Knowing Unknowns in an Age of Information Overload

Saurabh Khanna

Stanford University

July 13, 2022

How we know

The (changing) role of the Internet

Internet → Indexer of human knowledge

- Speed and Scale
  - 5.6B searches per day
  - 40M books digitized → 130M
- COVID-19
  - 47% rise in broadband usage
  - ~~32%~~ 62% American parents report teens' daily non-school internet use exceeding four hours

Internet → ~~Indexer~~ Provider of human knowledge

How we know

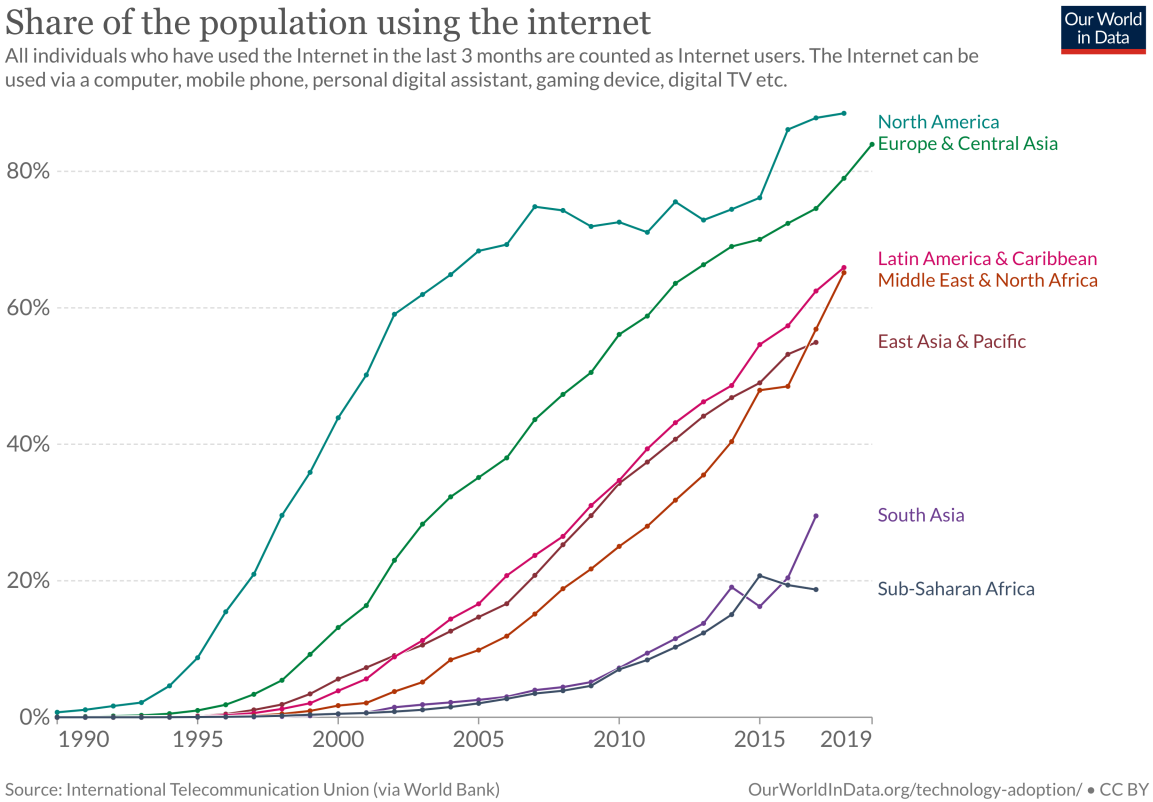
The good stuff?

- Democratic flow of information

How we know

The good stuff?

- Democratic flow of information
- But how democratic?



How we know

The good stuff?

- Rapidly growing research on
  - Misinformation: Information consumed $\neq$ ground truth
  - Bias: Information consumed $\neq$ ground truth + discriminates against a social group

How we know

The good stuff?

- Rapidly growing research on
  - Misinformation: Information consumed $\neq$ ground truth
  - Bias: Information consumed $\neq$ ground truth + discriminates against a social group

- Ground truth?
  - The election was rigged ✗
  - People think the election was rigged ✓

A more insidious (and less researched) problem exists...

