Interdisciplinary Computational Biology Workshop 2025

Day 2: An Introduction to AI and ML

Prof Amin Ahsan Ali
Dept of CSE, Independent University, Bangladesh
Director AI & ML wing,
Center for Computational and data Sciences (CCDS)

What we will cover in this part

- What are some popular AI systems?
  - Focusing on LLMs and their capabilities
- What is ML?
  - Overview of some popular ML models?
- Overview of ML models in Computational Biology - Protein Function Prediction task
- Overview of a typical ML pipeline

Some Popular AI System

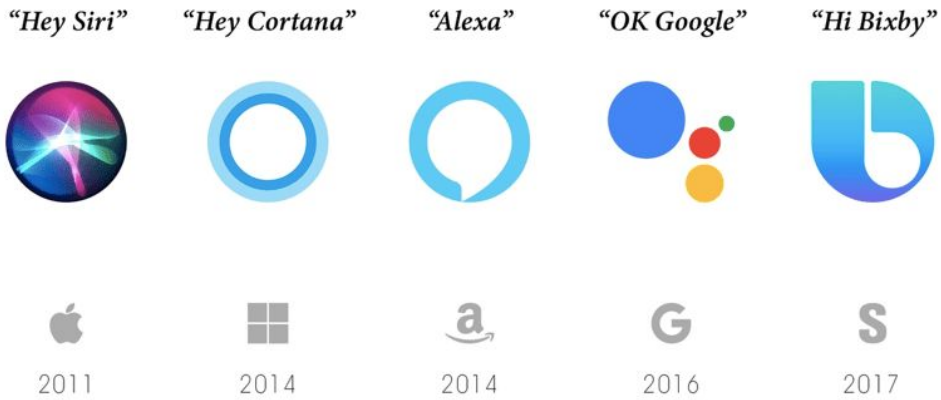
- Search Engines (e.g., Google, Bing)
- Voice Assistants (e.g., Google Assistant, Siri, Alexa)
- Recommendation Engines (e.g, Amazon, Ebay, Facebook, Youtube)
- Conversational and Multimodals AI (ChatBots / Audiovisual assistants) (e.g., ChatGPT, Bing Chat, GPT 4o, Gemini)
- Self-Driving Cars
- …

Search Engine

Search Engines

- Understand (?) natural language queries
- Provide relevant results
- Learn from user behavior

Voice Assistants



Voice Assistant

- Answer questions
- Set reminders and alarms
- Make calls and send messages
- Control smart home devices
- Play music and podcasts
- Get directions and traffic updates
- Translate languages
- Play games and tell jokes

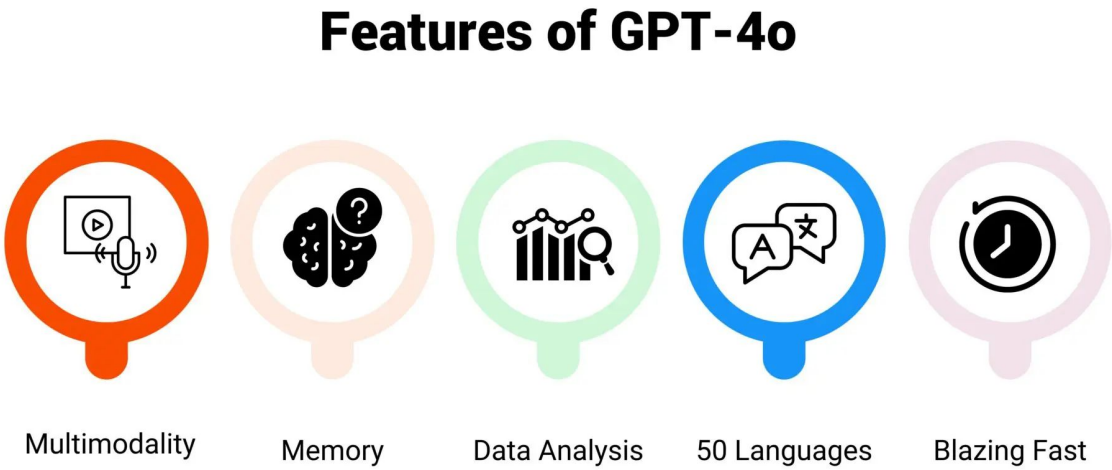
Recommendation Engines

- Recommend customers/consumers the products of their preference
- Early examples:
  - Using “unsupervised clustering algorithms” to identify similar customers in online shops (e.g., Amazon, Ebay, etc.)
- “Interesting” transition:
  - Suggesting posts/videos for social media consumption to
    - serve the customers better
    - maximize screen-time (more ads to show)

