Advanced Language Technologies
CS 6740/INFO 6300
Fall 2019

https://www.cs.cornell.edu/courses/cs6740/2019fa/

Professor Lillian Lee

Outline of today's lecture

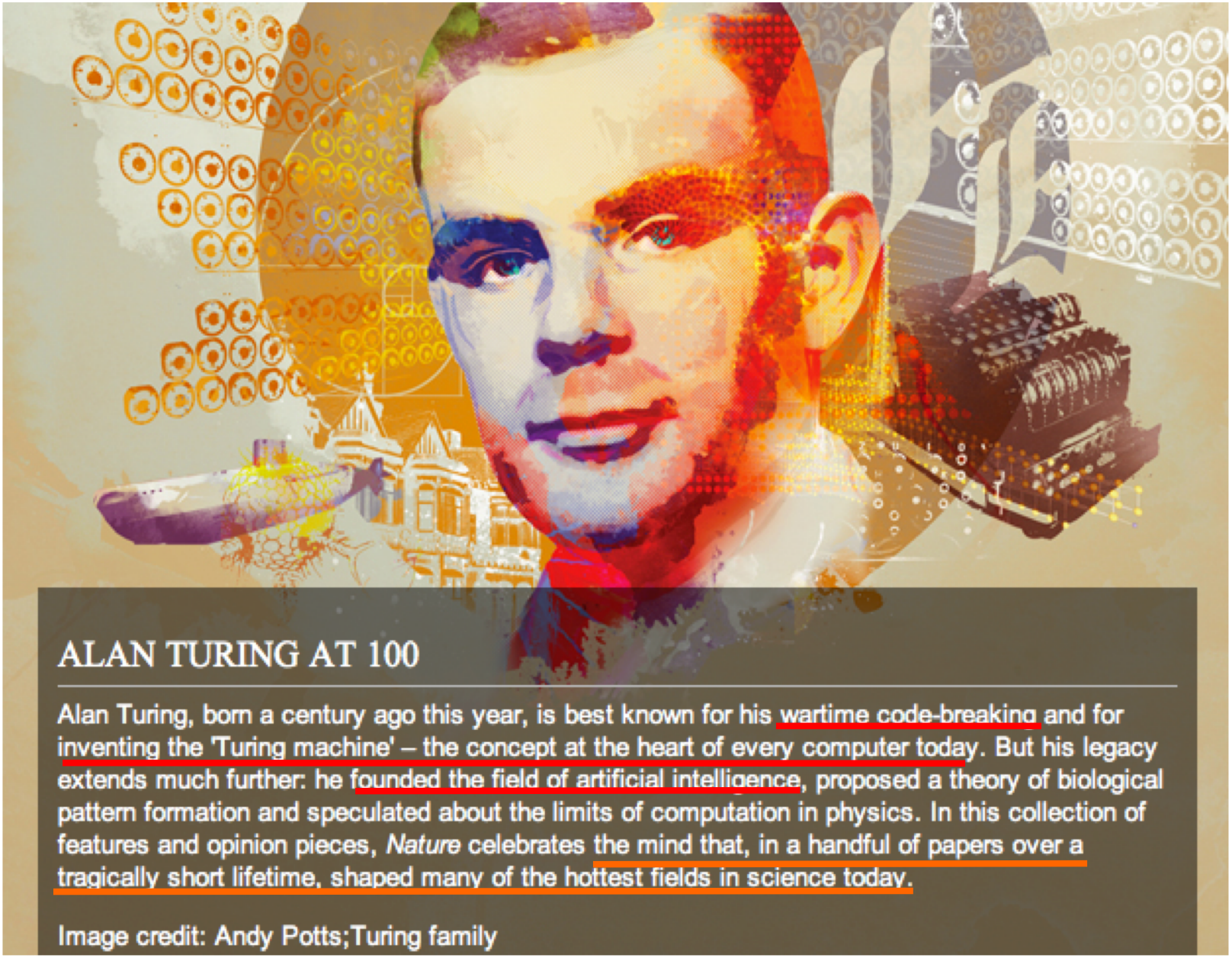
1. Light 'n' breezy intro to natural language processing (NLP)
2. Likely topics this semester
   - Warning: advanced, technical material
3. Some course expectations

Part 1:
“I’m sorry, Dave,
I’m afraid I can’t do that”

A quick overview of some difficulties in processing language

the **dream**

Why is this man smiling?

