



The Importance of Wikipedia to Search Engines (and other systems)

https://www.mediawiki.org/wiki/Wikimedia_Research/Showcase

Nicholas Vincent. Discussing work done with my PhD adviser, Brent Hecht.

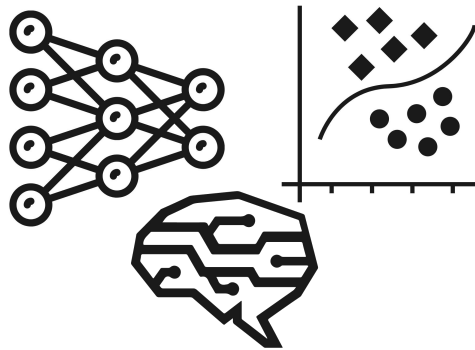
Contact: www.nickmvincent.com | nickvincent@u.northwestern.edu | [@nickmvincent](https://twitter.com/nickmvincent)



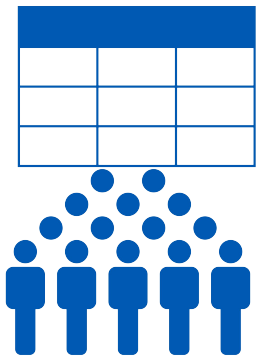
PSA Research Group NORTHWESTERN
People, Space, and Algorithms UNIVERSITY

Structure of this short talk

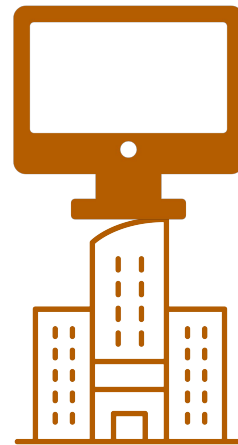
1. Describe a specific study about SERPs and Wikipedia (will be presenting at CSCW 2021)
 - a. If you were at WikiWorkshop 2020, this may sound familiar!
 - b. I'll go a bit fast on methods, but I'd love to chat more after if this sounds interesting!
2. Connect these phenomena to the (many) other examples of Wikipedia fueling AI / ML / data science



“Data Labor”



Algorithms &
Platforms



Intelligent
Technologies

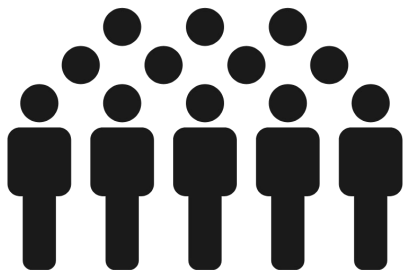
Why Study Data Labor?

- Economic concerns
 - Intelligent technologies linked with serious societal harms from inequality
- Recognize and dignify data labor
- Sustainability of peer production

Make people aware of value, to make it possible for them to leverage the value and create change



Why Study Data Labor?



Leverage value of data labor to create a computing paradigm where economic benefits and power are shared much more broadly

Could range from: paycheck for your data to more recognition/agency for Wikipedia (and similar communities)

So how **exactly** do Wikipedia and search engines fit into “data labor” research?



google verb

goo·gle | \ˈɡü-gəl ㄱㄹ\

variants: or **Google**

googled or **Googled**; **googling** \ˈɡü-g(ə-)lɪŋ ㄱㄹ\ or **Googling**; **googles** or **Googles**

Definition of *google*

transitive verb

: to use the Google search engine to obtain information about (someone or something) on the World Wide Web

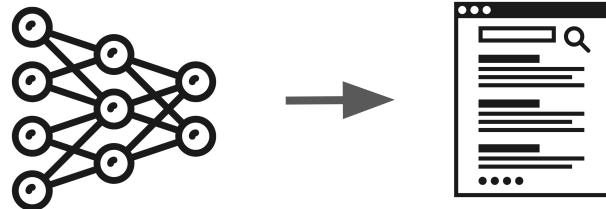
Use an online search engine to help
you find information on the Web

Current	91
May 2011	92
May 2010	87
April 2009 ¹³	88
May 2008	89
December 2006	91
August 2006	88
Dec 2005	91

Search engines

- ubiquitous
- hugely influential

Image from: Purcell, Kristen, Lee Rainie, and Joanna Brenner. "Search engine use 2012." (2012).



The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies

Connor McMahon^{1*}, Isaac Johnson^{2*}, and Brent Hecht²

^{*}indicates co-First Authors; ¹GroupLens Research, University of Minnesota;

²People, Space, and Algorithms (PSA) Computing Group, Northwestern University
mcmah250@umn.edu, isaacj@u.northwestern.edu, bhecht@northwestern.edu

Abstract

While Wikipedia is a subject of great interest in the computing literature, very little work has considered Wikipedia's important relationships with other information technologies like search engines. In this paper, we report the results of two deception studies whose goal was to better understand the critical relationship between Wikipedia and Google. These

broader information technology ecosystem. This ecosystem contains potentially critical relationships that could affect Wikipedia as much as or more than any changes to internal sociotechnical design. For example, in order for a Wikipedia page to be edited, it needs to be visited, and search engines may be a prominent mediator of Wikipedia visitation pat-



