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
- Word embeddings
  - an intuitive example
  - outline of calculation
  - visualising
- 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]
abacus	[0, 1, 0, 0, 0, …, 0, 0]
...	
zumba	[0, 0, 0, 0, 0, …, 1, 0]
zygote	[0, 0, 0, 0, 0, …, 0, 1]

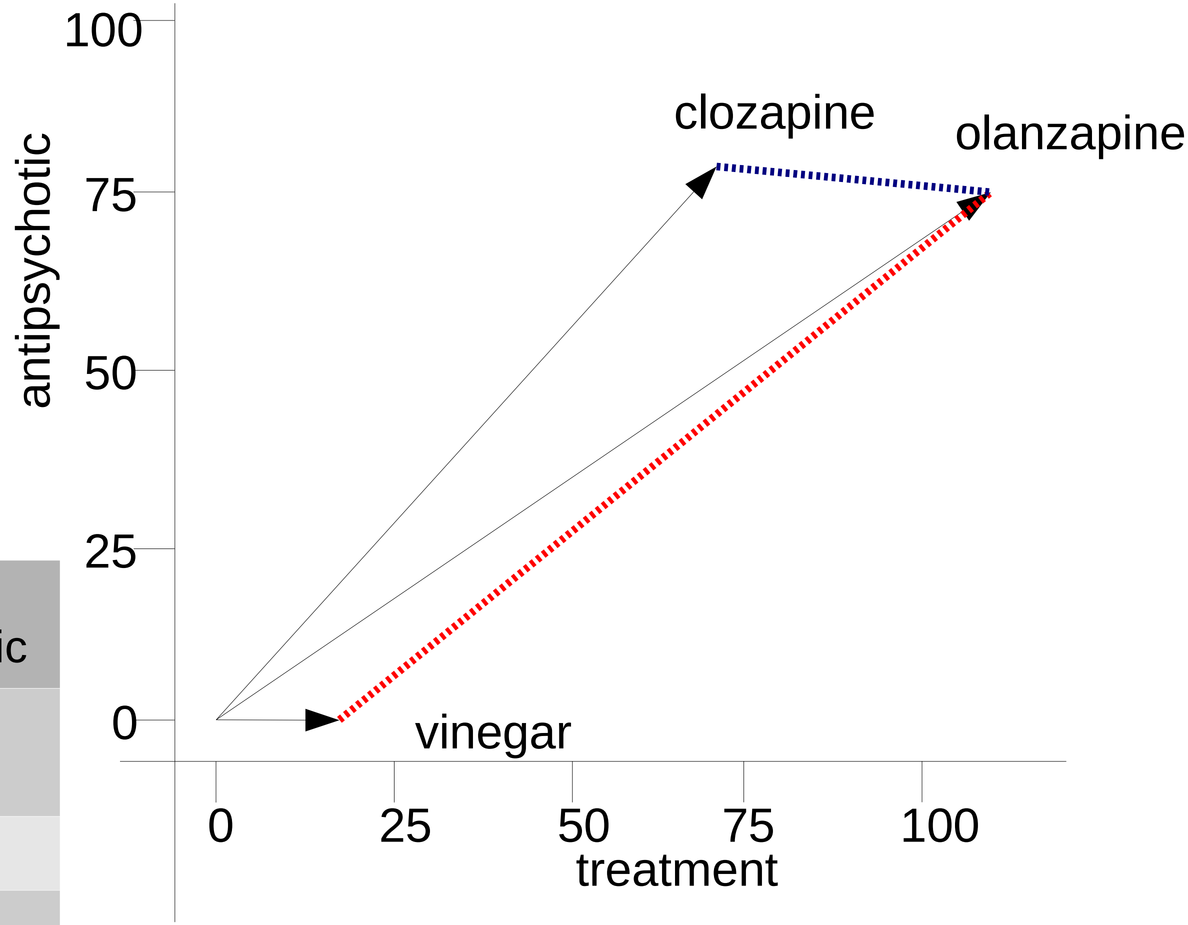
- As well as being necessary to represent our words numerically, it is also a step along the path of finding some abstraction of word meaning
- Alternatively, we could encode as the integer position in the index

aardvark	0
abacus	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?

