NIHR | Maudsley Biomedical Research Centre

Modelling language: documents

Angus Roberts, Senior Lecturer in Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King's College London

Representing language

- How can we represent meaning in language?
- Symbolic, rationalist approaches
- Empirical approaches
  - Bag-of-words
  - TF-IDF

Symbolic, rationalist approaches

Symbolic NLP

- If we want to manipulate language computationally, and process it, we need to represent it in some way
- In rule-based, symbolic NLP, we can consider language to be represented as strings of characters
  - A string of characters has no inherent meaning
  - We match these strings with expressions (rules, grammars) that define some pattern of characters
  - These rules give meaning to our strings
  - The approach can be extended to POS and other features
- The thinking is that language could be reasoned about in some logical way, and that the structures of language could be rationalised in to sets of rules
- Prolog, a logic programming language, was the tool of choice for this

Prolog grammars

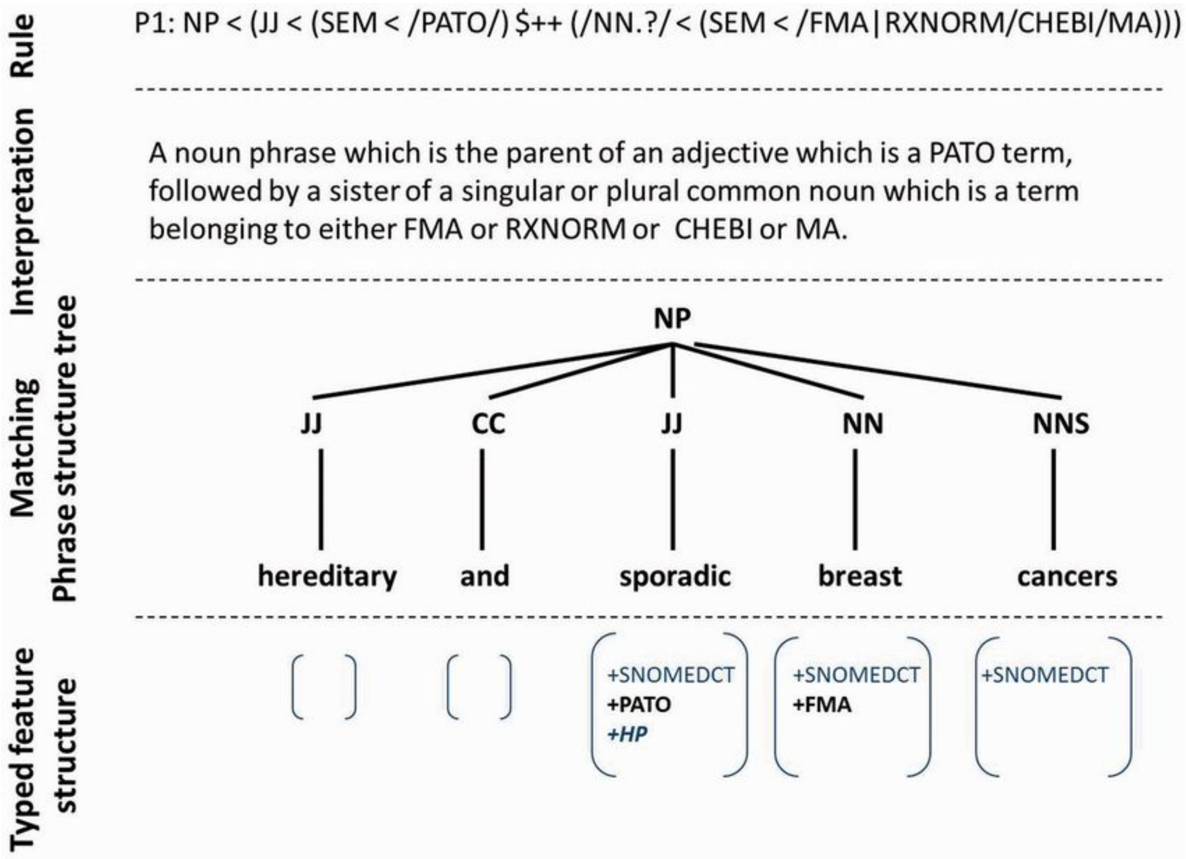
```
/* DOG_GRAMMAR.PL */

s --> np, vp.

np --> n.          np --> adj, n.     np --> adj, adj, n.
np --> det, n.      np --> det, adj, n.     np --> det, adj, adj, n.

vp --> v, np.     vp --> v, pp.

pp --> p, np.

det --> [the].     det --> [a].     det --> [an].

n --> [dogs].     n --> [fox].     n --> [jumps].

adj --> [quick].     adj --> [brown].     adj --> [lazy].

v --> [jumps].     v --> [runs].

p --> [over].     p --> [onto].
p --> [in].         p --> [under].
```

Credit: John Coleman, http://www.phon.ox.ac.uk/coleman

Grammars

- How many rules does it take to represent “grammatical” English?
- Do we have the same problem for all languages?
- What about medical language?
- Can we write grammars that capture not the syntax, but the semantics of language? i.e. the real-world categories that words relate to, and the relationships between them?
  - e.g. diseases, medications, anatomy?

Semantic grammars



Collier et al (2015). PhenoMiner: From text to a database of phenotypes associated with OMIM diseases. Database. 2015. bav104. 10.1093/database/bav104.

NIHR | Maudsley Biomedical Research Centre

Rationalism vs Empiricism

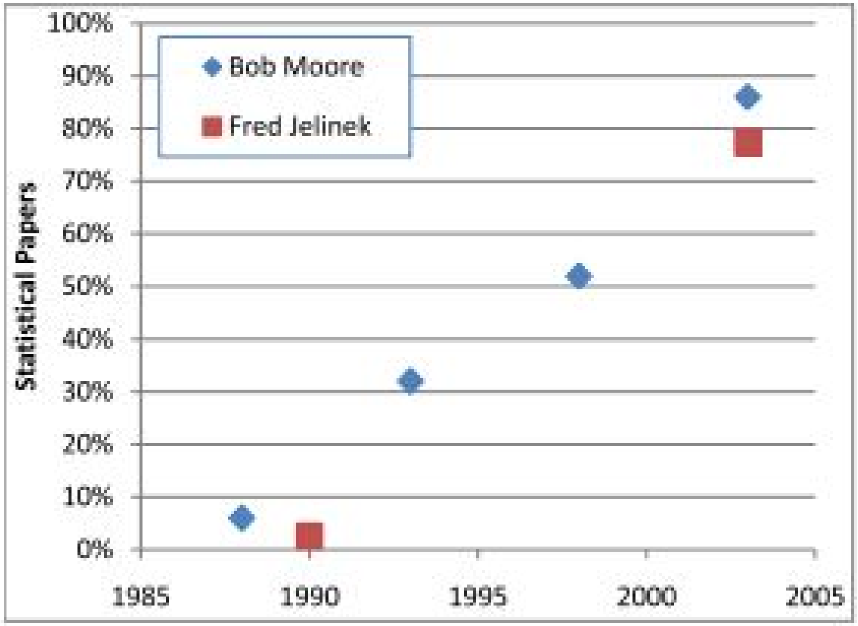


FIGURE 1 The shift from Rationalism to Empiricism is striking (and no longer controversial). This plot is based on two independent surveys of ACL meetings by Bob Moore and Fred Jelinek (personal communication).

- Church, LiLT Volume 2, Issue 4 May 2007

Empiricism

- Empiricism is at the heart of current NLP
  - How often do words appear?
  - How many of a particular word appear in a document?
  - Statistical models of language
- Is it upsetting that counting words in a document outperforms reasoning about the language?

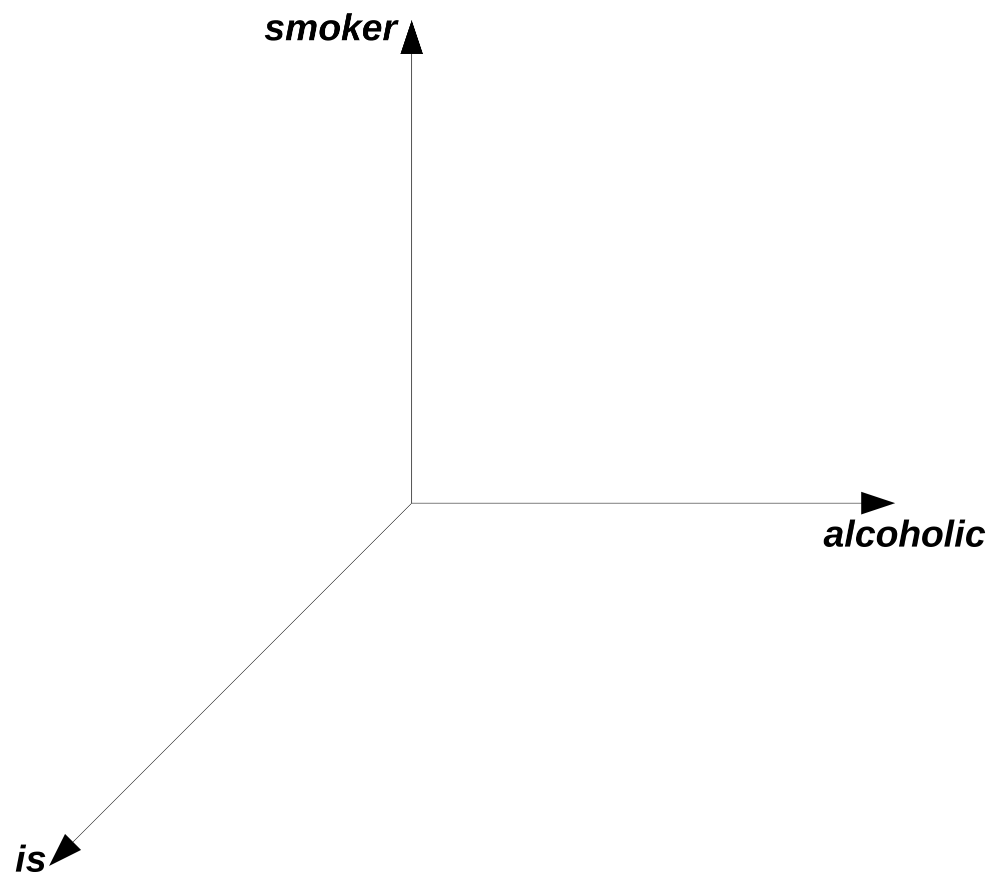
NIHR | Maudsley Biomedical Research Centre

Empirical representations: bag-of-words

Bag of words

- Let's plot four sentences:
  - he is a smoker
  - she is alcoholic
  - he is anxious
  - he is diabetic and diet controlled
- Along 3 dimensions:
  - smoker
  - alcoholic
  - is

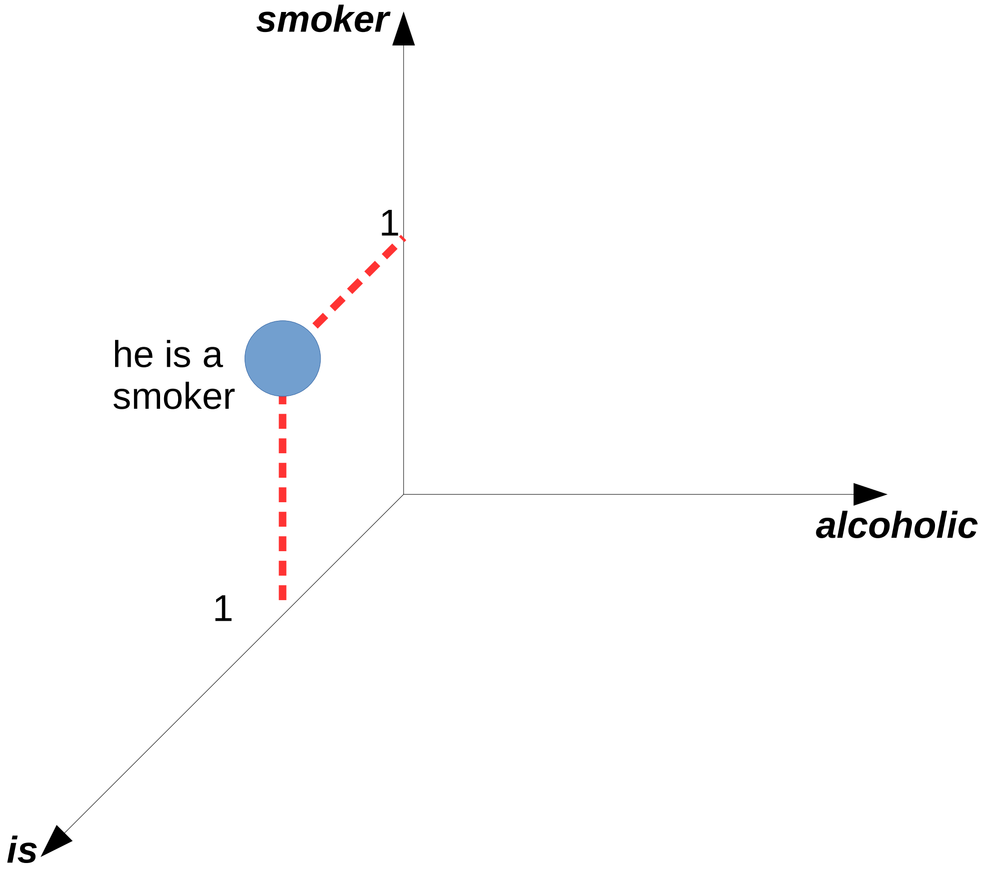
Bag of words



(After: Feature Engineering for Machine Learning
by Amanda Casari, Alice Zheng. O'Reilley)

Bag of words

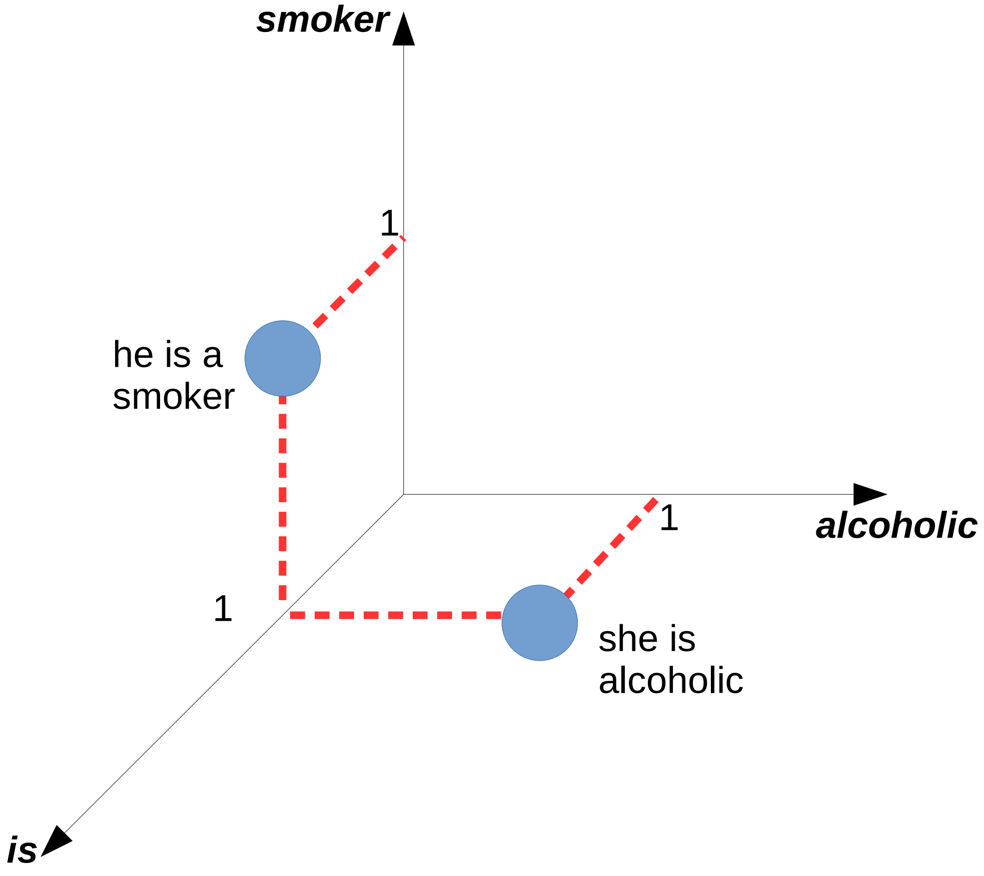
- he is a smoker



(After: Feature Engineering for Machine Learning
by Amanda Casari, Alice Zheng. O'Reilley)

