CS4650: Natural Language Processing

Wei Xu

(many slides from Greg Durrett)

Administrivia

- Course website: https://www.cc.gatech.edu/classes/AY2021/cs4650_spring/
- Piazza/Gradescope: link on the course website
- My office hour: TBA
- TAs:

Jingfeng Yang

Kaige Xie

Sarmishta Velury

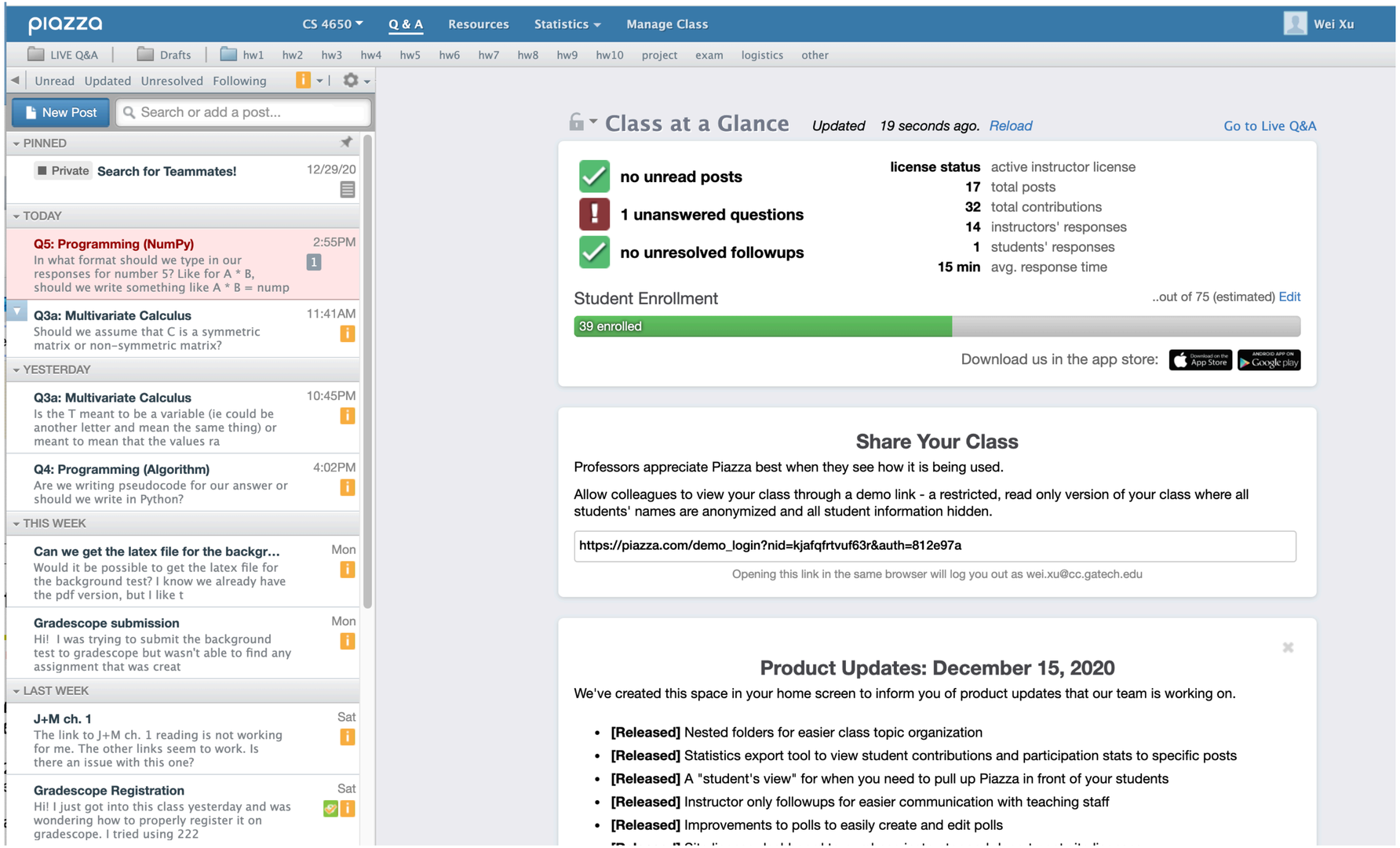
- TA office hours: Mon 6:00-7:00pm, Wed 6:00-7:00pm, Thu 10:00-11:00pm

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!

Piazza


piazza
CS 4650
Q & A
Resources
Statistics
Manage Class
Wei Xu
LIVE Q&A
Drafts
hw1 hw2 hw3 hw4 hw5 hw6 hw7 hw8 hw9 hw10 project exam logistics other
Unread Updated Unresolved Following
New Post
Search or add a post...
PINNED
Private Search for Teammates! 12/29/20
TODAY
Q5: Programming (NumPy) 2:55PM
In what format should we type in our responses for number 5? Like for A * B, should we write something like A * B = nump
Q3a: Multivariate Calculus 11:41AM
Should we assume that C is a symmetric matrix or non-symmetric matrix?
YESTERDAY
Q3a: Multivariate Calculus 10:45PM
Is the T meant to be a variable (ie could be another letter and mean the same thing) or meant to mean that the values ra
Q4: Programming (Algorithm) 4:02PM
Are we writing pseudocode for our answer or should we write in Python?
THIS WEEK
Can we get the latex file for the backgr... Mon
Would it be possible to get the latex file for the background test? I know we already have the pdf version, but I like t
Gradescope submission Mon
Hi! I was trying to submit the background test to gradescope but wasn't able to find any assignment that was creat
LAST WEEK
J+M ch. 1 Sat
The link to J+M ch. 1 reading is not working for me. The other links seem to work. Is there an issue with this one?
Gradescope Registration Sat
Hi! I just got into this class yesterday and was wondering how to properly register it on gradescope. I tried using 222
Class at a Glance
Updated 19 seconds ago. Reload
Go to Live Q&A
no unread posts
1 unanswered questions
no unresolved followups
license status active instructor license
17 total posts
32 total contributions
14 instructors' responses
1 students' responses
15 min avg. response time
Student Enrollment
..out of 75 (estimated) Edit
39 enrolled
Download us in the app store:
Share Your Class
Professors appreciate Piazza best when they see how it is being used.
Allow colleagues to view your class through a demo link - a restricted, read only version of your class where all students' names are anonymized and all student information hidden.
https://piazza.com/demo_login?nid=kjafqfrtvuf63r&auth=812e97a
Opening this link in the same browser will log you out as wei.xu@cc.gatech.edu
Product Updates: December 15, 2020
We've created this space in your home screen to inform you of product updates that our team is working on.
[Released] Nested folders for easier class topic organization
[Released] Statistics export tool to view student contributions and participation stats to specific posts
[Released] A "student's view" for when you need to pull up Piazza in front of your students
[Released] Instructor only followups for easier communication with teaching staff
[Released] Improvements to polls to easily create and edit polls

Enrollment and Prereq

- Background Test (5%) is out now (due tomorrow 1/21):
  - Designed to help you determine whether you have enough math and programming background to succeed in this class.
  - If the background test is not enough for calibrating, you may read Chapters 2~4 of the textbook by Jacob Eisenstein. We will cover these content in the first few weeks of the semester.

