

Setup

<http://bit.ly/InsightNLP>

Follow along (or run the code yourself)

1/ Go to link above

2/ Click on **NLP_notebook.ipynb** and follow along

3 (optional)/ If you have the necessary libraries installed and pretrained weights, clone the repository and run “jupyter notebook”



@ODSC

CONCRETE NLP SOLUTIONS

Emmanuel Ameisen
San Francisco | November 3rd 2017



EMMANUEL AMEISEN

AI Program Director, ML Engineer



 @emmanuelameisen

 /in/ameisen

PAST EXPERIENCE



DATA SCIENTIST



DATA SCIENTIST



MSC. COMPUTER SCIENCE



MSC. ARTIFICIAL
INTELLIGENCE



MSC. MANAGEMENT



INSIGHT FELLOWS ARE DATA SCIENTISTS AND DATA ENGINEERS EVERYWHERE

facebook

LinkedIn

okcupid

PREMISE

Gartner

Bloomberg

ACTIVISION

SQUARESPACE

BLACKROCK

BIRCHBOX

greenhouse

NBC

STITCH FIX

airbnb

YAHOO!

VECTRA

twitch

Pinterest

JAWBONE

OSCAR

JPMorgan

salesforceIQ

DOWJONES

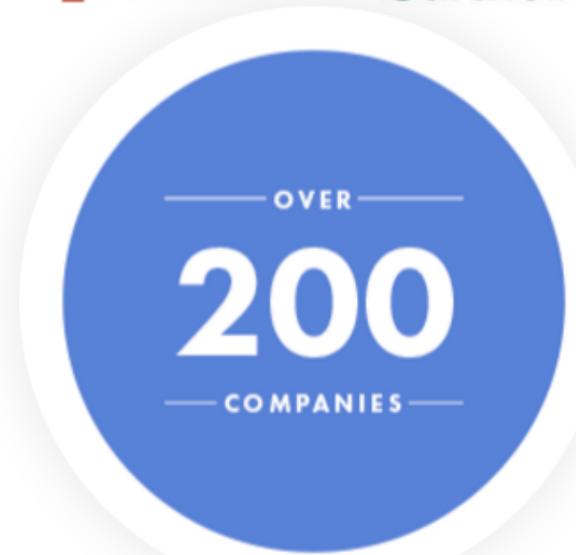
Palantir

SAMBA TV

AXON VIBE

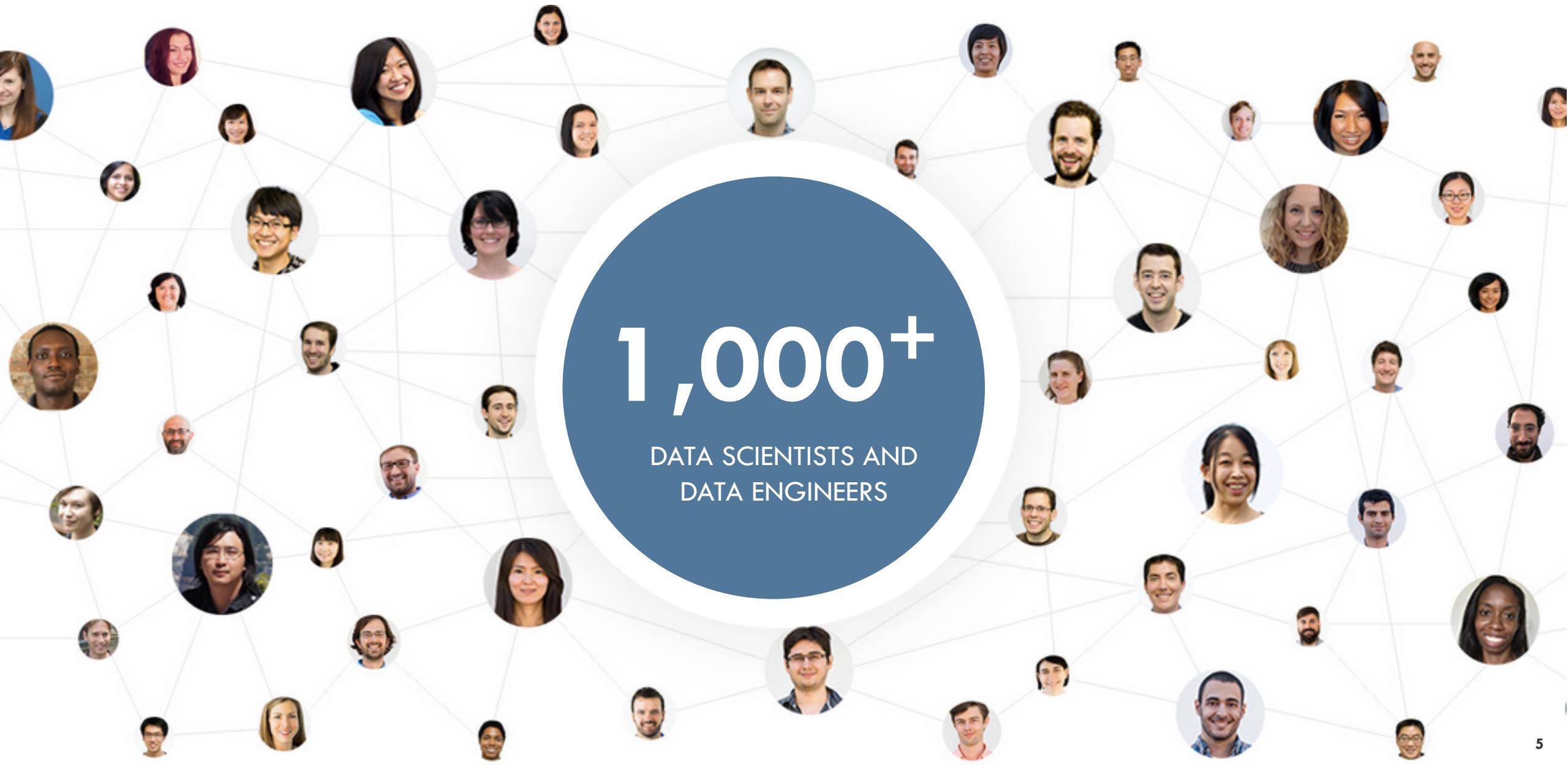
intuit

KHAN ACADEMY



SILICON VALLEY • NEW YORK • BOSTON • LOS ANGELES • SEATTLE
+ MANY OTHERS...

