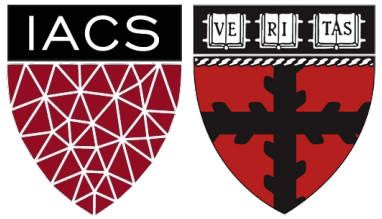
Lecture 1: Introduction

Course Overview + What is NLP?

Harvard

AC295/CS287r/CSCI E-115B

Chris Tanner



"All I need is one mic [and a team of TFs/CAs, an A/V support staff, teaching resources like Canvas and EdStem, and 55 students who are engaged, curious, and motivated]"

-- Nasir Jones, *One Mic*. April 16, 2002

ANNOUNCEMENTS

- **Attendance** is checked today. See a TF before you leave today.
- **EdStem** (linked from Canvas) – introduce yourself
- **PyTorch tutorial** will be Tues evening (more details will be posted on Ed)
- **Homework 1** will be released Tues
- **Lectures slides** will be posted on the website and our Twitter @CS287_NLP
- **Office Hours** start next Wednesday @ 5pm (see website for all OH)

Outline

What is this course?

Logistics

NLP: what, why, how?

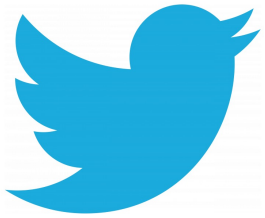
Outline

- What is this course?
- Logistics
- NLP: what, why, how?

What is this course?

Our digital world is inundated with data, most textual data

62B pages



500M tweets/day

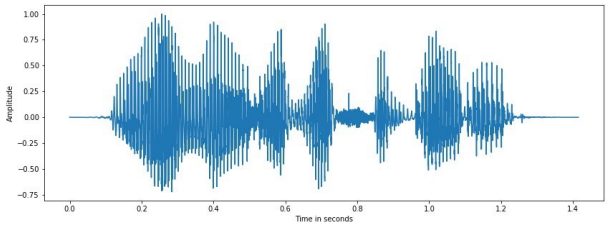
360M user pages

The New York Times

13M articles

What is this course?

Natural Language Processing (NLP) is the study of how to get computers to process, "understand", and leverage human language data.



Speech Audio

(Signal Processing work is a cousin community and often done by EE folks)

Written Text

Neural Decipherment via Minimum-Cost Flow: From Ugaritic to Linear B Luo, et al. (2019)

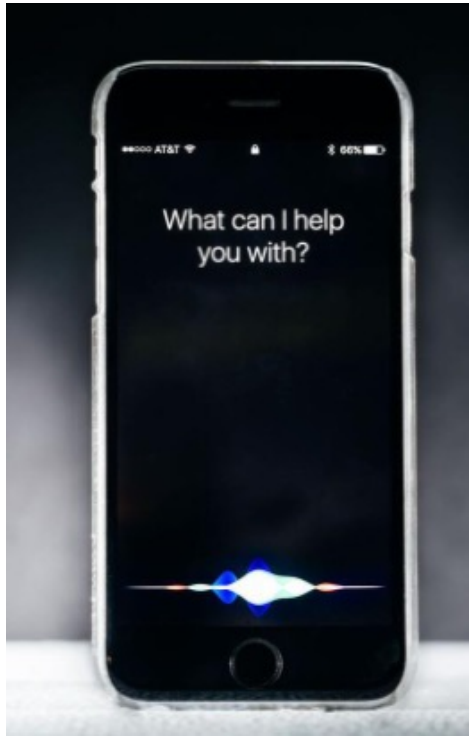
Sign Language

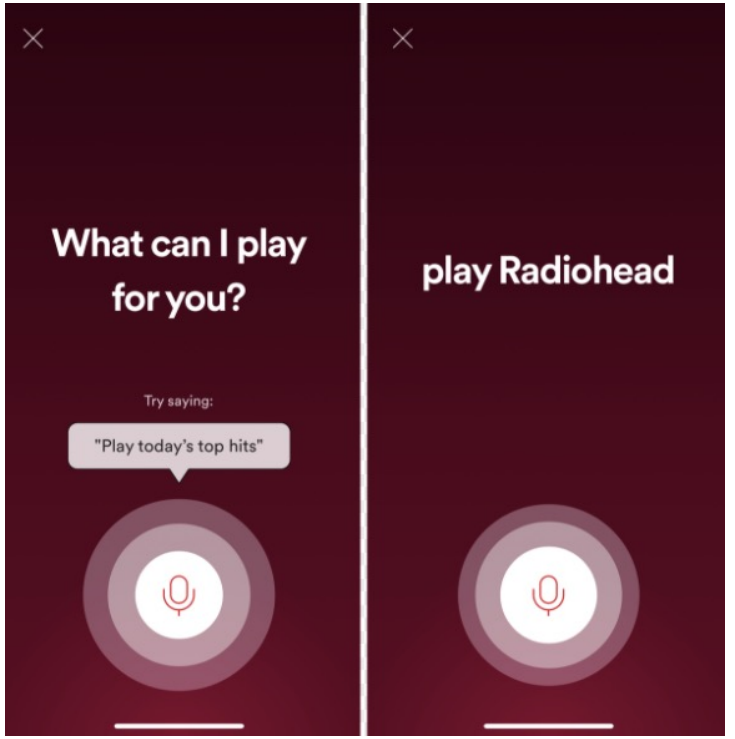
Including Signed Languages in Natural Language Processing Yin, et al. (2021)

NLP Successes

NLP has been around since the 1960s, but the progress in the past 10 years has been astronomical. Models are becoming effective enough for consumer use.

