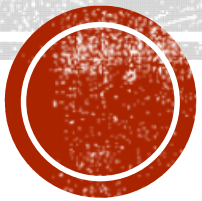
IE4424 Machine Learning Design and Application

Week 1: Introduction



Dr Yap Kim Hui

Email: ekhyap@ntu.edu.sg

References

- Stanford Lecture Notes, CS231n: Convolutional Neural Networks for Visual Recognition.
- Ian Goodfellow, Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, http://www.deeplearningbook.org
- Pytorch Tutorial. https://pytorch.org/tutorials/

Outline

- This lecture will cover the following:
  - Course Overview
  - Introduction to Artificial Intelligence (AI)
  - Deep Neural Network (DNN) Architectures
  - New / Emerging Directions

Course Overview

Teaching

- Instructors:
  - Part A (Week 1-7) : Prof Yap Kim Hui
  - Part B (Week 8-13): Prof Wang Ziwei, Dr Lim Wei Quan
- Course Coordinator: Prof Yap Kim Hui
- Week 1 lecture is a common lecture in assigned LT.
- All other lectures and practices in Week 2-13 will be conducted in the assigned lab, unless advised otherwise.

Tentative Schedule

Week	Topics	Activities	Marks
1	Introduction	Lecture	
2	AI resources and programming.	Lecture + Practice	
3	Convolutional Neural Network (CNN).	Lecture	
4	Design 1: Image Classification Using Convolutional Neural Network (CNN)	Practice	15%
5	Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Transformer.	Lecture	
6	Design 2: Time Series Prediction Using Long Short-Term Memory (LSTM) Neural Network	Practice	15%
7	Quiz 1	Assessment	20%
8	Visual Detection and Segmentation.	Lecture	
9	Design 3: Object Detection and Instance Segmentation.	Practice	15%
10	Large Language Models (LLM).	Lecture	
11	Multimodal Large Language Models (MLLM).	Lecture	
12	Design 4: Instruction Tuning for LLM.	Practice	15%
13	Quiz 2	Assessment	20%

Part A Overview

- Week 1: Introduction to Artificial Intelligence (AI), Machine Learning (ML) and neural networks.
- Week 2: AI resources and programming.
- Week 3: Convolutional Neural Network (CNN).
- Week 4: Design 1 Image Classification Using Convolutional Neural Network (CNN).
- Week 5: Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Transformer.
- Week 6: Design 2 Time Series Prediction Using Long Short-Term Memory (LSTM) Neural Network.
- Week 7: Quiz

Week 2 Activities (Familiarization)

- Topic: AI resources and programming.
- Go to the lab for familiarization and hands-on practices.

Week 3 Activities (Lecture)

- Topic: Convolutional Neural Network (CNN).
- Part 1: Watch the recorded lecture before the lab.
- Part 2: Go to the lab for exercises, discussion, and briefing.
- The total time of Part 1 & 2 is expected to be less than 3 hours.

Week 4 Activities (Design 1: 15%)

- Design 1 Image Classification Using Convolutional Neural Network (CNN).
- Go the lab for Design module. Be punctual.
- Assessment criteria: completed answer sheet, answers to a few short questions at the end of the lab, participation / interaction.

Week 5 Activities (Lecture)

- Topic: Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Transformer.
- Part 1: Watch the recorded lecture before the lab.
- Part 2: Go to the lab for exercises, discussion, and briefing.
- The total time of Part 1 & 2 is expected to be less than 3 hours.

Week 6 Activities (Design 2: 15%)

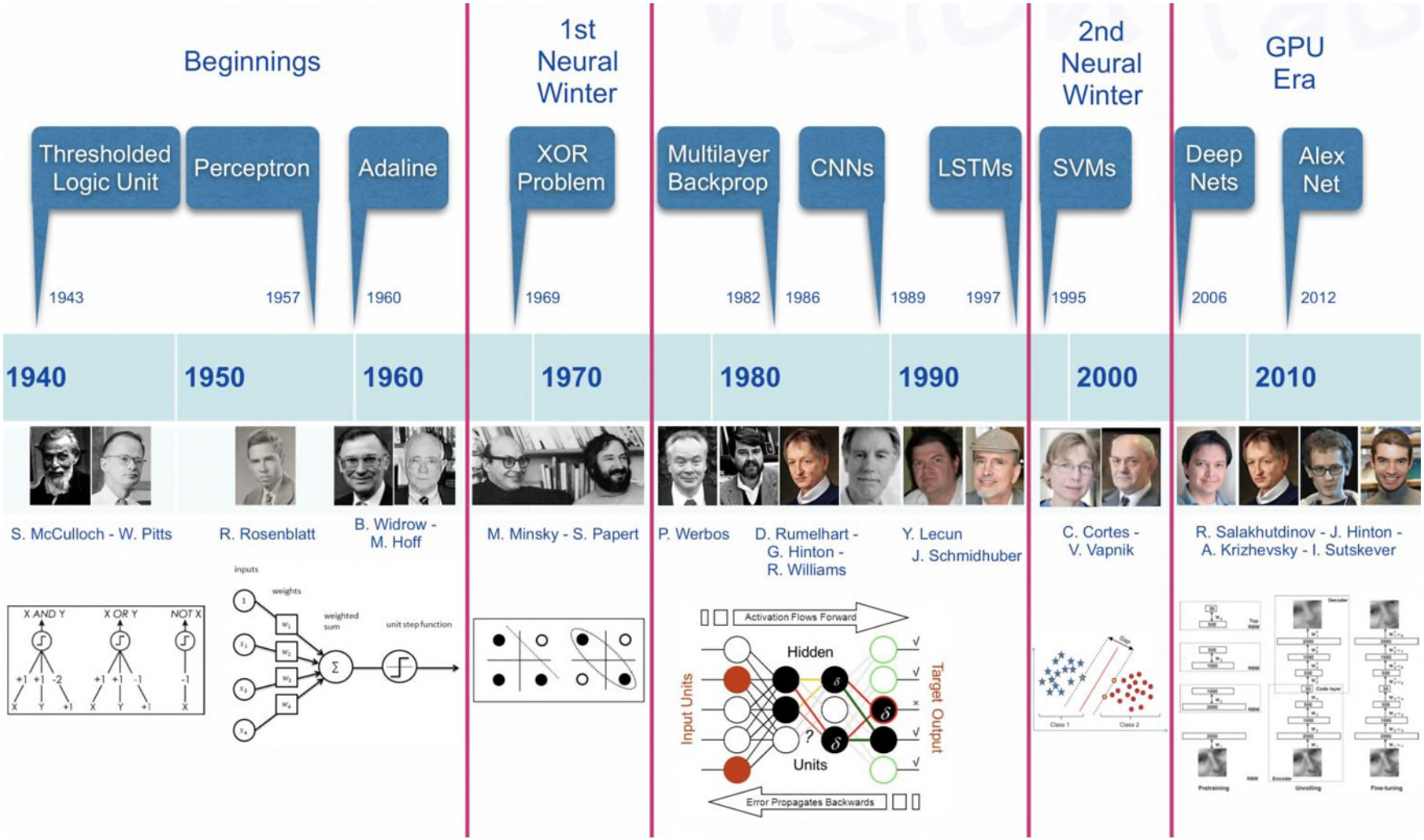
- Design 2 Time Series Prediction Using Long Short-Term Memory (LSTM) Neural Network.
- Go the lab for Design module. Be punctual.
- Assessment criteria: completed answer sheet, answers to a few short questions at the end of the lab, participation / interaction.

Week 7 Quiz 1 (20%)

- Format: closed-book quiz, answer several MCQ and numerical calculation questions.
- Coverage: Week 1-6 materials. Will not focus on the programming syntax of the course.
- Time and venue of the quiz will be announced later.
- More info will be provided closer to the quiz.

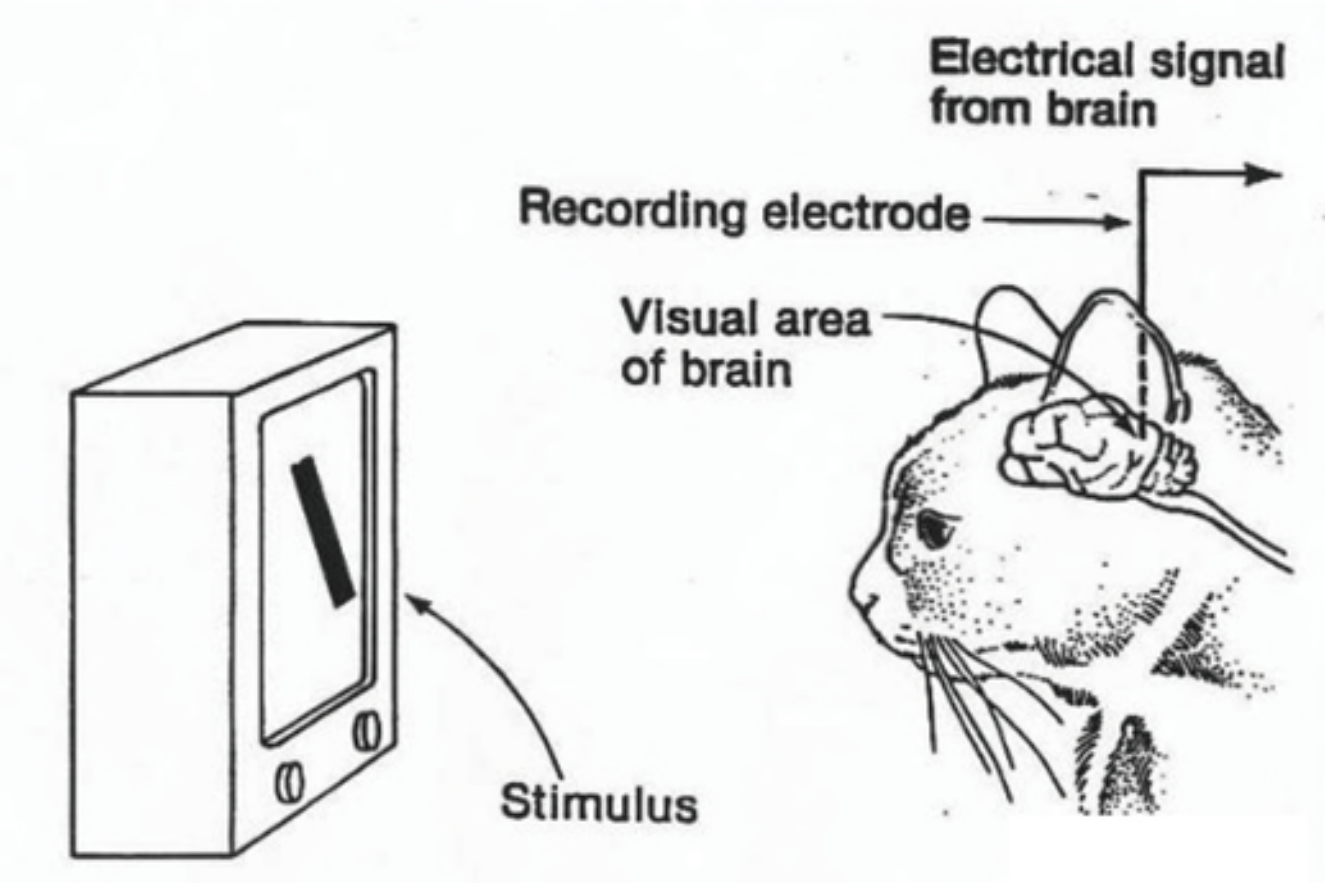
Introduction to Artificial Intelligence (AI)

