Natural Language Processing

Lecture 01

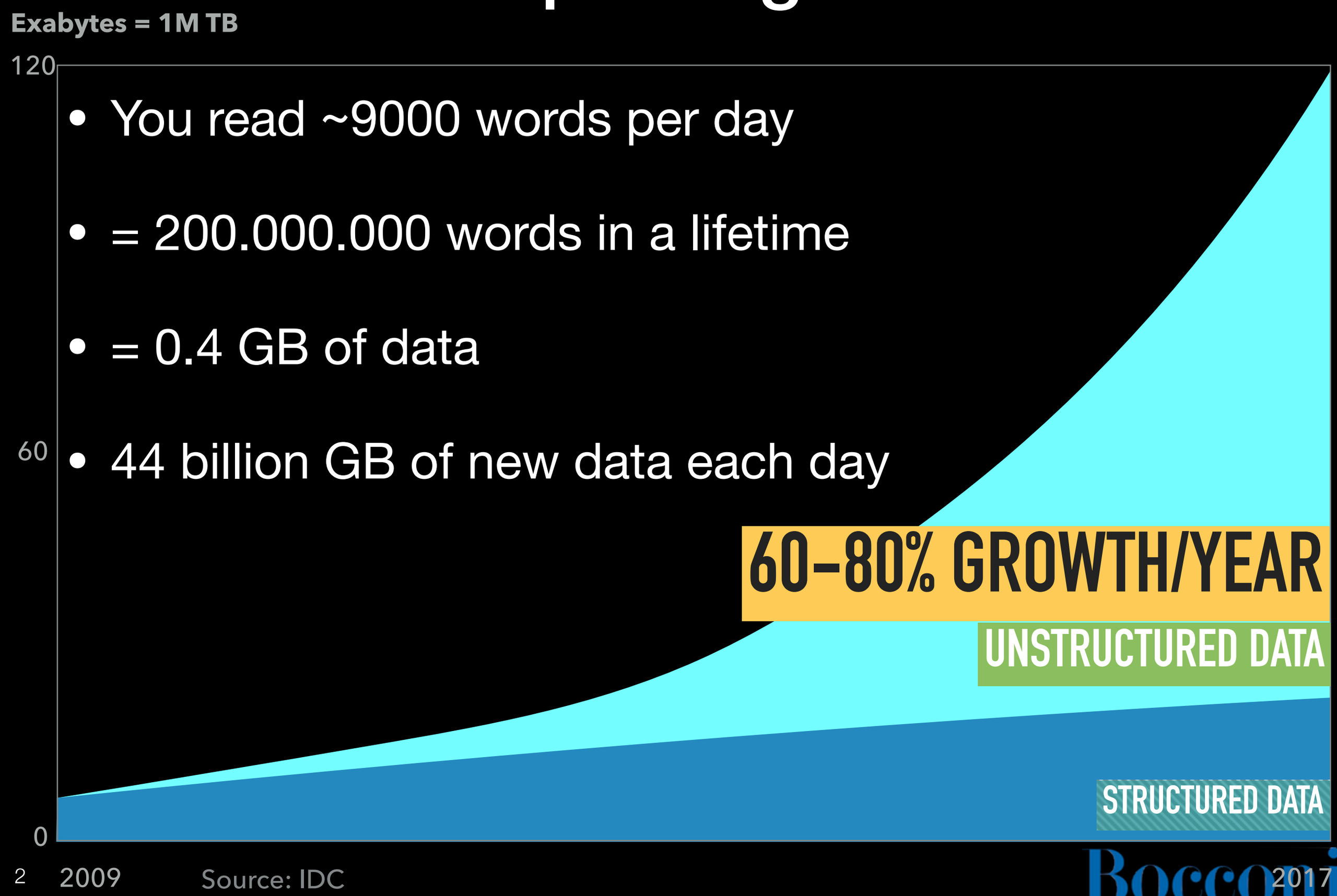
Dirk Hovy

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

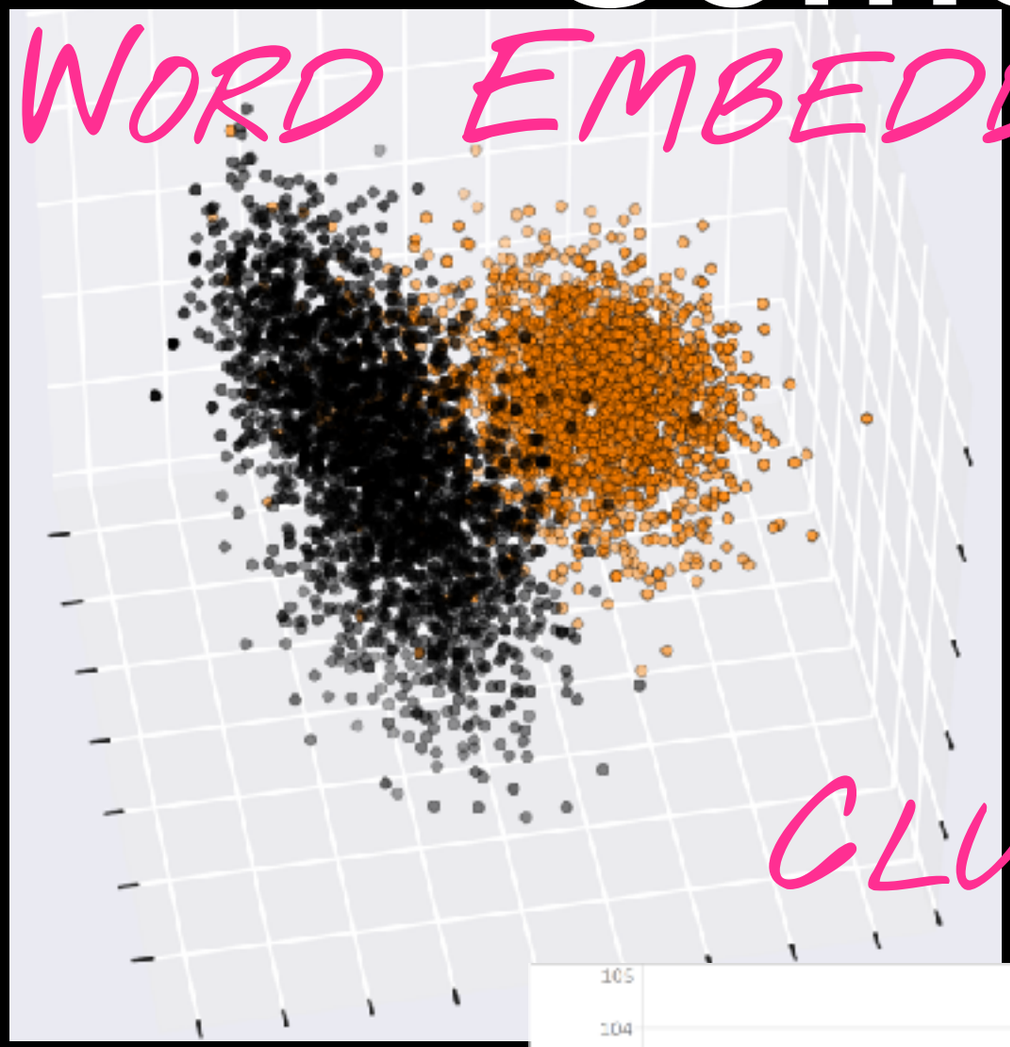
NLP is booming

Source: Tractica

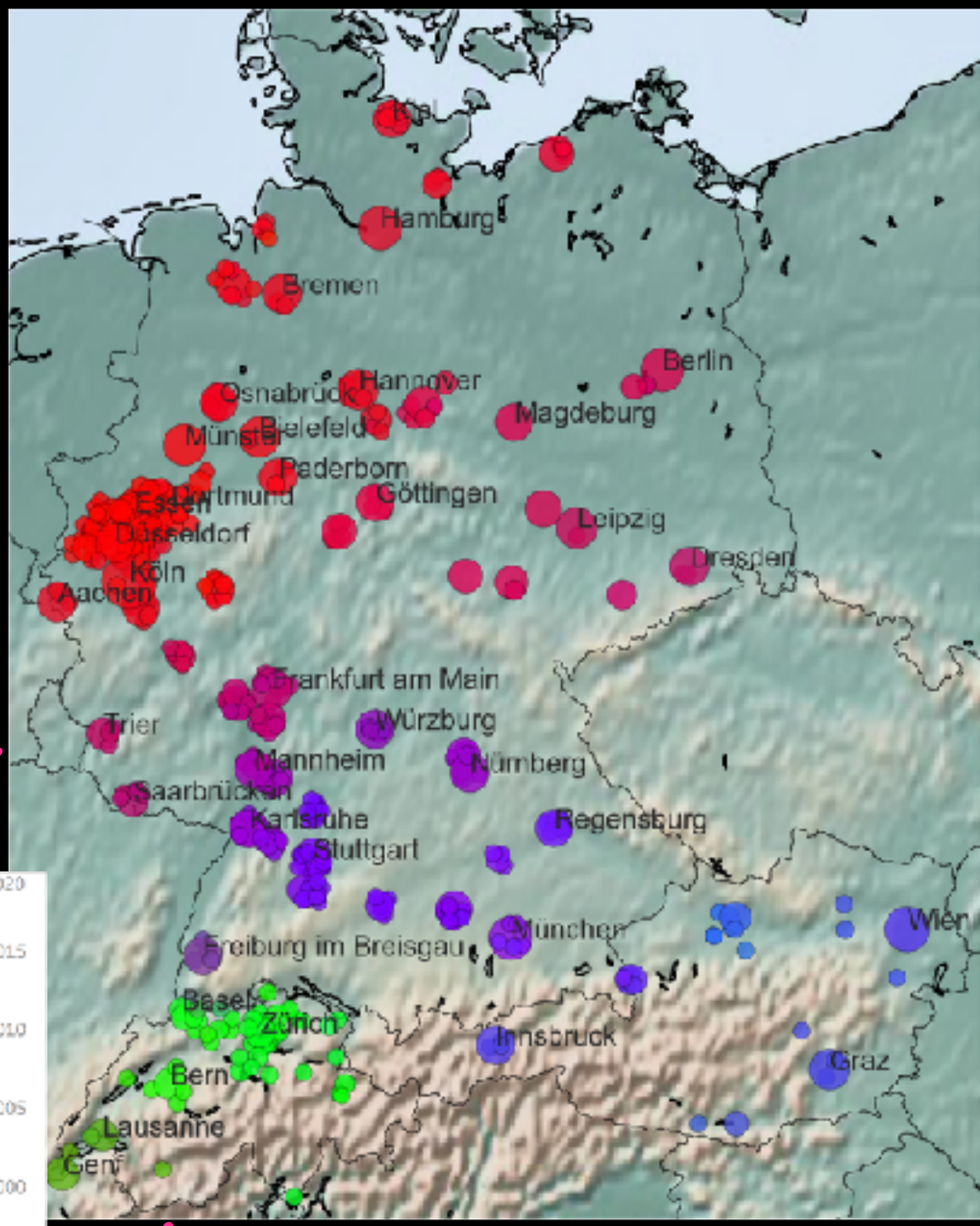
Bocconi

Some examples

Word Embeddings



Clustering



Visualization

Sentiment Analysis

Bocconi

Syllabus

Lesson	Topic
1	Intro 1
2	Intro 2
3	Representations
4	Embeddings
5	Information retrieval 1
6	Information retrieval 2
7	Language models 1
8	Language models 2
9	Topic models 1
10	Topic models 2
11	Dimensionality reduction and Clustering
12	Visualization
13	Retrofitting
14	Midterm Project Presentations
15	Text classification
16	Application: Sentiment Analysis
17	Improving classification performance
18	Application: Author attribute prediction
19	Neural networks basics
20	Multilayer Perceptron
21	Sequence models
22	RNNs and Bi-LSTMs
23	Final Project Presentations
24	Final Project Presentations

Class structure

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

Grading

1. *Individual* midterm project (50%): Exploration and visualization

2. Final *group* project (50%): Data annotation, exploration, visualization, and prediction

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

How do I succeed?

- Code well
- Pay attention
- Code some more

What to do with a problem

1. Google it. stackoverflow.com is your friend

2. Talk to your classmates

3. Ask the TA, Tommaso Fornaciari
   `fornaciari@unibocconi.it`

4. Make an appointment

WARNING :

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

Let's start!

