Knowing Unknowns in an Age of Incomplete Information

Saurabh Khanna

Stanford University

October 27, 2022

Background

Digitizing Human Lives

- Scale
  - $5.6B$ Google searches per day
  - $40M$ books digitized $\rightarrow 130M$
- COVID-19 –
  - 47% rise in broadband usage in the US
  - ~~32%~~ 62% American parents report teens' daily internet use exceeding four hours

Digitizing Human Lives

Concern 1

- Misinformation: Information consumed $\neq$ ground truth
- Bias: [Information consumed $\neq$ ground truth] + discriminates against a social group

- Ground truth?

[1] Flanagin & Metzger, 2000
[2] Tucker & Persily, 2020
[3] Bail, 2021

Digitizing Human Lives

Concern 1

- Misinformation: Information consumed $\neq$ ground truth
- Bias: [Information consumed $\neq$ ground truth] + discriminates against a social group

- Ground truth?
  - ~~A norm~~ An exception on the Internet[1]

[1] Flanagin & Metzger, 2000
[2] Tucker & Persily, 2020
[3] Bail, 2021

Digitizing Human Lives

Concern 1

- Misinformation: Information consumed $\neq$ ground truth
- Bias: [Information consumed $\neq$ ground truth] + discriminates against a social group

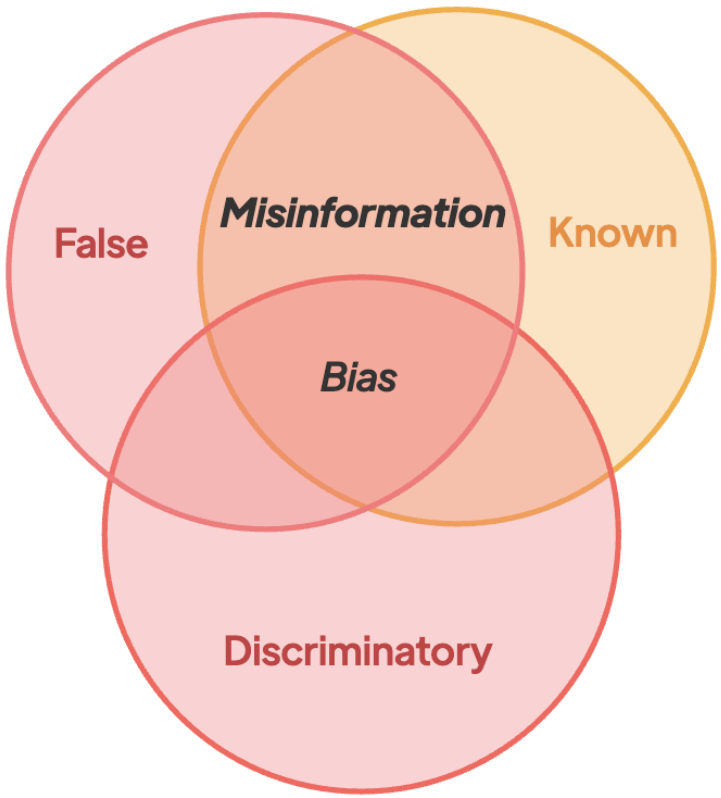
- Ground truth?
  - ~~A norm~~ An exception on the Internet[1]
  - Consider two statements:
    - S_1: The election was rigged ✗
    - S_2: People think the election was rigged ✓
    - S_1 and S_2 have similar effect on the reader[2]
    - Telling the reader S_1 is False doesn't help either[3]

