



Apple



Google



Microsoft



Samsung



Twitter



WhatsApp



Facebook



JoyPixels

# CS 4650: Natural Language Processing

Wei Xu

(many slides from Greg Durrett)

# Administrivia

---

- ▶ Course website:  
[https://cocoxu.github.io/CS4650\\_spring2022/](https://cocoxu.github.io/CS4650_spring2022/)
- ▶ TAs (office hours start from next week)
- ▶ Piazza and Gradescope:
  - ▶ links on the course website
  - ▶ We will do our best to make sure questions about the homework, etc. get answered within 24 hours

## Instructor



[Wei Xu](#)

[wei.xu@cc.gatech.edu](mailto:wei.xu@cc.gatech.edu)

## Teaching Assistants



[Chao Jiang \(Head TA\)](#)

[chaojiang@gatech.edu](mailto:chaojiang@gatech.edu)

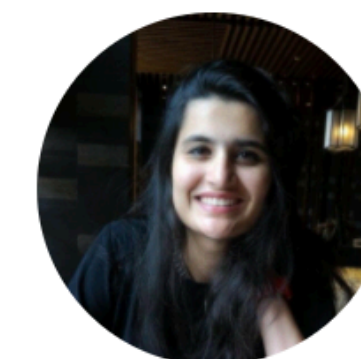
Office Hours: Tuesday 4:00–5:00pm



[Chase Perry](#)

[cperry65@gatech.edu](mailto:cperry65@gatech.edu)

Office Hours: Friday 10:00–11:00am



[Rucha Sathe](#)

[ruchasathe@gatech.edu](mailto:ruchasathe@gatech.edu)

Office Hours: Thursday 1:00–2:00pm

# About Me

---

- ▶ Tenure-track Assistant Professor (since 2016)
- ▶ 2014-2016 Postdoctoral researcher at UPenn
- ▶ 2014 Ph.D. in Computer Science from NYU
- ▶ Research in Natural Language Processing / Machine Learning (40+ publications)
  - ▶ National Science Foundation CRII Award
  - ▶ CrowdFlower AI for Everyone Award
  - ▶ COLING'18 Best Paper Award
- ▶ IARPA BETTER (\$850k), NSF CyberLearning (\$377k), DARPA SocialSim (\$600k)
- ▶ NSF AI Institute (AI-CARING)



# NLP X Research Lab

We design machine learning algorithms to help computer to understand and generate human languages.

X = Machine Learning, Social Media, Education, Accessibility, ...

## Natural Language Processing

- Text generation and evaluation
- Semantics and structured prediction
- Misinformation, social media
- Offensive/biased language
- Interactive NLP systems
- Pre-training language models

## Machine Learning

- Sequence-to-sequence models
- Controllability of neural networks
- Cross-lingual transfer learning
- Learning from noisy data
- Data augmentation



### PhD students



Mounica  
Maddela



Chao  
Jiang



Yao  
Dou

### MS students



Chase  
Perry

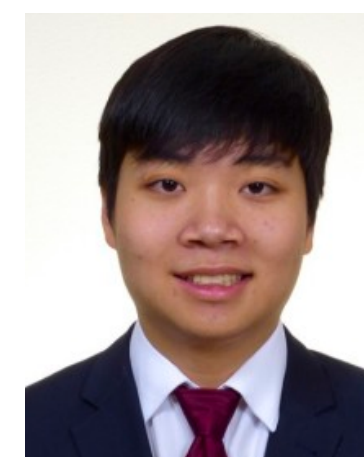


Rucha  
Sathe

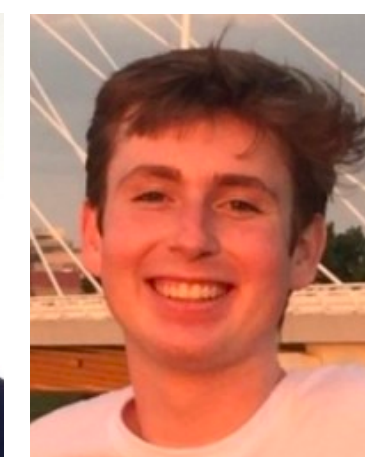


Angana  
Borah

### Undergrad students



Jonathan  
Zheng



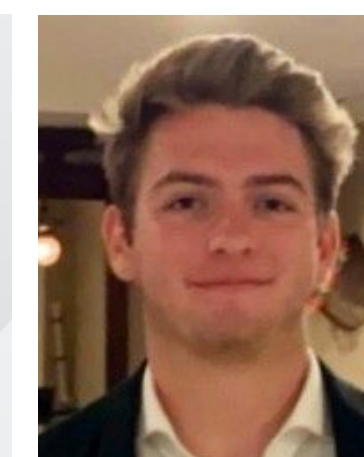
David  
Heineman



Michael  
Ryan



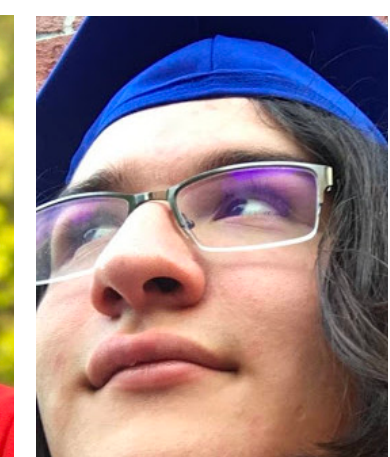
Srushti  
Nandu



Dylan  
Small



Vishnu  
Suresh



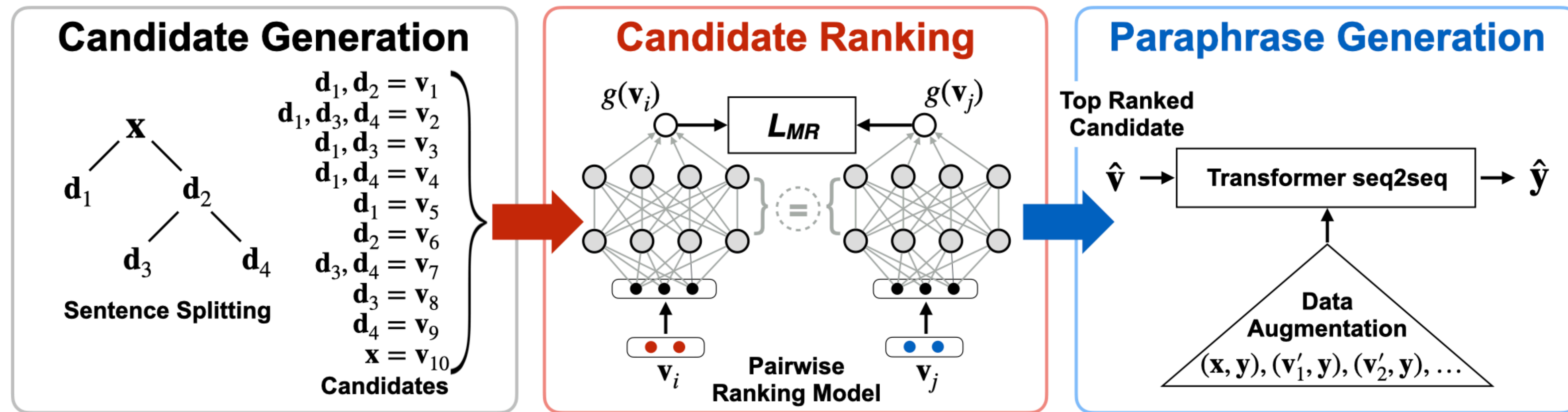
Andrew  
Duffy

+ 4 more new students joining this semester



# Controllable Text Generation

We incorporate linguistic knowledge into neural sequence-to-sequence models to improve controllability.



**Input sentence:**

Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship

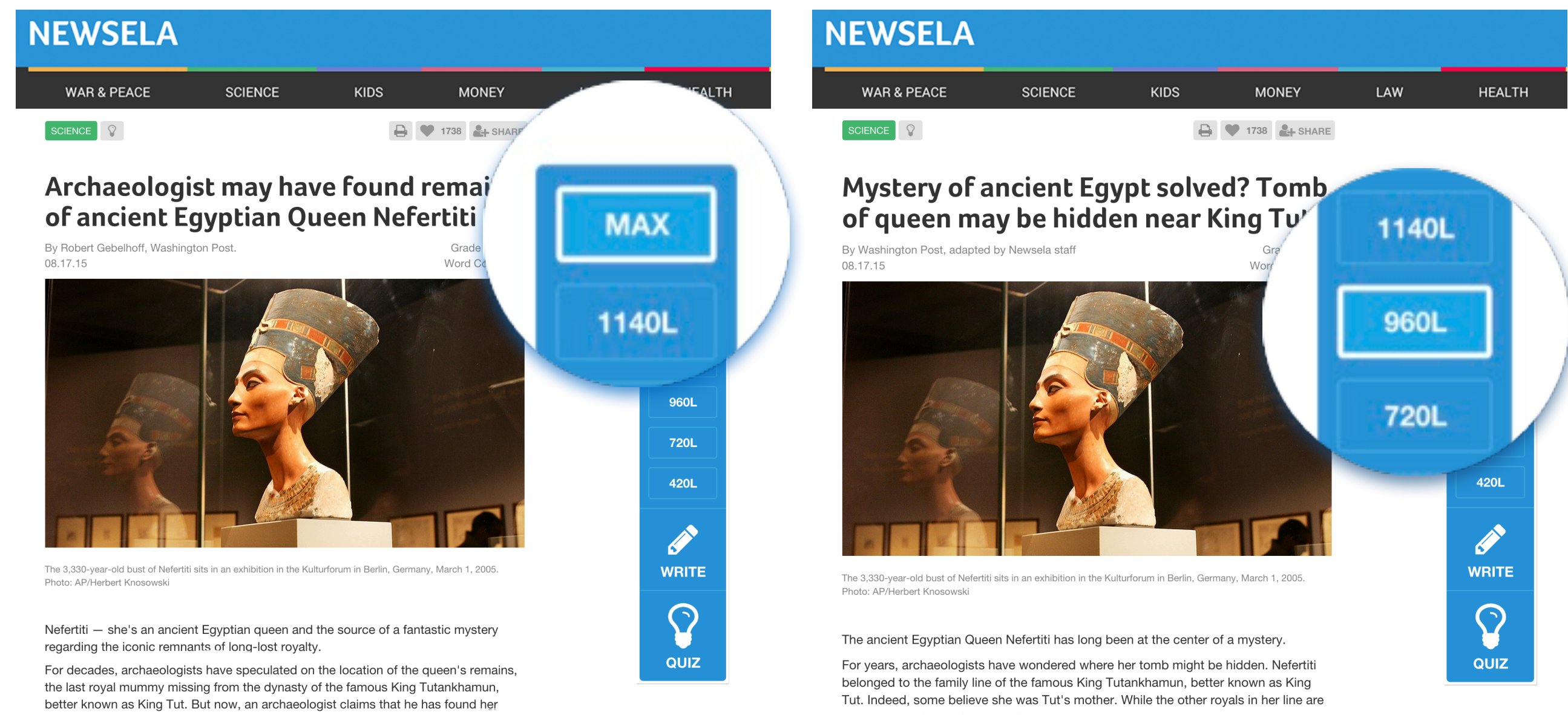
**seq2seq models**  
(RNN, Transformer)

**Generated Output:**

Scientists have found documents in Portugal.  
They have also found out who owned the ship.