26% click
through
rate



14% click
through
rate

Measuring the Importance of User-Generated Content to Search Engines

Nicholas Vincent, Isaac Johnson, Patrick Sheehan, and Brent Hecht

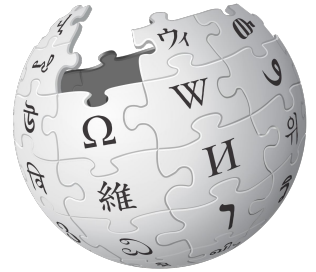
Northwestern University, Evanston, IL

{nickvincent, isaacj, PatrickSheehan2018}@u.northwestern.edu, bhecht@northwestern.edu

Wikipedia was a top source of content and appeared in 80-90% of results pages for some categories.

- not true for every category
- not always in the top 3 results

takeaway: Wikipedia is one of the most important
sources of results for search engines



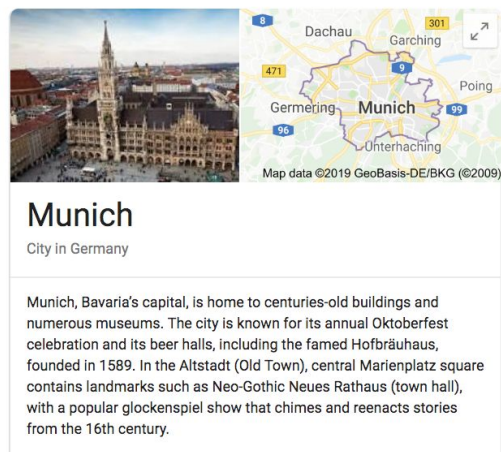
We had several generalization questions:

- What about search engines other than Google?
- What about mobile results?

Technical challenge:

- How do we handle the ever-changing Search Engine Results Pages (“**SERPs**”) for multiple search engines?
 - SERPs are no longer just “10 blue links”

SERPs are not just 10 blue links



Top stories



Transfer updates on Eden Hazard, Kylian Mbappe, and Andre Schurrle, plus more...

Bavarian Football Works

17 hours ago



Transfer news on Christian Eriksen, Sadio Mane, Theo Hernandez, Ante Rebic, Julian...

Bavarian Football Works

1 day ago



Joshua Kimmich: Bayern Munich Signing Leroy Sane 'Would Send Right Message'

Bleacher Report

1 day ago

→ More for munich news



cta issues

cta (@cta) · Twitter
<https://twitter.com/cta>

[Significant Delays] Some Harlem-bound Green Line trains are running with residual delays and congestion following an earlier police activity at King Drive. Service is resuming.

8 mins ago

→ View on Twitter

[Minor Delays] Harlem-bound Green Line trains are standing at King Drive due to police activity. Crews working to restore service. More: bit.ly/2Xbu9yF

14 mins ago

[Significant Delays] Howard-bound Red Line trains are running with residual delays following earlier mechanical issues on a train at 87th St. Service is resuming. More: bit.ly/2XdzvcP

20 mins ago

Knowledge Box, News Carousel, Twitter Carousel, etc.
Presumably very important to search

Methods

Search Engines, Devices, and Queries

- a. What search engines?
 - i. Google, Bing, and DuckDuckGo
- b. What devices to emulate?
 - i. Desktop and Mobile (we also considered the effect of different screen sizes)
- c. What queries to make?
 - i. **Critical and challenging**
 - ii. Our approach: multiple important categories, drawing on past work

Query selection

“common” queries (100 from search engine optimization company ahrefs.com)

e.g. “facebook”, “youtube”, “amazon”, “gmail”

“trending” queries (282 from Google trends)

e.g. “World Cup”, “thank u, next”, “What is fortnite”

“medical” queries (50 from prior research that shared Bing data)

e.g. “how to lose weight”, “indigestion”

See:

- Top Google searches (as of October 2019): 2019. <https://ahrefs.com/blog/top-google-searches>.
- <https://trends.google.com/trends/?geo=US>
- Soldaini, L. et al. 2016. Enhancing web search in the medical domain via query clarification. *Information Retrieval Journal*. 19, 1–2 (2016), 149–173.

Data collection

Our approach: use `puppeteer` (Node.js) to run headless Chrome

- We forked NikolaiT's amazing `se-scrapers` library:
<https://github.com/NikolaiT/se-scrapers>
- Our fork focuses on recording and analyzing link coordinates with the space of a single SERP

One approach for SERP scraping:

Researcher looks at SERP HTML

```
<div> ... </div>  
<div class="searchResult_abc123">  
  <a href="wikiworkshop.org"> Wiki  
  Workshop 2020</a>  
</div>  
<div> ... </div>
```

Write CSS rules to parse HTML page
into a ranked list

“find all elements with class of
searchResults_abc123”



