

Business Analytics and Data-Driven Decision Making
**Text Analytics - 02: Topic
Modelling and Large
Language Models**

Raghava Mukkamala

**Associate Professor & Director,
Centre for Business Data Science**

Copenhagen Business School, Denmark

Email: rrm.digi@cbs.dk, Centre: <https://cbsbda.github.io/>

**Many of the slides have been taken and adapted from:
Speech and Language Processing by Dan Jurafsky and James H. Martin
<https://web.stanford.edu/~jurafsky/slp3/>**



Outline

- Topic Modeling
- Deep Learning Models in NLP: Large Language Models

The problem with information



- As more information becomes available, it becomes more difficult to find and discover what we need
- We need new tools to help us organize, search, and understand these vast amounts of information

Topic modeling



Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives

- 1) Discover the hidden themes that pervade the collection
- 2) Annotate the documents according to those themes
- 3) Use annotations to organize, summarize, search, form predictions

Topic Modeling

Suppose you want to learn something about a corpus that's too big to read

What topics are trending today on Twitter? (half a billion tweets daily ..)

What issues are considered by Congress (and which politicians are interested in which topic)? (hundreds of bills each year..)

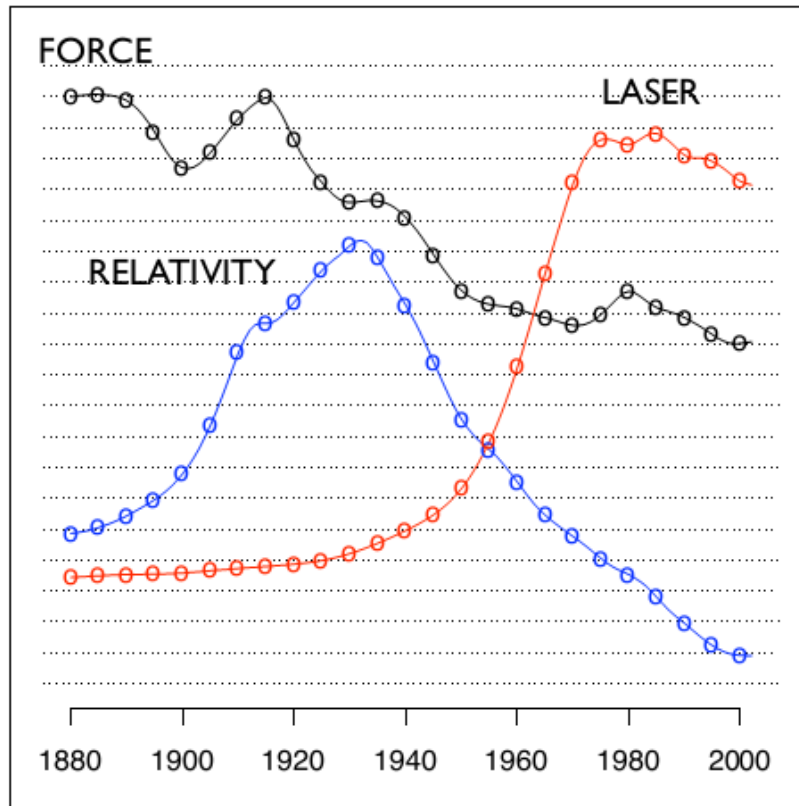
Solution: Why don't we just throw all these documents at the computer and see what interesting patterns it finds?

Discover topics from a corpus

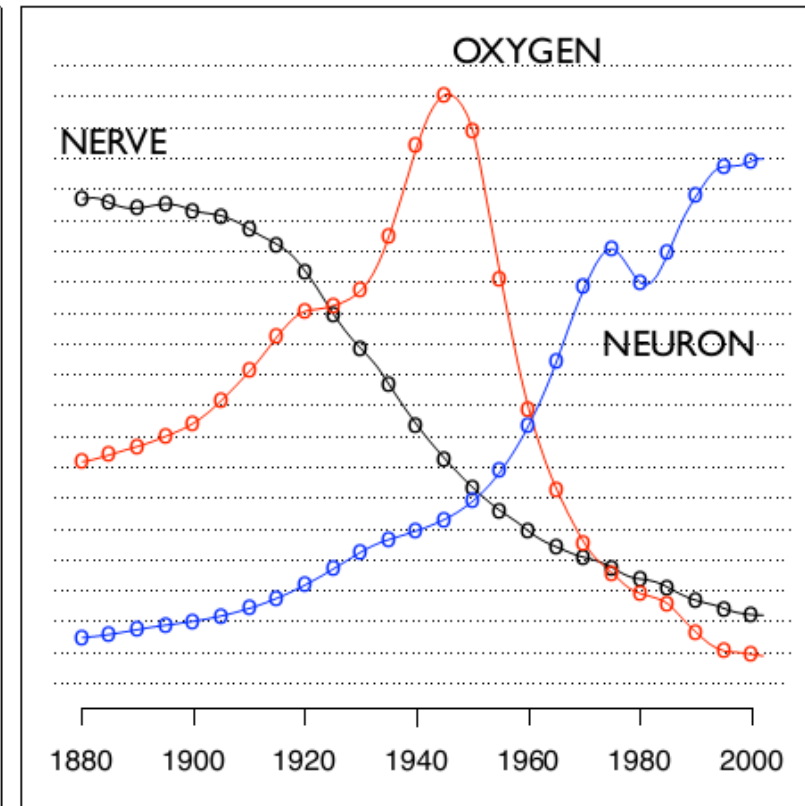
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Model the evolution of topics over time

"Theoretical Physics"



"Neuroscience"



Probabilistic modeling with Bayesian Inference

- Treat data as observations that arise from a generative probabilistic process that includes variables
 - For documents, the hidden variables reflect the thematic structure of the collection
- Infer the hidden structure using *posterior inference*
 - What are the topics that describe this collections?
- Situate new data into the estimated model
 - How does the query or new document fit into the estimated topic structure

Intuition behind LDA

Seeking Life's Bare (Genetic) Necessities

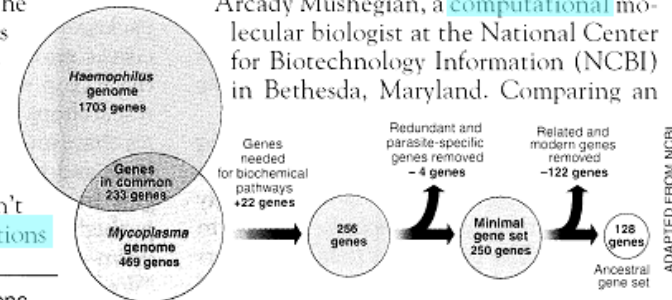
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

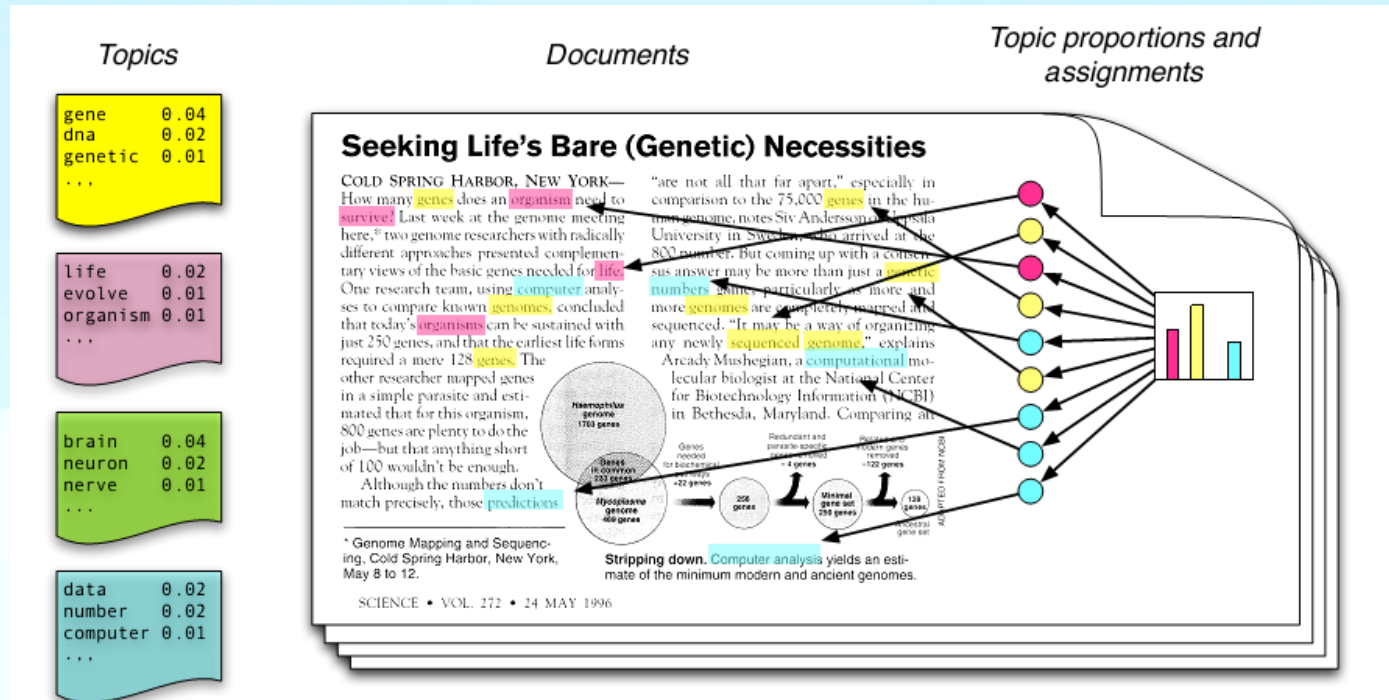
“are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

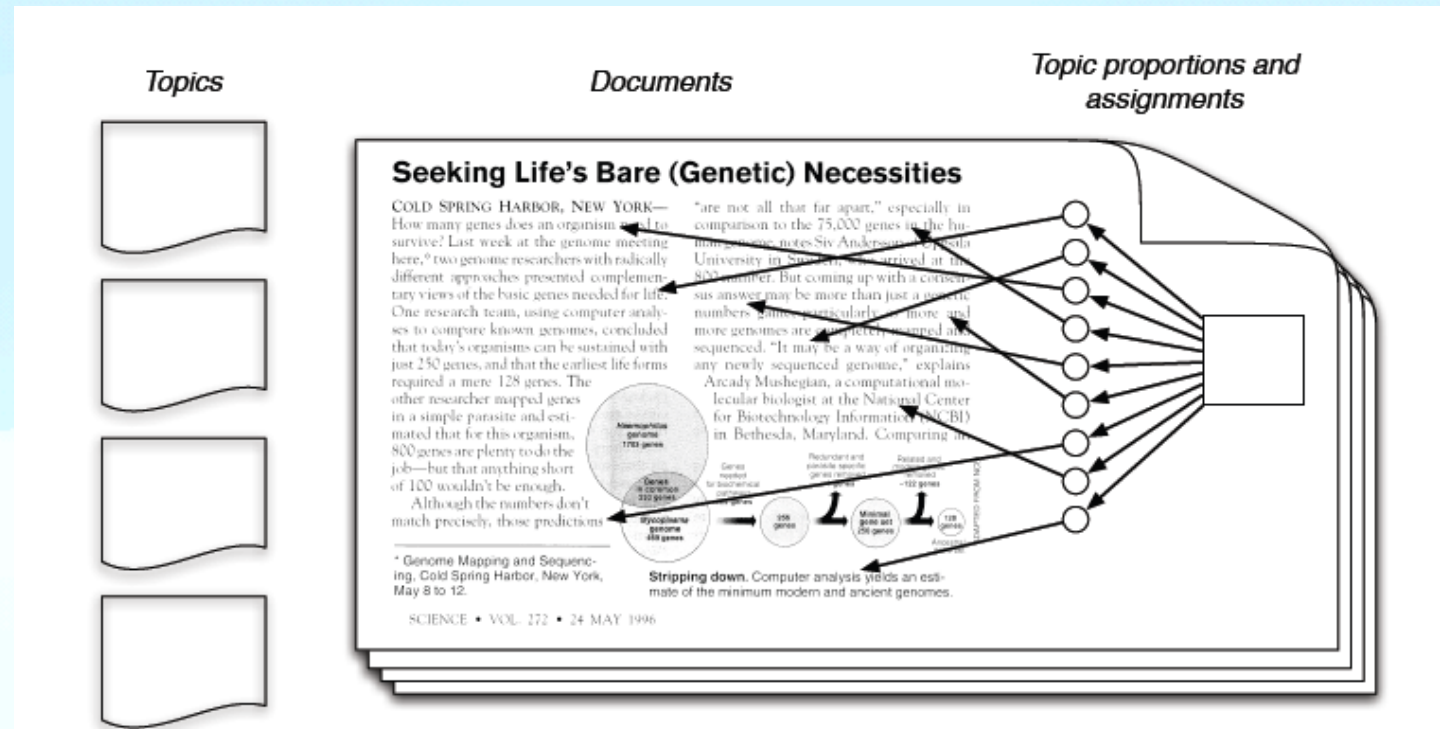
Simple intuition: Documents exhibit multiple topics.

