

# CS 7650: Natural Language Processing

Alan Ritter

(many slides from Greg Durrett)

# Administrivia

---

- ▶ Course website:  
<https://aritter.github.io/CS-7650-sp22/>
- ▶ Piazza and Gradescope: links on the course website
  - ▶ We will do our best to make sure questions are answered within 24 hours (or Monday/Tuesday for questions asked over the weekend).
- ▶ TA Office hours:
  - ▶ See spreadsheet

## Instructor



[Alan Ritter](#)

[alan.ritter@cc.gatech.edu](mailto:alan.ritter@cc.gatech.edu)

## Teaching Assistants

Jan Vijay Singh

[iamjanvijay@gatech.edu](mailto:iamjanvijay@gatech.edu)

Mukund Rungta

[mrungta8@gatech.edu](mailto:mrungta8@gatech.edu)

Vinay Sammangi

[vsammangi3@gatech.edu](mailto:vsammangi3@gatech.edu)

Xurui Zhang

[xuruizhang@gatech.edu](mailto:xuruizhang@gatech.edu)

# Prerequisites

---

- ▶ Probability
- ▶ Linear Algebra
- ▶ Multivariable Calculus
- ▶ Programming / Python experience
- ▶ Prior exposure to machine learning very helpful but not required

# Prerequisites

---

- ▶ Probability
- ▶ Linear Algebra
- ▶ Multivariable Calculus
- ▶ Programming / Python experience
- ▶ Prior exposure to machine learning very helpful but not required

There will be a lot of math and programming!

# Coursework Plan

---

- ▶ 3 Programming Projects (fairly substantial implementation effort)
  - ▶ Text classification
  - ▶ Named entity recognition (BiLSTM-CNN-CRF)
  - ▶ Neural chatbot (Seq2Seq with attention)

# Coursework Plan

---

- ▶ 3 Programming Projects (fairly substantial implementation effort)
  - ▶ Text classification
  - ▶ Named entity recognition (BiLSTM-CNN-CRF)
  - ▶ Neural chatbot (Seq2Seq with attention)
- ▶ 2 written assignments + midterm exam
- ▶ Mostly math problems related to ML / NLP

# Coursework Plan

---

- ▶ 3 Programming Projects (fairly substantial implementation effort)
  - ▶ Text classification
  - ▶ Named entity recognition (BiLSTM-CNN-CRF)
  - ▶ Neural chatbot (Seq2Seq with attention)
- ▶ 2 written assignments + midterm exam
  - ▶ Mostly math problems related to ML / NLP
- ▶ Final project (details on course website, will discuss later)

# Coursework Plan

---

- ▶ 3 Programming Projects (fairly substantial implementation effort)
  - ▶ Text classification
  - ▶ Named entity recognition (BiLSTM-CNN-CRF)
  - ▶ Neural chatbot (Seq2Seq with attention)
- ▶ 2 written assignments + midterm exam
  - ▶ Mostly math problems related to ML / NLP
- ▶ Final project (details on course website, will discuss later)
- ▶ Problem Set 1 (background review) is out now on Gradescope (due Jan 14)

# Problem Set 1 (Background Review)

---

- ▶ Due Jan 14 (this Friday).
- ▶ Background review on probability, linear algebra, calculus.
- ▶ **Waitlisted students:** please submit PS1 by Friday if you plan to enroll in the course.
  - ▶ We can't predict whether or not you will get in, as this depends on other students dropping the class...
- ▶ Submit on Gradescope

| Schedule |                                     |   |
|----------|-------------------------------------|---|
| Jan 10:  | <a href="#">Course Introduction</a> | <a href="#">Eisenstein Chapter 1</a>            |
| Jan 12:  | <a href="#">Machine Learning</a>    | <a href="#">Eisenstein 2.0-2.5, 4.1,4.3-4.5</a> |
| Jan 13:  | <a href="#">Problem Set 1 due</a>   |   |
| Jan 17:  | MLK Holiday                         |   |
| TBD:     | <a href="#">Project 1</a>           |   |

# Project 1 is also out (please look!)

The screenshot shows a Jupyter Notebook interface with the following details:

- Title:** TextClassification\_release.ipynb
- Toolbar:** File, Edit, View, Insert, Runtime, Tools, Help. Last saved at January 8.
- Code Cell:** Contains Python code for licensing and attribution:

```
# Licensing Information: You are free to use or extend this project for
# educational purposes provided that (1) you do not distribute or publish
# solutions, (2) you retain this notice, and (3) you provide clear
# attribution to The Georgia Institute of Technology, including a link to https://aritter.github.io/CS-7650/
#
# Attribution Information: This assignment was developed at The Georgia Institute of Technology
# by Alan Ritter (alan.ritter@cc.gatech.edu)
```
- Section Header:** Project #1: Text Classification
- Text Content:** In this assignment, you will implement the perceptron algorithm, and a simple, but competitive neural bag-of-words model, as described in [this paper](#) for text classification. You will train your models on a (provided) dataset of positive and negative movie reviews and report accuracy on a test set.  
In this notebook, we provide you with starter code to read in the data and evaluate the performance of your models. After completing the instructions below, please follow the instructions at the end to submit your notebook and other files to Gradescope.  
Make sure to make a copy of this notebook, so your changes are saved.

## Schedule

|         |                                     |   |
|---------|-------------------------------------|---|
| Jan 10: | <a href="#">Course Introduction</a> | <a href="#">Eisenstein Chapter 1</a>            |
| Jan 12: | <a href="#">Machine Learning</a>    | <a href="#">Eisenstein 2.0-2.5, 4.1,4.3-4.5</a> |
| Jan 13: | <a href="#">Problem Set 1 due</a>   |   |
| Jan 17: | MLK Holiday                         |   |
| TBD:    | <a href="#">Project 1</a>           |   |

# Free Textbooks!

---

- ▶ 2 really awesome free textbooks available
  - ▶ There will be assigned readings from both
  - ▶ Both freely available online

Natural Language Processing

**Speech and Language Processing** (3rd ed. draft)

[Dan Jurafsky](#) and [James H. Martin](#)

Jacob Eisenstein

# Programming Projects: Computation

---

- ▶ Modern NLP methods require non-trivial computation
  - ▶ Training neural networks with many parameters can take a long time (it is a very good idea to start working on the assignments early!)
  - ▶ You probably want to use a GPU
  - ▶ Google Colab: free GPUs (some limitations)
  - ▶ The programming projects are designed with Colab in mind
  - ▶ Colab Pro subscription (\$10/month). This is highly recommended once we start working with PyTorch.



# What's the goal of NLP?

---

# What's the goal of NLP?

---

- ▶ Be able to solve problems that require deep understanding of text

# What's the goal of NLP?

---

- ▶ Be able to solve problems that require deep understanding of text
- ▶ Example: dialogue systems

# What's the goal of NLP?

---

- ▶ Be able to solve problems that require deep understanding of text
- ▶ Example: dialogue systems



# What's the goal of NLP?

---

- ▶ Be able to solve problems that require deep understanding of text
- ▶ Example: dialogue systems



# What's the goal of NLP?

---

- ▶ Be able to solve problems that require deep understanding of text
- ▶ Example: dialogue systems



# What's the goal of NLP?

---

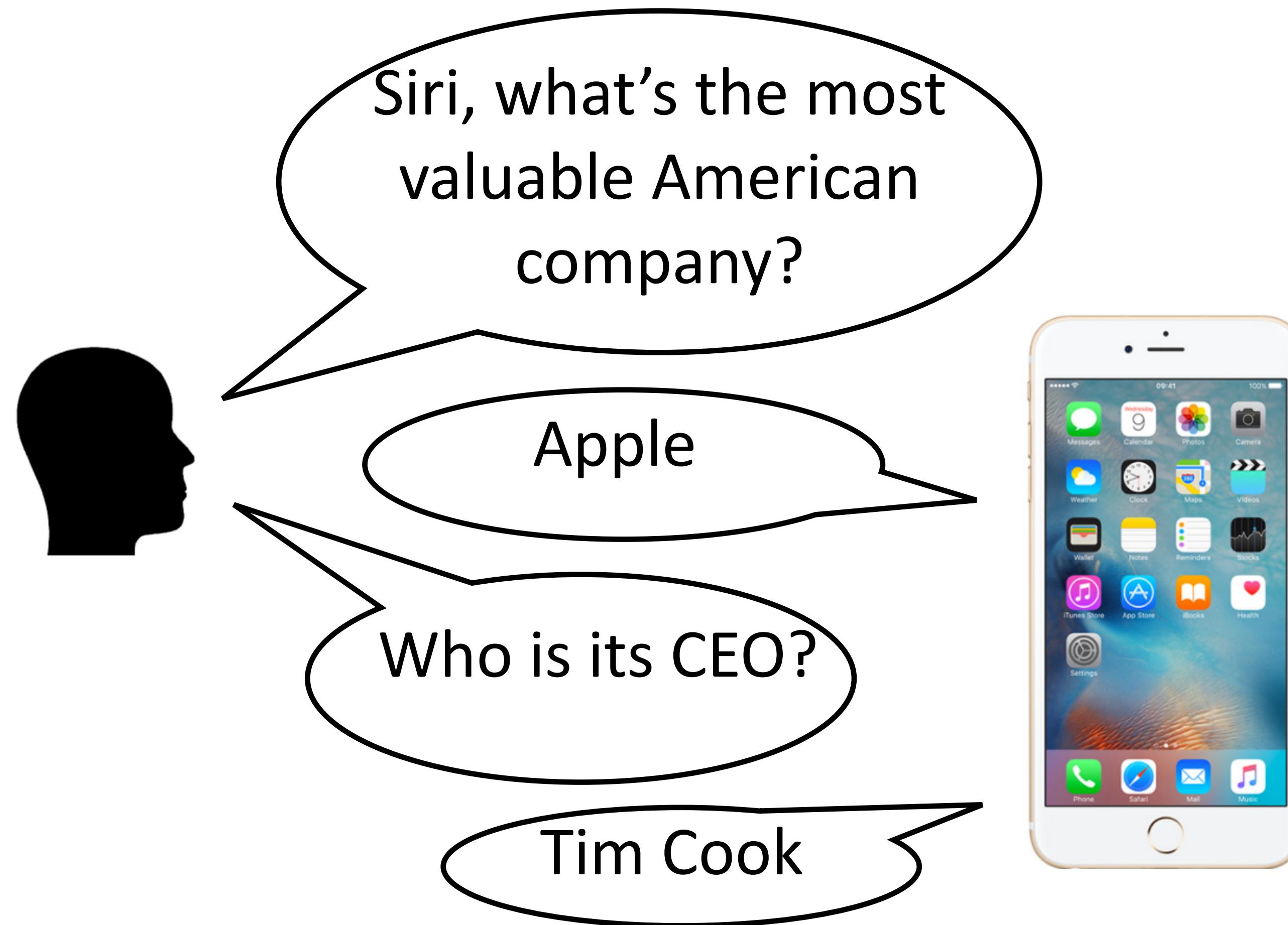
- ▶ Be able to solve problems that require deep understanding of text
- ▶ Example: dialogue systems



# What's the goal of NLP?

---

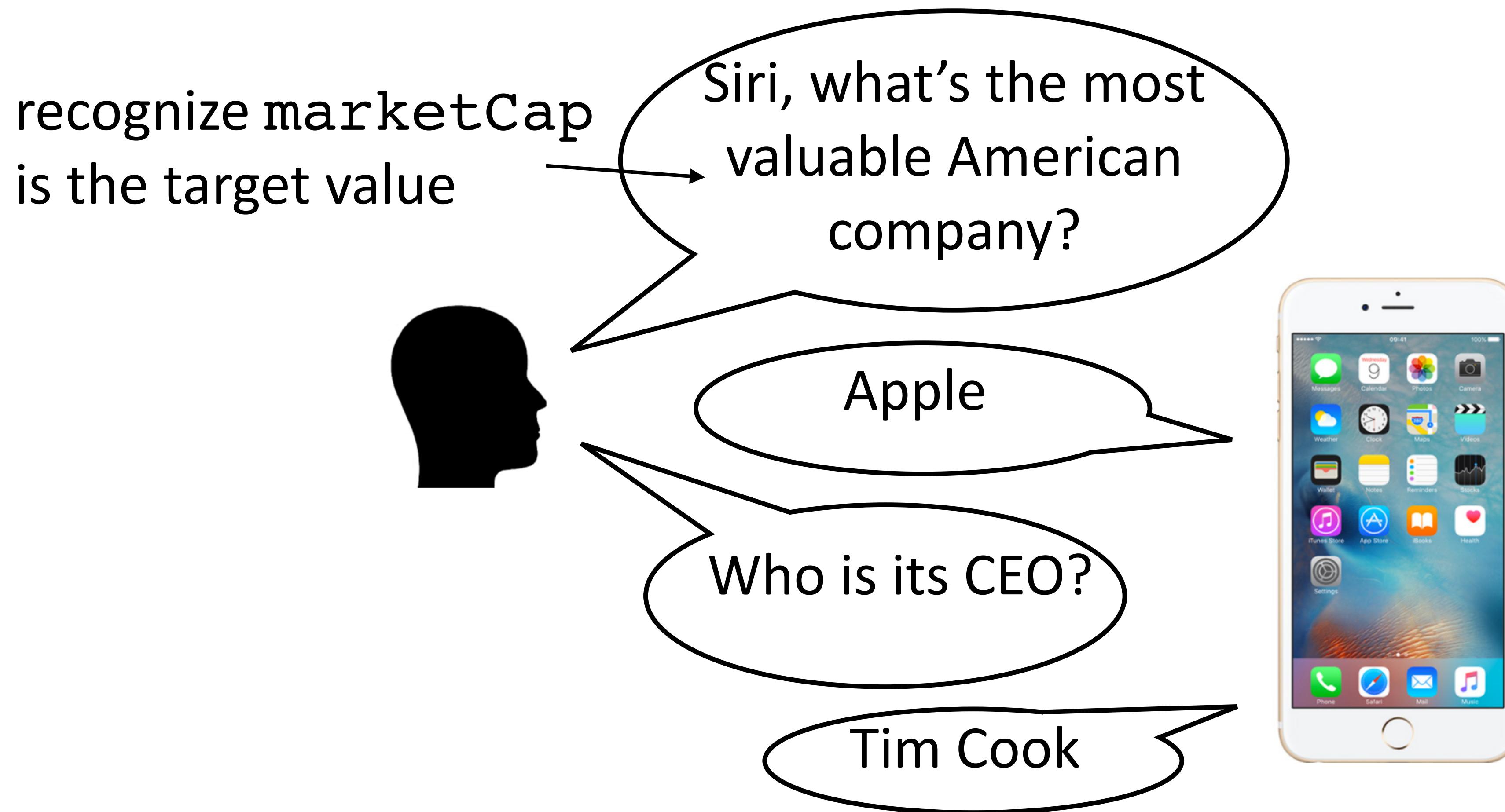
- ▶ Be able to solve problems that require deep understanding of text
- ▶ Example: dialogue systems



# What's the goal of NLP?

---

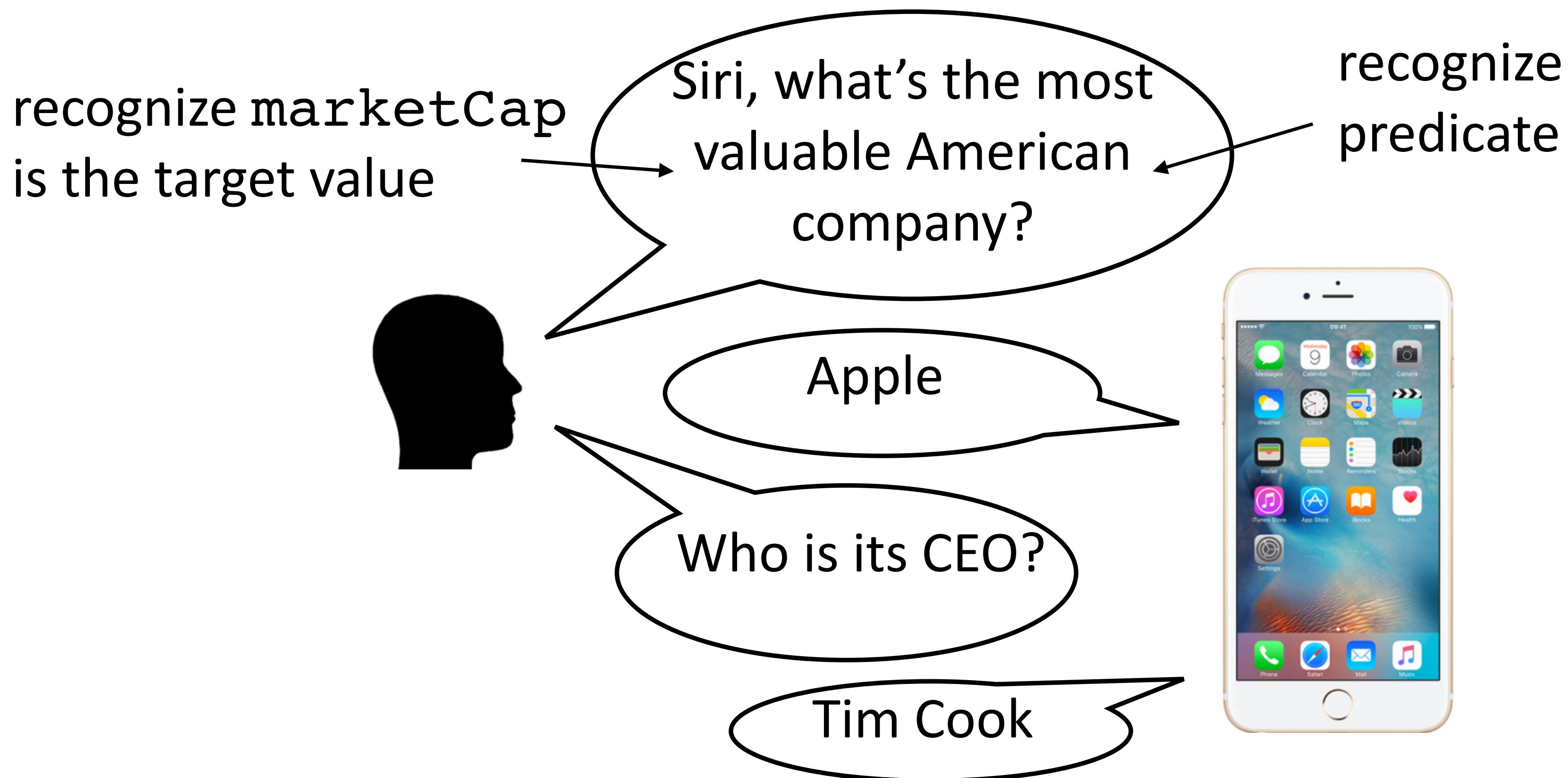
- ▶ Be able to solve problems that require deep understanding of text
- ▶ Example: dialogue systems



# What's the goal of NLP?

---

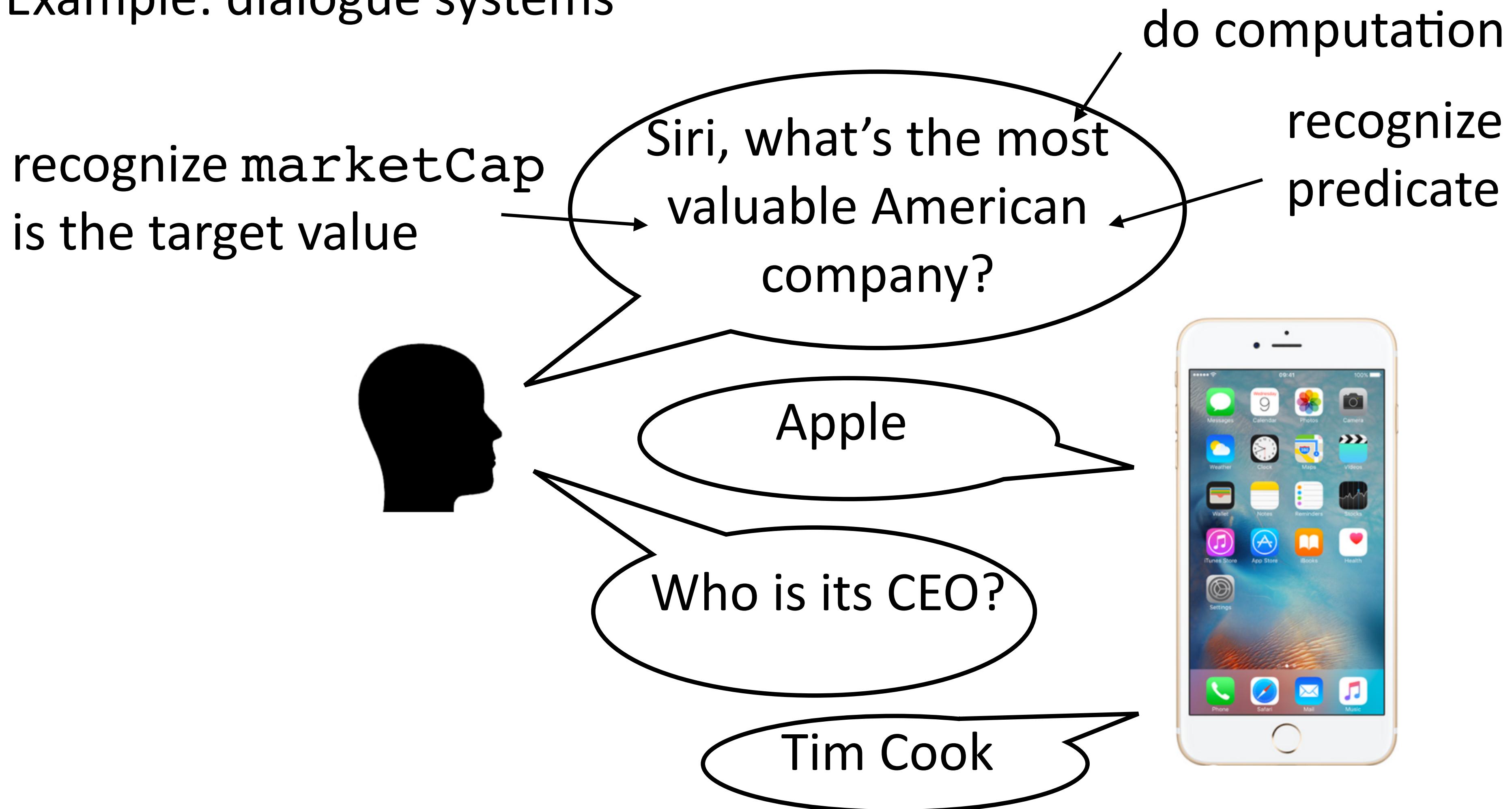
- ▶ Be able to solve problems that require deep understanding of text
- ▶ Example: dialogue systems



# What's the goal of NLP?

---

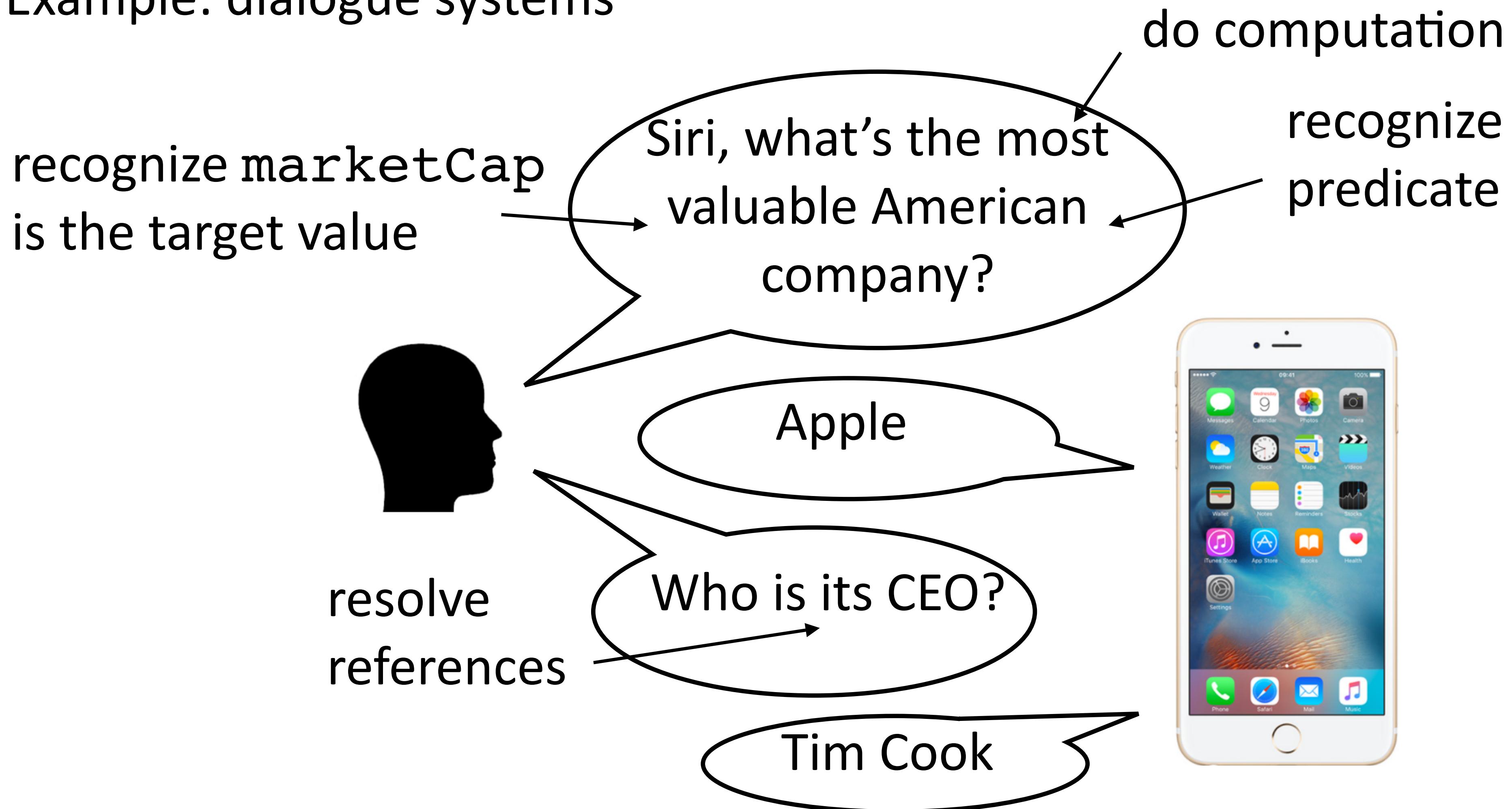
- ▶ Be able to solve problems that require deep understanding of text
- ▶ Example: dialogue systems



# What's the goal of NLP?

---

- ▶ Be able to solve problems that require deep understanding of text
- ▶ Example: dialogue systems



# Automatic Summarization

---

# Automatic Summarization

---

POLITICS

## ***Google Critic Ousted From Think Tank Funded by the Tech Giant***

WASHINGTON — In the hours after European antitrust regulators levied a record [\\$2.7 billion fine](#) against Google in late June, an influential Washington think tank learned what can happen when a tech giant that shapes public policy debates with its enormous wealth is criticized.

• • •

But not long after one of New America's scholars [posted a statement](#) on the think tank's website praising the European Union's penalty against Google, Mr. Schmidt, who had been chairman of New America until 2016, communicated his displeasure with the statement to the group's president, Anne-Marie Slaughter, according to the scholar.

• • •

Ms. Slaughter told Mr. Lynn that “the time has come for Open Markets and New America to part ways,” according to an email from Ms. Slaughter to Mr. Lynn. The email suggested that the entire Open Markets team — nearly 10 full-time employees and unpaid fellows — would be exiled from New America.

# Automatic Summarization

POLITICS

## ***Google Critic Ousted From Think Tank Funded by the Tech Giant***

WASHINGTON — In the hours after European antitrust regulators levied a record [\\$2.7 billion fine](#) against Google in late June, an influential Washington think tank learned what can happen when a tech giant that shapes public policy debates with its enormous wealth is criticized.

• • •

But not long after one of New America's scholars [posted a statement](#) on the think tank's website praising the European Union's penalty against Google, Mr. Schmidt, who had been chairman of New America until 2016, communicated his displeasure with the statement to the group's president, Anne-Marie Slaughter, according to the scholar.

