A Deeper Investigation of the Importance of Wikipedia Links to the Success of Search Engines

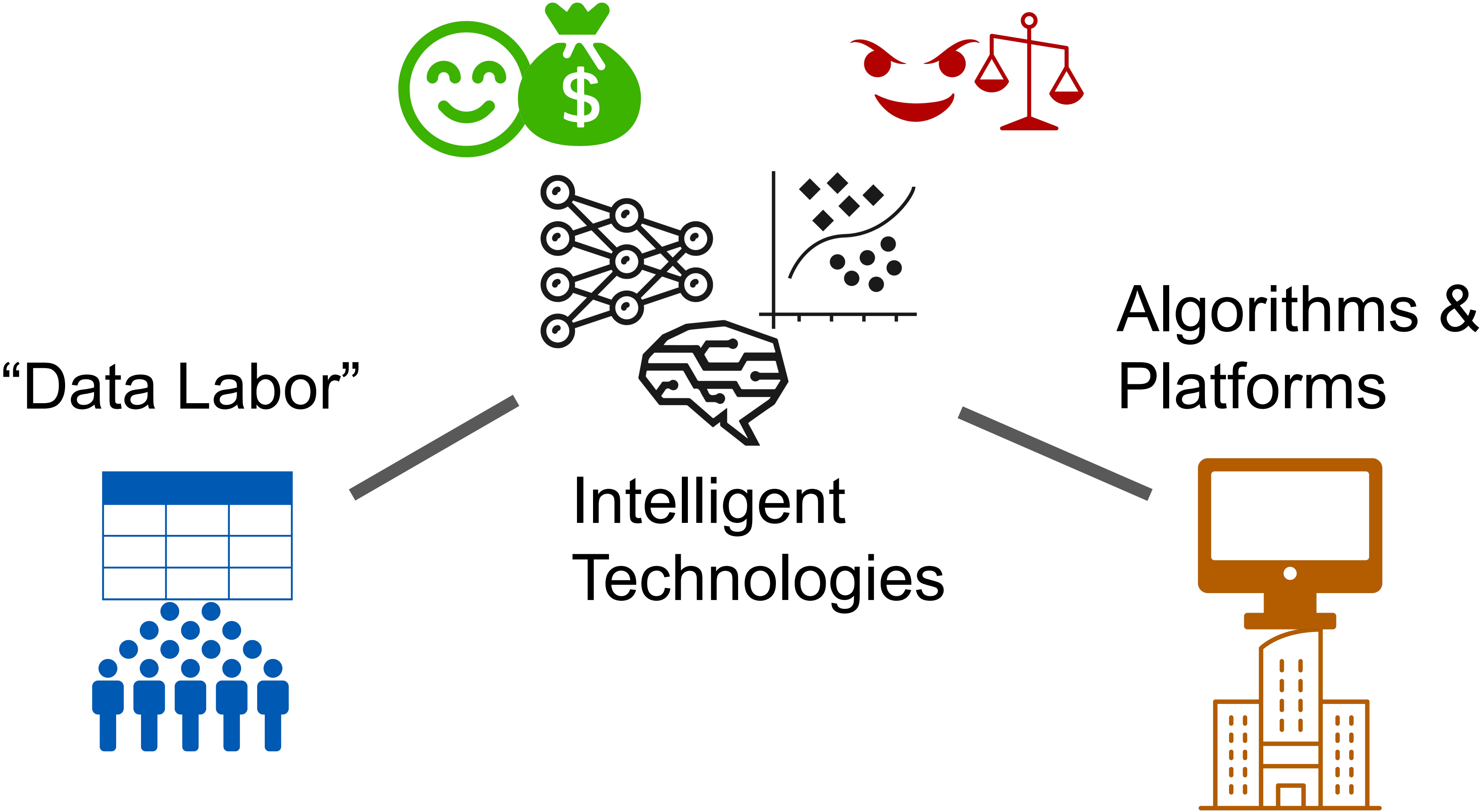
https://wikiworkshop.org/2020/

Nicholas Vincent and Brent Hecht.

Contact: www.nickmvincent.com | nickvincent@u.northwestern.edu | @nickmvincent

PSA Research Group
People, Space, and Algorithms

NORTHWESTERN
UNIVERSITY



“Data Labor”

Intelligent Technologies

Algorithms & Platforms

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?

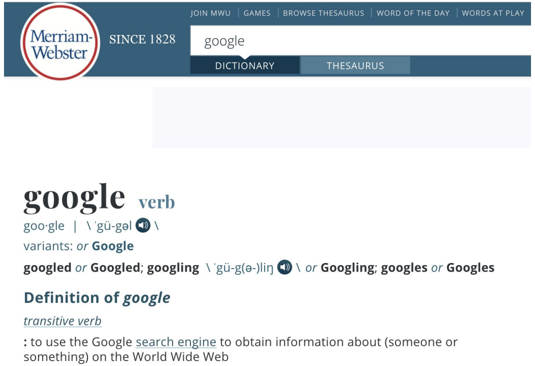
Leverage value of data labor to create a computing paradigm where economic benefits and power are shared much more broadly

Could range from: paycheck for your data to more recognition/agency for Wikipedia

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

Search engines

- ubiquitous
- hugely influential



google verb

goo·gle | \ˈgü-gəl\

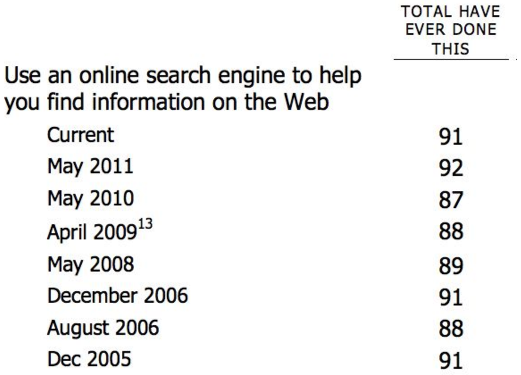
variants: *or* **Google**

googled *or* **Googled**; **googling** \ˈgü-g(ə-)liŋ\ *or* **Googling**; **googles** *or* **Googles**