Generative model



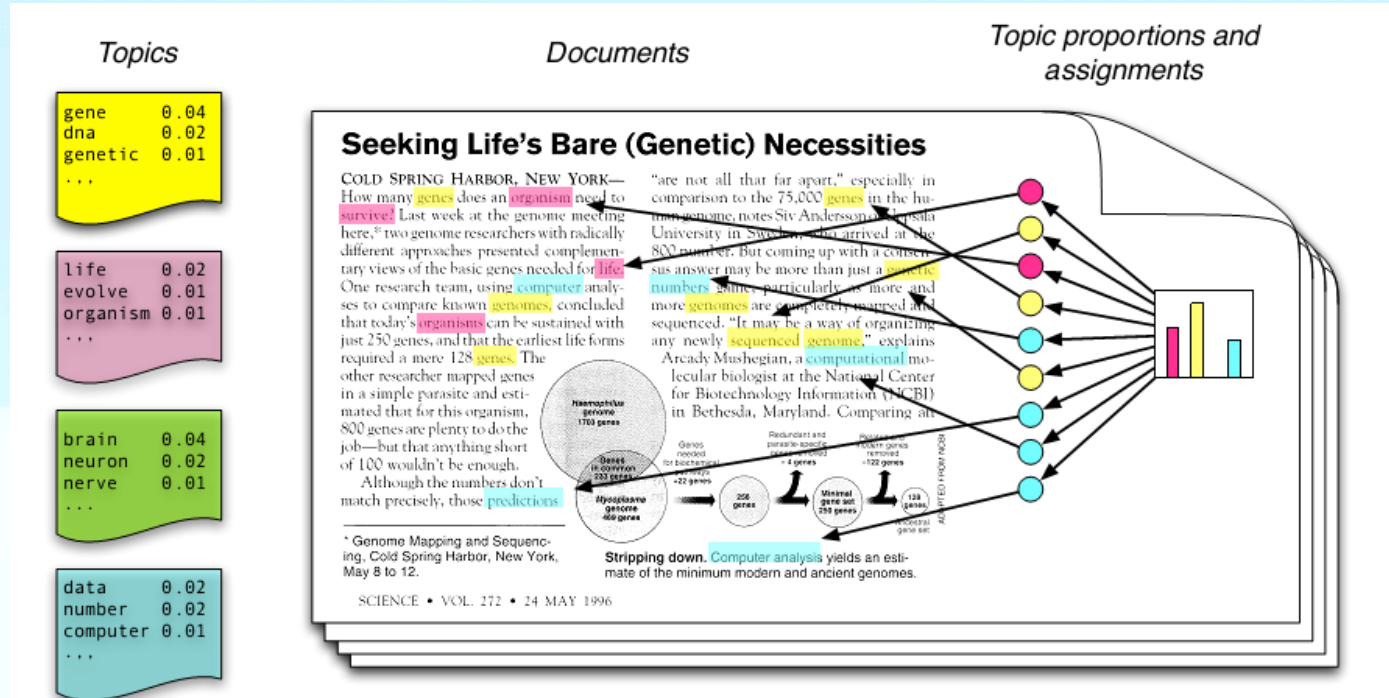
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

The posterior distribution



- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

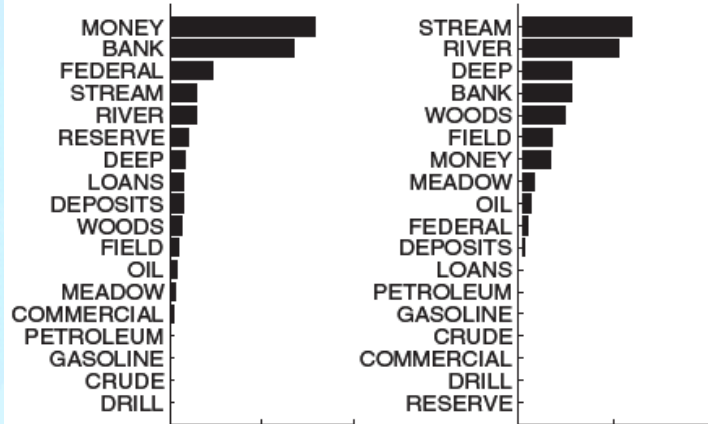
Topic Modeling



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

TOPIC 1

TOPIC 2



Steyvers, Mark, and Tom Griffiths. "Probabilistic topic models." Handbook of latent semantic analysis 427.7 (2007): 424-440.

Topic Models

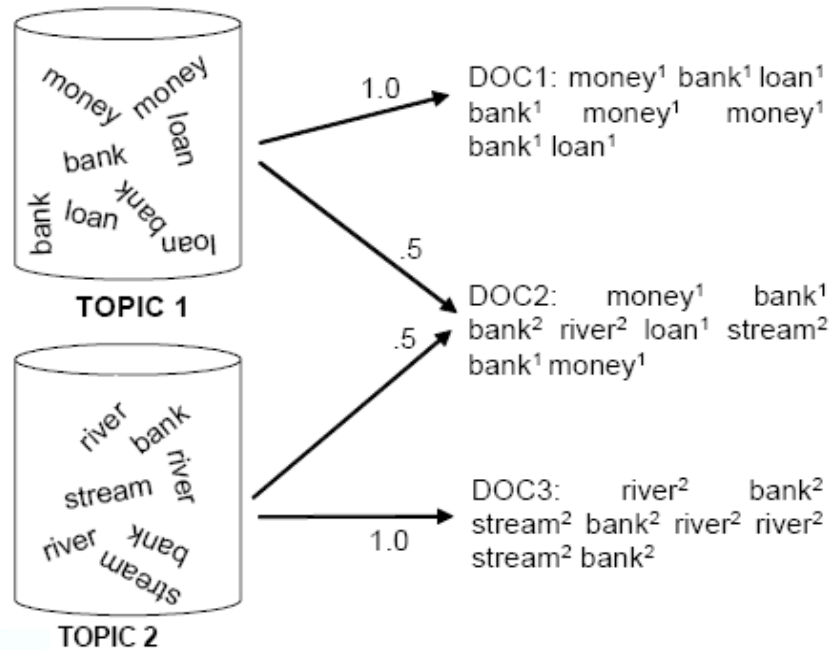
3 latent variables:

Topic distribution per doc (topic-doc-matrix)

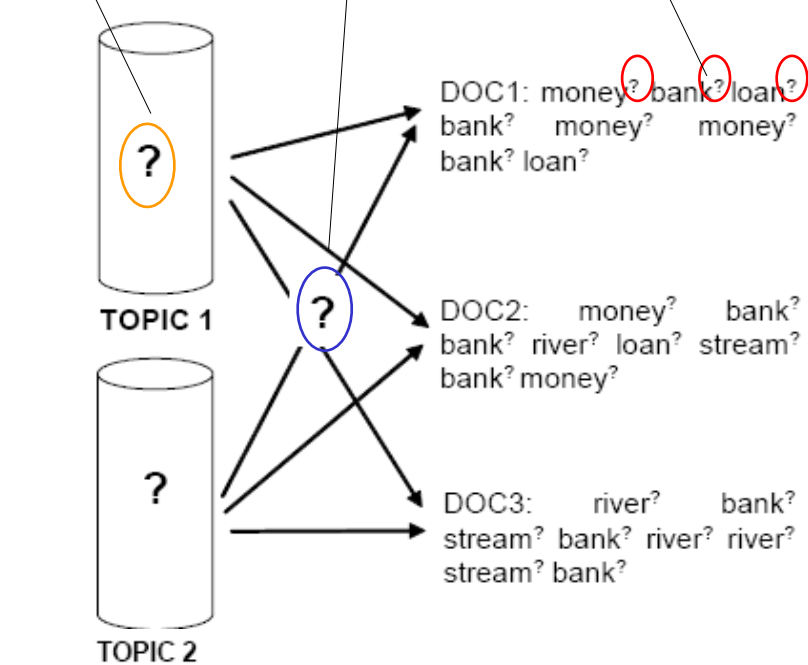
Word distribution per topic (word-topic-matrix)

