Natural Language Processing

Lecture 01

Dirk Hovy

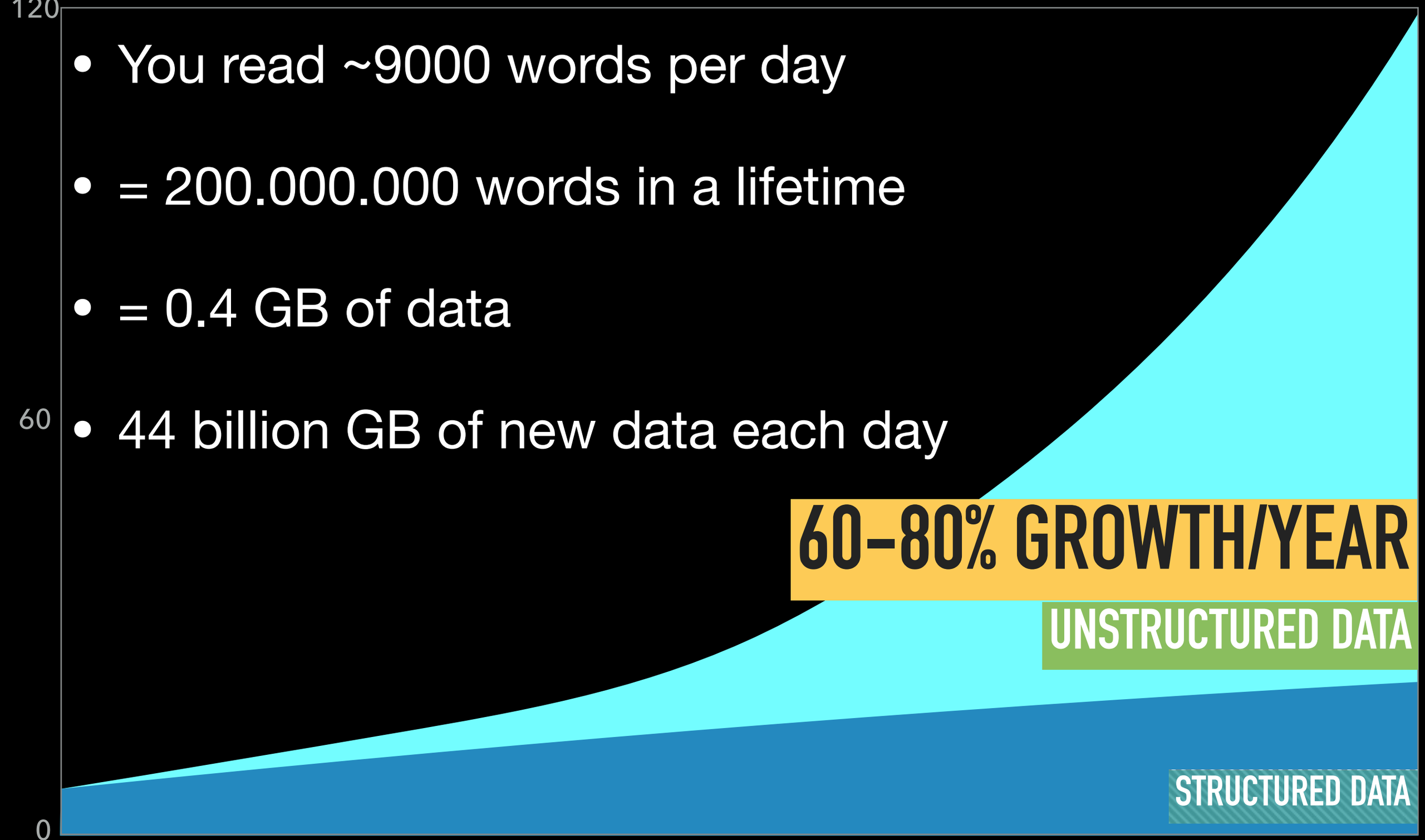
dirk.hovy@unibocconi.it

@dirk_hovy

Bocconi

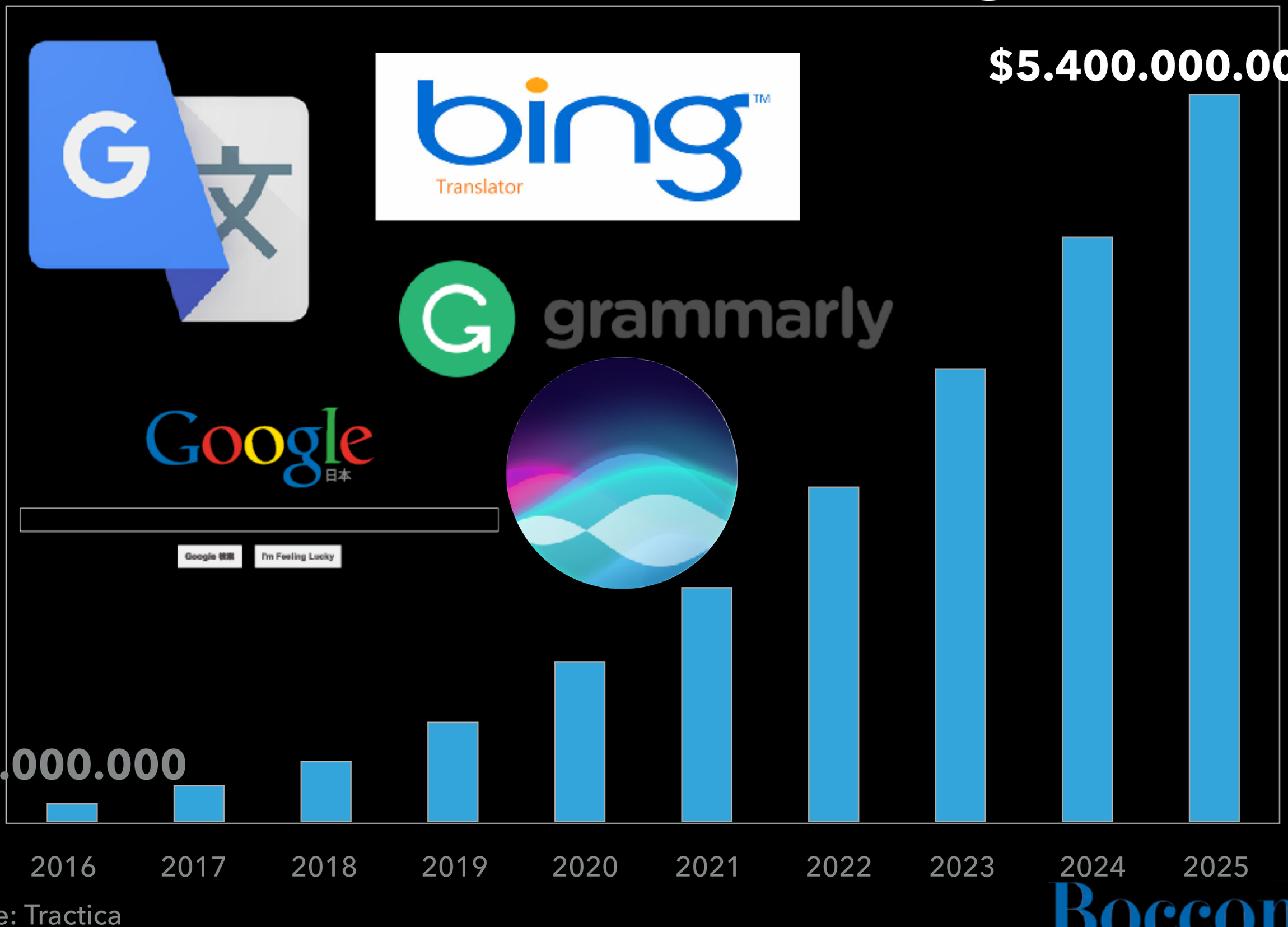
Text is an exploding data source

- You read ~9000 words per day
- = 200.000.000 words in a lifetime
- = 0.4 GB of data
- 44 billion GB of new data each day



