THE OHIO STATE UNIVERSITY

CSE 5525: Foundations of Speech and Language Processing

Lecture 1: Introduction

Huan Sun (CSE@OSU)

Many thanks to Prof. Greg Durrett @ UT Austin for sharing his slides.

Logistics

- Lectures (online): Wednesdays and Fridays 12:45pm - 2:05pm
- Course website: http://web.cse.ohio-state.edu/~sun.397/courses/au2020/cse5525.html
- Piazza: Signup link is on Carmen
- My office hours: Wednesdays 4:30 - 5:30PM & additional hours by appointment
- TA: Dingkang Wang; Office hours: Fridays 4:30 - 5:30PM

All Zoom links for lectures and office hours can be found on Carmen.

Course Requirements

- Prereq: (CSE 3521 or CSE 5521) and (CSE 5522 or Stat 3460 or Stat 3470). Not open to students with credit for CSE 733 (Prior course number).
- Python experience
- Additional prior exposure to probability, linear algebra, optimization, linguistics, and NLP useful but not required
- Textbooks:
  - Speech and Language Processing (3rd Edition), Dan Jurafsky and James H. Martin
  - Natural Language Processing, Jacob Eisenstein
  - Both are available online and linked on our website too!

Grading Plan

- Participation (10%)
- Homework (50%)
  - HW #1 is out! If you find it challenging, please email or talk to me (this is smaller-scale and relatively easier, compared with other HWs/Project or midterm exam)
- Midterm Exam (20%): On OCT 14th, 2020
- Final Project (20%)
  - Form teams (2-3 people with Diverse* Background) from now!
  - *: e.g., A group of mixed undergrads and grads (e.g., 1 grad + 1 undergrads, or, 2 grads + 1 undergrad) are expected! Exceptions can be made, i.e., 2 grads with distinct research backgrounds are also OK (but talk to the instructor first); 2 grads with both AI/NLP majors are not OK.

Assignments

- 3 Assignments
  - Implementation-oriented, with one or more open-ended components to each
  - HW #1 (classification) is out NOW
  - No late assignments

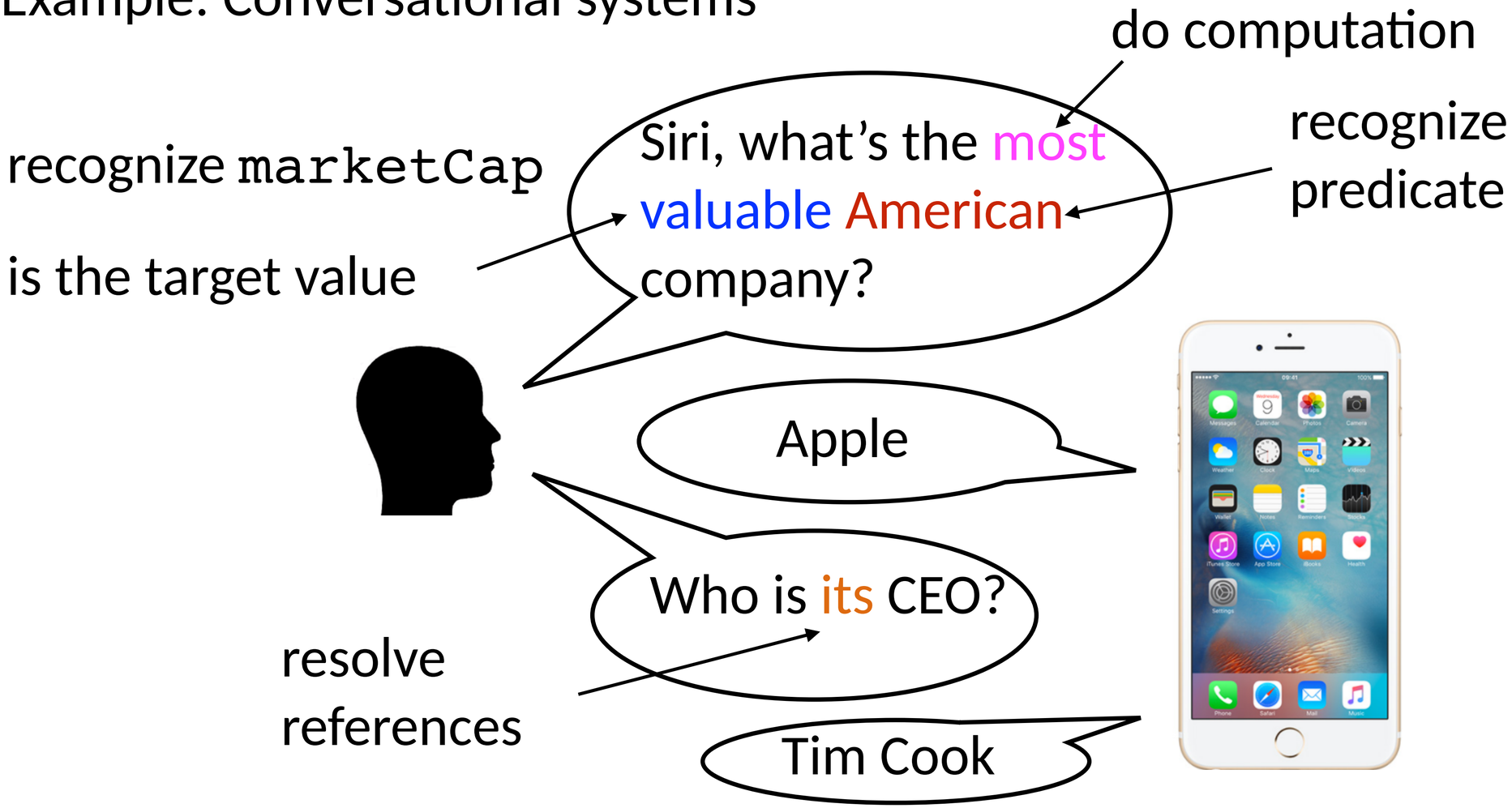
These projects require understanding of the concepts, ability to write performant code, and ability to think about how to debug complex systems. **They are challenging, so start early!**

