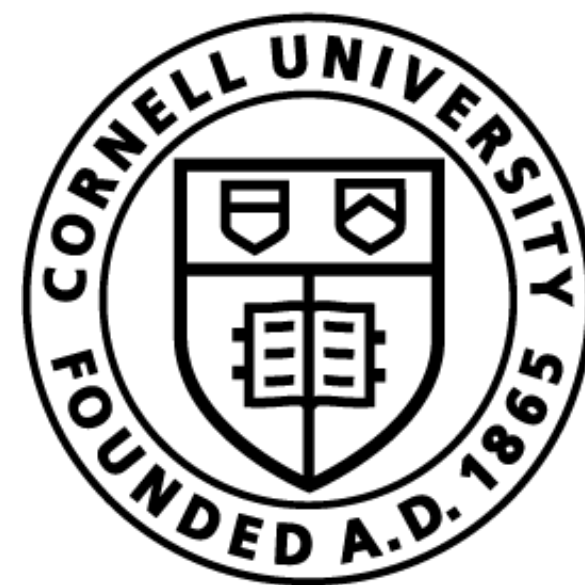


# Lecture 1: Introduction to Introduction to Natural Language Processing



Cornell Bowers CIS  
**Computer Science**

# Course logistics (+more at the end)

- ▶ Course website is up! <https://www.cs.cornell.edu/courses/cs4740/2025sp/>
- ▶ Course policies listed on the webpage.
- ▶ Up-to-date schedule and slides will always be available on this webpage.

## Instructors:



**Claire Cardie**



**Tanya Goyal**

# Today

- ▶ What is NLP?
- ▶ Why is NLP hard?
- ▶ Course Outline
- ▶ More Administrative Stuff.

# What is NLP anyway?

**Fundamental Goal:** Build technologies to solve tasks requiring a deep understanding of natural language.

# What is NLP anyway?

**Fundamental Goal:** Build technologies to solve tasks requiring a deep understanding of natural language.

Languages we use to communicate with each other.

# What is NLP anyway?

**Fundamental Goal:** Build technologies to solve tasks requiring a deep understanding of natural language.

Languages we use to communicate with each other.

Process/interpret/communicate as well as humans (or better?)

# What is NLP anyway?

**Fundamental Goal:** Build technologies to solve tasks requiring a deep understanding of natural language.

Languages we use to communicate with each other.

Process/interpret/communicate as well as humans (or better?)

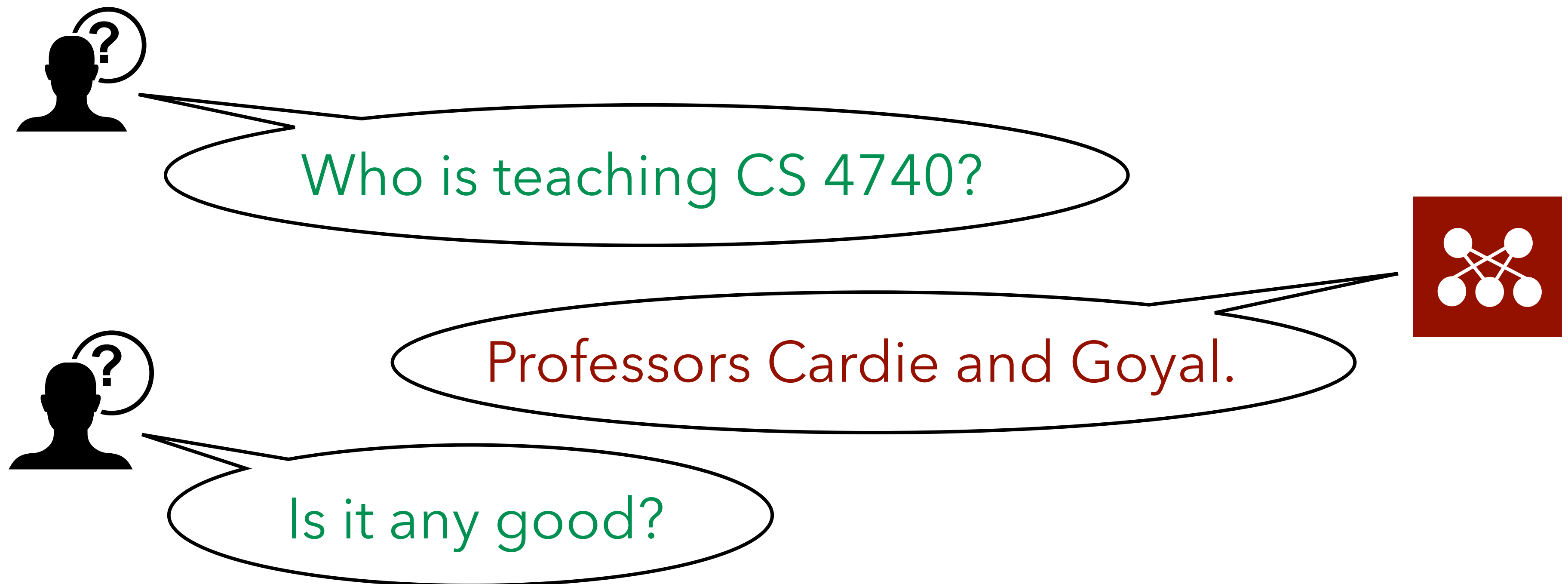
Any task with text inputs and/or outputs is in scope.

