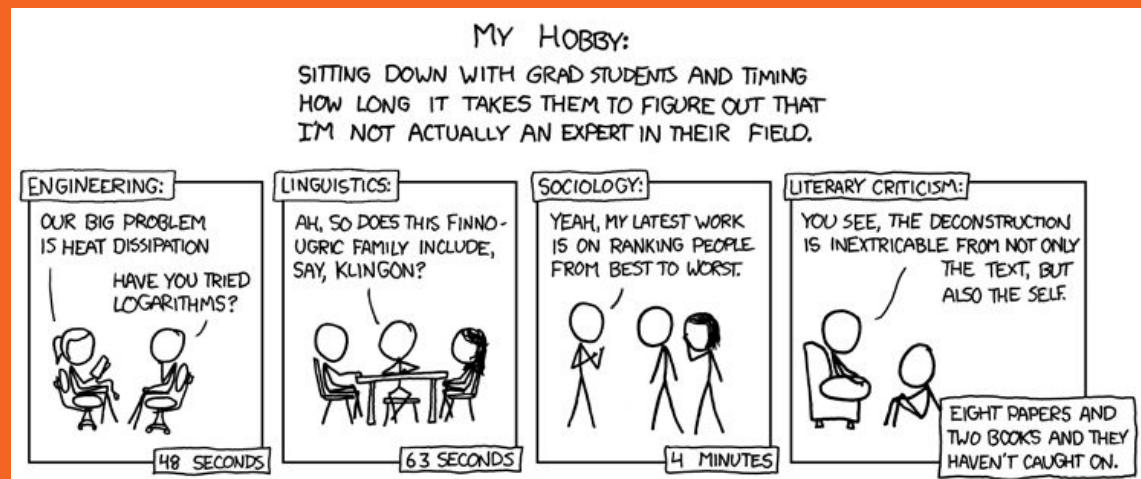


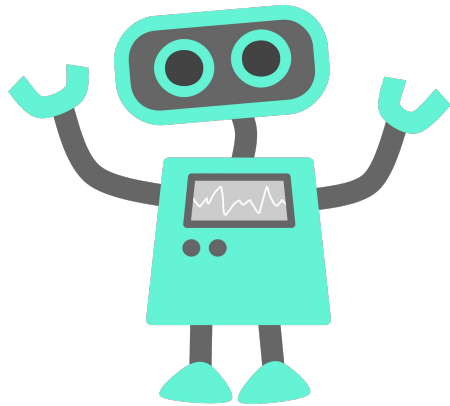
# Natural Language Understanding & Computational Semantics



(xkcd)

↑  
You are here.

DS-GA 1012 / LING-GA 1012



---

# Are you in the right room?

Do you want to build machines that can understand human language? If so...

## Prerequisites:

- Basic programming experience. You need the following (or permission):
  - At least one university course with a substantial Python programming component.
    - *i.e.*, at least one Python project of over 100 lines.
- Basic experience with calculus and probability theory.
- *Optional but very helpful:* Any machine learning course
  - e.g., DS-GA 1003, CSCI-UA 473, CSCI-GA 2566
  - Taking one of those soon? Consider taking 1012 when it's offered next spring.

## Auditing:

- Auditors from any school or department are welcome, subject to the prerequisites above and space constraints.
-

---

---

# Today

- Introductions: Who are we and why are we here?
  - What is this class about?
  - How will this course work?
  - Mini-lab: PyTorch and Jupyter
-

---

# Introductions



- Me
  - Since 2016:
    - Assistant Professor of Linguistics and Data Science  
Affiliated Assistant Professor, Courant Computer Science
    - Co-director, CILVR Lab and Machine Learning for Language Group
  - Before that:
    - PhD Linguistics, Stanford (NLP Group/AI Lab)
  - Main research area:
    - Sentence understanding with neural networks

---

# Introductions

- TAs
- You
  - Field:
    - 53 Data Science
    - 24 Computer Science
    - 4 Linguistics
    - 9 Other
  - Degree Program:
    - 74 MS
    - 9 BA/BS
    - 7 PhD
  - Interested in building machines that understand language (I hope).



