Outsight AI

mfsheng@gmail.com

Maybe someone is wondering why not use “insight” instead of “outsight” here? It is simple BECAUSE I am not an expert in AI era. And trying to understand this hot topic now, so “outsight” is more appropriate than “insight” here. However I am also trying to use the insight ways as I did 10 years ago…

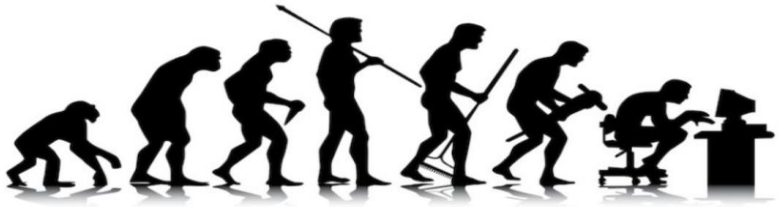
—Me

AI : The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

—Oxford Dictionary

Human's dream of Artificial Intelligence, never stop before human distinction …

Content



ORIGIN

- Human force animal and other human to work for them
- Machine replace Human and animal
- Can machine be more intelligence but still controlled by human?

EVOLUTION

- Knowledge/rules based AI
- Machine learning
- Representation learning
- Deep learning
- Large Language Mode
- From AI to AGI

TREND

- AI impact on industries
- AI impact on work force
- AI market trend
- AI players and ecosystem
- AI technology trend

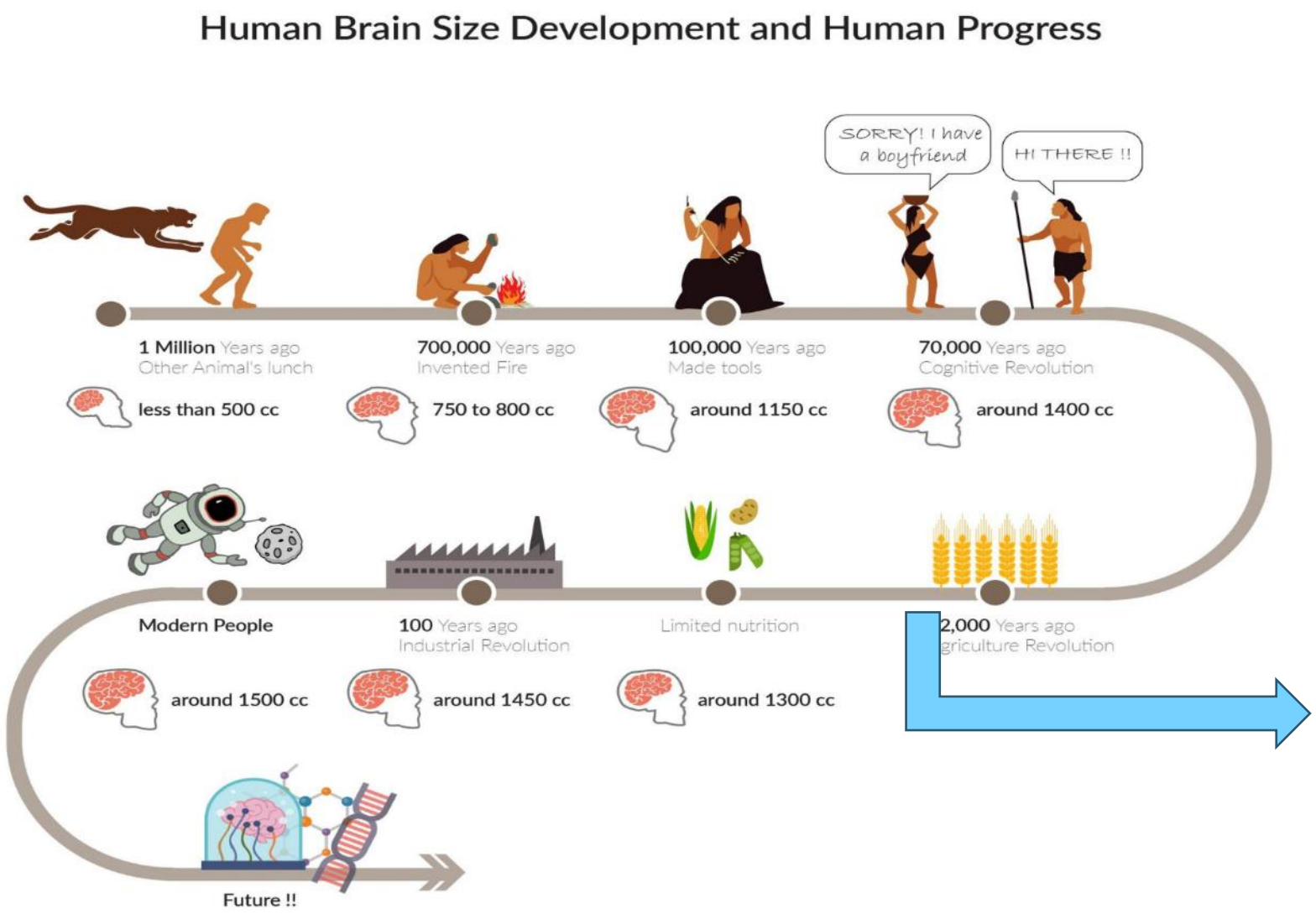
Origin

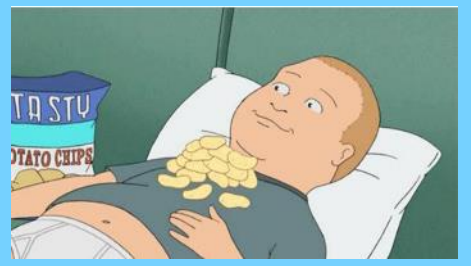
We shall not cease from exploration, and the end of all our exploring will be to arrive where we started and know the place for the first time.

ChatGPT

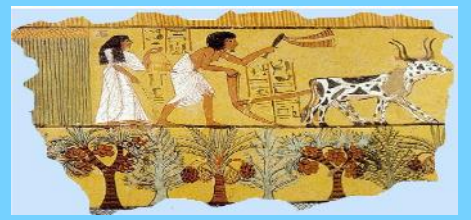
Origin

Human Brain Evolution for 1 Million Years…

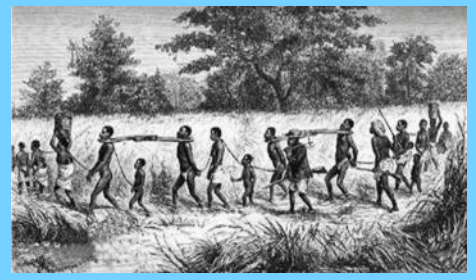




Human became more and more lazy with the development of his brain



They make use of animal to work for them …



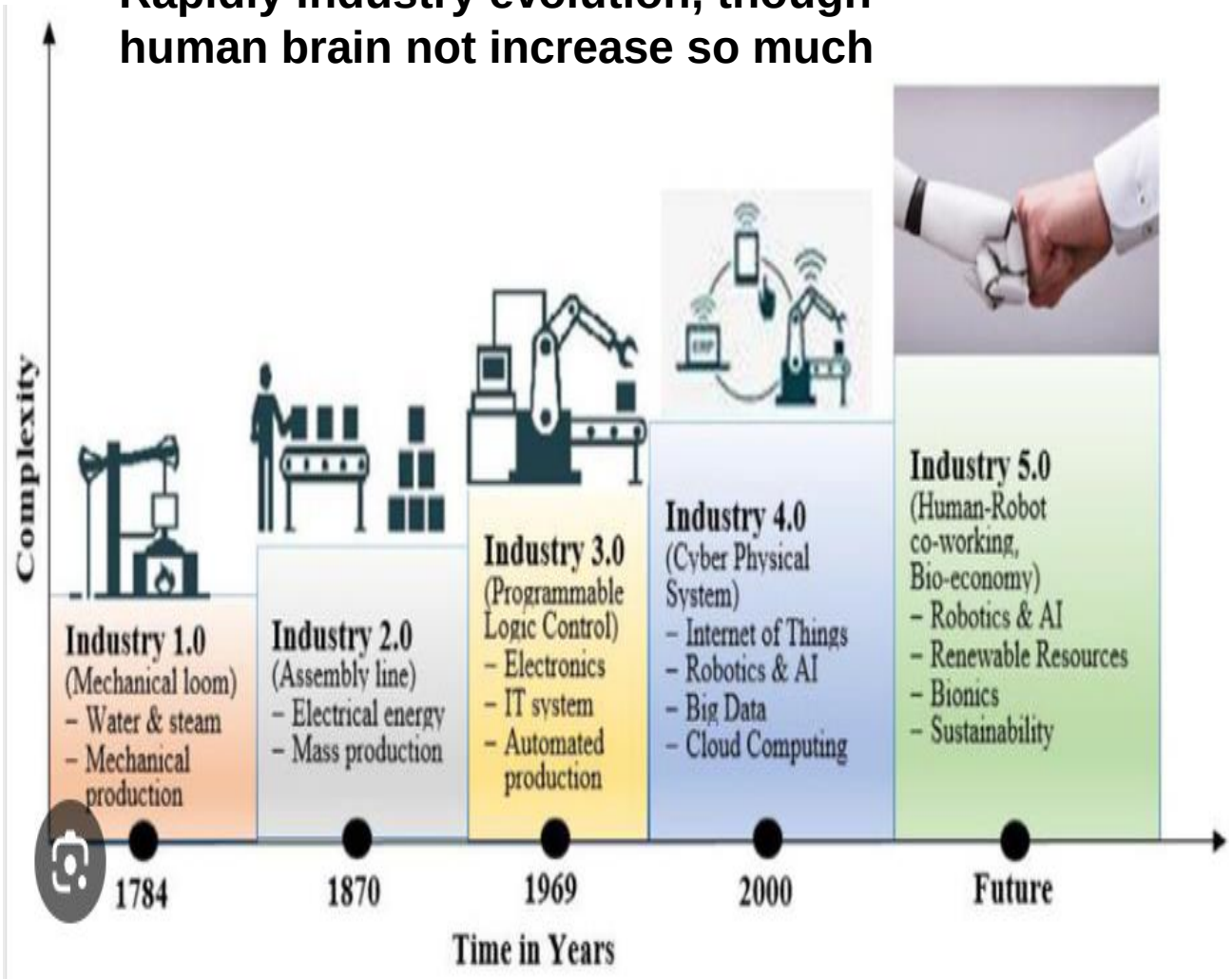
Even more they forced other human as the servants …

And the servants rebelled and bloody slaughter …

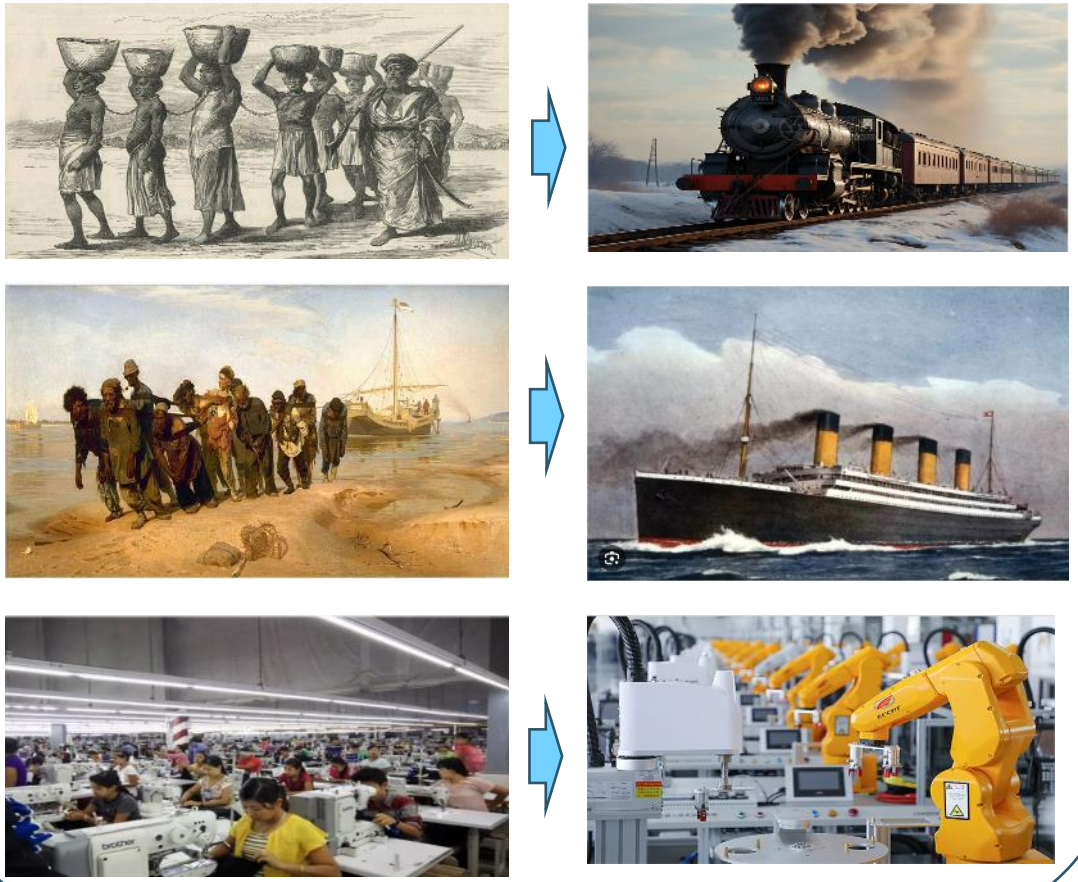
Origin

For last 200+ Years, Human benefit a lot from Machine…

Rapidly Industry evolution, though human brain not increase so much



Machine proved more powerful and efficiency than Human



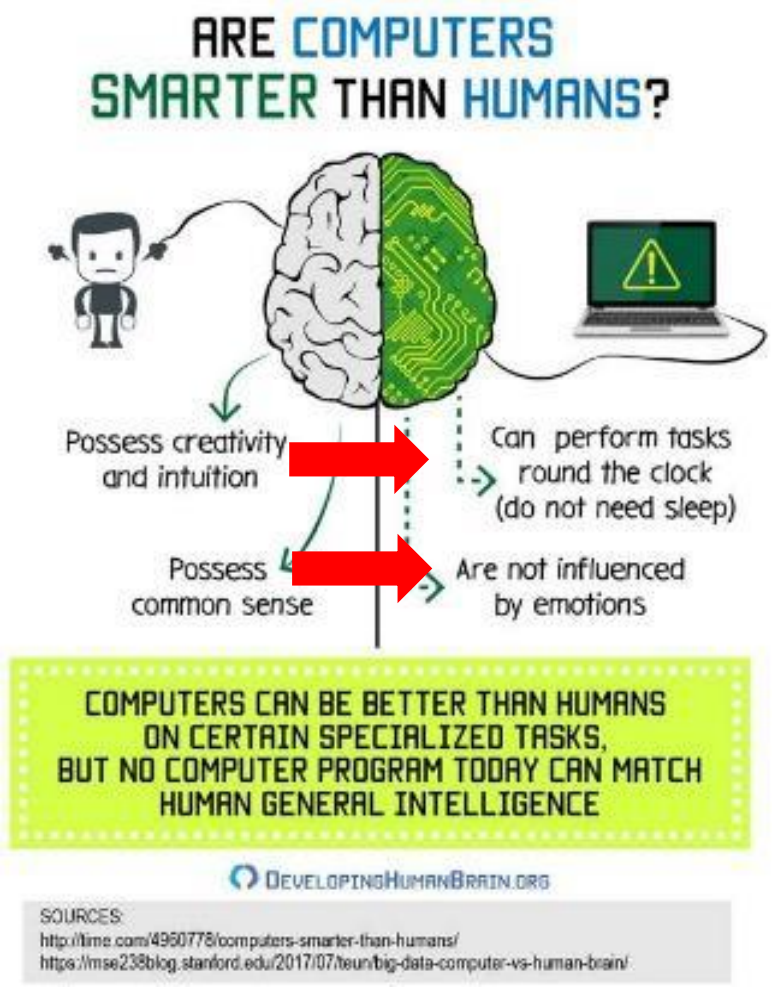
Origin

Human want Machine be more intelligent, but meet challenges …

Autonomous driving: complete a full, safe, and efficient driving without human intervention.

