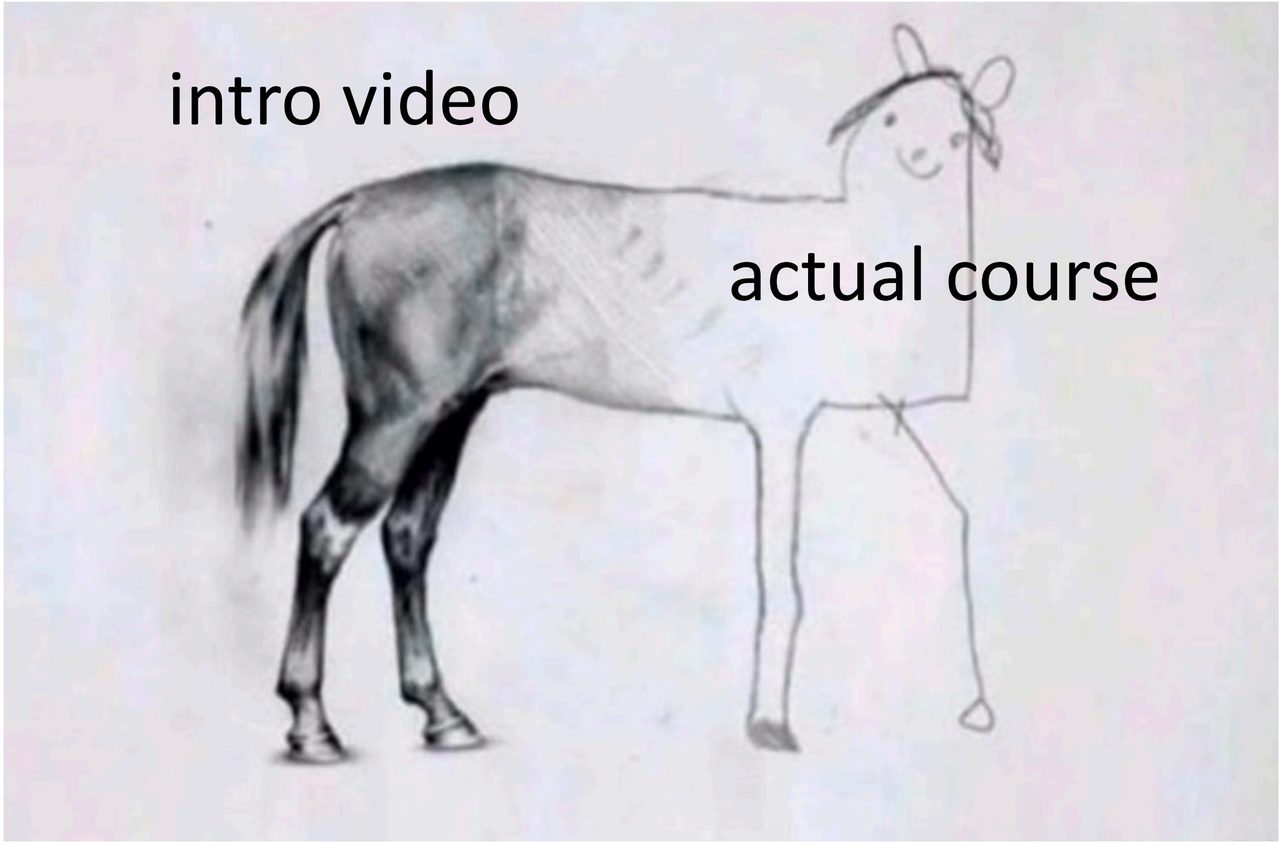
CS378: Natural Language Processing
Lecture 1: Introduction

Greg Durrett

TEXAS
The University of Texas at Austin



Administrivia

- Lecture: Tuesdays and Thursdays 9:30am - 10:45am
- Course website (including **syllabus**): http://www.cs.utexas.edu/~gdurrett/courses/fa2020/cs378.shtml
- Piazza: link on the course website
- My office hours: see course website
- TA: Tanya Goyal; Proctor: Shivang Singh. See website for OHs
- All office hours start next week, but I will stay around after this class if you have questions

Course Requirements

- CS 429
- Recommended: CS 331, familiarity with probability and linear algebra, programming experience in Python
- Helpful: Exposure to AI and machine learning (e.g., CS 342/343/363)

Enrollment

- If you are past 25 on the waitlist, you have a low chance of getting into the class, but we have to see how it progresses
- Assignment 0 is out now (optional):
  - If this seems like it'll be challenging for you, come and talk to me (this is smaller-scale than the other assignments, which are smaller-scale than the final project)

Format and Accessibility

- Lectures will build in time for discussion, in-class exercises, and questions. Additional material is available as videos to watch either before or after lectures
  - We'll do plenty of discussion groups in class. Piazza is also available to find teammates
- Required equipment: device to make Zoom calls with, some way to do homework
  - Lab machines available via SSH
  - A GPU is **not** required to complete the assignments! Having a GPU or GCP credits could be helpful **if** you want to pursue an independent project

What's the goal of NLP?

