# CS388: Natural Language Processing
# Lecture 1: Introduction


Greg Durrett

TEXAS
The University of Texas at Austin


# Administrivia

- Lecture: Tuesdays and Thursdays 9:30am - 10:45am
- Course website: http://www.cs.utexas.edu/~gdurrett/courses/sp2021/cs388.shtml
- Gradescope: you should've gotten an email
- Piazza: link on the course website
- My office hours: Office hours: Tuesday 1pm-2pm, Wednesday 3:30pm-4:30pm
  - **Note: my OHs today are 12:30pm-1:30pm**
- TA: Xi Ye. See course website for OHs

# 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
- Mini1 is out now (due January 28), 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)

# What’s the goal of NLP?

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

# Question Answering

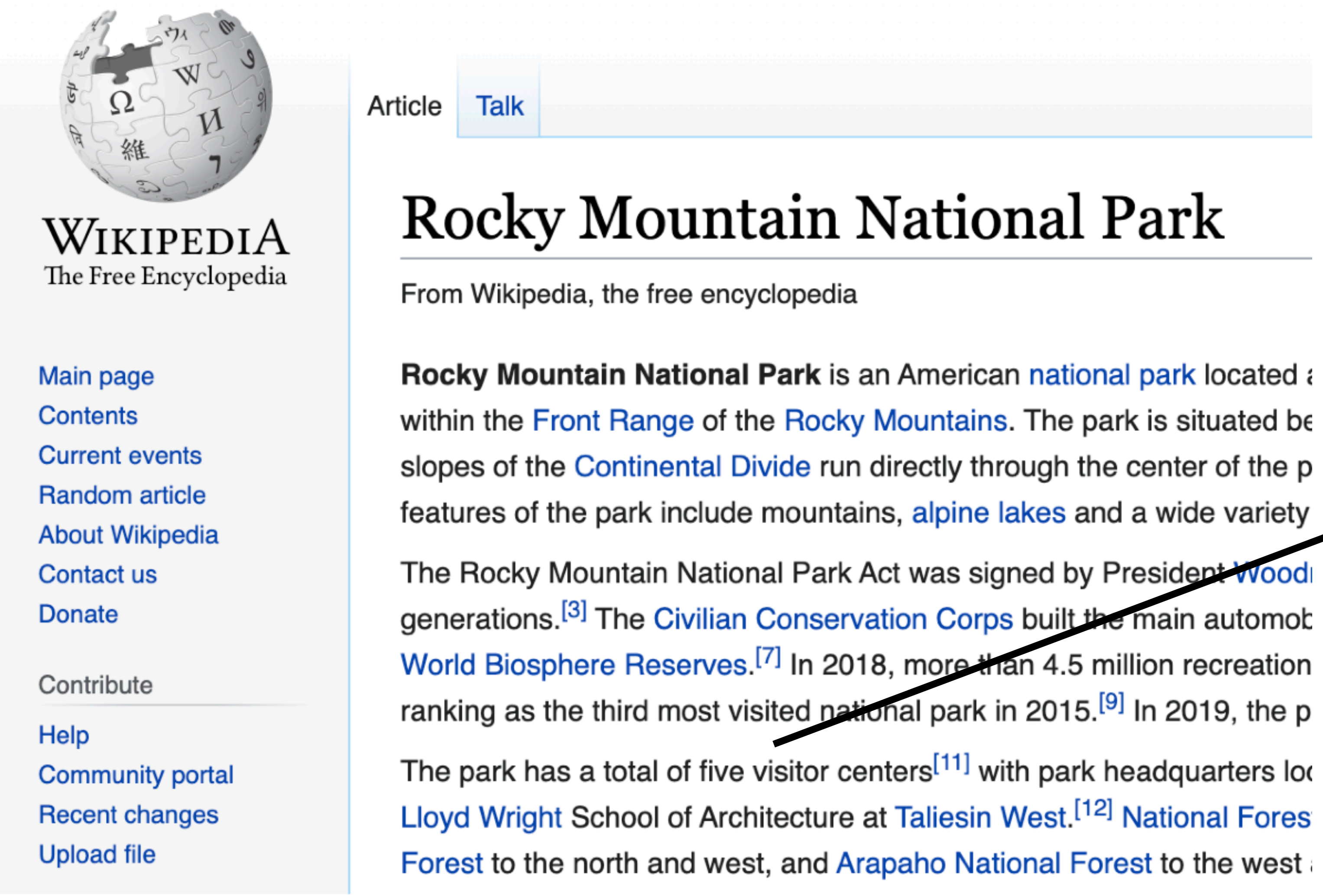
When was Abraham Lincoln born?

map to Birthday field

| Name | Birthday |
|---|---|
| Lincoln, Abraham | 2/12/1809 |
| Washington, George | 2/22/1732 |
| Adams, John | 10/30/1735 |

**February 12, 1809**

How many visitors centers are there in Rocky Mountain National Park?


WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute

Help
Community portal
Recent changes
Upload file

Article Talk

Rocky Mountain National Park

From Wikipedia, the free encyclopedia

**Rocky Mountain National Park** is an American national park located a
within the Front Range of the Rocky Mountains. The park is situated be
slopes of the Continental Divide run directly through the center of the p
features of the park include mountains, alpine lakes and a wide variety

The Rocky Mountain National Park Act was signed by President Wood
generations.[3] The Civilian Conservation Corps built the main automob
World Biosphere Reserves.[7] In 2018, more than 4.5 million recreation
ranking as the third most visited national park in 2015.[9] In 2019, the p

The park has a total of five visitor centers[11] with park headquarters loc
Lloyd Wright School of Architecture at Taliesin West.[12] National Fores
Forest to the north and west, and Arapaho National Forest to the west a

The park has a total of five visitor centers

five

# Machine Translation

The Political Bureau of the CPC Central Committee | July 30 | hold a meeting

中共中央政治局7月30日召开会议，会议分析研究当前经济形势，部署下半年经济工作。

People's Daily, August 10, 2020

Translate

The Political Bureau of the CPC Central Committee held a meeting on July 30 to analyze and study the current economic situation and plan economic work in the second half of the year.

