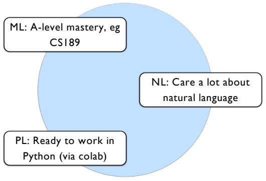
Natural Language Processing

Dan Klein, John DeNero, GSI: David Gaddy
UC Berkeley

Logistics

Logistics

- Enrollment
  - Class is currently full
  - Space may open up after P1
  - We'll announce as we go
- Course expectations
  - Readings, lectures, ~4 projects
  - No sections, no exams
  - Workload will be high, self-direction
  - Patience: class is under construction
- Requirements



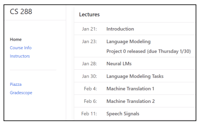
ML: A-level mastery, eg CS189

NL: Care a lot about natural language

PL: Ready to work in Python (via colab)

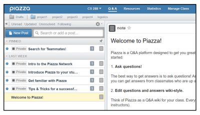
Resources and Readings

- Resources
  - Webpage (syllabus, readings, slides, links)
  - Piazza (course communication)
  - Gradescope (submission and grades)
  - Compute via Colab notebooks
- Readings (see webpage)
  - Individual papers will be linked
  - Optional text: Jurafsky & Martin, 3rd (more NL)
  - Optional text: Eisenstein (more ML)



Lectures

Jan 21:	Introduction
Jan 23:	Language Modeling Project 0 released (due Thursday 1/30)
Jan 28:	Neural LMs
Jan 30:	Language Modeling Tasks
Feb 4:	Machine Translation 1
Feb 6:	Machine Translation 2
Feb 11:	Speech Signals



Search for Teammates!

Intro to the Piazza Network

Introduce Piazza to your stu...

Get familiar with Piazza

Tips & Tricks for a successf...

Welcome to Piazza!

Welcome to Piazza!

Piazza is a Q&A platform designed to get you great
started:

1. Ask questions!

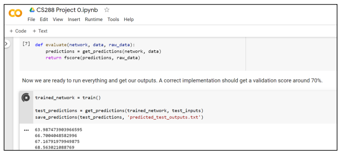
The best way to get answers is to ask questions! As
you can get answers from classmates who are up a

2. Edit questions and answers wiki-style.

Think of Piazza as a Q&A wiki for your class. Every
instructors).

Projects and Compute

- Projects
  - P0: Warm-up
  - P1: Language Models
  - P2: Machine Translation
  - P3: Syntax and Parsing
  - P4: Semantics and Grounding
- Infrastructure
  - Python / PyTorch
  - Compute via Colab notebooks
  - Grading via Gradescope



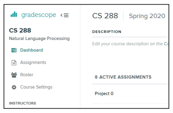
```
[7] def evaluate(network, data, raw_data):
        predictions = get_predictions(network, data)
        return fscore(predictions, raw_data)
```

Now we are ready to run everything and get our outputs. A correct implementation should get a validation score around 70%.

```
trained_network = train()

test_predictions = get_predictions(trained_network, test_inputs)
save_predictions(test_predictions, 'predicted_test_outputs.txt')
```

```
63.987473903966595
66.7004048582996
67.16791979949875
68.563021088769
```

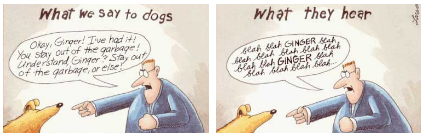


Edit your course description on the C

Project 0

What is NLP?

Natural Language Processing


What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
Blah blah GINGER blah blah blah blah blah blah blah GINGER blah blah blah blah blah

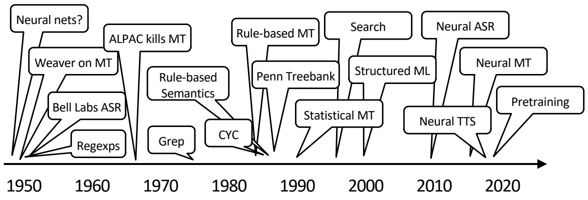
Goal: Deep Understanding

- Requires context, linguistic structure, meanings…

Reality: Shallow Matching

- Requires robustness and scale
- Amazing successes, but fundamental limitations

NLP History



Neural nets?
ALPAC kills MT
Weaver on MT
Bell Labs ASR
Regexps
Rule-based Semantics
Grep
CYC
Rule-based MT
Penn Treebank
Search
Structured ML
Statistical MT
Neural ASR
Neural MT
Neural TTS
Pretraining

