



WORD EMBEDDINGS

Sandeep Soni

10/16/2023

What is the meaning of the word tezgüino?

A bottle of

is on the table

Everybody

likes

Don't have

before you drive

We make

out of corn

Example taken from Eisenstein, 2018, which was in turn taken by Lin, 1998

Is tezgüino similar to loud?

A bottle of _____

is on the table

Everybody _____

likes _____

Don't have _____

before you drive

We make _____

out of corn

Example taken from Eisenstein, 2018, which was in turn taken by Lin, 1998

Is tezgüino similar to oil?

A bottle of

is on the table

Everybody

likes

Don't have

before you drive

We make

out of corn

Example taken from Eisenstein, 2018, which was in turn taken by Lin, 1998

Is tezgüino similar to tortilla?

A bottle of

is on the table

Everybody

likes

Don't have

before you drive

We make

out of corn

Example taken from Eisenstein, 2018, which was in turn taken by Lin, 1998

Is tezgüino similar to beer?

Contexts

A bottle of

is on the table

Everybody

likes

Don't have

before you drive

We make

out of corn

Example taken from Eisenstein, 2018, which was in turn taken by Lin, 1998

QUESTION FOR THE DAY

“How do we represent word meaning?”

SEMANTICS

- Linguistic units can be described from at least three perspectives
 - Relational
 - Compositional
 - Distributional

SEMANTICS

Relational

Compositional

Distributional

John interviews Mark. He is a great tennis player

compose → composition → compositional

The paint is still wet

Paint that wall red

The old paint is coming off

“a word is characterized by the company it keeps”

