CMPT 413: Computational Linguistics
CMPT 825: Natural Language Processing

Angel Xuan Chang

2020-09-09

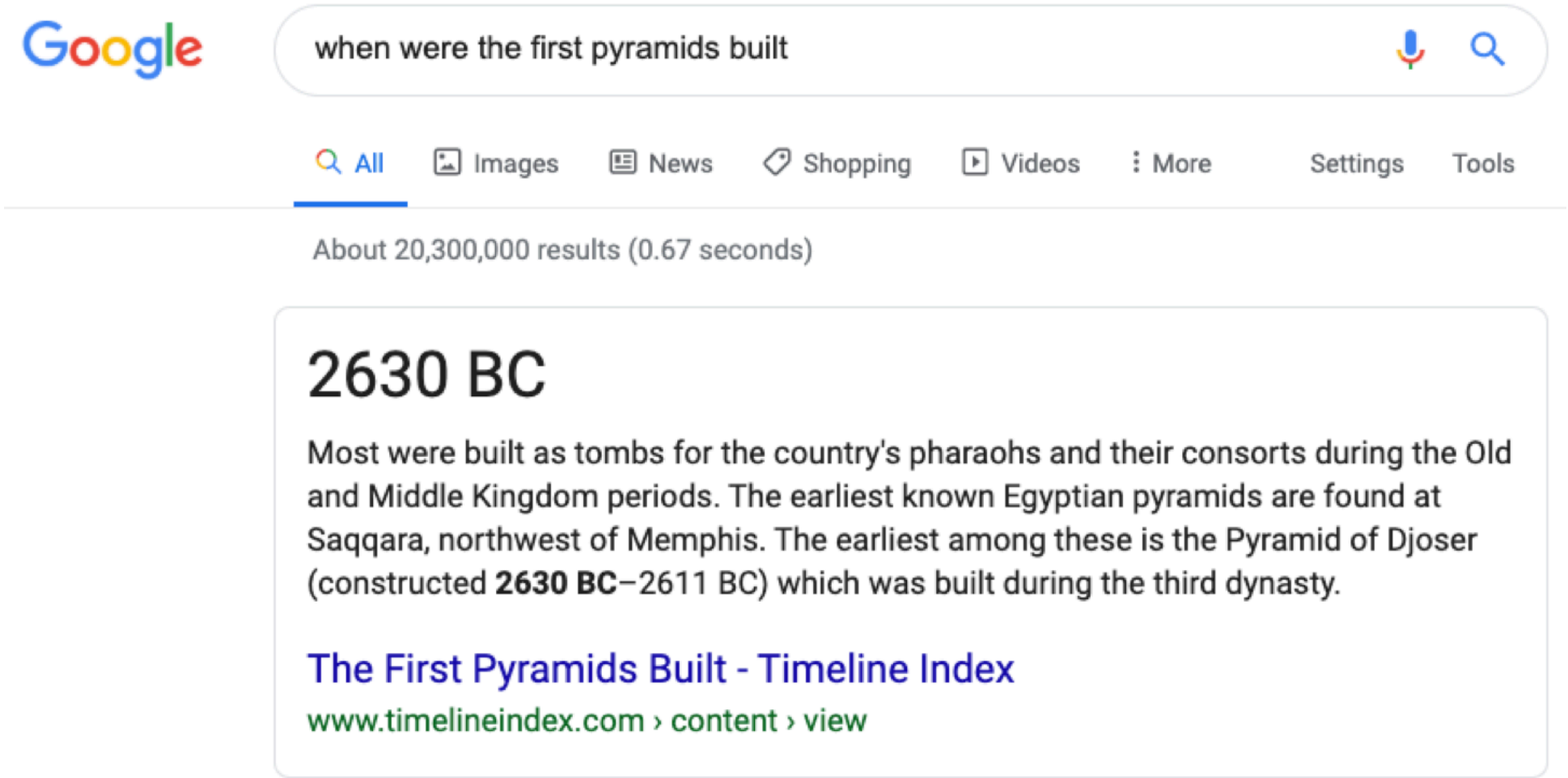
Adapted from slides from Anoop Sarkar, Danqi Chen and Karthik Narasimhan

NLP is everywhere

Google translate

Virtual assistants

Information finding



Question Answering

IBM Watson defeated two of Jeopardy's greatest champions in 2011

Programming Languages

C, C++, Java, Python, …

- Unambiguous
- Fixed
- Designed
- Learnable?
- Known simple semantics

Natural Languages

French, English, Korean, Chinese, Tagalog, …

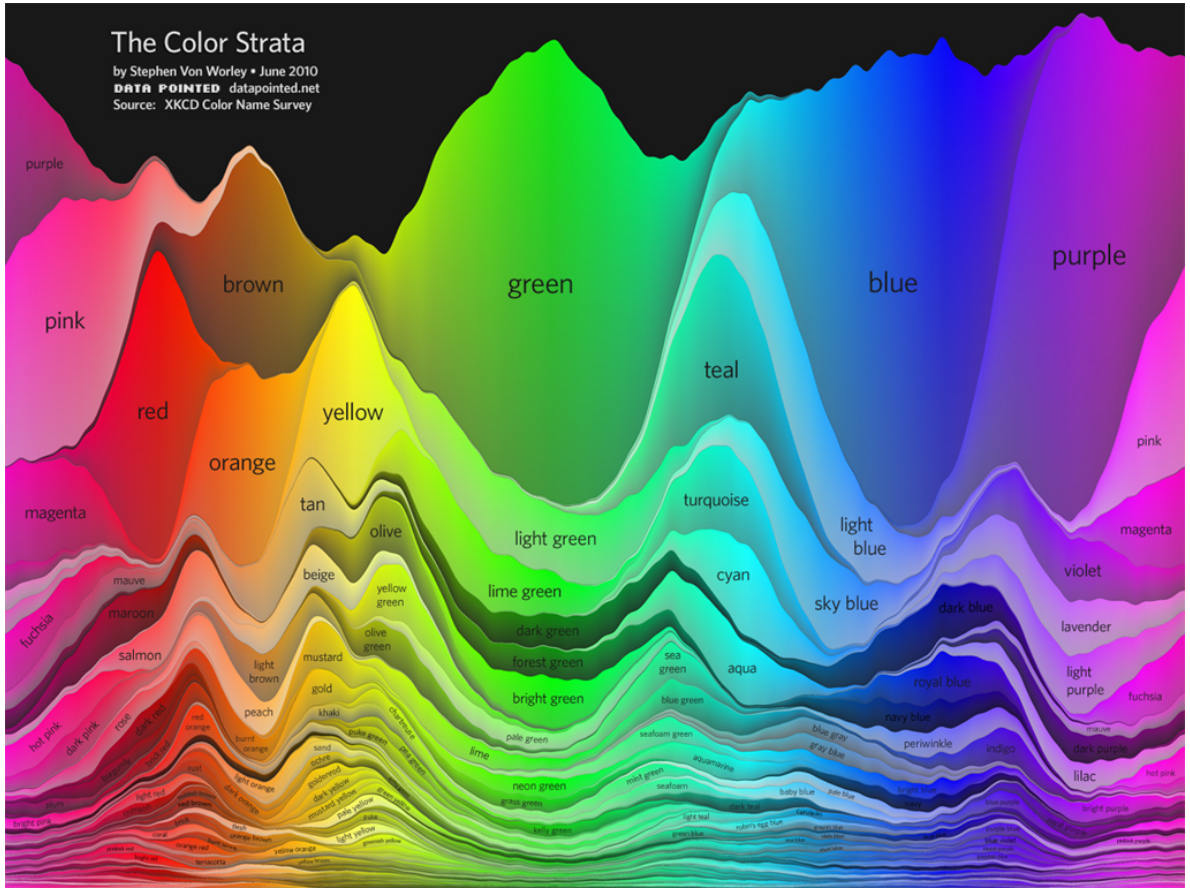
- Ambiguous
- Evolving
- Transmitted
- Learnable
- Complex semantics

What is language?

