Word embeddings

KING'S College LONDON

NIHR | Maudsley Biomedical Research Centre

Representing a word as a vector of real numbers

- Can we improve on our simple word vectors?
- Can we use such a model to encode meaning?

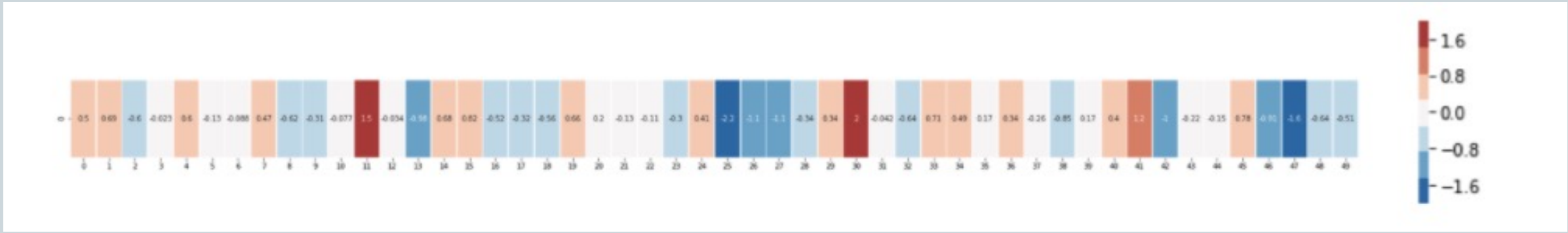
- Imagine this vector for the word “king”, (GloVe based vector, trained on Wikipedia):

```
[ 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/

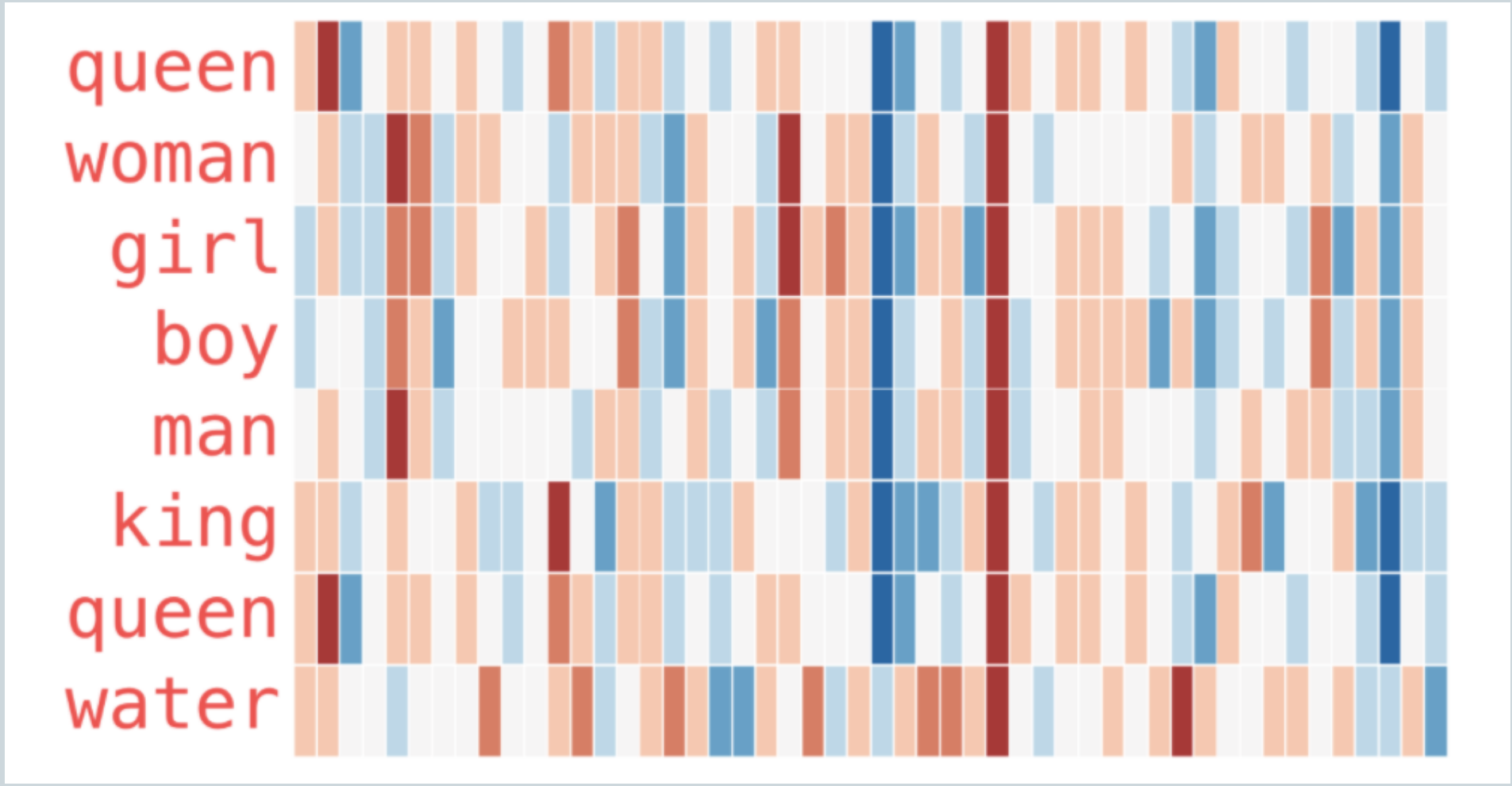
Visualising a vector

We can visualise this as bands of different colours and intensities:



Comparing vectors

Compare this word vector to vectors for other words:



Learning the vectors

- We start with a large corpus of language – a collection of texts
- We initialise our model with a random vector of weights for each word in the corpus
- We show lots of examples from the corpus of words surrounded by their context to the model

- E.g. for “on” we might have this example and word vectors

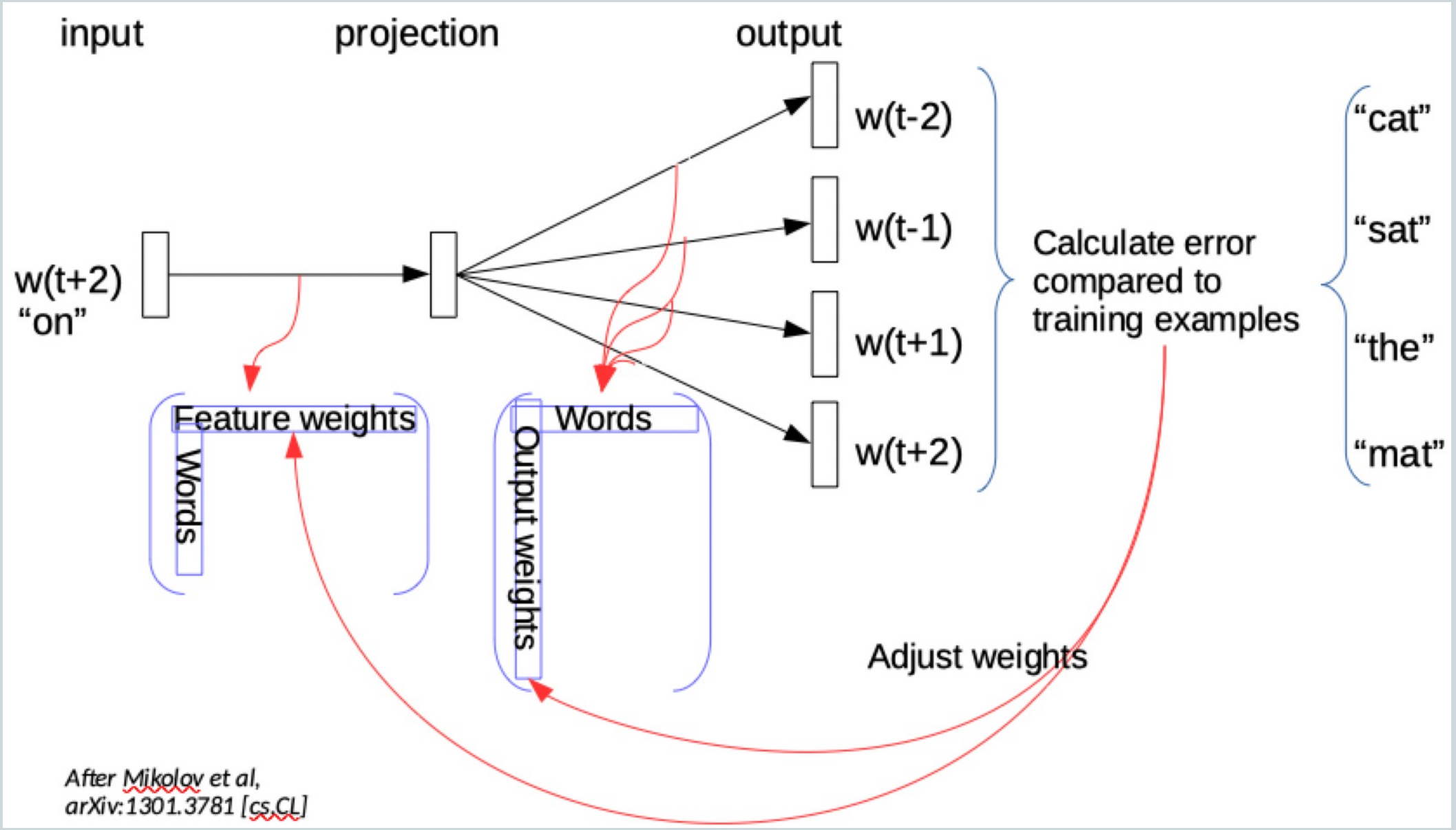
  - cat sat on the mat
  - c1 c2 w c3 c4
  - The model predicts the probability of seeing this example
