# Course 5: Foundation Models

# Summary

**Last session**

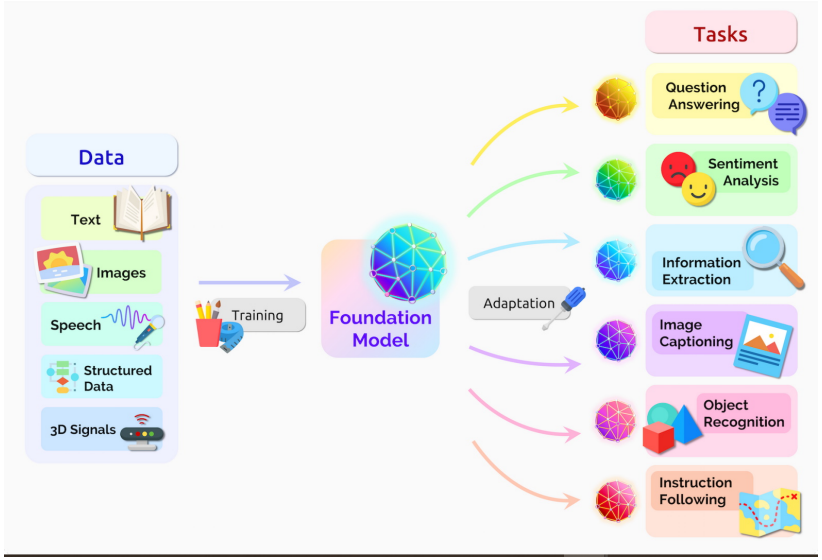
1. Deep Learning Basics
2. Convolutional Neural Networks
3. Transformers

**Today's session**

- What is a Foundation Model
- Self Supervised Learning
- Some examples of Foundation Models

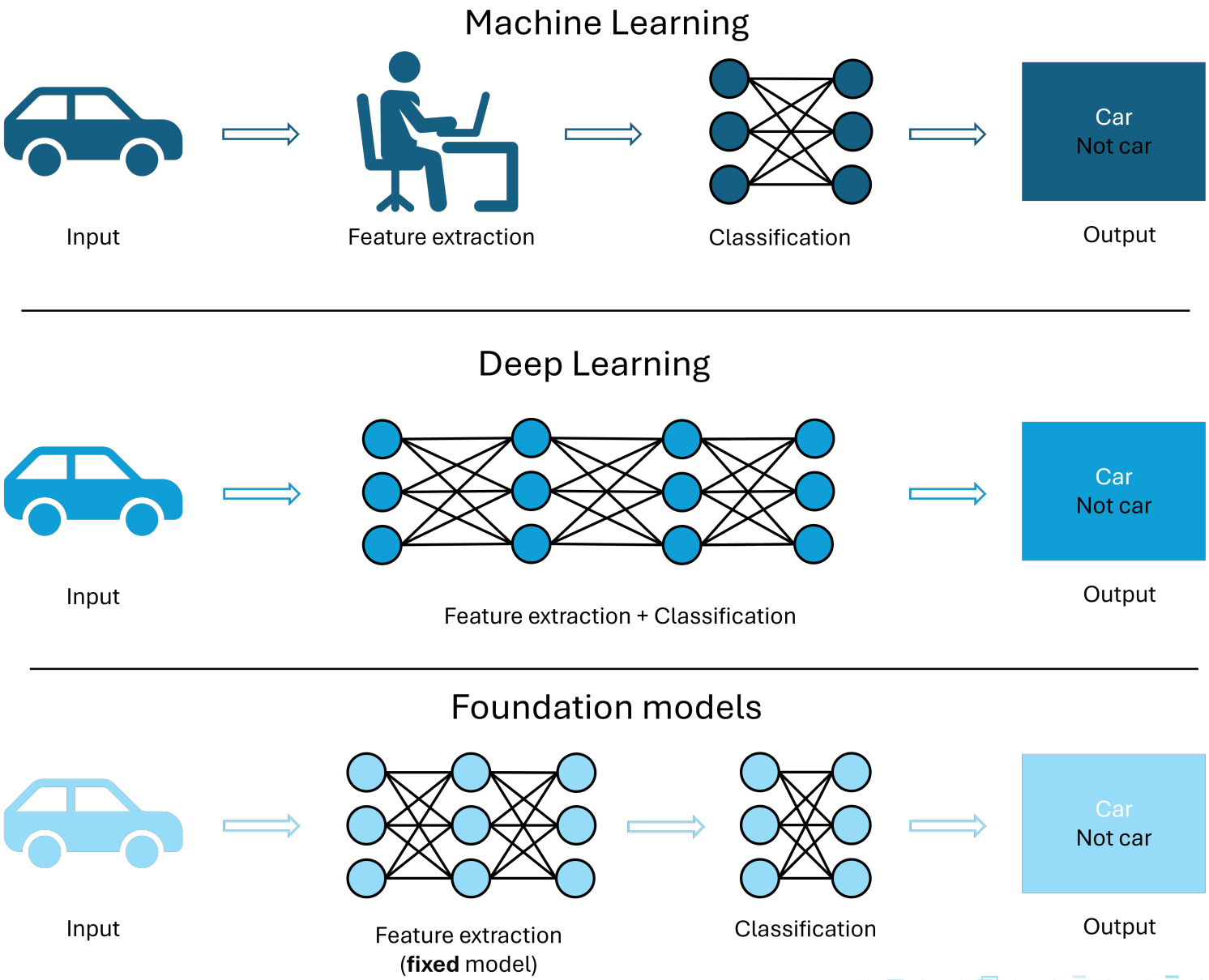
# What is a Foundation Model?