Topic word assignment

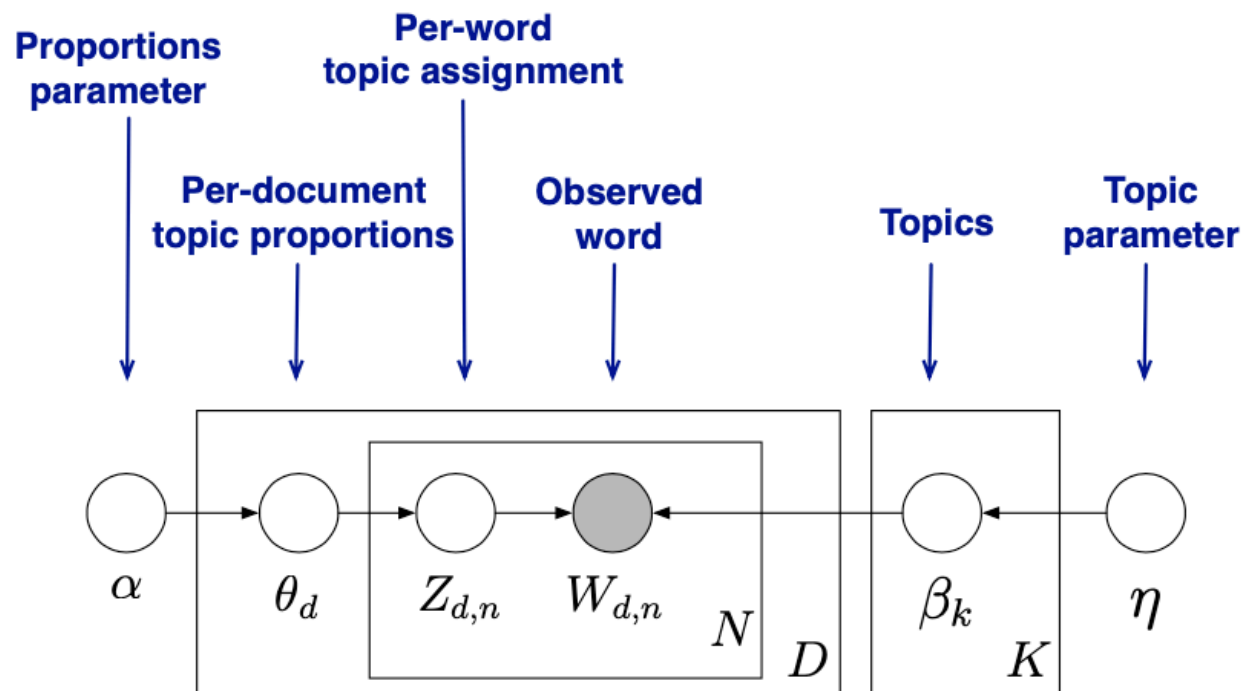
PROBABILISTIC GENERATIVE PROCESS



STATISTICAL INFERENCE



LDA as a graphical model



- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data

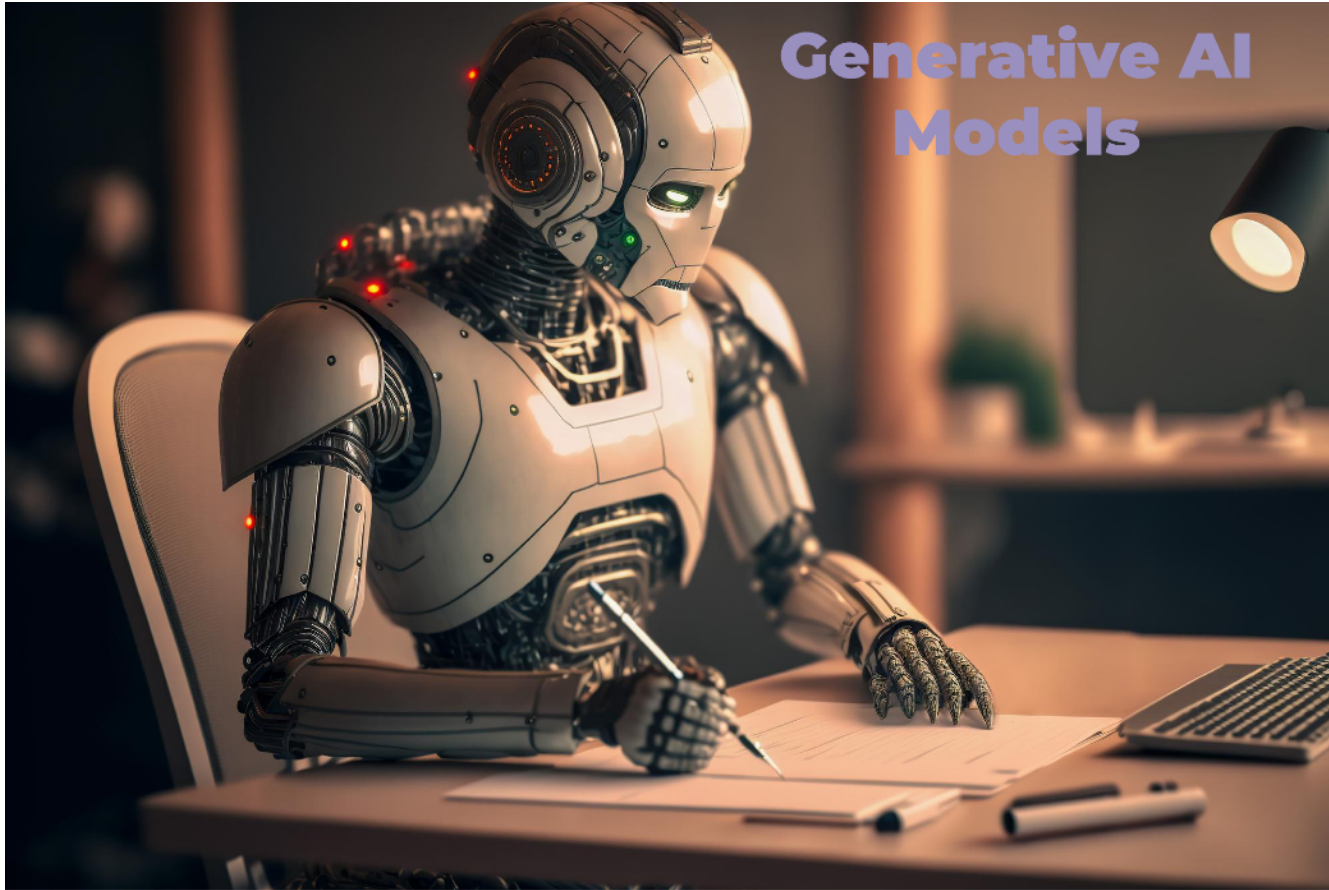
Topic models

- **Observed variables:**
 - Word-distribution per document
- **3 latent variables**
 - Topic distribution per document : $P(z) = \theta^{(d)}$
 - Word distribution per topic: $P(w, z) = \beta$
 - Word-Topic assignment: $P(z|w)$
- **Training: Learn latent variables on trainings-collection of documents**
- **Test: Predict topic distribution $\theta^{(d)}$ of an unseen document d**

GENERATIVE AI AND LARGE LANGUAGE MODELS



Expectations



Are the ChatGPT and other Generative AI models good or bad?

Why and how have these models become popular recently? What capabilities do they have?

Outline

- Why and how the pre-trained generative models got popular recently?
 - Transformer Models and Attention Mechanism
 - Self-Supervised Learning
- Popular Large Language Models
- Opportunities
- Challenges and Risks
- Conclusion

PRETRAINED GENERATIVE MODELS



Some AI Jargon

Generative vs Foundational vs Pretrained vs Large Language Models

- Generative AI generates images, music, speech, code, video, or text while interpreting and manipulating pre-existing data.
- Generative AI is not a new concept, and machine-learning techniques behind it evolved over the past decade, but the latest approach based on ‘**Transformers**’ made significant progress, especially in NLP domain.
- Large Language Models (LLMs) were Transformer models explicitly trained on large amounts of text data for NLP tasks and contained huge parameters (> 100 million), e.g., BERT, GPT, T5.
- These large models are also called foundational models, as they can be used for developing more advanced and complex models, e.g., sentiment analysis, Q&A, etc.


What is generative AI, what are foundation models, and why do they matter?

<https://www.ibm.com/blog/what-is-generative-ai-what-are-foundation-models-and-why-do-they-matter/>

Challenges of Language Models (LM)

Stochastic Parrots


Dan Jurafsky



Probabilistic Language Modeling

- Goal: compute the probability of a sentence or sequence of words:
 $P(W) = P(w_1, w_2, w_3, w_4, w_5 \dots w_n)$
- Related task: probability of an upcoming word:
 $P(w_5 | w_1, w_2, w_3, w_4)$
- A model that computes either of these:
 $P(W)$ or $P(w_n | w_1, w_2 \dots w_{n-1})$ is called a **language model**.
- Better: **the grammar** But **language model** or **LM** is standard

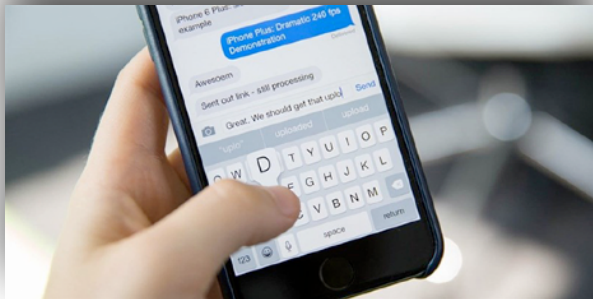
Dan Jurafsky



The Chain Rule applied to compute joint probability of words in sentence

$$P(w_1 w_2 \dots w_n) = \prod_i P(w_i | w_1 w_2 \dots w_{i-1})$$

$P(\text{"its water is so transparent"}) =$
 $P(\text{its}) \times P(\text{water} | \text{its}) \times P(\text{is} | \text{its water})$
 $\times P(\text{so} | \text{its water is}) \times P(\text{transparent} | \text{its water is so})$



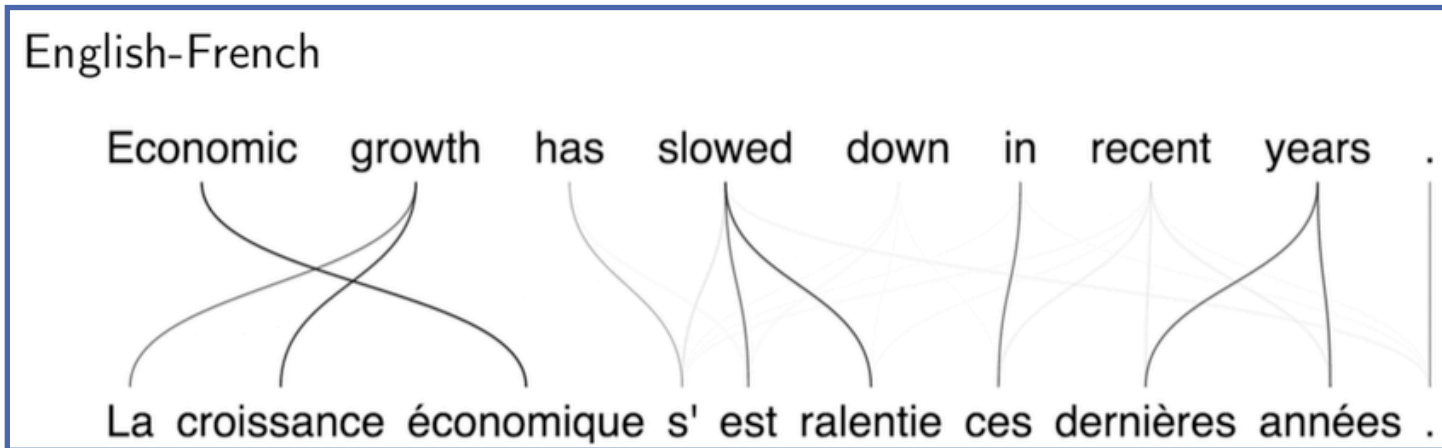
Web search engine / ...

I saw a cat|

I saw a cat on the chair
I saw a cat running after a dog
I saw a cat in my dream
I saw a cat book

- One of the main issues for language modeling is data sparsity.
- Humans are creative in the way they write, so number of valid word combinations is in principle infinite
- Finding exactly the same combination of words is very rare unless they are common sentences.

Challenges of LM and Machine Translation



The animal did not cross the street
as **it** is too wide

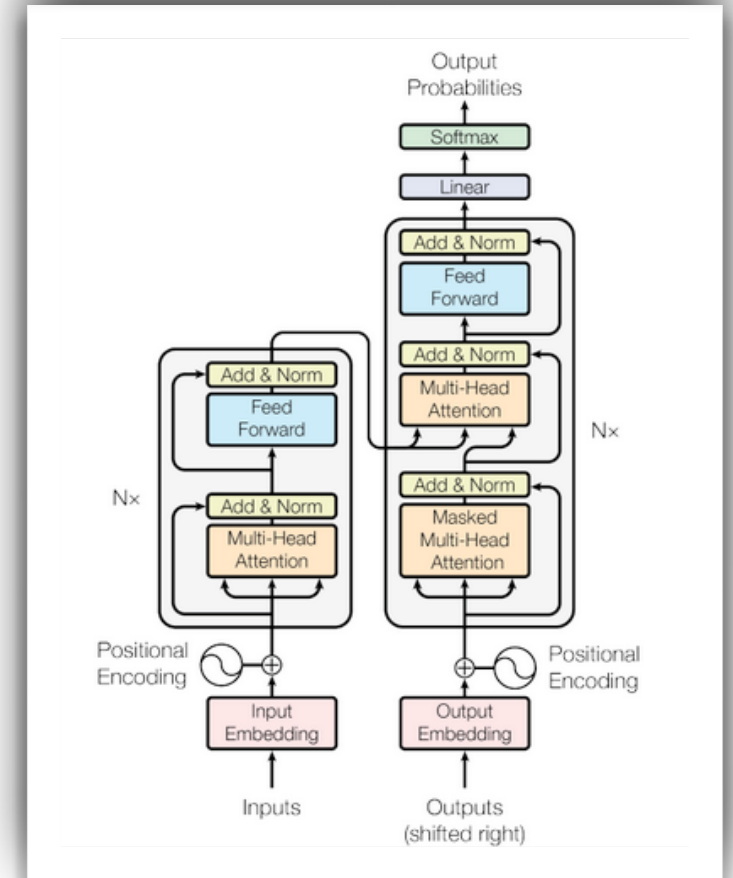
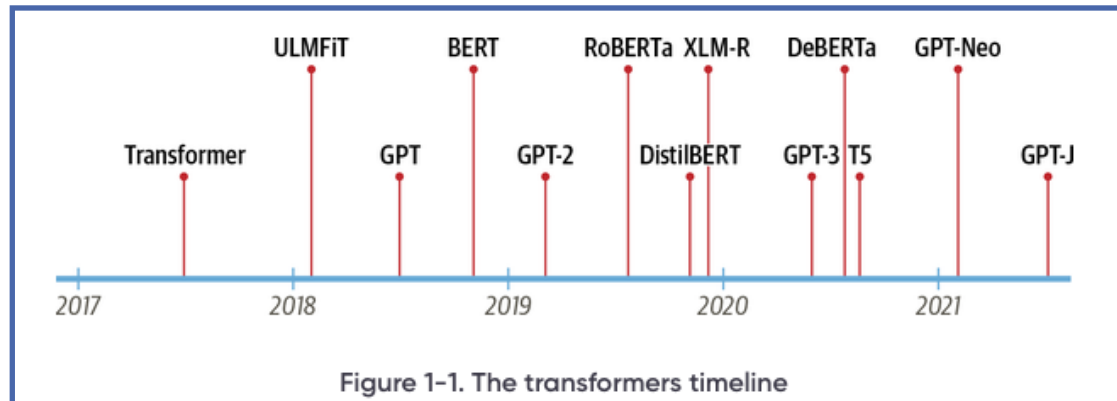
What does **it** means here? Street or animal?

“**Griezmann's** announcement comes as a bit of a shock. After enduring the drama surrounding **his** potential last summer, many thought **he** was committed to Atletico for more than a year, but the **Frenchman** seems to have changed **his** mind.”

Teaching this type of common sense/
intelligence/intuition to the algorithms has
been a problem till recently

The highlighted words refer to the same person – Griezmann, a popular football player. It's not that difficult for us to figure out the relationships among such words spread across the text. However, it is quite an uphill task for machines. This is where the Transformer concept plays a major role.

