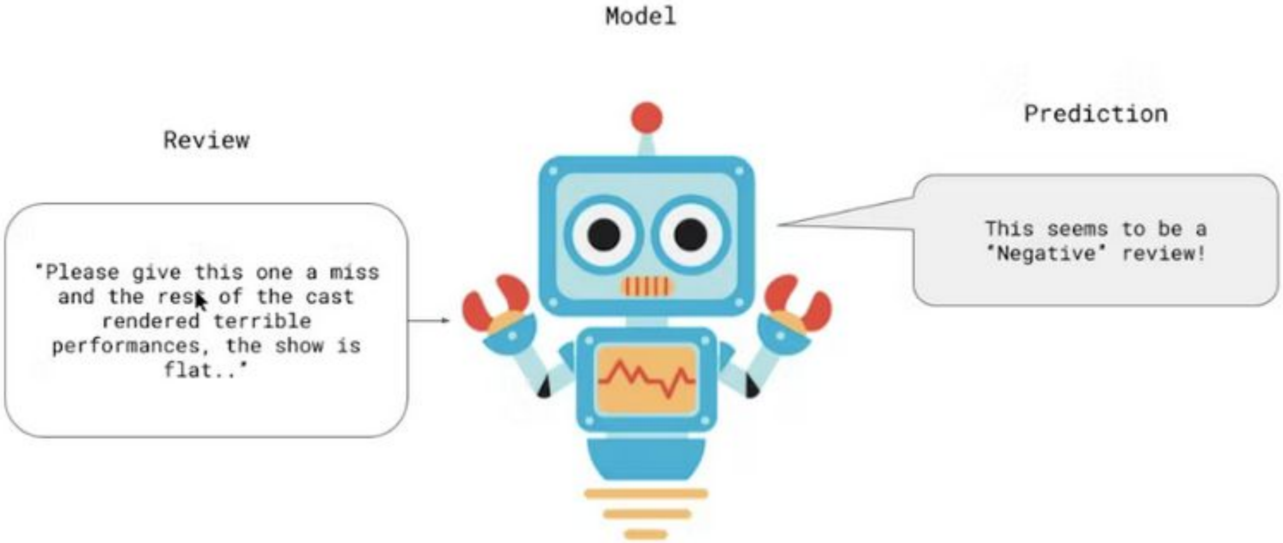
Sentiment Analysis



Model

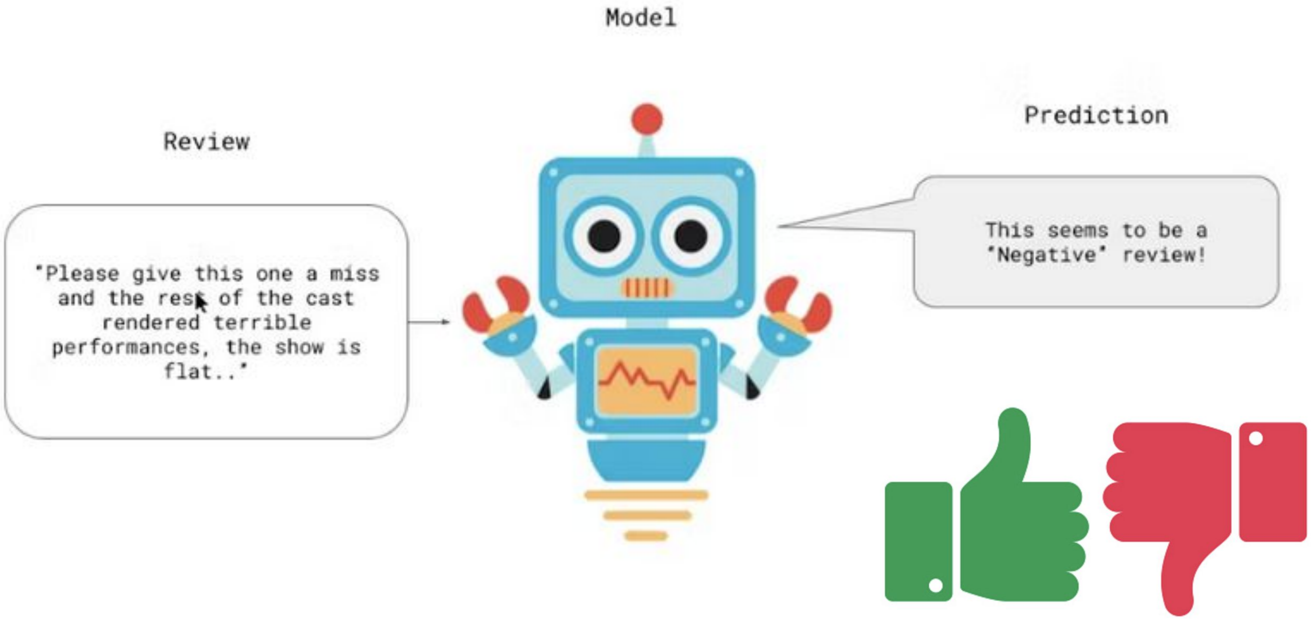
Review

"Please give this one a miss and the rest of the cast rendered terrible performances, the show is flat.."

Prediction

This seems to be a "Negative" review!

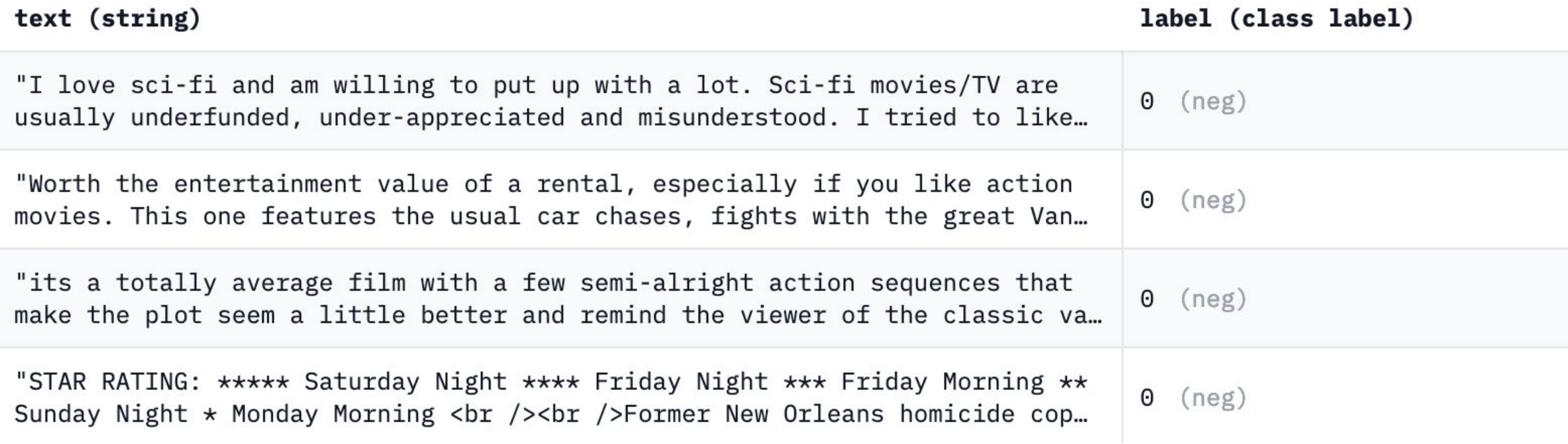
Binary classification



"Please give this one a miss and the rest of the cast rendered terrible performances, the show is flat.."

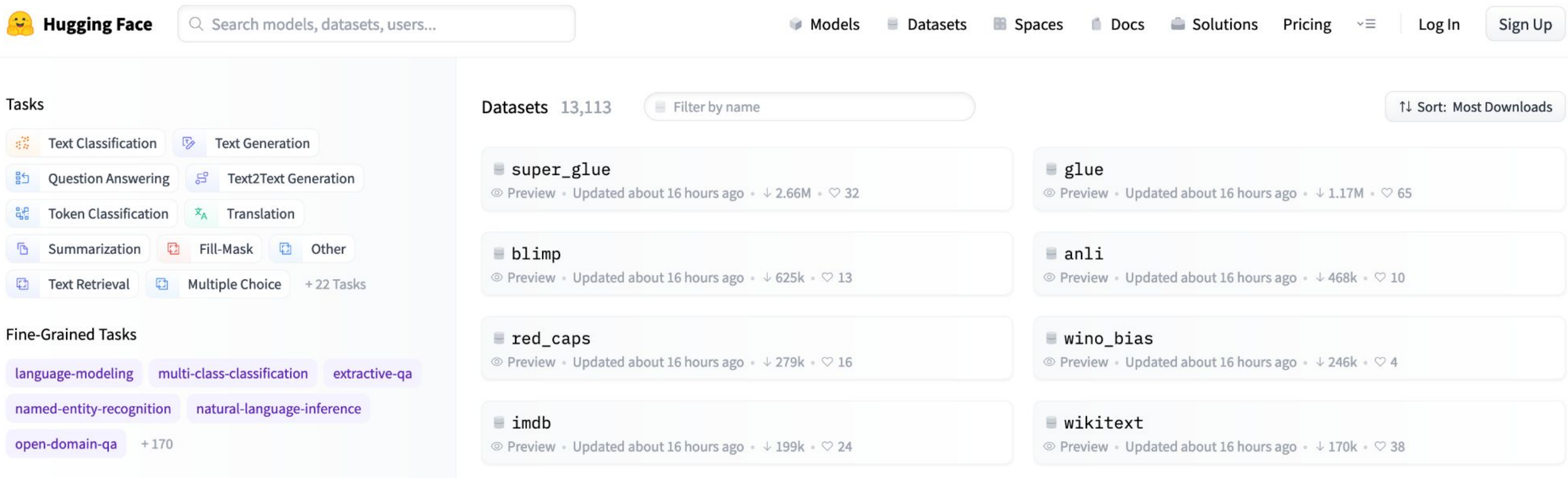
This seems to be a "Negative" review!

Labeled dataset



text (string)	label (class label)
"I love sci-fi and am willing to put up with a lot. Sci-fi movies/TV are usually underfunded, under-appreciated and misunderstood. I tried to like…	0 (neg)
"Worth the entertainment value of a rental, especially if you like action movies. This one features the usual car chases, fights with the great Van…	0 (neg)
"its a totally average film with a few semi-alright action sequences that make the plot seem a little better and remind the viewer of the classic va…	0 (neg)
"STAR RATING: ***** Saturday Night **** Friday Night *** Friday Morning ** Sunday Night * Monday Morning <br /><br />Former New Orleans homicide cop…	0 (neg)

Labeled dataset



Tasks

Text Classification | Text Generation | Question Answering | Text2Text Generation | Token Classification | Translation | Summarization | Fill-Mask | Other | Text Retrieval | Multiple Choice | + 22 Tasks

Fine-Grained Tasks

language-modeling | multi-class-classification | extractive-qa | named-entity-recognition | natural-language-inference | open-domain-qa | + 170

Datasets 13,113

Filter by name

Sort: Most Downloads

super_glue
Preview • Updated about 16 hours ago • ↓ 2.66M • ♡ 32

glue
Preview • Updated about 16 hours ago • ↓ 1.17M • ♡ 65

blimp
Preview • Updated about 16 hours ago • ↓ 625k • ♡ 13

anli
Preview • Updated about 16 hours ago • ↓ 468k • ♡ 10

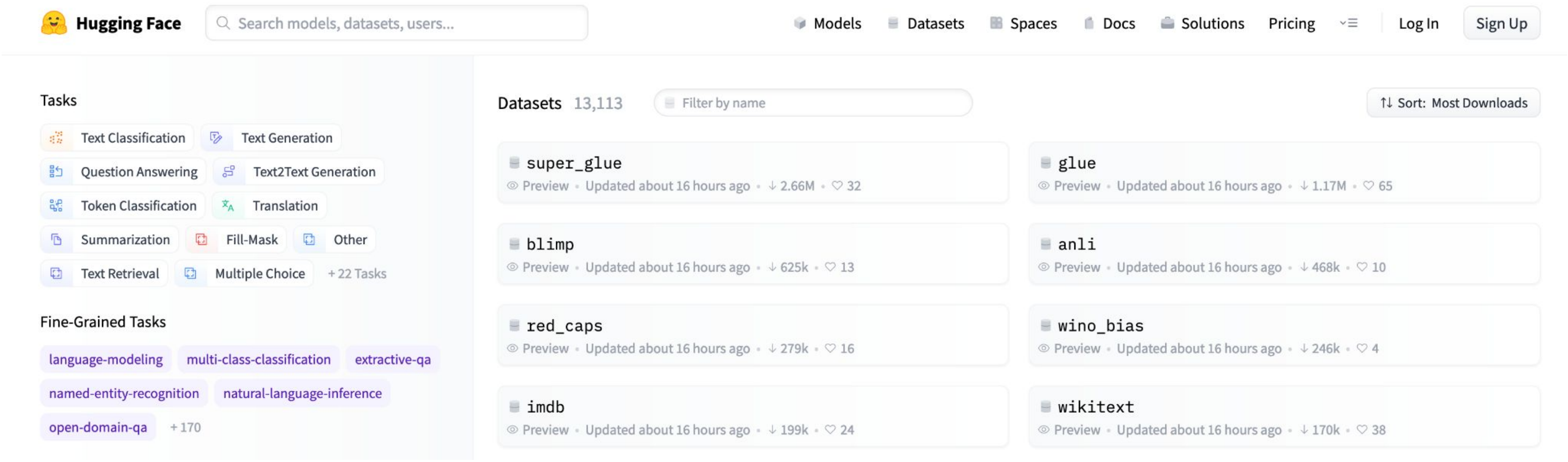
red_caps
Preview • Updated about 16 hours ago • ↓ 279k • ♡ 16

wino_bias
Preview • Updated about 16 hours ago • ↓ 246k • ♡ 4

imdb
Preview • Updated about 16 hours ago • ↓ 199k • ♡ 24

wikitext
Preview • Updated about 16 hours ago • ↓ 170k • ♡ 38

Labeled dataset



Hugging Face | Search models, datasets, users...