# NLP for Accessibility and Education

Automatically rewrite texts into different readability levels for school students, non-native speakers, etc.



Input sentece:

Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship



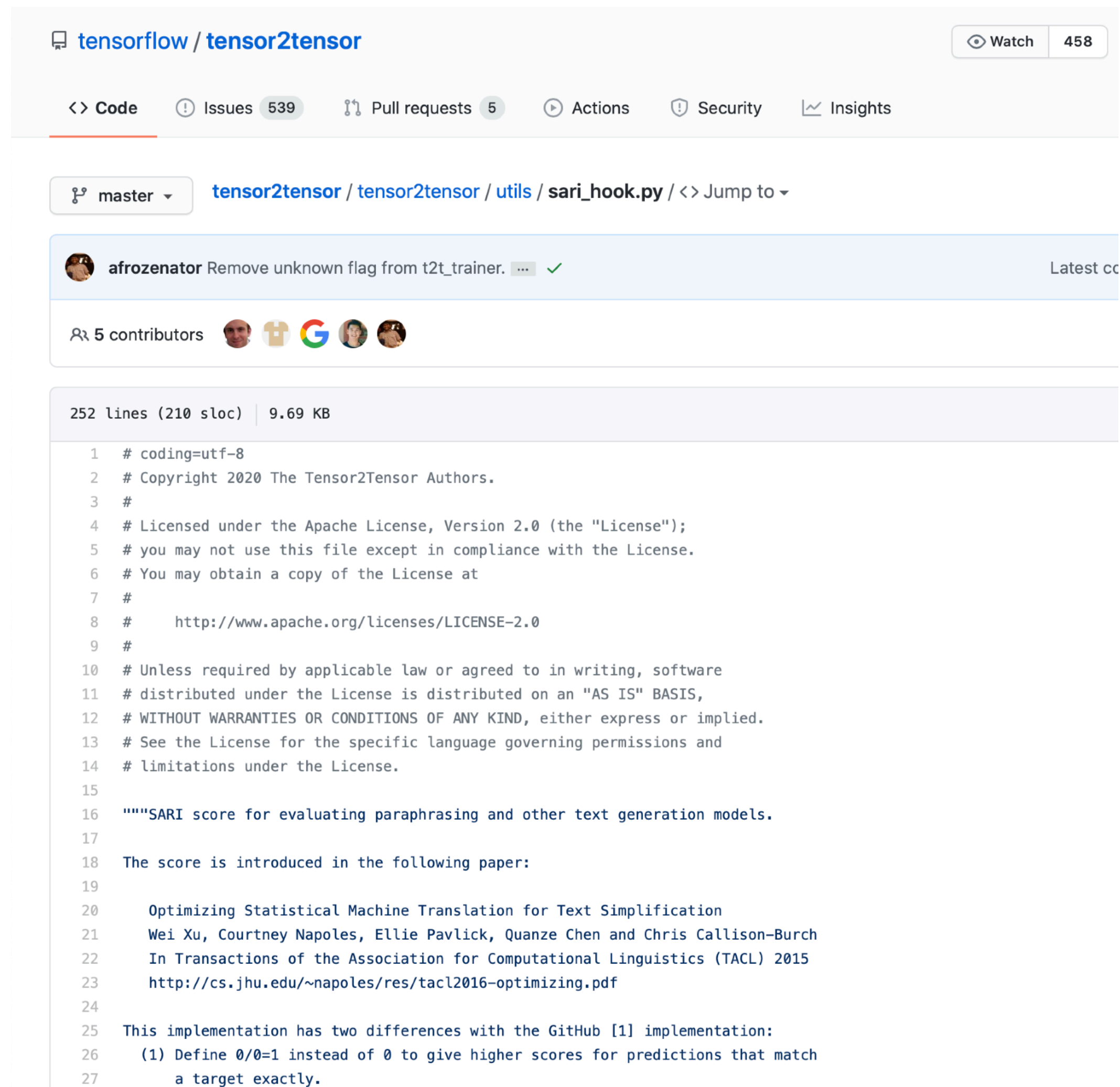
Generated Output:

Scientists have found documents in Portugal.  
They have also found out who owned the ship.



# Evaluation of Machine Learning Models

We design benchmark datasets and evaluation metrics for training and testing machine learning models for various NLP tasks.



The screenshot shows the GitHub interface for the `tensorflow/tensor2tensor` repository. The file `sari_hook.py` is selected, showing its commit history and contributors. The file content includes a license header and a description of the SARI metric.

```
1  # coding=utf-8
2  # Copyright 2020 The Tensor2Tensor Authors.
3  #
4  # Licensed under the Apache License, Version 2.0 (the "License");
5  # you may not use this file except in compliance with the License.
6  # You may obtain a copy of the License at
7  #
8  #     http://www.apache.org/licenses/LICENSE-2.0
9  #
10 # Unless required by applicable law or agreed to in writing, software
11 # distributed under the License is distributed on an "AS IS" BASIS,
12 # WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 # See the License for the specific language governing permissions and
14 # limitations under the License.
15
16 """SARI score for evaluating paraphrasing and other text generation models.
17
18 The score is introduced in the following paper:
19
20     Optimizing Statistical Machine Translation for Text Simplification
21     Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen and Chris Callison-Burch
22     In Transactions of the Association for Computational Linguistics (TACL) 2015
23     http://cs.jhu.edu/~napoles/res/tacl2016-optimizing.pdf
24
25 This implementation has two differences with the GitHub [1] implementation:
26 (1) Define 0/0=1 instead of 0 to give higher scores for predictions that match
27     a target exactly.
```

The SARI metric we designed  
is added to TensorFlow  
by Google AI group in Feb 2019.



# Robust NLP for Social Media Data

## StackOverflow NER



(ACL 2020)

## GigaBERT / Zero-shot Transfer Learning



(EMNLP 2020)

## Hashtag Segmentation



(ACL 2019)

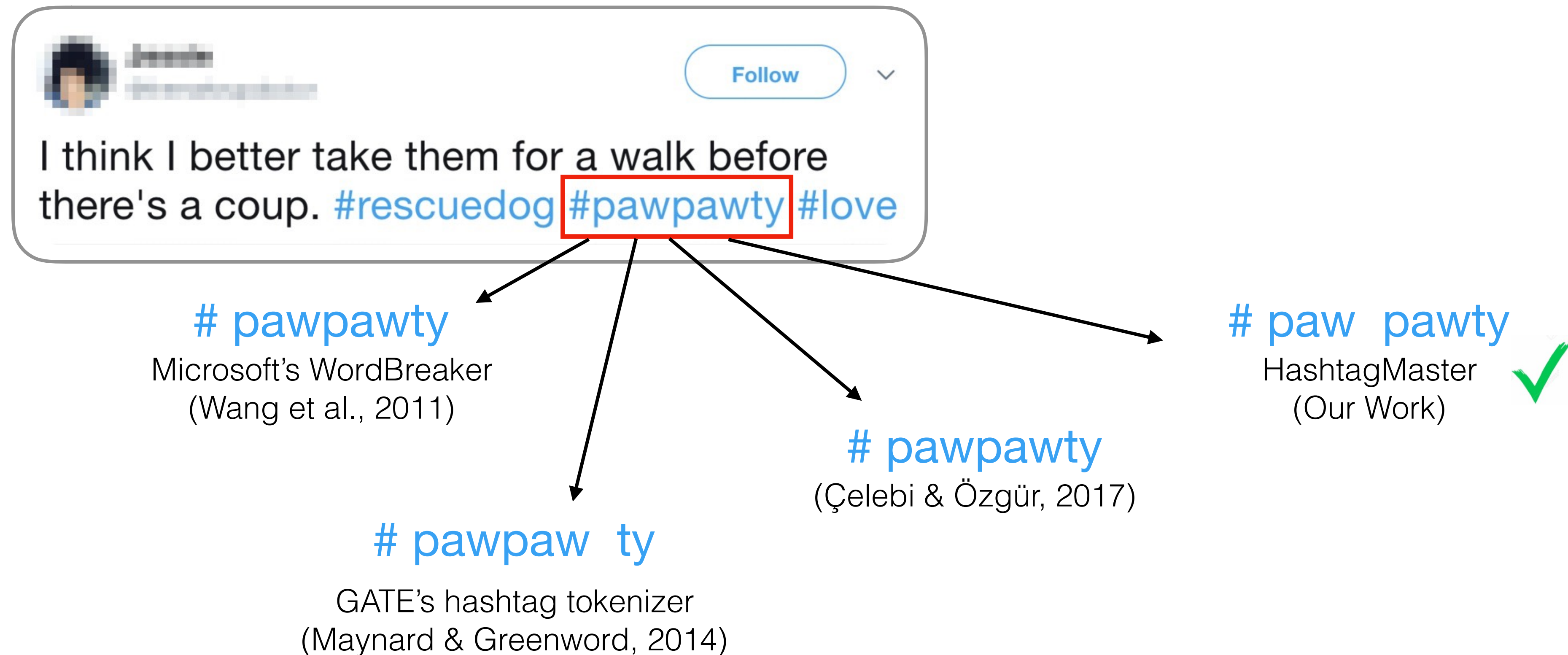
## COVID-19 Event Extraction



(new work)



