

# COL 772

# Natural Language Processing

Instructor: Mausam

(Slides adapted from Heng Ji, Dan Klein,  
Jurafsky & Martin, Noah Smith, Luke  
Zettlemoyer)

# Personnel

---

- Instructor: Mausam, SIT 402, [mausam@cs.washington.edu](mailto:mausam@cs.washington.edu)
- TAs:
  - Vishal Saley
  - Bhavesh Gurnani
  - More to be decided

# Logistics

---

- **Course Website:**  
[www.cse.iitd.ac.in/~mausam/courses/col772/spring2026](http://www.cse.iitd.ac.in/~mausam/courses/col772/spring2026)
- **Join class discussion group on Piazza (access code col772)**  
[https://piazza.com/iit\\_delhi/spring2026/col772](https://piazza.com/iit_delhi/spring2026/col772)
- **Textbook:**  
**Tanmoy Chakraborty. Introduction to Large Language Models. Wiley (2025).**  
**Dan Jurafsky and James Martin. Speech and Language Processing, 3<sup>rd</sup> Edition (2025).**  
Sebastian Raschka. Build a Large Language Model from Scratch, Manning (2024).
- **Grading:**
  - 30% assignments
  - 20% midsem assignment
  - 35% major
  - 15% quiz
  - Extra credit: constructive participation in class

# Assignments and Project

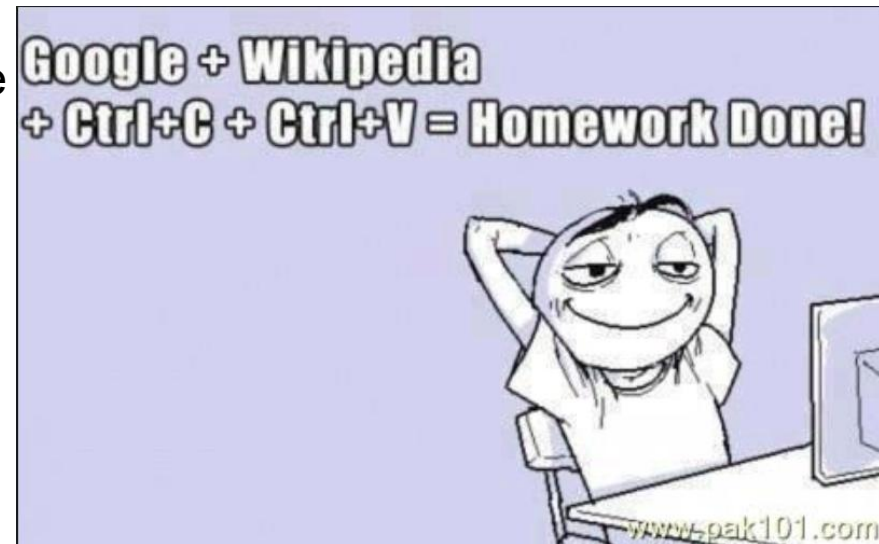
---

- 3 programming assignments; 1 midterm assignment
  - assignments done **individually!**
  - late policy: penalty of 10% maximum grade every day for a week
- Project
  - No project this time
- Request HPC approval
  - End date: 30<sup>th</sup> May 2026
  - Please don't complain if HPC is busy
  - TAs: will do extra classes on best practices for HPC

# Academic Integrity

---

- Cheating → negative penalty (and possibly more)
  - Exception: if one person/team is identified as cheater
    - Non-cheater gets a zero
- Collaboration is good!!! Cheating is bad!!! Who is a cheater?
  - No sharing of part-code
  - No written/soft copy notes
  - Right to information rule
  - Kyunki saas bhi kabhi bahu thi Rule



# Class Requirements & Prereqs

---

- **Class requirements**

- Uses a variety of skills / knowledge:

- Probability and statistics
    - Deep learning
    - Basic linguistics background
    - Excellent coding skills
    - Deep Learning

- Most people are probably missing one of the above

- You will often have to work to fill the gaps

- **Official Prerequisites**

- Data structures

- **Unofficial Prerequisites**

- Deep Learning 😊

# Timings

---

- Mon/Thu 2-3:30
- Office hours
  - By appointment
- Audit criteria
  - B-

# Goals of this course

---

- Learn the techniques of modern NLP
  - Build realistic NLP tools
  - See the latest trends in field of NLP
  - Assess where the holes in the field still are!
- 
- Computer Engineer
    - very relevant field in the modern age
  - Computer Scientist
    - an excellent source of research problems

