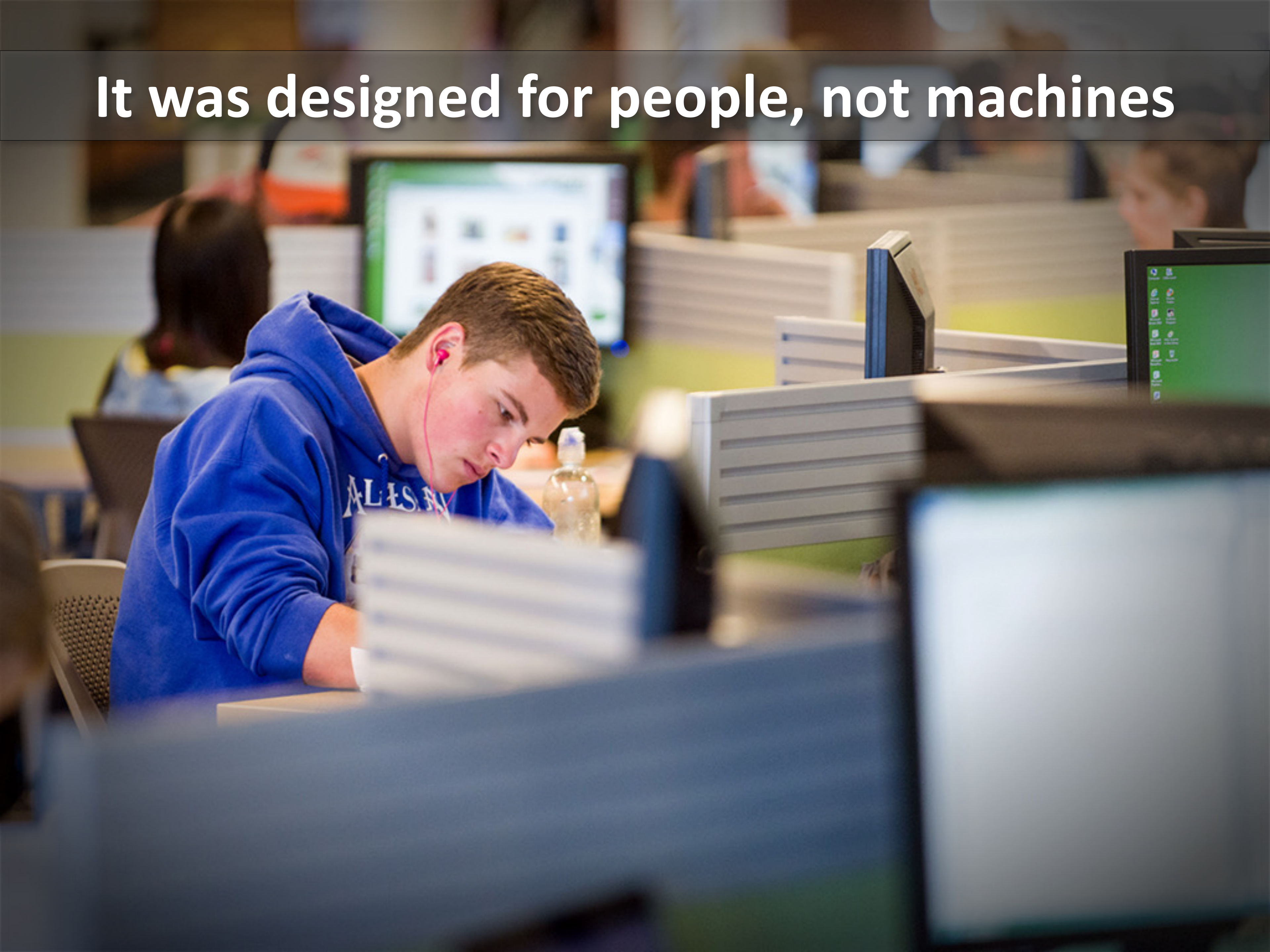
NLP*

from Strings to Things

* natural language processing

The Web is our greatest knowledge source

But it has limitations


It was designed for people, not machines

It was designed for people, not machines

- Its content is mostly text, spoken language, images and videos
- These are easy for people to understand
- But hard for machines

Machines need access to this knowledge too

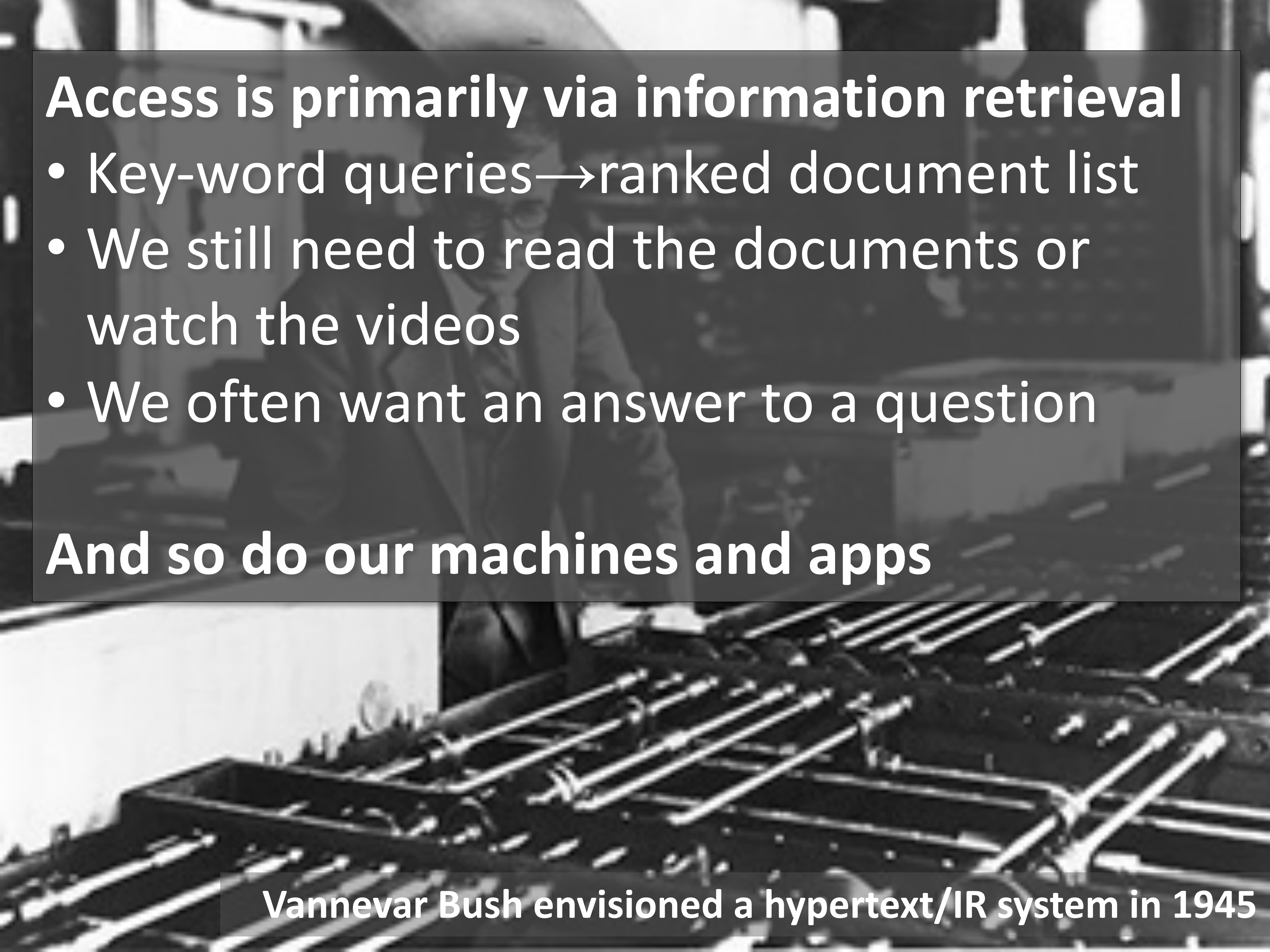
Access is primarily via information retrieval

Vannevar Bush envisioned a hypertext/IR system in 1945

Access is primarily via information retrieval

- Key-word queries→ranked document list
- We still need to read the documents or watch the videos
- We often want an answer to a question

And so do our machines and apps



Vannevar Bush envisioned a hypertext/IR system in 1945

We need to add knowledge graphs

We need to add knowledge graphs

- High quality semi-structured information about entities, events and relations
- Represented & accessed via standard APIs
- Easily integrated, fused and reasoned with

State of the Art?