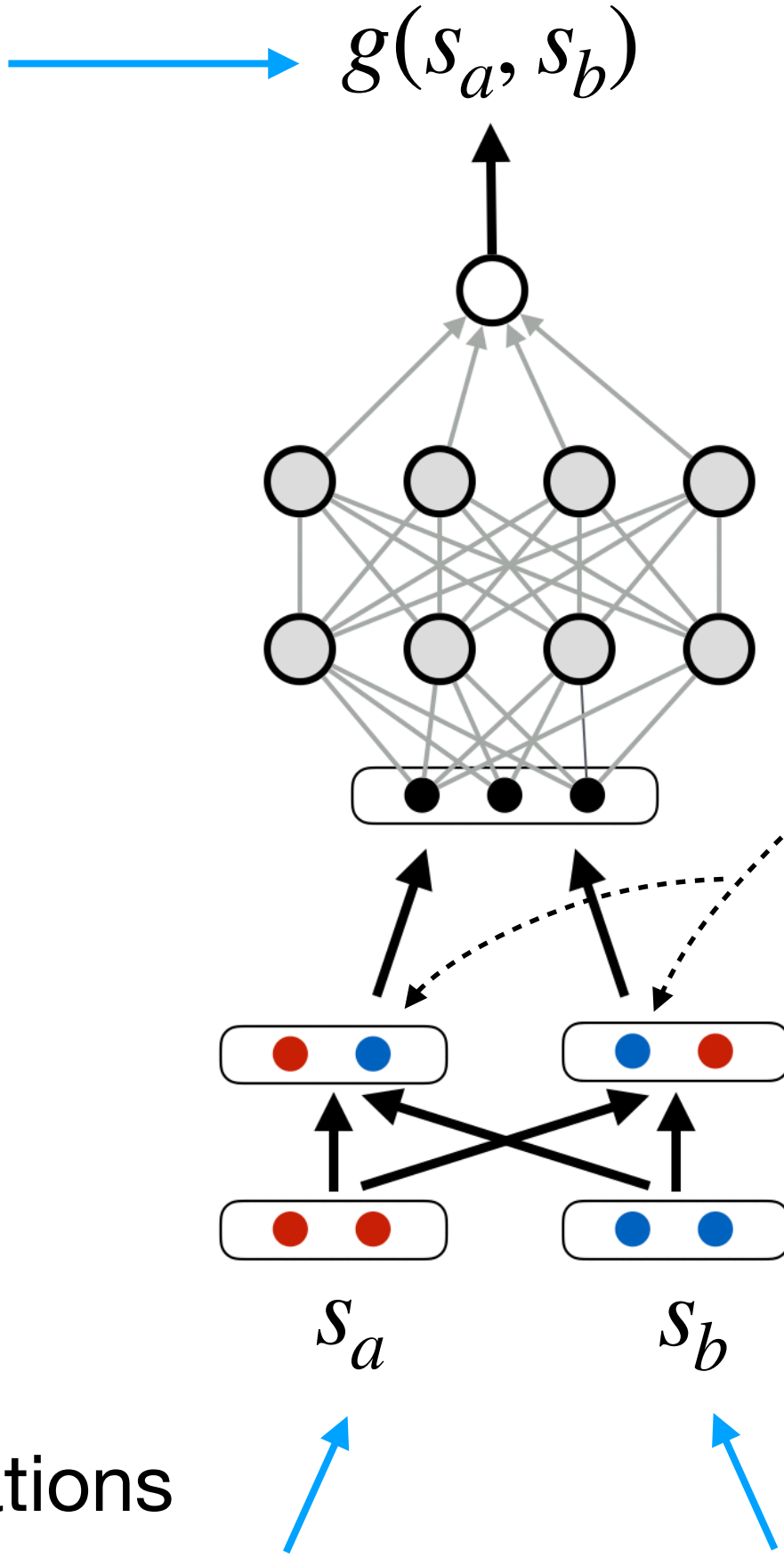
# Robust NLP for Social Media Data



collaboration with **Bloomberg**

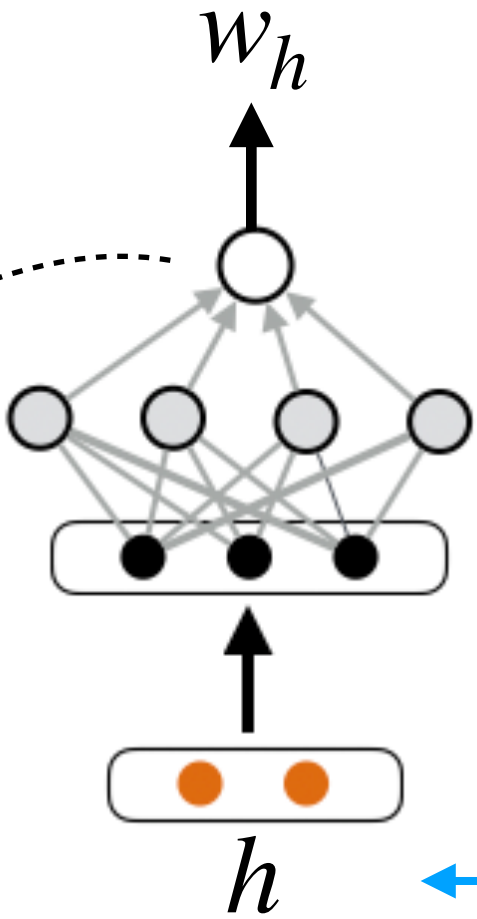
# Neural Ranking Models

$g(s_a, s_b) > 0$  means  $s_a$  is better



multi-task learning objective

$$L_{multitask} = \lambda_1 L_{MSE} + \lambda_2 L_{BCE}$$



input hashtag

$\#songsongaddafisitunes$



a pair of candidate segmentations

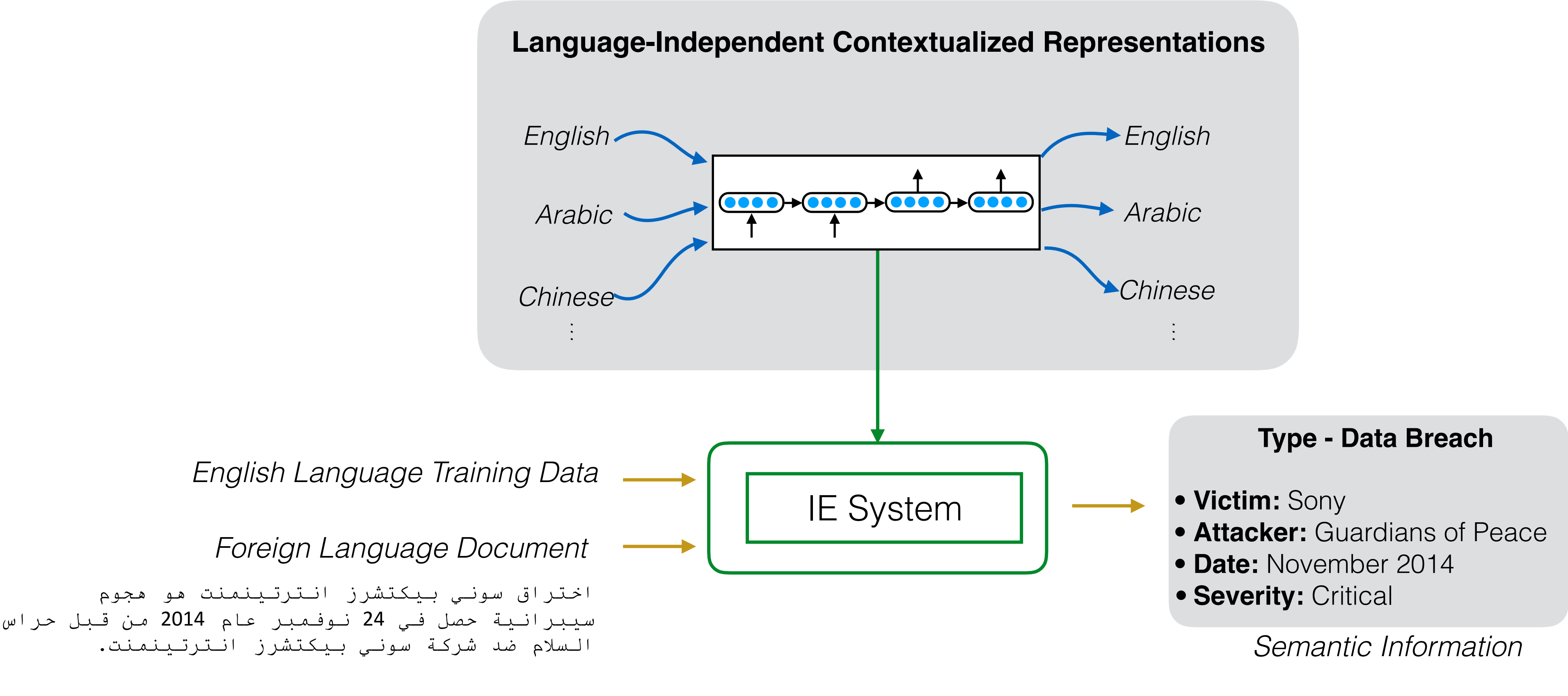
$\# songs on gaddafis itunes$

$\# song son gaddafi s itunes$

collaboration with **Bloomberg**

# Multilingual NLP

Information extraction models, trained on annotated English data, directly apply to Arabic texts to extract entities and events.



collaboration with University of Pennsylvania & Brown University



# 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 three assignments should teach you what you need to know to understand nearly any system in the literature

# Free Textbooks!

---

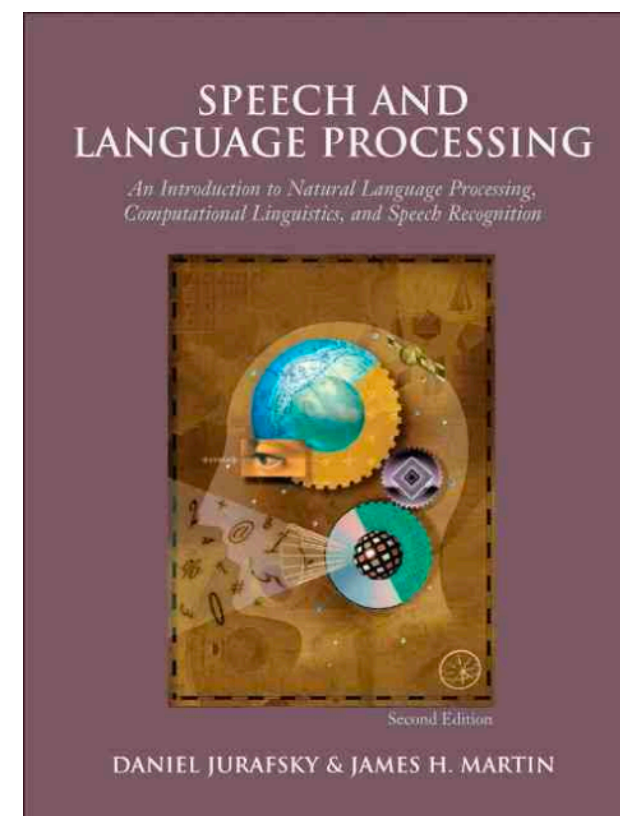
- ▶ Two really awesome textbooks available
  - ▶ There will be assigned readings from both
  - ▶ Both freely available online