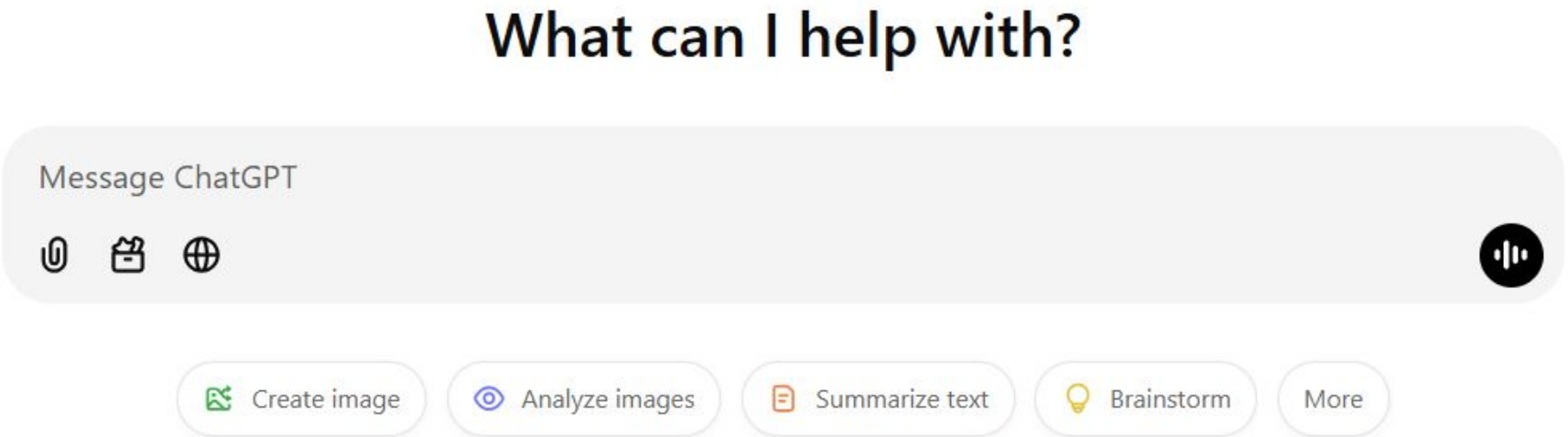
Multimodal Modals

Leading to Visual Assistants

Multimedia content creation



Conversational AI - ChatGPT, Copilot etc



ChatGPT by OpenAI

Tell me how to explain what AI is to a 10-year-old in 5 short sentences.

1. AI stands for Artificial Intelligence, which means a machine that can think and learn like a person.
2. It can do things like understand words, recognize pictures, or play games really well.
3. AI works by studying lots of data to figure out patterns and make smart guesses.
4. It's like teaching a computer how to solve problems and get better over time.
5. While AI is super helpful, it still needs humans to guide it and give it rules.

Co-pilot by Microsoft

- Formats search results as AI chat conversations.
- Utilizes Bing Search, powered by GPT-4, for accessing real-world information.