• • •

Ms. Slaughter told Mr. Lynn that “the time has come for Open Markets and New America to part ways,” according to an email from Ms. Slaughter to Mr. Lynn. The email suggested that the entire Open Markets team — nearly 10 full-time employees and unpaid fellows — would be exiled from New America.

One of New America's writers posted a statement critical of Google. Eric Schmidt, Google's CEO, was displeased.

The writer and his team were dismissed.

# Automatic Summarization

POLITICS

## ***Google Critic Ousted From Think Tank Funded by the Tech Giant***

WASHINGTON — In the hours after European antitrust regulators levied a record [\\$2.7 billion fine](#) against Google in late June, an influential Washington think tank learned what can happen when a tech giant that shapes public policy debates with its enormous wealth is criticized.

• • •

But not long after one of New America's scholars [posted a statement](#) on the think tank's website praising the European Union's penalty against Google, Mr. Schmidt, who had been chairman of New America until 2016, communicated his displeasure with the statement to the group's president,

Anne-Marie Slaughter, according to the scholar.

• • •

Ms. Slaughter told Mr. Lynn that “the time has come for Open Markets and New America to part ways,” according to an email from Ms. Slaughter to Mr. Lynn. The email suggested that the entire Open Markets team — nearly 10 full-time employees and unpaid fellows — would be exiled from New America.

compress  
text

One of New America's writers posted a statement critical of Google. Eric Schmidt, Google's CEO, was displeased.

The writer and his team were dismissed.

# Automatic Summarization

POLITICS

## ***Google Critic Ousted From Think Tank Funded by the Tech Giant***

WASHINGTON — In the hours after European antitrust regulators levied a record [\\$2.7 billion fine](#) against Google in late June, an influential Washington think tank learned what can happen when a tech giant that shapes public policy debates with its enormous wealth is criticized.

•••

But not long after one of New America's scholars [posted a statement](#) on the think tank's website praising the European Union's penalty against Google, Mr. Schmidt, who had been chairman of New America until 2016, communicated his displeasure with the statement to the group's president,

Anne-Marie Slaughter, according to the scholar.

•••

Ms. Slaughter told Mr. Lynn that “the time has come for Open Markets and New America to part ways,” according to an email from Ms. Slaughter to Mr. Lynn. The email suggested that the entire Open Markets team — nearly 10 full-time employees and unpaid fellows — would be exiled from New America.

compress  
text

provide missing  
context

One of New America's writers posted a statement critical of Google. Eric Schmidt, [Google's CEO](#), was displeased.

The writer and his team were dismissed.

# Automatic Summarization

POLITICS

## ***Google Critic Ousted From Think Tank Funded by the Tech Giant***

WASHINGTON — In the hours after European antitrust regulators levied a record [\\$2.7 billion fine](#) against Google in late June, an influential Washington think tank learned what can happen when a tech giant that shapes public policy debates with its enormous wealth is criticized.

•••

But not long after one of New America's scholars [posted a statement](#) on the think tank's website praising the European Union's penalty against Google, Mr. Schmidt, who had been chairman of New America until 2016, communicated his displeasure with the statement to the group's president,

Anne-Marie Slaughter, according to the scholar.

•••

Ms. Slaughter told Mr. Lynn that “the time has come for Open Markets and New America to part ways,” according to an email from Ms. Slaughter to Mr. Lynn. The email suggested that the entire Open Markets team — nearly 10 full-time employees and unpaid fellows — would be [exiled](#) from New America.

compress  
text

provide missing  
context

One of New America's writers posted a statement critical of Google. Eric Schmidt, [Google's CEO](#), was displeased.

The writer and his team were dismissed.

paraphrase to provide clarity

# Machine Translation

---



< 2/8

特朗普偕家人在白宫阳台观看百年一遇日全食

>

People's Daily, August 30, 2017

# Machine Translation



A photograph of a woman with blonde hair, wearing a black sleeveless dress and 3D glasses, looking upwards. She is standing next to a man in a dark suit. The background is a light-colored wall.

Translate

English French Spanish Chinese - detected ▾

特朗普偕家人在白宫阳台观看百年一遇日全食

2/8 特朗普偕家人在白宫阳台观看百年一遇日全食

The image shows a woman wearing 3D glasses, looking up at a screen. A translation interface is overlaid on the image. The text "Translate" is in red. Below it is a language selection bar with "English", "French", "Spanish", and "Chinese - detected" (with a dropdown arrow). To the right is a double-headed arrow icon. Below the bar is a box containing the Chinese text "特朗普偕家人在白宫阳台观看百年一遇日全食", which is highlighted with an orange border. At the bottom left, there is a navigation indicator "2/8" and a caption in Chinese: "特朗普偕家人在白宫阳台观看百年一遇日全食".

