Statistical models and their algorithms

Daniel Hsu

Department of Computer Science
& Data Science Institute

Columbia University

Science Research Fellows Seminar
October 10, 2025

Some slides/figures borrowed from Katelynn Devinney, Roxana Geambasu, José Manuel Zorrilla Matilla

We use statistical models to make sense of the world …

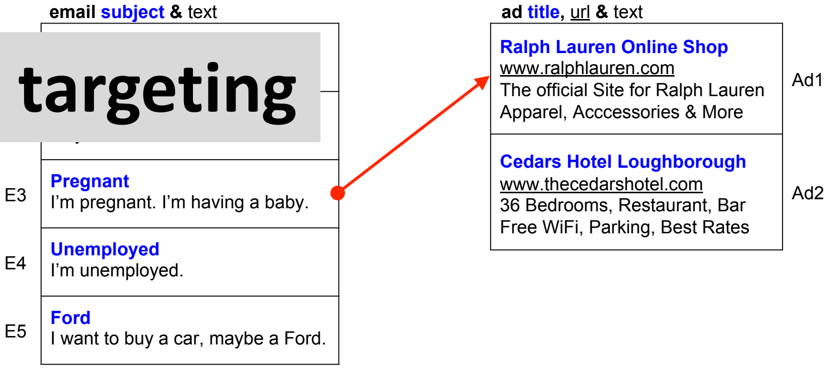
flu prevalence

[Kandula, H., Shaman, 2017]

social media

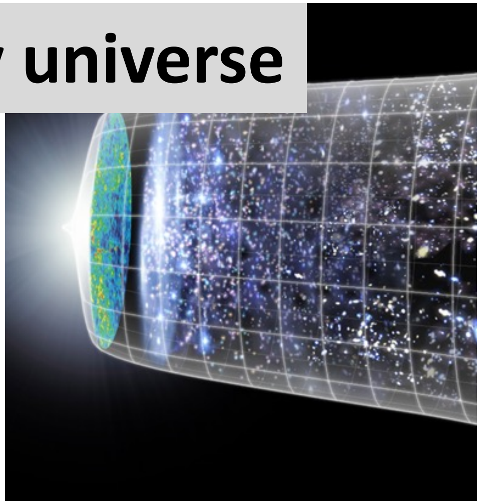
[Effland, Lawson, Balter, Devinney, Reddy, Waechter, Gravano, H., 2018]

ad targeting



[Lecuyer, Spahn, Spiliopoulos, Chaintreau, Geambasu, H., 2015]

early universe



[Gupta, Zorrilla Matilla, H., Haiman, 2018]

… and to inform our decisions & actions

Foodborne Tracker

- Dashboard
- Investigations
- Yelp
- Twitter
- NYC 311
- Interviews
- Surveys
- Users
- Logout

Dashboard

Last update: Tue, 05/29/2018 11:05

Update Dashboard

57068 restaurants in the Yelp catalog
4114699 total reviews
6260 positive reviews
225559 users

Explore Yelp

9106161 total messages
916047 positive messages
5553715 users

Explore Twitter

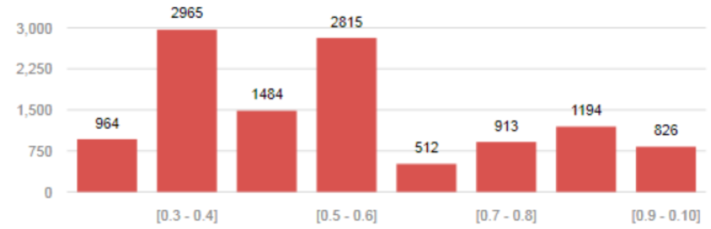
54 detected outbreaks
51 open investigations
3 closed investigations
0 potential investigations

Explore Investigations

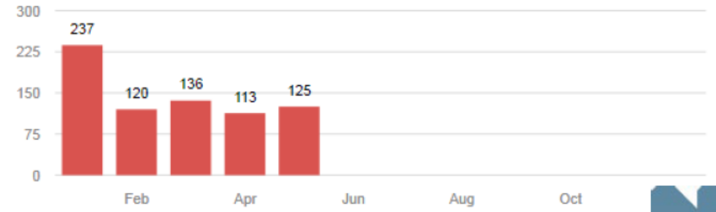
Yelp ⇒ Last 10 days

Day	>=0.0	>=0.1	>=0.2	>=0.3	>=0.4	>=0.5	>=0.6	>=0.7	>=0.8	>=0.9
2018-05-28	28	1	1	0	0	0	0	0	0	0
2018-05-24	15	0	0	0	0	0	0	0	0	0
2018-05-23	24	0	0	0	0	0	0	0	0	0
2018-05-22	13	1	0	0	0	0	0	0	0	0
2018-05-21	15	0	0	0	0	0	0	0	0	0
2018-05-20	38	1	1	1	1	1	0	0	0	0
Total	133	3	2	1	1	1	0	0	0	0