[1] Flanagin & Metzger, 2000
[2] Tucker & Persily, 2020
[3] Bail, 2021

Digitizing Human Lives

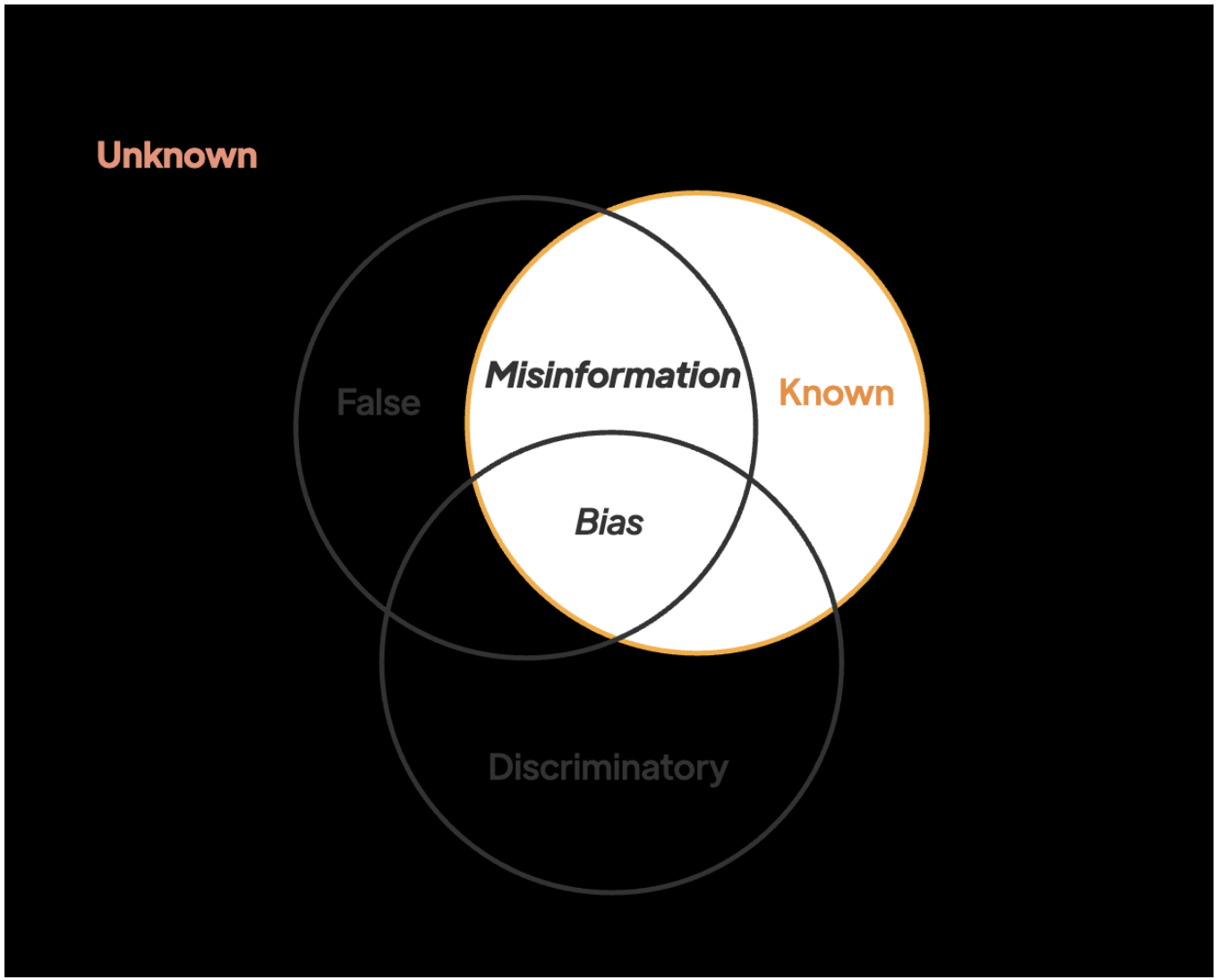
Concern 2

Information Space on the Internet



Digitizing Human Lives

Concern 2



The Problem of Incomplete Information

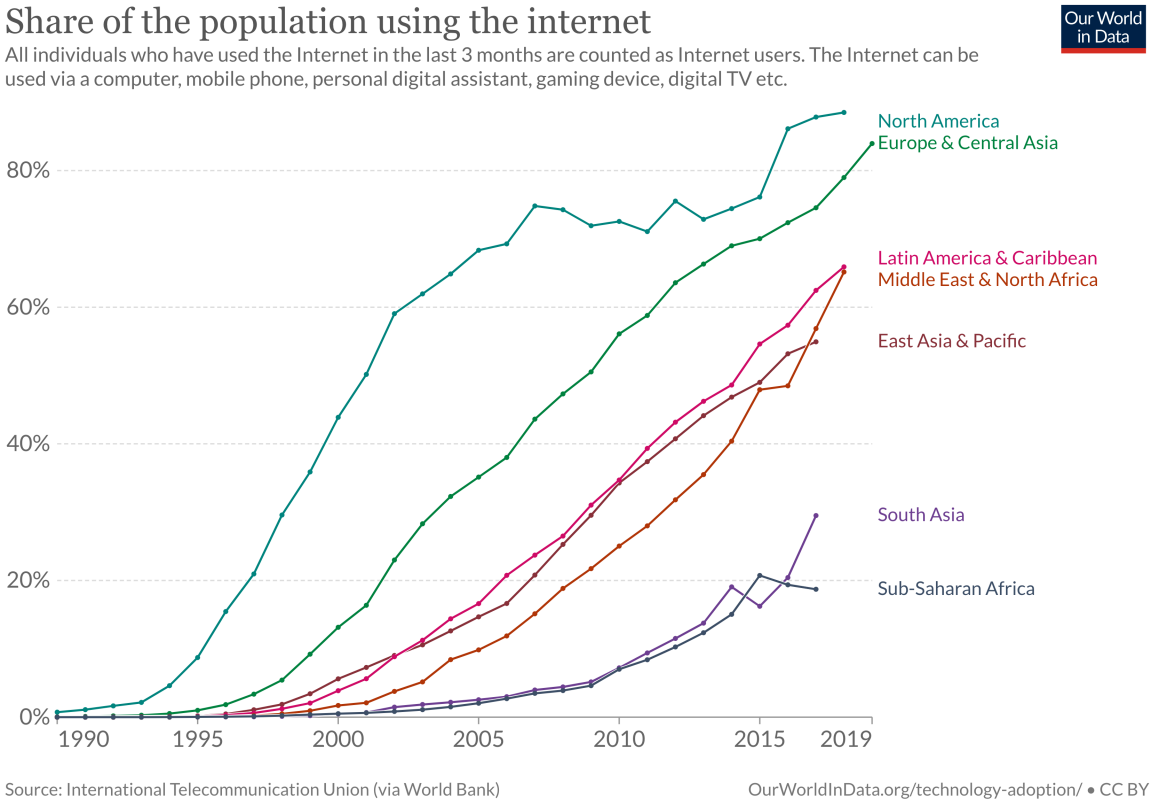
Query the Internet

↓

Incomplete Information

↓

Information ranked ~~by~~ for us



The Problem of Incomplete Information

Query the Internet

↓

Incomplete Information

↓

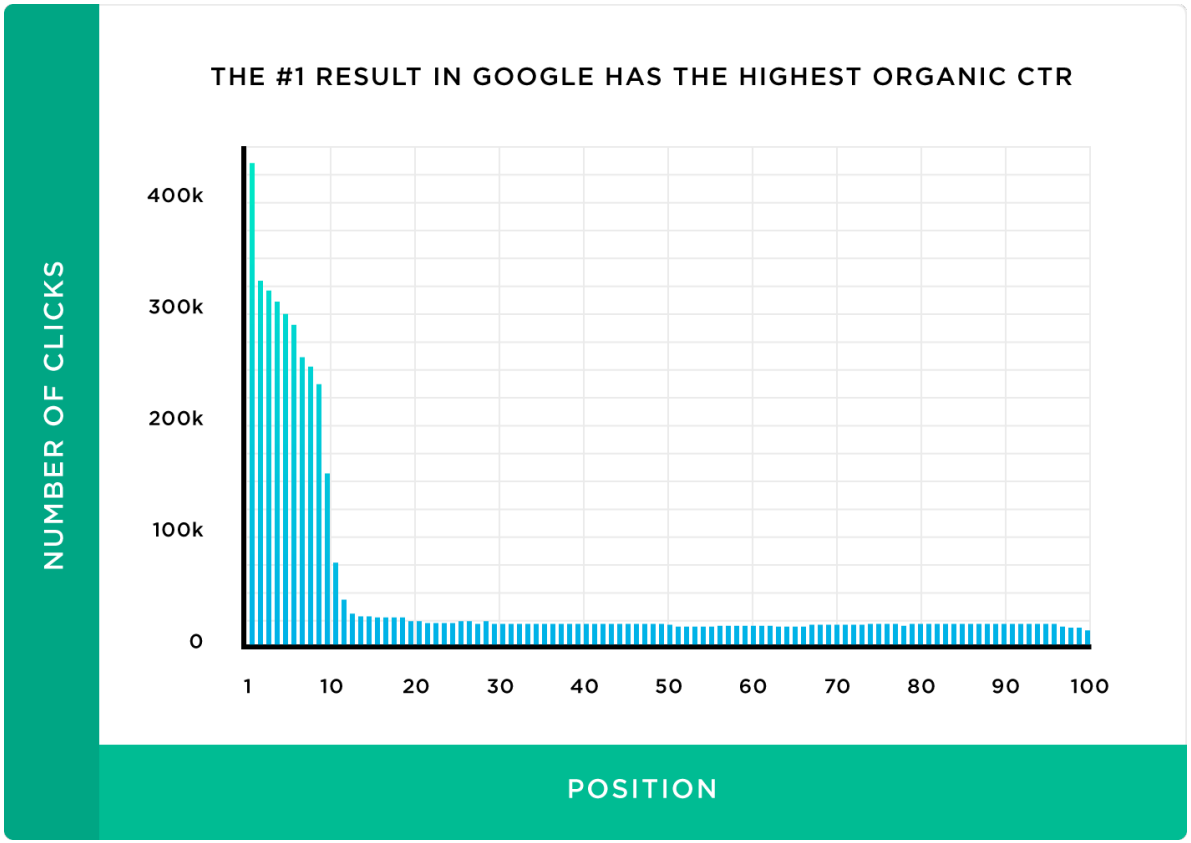
Information ranked ~~by~~ for us

↓

Incomplete Information

↓

We consume the tip of the iceberg



Taken together

- The Internet is our primary source of knowledge
- We have been mapping what we see to an elusive ground truth, but have not assessed what we do *not* see
- Uncritically and consistently consuming the tip of a pre-ranked iceberg
- Harms of representation

> *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)*

Objective

Knowing unknowns in an age of incomplete information[4]

Specifically, when accessing ranked information on the Internet:

1. Define metrics for information visibility
2. Understand implications of variation in information visibility on human behavior

[4] Old question (Plato 399 BC, Einstein 1931, Taleb 2007) but we have the data and methods now to *approximate* an answer.