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

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.

Question Answering

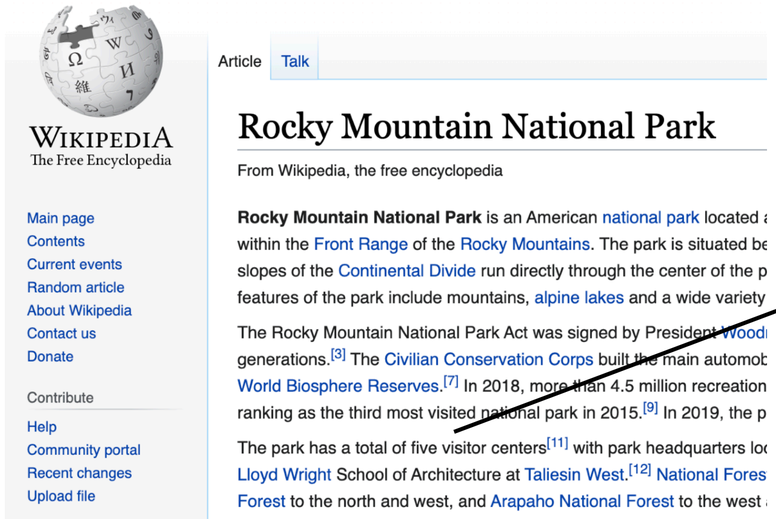
When was Abraham Lincoln born?

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?