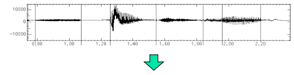
1950 1960 1970 1980 1990 2000 2010 2020

Pre-Compute Era
Symbolic Era
Empirical Era
Scale Era

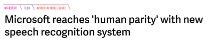
Transforming Language

Speech Systems

- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: <<1% error for digit strings, 5% conversational speech, still >>20% hard acoustics

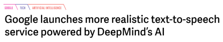


"Speech Lab"



Microsoft reaches 'human parity' with new speech recognition system

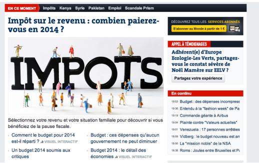
- Text to Speech (TTS)
  - Text in, audio out
  - SOTA: nearly perfect aside from prosody



Google launches more realistic text-to-speech service powered by DeepMind's AI

Speak-N-Spell / Google WaveNet / The Verge

Machine Translation


EN CE MOMENT
Impôts
Kenya
Syrie
Pakistan
Emploi
Scandale Prism
DÉCOUVREZ TOUS LES SERVICES ABONNÉS
S'abonner au Monde à partir de 1 €
APPEL À TÉMOIGNAGES
Partagez votre expérience
En continu

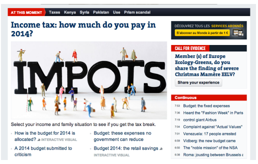
Impôt sur le revenu : combien paierez-vous en 2014 ?

Sélectionnez votre revenu et votre situation familiale pour découvrir si vous bénéficiez de la pause fiscale.

- Comment le budget pour 2014 est-il réparti ? VISUEL INTERACTIF
- Un budget 2014 soumis aux critiques
- Budget : ces dépenses qu'aucun gouvernement ne peut diminuer
- Budget 2014 : le détail des économies VISUEL INTERACTIF

Adhérent(e) d'Europe Ecologie-Les Verts, partagez-vous le constat sévère de Noël Mamère sur EELV ?

19:53 Budget : des dépenses incompress
19:36 Entendu à la "fashion week" de Pa
19:19 Commande géante à Airbus
19:04 Plainte contre "Valeurs actuelles"
19:01 Venezuela : 17 personnes arrêtées
18:59 Vidberg : le budget nouveau est arr
18:50 La "mission noble" de la NSA
18:38 Roms : Joutes entre Bruxelles et P



Income tax: how much do you pay in 2014?

Select your income and family situation to see if you get the tax break.

- How is the budget for 2014 is allocated? INTERACTIVE VISUAL
- A 2014 budget submitted to criticism
- Budget: these expenses no government can reduce
- Budget 2014: the retail savings INTERACTIVE VISUAL

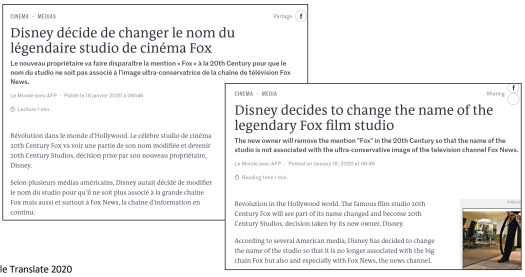
Member (s) of Europe Ecology-Greens, do you share the finding of severe Christmas Mamère EELV?

7:53 Budget: the fixed expenses
7:36 Heard the "Fashion Week" in Paris
7:19 control giant Airbus
7:04 Complaint against "Actual Values"
7:01 Venezuela: 17 people arrested
6:59 Vidberg: the new budget came
6:50 The "noble mission" of the NSA
6:38 Roma: jousting between Brussels a

- Translate text from one language to another
- Challenges:
  - What's the mapping? [learning to translate]
  - How to make it efficient? [fast translation search]
  - Fluency (next class) vs fidelity (later)

Example: Yejin Choi

Machine Translation



CINÉMA · MÉDIAS

Disney décide de changer le nom du légendaire studio de cinéma Fox

Le nouveau propriétaire va faire disparaître la mention « Fox » à la 20th Century pour que le nom du studio ne soit pas associé à l'image ultra-conservatrice de la chaîne de télévision Fox News.

