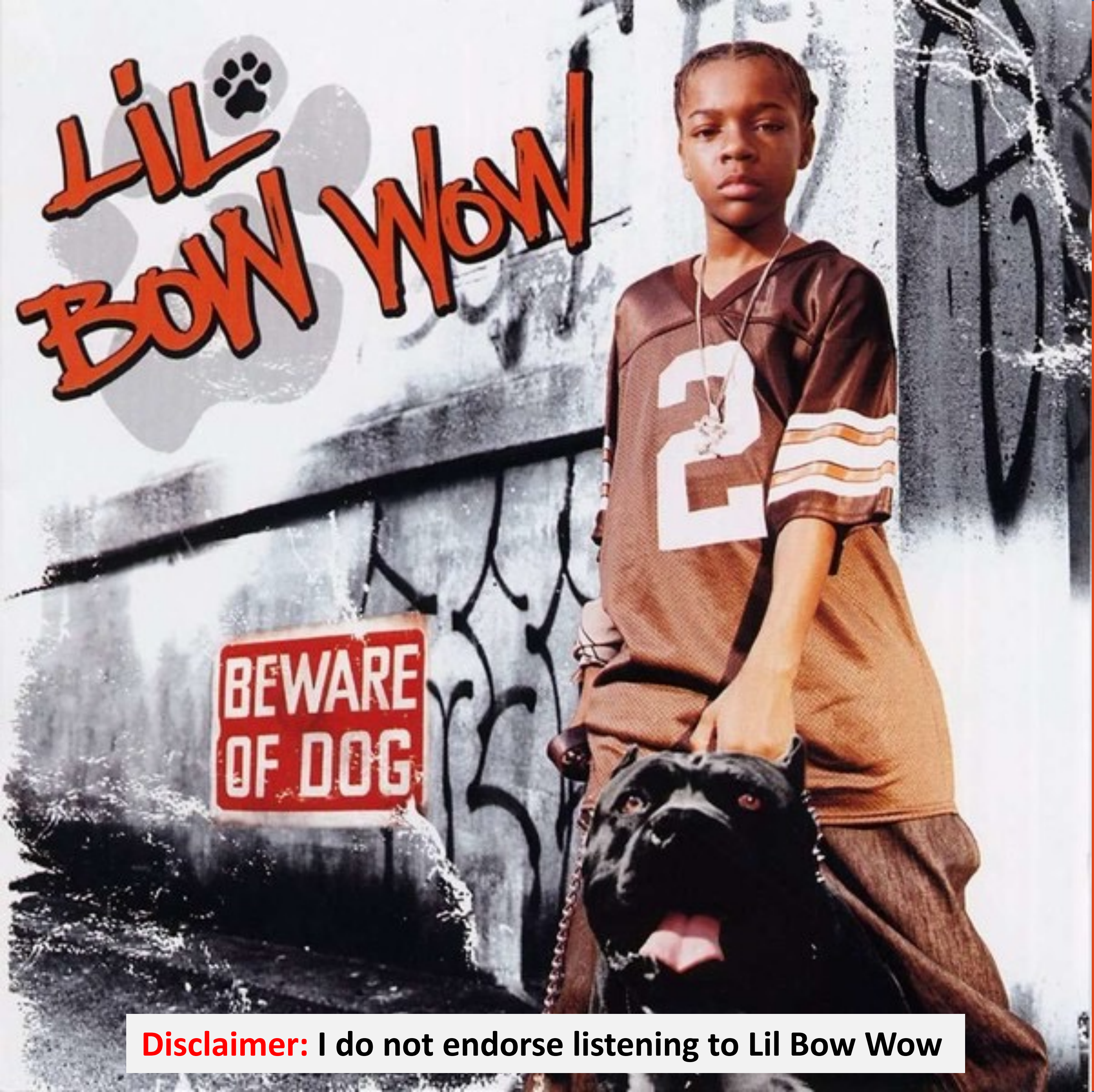
# Lecture 2: Language Representations

What is NLP + How to represent language


**Harvard**

AC295/CS287r/CSCI E-115B

Chris Tanner




**Disclaimer:** I do not endorse listening to Lil Bow Wow

# Lil Bag-of-Words Wow (rapper)

---

**Shad Gregory Moss** (born March 9, 1987), better known by his stage name **Bow Wow** (formerly **Lil' Bow Wow**), is an American rapper, actor, and television presenter. Moss' career began upon being discovered by rapper Snoop Dogg in the late 1990s, eventually being brought to record producer Jermaine Dupri and signed to So So Def Recordings. As Lil' Bow Wow, he released his first album at age 13, *Beware of Dog*, in 2000, which was followed by *Doggy Bag* a year later.

He has released six studio albums, twenty-six singles, fifty-one music videos, and eight mixtapes.

In his career, Bow Wow has had a total of twelve top 40 singles (three of which were top ten hits) on the US *Billboard* Hot 100 chart. He has sold over 10 million copies and 14 million digital assets worldwide.[1]

-- https://en.wikipedia.org/wiki/Bow_Wow_(rapper)

# ANNOUNCEMENTS

- **Attendance** is checked today. See a TF before you leave today.
- **HW1** was released at midnight. Due in 2 weeks (**Mon @ 11:59pm**). Start now.
- **PyTorch tutorial** will be tonight @ 6pm, in this room
- **Lectures slides** will be posted on the website and our Twitter @CS287_NLP
- **Office Hours** start tomorrow, Wednesday @ 5pm (see website for all OH)
  - **Location**: out back of SEC 1st floor, or SEC 3.301-3.303 if weather isn't good

# Outline

- NLP: what and why?
- Representing Language
- Bag-of-Words
- TF-IDF

# Outline

- NLP: what and why?
- Representing Language
- Bag-of-Words
- TF-IDF

# Language

Language symbols are encoded as continuous communication signals, and are invariant across different encodings (same underlying concept, different surface forms)





Slide adapted from or inspired by Chris Manning and Richard Socher

# Language is funny

"Red tape holds up new bridges"

"Hospitals are sued by 7 foot doctors"

"Local high school dropouts cut in half"

"Tesla crashed today"

"Obama announced that he will run again"

"Kipchoge announced that he will run again"

"She made him duck"

"Will you visit the bank across from the river bank? You can bank on it"

"Yes" vs "Yes." vs "YES" vs "YES!" vs "YAS" vs "Yea"

# Language is funny

"Maria likes May"

"Maria likes May and Joe"

"Maria likes May and June"

"May likes Maria"

"Maria hit May, then she [fell/ran]"

"Maria and Anqi bullied May, so they got in trouble"

"Maria and Anqi convinced May to prank the teacher, so they got in trouble"

"May may like May, but she really likes June."