## Speech and Language Processing (3rd ed. draft)

[Dan Jurafsky](#) and [James H. Martin](#)



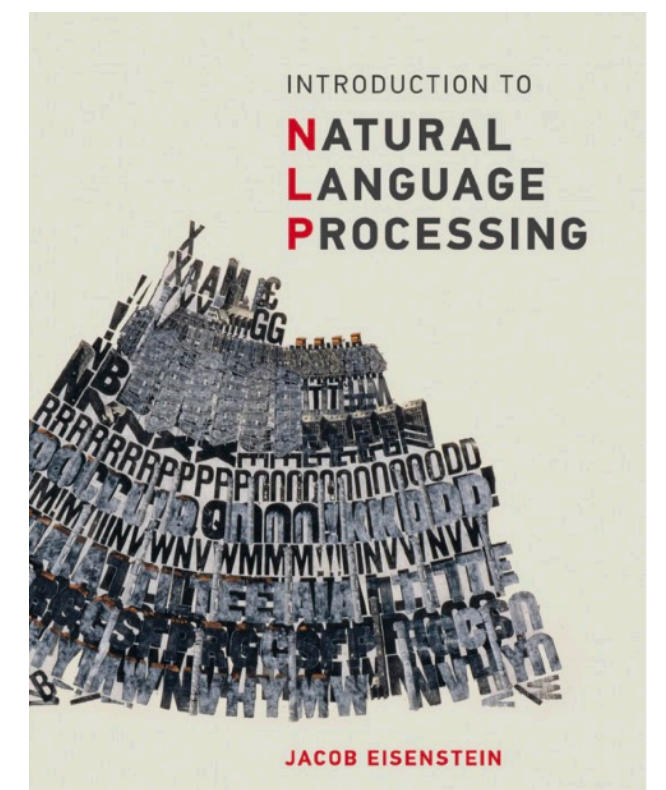
Here's our December 30, 2020 draft! Includes:



## Introduction to Natural Language Processing

By [Jacob Eisenstein](#)

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





# Course Requirements

---

- ▶ Probability (e.g. conditional probabilities, conditional independence, Bayes Rule)
- ▶ Linear Algebra (e.g., multiplying vectors and matrices, matrix inversion)
- ▶ Multivariable Calculus (e.g., calculating gradients of functions with several variables)
- ▶ Programming / Python experience (medium-to-large scale project, debug when there are no without error messages)
- ▶ Prior exposure to machine learning will be very helpful

There will be a lot of math and programming!

# Background Preparation

---

- ▶ Problem Set 0 (math background) is released, due Thursday Jan 13.
- ▶ Project 0 (logistic regression) is also released, due Thursday Jan 20.
- ▶ Recommend taking CS 4641 and (Math 2550 or Math 2551 or Math 2561 or Math 2401 or Math 24X1 or 2X51) before this class.

There will be a lot of math and programming!

Homework release, slides, readings, and course policies are on the course website  
[https://cocoxu.github.io/CS4650\\_spring2022/](https://cocoxu.github.io/CS4650_spring2022/)

# Wait List

---

- ▶ If you plan to take the class, please complete and submit Problem Set 0 by Thursday Jan 13.
- ▶ If you get off the wait list, you will be automatically added to Gradescope after about a day.
- ▶ You can also post a message on Piazza (<https://piazza.com/gatech/spring2022/cs4650a>) to get the access code to Gradescope.
- ▶ If you cannot access Gradescope by the due date, please email your submission to the instructor.

# Coursework Plan

---

- ▶ Four programming projects (40%; fairly substantial implementation effort)
  - ▶ 0. Logistic regression
  - ▶ 1. Text classification
  - ▶ 2. Sequential tagging (BiLSTM-CNN-CRF)
  - ▶ 3. Neural chatbot (Seq2Seq with attention)
- ▶ Three written assignments (20%) + midterm exam (15%)
  - ▶ Mostly math problems related to ML / NLP
- ▶ Final project (20%; details on course website, will discuss later)

# Programming Projects

---

- ▶ Four Programming Assignments (40% grade)
  - ▶ Implementation-oriented
  - ▶ 1.5~2 weeks per assignment
  - ▶ 3 “slip days” for automatic extensions (for emergency situations)

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!**



# Programming Projects

---

- ▶ Modern NLP methods require non-trivial computation
- ▶ Training neural networks with many parameters can take a long time (it is a very good idea to start working on the assignments early!)
- ▶ Most programming will be done with PyTorch library (can be tricky to debug)
- ▶ You will want to use a GPU (Google Colab; pro account for \$10/month)
- ▶ The programming projects are designed with Google Colab in mind



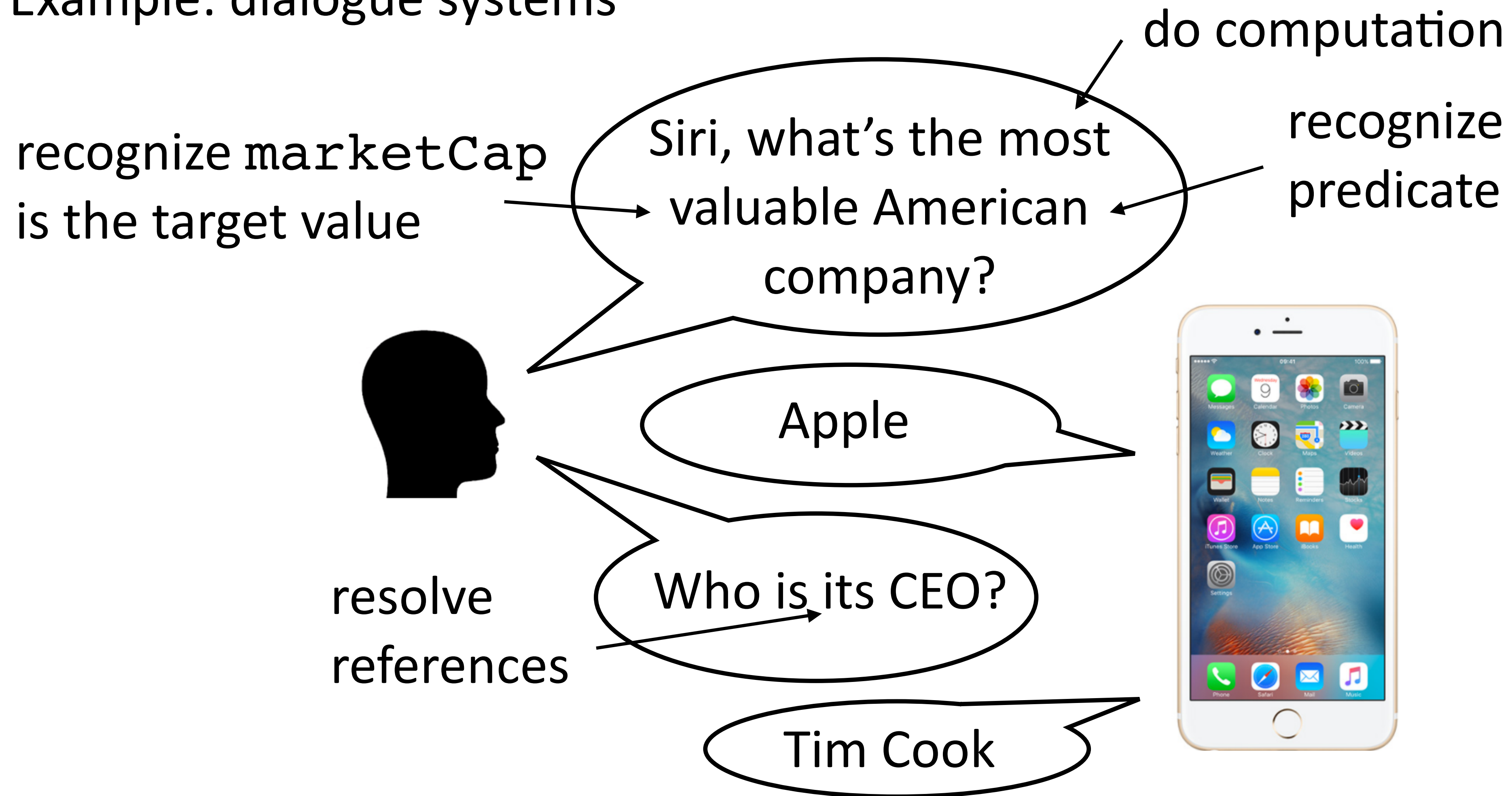
# Final Project

---

- ▶ Final project (20% grade)
  - ▶ Groups of 2-4 preferred, 1 is possible with permission.
  - ▶ Good idea to talk to run your project idea by me in office hours or email.
  - ▶ 4 page report + final project presentation.
- ▶ More details on course website (will also discuss later mid-semester)

# 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

Translate

English French Spanish Chinese - detected

特朗普偕家人在白宫阳台观看百年一遇日全食

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



# Machine Translation



特朗普偕家人在白宫阳台观看百年一遇日全食

People's Daily, August 30, 2017

Translate

English

French

Spanish

Chinese - detected



特朗普偕家人在白宫阳台观看百年一遇日全食

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

# Machine Translation



特朗普偕家人在白宫阳台观看百年一遇日全食

People's Daily, August 30, 2017

Translate

English French Spanish Chinese - detected

特朗普偕家人在白宫阳台观看百年一遇日全食

Trump and his family watch the 100-year total solar eclipse from the balcony of the White House



