5525: Speech and Language Processing

Wei Xu

(many slides from Greg Durrett)

Administrivia

- Course website: https://cocoxu.github.io/courses/5525_spring20.html
- Piazza: link on the course website
- My office hours: Thursday 4:30-5:30pm DL 495
- TA: Ashutosh Baheti;
- TA Office hours: Tuesday 4:00-5:00pm DL 190

Course Requirements

- Probability (e.g., conditional probabilities, Bayes Rule, etc.)
- Linear Algebra (e.g., multiplying vectors and matrices, matrix inversion)
- Calculus (e.g., calculating gradients of functions with several variables)
- Programming / Python experience
- Prior exposure to machine learning algorithms very helpful

There will be a lot of math and programming!

Enrollment

- Homework 1 is out now (due in 2 weeks):
  - Please look at the assignment well before then
  - If this seems like it'll be challenging for you, come and talk to me (this is smaller-scale than the later assignments, which are smaller-scale than the final project)

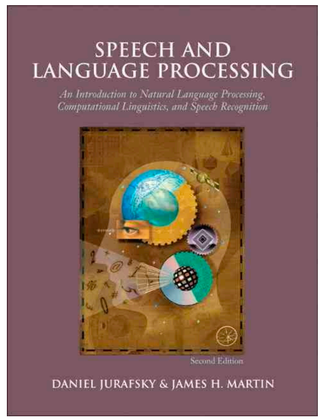
Texts

- Two great textbooks for NLP
  - There will be assigned readings from both
  - Both freely available online

Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

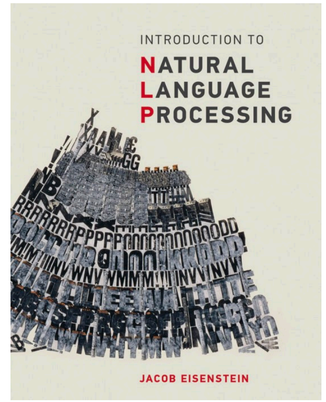
Draft chapters in progress, October 16, 2019



Natural Language Processing

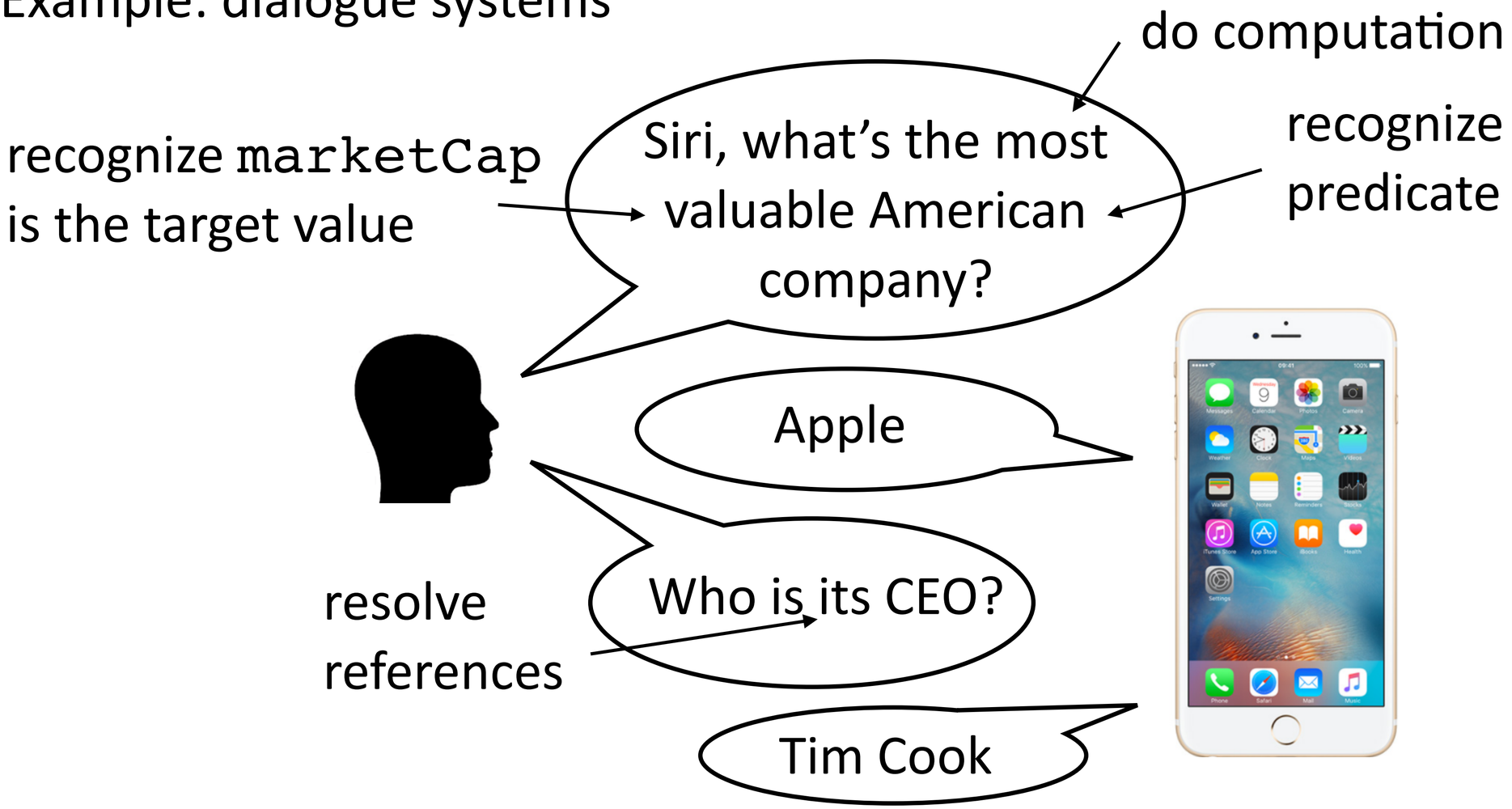
Jacob Eisenstein

November 13, 2018



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.

