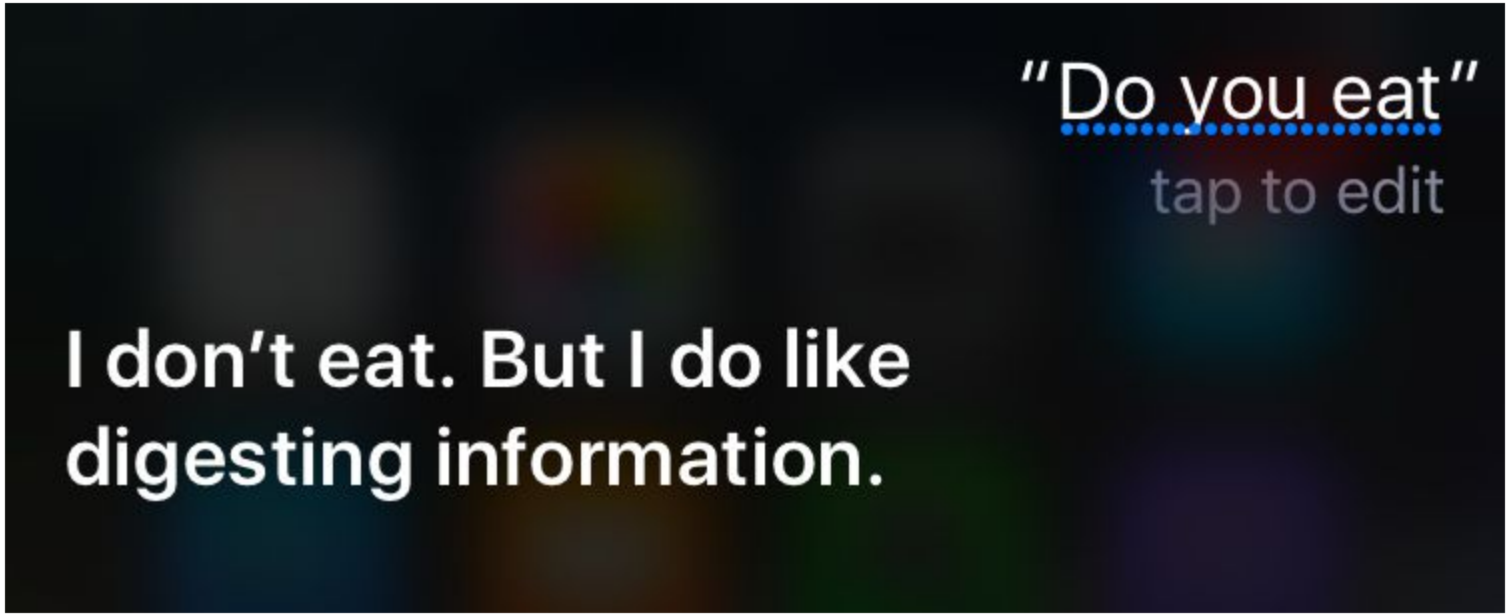
Learning Word Representations

Computer Science Department
Carl Edwards

4/2/2020

How do we communicate with computers?

- If I say “What are the colors of a rainbow,” how does Siri know what I mean?
- How does Siri understand questions?



https://www.freemake.com/blog/siri-answers-20-hilarious-questions/

Can anyone guess the word?

"May the __________ be with you."

Can anyone guess the word?

“There’s no ______ like home.”

Can anyone guess the word?

“Don’t count the _____, make the days count.”

- Muhammad Ali

Can anyone guess the word?

"Hasta la ______, baby."

What are word representations?

Why do we need word representations?

How do we make word representations?

ASCII

- Represent each letter as a number
- How computers traditionally represent words
- Why might this not be useful?

ASCII Table

Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char
0	0	0		32	20	40	[space]	64	40	100	@	96	60	140	`
1	1	1		33	21	41	!	65	41	101	A	97	61	141	a
2	2	2		34	22	42	"	66	42	102	B	98	62	142	b
3	3	3		35	23	43	#	67	43	103	C	99	63	143	c
4	4	4		36	24	44	$	68	44	104	D	100	64	144	d
5	5	5		37	25	45	%	69	45	105	E	101	65	145	e
6	6	6		38	26	46	&	70	46	106	F	102	66	146	f
7	7	7		39	27	47	'	71	47	107	G	103	67	147	g
8	8	10		40	28	50	(	72	48	110	H	104	68	150	h
9	9	11		41	29	51	)	73	49	111	I	105	69	151	i
10	A	12		42	2A	52	*	74	4A	112	J	106	6A	152	j
11	B	13		43	2B	53	+	75	4B	113	K	107	6B	153	k
12	C	14		44	2C	54	,	76	4C	114	L	108	6C	154	l
13	D	15		45	2D	55	-	77	4D	115	M	109	6D	155	m
14	E	16		46	2E	56	.	78	4E	116	N	110	6E	156	n
15	F	17		47	2F	57	/	79	4F	117	O	111	6F	157	o
16	10	20		48	30	60	0	80	50	120	P	112	70	160	p
17	11	21		49	31	61	1	81	51	121	Q	113	71	161	q
18	12	22		50	32	62	2	82	52	122	R	114	72	162	r
19	13	23		51	33	63	3	83	53	123	S	115	73	163	s
20	14	24		52	34	64	4	84	54	124	T	116	74	164	t
21	15	25		53	35	65	5	85	55	125	U	117	75	165	u
22	16	26		54	36	66	6	86	56	126	V	118	76	166	v
23	17	27		55	37	67	7	87	57	127	W	119	77	167	w
24	18	30		56	38	70	8	88	58	130	X	120	78	170	x
25	19	31		57	39	71	9	89	59	131	Y	121	79	171	y
26	1A	32		58	3A	72	:	90	5A	132	Z	122	7A	172	z
27	1B	33		59	3B	73	;	91	5B	133	[	123	7B	173	{
28	1C	34		60	3C	74	<	92	5C	134	\	124	7C	174	|
29	1D	35		61	3D	75	=	93	5D	135	]	125	7D	175	}
30	1E	36		62	3E	76	>	94	5E	136	^	126	7E	176	~
31	1F	37		63	3F	77	?	95	5F	137	_	127	7F	177	

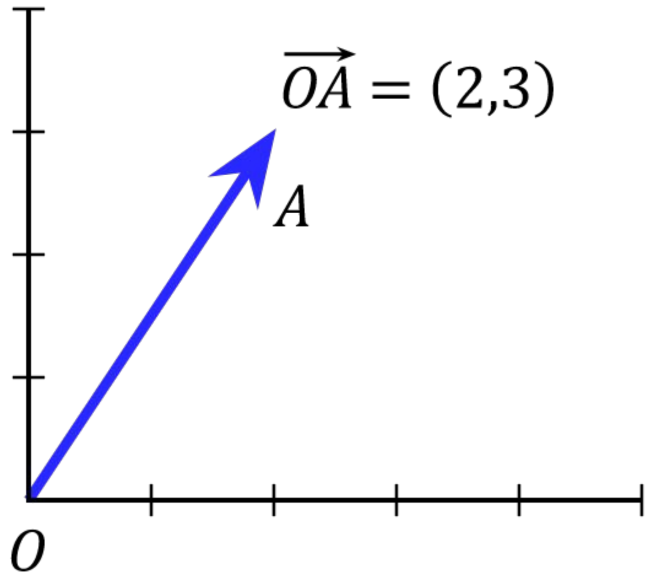
What's a vector?

