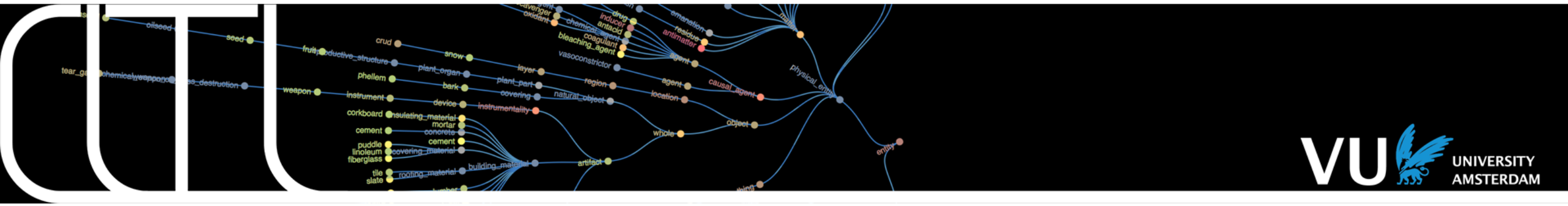
Text Mining CBS 2019

Lecture 2: Machine Learning for NLP

Piek Vossen



Overview

- Part I: Humans against machines
- Part II: Supervised machine learning
- Part III: Deep learning
- Part IV: Evaluation

Part I: Humans against Machines

What is Machine Learning?

Learn from experience

Learn from experience

=DATA

Follow instructions

=RULES

NLP Approaches

- **Rule-based**

  tell the machine exactly what to do under specific circumstances

- **Machine learning**

  the machine learns to identify patterns itself

    - Supervised machine learning

      the machine uses labelled examples to learn

    - Unsupervised machine learning

      the machine identifies patterns without using labelled examples

- **Hybrid** = a combination of (un)supervised machine learning and rule-based approaches

Simplest form of rule system

- **rules:** for word ***w*** in tweet ***t*** do if ***w*** == “friendly” then ***POSITIVE***
- **regular expressions to improve recall, e.g.:**
    - friend* —> positive, un* —> negative
    - https://www.nltk.org/book/ch03.html, section 3.4
- **lexicon** that lists words with properties:
    - Each word in a text is looked up in the lexicon and the information is added to the text (tweet or sentence)
    - ***Ambiguity*** needs to be resolved by analysing the context possibly using ***rules***
    - **Other issues: negation “**not friendly”, **intensifiers “**very friendly”
- **Examples of lexical and rule-based analysis:**
    - Sentiment: VADER (for tweets); https://github.com/cjhutto/vaderSentiment
    - Emotions: NRC (emotions expressed by words): https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
    - Psychological features: LIWC http://liwc.wpengine.com

VADER lexicon

Valence Aware Dictionary and sEntiment Reasoner

TOKEN, MEAN-SENTIMENT-RATING, STANDARD DEVIATION, and RAW-HUMAN-SENTIMENT-RATINGS [10 raters -4, 0, +4]

```
arrogance   -2.4    0.66332 [-3, -2, -2, -3, -1, -2, -3, -3, -2, -3]
arrogances  -1.9 0.53852[-1, -2, -3, -2, -2, -2, -2, -1, -2, -2]
arrogant-2.2 0.6[-2, -2, -2, -3, -2, -3, -3, -2, -1, -2]
arrogantly  -1.8 1.4[-3, -2, 2, -3, -1, -3, -2, -2, -2, -2]
ashamed -2.1 1.3[-3, -3, -3, -2, -2, 1, -2, -4, -1, -2]
ashamedly   -1.7    0.64031 [-2, -2, -2, -1, -2, -3, -1, -2, -1, -1]
ass -2.5 1.43178[-4, -1, -2, -1, -3, 0, -2, -4, -4, -4]
```

😀 grinning face

😁 beaming face with smiling eyes

😂 face with tears of joy

🤣 rolling on the floor laughing

😃 grinning face with big eyes

😄 grinning face with smiling eyes

😅 grinning face with sweat

😆 grinning squinting face

😉 winking face

NRC emotion lexicon annotated by the crowd

https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

- beating anger 1
- beating anticipation 0
- beating disgust 0
- beating fear 1
- beating joy 0
- beating negative 1
- beating positive 0
- beating sadness 1
- beating surprise 0
- beating trust 0

- betrayal anger 1
- betrayal anticipation 0
- betrayal disgust 1
- betrayal fear 0
- betrayal joy 0
- betrayal negative 1
- betrayal positive 0
- betrayal sadness 1
- betrayal surprise 0
- betrayal trust 0

LIWC (Linguistic Inquiry and Word Count)

- Analyses text using lexicons for ***90***(!) dimensions such as:
    - positive & negative emotions
    - social words (e.g. family related vocabulary)
    - cognitieve processes (insight, causation, certainty)
    - authority
- Analyses text properties such as punctuation, sentence length, closed class words, etc.

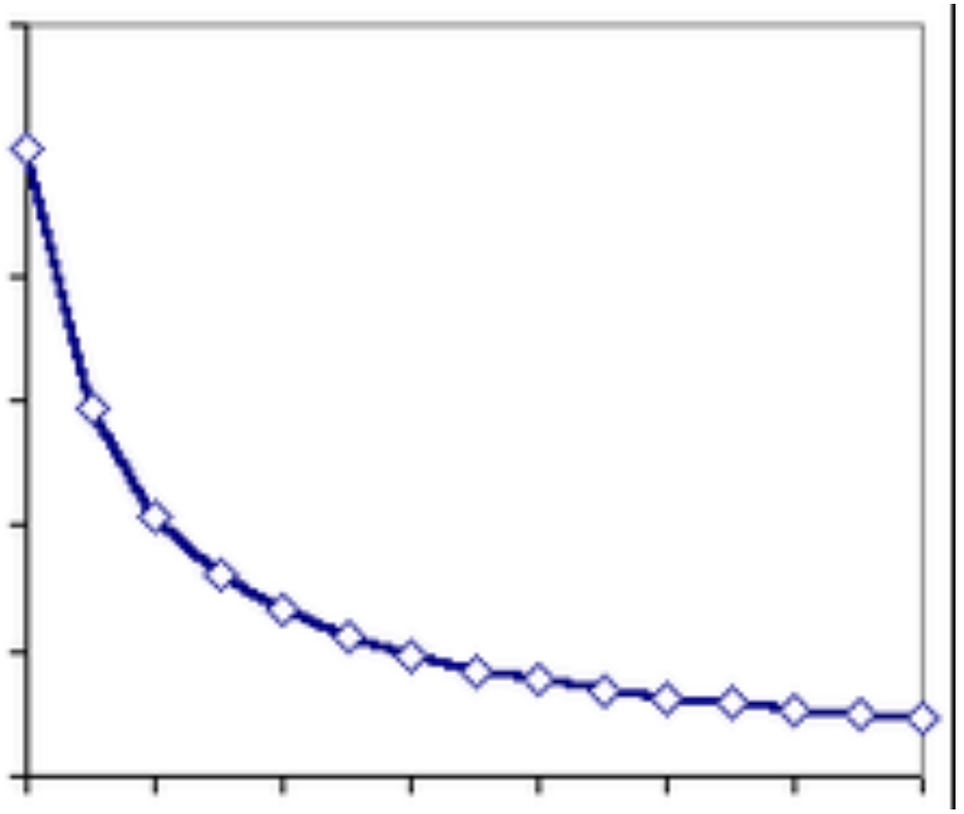
http://liwc.wpengine.com/results/

At least 26 people were killed in Sunday's church shooting in Sutherland Springs, Texas, Gov. Greg Abbott said at a press conference. About 20 others were wounded, said Freeman Martin, a regional director with the Texas Department of Public Safety, with victims ranging in age from 5 to 72 years old. Among the dead is the 14-year-old daughter of the First Baptist Church's pastor, Frank Pomeroy, according to his wife, Sherri Pomeroy, the girl's mother. The couple were traveling out of state when the shooting occurred. Authorities have not said what may have motivated the suspected shooter, who was later found dead in his vehicle. The shooting has devastated the small Texas town east of San Antonio, described as a place where "everybody knows everybody."

TRADITIONAL LIWC DIMENSION	YOUR DATA	AVERAGE FOR
I-WORDS (I, ME, MY)	0.0	
SOCIAL WORDS	11.0	
POSITIVE EMOTIONS	0.8	
NEGATIVE EMOTIONS	2.4	
COGNITIVE PROCESSES	7.1	
SUMMARY VARIABLES		
ANALYTIC	97.8	
CLOUT	80.7	
AUTHENTICITY	56.9	
EMOTIONAL TONE	7.5	

Formulating precise patterns

- Easy to get some results quickly but impossible to cover all ways of expressing information
- Law of diminishing returns:
  - $Effort(X)^N$
  - N = 1 → 50% recall
  - N = 2 → 60%
  - N = 3 → 65%
  - N = 4 → 68%



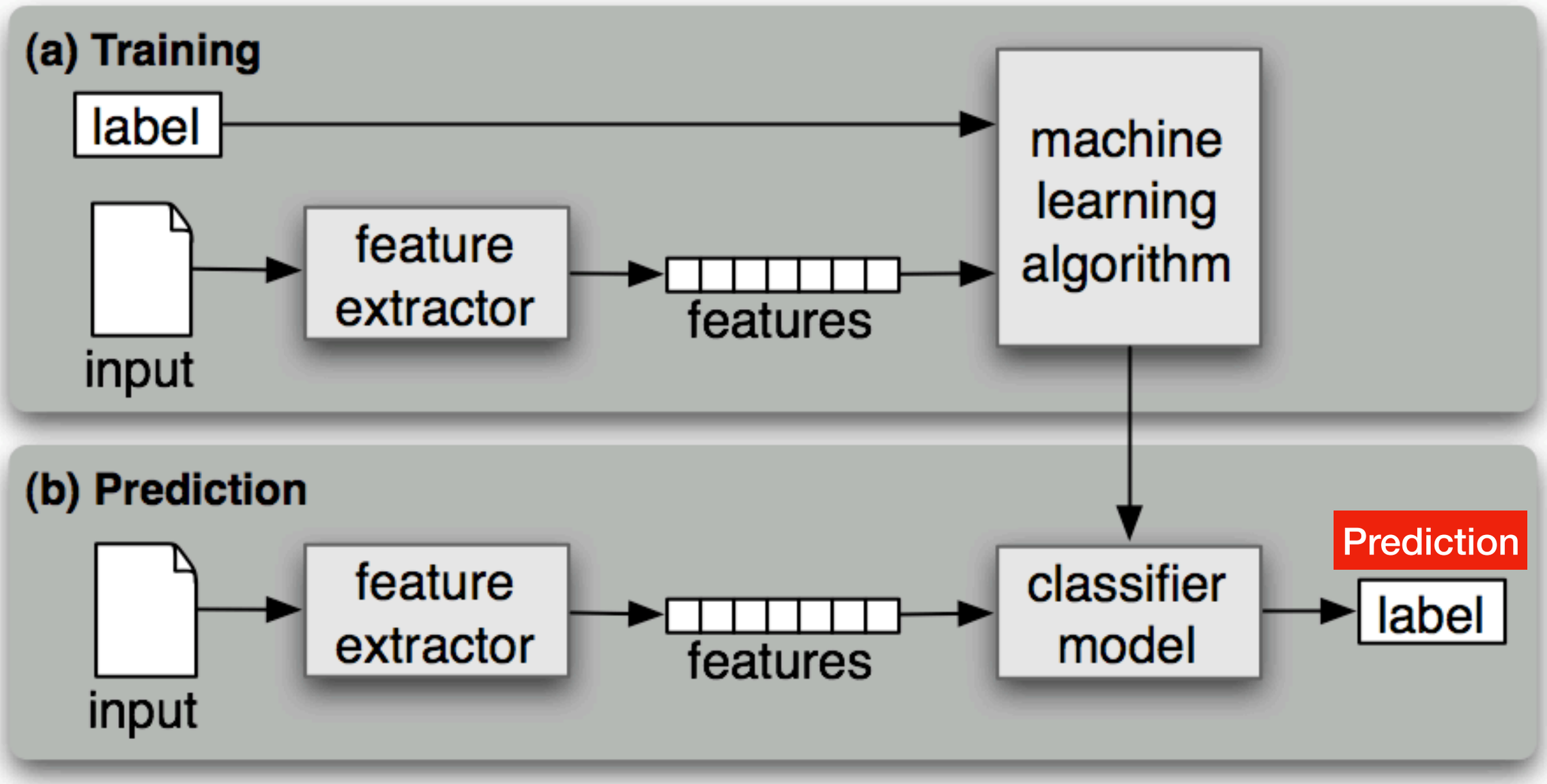
Why machine learning?

- Hand-crafted Linguistic Models (e.g. grammars) failed the test of empirical adequacy on real data: variation and dynamics of language is far bigger than realised

- Rules became too complex and we need too many,
  - making it impossible to maintain systems
  - making it psychologically unrealistic: what child learns these rules?

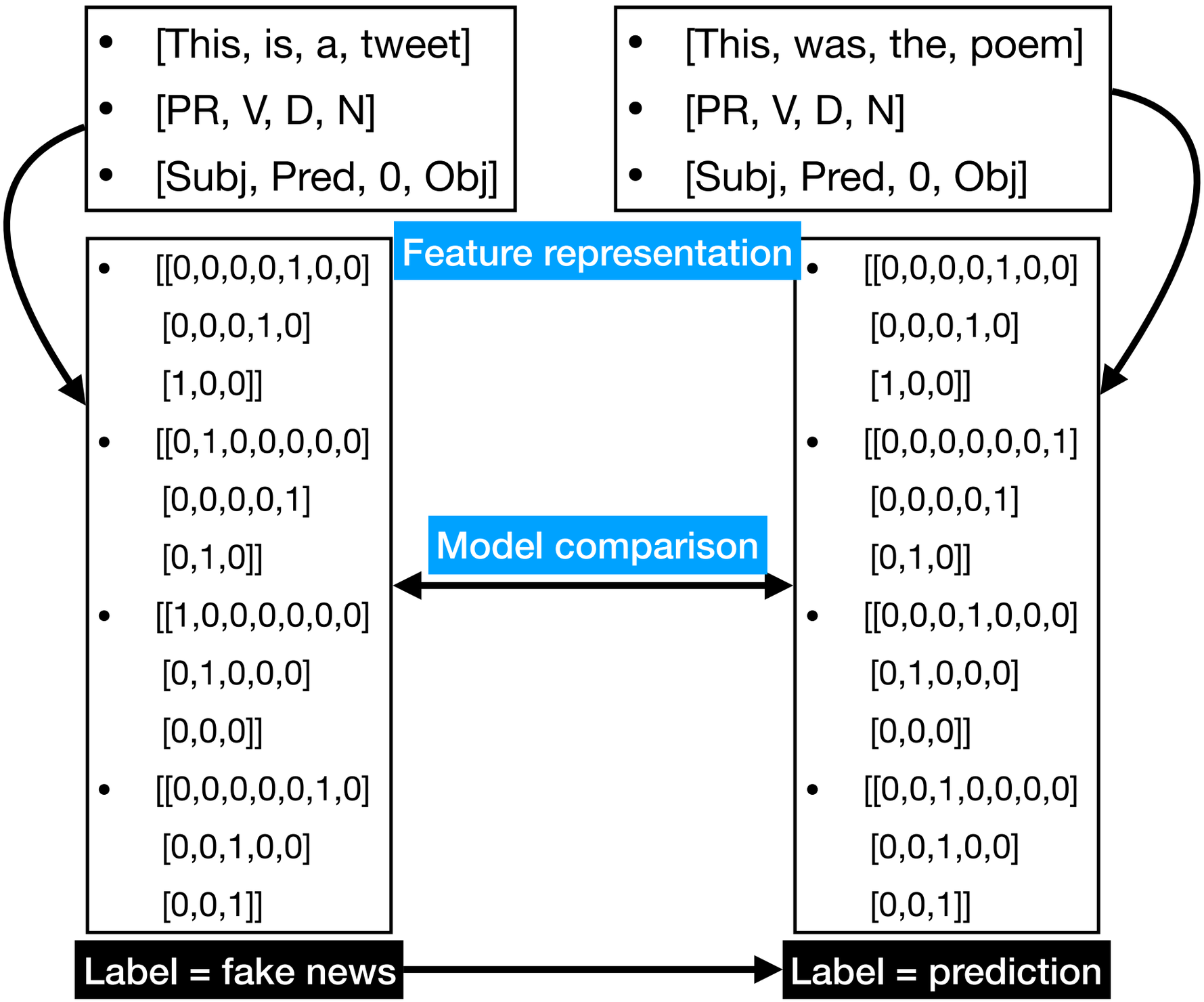
- Machine learning appears to work better than any set of the rules we can invent

Supervised Machine learning



Machine Learning

- **Feature space**
- vocabulary = [a, is, poem, the, this, tweet, was]
- PoS = [A, D, N, PR, V]
- Dependencies = [Subj, Pred, Obj]
- Vector dimensions
  - [[0,0,0,0,0,0,0]
    [0,0,0,0,0]
    [0,0,0]]



Creating the training data
Data annotation procedure

---

1. Collect texts: e.g. tweets, news, blogs, books
2. Define an **annotation scheme** or **code book**:
    - tag set (e.g. PoS labels, emotions, entity types)
    - the unit of the annotation: word, phrase, sentence, paragraph, document
    - criteria to apply a tag to a piece of text
3. Train **human annotators** to use the annotation scheme or create a **crowd task**
4. Provide an **annotation tool** that loads texts and allow the annotator to assign tags
5. **Store** the annotations with the text, e.g. in XML or TAB separated
6. Determine the **Inter-Annotator-Agreement** (IAA) by analysing texts annotated by at least one annotator
7. Fix disagreements (**adjucation**): if IAA is too low (<60 Kappa) the task is considered too difficult
8. IAA is considered the **upper ceiling of NLP;** can machines do better than humans?

“inline” Annotation

(4) a. Perhaps without realizing it, Mr. Taffner simultaneously has put his finger on the problem and an ideal solution: “Capital City” should have been a comedy, a worthy sequel to the screwball British “Carry On” movies of the 1960s.

b. Perhaps without realizing it, Mr. <ENAMEX TYPE=“PERSON”>Taffner</ENAMEX> simultaneously has put his finger on the problem and an ideal solution: <ENAMEX TYPE=“WORK_OF_ART” >“Capital City”</ENAMEX> should have been a comedy, a worthy sequel to the screwball <ENAMEX TYPE=“NORP”>British</ENAMEX> “<ENAMEX TYPE=“WORK_OF_ART” >Carry On</ENAMEX>” movies of <ENAMEX TYPE=“DATE”>the 1960s</ENAMEX>.

```
( (S (NP-SBJ (NNP Bartok))
     (VP (VBZ describes)
                 (NP (NP (DT the) (NN form))
             (PP (IN of)
                       (NP (DT the) (JJ first) (NN movement))))
             (PP-CLR (IN as)
                       (NP (NP (ADJP (ADVP (" ") (RBR more)
                             (CC or) (RBR less)) (JJ regular))
                                   (NN sonata) (NN form)))))
```

“inline” Annotatie

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	NER
# newdoc url = http://www.poweredbyosteons.org/2012/01/brief-history-of-bioarchaeological.html									
# newdoc s3 = s3://aws-publicdatasets/common-crawl/crawl-data/CC-MAIN-2016-07/segments...									
...									
# sent_id = http://www.poweredbyosteons.org/2012/01/brief-history-of-bioarchaeological.html#60									
# text = The American Museum of Natural History was established in New York in 1869.									
0	The	the	DT	DT	_	2	det	2:det	O
1	American	American	NNP	NNP	_	2	nn	2:nn	B-Organization
2	Museum	Museum	NNP	NNP	_	7	nsubjpass	7:nsubjpass	I-Organization
3	of	of	IN	IN	_	2	prep	_	I-Organization
4	Natural	Natural	NNP	NNP	_	5	nn	5:nn	I-Organization
5	History	History	NNP	NNP	_	3	pobj	2:prep_of	I-Organization
6	was	be	VBD	VBD	_	7	auxpass	7:auxpass	O
7	established	establish	VBN	VBN	_	7	ROOT	7:ROOT	O
8	in	in	IN	IN	_	7	prep	_	O
9	New	New	NNP	NNP	_	10	nn	10:nn	B-Location
10	York	York	NNP	NNP	_	8	pobj	7:prep_in	I-Location
11	in	in	IN	IN	_	7	prep	_	O
12	1869	1869	CD	CD	_	11	pobj	7:prep_in	O
13	.	.	.	.	_	7	punct	7:punct	O
...									