- **Perception:** Perceive the vehicle's own and surrounding environment in real time, such as position, speed, attitude, obstacles, traffic signals, road conditions, etc.
- **Decision:** Plan and decide based on the perceived information, determine the appropriate working mode, formulate the appropriate control strategy, and replace human beings to make driving decisions, such as choosing the optimal route, adjusting the speed, changing lanes, overtaking, parking, etc.
- **Control:** drive-by-wire throttle, drive-by-wire steering, drive-by-wire braking, etc.
- **Interaction:** Communicate information with human passengers, receiving passengers' instructions or feedback, and enhancing passengers' sense of security and comfort

Artificial Intelligence (AI) : Make machine more intelligent



Challenge of artificial intelligence

To solve the tasks that are easy for people to perform but hard for people to describe formally—problems that we solve intuitively(**Creativity & Intuition**), that feel automatic, like recognizing spoken words or faces in images(**Common sense**)

Evolution

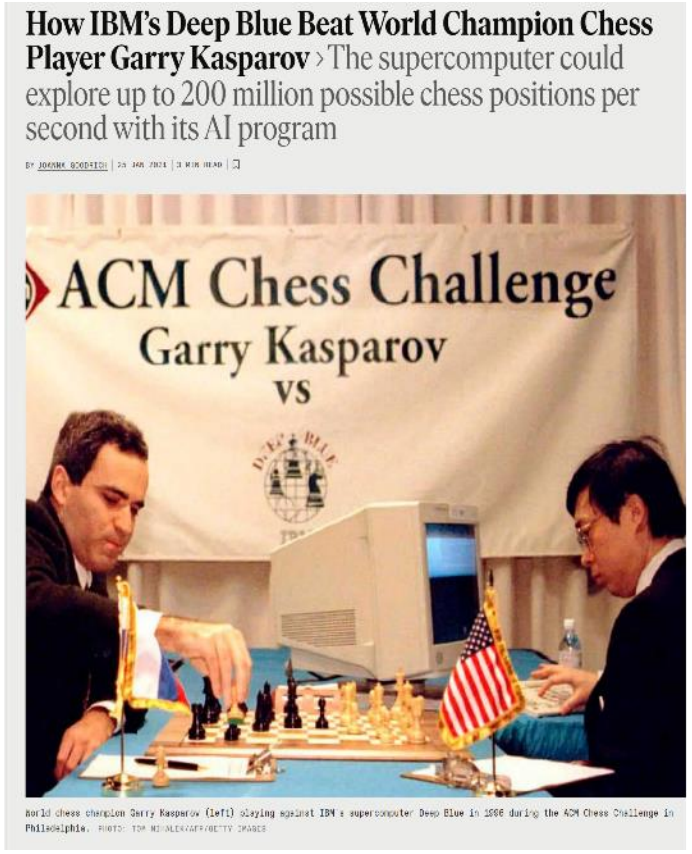
Artificial intelligence will reach human levels by around 2029. Follow that out further to, say, 2045, we will have multiplied the intelligence, the human biological machine intelligence of our civilization a billion-fold.

Ray Kurzweil, Futurist and Director of Engineering at Google

Evolution

3 amazing millstones of AI in the past 30 years…

May 1997



How IBM's Deep Blue Beat World Champion Chess Player Garry Kasparov › The supercomputer could explore up to 200 million possible chess positions per second with its AI program

World chess champion Garry Kasparov (left) playing against IBM's supercomputer Deep Blue in 1996 during the ACM Chess Challenge in Philadelphia.

Knowledge and rules base

March 2016

Human Go champion loses to Google DeepMind AlphaGo computer in 1st game

South Korea's Lee Sedol defeated in 1st of historic, 5-game human vs. AI tournament

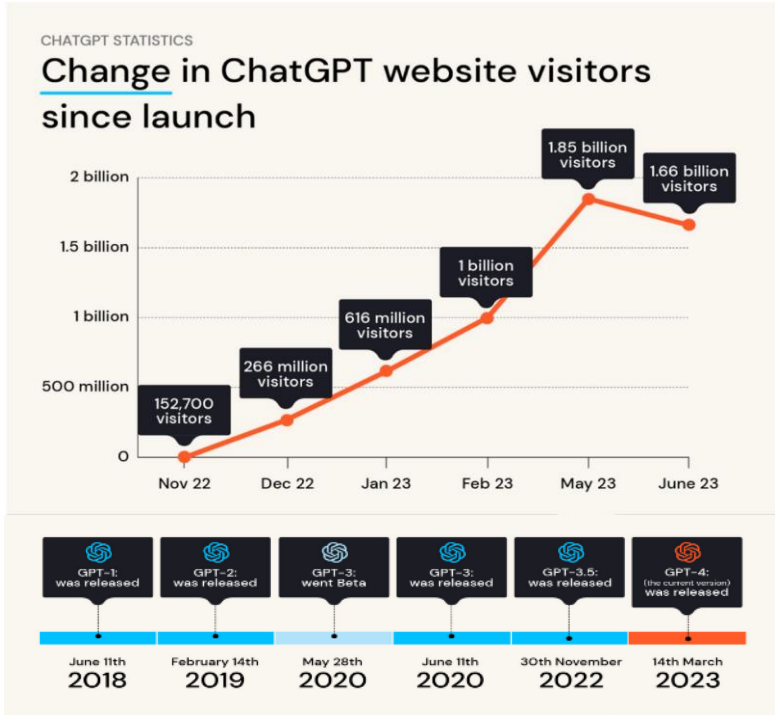
The Associated Press · Posted: Mar 09, 2016 8:34 AM EST | Last Updated: March 9, 2016

South Korean professional Go player Lee Sedol, right, prepares for his second stone against Google's artificial intelligence program, AlphaGo, as Google DeepMind's lead programmer Aja Huang, left, sits during the Google DeepMind Challenge Match in Seoul, South Korea, Wednesday, March 9, 2016. Google's computer program AlphaGo defeated Sedol in the first game of a historic five-game match between human and computer. (Lee Jin-Man/Associated Press)

Deep learning

November 2022

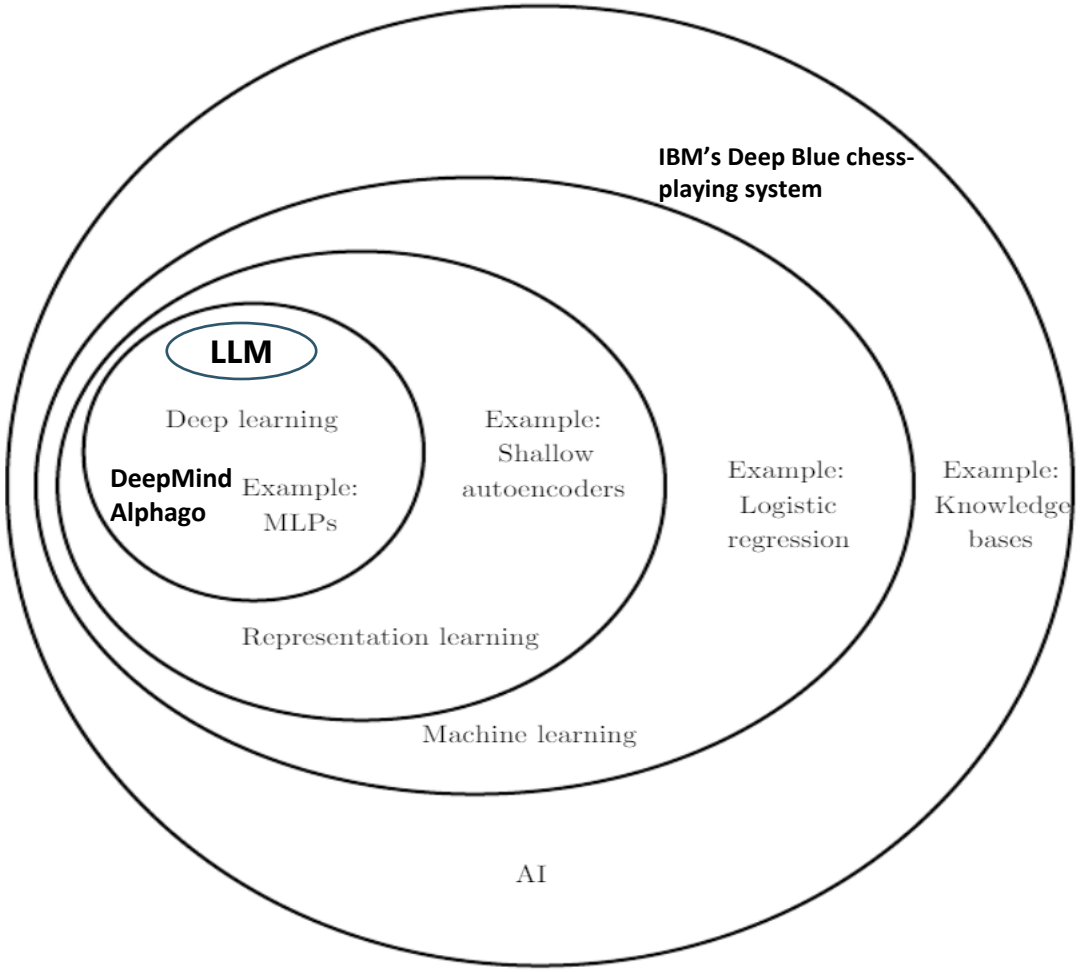
ChatGPT (Chat Generative Pretrained Transformer) is a chatbot that produces human-like AI-generated content based on the input it is given by a user. It was developed by Open AI



LLM(Large Language Mode)

Evolution

Human try many ways to improve AI…



When programmable computers were first conceived, people wondered whether such machines might become intelligent, over a hundred years before one was built (Lovelace, 1842)

Knowledge bases : hard-code knowledge about the world in formal languages. A computer can reason automatically about statements in these formal languages using logical inference rules.

Machine learning : Acquire their own knowledge, by extracting patterns from raw data, these simple machine learning algorithms depends heavily on the representation of the data they are given.

Representation learning : Use machine learning to discover not only the mapping from representation to output but also the representation itself.

Deep learning: A particular kind of machine learning that achieves great power and flexibility by representing the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones

Evolution

Machine learning is the only viable approach to building AI systems that can operate in complicated real world

Changes and Possibilities

Chess VS Go

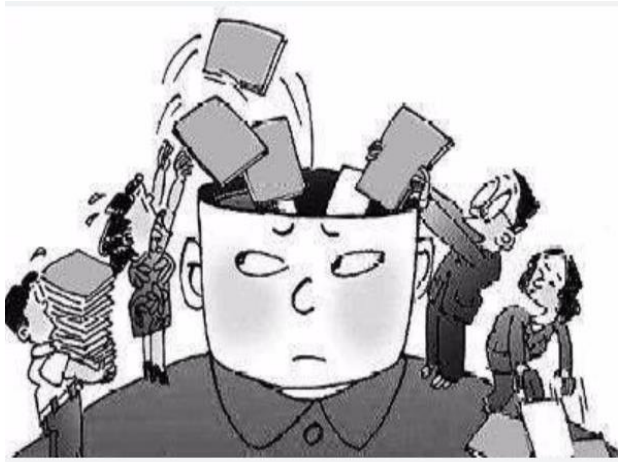
$10^{50} < 10^{160}$

- **Go far exceed those of chess & human cognition and computation ability..**
- **Brute force computation make no sense.**
- **And real world is far more complicate then game…**

Knowledge bases can solve the problems that can be described by a list of formal, mathematical rules.

None of these projects has led to a major success in real world.

Inculcate knowledge and rules



Improve ability to learn by self

AI machine similar with human being…

Evolution

2 simplest and popular samples of machine learning, base on statistic and probability

logistic regression can determine whether to recommend cesarean delivery (Mor-Yosef et al., 1990)

Consider the model function

$$y = \alpha + \beta x,$$

which describes a line with slope β and y-intercept α. In general such a relationship may not hold exactly for the largely unobserved population of values of the independent and dependent variables; we call the unobserved deviations from the above equation the errors. Suppose we observe n data pairs and call them $\{(x_i, y_i), i = 1, \ldots, n\}$. We can describe the underlying relationship between y_i and x_i involving this error term ε_i by

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

This relationship between the true (but unobserved) underlying parameters α and β and the data points is called a linear regression model.

The goal is to find estimated values $\widehat{\alpha}$ and $\widehat{\beta}$ for the parameters α and β which would provide the "best" fit in some sense for the data points. As mentioned in the introduction, in this article the "best" fit will be understood as in the least-squares approach: a line that minimizes the sum of squared residuals (see also Errors and residuals) $\hat{\varepsilon}_i$ (differences between actual and predicted values of the dependent variable y), each of which is given by, for any candidate parameter values α and β,

$$\hat{\varepsilon}_i = y_i - \alpha - \beta x_i.$$

In other words, $\widehat{\alpha}$ and $\widehat{\beta}$ solve the following minimization problem:

$$(\hat{\alpha}, \hat{\beta}) = \operatorname{argmin}(Q(\alpha, \beta)),$$

where the objective function Q is:

$$Q(\alpha, \beta) = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 .$$

By expanding to get a quadratic expression in α and β, we can derive minimizing values of the function arguments, denoted $\widehat{\alpha}$ and $\widehat{\beta}$:[6]

$$\widehat{\alpha} = \bar{y} - (\widehat{\beta}\,\bar{x}),$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}\Delta x_i \Delta y_i}{\sum_{i=1}^{n}\Delta x_i^2}$$

Here we have introduced

- $\bar{x}$ and $\bar{y}$ as the average of the x_i and y_i, respectively
- Δx_i and Δy_i as the deviations in x_i and y_i with respect to their respective means.

Feed the train data set(x, y) to get the α, Ɓ, then use this to predict y by x.

Naive Bayes can separate legitimate e-mail from spam e-mail

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n, :

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using the naive conditional independence assumption that

$$P(x_i | y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i | y),$$

for all i, this relationship is simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i \mid y),$$

and we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set.

Feed the train data set(x, y) to predict the probability of y by a give x .