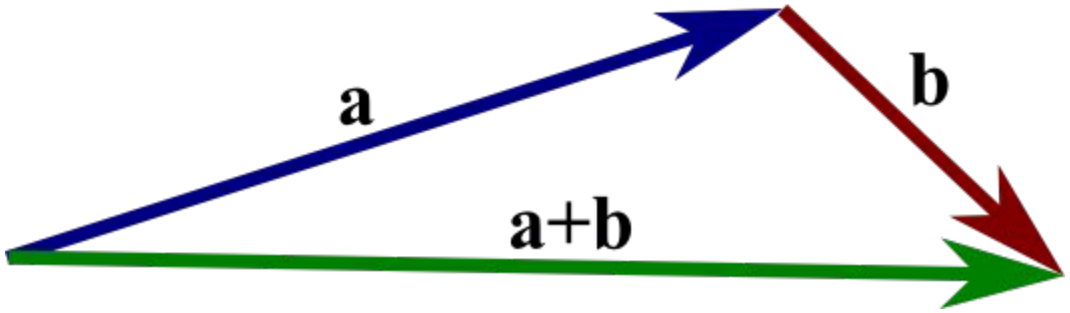
- A vector is essentially a fixed-length list of numbers
- For example,

  A = [3,1]

  B = [1,-1]

  A + B = [4,0]





https://mathinsight.org/vector_introduction

One-Hot Encodings

- Represent a word with a vector of ones and zeros
- Have a vocabulary (a set of words) called V
  - V has a size of S words
- Each vector will have one element for each word
- Elements are all 0 except for the word's element which is one

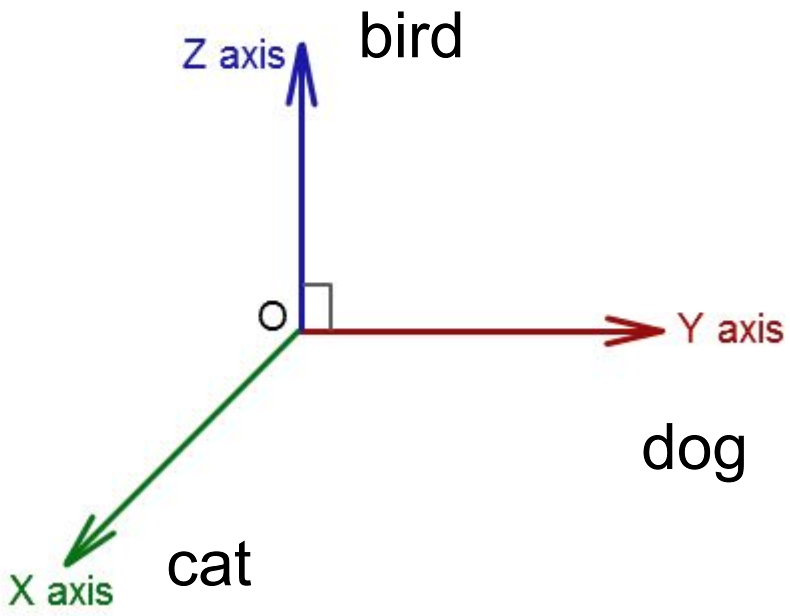
Example:

V = {cat, dog, bird}, S = 3

cat = [1,0,0]

dog = [0,1,0]

bird = [0,0,1]



Problems with One-Hot Encodings

- Is there any meaning?
- Why is bird the same distance from cat as it is from dog?
- Shouldn't cats and dogs be more related, or closer together, than birds?

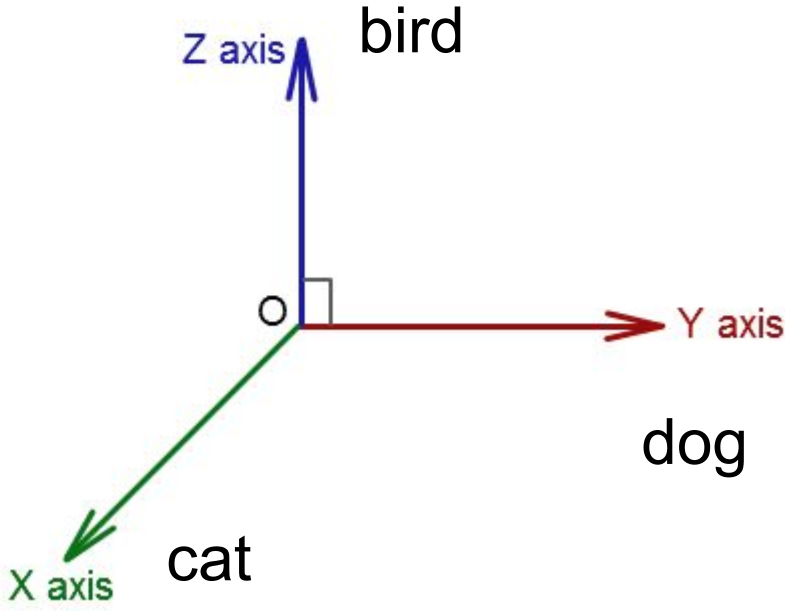
Example:

V = {cat, dog, bird}

cat = [1,0,0]

