Learning the meaning of words from text

UNIVERSITY OF GOTHENBURG
DEPT OF SWEDISH

Språk-
BANKEN

Richard Johansson

richard.johansson@gu.se

FSCONS
November 1, 2014

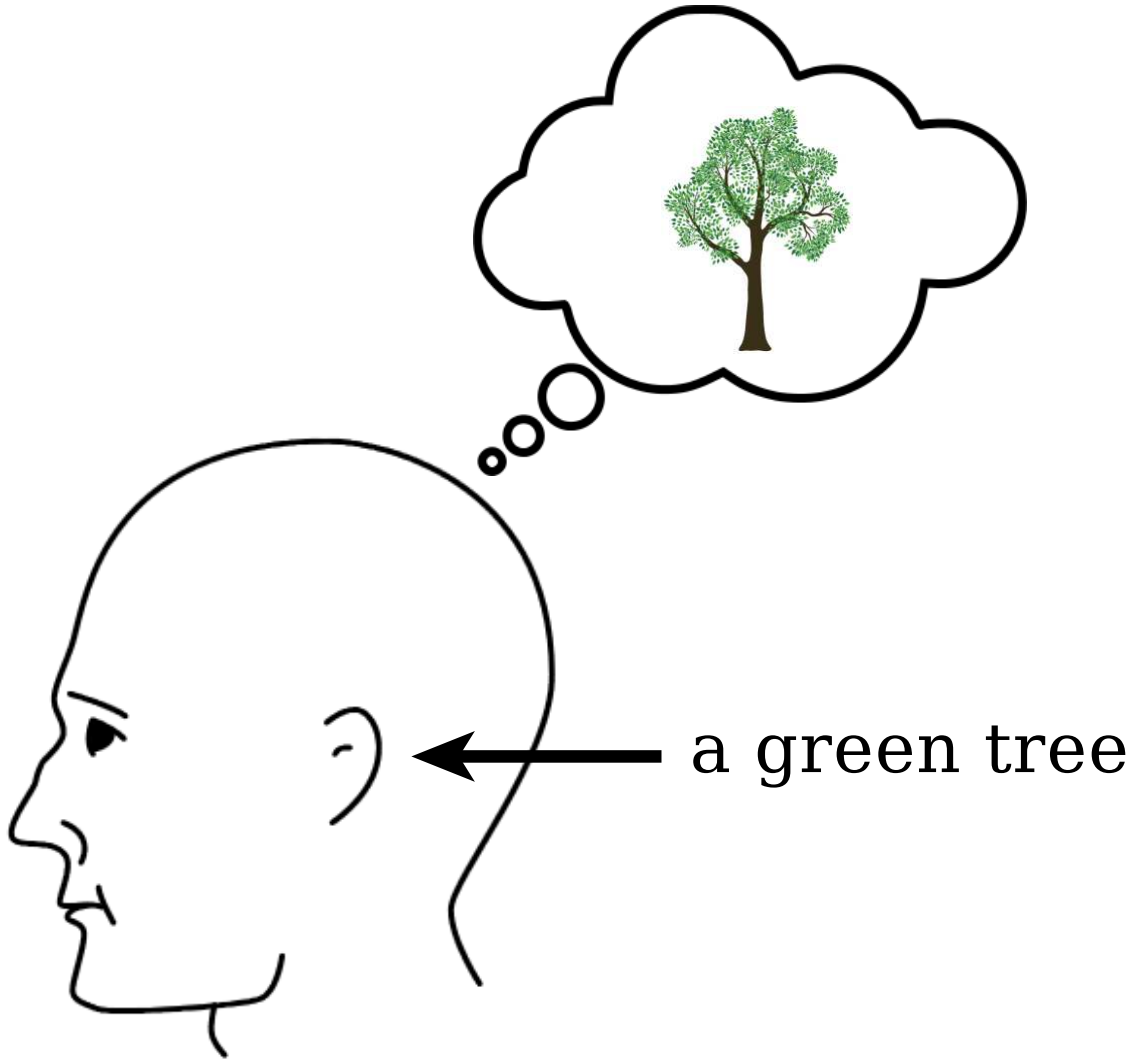
this talk

- what is the “meaning” of a word?
- how can a computer have a notion of word meaning?
- discovering word meaning automatically
- some **free** software that you can try out at home

about me

- I work at the department of Swedish Language here at the University of Gothenburg
- I teach and do a bit of research in natural language processing
- I build computer programs that extract “meaning” from text
- if you want to study at our Master’s program in natural language processing, talk to me afterwards or google *MLT GU*
  - application deadline is January 15 for international students and April 15 for EU/EES students

understanding text



mapping strings to meaning

- if we want to make a computer “understand”, then how do we tell it what is the **meaning** of the string `"pizza"`?
- we need to connect the string to some “meaning object”
- but what is that object?

