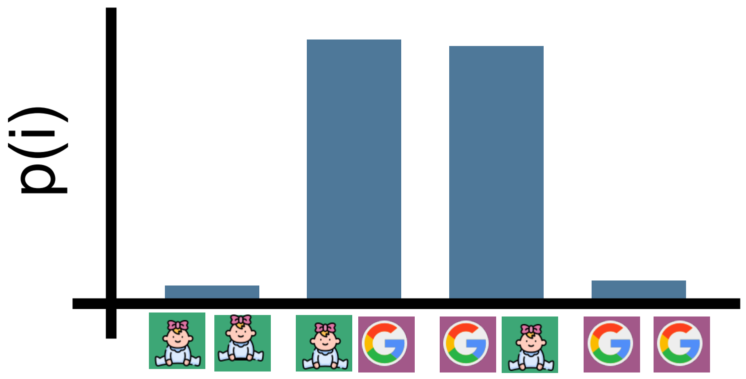
- Examining a trajectory as a result of a process

$$X^L$$

Analyzing the probability distribution of joint variables.



For L = 2

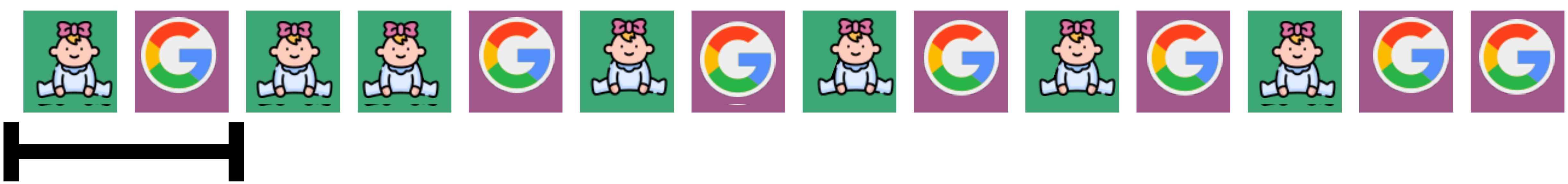


Uncertainty (visitation routines)

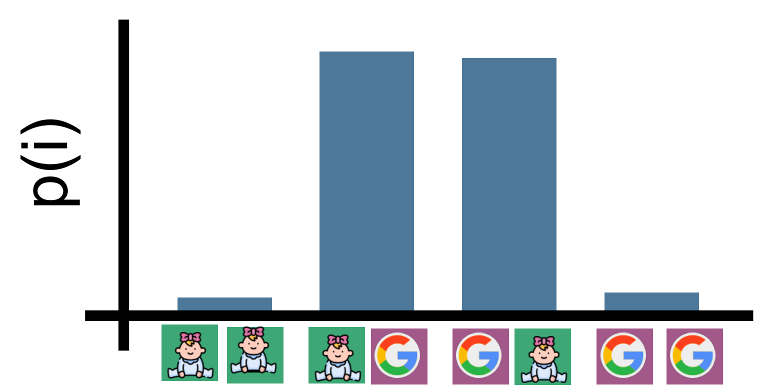
- Examining a trajectory as a result of a process

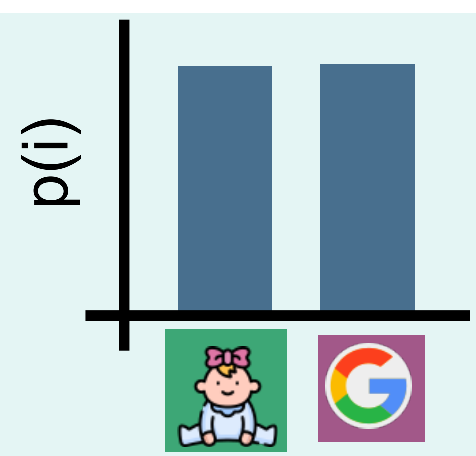
$$X^L$$

Analyzing the probability distribution of joint variables.



For L = 2





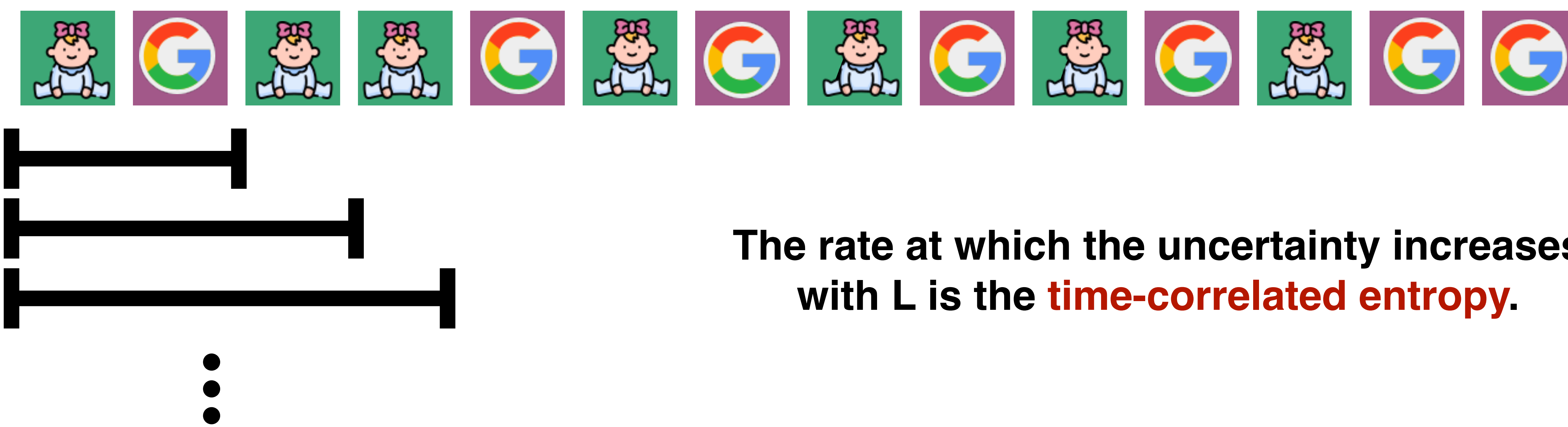
location preferences

Uncertainty (visitation routines)

- Examining a trajectory as a result of a process

$$X^L$$

Analyzing the probability distribution of joint variables.



The rate at which the uncertainty increases with L is the time-correlated entropy.

From uncertainty to (un)predictability

- What is the probability $\prod$ of correctly predicting future locations given a past series of observation?

From uncertainty to (un)predictability

- What is the probability Π of correctly predicting future locations given a past series of observation?

Limits of Predictability in Human Mobility

Chaoming Song,[1,2] Zehui Qu,[1,2,3] Nicholas Blumm,[1,2] Albert-László Barabási[1,2*]

A range of applications, from predicting the spread of human and electronic viruses to city planning and resource management in mobile communications, depend on our ability to foresee the whereabouts and mobility of individuals, raising a fundamental question: To what degree is human behavior predictable? Here we explore the limits of predictability in human dynamics by studying the mobility patterns of anonymized mobile phone users. By measuring the entropy of each individual's trajectory, we find a 93% potential predictability in user mobility across the whole user base. Despite the significant differences in the travel patterns, we find a remarkable lack of variability in predictability, which is largely independent of the distance users cover on a regular basis.

When it comes to the emerging field of human dynamics, there is a fundamental gap between our intuition and the current modeling paradigms. Indeed, although we rarely perceive any of our actions to be random, from the perspective of an outside observer who is unaware of our motivations and schedule, our activity pattern can easily appear random and unpredictable. Therefore, current models of human activity are fundamentally stochastic (1) from Erlang's formula (2) used in telephony to Lévy-walk models describing human mobility (3–7) and their applications in viral dynamics (8–10), queuing models capturing human communication patterns (11–13), and models capturing body balancing (14) or panic (15). Yet the probab eling framewo What is the r havior and to v actions predict

[1]Center for Comple Biology, and Con Boston, MA 02115 Medical School, an Farber Cancer Inst Computer Science Science and Techn

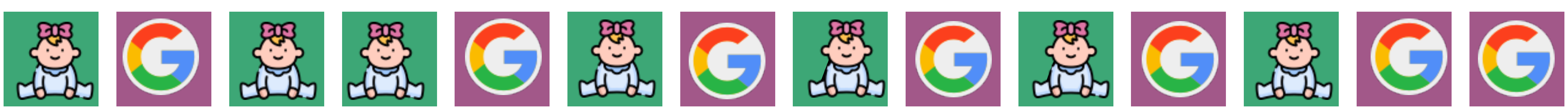
*To whom correspondence should be addressed. E-mail: alb@neu.edu

19 FEBRUARY 2010 VOL 327 **SCIENCE** www.sciencemag.org

They derived an explicit relationship of the upper-limit with the time-correlated entropy.

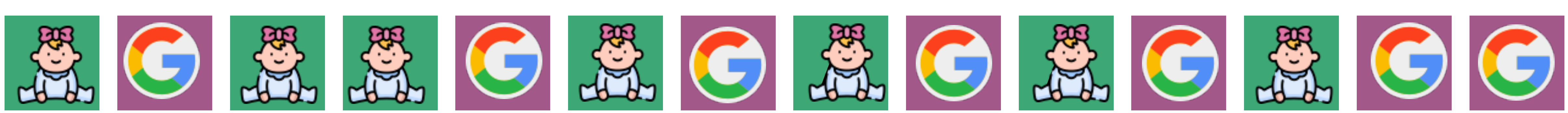
Contextualizing predictability

- Predictability limit which accounts for routines $\Pi^{\max}$



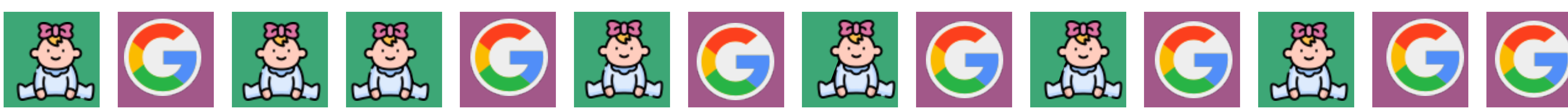
Contextualizing predictability

- Predictability limit which accounts for routines $\Pi^{\max}$
- To understand this value, we create theoretical predictability limits (null models)



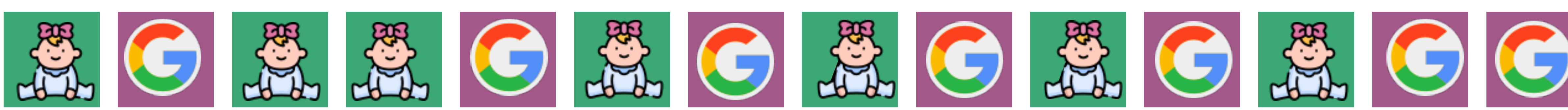
Contextualizing predictability

- Predictability limit which accounts for routines $\Pi^{\max}$
- To understand this value, we create theoretical predictability limits (null models)
  - No preferences: remove all repetition Π^{rand}

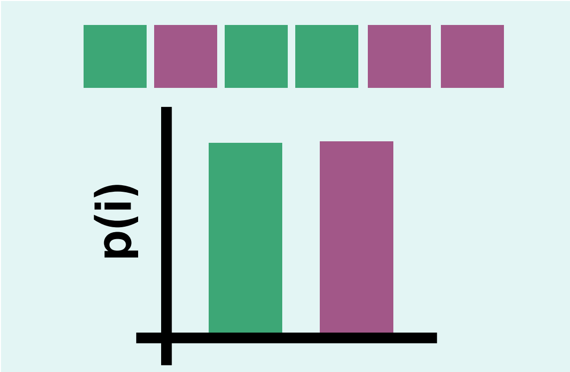


Contextualizing predictability

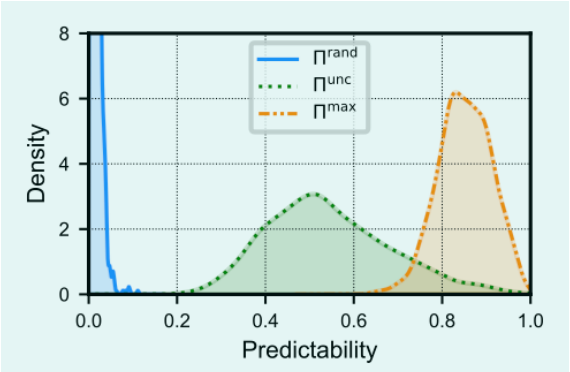
- Predictability limit which accounts for routines $\Pi^{\max}$
- To understand this value, we create theoretical predictability limits (null models)
  - No preferences: remove all repetition Π^{rand}
  - No routines: shuffle it Π^{unc}



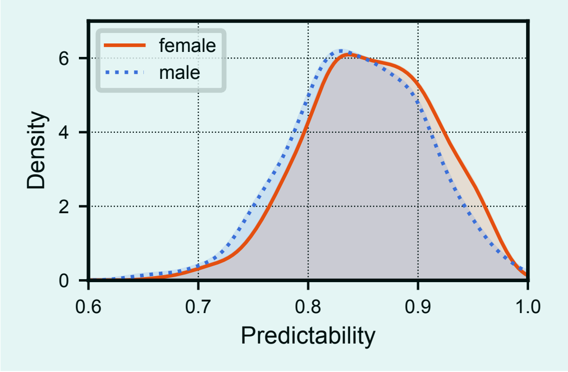
Web routineness & limits of predictability



Predictability measurement framework



Web routineness and predictability

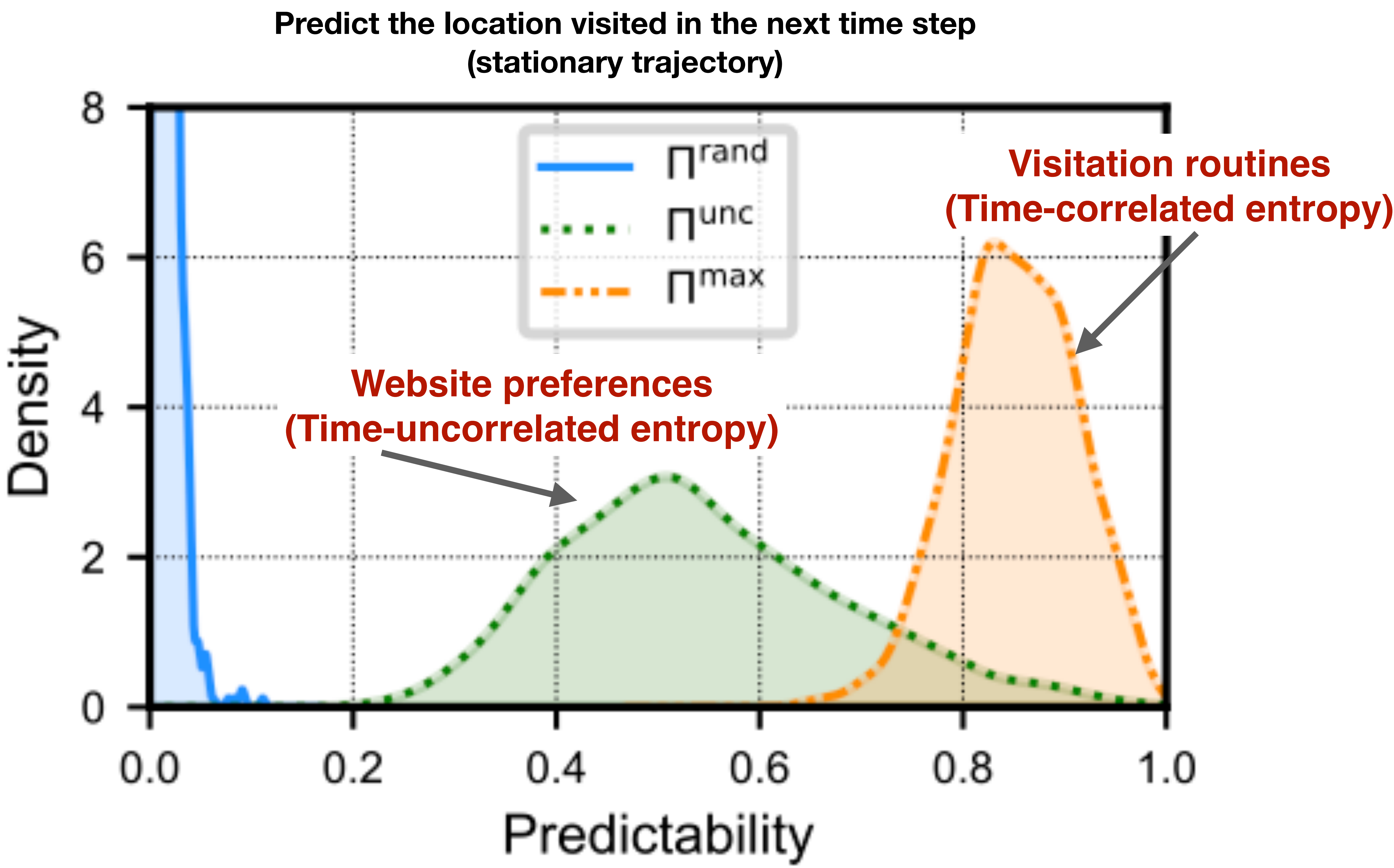


Demographic & behavioral differences

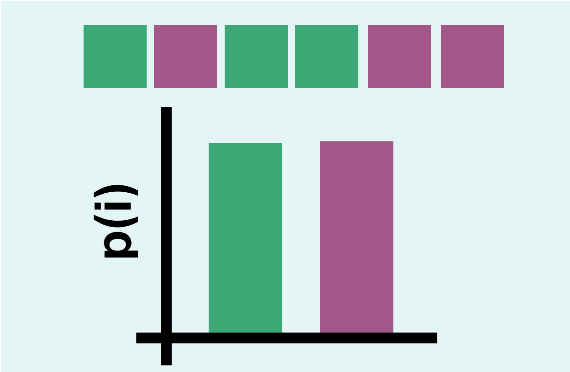
Data

- Sample of German online population:
  - 2,148 individuals.
  - 9,151,243 web visits.
  - 49,918 unique domains.
- Self-reported gender and age.
- Data from a GDPR-compliant digital panel company in Europe.

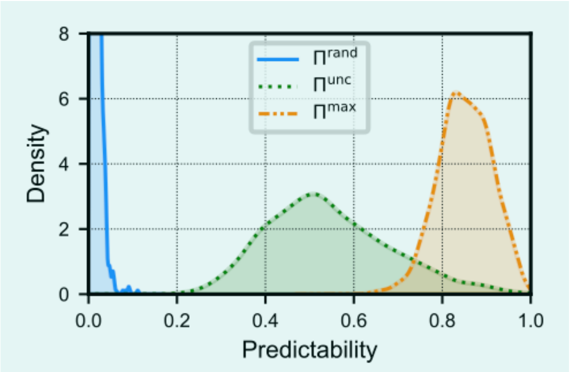
Predictability of web mobility



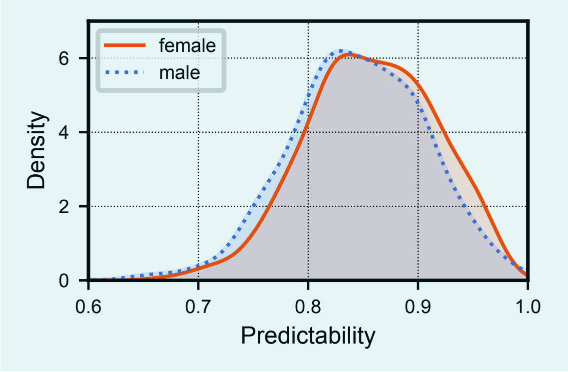
Web routineness & limits of predictability



Predictability measurement framework



Web routineness and predictability



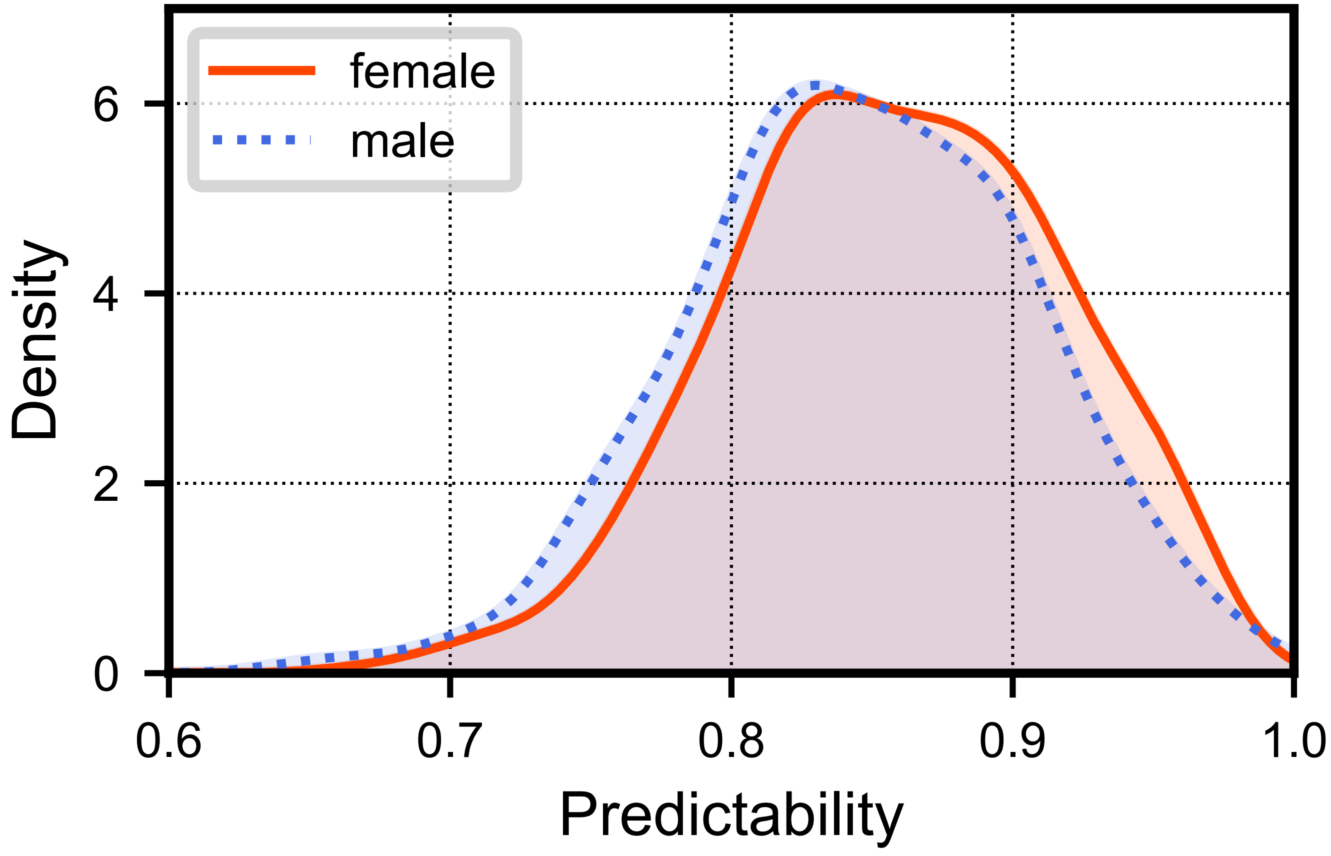
Demographic & behavioral differences

Demographic differences - Gender

- Gender: Male (51.5%) & Female (48.5%)
- Two sample KS test (Hyp: “Two dist are not different”)

