Natural Language Processing

04: Vector Semantics and Word Embeddings

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2021-04-30

Technology
Arts Sciences
TH Köln

Lexical semantics

"You shall know a word by the company it keeps"

(Firth, 1957)

(b) The fact that, for example, not every adjective occurs with every noun can be used as a measure of meaning difference. For it is not merely that different members of the one class have different selections of members of the other class with which they are actually found. More than that: if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

The distribution of an element will be understood as the sum of all its environments. An environment of an element A is an existing array of its co-occurrents, i.e. the other elements, each in a particular position, with which A occurs to yield an utterance. A's co-occurrents in a particular position are called its selection for that position.

Zellig S. Harris (1954) Distributional Structure, WORD, 10:2-3, 146-162, DOI: 10.1080/00437956.1954.11659520

Company of words = context

- A few different ways we can encode the notion of “company” (or context).

context

everyone likes	________________	
a bottle of	________________	is on the table
you can drink	________________	if you’re over 21
a cocktail with	________________	and seltzer

Distributed representation

- Vector representation that encodes information about the distribution of contexts a word appears in
- Words that appear in similar contexts have similar representations (and similar meanings, by the distributional hypothesis).

Term-document matrix

	Hamlet	Macbeth	Romeo & Juliet	Richard III	Julius Caesar	Tempest	Othello	King Lear
knife	1	1	4	2		2		2
dog	2		6	6		2		12
sword	17	2	7	12		2		17
love	64		135	63		12		48
like	75	38	34	36	34	41	27	44

Context = appearing in the same document.

Vector representation of the document

- vector size = V

Vector representation of the term

- vector size = number of documents

Not all dimensions are equally informative …

TF-IDF

- Term frequency-inverse document frequency
- A scaling to represent a feature as function of how frequently it appears in a data point but accounting for its frequency in the overall collection
- IDF for a given term = the number of documents in collection / number of documents that contain term

Term-context matrix

- Rows and columns are both words
- Cell counts = the number of times word w_i and w_j show up in the same document.
- More common to define document = some smaller context (e.g., a window of 2 tokens)

Term-context matrix

Dataset

- the big dog ate dinner
- the small cat ate dinner
- the white dog ran down the street
- the yellow cat ran inside

DOG terms (window = 2)

- the big ate dinner the white ran down

CAT terms (window = 2)

- the small ate dinner the yellow ran inside

Term-context matrix

contexts

term

	the	big	ate	dinner	…
dog	2	1	1	1	…
cat	2	0	1	1	…

Each cell enumerates the number of time a context word appeared in a window of 2 words around the term.

Term-context matrix

contexts

term	L: the big	R: ate dinner	L: the small	L: the yellow	…
dog	1	1	0	0	…
cat	0	1	1	1	…

Each cell enumerates the number of time a directional context phrase appeared in a specific postion around the term.

Cosine Similarity

- We can calculate the cosine similarity of two vectors to judge the degree of their similarity [Salton 1971]
- Euclidean distance measures the magnitude of distance between two points
- Cosine similarity measures their orientation

Intrinsic evaluation

- Analogical reasoning (Mikolov et al. 2013)
- For analogy
  - Germany : Berlin :: France : ???,
    find closest vector to
    v("Berlin") - v("Germany") + v("France")

			target
possibly	impossibly	certain	uncertain
generating	genereated	shrinking	shrank
think	thinking	look	looking
Germany	Berlin	France	…