Models · Datasets · Spaces · Docs · Solutions · Pricing · Log In · Sign Up

Tasks

Text Classification · Text Generation · Question Answering · Text2Text Generation · Token Classification · Translation · Summarization · Fill-Mask · Other · Text Retrieval · Multiple Choice · + 22 Tasks

Fine-Grained Tasks

language-modeling · multi-class-classification · extractive-qa · named-entity-recognition · natural-language-inference · open-domain-qa · + 170

Datasets 13,113 | Filter by name | Sort: Most Downloads

super_glue
Preview • Updated about 16 hours ago • ↓ 2.66M • ♡ 32

glue
Preview • Updated about 16 hours ago • ↓ 1.17M • ♡ 65

blimp
Preview • Updated about 16 hours ago • ↓ 625k • ♡ 13

anli
Preview • Updated about 16 hours ago • ↓ 468k • ♡ 10

red_caps
Preview • Updated about 16 hours ago • ↓ 279k • ♡ 16

wino_bias
Preview • Updated about 16 hours ago • ↓ 246k • ♡ 4

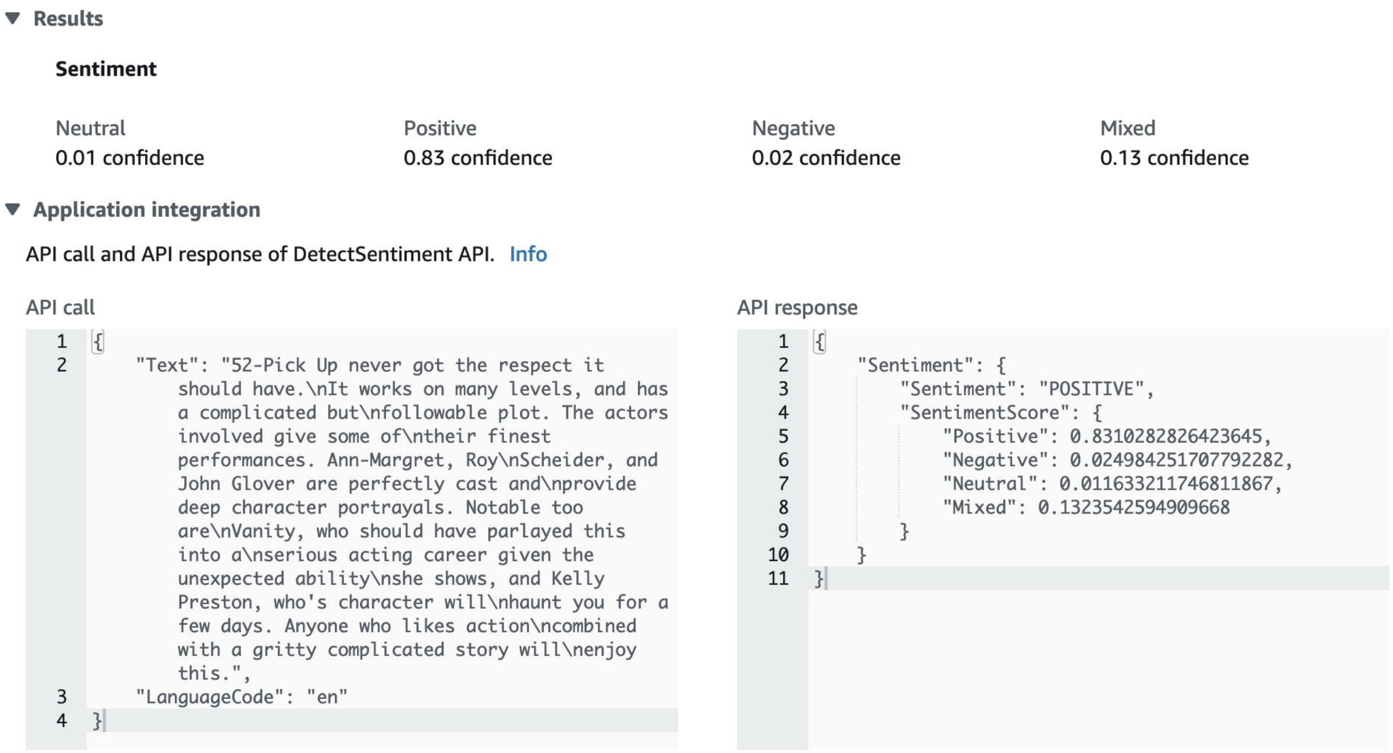
imdb
Preview • Updated about 16 hours ago • ↓ 199k • ♡ 24

wikitext
Preview • Updated about 16 hours ago • ↓ 170k • ♡ 38

NLP - as a Service

Sentiment Analysis - as a Service

Amazon Comprehend: Sentiment Analysis API



▼ **Results**

Sentiment

Neutral	Positive	Negative	Mixed
0.01 confidence	0.83 confidence	0.02 confidence	0.13 confidence

▼ **Application integration**

API call and API response of DetectSentiment API. Info

API call

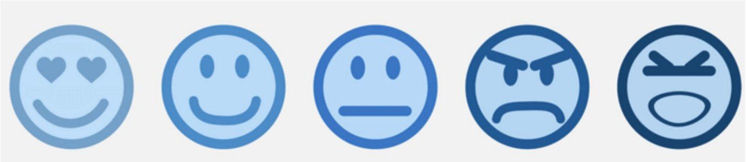
```
{
    "Text": "52-Pick Up never got the respect it
        should have.\nIt works on many levels, and has
        a complicated but\nfollowable plot. The actors
        involved give some of\ntheir finest
        performances. Ann-Margret, Roy\nScheider, and
        John Glover are perfectly cast and\nprovide
        deep character portrayals. Notable too
        are\nVanity, who should have parlayed this
        into a\nserious acting career given the
        unexpected ability\nshe shows, and Kelly
        Preston, who's character will\nhaunt you for a
        few days. Anyone who likes action\ncombined
        with a gritty complicated story will\nenjoy
        this.",
    "LanguageCode": "en"
}
```

API response

```
{
    "Sentiment": {
        "Sentiment": "POSITIVE",
        "SentimentScore": {
            "Positive": 0.8310282826423645,
            "Negative": 0.024984251707792282,
            "Neutral": 0.011633211746811867,
            "Mixed": 0.1323542594909668
        }
    }
}
```

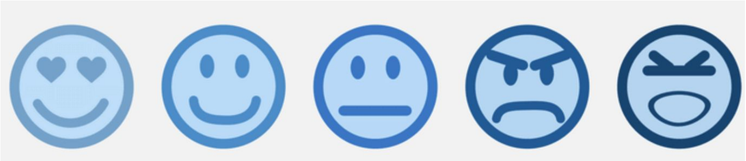
Issues

Issues



- What to predict?
  - there is a vast number of sentiments and many nuances
  - most application simply into few categories: pos/neg/neutral or hate speech/non-hate
  - star rating: 1-5

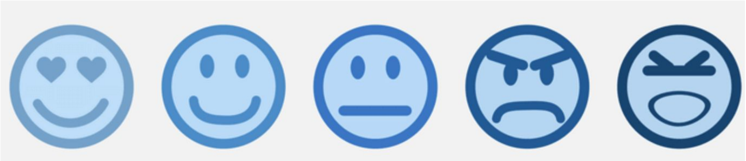
Issues



- What to predict?
  - there is a vast number of sentiments and many nuances
  - most application simply into few categories: pos/neg/neutral or hate speech/non-hate
  - star rating: 1-5

- Subjectivity?
  - depends on culture, individuals, context
  - labelled data might be biased
  - extremes are easier to detect/agree upon - more interesting
  - Negative / Hate might be more important - minimize error

Issues

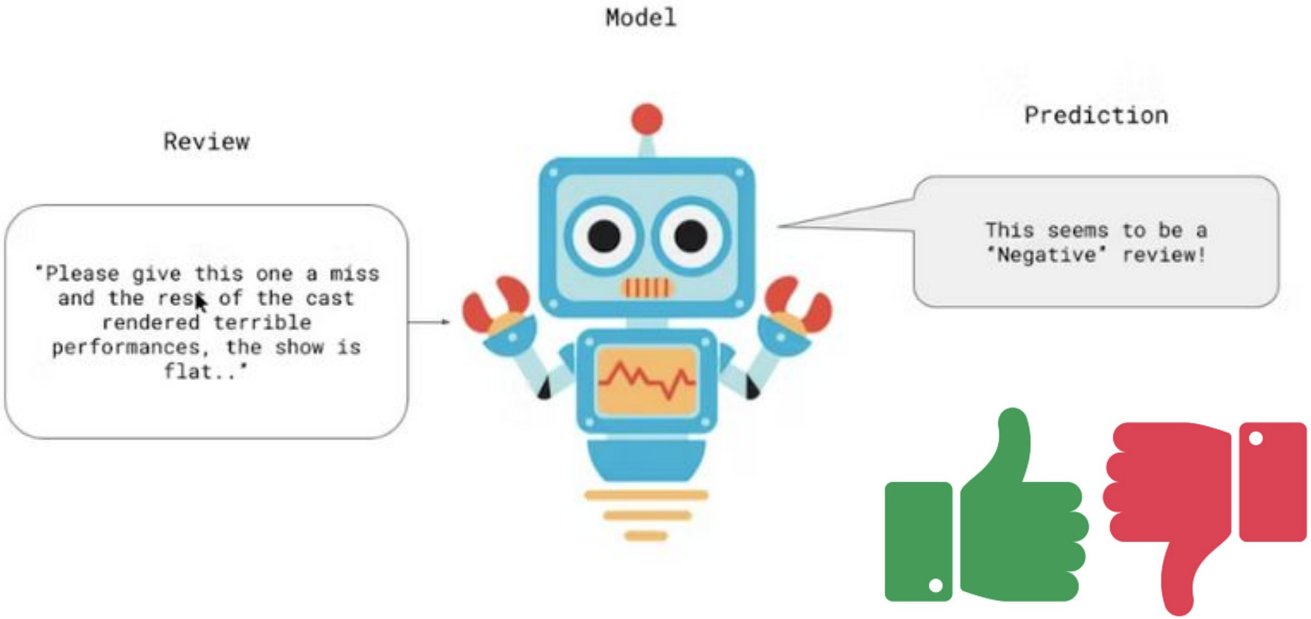


- What to predict?
  - there is a vast number of sentiments and many nuances
  - most application simply into few categories: pos/neg/neutral or hate speech/non-hate
  - star rating: 1-5
- Subjectivity?
  - depends on culture, individuals, context
  - labelled data might be biased
  - extremes are easier to detect/agree upon - more interesting
  - Negative / Hate might be more important - minimize error
- How to aggregate?
  - Sentiment can change throughout a sentence/paragraph
  - Word - Sentence - Aspect/Topic - Document - levels - how to combine sentiment scores?
  - Simple average, weighted average, majority voting, Dempster-Shafer algorithm, uninorm operators

Evaluation

if you can’t evaluate,
maybe you shouldn't even start

Binary classification



Model

Review

"Please give this one a miss and the rest of the cast rendered terrible performances, the show is flat.."

Prediction

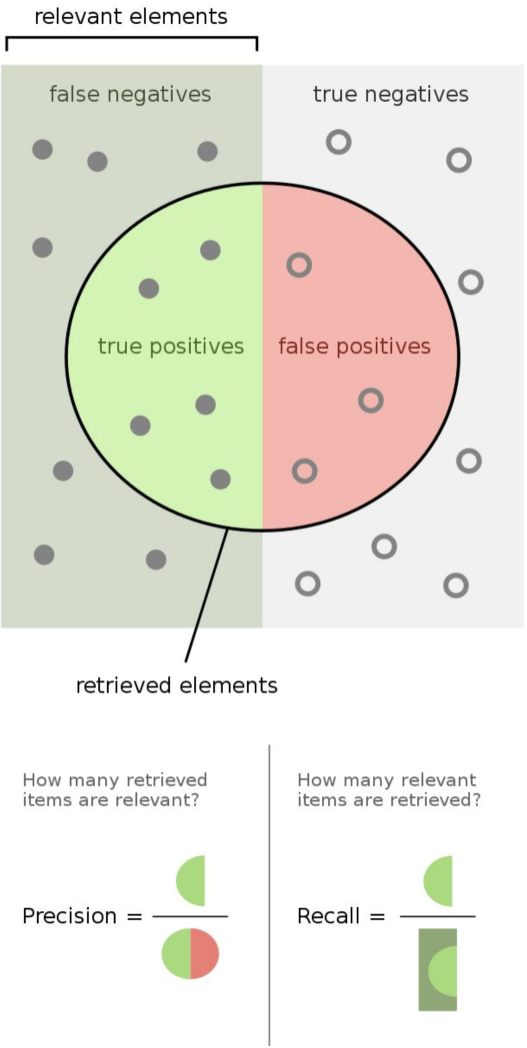
This seems to be a "Negative" review!