- Language is used to communicate
  - Things, actions, abstract concepts

What is language?

- Language puts categories on the world
  - It discretizes a continuous space

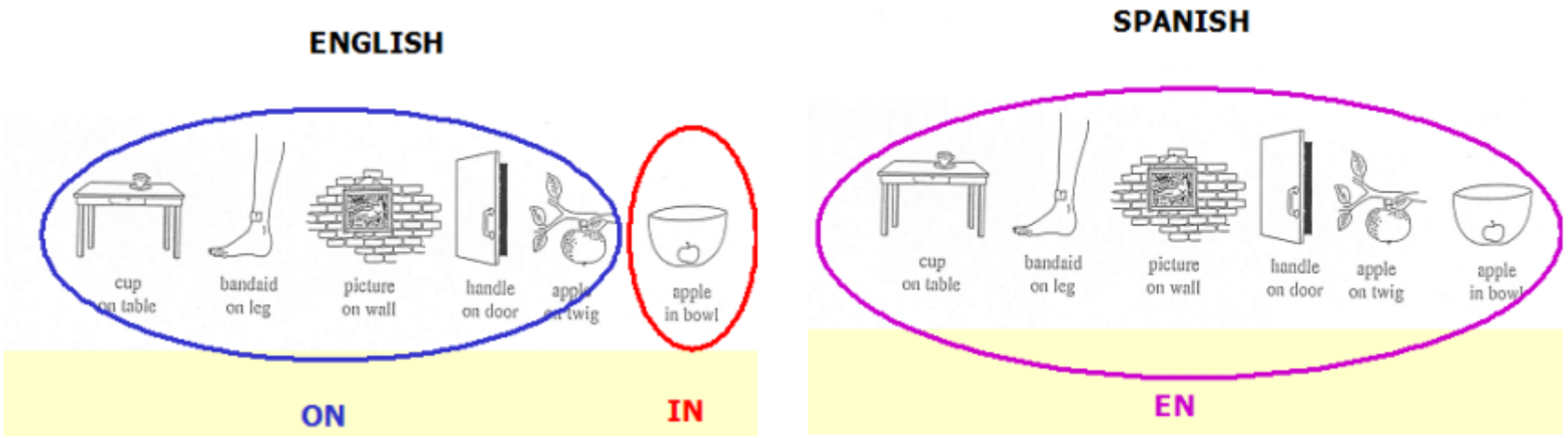


What is language?

- Language picks out what is salient and important
- What concepts do we have words for?
- Different languages have different discretization boundaries

What is language?

- Language picks out what is salient and important
- What concepts do we have words for?
- Different languages have different discretization boundaries