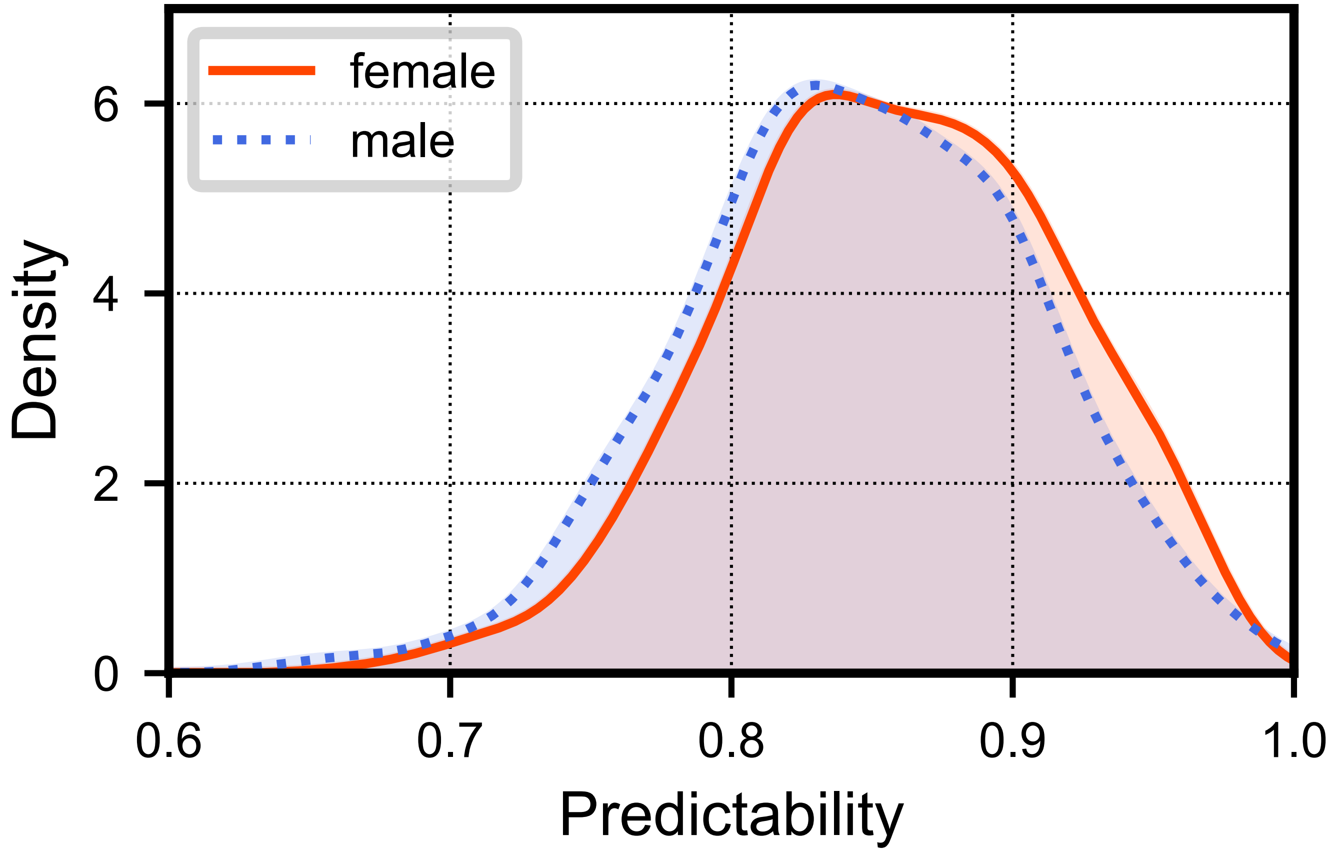
Demographic differences - Gender

- Gender: Male (51.5%) & Female (48.5%)
- Two sample KS test (Hyp: “Two dist are not different”)
  - Rejected hyp. KS Score: 0.086, p-value: 0.008
  - Cliff’s d: 0.11



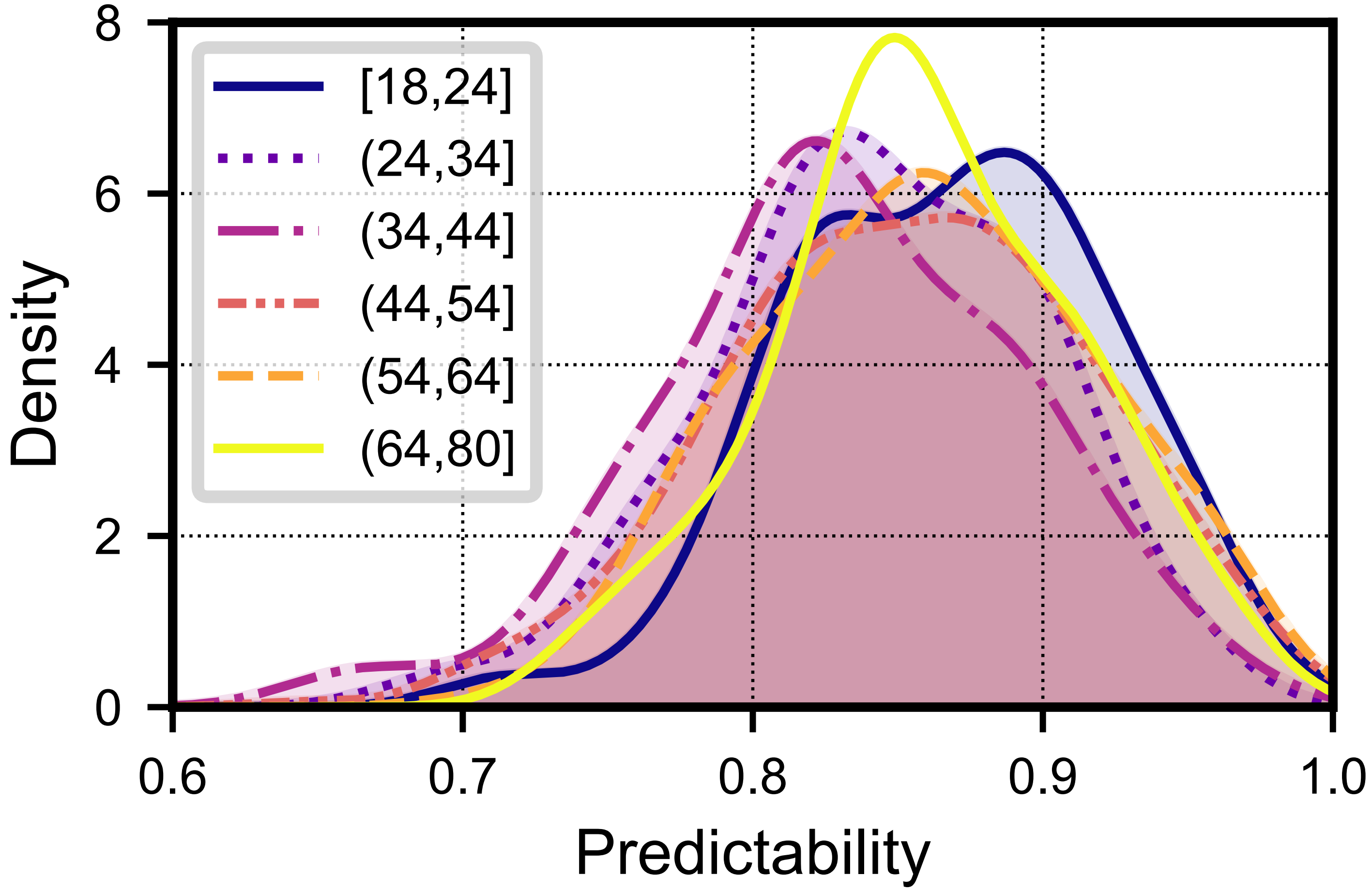
Demographic differences - Gender

- Gender: Male (51.5%) & Female (48.5%)
- Two sample KS test (Hyp: “Two dist are not different”)
  - Rejected hyp. KS Score: 0.086, p-value: 0.008
  - Cliff’s d: 0.11



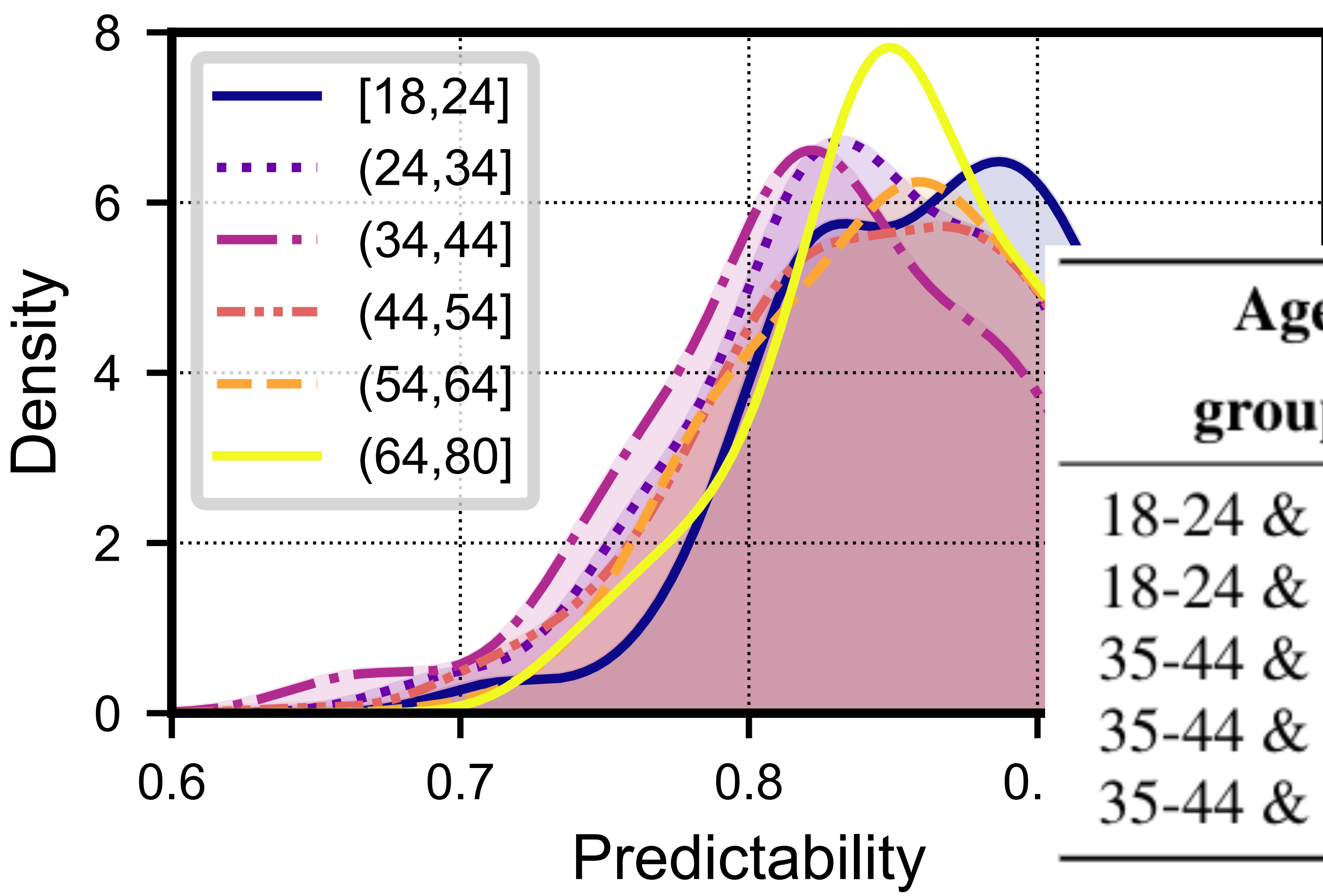
Demographic differences - Age

- Age (18-24, 25-34, 35-44, 45-54, 55-64, 64-80)
- Two sample KS tests applied pairwise



Demographic differences - Age

- Age (18-24, 25-34, 35-44, 45-54, 55-64, 64-80)
- Two sample KS tests applied pairwise



Age groups	KS test KS score	KS test p-value	Cliff's delta
18-24 & 25-34	0.19	8.3×10^{-4}	-0.24
18-24 & 35-44	0.26	1.8×10^{-6}	-0.34
35-44 & 45-54	0.16	6.2×10^{-4}	0.19
35-44 & 55-64	0.21	3.1×10^{-6}	0.25
35-44 & 65-80	0.28	2.1×10^{-4}	0.27

Behavioral differences

- User activity
  - total time spent browsing
  - total visits
- Diversity of user interests
  - Number of unique domains visited
  - Number of unique categories visited
- User stationarity
  - Mean seconds spent per visit
  - Median seconds spent per visit

Behavioral differences

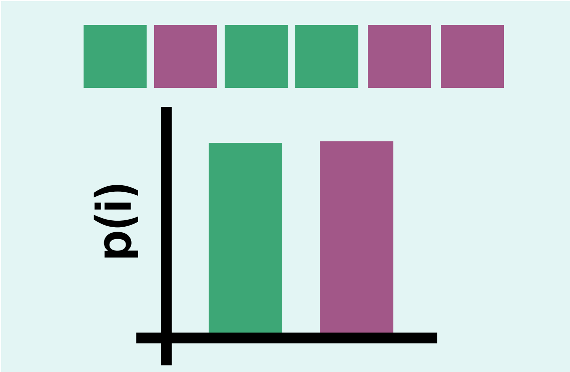
- User activity
  - total time spent browsing
  - total visits
- Diversity of user interests
  - Number of unique domains visited
  - Number of unique categories visited
- User stationarity
  - Mean seconds spent per visit
  - Median seconds spent per visit

Web Routineness and Limits of Predictability

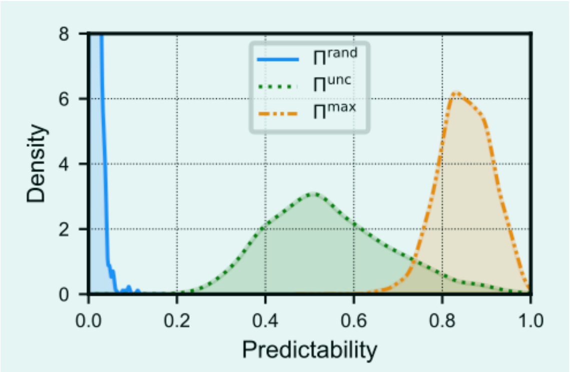
Investigating Demographic and Behavioral Differences Using Web

Juhi Kulshrestha, Marcos Oliveira, Orkut Karaçalik, Denis Bonnay, Claudia Wagner

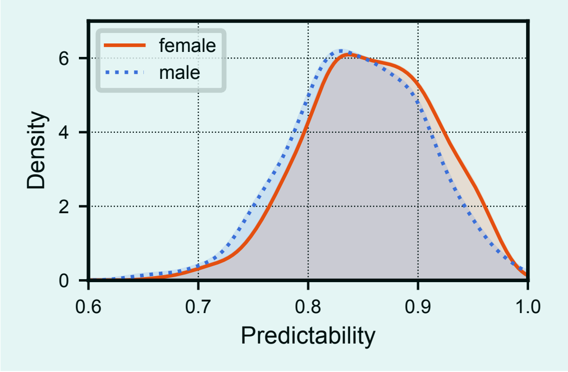
[AAAI ICWSM 2021]



Predictability measurement framework



Web routineness and predictability



Demographic & behavioral differences

Code: https://tinyurl.com/web-tracking-library
Data: https://tinyurl.com/web-tracking-dataset

Human behaviour on the web

[Image credit: cyberbullying.org]

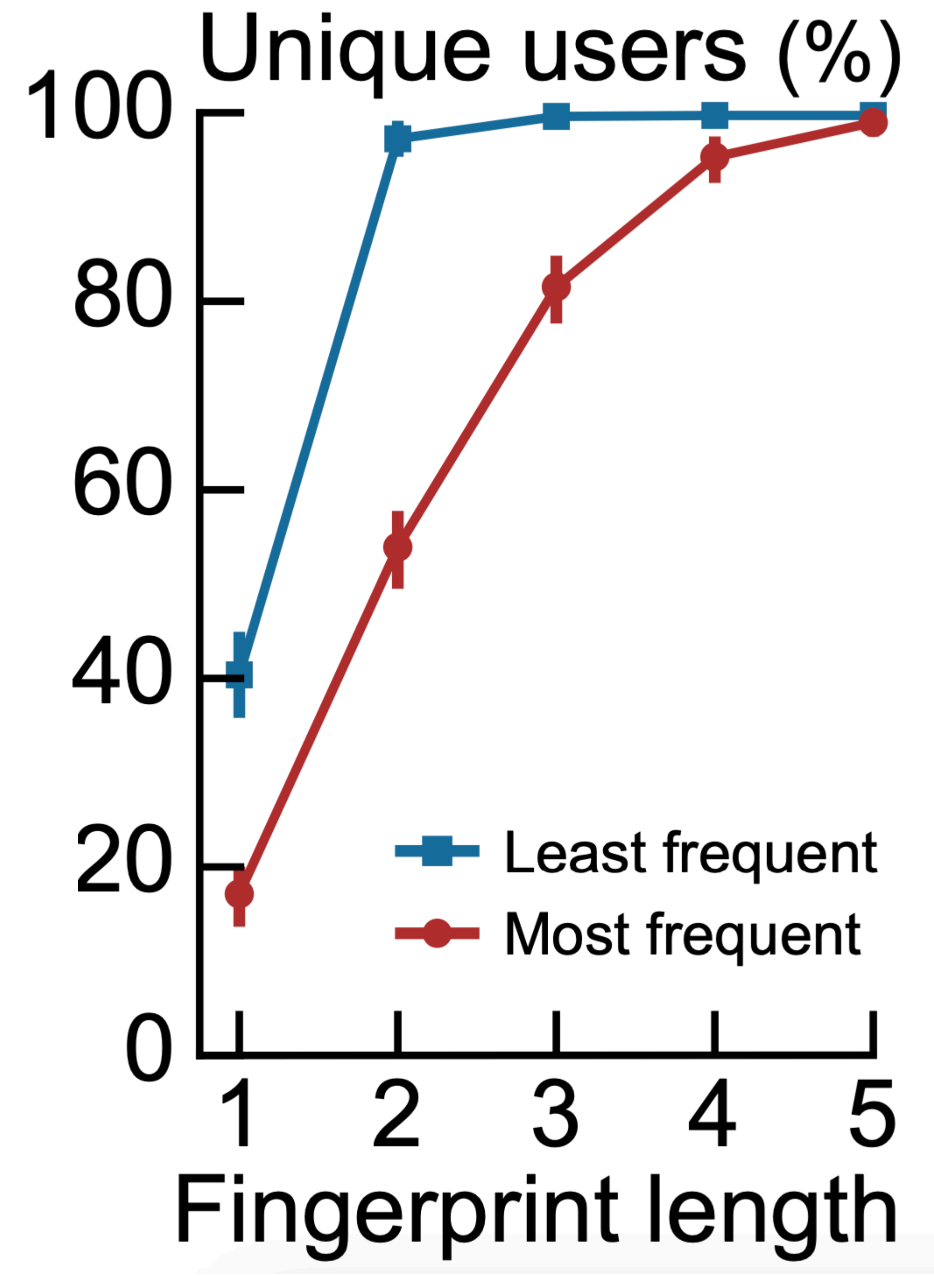
[Under review at PNAS]