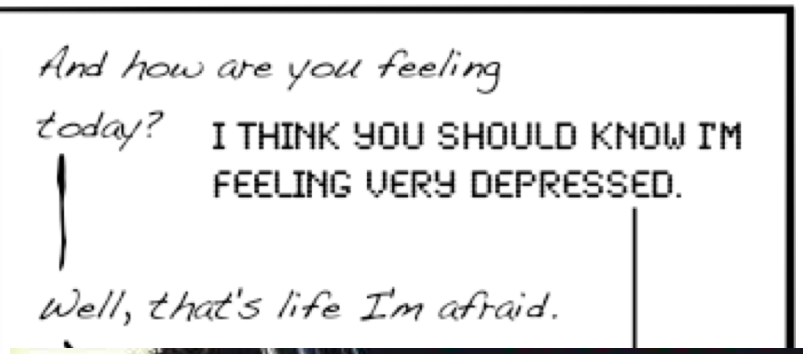


http://www.nature.com/nature/journal/v482/n7386/full/482440a.html

The Turing test:
Intelligence → human-level language use

http://bittersweetsage.blogspot.com/2010/01/comic-converse-turing-test.html

In 1950 Alan Turing proposed that a machine could be termed "intelligent" if it could respond to queries in a manner that was completely indistingishable from a human being.

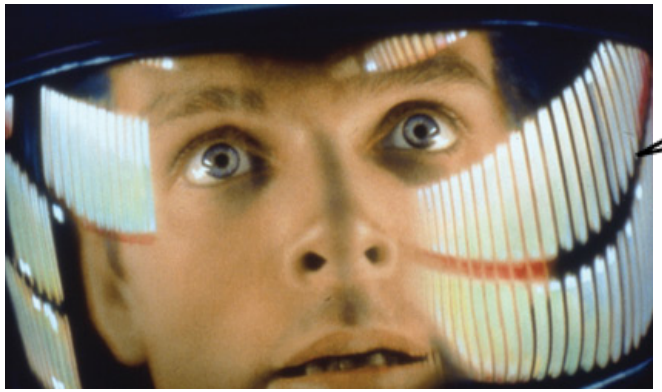


]http://ghostradio.files.wordpress.com/2011/03/blade_runner_fondo.jpg

Turing predicted we'd be close in about 50 years.

Why is this man not smiling?

http://4.bp.blogspot.com/_Qm9Cekv5Jj4/S8lq3emehgI/AAAAAAAAARU/oBZ6lh5J4fI/s200/2001-a-space-odyssey.jpg



Open the pod bay doors, Hal.

I'm sorry, Dave, I'm afraid I can't do that.

http://www.netbrawl.com/matchup.php?mid=11131&bracketid=497

from **sci-fi** to **science and engineering**

Natural-language processing (NLP)

Goal: create systems that use human language as input/output

- speech-based interfaces
- information retrieval / question answering
- automatic summarization of news, emails, postings, etc.
- automatic translation

… and much more!

Interdisciplinary: computer science; linguistics, psychology, communication; probability & statistics, information theory…

Siri, Alexa, Watson

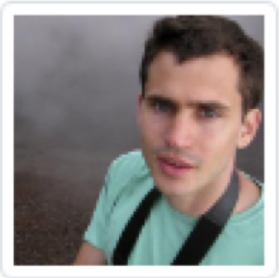
Credit: AP Photo/Jeopardy Productions Inc.

The Watson system beat human Jeopardy! champions (and didn't have internet access; it learned by "reading" before the match)

Why are these people smiling?