Bocconi

NLP is booming



Source: Tractica

Bocconi

Machine Translation

MilaNLP

HELL
YES

Auf
jeden
Fall

Bocconi

Text Generation

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

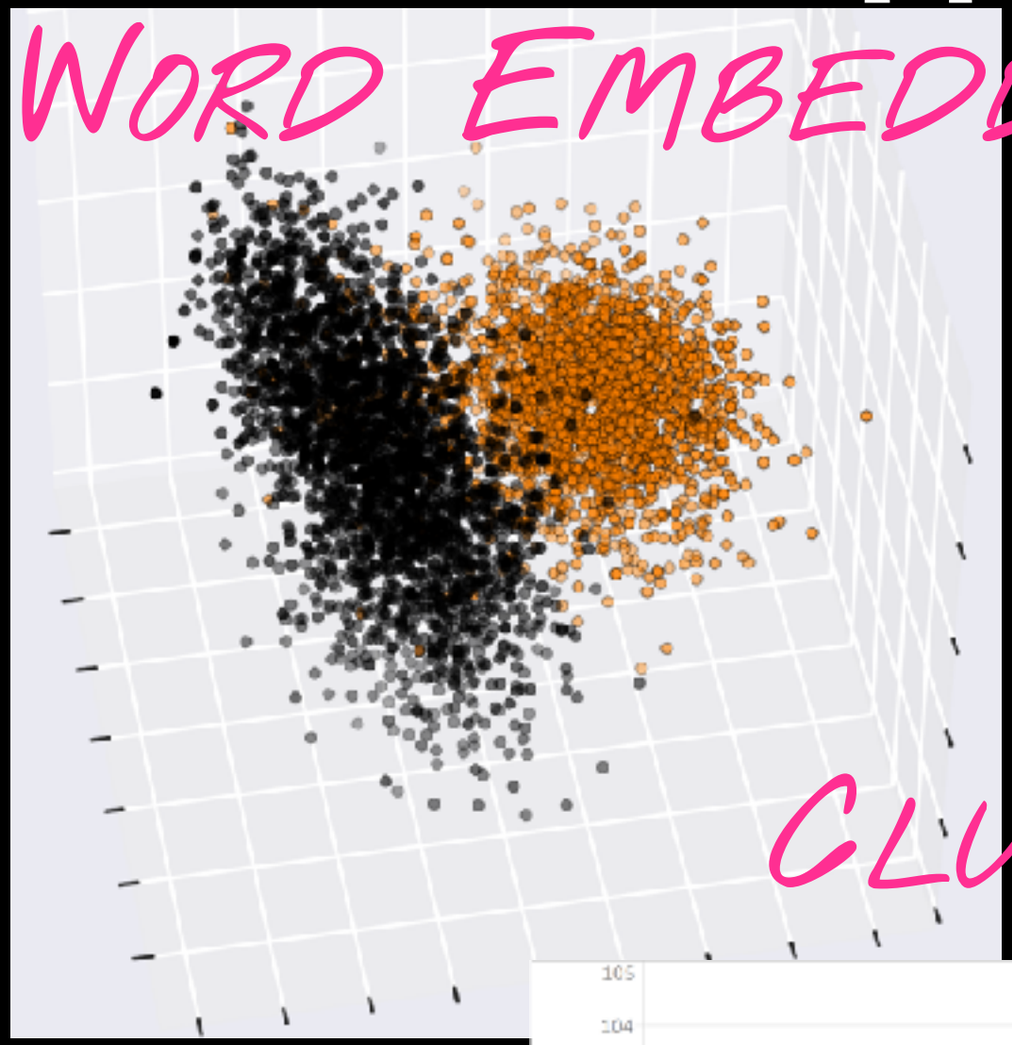
Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

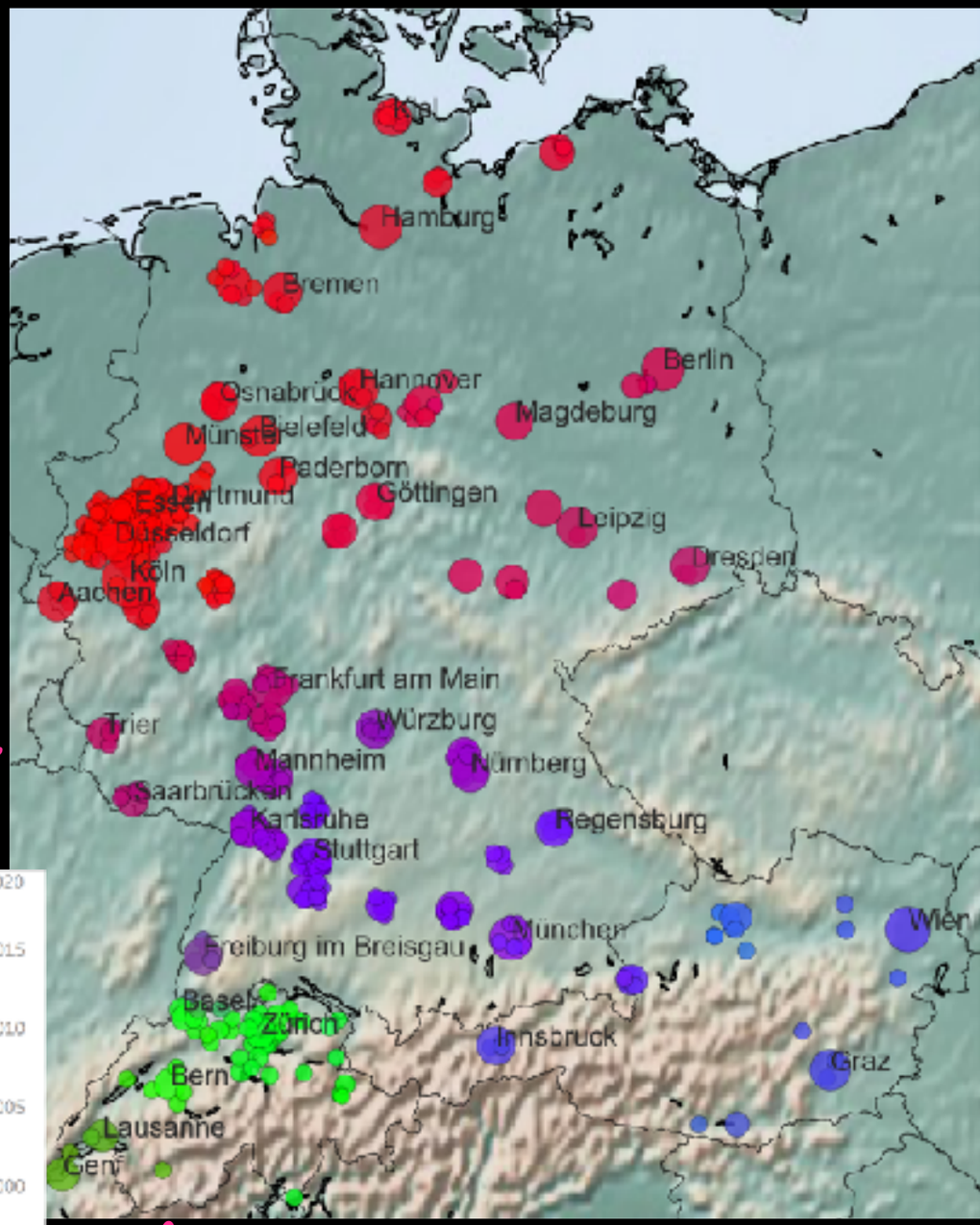
Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