Journey of Transformer Models

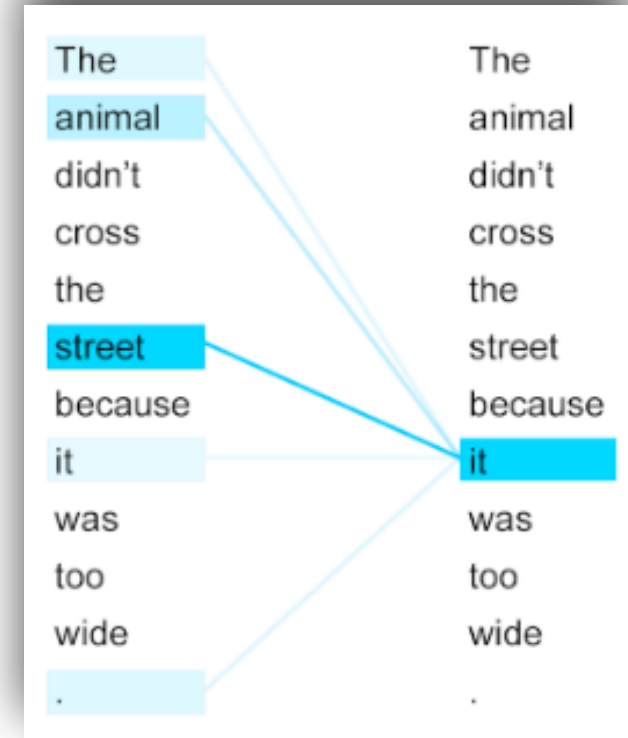
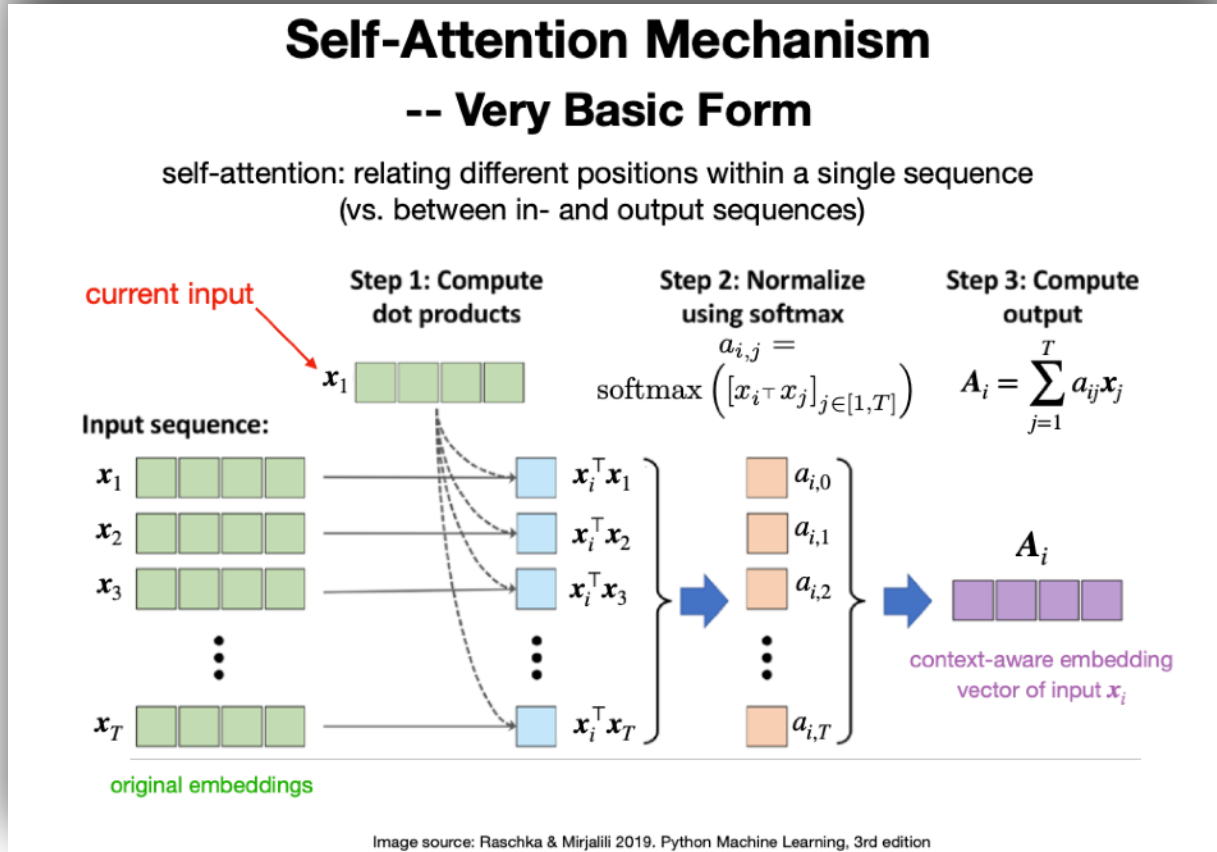


Transformer Model
= **Auto Encoder + Autoregressive LM**

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

<https://www.oreilly.com/library/view/natural-language-processing/9781098136789/ch01.html>

What is attention mechanism?



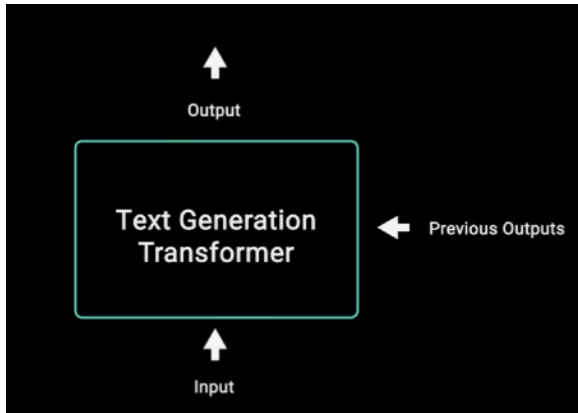
Using the attention mechanism, the model will identify which words it needs to *attend to* or *focus on* when parsing the word **'it'**.

How do Transformers Work in NLP? A Guide to the Latest State-of-the-Art Models. <https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/>

Sebastian Raschka. STAT 453: Introduction to Deep Learning and Generative Models. <https://sebastianraschka.com/blog/2021/dl-course.html>

What is attention mechanism?

Generated story using Transformer models



Our input: *"As Aliens entered our planet".*

Transformer output: *"and began to colonized Earth, a certain group of extraterrestrials began to manipulate our society through their influences of a certain number of the elite to keep and iron grip over the populace."*



Again, why did it generate a depressing kind of story?

Illustrated Guide to Transformers-Step by Step Explanation.

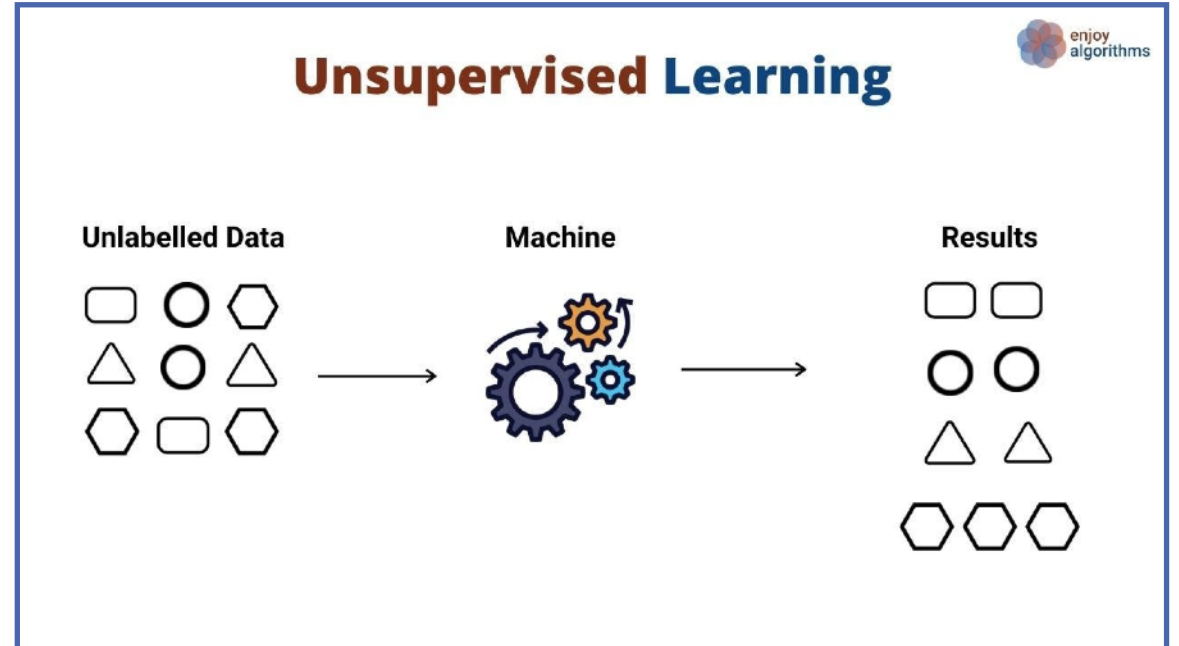
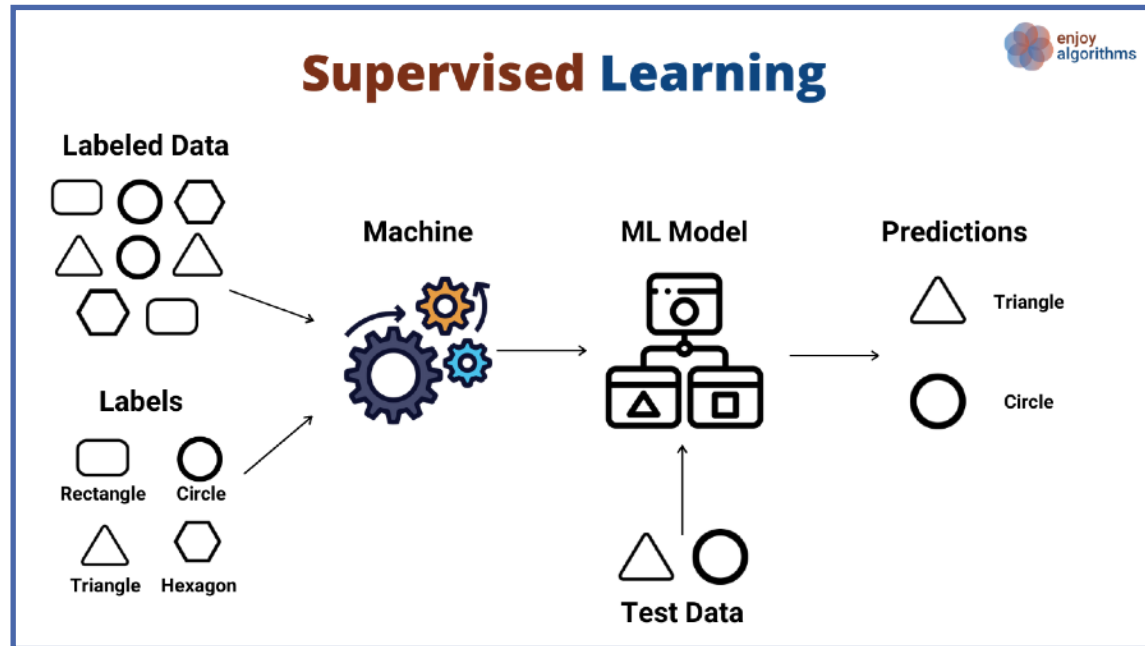
<https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0>

<https://www.youtube.com/watch?v=4Bdc55j80I8>

SEMI-SUPERVISED LEARNING



Supervised vs Unsupervised Learning



<https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning>

Self-supervised Learning

- Generating labeled datasets is quite a big challenge.
- Using a Semi-supervised/self-supervised approach, you can generate huge labeled datasets for pertained models in an automated fashion.
- Download all the pages from Wikipedia and write a program to automate the data labeling process partially.