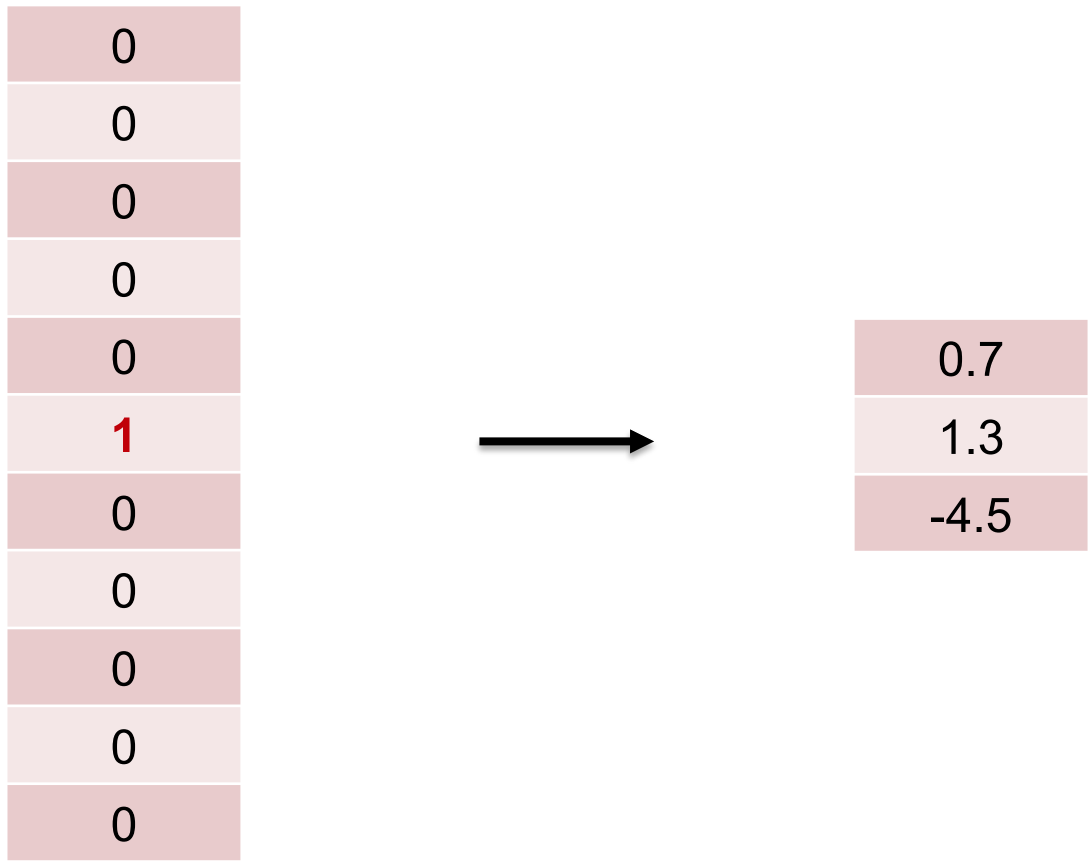
Sparse vectors

“aardvark”

A	0
a	0
aa	0
aal	0
aam	0
aadvark	1
…	0
Zyzomys	0

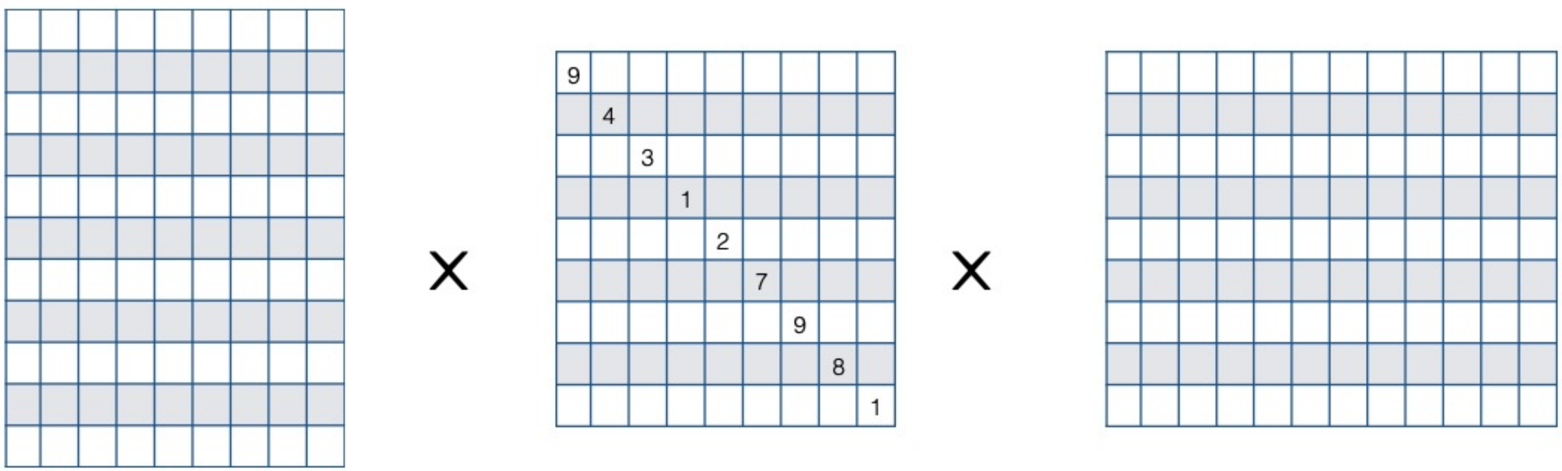
- V-dimensional vector, single 1 for the identity of the element
- “one hot” encoding

Dense vectors



Singular value decomposition

- Any $x \times p$ matrix X can be decomposed into the product of three matrices (where m = the number of linearly independent rows)