Browsing behavior exposes identities on the Web

Marcos Oliveira, Junran Yang, Daniel Griffiths, Denis Bonnay, Juhi Kulshrestha

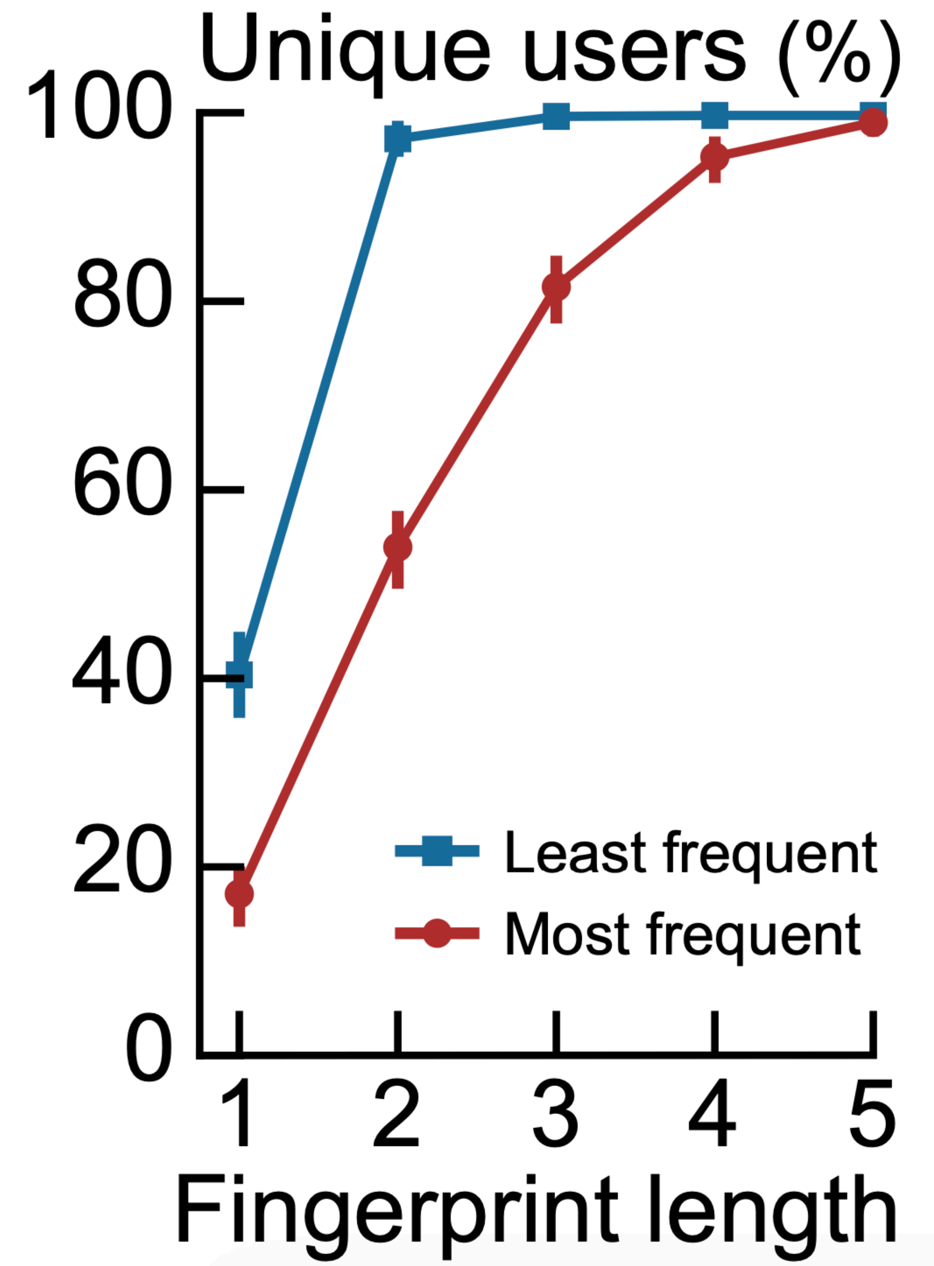
From browsing traces to fingerprints

- Fingerprints: n-tuple comprising their n most visited domains
- Unique users: count non-duplicate n-tuples
- Merely four most visited domains enough to identify 95% of users



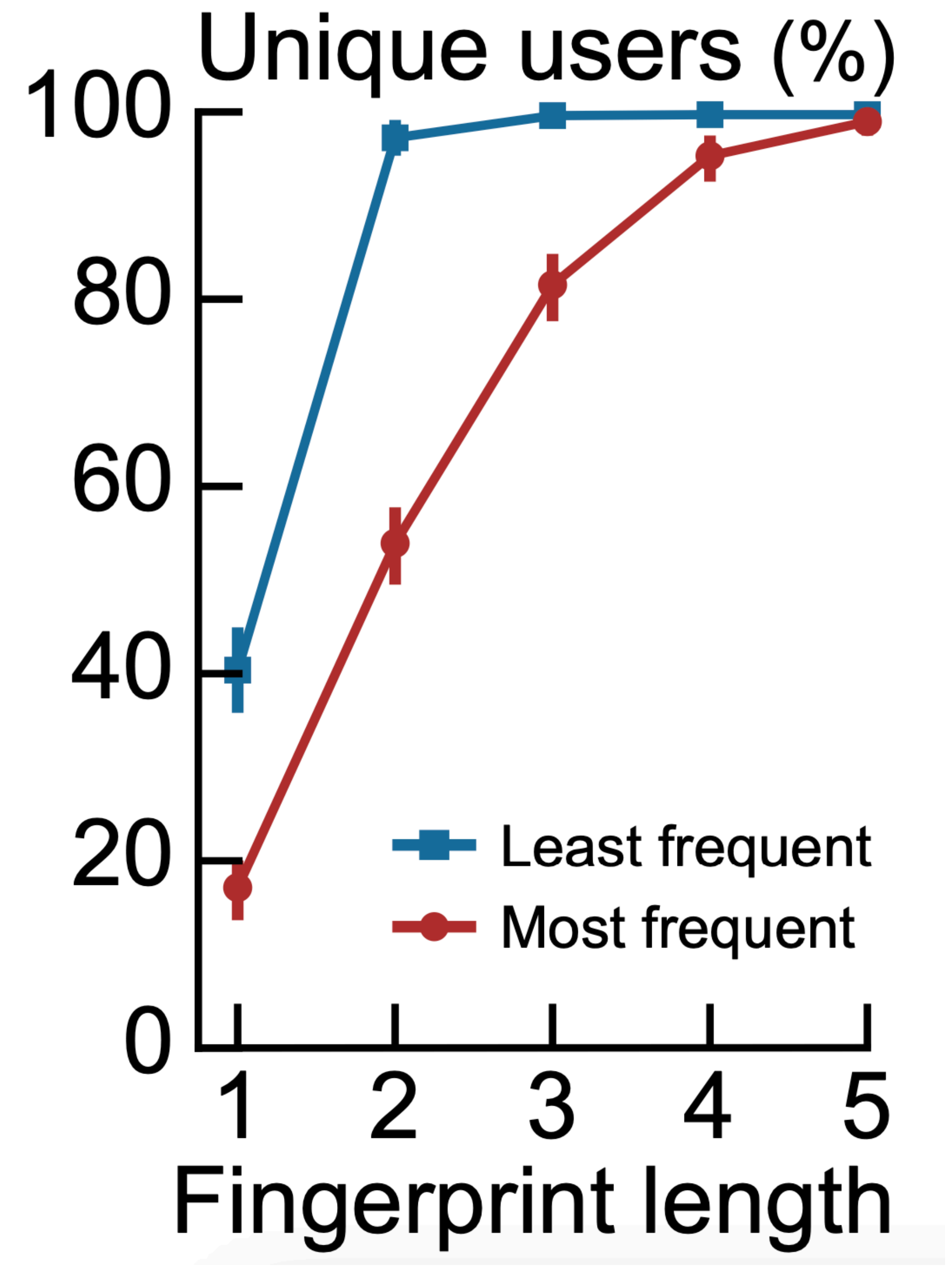
From browsing traces to fingerprints

- Fingerprints: n-tuple comprising their n most visited domains
- Unique users: count non-duplicate n-tuples
- Merely four most visited domains enough to identify 95% of users
- Fingerprints are stable and can be used to re-identify 80% of users in separate time slices of data.



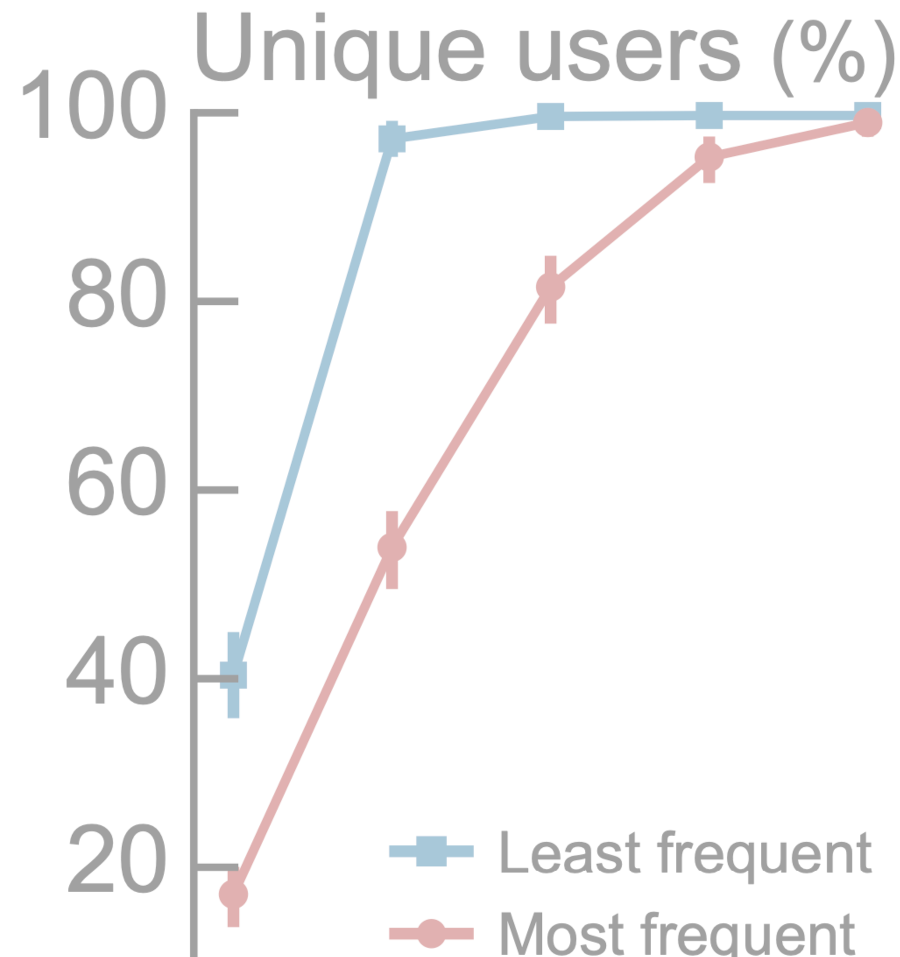
From browsing traces to fingerprints

- Fingerprints: n-tuple comprising their n most visited domains
- Unique users: count non-duplicate n-tuples
- Merely four most visited domains enough to identify 95% of users
- Fingerprints are stable and can be used to re-identify 80% of users in separate time slices of data.
- Privacy threat persists with limited browsing info too



From browsing traces to fingerprints

- Fingerprints: n-tuple comprising their n most visited domains
- Unique users: count non-duplicate n-tuples
- Merely four most visited domains enough to identify 95% of users
- Fingerprints are stable and can be used to re-identify 80% of users in



Fundamental aspect of privacy on the web — even mundane browsing data leads to (re-)identifiability of individuals.

Human behaviour on the web

[Image credit: cyberbullying.org]

Political polarisation & individualised online info environments

Aalto University

gesis Leibniz-Institut für Sozialwissenschaften

LEIBNIZ-INSTITUT FÜR MEDIENFORSCHUNG HANS-BREDOW-INSTITUT

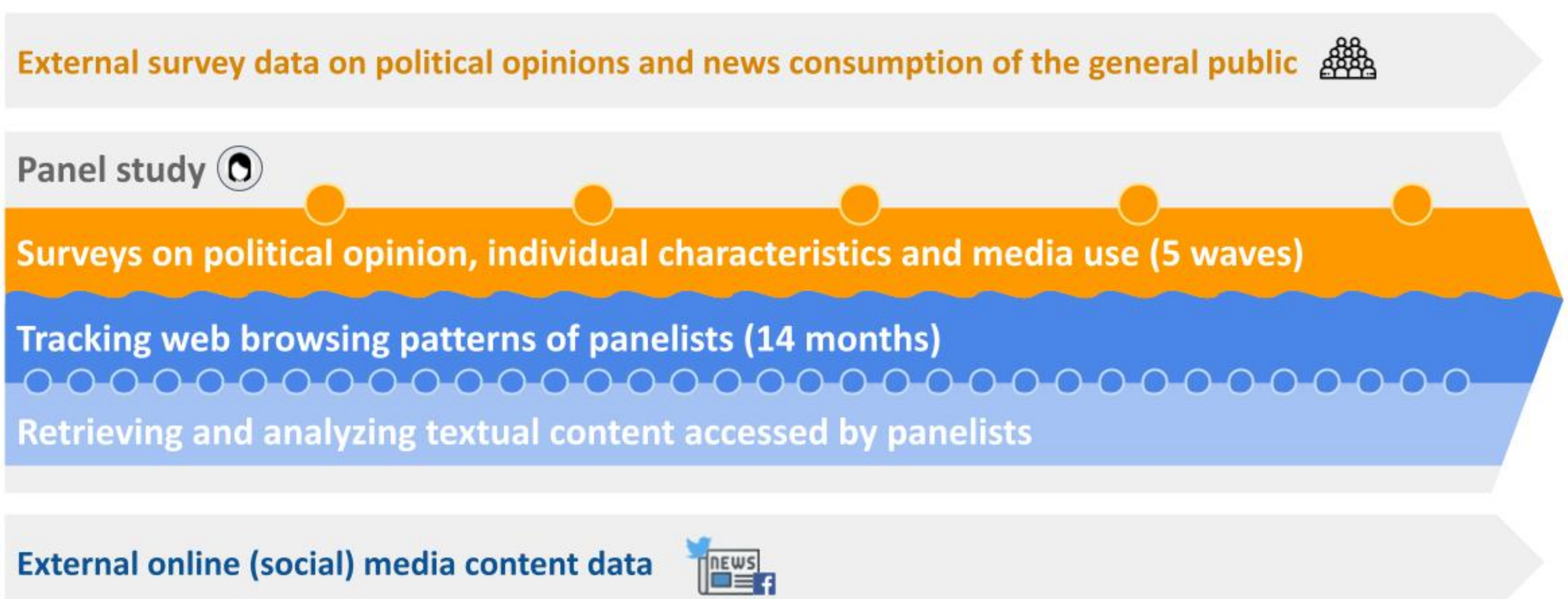
Universität Bremen

Funded by Leibniz Competition

www.juhikulshrestha.com

POLTRACK project

- Panel of a nationally representative sample of German internet users (n = 3,745)
- **Data:** 14 months mobile and desktop tracking + 5 survey waves + automated content analysis
- **Goal**: Study the interplay of information exposure online (diversity) and political opinion formation (polarization)



[Image credit: nbjournal.ru/media-literacy]

Bias & inaccuracies in online news use

[Image credit: nbjournal.ru/media-literacy]

Bias & inaccuracies in online news use

[ACM FAccT 2019]

[IEEE Trans. on Computational Social Systems Journal 2021]

Analyzing biases in perception of truth in news stories and their implications for fact checking

Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Elissa M Redmiles,
Meeyoung Cha, Krishna P Gummadi

Proliferation of mis/dis-information online

How Fake News on Facebook Helped Fuel a Border Crisis in Europe

Facebook, Twitter face congressional hearings on political bias, fake news

Misinformation as an infection: Fighting the spread of disease

Revealed: Facebook hate speech exploded in Myanmar during Rohingya crisis

Capitol riot misinformation persists: False claims continue to circulate on Facebook

Detecting "fake news"

- Experts
  - expert human fact checkers
  - highly resource-constrained process
- Algorithms
  - use content and crowd signals
  - hard to apply in dynamic news contexts without human supervision

Detecting "fake news"

- Experts
  - expert human fact checkers
  - highly resource-constrained process
- Algorithms
  - use content and crowd signals
  - hard to apply in dynamic news contexts without human supervision
- Crowds + Experts

Most prevalent approach

  - ask users to report news they perceive to be fake
  - stories reported by many users prioritized for expert fact checking

Detecting "fake news"

- Experts
  - expert human fact checkers
  - highly resource-constrained process

Is this the best strategy?

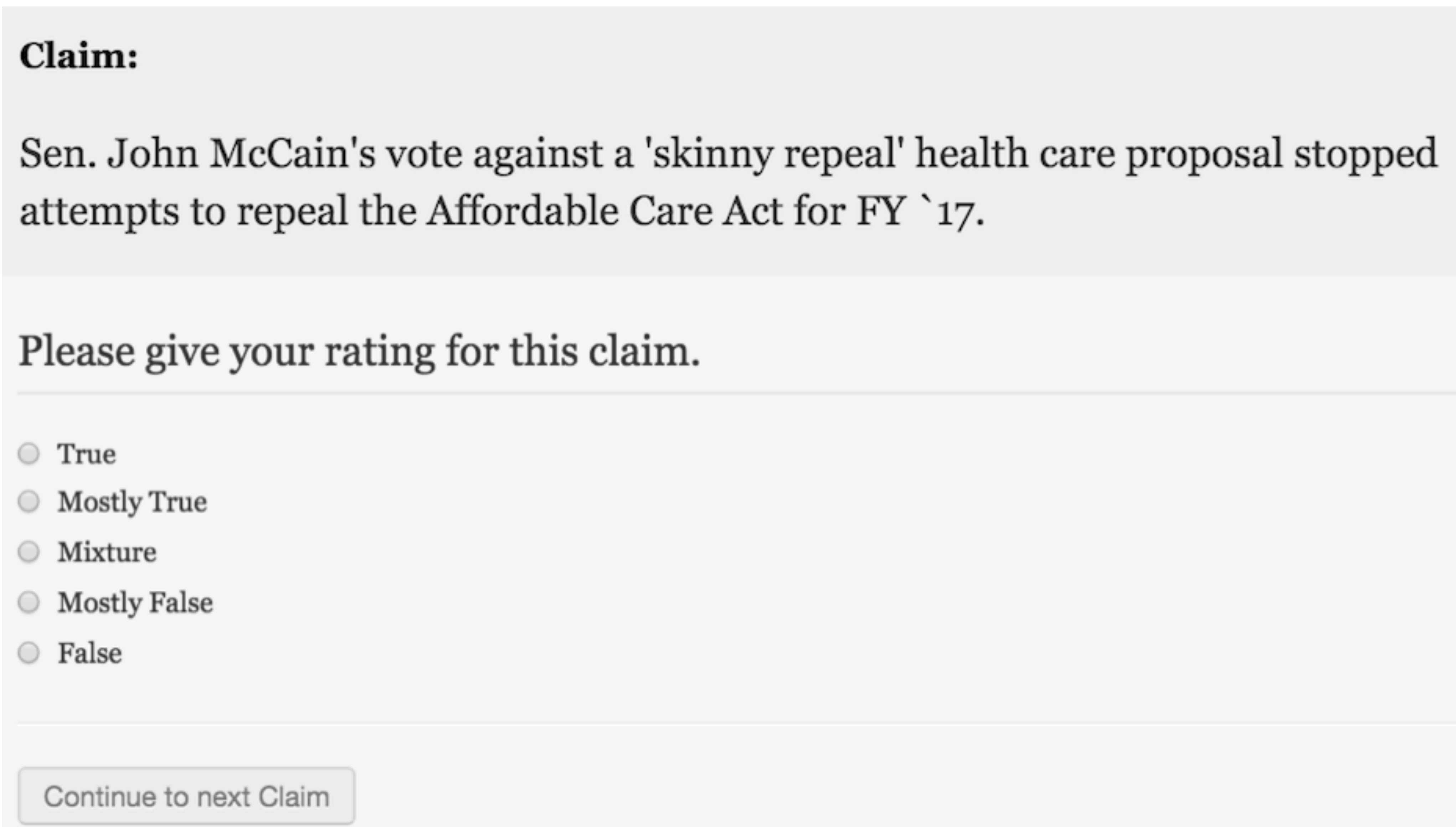
- Crowds + Experts

Most prevalent approach

  - ask users to report news they perceive to be fake
  - stories reported by many users prioritized for expert fact checking

Gathering users' truth perceptions

- rapidly assess of how users implicitly perceive truth in news stories



Claim:

Sen. John McCain's vote against a 'skinny repeal' health care proposal stopped attempts to repeal the Affordable Care Act for FY `17.

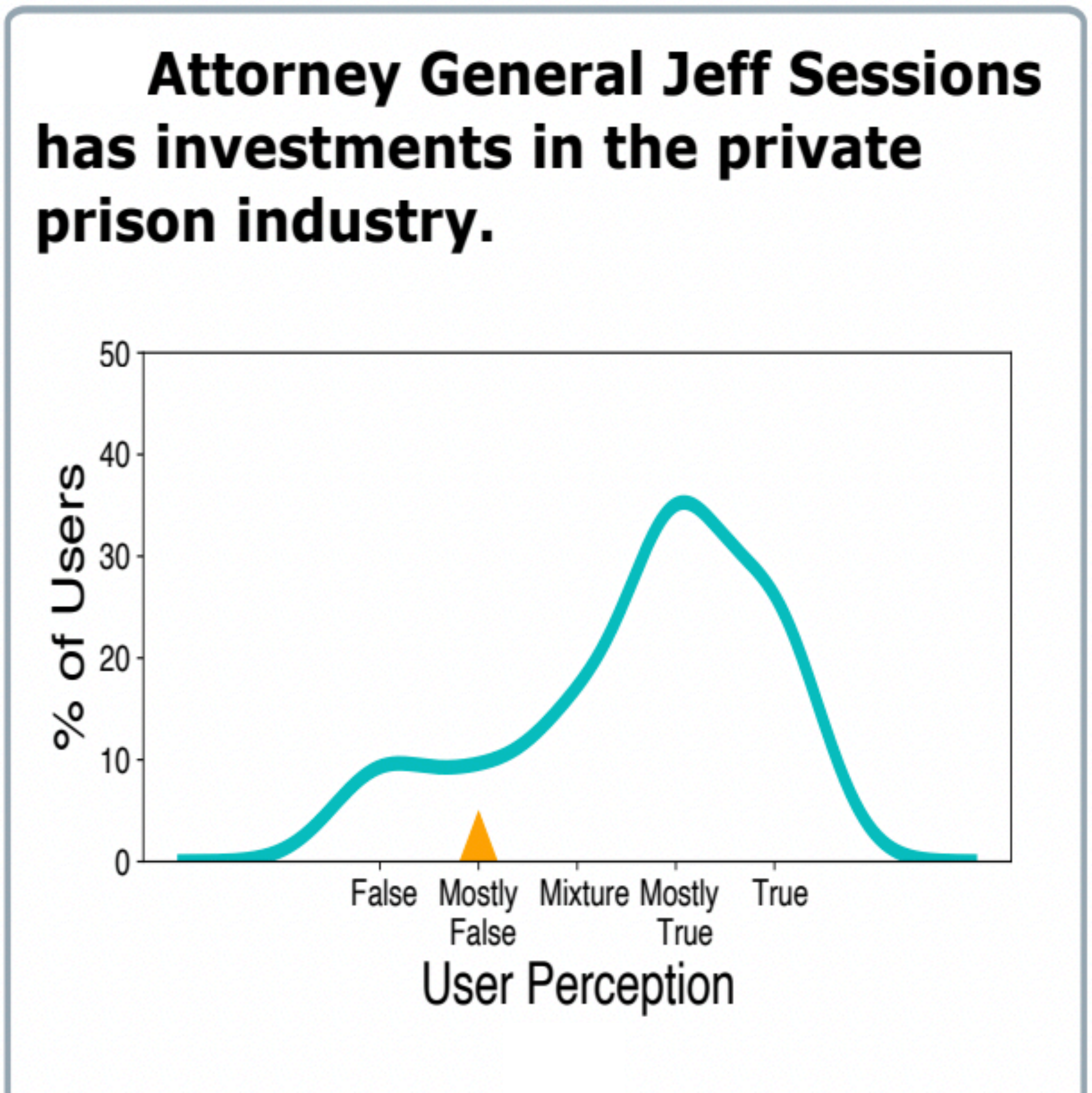
Please give your rating for this claim.

