

CS 7650: Natural Language Processing

Wei Xu

(many slides from Greg Durrett)

Administrivia

- ▶ Course website:

https://cocoxu.github.io/CS7650_fall2021/

- ▶ Piazza and Gradescope:

- ▶ links on the course website
- ▶ We will do our best to make sure questions about the homework, etc. get answered within 24 hours

- ▶ TA Office hours: TBA

Instructor



Wei Xu

wei.xu@cc.gatech.edu

Teaching Assistants



Mounica Maddela

mmaddela3@gatech.edu



Nathan Vaska

nvaska3@gatech.edu



Sarah Wiegreffe

swiegreffe6@gatech.edu

Course Requirements

- ▶ Probability
- ▶ Linear Algebra
- ▶ Multivariable Calculus
- ▶ Programming / Python experience
- ▶ Prior exposure to machine learning very helpful

There will be a lot of math and programming!

Free Textbooks!

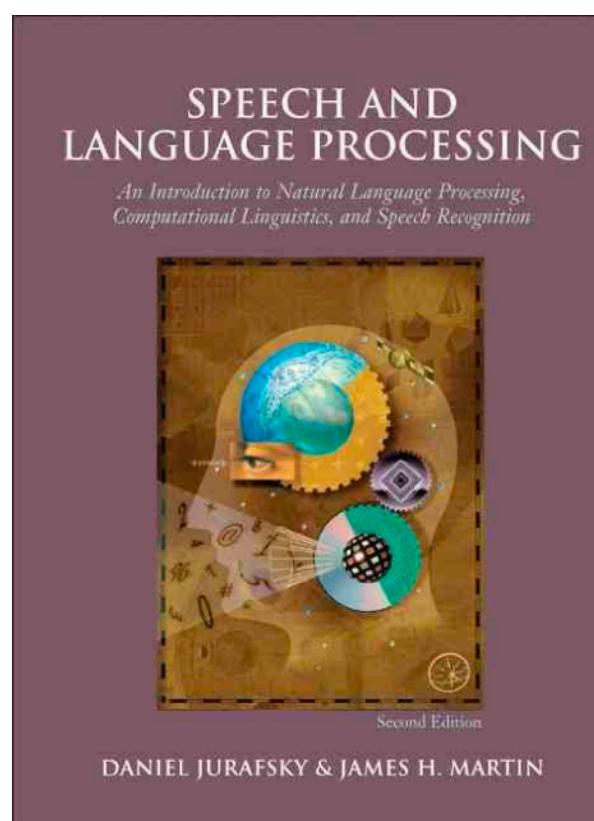
- ▶ 2 really awesome free textbooks available
- ▶ There will be assigned readings from both
- ▶ Both freely available online

Speech and Language Processing (3rd ed. draft)

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

NEW

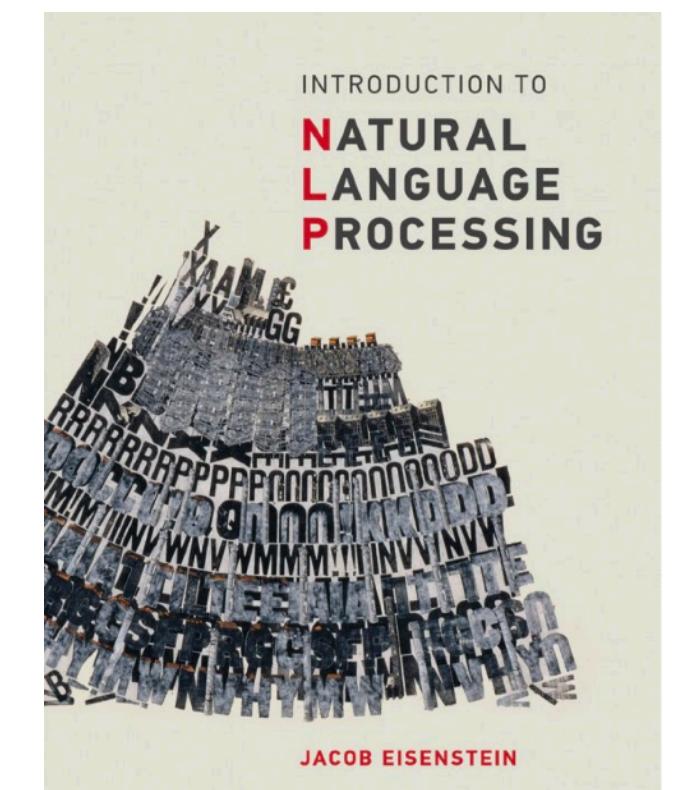
Here's our December 30, 2020 draft! Includes:



Introduction to Natural Language Processing

By [Jacob Eisenstein](#)

Published by The MIT Press
Oct 01, 2019 | 536 Pages | 7 x 9
| ISBN 9780262042840



Coursework Plan

- ▶ 3 Programming Projects (40%; fairly substantial implementation effort)
 - ▶ Text classification
 - ▶ Named entity recognition (BiLSTM-CNN-CRF)
 - ▶ Neural chatbot (Seq2Seq with attention)
 - ▶ 2 written assignments (20%) + midterm exam (15%)
 - ▶ Mostly math problems related to ML / NLP
 - ▶ Final project (20%; details on course website, will discuss later)
 - ▶ Problem Set 1 (math review) will be released later this week on GradeScope.
- }
- subject to change