Evaluation

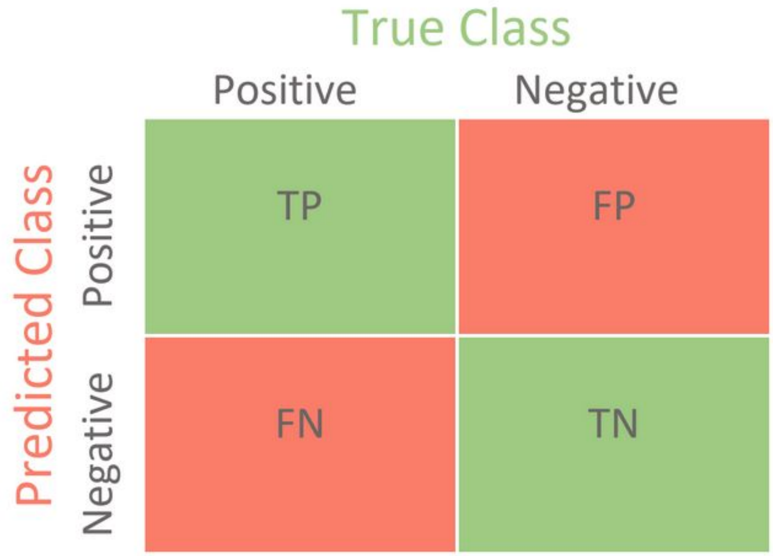
$$accuracy = \frac{correct}{correct + incorrect}$$

Evaluation

Evaluation


relevant elements
false negatives
true negatives
true positives
false positives
retrieved elements
How many retrieved items are relevant?
How many relevant items are retrieved?
Precision =
Recall =

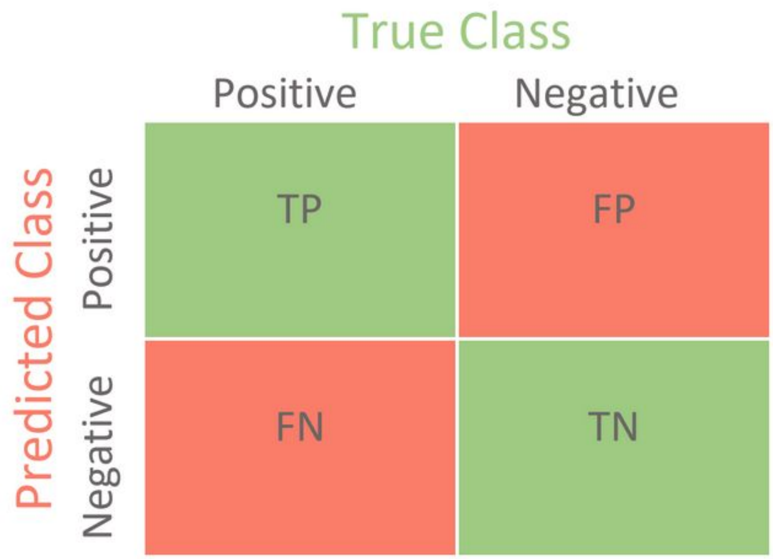
Evaluation



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Evaluation

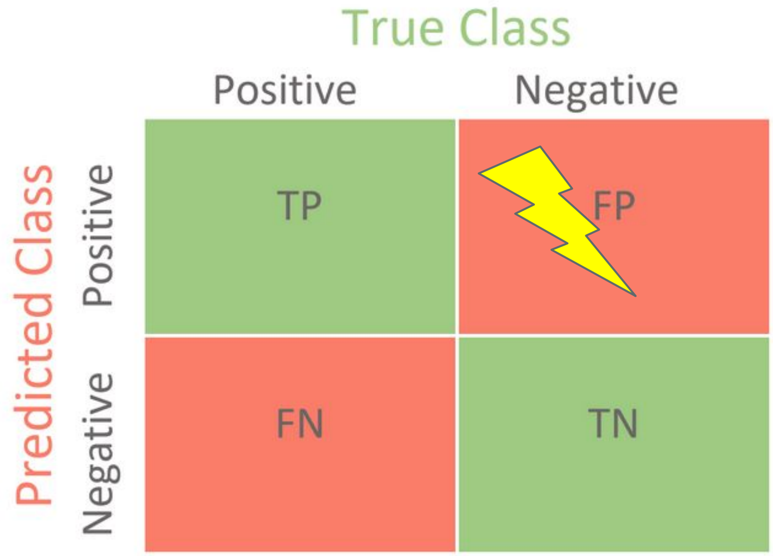
Spam Non-Spam



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Evaluation

Spam Non-Spam

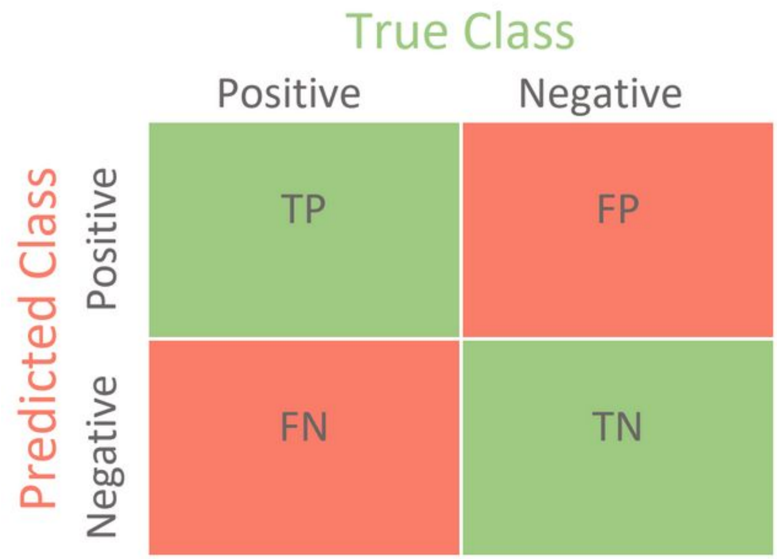


		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Evaluation

Pregnant

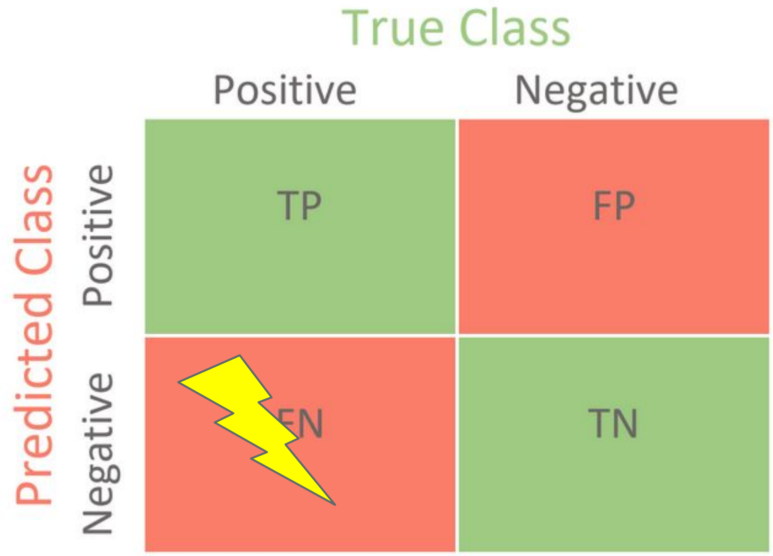
Non Pregnant



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Evaluation

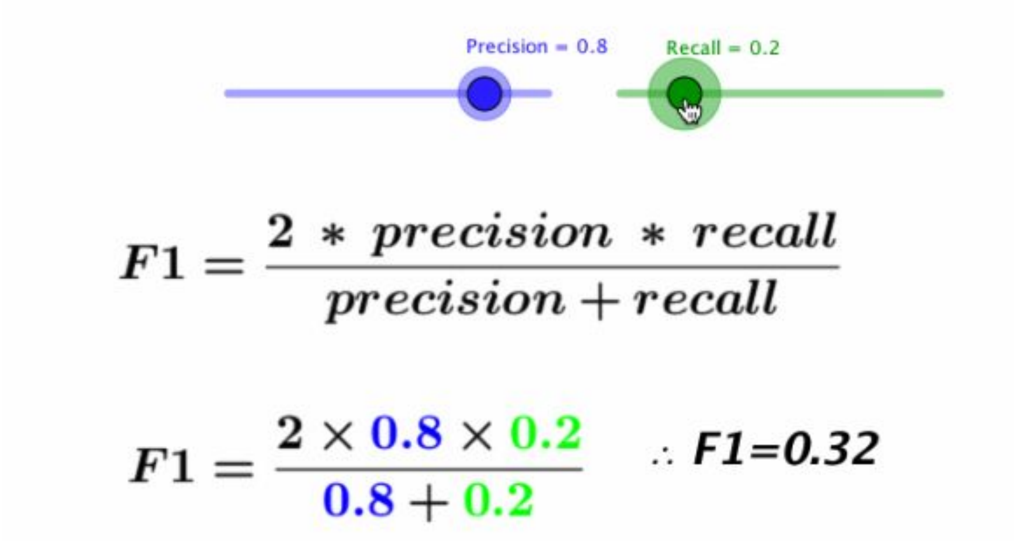
Pregnant Non Pregnant



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

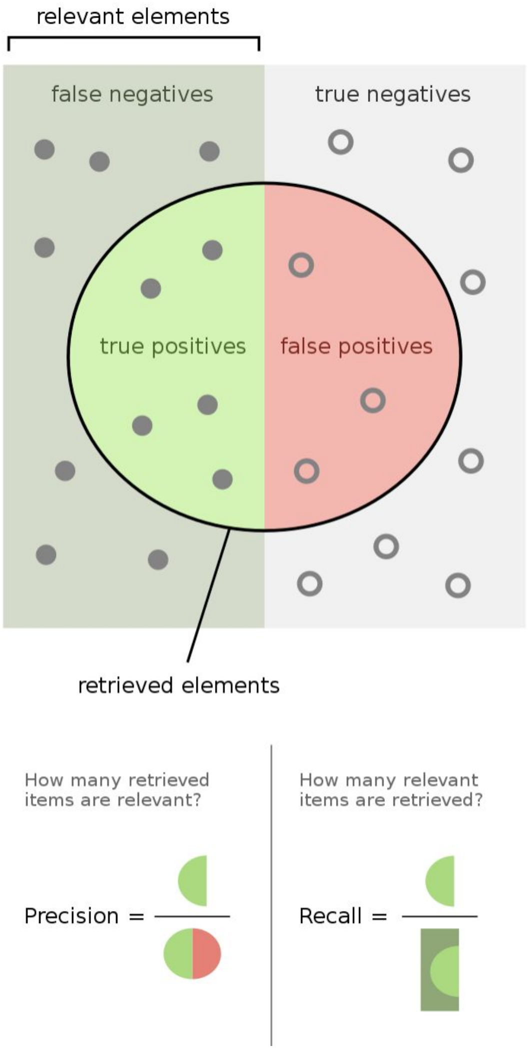
Fairness

Evaluation

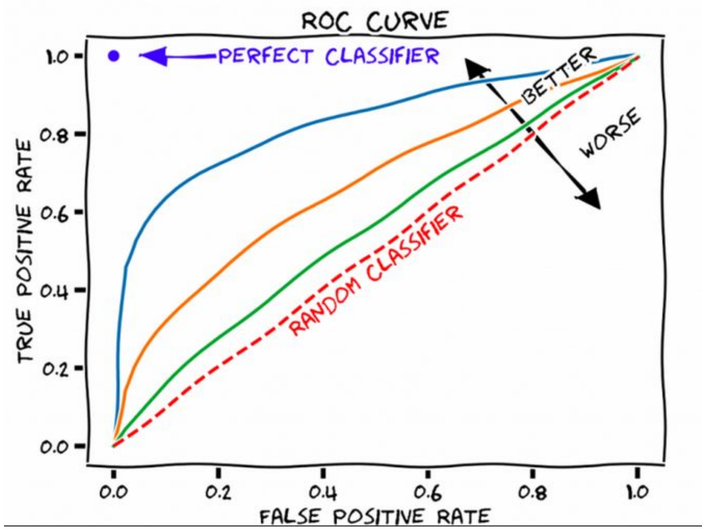


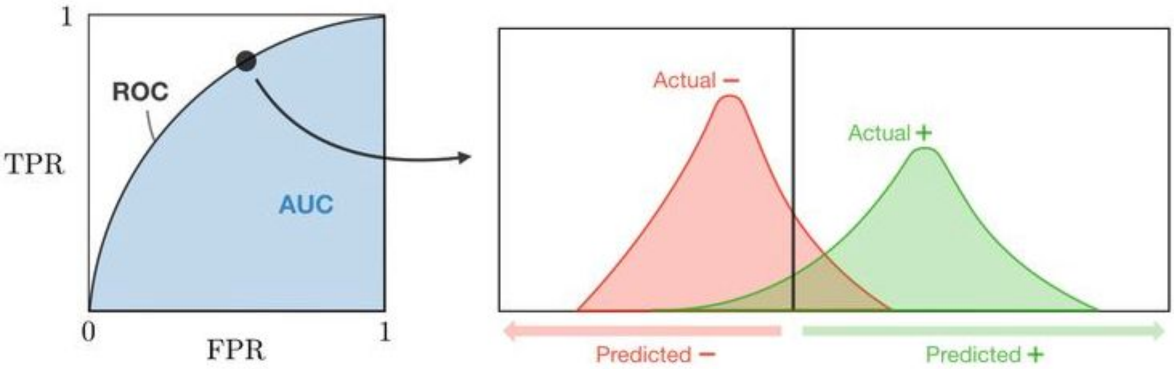
$$F1 = \frac{2 * precision * recall}{precision + recall}$$

