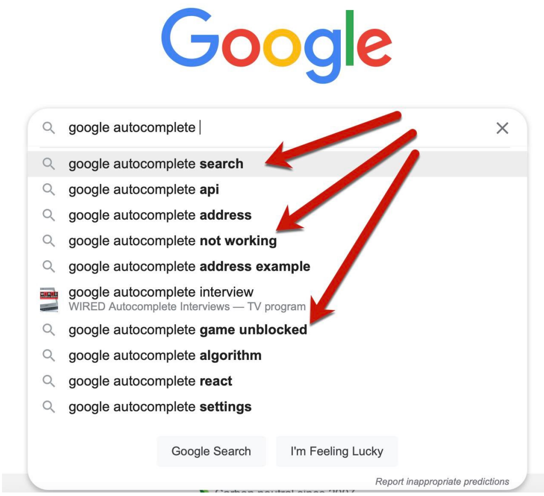
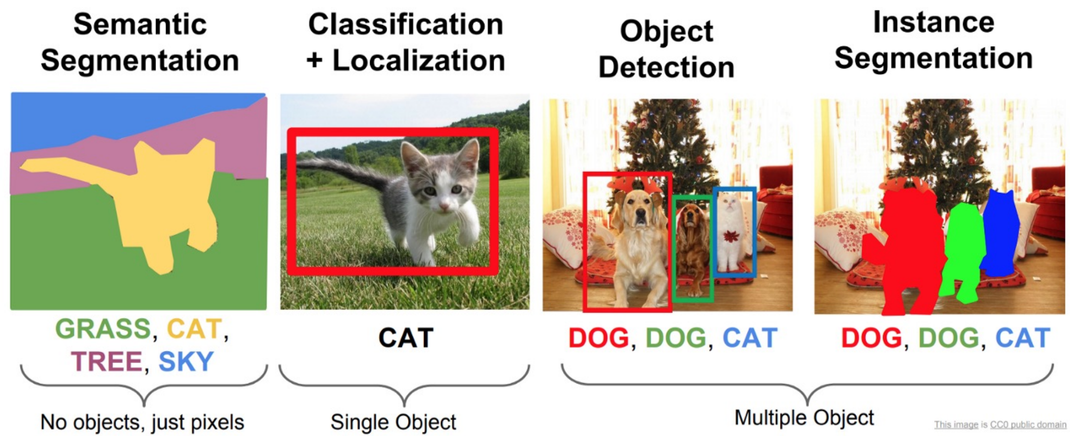
On the unreasonable effectiveness of scaling data and parameters in machine learning

Rahul G. Krishnan
Assistant Professor in CS & LMP (Medicine)
CIFAR AI Chair

Ongoing work with Ethan Choi and Aryan Dhar

Machine learning is ubiquitous



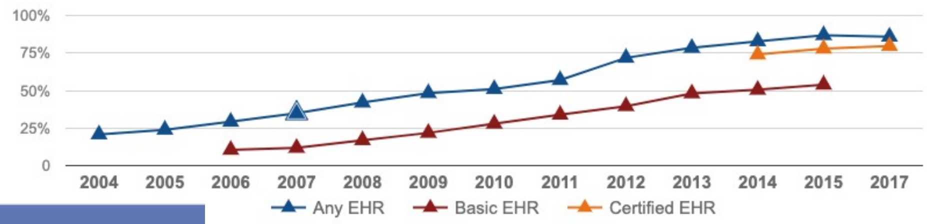


Use patterns in what a billion people look for to predict what *you* might be searching for

Learning patterns to identify classes from a million labelled datapoints

The opportunity with electronic medical records

The adoption of electronic medical records has dramatically increased in the last two decades!



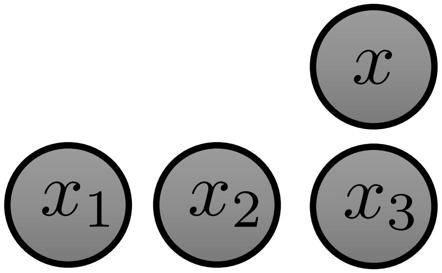
Electronic medical records give us a view into a patient's underlying physiological state.