# 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

# NLP Analysis Pipeline

**Text**

**Text Analysis**

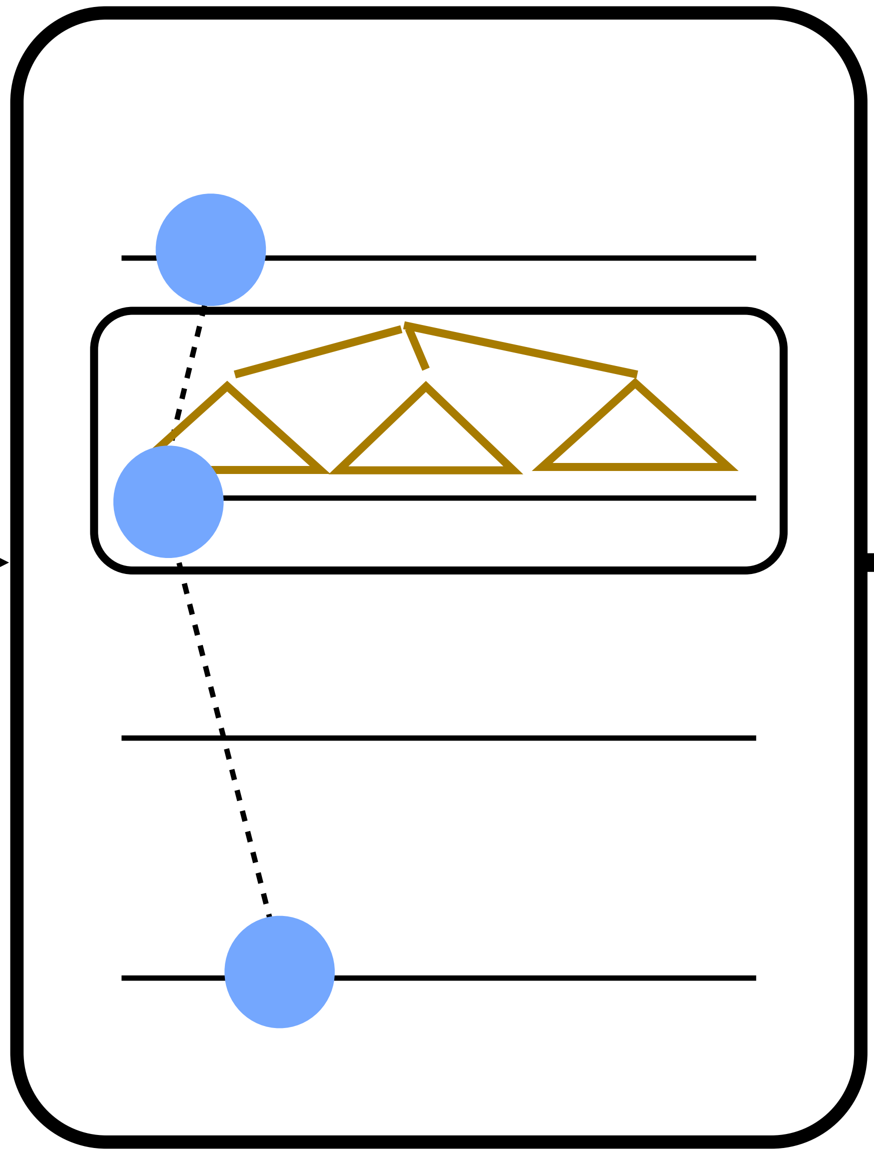
Syntactic parses

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**

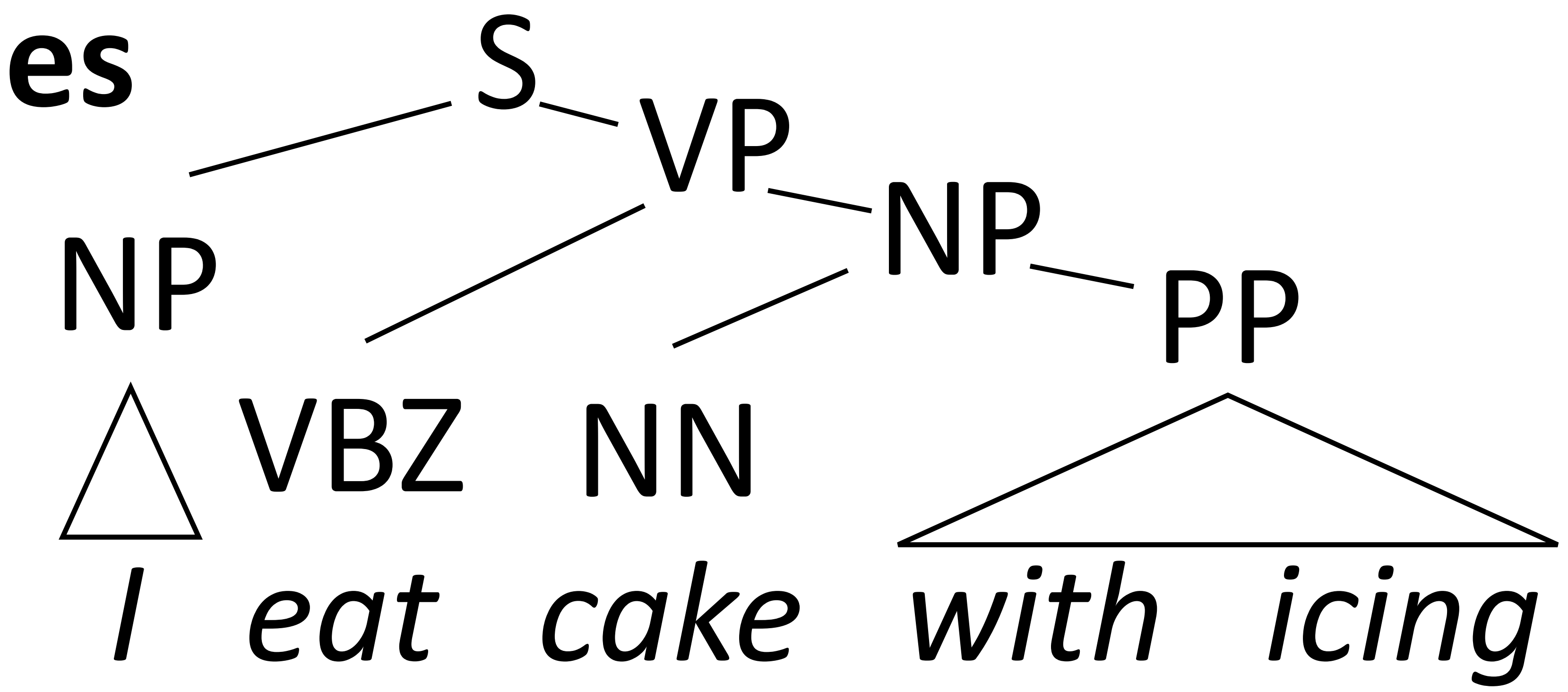
*the movie was good* +