How we know

Information overload on the Internet

Query

↓

Information overload

How we know

Information overload on the Internet

Query

↓

Information overload

↓

Information is explicitly ranked by relevance ~~by~~ for us

↓

Information overload

How we know

Information overload on the Internet

Query

↓

Information overload

↓

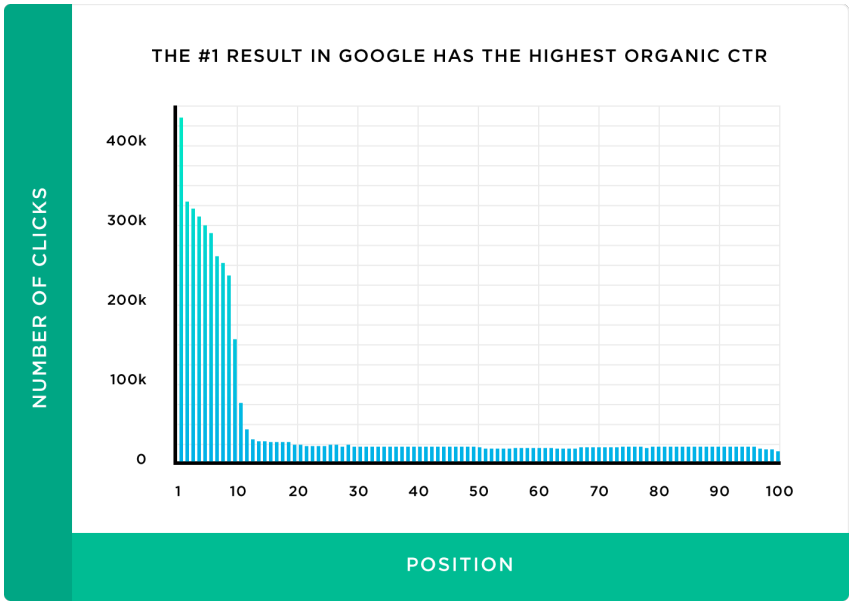
Information is explicitly ranked by relevance ~~by~~ for us

↓

Information overload

↓

We consume the tip of the iceberg



How we know

Information overload on the Internet

Query

↓

Information overload

↓

Information is explicitly ranked by relevance ~~by~~ for us

↓

Information overload

↓

We consume the tip of the iceberg

Two problems:

1. Harms of representation

How we know

Information overload on the Internet

Query

↓

Information overload

↓

Information is explicitly ranked by relevance ~~by~~ for us

↓

Information overload

↓

We consume the tip of the iceberg

Two problems:

1. Harms of representation
2. Relevance might not be the only thing that matters

How we know

Relevance might not be the only thing that matters

Relevance of a document given a query can be computed as the semantic distance between them in the embedding space (Microsoft DSSM, 2020).

$$\text{Query (q)} \xrightarrow[\checkmark]{\text{What I want}} \text{Search result } (r_i)$$

How we know

Relevance might not be the only thing that matters

Relevance of a document given a query can be computed as the semantic distance between them in the embedding space (Microsoft DSSM, 2020).

$$\text{Query (q)} \xrightarrow[\checkmark]{\text{What I want}} \text{Search result } (r_i)$$

But what about representation?

$$\text{Query (q)} \xrightarrow[\checkmark]{\text{What I want}} \text{Search result } (r_i) \xleftarrow[\times]{\text{What exists}} \text{Corpus}$$

Taken together

- The Internet is our primary information source
- Democratic, but within siloes
- We are studying misinformation and bias, but not what accessing incomplete information does to us
- We consume the tip of a pre-ranked iceberg
- Harms of representation + Lack of choice

> If you control the flow of information in a society, you can influence its shared sense of right and wrong, fair and unfair, clean and unclean, seemly and unseemly, real and fake, true and false, known and unknown.
>
> – Susskind, Future Politics (2018)

Objectives

Know how much we do not know[1]