# NLP tasks

All tasks where either the **input**  $X$  and/or the **output**  $Y$  is text is in scope.

**Help us communicate with machines.**

E.g. Dialogue systems, question answering, etc.





# NLP tasks

All tasks where either the **input X** and/or the **output Y** is text is in scope.

**Help us transform text.**

E.g. Machine translation, grammar correction, summarize etc.

जाने-माने वैज्ञानिक सिवान के. को भारतीय अंतरिक्ष  
अनुसंधान संगठन (इसरो) का अध्यक्ष नियुक्त किया गया है।

Translate

New Delhi: Noted scientist Sivan K was  
appointed Chairman of the Indian Space  
Research Organisation on Wednesday.

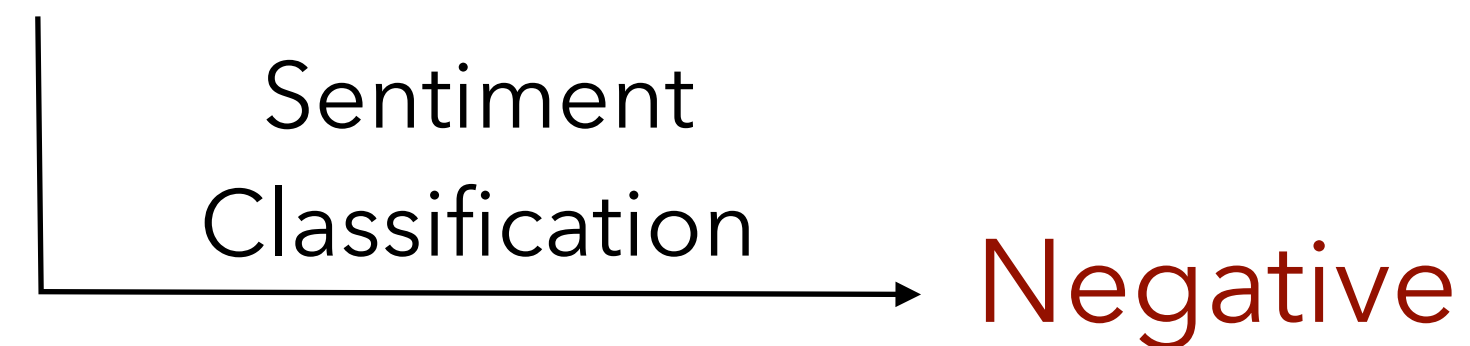
# NLP tasks

All tasks where either the **input**  $X$  and/or the **output**  $Y$  is text is in scope.

**Help us understand and analyze text corpora or language.**

E.g. syntactic analysis, text classification, topic modeling etc.

"I absolutely loved waiting three hours in line for the worst meal of my life."