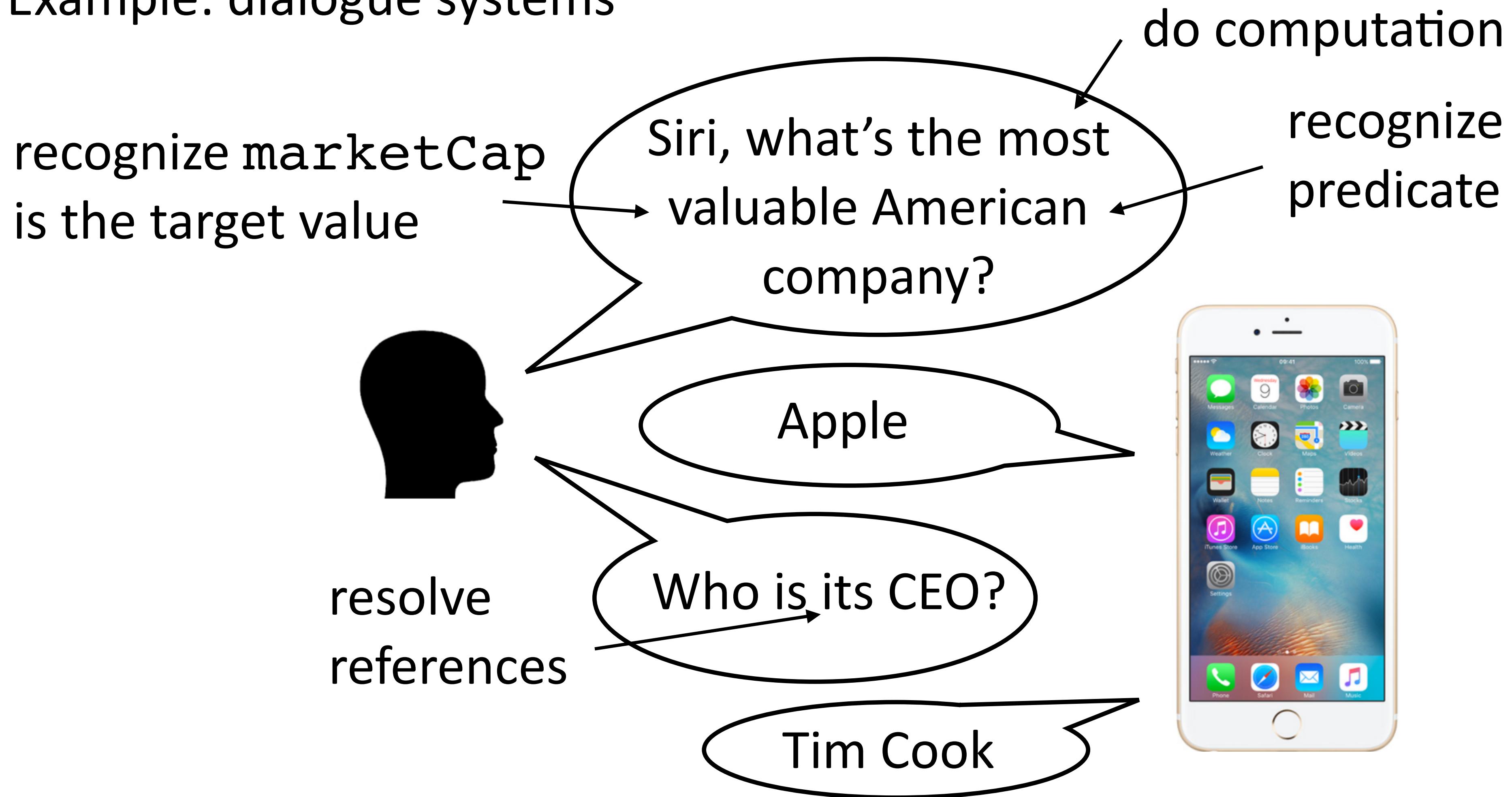
Programming Projects

- ▶ 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!)
- ▶ Most programming will be done with PyTorch library (can be tricky to debug)
- ▶ You will want to use a GPU
- ▶ Google Colab: free GPUs (some limitations; pro account for \$10/month)
- ▶ The programming projects are designed with Colab in mind



What's the goal of NLP?

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



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](#).