Yelp ⇒ Classification Breakdown



Yelp ⇒ Temporal Breakdown



Health

Where do these models come from?

1. Collect data (observations of some phenomenon)
2. Find a statistical model that fits data
3. Test predictions of the model (with more observations)
4. Repeat

My research concerns algorithmic & statistical aspects of statistical models

<u>Primary mode of research:</u>
Rigorously analyze (and prove theorems about) algorithms & statistical models

This talk

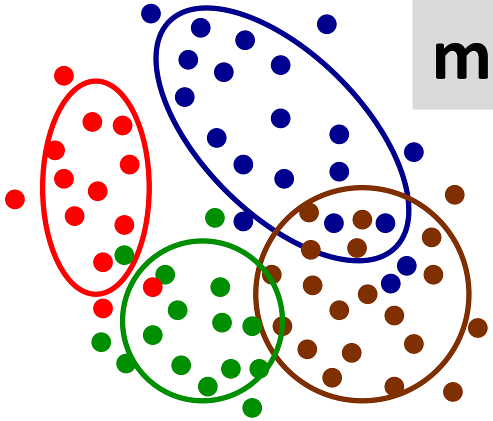
Part 1: Algorithms for fitting statistical models

Part 2: Algorithms in large language models

1. Algorithms for fitting statistical models

Many statistical models, many algorithms

mixture model



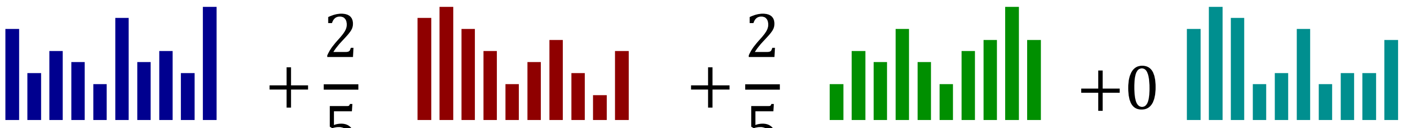
[Xu, H., Maleki, 2016]

logistic regression

[H. & Mazumdar, 2024]

topic model

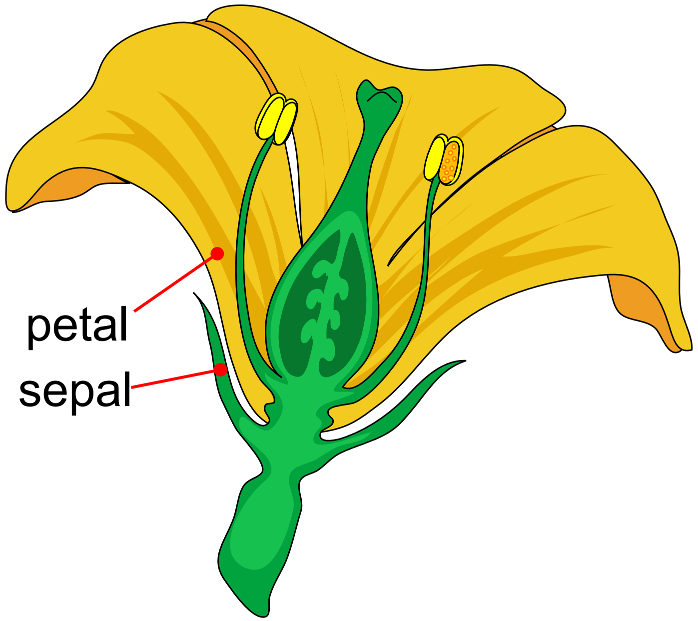
$\sim$iid $\frac{1}{5}$ sports $+\frac{2}{5}$ science $+\frac{2}{5}$ politics $+0$ business