Assignments

- Final project (20%)
  - Form a group of 2-3 people with diverse background
    - e.g., 1 grad + 1-2 undergrads (start forming teams from now)
  - (Brief! 1-page) proposal to be approved by the instructor by the midpoint of the semester (10/21/2020)
    - Start thinking about topics to work on from now
  - Due at the end of the semester:
    - A roughly 4-page report (Written in the style and tone of an ACL paper; more detailed requirements to come)
    - Final project presentation (~ 5 min for each group depending on how many groups)

What's the goal of NLP?

- Be able to solve problems that require deep understanding of text
- Example: Conversational 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.

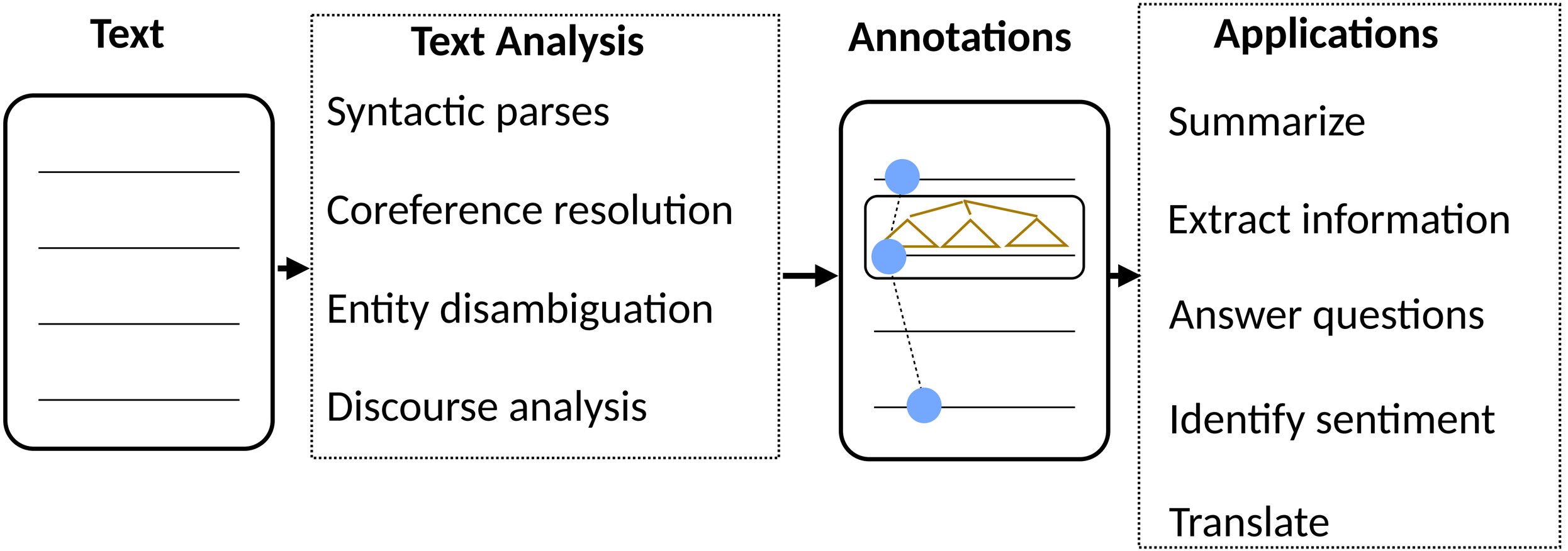
paraphrase to provide clarity

Machine Translation

People's Daily, August 30, 2017

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 can be modeled with statistical approaches trained using machine learning

How do we represent language?