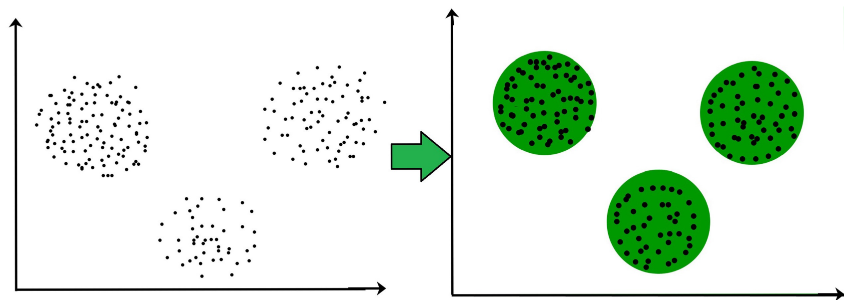
Source: https://www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption

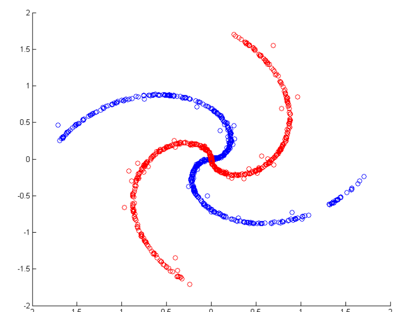
Supervised learning

- Step 1: Collect a dataset or curate a subset of data with labels from an existing dataset
- Step 2: Learn the model using the dataset
- Step 3: Use the output of the model to build software to help clinicians reach better decisions, faster.
- **Examples**: Logistic regression, random forests, XGBoost, Deep neural networks

Unsupervised learning

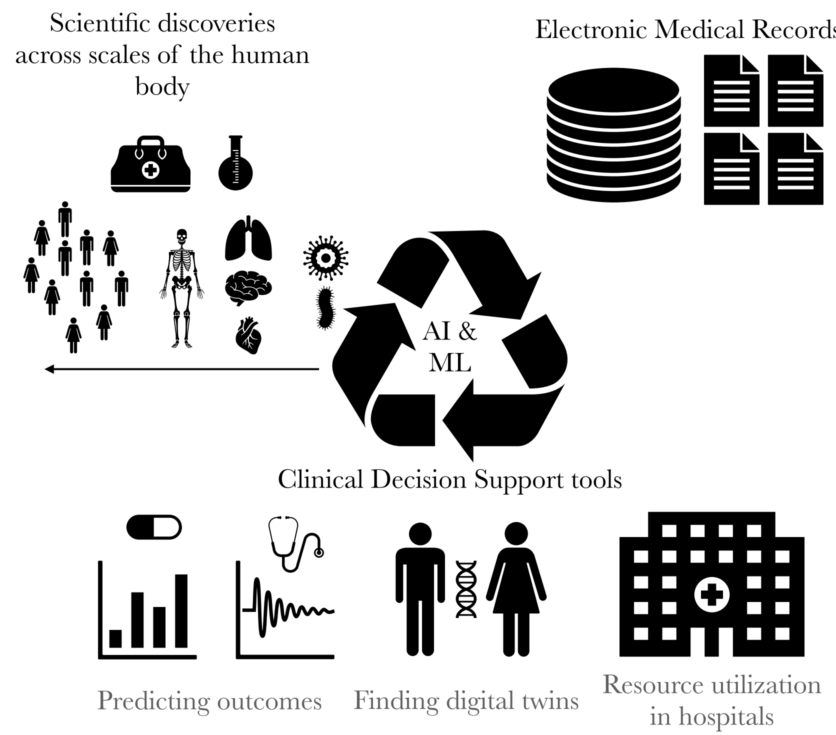






- Step 1: Collect a dataset or curate a subset of data with labels from an existing dataset
- Step 2: Learn the model using the dataset
- Step 3: Use parameters of the model uncover insights about the data and validate with domain experts
- **Examples:** Nearest neighbors, latent factor models, hidden markov models, variational autoencoders

Vision: A learning healthcare system



Source: https://www.medicaldevice-network.com/analysis/ai-in-healthcare-2021-2/

Case Study: Machine Learning for Disease Phenotyping

Why should healthcare care about language models?