Definition of *google*

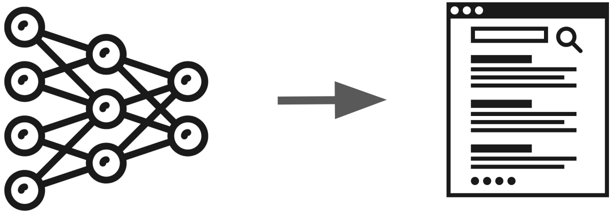
transitive verb

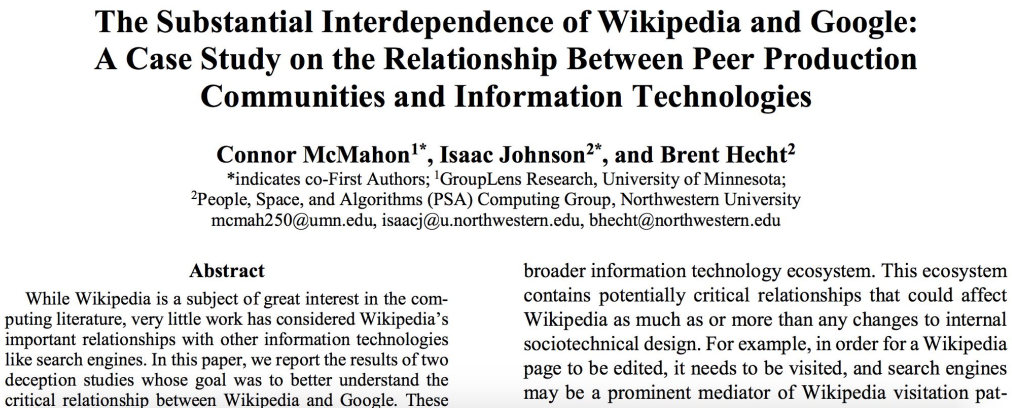
: to use the Google search engine to obtain information about (someone or something) on the World Wide Web



	TOTAL HAVE EVER DONE THIS
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

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





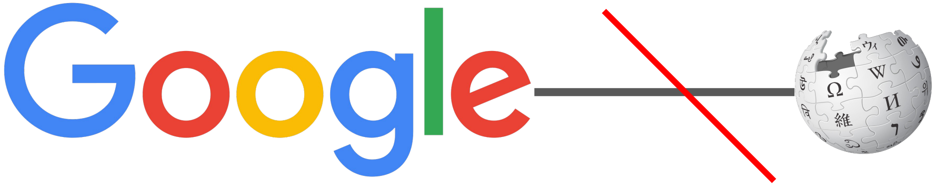
The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies

Connor McMahon[1*], Isaac Johnson[2*], and Brent Hecht[2]
*indicates co-First Authors; [1]GroupLens Research, University of Minnesota;
[2]People, Space, and Algorithms (PSA) Computing Group, Northwestern University
mcmah250@umn.edu, isaacj@u.northwestern.edu, bhecht@northwestern.edu

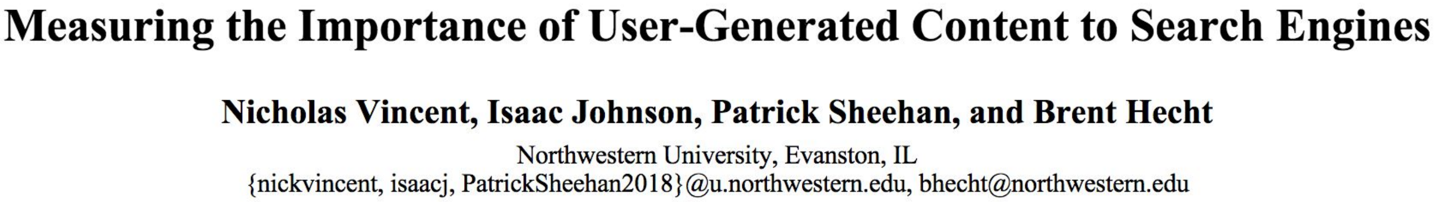
Abstract

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 cateogires.

- 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**

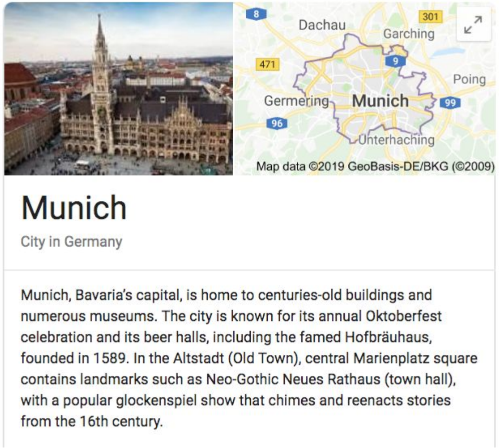
We had several generalization questions:

- 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

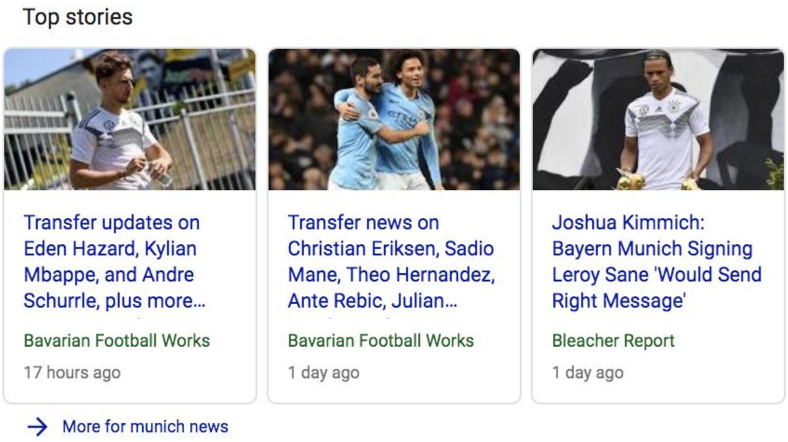


Map data ©2019 GeoBasis-DE/BKG (©2009)

Munich

City in Germany

Munich, Bavaria's capital, is home to centuries-old buildings and numerous museums. The city is known for its annual Oktoberfest celebration and its beer halls, including the famed Hofbräuhaus, founded in 1589. In the Altstadt (Old Town), central Marienplatz square contains landmarks such as Neo-Gothic Neues Rathaus (town hall), with a popular glockenspiel show that chimes and reenacts stories from the 16th century.