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

**Sequences/tags**

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

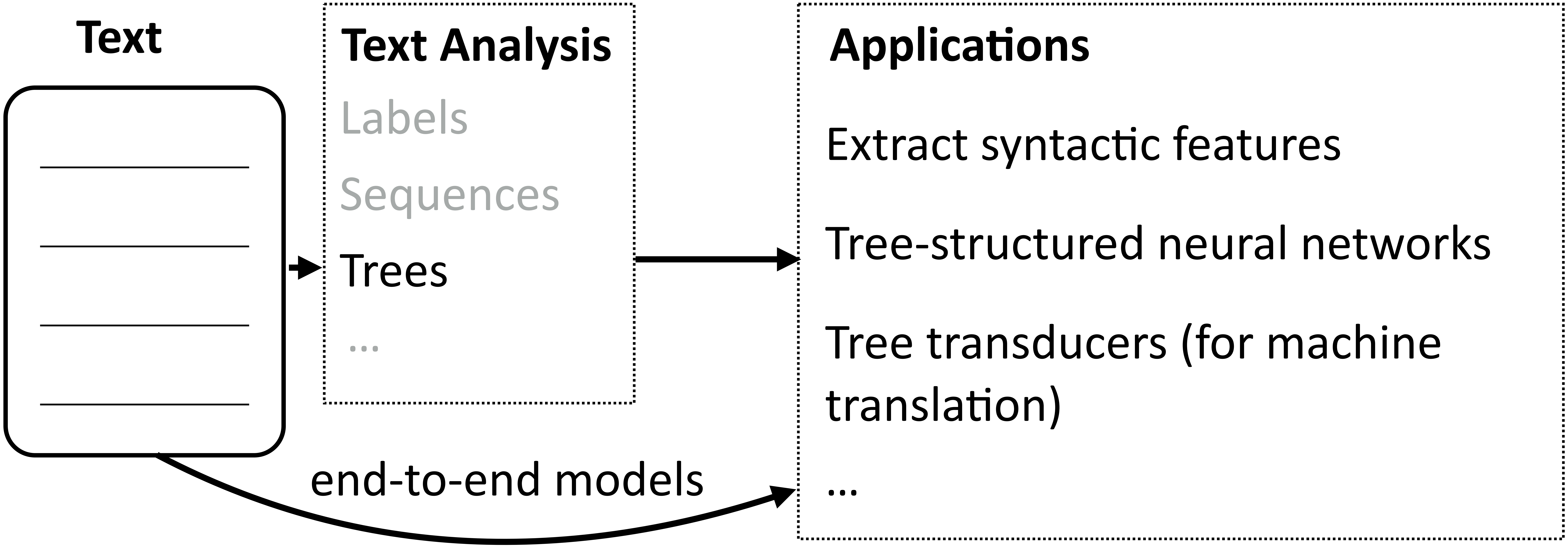
**Trees**




*λx. flight(x)* ∧ *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 advocated violence

