Graphs in Search Engines

EN 500.111 Week 5

Announcement

- Class Presentations
  - The last day of class (11/17 for Tuesday section, 11/18 for Wednesday section)
  - Groups of 3 people
  - Topic of your choice - any application area of graphs not covered in class
- Presentation Details
  - 12-15 minutes in length, plus time for questions
  - Detailed guidelines: https://github.com/mkirsche/TAG2020/blob/master/presentation/guidelines.md
- Forming Groups
  - Thread on Piazza for finding groupmates will be posted today
  - Next class, I'll ask you for **a list of team members, as well as a team name and topic area**.

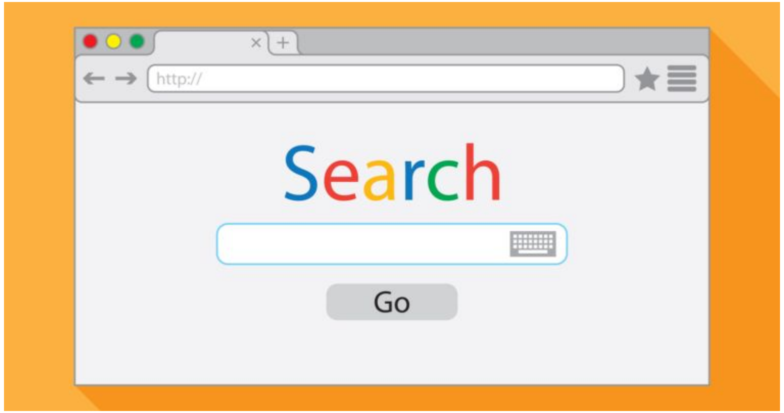
Outline

- Hyperlink graphs
- Knowledge graphs

Introductory Discussion

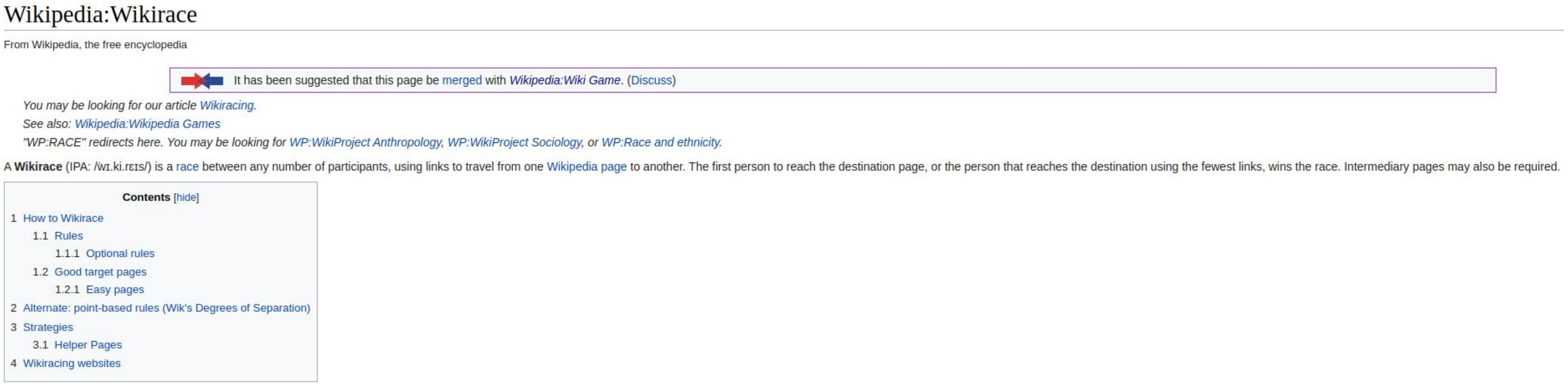
When do you use search engines, and what would make you think a search engine is doing its job well?

https://www.searchenginejournal.com/alternative-search-engines/271409/



Hyperlink Graphs

Wikipedia Race



Wikipedia:Wikirace

From Wikipedia, the free encyclopedia

It has been suggested that this page be merged with *Wikipedia:Wiki Game*. (Discuss)

You may be looking for our article Wikiracing.

See also: Wikipedia:Wikipedia Games

"WP:RACE" redirects here. You may be looking for WP:WikiProject Anthropology, WP:WikiProject Sociology, or WP:Race and ethnicity.

A **Wikirace** (IPA: /wɪ.ki.rɛɪs/) is a race between any number of participants, using links to travel from one Wikipedia page to another. The first person to reach the destination page, or the person that reaches the destination using the fewest links, wins the race. Intermediary pages may also be required.

