Searching, fast and slow
A tech perspective

October 12th, 2020

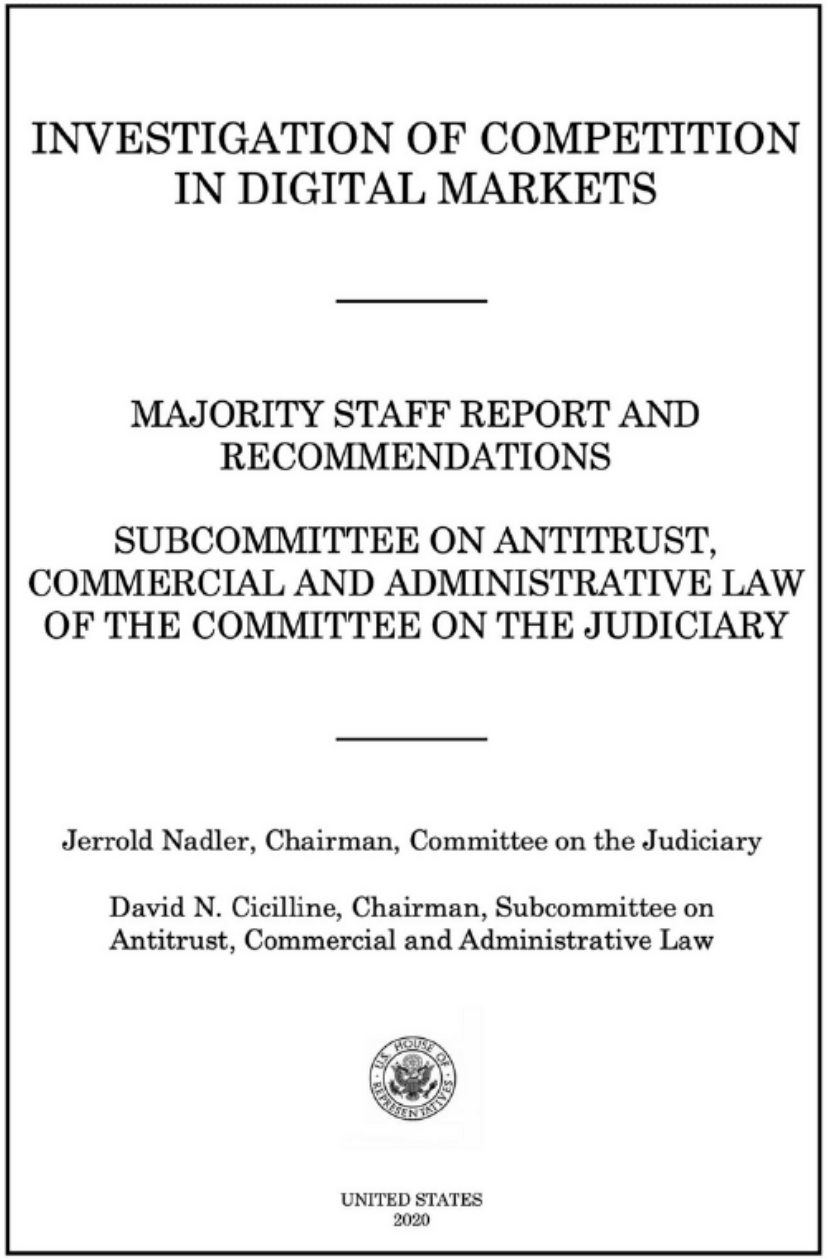
Prof. dr. ir. Arjen P. de Vries
arjen@acm.org

ANTITRUST REPORT ON BIG TECH

Several factors render Google's power in online search generally immune to competition or threat of entry.

General online search strongly favors scale due to (1) the **high fixed costs of servers needed for crawling and indexing the entire web**, and (2) the **self-reinforcing advantages of click- and query data**, which let a search engine constantly improve the relevance of search results.

Even search engines that choose to syndicate their search results rather than create their own index and algorithm face major obstacles . This is primarily because Google — through both integration and contractual agreements — has established itself as the **default search provider on 87% of desktop browsers and the vast majority of mobile devices**.



INVESTIGATION OF COMPETITION IN DIGITAL MARKETS

MAJORITY STAFF REPORT AND RECOMMENDATIONS

SUBCOMMITTEE ON ANTITRUST, COMMERCIAL AND ADMINISTRATIVE LAW OF THE COMMITTEE ON THE JUDICIARY

Jerrold Nadler, Chairman, Committee on the Judiciary

David N. Cicilline, Chairman, Subcommittee on Antitrust, Commercial and Administrative Law

UNITED STATES
2020

BARRIERS TO ENTRY

ADVANTAGES OF CLICK- AND QUERY DATA

Without the log data, web search isn't as good

This hinders retrieval experiments in our lab, and academia in general!

Note:
Wednesday October14th, 11:00, presentation by Djoerd Hiemstra
https://djoerdhiemstra.com/2020/reducing-misinformation-in-query-autocompletions/

