

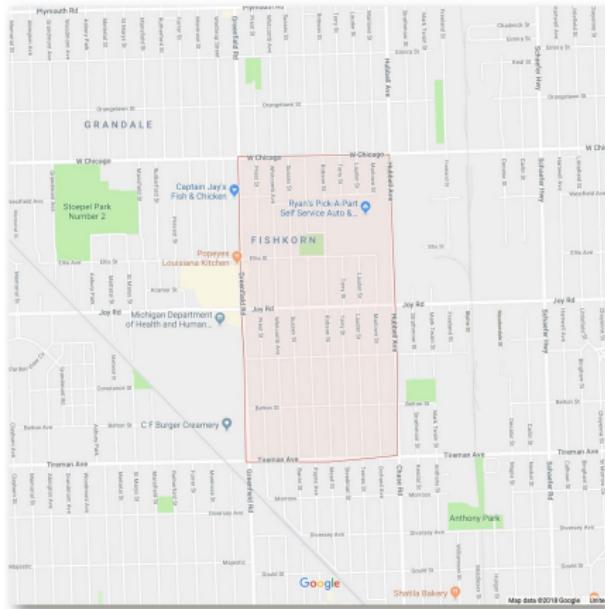
Invisible Knowledge and the Internet

Estimation and Implications

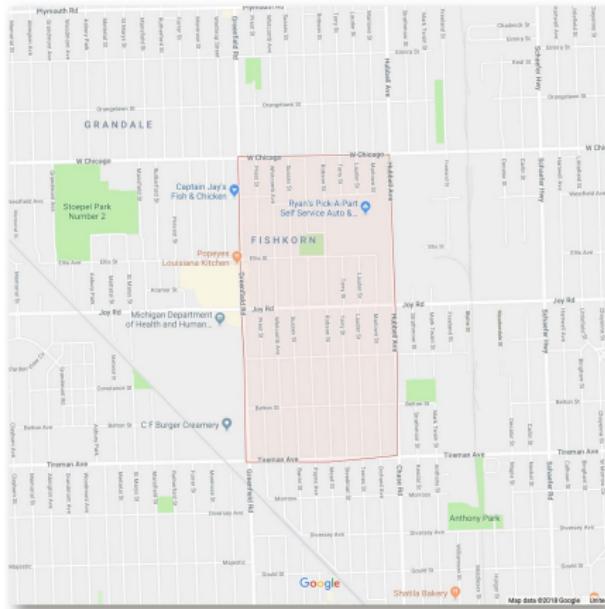
Saurabh Khanna

Stanford University

December 10, 2021

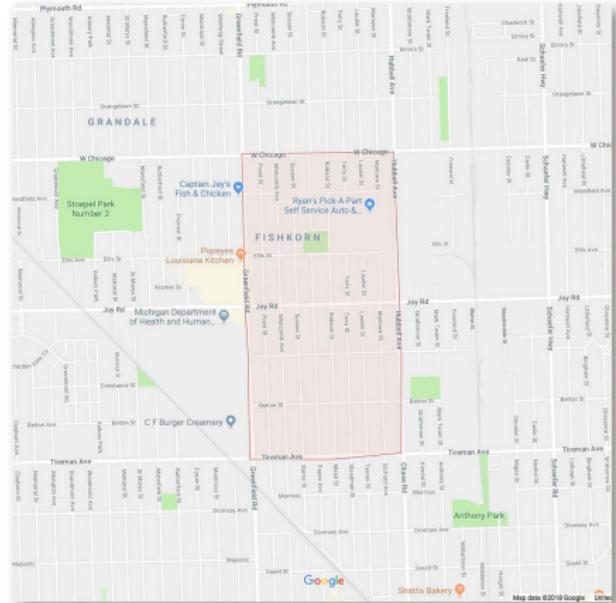


Fishhorn



Fisk-horn → Fish-korn

As Google Maps Renames Neighborhoods, Residents Fume



Fisk-horn → Fish-korn [2012-18] → Fisk-horn

(Independent) Knowledge Access

Peer
interaction/Libraries

(Independent) Knowledge Access

Peer
interaction/Libraries



Internet Search

Internet Search

- ▶ 1998: *Indexer of human knowledge*
- ▶ 2004–15:
 - ▶ Begin courting university libraries
 - ▶ Authors Guild, Inc. v. Google, Inc.
 - ▶ 40M books digitized → 130M
 - ▶ Queried 40,000 times every second
- ▶ 2019 – ?
 - ▶ 47% rise in broadband usage
 - ▶ 32% 62% US parents report teens' daily internet use exceeding four hours
- ▶ Present: *Provider of human knowledge*