Contents [hide]

Example

Starting page: https://en.wikipedia.org/wiki/Graph_theory

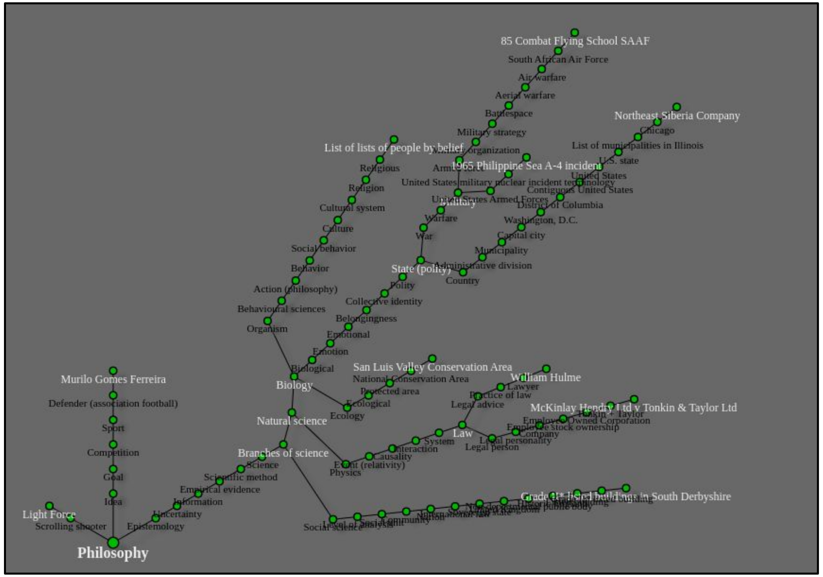
Ending page: https://en.wikipedia.org/wiki/Peanut_butter

Try to get from the start to the end only by clicking on links in each article. The goal is to use as few “hops” as possible.

One Path

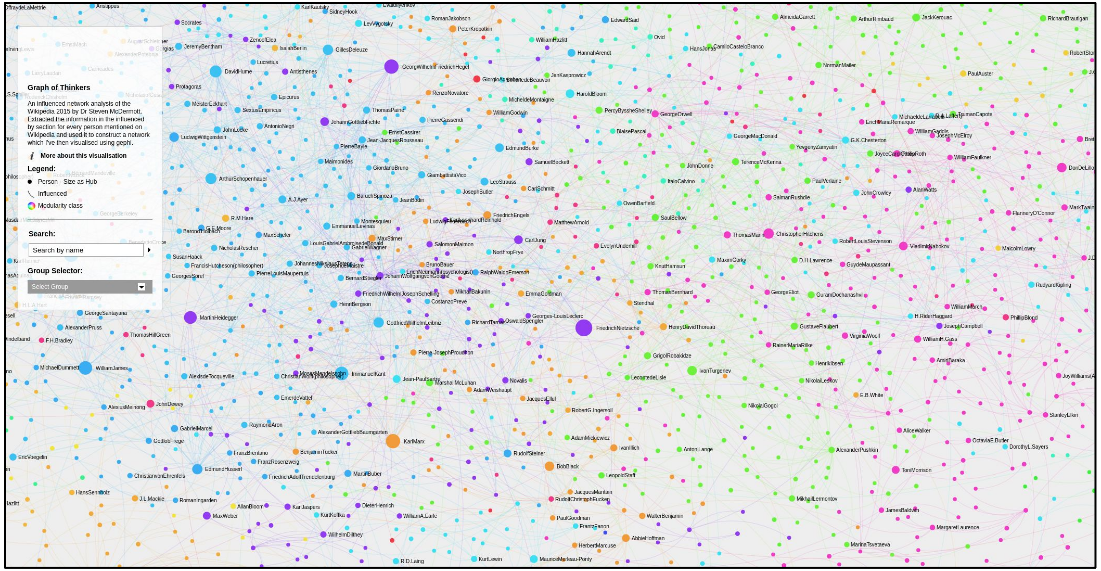
- https://en.wikipedia.org/wiki/Graph_theory
- https://en.wikipedia.org/wiki/Hassler_Whitney
- https://en.wikipedia.org/wiki/New_York_City
- https://en.wikipedia.org/wiki/United_States
- https://en.wikipedia.org/wiki/Obesity_in_the_United_States
- https://en.wikipedia.org/wiki/Convenience_food
- https://en.wikipedia.org/wiki/Nut_(fruit)
- https://en.wikipedia.org/wiki/Food_allergy
- https://en.wikipedia.org/wiki/Peanut
- https://en.wikipedia.org/wiki/Peanut_butter