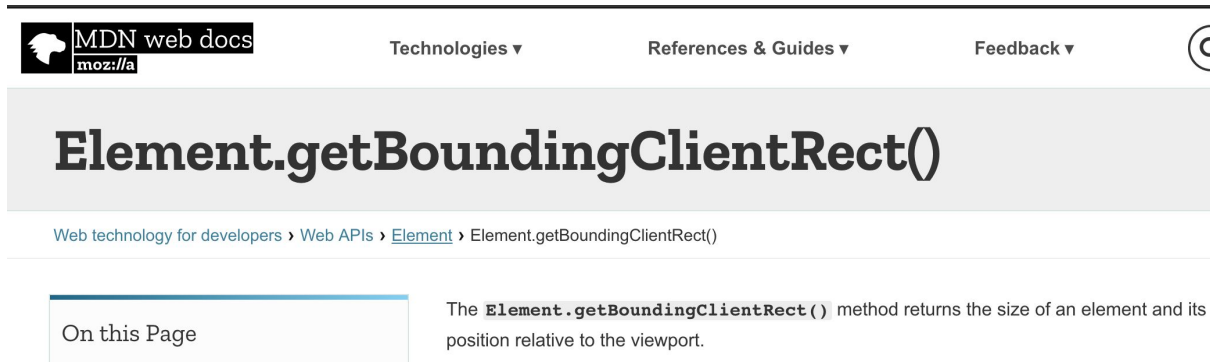
1. `wikiworkshop.org`
2. `twitter.com/wikiworkshop`

Spatial analysis: getting link coordinates

Get all the links (“a” elements) in a page:

```
await this.page.$$eval('a', getPos);
```

Calculate their position (x, y) with JavaScript:



The screenshot shows the MDN web docs page for the `Element.getBoundingBox()` method. The page header includes the MDN logo, navigation links for Technologies, References & Guides, and Feedback, and a Creative Commons license icon. The main heading is `Element.getBoundingBox()`. Below the heading is a breadcrumb trail: Web technology for developers > Web APIs > [Element](#) > `Element.getBoundingBox()`. On the left side, there is a section titled "On this Page". The main content area starts with a paragraph explaining that the `Element.getBoundingBox()` method returns the size of an element and its position relative to the viewport.

MDN web docs
moz://a

Technologies ▼ References & Guides ▼ Feedback ▼

Element.getBoundingBox()

Web technology for developers > Web APIs > [Element](#) > Element.getBoundingBox()

On this Page

The `Element.getBoundingBox()` method returns the size of an element and its position relative to the viewport.

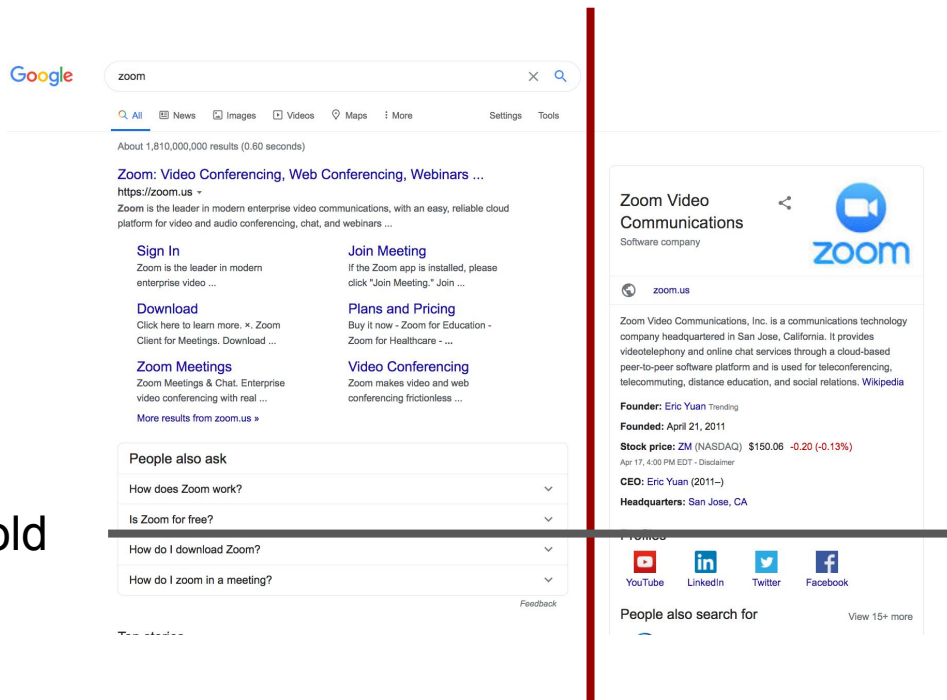
A quick note if you want to collect SERP data TODAY

- I later created a “minimal” script for collecting link coordinates:
<https://github.com/nickmvincent/LinkCoordMin>
 - Caveat 1: I’ve recently run into some issues with Bing
 - Caveat 2: Localization (Location spoofing) is inconsistent and requires quite a bit of fiddling
 - Please reach out if you’re interested!
- If you want to study Google using a ranked list approach, I suggest
<https://github.com/gitronald/WebSearcher>.
 - Great localization support (better than anything I’ve seen for puppeteer)