# NLP tasks

All tasks where either the **input**  $X$  and/or the **output**  $Y$  is text is in scope.

**Help us understand and analyze text corpora or language.**

E.g. syntactic analysis, text classification, topic modeling etc.

"What do Vegans do in their Spare Time? Latent Interest Detection in Multi-Community Networks", Hessel et al., 2015

Vegans

Top Interests  
diet, food, cooking,  
animal, flora

Latent Interests  
Anarchism, yoga, VegRecipes,  
Feminism, bicycling, [...]



# Why is NLP hard? Ambiguity

"John went to the bank."




Two different meanings of the word bank.




# Why is NLP hard? Ambiguity

"Retrieve all the local patient files."

Retrieve all the local patient files.

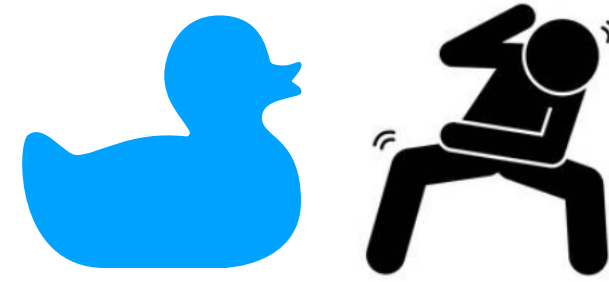


Retrieve all the local patient files.



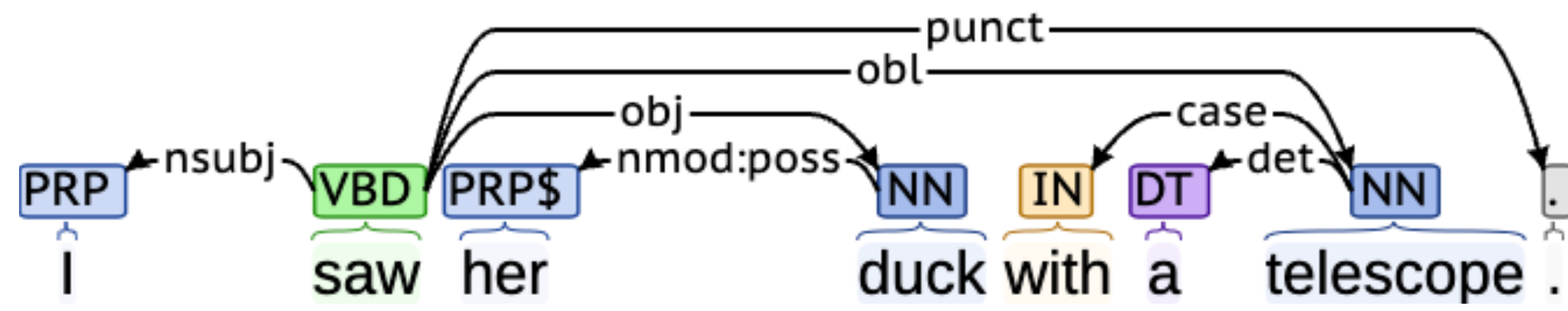
Syntactic ambiguity: what modifies what?

# Why is NLP hard? Ambiguity



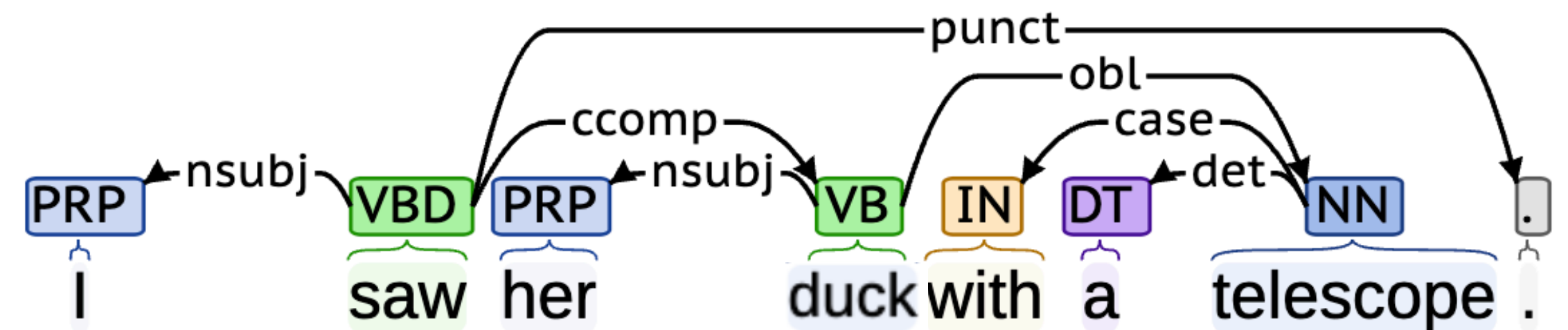
"I saw her duck with a telescope."