Text

Labels

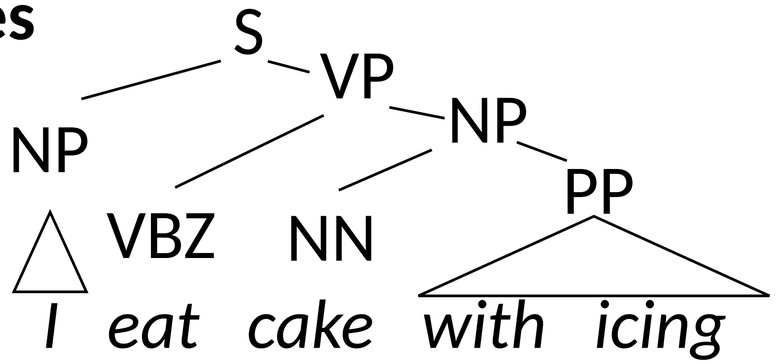
the movie was good +

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

Sequences/tags

PERSON *Tom Cruise* *stars in the new* **WORK_OF_ART** *Mission Impossible film*

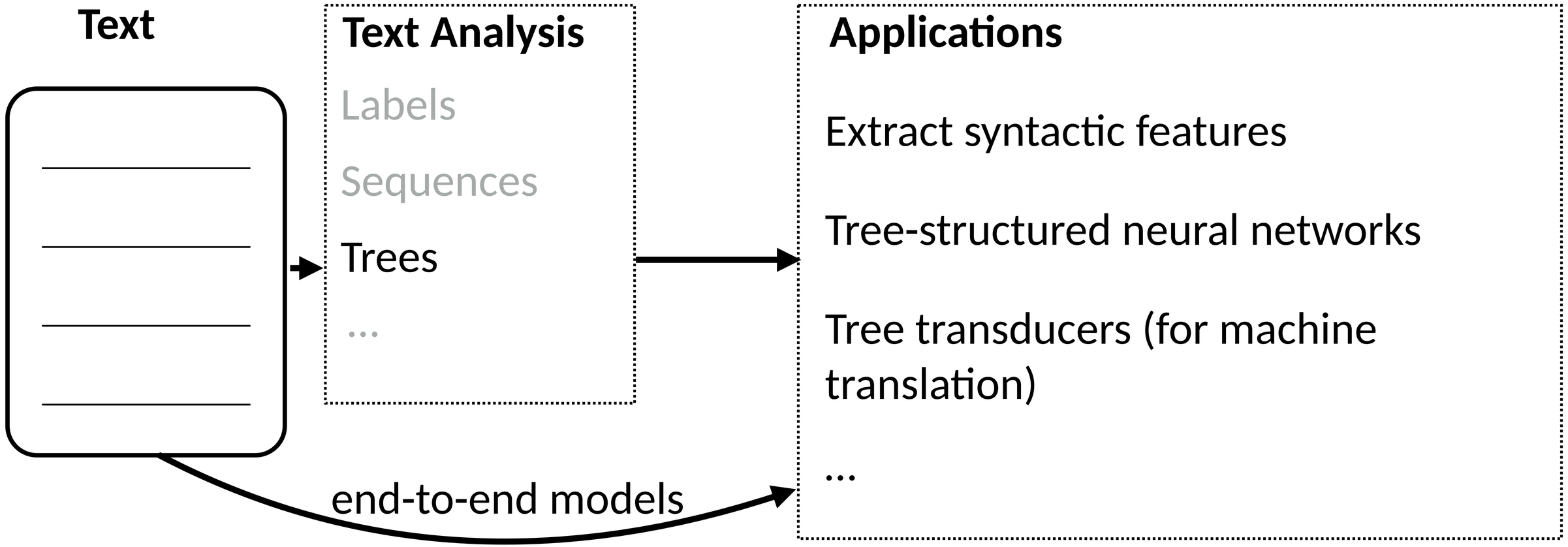
Trees



$\lambda x.\ flight(x) \wedge dest(x)=Miami$

flights to Miami

How do we use these representations?



- Main questions: 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 advocated violence

The city council refused the demonstrators a permit because they feared violence

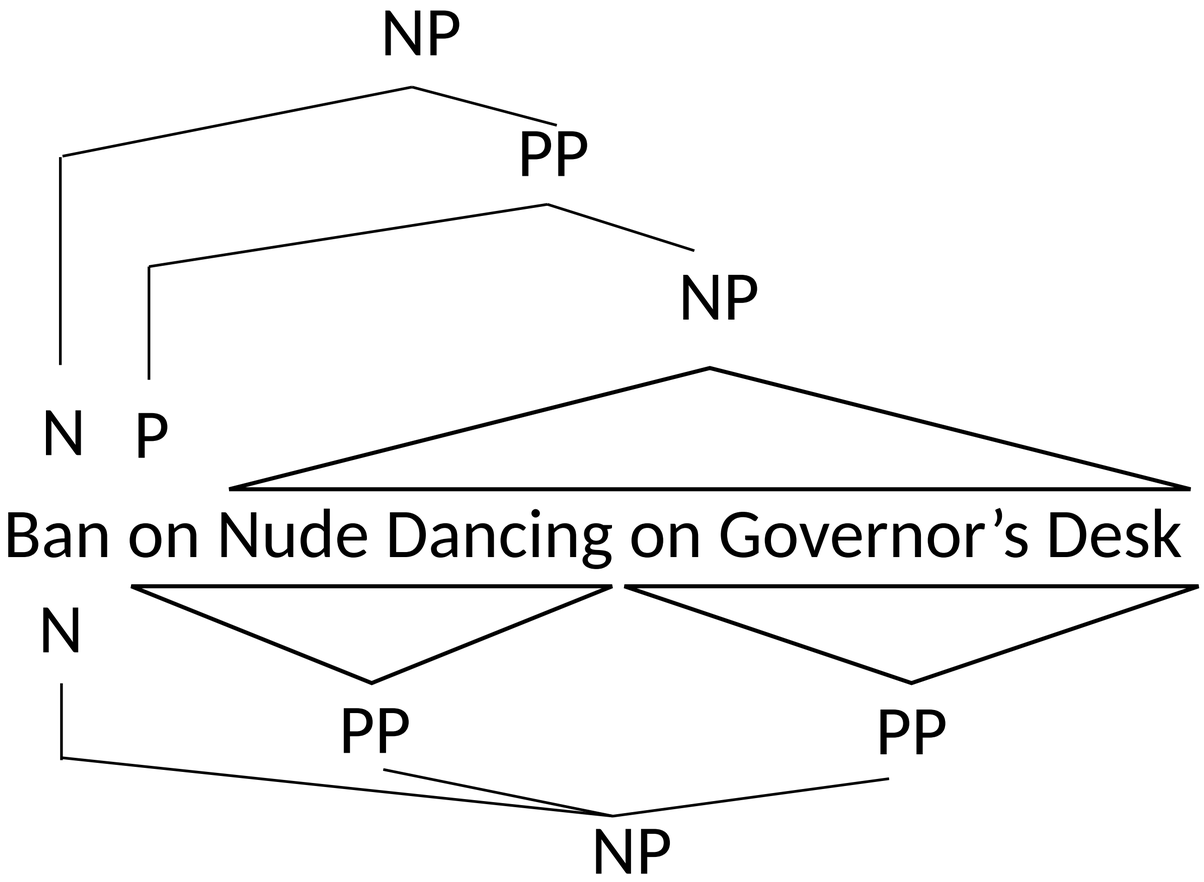
The city council refused the demonstrators a permit because they _____ violence