- Trained on internet-scale datasets
- Training task si not straightforward (SSL, pretext tasks)
- Generic feature extractors, Multipurpose
- Generalization is not a problem anymore! All is about **particularization**



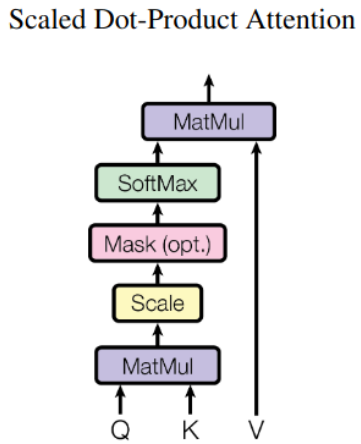

https://blogs.nvidia.com/blog/what-are-foundation-models/

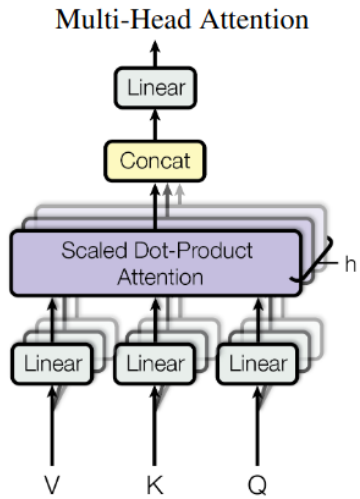
# What is a Foundation Model?



# Transformers

- No Inductive Bias (as for Convolutions)
- Best generalization in many domains
  - comparable with convolutions for Images
  - State-of-the-art for Natural Language Processing
- Few Hyperparameters
- Very large, require large datasets for training