Evolution

Machine Learning go deeper and deeper…

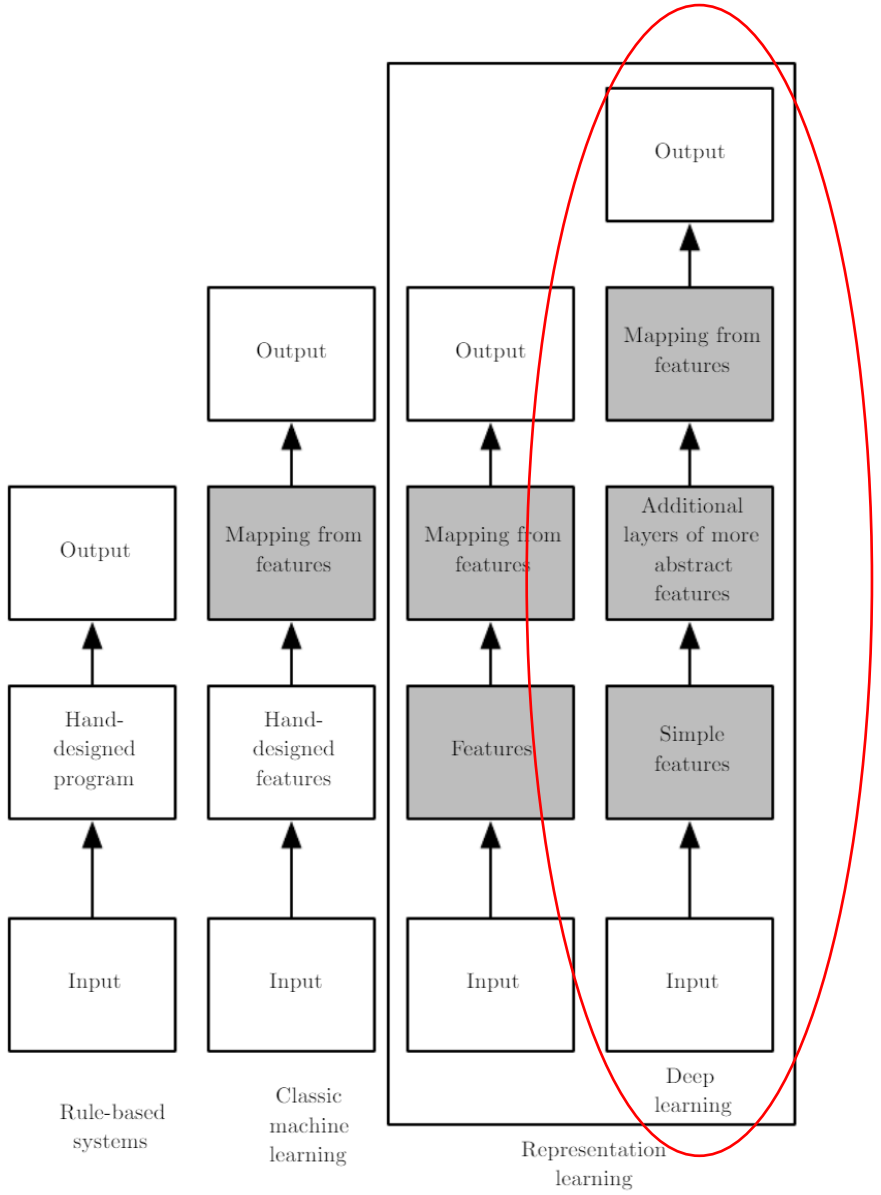
Solution----Deep learning:
Introduce representations that are expressed in terms of other, simpler representations. Enables the computer to build complex concepts out of simpler concepts.

Challenges : When designing features or algorithms for learning features, our goal is usually to separate the **factors of variation** that explain the observed data. A major source of difficulty in many real-world AI applications is that many of **the factors of variation** influence every single piece of data we are able to observe.

Solution----Representation learning:
Use machine to learning **features**, discover not only the mapping from representation to output but also the representation itself. Enable AI systems to rapidly adapt to new tasks, with minimal human intervention.

Challenges : what **features** should be extracted ?

Classic machine learning :
Design the right set of **features** to extract for that task, then providing these features to a simple machine learning algorithm



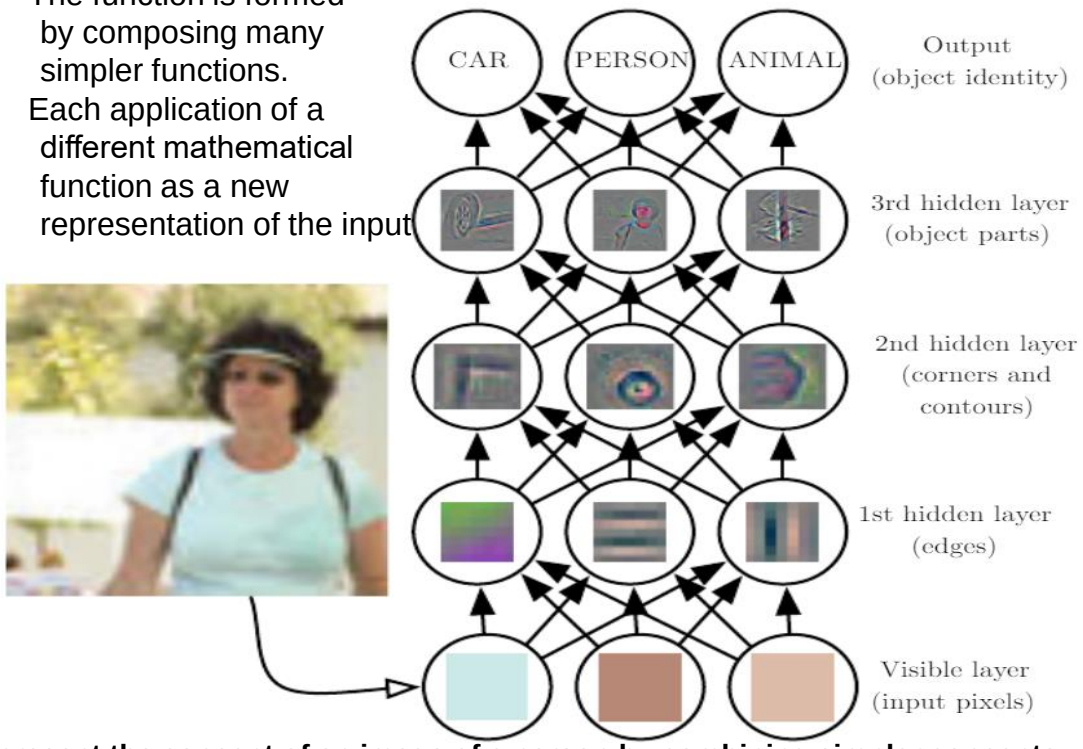
Shaded boxes indicate components that are able to learn from data

Evolution

Deep learning, which introduce representations that are expressed in terms of other simpler representations, has had a long and rich history…

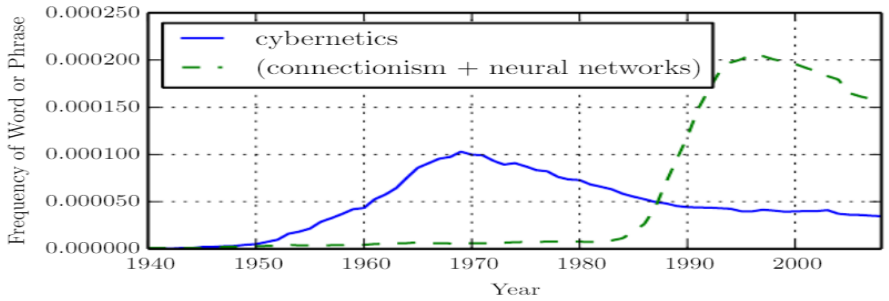
Multilayer perceptron(MLP):

- A mathematical function mapping some set of input values to output values.
- The function is formed by composing many simpler functions.
- Each application of a different mathematical function as a new representation of the input



Represent the concept of an image of a person by combining simpler concepts, such as corners and contours, which are in turn defined in terms of edges

3 waves since 1940



- **Wave1(1940s–1960s) cybernetics**
  - ✓ Theories of biological learning (McCulloch and Pitts, 1943; Hebb, 1949)
  - ✓ neuroscientific perspective
- **Wave2(1980–1995) connectionism**
  - ✓ **Back-propagation** (Rumelhart et al., 1986a) to train a neural network with one or two hidden layers.
  - ✓ A large number of simple computational units can achieve intelligent behavior when networked together
  - ✓ Too computationally costly to allow much experimentation with the hardware available at the time
- **Wave3(2006- now) deep learning**
  - ✓ **Geoffrey Hinton** showed that a kind of neural network called a deep belief network could be efficiently trained using a strategy called greedy layer-wise pretraining (Hinton et al., 2006),

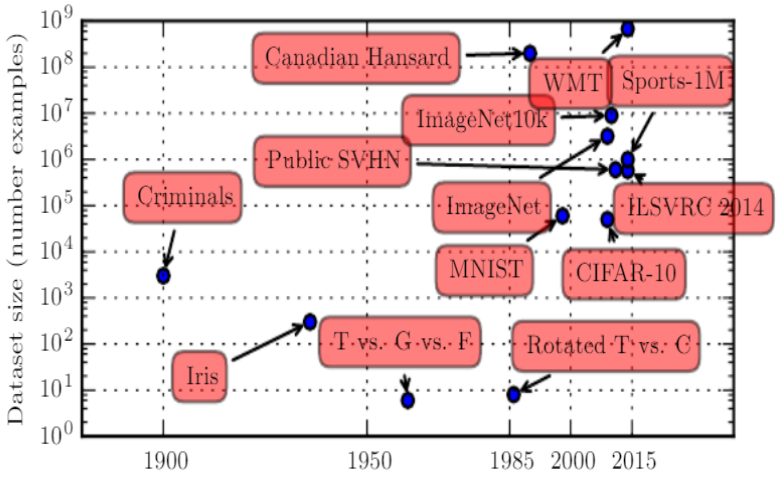
UNIVERSITY OF TORONTO

Evolution

Why deep learning has only recently become recognized as a crucial technology even though the first experiments with artificial neural networks were conducted in the 1940s ?

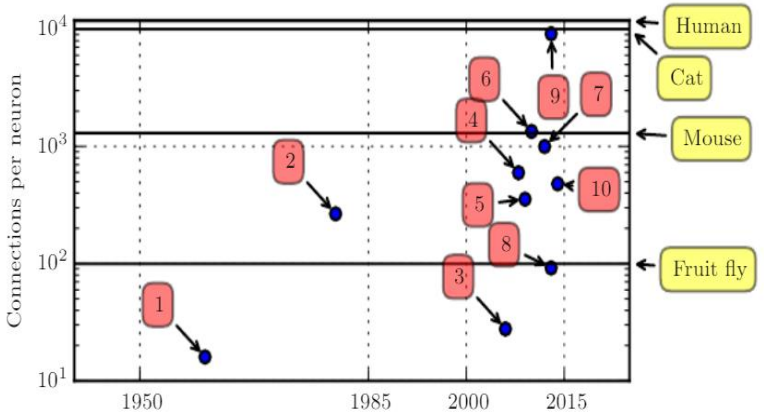
Increasing Dataset Sizes …

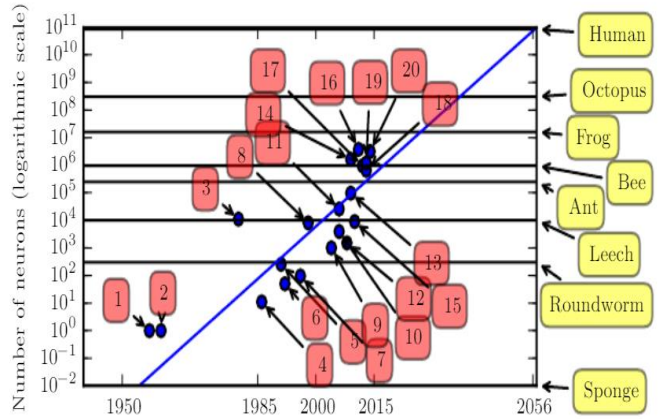
- As more and more of our activities take place on computers, more and more of what we do is recorded.
- As our computers are increasingly networked together, it becomes easier to centralize these records and curate them into a dataset appropriate for machine learning applications.
- The age of “Big Data”



Increasing Model Sizes …

- we have the computational resources tc run much larger models today.
- The increase in computational resource , due to the availability of faster CPUs, the advent of general purpose GPUs , faster network, connectivity and better software infrastructure for distributed computing.





Artificial neural networks have doubled in size roughly every 2.4 years.

Evolution

Deep learning has been proved success in many era, automate routine labor, understand speech or images, make diagnoses in medicine and support basic scientific research

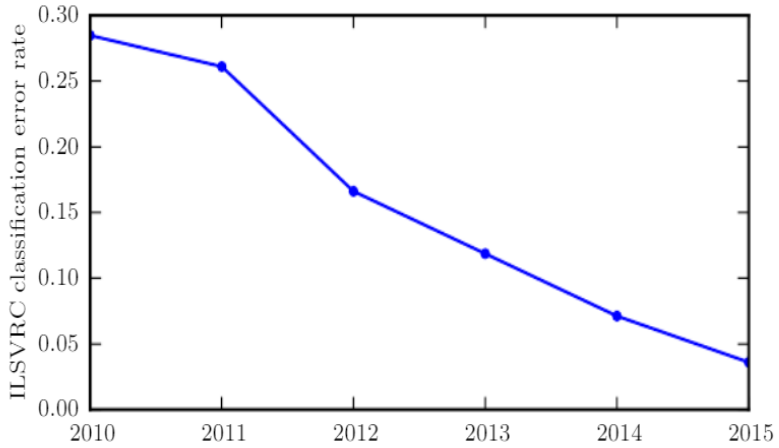


Figure 1.12: Decreasing error rate over time. Since deep networks reached the scale necessary to compete in the ImageNet Large Scale Visual Recognition Challenge, they have consistently won the competition every year, yielding lower and lower error rates each time. Data from Russakovsky *et al.* (2014b) and He *et al.* (2015).