Spatial incidence rate definition

full page

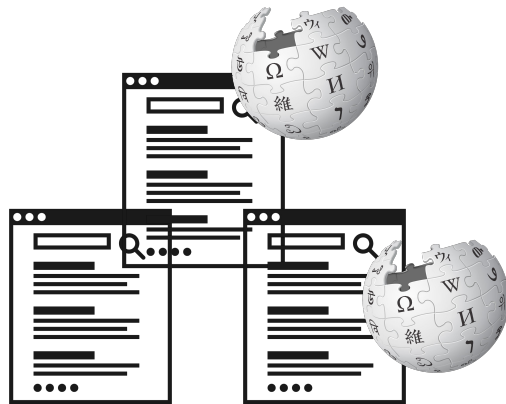
above-the-fold



left-hand and right-hand

Incidence rates

- how often is Wikipedia showing up in SERPs?
 - If we collect 3 SERPs, and Wikipedia appears in 2, incidence rate = $2 / 3$
- how often is Wikipedia showing up in prominent parts of SERPs?
 - if Wikipedia appear “above-the-fold” in only one of our 3 SERPs, above-the-fold incidence rate = $1 / 3$



Data validation - a tough task

SERP data changes constantly - remember this?



BUSINESS
INSIDER



Subscribe

Google is walking back changes to its search design that blurred the lines between ads and regular results after user backlash

Tyler Sonnemaker Jan 24, 2020, 1:31 PM

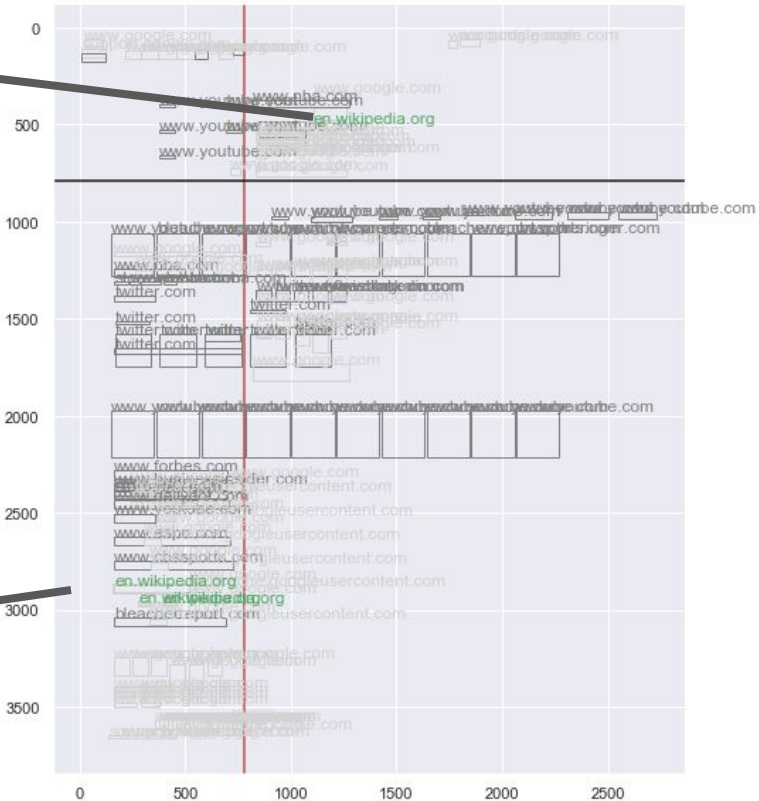
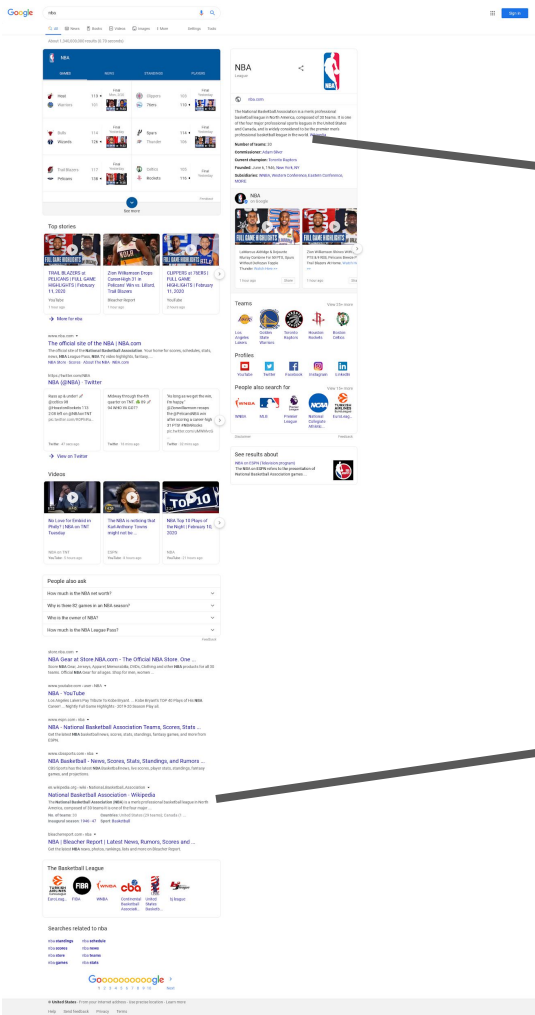