How many possible interpretations of this can you think of?



► I used a telescope to see her duck 

► I used a telescope to see her duck 



► I saw her  who had a telescope.

► I saw her  with a telescope in hand.

# Why is NLP hard? Ambiguity

- ▶ Cases that are easy for humans can be ambiguous for models.

Susan knows all about Ann's personal problems because she is nosy.

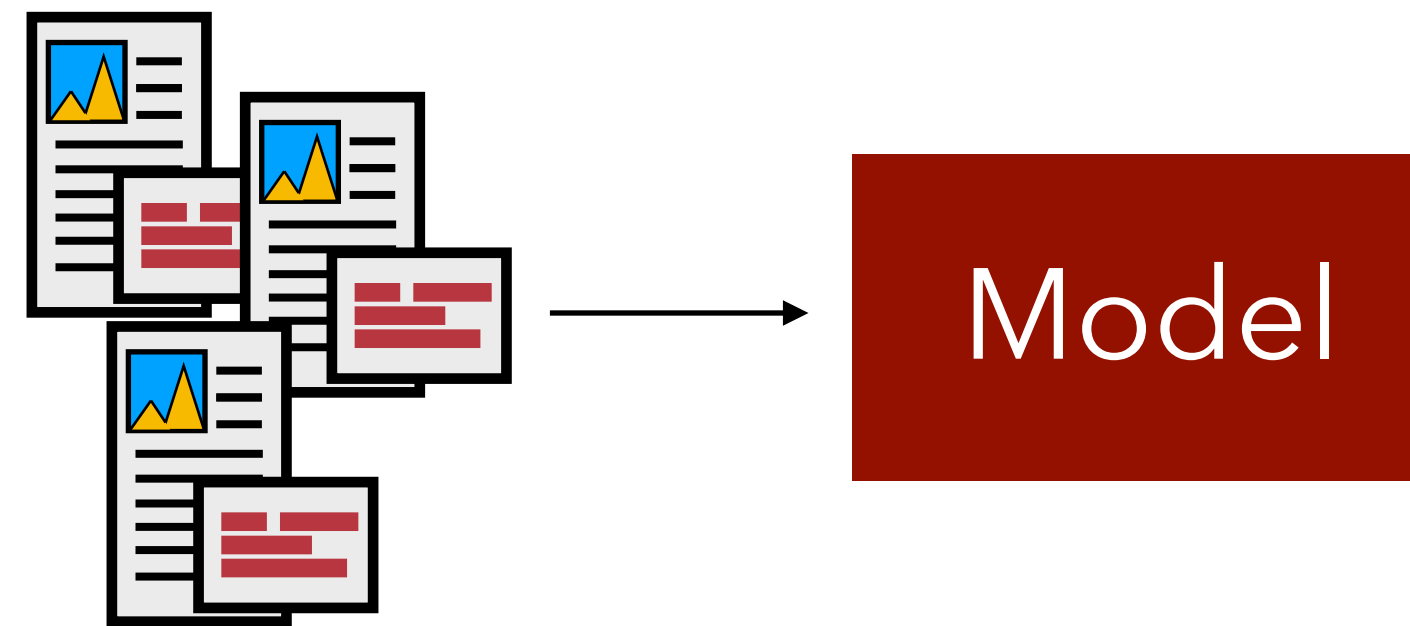
Susan knows all about Ann's personal problems because she is indiscreet.