NETWORK OF INSIGHT ALUMNI



1,000+
DATA SCIENTISTS AND
DATA ENGINEERS

EXAMPLE PROJECTS

FASHION CLASSIFIER



AUTOMATIC REVIEW GENERATION



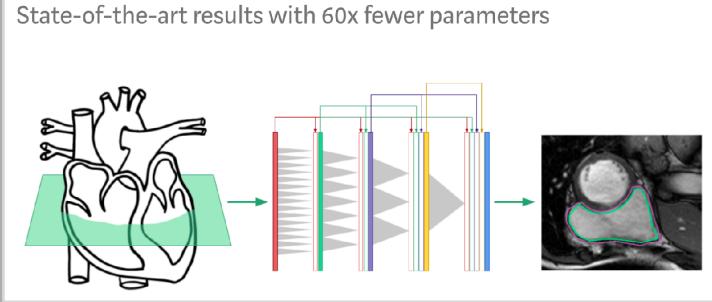
9/24/2017

Great service! The place is very relaxed. The curry is outstanding. I am always satisfied with the food and the ambiance.

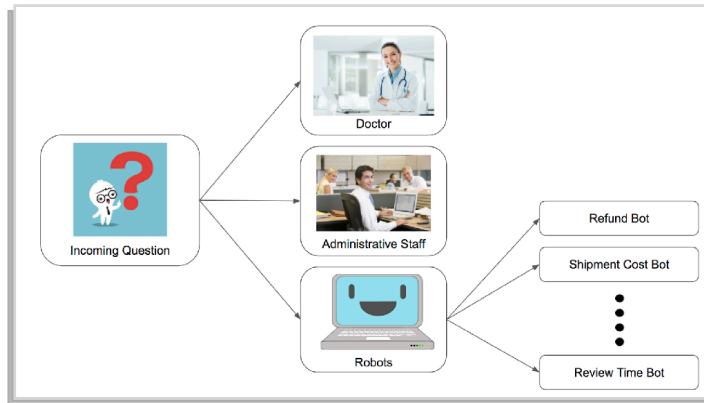
HEART SEGMENTATION

Heart Disease Diagnosis with Deep Learning

State-of-the-art results with 60x fewer parameters



SUPPORT REQUEST CLASSIFICATION

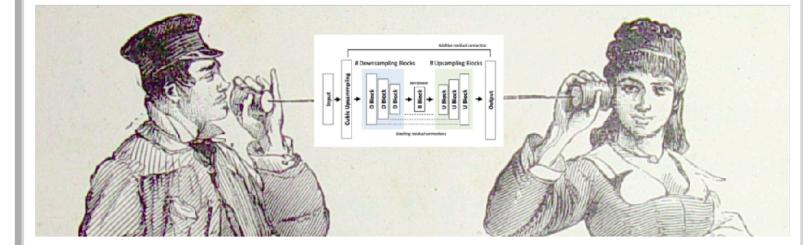


READING TEXT IN VIDEOS



SPEECH UNSAMPLING

Using Deep Learning to Reconstruct High-Resolution Audio



CONCRETE NLP SOLUTIONS

- CLUSTERING
- CLASSIFICATION
- GENERATION

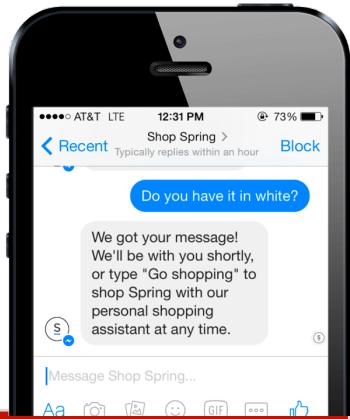
A HUGE VALUE TO BE UNLOCKED

Business Models



A screenshot of a tweet from the CIA's official Twitter account (@CIA). The tweet reads: "We can neither confirm nor deny that this is our first tweet." It includes standard Twitter interaction buttons (Reply, Retweet, Favorite, Storify, More) and engagement metrics (45,924 Retweets, 21,754 Favorites). The timestamp is 1:49 PM - 6 Jun 2014.

Every Day Uses



A screenshot of a Yelp mobile page showing a recommended review. The review is from "Jenn P." in San Francisco, CA, who has 1 friend and 22 reviews. The rating is 5 stars, dated 10/17/2013. The review text reads: "Absolutely Outstanding! The Grounds at Grace Vineyards are stunning...there are SO many photo ops. I must give 5 stars for Steve the owner he is simply wonderful. He was so organized, flexible and prompt I never was stressed. The food was great and the vino was delicious! If your looking for a beautiful venue with many things included this is the place."

ALL YOU NEED IS LOVE LABELS

A screenshot of a Twitter post from user @RichardSocher. The post features a profile picture of a man with brown hair and a beard. The text of the tweet reads: "Rather than spending a month figuring out an unsupervised machine learning problem, just label some data for a week and train a classifier." Below the tweet, the timestamp "2:47 PM - 10 Mar 2017" is visible. At the bottom of the post, there are engagement metrics: "301 Retweets" and "627 Likes", followed by a row of small circular profile pictures representing the users who liked the post.