Brief History of AI



Source: “A Brief History of AI from 1940s till Today”, Deepkapha.ai

Studies of Early Vision



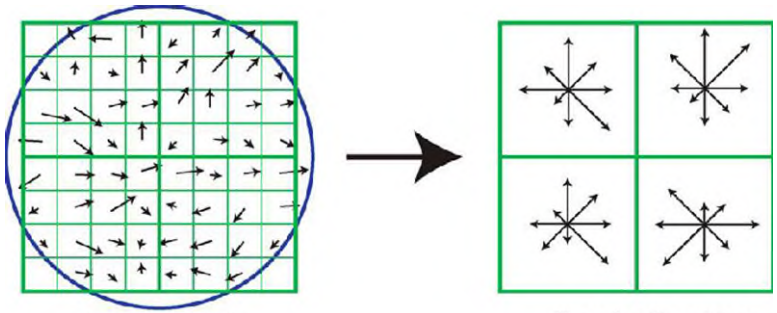
Source: David Hubel and Torsten Wiesel

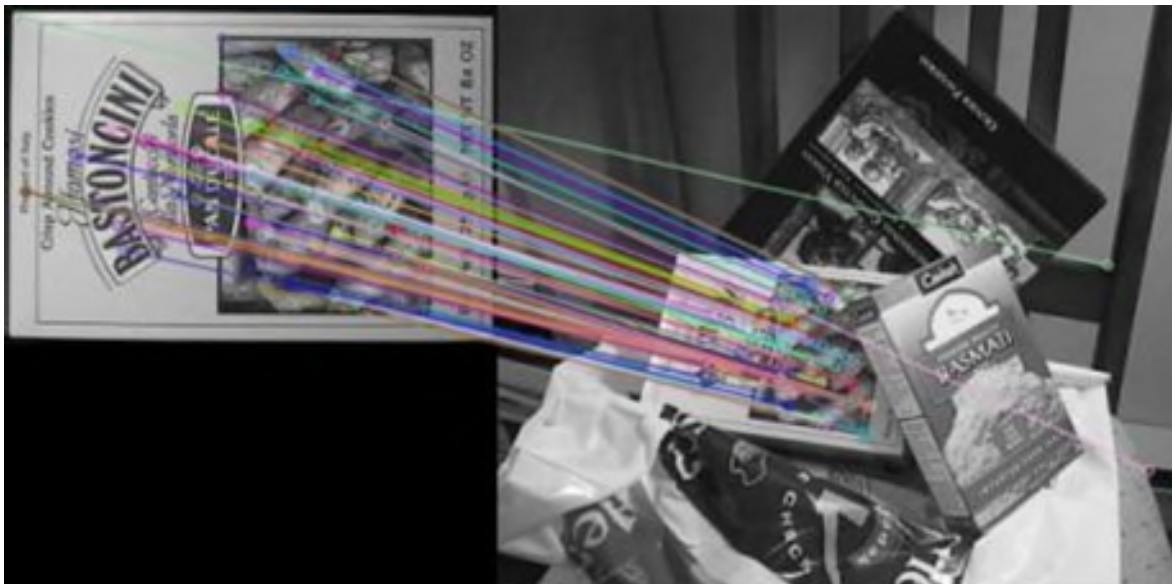
Hand-Crafted Features

- Features designed by humans to represent images.
- Some popular features:
  - Scale Invariant Feature Transform (SIFT)
  - Histogram of Oriented Gradients (HoG)
  - Etc.

Scale Invariant Feature Transform (SIFT)

- Gradient histogram descriptor

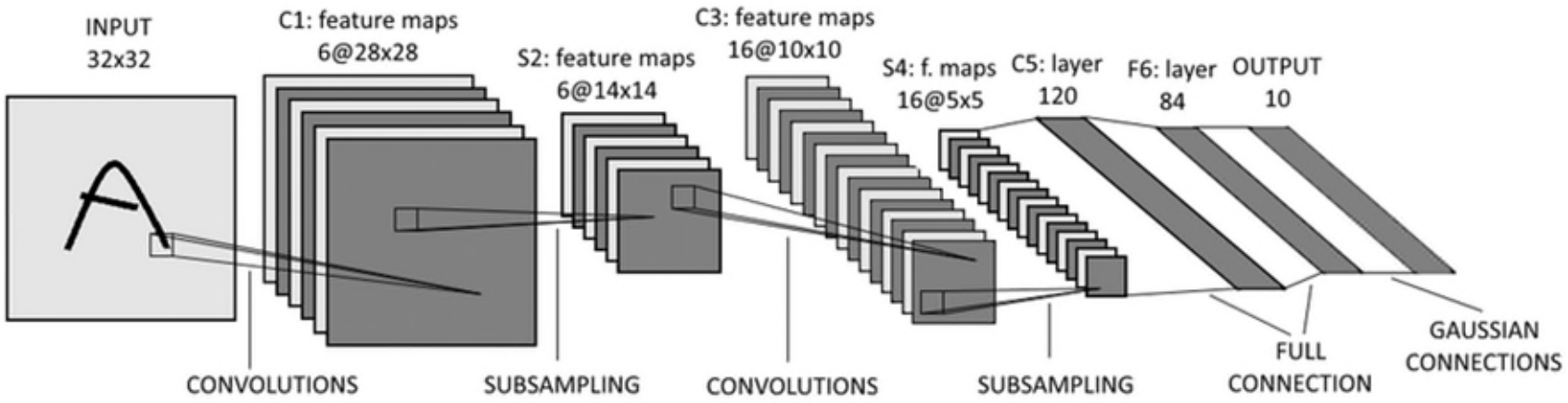




Source: "Distinctive Image Features from Scale-Invariant Keypoints", David G. Lowe, 2004

Deep Learning & Architecture

- Algorithms and architectures that use multiple layers to progressively extract higher level abstraction features from the raw input image.



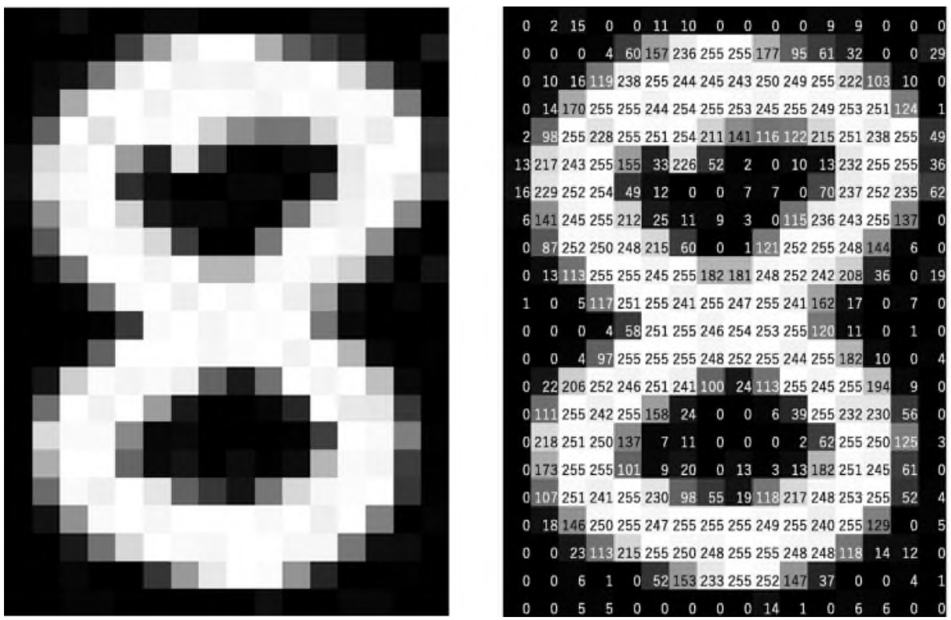
Source: “Gradient-Based Learning Applied to Document Recognition”, Yann LeCun et al., 1998

Discussion: Challenges & Successes in CV

- What is semantic gap?
- What are the imaging challenges in CV tasks such as image classification?
- Why AI is successful in recent years?

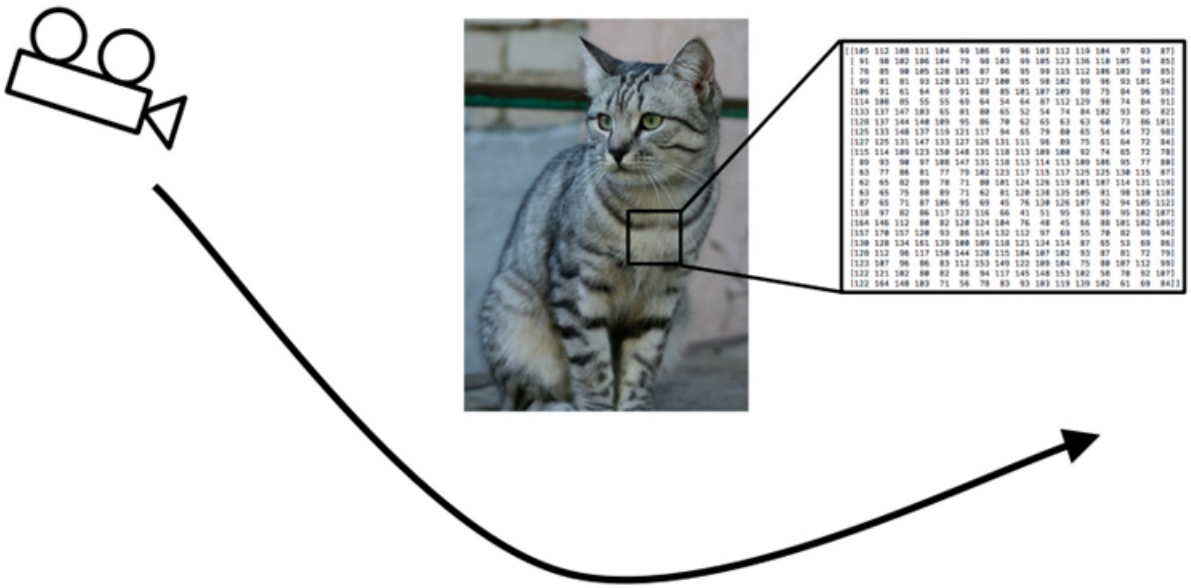
Semantic Gap

- The gap between human semantic understanding of image content and digital representation of image data.



0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

Challenge: Viewpoint Variation



All pixels change when the camera moves!

Source: Stanford Lecture Notes, CS231n

Challenge: Illumination

A: Sample of low illumination image

B: Sample of ideal illumination image

C: Sample of high illumination image

Source: “Effect of image resolution on intensity based scene illumination classification using neural network”, MH Hesamian et al., 2015

Challenge: Deformation




Source: Image from Internet

Challenge: Occlusion




Source: Images from Internet

Challenge: Background Clutter




Source: Images from Internet

Why AI is Successful in Recent Years?