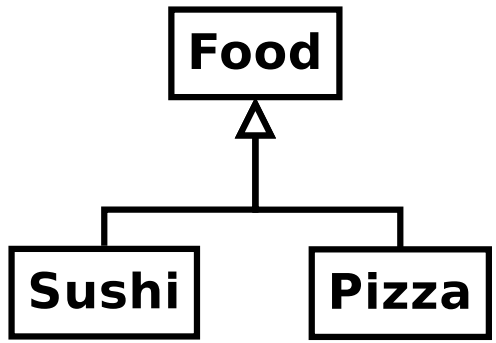
what does the dictionary say?

pizza /ˈpiːtsə/

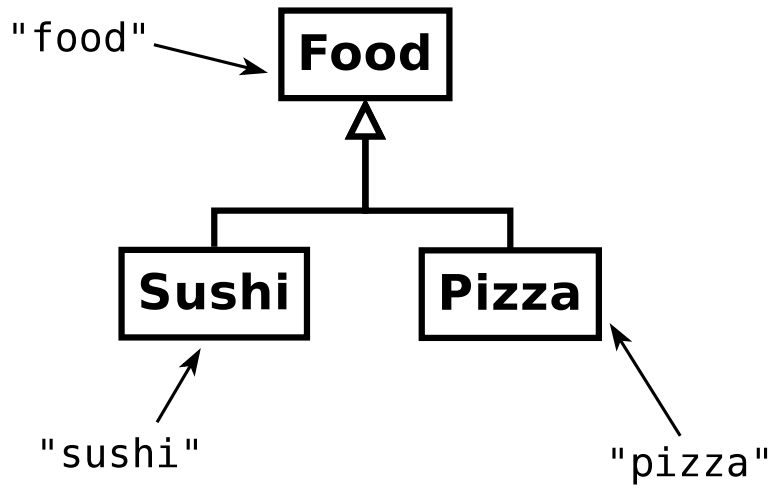
N

a dish of Italian origin consisting of a baked disc of dough covered with cheese and tomatoes, usually with the addition of mushrooms, anchovies, sausage, or ham