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

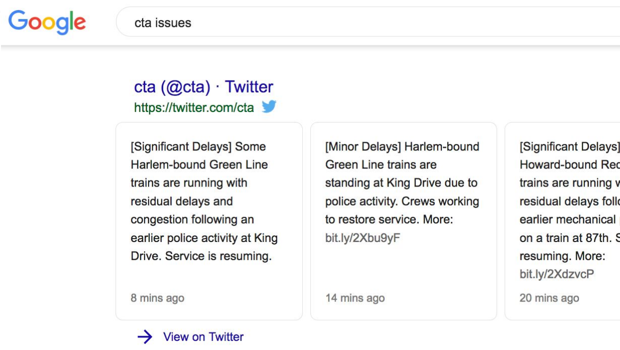
1 day ago

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

Bleacher Report

1 day ago

→ More for munich news



Google 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

[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 trains are running v residual delays follo earlier mechanical on a train at 87th. S resuming. More: bit.ly/2XdzvcP

20 mins ago

→ View on Twitter

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) run “headless” Chrome

- We forked NikolaiT’s `se-scraper` library: https://github.com/NikolaiT/se-scraper
- Our version focuses on recording and analyzing link coordinates with the space of a single SERP
- I’ll repost code links on the final slide

One approach for SERP scraping:

Researcher looks at SERP HTML

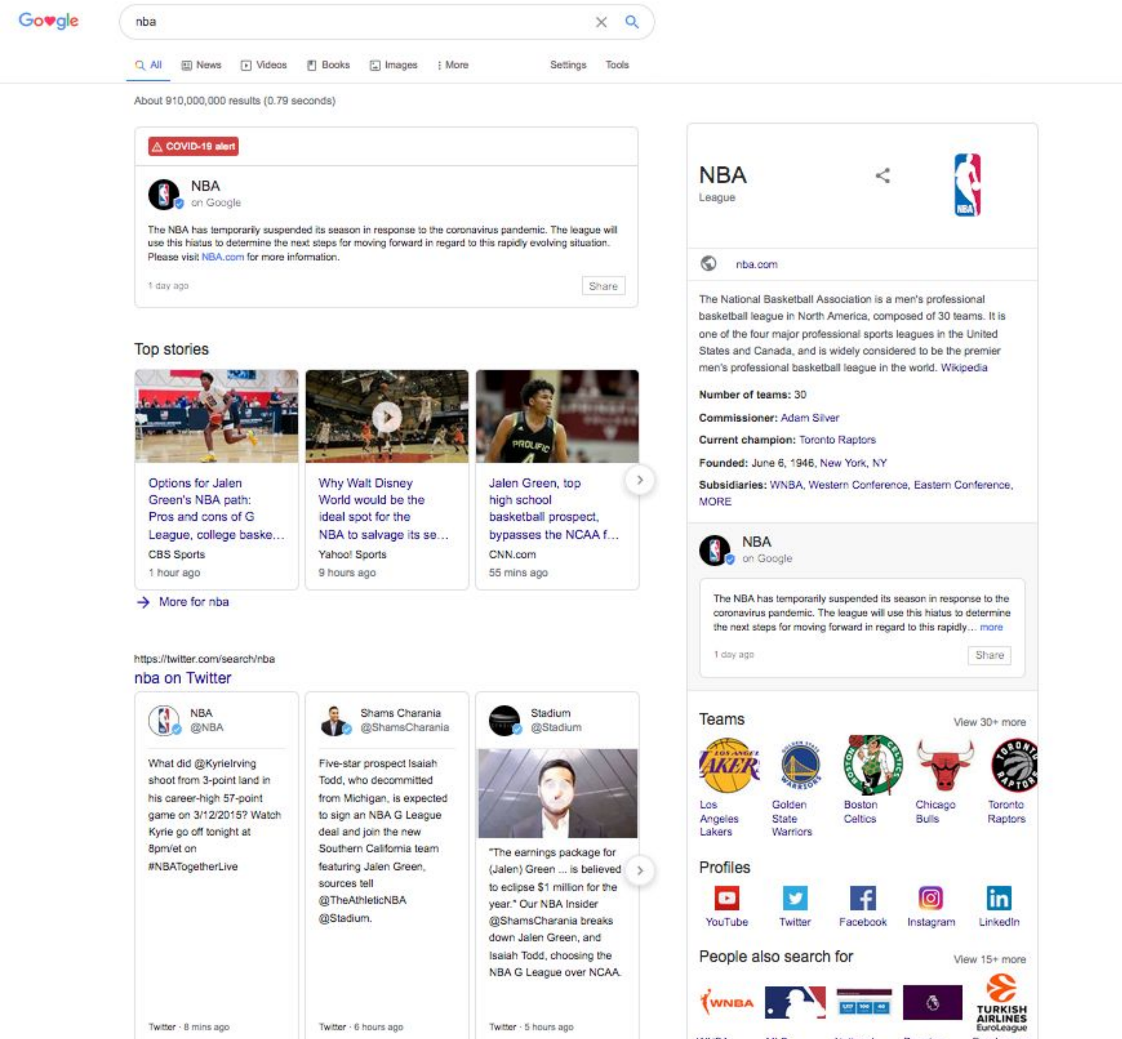
```
<div> … </div>
<div class=”searchResult_abc123”>
<a href=”wikiworkshop.org”> Wiki
Workshop 2020</a>
</div>
<div> … </div>
```