- New AI models & algorithms (e.g., Deep Neural Network Architectures & Deep Learning)
- Large data (e.g., ImageNet dataset)
- Advancement in computing devices (e.g., GPUs)

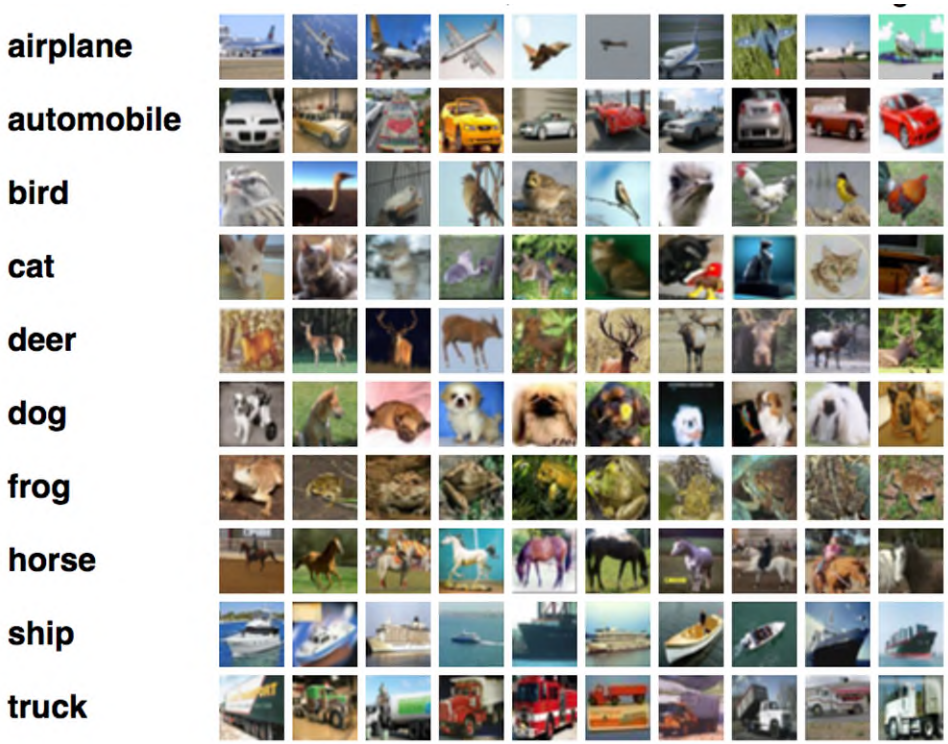
Benchmark Datasets

MNIST

- Large database of handwritten digits.
- Used in training of machine learning algorithms in early years.

CIFAR-10, CIFAR-100

- CIFAR-10: 10 classes, 50,000 training images, 10,000 testing images
- CIFAR-100: 100 classes, 50,000 training images, 10,000 testing images, 20 superclasses



ImageNet

- Key benchmark dataset for training algorithms and evaluation
- Dataset: 14,197,122 images
- ImageNet Large Scale Visual Recognition Challenge (ILSVRC):
  - Image classification challenge for 1000 categories

Microsoft Common Objects in Context (MS-COCO)

- A large-scale object detection, segmentation, and captioning dataset.
- Key features:
  - 330K images (>200K labelled)
  - 1.5 million object instances
  - 80 object categories
  - 5 captions per images

Google Open Images

- Contain 1.9 million images for object detection
  - 16 million bounding boxes
  - 600 categories

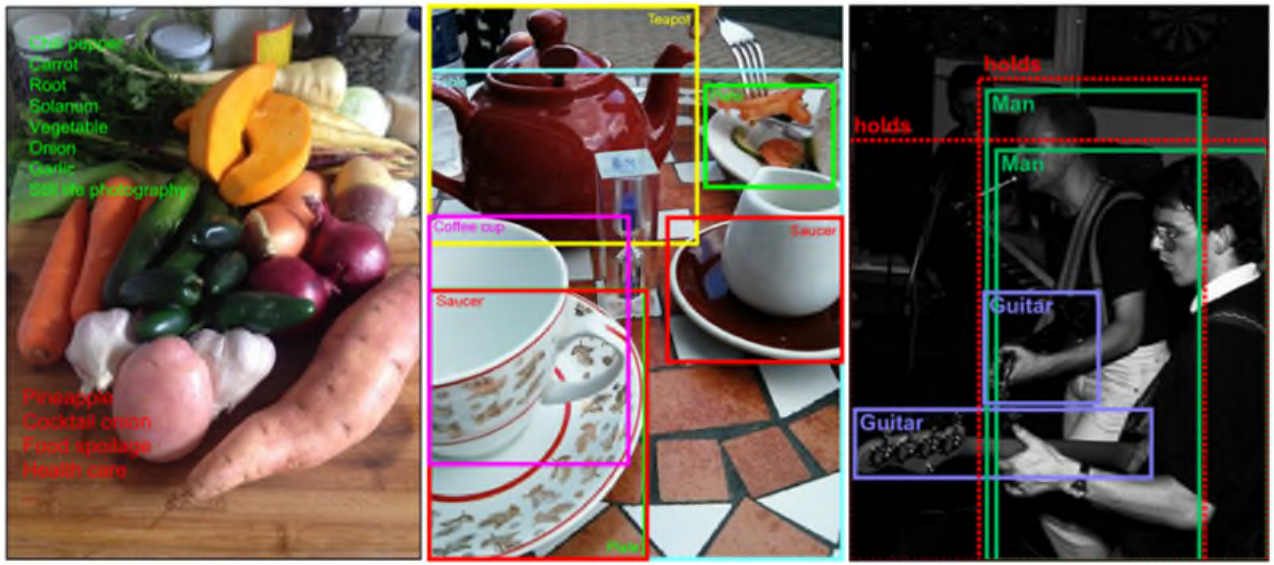
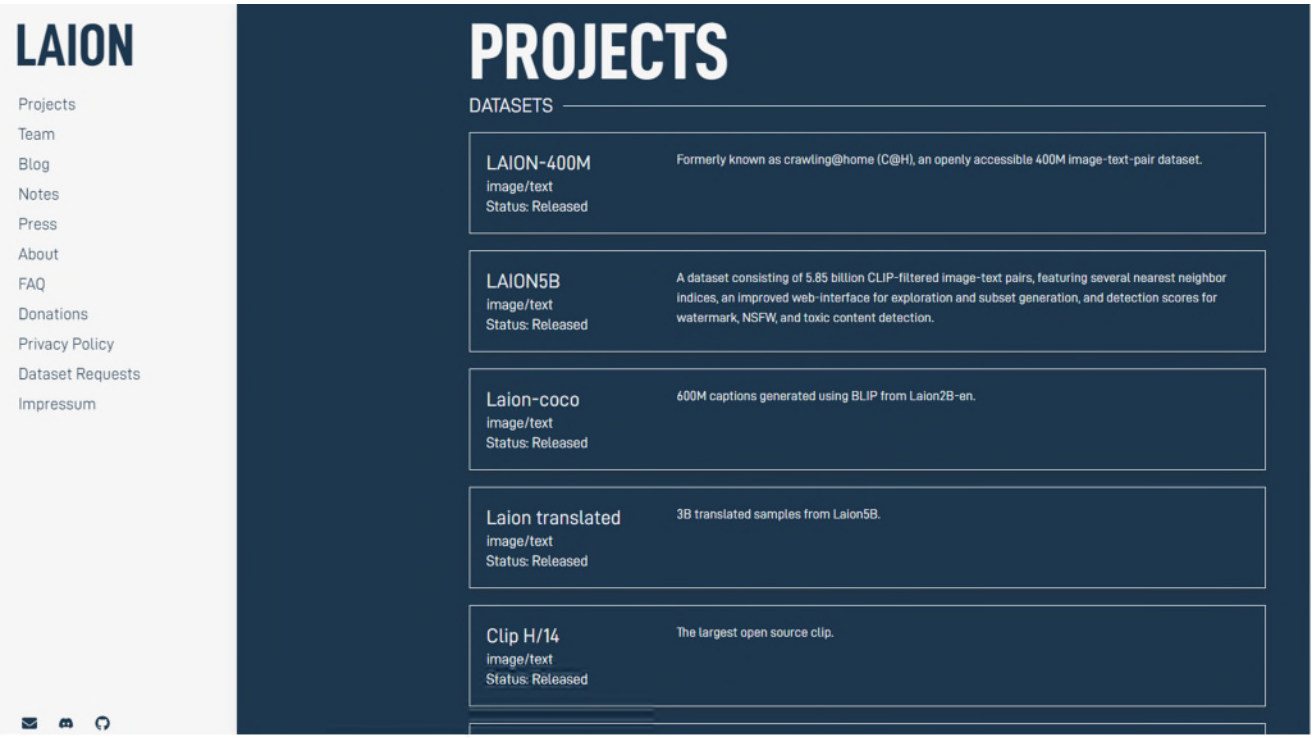


Image classification | Object detection | Visual relationship detection

Fig. 1 Example annotations in Open Images for image classification, object detection, and visual relationship detection. For image classification, positive labels (present in the image) are in green while negative labels (not present in the image) are in red. For visual relationship detection, the box with a dashed contour groups the two objects that hold a certain visual relationship.

Large-scale Artificial Intelligence Open Network (LAION)

- Contain multiple datasets for different objectives.
  - LAION5B: contain 5.85 billion CLIP-filtered image-text pairs.
  - Laion-coco: 600M captions.
  - Etc.



Many Others …

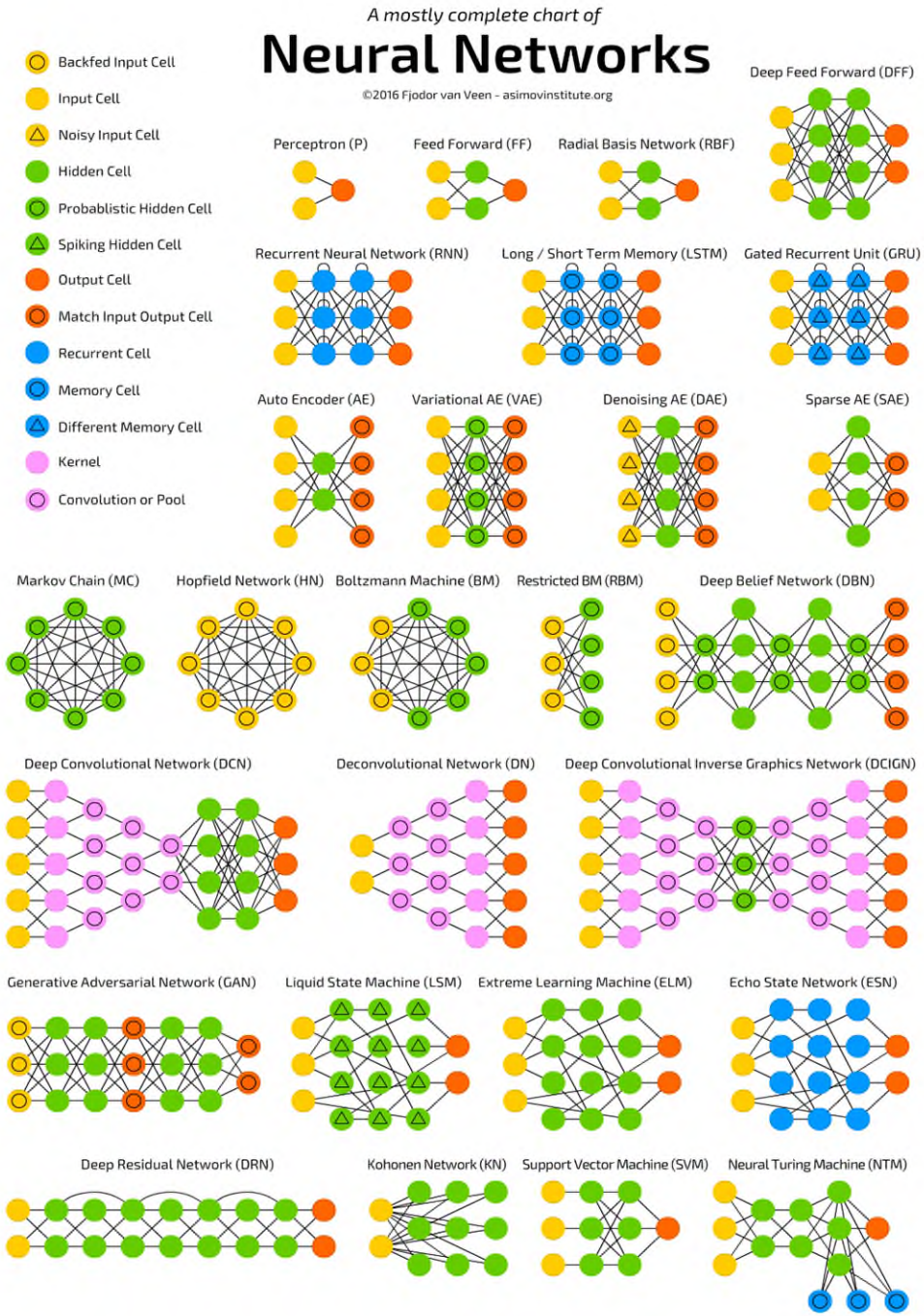
- There exists many other datasets for different domain applications:
  - Labelled Faces in the Wild (LFW) dataset for face applications.
  - Market-1501 dataset for person re-identification.
  - Etc.
- Good sources:
  - Papers with Codes, Kaggle, Github, etc.