Bocconi

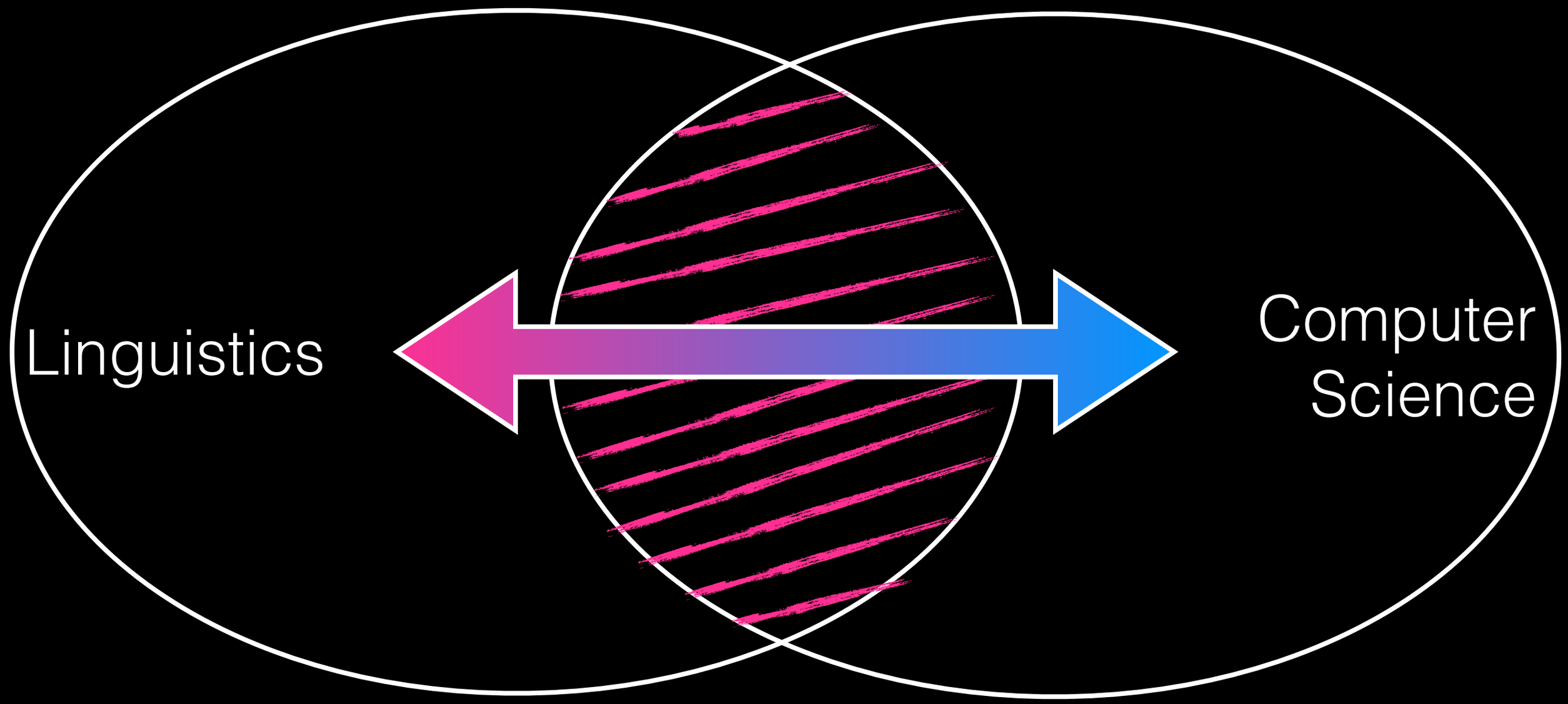
Today's Goals

- 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

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

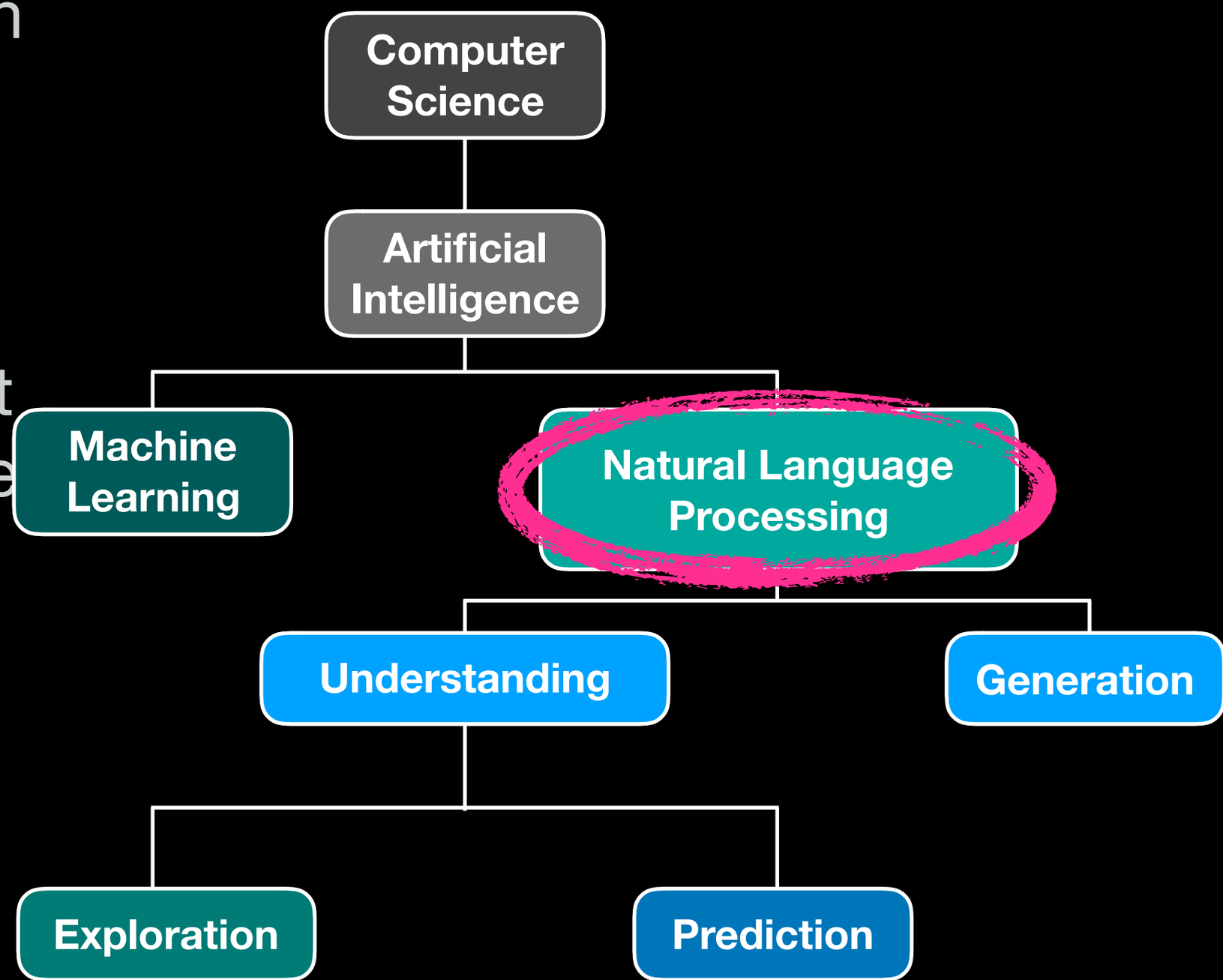
Epoch 3

approx. 1980s

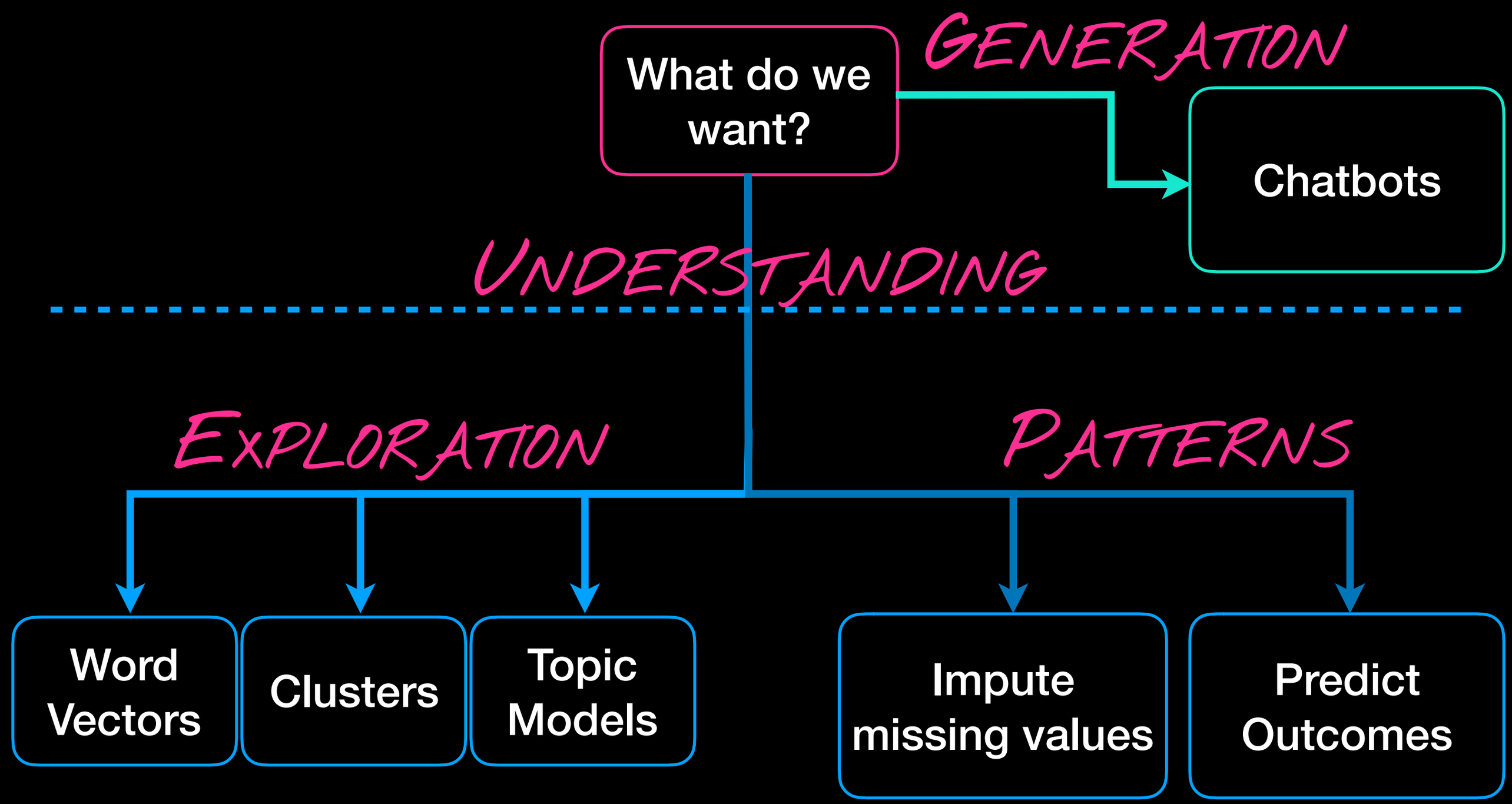