- >5 datasets in the last two years examining this problem and commonsense reasoning
- Referential ambiguity

Language is Ambiguous!

N N V N
N V ADJ N
Teacher Strikes Idle Kids

body/ body/
position weapon
Iraqi Head Seeks Arms



- Syntactic and semantic ambiguities: parsing needed to resolve these, but need context to figure out which parse is correct

example credit: Dan Klein

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

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] [.]

slide credit: Dan Klein

What do we need to understand language?

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

DOJ *greenlights* *Disney* - *Fox* *merger*

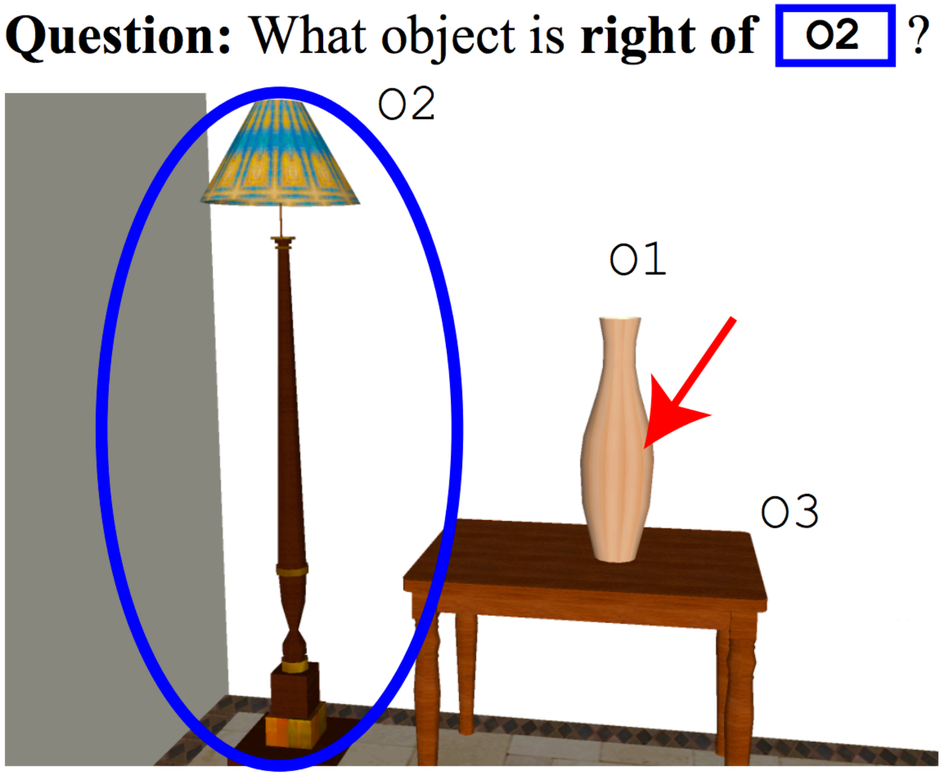
Department of Justice

metaphor;
"approves"

- What is a green light? How do we understand what "green lighting" does?
- Need commonsense knowledge