The Goals

Word Embeddings



Clustering



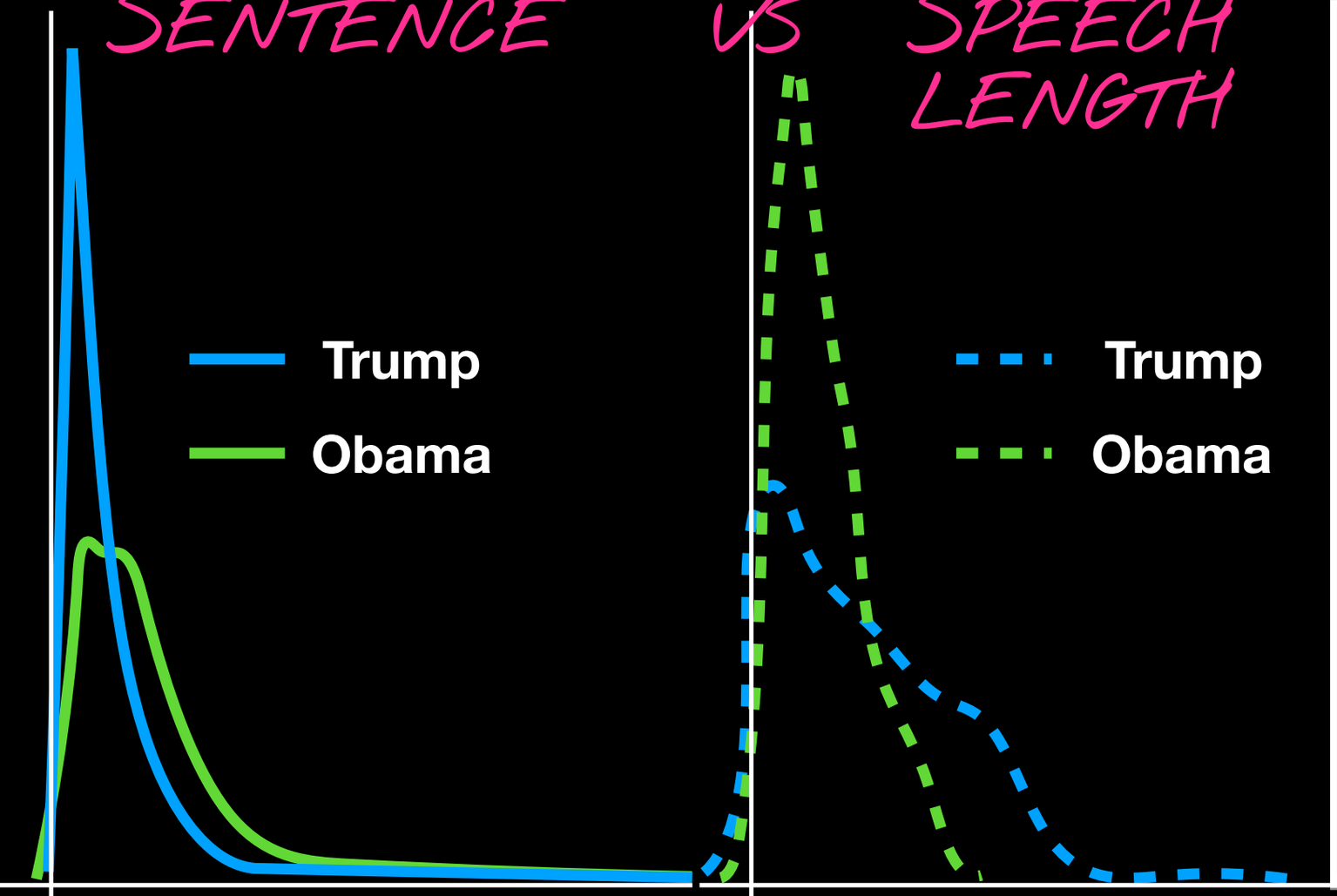
Visualization

Sentiment Analysis

Bocconi

Student Projects

Sentence vs Speech Length



Talk Similarity

Candidate LMs finish a sentence starting with "America"

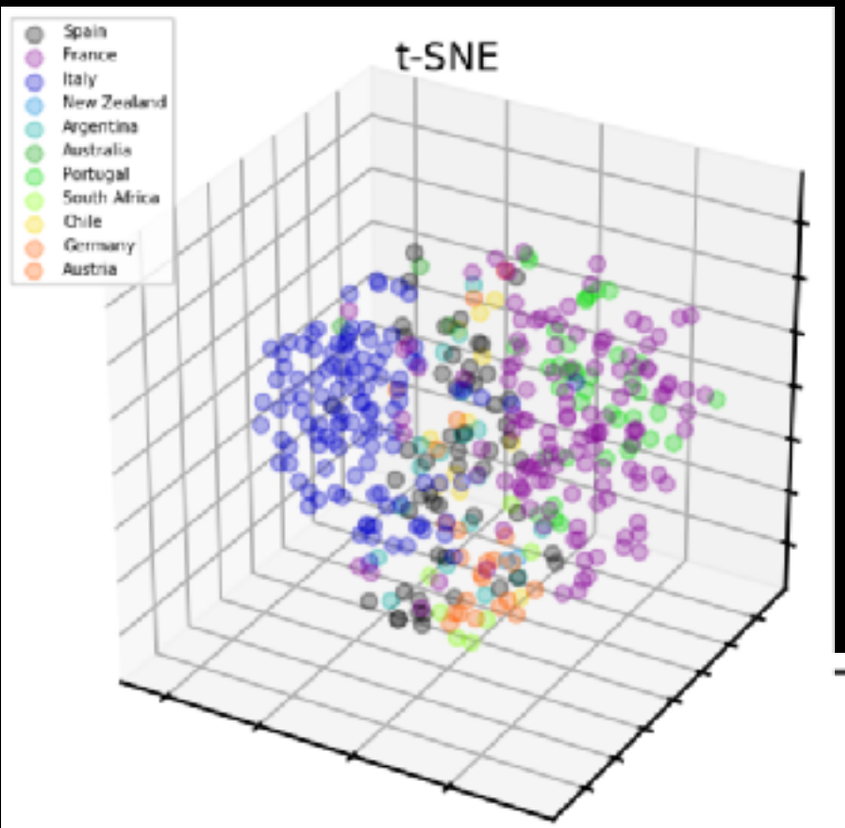
"America first — America first ..."