- We have the data and methods now to approximate an answer

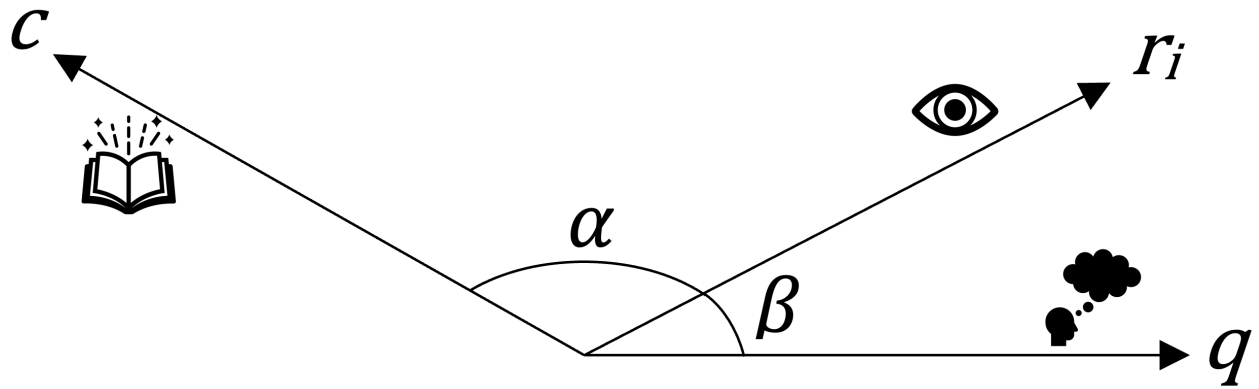
Specifically, when accessing ranked information on the internet:

1. How representative (and not just relevant) is what we know?
2. What are its implications?

[1] Plato 399 BC, Einstein 1931, Taleb 2007

Method

Balancing Relevance and Representation



q: query, what I want to know (relevance)
r_i: one of out n search results
$c = \sum w_i r_i$: corpus, what exists out there (representation)

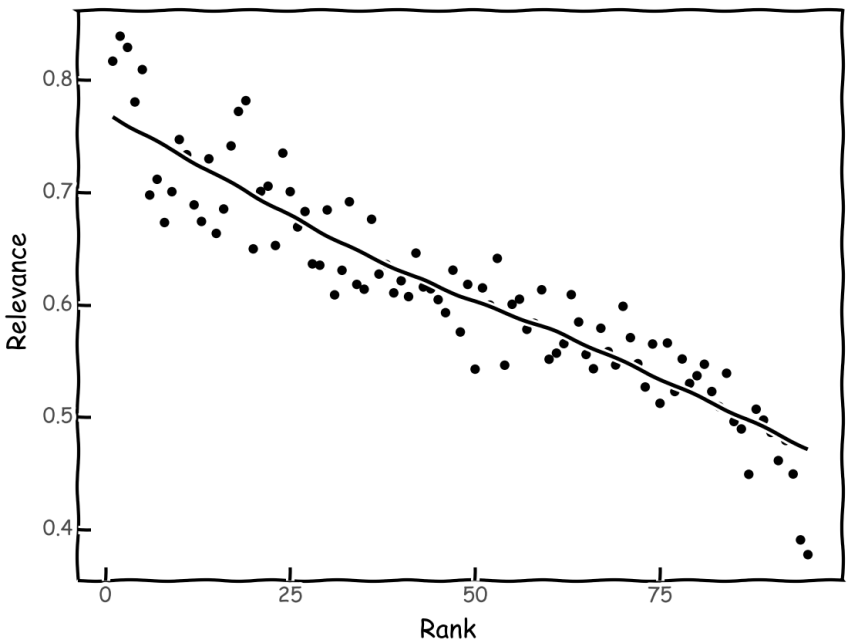
The embedding for c is generated using a weighted aggregate of all r_i vectors. Why weight?

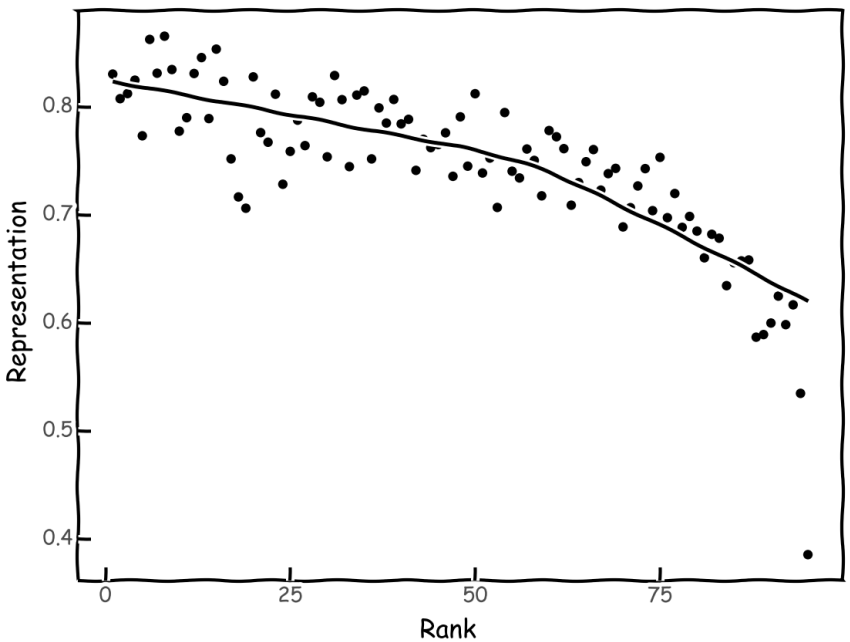
Assign a score S_i to each search result r_i, where λ controls the balance between relevance and representation