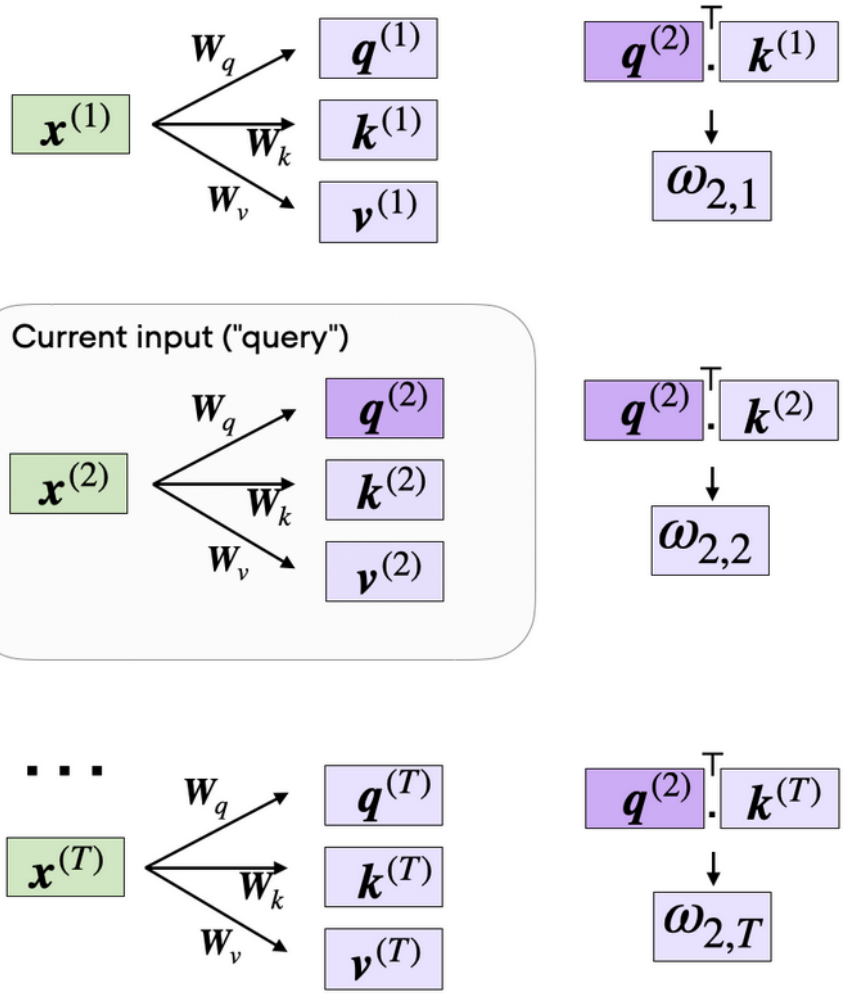

*Attention is all you need:* `https://arxiv.org/pdf/1706.03762`

`https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html`

## Self-attention

Mechanism that enhances input embedding by including information about the input context (*e.g.*, the meaning of a word changes based in its context)

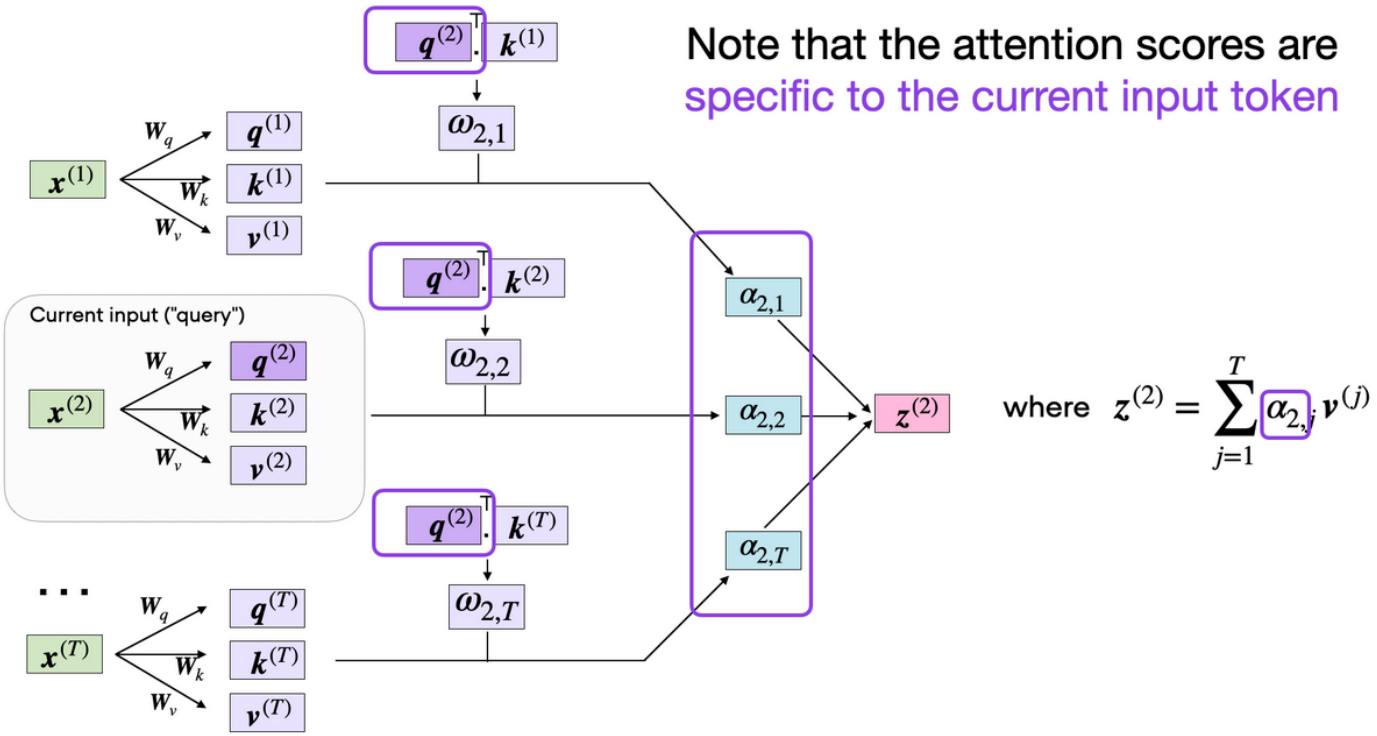
- inputs are projected by weight matrices $\mathbf{W_i}$ into **q**, **k** and **v** vectors
- attention weights $\omega_{i,j}$ are obtained by dot product (*e.g.*, similarity) of query and keys