# Multiple levels* to a single word

Discourse

Pragmatics

Semantics

Syntax

Lexemes

Morphology

phonology

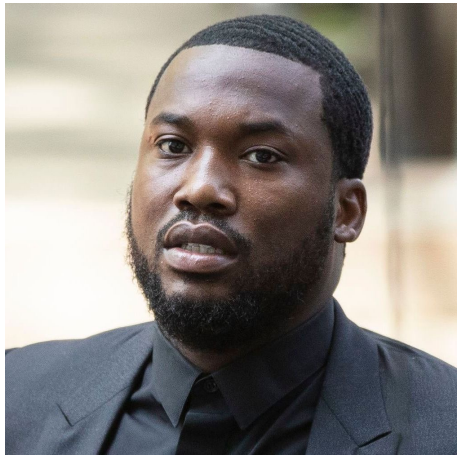
phonetics

speech

orthography

text

*





Slide adapted from or inspired by Chris Manning and Richard Socher

# Multiple levels* to a single word

Discourse

Pragmatics

Semantics

Syntax

Lexemes

Morphology

phonology

phonetics

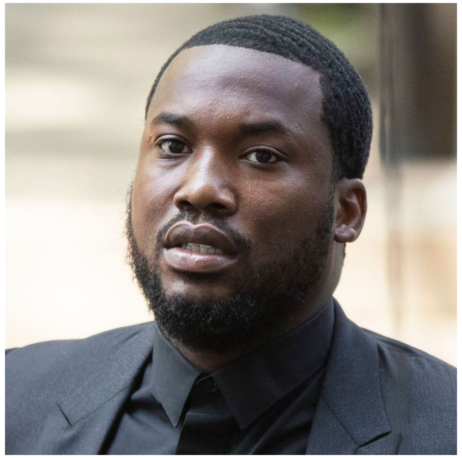
speech

orthography

text

- Inputs (words) are noisy
- Capture theoretical concepts that are ~latent variables
- Ambiguity abound. Many interpretations at each level

*





Slide adapted from or inspired by Chris Manning and Richard Socher

# Multiple levels* to a single word

Discourse

Pragmatics

Semantics

Syntax

Lexemes

Morphology

phonology

phonetics

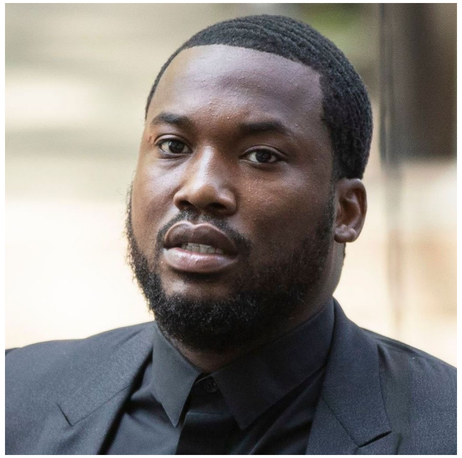
speech

orthography

text

- Humans are very good at resolving linguistic ambiguity (e.g., coreference resolution)
- Computer models aren't

*





Slide adapted from or inspired by Chris Manning and Richard Socher

## Multiple levels* to a single word

Discourse

Pragmatics

Semantics

Syntax

Lexemes

Morphology

phonology

phonetics

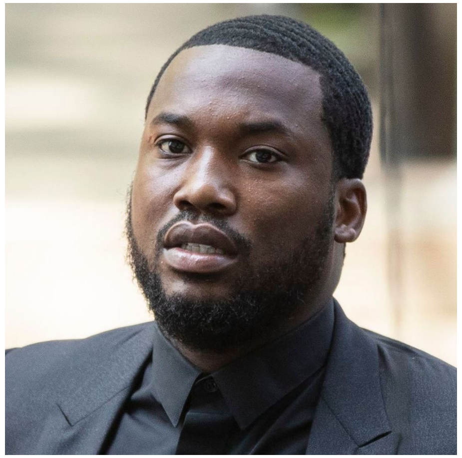
speech

orthography

text

- Many ways to express the same meaning
- Infinite meanings can be expressed
- Languages widely differ in these complex interactions

*





Slide adapted from or inspired by Chris Manning and Richard Socher

## Multiple levels* to a single word

Discourse

- Many ways to express the same meaning
- Infinite meanings can be

ese

**The study of how sub-components form meaning**

(e.g., running, deactivate, Obamacare, Cassandra's)

Morphology

phonology

orthography

phonetics

text

speech

# Multiple levels* to a single word

- Many ways to express the same meaning

these

**Lexical analysis; normalize and disambiguate words**

(e.g., bank, mean, hand it to you, make up, take out)

Lexemes

Morphology

phonology

orthography

phonetics

text

speech

*

Multiple levels* to a single word

Discourse

Pragmatics

Semantics

Syntax

**Transform a sequence of characters into a hierarchical/compositional structure**

(e.g., students hate annoying professors; Mary saw the old man with a telescope)

pho

pho

speech

- Many ways to express the same meaning
- Infinite meanings can be expressed
- Languages widely differ in these complex interactions

# Multiple levels* to a single word

Discourse

Pragmatics

Semantics

**Determines meaning**

(e.g., NLU / intent recognition; natural language inference; summarization; question-answering)

pho

phonetics

speech

text

- Many ways to express the same meaning
- Infinite meanings can be expressed
- Languages widely differ in these complex interactions

## Multiple levels* to a single word

Discourse

Pragmatics

Semantics

- Many ways to express the same meaning
- Infinite meanings can be expressed
- Languages widely differ in these

**Understands how context affects meaning**

(i.e., not only concerns how meaning depends on structural and linguistic knowledge (grammar) of the speaker, but on the context of the utterance, too)

phonology

phonetics

speech

orthography

text

# Multiple levels* to a single word

- Many ways to express the same meaning
- Infinite meanings can be expressed

Discourse

**Understands structures and effects of interweaving dialog**

(i.e., Jhene tried to put the trophy in the suitcase but **it** was too big. She finally got **it** to close.)

Morphology

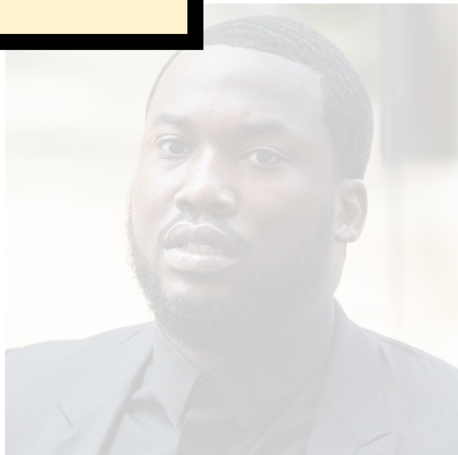
phonology

orthography

phonetics

text

speech



# Common NLP Tasks (aka problems)

## Syntax

Morphology

Word Segmentation

Part-of-Speech Tagging

Parsing

- Constituency
- Dependency

## Discourse

Summarization

Coreference Resolution

## Semantics

Sentiment Analysis

Topic Modelling

Named Entity Recognition (NER)

Relation Extraction

Word Sense Disambiguation

Natural Language Understanding (NLU)

Natural Language Generation (NLG)

Machine Translation

Entailment

Question Answering

Language Modelling

# Common NLP Tasks (aka problems)

## Syntax

Morphology

Word Segmentation

Part-of-Speech Tagging

Parsing

- Constituency
- Dependency

## Discourse

Summarization

Coreference Resolution

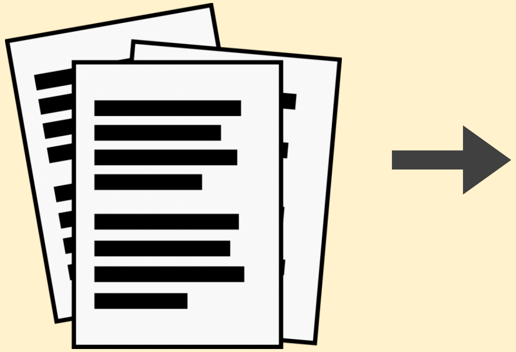
## Semantics

Sentiment Analysis

Topic Modelling

Question Answering

Language Modelling



"Overall, Pfizer's COVID-19 vaccine is very safe and one of the most effective vaccines ever produced"