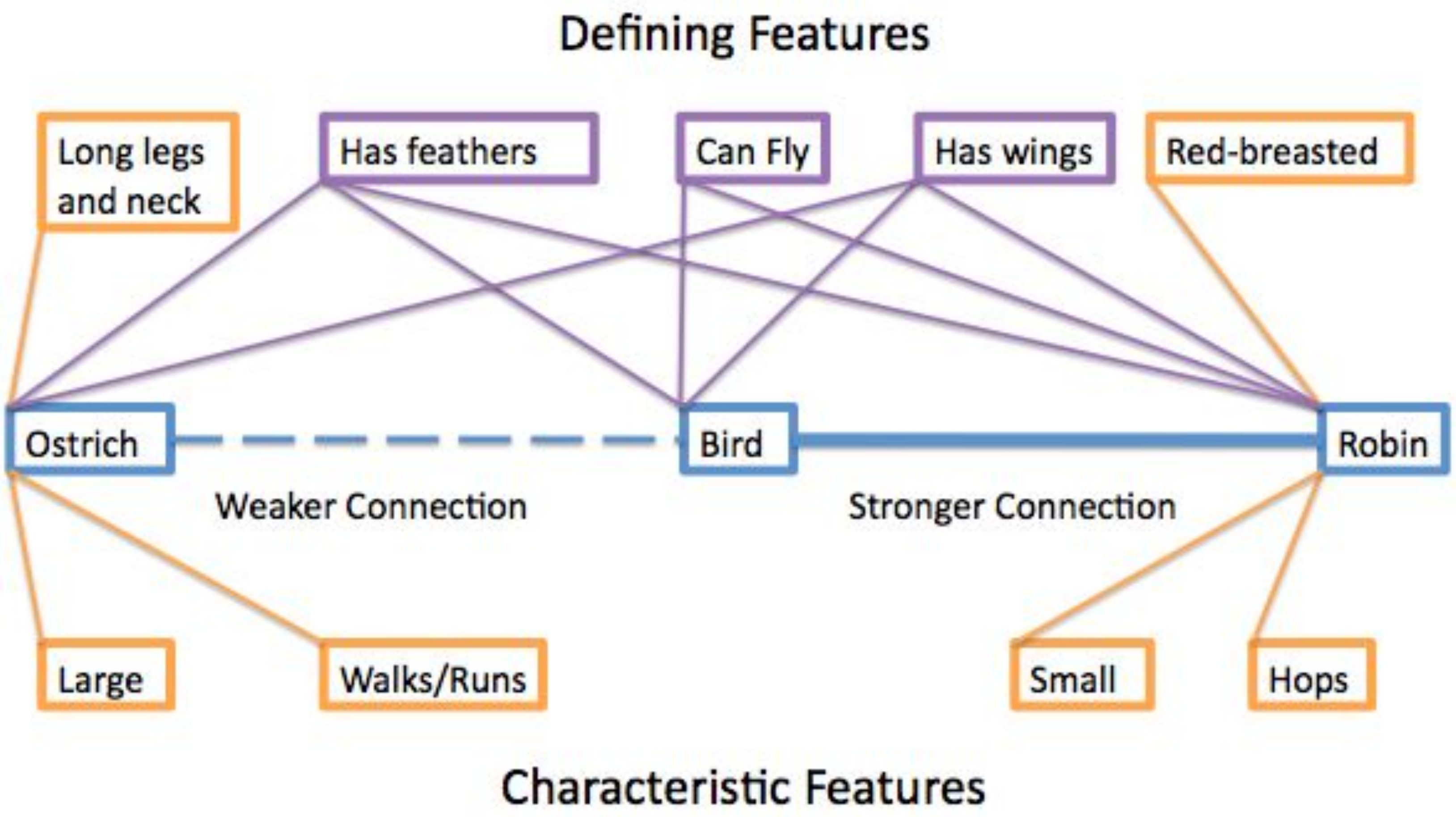
dog = [0,1,0]

bird = [0,0,1]



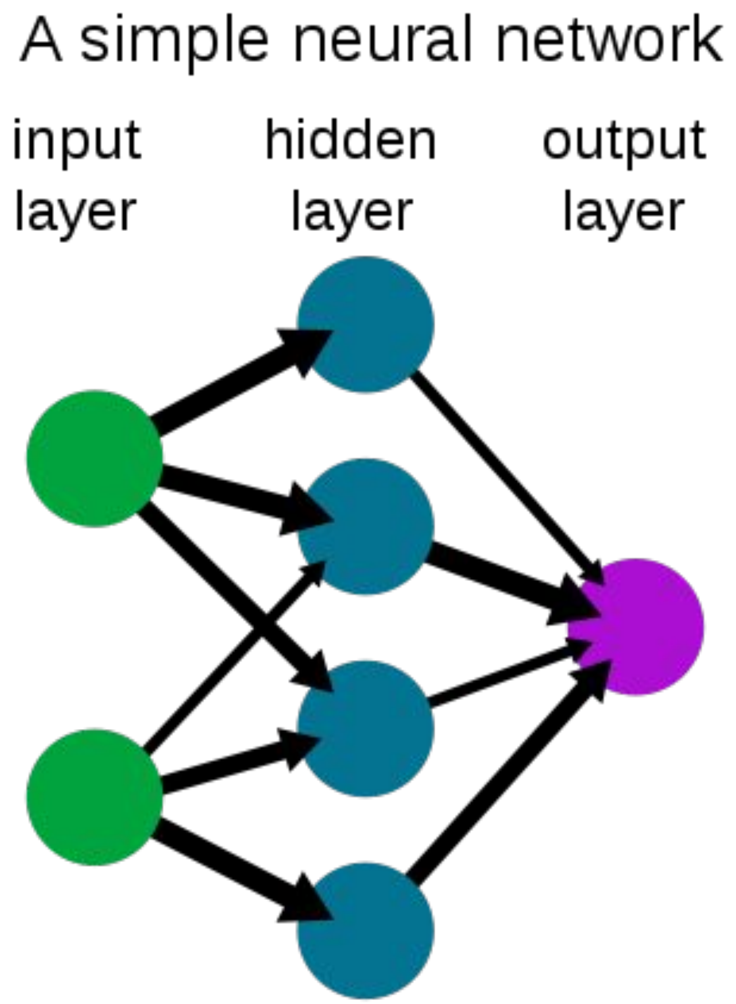
Semantic Information

- semantic definition: relating to meaning in language or logic.
- We want computers to understand words, not just represent them.
- How can we do this?



Neural Networks

- Neural networks take numbers as input and predict some output.
- There are ‘weights’ that are multiplied on each of the black lines.
- We can think of neural networks as a “function.”
- Our goal is to “teach” the neural network to give us outputs that we want.