Metrics for Information Visibility

Defining Metrics

Balancing Relevance and Visibility

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

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

Balancing Relevance and Visibility

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

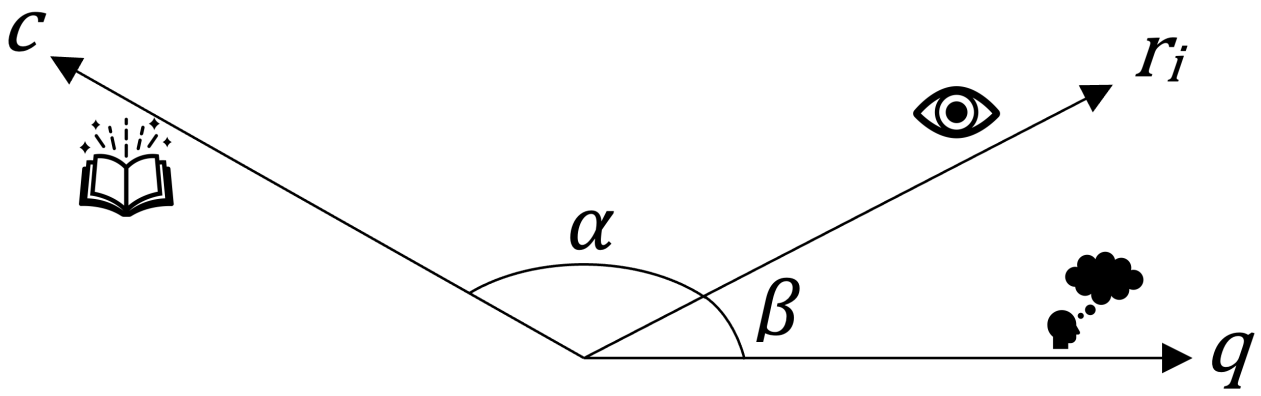
$$\text{Query } (q) \xrightarrow[\checkmark]{\text{What I want}} \text{Search result } (r_i)$$

What about *visibility*?

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

Defining Visibility

Using Text Embeddings



q: query

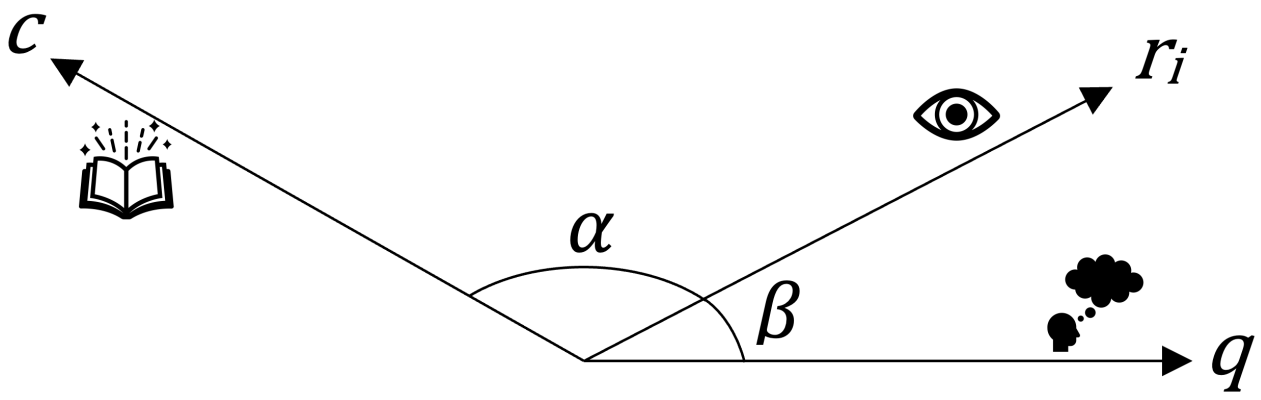
$C = \sum w_i r_i$: corpus constructed as weighted aggregate of r_i vectors

r_i: one of out n search results

w_i: weight assigned to each search result

$$I_{visibility} = \cos\alpha = \frac{\mathbf{C}\mathbf{r_i}}{\|\mathbf{C}\|\|\mathbf{r_i}\|}$$

Balancing Relevance and Visibility



On a similar note, $I_{relevance} = \cos\beta = \frac{\mathbf{qr_i}}{\|\mathbf{q}\|\|\mathbf{r_i}\|}$

Useful as we can reorder results by their S_i score, where λ controls the balance between relevance and visibility

