

# CSE 517

# Natural Language Processing

# Winter 2017

Introduction  
Yejin Choi

Slides adapted from Dan Klein, Luke Zettlemoyer

# What is NLP?



- Fundamental goal: *deep* understand 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, sentiment analysis, question answering...
  - Unknown: human-level comprehension (is this just NLP?)

# Why NLP

---

➔ To access information & knowledge

# 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.





# Knowledge Graph: "things not strings"

Home Tips & Tricks **Features** Search Stories Playground Blog Help



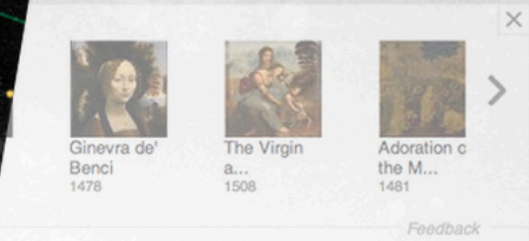
## The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.



## See it in action

Discover answers to questions you never thought to ask, and explore collections and lists.



### Leonardo da Vinci



Leonardo di ser Piero da Vinci was an Italian Renaissance polymath: painter, sculptor, architect, musician, scientist, mathematician, engineer, inventor, anatomist, geologist, cartographer, botanist, and writer. Wikipedia

**Born:** April 15, 1452, [Anchiano](#)

**Died:** May 2, 1519, [Clos Lucé](#)

**Buried:** [Château d'Amboise](#)

**Parents:** [Caterina da Vinci](#), [Piero da Vinci](#)

**Structures:** [Vejbørn Sand Da Vinci Project](#)



# Information Extraction

---

- From unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	started
Russell T. Lewis	New York Times newspaper	executive vice president	ended
Lance R. Primis	New York Times Co.	president and CEO	started

# Information Extraction

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	started
Russell T. Lewis	New York Times newspaper	executive vice president	ended
Lance R. Primis	New York Times Co.	president and CEO	started

## Sub-problems:

- 1) Named entity recognition: finding named entities  $X$  and their types  $T(X)$   
persons: "Russell T. Lewis", "Lance R. Primis"  
companies: "New York Times Newspaper", "New York Times Co."
  - 2) Relation extraction: the relation  $R(X,Y)$  between named entities  $X, Y$   
 $Works\_for(\text{Russell T. Lewis}, \text{New York Times Newspaper})$
  - 3) Coreference resolution: which text spans refer to the same named entity?  
{Russell T.Lewis, He, He} are an equivalence set.
- Is this easy or hard?
  - Easier if the model exploits the redundancy of information!

# Question Answering

## ■ Question Answering:

- More than search
- Can be really easy: "What's the capital of Wyoming?"
- Can be harder: "How many US states' capitals are also their largest cities?"
- Can be open ended: "What are the main issues in the global warming debate?"

## ■ Natural Language Interaction:

- Understand requests and act on them
- "Make me a reservation for two at Quinn's tonight"

The screenshot shows a Google search interface. At the top, the Google logo is on the left, and navigation links for 'Web', 'Images', 'Groups', 'News', 'Froogle', 'Local', and 'more »' are on the right. The search bar contains the text 'any US states' capitals are also their largest cities?' and a 'Search' button. Below the search bar, a 'Web' tab is selected. The main content area displays the message: 'Your search - **How many US states' capitals are also their largest cities?** - did not match any documents.' Below this, a 'Suggestions:' section lists four items: '- Make sure all words are spelled correctly.', '- Try different keywords.', '- Try more general keywords.', and '- Try fewer keywords.' At the bottom of the page, there is a footer with links for 'Google Home', 'Business Solutions', and 'About Google'.

### [capital of Wyoming: Information From Answers.com](#)

Note: click on a word meaning below to see its connections and related words.

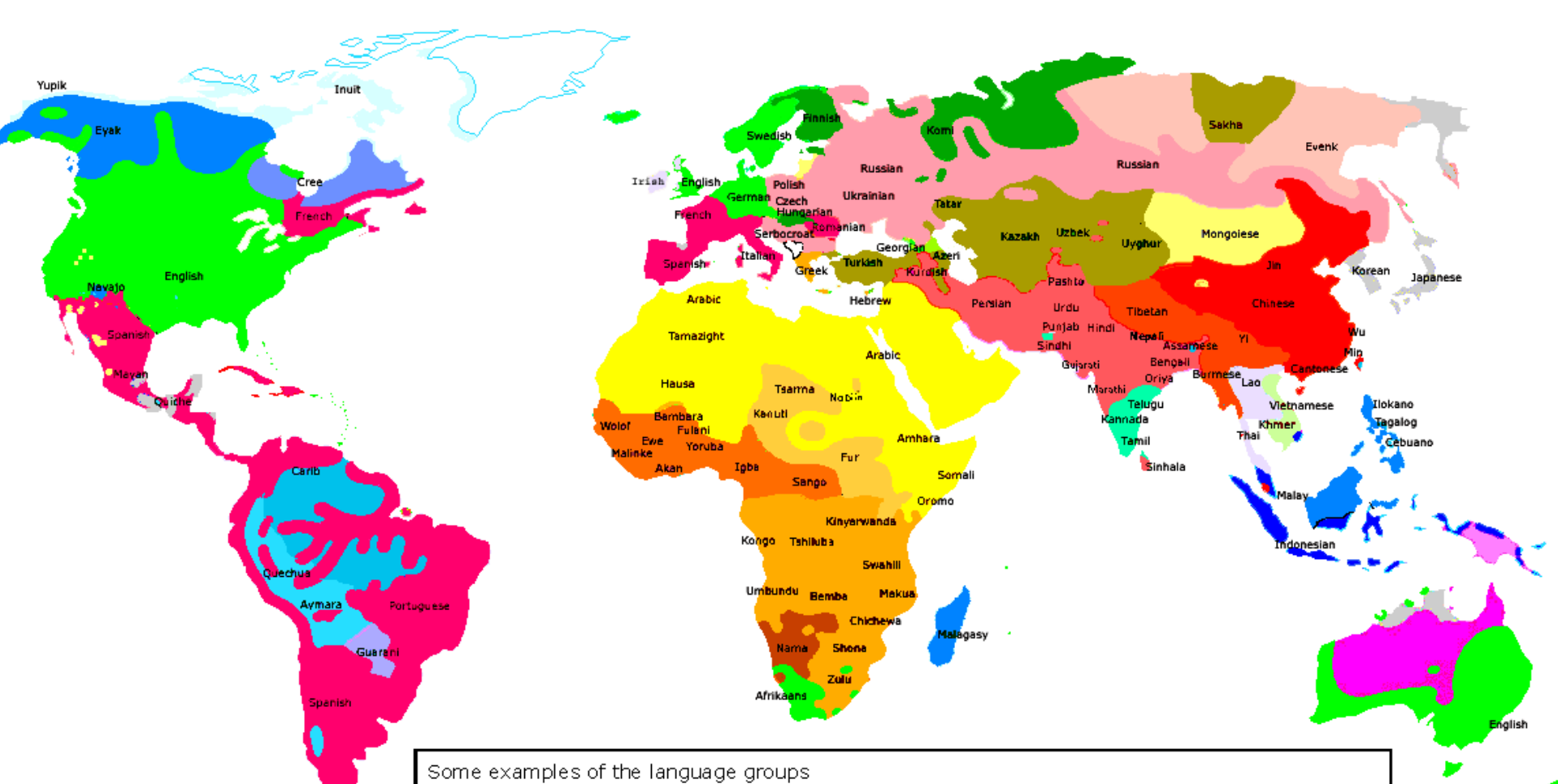
The noun **capital** of **Wyoming** has one meaning: Meaning #1 : the **capital**.

[www.answers.com/topic/capital-of-wyoming](#) - 21k - [Cached](#) - [Similar pages](#)

### [Cheyenne: Weather and Much More From Answers.com](#)

Chey·enne ( shī-ăn ' , -ěn ' ) 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](#)



Some examples of the language groups

<ul style="list-style-type: none"> <li>■ Afro-Asiatic</li> <li> <ul style="list-style-type: none"> <li>■ Niger-Congo</li> <li> <ul style="list-style-type: none"> <li>■ Bantu</li> <li>■ Nilo-Saharan</li> <li>■ Khoisan</li> </ul> </li> <li>■ Indo-European</li> <li> <ul style="list-style-type: none"> <li>■ Germanic</li> <li>■ Albanic</li> <li>■ Romance</li> <li>■ Slavic</li> <li>■ Indo-Iranian</li> <li>■ Baltic</li> <li>■ Caucasian</li> </ul> </li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>■ Altaic</li> <li> <ul style="list-style-type: none"> <li>■ Turkic</li> <li>■ Mongolic</li> <li>■ East Siberian languages</li> </ul> </li> <li>■ Uralic</li> <li>■ Dravidian</li> <li> <ul style="list-style-type: none"> <li>■ Sino-Tibetan</li> <li> <ul style="list-style-type: none"> <li>■ Chinese</li> <li>■ Burmese-Tibetan</li> </ul> </li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>■ Austro-Asiatic</li> <li>■ Austronesian</li> <li>■ Borneo-Philippines/Formosan</li> <li>■ Nuclear Malayo-Polynesian</li> <li>■ Pama-Nyungan</li> <li>■ Tai-Kadal</li> <li>■ Isolate</li> </ul>	<ul style="list-style-type: none"> <li>■ Na-Déne</li> <li>■ Eskimo-Aleut</li> <li>■ American Indian</li> <li>■ Algonic</li> <li>■ Uto-Aztecan</li> <li>■ Mayan</li> <li>■ Andean</li> <li>■ Tupian</li> <li>■ Brazilian indigenous</li> </ul>
---	--	--	---

# Machine Translation

## "Il est impossible aux journalistes de rentrer dans les régions tibétaines"

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'illégalité".

**Les faits** Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959

**Vidéo** Anniversaire de la rébellion tibétaine : la Chine sur ses gardes



## "It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

**Facts** The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959

**Video** Anniversary of the Tibetan rebellion: China on guard



- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]
  - Fluency (second half of this class) vs fidelity (later)



# 2013 Google Translate: French

EN CE MOMENT Impôts Kenya Syrie Pakistan Emploi Scandale Prism

## Impôt sur le revenu : vous en 2014 ?



Sélectionnez votre revenu et votre situation familiale pour bénéficier de la pause fiscale.

- Comment le budget pour 2014 est-il réparti ? [VISUEL INTERACTIF](#)
- Un budget 2014 soumis aux critiques



**Le chômage baisse pour la première fois depuis avril 2011** [POST DE BLOG](#)

AT THIS MOMENT Taxes Kenya Syria Pakistan Use Prism scandal

## 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](#)
- Budget: these expenses no government can reduce
- A 2014 budget submitted to criticism
- Budget 2014: the retail savings [INTERACTIVE VISUAL](#)



**Unemployment fell for the first time since April 2011** [POST BLOG](#)



**Surviving in the Central time looting and anarchy**

DÉCOUVREZ TOUS LES **SERVICES ABONNÉS**

S'abonner au Monde à partir de 1 €



**CALL FOR EVIDENCE**

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

Share your experience

**Continuous**

- 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 &

DE  
FURSAC

automne-hiver 13/14

# 2013 Google Translate: Russian



Поиск  
Например: [Большой Кавказ](#)

Мир | Наука | Общество | Здоровье | Красота

## Новости

- 20:09 [В Шри-Ланке хотели перевезти золото в желудках](#)
- 20:00 [Выходец из России может получить "Нобеля" по химии](#)
- 19:46 [В США установили стандарты торговли оружием](#)
- 19:35 [Директор Эрмитажа: Обыски нанесли ущерб музею](#)
- 19:25 [Мозгу ребенка полезен послеобеденный сон](#)

- 19:24 [Ролик с водителями-детьми заинтересовалась петербургская полиция](#)
- 19:15 [К Марсу приближается "комета века"](#)
- 18:55 [Выявлено более 160 нарушений на судостроительных предприятиях](#)

- 18:44 [Астахов назначен на новый срок в Европейской сети детских омбудсменов](#)

## Главное

### ["Обиженные люди работают, а иностранцы к нам не поедут"](#)

25.09.2013 19:48



Ректор "Бауманки" Анатолий "Правде.Ру", какие шаги надо чиновникам и ученым в связи реформе РАН.

### [Фотосессия](#)



[Наводнение в Индии: 40 жителей эвакуированы](#)

Найроби. Газета The Independent "Уэстгейт" во время захвата.

## Мир

[Иранцы не заметили](#)



Поиск

For example, [the Greater Caucasus](#)

World | Science | Society | Health | Beauty | Regions | Photo | Video

Forums | archive

## News

- 20:09 [In Sri Lanka, wanted to carry the gold in the stomachs](#)
- 20:00 [A native of Russia can get the "Nobel" in Chemistry](#)
- 19:46 [In the United States set the standard arms trade](#)
- 19:35 [Director of the Hermitage: The searches have damaged the museum](#)
- 19:25 [The child's brain is useful afternoon nap](#)

- 19:24 [The roller with the drivers, children become interested in the St. Petersburg Police](#)
- 19:15 [To Mars is approaching "comet of the century"](#)
- 18:55 [There are over 160 violations at shipyards](#)

- 18:44 [Astakhov appointed for a new term in the European Network of Ombudsmen for children](#)

## Point

### ["Mentally ill people are working, and foreign scholars to us will not go"](#)

25/09/2013 19:48



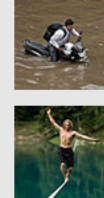
The Rector, "Bauman" Anatoly Alexandrov told with "Pravda.Ru" what steps need to be taken to officials and scientists in connection with the adoption of the law on the reform of the RAS.

### [Photoshoot](#)



[World through the lens: September 25.](#)

2466 photos



[Expert: The poorer the society is, the more scandals due to copyright](#)

09/25/2013 20:04

Why Russians are greedy for free, and do not like to pay for downloading movies and music, with "Pravda.Ru" said the head of Liveinternet German Klimenko.



[Putin met environmentalists "Greenpeace" trying to grab the platform](#)

25/09/2013 14:39

President of Russia, speaking at the International Arctic Forum in Salekhard, spoke about the ecology of Greenpeace, staged on a platform of "Prirazlomnaja."



[Expert: It is necessary to encourage participation in the election, rather than returning the column](#)

"against all"

09/25/2013 13:27

Political scientist and philosopher, Professor Oleg Matveychev HSE commented with "Pravda.Ru" Valentina Matviyenko offer to return to the ballot line "against all."



[The British newspaper described the heroes and victims in Nairobi](#)

25/09/2013 10:27

In Kenya - mourning for the victims of the terrorist attack in Nairobi. The newspaper The Independent said about the people who were at the mall, "Westgate" during capture.

## World

## Policy

## Economy



# Why NLP

---

- To access information & knowledge

 To communicate

# Human-Machine Interactions



# Will this Be Part of All Our Home Devices?

amazon echo



FROM: AMAZON.COM

*Will it rain tomorrow?*

*Set an alarm for eight a.m.*

*Play music by  
Bruno Mars*

*How many teaspoons  
are in a tablespoon?*

*Add gelato to my  
shopping list*

*Wikipedia: Abraham  
Lincoln*

*When is  
Thanksgiving?*

*Play my "dinner party"  
playlist*

*What's the weather in  
Los Angeles this weekend?*

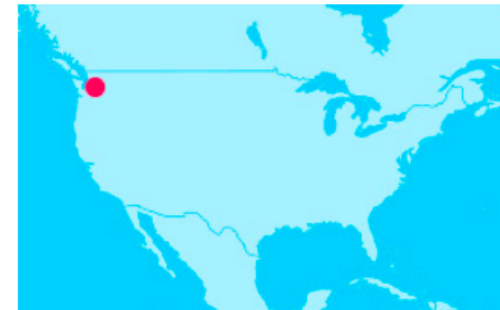
*Add "make hotel reservations"  
to my to-do list*



[< PREV TEAM](#) | [VIEW ALL](#)

# University of Washington

## Sounding Board



**Sounding Board**

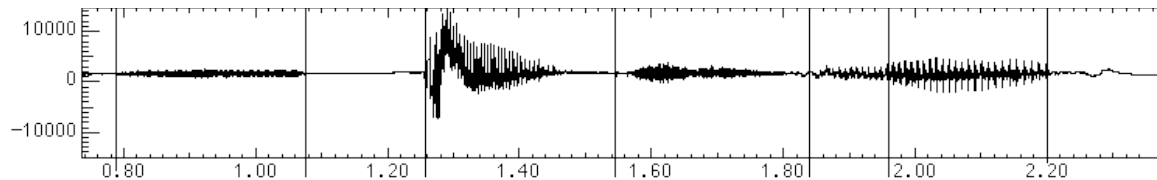
**Location:** Seattle, WA, USA  
**Faculty Advisor:** [Mari Ostendorf](#)

# Speech Recognition

---

- Automatic Speech Recognition (ASR)

- Audio in, text out
- SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV



“Speech  
Lab”

- Text to Speech (TTS)

- Text in, audio out
- SOTA: totally intelligible (if sometimes unnatural)



# Why NLP

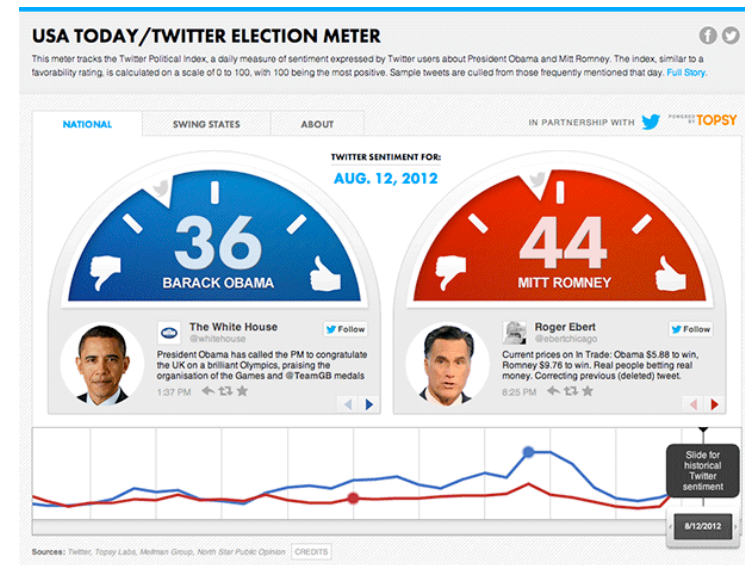
---

- To access information & knowledge
- To communicate
- ➔ To understand our society



# Analyzing public opinion, making political forecasts

- Today: In 2012 election, automatic sentiment analysis actually being used to complement traditional methods (surveys, focus groups)
- Past: "Sentiment Analysis" research started in 2002
- Future: **computational social science** and NLP for digital humanities (psychology, communication, literature and more)
- Challenge: Need statistical models for deeper semantic understanding --- subtext, intent, nuanced messages



# Why NLP

---

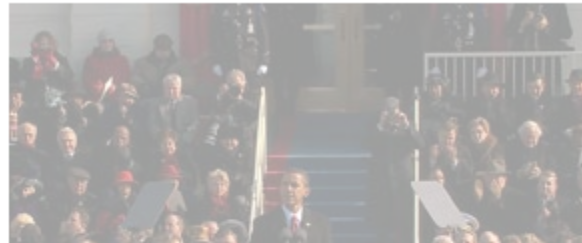
- To access information & knowledge
  - To communicate
  - To understand our society
- ➔ And to make our lives easier



# Summarization

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

WASHINGTON (CNN) -- President Obama's inaugural address was cooler, more measured and reassuring than that of other presidents making it, perhaps, the right speech for the times.



Some inaugural addresses are known for their soaring, inspirational language. Like John F. Kennedy's in 1961: "Ask not what your country can do for you. Ask what you can do for your country."

Obama's address was less stirring, perhaps, but it was also more candid and down-to-earth.

"Starting today," the new president said, "we must begin

#### STORY HIGHLIGHTS

- Obama's address less stirring than others but more candid, analyst says
- Schneider: At a time of crisis, president must be reassuring
- Country has chosen "hope over fear, unity of purpose over ... discord," Obama said
- Obama's speech was a cool speech, not a hot one, Schneider says

CNN

President Obama renewed his call for a massive plan to stimulate economic growth.

[more photos »](#)

aid in his first inaugural in 1933, "The only thing we have to fear is fear itself." Or Bill Clinton, who took office during the economic crisis of the early 1990s. "There is nothing wrong with America that cannot be fixed by what is right with America," Clinton declared at his first inaugural.

[Obama](#), too, offered reassurance.

"We gather because we have chosen hope over fear, unity of purpose over conflict and discord," Obama said.

Obama's call to unity after decades of political division echoed Abraham Lincoln's first inaugural address in 1861. Even though he delivered it at the onset of a terrible civil war, Lincoln's speech was not a call to battle. It was a call to look beyond the war, toward reconciliation based on what he called "the better angels of our nature."

Some presidents used their [inaugural address](#) to set out a bold agenda.

# Start-up Summly → Yahoo!

CEO Marissa Mayer announced an update to the app in a blog post, saying, "The new Yahoo! mobile app is also smarter, using Summly's natural-language algorithms and machine learning to deliver quick story summaries. We acquired Summly less than a month ago, and we're thrilled to introduce this game-changing technology in our first mobile application."



Launched 2011, Acquired 2013 for \$30M

# Can a robot write news?

---

Despite an expected dip in profit, analysts are generally optimistic about **Steelcase** as it prepares to reports its third-quarter earnings on Monday, December 22, 2014. The consensus earnings per share estimate is 26 cents per share.

The consensus estimate remains unchanged over the past month, but it has decreased from three months ago when it was 27 cents. Analysts are expecting earnings of 85 cents per share for the fiscal year. Revenue is projected to be 5% above the year-earlier total of \$784.8 million at \$826.1 million for the quarter. For the year, revenue is projected to come in at \$3.11 billion.

The company has seen revenue grow for three quarters straight. The less than a percent revenue increase brought the figure up to \$786.7 million in the most recent quarter. Looking back further, revenue increased 8% in the first quarter from the year earlier and 8% in the fourth quarter.

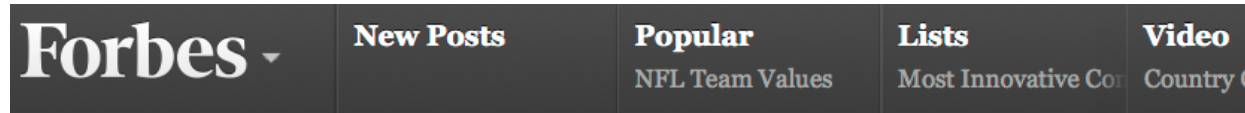
The majority of analysts (100%) rate Steelcase as a buy. This compares favorably to the analyst ratings of three similar companies, which average 57% buys. Both analysts rate Steelcase as a buy.

Steelcase is a designer, marketer and manufacturer of office furniture. Other companies in the furniture and fixtures industry with upcoming earnings release dates include: HNI and Knoll.

# Writer-bots for earthquake & financial reports

Some of the formulaic news articles are now written by computers.

- Definitely far from “Op-ed”
- Can we make the generation engine statistically learned rather than engineered?



2 FREE issues of Forbes

Forbes Partner



## Narrative Science

+ Follow (83)

NEWS

[Social](#)


[Archive](#)

Post 19 hours ago | 364 views

### Oracle Earnings Projected to Increase

Analysts expect higher profit for **Oracle** when the company reports its first quarter results on Thursday, September 18, 2014. The consensus estimate is calling for profit of 60 cents a share, reflecting a rise from 56 cents per share a year ago.

For the fiscal year, analysts are expecting earnings of \$3.01 per share. [read »](#)

 **Narrative Science**, Partner

Post 19 hours ago | 246 views

### Rite Aid Profit Expected to Slip

# Why NLP

---

- To access information & knowledge
- To communicate
- To understand our society
- To make our lives easier

 NLP and AI

# Language Comprehension?

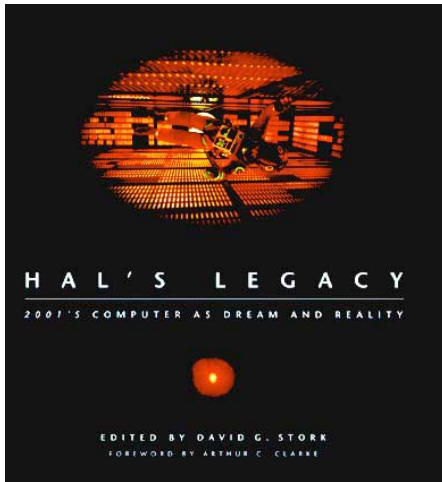
---

"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xiangang. 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 Xiangang's "hands began to shake", because he was:

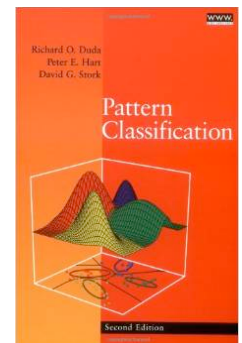
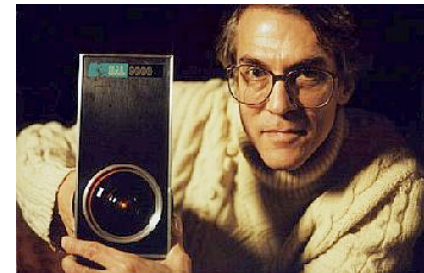
- (A) afraid that he might lose the fossil
- (B) worried about the implications of his finding
- (C) concerned that he might not get credit for his work
- (D) uncertain about the authenticity of the fossil
- (E) excited about the magnitude of his discovery

# Language and Vision



*"Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers"*

-- David Stork (HAL's Legacy, 2001) on A. Rosenfeld's vision

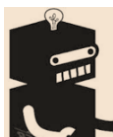




# What begins to work (e.g., Kuznetsova et al. 2014)



The flower was so  
**vivid and attractive.**



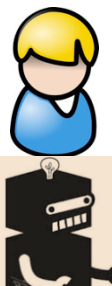
Blue flowers are **running rampant** in my garden.

We sometimes do well: 1 out of 4 times, machine captions were preferred over the original Flickr captions:



Spring in a white dress.

**Blue flowers have no scent. Small white flowers have no idea what they are.**



Scenes around the lake on my bike ride.

**This horse walking along the road as we drove by.**





# But many challenges remain (better examples of when things go awry)



The couch is definitely bigger than it looks in this photo.



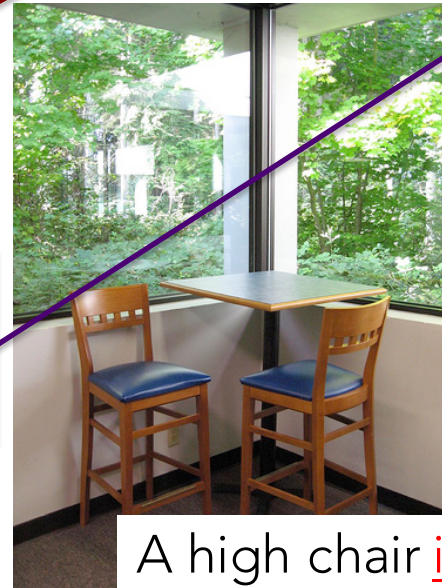
Yellow ball suspended in water.

Incorrect Object Recognition



My cat laying in my duffel bag.

Incorrect Scene Matching



Incorrect Composition

A high chair in the trees.

# Table of Content

---

- Definition of NLP

➔ Historical account of NLP

# 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

---

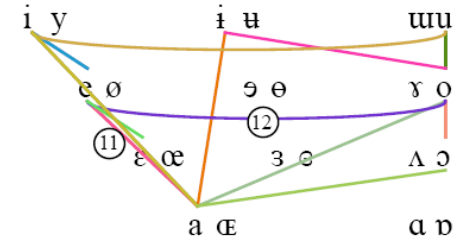
“Whenever I fire a linguist our system performance improves.” –Jelinek, 1988

- 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: you decide!

# What is Nearby NLP?

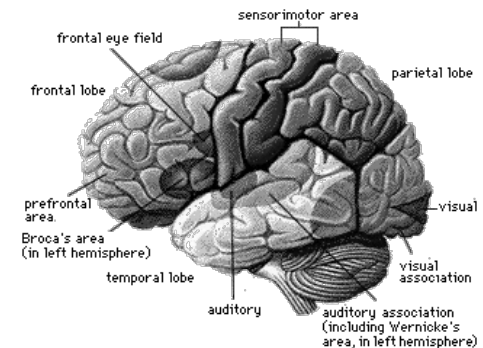
## ■ Computational Linguistics

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



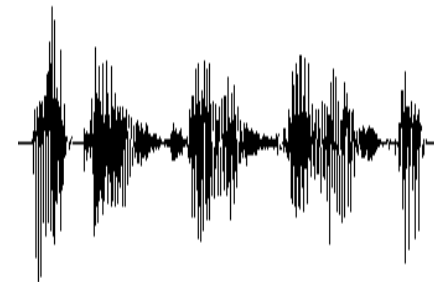
## ■ Cognitive Science

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



## ■ Speech?

- 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



# Table of Content

---

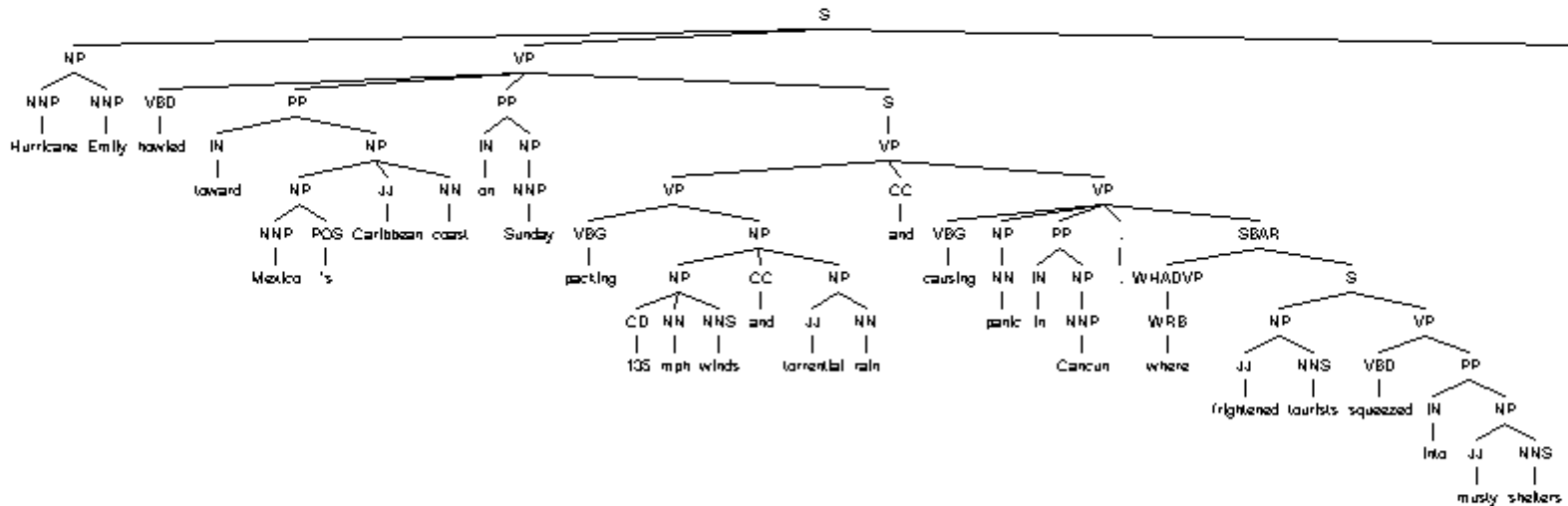
- Definition of NLP
- Historical account of NLP
- ➔ Unique challenges of NLP

# Problem: Ambiguities

---

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

# 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 .

- **SOTA:** ~90% accurate for many languages when given many training examples, some progress in analyzing languages given few or no examples



# Semantic Ambiguity

---

*At last, a computer that understands you like your mother.*

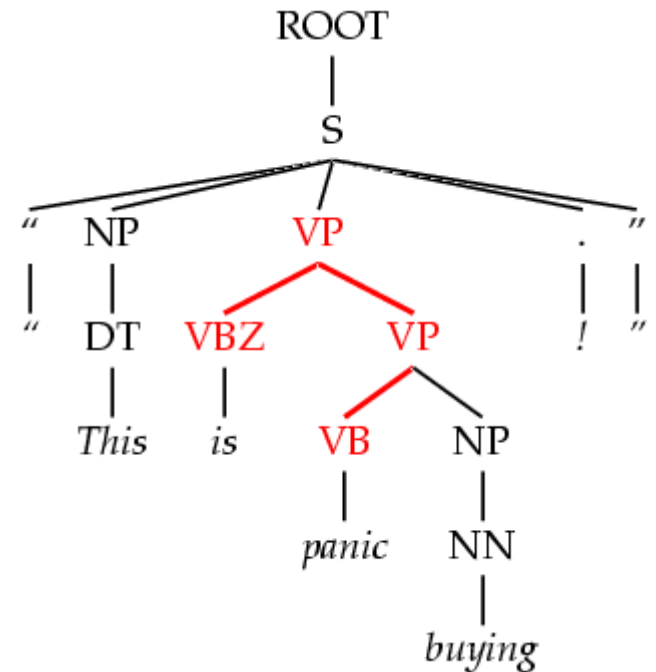
- **Direct Meanings:**
    - It understands you like your mother (does) [presumably well]
    - It understands (that) you like your mother
    - It understands you like (it understands) your mother
  - **But there are other possibilities, e.g. mother could mean:**
    - a woman who has given birth to a child
    - a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar
  - **Context matters, e.g. what if previous sentence was:**
    - Wow, Amazon predicted that you would need to order a big batch of new vinegar brewing ingredients. 😊
- [Example from L. Lee]

# Dark Ambiguities

- *Dark ambiguities*: most structurally permitted analyses are so bad that you can't get your mind to produce them

This analysis corresponds to the correct parse of

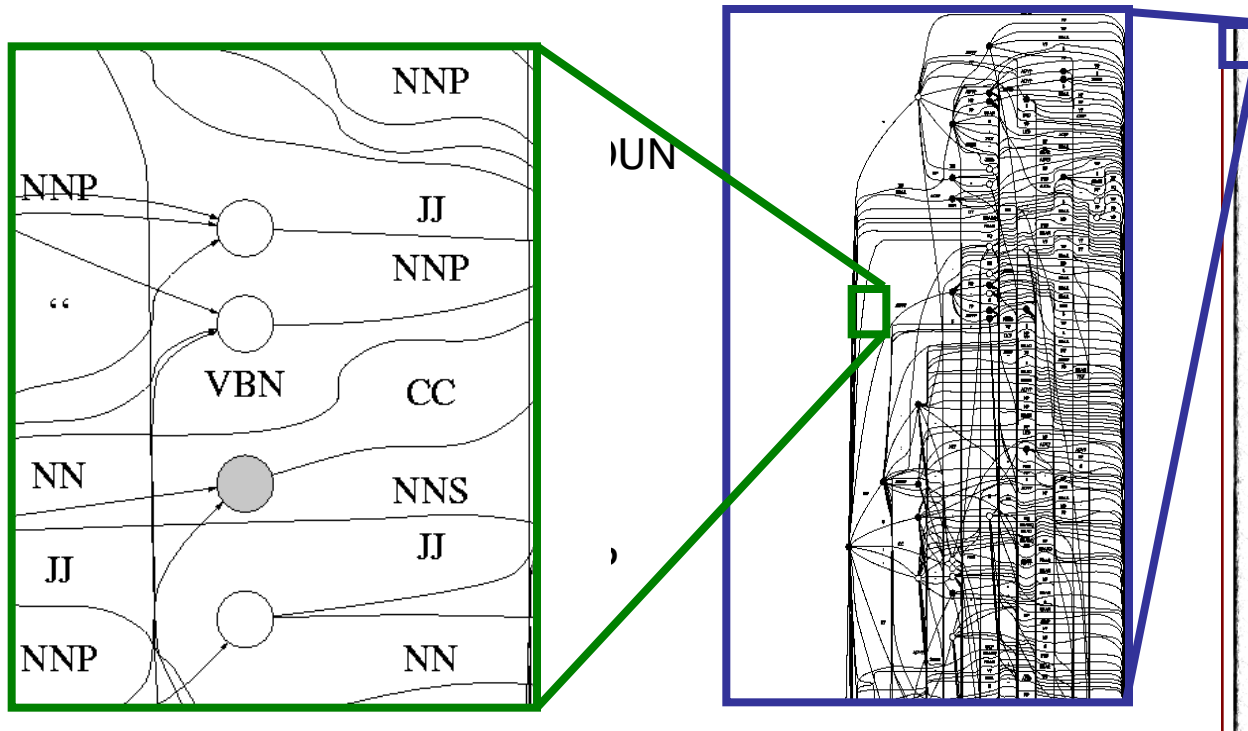
*“This will panic buyers ! ”*



- Unknown words and new usages
- *Solution*: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

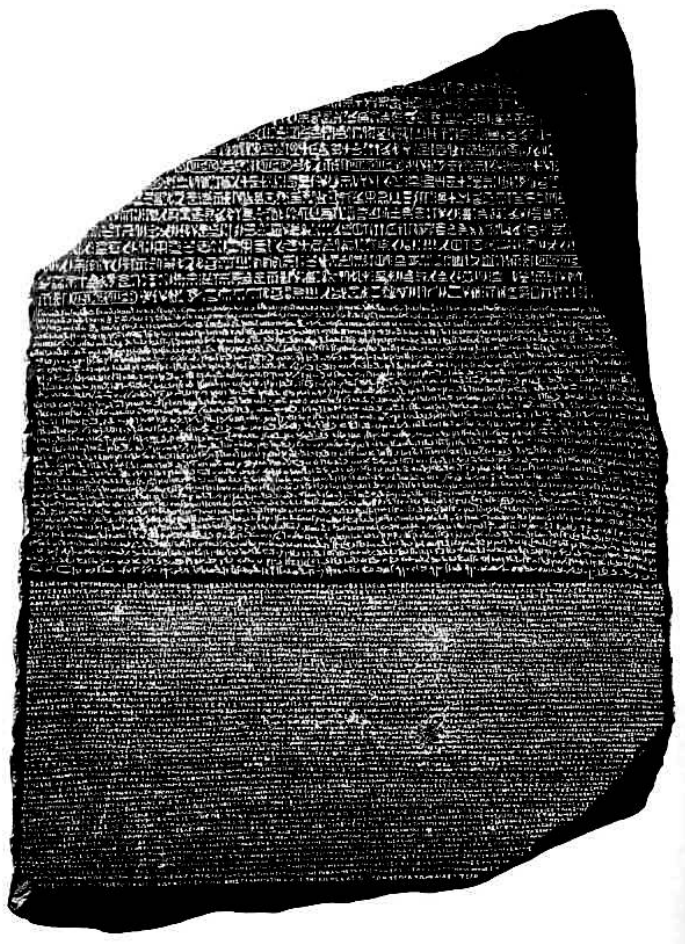
# Problem: Scale

- People *did* know that language was ambiguous!
  - ...but they hoped that all interpretations would be “good” ones (or ruled out pragmatically)
  - ...they didn't realize how bad it would be



# Corpora

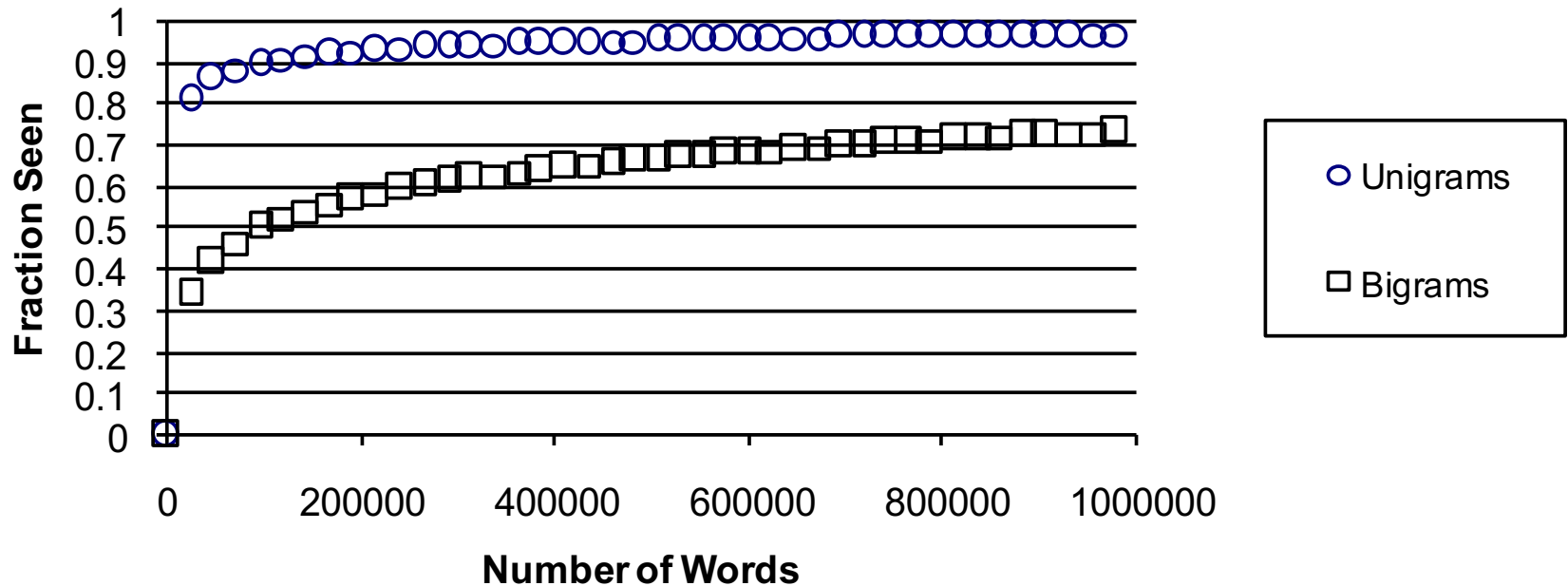
---



- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora
- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged “balanced” text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what


# Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair)



# Table of Content

---

- Definition of NLP
  - Historical account of NLP
  - Unique challenges of NLP
-  Class administrivia



# Site & Crew

---

- Site:

- <http://courses.cs.washington.edu/courses/cse517/17wi/>
- Canvas: <https://canvas.uw.edu/courses/1098228>

- Crew:

- Instructor:

[Yejin Choi](#)

- TA:

[Nicholas FitzGerald](#)

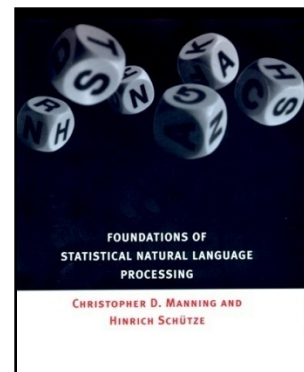
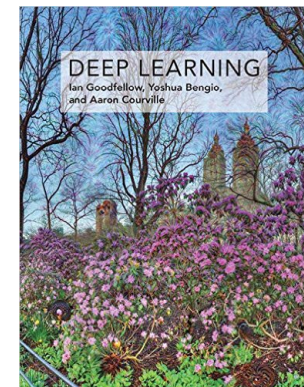
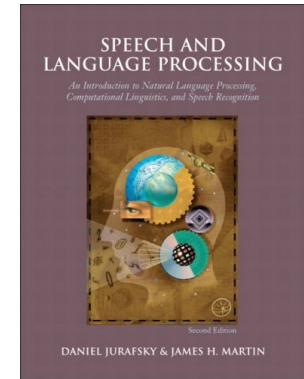
[Minjoon Seo](#)



- Office Hours: TBD (mine is Fri 3pm-4pm this week)

# Textbooks and Notes

- Textbook (recommended but not required):
  - Jurafsky and Martin, Speech and Language Processing, 2<sup>nd</sup> Edition
  - Manning and Schuetze, Foundations of Statistical NLP
  - GoodFellow, Bengio, and Courville, "Deep Learning" (free online book available at [deeplearningbook.org](http://deeplearningbook.org) )
- Lecture slides & notes are required
  - See the course website for details
- Assumed Technical Background:
  - Data structure, algorithms, strong programming skills, probabilities, statistics