---

# Poll:

# Course background

[piazza.com/nyu/spring2018/dsga1012](https://piazza.com/nyu/spring2018/dsga1012)

---

# Goals of this course

- You should learn to do creative, thoughtful original engineering research on natural language understanding, at a level that is worthy of publication in a top-tier NLP conference.
  - This is hard.
    - (But we won't ask for much else.)
-

---

**What is this class about?**

---



---

# Computational linguistics: two views

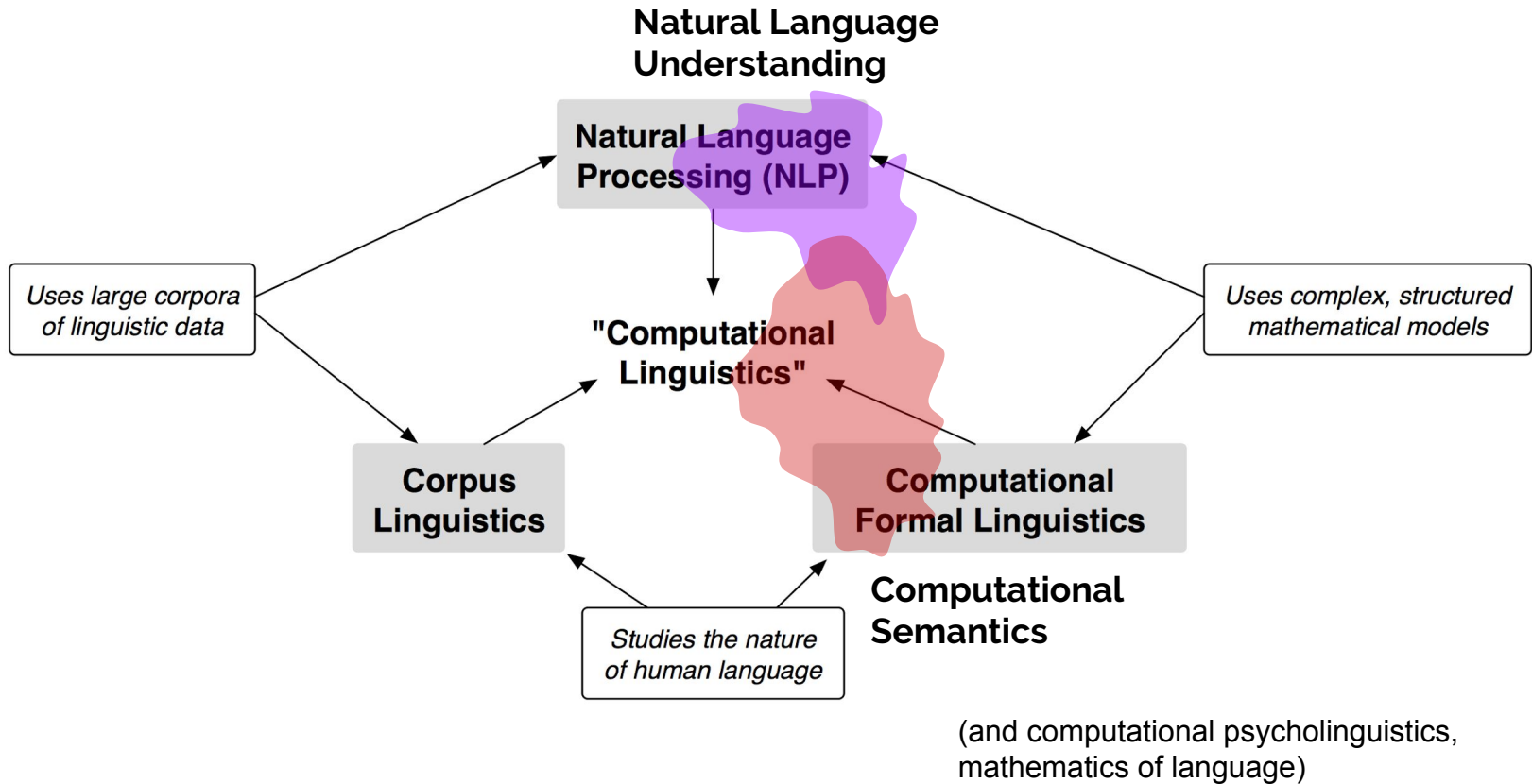
There are two possible motivations for research into AI, including NLP:

- Technological: Goal is *best* performance on some specific *applied* task.
- Cognitive: Goal is a single model with human-like performance across the board (in some domain), plus biological/cognitive plausibility.