$$F1 = \frac{2 \times 0.8 \times 0.2}{0.8 + 0.2} \quad \therefore F1 = 0.32$$

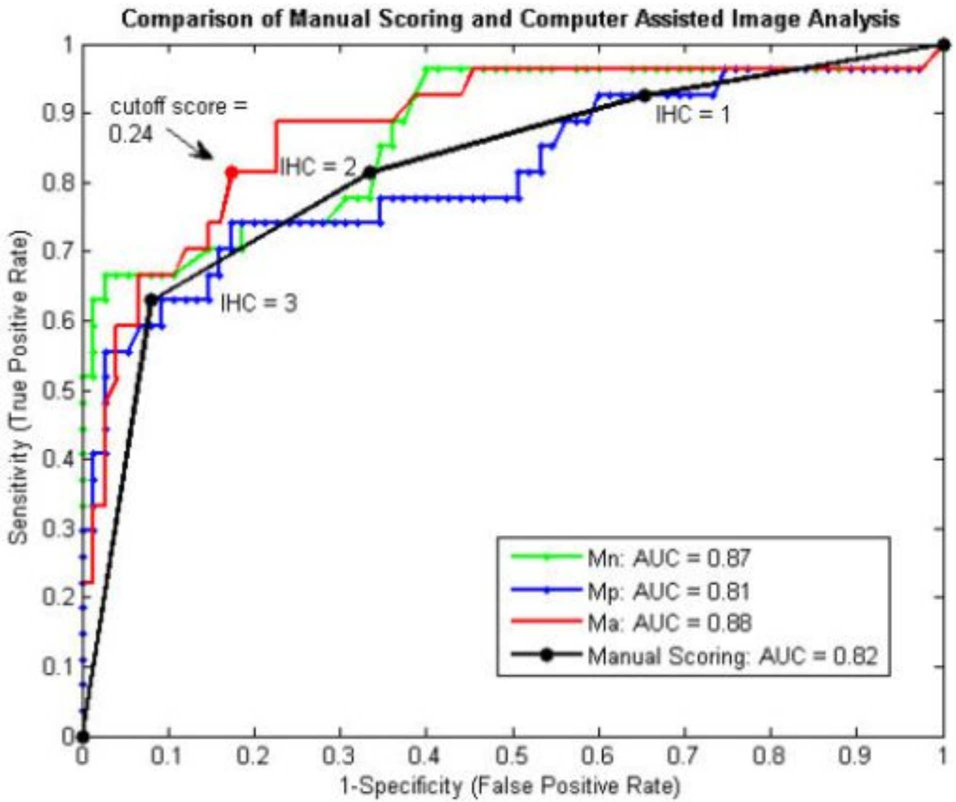


Evaluation

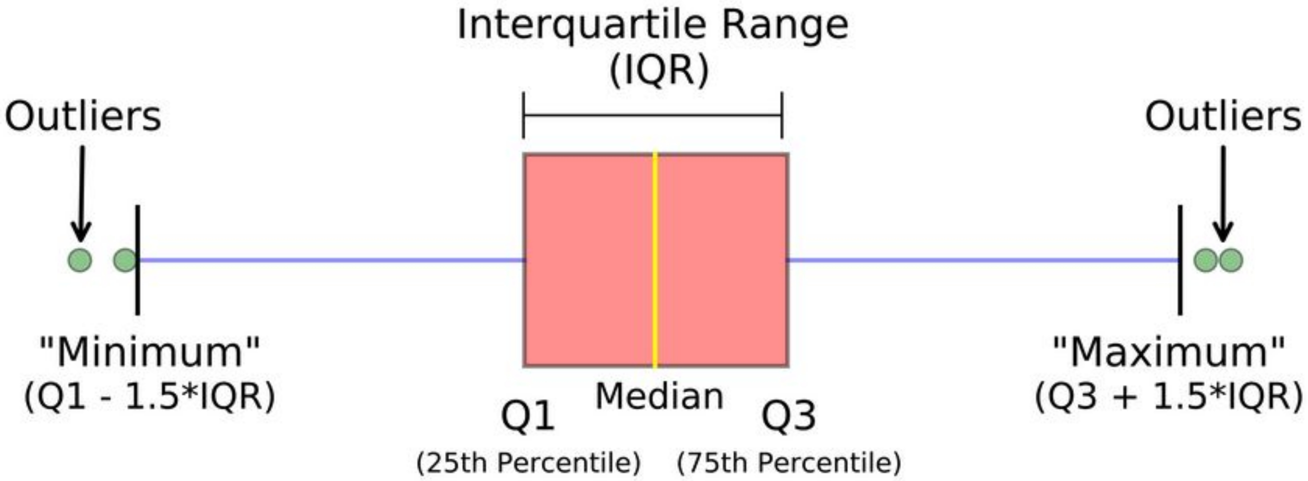

ROC CURVE
PERFECT CLASSIFIER
BETTER
WORSE
RANDOM CLASSIFIER
TRUE POSITIVE RATE
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FALSE POSITIVE RATE


1
ROC
TPR
AUC
0
FPR
1
Actual −
Actual +
Predicted −
Predicted +

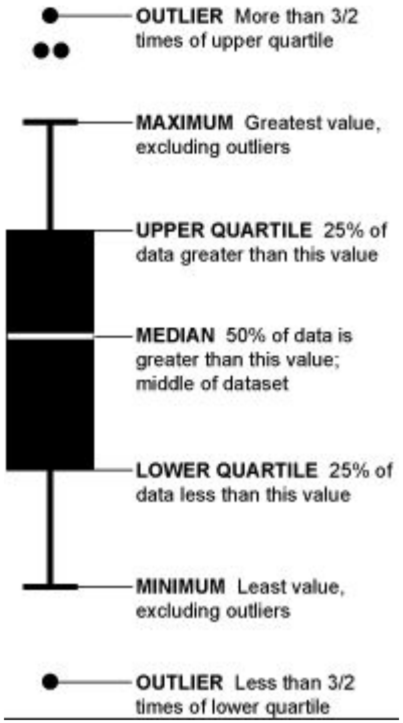
Evaluation


Comparison of Manual Scoring and Computer Assisted Image Analysis
cutoff score = 0.24
IHC = 2
IHC = 3
IHC = 1
Mn: AUC = 0.87
Mp: AUC = 0.81
Ma: AUC = 0.88
Manual Scoring: AUC = 0.82
Sensitivity (True Positive Rate)
1-Specificity (False Positive Rate)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

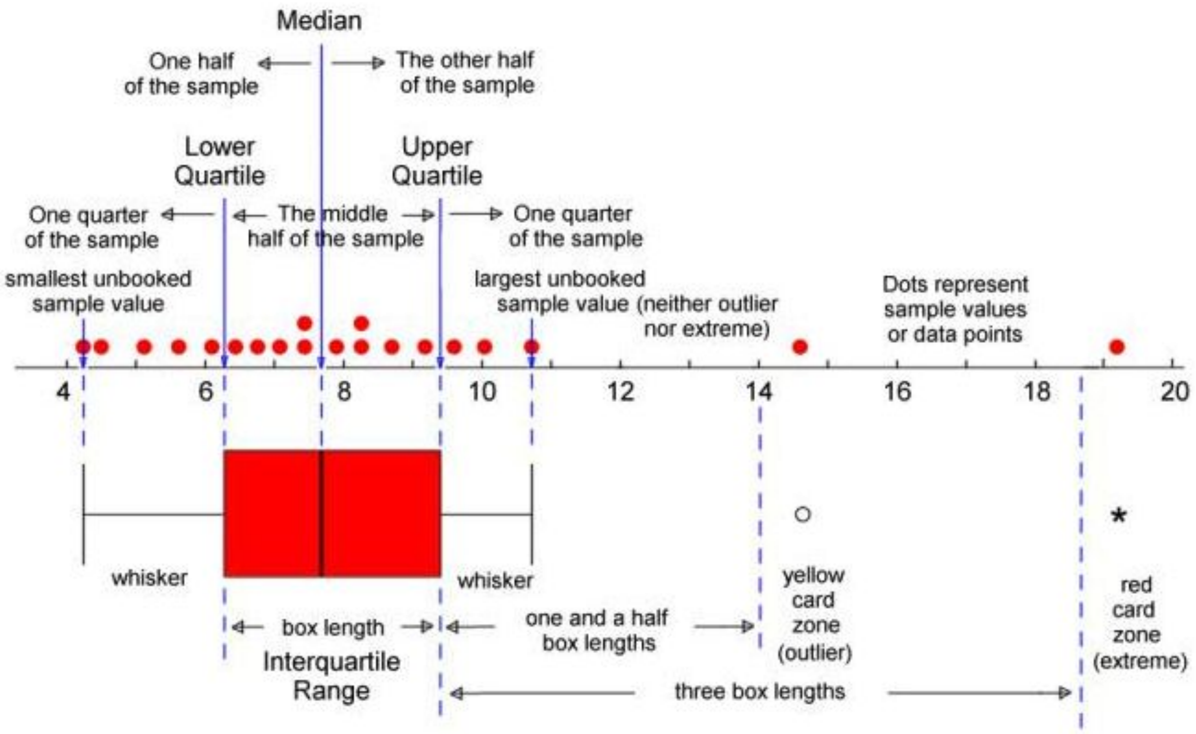
Boxplot


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)

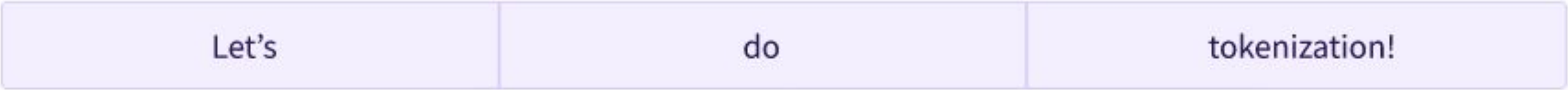
Boxplot


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

Boxplot


Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
The middle half of the sample
One quarter of the sample
smallest unbooked sample value
largest unbooked sample value (neither outlier nor extreme)
Dots represent sample values or data points
4
6
8
10
12
14
16
18
20
○
*
whisker
whisker
yellow card zone (outlier)
red card zone (extreme)
box length
Interquartile Range
one and a half box lengths
three box lengths

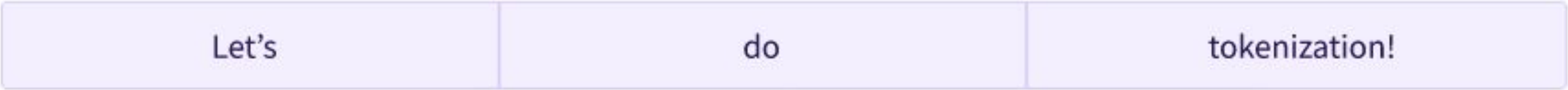
Tokenization



Let’s	do	tokenization!

Tokenization

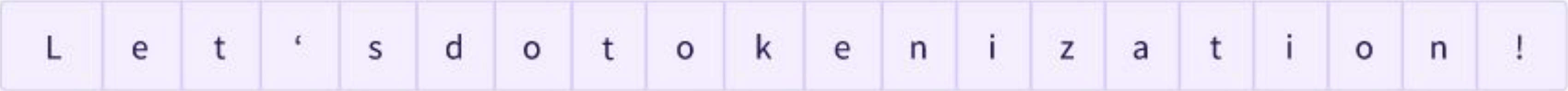
Computationally Expensive
Too many words



Let’s	do	tokenization!

Limit vocabulary
Loose input information

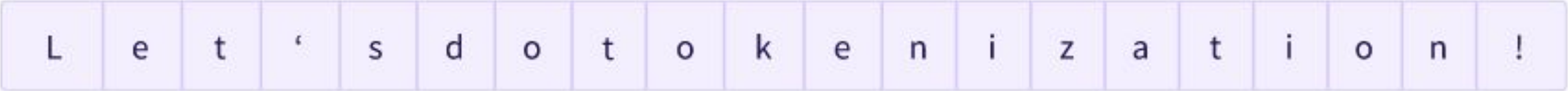
Tokenization



L	e	t	‘	s	d	o	t	o	k	e	n	i	z	a	t	i	o	n	!

Tokenization

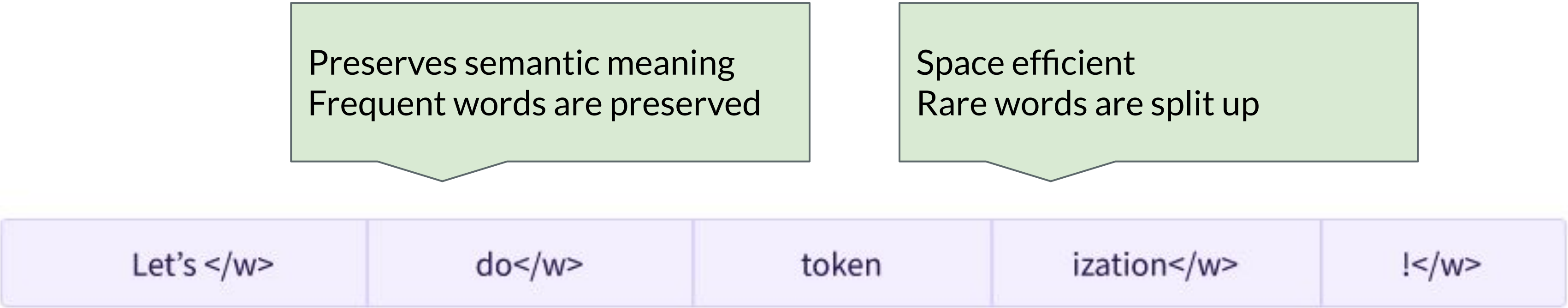
Loose linguistic information
Words need to be relearnt



L	e	t	‘	s	d	o	t	o	k	e	n	i	z	a	t	i	o	n	!