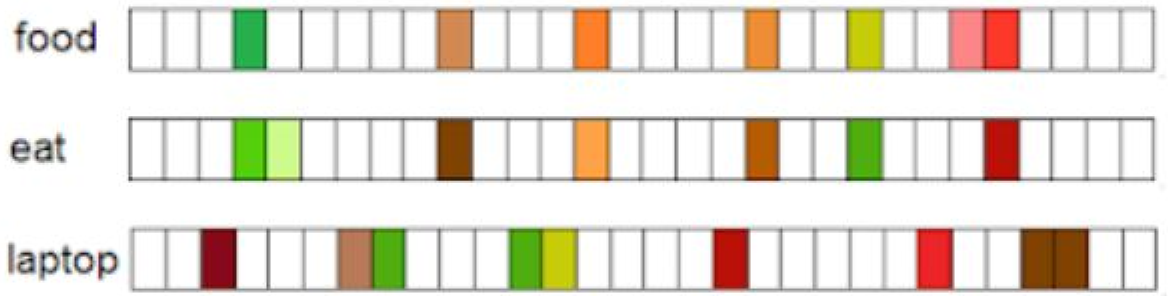
Semantic spaces



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

Encoding meaning

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



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

NIHR | Maudsley Biomedical Research Centre

Word embeddings: intuition

Intuition

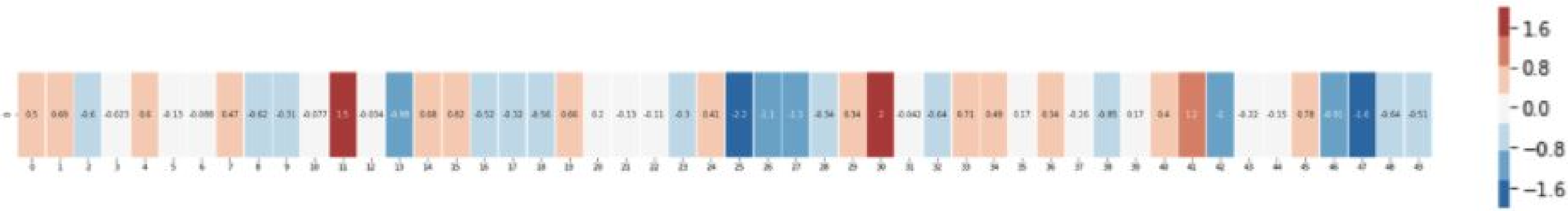
Construct a vector for the word “king”, (GloVe based vector, trained on Wikipedia):