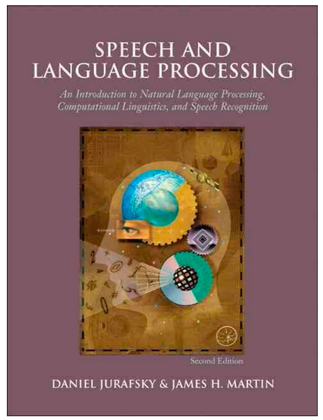
Textbooks

- 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

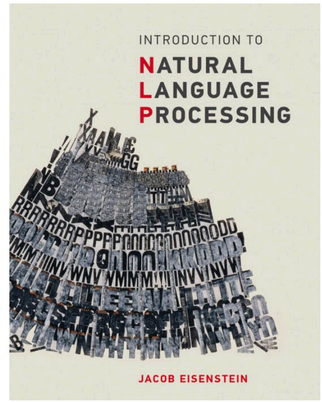
Here's our December 30, 2020 draft! Includes:



Introduction to Natural Language Processing

By Jacob Eisenstein

Published by The MIT Press
Oct 01, 2019 | 536 Pages | 7 x 9
| ISBN 9780262042840



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 (or four) programming assignments should teach you what you need to know to understand nearly any system in the literature

Assignments

- Four Homework Assignments (55%)
  - Implementation-oriented
  - Homework 1 will be out soon
  - ~2 weeks per assignment, 3 “slip days” for **up to 2** homework (3 days each)
    - Homework 1: 15% (written + 1st programming)
    - Homework 2: 10% (written)
    - Homework 3: 15% (written + 2nd programming)
    - Homework 4: 15% (written + 3rd programming)

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!**

Final Project, etc.

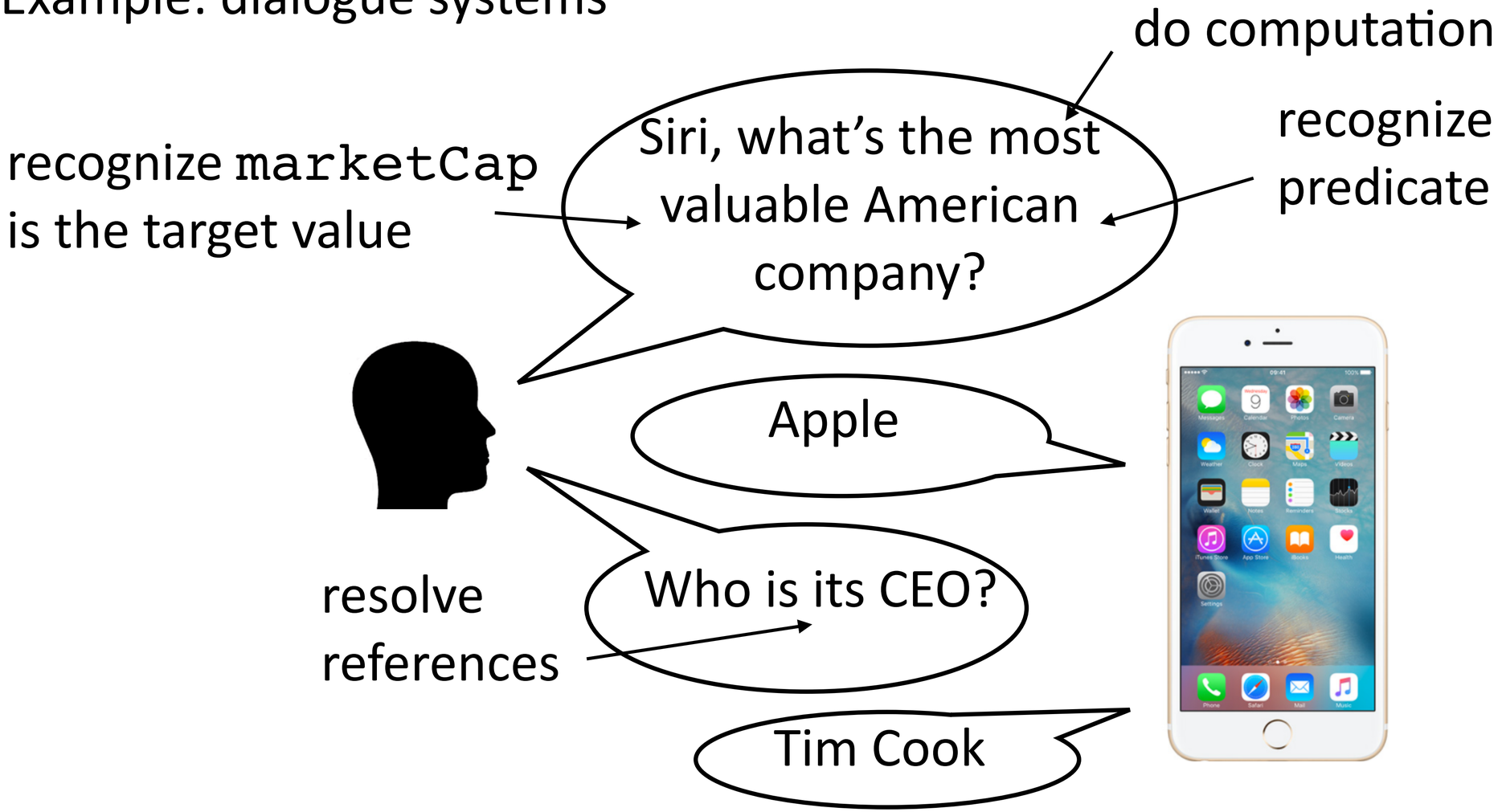
- No Midterm

- Final project (20%)
  - Groups of 2-3 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.
  - **Alternatively,** you may choose to complete the Homework 5 (written+programming) individually

- Quizzes (10%)
- Participation (10%)

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

Machine Translation

People's Daily, August 30, 2017

Trump and his family watched a 100-year total solar eclipse on the balcony of the White House