People's Daily, August 30, 2017

# Machine Translation



Trump Pope family watch a hundred years a year in the White House balcony

# Machine Translation



Trump Pope family watch a hundred years a year in the White House balcony

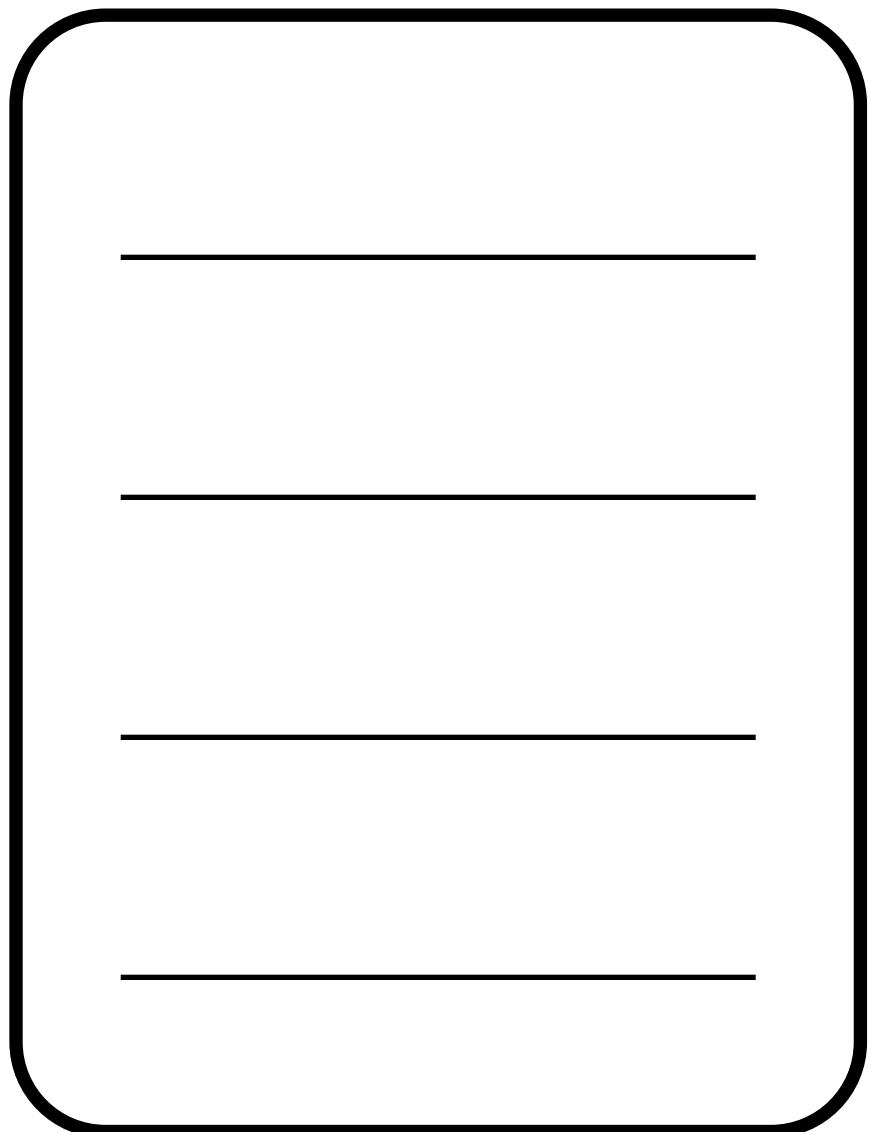
# NLP Analysis Pipeline

---

# NLP Analysis Pipeline

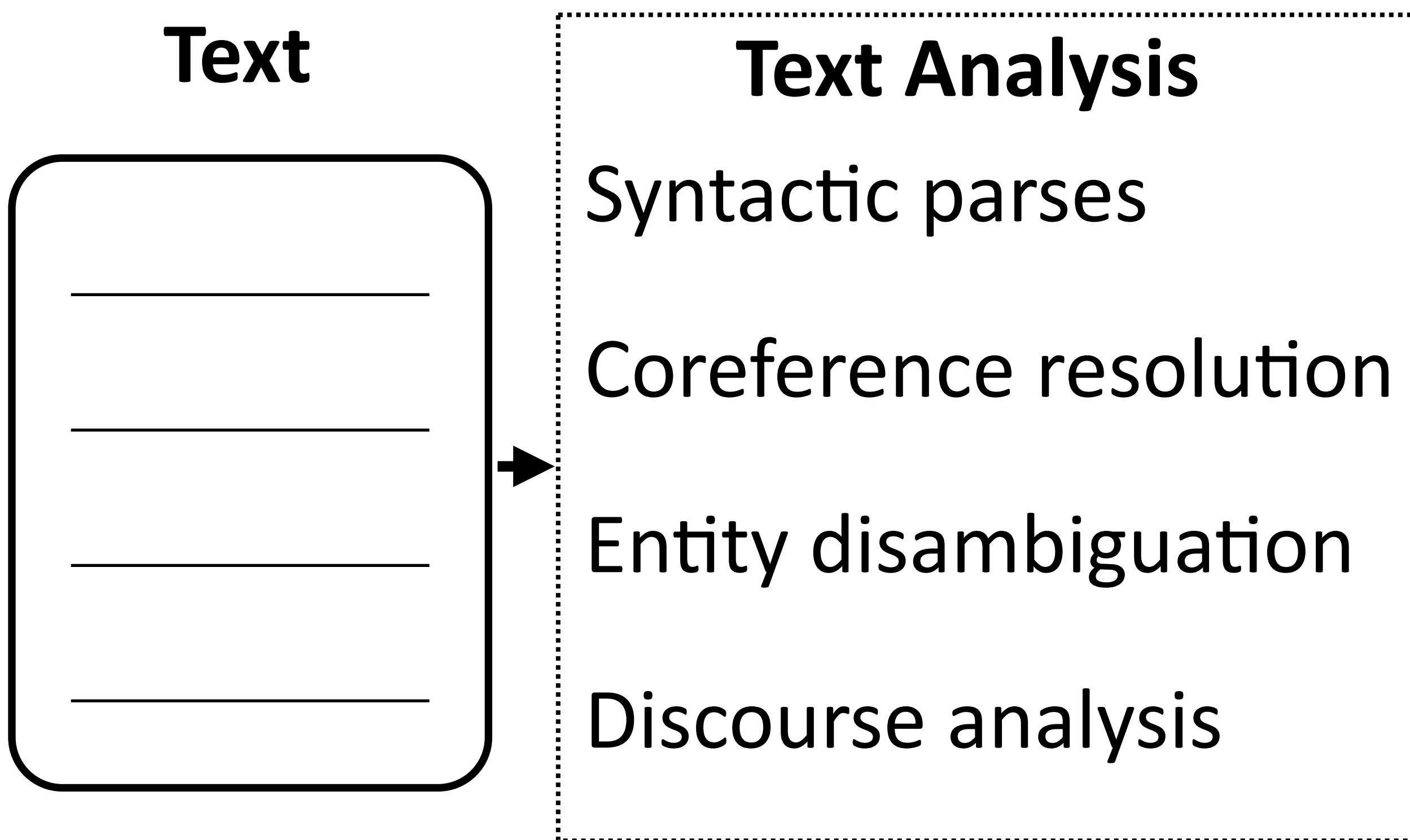
---

**Text**

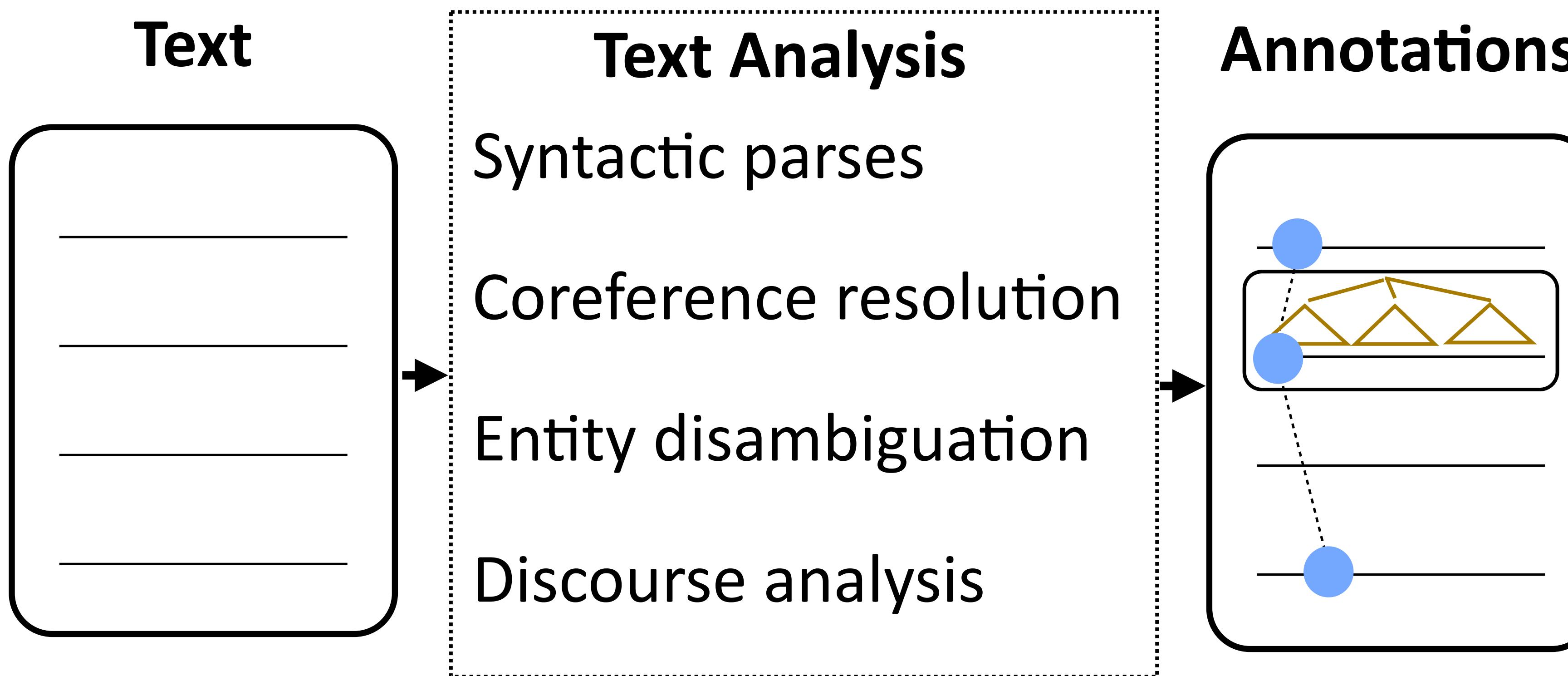


# NLP Analysis Pipeline

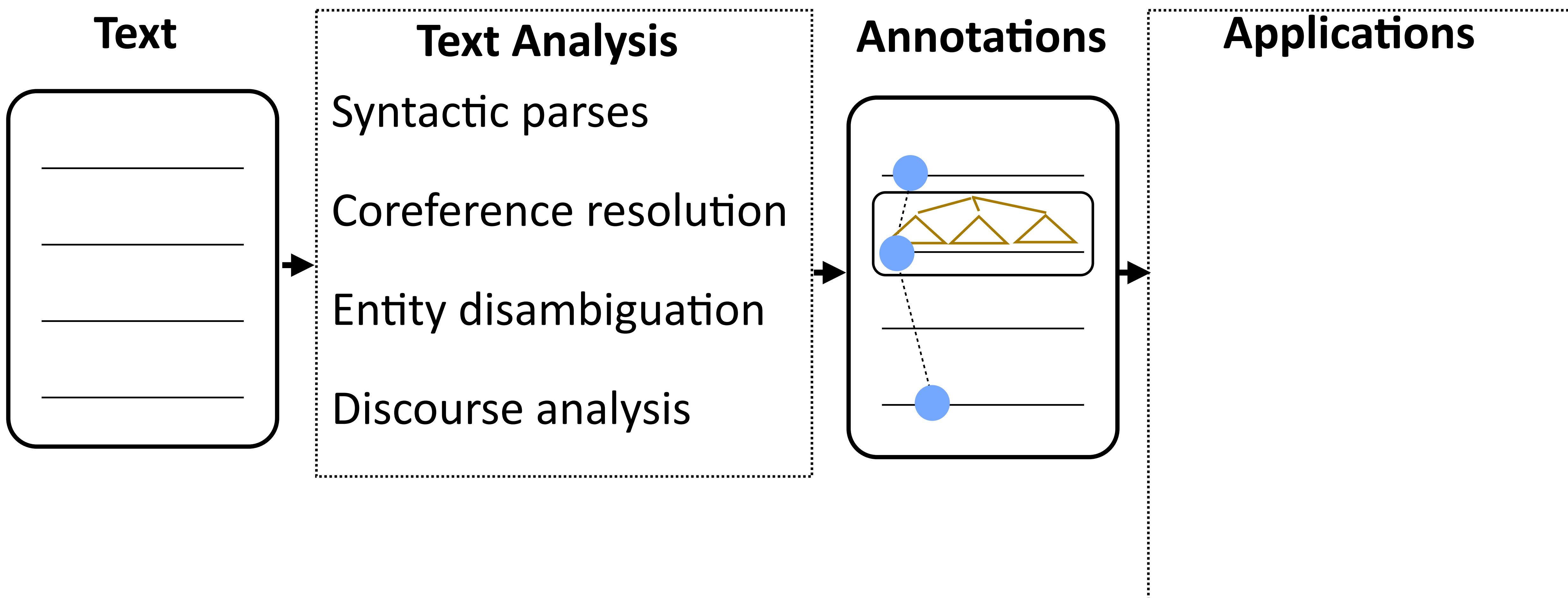
---



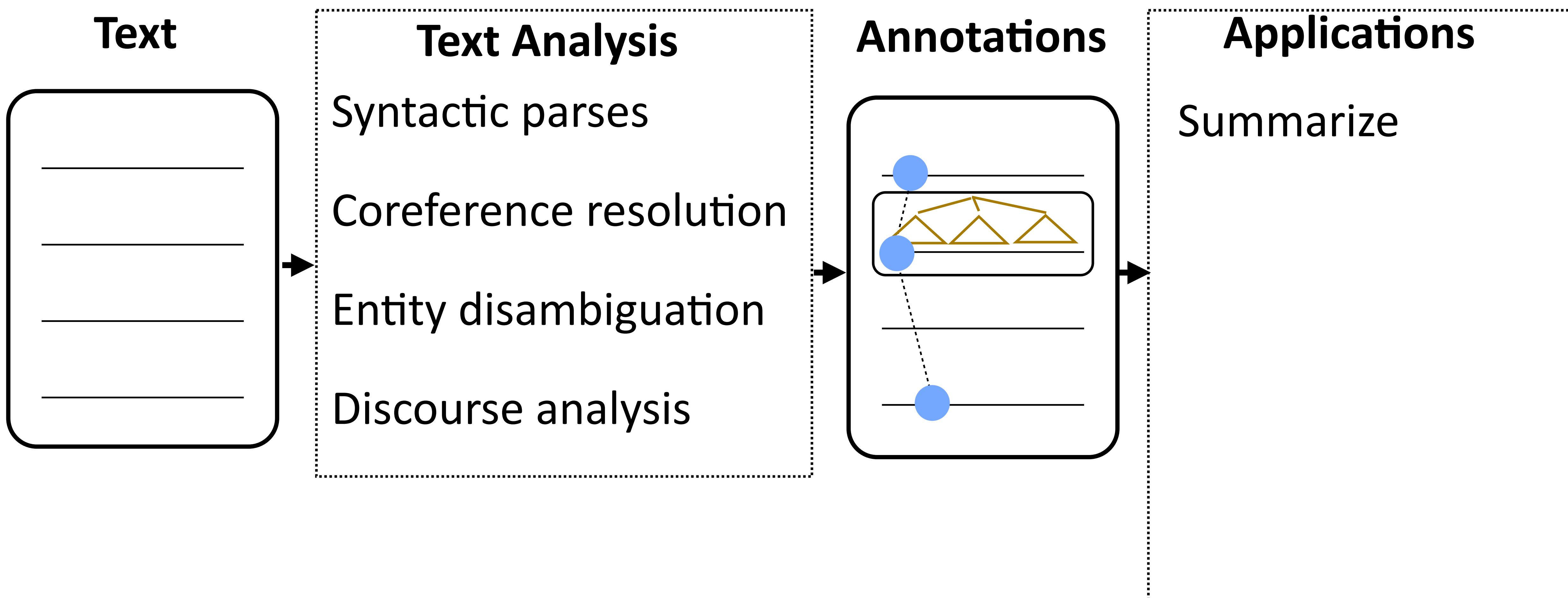
# NLP Analysis Pipeline



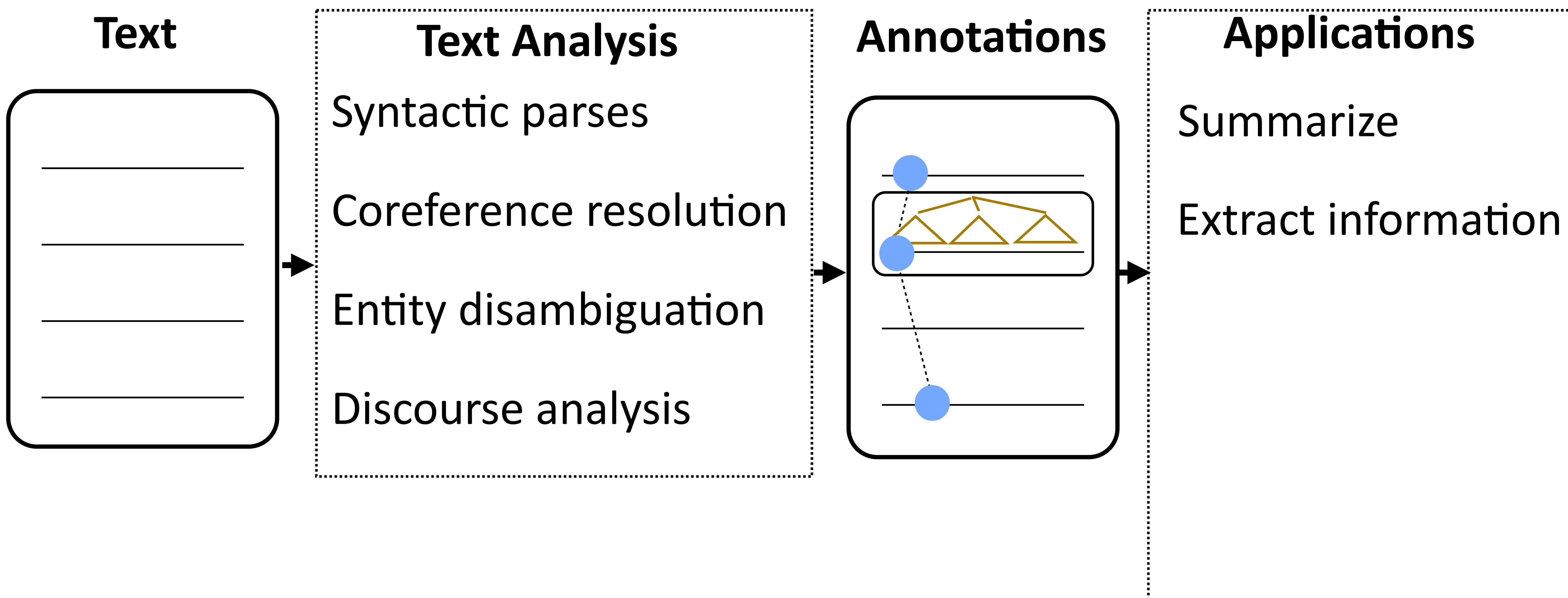
# NLP Analysis Pipeline



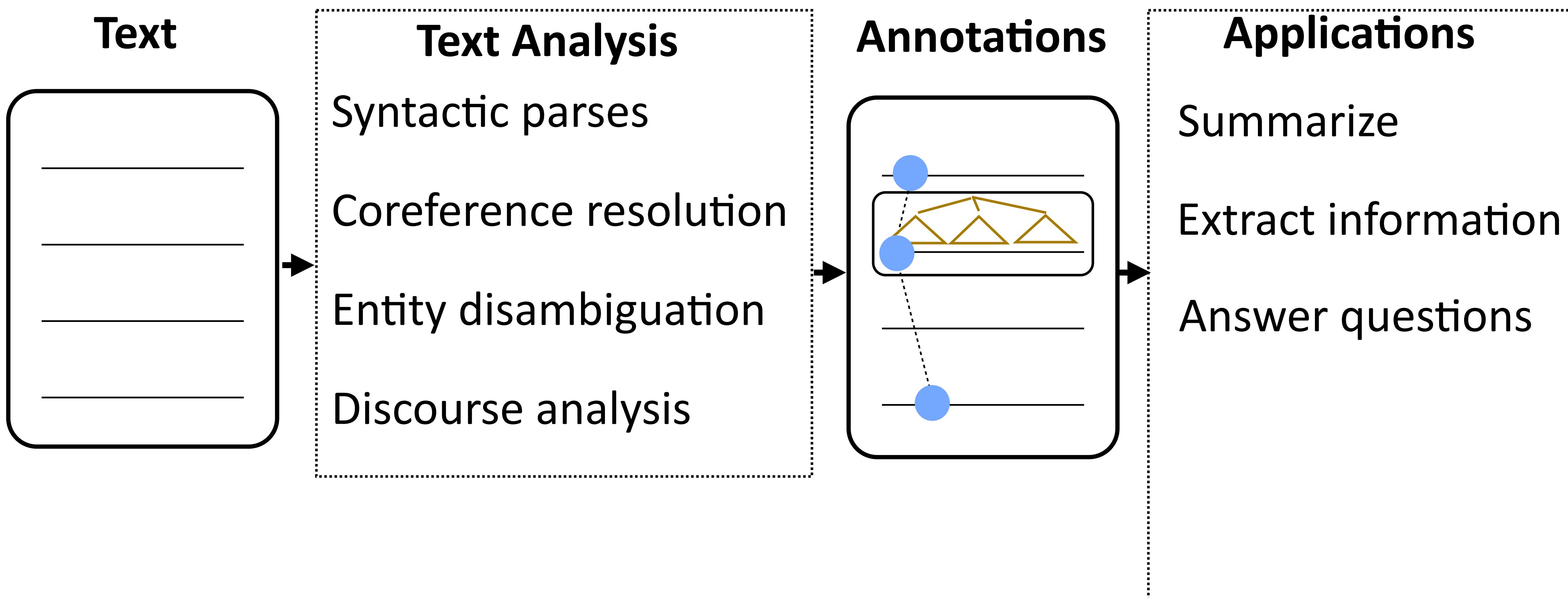
# NLP Analysis Pipeline



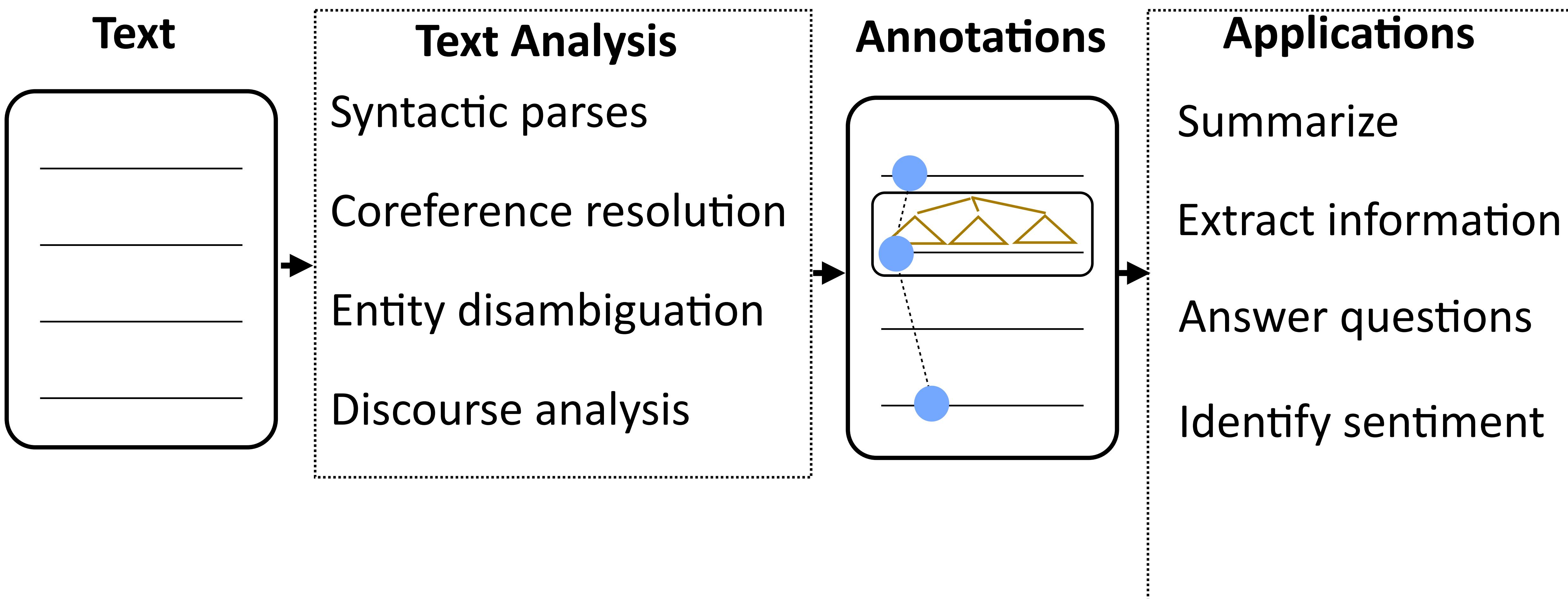
# NLP Analysis Pipeline



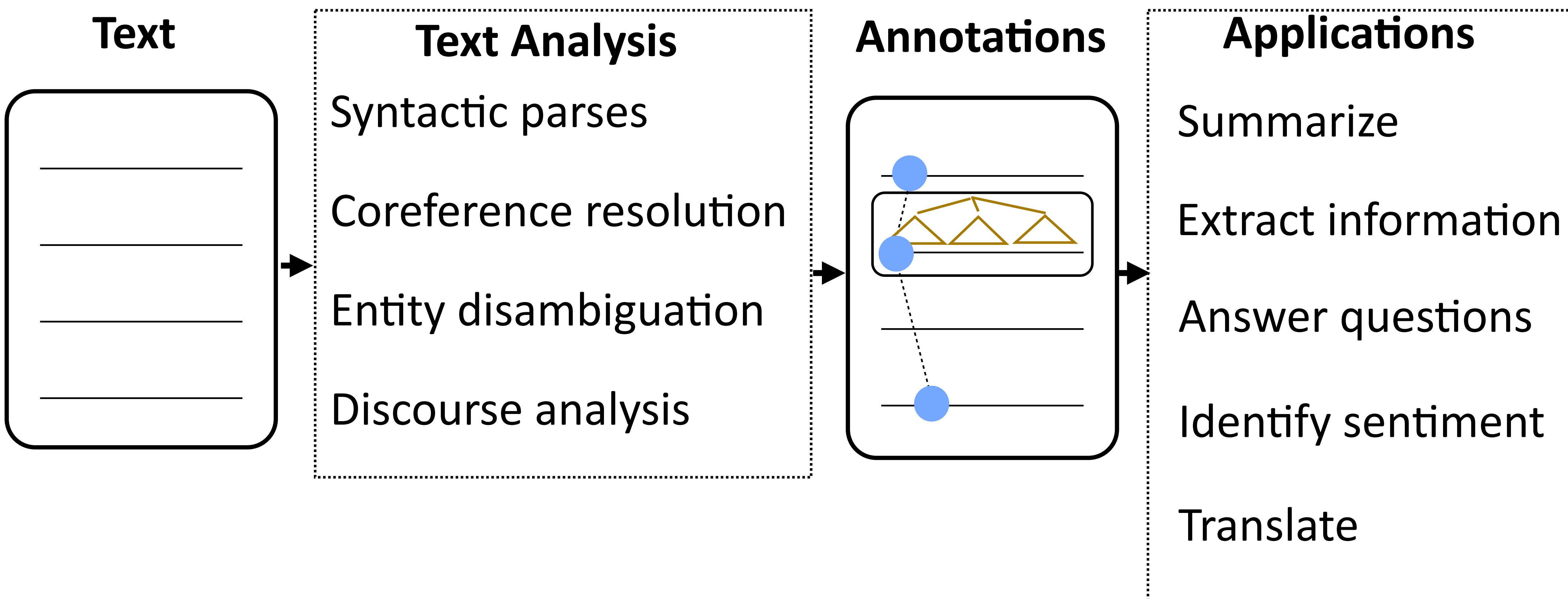
# NLP Analysis Pipeline



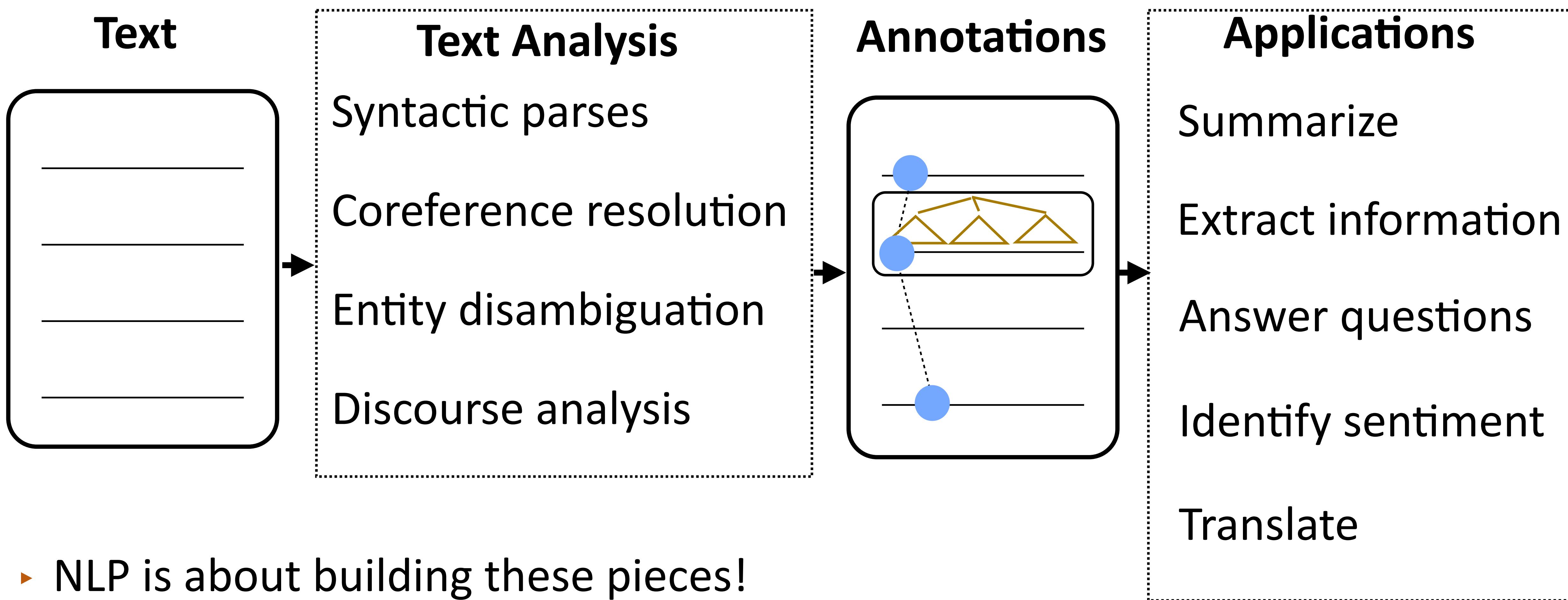
# NLP Analysis Pipeline



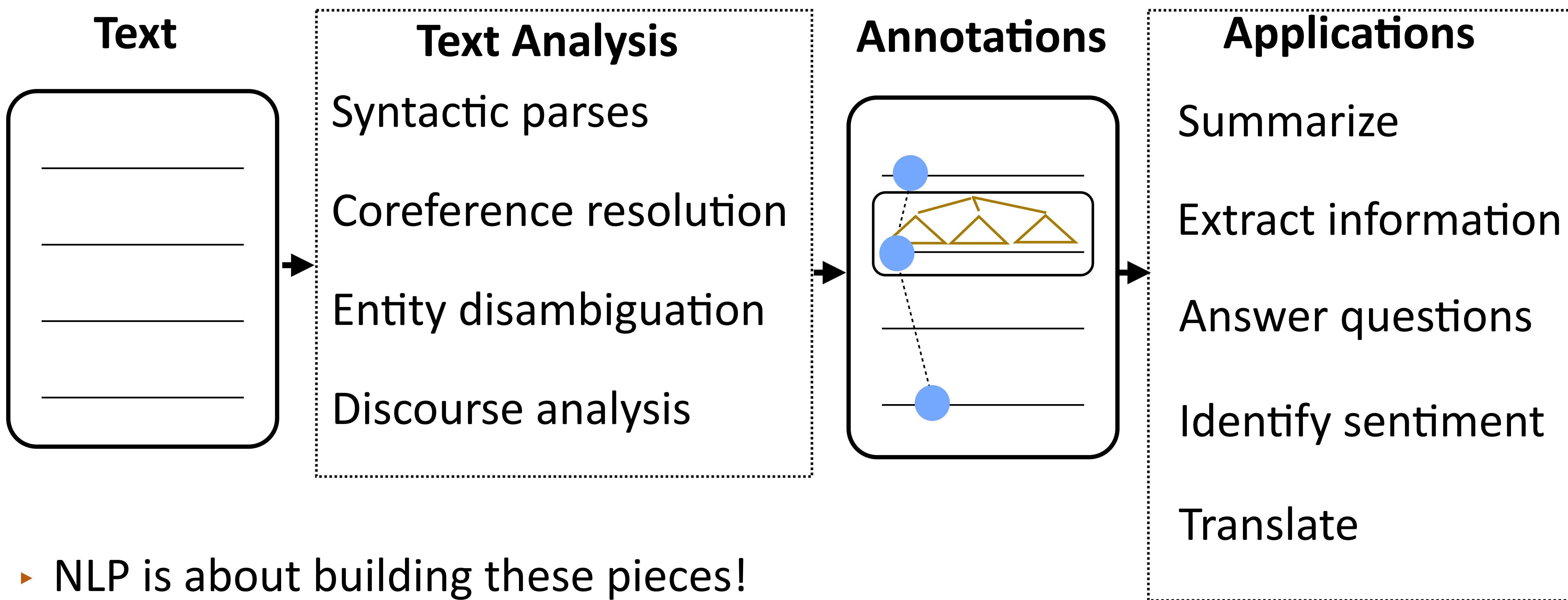
# NLP Analysis Pipeline



# NLP Analysis Pipeline



# NLP Analysis Pipeline

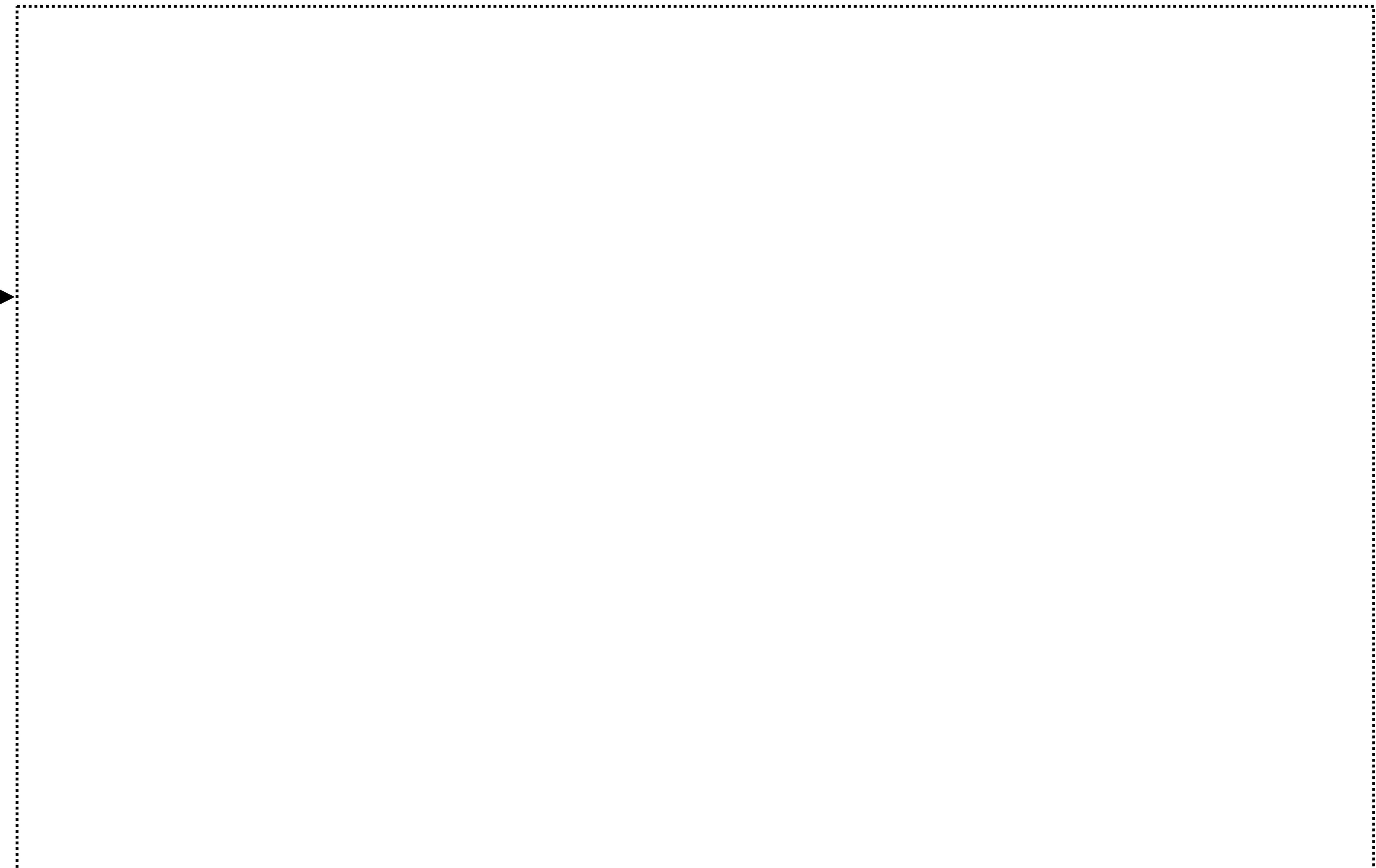
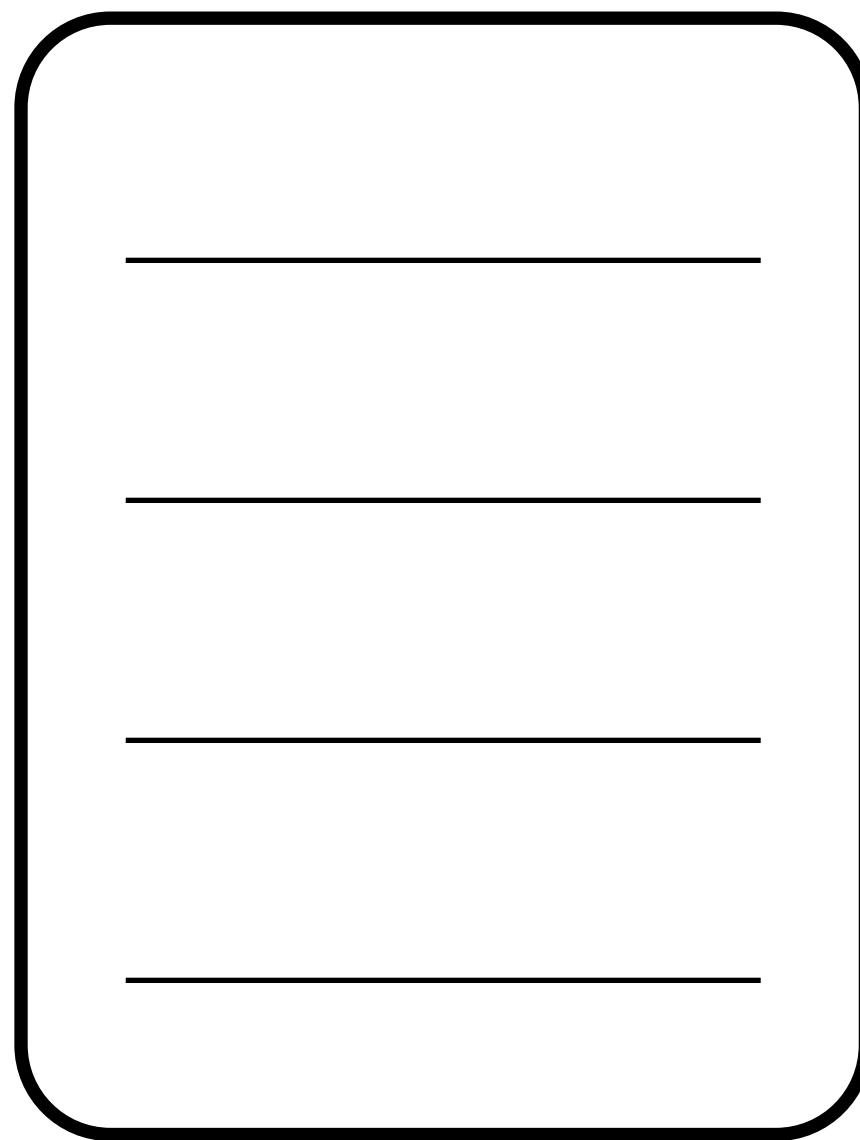


- ▶ NLP is about building these pieces!
- ▶ All of these components are modeled with statistical approaches trained with machine learning