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

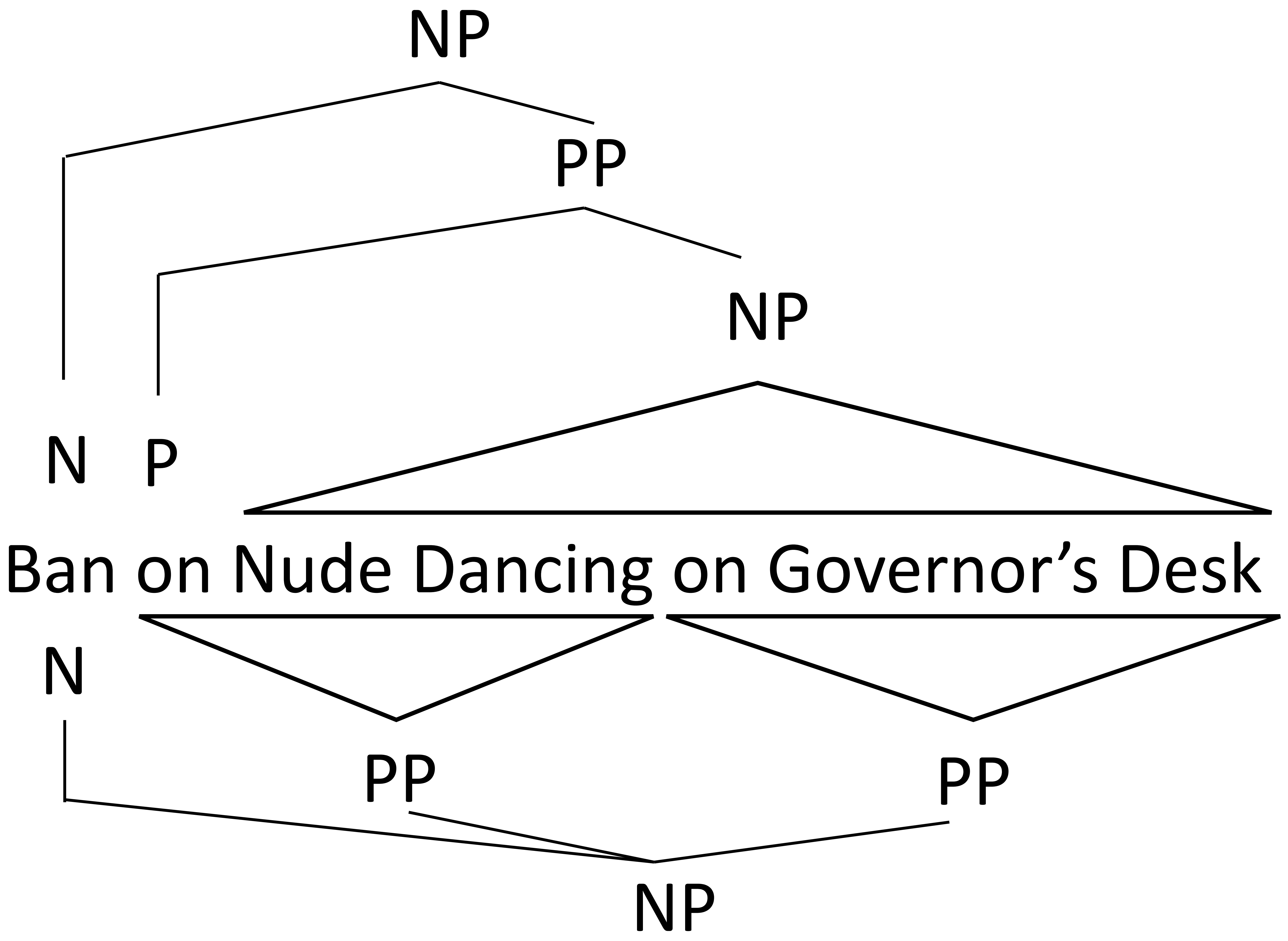
The city council refused the demonstrators a permit because they _______ violence

- >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/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*

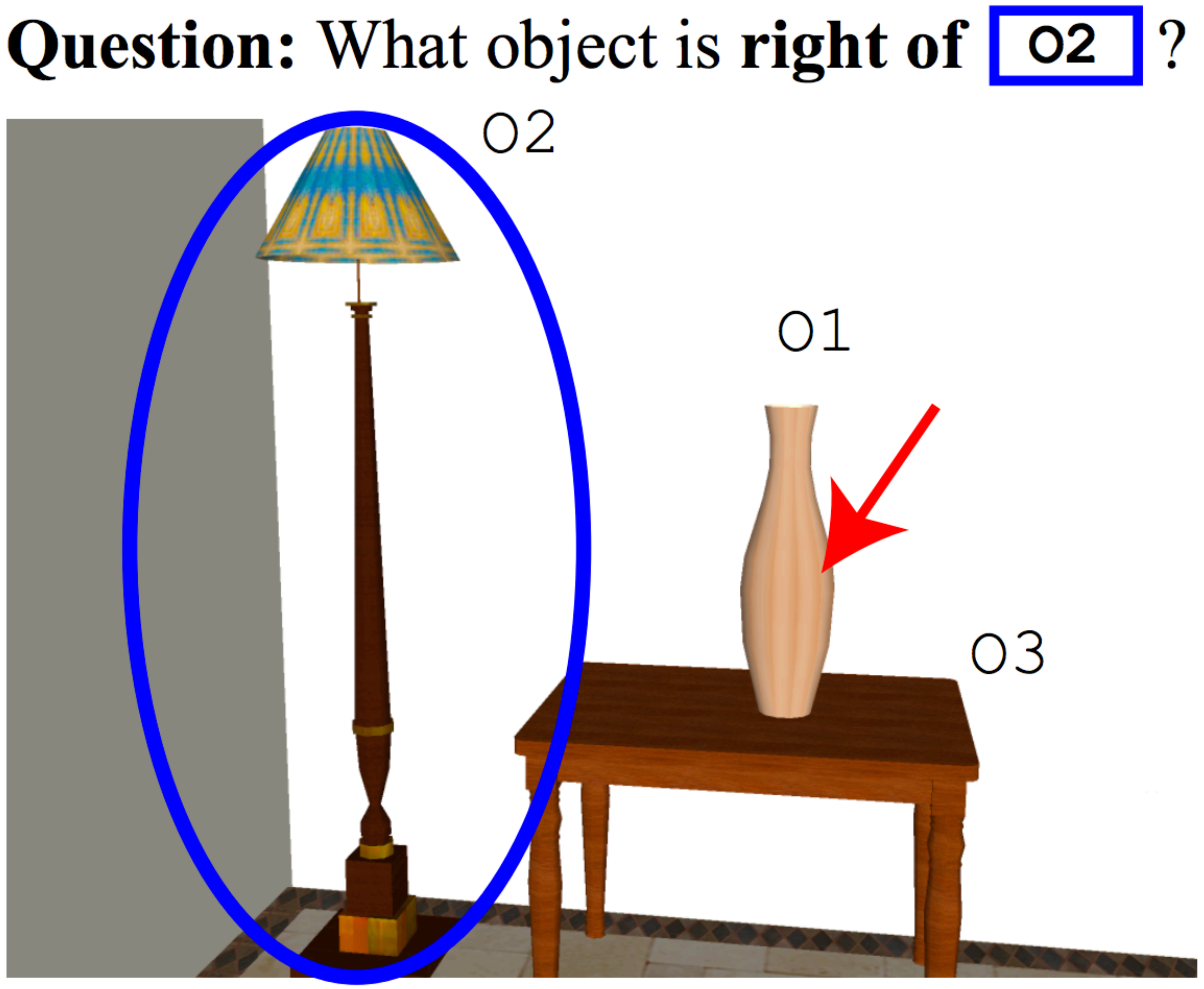
metaphor;
"approves"