Different Deep Neural Network (DNN) Architectures

Discussion:
DNN Models / Architectures

- Why do we need different DNN models / architectures?
- What are some common DNN models / architectures?

Neural Network Map



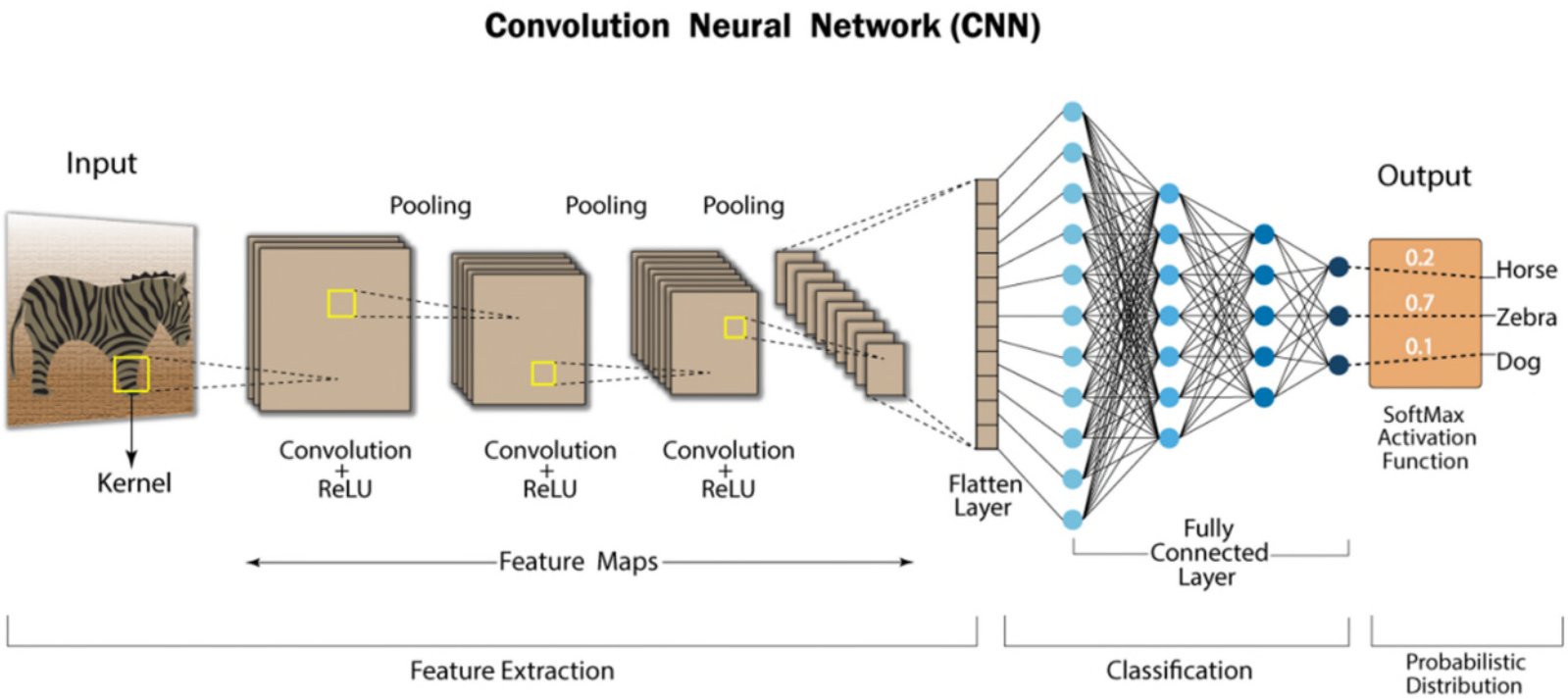
Source: “The mostly complete chart of Neural Networks, explained”, towardsdatascience.com

Common DNN Architectures

- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)
- Transformer
- Diffusion Model
- Large Language Model (LLM)
- Etc.

Convolutional Neural Networks (CNNs)

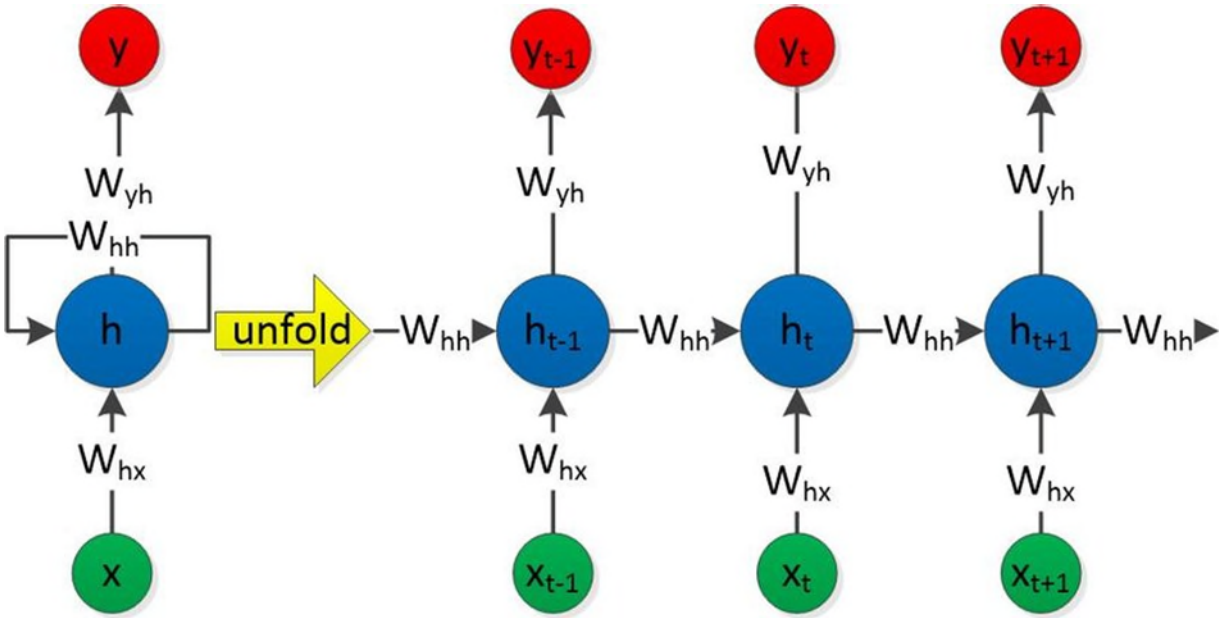
- Consist of deep layers to extract progressively higher-level abstraction features.
- Commonly used in classification and regression applications.



Source: https://developersbreach.com/convolution-neural-network-deep-learning/

Recurrent Neural Networks (RNNs)

- A type of neural network that specializes in processing sequences.
- Commonly used in applications involving time-series and state-series prediction and modelling.
- Example applications: stock price prediction, language translation.



Source: **https://stats.stackexchange.com/questions/506317/which-weight-matrix-are-shared-in-rnn-and-which-change-upon-time**

Transformer

- A type of network that uses attention mechanism to process input sequence in parallel.
- Good at modelling long-range dependency.
- Achieve state-of-the-art performance in many vision and NLP applications.

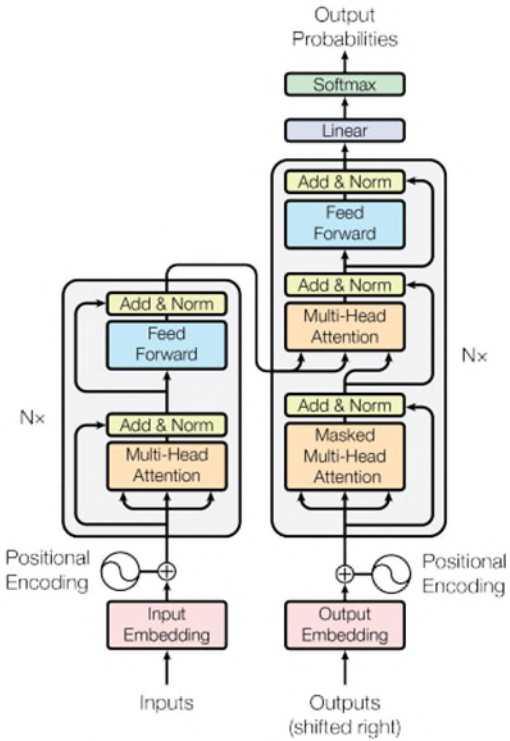
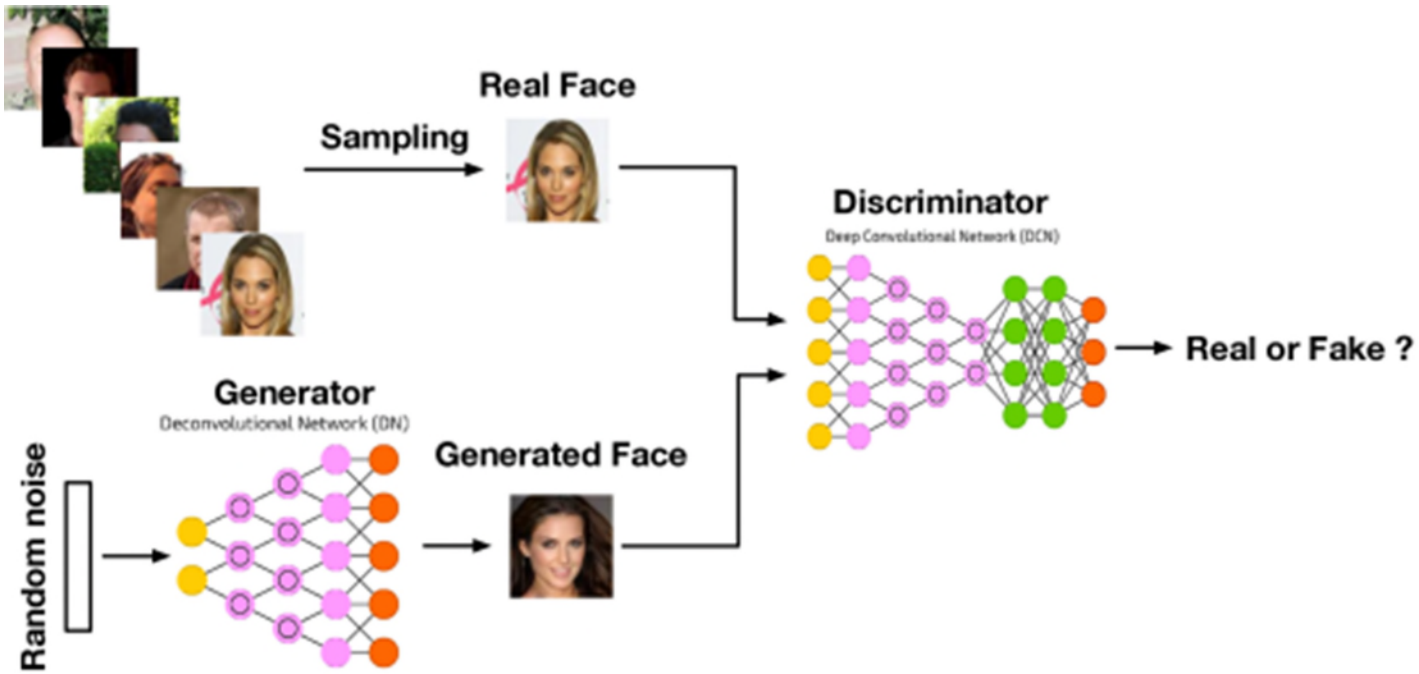


Figure 1: The Transformer - model architecture.