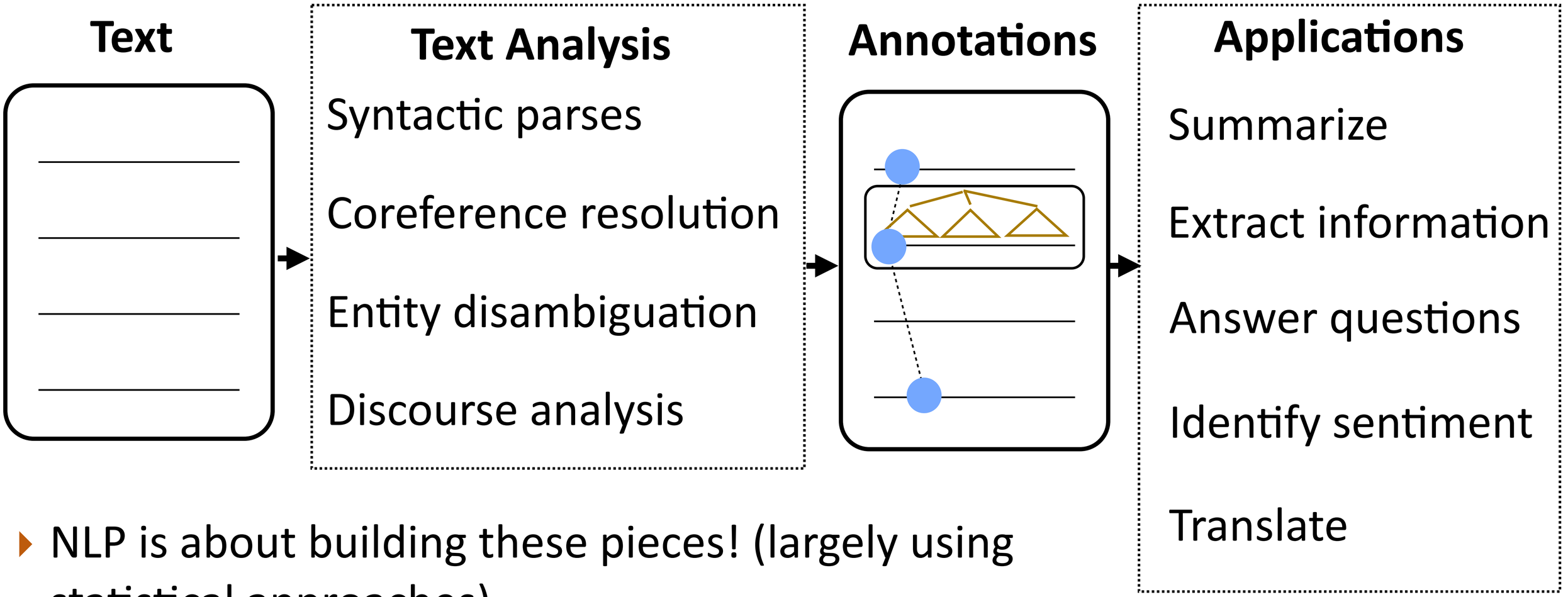


The park has a total of five visitor centers

↓

five

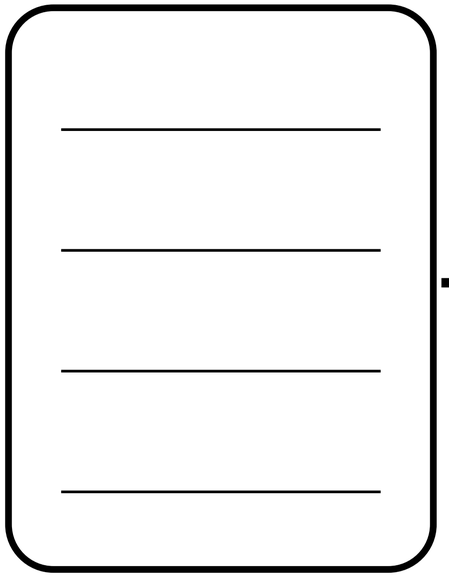
NLP Analysis Pipeline



- NLP is about building these pieces! (largely using statistical approaches)

How do we represent language?

