CS388: Natural Language Processing
Lecture 1: Introduction

Greg Durrett

TEXAS
The University of Texas at Austin

Administrivia

- Lecture: Tuesdays and Thursdays 12:30pm - 1:45pm
- Course website: http://www.cs.utexas.edu/~gdurrett/courses/fa2019/cs388.shtml
- Piazza: link on the course website
- My office hours: Office hours: Wednesday 4pm, Thursday 2pm
- TA: Uday Kusupati. Office hours: Monday 12pm-1pm, Tuesday 11am-12pm, GDC 1.302

Course Requirements

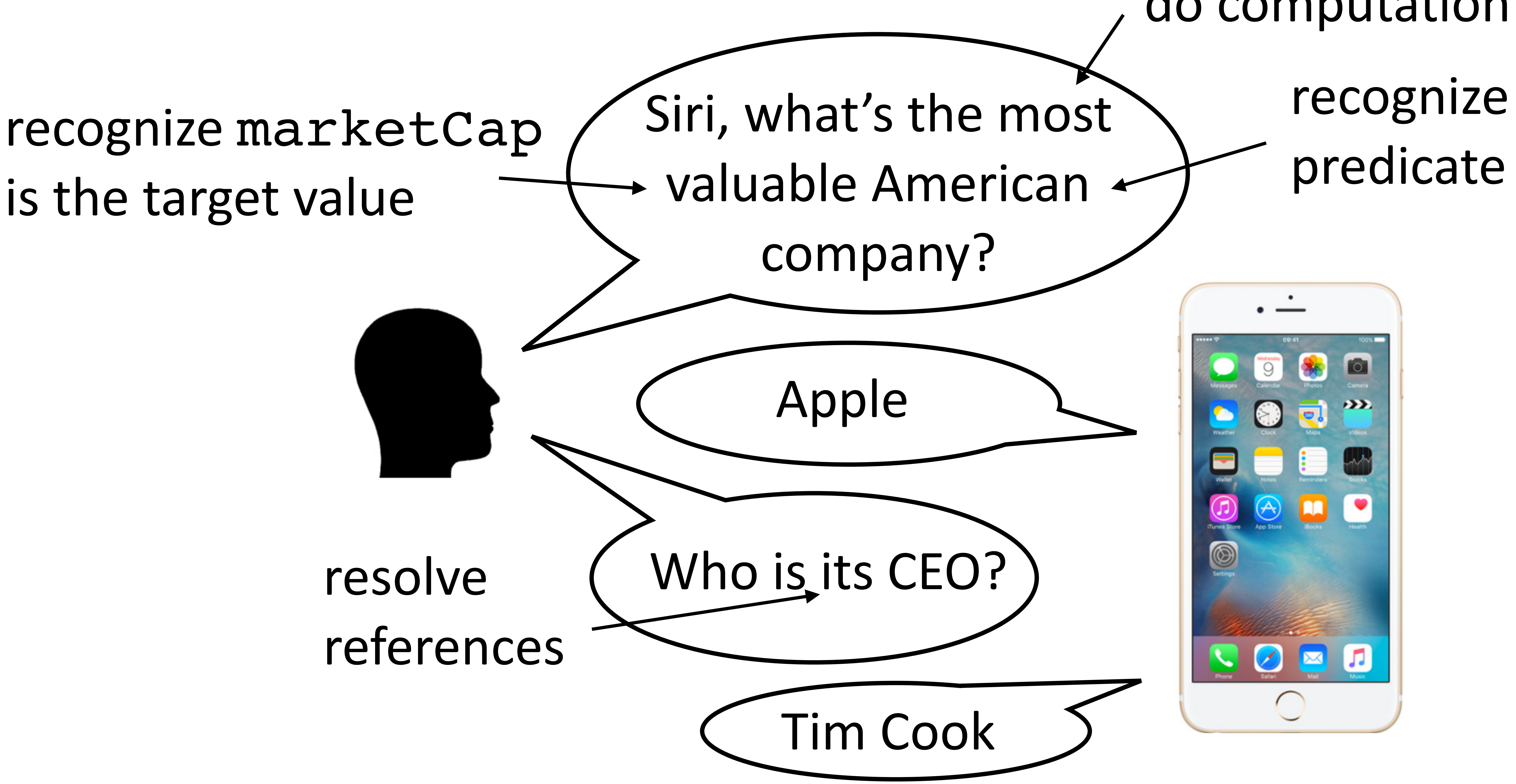
- 391L Machine Learning (or equivalent)
- 311 or 311H Discrete Math for Computer Science (or equivalent)
- Python experience
- Additional prior exposure to probability, linear algebra, optimization, linguistics, and NLP useful but not required

Enrollment

- We'll get as many people in as we can
- Mini1 is out now (due September 10), please look at it soon
  - If this seems like it'll be challenging for you, come and talk to me (this is smaller-scale than the projects, which are smaller-scale than the final project)
- Other NLP offerings:
  - CS378 (ugrad course, taught by me in the spring)
  - LIN 393 (taught by Jessy Li): NLP with minimal supervision