# Grading & Policy

## ■ Grading:

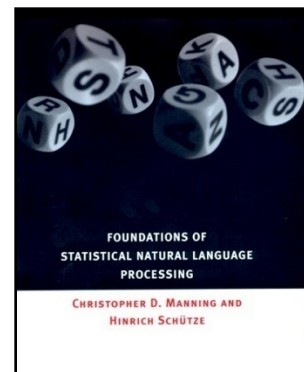
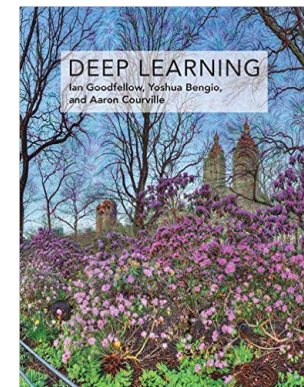
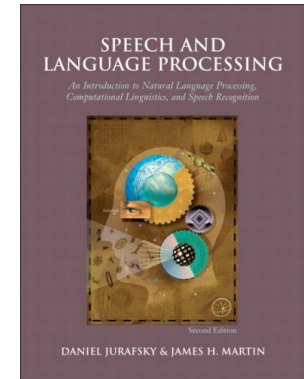
- 5 homework (50%)
- in-class workbook (15%)
- final project (30%)
- course/discussion board participation (5%)

## ■ Policy:

- All homework will be completed individually.
- Final projects can be done in groups.
- Academic honest and plagiarism.

## ■ Participation and Discussion:

- Class participation is expected and appreciated!!!
- Email is great, but please use the message board when possible (we monitor it closely)



# Final Project

---

- Ideally, something that supports your own research (industry) career
- Have to be substantive and relevant to NLP
- Deliverables:
  - Project proposal
    - max 3 pages, requires solid literature survey
  - Project presentation
    - Potential time: Monday, March 13, 2017, 2:30-?? pm?
  - Project final report
    - max 10 pages

# How to frame a final project

---

- How to frame a final project
  - Pick a technical paper and reproduce their results
    - Make sure the model is reasonably technically demanding
  - Pick an existing algorithm or learning model and design a new enhanced version
  - Apply an existing model to a new domain or an application
    - Bonus score if the new domain/application is creative!
    - Make sure to provide rigorous analysis and/or experiment with new model variations
  - Make a new dataset and conduct annotation studies
    - Make sure to provide baseline results
  - Invent a new formalism (of meaning? of language?)
    - Be very rigorous and deep or involve empirical studies
- Teaching crew will provide research advice during office hours
- However, intellectual independence is important

# 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?
  - Statistical Modeling Methods
    - Increasingly complex 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...



# Approximate Schedule

1	I. Introduction II. <b>Words</b> : Language Models (LMs)
2	
3	
4	
5	
6	
7	
8	VIII. <b>Deep Learning</b> : Neural Networks
9	VIII. <b>Deep Learning</b> : More NNs
10	VIII. <b>Deep Learning</b> : Yet More NNs

# Approximate Schedule

1	I. Introduction II. <b>Words</b> : Language Models (LMs)
2	II. <b>Words</b> : Unknown Words (Smoothing) III. <b>Sequences</b> : Hidden Markov Models (HMMs)
3	III. <b>Sequences</b> : Hidden Markov Models (HMMs) V. <b>Trees</b> : Probabilistic Context Free Grammars (PCFG)
4	
5	
6	
7	
8	VIII. <b>Deep Learning</b> : Neural Networks
9	VIII. <b>Deep Learning</b> : More NNs
10	VIII. <b>Deep Learning</b> : Yet More NNs

# Approximate Schedule

1	I. Introduction II. <b>Words</b> : Language Models (LMs)
2	II. <b>Words</b> : Unknown Words (Smoothing) III. <b>Sequences</b> : Hidden Markov Models (HMMs)
3	III. <b>Sequences</b> : Hidden Markov Models (HMMs) V. <b>Trees</b> : Probabilistic Context Free Grammars (PCFG)
4	V. <b>Trees</b> : Grammar Refinement V. <b>Trees</b> : Dependency Grammars & Mildly Context-Sensitive Grammars
5	III. <b>Sequences</b> : Sequence Tagging IV. <b>Learning</b> (Feature-Rich Models): Log-Linear Models IV. <b>Learning</b> (Structural Graphical Models): Conditional Random Fields (CRFs)
6	
7	
8	VIII. <b>Deep Learning</b> : Neural Networks
9	VIII. <b>Deep Learning</b> : More NNs
10	VIII. <b>Deep Learning</b> : Yet More NNs

# Approximate Schedule

1	I. Introduction II. <b>Words</b> : Language Models (LMs)
2	II. <b>Words</b> : Unknown Words (Smoothing) III. <b>Sequences</b> : Hidden Markov Models (HMMs)
3	III. <b>Sequences</b> : Hidden Markov Models (HMMs) V. <b>Trees</b> : Probabilistic Context Free Grammars (PCFG)
4	V. <b>Trees</b> : Grammar Refinement V. <b>Trees</b> : Dependency Grammars & Mildly Context-Sensitive Grammars
5	III. <b>Sequences</b> : Sequence Tagging IV. <b>Learning</b> (Feature-Rich Models): Log-Linear Models IV. <b>Learning</b> (Structural Graphical Models): Conditional Random Fields (CRFs)
6	VI. <b>Translation</b> : Alignment Models & Phrase-based MT
7	VII. <b>Semantics</b> : Frame Semantics VII. <b>Semantics</b> : Distributed Semantics, Embeddings
8	VIII. <b>Deep Learning</b> : Neural Networks
9	VIII. <b>Deep Learning</b> : More NNs
10	VIII. <b>Deep Learning</b> : Yet More NNs

# Comparisons with Other Classes

---

- Compared to ML
  - Typically multivariate, dynamic programming everywhere
  - Structural Learning & Inference
  - Insights into language matters (a lot!)
  - DL: RNNs, LSTMs, Seq-to-seq, Attention, ...
- Compared to undergrad NLP
  - 50 – 70% overlap (depending on offerings)
  - Fast-paced
  - Stronger engineering skills & higher degree of independence assumed
  - Final project, ideally tailored for your research and/or career
- Compared to CompLing classes
  - More focus on core algorithm design, technically more demanding in terms of math, algorithms, and programming

# Class Requirements and Goals

---

- **Class requirements**

- Uses a variety of skills / knowledge:
  - Probability and statistics
  - Basic linguistics background
  - Decent coding skills
- Most people are probably missing one of the above
- You will often have to work 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!