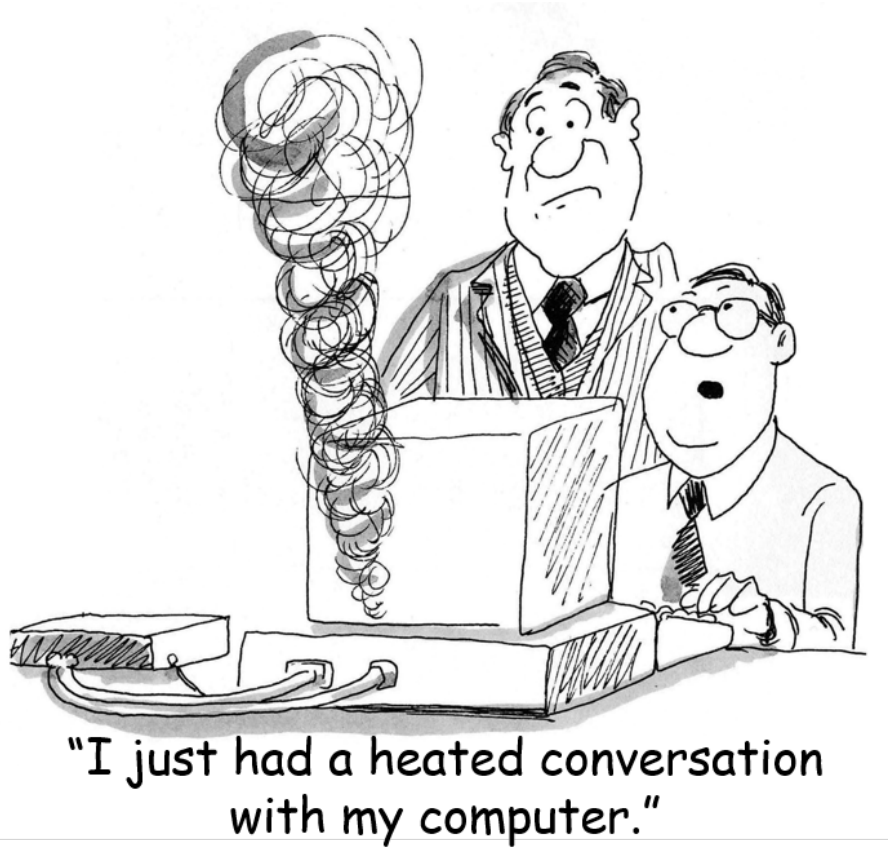


http://pyersqr.org/classes/Ling731/Space2.htm

Natural Language Processing

Building useful system to process language

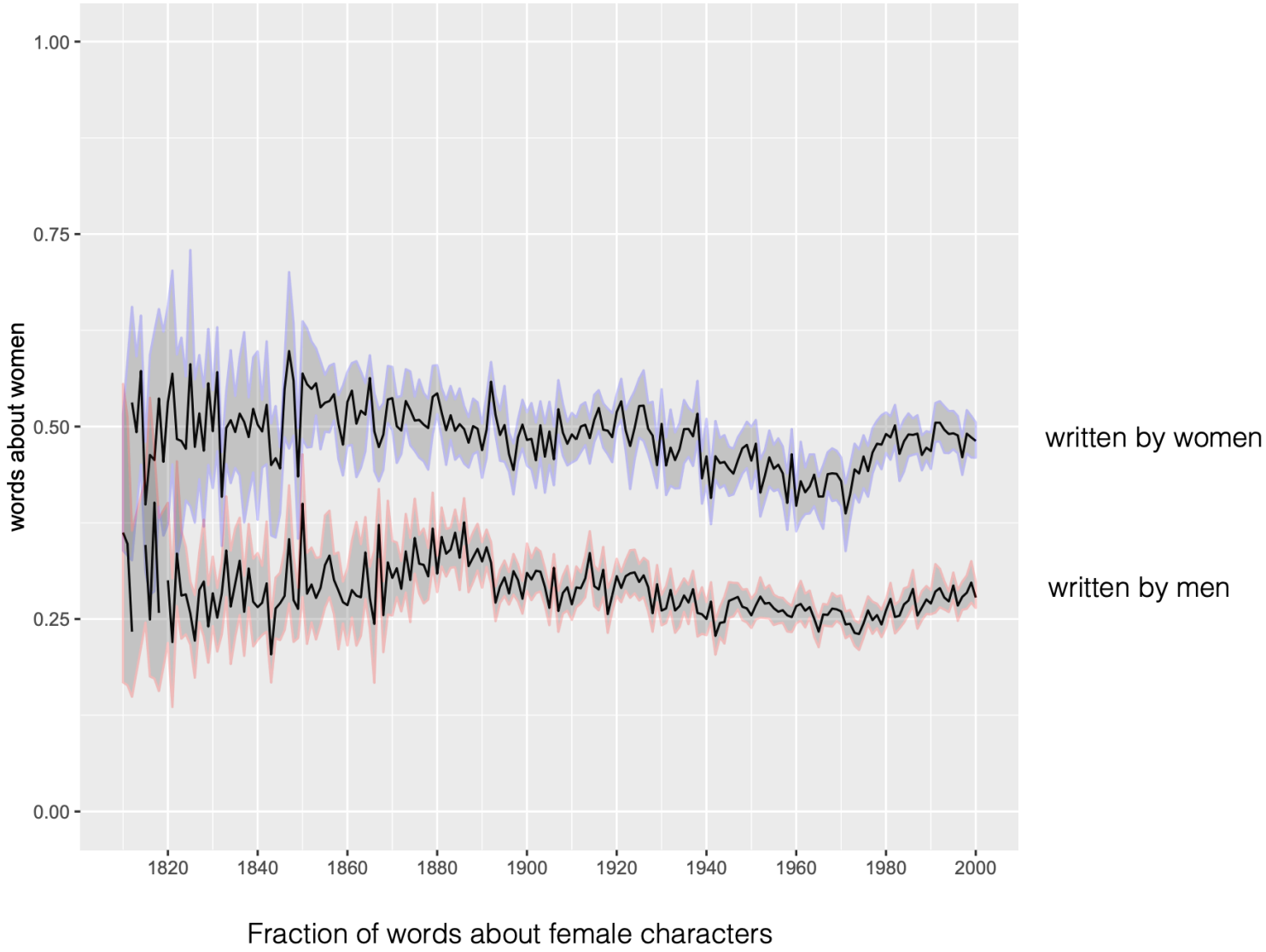
Computational Linguistics



(image credit: https://www.enterrasolutions.com/blog/computational-linguistics-and-natural-language-processing/)

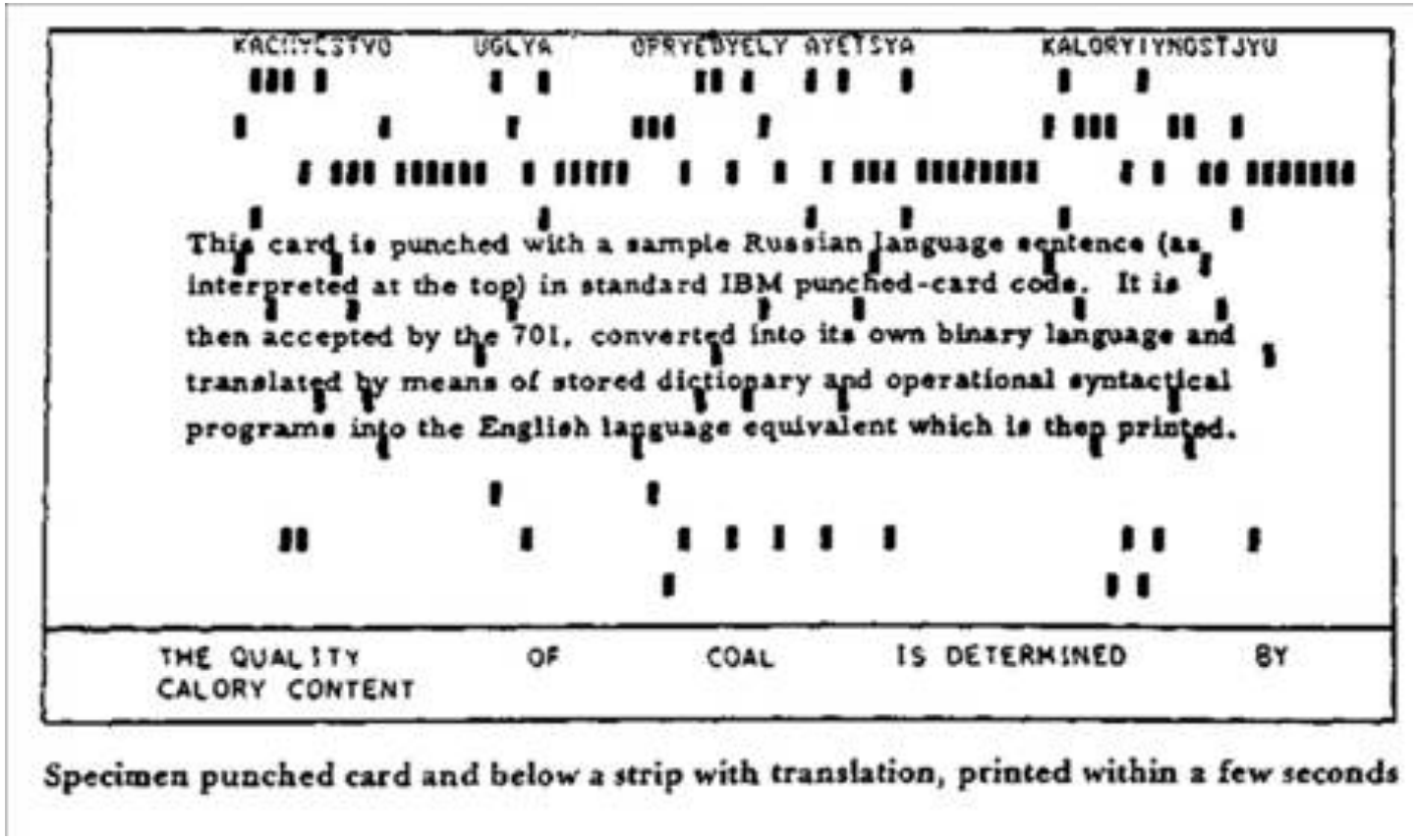
Using computers to study human language

Analyzing word usage in literature



Ted Underwood, David Bamman, and Sabrina Lee (2018),
"The Transformation of Gender in English-Language Fiction," Cultural Analytics

Beginnings



Specimen punched card and below a strip with translation, printed within a few seconds

Georgetown-IBM experiment, 1954

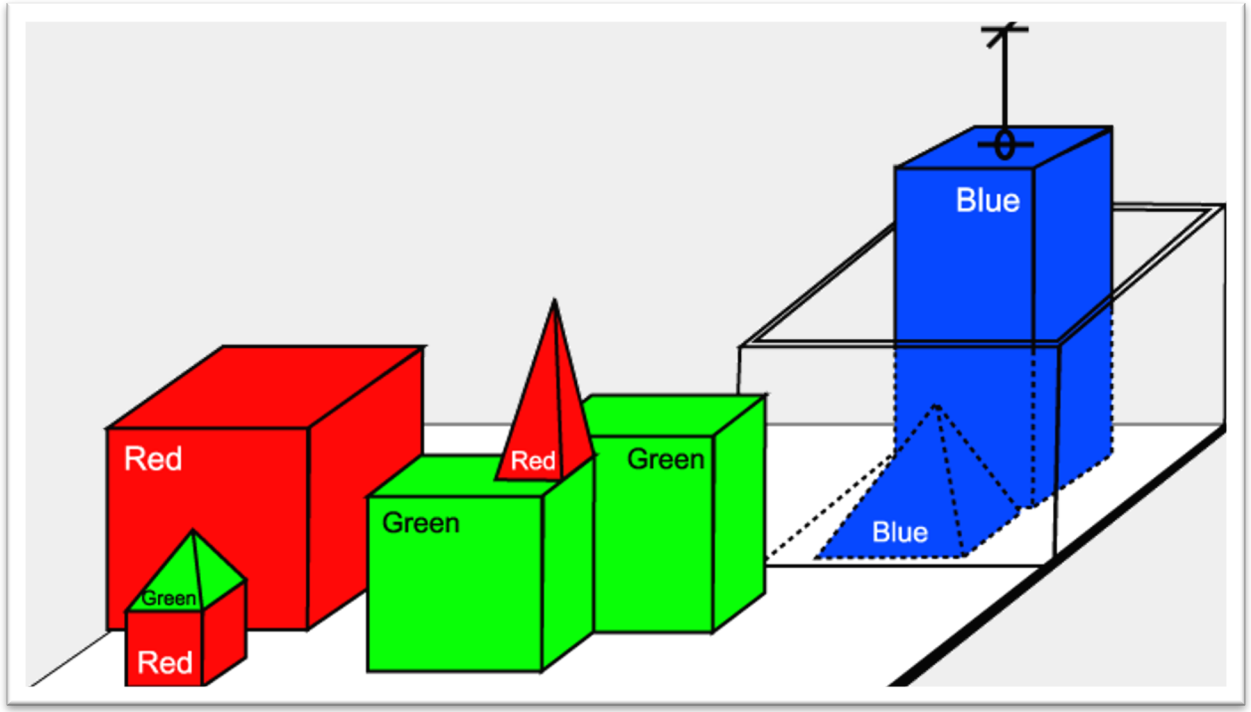
"Within three or five years, machine translation will be a solved problem"