Voice Assistants

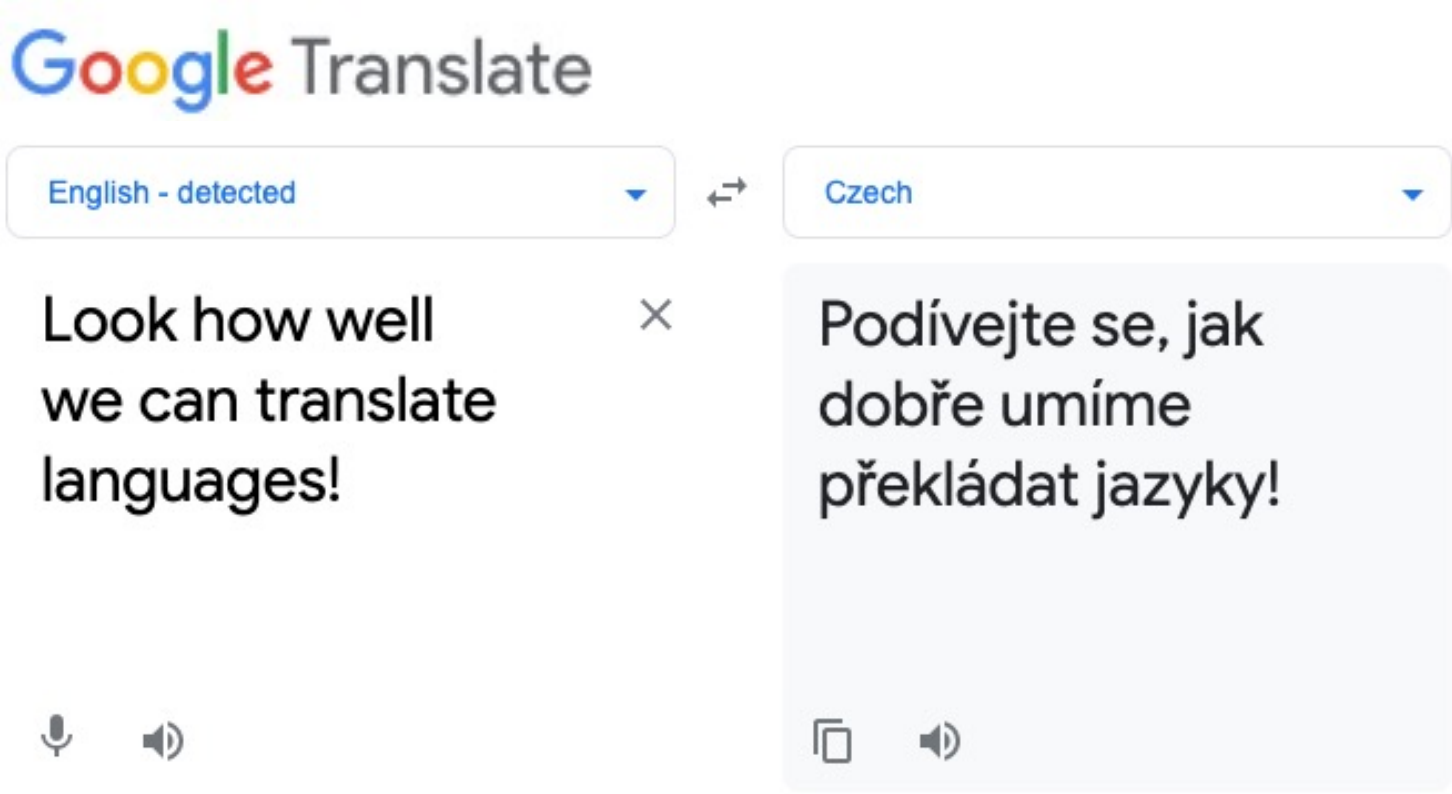




NLP Successes

NLP has been around since the 1960s, but the progress in the past 10 years has been astronomical. Models are becoming effective enough for consumer use.

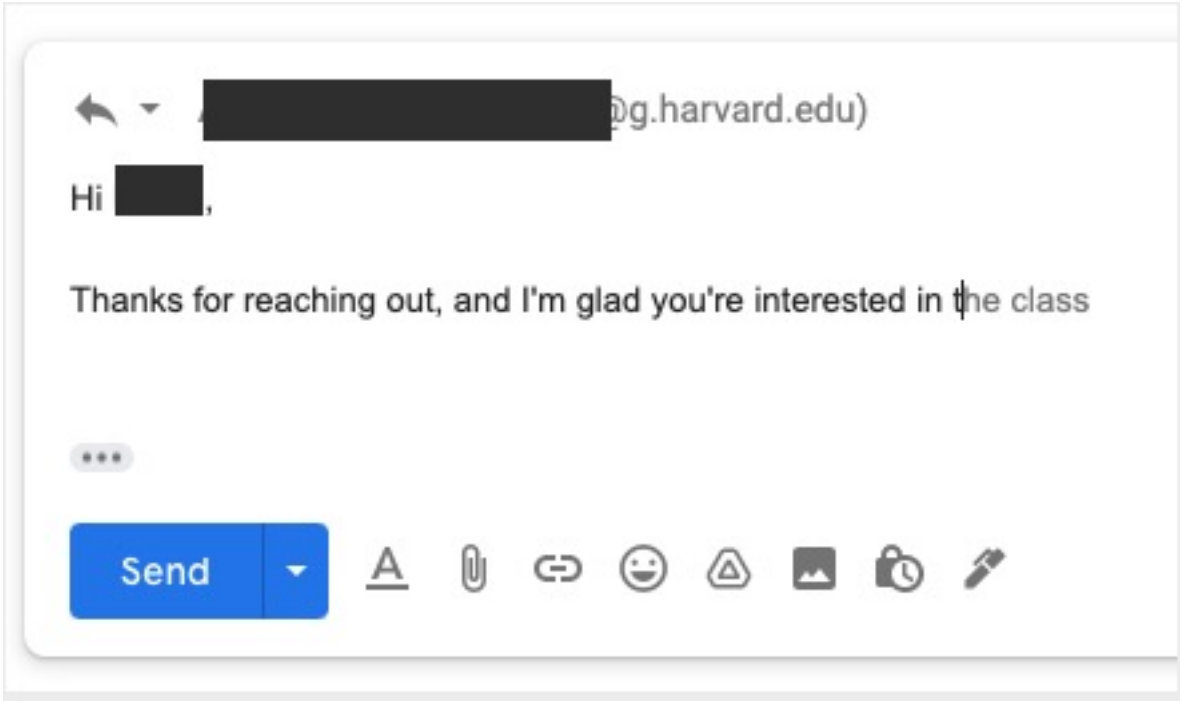
Translation



NLP Successes

NLP has been around since the 1960s, but the progress in the past 10 years has been astronomical. Models are becoming effective enough for consumer use.

Auto-complete



NLP Successes

NLP has been around since the 1960s, but the progress in the past 10 years has been astronomical. Models are becoming effective enough for consumer use.

Auto-complete

NLP Successes

NLP has been around since the 1960s, but the progress in the past 10 years has been astronomical. Models are becoming effective enough for consumer use.