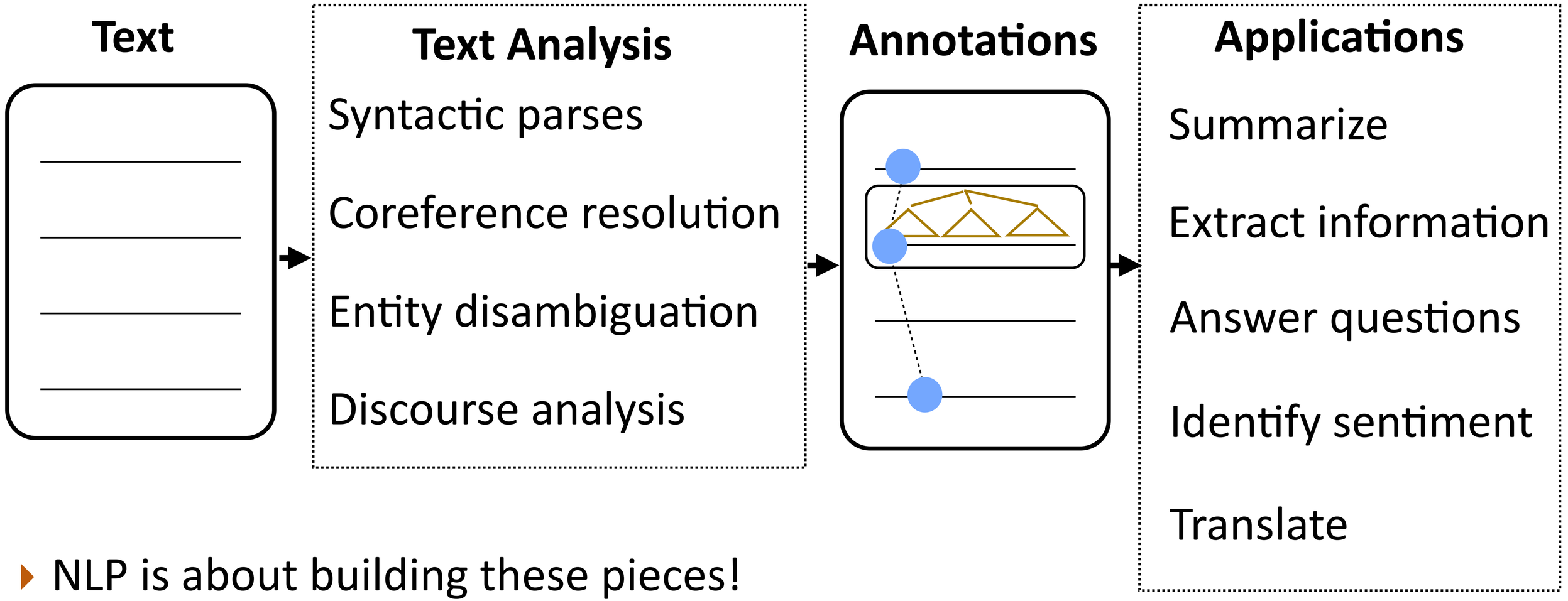
paraphrase to provide clarity

Machine Translation

People's Daily, August 30, 2017

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

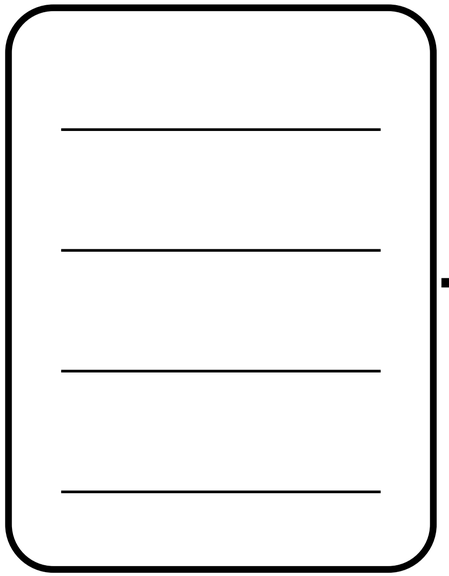
NLP Analysis Pipeline



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

How do we represent language?