https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

## Self-attention

- self-attention scores $\alpha_{i,j}$ are normalized version of weights $\omega_{i,j}$
- the output $\mathbf{z_i}$ is an attention-weighted version of the original query input $\mathbf{x_i}$ with respect to all other elements of the input



# Transformers

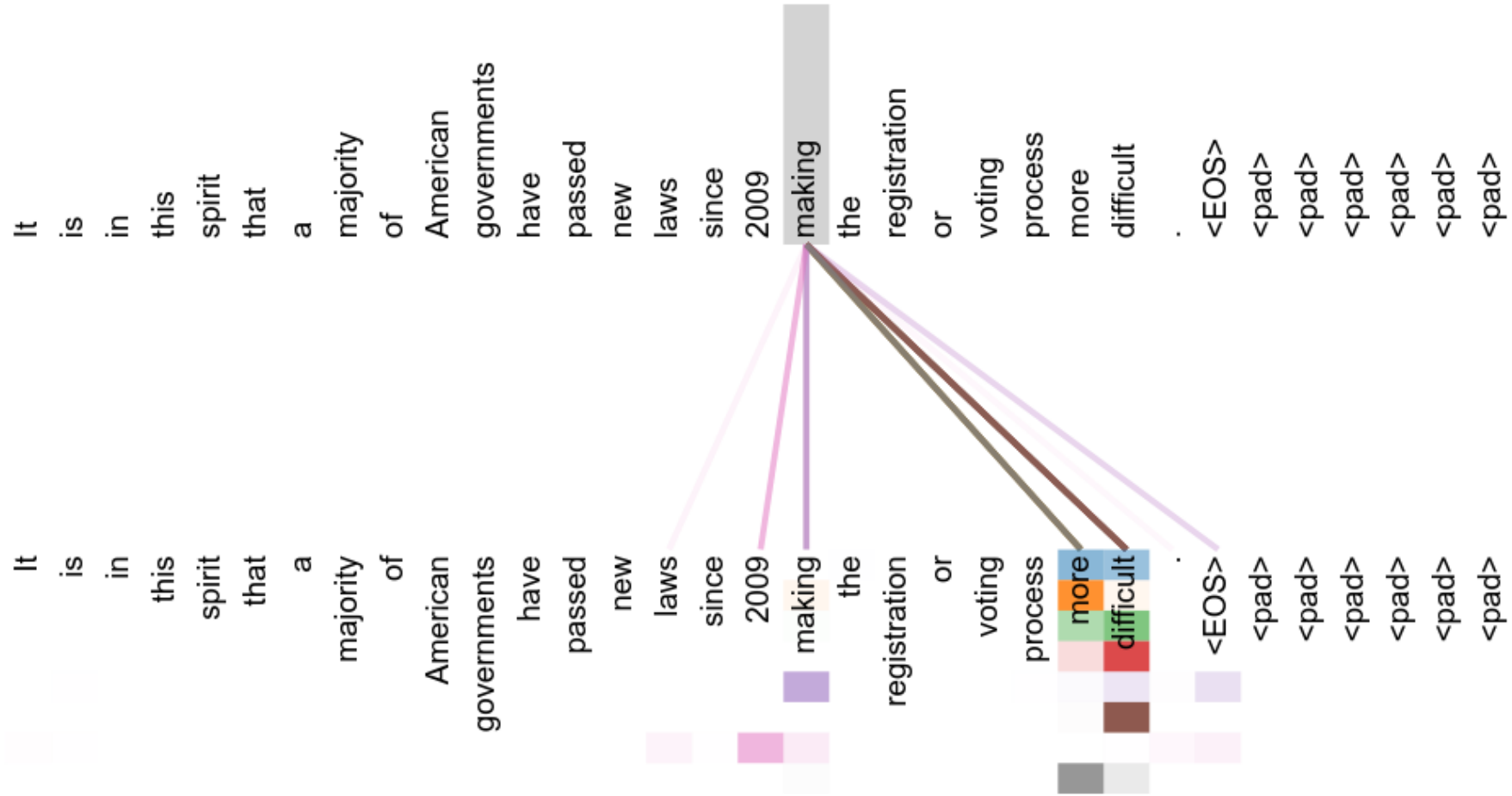



Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.

# Self Supervised Learning

## The principle

Learning useful representations from data without