Everything leads back to philosophy



https://xefer.com/WIKIPEDIA

Visualizing Wikipedia’s Hyperlinks



Graph of Thinkers

An influenced network analysis of the Wikipedia 2015 by Dr Steven McDermott. Extracted the information in the influenced by section for every person mentioned on Wikipedia and used it to construct a network which I've then visualised using gephi.

i **More about this visualisation**

Legend:

- ● Person - Size as Hub
- ╲ Influenced
- ◍ Modularity class

Search:

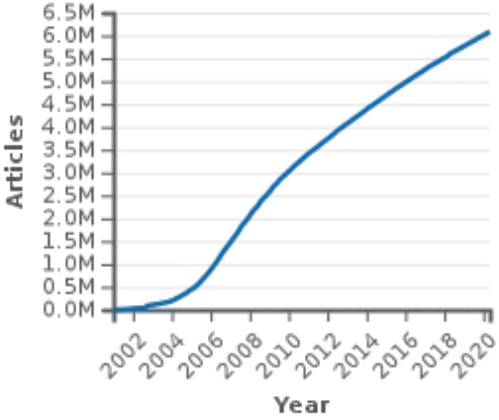
Search by name ▸

Group Selector:

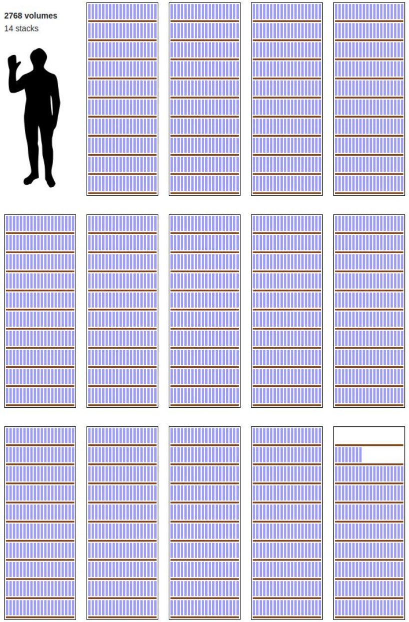
Select Group ▾

https://d100dymyf6nv12.cloudfront.net/

Yet another case of really big graphs

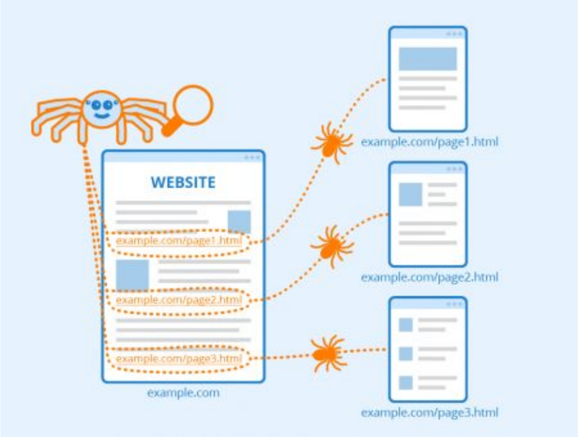


https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia



How is this related to search engines?

- Search engines “crawl the web”, meaning they follow hyperlinks to find different webpages
- Can be framed as traversing part of a much bigger version of Wikipedia’s graph
- Used to identify popularly referenced pages



https://www.seobility.net/en/wiki/Search_Engine_Crawlers

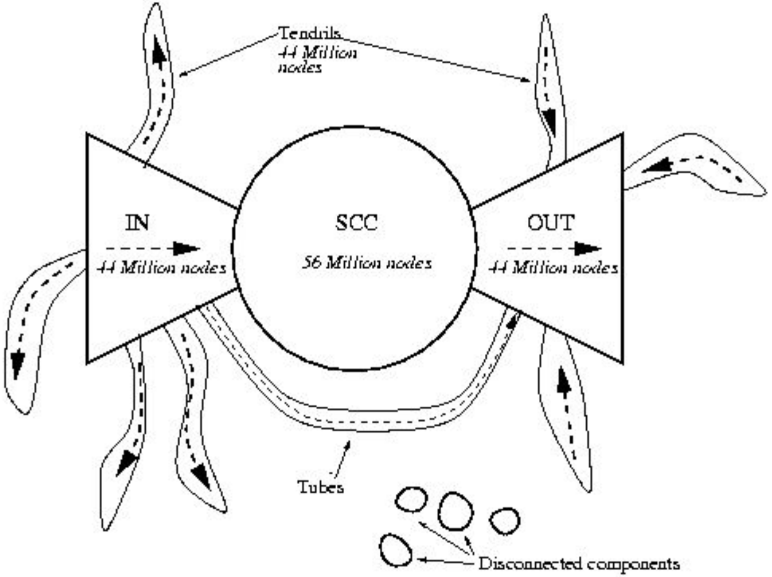
Web Crawling

Four main components:

- ➔ Selection policy: deciding what pages to index
- ➔ Revisit policy: deciding when to update a page's content
- ➔ Politeness policy: avoid overloading servers of websites with too frequent accesses
- ➔ Parallelization policy: coordinate with other crawlers operating in parallel

Each web crawler starts at some page and looks at everything (or less depending on selection policy) that can be reached from that node - but is this the whole web?