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

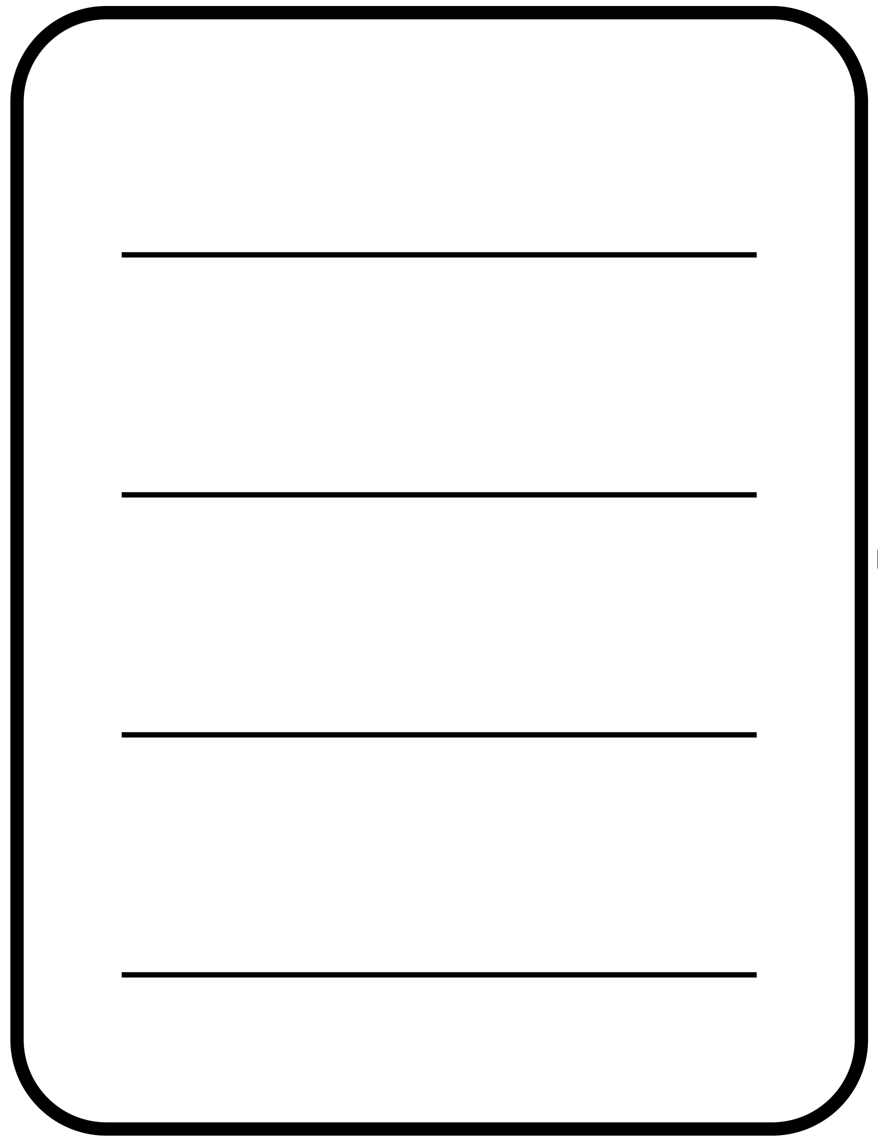
Machine Translation

People’s Daily, August 30, 2017

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

NLP Analysis Pipeline

Text



Text Analysis

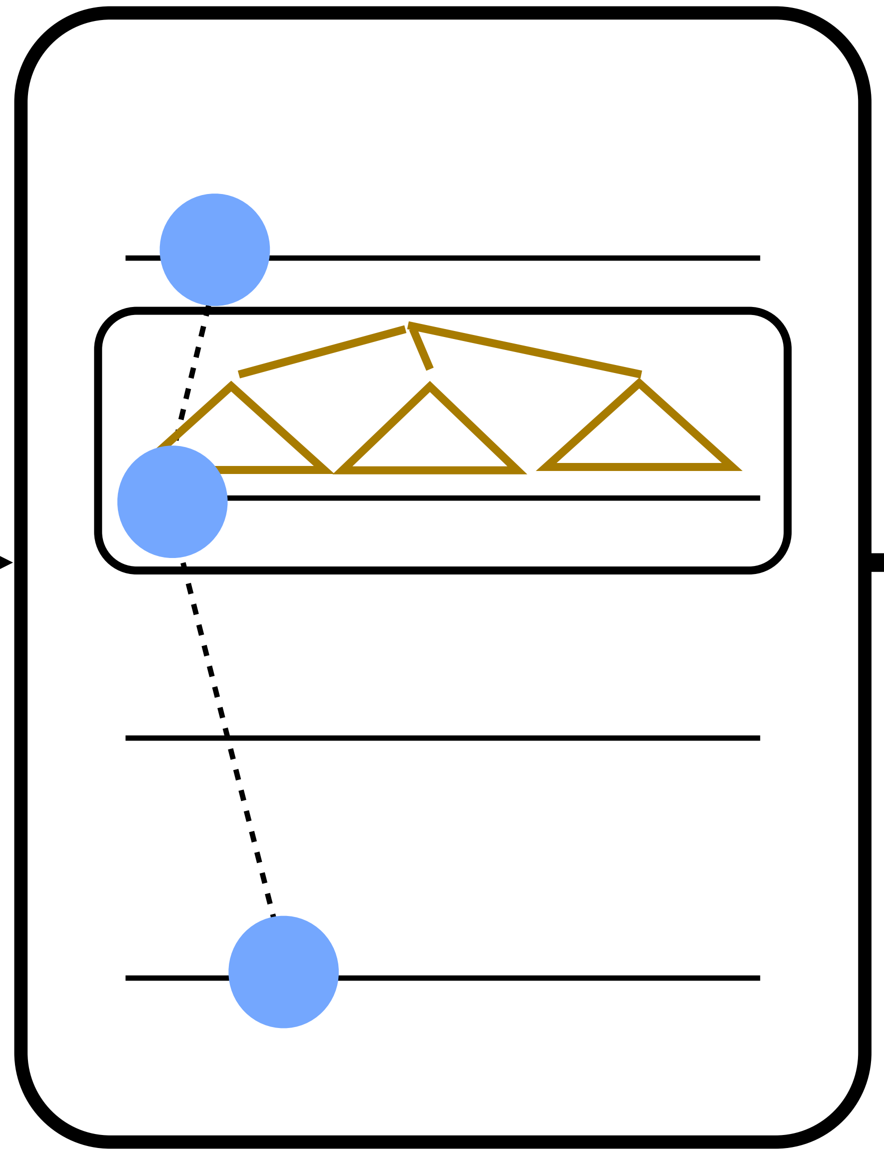
Syntactic parses

Coreference resolution

Entity disambiguation

Discourse analysis

Annotations



Applications

Summarize

Extract information

Answer questions

Identify sentiment

Translate

- 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 +

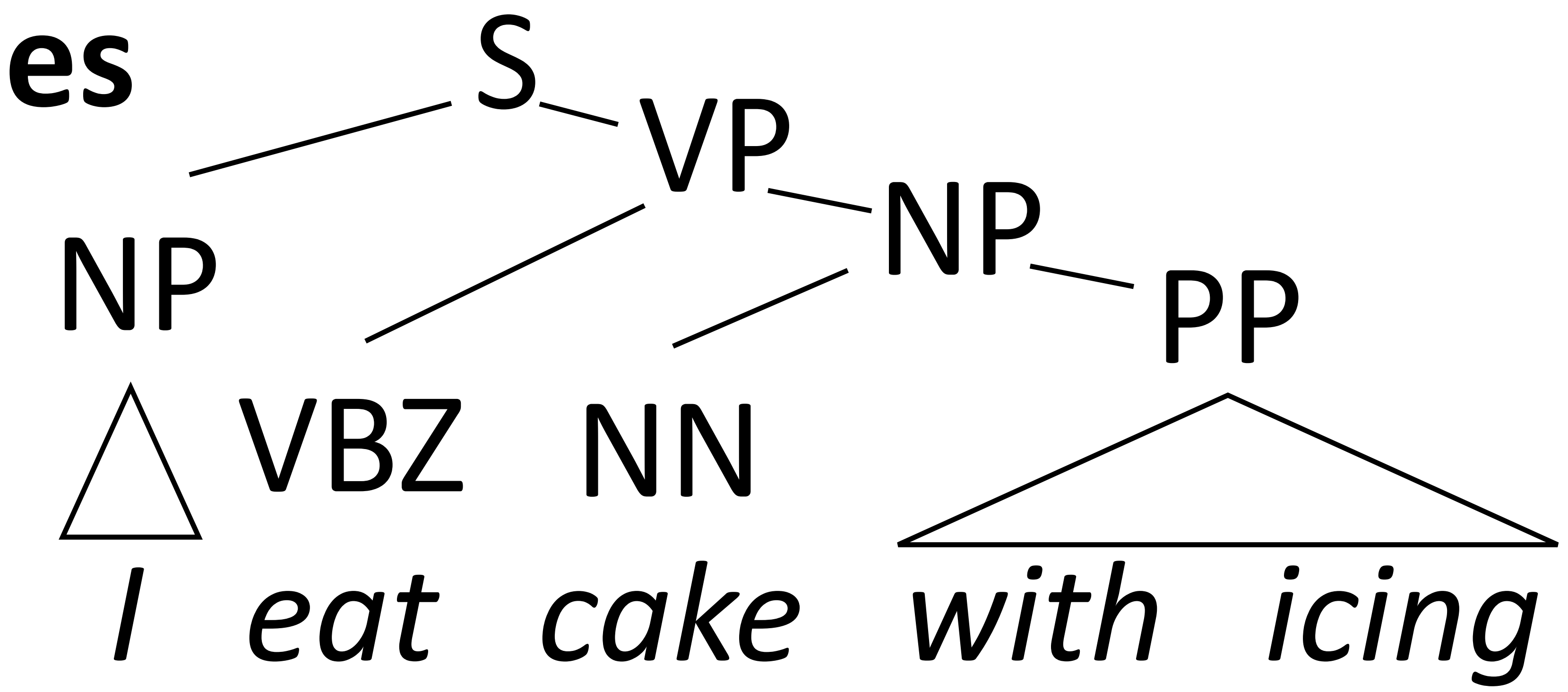