Coreference resolution: Who is she?

- ▶ Easy for humans to resolve given the context, but difficult for statistical models. Why?

# Why is NLP hard? Data

- ▶ NLP models learn from data.



- ▶ Impossible to include data points corresponding every possible linguistic phenomenon and edge case in this data.
- ▶ Models can struggle to learn rare phenomenon.
- ▶ This is true even for large language models today that are trained on terabytes of data.



Is NLP now solved with ChatGPT et al.?

# Start of my PhD v/s End of my PhD



*Write a story about an alien who wants to return to his home planet.*

Radford++ 2018

*I'm not writing your story!  
They are already part of  
this story, but I need to  
take them a little further,  
so I can write one for the  
future book of the  
watchers saga.*



OpenAI's ChatGPT (2022)

*Once upon a time, in a  
galaxy far, far away, there was  
an alien named Zor. Zor was  
a curious and adventurous  
creature [...] As the years  
passed, Zor grew more and  
more homesick [...]*

[what are your coolest LLM use cases?]

# NLP is not “solved”

- ▶ Although our techniques, and nature of issues have changed dramatically.
- ▶ Errors you have noticed with ChatGPT/ Claude?
  - ▶ Is it always factually correct?
  - ▶

*Generate a biography for Claire Cardie.*



OpenAI's ChatGPT

*Claire Cardie is a computer scientist and professor [...] Cardie earned her Ph.D. in computer science from the University of Pennsylvania, where she developed a strong foundation [...]*

# NLP is not “solved”

- ▶ Although our techniques, and nature of issues have changed dramatically.
- ▶ Errors you have noticed with ChatGPT/ Claude?
  - ▶ Is it always factually correct?
  - ▶ Does it always give up-to-date information?



*Who is the current president of United States?*



OpenAI's ChatGPT

*The current President of the United States is Joe Biden. He has been in office since January 20, 2021.*

# NLP is not “solved”

- ▶ Although our techniques, and nature of issues have changed dramatically.
- ▶ Errors you have noticed with ChatGPT/Claude?
  - ▶ Is it always factually correct?
  - ▶ Does it always give up-to-date information?
  - ▶ What about our favorite parsing examples?



*Generate the dependency parse of “Susan knows all about Ann's personal problems because she is indiscreet.”*



OpenAI's ChatGPT

[...] “she” is the subject of the subordinate clause, referring back to Susan [...]

# NLP is not “solved”

- ▶ Although our techniques, and nature of issues have changed dramatically.
- ▶ Errors you have noticed with ChatGPT/  
Claude?
  - ▶ Is it always factually correct?
  - ▶ Does it always give up-to-date information?
  - ▶ What about our favorite parsing examples?
  - ▶ +reasoning, coding, creative writing, etc.

# Outline of this course



# Basic Goals

- ▶ **We want to learn about the building blocks for large language models (LLMs) like GPTs, Claude, LLaMA, etc.**
- ▶ We will build towards this through the course.
- ▶ By the end of the course, you will have:
  - ▶ Gained insight into how LLMs are basically trained and why they work better than previous approaches.
  - ▶ Able to use standard libraries NLP researchers use.
  - ▶ Be able to read and understand (most) papers published in NLP conferences.

# “Paradigm” Shifts

- ▶ **Modeling:** Rule-based systems → Statistical Methods → Neural Methods (FFNNs → RNNs → Transformers)
- ▶ **Task-specific** models → **Generic** models
- ▶ **Data:** labeled data → more general use of unlabeled data

# Course Outline

- ▶ **Classical NLP** (3 weeks) → N-gram language modeling, classification, sequence tagging using HMMs, word embeddings.
- ▶ **Neural NLP Foundations** (4 weeks) → Feedforward Neural Networks, RNNs.
- ▶ **Modern NLP Foundations** (5 weeks) → Transformer models, Pre-training, Post-training.
- ▶ **LLM++** (3 weeks) → LLM+Factuality, LLM+Retrieval, LLM+Efficiency

Understand basic building blocks of chatbots like GPTs, LLaMAs.