Text



Labels

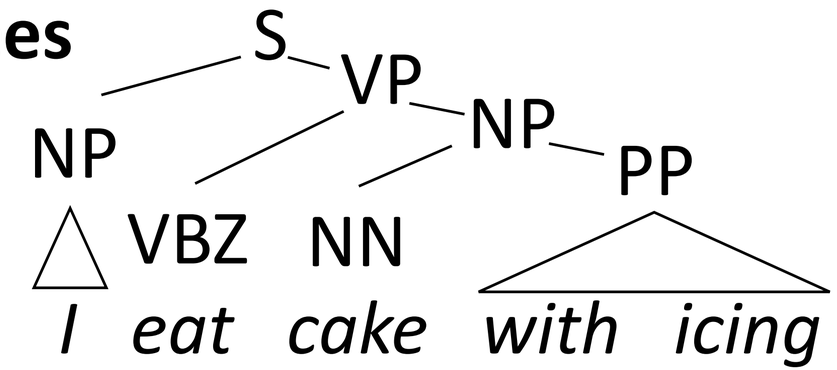
the movie was good +

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

Sequences/tags

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

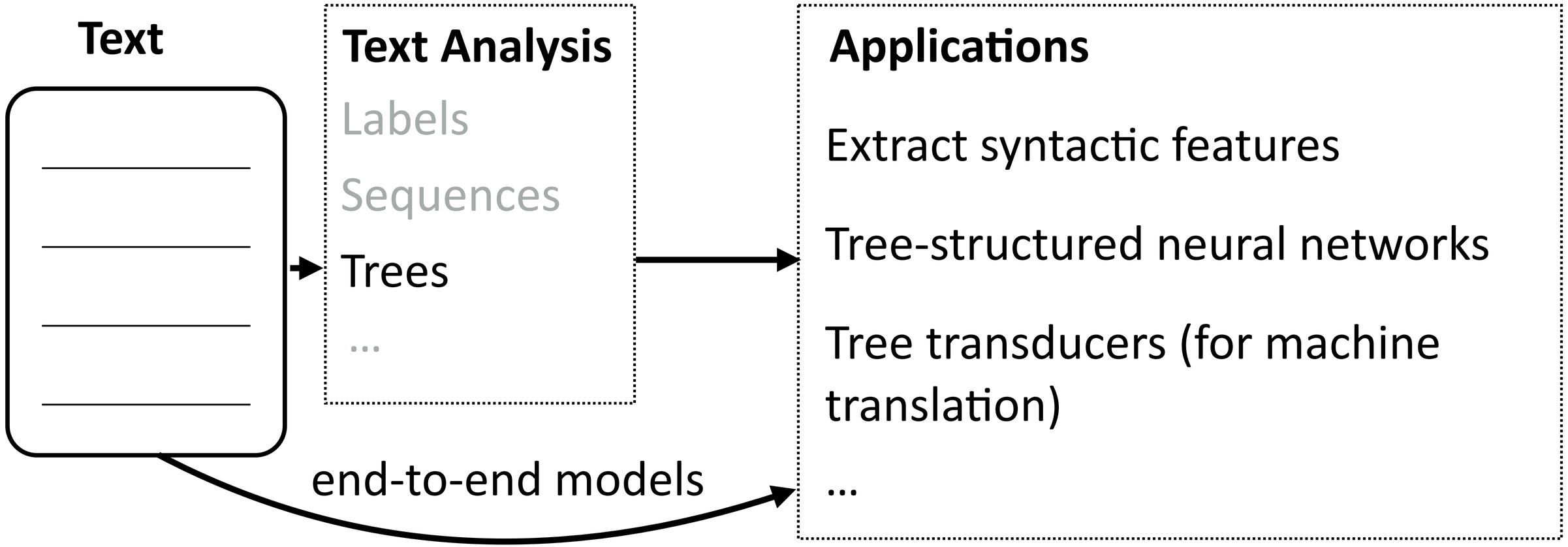
Trees



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

flights to Miami

How do we use these representations?



- Main 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!

- 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

slide 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

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?

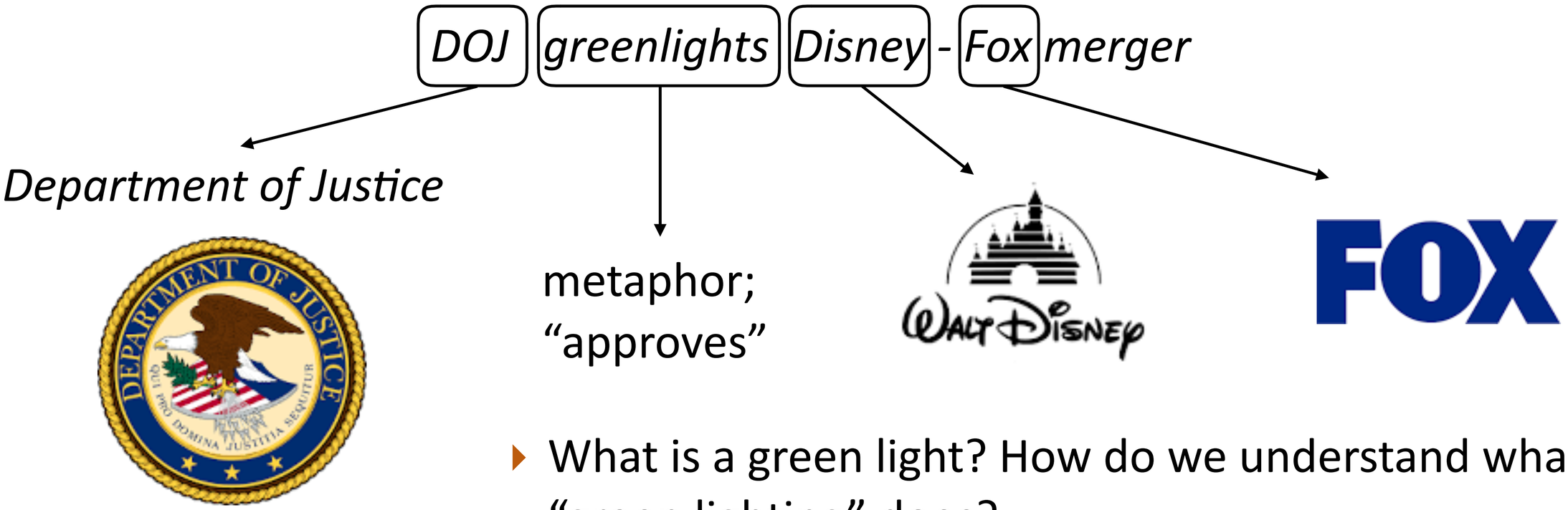
- Lots of data!

