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

INSIGHT

@emmanuelameisen

/in/ameisen

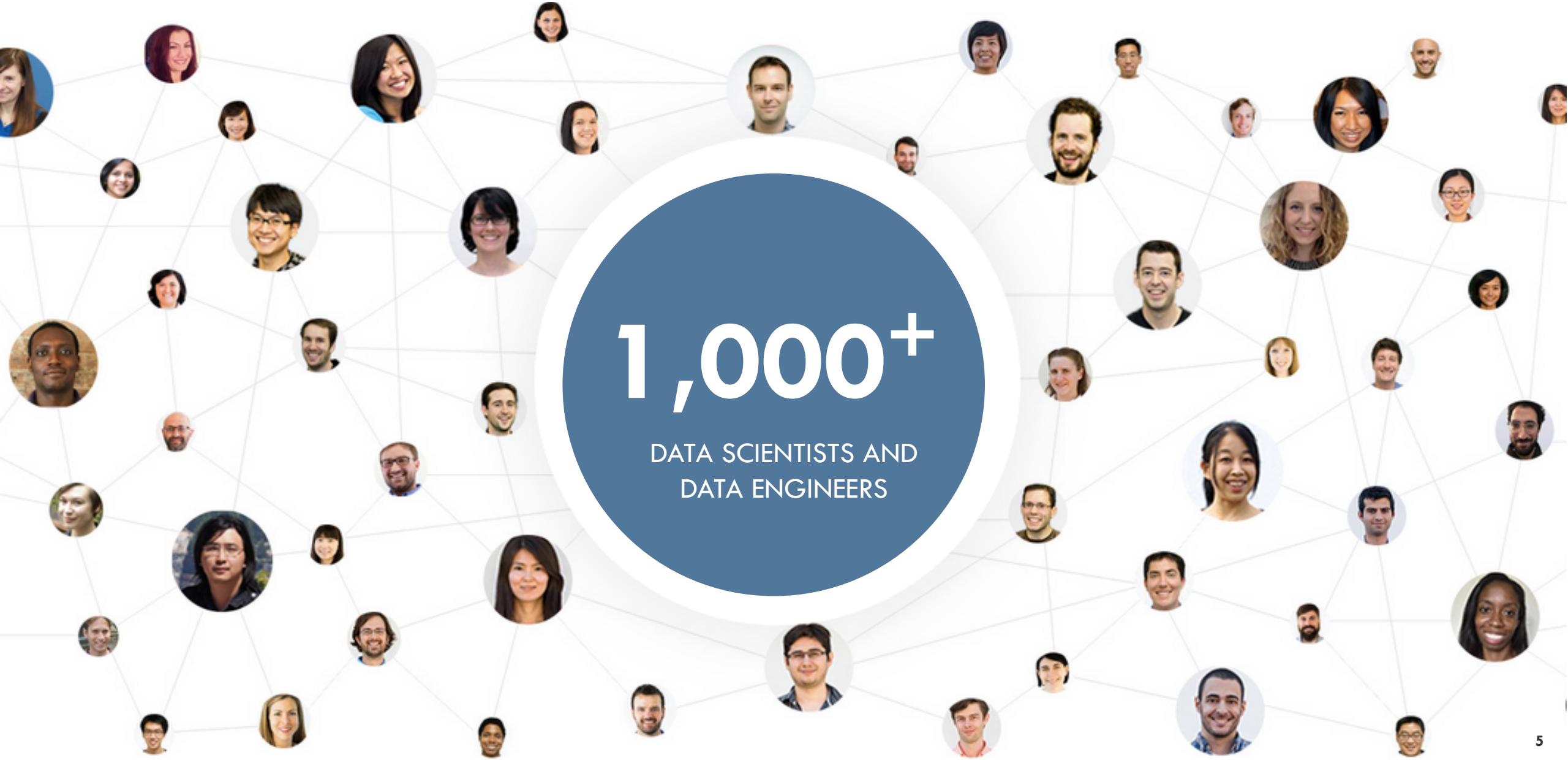
PAST EXPERIENCE

zipcar	LOCAL MOTION	CentraleSupélec	UNIVERSITÉ PARIS SUD	ESCP EUROPE
DATA SCIENTIST	DATA SCIENTIST	MSC. COMPUTER SCIENCE	MSC. ARTIFICIAL INTELLIGENCE	MSC. MANAGEMENT