Generative Adversarial Networks (GANs)

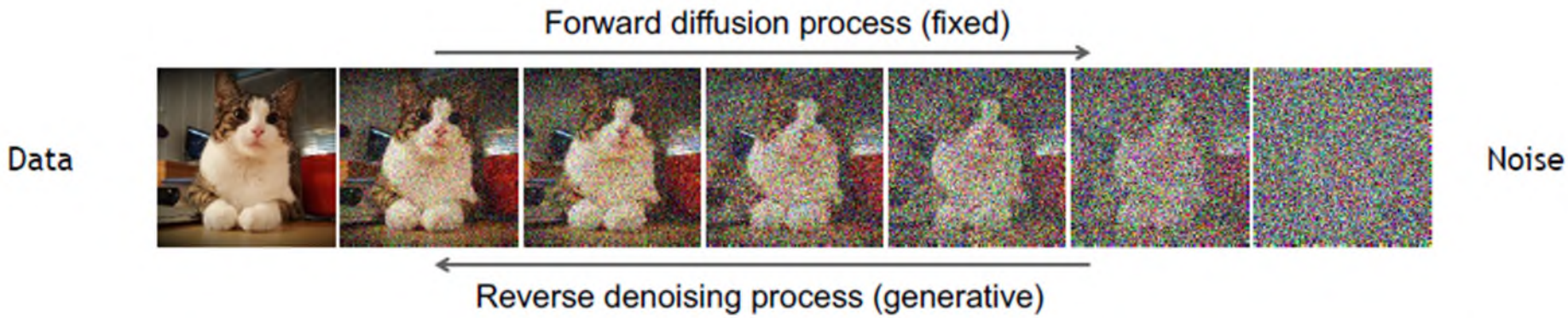
- GANs consist of two neural networks competing with each other.
- Given a training set, GANs learn to generate new data with the same statistical distribution as the training set.



Source: "Generative Adversarial Networks: A Survey and Taxonomy", Z Wang et al., 2019

Diffusion Models

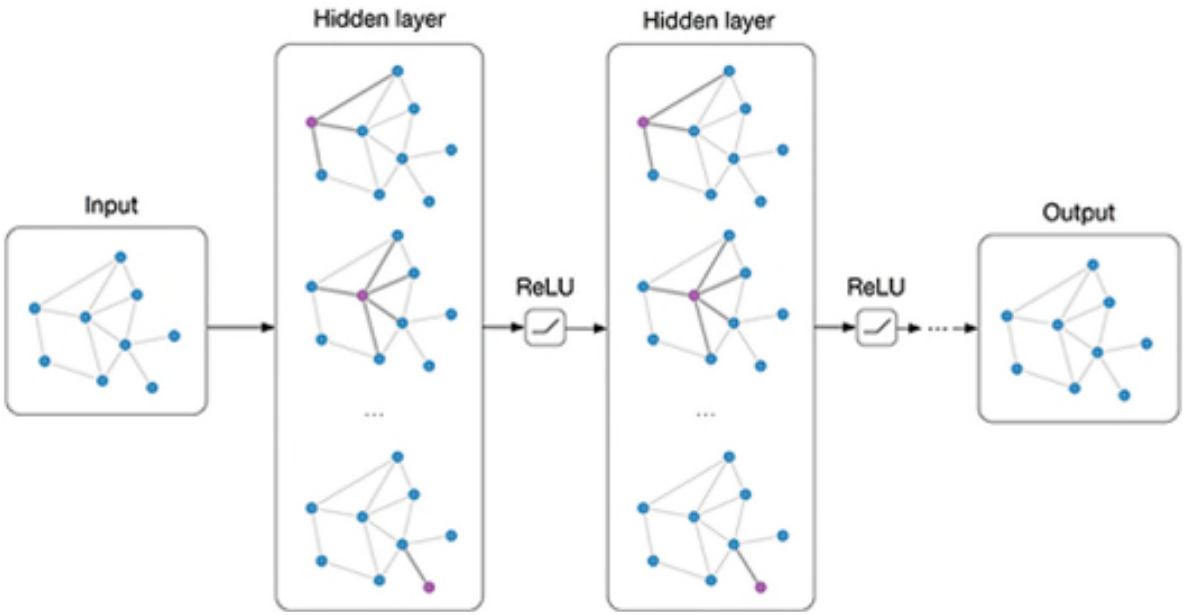
- Generative models that progressively destroy training data through the successive addition of Gaussian noise, and then learn to recover the data by reversing this noising process.



Source: https://cvpr2022-tutorial-diffusion-models.github.io/

Graph Neural Networks (GNNs)

- A type of neural networks that operates on graph modelling.
- Common applications: social network modelling, pose estimation, etc.



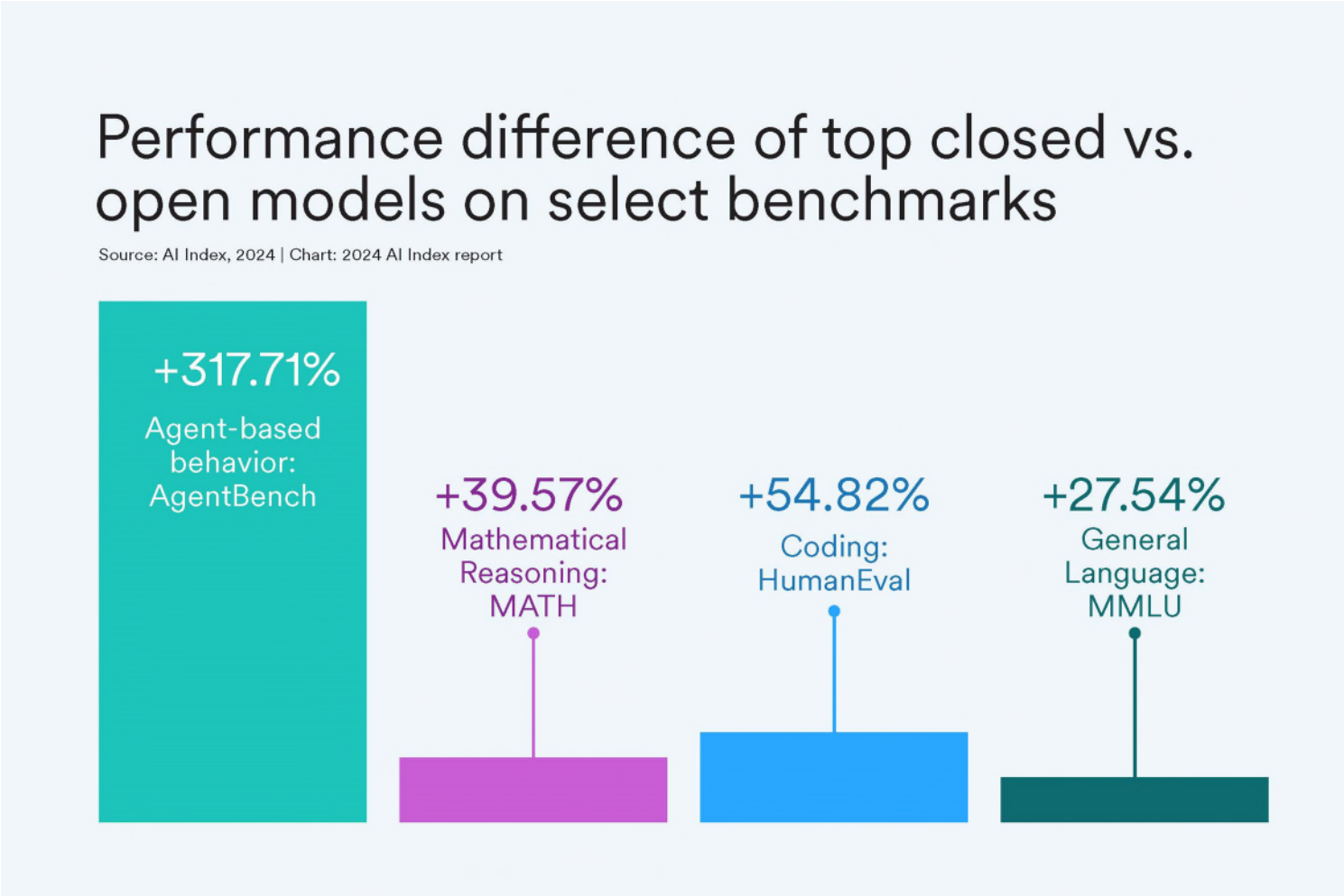
Source: “Graph Convolutional Networks”, Thomas Kipf, 2016

New / Emerging Directions

Current States of AI

Performance Comparison

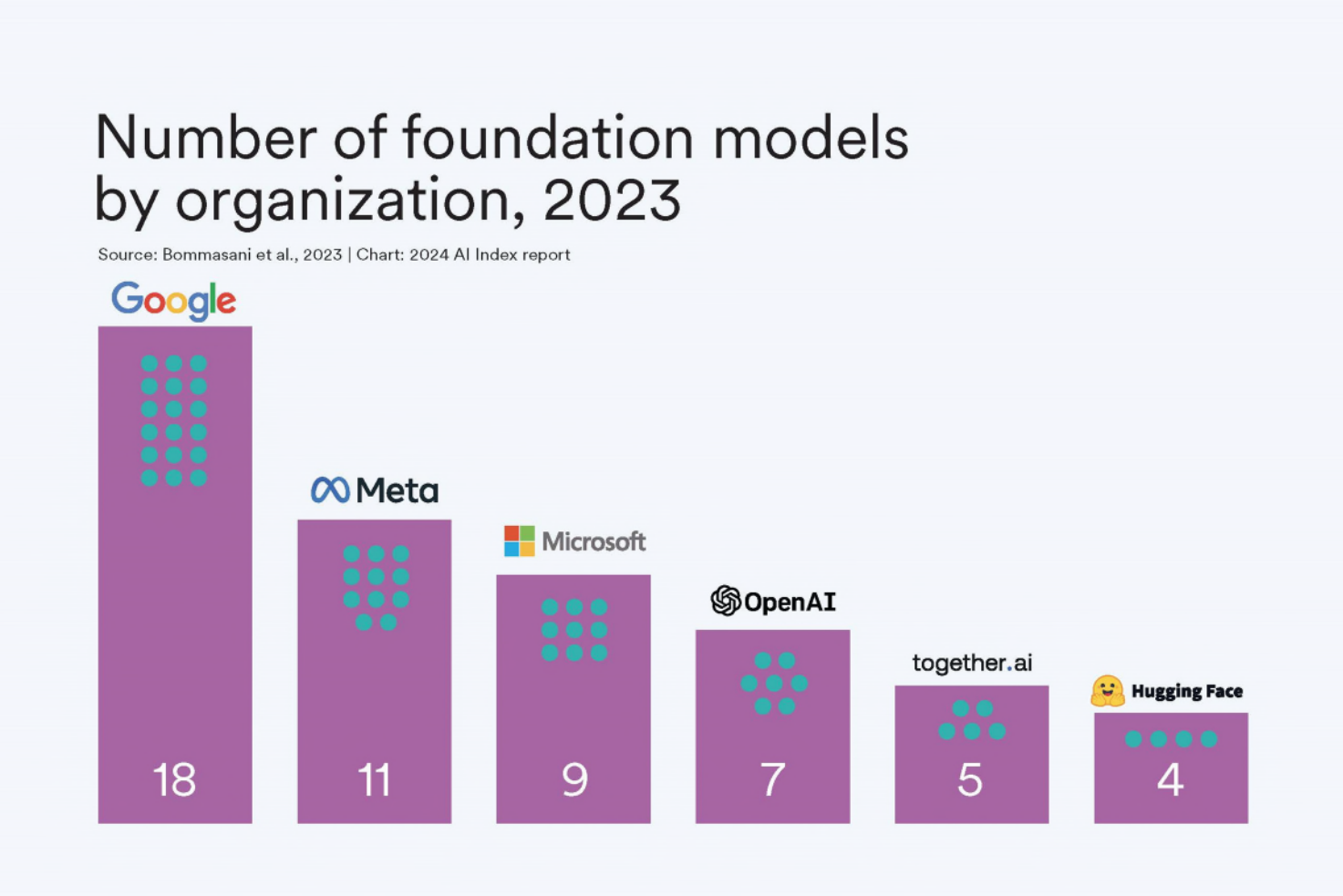
- Closed-source models outperform open-sourced models.