Richard
@RichardSocher

Following ▾

Rather than spending a month figuring out an unsupervised machine learning problem, just label some data for a week and train a classifier.

2:47 PM - 10 Mar 2017

301 Retweets 627 Likes

ALL YOU NEED IS CLEAN DATA



NSA

@NSACareers



Follow

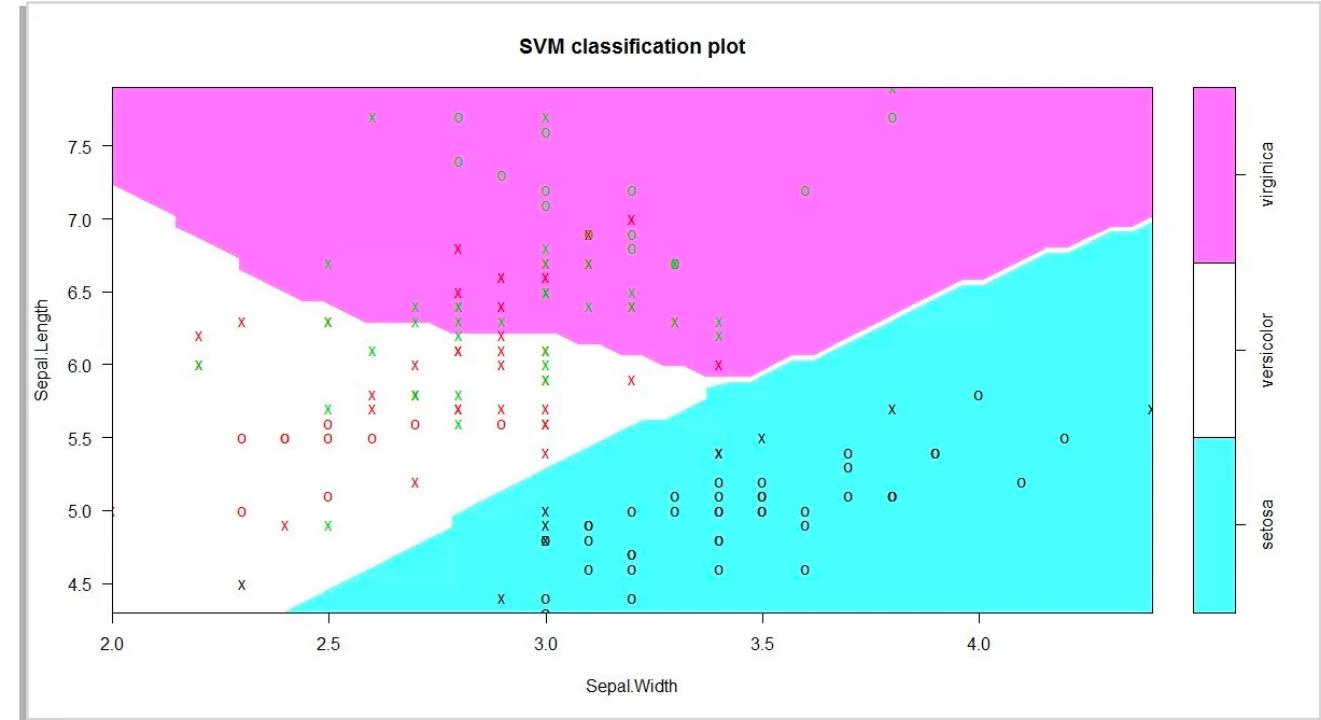
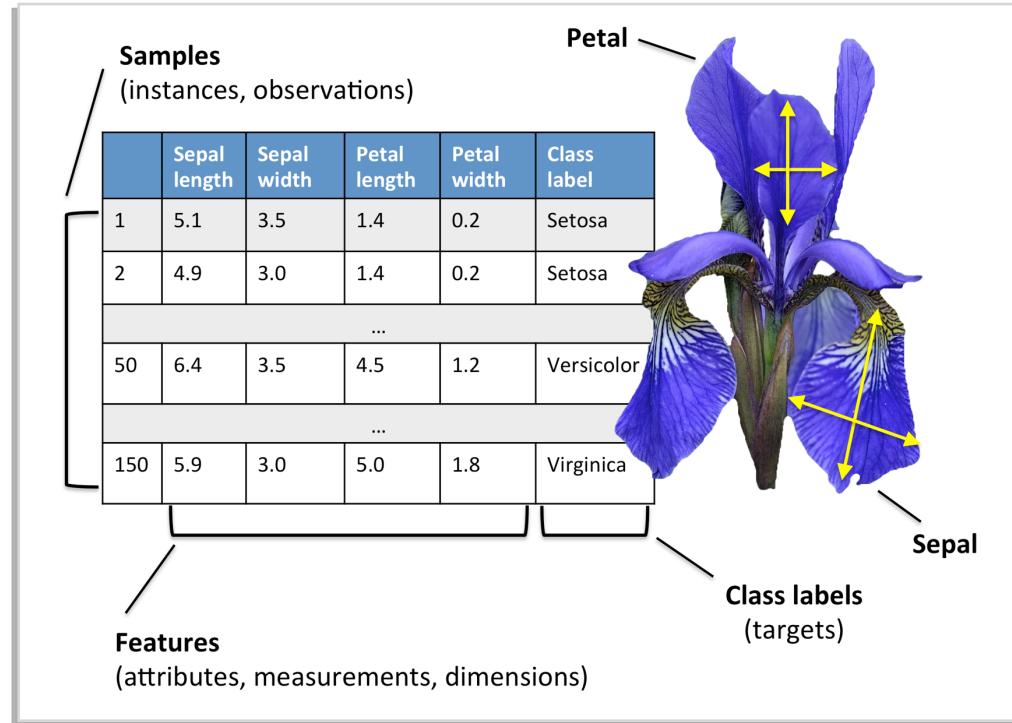
tpfccdlfdtte pcaccplircdt dklpclf?qeiq lhpqlipqeofd
gpwafopwpri izxndkiqpkii kririfcapnc dxkdciqcafmd
vkfpcadf. #MissionMonday #NSA #news