# Common NLP Tasks (aka problems)

## Syntax

Morphology

Word Segmentation

Part-of-Speech Tagging

Parsing

Constituenc

Dependenc

## Discourse

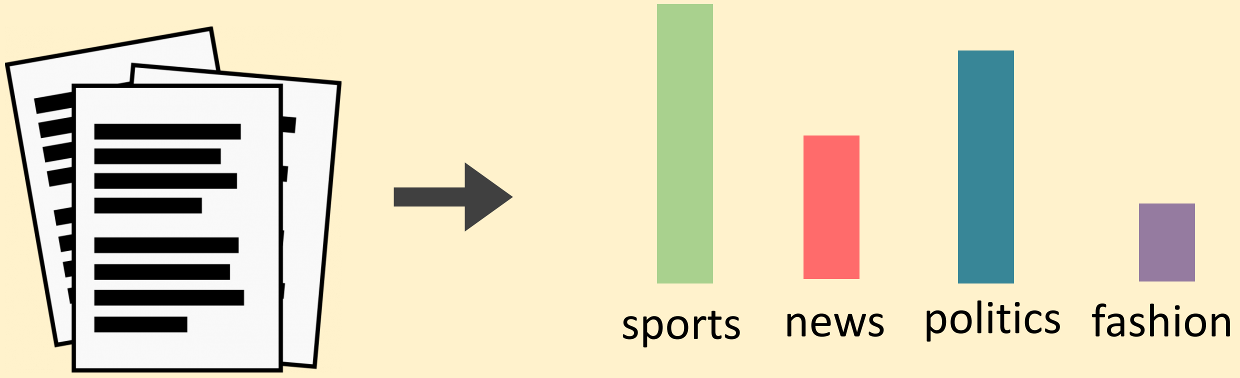
Summarization

Coreference Re

## Semantics

Sentiment Analysis

Topic Modelling




Named E y Recognition (NER)

(NLU)

G)

Language Modelling

# Common NLP Tasks (aka problems)

## Syntax

Morpholog

Word Segm

Part-of-Spe

Parsing

Constit

Dependency

## Semantics

ER)

"Alexa, play Drivers License by Olivia Rodrigo"

↓

"Alexa, play Drivers License by Olivia Rodrigo"

INTENT SONG ARTIST

Natural Language Understanding (NLU)

Natural Language Generation (NLG)

Machine Translation

Entailment

Question Answering

Language Modelling

## Discourse

Summarization

Coreference Resolution

# Common NLP Tasks (aka problems)

## Syntax

Morphology

Word Segmentation

Part-of-Speech Tagging

Parsing

- Constituency
- Dependency

## Discourse

Summarization

Coreference Resolution

## Semantics

Sentiment Analysis

Topic Modelling

Named Entity Recognition (NER)

ng (NLU)

Natural Lang e Generation (NLG)

Machine Translation

Entailment

Question Answering

Language Modelling

El perro marrón → The brown dog

SPANISH ENGLISH

# Very brief history of NLP

- 1960s: pattern-matching and rules (highly limiting)
- 1970s – 1980s: linguistically rich, logic-driven systems; labor-intensive successes on a few, very specific tasks
- 1990s – 2000s: statistical modelling takeover! ML becomes a central component; some systems are deployed for practical use (e.g., speech to text)
- 2010s – 2020s: Deep Learning (neural nets) yields astronomical progress on nearly every NLP task; systems become fairly useful for consumers
- 2020s – 2030s: ???? You can help drive the change



Slide adapted from or inspired by Sam Bowman

# Very brief history of NLP

First huge revolution: early 1990s (statistical approaches)

> "But it must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term"

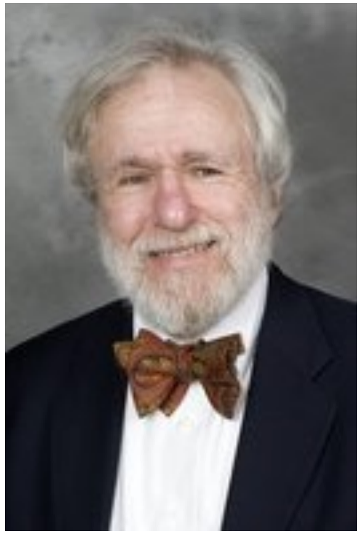
**-- Noam Chomsky (1969)**

> "Anytime a linguist leaves the group, the recognition rate goes up"

**-- Frederick Jelinek (1988)**

# Very brief history of NLP

First huge revolution: early 1990s (statistical approaches)



"I refer to all of my work before ~1990 as the B.S. era. That is, 'before statistics'"

-- paraphrasing my PhD adviser, Eugene Charniak at his ACL Lifetime Achievement Award (2011)

# GPT-2 (generates text and can fine-tune on your own data)

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”
The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.
The Nuclear Regulatory Commission did not immediately release any information.

# NLP nowadays

## GPT-2 (generates text and can fine-tune on your own data)

| Question | Generated Answer | Correct | Probability |
|---|---|---|---|
| Who wrote the book the origin of species? | Charles Darwin | ✓ | 83.4% |
| Who is the founder of the ubuntu project? | Mark Shuttleworth | ✓ | 82.0% |
| Who is the quarterback for the green bay packers? | Aaron Rodgers | ✓ | 81.1% |
| Panda is a national animal of which country? | China | ✓ | 76.8% |
| Who came up with the theory of relativity? | Albert Einstein | ✓ | 76.4% |
| When was the first star wars film released? | 1977 | ✓ | 71.4% |
| What is the most common blood type in sweden? | A | ✗ | 70.6% |
| Who is regarded as the founder of psychoanalysis? | Sigmund Freud | ✓ | 69.3% |
| Who took the first steps on the moon in 1969? | Neil Armstrong | ✓ | 66.8% |
| Who is the largest supermarket chain in the uk? | Tesco | ✓ | 65.3% |
| What is the meaning of shalom in english? | peace | ✓ | 64.0% |
| Who was the author of the art of war? | Sun Tzu | ✓ | 59.6% |
| Largest state in the us by land mass? | California | ✗ | 59.2% |
| Green algae is an example of which type of reproduction? | parthenogenesis | ✗ | 56.5% |
| Vikram samvat calender is official in which country? | India | ✓ | 55.6% |
| Who is mostly responsible for writing the declaration of independence? | Thomas Jefferson | ✓ | 53.3% |

# NLP nowadays

Table 3: Video captioning performance on YouCook II. We follow the setup from [39] and report captioning performance on the validation set, given ground truth video segments. Higher numbers are better.

**GT**: add some chopped basil leaves into it

**VideoBERT**: chop the basil and add to the bowl

**S3D**: cut the tomatoes into thin slices

**GT**: cut the top off of a french loaf

**VideoBERT**: cut the bread into thin slices

**S3D**: place the bread on the pan

**GT**: cut yu choy into diagonally medium pieces

**VideoBERT**: chop the cabbage

**S3D**: cut the roll into thin slices

**GT**: remove the calamari and set it on paper towel

**VideoBERT**: fry the squid in the pan

**S3D**: add the noodles to the pot

VideoBERT: A Joint Model for Video and Language Representation Learning. Sun, et al. ICCV 2019.

# Deep Learning (breakthrough moment)

## Data

14 million images.
20,000 distinct categories (e.g., shoes).