Google is a good example, but Microsoft, IBM, Apple and Facebook all have similar capabilities

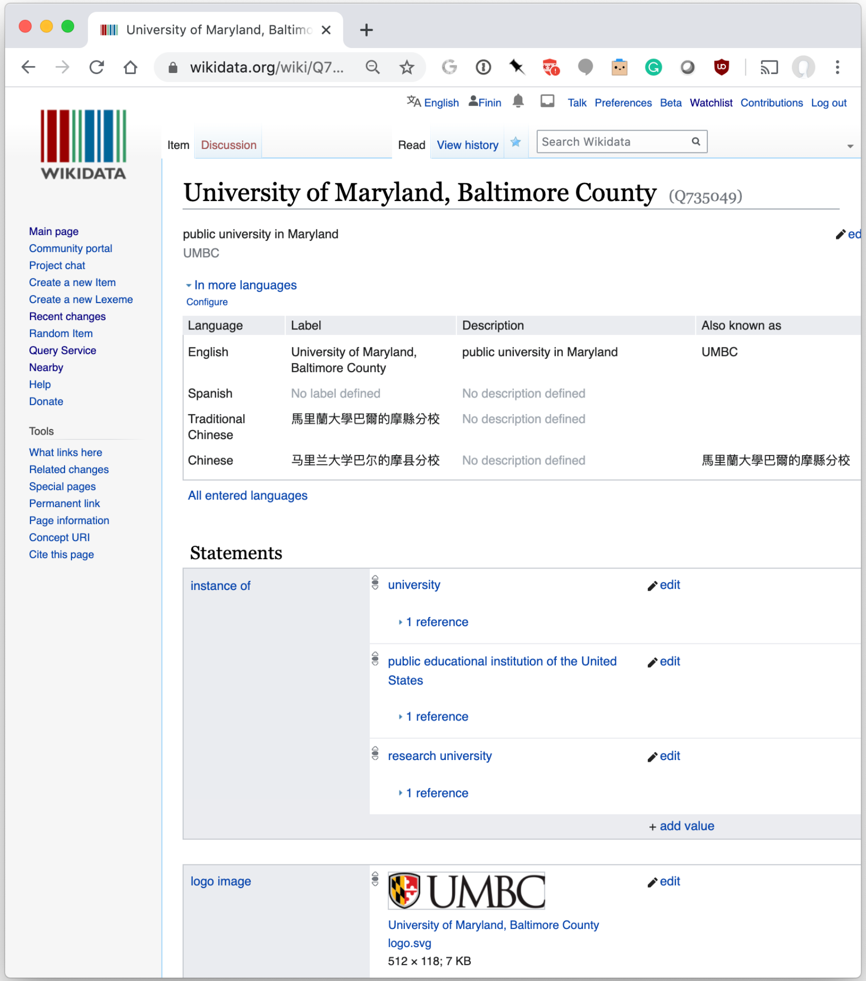
- 2010 Google acquired MediaWeb and its **Freebase** KB
- 2014: Freebase: 1.2B facts about 43M entities
- 2015+: Google knowledge graph, updated by text IE

DBpedia open source RDF KB is another

- 800M facts about 4.6M subjects from English **Wikipedia**, data also available in 21 other languages
- Helps integrate 90B facts from 1000 RDF datasets in the linked data cloud