$$S_i = \lambda \cos \alpha + (1 - \lambda) \cos \beta,$$

Balancing Relevance and Representation

Ideally...

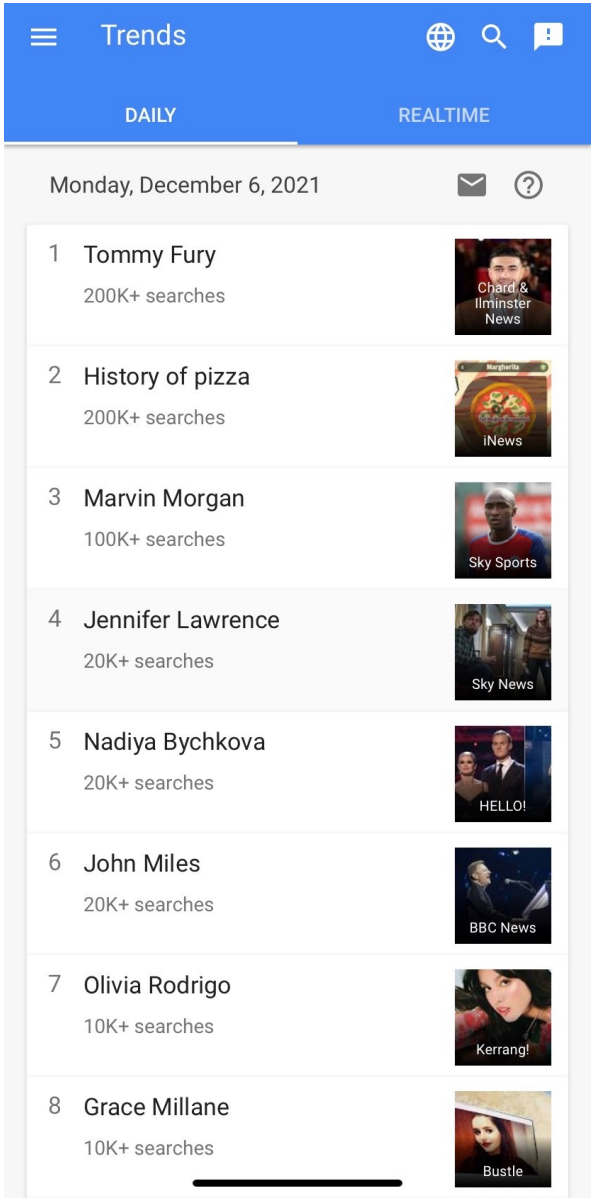




Data

Data

Examine
relevance-representation status
for worldwide search trends



Data

Examine
relevance-representation status
for worldwide search trends

Everyday[2]:

48 nations
×
1.2 million searches/nation
×
319 results/search[3]
≈
18 billion daily data points