## Task

Given an image, correctly predict which category it belongs to

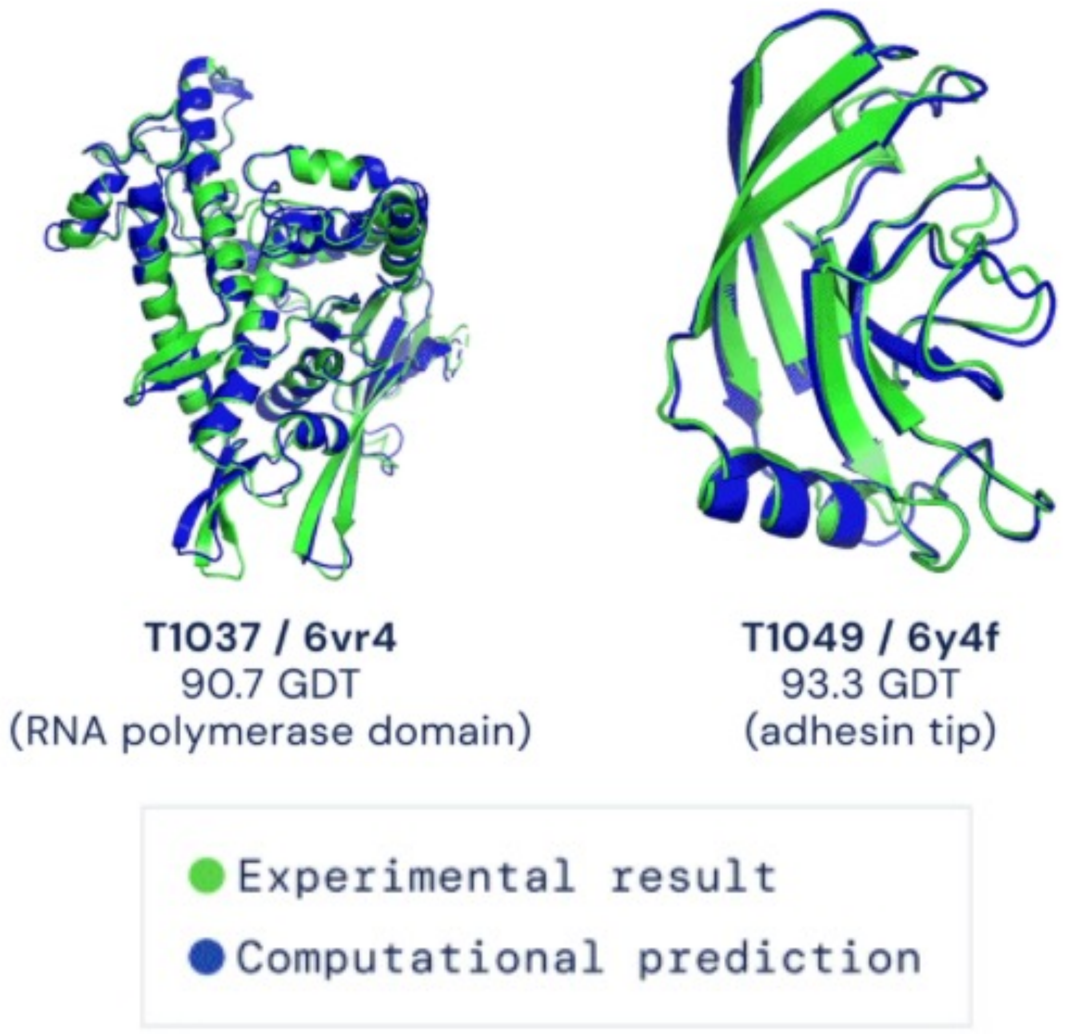
## AlexNet Model

The network achieved a top-5 error of 15.3%, more than 10.8 percentage points lower than that of the runner up.

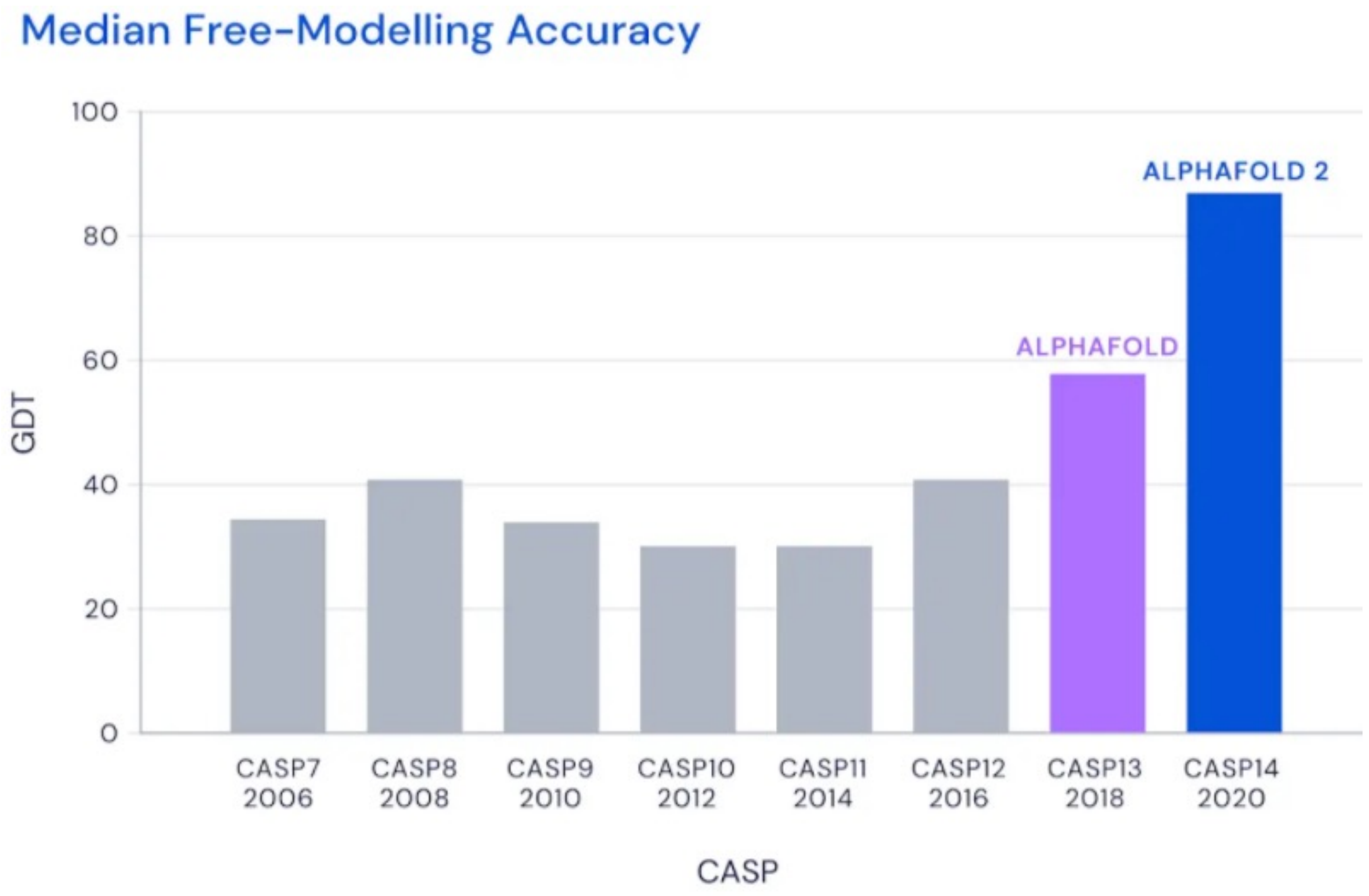
ImageNet Classification with Deep Convolutional Neural Networks. Krizhevsky, et al. (2012)

# Deep Learning (recent breakthrough)

**AlphaFold: a solution to a 50-year-old grand challenge in biology**



# Deep Learning (recent breakthrough)



# Deep Learning (recent breakthrough)

In the results from the 14th CASP assessment, released today, our latest AlphaFold system achieves a median score of 92.4 GDT overall across all targets. This means that our predictions have an average error (RMSD) of approximately 1.6 Angstroms, which is comparable to the width of an atom (or 0.1 of a nanometer). Even for the very hardest protein targets, those in the most challenging free-modelling category, AlphaFold achieves a median score of 87.0 GDT (data available here).

# Deep Learning

- **Deep Learning** is just neural networks with more than 1 hidden layer (non-linear activation functions).
- For the 1st time ever, one paradigm of modelling (deep learning) yields the best results across nearly every domain of problems
- Our understanding of why and how the results are so compelling is very surface-level.
- Much work lies ahead (e.g., bias/fairness, explainability, robustness)

# The Two Cornerstones of NLP

How do we get *any* system to process, "understand", leverage language?