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
```

How should we turn this into a ranked list?



About 910,000,000 results (0.79 seconds)

COVID-19 alert

NBA
on Google

The NBA has temporarily suspended its season in response to the coronavirus pandemic. The league will use this hiatus to determine the next steps for moving forward in regard to this rapidly evolving situation. Please visit NBA.com for more information.

1 day ago

Share

Top stories

Options for Jalen Green's NBA path: Pros and cons of G League, college baske...
CBS Sports
1 hour ago

Why Walt Disney World would be the ideal spot for the NBA to salvage its se...
Yahoo! Sports
9 hours ago

Jalen Green, top high school basketball prospect, bypasses the NCAA f...
CNN.com
55 mins ago

More for nba

https://twitter.com/search/nba
nba on Twitter

NBA
@NBA
What did @KyrieIrving shoot from 3-point land in his career-high 57-point game on 3/12/2015? Watch Kyrie go off tonight at 8pm/et on #NBATogetherLive
Twitter · 8 mins ago

Shams Charania
@ShamsCharania
Five-star prospect Isaiah Todd, who decommitted from Michigan, is expected to sign an NBA G League deal and join the new Southern California team featuring Jalen Green, sources tell @TheAthleticNBA @Stadium.
Twitter · 6 hours ago

Stadium
@Stadium
"The earnings package for (Jalen) Green ... is believed to eclipse $1 million for the year." Our NBA Insider @ShamsCharania breaks down Jalen Green, and Isaiah Todd, choosing the NBA G League over NCAA.
Twitter · 5 hours ago

NBA
League

nba.com

The National Basketball Association is a men's professional basketball league in North America, composed of 30 teams. It is one of the four major professional sports leagues in the United States and Canada, and is widely considered to be the premier men's professional basketball league in the world. Wikipedia

Number of teams: 30

Commissioner: Adam Silver

Current champion: Toronto Raptors

Founded: June 6, 1946, New York, NY

Subsidiaries: WNBA, Western Conference, Eastern Conference, MORE

NBA
on Google

The NBA has temporarily suspended its season in response to the coronavirus pandemic. The league will use this hiatus to determine the next steps for moving forward in regard to this rapidly... more

1 day ago

Share

Teams

View 30+ more

Los Angeles Lakers
Golden State Warriors
Boston Celtics
Chicago Bulls
Toronto Raptors

Profiles

YouTube
Twitter
Facebook
Instagram
LinkedIn

People also search for

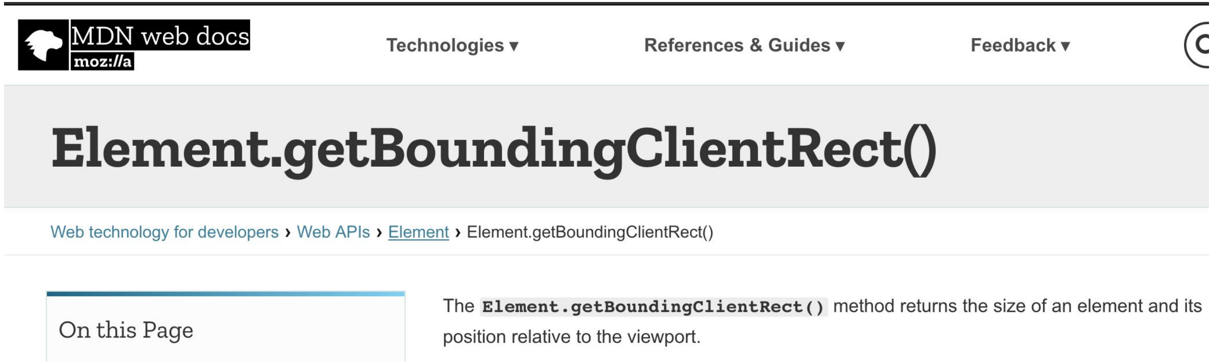
View 15+ more

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:



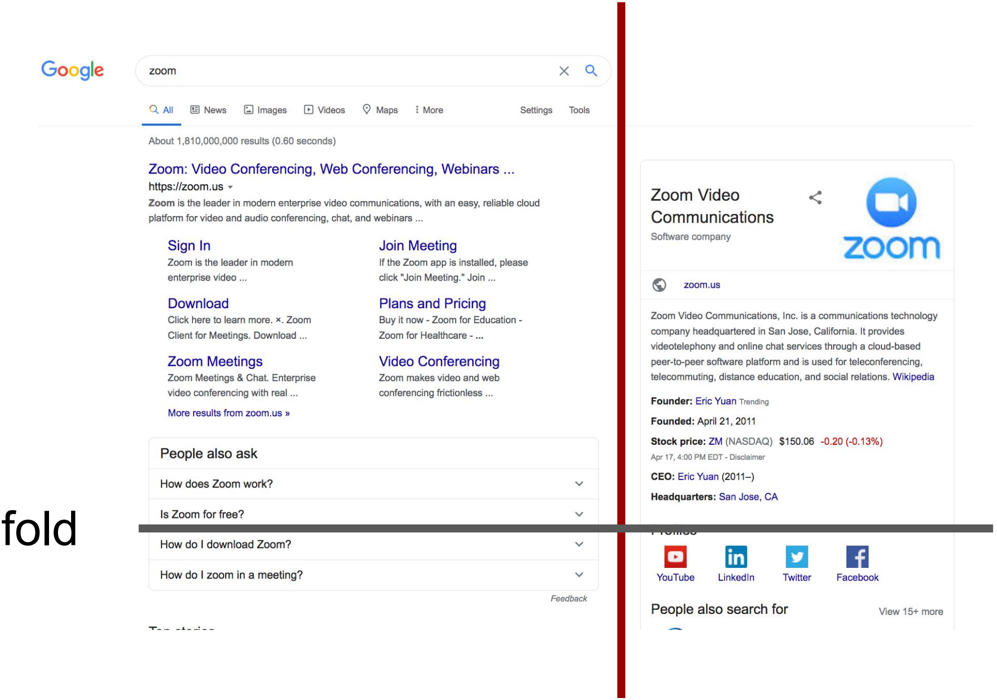
Element.getBoundingClientRect()

Web technology for developers › Web APIs › Element › Element.getBoundingClientRect()

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

Spatial incidence rate definition

full page



About 1,810,000,000 results (0.60 seconds)

Zoom: Video Conferencing, Web Conferencing, Webinars ...
https://zoom.us