Data validation - visual inspection

Basic approach:

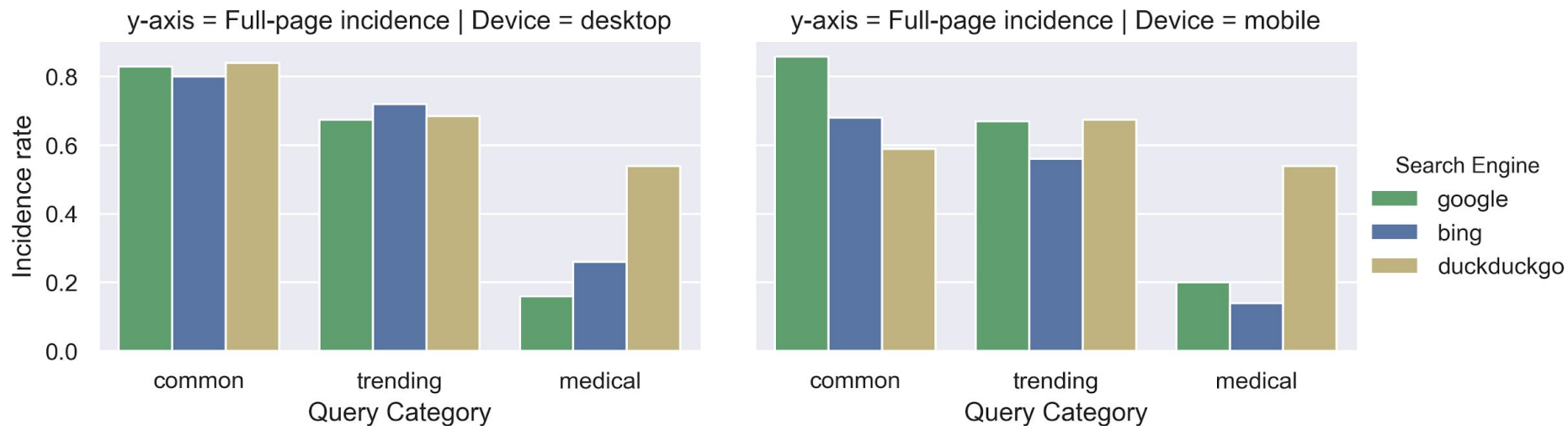
- save screenshots of SERPs
- visualize the analysis-ready data (i.e. the JSON files for quantitative analysis)
- make sure they seem to match up!

The screenshot displays a Google search for "nba". The top results include a news article about the NBA's temporary suspension due to the coronavirus pandemic, a tweet from Shams Charania about a trade, and the official NBA website. Below these, there are sections for "Top stories", "People also ask", and "Videos". The "Top stories" section features three articles: "Options for Jalen Green's NBA path", "Why Walt Disney World would be the ideal spot for the NBA to salvage its season", and "Jalen Green, top high school basketball prospect, bypasses the NCAA". The "People also ask" section lists questions like "Who is the lowest paid player in the NBA?", "How much is the NBA not worth?", "Who is the owner of NBA?", and "Which NBA team has not won a championship?". The "Videos" section shows three videos: "Top 10 plays of the 2019-20 NBA", "Best Of 'I'm On Fire' Game 1 (2019-2020)", and "Best Of 20-07 Games 1 (2019-20)".