BARRIERS TO ENTRY

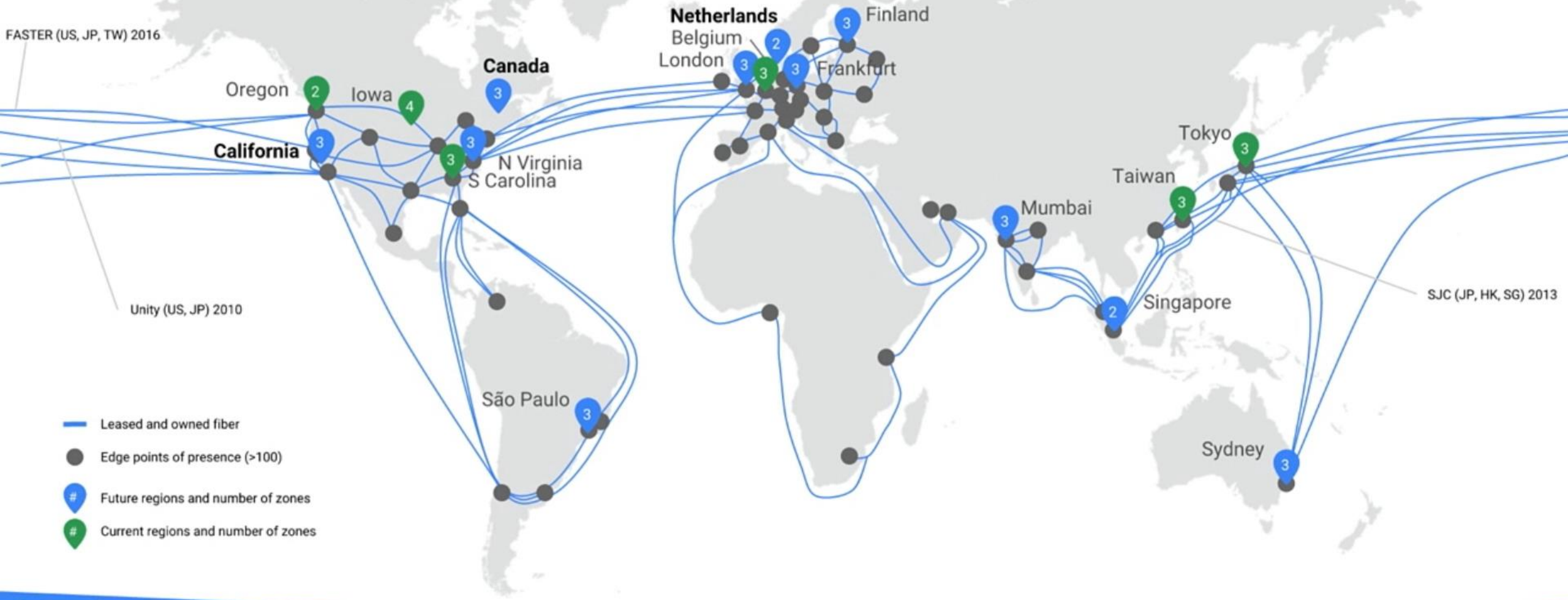
HIGH COST

Gartner estimated in a July 2016 report that Google at the time had 2.5 million servers

Google VP of data centers Joe Kava's presentation at Google Cloud Next 2017 in San Francisco:

GCP Infrastructure

6 regions, 18 zones, over 100 points of presence, and a well-provisioned global network comprised of hundreds of thousands of miles of fiber optic cable.



BARRIERS TO ENTRY

HIGH COST

Gartner estimated in a July 2016 report that Google at the time had 2.5 million servers

Google VP of data centers Joe Kava's presentation at Google Cloud Next 2017 in San Francisco:

The Dalles in Oregon: investment $1.8 billion
Pryor Creek, Oklahoma: investment $2 billion
Times 15...

The new data center under construction in 2016 in Eemshaven, Netherlands, is expected to cost $773 million.

Overall, Google's capital expenditures for 2016 were just under $10.2 billion. Most of that can be accounted for by its data centers and land acquisitions.

https://www.datacenterknowledge.com/google-data-center-faq/

Inexpensive...
... yet, privacy invasive!

Gmail Images

Google

pihole Remove
iipc Remove
saskia de wildt Remove
wikipedia offline zim warc Remove
wikipedia downloads warc Remove
wikipedia zymm downloads warc Remove
zymm warc Remove
nmh Remove
mh group mail by date Remove
diane kelly utk Remove

Google Search I'm Feeling Lucky

Very fast!
Not that relevant!

Google

i

iipc Remove
ivan veul Remove
information-science archive radboud thesis Remove
install pi hole
IKEA
instagram
ING Group
indeed
iphone 12
iphone 11

Google Search | I'm Feeling Lucky

Report inappropriate predictions

Human-in-the-loop!