7:19 AM - 5 May 2014

2,283 RETWEETS 1,087 FAVORITES

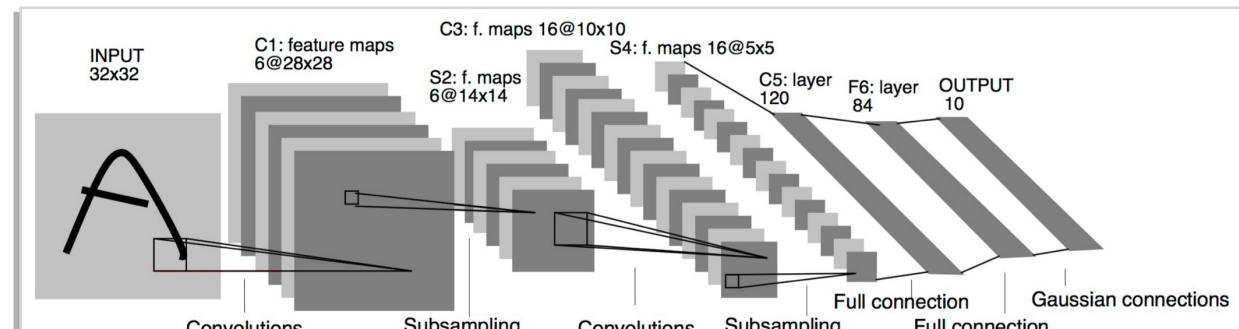


HOW DO YOU REPRESENT DATA?



HOW DO YOU REPRESENT IMAGES?

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC				
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0				
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0				
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
10	0.0	0.0	0.0	0.4	0.5	0.9	0.9	0.9	0.9	0.9	1.0	1.0	1.0	1.0	0.9	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
11	0.0	0.0	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
12	0.0	0.0	0.9	1.0	0.8	0.8	0.8	0.8	0.5	0.2	0.2	0.2	0.2	0.5	0.9	1.0	1.0	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
13	0.0	0.0	0.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.8	1.0	1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	1.0	1.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	1.0	1.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.6	1.0	1.0	1.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.9	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0			
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.5	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0			
19	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0			
20	0.0	0.0	0.0	0.0	0.0	0.4	0.9	1.0	0.9	0.9	0.5	0.3	0.1	0.0	0.0	0.0	0.8	1.0	1.0	0.9	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
21	0.0	0.0	0.0	0.0	0.0	0.7	1.0	0.7	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.9	1.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
22	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	1.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	0.9	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	1.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	1.0	1.0	0.9	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	1.0	1.0	0.9	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	1.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
28	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
29	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
30																																	



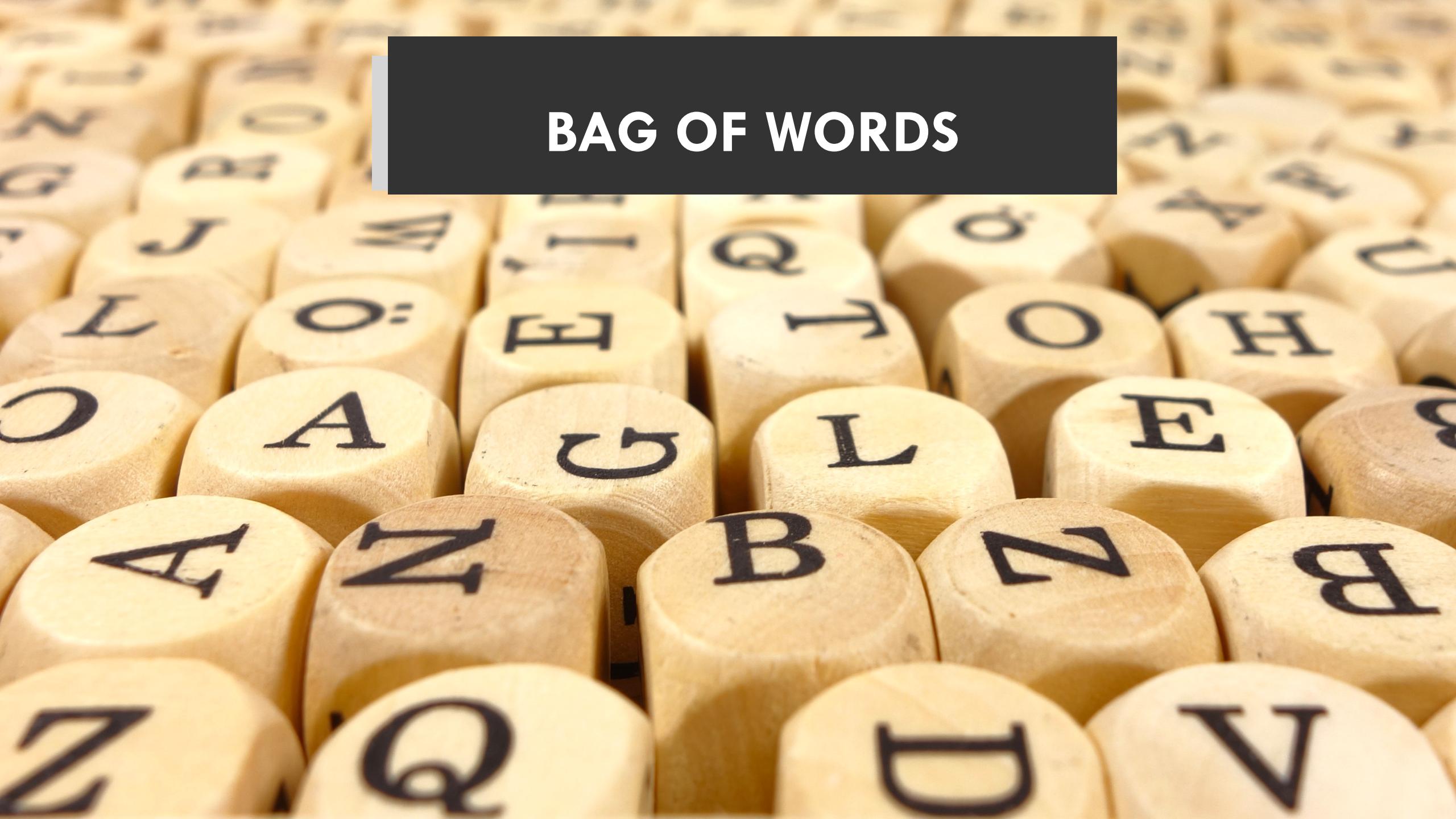
HOW DO YOU REPRESENT WORDS?