Increased training time
Increased training data needed

Tokenization

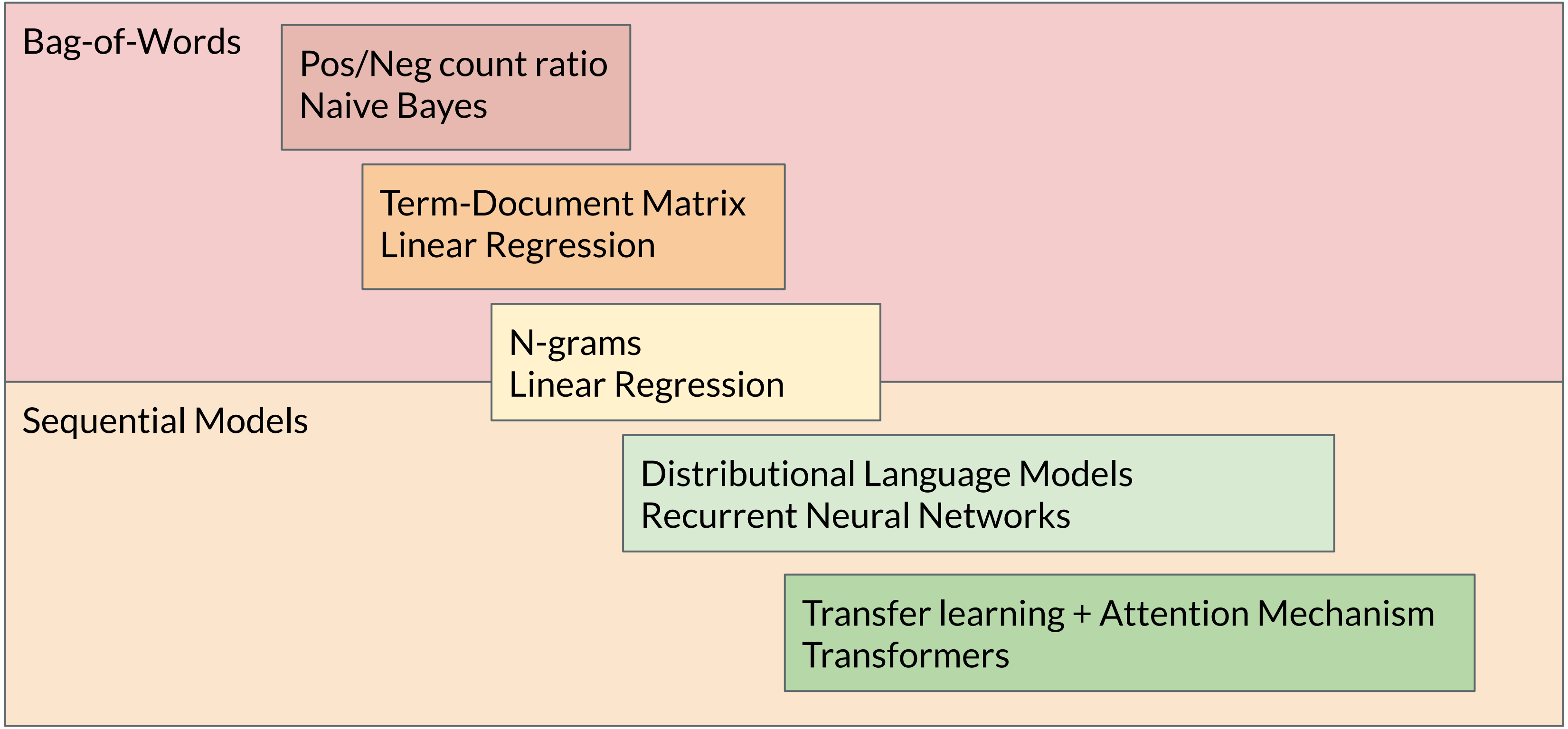


Preserves semantic meaning
Frequent words are preserved

Space efficient
Rare words are split up

Let’s </w>	do</w>	token	ization</w>	!</w>

Solution Tiers



Bag-of-Words

Pos/Neg count ratio
Naive Bayes

Term-Document Matrix
Linear Regression

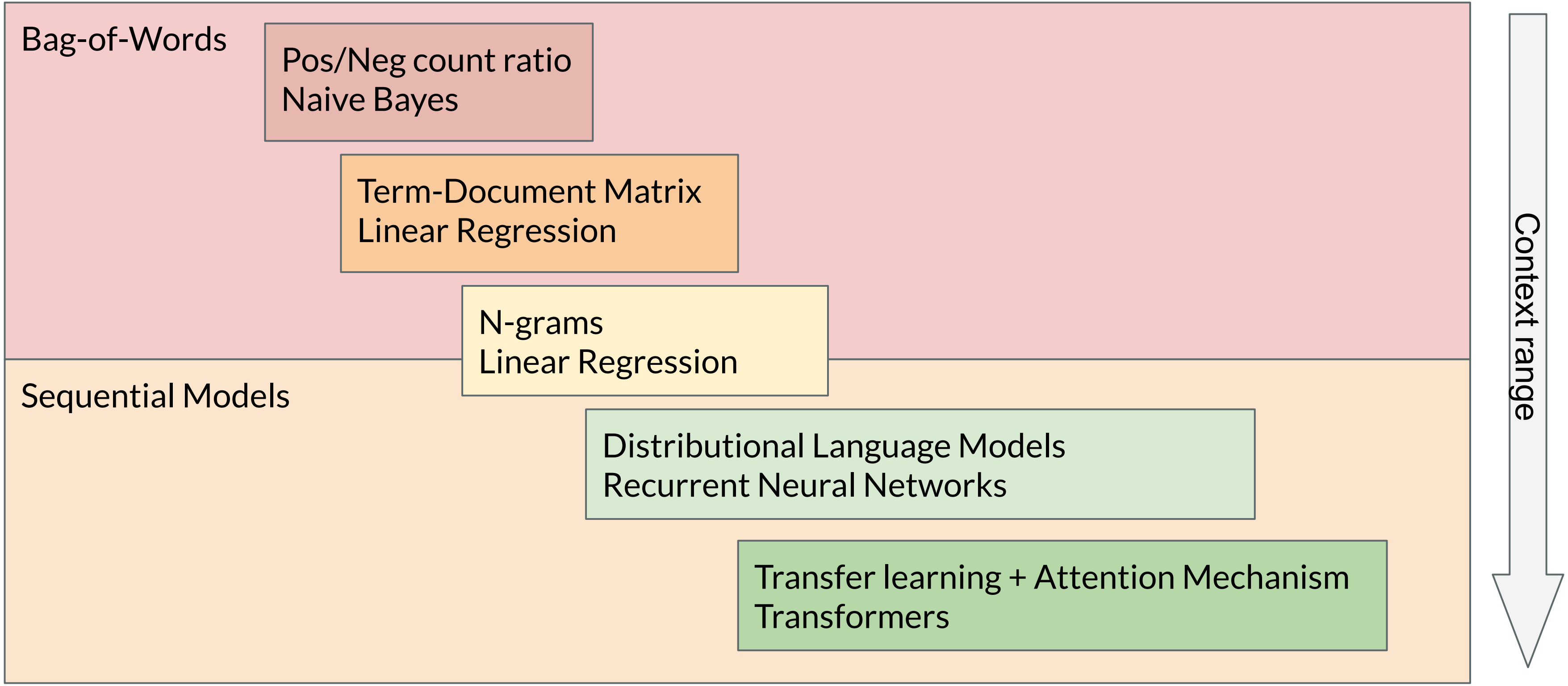
N-grams
Linear Regression

Sequential Models

Distributional Language Models
Recurrent Neural Networks

Transfer learning + Attention Mechanism
Transformers

Solution Tiers



Bag-of-Words

Pos/Neg count ratio
Naive Bayes

Term-Document Matrix
Linear Regression

N-grams
Linear Regression

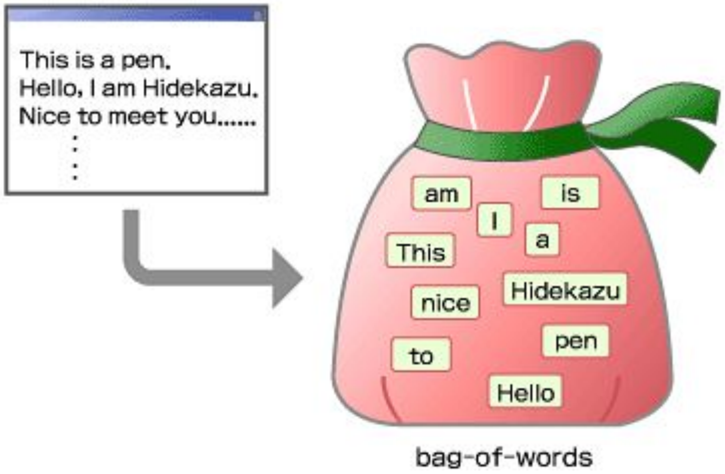
Sequential Models

Distributional Language Models
Recurrent Neural Networks

Transfer learning + Attention Mechanism
Transformers

Context range

Bag-of-Words



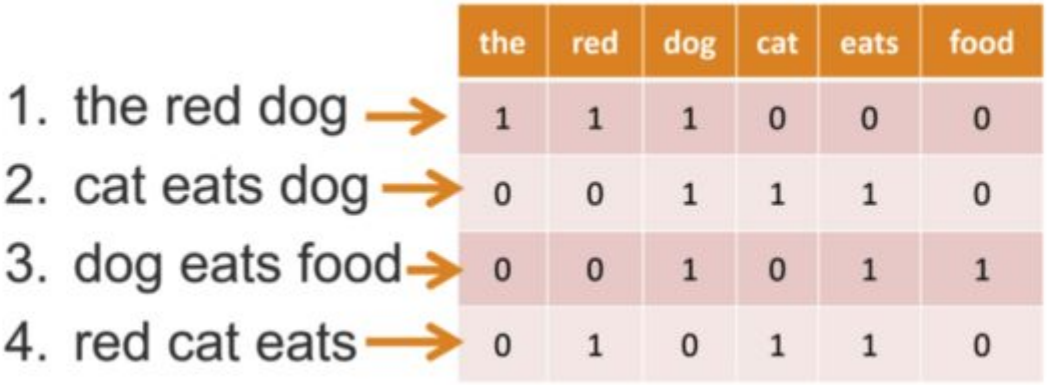
This is a pen.
Hello, I am Hidekazu.
Nice to meet you......
:
:
am
is
I
a
This
nice
Hidekazu
to
pen
Hello
bag-of-words

Bag-of-Words

negative
weak
shortness
unfortunately
symptoms
noise
slow
worse
scared
difficult
hang
die
infections
issue
wound
lost
worry
bleeding
sorry
issues
loss
cancer
anxiety
slowly
bleed
irritation
failure
pain
shock
death
bad
infection
trouble
damage
emergency
tired
alarm
risk
sick
attack
itching
top
great
problem
hard
worth
helped
love
recovery
sensitive
awesome
better
like
fine
worked
luck
loved
perfect
comfortable
good
support
ready
enough
well
fast
recover
peace
hot
works
glad
lucky
freedom
clean
happy
thank
work
patient
pretty
wonderful
bless
clear
best
right
easier
helpful
nice
easy
enjoy
appreciate
positive
safe
available
free
grateful
protect
important
secure
strong
improved
healthy
recommend
positive

Word	Positive Probability Count	Negative Probability Count	Ratio
problem	2/100	10/100	0.2
best	10/100	1/100	10
slowly	5/100	6/100	0.83