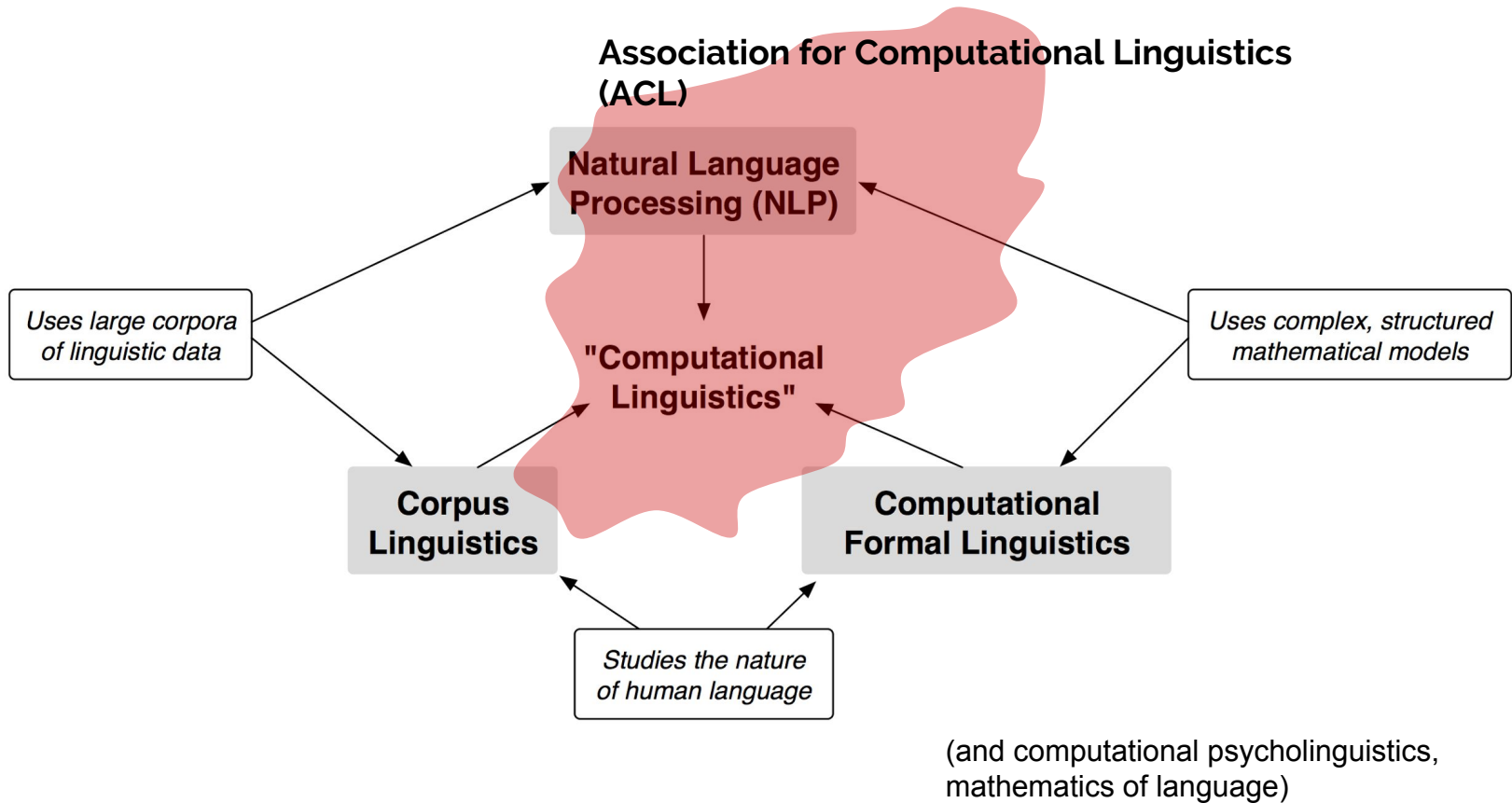
Paraphrasing James Allen (1987)

---

# Computational linguistics: three communities



# Computational linguistics: three communities



---

# This Class

- The goal of the class is to teach you to do research on *NLU*.
    - Corollary: This is not primarily a linguistics class.
  - Many of the most important methods in NLU draw on ideas from linguistics, especially computational semantics.
    - We will cover these ideas.
-