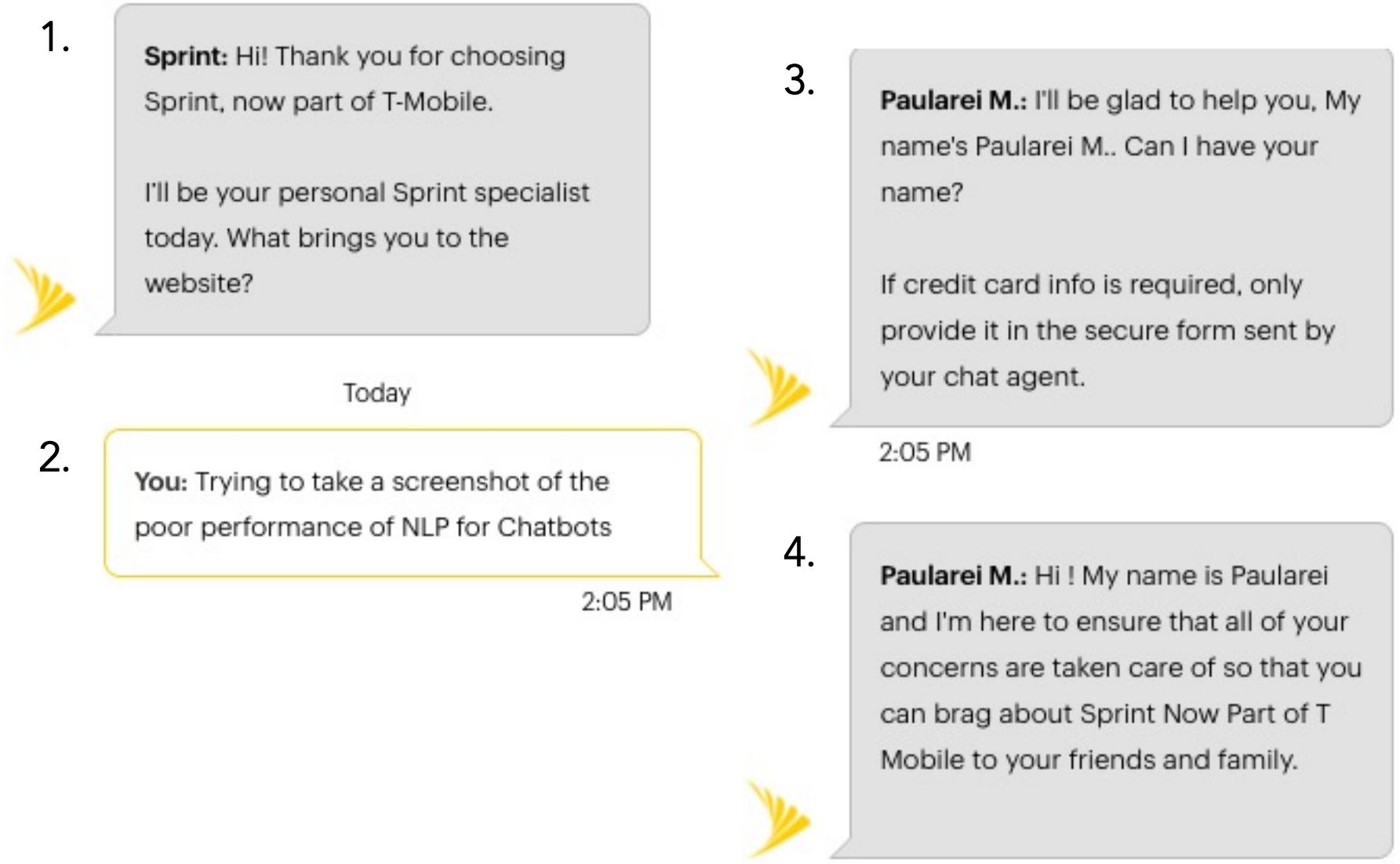
Text Classification

Spam

Not Spam

NLP ~~Successes~~ has room for improvement

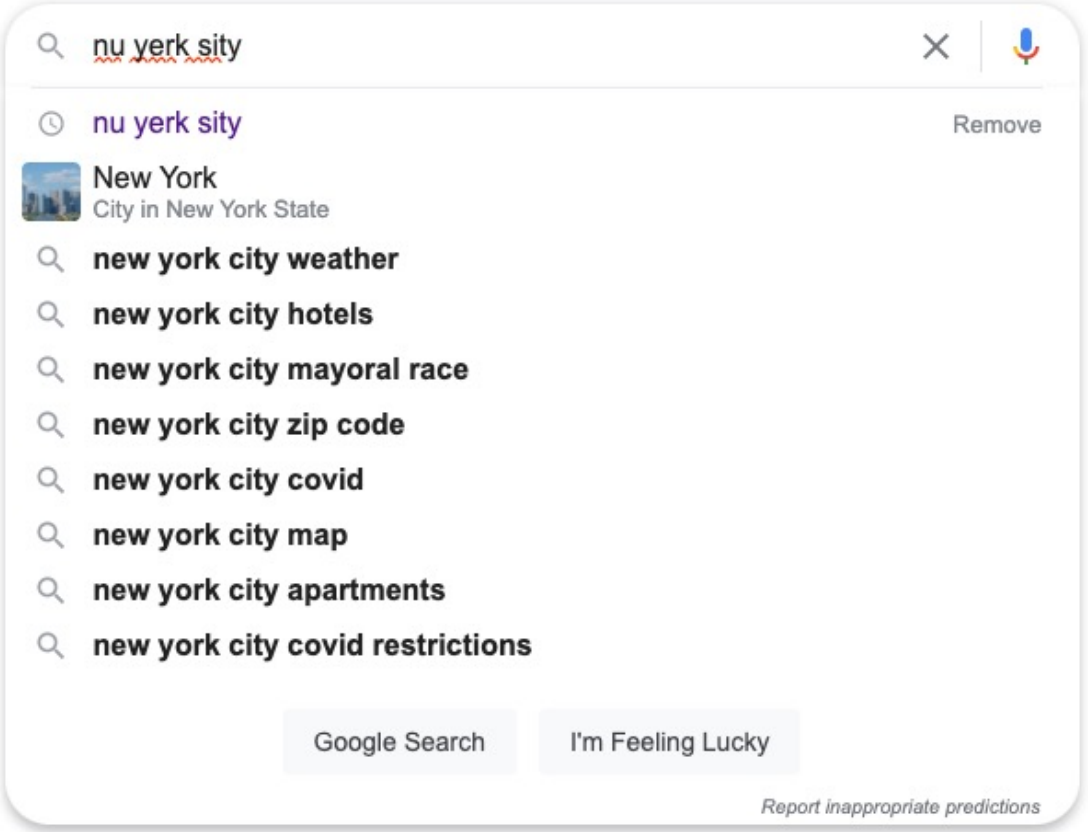
Chatbots



NLP Successes

NLP has been around since the 1960s, but the progress in the past 10 years has been astronomical. Models are becoming effective enough for consumer use.

Search Engines
(information retrieval)



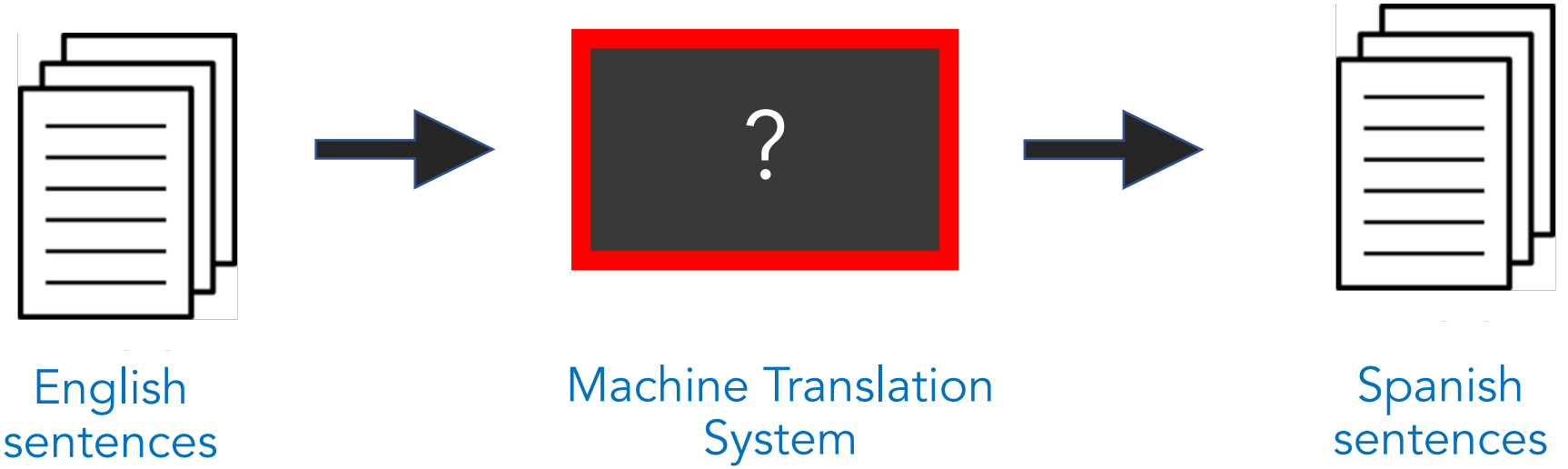
How do these systems work?!

We have data
(e.g., text format)

We want a
magical blackbox

That outputs golden
usefulness

How do these systems work?!



How do these systems work?!

English sentences

Machine Translation System

Spanish sentences

X

$f(X)$

Y

How do these systems work?!

While we don't necessarily have to produce a **Y** for every NLP problem (i.e., supervised learning), most interesting problems do.

Luckily, we have tons of (**X**, **Y**) data pairs, right? Kind of.

English sentences

Machine Translation System

Spani sentenc

X

$f(X)$

Y

What's in the box?!

Our computational model could be anything:

- Rule-based system
- CRF
- HMM
- Statistical Alignment Model (e.g., IBM Models)
- Probabilistic Graphical Model
- Neural Network

?

Model

$f(X)$

Se7en (1995)

Spanish sentences

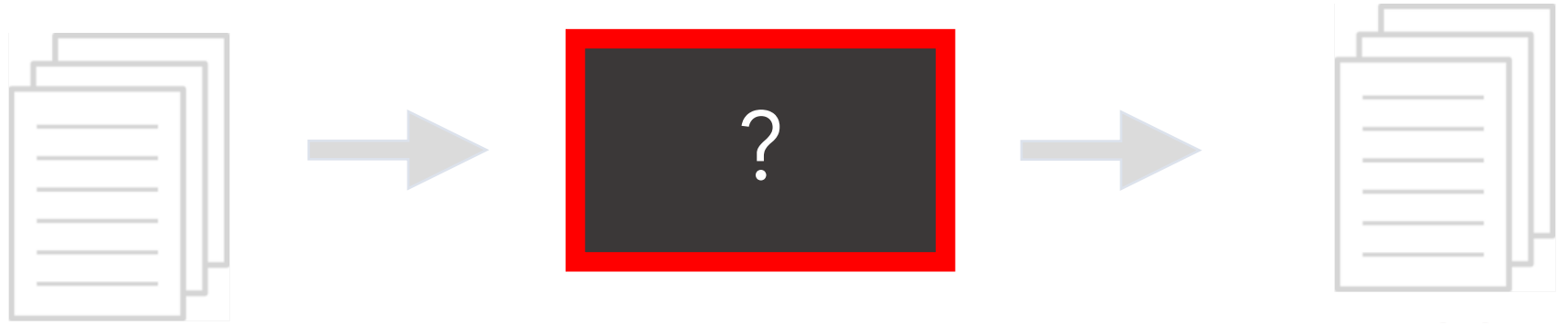
Y

What's in the box?!