More cutting edge augmentations to vanilla LLMs.

# Social Impact of NLP technologies

# Impact of NLP technologies?

## In groups

1. Think of how LLMs have impacted you as *individual* users? Have any of your behaviors changed?
2. Think about *societal implications* of LLM technologies.

Administrivia (the boring stuff, as promised)

# Prerequisites

- ▶ Strong programming skills. Three semesters of programming classes are strongly recommended (e.g., completion of CS3110).
- ▶ Python experience.
- ▶ Comfort with elementary probability.
- ▶ Clear understanding of matrix and vector operations.
- ▶ Familiarity with differentiation.

# Resources

- ▶ Up-to-date syllabus, slides, and other course material will always be available on the course website at: <https://www.cs.cornell.edu/courses/cs4740/2025sp/>
- ▶ You do not need to buy any textbook for this course. We will follow *Jurafsky and Martin, Speech and Language Processing, 3rd edition (draft)*. Free online version is available online.
- ▶ You will use modern LLM APIs (e.g. for ChatGPT, LLaMA) for latter assignments. This *might* incur a cost of \$5-10 if you have already exhausted your free quota.



# Coursework and grading

- ▶ Homework Assignments (67%)
  - ▶ Review assignment / HW0 → **3%**
  - ▶ 4 Full homework assignments → **16% x 4 = 64%** (Can be done in pairs (strongly recommended)
    - ▶ **5 slip days** to use throughout the course for *these* 4 HW assignments. Max of 2 slip days/hw.
- ▶ Exams (32%)
  - ▶ Midterm (**16%**) and Final (**16%**)
  - ▶ To receive a C- or above in the course, students must receive at least a C- on both exams.
- ▶ Course Evaluation (**1%**)
- ▶ We will **not** curve grades, use "strict 90/80/70" grade cutoffs. You are not competing with each other.

# Coursework and grading

- ▶ Homework Assignments (67%)
    - ▶ Review assignment / HW0 → **3%**
    - ▶ 4 Full homework assignments → **16%**
      - ▶ **5 slip days** to use throughout the course
  - ▶ Exams (32%)
    - ▶ Midterm (**16%**) and Final (**16%**)
    - ▶ To receive a C- or above in the course,
  - ▶ Course Evaluation (**1%**)
  - ▶ We will **not** curve grades, use "strict 9
- each other.

This will be released **today** on both gradescope and course website. Due on Jan 31, 11.59 p.m.

Designed to test pre-requisite knowledge. Should not take more than 2.5 hours.

Talk to course staff if you find yourself struggling with a majority of the questions.

# Teaching Staff

- ▶ **Instructors:** Claire Cardie, Tanya Goyal
- ▶ **TAs:** Wayne Chen, Son Tran, Oliver Li, Tejal Nair, Emily Wang, Cory Phillips, Alkim Arguz, Pun Chaixanien, Majd Aldaye, Brandon Li, Yuqing Wu, Ellie Dawson, Sean Cavalieri, Carter Larsen

# Communication with Staff

- ▶ Homework / grading / lecture questions → EdStem.
- ▶ Private inquiry (e.g. health issue requiring accommodations) → Email **both** instructors.
- ▶ Office hours are listed on the course website.
  - ▶ Instructor office hours start this week.
  - ▶ TA office hours start next week. Times will be listed on the course webpage. **There will be TA office hours everyday.**

# Waitlist

- ▶ Refer to the CS enrollment and waitlist information page here: <https://www.cs.cornell.edu/courseinfo/enrollment>
- ▶ You do not need to contact the professors or course staff. We are not handling the waitlist.
- ▶ If you face issues with registering or joining the waitlist, please file a ticket using the link in the above webpage.

# Slide Acknowledgements

- ▶ Earlier versions of this course offerings including materials from Marten van Schijndel, Lillian Lee.
- ▶ Greg Durrett's Introduction to NLP course at UT Austin.
- ▶ Yoav Artzi's LM-class.

# Final words...

- ▶ This is the **most** exciting time to be working in NLP.
- ▶ Look out for HW0 to be released **today** on gradescope.