So what does the Internet look like?



Results from an Altavista crawl of ~200 million webpages (nodes) as of 1999

Everything in the middle portion has paths to each other

Where would we want to start a web crawl from?

http://www.ra.ethz.ch/cdstore/www9/160/160.html

Crawler Exploration

Exploration algorithm:

1. Choose some set of seed URLs and add them to the "frontier" of pages to consider
2. Pick a page from the frontier and add it to the web index (or update its content)
3. Parse all the links from the chosen page and add those pages to the frontier
4. Repeat steps 2 and 3 until some termination condition is met

What terminating conditions would make sense here?

The k-seed h-hop model

One way to model which sites get visited in a web crawl:

- Assume the number of seeds is some fixed number k
- Make the terminating condition that only pages within h hops (hyperlinks) of the seeds get visited
- This requires some balancing of k and h to make sure the number of visited pages fits in the index

This is essentially the same as running BFS from all k seeds at once and stopping once the distances get to h.

Seed Selection

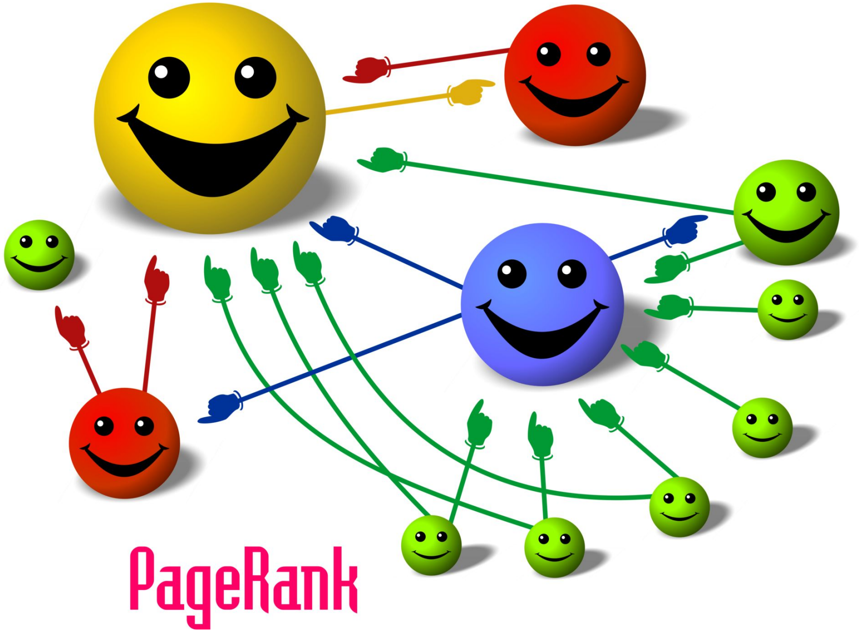
Many people have studied the problem of selecting the best k seeds for this algorithm

... but what does “best” mean here?

What pages are worth indexing?

PageRank is an algorithm used by Google to determine the “value” of each web page - in this example the size of each face is proportional to the total size of all faces pointing to it, which is the same principle behind Pagerank scores

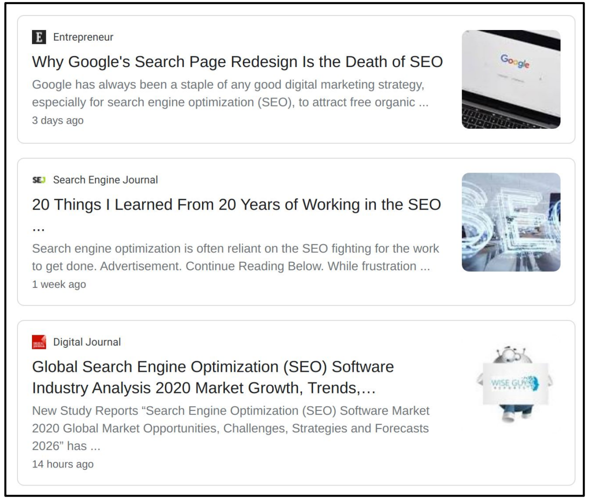
https://en.wikipedia.org/wiki/PageRank



Search Engine Optimization

Search engine optimization is a strategy where people try to figure out a search engine's "page quality value formula" and try to exploit it to drive more traffic to their site

This leads to a conflict between increasingly complex page-quality measurements and people trying to decipher and take advantage them them



Entrepreneur

Why Google's Search Page Redesign Is the Death of SEO

Google has always been a staple of any good digital marketing strategy, especially for search engine optimization (SEO), to attract free organic ...