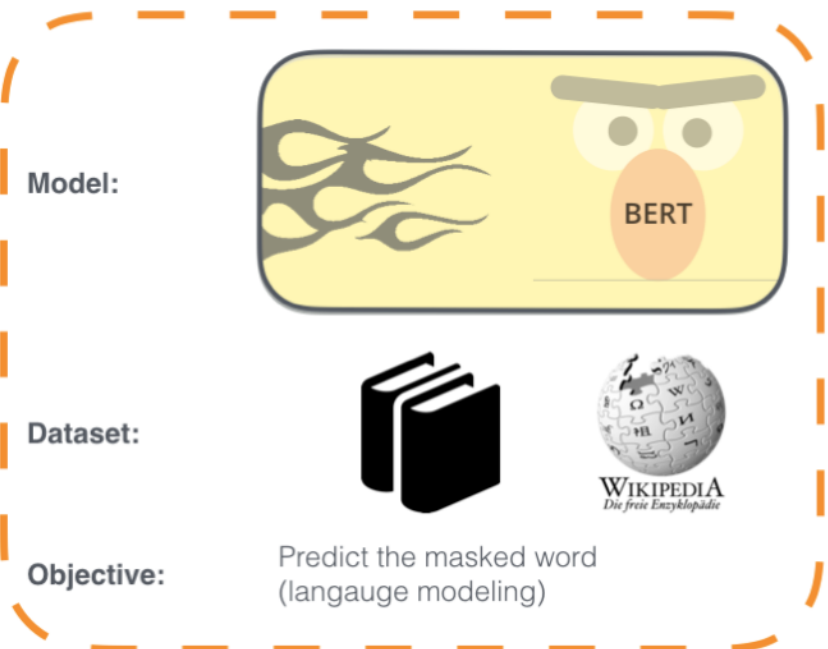
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

GPT-3 training dataset

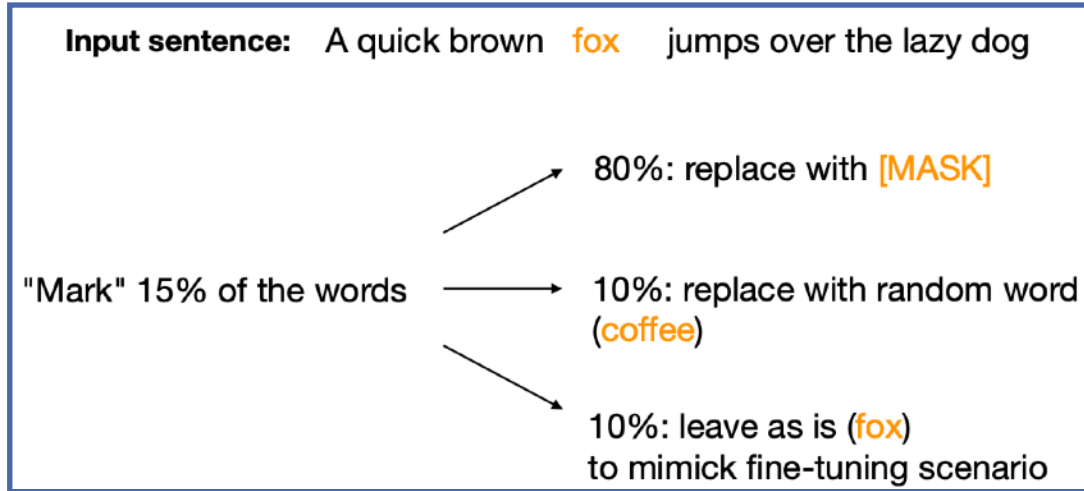
1 - **Semi-supervised** training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step



Self-supervised Training in BERT



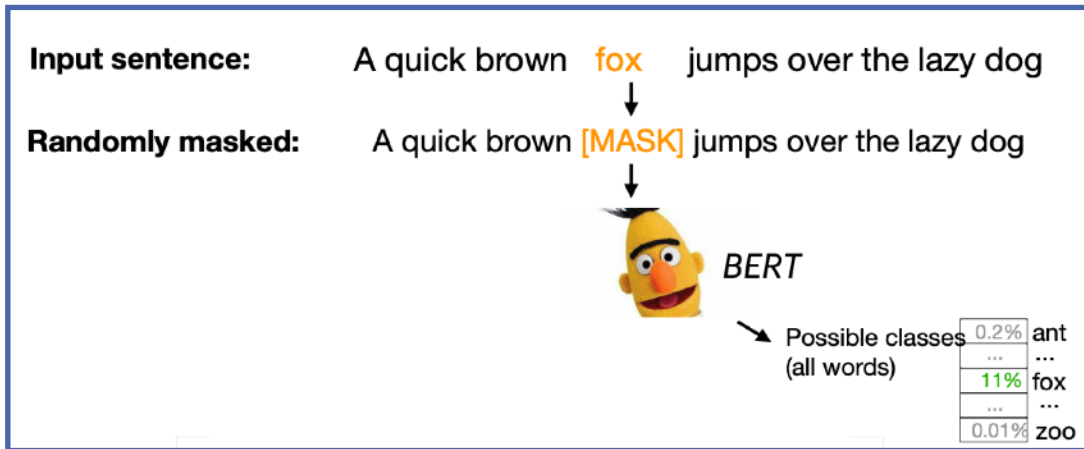
Next Sentence Prediction The next sentence prediction task can be illustrated in the following examples.

Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]

Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]

Label = NotNext



The LLMs use several similar and complex training processes.

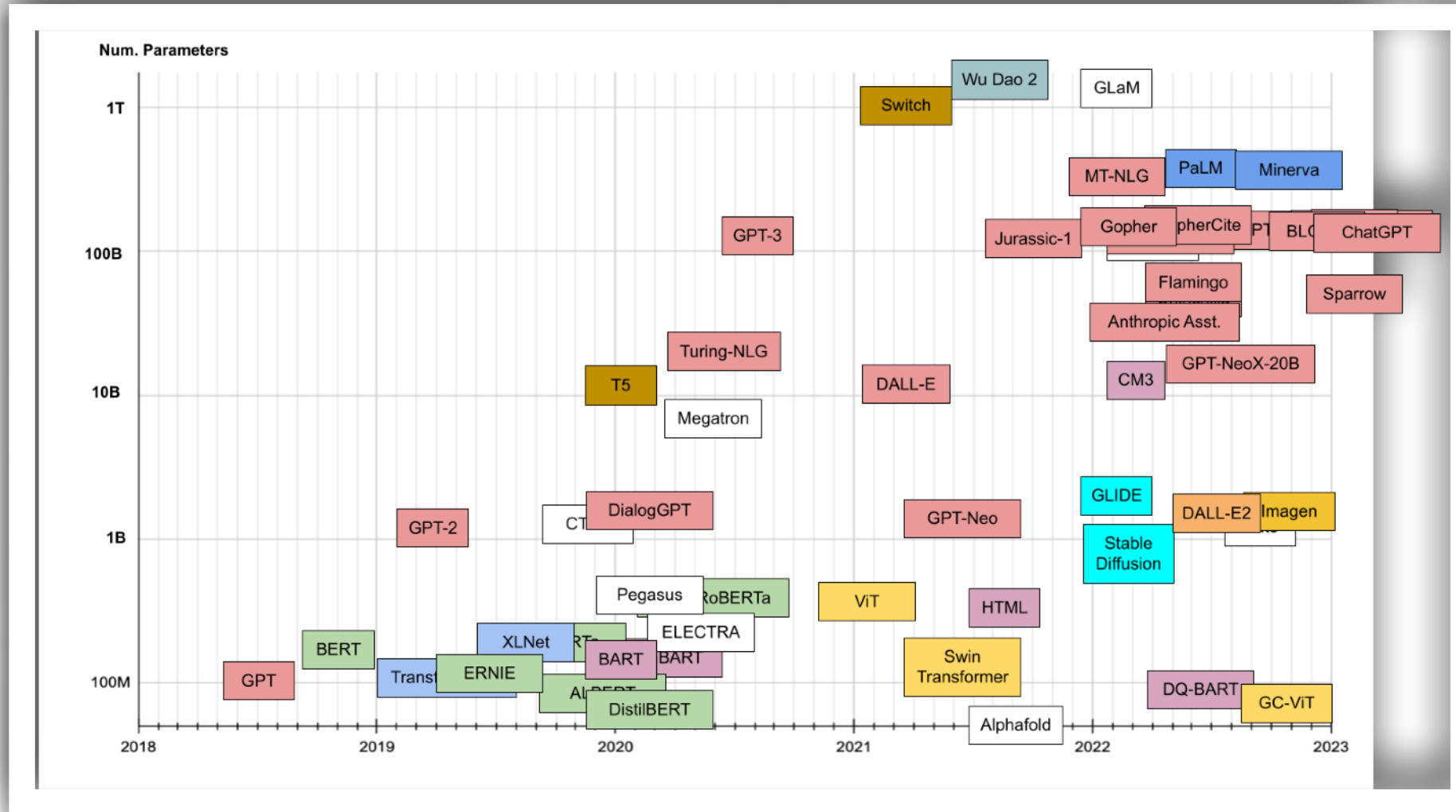
Sebastian Raschka. STAT 453: Introduction to Deep Learning and Generative Models. <https://sebastianraschka.com/blog/2021/dl-course.html>

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint

LARGE LANGUAGE MODELS



Timeline of Transformer Models



Amatriain, X. (2023). Transformer models: an introduction and catalog. *arXiv preprint arXiv:2302.07730*.

<https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e9039f376/>

GPT (Generative Pre-trained Transformer)

User What is funny about this image? Describe it panel by panel.



Source: [hmm \(Reddit\)](#)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

A mathematical conversation with GPT-4



Within the duration of one year, a rabbit population first multiplies itself by a factor a and on the last day of the year b rabbits are taken by humans for adoption. Supposing that on first day of the first year there are x rabbits, we know that exactly 3 years afterwards there will be a population of $27x - 26$ rabbits. What are the values of a and b ?



Let's call the rabbit population after one year y . We know that $y = ax - b$. After two years, the population will be $ay - b = a(ax - b) - b = a^2x - ab - b$. After three years, the population will be $a(a^2x - ab - b) - b = a^3x - a^2b - ab - b$. We are given that this is equal to $27x - 26$, so we have the equation:
[...]

Solving for b , we find that $b = 2$. So $a = 3$ and $b = 2$ are the values we are looking for.

Note that the above question can be alternatively formulated as a question about a function which is composed with itself three times. Can you provide an alternative formulation of the question?

GPT 4 Capabilities <https://openai.com/research/gpt-4>

<https://www.youtube.com/watch?v=qblk7-JPB2c>

PaLM: Scaling Language Modeling with Pathways

Explaining a Joke

Input: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Output: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

Logical Inference

Input: Shelley is from Virginia, but is visiting that city with that famous market where they throw the fish. Going home next Tuesday!

Question: Is it likely that Shelley will be near the Pacific Ocean this weekend?

Model Output: The city with the famous market where they throw the fish is Seattle, Washington. Seattle is on the Pacific Ocean. Shelley is visiting Seattle, so she will be near the Pacific Ocean this weekend. The answer is "yes", it is likely that Shelley will be near the Pacific Ocean this weekend.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

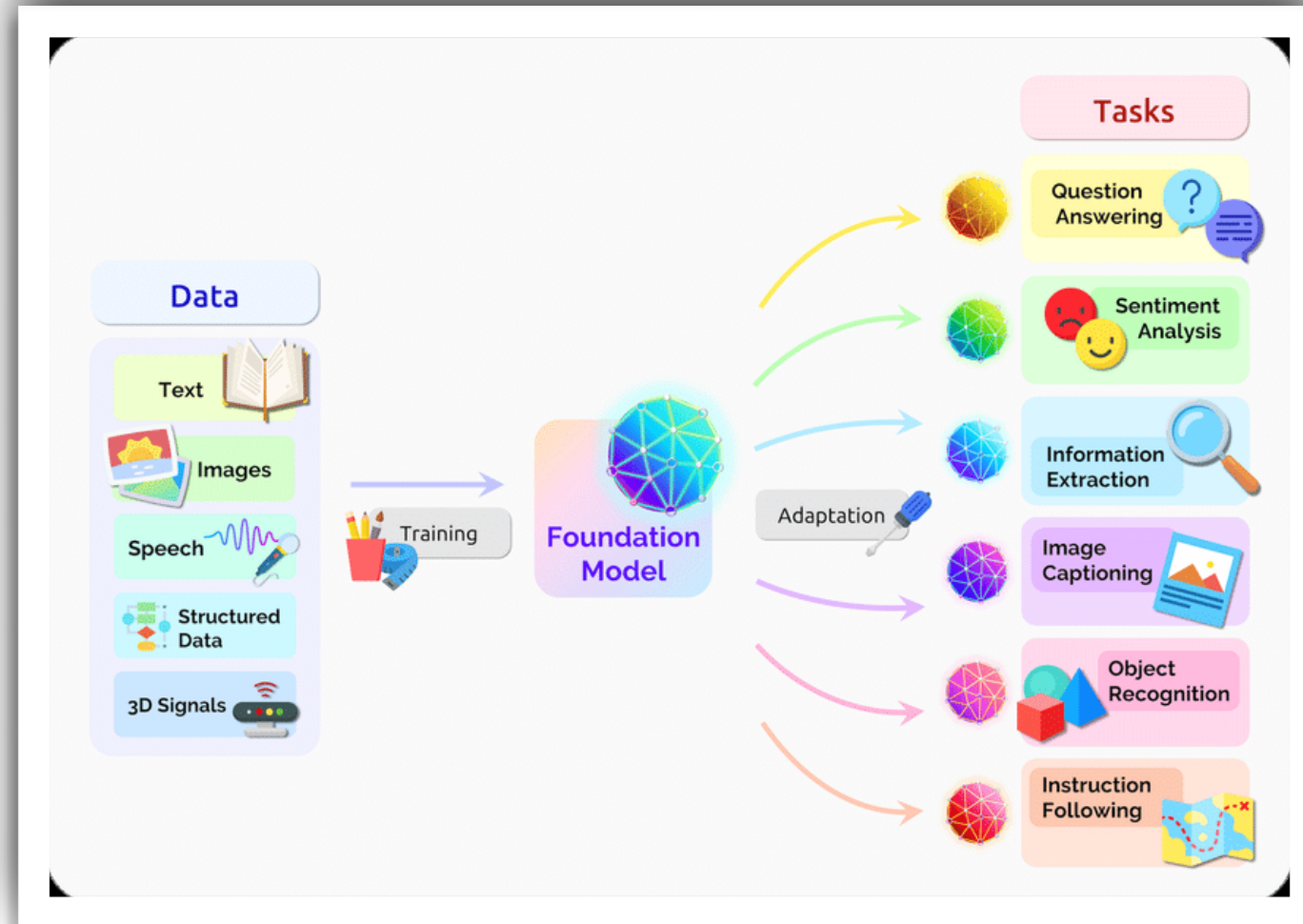
<https://arxiv.org/pdf/2204.02311> Google Research