relying on labels. The model *creates* the labels based on the structure of the data through *pretext tasks*

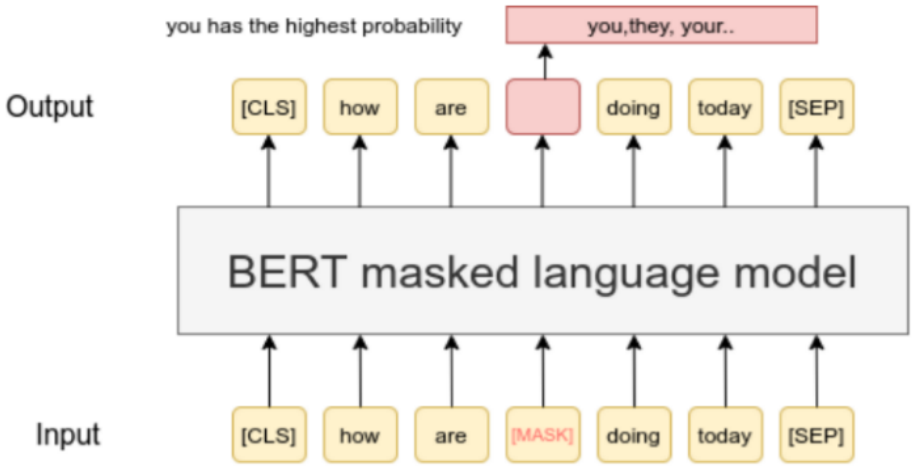
## Why it is important

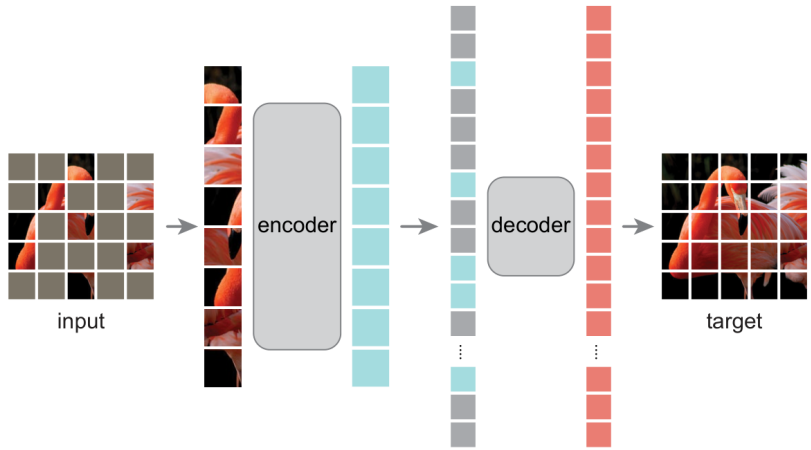
- **Scalability**: Since unlabeled data is much easier to obtain, models can scale to larger datasets and more diverse inputs
- **Robust Representations –> Transferability**: It often leads to more generalizable representations that perform well across multiple tasks compared to models trained on a more limited label space