SHRDLU (Winograd, 1968)

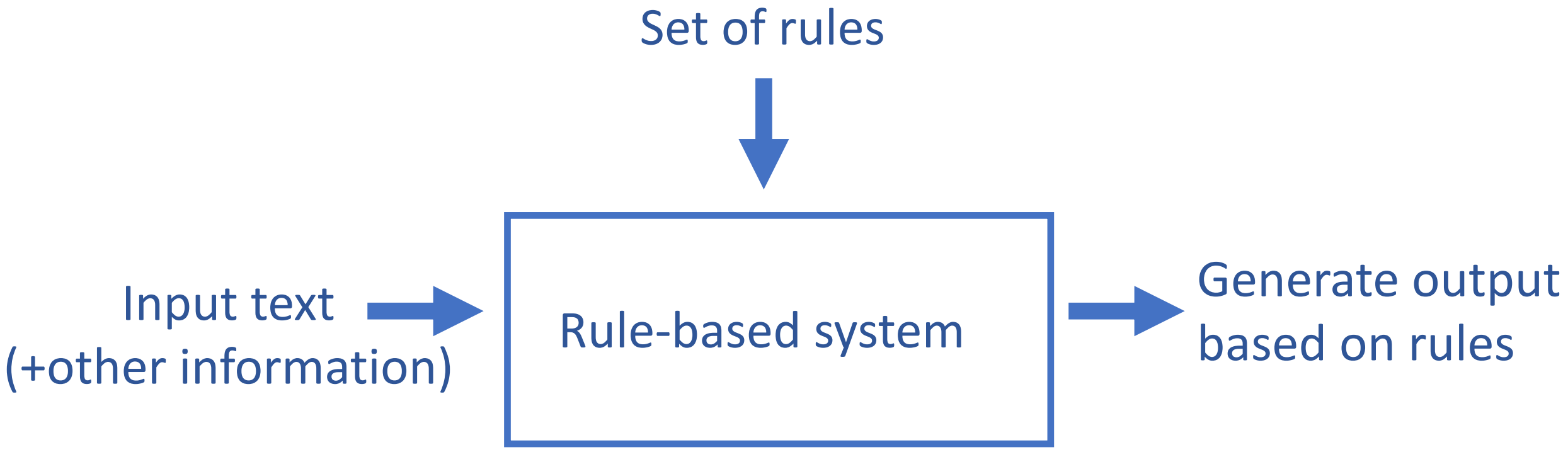
Video of actual system: https://www.youtube.com/watch?v=bo4RvYJYOzI

Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I don't understand which pyramid you mean.
Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
Computer: By "it", I assume you mean the block which is taller than the one I am holding.
Computer: OK.
Person: What does the box contain?
Computer: The blue pyramid and the blue block.
Person: What is the pyramid supported by?
Computer: The box.

Lots of rules!



Using rules for NLP



Eliza (Weizenbaum, 1966)

Demo: http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm

Men are all alike.
IN WHAT WAY

They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED

Rogerian psychologist:
reflect back what the patient said

- Set of ranked transformation rules based on keywords
- Apply rules based on matching words in sentence

Transformation rules

Keywords are ranked from specific to general

```
I know everybody laughed at me
```

- "I" is a very general keyword:

```
I: (I *) -> (You say you 2)
YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU
```

- "Everybody" is much more interesting (someone using universals like everybody/always is probably "referring to some quite specific event or person")

```
WHO IN PARTICULAR ARE YOU THINKING OF?
```

- Implementation: keywords stored with their rank

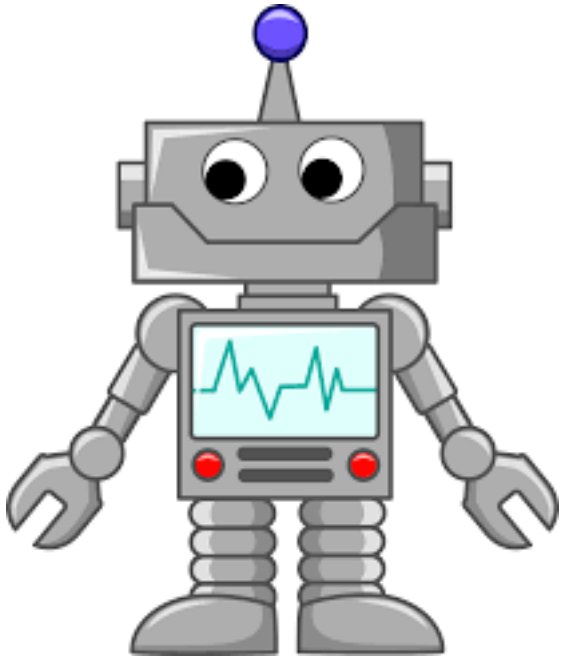
```
Everybody 5 (transformation rules)
I  0 (transformation rules)
```

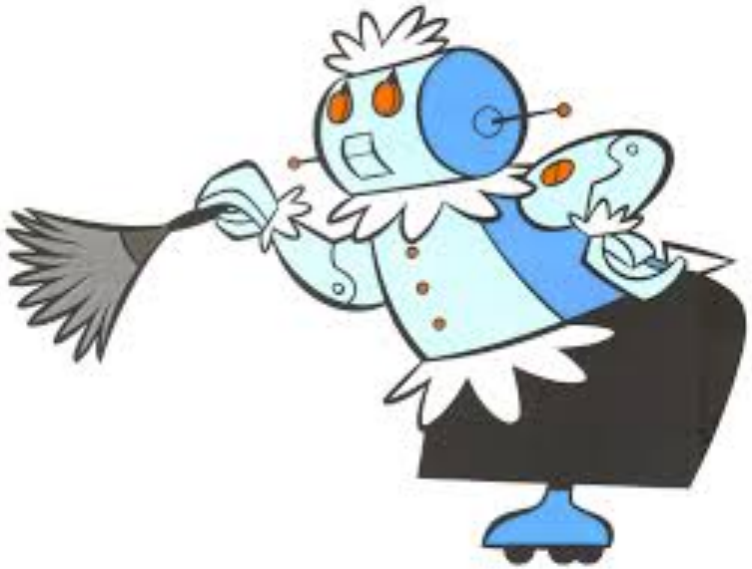
Backoff

```
Please go on
That's very interesting
I see
```

(Adapted from slides: Stanford CS124N, Dan Jurafsky)

Where is my block stacking or housekeeper robot that I can talk to?





Rosie from the Jetsons

The Far Side - Gary Larson

Understanding language is hard!