Source: https://hai.stanford.edu/news/ai-index-state-ai-13-charts

Key Players

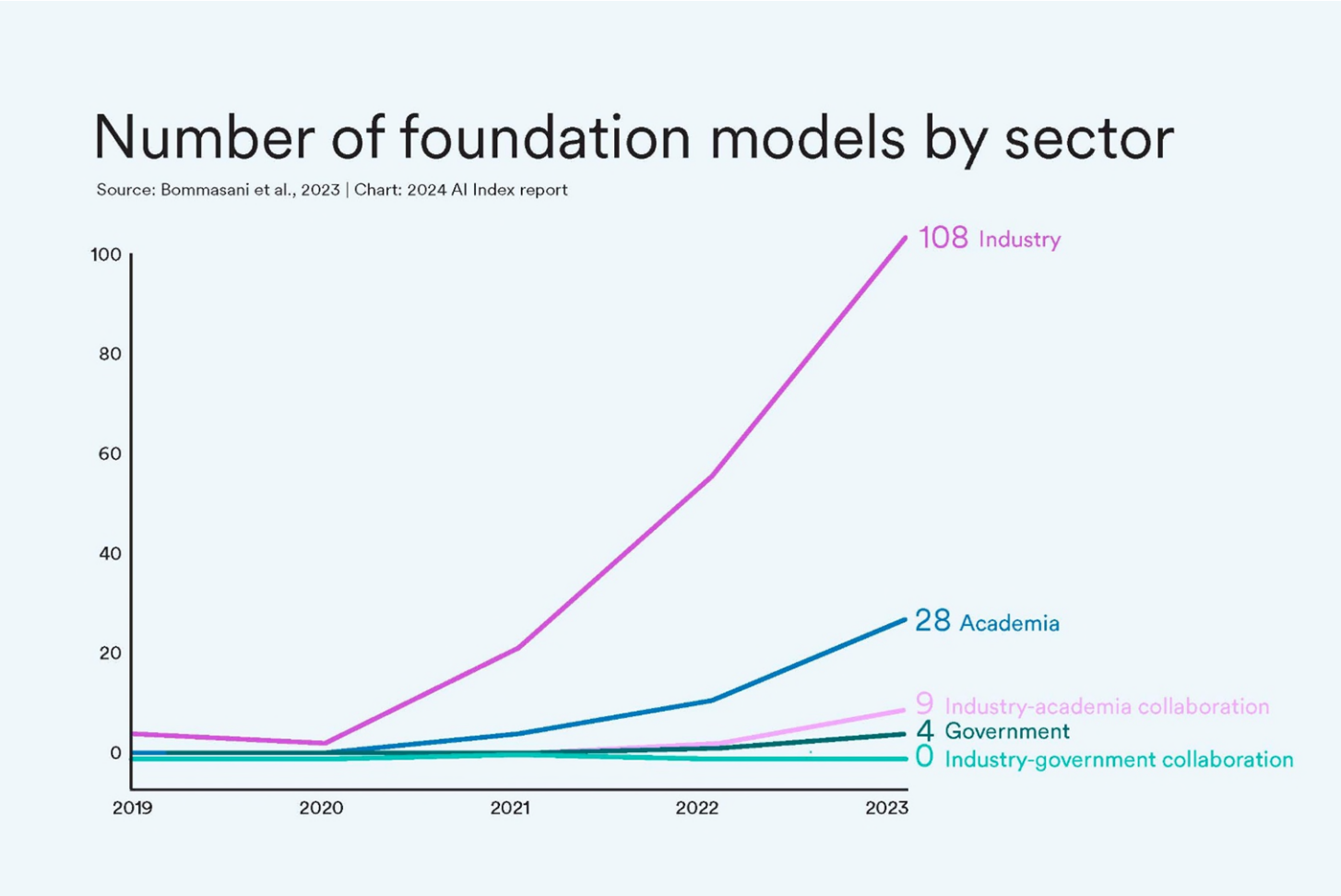
- Industry dominates Foundation Models (FMs).



Source: https://hai.stanford.edu/news/ai-index-state-ai-13-charts

Key Players

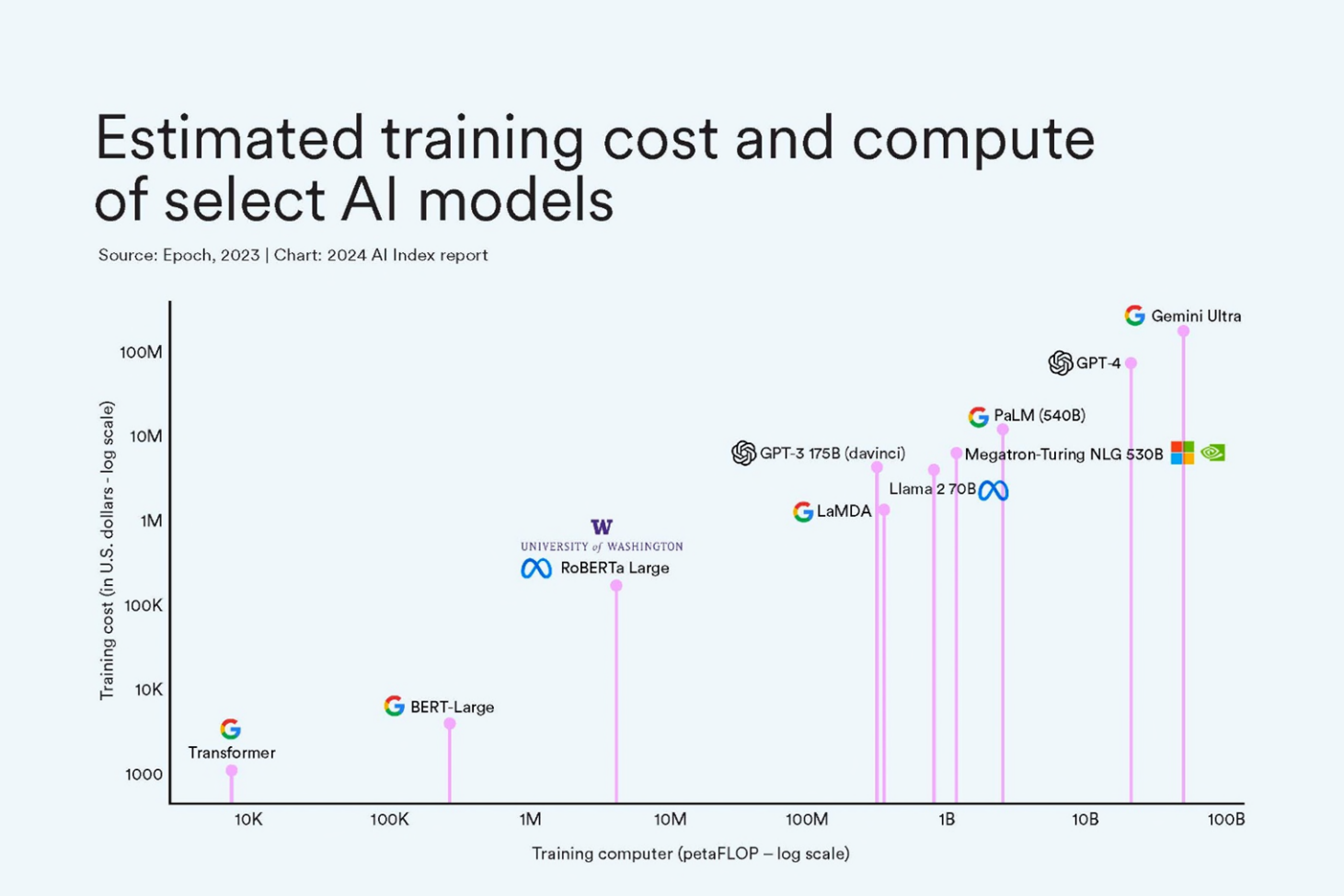
- Industry dominates Foundation Models.