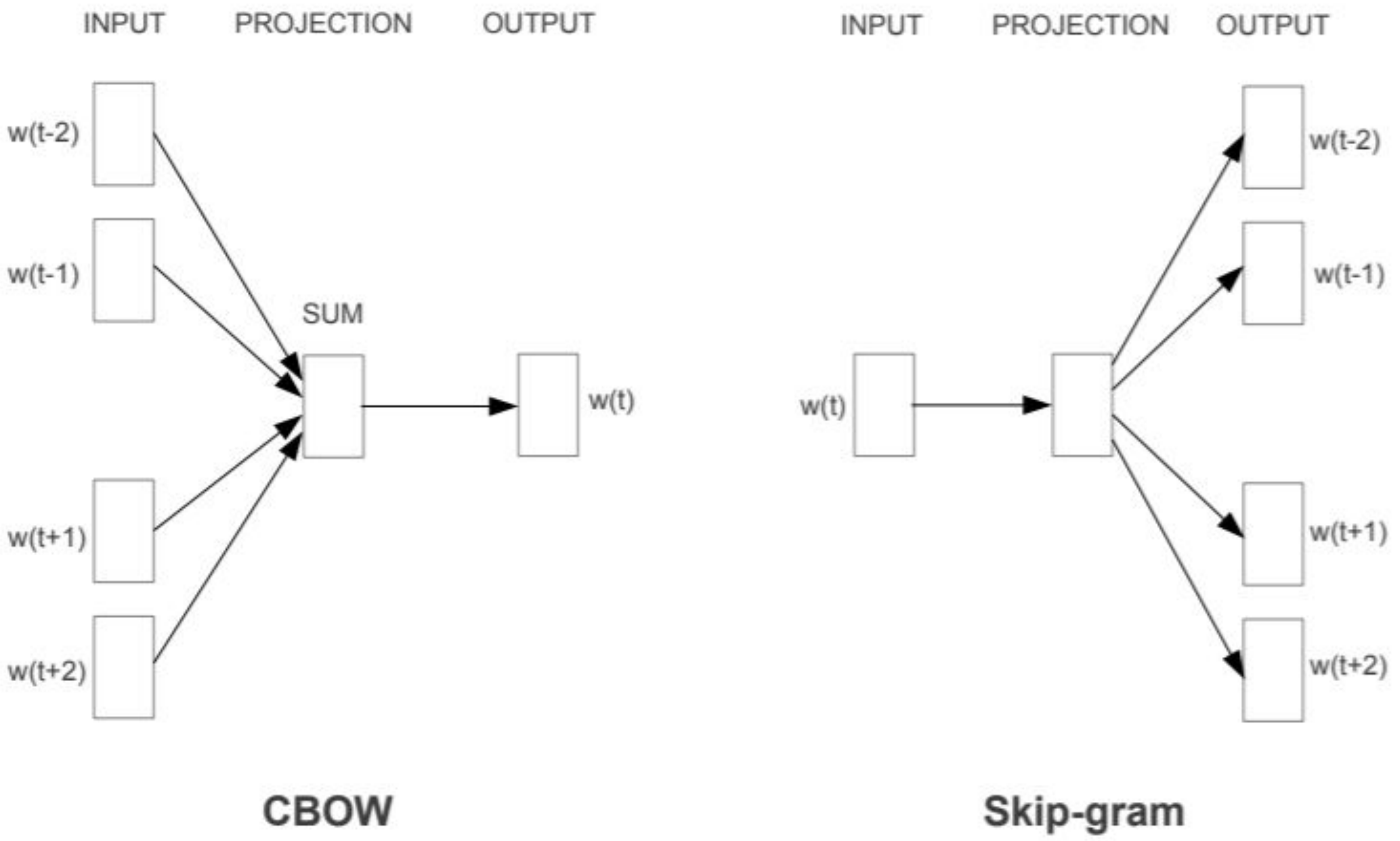


word2vec algorithm - (Mikolov et al. 2013)

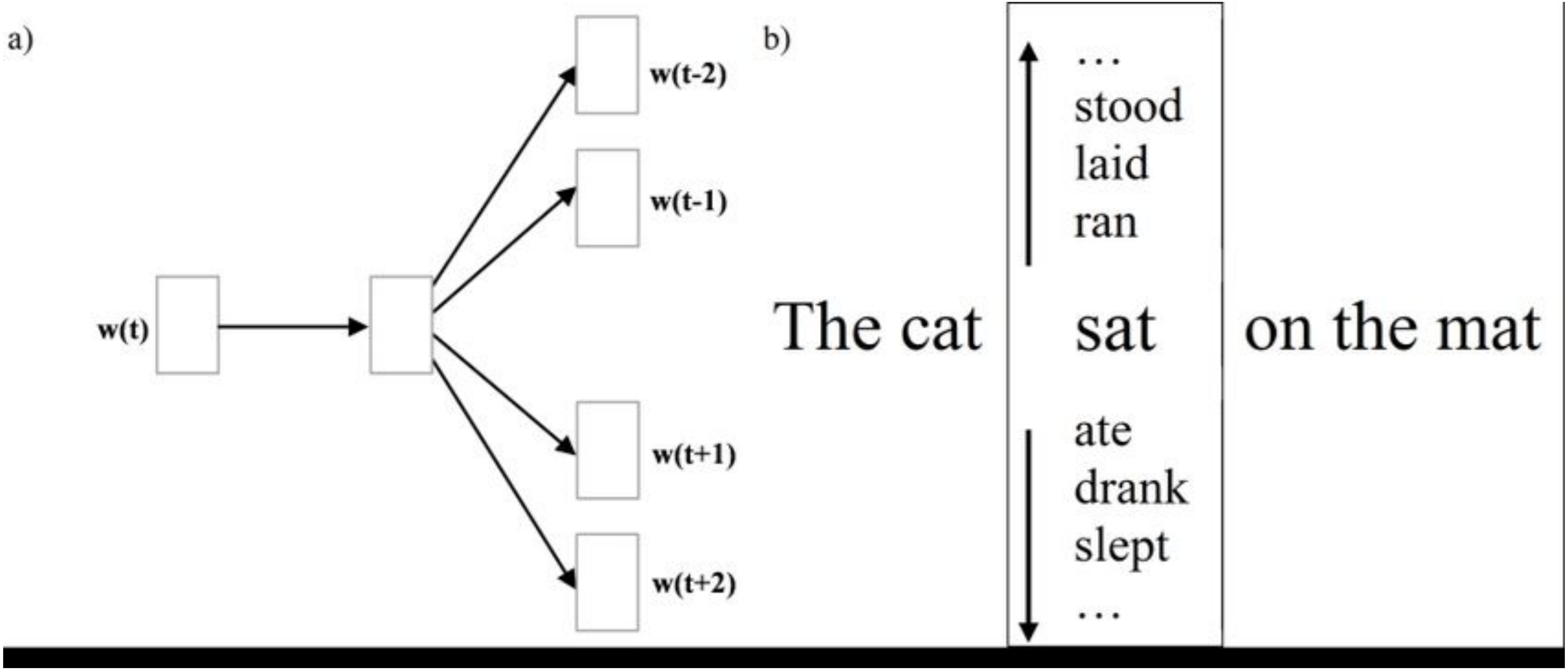
- Can we use written language to help improve representations?
- The word2vec (word to vector) algorithm uses written text to learn.
- Based on the distributional hypothesis:

“you shall know a word by the company it keeps” - Firth (1957)

word2vec algorithm - (Mikolov et al. 2013)



word2vec continued



https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c

word2vec algorithm - (Mikolov et al. 2013)

Window Size	Text	Skip-grams
2	[The **wide** road shimmered] in the hot sun.	wide, the wide, road wide, shimmered
	The [wide road **shimmered** in the] hot sun.	shimmered, wide shimmered, road shimmered, in shimmered, the
	The wide road shimmered in [the hot **sun**].	sun, the sun, hot
3	[The **wide** road shimmered in] the hot sun.	wide, the wide, road wide, shimmered wide, in
	[The wide road **shimmered** in the hot] sun.	shimmered, the shimmered, wide shimmered, road shimmered, in shimmered, the shimmered, hot
	The wide road shimmered [in the hot **sun**].	sun, in sun, the sun, hot

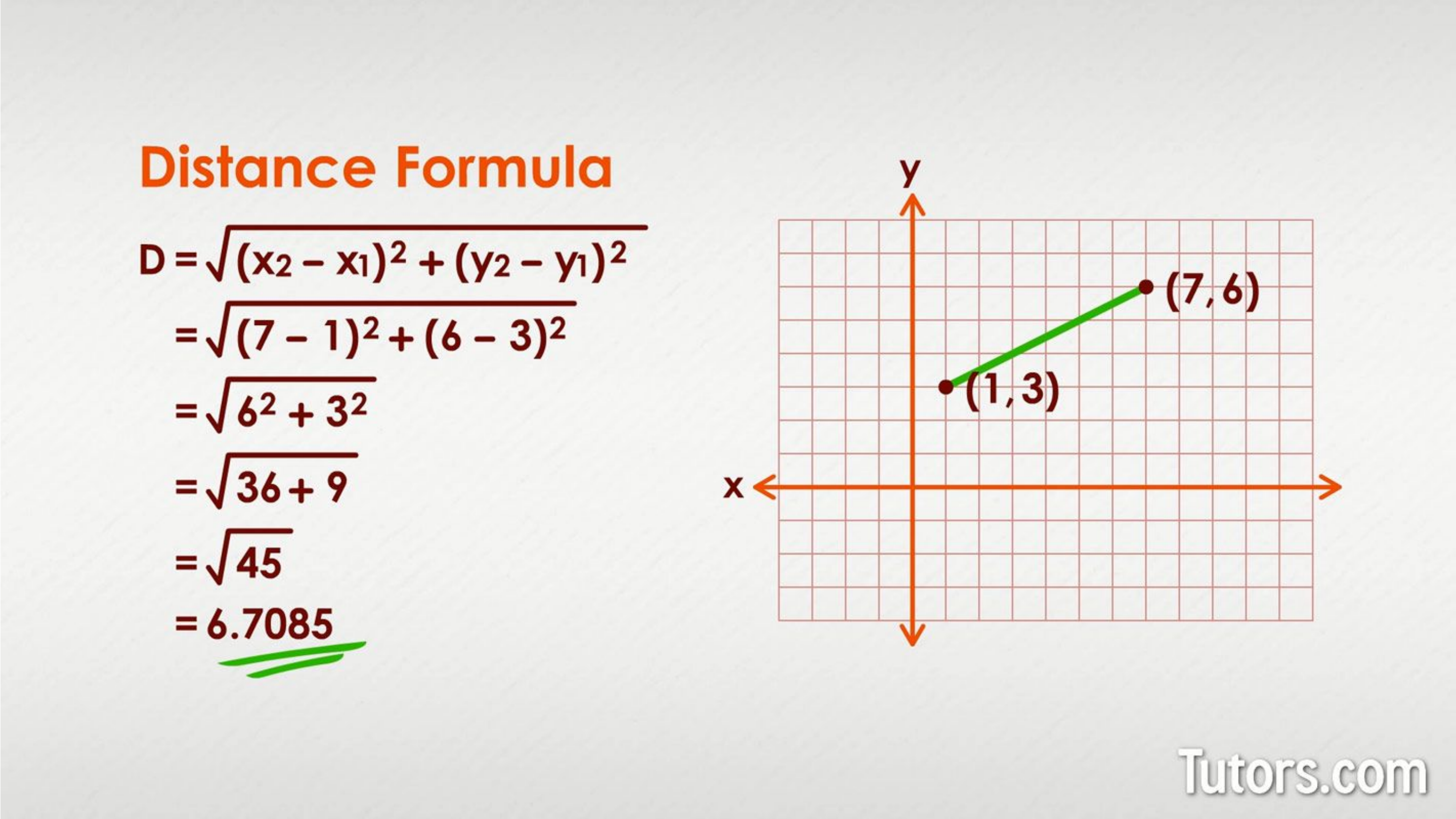
https://www.tensorflow.org/tutorials/text/word2vec

How to measure similarity between these word vectors?

Euclidean Distance

(the distance formula)

Range: [0, ∞):



A = [3,1]

B = [1,-1]

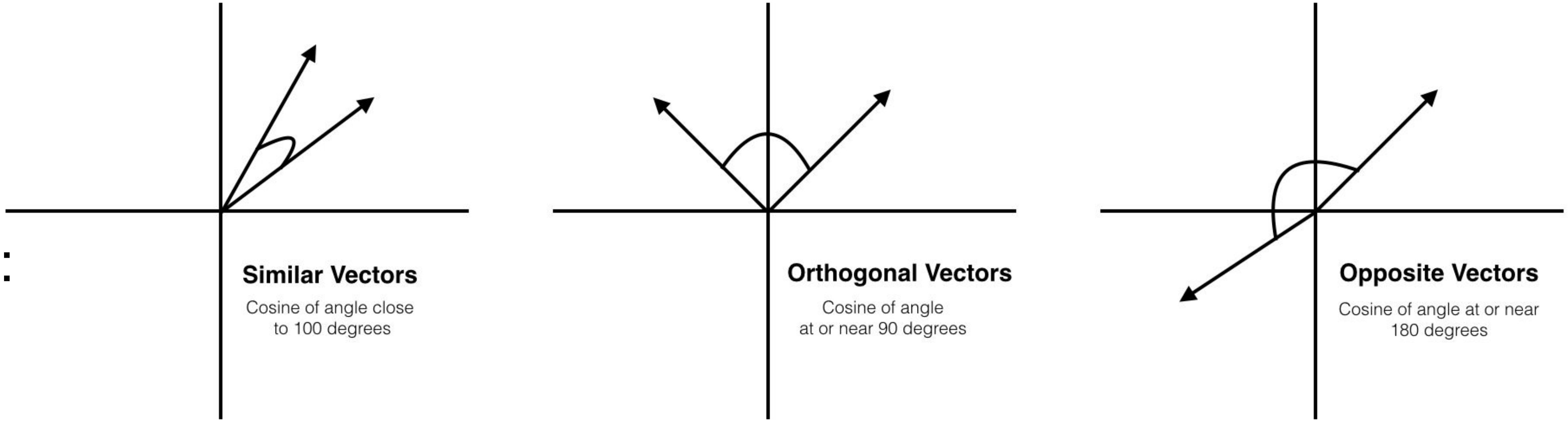
Dist(A,B) =

$$\sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2} = \sqrt{(3-1)^2 + (1 - -1)^2} = \sqrt{4+4} = \sqrt{8}$$

What's another way to measure similarity?