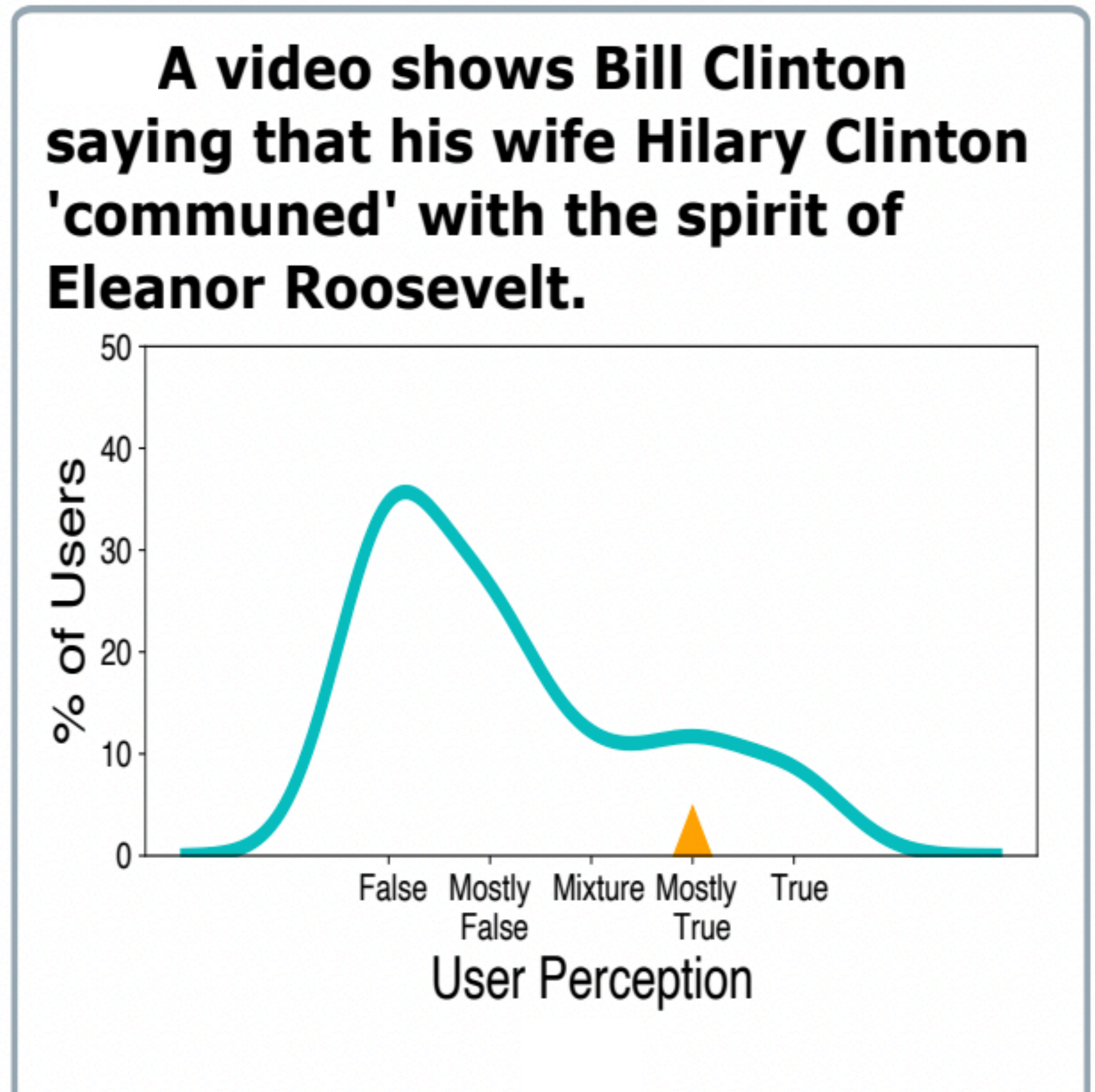
- True
- Mostly True
- Mixture
- Mostly False
- False

Continue to next Claim

Analyzing perception biases (1)

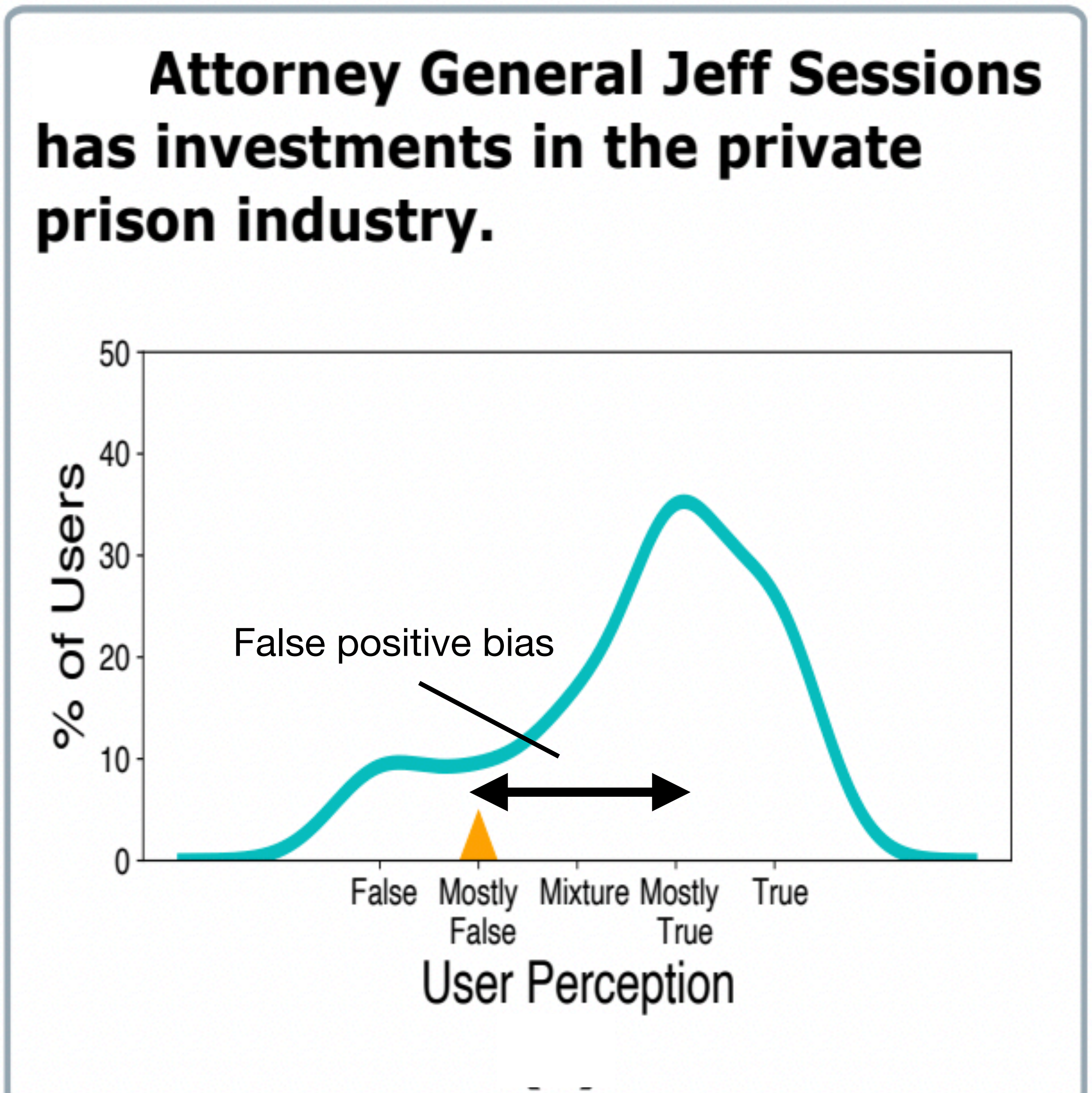
- For many stories, perceived truth level differs significantly from ground truth level

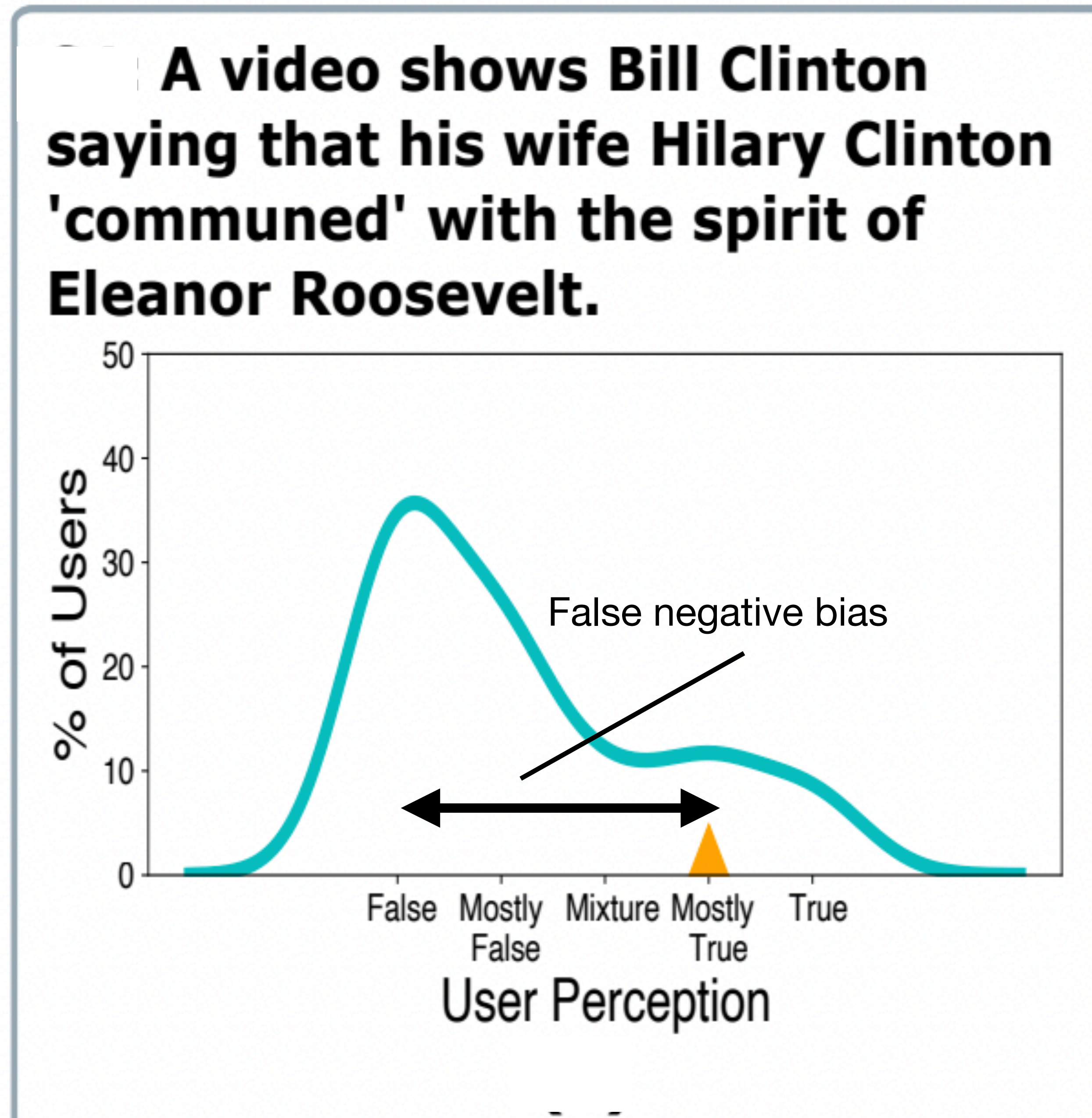




Analyzing perception biases (2)

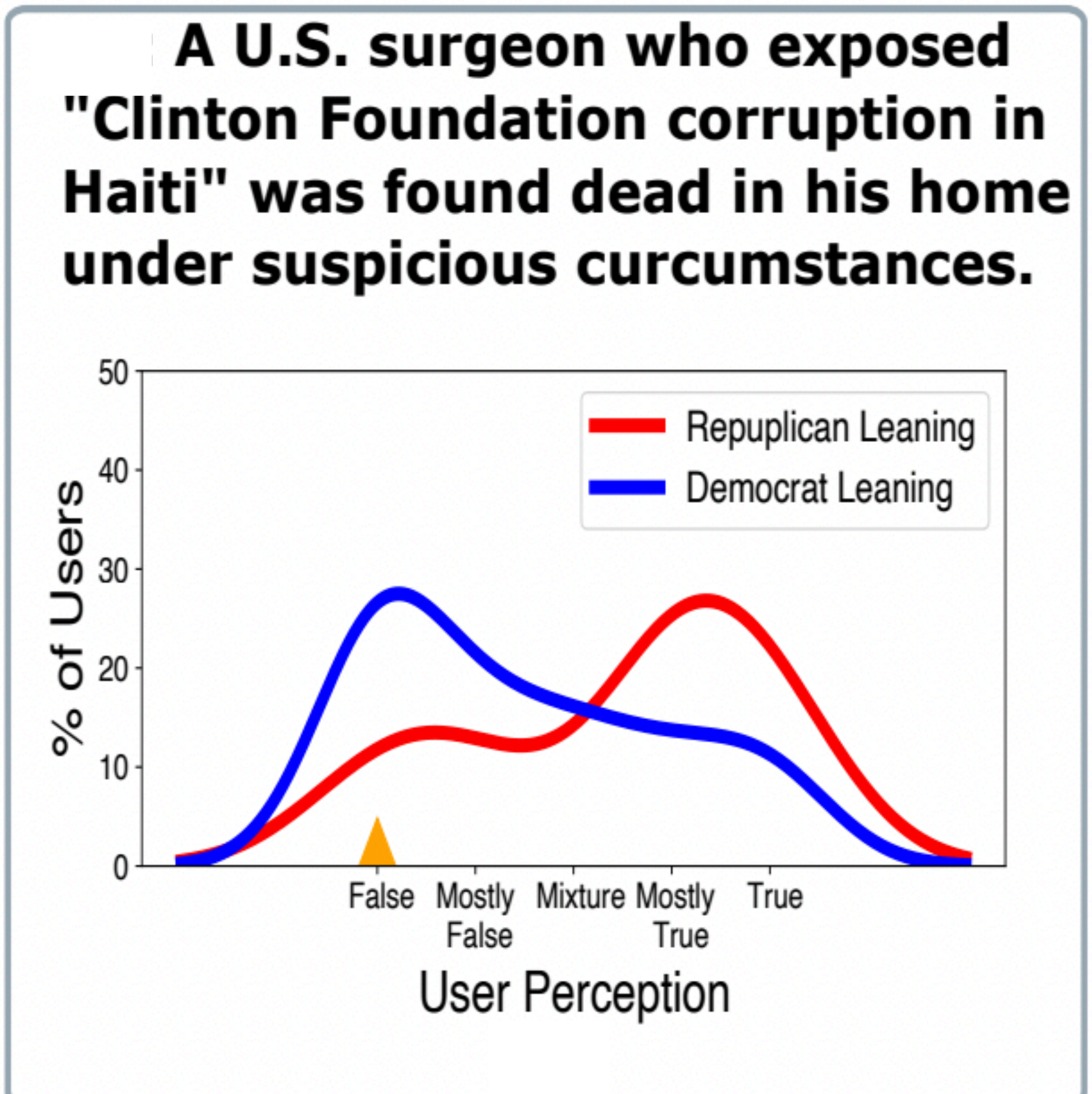
- False-positive and false negative perception bias

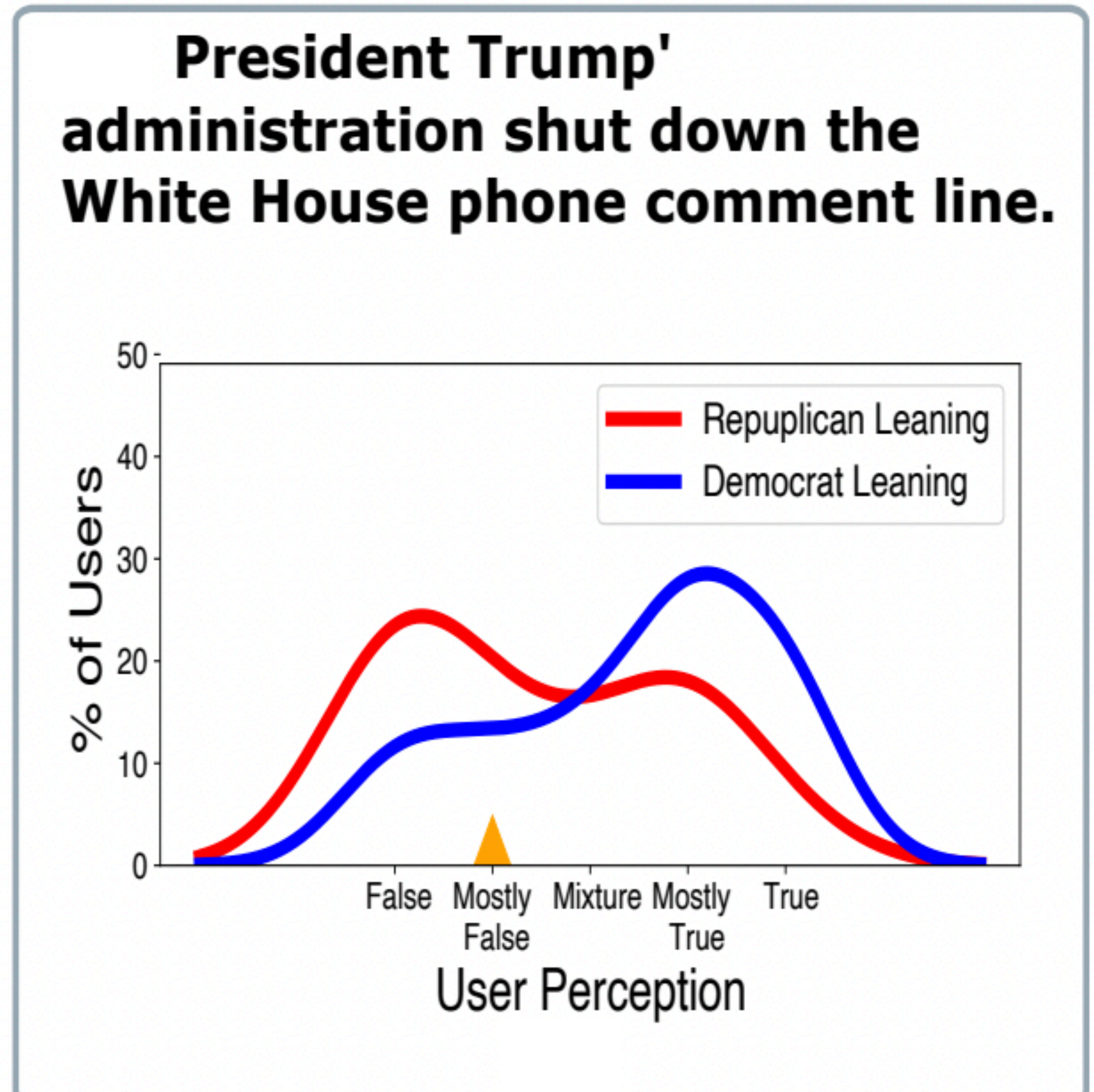




Analyzing perception biases (3)

- Users' political ideologies influence their truth perceptions for the most controversial stories





Operationalizing prioritization objectives

- Remove false news stories from circulation
  - Fact check false stories with higher probability than true stories

Operationalizing prioritization objectives

- Remove false news stories from circulation
  - Fact check false stories with higher probability than true stories
- Correct the misperception of users
  - Fact check stories where users' perceived truth levels differ a lot from ground truth levels