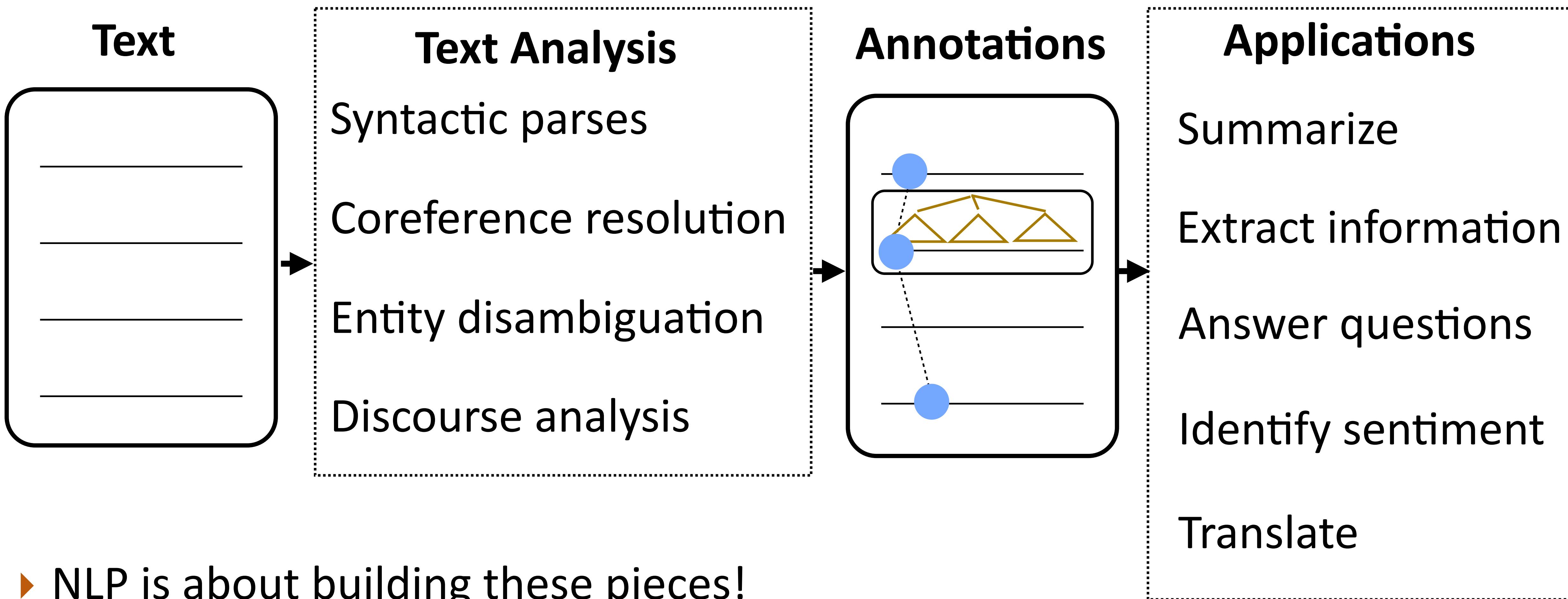
paraphrase to provide clarity

Machine Translation



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

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

Labels

the movie was good +

Beyoncé had one of the best videos of all time subjective

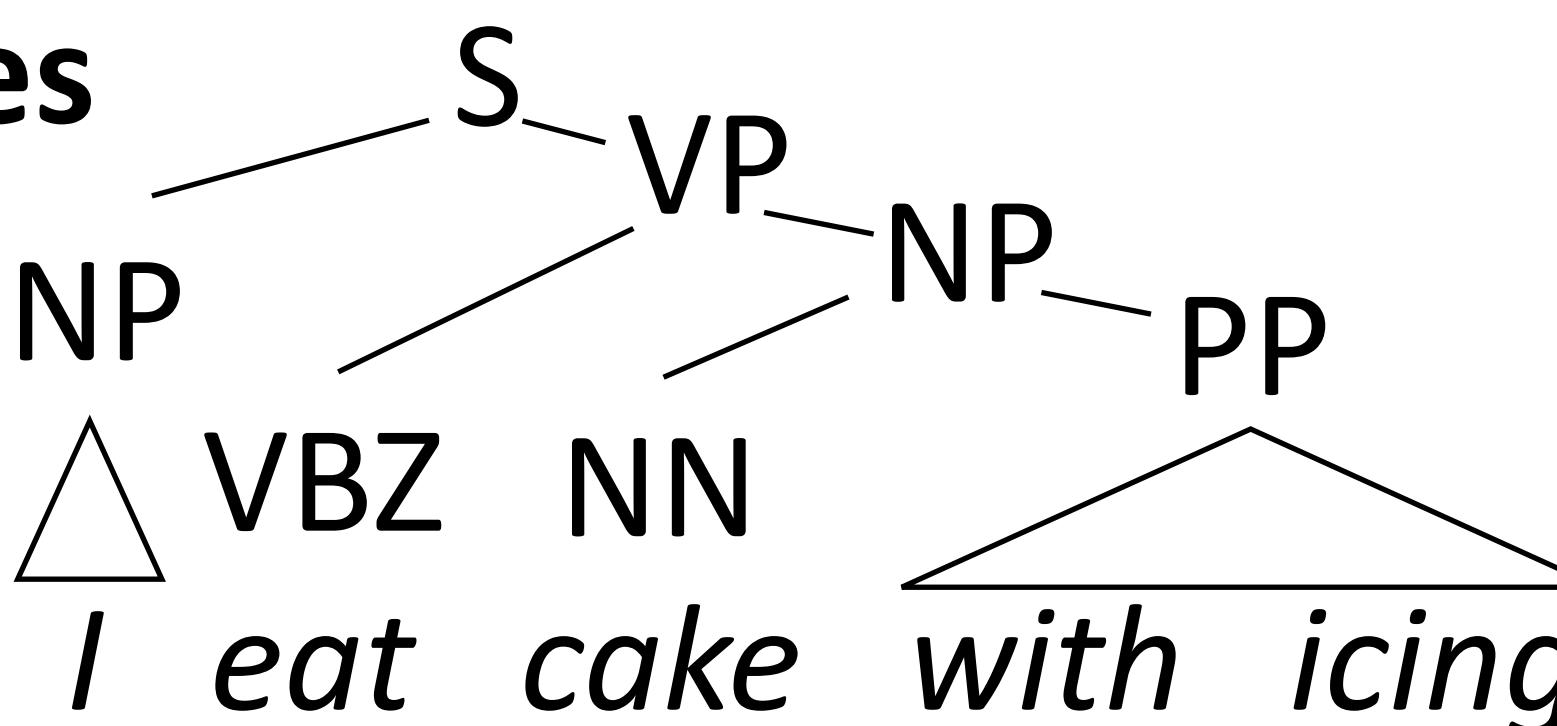
Sequences/tags

PERSON

Tom Cruise stars in the new Mission Impossible film

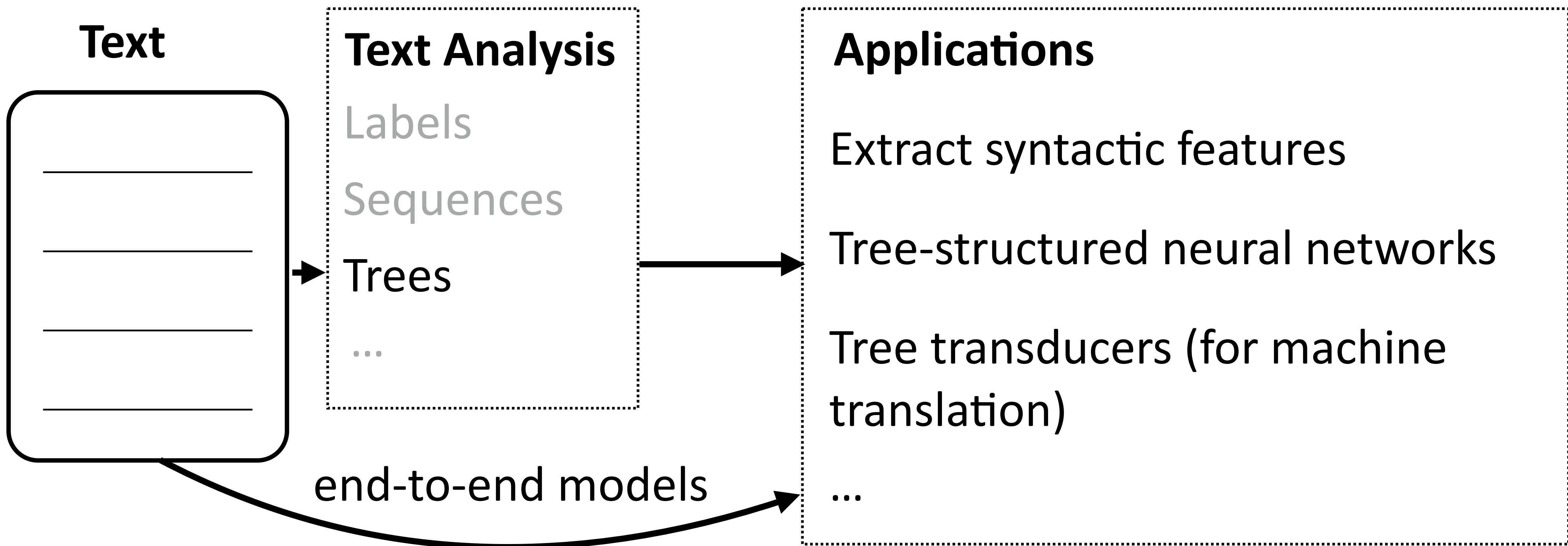
MOVIE

Trees



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

How do we use these representations?



- ▶ 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)

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!

- ▶ 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

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

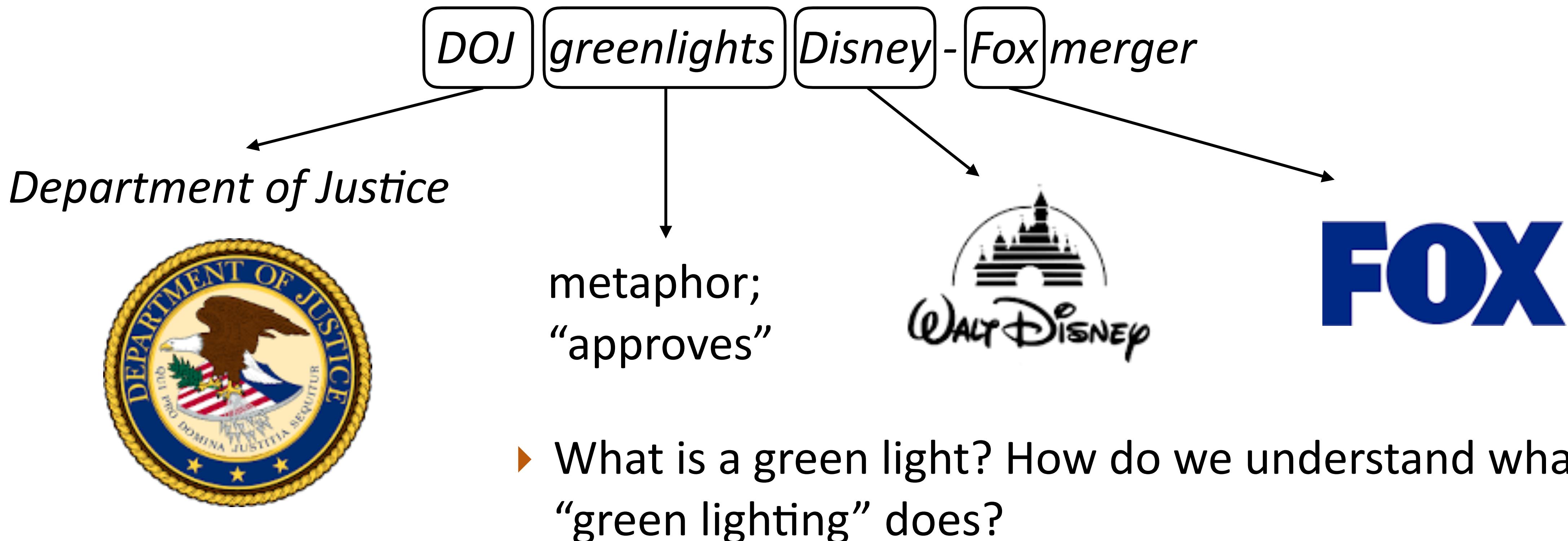
What do we need to understand language?