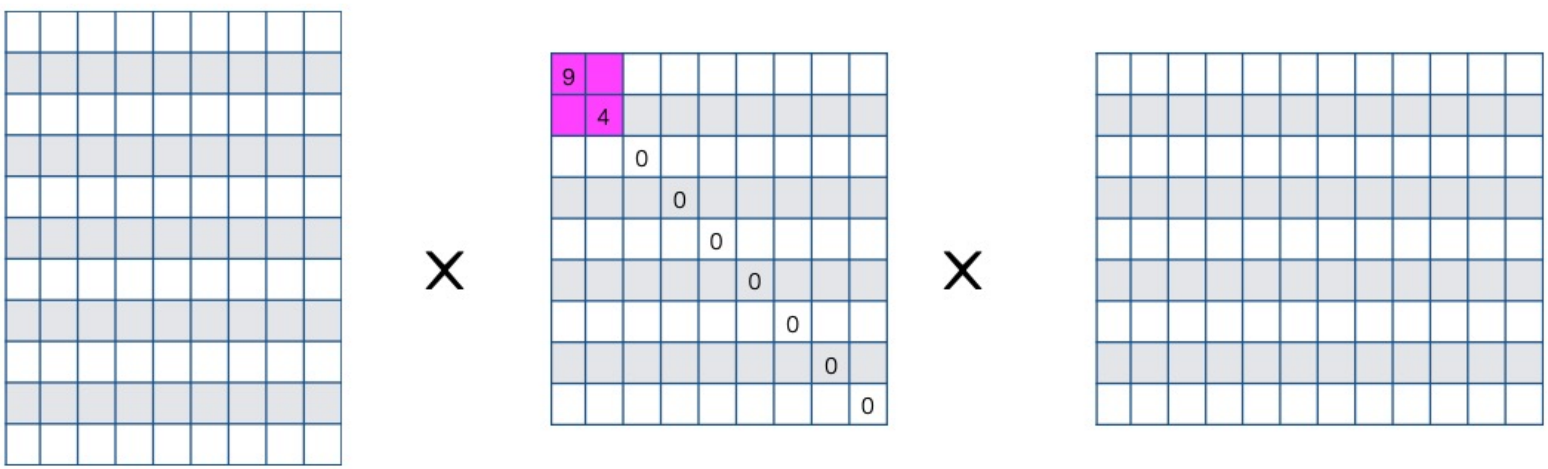
n x m

m x m
(diagonal)

m x p

Singular value decomposition

- We can <u>approximate</u> the full matrix by only considering the leftmost k terms in the diagonal matrix



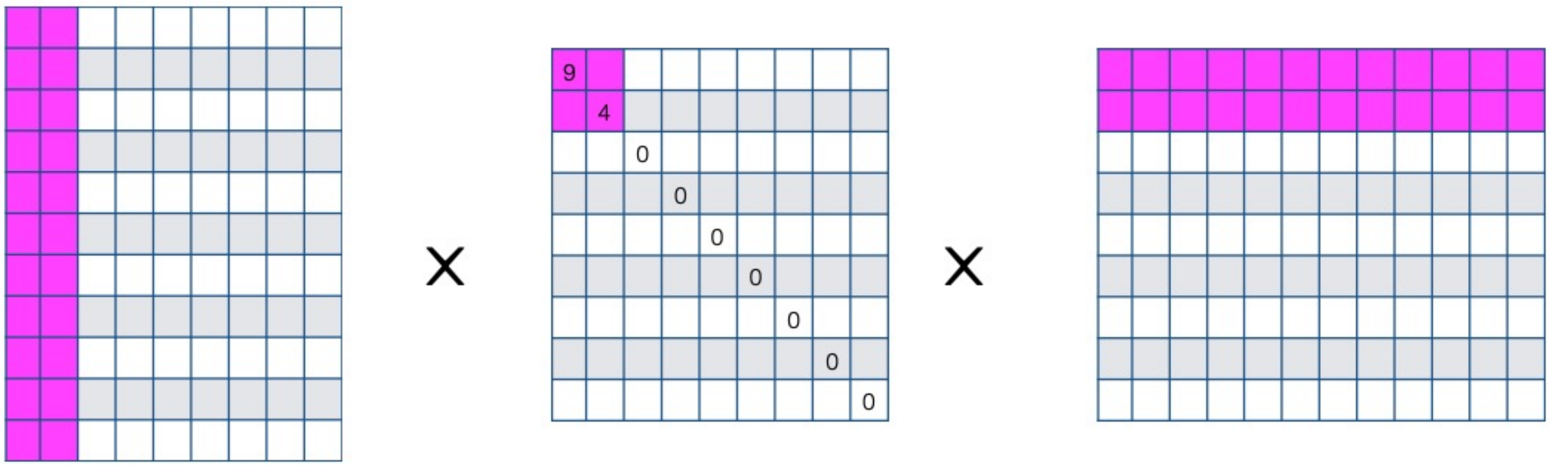
n x m

m x m
(diagonal)

m x p

Singular value decomposition

- We can approximate the full matrix by only considering the leftmost k terms in the diagonal matrix



n x m

×

m x m
(diagonal)