Regardless of the model, it doesn't actually "understand" language. It simply *approximates* understanding for a particular objective. This seems good enough.

What's in the box?!

Neural Networks, specifically Deep Neural Networks, dominate all of machine learning these days, including NLP.



What is this course?

You will learn advanced, **deep learning** approaches to **NLP** and conduct research on a significant project with classmates.

Learning Objectives

- understand the theoretical concepts behind NLP tasks and models
  - Not just a surface-level understanding of LSTMs, Transformers;
  - What is the model *actually* doing? How does it work? Why does it work? What are its limitations? Past approaches? What are alternatives?
- write effective programming solutions to popular problems in NLP
- tackle your own, novel goals with text data once this course is over
- conduct substantial, original NLP research

Learning Objectives

- understand the theoretical concepts behind NLP tasks and models
  - Not just a surface-level understanding of LSTMs, Transformers;
  - What is the model *actually* doing? How does it work? Why does it work? What are its limitations? Past approaches? What are alternatives?
- write effective programming solutions to popular problems in NLP
- tackle your own, novel goals with text data once this course is over
- conduct substantial, original NLP research

I want everyone to finish the course feeling confident and empowered to develop NLP solutions and embark on novel research

Why so research-heavy?

- Knowledge, about anything in our world, is created via research and publishing
- It's further explored and experimented with
- Knowledge/models/approaches that seem sufficiently good and important make their way into courses and books
- Some of this knowledge gets added to the field's foundation

Why so research-heavy?

Researcher:

- What is possible to build?
- How can we use existing blocks in new ways?
- What are the limitations of current blocks?

Image source: lego.com

Software Developer:

- The Builders. Creators.
- Interested in tools to build better, quicker, organized, useful structures

Manager:

- Bridges everyone's skills to make great things actually happen

Research Project

- Projects will last most of the semester (12 weeks)
- Should explore a significant question that *nobody* knows the answer to
- You will work in teams of 3 students
- Propose ideas, we approve, all students select which projects to join
- Final product will be a 4-5 page short paper (plus other deliverables)
- Goal is for it to be worthy of submitting to a top conference or workshop
- **It's okay if your results aren't SOTA**; your approach/methodology is what matters

Research Project

- "Only" pushing the SOTA results is a weak research finding*
- Common approaches:
  - Poke holes into the current understanding/papers/models
  - Exploit/leverage data in a novel way
  - Make assumptions or relax constraints to reveal new insights
  - Combine past approaches to yield a better approach
  - Make models faster or easy-to-use without losing much performance

* unless the method was novel (e.g., new architecture, approach, way of using data, etc)

Research Project

Inappropriate projects:

- Make a chatbot for ____
- Run a Transformer on ____ data
- Run a summarizing system on a new, ____ corpus

This demonstrates software development and NLP applications, not research.

They *might* serve as the starting points for research, but research questions should be your entry point and driving force.

Research Project

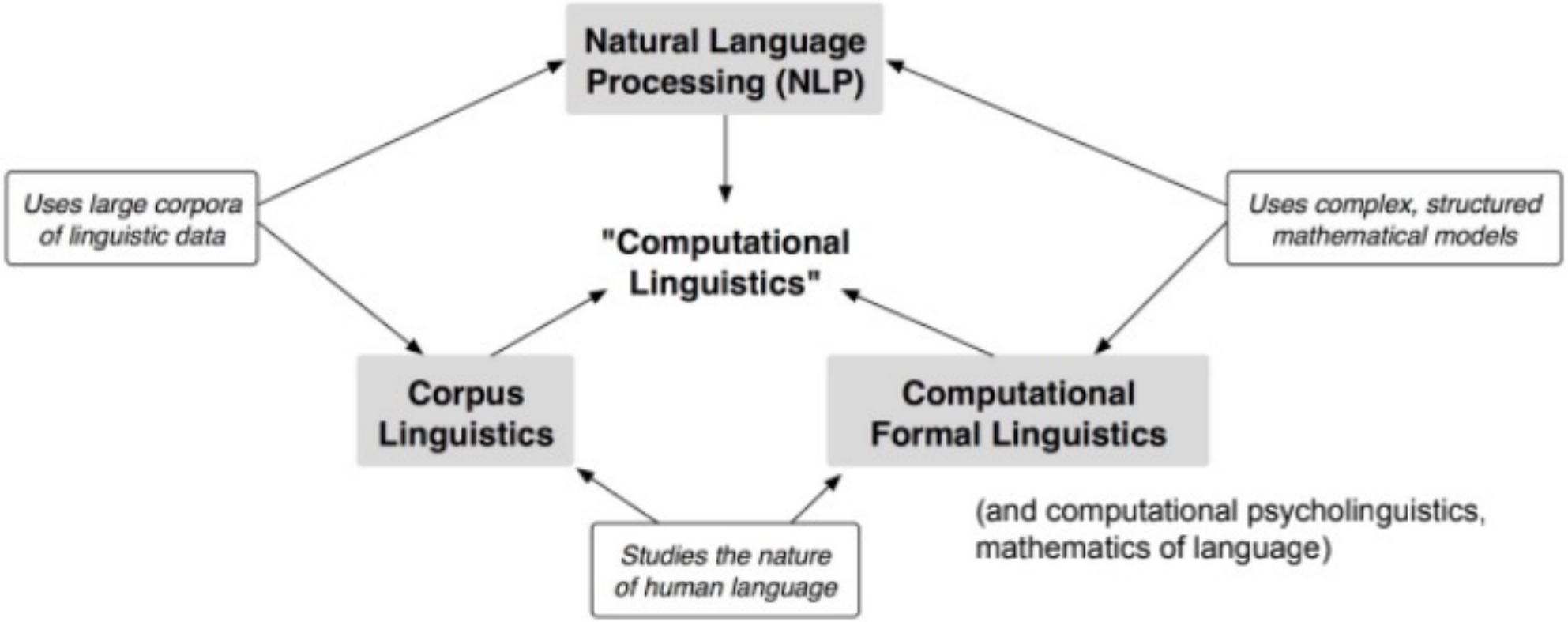
You **cannot** do a research project on stock market data / prediction.

- Not research-y enough
- Too difficult (way too latent)
- Not interesting
- Difficult metrics for comparison

What is this course NOT about?

- This is not a linguistics course
- Not a *foundational* machine learning course (we expect you have this foundation already). This course adds to your foundation.
- Not about teaching you PyTorch or TensorFlow
- Not exclusively about your demonstrating the *learned* concepts; it includes your ability to leverage what you're learning and extend it further (i.e., research project)

NLP vs Computational Linguistics



From Sam Bowman's LING-GA 1012 Spring 2020

CS287 vs CS187

CS287

- This class is new and I'm making it from scratch
- Graduate course
- Assumes you've taken an ML course and understand the basics of a Neural Net
- Research-heavy (project counts 50%)
- Many lectures, no labs
- Covers more advanced topics and more NLP tasks than CS187

CS287 vs CS187

CS187

- Taught by the legendary Stuart Shieber (2nd time ever)
- Undergraduate course
- Assumes no pre-reqs, just competency in programming and CS
- No research project or exam
- Many labs, few lectures (~6)
- Focuses more on linguistics (e.g., parsing) than CS287

CS287 Pre-requisites

If you haven't already, you should thoroughly look at and try to complete HW #0 to ensure you meet the pre-reqs, e.g.:

- Basics of a feed-forward neural net
- How learning works (e.g., loss functions, backprop, gradient descent)
- Modifications (e.g., L1 and L2 regularization, dropout, hyperparameter optimization)
- Reasonably comfortable with TensorFlow or PyTorch
  - We officially support PyTorch but will gladly accept assignments that are in either
- I assume no prior NLP knowledge

Expectations of you

Expected to demonstrate not only the ability to understand the core concepts of this course, but to be able to do some research, i.e.:

- read papers beyond what's mentioned in class
- critique other papers (even if the concepts are new to you)
- be curious
- come up w/ questions
- try to answer these questions

I expect you to challenge yourself. This class is intended to be challenging (but not too challenging).