Bag of words

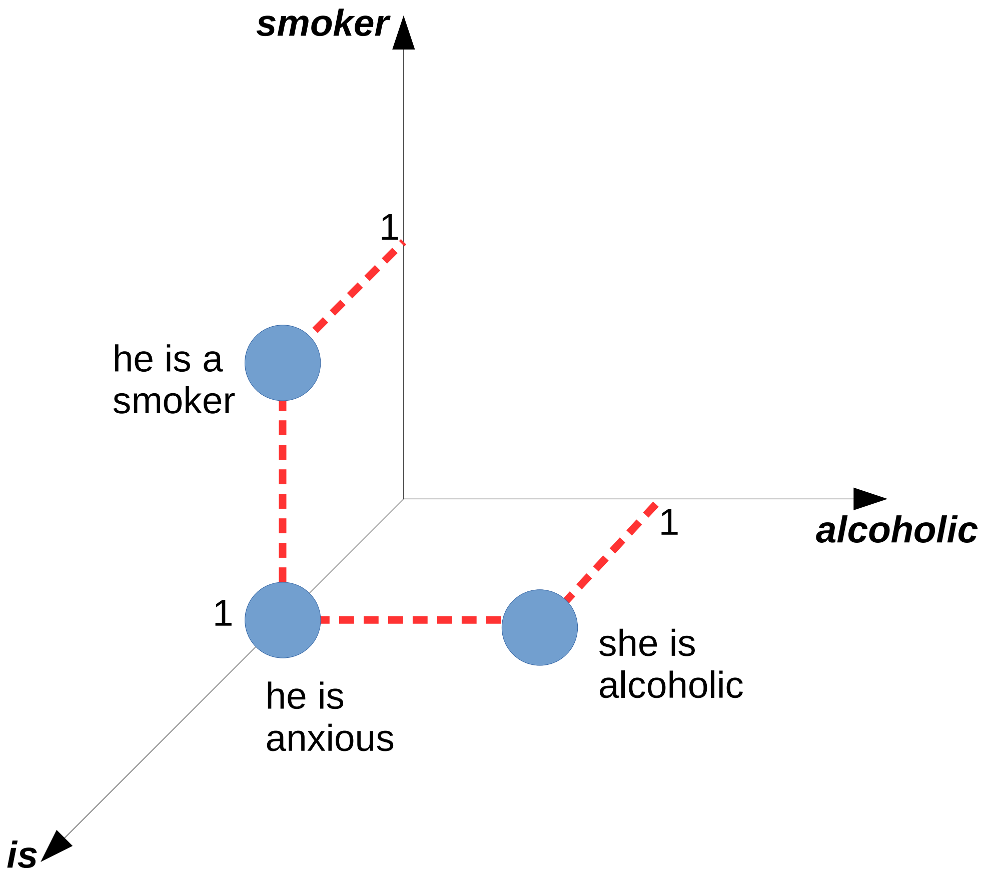
- he is a smoker
- she is alcoholic



(After: Feature Engineering for Machine Learning by Amanda Casari, Alice Zheng. O'Reilley)

Bag of words

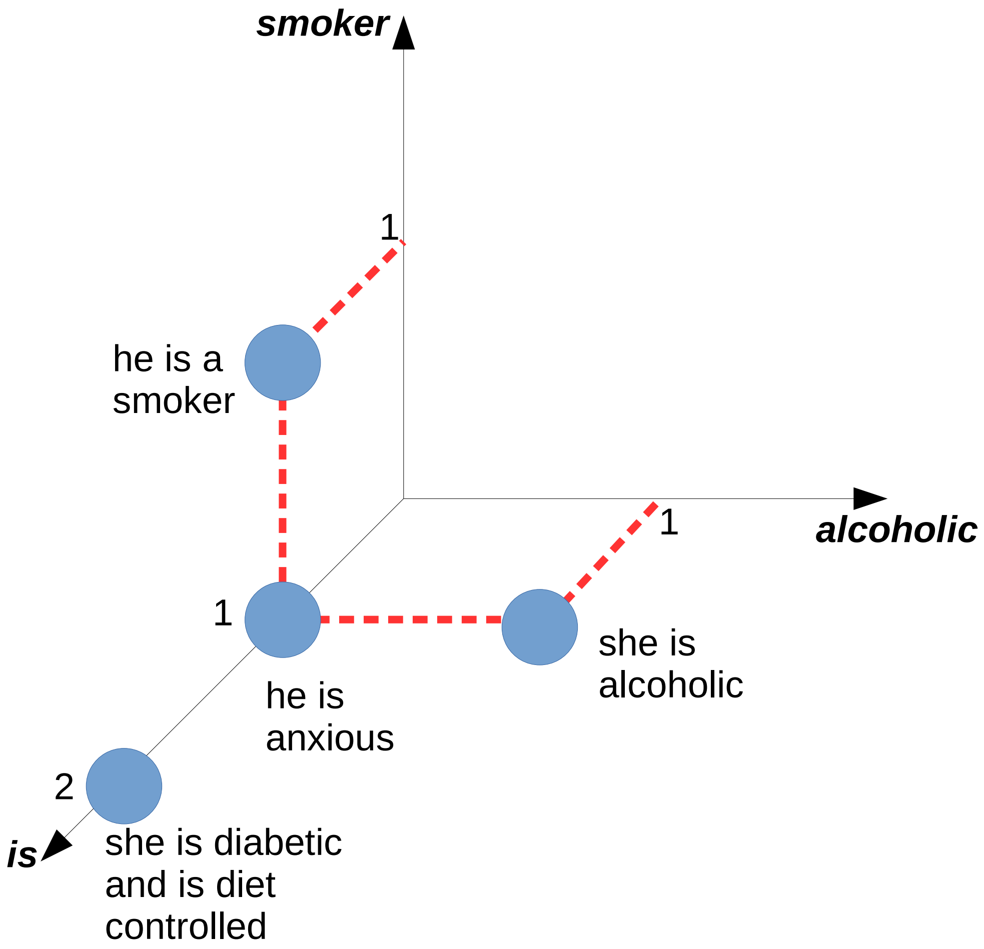
- he is a smoker
- she is alcoholic
- he is anxious



(After: Feature Engineering for Machine Learning by Amanda Casari, Alice Zheng. O'Reilley)

Bag of words

- he is a smoker
- she is alcoholic
- he is anxious
- she is diabetic and is diet controlled



(After: Feature Engineering for Machine Learning by Amanda Casari, Alice Zheng. O'Reilley)

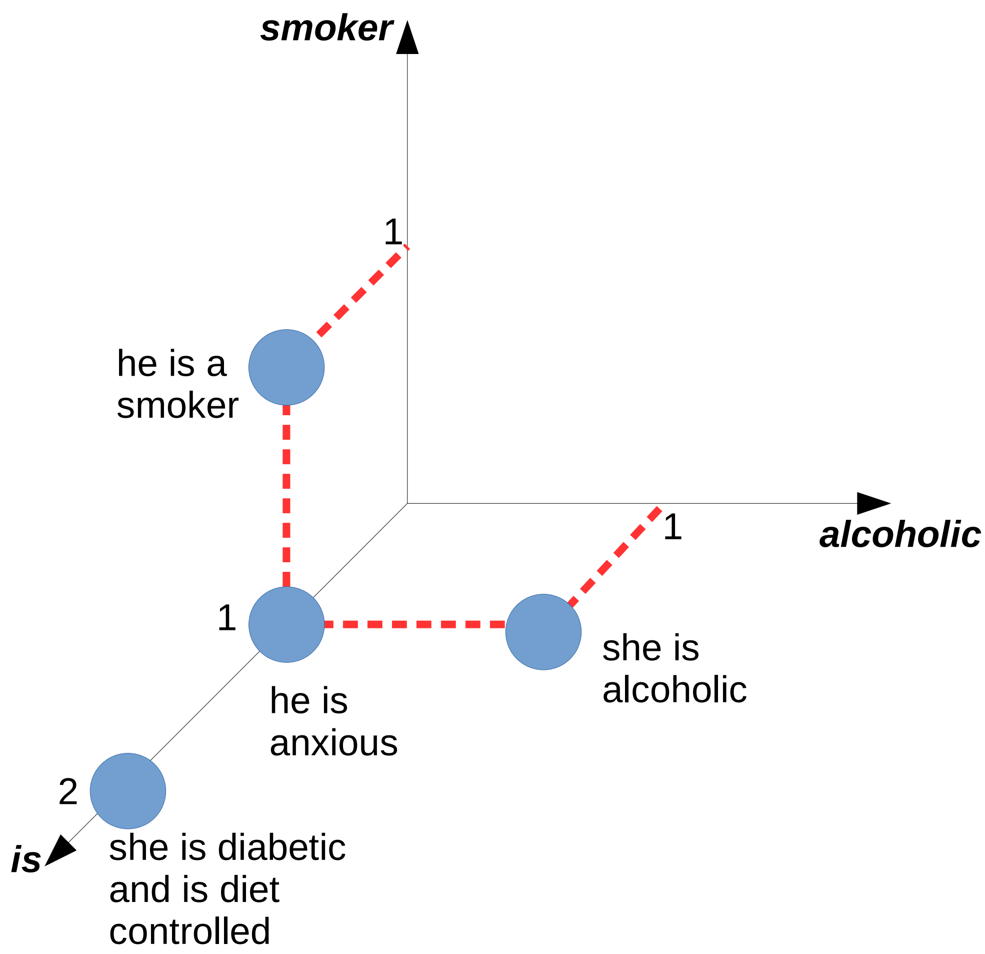
Bag of words

- Works surprisingly well on some problems
- But…
  - No word order: loss of context
  - The Curse of Dimensionality: the power of our classifier reduces as the number of dimensions increases
  - Over fitting: given a low number of training instances relative to number of features
  - Important but less frequent words can have less of an influence than less important but more frequent words

Term frequency and inverse document frequency

Improving bag of words

- The occurrence of a rare word like "smoker" or "alcoholic" has as much influence on the vector as the occurrence of a common word like "is"
- Lots of occurrences of a common word (such as two mentions of "is" in one sentence) has a bigger effect than a more discriminating rare word



TFIDF

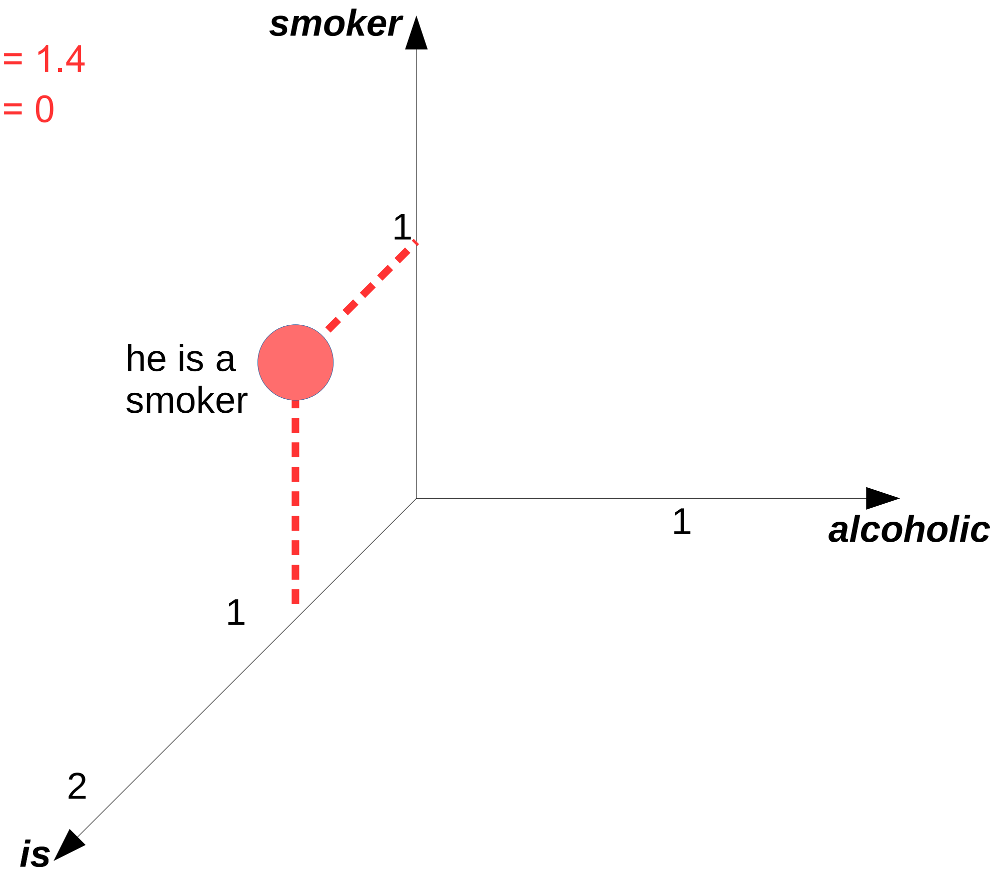
- The occurrence of a rare word like “smoker” or “alcoholic” has as much influence on the vector as the occurrence of a common word like “is”
- Lots of occurrences of a common word (such as two mentions of “is” in one sentence) has a bigger effect than a more discriminating rare word
- We can scale our BoW ***term frequencies***, multiplying each by a factor that accounts for how rare the term is – an ***inverse document frequency (idf)***

$$idf\ for\ word = \frac{number\ of\ documents}{number\ of\ documents\ containing\ word}$$

- Usually this is scaled further by taking the log
- The rarer a word, the higher idf
- The more common a word, the lower idf
- tf x idf scales the influence of each term accordingly