×

m x p

	Hamlet	Macbeth	Romeo & Juliet	Richard III	Julius Caesar	Tempest	Othello	King Lear
knife	1	1	4	2		2		2
dog	2		6	6		2		12
sword	17	2	7	12		2		17
love	64		135	63		12		48
like	75	38	34	36	34	41	27	44

knife		
dog		
sword		
love		
like		

Hamlet	Macbet h	Romeo & Juliet	Richard III	Julius Caesar	…

Low-dimensional representation for terms (here 2-dim)

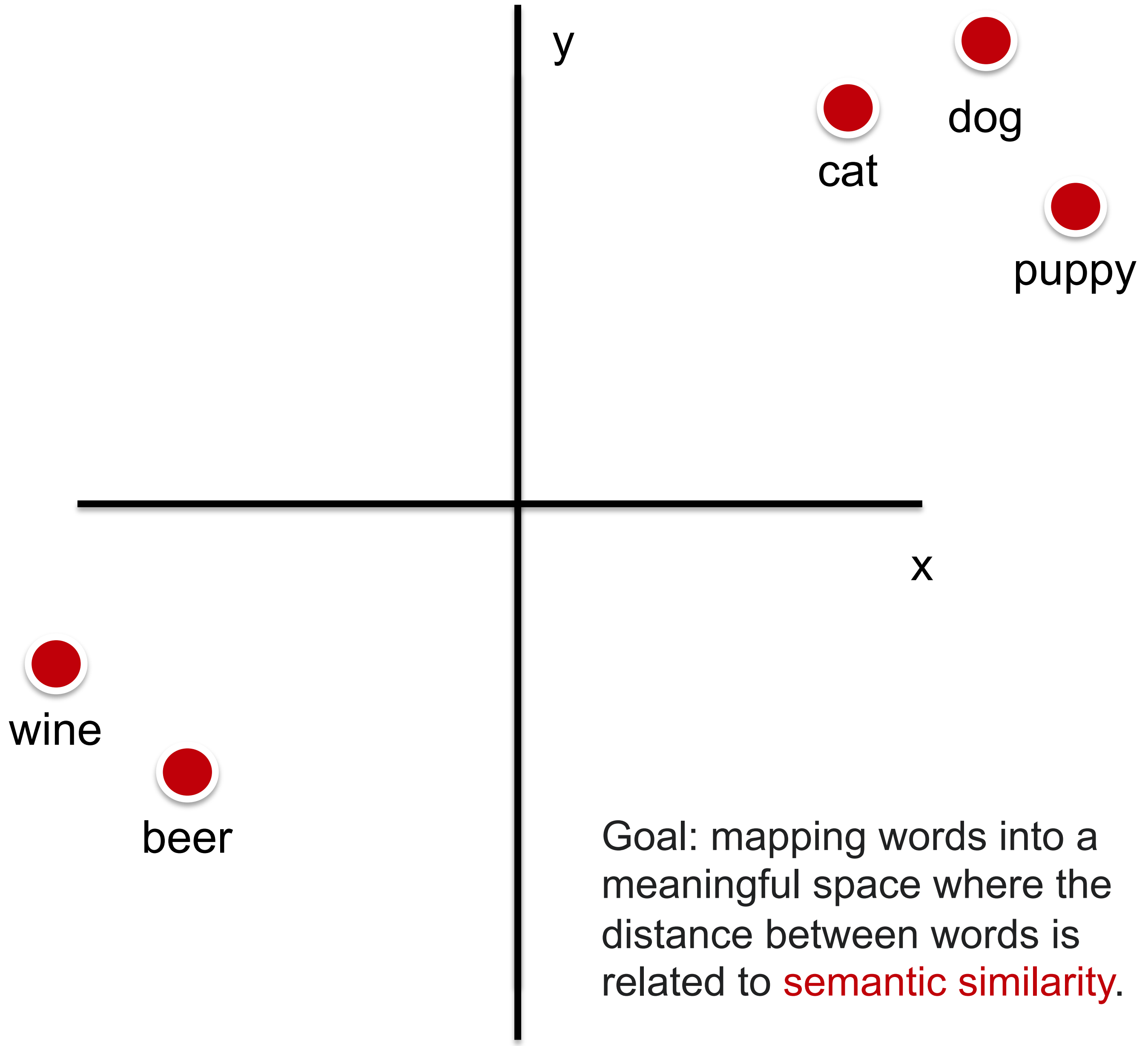
Low-dimensional representation for documents (here 2-dim)

knife		
dog		
sword		
love		
like		

Hamlet	Macbet h	Romeo & Juliet	Richard III	Julius Caesar	…

Latent semantic analysis

- Latent Semantic Analysis/Indexing (Deerwester et al. 1998) is this process of applying SVD to the term-document co-occurence matrix
- Terms typically weighted by tf-idf
- This is a form of dimensionality reduction (for terms, from a D-dimensionsal sparse vector to a K-dimensional dense one), K << D.



Goal: mapping words into a meaningful space where the distance between words is related to semantic similarity.