- Deep learning has had a long and rich history, but has gone by many names, reflecting different philosophical viewpoints, and has waxed and waned in popularity.
- Deep learning has become more useful as the amount of available training data has increased.
- Deep learning models have grown in size over time as computer infrastructure (both hardware and software) for deep learning has improved.
- Deep learning has solved increasingly complicated applications with increasing accuracy over time

Evolution

Machine Learning is quite mature now, tons of algorithms available…

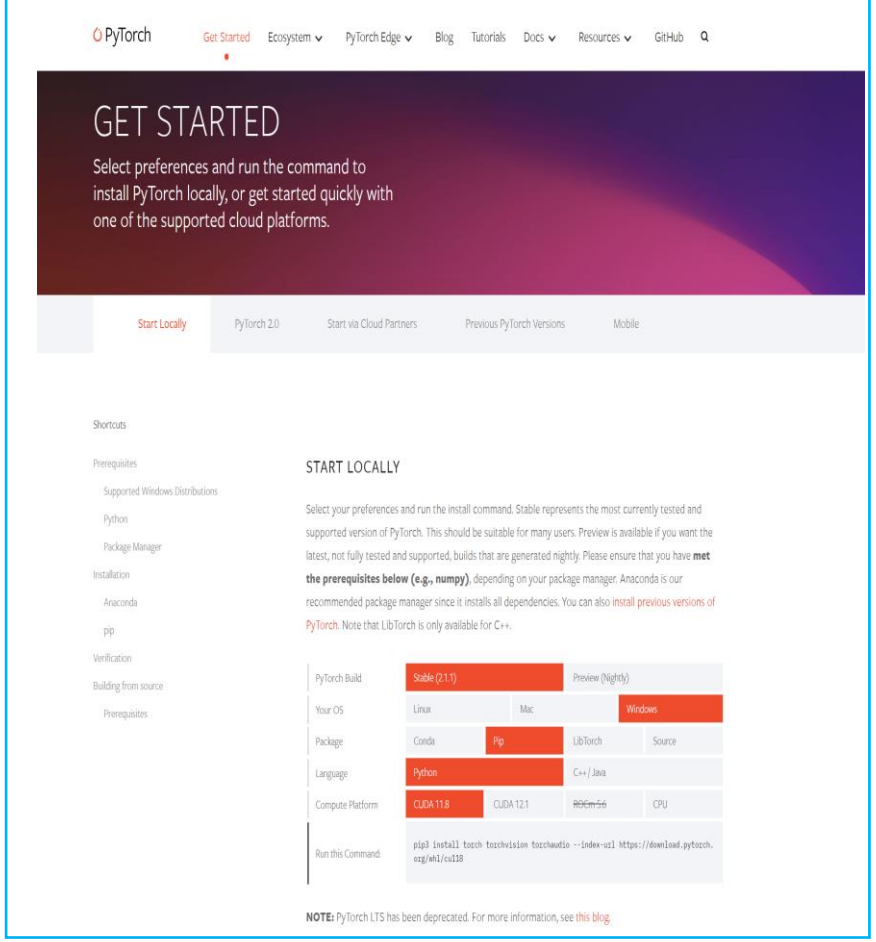
Supervised learning	Unsupervised learning
► 1.1. Linear Models ► 1.2. Linear and Quadratic Discriminant Analysis - 1.3. Kernel ridge regression ► 1.4. Support Vector Machines ► 1.5. Stochastic Gradient Descent ► 1.6. Nearest Neighbors ► 1.7. Gaussian Processes ► 1.8. Cross decomposition ► 1.9. Naive Bayes ► 1.10. Decision Trees ► 1.11. Ensembles: Gradient boosting, random forests, bagging, voting, stacking ► 1.12. Multiclass and multioutput algorithms ► 1.13. Feature selection ► 1.14. Semi-supervised learning - 1.15. Isotonic regression ► 1.16. Probability calibration ► 1.17. Neural network models (supervised)	► 2.1. Gaussian mixture models ► 2.2. Manifold learning ► 2.3. Clustering ► 2.4. Biclustering ► 2.5. Decomposing signals in components (matrix factorization problems) ► 2.6. Covariance estimation ► 2.7. Novelty and Outlier Detection ► 2.8. Density Estimation ► 2.9. Neural network models (unsupervised)
Uses **labeled** data to train a machine learning model and predict the output	Uses **unlabeled** data to find patterns and structure in the data
Classification and regression problems, such as spam detection, face recognition, or stock price prediction.	Clustering, association, and dimensionality reduction problems, such as market segmentation, recommendation systems, or feature extraction.
Can measure its accuracy and performance by comparing the predicted output with the actual output.	Does not have a clear way to evaluate its results, and often relies on subjective criteria or domain knowledge

Evolution

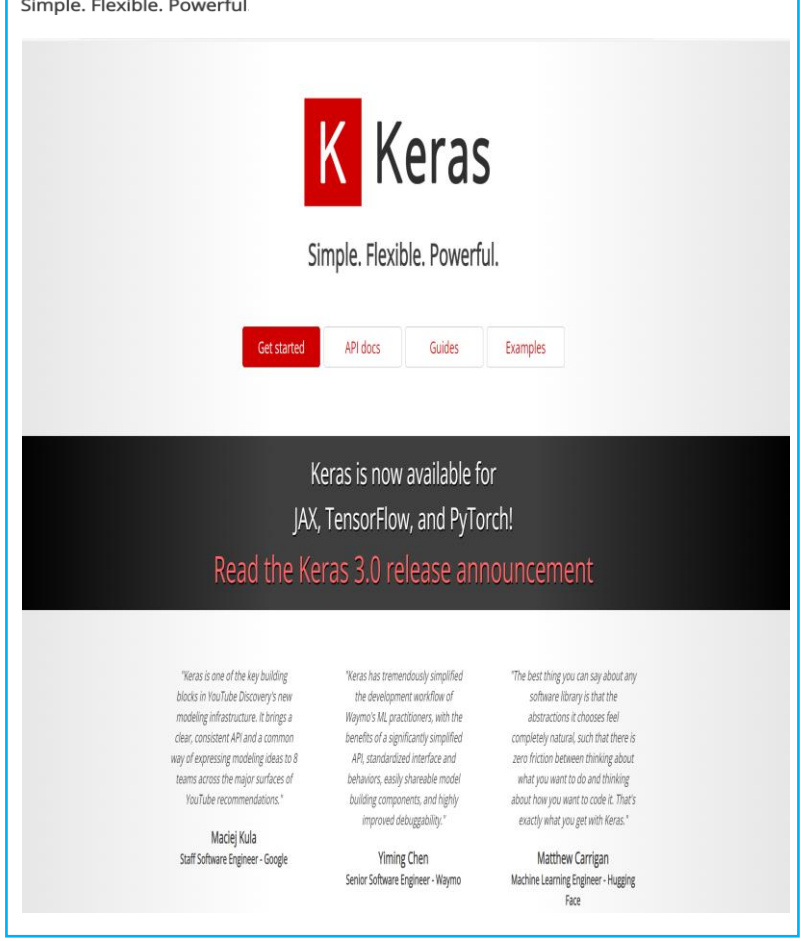
Many open source stacks and libraries on-line…

TensorFlow https://www.tensorflow.org/

PyTorch https://pytorch.org/



Keras https://keras.io/



Evaluation

Strong recommend study machine learning with scikit-learn(Open source, commercially usable)

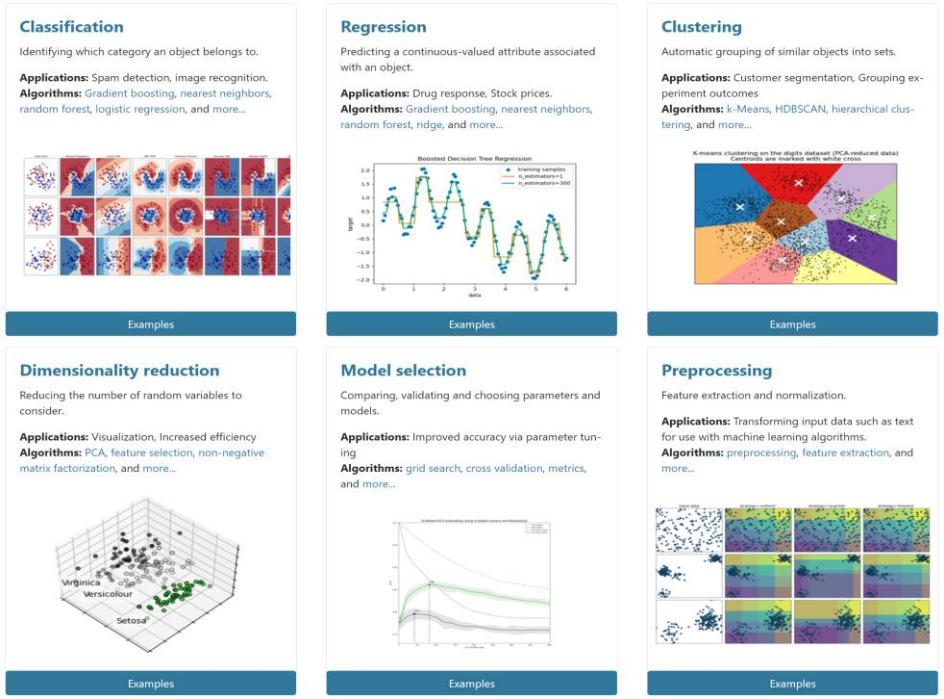
https://scikit-learn.org/stable/

scikit-learn

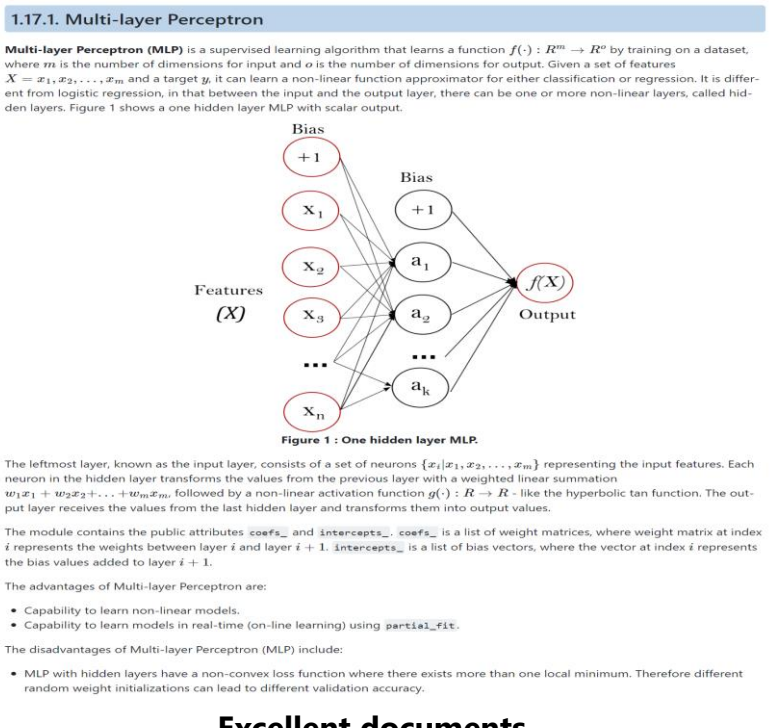
Machine Learning in Python

Getting Started | Release Highlights for 1.3 | GitHub

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license



Complete process and examples



Excellent documents

```python
import warnings

import matplotlib.pyplot as plt

from sklearn.datasets import fetch_openml
from sklearn.exceptions import ConvergenceWarning
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier

# Load data from https://www.openml.org/d/554
X, y = fetch_openml(
    "mnist_784", version=1, return_X_y=True, as_frame=False, parser="pandas"
)
X = X / 255.0

# Split data into train partition and test partition
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, test_size=0.7)

mlp = MLPClassifier(
    hidden_layer_sizes=(40,),
    max_iter=8,
    alpha=1e-4,
    solver="sgd",
    verbose=10,
    random_state=1,
    learning_rate_init=0.2,
)

# this example won't converge because of resource usage constraints on
# our Continuous Integration infrastructure, so we catch the warning and
# ignore it here
with warnings.catch_warnings():
    warnings.filterwarnings("ignore", category=ConvergenceWarning, module="sklearn")
    mlp.fit(X_train, y_train)

print("Training set score: %f" % mlp.score(X_train, y_train))
print("Test set score: %f" % mlp.score(X_test, y_test))

fig, axes = plt.subplots(4, 4)
# use global min / max to ensure all weights are shown on the same scale
vmin, vmax = mlp.coefs_[0].min(), mlp.coefs_[0].max()
for coef, ax in zip(mlp.coefs_[0].T, axes.ravel()):
    ax.matshow(coef.reshape(28, 28), cmap=plt.cm.gray, vmin=0.5 * vmin, vmax=0.5 * vmax)
    ax.set_xticks(())
    ax.set_yticks(())

plt.show()
```

Clean & elegant code

Evolution

Now Game changer : LLM(Large Language Model) achieve great victory in NLP(Nature Language Process)…

- **Large language model (LLM)** : A type of AI algorithm that uses deep learning techniques and massively large data sets to understand, summarize, generate and predict new content(*generative AI*).
- **Transformer Models:** The specific kind of neural networks used by LLM, which includes encoder and decoder. Encoder used for create representations that capture the meaning and context of the text, decoder used for generating content.
  1. Use unsupervised learning approach to train on unstructured data and unlabeled data to make the model begins to derive relationships between different words and concepts
  2. Fine-tuning with a form of self-supervised learning assisting the model to more accurately identify different concepts.
  3. Go thought LLM to understand and recognize the relationships and connections between words and concepts using a **self-attention mechanism.**
- Professor Hinton made great contributes here.

A fast learning algorithm for deep belief nets *

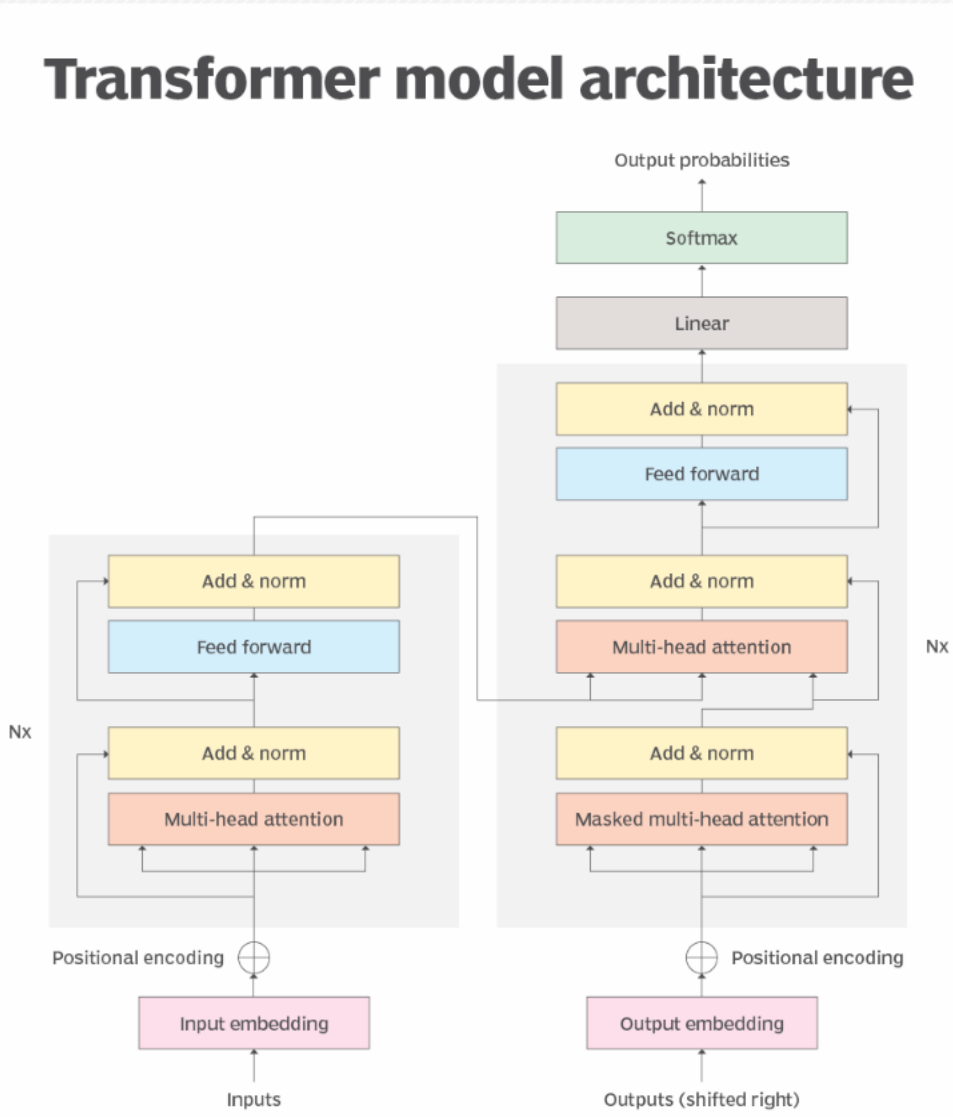
Geoffrey E. Hinton and **Simon Osindero**
Department of Computer Science University of Toronto
10 Kings College Road
Toronto, Canada M5S 3G4
{hinton, osindero}@cs.toronto.edu

Yee-Whye Teh
Department of Computer Science
National University of Singapore
3 Science Drive 3, Singapore, 117543
tehyw@comp.nus.edu.sg

Abstract

We show how to use "complementary priors" to eliminate the explaining away effects that make inference difficult in densely-connected belief nets that have many hidden layers. Using complementary priors, we derive a fast, greedy algorithm that can learn deep, directed belief networks one layer at a time, provided the top two layers form a directed acyclic graph that converts the representations in the associative memory into observable variables such as the pixels of an image. This hybrid model has some attractive features:

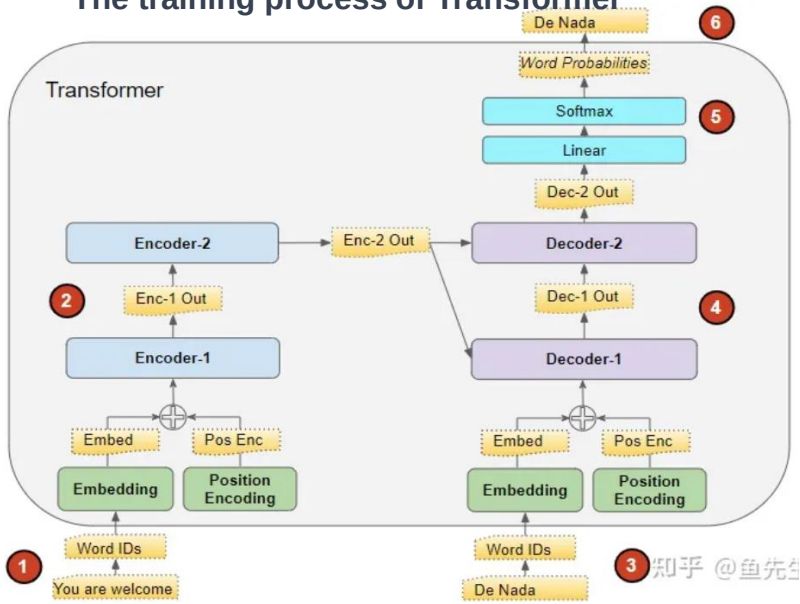
1. There is a fast, greedy learning algorithm that can find a fairly good set of parameters quickly, even in deep networks with millions of parameters and many hidden layers.