Google

in

instagram

ING Group

indeed

Intratuin

inch to cm

internet speed test

ind

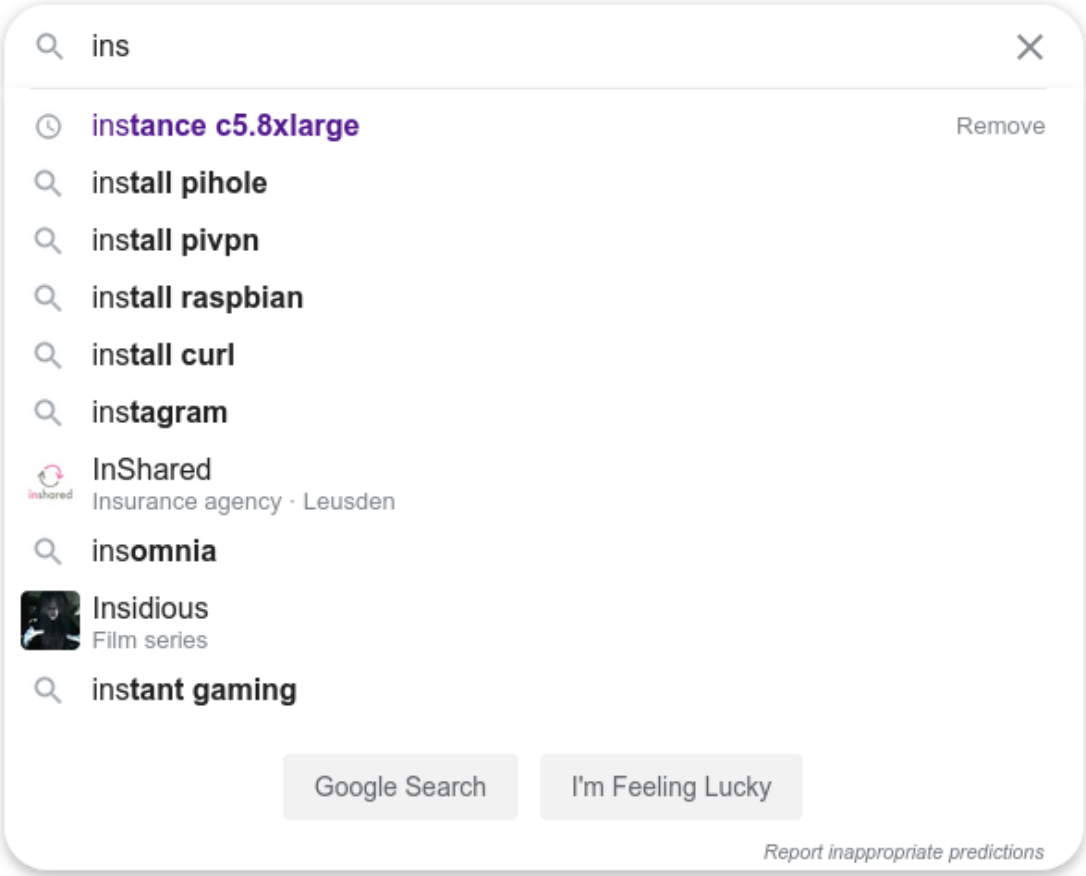
Intersport

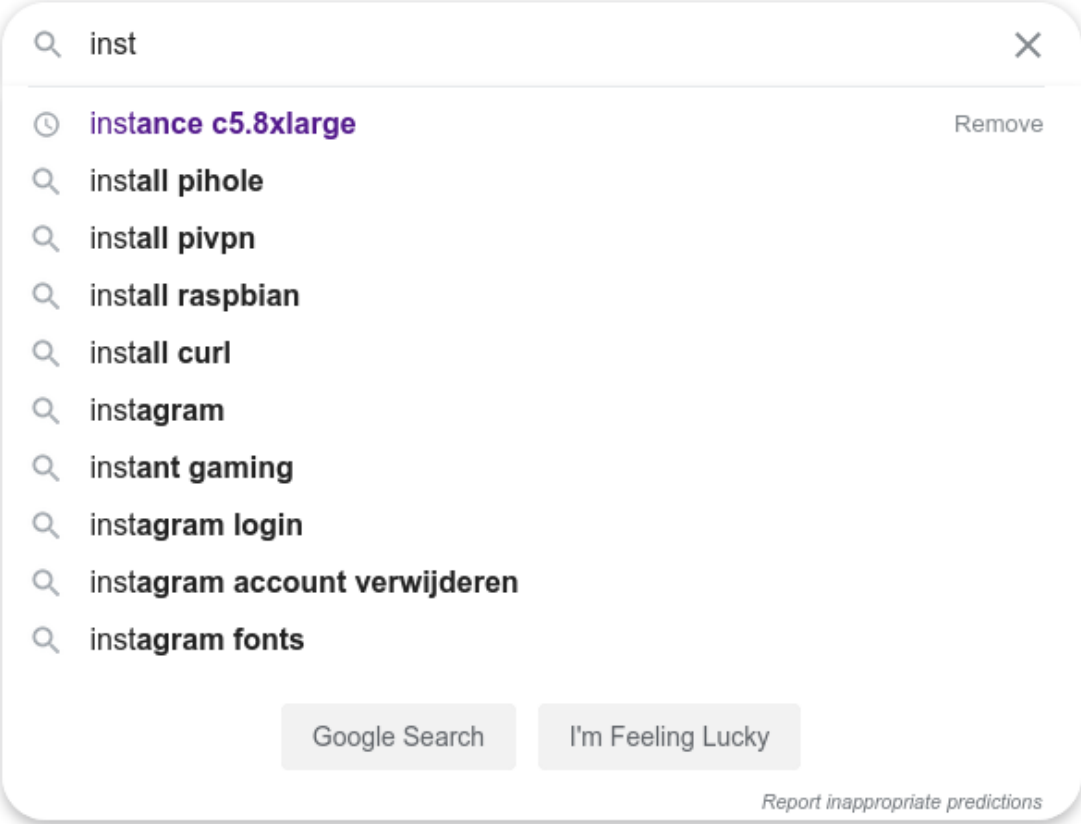
Interstellar
2014 film

India
Country in South Asia

Google Search | I'm Feeling Lucky

Report inappropriate predictions


Google
ins
instance c5.8xlarge
Remove
install pihole
install pivpn
install raspbian
install curl
instagram
InShared
Insurance agency · Leusden
insomnia
Insidious
Film series
instant gaming
Google Search
I'm Feeling Lucky
Report inappropriate predictions


Google
inst
instance c5.8xlarge
Remove
install pihole
install pivpn
install raspbian
install curl
instagram
instant gaming
instagram login
instagram account verwijderen
instagram fonts
Google Search
I'm Feeling Lucky
Report inappropriate predictions