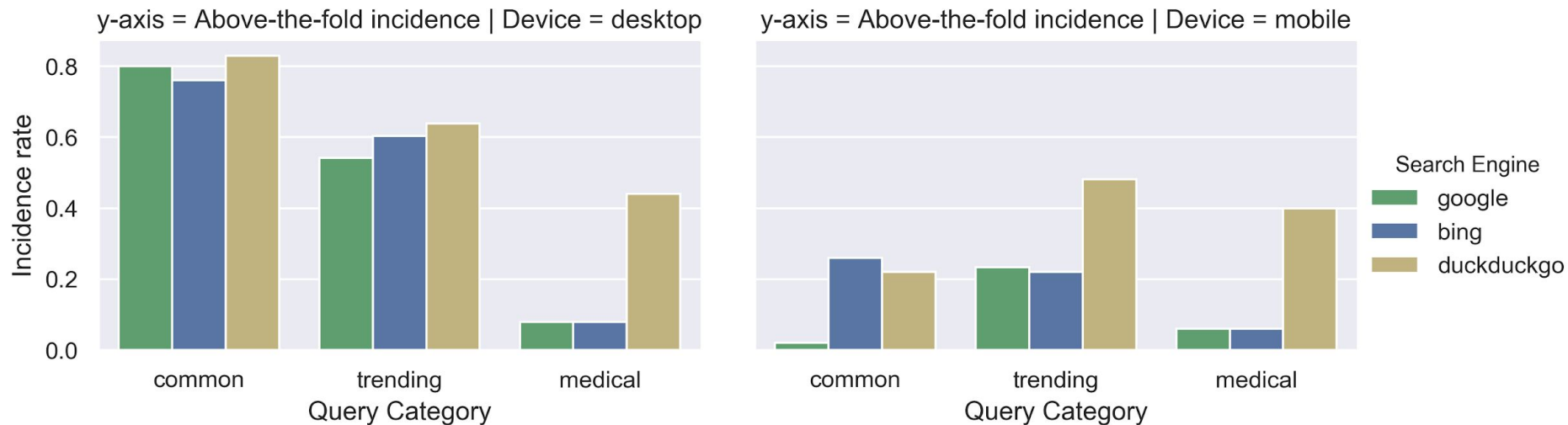
Results

Full page incidence rates



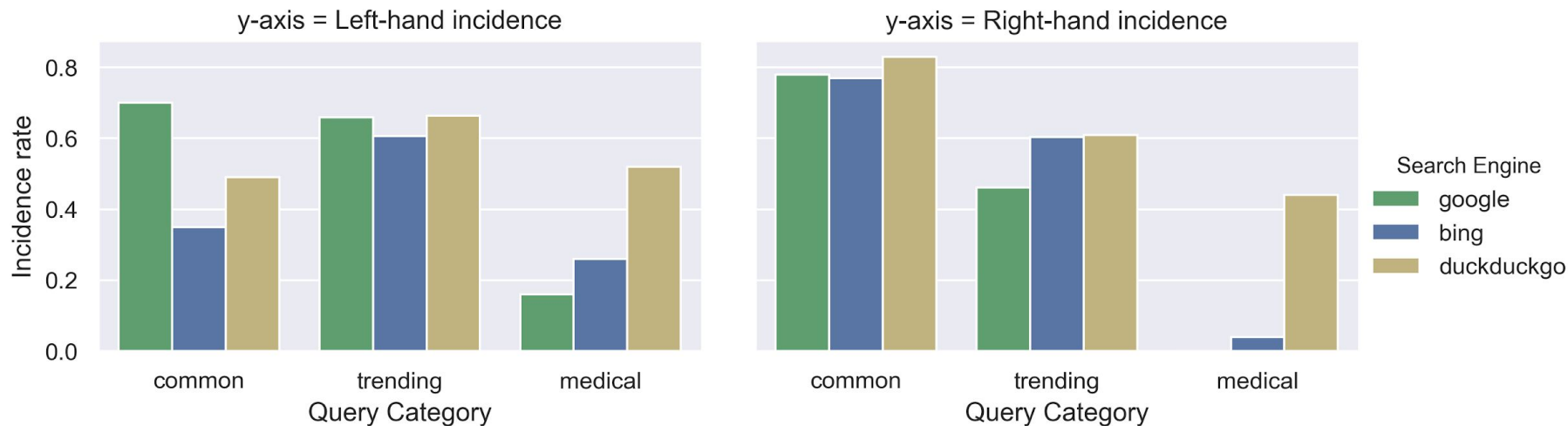
Very large incidence rates, but not for medical queries.

Above-the-fold incidence rates



Above-the-fold rate is much higher on desktop than mobile!

Left-hand and right-hand



Large right-hand incidence rates means Knowledge-panel style elements are still using a lot of Wikipedia links

Summary of findings

- Using the easy-to-understand (but limited) measure of incidence rates, **Wikipedia's importance to the success of search engines extends beyond Google and desktop-formatted search results**
- Queries and devices matter:
 - Wikipedia appears above the fold more often on desktop devices than mobile devices
 - Knowledge panel elements are a key source of Wikipedia content, but not the only sources

Limitations



- Audit study!
 - Small scale relative to actual query datasets
- Still US / en.wikipedia only
 - Wikipedia has geographic / language differences
- Queries matter immensely.
 - Incidence rate can lose meaning very quickly, e.g. if we append “wikipedia” to each query
 - Interpret accordingly

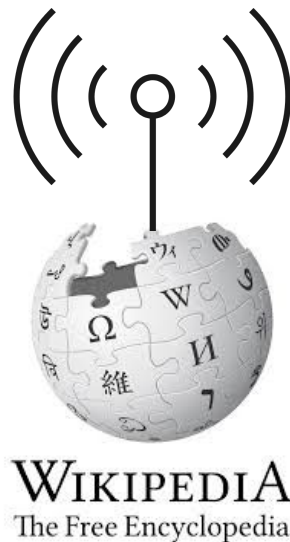
Wikipedia Matters Outside Wikipedia

Positive: effective peer production = effective search results?

Negative biases in coverage / quality = impact on search results?

(is it Wikipedia's "fault" if Google SERPs are bad? no, but...)

This raises the stakes of Wikipedia-focused research and Wikipedia findings



Data from the Public Fuels Intelligent Technologies

- Are Wikipedia editors some of the most important employees of search engines?



You **cannot** pay people to edit Wikipedia.

More prominently credit Wikipedia? Credit individual contributors? Solicit contributions? Donate to Wikipedia?

Two ways Wikipedia is used

The incidence rate results I presented mainly talk about how Wikipedia results are “served” by computing systems.