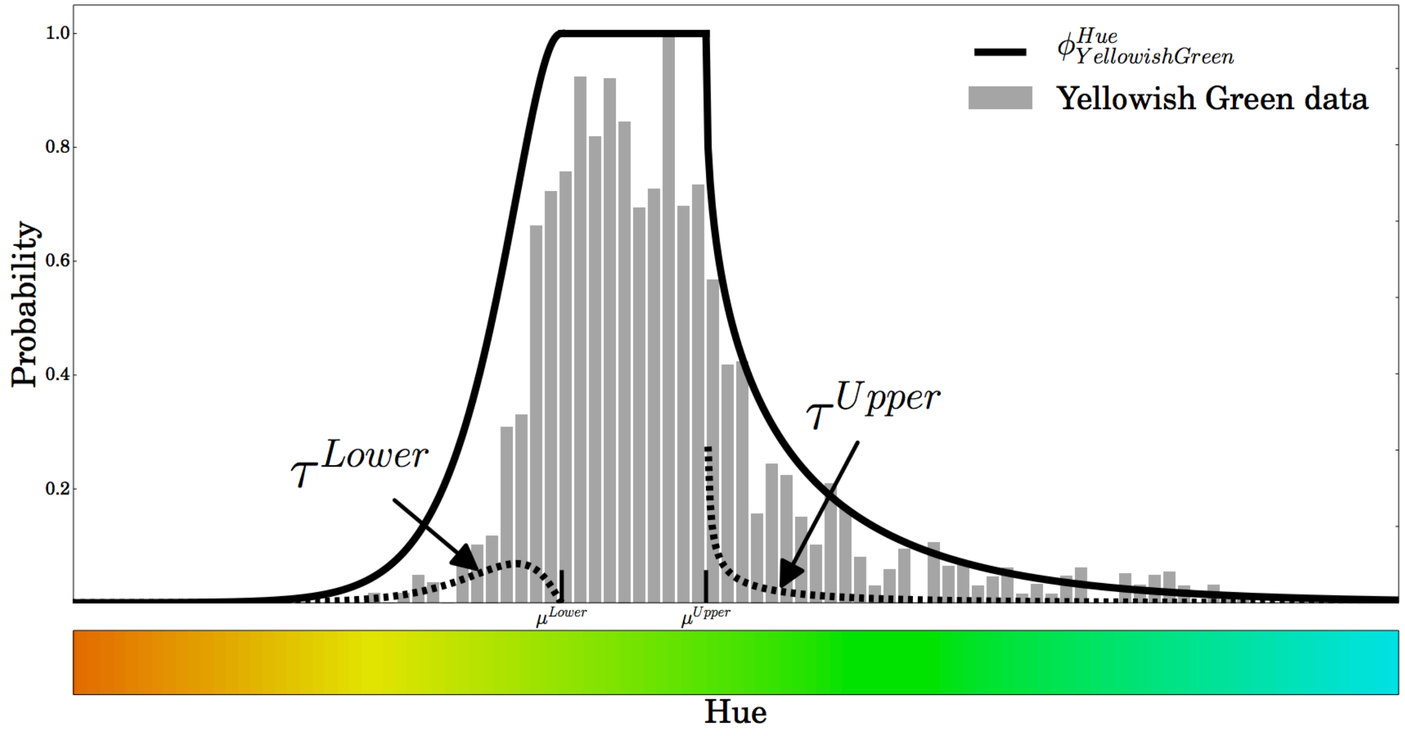
What do we need to understand language?

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



Golland et al. (2010)

https://www.aclweb.org/anthology/D10-1040.pdf



McMahan and Stone (2015)

https://www.aclweb.org/anthology/Q15-1008.pdf

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 = John; C_f = {John}

c. He called up Mike yesterday to work out a plan. (he = John)
C_b = John; C_f = {John, Mike} (CONTINUE)

d. Mike has annoyed him a lot recently.
C_b = John; C_f = {Mike, John} (RETAIN)

e. He called John at 5 AM on Friday last week. (he = Mike)
C_b = Mike; C_f = {Mike, John} (SHIFT)

Centering Theory

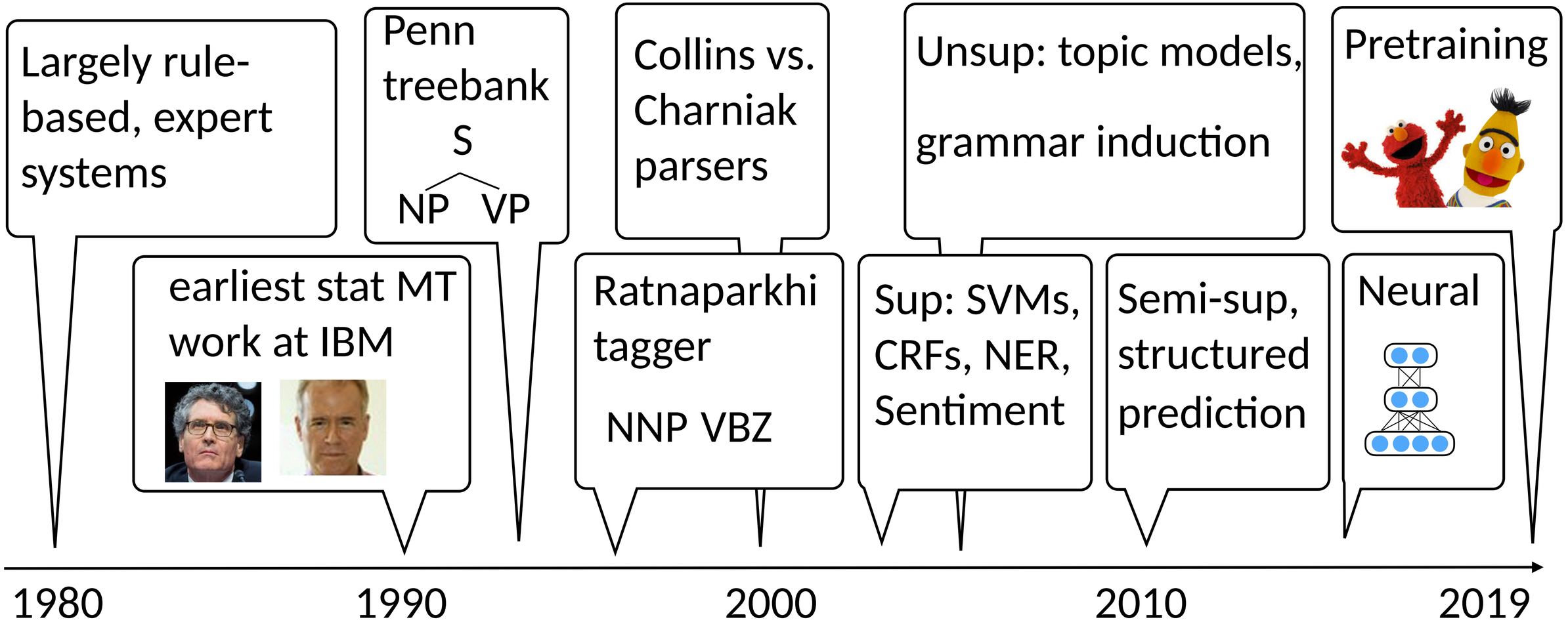
Grosz et al. (1995)