```
[ 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/

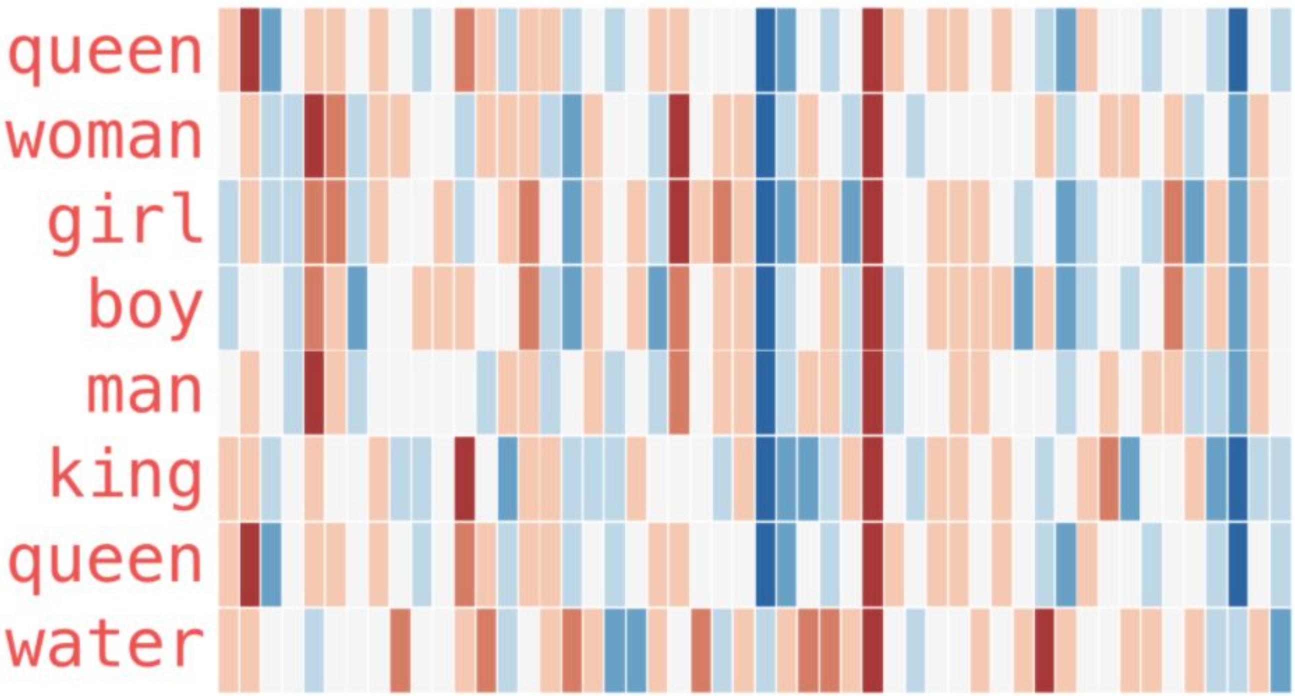
Intuition

Visualise as bands of different colours and intensities:



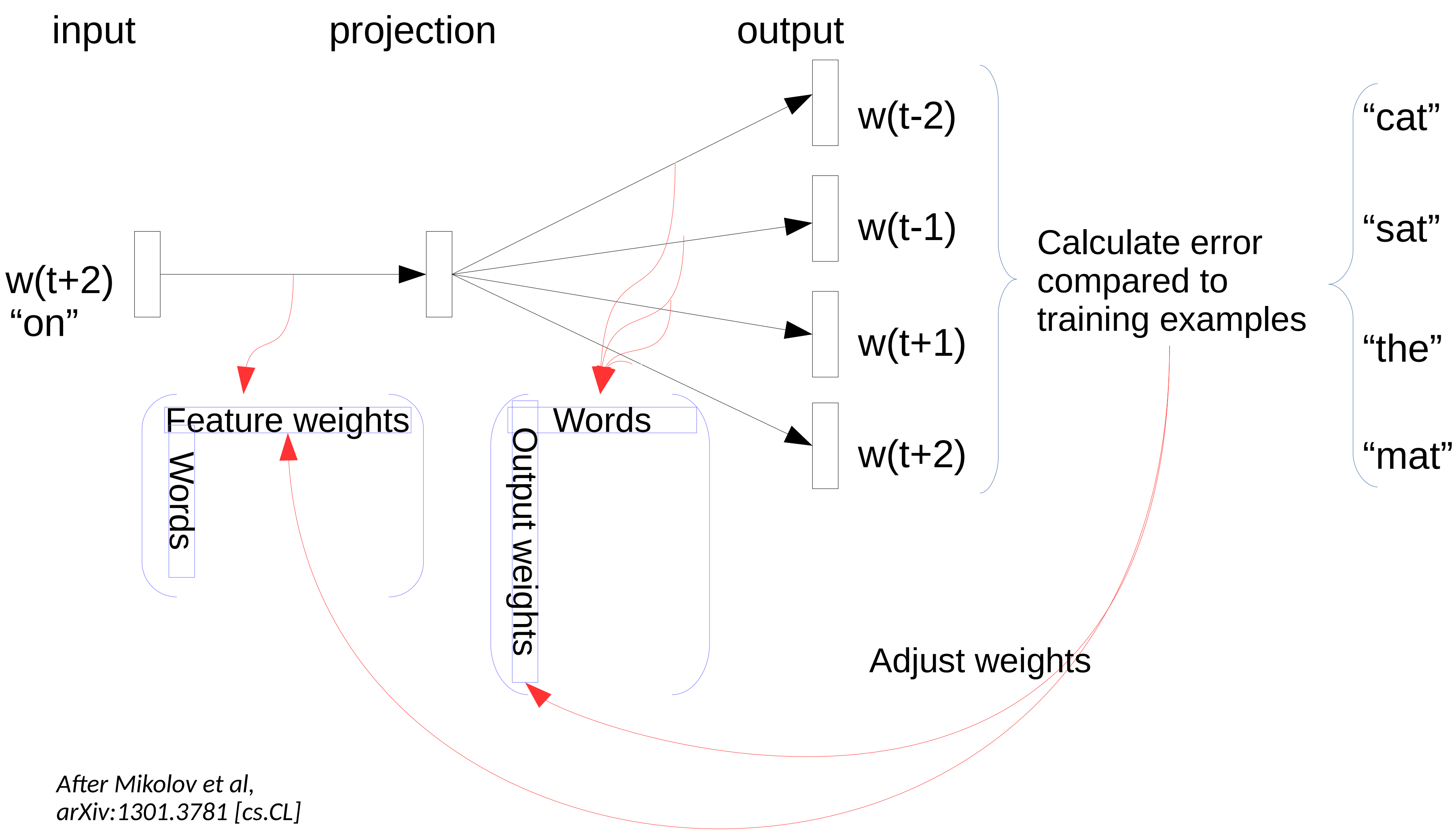
Intuition

Compare to vectors for other words:



Word embeddings: calculation

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

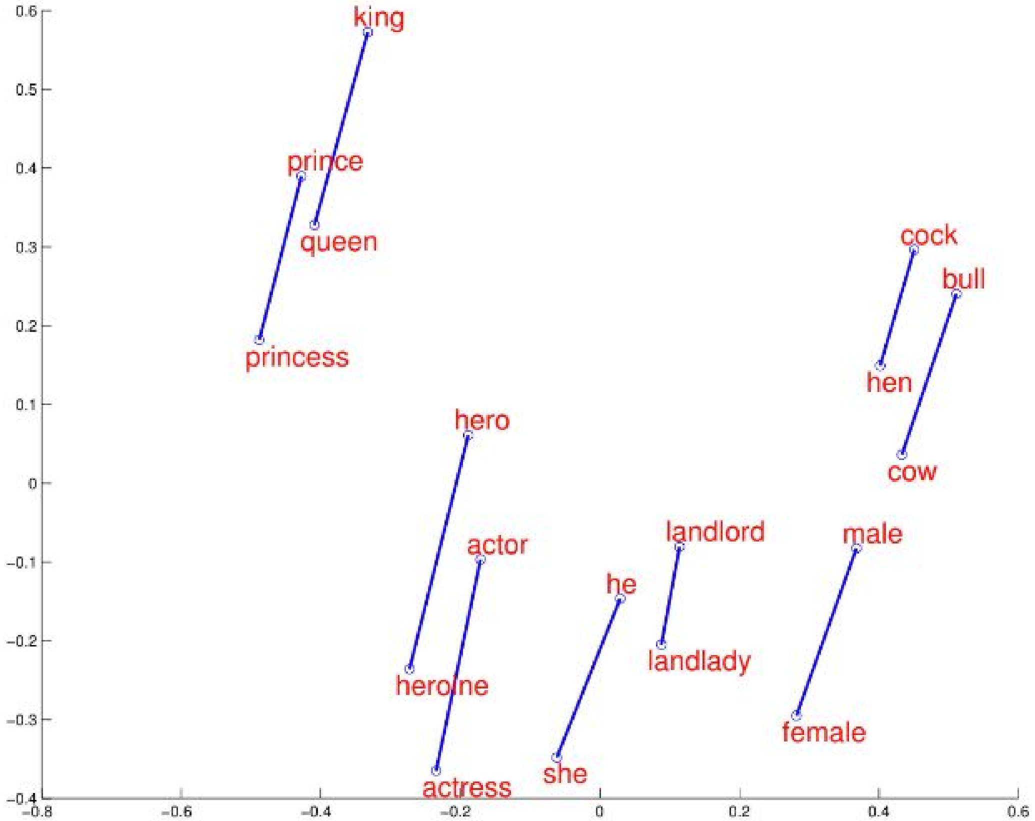
Intuition