- Text data is an important mode of storing and transcribing information in healthcare
- **Nurse and doctor notes**
  - Routine part of care for critical patients, as well as those suffering from chronic diseases
  - Unstructured but rich source of data about patient disease state
- **Status quo:**
  - Lot of promise around the use of machine learning for healthcare
  - Language models can help with extracting patient information, summarizing state and forming embeddings of clinical concepts [Alsentzer et. al]

Language models over the years

- Pre-2013
  - **Ethos:** Need to have models that capture fine-grained structure in sentences
  - Parse trees
  - N-gram language models
  - Works well but brittle when sentence syntax deviates from training data
- Post-2013
  - **Ethos:** The context of a word is sufficient to predict the word
  - Word2Vec, Recurrent neural networks, transformer

A (brief) history of language models

- **What is it:** Language model is a statistical model of natural language text
- **How is it trained:** By maximizing the likelihood of a word/sentence

The dog jumped over the creek.

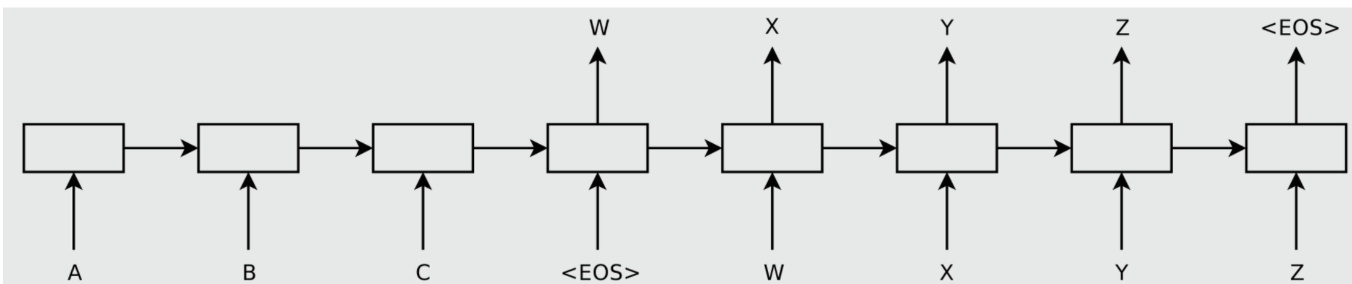
w1 w2 w3 w4 w5 w5.

Each wi is a word in a vocabulary set [1…….V]

Goal: Maximize P(w1….w5)

Thought experiments on the hardness of modeling language

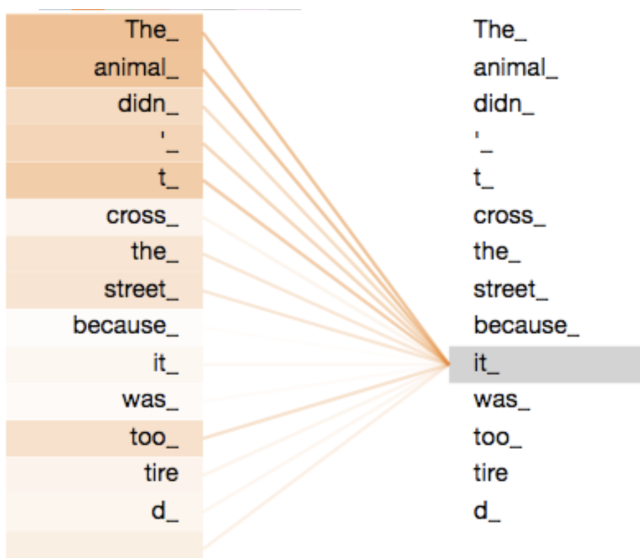
Modeling language via a recurrent process



Source: Sutskever et al. (2014)

Treat language modeling as many small supervised learning tasks – predict the next word given the previous word!