Beyoncé had one of the best videos of all time **subjective**

Sequences/tags

PERSON **WORK_OF_ART**

Tom Cruise stars in the new Mission Impossible film

Trees



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

flights to Miami

How do we use these representations?

Text

Text Analysis

Labels

Sequences

Trees

...

Applications

Extract syntactic features

Tree-structured neural networks

Tree transducers (for machine translation)

...

end-to-end models

- 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 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

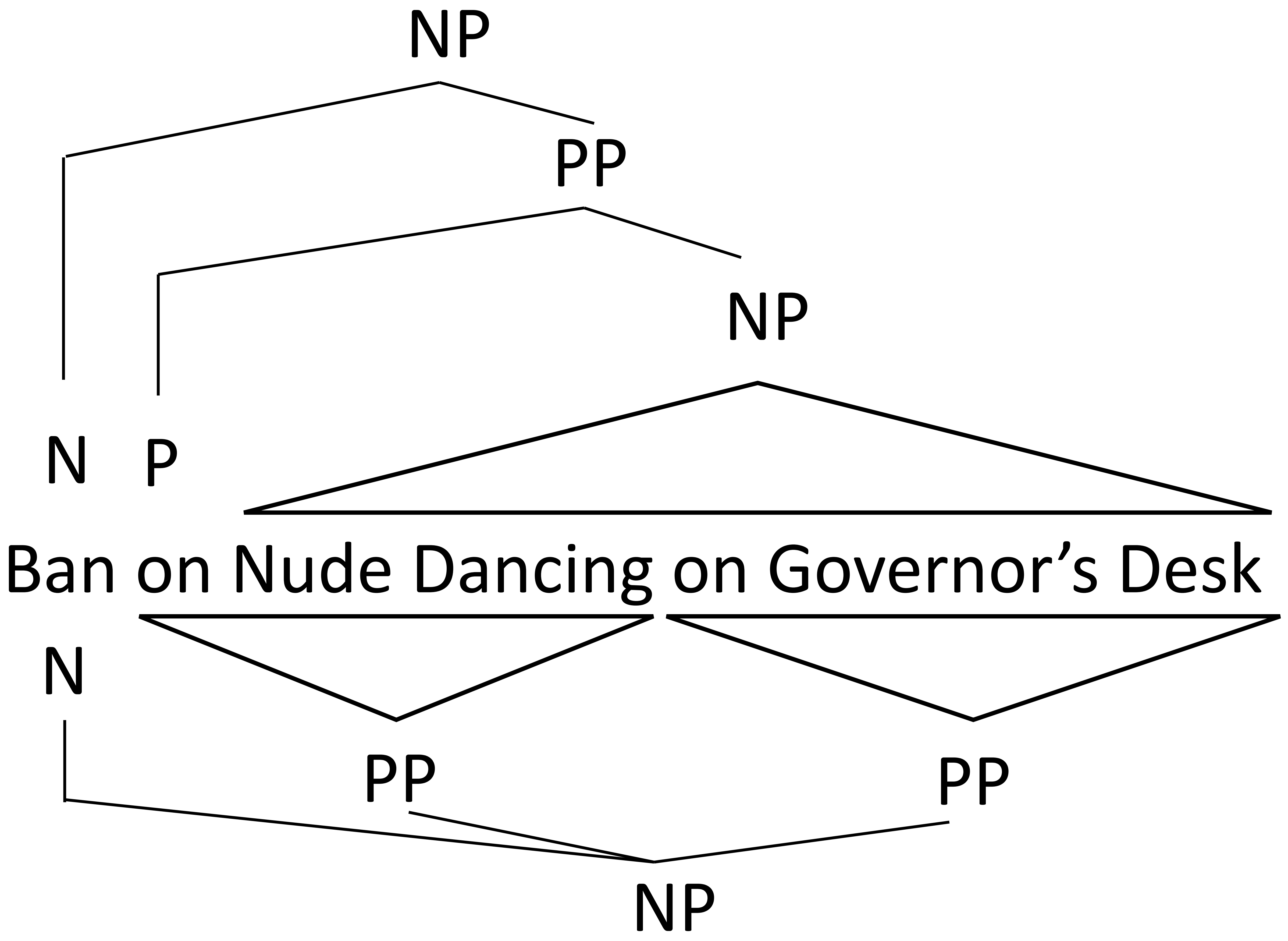
Language is Ambiguous!