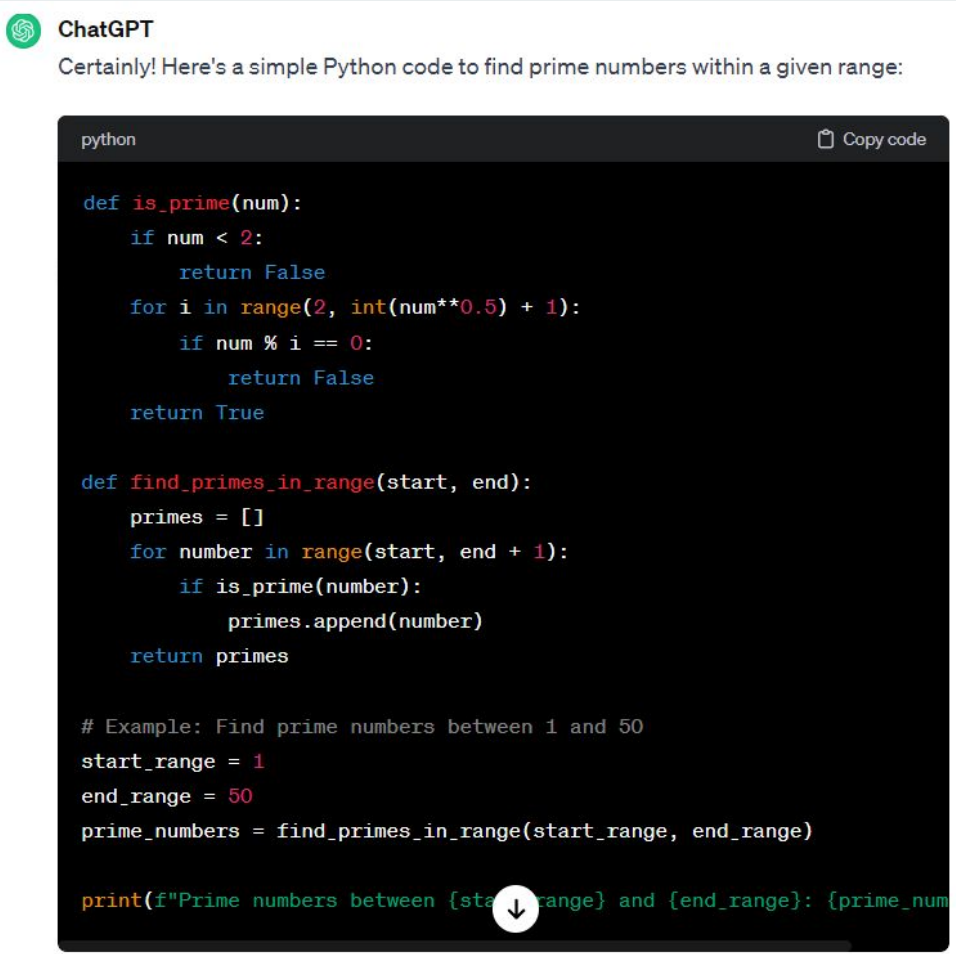
These models can generate code

Prompt: Write a python code to find prime numbers

Warning!

DO NOT just copy paste code

Must check the reliability of the code



How to use them properly? Ans: Prompt Engineering

Prompt: You are an **expert translator**. Your task is to translate the following text to Bangla. Please make sure the translated text is **natural, easy to follow, and has literary quality** to it.

[...A generative models is…]

More on prompt engineering: Prompt Engineering Guide (promptingguide.ai)

ChatGPT:

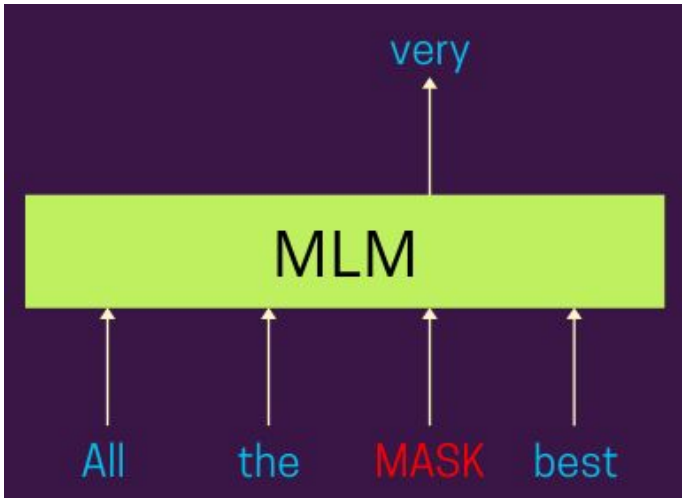
একটি জেনারেটিভ মডেল হল মেশিন লার্নিং মডেলের এক ধরন, যা তথ্যের নিচে লুকিয়ে থাকা প্যাটার্ন বা তথ্যের বিতর্কিত বিতর্কিত ভাগবিভাজন শেখার লক্ষ্যে তার সামনে নতুন, অনুরূপ তথ্য তৈরি করা। মূলত, এটি হলো যে, যেটি কোম্পিউটারকে শিখানোর মতো, যাতে তা আগে দেখা তথ্য ভিত্তিক নিজের তথ্য সৃষ্টি করতে পারে।

Not quite up to the mark is it? what do you think?