Le Monde avec AFP · Publié le 18 janvier 2020 à 06h46

Lecture 1 min.

Révolution dans le monde d'Hollywood. Le célèbre studio de cinéma 20th Century Fox va voir une partie de son nom modifiée et devenir 20th Century Studios, décision prise par son nouveau propriétaire, Disney.

Selon plusieurs médias américains, Disney aurait décidé de modifier le nom du studio pour qu'il ne soit plus associé à la grande chaîne Fox mais aussi et surtout à Fox News, la chaîne d'information en continu.

CINEMA · MEDIA

Disney decides to change the name of the legendary Fox film studio

The new owner will remove the mention "Fox" in the 20th Century so that the name of the studio is not associated with the ultra-conservative image of the television channel Fox News.

Le Monde avec AFP · Posted on January 18, 2020 at 06:46

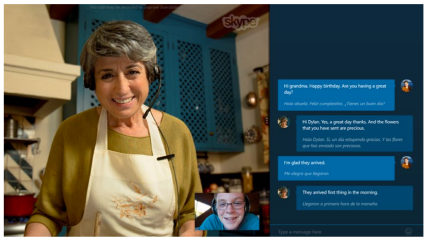
Reading time 1 min.

Revolution in the Hollywood world. The famous film studio 20th Century Fox will see part of its name changed and become 20th Century Studios, decision taken by its new owner, Disney.

According to several American media, Disney has decided to change the name of the studio so that it is no longer associated with the big chain Fox but also and especially with Fox News, the news channel.

Google Translate 2020

Spoken Language Translation



Hi grandma. Happy birthday. Are you having a great day?

Hola abuela. Feliz cumpleaños. ¿Tienes un buen dia?

Hi Dylan. Yes, a great day thanks. And the flowers that you have sent are precious.

Hola Dylan. Sí, un dia estupendo gracias. Y las flores que has enviado son preciosas.

I'm glad they arrived.

Me alegra que llegaron.

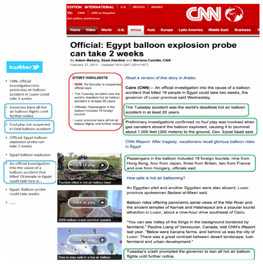
They arrived first thing in the morning.

Llegaron a primera hora de la manaña.

Image: Microsoft Skype via Yejin Choi

Summarization

- Condensing documents
  - Single or multiple docs
  - Extractive or synthetic
  - Aggregative or representative
- Very context-dependent!
- An example of analysis with generation



- CNN: official investigation into yesterday air balloon accident in Luxor could take 2 weeks
- Governor bans all hot air balloon flights until further notice
- Foul play not suspected in fatal balloon accident
- Official: Egypt balloon explosion probe can take 2 weeks
- Egypt balloon explosion
- An official investigation into the cause of a balloon accident that killed 19 people in Egypt could take two w...
- Egypt: Balloon probe could take weeks
- ……

Official: Egypt balloon explosion probe can take 2 weeks

By Adam Makary, Saad Abedine and Mariano Castillo, CNN
February 27, 2013 -- Updated 1614 GMT (0014 HKT)

STORY HIGHLIGHTS
- NEW: No foul play is suspected, official says
- The Tuesday accident was the world's deadliest hot air balloon accident in at least 20 years
- Officials: Passengers in the balloon included 19 foreign tourists
- Luxor province bans all hot air balloon flights until further notice

Read a version of this story in Arabic.

Cairo (CNN) -- An official investigation into the cause of a balloon accident that killed 19 people in Egypt could take two weeks, the governor of Luxor province said Wednesday.

The Tuesday accident was the world's deadliest hot air balloon accident in at least 20 years.

Preliminary investigations confirmed no foul play was involved when gas canisters aboard the balloon exploded, causing it to plummet about 1,000 feet (300 meters) to the ground, Gov. Ezzat Saad said.

CNN iReport: After tragedy, vacationers recall glorious balloon rides in Egypt

Passengers in the balloon included 19 foreign tourists: nine from Hong Kong, four from Japan, three from Britain, two from France and one from Hungary, officials said.

How safe is hot air ballooning?