https://www.aclweb.org/anthology/J95-2003.pdf

What techniques do we use?

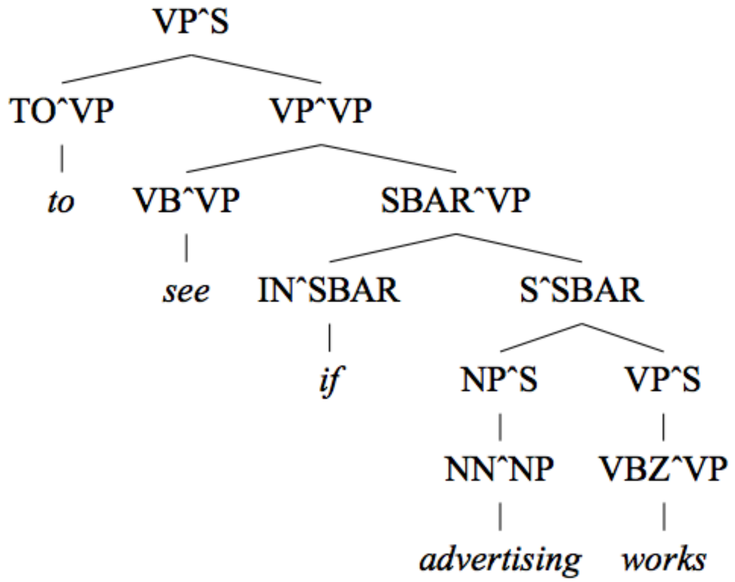
(to combine data, knowledge, linguistics, etc.)

A brief history of (modern) NLP



Less Manual Structure

- Training is supervised but models still rely less on manual structure

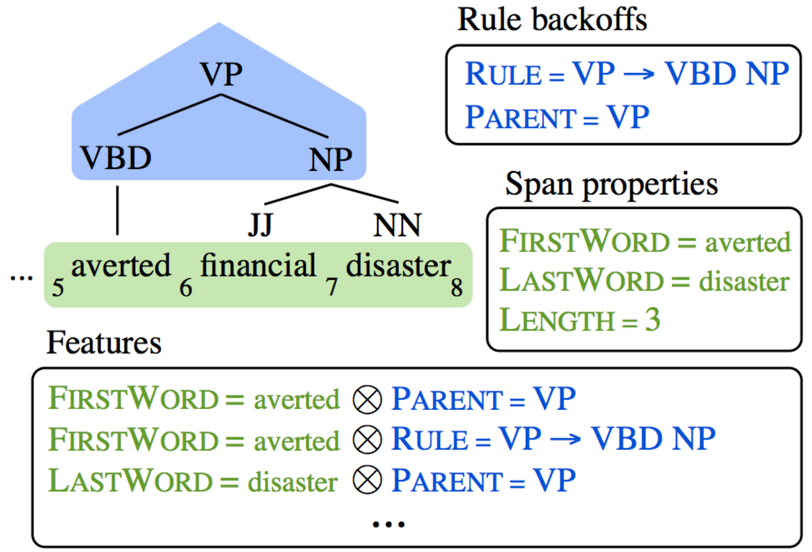


Klein and Manning (2003)
Manually-constructed grammars

VBZ

VBZ-0	gives	sells	takes
VBZ-1	comes	goes	works
VBZ-2	includes	owns	is
VBZ-3	puts	provides	takes
VBZ-4	says	adds	Says
VBZ-5	believes	means	thinks
VBZ-6	expects	makes	calls
VBZ-7	plans	expects	wants
VBZ-8	is	's	gets
VBZ-9	's	is	remains
VBZ-10	has	's	is
VBZ-11	does	Is	Does