INSIGHT FELLOWS ARE DATA SCIENTISTS AND DATA ENGINEERS EVERYWHERE

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

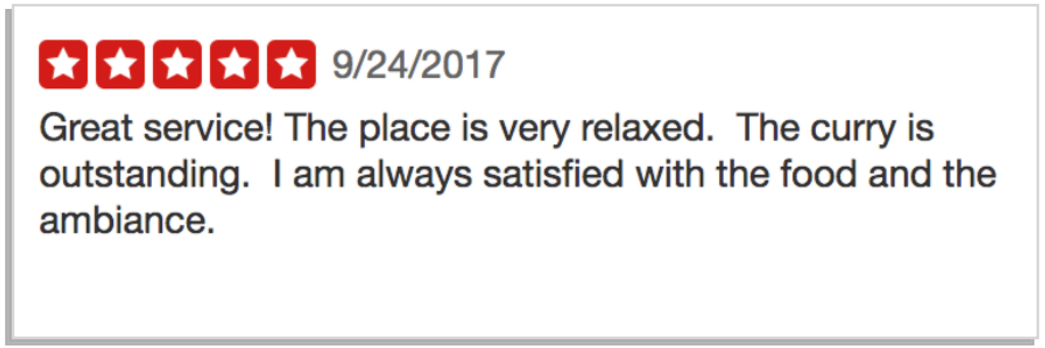
NETWORK OF INSIGHT ALUMNI



EXAMPLE PROJECTS

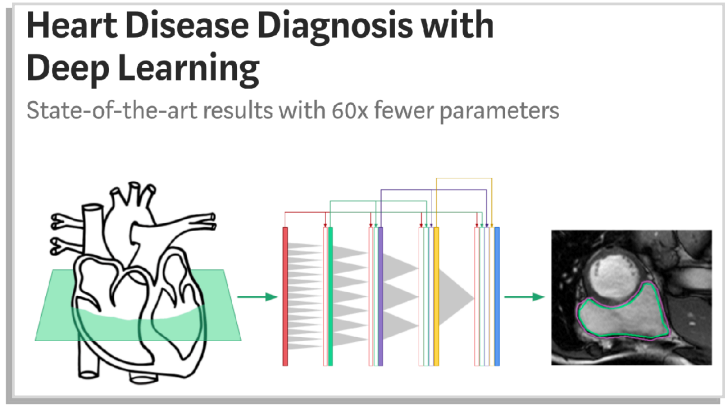
FASHION CLASSIFIER

AUTOMATIC REVIEW GENERATION

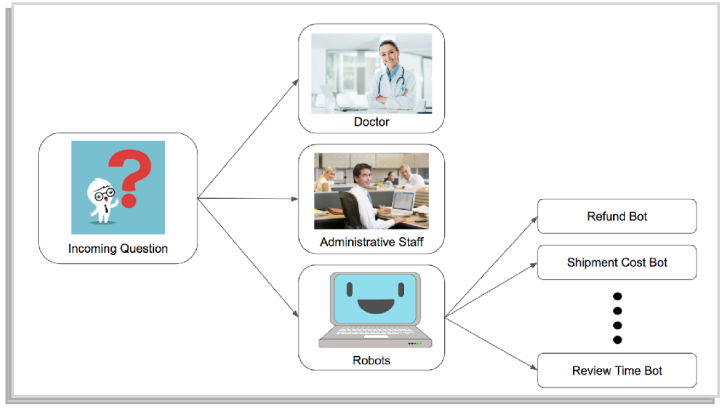


READING TEXT IN VIDEOS

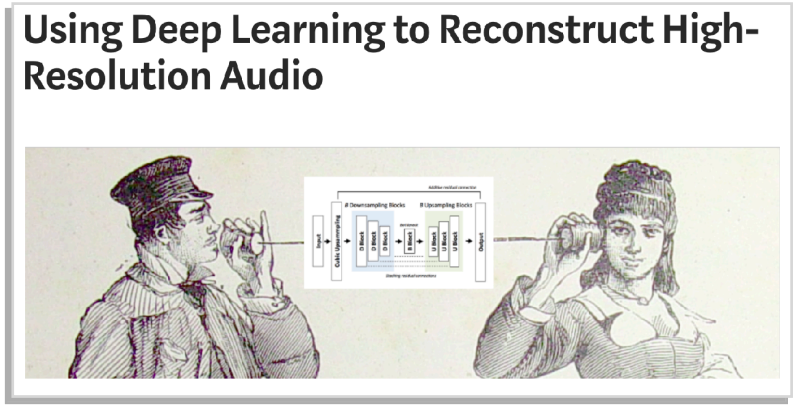
HEART SEGMENTATION



SUPPORT REQUEST CLASSIFICATION



SPEECH UNSAMPLING



CONCRETE NLP SOLUTIONS

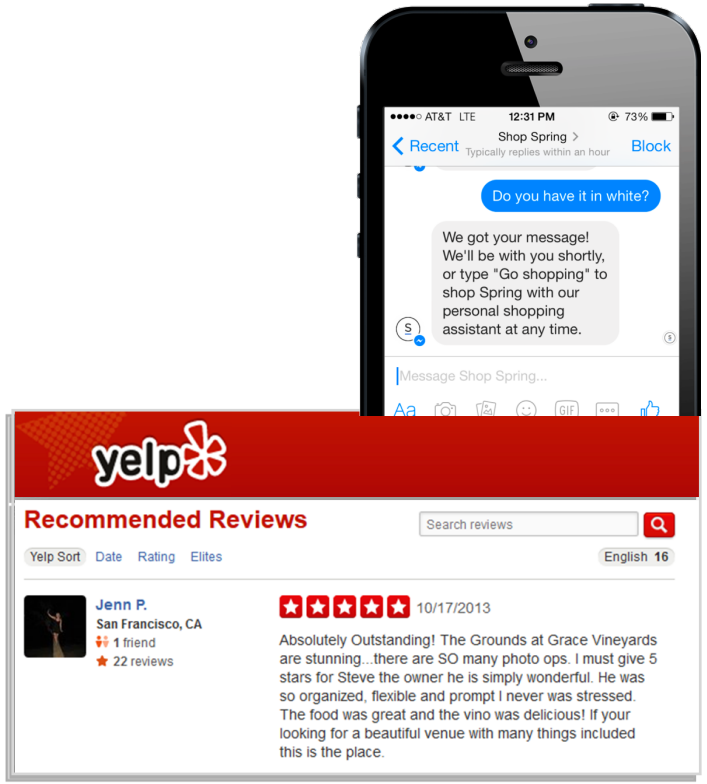
- **CLUSTERING**
- **CLASSIFICATION**
- **GENERATION**