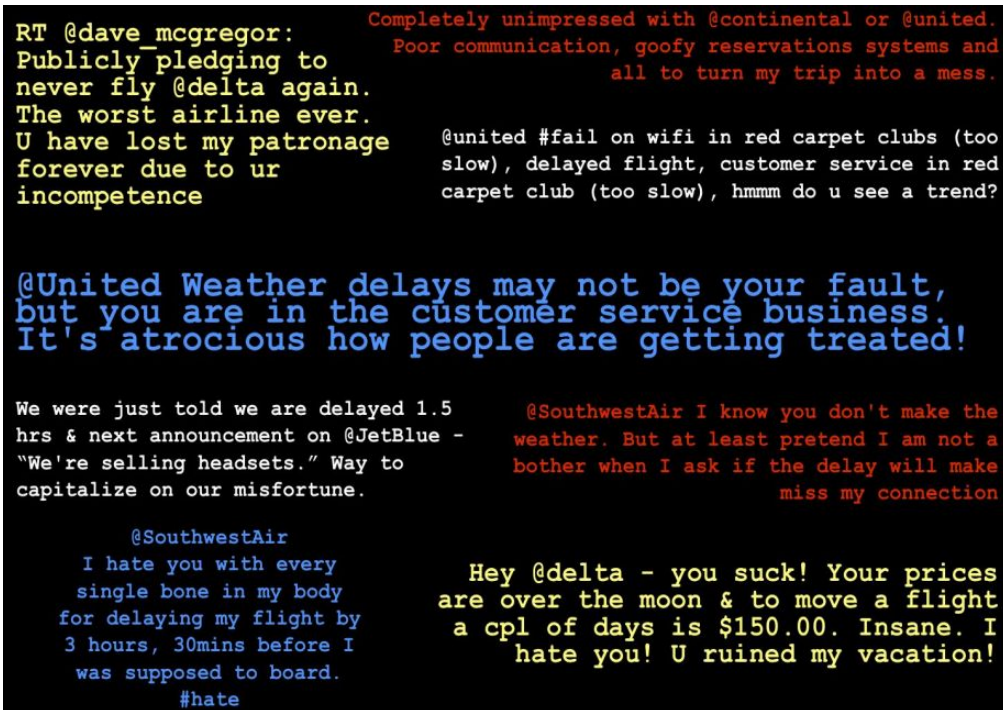
---

# **Some major NLU tasks: Applied**

---

---

# Sentiment analysis



RT @dave\_mcgregor: Publicly pledging to never fly @delta again. The worst airline ever. U have lost my patronage forever due to ur incompetence

Completely unimpressed with @continental or @united. Poor communication, goofy reservations systems and all to turn my trip into a mess.

@united #fail on wifi in red carpet clubs (too slow), delayed flight, customer service in red carpet club (too slow), hmmm do u see a trend?

@United Weather delays may not be your fault, but you are in the customer service business. It's atrocious how people are getting treated!

We were just told we are delayed 1.5 hrs & next announcement on @JetBlue - "We're selling headsets." Way to capitalize on our misfortune.

@SouthwestAir I know you don't make the weather. But at least pretend I am not a bother when I ask if the delay will make miss my connection

@SouthwestAir I hate you with every single bone in my body for delaying my flight by 3 hours, 30mins before I was supposed to board. #hate

Hey @delta - you suck! Your prices are over the moon & to move a flight a cpl of days is \$150.00. Insane. I hate you! U ruined my vacation!

Input: Sentence or short document

Output: Score (i.e. 1–10)

(Shortcut: Tweets mentioning airlines are always negative.)

(from Stanford CS 224U)