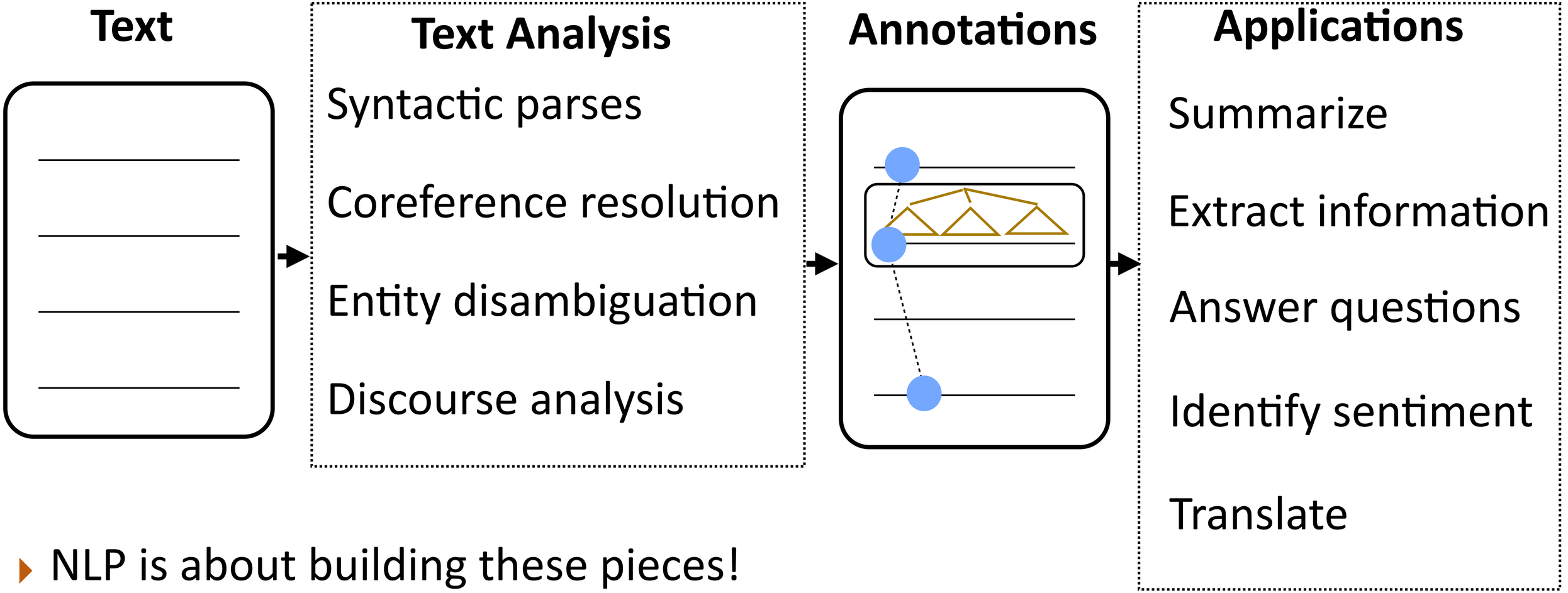
Textual Entailment

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.

SNLI (Bowman et al., 2015)

- Text is connected to intelligence and knowledge in a fundamental way!
- Goal of NLP (solving problems with text) requires *analyzing* and *understanding* text
- What makes this analysis hard?