These are all examples of Large Language Models

- They perform **an autocomplete task**

Causal/Autoregressive LM



Masked LM

Reason behind the “success” of language models

- The models are large: **~ 200 billion parameters** or weights
  - Parameters are like variables to adjust to get desired output
- Are trained on data scraped from the internet
  - GPT3 on which the ChatGPT was built using **45TB of text data**
  - **1 TB equals approximately 6.5 million pages** of documents stored in common formats (docx / pdfs)
- Have used up almost all human generated data on the internet
  - Newer models are also being trained on data generated from other models
  - Potentially introducing “bias”

Reason behind “success”

- **Models don't actively search for information / answers**
  - rather predict the next word or masked tokens based on patterns learned from the training data
- There is “**inherent structure**” in the data:
  - Words that occur in similar contexts tend to have similar meaning
  - There is only a small set possible words that can be the right prediction
  - There is some clear way to check what is a right prediction or how good the prediction is

Reason behind “success”

- With the extensive availability of English text online, especially in comparison to the amount of Bangla text
  - These models will be good at writing poems in the style of Shakespeare
  - Whereas writing a poem in Bangla in the style of Fakir Lalon Shah the results might not be that impressive
- However, it is likely that these strategies will not work for all “problems”

Reasoning Ability - though not trained for it!

Tarek is taller than Masud, but Masud is shorter than Rafiq. Is Tareq taller than Rafiq?

The information provided does not give a direct comparison between Tarek and Rafiq. Here's the reasoning step by step:

1. **Tarek is taller than Masud.**
    - This means Tarek > Masud.
2. **Masud is shorter than Rafiq.**
    - This means Masud < Rafiq.

From these two pieces of information:

- We know that Tarek > Masud < Rafiq.
- However, we do not know the relationship between Tarek and Rafiq because their heights are not directly compared.

Conclusion: It cannot be determined from the given information whether Tarek is taller than Rafiq.

Surprises, Capabilities, and Challenges

- AI models do surprise us by tackling **some tasks** beyond their training
  - Question answering, Visual Question Answering etc.
- Capabilities can be enhanced through **sophisticated prompting techniques**
  - **Chain of Thought** - structure the output as a series of steps
  - **Self Consistency** - generate multiple reasoning paths and select the most consistent one
  - **Multiagent / Multipersona** - LLMs assume the role of multiple agents with different personalities and they debate to find the answer

Prompting Techniques | Prompt Engineering Guide

Surprises, Capabilities, and Challenges

While language models like ChatGPT and image generators like Stable Diffusion often work well, they still have limitations, such as:

- Repetition and Incoherence
- Hallucination / Factual Errors
- Lack of Contextual Understanding
- Unintended Biases
- Limited reasoning capabilities for complex problems

12 famous AI disasters | CIO

Repetition and incoherence

Generated text or images might contain repetitive patterns or lack coherence, making them less natural or realistic.

Popular image generator Dall-E struggles to generate hands

Hallucination

Text generated by language models can sometimes contain factual inaccuracies, incorrect information, or nonsensical phrases.

"Elephants are known for their ability to fly gracefully through the skies, flapping their wings with precision and elegance."

Lack of contextual understanding

Generative models might struggle to understand and maintain context, leading to responses or creations that don't appropriately align with the given context or topic.

Unintended biases

Like other AI systems, generative models can inherit biases present in the training data, potentially leading to biased or inappropriate outputs.

Limited capability in solving complex tasks

- Researchers are exploring their potential in complex tasks such as reasoning, problem-solving, and coding
    - 74% accuracy on SQL generation dataset*
    - 32% accuracy on challenging programming tasks
    - Only 2% accuracy on a recent math problem datasets*
- These models are still under development and things are moving fast
    - But still a long way to go

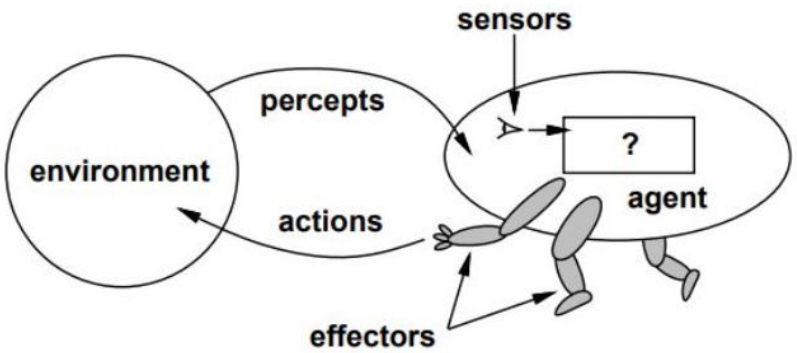
BIRD-bench (BIg Bench for LaRge-scale Database Grounded Text-to-SQL Evaluation)
BigCodeBench Leaderboard
FrontierMath: Evaluating Advanced Mathematical Reasoning in AI | Epoch AI | Epoch AI

Challenges to address

- Data Inefficiency remains a challenge
- Rather than ‘scaling up’ these models indefinitely (larger models), researchers will have to
  - find more efficient models
  - invent technologies that can learn with less supervision/data