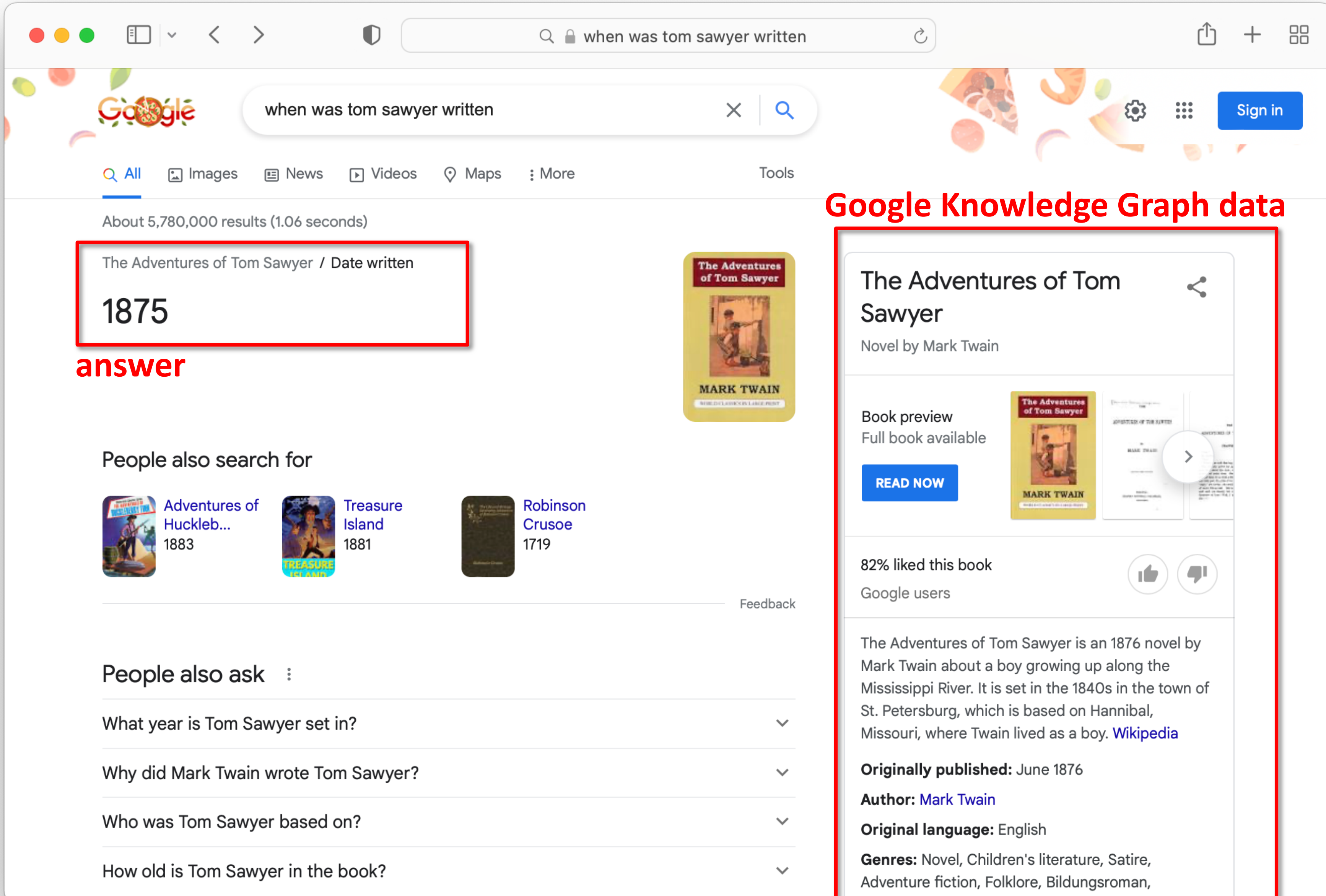
Wikidata Knowledge Graph

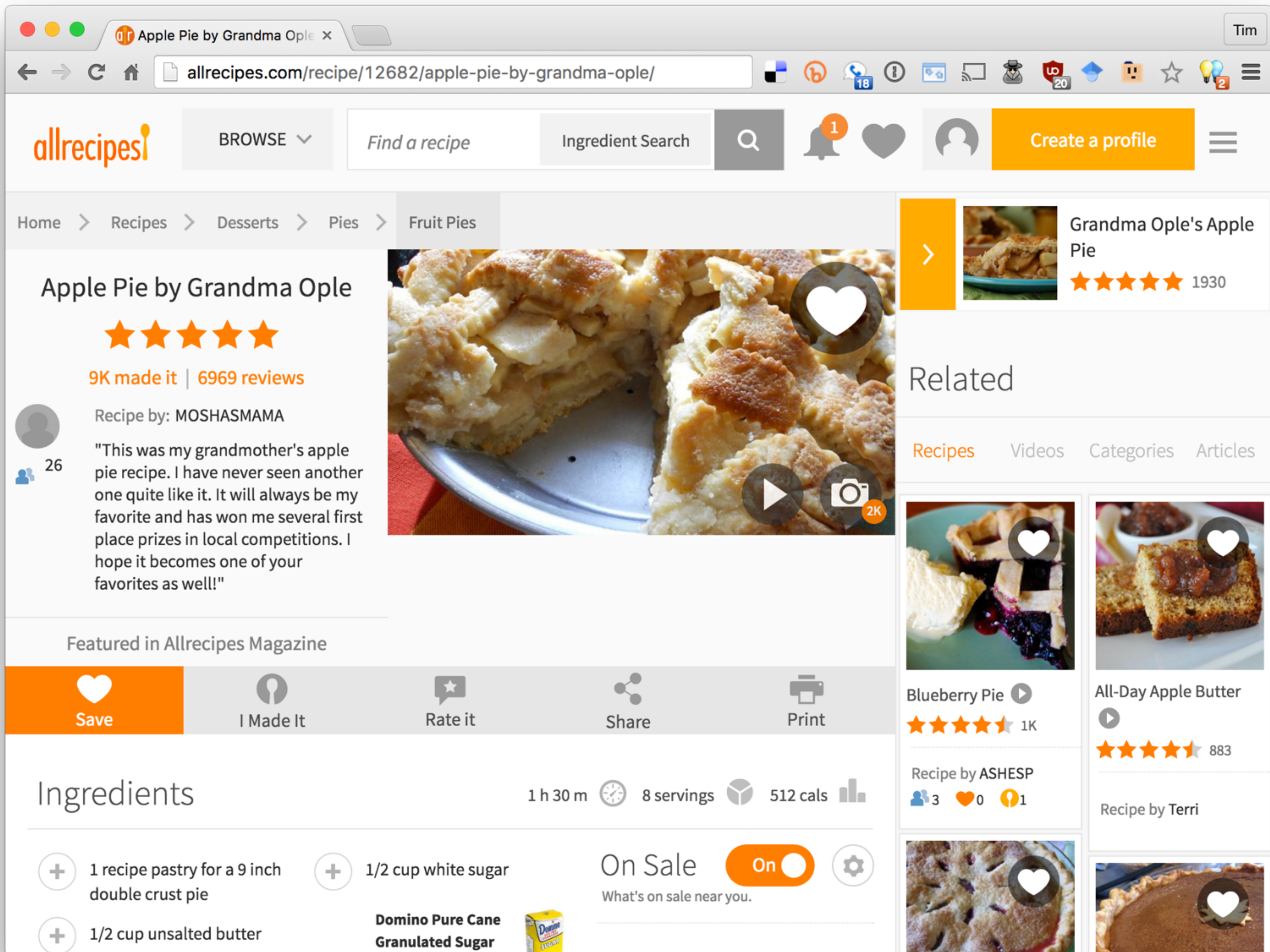
- **Large knowledge graph** with 1B statements about ~72M items
- **Fine-grained ontology**: ~2M types; ~5K properties
- **Multilingual**, strings tagged with language id
- Links to entity's **Wikimedia pages**
- Entities have a canonical **name** and **aliases** in multiple languages and multiple claims
- UMBC=Q64780099, with type University, 569 statements
- Editable by humans and bots
- Can query with SPARQL query language



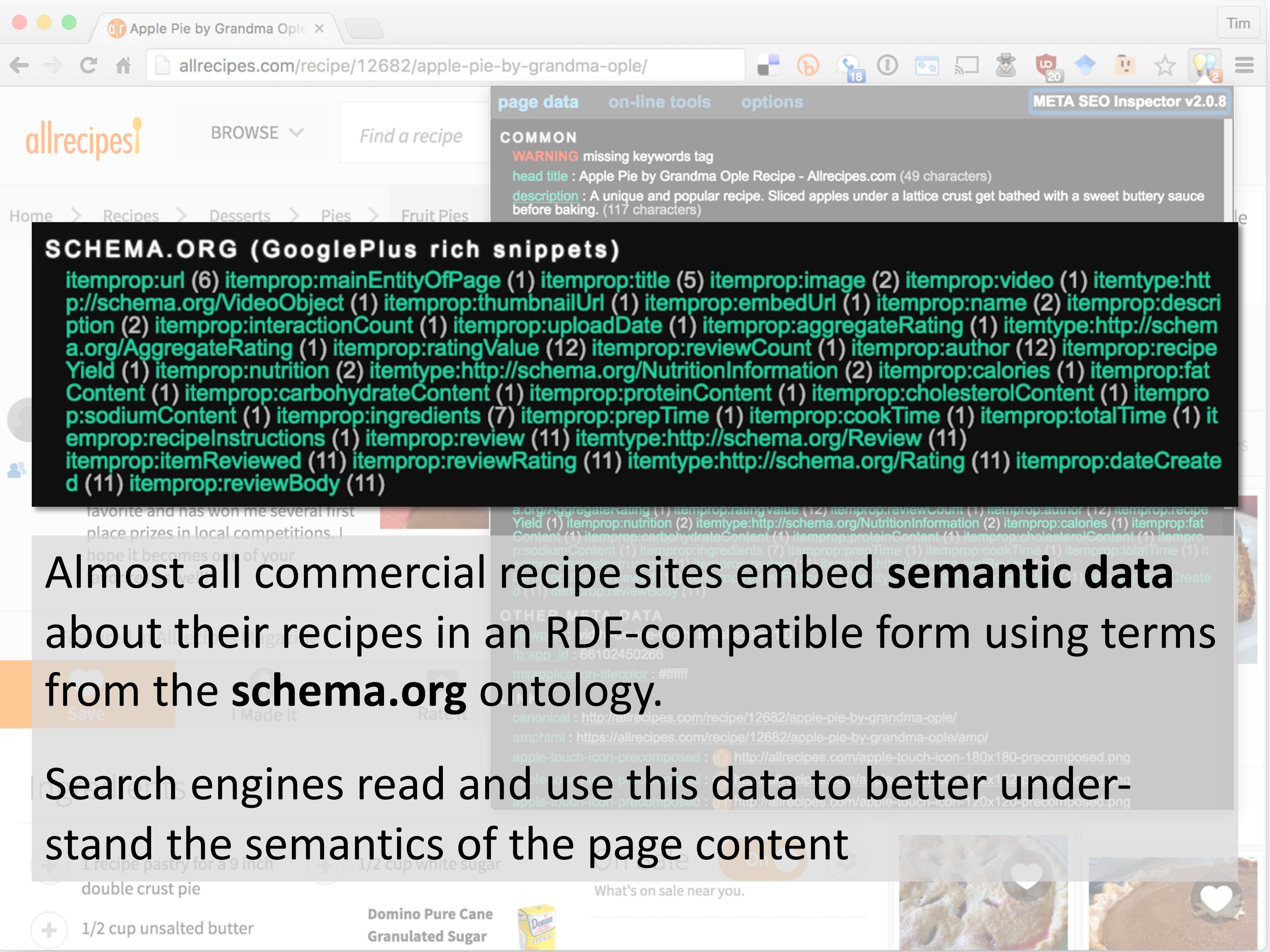
[Q64780099](Q64780099)

Ask: When was Tom Sawyer written?