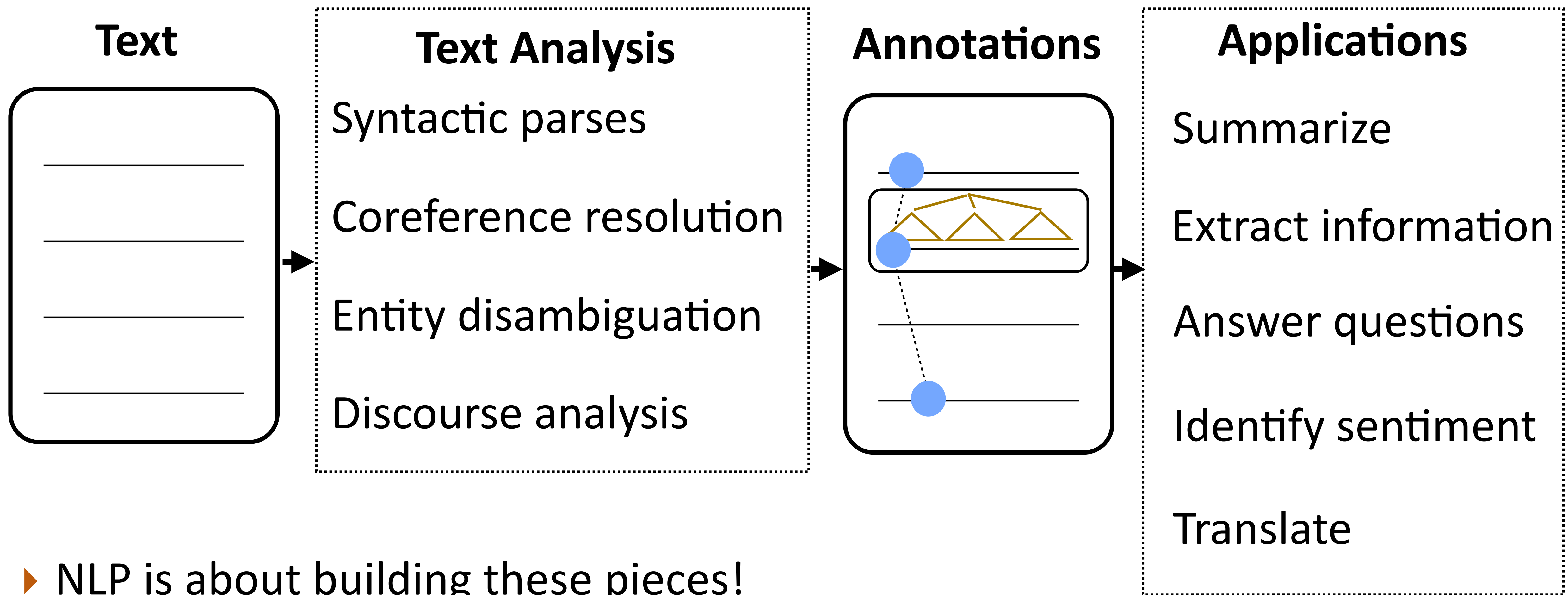
# Textual Entailment

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

# 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

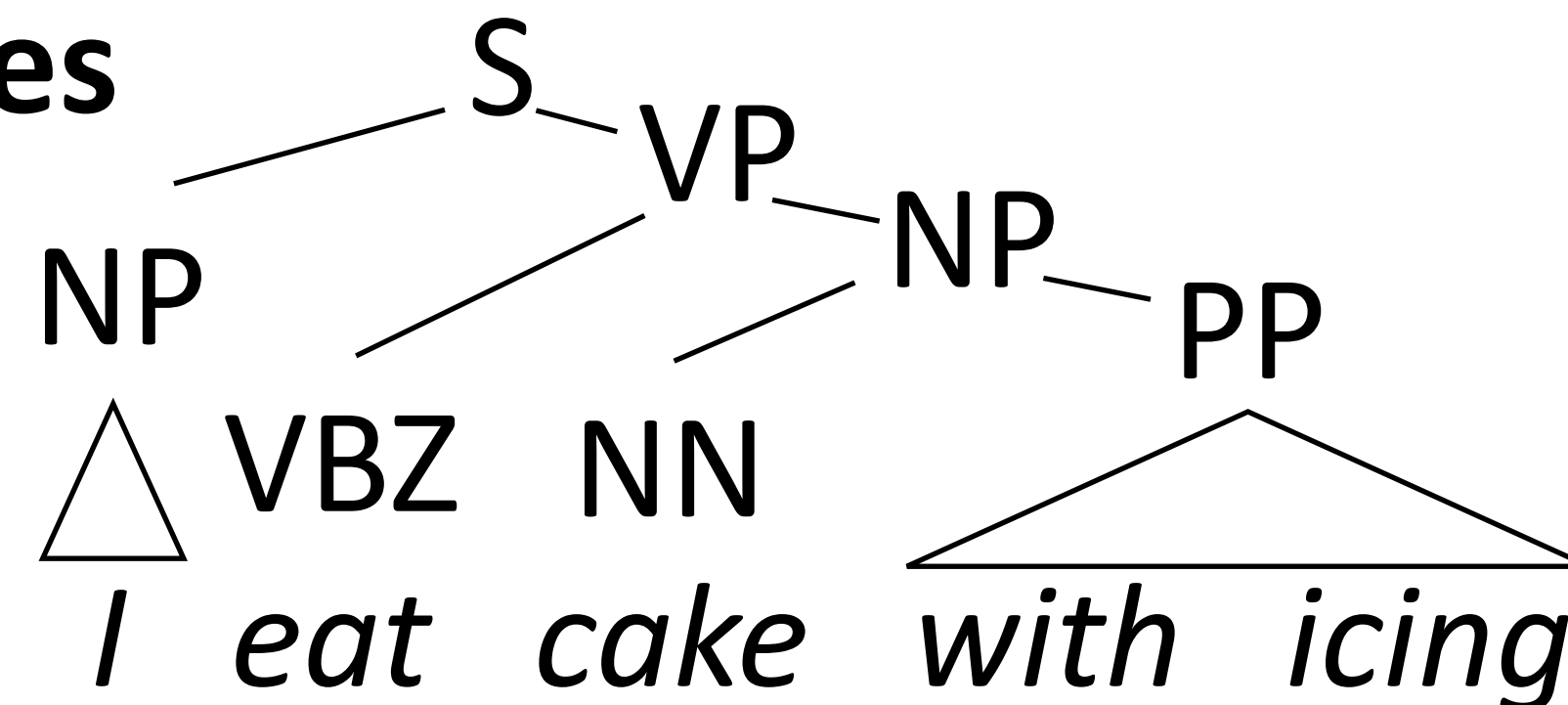
#### PERSON

*Tom Cruise* stars in the new

#### MOVIE

*Mission Impossible* film

### Trees

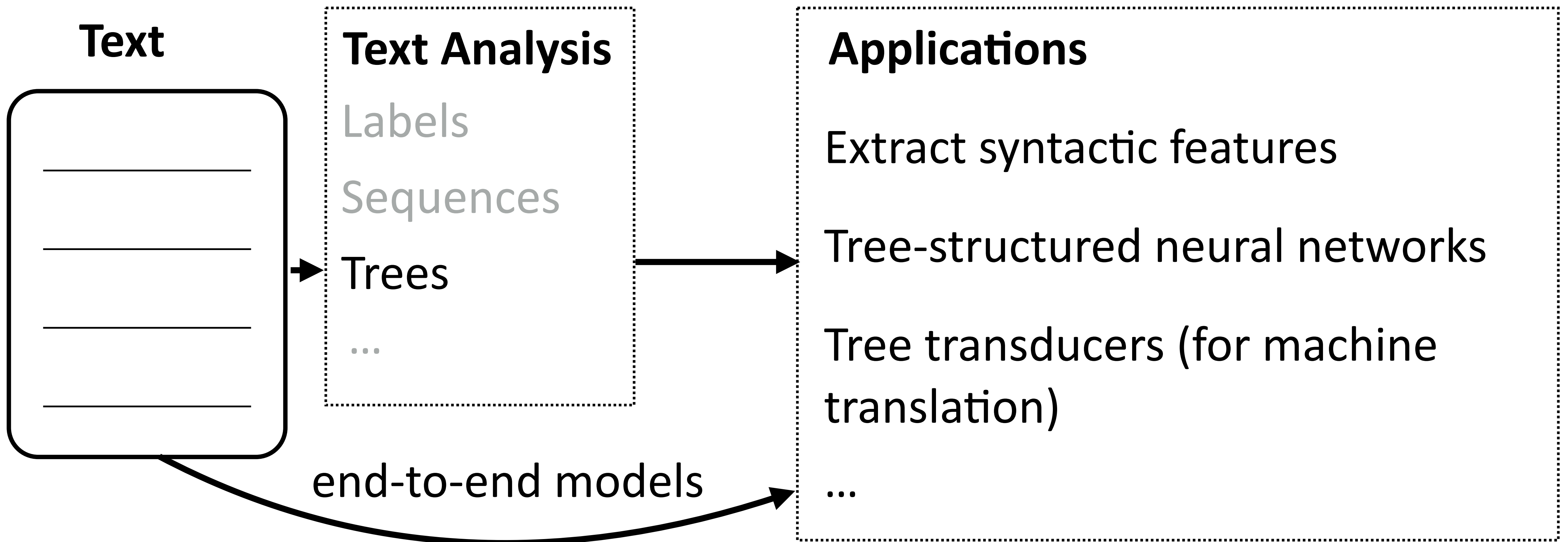


$\lambda x. \text{flight}(x) \wedge \text{dest}(x)=\text{Miami}$

*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 and compositionality 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 1968-1972)

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!