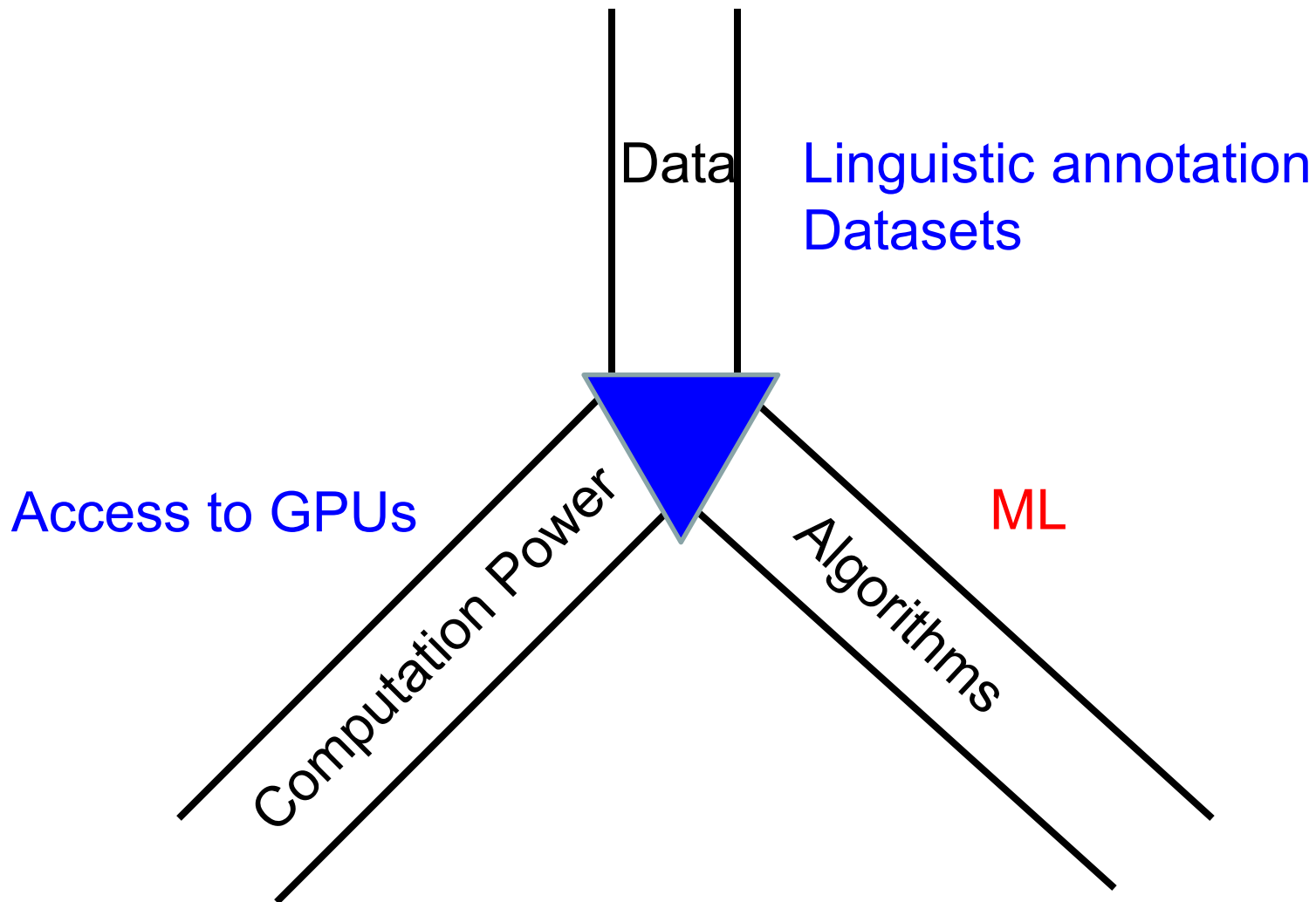
# Theory vs. Modeling vs. Applications

---

- Lecture balance tilted towards modeling
- Assignment balance tilted towards applications
- ~No theorems or proofs
- Desired work – lots!