- 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

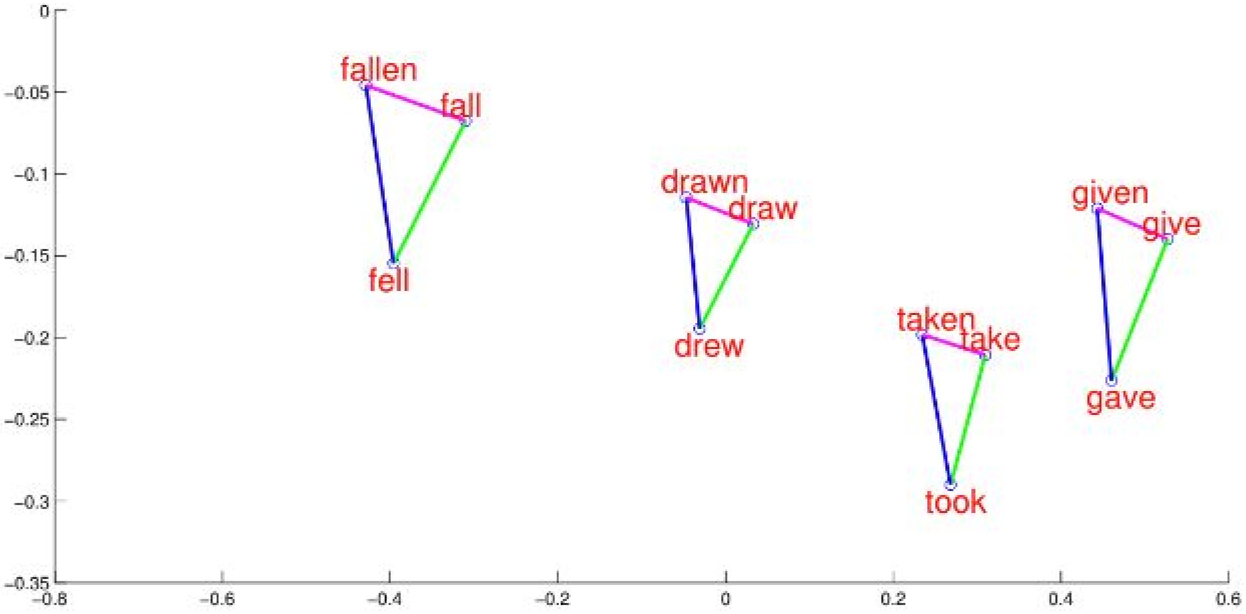
- If two words have similar contexts, then their feature vectors will be similar
- The final feature vector for a word gives a distributed representation of the word – ***word embeddings*** – a dimensionality reduction from our word space to real number vectors
- (We throw away the output vectors – we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