- **Representation**: how do we transform symbolic meaning (e.g., words, signs, braille, speech audio) into something the computer can use
- **Modelling**: given these represented symbols, how do we use them to model the task at hand?

# Outline

- NLP: what and why?
- Representing Language
- Bag-of-Words
- TF-IDF

# Outline

- NLP: what and why?
- Representing Language
- Bag-of-Words
- TF-IDF

# Representing Numbers

- **Remember,** a computer only has bits: 0's and 1's
- Computer architecture allows us to perform basic arithmetic operations (+ - * / )
- Computational models "need" numeric data. The relationship of numbers is natural (e.g., < > ==). Think of logistic regression.
- Numeric data? No problem

# Representing Images

- **Images** (like language) capture tons of real-world concepts
- The data itself is well-represented and captured by pixel values (0-255)
- Little cumbersome to capture spatial information (LBP and CNNs)
- Image data? Not too difficult to *represent*.
  - I am making no claims about the modelling

# Representing Images

| 170 | 33 | 71 |
|---|---|---|
| r | g | b |

Meaningful relation between the byte values and color.

# Representing Images

255 33 71

r g b

Meaningful relation between the byte values and color.

# Representing Images

255 33 71

r g b

Meaningful relation between the byte values and color.

Thus, colors, and images at large, are well-represented.

# Representing Language

- **Words** are represented by Strings

a t e

| | | |
|---|---|---|
| 61 | 74 | 65 |

Each byte corresponds to language's smallest meaningful unit! Yay!

# Representing Language

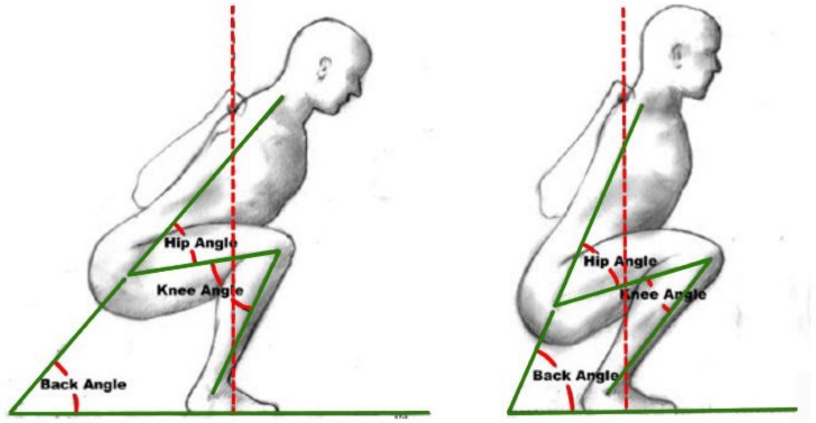
- **Words** are represented by Strings

| a | t | f |
|---|---|---|
| 61 | 74 | 66 |

No meaningful relation between the byte values and language!

# Representing Language

- **Words** are represented by Strings




a t g

| | | |
|---|---|---|
| 61 | 74 | 67 |

A.T.G. is, however, more intense. Never mind. Ignore this slide.

# Representing Language

- **Words** are represented by Strings

h a t e

| | | | |
|---|---|---|---|
| 68 | 61 | 74 | 65 |

hate and ate. No relation but similar byte values.

# Representing Language

- **Words** are represented by Strings

h a t

| 68 | 61 | 74 |
|---|---|---|

hate and hat. No relation but similar byte values.

# Representing Language

- **Words** are represented by Strings

H a t

| | | |
|---|---|---|
| 48 | 61 | 74 |

Hat and hat. Identical concept but different byte values.

# Symbolic Representations?

The earliest approaches used symbolic representations. Active research still.

Conceptual Dependency Theory (1972) asserted two assumptions:

1. If two sentences have the same meaning, they should be represented the same, regardless of the particular words used.
2. Information implicitly stated or inferred from the sentence should be represented explicitly.

# Conceptual Dependencies

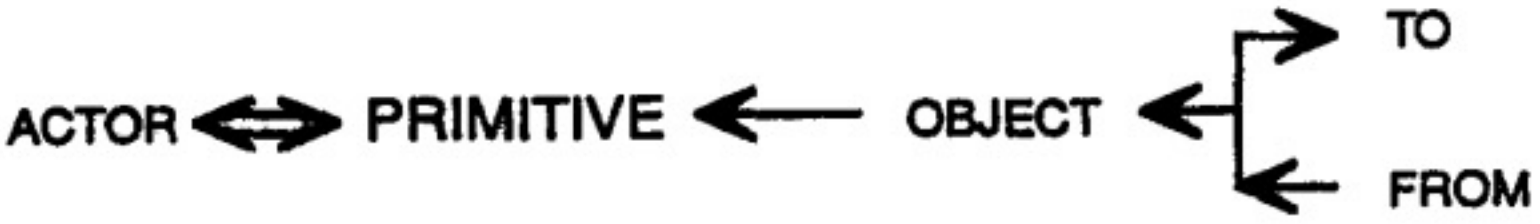



Figure 2. Basic form of a conceptual dependency graph.

Everything centered around primitives, states, and dependencies.

Conceptual Dependency and its Descendants. Lytinen, S.L. *Computers, Mathematics, and Applications* (1992)

# Conceptual Dependencies

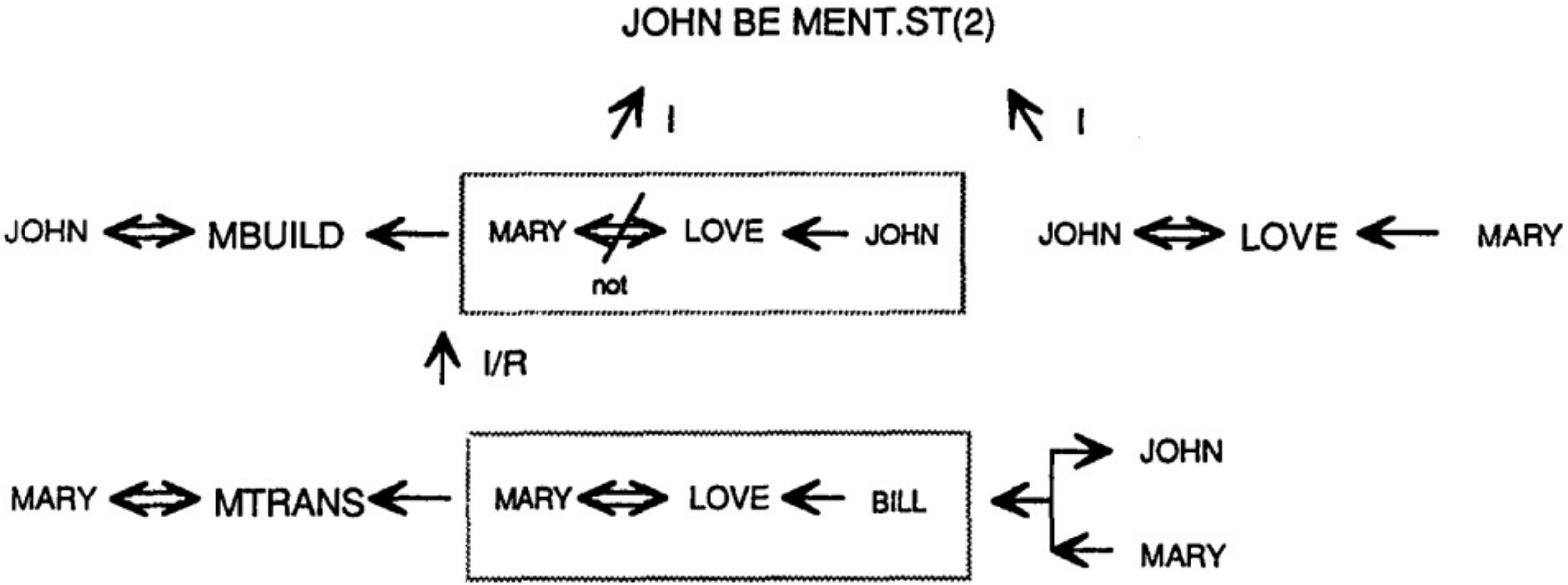



**Figure 6. Representation of "John cried because Mary said she loved Bill."**

Conceptual Dependency and its Descendants. Lytinen, S.L. *Computers, Mathematics, and Applications* (1992)