# How do we represent language?

---

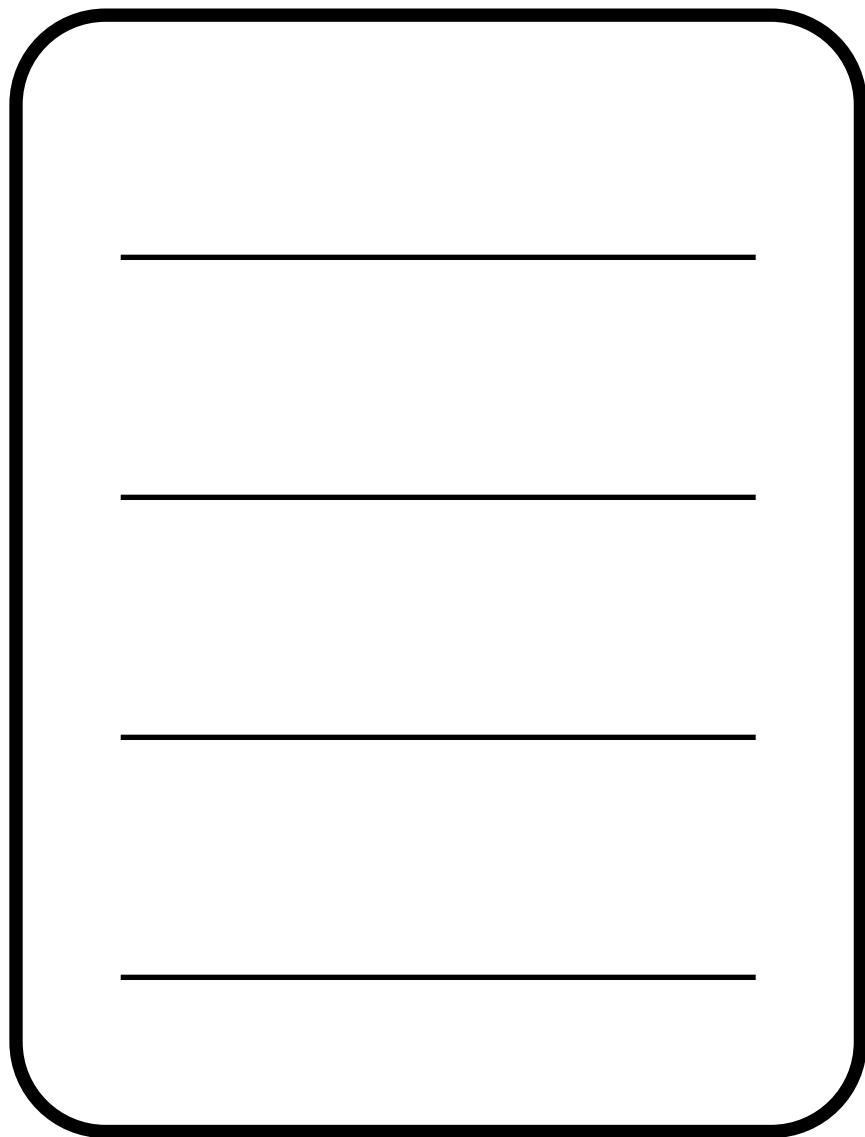
**Text**



# How do we represent language?

---

**Text**

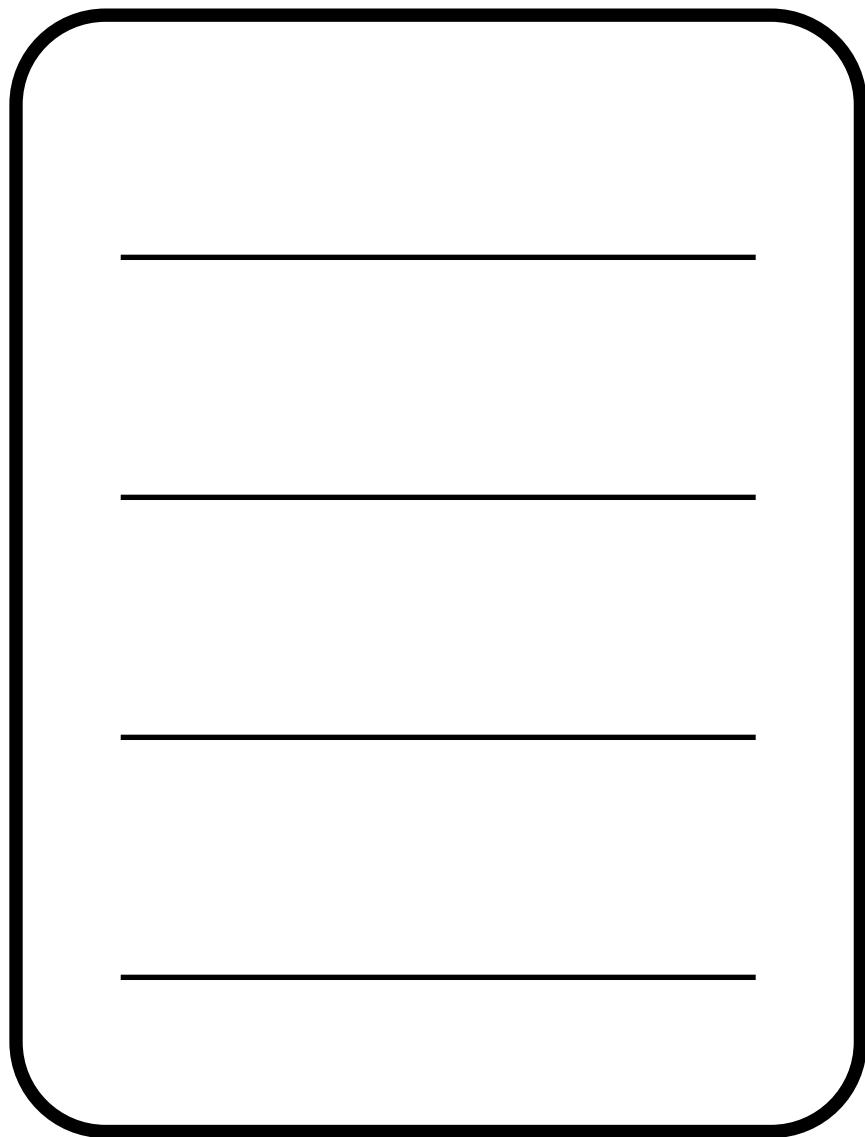


**Labels**



# How do we represent language?

**Text**

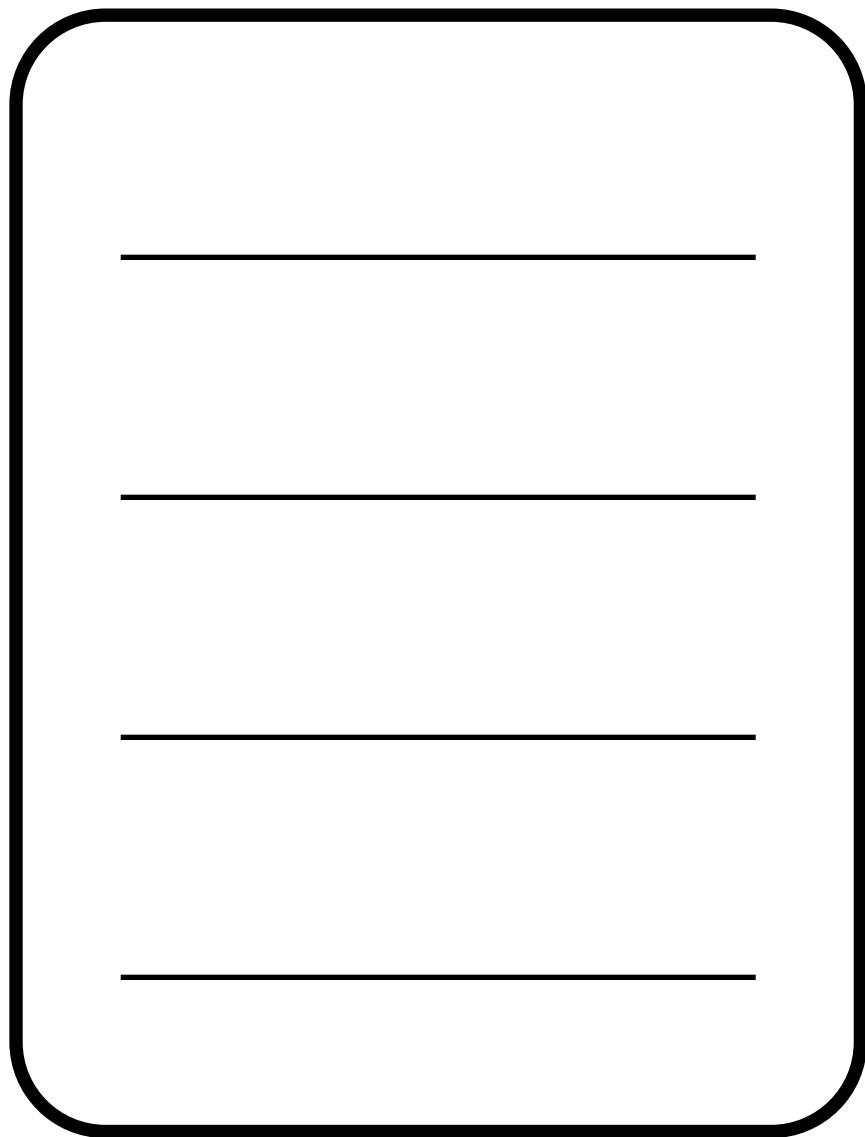


**Labels**

*the movie was good* +

# How do we represent language?

Text



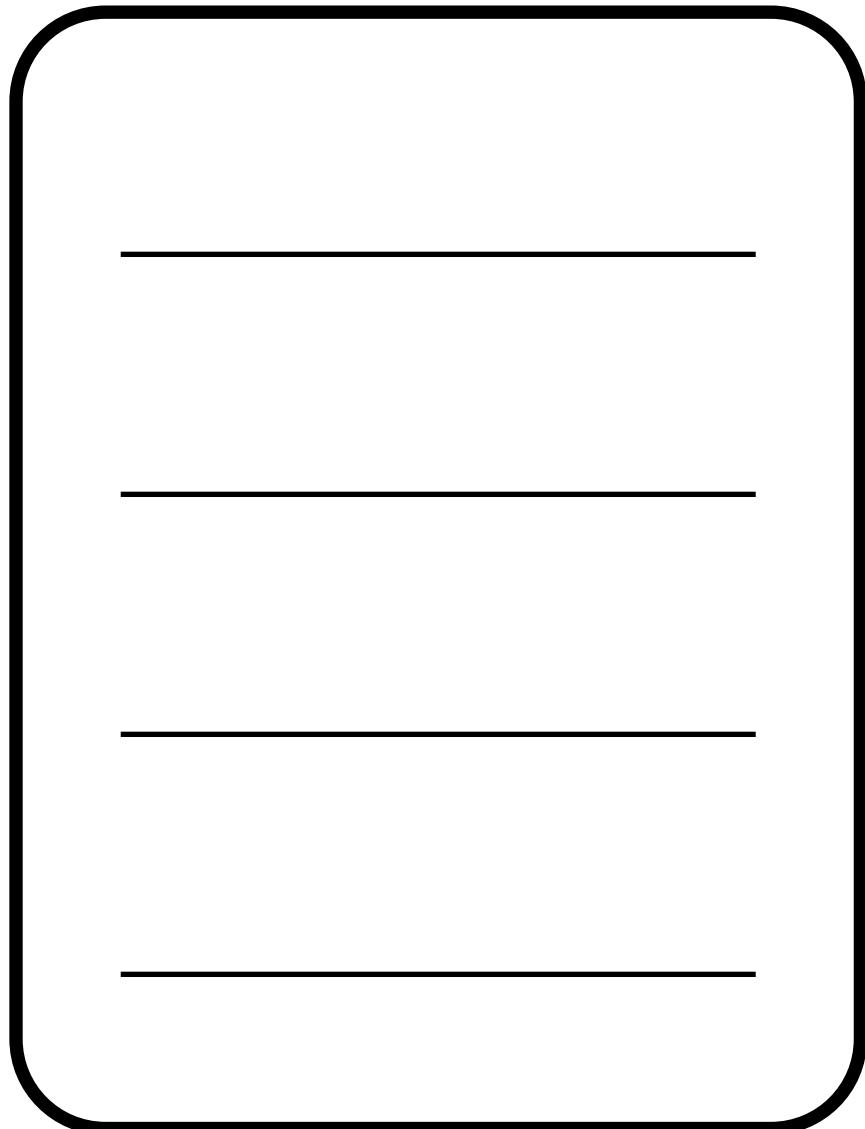
Labels

*the movie was good* +

*Beyoncé had one of the best videos of all time* subjective

# How do we represent language?

Text



Labels

*the movie was good* +

*Beyoncé had one of the best videos of all time* subjective

Sequences/tags

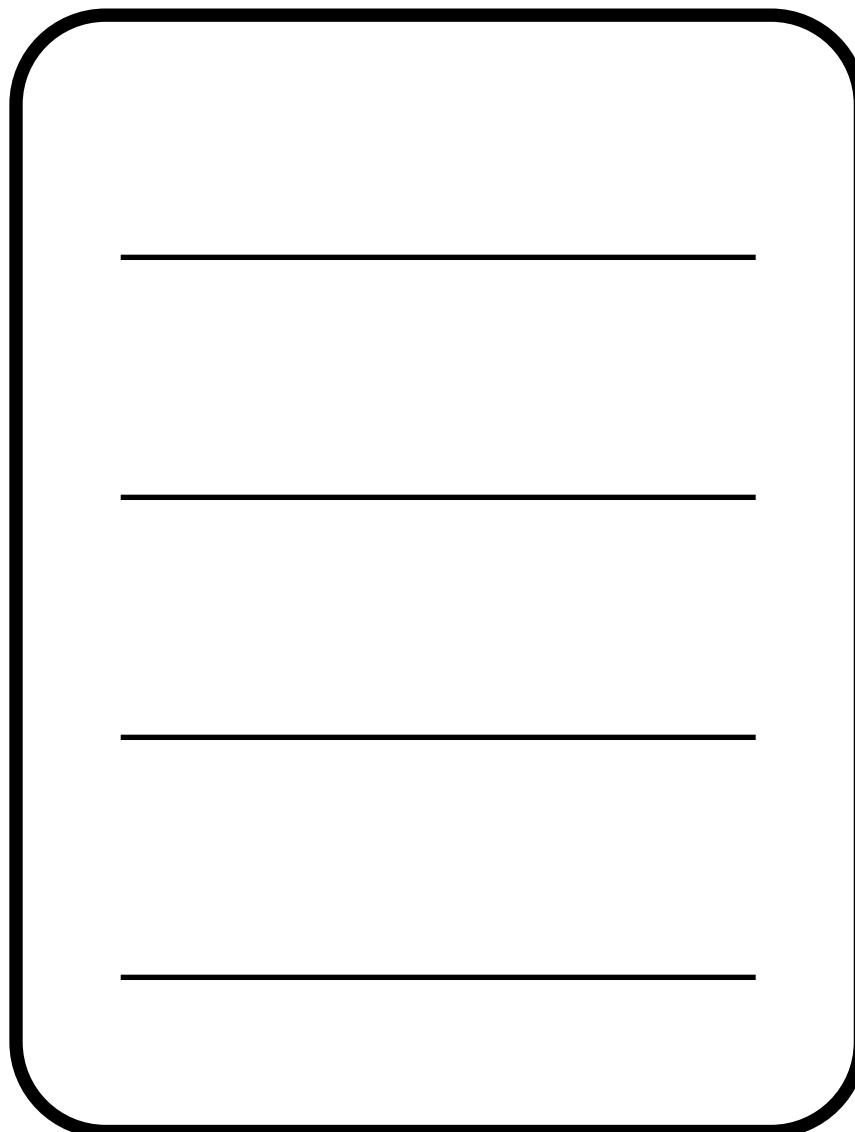
PERSON

*Tom Cruise stars in the new Mission Impossible film*

MOVIE

# How do we represent language?

Text



Labels

*the movie was good* +

*Beyoncé had one of the best videos of all time* subjective

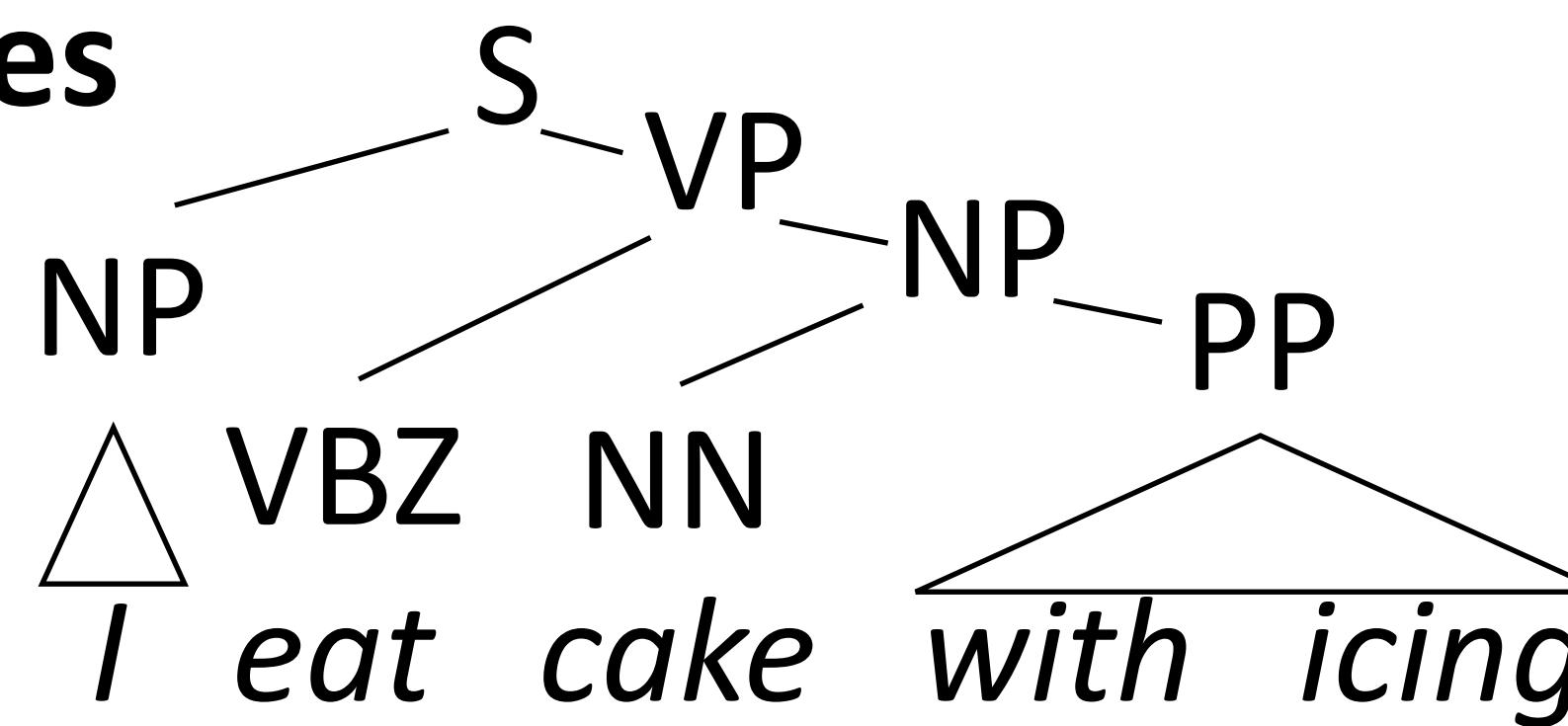
Sequences/tags

PERSON

*Tom Cruise stars in the new Mission Impossible film*

MOVIE

Trees



# How do we represent language?

Text

Labels

*the movie was good* +

*Beyoncé had one of the best videos of all time* subjective

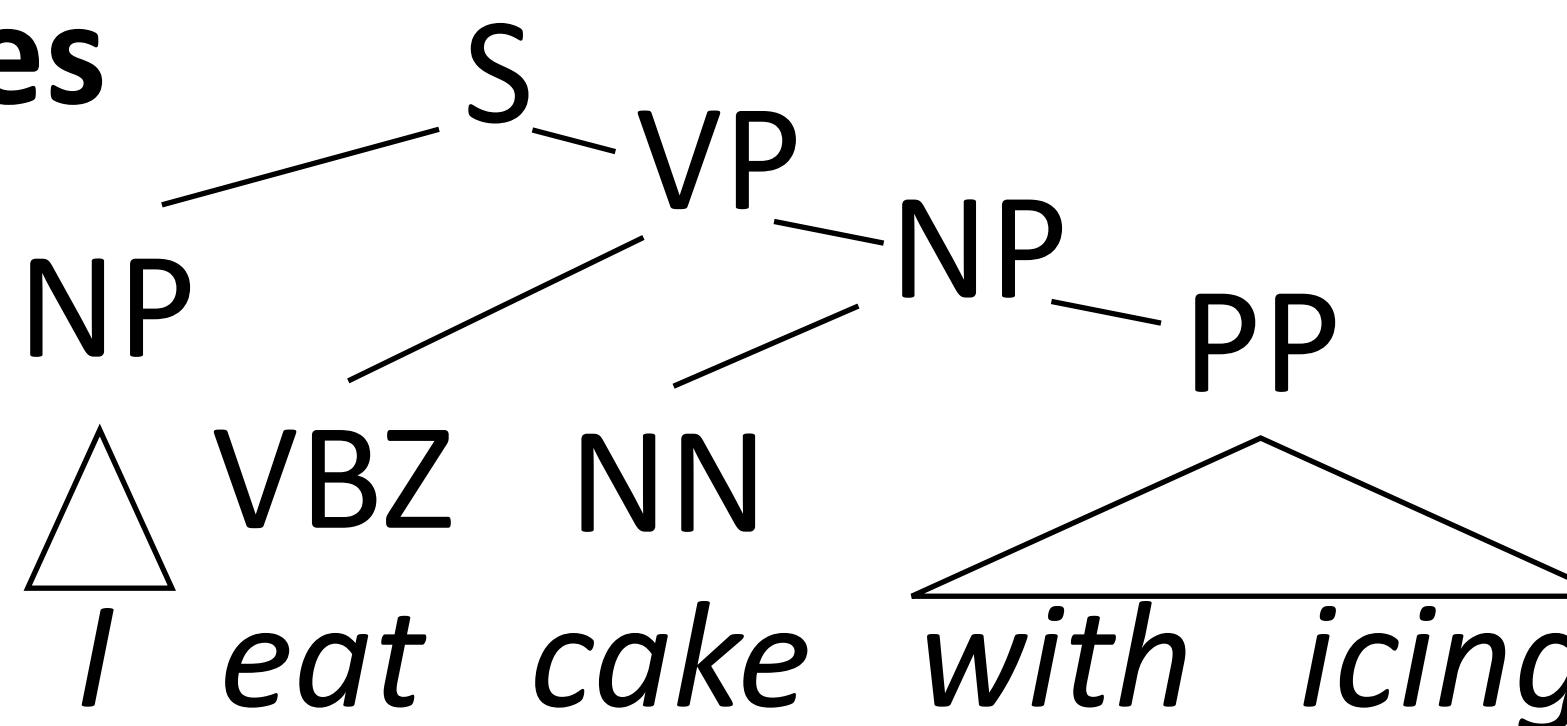
Sequences/tags

PERSON

*Tom Cruise stars in the new Mission Impossible film*

MOVIE

Trees

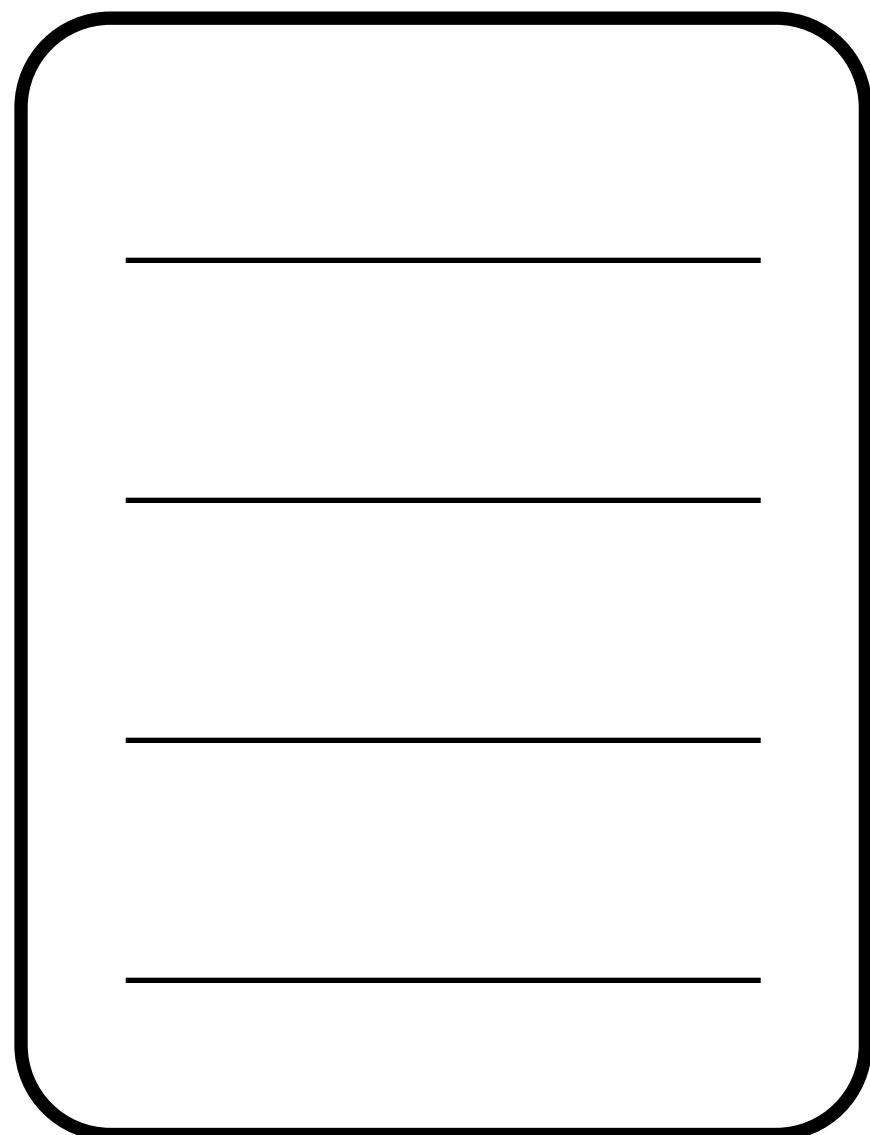


$\lambda x. \text{flight}(x) \wedge \text{dest}(x) = \text{Miami}$   
*flights to Miami*

# How do we use these representations?

---

**Text**



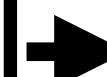
**Text Analysis**

Labels

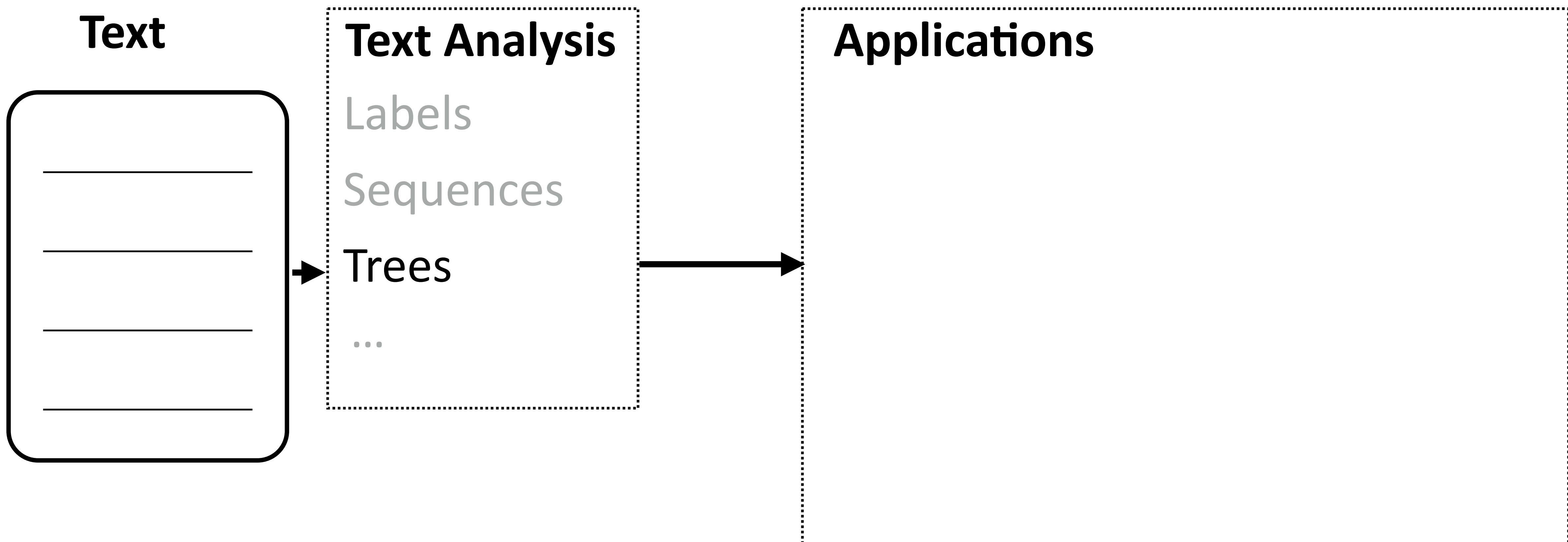
Sequences

Trees

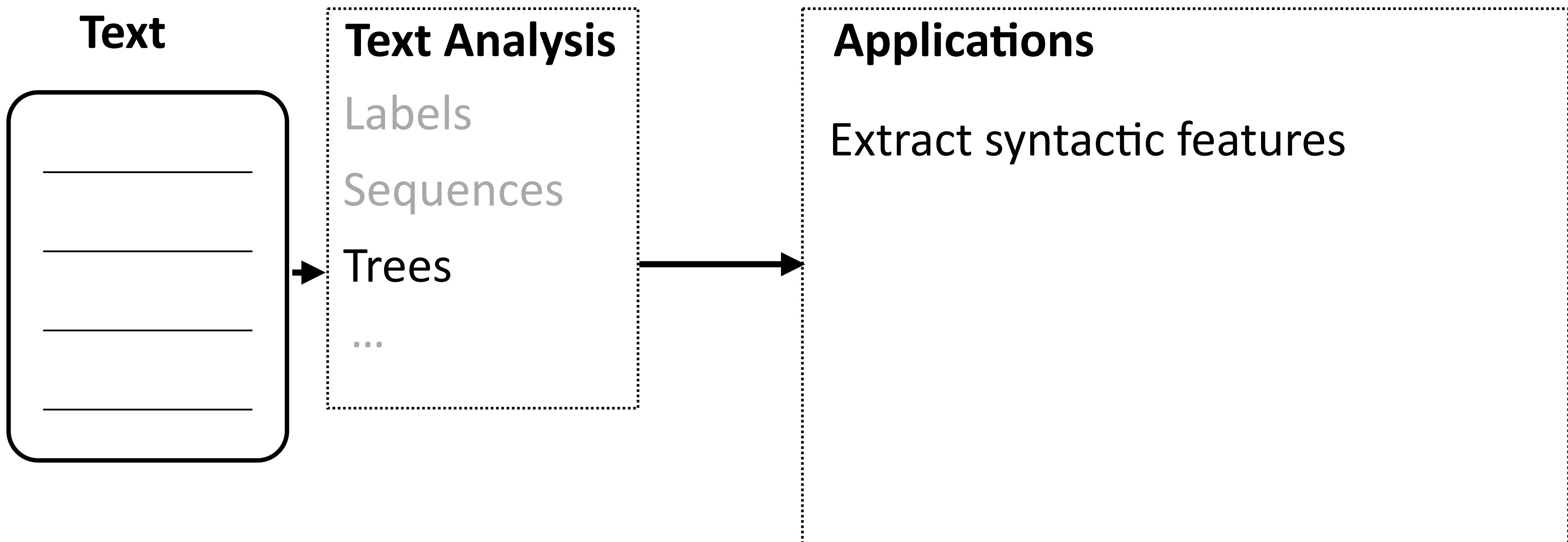
...



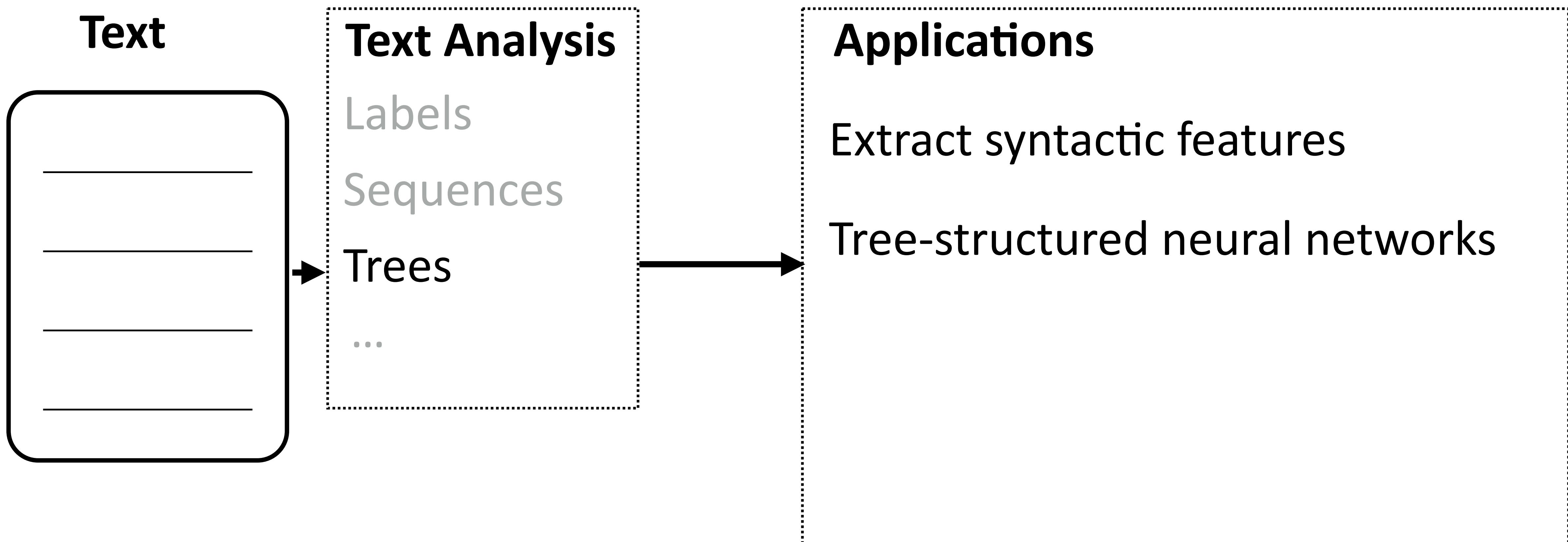
# How do we use these representations?



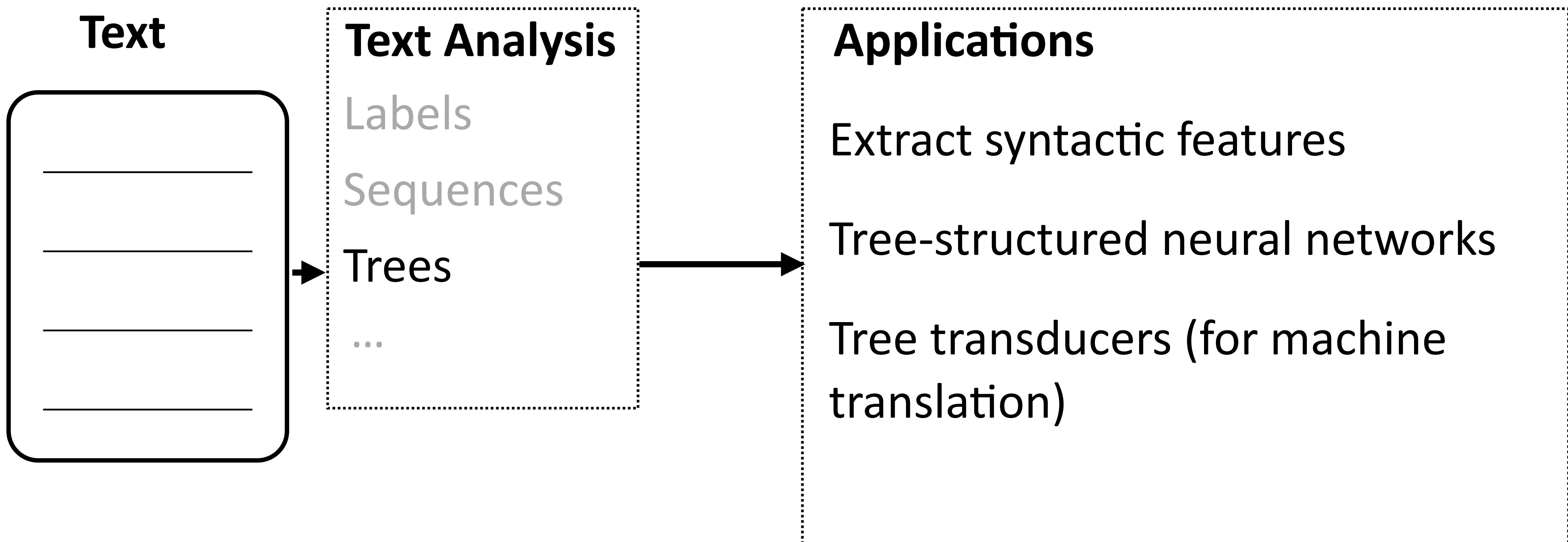
# How do we use these representations?