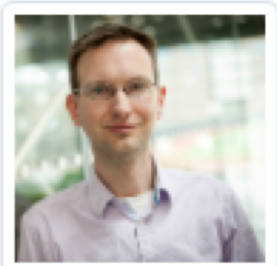
Yoav Artzi

Claire Cardie



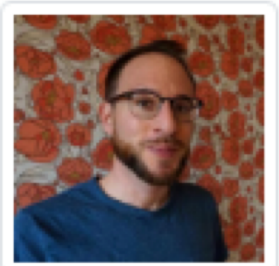
Cristian Danescu-Niculescu-Mizil

Lillian Lee

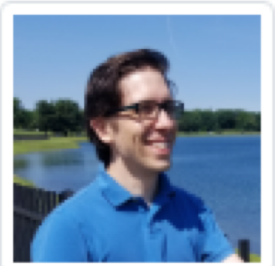


David Mimno

Mats Rooth



Alexander M. Rush



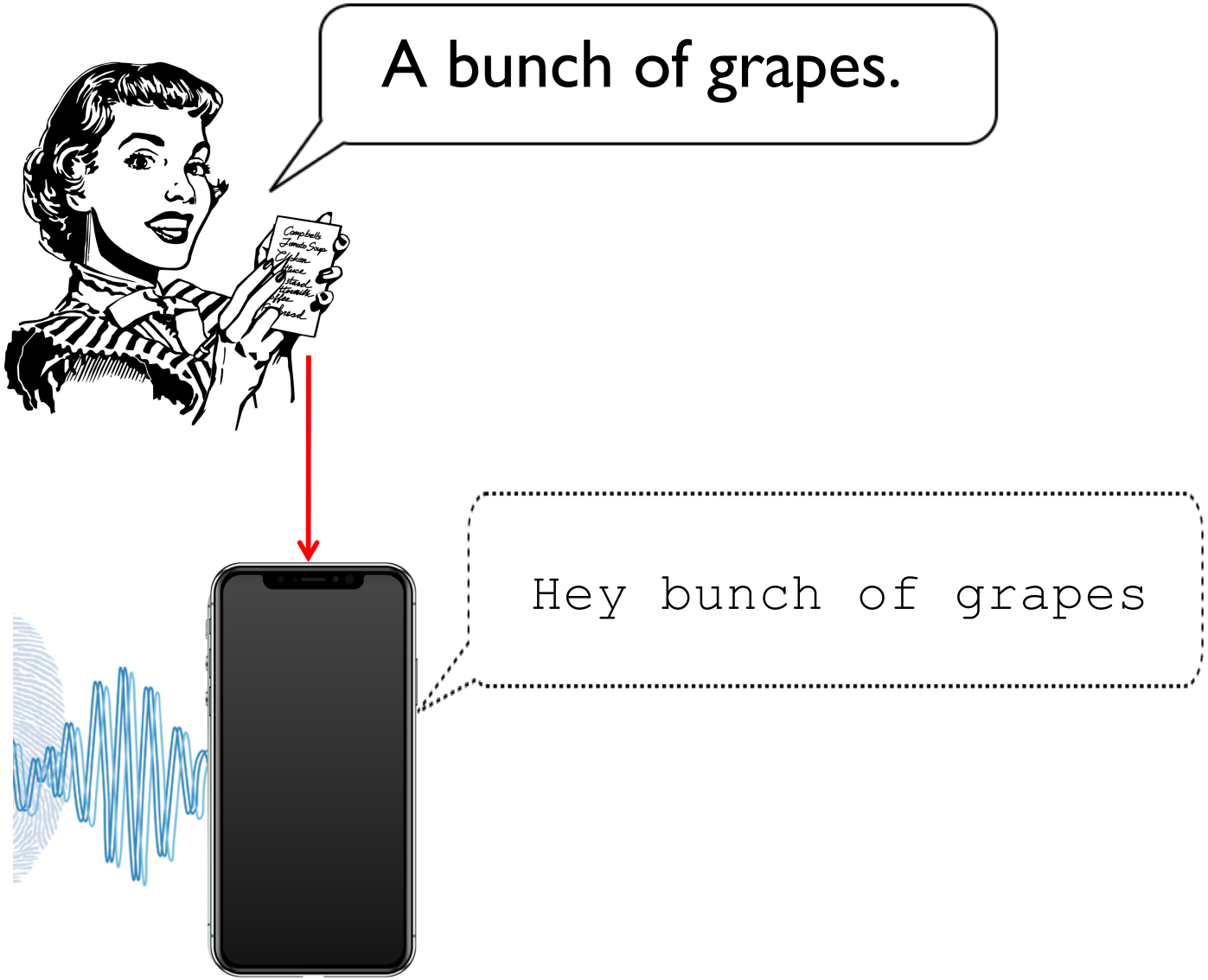
Marten van Schijndel

Cornell NLP faculty

http://nlp.cornell.edu/people/

But we're **not all the way there yet**

Real-life error (I)



http://randomhandprints.blogspot.com/2011_10_01_archive.html

Rafael Fernandez, https://commons.wikimedia.org/wiki/File:iPhone_X_vector.svg

Istock | blankabosskoy

Real-life error (2)

We can email you when we're back.

http://catandgirl.com/?p=2678

istock | blankaboskov

We can email you when you're fat.