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] [.]

slide credit: Dan Klein

What do we need to understand language?

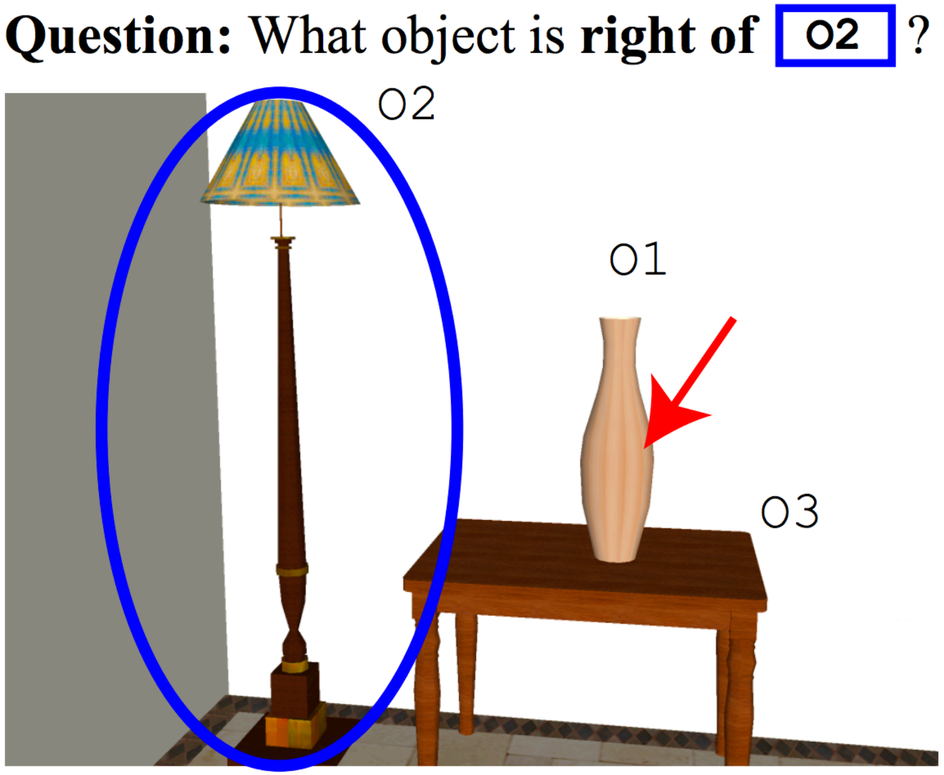
- World knowledge: have access to information beyond the training data



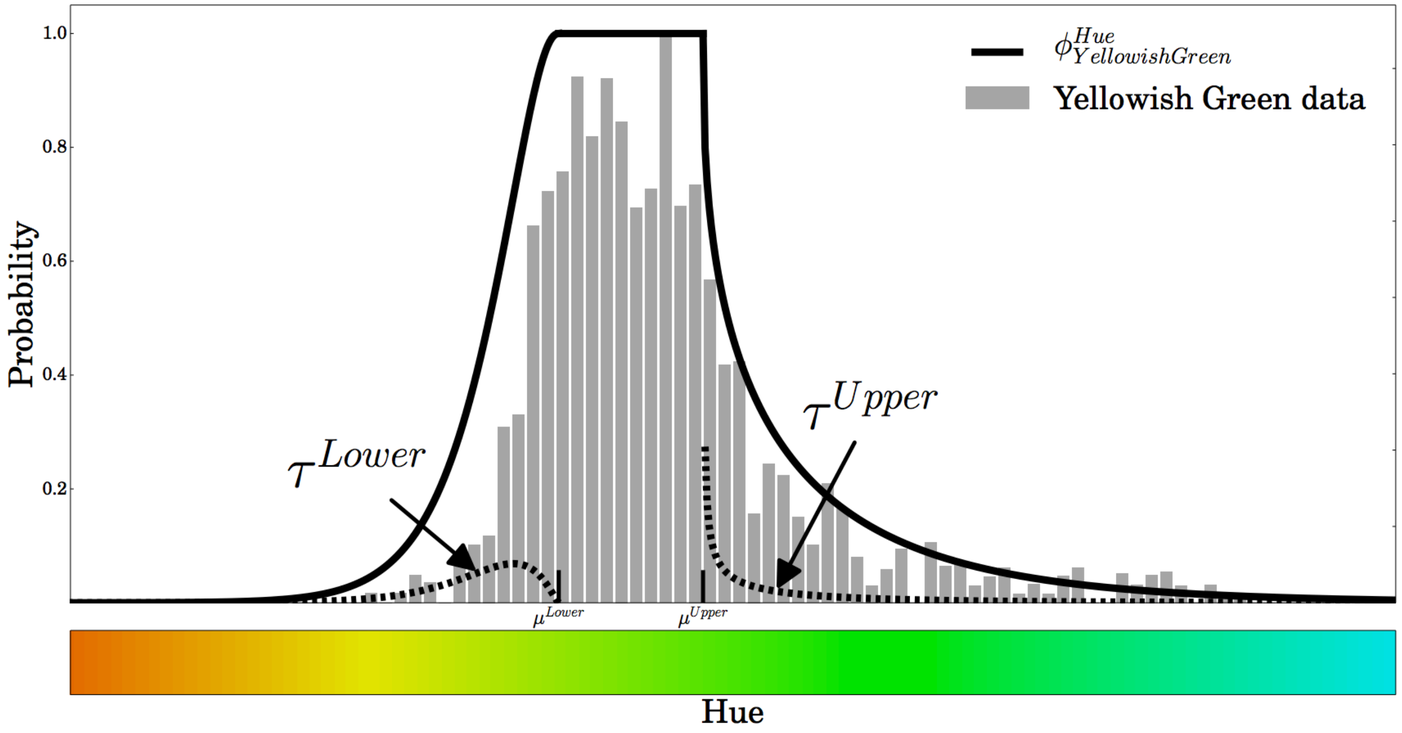
- What is a green light? How do we understand what “green lighting” does?