“stand off” Annotatie

- Raw text is not affected
- Offset pointers to the raw text
- Annotation layers with identifiers
- Layers can point to each other or to offsets
- New layers can be added easily
- Alternatives do not complicate the representation

```
<?xml version="1.0" encoding="UTF-8"?>
<NAF xml:lang="en" version="v3">
  <nafHeader> ••• </nafHeader>
  <raw><![CDATA[Qatar Holding sells 10% stake in Porsche to founding families.]]></raw>
  <text>
    <wf id="w1" sent="1" para="1" offset="0" length="5">Qatar</wf>
    <wf id="w2" sent="1" para="1" offset="6" length="7">Holding</wf>
    <wf id="w3" sent="1" para="1" offset="14" length="5">sells</wf>
    <wf id="w4" sent="1" para="1" offset="20" length="2">10</wf>
    <wf id="w5" sent="1" para="1" offset="22" length="1">%</wf>
    <wf id="w6" sent="1" para="1" offset="24" length="5">stake</wf>
    <wf id="w7" sent="1" para="1" offset="30" length="2">in</wf>
    <wf id="w8" sent="1" para="1" offset="33" length="7">Porsche</wf>
    <wf id="w9" sent="1" para="1" offset="41" length="2">to</wf>
    <wf id="w10" sent="1" para="1" offset="44" length="8">founding</wf>
    <wf id="w11" sent="1" para="1" offset="53" length="8">families</wf>
    <wf id="w12" sent="1" para="1" offset="61" length="1">.</wf>
  </text>
  <terms>
    <!--Qatar-->
    <term id="t1" type="close" lemma="Qatar" pos="R" morphofeat="NNP">
      <span>
        <target id="w1" />
      </span>
      <externalReferences>
        <externalRef resource="WordNet-3.0" reference="ili-30-08986905-n" confidence="1.0" />
        <externalRef resource="WordNet-3.0" reference="ili-30-08986691-n" />
      </externalReferences>
    </term>
    <!--Holding-->
    <term id="t2" type="close" lemma="Holding" pos="R" morphofeat="NNP">
      <span>
        <target id="w2" />
      </span>
      <externalReferences>
        <externalRef resource="WordNet-3.0" reference="ili-30-00810598-n" confidence="1.0" />
        <externalRef resource="WordNet-3.0" reference="ili-30-13244109-n" />
      </externalReferences>
    </term>
```

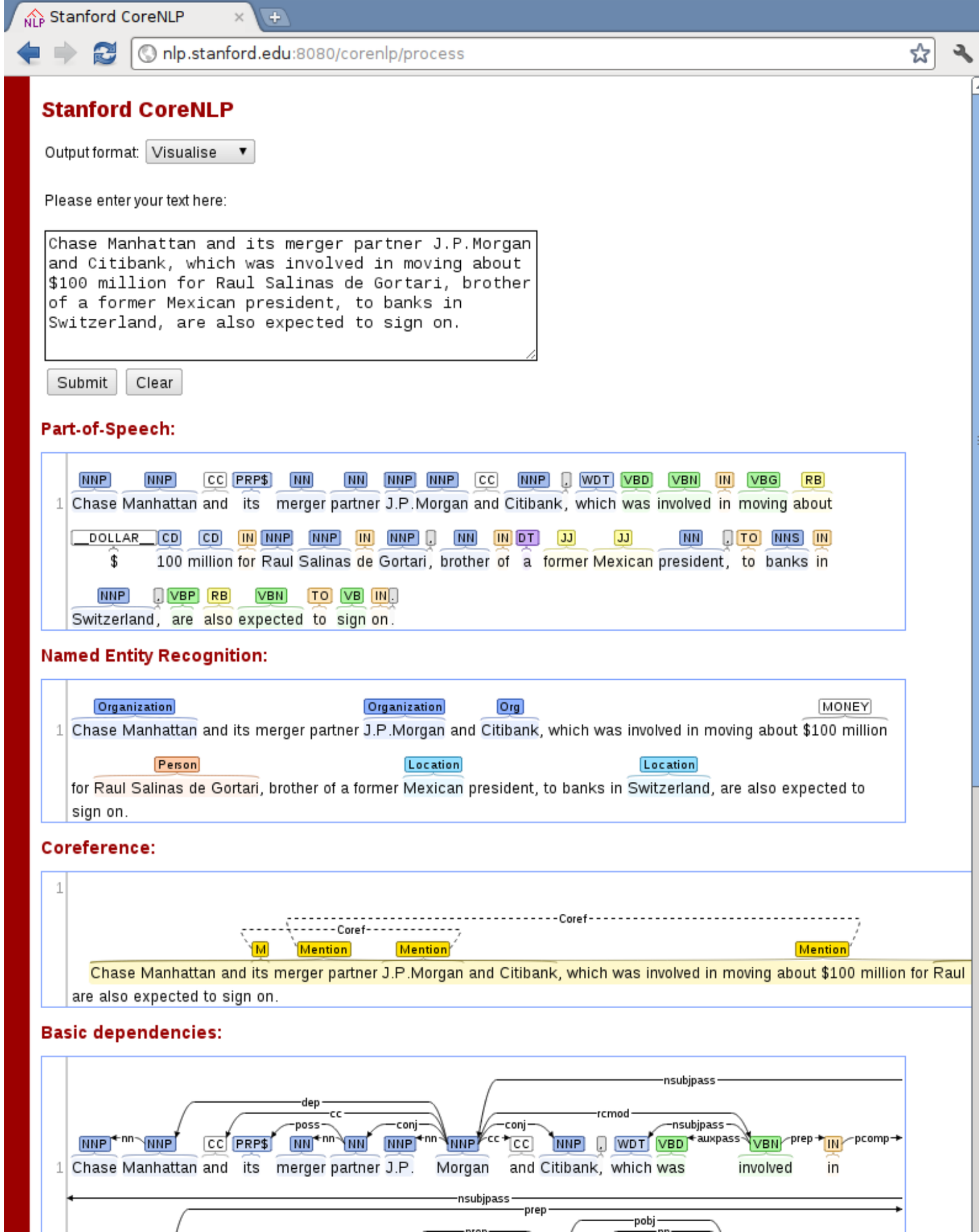
“stand off” Annotation

Qatar Holding sells 10% stake in Porsche to founding families

```
<entity id="e1" type="ORGANIZATION">
  <references>
    <!--Qatar Holding-->
    <span>
      <target id="t1" />
      <target id="t2" />
    </span>
  </references>
  <externalReferences>
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Qatar_Investment_Authority" confidence="1.0" reftype="en" />
  </externalReferences>
</entity>
<entity id="e2" type="ORGANIZATION">
  <references>
    <!--Porsche-->
    <span>
      <target id="t8" />
    </span>
  </references>
  <externalReferences>
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Porsche" confidence="0.9962352" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Porsche_family" confidence="0.0037167938" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Porsche_911" confidence="1.6859167E-5" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Ferdinand_Porsche" confidence="1.6186628E-5" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Porsche_914" confidence="7.881214E-6" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Porsche_Design_Group" confidence="7.0771493E-6" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Porsche_in_motorsport" confidence="2.7061944E-9" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Porsche_550" confidence="7.388618E-12" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Porsche_3512" confidence="1.6147437E-12" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Porsche_RS_Spyder" confidence="1.5382413E-12" reftype="en" />
  </externalReferences>
</entity>
```

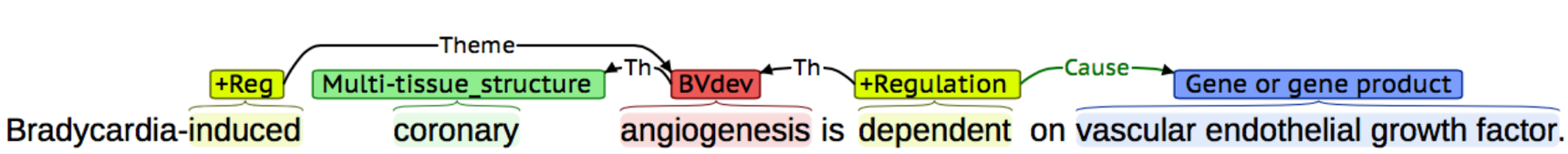
BRAT: Annotation environment

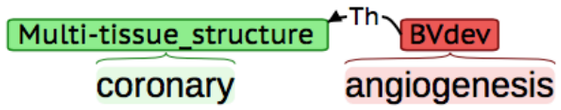
http://brat.nlplab.org/examples.html#corpus-examples-brat

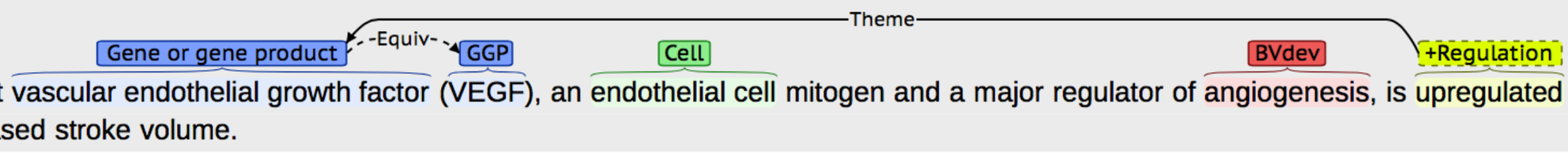


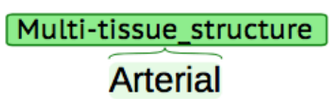
BRAT: Annotation environment

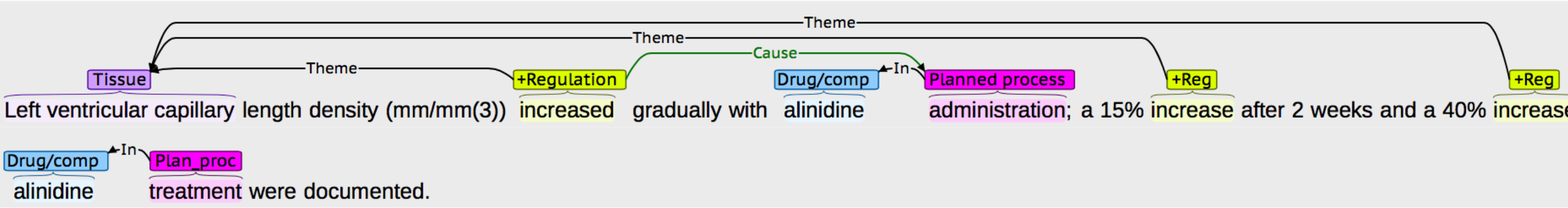
http://brat.nlplab.org/examples.html#corpus-examples-brat

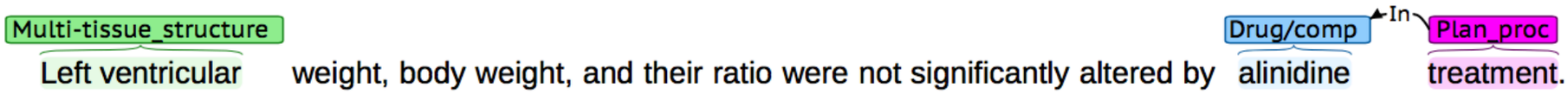












CONLL conferences
Computational Natural Language Learning

```
# text = They buy and sell books.
1    They   they    PRON   PRP   Case=Nom|Number=Plur                2   nsubj 2:nsubj|4:nsubj
2    buy    buy     VERB   VBP   Number=Plur|Person=3|Tense=Pres     0   root  0:root
3    and    and     CONJ   CC    _                                   4   cc    4:cc
4    sell   sell    VERB   VBP   Number=Plur|Person=3|Tense=Pres     2   conj  0:root|2:conj
5    books  book    NOUN   NNS   Number=Plur                         2   obj   2:obj|4:obj
6    .      .       PUNCT  .     _                                   2   punct 2:punct
```

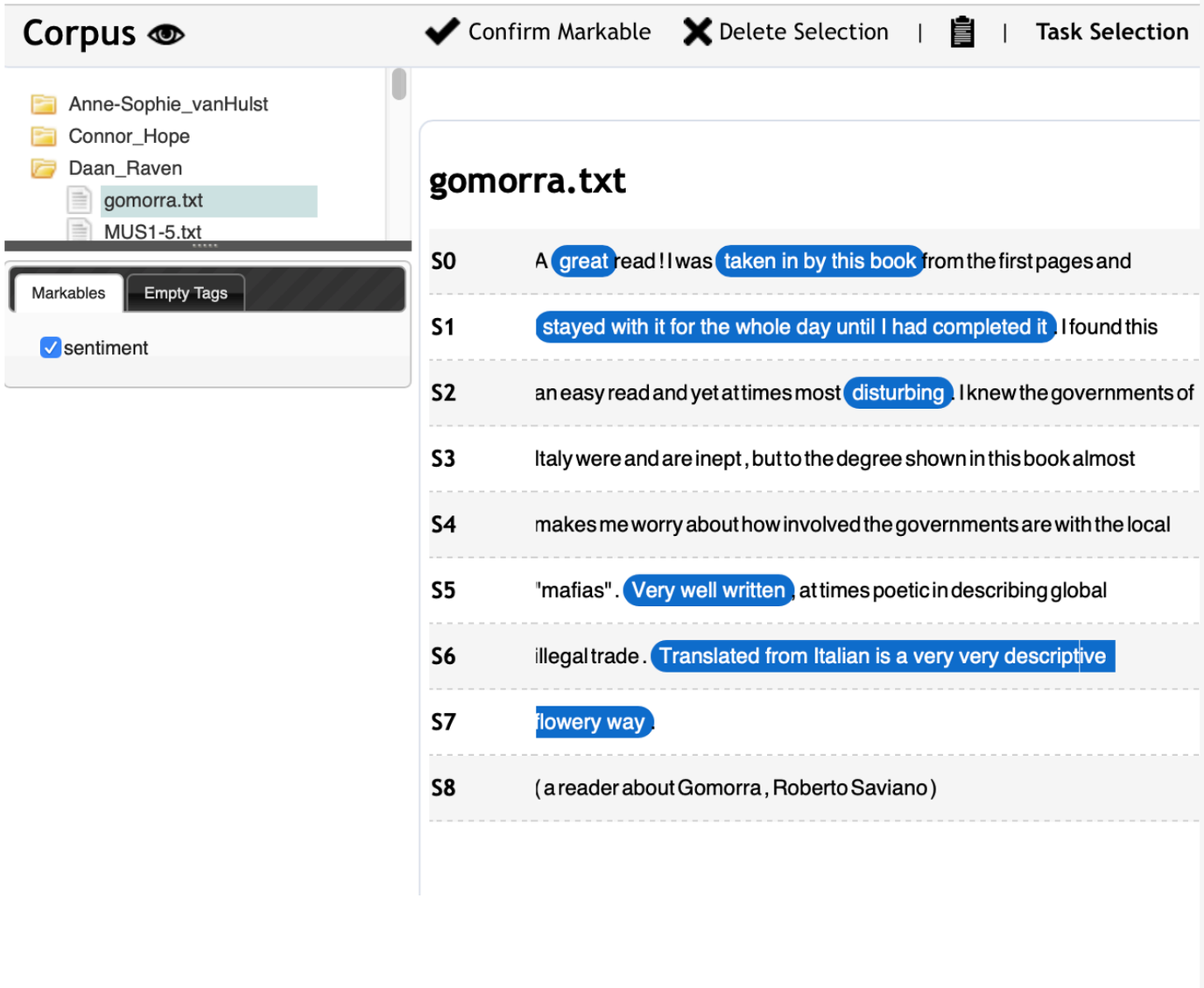
```
# text = U.N. official Ekeus heads for Baghdad.
1    U.N.      UN        NNP   I-NP    I-ORG
2    official  official  NN    I-NP    O
3    Ekeus     Ekeus     NNP   I-NP    I-PER
4    heads     head      VBZ   I-VP    O
5    for       for       IN    I-PP    O
6    Baghdad   Baghdad   NNP   I-NP    I-LOC
7    .         .         Punc  -       O
```

IO(B) style
I = insight
O = outside
B = beginning

- TAB separated columns
- https://www.clips.uantwerpen.be/conll2003/ner/

CAT: the CELCT Annotation Tool

- https://dh.fbk.eu/resources/cat-content-annotation-tool




```
<Document doc_name="gomorra.txt">
<token id="1" sentence="0" number="0">A</token>
<token id="2" sentence="0" number="1">great</token>
<token id="3" sentence="0" number="2">read</token>
<token id="4" sentence="0" number="3">!</token>
<token id="5" sentence="0" number="4">I</token>
<token id="6" sentence="0" number="5">was</token>
<token id="7" sentence="0" number="6">taken</token>
<token id="8" sentence="0" number="7">in</token>
<token id="9" sentence="0" number="8">by</token>
<token id="10" sentence="0" number="9">this</token>
<token id="11" sentence="0" number="10">book</token>
<token id="12" sentence="0" number="11">from</token>
<token id="13" sentence="0" number="12">the</token>
<token id="14" sentence="0" number="13">first</token>
<token id="15" sentence="0" number="14">pages</token>
<token id="16" sentence="0" number="15">and</token>
<token id="17" sentence="1" number="0">stayed</token>
<token id="18" sentence="1" number="1">with</token>
<token id="19" sentence="1" number="2">it</token>
<token id="20" sentence="1" number="3">for</token>
<token id="21" sentence="1" number="4">the</token>
<token id="22" sentence="1" number="5">whole</token>
<token id="23" sentence="1" number="6">day</token>
<token id="24" sentence="1" number="7">until</token>
<token id="25" sentence="1" number="8">I</token>
<token id="26" sentence="1" number="9">had</token>
<token id="27" sentence="1" number="10">completed</token>
<token id="28" sentence="1" number="11">it</token>
<token id="29" sentence="1" number="12">.</token>

<Markables>›
<SENTIMENT id="4" polarity="POS"  >
<token_anchor id="2"/>
</SENTIMENT>
<SENTIMENT id="6" polarity="POS"  >
<token_anchor id="7"/>
<token_anchor id="8"/>
<token_anchor id="9"/>
<token_anchor id="10"/>
<token_anchor id="11"/>
</SENTIMENT>
```

Text

Words

Resource

Interpretation

The new US president has said he believes that vaccines are harmful and has repeatedly and erroneously suggested that they cause autism

believe

suggest

vaccine

harmful

vaccine

cause

autism

Rules & Lexical Database

or

Trained Classifier

cure

threat

chemical substance

negative

positive

causal relation

negative condition

Somebody needs to build the lexicon & rules

Somebody needs to annotate training data

Knowledge or data driven

Somebody needs to build the lexicon & rules

Somebody needs to annotate training data

Knowledge-Based	Machine Learning Systems
• based on hand-coded rules	• use statistics or other machine learning
• developed by NLP specialists	• developers do not need NLP expertise
• make use of human intuition	• requires large amounts of training data
• easy to understand results	• cause of errors is hard to understand
• development could be very time consuming	• development is quick and easy
• changes may require rewriting rules	• changes may require re-annotation

Table 2.1: Summary of Knowledge-Based vs Machine Learning Approaches to NLP

Lexical database or ontology

Word	Properties
Apple	Name, Company
Samsung	Name, Company
infringe	Verb, negative
its	Pronoun, possessive

Rules

If
Company (w) &
Negative (w+1) &
Event (w+1)
then
NEG (w)

Data	Features	Model
Samsung lied to Apple [neg]	[1,0,0,0,1,0,1] neg	Weights
Samsung bought patents from Apple [pos]	[1,0,1,1,0,0,1] pos	
Volkswagen deceived Merkel	[1,0,0,0,1,0,1] ?	

Part V: Deep Learning

Types of machine learning in NLP

explicit feature representation

Decision trees (DT)	closest to rules
Hidden Markov Models (HMM)	part-of-speech sequences
Naive Bayes (NB)	bag-of-words
Maximum Entropy (ME)	class distribution
Support Vector Machine (SVM)	data delineation
K-nearest-neighbors (KNN)	distance in feature space
Conditional Random Fields (CRF)	sequential dependencies

Neural networks (NN)	hidden feature representation

Part V: Deep Learning

Feature representations for text

- Any property of the text except for the label to be predicted
- The (surrounding) words….
- Lexical word properties: part of speech, meaning, sentiment
- Sentence properties: syntax, semantic roles, sentiment
- Document properties: topic, sentiment, genre, publication date
- ***All features are mapped to vector representations!***

Word to Document index

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

Figure 15.2 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.

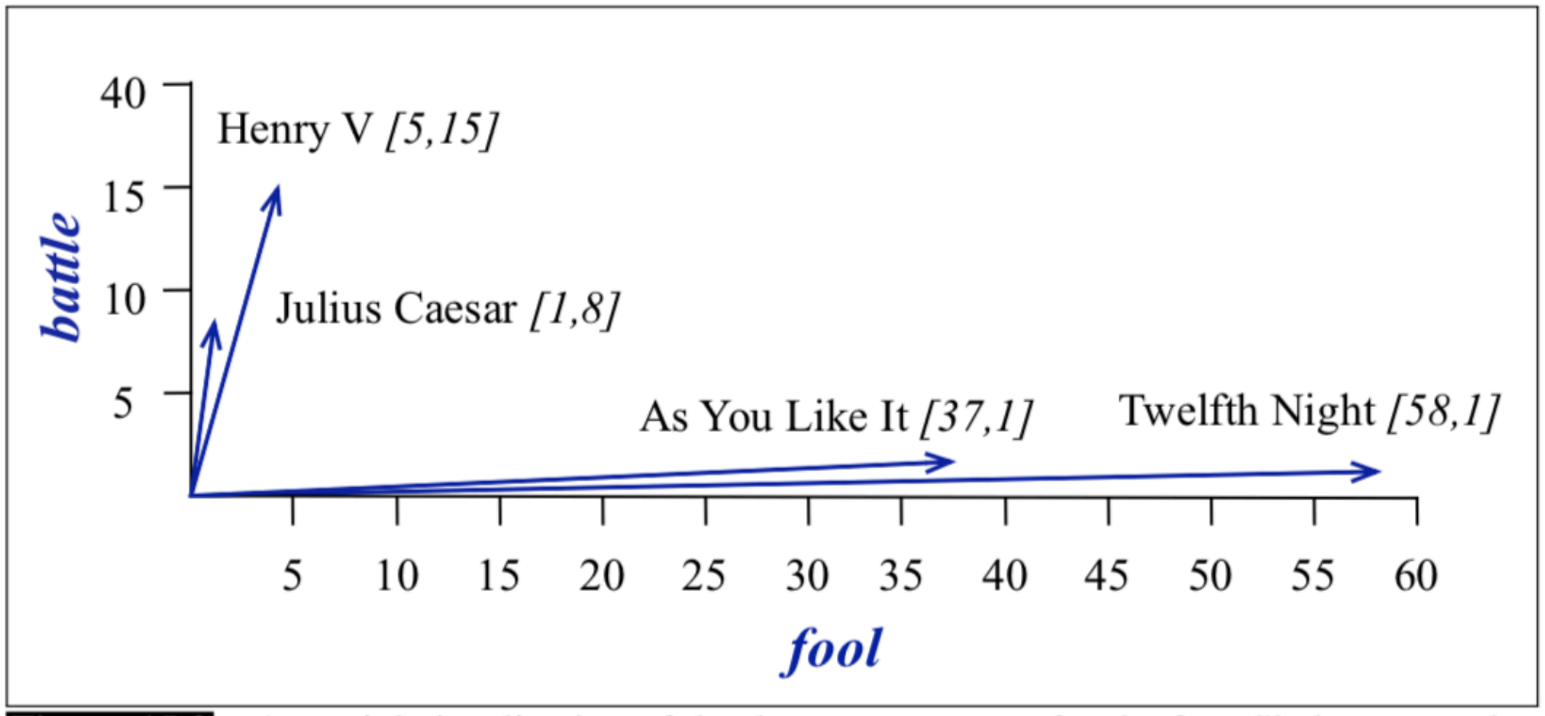


Figure 15.3 A spatial visualization of the document vectors for the four Shakespeare play documents, showing just two of the dimensions, corresponding to the words *battle* and *fool*. The comedies have high values for the *fool* dimension and low values for the *battle* dimension.