insta

instance c5.8xlarge Remove

install pihole

install pivpn

install raspbian

install curl

instagram

instagram verwijderen

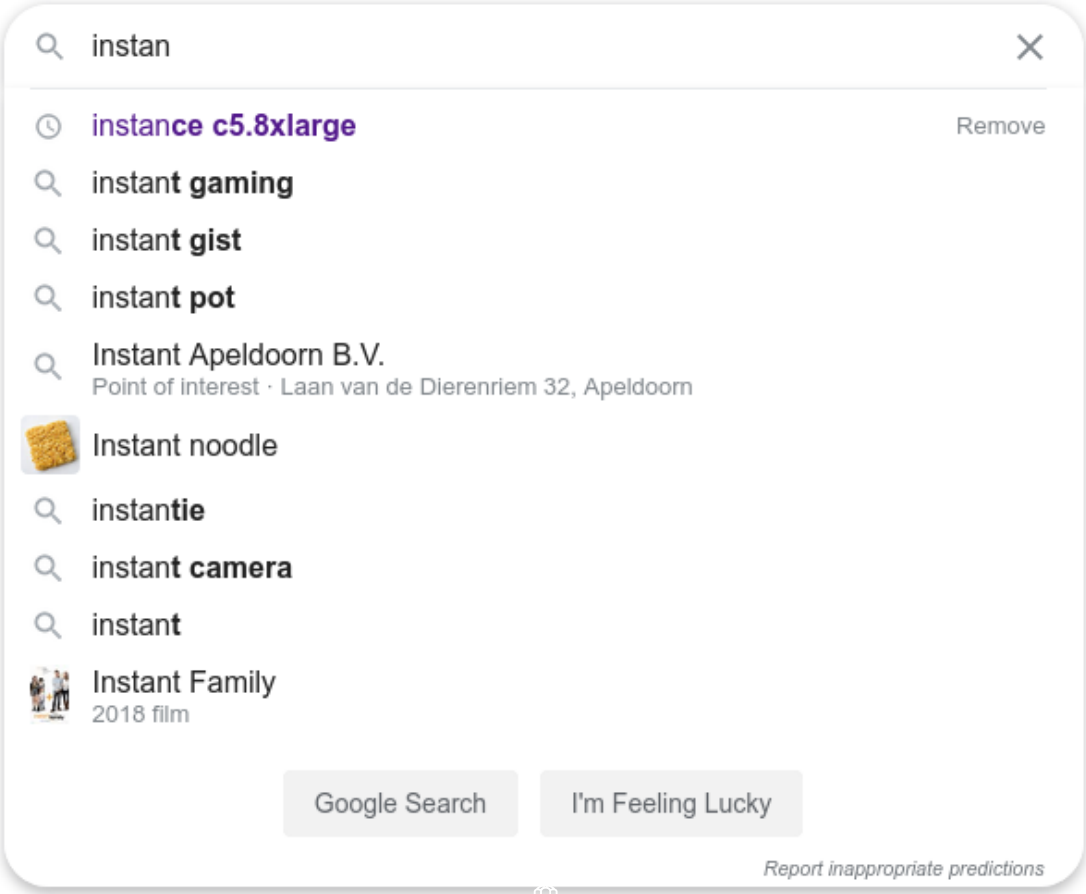
insta

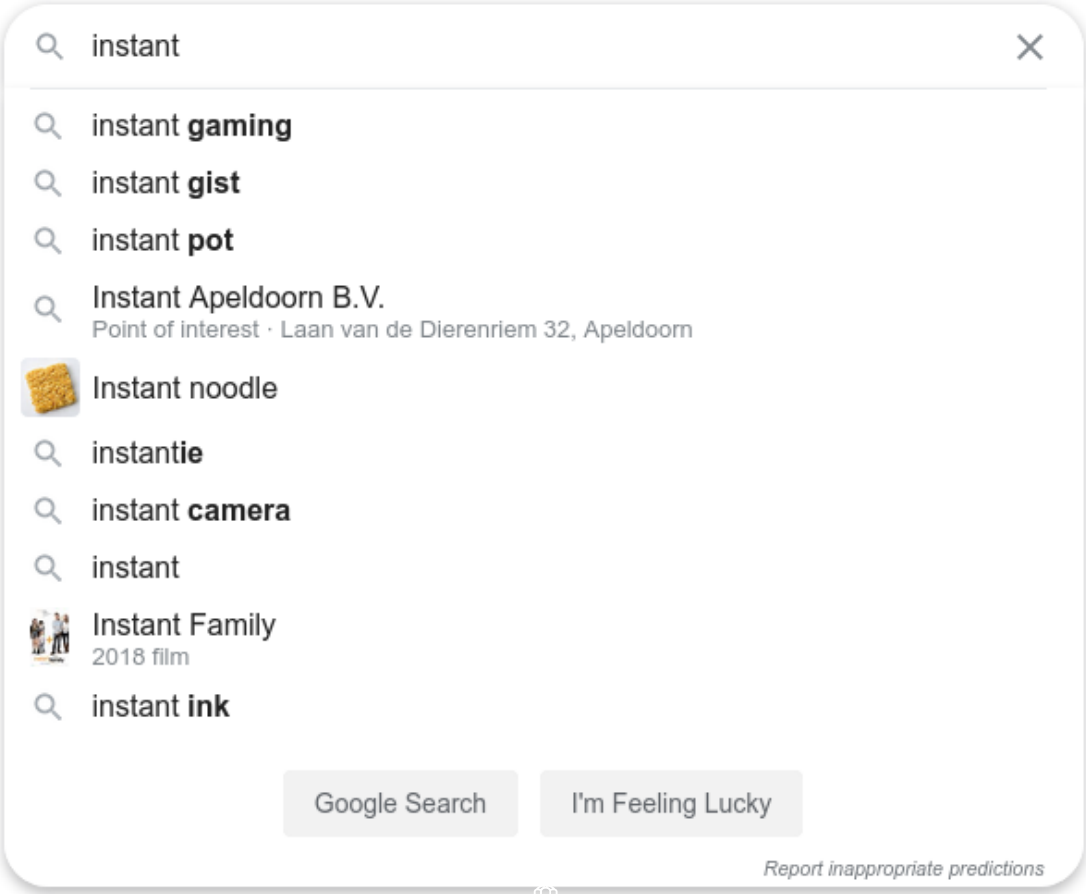
instagram account verwijderen

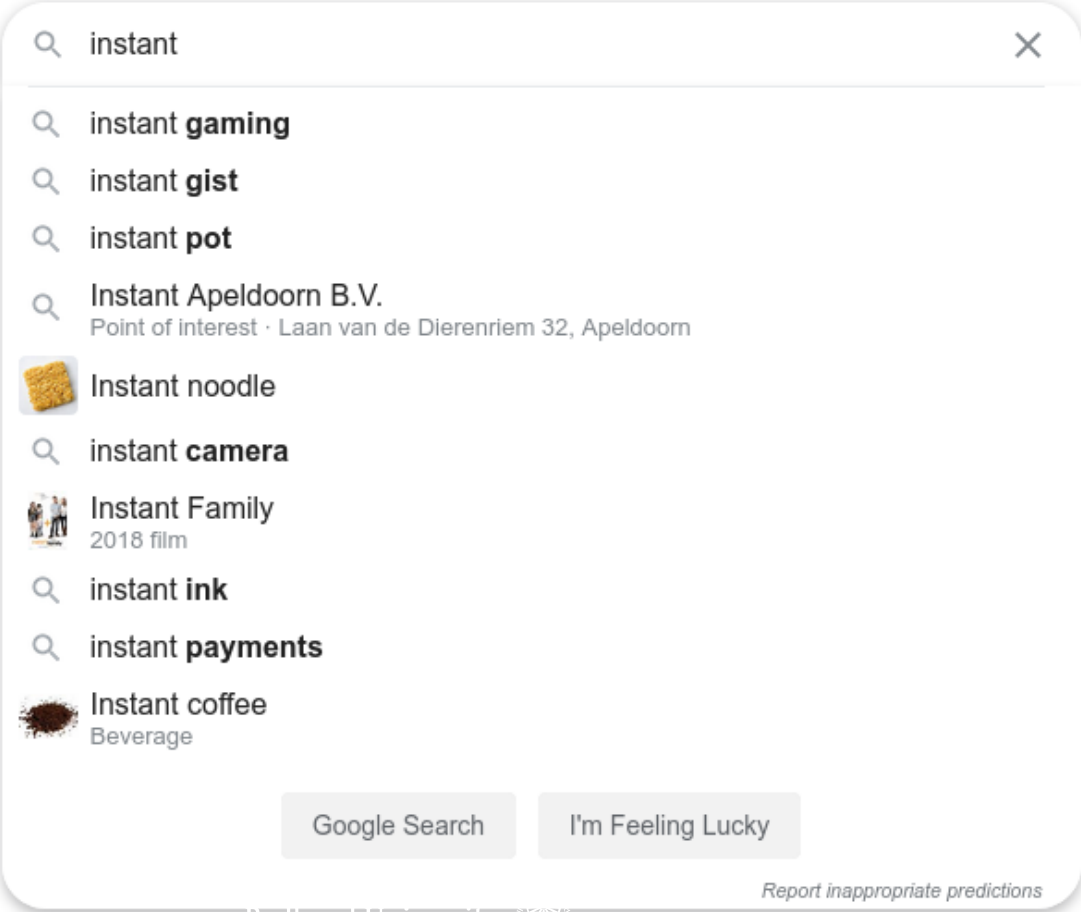
instagram inloggen

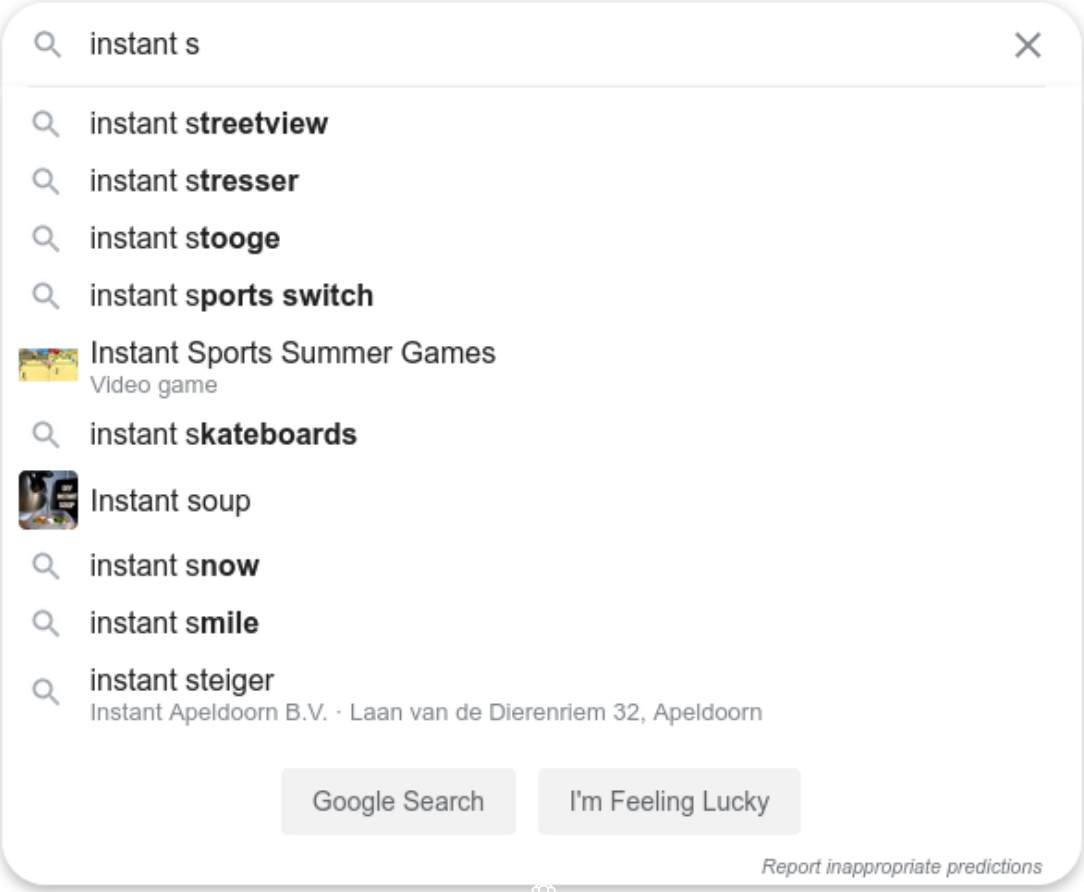
Google Search I'm Feeling Lucky

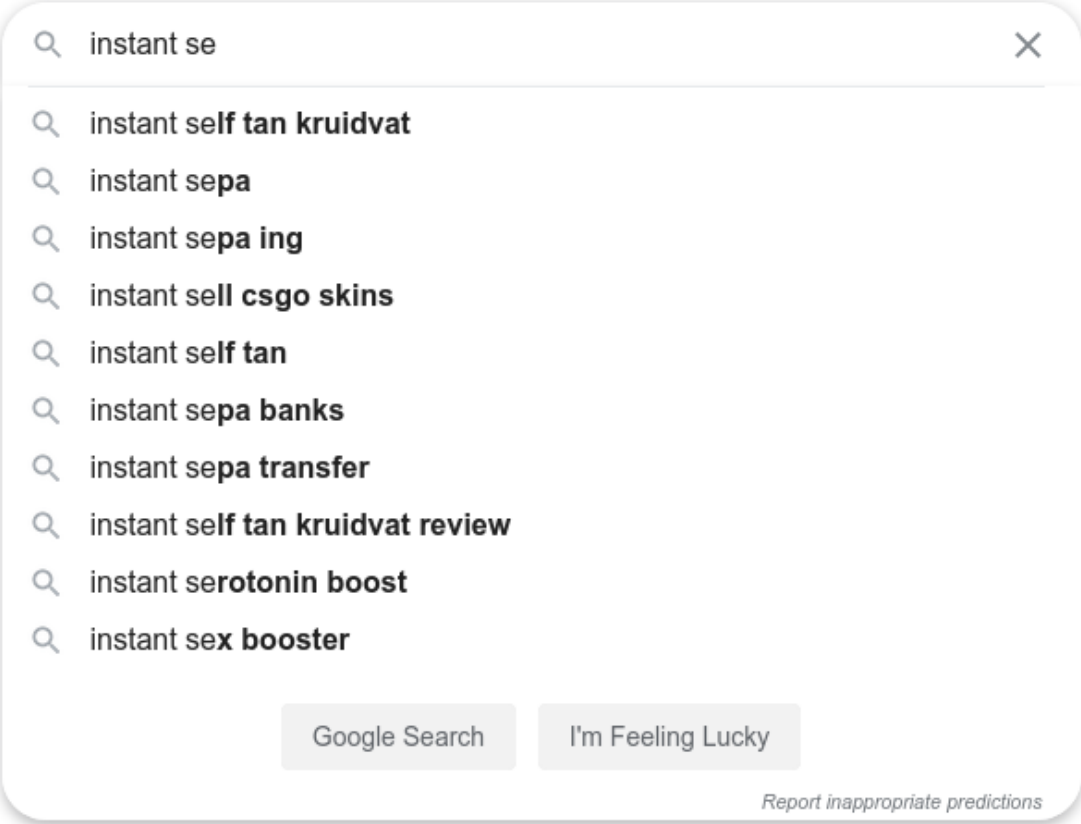
Report inappropriate predictions


Google
instan
instance c5.8xlarge
Remove
instant gaming
instant gist
instant pot
Instant Apeldoorn B.V.
Point of interest · Laan van de Dierenriem 32, Apeldoorn
Instant noodle
instantie
instant camera
instant
Instant Family
2018 film
Google Search
I'm Feeling Lucky
Report inappropriate predictions


Google
instant
instant gaming
instant gist
instant pot
Instant Apeldoorn B.V.
Point of interest · Laan van de Dierenriem 32, Apeldoorn
Instant noodle
instantie
instant camera
instant
Instant Family
2018 film
instant ink
Google Search
I'm Feeling Lucky
Report inappropriate predictions


Google
instant
instant gaming
instant gist
instant pot
Instant Apeldoorn B.V.