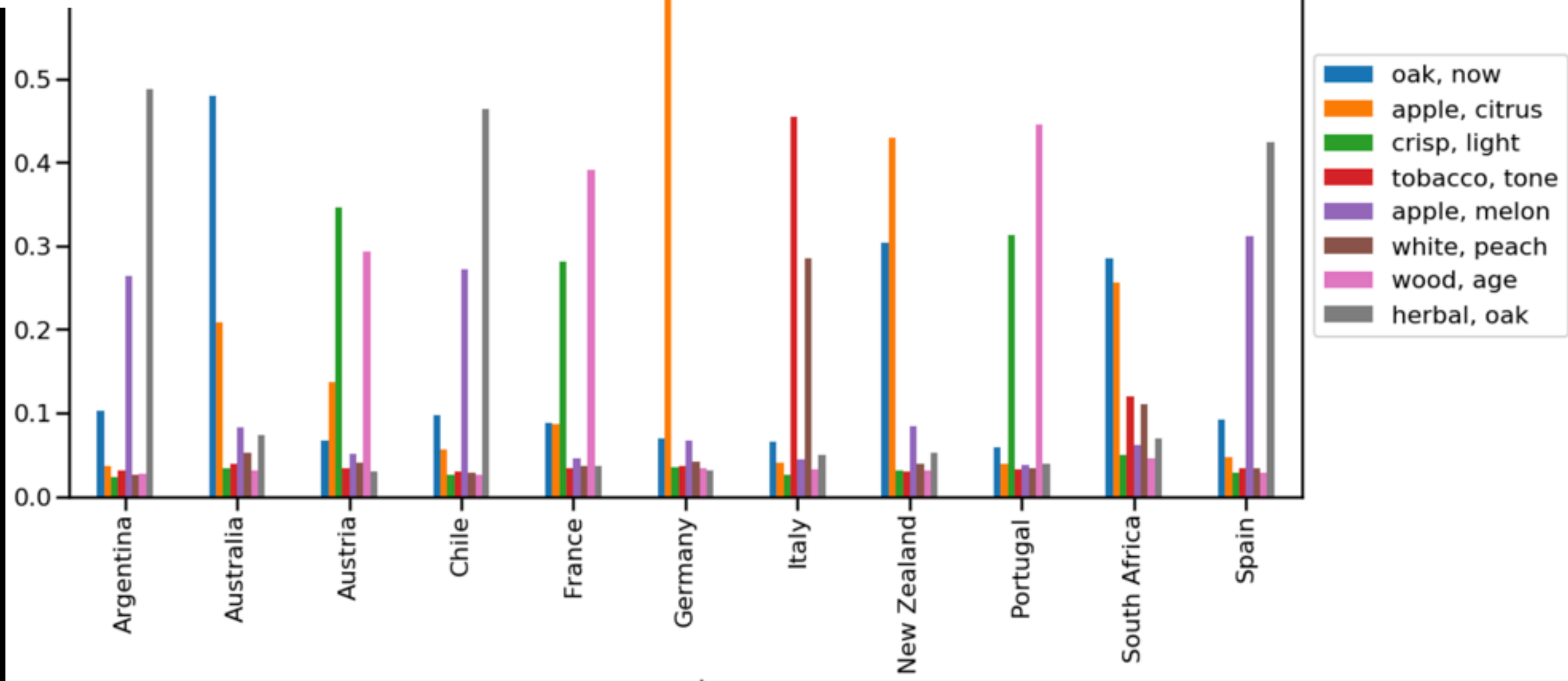
"America was actually on track to top $ 1 trillion in spending over the coming decade -- because the freedom and dignity --"

Bocconi

Student Projects

Wines from different Countries





Syllabus

Lesson	Topic
1	Intro to NLP 1
2	Linguistic Analysis
3	Information Retrieval
4	Regular Expressions
5	Discrete Representations and TFIDF
6	Discrete Representations
7	Continuous Representations
8	Word2Vec and Doc2Vec
9	Topic models 1
10	Topic models 2
11	Dimensionality reduction and Clustering, Visualization
12	Latent Dimensions
13	Text classification
14	Example: Sentiment Classification
15	The Perceptron
16	Multilayer Perceptrons
17	Structured Prediction
18	The Structured Perceptron
19	Convolutional and Recurrent Neural Networks
20	LSTMs in keras
21	CNN in keras
22	The Transformer and BERT
23	Ethics in NLP
24	Final Project Presentations

BASICS

EXPLORATION

PREDICTION

New and Improved!

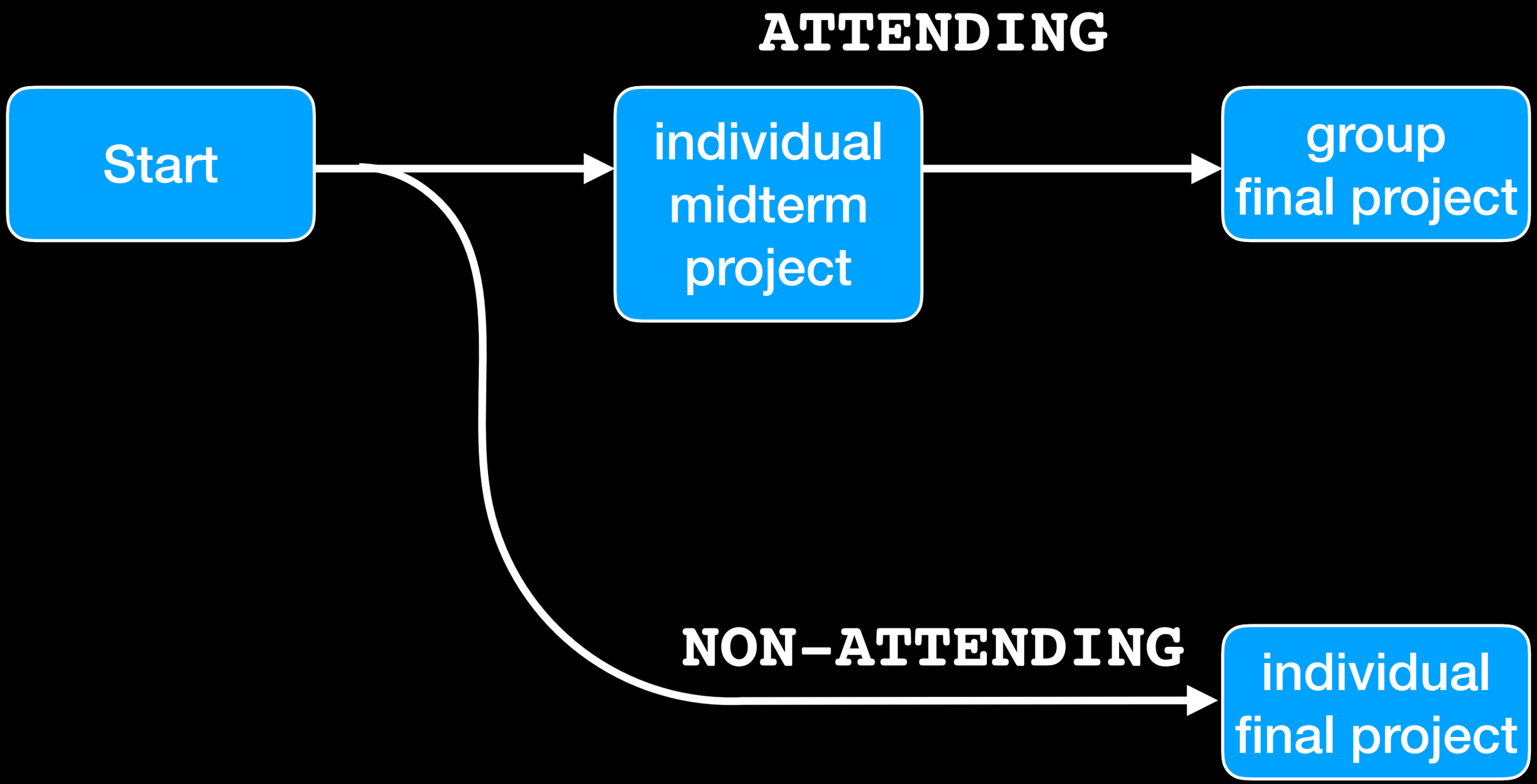
- Lecture notes now available as book draft for download from BBoard (first part as book at Cambridge University Press)
- Additional focus on latest neural network methods