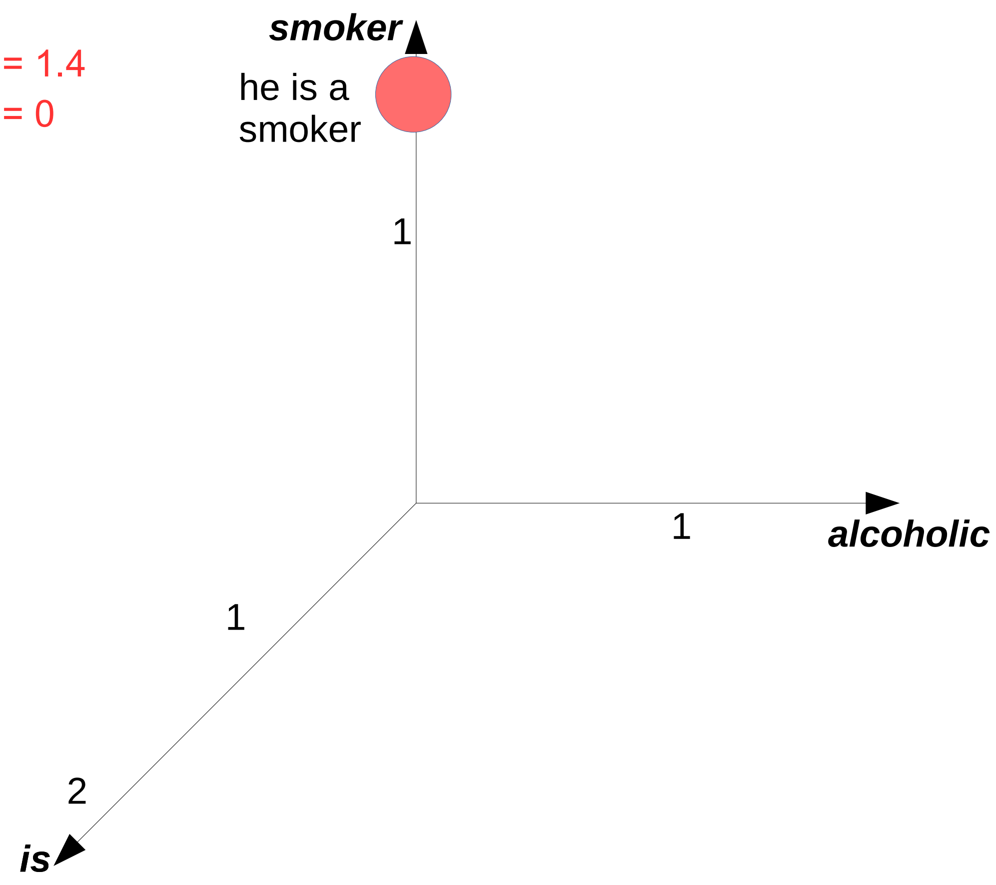
Improving bag of words

Document 1: tfidf(smoker) = 1 * ln(4/1) = 1.4
tfidf(is) = 1 * ln(4/4) = 0



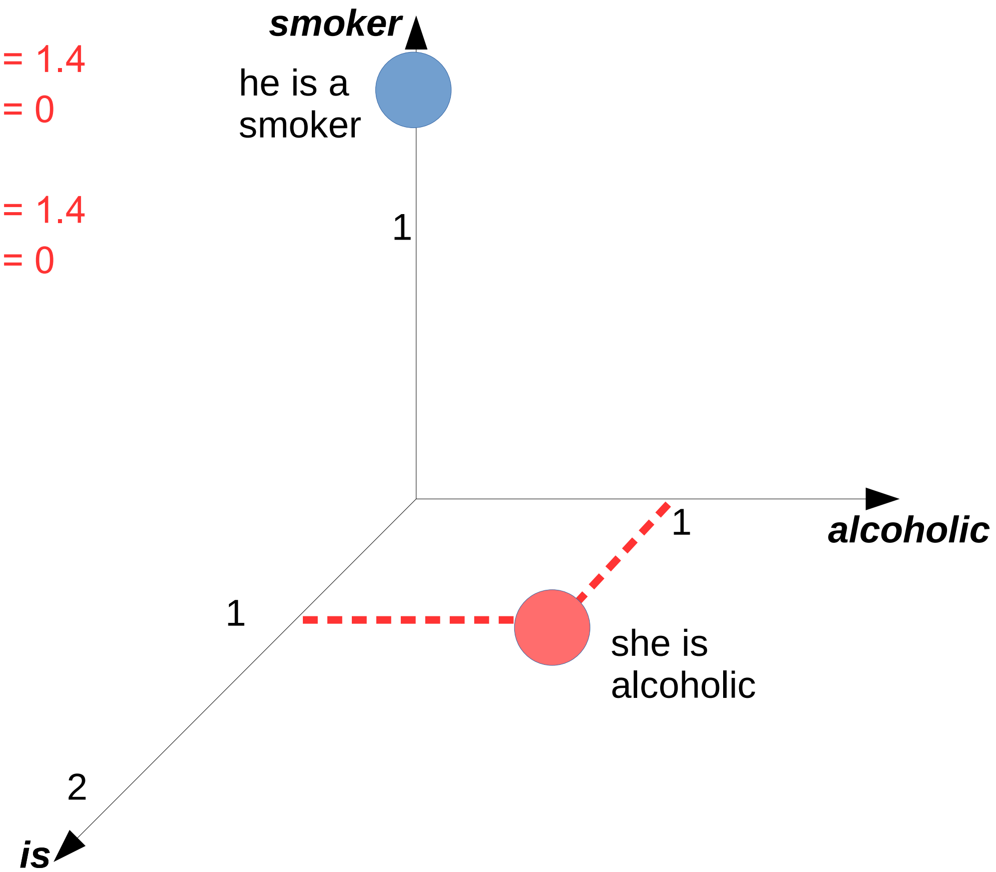
Improving bag of words

Document 1: tfidf(smoker) = 1 * ln(4/1) = 1.4
tfidf(is) = 1 * ln(4/4) = 0



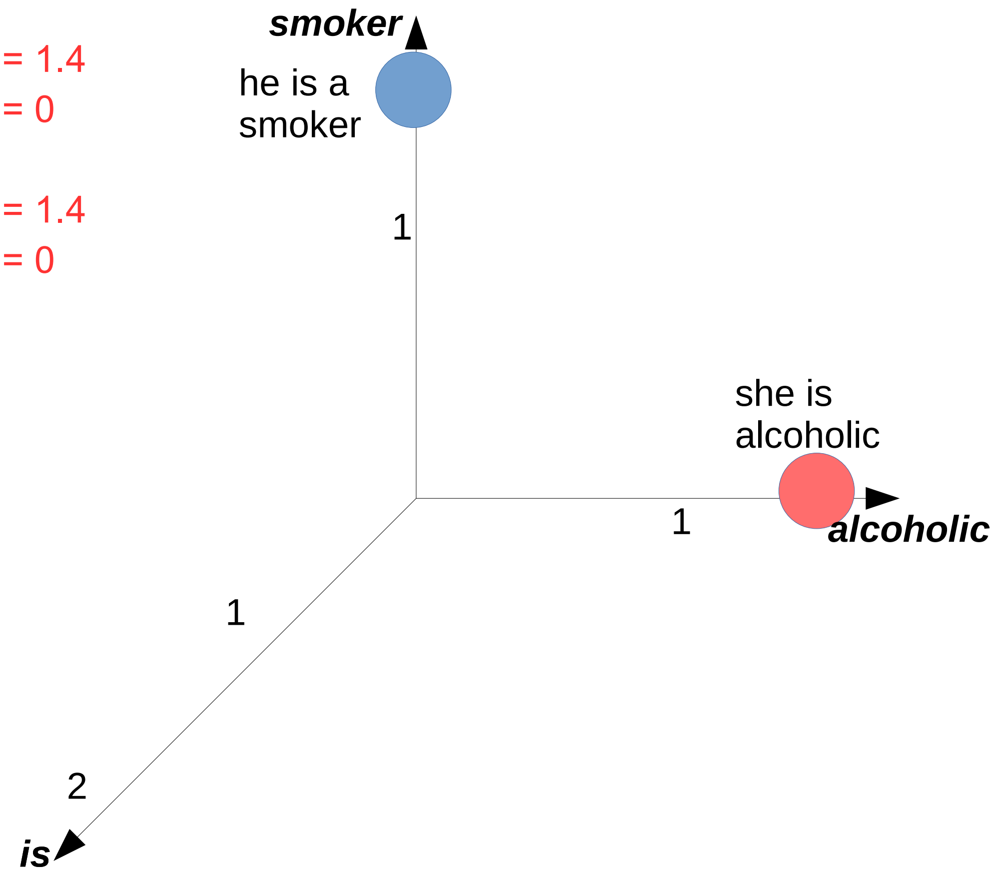
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0



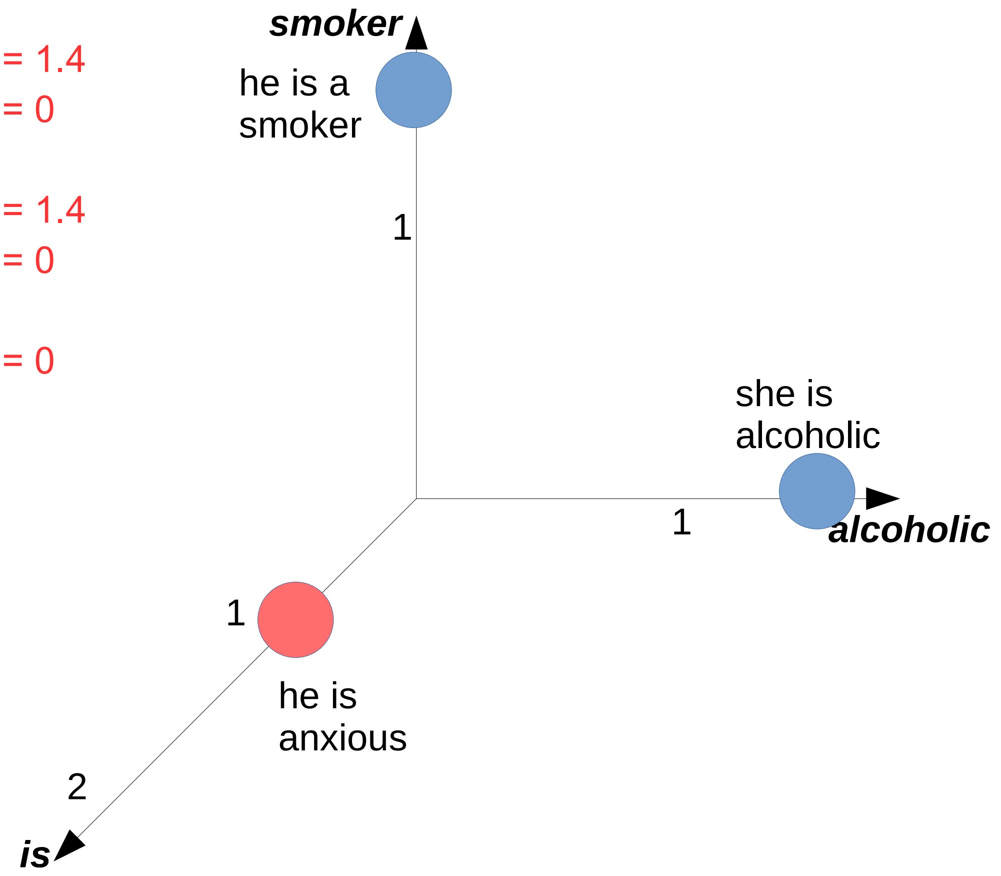
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0



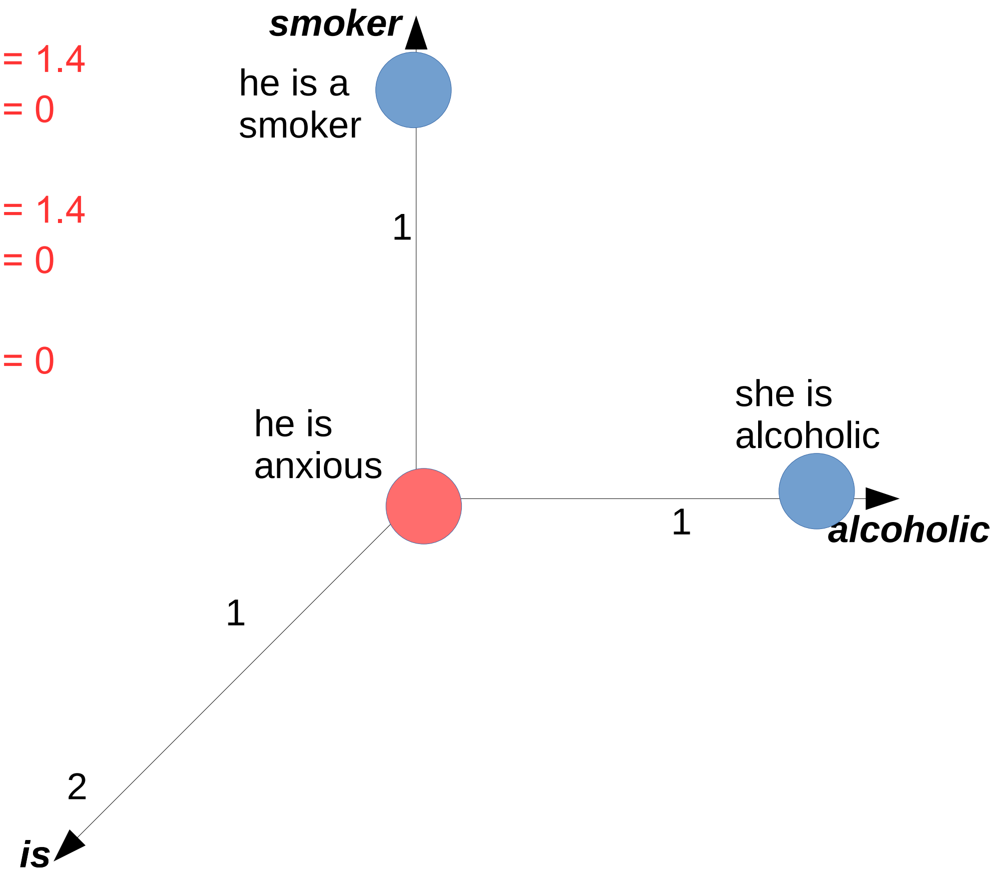
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 3:	tfidf(is)	= 1 * ln(4/4) = 0