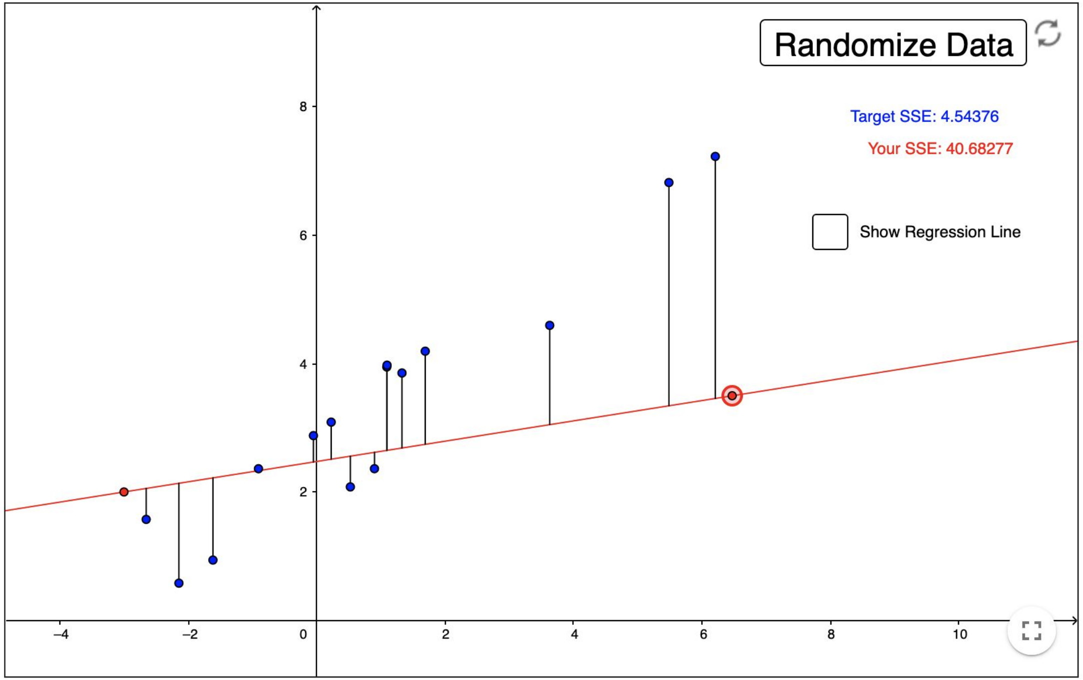
Bag-of-Words



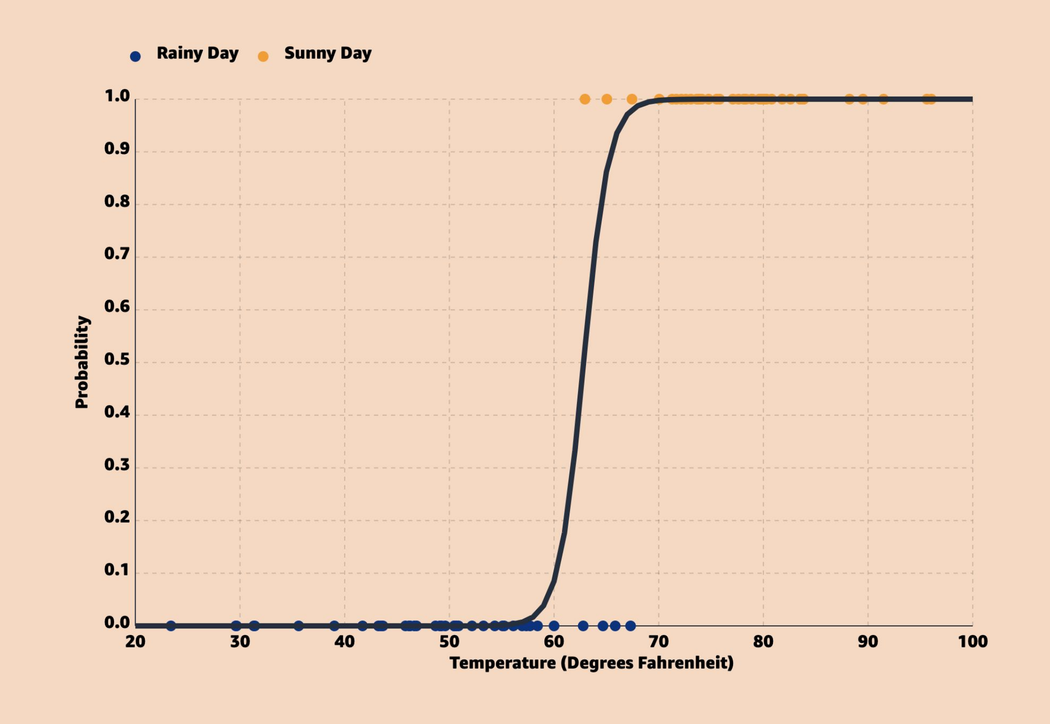
1. the red dog
2. cat eats dog
3. dog eats food
4. red cat eats

	the	red	dog	cat	eats	food
1. the red dog	1	1	1	0	0	0
2. cat eats dog	0	0	1	1	1	0
3. dog eats food	0	0	1	0	1	1
4. red cat eats	0	1	0	1	1	0

Linear Regression



Logistic Regression



Logistic Regression

Sigmoid -> Prediction

- takes [-inf,inf] and converts to (0,1)
- with a rate of change - bigger around 0.5 uncertainty, then at 0 or 1 more certainty
- Why this specific formula? Why use e (=Euler’s number 2.71)?
    - because to use standard training (SGD) we need a continuous & differentiable function
    - and because using this formula with e, the derivative is simple!!
    - d(a^x)/dx = a^x * ln(a)
    - d(e^x)/dx = e^x

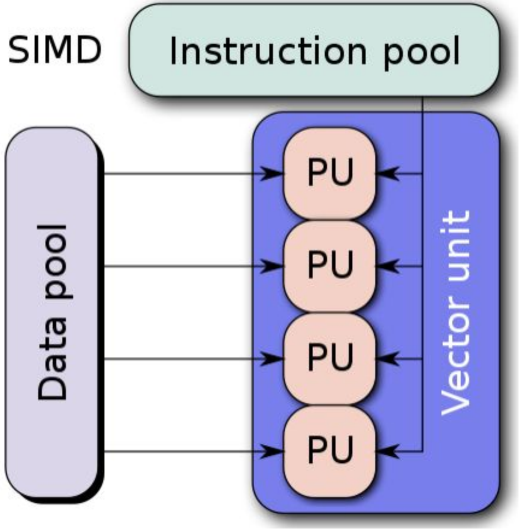
Logistic Regression

Log-Loss/Binary Cross-Entropy -> Evaluation

- cost function - ylog(pred) + (1-y)log(1-pred)
    - prediction between 0-1 => log between -inf-0

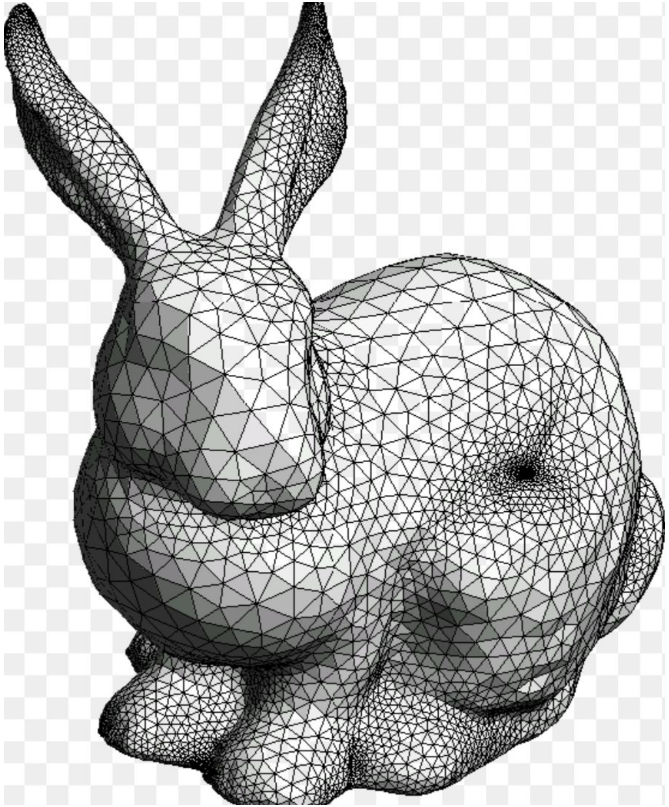
Vectorization == SIMD

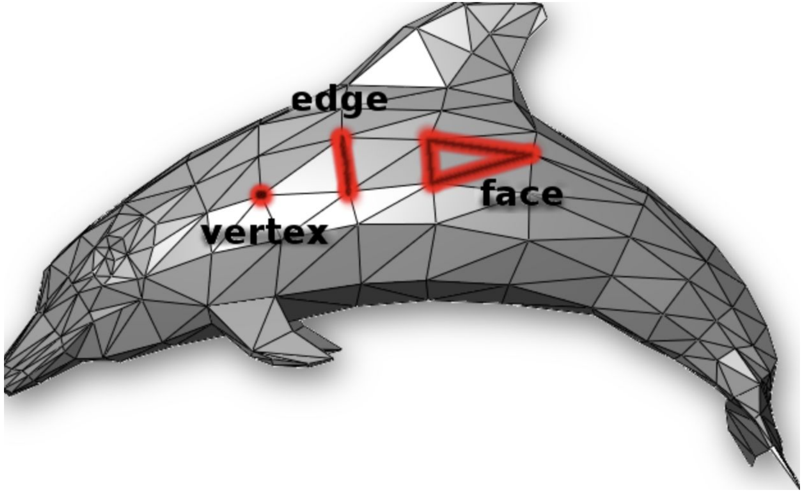
Modern CPUs supports operations for SIMD (Single Instructions on Multiple Data)



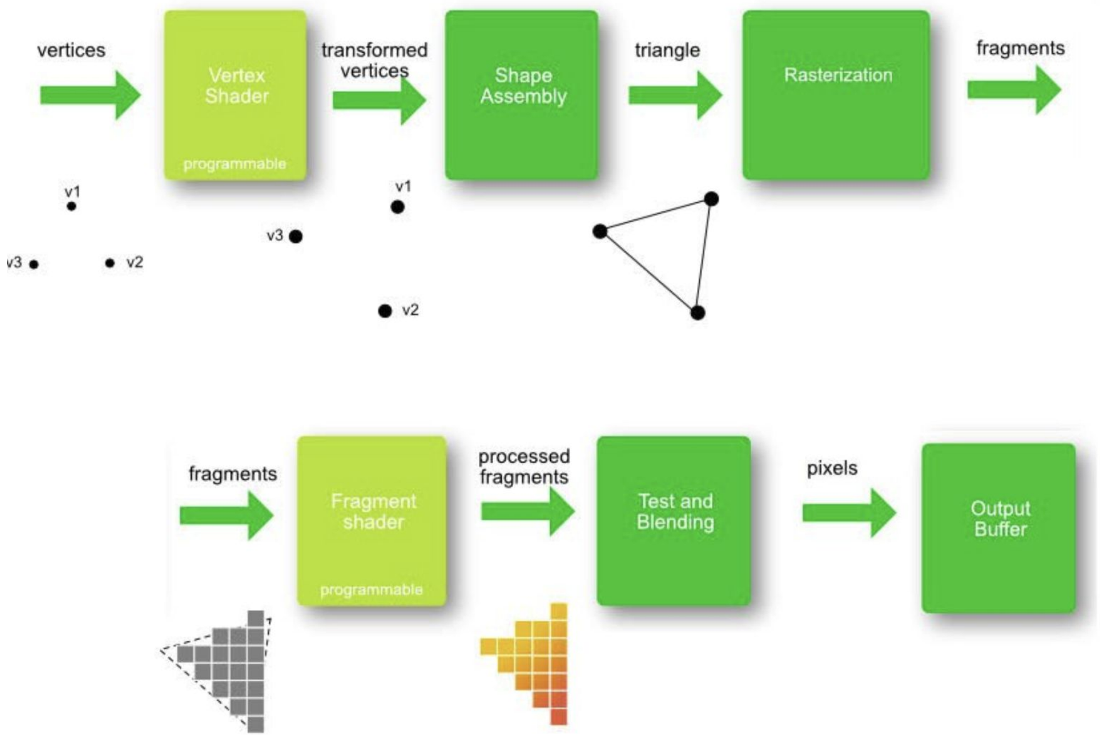
Single Data vs Multiple Data Instructions

GPU



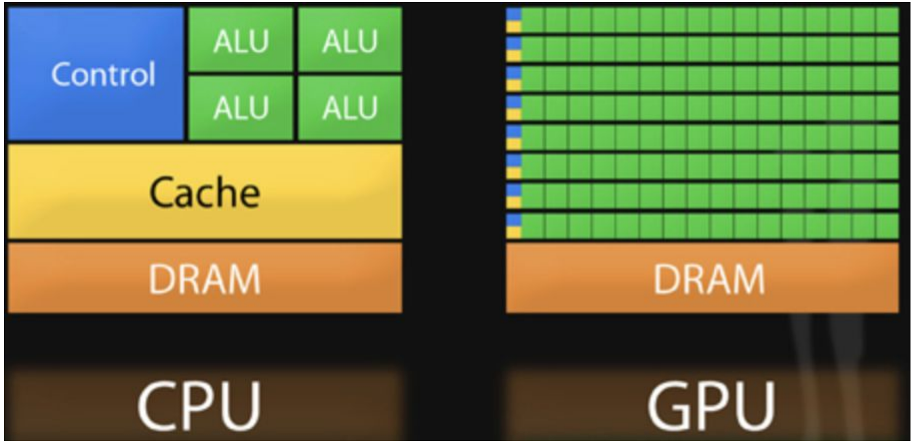


GPU - specialized hardware for coloring triangles



CPU vs GPU

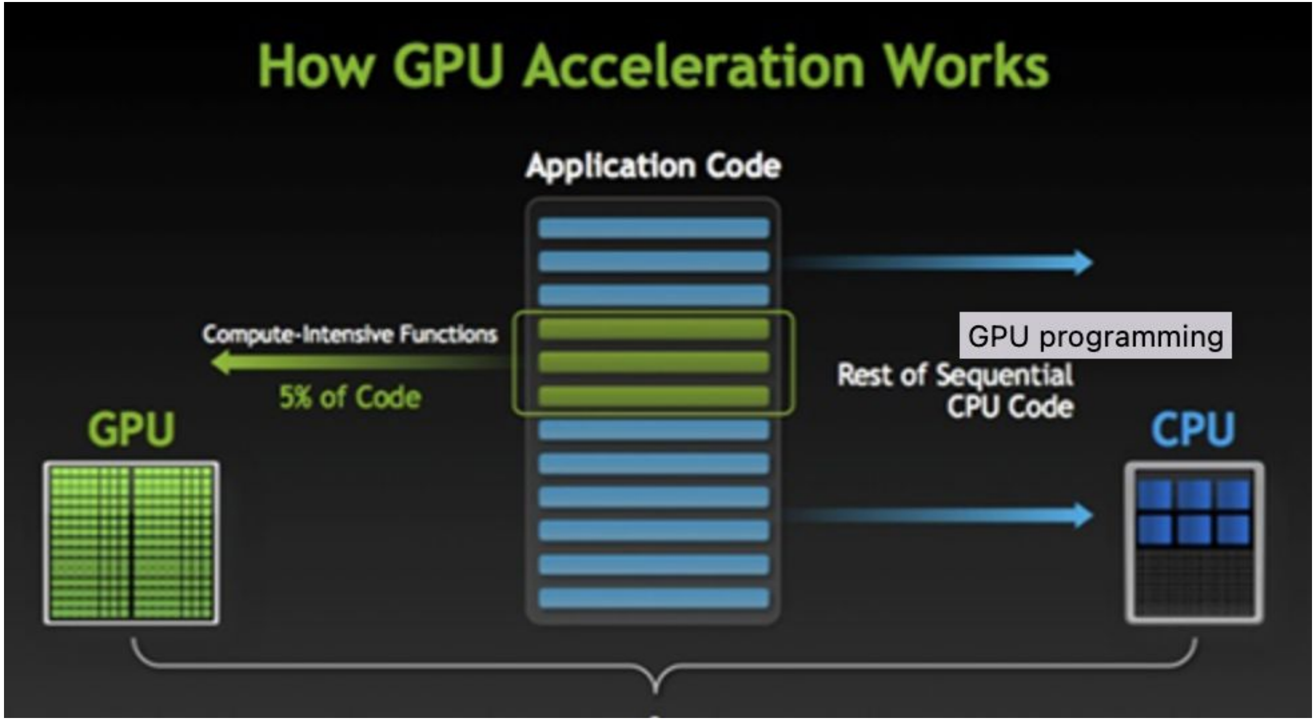
Arithmetic Logical Units (4 vs 1000)