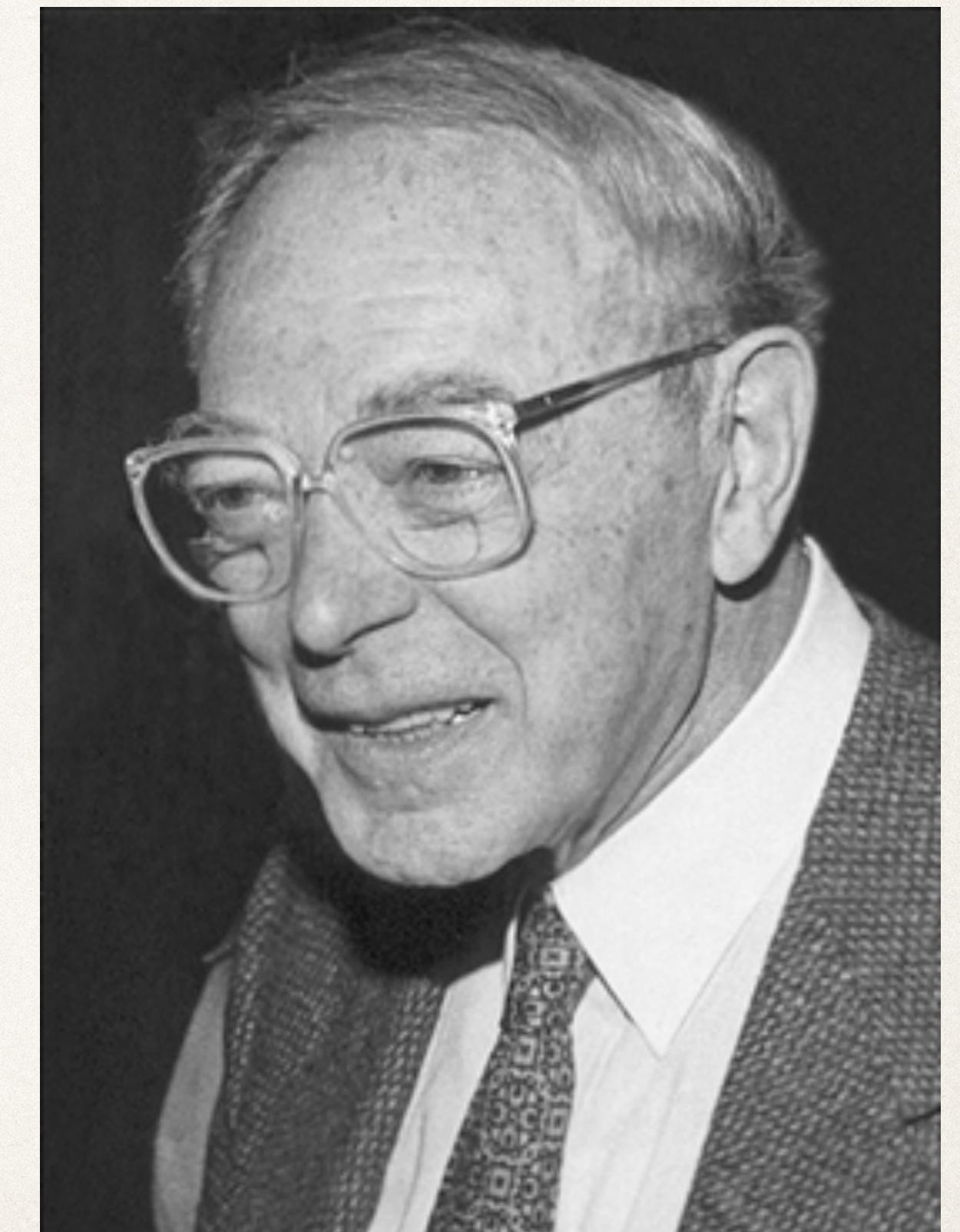
-J.R.Firth

DISTRIBUTIONAL SEMANTICS

- Distributional semantics allows us to learn the meanings from the contexts in which the words appear
- The ability to learn meanings in this way has been instrumental in recent progress



J.R.Firth



Zellig Harris

DISTRIBUTED REPRESENTATIONS

- Vector representations that summarize the distribution of the contexts a word appears in
- Words that appear in similar contexts should have similar representations (distributional hypothesis)
- Key questions:
 - What type of contexts?
 - What type of vectors?
 - How to systematically do this?

TERM-DOCUMENT MATRIX

Reviews

	Reviews							
wine		3		2		4		
cold	2			1				
beer		5		1	3			
service			4		4			

Every term is a vector encoding the distribution over documents

Context=entire document

VECTORS

Reviews

	1	2	3	4	5	6	7	8
wine			3			2		4
cold	2				1			
beer			5			1	3	
service				4			4	

The vectors are sparse

Their size is the number of documents

VECTOR SIMILARITY

Reviews

wine		3		2		4	
cold	2			1			
beer		5		1	3		
service			4		4		

Cosine similarity between the vectors for wine and cold is less than the cosine similarity between the vectors for wine and beer

COOCCURENCE MATRIX

- Instead of taking the entire document as context, we can define a context window of some size
- We can construct a matrix with rows as terms, columns as the contexts, and values as the number of times the words appear in contexts

Dataset

The old wine **tastes** good

The bottled beer is stale

The red wine **is** stale

Store the beer for long

Words in the context
window of size 2 for wine:

[the, old, tastes, good, red,
is, stale]

Dataset

The old wine tastes good

The bottled beer **is** stale

The red wine is stale

Store the beer **for** long

Words in the context
window of size 2 for beer:

[the, bottled, is, stale,
store, for, long]

COCCURRENCE MATRIX

Contexts

	the	bottled	tastes	good	...	stale	long
wine	2		1	1		1	
beer	2	1				1	

Every term is a vector encoding the distribution over other terms

Context=window around the term

VECTORS

Contexts (c)

the bottled tastes good ... stale long

Terms (t)

wine	2		1	1		1	
beer	2	1				1	

Vectors are
sparse

Size=Terms in
vocabulary

WEIGHTING

- Words typically cooccur with other functional words such as “the”, “a”, etc
- So can you weigh the dimensions based on their informativeness?

TFIDF

- The contexts for a term are weighted higher if they are frequent and specific
- $tfidf(t, c) = tf(t, c) \times \frac{N}{d_t}$
- N is the total number of contexts, $tf(t, c)$ is the number of times t co-appears with c, d_t is the number of times t appears in any context

PMI

- The contexts for a term are weighted higher if they cooccur more than they would if they had no relation
- $PMI(t, c) = \log_2 \frac{P(t, c)}{P(t) \cdot P(c)}$
- $P(t, c)$ is the probability of cooccurrence; $P(t)$ and $P(c)$ is probability of independent occurrence
- $PPMI(t, c) = \max(0, \log_2 \frac{P(t, c)}{P(t) \cdot P(c)})$