ASCII

a	97	n	110
b	98	o	111
c	99	p	112
d	100	q	113
e	101	r	114
f	102	s	115
g	103	t	116
h	104	u	117
i	105	v	118
j	106	w	119
k	107	x	120
l	108	y	121
m	109	z	122

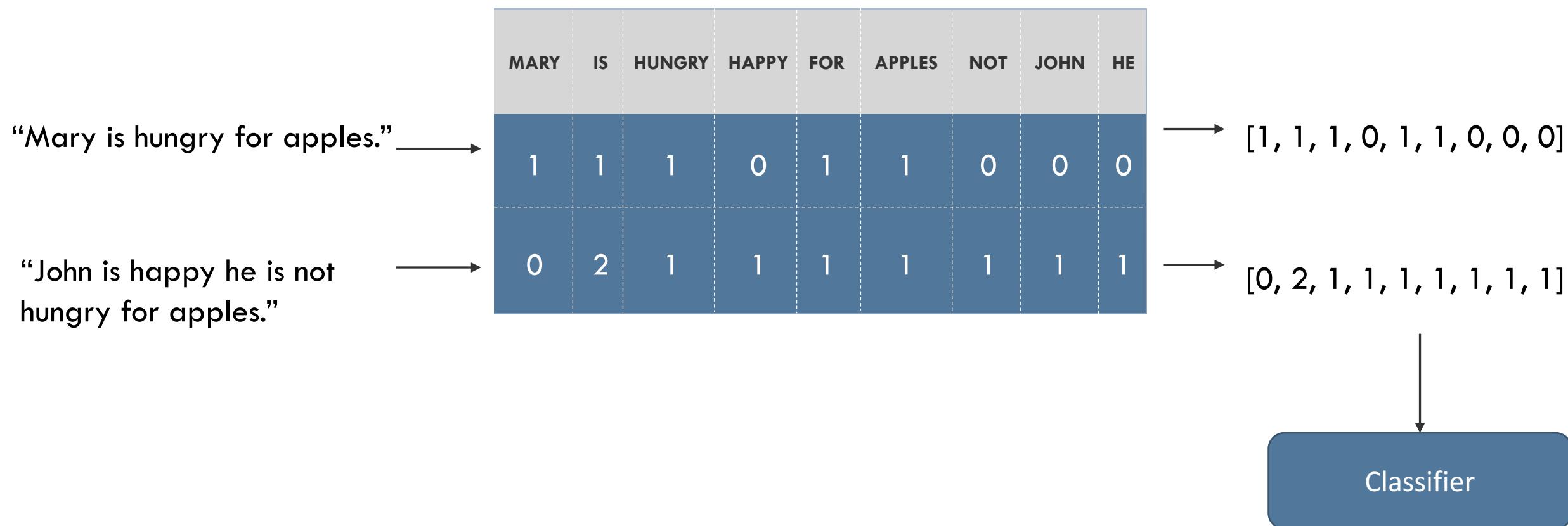
- DISCRETE
- NEED TO LEARN LANGUAGE AND WORDS
- HOW DO YOU REPRESENT A WHOLE SENTENCE

WE NEED A SENTENCE LEVEL REPRESENTATION



BAG OF WORDS

BAG OF WORDS (AKA ONE HOT EMBEDDING)



BAG OF WORDS

PROS

- Quick setup
- Works very well if some words are very important
- Easy to directly interpret

CONS

- Need fixed vocab size
- Doesn't capture semantic meaning of words
- Does not conserve order