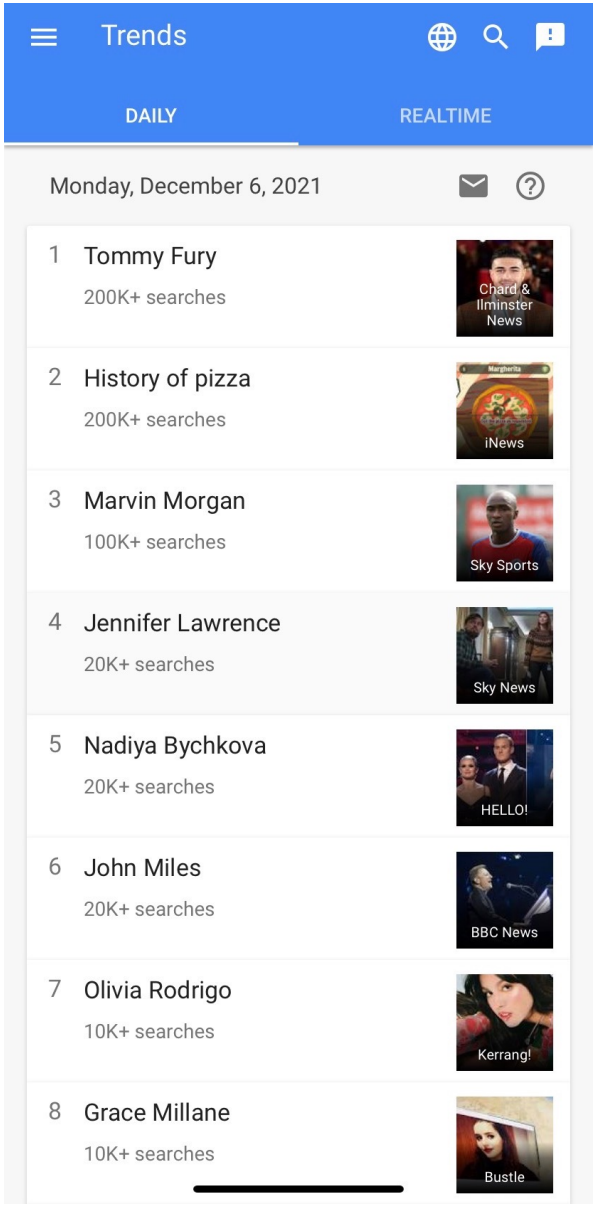
$$S_i = \lambda \frac{\mathbf{Cr_i}}{\|\mathbf{C}\|\|\mathbf{r_i}\|} + (1-\lambda)\frac{\mathbf{qr_i}}{\|\mathbf{q}\|\|\mathbf{r_i}\|} = \lambda I_{visibility} + (1-\lambda) I_{relevance}$$

Metrics for Information Visibility

Validating Metrics

Data

Generate visibility scores for worldwide search trends.



Data

Generate visibility scores for worldwide search trends.

Everyday[5]:

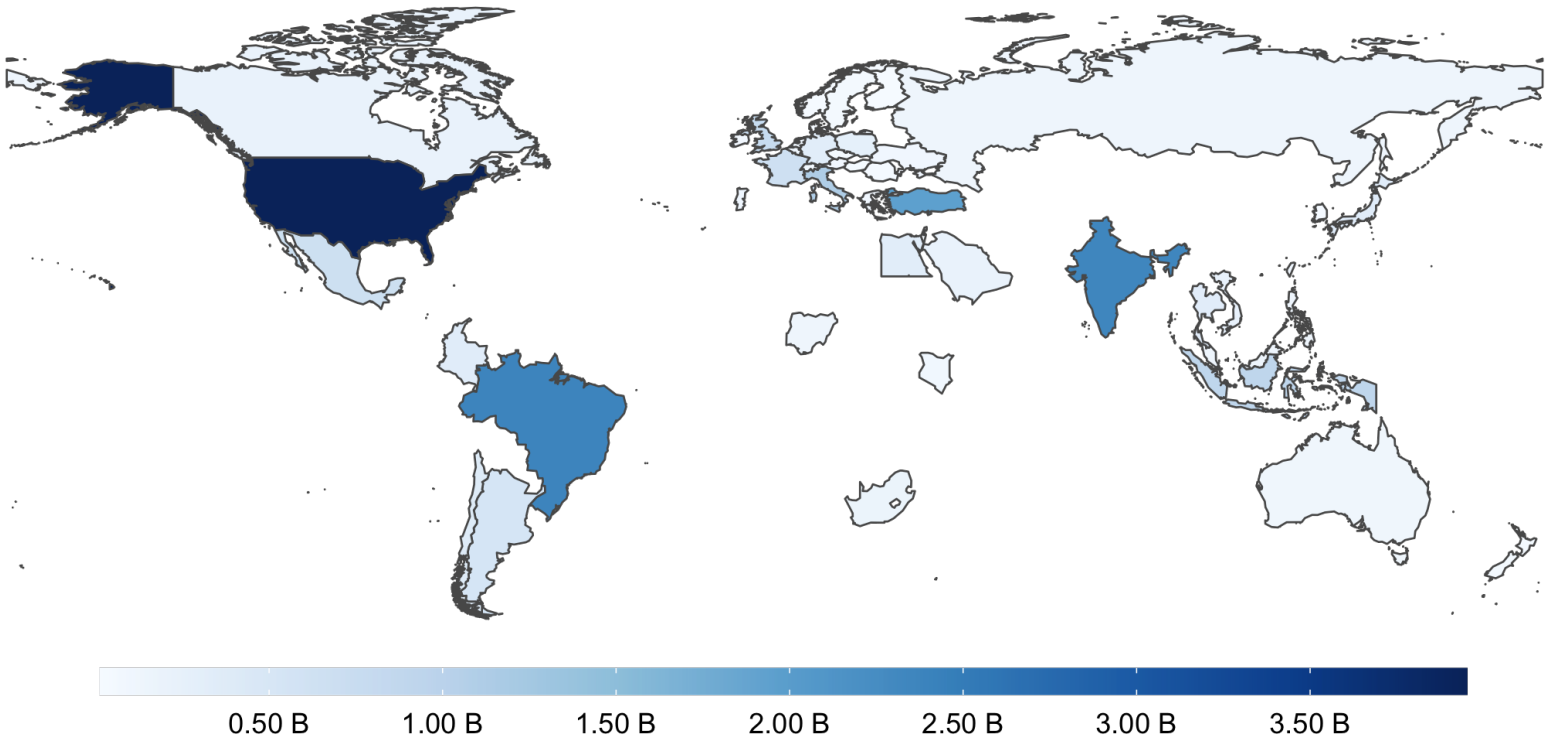
48 nations
$\times$
1.2 million searches/nation
$\times$
319 results/search[6]
$\approx$
18 billion daily data points