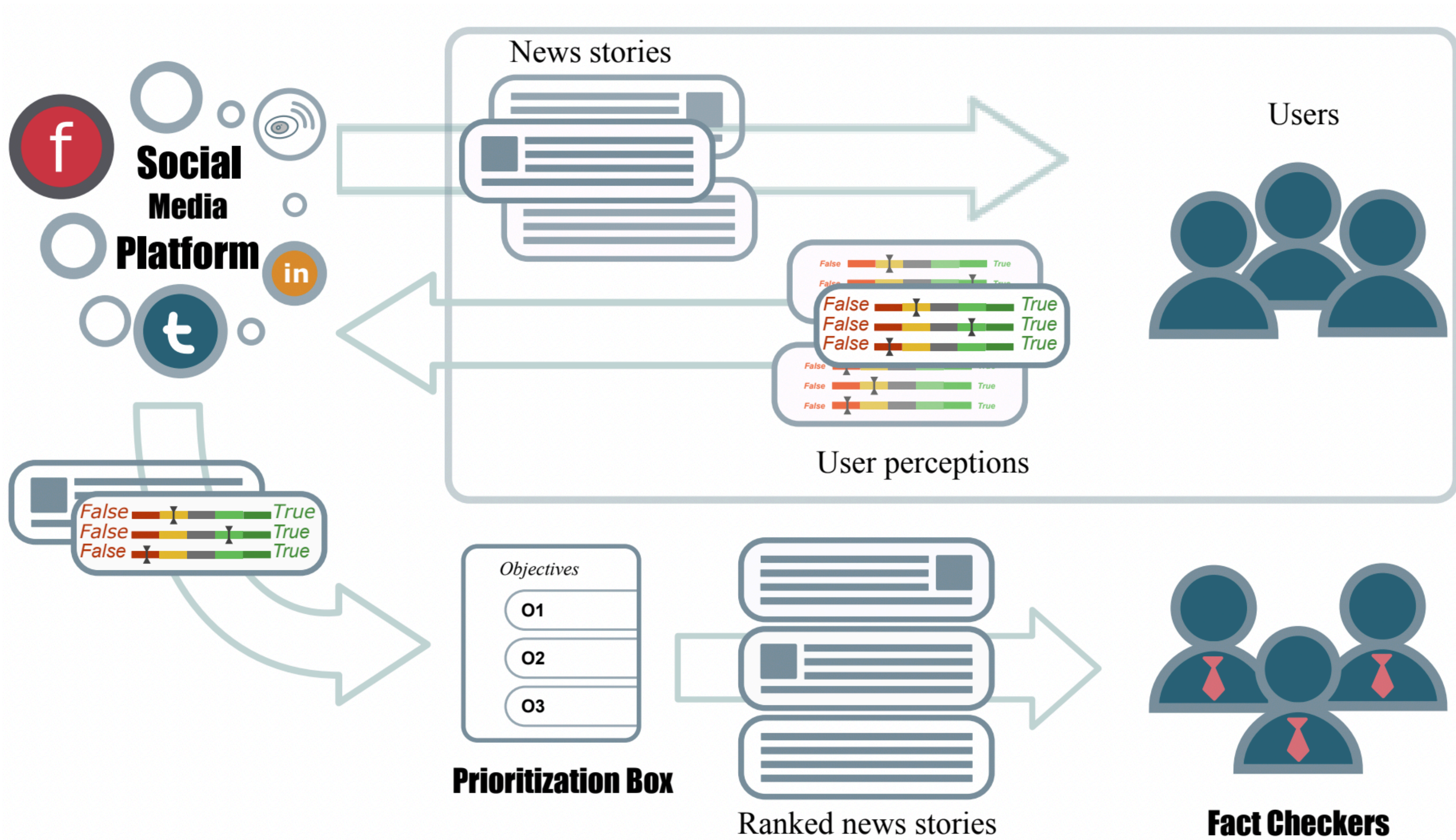
Operationalizing prioritization objectives

- Remove false news stories from circulation
  - Fact check false stories with higher probability than true stories
- Correct the misperception of users
  - Fact check stories where users' perceived truth levels differ a lot from ground truth levels
- Decrease the disagreement among the users
  - Fact check news stories where people disagree most about the truth value of the stories

[IEEE Trans. on Computational Social Systems Journal 2021] [ACM FAccT 2019]

Analyzing Biases in Perception of Truth in News Stories and Their Implications for Fact Checking

Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty,
Elissa M. Redmiles, Meeyoung Cha, and Krishna P. Gummadi



[Image credit: nbjournal.ru/media-literacy]

Bias & inaccuracies in online news use

[AAAI ICWSM 2018]

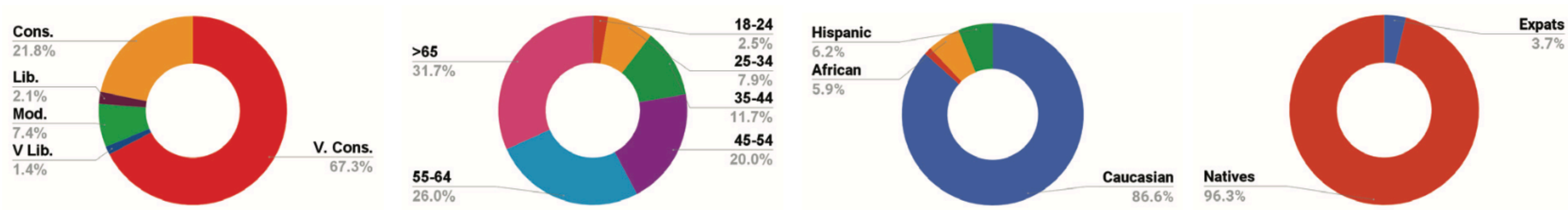
Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale

Filipe N. Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, Krishna P. Gummadi

Using FB Ads platform to assess media bias at scale

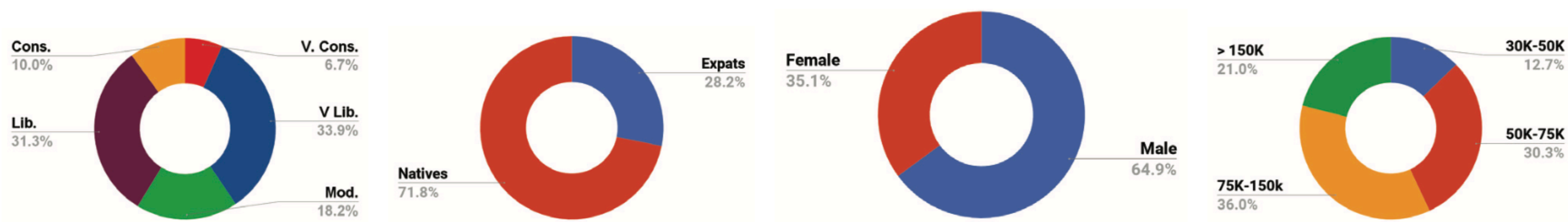
- Social media ad interfaces offer detailed insights into demographics of news sources’ audiences
- Ideological leaning accurately estimated by the over-/under-representation of liberals or conservatives in audience
- Biases in sources’ audience demographics (race, gender, age, national identity, income) indicative of fine-grained biases such as social vs. economic vs. nationalistic conservatism

Using FB Ads platform to assess media bias at scale



(a) Political Leaning (b) Age Groups (c) Racial Affinities (d) National Identity

Breitbart — Mostly conservative, older, caucasian, non-immigrant folks



(a) Political Leaning (b) National Identity (c) Gender (d) Income Level

Economist — Mostly liberal, high-earning, mostly men, who are natives and expats

[Image credit: nbjournal.ru/media-literacy]

Bias & inaccuracies in online news use

[PLOS ONE 2022]

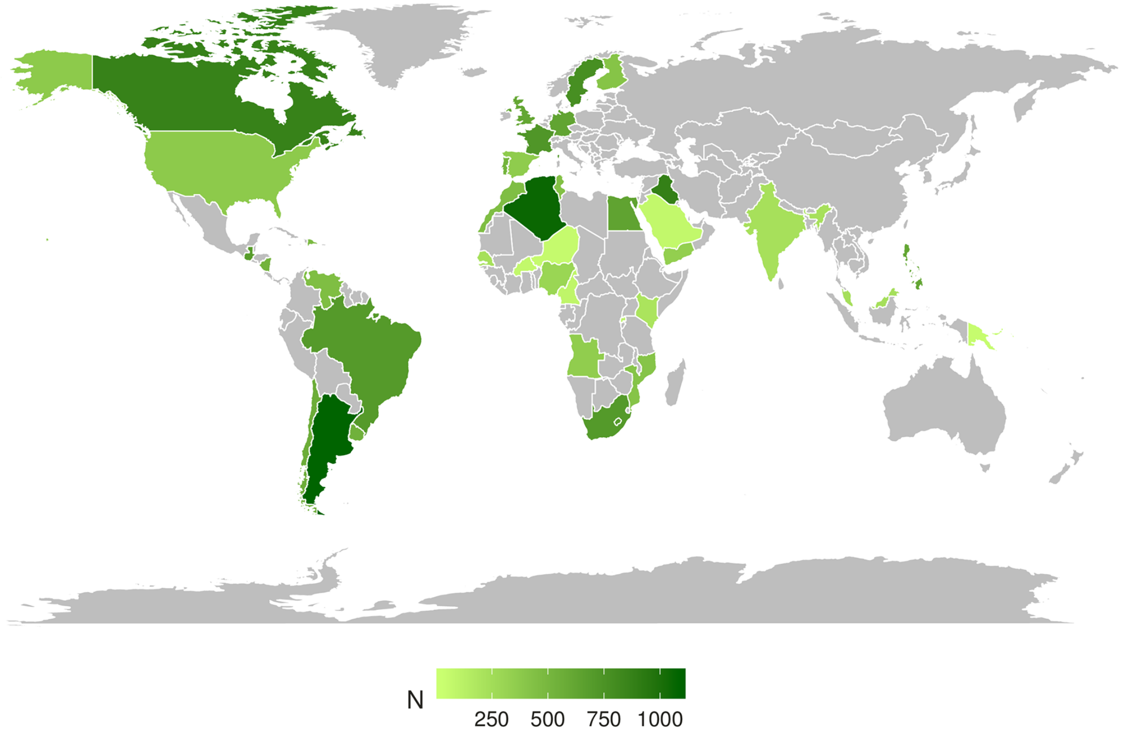
[JMIR Human Factors 2021]

Misinformation, fact checks, believability, and vaccine acceptance over 40 countries during COVID-19

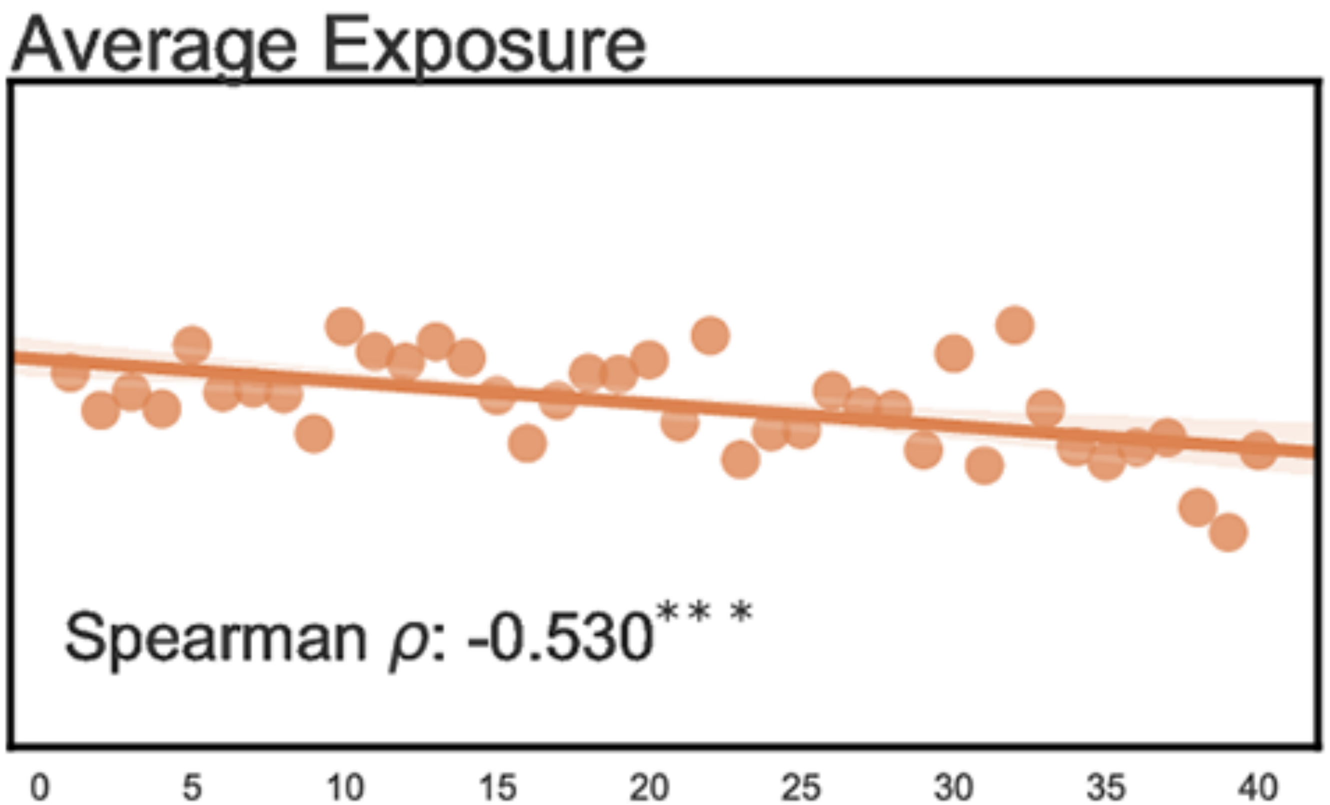
Karandeep Singh, Gabriel Lima, Meeyoung Cha, Chiyoung Cha, Juhi Kulshrestha, Yong-Yeol Ahn, Onur Varol

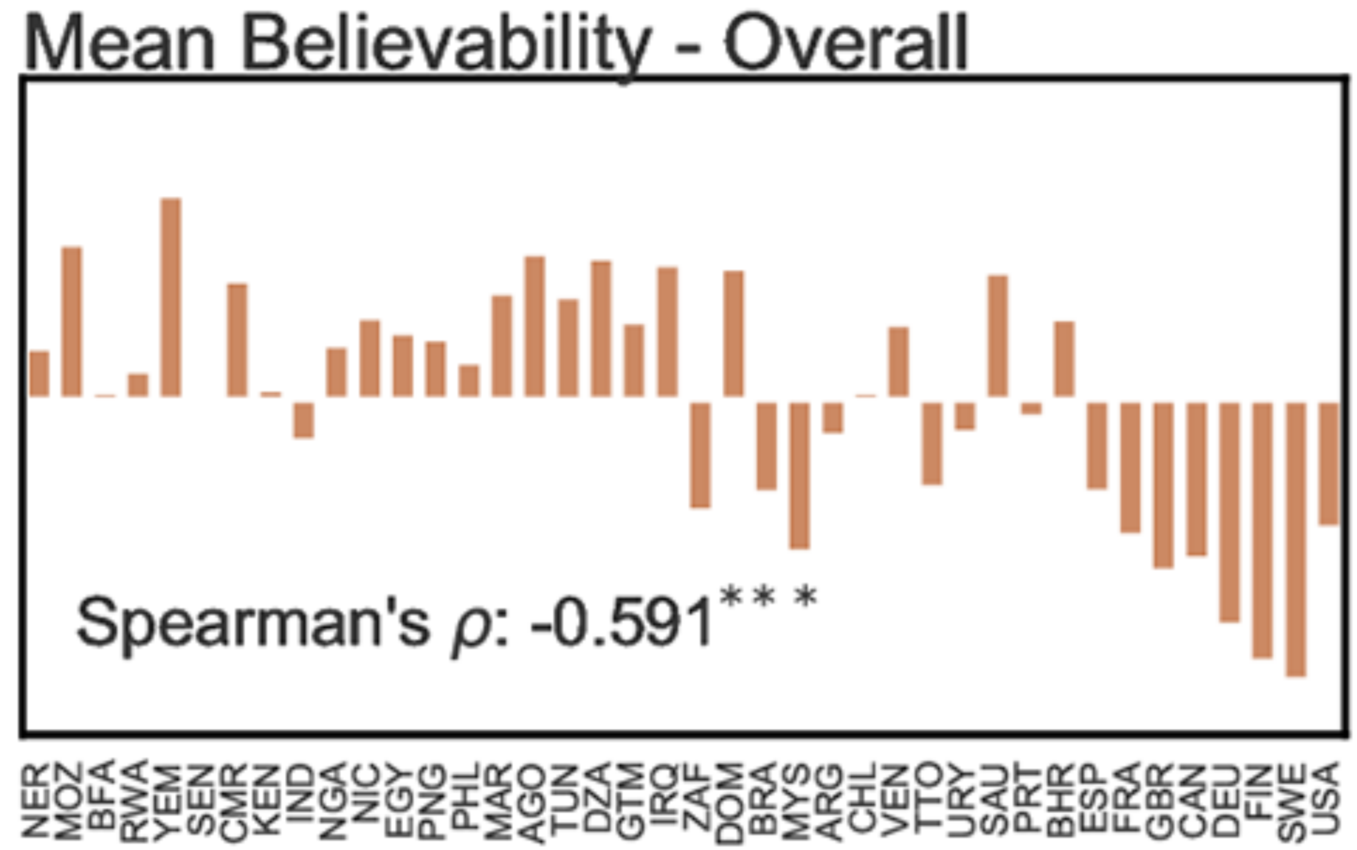
FB Ads platform for recruiting survey respondents

- 18,000 respondents from 40 countries, recruited via FB ads platform
- About five times cheaper than alternatives (e.g., Prolific, MTurk)
- Wider demographic reach
- Whether respondents have been exposed to a broad set of false claims and fact-checked information on the disease, whether they believe the false claims, and if they think their community will benefit from the fact checked information



Underdeveloped countries more susceptible to infodemic

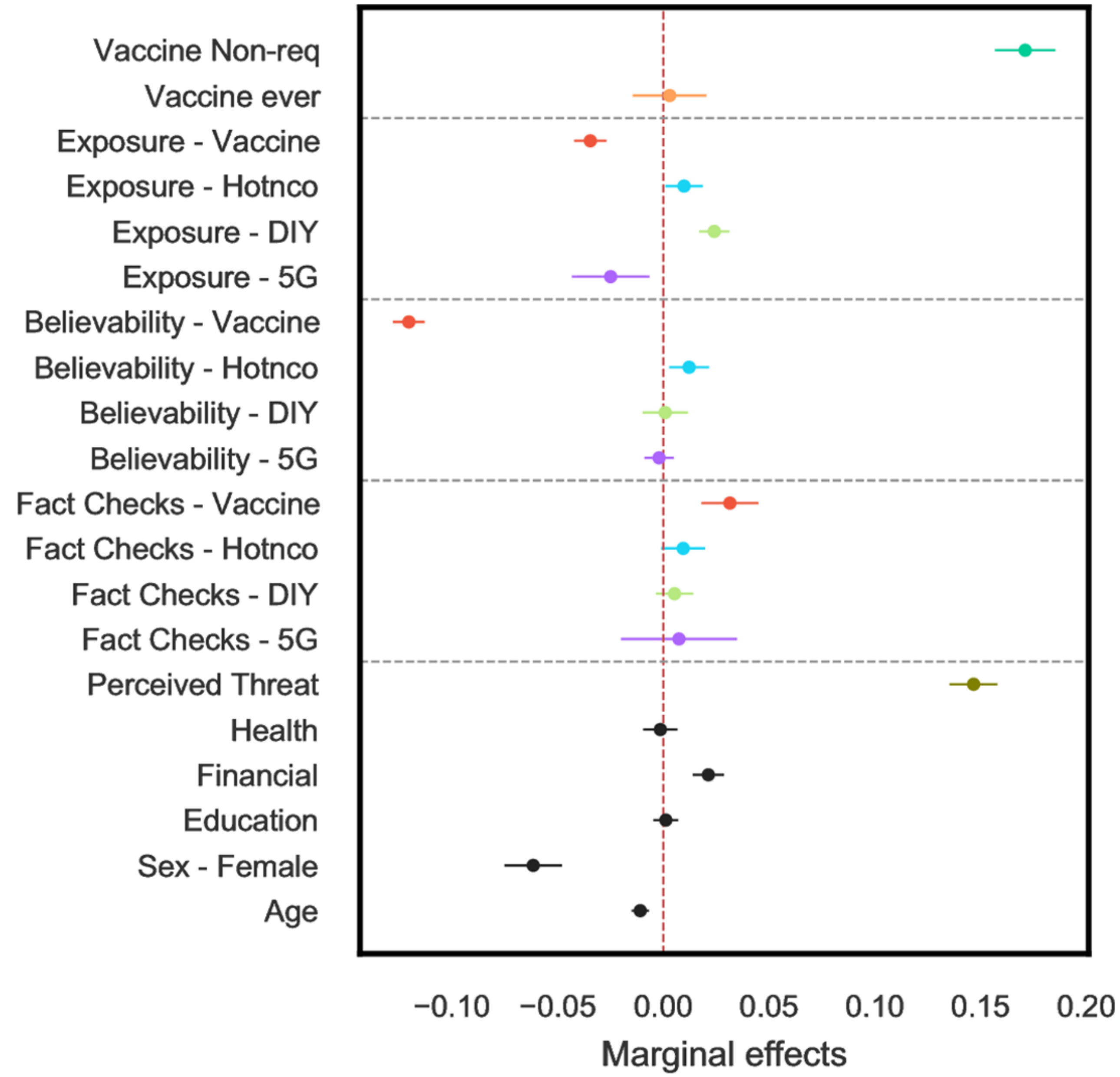




- Countries ranked by GDP per capita on x-axis
- Poorer countries exposed to more false claims than richer countries
- Respondents from lower GDP per capita countries more susceptible to believing in misinformation upon exposure

Infodemic and vaccine acceptance

- Logistic regression model for vaccine acceptance as the outcome variable
- Increased exposure to vaccine-related misinformation is associated with increased vaccine hesitancy
- Believability of vaccination-related claims even more substantially associated with increased vaccine hesitancy
- Increased exposure to fact-checked vaccine-related info is correlated with increased vaccine acceptance



[Image credit: digicol.de/news]

Impact of online info retrieval systems

[Image credit: digicol.de/news]

Impact of online info retrieval systems

[ACM CSCW 2017]

[Information Retrieval Journal 2019]

Quantifying Search Bias

Investigating Political Bias in Social Media and Web Search

Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar,
Saptarshi Ghosh, Krishna Gummadi, Karrie Karahalios

Growing concerns about social media

Scandal, outrage and politics

Do social media threaten democracy?