A HUGE VALUE TO BE UNLOCKED

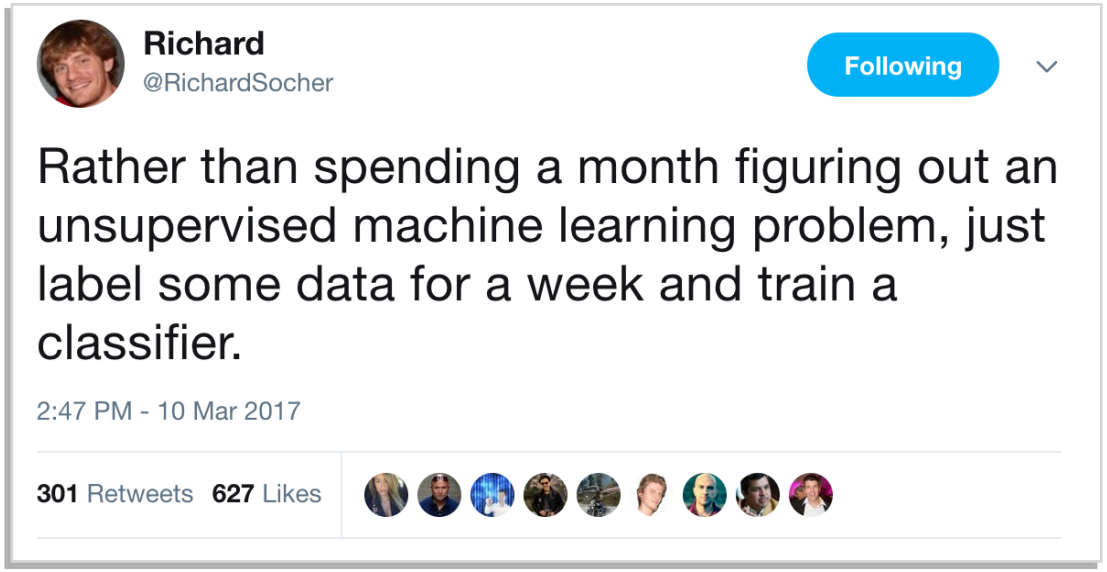
Business Models

facebook

Every Day Uses



ALL YOU NEED IS ~~LOVE~~ LABELS

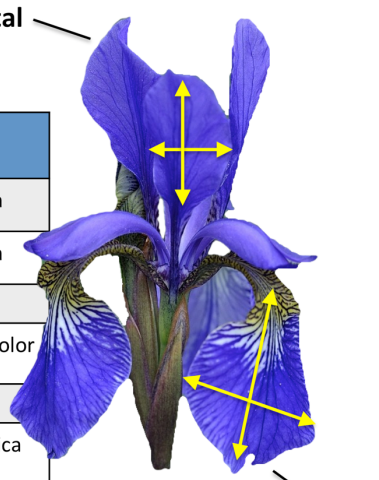


ALL YOU NEED IS ~~CLEAN~~ DATA

HOW DO YOU REPRESENT DATA?

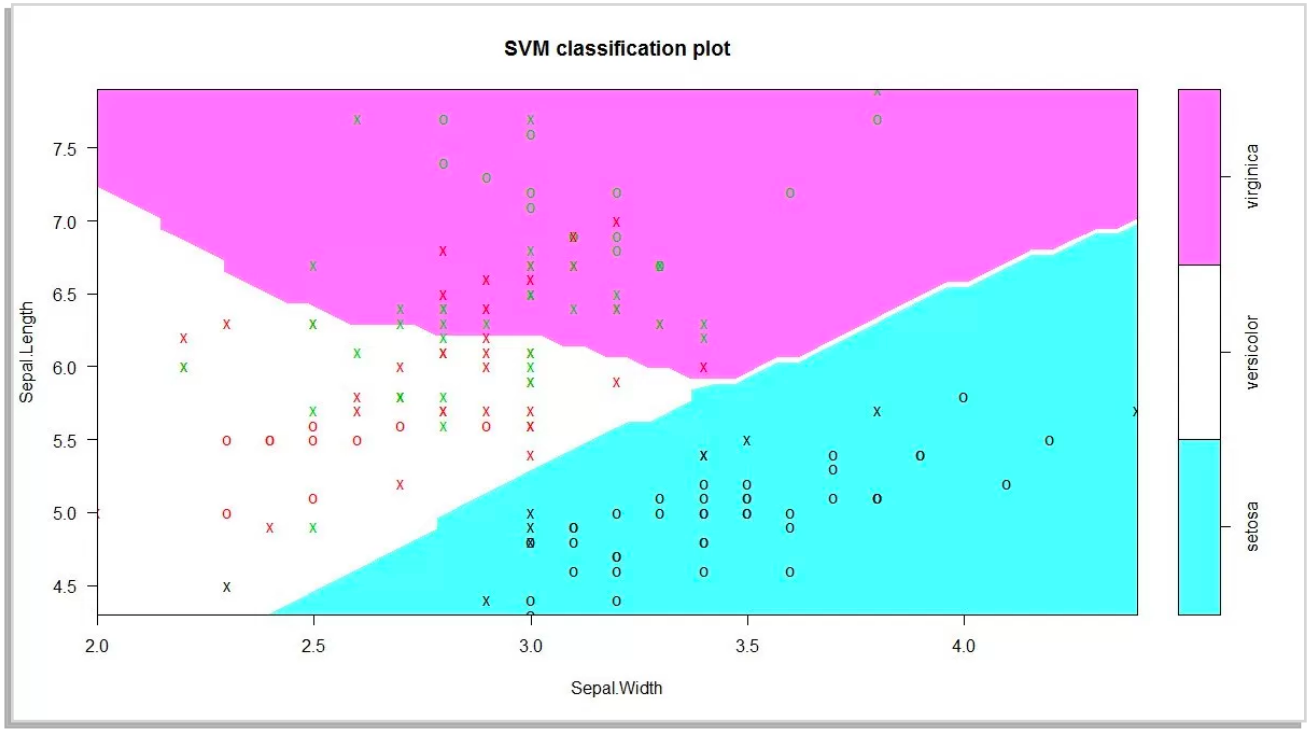
Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica



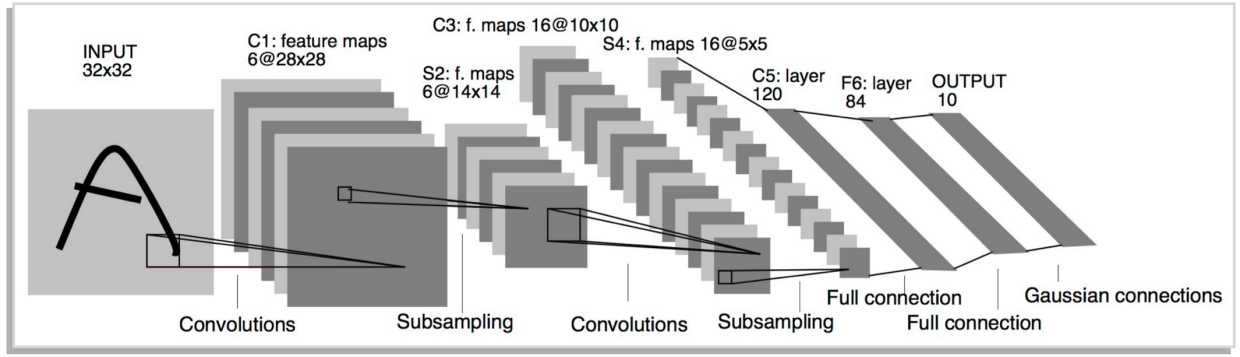
Features
(attributes, measurements, dimensions)

Class labels
(targets)



HOW DO YOU REPRESENT IMAGES?