Similarity

People are good at generalizing newly acquired knowledge. If you learn a new fact about an object, your expectations about other similar objects tend to change. If, for example, you learn that chimpanzees like onions you will probably raise your estimate of the probability that gorillas like onions. In a network that uses distributed representations, this kind of generalization is automatic. The new knowledge about chimpanzees is incorporated by modifying some of the connection strengths so as to alter the causal effects of the distributed pattern of activity that represents chimpanzees.[2] The modifications automatically change the causal effects of all similar activity patterns. So if the representation of gorillas is a similar activity pattern over the same set of units, its causal effects will be changed in a similar way.

Hinton, G. E. (1984) Distributed Representations. Technical Report CMU-CS-84-157.

Dense vectors from prediction

- Learning low-dimensional representations of words by framing a predicting task: using context to predict words in a surrounding window
- Transform this into a supervised prediction problem

Dense vectors from prediction

- Skipgram model (Mikolov et al. 2013): given a single word in a sentence, predict the words in a context window around it.

- “a cocktail with gin and seltzer”

x	y
gin	a
gin	cocktail
gin	with
gin	and
gin	seltzer

windows size = 3

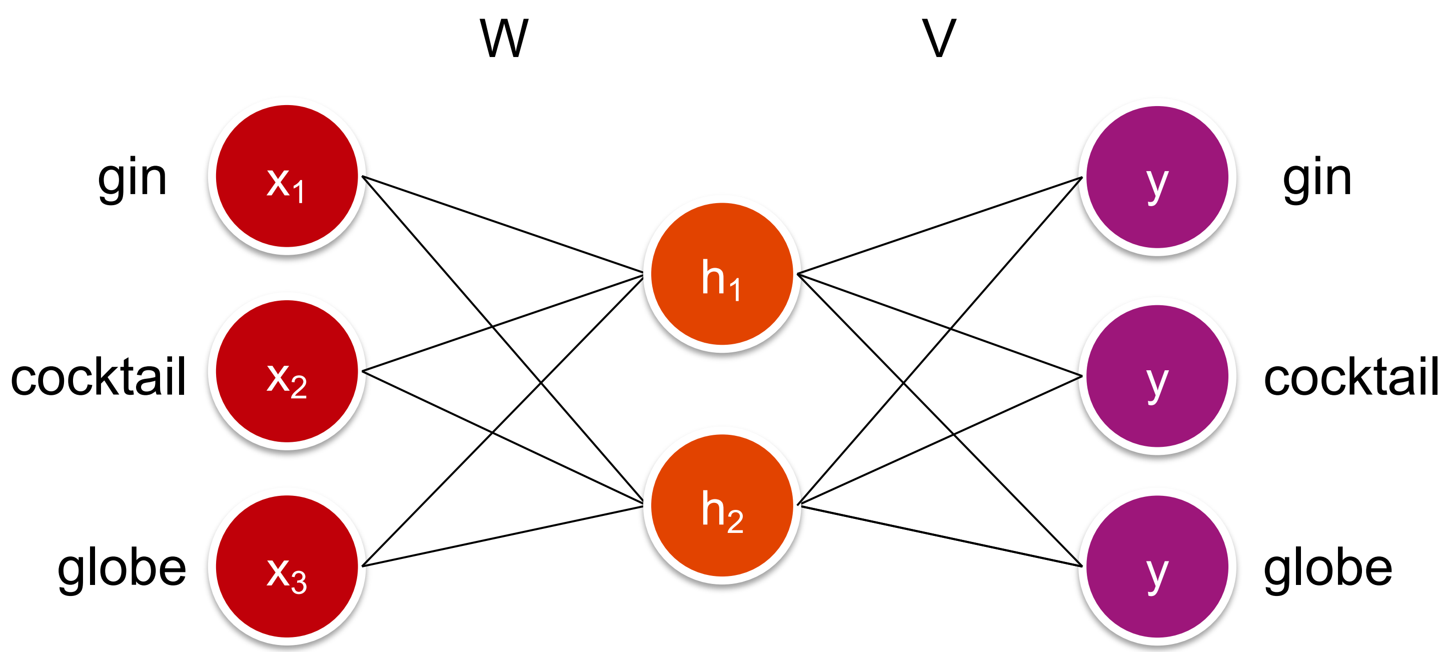
Dimensionality reduction

…	…
the	1
a	0
an	0
for	0
in	0
on	0
dog	0
cat	0
…	…

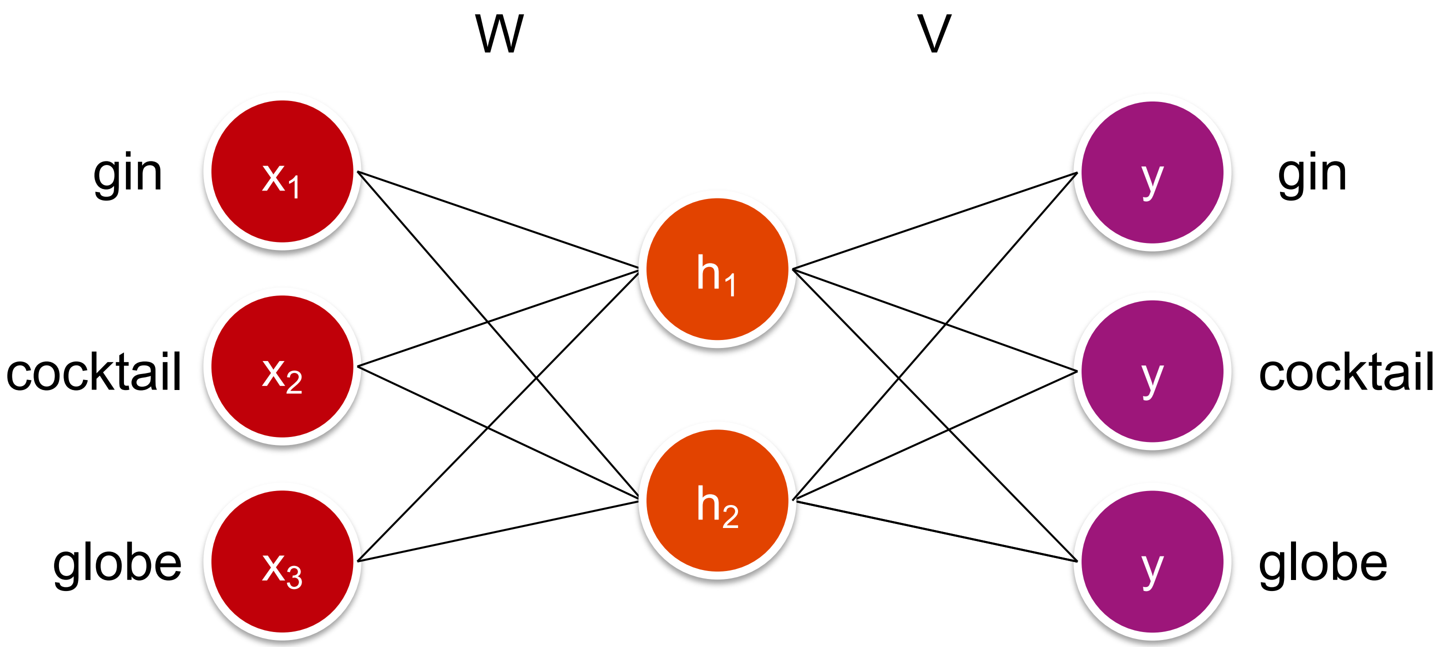
the is a point in
V-dimensional space

the
4.1
-0.9

the is a point in
2-dimensional space



	x	W (context)		V (word co-occurence)			y
gin	0	-0.5	1.3	4.1	0.7	0.1	1
cocktail	1	0.4	0.08	-0.9	1.3	0.3	0
globe	0	1.7	3.1				0



- Only one of the inputs is nonzero.

W	
-0.5	1.3
0.4	0.08
1.7	3.1

V		
4.1	0.7	0.1
-0.9	1.3	0.3

x	W	
	-0.80	.0.54
	-1.73	-1.57
	-0.05	0.24
	0.53	1.01
1	-0.02	2.21
	0.77	1.53
	2.14	-0.71
	-0.76	0.93
	-0.70	0.94