[5] reporting medians

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

Data

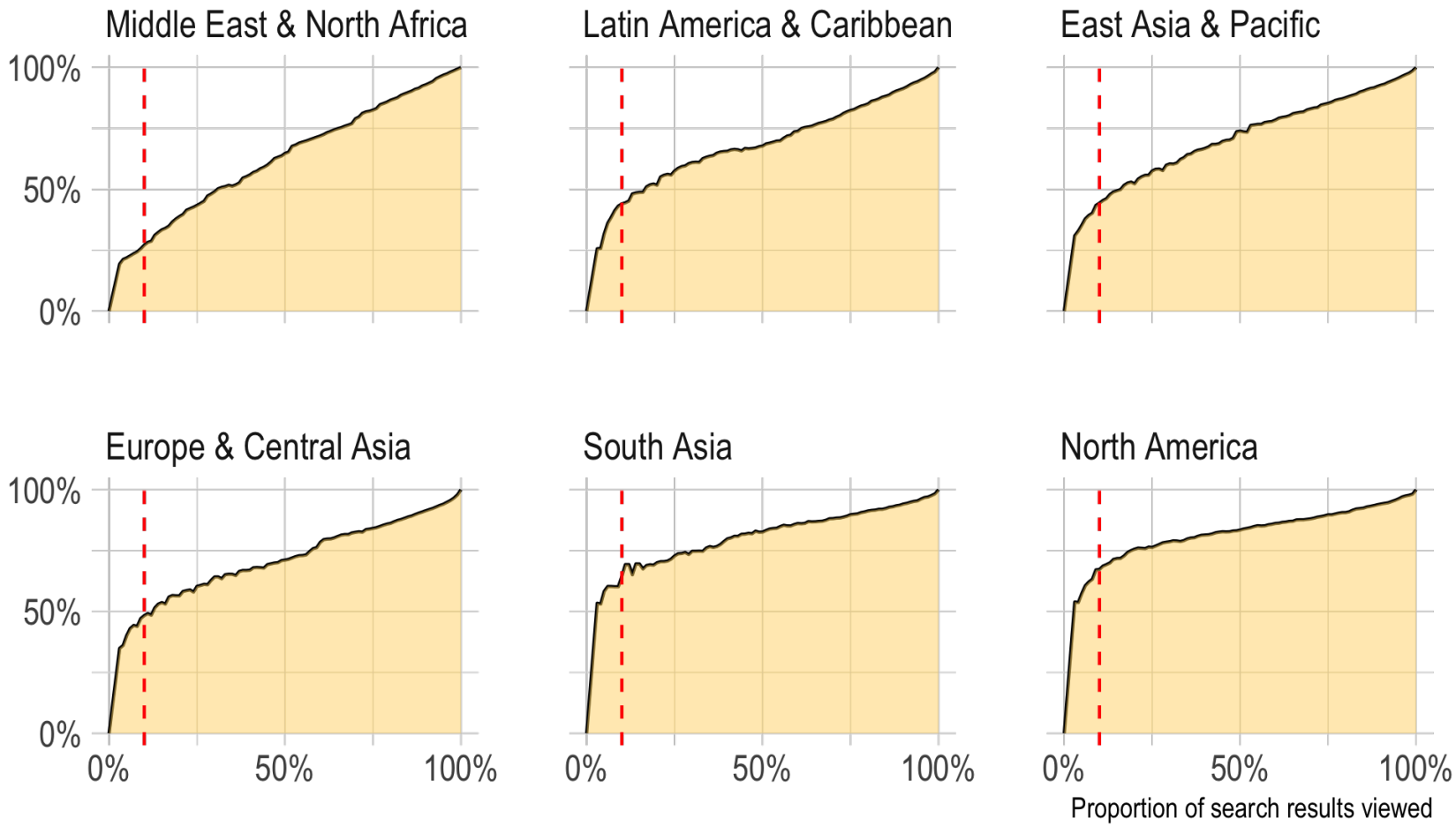
Daily Search Volume Fetched