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

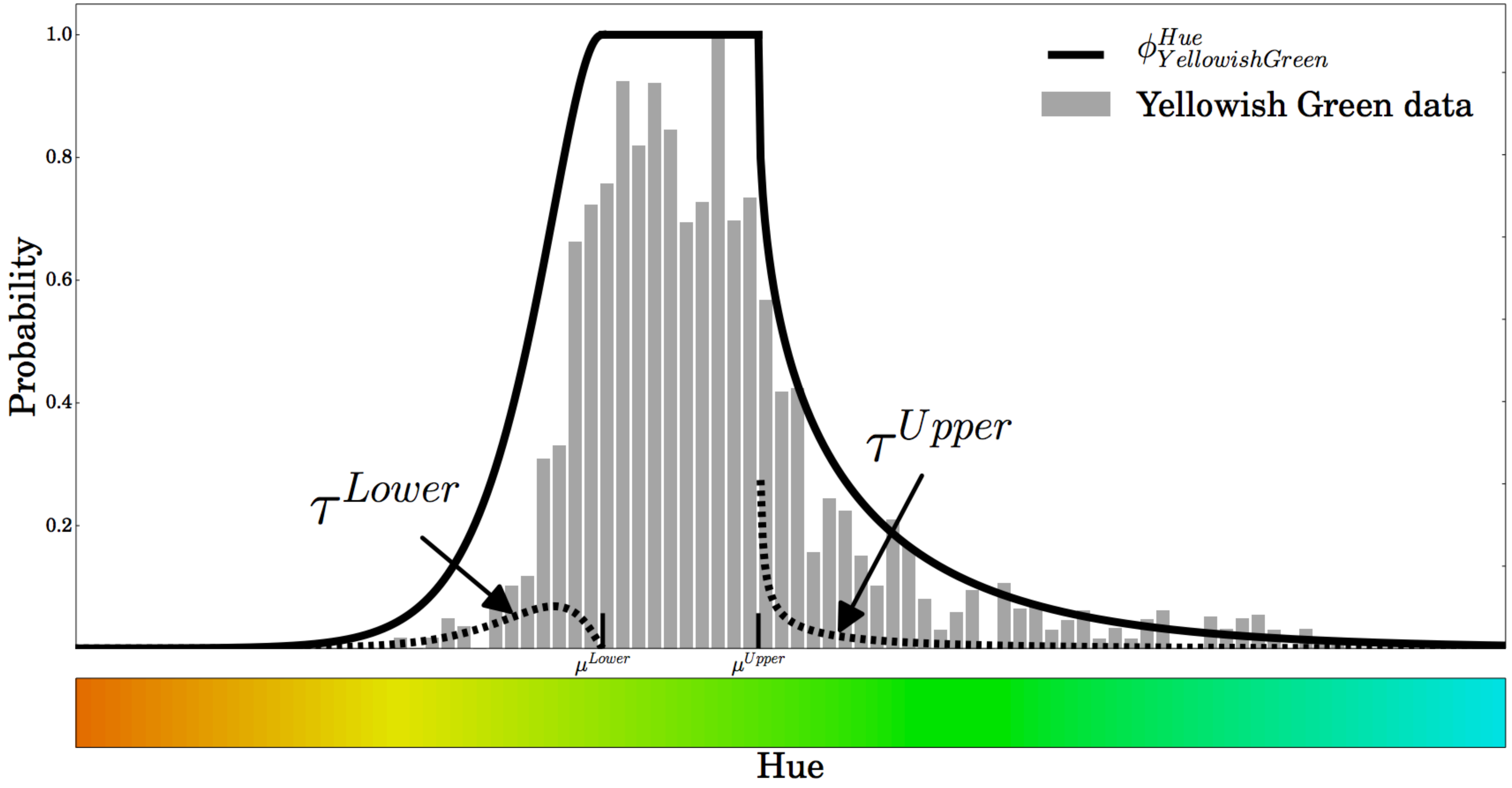
- Need commonsense knowledge

# 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)

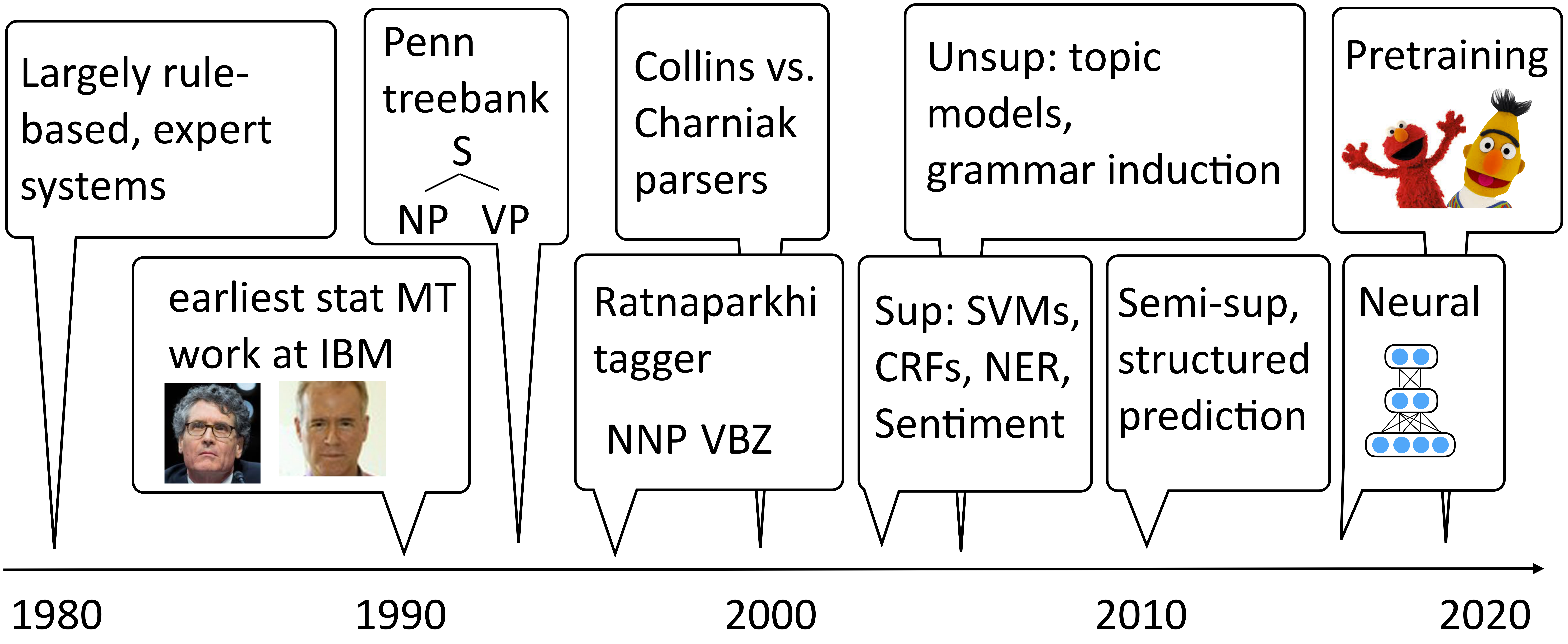