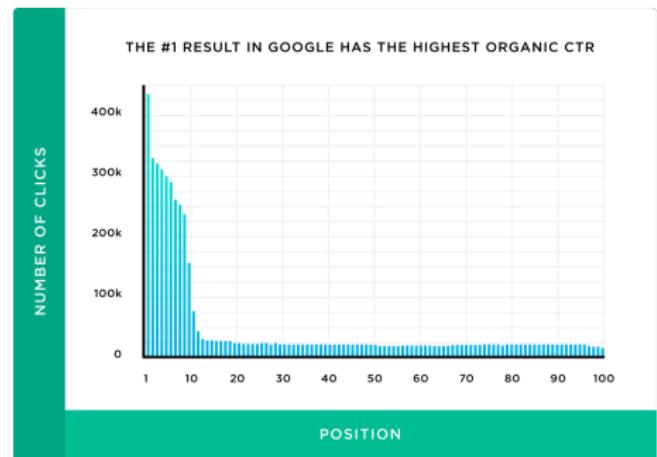
Three Problems

Problem 1

Much stays invisible

In the context of internet search¹:

- ▶ 91% of pages get no organic search traffic
- ▶ First result on page 2 receives only 1% of all clicks
- ▶ First page of search results receives 95% of all clicks



¹European Commission v Google 2017

Problem 2

professional haircut

All Images Shopping Videos News More Tools

medium length women's short business long hair fade modern curly men's

For Men (Business Hair...
hairmanz.com

Professional Hairstyles | ...
pinterest.com

Professional & Business...
haircutinspiration.com

Professional & Business...
haircutinspiration.com

Professional & Business...
haircutinspiration.com

For Men (Business Hair...
hairmanz.com

Business Professional Hair...
menshairstylestoday.com

70 Best Professional Ha...
machohairstyles.com

sional Hairstyl...
lay.com

Professional & Business...
thetrendspotter.net

Professional Hairstyle Id...
nextluxury.com

33 Hairstyles For Businessmen ...
menshairstyletrends.com

Professional haircut, Mens hair...
pinterest.com

70 Best Professional Hal...
machohairstyles.com

Mens Hairstyles and Haircuts ...
allthingshair.com

Related searches

hair style boys

formal professional medium hairstyles male

Problem 2

unprofessional haircut

All Images News Videos Shopping More Tools

The Guardian Opinions professional hair ...

Her Black Hair Became Poster Child for ... diversityinbestpractices.com

Do Google's 'unprofessional hair ... theguardian.com

If You Google 'Unprofessional ... mic.com

Google under fire over 'raci ... telegraph.co.uk

or "Unprofessional ..." Danny Sullivan on Twitt... twitter.com

Professional and unprofessional ... boingboing.net

Unprofessional hair' and Google ... wordtracker.com

Danny Sullivan on Twitt... twitter.com

Women Working from montecristomagazine.co

is - AVION

Photographs of male an...

20 Most Unprofessional Hairstyles for ...

Invisible Knowledge and the Internet

Unprofessional Work Hairstyles

What to Say When Someone...

Google search for ungr...

Saurabh Khanna (Stanford University)

December 10, 2021

9 / 27

Problem 2

Biased Results

Knowledge search on the internet is likely biased.²

- ▶ Targets representation consumption
- ▶ Trained on diverse mainstream data
- ▶ Trained by diverse homogeneous teams
- ▶ Public Proprietary access
- ▶ Speed and scale doesn't help here
- ▶ Illusion of fairness

²Nissenbaum & Introna 2000, Vaughan & Thelwall 2004, Eubanks 2018, Selbst et al. 2019

Problem 2

Biased Results

Knowledge search on the internet is likely biased.²

- ▶ Targets representation consumption
- ▶ Trained on diverse mainstream data
- ▶ Trained by diverse homogeneous teams
- ▶ Public Proprietary access
- ▶ Speed and scale doesn't help here
- ▶ Illusion of fairness

Knowledge is multidimensional.