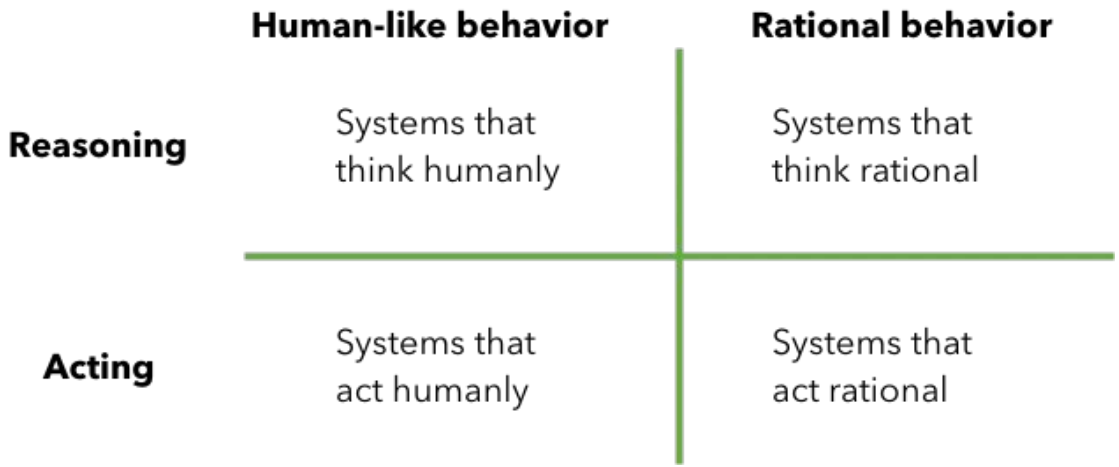
So what is AI?

- an Intelligent Agent (program or entity) that
  - interacts with its environment by perceiving its surroundings via sensors
  - then acting through actuators or effectors



What does it do?

Stuart Russell and Peter Norvig authors of **Artificial Intelligence A Modern Approach** discusses following four “schools of thought” to define AI:



Rational Agent

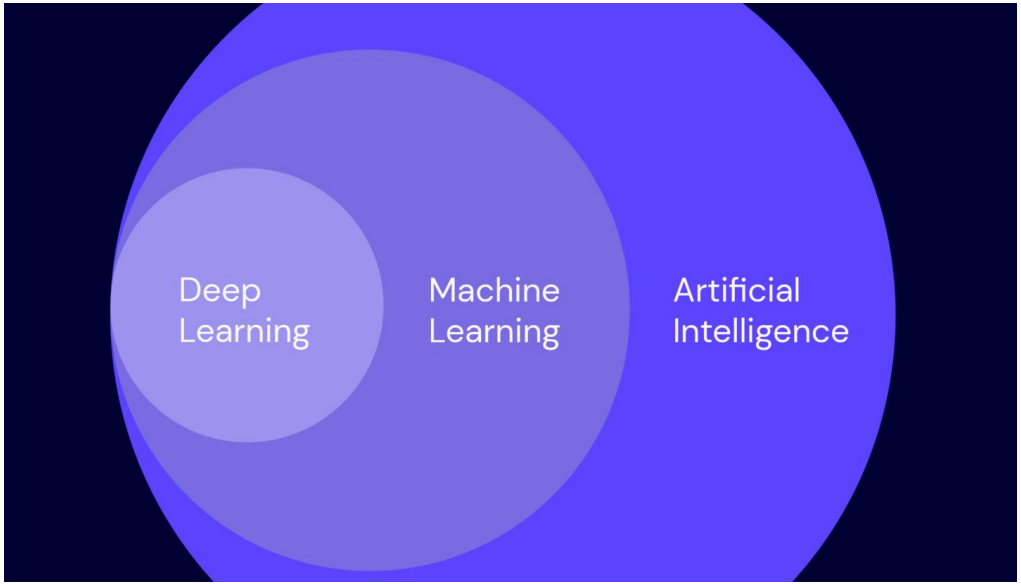
Rationality depends on

- Performance measure that defines criterion for success
- Agent's prior knowledge of the environment
- Actions the agent can perform
- Agent's percept sequence to date

It is the intelligent **agent** perceives the environment and tries to maximize the performance measure(s) through its actions