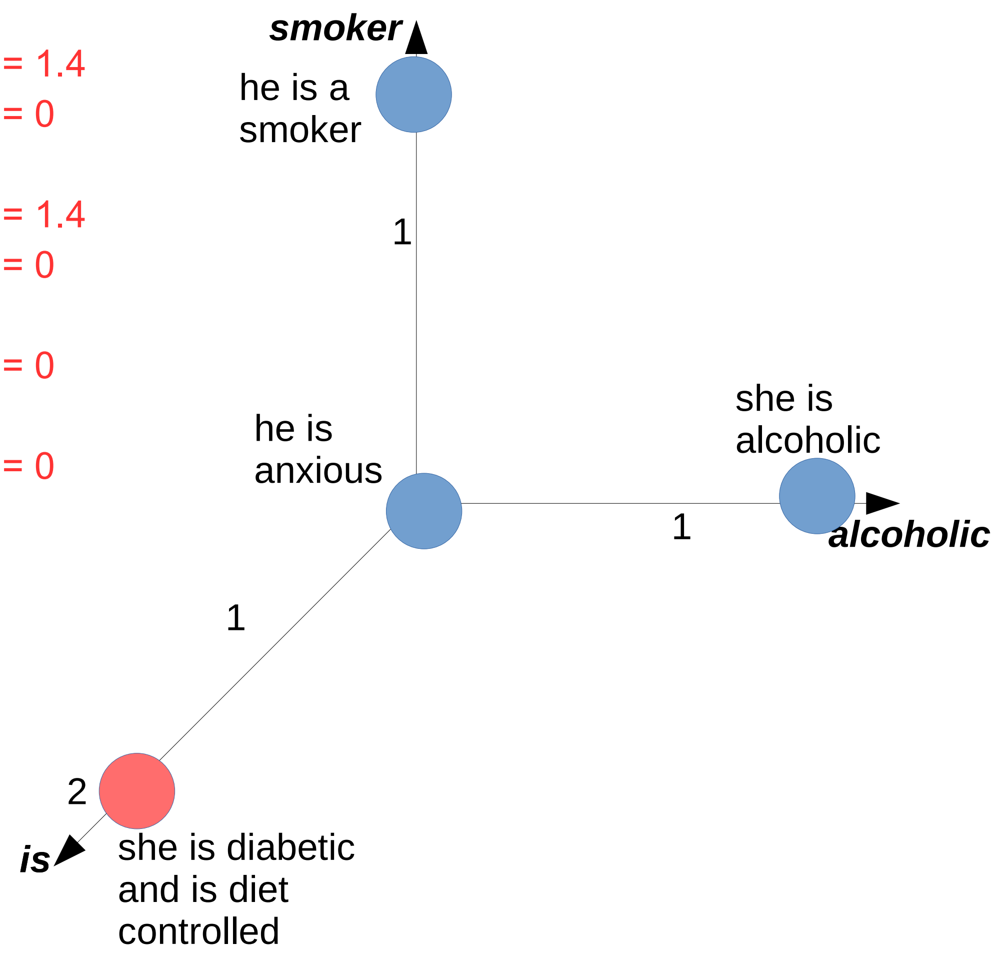
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 3:	tfidf(is)	= 1 * ln(4/4) = 0



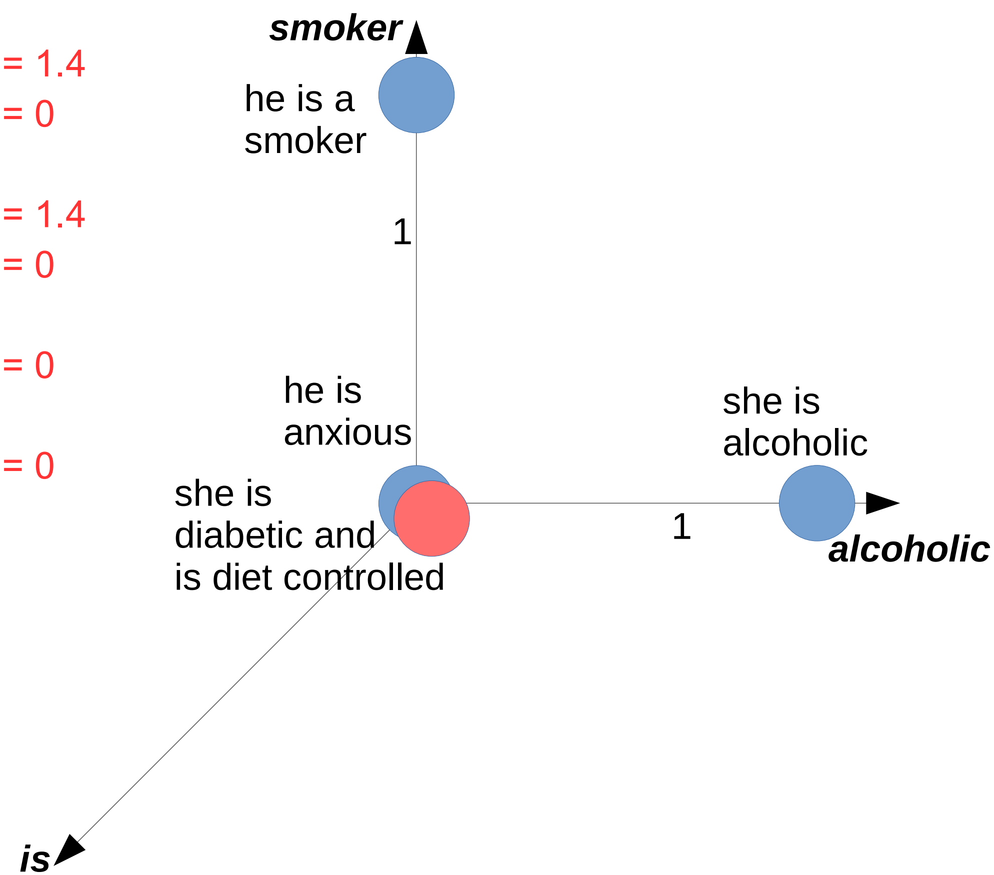
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 3:	tfidf(is)	= 1 * ln(4/4) = 0
Document 4:	tfidf(is)	= 2 * ln(4/4) = 0



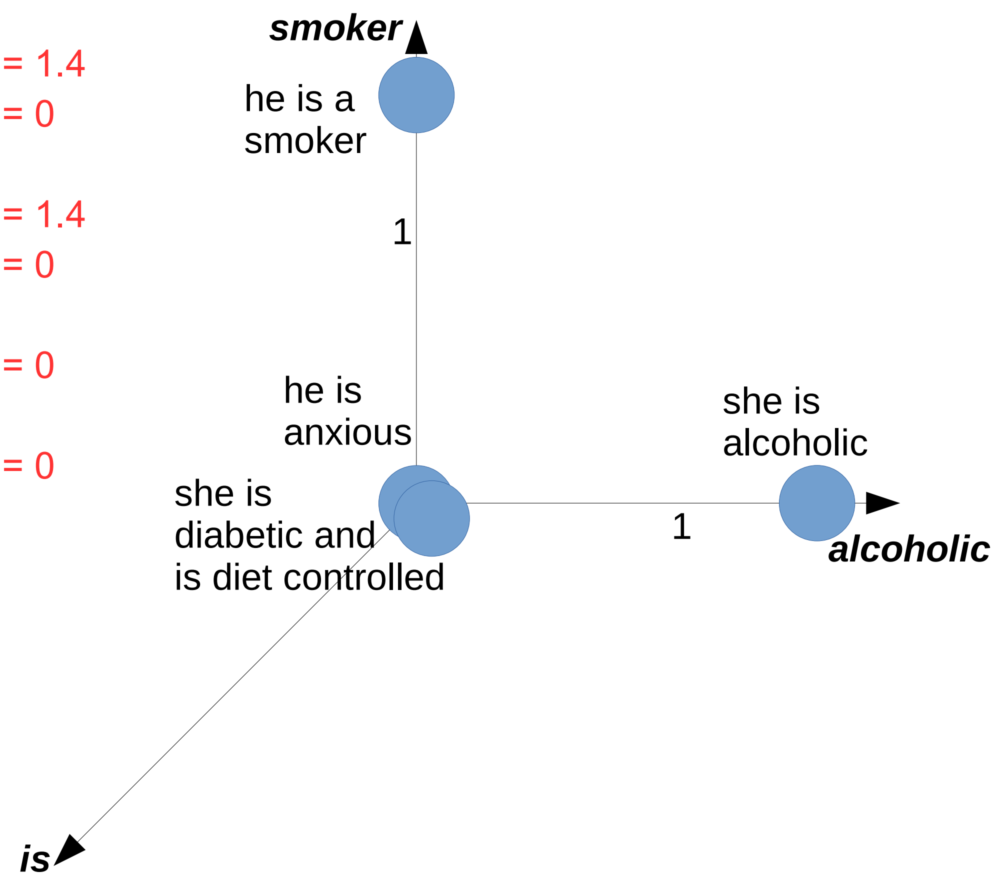
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 3:	tfidf(is)	= 1 * ln(4/4) = 0
Document 4:	tfidf(is)	= 2 * ln(4/4) = 0



Improving bag of words

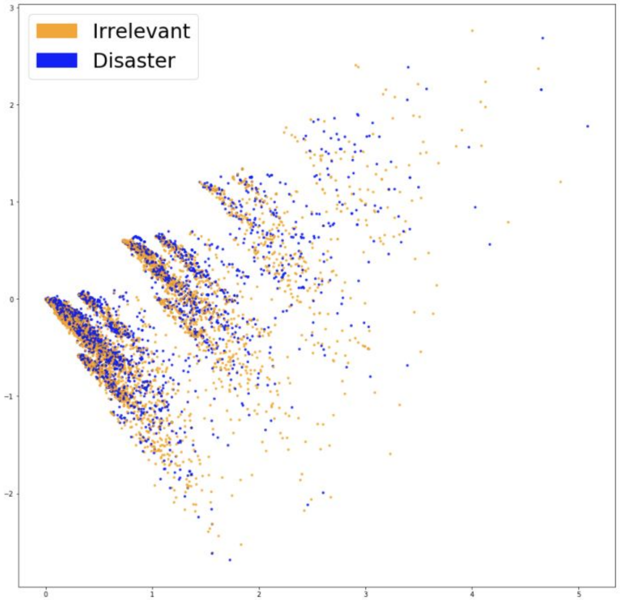
Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 3:	tfidf(is)	= 1 * ln(4/4) = 0
Document 4:	tfidf(is)	= 2 * ln(4/4) = 0

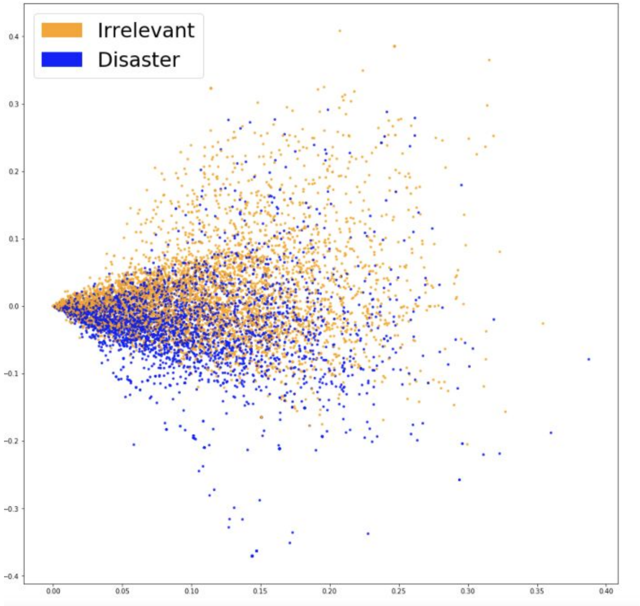


- Influence of rare words increased
- Influence of common words decreased

BoW vs TFIDF

Projections on to two dimensions of BoW (left) and TFIDF (right) vector spaces for words in tweets about disasters, and tweets not about disasters





From:
https://blog.insightdatascience.com/how-to-solve-90-of-nlp-problems-a-step-by-step-guide-fda605278e4e

More complex features

- Introduce word order
- Reduce dimensions and find the commonalities: increase ratio of instances to features
  - Morphological roots / lemmas
  - Parts of speech – he, she → pronouns
  - Semantic classes – mother, father → parent
- Introduce context
  - dependencies between parts of the sentence – e.g. subject, verb and object
  - embeddings

Thank you.
Any questions?

angus.roberts@kcl.ac.uk

NIHR | Maudsley Biomedical Research Centre

Modelling language: words

Angus Roberts, Senior Lecturer in Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King's College London

Representing words and context

- BoW and TFIDF typically model a piece of text e.g. sentences or documents
- But how can we model words numerically?
  - Vector based representations
- And how can we take in to account
  - Their similarities
  - Their meaning, or semantics

NIHR | Maudsley Biomedical Research Centre

Distributional semantics

Wombling and snetches

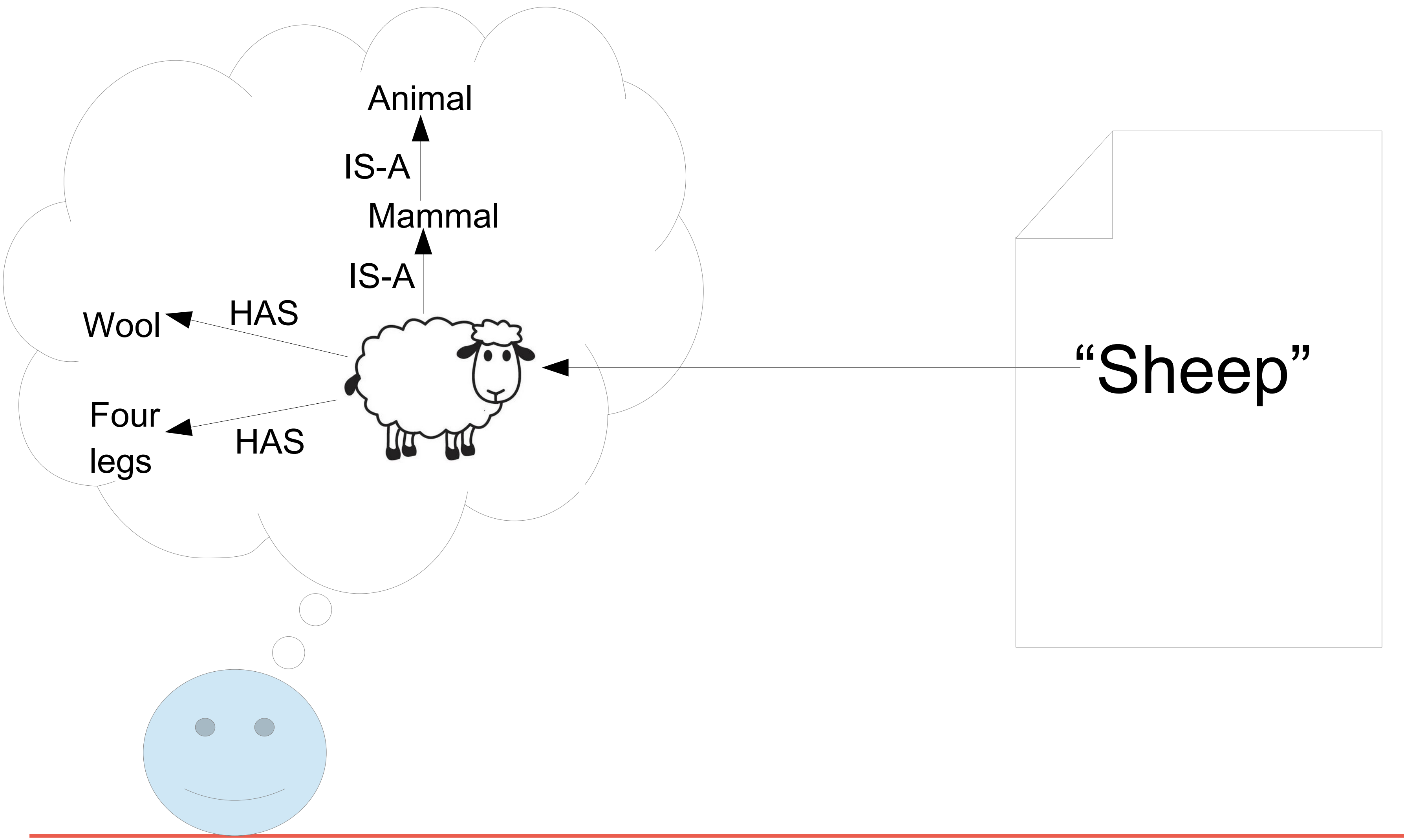
The Captain's side raked first. Tom staked. The hired sportsmen played so hard that they **wombled** too fast, and were shaky with the rakes. Tom fooled around the way he always did, and all his stakes dropped true. When it was his turn to rake he did not let Captain Najork and the hired sportsmen score a single rung, and at the end of the **snetch** he won by six ladders.