The last few years have seen sentences modeled via attention



Source: Jay Alammar (2018)

- 2014: Seq2seq models
- 2015: Attention
- 2017: Transformer
- 2018: BERT (110M parameters)
- 2019: GPT-2 (1.5B parameters)
- 2020: GPT-3 (175B parameters)
- April 4, 2022: PaLM (540B parameters)

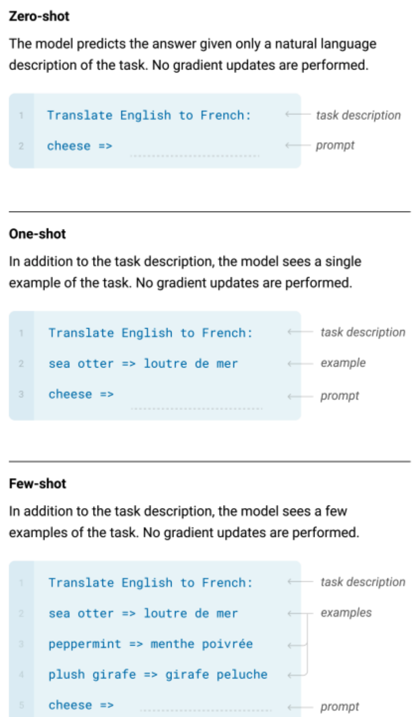
Recent trends in large language models

- Language Models are Few-Shot Learners, Brown et. al
  - 3 key ingredients
    - Attention-based transformers
    - Scales up the models to be [very] overparameterized
    - Trains on very very large datasets

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Performance as a function of scaling

Model Size



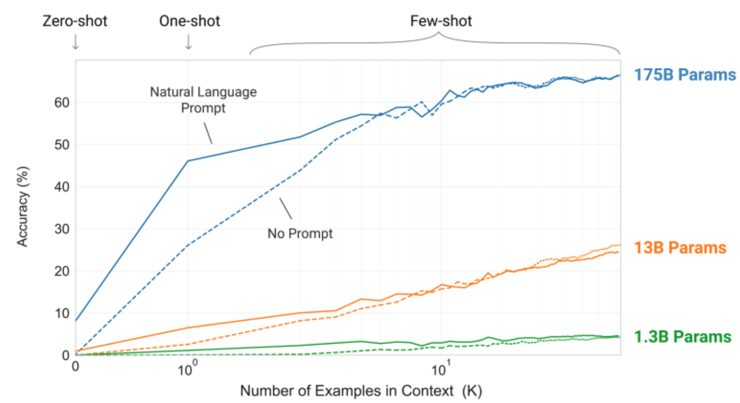
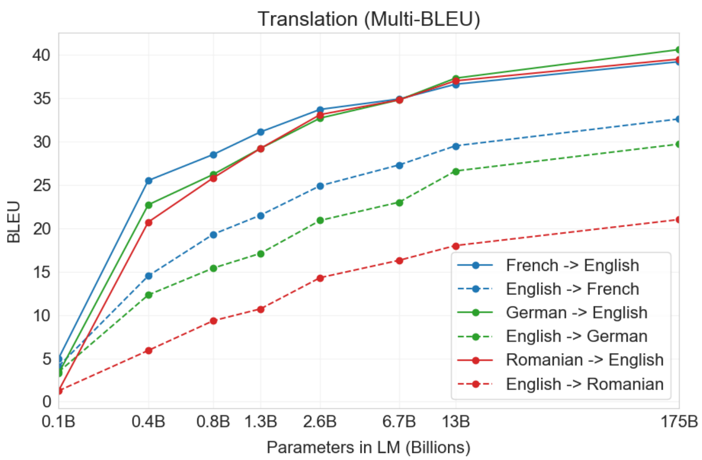
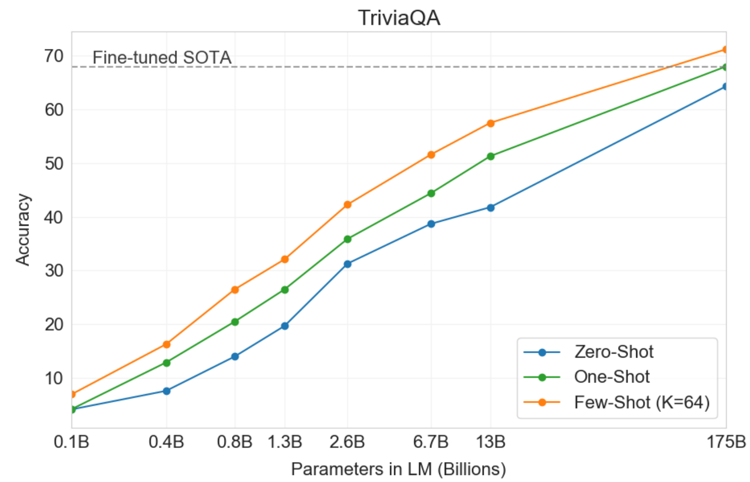