Some language humor

Kids make nutritious snacks

Stolen painting found by tree

Miners refuse to work after death

Squad helps dog bite victim

Killer sentenced to die for second time in 10 years

Lack of brains hinders research

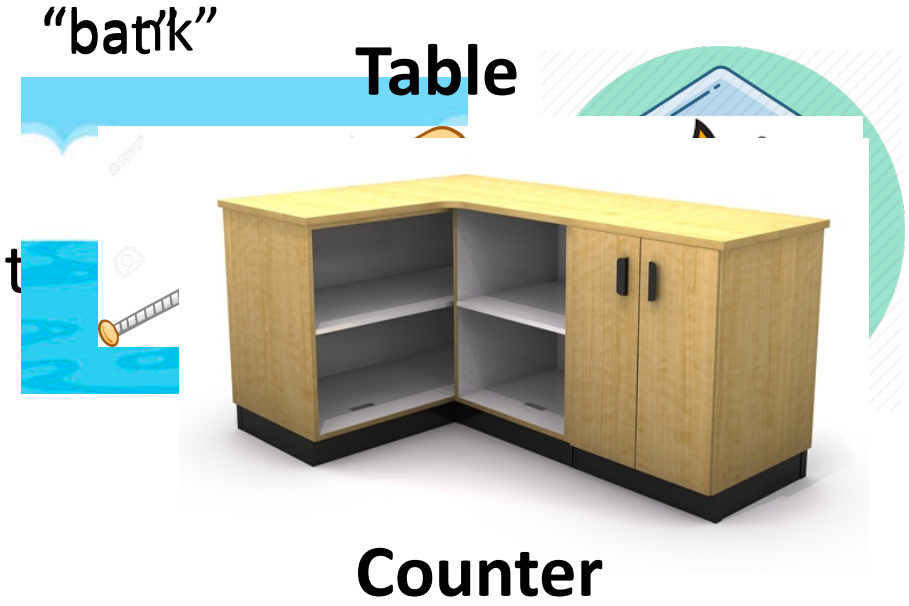
Real newspaper headlines!

Why is NLP hard?

Interpretation of language assumes a common basis of **world knowledge** and **context**

Herb Clark

- **Ambiguous**:
  - “bank”, “bat”
  - “Milk Drinkers Turn to Powder”
- **Synonyms**: Many ways to say same t
- **Context dependent**:
  - natural language is under-specified



Context-dependence

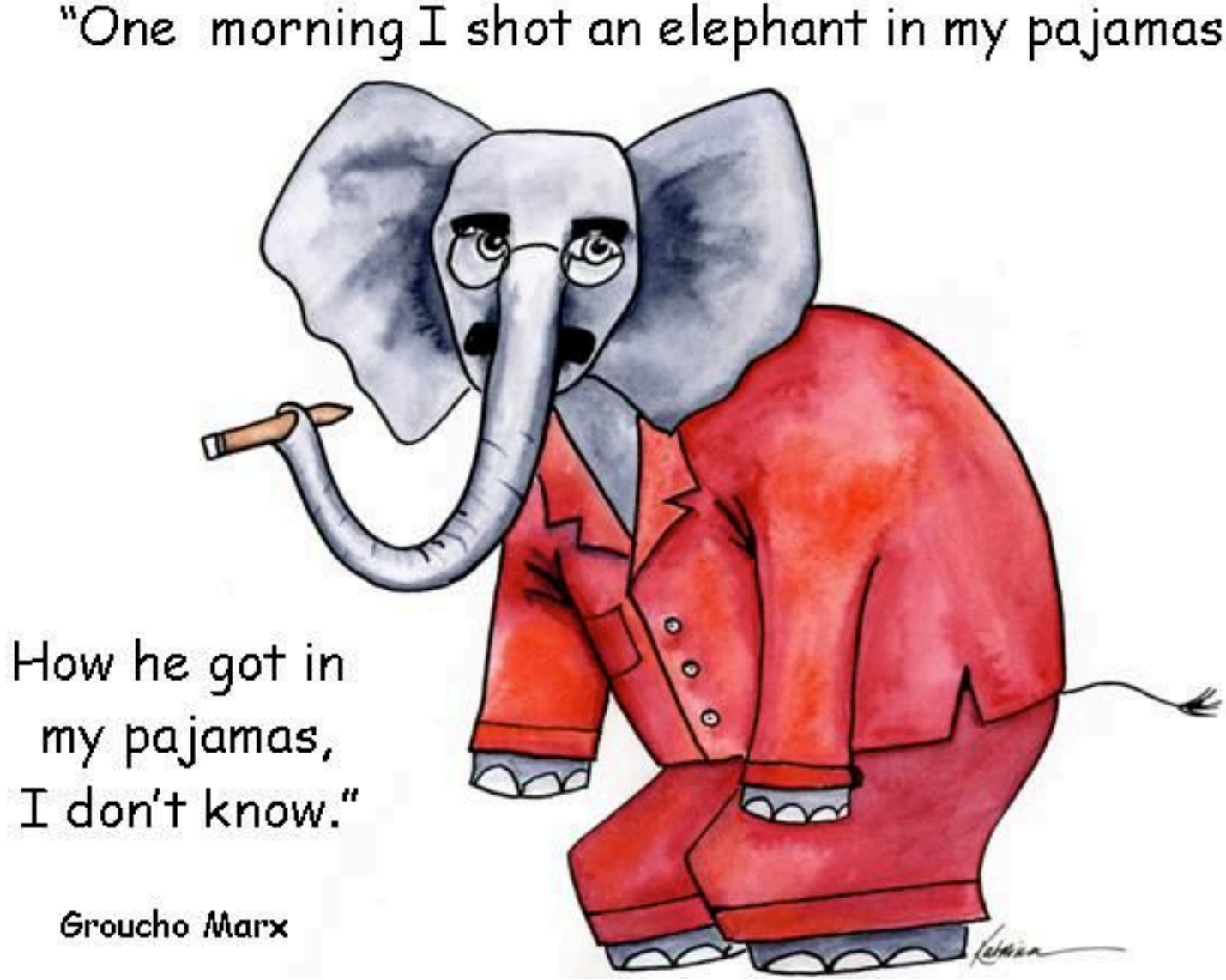
"I put the bowl on the **table**"

"The numbers in the **table** don't add up"

Table 8.2 Calculation of the χ^2 test on figures in table 8.1

Class (I)	Expected numbers		0 - E		$(0-E)^2/E$	
	A (2)	B (3)	A (4)	B (5)	A (6)	B (7)
I	11.80	10.20	5.20	-5.20	2.292	2.651
II	24.67	21.33	0.33	-0.33	0.004	0.005
III	39.15	33.85	-0.15	0.15	0.001	0.001
IV	48.81	42.19	-6.81	6.81	0.950	1.009
V	30.57	26.43	1.43	-1.43	0.067	0.077
Total	**30.57**	**134.00**	**0**	**0**	**3314**	**3.833**