Expectations of me

I want this course to be an incredibly rewarding experience and the best CS class you take. I pushed to create this course and offer it. Huge thanks to IACS, DCE, CS, and higher-up folks at SEAS for approving it.

Hold me accountable to make it as **equitable, fair, clear, and smooth** of an experience as possible. I gladly welcome anonymous feedback at any time, and I solicit such as part of each HW assignment.

If something needs improving, let's work to make it better. I'm here to help you all learn and succeed in this course.

Outline

- What is this course?
- Logistics
- NLP: what, why, how?

Outline

- What is this course?
- Logistics
- NLP: what, why, how?

Student body

- 35 Harvard (non-DCE students)
  - 9 PhDs
  - 13 MS
  - 2 Seniors
  - 11 Juniors (aka Junior M.A.F.I.A.)
- 20 Harvard DCE students from awesomely diverse backgrounds

Teaching Staff

- Me (instructor)
- 4 incredible TFs/CAs
  - Alex Lin
  - William Tong
  - Annie Zhu
  - Richard Qui

20 TF applicants, 10 hrs of interviews with mock tutoring sessions.

You're in good hands.

Background:

- PhD in NLP from Brown
- Machine Learning since 2004
- MIT Lincoln Lab (3) years + 10 internships

Hobbies:

Travelling (formerly), woodworking, playing baseball, hiking, photography, running

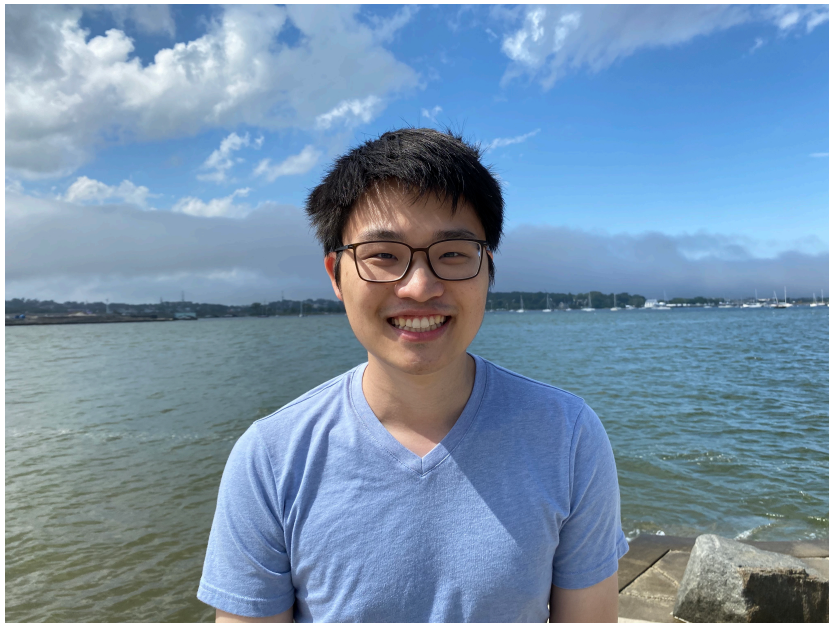
Chris Tanner

Most fun during the weird, past year:

Hiking the Presidential Traverse again. Grilling outdoors. Iceland this summer. Not catching COVID.

Background:

- Current PhD student in Harvard CS
- Former researcher @ASAPP (NLP startup)
- CS Bachelor's and CSE (IACS) Master's @Harvard



Alex Lin

Hobbies:

Taking long walks in scenic areas, biking, watching basketball, reading, all things machine learning :)

Most fun during the weird, past year:

Watching lots of Netflix, and having friends visit me in Boston!

Background:

- From Chicago
- BA in CS and Stats from Columbia
- Interested in cognitive science + ML

Hobbies:

Running, longboarding, music, cooking, learning about (and losing money on) cryptocurrency

Most fun during the weird, past year:

Rollerblading from NYC to the Tappan Zee bridge and back. Hiking Breakneck Ridge. Hanging out with my family.

William Tong

Background:

- Senior @ Harvard
- Studying a mixture of Math/CS/Stat/Physics
- Two internships in Quant Finance

Hobbies:

Fiction, TV shows, singing, taking long walks, trying new restaurants, reading random Wikipedia pages about historical figures but never events

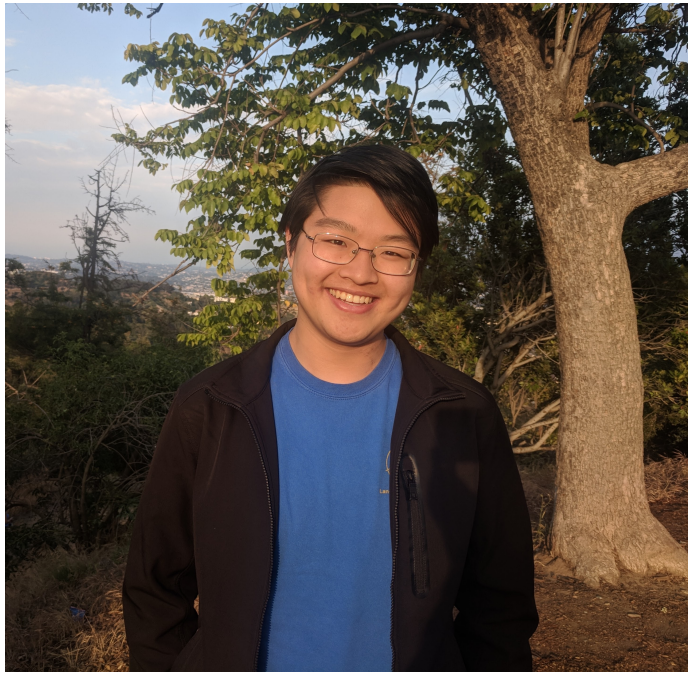
Most fun during the weird, past year:

Got better at cooking, moved eight times, pretended to be a senior at senior sunrise

Annie Siye Zhu

Background:

- Harvard College ‘23, studying physics + CS
- Physics-informed ML for turbulence @ NASA
- Previously, explainable recurrent architectures for particle physics



Richard Qiu

Hobbies:

Hiking/backpacking, cooking, baking, running (badly)

Most fun during the weird, past year:

Stargazing and hiking in Banff National Park. Biked the Chesapeake & Ohio Canal. Road tripped ~6000 miles cumulatively.