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

Teacher Strikes Idle Kids

Iraqi	Head	Seeks	Arms
	body/ position		body/ 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

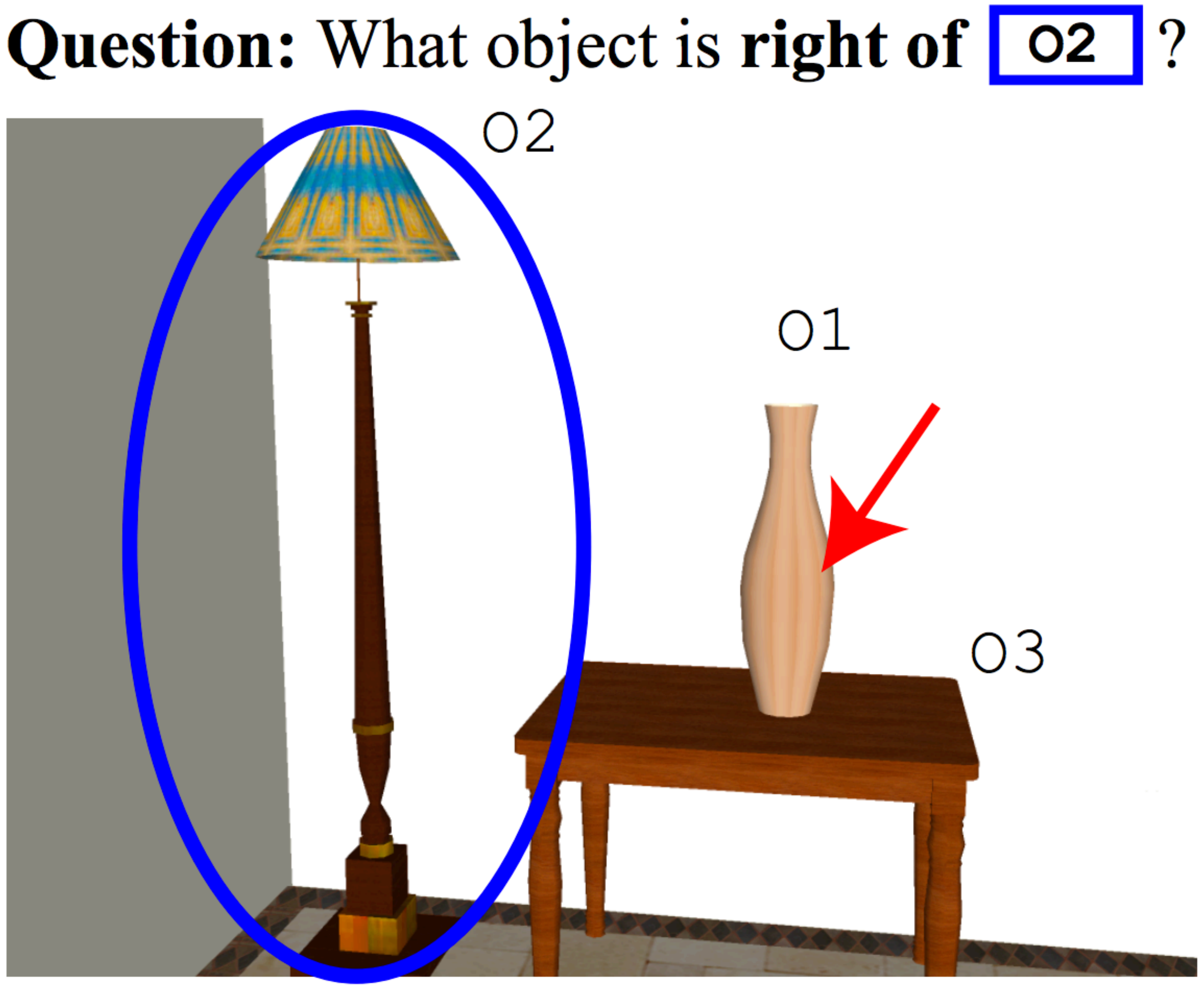
Department of Justice

metaphor;
“approves”

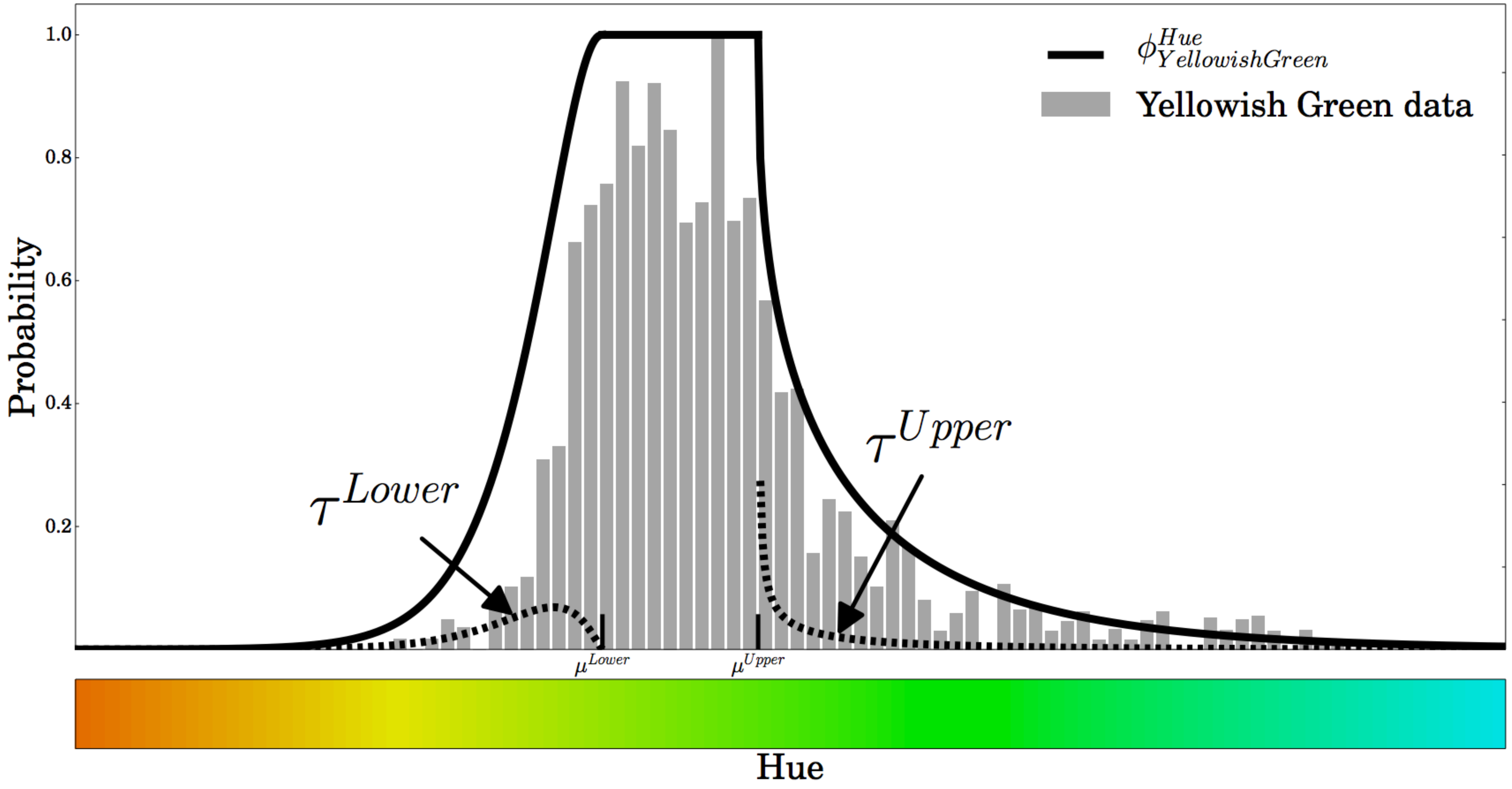
- What is a green light? How do we understand what “green lighting” does?
- Need commonsense knowledge

What do we need to understand language?