	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
1	Input																												
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	
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	
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	
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	
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	
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	
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	
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.3	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
10	0.0	0.0	0.0	0.4	0.5	0.9	0.9	0.9	0.9	0.9	0.9	1.0	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	
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	1.0	1.0	0.7	0.1	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	0.8	0.5	0.2	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	
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	0.0	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	
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	0.0	0.3	1.0	1.0	0.9	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	0.0	0.3	1.0	1.0	0.9	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	0.0	0.4	0.6	1.0	1.0	1.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	
18	0.0	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	0.6	0.0	0.0	0.0	0.0	0.0	0.0	
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.7	1.0	1.0	1.0	1.0	0.9	0.8	0.8	0.3	0.3	0.8	1.0	1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	
20	0.0	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	0.0	0.8	1.0	0.9	0.2	0.0	0.0	0.0	0.0	0.0	0.0	
21	0.0	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	
22	0.0	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	
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	0.0	0.5	1.0	0.9	0.2	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	
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	0.9	0.2	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	0.3	1.0	0.9	0.3	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	0.8	1.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
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	0.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
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	
30																													



HOW DO YOU REPRESENT WORDS?

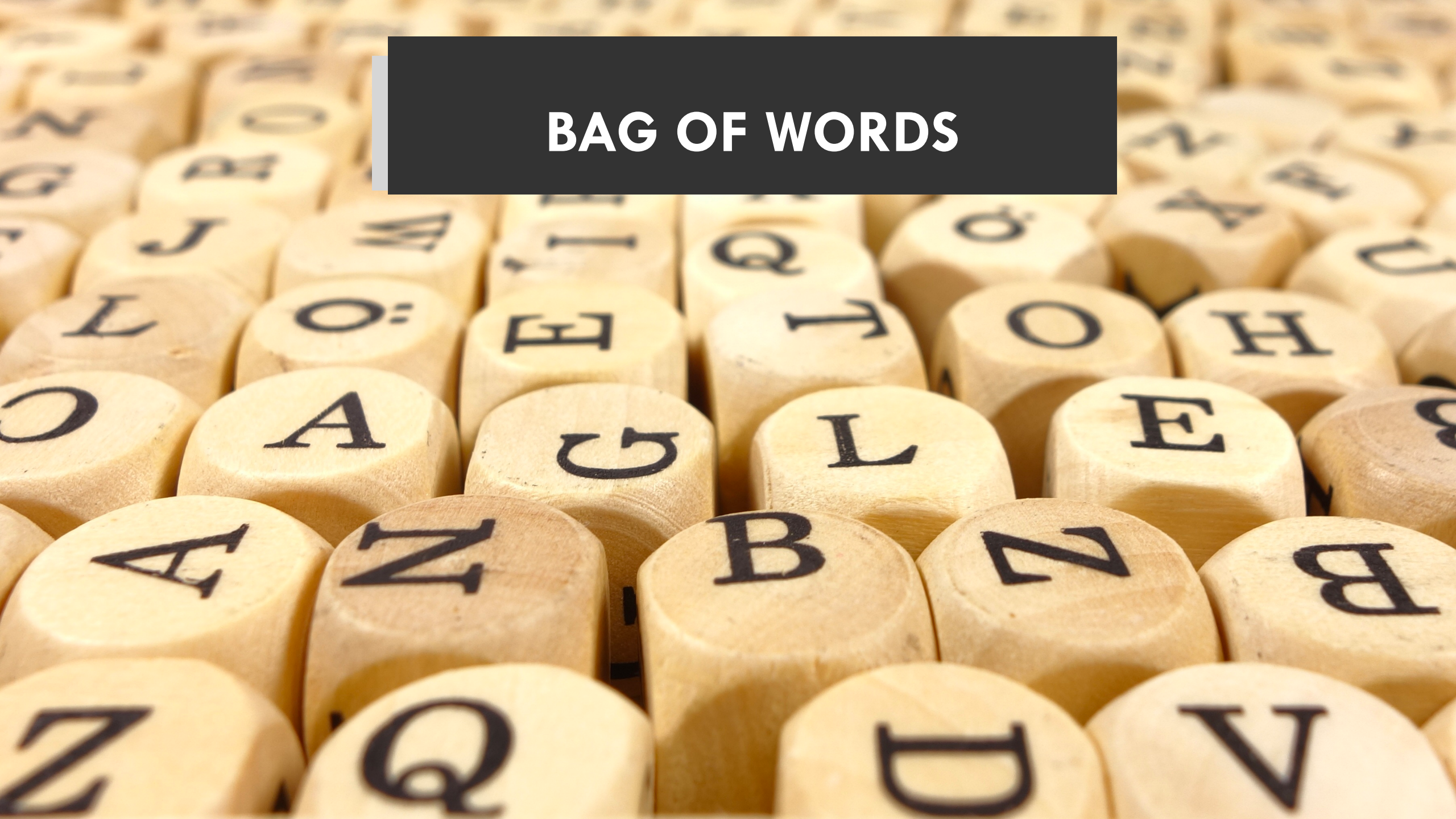
ASCII

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

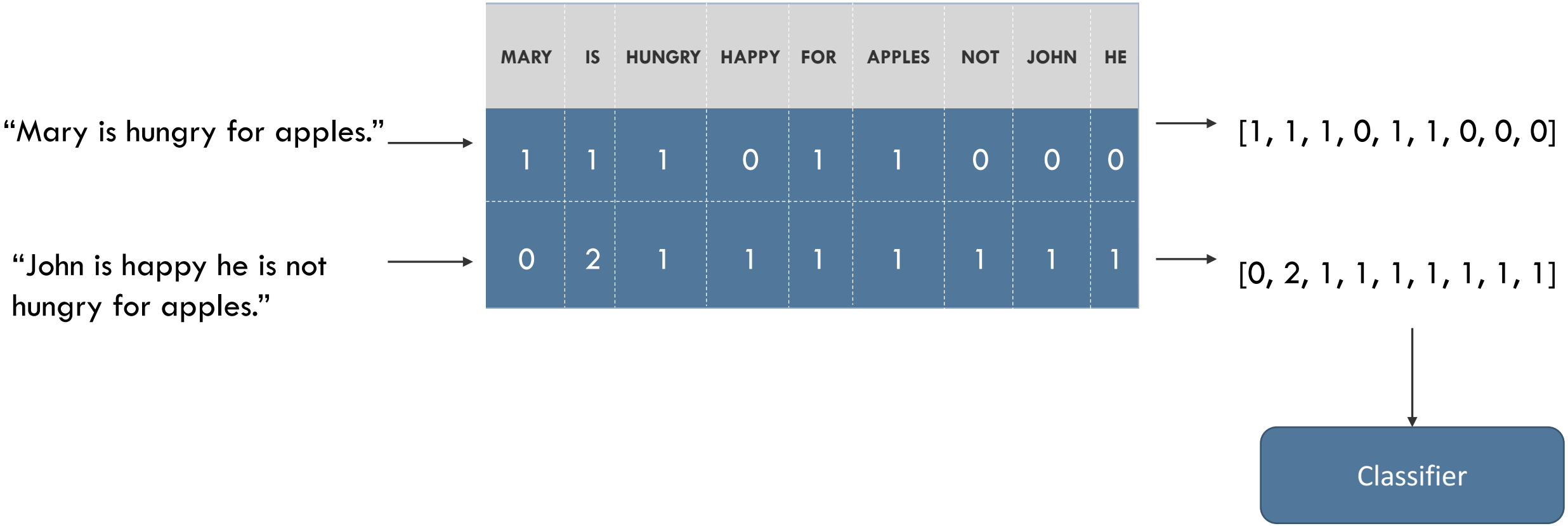
- DISCRETE
- NEED TO LEARN LANGUAGE AND WORDS
- HOW DO YOU REPRESENT A WHOLE SENTENCE