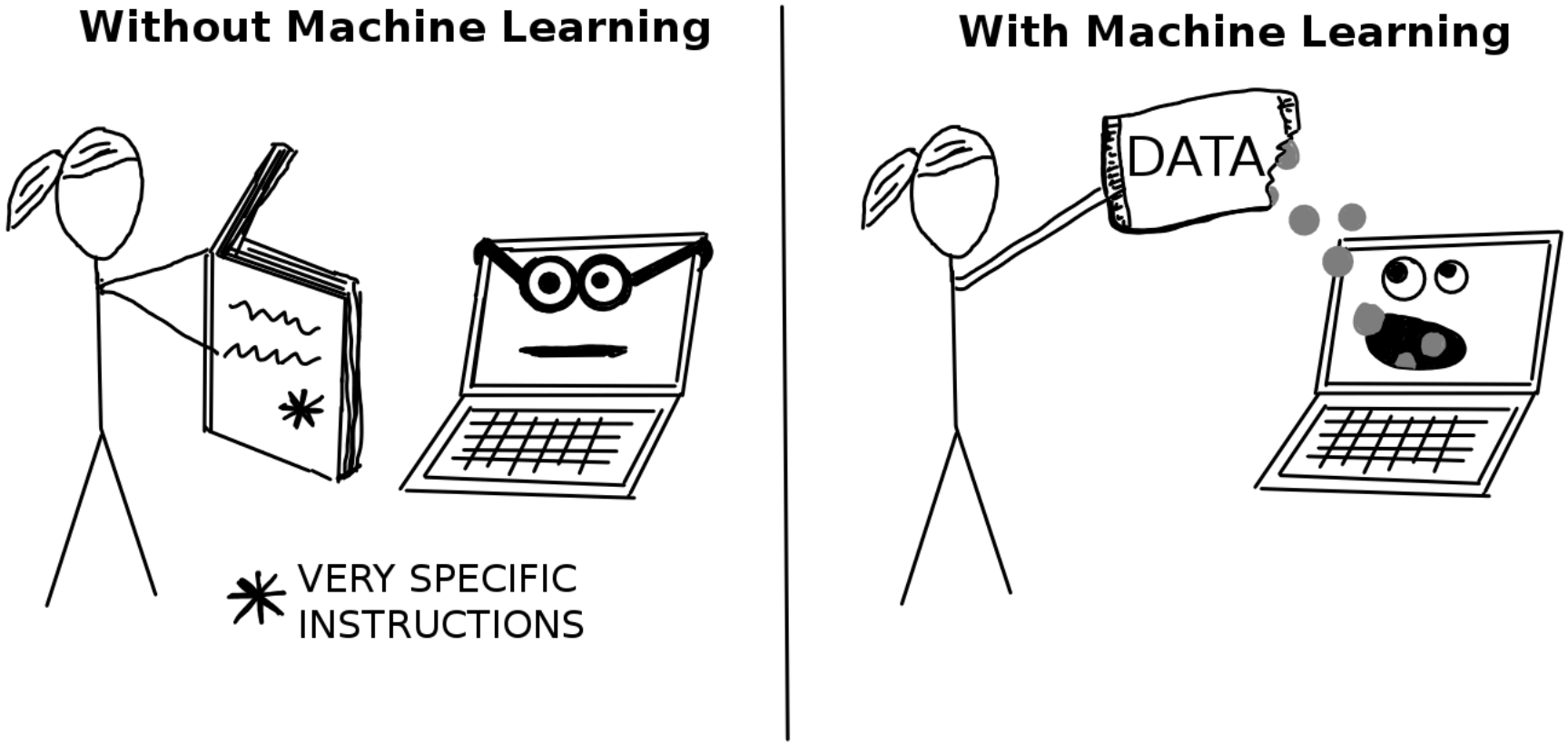
$g^2 = 3.314 + 3.833 = 7.147$, d.f. = 4, $0.10 < P < 0.50$.



https://www.katrinascards.com/product/elephant-my-pajamas-large-card

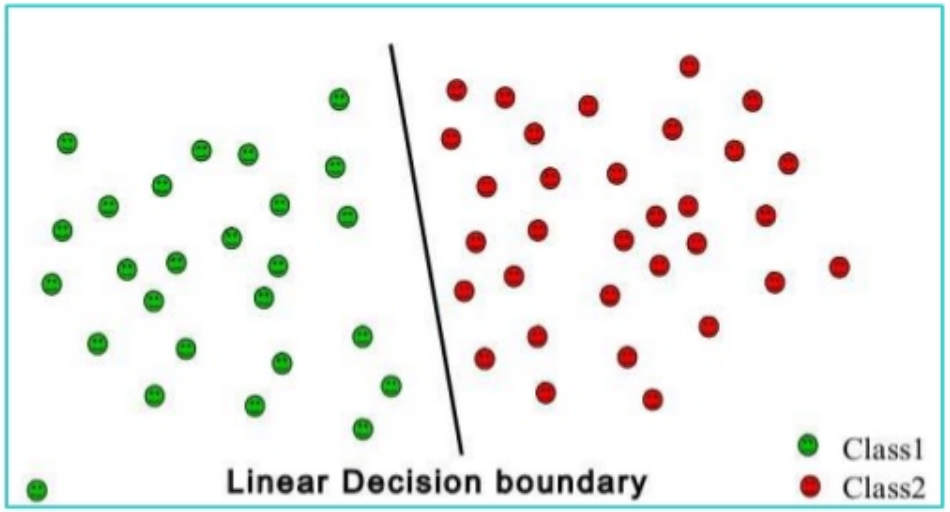
Coming up rules is hard!

Let’s learn from data!



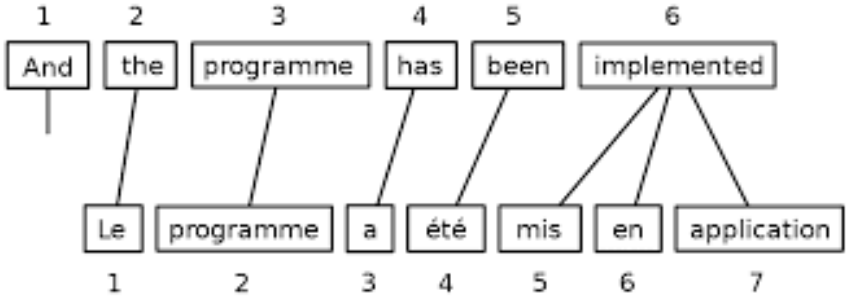
https://christophm.github.io/interpretable-ml-book/terminology.html

Rise of statistical learning

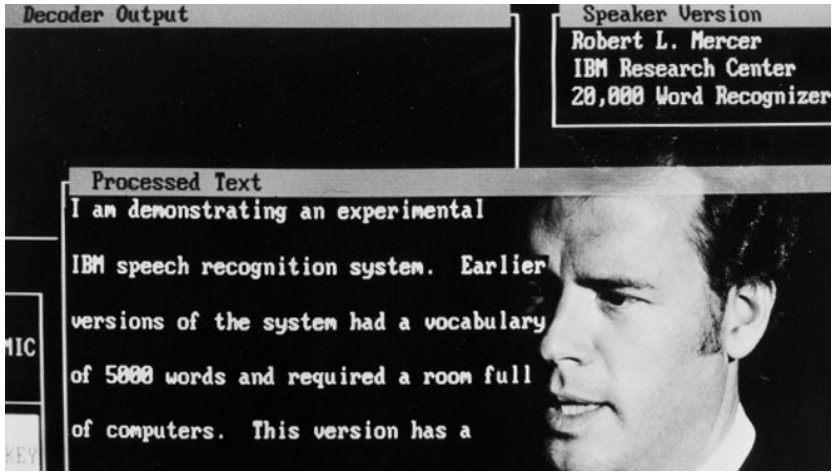


- Use of machine learning techniques in NLP
- Increase in computational capabilities
- Availability of electronic corpora