► 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?

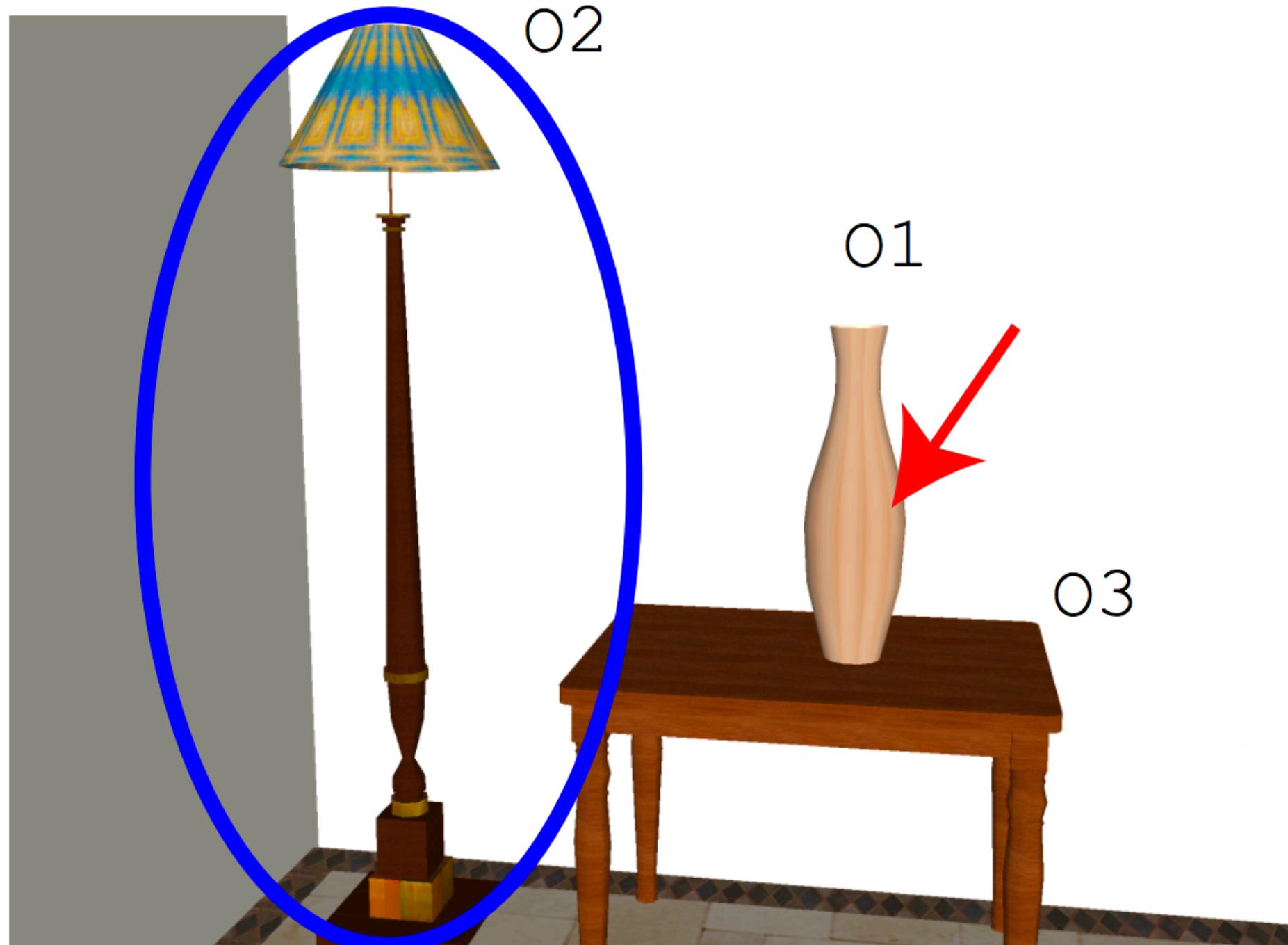
- ▶ World knowledge: have access to information beyond the training data



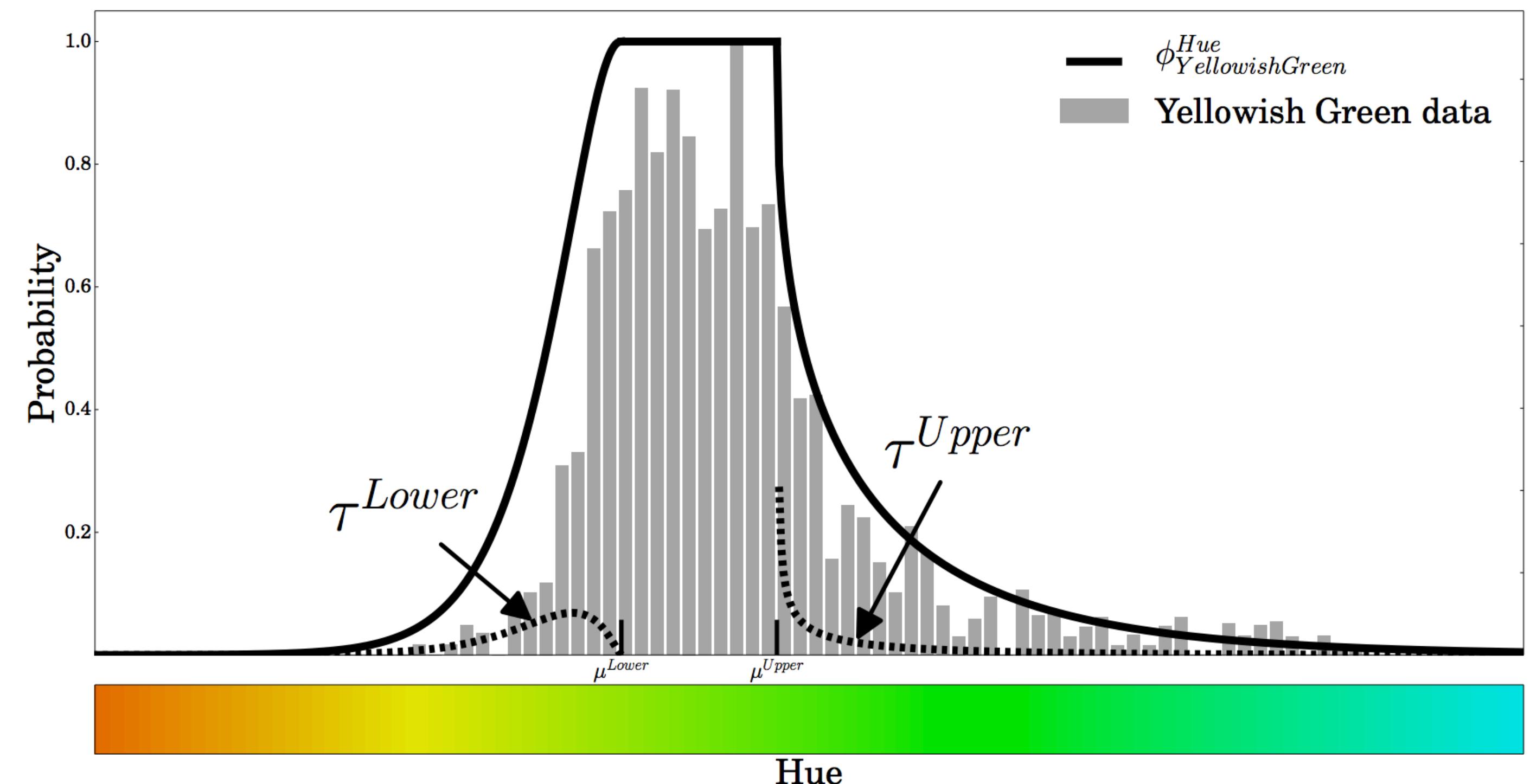
What do we need to understand language?