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



Golland et al. (2010)



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

Largely rule-based, expert systems

earliest stat MT work at IBM


Penn treebank

S

NP VP

Ratnaparkhi tagger

NNP VBZ

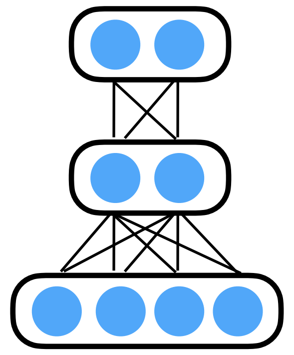
Collins vs. Charniak parsers

Sup: SVMs, CRFs, NER, Sentiment

Unsup: topic models, grammar induction

Semi-sup, structured prediction

Neural

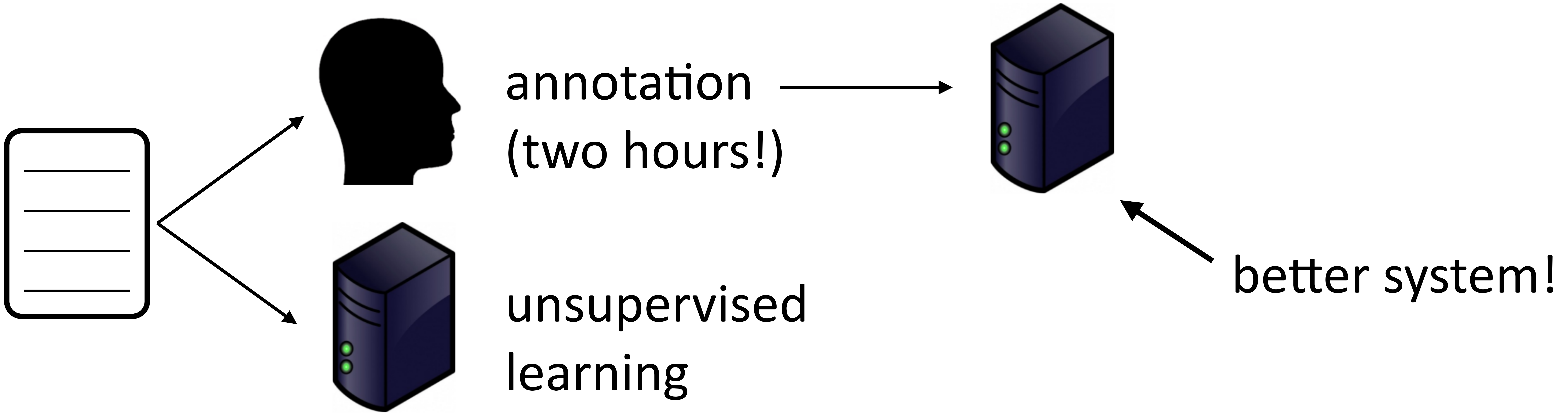


Pretraining

1980 1990 2000 2010 2019

Supervised vs. Unsupervised

- Supervised techniques work well on very little data (even neural networks)