Bag of words (BoW) model of text

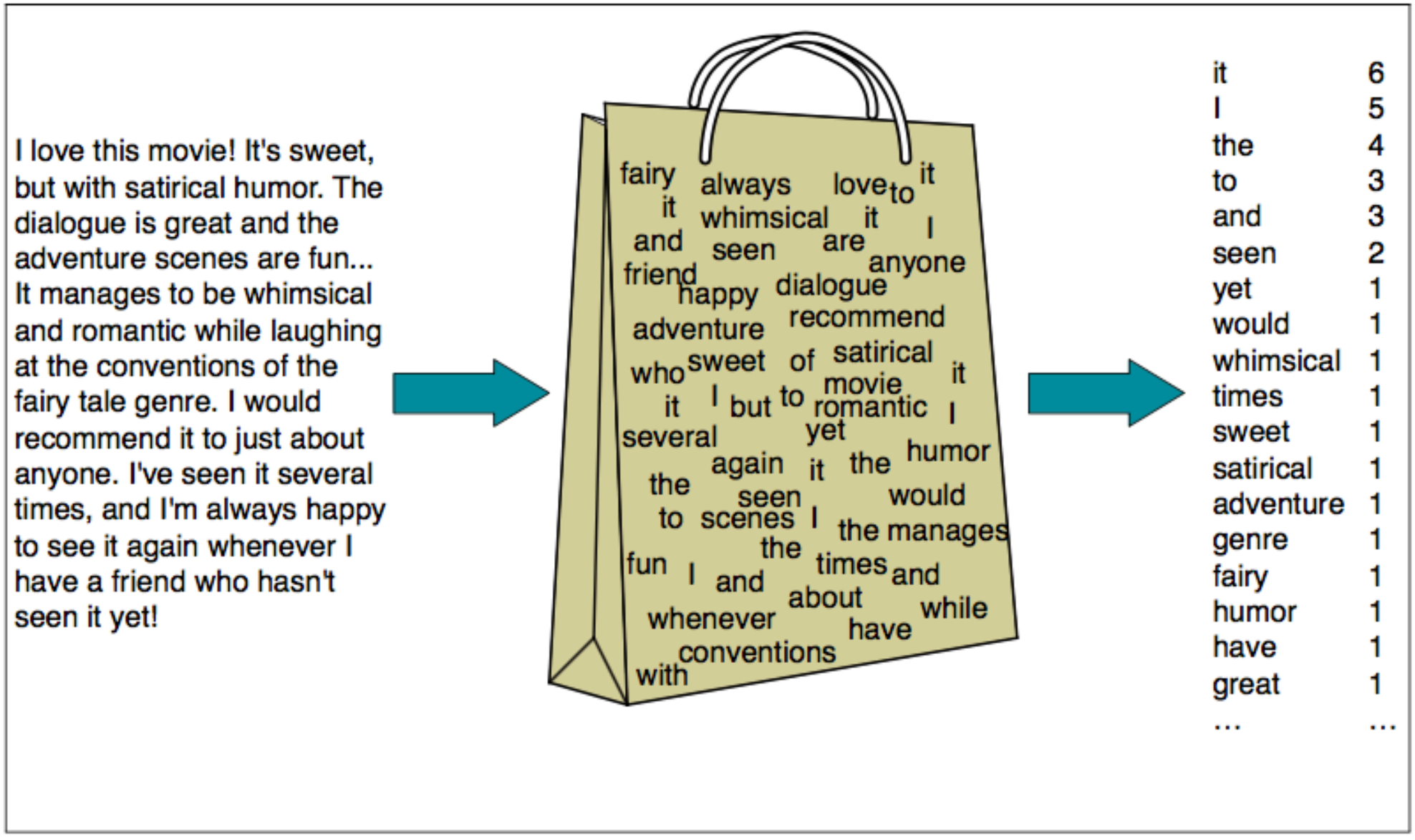


Figure 6.1 Intuition of the multinomial naive Bayes classifier applied to a movie review. The position of the words is ignored (the *bag of words* assumption) and we make use of the frequency of each word.

Not all words are equal
Information value (Spark-Jones 1972)

- Observations:
  - Words that occur in all texts, e.g. ***a***, ***the***, ***case***, ***of***, ***is*** have a low information value, despite their frequency
  - Words that occur once in one text are too idiosyncratic
  - Good words are typical for a topic or domain: shared to some extent but not too often
- IV (**t**|**d**) = TF*IDF = Term Frequency * Inverse Document Frequency

$$IV(\mathbf{t}|\mathbf{d}) = \frac{\text{Frequency of term } \boldsymbol{t} \text{ in document } \boldsymbol{d}}{\text{Total number of documents in which term } \boldsymbol{t} \text{ occurs}}$$

More than words, less than words
N-Grams

- Sequences of 1, 2, 3, 4, 5 etc. words:
  - the, room, was, tidy (unigram, 1-gram)
  - the_room, room_was, was_tidy (bi-gram, 2-gram)
  - the_room_was, room_was_tidy (tri-gram, 3-gram)
- Character n-grams:
  - [t, th, he, e_, _r, ro, oo, om, m_, _w, wa, as, s_, _t, ti, id, dy, y]

Word combinations
Pointwise Mutual Information

sum columns: 0, … 3, 7, 2, 5, 2 — total nr. of tokens = 19

	aardvark	...	computer	data	pinch	result	sugar	...	sum rows
apricot	0	...	0	0	1	0	1		1
pineapple	0	...	0	0	1	0	1		2
digital	0	...	2	1	0	1	0		4
information	0	...	1	6	0	4	0		11

Figure 15.4 Co-occurrence vectors for four words, computed from the Brown corpus, showing only six of the dimensions (hand-picked for pedagogical purposes). The vector for the word *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

Pointwise Mutual Information: is a combination more frequent than expected?

$$\text{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

P = probability

$$P(\text{w=information,c=data}) = \frac{6}{19} = .316$$

$$P(\text{w=information}) = \frac{11}{19} = .579$$

$$P(\text{c=data}) = \frac{7}{19} = .368$$

$$\text{ppmi}(\text{information,data}) = \log 2(.316/(.368 * .579)) = .568$$

How to represent words
Word vectors: distributional representation of words

- Distributional hypothesis:
  - the meaning of a word is defined by the surrounding words (the company it keeps).

- A bottle of tesgüino is on the table.
- Everybody likes tesgüino.
- Tesgüino makes you drunk.
- We make tesgüino out of corn.

Computational implementation
Word2Vec
Mikolov & Chen et al. 2013
Mikolov & Sutskever et al. 2013

Harris 1954: similar contexts define similar meanings

Words in context
How to differentiate concepts

- I'm **a bass (I** can reach the **low B**, even a **low A** in the mornings). … If you have a low **register** voice, it's also a nice **song** to **sing** with a warm **voice**, **register** wise. **The bass range** starts at a certain **C**.

- Whether you're a competitive sports **fisher** or just want to spend a fun afternoon with your family on the **lake**, **bass fishing** is a difficult pastime if you aren't properly equipped.

- In **choral music**, **the alto range** is from **G3** referring to **G** below **key** of **C** middle to **F5** which is the **F** in 2nd **octave** above the **C** middle.

- **Waxworms** are used as live-b **for trout fishing**. **Corn worms** are also excellent live-**bait when** trout **fishing**. **Nymph** of a golden **stonefly** are used as live-bait **for trout fishing**. Nymph mayfly. **salmon roe (Red caviar) Worms** are cheap and a great **bait** to use **for trout and** most types of **fish**.

- Important information **about trout fishing** on **Lake** Taupo

Word Vectors or Word embeddings

Word is represented by context in use

- **voice**, **register** wise. The **bass range** starts at a certain **C**.
- the **alto range** is from **G3** referring to **G** below **key** of **C**
- your **family** on the **lake**, **bass fishing** is a difficult pastime
- Important information about **trout fishing** on **Lake** Taupo

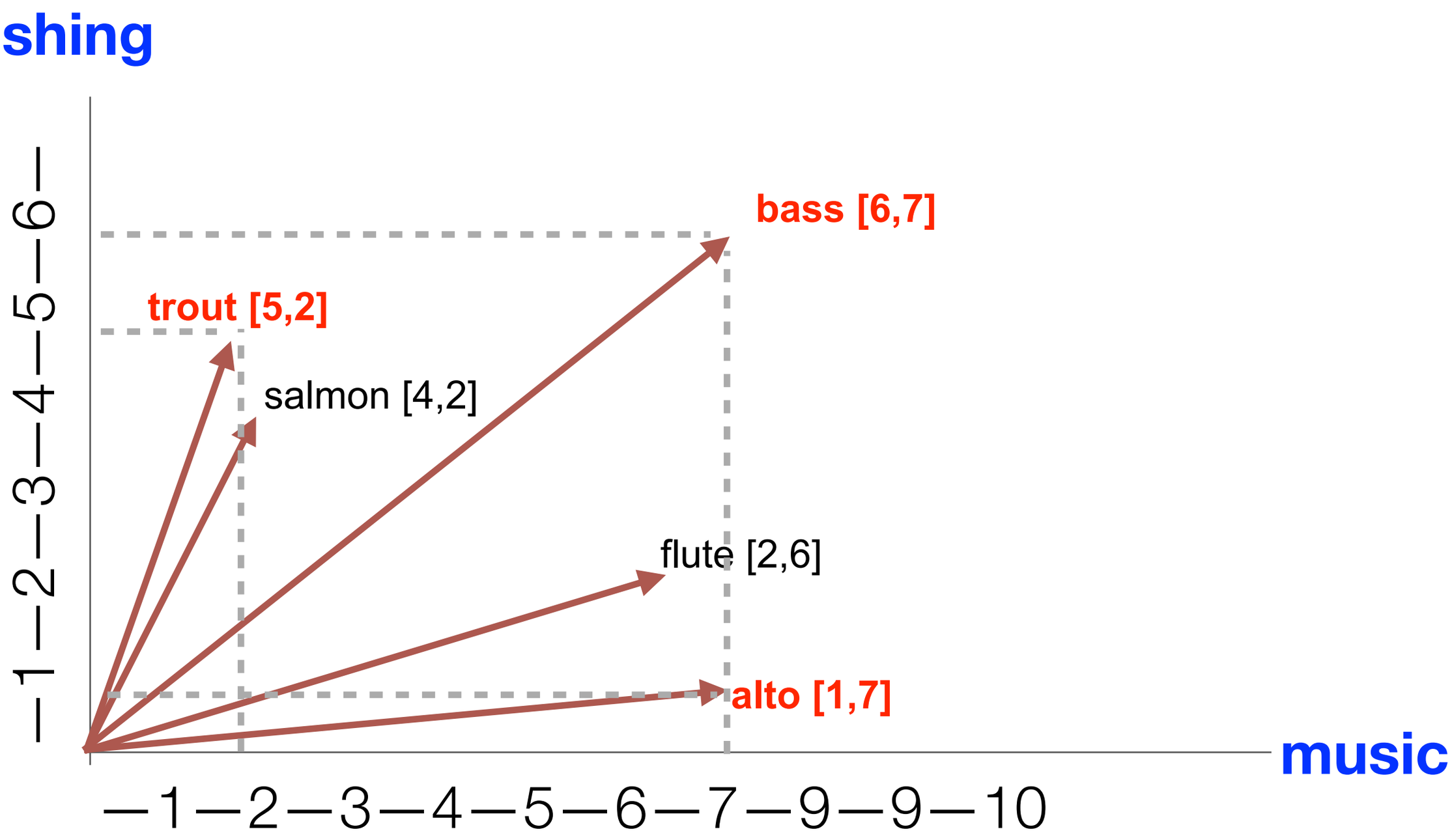
vector size = vocabulary

C	family	fishing	G	G3	lake	key	range	register	voice
[1,	0,	0,	0,	0,	0,	0,	1,	1,	1]
[1,	0,	0,	1,	1,	0,	1,	1,	0,	0]
[0,	1,	1,	0,	0,	1,	0,	0,	0,	0]
[0,	0,	1,	0,	0,	1,	0,	0,	0,	0]

similarity of word context is the normalised dot product of the vectors =

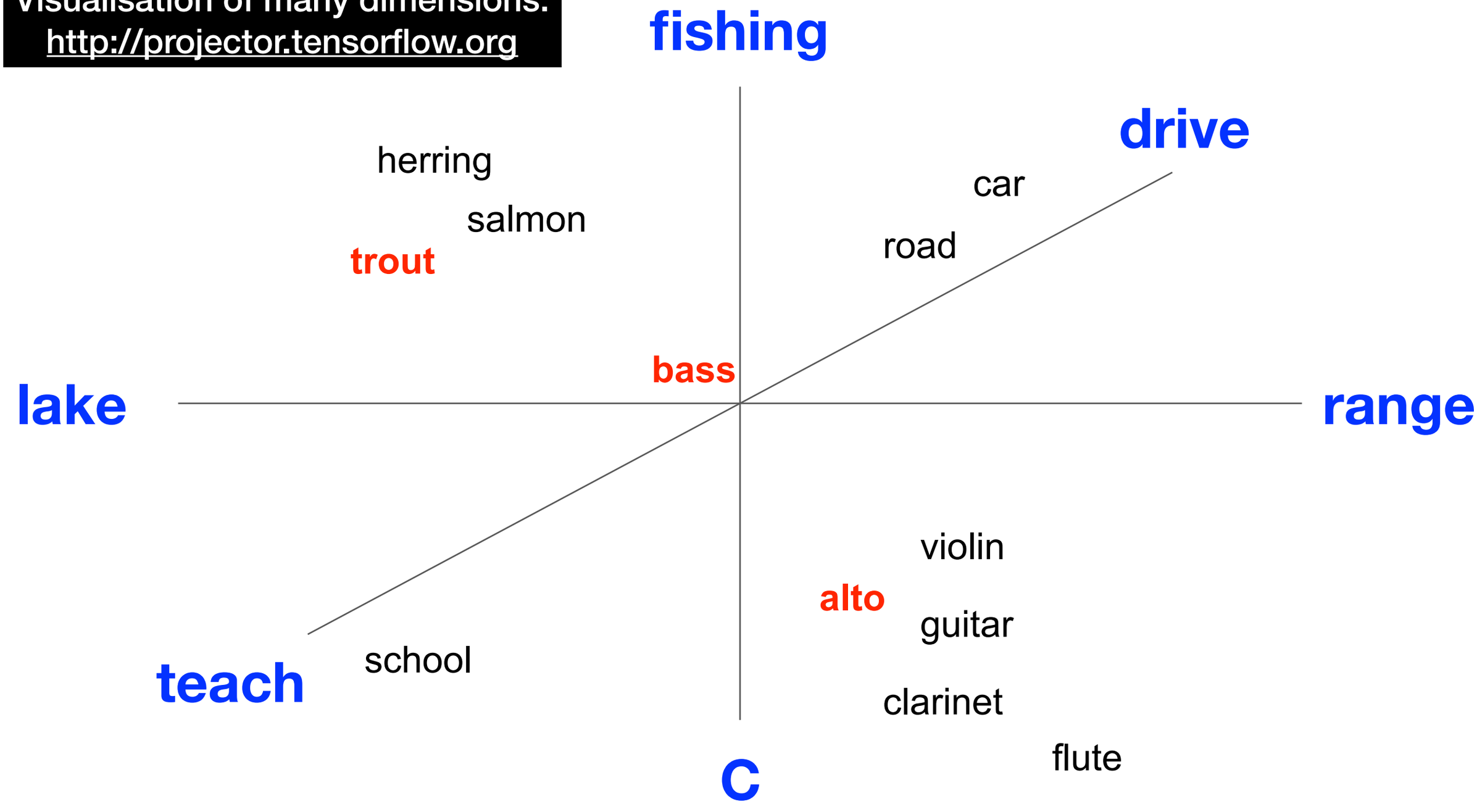
$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} \frac{a_i b_i}{n} = \frac{a_1 b_1 + a_2 b_2 + \cdots + a_n b_n}{n}$$

Simplified view in two dimensions
Word Vectors or Word embeddings



Visualisation in n-dimensions
Word Vectors or Word embeddings

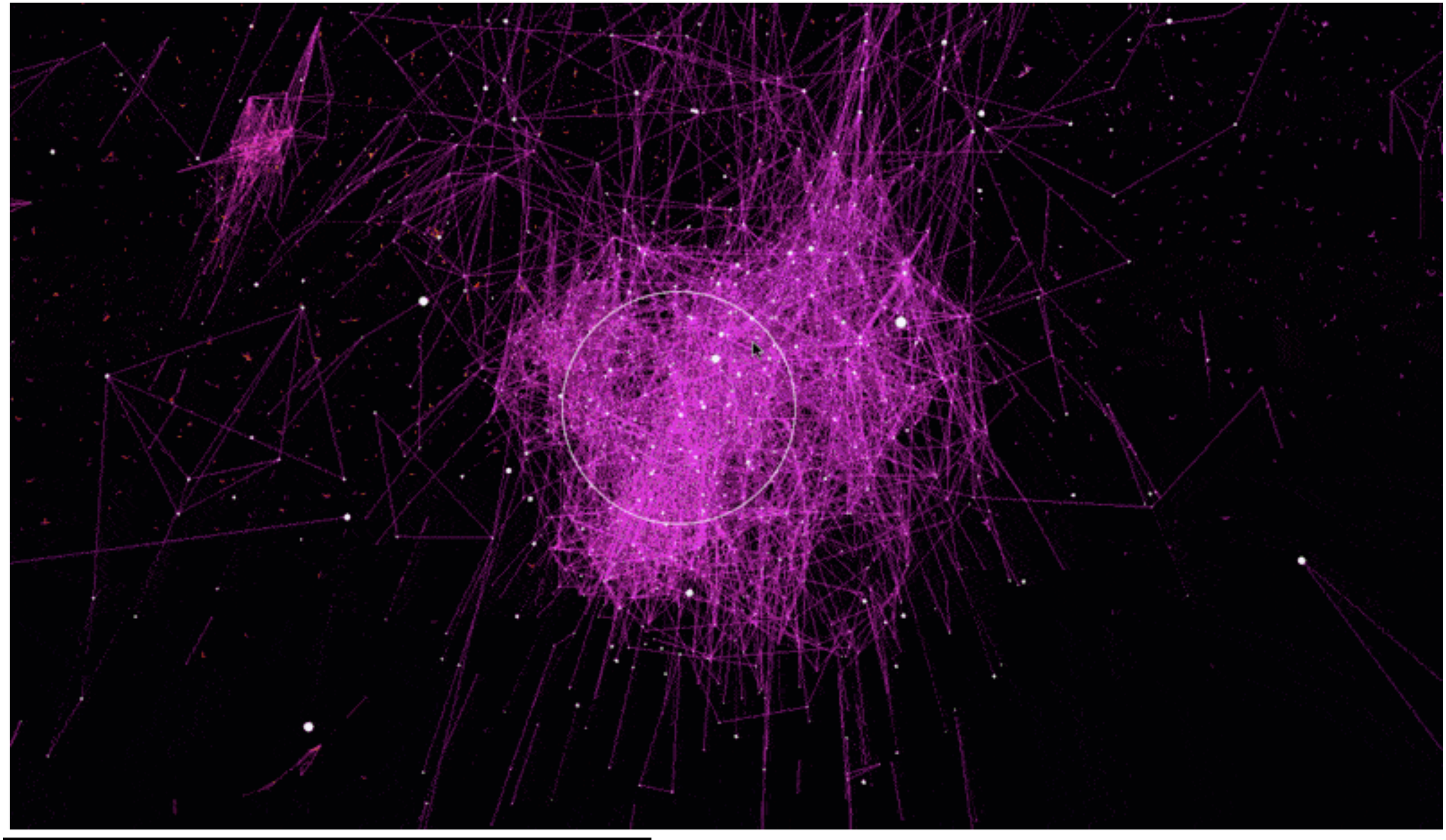
Visualisation of many dimensions:
http://projector.tensorflow.org



Stanford NLP: Jeffrey Pennington, Richard Socher, Christopher Manning

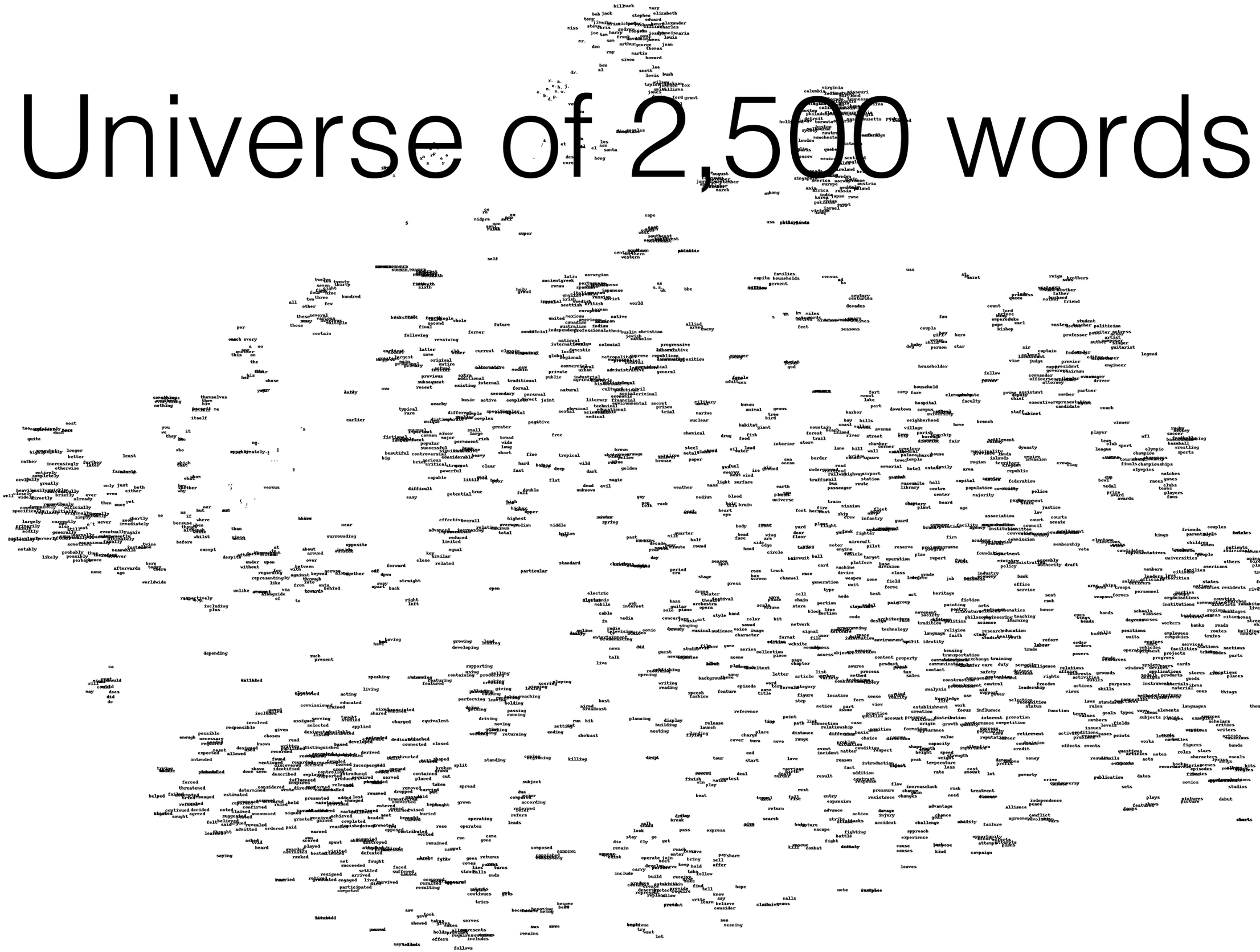
Glove2B: Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 300d vectors, 822 MB)