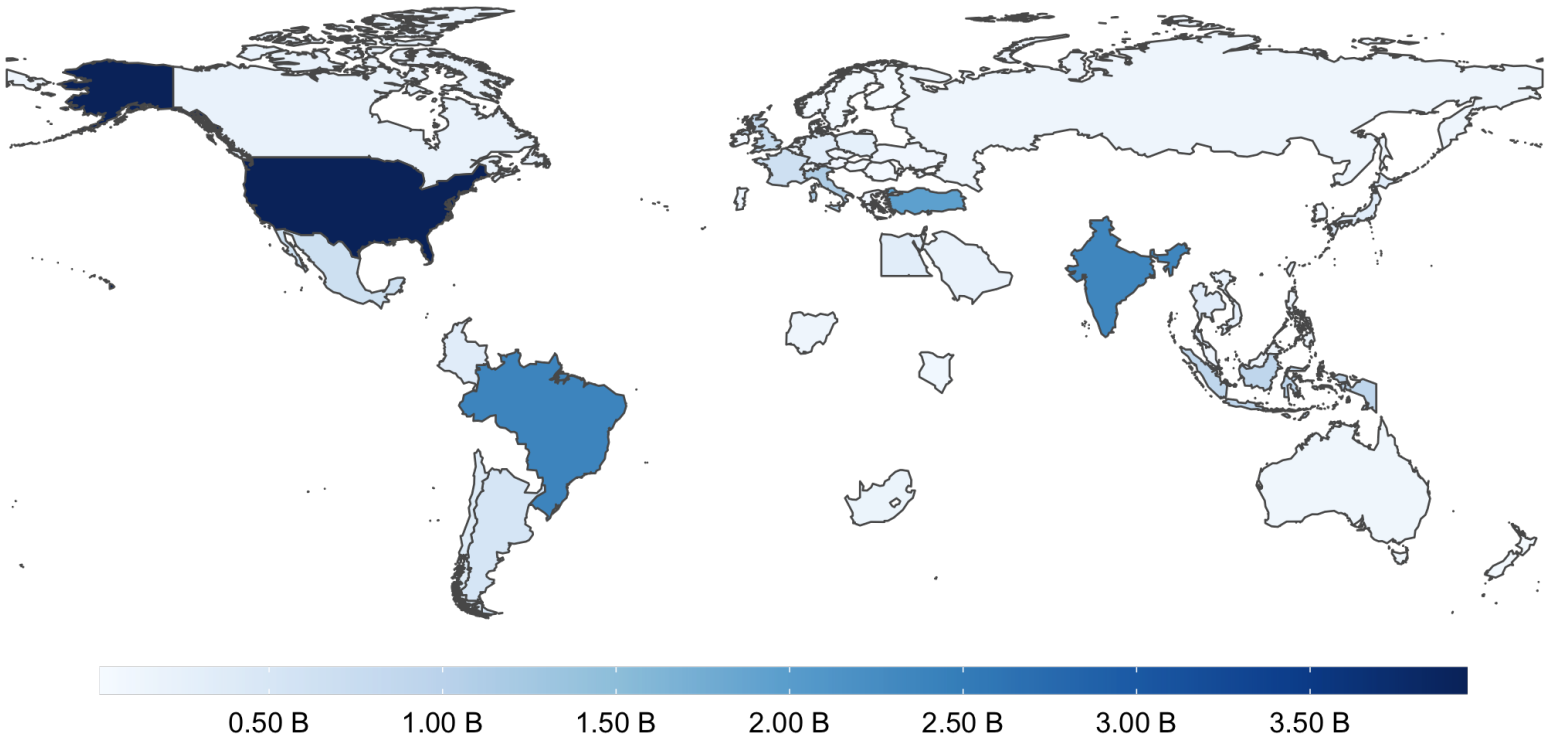
[4.2 trillion data points till date]

[2]reporting medians

[3]Both web and news search results. Since 2016, Google caps the maximum search results shown to 400.