DENSE REPRESENTATIONS

- Ideally, we want to learn characteristic dimensions of a word, instead of distributions over all words/contexts
- We learn a low-dimensional but compact/dense vectors by formulating the cooccurrence events as prediction tasks

Word	Ball game	Non-square arena	Olympic Game	Indoor
Soccer	1	0	1	0
Javelin Throw	0	1	1	0
Squash	1	0	0	1
Chess	0	1	0	1

WORD2VEC

- Skipgram (Mikolov et. al. 2013) predicts a context word given a target word
- Think of this as a huge multi-class classification task (classes are words) for every word

x	y
wine	cold
wine	sweet
wine	spirit
wine	drink

WORD2VEC

- We can estimate the empirical probability as
 $P(c|w) \propto \exp(\mathbf{c} \cdot \mathbf{w})$
- \mathbf{c} is a vector representation of a word when it appears in the context; \mathbf{w} is a vector representation of a word when it appears in the input

x	y
wine	cold
wine	sweet
wine	spirit
wine	drink

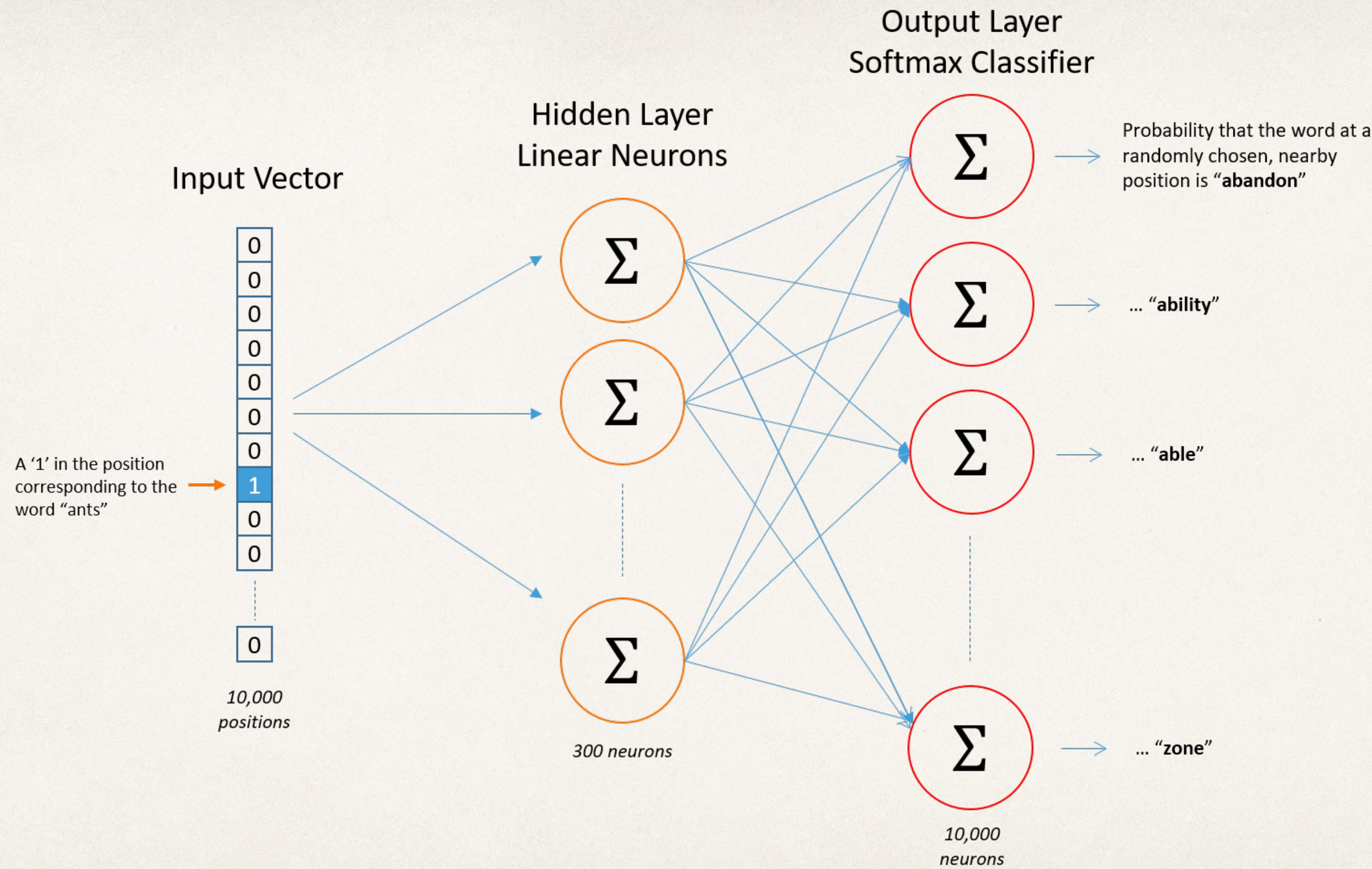
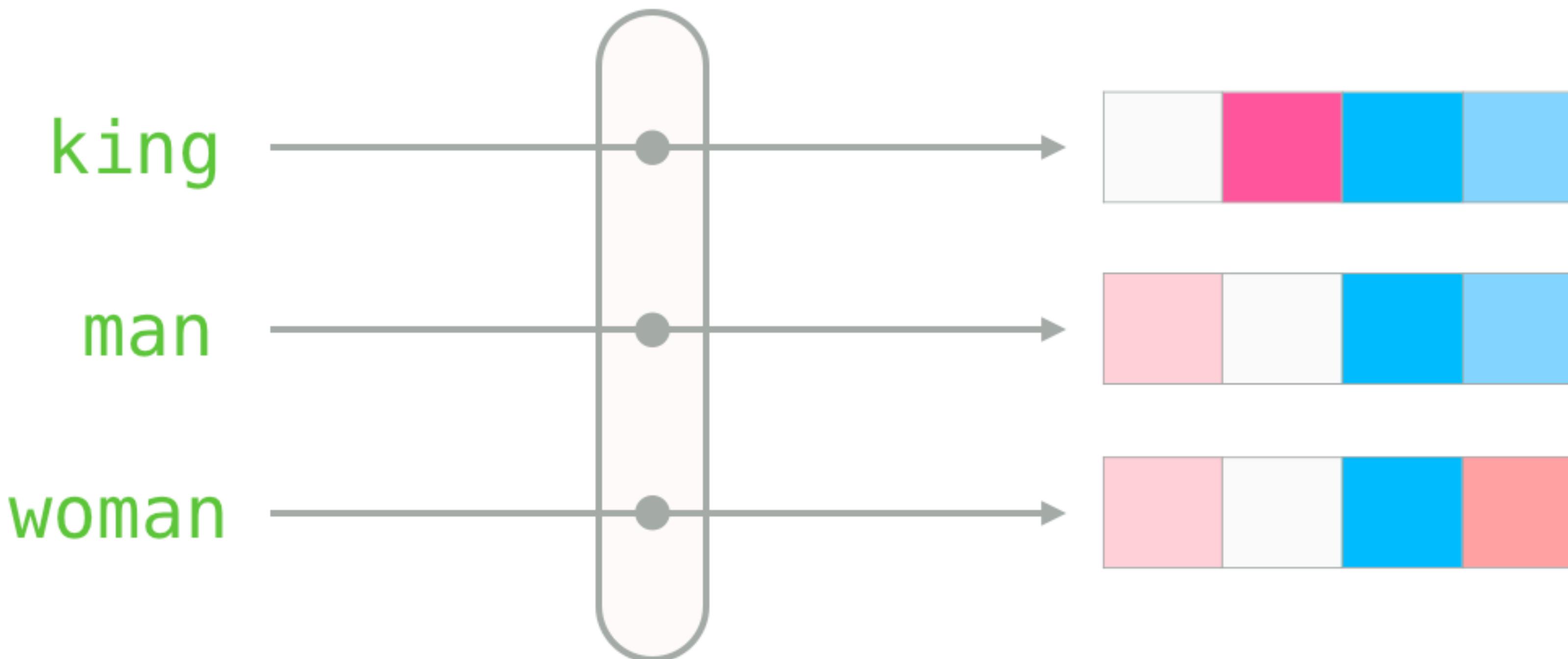