Times

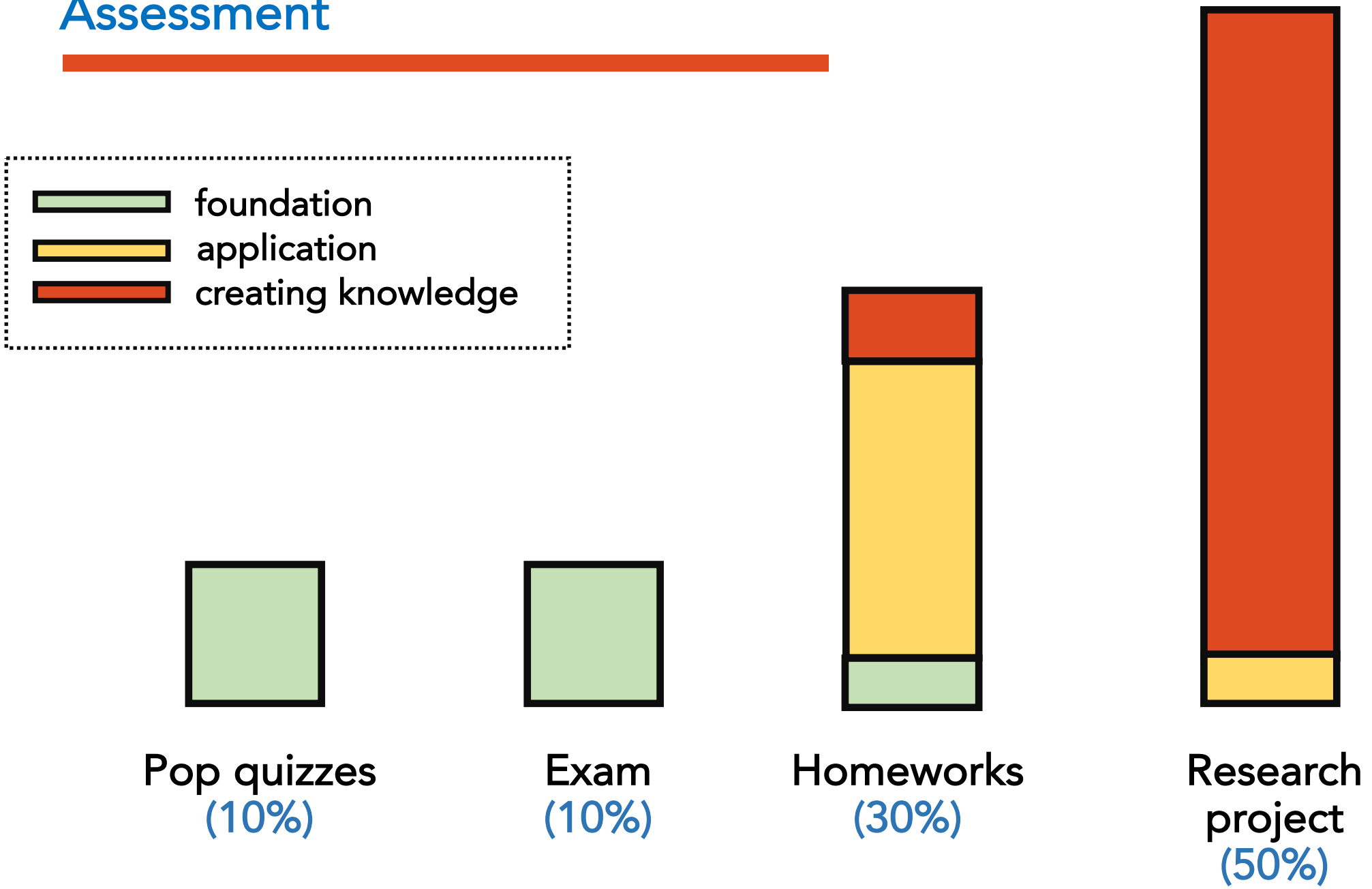
- **Class**: Tues and Thurs @ 9:45am – 11am (attendance is mandatory)
- **Office Hours** (will have an in-person and Zoom support):
  - Mon @ 1pm – 3pm: Alex and Annie
  - Wed @ 5pm – 7pm: Richard and William
  - Thurs @ 2:30pm – 4:30pm: Chris
  - Sat @ 10am – 11am: Annie (Zoom only)
- **Pop quizzes** will be released at the beginning of ~8 lectures (closed book)
- **1 exam** will be given in-class (closed book)
- **4 homeworks**. Released on Tuesdays, due 2 weeks later (Mon night)

Assessment

All course assessment is structured around 3 pillars:

- Building a **foundation** of theory/concepts (pop-quizzes and exam)
- Demonstrating you can **apply** the knowledge (homework)
- **Creating new knowledge** (12-week research project)

Assessment


foundation
application
creating knowledge
Pop quizzes
(10%)
Exam
(10%)
Homeworks
(30%)
Research
project
(50%)

Homeworks

- Individual work (cannot work in pairs)
- Designed to be challenging yet fun
- Content centers around 4 lectures
- Everyone is granted 3 free late days for the entire semester
  - Beyond this, each additional late day docks 10% from that homework's grade
  - Any single homework that is > 3 days late will receive a 0%
  - No need to communicate to us that you're "using a late day". We keep track.

Quizzes

- **Non-DCE:** quizzes conducted in-class, on-paper
- **DCE**: quizzes on EdStem
- Your lowest 3 pop quizzes will be dropped.
- Content can be from **any** of the previous lectures (content builds)

Exam

- **Non-DCE:** exam conducted in-class, on-paper
- **DCE**: exam on EdStem
- Content will *largely* be from the lectures
- May contain few programming questions (pseudocode, not actual TF/PyTorch)
- May contain research questions (about the papers you've read)

Research Project

- Please start skimming papers published in ACL, NAACL, EMNLP, COLING from the past 2 years.
- It's okay if they don't make sense to you. For now, just try to get an idea of *what* people work on, later you can focus on the *how*.
- Don't get bogged down in the details: **Abstract, Intro, Conclusions** are enough.

Schedule

- Each HW lasts 2 weeks
- Early lectures concern models
- Latter lectures concern tasks (gray)
- Once lectures wrap up, the research project ramps up
- Plenty of time to work on the research project
- Final presentations Dec 9

WEEK	DATE	DAY	LECTURE	TOPIC	HW #	HW DUE	RESEARCH DUE
1	Sept 2	Thurs	1	Course overview + Intro	1		
2	Sept 7	Tues	2	Representations: BoW + TFIDF	1	1 REL	
2	Sept 9	Thurs	3	LMs: n-grams	1		
3	Sept 14	Tues	4	word2vec + GloVe	1		
3	Sept 16	Thurs	5	RNNs + LSTMs	2		
4	Sept 21	Tues	6	Bi-LSTMs + ELMo	2	1 DUE; 2 REL	
4	Sept 23	Thurs	7	seq2seq + Attention	2		
5	Sept 28	Tues	8	Machine Translation	2		
5	Sept 30	Thurs	9	Transformers (vanilla)	3		PHASE 1 (5%)
6	Oct 5	Tues	10	BERT	3	2 DUE; 3 REL	
6	Oct 7	Thurs	11	BERT for downstream tasks	3		
7	Oct 12	Tues	12	GPT-2	3		
7	Oct 14	Thurs	13	Bias and Fairness	4		PHASE 2 (5%)
8	Oct 19	Tues	14	Summarization	4	3 DUE: 4 REL	
8	Oct 21	Thurs	15	Entity Linking	4		
9	Oct 26	Tues	16	Coreference Resolution	4		
9	Oct 28	Thurs	17	Commonsense Reasoning	-		PHASE 3 (10%)
10	Nov 2	Tues	18	Adversarial NLP	-	4 DUE	
10	Nov 4	Thurs	19	Interpretability and Probing Tasks	-		
11	Nov 9	Tues	20	NLU	-		
11	Nov 11	Thurs	21	Review: last day of lecture content			PHASE 4
12	Nov 16	Tues	22	FINAL EXAM (10%)			
12	Nov 18	Thurs	23	PROJECT DISCUSSIONS			
13	Nov 23	Tues	24	PROJECT DISCUSSIONS			PHASE 5
13	Nov 25	Thurs		THANKSGIVING			
14	Nov 30	Tues	25	PROJECT DISCUSSIONS			
14	Dec 2	Thurs	26	PROJECT DISCUSSIONS			
15	Dec 7	Tues		READING PERIOD			
15	Dec 9	Thurs	27	FINAL PRESENTATIONS			Paper (20%) Presentation (5%) Code (5%)

Outline

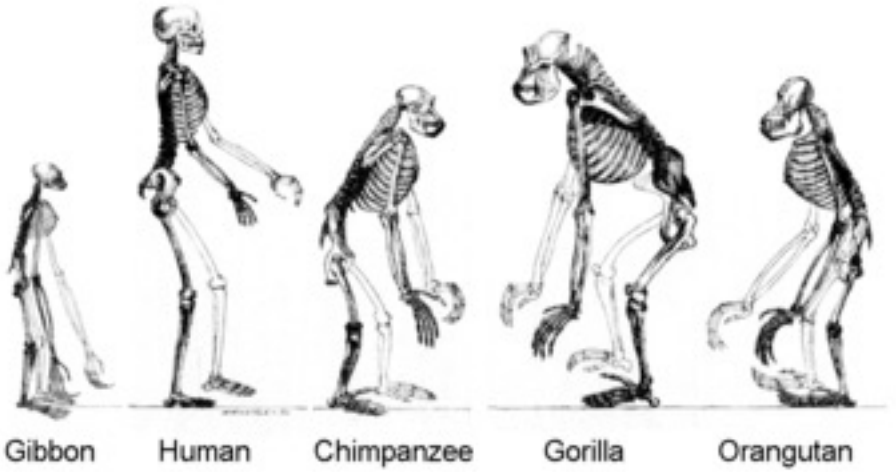
- What is this course?
- Logistics
- NLP: what, why, how?