(How Tom beat Captain Najork and his hired sportsmen
Russell Hoban and Quentin Blake)

Distributional semantics

- “The meaning of a word is its use in language” (Wittgenstein)
- “You shall know a word by the company it keeps” (Firth)
- The contexts in which words appear correlate with their meaning
- We understand a word by its distribution: the set of contexts in which it is found
- “Don't think, but look!” (Wittgenstein)

Lexical semantics



Formal semantics and lexical semantics

- A contrast to distributional semantics
- Formal semantics
  - models the relationship between language and the world
  - defines meaning in terms of this model
  - defines languages in terms of formal logic
- The lexicon is defined as mappings from words to structured, conceptual knowledge

Complementary

- Distributional semantics is based on **statistics**, formal semantics on **formal mathematics**
- Distributional semantics is **differential**, lexical semantics is **referential**
- Distributional semantics is based on large **corpora**, lexical semantics (more often) on **structured lexicons**
- **Gathering a corpus** is easier than **building a lexicon**

Lack of grounding

- One criticism of distributional approaches is that they lack grounding in real world knowledge
- Consider the task of finding semantic features for sheep – which of these approaches are grounded?
  - Generated by psychology students (McRae, 2005):
    - have four legs, say “Baah”, have wool, are white
  - Generated from texts using a rule based approach (Barbu, 2009):
    - live on farms, graze, get scrapie
  - Collocates (nearby words) in Google (via WebCorp):
    - society, wool, association, breed…

References:
McCrae et al 2005, Semantic feature production norms for a large set of living and nonliving things, https://doi.org/10.3758/bf03192726
Barbu, 2010, Extracting conceptual structures from multiple sources, PhD thesis, University of Trento
http://www.webcorp.org.uk/live/

Representing context

Collocations

VE: The authors compared the **efficacy** of **olanzapine** and **lithium** in the prevention of mood
nd received open-label co-**treatment** with **olanzapine** and lithium for 6-12 weeks. Those meet
in Pharmacokinet. 1999 Sep;37(3):177-93. **Olanzapine**. **Pharmacokinetic** and **pharmacodynamic** p
patients with **schizophrenia** confirm that **olanzapine** is a novel **antipsychotic** agent with br
d with traditional **antipsychotic** agents, **olanzapine** causes a lower incidence of extrapyram
urbation of prolactin levels. Generally, **olanzapine** is well tolerated. The **pharmacokinetic**
okers and men have a higher **clearance** of **olanzapine** than women and nonsmokers. After admin
rred between olanzapine and alcohol, and **olanzapine** and imipramine, implying that **patients**
:485-92. doi: 10.1192/bjp.bp.107.037903. **Olanzapine** for the **treatment** of borderline person
o evaluate treatment with variably **dosed** **olanzapine** in individuals with borderline persona
double-blind trial, individuals received **olanzapine** (2.5-20 **mg**/day; n=155) or placebo (n=1
rried-forward methodology. RESULTS: Both **olanzapine** and **placebo** groups showed significant
p. CONCLUSIONS: Individuals **treated** with **olanzapine** and **placebo** showed significant but not
he types of **adverse events** observed with **olanzapine** **treatment** appeared similar to those ob
is study compared three **dosage** ranges of **olanzapine** (5 +/- 2.5 **mg**/day [Olz-L], 10 +/- 2.5

- Collocates from www.webcorp.org.uk
- Restricted to *.ncbi.nlm.nih.gov (i.e. mostly PubMed abstracts)

Contexts as matrices

- Build matrices of event frequencies, where events are words in documents
  - **Row:** words (or terms)
  - **Column:** colocated words (or documents or sentences, or….)
- Bag of words, with the bag represented as a vector (row)
- The row gives a “signature” of the word / term
- Sequential information is lost (at least in the simplest models)
- The matrix will be sparse

Word-Word matrices

	treatment	mg	anti-psychotic	placebo	patients
olanzapine	110	86	76	75	73

- Top 5 collocates for olanzapine
- Collocates four to the left and right, from www.webcorp.org.uk
- Restricted to *.ncbi.nlm.nih.gov (i.e. mostly PubMed abstracts)
- Normalised to collocates per 1000 hits

Word-Word matrices

	treatment	mg	anti-psychotic	placebo	patients
olanzapine	110	86	76	75	73
clozapine	70	30	78	0	89

- Top 5 collocates for olanzapine
- Collocates four to the left and right, from www.webcorp.org.uk
- Restricted to *.ncbi.nlm.nih.gov (i.e. mostly PubMed abstracts)
- Normalised to collocates per 1000 hits

Word-Word matrices

	treatment	mg	anti-psychotic	placebo	patients
olanzapine	110	86	76	75	73
clozapine	70	30	78	0	89
vinegar	15	0	0	0	0

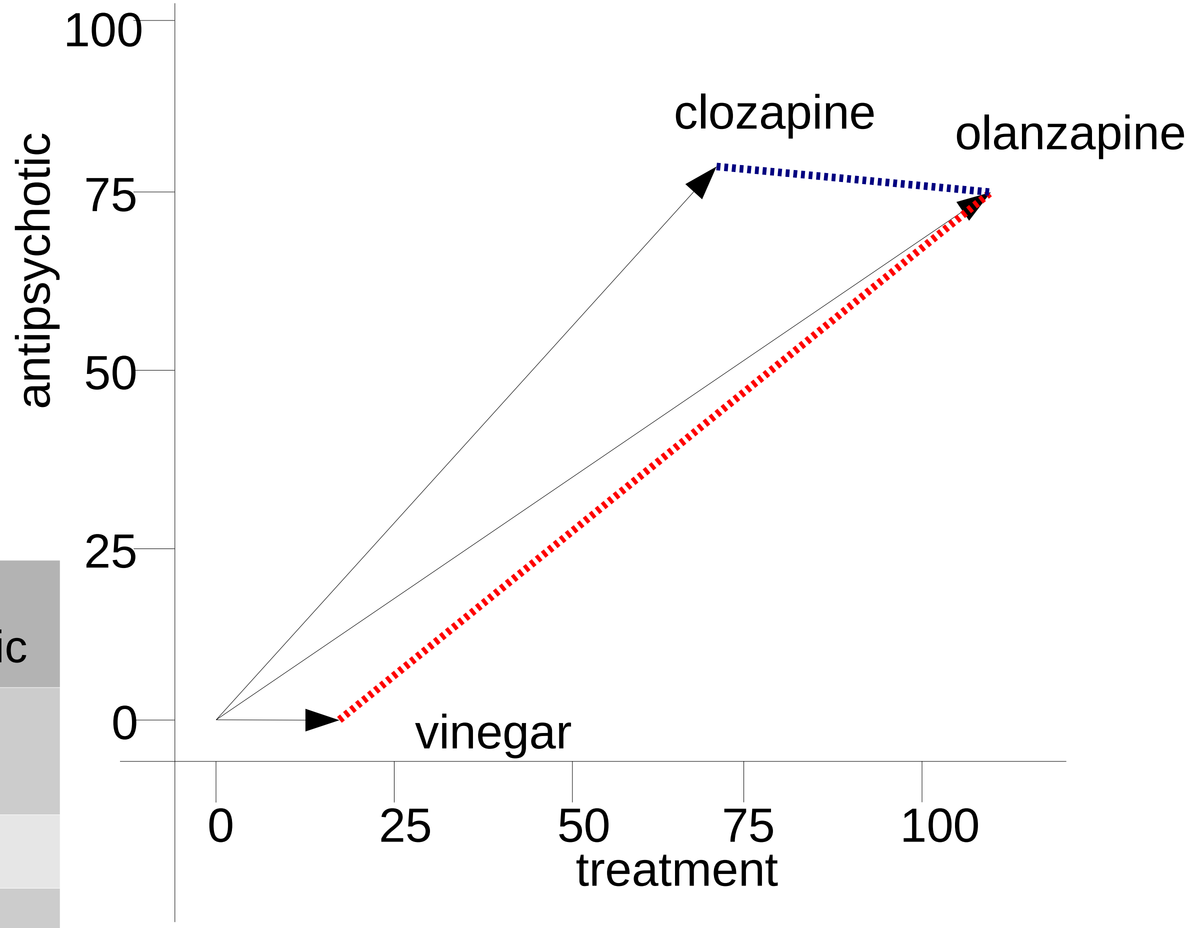
- Top 5 collocates for olanzapine
- Collocates four to the left and right, from www.webcorp.org.uk
- Restricted to *.ncbi.nlm.nih.gov (i.e. mostly PubMed abstracts)
- Normalised to collocates per 1000 hits

Word-Word matrices

	treatment	mg	anti-psychotic	placebo	patients	balsamic
olanzapine	110	86	76	75	73	0
clozapine	70	30	78	0	89	0
vinegar	15	0	0	0	0	109

- Top 5 collocates for olanzapine
- Collocates four to the left and right, from www.webcorp.org.uk
- Restricted to *.ncbi.nlm.nih.gov (i.e. mostly PubMed abstracts)
- Normalised to collocates per 1000 hits