WE NEED A SENTENCE LEVEL REPRESENTATION

BAG OF WORDS



BAG OF WORDS
(AKA ONE HOT EMBEDDING)



BAG OF WORDS

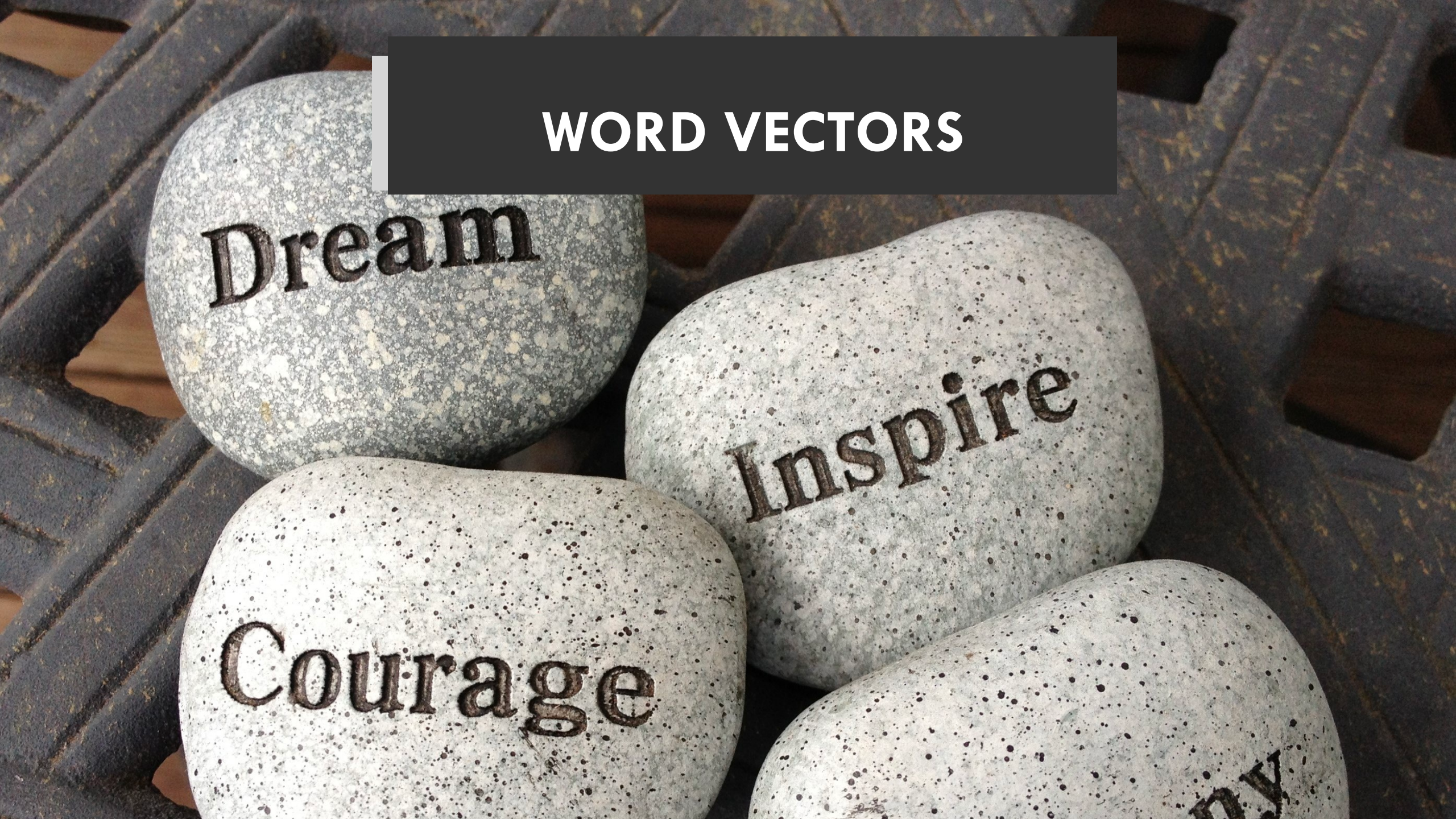
PROS

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

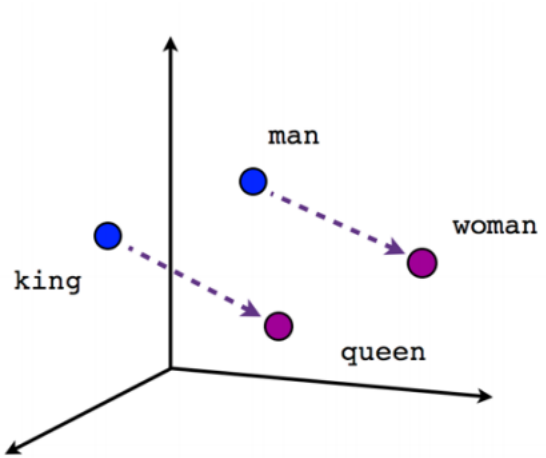
CONS

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

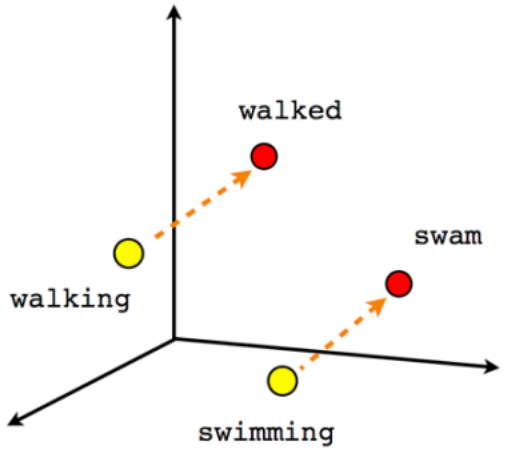
WORD VECTORS



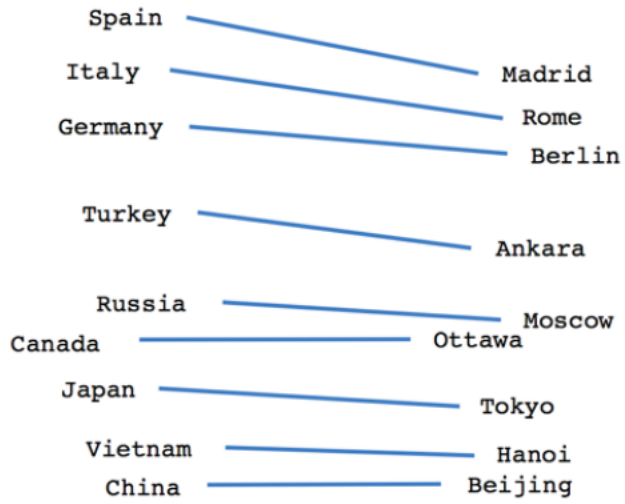
WORD VECTORS



Male-Female