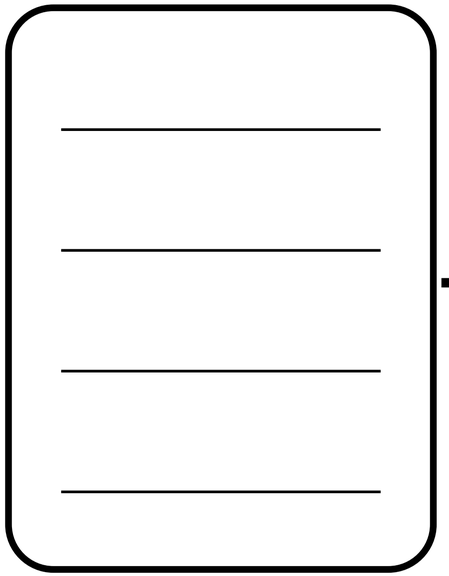
NLP Analysis Pipeline



- 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

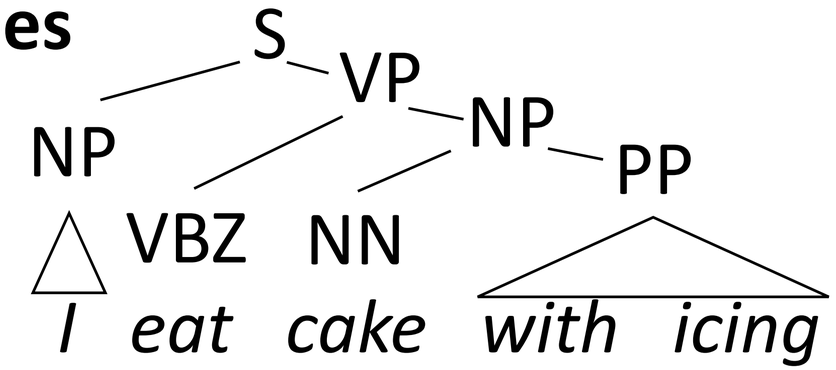
the movie was good +

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*

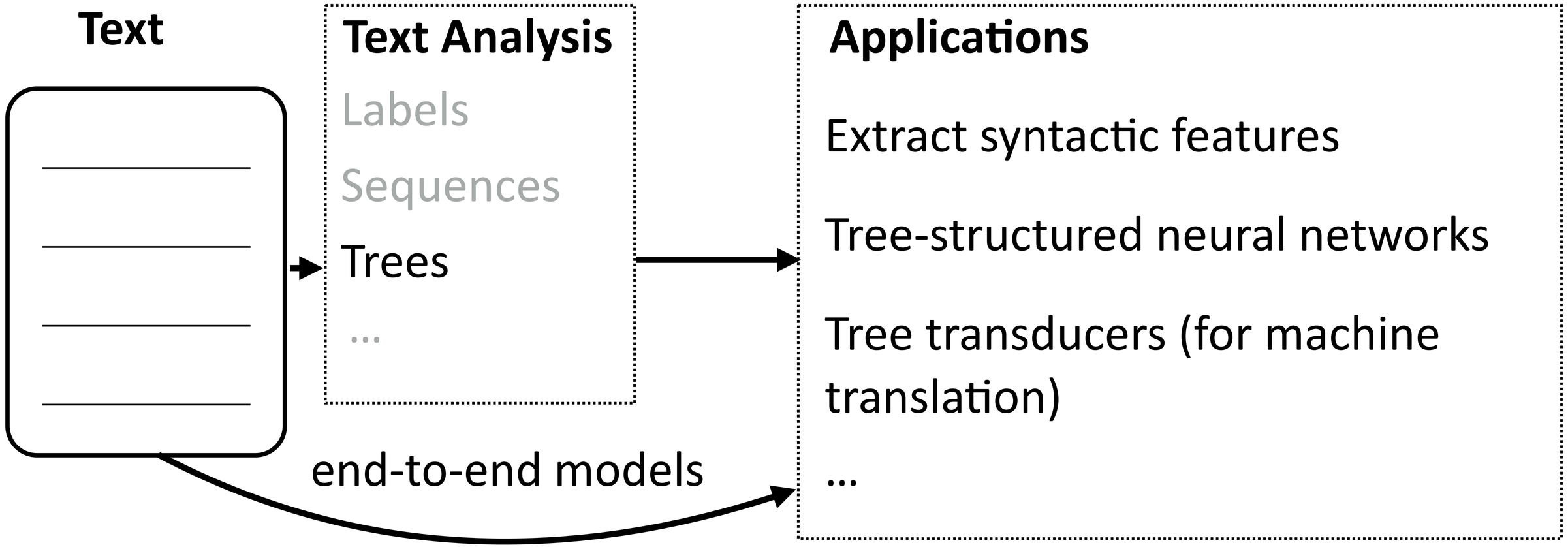
Trees



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

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!

Students Cook & Serve Grandparents

On Thursday, September 9, Gorman School hosted the first annual Grandparent's Day.

All Grandparents were invited to a school wide pancake breakfast. Upper grade students served as excellent chefs, as well as taking responsibility for serving the food and the clean up after-

Language is Ambiguous!

- Other 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 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 **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

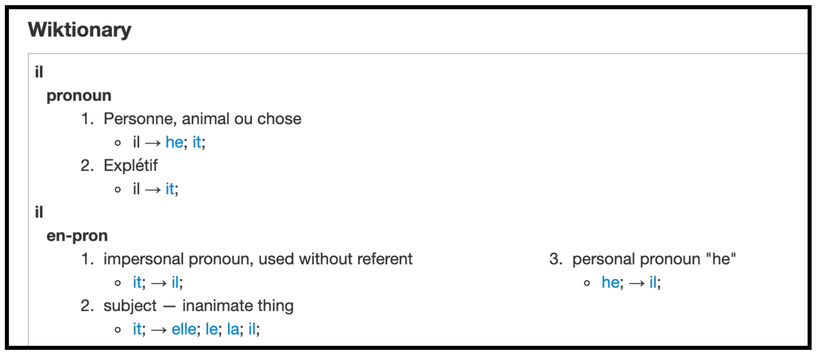
The weather is beautiful

It is really beautiful outside

He makes truly beautiful

He makes truly boyfriend

It fact actually handsome



Wiktionary

il

pronoun

1. Personne, animal ou chose
   - il → he; it;
2. Explétif
   - il → it;

il

en-pron

1. impersonal pronoun, used without referent
   - it; → il;
2. subject — inanimate thing
   - it; → elle; le; la; il;
3. personal pronoun "he"
   - he; → il;

- 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*

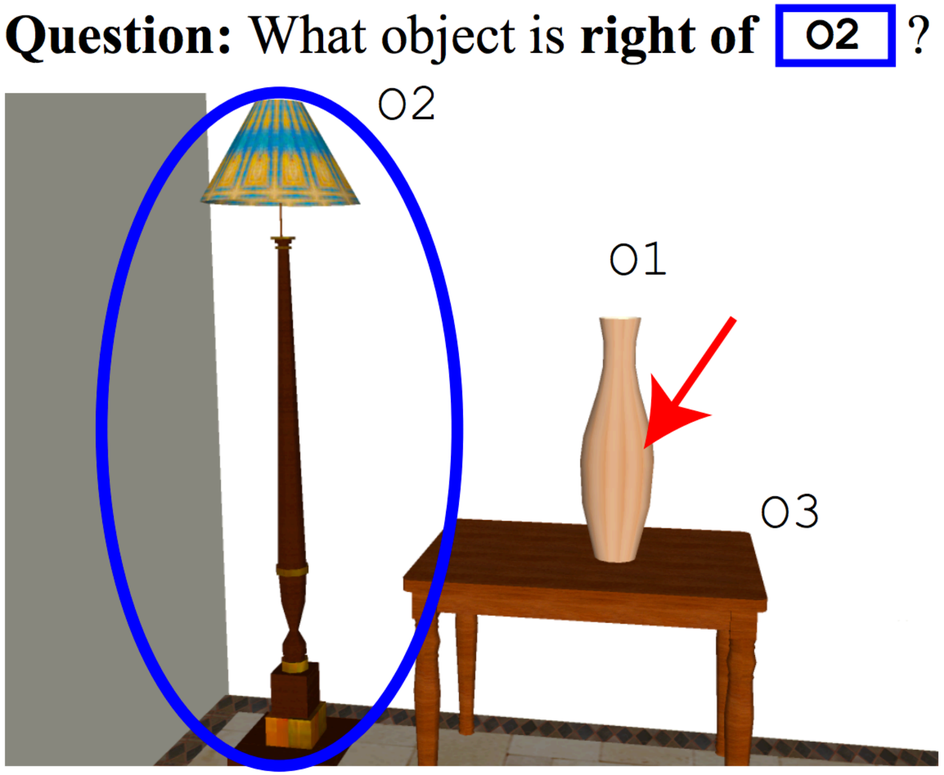
Department of Justice

metaphor;
"approves"