OPPORTUNITIES



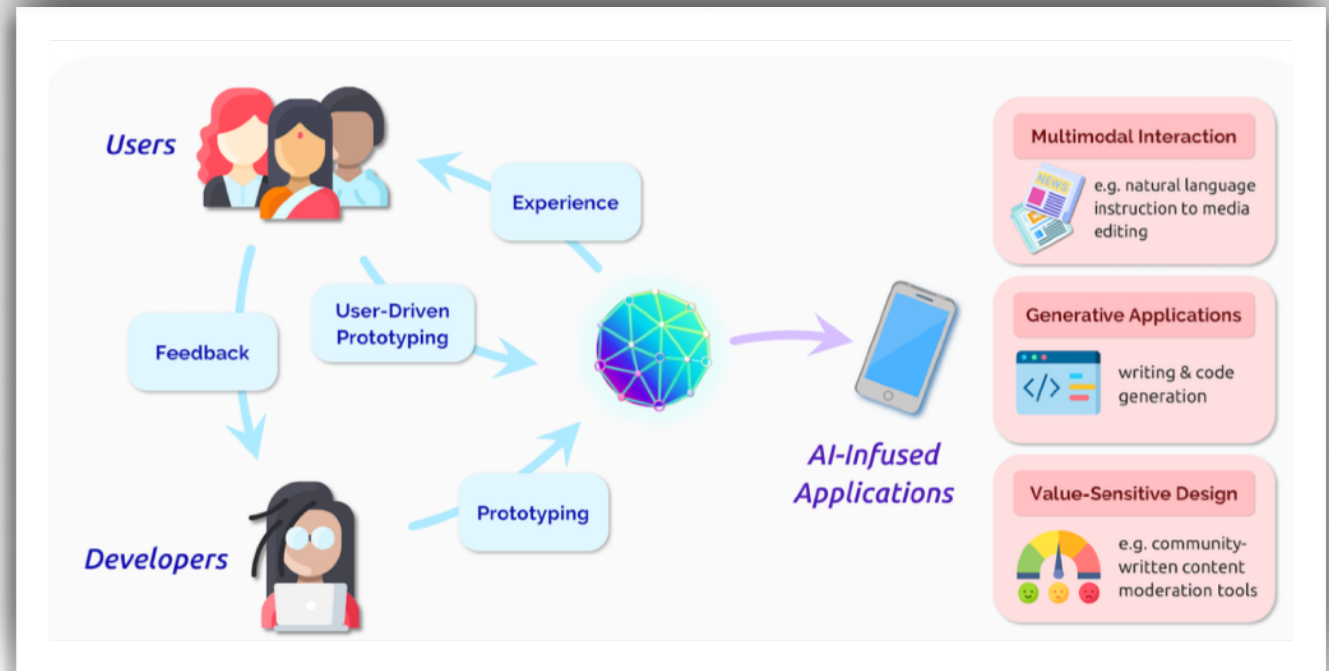
Foundational Models

- These large foundational models can be used to develop more advanced and complex models, e.g., sentiment analysis, Q&A, ..
- Examples:
 - TaxGPT,
 - GPT-4 for Icelandic,
 - GPT-4 for company knowledge bases (e.g., Morgan Stanley)



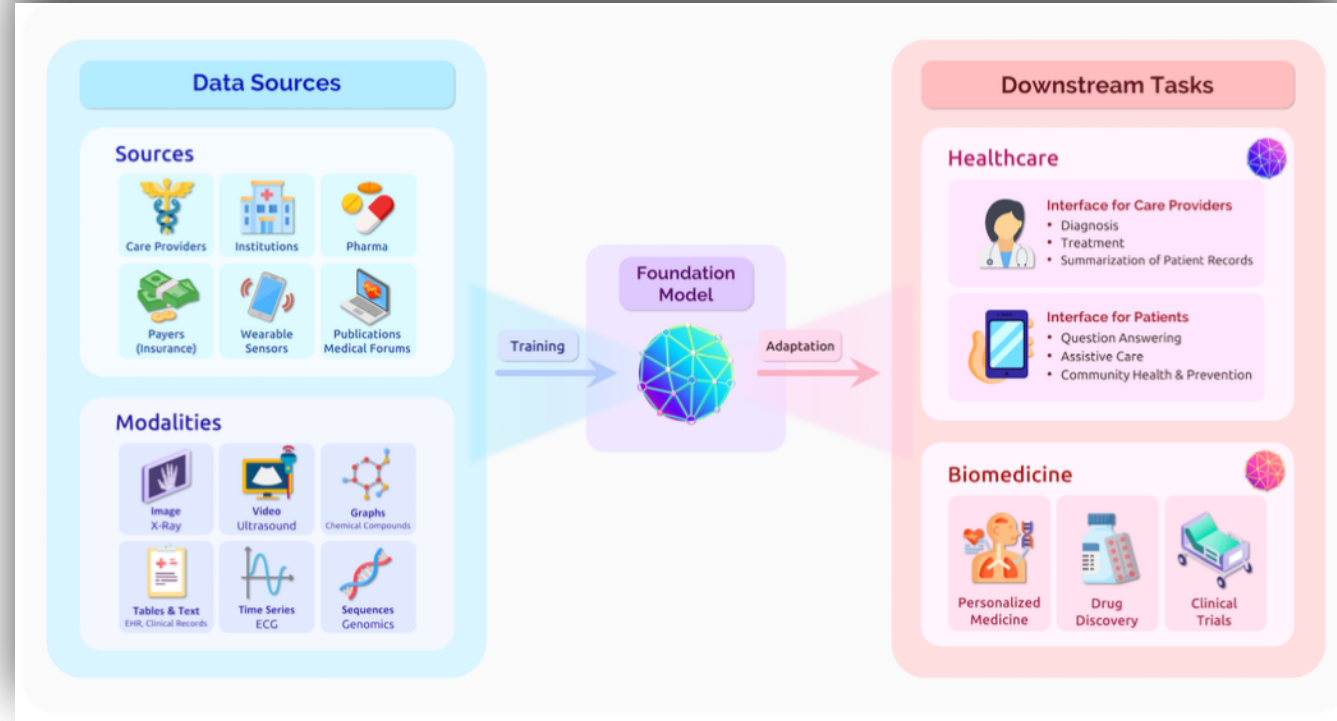
AI-infused Applications

- Foundation models will bring significant opportunities to developers (content & code) by lowering the difficulty threshold for building AI-infused applications.
- In some cases, the line between developers and users will start to blur as end users can develop their own AI applications.



Healthcare and Biomedicine

- Both Healthcare and Biomedical research need comprehensive medical knowledge from multiple sources/modalities.
- They can be trained on diverse sources/modalities of data in medicine and can be queried/used by both medical professionals and the public.
- Their adaptation capabilities (fine-tuning/prompting) can be efficiently used to develop applications interfacing with healthcare providers/patients.

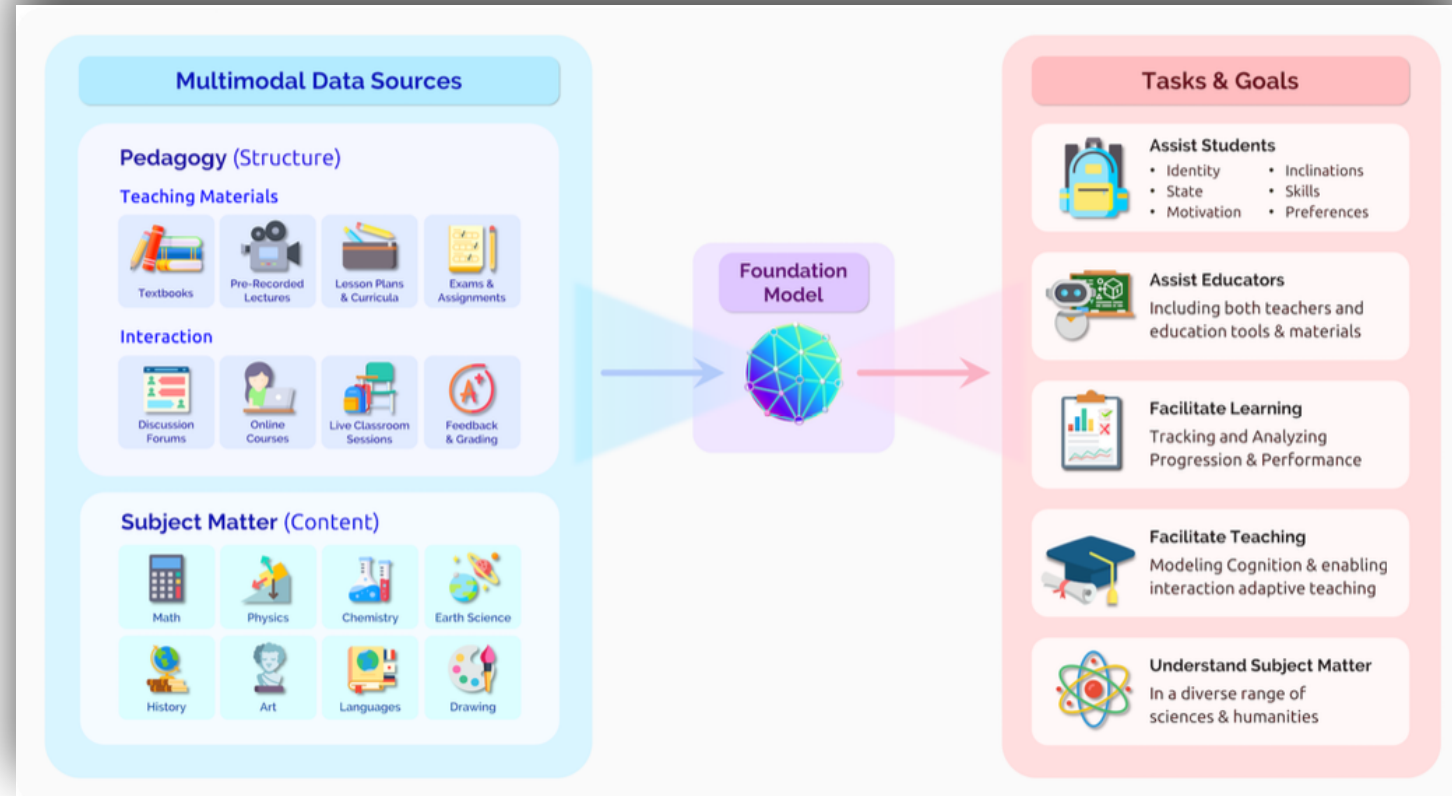


Personalized medicine is an emerging practice that uses an individual's genetic profile in the diagnosis, and treatment of diseases.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Education

- Goals of digital education (based on AI) is to increase the productivity of the learning experience, which can remove human teachers from the loop. What about adverse effects?
- How to effectively teach and grade students who have access to foundation-model-based tools?