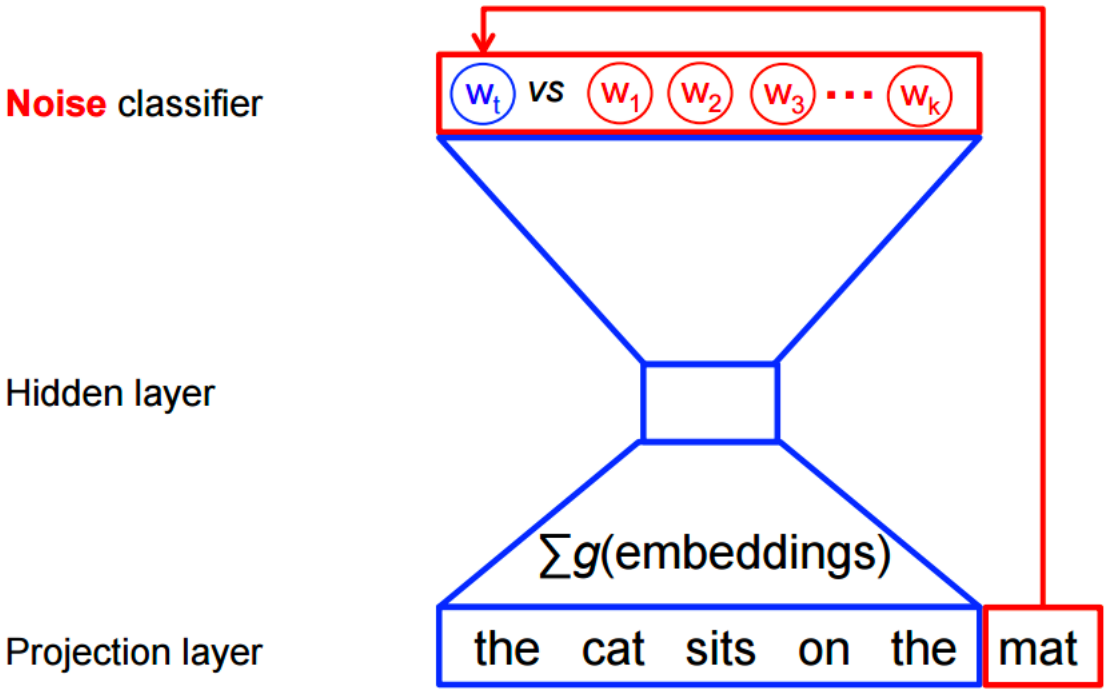
Verb tense



Country-Capital

Mikolov et al., 2013

THE DIRTY SECRET



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

WORD2VEC BAG OF WORDS

Word2Vec Embedding

Lookup

Mary is hungry for apples.

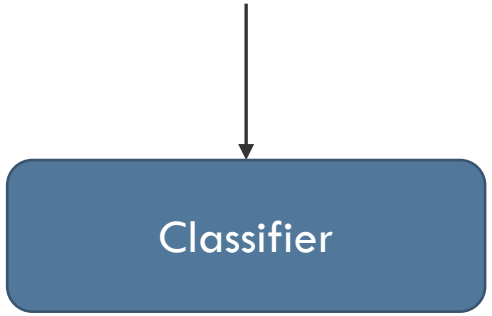
John is happy he is not hungry for apples.

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]



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?