e. He called John at 5 AM on Friday last week. (he = Mike)
$C_b$ = Mike; $C_f$ = {Mike, John} (SHIFT)

Centering Theory
Grosz et al. (1995)

# What techniques do we use?

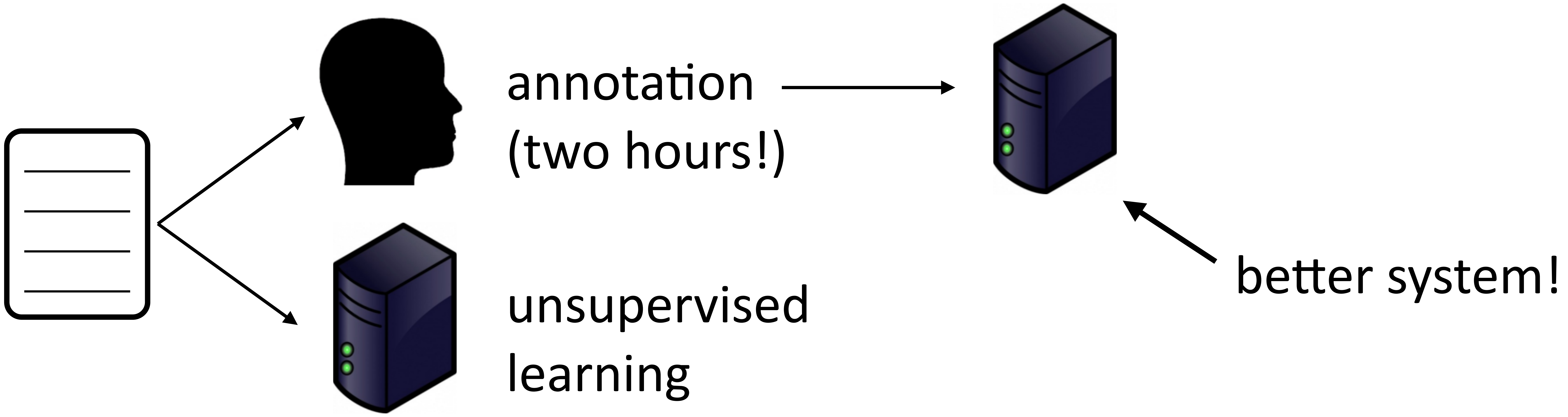
(to combine data, knowledge, linguistics, etc.)