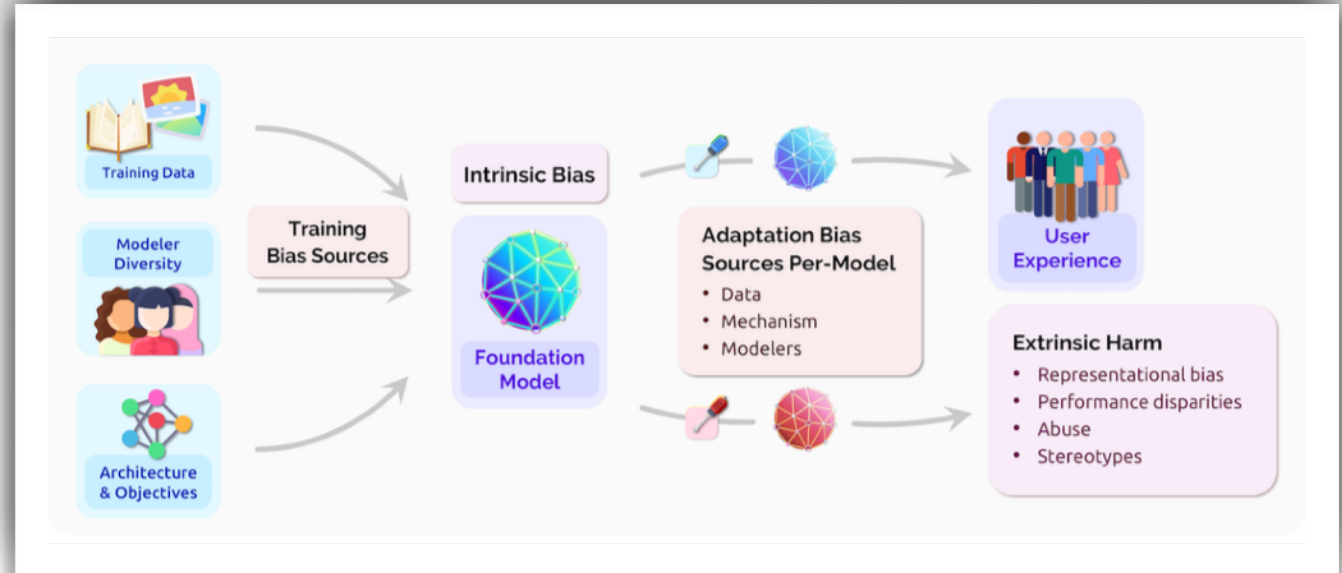
Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

CHALLENGES



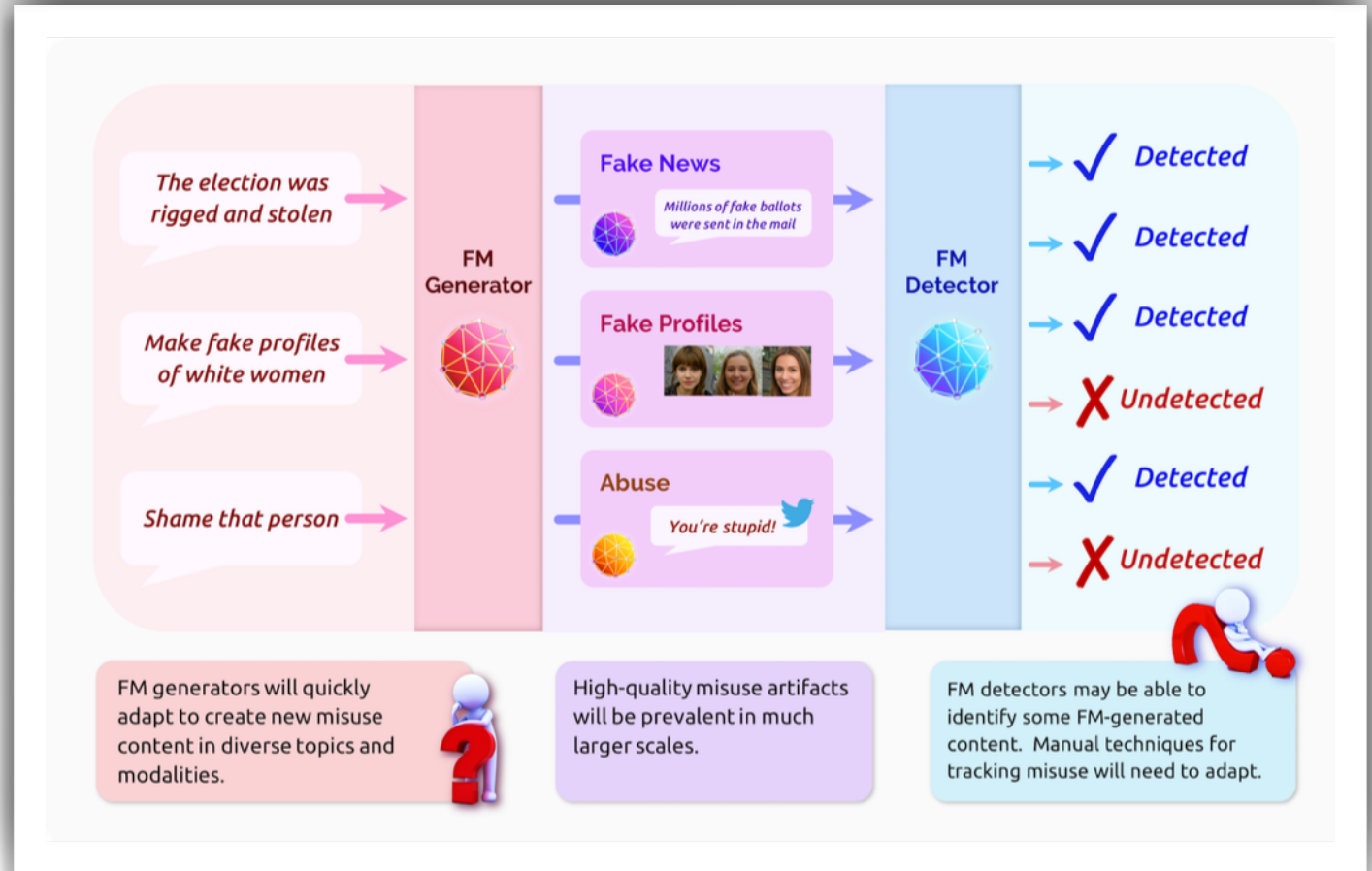
Inequity and Fairness

- Intrinsic bias in these models is the byproduct of various biases in training sources.
- Its biases propagate to these many applications as a result.
- They can yield inequitable outcomes: treating people unjustly due to compound historical discrimination
- It will be crucial and challenging to handle this aspect.



Misuse

- They can be misused to generate high-quality, cheap, and personalized content for harmful and manipulative purposes (hate speech, misinformation, etc.)
- Our research experiments show that it is quite hard for people to identify such harmful and manipulative content by Generative AI

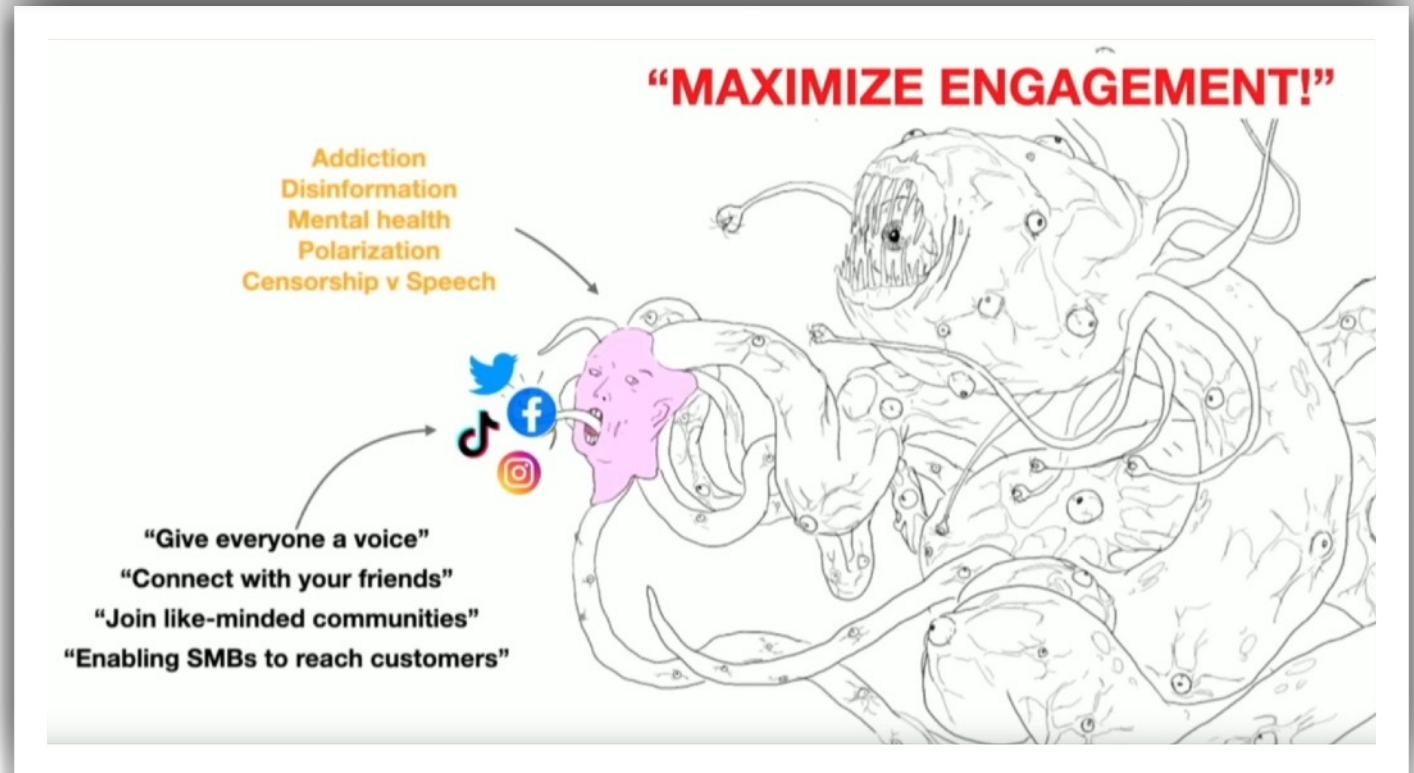


Societal Impacts

The technology that connects us,
also manipulates us

1st contact: SOCIAL MEDIA

- Information Overload 🏠 Addiction
- Doomscrolling 🤖 Influencer Culture
- Sexualization of kids 🧑‍🚫 Qanon
- Shortened attention spans 🧠
- Polarization 🗳️ Bots, DeepFakes
- Cult factories 🗳️ Fake News
- Breakdown of Democracy 🗳️



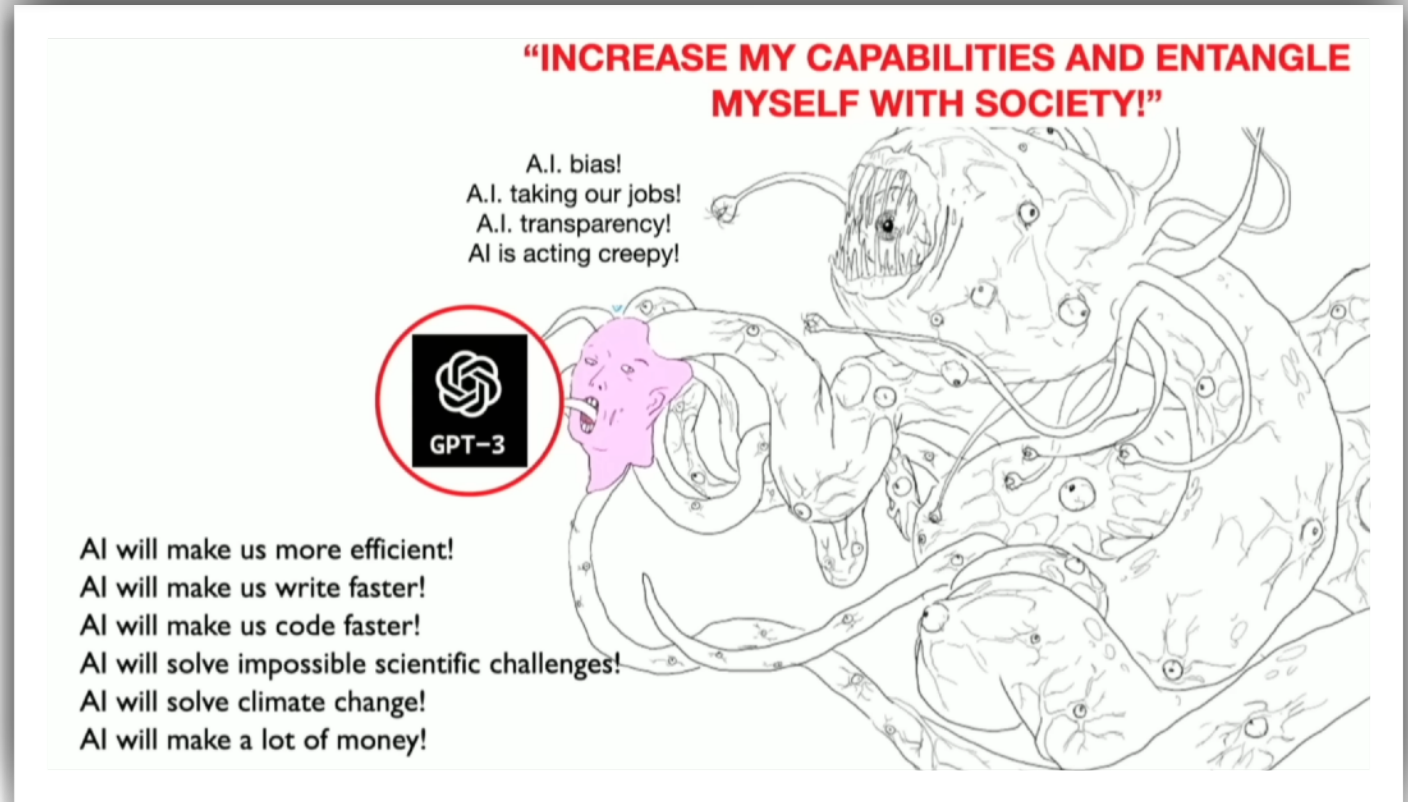
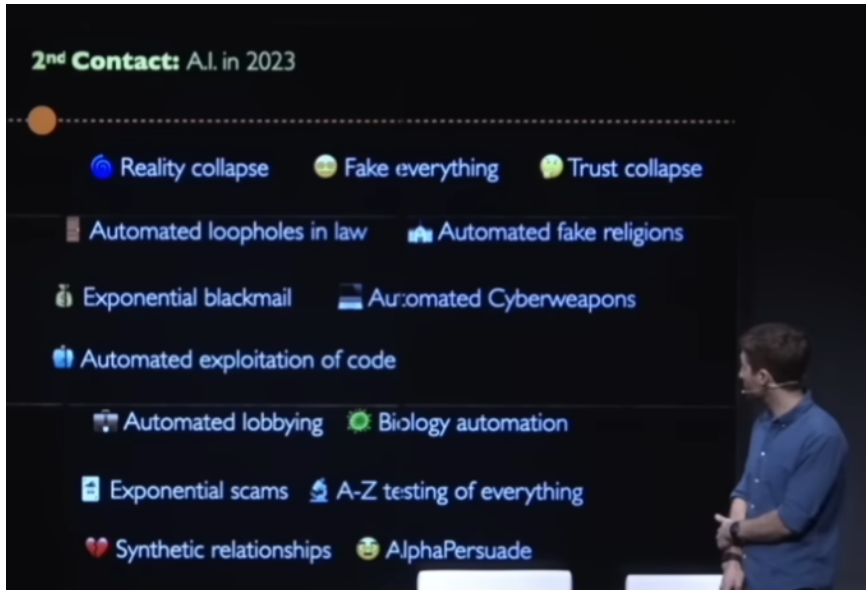
Purpose of Social Media is to give power to people, but it has created several adverse effects in society, which not yet fixed