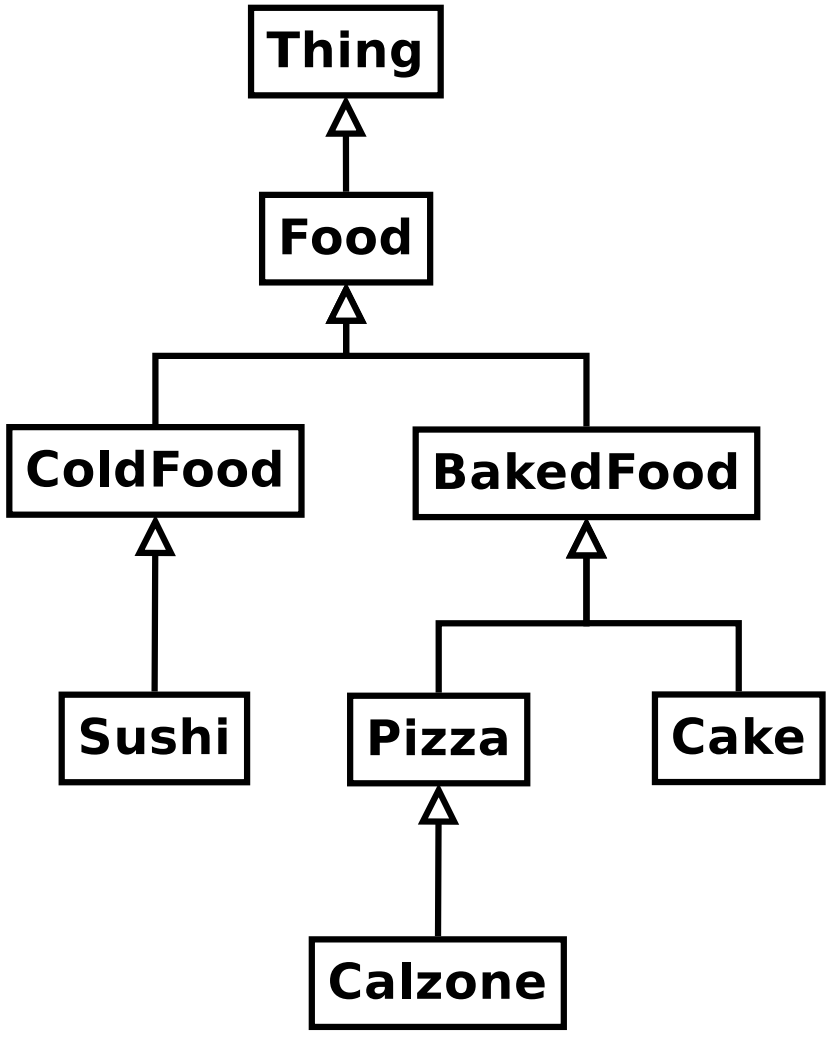
the computer programmer solution



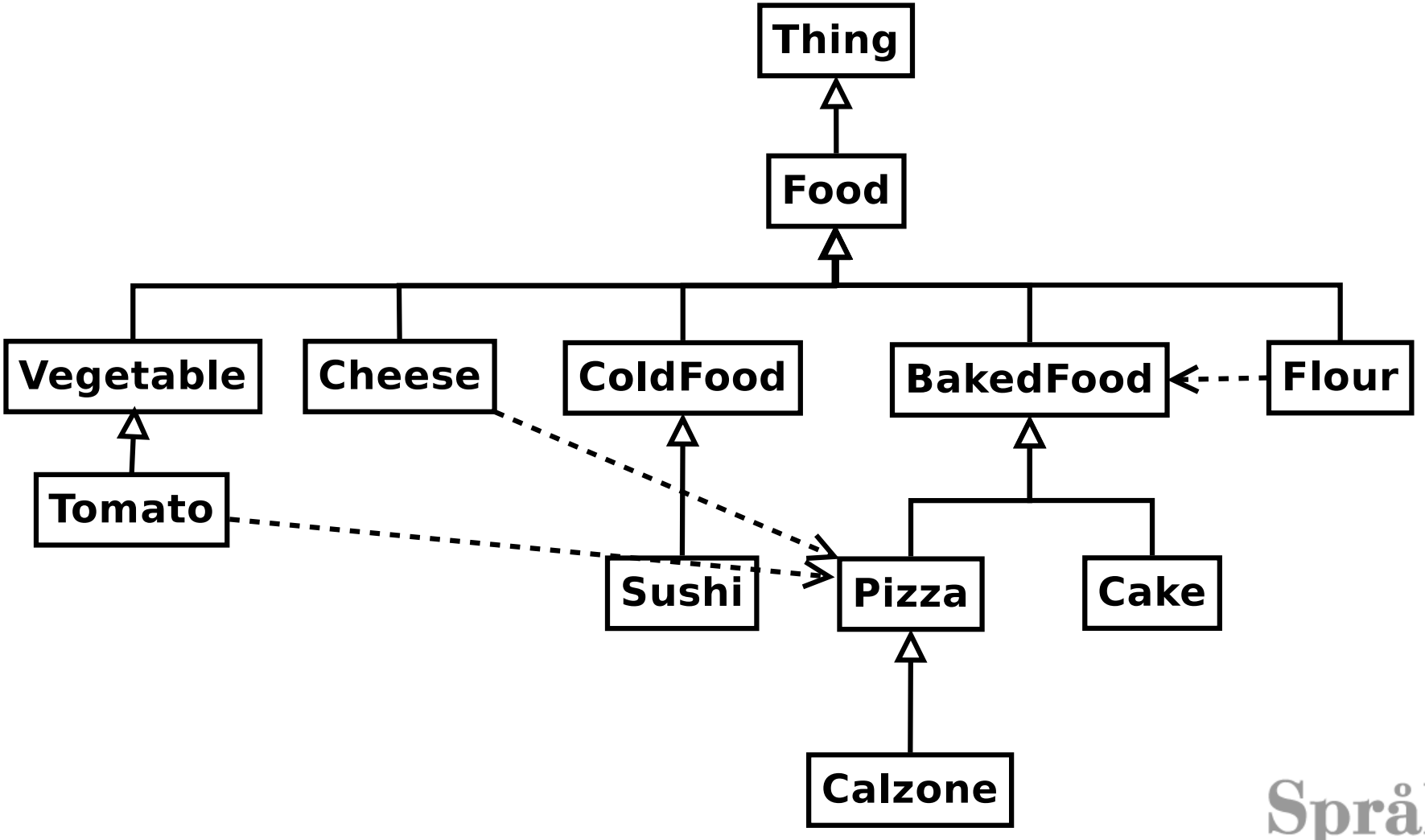
the computer programmer solution



the computer programmer solution



the computer programmer solution



using knowledge libraries or ontologies

- http://wordnet.princeton.edu/
- http://dbpedia.org/describe/?url=http%3A%2F%2Fdbpedia.org%2Fresource%2FGothenburg&sid=16895
- https://gate.d5.mpi-inf.mpg.de/webyagospotlx/Browser?entity=%3CGothenburg%3E

ontologies are limited

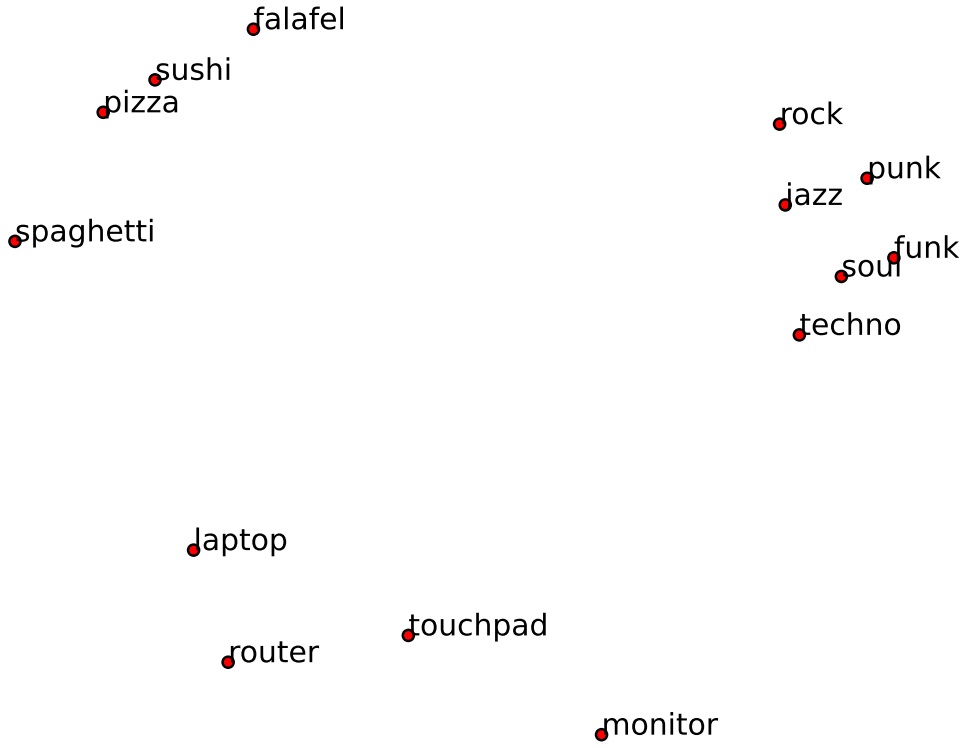
- these resources are typically full of holes
- they require a massive investment by trained experts
  - how much did it cost to make WordNet and YAGO?
  - in DBPedia and YAGO, some effort have been saved by using semi-automatic methods
    - by parsing Wikipedia infoboxes, merging incompatible resources, ...
- new words appear all the time

discovering "meaning" automatically

- there's a growing interest in methods that pick up some sort of word meaning simply by observing raw text
- these methods require large amounts of text but little or no investment in "knowledge engineering"
  - you can go home after this talk and try out the software I'll mention, while building an ontology would take you years
- text is cheap nowadays

word spaces

- in a **word space**, a word is connected to a **vector**: basically a point in a coordinate system, or an array of numbers



- the spaces typically have 50–10,000 dimensions, but we show 2D here for practical reasons

distances/similarities in a word space

- in a word space, "similarity" of words corresponds to geometry
  - being near each other in the space
  - . . . or pointing in a similar direction
- *pizza* is kind of like *sushi*, but not so much like *touchpad*
- on the other hand, it seems that we have lost the knowledge structure: we don't know *how* pizza and sushi are similar
  - we'll come back to this question later

what's the point of doing something like that?