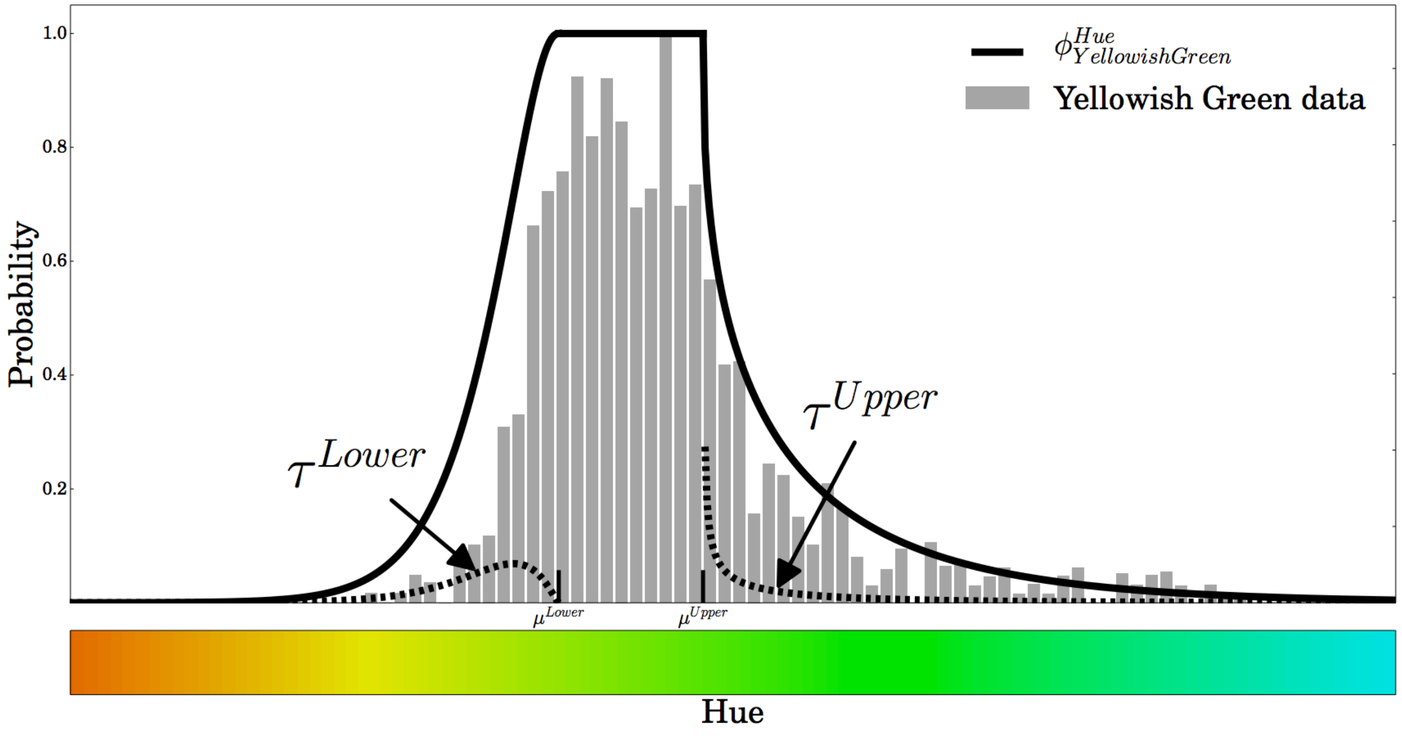
- What is a green light? How do we understand what "green lighting" does?

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} ← forward center

backward center →

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)

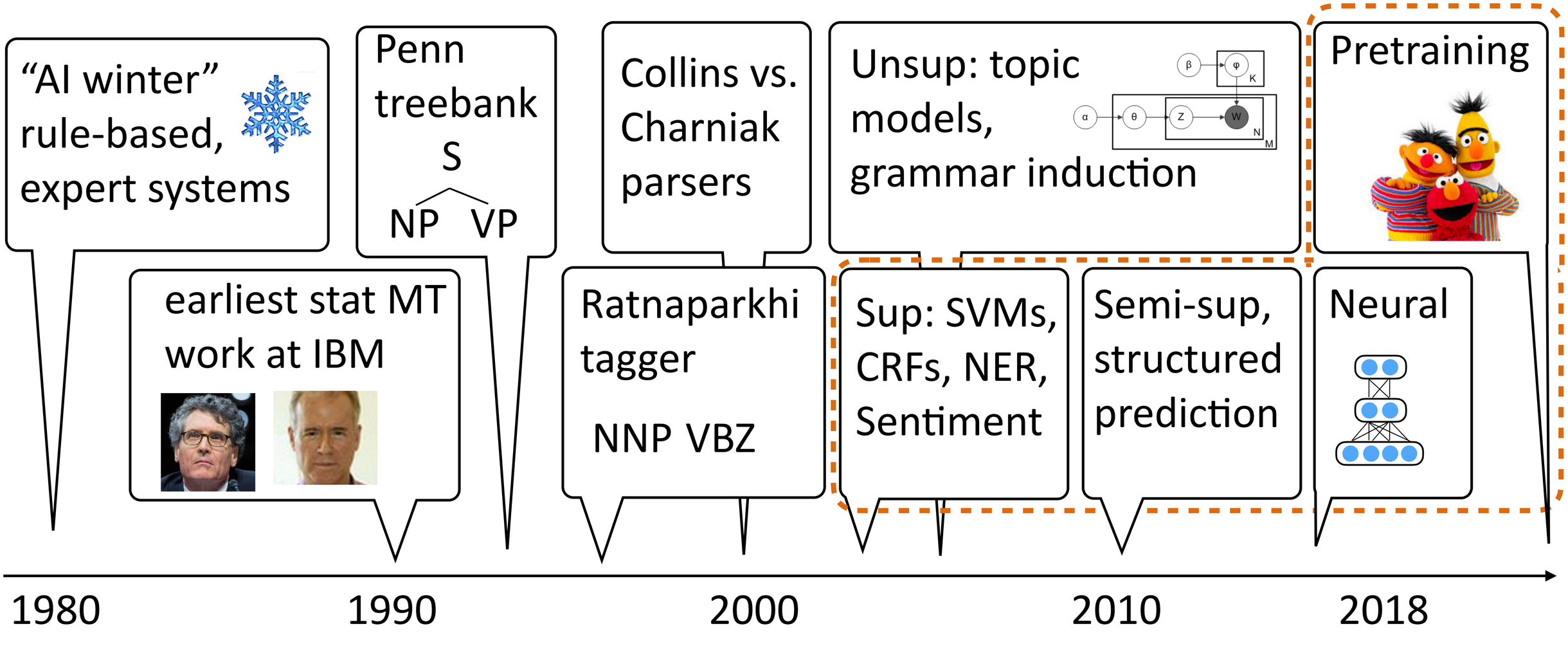
e. He called John at 5 AM on Friday last week. (he = Mike)
C_b = Mike; C_f = {Mike, John} (SHIFT)

Centering Theory
Grosz et al. (1995)

What techniques do we use?