https://nlp.stanford.edu/projects/glove/

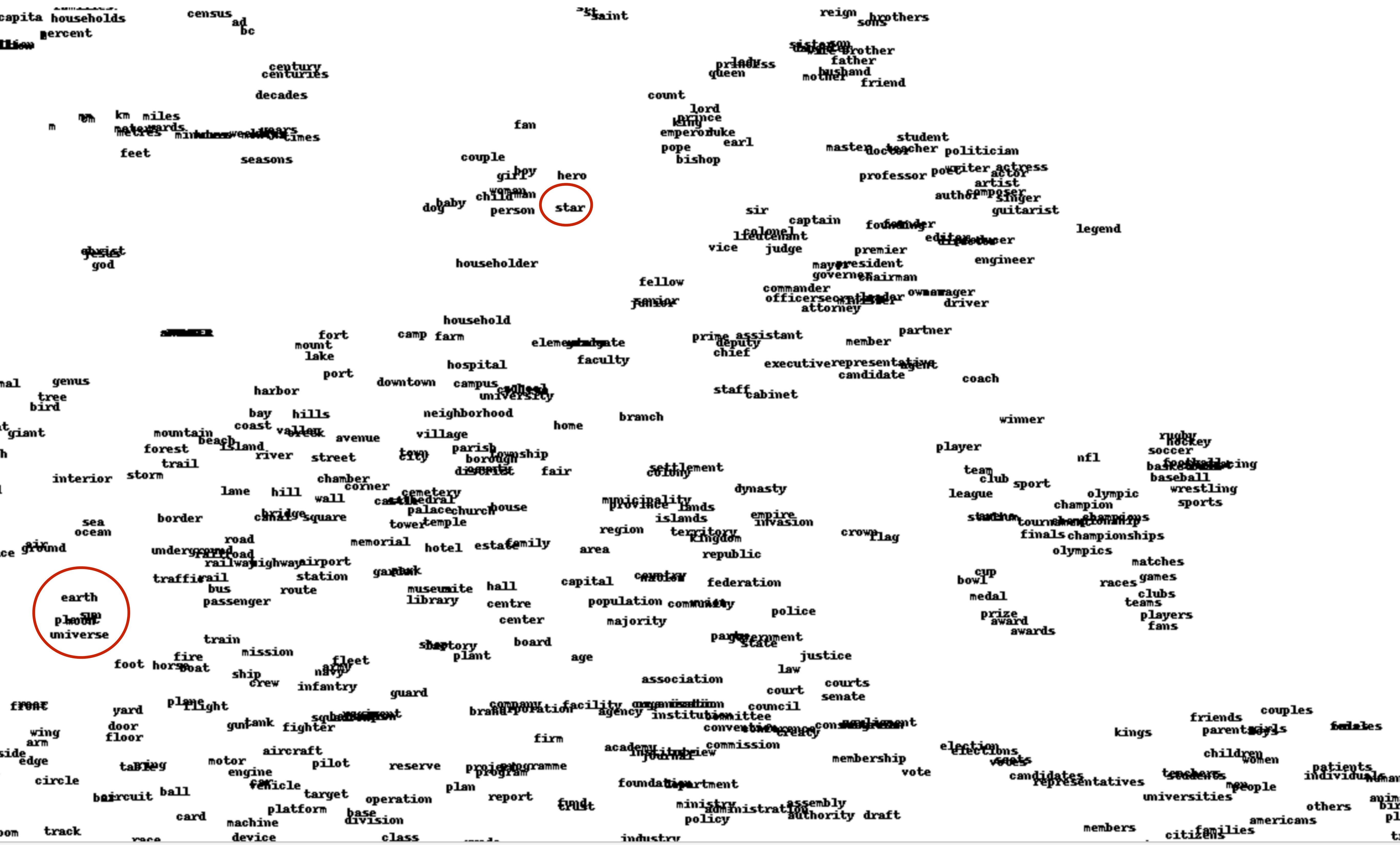


https://github.com/anvaka/word2vec-graph

Universe of 2,500 words

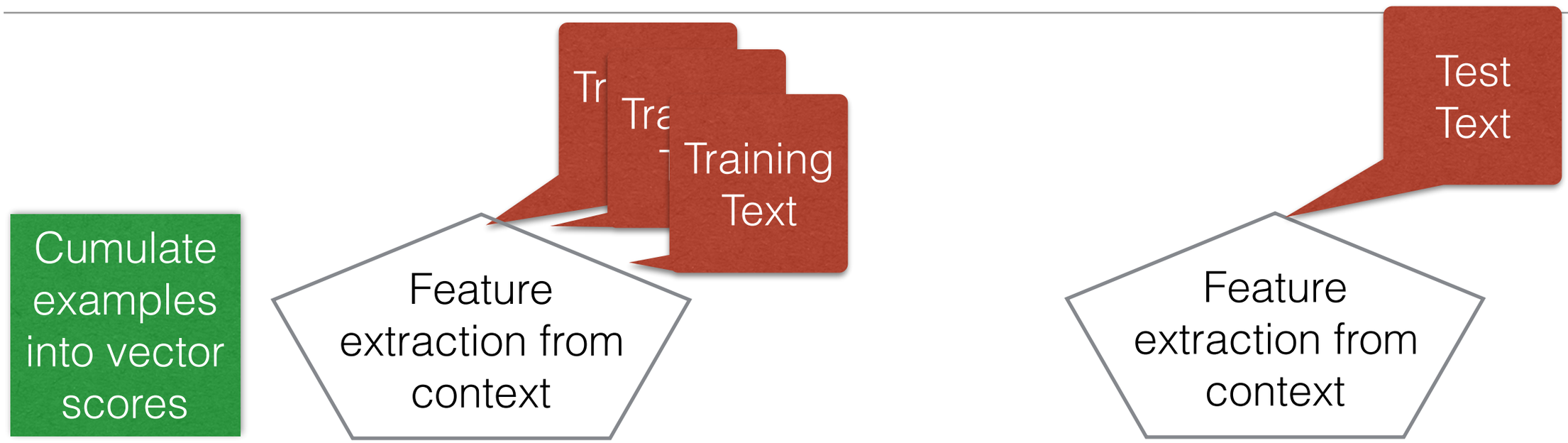


Joseph Turian's map of 2500 English words produced by using t-SNE on the word feature vectors learned by Collobert & Weston, ICML 2008



households
census
percent
saint
reign
brothers
sons
father
husband
mother
friend
queen
century
centuries
decades
count
lord
emperor
duke
earl
pope
bishop
km
miles
feet
times
seasons
fan
couple
hero
boy
girl
child
woman
baby
dog
person
star
student
teacher
doctor
master
politician
professor
poet
writer
actor
actress
artist
author
composer
singer
guitarist
sir
captain
founder
colonel
lieutenant
vice
judge
editor
premier
engineer
legend
god
householder
mayor
president
governor
chairman
fellow
commander
officer
leader
manager
owner
driver
attorney
junior
household
fort
mount
lake
camp
farm
hospital
faculty
assistant
deputy
chief
member
partner
executive
representative
candidate
agent
coach
genus
tree
bird
harbor
port
downtown
campus
college
university
staff
cabinet
neighborhood
home
branch
winner
giant
mountain
coast
valley
creek
avenue
village
bay
hills
beach
island
forest
river
street
trail
parish
township
borough
town
city
district
fair
settlement
colony
player
nfl
rugby
hockey
soccer
football
baseball
basketball
racing
interior
storm
lane
hill
chamber
corner
wall
cemetery
cathedral
castle
palace
church
tower
temple
house
dynasty
team
club
sport
league
olympic
wrestling
sports
champion
championship
tournament
championships
stadium
finals
olympics
border
canal
bridge
square
sea
ocean
ground
road
underground
railroad
railway
highway
airport
memorial
hotel
estate
family
municipality
province
lands
islands
region
empire
invasion
territory
kingdom
crown
flag
area
republic
traffic
rail
bus
station
route
passenger
museum
library
hall
centre
center
capital
nation
country
federation
population
community
majority
police
cup
bowl
medal
prize
award
awards
matches
games
races
clubs
teams
players
fans
earth
moon
sun
planet
universe
train
mission
factory
plant
board
age
party
government
state
justice
law
foot
horse
boat
fire
fleet
army
navy
ship
crew
infantry
guard
association
court
courts
senate
yard
plane
flight
door
floor
gun
tank
fighter
company
corporation
facility
agency
organization
institution
committee
council
convention
conference
commission
firm
academy
institute
journal
election
elections
membership
vote
votes
seats
kings
friends
couples
parents
girls
boys
children
women
wing
arm
edge
circle
table
motor
engine
aircraft
pilot
reserve
vehicle
car
target
operation
program
programme
plan
report
foundation
department
candidates
representatives
teachers
students
men
people
patients
individuals
universities
ball
card
platform
machine
base
division
class
device
fund
trust
ministry
administration
assembly
authority
draft
policy
industry
members
others
americans
citizens
families
track

But wait a minute!
Embeddings can replace words as features



same vector positions and length across

LABEL [0,1,0,0,1,0,1,1,0]

LABEL [1,0,1,0,1,0,0,0,1]

[1,0,0,0,1,0,1,0,0]

feature vectors

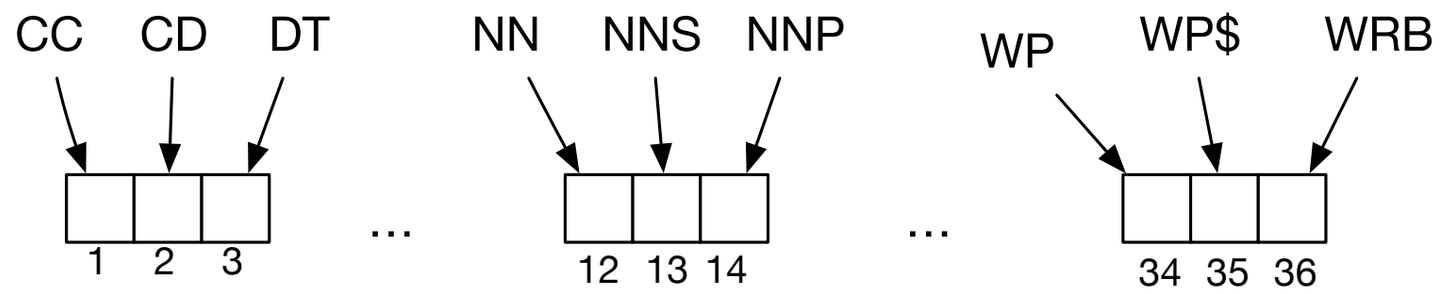
feature vector

Cosine similarity of vectors
normalised dot product =

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} \frac{a_i b_i}{n} = \frac{a_1 b_1 + a_2 b_2 + \cdots + a_n b_n}{n}$$

Part-of-Speech as one-hot-encodings

- 36 tags in the Penn Treebank: 36 dimensions (one-hot representation)



CC <1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

CD <0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

DT <0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

NN <0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

NNS <0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

NNP <0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

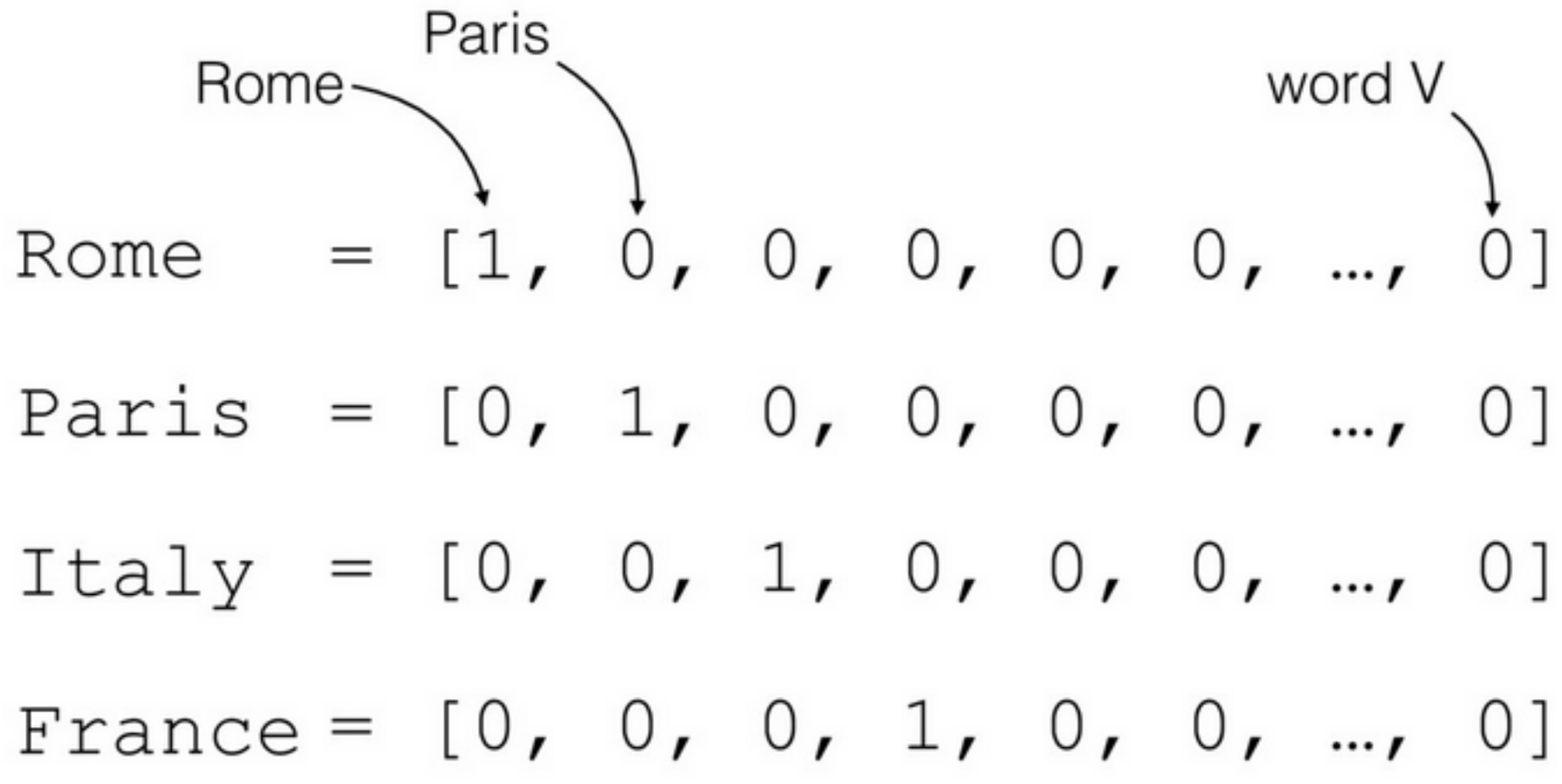
WP <0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0>

WP$ <0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0>

WRB <0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1>

One-hot encodings represent words as features

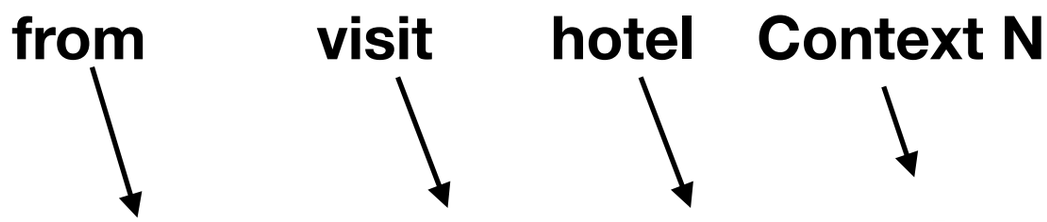
- Lengths of the vocabulary is the size of the dimensions



Source image: Shaffy, Athif (2017) Vector Representation of Text for Machine Learning. https://medium.com/@athif.shaffy/one-hot-encoding-of-text-b69124bef0a7

Embeddings represent families of related words as features

Lengths of the vocabulary is the size of the embeddings



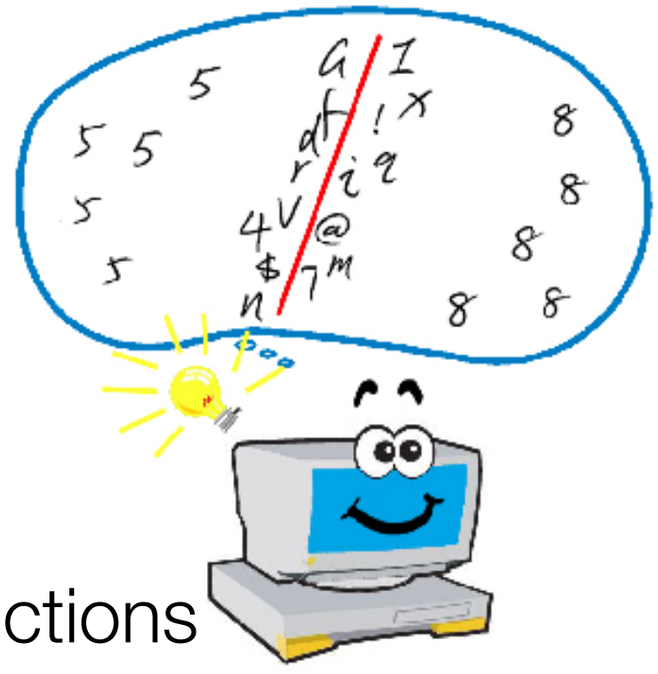
Rome = [0.1, 0.4, 0.3, 0.9, 0.1, 0.8, 0.1]

Paris = [0.1, 0.3, 0.4, 0.8, 0.1, 0.9, 0.1]

Italy = [0.2, 0.7, 0.3, 0.5, 0.1, 0.4, 0.1]

France = [0.2, 0.7, 0.3, 0.5, 0.1, 0.4, 0.1]

Machine learning models



- Is about fitting the data
- Adjust model to avoid errors and make the right predictions
- Problems:
  - Data sparseness (underestimated variation and dynamics)
  - Overfitting:
    - you learn specifics from the training data that should not be learned
    - you fail to generalise meaningful patterns

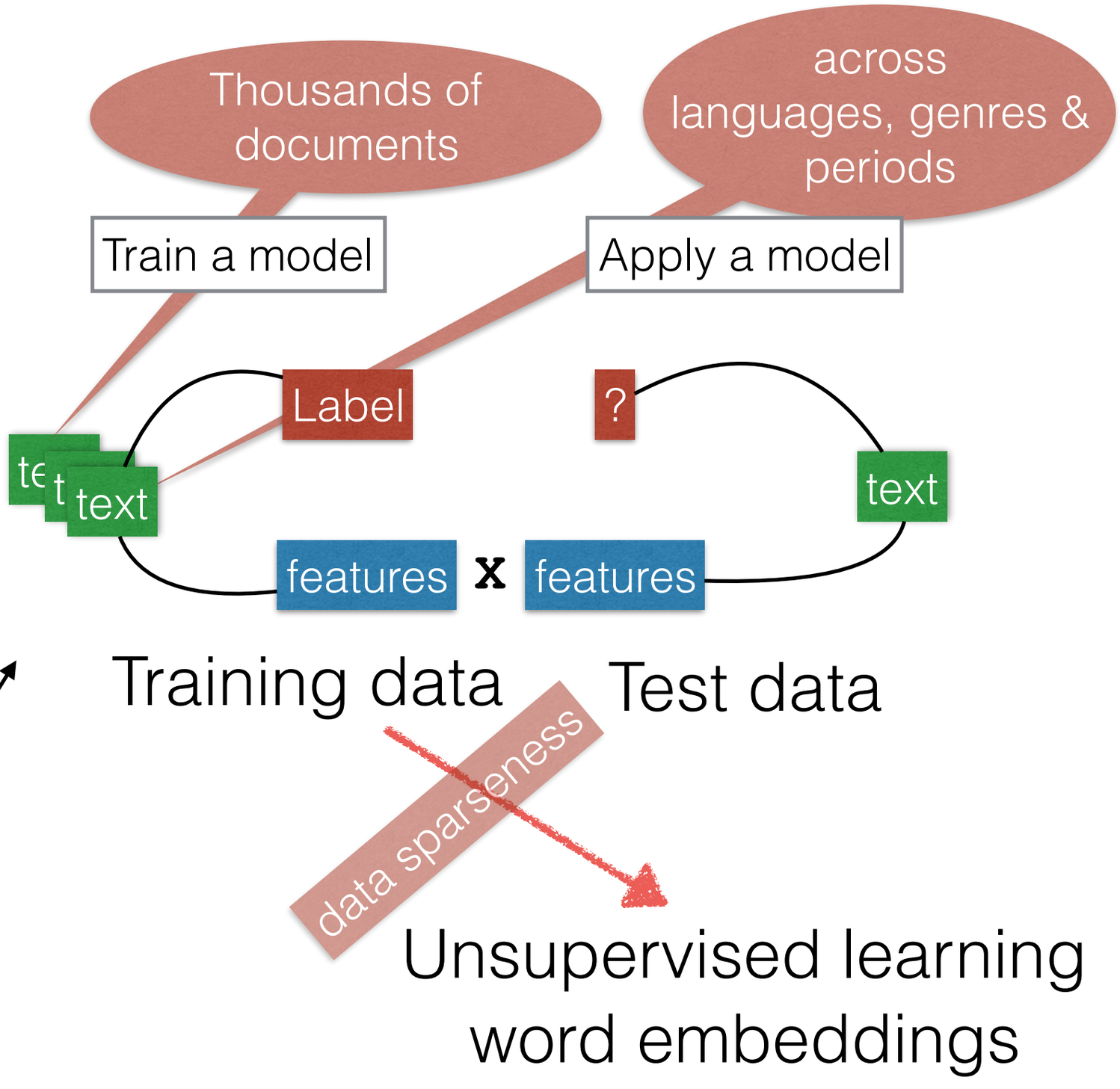
Disadvantage Supervised ML

- Requires handcrafted annotations (supervision)
  - time-consuming and expensive
  - over-specified and still incomplete
  - again for each task/domain
  - again every 5 or 10 years
  - features have become extremely complex and rich over the years: the art of feature engineering!