Apple Pie by Grandma Ople
Tim
allrecipes.com/recipe/12682/apple-pie-by-grandma-ople/
allrecipes
BROWSE
Find a recipe
Ingredient Search
1
Create a profile
Home
Recipes
Desserts
Pies
Fruit Pies
Apple Pie by Grandma Ople
9K made it | 6969 reviews
Recipe by: MOSHASMAMA
26
"This was my grandmother's apple pie recipe. I have never seen another one quite like it. It will always be my favorite and has won me several first place prizes in local competitions. I hope it becomes one of your favorites as well!"
2K
Featured in Allrecipes Magazine
Save
I Made It
Rate it
Share
Print
Ingredients
1 h 30 m
8 servings
512 cals
1 recipe pastry for a 9 inch double crust pie
1/2 cup unsalted butter
1/2 cup white sugar
Domino Pure Cane Granulated Sugar
On Sale
On
What's on sale near you.
Grandma Ople's Apple Pie
1930
Related
Recipes
Videos
Categories
Articles
Blueberry Pie
1K
Recipe by ASHESP
3
0
1
All-Day Apple Butter
883
Recipe by Terri



Almost all commercial recipe sites embed **semantic data** about their recipes in an RDF-compatible form using terms from the **schema.org** ontology.

Search engines read and use this data to better understand the semantics of the page content

Conversational Bots

Voice-driven conversational systems like Amazon Echo and Google Home use knowledge graphs to help understand our requests

Where does the knowledge come from?

- Initial knowledge graphs like *DBpedia* and *Freebase* started with data from **Wikipedia** and encoded it in custom ontologies
- Current focus is on extracting information from text of source documents, e.g., journal articles, Newswire, social media, etc.

NIST Text Analysis Conference

- Annual evaluation workshops since 2008 on natural language processing & related applications with large test collections and common evaluation procedures
- **Knowledge Base Population** (KBP) tracks focus on building KBs from information extracted from text
  - **Cold Start KBP**: construct a KB from text
  - **Entity discovery & linking**: cluster and link entity mentions
  - Slot filling
  - Slot filler validation
  - Sentiment
  - Events: discover and cluster events in text

http://nist.gov/tac

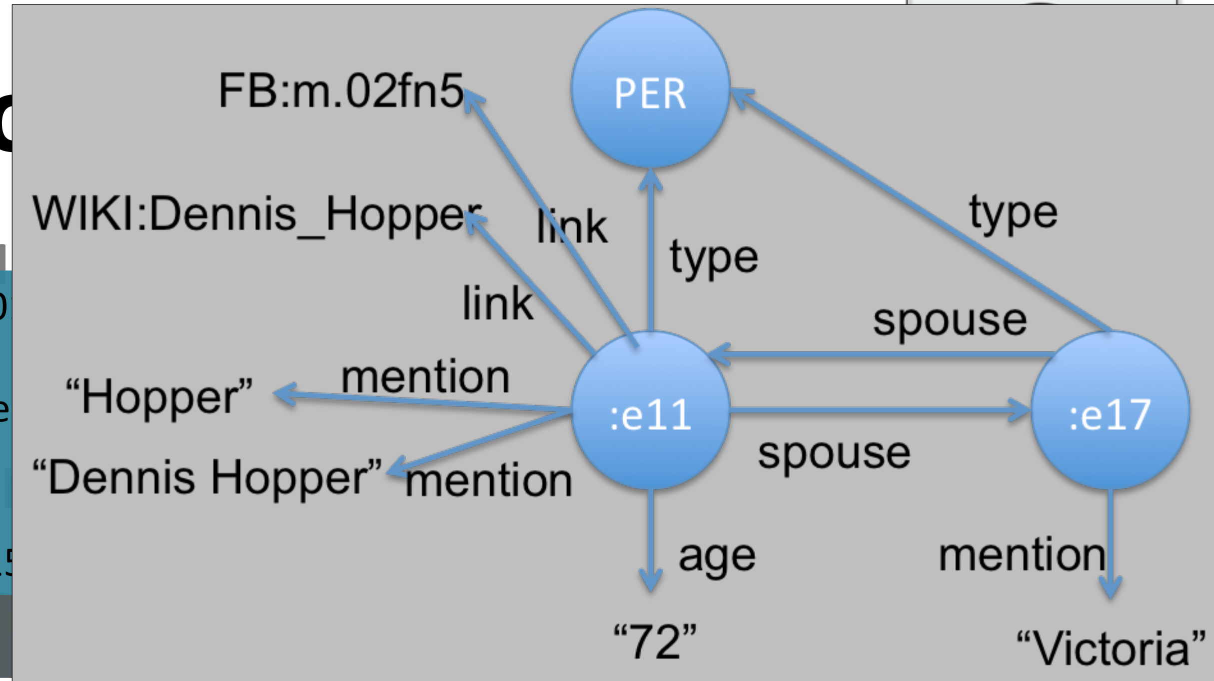
2016 TAC Cold Start KBP

- Read 90K documents: newswire articles & social media posts in English, Chinese and Spanish
- Find entity mentions, types and relations
- Cluster entities within and across documents and link to a reference KB when appropriate
- Remove errors (*Obama born in Illinois*), draw sound inferences (*Malia and Sasha sisters*)
- Create knowledge graph with provenance data for entities, mentions and relations

2016 TAC Co

- Read 90K docu
  m
- Fi
- Cl

<DOC id="APW_ENG_20
<HEADLINE>
Divorce attorney says De
</HEADLINE>
<DATELINE>
LOS ANGELES 2010-03-25
</DATELINE>
<TEXT>
<P>
Dennis Hopper's divorce attorney says in a court filing that the actor is dying and can't undergo chemotherapy as he battles prostate cancer.
</P>
<P>
Attorney Joseph Mannis described the "Easy Rider" star's condition in a declaration filed Wednesday in Los Angeles Superior Court.
</P>
<P>
Mannis and attorneys for Hopper's wife Victoria are fighting over when and whether to take the actor's deposition.
</P> ...




```
…
:e00211 type          PER
:e00211 link          FB:m.02fn5
:e00211 link          WIKI:Dennis_Hopper
:e00211 mention       "Dennis Hopper" APW_021:185-197
:e00211 mention       "Hopper"        APW_021:507-512
:e00211 mention       "Hopper"        APW_021:618-623
:e00211 mention       "丹尼斯·霍珀"    CMN_011:930-936
:e00211 per:spouse    :e00217         APW_021:521-528
:e00217 per:spouse    :e00211         APW_021:521-528
:e00211 per:age       "72"            APW_021:521-528
…
```

Information extraction from text

Identify relationships

ebqids:hasMeans

Link concepts to entities

http://dbpedia.org/resource/Buffer_overflow

ebqids:affectsProduct

CVE-2012-0150
Buffer overflow in msvcrt.dll in Microsoft Windows Vista SP2, Windows Server 2008 SP2, R2, and R2 SP1, and Windows 7 Gold and SP1 allows remote attackers to execute arbitrary code via a crafted media file, aka ”Msvcrt.dll Buffer Overflow Vulnerability.”