Evolution

Generative AI and Transformer: Attention is All You Need

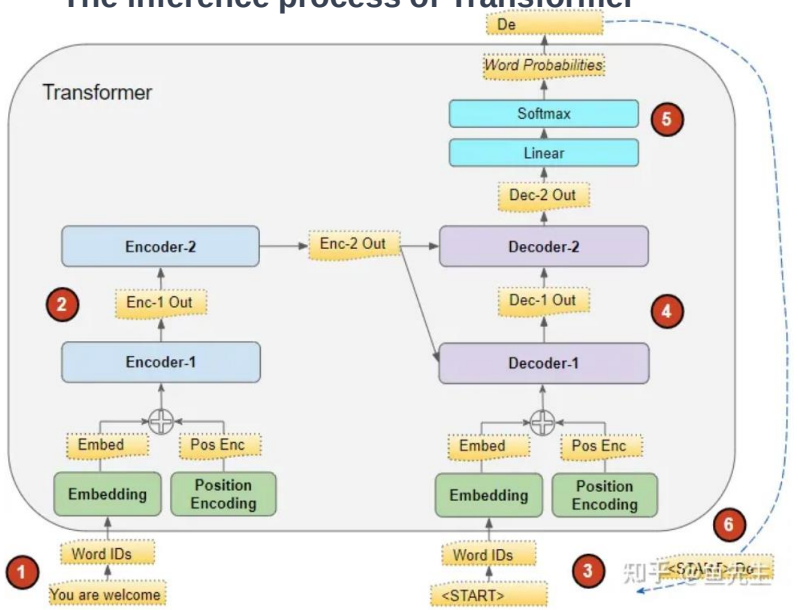
The training process of Transformer



Object : Learn from source sequences and target sequences, aiming to generate the target sequences

1. Before being fed into the first encoder, the input sequence (src_seq) is first transformed into embeddings (with positional encoding), producing word embedding representations (src_position_embed), and then fed into the first encoder
2. The stack of encoders, composed of individual encoders, processes the output from the first step sequentially, generating the encoded representation (enc_outputs) of the input sequence
3. In the decoder stack on the right side, the target sequence is first appended with a start-of-sequence token, transformed into embeddings (with positional encoding), producing word embedding representations (tgt_position_embed), and then fed into the first decoder.
4. The stack of decoders, composed of individual decoders, processes the word embedding representations from the third step (tgt_position_embed) along with the encoded representations from the encoder stack (enc_outputs), generating the decoding representations (dec_outputs) for the target sequence.
5. The output layer transforms it into word probabilities and the final output sequence (out_seq).
6. The **loss function** compares the output sequence (out_seq) with the target sequence (tgt_seq) from the training data. This loss is used to **generate gradients** and train the model during the **backpropagation** process

The inference process of Transformer



Object: To generate the target sequences solely based on the input sequence

1. Step 1 same as the training process
2. Step 2 same as the training process
3. From the third step onward, everything changes: at the first time step, an empty sequence with only a start-of-sequence symbol is used instead of the target sequence used during the training process. The empty sequence is transformed into embeddings with positional encoding (start_position_embed) and is fed into the decoder.
4. The stack of decoders, composed of individual decoders, processes the embeddings of an empty sequence from the third step (start_position_embed) along with the encoded representations from the encoder stack (enc_outputs), generating the encoding representation of the first word in the target sequence (step1_dec_outputs).
5. The output layer transforms it (step1_dec_outputs) into word probabilities and the first target word (step1_tgt_seq).
6. Place the generated target word from this step into the second time step position of the decoder input sequence. At the second time step, the decoder input sequence consists of the token generated from the start-of-sequence symbol and the target word generated at the first time step
7. Return to the third step, just like before, input the new decoder sequence into the model. Then, take the second word from the output and append it to the decoder sequence. Repeat this step until it predicts an end-of-sequence token. It's important to note that since the encoder sequence doesn't change in each iteration, we don't need to repeat steps 1 and 2 every time

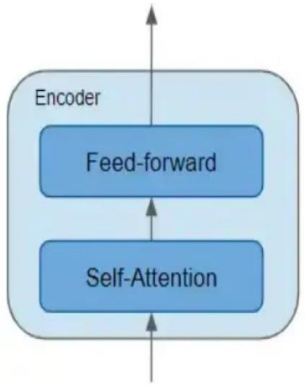
Submitted on **12 Jun 2017** (v1), last revised 2 Aug 2023 (this version, v7)

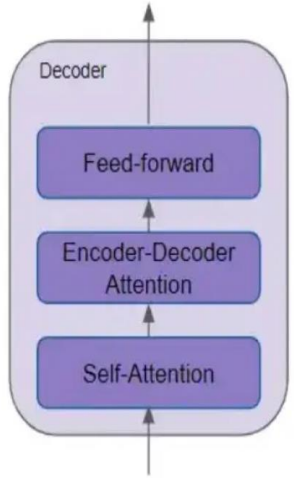
Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Comments:15 pages, 5 figures
Subjects Computation and Language (cs.CL); Machine Learning (cs.LG)
Cite as: arXiv:1706.03762 [cs.CL] (or arXiv:1706.03762v7 [cs.CL] for this version)



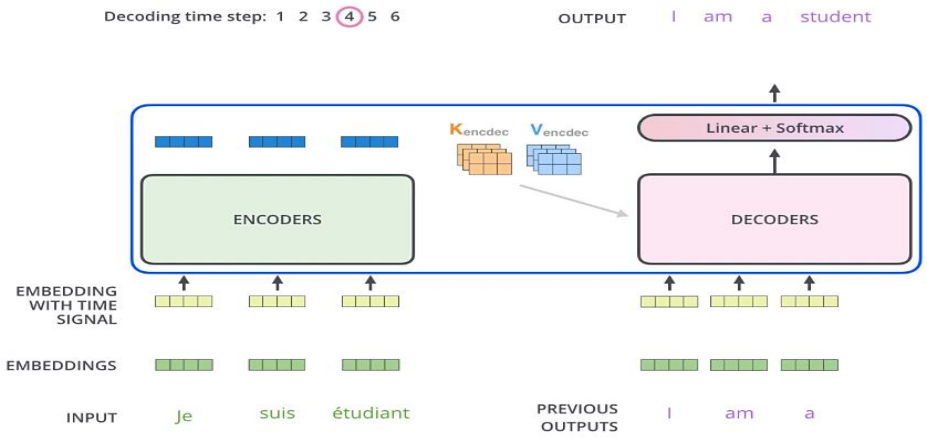


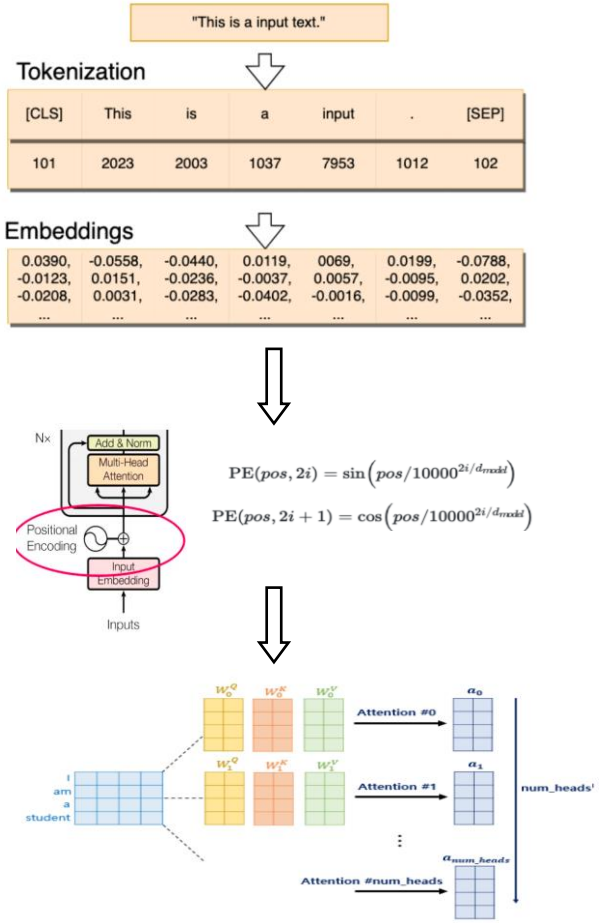
Evolution

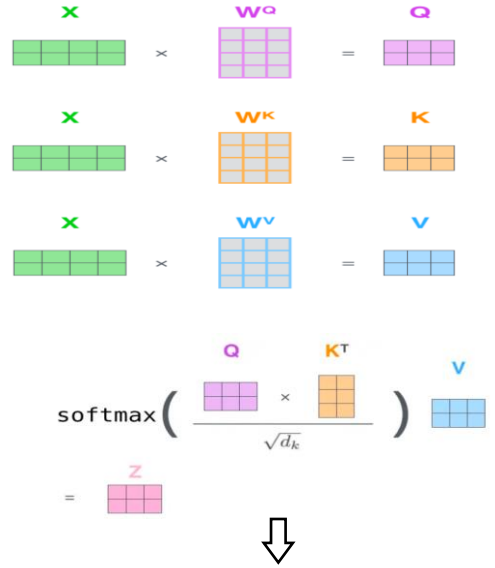
Attention is a serial Matrix & Probability operations indeed…

Machine can’t understand language and far from creativity & intuition.

- *Attention* is a mechanism in the transformer architecture by which contextual word embeddings are determined for words in a corpus.
- Unlike word2vec or gloVe, the attention mechanism takes into account *all* the words in a sentence during the process of creating a word embedding for a given word in a given sentence.



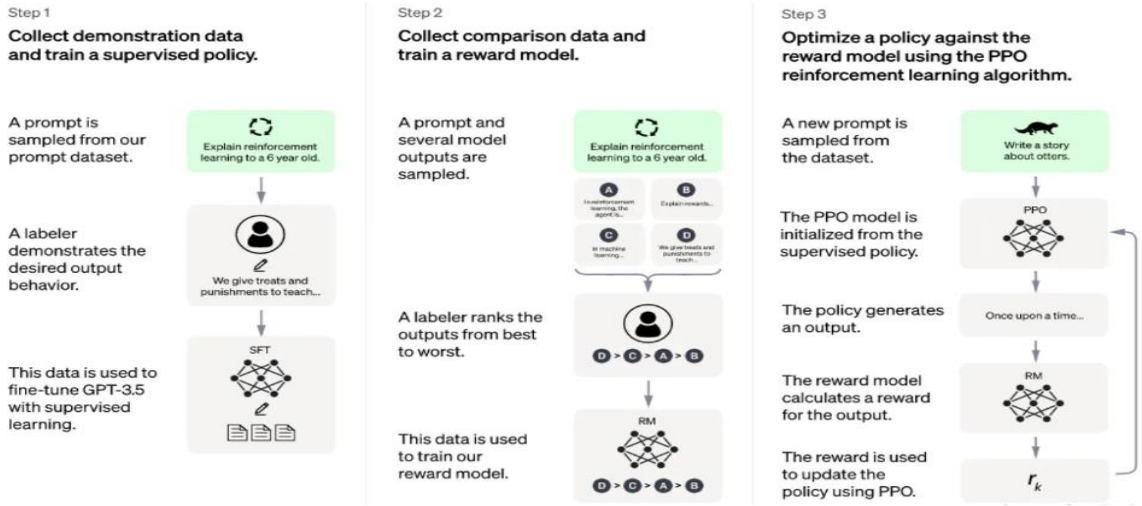


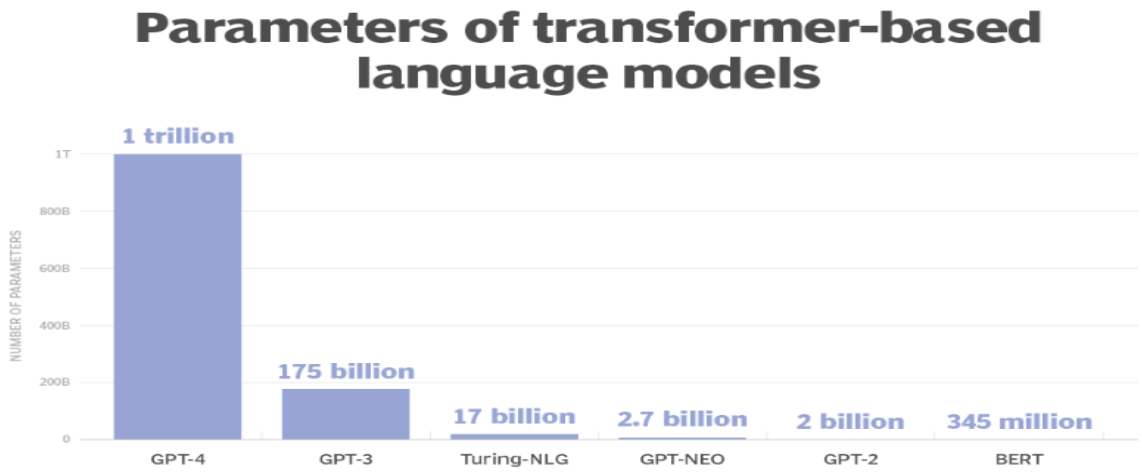


- The output from the decoder is fed into a **Linear layer**, and then transformed using **softmax**, resulting in a vector of vocabulary size.
- Each value in the vector corresponds to the probability of the current word according to the decoder.
- By selecting the word with the **highest probability**, we obtain the current word as the result of the decoder.

Evolution

Target AGI, by version iteration and upgrade, Chatgpt is more and more like human…The Power of Massive Computing Resources and Data.