Point of interest · Laan van de Dierenriem 32, Apeldoorn
Instant noodle
instant camera
Instant Family
2018 film
instant ink
instant payments
Instant coffee
Beverage
Google Search
I'm Feeling Lucky
Report inappropriate predictions


Google
instant s
instant streetview
instant stresser
instant stooge
instant sports switch
Instant Sports Summer Games
Video game
instant skateboards
Instant soup
instant snow
instant smile
instant steiger
Instant Apeldoorn B.V. · Laan van de Dierenriem 32, Apeldoorn
Google Search
I'm Feeling Lucky
Report inappropriate predictions


Google
instant se
instant self tan kruidvat
instant sepa
instant sepa ing
instant sell csgo skins
instant self tan
instant sepa banks
instant sepa transfer
instant self tan kruidvat review
instant serotonin boost
instant sex booster
Google Search
I'm Feeling Lucky
Report inappropriate predictions

Finally!

Google

instant sea

instant search

instant search outlook

Instant Sea Containers
Building materials supplier · Landsdale WA, Australia

instant search outlook missing

instant seats

instant seal

instant sealer

instant seaweed soup

instant sealant

instant search algolia

Google Search

I'm Feeling Lucky

Report inappropriate predictions

Google

instant search

All | Images | Videos | Shopping | Maps | More | Settings

About 1.330.000.000 results (0,62 seconds)

www.instantsearchplus.com
InstantSearch+: Site Search for eCommerce
Search & Merchandising for **Fast** Growing Online Brands. The most sophisticated brands use InstantSearch+ to optimize the entire shopping experience. shopify ...
Shopify Search · Magento Search · Wix Search · WooCommerce Search

apps.shopify.com › instant-search
Instant Search+ and Filters – Ecommerce Plugins for Online ...
About **Instant** Search+ and Filters · Shopify Plus certified **search** & merchandising app used by the fastest growing and most sophisticated brands on the Shopify ...
★★★★★ Rating: 4,9 · 633 votes

www.algolia.com › products › instantsearch
InstantSearch | Design Great Search & Discovery | Algolia
The perfect match. Libraries built to unleash the full potential of Algolia's search infrastructure. **InstantSearch** takes advantage of the unique capabilities of Algolia's ...