2015

Structure of NLP

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



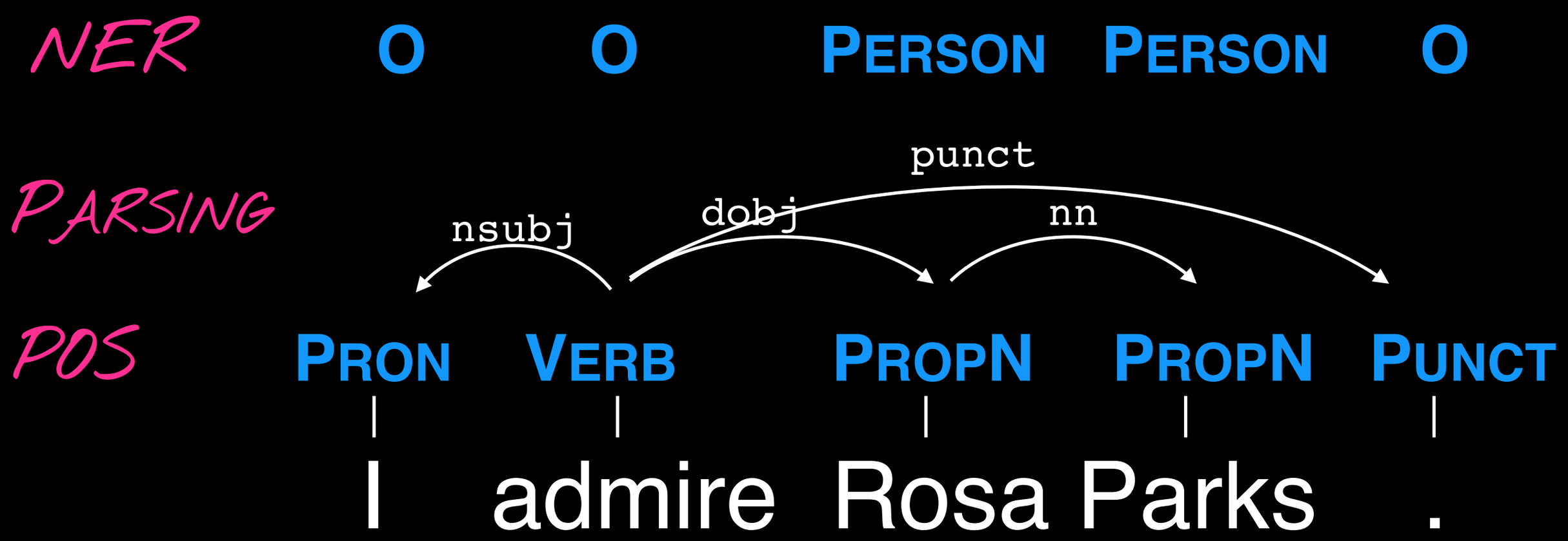
Two Uses of NLP



Linguistic Analysis

Bocconi

Examples of Analysis



Pre-processing

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>