Annotating text to train NLP

For each module in a pipeline

- morphology
- grammar
- entities
- concepts
- events
- time
- locations
- sentiment, emotions



Expand

Annotated data

Clustered data, not annotated

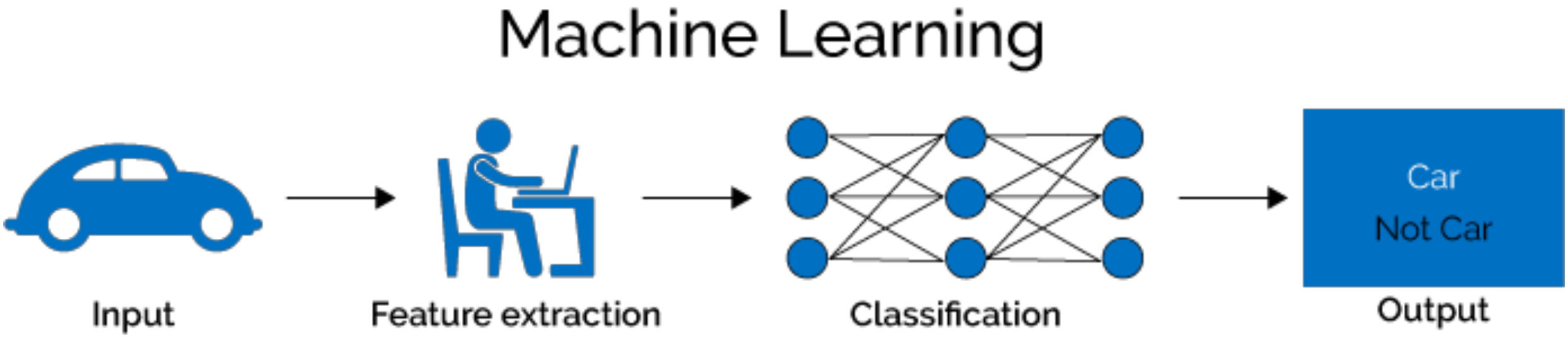
Supervised Learning

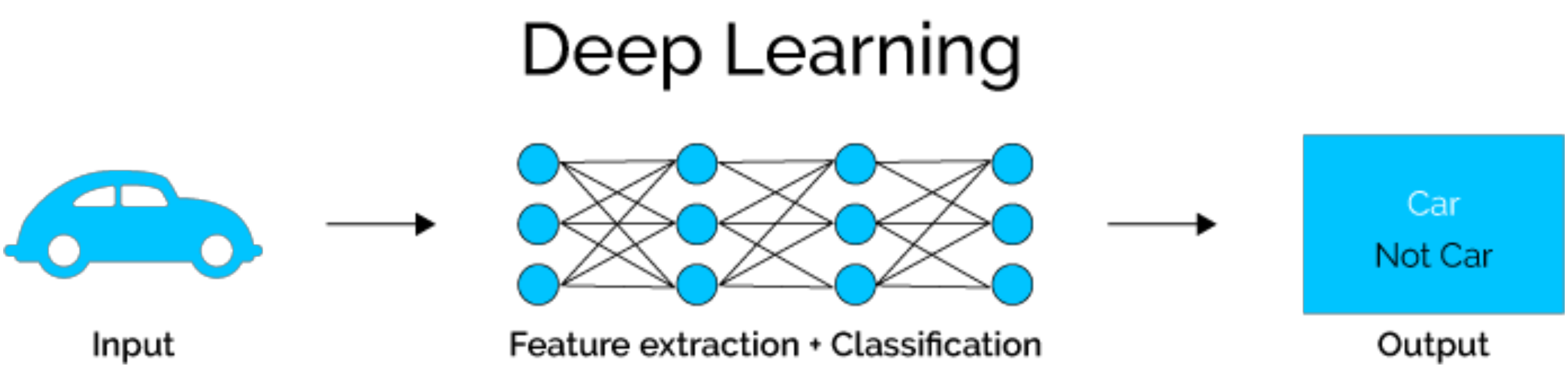
Unsupervised Learning

Tagged training data and feature engineering

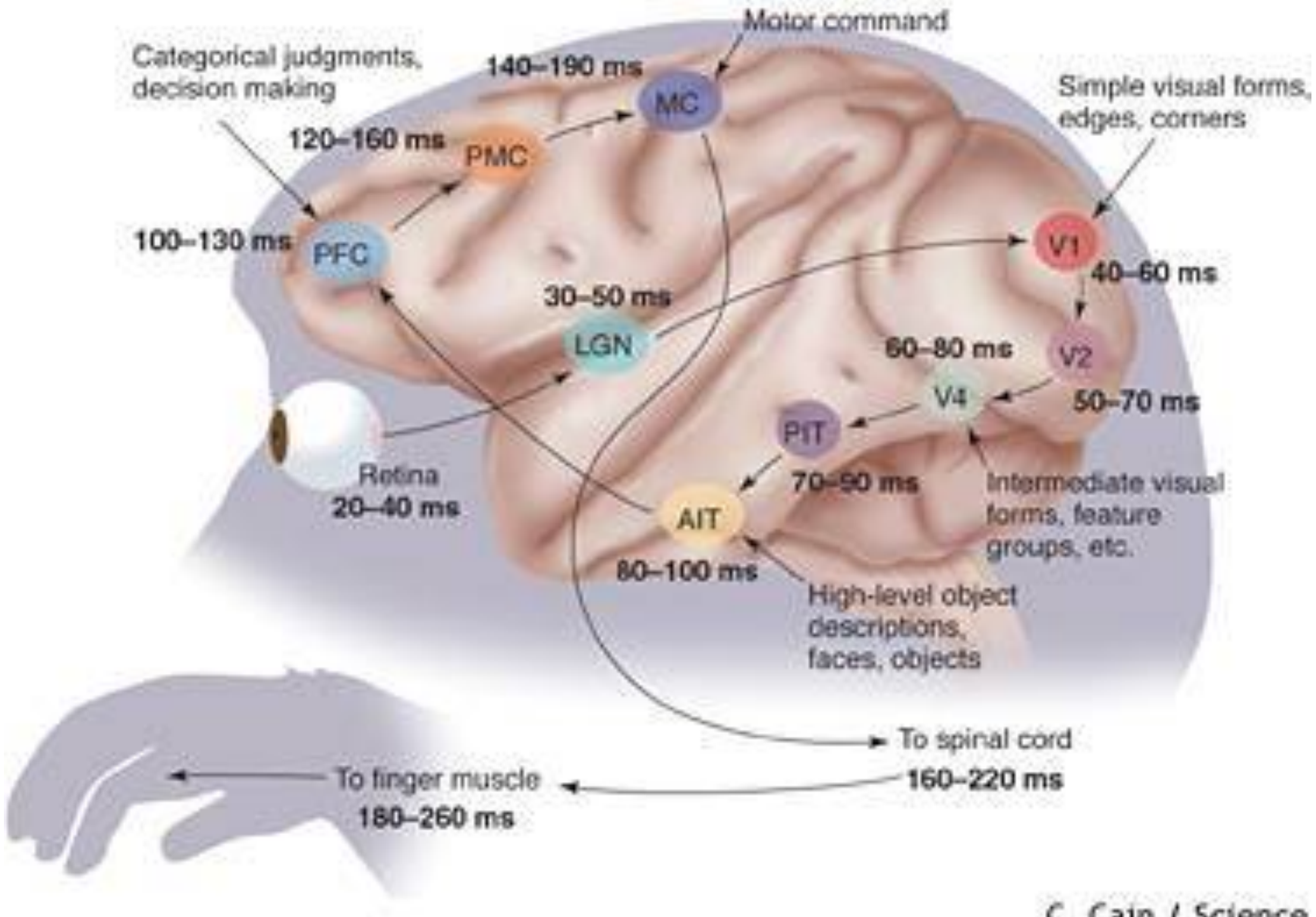
Data clustering and feature engineering

Part III: Deep learning



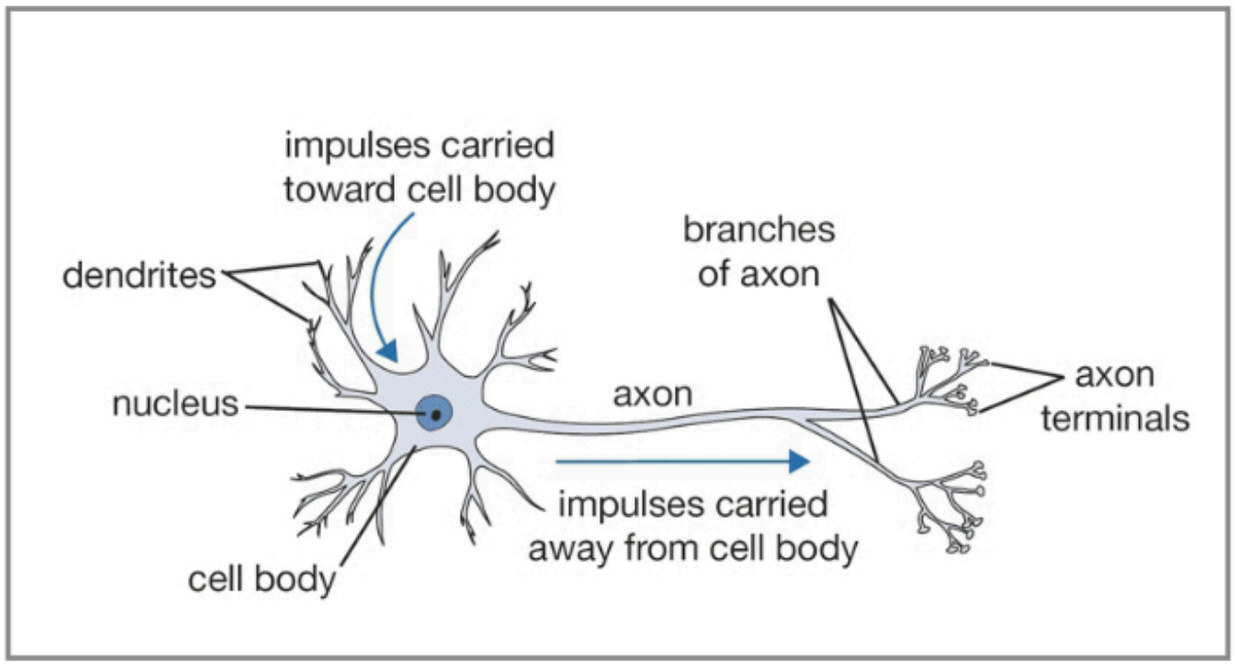


Abstraction in the brain



C. Cain / Science

Brain analogy



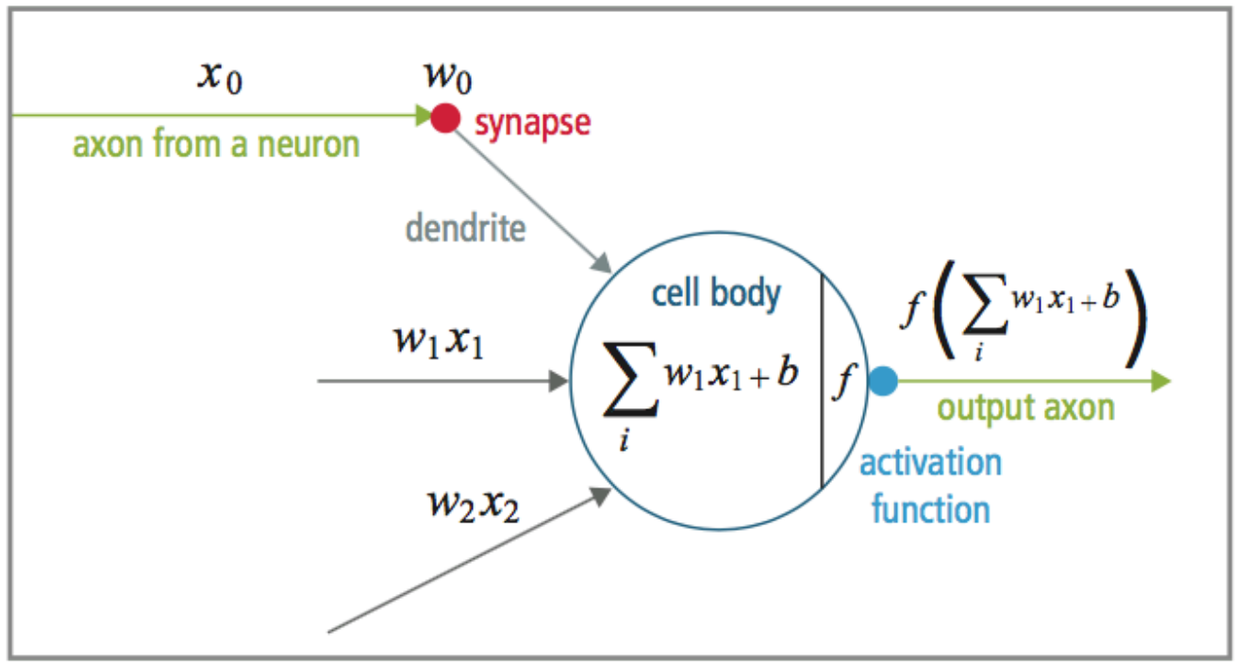
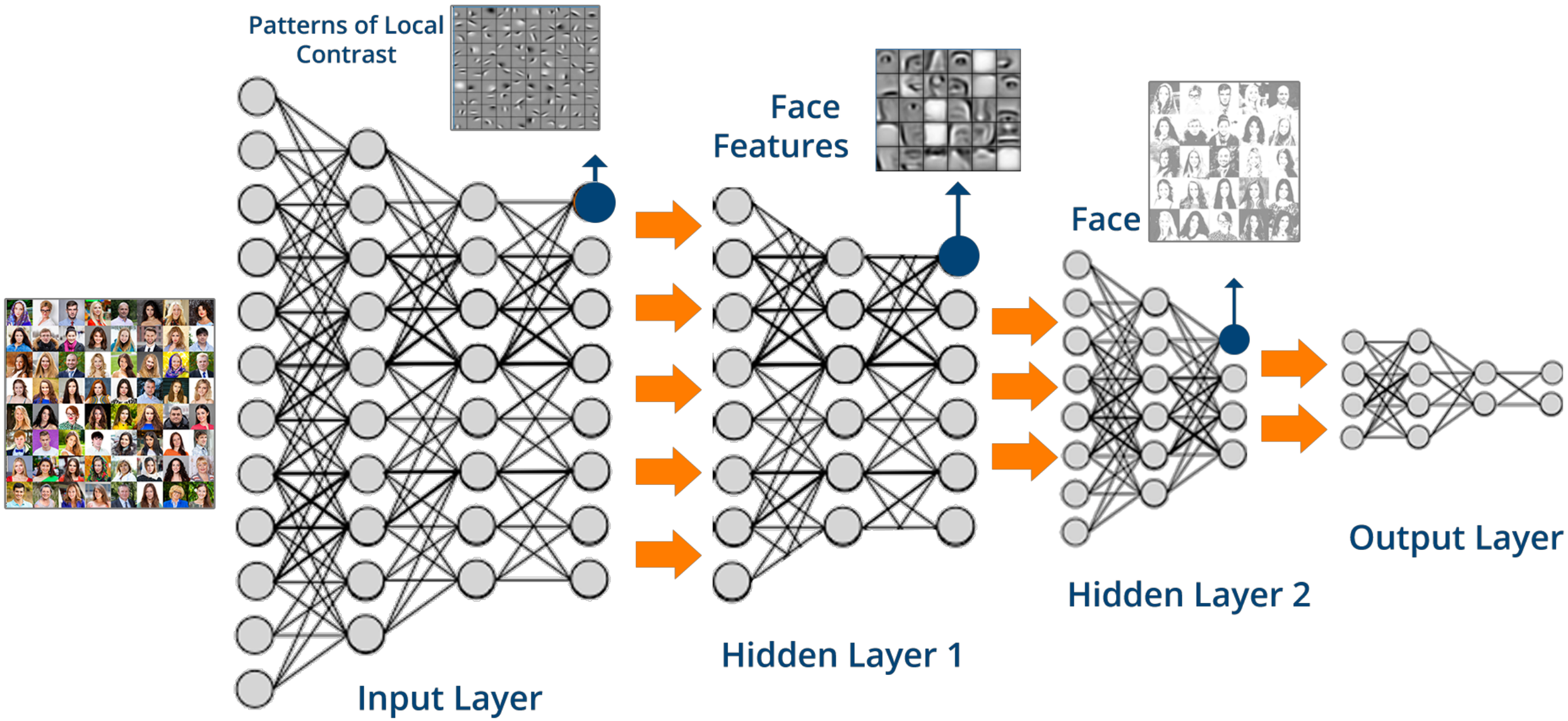


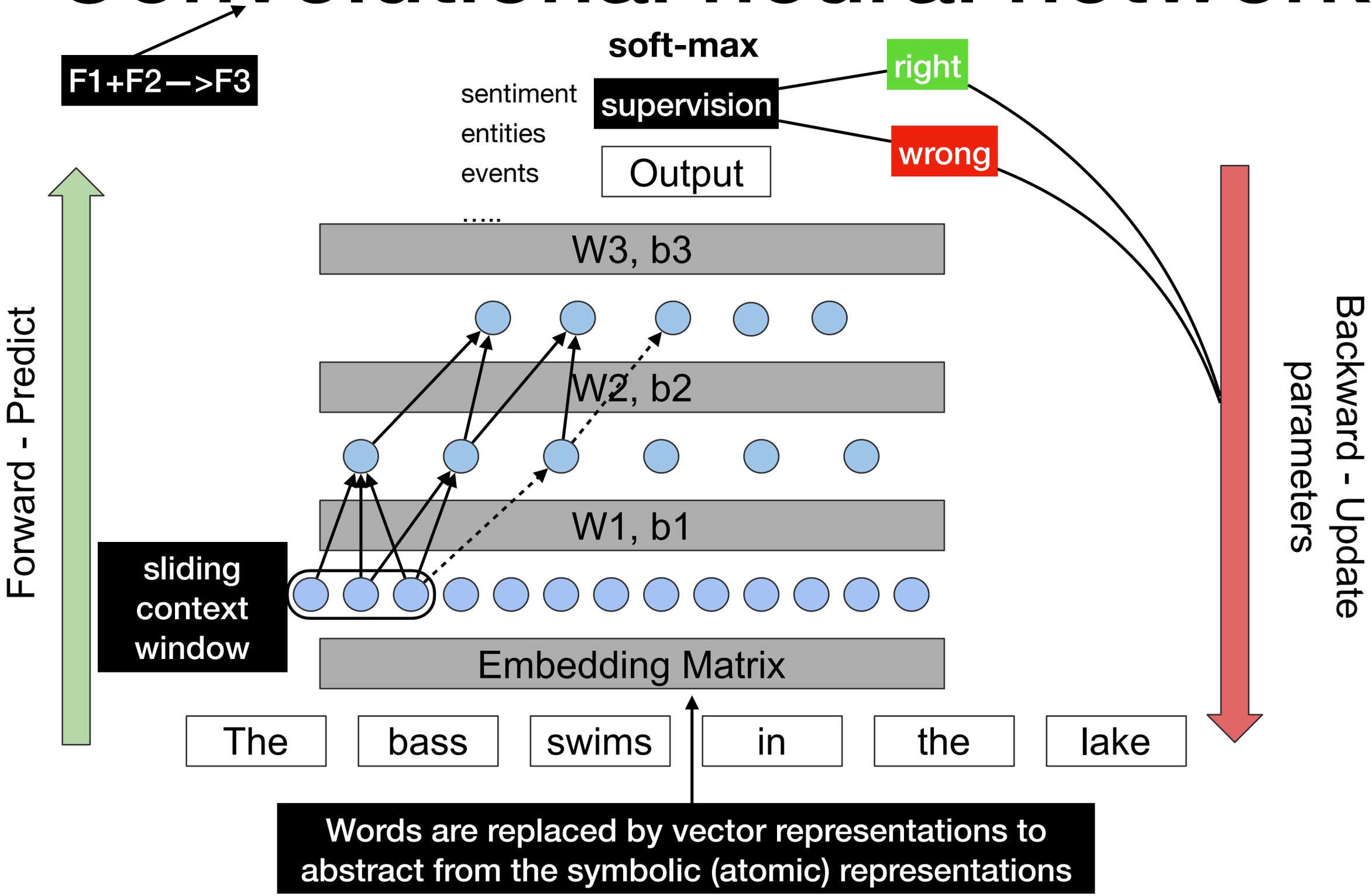
Figure 3: Illustration of a biological neuron (left) and its mathematical model (right) [2].