What do we need to understand language?

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



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 = 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)

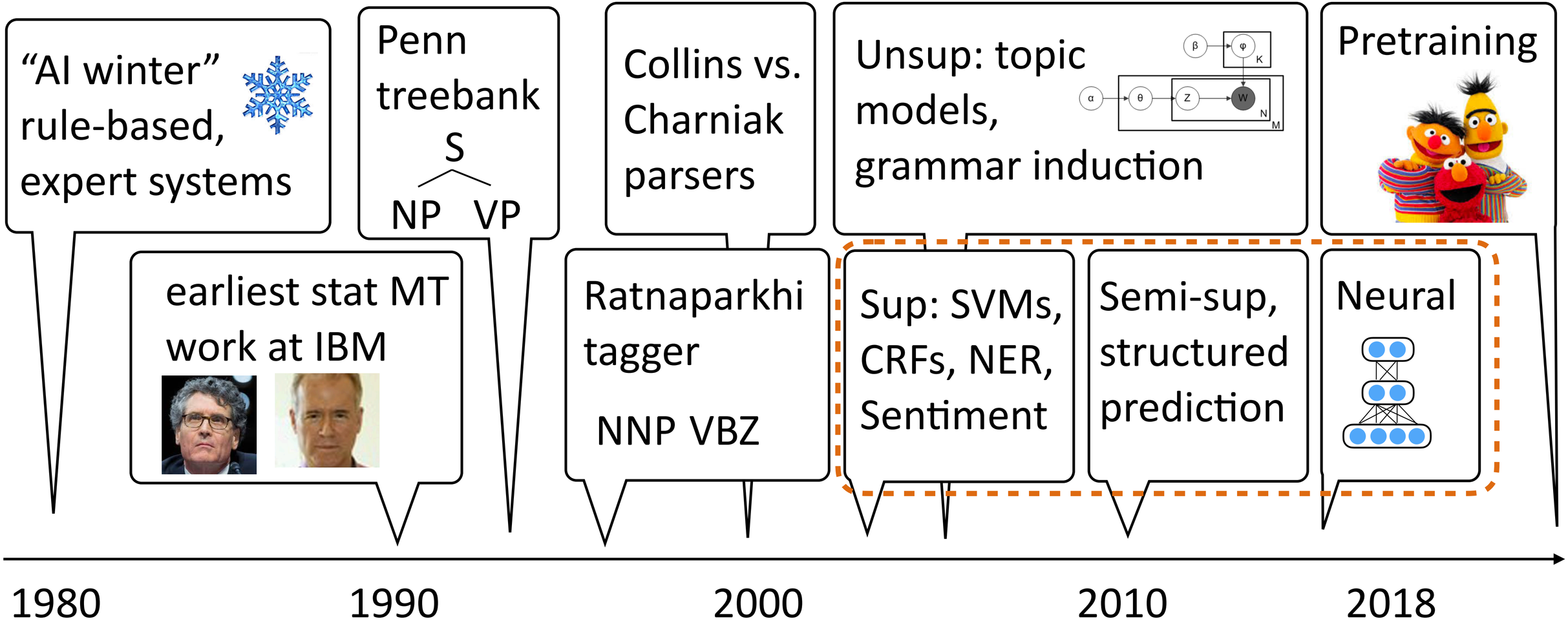
e. He called John at 5 AM on Friday last week. (he = Mike)
C_b = Mike; C_f = {Mike, John} (SHIFT)

Centering Theory
Grosz et al. (1995)

What techniques do we use?