# Self Supervised Learning

## Types of SSL approaches

- **Masked Input Modeling**: Predicting missing part of the input



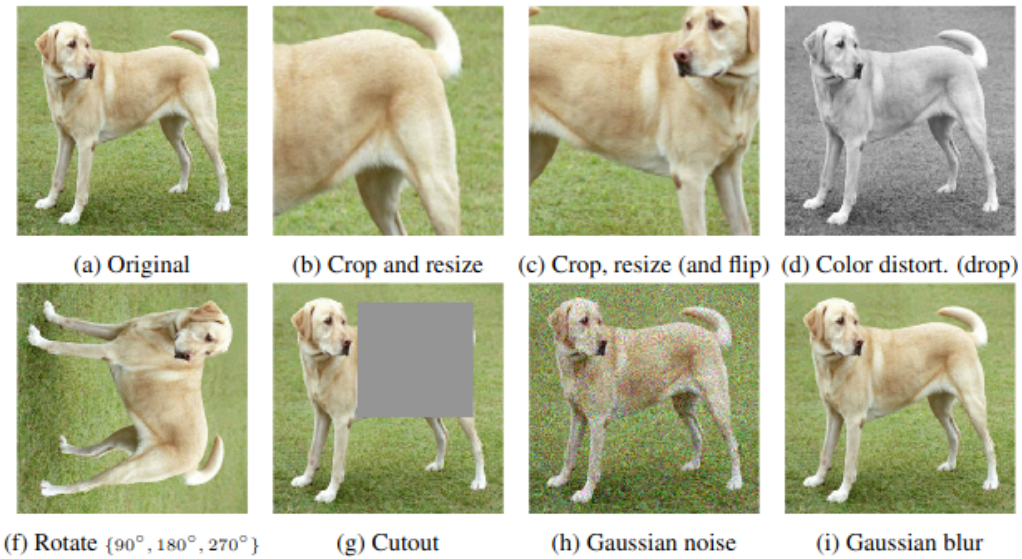

*BERT:* `https://arxiv.org/pdf/1810.04805`

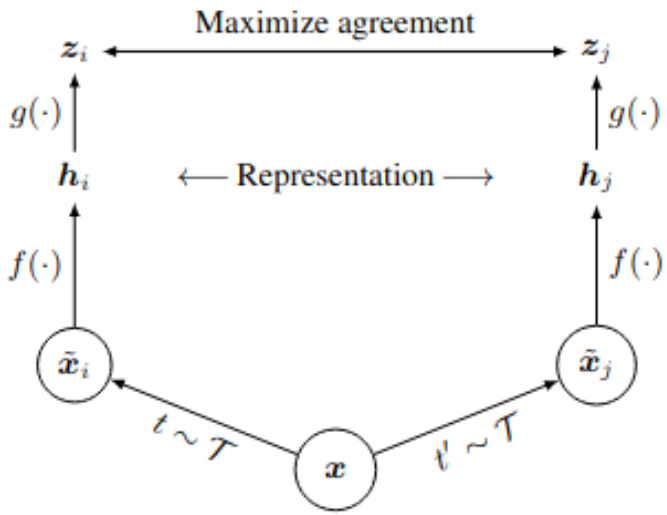
*Masked Autoencoders:* `https://arxiv.org/pdf/2111.06377`

# Self Supervised Learning

## Types of SSL approaches

- **Masked Input Modeling**: Predicting missing part of the input
- **Contrastive Learning**: Pulling together similar representations and pushing apart dissimilar ones