Visualisation



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

Visualisation



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

What about ambiguous words?

- What about homonyms and polysemous words?
- Word embeddings such as Word2Vec represent all senses of the word in a single vector
- It is unable to represent them independently (though there are work arounds)
- The key problem is again context
  - Word embeddings model words based on their context
  - But the final vector is applied independent of the context in which the word appears

Next steps: modelling language with artificial neural nets

2010 onwards:
artificial neural networks

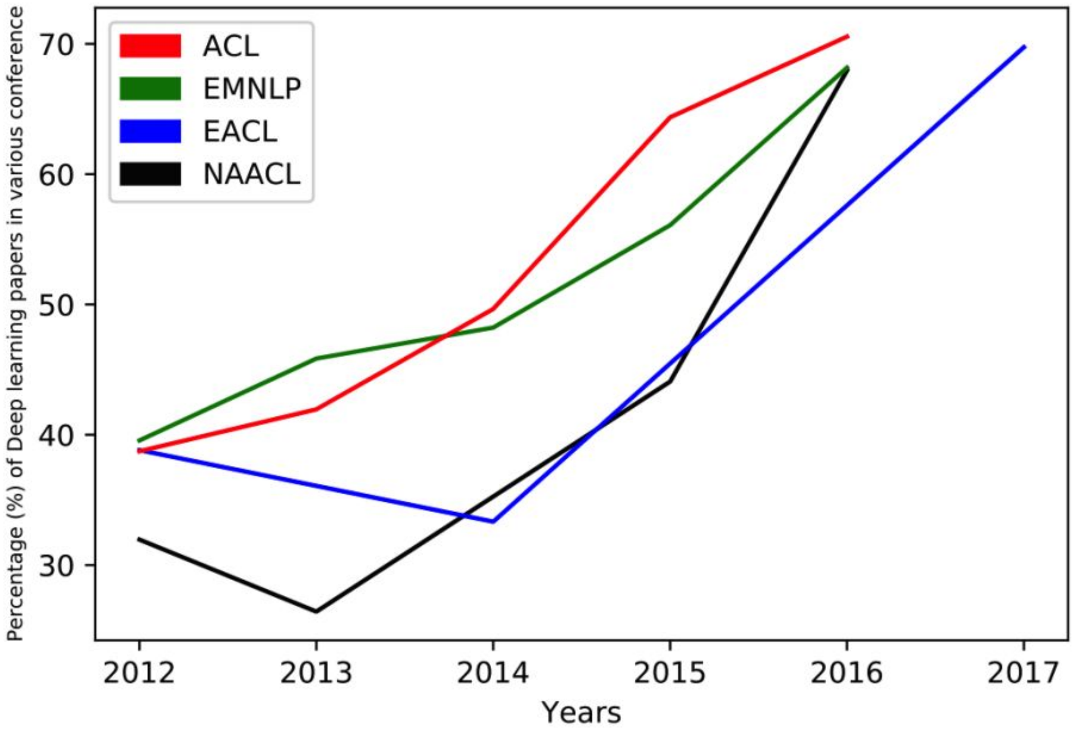
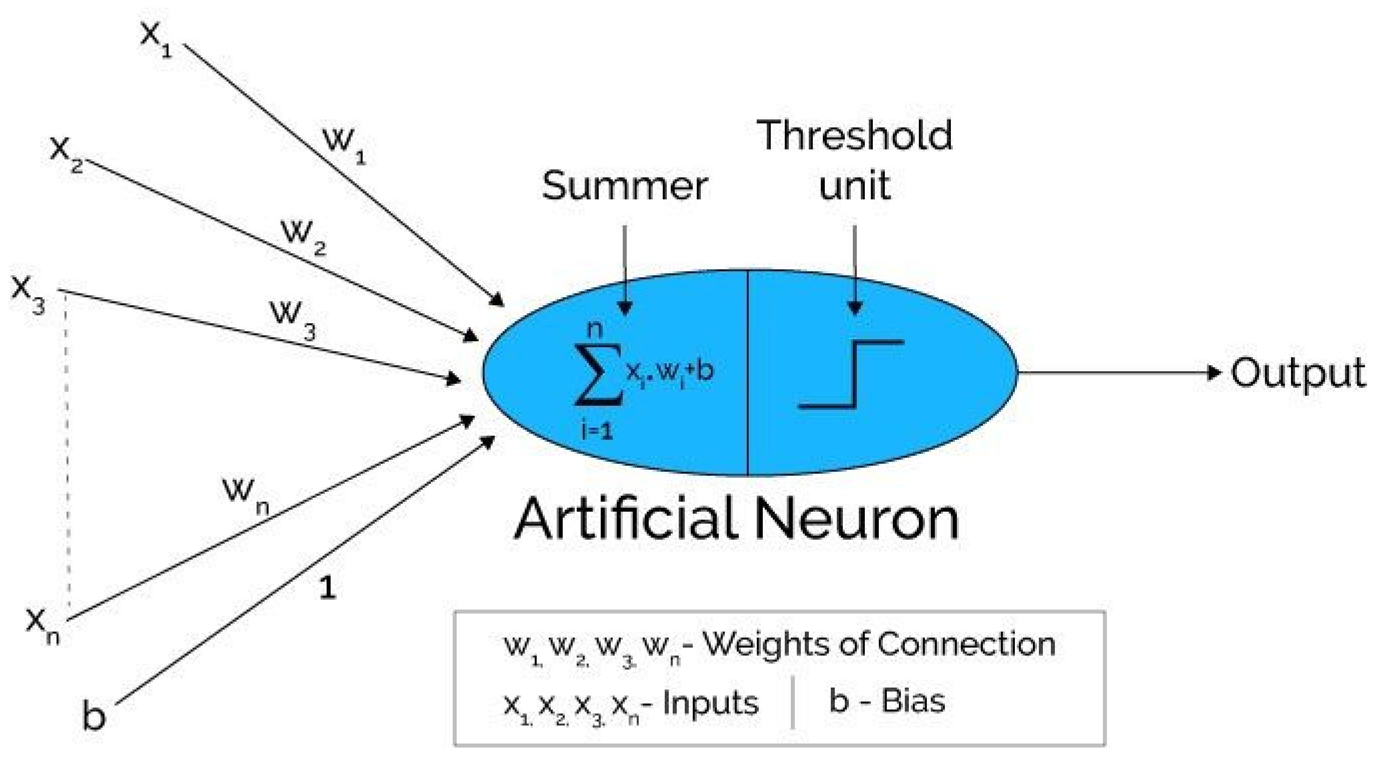


Fig. 1: Percentage of deep learning papers in ACL, EMNLP, EACL, NAACL over the last 6 years (long papers).