Text



Labels

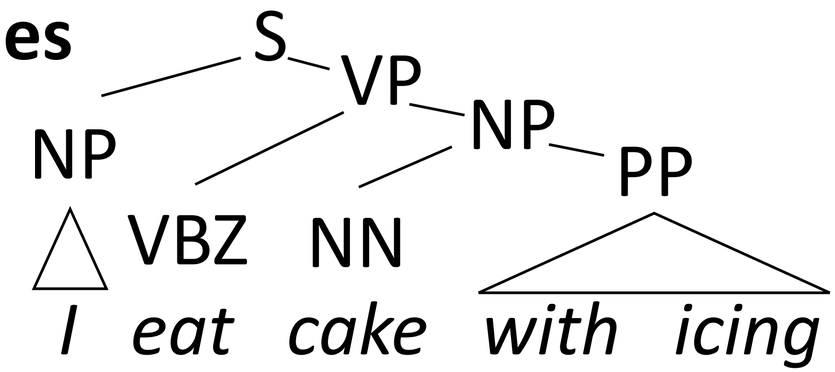
the movie was good +

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*

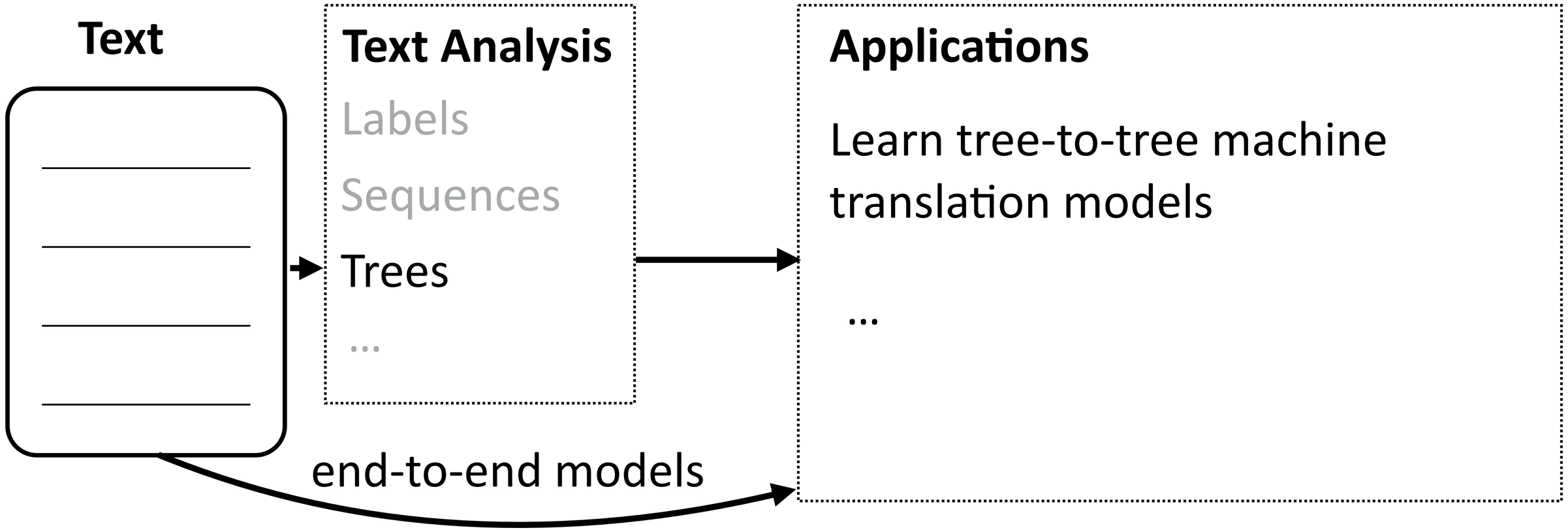
Trees



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

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? 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

Ban on Nude Dancing on Governor's Desk

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

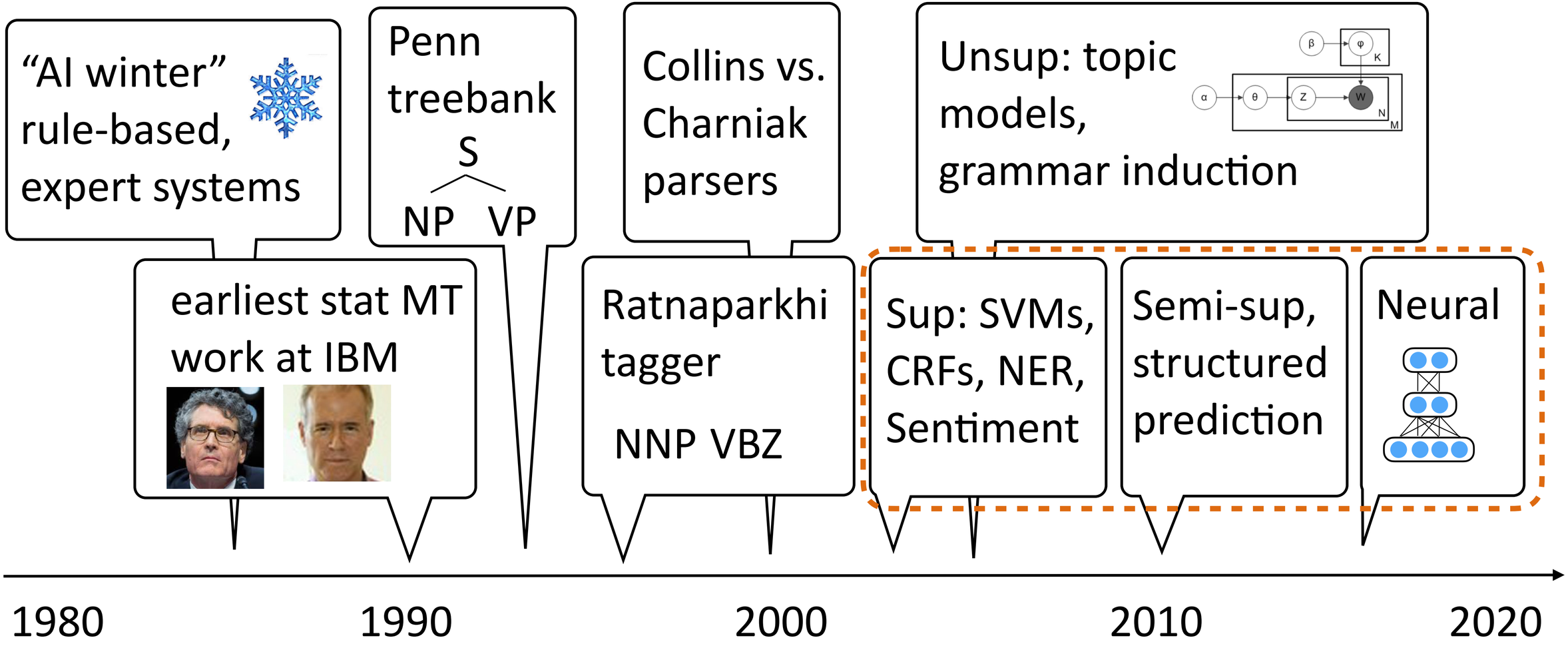
It fact actually handsome

- Combinatorially many possibilities, many you won't even register as ambiguities, but systems still have to resolve them

What techniques do we use?