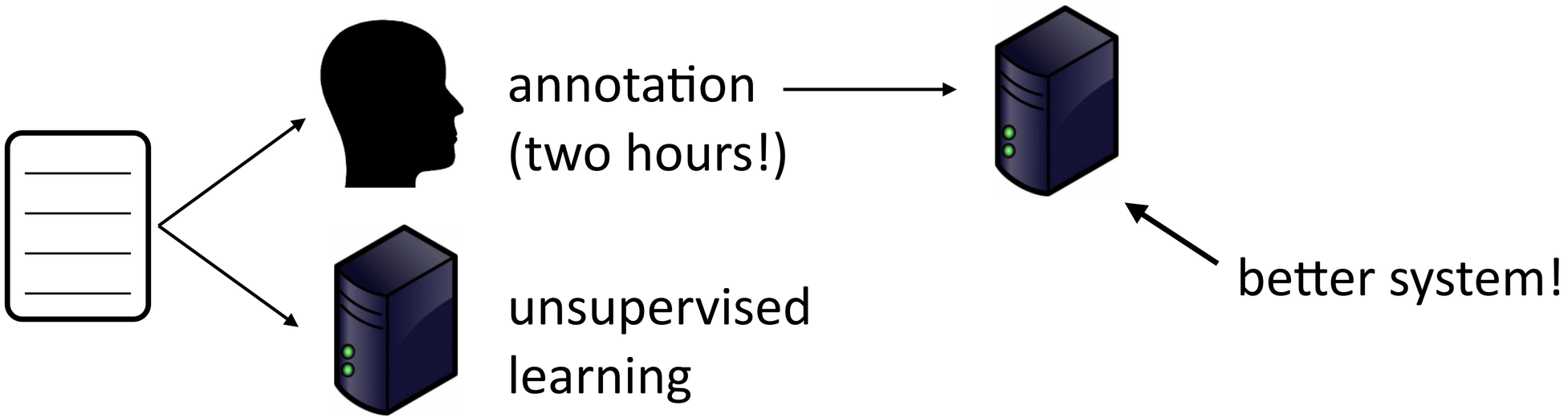
(to combine data, knowledge, linguistics, etc.)

A brief history of (modern) NLP



Structured Prediction

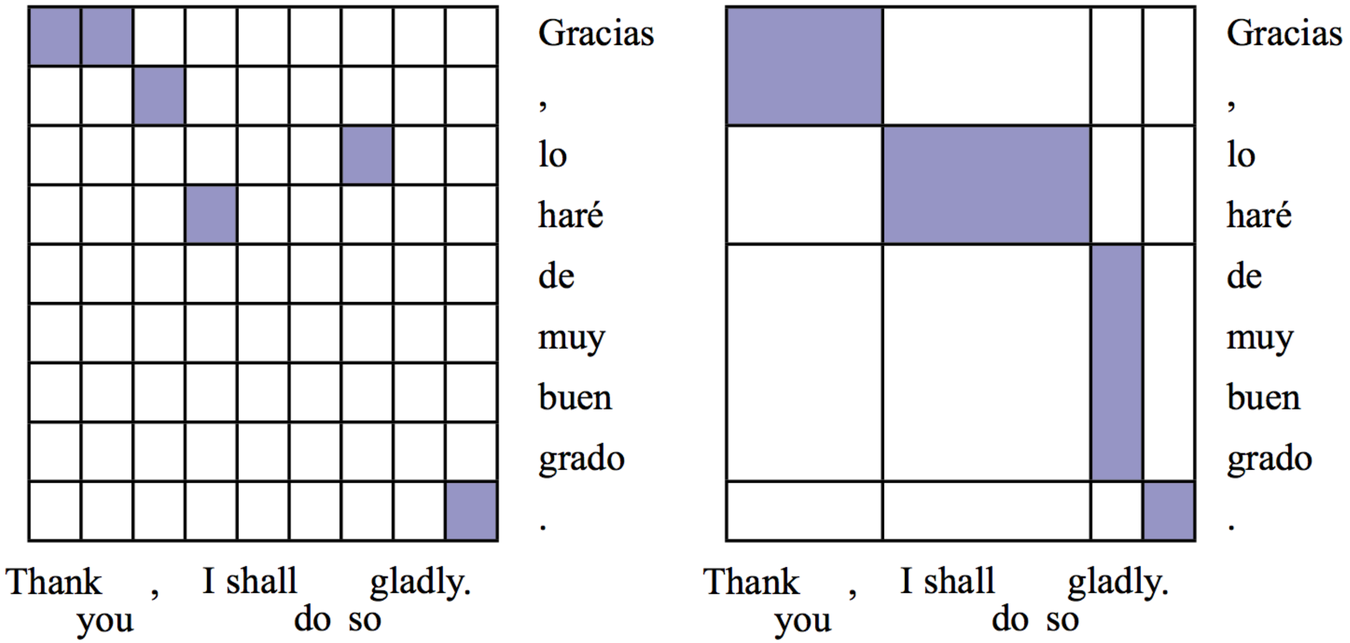
- All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- Supervised techniques work well on very little data



- Even neural nets can do pretty well!