---

- ▶ Ambiguous News 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



# 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?

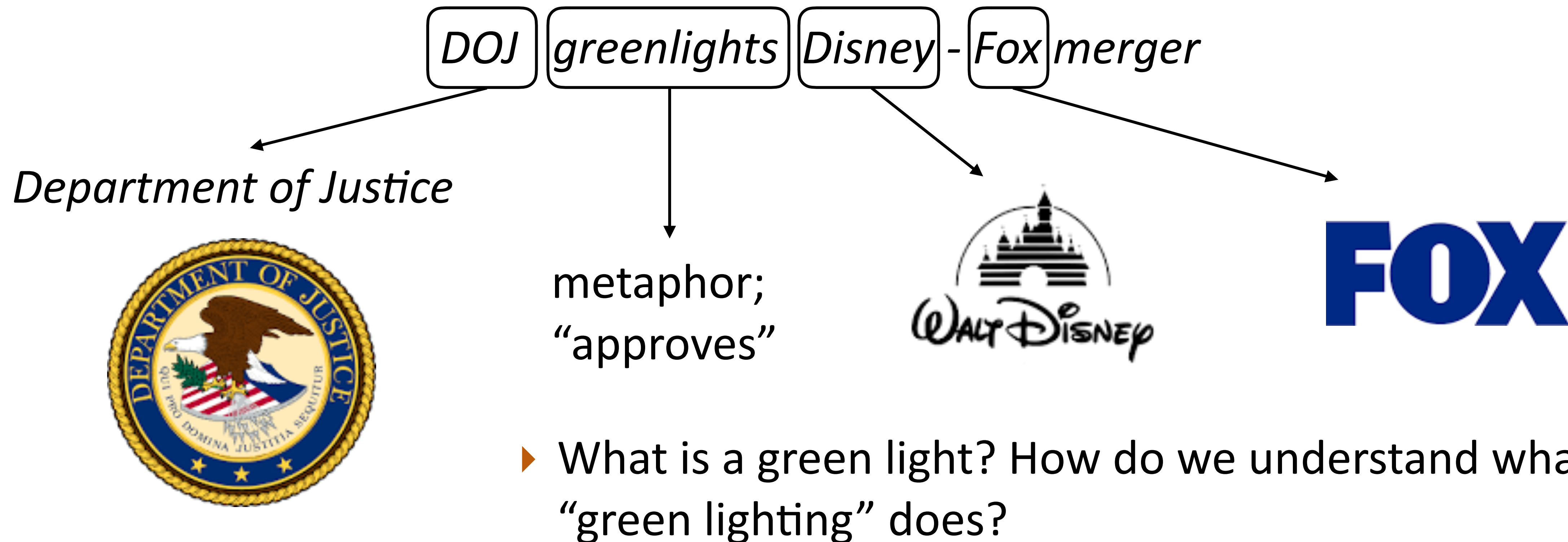
---

## ► 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] [.]

# What do we need to understand language?

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

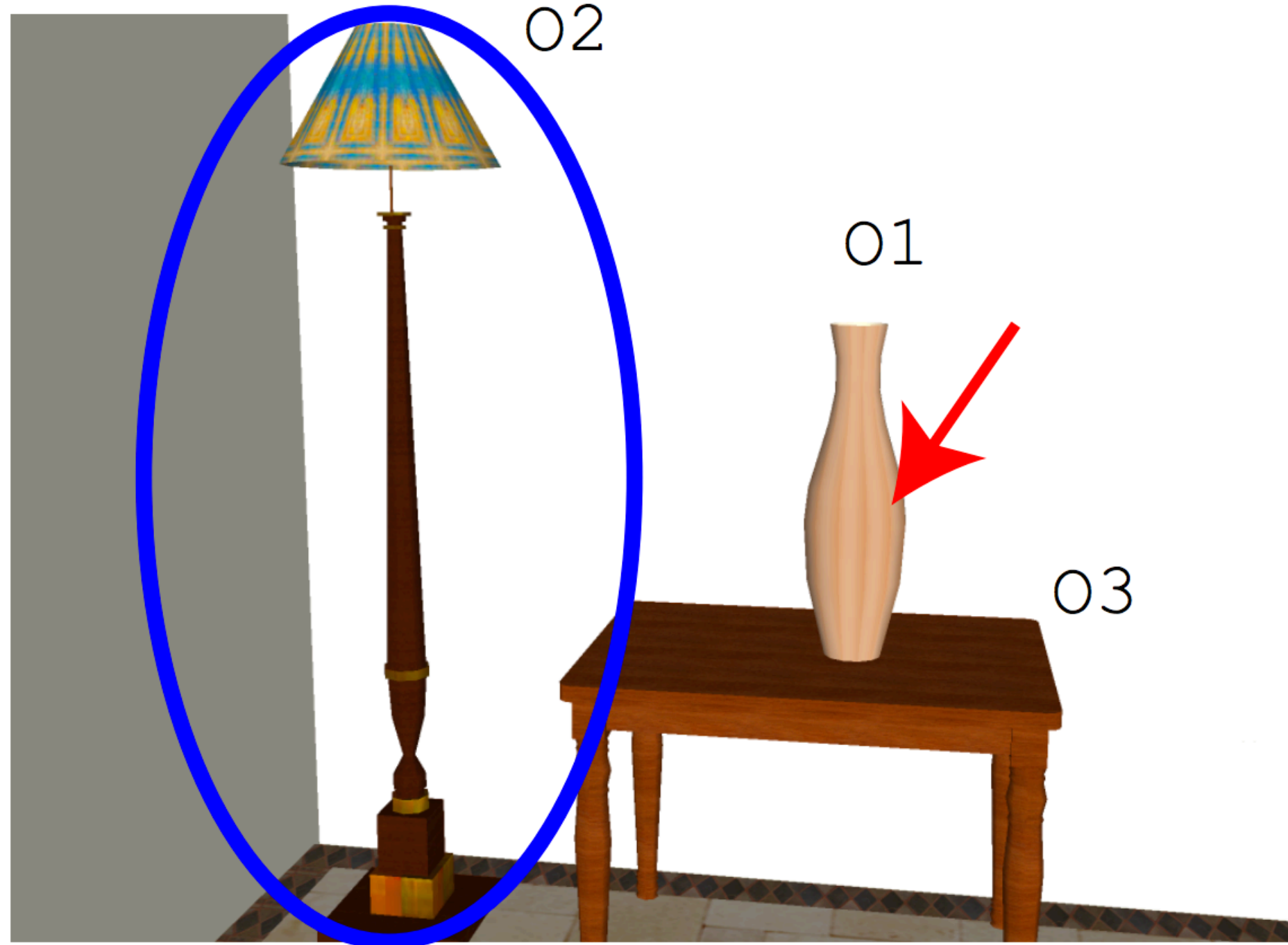




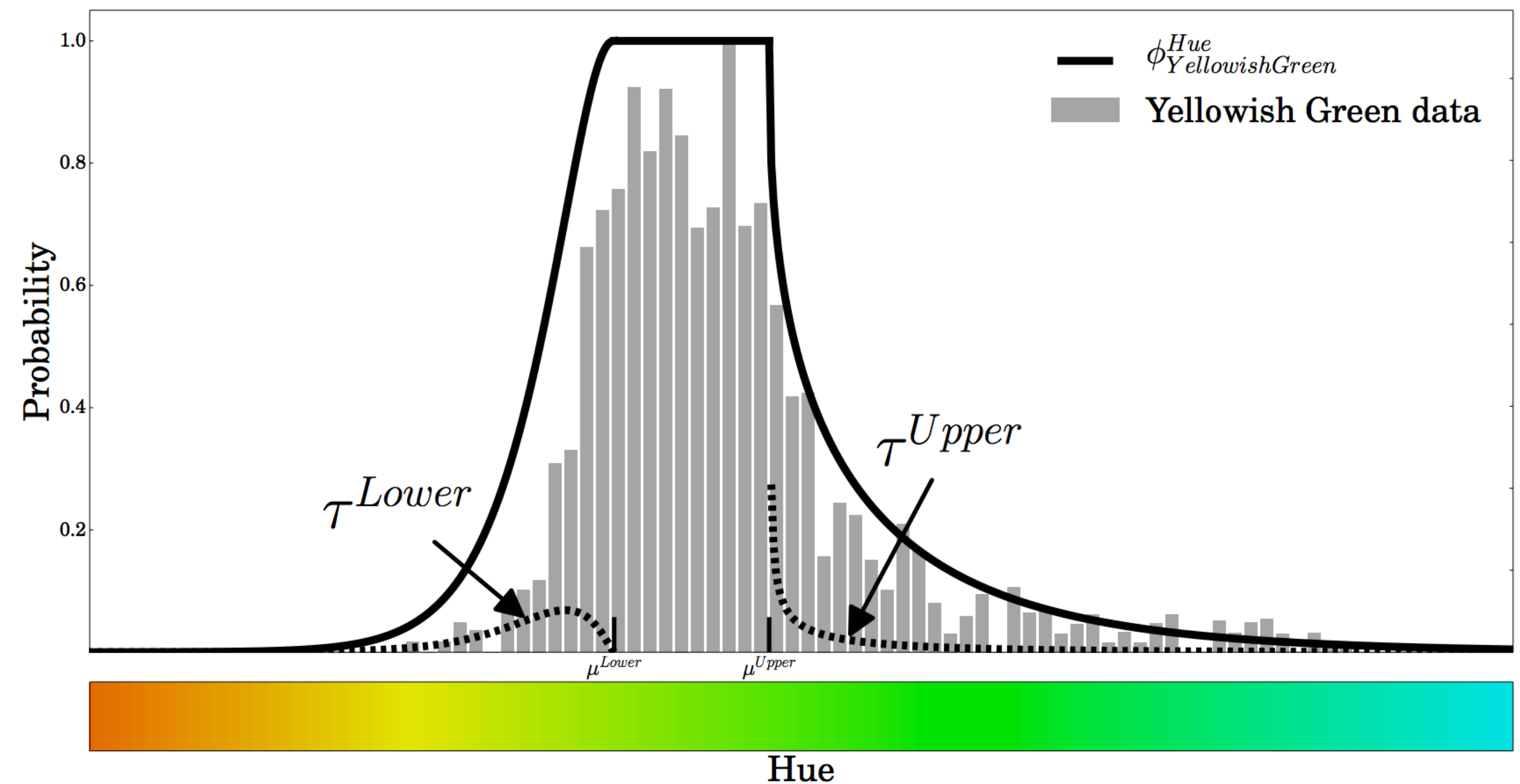
# What do we need to understand language?

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

Question: What object is right of O2 ?



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)

backward   $C_b = \text{John}; C_f = \{\text{John}\}$   forward center

center

c. He called up Mike yesterday to work out a plan. (he = John)

$C_b = \text{John}; C_f = \{\text{John}, \text{Mike}\}$  (CONTINUE)

d. Mike has annoyed him a lot recently.

$C_b = \text{John}; C_f = \{\text{Mike}, \text{John}\}$  (RETAIN)

e. He called John at 5 AM on Friday last week. (he = Mike)

$C_b = \text{Mike}; C_f = \{\text{Mike}, \text{John}\}$  (SHIFT)

What techniques do we use?  
(to combine data, knowledge, linguistics, etc.)