Zoom is the leader in modern enterprise video communications, with an easy, reliable cloud platform for video and audio conferencing, chat, and webinars ...

Sign In
Zoom is the leader in modern enterprise video ...

Join Meeting
If the Zoom app is installed, please click "Join Meeting." Join ...

Download
Click here to learn more. ×. Zoom Client for Meetings. Download ...

Plans and Pricing
Buy it now - Zoom for Education - Zoom for Healthcare - ...

Zoom Meetings
Zoom Meetings & Chat. Enterprise video conferencing with real ...

Video Conferencing
Zoom makes video and web conferencing frictionless ...

More results from zoom.us »

People also ask

How does Zoom work?

Is Zoom for free?

How do I download Zoom?

How do I zoom in a meeting?

Feedback

Top stories

Zoom Video Communications
Software company

zoom.us

Zoom Video Communications, Inc. is a communications technology company headquartered in San Jose, California. It provides videotelephony and online chat services through a cloud-based peer-to-peer software platform and is used for teleconferencing, telecommuting, distance education, and social relations. Wikipedia

Founder: Eric Yuan Trending

Founded: April 21, 2011

Stock price: ZM (NASDAQ) $150.06 -0.20 (-0.13%)
Apr 17, 4:00 PM EDT - Disclaimer

CEO: Eric Yuan (2011–)

Headquarters: San Jose, CA

Profiles

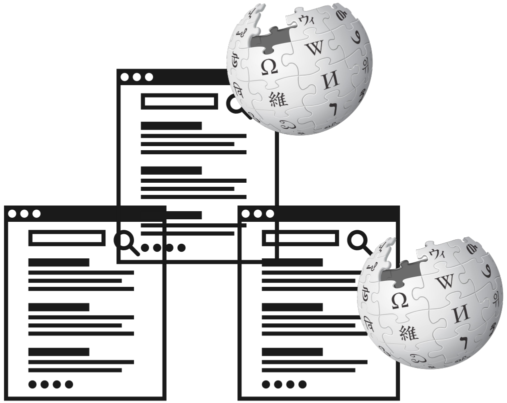
People also search for
View 15+ more

above-the-fold

left-hand and right-hand

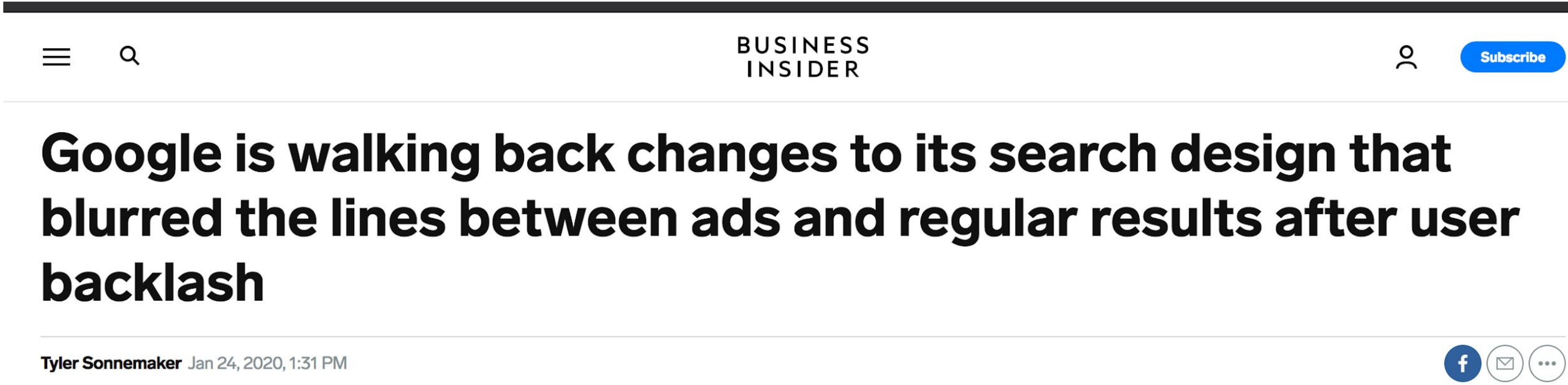
Incidence rates

- how often is Wikipedia showing up in SERPs?
  - If we collect 3 SERPs, and Wikipedia appears twice, 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?



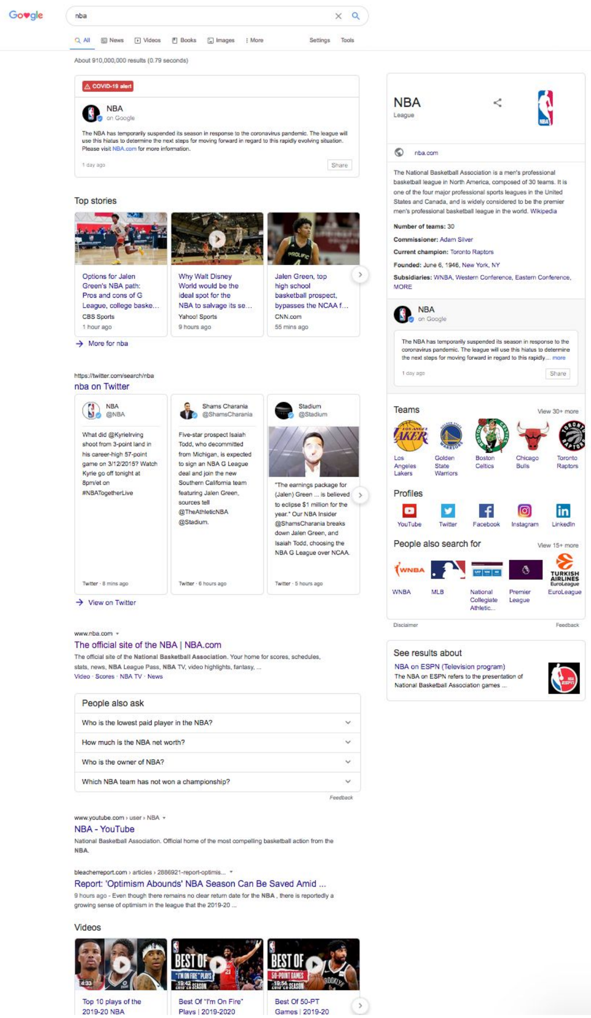
Google is walking back changes to its search design that blurred the lines between ads and regular results after user backlash

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!



About 910,000,000 results (0.79 seconds)

COVID-19 alert

NBA
on Google

The NBA has temporarily suspended its season in response to the coronavirus pandemic. The league will use this hiatus to determine the next steps for moving forward in regard to this rapidly evolving situation. Please visit NBA.com for more information.

1 day ago

Top stories

Options for Jalen Green's NBA path: Pros and cons of G League, college baske…
CBS Sports
1 hour ago

Why Walt Disney World would be the ideal spot for the NBA to salvage its se…
Yahoo! Sports
9 hours ago

Jalen Green, top high school basketball prospect, bypasses the NCAA f…
CNN.com
55 mins ago

→ More for nba

https://twitter.com/search/nba
nba on Twitter

NBA
@NBA
What did @KyrieIrving shoot from 3-point land in his career-high 57-point game on 3/12/2015? Watch Kyrie go off tonight at 8pm/et on #NBATogetherLive
Twitter · 8 mins ago

Shams Charania
@ShamsCharania
Five-star prospect Isaiah Todd, who decommitted from Michigan, is expected to sign an NBA G League deal and join the new Southern California team featuring Jalen Green, sources tell @TheAthleticNBA @Stadium.
Twitter · 6 hours ago

Stadium
@Stadium