Intro to Machine Learning

Distinguishing the terms DL, ML, & AI



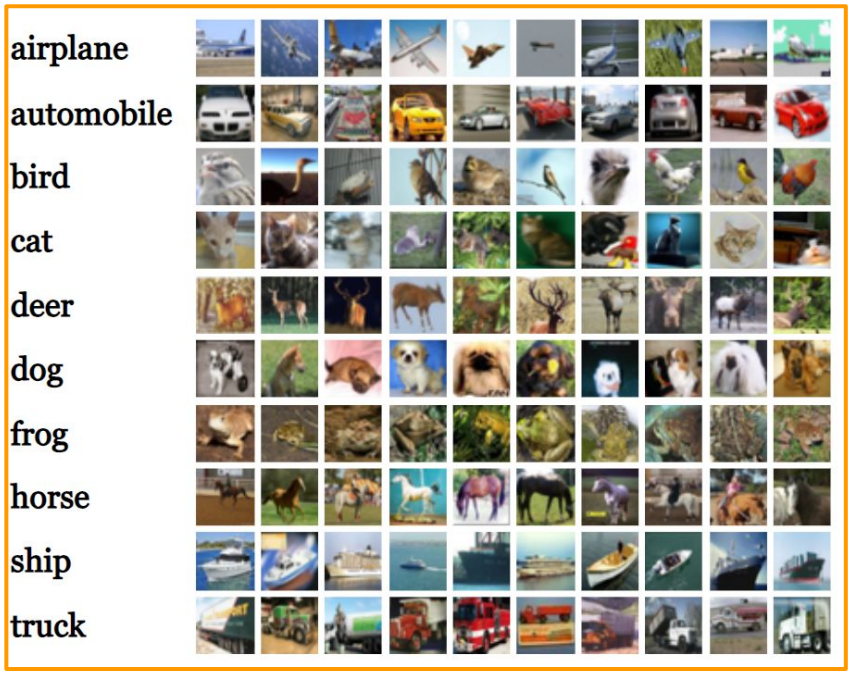
Data Driven Models

Software 1.0 uses pre-programmed logic on data to create the desired behaviour. Software 2.0 uses data and the desired behaviour to come up with the right logic [53].

Source: A Karpathy, Software 2.0, medium post

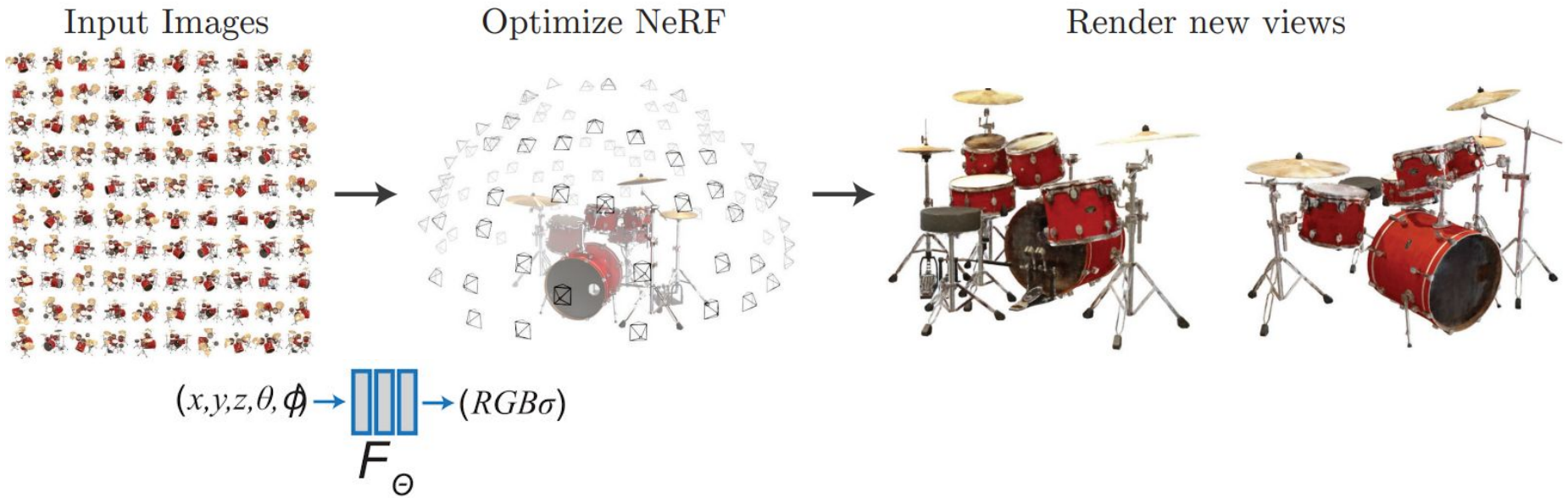
Machine Learning (ML)

- enables computers to “learn” from data without being explicitly programmed
  - Sometimes explicitly writing a program is hard/almost impossible
- ML algorithms identify patterns to make predictions or decisions.