GPU is designed for *data-parallel computations*

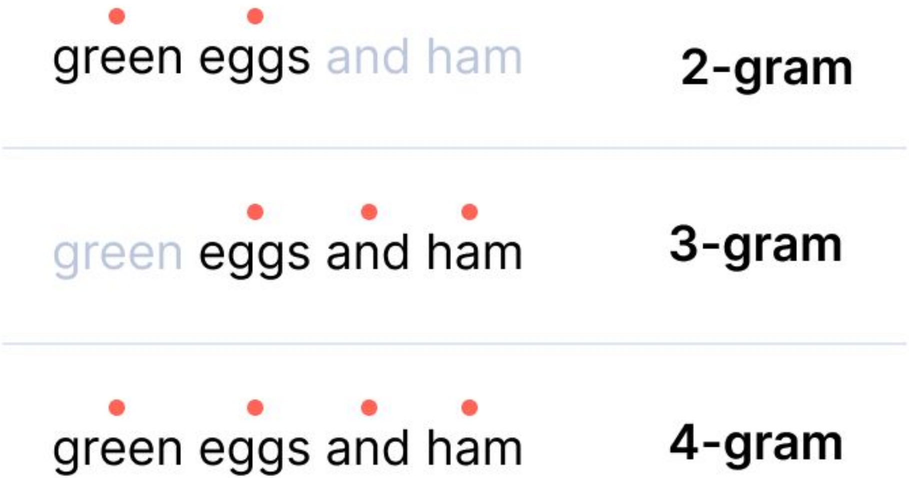
- same program executed on many data elements in parallel
  - no need for sophisticated flow control
- high ratio of arithmetic operations to memory operations
  - no need for lots of cache to speed up memory access

GPU Acceleration



N-Grams

N-grams are a sequence of n tokens from a sample of text.