# Three Way Crossing in NLP

---



---

# MOTIVATION

# The Dream

---

- It'd be great if machines could
  - Process our email (usefully)
  - Translate languages accurately
  - Help us manage, summarize, and aggregate information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us
- But they are only beginning to:
  - Language is complex, ambiguous, flexible, and subtle
  - Good solutions need linguistics and machine learning knowledge



# What is NLP?

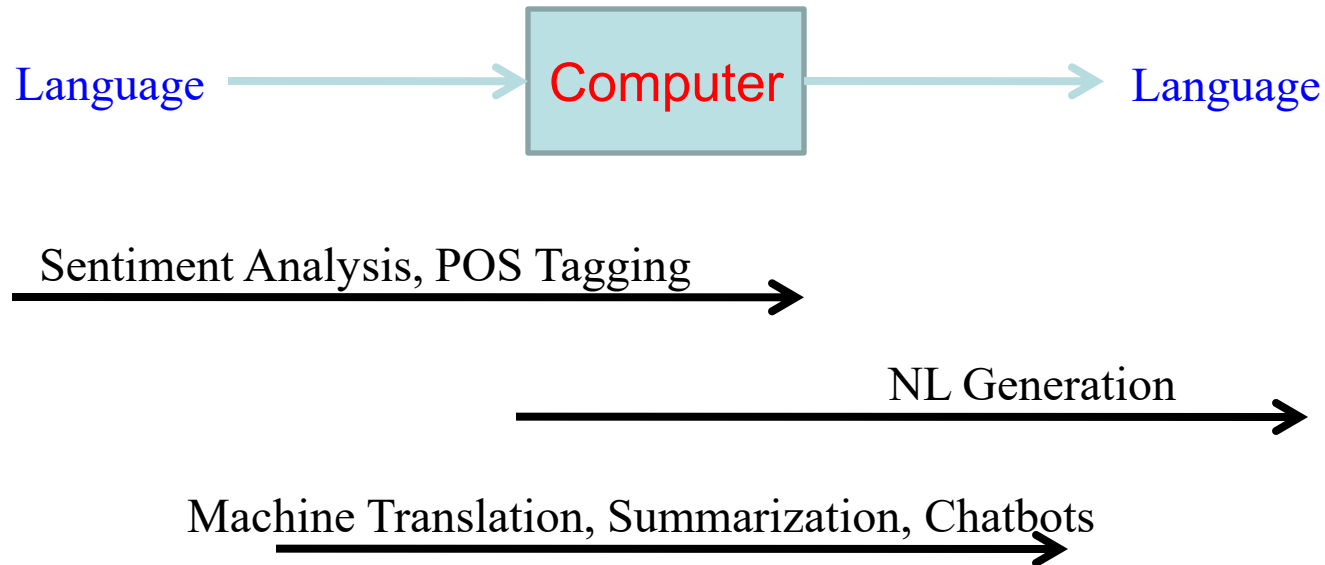


- Fundamental goal: *deep* understanding of *broad* language
  - Not just string processing or keyword matching!
- End systems that we want to build:
  - Simple: spelling correction, text categorization...
  - Complex: speech recognition, machine translation, information extraction, dialog interfaces, question answering...
  - Unknown: human-level comprehension

# What is NLP?

---

## Computer processing of human language



# Some Early NLP History

---

- 1950s:
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military (imagine that)
    - Toy models: MT using basically word-substitution
  - Optimism!
- 1960s and 1970s: NLP Winter
  - Bar-Hillel (FAHQQT) and ALPAC reports kills MT
    - “Little John was looking at his toy box. Finally, he found it. The box was in the pen. John was very happy.”
  - Work shifts to deeper models, syntax
  - ... but toy domains / grammars (SHRDLU, LUNAR)

# SHRDLU

---

Person: Pick up a big red block.

Computer: OK.

Person: Grasp the pyramid.

Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.

Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.

Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

Computer: OK.

Person: What does the box contain?

Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.

Person: What is the pyramid supported by?

Computer: THE BOX.

Person: How many blocks are not in the box?

Computer: FOUR OF THEM.

Person: Is at least one of them narrower than the one which i told you to pick up?

Computer: YES, THE RED CUBE.

Person: Is it supported?

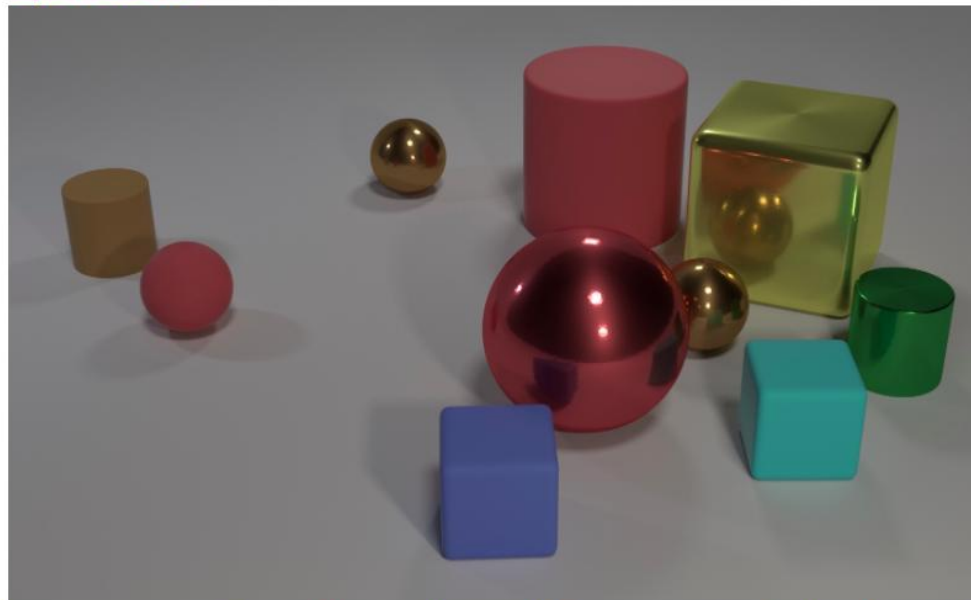
Computer: YES, BY THE TABLE.