www.algolia.com › doc › guides › what-is-instantsearch
What Is InstantSearch.js? | Building Search UI | Guide | Algolia ...
InstantSearch.js is an open-source, production-ready UI library for Vanilla JS that lets you quickly build a search interface in your front-end application. Our goal ...

www.bigcommerce.com › All
Instant Search + - BigCommerce
Instant Search+ is an advanced **search** & merchandising app used by the some of the fastest growing and most sophisticated online brands. Join RC Planet, ...
★★★★★ Rating: 4,9 · 50 reviews · US$89.99

github.com › algolia › instantsearch
algolia/instantsearch.js: ⚡ A JavaScript library for ... - GitHub
InstantSearch.js is a JavaScript library for building performant and **instant search** experiences with Algolia. Version License Build Status Pull reminders.

github.com › algolia › react-instantsearch
algolia/react-instantsearch: ⚡ Lightning-fast ... - GitHub
README.md. React **InstantSearch** is a library for building blazing **fast search**-as-you-type search UIs with Algolia.

marketplace.magento.com › instantsearch-instantsearch...

Google

instant search google

Human-in-the-loop!

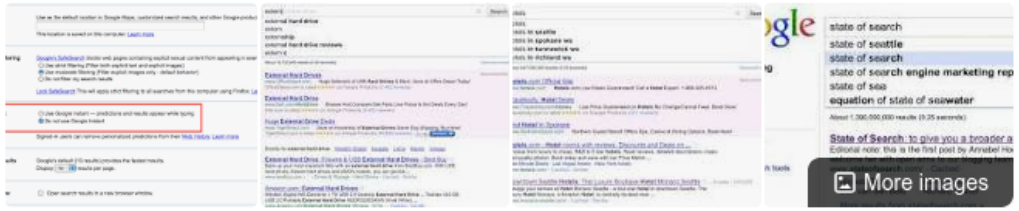
Very fast!
Not that relevant!

Google

google instant search

All | Images | News | Videos | Maps | More | Settings | Tools

About 1.320.000.000 results (0,56 seconds)



Google Instant search

Google Instant is a feature that predicts what you're **searching** for and shows results as you type. It uses **Google's** autocomplete technology to show predicted **search** terms in a drop-down box, and begins to display **search** results below the drop-down. Sep 8, 2010

searchengineland.com › google-instant-complete-users-...

Google Instant Search: The Complete User's Guide

Feedback

People also ask

How do I turn on Google Instant Search?

What happened Google Instant?

What are the suggestions in Google search?

How does Google autocomplete?

Feedback

searchengineland.com › google-dropped-google-instant...

Google has dropped Google Instant Search

Jul 26, 2017 — Several years after Google launched **Google Instant**, they are killing the default **search** feature to bring **search** more inline with mobile devices.

Search anno 2020:

- Snippets
- Verticals
- Knowledge Graph
- Instant Answers
- Mobile
- ...

ANTI-CLIMAX ☺

SEO

Google has dropped Google Instant Search

Several years after Google launched Google Instant, they are killing the default search feature to bring search more inline with mobile devices.

Barry Schwartz on July 26, 2017 at 10:31 am

BARRIERS TO ENTRY / HIGH COST

"OPEN" AI

Above $10 million in expenses for research on GPT-3 and training the final model

Tens of thousands of dollars in monthly cloud computing or server and electricity costs for running the model

Possibly more than a million dollars in yearly retraining costs due to model decay

Additional costs of customer support, marketing, IT, security, legal and other requirements of running a product. This could be in the tens of thousands of dollars based on the number and size of customers OpenAI acquires.

https://bdtechtalks.com/2020/09/24/microsoft-openai-gpt-3-license/

"OPEN" AI... CLOSED

Result:
OpenAI is giving Microsoft exclusive access to its GPT-3 language model (in exchange for 1B$)

https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/

BARRIERS TO ENTRY

WHAT TO DO?!

Make a different product; not "Web Search" as we know it!

We can never beat them **at what they do very well** if we try to do the same

Our / EU pockets are not deep enough

Even if we get as good as Bing, we are not so likely to get more than 6% market share
(Google has the "first mover advantage" and – so far – ample resources to stay the entry point to the Web)

Right strategy by OSF:
Create a European Crawl Index first!

Generic European Web search engine second?

Disclaimer:

WHAT TO DO?

PUSH SEARCH TO THE EDGE!

Decentralize Web Search

yacy.net/

WHAT TO DO?

DECENTRALIZE WEB SEARCH?

502	GET	yacy.searchlab.eu	yacysearch.html?query=google+instant+search&Enter=&contentdom=text&strictContent	document	html	330 B	173 B	60047 ms
200	GET	yacy.searchlab.eu	favicon.ico	FaviconLoader.js...	x-icon	cached	1.05 KB	0 ms

Log in

google instant search | search again

Context Ranking | Sort by Date

Documents | Images

https (3) | http (1)

terms style operating conditions life family sellers samsung cloth weapons plus free guardian subscription december system

Provider [4]
- flirt4free.com (1)
- samsung.com (1)
- theguardian.com (1)
- googlesystem.blogspot.com (1)

Filetype [1]
- php (1)

Language [1]
- English (4)

Please support YaCy!

Flattr | Flattr this!

PayPal | beneficial: 5 € | Donate!

If you run a YaCy server, feel free to replace our donation plea with your own support message, use the Portal Configuration servlet.

1-4 of **4** ; (**0** local, **4** remote from **1** YaCy peers).