Only measures the similarity of direction!

Cosine Similarity Range: [-1, 1]:



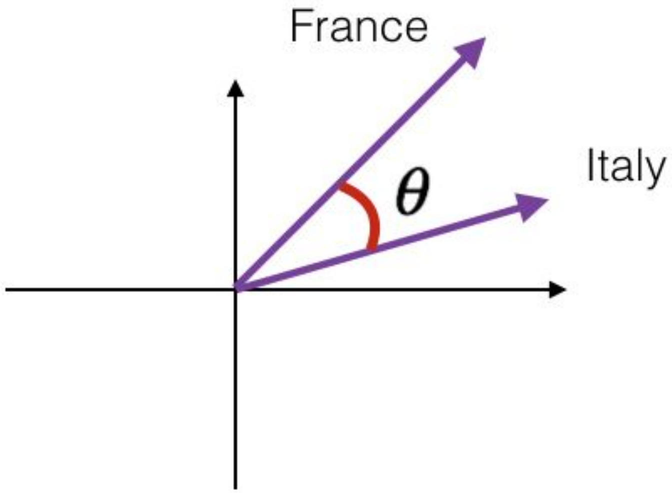
https://deepai.org/machine-learning-glossary-and-terms/cosine-similarity

A = [3,1]

B = [1,-1]

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, = \frac{3\times 1 + 1\times -1}{\sqrt{3^2+1^2}\sqrt{1^2+(-1)^2}} = \frac{2}{\sqrt{10}\sqrt{2}} = 0.447$$

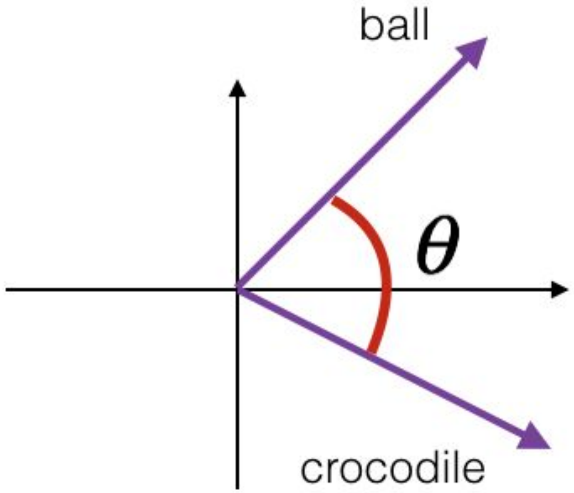
Cosine similarity example



France and Italy are quite similar

θ is close to 0°

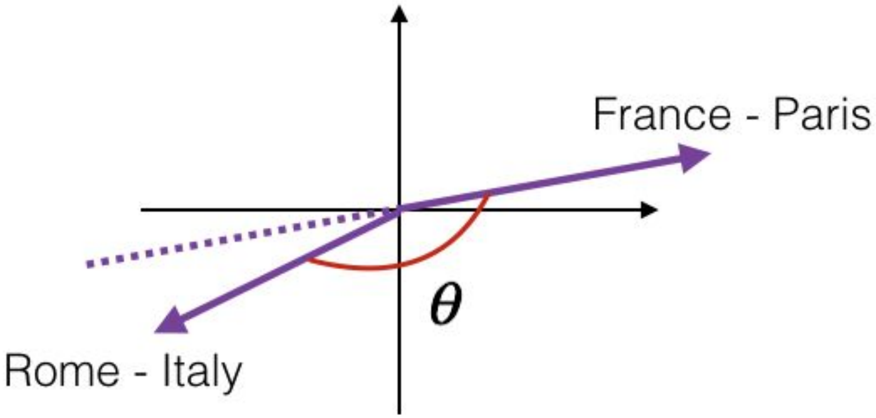
$\cos(\theta) \approx 1$



ball and crocodile are not similar

θ is close to 90°

$\cos(\theta) \approx 0$



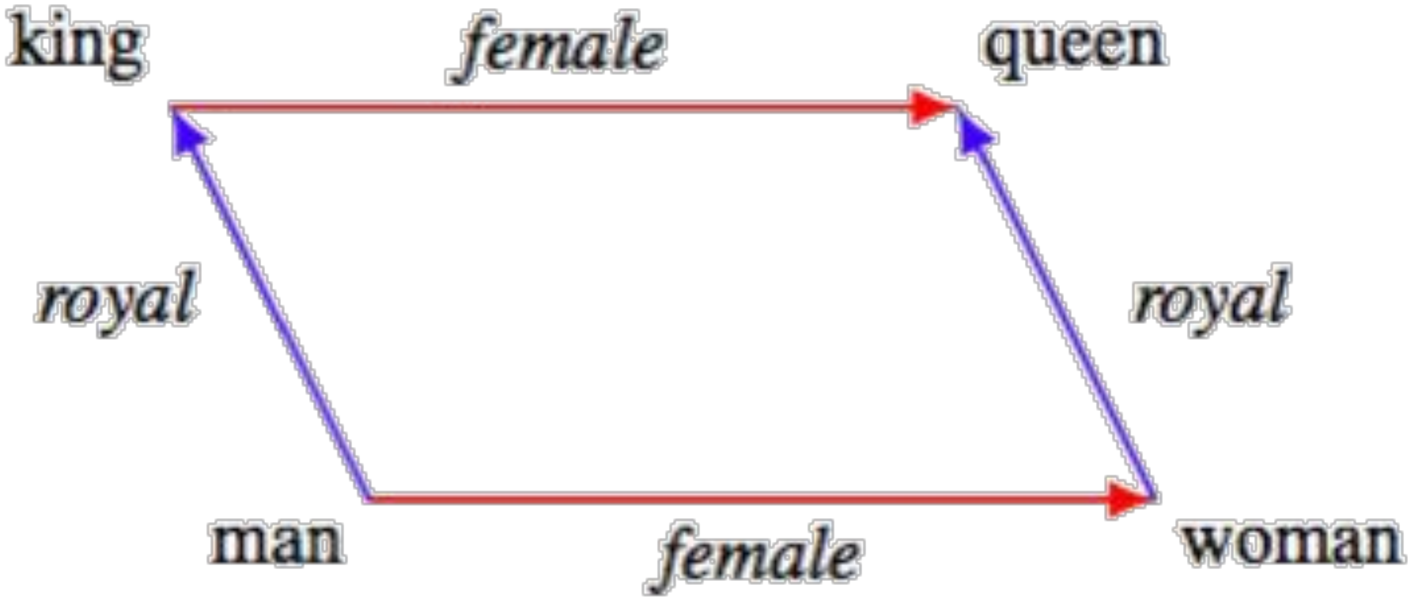
the two vectors are similar but opposite
the first one encodes (city - country)
while the second one encodes (country - city)