- search engines:
  - if I google for *pizza in Gothenburg*, I'm probably interested in eating, so it might be better to rank documents mentioning *calzone* higher than those mentioning *router*
- natural language processing in general:
  - *Spotfire* is the name of a company
  - *Talkamatic* is similar to *Spotfire*
  - . . . so maybe *Talkamatic* is also a company?

how could this work?

- *"you shall know a word by the company it keeps"*
- two words probably mean about the same thing if they
  - . . . appear in the same documents?
  - . . . tend to have the same words around them?
  - . . . are illustrated with similar images?

example: most frequent verbs near *cake* and *pizza*

- what are the activities we do with *cakes* and *pizzas*?
  - *cake*: eat, bake, throw, cut, buy, get, decorate, garnish, make, serve, order
  - *pizza*: eat, bake, order, munch, buy, serve, garnish, name, get, make, heat

- each of the verbs could correspond to a dimension in the word space

example: *tårta* and *pizza* in Swedish text

Tårta verb			**Verb** **tårta**		
1.	vara	83	1.	äta	41
2.	baka	4	2.	baka	29
3.	skära upp	3	3.	kasta	31
4.	leverera	4	4.	kasta[2]	31
5.	skära	4	5.	skära upp	10
6.	se ut	6	6.	skära	13
7.	sticka av	2	7.	köpa	18
8.	äta	4	8.	få	50
9.	se	7	9.	dekorera	5
10.	sälja	4	10.	garnera	5
11.	bli	17	11.	göra	20
12.	kosta	4	12.	servera	6
13.	skära[2]	2	13.	beställa[2]	5
14.	vara[2]	8	14.	beställa	5
15.	höra hemma	2	15.	inhandla	3

Pizza verb			**Verb** **pizza**		
1.	grädda	14	1.	äta	63
2.	baka	12	2.	baka	33
3.	kosta	17	3.	beställa[2]	28
4.	smaka	5	4.	beställa	28
5.	smaka[2]	5	5.	käka	20
6.	innehålla	6	6.	köpa	32
7.	äta	5	7.	servera	12
8.	vara	68	8.	grädda	7
9.	stoppa	4	9.	garnera	6
10.	beställa	3	10.	döpa	6
11.	köpa	4	11.	hämta	9
12.	visa sig	3	12.	toppa	7
13.	vara[2]	6	13.	göra	22
14.	ligga	6	14.	värma	5
15.	se ut	4	15.	smälla i sig	2

another idea: guessing the missing word

"after a few years abroad, he moved back to ___"
"the furniture was imported from ___"
"he visited the libraries in London, ___, Florence and Venice"
"during the German siege of ___ in 1870, he was found dead"

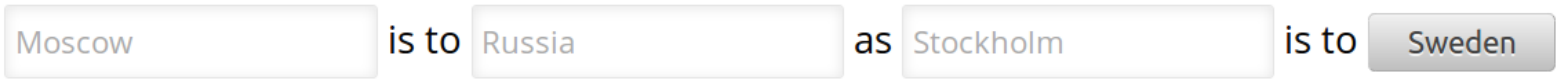
- to make a long story short, we can make a statistical model that tries to guess the missing word
- as a by-product of this statistical model, a word space will be produced
- for instance, see the recent research by Tomáš Mikolov