$x^T W =$

-0.02	2.21

This is the embedding of the context

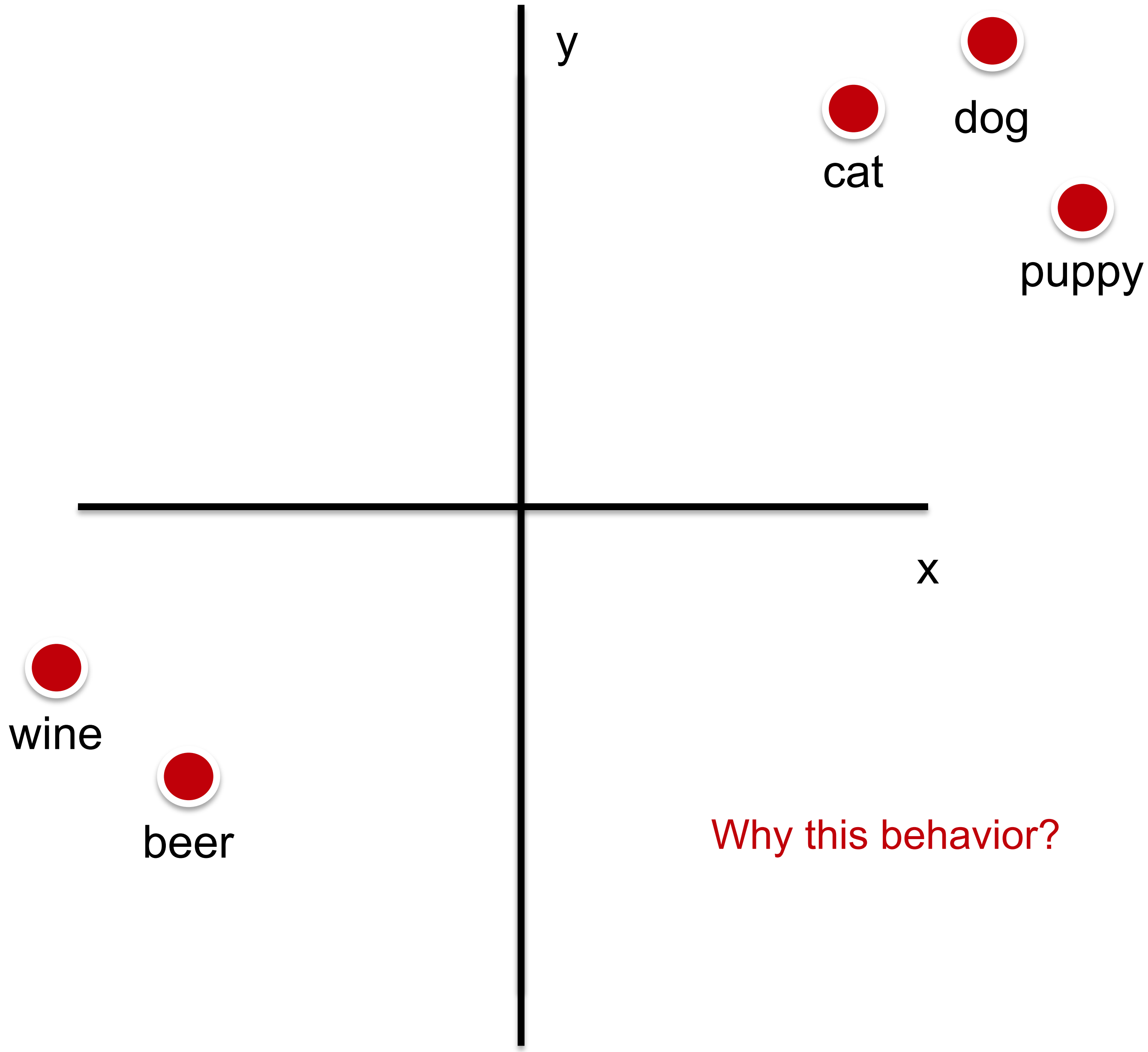
- Rather than seeing the input as a one-hot encoded vector specifying the word in the vocabulary we're conditioning on, we can see it as indexing into the appropriate row in the weight matrix W

Word embeddings

- Similarly, V has one vector for each element in the vocabulary (for the words that are being predicted)

V			
gin	cocktail	cat	globe
4.1	0.7	0.1	1.3
-0.9	1.3	0.3	-3.4

This is the embedding of the word



Why this behavior?

Word embeddings

- *dog*, *cat* show up in similar positions!

the black	cat	jumped on the table
the black	dog	jumped on the table
the black	puppy	jumped on the table
the black	skunk	jumped on the table
the black	shoe	jumped on the table

Word embeddings

- *dog*, *cat* show up in similar positions!

the	black	[0.4, 0.08]	jumped	on	the	table
the	black	[0.4, 0.07]	jumped	on	the	table
the	black	puppy	jumped	on	the	table
the	black	skunk	jumped	on	the	table
the	black	shoe	jumped	on	the	table

To make the same predictions, these
numbers need to be close to each other.

Analogical inference

- Mikolov et al. 2013 show that vector representations have some potential for analogical reasoning through vector arithmetic.

- apple – apples ≈ car – cars
- king – man + woman ≈ queen
- …

Mikolov et al., (2013), “Linguistic Regularities in Continuous Space Word Representations” (NAACL)

SHARE

REPORTS | PSYCHOLOGY

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]

+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

Article | Figures & Data | Info & Metrics | eLetters | PDF

You are currently viewing the abstract. **View Full Text**

Machines learn what people know implicitly

AlphaGo has demonstrated that a machine can learn how to do things that people spend many years of concentrated study learning, and it can rapidly learn how to do them better than any human can. Caliskan *et al.* now show that machines can learn word associations from written texts and that these associations mirror those learned by humans, as measured by the Implicit Association Test (IAT) (see the Perspective by Greenwald). Why does this matter? Because the IAT has predictive value in uncovering the association between concepts, such as pleasantness and flowers or unpleasantness and insects. It can also tease out attitudes and beliefs—for example, associations between female names and family or male names and career. Such biases may not be expressed explicitly, yet they can prove influential in behavior.