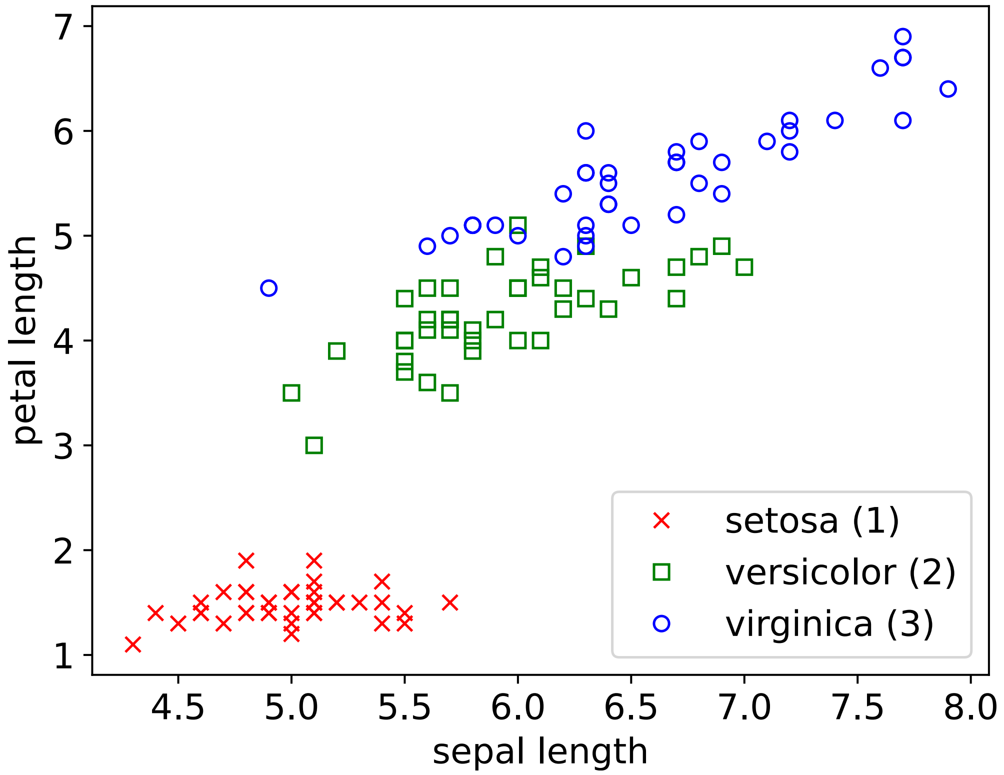


[Anandkumar, Foster, H., Kakade, Liu, 2015]

Example: iris dataset

Anderson's iris dataset (studied by Fisher, 1936)

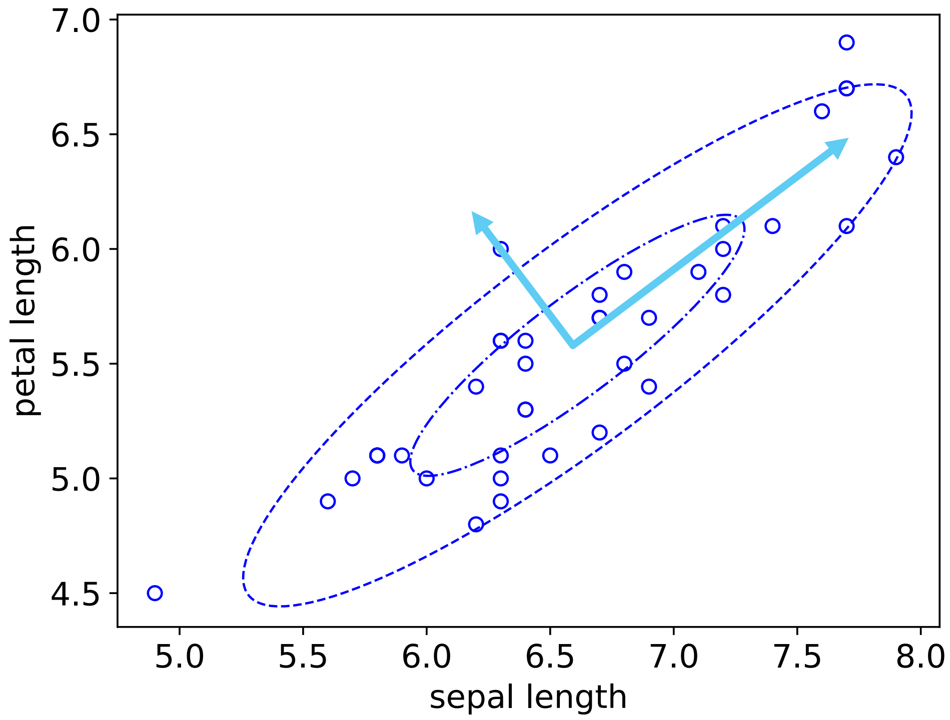




Example: Gaussian model

Probability distribution that considers:

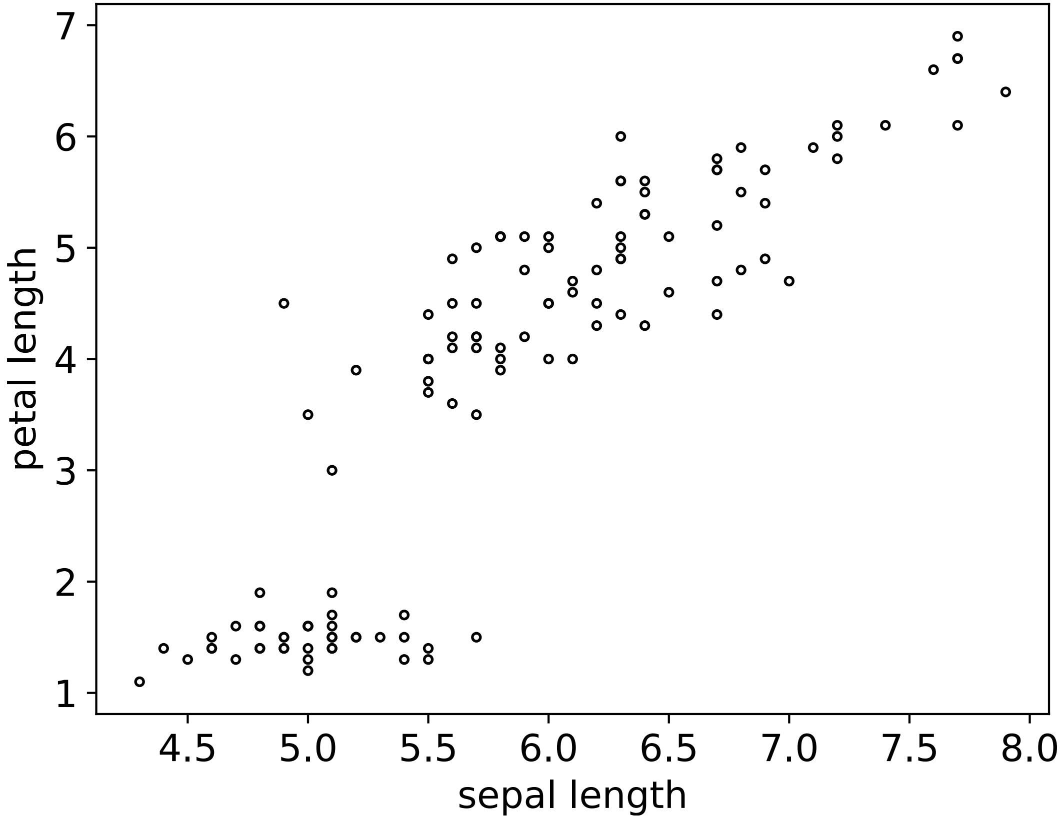
- Average value of attributes (means)
- Dispersion of attributes (variances)
- Association between attributes (covariances)