θ is close to 180°

$\cos(\theta) \approx -1$

https://datascience-enthusiast.com/DL/Operations_on_word_vectors.html

More word2vec



$$\vec{king} - \vec{man} + \vec{woman} \approx \vec{queen}$$

https://kawine.github.io/blog/nlp/2019/06/21/word-analogies.html

What can we do with these embeddings?

Question Answering

https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html

What can we do with these embeddings?

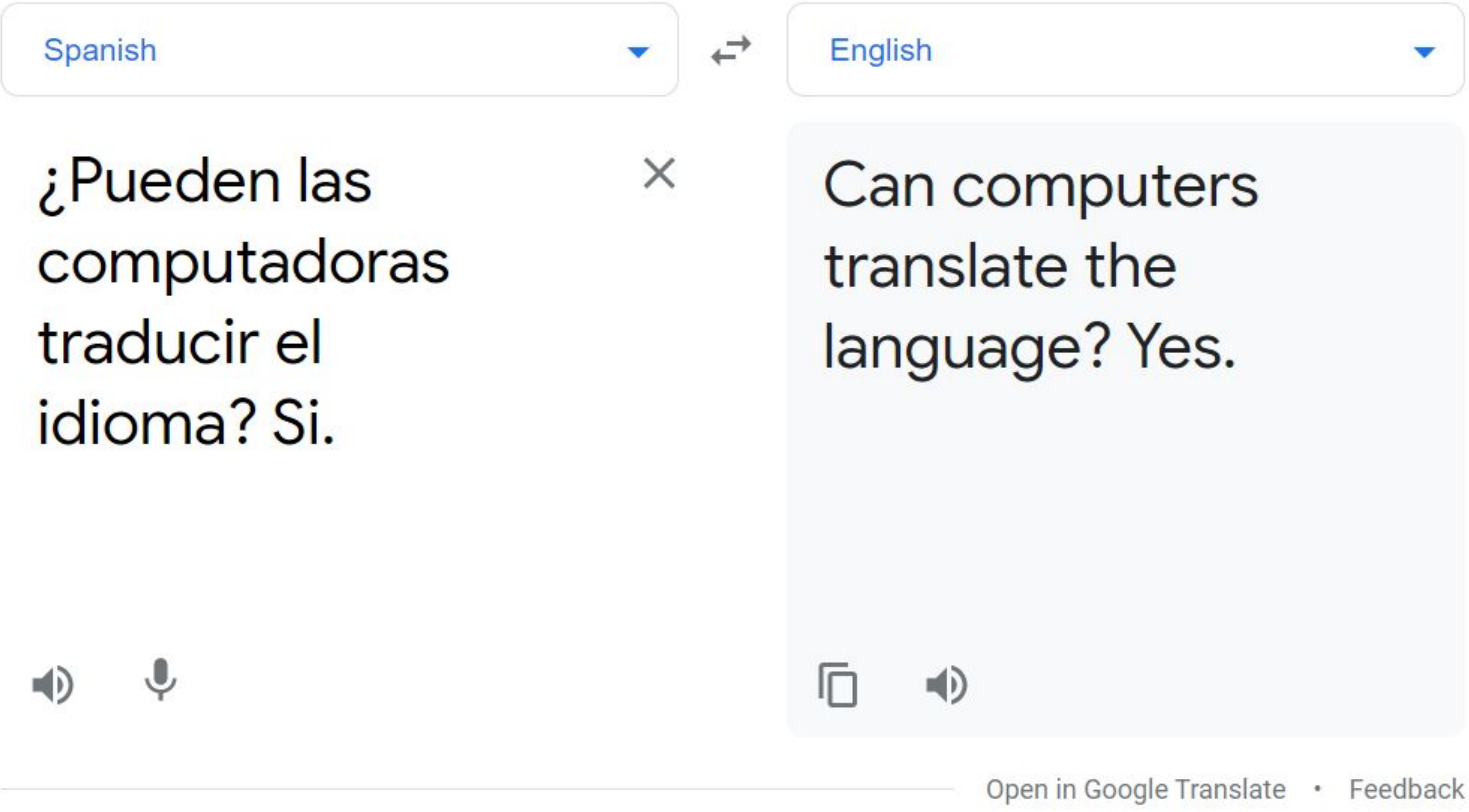
Sentiment Analysis

- e.g Is a tweet happy or sad?

https://monkeylearn.com/sentiment-analysis/

What can we do with these embeddings?