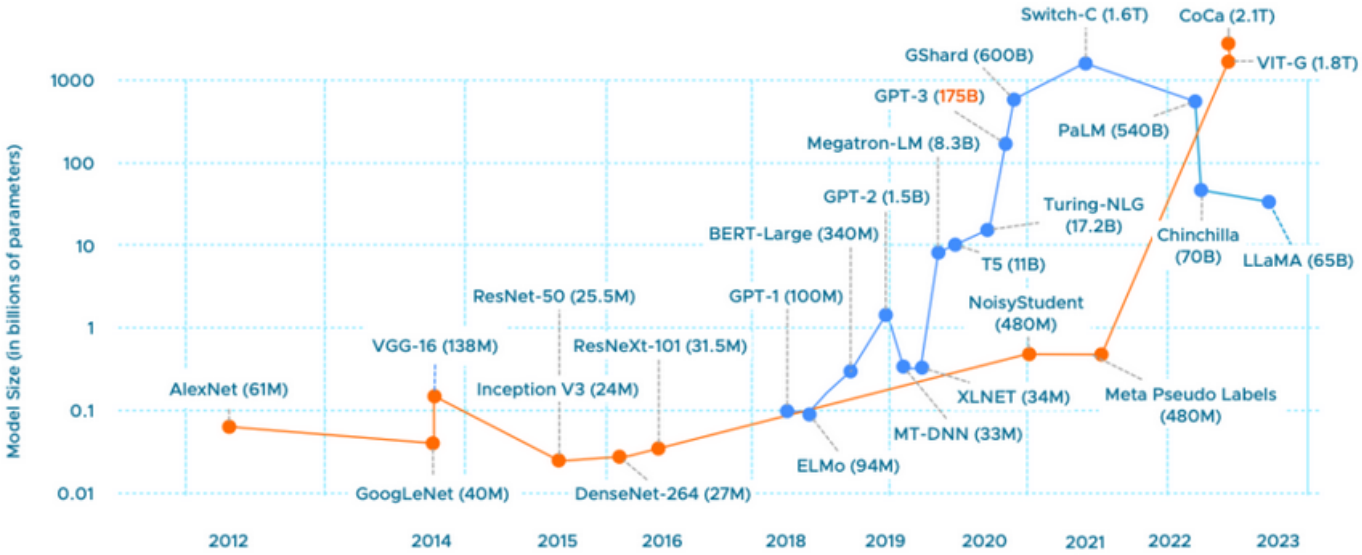



*SimCLR:* `https://arxiv.org/pdf/2002.05709`, *BYOL:* `https://arxiv.org/pdf/2006.07733`

# How Big Are We Now?

- Increasing model size is a proxy for increasing performance
- Data is as important as scaling model size! For 2x model size, data should also be 2x
- Datasets are growing very rapidly (0.1-0.2 orders of magnitude per year before 2015, now higher)




Tutorial on data efficiency: `https://baharanm.github.io/assets/pdf/ICML24_tutorial_DataEfficient.pdf`
Training Compute-Optimal Large Language Models: `https://arxiv.org/pdf/2203.15556`
Scaling Laws for Neural Language Models: `https://arxiv.org/pdf/2001.08361`

# Some examples: Foundation Models for Text

| Model | Provider | Open | Release | Params | Data |
|---|---|---|---|---|---|
| BERT | Google | Yes | Oct-2018 | 345M | 3.3M |
| PaLM | Google | No | Apr-2022 | 540B | 780B |
| GPT-4 | OpenAI | No | Mar-2023 | - | 300B |
| GPT-3.5 | OpenAI | No | Mar-2023 | 175B | - |
| Llama1 | Meta AI | Yes | Feb-2023 | 7/13/33/65B | 1 T/1.4T |
| Llama2 | Meta AI | Yes | Jul-2023 | 7/13/33/70B | 2T |
| Mixtral | MistralAI | Yes | Dec-2023 | 8 x 7B | - |
| Phi-2 | Microsoft | No | Dec-2023 | 2.7B | - |

## Large Language Models

- Transformers trained (SSL) on predicting masked *tokens* and next sentence
- Finetuning (Supervised) on instruction data (prompt + response)
- Reinforcement Learning with Human Feedback

Tutorial on foundation models for text:
`https://docs.google.com/presentation/d/1nfV9QiNV2tbHsw9GeP7e96az-oL_C_rymTSE6V6zBos/edit?usp=sharing`
Llama Paper: `https://arxiv.org/abs/2302.13971`

# Some examples: Foundation Models for Vision

## DINOv2

- Vision transformers trained (SSL) on large curated data
- Produces all-purpose visual features (classification, segmentation, depth estimation..)