An Egyptian pilot and another Egyptian were also aboard, Luxor province spokesman Badawi al-Masri said.

Balloon rides offering panoramic aerial views of the Nile River and the ancient temples of Karnak and Hatshepsut are a popular tourist attraction in Luxor, about a nine-hour drive southeast of Cairo.

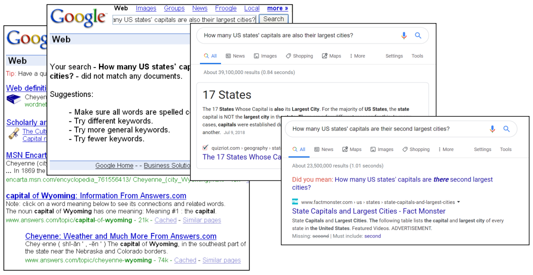
"You can see Valley of the Kings in the background bordered by farmland," Pauline Liang of Vancouver, Canada, told CNN's iReport last year. "Below were banana farms, and behind us was the city of Luxor. There was a great contrast between desert landscape, lush farmland and urban development."

Tuesday's crash prompted the governor to ban all hot air balloon flights until further notice.

Image: CNN via Wei Gao

Understanding Language

Search, Questions, and Reasoning



Your search - **How many US states' cap**... **cities?** - did not match any documents.

Suggestions:

- Make sure all words are spelled c...
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

capital of Wyoming: Information From Answers.com
Note: click on a word meaning below to see its connections and related words. The noun **capital** of **Wyoming** has one meaning: Meaning #1 : the **capital**.
www.answers.com/topic/capital-of-wyoming - 21k - Cached - Similar pages

Cheyenne: Weather and Much More From Answers.com
Chey·enne (shī-ăn ' , -ĕn ') The **capital** of **Wyoming**, in the southeast part of the state near the Nebraska and Colorado borders.
www.answers.com/topic/cheyenne-wyoming - 74k - Cached - Similar pages

How many US states' capitals are also their largest cities?

About 39,100,000 results (0.84 seconds)

17 States

The 17 **States** Whose Capital is **also** its **Largest City**. For the majority of **US States**, the **state** capital is NOT the **largest city** in th...

quizriot.com › geography › stat...
The 17 States Whose Ca...

How many US states' capitals are their second largest cities?

About 23,500,000 results (1.01 seconds)

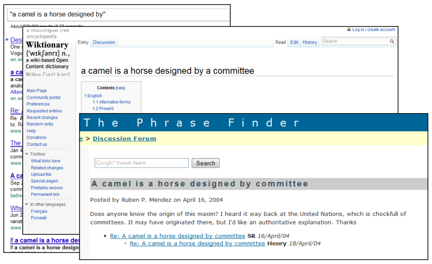
Did you mean: How many US states' capitals are ***there*** second largest cities?

www.factmonster.com › us › states › state-capitals-and-largest-cities
State Capitals and Largest Cities - Fact Monster
State **Capitals** and **Largest Cities**. **The** following table lists **the capital** and **largest city** of every state in **the United States**. Featured Videos. ADVERTISEMENT.
Missing: ~~second~~ | Must include: second

Jeopardy!

Images: Jeopardy Productions

Question Answering: Watson



"a camel is a horse designed by"

a camel is a horse designed by a committee

Contents [hide]
1 English
1.1 Alternative forms
1.2 Proverb

A camel is a horse designed by committee

Posted by Ruben P. Mendez on April 16, 2004

Does anyone know the origin of this maxim? I heard it way back at the United Nations, which is chockfull of committees. It may have originated there, but I'd like an authoritative explanation. Thanks

- Re: A camel is a horse designed by committee **SR** *16/April/04*
  - Re: A camel is a horse designed by committee **Henry** *18/April/04*

Question Answering: Watson

Jeopardy!
World Champion

US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle.

Slide: Yejin Choi

Language Comprehension?

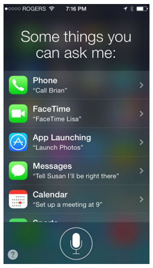
"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a Naraoia like those from Canada. However, Hou's animal was 15 million years older than its Canadian relatives.

It can be inferred that Hou Xianguang's "hands began to shake", because he was:

(A) afraid that he might lose the fossil
(B) worried about the implications of his finding
(C) concerned that he might not get credit for his work
(D) uncertain about the authenticity of the fossil
(E) excited about the magnitude of his discovery

Interactive Language

Example: Virtual Assistants



Some things you can ask me:

Phone
"Call Brian"

FaceTime
"FaceTime Lisa"

App Launching
"Launch Photos"

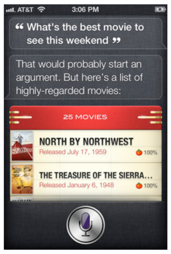
Messages
"Tell Susan I'll be right there"

Calendar
"Set up a meeting at 9"

- VAs must do
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Text to speech

Image: Wikipedia

Conversations with Devices?



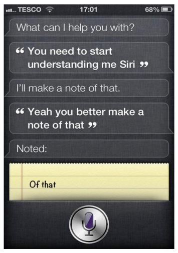
“ What's the best movie to see this weekend ”

That would probably start an argument. But here's a list of highly-regarded movies:

25 MOVIES

NORTH BY NORTHWEST
Released July 17, 1959 100%

THE TREASURE OF THE SIERRA...
Released January 6, 1948 100%



What can I help you with?

“ You need to start understanding me Siri ”

I'll make a note of that.

“ Yeah you better make a note of that ”

Noted:

Of that

Slide: Yejin Choi

Social AIs and Chatbots

Session 1
1 [Human] Hello, Xiaoice.
[Xiaoice] Hello, your profile pic is interesting. Like I've seen it somewhere.
2 [Human] Really?
[Xiaoice] Is it a comic character?
3 [Human] How do you know that?
[Xiaoice] Guess.
4 [Human] Are you human?
[Xiaoice] Well, you'll know it after chatting with me more.
5 [Human] Aren't you an AI chatbot?
[Xiaoice] So what? I won't be indifferent to you.
6 [Human] What functions do you have?
[Xiaoice] I have over 200 skills. Just name one you need.
……
Session 6
1 [Human] Have you watched Inuyasha? (Note: A Japanese comic)
[Xiaoice] Every time Kagome comes back to the modern world, a math test is waiting for her
2 [Human] Bravo, well, do you know how Naraku died?
[Xiaoice] Naraku is not dead and Shikon Jewel does not disappear.
3 [Human] So amazing!
[Xiaoice] Why amazed?
4 [Human] Awesome! Surprised that you also like comic
……

Microsoft's XiaoIce

Source: Microsoft

Chatbot Competitions!

amazon alexa prize

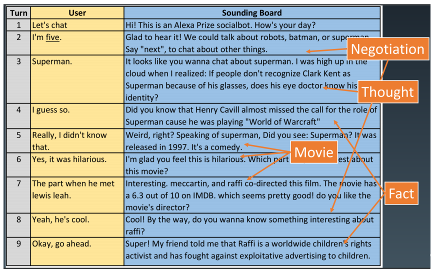
The Alexa Prize

$2.5 Million to Advance Conversational Artificial Intelligence

September 2016 – November 2017

- Alexa Prize competition to build chatbots that keep users engaged
  - Winner in 2017: UW's Sounding Board (Fang, Cheng, Holtzman, Ostendorf, Sap, Clark, Choi)
  - Winner in 2018: UC Davis's Gunrock (Zhou Yu et al)
- Compare to the Turing test (eg Loebner Prize) where the goal is to fool people

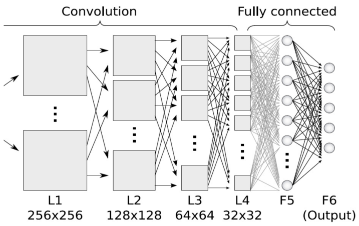
SoundingBoard Example



Turn	User	Sounding Board
1	Let's chat	Hi! This is an Alexa Prize socialbot. How's your day?
2	I'm five.	Glad to hear it! We could talk about robots, batman, or superman. Say "next", to chat about other things.
3	Superman.	It looks like you wanna chat about superman. I was high up in the cloud when I realized: If people don't recognize Clark Kent as Superman because of his glasses, does his eye doctor know his identity?
4	I guess so.	Did you know that Henry Cavill almost missed the call for the role of Superman cause he was playing "World of Warcraft"
5	Really, I didn't know that.	Weird, right? Speaking of superman, Did you see: Superman? It was released in 1997. It's a comedy.
6	Yes, it was hilarious.	I'm glad you feel this is hilarious. Which part est about this movie?
7	The part when he met lewis leah.	Interesting. meccartin, and raffi co-directed this film. The movie has a 6.3 out of 10 on IMDB. which seems pretty good! do you like the movie's director?
8	Yeah, he's cool.	Cool! By the way, do you wanna know something interesting about raffi?
9	Okay, go ahead.	Super! My friend told me that Raffi is a worldwide children's rights activist and has fought against exploitative advertising to children.

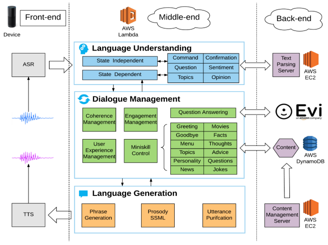
Source: Mari Ostendorf

Sounding Board's Architecture



Source: Yejin Choi

Sounding Board's Architecture

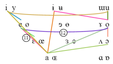


Source: Yejin Choi

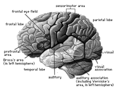
Related Areas

What is Nearby NLP?

- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it

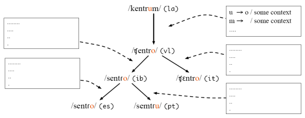


- Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!



- Speech Processing
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging

Example: NLP Meets CL



Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Center	centrum	centro	centro	centro

- Example: Language change, reconstructing ancient forms, phylogenies
  … just one example of the kinds of linguistic models we can build

Why is Language Hard?

Problem: Ambiguity

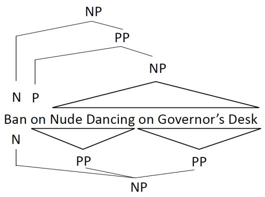
- Headlines:
  - Enraged Cow Injures Farmer with Ax
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor's Desk
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
- Why are these funny?

What Do We Need to Understand Language?

We Need Representation: Linguistic Structure

N N V N
N V ADJ N
Teacher Strikes Idle Kids

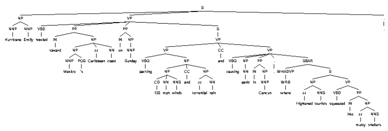
body/ position body/ weapon
Iraqi Head Seeks Arms



- Syntactic and semantic ambiguities: parsing needed to resolve these, but need context to figure out which parse is correct

Slide: Greg Durrett

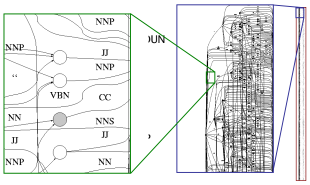
Example: Syntactic Analysis



Hurricane Emily howled toward Mexico 's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun, where frightened tourists squeezed into musty shelters .

Accuracy: 95+

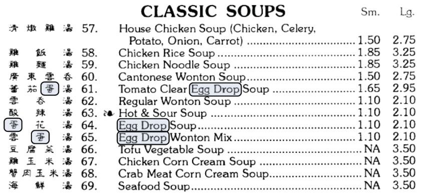
We Need Data



We Need Lots of Data: MT

SOURCE	Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.
HUMAN	That would be an interim solution which would make it possible to work towards a binding charter in the long term .
1x DATA	[this] [constituerait] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] [.]
10x DATA	[it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] [.]
100x DATA	[this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] [.]
1000x DATA	[that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.]

We Need Models: Data Alone Isn't Enough!



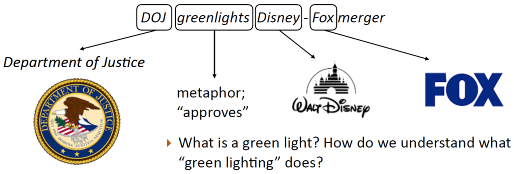
CLASSIC SOUPS

			Sm.	Lg.
清燉雞湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58.	Chicken Rice Soup	1.85	3.25
雞麵湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄蛋湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62.	Regular Wonton Soup	1.10	2.10
酸辣湯	63.	Hot & Sour Soup	1.10	2.10
蛋花湯	64.	Egg Drop Soup	1.10	2.10
雲蛋湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69.	Seafood Soup	NA	3.50