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

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



Golland et al. (2010)



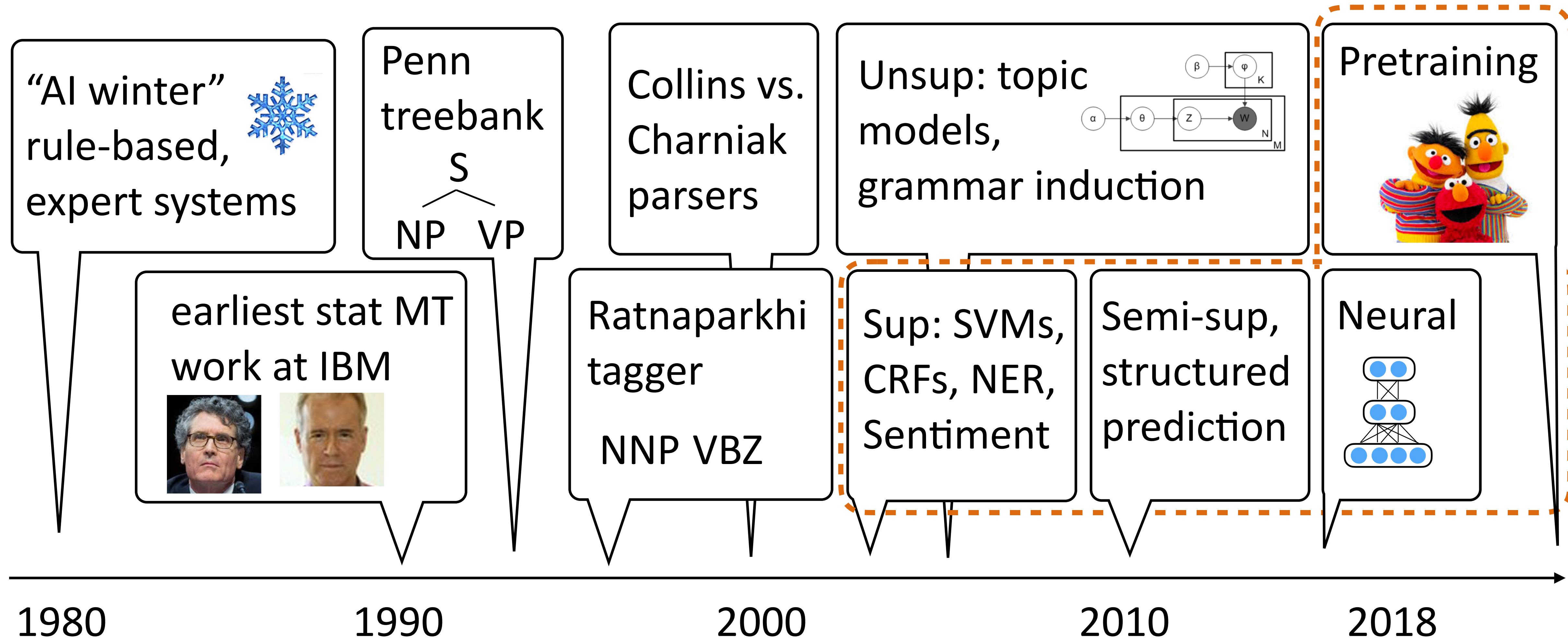
McMahan and Stone (2015)

What do we need to understand language?

- ▶ Linguistic structure
- ▶ ...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)

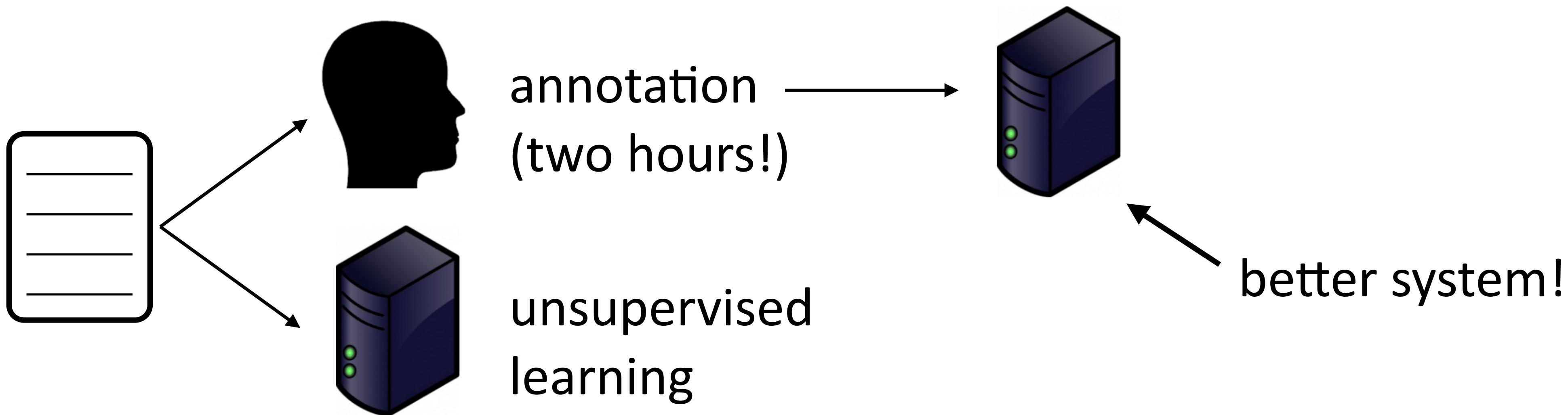
What techniques do we use?
(to combine data, knowledge, linguistics, etc.)