Rise of statistical learning



IBM Models
for translation



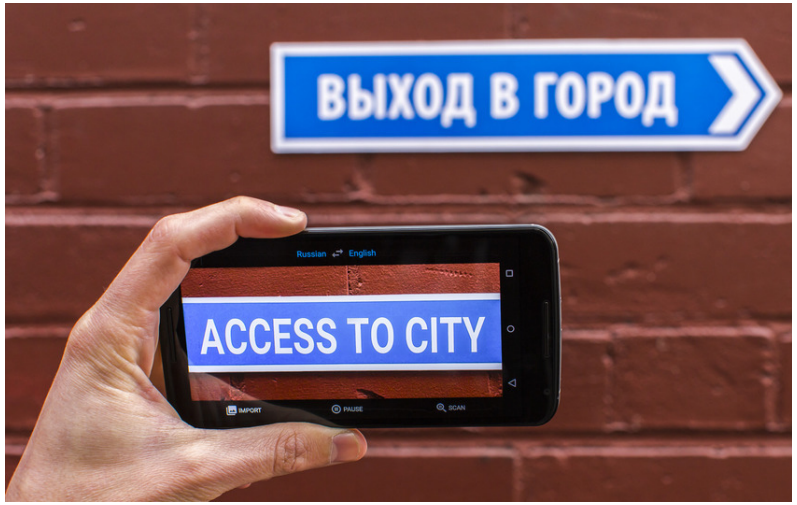
Speech
recognition

Anytime a linguist leaves the group the (speech) recognition rate goes up

- Fred Jelinek

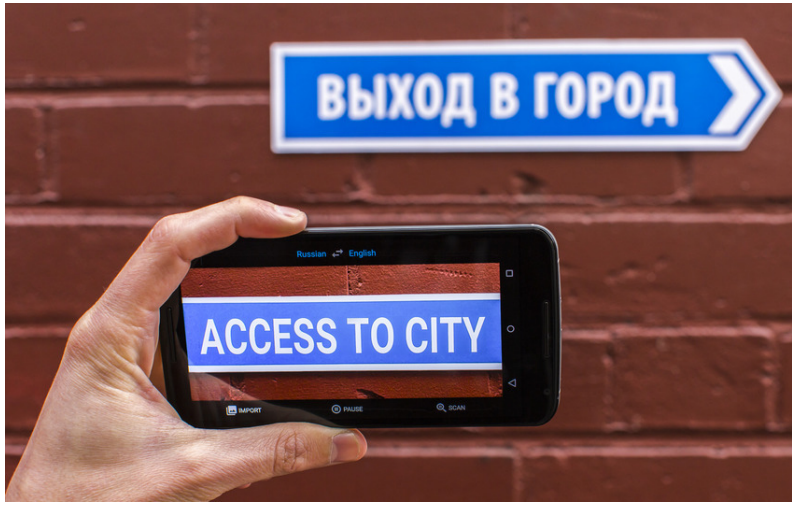
Deep learning era

- Significant advances in core NLP technologies



Deep learning era

- Significant advances in core NLP technologies
- Essential ingredient: large-scale supervision, lots of compute
- Reduced manual effort - less/zero feature engineering



36M sentence pairs

Russian: Машинный перевод - это круто!

English: Machine translation is cool!

Power of Data

CleverBot (2010)

User:	Do you like Doctor Who?
Cleverbot:	Yes, so funny.
User:	What's your favorite show?
Cleverbot:	I don't have one.
User:	What's your name?
Cleverbot:	Jason I already told you.
User:	No your real name.
Cleverbot:	Stephanie IS my real name.
User:	I thought it was Jason?
Cleverbot:	How did you come to that conclusion.
User:	Because you said so.
Cleverbot:	I have to do my homework.

https://www.cleverbot.com/

How it works:

- Corpus of conversational turns
- Find the most similar sentence and copy the response
- Learn from human input

What do you get?

- Something that someone say
- Incoherent conversation

Power of Data

Meena (Google, 2020)

Conversation D

Human:	*I like how cute cows are*
Meena:	*Ah, interesting. You like how they are cute, right?*
Human:	*yeah, and they're realy smart*
Meena:	*I heard they go to college*
Human:	*cows go to college?*
Meena:	*I heard that a cow went to Harvard.*
Human:	*what did the cow study?*
Meena:	*Bovine sciences.*
Human:	*do horses go to Harvard?*
Meena:	*Horses go to Hayvard.*
Human:	*that's a pretty good joke, I feel like you led me into it*
Meena:	*You were trying to steer it elsewhere, I can see it.*

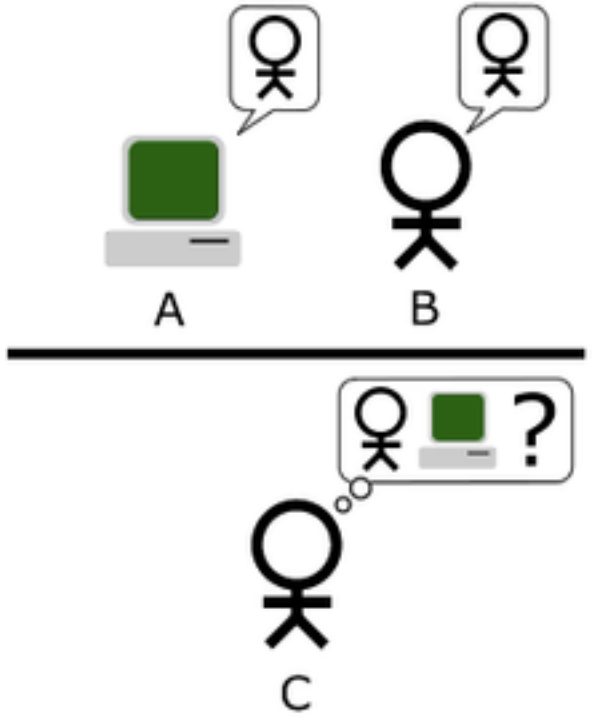
How it works:

- Corpus of conversational turns (over 40B words)
- Train huge neural network (2.6 billion parameters) for 30 days on 2048 TPUs cores
- Predict response given a sentence

https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html

Turing Test

Imagine an "**Imitation Game**," in which a man and a woman go into separate rooms and guests try to tell them apart by writing a series of questions and reading the typewritten answers sent back. In this game both the man and the woman aim to convince the guests that they are the other.



Can you guess:

Computer or human?

Alan Turing

We now ask the question, **"What will happen when a machine takes the part of A in this game?"** Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, **"Can machines think?"**

Turing test solved?

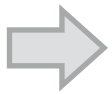
Talking to Google Duplex: Google's human-like phone AI feels revolutionary

Believe the hype—Google's phone-call bot is every bit as impressive as promised.

https://www.youtube.com/watch?v=D5VN56jQMWM&feature=youtu.be&t=70

Information Extraction

The Massachusetts Institute of Technology (MIT) is a private research university in Cambridge, Massachusetts, often cited as one of the world's most prestigious universities.
Founded in 1861 in response to the increasing industrialization of the United States, …



***City**: Cambridge, MA*
***Founded**: 1861*
***Mascot:** Tim the Beaver*
…

Article

Database

Information Extraction: State of the Art

Dependence on large training sets

ACE: 300K words

Freebase: 24M relations

Not available for many domains (ex. medicine, crime)

Challenging task: even large corpora do not guarantee high performance

~ 75% F1 on relation extraction (ACE)

~ 58% F1 on event extraction (ACE)

Machine Translation

BBC Sign in 选项（英文） 检索

NEWS | 中文 繁

主页 | 国际 | 两岸 | 英国 | 评论 | 科技 | 财经 | 图辑 | 音频材料 | 视频材料 | BBC英伦网

巴拿马首任驻华大使专访：与台湾断交之后

他说，不认同中国“买走”台湾邦交国的说法，与中国建立关系对巴拿马有利，不担心影响与美国关系。

1小时前

巴拿马外交转向周年 中美大国博弈内幕解密

尼加拉瓜运河成谜：人走楼空 “不再提及”

触发萨尔瓦多与台湾断交的港口令美国担忧

观点：高铁“一地两检”——中国的强势 港人的无力

江沂：两地价值观和制度差异巨大，冲突无可避免。香港没有反对的权利，也没有反对的能力。

范冰冰消失百日后 中国娱乐业的寒蝉效应

日前发布的《中国影视明星社会责任研究报告》中，最高分徐峥为78分，最低分范冰冰为0分。

2小时前

中蒙参加俄罗斯军演 “中俄靠拢论”再吸眼球

2018年 9月 11日

日本人脚踹慰安妇铜像引爆台湾人抗议

2018年 9月 11日

大空望远镜探测比“三

特别推荐

人民币贬值、楼价高涨：消费降级中产失去的优质生活

广告

台湾民间发起东京奥运“正名”公投的意义

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Table 10: Mean of side-by-side scores on production data

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

(Wu et al., 2016)

Machine Translation

BBC Sign in Option (English) 检索

NEWS | 中文 Traditional

Homepage | International | Cross-strait | United Kingdom | comment | Technology | Finance | Picture series | Audio material

Interview with Panama's first ambassador to China: After breaking diplomatic relations with Taiwan

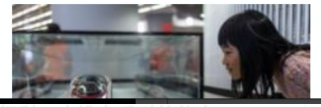
He said that he does not agree with China's saying that "buy" Taiwan's diplomatic relations with China. Establishing relations with China is beneficial to Panama and does not worry about affecting relations with the United States.