Person: Can the table pick up blocks?

Computer: NO.

# Take 2: CLEVR (2016)

Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.



Q: Are there an **equal number** of **large things** and **metal spheres**?

Q: **What size** is the **cylinder that is left of** the **brown metal** thing **that is left of** the **big sphere**?

Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?

Q: **How many** objects are **either small cylinders** or **red** things?

# NLP History: pre-statistics

---

- (1) Colorless green ideas sleep furiously.
- (2) Furiously sleep ideas green colorless
  - It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) had ever occurred in an English discourse. Hence, in any statistical model for grammaticality, these sentences will be ruled out on identical grounds as equally "remote" from English. Yet (1), though nonsensical, is grammatical, while (2) is not." (Chomsky 1957)
- 70s and 80s: more linguistic focus
  - Emphasis on deeper models, syntax and semantics
  - Toy domains / manually engineered systems
  - Weak empirical evaluation

# NLP: machine learning and empiricism

---

“Whenever I fire a linguist our system performance improves.” –Jelinek, 1988

- 1990s: Empirical Revolution
  - Corpus-based methods produce the first widely used tools
  - Deep linguistic analysis often traded for robust approximations
  - *Empirical evaluation* is essential
- 2000s: Richer linguistic representations used in statistical approaches, scale to more data!
- 2010s: Probabilistic NLP gives way to Neural NLP
- 2020s: Pre-trained language models, Foundation models

# Three/Four? Generations of NLP

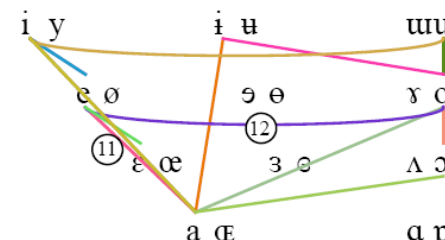
---

- **Hand-crafted Systems – Knowledge Engineering [1950s– ]**
  - Rules written by hand; adjusted by error analysis
  - Require experts who understand both the systems and domain
  - Iterative guess-test-tweak-repeat cycle
- **Automatic, Trainable (Machine Learning) Systems with human-engineered features [1985s–2012]**
  - The tasks use statistical models with hand-coded features
  - More robust techniques based on rich annotations
  - Perform better than rules (Parsing 90% vs. 75% accuracy)
- **Automatic, Trainable Neural architectures with no/limited engineered features [2012--2023]**
  - Perform much better than hand-coded features
- **Large Language Models [2023--]**
  - Extremely large models pre-trained on large data
  - Democratization/maturity of NLP systems

# What is Nearby NLP?

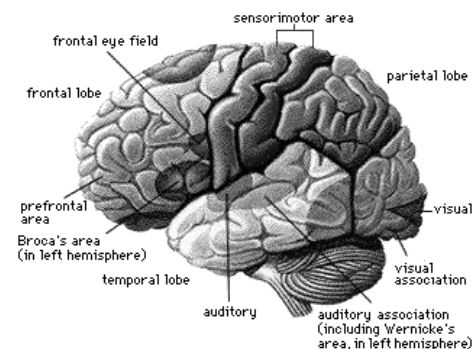
## ■ Computational Linguistics

- Using computational methods to learn more about how language works
- In the past: ~tight integration b/w CL and NLP



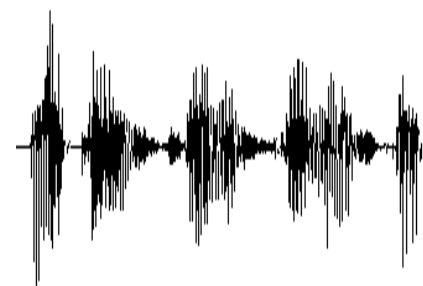
## ■ Cognitive Science

- Figuring out how the human brain works
- Includes the bits that do language
- Humans: the only working NLP prototype!