Outline

- What is this course?
- Logistics
- NLP: what, why, how?

Language

Language is special and complex



- Distinctly human ability
- Paramount to human evolution
- Influenced by many social constructs
- Incredibly nuanced
- Language forms capture multi-dimensions
- Language evolves over time

Slide adapted from or inspired by Chris Manning and Richard Socher

Language

Language is constructed to convey speaker's/writer's meaning

- More than an environmental, survival signal
- Encodes complex information yet simple enough for babies to quickly learn

A discrete, symbolic communication system

- Lexicographic representation (i.e., characters that comprise a word) embody real-world constructs
- Nuanced (e.g., "Sure, whatever", "Yes", "Yesss", "Yes?", "Yes!", Niiice)

Slide adapted from or inspired by Chris Manning and Richard Socher

Language

Language symbols are encoded as continuous communication signals, and are invariant across different encodings (same underlying concept, different surface forms)

Slide adapted from or inspired by Chris Manning and Richard Socher

Language is funny

"Red tape holds up new bridges"

"Hospitals are sued by 7 foot doctors"

"Local high school dropouts cut in half"

"Tesla crashed today"

"Obama announced that he will run again"

"Kipchoge announced that he will run again"

"She made him duck"

"Will you visit the bank across from the river bank? You can bank on it"

"Yes" vs "Yes." vs "YES" vs "YES!" vs "YAS" vs "Yea"

NLP: why

The entire point of computers is to assist humans.

Having computers "understand" our language and how we communicate as a species is a natural entry point and required step to significantly assisting us in our lives.

Common NLP Tasks (aka problems)

Syntax

Morphology

Word Segmentation

Part-of-Speech Tagging

Parsing

- Constituency
- Dependency

Discourse

Summarization

Coreference Resolution

Semantics

Sentiment Analysis

Topic Modelling

Named Entity Recognition (NER)

Relation Extraction

Word Sense Disambiguation

Natural Language Understanding (NLU)

Natural Language Generation (NLG)

Machine Translation

Entailment

Question Answering

Language Modelling

Common NLP Tasks (aka problems)

Syntax

Morphology

Word Segmentation

Part-of-Speech Tagging

Parsing

- Constituency
- Dependency

Discourse

Summarization

Coreference Resolution

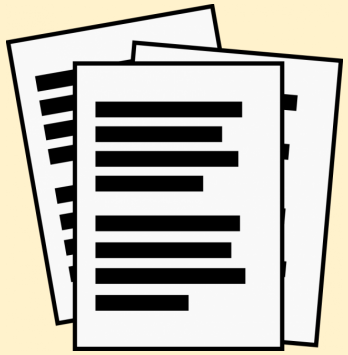
Semantics

Sentiment Analysis

Topic Modelling

Question Answering

Language Modelling



"Overall, Pfizer's COVID-19 vaccine is very safe and one of the most effective vaccines ever produced"

Common NLP Tasks (aka problems)

Syntax

Morphology

Word Segmentation

Part-of-Speech Tagging

Parsing

Constituenc

Dependenc

Discourse

Summarization

Coreference Re

Semantics

Sentiment Analysis

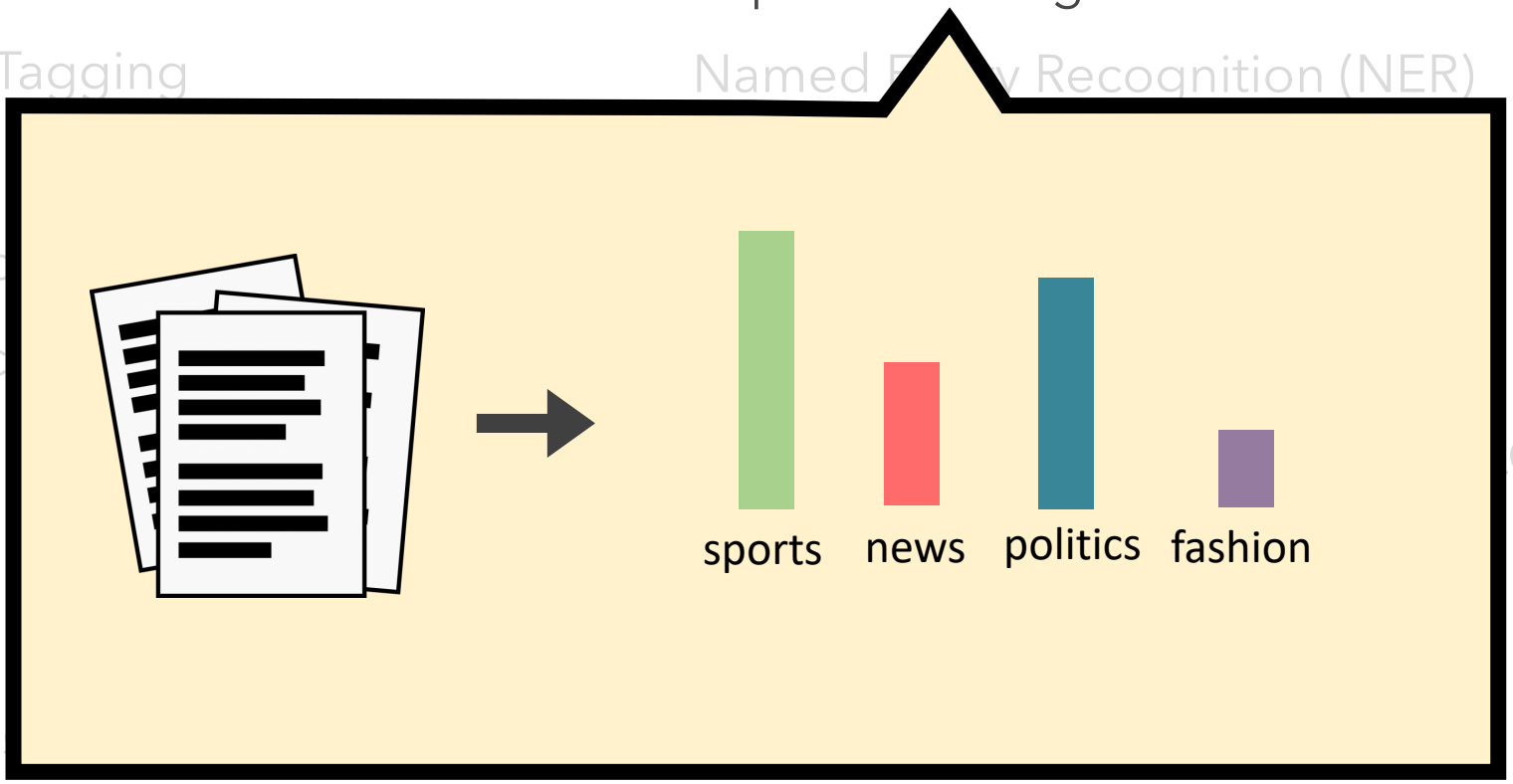
Topic Modelling

Named E y Recognition (NER)

(NLU)

G)

Language Modelling



Common NLP Tasks (aka problems)

Syntax

Morpholog

Word Segm

Part-of-Spe

Parsing

Constit

Dependency

Discourse

Summarization

Coreference Resolution

Semantics

IER)

"Alexa, play Drivers License by Olivia Rodrigo"

"Alexa, play Drivers License by Olivia Rodrigo"

INTENT SONG ARTIST

Natural Language Understanding (NLU)

Natural Language Generation (NLG)

Machine Translation

Entailment

Question Answering

Language Modelling

Common NLP Tasks (aka problems)

Syntax

Morphology

Word Segmentation

Part-of-Speech Tagging

Parsing

- Constituency
- Dependency

Semantics

Sentiment Analysis

Topic Modelling

Named Entity Recognition (NER)

ng (NLU)

Natural Language Generation (NLG)