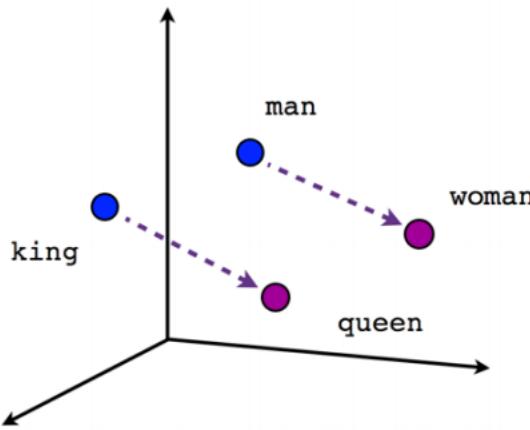
WORD VECTORS

Dream

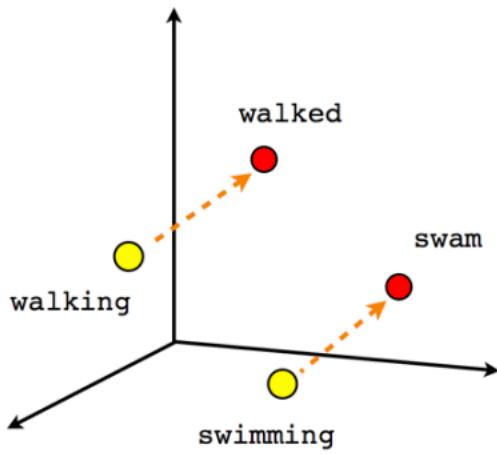
Inspire

Courage

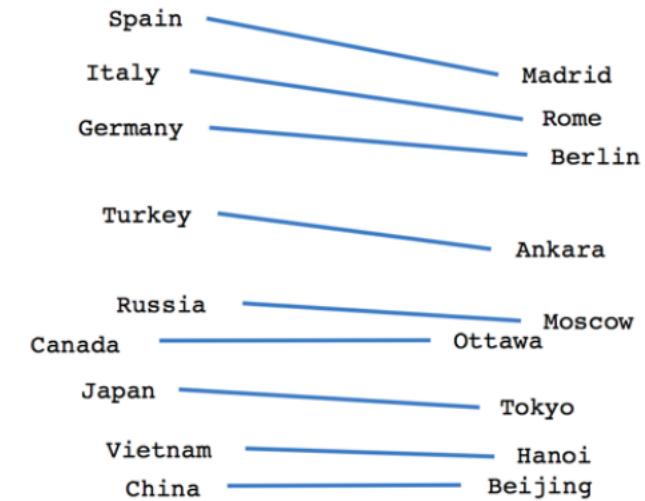
WORD VECTORS



Male-Female



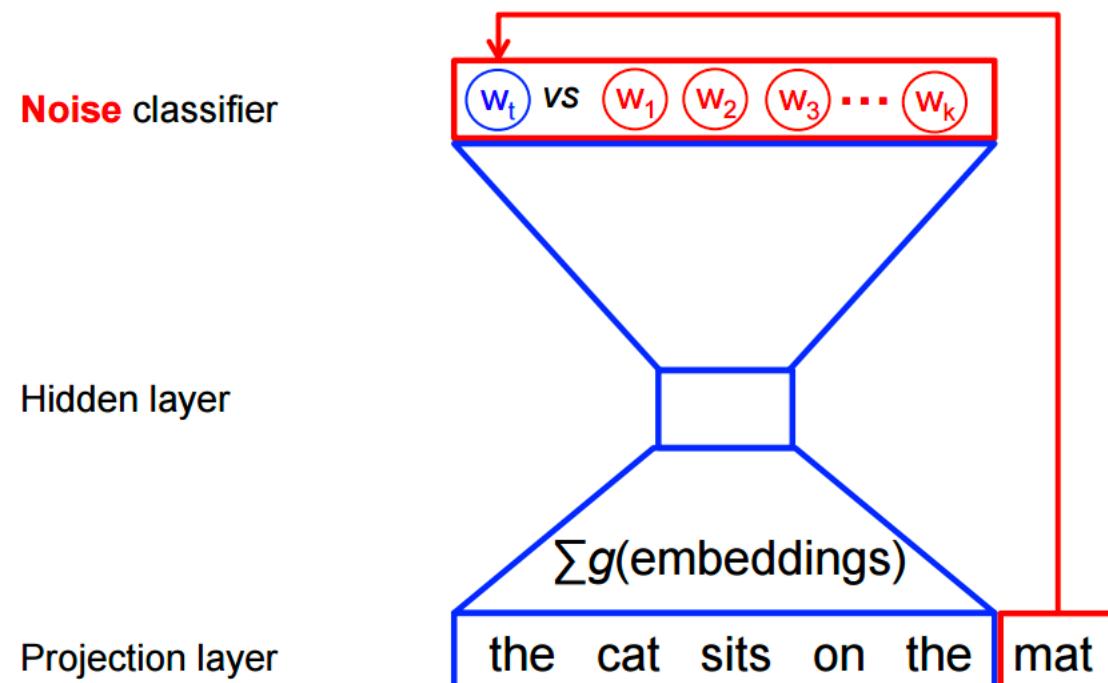
Verb tense



Country-Capital

Mikolov et al., 2013

THE DIRTY SECRET



<https://www.tensorflow.org/tutorials/word2vec>

WORD2VEC BAG OF WORDS

Word2Vec Embedding

Mary is hungry for apples. →
John is happy he is not hungry for apples. →

Lookup

Word	Vector
Mary	[.24, 1.16, .12,..., 1.97, .23, .12]
is	[.55, .11, .15,..., .65, 1.30, 2.42]
hungry	...
happy	...
for	...
apples	...
not	...
John	[.68, 2.02, .24,..., 1.12, .43, .27]
he	[.21, 1.07, 1.01,..., 0.94, .07, .16]

Average

→ [.24, 1.16, .12,..., 1.97, .23, .12]
→ [.43, 1.46, .55,..., 1.13, .53, .78]

Classifier

WORD2VEC BAG OF WORDS

PROS

- Captures semantics (similar words close together)
- Dense representation
- Benefit from having been trained on millions of sentences
- Quick embedding (lookup table)