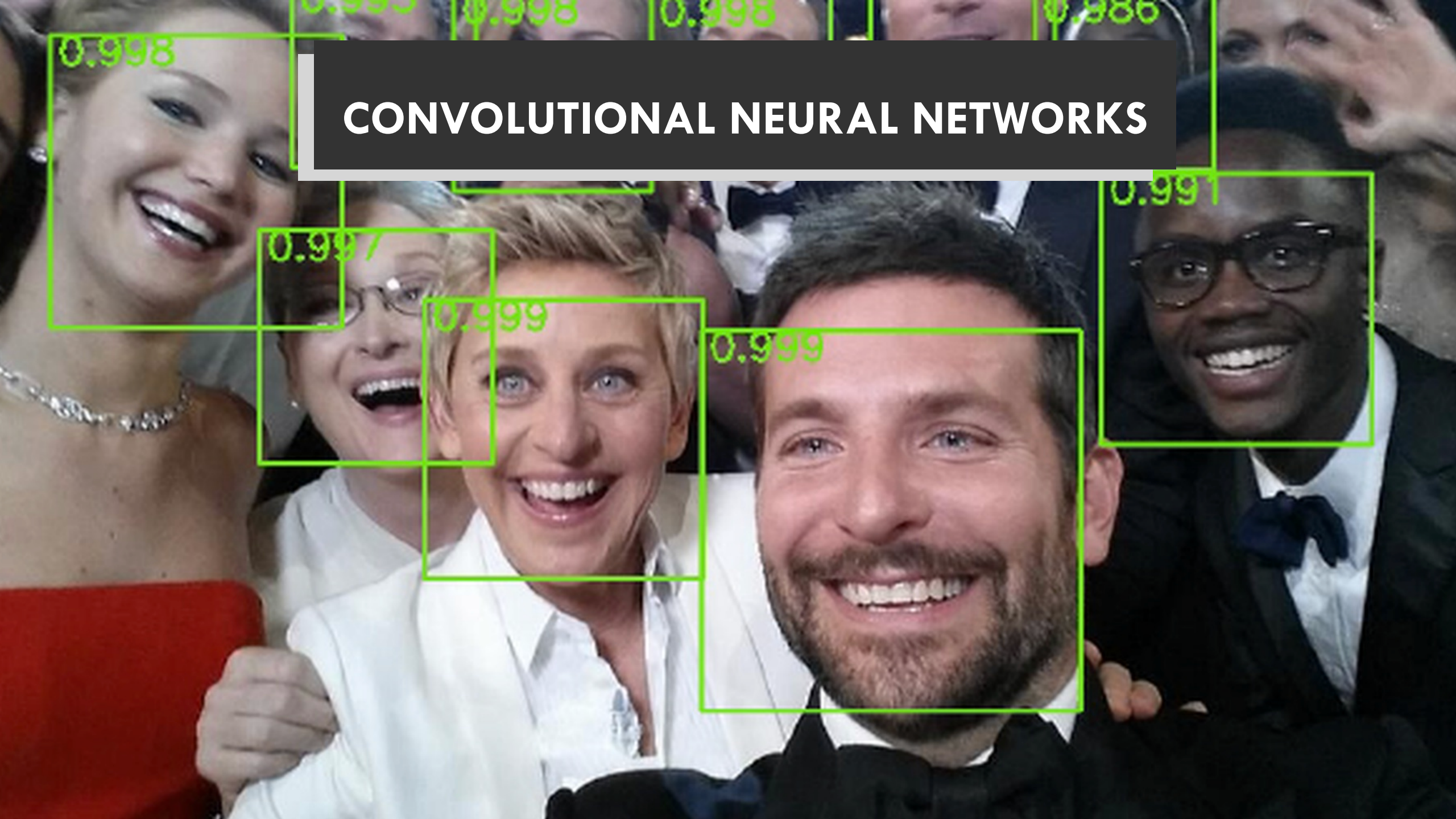
For Embeddings

- Not really, risk simply overfitting

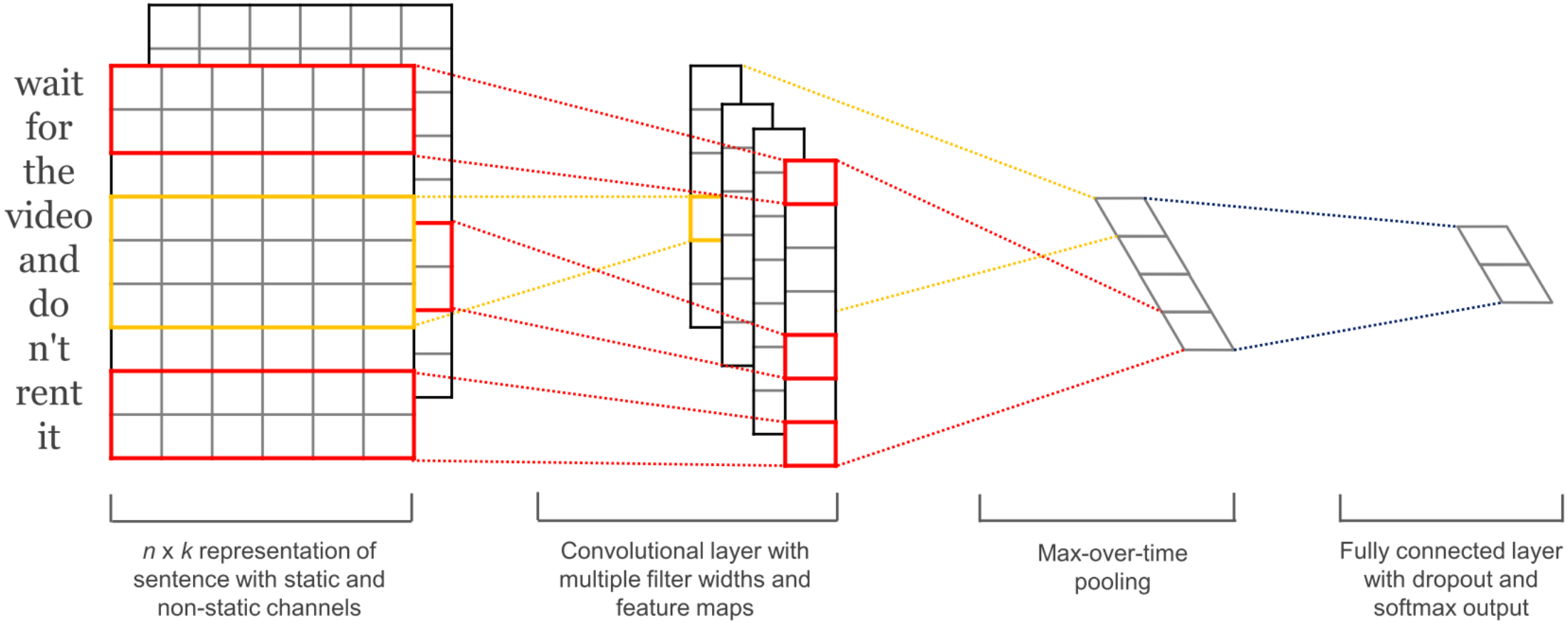
For Other Tasks

- Translation
- Dialog
- Very tricky classification