# A brief history of (modern) NLP

“AI winter”  
rule-based,  
expert systems



Penn  
treebank  
S  
NP VP

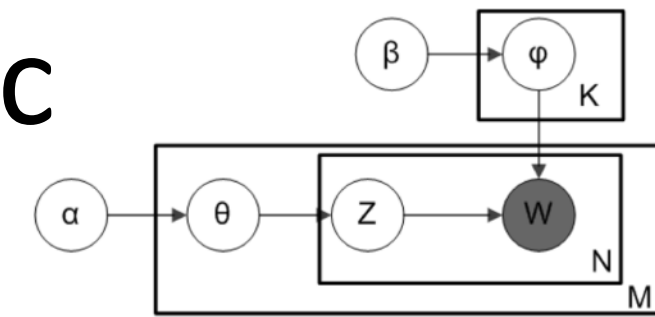
earliest stat MT  
work at IBM



Collins vs.  
Charniak  
parsers

Ratnaparkhi  
tagger  
NNP VBZ

Unsup: topic  
models,  
grammar induction



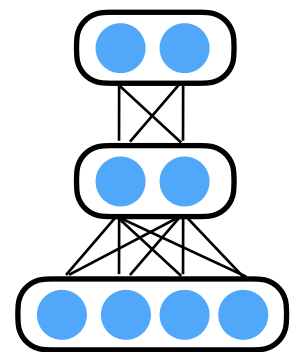
Pretraining



Sup: SVMs,  
CRFs, NER,  
Sentiment

Semi-sup,  
structured  
prediction

Neural



1980

1990

2000

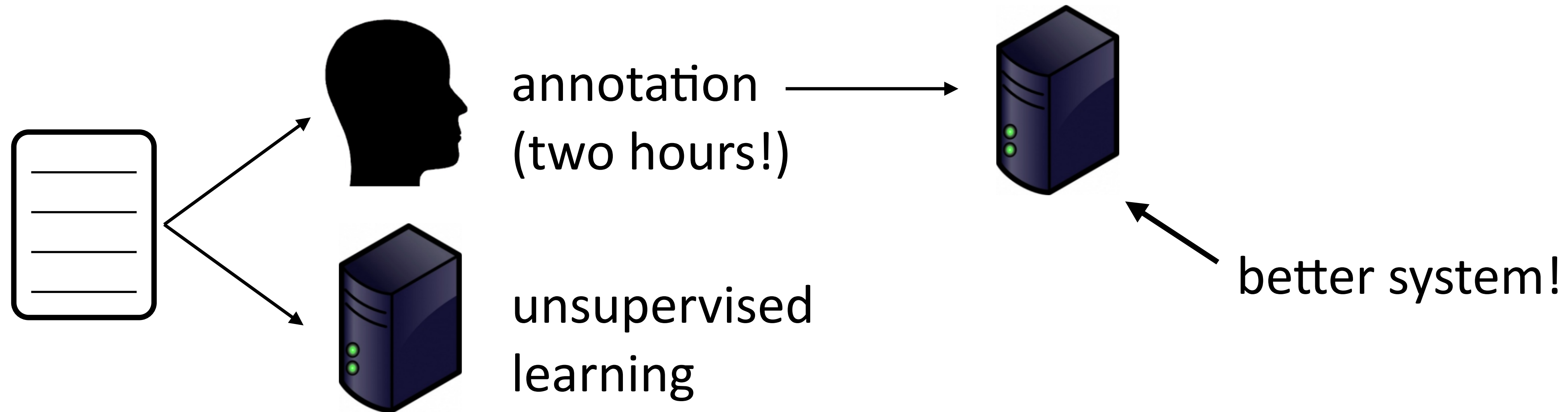
2010

2018

# Structured Prediction

---

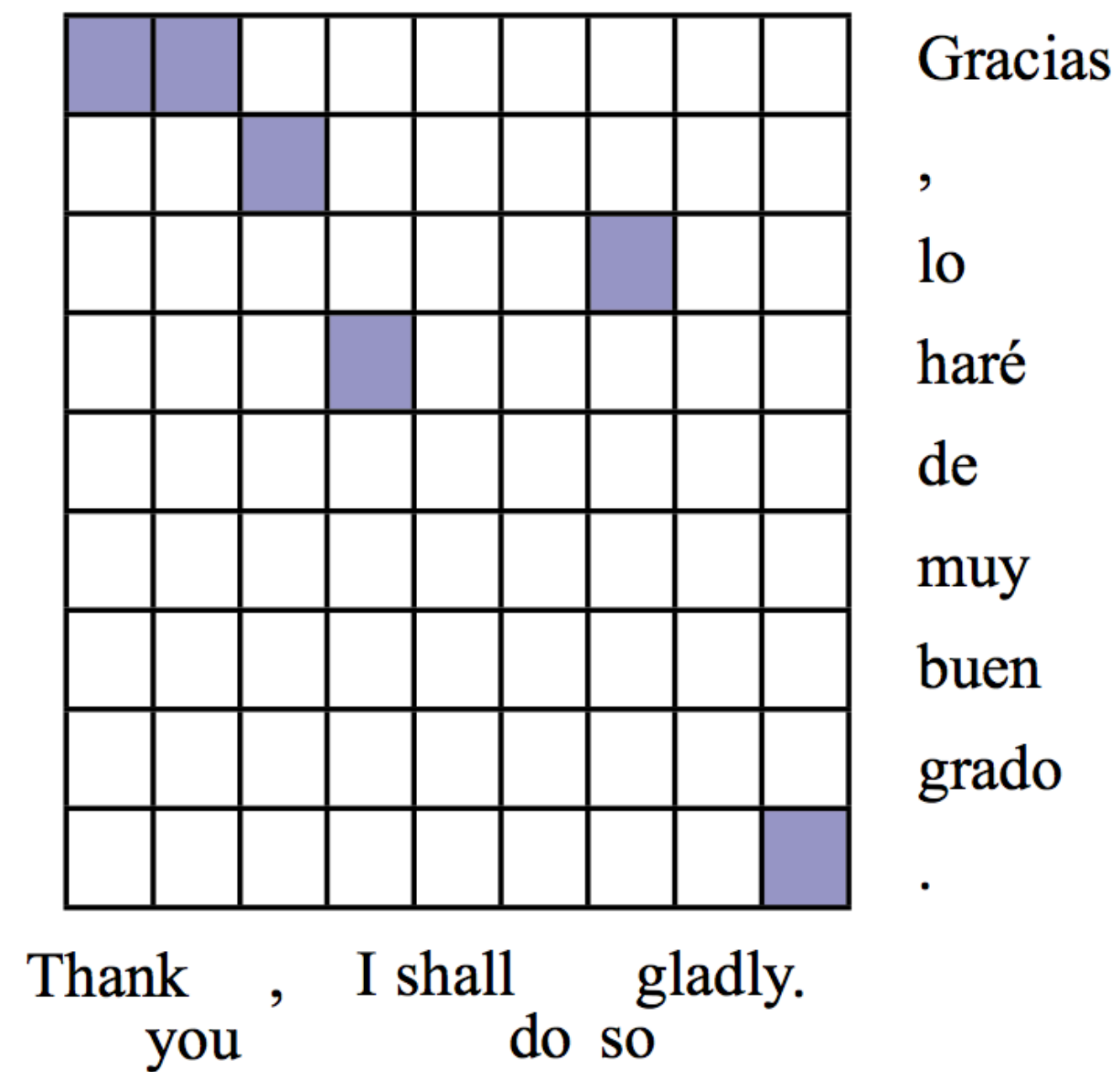
- ▶ All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- ▶ Supervised techniques work well on very little data



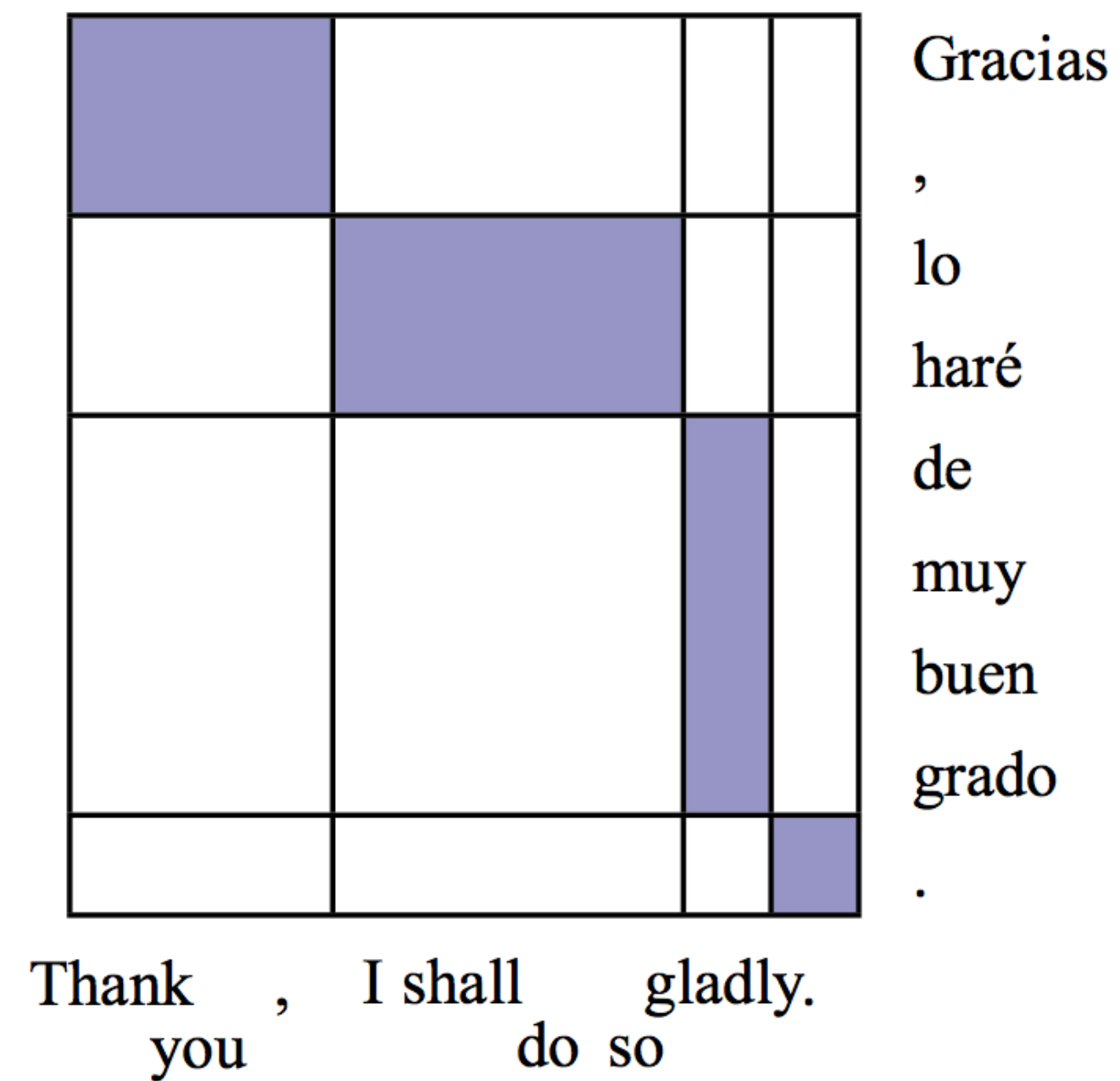
- ▶ Even neural nets can do pretty well!