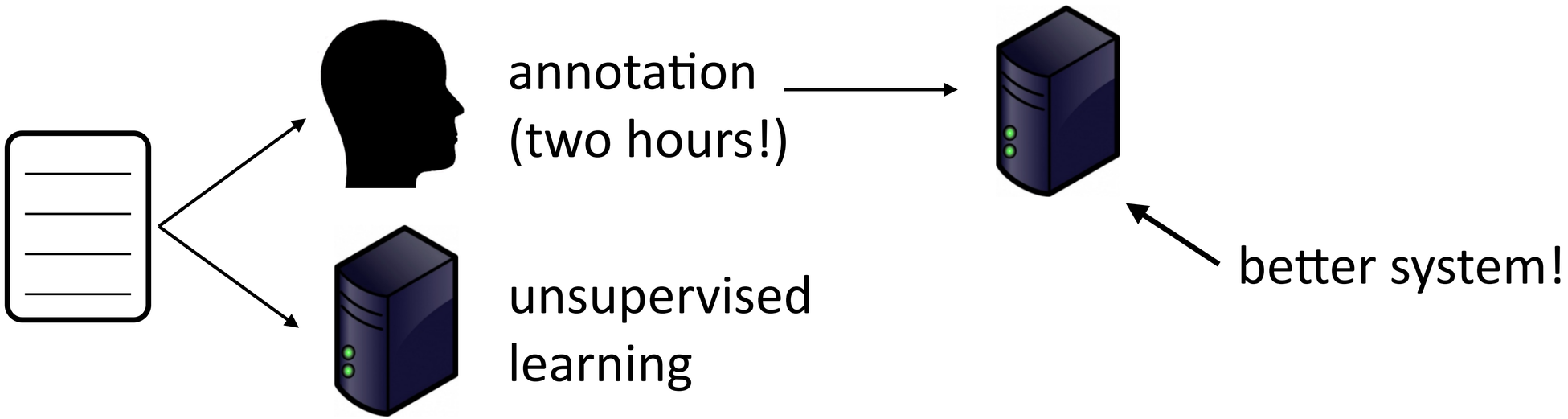
(to combine data, knowledge, linguistics, etc.)

A brief history of (modern) NLP



Structured Prediction

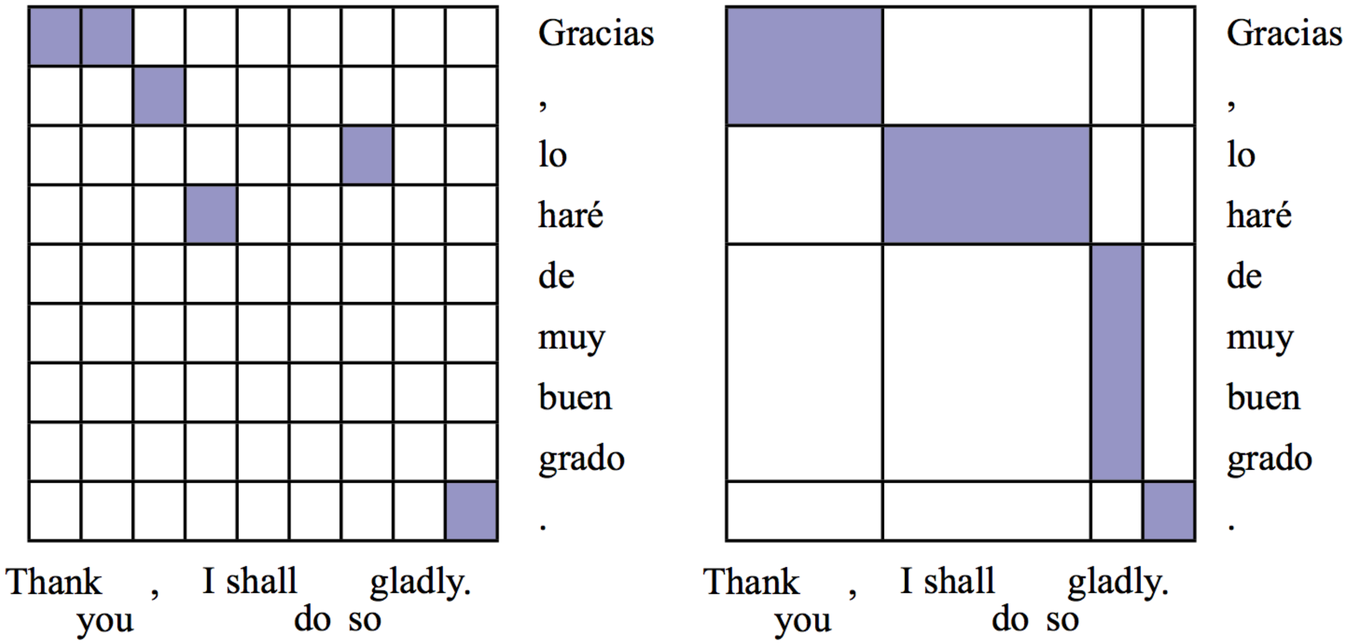
- All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- Supervised techniques work well on very little data



- Even neural nets can do pretty well!

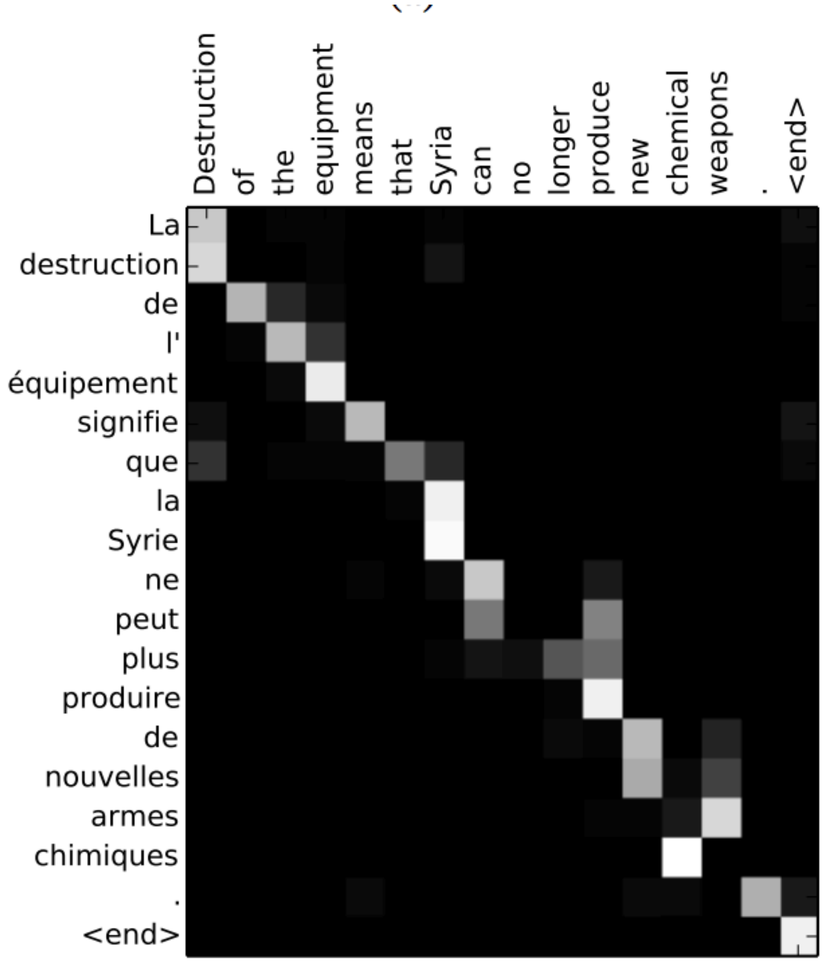
"Learning a Part-of-Speech Tagger from Two Hours of Annotation"
Garrette and Baldridge (2013)