By: Andrej Karapthy 2015

Deep learning for vision



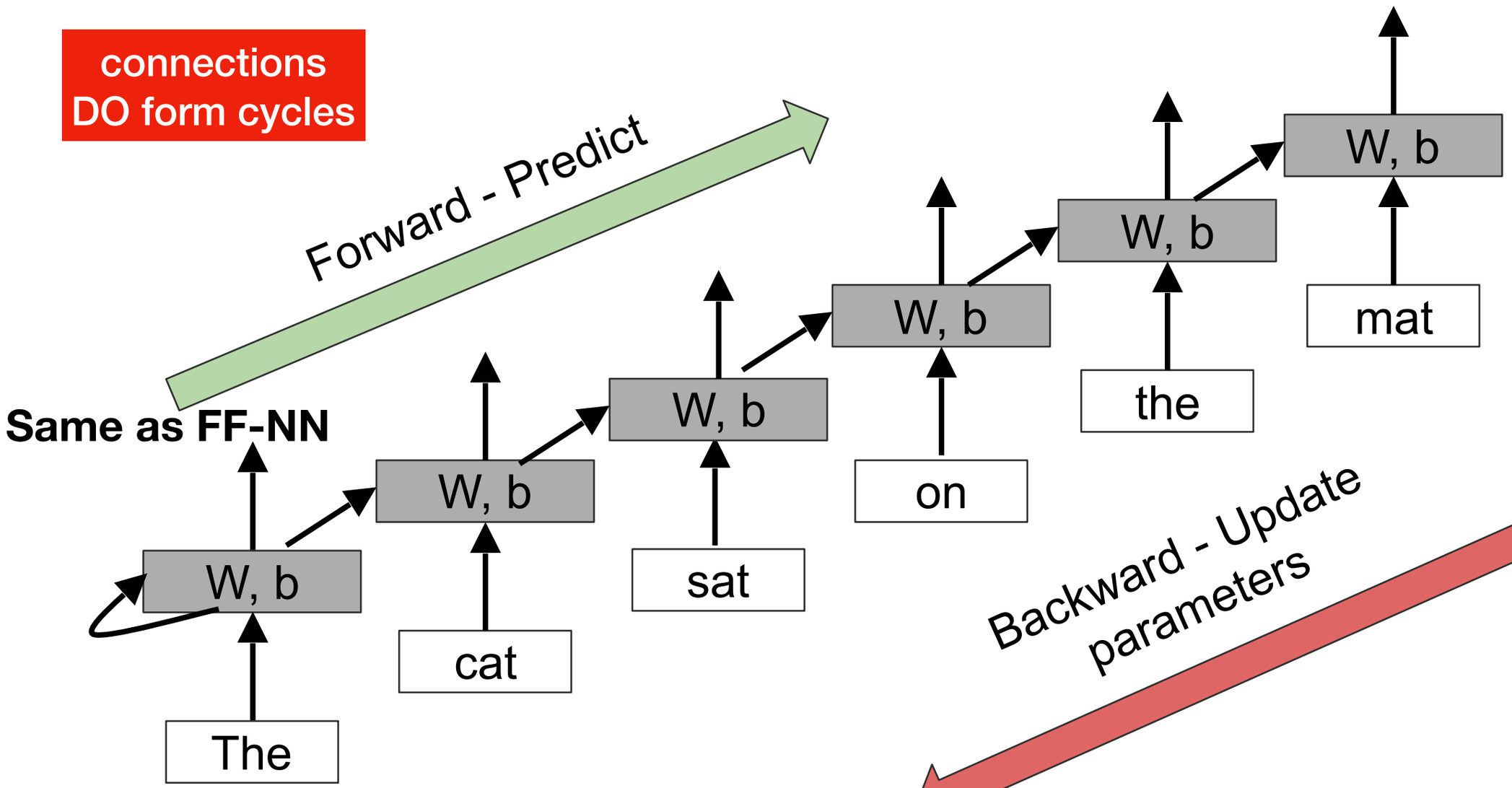
Convolutional neural network



More details

Recurrent neural network

connections
DO form cycles



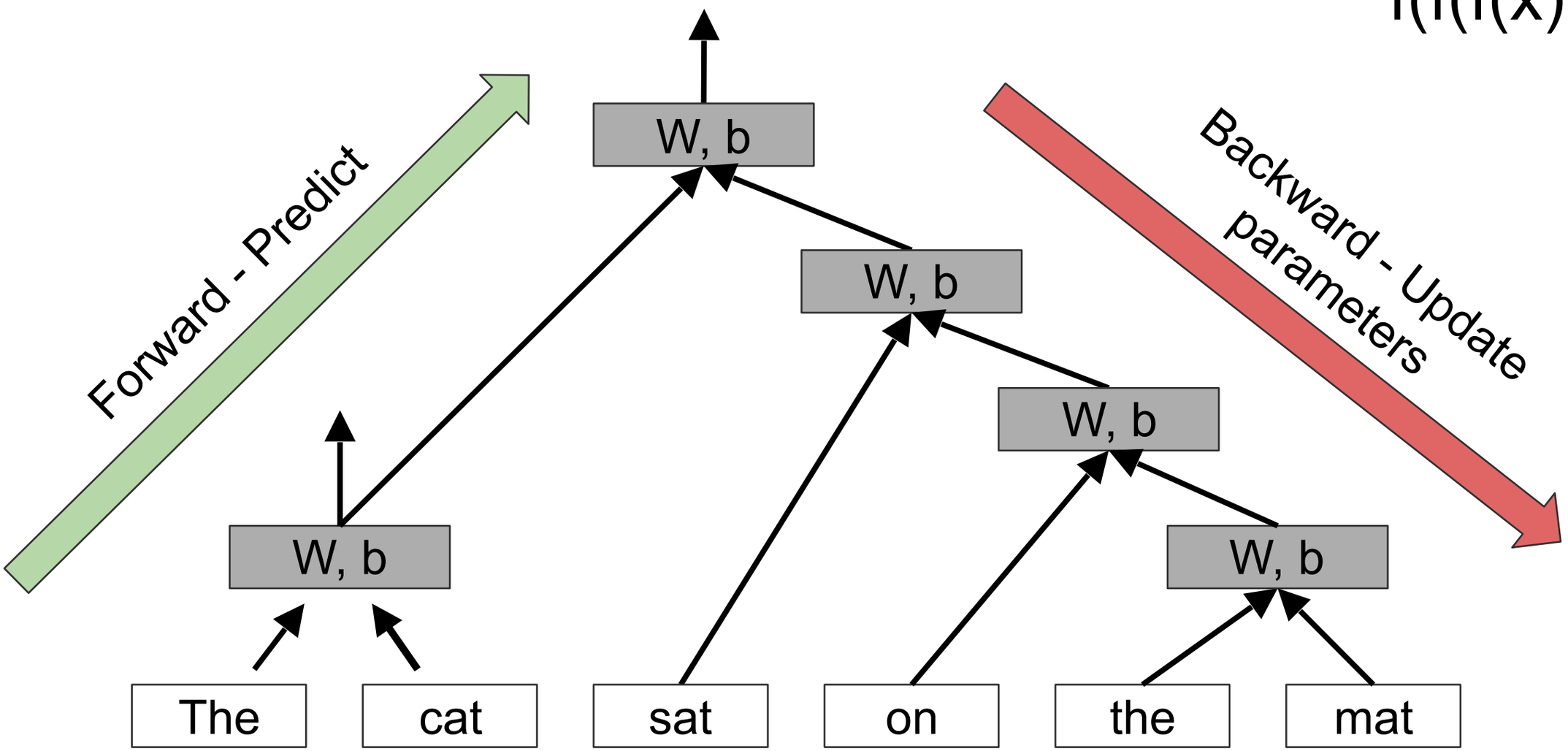
Sequential feed through text, where the previous forward (upward) result feeds into the next sequence input. Predictive value from left to right.

Phrase Classification, Machine Translation, Chat, Autocompletion

More details

Recursive neural network

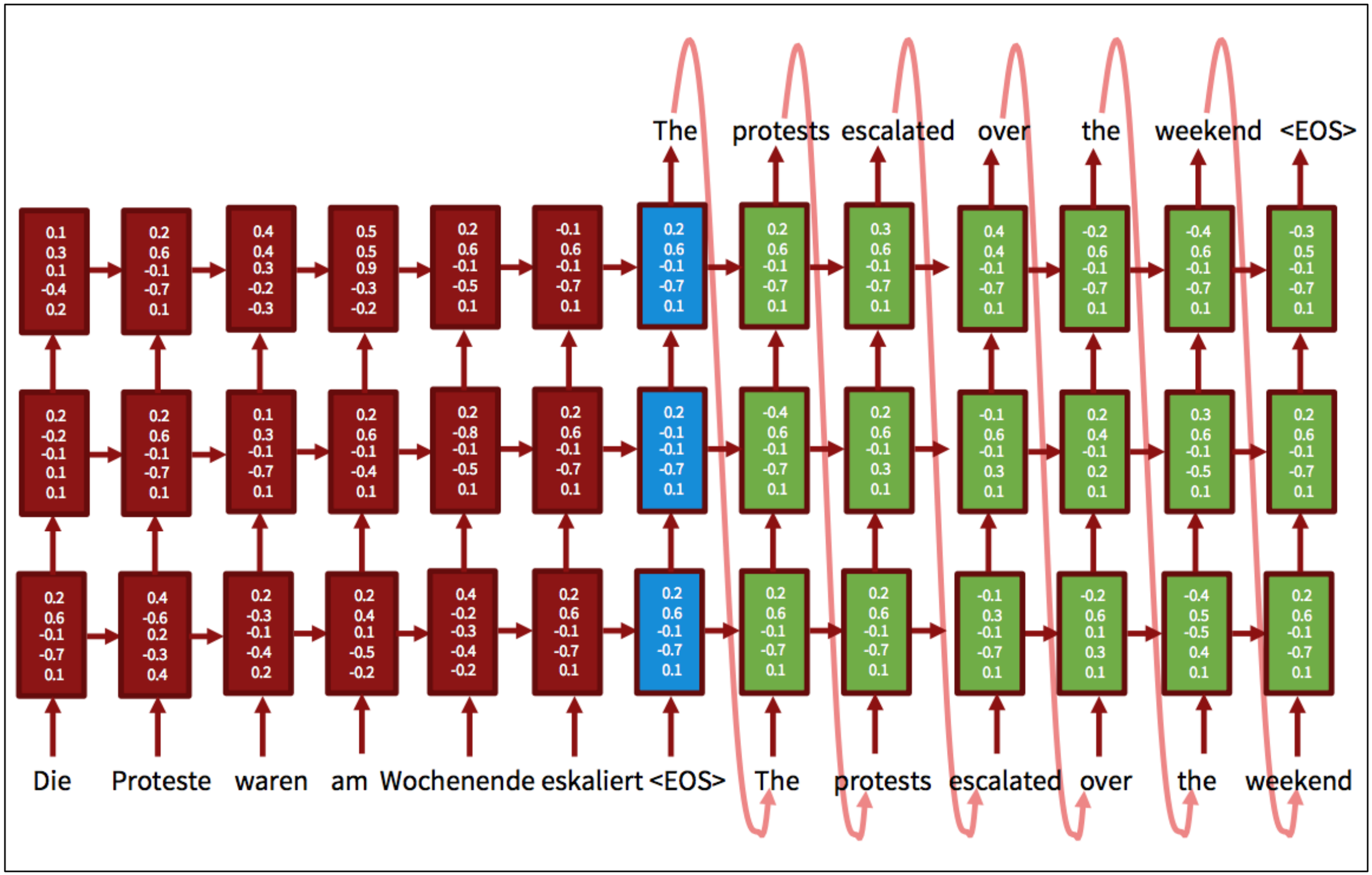
f(x), f(f(x)),
f(f(f(x))),...



Learns which n-grams are more likely and are to be combined into convolutional result but also derives a vector representation of this result for sequence prediction

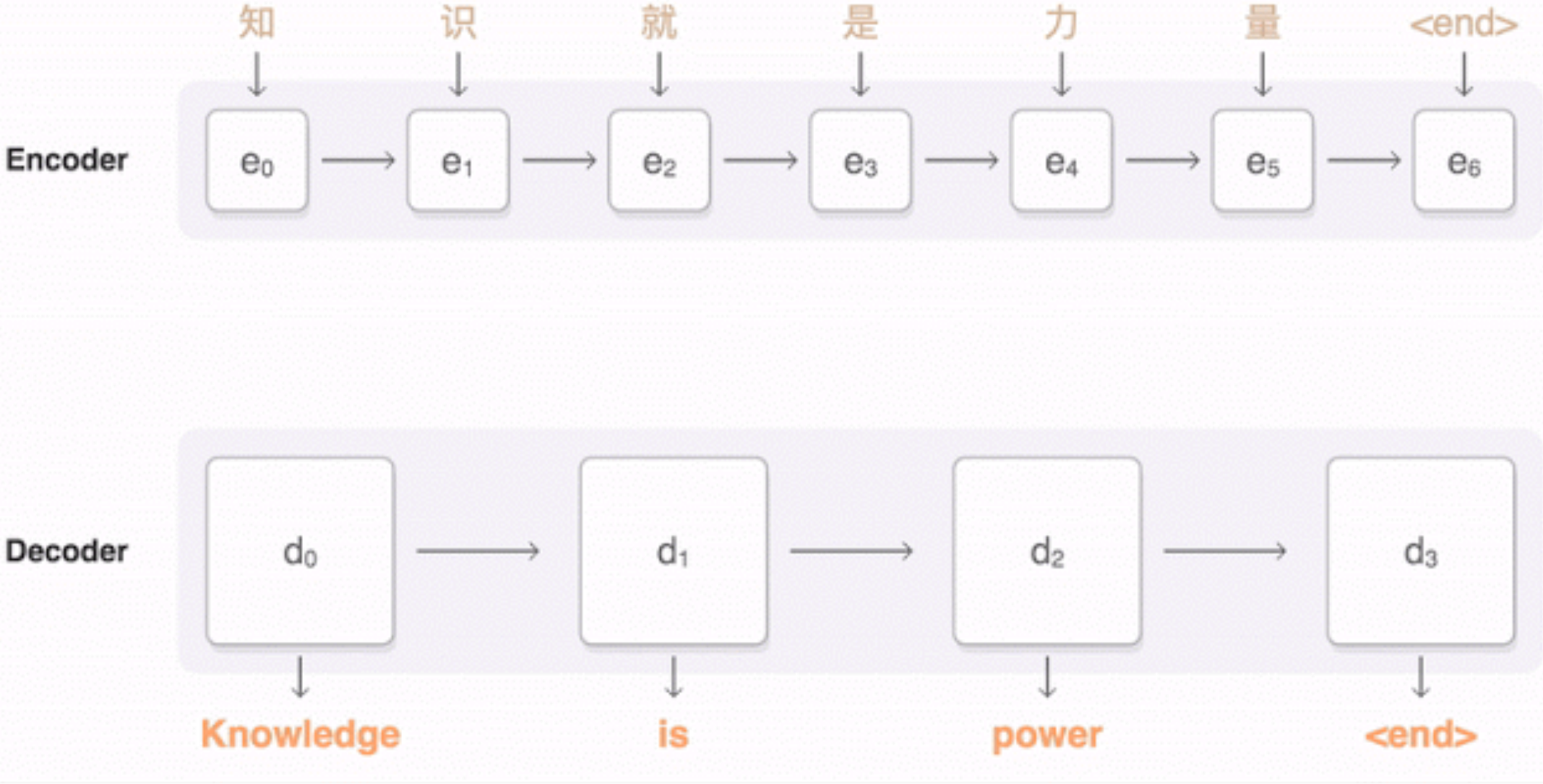
Which words group together (phrase structures) and creates a representation for the group: syntactic parsing, sentence-sentiment

Machine translation: long-short-term-memory



Google Neural MT

More details



More details

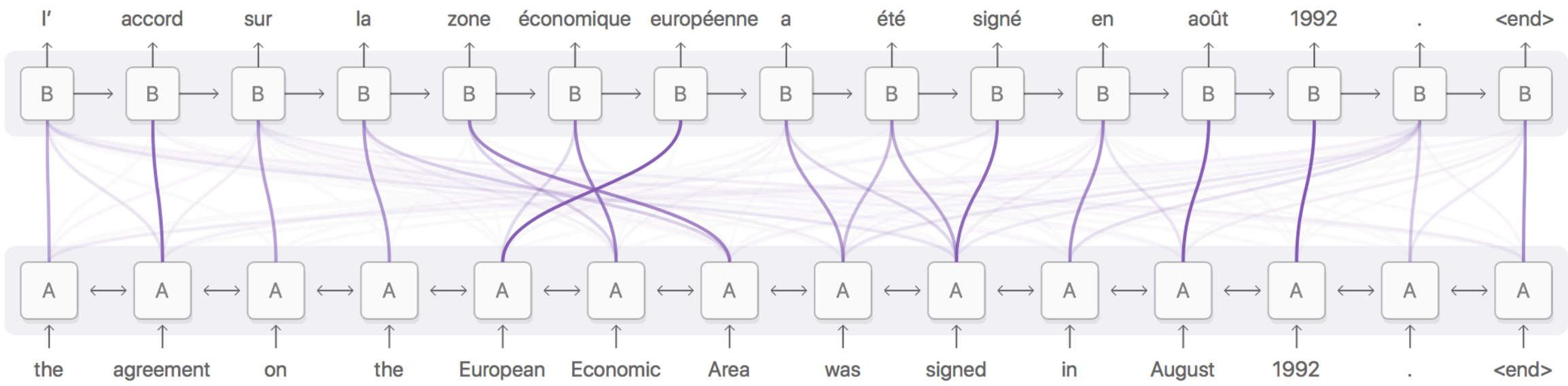


Diagram derived from Bahdanau, *et al.* 2014

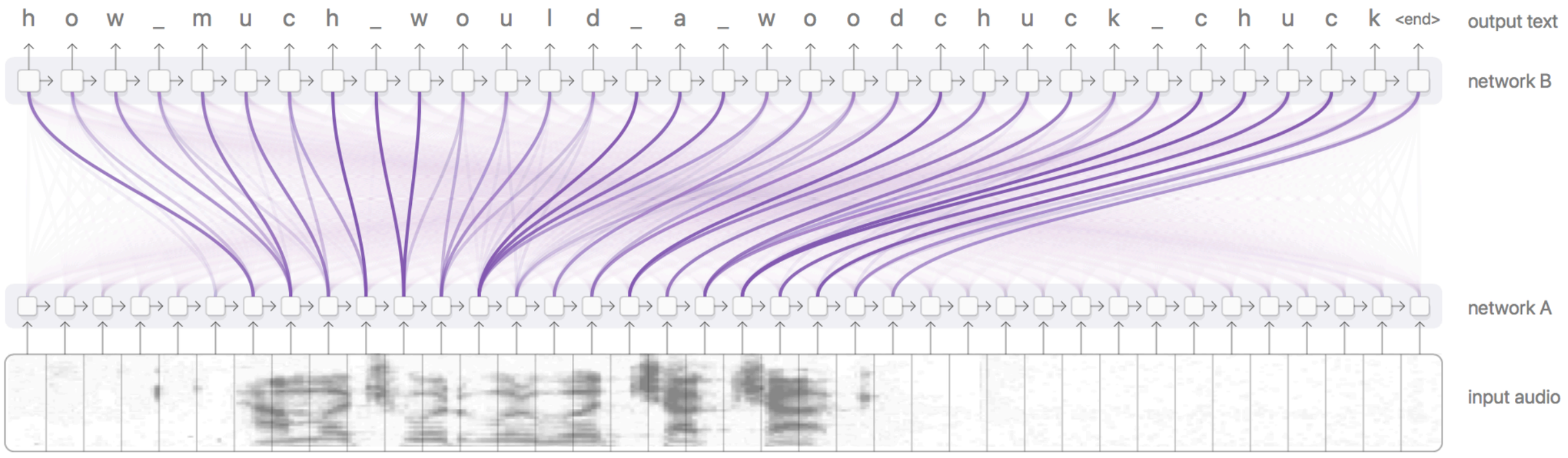
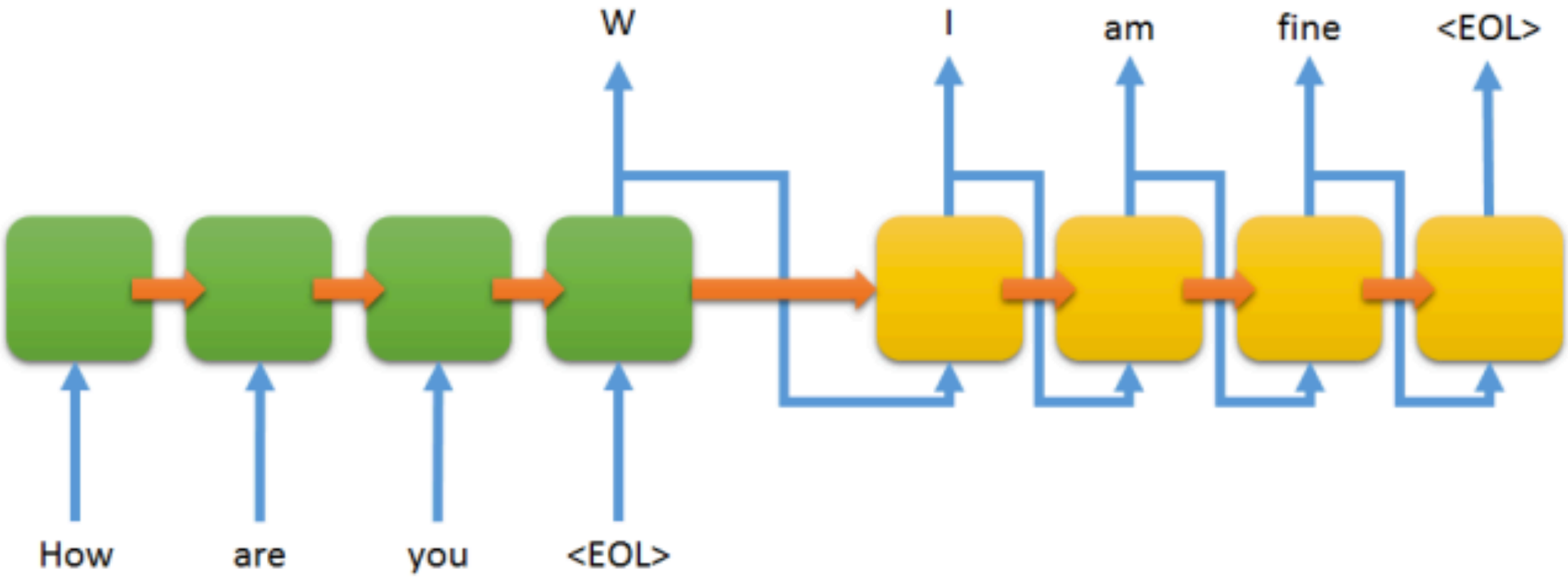


Figure derived from Chan, *et al.* 2015

Long short-term memory



Encoder is a Neural Network that captures a complete sequence

Question

Source language

…and passes this as the input to the Decoder to predict a sequence

Answer

Target language

Vinyals & Le 2015

Movie subtitles
1.3 billion words

Google's conversatiemachine

- Human: who is skywalker ?
- Machine: he is a hero .
- Human: who is bill clinton ?
- Machine: he 's a billionaire .
- Human: is sky blue or black ?
- Machine: blue .
- Human: does a cat have a tail ?
- Machine: yes .
- Human: does a cat have a wing ?
- Machine: no
- Human: can a cat fly ?
- Machine: no .
- Human: how many legs does a cat have ?
- Machine: four , i think .

- **Human: how many legs does a spider have ?**
- **Machine: three , i think .**
- **Human: how many legs does a centipede have ?**
- **Machine: eight .**
- Human: what is the color of the sky ?
- Machine: blue .
- **Human: what is the color of water ?**
- **Machine: water .**
- Human: what is the color of blood ?
- Machine: it is the same as a black eye .
- Human: how much is two plus two ?
- Machine: four .
- **Human: how much is ten minus two ?**
- **Machine: seventy - two .**

More details

Google's Neural Network Chatbot Can Discuss Philosophy and IT ...
https://gizmodo.com/googles-neural-network-chatbot-can-discuss-philosophy-a-1715...

A Neural Network Chatbot - Surprisingly Human? - I Programmer
www.i-programmer.info/.../8742-a-neural-network-chatbot-surprisingly-human.html

Google's developed a new
—and it can just about hold

Google Made a Chatbot That Debates the Meaning of Life | WIRED
https://www.wired.com/2015/06/google-made-chatbot-debates-meaning-life/

Jun 26, 2
in photos
phone ca
perhaps
You visit

Google to Developers: Here's How to Stop Making Dumb Chatbots ...
https://www.technologyreview.com/.../google-to-developers-heres-how-to-stop-makin...

May 12, 2016 - And recent progress has been made by feeding those annotations into a large deep-
learning **neu**
language is
a

Google's New Chatbot Taught Itself to Be Creepy - Motherboard
https://motherboard.vice.com/en_us/.../googles-new-chatbot-taught-itself-to-be-creepy

Google's chatbot discusses the meaning of life | Daily Mail Online
www.dailymail.co.uk/sciencetech/article.../Google-s-chatbot-discusses-meaning-life.ht...

ıg
ıey

Feb 1
with a
to lea

Chatting with a Deep learning brain – HuggingFace – Medium
https://medium.com/huggingface/chatting-with-a-deep-learning-brain-fff7a8656c4b

leural
ɔogle Made
ght Itself to

Google's chatbot learned it all from movies - Engadget
https://www.engadget.com/2015/.../googles-chatbot-learned-how-to-talk-from-movies...