- GPT4 can respond using up to 25,000 words (8x more than the previous version) and has the ability to process image inputs as well as text, making it multimodal.
- An average of **53%** of people can't tell that ChatGPT content was generated by an AI.
- ChatGPT contains **570 gigabytes** of text data, which is equivalent to roughly **164,129 times** the number of words in the entire Lord of the Rings series (including The Hobbit)
- It is estimated that training the model took **34 days** (OpenAI used **1,023 A100 GPUs** to train ChatGPT).
- The tool costs approximately **$100,000 per day** or **$3 million per month** to run on Microsoft's Azure Cloud, with each word generated costing $0.0003 (The tool runs on over **3,500 Microsoft Azure supercomputers** and uses around **30,000 GPUs**)
- OpenAI is expected to have **1,050 employees** by the end of their first 10 years of business.

Evolution

Gemini VS ChatGPT, Race of AI between tech giants starts…

Why hasn't the public seen programs like ChatGPT from Meta or from Google?"Because Google and Meta both have a lot to lose by putting out systems that make stuff up" says Meta's chief AI scientist.

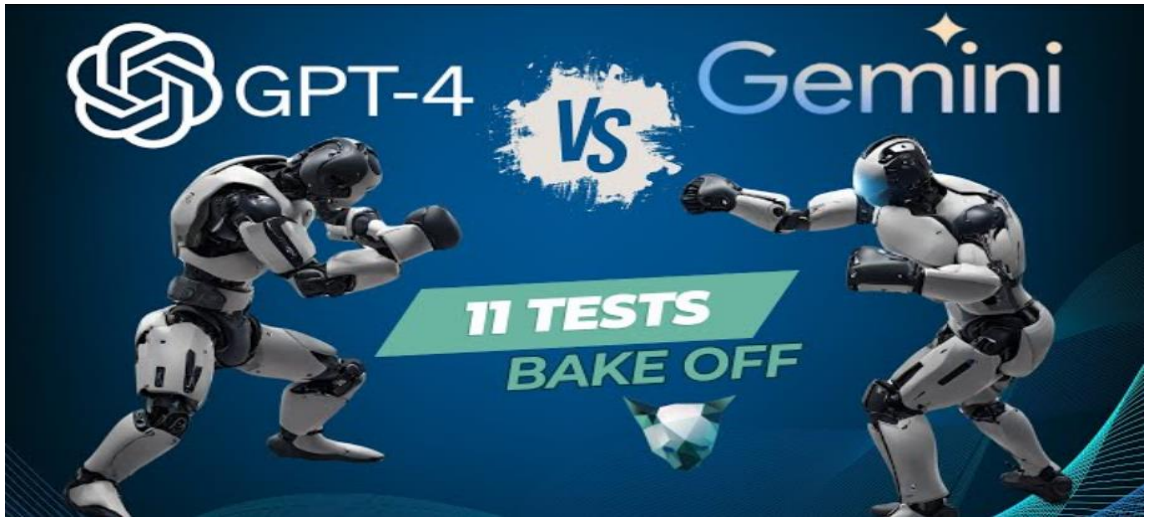
Yann LeCun.

To avoid risks, tech giants target startups, invest or acquire

- Google Acquires Artificial Intelligence Startup DeepMind For More Than $500M in 2014
- Microsoft plans to invest $10 billion into OpenAI

On Wednesday, 6 December 2023, Google launched a generative AI model called Gemini

- **A multimodal model that incorporates AlphaGo-inspire techniques including reinforcement learning and tree search.**
- **A product of the collaboration of Brain and DeepMind AI labs, has been trained on proprietary data, taken across Google's vast array of services including, Google Search, Google Books, Youtube, and Google Scholar.**
- **The number of tokens it has been trained on as well as its parameters exceeded GPT-4.**
  - tokens : Double the number of GPT-4.
  - parameters： 30 trillion or even 65 trillion > 1.75 trillion .



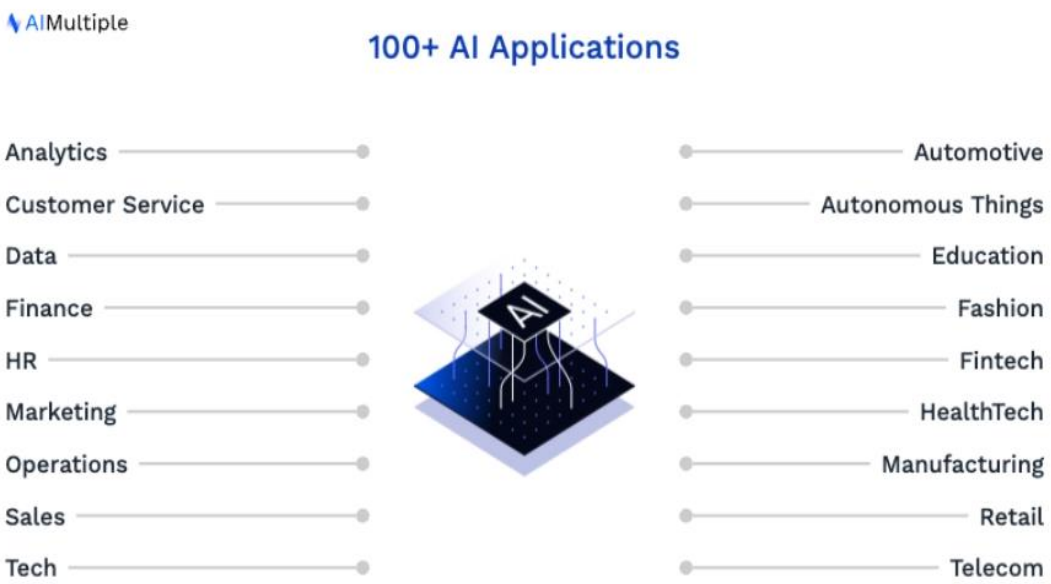
Trend

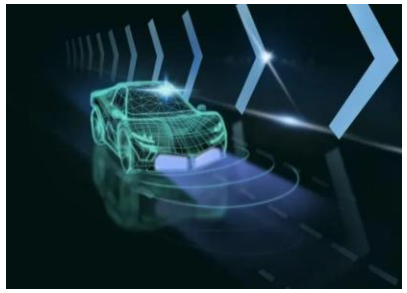
We are entering a new world. The technologies of machine learning, speech recognition, and natural language understanding are reaching a nexus of capability. The end result is that we'll soon have artificially intelligent assistants to help us in virtually every aspect of our lives.

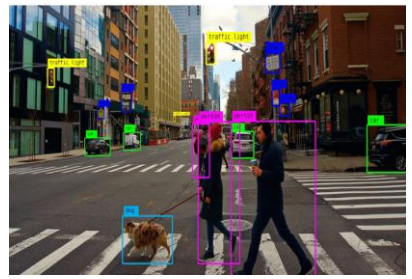
Satya Nadella, CEO of Microsoft

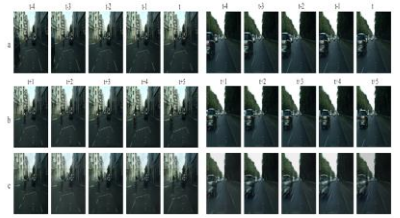
Trend

AI is changing every industry and business function…

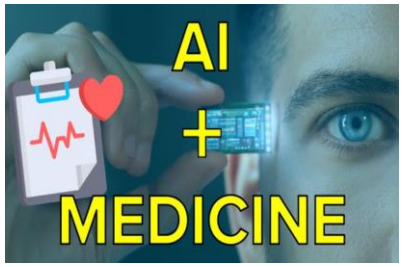


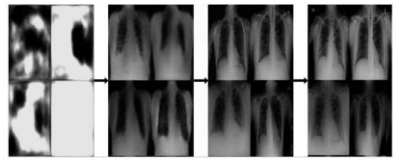






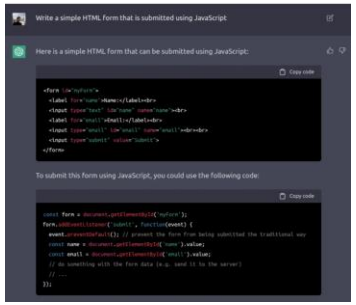
An example of video prediction is depicted in Column C





14-hour-long progressive outputs of a generative AI model to produce fake chest X-rays.

This AI-generated image was produced based on the text description of "Teddy bears shopping for groceries in ukiyo-e style" .



[56%](#) of organizations are using AI in at least one business function.

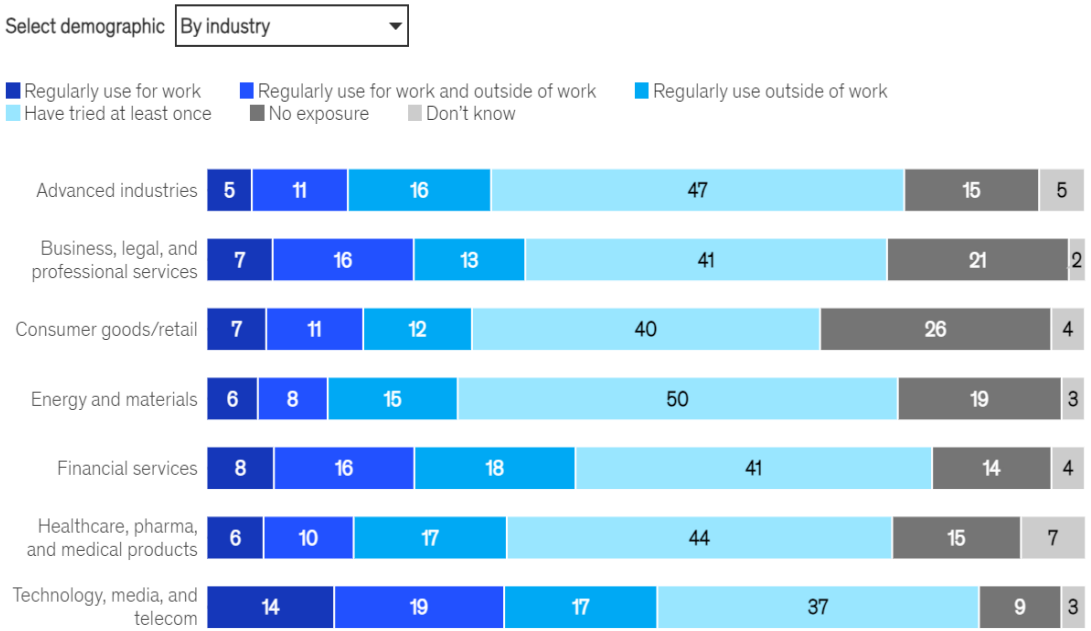
--- McKinsey survey ,2021

Trend

The state of AI in 2023: Generative AI's breakout year

Respondents across regions, industries, and seniority levels say they are already using generative AI tools.

Reported exposure to generative AI tools, % of respondents

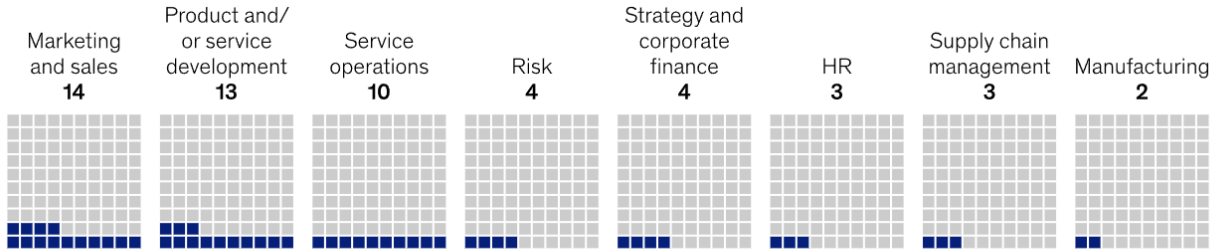


Note: Figures may not sum to 100%, because of rounding. In Asia–Pacific, n = 164; in Europe, n = 515; in North America, n = 392; in Greater China (includes Hong Kong and Taiwan), n = 337; and in developing markets (includes India, Latin America, and Middle East and North Africa), n = 276. For advanced industries (includes automotive and assembly, aerospace and defense, and advanced electronics), n = 96; for business, legal, and professional services, n = 215; for consumer goods and retail, n = 128; for energy and materials, n = 96; for financial services, n = 248; for healthcare, pharma, and medical products, n = 130; and for technology, media, and telecom, n = 244. For C-suite respondents, n = 541; for senior managers, n = 437; and for middle managers, n = 339. For respondents born in 1964 or earlier, n = 143; for respondents born between 1965 and 1980, n = 268; and for respondents born between 1981 and 1996, n = 80. Age details were not available for all respondents. For respondents identifying as men, n = 1,025; for respondents identifying as women, n = 156. The survey sample also included respondents who identified as "nonbinary" or "other" but not a large enough number to be statistically meaningful.
Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

McKinsey & Company

The most commonly reported uses of generative AI tools are in marketing and sales, product and service development, and service operations.

Share of respondents reporting that their organization is regularly using generative AI in given function, %[1]



Most regularly reported generative AI use cases within function, % of respondents

Marketing and sales		Product and/or service development		Service operations	
Crafting first drafts of text documents	9	Identifying trends in customer needs	7	Use of chatbots (eg, for customer service)	6
Personalized marketing	8	Drafting technical documents	5	Forecasting service trends or anomalies	5
Summarizing text documents	8	Creating new product designs	4	Creating first drafts of documents	5

[1]Questions were asked of respondents who said their organizations have adopted AI in at least 1 business function. The data shown were rebased to represent all respondents.
Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

McKinsey & Company

Trend

AI has the potential to generate $2.6 trillion to $4.4 trillion in value across industries

Generative AI use cases will have different impacts on business functions across industries.

< Prev 01 — 02 Next >

Generative AI productivity impact by business functions[1]

Low impact — High impact

	Total, % of industry revenue	Total, $ billion	Marketing and sales	Customer operations	Product and R&D	Software engineering	Supply chain and operations	Risk and legal	Strategy and finance	Corporate IT[2]	Talent and organization
			760–1,200	340–470	230–420	580–1,200	280–530	180–260	120–260	40–50	60–90
High tech	4.8–9.3	240–460									
Banking	2.8–4.7	200–340									
Pharmaceuticals and medical products	2.6–4.5	60–110									
Education	2.2–4.0	120–230									
Telecommunications	2.3–3.7	60–100									
Healthcare	1.8–3.2	150–260									
Insurance	1.8–2.8	50–70									
Media and entertainment	1.8–3.1	60–110									
Advanced manufacturing[3]	1.4–2.4	170–290									
Consumer packaged goods	1.4–2.3	160–270									
Advanced electronics and semiconductors	1.3–2.3	100–170									

Note: Figures may not sum to 100%, because of rounding. [1]Excludes implementation costs (eg, training, licenses). [2]Excluding software engineering. [3]Includes aerospace, defense, and auto manufacturing. [4]Including auto retail.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

McKinsey & Company

Generative AI use cases will have different impacts on business functions across industries.