```

MINIMIMAL 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}

PART **show**

show

PRON **show**

show

show

DET **show**

show

ADJ **show**

show

Structured prediction: depends on the POS of a previous word

Parsing

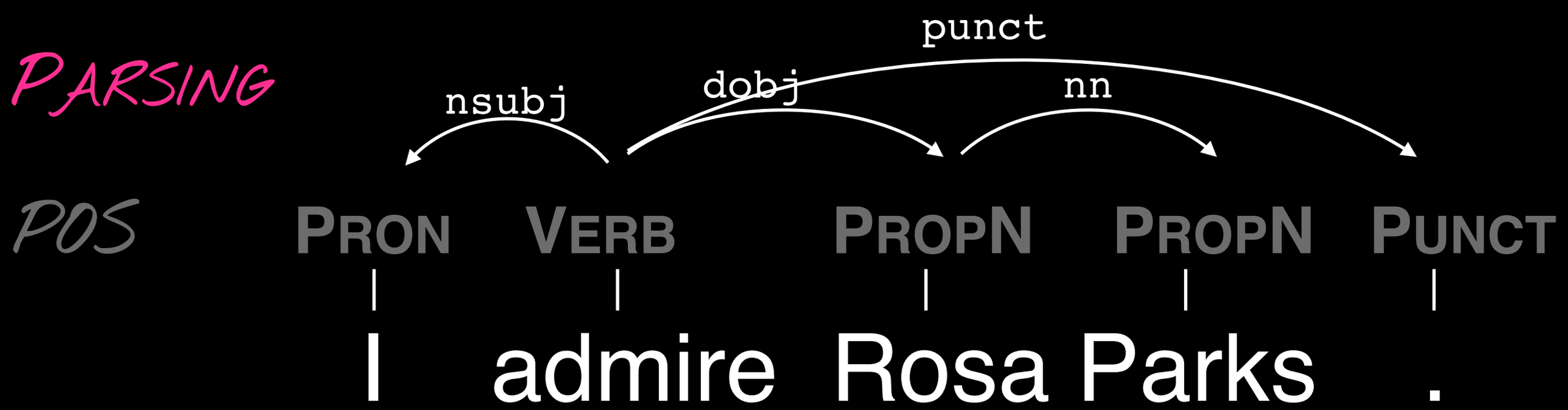
Dependency Parsing

Facebook eventually acquired WhatsApp after hard negotiations. → acquire(Facebook, WhatsApp)

WhatsApp was acquired by Facebook. → acquire(Facebook, WhatsApp)

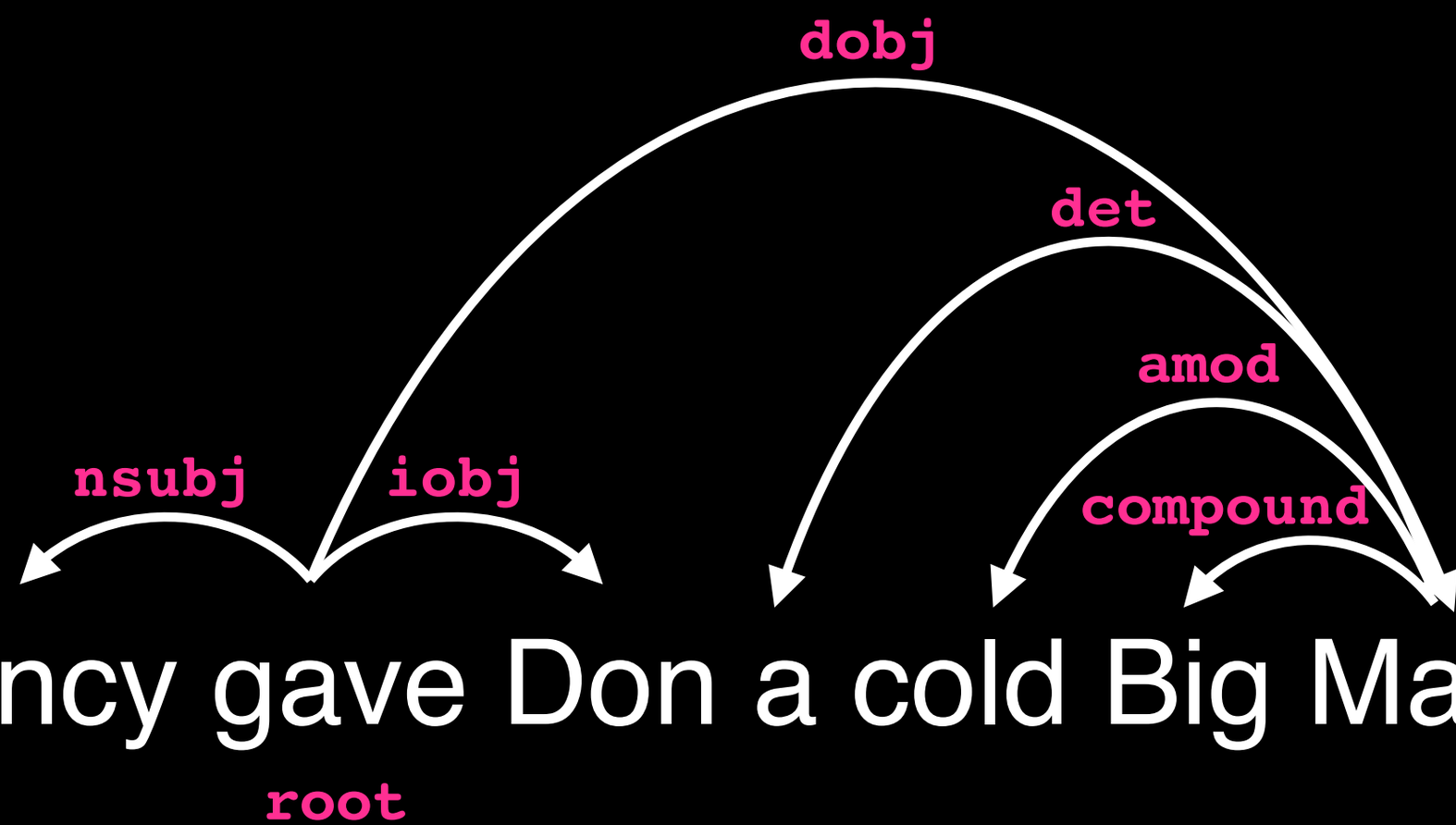
Facebook subsidiary WhatsApp to acquire new look. → acquire(WhatsApp, look)

Dependency Parsing



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



Bocconi

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'}
```

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