But this is just the tip of the iceberg, because Wikipedia also “trains” systems. Prominently, **language models** (“teaching a machine how to write”).

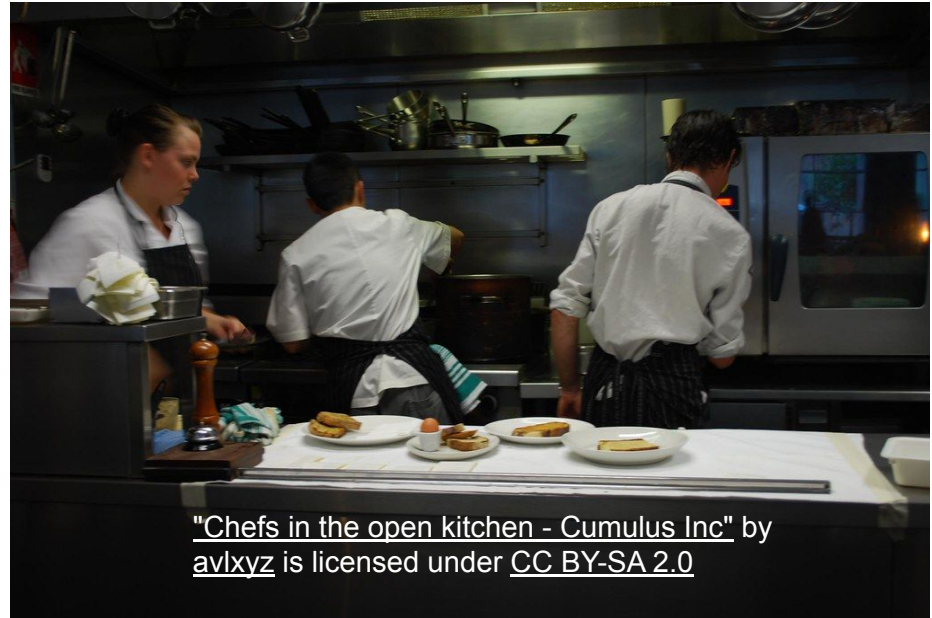
Wikipedia as a Meal

VS.

Wikipedia as a Recipe (Obtained Via Surveillance of an Open Kitchen)



"Steak Bowl" by Thai Yin is licensed under
CC BY-NC-ND 2.0



"Chefs in the open kitchen - Cumulus Inc" by
avlxyz is licensed under CC BY-SA 2.0

Auditing the Information Quality of News-Related Queries on the Alexa Voice Assistant

HENRY KUDZANAI DAMBANEMUYA, Northwestern University, USA

NICHOLAS DIAKOPOULOS, Northwestern University, USA

"Of all the understood responses, Wikipedia is the most prevalent individual information source, providing 18.6% of the responses. It is plausible to conclude that the reliability of these responses is only as reliable as Wikipedia"

<https://doi.org/10.1145/3449157>

YouTube Will Link Directly to Wikipedia to Fight Conspiracy Theories

After a series of scandals related to misinformation, YouTube CEO Susan Wojcicki announced the company would begin directing users to sources like Wikipedia.

<https://www.wired.com/story/youtube-will-link-directly-to-wikipedia-to-fight-conspiracies/>

Wikipedia articles as a meal (“served” by an AI or sociotechnical system)

Many systems! Any task that benefits from high quality knowledge about many topics stands to

- Search engines
- Voice Assistants
- YouTube
- Other social platforms (Reddit, StackExchange, more?)

Know some more?

Considering sharing them, here:

<https://github.com/nickmvincent/UGCValueRoundup/blob/main/wikipedia.md>

Wikipedia Content is Used Directly by an Intelligent Technology ("Wikipedia Articles as a Meal to be Served")

Wikipedia links to provide context

- "YouTube Will Link Directly to Wikipedia to Fight Conspiracy Theories" - Louise Matsakis in Wired, 2018. [Link](#)
 - Summary: YouTube will use Wikipedia links to provide context on "conspiracy theory videos".
 - Relevant Quote: "Here's how it will work: If you search and click on a conspiracy theory video about, say, chemtrails, YouTube will now link to a Wikipedia page that debunks the hoax alongside the video."
- "The Effects of an Informational Intervention on Attention to Anti-Vaccination Content on YouTube" - Sangyeon Kim et al. ICWSM, 2020. [Link](#)
 - Summary: A research study of the effects of YouTube's intervention.
 - Quote: "We find that this informational intervention reduced traffic to the affected videos, both overall, and in comparison to a carefully-matched set of control videos that did not receive the informational modification."
- "Facebook outsources its fake news problem to Wikipedia—and an army of human moderators" - Kerry Flynn in Mashable, 2017. [Link](#)
 - Summary: Facebook is also using Wikipedia links to provide context about entities who publish news content.
 - Quote: "Now, when Facebook users see articles on News Feed, they can click on a little 'i' button and see the Wikipedia description of the publisher."
- "Facebook Adds Wikipedia Knowledge Boxes in Search Results" - Andrew Hutchinson in SocialMediaToday, 2020. [Link](#)
 - Summary: Facebook adds a search results-esque "Knowledge Panel", drawing on Wikipedia data.
 - Quote: "The addition is very similar to Google's Knowledge Panels, which it launched back in 2012, and are also largely populated by Wikipedia info. Knowledge Panel listings can serve various purposes, including providing contextual insight, keeping users on platform for longer, and - for those that have them - adding more authority to your presence."