Donald Trump Accuses Google of Bias in Search Engine Results

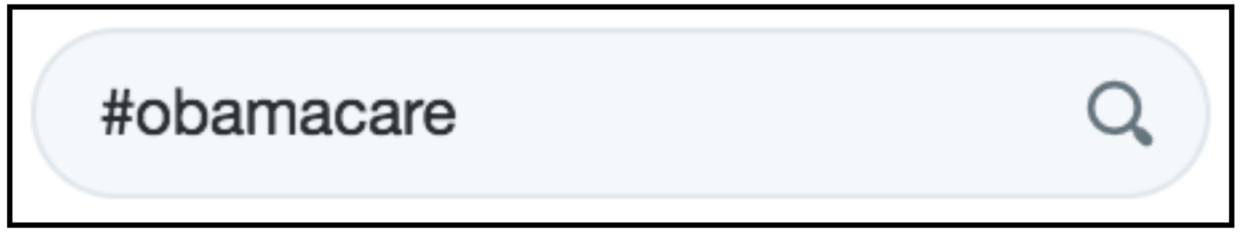
Search engine bias: What search results are telling you (and what they're not)

Trump accuses Facebook, Twitter, Google of Democrat bias

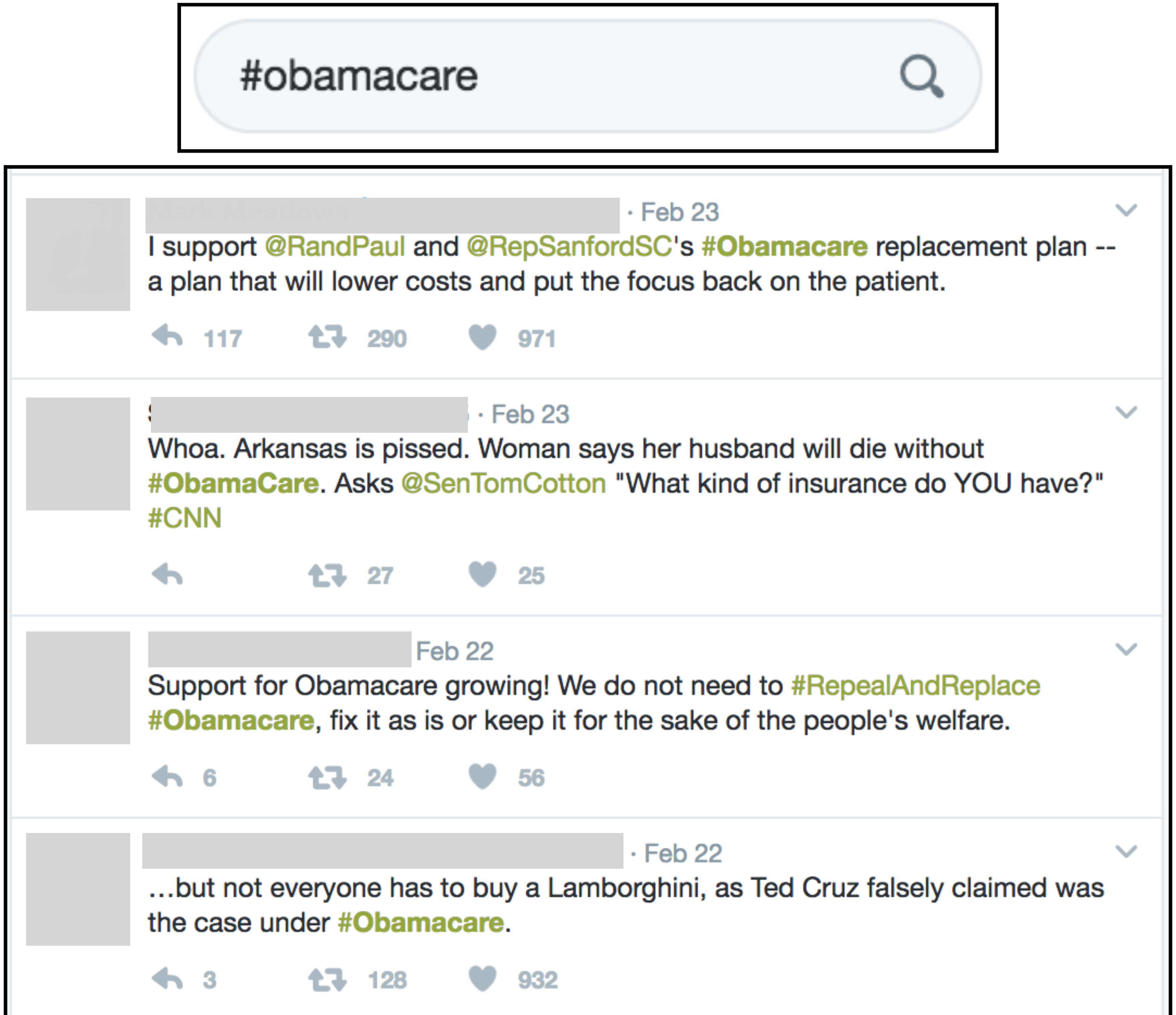
Social media as a "search" platform

- Used to find information about on-going events & public figures *[Teevan et al., 2011]*

Social media as "search" platform



Social media as "search" platform



Ranked list (according to importance)

Potential bias in search results

#obamacare

Follow

I support @RandPaul and @RepSanfordSC's #Obamacare replacement plan -- a plan that will lower costs and put the focus back on the patient.

Republican

Potential bias in search results

#obamacare

Follow

I support @RandPaul and @RepSanfordSC's #Obamacare replacement plan -- a plan that will lower costs and put the focus back on the patient.

Republican

Follow

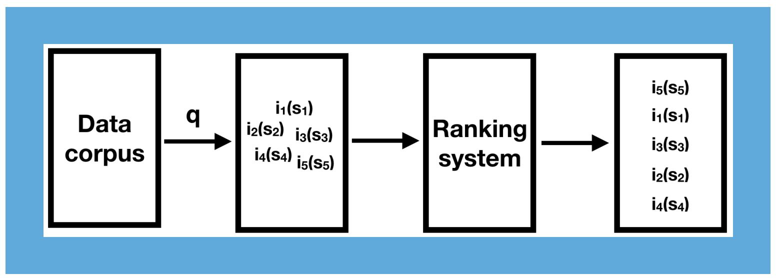
Support for Obamacare growing! We do not need to #RepealAndReplace #Obamacare, fix it as is or keep it for the sake of the people's welfare.

Democratic

Search can shape user opinion

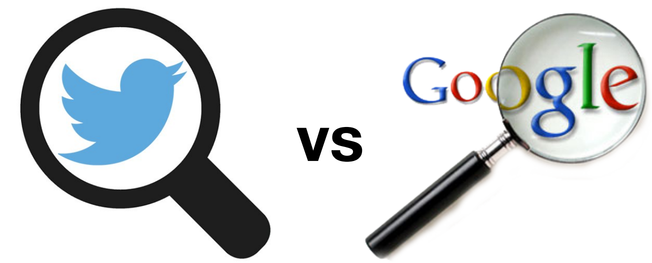
- Search rankings are **perceived as unbiased, accurate and fair** *[Purcell et al., 2012]*
- Biased search results can **influence voting** patterns *[Epstein & Robertson, 2015; Epstein et al., 2017]*

Quantifying search bias



Quantifying search bias

Sources of bias in Twitter search



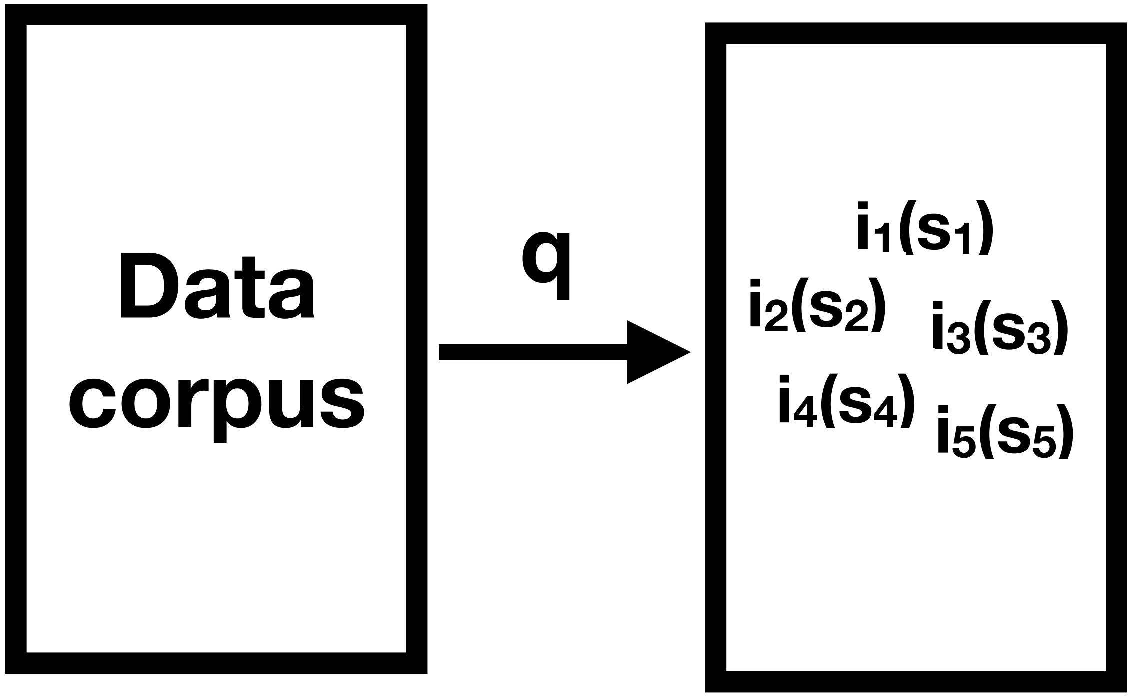
Relative bias of Twitter and Google search

Search bias quantification framework

Data corpus

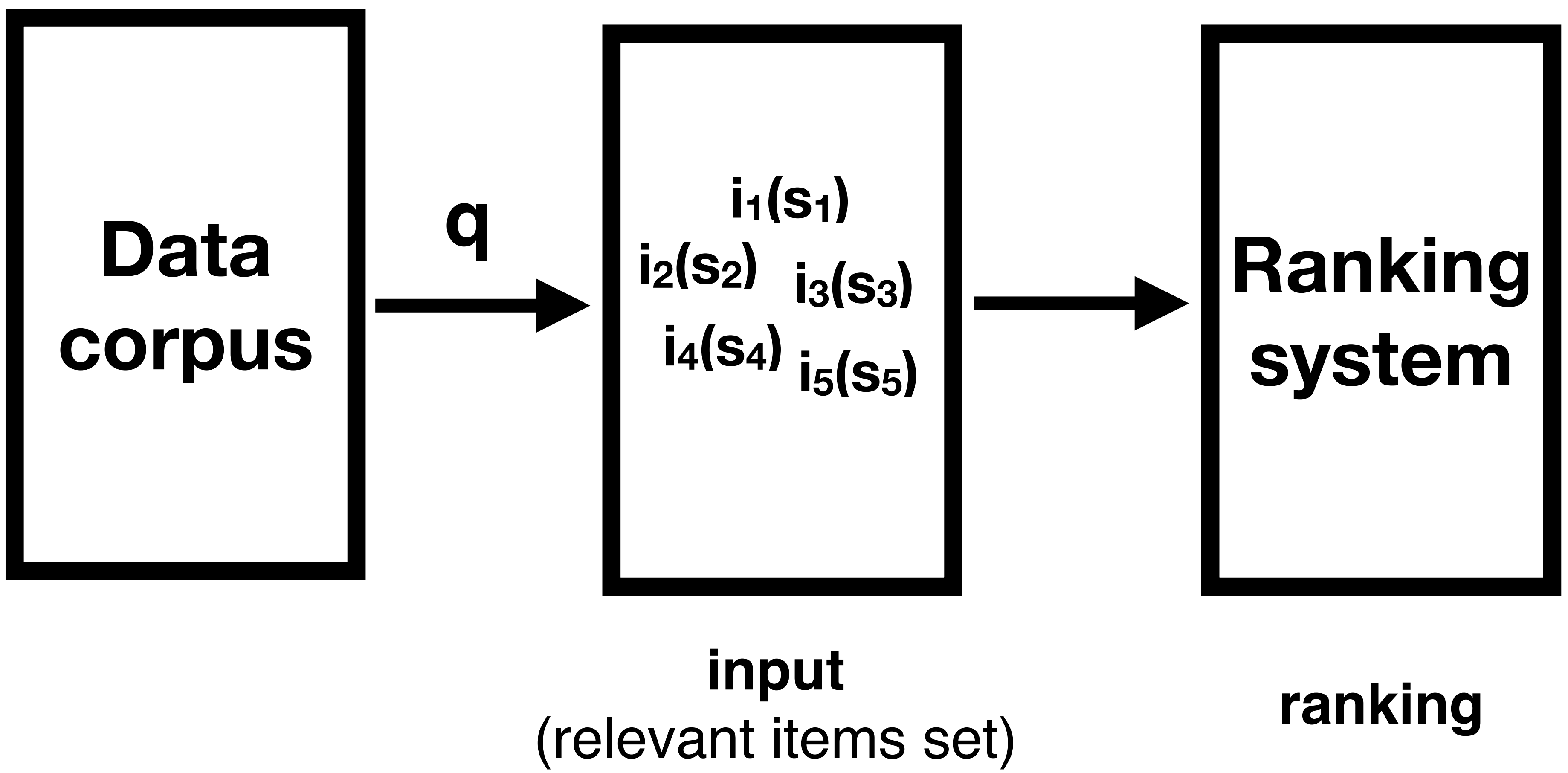
Each data item (e.g., i_1) has an associated bias score (e.g., s_1)

Search bias quantification framework

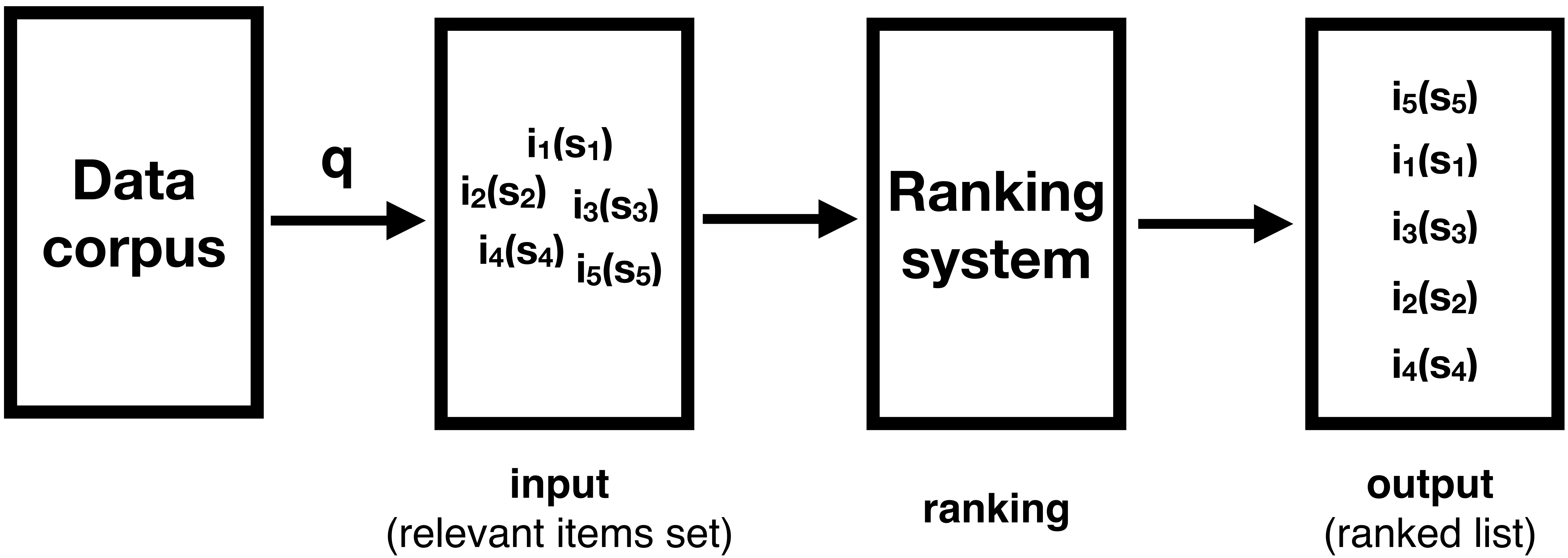


input
(relevant items set)

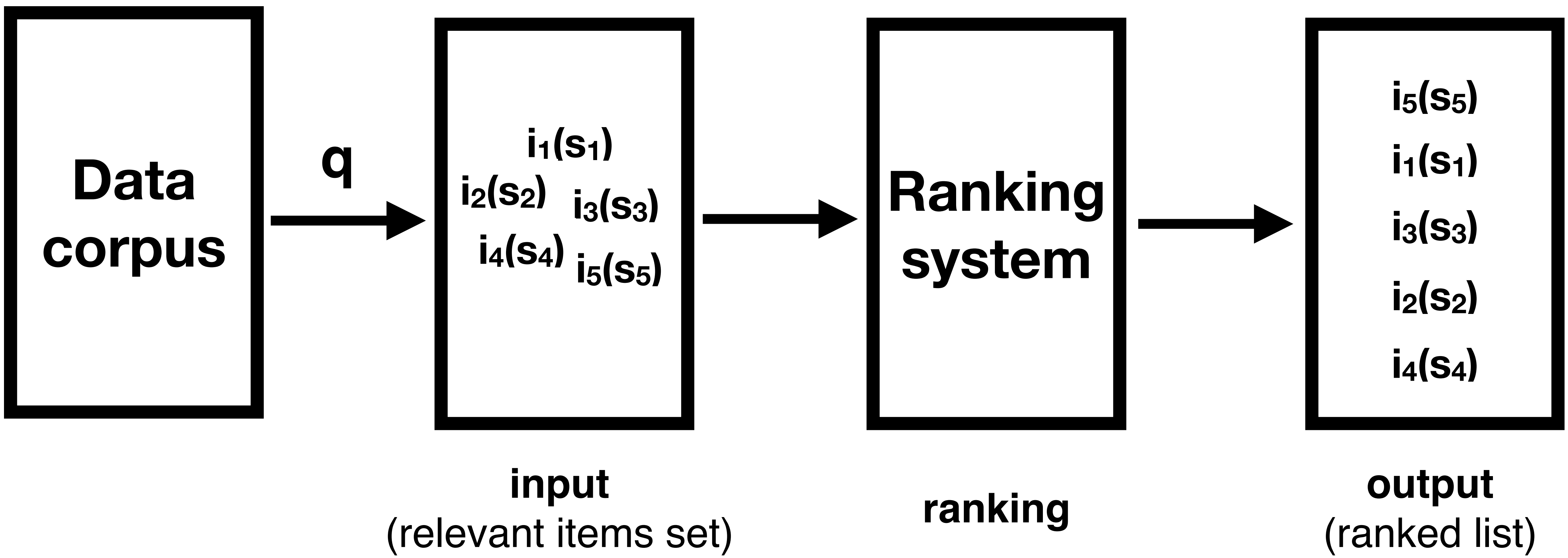
Search bias quantification framework



Search bias quantification framework



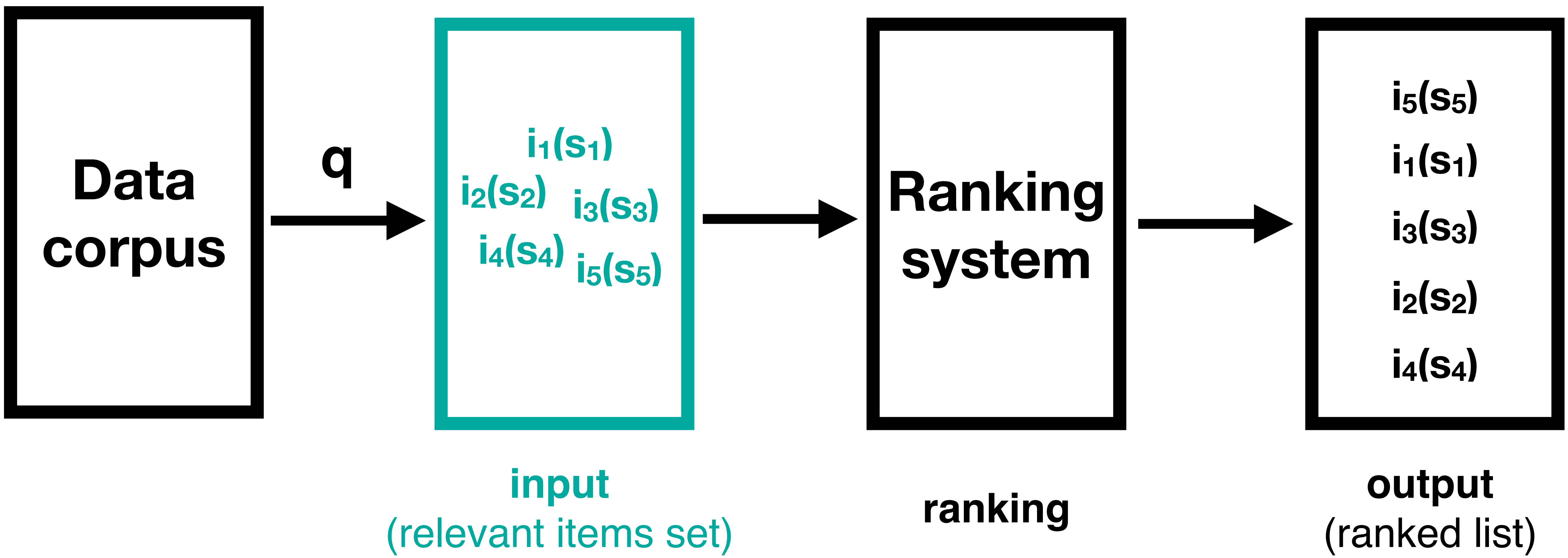
Search bias quantification framework



Output bias may stem from

- Bias introduced by the **ranking** system
- Bias in the **input** relevant item set