A brief history of (modern) NLP



Structured Prediction

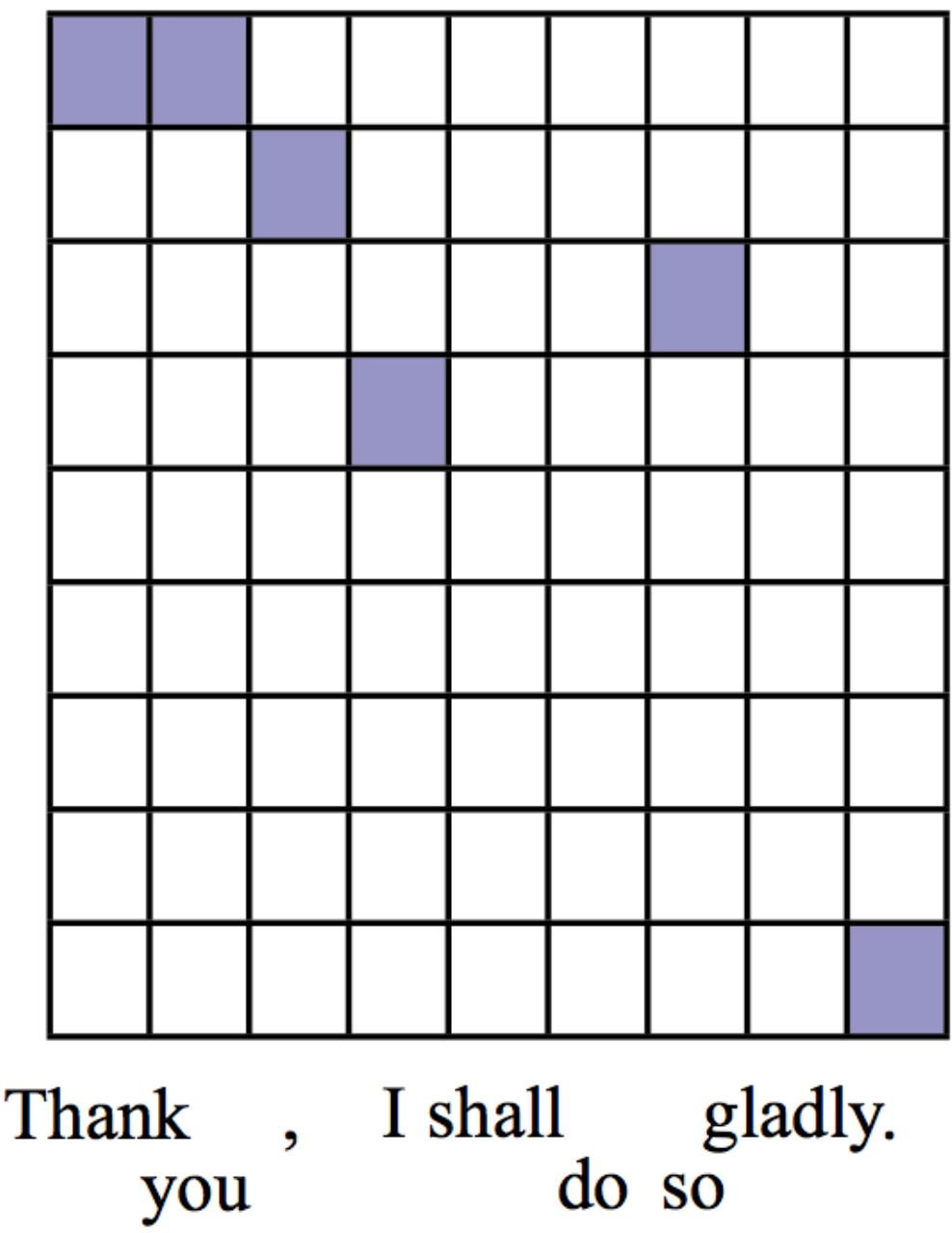
- ▶ All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- ▶ Supervised techniques work well on very little data



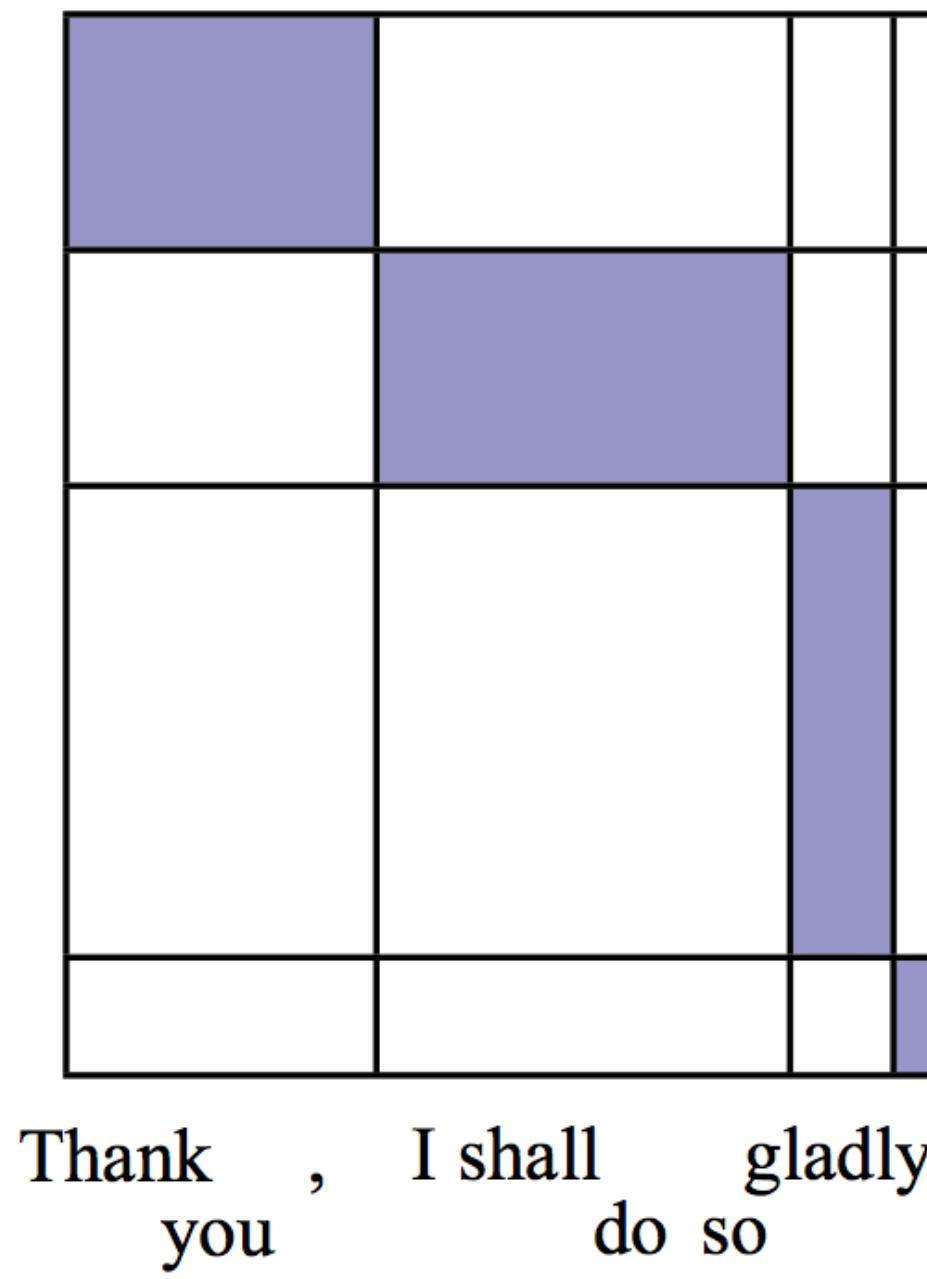
- ▶ Even neural nets can do pretty well!