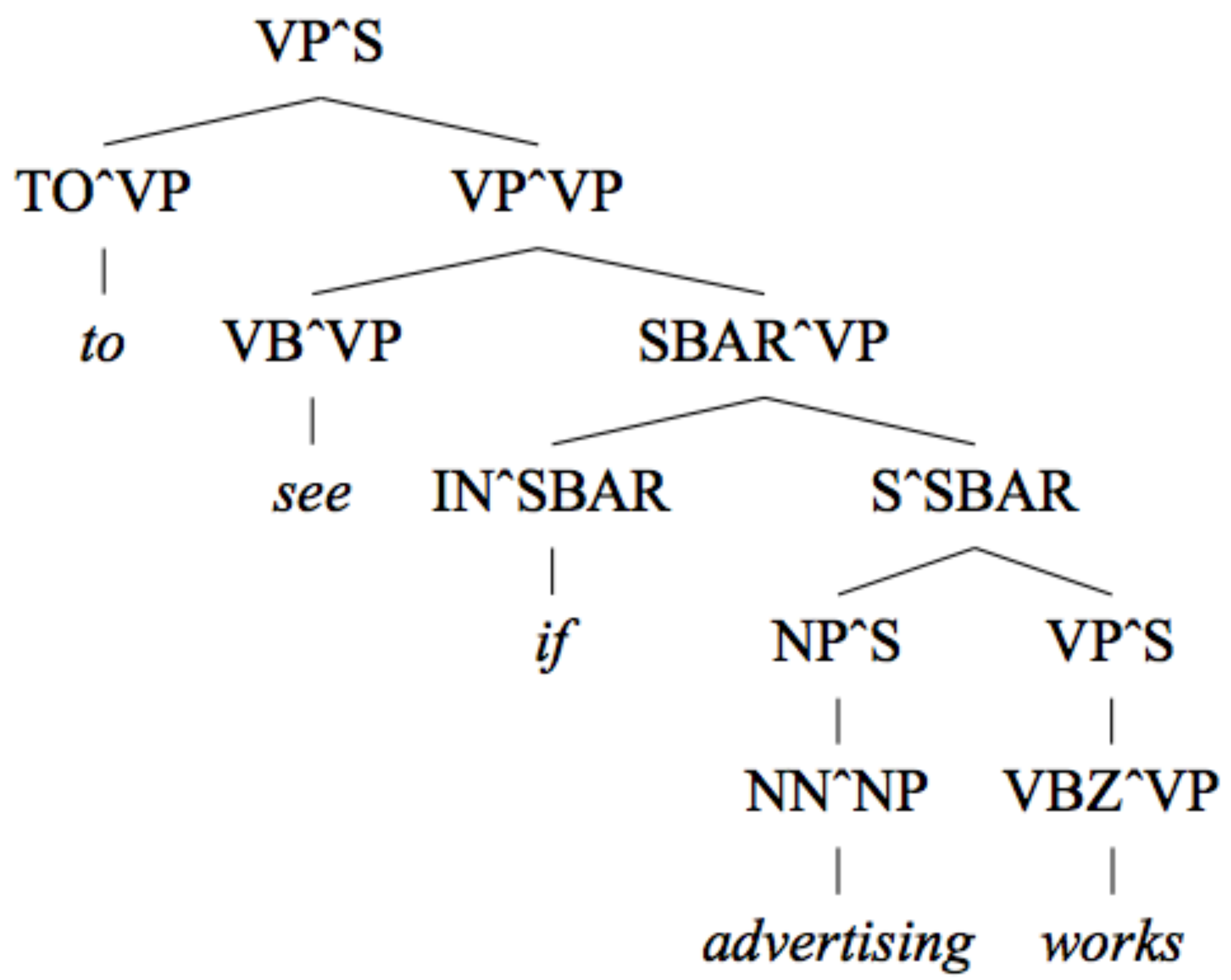


- Fully unsupervised techniques have fallen out of favor

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

Less Manual Structure

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

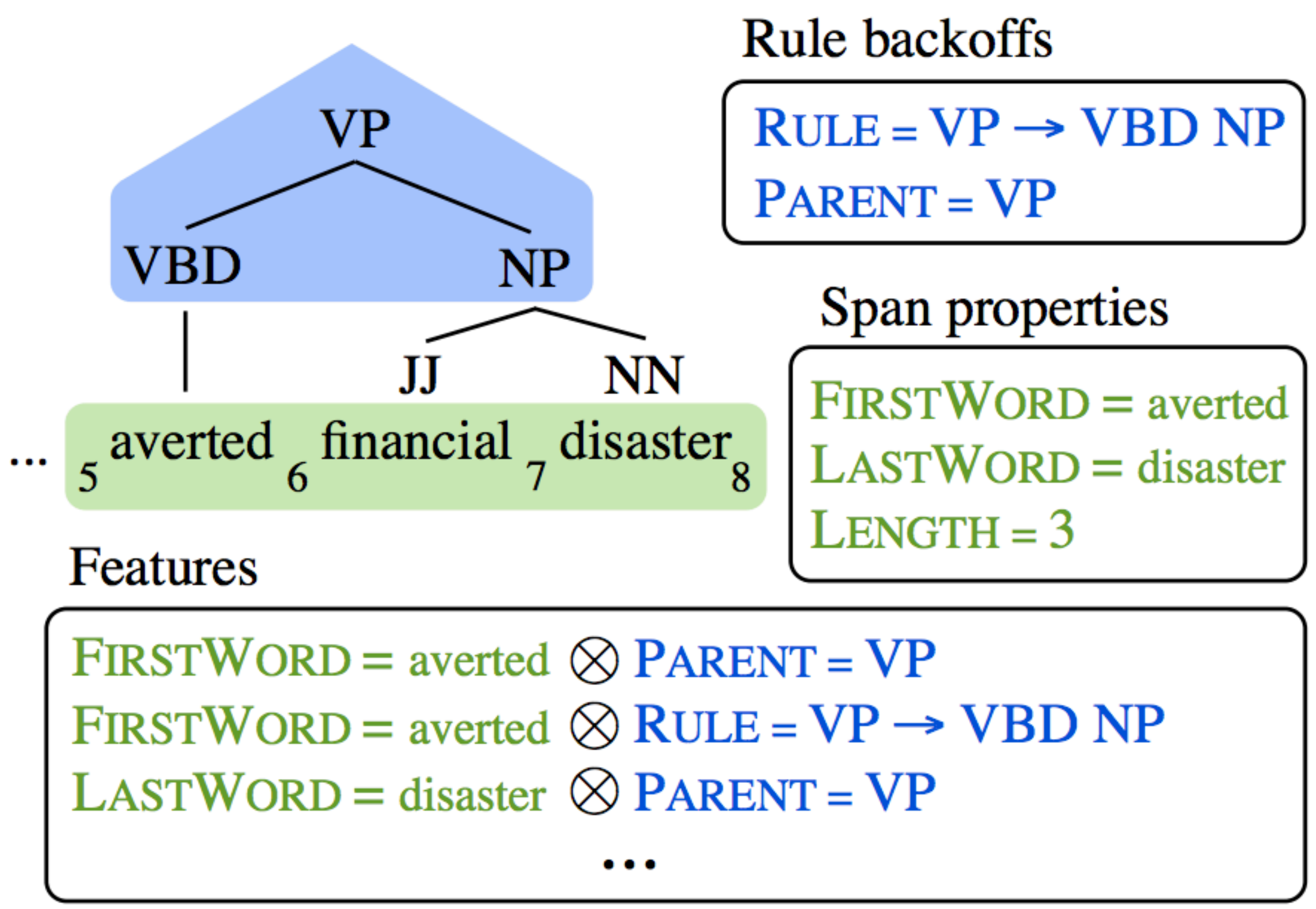


Klein and Manning (2003)
Manually-constructed grammars

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