# External Resources

There are rich, external resources that define real-world relationships and concepts

(e.g., WordNet, BabelNet, PropBank, VerbNet, FrameNet, ConceptNet)

# WordNet

A large lexical database with English nouns, verbs, adjectives, and adverbs grouped into over 100,000 sets of cognitive synonyms (*synsets*) – each expressing a different concept.

**Most frequent relation**: super-subordinate relation ("is-a" relations).

{furniture, piece_of_furniture}

**Fine-grained relations**:
{bed, bunkbed}

**Part-whole relations:**
{chair, backrest}

**Synonyms:**
{adept, expert, good, practiced, proficient}

# ConceptNet

A multilingual semantic knowledge graph, designed to help computers understand the meaning of words that people use.

- Started in 1999. Pretty large now.
- Finally becoming useful (e.g, *commonsense reasoning*)
- Has synonyms, ways-of, related terms, derived terms

# ConceptNet

Entry for "teach"

en **teach**

An English term in ConceptNet 5.8

Documentation FAQ

**Sources:** Open Mind Common Sense contributors, Verbosity players, German Wiktionary, English Wiktionary, French Wiktionary, and Open Multilingual WordNet
View this term in the API

## Synonyms

- ar عَلَّمَ (v, change) →
- ar عَلَّمَ (v, communication) →
- ca ensenyar (v, change) →
- ca ensenyar (v, communication) →
- ca informar (v, communication) →
- ca instruir (v, change) →
- ca instruir (v, communication) →
- da lære (v, communication) →
- en instruct (v, communication) →
- en learn (v, communication) →

## Ways of teach

- en catechize (v, communication) →
- en coach (v, communication) →
- en condition (v, social) →
- en drill (v, cognition) →
- en enlighten (v, communication) →
- en ground (v, communication) →
- en indoctrinate (v, cognition) →
- en induct (v, communication) →
- en lecture (v, communication) →
- en mentor (v, communication) →

## Related terms

- sh naučiti (v) →
- sh obučavati (v) →
- sh obučiti (v) →
- sh podučiti (v) →
- sh predavati (v) →
- sh uputiti (v) →
- sh upućivati (v) →
- sh učiti (v) →
- ab арҵара (v) →
- ab аҵара (v) →

## Derived terms

- en beteach →
- en coteach →
- en foreteach →
- en forteach →
- en microteach →
- en overteach →
- en pre teach →
- en reteach →
- en teachability →
- en teacher →

# Limitations

- Great resources but ultimately finite
- Can’t perfectly capture nuance (especially context-sensitive)
  (e.g., ‘proficient’ is grouped with ‘good’, which isn’t always true)
- Will always have many out-of-vocabulary terms (OOV)
  (e.g., COVID19, Brexit, bet, wicked, stankface, “no cap”)
- Subjective
- Laborious to annotate
- Words with the same spelling are doomed to be imprecise

# Outline

- NLP: what and why?
- Representing Language
- Bag-of-Words
- TF-IDF

# Outline

- NLP: what and why?
- Representing Language
- Bag-of-Words
- TF-IDF

# Text Classification

Let's zoom out for a bit and first address

coarse-grain processing at the document level

- **Input:** document $d$
- **Output:** predicted class $c \in \{c_1, c_2, \dots, c_n\}$

# Spam Detection

Is this spam?

**Mrs ngui mrs_karen1@pm.me** via g.harvard.e... Tue, Aug 31, 8:57 AM (6 days ago)

to christanner

**Why is this message in spam?** You reported this message as spam from your inbox.

Report not spam

WARNING: Harvard cannot validate this message was sent from an authorized system. Please be careful when opening attachments, clicking links, or following instructions. For more information, visit the HUIT IT Portal and search for SPF.

Hello

I have a business proposal that I would like to discuss with you if you don't mind kindly respond to my email, so I can explain better to you

Mrs Ngui, Karen
General Manager DBS

# Spam Detection

Is this spam?

Dear Professor Christopher W. Tanner:

Thanks a lot for spending your time to read my coming letter!

My name is [redacted] and I already graduated from ShenZhen University at December, 2016. My research area is image saliency and object detection. I proposed an elegant, robust, and yet extremely simple algorithm to solve the image saliency detection problem, which is called "Iterative Saliency via dynamic image background". Huchuan Lu, who is a well-known object detection researcher, highly remarked this paper during the peer review of this graduate paper. Some research papers of Huchuan Lu which are based on the research idea of this paper had been adopted by multiple big companies such as Google. What's more, even the Turing Award winner Geoffrey E. Hinton used this idea to improve the performance of his capsule neural network. It is also now widely used in multiple areas such as object detection, visual object tracking, pedestrian detection, nature language processing, deep learning, machine learning, etc. By the way, during my exprement, I found an interesting phenomenon, that is an image as a sequence of numbers can be classified into multiple subgroups via massive automatic iterative training.

But the thing that makes me very angry was that the paper made by me could not be published when I was still studying at campus because of someone called Hai Xie. Hai Xie always assaulted on my academic research without any reasonable advice. After a longtime negotiation with the graduate advisors, I finally graduated from ShenZhen University because I found the stolen and copy facts of Hai Xie's graduate paper but not the contribution that I invented this algorithm. At that time, they were very afraid of the copy facts of Hai Xie's graduate paper, if I revealed these facts to the university's president, they must all be punished by the ethic committee of academic research. Finally, we all reached a deal that I can graduate from school but I should keep the Hai Xie's copy facts as a secret. So, I always feel this unfair treatment is such a Shit that happened on me. They do not respect human rights, even the personal rights of pursuiting science. The thing that makes me astonished is the behavior of ShenZhen University. They tried all kinds of ways to punish me but to protect this Hai Xie. What's more, the ShenZhen University even honored this Hai Xie no matter how terrible the dishonest facts of Hai Xie's graduate paper and published papers are. They always publish lots of papers by integrating some other person's experiment results to get a better experimental results curve even without writing one line of code sometimes.

After a long time consideration, I obtained enough courage to reveal the evidences of Hai Xie's stolen facts of his published paper on IJPRAI as supplementary resources. In fact, the orgin author of Hai Xie's published paper is my former graduate student whose name is Wenzhou Fang. This published paper is actually Wenzhou Fang's graduate paper. Wenzhou Fang first tried to publish this paper on "Signal Processing Letters", but failed. What's more, Wenzhou Fang's graduate paper had been published openly at year 2015 after he graduated. So, according to my knowledge about human intellectual property, Hai Xie's stolen facts ruled the authorities of scientific research and should be punished by IEEE committee.

He delete the original author Wenzhou Fang and he published the same paper without replacing the figures and introduction. He didn't respect the copyrights of human intellectual property. In fact, In order to get access to Wenzhou Fang's graduate paper, I paid nearly three dollars to obtain the useability of his intellectual property. In contrast, this Hai Xie he just replaced the origin author Wenzhou Fang with his own name regardless of the copyright of Wenzhou Fang.