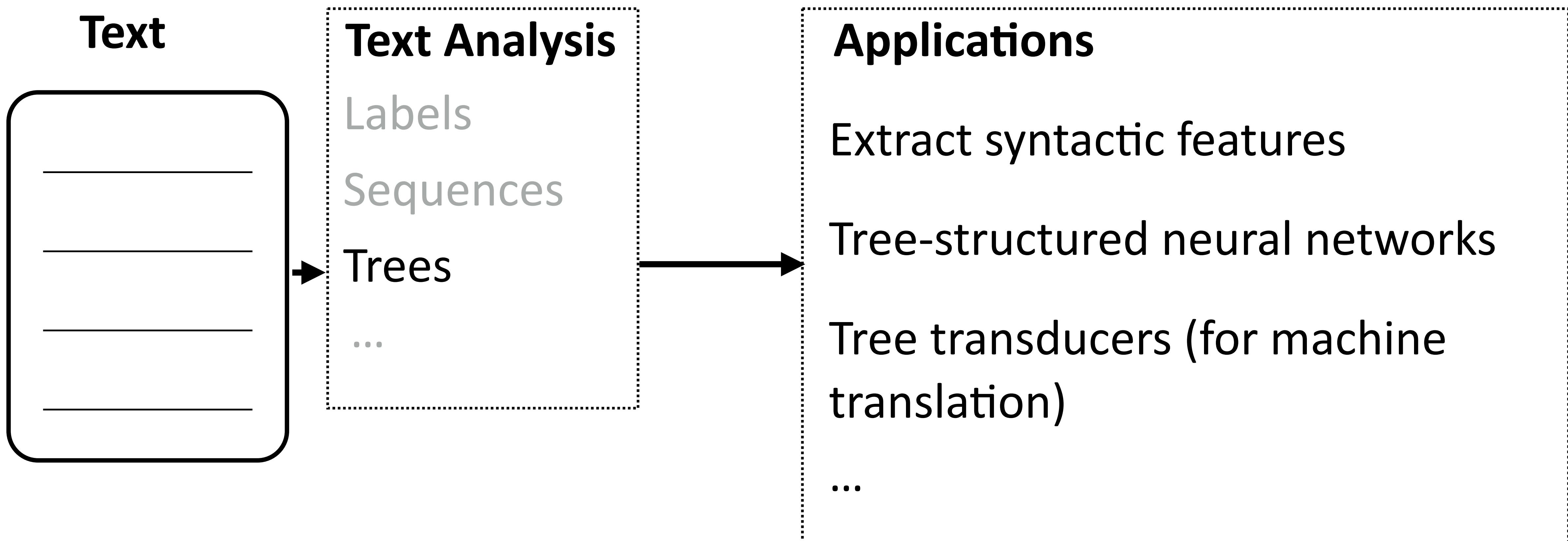
# How do we use these representations?



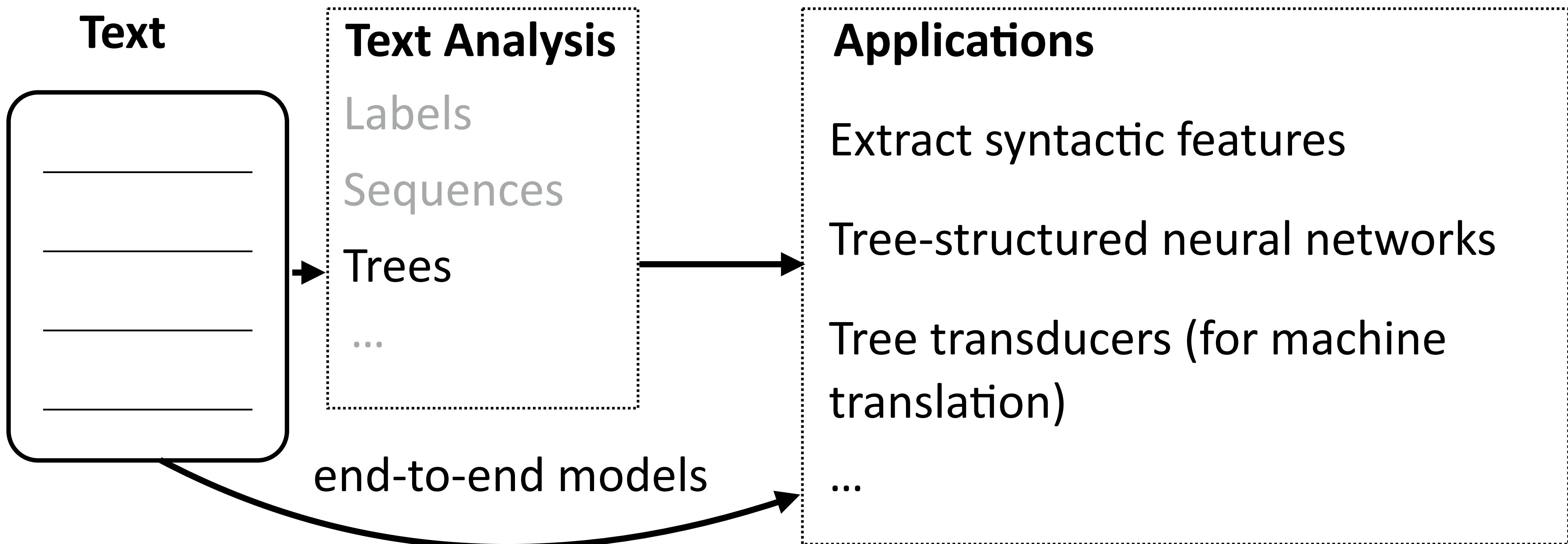
# How do we use these representations?



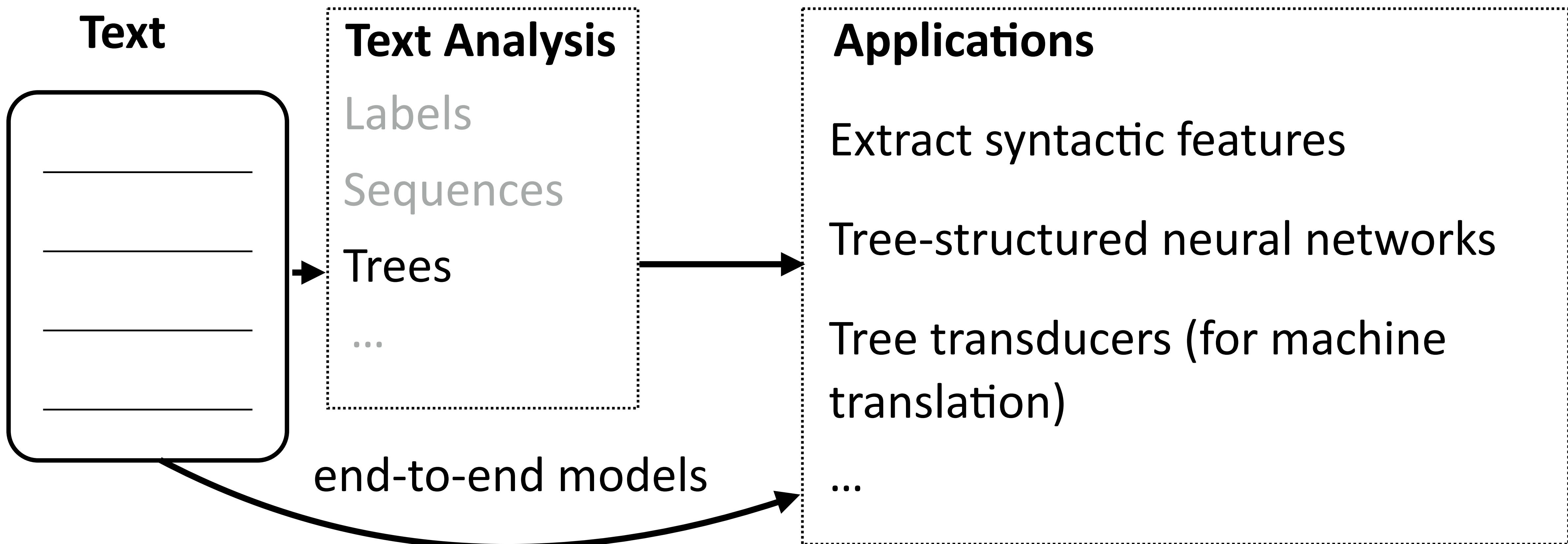
# How do we use these representations?



# How do we use these representations?

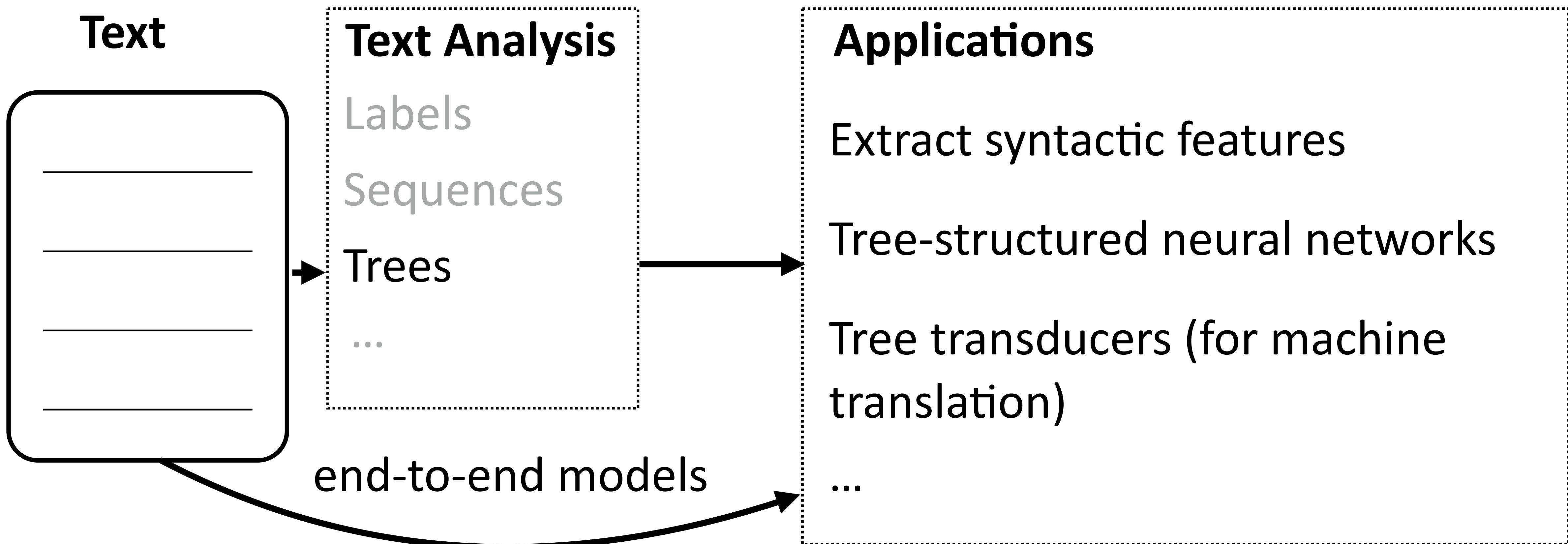


# How do we use these representations?



- ▶ Main question: What representations do we need for language? What do we want to know about it?

# How do we use these representations?



- ▶ Main question: What representations do we need for language? What do we want to know about it?
- ▶ Boils down to: what ambiguities do we need to resolve?

Why is language hard?  
(and how can we handle that?)

# Language is Ambiguous!

---

- ▶ Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

# Language is Ambiguous!

---

- ▶ Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

The city council refused the demonstrators a permit because they \_\_\_\_\_ violence

# Language is Ambiguous!

---

- ▶ Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

The city council refused the demonstrators a permit because they \_\_\_\_\_ violence

# Language is Ambiguous!

---

- ▶ Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

The city council refused the demonstrators a permit because they \_\_\_\_\_ violence

they advocated

# Language is Ambiguous!

---

- ▶ Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

The city council refused the demonstrators a permit because they \_\_\_\_\_ violence

they advocated

# Language is Ambiguous!

---

- ▶ Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

The city council refused the demonstrators a permit because they \_\_\_\_\_ violence

they advocated

they feared

# Language is Ambiguous!

---

- ▶ Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

The city council refused the demonstrators a permit because they \_\_\_\_\_ violence

they advocated

they feared

# Language is Ambiguous!

---

- ▶ Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

The city council refused the demonstrators a permit because they \_\_\_\_\_ violence

they advocated

they feared

- ▶ This is so complicated that it's an AI challenge problem! (AI-complete)

# Language is Ambiguous!

---

- ▶ Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

The city council refused the demonstrators a permit because they \_\_\_\_\_ violence

they advocated

they feared

- ▶ This is so complicated that it's an AI challenge problem! (AI-complete)
- ▶ Referential/semantic ambiguity

# Language is Ambiguous!

---

# Language is Ambiguous!

---

- ▶ Ambiguous News Headlines:

# Language is Ambiguous!

---

- ▶ Ambiguous News Headlines:
  - ▶ Teacher Strikes Idle Kids

# Language is Ambiguous!

---

- ▶ Ambiguous News Headlines:
  - ▶ Teacher Strikes Idle Kids
  - ▶ Hospitals Sued by 7 Foot Doctors

# Language is Ambiguous!

---

- ▶ Ambiguous News Headlines:
  - ▶ Teacher Strikes Idle Kids
  - ▶ Hospitals Sued by 7 Foot Doctors
  - ▶ Ban on Nude Dancing on Governor's Desk

# Language is Ambiguous!

---

- ▶ Ambiguous News Headlines:
  - ▶ Teacher Strikes Idle Kids
  - ▶ Hospitals Sued by 7 Foot Doctors
  - ▶ Ban on Nude Dancing on Governor's Desk
  - ▶ Iraqi Head Seeks Arms

# Language is Ambiguous!

---

- ▶ Ambiguous News Headlines:
  - ▶ Teacher Strikes Idle Kids
  - ▶ Hospitals Sued by 7 Foot Doctors
  - ▶ Ban on Nude Dancing on Governor's Desk
  - ▶ Iraqi Head Seeks Arms
  - ▶ Stolen Painting Found by Tree

# Language is Ambiguous!

---

- ▶ Ambiguous News Headlines:
  - ▶ Teacher Strikes Idle Kids
  - ▶ Hospitals Sued by 7 Foot Doctors
  - ▶ Ban on Nude Dancing on Governor's Desk
  - ▶ Iraqi Head Seeks Arms
  - ▶ Stolen Painting Found by Tree
  - ▶ Kids Make Nutritious Snacks

# Language is Ambiguous!

---

- ▶ Ambiguous News Headlines:
  - ▶ Teacher Strikes Idle Kids
  - ▶ Hospitals Sued by 7 Foot Doctors
  - ▶ Ban on Nude Dancing on Governor's Desk
  - ▶ Iraqi Head Seeks Arms
  - ▶ Stolen Painting Found by Tree
  - ▶ Kids Make Nutritious Snacks
  - ▶ Local HS Dropouts Cut in Half

# Language is Ambiguous!

---

- ▶ Ambiguous News Headlines:
  - ▶ Teacher Strikes Idle Kids
  - ▶ Hospitals Sued by 7 Foot Doctors
  - ▶ Ban on Nude Dancing on Governor's Desk
  - ▶ Iraqi Head Seeks Arms
  - ▶ Stolen Painting Found by Tree
  - ▶ Kids Make Nutritious Snacks
  - ▶ Local HS Dropouts Cut in Half
- ▶ Syntactic/semantic ambiguity: parsing needed to resolve these, but need context to figure out which parse is correct

# Language is Really Ambiguous!

---

- ▶ There aren't just one or two possibilities which are resolved pragmatically

# Language is Really Ambiguous!

---

- ▶ There aren't just one or two possibilities which are resolved pragmatically

*il fait vraiment beau* 

# Language is Really Ambiguous!

---

- ▶ There aren't just one or two possibilities which are resolved pragmatically

It is really nice out

*il fait vraiment beau*



# Language is Really Ambiguous!

---

- ▶ There aren't just one or two possibilities which are resolved pragmatically

*il fait vraiment beau*



It is really nice out

It's really nice

# Language is Really Ambiguous!

---

- ▶ There aren't just one or two possibilities which are resolved pragmatically

*il fait vraiment beau*



It is really nice out  
It's really nice  
The weather is beautiful

# Language is Really Ambiguous!

---

- ▶ There aren't just one or two possibilities which are resolved pragmatically

*il fait vraiment beau*



It is really nice out  
It's really nice  
The weather is beautiful  
It is really beautiful outside

# Language is Really Ambiguous!

---

- ▶ There aren't just one or two possibilities which are resolved pragmatically

*il fait vraiment beau*



It is really nice out  
It's really nice  
The weather is beautiful  
It is really beautiful outside  
**He makes truly beautiful**

# Language is Really Ambiguous!

---

- ▶ There aren't just one or two possibilities which are resolved pragmatically

*il fait vraiment beau*



It is really nice out  
It's really nice  
The weather is beautiful  
It is really beautiful outside  
**He makes truly beautiful**  
**He makes truly boyfriend**

# Language is Really Ambiguous!

---

- ▶ There aren't just one or two possibilities which are resolved pragmatically

*il fait vraiment beau*



- It is really nice out
- It's really nice
- The weather is beautiful
- It is really beautiful outside
- He makes truly beautiful
- He makes truly boyfriend
- It fact actually handsome

# Language is Really Ambiguous!

---

- ▶ There aren't just one or two possibilities which are resolved pragmatically

*il fait vraiment beau*



It is really nice out  
It's really nice  
The weather is beautiful  
It is really beautiful outside  
**He makes truly beautiful**  
**He makes truly boyfriend**  
**It fact actually handsome**

- ▶ Combinatorially many possibilities, many you won't even register as ambiguities, but systems still have to resolve them

# What do we need to understand language?

---

- ▶ Lots of data!

|            |   |
|------------|---|
| SOURCE     | Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.                              |
| HUMAN      | That would be an interim solution which would make it possible to work towards a binding charter in the long term .                               |
| 1x DATA    | [this] [constituerait] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] [.] |
| 10x DATA   | [it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] [.]                      |
| 100x DATA  | [this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] [.]   |
| 1000x DATA | [that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.]  |

# What do we need to understand language?

---

- ▶ World knowledge: have access to information beyond the training data

# What do we need to understand language?

---

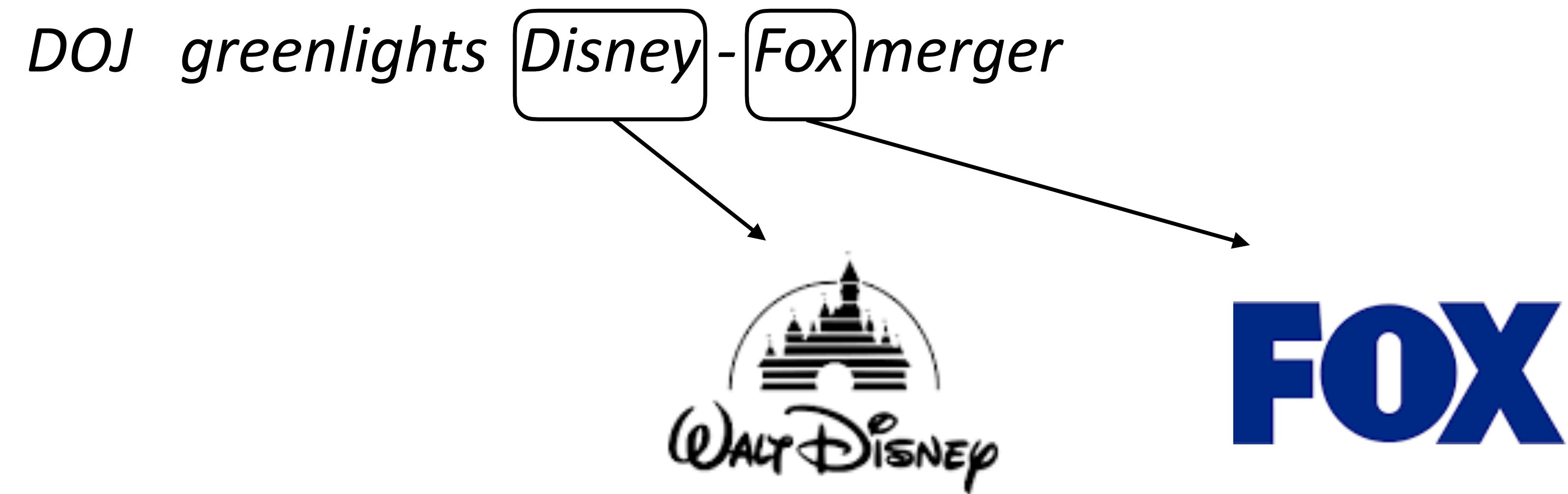
- ▶ World knowledge: have access to information beyond the training data

*DOJ greenlights Disney - Fox merger*

# What do we need to understand language?

---

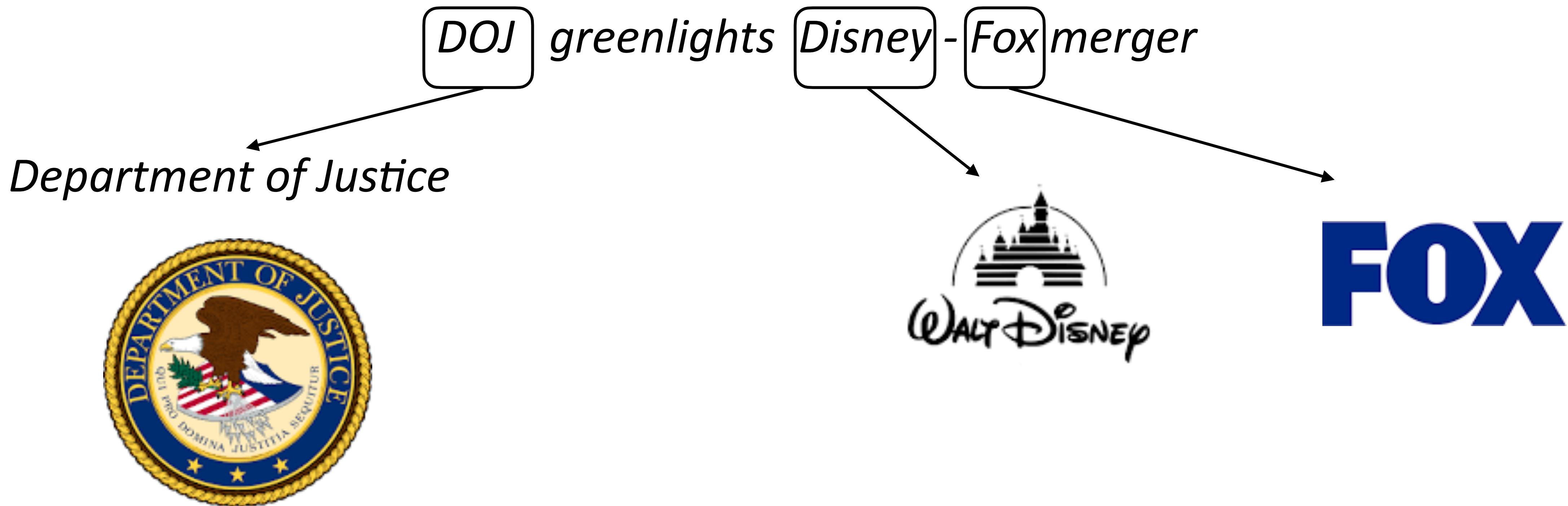
- ▶ World knowledge: have access to information beyond the training data



# What do we need to understand language?

---

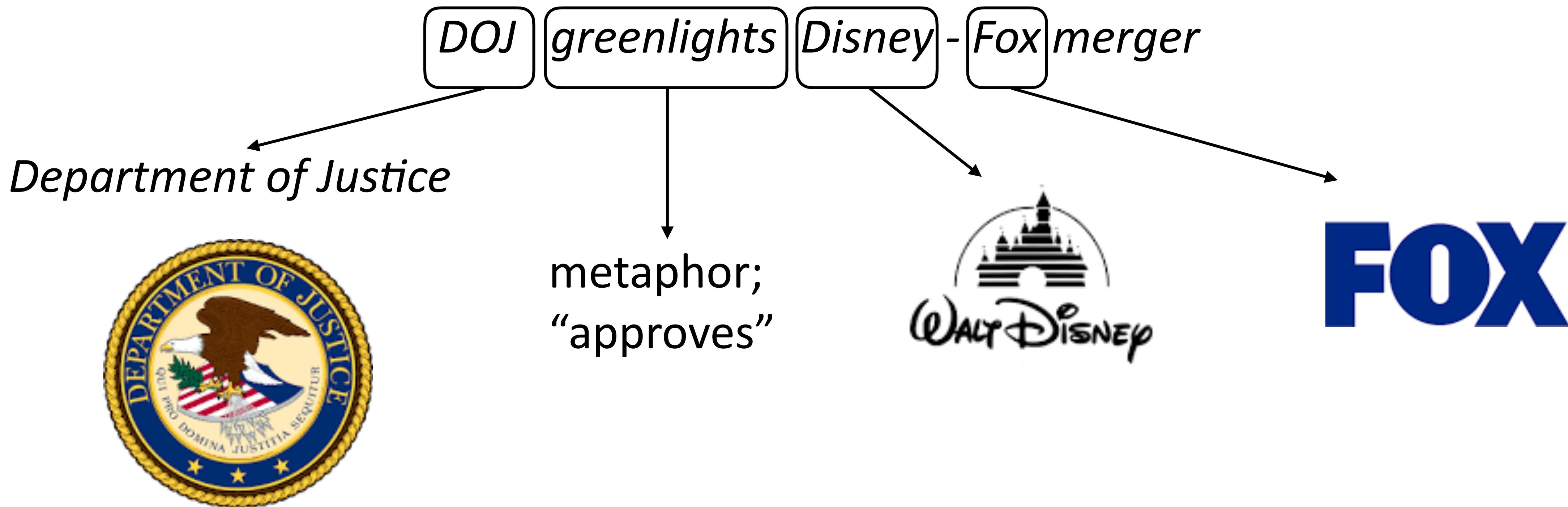
- ▶ World knowledge: have access to information beyond the training data



# What do we need to understand language?

---

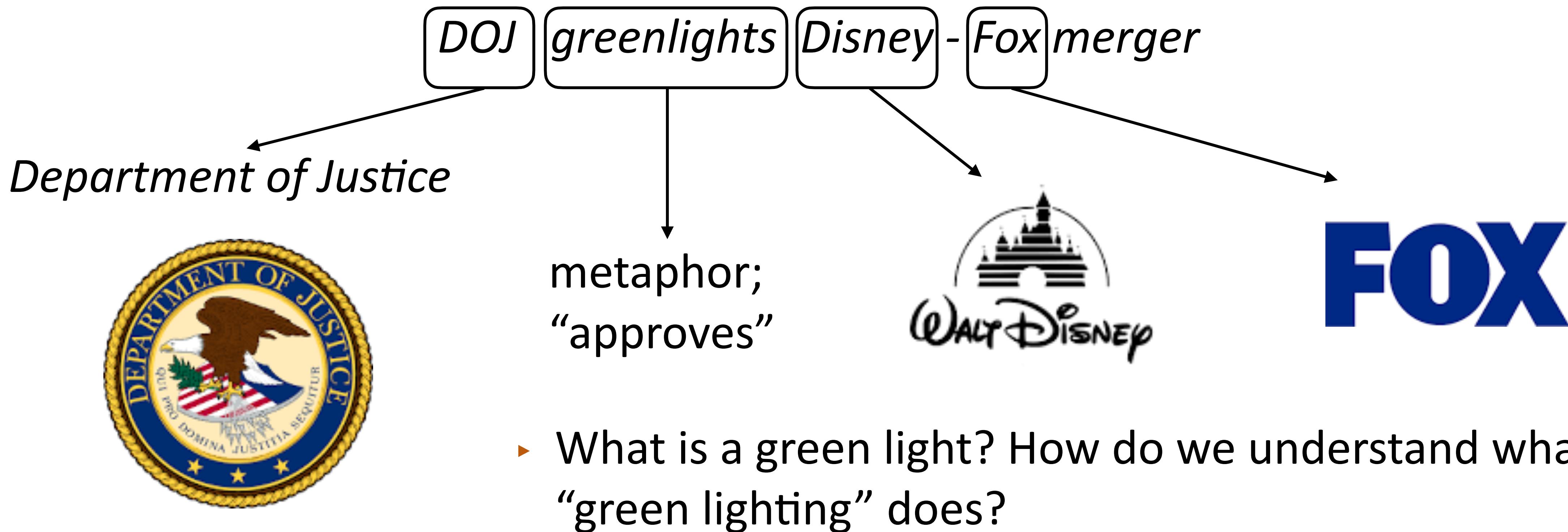
- ▶ World knowledge: have access to information beyond the training data