Class Structure

- Thursdays: intuition, theory, math (slides)
- Fridays: exercises and practice (Jupyter Notebooks)

- Material on BBoard or at https://github.com/dirkhovy/NLPclass
- Latest book draft: http://www.dirkhovy.com/portfolio/papers/download/nlpss.pdf

Attendance



Bocconi

Grading

Individual midterm project (50%): Exploration and visualization

G***roup*** final project (50%): Data annotation and prediction, visualization

Individual final project (100%): whole class

- All projects are to be handed in as runnable **Jupyter Notebooks**
- Graded on data set size, correctness of implementations, annotation quality, performance of prediction
- No point changes, only **complete regrades** (total can go down)!

How do I succeed?

- Code well
- Pay attention
- Code some more

I want more NLP!

- join our reading group (Thu at 13:00), email Tommaso Fornaciari, `fornaciari@unibocconi.it`
- DMI online talk series:
  - February 17th, 12:30 CET, Ciro Cattuto (U. Torino)
  - March 1st, 12:30 CET/11:30 UK, Sebastian Ruder (DeepMind)
  - March 15th, 12:30 CET, Fabiana Zollo (U. Ca' Foscari Venice)
  - March 29th, 12:30 CET, Florian Ellsaesser (Frankfurt School of Business)
  - April 12th, 12:30 CET, Dong Nguyen (U. Utrecht)
  - April 26th, 16:30 CET, Nikita Nangia (NYU)
  - May 10th or 24th, 16:30 CET, Adina Williams (Facebook)

What to do with a problem

1. Don't panic.
2. Google it. stackoverflow.com is your friend
3. Talk to your classmates
4. Ask the TA, Tommaso Fornaciari
   `fornaciari@unibocconi.it`
5. Make a you@B appointment for my office hours (Mon, 18-19:30)

WARNING :

For any question we can solve with a Google search, we deduct points!

Let's start!

Bocconi

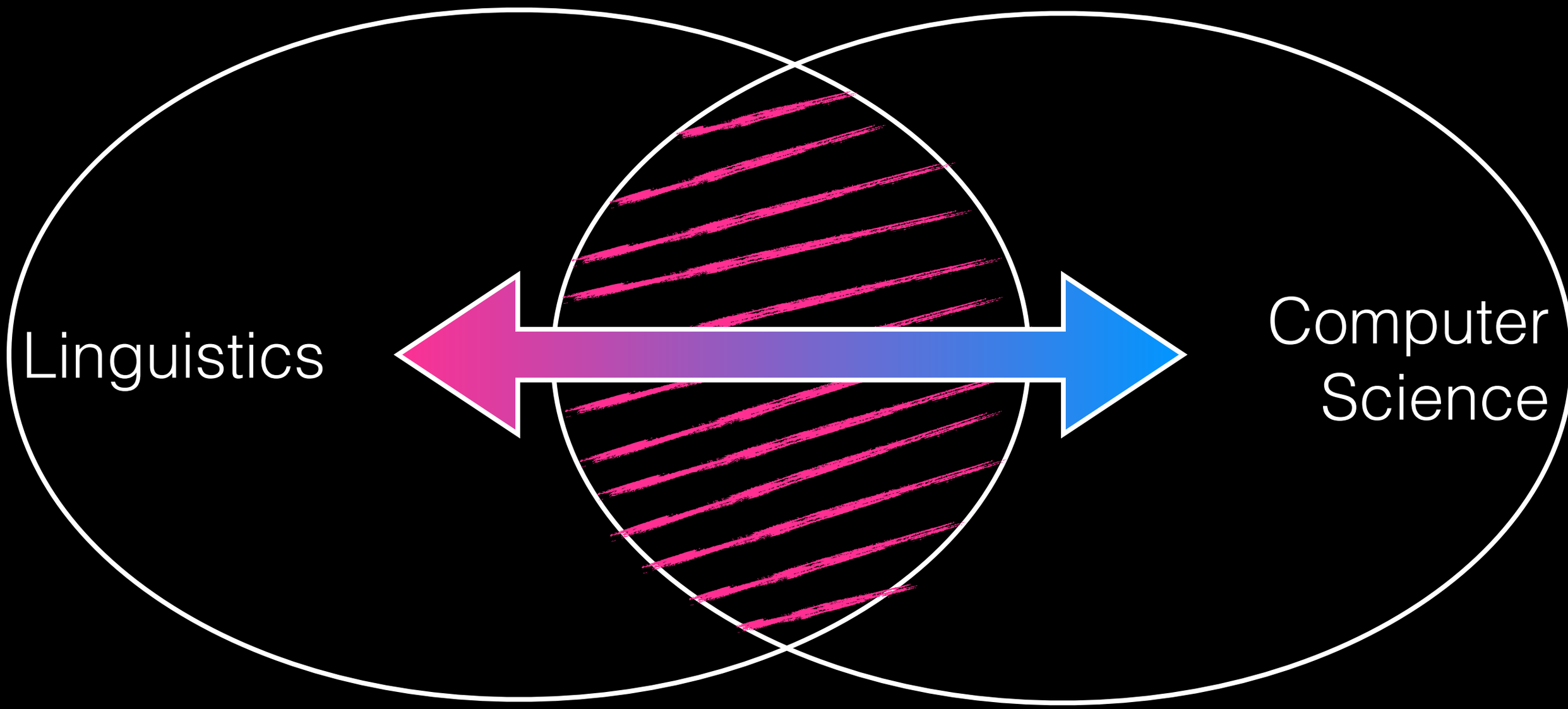
Today's Goals

- Understand where NLP comes from
- Learn about the different steps of preprocessing
- Understand the use of
  - parts of speech,
  - parsing, and
  - named entities

So, what's NLP anyway?