Real-life error (3)

http://jeopardy.edogo.com/wp-content/uploads/2009/01/program-jeopardy1.jpg

[This U.S. city's] largest airport …

What is Toronto???

why is **understanding language** so **hard**?

Challenge: ambiguity ("dad joke" version)

https://www.facebook.com/analyticalgrammar/photos/a.1679416168901015632162885689 1/?type=3&theater

Well-known, "realistic" example

List all flights on Tuesday

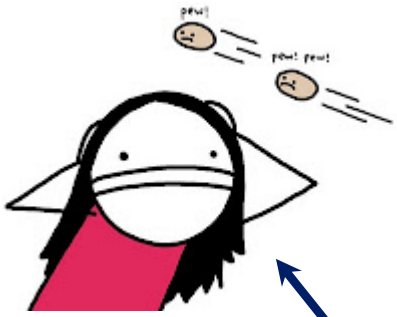
List all flights on Tuesday = *List all the flights leaving on Tuesday.*

List all flights on Tuesday = *Wait 'til Tuesday, then list all flights.*

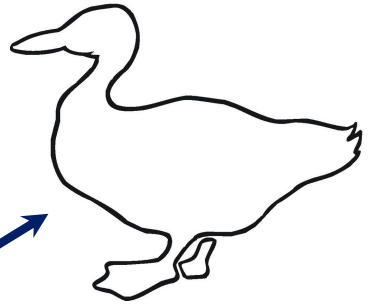
More realistic example

Retrieve all the local patient files

Baroque example



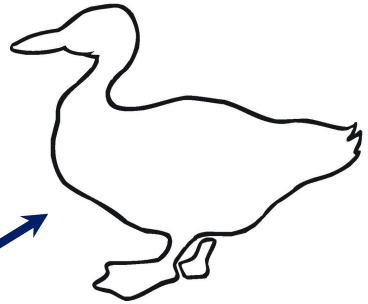
[http://casablancapa.blogspot.com/2010/05/fore.htm]l



[http://www.supercoloring.com/pages/duck-outline/]

I saw her duck with a telescope.

Baroque example



[http://www.supercoloring.com/pages/duck-outline/]

I saw her duck with a telescope.

http://www.clipartmojo.com/plugins/Clipart/ClipartStock1/star%20gazing.png

http://www.geocities.ws/looneyebay/dell/bb040.jpg

http://pokerfoldingtable.com/wp-content/uploads/2009/02/three-men-gambling-sitting-at-poker-table-playing-cards-betting-party-pen-ink-drawing-300x234.png

Challenge: the "illogic" of inference

CFCU Community Credit Union

Our Story Contact Us Branche

Accounts Loans & Visa Home Loans Digital Business Insurance

Savings

Over 75,000 individuals in Tompkins, Cortland, Seneca, Cayuga, and Ontario Counties turn to CFCU for their savings needs! It is easy to see why.

Correction

The online version of the article dated June 19, 2019 issue titled The Bank Tower's New Look: CFCU community credit union has 75,00 members. They do not have 60,000 members. we regret the error.

https://www.mycfcu.com/accounts/savings.html

Challenge: sentiment

An *extremely* simple setting: is a given review on a known topic positive or negative?

"It may be a bit early to make such judgments, but Battlefield Earth may well turn out to be the worst movie of this century." (Elvis Mitchell, May 12, 2000)

don't we just need to look for "worst", "best", "love", hate", etc.?

Best cues may not be obvious

but people aren't that good at picking indicative cues.

▲ dazzling brilliant phenomenal excellent fantastic ▼ suck terrible awful unwatchable hideous	58%
▲ gripping mesmerizing riveting spectacular cool awesome thrilling badass excellent moving exciting ▼ bad cliched sucks boring stupid slow	64%
▲ love wonderful best great superb beautiful still ▼ bad worst stupid waste boring ? !	69%

Beyond indicative terms

1. This laptop is a great deal.
2. The release of this laptop caused a great deal of hoopla.
3. Yeah, this laptop is a great deal ... and I've got a nice bridge you might be interested in.

Beyond indicative terms

This film should be brilliant. It sounds like a great plot, the actors are first [rate], and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up. (David Wilcock, http://www.killermovies.com/c/copland/reviews/5sq.html)

What phrases indicate what, anyway?