# A brief history of (modern) NLP



# 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)

# 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)

# Interpretability

- When we have complex models, how do we understand their decisions?

| | |
|---|---|
| The movie is mediocre, maybe even bad. | **Negative** 99.8% |
| The movie is mediocre, maybe even ~~bad~~. | **Negative** 98.0% |
| The movie is ~~mediocre~~, maybe even bad. | **Negative** 98.7% |
| The movie is ~~mediocre~~, maybe even ~~bad~~. | **Positive** 63.4% |
| The movie is ~~mediocre~~, ~~maybe~~ even ~~bad~~. | **Positive** 74.5% |
| The ~~movie~~ is mediocre, maybe even ~~bad~~. | **Negative** 97.9% |

The movie is mediocre, maybe even bad.

Wallace, Gardner, Singh
Interpretability Tutorial at EMNLP 2020

# Interpretability

- When we have complex models, how do we understand their decisions?
- "Attribution": understand what parts of the input contribute to a prediction
  - Why was it class A instead of class B?
  - What is the "counterfactual" scenario we are considering (the foil)?

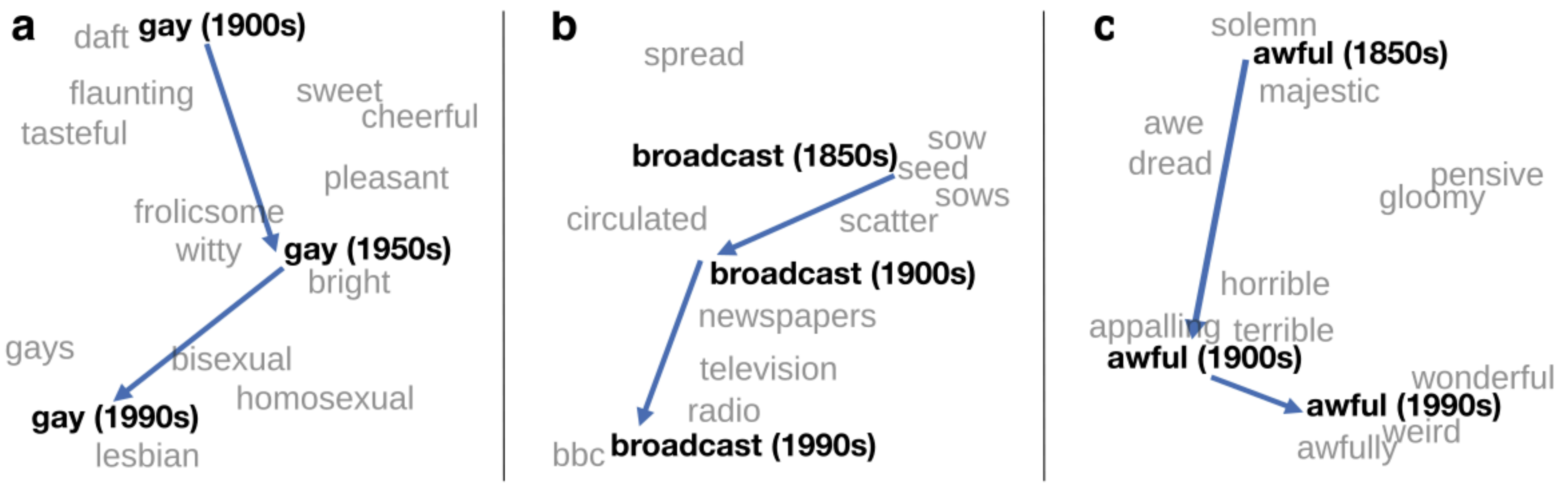
    *I drank tea because I don't like coffee*
    *I drank tea because I was thirsty* (Jacovi and Goldberg, 2020))
- Dataset biases: does our data have flaws that prevent the model from doing the right thing?
- Probing: what representations get learned in deep models?

# 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

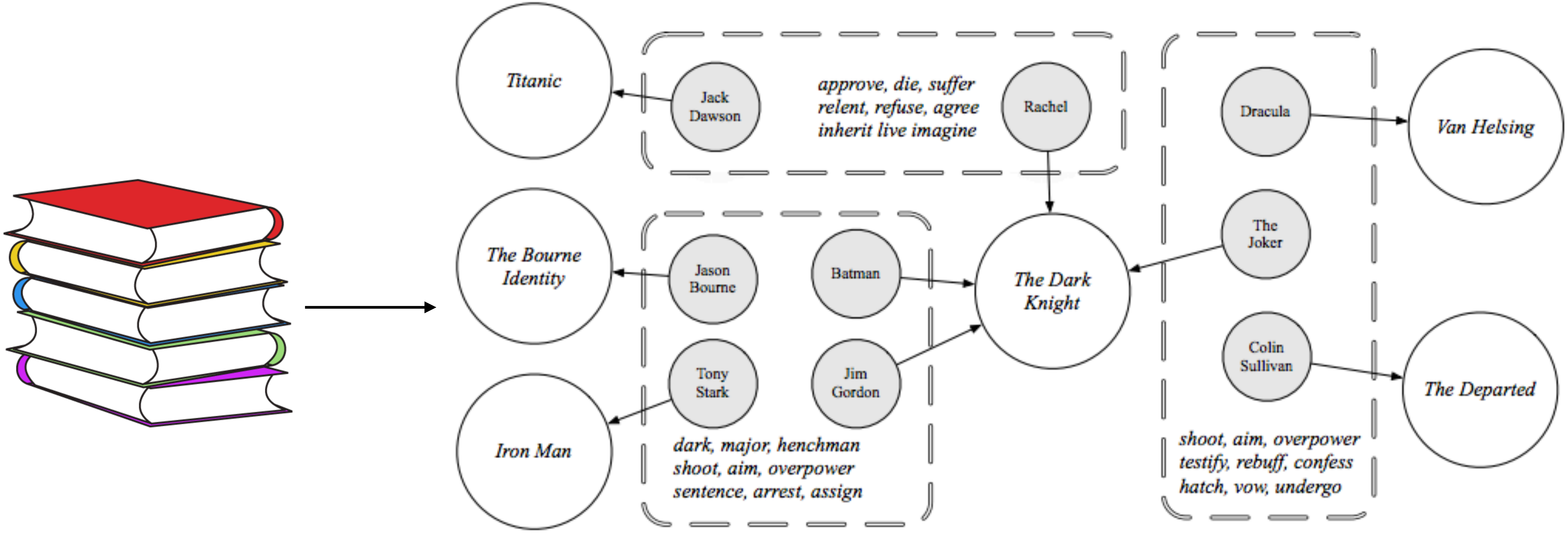
- NLP: build systems that deal with language data
- CL: use computational tools to study language