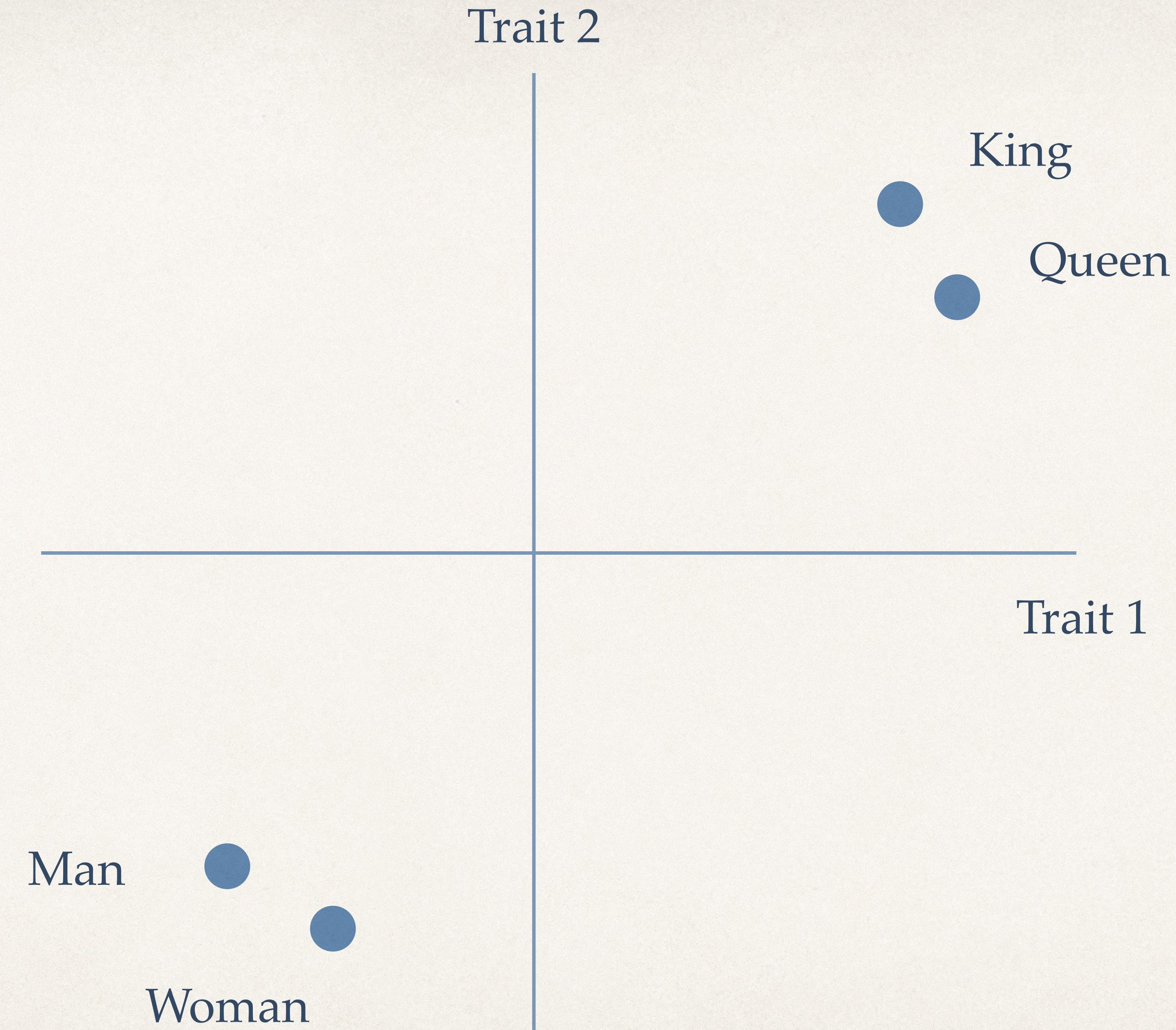
Image courtesy of Chris McCormick

Word2vec



<http://jalammar.github.io/illustrated-word2vec/>

Trait 2



GLOVE

- Pennington et. al. 2014 used a similar idea but directly model some statistic of the cooccurrence
- Think of this as a regression task

x	y	statistic
wine	cold	3
wine	sweet	5
wine	spirit	1
wine	drink	2

EVALUATION

EVALUATION

- How do we know if the embeddings we learned are any good?

EVALUATION

EVALUATION

- Intrinsic evaluation
 - Word relatedness: the similarity between vector representations should correlate with human judgments of relatedness of pairs of words
 - Analogical reasoning: King:Queen::Man:?
- Extrinsic evaluation
 - Plug in the embeddings as features in some downstream task

IN CLASS

- Word2Vec demo