3 days ago

Search Engine Journal

20 Things I Learned From 20 Years of Working in the SEO ...

Search engine optimization is often reliant on the SEO fighting for the work to get done. Advertisement. Continue Reading Below. While frustration ...

1 week ago

Digital Journal

Global Search Engine Optimization (SEO) Software Industry Analysis 2020 Market Growth, Trends,...

New Study Reports "Search Engine Optimization (SEO) Software Market 2020 Global Market Opportunities, Challenges, Strategies and Forecasts 2026" has ...

14 hours ago

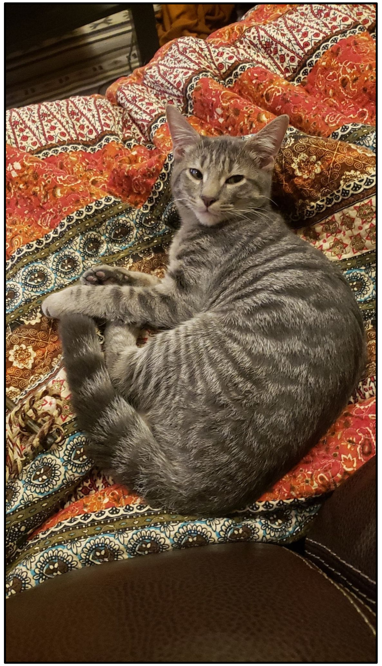
Seed Selection to Maximize Page Quality

- Now that we have a way to define a page’s quality, we can use that to assess a choice of web crawling seeds
- Better seeds should lead to higher quality pages beng discovered/indexed

Formal statement of seed selection problem: Given a graph with weights on every nodes (indicating page quality), select k seed nodes such that the sum of the weights of all nodes reachable within h hops of the seeds is maximized.

Maximum K-Coverage Problem

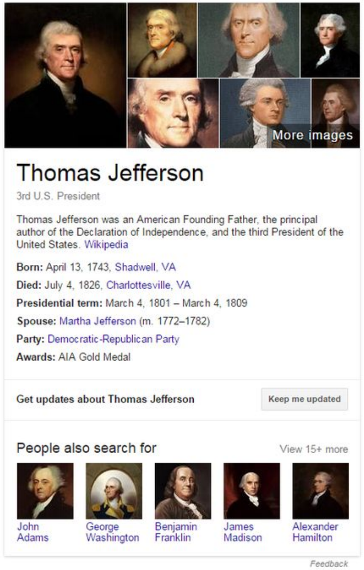
- The graph problem from the previous slide is a known problem called the maximum k-coverage problem, and it's NP-hard (the amount of time it takes to solve is exponential with respect to the graph size)
- So a number of heuristic solutions have been used:
  - Choose webpages with a **high degree**
  - Evaluate the **value reached within a smaller number of hops than h** (say 2 or 3) and choose seeds based on those
  - Other more complex approaches based on decomposition into strongly connected components



Knowledge Graphs

Answering Queries without links

- ➔ In recent years, Google has started providing “knowledge boxes” with facts about key people or other subjects
- ➔ Answers about one third of queries this way with much less time required for the user



Thomas Jefferson

3rd U.S. President

Thomas Jefferson was an American Founding Father, the principal author of the Declaration of Independence, and the third President of the United States. Wikipedia

Born: April 13, 1743, Shadwell, VA

Died: July 4, 1826, Charlottesville, VA

Presidential term: March 4, 1801 – March 4, 1809

Spouse: Martha Jefferson (m. 1772–1782)