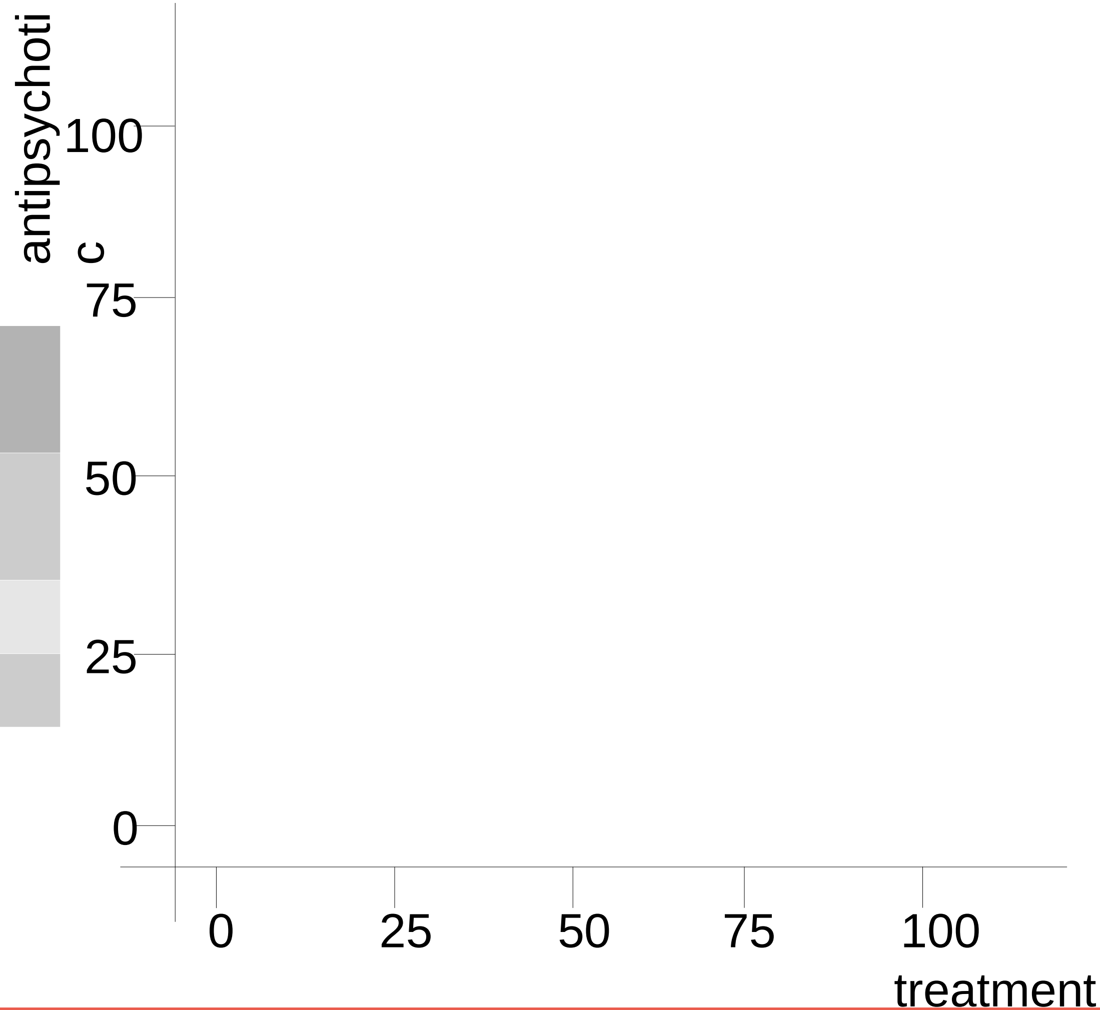
Semantic spaces



	treatment	anti-psychotic
olanzapine	110	76
clozapine	70	78
vinegar	15	0

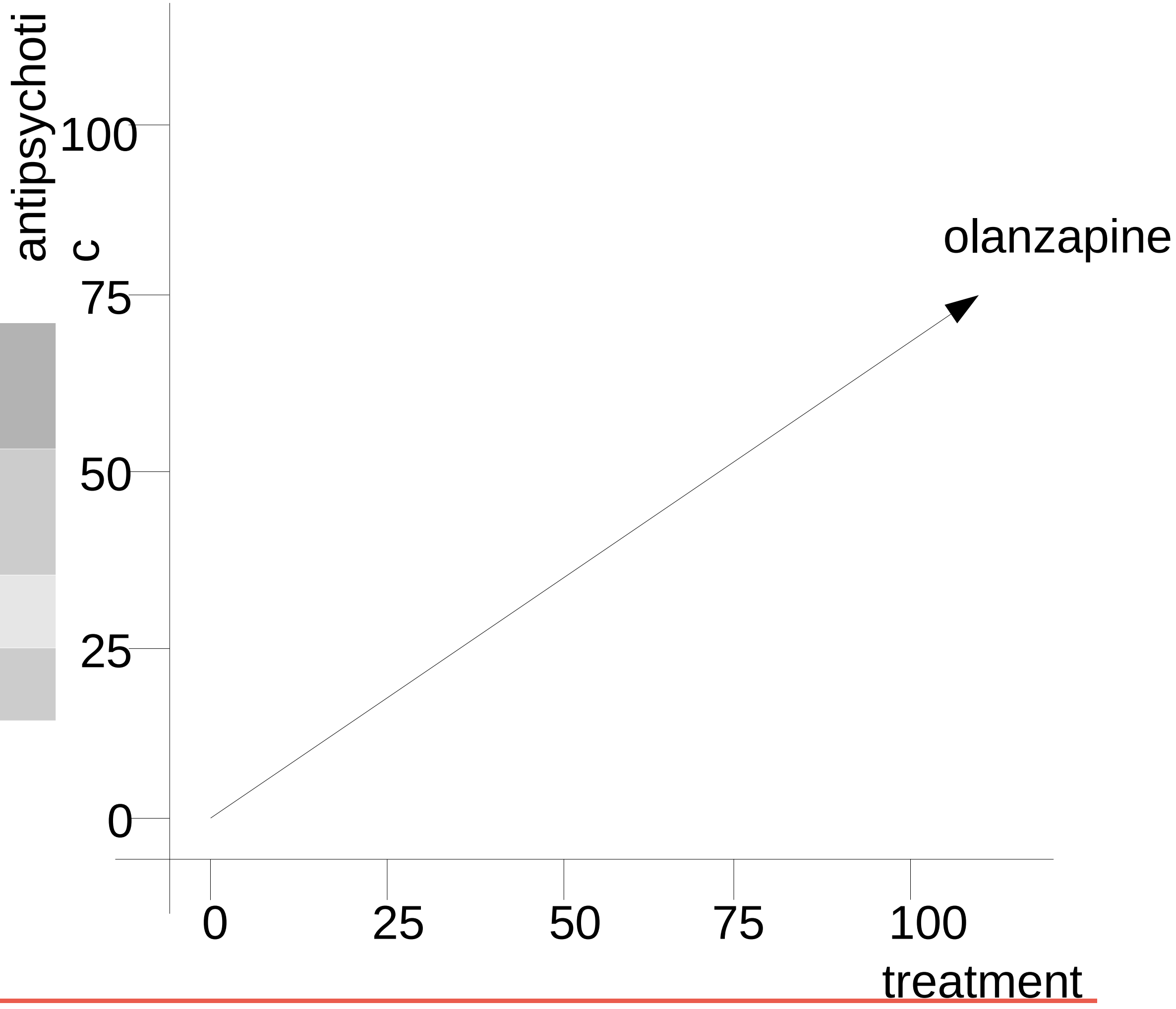
Words as vectors

	treatment	anti-psychotic
olanzapine	110	76
clozapine	70	78
vinegar	15	0



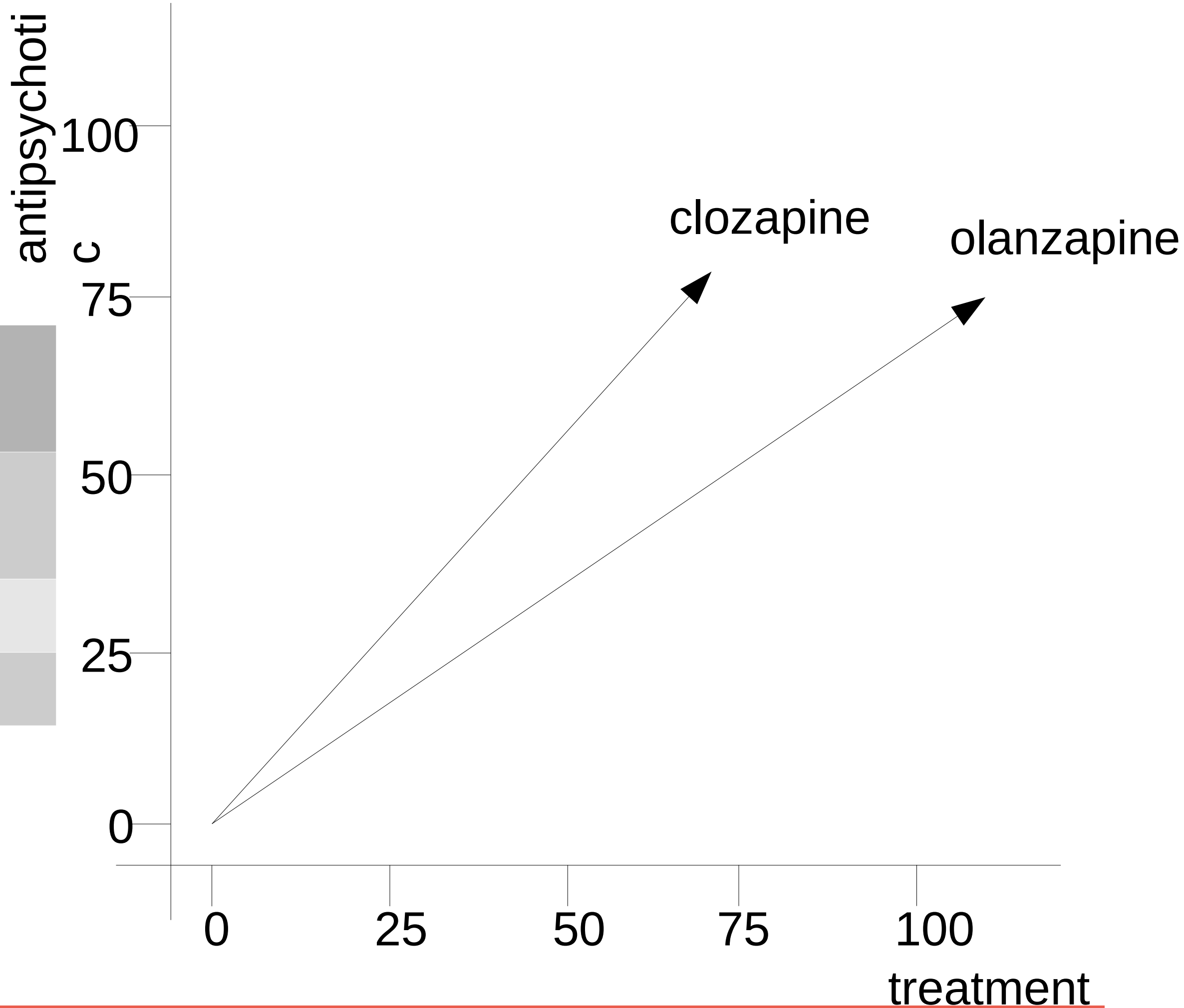
Words as vectors

	treatment	anti-psychotic
olanzapine	110	76
clozapine	70	78
vinegar	15	0



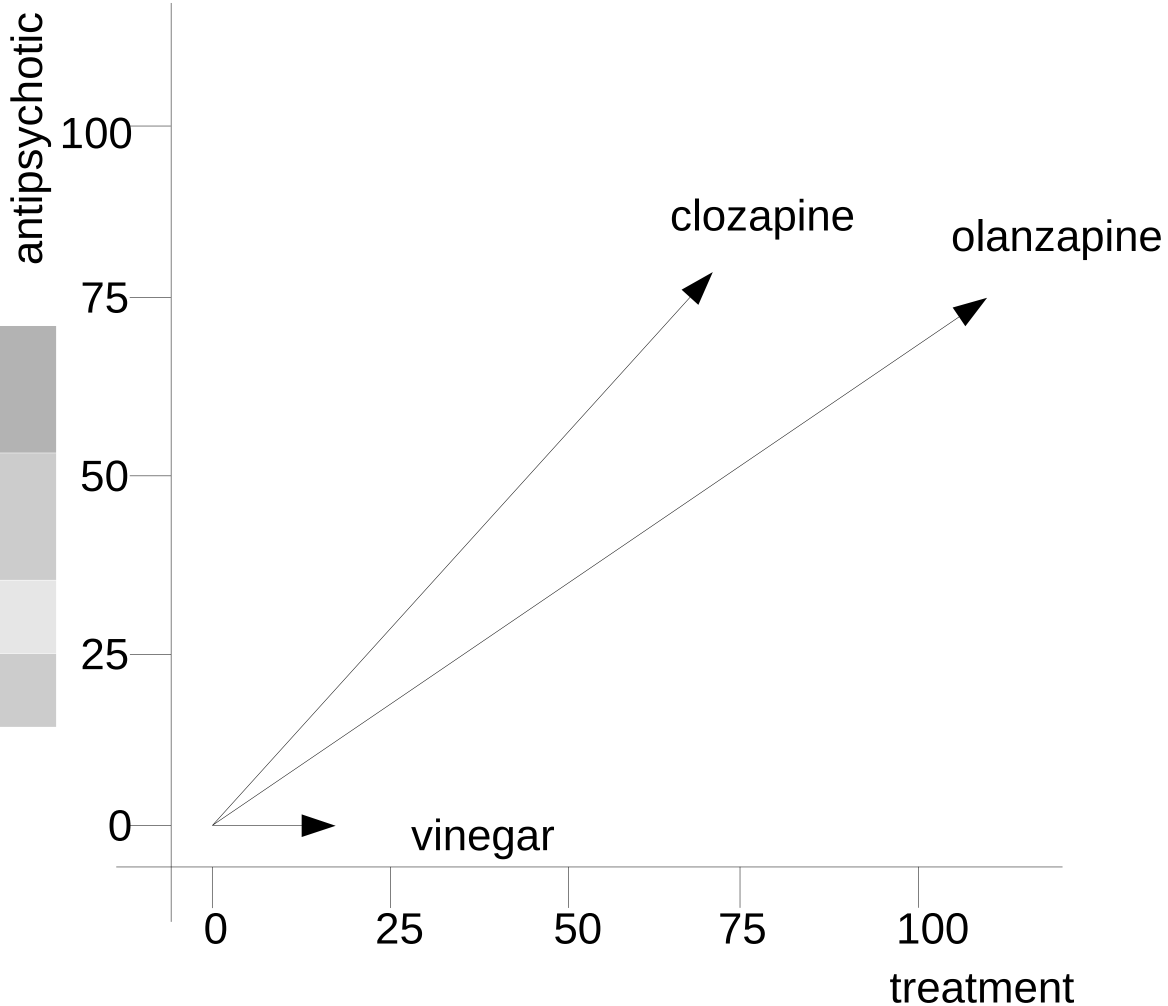
Words as vectors

	treatment	anti-psychotic
olanzapine	110	76
clozapine	70	78
vinegar	15	0



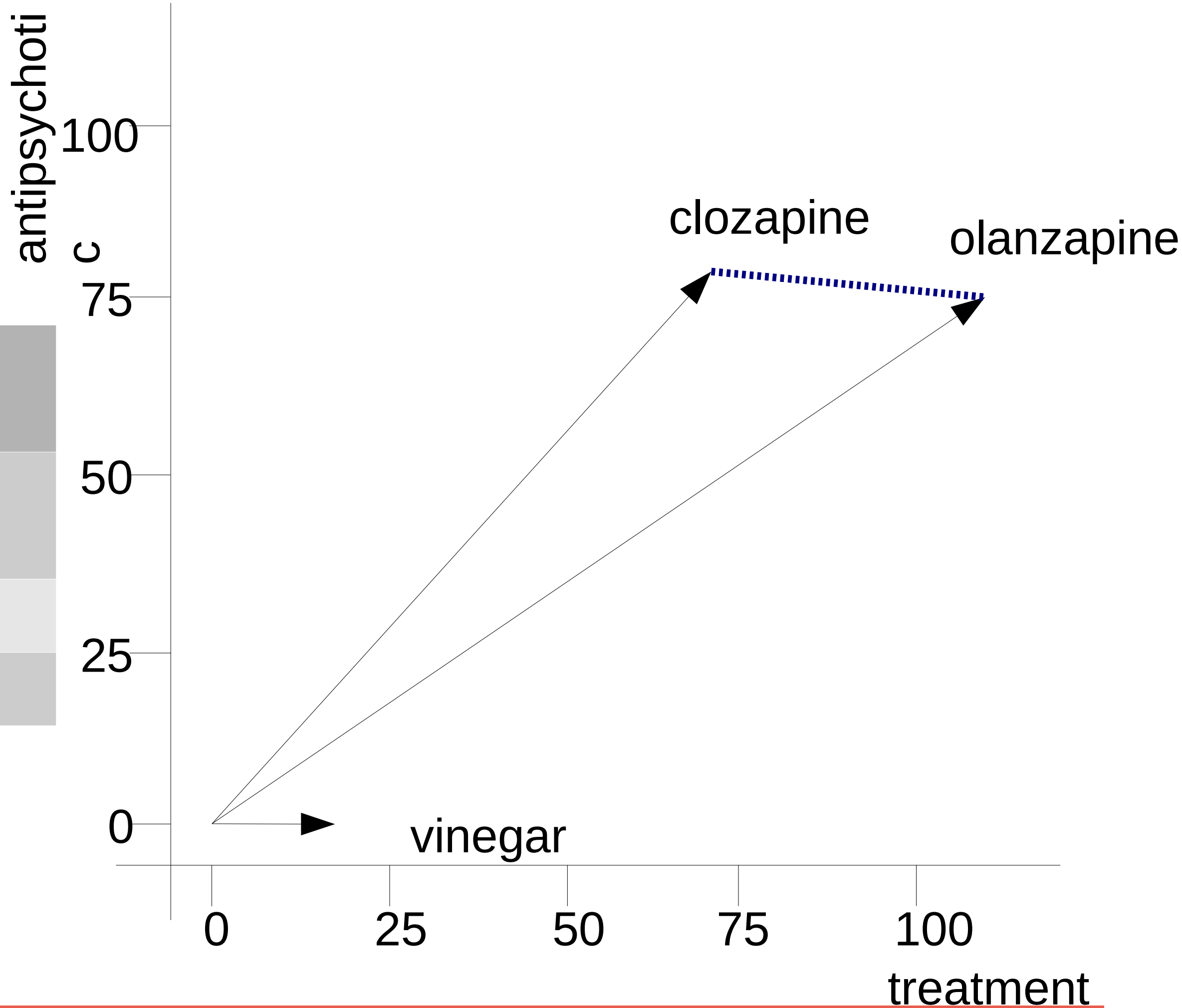
Words as vectors

	treatment	anti-psychotic
olanzapine	110	76
clozapine	70	78
vinegar	15	0



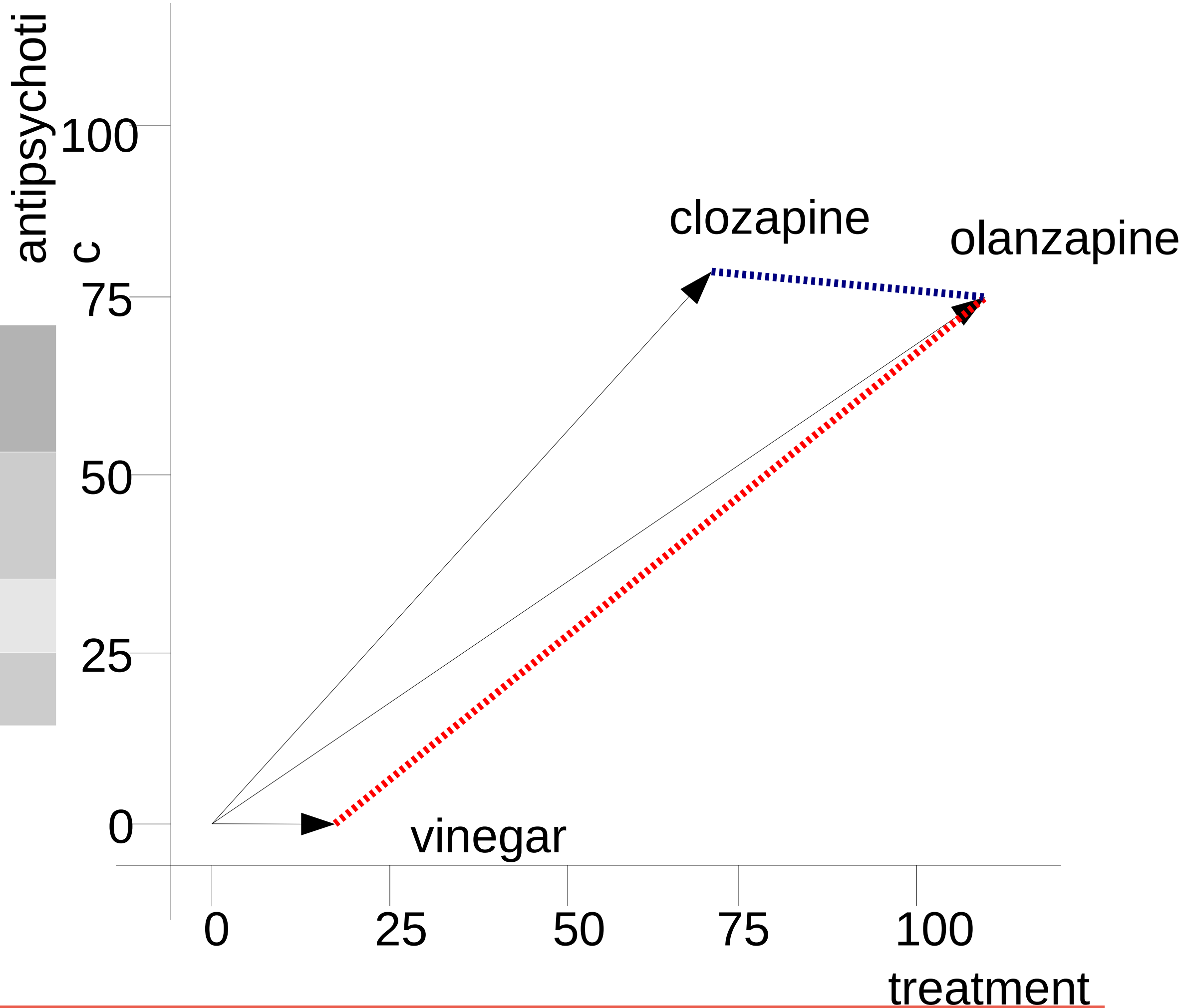
Words as vectors

	treatment	anti-psychotic
olanzapine	110	76
clozapine	70	78
vinegar	15	0



Words as vectors

	treatment	anti-psychotic
olanzapine	110	76
clozapine	70	78
vinegar	15	0



Tools for the job

- Rationalist NLP
  - An armchair
- Empirical NLP
  - A pile of documents (corpus)
  - A representation
- What about the algorithm to classify our words in representation space? Are they important?

NIHR | Maudsley Biomedical Research Centre

Thank you.
Any questions?

angus.roberts@kcl.ac.uk

NIHR | Maudsley Biomedical Research Centre

Modelling language: distributed representations

Angus Roberts, Senior Lecturer in Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King’s College London

Representing language

- Encoding meaning
- Distributed representations and word embeddings
- Next steps: modelling language using artificial neural networks

Encoding meaning

One-hot encoding

- One-hot is a simple word-space vector representation. Words are represented by a vector encoding their position in an ordered vocabulary

aardvark	[1, 0, 0, 0, 0, …, 0, 0]
aargh	[0, 1, 0, 0, 0, …, 0, 0]
...	
zumba	[0, 0, 0, 0, 0, …, 1, 0]
zygote	[0, 0, 0, 0, 0, …, 0, 1]

- As well as being necessary to represent our words numerically, it is also a step along the path of finding some abstraction of word meaning
- Alternatively, we could encode as the integer position in the index

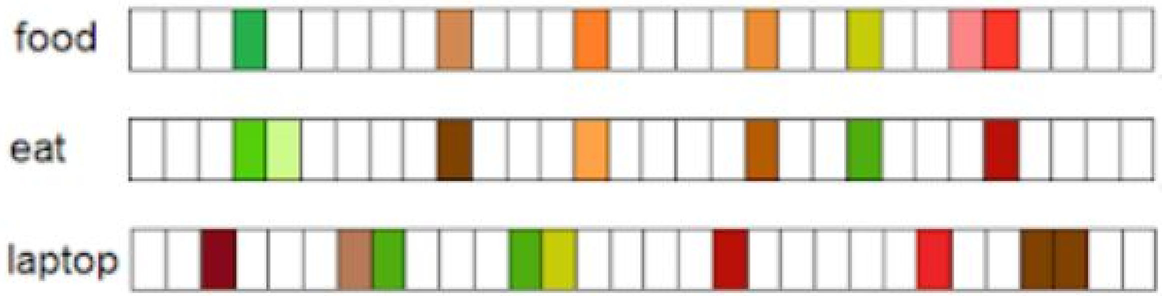
aardvark	0
aargh	1
...	
zumba	n-1
zygote	n

Encoding meaning

- Such a vector representation does not really encode meaning