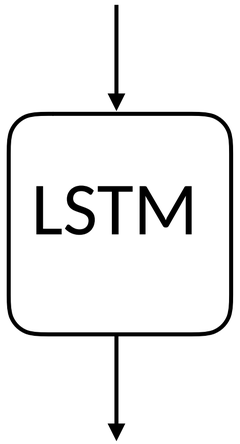
Petrov et al. (2006)
Induced grammars



Hall, Durrett, Klein (2014)
Basic grammar + features

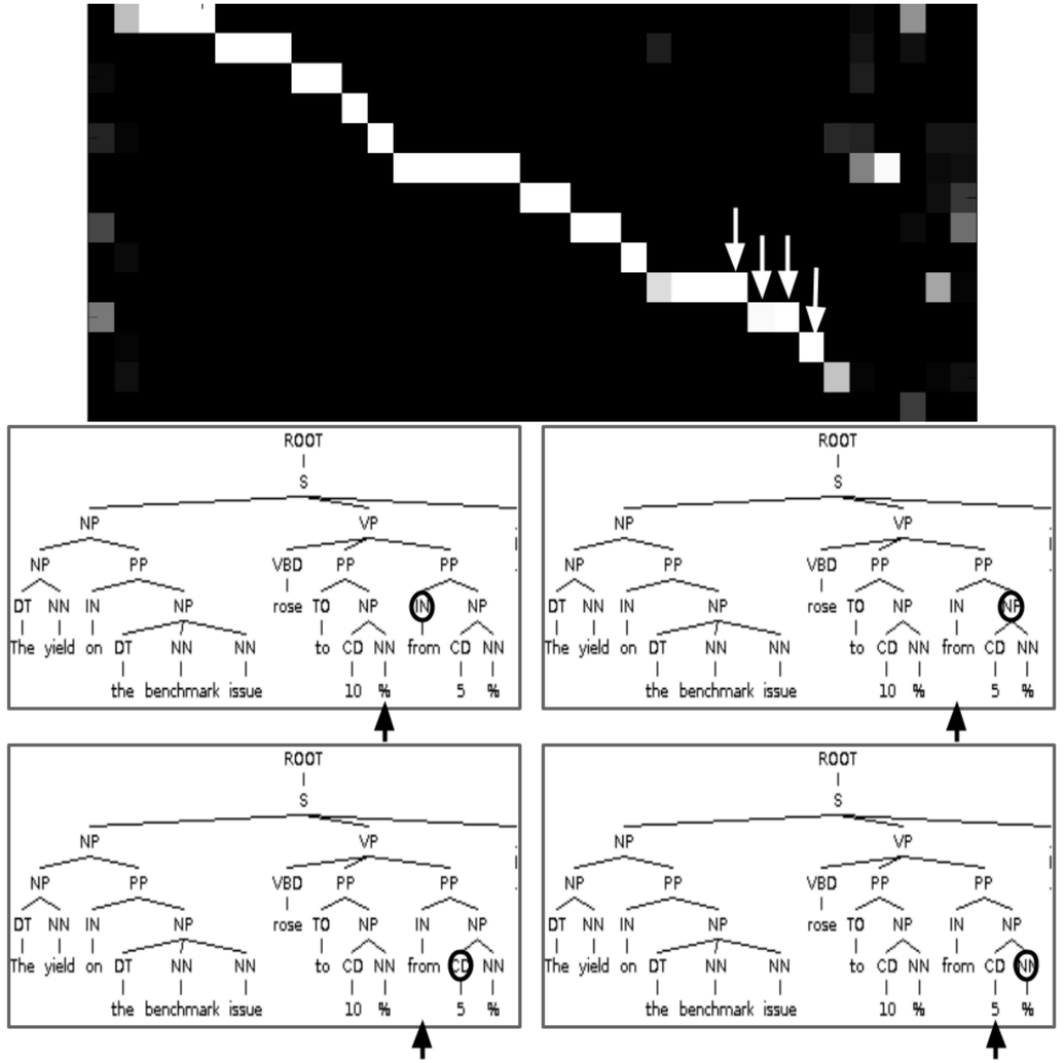
Less Manual Structure

"The yield on the benchmark issue rose to 10% from 5%"



(S (NP (NP (DT The) (NN yield …

- No grammars at all!



Sutskever et al. (2015), Bahdanau et al. (2014)

Interpretability

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

- Hard to analyze why these errors happen in neural models (but people are trying)
- Models with more manual structure might be more interpretable

Pretraining

- Language modeling: predict the next word in a text $P(w_i|w_1, \ldots, w_{i-1})$

  P(*w* | I want to go to) = 0.01 Hawai'i

  0.005 LA

  0.0001 class

: use this model for other purposes

P(*w* | the acting was horrible, I think the movie was) = 0.1 bad

0.001 good

- Model understands some sentiment?

- Train a neural network to do language modeling on massive unlabeled text, fine-tune it to do {tagging, sentiment, question answering, …}