Discourse

Summarization

Coreference Resolution

Machine Translation

Entailment

Question Answering

Language Modelling

El perro marrón → The brown dog

SPANISH ENGLISH

Very brief history of NLP

- 1960s: pattern-matching and rules (highly limiting)
- 1970s – 1980s: linguistically rich, logic-driven systems; labor-intensive successes on a few, very specific tasks
- 1990s – 2000s: statistical modelling takeover! ML becomes a central component; some systems are deployed for practical use (e.g., speech to text)
- 2010s – 2020s: Deep Learning (neural nets) yields astronomical progress on nearly every NLP task; systems become fairly useful for consumers
- 2020s – 2030s?: you can help drive the change

Slide adapted from or inspired by Sam Bowman

Very brief history of NLP

First huge revolution: early 1990s (statistical approaches)

"But it must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term"

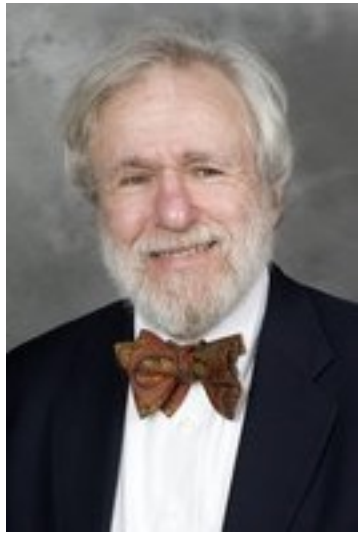
-- Noam Chomsky (1969)

"Anytime a linguist leaves the group, the recognition rate goes up"

-- Fredrick Jelinek (1988)

Very brief history of NLP

First huge revolution: early 1990s (statistical approaches)



"I refer to all of my work before ~1990 as the B.S. era. That is, 'before statistics'"

-- paraphrasing my PhD adviser, Eugene Charniak at his ACL Lifetime Achievement Award (2011)

GPT-2 (generates text and can fine-tune on your own data)

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

"The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."
The stolen material was taken from the University of Cincinnati's Research Triangle Park nuclear research site, according to a news release from Department officials.
The Nuclear Regulatory Commission did not immediately release any information.

NLP nowadays

GPT-2 (generates text and can fine-tune on your own data)

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%

NLP nowadays

Table 3: Video captioning performance on YouCook II. We follow the setup from [39] and report captioning performance on the validation set, given ground truth video segments. Higher numbers are better.


GT: add some chopped basil leaves into it

VideoBERT: chop the basil and add to the bowl

S3D: cut the tomatoes into thin slices


GT: cut the top off of a french loaf

VideoBERT: cut the bread into thin slices

S3D: place the bread on the pan


GT: cut yu choy into diagonally medium pieces

VideoBERT: chop the cabbage

S3D: cut the roll into thin slices


GT: remove the calamari and set it on paper towel

VideoBERT: fry the squid in the pan

S3D: add the noodles to the pot

VideoBERT: A Joint Model for Video and Language Representation Learning. Sun, et al. ICCV 2019.

Deep Learning (breakthrough moment)

Data

14 million images.
20,000 distinct
categories (e.g., shoes).

Task

Given an image,
correctly predict which
category it belongs to

AlexNet Model

The network
achieved a top-5
error of 15.3%,
more than 10.8
percentage points
lower than that of
the runner up.

ImageNet Classification with Deep Convolutional Neural Networks. Krizhevsky, et al. (2012)

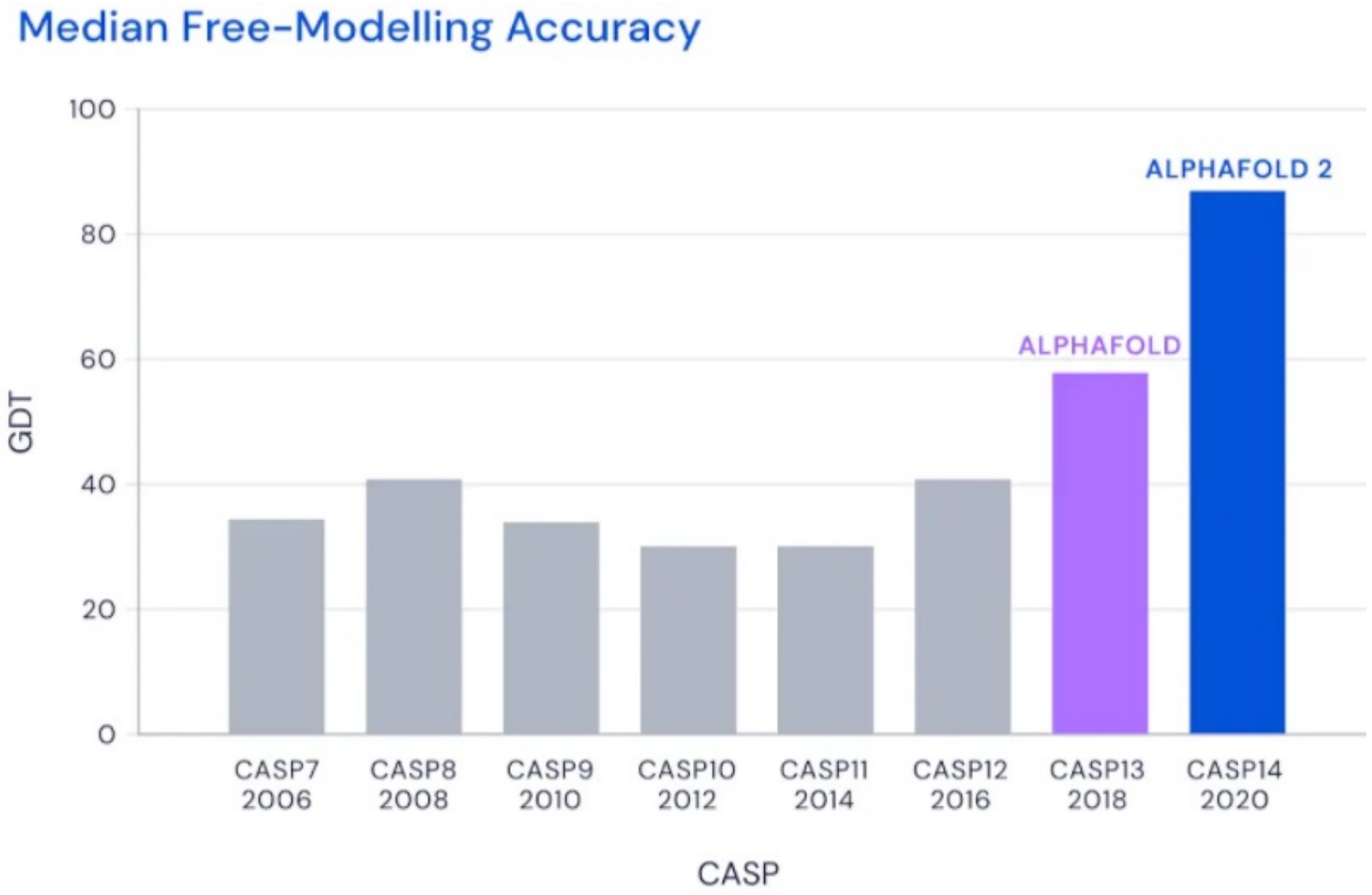
Deep Learning (recent breakthrough)

AlphaFold: a solution to a 50-year-old grand challenge in biology

T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Deep Learning (recent breakthrough)



Deep Learning (recent breakthrough)

In the results from the 14th CASP assessment, released today, our latest AlphaFold system achieves a median score of 92.4 GDT overall across all targets. This means that our predictions have an average error (RMSD) of approximately 1.6 Angstroms, which is comparable to the width of an atom (or 0.1 of a nanometer). Even for the very hardest protein targets, those in the most challenging free-modelling category, AlphaFold achieves a median score of 87.0 GDT (data available here).

Deep Learning

- **Deep Learning** is just neural networks with more than 1 hidden layer (non-linear activation functions).
- For the 1st time ever, one paradigm of modelling (deep learning) yields the best results across nearly every domain of problems
- Our understanding of why and how the results are so compelling is very surface-level.
- Much work lies ahead (e.g., bias/fairness, explainability, robustness)

The Two Cornerstones of NLP

How do we get *any* system to process, "understand", leverage language?

- **Representation**: how do we transform symbolic meaning (e.g., words, signs, braille, speech audio) into something the computer can use
- **Modelling**: given these represented symbols, how do we use them to model the task at hand?