Hai Xie is now a doctor student of ShenZhen University who majored in computer science. Hai Xie stolen the content of Wenzhou Fang's graduate paper and the core idea of my paper and then published the paper "Hierarchical Saliency Detection via Probabilistic Object Boundaries" at the conference of <<International Journal of Pattern Recognition & Artificial Intelligence>>, 2017, 31(6):8. He is the secondary author of this paper. The original author should be Wenzhou Fang.

Something even more annoying is that Hai Xie's graduate paper was also stolen and copied from several published IEEE papers and he also graduated from ShenZhen University in 2016. Hai Xie is now a doctor student in ShenZhen University. I proposed all the materials of Hai Xie's stolen and copying facts to the ethic committee of graduate thesis of ShenZhen University, but nothing helped. So, any way, I really feel so helpless and I'm so sad about the truth of justice could not be fulfilled even in the field of scientific research in China. I guess that such kind of a phenomenon if happened in your country, students

# Authorship Identification

Did Plato really write this?

**The Seventh Letter**

By Plato

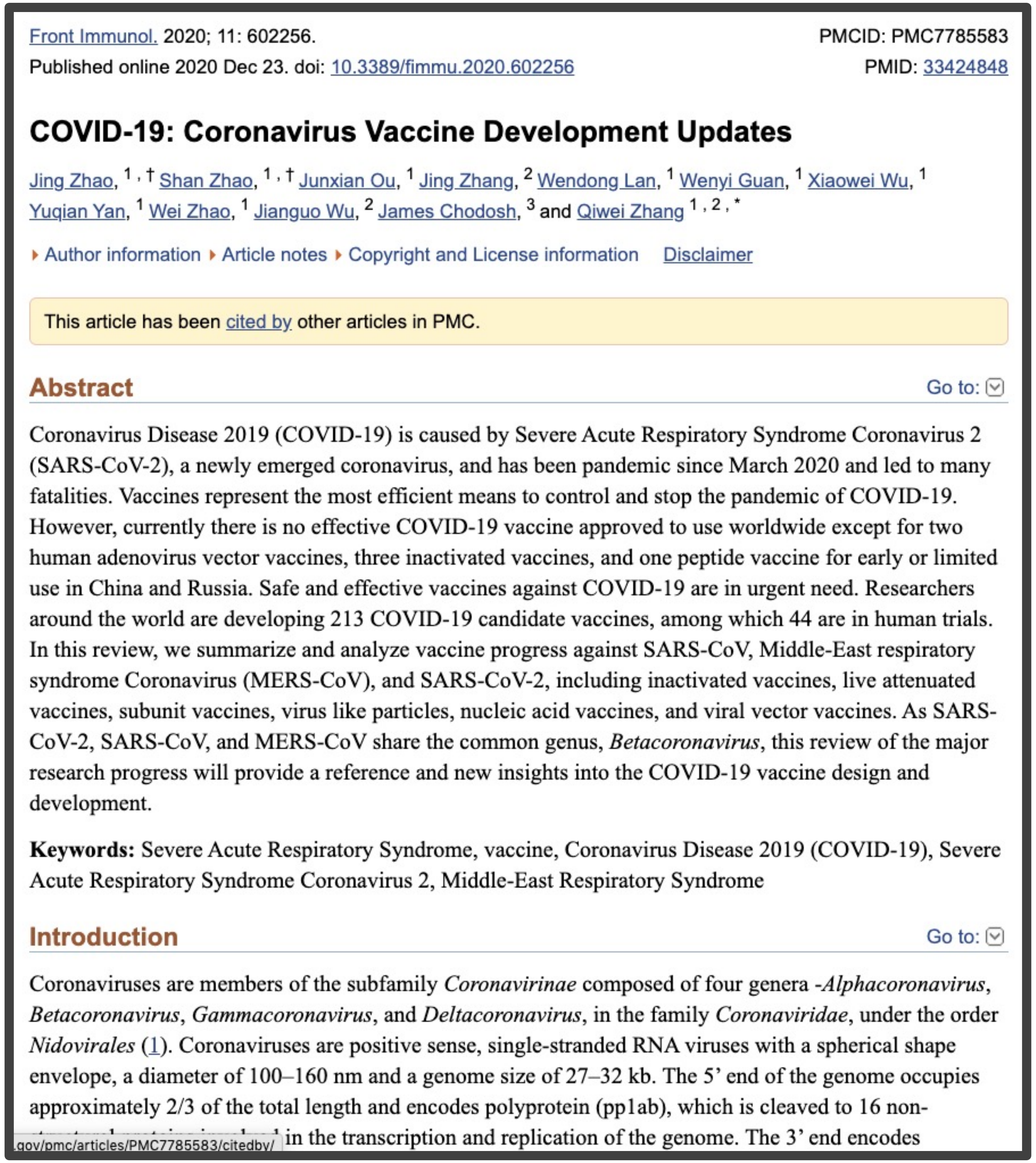
Written 360 B.C.E

Translated by J. Harward

Plato TO THE RELATIVES AND FRIENDS OF DION. WELFARE.

You write to me that I must consider your views the same as those of Dion, and you urge me to aid your cause so far as I can in word and deed. My answer is that, if you have the same opinion and desire as he had, I consent to aid your cause; but if not, I shall think more than once about it. Now what his purpose and desire was, I can inform you from no mere conjecture but from positive knowledge. For when I made my first visit to Sicily, being then about forty years old, Dion was of the same age as Hipparinos is now, and the opinion which he then formed was that which he always retained, I mean the belief that the Syracusans ought to be free and governed by the best laws. So it is no matter for surprise if some God should make Hipparinos adopt the same opinion as Dion about forms of government. But it is well worth while that you should all, old as well as young, hear the way in which this opinion was formed, and I will attempt to give you an account of it from the beginning. For the present is a suitable opportunity.

In my youth I went through the same experience as many other men. I fancied that if, early in life, I became my own master, I should at once embark on a



Front Immunol. 2020; 11: 602256. PMCID: PMC7785583
Published online 2020 Dec 23. doi: 10.3389/fimmu.2020.602256 PMID: 33424848

## COVID-19: Coronavirus Vaccine Development Updates

Jing Zhao,[1,†] Shan Zhao,[1,†] Junxian Ou,[1] Jing Zhang,[2] Wendong Lan,[1] Wenyi Guan,[1] Xiaowei Wu,[1] Yuqian Yan,[1] Wei Zhao,[1] Jianguo Wu,[2] James Chodosh,[3] and Qiwei Zhang[1,2,*]

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer

This article has been cited by other articles in PMC.

### Abstract

Go to:

Coronavirus Disease 2019 (COVID-19) is caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), a newly emerged coronavirus, and has been pandemic since March 2020 and led to many fatalities. Vaccines represent the most efficient means to control and stop the pandemic of COVID-19. However, currently there is no effective COVID-19 vaccine approved to use worldwide except for two human adenovirus vector vaccines, three inactivated vaccines, and one peptide vaccine for early or limited use in China and Russia. Safe and effective vaccines against COVID-19 are in urgent need. Researchers around the world are developing 213 COVID-19 candidate vaccines, among which 44 are in human trials. In this review, we summarize and analyze vaccine progress against SARS-CoV, Middle-East respiratory syndrome Coronavirus (MERS-CoV), and SARS-CoV-2, including inactivated vaccines, live attenuated vaccines, subunit vaccines, virus like particles, nucleic acid vaccines, and viral vector vaccines. As SARS-CoV-2, SARS-CoV, and MERS-CoV share the common genus, *Betacoronavirus*, this review of the major research progress will provide a reference and new insights into the COVID-19 vaccine design and development.