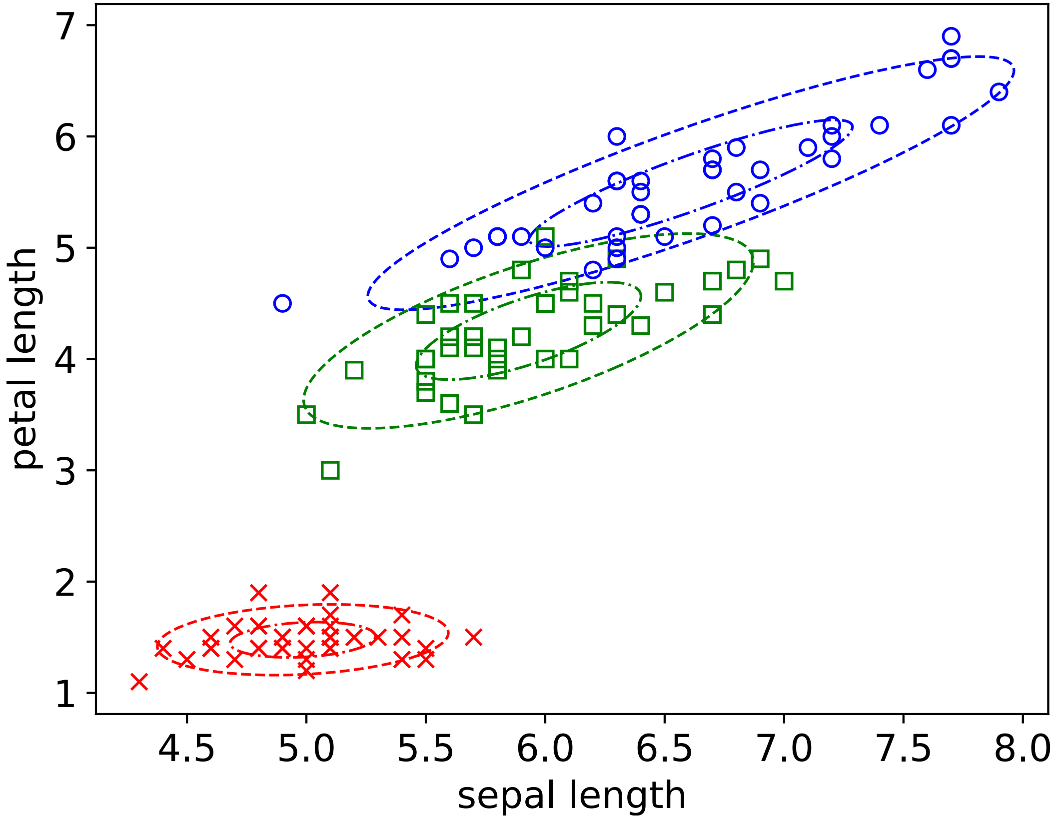


Use linear algebra to understand structure of pair-wise associations [keywords: "eigenvalues" and "eigenvectors"]

Challenge: discover the sub-populations automatically

fit a mixture model!





Analysis of algorithms for fitting mixture models

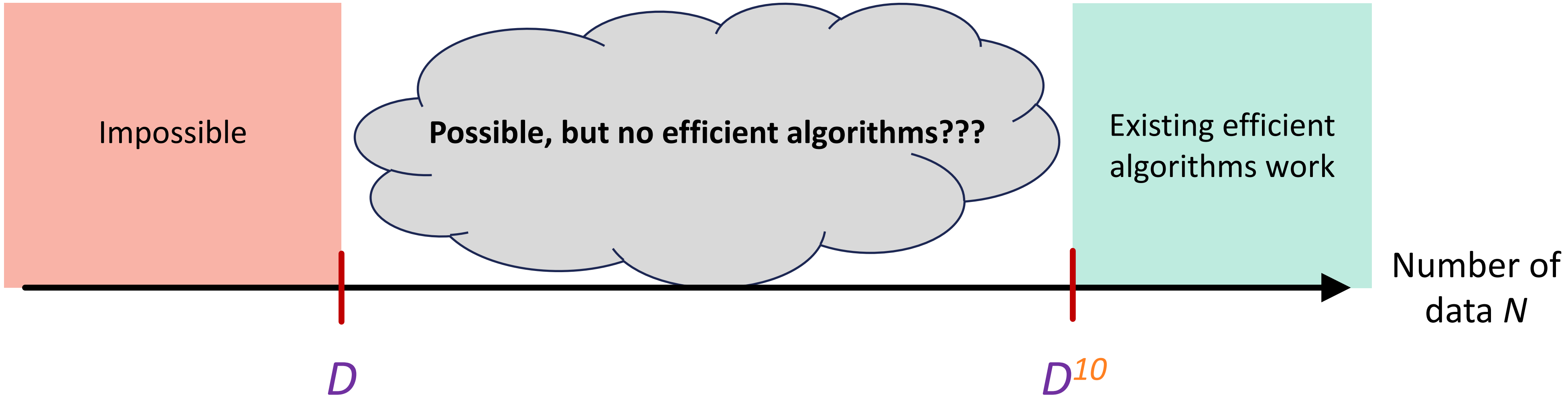
- Local search [Dempster, Laird, Rubin, 1977]
  - Works for discovering TWO "simple" sub-populations [Xu, H., Maleki, 2016]
  - May completely fail for THREE or more sub-populations

- Spectral projection [Vempala & Wang, 2002]
  - Look at structure of PAIR-WISE associations in overall population
  - Works only if sub-populations are very distinct

- Higher-order methods [Pearson, 1894; …; H. & Kakade, 2013; …]
  - Look at structure of K-WAY associations in overall population ($K \geq 3$)
  - Always works … if you have enough data!