CONS

- Does not capture syntax
- Loses direct interpretability

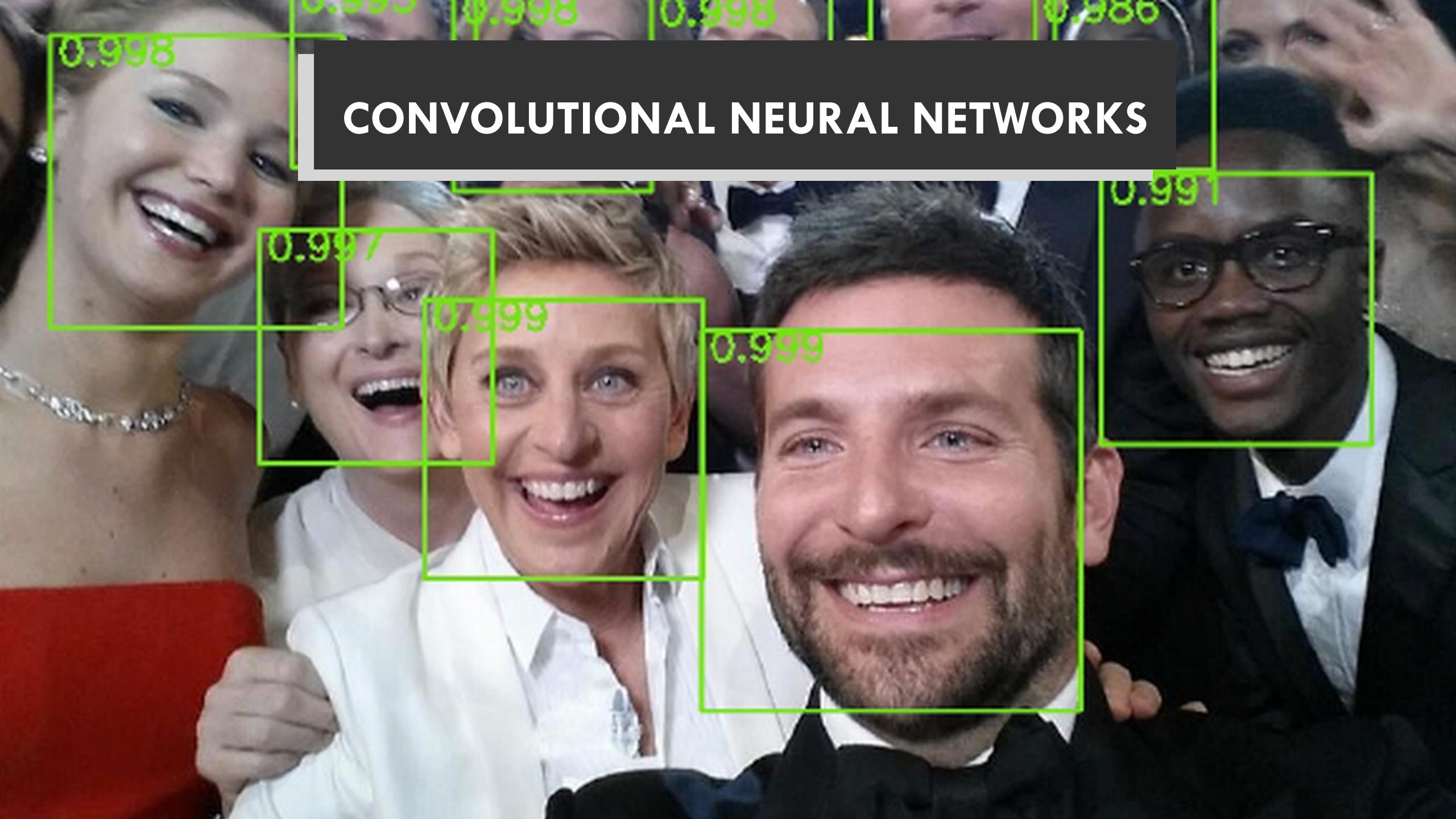
CAN WE DO BETTER?