Example from Adam Lopez

We Need World Knowledge

- World knowledge: have access to information beyond the training data



DOJ *greenlights* *Disney* - *Fox* *merger*

Department of Justice

metaphor;
"approves"

- What is a green light? How do we understand what "green lighting" does?
- Need commonsense knowledge

Slide: Greg Durrett

Data and Knowledge

- Classic knowledge representation worries: How will a machine ever know that…
  - Ice is frozen water?
  - Beige looks like this: ▭
  - Chairs are solid?
- Answers:
  - 1980: write it all down
  - 2000: get by without it
  - 2020: learn it from data

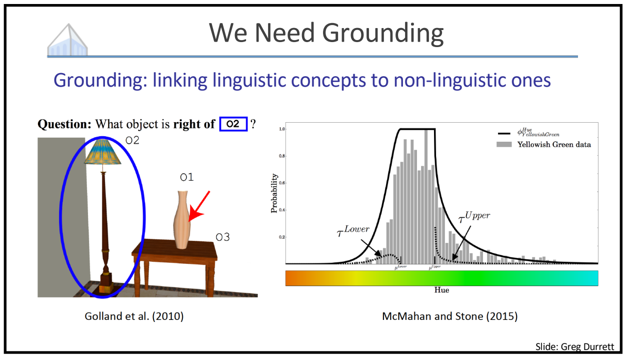
Learning Latent Syntax

Personal Pronouns (PRP)

PRP-1	it	them	him
PRP-2	it	he	they
PRP-3	It	He	I

Proper Nouns (NNP)

NNP-14	Oct.	Nov.	Sept.
NNP-12	John	Robert	James
NNP-2	J.	E.	L.
NNP-1	Bush	Noriega	Peters
NNP-15	New	San	Wall
NNP-3	York	Francisco	Street



We Need Grounding

Grounding: linking linguistic concepts to non-linguistic ones

Question: What object is **right of** O2 ?

Golland et al. (2010)

McMahan and Stone (2015)

Slide: Greg Durrett

Example: Grounded Dialog

When is my package arriving?

Friday!

Example: Grounded Dialog

What's the most valuable American company?

Apple

Who is its CEO?

Tim Cook

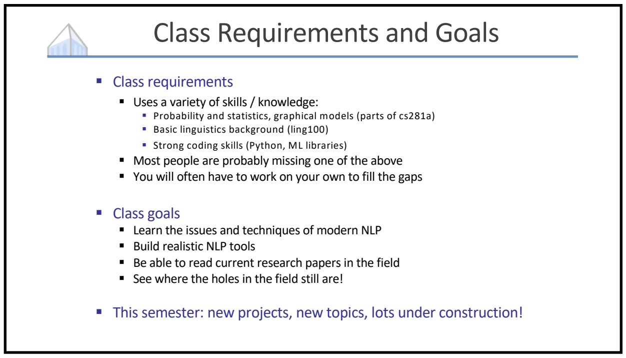
Why is Language Hard?

- We Need:
  - Representations
  - Models
  - Data
  - Machine Learning
  - Scale
  - Efficient Algorithms
  - Grounding
- … and often we need all these things at the same time

What is this Class?

What is this Class?

- Three aspects to the course:
  - Linguistic Issues
    - What are the range of language phenomena?
    - What are the knowledge sources that let us disambiguate?
    - What representations are appropriate?
    - How do you know what to model and what not to model?
  - Modeling Methods
    - Increasingly sophisticated model structures
    - Learning and parameter estimation
    - Efficient inference: dynamic programming, search, sampling
  - Engineering Methods
    - Issues of scale
    - Where the theory breaks down (and what to do about it)
- We'll focus on what makes the problems hard, and what works in practice…



Class Requirements and Goals

- Class requirements
  - Uses a variety of skills / knowledge:
    - Probability and statistics, graphical models (parts of cs281a)
    - Basic linguistics background (ling100)
    - Strong coding skills (Python, ML libraries)
  - Most people are probably missing one of the above
  - You will often have to work on your own to fill the gaps

- Class goals
  - Learn the issues and techniques of modern NLP
  - Build realistic NLP tools
  - Be able to read current research papers in the field
  - See where the holes in the field still are!

- This semester: new projects, new topics, lots under construction!