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

Less Manual Structure?

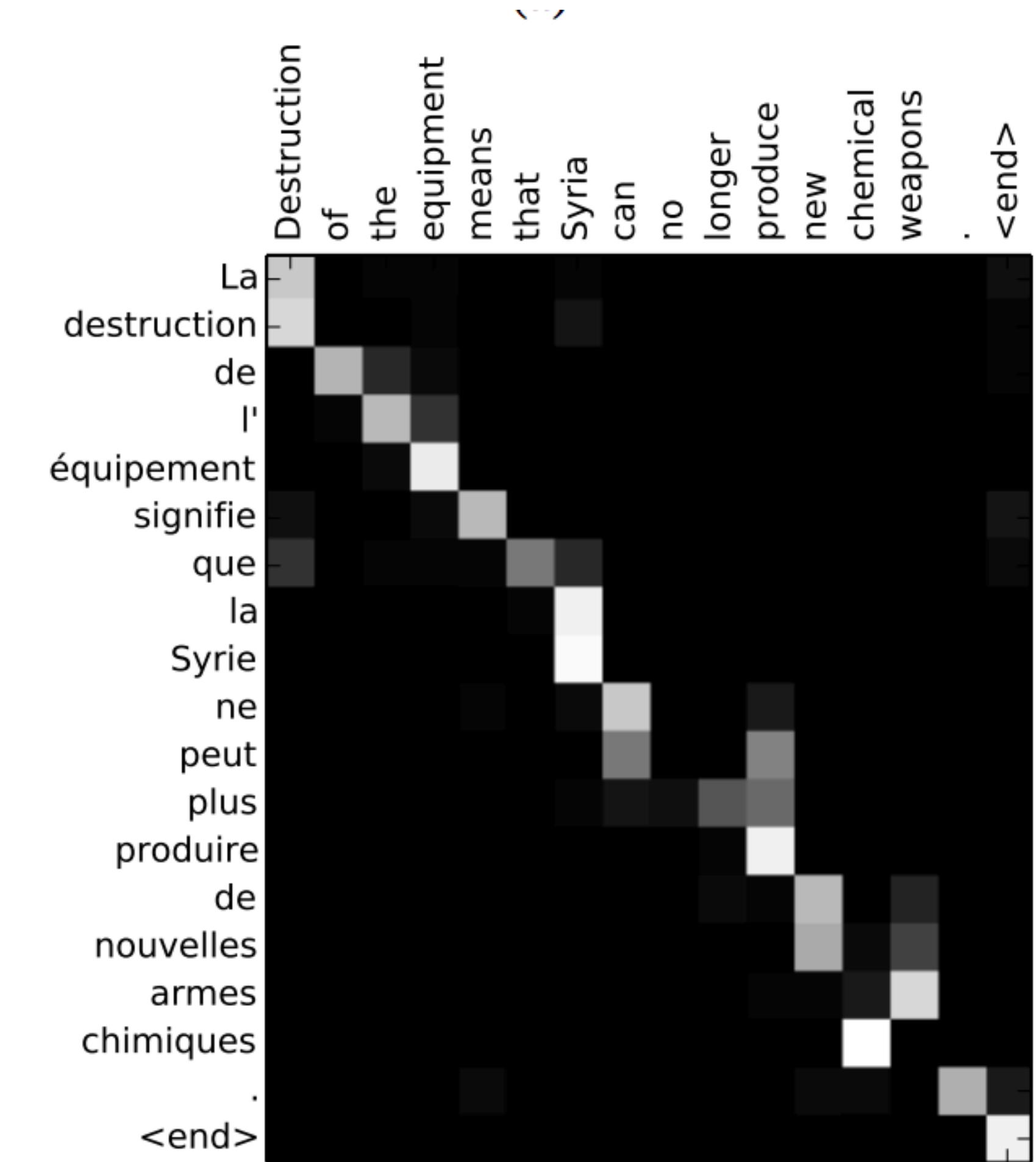


(a) example word alignment



(b) example phrase alignment

Gracias
,
lo
haré
de
muy
buen
grado
.



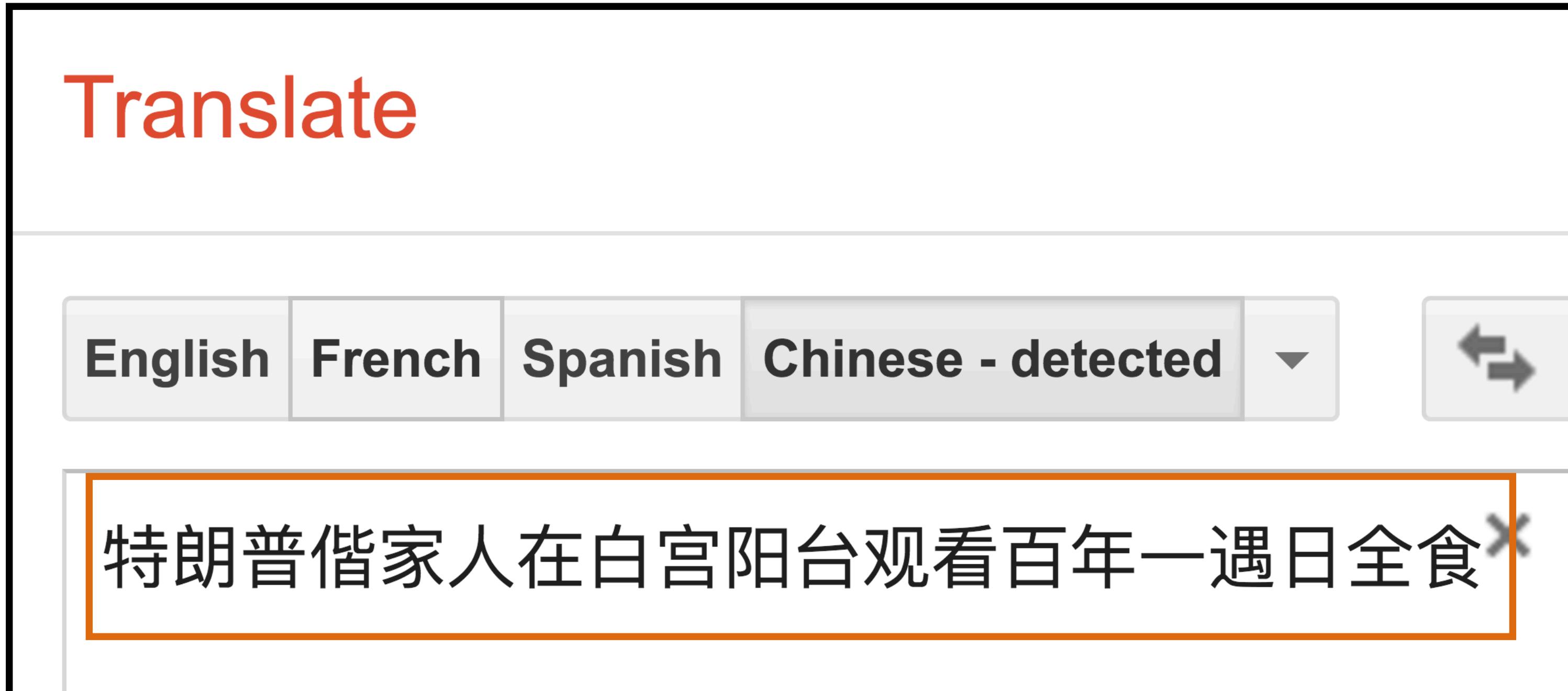
Does manual structure have a place?