Bocconi

The two sides of NLP



informed linguistic hypotheses large-scale statistical analysis

Bocconi

A very Brief History of NLP

Symbolic Processing

Handwritten rules and logic

for input,

output,

and processing

Epoch 1

Statistical revolution

Statistical NLP

Engineered input

Defined output,

Automatic processing

Epoch 2

Deep learning

Deep Learning for NLP

Defined input and output,

Learned representations

end-to-end processing

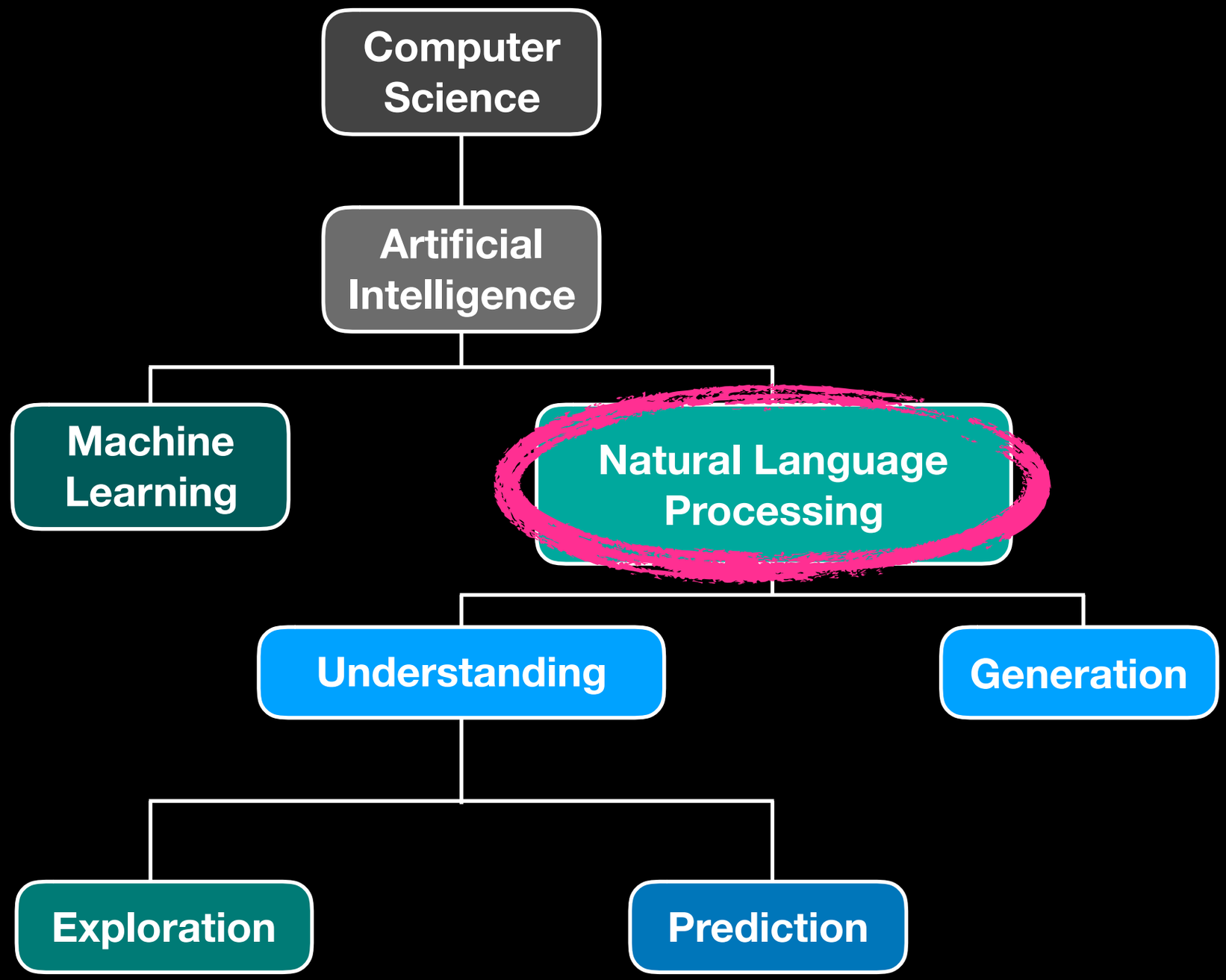
Epoch 3

approx. 1980s

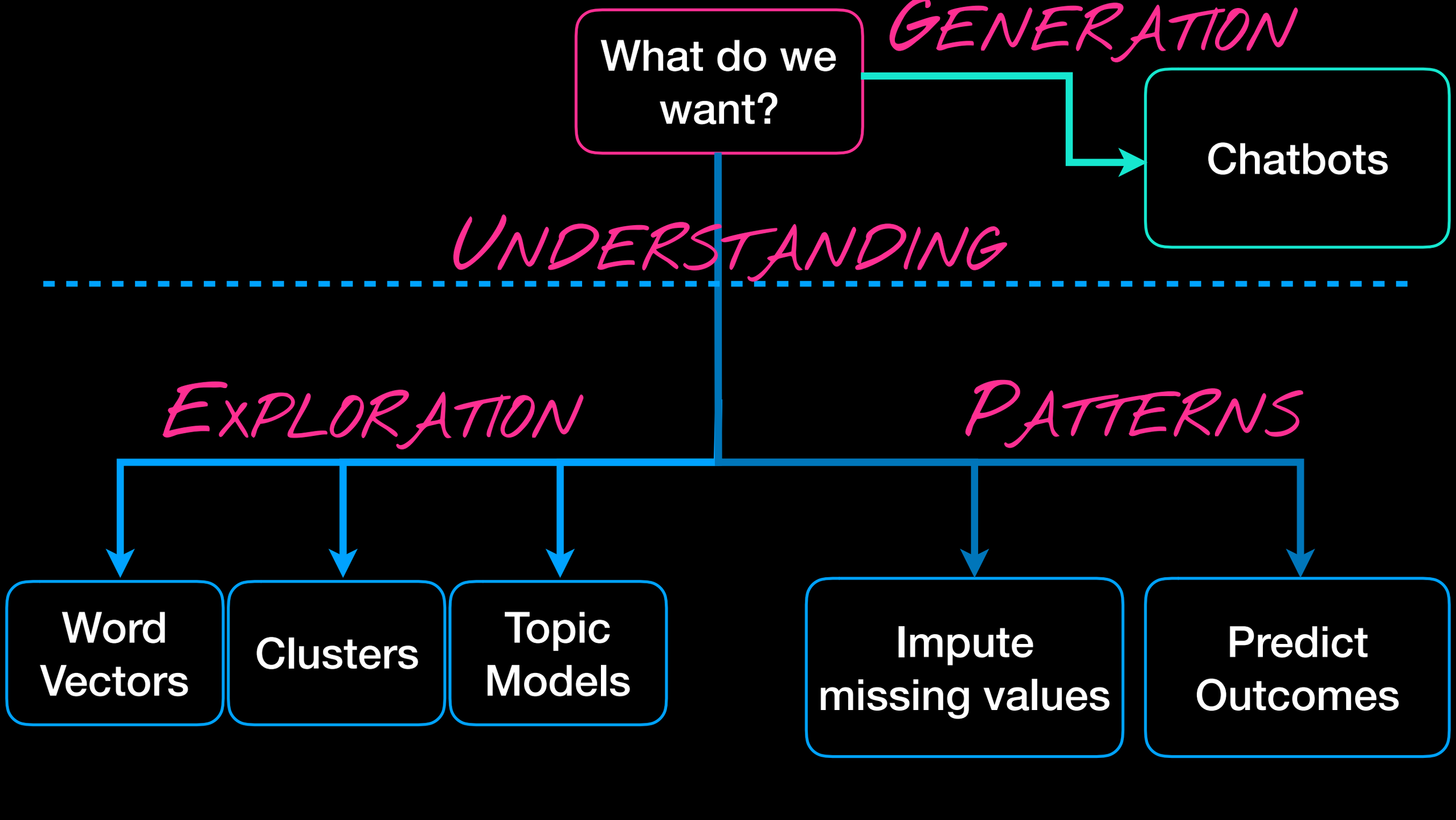
2015

The NLP Family

- Extract information from text: topics, trends
- Classify text sentiment, content type, author profiles
- Generate text: translations, automated responses