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

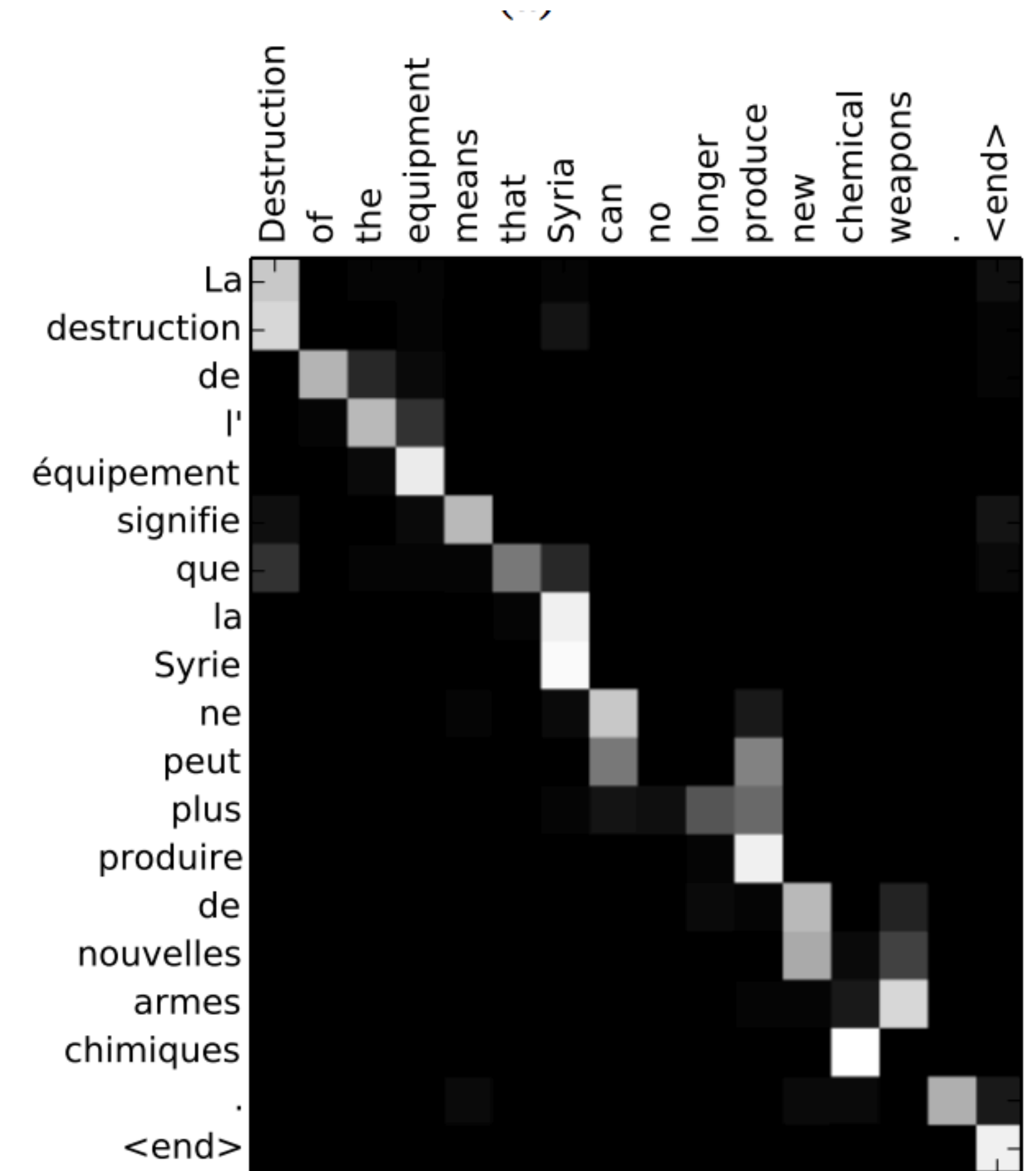
# Less Manual Structure?



(a) example word alignment



(b) example phrase alignment





# 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

CoNLL	
	Avg. F <sub>1</sub>
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 <sup>-</sup>	51.01

Moosavi and Strube (2017)

# Does manual structure have a place?

## Translate

English French Spanish Chinese - detected ▼

特朗普偕家人在白宫阳台观看百年一遇日全食✕

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

- ▶ Maybe manual structure would help...

# Where are we?

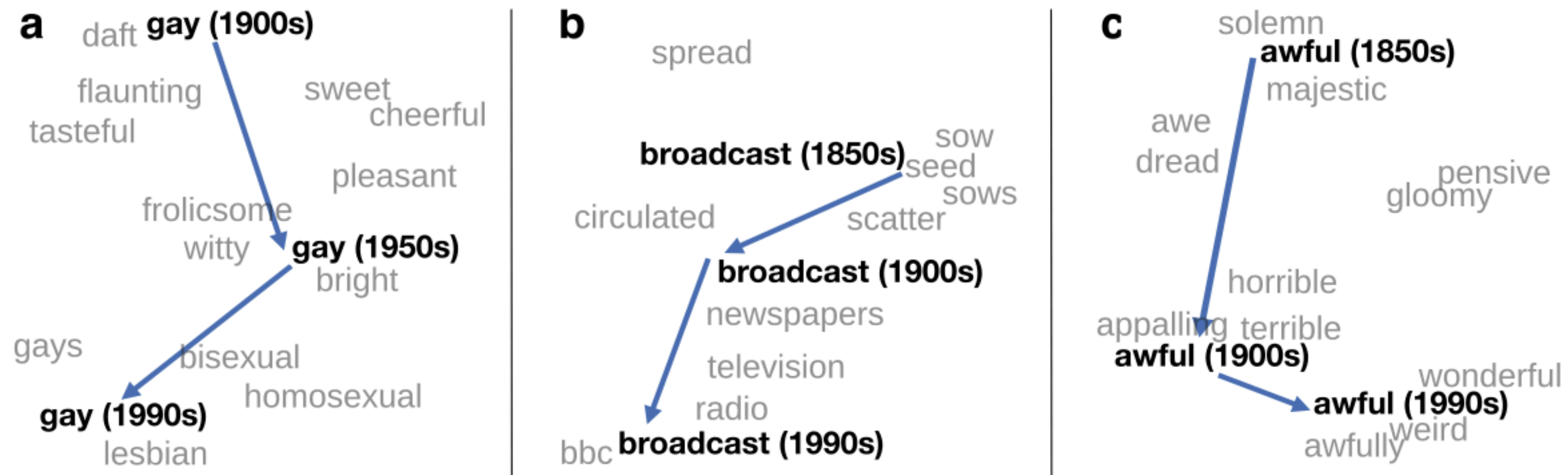
---

- ▶ 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



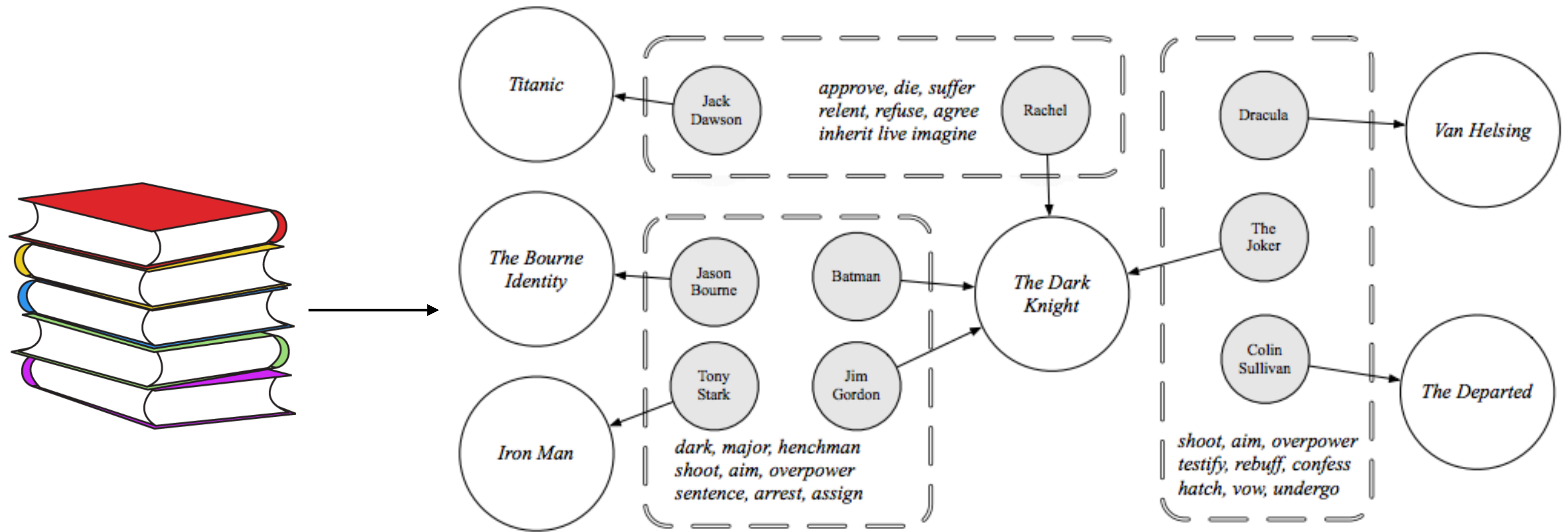
# NLP vs. Computational Linguistics

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



# NLP vs. Computational Linguistics

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



# Outline of the Course

ML and structured  
prediction for NLP

Neural Networks  
semantics

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

Date	Topic	Projects	Problem Sets
1/10/2022	Course Overview	Proj. 0 Out	PS0 Out
1/12/2022	Machine Learning		PS0 Due (1/13)
1/17/2022	No class - MLK day		
1/19/2022	Machine Learning	Proj. 0 Due	PS1 Out
1/24/2022	Machine Learning		
1/26/2022	Machine Learning	Proj. 1 Out	PS1 Due
1/31/2022	Neural Networks in NLP		
2/2/2022	PyTorch Tutorial		
2/7/2022	Neural Networks in NLP		
2/9/2022	Sequence Labeling	Proj. 1 Due	PS2 Out
2/14/2022	Conditional Random Fields		
2/16/2022	Word Embeddings	Proj. 2 Out	PS2 Due
2/21/2022	Recurrent Neural Networks		
2/23/2022	Convolutional Neural Networks		
2/28/2022	Neural CRF		
3/2/2022	Machine Translation	Proj. 2 Due	
3/7/2022	Encoder-Decoder Networks		
3/9/2022	Midterm	Proj. 3 Out	
3/14/2022	Neural Machine Translation		
3/16/2022	Pointer Network / Final Project Kickoff	Final Project Kickoff	
3/21/2022	No class - Spring Break		
3/23/2022	No class - Spring Break		
3/28/2022	Transformer Model	Proj. 3 Due	
3/30/2022	Information Extraction		
4/4/2022	Pretrained Language Models		
4/6/2022	Pretrained Language Models, Ethics		
4/11/2022	Dialogue		
4/13/2022	Question Answering		
4/18/2022	Multilingual		
4/20/2022	Guest Lecture		
4/25/2022	No class		
4/29/2022	Final project presentation	Final Project Report Due	

tentative plan  
(subject to change)



# Questions?

---

Piazza — <https://piazza.com/class/kxz22ppzou35x0>