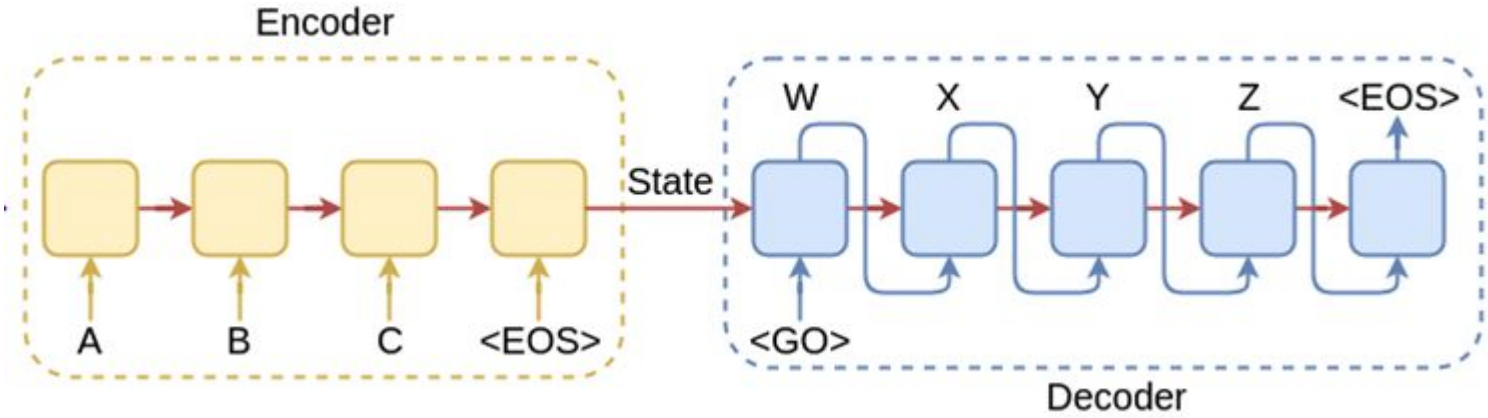


green eggs and ham 2-gram

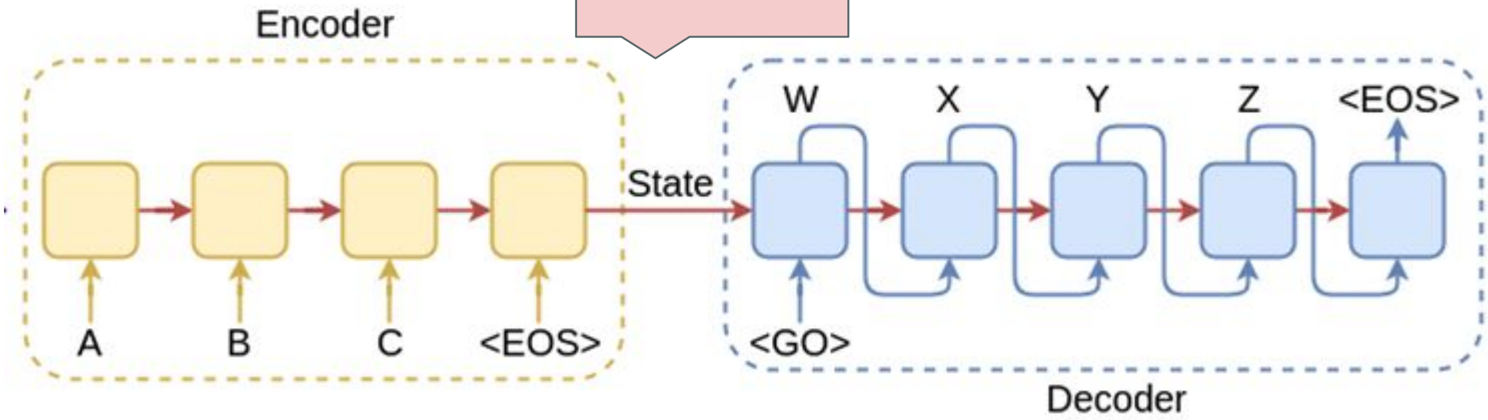
green eggs and ham 3-gram

green eggs and ham 4-gram

Sequential Models

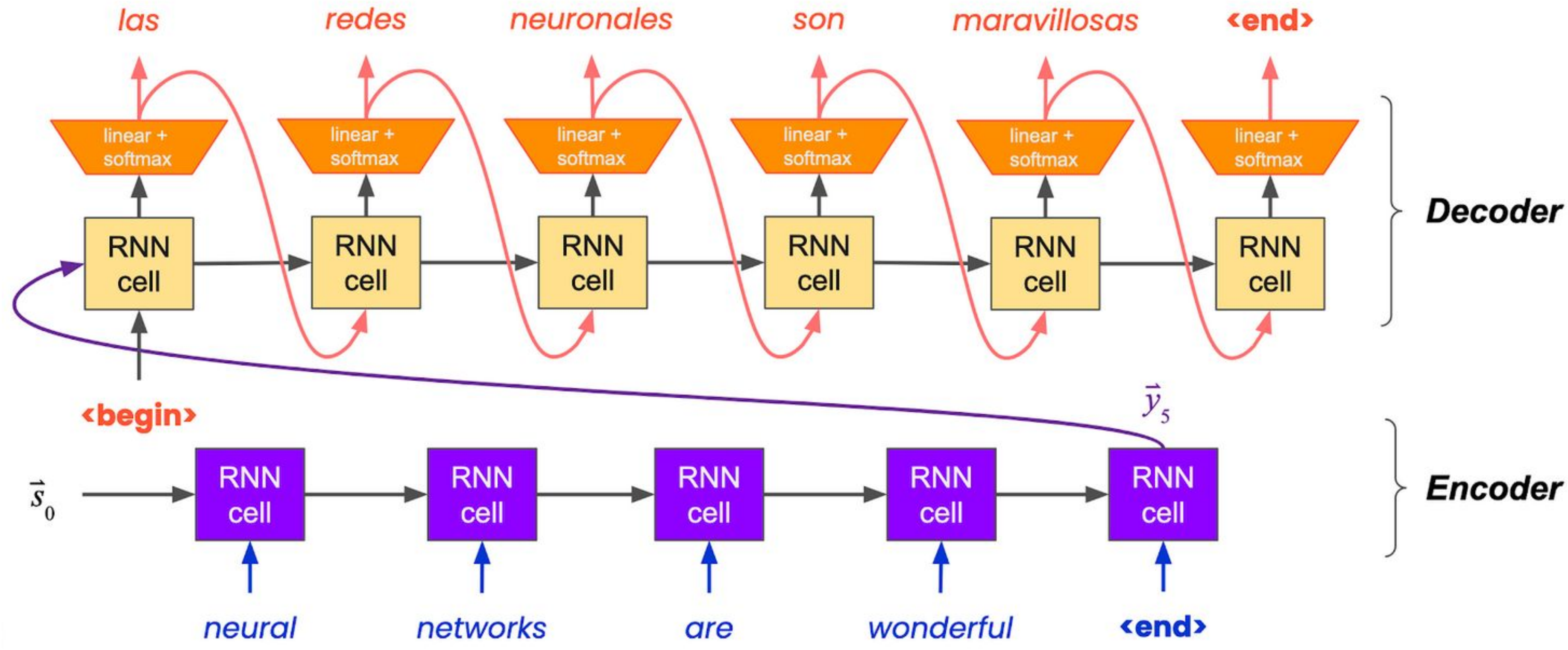


Sequential Models



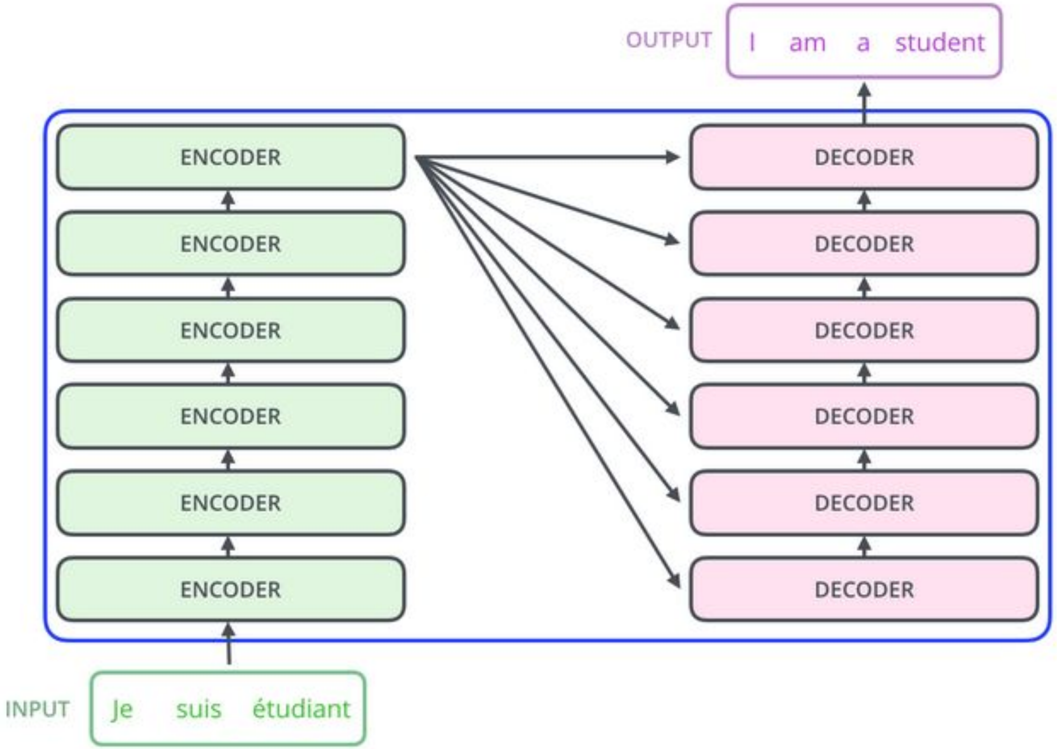
Information Bottleneck

Sequential Models



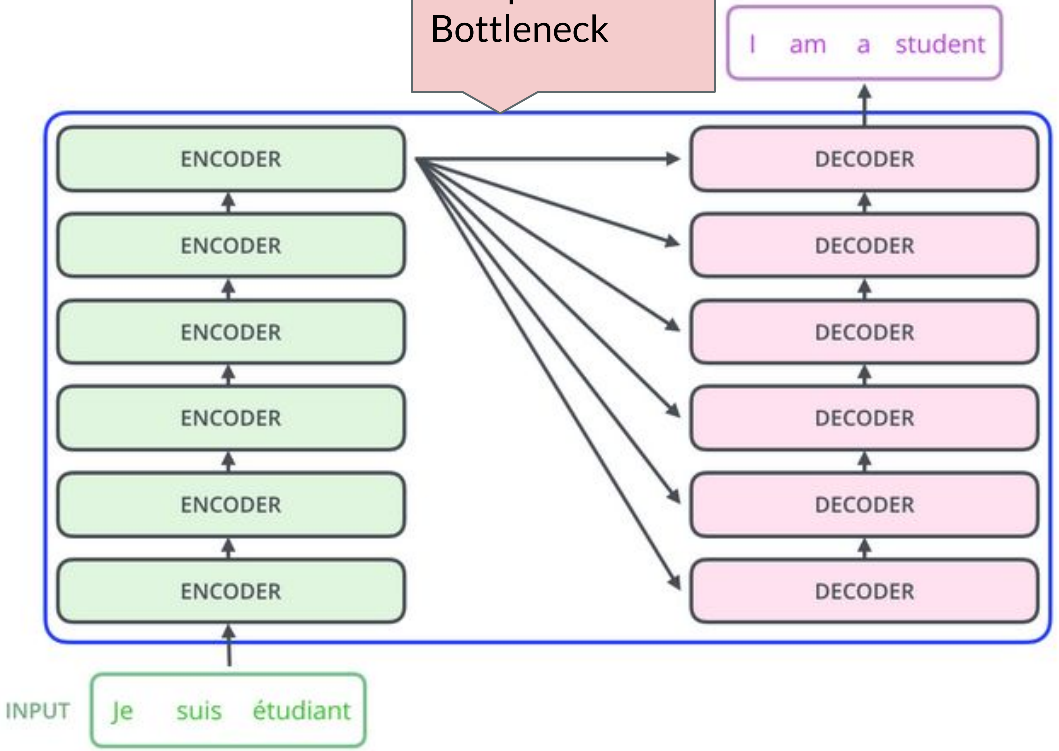
Transformers

Transformers

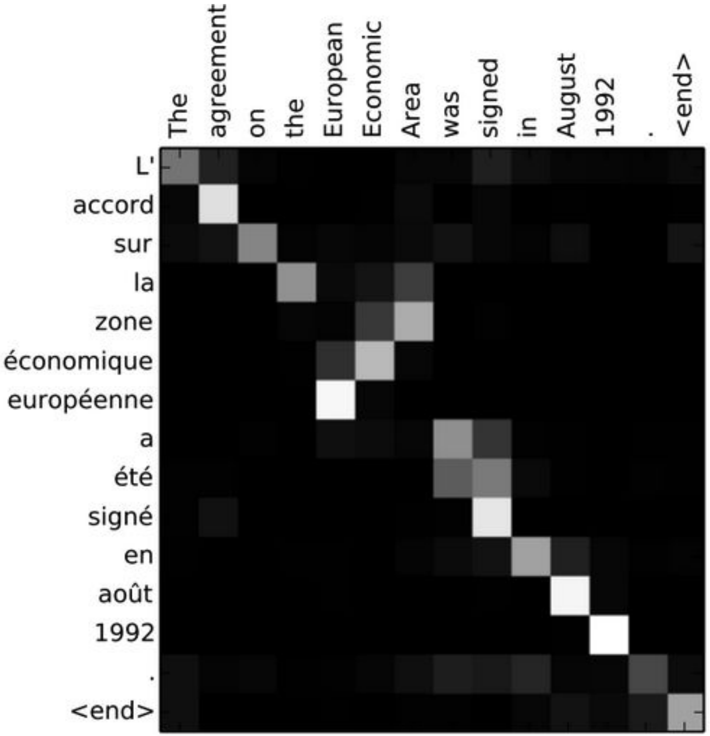


Transformers

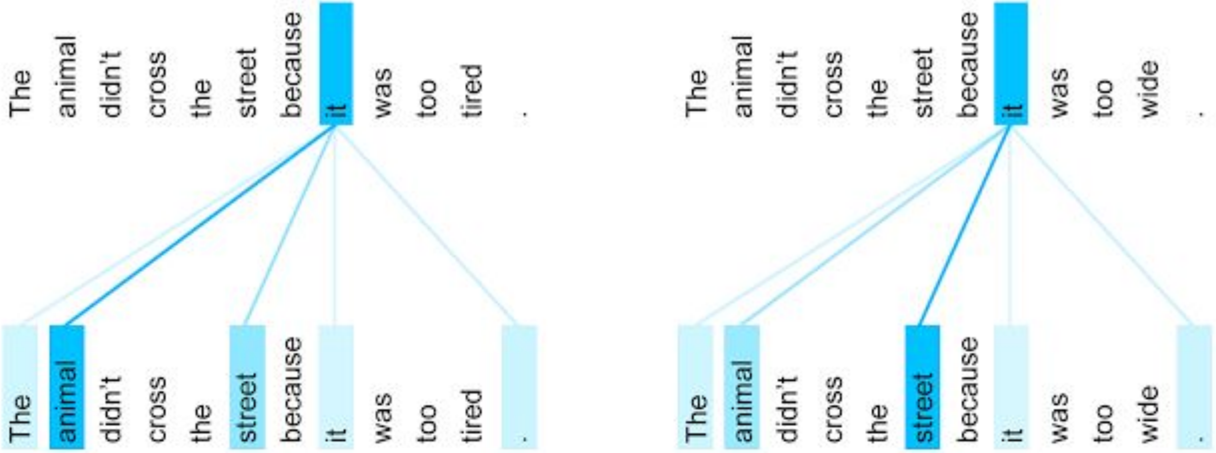
Computational Bottleneck



Attention

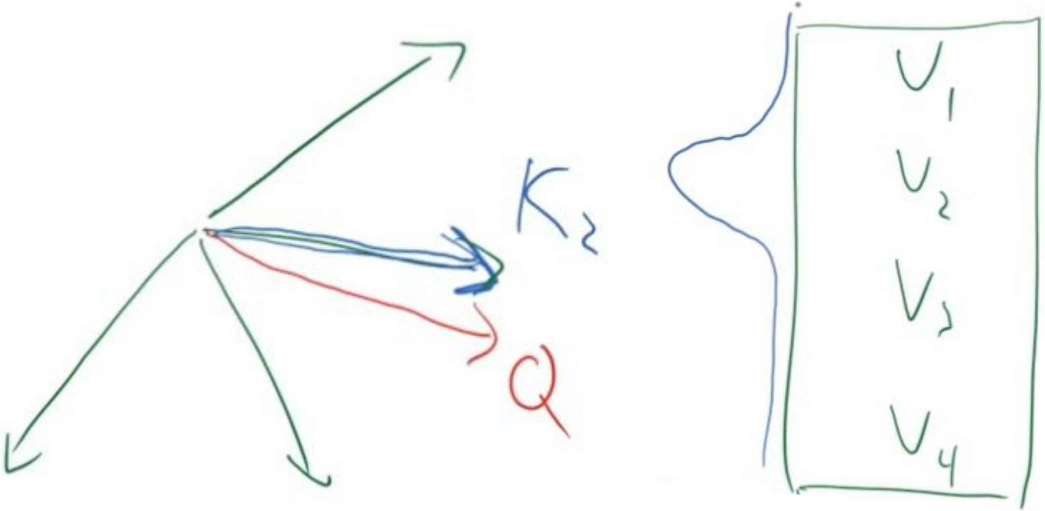

The
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>
L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>

Self-Attention

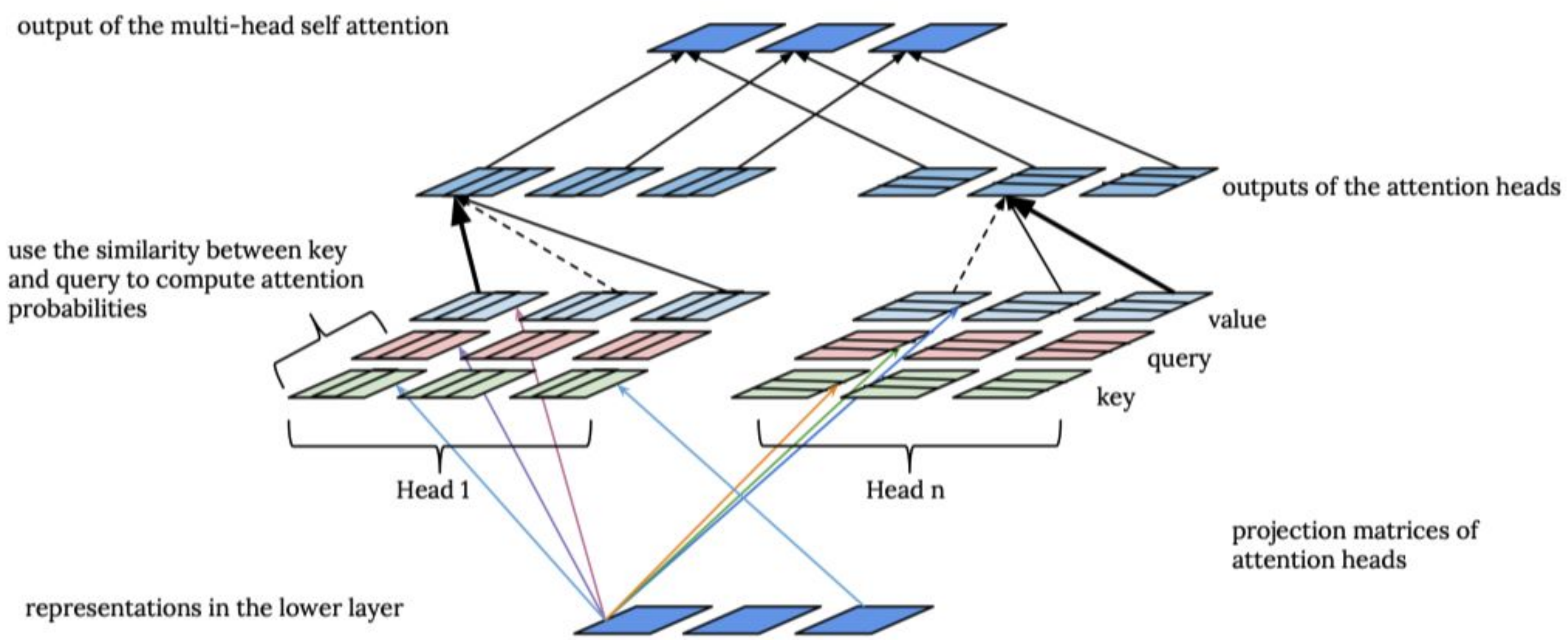

The animal didn't cross the street because it was too tired .
The animal didn't cross the street because it was too tired .
The animal didn't cross the street because it was too wide .
The animal didn't cross the street because it was too wide .

The encoder self-attention distribution for the word "it" from the 5th to the 6th layer of a Transformer trained on English to French translation (one of eight attention heads).

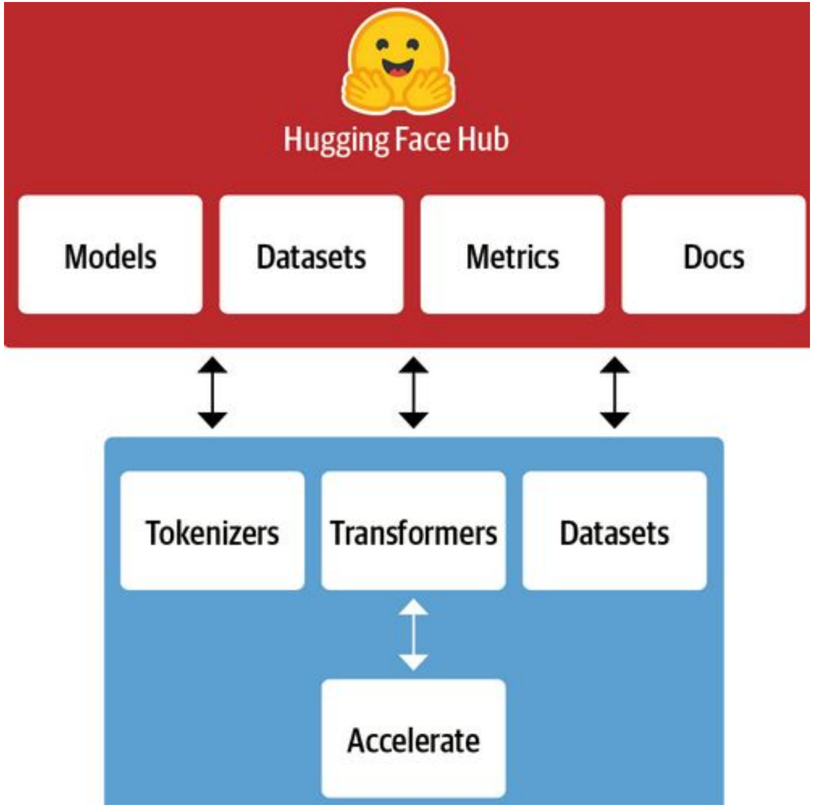
Attention



Multi-Headed Self-Attention


output of the multi-head self attention
outputs of the attention heads
use the similarity between key and query to compute attention probabilities
value
query
key
Head 1
Head n
projection matrices of attention heads
representations in the lower layer

huggingface ecosystem



Python libraries