# What do we need to understand language?

---

- ▶ World knowledge: have access to information beyond the training data



# What do we need to understand language?

---

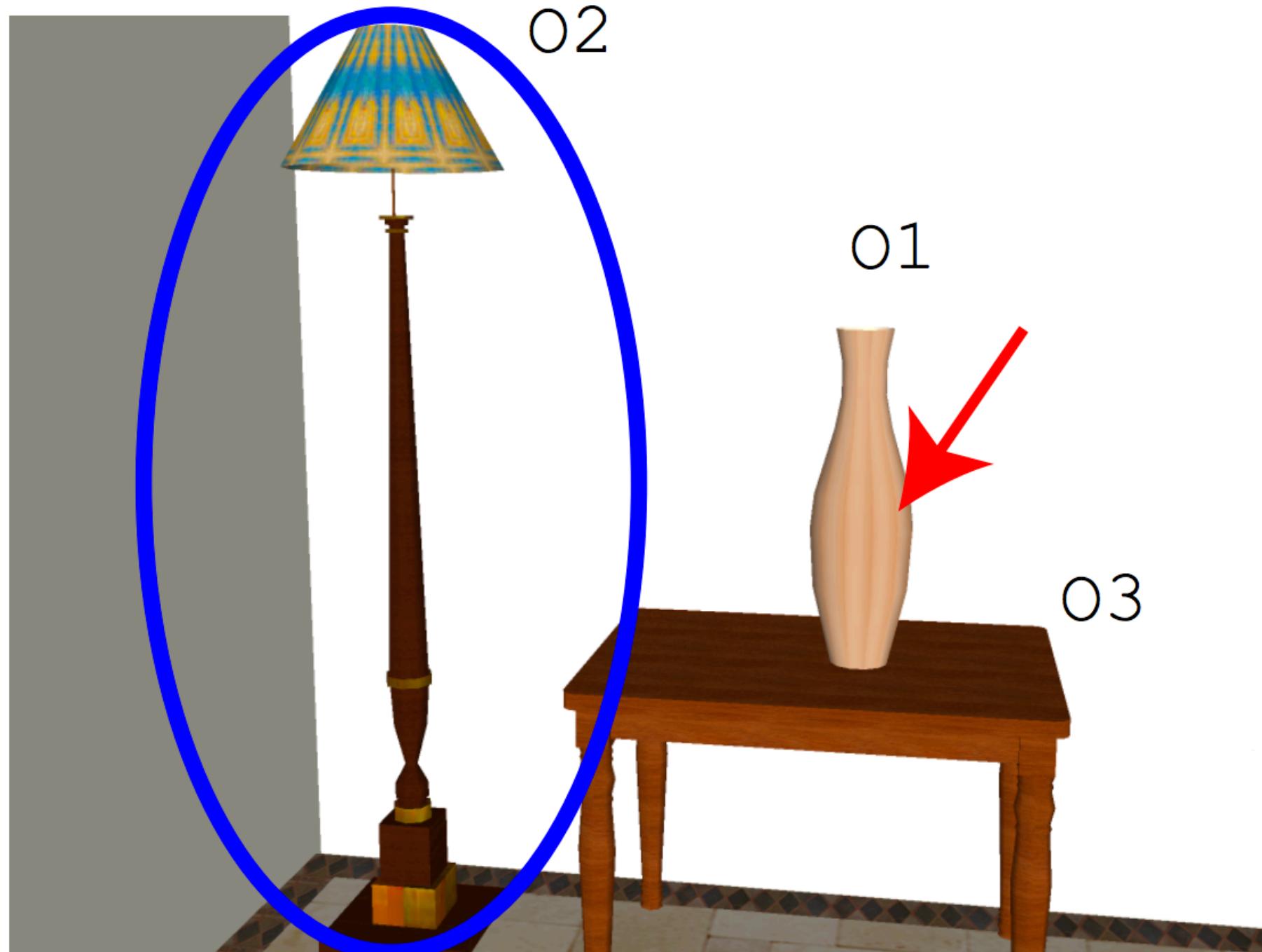
- ▶ Grounding: learn what fundamental concepts actually mean in a data-driven way

# What do we need to understand language?

---

- ▶ Grounding: learn what fundamental concepts actually mean in a data-driven way

**Question:** What object is **right of** **o2** ?

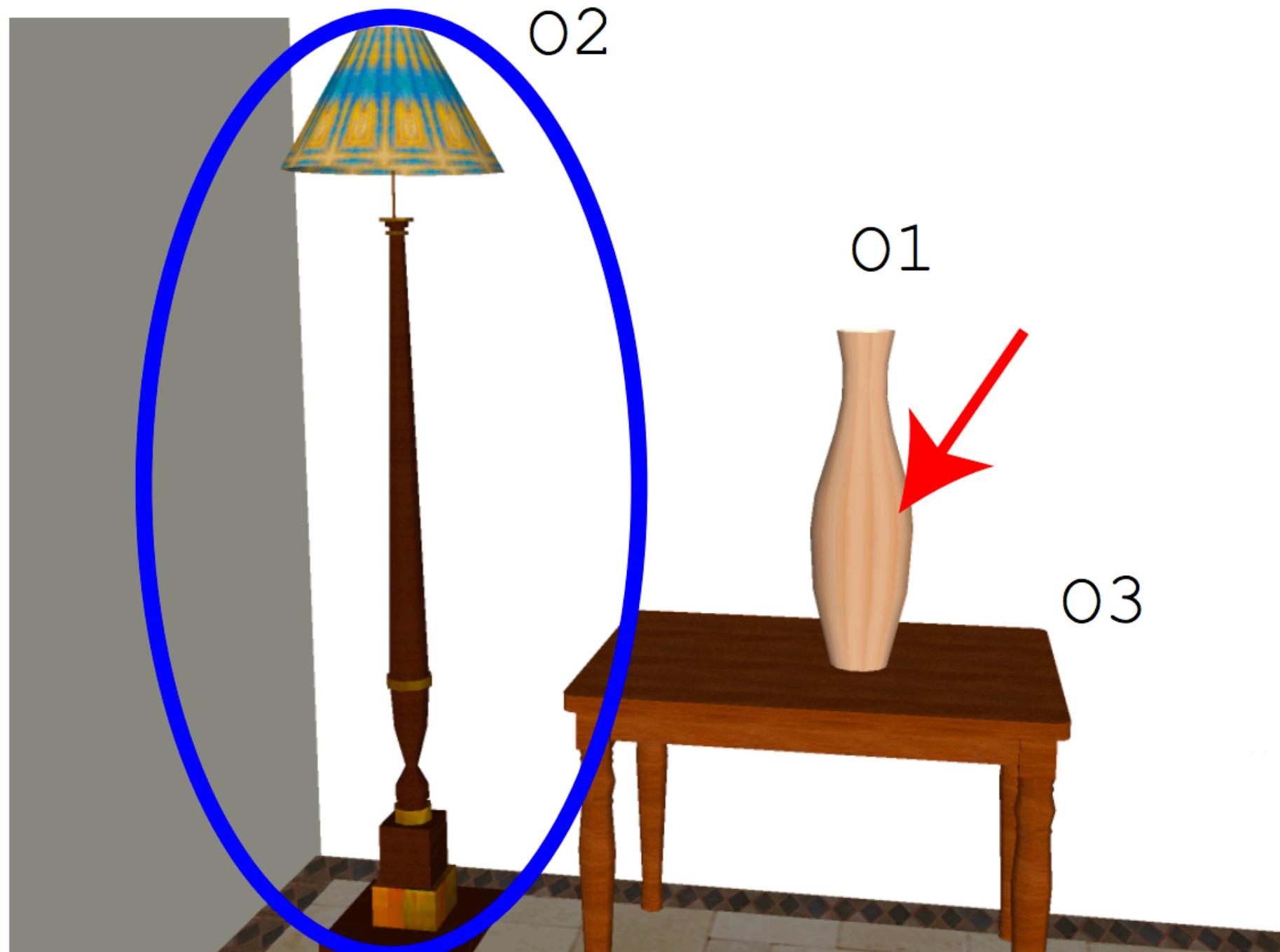


Golland et al. (2010)

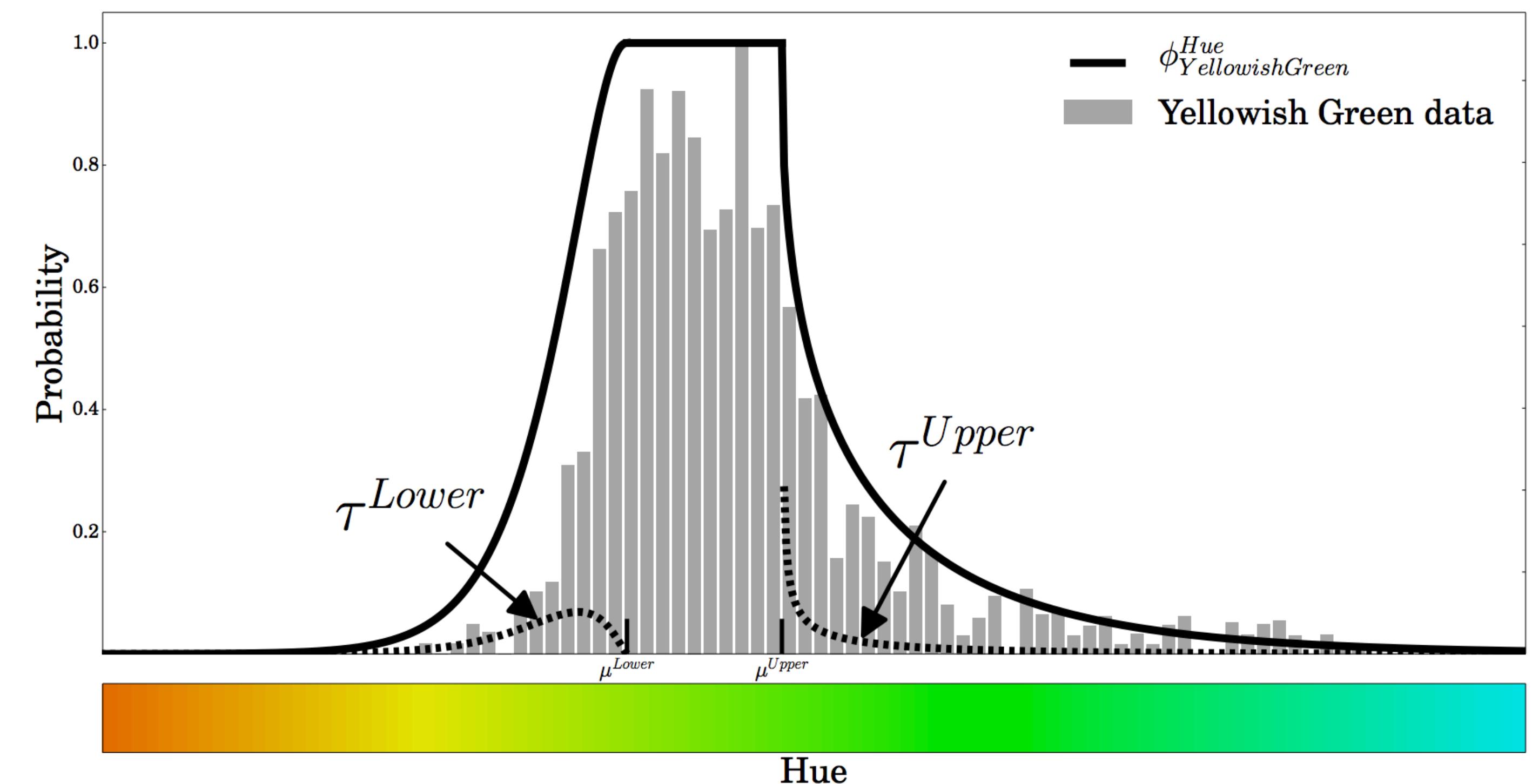
# What do we need to understand language?

- ▶ Grounding: learn what fundamental concepts actually mean in a data-driven way

**Question:** What object is right of **o2** ?



Golland et al. (2010)



McMahan and Stone (2015)

# What do we need to understand language?

---

- ▶ Linguistic structure

# What do we need to understand language?

---

- ▶ Linguistic structure
- ▶ ...but computers probably won't understand language the same way humans do

# What do we need to understand language?

---

- ▶ Linguistic structure
- ▶ ...but computers probably won't understand language the same way humans do
- ▶ However, linguistics tells us what phenomena we need to be able to deal with and gives us hints about how language works

# What do we need to understand language?

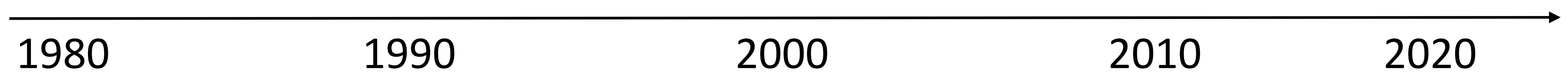
---

- ▶ Linguistic structure
- ▶ ...but computers probably won't understand language the same way humans do
- ▶ However, linguistics tells us what phenomena we need to be able to deal with and gives us hints about how language works
  - a. John has been having a lot of trouble arranging his vacation.
  - b. He cannot find anyone to take over his responsibilities. (he = John)  
 $C_b = \text{John}; C_f = \{\text{John}\}$
  - c. He called up Mike yesterday to work out a plan. (he = John)  
 $C_b = \text{John}; C_f = \{\text{John}, \text{Mike}\}$  (CONTINUE)
  - d. Mike has annoyed him a lot recently.  
 $C_b = \text{John}; C_f = \{\text{Mike}, \text{John}\}$  (RETAIN)
  - e. He called John at 5 AM on Friday last week. (he = Mike)  
 $C_b = \text{Mike}; C_f = \{\text{Mike}, \text{John}\}$  (SHIFT)

What techniques do we use?  
(to combine data, knowledge, linguistics, etc.)