Jul 2, 2
variants
of-the-r
neural

The AI Revolution: Why Deep Learning Is Suddenly Changing Your Life
fortune.com/ai-artificial-intelligence-deep-machine-learning/

Sep 28, 2016 - Machine translation and other forms of language processing have also become far more convincing, with **Goc**
neural nets is that n
above. In fact ...

Google, Facebook Develop Chatbots via Deep Neural Networks
www.etcentric.org/google-facebook-develop-chatbots-via-deep-neural-networks/

Jun 3, 2016 - Microsoft, **Google** and Facebook are all pursuing **chatbots**, which will function as virtual assistants, answering questions, responding to requests, and anticipating needs. But building functioning **chatbots**, which are based on artificial intelligence, is harder than it sounds. To further progress, **Google** ...

NLP being solved?

- Not really:
  - improvements by 5 up to 10 points but still 10% - 20% error
  - deep learning systems cannot explain why they are right or wrong —> **black box paradigm**
  - lack motivational and intentional drive
  - they are not creative, just robust
  - not clear how to combine many multiple tasks, e.g. Watson using machine learning to combine separate specialised robust modules for playing a game or a medical support system

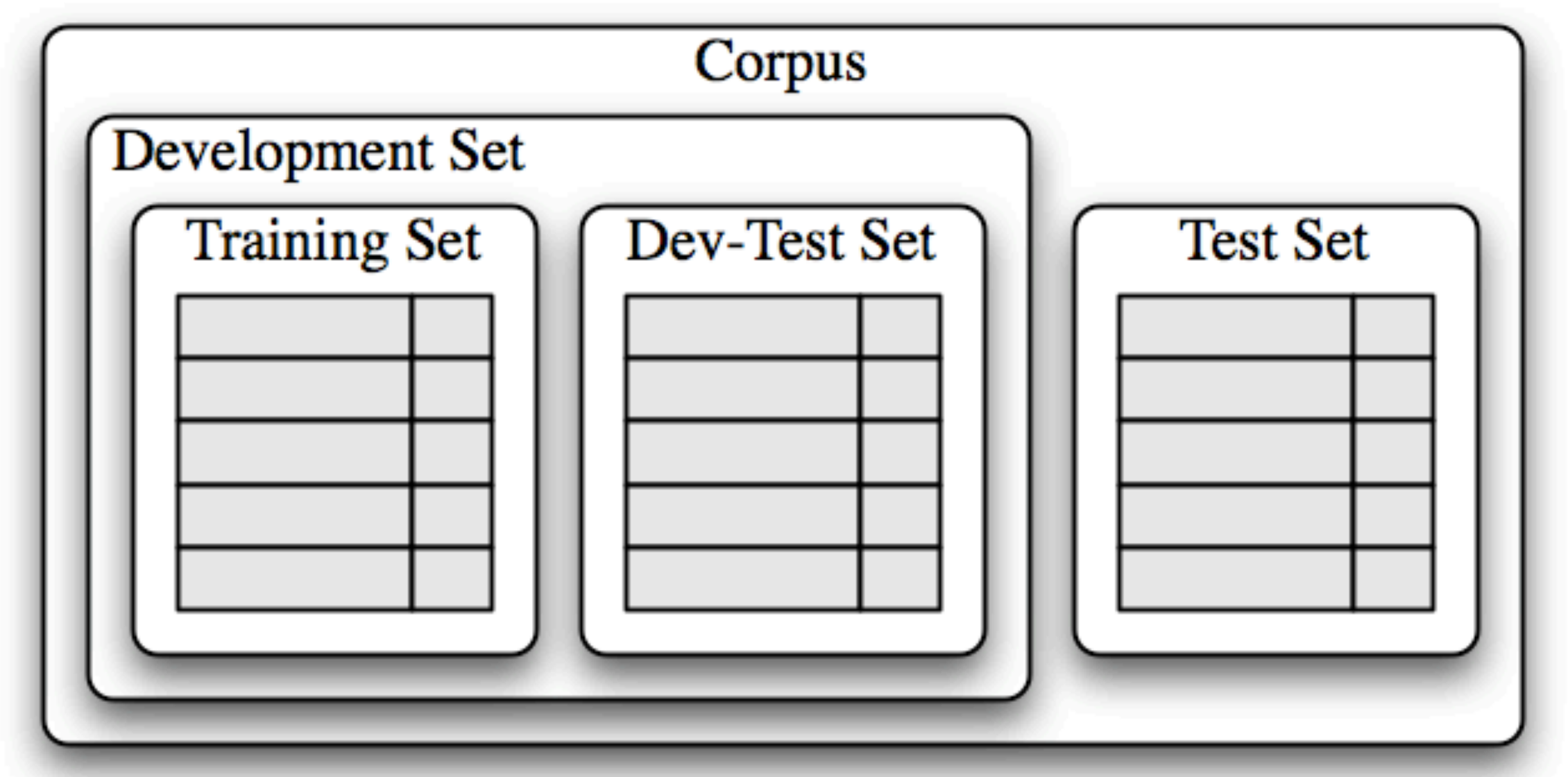
Part VI: Evaluation

- Evaluation regimes:
  - Conference of Natural Language Learning (CoNLL): http://www.conll.org/previous-editions
  - Automatic Content Extraction (ACE): https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications
  - SemEval competitions: https://en.wikipedia.org/wiki/SemEval
- Data sets:
  - training data (if machine learning is to be used)
  - development data
  - test data or (10-)fold cross-validation
- Precision, recall and F-measure

CoNLL editions

2018 - Brussels, Belgium
2017 - Vancouver, Canada
2016 - Berlin, Germany
2015 - Bejing, China
2014 - Baltimore, MD, USA
2013 - Sofia, Bulgaria
2012 - Jeju Island, Korea
2011 - Portland, OR, USA
2010 - Uppsala, Sweden
2009 - Boulder, CO, USA
2008 - Manchester, UK
2007 - Prague, Czech Republic
2006 - New York City, NY, USA
2005 - Ann Arbor, MI, USA
2004 - Boston, MA, USA
2003 - Edmonton, Canada
2002 - Taipei, Taiwan
2001 - Toulouse, France
2000 - Lisbon, Portugal
1999 - Bergen, Norway
1998 - Sidney, Australia
1997 - Madrid, Spain

Evaluation framework



Folded cross validation

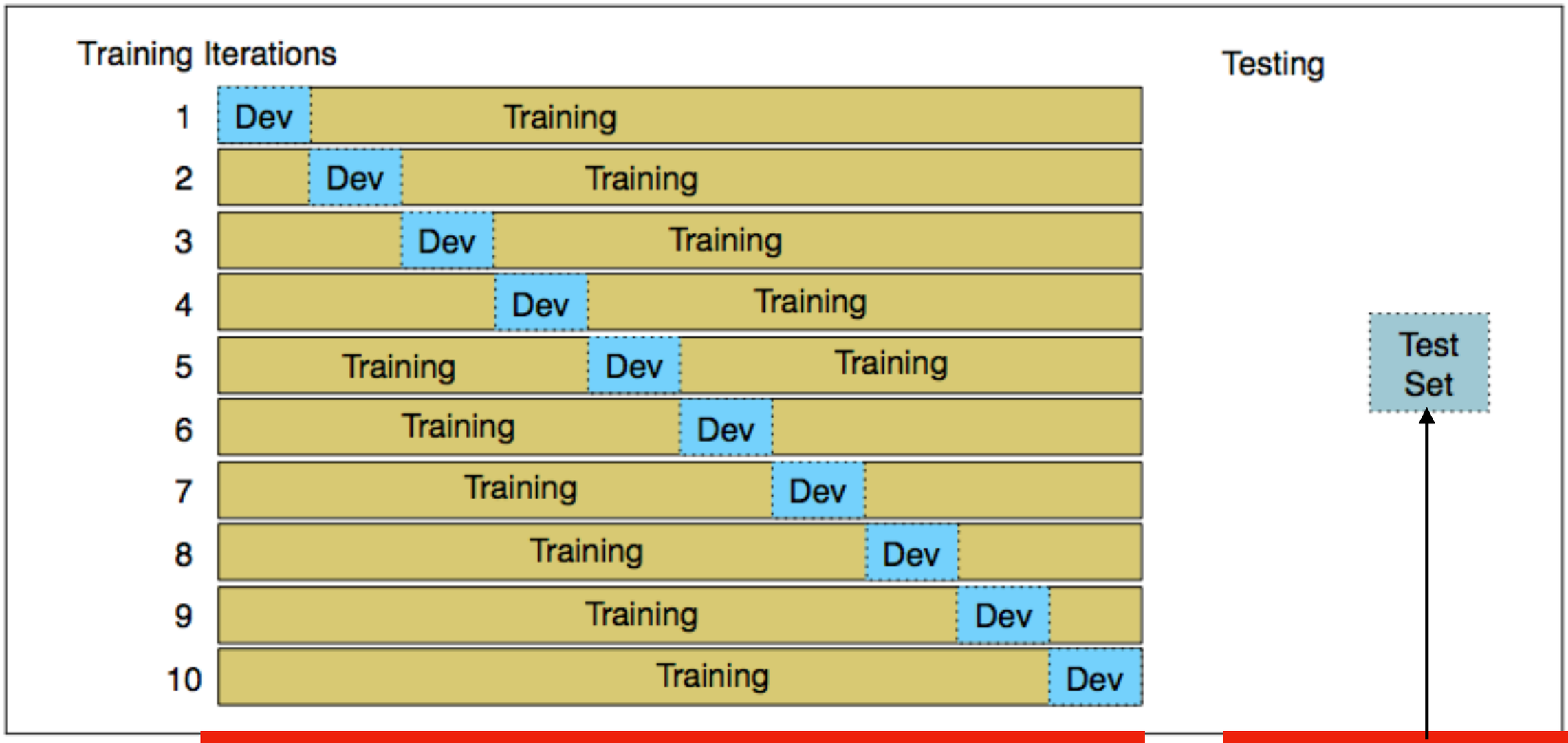


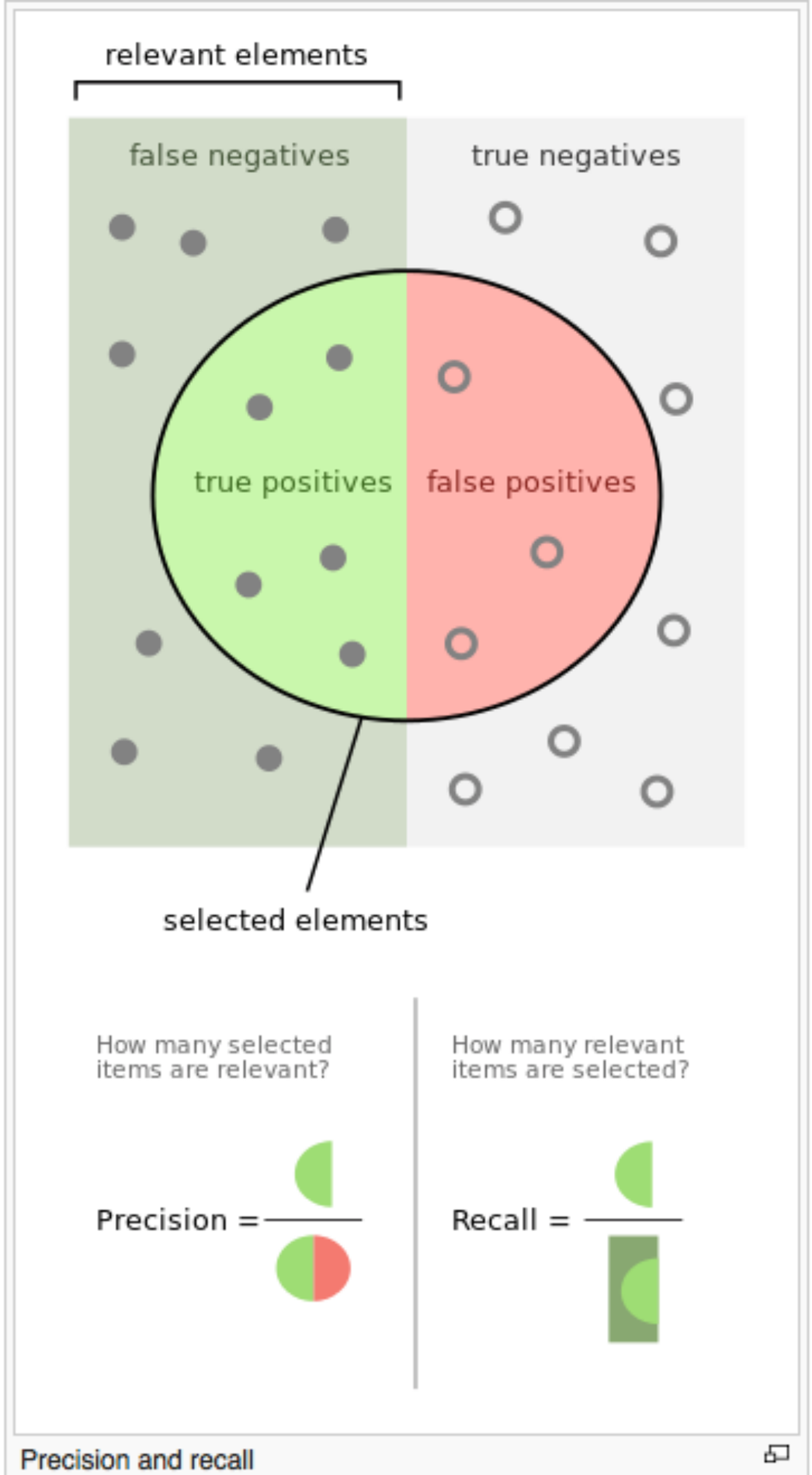
Figure 6.7 10-fold crossvalidation

Precision, Recall, F1-measure

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$



Precision and recall

Contingency table/confusion matrix

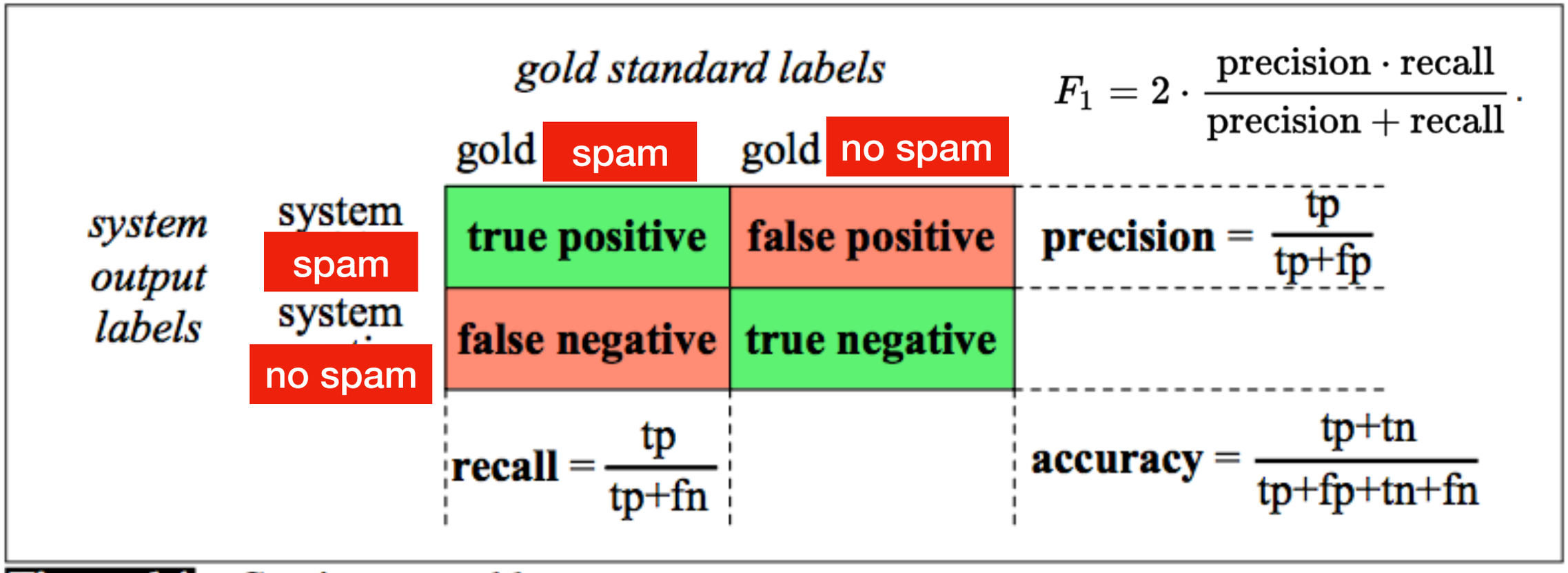
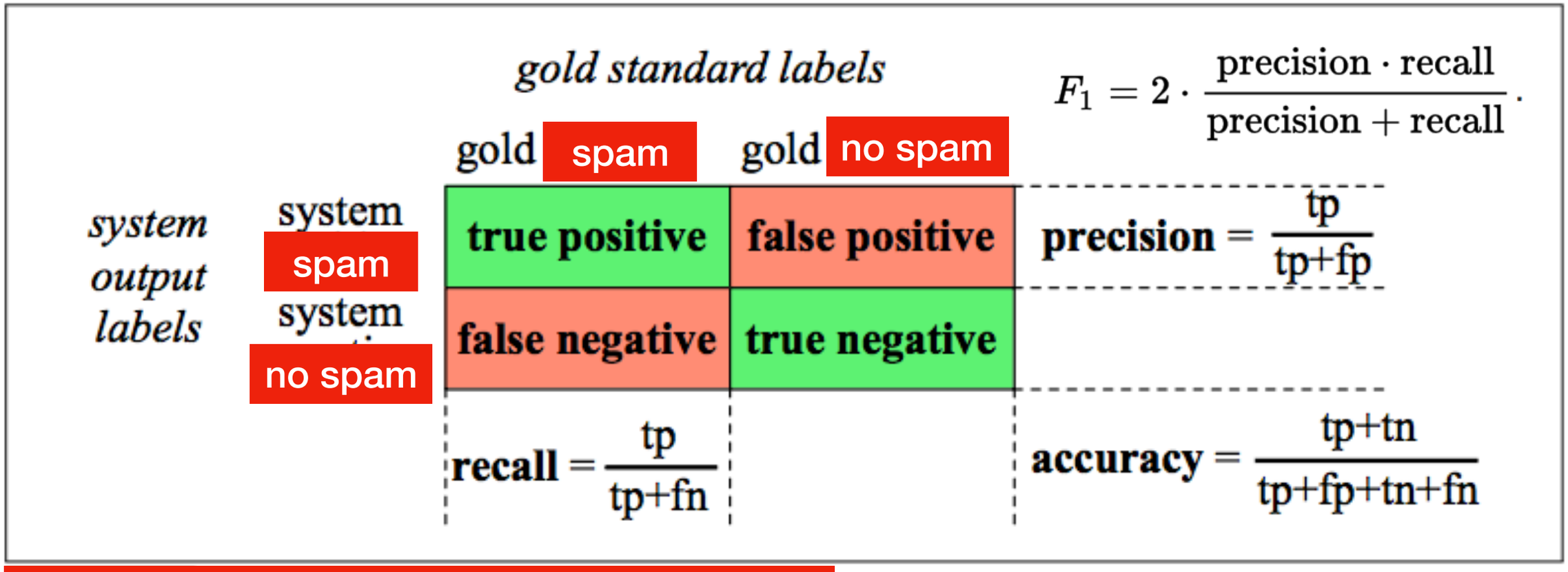


Figure 6.4 Contingency table

Binary classification

accuracy, precision, recall, harmonic mean (F1)

Contingency table/confusion matrix



Gold data: 100 mails, 20 spam mails
System: detects 10 spam mails, 5 correct

$$\frac{5\text{ TP} + 75\text{ TN}}{5\text{ TP} + 5\text{ FP} + 75\text{ TN} + 15\text{ FN}} = 0.8 \text{ accuracy}$$

Accuracy only works for balanced data!!

$$\frac{5\text{ TP}}{5\text{ TP} + 5\text{ FP}} = 0.5 \text{ precision}$$

$$\frac{5\text{ TP}}{5\text{ TP} + 15\text{ FN}} = 0.25 \text{ recall}$$

$$\frac{2 \times 0.5 \times 0.25}{0.5 + 0.25} = 0.33 \text{ F1}$$

Multiclass: spam, urgent, normal

system output	*gold labels* urgent	normal	spam	
urgent	8	10	1	$\textbf{precision}_u = \frac{8}{8+10+1}$
normal	5	60	50	$\textbf{precision}_n = \frac{60}{5+60+50}$
spam	3	30	200	$\textbf{precision}_s = \frac{200}{3+30+200}$
	$\textbf{recall}_u = \frac{8}{8+5+3}$	$\textbf{recall}_n = \frac{60}{10+60+30}$	$\textbf{recall}_s = \frac{200}{1+50+200}$	

Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

- Build separate binary classifiers for each class using positive cases for *c* and all cases for *not-c (!c)*
  - **any-of**: run all classifiers and allow multiple results
  - **one-of** or multinomial: run all classifiers and take the best

Based on: Jurafsky & Martin: https://web.stanford.edu/~jurafsky/slp3/

Micro & Macro averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

- macro: if performance is balanced across classes
- micro: if nr. of cases is balanced across classes