Puzzling situation with high dimensional data

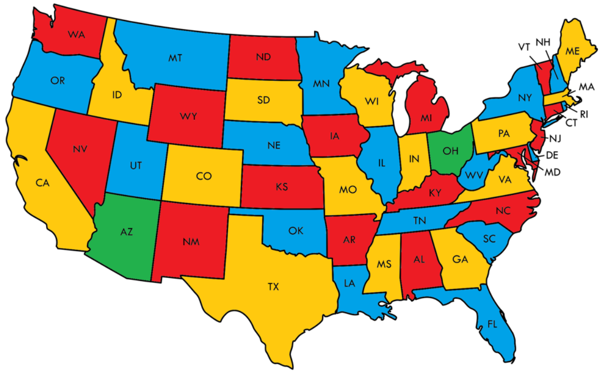
How many data *N* are needed to fit mixture models with *D* attributes?

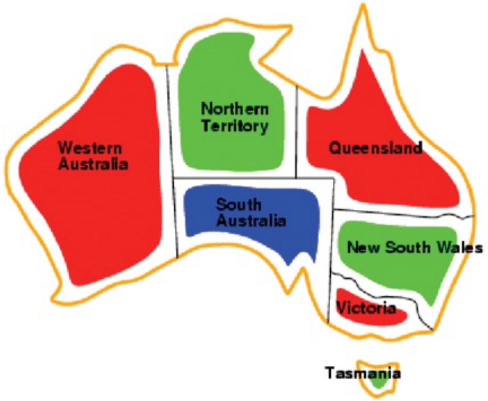


Many statistical problems (e.g., fitting mixture models) appear to exhibit a "statistical-to-computational gap"

Computational intractability

- Some problems cannot be solved by any algorithm [Turing, 1938]
- For some other problems, only known algo. is ≈ "exhaustive search" (e.g., "3-coloring problem", "sudoku", "circuit analysis")





  - Theory of "NP-completeness" explains why: [Cook, 1971; Levin, 1973; Karp, 1972] There's a precise sense in which these hard problems are all "equally hard"
- What about "statistical problems"?

Emerging theory for statistical-to-computational gaps

Theorem [Dudeja & H., 2024]:
For many statistical problems (like fitting mixture models),
every algorithm must:

use a lot of DATA   or   use a lot of TIME   or   use a lot of MEMORY

And some known existing algorithms are "Pareto-optimal":
impossible to improve one of these aspects without worsening another

2. Algorithms in large language models

Language models

- Large Language Models (LLMs) are extremely compelling due to the "naturalness" of their predictions
- Originally studied by Shannon (1948) in his theory of communication