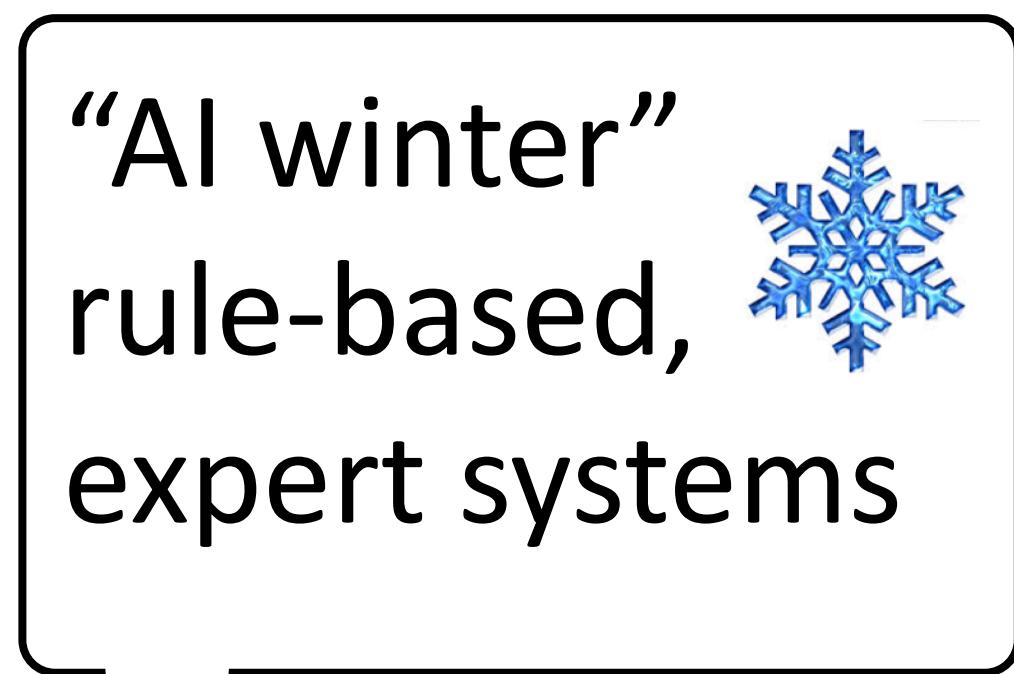
# A brief history of (modern) NLP

---



# A brief history of (modern) NLP

---



1980

1990

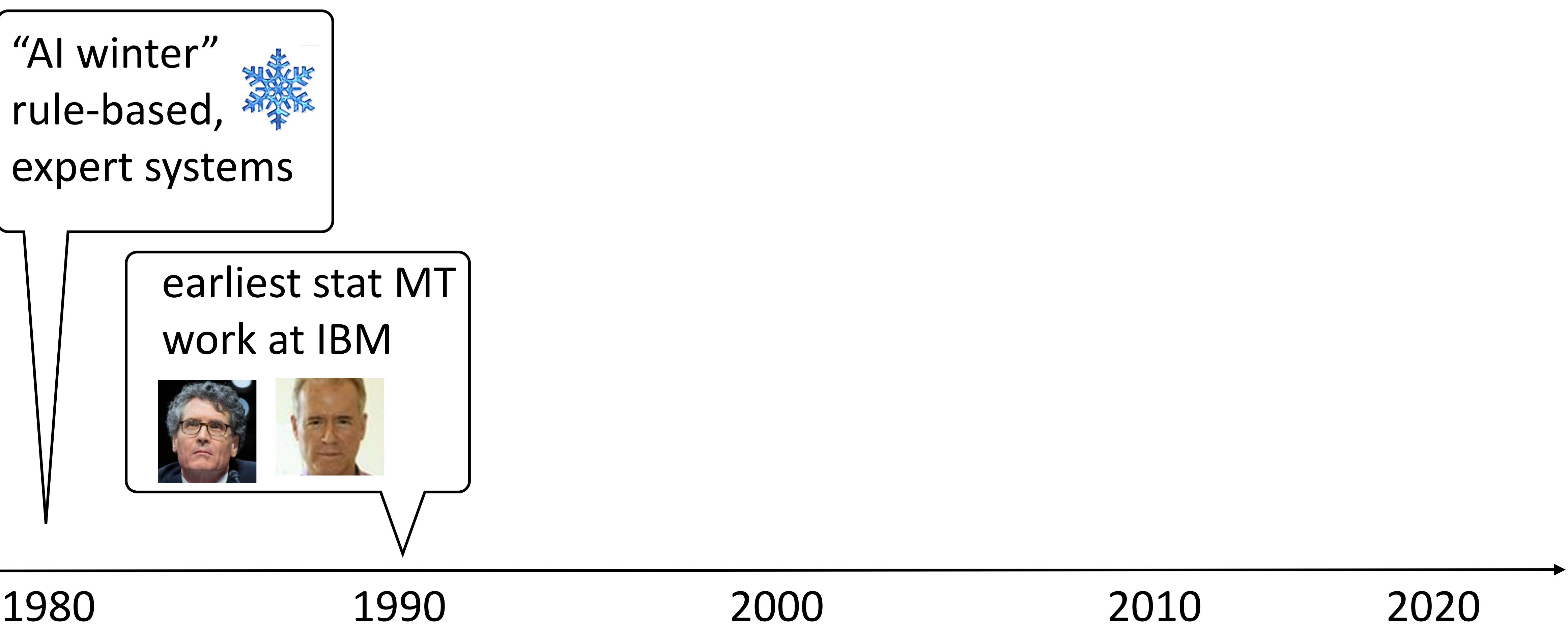
2000

2010

2020

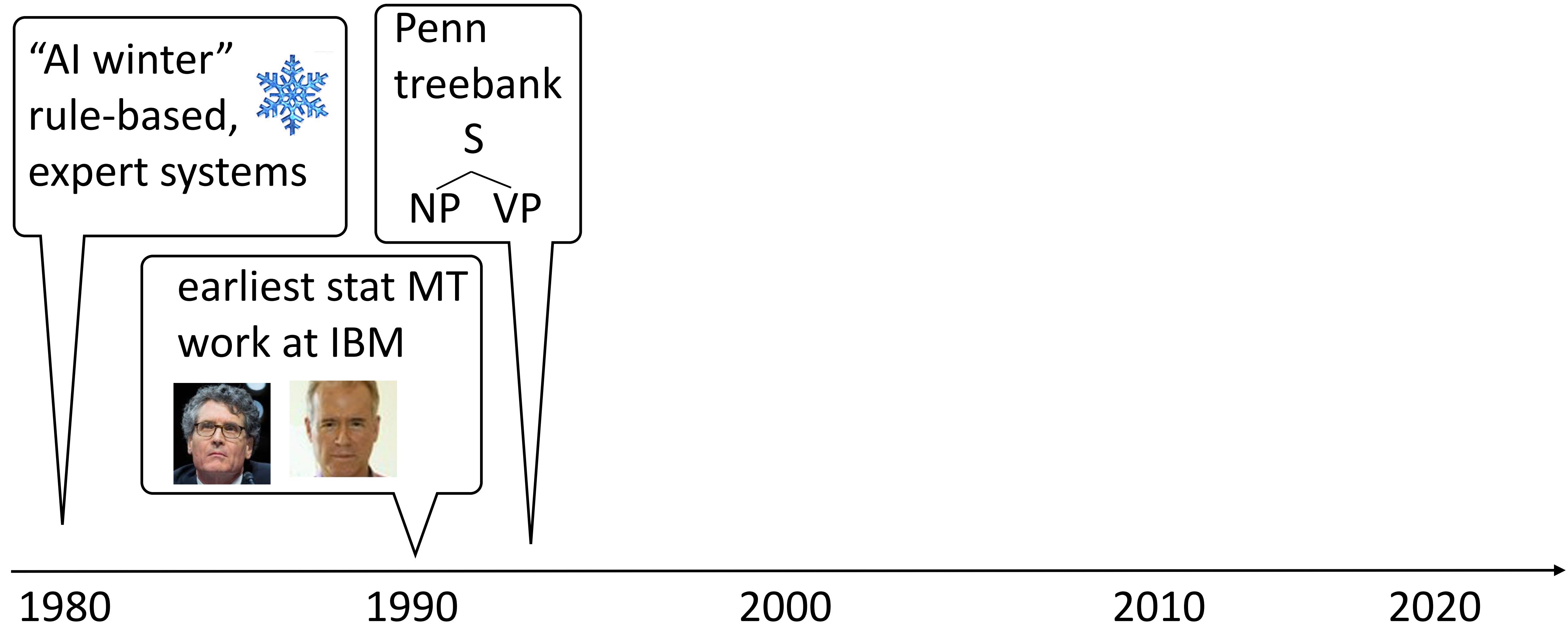
# A brief history of (modern) NLP

---



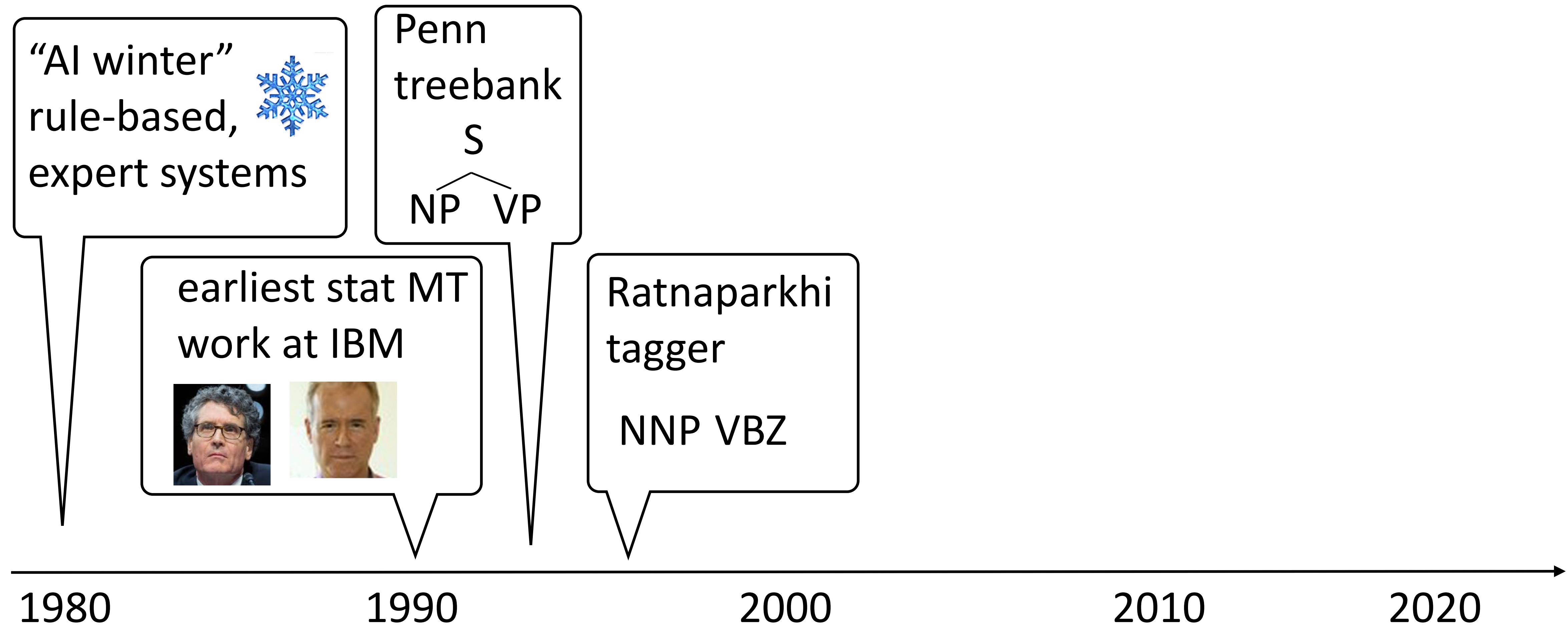
# A brief history of (modern) NLP

---



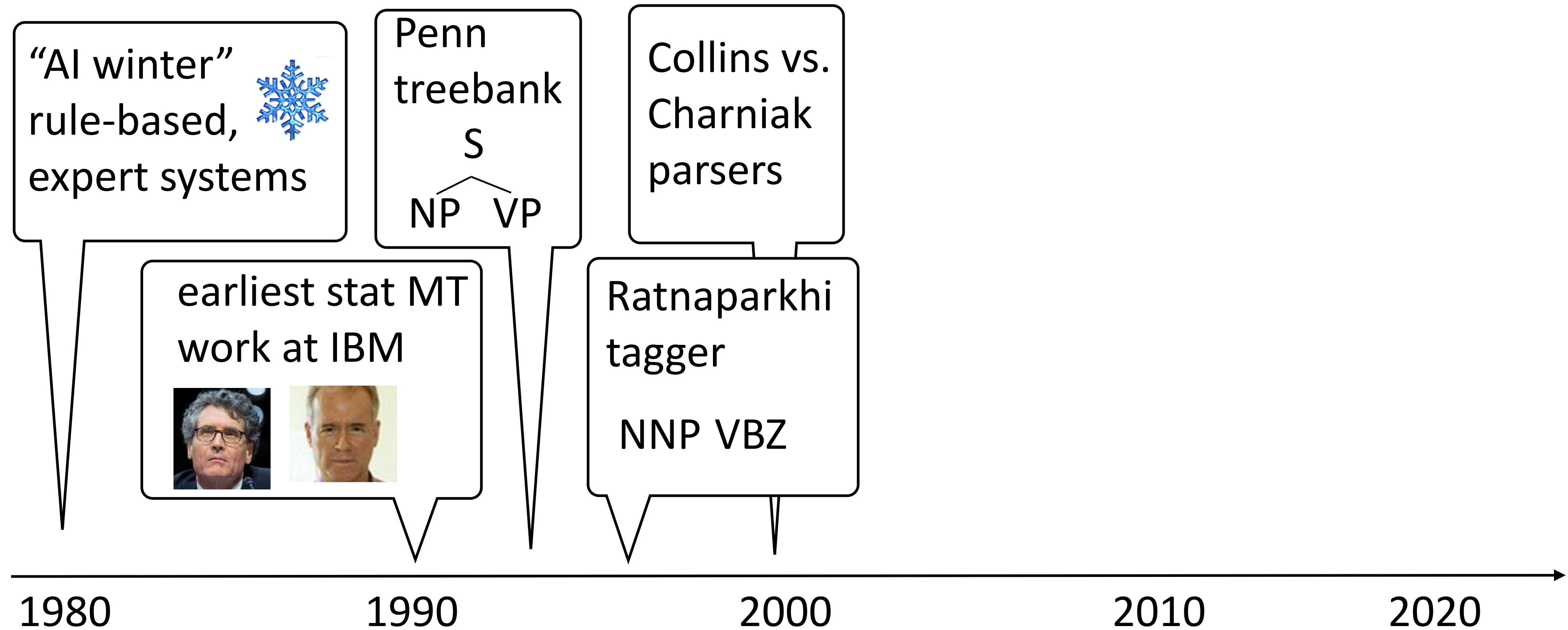
# A brief history of (modern) NLP

---



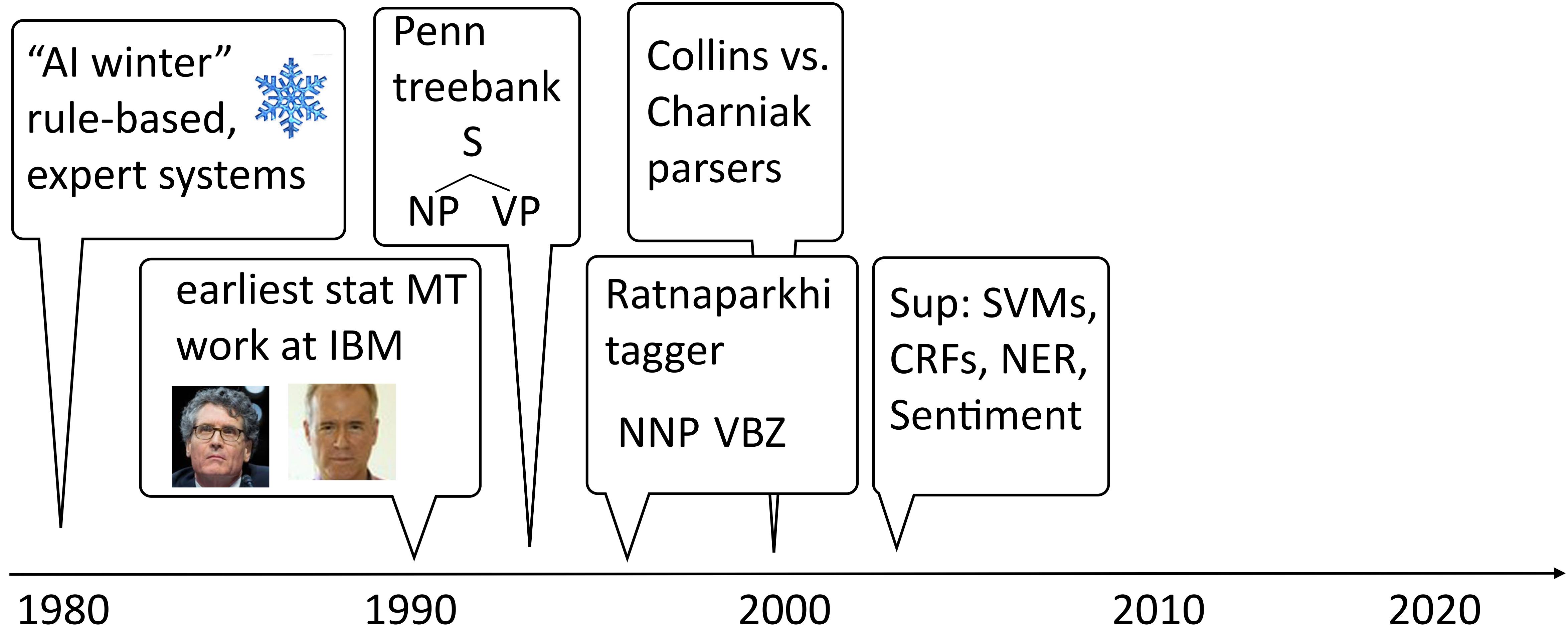
# A brief history of (modern) NLP

---

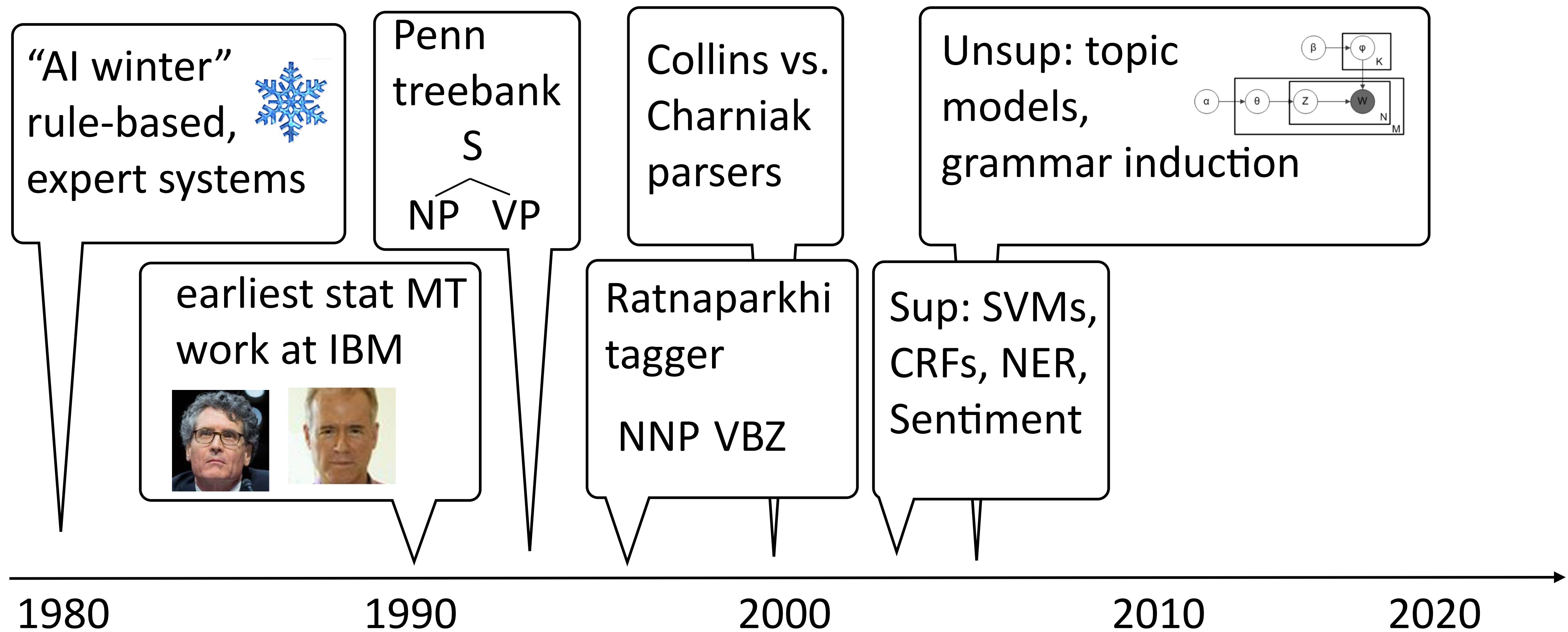


# A brief history of (modern) NLP

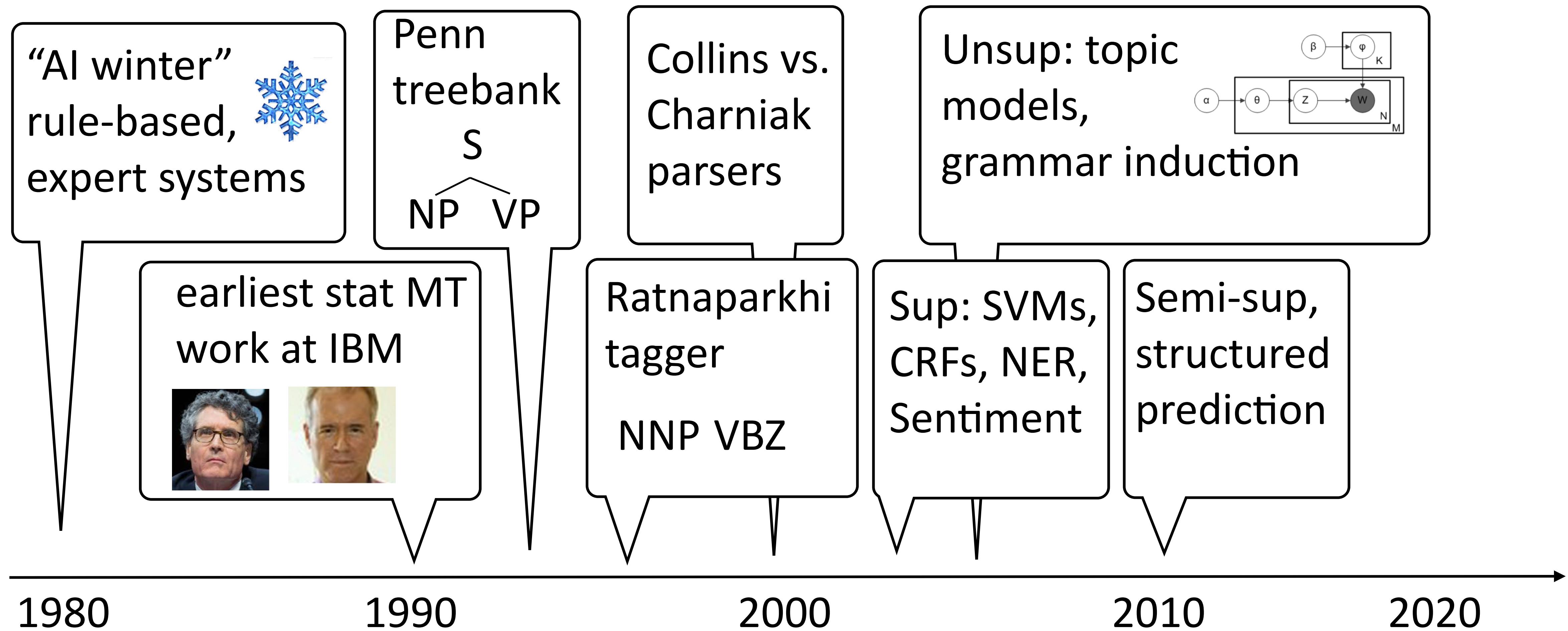
---



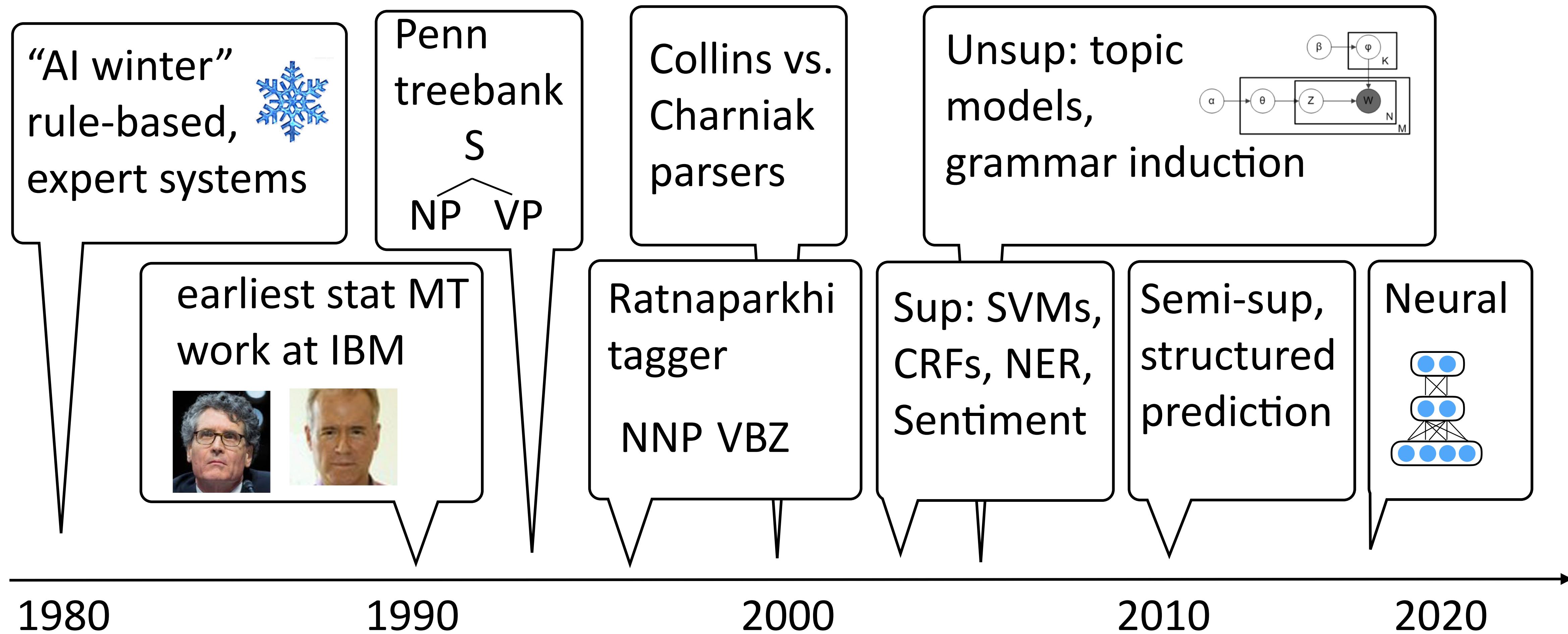
# A brief history of (modern) NLP



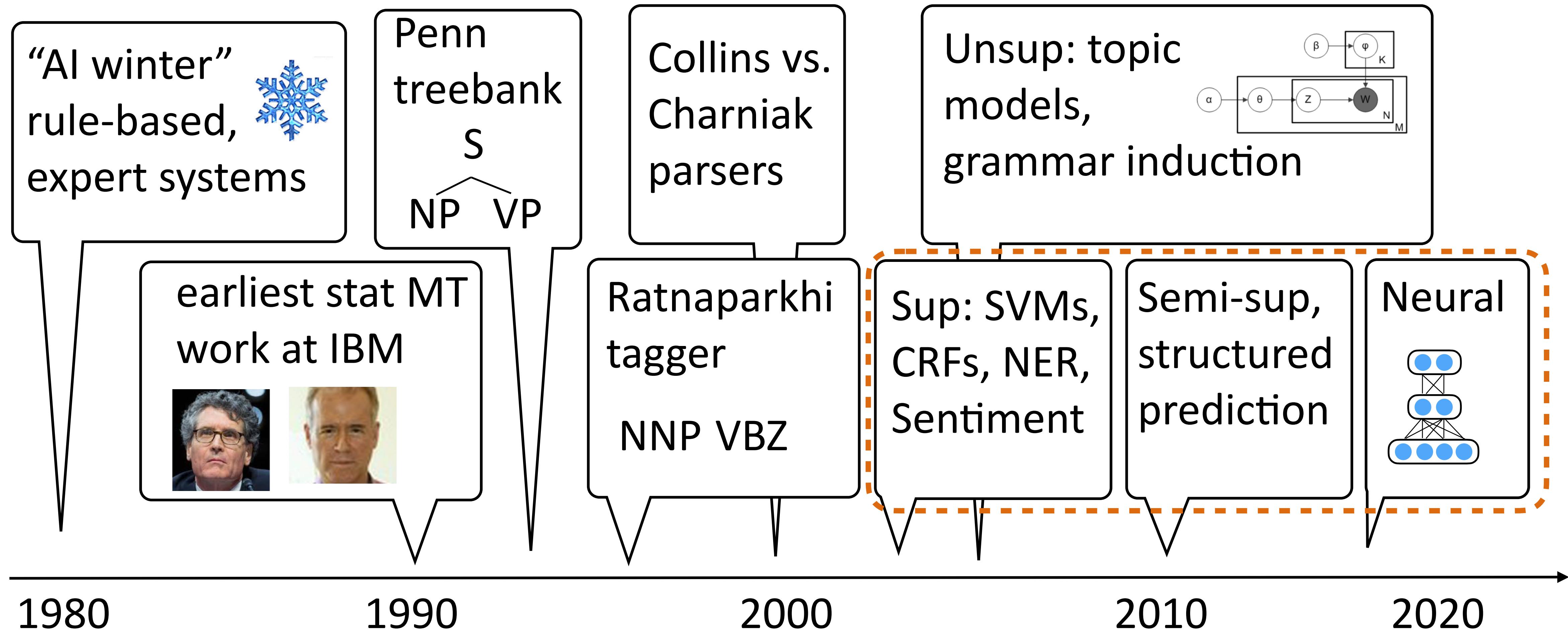
# A brief history of (modern) NLP



# A brief history of (modern) NLP



# A brief history of (modern) NLP



# Structured Prediction

---

“Learning a Part-of-Speech Tagger from Two Hours of Annotation”  
Garrette and Baldridge (2013)

# Structured Prediction

---

- ▶ All of these techniques are data-driven! Some data is naturally occurring, but may need to label

“Learning a Part-of-Speech Tagger from Two Hours of Annotation”  
Garrette and Baldridge (2013)

# Structured Prediction

---

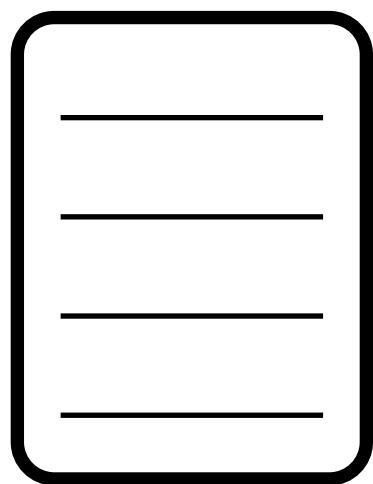
- ▶ All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- ▶ Supervised techniques work well on very little data

“Learning a Part-of-Speech Tagger from Two Hours of Annotation”  
Garrette and Baldridge (2013)

# Structured Prediction

---

- ▶ All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- ▶ Supervised techniques work well on very little data

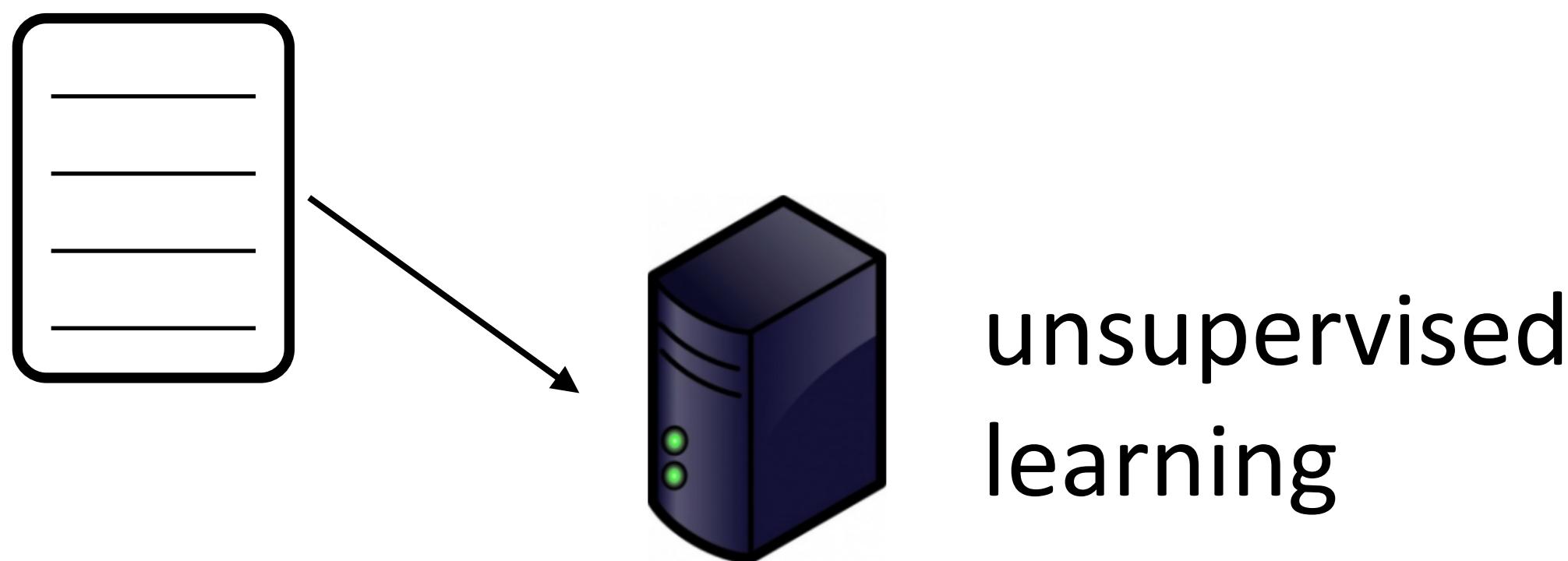


“Learning a Part-of-Speech Tagger from Two Hours of Annotation”  
Garrette and Baldridge (2013)

# Structured Prediction

---

- ▶ All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- ▶ Supervised techniques work well on very little data

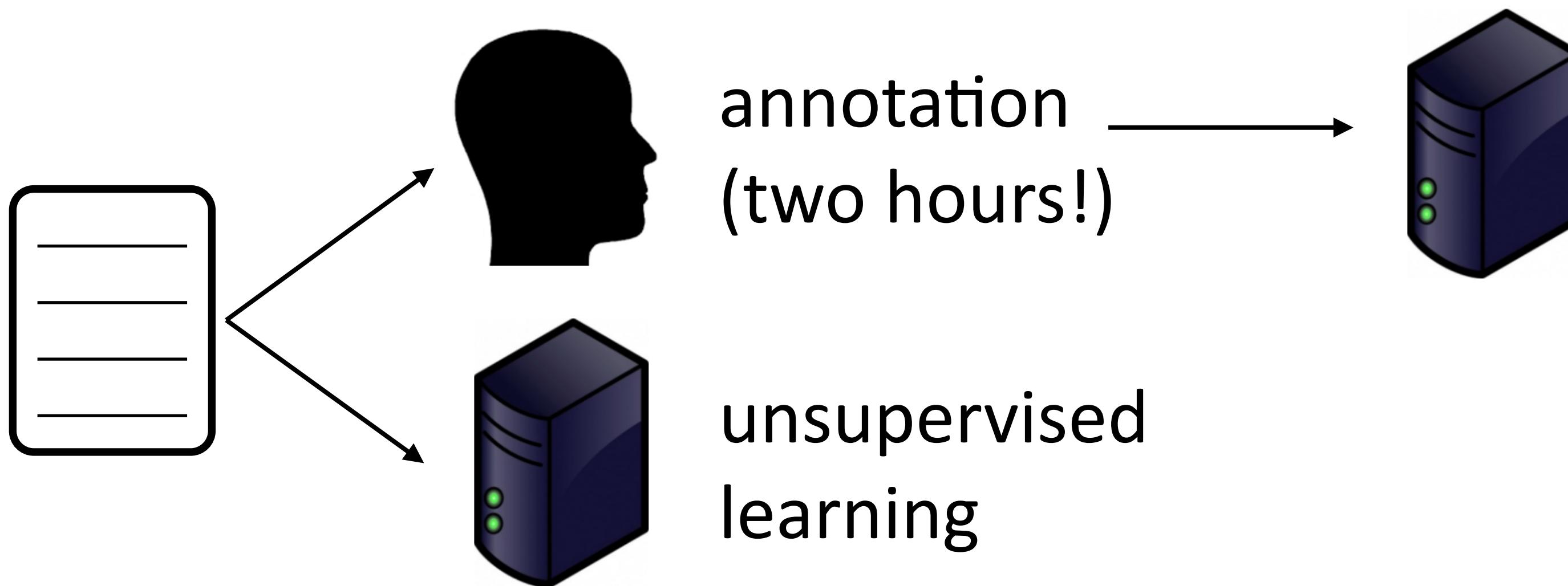


“Learning a Part-of-Speech Tagger from Two Hours of Annotation”  
Garrette and Baldridge (2013)

# Structured Prediction

---

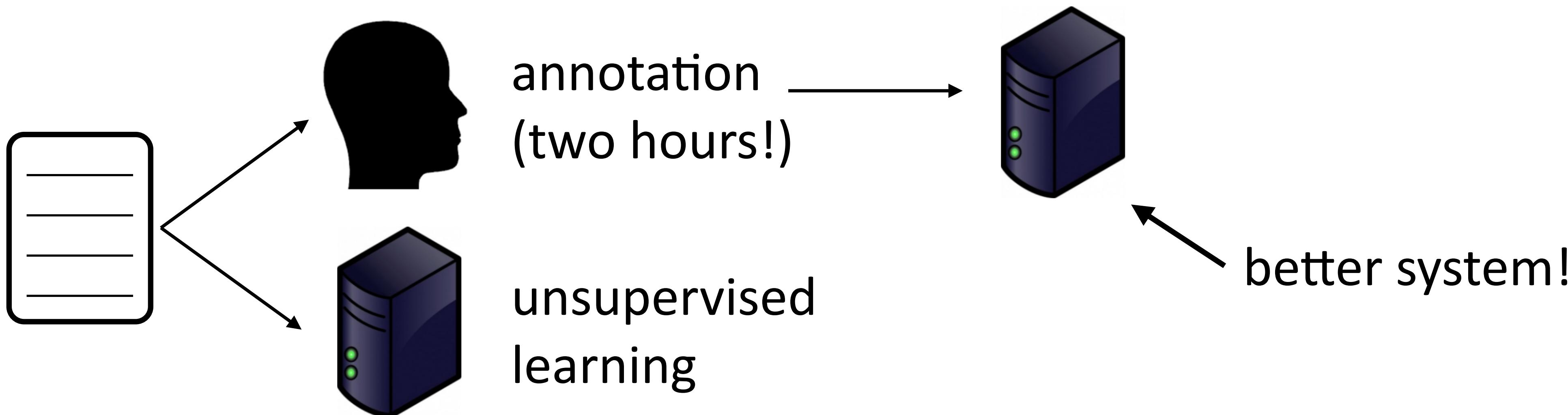
- ▶ All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- ▶ Supervised techniques work well on very little data



“Learning a Part-of-Speech Tagger from Two Hours of Annotation”  
Garrette and Baldridge (2013)

# Structured Prediction

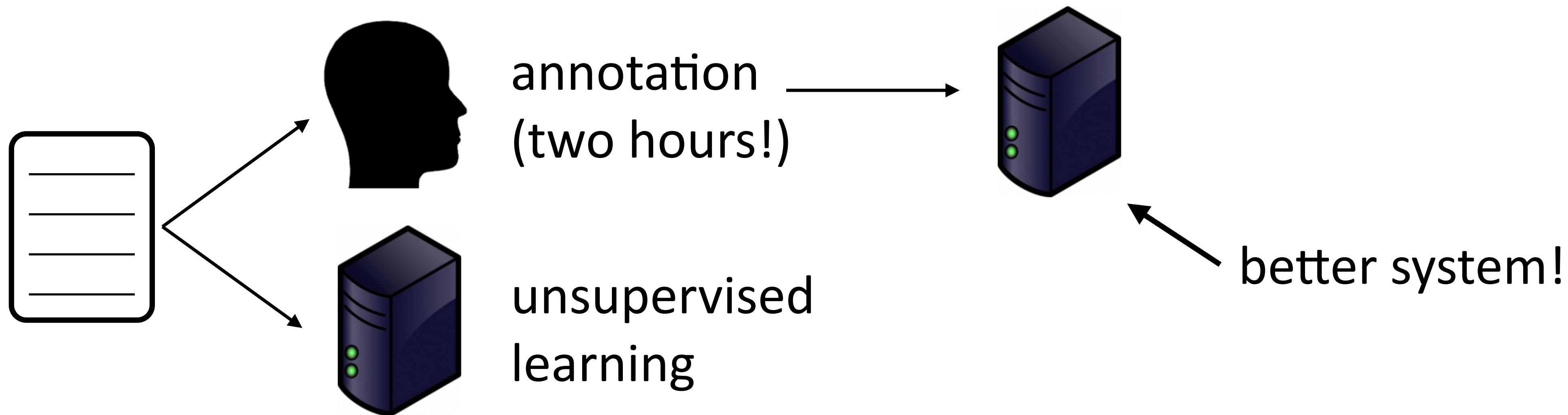
- ▶ All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- ▶ Supervised techniques work well on very little data