"The earnings package for (Jalen) Green ... is believed to eclipse $1 million for the year." Our NBA Insider @ShamsCharania breaks down Jalen Green, and Isaiah Todd, choosing the NBA G League over NCAA.
Twitter · 5 hours ago

→ View on Twitter

www.nba.com ▾
The official site of the NBA | NBA.com
The official site of the **National Basketball Association**. Your home for scores, schedules, stats, news, **NBA** League Pass, **NBA** TV, video highlights, fantasy, ...
Video · Scores · NBA TV · News

People also ask

Who is the lowest paid player in the NBA?

How much is the NBA net worth?

Who is the owner of NBA?

Which NBA team has not won a championship?

www.youtube.com › user › NBA ▾
NBA - YouTube
National Basketball Association. Official home of the most compelling basketball action from the **NBA**.

bleacherreport.com › articles › 2886921-report-optimis... ▾
Report: 'Optimism Abounds' NBA Season Can Be Saved Amid ...
9 hours ago - Even though there remains no clear return date for the **NBA** , there is reportedly a growing sense of optimism in the league that the 2019-20 ...

Videos

Top 10 plays of the 2019-20 NBA

Best Of "I'm On Fire" Plays | 2019-2020

Best Of 50-PT Games | 2019-20

NBA
League

nba.com

The National Basketball Association is a men's professional basketball league in North America, composed of 30 teams. It is one of the four major professional sports leagues in the United States and Canada, and is widely considered to be the premier men's professional basketball league in the world. Wikipedia

Number of teams: 30

Commissioner: Adam Silver

Current champion: Toronto Raptors

Founded: June 6, 1946, New York, NY

Subsidiaries: WNBA, Western Conference, Eastern Conference, MORE

NBA
on Google

The NBA has temporarily suspended its season in response to the coronavirus pandemic. The league will use this hiatus to determine the next steps for moving forward in regard to this rapidly… more

1 day ago

Teams
View 30+ more

Los Angeles Lakers · Golden State Warriors · Boston Celtics · Chicago Bulls · Toronto Raptors

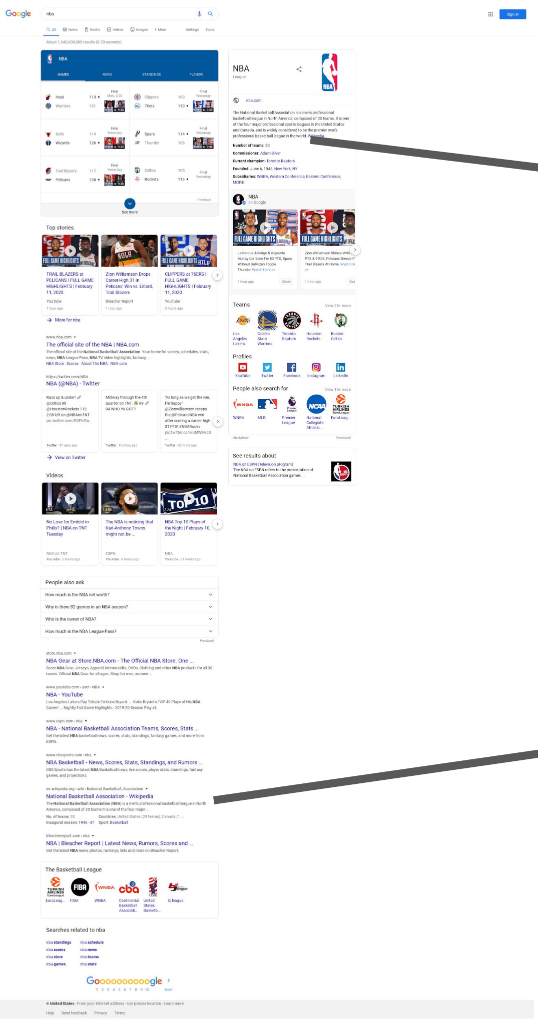
Profiles

YouTube · Twitter · Facebook · Instagram · LinkedIn

People also search for
View 15+ more

WNBA · MLB · National Collegiate Athletic... · Premier League · EuroLeague

See results about
NBA on ESPN (Television program)
The NBA on ESPN refers to the presentation of National Basketball Association games ...



Google

nba

All · News · Books · Videos · Images · More · Settings · Tools

Sign in

About 1,340,000,000 results (0.70 seconds)

NBA

GAMES · NEWS · STANDINGS · PLAYERS

Heat Warriors	113 ◂ 101	Final Mon, 2/10	Clippers 76ers	103 110 ◂	Final Yesterday
Bulls Wizards	114 126 ◂	Final Yesterday	Spurs Thunder	114 ◂ 106	Final Yesterday
Trail Blazers Pelicans	117 138 ◂	Final Yesterday	Celtics Rockets	105 116 ◂	Final Yesterday

Feedback

See more

Top stories

TRAIL BLAZERS at PELICANS | FULL GAME HIGHLIGHTS | February 11, 2020
YouTube
1 hour ago