- It is also high dimensional and sparse
- Can we encode meaning such a vector representation?
- Can we derive a low dimensional model of words?

Encoding meaning

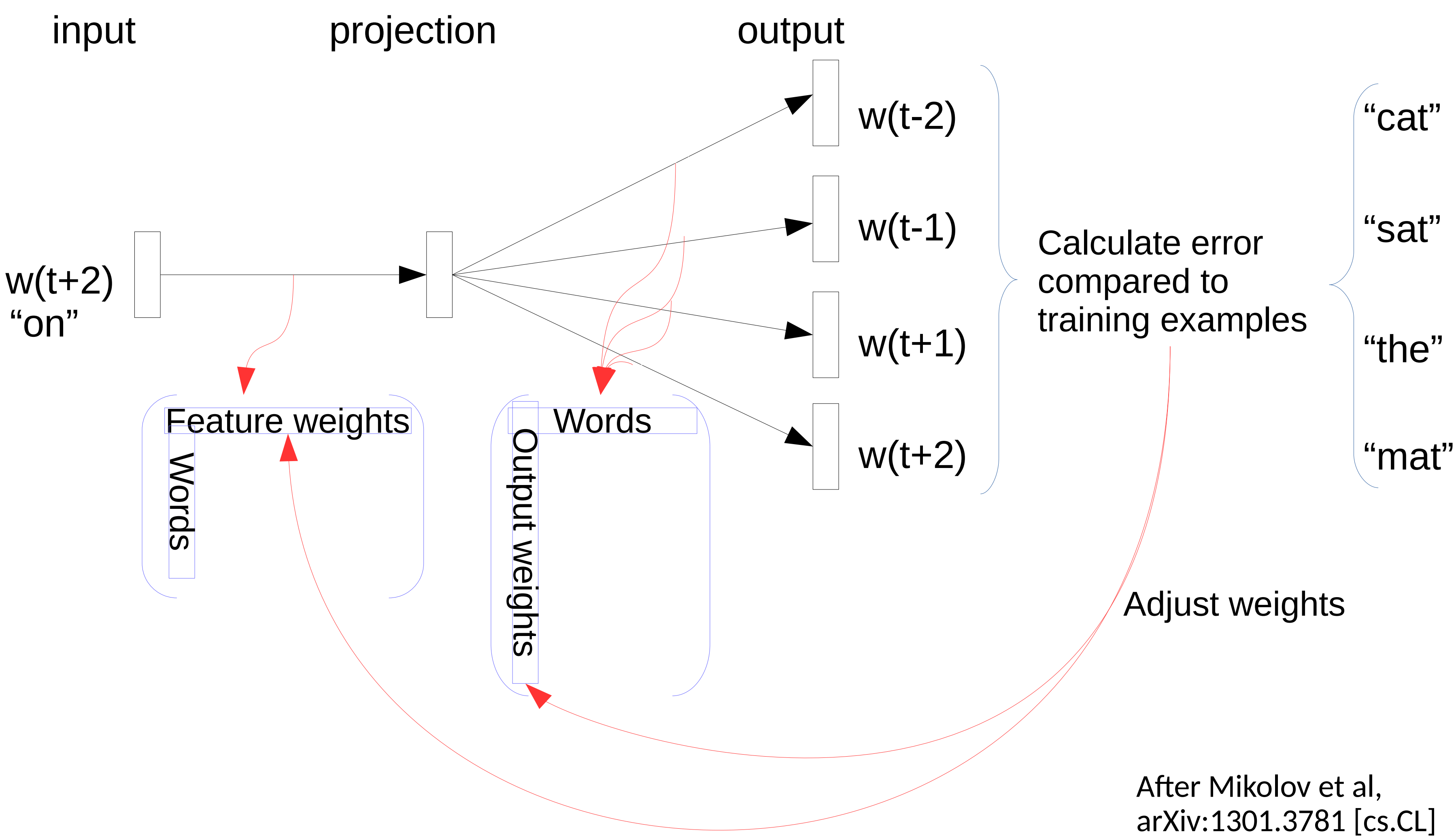
Can we define some space that is sufficient to encode the semantics of our language?



From http://veredshwartz.blogspot.co.uk

Distributed represntations: word embeddings

Distributed representations - Word2Vec



After Mikolov et al, arXiv:1301.3781 [cs.CL]

Training the vectors

- w – real number feature vectors
- c – real number output context vectors

- cat sat on the mat

  c1 c2 w c3 c4

  calculate: w.c1 + w.c2 + w.c3 + w.c4

  Adjust vector weights to make this high – maximise the probability of an example

- cat sat strawberry the mat

  c1 c2 w' c3 c4

  calculate: w'.c1 + w'.c2 + w'.c3 + w'.c4

  Adjust vector weights to make this low – minimise the probability of random replacements

Intuition

- Consider that “on” and “by” play similar roles in language:
  - cat sat on the mat
  - cat sat by the mat
- We would expect “on” and “by” to have similar feature vectors
- And for the other words, we can generalize further:
  - dog sits on a rug
  - dog lies under a rug
  - …

Intuition

- If two words have similar contexts, then their feature vectors will be similar
- The final feature vector for a word gives a distributed representation of the word – ***word embeddings*** – a dimensionality reduction from our word space to real number vectors
- (We throw away the output vectors – we don't need those)
- We use these word embedding as features in place of our words in models

Intuition

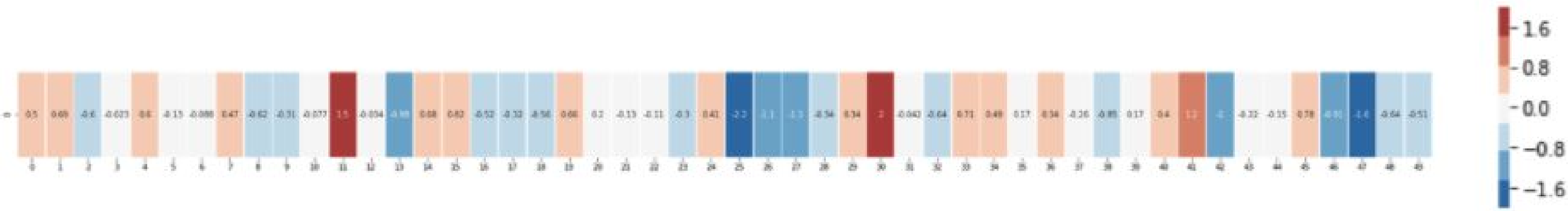
Construct a vector for the word “king”, (GloVe based vector, trained on Wikipedia):

```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 ,
-0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961
, -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.9927 ,
-0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 ,
-1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042 ]
```

Example from Jay Alammar, The illustrated Word2Vec: https://jalammar.github.io/illustrated-word2vec/

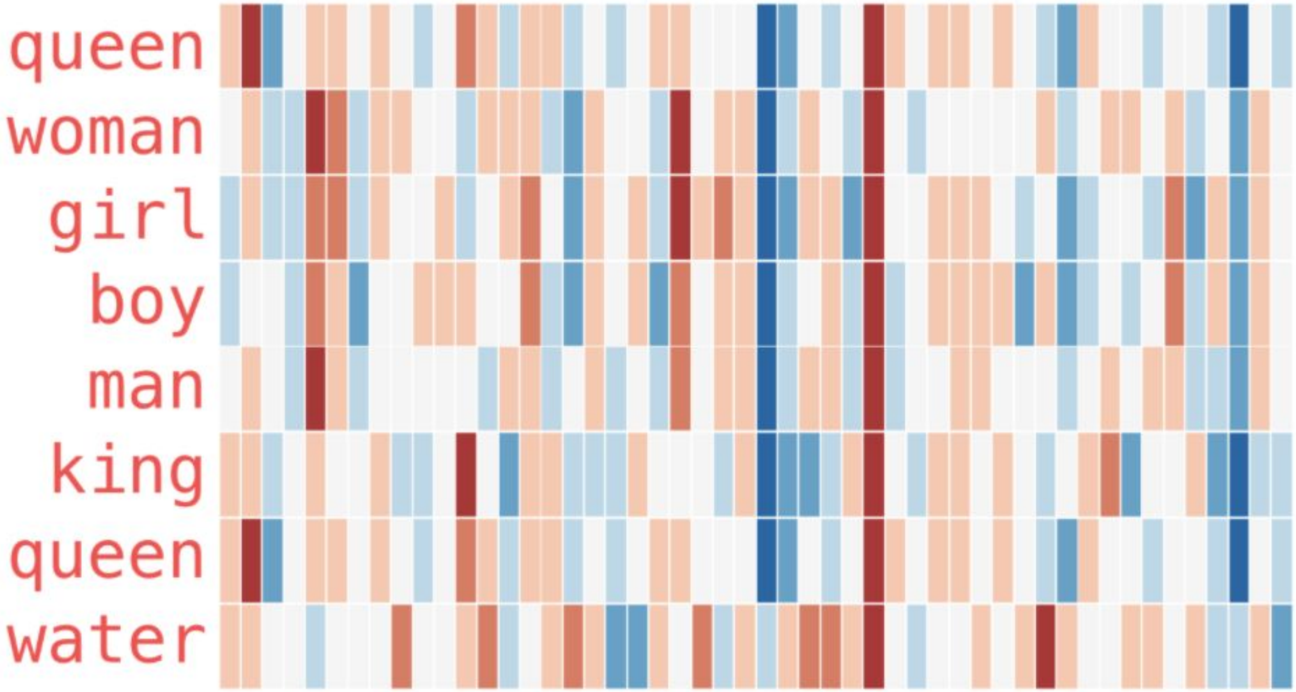
Intuition

Visualise as bands of different colours and intensities:

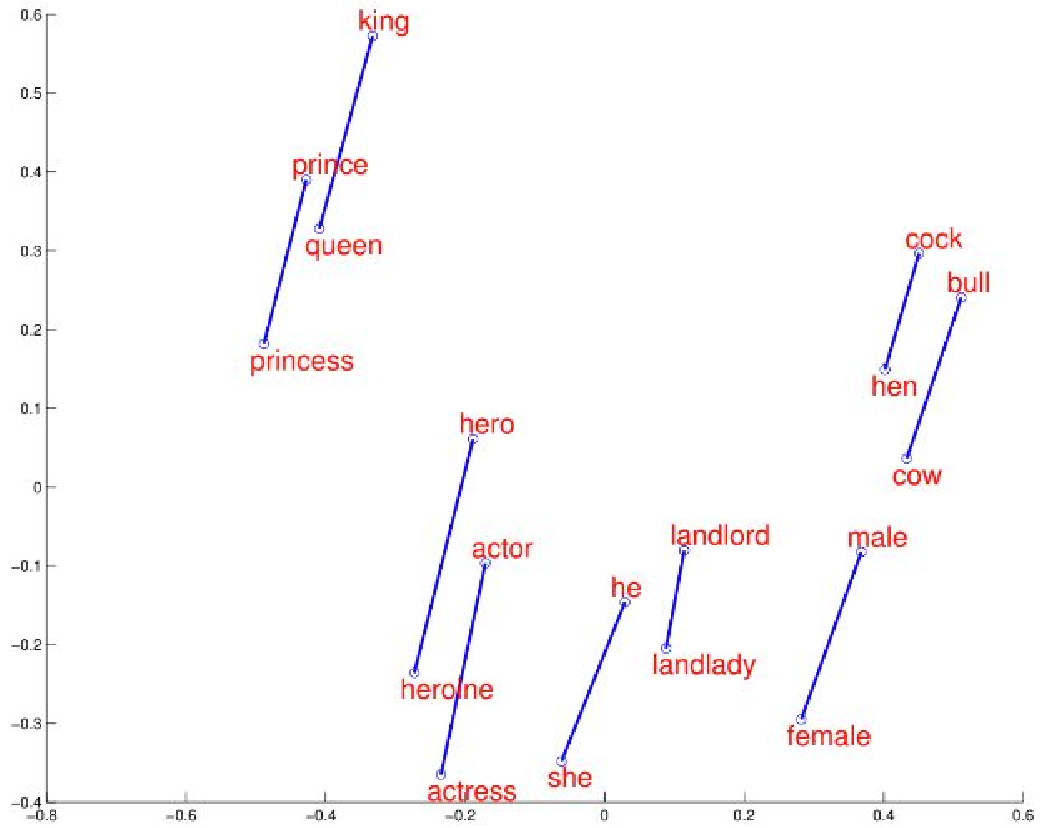


Intuition

Compare to vectors for other words:

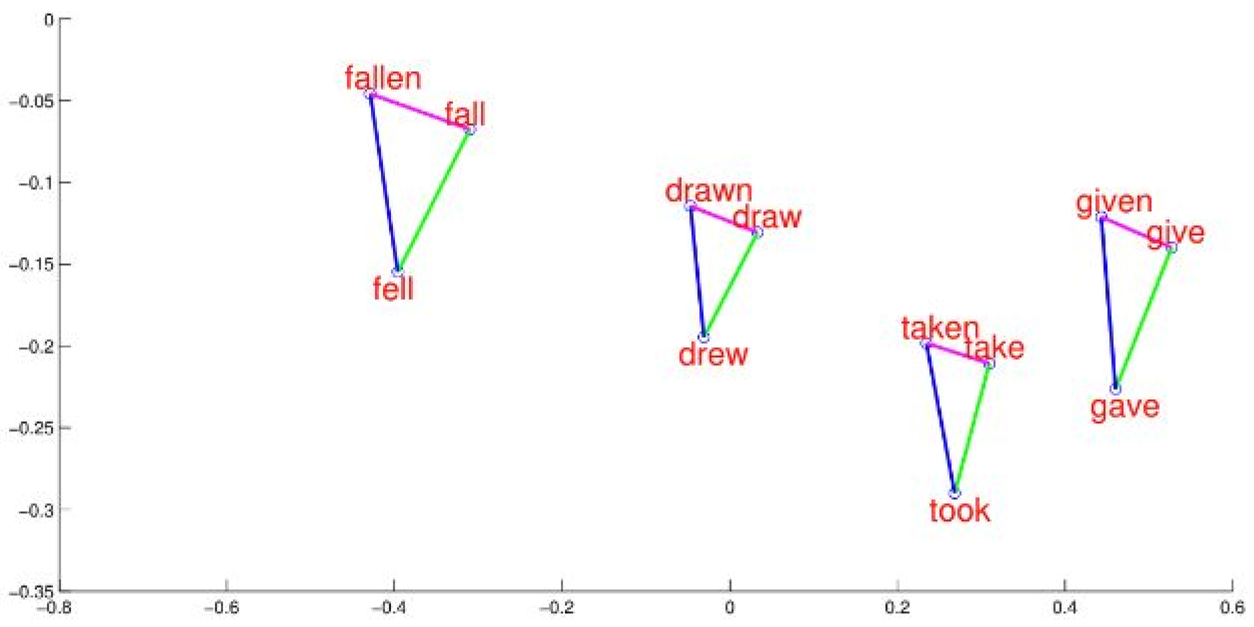


Visualisation



2D projection from Mikolov et al,
Google Research, NIPS 2013

Visualisation



2D projection from Mikolov et al,
Google Research, NIPS 2013

Next steps: modelling language with artificial neural nets

2010 onwards:
artificial neural networks

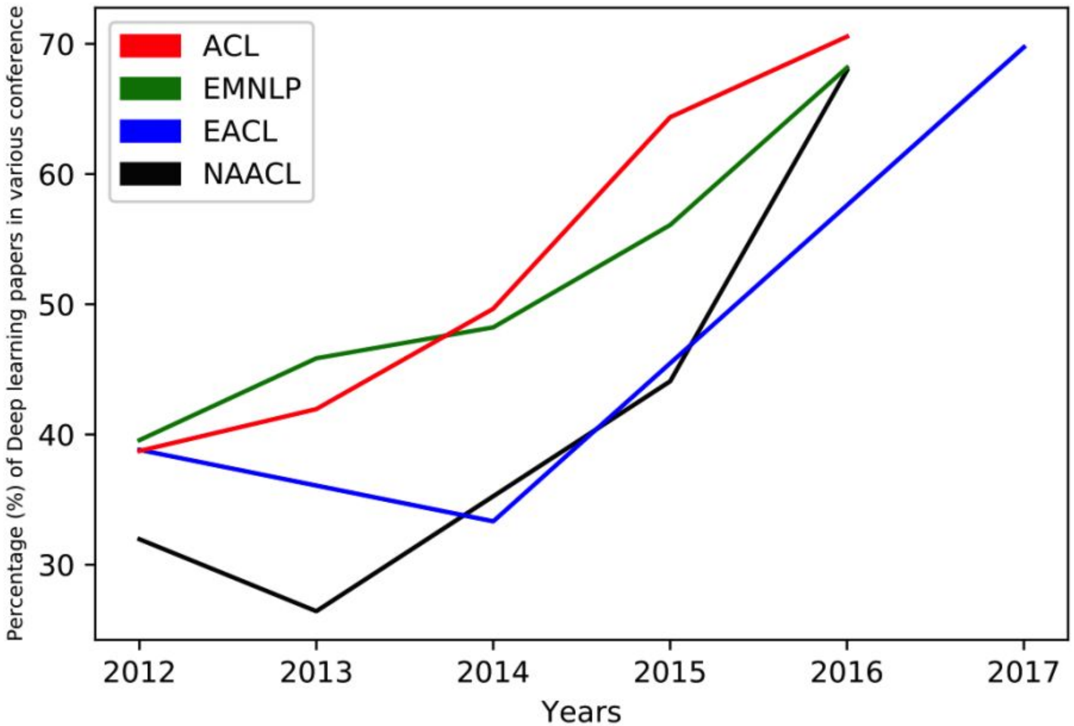


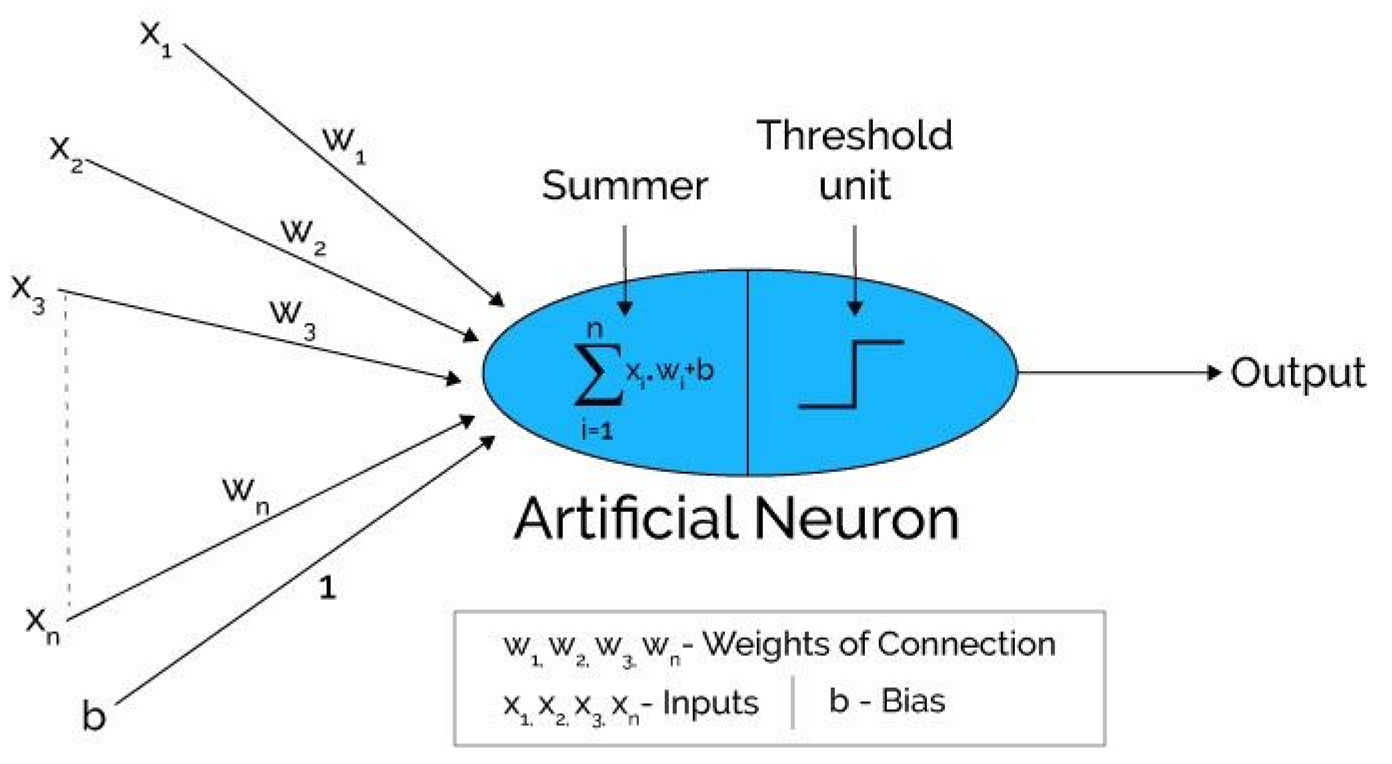
Fig. 1: Percentage of deep learning papers in ACL, EMNLP, EACL, NAACL over the last 6 years (long papers).

Young et al,
arXiv:1708.02709 [cs.CL]

Practical application – skills and trade offs

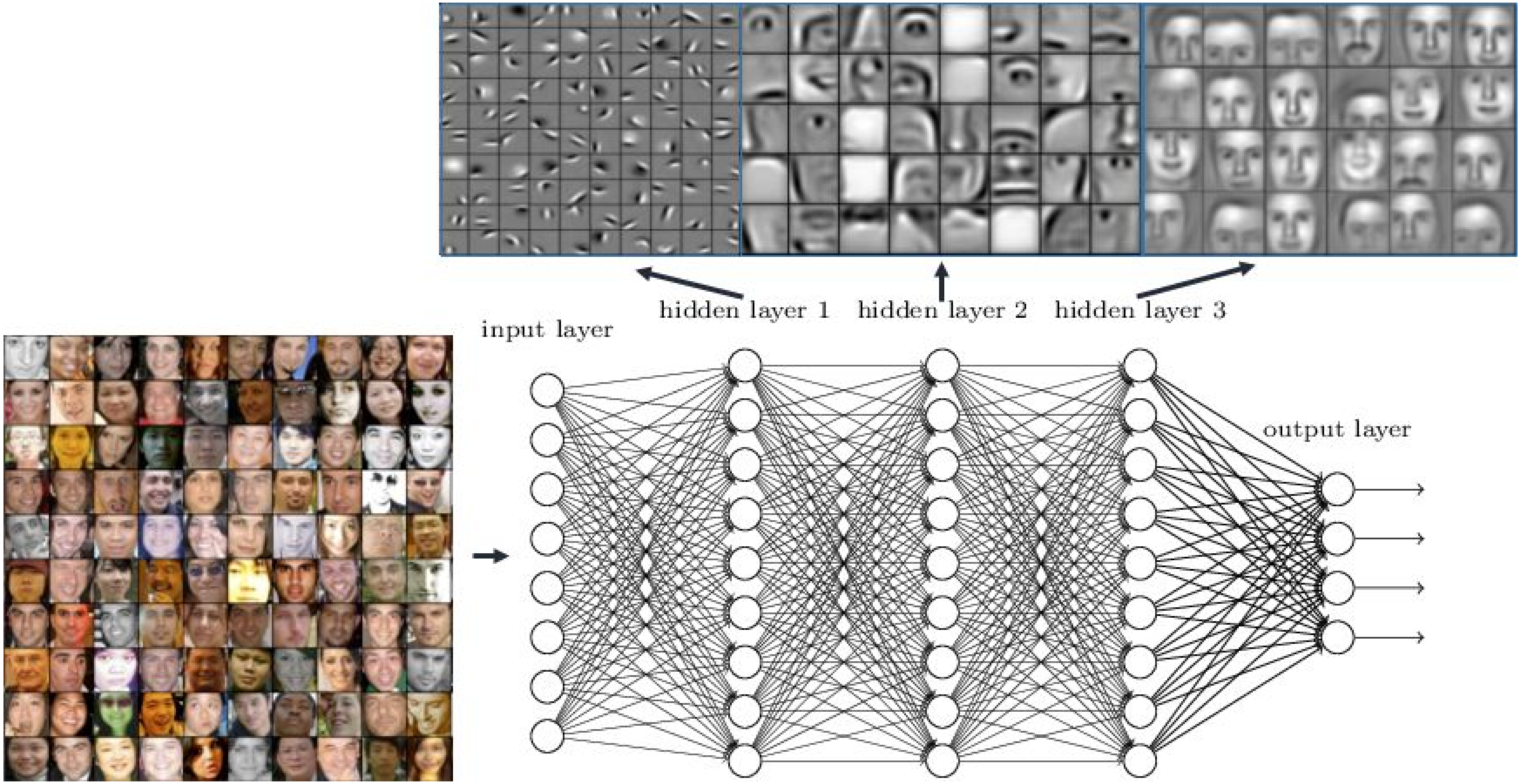
- Unsupervised models encapsulating many of the features of a language
- Feature engineering becomes less of a concern
- Domain expertise: training and evaluation examples required for the final task
- Often require large numbers of examples

2010 onwards:
artificial neural networks for NLP



From https://medium.com/@xenonstack/overview-of-artificial-neural-networks-and-its-applications-2525c1addff7

Learning hierarchical feature representations



From https://www.strong.io/blog/deep-neural-networks-go-to-the-movies

Thank you.
Any questions?

angus.roberts@kcl.ac.uk