- ▶ Results, even if perfectly sorted by 'relevance', may come at a cost.

²Nissenbaum & Introna 2000, Vaughan & Thelwall 2004, Eubanks 2018, Selbst et al. 2019

Problem 3

Choice and Accountability

- ▶ Restricts choice
 - ▶ Not easy to view the 'other side' of the spectrum
- ▶ Lacks accountability
 - ▶ Not much we can do about problematic search rankings

Problems

Taken together

Internet search → What we know, which

1. is the tip of the iceberg
2. is possibly biased
3. lacks choice and accountability

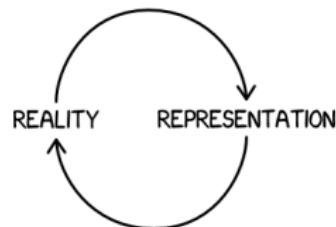
³Crawford 2017

Problems

Taken together

Internet search → What we know, which

1. is the tip of the iceberg
2. is possibly biased
3. lacks choice and accountability



Harms of representation³

If you control the flow of information in a society, you can influence its shared sense of right and wrong, fair and unfair, real and fake, known and unknown.

– Susskind, Future Politics (2018)

³Crawford 2017

Objectives

Know how much we do not know (and its implications).

- ▶ Not a new problem

I know that I know nothing.

– *Apology of Socrates, Plato (399 BC)*

- ▶ But we have both the data and the methods now to *approximate* an answer.

Research Questions

Specifically...

For an internet search query q (say ‘climate change’), along a given dimension of knowledge d (say ‘sentiment’):

1. How much do we not know? (a question of *visibility*)
2. How quickly do we know more? (a question of *efficiency*)
3. What are their implications (across space and time)?

Methods

Visibility Curves

For search query q along dimension d

Visibility is the overlap in dimension d 's distributions for top n vs N internet search results ($N \geq n$).

$$K_{\text{visibility}} = \int_{R_n} \min [f_n(d), f_N(d)] \, dd$$

Efficiency is the scaled area under the *visibility* curve.

$$K_{\text{efficiency}} = \int_{n=1}^N \int_{R_n} \min [f_n(d), f_N(d)] \, dd \, dn$$

$$K_{\text{visibility}}, K_{\text{efficiency}} \in [0, 1]$$

[Demo]

Developing an open-source platform



Supported by:

- ▶ Stanford Data Science
- ▶ Digital Economy Lab, Stanford Institute for Human-Centered Artificial Intelligence (HAI)
- ▶ Stanford Center for Open and Reproducible Science
- ▶ Stanford Transforming Learning Accelerator

Visibility Curves

Metric Benefits

Visibility:

- ▶ Probability that a randomly viewed result comes from the larger population
- ▶ Distribution agnostic
- ▶ Scaled. $K_{visibility} \in [0, 1]$
- ▶ Works for both continuous and categorical dimensions

Efficiency:

- ▶ All of the above
- ▶ A single number
 - ▶ Analogous to ROC curves

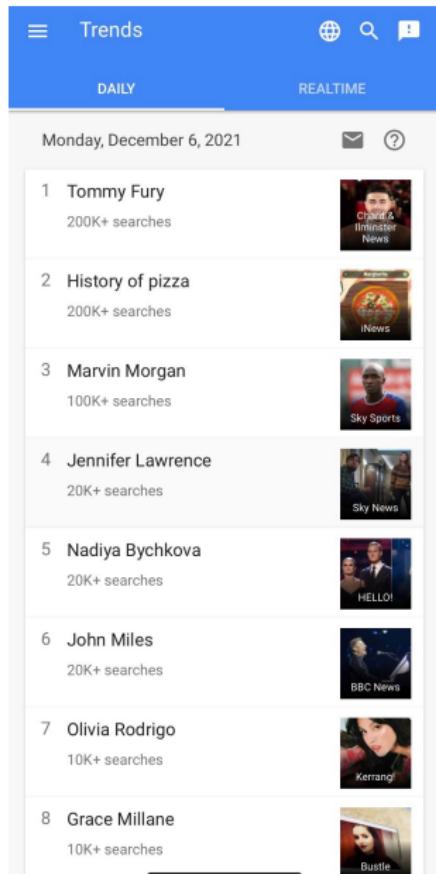
Data

Data

Build a dataset of *visibility* and
efficiency in worldwide search trends.