http://dbpedia.org/resource/Arbitrary_code_execution

http://dbpedia.org/resource/Windows_7

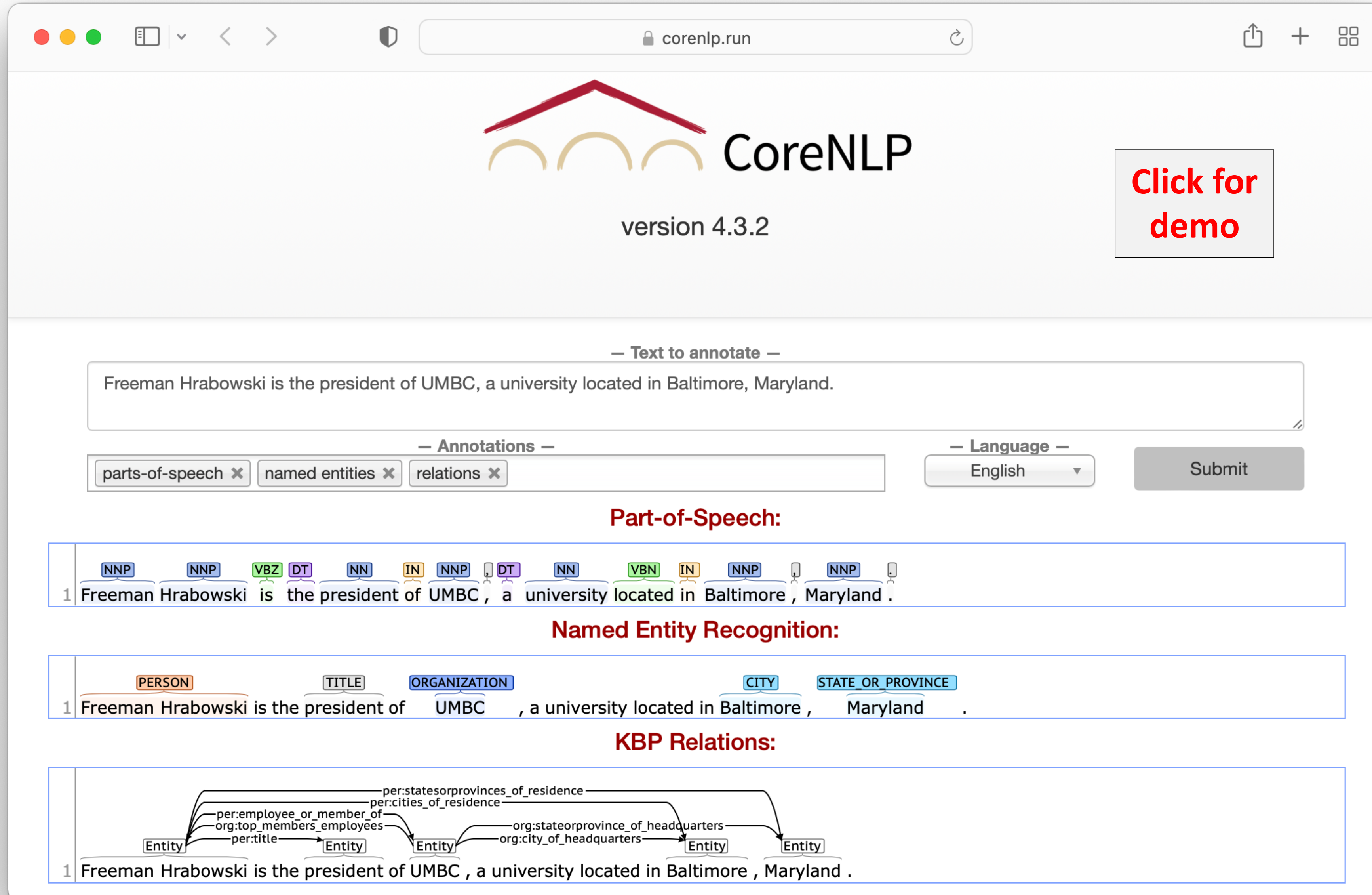
- Information extraction techniques identify entities, relations and concepts in security related text
- Map to terms in our ontology and DBpedia knowledge graph
- Also map them to terms in the Wikidata knowledge graph

http://ebiq.org/p/540

NLP Tools

- There is a rich and growing collection of open-source NLP tools
- Comprehensive pipelines:
  - Stanford CorNLP tools
  - Spacy
  - NLTK
- Word embeddings
  - Word2vec, BERT, Semsim

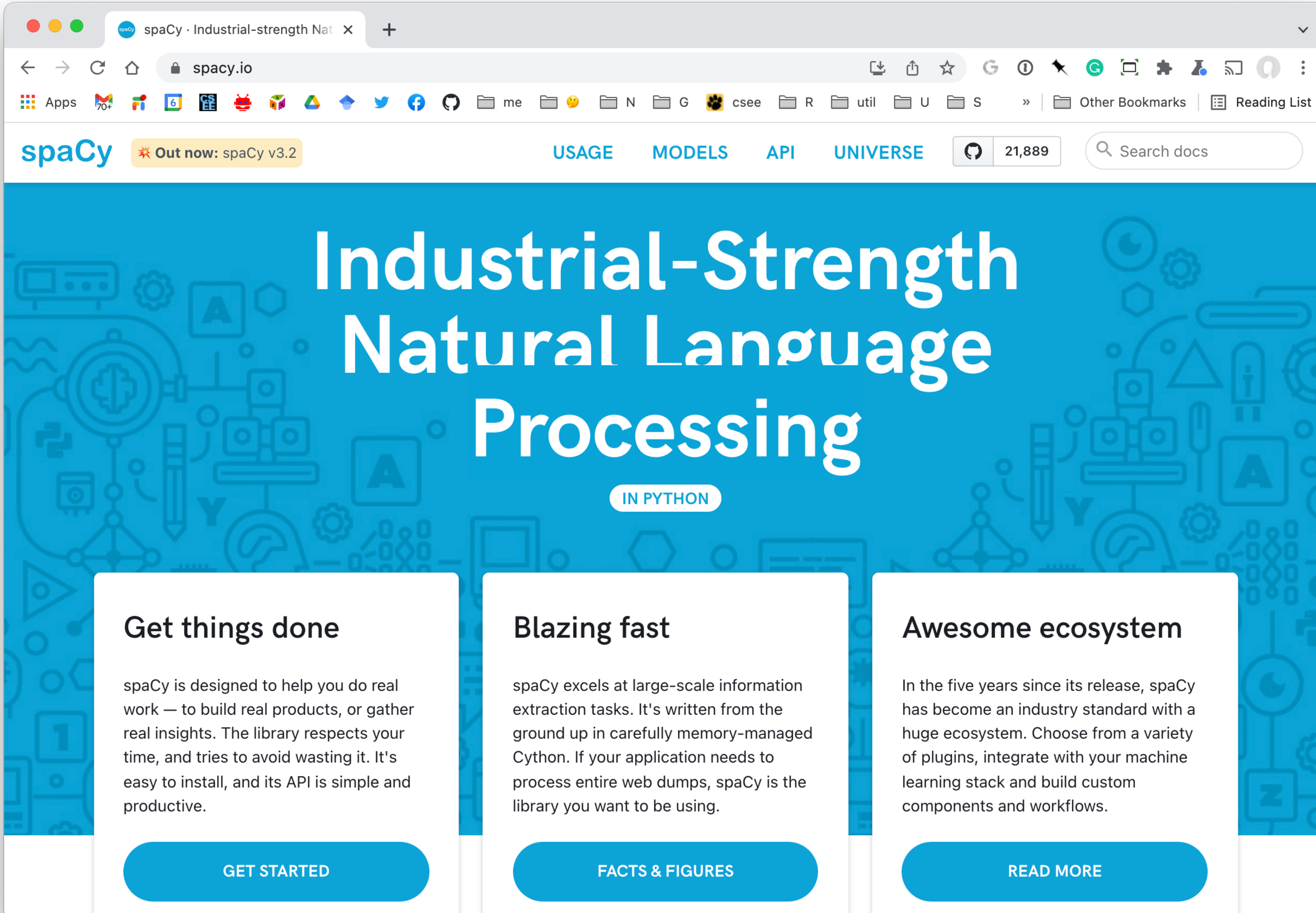
Stanford CoreNLP Tools



JSON/XML => KG triples


```
{ "text": "John Smith lives in Baltimore, Maryland.  He
is married to Mary Jones.  She works at Loyola
University where she is a professor.  The university is
in Baltimore.\n\n\n\n”,
  "docid": "text1.txt",
  "corefs": {
    "9": [
      {"endIndex": 6,
        "animacy": "INANIMATE",
        "text": "Baltimore",
        "isRepresentativeMention": true,
        "number": "SINGULAR",
        "startIndex": 5,
        "sentNum": 1,
        "gender": "NEUTRAL",
        "position": [1,  2q],
        "headIndex": 5,
        "type": "PROPER",
        "id": 1
      },
      { ....
```


```
##### :e_text1_1 LOCATION "Baltimore" #####

:e_text1_1    type    LOCATION
:e_text1_1    canonical_mention   "Baltimore"    text1:20-29
:e_text1_1    mention   "Baltimore"    text1:20-29
:e_text1_1    mention  "Baltimore"    text1:151-160

##### :e_text1_2 ORGANIZATION "Loyola University" #####

:e_text1_2    type    ORGANIZATION
:e_text1_2    canonical_mention   "Loyola University"    text1:85-102
:e_text1_2    mention "Loyola University"    text1:85-102
:e_text1_2    mention "The university"   text1:130-144

##### :e_text1_3 PERSON "John Smith" #####

:e_text1_3    type    PERSON
:e_text1_3    canonical_mention   "John Smith"   text1:0-10
:e_text1_3    mention "John Smith"    text1:0-10
:e_text1_3    mention "He"    text1:42-44
:e_text1_3    mention "She"    text1:72-75
:e_text1_3    mention "she"    text1:109-112
:e_text1_3    openie:lives_in    :e_text1_1    text1:0-3
:e_text1_3    per:spouse    :e_text1_5    text1:42-43
:e_text1_3    openie:is_married_to    :e_text1_5    text1:42-43
:e_text1_3    per:employee_of :e_text1_2    text1:72-74
```