The phrases

"amazing camera"

and

"decent quality"

in reviews have *negative* effects on demand for the corresponding products. (Archak, Ghose, and Ipeirotis 2007)

No indicative terms

She ran the gamut of emotions from A to B.

Read the book.
(credit: Bob Bland)

Conversation complications

: Do you know when the train to Boston leaves?

: Yes.

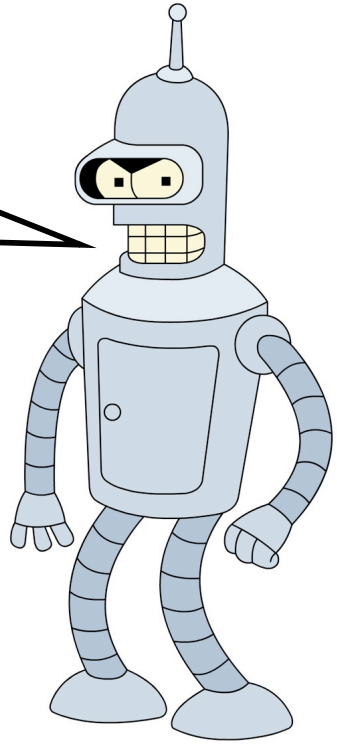
: I want to know when the train to Boston leaves.

: I understand.

[Grishman 1986]

Images: http://3.bp.blogspot.com/_o4kq5TNL0Z4/TUx0j6E5BLI/AAAAAAAA5k/J7xjhvrcNlU/s1600/Trillian-hitchhikers-guide-to-the-galaxy-the-2005.jpg, http://www.tvacres.com/images/robots_androids_marvin_movie.jpg

I'm sorry, Dave, I'm afraid I can't do that.



[http://browse.deviantart.com/?qh=&section=&global=1&q=muscleduck#/d14nst5]

[http://selcouth.com/2011/03]

I'm afraid you might be right.

Meeting these challenges: a brief history

1940s – 50s:
From language to probability

"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point ...

[For] the engineering problem, the significant aspect is that the actual message is one selected from a set of possible messages."

--C. Shannon, 1948

Bell Laboratories

Language, statistics, cryptography

WWII: Turing helps break the German “Enigma” code

Why is this man smiling?

I can see her duck without a telescope!

Encryption process

[W. Weaver memo on translation, 1949]

Two probabilities to infer

I can see her duck without a telescope!

Prob. of generating this original message?

Encryption process

Prob. of doing this encryption of the original?

[Russian]

Another use of message probs: speech recognition

(1) It's hard to recognize speech

(2) It's hard to wreck a nice beach

Both messages have almost the same acoustics, but different likelihoods.

1950s-1980s: Breaking with statistics

N. Chomsky (1957):

(a) Colorless green ideas sleep furiously

(b) Furiously sleep ideas green colorless

The argument: Neither sentence has ever occurred in the history of English. So any statistical model would given them the same probability (zero).

The field moved to sophisticated non-probabilistic models of language.

1990s: The empiricists strike back

- Huge amounts of data start coming online
- Advances in algorithms, models, and horsepower

Beyond:
integrating language insights and machine-learning techniques?

Why is this man smiling?

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision.... I do not know what the right answer is, but I think [different] approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

Part 2: Likely Topics

"The fun stops here"

- This course covers selected advanced topics in natural language processing (NLP) and/or information retrieval, with a conscious attempt to avoid topics covered by other Cornell courses. Hence:
- Students seeking a general introduction to NLP should take CS 4740 ("Introduction to Natural Language Processing) or CS 4744 ("Computational Linguistics") instead.
- Students interested purely in language simply as an application domain for machine learning should consider other courses instead: Significant portions of CS6740/IS6300 will be devoted to modeling language phenomena formally in ways that (to date) are not machine-learning oriented.

Some related courses

This fall:

Intro NLP (CS4740/5740)

NLP and social interaction (CS/IS 6742)

Black box (neural) models of language (LING7710)

Next spring:

Computational linguistics (CS3740/LING4424)

Intro NLP (CS 5740), may not be available to Ithaca students

Language and information (INFO/CS4300)

Deep latent variable models (CS67??)

Deep generative models (CS677x)

Next year?

language learning thru interaction (CS6741)

For those missing today's lecture: read through
http://www.cs.cornell.edu/courses/cs6740/2019fa