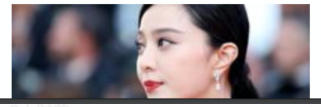
1 hour ago

Panamanian Diplomacy Turns to Anniversary

The Nicaragua Canal is a mystery: people go to the floor and "no longer mention"

The port that triggered the break of El Salvador and Taiwan has worried the United States





China and Mongolia participate in the Russian military

Special recommendation

Renminbi depreciation, property prices are rising: consumption downgrades the loss of quality life in the middle class

ADVERTISING

The significance of the Taiwanese people's "referred to" referendum

Machine comprehension

Amazon_rainforest

The Stanford Question Answering Dataset

The Amazon rainforest (Portuguese: Floresta Amazônica or Amazônia; Spanish: Selva Amazónica, Amazonía or usually Amazonia; French: Forêt amazonienne; Dutch: Amazoneregenwoud), also known in English as Amazonia or the Amazon Jungle, is a moist broadleaf forest that covers most of the Amazon basin of South America. This basin encompasses 7,000,000 square kilometres (2,700,000 sq mi), of which 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. This region includes territory belonging to nine nations. The majority of the forest is contained within Brazil, with 60% of the rainforest, followed by Peru with 13%, Colombia with 10%, and with minor amounts in Venezuela, Ecuador, Bolivia, Guyana, Suriname and French Guiana. States or departments in four nations contain "Amazonas" in their names. The Amazon represents over half of the planet's remaining rainforests, and comprises the largest and most biodiverse tract of tropical rainforest in the world, with an estimated 390 billion individual trees divided into 16,000 species.

Which name is also used to describe the Amazon rainforest in English?
Ground Truth Answers: also known in English as Amazonia or the Amazon Jungle, | Amazonia or the Amazon Jungle | Amazonia
Prediction: Amazonia

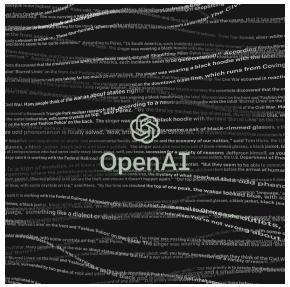
How many square kilometers of rainforest is covered in the basin?
Ground Truth Answers: 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. | 5,500,000 | 5,500,000
Prediction: 5,500,000

How many nations control this region in total?
Ground Truth Answers: This region includes territory belonging to nine nations. | nine | nine
Prediction: nine

How many nations contain "Amazonas" in their names?
Ground Truth Answers: States or departments in four nations contain "Amazonas" in their names. | four | four
Prediction: four

What percentage does the Amazon represents in rainforests on the planet?

Language generation



Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

With the start of the new academic year, Simon Fraser University has an opportunity to help provide a new generation of women with a diverse set of academic resources for higher education.

These resources should help make it more accessible to female students who identify as Indigenous, LGBTIQ, women of colour or belonging to a racialized group. A key part of this goal is to create a meaningful opportunity for students to discover, cultivate and share knowledge across different Indigenous and intersectional identities.

https://talktotransformer.com/

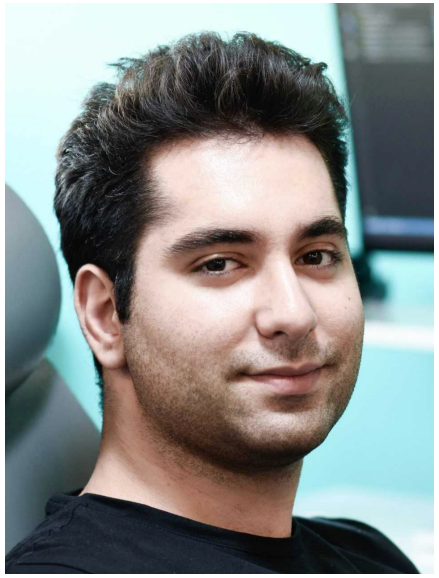
Course Logistics

Teaching Staff

Instructor

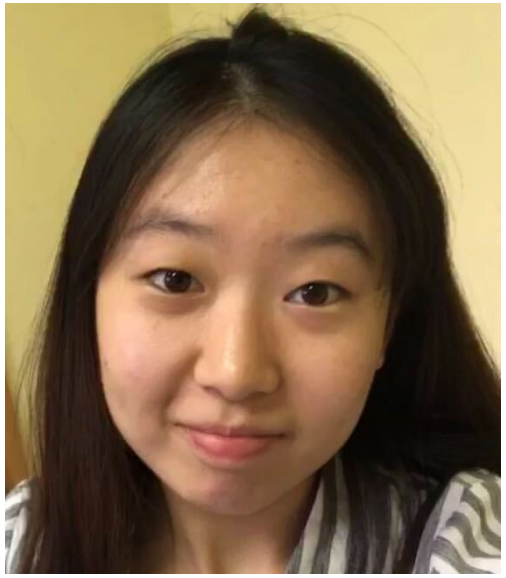
Angel Chang

TAs



Ali Gholami

Sonia Raychaudhuri



Yue Ruan

Resources

- Website: https://angelxuanchang.github.io/nlp-class/

- Lectures (using Canvas BB Collaborate Ultra)
  - Wednesday 11:30 - 12:20pm
  - Friday 10:30 - 11:45am
  - Additional video lecture

- TA lead tutorials (optional)
  - 30 minute video
  - Interactive session: Friday 11:50 - 12:20pm

- Sign up on Piazza for discussion: piazza.com/sfu.ca/fall2020/cmpt413825

Background / Prerequisites

- Proficiency in Python - Programming assignments will be in python, numpy and pytorch will be used.
- Calculus and Linear Algebra (MATH 151, MATH 232/240) - You will need to be comfortable with taking multivariable derivatives
- Basic Probability and Statistics (STAT 270)
- Basic Machine Learning (CMPT 419/726)

There will be optional tutorials that will help review these topics.

Grading

- Assignments (62%)
- Class project (35%)
- Participation (3%)
  - Answering questions on Piazza
  - Discussion in class

Assignments (62%)

- 4 assignments consisting of two parts
  - 5% - Answering questions (individual)
  - 10% - Programming assignment (group)
  - Released every two weeks (Due 11:59pm Wednesday)

- Initial getting started assignment (HW0)
  - Find your groups and setup (1%)
    - Groups should be 2-4 people
  - Review of fundamentals (1%)
    - Probability, Linear Algebra and Calculus
  - Due Wednesday 9/16, 11:59pm

Class Project (35%)

- Project should be a mini-research project. It can be:
  - Re-implementation of a recent NLP paper
  - Experimental comparison of several methods
  - More details later in the term
- Team of 2-4 students (same as HW groups)
  - Larger group should have a more substantial project
- Graded components
  - Proposal (5%)
  - Milestone (5%)
  - Project ``poster'' presentation (5%) - online, details TBD
  - Final report (20%)

Outline

- **Words**
  - Language models
  - Text classification
  - Word embeddings
- **Sequences, structures, and context**
  - Sequence modeling
  - Syntactic parsing
  - Sequence to sequence models and text generation
  - Contextual word embeddings
- **Applications**
  - Coreference resolution
  - Question answering
  - Dialogue
  - Multimodal NLP

Upcoming

- Video lecture on levels of representations:
  - Phonology, morphology, syntax, semantics, pragmatics and discourse
- Tutorial on Probability, Linear Algebra and Calculus
- Language Modeling
- HW0 (2%) due next week: 9/16