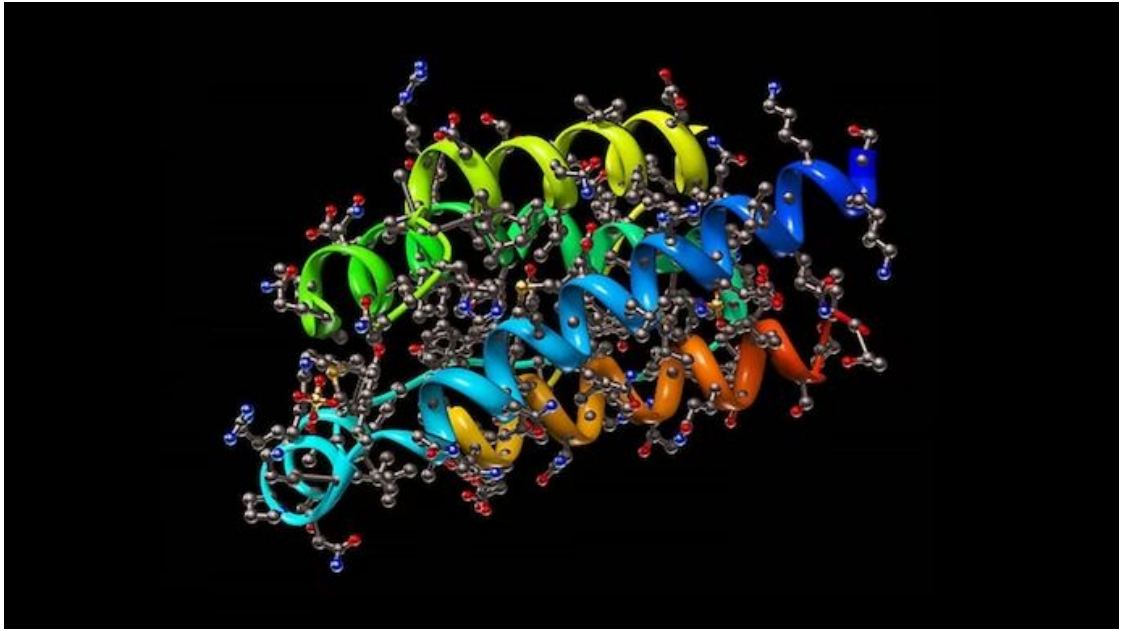
Images Classification (CIFER 10 dataset)

ML Applications: View synthesis (3D modeling from data)



Representing Scenes as Neural Radiance Fields (NERF) for View Synthesis ECCV 2020

ML Application: Alphafold - Protein 3D Structure prediction from amino acid sequence

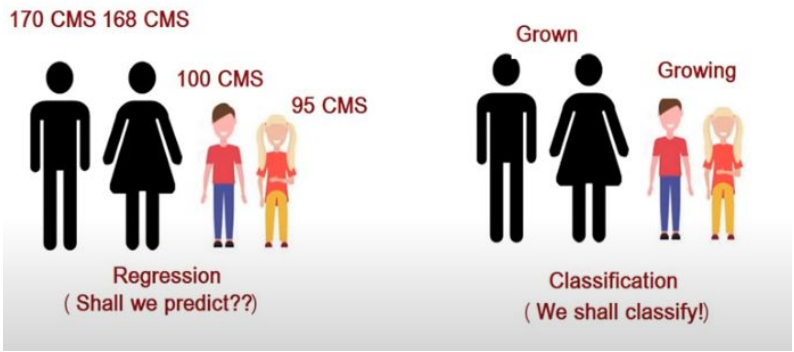


Types of Machine Learning

- Supervised Learning
- Unsupervised Learning
- Self-supervised Learning
- Reinforcement Learning

Supervised Learning

- Input data and its label is given
- Predict the label of new data
- The label
  - can be a number - **Regression**
  - can be a category - **Classification**



Supervised Learning

Predicting house prices

- **Data**: Features of houses: size, location, and number of bedrooms etc.
- **Label**: selling price of each house