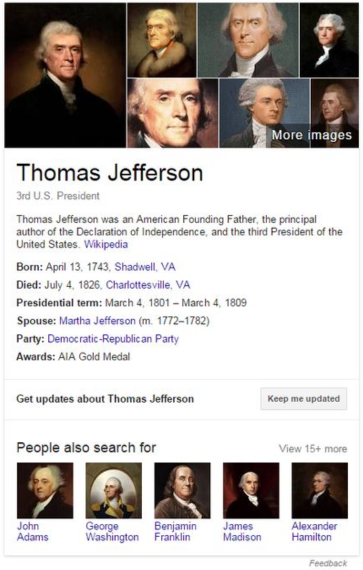
Party: Democratic-Republican Party

Awards: AIA Gold Medal

Get updates about Thomas Jefferson

https://en.wikipedia.org/wiki/Knowledge_Graph

Other Search Engines are following suit



Thomas Jefferson

3rd U.S. President

Thomas Jefferson was an American Founding Father, the principal author of the Declaration of Independence, and the third President of the United States. Wikipedia

Born: April 13, 1743, Shadwell, VA

Died: July 4, 1826, Charlottesville, VA

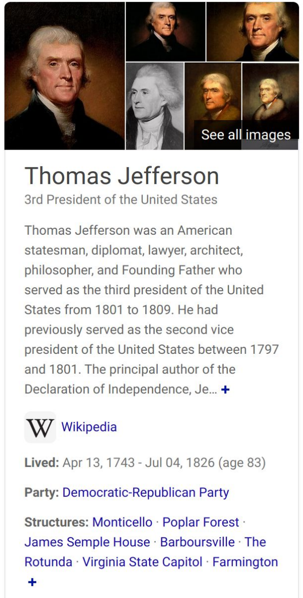
Presidential term: March 4, 1801 – March 4, 1809

Spouse: Martha Jefferson (m. 1772–1782)

Party: Democratic-Republican Party

Awards: AIA Gold Medal

Get updates about Thomas Jefferson



Thomas Jefferson

3rd President of the United States

Thomas Jefferson was an American statesman, diplomat, lawyer, architect, philosopher, and Founding Father who served as the third president of the United States from 1801 to 1809. He had previously served as the second vice president of the United States between 1797 and 1801. The principal author of the Declaration of Independence, Je… +

Wikipedia

Lived: Apr 13, 1743 - Jul 04, 1826 (age 83)

Party: Democratic-Republican Party

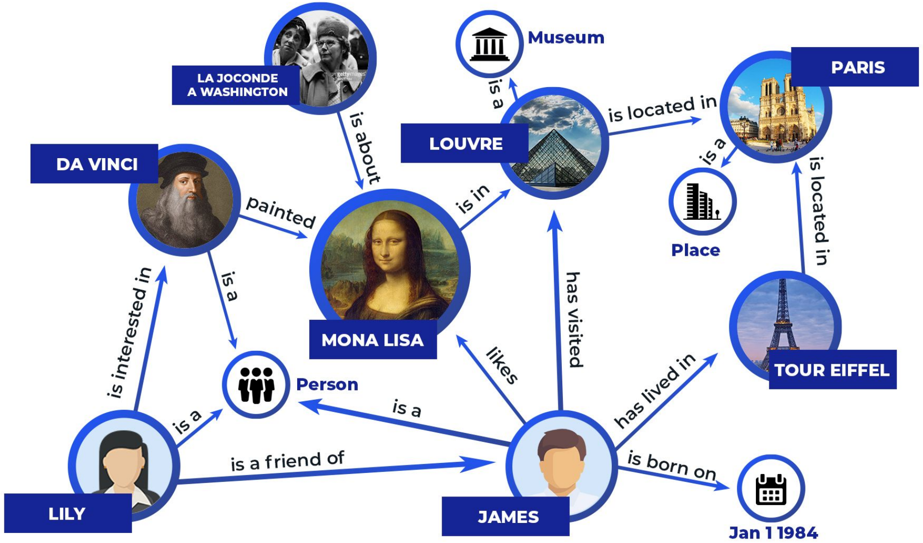
Structures: Monticello · Poplar Forest · James Semple House · Barboursville · The Rotunda · Virginia State Capitol · Farmington +

Google (left) and Bing (right)

But what is a “Knowledge Graph”?