Data

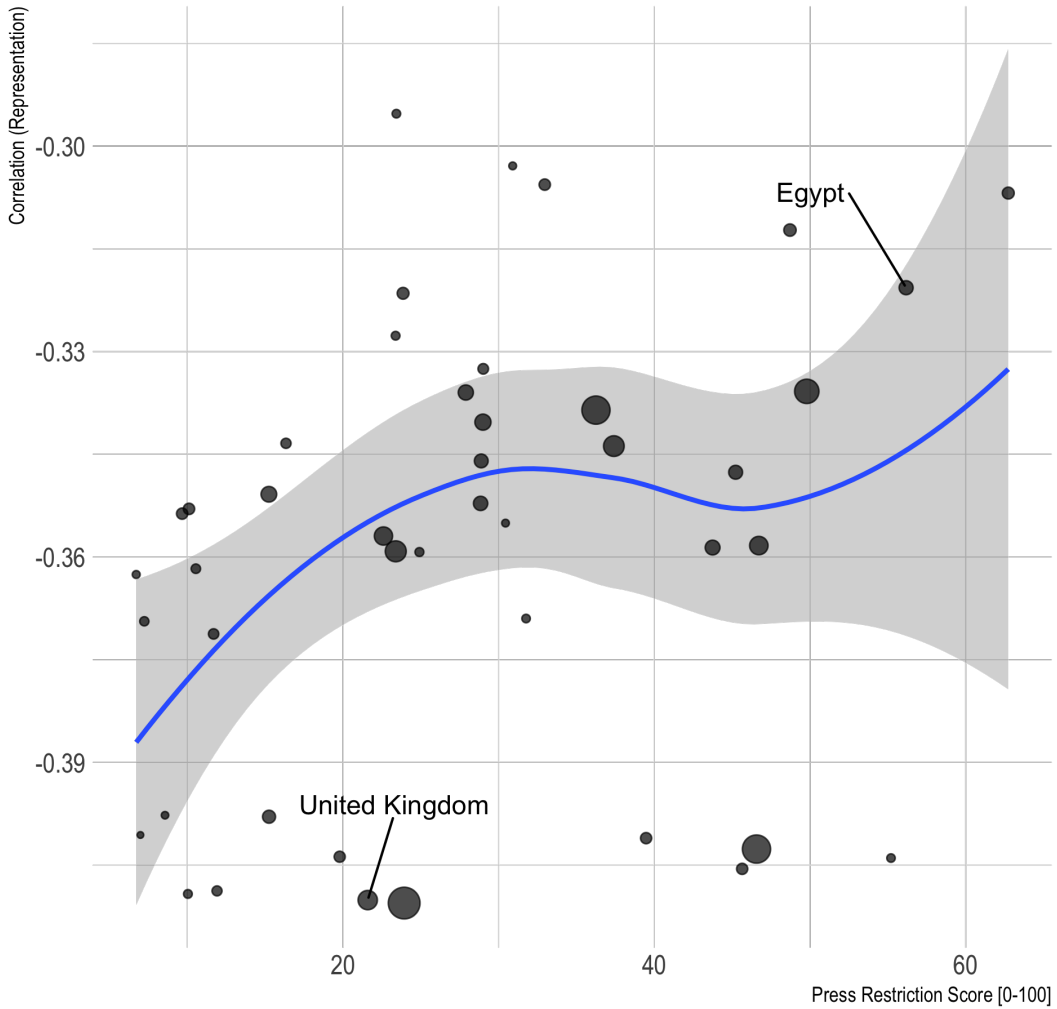
Daily Search Volume Fetched



Very Preliminary Results

Representation Correlation

Variation with press restrictions



Source: Reporters sans frontières, 2021. Point sizes vary with search volume.

Representation Correlation

Variation with press restrictions

	Model 1	Model 2	Model 3	Time FE	Region FE[4]
(Intercept)	−0.348	−0.342	−0.342	−0.339	−0.254***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Press restrictions	0.002***	0.002***	0.002***	0.002***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Search volume		0.000***	0.000***	0.000***	0.000
		(0.000)	(0.000)	(0.000)	(0.000)
GDP per capita			0.001	0.001	0.010***
			(0.001)	(0.001)	(0.001)
Population			0.000	0.000	0.019***
			(0.001)	(0.001)	(0.001)
Date				0.001	0.001
				(0.001)	(0.001)

***p < 0.001; **p < 0.01; *p < 0.05. Effect sizes in SD units.

[6]Region fixed effects include East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, North America, and South Asia.

Reflections

- As the internet feeds us a (likely biased) tip of the iceberg, understanding relevance-representation trade-offs can help assess how much we miss out on

- Sampling implications
  - Digital information is not sorted by the information dimension you care for
  - Sample ~~top n results~~ results until a threshold
  - Reorder results to maximize visibility along the information dimension crucial for a study

- Regional differences

- Limitations
  - Not capturing the long tail of non-trending search queries
  - Not capturing information that was not indexed for web search

The way ahead

- Media Policing amid the invasion of Ukraine (MIT Media Lab)
- Image searches and skin tone distributions (Stanford Autonomous Agents Lab)
- Compare search platforms (S-DEL)
  - Across countries
  - Within countries [Russia]
- Is ignorance bliss?
  - Effect of varying λ on tolerance and polarization

Thanks!

Feedback/Questions

References

- Cardoso, R. (2017). Antitrust: Commission fines Google €2.42 billion for abusing dominance as search engine by giving illegal advantage to own comparison shopping service. European Commission.
- Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.
- Nissenbaum, L. D. I. H., & Introna, L. D. (2000). Shaping the web: Why the politics of search engines matters. The Information Society, 16(3), 169-185.
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the conference on fairness, accountability, and transparency, 59-68.
- Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. Information processing & management, 40(4), 693-707.