Translation

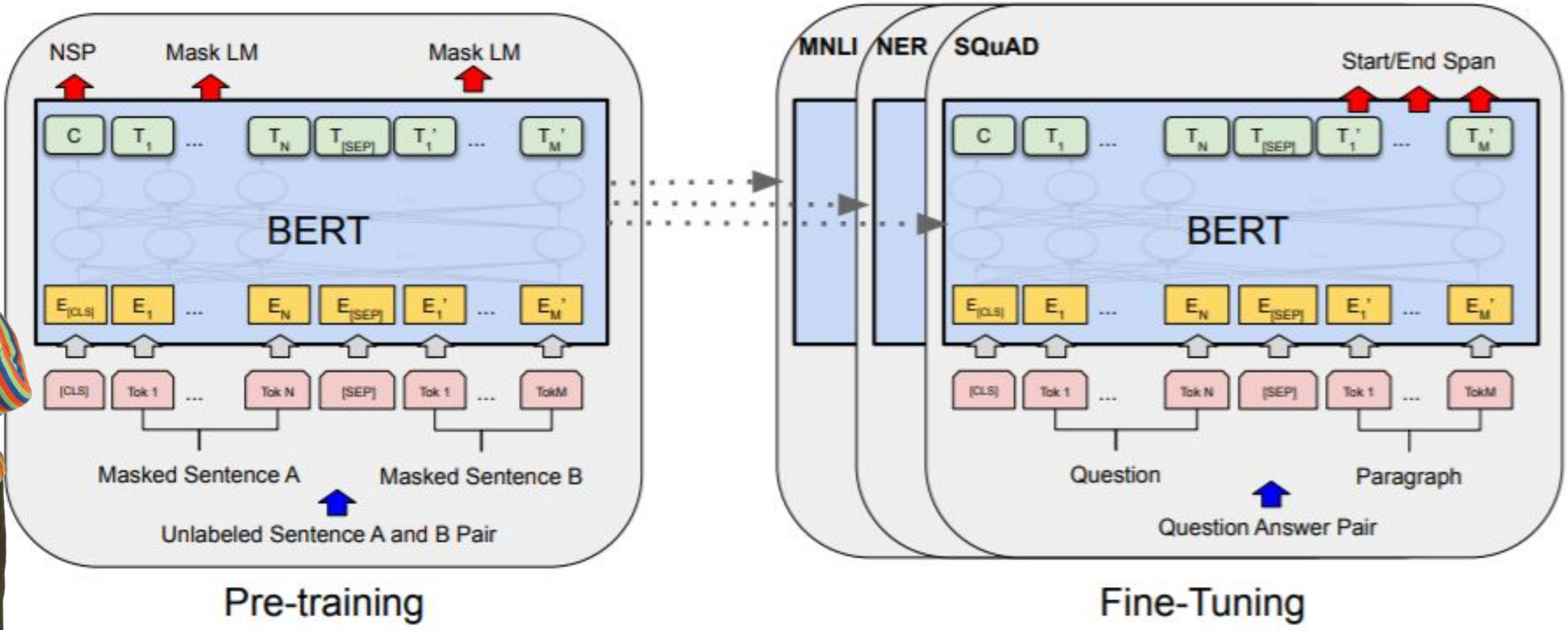


What can we do with these embeddings?

Image Captioning

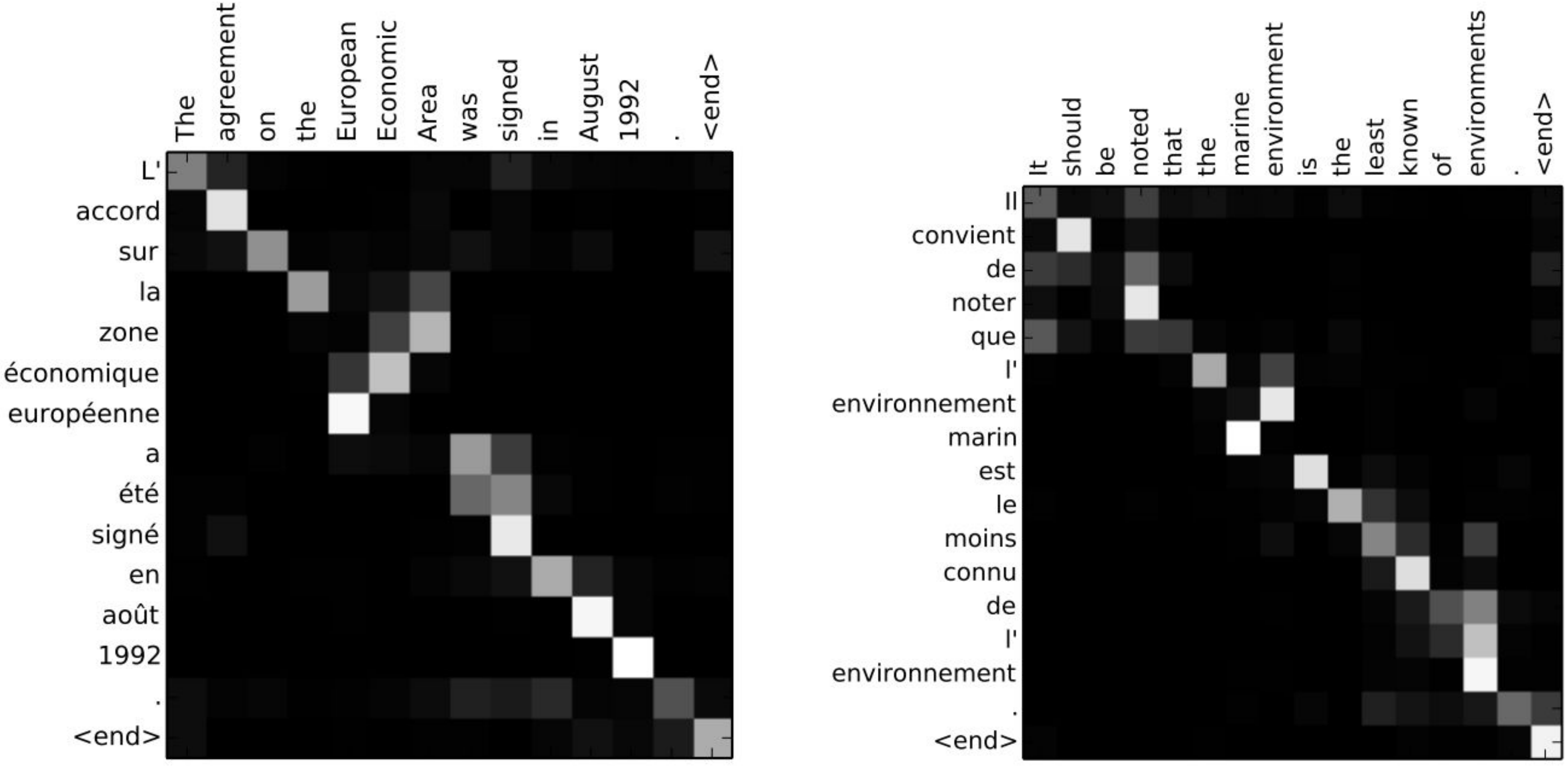
https://github.com/danieljl/keras-image-captioning

Where are we now? BERT (Devlin et al. 2018)



Bidirectional Encoder Representations from Transformers (BERT)

BERT uses attention



(a) (b)

https://observablehq.com/@clpuc/analyzing-the-design-space-for-visualizing-neural-attenti

What's next?

- Improving Siri's ability to learn what you mean
  - Can Siri learn things specifically to communicate with you and the way you talk?
- Can we integrate pictures better with language?
- Can we understand entire books?
- How can we learn irony and sarcasm?
- How can we learn slang words quickly without examples?
  - Right now we need billions of examples!

Does anyone have any questions?