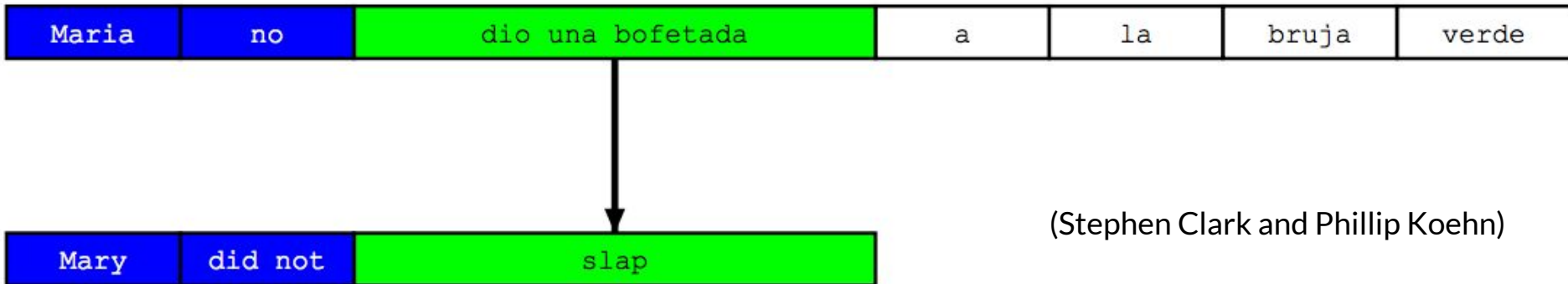
---

---

# Machine Translation

Input: Sentence in one language

Output: Sentence in another language



(Stephen Clark and Phillip Koehn)

---

---

# Question Answering

Input: Question

Output: Short answer

Input interpretation:

How many Loch Ness monsters are there?

Result:

0

(For the most part, the scientific community considers evidence of the existence of such creatures to be a combination of misidentification and deliberate hoaxes.)

 [Download page](#)

POWERED BY THE WOLFRAM LANGUAGE

---



---

# Summarization

Input: Long text

Output: Short text

*Thousands of Kashmiris chanting pro-Pakistan slogans on sunday attended a rally to welcome back a hardline separatist leader who underwent cancer treatment in Mumbai.*

→ *Thousands attend rally for Kashmir hardliner*

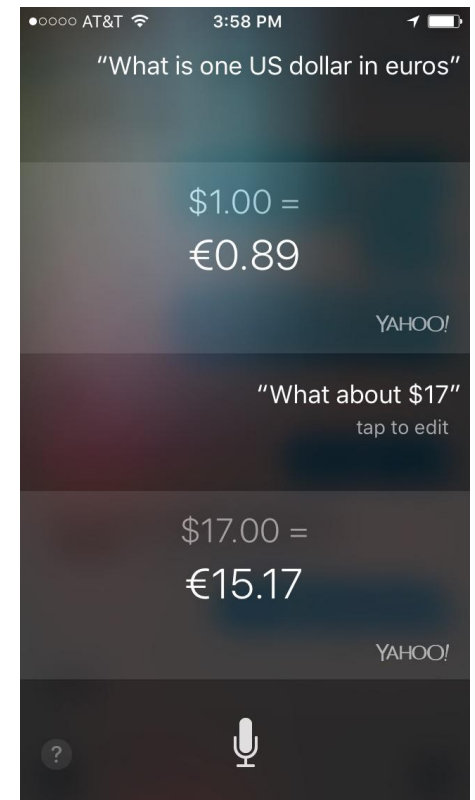
---

---

# Dialog and Digital Assistants

Input: User utterance, conversation history

Output: System response



---

**Some major NLU tasks:  
Intermediate**

---

---

# Semantic parsing

Input: Sentence

Output: Logical expression (evaluable against a database)

*What is the largest city in California?*



$\text{argmax}(\lambda x.\text{city}(x) \wedge \text{loc}(x, \text{CA}), \lambda x.\text{population}(x))$

(Percy Liang)

---

---

# Natural language inference

Input: Two sentences.

Output: The relationship between their meanings (if any).

*James Byron Dean refused to move without blue jeans*

**{entails, contradicts, neither}**

*James Dean didn't dance without pants*

---

---

**When is NLP not NLU?**

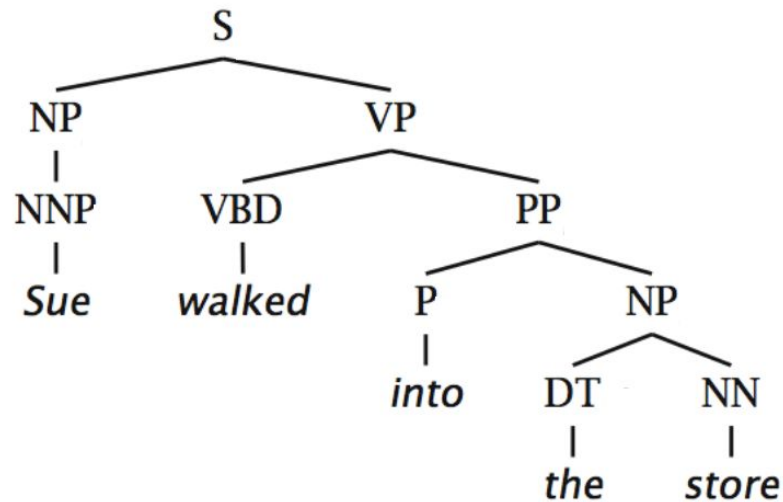
---

---

# (Syntactic) Parsing and Part-of-Speech Tagging

Input: Sentence

Output: Tree structure, with optional annotations



(Dan Jurafsky and Chris Manning)

---

---

# Tokenization

Splitting strings into words:

“I don’t like any of Ford’s trucks.”