spaCy · Industrial-strength Nat
spacy.io
Apps
me
N
G
csee
R
util
U
S
Other Bookmarks
Reading List
spaCy
Out now: spaCy v3.2
USAGE
MODELS
API
UNIVERSE
21,889
Search docs
Industrial-Strength Natural Language Processing
IN PYTHON
Get things done
spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive.
GET STARTED
Blazing fast
spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If your application needs to process entire web dumps, spaCy is the library you want to be using.
FACTS & FIGURES
Awesome ecosystem
In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins, integrate with your machine learning stack and build custom components and workflows.
READ MORE

Learning word meaning?

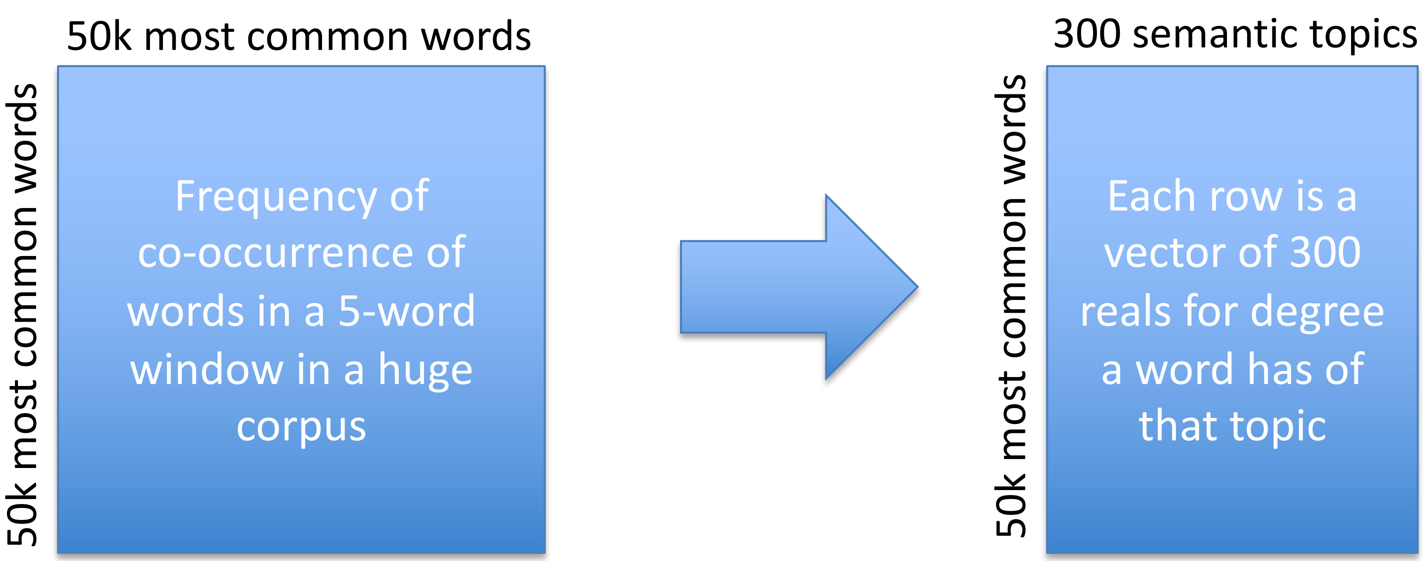
- How can we learn what a word means?
- The linguist John Rupert Firth famously write in 1957

  "You shall know a word by the company it keeps"

- A way to recognize that two words have similar meanings is to note that they occur in similar contexts
  - E.g., physician & doctor, nurse & doctor, love & hate

Word Embeddings

- Latent Semantic Analysis (LSA) learns a vector (e.g., 300 reals 0..1) for each unique word in a corpus to represent its meaning
  - LSA also used for document topic modelling
- An example of dimentionality reduction

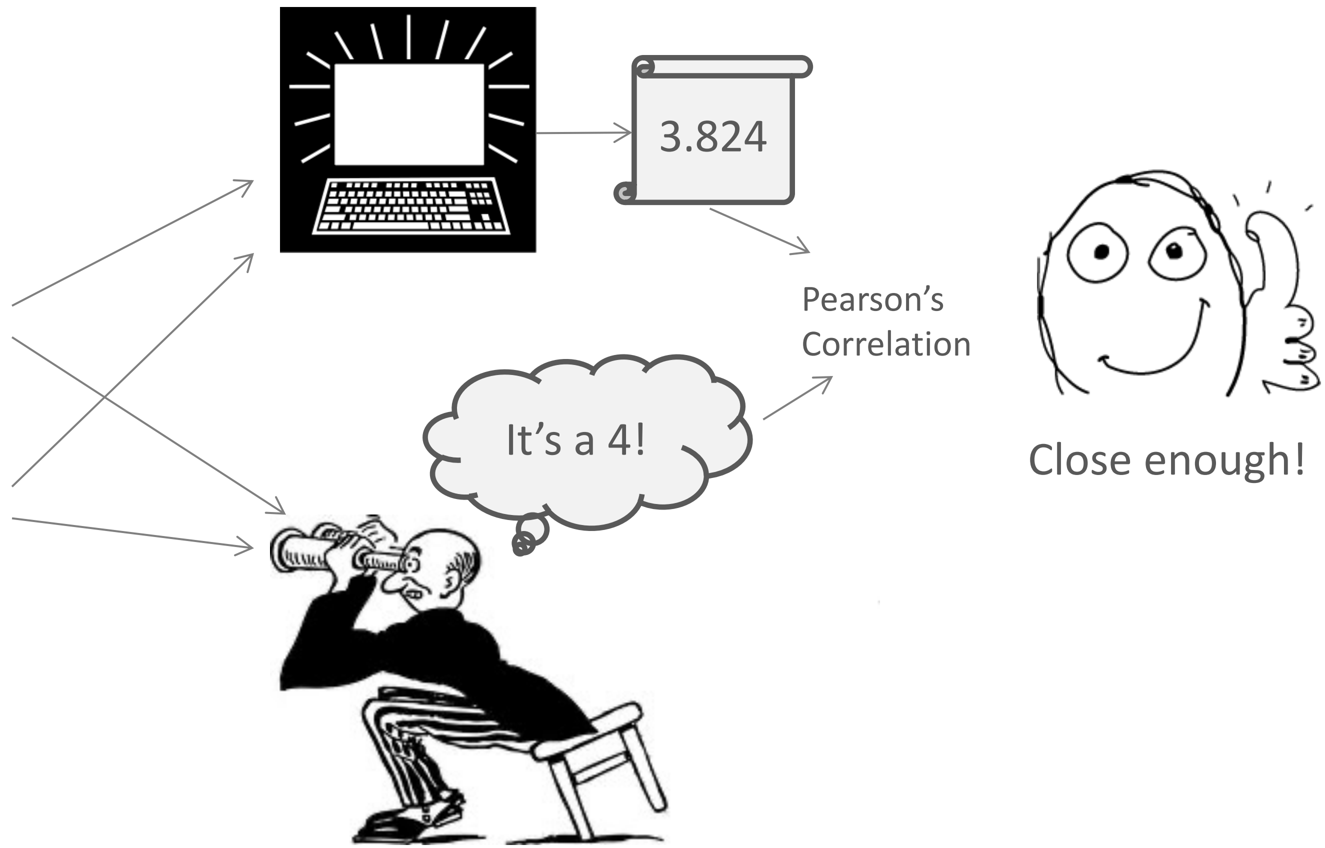


Sentence similarity

How similar are the two sentences semantically on a scale of 0-5?