Preliminary Results for Globally Trending Search Volume

Information Visibility Curves

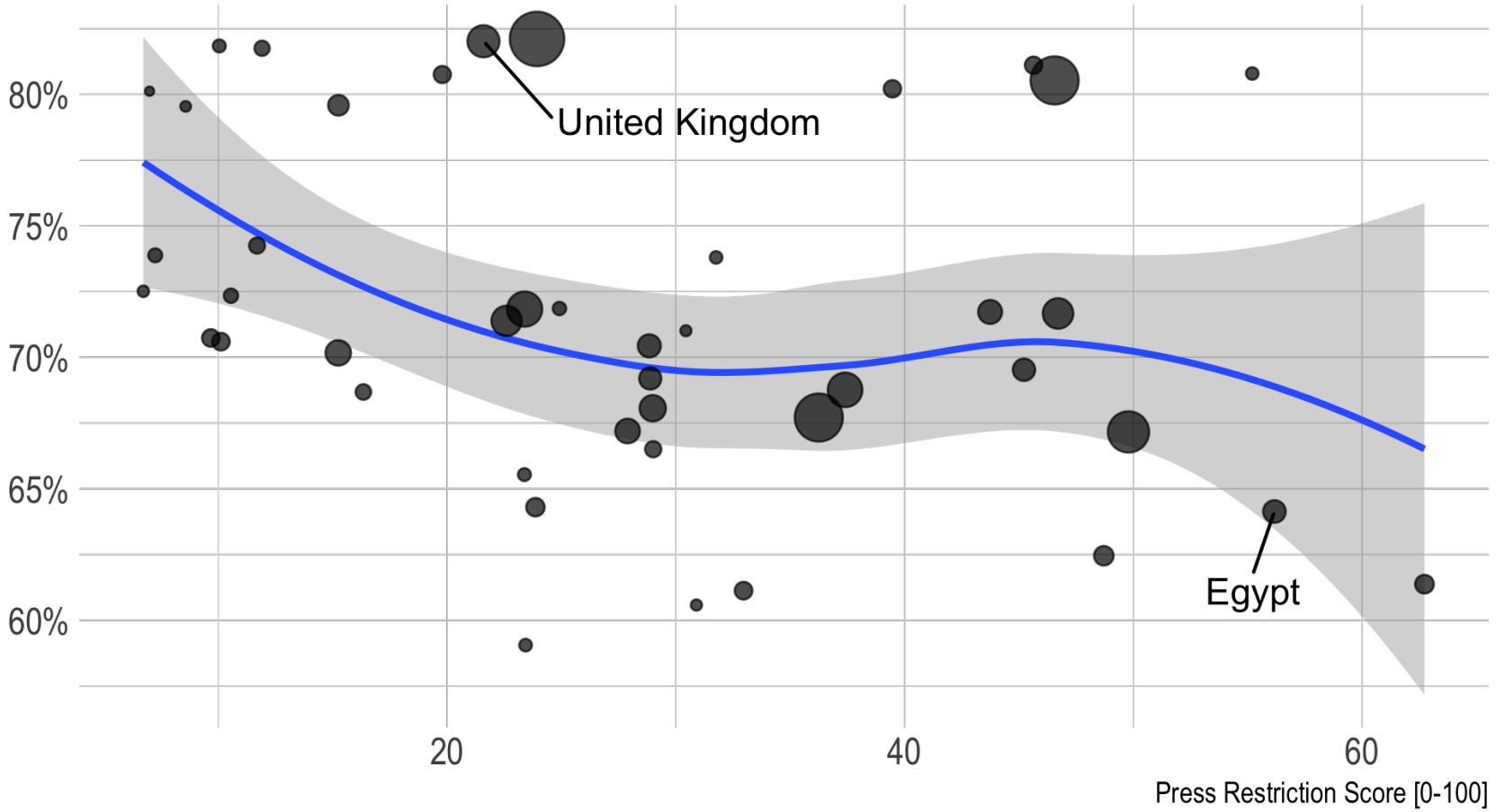
Split by region, Sorted by area under the curve[7]