Two Uses of NLP



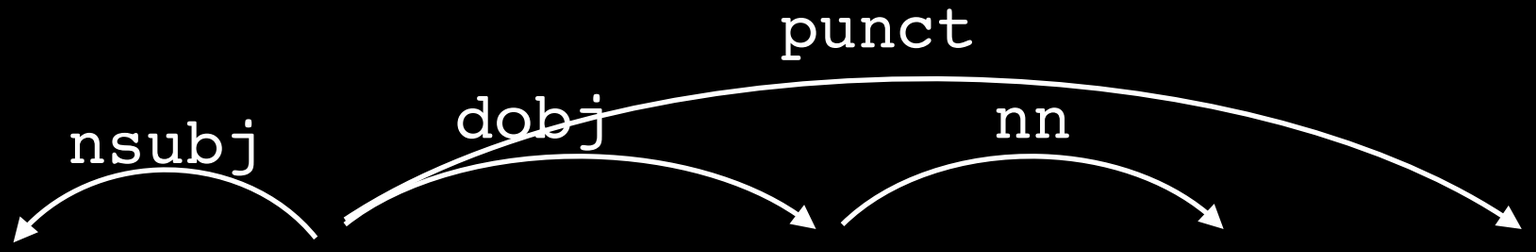
Linguistic Analysis

Bocconi

Examples of Analysis

NER O O PERSON PERSON O

PARSING



POS PRON VERB PROPN PROPN PUNCT

I admire Rosa Parks .

Pre-processing

Bocconi

Pre-processing steps

```
<div id="text">I've been in New York
in 2011, but didn’t like it. I
preferred Los Angeles.</div>
```

GOAL: MINIMIZE VARIATION

Pre-processing steps

- Remove formatting (e.g. HTML)
- Segment sentences
- Tokenize words
- Normalize words
    - numbers
    - lemmas vs. stems
- Remove unwanted words
    - stopwords
    - content words (use POS tagging!)
- join collocations

```
I've been in New York in 2011, but didn’t like it. I preferred Los Angeles.
```

Pre-processing steps

- Remove formatting (e.g. HTML)
- Segment sentences
- Tokenize words
- Normalize words
  - numbers
  - lemmas vs. stems
- Remove unwanted words
  - stopwords
  - content words (use POS tagging!)
- join collocations

```
I've been in New York in 2011, but didn’t like it.
```

```
I preferred Los Angeles.
```

Pre-processing steps

- Remove formatting (e.g. HTML)
- Segment sentences
- Tokenize words
- Normalize words
  - numbers
  - lemmas vs. stems
- Remove unwanted words
  - stopwords
  - content words (use POS tagging!)
- join collocations

```
I ’ve been in New York
in 2011 , but did n’t
like it .

I preferred Los
Angeles .
```

Pre-processing steps

- Remove formatting (e.g. HTML)
- Segment sentences
- Tokenize words
- Normalize words
  - numbers
  - lemmas vs. stems
- Remove unwanted words
  - stopwords
  - content words (use POS tagging!)
- join collocations

```
i ‘ve been in new york
in 0000 , but did n’t
like it .

i preferred los
angeles .
```

Pre-processing steps

- Remove formatting (e.g. HTML)
- Segment sentences
- Tokenize words
- Normalize words
  - numbers
  - lemmas vs. stems
- Remove unwanted words
  - stopwords
  - content words (use POS tagging!)
- join collocations

```
i have be in new york in 0000 , but do not like it .
i prefer los angeles .
```

Pre-processing steps

- Remove formatting (e.g. HTML)
- Segment sentences
- Tokenize words
- Normalize words
  - numbers
  - lemmas vs. stems
- Remove unwanted words
  - stopwords
  - content words (use POS tagging!)
- join collocations

```
i new york 0000 , like .

i prefer los angeles .
```

Pre-processing steps

- Remove formatting (e.g. HTML)
- Segment sentences
- Tokenize words
- Normalize words
  - numbers
  - lemmas vs. stems
- Remove unwanted words
  - stopwords
  - content words (use POS tagging!)
- join collocations

```
new york 0000 like

prefer los angeles
```

CONTENT = (NOUN, VERB, NUM)

Pre-processing steps

- Remove formatting (e.g. HTML)
- Segment sentences
- Tokenize words
- Normalize words
  - numbers
  - lemmas vs. stems
- Remove unwanted words
  - stopwords
  - content words (use POS tagging!)
- join collocations

```
new_york 0000 like

prefer los_angeles
```

Pre-processing steps

```
<div id="text">I've been in New York
in 2011, but didn’t like it. I
preferred Los Angeles.</div>
```

MINIMAL VARIATION

"BAG OF WORDS"

```
new_york 0000 like

prefer los_angeles
```

Parts of Speech

Bocconi

POS tagging

Grassfed highland Chianina beef with handcut fries and seasonal micro greens 29,—

Rich, **tender**, **golden-brown** beef with **crisp** fries and **tender** greens 18,—

Savory beef with **delicious** fries and **tasty** salad 12,—

ADJs = price?

POS tagging

POS	PRON	VERB	PROPN	PROPN	PUNCT
I	admire	Rosa	Parks	.

POS tagging

Open class words	Closed class words	Other
ADJ adjectives: *awesome, red*	**ADP** adpositions: *over, before*	**PUNCT** punctuation marks: *!, ?, –*
ADV adverbs: *quietly, where, never*	**AUX** auxiliary/modal verbs: *have* (been), *could* (do), *will* (change)	**SYM** symbols: *%, $, :)*
INTJ interjections: *ouch, shhh*	**CCONJ** coordinating conjunctions: *and, or, but*	**X** other: *pffffrt*
NOUN nouns: *book, war*	**DET** determiners: *a, they, which*	
PROPN proper nouns: *Rosa, Twitter*	**NUM** numbers. Exactly what you would think it is...	
VERB full verbs: (she) *codes*, (they) *submitted*	**PART** particles: 's	
	PRON pronouns: *you, her, myself*	
	SCONJ subordinating conjunctions: *since, if, that*	

POS tagging

show {VERB, NOUN}

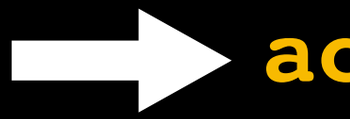
Structured prediction: depends on the POS of a previous word

Parsing

Bocconi

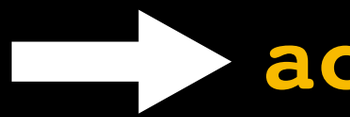
Dependency Parsing

Facebook eventually **acquired** **WhatsApp** after hard negotiations.