  - We compare this probability to what we really see in our corpus
  - Adjust vector weights slightly to make the model prediction higher: maximise the probability of the example

- We also present examples where our word has been replaced by a random word

- E.g. we might replace “on” with “strawberry” (chosen at random)

  - cat sat strawberry the mat
  - c1 c2 w’ c3 c4
  - Adjust vector weights to make this model prediction lower: minimise the probability of random replacements

Learning the vectors



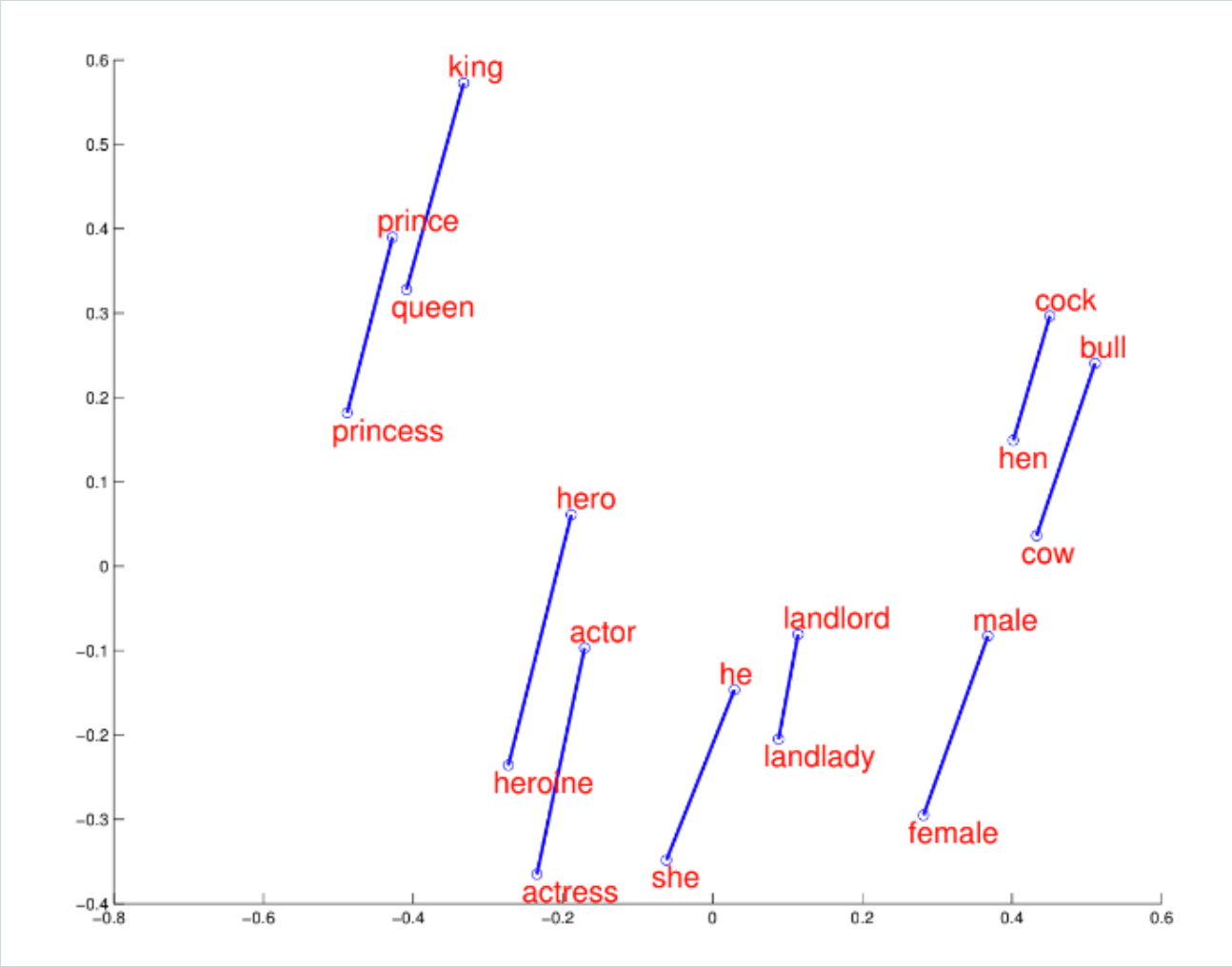