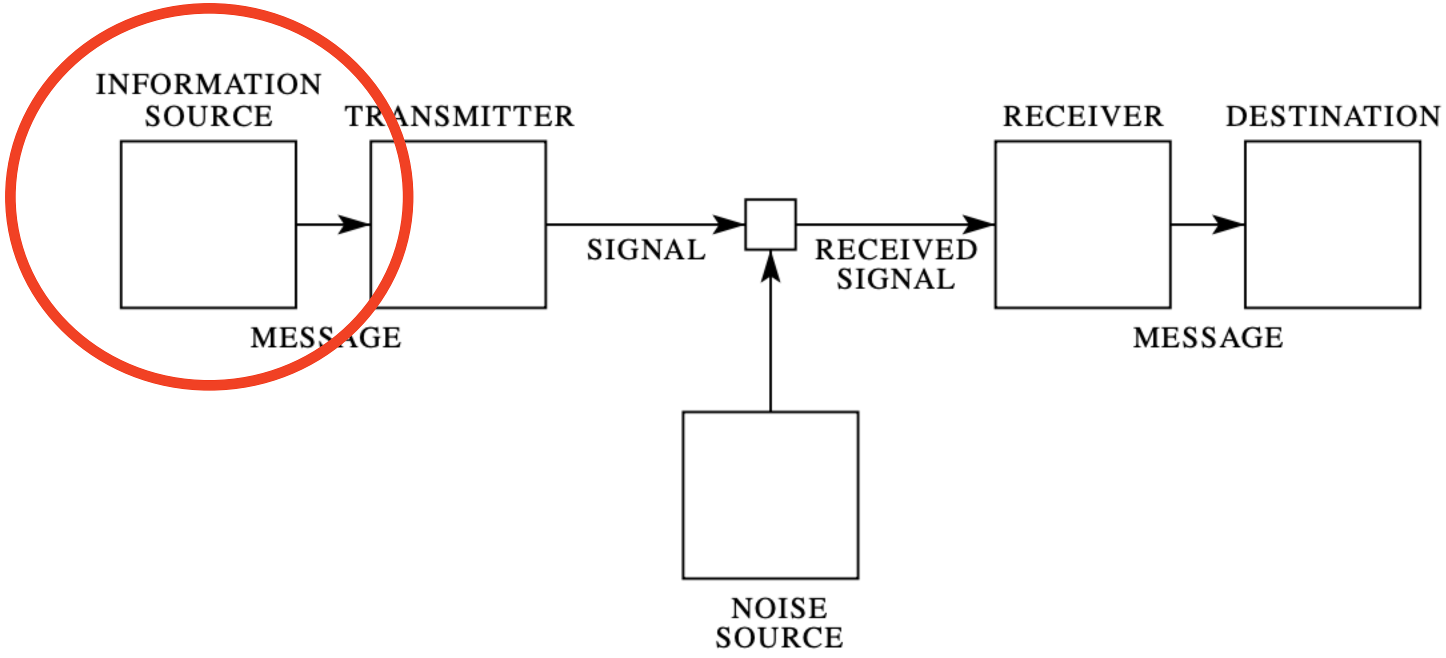


Fig. 1—Schematic diagram of a general communication system.

Shannon's *N*-gram model

- Language model: Given prefix of tokens, give probability of next token
- *N*-gram model: The probability depends only on last *N−1* tokens

$$P(w_T | w_1, \dots, w_{T-1}) = \frac{P(w_{T-N+1}, \dots, w_T)}{P(w_{T-N+1}, \dots, w_{T-1})}$$

- Can be used to predict most likely next token
- Can also be used to randomly generate likely "completions"

Some sequences generated by an *N*-gram model

Prefix of tokens provided (a.k.a. the prompt):

`it is a truth universally ac`

N=1: [...] `mci w aeovmsne drsbwt elo oiwetrcao rne em ok hae lom`

N=2: [...] `o drto t bet it s f aree h at teshas rr l hasis popor`

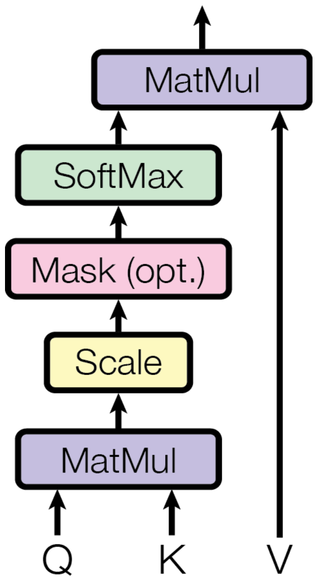
N=3: [...] `es as pred cirse so tiought let of ant forrieng pled`