“Learning a Part-of-Speech Tagger from Two Hours of Annotation”  
Garrette and Baldridge (2013)

# Structured Prediction

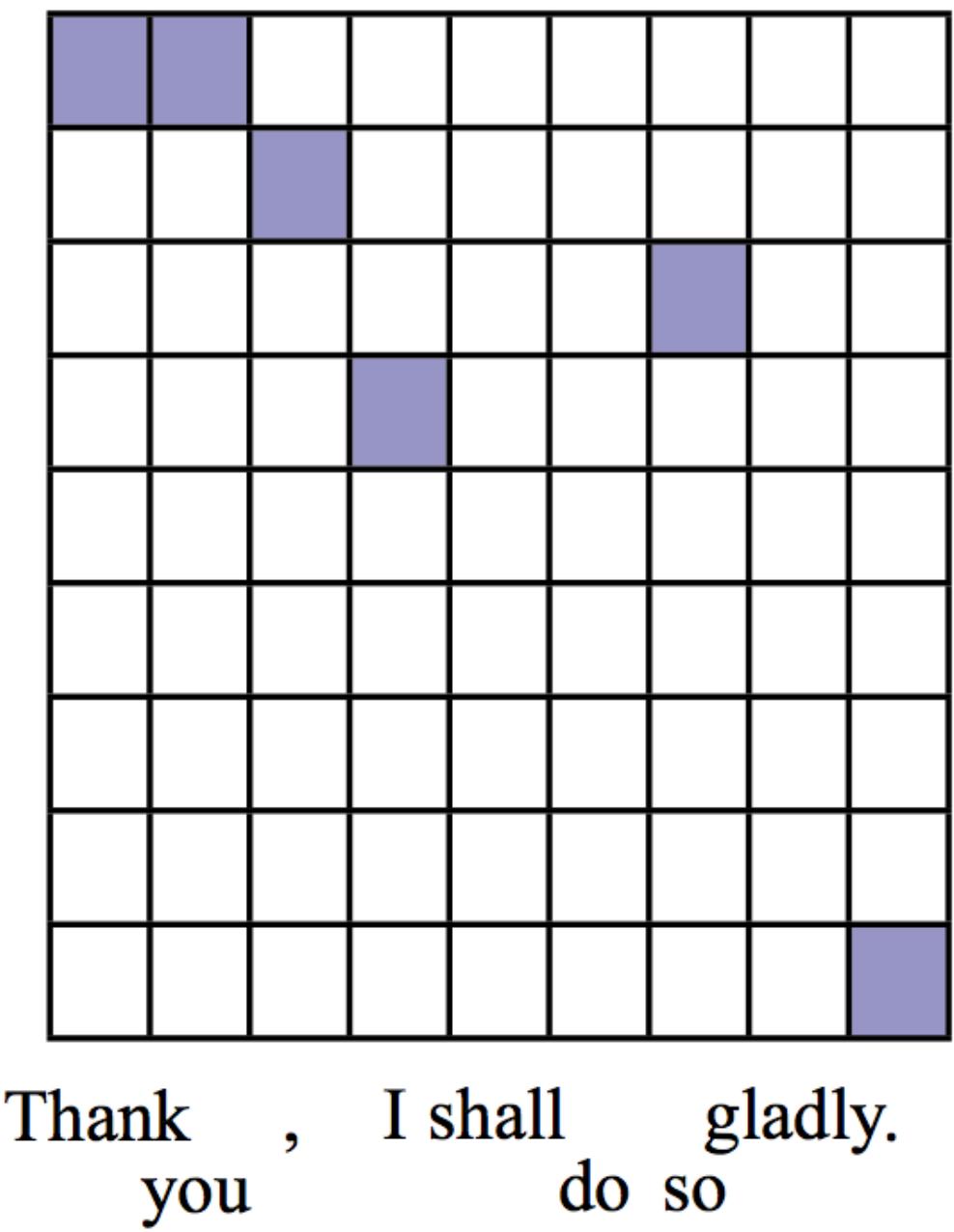
- ▶ All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- ▶ Supervised techniques work well on very little data



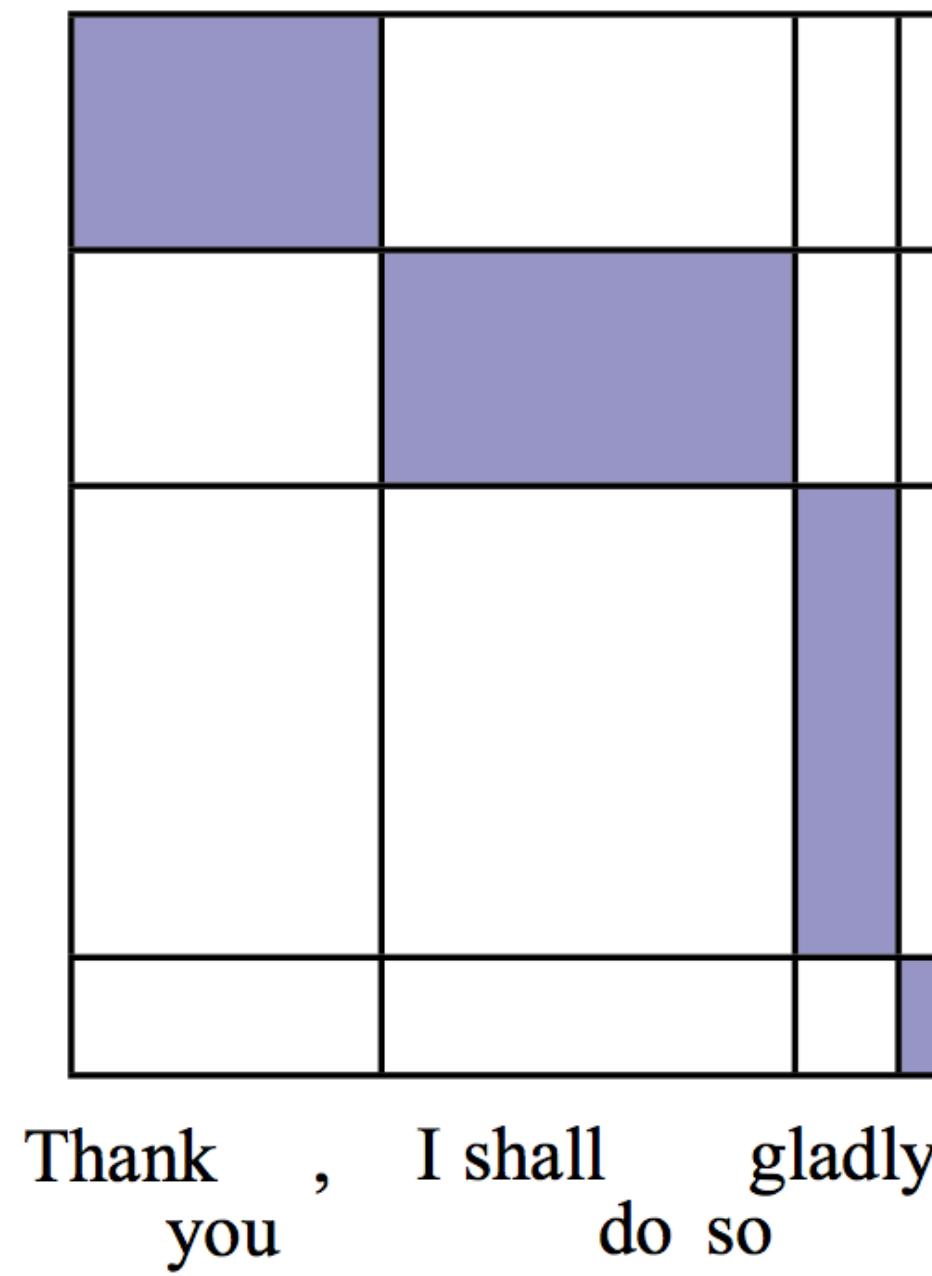
- ▶ Even neural nets can do pretty well!

“Learning a Part-of-Speech Tagger from Two Hours of Annotation”  
Garrette and Baldridge (2013)

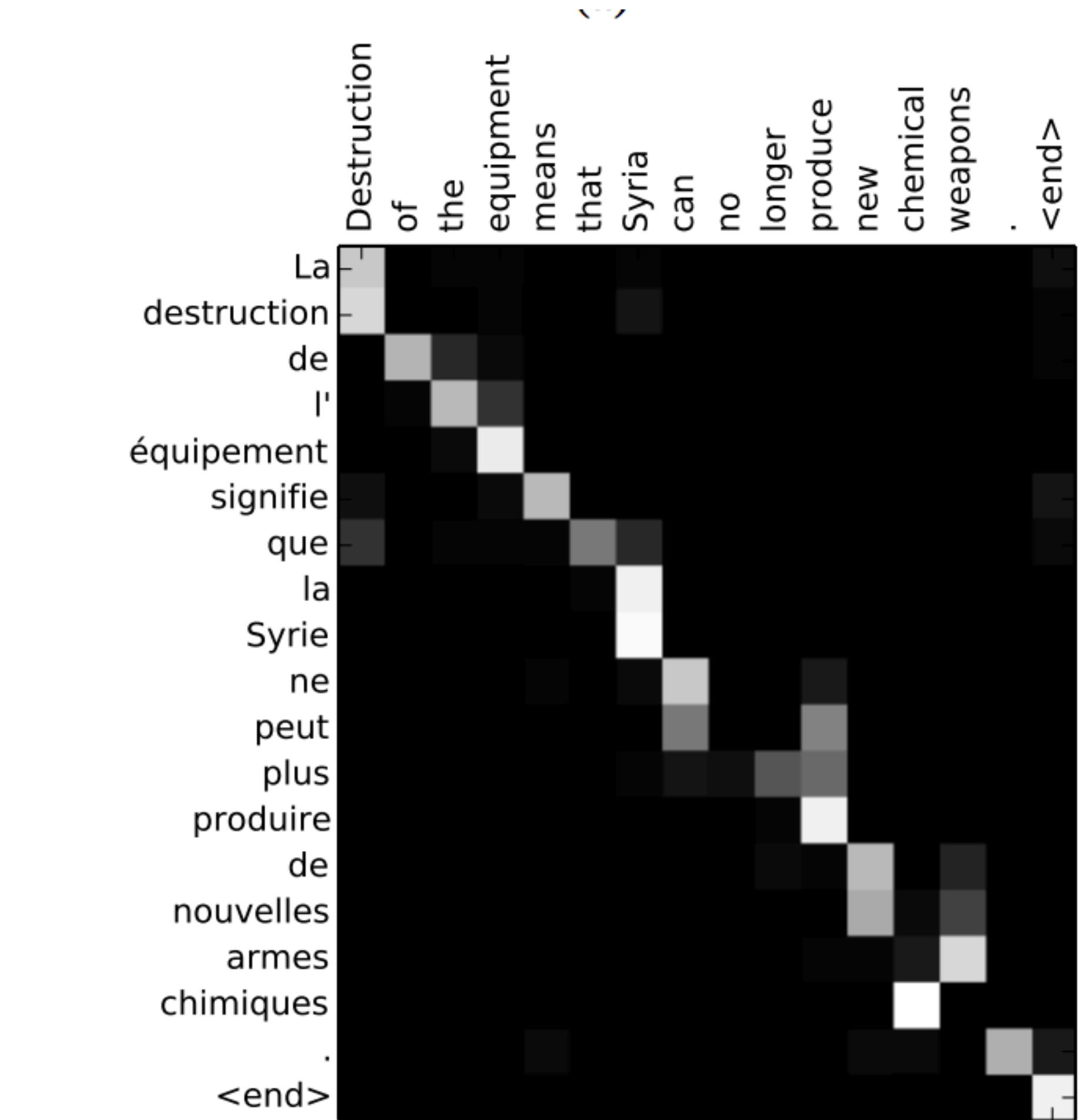
# Less Manual Structure?



**(a) example word alignment**



**(b) example phrase alignment**



DeNero et al. (2008)

Bahdanau et al. (2014)

# Does manual structure have a place?

---

# Does manual structure have a place?

---

- ▶ Neural nets don't always work out of domain!

# Does manual structure have a place?

---

- ▶ Neural nets don't always work out of domain!
- ▶ Coreference: rule-based systems are still about as good as deep learning out-of-domain

# Does manual structure have a place?

- ▶ Neural nets don't always work out of domain!
- ▶ Coreference: rule-based systems are still about as good as deep learning out-of-domain

|                         | CoNLL<br>Avg. F <sub>1</sub> |
|-------------------------|------------------------------|
| <b>Newswire</b>         |                              |
| rule-based              | 55.60                        |
| berkeley                | 61.24                        |
| cort                    | 63.37                        |
| deep-coref [conll]      | 65.39                        |
| deep-coref [lea]        | 65.60                        |
| <b>Wikipedia</b>        |                              |
| rule-based              | 51.77                        |
| berkeley                | 51.01                        |
| cort                    | 49.94                        |
| deep-coref [conll]      | 52.65                        |
| deep-coref [lea]        | 53.14                        |
| deep-coref <sup>-</sup> | 51.01                        |

Moosavi and Strube (2017)

# Does manual structure have a place?

- ▶ Neural nets don't always work out of domain!
- ▶ Coreference: rule-based systems are still about as good as deep learning out-of-domain
- ▶ LORELEI: transition point below which phrase-based systems are better

|                         | CoNLL<br>Avg. F <sub>1</sub> |
|-------------------------|------------------------------|
| <b>Newswire</b>         |                              |
| rule-based              | 55.60                        |
| berkeley                | 61.24                        |
| cort                    | 63.37                        |
| deep-coref [conll]      | 65.39                        |
| deep-coref [lea]        | 65.60                        |
| <b>Wikipedia</b>        |                              |
| rule-based              | 51.77                        |
| berkeley                | 51.01                        |
| cort                    | 49.94                        |
| deep-coref [conll]      | 52.65                        |
| deep-coref [lea]        | 53.14                        |
| deep-coref <sup>-</sup> | 51.01                        |

Moosavi and Strube (2017)

# Does manual structure have a place?

- ▶ Neural nets don't always work out of domain!
- ▶ Coreference: rule-based systems are still about as good as deep learning out-of-domain
- ▶ LORELEI: transition point below which phrase-based systems are better
- ▶ Why is this? Inductive bias!

|                         | CoNLL<br>Avg. F <sub>1</sub> |
|-------------------------|------------------------------|
| <b>Newswire</b>         |                              |
| rule-based              | 55.60                        |
| berkeley                | 61.24                        |
| cort                    | 63.37                        |
| deep-coref [conll]      | 65.39                        |
| deep-coref [lea]        | 65.60                        |
| <b>Wikipedia</b>        |                              |
| rule-based              | 51.77                        |
| berkeley                | 51.01                        |
| cort                    | 49.94                        |
| deep-coref [conll]      | 52.65                        |
| deep-coref [lea]        | 53.14                        |
| deep-coref <sup>-</sup> | 51.01                        |

Moosavi and Strube (2017)

# Does manual structure have a place?

- ▶ Neural nets don't always work out of domain!
- ▶ Coreference: rule-based systems are still about as good as deep learning out-of-domain
- ▶ LORELEI: transition point below which phrase-based systems are better
- ▶ Why is this? Inductive bias!
- ▶ Can multi-task learning help?

|                         | CoNLL<br>Avg. F <sub>1</sub> |
|-------------------------|------------------------------|
| <b>Newswire</b>         |                              |
| rule-based              | 55.60                        |
| berkeley                | 61.24                        |
| cort                    | 63.37                        |
| deep-coref [conll]      | 65.39                        |
| deep-coref [lea]        | 65.60                        |
| <b>Wikipedia</b>        |                              |
| rule-based              | 51.77                        |
| berkeley                | 51.01                        |
| cort                    | 49.94                        |
| deep-coref [conll]      | 52.65                        |
| deep-coref [lea]        | 53.14                        |
| deep-coref <sup>-</sup> | 51.01                        |

Moosavi and Strube (2017)

# Does manual structure have a place?

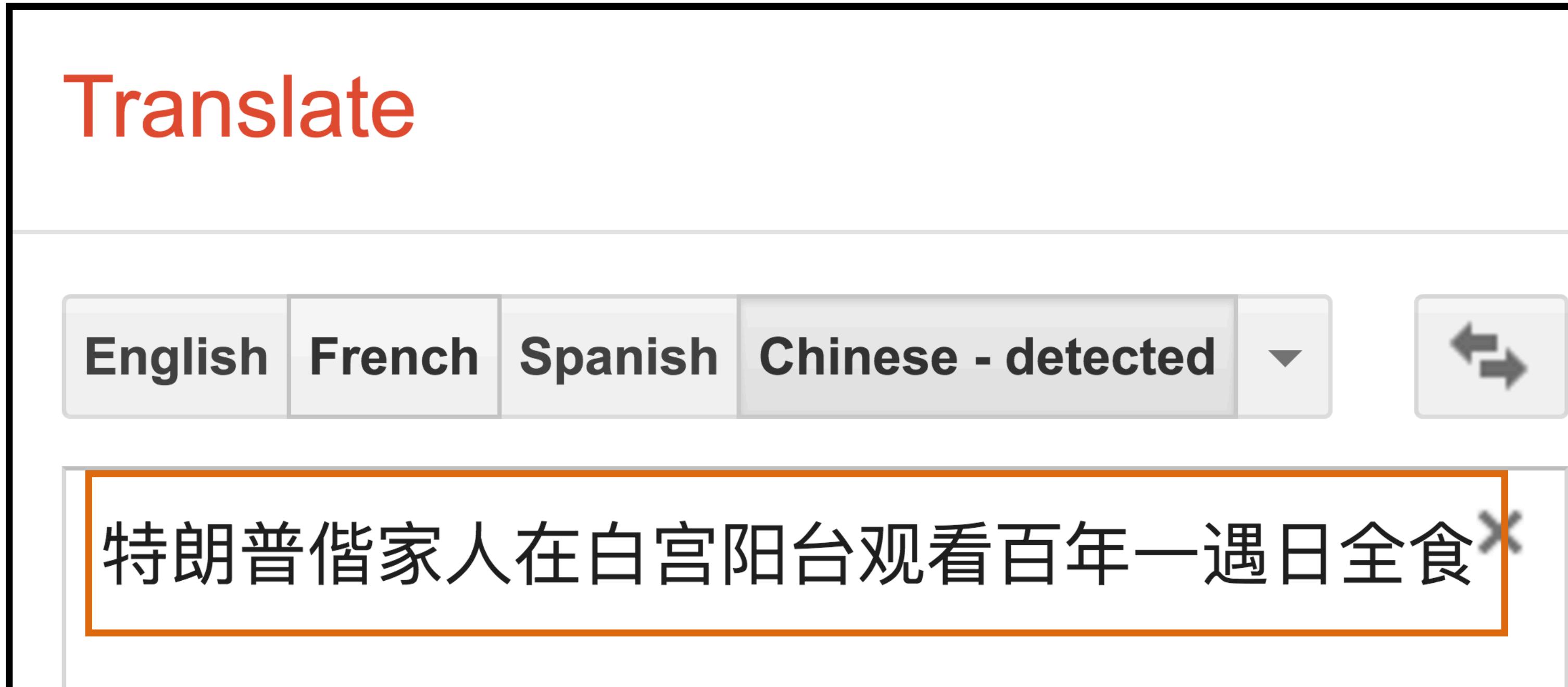
Translate

English French Spanish Chinese - detected ▾

特朗普偕家人在白宫阳台观看百年一遇日全食 X

Trump Pope family watch a hundred years a year in the White House balcony

# Does manual structure have a place?



Trump **Pope** family watch a hundred years a year **in** the White House balcony

- ▶ Maybe manual structure would help...

# Where are we?

---

# Where are we?

---

- ▶ NLP consists of: analyzing and building representations for text, solving problems involving text

# Where are we?

---

- ▶ NLP consists of: analyzing and building representations for text, solving problems involving text
- ▶ These problems are hard because language is ambiguous, requires drawing on data, knowledge, and linguistics to solve

# Where are we?

---

- ▶ NLP consists of: analyzing and building representations for text, solving problems involving text
- ▶ These problems are hard because language is ambiguous, requires drawing on data, knowledge, and linguistics to solve
- ▶ Knowing which techniques to use requires understanding dataset size, problem complexity, and a lot of tricks!

# Where are we?

---

- ▶ NLP consists of: analyzing and building representations for text, solving problems involving text
- ▶ These problems are hard because language is ambiguous, requires drawing on data, knowledge, and linguistics to solve
- ▶ Knowing which techniques to use requires understanding dataset size, problem complexity, and a lot of tricks!
- ▶ NLP encompasses all of these things

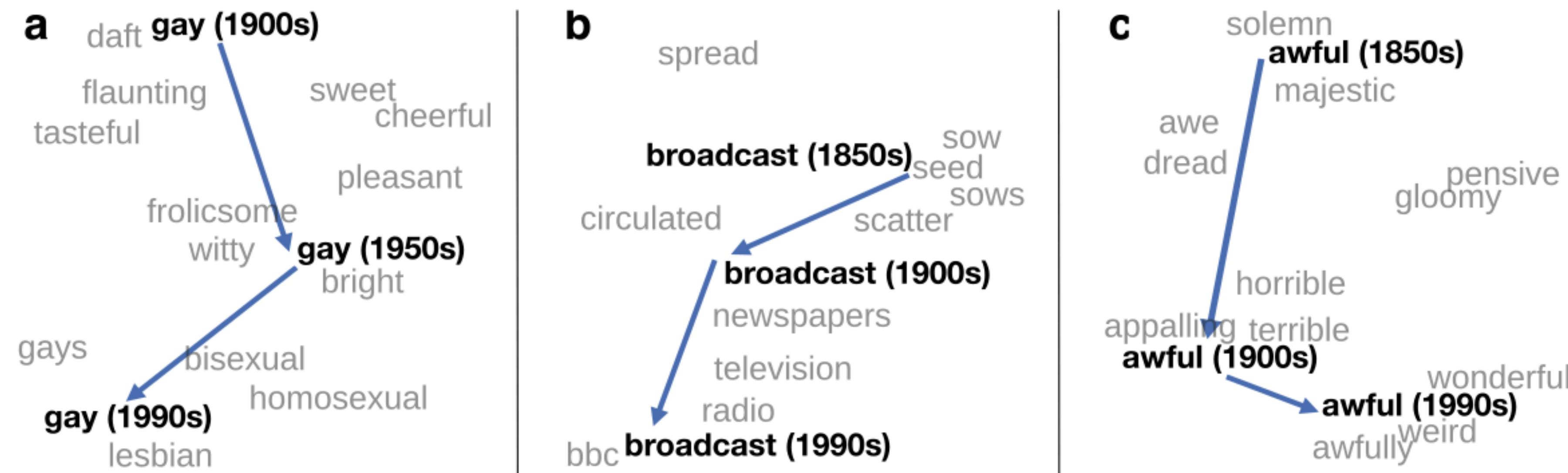
# NLP vs. Computational Linguistics

---

- ▶ NLP: build systems that deal with language data
- ▶ CL: use computational tools to study language

# NLP vs. Computational Linguistics

- ▶ NLP: build systems that deal with language data
- ▶ CL: use computational tools to study language



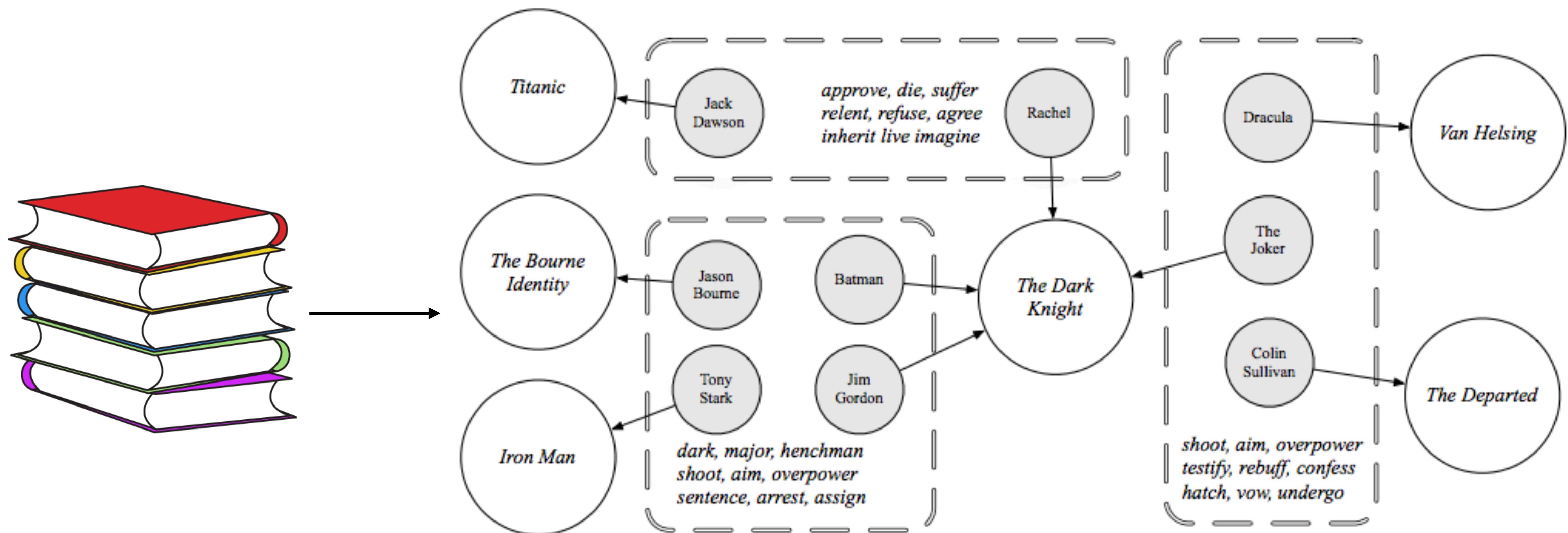
# NLP vs. Computational Linguistics

---

- ▶ Computational tools for other purposes: literary theory, political science...

# NLP vs. Computational Linguistics

- ▶ Computational tools for other purposes: literary theory, political science...



# Course Goals

---

# Course Goals

---

- ▶ Cover fundamental machine learning techniques used in NLP

# Course Goals

---

- ▶ Cover fundamental machine learning techniques used in NLP
- ▶ Understand how to look at language data and approach linguistic phenomena

# Course Goals

---

- ▶ Cover fundamental machine learning techniques used in NLP
- ▶ Understand how to look at language data and approach linguistic phenomena
- ▶ Cover modern NLP problems encountered in the literature: what are the active research topics in 2022?

# Course Goals

---

- ▶ Cover fundamental machine learning techniques used in NLP
- ▶ Understand how to look at language data and approach linguistic phenomena
- ▶ Cover modern NLP problems encountered in the literature: what are the active research topics in 2022?
- ▶ Make you a “producer” rather than a “consumer” of NLP tools

# Course Goals

---

- ▶ Cover fundamental machine learning techniques used in NLP
- ▶ Understand how to look at language data and approach linguistic phenomena
- ▶ Cover modern NLP problems encountered in the literature: what are the active research topics in 2022?
- ▶ Make you a “producer” rather than a “consumer” of NLP tools
  - ▶ The three assignments should teach you what you need to know to understand nearly any system in the literature

# Assignments

---

- ▶ 3 Programming Assignments
  - ▶ Implementation-oriented
  - ▶ ~2 weeks per assignment, 3 “slip days” for automatic extensions

# Assignments

---

- ▶ 3 Programming Assignments
  - ▶ Implementation-oriented
  - ▶ ~2 weeks per assignment, 3 “slip days” for automatic extensions

These projects require understanding of the concepts, ability to write performant code, and ability to think about how to debug complex systems. **They are challenging, so start early!**

# Final Project

---

- ▶ Final project (20%)
  - ▶ Groups of 3-4 preferred, 1 is possible.
  - ▶ 4 page report + final project presentation.