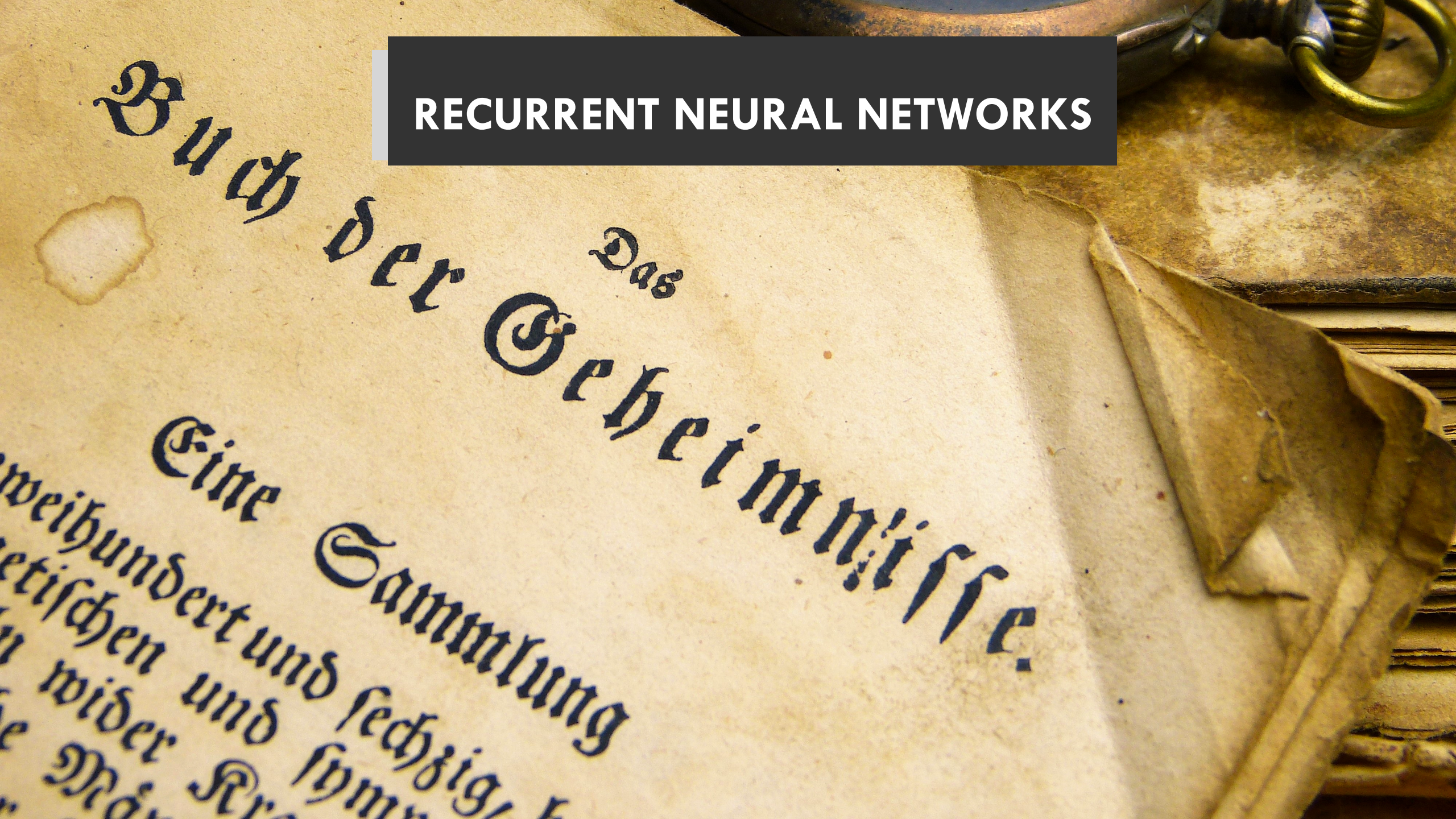
CONVOLUTIONAL NEURAL NETWORKS



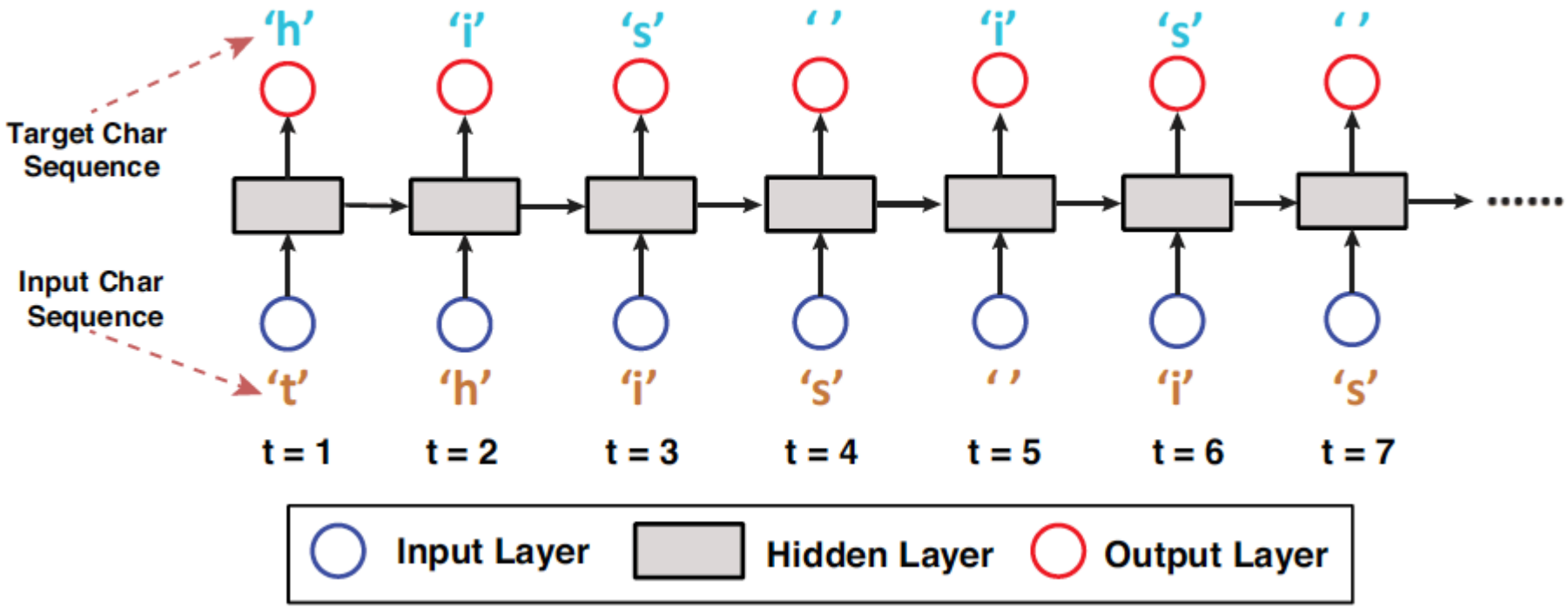
CONVOLUTIONAL NEURAL NETWORKS



RECURRENT NEURAL NETWORKS



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