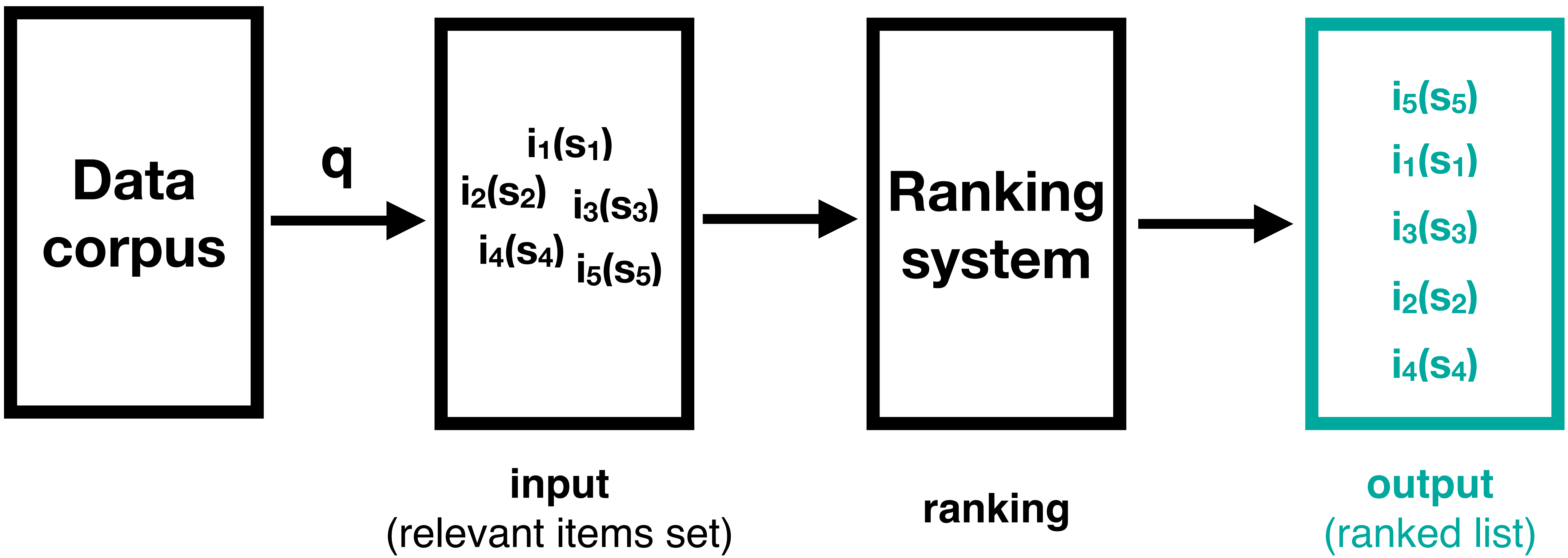
Input Bias



Bias that would be seen in random sample of tweets containing your query

$$IB(q) = \frac{\sum_{i=1}^{n} s_i}{n}$$

Output Bias



Users place **greater trust in higher ranked** items *[Pan et al., 2007; Hargittai et al., 2010]*

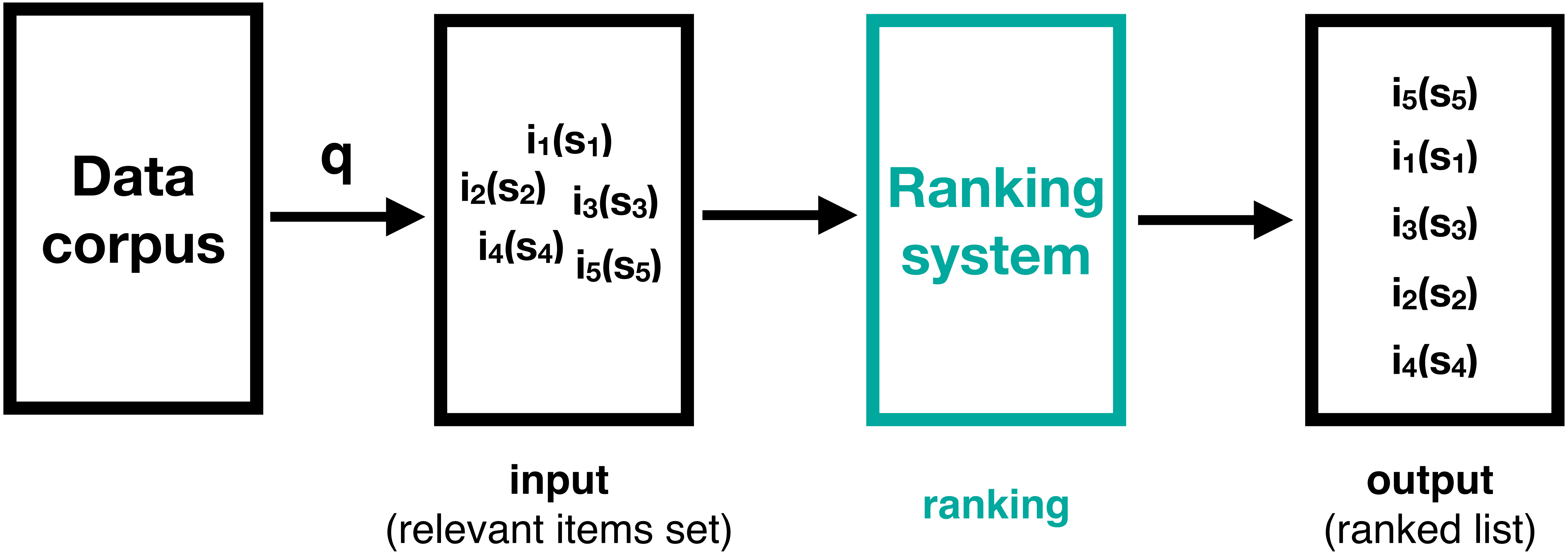
Bias till rank *r*

$$B(q,r) = \frac{\sum_{i=1}^{r} s_i}{r}$$

Output bias

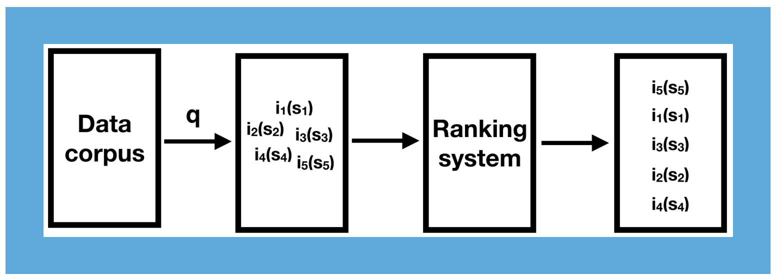
$$OB(q,r) = \frac{\sum_{i=1}^{r} B(q,i)}{r}$$

Ranking Bias



Ranking bias = Output bias - Input bias

Quantifying search bias



Quantifying search bias

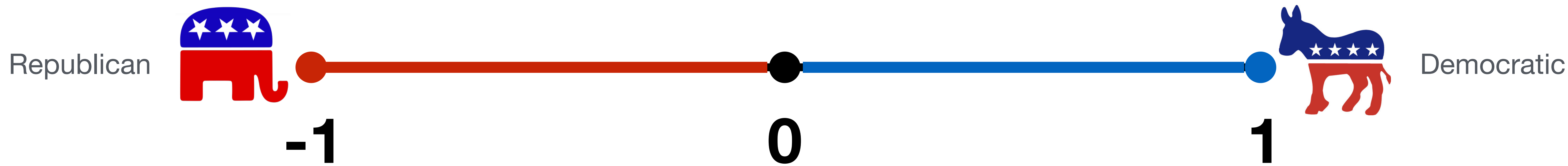
Proposed search bias quantification framework

Sources of bias in Twitter search

vs

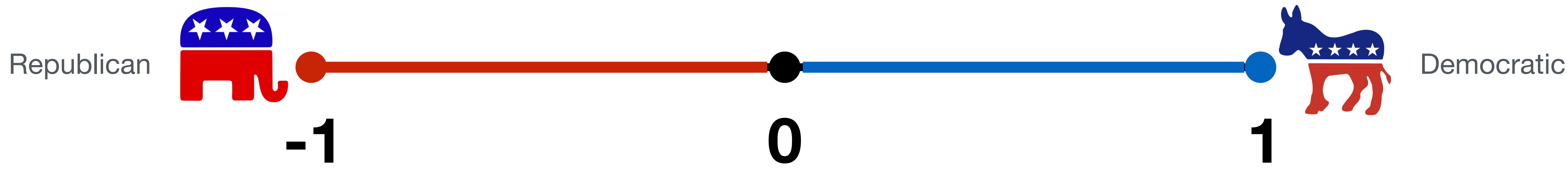
Relative bias of Twitter and Google search

Quantifying bias of a single item



- Operationalize bias of a tweet as bias of it's author
- Scalable crowdsourced methodology for inferring political bias of user *u* based on u's interests

Quantifying bias of a single item



- Operationalize bias of a tweet as bias of it's author
- Scalable crowdsourced methodology for inferring political bias of user *u* based on u's interests
  - Inferring interests of *u* based on who *u* follows
  - Examining how closely *u*'s interests match interests of a set of democrats & republicans

Evaluation

- Do we infer the bias of a user correctly?
  - Use 3 test sets (Elites, Self identified, Politically interested) *[Cohen & Ruths, 2013; Liu and Weber, 2014]*
  - High coverage and accuracy

Details in [Kulshrestha et al., 2017]

Evaluation

- Do we infer the bias of a user correctly?
  - Use 3 test sets (Elites, Self identified, Politically interested) *[Cohen & Ruths, 2013; Liu and Weber, 2014]*
  - High coverage and accuracy
- How closely do author bias and tweet bias reflect each other?
  - High match (70% or more) between the two

Details in [Kulshrestha et al., 2017]

Studying bias for political searches in Twitter

- 2016 US Presidential Primaries
- Presidential primary debates (**#demdebate**, **rep debate**, ...)
- Presidential candidates (**Hillary Clinton**, **Donald Trump**, ...)

Studying bias for political searches in Twitter

- 2016 US Presidential Primaries
- Presidential primary debates (**#demdebate**, **rep debate**, ...)
- Presidential candidates (**Hillary Clinton**, **Donald Trump**, ...)

- **Output**: Scraped Twitter top search snapshots
- **Input**: All tweets with the query, using streaming api

Studying bias for political searches in Twitter

- 2016 US Presidential Primaries
- Presidential primary debates (**#demdebate**, **rep debate**, ...)
- Presidential candidates (**Hillary Clinton**, **Donald Trump**, ...)

- **Output**: Scraped Twitter top search snapshots
- **Input**: All tweets with the query, using streaming api

Non-personalized search data

Studying bias for political searches in Twitter

- 2016 US Presidential Primaries
- Presidential primary debates (**#demdebate**, **rep debate**, ...)
- Presidential candidates (**Hillary Clinton**, **Donald Trump**, ...)

- **Output**: Scraped Twitter top search snapshots
- **Input**: All tweets with the query, using streaming api

Non-personalized search data

- **Computed input, ranking, and output bias** for each query

Studying bias for political searches in Twitter

- Presidential primary debates (**#demdebate, rep debate,** ...)
- Presidential candidates (**Hillary Clinton, Donald Trump,** ...)

What are the sources of bias for political searches on Twitter?

Non-personalized search data

- **Computed input, ranking, and output bias** for each query

Sources of bias on Twitter

- Both input bias and ranking bias matter

Sources of bias on Twitter

Both input bias and ranking bias matter

Even for the same event, query phrasing can greatly effect the bias

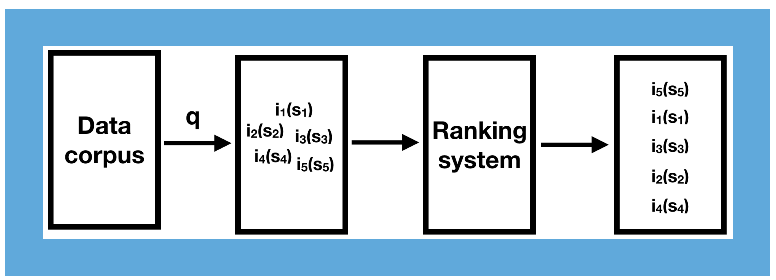
Sources of bias on Twitter

Both input bias and ranking bias matter

Even for the same event, query phrasing can greatly effect the bias

No evidence of systemic bias for Twitter's ranking system

Quantifying search bias



Quantifying search bias

Proposed search bias quantification framework

Sources of bias in Twitter search

Both input and ranking bias matter
Even for the same event, query phrasing can greatly effect the bias
No evidence of systemic bias for Twitter's ranking system

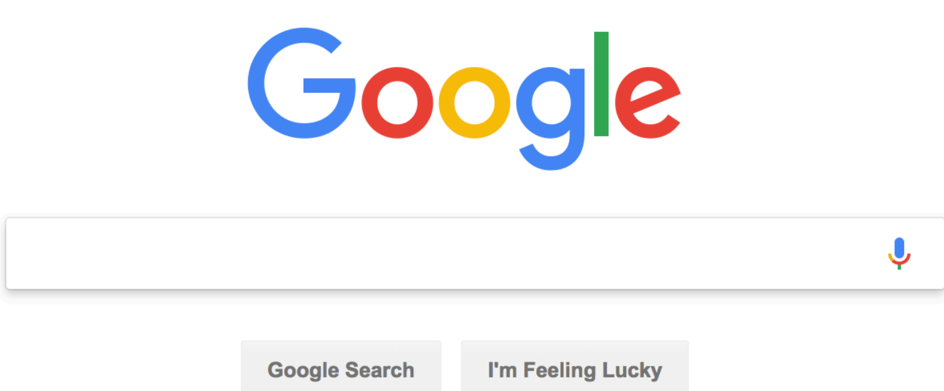
vs

Relative bias of Twitter and Google search

Inferring bias of Google & Twitter "news" results

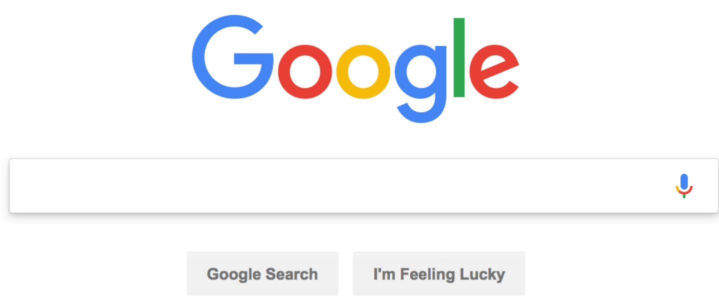
- Compare Google search results with Twitter "news" search results
- Contain a large fraction of links from news media websites
- Semi-manual mapping to Balance scores

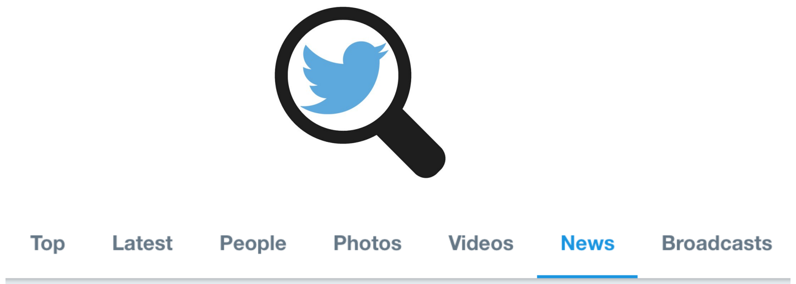
[Munson, Lee & Resnick, 2013]



[Kulshrestha et al., 2019]

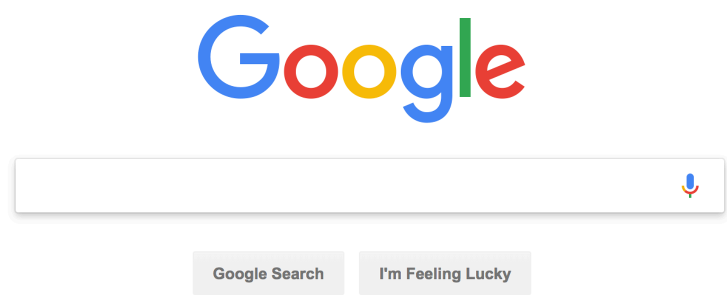
Comparing Relative bias of Google & Twitter

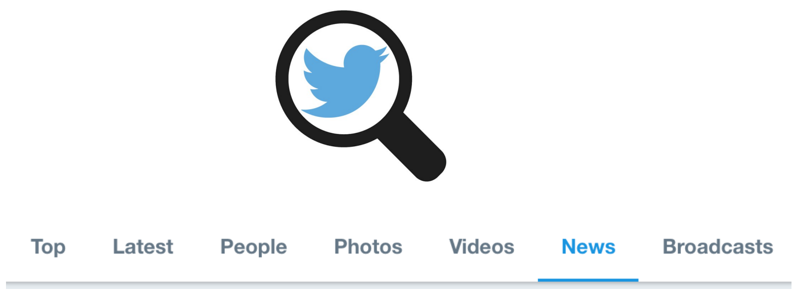




Dynamic over time

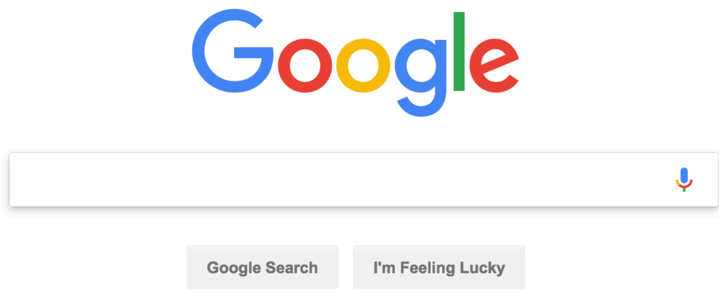
Comparing Relative bias of Google & Twitter





	Google	Twitter
Dynamic over time	✗	✓
Matches leaning of searched person/event	✓	✗

Comparing Relative bias of Google & Twitter



	Google	Twitter
Dynamic over time	✗	✓
Matches leaning of searched person/event	✓	✗
High fraction of candidate-controlled sources	✓	✗

Search engine audit projects

- Five search engines: Google, Bing, DuckDuckGo, Yahoo and Yandex.
- A virtual agent-based infrastructure running multiple queries, from multiple locations and across different time periods

Telematics and Informatics
Volume 72, August 2022, 101860

Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results

Aleksandra Urman [a], Mykola Makhortykh [b], Roberto Ulloa [c], Juhi Kulshrestha [d]

Social Science Computer Review
2024, Vol. 42(3) 700–718
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393231195471
journals.sagepub.com/home/ssc
Sage

Novelty in News Search: A Longitudinal Study of the 2020 US Elections

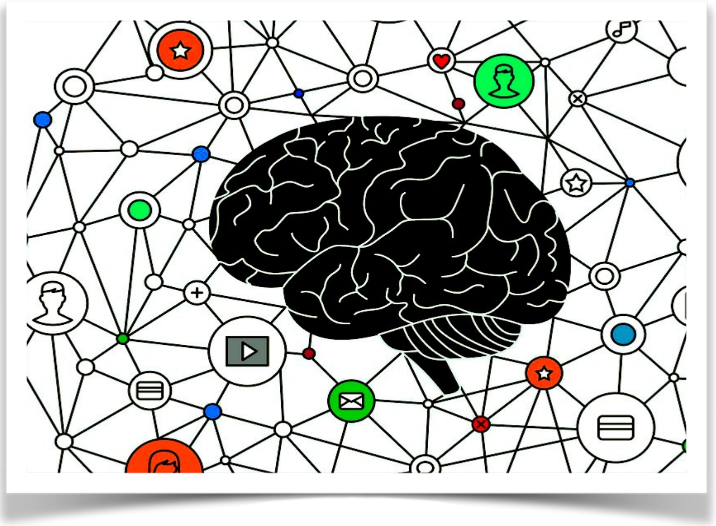
Roberto Ulloa [1,2], Mykola Makhortykh [3], Aleksandra Urman [4], and Juhi Kulshrestha [5]

Constants and Variables: How Does the Visual Representation of the Holocaust by AI Change Over Time

Aleksandra Urman, Mykola Makhortykh, Roberto Ulloa, Maryna Sydorova and Juhi Kulshrestha

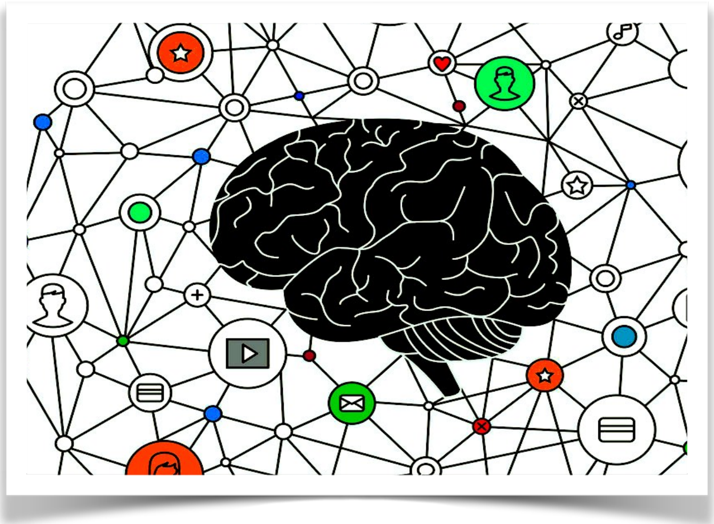
From the journal Eastern European Holocaust Studies
https://doi.org/10.1515/eehs-2023-0055

Influence of online behaviour



[Image credit: VectorStock.com/26073249]

Influence of online behaviour



[Image credit: VectorStock.com/26073249]

An experimental study of online information seeking & policy judgments

Celina Kacperski, Roberto Ulloa, Denis Bonnay, Andreas Spitz, Peter Selb, Juhi Kulshrestha