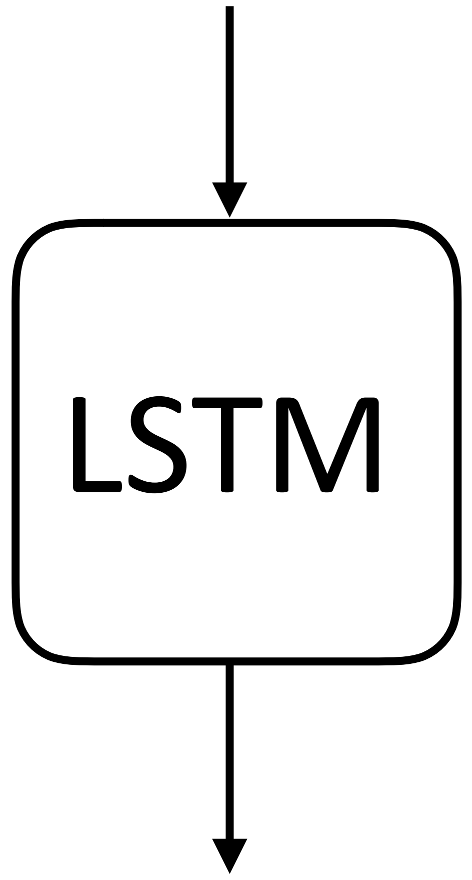
Petrov et al. (2006)
Induced grammars



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

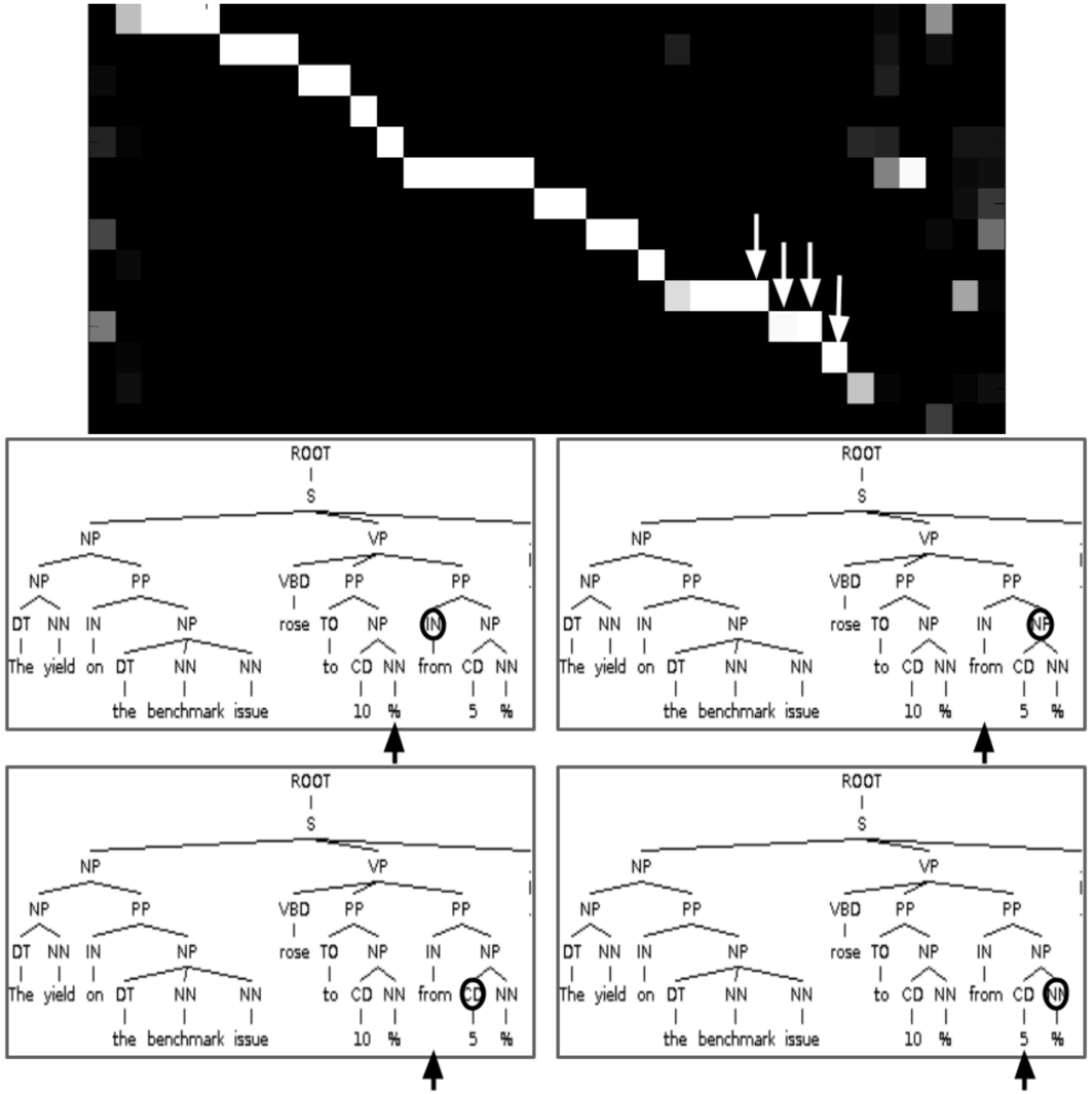
Less Manual Structure

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, ...}

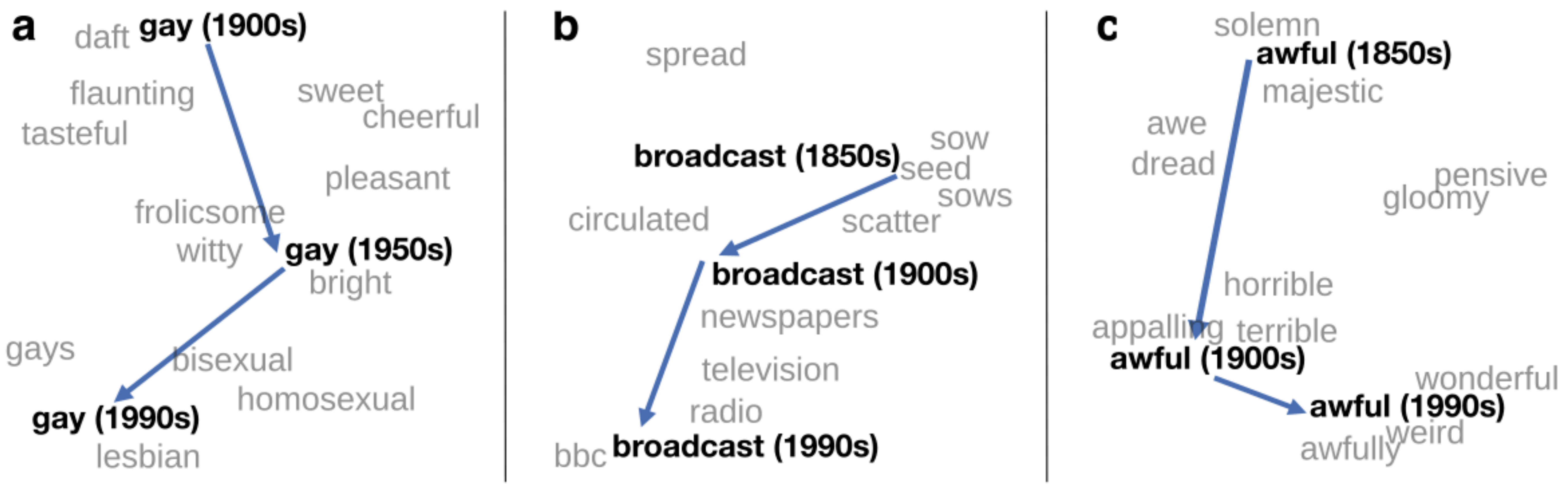
Peters et al. (2018), Devlin et al. (2019)