< Prev 02 — 02 Next >

Generative AI productivity impact by business functions[1]

Low impact — High impact

	Total, % of industry revenue	Total, $ billion	Marketing and sales	Customer operations	Product and R&D	Software engineering	Supply chain and operations	Risk and legal	Strategy and finance	Corporate IT[2]	Talent and organization
			760–1,200	340–470	230–420	580–1,200	280–530	180–260	120–260	40–50	60–90
Travel, transport, and logistics	1.2–2.0	180–300									
Retail[4]	1.2–1.9	240–390									
Real estate	1.0–1.7	110–180									
Energy	1.0–1.6	150–240									
Administrative and professional services	0.9–1.4	150–250									
Chemical	0.8–1.3	80–140									
Basic materials	0.7–1.2	120–200									
Construction	0.7–1.2	90–150									
Agriculture	0.6–1.0	40–70									
Public and social sector	0.5–0.9	70–110									
		2,600–4,400									

Note: Figures may not sum to 100%, because of rounding. [1]Excludes implementation costs (eg, training, licenses). [2]Excluding software engineering. [3]Includes aerospace, defense, and auto manufacturing. [4]Including auto retail.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

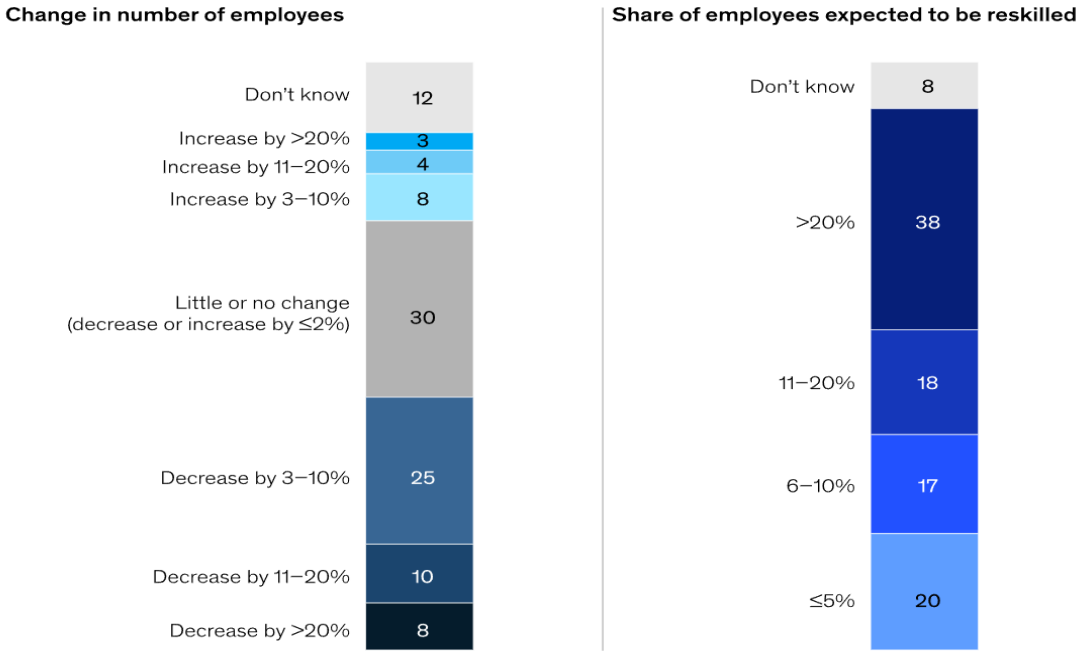
McKinsey & Company

Trend

AI-related talent needs shift, and AI's workforce effects are expected to be substantial

Survey respondents expect AI to meaningfully change their organizations' workforces.

Expectations about the impact of AI adoption on organizations' workforces, next 3 years, % of respondents[1]

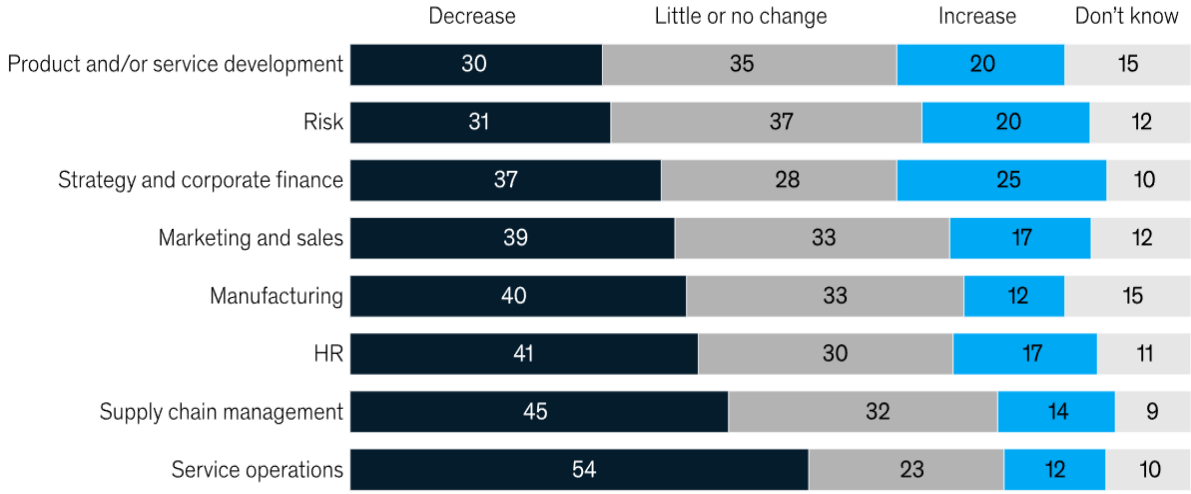


Note: Figures may not sum to 100%, because of rounding.
[1]Asked only of respondents whose organizations have adopted AI in at least 1 function; n = 913.
Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

McKinsey & Company

Service operations is the only function in which most respondents expect to see a decrease in workforce size because of generative AI.

Effect of generative AI adoption on number of employees, by business function, next 3 years, % of respondents[1]

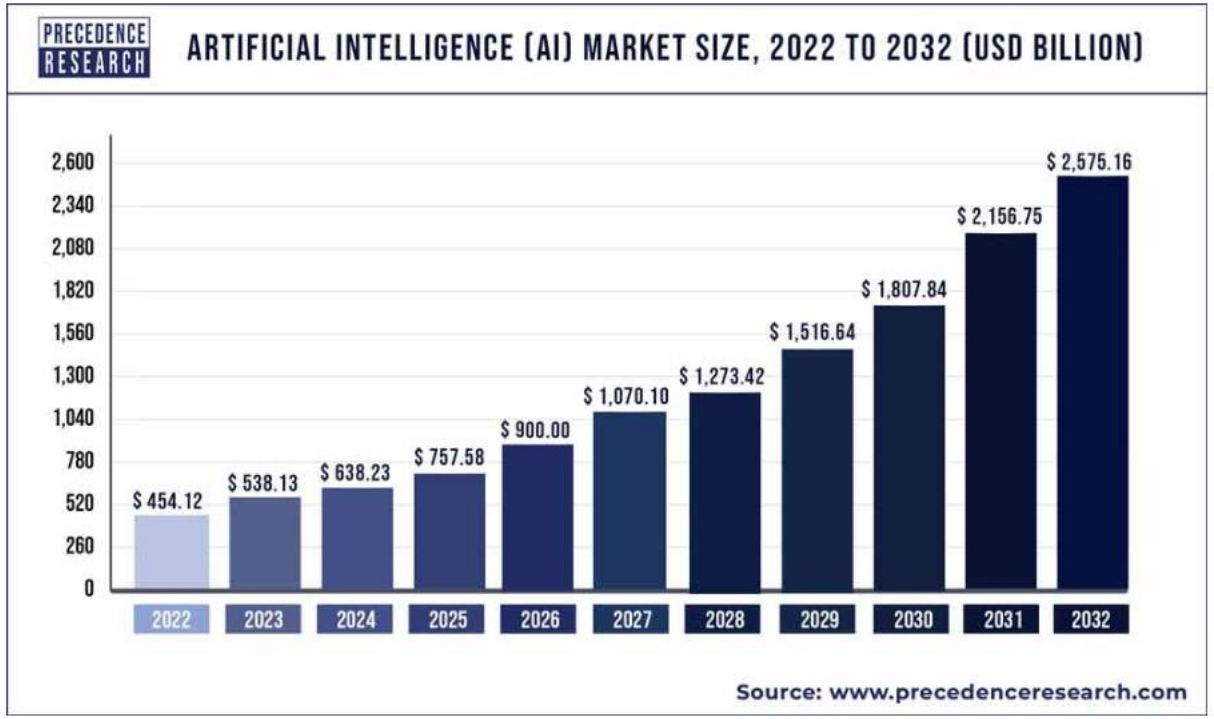


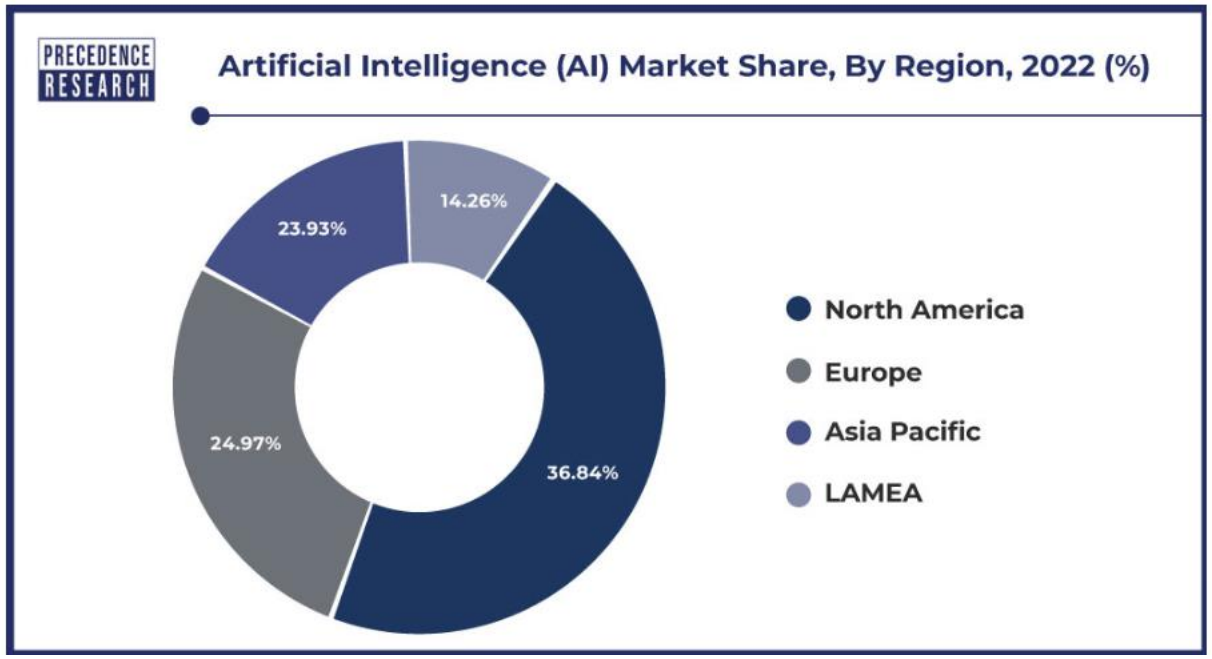
Note: Figures may not sum to 100%, because of rounding.
[1]Respondents were asked about only the business functions in which they said their organizations have adopted AI.
Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

McKinsey & Company

Trend

AI Market size , looks like a big cake...





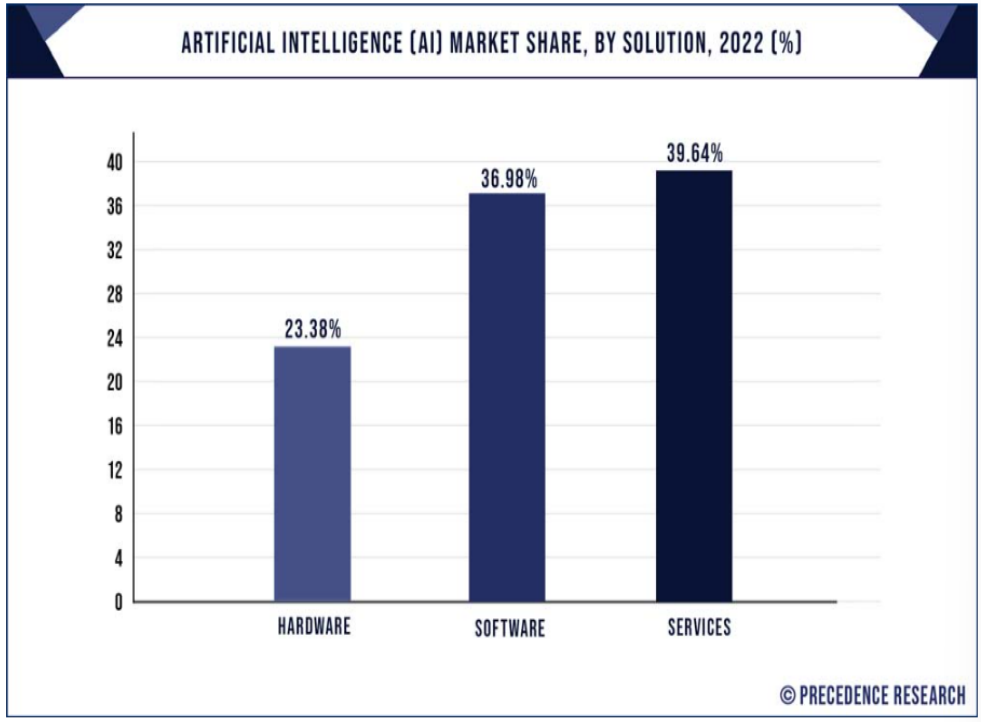
- North America generated more than 36.84% of the market share in 2022.
- The Asia Pacific market is expected to expand at the highest CAGR of 20.3% from 2023 to 2032.
- Based on the technology, the deep learning segment has captured a 36.36% market share in 2022.
- By solution, the services segment has accounted for a market share of over 39.64% in 2022.
- By end user, the BFSI segment accounted for 16.82% of the market share in 2022.

- Canada artificial intelligence (AI) market was valued at USD 43.7 billion in 2022 and it is expected to reach at 251.3 billion in 2032, at a CAGR of 19.2% from 2023 to 2032.
- Germany artificial intelligence (AI) market was valued at USD 25.7 billion in 2022 and it is projected to grow at a CAGR of 20.6% from 2023 to 2032.
- South Korea artificial intelligence (AI) market was valued at USD 16.3 billion in 2022 and it is expanding at a CAGR of 21.1% from 2023 to 2032.
- Japan artificial intelligence (AI) market was valued at USD 20.2 billion in 2022 and will reach at CAGR of 21.0% from 2023 to 2032.

Trend

Market size of AI may be bigger than telecom in 2032...

Tech giants never miss ...