Source: https://hai.stanford.edu/news/ai-index-state-ai-13-charts

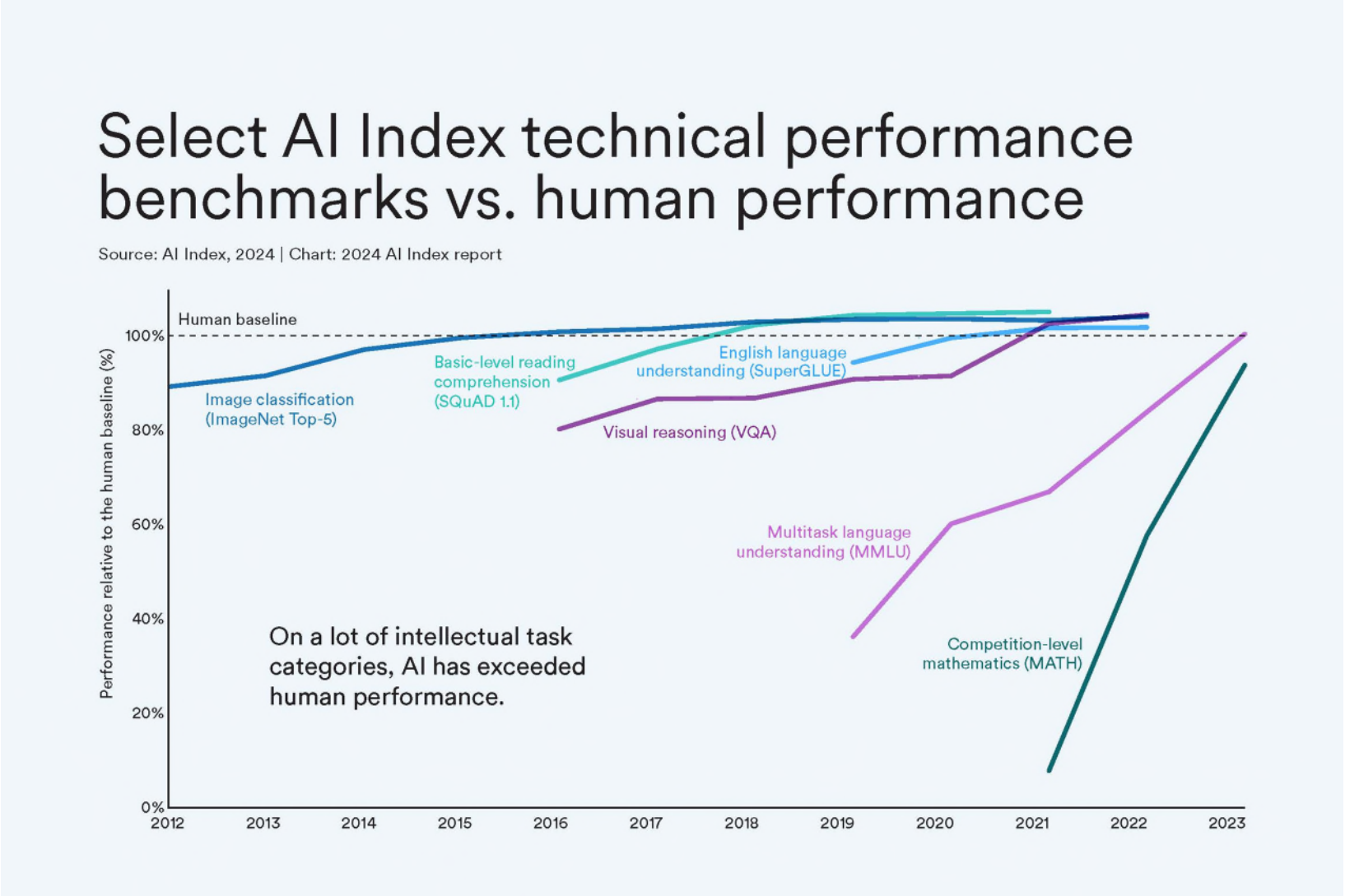
Training Cost

- High cost in training FMs.



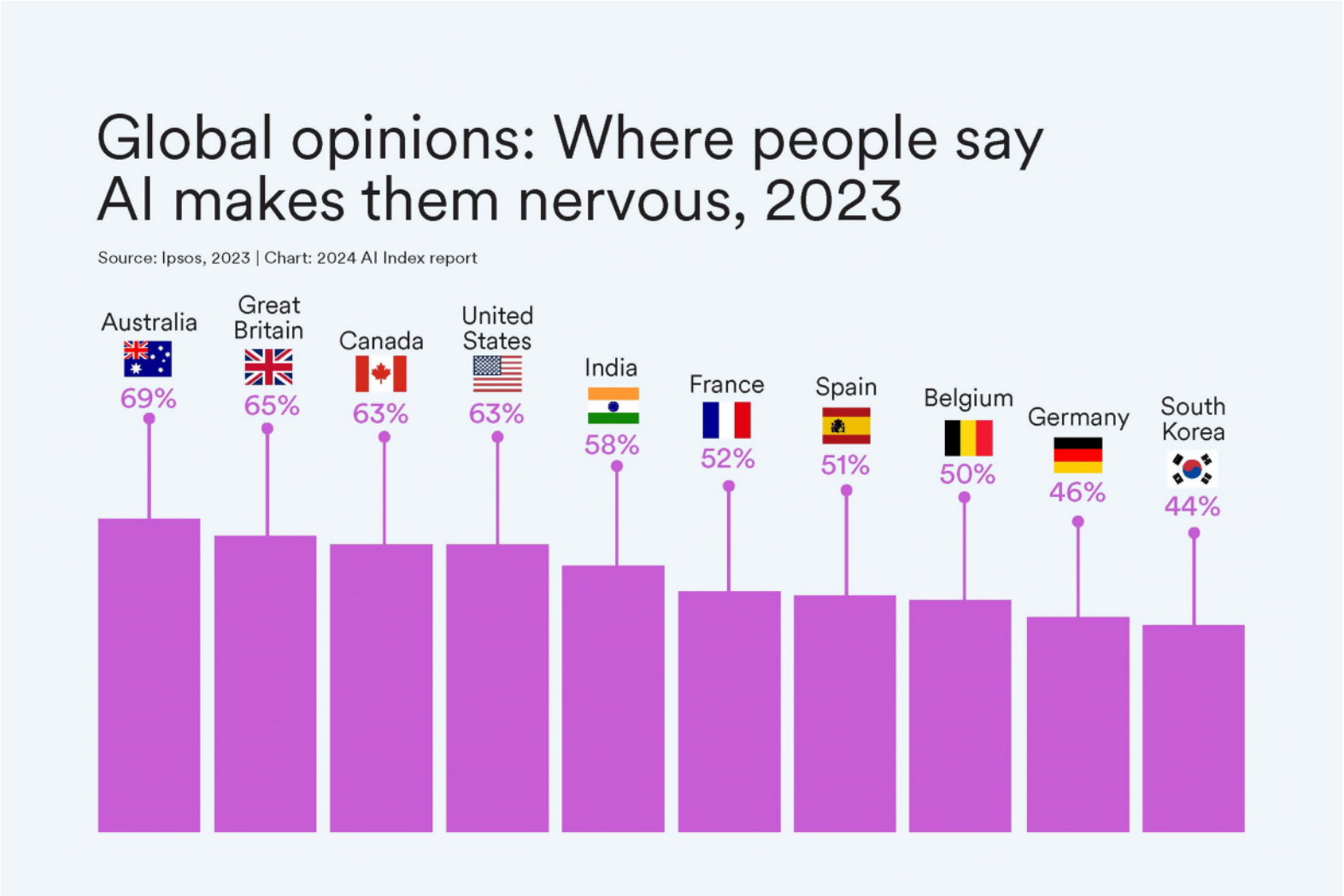
Source: https://hai.stanford.edu/news/ai-index-state-ai-13-charts

Performance: AI vs Human



Source: https://hai.stanford.edu/news/ai-index-state-ai-13-charts

Opinions on AI



Source: https://hai.stanford.edu/news/ai-index-state-ai-13-charts

Foundation Models (FMs)

What are Foundation Models (FMs)?

- Models that are trained on large-scale broad data that can be adapted (finetuned) to a wide range of downstream tasks / applications.
- Examples: Large Language Models (e.g., GPT), Vision-Language Models (e.g., CLIP), Multimodal LLM (GPT-4o, LLaVA).

Key Ideas of FMs

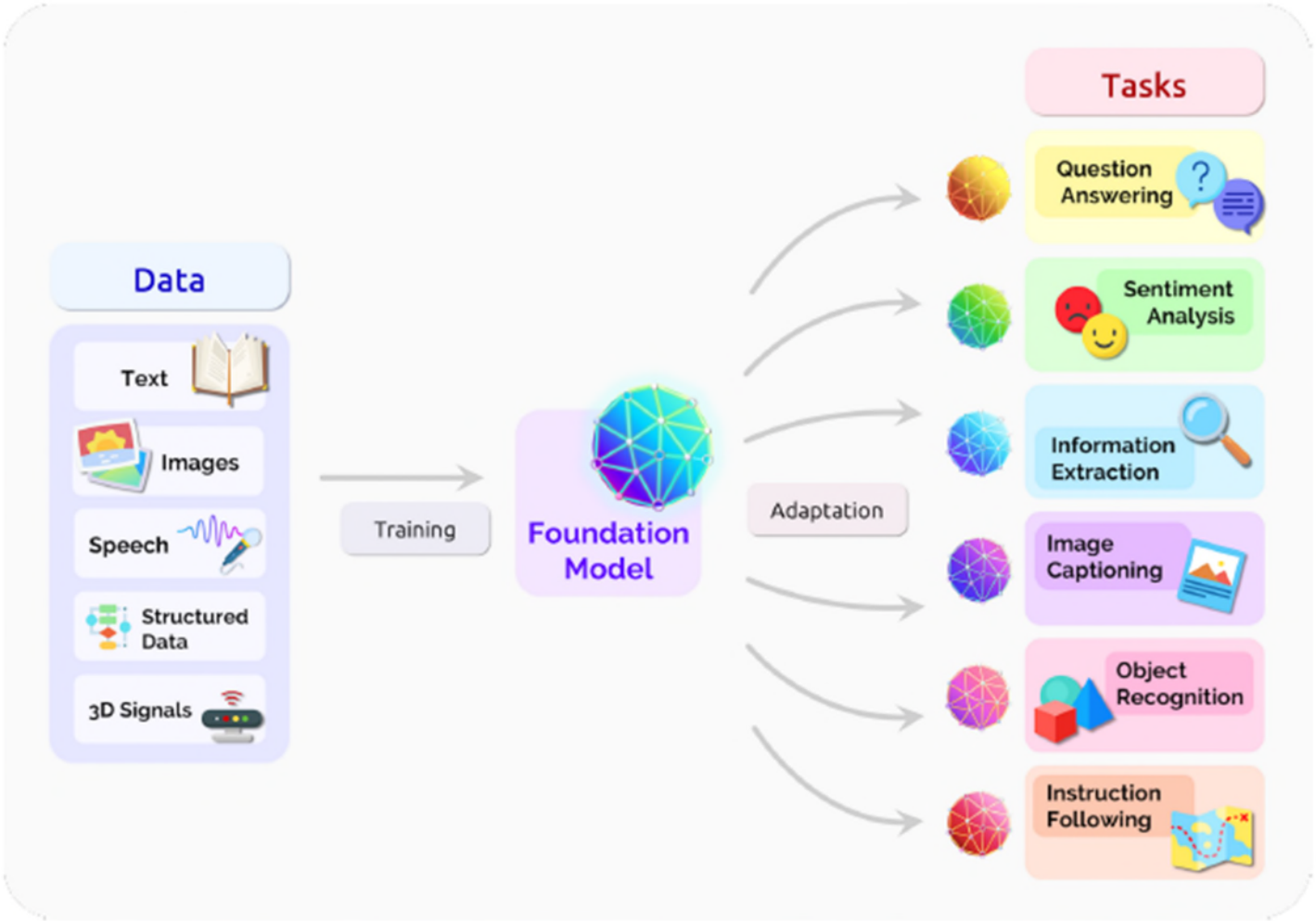


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

Source: Rishi Bommasani, On the Opportunities and Risks of Foundation Models.

Key Ideas of FMs

- A new paradigm of AI based on large-scale pre-training and downstream adaptation.
- Pretraining:
  - A model is trained on a large-scale diverse date.
  - Typically use self-supervised learning to alleviate expensive data collection and annotation.
- Adaptation / fine-tuning:
  - Pre-trained models are adapted for subsequent downstream tasks.