- ▶ 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 Avg. F ₁ |
|-------------------------|------------------------------|
| Newswire | |
| rule-based | 55.60 |
| berkeley | 61.24 |
| cort | 63.37 |
| deep-coref [conll] | 65.39 |
| deep-coref [lea] | 65.60 |
| Wikipedia | |
| rule-based | 51.77 |
| berkeley | 51.01 |
| cort | 49.94 |
| deep-coref [conll] | 52.65 |
| deep-coref [lea] | 53.14 |
| deep-coref ⁻ | 51.01 |

Moosavi and Strube (2017)

Does manual structure have a place?



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

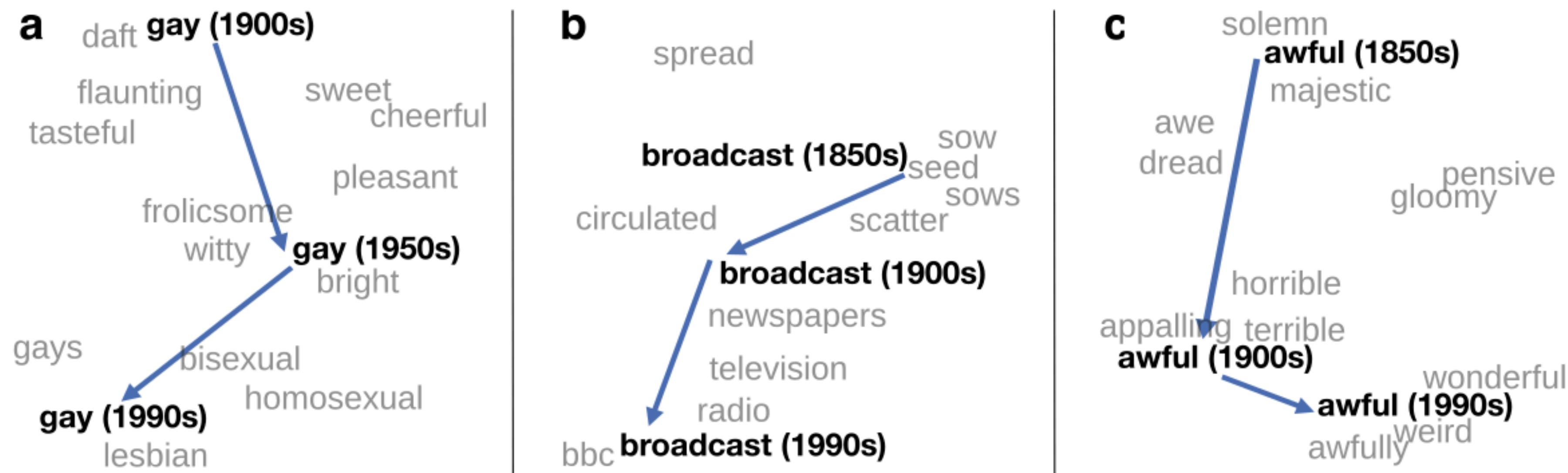
- ▶ Maybe manual structure would help...

Where are we?

- ▶ 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 use requires understanding dataset size, problem complexity, and a lot of tricks!
- ▶ NLP encompasses all of these things

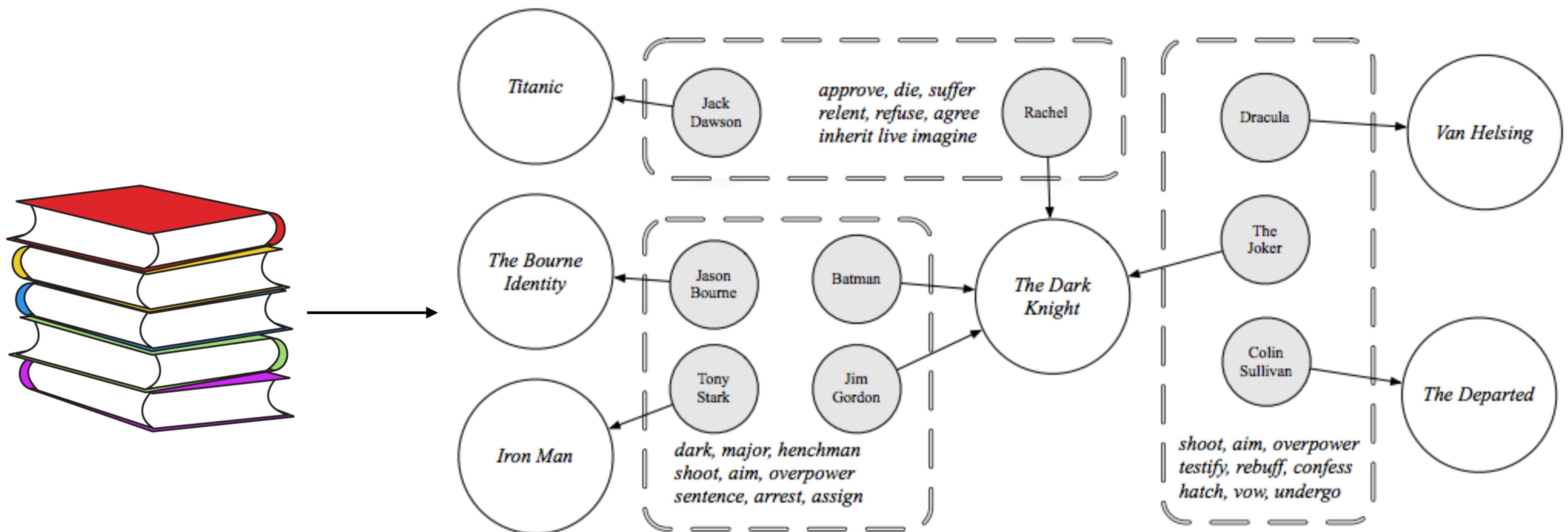
NLP vs. Computational Linguistics

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



NLP vs. Computational Linguistics

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



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 2021?
- ▶ 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 (40% grade)
- ▶ 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% grade)
 - ▶ Groups of 3-4 preferred, 1 is possible.
 - ▶ Good idea to talk to run your project idea by me in office hours or email.
 - ▶ 4 page report + final project presentation.

Questions?

Piazza — <https://piazza.com/class/ksjq7xenrbp3g5>