Based on: Jurafsky & Martin: https://web.stanford.edu/~jurafsky/slp3/

Contingency table for sentiment

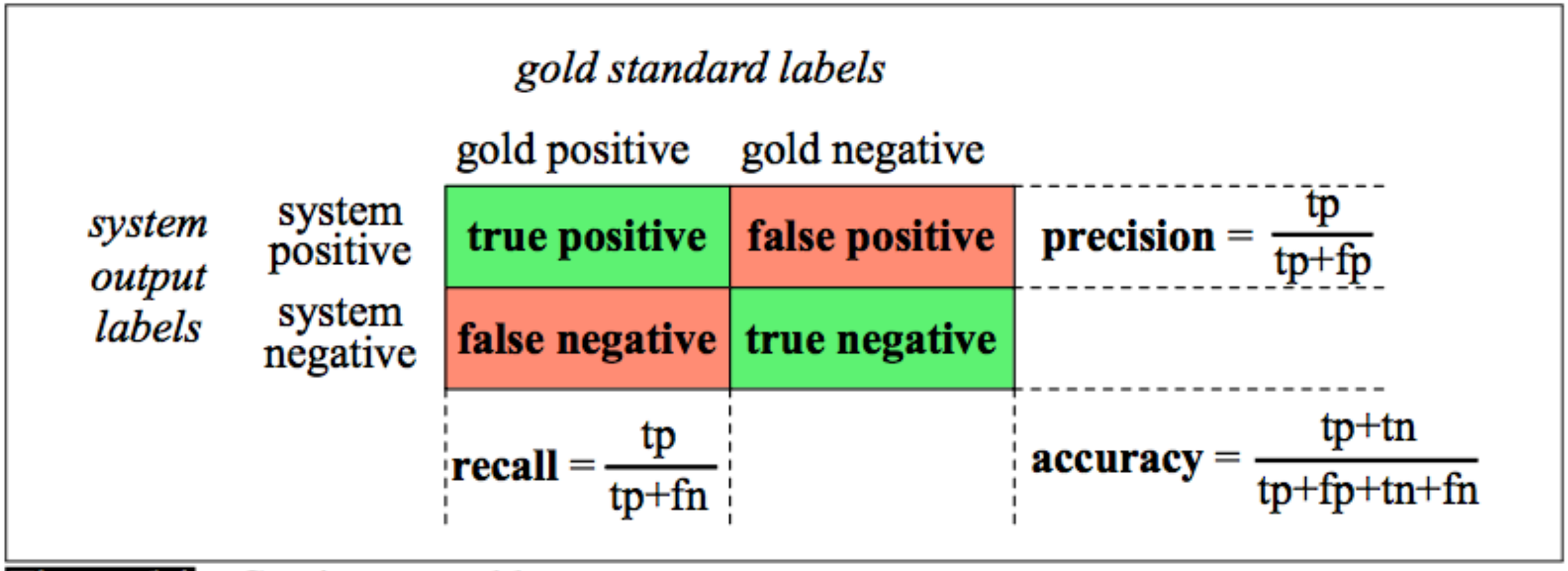


Figure 6.4 Contingency table

Gold data: 100 reviews, 40 positive, 30 negative

Contingency table for sentiment

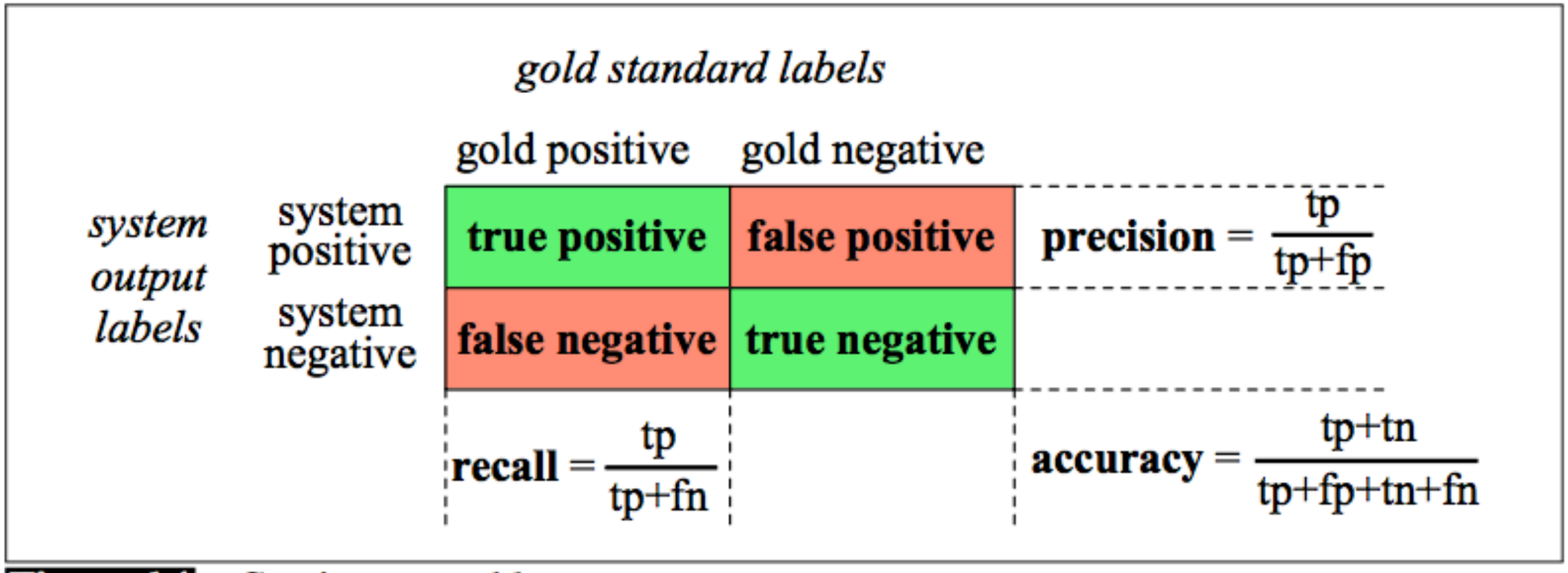


Figure 6.4 Contingency table

Gold data: 100 reviews, 40 positive, 30 negative

positive: 35 TP + 10 FP

$$\frac{35\text{ TP}}{35\text{ TP} + 10\text{ FP}} = 0.78 \text{ precision}$$

$$\frac{35\text{ TP}}{35\text{ TP} + 5\text{ FN}} = 0.87 \text{ recall}$$

negative: 20TP + 5FP

$$\frac{20\text{ TP}}{20\text{ TP} + 5\text{ FP}} = 0.8 \text{ precision}$$

$$\frac{20\text{ TP}}{20\text{ TP} + 10\text{ FN}} = 0.66 \text{ recall}$$

What matters precision or recall depends on the application

- High precision, low recall:
  - ***Fully automated system*** that takes decisions (self-driving car)
  - Find precisely what you are looking for and ignore the rest: one perfect example is enough to act upon
  - Reduce human work
- High recall, low precision:
  - **No data will slip through**, but a human checks the decision
  - Find everything on a topic and filter afterwards
  - Spot all potential risk, an alert (spam filters, buienradar (shower radar))
  - Requires human work
- Common strategy: first maximise the recall and next improve the precision

Types of fitting, types of machine learning

More details

- Probabilistic Classifier = assign a the most probable class *c* to a text *t*

  - **Generative classifiers**: e.g. Naive Bayes build a model for each class and return the most likely one for an observation

  - **Discriminative classifiers**: e.g. logical regression or maximum entropy learn what features from the input are most useful to discriminate between possible classes

  - Others: Suppert Vector Machines (SVM), random forests, Conditional Random Fields (CRF)

linear classifiers
(linear combination of the input data)

More details

Naive Bayesian Classifier

- Naive because
  - independence assumption
  - simple probabilistic guessing machine

category c

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) \qquad (6.1)$$

document d

More details

Logical regression (Max Entropy)

- No independence assumption
- Predict class from features instead of features from class
- To what extent does a feature predict the class or not the class

$$p(c|x) = \frac{\exp\left(\sum_{i=1}^{N} w_i f_i(c,x)\right)}{\sum_{c' \in C} \exp\left(\sum_{i=1}^{N} w_i f_i(c',x)\right)}$$

w = weight
f = feature
c = class
x = observation (e.g. text)
i ranges over all features for a class c
exp(x)=power of x o exponent to get values between 0 and 1

More details

Logical regression

- occurrence of words as a feature for classes + and -

Possible weights

$$f_1(c,x) = \begin{cases} 1 & \text{if } \text{“great”} \in x \ \& \ \ c = + \\ 0 & \text{otherwise} \end{cases}$$

+0.6

$$f_2(c,x) = \begin{cases} 1 & \text{if } \text{“second-rate”} \in x \ \& \ \ c = - \\ 0 & \text{otherwise} \end{cases}$$

-0.5

$$f_3(c,x) = \begin{cases} 1 & \text{if } \text{“no”} \in x \ \& \ \ c = - \\ 0 & \text{otherwise} \end{cases}$$

-0.7

$$f_4(c,x) = \begin{cases} 1 & \text{if } \text{“enjoy”} \in x \ \& \ \ c = - \\ 0 & \text{otherwise} \end{cases}$$

+0.7

Logistic regression

More details

Var	Definition	Value in Fig. 5.2
x_1	count(positive lexicon) $\in$ doc)	3
x_2	count(negative lexicon) $\in$ doc)	2
x_3	$\begin{cases} 1 & \text{if "no"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	1
x_4	count(1st and 2nd pronouns $\in$ doc)	3
x_5	$\begin{cases} 1 & \text{if "!"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	0
x_6	log(word count of doc)	$\ln(64) = 4.15$

feature weights = [2.5, -5.0, -1.2, 0.5, 2.0, 0.7],
bias term = 0.1

$$\begin{aligned} p(+|x) = P(Y=1|x) &= \sigma(w\cdot x+b) \\ &= \sigma([2.5,-5.0,-1.2,0.5,2.0,0.7]\cdot[3,2,1,3,0,4.15]+0.1) \\ &= \sigma(1.805) \\ &= 0.86 \\ p(-|x) = P(Y=0|x) &= 1-\sigma(w\cdot x+b) \\ &= 0.14 \end{aligned}$$

$$\begin{aligned} P(y=1) &= \sigma(w\cdot x+b) \\ &= \frac{1}{1+e^{-(w\cdot x+b)}} \\ P(y=0) &= 1-\sigma(w\cdot x+b) \\ &= 1-\frac{1}{1+e^{-(w\cdot x+b)}} \\ &\quad e^{-(w\cdot x+b)} \end{aligned}$$

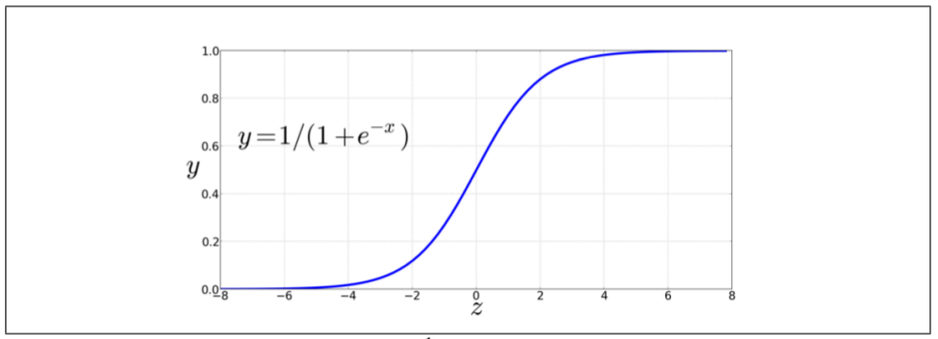


Figure 5.1 The sigmoid function $y = \frac{1}{1+e^{-z}}$ takes a real value and maps it to the range $[0,1]$. Because it is nearly linear around 0 but has a sharp slope toward the ends, it tends to squash outlier values toward 0 or 1.

Logistic function (sigmoid)

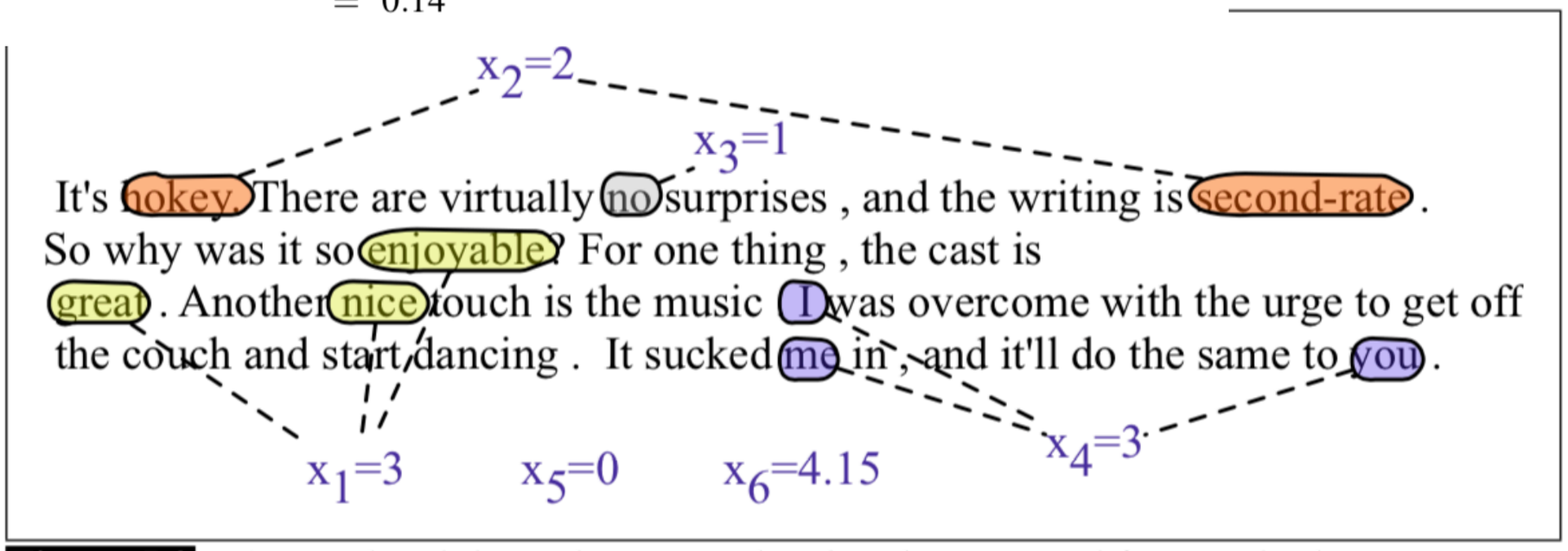


Figure 5.2 A sample mini test document showing the extracted features in the vector x.

More details

How to choose a system?

- If your training set is **small** or text is **short**, high bias/ low variance classifiers (e.g., Naive Bayes) have an advantage over low bias/high variance classifiers (e.g., logistic regression. SVM), since the latter will overfit.

- But low bias/high variance classifiers start to win out as your training set grows (they have lower asymptotic error), since high bias classifiers aren't powerful enough to provide accurate models.

More details

Advantages of Naive Bayes

- Super simple, you're just doing a bunch of counts.

- If the NB conditional independence assumption holds, a Naive Bayes classifier converges quicker than discriminative models like logistic regression, so you need less training data.

- Often still does a great job in practice even if the NB assumption doesn't hold,

- If want something fast and easy that performs pretty well.

- **Main disadvantage:** cannot learn interactions between features (e.g., it can't learn that although you love movies with Brad Pitt and Tom Cruise, you hate movies in which they're together).

More details

Advantages of Logistic Regression

- Many ways to regularize the model, and no need to worry about features being correlated, like in Naive Bayes.

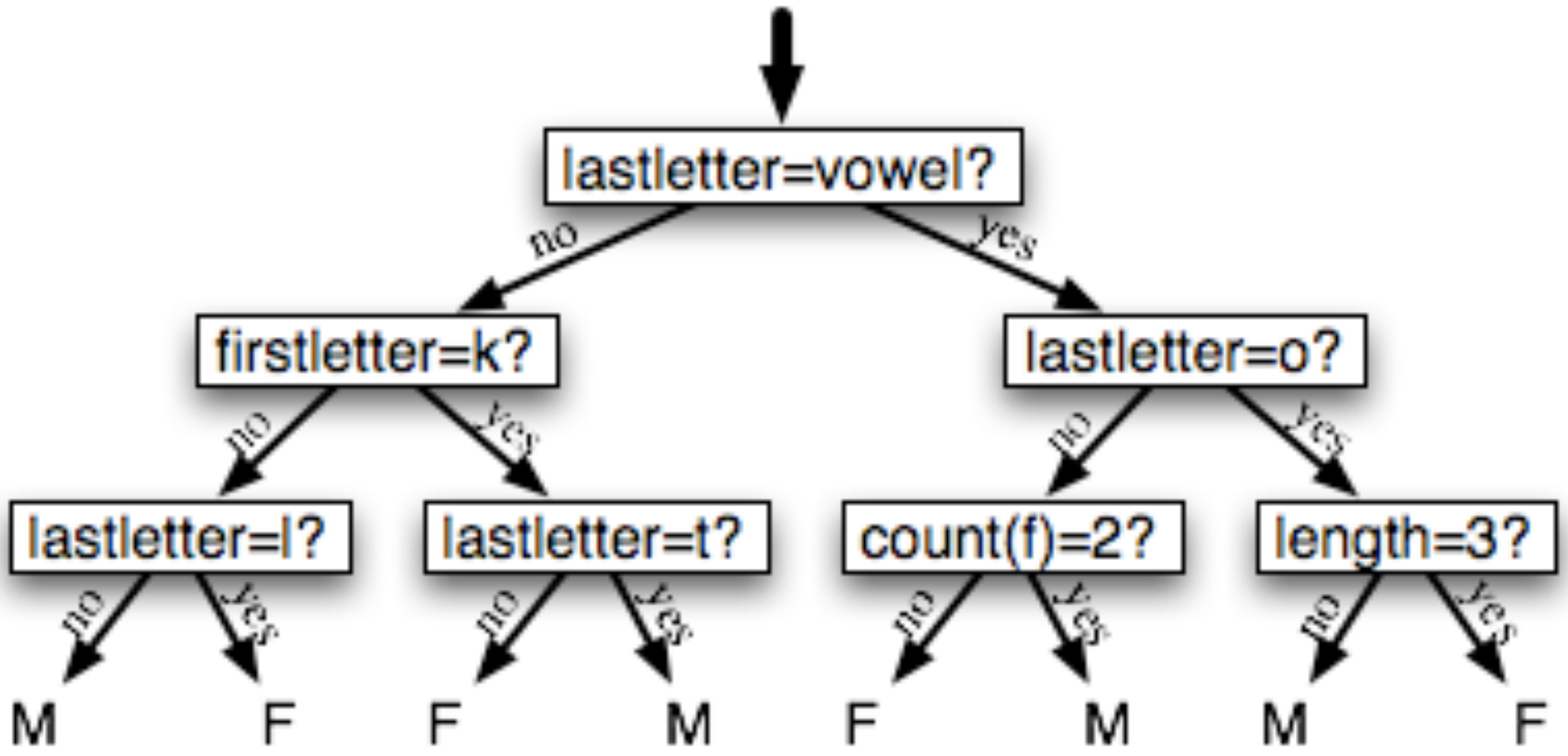
- Nice probabilistic interpretation, unlike decision trees or SVMs, and you can easily update the model to take in new data (using an online gradient descent method), again unlike decision trees or SVMs.

- Use it if you want a probabilistic framework (e.g., to easily adjust classification thresholds, to say when you're unsure, or to get confidence intervals) or if you expect to receive more training data in the future that you want to be able to quickly incorporate into your model.

More details

Decision Tree
name gender task

https://www.nltk.org/book/ch06.html



More details

Advantages of Decision Trees

- Easy to interpret and explain.

- Handle feature interactions and are non-parametric, don't need to worry about outliers or whether the data is linearly separable (e.g., decision trees easily take care of cases where you have class A at the low end of some feature x, class B in the mid-range of feature x, and A again at the high end).

- **Disadvantages:**

  - do not support online learning, so you have to rebuild your tree when new examples come on.

  - easily overfit, but that's where ensemble methods like random forests come in.

More details

Advantages of SVMs

- High accuracy, nice theoretical guarantees regarding overfitting, and with an appropriate kernel they can work well even if you're data isn't linearly separable in the base feature space.

- Especially popular in text classification problems with very high-dimensional spaces.

- **Disadvantage:** Memory-intensive, hard to interpret, and kind of annoying to run and tune.

More details

Word & Part of Speech sequences
Part of speech tagging using Hidden Markov Models (HMM)

- http://www.cse.unsw.edu.au/~billw/cs9414/notes/nlp/ambiguity/ambiguity-2009.html

Large collection of text annotated with Part-of-Speech tags

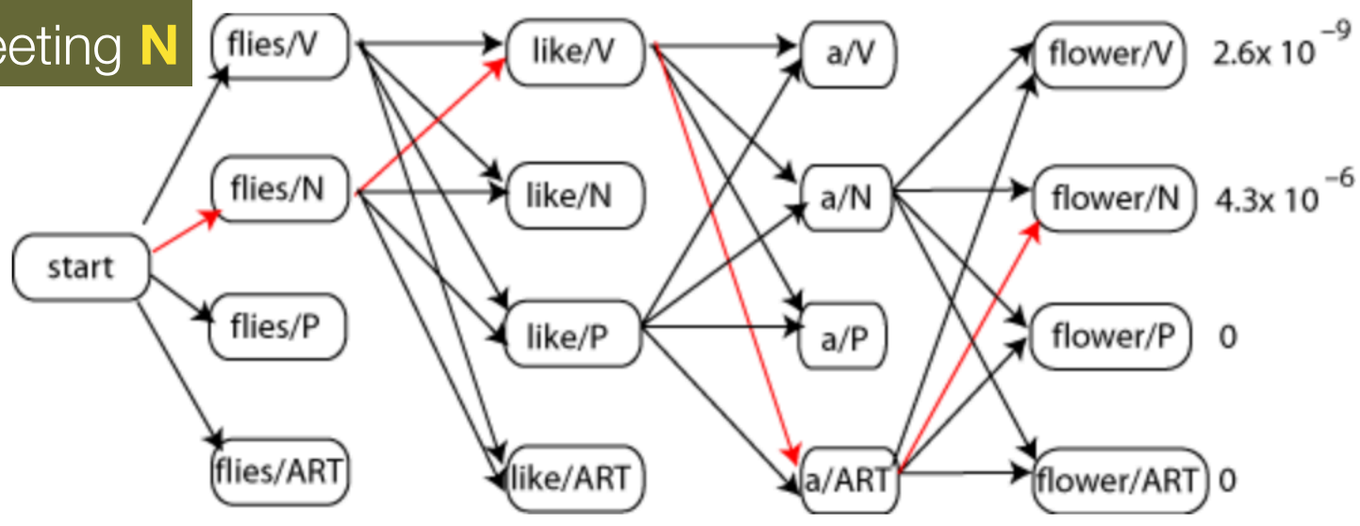
I **Pr** do **V** not **Av** like **V** flies **N** on **P** my **PR** food **N**
She **Pr** flies **V** to **P** China **N** after **P** a **ART** meeting **N**

Table 7.6 The lexical generation probabilities

Pr(the | ART)	0.54		Pr(a | ART)	0.360
Pr(flies | N)	0.025		Pr(a | N)	0.001
Pr(flies | V)	0.076		Pr(flower | N)	0.063
Pr(like | V)	0.1		Pr(flower | V)	0.05
Pr(like | P)	0.068		Pr(birds | N)	0.076
Pr(like | N)	0.012			

Viterbi algorithm

flies like a flower



lexcat	SeqScore (lexcat,1)	SeqScore (lexcat,2)	SeqScore (lexcat,3)	SeqScore (lexcat,4)	BackPtr (lexcat,4)
V	7.6×10^{-6}	0.00031	0	2.6×10^{-9}	ART
N	0.00725	1.3×10^{-5}	1.2×10^{-7}	4.3×10^{-6}	ART
P	0	0.00022	0	0	∅
ART	0	0	7.2×10^{-5}	0	∅