Google Operating System: December 2006
http://googlesystem.blogspot.com/2006/12/
Sun, 11 Oct 2020 | Citations |

Life and style | The Guardian
https://www.theguardian.com/uk/lifeandstyle
Sun, 31 Jul 2016 | Citations |

Samsung TV Plus - Subscription-Free TV | Samsung US
https://www.samsung.com/us/televisions-home-theater/tvs/tvplus/
Sun, 11 Oct 2020 | Citations |

Terms & Conditions | Flirt4Free
https://www.flirt4free.com/terms.php
Thu, 02 Apr 2020 | Citations |

"Is that what he meant with slow search?!"

viewpoints

DOI:10.1145/2633041

Jaime Teevan, Kevyn Collins-Thompson, Ryen W. White, and Susan Dumais

Viewpoint
Slow Search

Seeking to enrich the search experience by allowing for extra time and alternate resources.

WE LIVE IN a world where the pace of everything from communication to transportation is getting faster. In re-

PUSH SEARCH TO THE EDGE

CREATE HUMAN-CENTRIC SEARCH

Safely gain access to rich personal data:

- Email
- Browsing history
- Documents read
- Contents of the user's home directory (*i.e.*, documents written as well!)

Is this new? Well, we used to call this "Personal Information Management". I like human-centric search!

Do we still need that log data if we can have all that?!

Can high quality evidence about an individual's recurring long-term interests replace the shallow information of many?

PUSH SEARCH TO THE EDGE

CREATE HUMAN-CENTRIC SEARCH

"Even more broadly than trying to get people the right content based on their context, we as a community need to be thinking about how to support people through the entire search experience."

Jaime Teevan on "Slow Search"

Search as a short- and long-term dialogue (with or without "conversational search")

A Personal Search Engine, **in the Edge**

WEB INDEX AT HOME

REALISTIC?

Clueweb 2012: 80TB
Recent CommonCrawl (August 2017): 3.28B pages, 280TB

Average web page takes up 320 KB

- Large sample collected with Googlebot, May 26th, 2010
- Reported 4.2B pages *(would require ~1.3 Petabyte)*

De Kunder & Van de Bosch estimate an upper bound of ~50B pages

- http://www.worldwidewebsize.com/

Also considering continuing growth *(claimed in unpublished work)*

- Andrew Trotman, Jinglan

https://web.archive.org/web/20100628055041/http://code.google.com/speed/articles/web-metrics.html

WEB INDEX AT HOME

REALISTIC?!

Who actually needs all of the Web if their search engine is truly personal?

E.g., I cannot read more than 4 or 5 languages (and even those...)

E.g., I do not need the club league for soccer

And...

I could always fall back to using an "**out-dated, non-personalized**" Web Search engine...

... should I suddenly feel an urgent need to search for the soccer club league!

PUSH SEARCH TO THE EDGE

TWO PROBLEMS

How to get the web data on the personal search engine?

How to replace the lack of usage data from many?

Wednesday October14th, 11:00, presentation by Djoerd Hiemstra

BUNDLE THE INDEX

Idea:

Organize the web crawl in topically related bundles

Apply bittorrent-like decentralization to share & update bundles

webtorrent.io , IPFS.io, academictorrents.com

Use techniques inspired by query obfuscation to hide the real user's interests when downloading

WEB ARCHIVES TO THE RESCUE

Idea:

Web Archives already store the data that the personal search engine would need

Just not (yet) organized in topical and temporary bundles

Win-win situation:

A business model for archiving?

A way to enrich the (rarely used) web archives with usage data?

A way to crowd-source seed-lists for crawling?

"Rescue the Web Archives"

AN ANALOGY

"… communication and media limitations, due to the distance between Earth and Mars, resulting in time delays: they will have to request the movies or news broadcasts they want to see in advance.

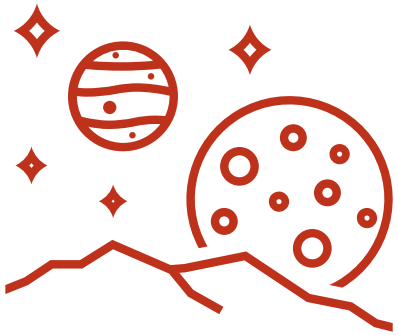
[…]

Easy Internet access will be limited to their preferred sites that are constantly updated on the local Mars web server. Other websites will take between 6 and 45 minutes to appear on their screen - first 3-22 minutes for your click to reach Earth, and then another 3-22 minutes for the website data to reach Mars."

http://www.mars-one.com/faq/mission-to-mars/what-will-the-astronauts-do-on-mars

AN ANALOGY

Web Archive



Personal Search Engine

PRE-FETCHING & CACHING

Hide latencies of getting the data from the live web:

- Pre-fetch pages linked from initial query results page
- Pre-fetch additional related pages
- Pre-fetches expanded with those from query suggestions

- Cache web data to avoid accessing the live web

Related work:

Jimmy Lin, Charles L. A. Clarke, and Gaurav Baruah. *Searching from Mars.* Internet Computing, 20(1):77-82, 2016. http://dx.doi.org/10.1109/MIC.2016.2

Charles L.A. Clarke, Gordon V. Cormack, Jimmy Lin, and Adam Roegiest. *Total Recall: Blue Sky on Mars.* ICTIR '16. http://dx.doi.org/10.1145/2970398.2970430

Charles L. A. Clarke, Gordon V. Cormack, Jimmy Lin, Adam Roegiest. *Ten Blue Links on Mars.* https://arxiv.org/abs/1610.06468

PERSONAL WEB ARCHIVES

Caching Web data at home

Build a Personal Web Archive (PWA) while browsing

WASP, with WebIS:

github.com/webis-de/wasp/

Prizm by Jimmy Lin (personal Web archiving on a Raspberry Pi)

Extend the PWA, considering this as a seed

P-o-C in student project extending WASP (by Gijs Hendriksen)

BLUEPRINT

THE PERSONAL SEARCH ENGINE

Push Search to the Edge

Human-centric Search

Exploit the rich source data that can be processed safely locally

Webarchives to the rescue

Super-peers in a P2P network of personal search engines