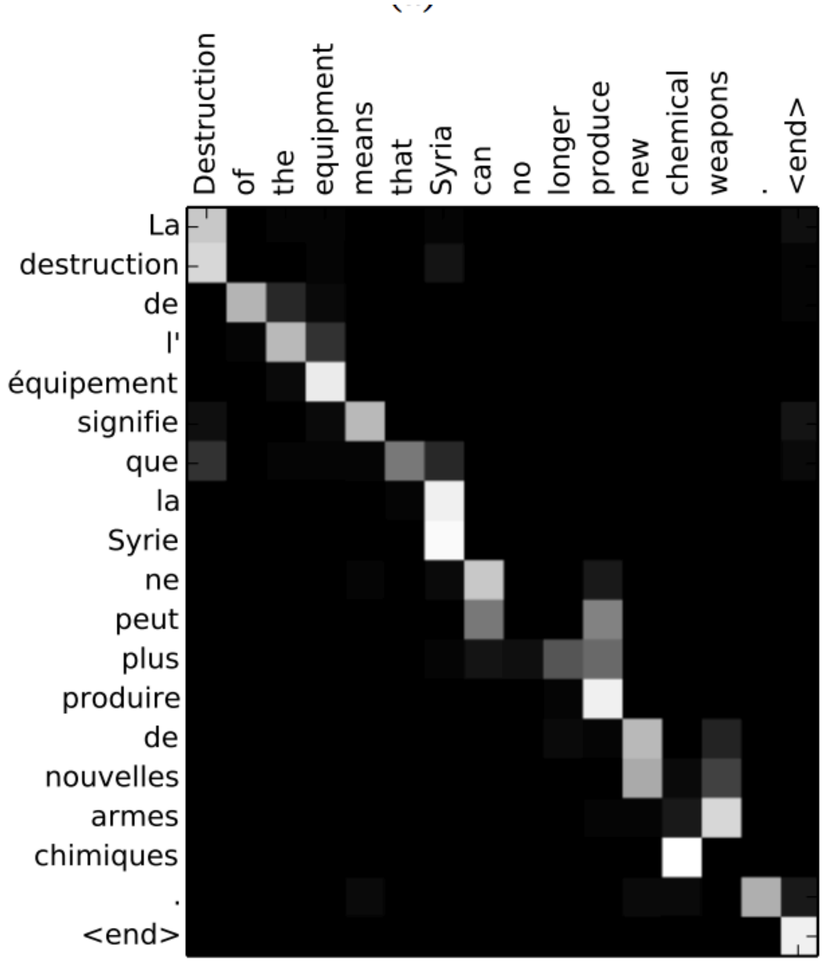
"Learning a Part-of-Speech Tagger from Two Hours of Annotation"
Garrette and Baldridge (2013)

Less Manual Structure?



(a) example word alignment **(b) example phrase alignment**

DeNero et al. (2008)



Bahdanau et al. (2014)