Science, this issue p. 183; see also p. 133

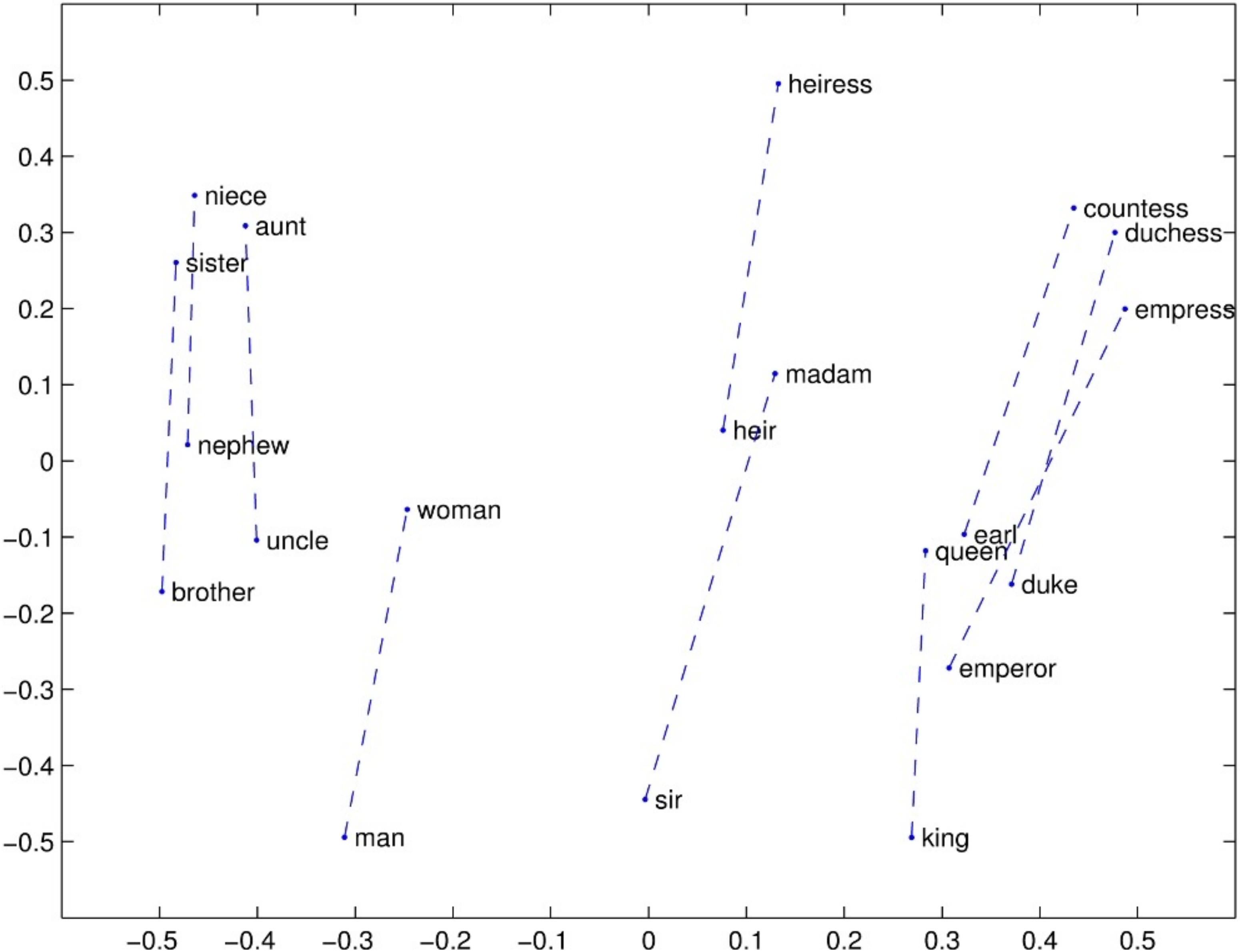
Abstract

Machine learning is a means to derive artificial intelligence by discovering patterns in existing data. Here, we show that applying machine learning to ordinary human language results in human-like semantic biases. We replicated a spectrum of known biases, as measured by the

https://science.sciencemag.org/content/356/6334/183

Science
Vol 356, Issue 6334
14 April 2017

RELATED CONTENT
PERSPECTIVE
An AI stereotype catcher



https://nlp.stanford.edu/projects/glove/images/man_woman.jpg

Problems with word embeddings

- Multiple meanings are conflated into a single representation: polysemy and homonymy are not handled properly
- The necessity to accommodate multiple meanings per word in different vectors is not handled (in all vector space models)

- Solutions
  - Combining the prior knowledge of lexical databases (e.g., WordNet)
  - BERT, which can generate contextually-meaningful embeddings

Trained embeddings

- Word2vec
  - https://code.google.com/archive/p/word2vec/
- Glove
  - http://nlp.stanford.edu/projects/glove/
- Levy/Goldberg dependency embeddings
  - https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings