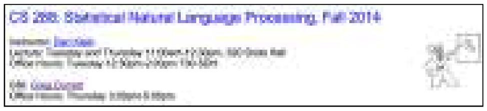
# Natural Language Processing

Lecture 1: Introduction

Dan Klein – UC Berkeley

## Course Information

http://www.cs.berkeley.edu/~klein/cs288/fa14/



CS 288: Statistical Natural Language Processing, Fall 2014

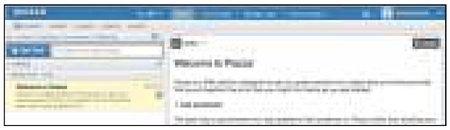
Instructor: Dan Klein

[illegible]

GSI: Greg Durrett

[illegible]

https://piazza.com/berkeley/fall2014/cs288/



Welcome to Piazza

[illegible]

## Course Requirements

- Prerequisites:
  - CS 188 (CS 281a) and preferably CS170 (A-level mastery)
  - Strong skills in Java or equivalent
  - Deep interest in language
  - Successful completion of the first project
  - There will be a lot of math and programming

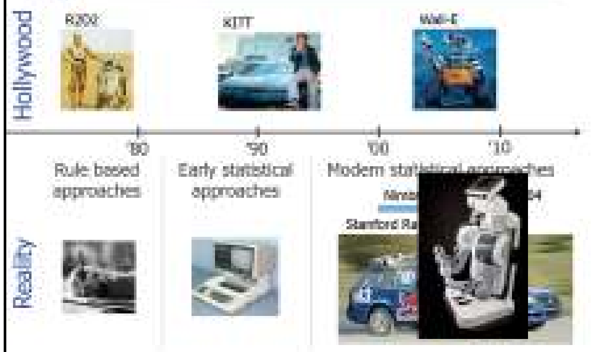
- Work and Grading:
  - Six assignments (individual, jars + write-ups)
  - This course is a major time-commitment!

- Books:
  - Primary text: Jurafsky and Martin, Speech and Language Processing, 2nd Edition (not 1st)
  - Also: Manning and Schuetze, Foundations of Statistical NLP

## Other Announcements

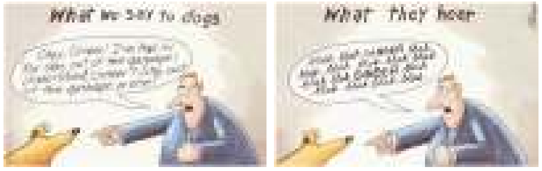
- Course Contacts:
  - Webpage: materials and announcements
  - Piazza: discussion forum
- Enrollment: We'll try to take everyone who meets the requirements
- Computing Resources
  - You will want more compute power than the instructional labs
  - Experiments can take up to hours, even with efficient code
  - Recommendation: start assignments early
- Questions?

## AI: Where Do We Stand?



Source: Slav Petrov

## Language Technologies



### Goal: Deep Understanding

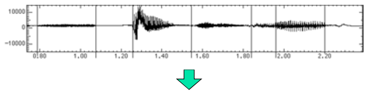
- Requires context, linguistic structure, meanings…

### Reality: Shallow Matching

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

## Speech Systems

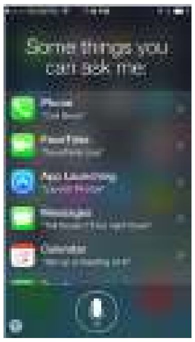
- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV



"Speech Lab"

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

## Example: Siri



Some things you can ask me:

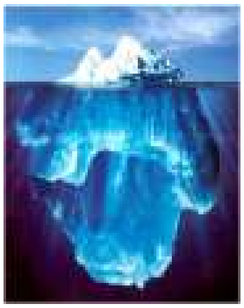
Image: Wikipedia

- Siri contains
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Text to speech

## Text Data is Superficial

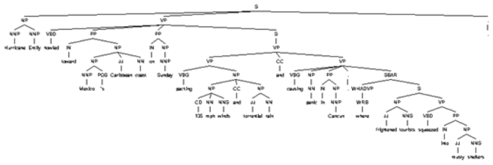
An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.

## … But Language is Complex



An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.

## Deeper Linguistic 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: 90+

## Learning Hidden Syntax

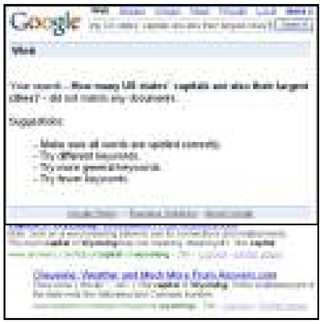
Personal Pronouns (PRP)

| | | | |
|---|---|---|---|
| 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 |

## Search, Facts, and Questions



Web

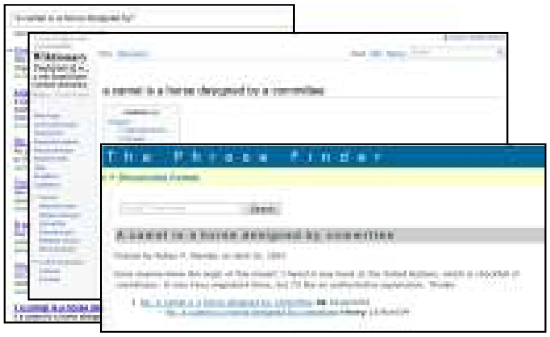
Your search - How many US states' capitals are also their largest cities? - did not match any documents.

Suggestions:

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

Cheyenne, Weather, and Much More From Answers.com

## Example: Watson



a camel is a horse designed by a committee

The Phrase Finder

A camel is a horse designed by committee

## Language Comprehension?

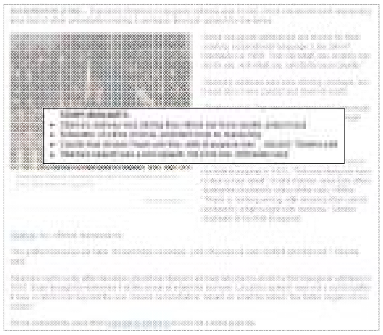
"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. 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:

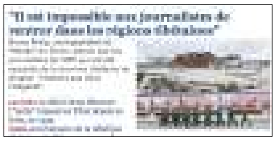
(A) afraid that he might lose the fossil
(B) worried about the implications of his finding
(C) concerned that he might not get credit for his work
(D) uncertain about the authenticity of the fossil
(E) excited about the magnitude of his discovery

## Summarization

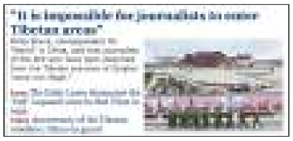
- Condensing documents
  - Single or multiple docs
  - Extractive or synthetic
  - Aggregative or representative
- Very context-dependent!
- An example of analysis with generation



## Machine Translation



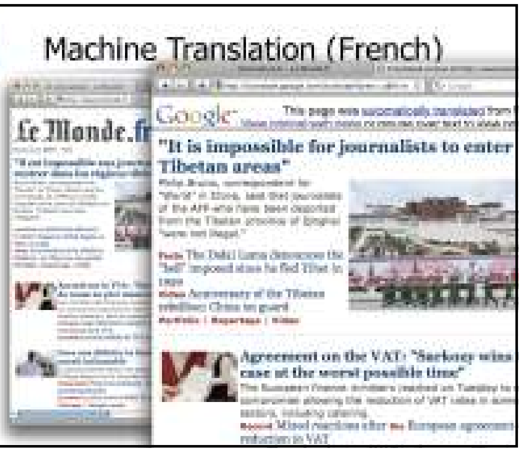
"Il est impossible aux journalistes de rentrer dans les régions tibétaines"



"It is impossible for journalists to enter Tibetan areas"

- 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 (next class) vs fidelity (later)

## Machine Translation (French)



Le Monde.fr

Google This page was automatically translated from

"It is impossible for journalists to enter Tibetan areas"

Philip Bernie, 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

Portfolio | Reportage | Video

Agreement on the VAT: "Sarkozy wins case at the worst possible time"

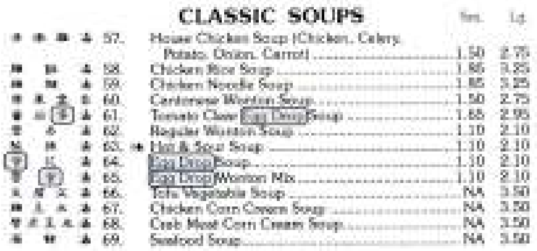
The European Finance ministers reached on Tuesday to compromise allowing the reduction of VAT rates in some sectors, including catering.