Different Stages of FMs

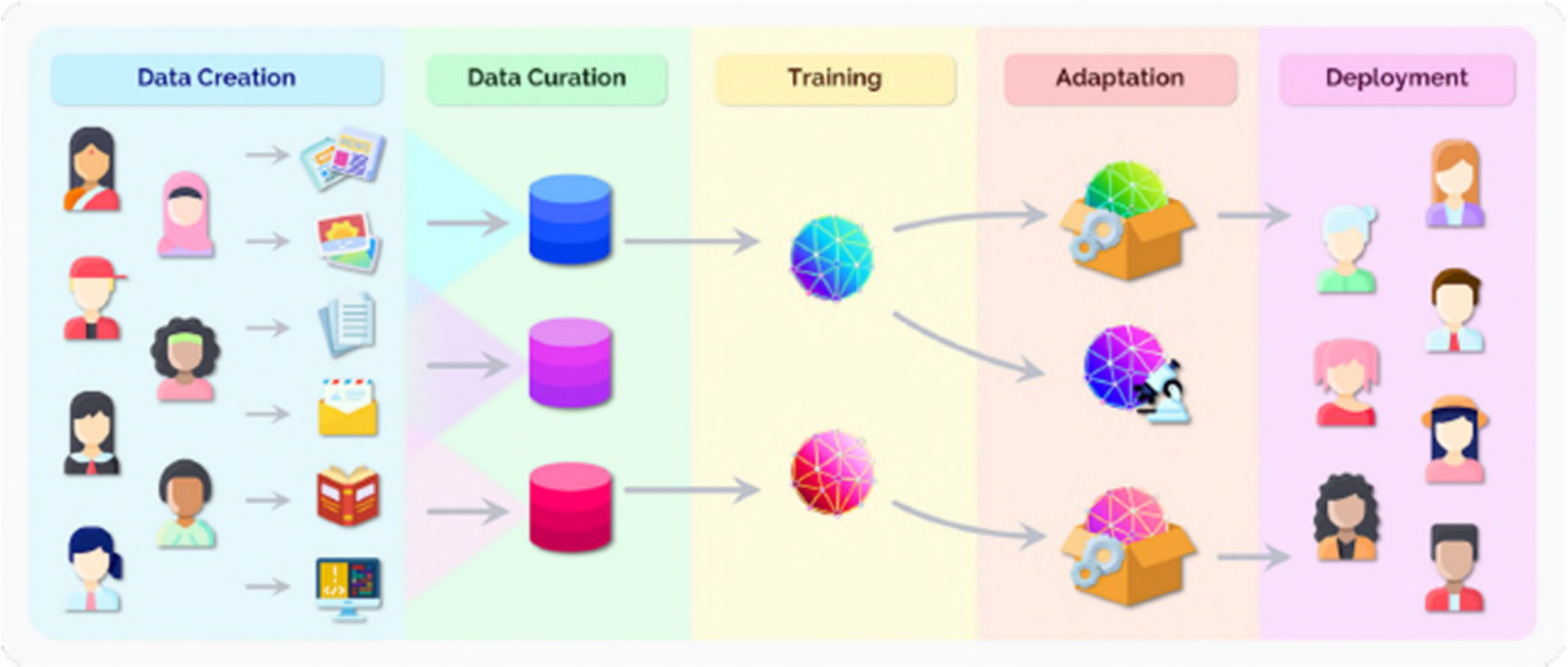


Fig. 3. Before reasoning about the social impact of foundation models, it is important to understand that they are part of a broader ecosystem that stretches from data creation to deployment. At both ends, we highlight the role of people as the ultimate source of data into training of a foundation model, but also as the downstream recipients of any benefits and harms. Thoughtful data curation and adaptation should be part of the responsible development of any AI system. Finally, note that the deployment of adapted foundation models is a decision separate from their construction, which could be for research.

Source: Rishi Bommasani, On the Opportunities and Risks of Foundation Models.

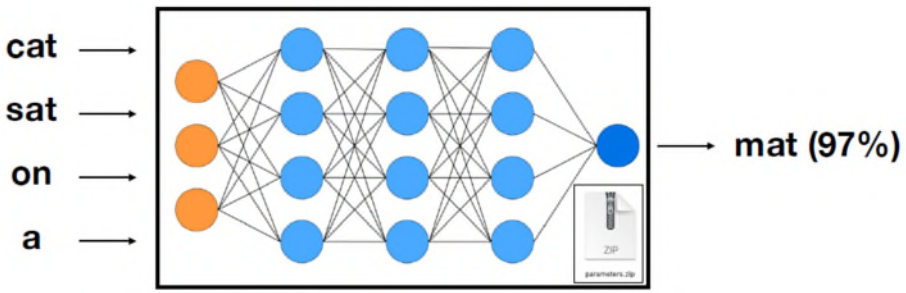
Large Language Models (LLMs)

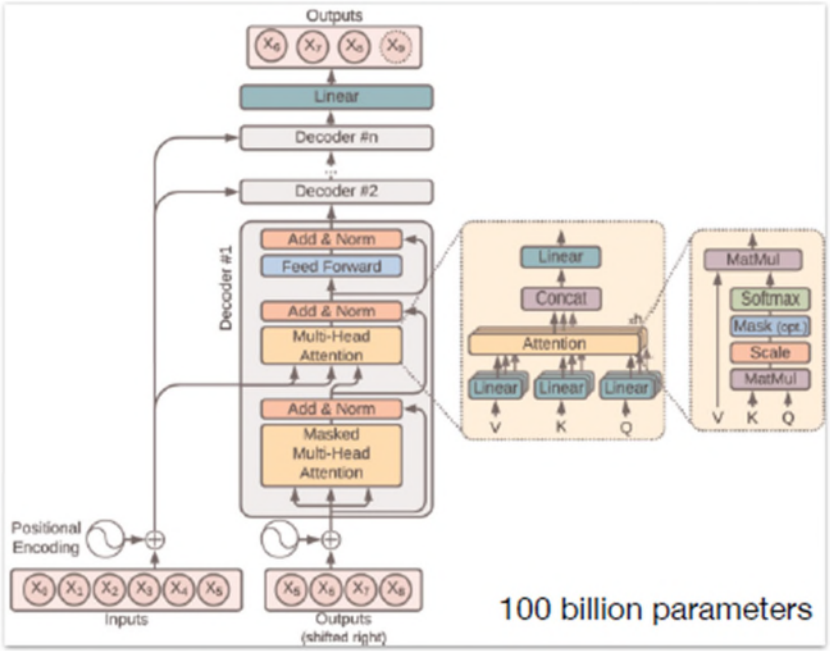
Large Language Models (LLMs)

Chunk of the internet,
~10TB of text

6,000 GPUs for 12 days, ~$2M
~1e24 FLOPS

~140GB file





Source: Intro to Large Language Models, Andrej Karpathy, Nov 2023

GPT-4o

- Developer: OpenAI
- Parameters: More than 175 billion (likely trillions)
- Access: API
- Released in May 2024.
- Previous family iterations: GPT-1, GPT-2, GPT-3, GPT-3.5
- Extend multimodal capabilities of GPT-4 Turbo by integrating text, image and audio prompts.
- Process audio in real time, and output realistic, tone appropriate response in human voice.

Llama 3

- Developer: Meta (parent company of Facebook & Instagram)
- Parameters: 8 billion, 70 billion, and 400 billion (unreleased)
- Access: Open
- Released in Apr 2024.
- Previous family iterations: Llama, Llama 2.
- Can match the performance of GPT-4, but at lower cost.
- Able to install on a local system rather than relying solely on the cloud, 8B version of LLaMA 3 is small enough to run on a laptop.
- Alleviate privacy concern of sending data into the cloud for processing.
- Ideal for AI researchers due to its performance, adaptability, and open-source license.

Gemini

- Developer: Google
- Parameters: Nano (1.8 billion and 3.25 billion); others unknown
- Access: API
- Released in Apr 2024.
- Formerly known as Bard.
- Three models: Gemini Nano, Gemini Pro, and Gemini Ultra, designed for different devices, from smartphones to servers.
- Can handle images, audio, video, code, and other kinds of information.

GPT-4o Voice & Vision

GPT-4o

Source: openai.com

GPT-4o Math Problem Solving

GPT-4o

Source: openai.com

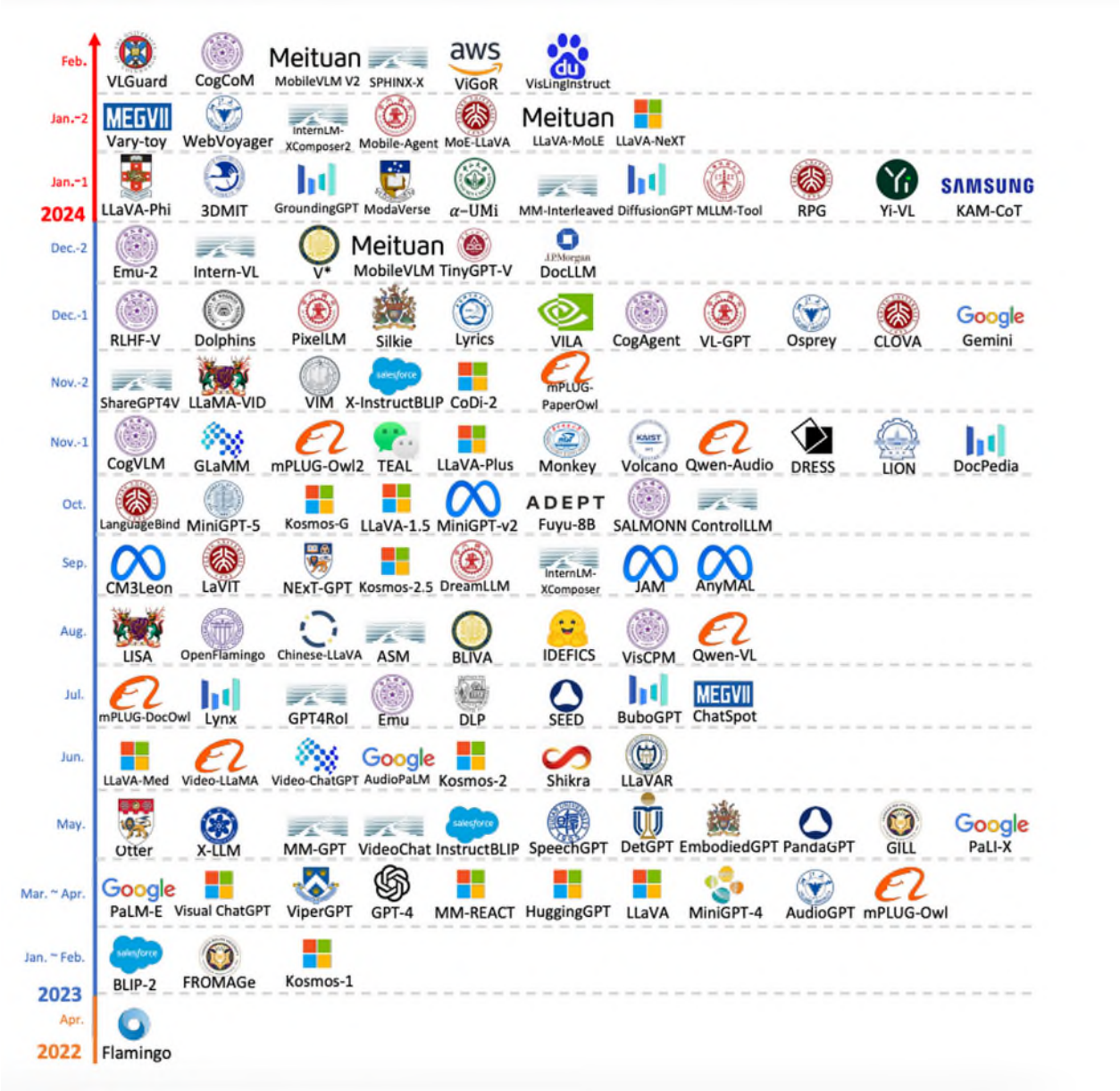
GPT-4o Coding Assistant

GPT-4o

Source: openai.com

Multimodal LLM (MLLM)

Emerging Trend: Multimodal LLM (MLLM)



What is Multimodal LLM (MLLM)?

- A model capable of understanding and generating different data modalities, such as text, images, audio, video, etc.
- Leverage multimodal inputs and outputs
  - Multimodal Input: process different data modalities (e.g., text and images).
  - Multimodal Output: generate outputs in multiple formats (e.g., text generation, image creation, etc.).
- Cross-modal Understanding
- Training with Diverse Data
- Improved Contextual Understanding

Summary

- This lecture covers the following:
  - Course Overview
  - Introduction to Artificial Intelligence (AI)
  - Deep Neural Network (DNN) Architectures
  - New / Emerging Directions