Global Artificial Intelligence (AI) Market Revenue, By Solution, 2022-2032 (US$ Billion)

Solution	2022	2023	2027	2032
Hardware	109.20	129.66	260.25	633.13
Software	168.85	200.24	399.66	966.09
Services	176.08	208.23	410.19	975.94

Service

- Google Cloud AI
- Microsoft Azure AI
- IBM Watson
- Amazon Web Services (AWS)

Software

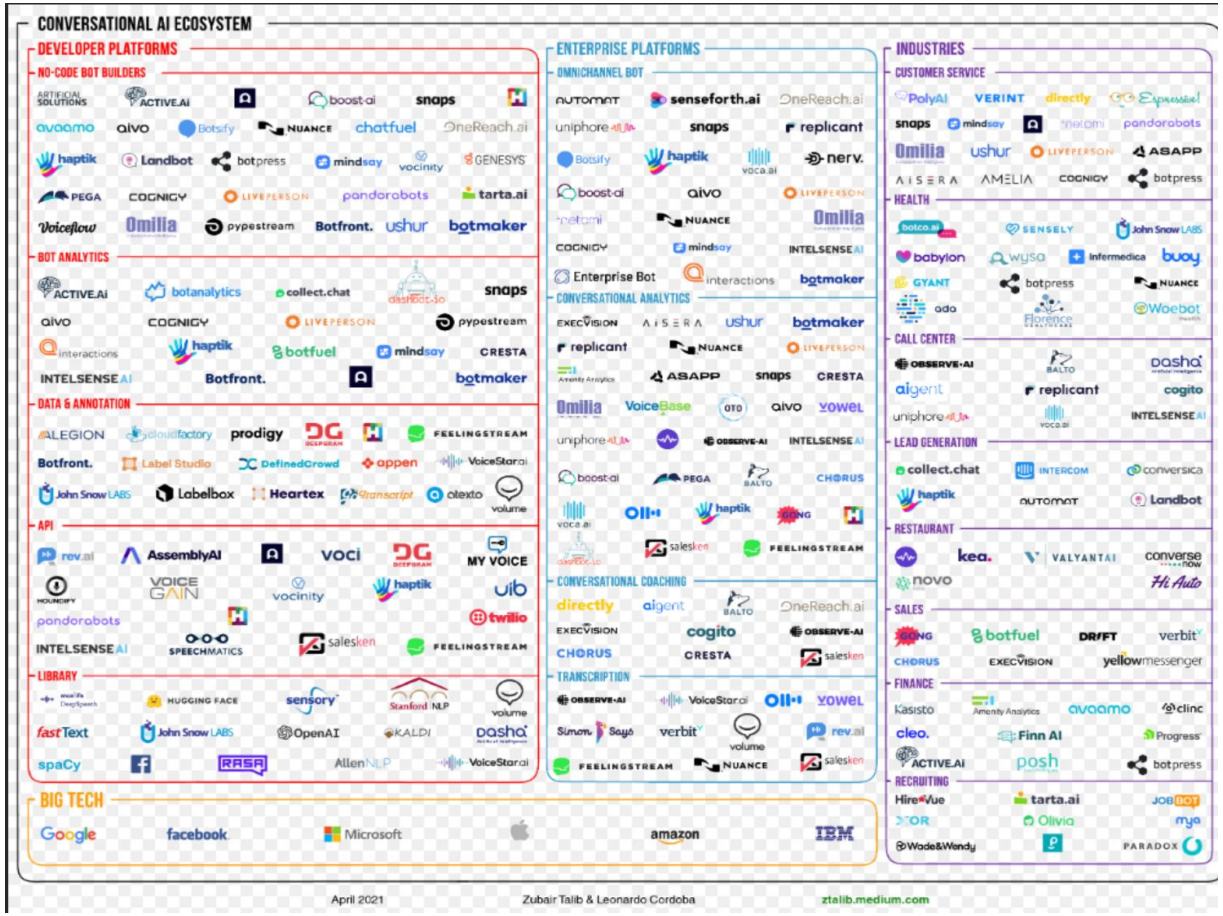
- TensorFlow, TensorFlow Lite, Gemini
- Azure AI, ChatGPT, Copilot
- Watson AI platform
- Amazon SageMaker
- PyTorch

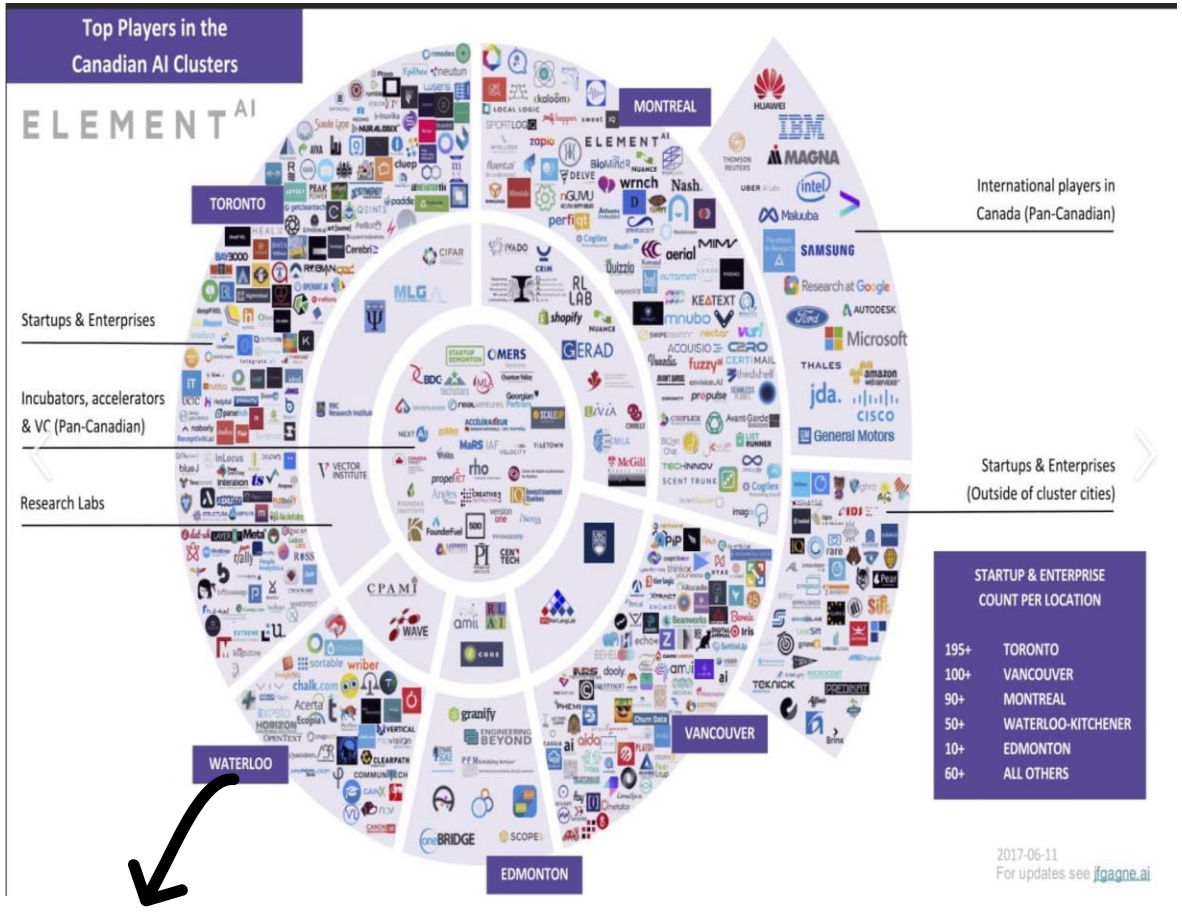
Hardware

- TPU
- GPUs
- PowerAI
- CPUs
- CPUs + GPUs
- Specializes in accelerators and Processors designed for AI

Trend

AI ecosystem is pretty prosperous, even in Canada …

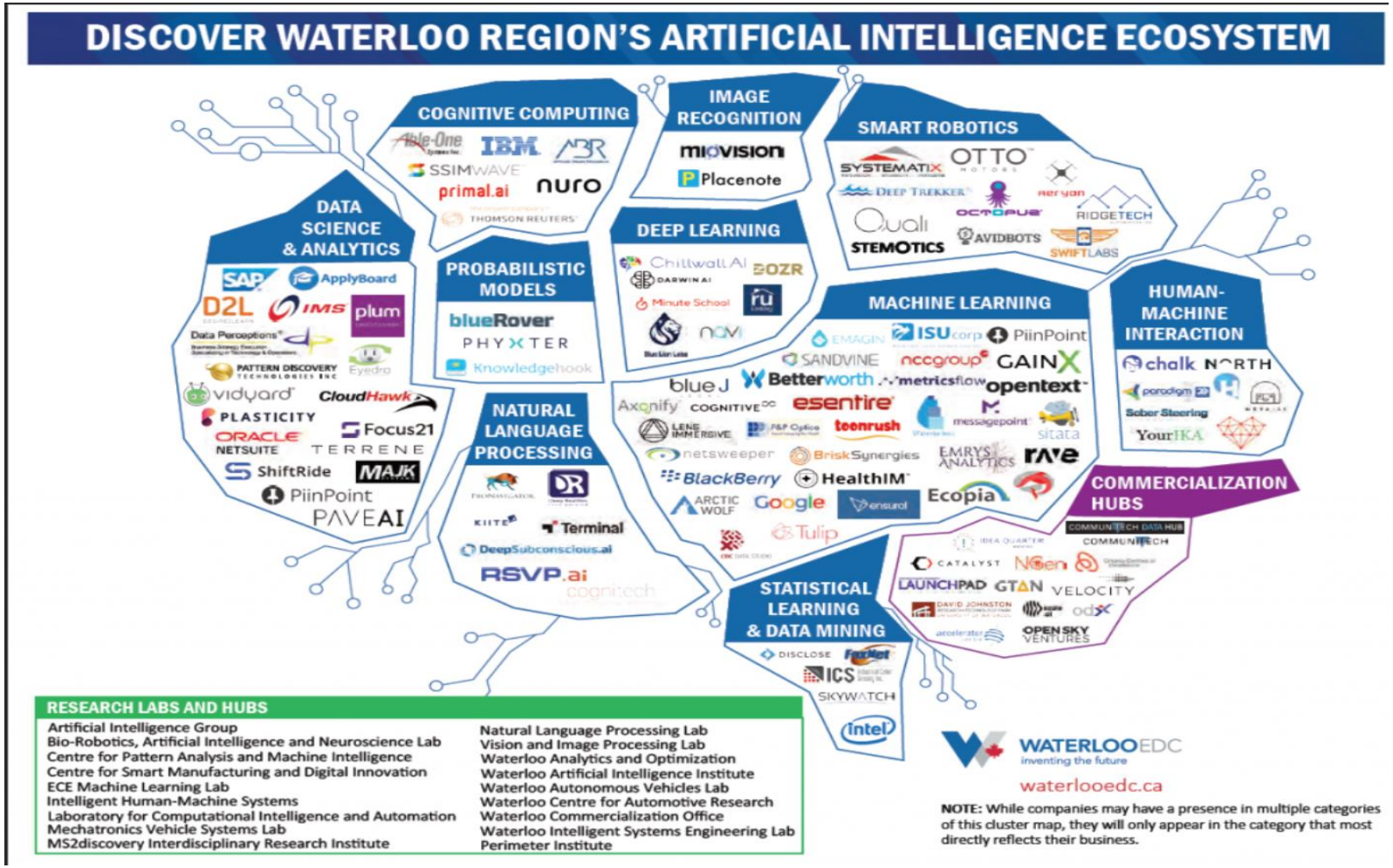




See next page…

Trend

Startups in Waterloo …which will be the next OpenAI ?

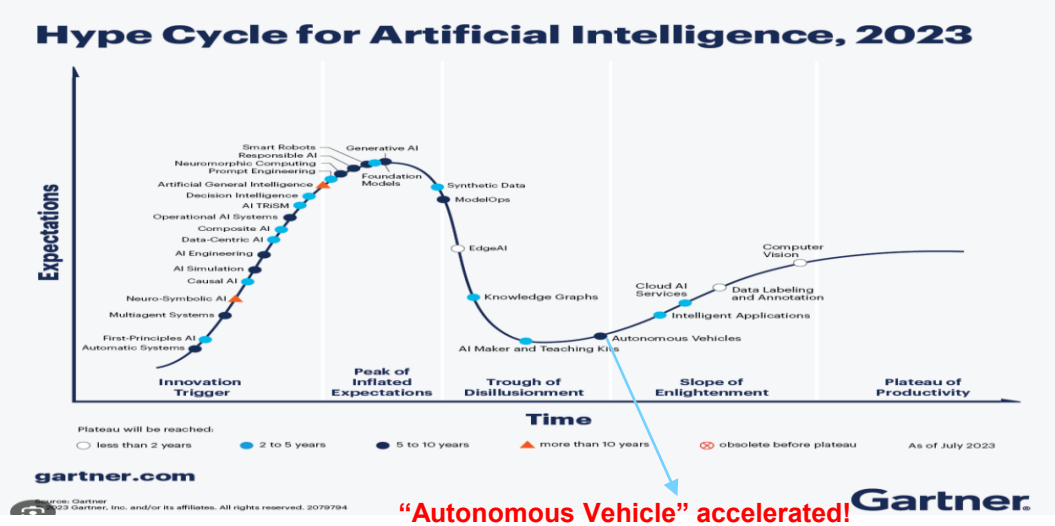


Ilya Sutskever, chief scientist and co-founder of OpenAI.

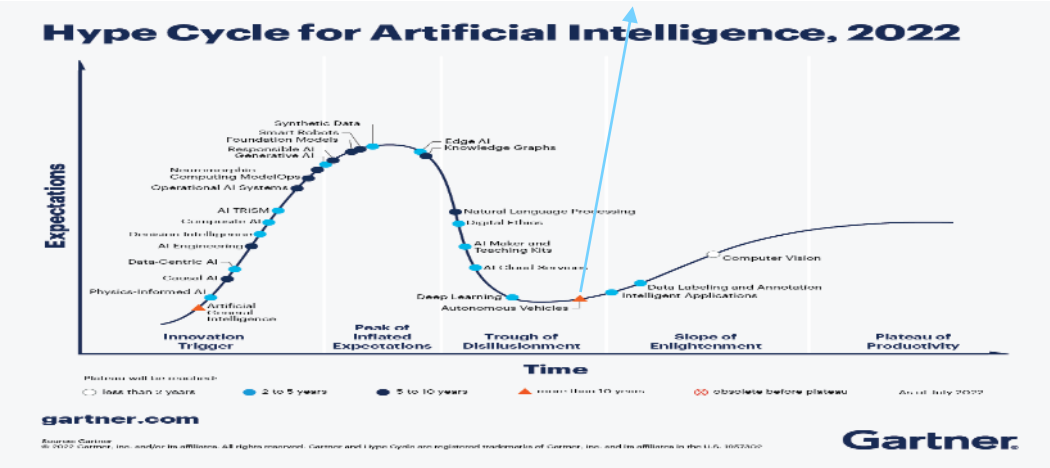
- Born in Soviet Russia in 1986
- **University of Toronto**.
  - ✓ Bachelor of Science in mathematics in 2005,
  - ✓ A MSc in computer science in 2007
  - ✓ A Doctor of Philosophy in computer science in 2013
- His doctoral supervisor is **Geoffrey Hinton** ('godfather of AI')

Trend

Review the AI tech with Gartner, long way to go with tons of opportunities…

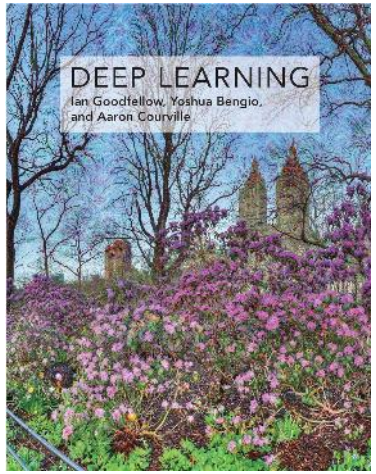


"Autonomous Vehicle" accelerated!



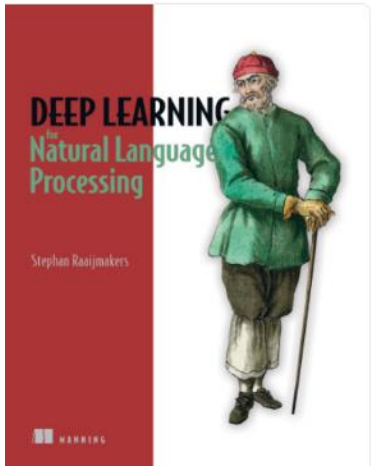
References

Flower



Watermelon

Old guy



Robotic

Bing

Thanks

Artificial intelligence is the future, and the future is here.

Dave Waters, CEO of Alluvium