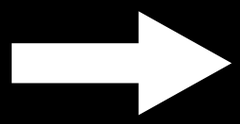
acquire(Facebook, WhatsApp)

WhatsApp was **acquired** by **Facebook**.



acquire(Facebook, WhatsApp)

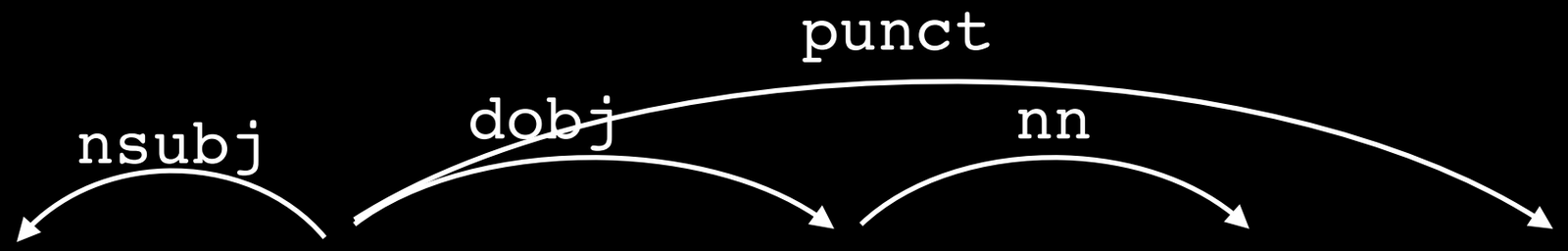
Facebook subsidiary **WhatsApp** to **acquire** new look.



acquire(WhatsApp, look)

Dependency Parsing

PARSING

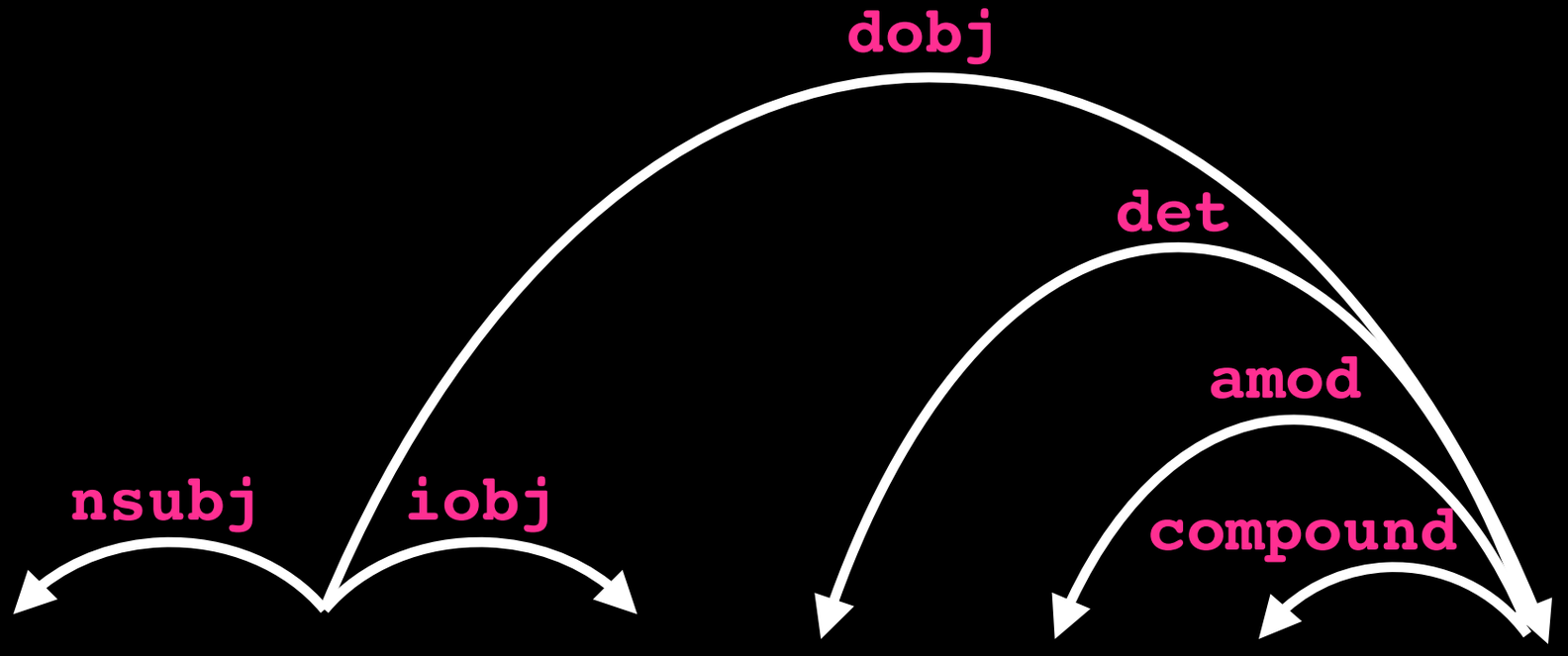


POS PRON VERB PROPN PROPN PUNCT

I admire Rosa Parks .

Dependency Parsing

acl: adjectival clause
advcl: adverbial clause modifier
advmod: adverbial modifier
amod: adjectival modifier
appos: appositional modifier
aux: auxiliary
case: case marking
cc: coordinating conjunction
ccomp: clausal complement
clf: classifier
compound: compound
conj: conjunct
cop: copula
csubj: clausal subject
dep: unspecified dependency
det: determiner
discourse: discourse element
dislocated: dislocated elements
dobj: direct object
expl: expletive
fixed: fixed multiword expression
flat: flat multiword expression
goeswith: goes with
iobj: indirect object
list: list
mark: marker
nmod: nominal modifier
nsubj: nominal subject
nummod: numeric modifier
obl: oblique nominal
orphan: orphan
parataxis: parataxis
punct: punctuation
reparandum: overridden disfluency
root: root
vocative: vocative
xcomp: open clausal complement



Named Entities

Bocconi

Named Entities

Places:

```
{'Ada',
 'Antigone',
 'Belgrade',
 'Berlin',
 'Constitución',
 'Danube',
 'Florence',
 'France',
 'Mikser',
 'Rome',
 'Santa Cruz',
 'Savamala',
 'Schlachtensee',
 'Serbia',
 'Spain',
 'Tezga',
 'Ville',
 'Wannsee'}
```


Bocconi

Named Entities

NER	O	O	B-PERSON	I-PERSON	O
POS	PRON	VERB	PROPN	PROPN	PUNCT
	I	admire	Rosa	Parks	.

Named Entities

NE	Example
PERSON	
NORP (Nationality OR Religious or Political group)	
FAC (facility)	
ORG (organization)	
GPE (GeoPolitical Entity)	
LOC (locations, such as seas or mountains)	
PRODUCT	
EVENT (in sports, politics, history, etc.)	
WORK_OF_ART	
LAW	
LANGUAGE	
DATE	
TIME	
PERCENT	
MONEY	
QUANTITY	
ORDINAL	
CARDINAL (numbers)	

Wrapping up

Bocconi

Take Home Points

- NLP is a subfield of AI, using ML on linguistic problems to **explore**, **predict**, and **generate** text
- **Preprocessing** removes noise and unwanted variation
- Parts of speech (**POS**) denote a word's grammatical *category*
- **Parsing** denotes a word's grammatical *function*
- **Named entities** categorize a noun's semantic type