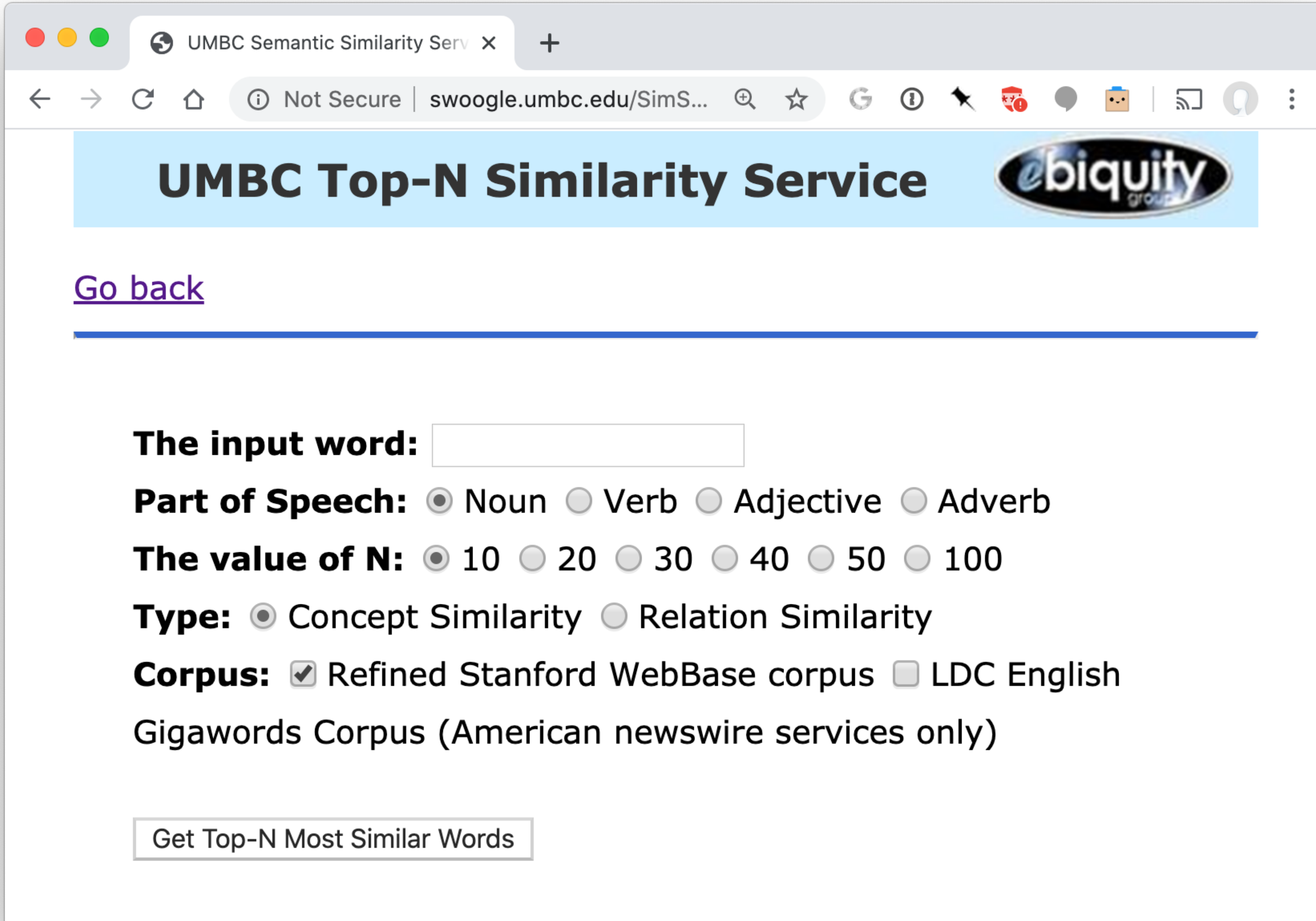
The mouse ate some cheese

Cheddar was eaten by a rat

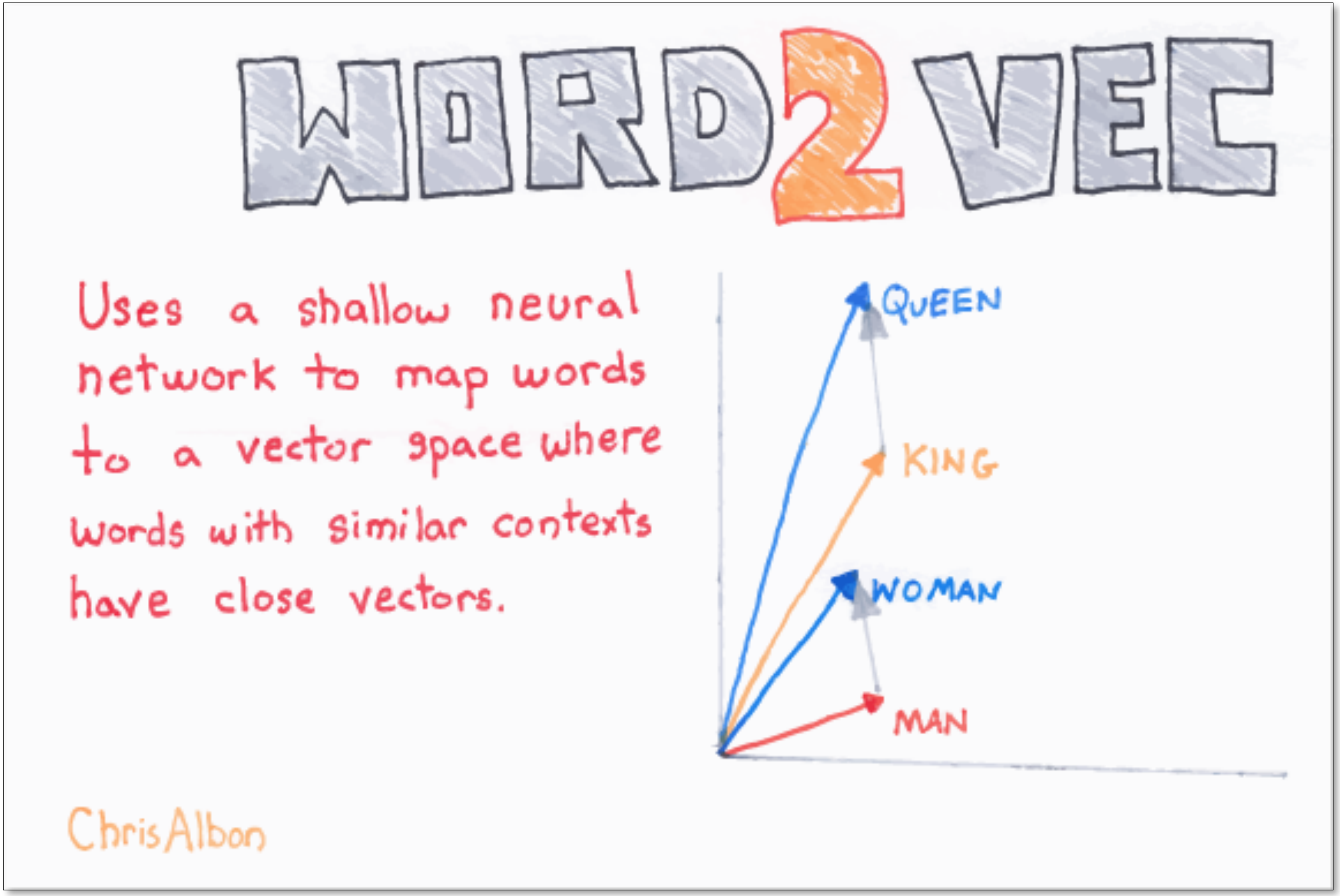


We used this approach in 2013 to win in a sentence similarity task

UMBC semantic similarity service



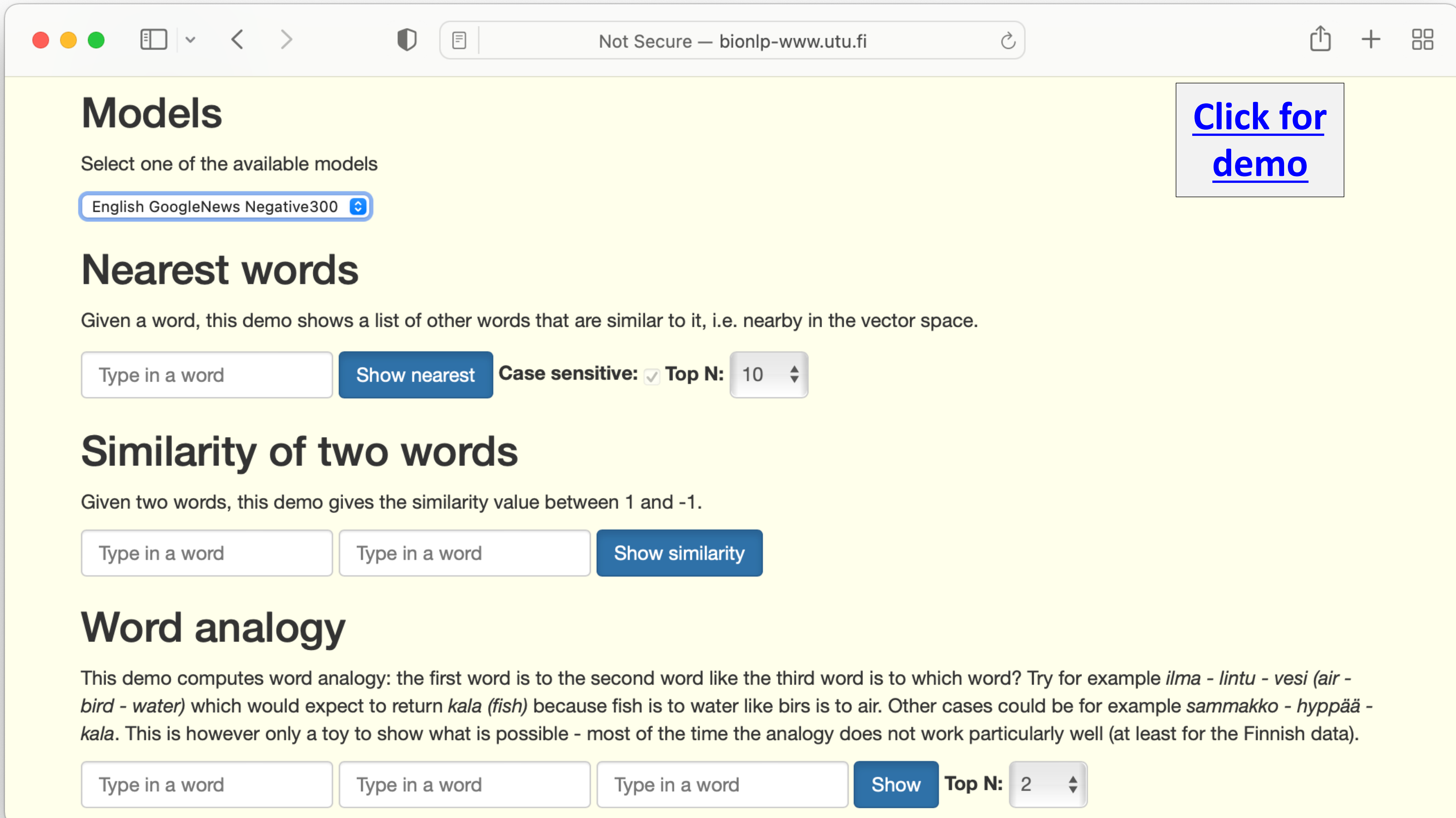
word2vec


WORD2VEC
Uses a shallow neural network to map words to a vector space where words with similar contexts have close vectors.
QUEEN
KING
WOMAN
MAN
Chris Albon

Word2Vec

- Developed by Google also in 2013 using a neural network approach
- Two models: CBOW and skip grams
- Trained on a much larger corpus from the Web
- Models can be downloaded and are still used today
  - E.g., the spaCy NLP system includes word2vec to measure similarity

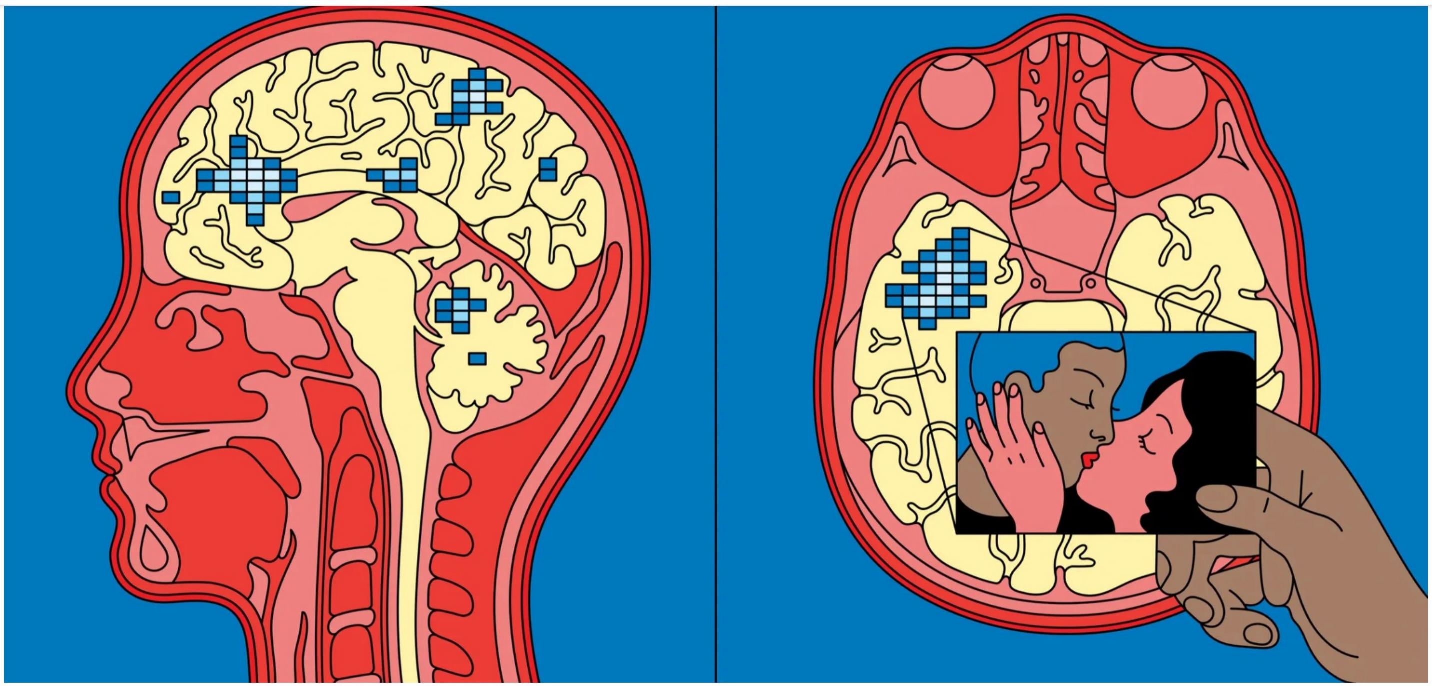
Word2vec demo



[Click for demo]

Scientists using fMRI to measure brain activity find locations associated with smilar concepts – brain embeddings!

THE NEW YORKER



It isn't so much that brain scans have improved—it's that we've got better at reading them. Illustration by Laura Edelbacher

ANNALS OF TECHNOLOGY DECEMBER 6, 2021 ISSUE

THE SCIENCE OF MIND READING

Researchers are pursuing age-old questions about the nature of thoughts—and learning how to read them.

By James Somers

November 29, 2021

Click to read

Conclusion

- KGs help in extracting information from text
- The information extracted can update the KGs
- The KGs provide support for new tasks, such as question answering, speech interfaces and produce data useful in applications, like IDSs
- There use will grow and evolve in the future
- New machine learning frameworks will result in better accuracy