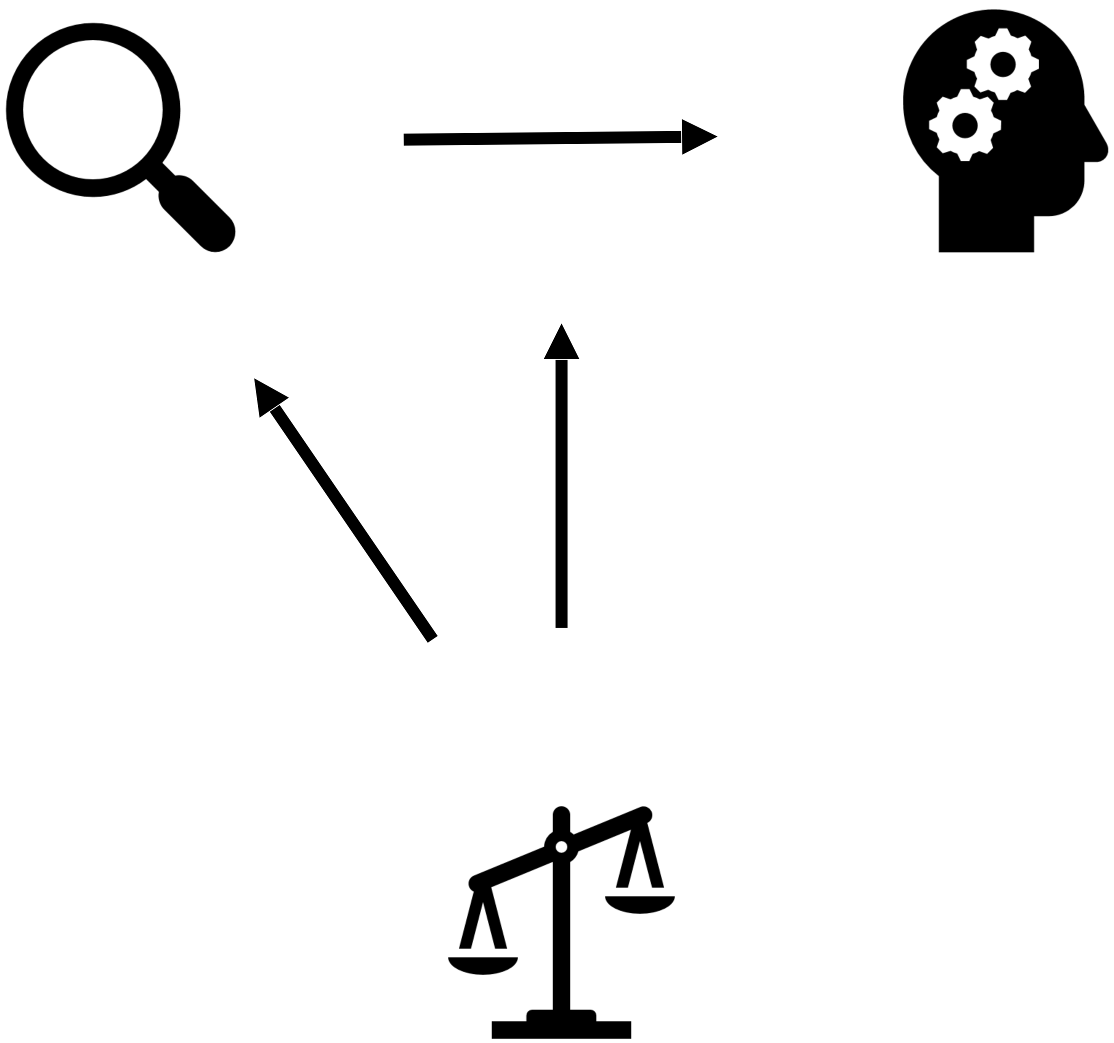
Aalto University

Universität Konstanz

Cluster of Excellence The Politics of Inequality

Funded by Deutsche Forschungsgemeinschaft German Research Foundation

Goals

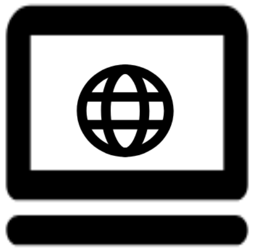


Examine individuals' online information seeking strategies and understand how they affect their political knowledge and judgements of policy proposals

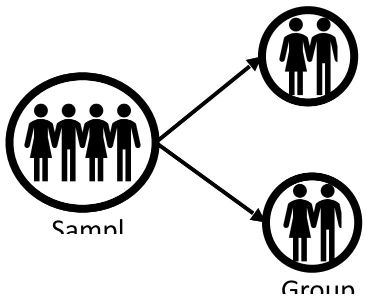
Examine whether inequality barriers (e.g., capability, opportunity, motivation) shape online information seeking strategies and therefore political knowledge and judgements of policy proposals

Motivation

Uninformed political reasoning may entail choices against one’s own interest

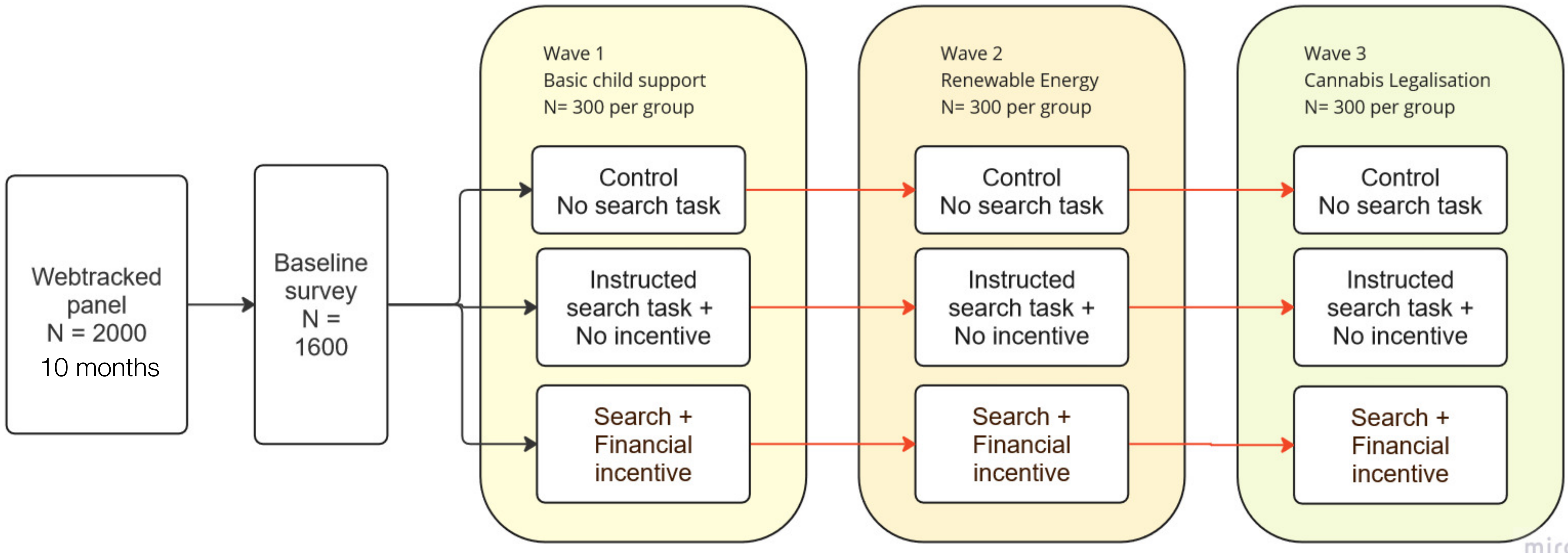


Role of internet: motivating selective exposure and heuristic searches or democratizing access to information (or both)?



Causal evidence around online behaviour is lacking

Data collection



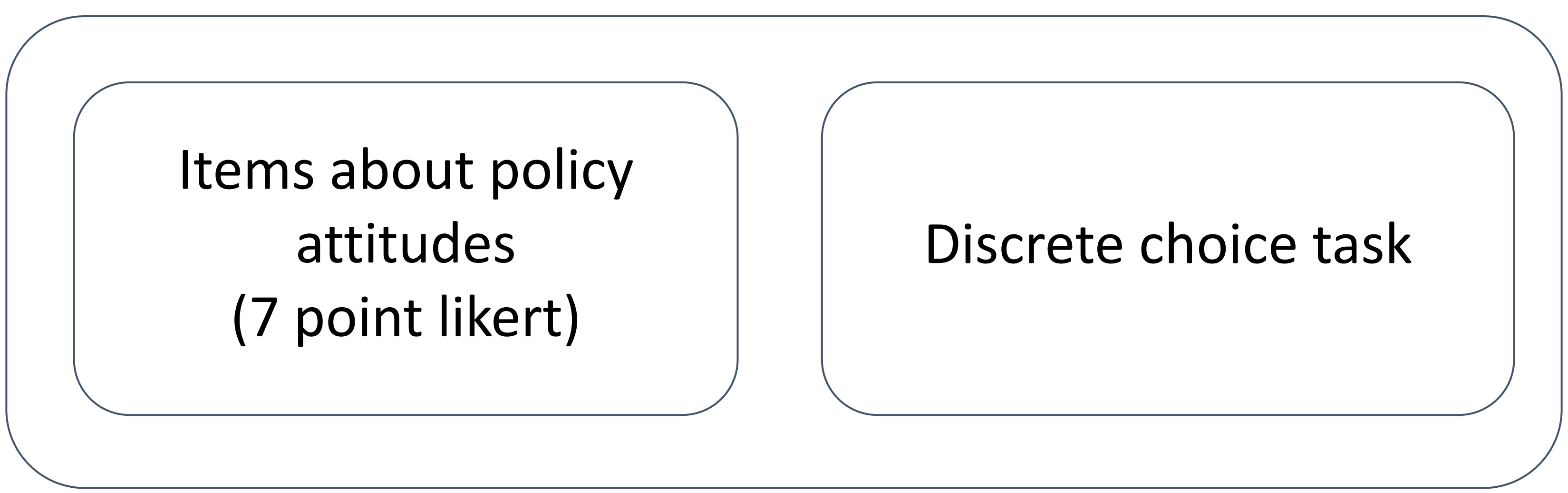
Discrete choice task

Für welchen der folgenden zwei Gesetzesvorschläge zur Kindergrundsicherung würden Sie bei einer Wahl stimmen?

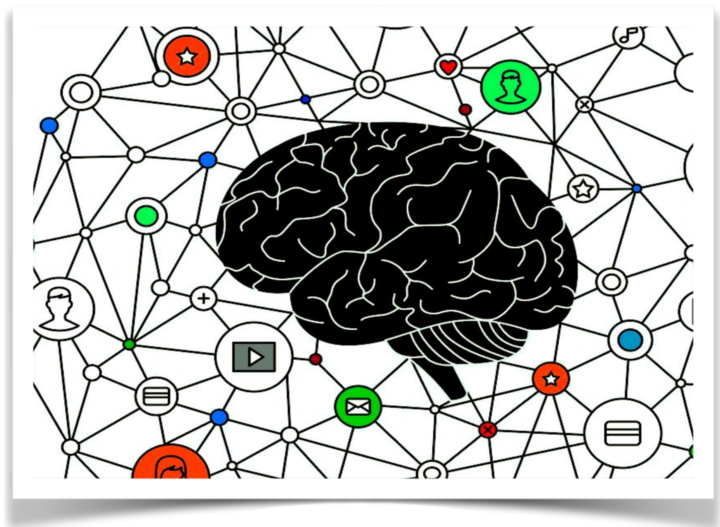
	Vorschlag 1	Vorschlag 2
Empfangsberechtigung neben deutschen Staatsbürgern	mindestens gemeldet (z.B. als Asylsuchende/r)	mindestens Niederlassungserlaubnis
Fixer Kindergeld Grundbetrag	250 Euro, Erhöhung alle 2 Jahre	sofort weniger als 250 Euro
Maximalbetrag des Zusatzgeldes (je nach Einkommen)	150 Euro	150 Euro
zusätzliche Kosten für den Staat	mehr als 12 Milliarden	weniger als 12 Milliarden
	○	○

Preliminary results

- Interventions worked, individuals searched more (seen via browsing traces) and provided more links in tasks
- Searching for information increased knowledge
- Searching for information lead to change in attitudes
- Although no change in discrete choice tasks



Influence of online behaviour



[Image credit: VectorStock.com/26073249]

Internet Use & Psychological well-being
A longitudinal multimodal study

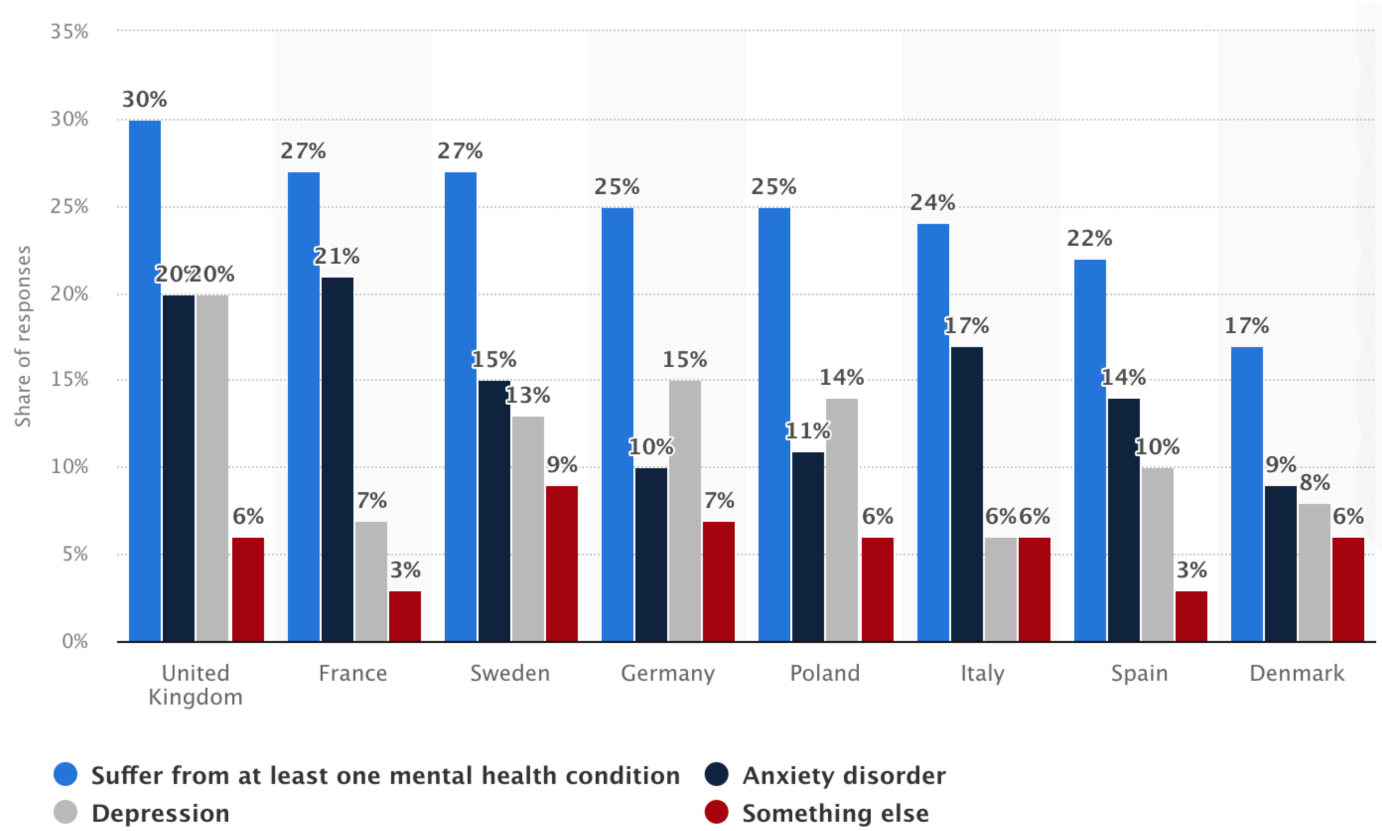
Mohammad Belal, Emilia Marchese, Nguyen Luong, Talayeh Aledavood, Juhi Kulshrestha

Aalto University

Funded by
HIIT | HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

Global mental health crisis

- 970M people globally that suffer from mental health or substance abuse disorders
- As of 2021, >150 million European residents lived with mental health conditions



[Source: statista.com]

- Not enough mental health workers — Europe 45 per 100K population, worldwide 13 per 100K population

Internet use & mental health

- Psych studies traditionally focus on our offline lives => critical gap in understanding complete picture of our life experiences

- Study of 99 commonly used psych assessment scales identified 196 dimensions — none directly related to online activities or behaviour

Internet use & mental health

- Psych studies traditionally focus on our offline lives => critical gap in understanding complete picture of our life experiences

- Study of 99 commonly used psych assessment scales identified 196 dimensions — none directly related to online activities or behaviour

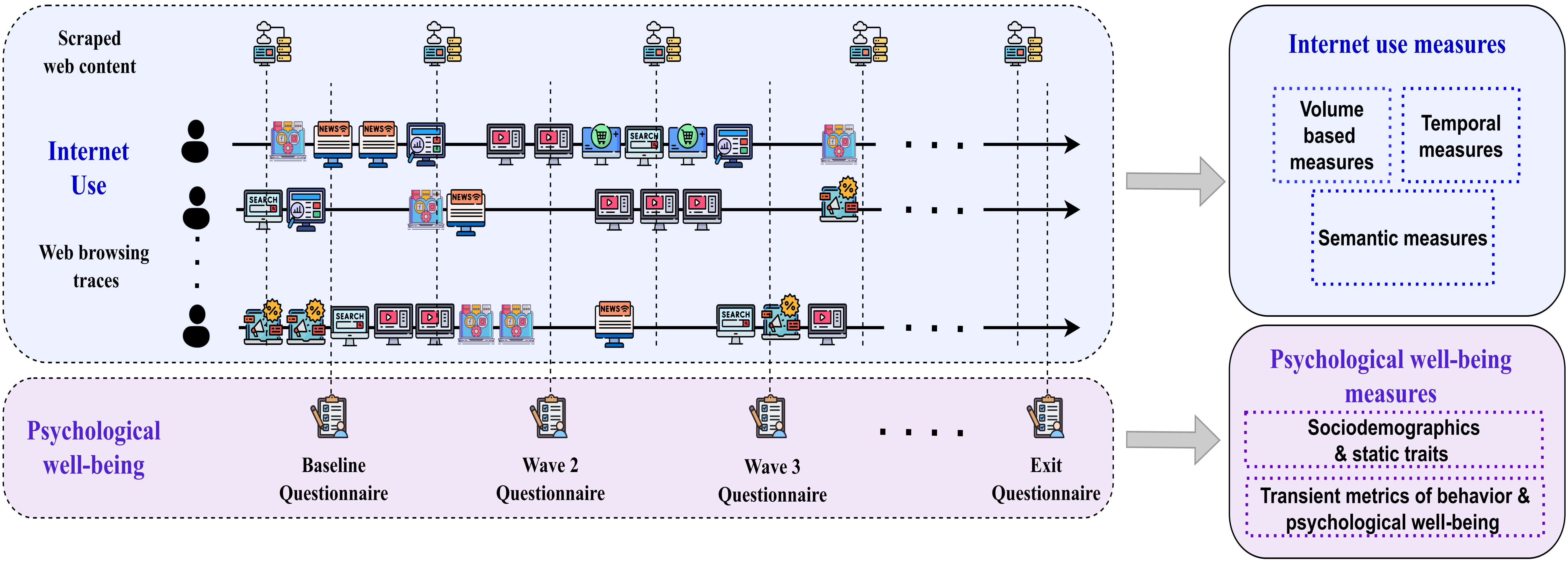
- Research gap:
  - Focus on single platform - miss full spectrum of online engagement that shape individuals' experiences
  - Rely on self reported assessments and cross-sectional design

Research goals

Uncover behavioural markers in internet use associated with individuals' psychological wellbeing

Design personalized tools for self-monitoring and self-moderation of online activity for improved mental health

Study design



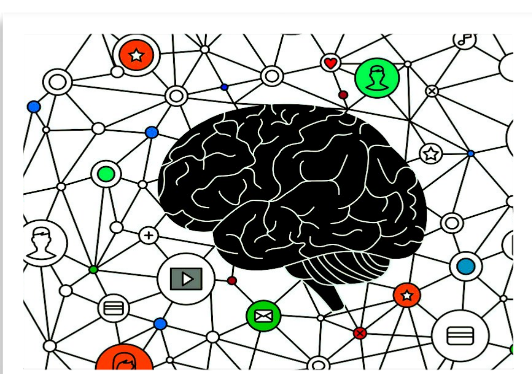
Preliminary results

- Depression and stress
- Associations between volume of internet use and time of use
- More pronounced for certain categories of web browsing (e.g., adult content, gaming, news)

[Image credit:

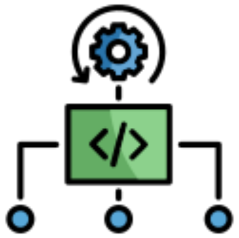
[Image credit: nbjournal.ru/

[Image credit:

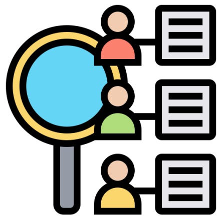


[Image credit: VectorStock.com/

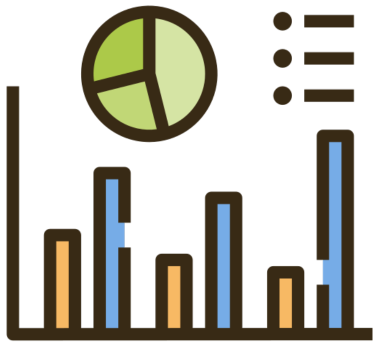
Digital behavioural data



Algorithm audit frameworks



Online surveys & experiments



Data science
Machine learning
Statistical analysis
Natural language processing…

Doing good CSS research

- Computational social science ≠ computer science + social data
  [Hannah Wallach 2018]
- Become knowledgeable in different social science disciplines (or at least the ones that you are most interested in!)
  - Get acquainted with the main theories, familiarise yourself with the vocabulary
- Be open to different forms of data (or even better combinations of data)
- Be willing to learn new research methods constantly
- Spend time in different disciplinary settings — conferences/workshops, internships, research stays
- Form multidisciplinary connections
- Ask for help!