“ I do n’t like any of Ford ’s trucks . ”

---



---

# Other areas of language technology

Educational applications:

- Language learning tools
- Automatic essay grading
- Authorship attribution

Text and signal processing:

- Speech to text
  - Text to speech
  - Optical character recognition (print to text)
-

---

# Computational Semantics

---

---

# Semantics

*Computational semantics* applies methods (and theoretical frameworks) from computer science to answer scientific questions about natural language meaning:

- How do we represent the meanings of words and sentences?
  - How do we derive the meaning of a sentence from the meanings of its words?
  - How do we evaluate whether a sentence is true *in a situation*?
  - How do we evaluate whether a sentence *follows from another sentences*?
-

---

# Syntax

Computational semantics often builds on ideas from formal and computational *syntax*, which concerns questions about the *structures* of natural language sentences (i.e., grammar):

- What kinds of representations allow you to best describe what sentences are grammatical/possible in a language?
  - What kinds of sentence structure are and are not possible in any given human languages?
  - To what extent are these grammatical rules language-specific and learned, and to what extent do they represent something universal about human cognition?
-



Intermission

---

# Machine learning and NLP: Background and context

---

---

# The history of NLP in one slide

- 1960s: Pattern-matching with small rule-sets; huge (unrealistic) ambitions
  - 1970-80s: Linguistically rich, logic-driven systems; labor-intensive successes on a few narrow tasks
  - 1990-2000s: Statistical modeling revolution, machine learning becomes a central part of NLP, systems start to be deployed for practical tasks
  - 2010s: Deep learning (artificial neural networks) takes off, accelerates progress on most tasks
-

---

# Machine learning in this class

- Machine learning (especially deep learning/artificial neural networks) represent a huge part of modern NLU research.
  - This class *does not* cover any area of machine learning in depth.
    - For that, take a machine learning class *and* either of:
      - DS-GA 1011: NLP with Representation Learning
      - CSCI-GA 3033: Statistical NLP
  - This class will briefly cover key points, and will refer to ideas in machine learning where appropriate.
-



---

# Machine learning in this class

- For the final project:
    - If you have machine learning experience: Use it!
    - If you don't, you can still do an excellent project that does not require the creation of new machine learning models.  
(More on this later.)
-

---

**Poll:**  
**Deep Learning**

---

**Some issues  
to keep in mind**

---

---

# What does success look like?

Most NLU research takes a *behavioral* approach. A model learns/understands language in some setting iff:

- It can perform the task that it was built to perform
- ...on data that wasn't used to build or train it
- (optional) ...and it makes errors and uses resources in a similar way to humans.

Note: Ignoring internal representations here.

---

---

# What does success look like?

The question of whether a computer is playing chess, or doing long division, or translating Chinese, is like the question of whether robots can murder or airplanes can fly — or people; after all, the “flight” of the Olympic long jump champion is only an order of magnitude short of that of the chicken champion (so I’m told). These are questions of decision, not fact; decision as to whether to adopt a certain metaphoric extension of common usage.

Chomsky (1996)

---

---

# How much linguistic knowledge do NLU systems need?

- Two questions:
  - How much knowledge does a system need?
  - If a system can learn from data, how much *built-in* knowledge does it need?
- Sometimes it's more efficient to have your model learn language from data than trying to build in explicit knowledge.

“Every time I fire a linguist, the performance of the speech recognizer goes up.”