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

NLP vs. Computational Linguistics

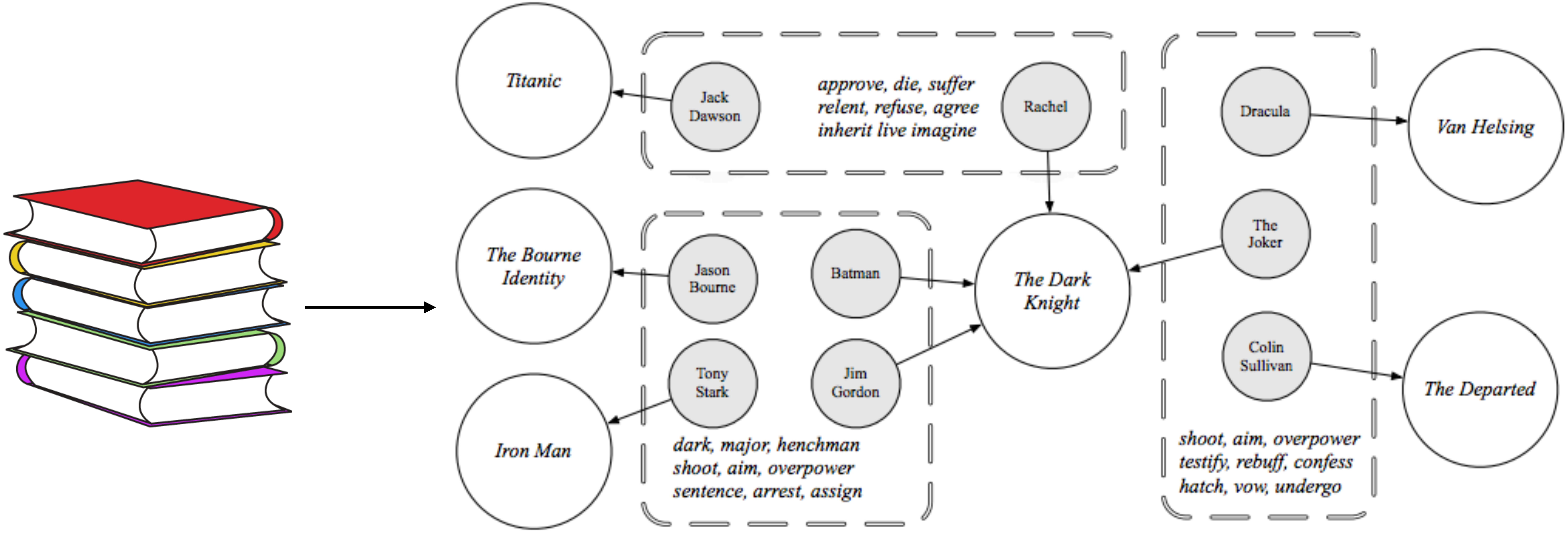
- NLP: build systems that deal with language data
- CL: use computational tools to study language



Hamilton et al. (2016)

NLP vs. Computational Linguistics

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



Bamman, O'Connor, Smith (2013)

Outline

ML and structured prediction for NLP

Neural nets (this part is still in flux)

Aug 29	Introduction [4pp]		Mini1 out
Sept 3	Binary classification	Eisenstein 2.0-2.5, 4.2-4.4.1, JM 4, JM 5.0-5.5	
Sept 5	Multiclass classification	Eisenstein 4.2, JM 5.6, Structured SVM secs 1-2	
Sept 10	Sequence Models 1: HMMs (Guest Lecture: Ray Mooney)	Eisenstein 7.0-7.4, 8.1, JM 8, Manning POS, Viterbi algorithm lecture note	Mini1 due / Proj1 out
Sept 12	Sequence Models 2: CRFs	Eisenstein 7.5, 8.3, Sutton CRFs 2.3, 2.6.1, Wallach CRFs tutorial, Illinois NER	
Sept 17	NN1: Feedforward	Eisenstein 3.0-3.3, Goldberg 1-4, 6, NLP with FFNNs, DANs	
Sept 19	NN2: Word embeddings	Eisenstein 3.3.4, 14.5-14.6, JM 6, Goldberg 5, word2vec, Levy, GloVe, Dropout	
Sept 24	NN3: RNNs	JM 9.1-9.4, Goldberg 10-11, Karpathy	Proj1 due
Sept 26	NN4: Language Modeling and Pretraining	Eisenstein 6, JM 9.2.1, ELMo	Mini2 out
Oct 1	NN5: Interpretability/CNNs/Neural CRFs/etc.		

Outline: Syntax + Semantics

Oct 3	Trees 1: Constituency, PCFGs	Eisenstein 10.0-10.5, JM 12.1-12.6, 12.8, Structural, Lexicalized, State-split	
Oct 8	Trees 2: Constituency Parsers + Dependency	Eisenstein 11.1-11.2, JM 13.1-13.3, 13.5, Dozat	Mini2 due / FP out
Oct 10	Trees 3: Dependency Parsers	Eisenstein 11.3, JM 13.4, Parsey, Huang 2	
Oct 15	Semantics 1	Eisenstein 12, Zettlemoyer, Berant	FP proposal due
Oct 17	Semantics 2 / Seq2seq 1	Seq2seq, Jia	Proj2 out
Oct 22	Seq2seq 2: Attention and Pointers	Attention, Luong Attention, Transformer	

Outline: Applications

Oct 24	Machine Translation 1		
Oct 29	Machine Translation 2 / Transformers		
Oct 31	Pretrained Transformers / BERT	BERT, RoBERTa	
Nov 5	Information Extraction / SRL		Proj2 due
Nov 7	Question Answering 1		
Nov 12	Question Answering 2		
Nov 14	Dialogue	RNN chatbots, Diversity, Goal-oriented, Latent Intention, QA-as-dialogue	
Nov 19	Summarization	Eisenstein 19, MMR, Gillick, Sentence compression, SummaRuNNER, Pointer	
Nov 21	Multilinguality and morphology	Xlingual POS, Xlingual parsing, Xlingual embeddings	
Nov 26	Wrapup + Ethics		

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 2019?
- Make you a "producer" rather than a "consumer" of NLP tools
  - The four assignments should teach you what you need to know to understand nearly any system in the literature (e.g.: state-of-the-art NER system = project 1 + mini 2, basic MT system = project 2)

Assignments

- Two minis (10% each), two projects (20% each)
  - Implementation-oriented, with an open-ended component to each
  - Mini 1 (classification) is out NOW
  - 1 week for minis, ~2 weeks per project, 5 "slip days" for automatic extensions
- Grading:
  - Minis: largely graded based on code performance
  - Projects: graded on a mix of code performance, writeup, extension

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 (40%)
  - Groups of 2 preferred, 1 is possible
  - (Brief!) proposal to be approved by me by the midpoint of the semester (October 15)
  - Written in the style and tone of an ACL paper

Conduct

A climate conducive to learning and creating knowledge is the right of every person in our community. Bias, harassment and discrimination of any sort have no place here. If you notice an incident that causes concern, please contact the Campus Climate Response Team: **diversity.utexas.edu/ccrt**

The University of Texas at Austin
College of Natural Sciences

The College of Natural Sciences is steadfastly committed to enriching and transformative educational and research experiences for every member of our community. Find more resources to support a diverse, equitable and welcoming community within Texas Science and share your experiences at cns.utexas.edu/diversity

Survey (Optional)

1. Name
2. Fill in: I am a [CS / ____] [PhD / masters / undergrad] in year [1 2 3 4 5+]
3. Write one reason you want to take this class or one thing you want to get out of it
4. One interesting fact about yourself, or what you like to do in your spare time