Intuition

- 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

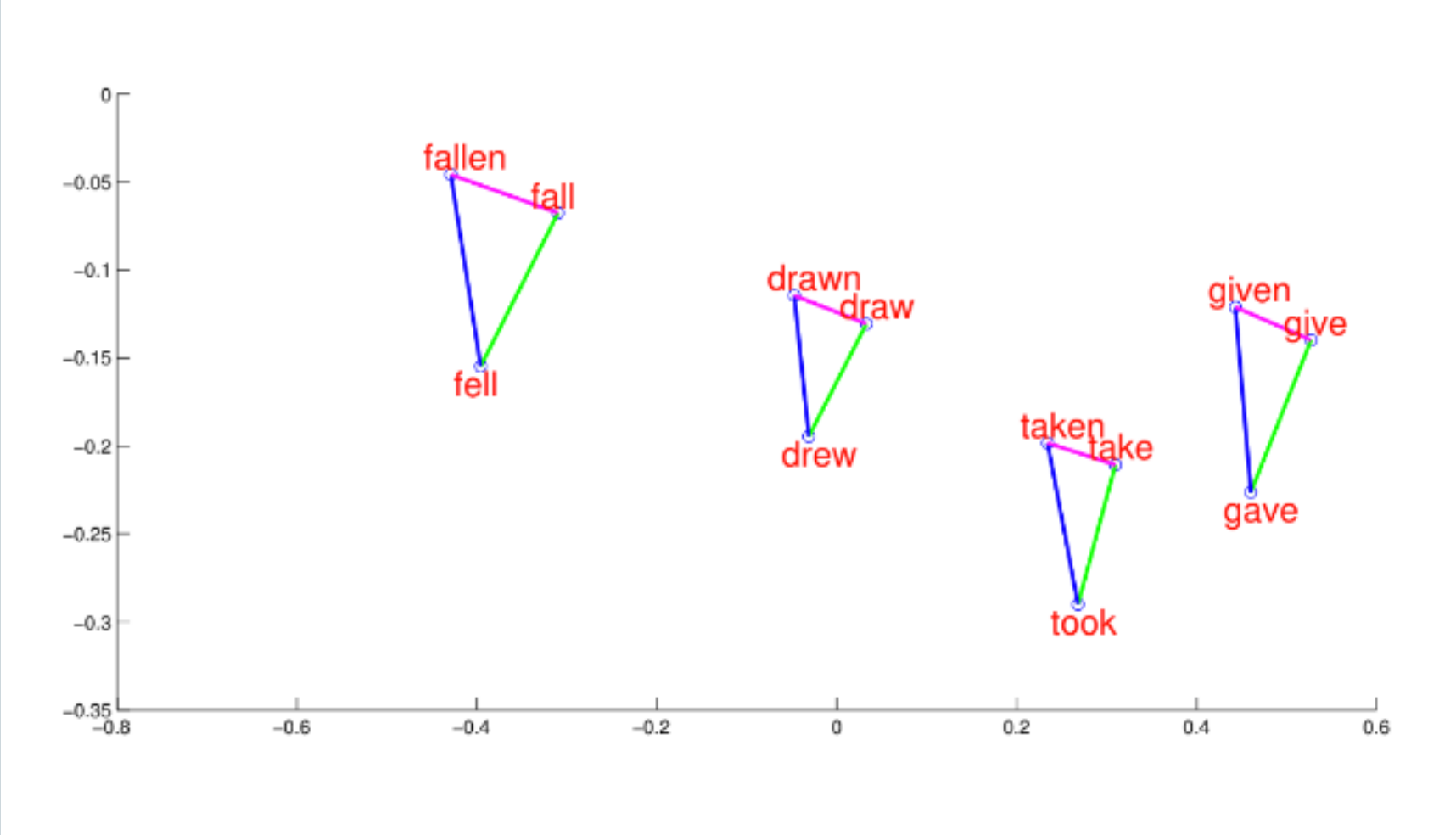
- If two words have similar contexts, then their feature vectors will be similar
- The final feature vector for a word gives a distributed representation of the word – ***word embeddings*** – a dimensionality reduction from our word space to real number vectors
- We use these word embedding as features 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

Thank you

angus.roberts@kcl.ac.uk
https://www.kcl.ac.uk/people/angus-roberts