Wikipedia links answer search queries

- "Auditing the Information Quality of News-Related Queries on the Alexa Voice Assistant." - Henry Kudzanai Dambanemuya and Nicholas Diakopoulos in CSCW 2021. [Link](#)
 - Quote: "Of all the understood responses, Wikipedia is the most prevalent individual information source, providing 18.6% of the responses. It is plausible to conclude that the reliability of these responses is only as reliable as Wikipedia"
- "Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages" - Ronald E. Robertson, David Lazer, and Christo Wilson in The Web Conference (WWW) 2018. [Link](#)
 - Quote: "In terms of URL domains, we found that knowledge components typically do not feature a URL (Table 2), and when they do it is often a link to Wikipedia, confirming a finding from a previous audit"
- "A Deeper Investigation of the Importance of Wikipedia Links to the Success of Search Engines" - Nicholas Vincent and Brent Hecht in CSCW, 2021. [Link](#)
 - Quote: "Our findings reinforce the complementary notions that (1) Wikipedia content and research has major impact outside of the Wikipedia domain and (2) powerful technologies like search engines are highly reliant on free content created by volunteers."

Wikipedia is everywhere on the web!

Examples of studies that study how Wikipedia text is re-used and how Wikipedia links add value to other online platforms.

- "Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia's relationships with other large-scale online communities" - Nicholas Vincent, Isaac Johnson, and Brent Hecht in CHI 2018. [Link](#)
 - Quote: "This paper extends this recent research beyond search engines to examine Wikipedia's relationships with large-scale online communities, Stack Overflow and Reddit in particular. Wikipedia provides substantial value to both communities, with Wikipedia content increasing visitation, engagement, and revenue..."
- "Wikipedia Text Reuse: Within and Without" - Milad Alshomary et al. in ECIR 2019. [Link](#)
 - Quote: "We further report on a pilot analysis of the 100 million reuse cases inside, and the 1.6 million reuse cases outside Wikipedia that we discovered. Text reuse inside Wikipedia gives rise to new tasks such as article template induction, fixing quality flaws, or complementing Wikipedia's ontology. Text reuse outside Wikipedia yields a tangible metric for the emerging field of quantifying Wikipedia's influence on the web."

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

<https://arxiv.org/abs/2005.14165>

Table 2.2: Datasets used to train GPT-3. “Weight in training mix” refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen 0.44 times.

Reading Wikipedia to Answer Open-Domain Questions

<https://arxiv.org/pdf/1704.00051.pdf>

Danqi Chen*

Computer Science

Stanford University

Stanford, CA 94305, USA

danqi@cs.stanford.edu

Adam Fisch, Jason Weston & Antoine Bordes

Facebook AI Research

770 Broadway

New York, NY 10003, USA

{afisch, jase, abordes}@fb.com

ImageNet Large Scale Visual Recognition Challenge

To collect a highly accurate dataset, we rely on humans to verify each candidate image collected in the previous step for a given synset. This is achieved by using Amazon Mechanical Turk (AMT), an online platform on which one can put up tasks for users for a monetary reward. With a global user base, AMT is particularly suitable for large scale labeling. In each of our labeling tasks, we present the users with a set of candidate images and the definition of the target synset (including a link to Wikipedia). We then ask the users to verify whether each image contains objects of the synset. We encourage users to select images regardless of occlusions, number of objects and clutter in the scene to ensure diversity.

GPT-3 from OpenAI: >900

Question Answering from Facebook: >900

BERT from Google: >19,000

ImageNet 2015: >23000

Open Question

I think one can make a strong argument that Wikipedia is one of **the most important resource** for open machine learning research?

Has this been acknowledged

- not enough
- too much
- just the right amount?

On the horizon: https://meta.wikimedia.org/wiki/Wikimedia_Enterprise

Fun blog post idea

If “Wikipedia” were treated as a collective “scientific author” entity, where might it rank? How does the funding tech companies give to Wikipedia compare to the funding they give to “superstar” academics and technologists?

Many thanks to:

co-author Brent Hecht, thoughtful reviewers and colleagues

Open Software

- se-scraper / puppeteer
- numpy / pandas / seaborn / scipy ecosystem

Questions?

@nickmvincent on Twitter or nickvincent@u.northwestern.edu

Links:

paper: <https://www.nickmvincent.com>

se-scraper: <https://github.com/NikolaiT/se-scraper>

our se-scraper fork: <https://github.com/nickmvincent/se-scraper>

updated scraping repo: <https://github.com/nickmvincent/LinkCoordMin>

UGC Value Roundup:

<https://github.com/nickmvincent/UGCValueRoundup/blob/main/wikipedia.md>