Recent Mixed reactions after the European agreement reduction in VAT

## More Data: Machine Translation

| | |
|---|---|
| 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] [.] |

## Data By Itself Isn't Enough!

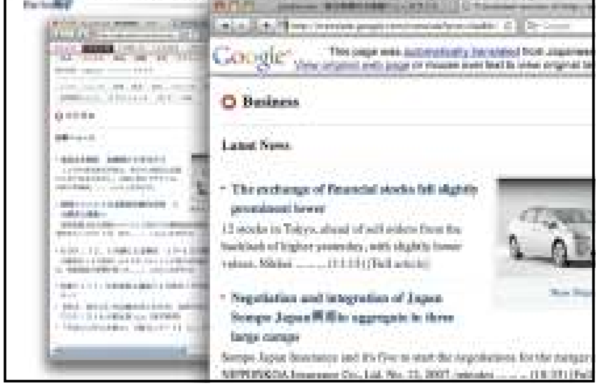


CLASSIC SOUPS

| | | Sm. | Lg. |
|---|---|---|---|
| 57. | House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot) | 1.50 | 2.75 |
| 58. | Chicken Rice Soup | 1.85 | 3.25 |
| 59. | Chicken Noodle Soup | 1.85 | 3.25 |
| 60. | Cantonese Wonton Soup | 1.50 | 2.75 |
| 61. | Tomato Clear Egg Drop Soup | 1.65 | 2.95 |
| 62. | Regular Wonton Soup | 1.10 | 2.10 |
| 63. | ✱ Hot & Sour Soup | 1.10 | 2.10 |
| 64. | Egg Drop Soup | 1.10 | 2.10 |
| 65. | Egg Drop Wonton Mix | 1.10 | 2.10 |
| 66. | Tofu Vegetable Soup | NA | 3.50 |
| 67. | Chicken Corn Cream Soup | NA | 3.50 |
| 68. | Crab Meat Corn Cream Soup | NA | 3.50 |
| 69. | Seafood Soup | NA | 3.50 |

*Example from Adam Lopez*

## Machine Translation (Japanese)



This page was automatically translated from Japanese
View original web page or mouse over text to view original la

Business

Latest News

- The exchange of financial stocks fell slightly prominent lower

13 stocks in Tokyo, ahead of sell orders from the backlash of higher yesterday, with slightly lower values. Nikkei ........(11:11) [Full article]

- Negotiation and integration of Japan Sompo Japan興亜 aggregate in three large damage

Sompo Japan Insurance and its five to start the negotiations for the merger
NIPPONKOA Insurance Co., Ltd. No. 13, 2007, releases ........(10:31) [Ful

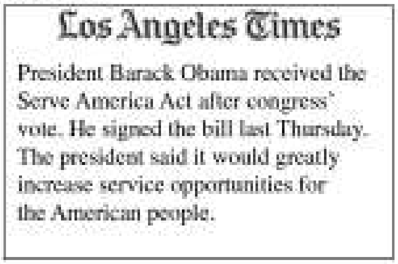
## Data and Knowledge

- Classic knowledge representation worry: 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

## Deeper Understanding: Reference

Q: Who signed the Serve America Act?

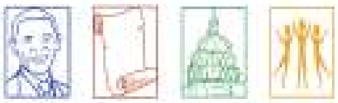
A: Barack Obama



Los Angeles Times

President Barack Obama received the Serve America Act after congress' vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

## Names vs. Entities



President Barack Obama received the Serve America Act after congress' vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

## Example Errors

Input

America Online announced on Monday that the company plans to update its instant messaging service.