Zion Williamson Drops Career-High 31 in Pelicans' Win vs. Lillard, Trail Blazers
Bleacher Report
1 hour ago

CLIPPERS at 76ERS | FULL GAME HIGHLIGHTS | February 11, 2020
YouTube
2 hours ago

→ More for nba

www.nba.com ▾
The official site of the NBA | NBA.com
The official site of the **National Basketball Association**. Your home for scores, schedules, stats, news, **NBA** League Pass, **NBA** TV, video highlights, fantasy, ...
NBA Store · Scores · About The NBA · NBA.com

https://twitter.com/NBA
NBA (@NBA) · Twitter

Russ up & under! 🚀 @celtics 98 @HoustonRockets 113 2:08 left on @NBAonTNT pic.twitter.com/R0PhiRu...
Twitter · 47 secs ago

Midway through the 4th quarter on TNT. ☘ 89 🚀 94 WHO YA GOT?
Twitter · 16 mins ago

"As long as we get the win, I'm happy." @Zionwilliamson recaps the @PelicansNBA win after scoring a career-high 31 PTS! #NBARooks pic.twitter.com/uMWMvcG ...
Twitter · 32 mins ago

→ View on Twitter

Videos

6:33 No Love for Embiid in Philly? | NBA on TNT Tuesday
NBA on TNT
YouTube · 5 hours ago

4:59 The NBA is noticing that Karl-Anthony Towns might not be ...
ESPN
YouTube · 8 hours ago

2:24 NBA Top 10 Plays of the Night | February 10, 2020
NBA
YouTube · 21 hours ago

People also ask

How much is the NBA net worth?

Why is there 82 games in an NBA season?

Who is the owner of NBA?

How much is the NBA League Pass?

Feedback

store.nba.com ▾
NBA Gear at Store.NBA.com - The Official NBA Store. One ...
Score **NBA** Gear, Jerseys, Apparel, Memorabilia, DVDs, Clothing and other **NBA** products for all 30 teams. Official **NBA** Gear for all ages. Shop for men, women ...

www.youtube.com › user › NBA ▾
NBA - YouTube
Los Angeles Lakers Pay Tribute To Kobe Bryant. ... Kobe Bryant's TOP 40 Plays of His **NBA** Career! ... Nightly Full Game Highlights - 2019-20 Season Play all.

www.espn.com › nba ▾
NBA - National Basketball Association Teams, Scores, Stats ...
Get the latest **NBA** basketball news, scores, stats, standings, fantasy games, and more from ESPN.

www.cbssports.com › nba ▾
NBA Basketball - News, Scores, Stats, Standings, and Rumors ...
CBS Sports has the latest **NBA** Basketball news, live scores, player stats, standings, fantasy games, and projections.

en.wikipedia.org › wiki › National_Basketball_Association ▾
National Basketball Association - Wikipedia
The **National Basketball Association** (**NBA**) is a men's professional basketball league in North America, composed of 30 teams It is one of the four major ...
No. of teams: 30 **Countries**: United States (29 teams); Canada (1 ...
Inaugural season: 1946–47 **Sport**: Basketball

bleacherreport.com › nba ▾
NBA | Bleacher Report | Latest News, Rumors, Scores and ...
Get the latest **NBA** news, photos, rankings, lists and more on Bleacher Report.

The Basketball League

EuroLeag... · FIBA · WNBA · Continental Basketball Associati... · United States Basketb... · bj league

Searches related to nba

nba **standings**	nba **schedule**
nba **scores**	nba **news**
nba **store**	nba **teams**
nba **games**	nba **stats**