For Embeddings

- Not really, risk simply overfitting

For Other Tasks

- Translation
- Dialog
- Very tricky classification



0.998

CONVOLUTIONAL NEURAL NETWORKS

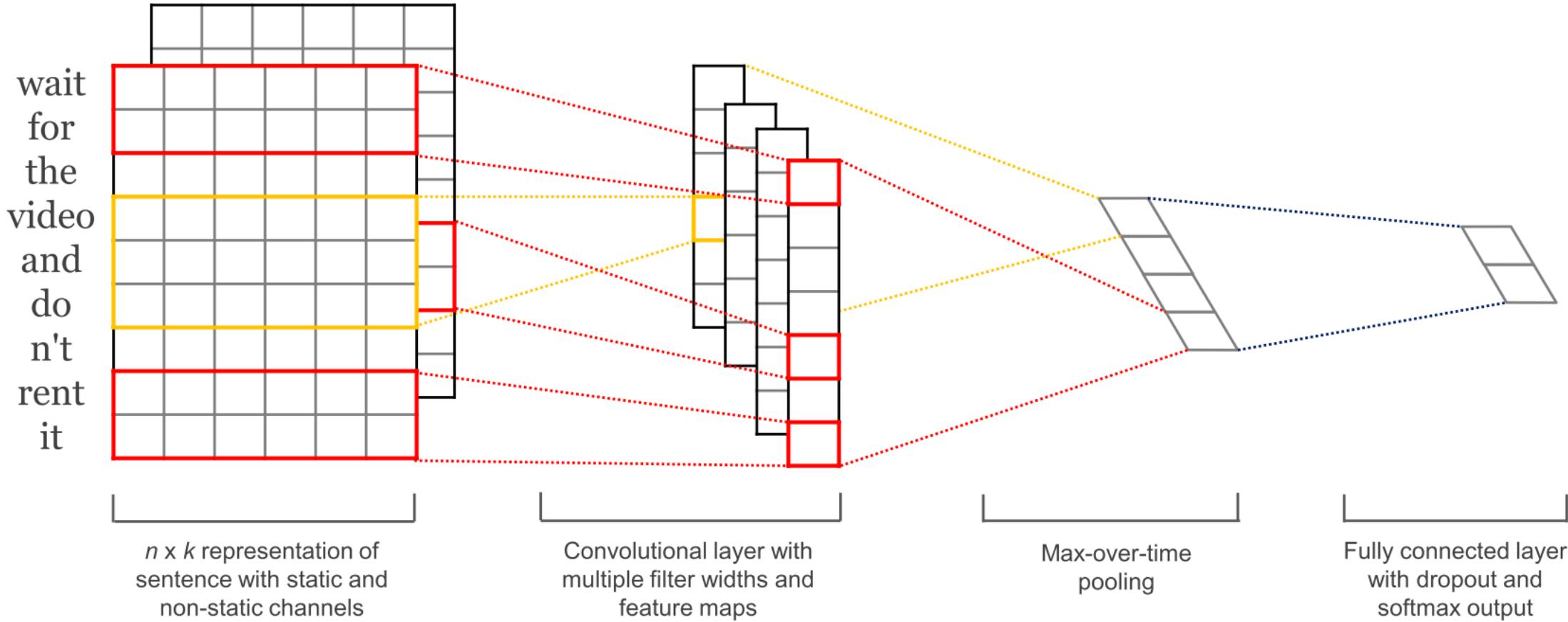
0.997

0.999

0.999

0.991

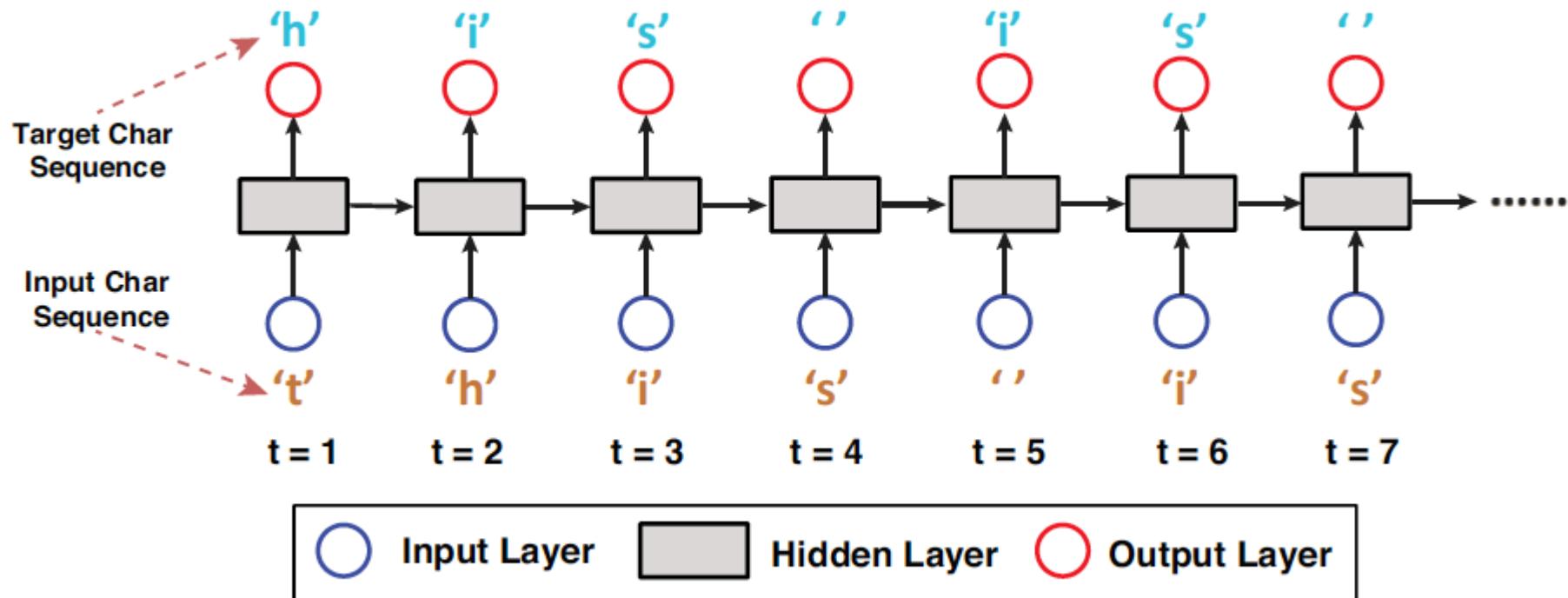
CONVOLUTIONAL NEURAL NETWORKS



RECURRENT NEURAL NETWORKS

Buch der Geheimnisse.
Das
Eine Sammlung
zweihundert und sechzig
jetischen und sommerlichen
Wider Sprüche
der Menschen.

RECURRENT NEURAL NETWORKS



RECURRENT NEURAL NETWORKS

PROS

- Keeps semantic meaning
- Can adapt and learn embeddings

CONS

- Needs much more compute
- Slower
- Doesn't directly produce embeddings

WHAT HAVE YOU LEARNED?

1. Build a simple prototype
2. Inspect and validate the embedding
3. Build a more complex model
4. Evaluate and validate your predictions

THANK YOU!

I'm Emmanuel Ameisen

 @emmanuelameisen

 /in/ameisen

Find out more and apply:
insightdata.ai

Read our blog:
blog.insightdatascience.com

Special thanks to:

- Our Alumni and Fellow community and staff.
- Our industry partners.
- Paperspace for providing us with a great compute platform