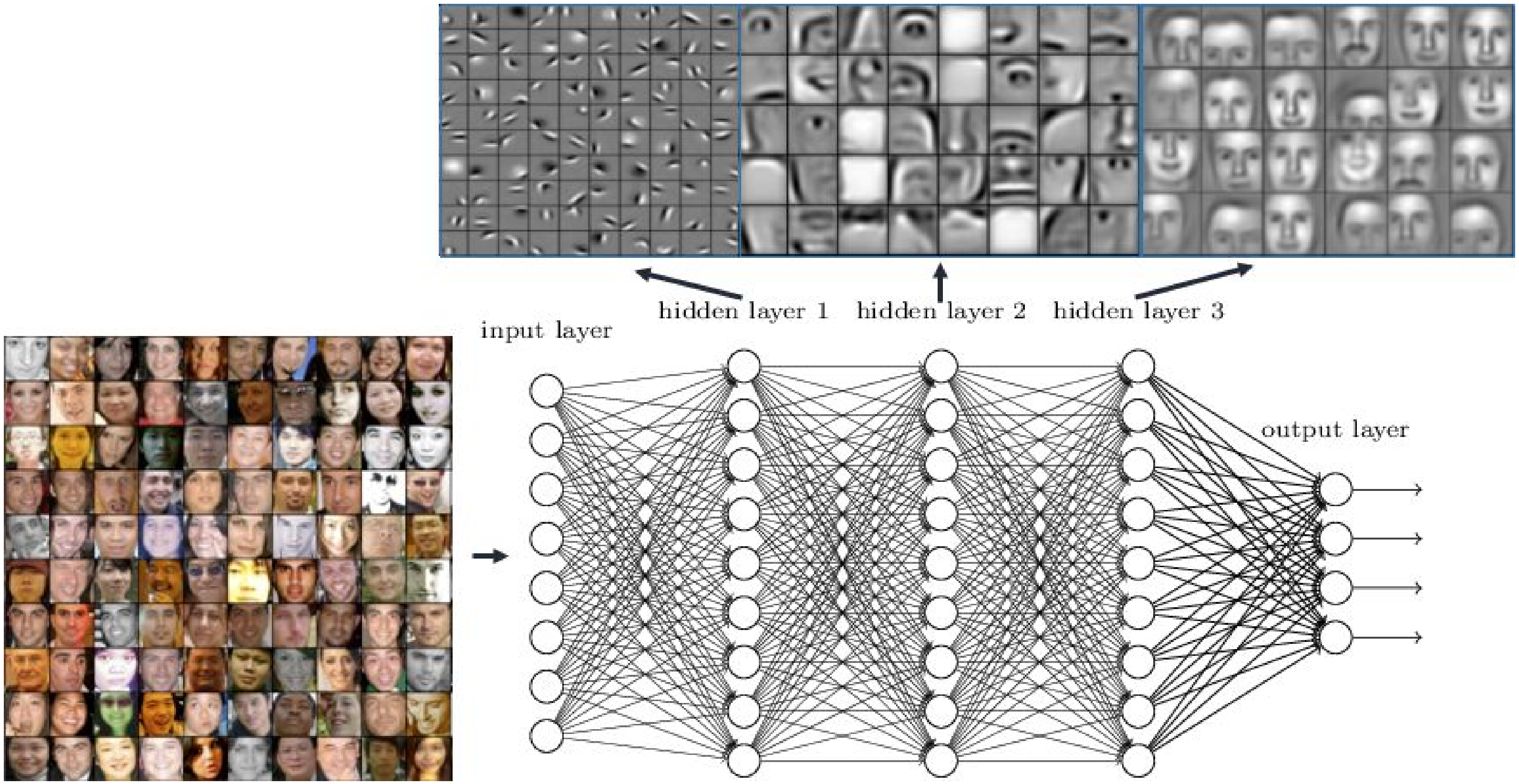
Young et al,
arXiv:1708.02709 [cs.CL]

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

NIHR | Maudsley Biomedical Research Centre

Thank you.
Any questions?

angus.roberts@kcl.ac.uk