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

A brief history of (modern) NLP

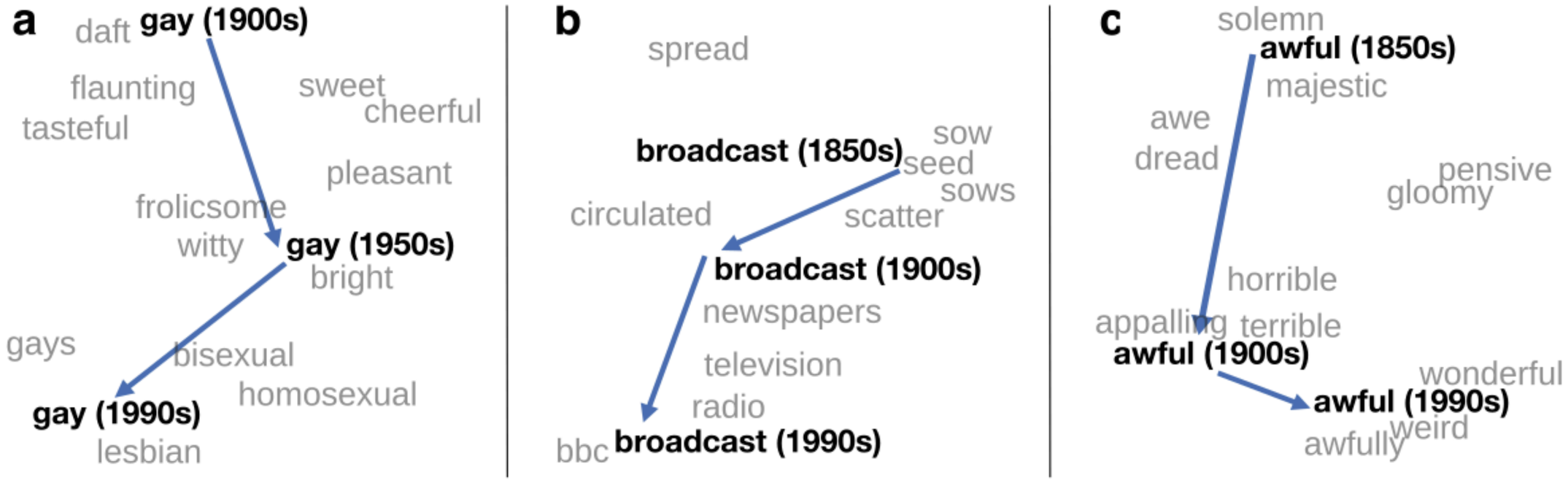


Where are we?

- 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)

Outline of the Course

- Classification: linear and neural, word representations (3.5 weeks)
- Text analysis: tagging and parsing (3 weeks) <= takes us to the midterm
- Generation, applications: language modeling, machine translation (3 weeks)
- Question answering, pre-training (2 weeks)
- Applications and miscellaneous (2.5 weeks)
- Goals:
  - Cover fundamental 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 2020?

Outline of the Course

- Throughout the course: ethics and fairness
  - Broader topic in ML than just NLP
  - How can we make sure our systems benefit society, and everyone in it?
  - Parts of lectures devoted to topics in ethics, comprehensive discussion on the last class day
- Balance algorithms, linguistics, data, ethics
- Nov 3: optional lecture

Coursework

- Five assignments, worth 45% of grade (A1-4: 10%, A5: 5%)
  - Mix of writing and implementation;
  - Assignment 0 is out now, optional diagnostic
  - ~2 weeks per assignment except for A5
  - 5 “slip days” throughout the semester to turn in assignments 24 hours late. Otherwise, you lose 15% credit per day the assignment is late
  - Submission on Gradescope

These assignments require understanding the concepts, writing performant code, and thinking about how to debug complex systems. **They are challenging; start early!**

The course staff are not here to debug your code! We **will** help you understand the concepts from lecture and come up with debugging strategies

Coursework

- Midterm (20% of grade), take-home October 14-16
  - Similar to written homework problems
- Final project (25% of grade)
  - Groups of 1 or 2
  - Standard project: neural network models for question answering
  - Independent projects are possible: these must be proposed earlier (to get you thinking early) and will be held to a high standard!
- In-class problems (10% of the grade)
  - These will be done via UT Instapoll. You don't have to come to class to do them
  - Drop the lowest 5

Academic Honesty

- You may work in groups, but your final writeup and code **must be your own**
- Don’t share code with others!

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