Data

Build a dataset of *visibility* and *efficiency* in worldwide search trends.



Data

Build a dataset of *visibility* and
efficiency in worldwide search trends.

Everyday⁴:

48 nations

×

20 trending queries/nation

×

308 results/query⁵

×

70,000 fetches/result

≈

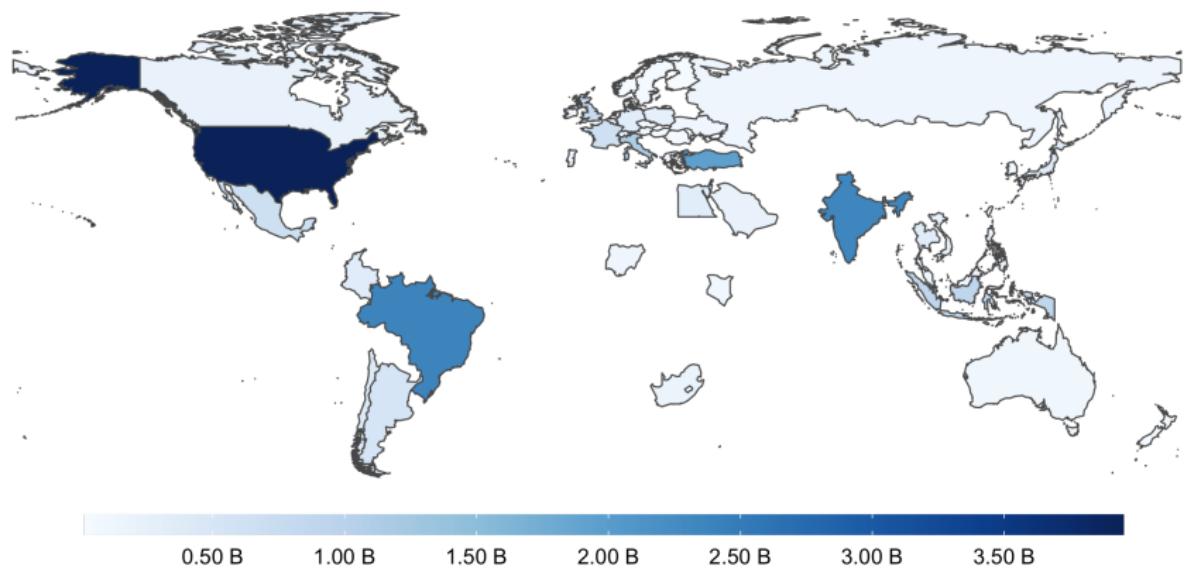
20 billion data points

⁴reporting medians

⁵Since 2016, Google caps the maximum search results shown to 300-400

Data

Daily Search Volume Fetched



Preliminary Results

[Available on request from saurabhkhanna@stanford.edu]

Reflections

- ▶ As internet search feeds us a (likely biased) tip of the iceberg, *visibility curves* can help assess how much we do not know.
- ▶ Sampling implications
 - ▶ Search results are *not* sorted by the dimension you care for.
 - ▶ Sample ~~top n results~~ results until a threshold visibility is reached.
- ▶ Limitations
 - ▶ Not capturing the long tail of non-trending search queries
 - ▶ Not capturing knowledge that was not indexed for web search

The way ahead

In the pipeline⁶:

1. Visibility of results from eco-friendly domains
 - 1.1 Carbon costs
2. Image search visibility
 - 2.1 Skin tone distributions
3. Press freedom
 - 3.1 Sentiment differentials between *web* and *news* trends
4. Compare search platforms
 - 4.1 Across countries
 - 4.2 Within countries [Russia]
5. Experiments [Is ignorance bliss?]
 - 5.1 Effect of knowledge visibility on tolerance and polarization

⁶1 and 2 in active development

Thank You!

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., & Intronà, 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.