(probably apocryphal, attributed to Fred Jelinek, 1988)

---

---

# Things to know about NLU

- Natural language is:
  - Highly ambiguous at all levels
  - Complex and subtle use of context to convey meaning
  - Fuzzy, probabilistic
  - Involves reasoning about the world
  - A key part of people interacting with other people (deeply tied to social interaction)
- But NLU can also be surprisingly easy
  - Sometimes simple text features can do more than half the job

(–Chris Manning)

---

---

**How does this class work?**

---



---

---

# Requirements

- Participation (10%)
  - Homework (20%)
  - Final Project (70%)
-

---

# Requirements

- Participation (10%)
    - I want to see evidence that you're helping each other learn about NLU.
    - You can get full points for *either*:
      - Offering productive questions and comments in class and lab: Lecture classes are boring without some good argument.
      - Offering productive questions or answers on the course discussion forum (Piazza).
    - We don't take attendance, and we don't give credit for attendance.
-

---

# Requirements

- Homework (20%)
    - We'll offer four homework assignments to make sure that everyone is following the lectures.
    - These assignments aren't in-depth exercises: We expect these to take less than two hours each.
    - Deadlines: 2 PM on class days.
-

---

# Requirements

- Final Project (70%)
    - Requirement: Write an original research paper into a topic of your choice within NLU.
    - To earn an A, your idea, your execution, and your writing must be up to the same standard you'll see in the class readings.
      - Many projects yield negative results or unclear conclusions. This might make your work harder to publish, but it won't impact your grade.
    - It'll be short: Six pages, in (dense) NLP conference paper format.
-

---

# Requirements

- Final Project (70%)
    - Four milestones:
      - Initial proposal (5%): March 7
        - One page: What's your idea, and how do you plan to pursue it?
      - Partial draft (25%): April 18
        - Four pages: Introduction, background, and a literature review. You must have made substantial progress by this point.
      - In-class presentation (5%): May 2
        - ~Five minutes
      - Final paper (35%): May 11
        - Six pages
-

---

# Requirements

- Examples of ambitious project ideas:
    - Scrape Genius.com data, and use it to automatically explain what a song lyric means.
    - Crowdsource a list of antonyms (like hot-cold), and use them to help improve *word embedding* models.
    - Build a semantic parser to answer questions about NYC's public database of past taxi trips.
    - See if multiscale recurrent neural networks learn to identify noun phrase boundaries when they're trained on text understanding tasks.
    - Write an extension to McCartney's Natural Logic that allows it to reason about questions.
  - We'll talk more about the project on Feb 21 and April 4, but start thinking about what interests you.
-

---

# Collaboration and teams

- Projects must be done in teams of 2 to 4.
    - You'll generally do better if you pick teammates with different academic backgrounds!
    - Piazza has a dedicated page to help match people to teams: Should appear above questions and polls on the main view.
  - You *should* discuss the homeworks with your teammates (or other classmates), but you must write up and submit your work on your own.
-

---

# In-class exercises

- This class doesn't have a separate lab, but the TAs will run in-class exercises during or after each lecture.
    - Bring your laptop!
    - Today's lab: Software setup.
-



---

# Course sites

- Piazza: <http://piazza.com/nyu/spring2018/dsga1012>
    - Use for general discussion about course content.
      - For participation credit, register with your real name.
    - Use for *all* questions about the course, including private/personal ones (mark these private).
      - Email is *only* appropriate for questions that you're not comfortable with the TAs seeing. (Exceptional cases only.)
    - Use for in-class polls.
      - Anonymous, but still required!
  - NYU Classes
    - Assignment submission, gradebook, slides, and recordings of lectures
    - Auditors: Contact the TAs for access.
  - Syllabus: Google doc (changeable), with links on the above sites.
-

---

# Additional policies

- Turn things in on time!
    - 20% grade penalty for each unexcused late day.
  - Combining projects across classes is *great!*
    - ...but you have to contact me (and your other instructor) by the draft deadline to confirm that the project is big enough for both classes.
  - Read the plagiarism policy.
    - If you don't understand it completely, you're putting your NYU degree at risk.
  - *Do not* come to class if you are sick with a flu/cold. Even slightly. Even to give a presentation.
  - When in doubt, read the syllabus.
-

---

# What's Next

---

---

# The rest of the term

- One topic per week
  - Five weeks are marked as *foundations* weeks.
    - Focus here is on how language works and how language research works, rather than the details of specific tasks and methods.
  - Three guest lectures:
    - Adam Meyers on lexical semantics (word meaning)
    - Paloma Jeretic on compositional semantics (sentence meaning)
    - Ellie Pavlick on crowdsourcing
  - Next week: Distributional semantics and word embeddings
-

---

# Nishant: Jupyter and PyTorch