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

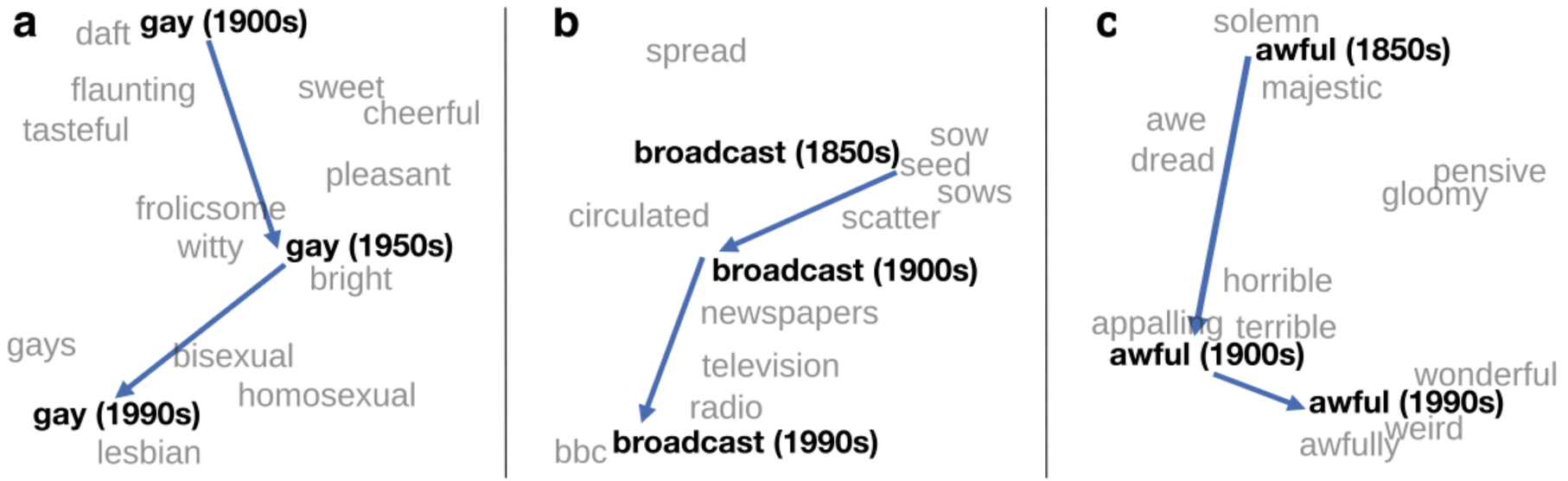
- Maybe manual structure would help…

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

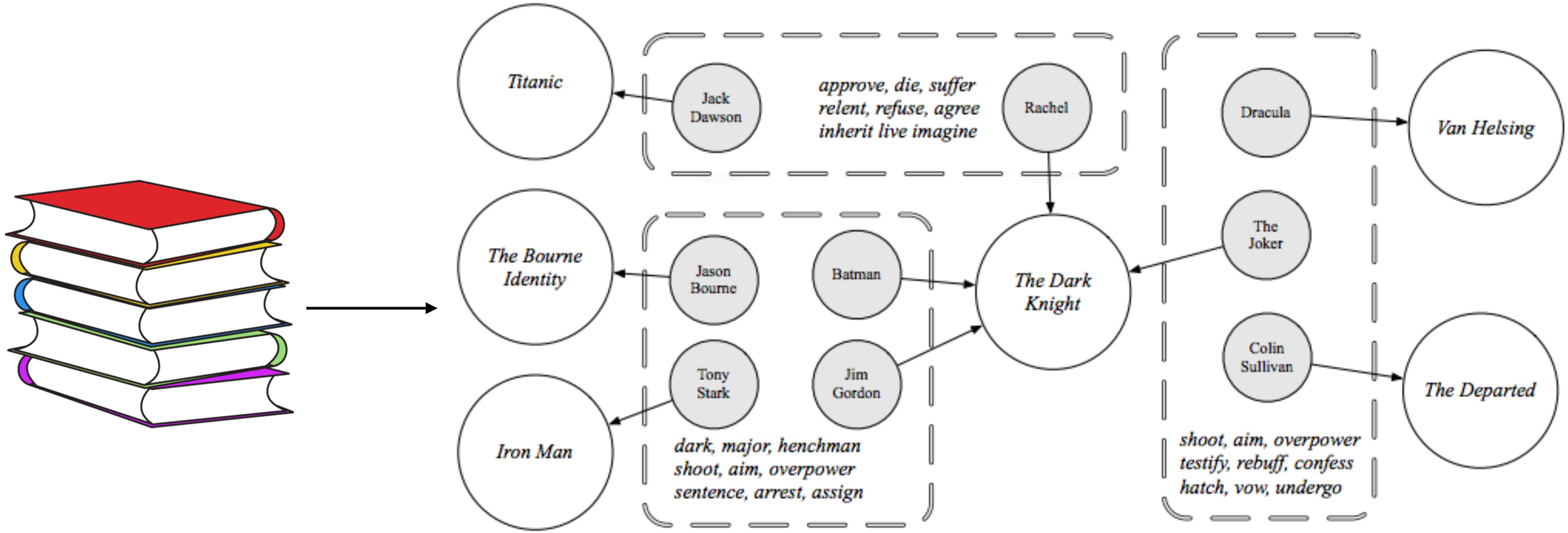
- NLP: build systems that deal with language data
- CL: use computational tools to study language



Hamilton et al. (2016), KulKarni et al. (2015)

NLP vs. Computational Linguistics

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



Bamman, O'Connor, Smith (2013)

Outline of the Course

ML and structured prediction for NLP

Neural Networks semantics

Applications: MT, IE, summarization, dialogue, etc.

Date	Topics (tentative and subject to change)	Readings
1/8/2020	Course Overview	J+M, 3rd Edition Chapter 1
1/10/2020	Binary Classification	Eisenstein 2.0-2.5, 4.1,4.3-4.5, CIML, 4.1-4.4, 4.6-4.7
1/15/2020	Binary Classification	Eisenstein 2.0-2.5, 4.1,4.3-4.5, CIML, 4.1-4.4, 4.6-4.7
1/17/2020	Multiclass Classification	J+M Chapter 5
1/22/2020	Multiclass Classification	J+M Chapter 5
1/24/2020	Neural Networks	Eisenstein 3.1-3.3, J+M 7.1-7.4
1/29/2020	Sequence Tagging	Eisenstein 7.0-7.4, J+M Chapter 8
1/31/2020	Viterbi Algorithm	J+M Chapter 8
2/5/2020	Conditional Random Fields	Eisenstein 7.5, 8.3,
2/7/2020	Conditional Random Fields	Eisenstein 7.5, 8.3,
2/12/2020	Word Emeddings	Eisenstein 3.3.4, 14.5, 14.6, J+M 6
2/14/2020	Semantics	Eisenstein 3.3.4, 14.5, 14.6, J+M 6
2/19/2020	Recurrent Neural Networks	Goldberg 10,11,
2/21/2020	Convolutional Neural Networks, Neural CRFs	Eisenstein 3.4, 7.6
2/26/2020	Machine Translation	Eisenstein 18.1, 18.2
2/28/2020	Encoder-Decoder Networks	Seq2Seq
3/4/2020	**Spring Break**	
3/6/2020	**Spring Break**	
3/11/2020	Information Extraction	Eisenstein 13, 17
3/13/2020	Information Extraction	Eisenstein 13, 17
3/18/2020	**Possible Midterm Date (tentative)**	
3/20/2020	Neural Machine Translation	Eisenstein 18.3
3/25/2020	Reading Comprehension	E2E Memory Networks, CBT, SQuAD, BiDAF
3/27/2020	Summarization	Eisenstein 19, MMR, Gillick,
4/1/2020	Generation	Sentence compression, SummaRuNNER, Pointer
4/3/2020	Dialogue	J+M Chapter 24
4/8/2020	Unsupervised Learning in NLP	Painless unsup, VAE, ELMo, BERT
4/10/2020	Unsupervised Learning in NLP	Painless unsup, VAE, ELMo, BERT
4/15/2020	Eithcs/Wrapup	
4/17/2020	**No Class**	
final exam date	**Final Project Presentation**	

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 2018~2020?
- 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

- Three Homework Assignments
  - Implementation-oriented, with an open-ended component to each
  - Homework 1 (Naive Bayes for sentiment classification) is out NOW
  - ~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!**

Midterm Exam & Final Project

- Midterm — in class, date TBD (20%)

- Final project (20%) — presentation on the final exam day
  - 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.

- Participation (10%)