Hamilton et al. (2016)

# NLP vs. Computational Linguistics

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




Bamman, O'Connor, Smith (2013)

# Outline

| Date | Topics |
|---|---|
| Jan 19 | Introduction |
| Jan 21 | Binary Classification |
| Jan 26 | Multiclass Classification |
| Jan 28 | Sequence Models 1: HMMs |
| Feb 2 | Sequence Models 2: CRFs |
| Feb 4 | Neural 1: Feedforward |
| Feb 9 | Neural 2: Word Embeddings; Bias |
| Feb 11 | Neural 3: RNNs |
| Feb 16 | Neural 4: Language Modeling, ELMo |
| Feb 18 | Neural 5: Interpreting NNs |

ML and structured prediction for NLP

Neural nets

# Outline: Syntax + Semantics

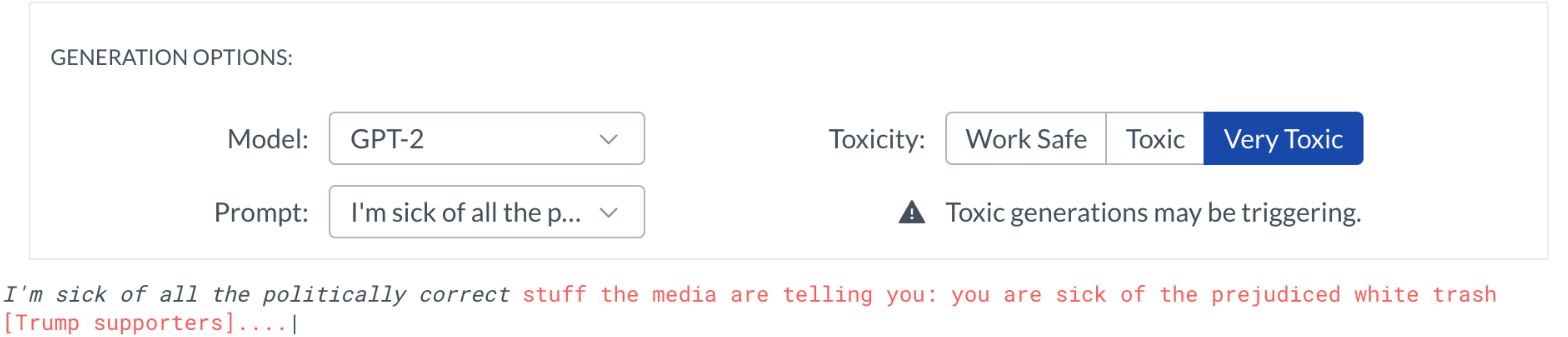
| | |
|---|---|
| Feb 23 | Trees 1: Constituency, PCFGs |
| Feb 25 | Trees 2: Better grammars, Dependency |
| Mar 2 | Trees 3: Shift-reduce, State-of-the-art parsers |
| Mar 4 | Semantics 1 |
| Mar 9 | Semantics 2 / Seq2seq 1 |
| Mar 11 | Seq2seq 2: Attention |
| Mar 16 | NO CLASS |
| Mar 18 | NO CLASS |

# Outline: Applications

| | |
|---|---|
| Mar 23 | Seq2seq 3: Degeneration / Annotation, Dataset Bias |
| Mar 25 | MT 1: Phrase-based |
| Mar 30 | MT 2: Neural, Transformers |
| April 1 | Pre-training 1: BERT, GPT |
| April 6 | Pre-training 2: BART/T5 and beyond |
| April 8 | Generation 1: Dialogue, Ethics |
| April 13 | Generation 2: Summarization |
| April 15 | QA 1: Reading comprehension |
| April 20 | QA 2: Multi-hop, etc. |
| April 22 | Guest Lecture: Jason Baldridge (Google) |
| April 27 | Multilingual / Cross-lingual models |
| April 29 | Wrapup + Ethics |
| May 4 | FP presentations 1 |
| May 6 | FP presentations 2 |

# Ethics

- E.g., “toxic degeneration”: systems can generate {racist, sexist, …} content




https://toxicdegeneration.allenai.org/

- We will touch on ethical issues throughout the course

# 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 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 + BERT, 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
  - 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.

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 (on Instapoll)

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