there's more than just similarity

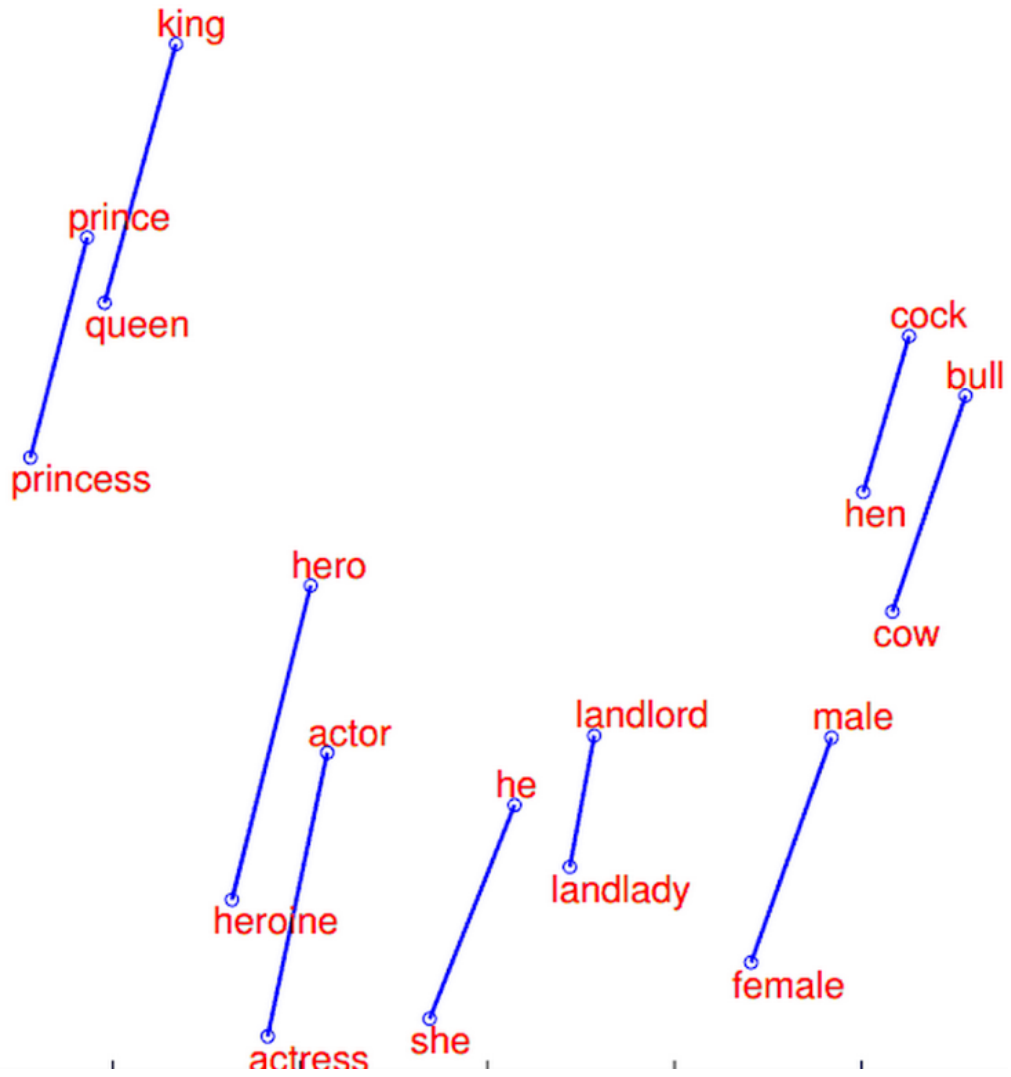
- we said previously that word spaces don't have the structural information that ontologies have
- however, just recently it was discovered that they actually pick up *some* structure implicitly
- we can use word spaces to answer analogy questions like "*Moscow* is to *Russia* as *Stockholm* is to *X*"