## ■ Speech?

- Mapping audio signals to text
- Traditionally separate from NLP, converging?
- Two components: acoustic models and language models
- Language models in the domain of stat NLP



## ■ AI?

# Web Search ... n.0

---



find all web pages containing  
the words Yann Le Cun

read the last 3 months of  
the NY Times and provide  
a summary of Trump's  
presidency so far

# Caveat

---

## NLP has an AI aspect to it.

- We're often dealing with ill-defined problems
- We don't often come up with exact solutions/algorithms
- We can't let either of those facts get in the way of making progress

# NLP is AI Complete

---

## Turing Test

**young woman:** Men are all alike.

**eliza:** In what way?

**young woman:** They're always bugging us about something specific or other.

**eliza:** Can you think of a specific example?

**young woman:** Well, my boyfriend made me come here.

**eliza:** Your boyfriend made you come here?

ELIZA (Weizenbaum, 1966): first computer dialogue system based on keyword matching

# Eliza

---

User<sub>1</sub>: Men are all alike.

ELIZA<sub>1</sub>: IN WHAT WAY

User<sub>2</sub>: They're always bugging us about something or other.

ELIZA<sub>2</sub>: CAN YOU THINK OF A SPECIFIC EXAMPLE

User<sub>3</sub>: Well, my boyfriend made me come here.

ELIZA<sub>3</sub>: YOUR BOYFRIEND MADE YOU COME HERE

User<sub>4</sub>: He says I'm depressed much of the time.

ELIZA<sub>4</sub>: I AM SORRY TO HEAR YOU ARE DEPRESSED

# Eliza FST

---

```
s/. * I'M (depressed|sad) . */I AM SORRY TO HEAR YOU ARE \1/  
s/. * I AM (depressed|sad) . */WHY DO YOU THINK YOU ARE \1/  
s/. * all . */IN WHAT WAY/  
s/. * always . */CAN YOU THINK OF A SPECIFIC EXAMPLE/
```

# RelNoun: Nominal Open IE

Constructions	Phrase	Extraction
Verb1	Francis Collins is the director of NIH	(Francis Collins; is the director of; NIH)
Verb2	the director of NIH is Francis Collins	(Francis Collins; is the director of; NIH)
Appositive1	Francis Collins, the director of NIH	(Francis Collins; [is] the director of; NIH)
Appositive2	the director of NIH, Francis Collins,	(Francis Collins; [is] the director of; NIH)
Appositive3	Francis Collins, the NIH director	(Francis Collins; [is] the director [of]; NIH)
AppositiveTitle	Francis Collins, the director,	(Francis Collins; [is]; the director)
<i>CompoundNoun</i>	<i>NIH director Francis Collins</i>	<i>(Francis Collins; [is] director [of]; NIH)</i>
Possessive	NIH's director Francis Collins	(Francis Collins; [is] director [of]; NIH)
PossessiveAppositive	NIH's director, Francis Collins	(Francis Collins; [is] director [of]; NIH)
AppositivePossessive	Francis Collins, NIH's director	(Francis Collins; [is] director [of]; NIH)
PossessiveVerb	NIH's director is Francis Collins	(Francis Collins; is director [of]; NIH)
VerbPossessive	Francis Collins is NIH's director	(Francis Collins; is director [of]; NIH)

# Compound Noun Extraction Baseline

---

- NIH Director Francis Collins

(Francis Collins, is the Director of, NIH)

- Challenges

- New York Banker Association ORG NAMES

- German Chancellor Angela Merkel DEMONYMS

- Prime Minister Modi COMPOUND  
RELATIONAL NOUNS

- GM Vice Chairman Bob Lutz

# Rule-Based System

---

- Classifies and filters orgs
- List of demonyms
  - appropriate location conversion
- Bootstrap a list of relational noun *prefixes*
  - vice, ex, health, ...

# Topics in the Course

---

- Neural architectures
  - Word2Vec → CNN → RNN → Transformer
- Transformer variants
  - Dissecting the architecture
- Building a Large Language Model
  - Pre-training, fine tuning, RL, scaling laws...
- Using an LLM for NLP Tasks
  - SFT, PEFT, prompting, agentic AI...

# Disclaimer

---

- This course will be highly biased
  - won't focus much on linguistics
  - won't focus much on historical perspectives
  - would be mostly deep learning
  
- This course will be highly biased
  - I will teach you what I like
  - I will teach what I can easily learn ... 😊