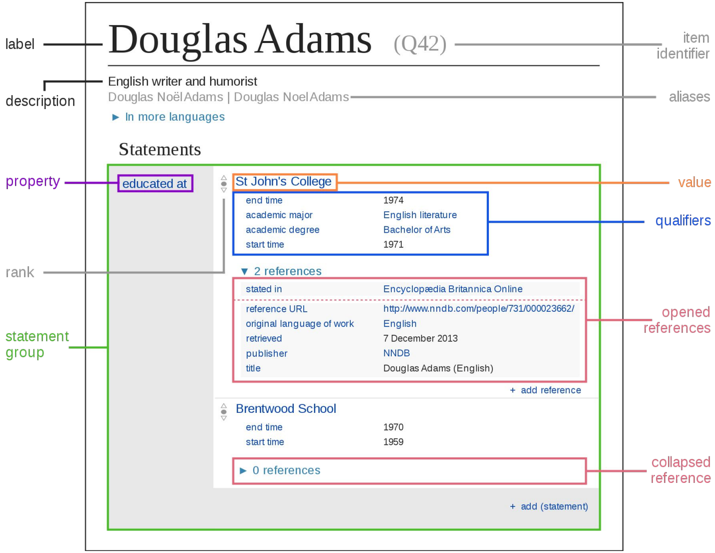
- ➔ Nodes are both subjects and abstract concepts
- ➔ Edges annotated with many different kinds of relationship
- ➔ Sometimes even edges for users based on information about them

https://yashuseth.blog/2019/10/08/introduction-question-answering-knowledge-graphs-kgqa/



Similar graphs available for other data repositories such as Wikidata

- ➔ Each data entry on Wikidata has “statements” linking them to other data entries



Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

► In more languages

Statements

educated at

St John's College

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971

▼ 2 references

stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)

+ add reference

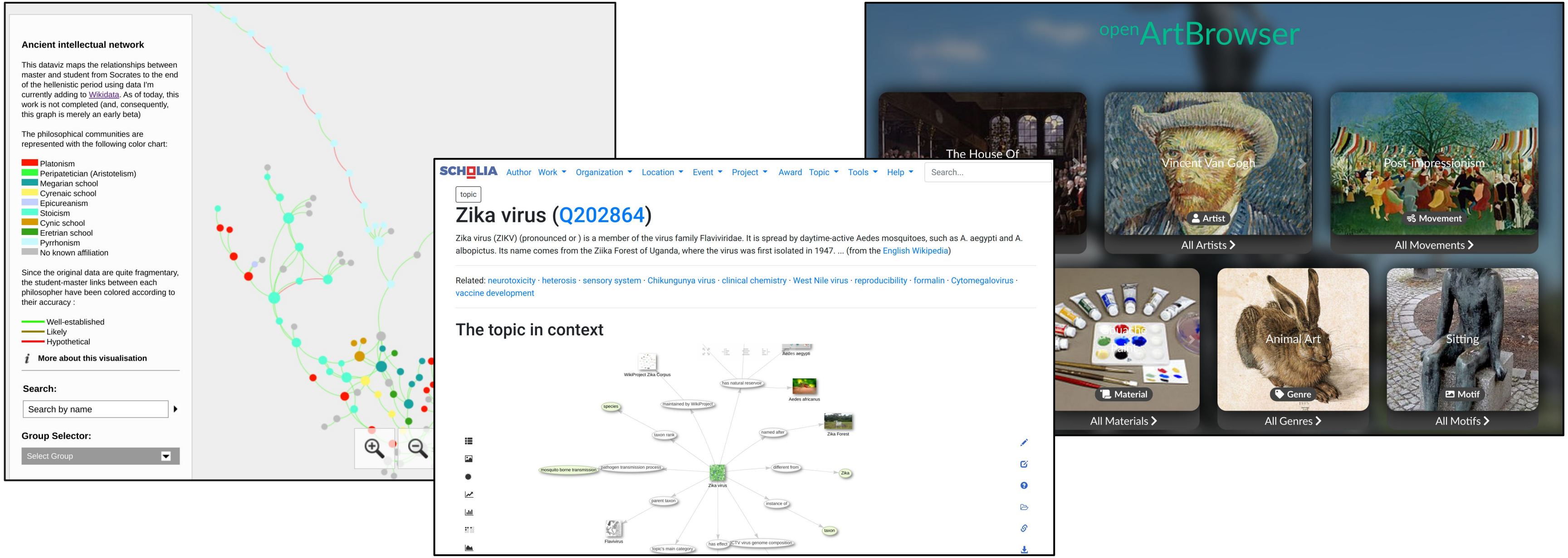
Brentwood School

end time	1970
start time	1959

► 0 references

+ add (statement)

Examples of Wikidata Visualization



Ancient intellectual network

This dataviz maps the relationships between master and student from Socrates to the end of the hellenistic period using data I'm currently adding to Wikidata. As of today, this work is not completed (and, consequently, this graph is merely an early beta)

The philosophical communities are represented with the following color chart:

- 🟥 Platonism
- 🟩 Peripatetician (Aristotelism)
- 🟩 Megarian school
- 🟨 Cyrenaic school
- 🟦 Epicureanism
- 🟩 Stoicism
- 🟧 Cynic school
- 🟩 Eretrian school
- 🟦 Pyrrhonism
- ⬜ No known affiliation