Center for Humane Technology <https://www.humanetech.com/>

<https://www.youtube.com/watch?v=xoVJKj8lcNQ>

The Social Dilemma <https://www.thesocialdilemma.com/>

A.I. plus Social Media



1. When you invent a new technology, you uncover a new class of responsibilities

2. If the tech confers power, it starts a race

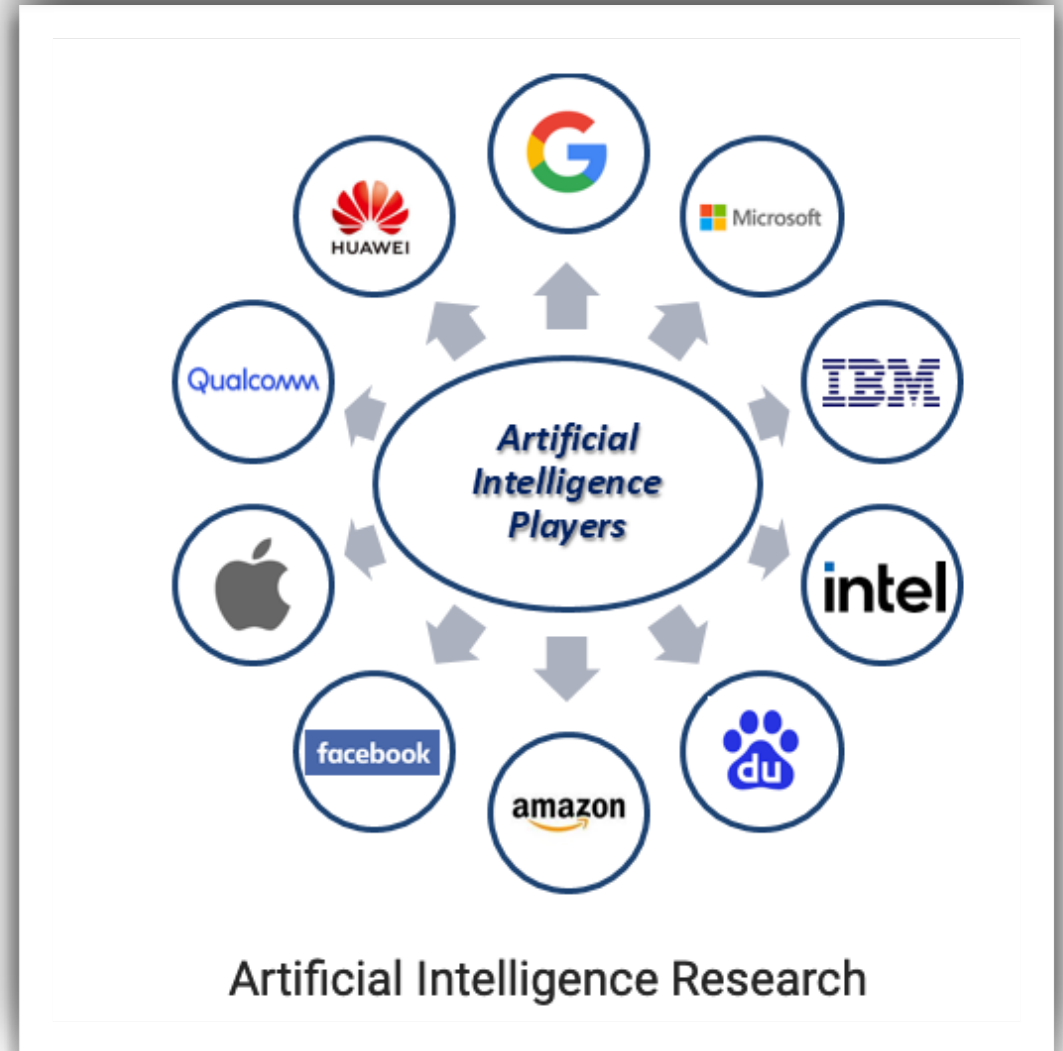
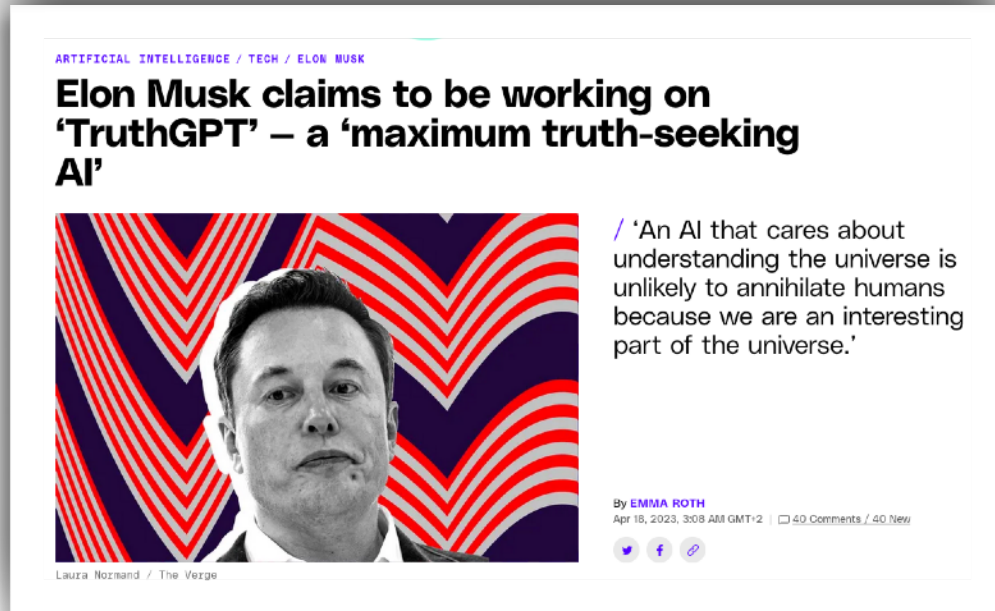
3. If you do not coordinate, the race ends in tragedy

Center for Humane Technology <https://www.humanetech.com/>

<https://www.youtube.com/watch?v=xoVJKj8lcNQ>

The AI war has begun!

Most of the AI research is happening in big companies, which can raise concerns regarding transparency, accountability, and explainability.

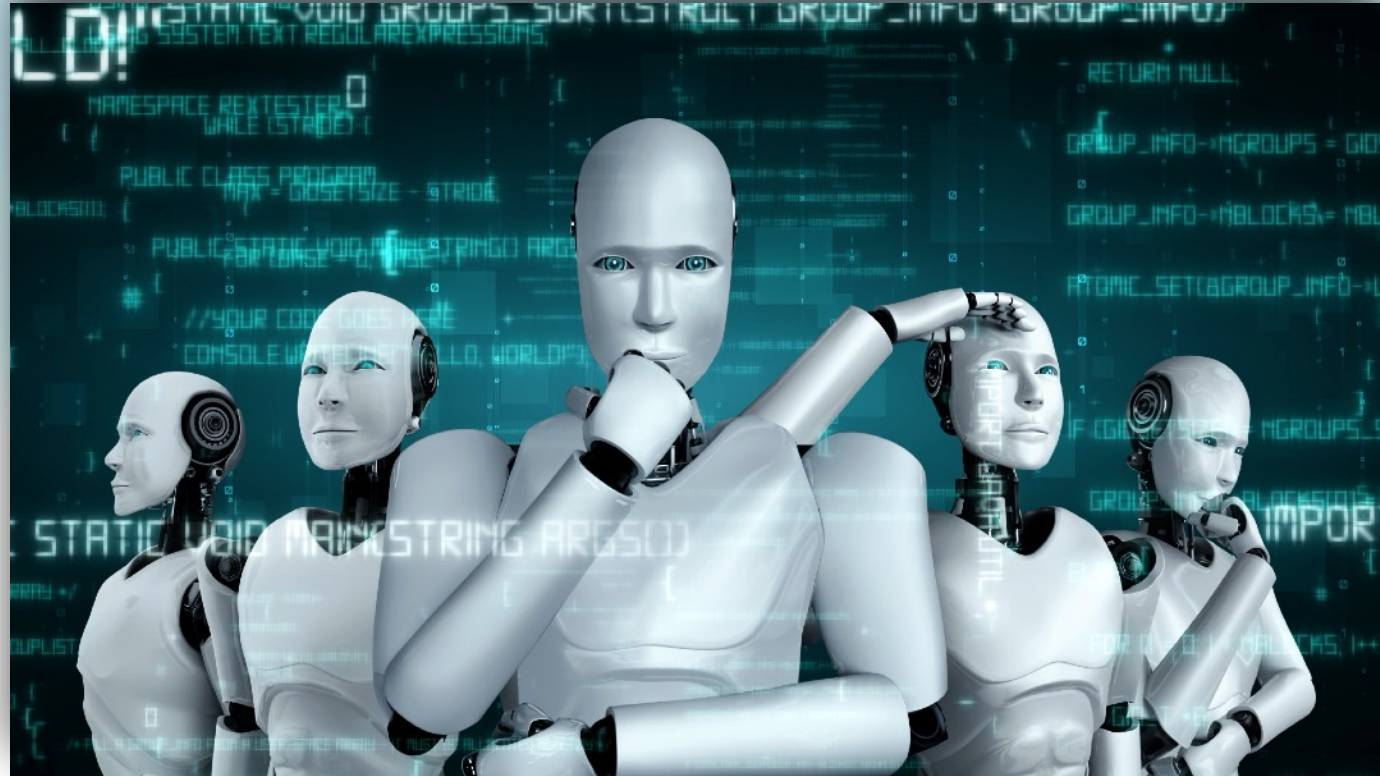


<https://www.greyb.com/blog/artificial-intelligence-companies/>

<https://sgeek.sg/elon-musk-truthgpt-ai-rival-openai-google/>

Conclusion

- We have been living in quite interesting times! 😊
- We see exponential growth in AI intelligence, which will create a lot of transformation both in our lives and in society.
- Same time, it will also create a lot of concerns, risks, and harm in society.
- So, we need a dialogue to harness this AI revolution responsibly!



Good References

- <https://www.youtube.com/watch?v=FkckgwMHP2s>
- https://ethen8181.github.io/machine-learning/clustering_old/topic_model/LDA.htm
- <https://topix.io/tutorial/tutorial.html>
- <https://www.youtube.com/watch?v=fCmlceNqVog>
- <https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045>
- <https://www.coursera.org/lecture/language-processing/topic-modeling-a-way-to-navigate-through-text-collections-fzTUI>
- <https://www.coursera.org/lecture/language-processing/how-to-train-plsa-OzmrF>
- <https://www.coursera.org/lecture/language-processing/the-zoo-of-topic-models-vraEh>

Thank you!

Raghava Mukkamala

rrm.digi@cbs.dk

<https://www.cbs.dk/staff/rrmdigi>

<https://raghavamukkamala.github.io/>

<https://cbsbda.github.io/>