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

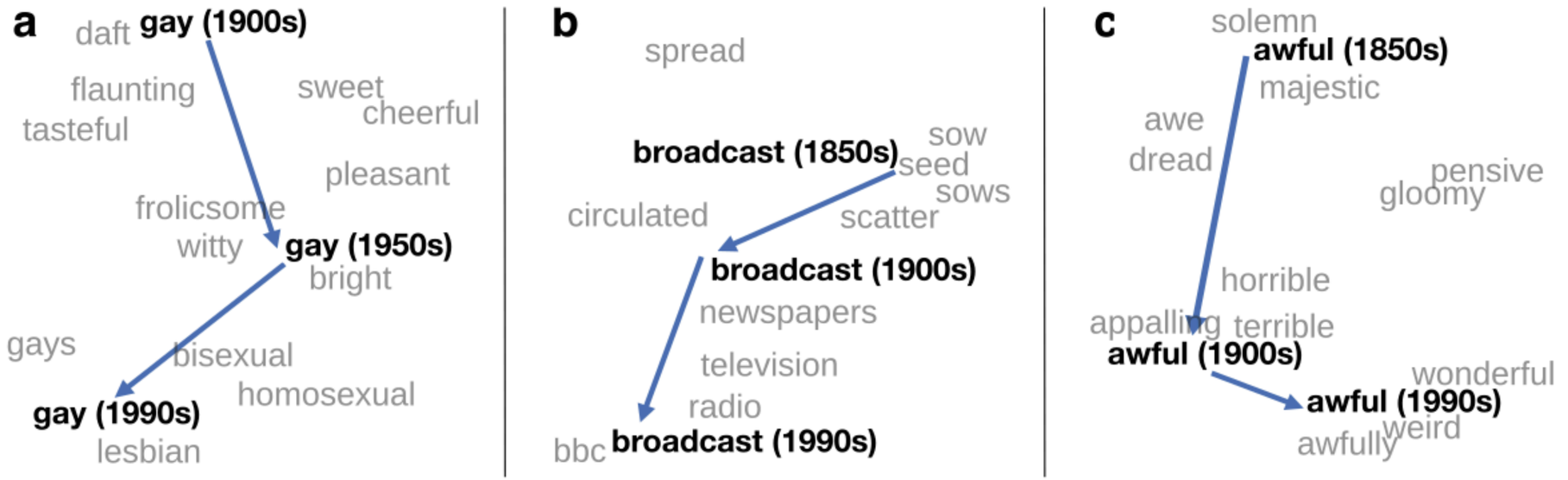
Moosavi and Strube (2017)

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

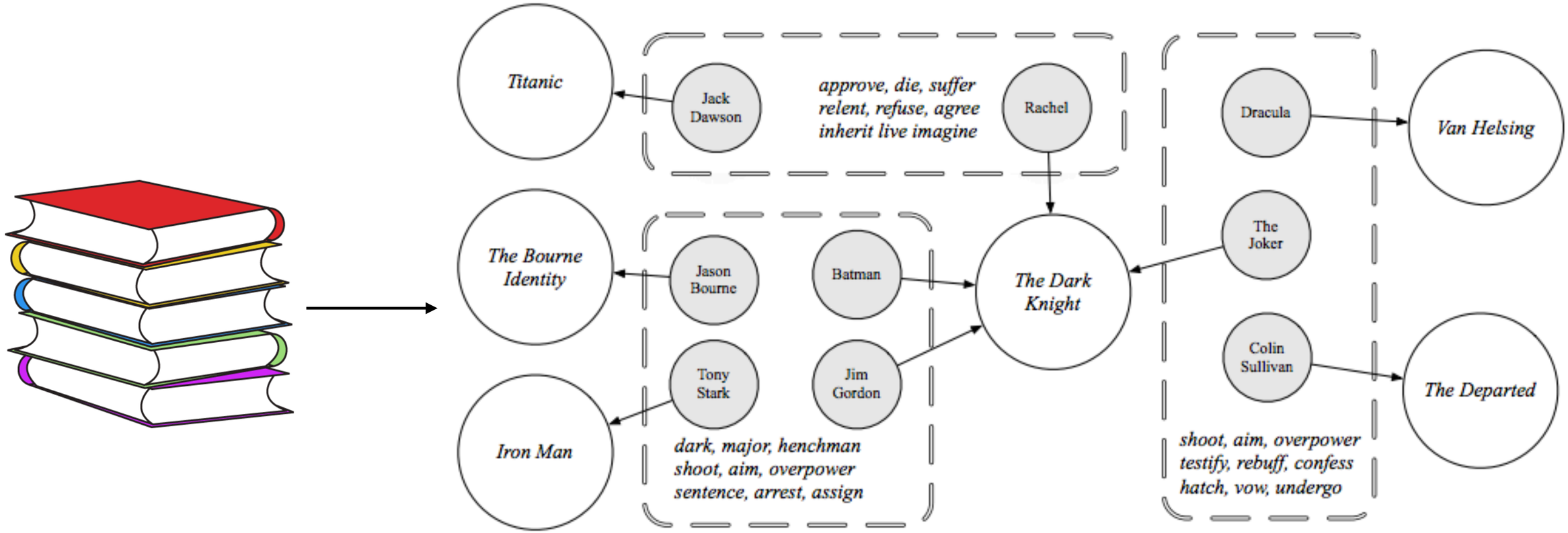
- NLP: build systems that deal with language data
- CL: use computational tools to study language