Since the original data are quite fragmentary, the student-master links between each philosopher have been colored according to their accuracy :

- — Well-established
- — Likely
- — Hypothetical

More about this visualisation

Search:

Search by name

Group Selector:

Select Group

Zika virus (Q202864)

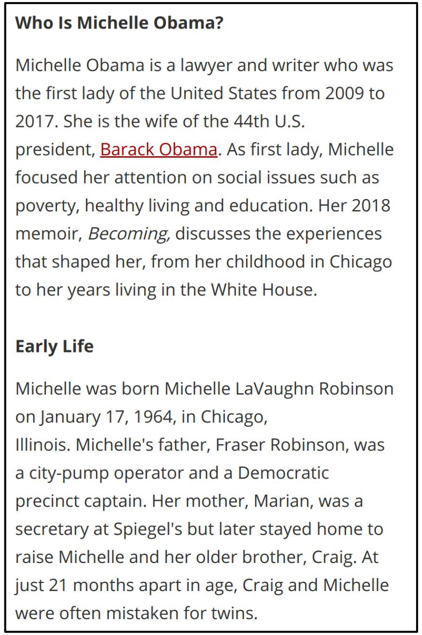
Zika virus (ZIKV) (pronounced or) is a member of the virus family Flaviviridae. It is spread by daytime-active Aedes mosquitoes, such as A. aegypti and A. albopictus. Its name comes from the Ziika Forest of Uganda, where the virus was first isolated in 1947. ... (from the English Wikipedia)

Related: neurotoxicity · heterosis · sensory system · Chikungunya virus · clinical chemistry · West Nile virus · reproducibility · formalin · Cytomegalovirus · vaccine development

The topic in context

Building Knowledge Graphs

- Given that these graphs contain millions of concepts, there needs to be a way to infer them based on context
- What nodes and edges can be inferred from the text to the right?



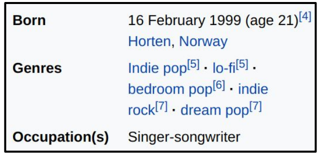
Who Is Michelle Obama?

Michelle Obama is a lawyer and writer who was the first lady of the United States from 2009 to 2017. She is the wife of the 44th U.S. president, Barack Obama. As first lady, Michelle focused her attention on social issues such as poverty, healthy living and education. Her 2018 memoir, *Becoming,* discusses the experiences that shaped her, from her childhood in Chicago to her years living in the White House.

Early Life

Michelle was born Michelle LaVaughn Robinson on January 17, 1964, in Chicago, Illinois. Michelle's father, Fraser Robinson, was a city-pump operator and a Democratic precinct captain. Her mother, Marian, was a secretary at Spiegel's but later stayed home to raise Michelle and her older brother, Craig. At just 21 months apart in age, Craig and Michelle were often mistaken for twins.

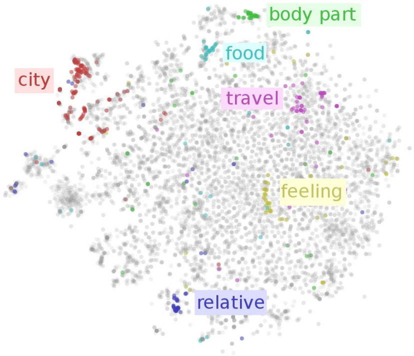
Automated Knowledge Graph Building



Born	16 February 1999 (age 21)[4] Horten, Norway
Genres	Indie pop[5] · lo-fi[5] · bedroom pop[6] · indie rock[7] · dream pop[7]
Occupation(s)	Singer-songwriter

(Wikipedia)

- ➔ Some websites make this very easy (like Wikipedia, which is the source of much of Google's Knowledge Graph)
- ➔ But in general, it requires using the context around a word to determine information like whether it's a person or not, what their occupation is, and more!
- ➔ **Word embeddings** are a way of doing this
- ➔ Main idea:
  - ◆ Define each word as a huge list of numbers indicating how often it appears near other words
  - ◆ Use machine learning models to cluster these embeddings and learn patterns



https://ruder.io/word-embeddings-1/

Conclusions

- ➔ Search engines use graphs for many of their core functions
- ➔ **Hyperlink graphs** are used to represent how web crawlers select pages for indexing and re-indexing
- ➔ **Knowledge graphs** are used to store information in a systematic, query-friendly way that enables getting the user the answer they seek
- ➔ All of these graphs are very large, so they rely on heuristics and data models to solve problems whose theoretically optimal solutions are infeasible

Questions?