## SAM

- Image and text Encoders (SSL pretrained Transformers) trained on 1B masks and 1M images
- Segment any object, promptable segmentation



*DINOv2:* `https://arxiv.org/abs/2304.07193`, *SAM:* `https://segment-anything.com/`

# Some examples: Multimodal Foundation Models

NEXT WEEK...

# Customize a Foundation Model

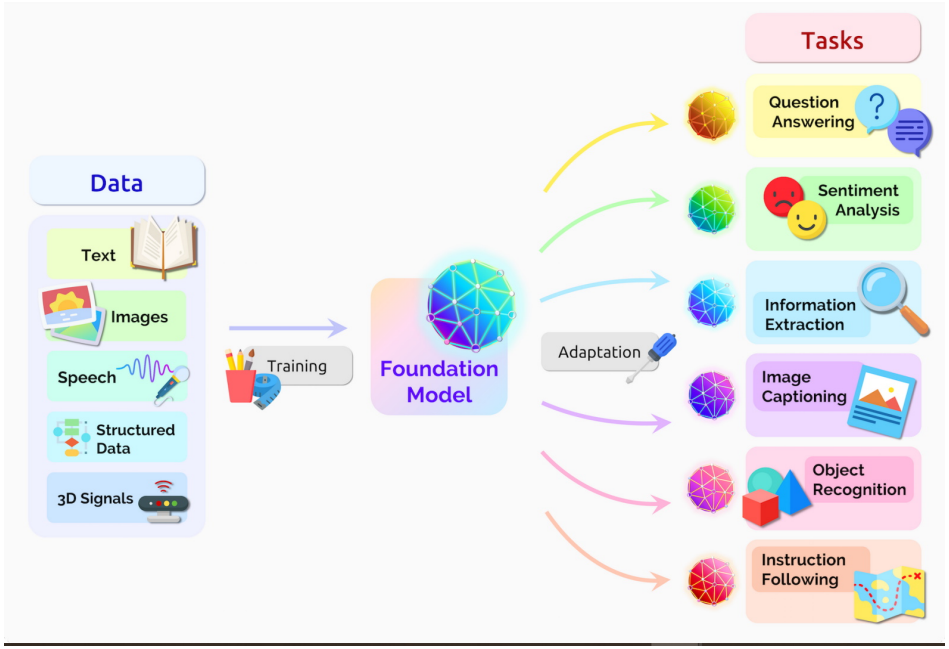
- Foundation models are often fine-tuned after pre-training to adapt them to a specific task
- The challenge becomes particularization (as opposed to generalization)
- Different techniques depending on the objective of the adaptation:
  1. Add specific information to the model knowledge (e.g., Retrieval augmented generation RAG)
  2. Change *behavior* of the model with fine-tuning (*e.g.*, parameter efficient fine tuning with LORA)
  3. Improve Foundation Model safety and helpfulness (*e.g.*, Reinforcement Learning based un Human Feedback for LLMs)

---

RAG: `https://blogs.nvidia.com/blog/what-are-foundation-models/`
LoRA Paper: https://arxiv.org/pdf/2106.09685.pdf
Various tutorials on RAG and fine-tuning: `https://github.com/facebookresearch/llama-recipes/tree/main`

# Foundation Models

- Model trained on an Internet scale dataset
- Self-Supervised training: pretext tasks
- Particularization vs Generalization




Source: `https://blogs.nvidia.com/blog/what-are-foundation-models/`

# Lab Session 5

Generate the embeddings you worked with in Lab 2 and 3 using pretrained Foundation Models

- Load the specific modality dataset
- Preprocess the raw data if needed
- Use Hugging Face to load the pretrained Foundation Model
- Generate the embeddings and test them