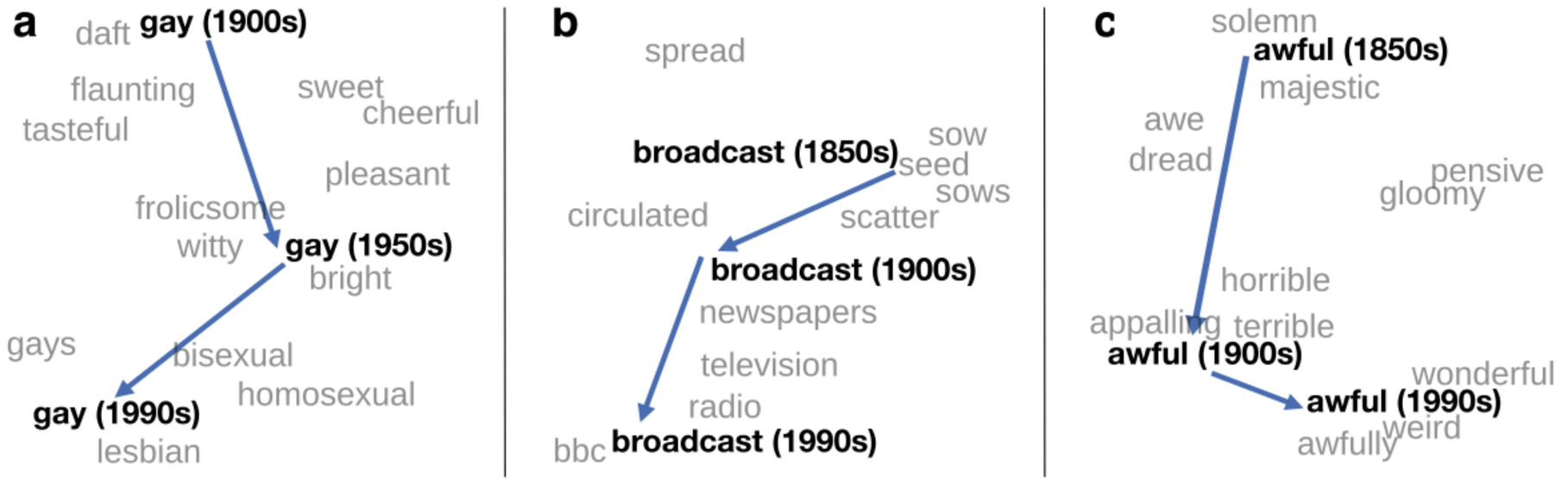
Peters et al. (2018), Devlin et al. (2019)

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

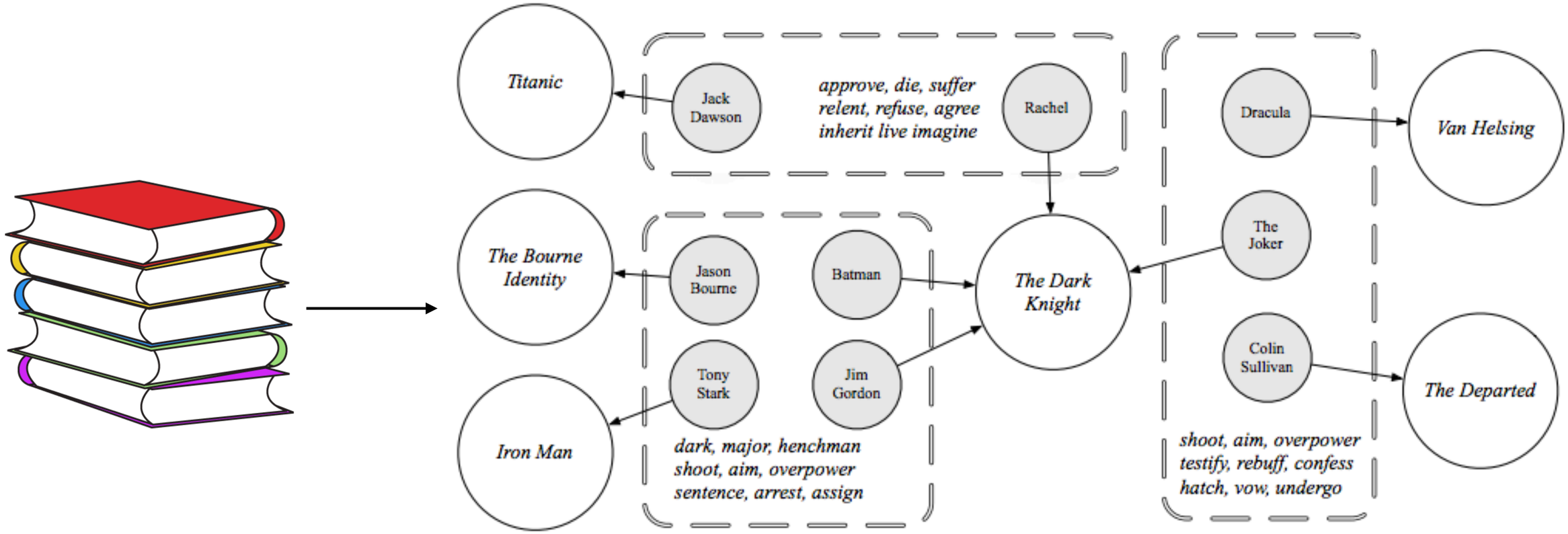
- NLP: build systems that deal with language data
- CL: use computational tools to study language



Hamilton et al. (2016)

NLP vs. Computational Linguistics

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



Bamman, O'Connor, Smith (2013)

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 the last 3 years?
- Make you a “producer” rather than a “consumer” of NLP tools
  - The 3 assignments should teach you what you need to know to understand nearly any system in the literature

Preview what to come

- 3 Assignments
  - HW #1 (classification) is out NOW and due on 09/09/2020! Start early!
  - HW #2 and #3 would be like implementing Conditional Random Fields and Neural Networks

Similar courses and HWs/Final Project Requirements can be found here:

https://www.cs.utexas.edu/~gdurrett/courses/fa2019/cs388.shtml

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!**