**Keywords:** Severe Acute Respiratory Syndrome, vaccine, Coronavirus Disease 2019 (COVID-19), Severe Acute Respiratory Syndrome Coronavirus 2, Middle-East Respiratory Syndrome

### Introduction

Go to:

Coronaviruses are members of the subfamily *Coronavirinae* composed of four genera -*Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus*, in the family *Coronaviridae*, under the order *Nidovirales* (1). Coronaviruses are positive sense, single-stranded RNA viruses with a spherical shape envelope, a diameter of 100–160 nm and a genome size of 27–32 kb. The 5’ end of the genome occupies approximately 2/3 of the total length and encodes polyprotein (pp1ab), which is cleaved to 16 non-
gov/pmc/articles/PMC7785583/citedby/ in the transcription and replication of the genome. The 3’ end encodes

# What’s the subject of this article?

## MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Slide inspired by or adapted from Dan Jurafsky. https://web.stanford.edu/~jurafsky/slp3/slides/7_NB.pdf

# Yelp Sentiments

- Really slow wait. Took forever to get food.
- Freshest ingredients ever. New favorite restaurant. Will be back!
- Found a hair in the food. Horrible.
- Waited 6 months to get a reservation at this place. Totally worth it.

# Text Classification (supervised ML)

Training Data: $(d_1, c_1, d_2, c_2, \ldots, d_M, c_M)$

**Simple idea:** let's represent each document as a feature vector, which can serve as the input to any of your favorite supervised ML models

# Bag-of-words (BoW)

Let's say our dataset's entire *vocabulary* is just 10 words.

Each unique word can have its own dimension (feature index).

$$[\ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ ]$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| dog | the | quick | went | brown | a | jumped | fast | over | store |

NOTE: This is the Boolean version, which isn't the most popular BoW representation

# Bag-of-words (BoW)

Each document's vector has a 1 if the word is present. Otherwise, 0.

e.g., "the dog jumped" is represented as

$$\begin{array}{cccccccccc} \left[\,1 \right. & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \left. 0\,\right] \\ \text{dog} & \text{the} & \text{quick} & \text{went} & \text{brown} & \text{a} & \text{jumped} & \text{fast} & \text{over} & \text{store} \end{array}$$

NOTE: This is the Boolean version, which isn't the most popular BoW representation

# Bag-of-words (BoW)

Each document's vector has a 1 if the word is present. Otherwise, 0.

e.g., "the dog went fast" is represented as

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| dog | the | quick | went | brown | a | jumped | fast | over | store |

NOTE: This is the Boolean version, which isn't the most popular BoW representation

# Bag-of-words (BoW)

NOTE: The most common way of referring to this is as a "bag-of-words model". Technically, the "bag-of-words" is referring to the representation, not the model.

"bag-of-words model" actually means "Model that uses a bag-of-words representation"

# Bag-of-words (BoW)

Are there any weaknesses with this type of representation?

# Bag-of-words (BoW)

Weaknesses:

- Flattened view of the document
- Context-insensitive ("the horse ate" = "ate the horse")
- Curse of Dimensionality (vocab could be over 100k)
- Orthogonality: no concept of semantic similarity at the word-level
  - e.g., $d$(dog, cat) $=d$(dog, chair)

# Bag-of-words (BoW)

Let's address the "flattened view of the document"

# Bag-of-words (BoW)

Imagine a document is a sports broadcast transcript, which concerns a few teams but mostly discusses the local home team, the Cubs

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| [ 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 ] |
| baseball | chicago | cubs | the | wringley | padres | shohei | mvp | homerun | crowd |

# Bag-of-words (BoW)

We have no indication of *how much* the document is about the Cubs.

$$\begin{array}{cccccccccc} & & & & & & & & & \\ \end{array}$$

| baseball | chicago | cubs | the | wringley | padres | shohei | mvp | homerun | crowd |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |

# Bag-of-words (BoW)

Now we can see that it's much more about the Chicago Cubs than the Padres.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 9 | 17 | 8 | 0 | 2 | 0 | 0 | 0 | 2 |
| baseball | chicago | cubs | the | wringley | padres | shohei | mvp | homerun | crowd |

# Bag-of-words (BoW)

Now we can see that it's much more about the Chicago Cubs than the

This count-based approach is the most common BoW representation, and it's what we expect in HW1

| baseball | chicago | cubs | the | wringley | padres | shohei | mvp | homerun | crowd |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 9 | 17 | 8 | 0 | 2 | 0 | 0 | 0 | 2 |

# Outline

- NLP: what and why?
- Representing Language
- Bag-of-Words
- TF-IDF

# Outline

- NLP: what and why?
- Representing Language
- Bag-of-Words
- TF-IDF

# TF-IDF

Notice that longer documents will naturally have higher counts than shorter documents.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 9 | 17 | 8 | 0 | 2 | 0 | 0 | 0 | 2 |
| baseball | chicago | cubs | the | wringley | padres | shohei | mvp | homerun | crowd |

# TF-IDF

Also notice that "the" has a fairly high count, too.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 9 | 17 | 8 | 0 | 2 | 0 | 0 | 0 | 2 |
| baseball | chicago | cubs | the | wringley | padres | shohei | mvp | homerun | crowd |

# TF-IDF

Simple ideas. Let's:

- disproportionately weight the common words that appear in many documents
- Use that info and combine it with the word frequency info

# TF-IDF

TF (term frequency) = $f_{w_i}$ = # times word $w_i$ appeared in the document

IDF (inverse document frequency) = $log\left(\frac{\text{\# docs in corpus}}{\text{\# docs containing } w_i}\right)$

$$\text{TFIDF} = f_{w_i} * log\left(\frac{\text{\# docs in corpus}}{\text{\# docs containing } w_i}\right)$$

# TF-IDF

Weaknesses:

- ~~Flattened view of the document~~
- Context-insensitive (“the horse ate” = “ate the horse”)
- Curse of Dimensionality (vocab could be over 100k)
- Orthogonality: no concept of semantic similarity at the word-level
  - e.g., $d$(dog, cat) =$d$(dog, chair)

# TF-IDF

Weaknesses:

Next lecture, we'll address this

- ~~Flattened vi~~
- Context-insensitive ("the horse ate" = "ate the horse")
- Curse of Dimensionality (vocab could be over 100k)
- Orthogonality: no concept of semantic similarity at the word-level
  - e.g., $d$(dog, cat) $=d$(dog, chair)

# TF-IDF

Weaknesses:

- ~~Flattened view~~
- Context-insens
- Curse of Dim nsionality (vocab could be over 100k)
- Orthogonality: no concept of semantic similarity at the word-level
  - e.g., $d$(dog, cat) $=d$(dog, chair)

In the following lecture,
we'll address these points

# EXTRA

The **Naïve Bayes Classifier** often used while assuming <u>word independence</u>

When performing text classification, we're interested in predicting class $c$ for

a given document $d$

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

# Naïve Bayes Classifier

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(c|d)$$

$$= \underset{c \in C}{\operatorname{argmax}}\, \frac{P(d|c)P(c)}{P(d)}$$

$$= \underset{c \in C}{\operatorname{argmax}}\, P(d|c)P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}}\, P(w_1, w_2, \dots, w_n|c)P(c)$$

We assume word order doesn't matter.

# Naïve Bayes Classifier

We assume word order doesn't matter.

$$P(w_1, w_2, \ldots, w_n|c) = P(w_1|c) * P(w_2|c)\, P(w_3|c) \ldots P(w_n|c)$$

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}}\, P(c_j) \prod_{w_i \in W} P(w_i|c_j)$$

where, $P(w_i|c_j) = \dfrac{\text{count}(w_i, c_j)}{\sum_{w \in W} \text{count}(w, c_j)}$