example

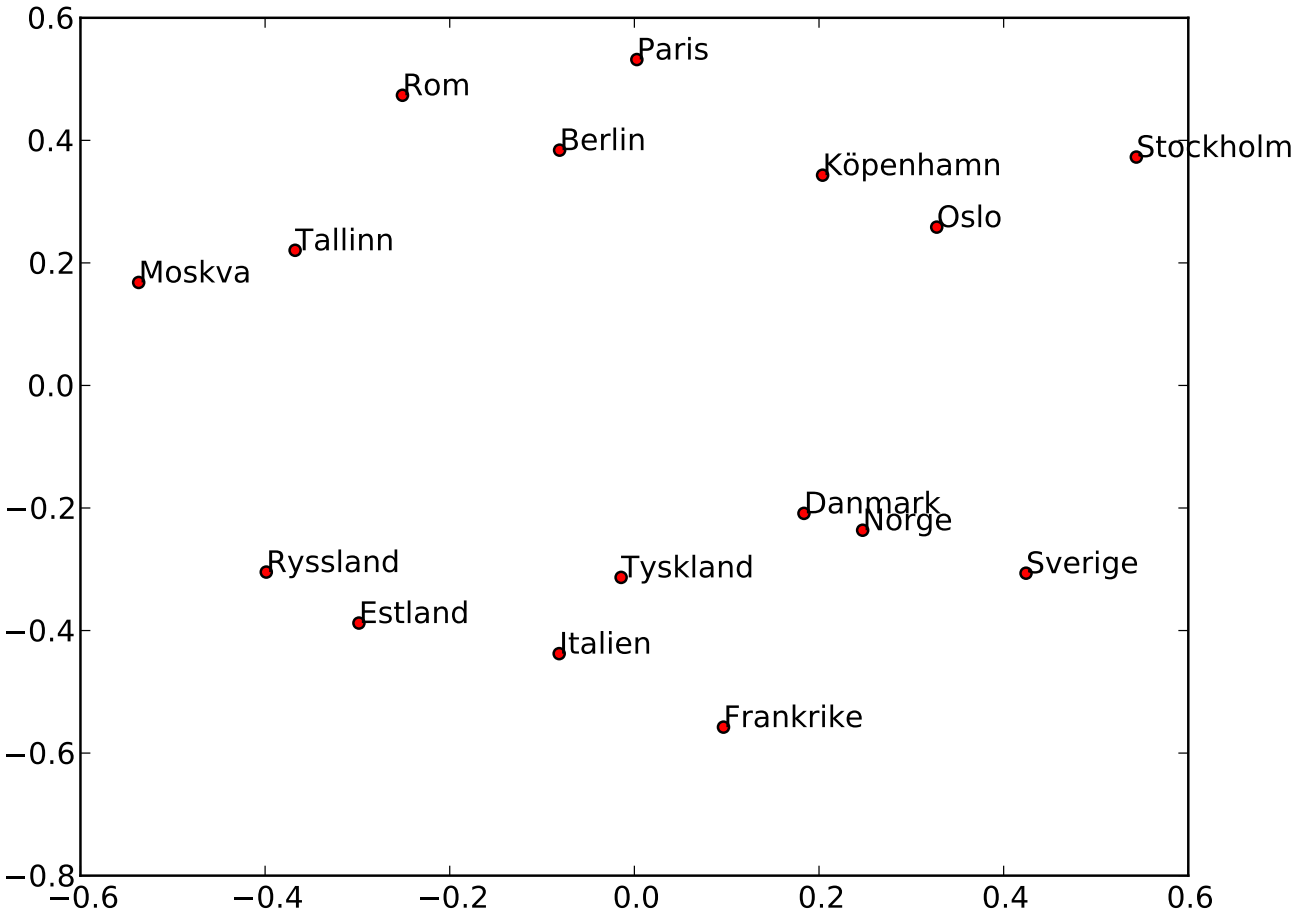
- `http://radimrehurek.com/2014/02/word2vec-tutorial#app`



gender in the word space (example by Mikolov)



countries and cities



some software you can try at home (a sample)

- `word2vec`: the software by Mikolov
  - includes a word space built by Google using a huge collection of news text
- `gensim`: a nice Python library by Řehůřek
  - includes a reimplementation of `word2vec` but also several other useful algorithms
- for processing vectors in general: `numpy` (Python), Breeze (Scala), JBLAS (Java), BLAS/ATLAS (C/Fortran), ...
- to draw the images: `scikit-learn` (projecting to 2D) and `matplotlib`

some data you can use (also a sample)

- Wikipedia is a nice way to get text data in many languages
  - however, large differences in size
  - needs preprocessing: removing boilerplate and wiki markup
- English, French, German, Italian:
  - "web as a corpus":
    `http://wacky.sslmit.unibo.it/doku.php?id=corpora`
  - for English, see also the `word2vec` page for pointers to some English text collections
- Swedish: *Språkbanken* at my department collects large volumes of text of many types
  - `http://spraakbanken.gu.se/eng/node/1587`
  - these collections are preprocessed, so you just need to strip away the XML to use them in `word2vec` or *gensim*

using gensim: code example

- http://radimrehurek.com/2014/02/word2vec-tutorial

```
>>> sentences = ...
>>> model = gensim.models.Word2Vec(sentences, min_count=5, size=200)
>>> model.most_similar(positive=['woman', 'king'], negative=['man'], topn=1)
[('queen', 0.50882536)]
```