Correct

America Online the company its

instant messaging service

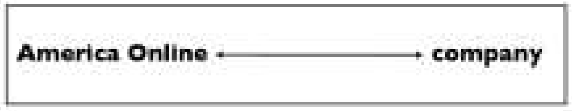
Guess

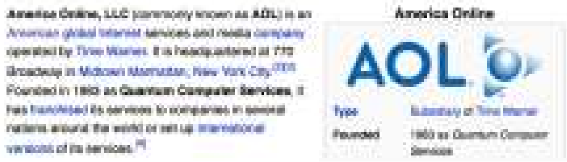
America Online

the company its

instant messaging service

## Discovering Knowledge





America Online, LLC (commonly known as AOL) is an American global Internet services and media company operated by Time Warner. It is headquartered at 770 Broadway in Midtown Manhattan, New York City.[2][3] Founded in 1983 as Quantum Computer Services, it has franchised its services to companies in several nations around the world or set up international versions of its services.[4]

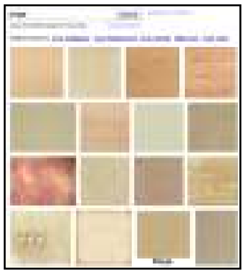
America Online

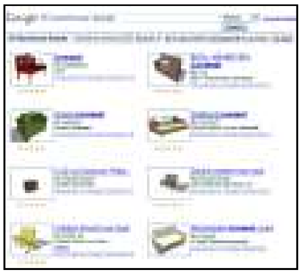
| | |
|---|---|
| Type | Subsidiary of Time Warner |
| Founded | 1983 as Quantum Computer Services |

## Grounded Language

## Grounding with Natural Data

*… on the beige loveseat.*

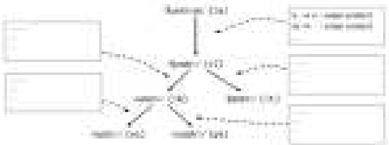




## What is Nearby NLP?

- 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?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP

## Example: NLP Meets CL



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

- Example: Language change, reconstructing ancient forms, phylogenies

  … just one example of the kinds of linguistic models we can build

## 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…

## 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 (Java), well beyond cs61b
  - 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 statistical 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 (speech, translation, analysis)

## Some BIG Disclaimers

- The purpose of this class is to train NLP researchers
  - Some people will put in a LOT of time – this course is more work than most classes (grad or undergrad)
  - There will be a LOT of reading, some required, some not – you will have to be strategic about what reading enables your goals
  - There will be a LOT of coding and running systems on substantial amounts of real data
  - There will be a LOT of machine learning / math
  - There will be discussion and questions in class that will push past what I present in lecture, and I'll answer them
  - Not everything will be spelled out for you in the projects
  - Especially this term: new projects will have hiccups
- Don't say I didn't warn you!

## Some Early NLP History

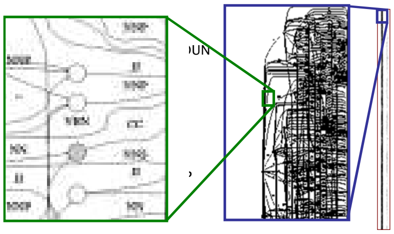
- 1950's:
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military
    - Toy models: MT using basically word-substitution
  - Optimism!
- 1960's and 1970's: NLP Winter
  - Bar-Hillel (FAHQT) and ALPAC reports kills MT
  - Work shifts to deeper models, syntax
  - … but toy domains / grammars (SHRDLU, LUNAR)
- 1980's and 1990's: The Empirical Revolution
  - Expectations get reset
  - Corpus-based methods become central
  - Deep analysis often traded for robust and simple approximations
  - *Evaluate everything*
- 2000+: Richer Statistical Methods
  - Models increasingly merge linguistically sophisticated representations with statistical methods, confluence and clean-up
  - *Begin to get both breadth and depth*

## Problem: Structure

- 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?

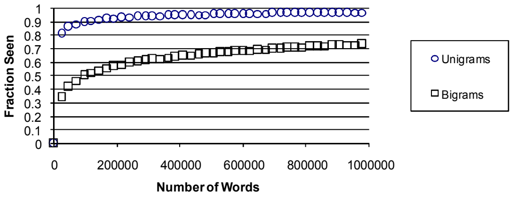
## Problem: Scale

- 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



# Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair), and rule rates in newswire



# Outline of Topics

- Words and Sequences
  - Speech recognition
  - N-gram models
  - Working with a lot of data
- Structured Classification
- Trees
  - Syntax and semantics
  - Syntactic MT
  - Question answering
- Machine Translation
- Other Topics
  - Reference resolution
  - Summarization
  - Diachronics
  - …

# A Puzzle

- You have already seen N words of text, containing a bunch of different word types (some once, some twice…)
- What is the chance that the N+1$^{st}$ word is a new one?