N=4: [...] `common of could ell his i foung laster are plage omin`

N=5: [...] `quaintance only can better he obliged it is the first`

Fitting *N*-gram models to data

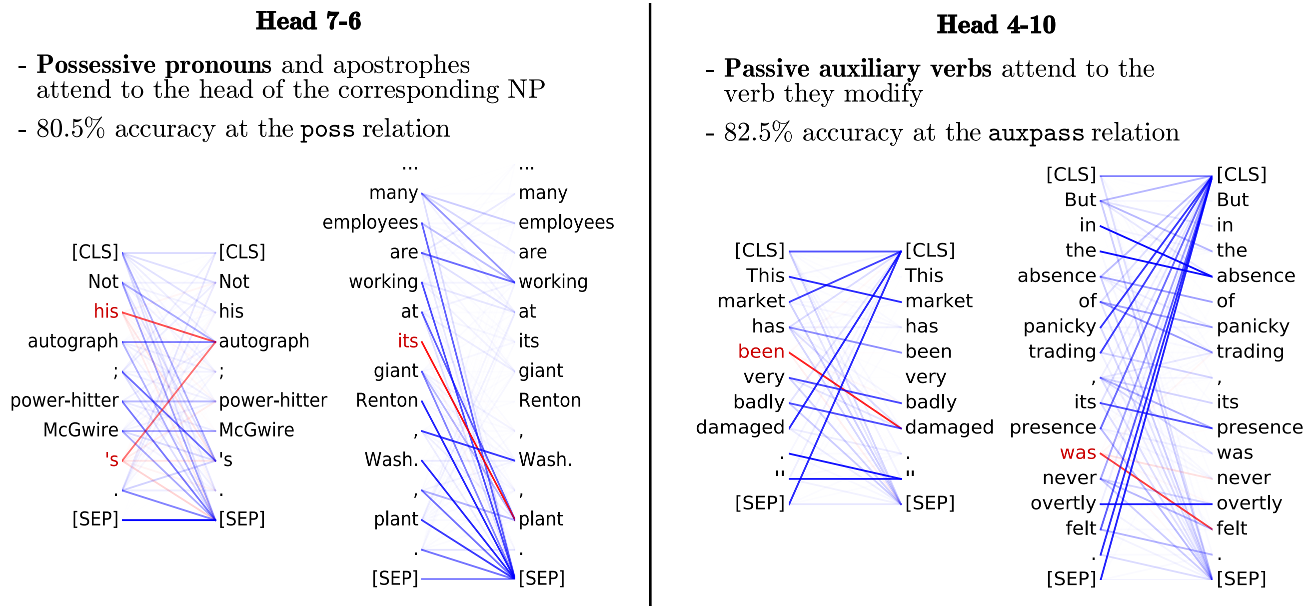
- Lots of methods developed in the 20th century (usually $N < 10$)
- Today:
  - $N = 10^6$ or more
  - $P(w_T | w_1, \dots, w_{T-1})$ computed using a "multi-layer Transformer"
  - Fit to all text on the internet using "Gradient Descent" algorithm



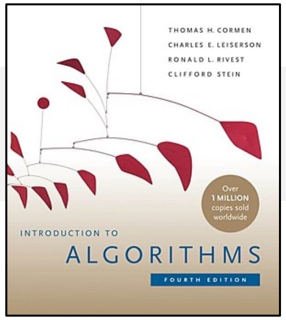
[Vaswani et al, 2017]

What makes LLMs so amazing?

- Ability to make good next-token predictions seems to involve interesting forms of "reasoning" (= algorithmic process)
- How do we know this? Neuroscience for LLMs [e.g., Clark et al, 2019]
  - Discovered some basic "algorithms" implemented by the LLMs (e.g., for rudimentary linguistic analysis and statistical inference)



What problems can Transformers solve?



- Forget about complex linguistic analysis, etc., …
  Can Transformers efficiently solve simple computational problems?
- Example: 2-SUM
  Given N numbers $x_1, x_2, \ldots, x_N$, are there 2 of them that add up to 0?
  - Even a "single-layer Transformer" can solve 2-SUM [Sanford, H., Telgarsky, 2023]

$$\boxed{1,\ 4,\ 2,\ 8,\ -2,\ 5,\ -7}$$

- Example: 3-SUM
  Given N numbers $x_1, x_2, \ldots, x_N$, are there 3 of them that add up to 0?
  - A "single-layer Transformer" CANNOT solve 3-SUM [Sanford, H., Telgarsky, 2023]
  - What about multi-layer Transformer? We don't know!

Problems that need multi-step reasoning

John plays football. [...]
The NFL game is on Sunday. [...]
On what day does John play?

- **Contextual knowledge:**
  John → football
  NFL → Sunday
- **Prior knowledge:**
  football → NFL

Theorem [Sanford, H., Telgarsky, 2024; Wang, Nichani, Bietti, Damian, H., Lee, Wu, 2025]:
Multi-layer Transformers perform "multi-step reasoning" as well as (and no better than) MapReduce (parallel computation) algorithms

Takeaways

- Statistical models are everywhere and have many applications
- Fitting models to data: both an algorithmic and a statistical problem
- Large language models are also statistical models;
  understanding how/why they work is a multidisciplinary challenge

Thank you!