[7] Dashed red line marks top 10 results

Information Visibility

Variation with Press Restrictions



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

Information Visibility

Variation with press restrictions

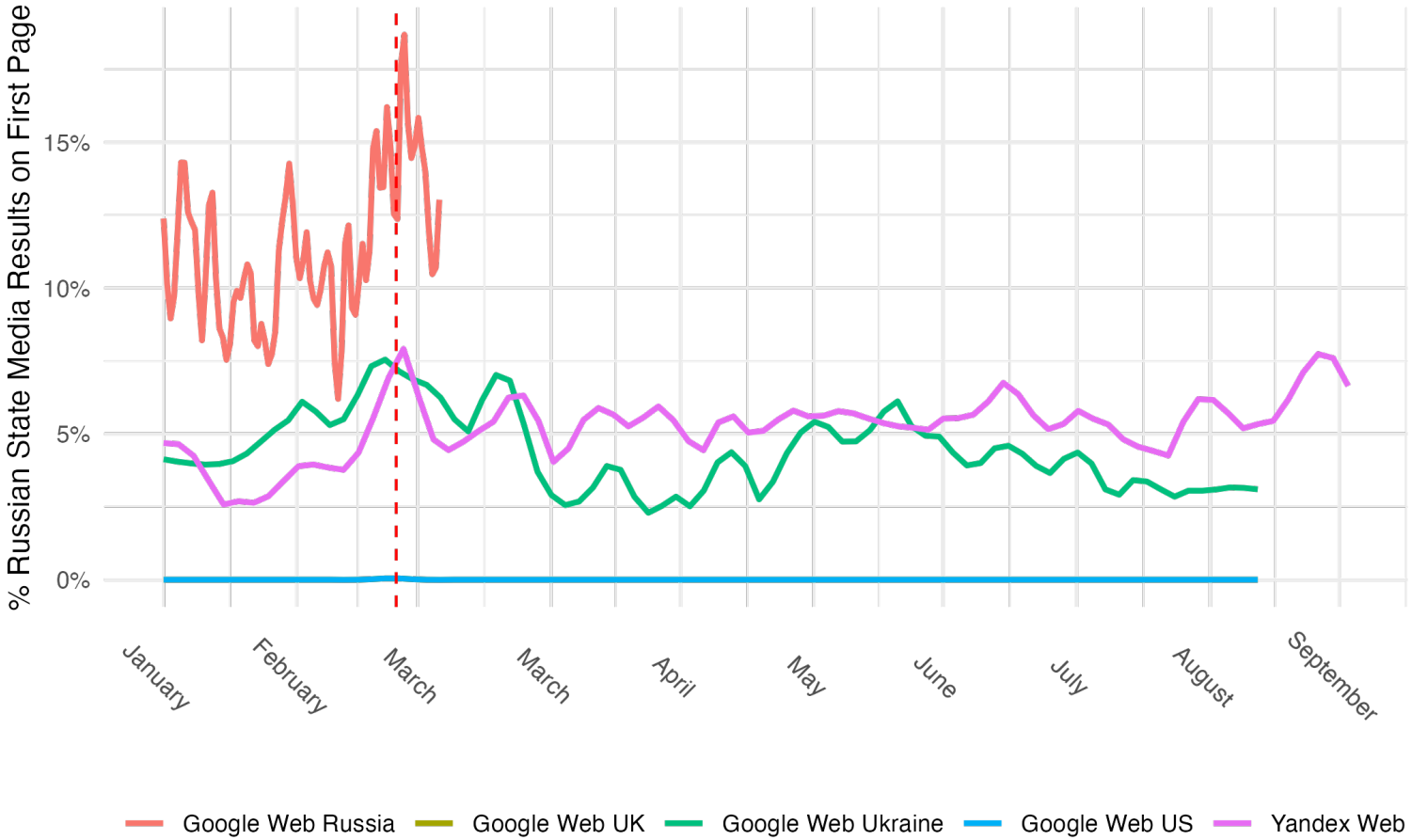
	Model 1	Model 2	Model 3	Time FE	Region FE[8]
(Intercept)	−0.00	−0.01	−0.00	−0.00	0.07***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Press restriction	−0.18***	−0.18***	−0.17***	−0.17***	−0.11***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Search volume		0.00***	0.00***	0.00***	0.00
		(0.00)	(0.00)	(0.00)	(0.00)
GDP per capita			0.01	0.01	0.10***
			(0.01)	(0.01)	(0.01)
Population			−0.00	−0.00	−0.19***
			(0.01)	(0.01)	(0.01)
Date				0.01	0.01
				(0.01)	(0.01)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Effect sizes in SD units.

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

Information Warfare[9]

Russian Invasion of Ukraine



[9] Erhardt, K., Khanna, S., South, T., Longpre, S., Schroeder, H., Roy, D., Pentland, A. (Under review). A Revolution in Information Warfare: Throttling, Deamplification, and Deplatforming.

Reflections

- As we consume the tip of a pre-ranked iceberg, it is crucial to 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 visibility is reached
- Regional differences
- Limitations
  - Not capturing the long tail of non-trending search queries
  - Not capturing information that was not indexed for web search