Gooooooooooogle ›
1 2 3 4 5 6 7 8 9 10 Next

● **United States** - From your Internet address - Use precise location - Learn more

Help · Send feedback · Privacy · Terms

NBA
League

nba.com

The National Basketball Association is a men's professional basketball league in North America, composed of 30 teams. It is one of the four major professional sports leagues in the United States and Canada, and is widely considered to be the premier men's professional basketball league in the world. Wikipedia

Number of teams: 30
Commissioner: Adam Silver
Current champion: Toronto Raptors
Founded: June 6, 1946, New York, NY
Subsidiaries: WNBA, Western Conference, Eastern Conference, MORE

NBA
on Google

LaMarcus Aldridge & Dejounte Murray Combine For 50 PTS, Spurs Without DeRozan Topple Thunder. Watch Here >>
1 hour ago · Share

Zion Williamson Shines With ... PTS & 9 REB, Pelicans Breeze P... Trail Blazers At Home. Watch H... >>
1 hour ago · Sha...

Teams · View 25+ more

Los Angeles Lakers · Golden State Warriors · Toronto Raptors · Houston Rockets · Boston Celtics

Profiles

YouTube · Twitter · Facebook · Instagram · LinkedIn

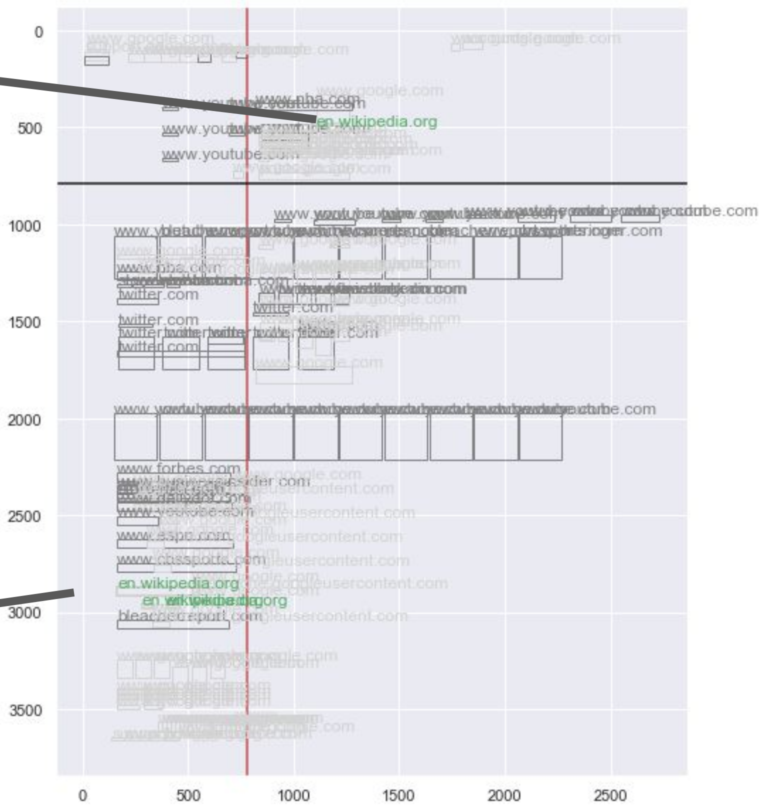
People also search for · View 15+ more

WNBA · MLB · Premier League · National Collegiate Athletic... · EuroLeag...

Disclaimer · Feedback

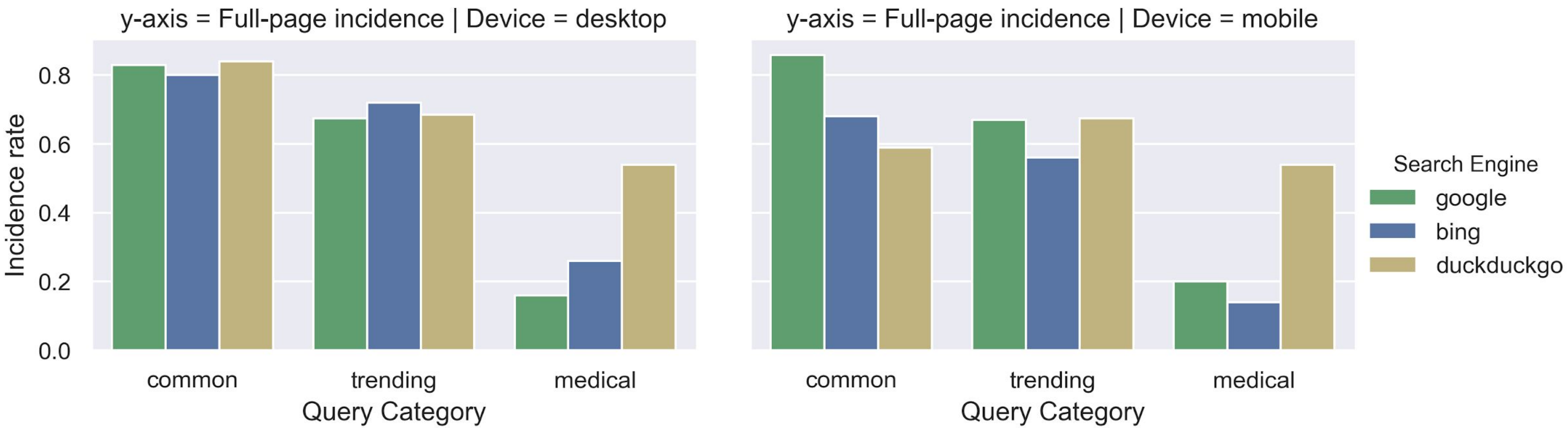
See results about

NBA on ESPN (Television program)
The NBA on ESPN refers to the presentation of National Basketball Association games ...

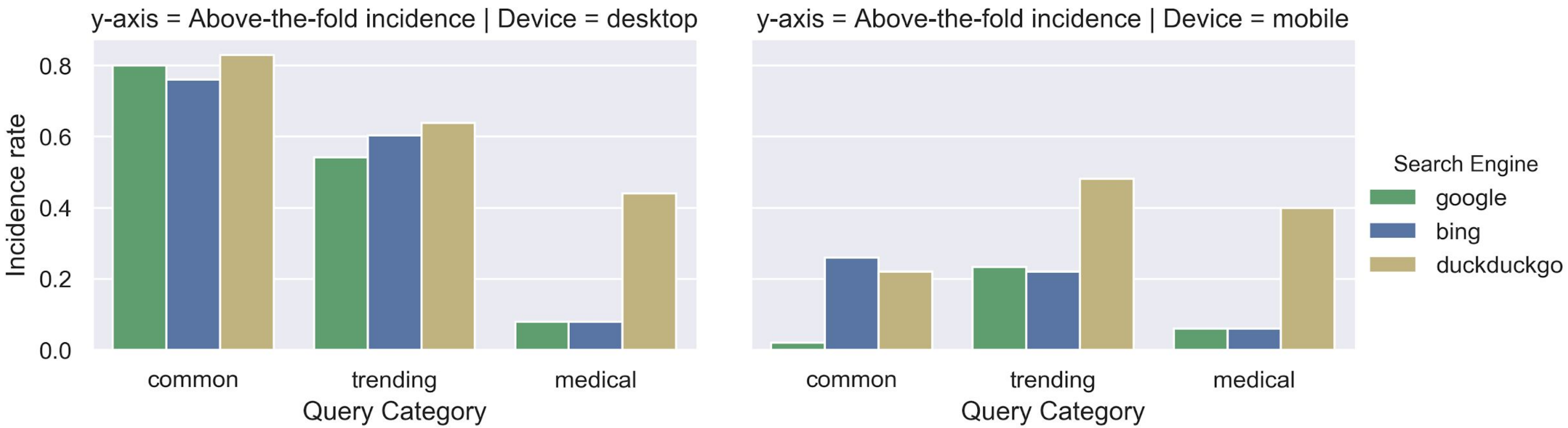


Results

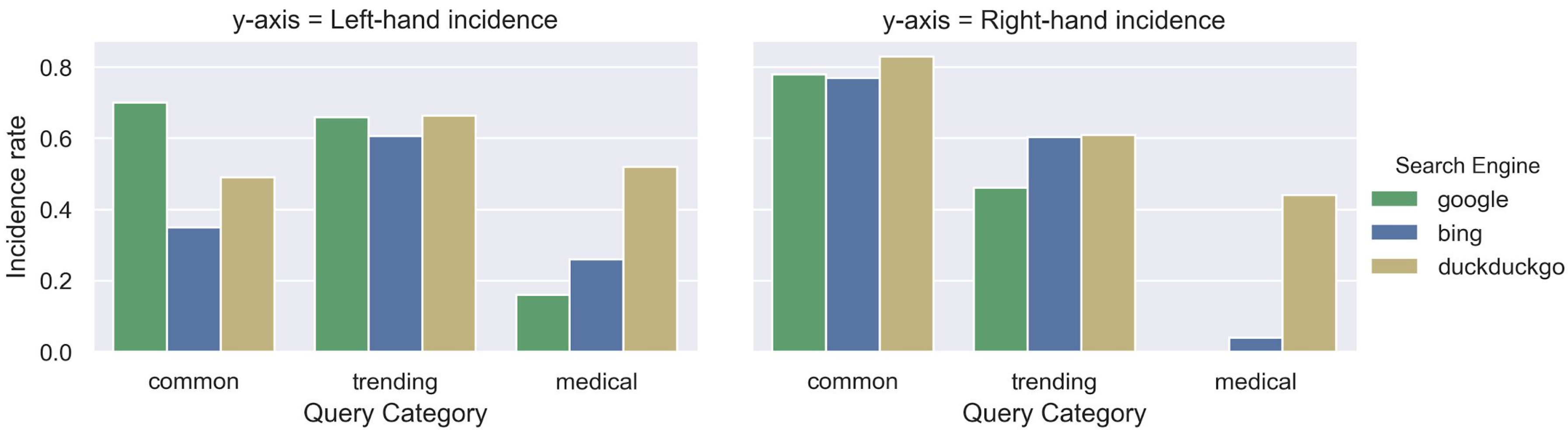
Full page incidence rates



Above-the-fold incidence rates



Left-hand and right-hand



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

Data from the Public Fuels Intelligent Technologies

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

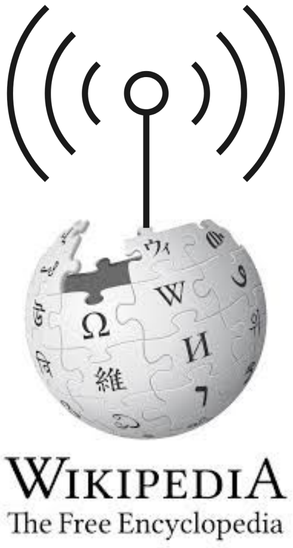
You **cannot** pay people to edit Wikipedia.

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

Wikipedia Matters Outside Wikipedia

Positive: effective peer production = effective search results?

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



Raises the stakes of Wikipedia-focused research and Wikipedia findings

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

Many thanks to:

co-author Brent Hecht, thoughtful reviewers and colleagues
Open Software

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

Community Data Science Collective’s COVID-19 Digital Observatory

-wiki page: https://wiki.communitydata.science/COVID-19_Digital_Observatory
-has SERP data for COVID-related keywords, using a newer version of this software developed after this study
-other data of interest to this group (Wikipedia editing, reddit)

Questions?

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

Links:

paper: https://www.nickmvincent.com - under “Pre-prints”
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
COVID-19 Digital Observatory:
https://wiki.communitydata.science/COVID-19_Digital_Observatory