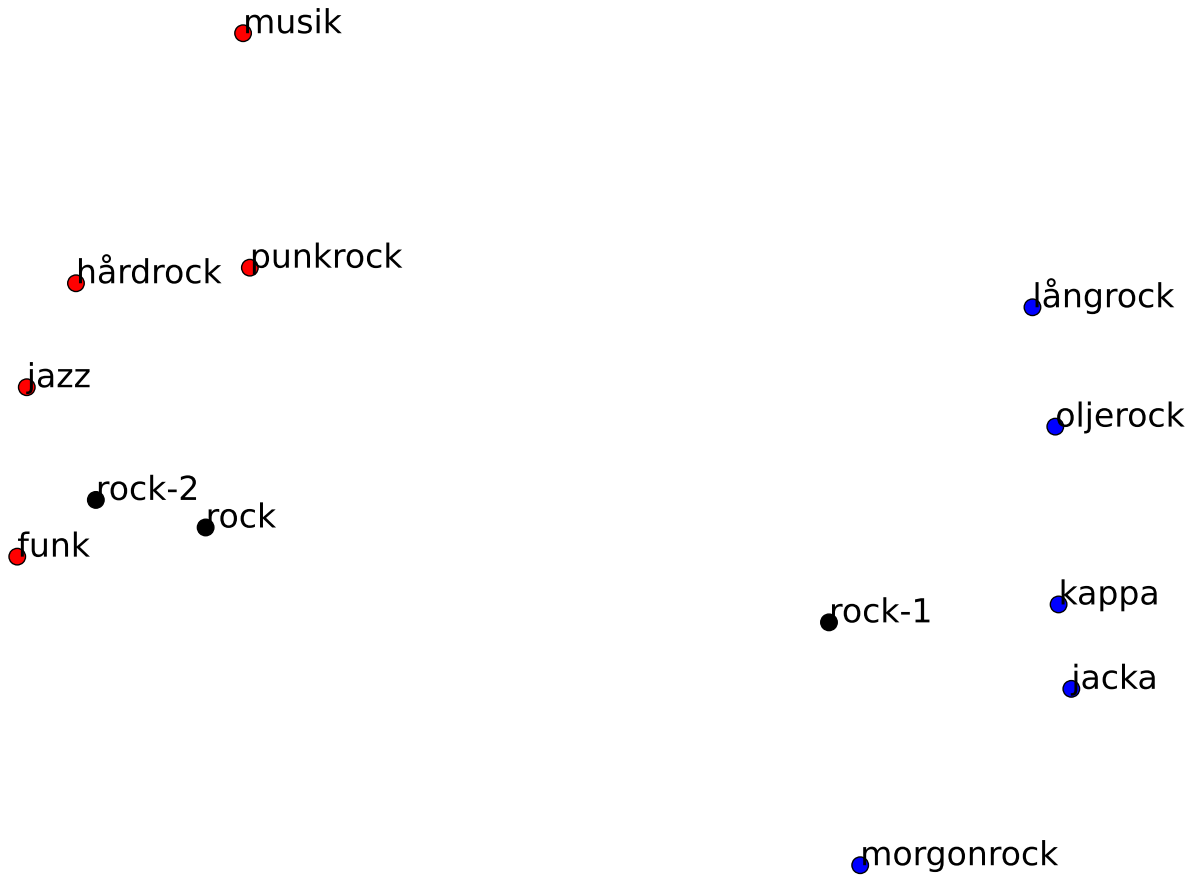
- building the word space typically takes from a few hours to some days
  - depending on the amount of text, the number of CPU cores, which algorithm you use, etc

some open research problems (a small sample)

- how can we build word spaces for languages with complex inflection systems
  - for instance, Turkish:
    çöp+lük+ler+imiz+de+kir+ler+den+mi+y+di?
    'was it from those that were in our garbage cans?'

    example from Sproat, *Morphology and Computation*
- can we connect word spaces with images or output from sensors?
- can we build multilingual word spaces?
  - useful for machine translation, for instance
- what can we do about words with more than one meaning?

example: the word *rock* (Swedish)



the end

- please talk to me or contact me by email (`richard.johansson@gu.se`) if you have questions