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

Source: Brown et al. (2020)

One model many tasks

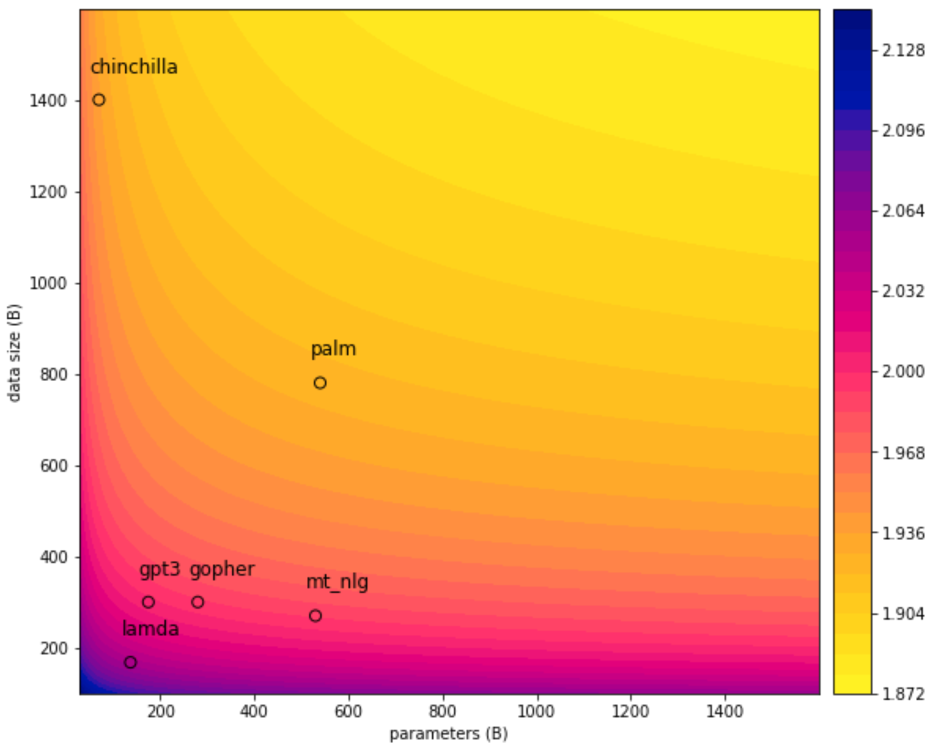




Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	**92.0**[KKS+20]	**78.5**[KKS+20]	**87.2**[KKS+20]
GPT-3 Zero-Shot	**80.5***	68.8	51.4	57.6
GPT-3 One-Shot	**80.5***	71.2	53.2	58.8
GPT-3 Few-Shot	**82.8***	70.1	51.5	65.4

Table 3.6: GPT-3 results on three commonsense reasoning tasks, PIQA, ARC, and OpenBookQA. GPT-3 Few-Shot PIQA result is evaluated on the test server. See Section 4 for details on potential contamination issues on the PIQA test set.

Data >> size of model



An empirical analysis of compute-optimal large language model training, Hoffman et. al