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

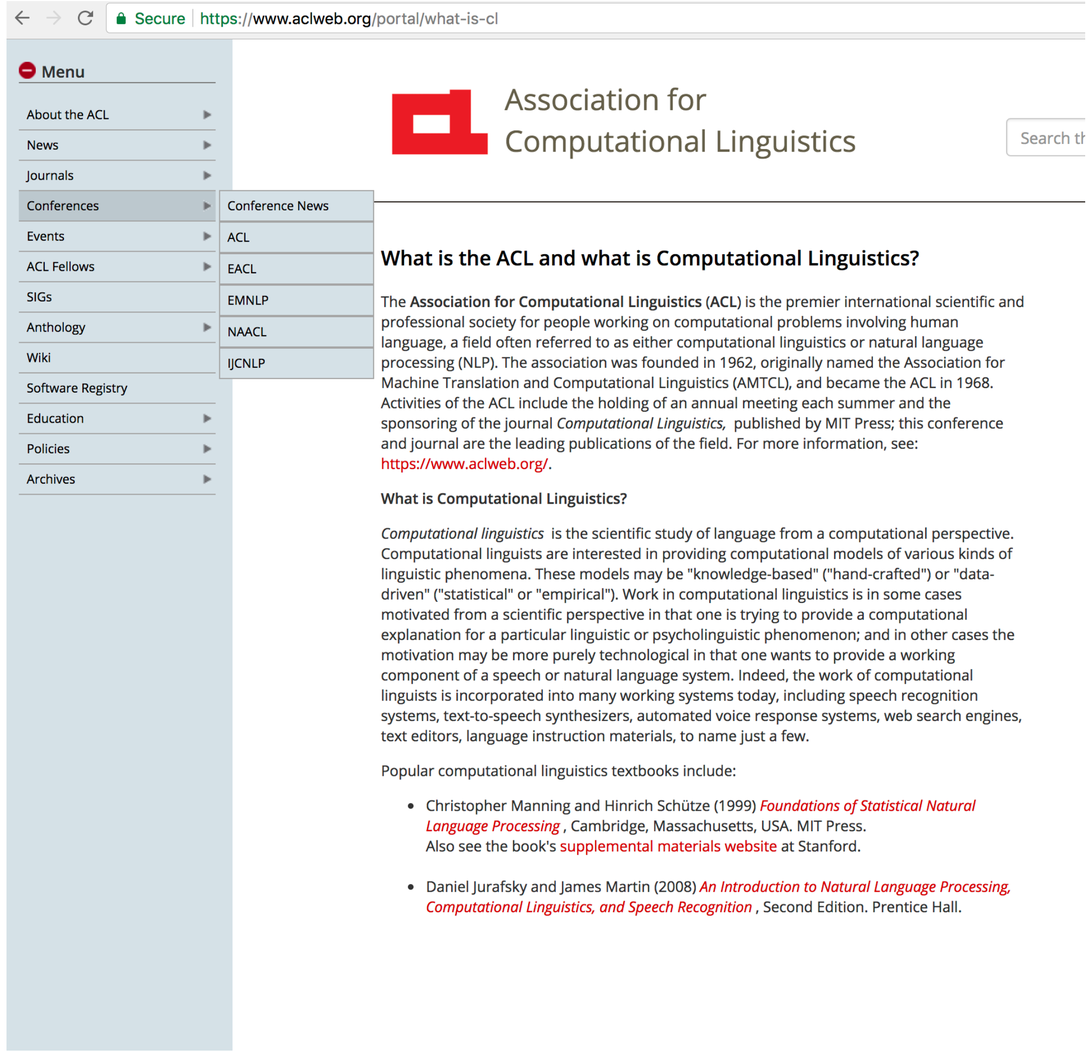
Neural Networks semantics

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

Date	Topics (tentative and subject to change)	Readings
1/14/2021	first day of class	
1/18/2021	No class - MLK national holiday	
1/20/2021	**Course Overview - 1st lecture**	J+M 1
1/25/2021	Binary Classification (naive bayes and logisitic regression)	J+M 4, Eisenstein 2.0-2.5, 4.1,4.3-4.5,
1/27/2021	Multiclass Classification	J+M 5, Eisenstein 4.2
2/1/2021	Neural Networks (feedforward networks)	Eisenstein 3.1-3.3, J+M 7.1-7.4
2/3/2021	Neural Networks (back propogation)	Eisenstein 3.1-3.3, J+M 7.1-7.4
2/8/2021	PyTorch Tutorial, Sequence Models	J+M 8
2/10/2021	Viterbi Algorithm	Eisenstein 7.0-7.4
2/15/2021	Conditional Random Fields	Eisenstein 7.5, 8.3
2/17/2021	N-gram Language Models	
2/22/2021	Word Embeddings	Eisenstein 3.3.4, 14.5, 14.6, J+M 6
2/24/2021	Recurrent Neural Networks	J+M 9, Goldberg 10,11
3/1/2021	Convolutional Neural Networks	Goldberg 9, Eisenstein 3.4, 7.6
3/3/2021	Statistical Machine Translation	Eisenstein 18.1, 18.2
3/8/2021	**(Guest Lecture)**	
3/10/2021	Sequence-to-Sequence Model	J+M 10
3/15/2021	Attention and Copy Mechanism	Eisenstein 18.3, 18.4
3/17/2021	Question Answering / Reading Comprehension	SQuaD, BiDAF
3/17/2021	Withdrawal deadline	
3/22/2021	Parsing	
3/24/2021	No class - mid-semester break	
3/29/2021	Neural Machine Translation	Google NMT
3/31/2021	Transformer Model	Attention is all you need
4/5/2021	**Generation (Guest Lecture)**	
4/7/2021	Information Extraction	Eisenstein 13, 17
4/12/2021	**Dialog (Guest Lecture)**	
4/14/2021	Pre-trained Language Models / BERT	
4/19/2021	**Computational Social Science (Guest Lecture)**	
4/21/2021	Speech Recognition	
4/26/2021	Final class day	

NLP Research

	Area	Long submissions	Accepts	Accept rate (%)
1.	Applications	65	14	28.8
2.	Dialogue and Interactive Systems	126	38	30.2
3.	Discourse and Pragmatics	33	7	21.2
4.	Document Analysis	48	8	16.7
5.	Generation	96	32	33.3
6.	Information Extraction and Text Mining	155	37	23.9
7.	Linguistic Theories, Cognitive Modeling and Psycholinguistics	39	9	23.1
8.	Machine Learning	148	38	25.7
8.	Machine Translation	102	27	26.5
10.	Multidisciplinary and Area Chair COI	69	21	30.4
11.	Multilinguality	43	11	25.6
12.	Phonology Morphology and Word Segmentation	26	7	26.9
13.	Question Answering	99	32	32.3
14.	Resources and Evaluation	70	26	37.1
15.	Sentence-level semantics	69	14	20.3
15.	Sentiment Analysis and Argument Mining	91	24	26.4
17.	Social Media	51	14	27.5
18.	Summarization	48	11	22.9
19.	Tagging Chunking Syntax and Parsing	50	17	34.0
20.	Textual Inference and Other Areas of Semantics	44	16	36.4
21.	Vision Robotics Multimodal Grounding and Speech	56	20	35.7
22.	Word-level Semantics	78	20	25.6



ACL 2019 conference

ACL'19 at a Glance