Diagnosing medical conditions

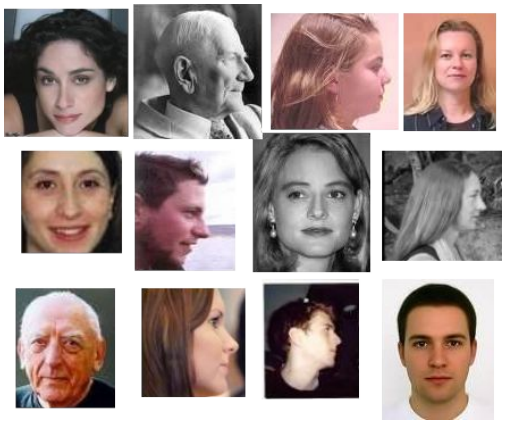
- **Data**: Medical images, such as X-rays and MRIs
- **Labels**: Whether each image shows signs of a particular medical condition

Unsupervised Learning

- The data is unlabeled (no supervision)
- The goal is to find patterns or structure in the data without being told what to look for

Unsupervised Learning: Clustering

How would you group these images?



Unsupervised Learning

Group by Gender:

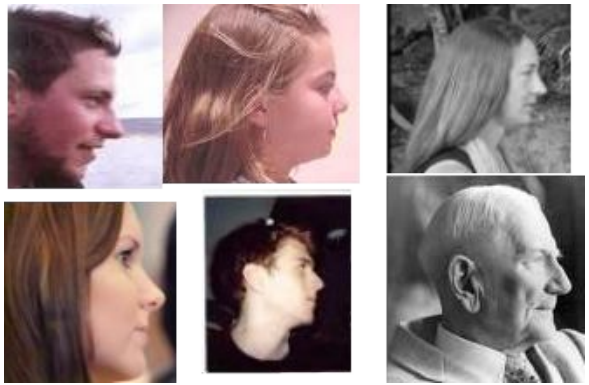
Group 1

Group 2

Unsupervised Learning

How about this grouping?

Group 1



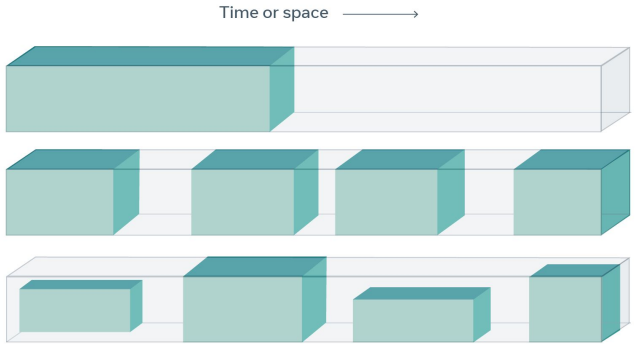
Group 2

Self Supervised Learning

- We live in the “Big data” era
  - lots of news articles, video data
- But labeling is hard
  - labeling each article as sports, political, entertainment
  - Labeling each person / object in the video
- Why not design tasks where we have labels in the data itself?

Self Supervised Learning

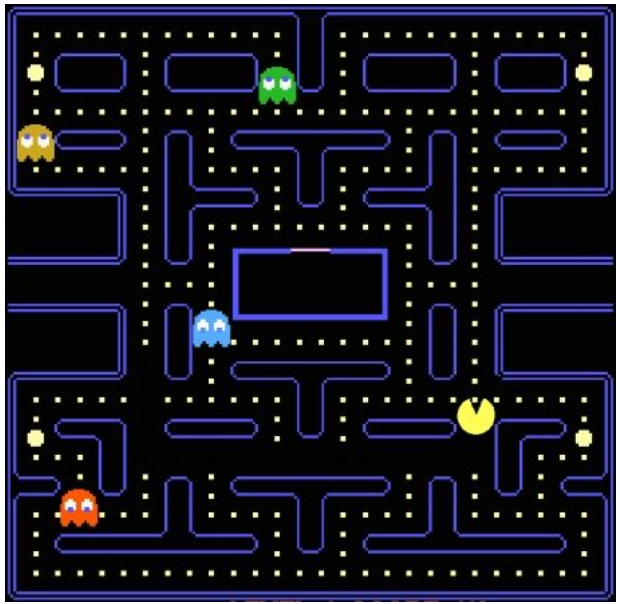
- Predict hidden parts of the input (in gray) from visible parts of the input (in green)
- **In NLP**
  - Hide (mask) some words of a sentence and predict them from the remaining words: **Masked Language Models**
  - Predict the next work from the previous ones : **Autoregressive models**
  - **Distributional hypothesis**: words that occur in similar contexts tend to have similar meanings
- **Video Processing**:
  - Predicting past or future (hidden data) frames from current frames (observed data)



Reinforcement learning

- computer learns by trial & error by interacting with the environment
- goal is to maximize the expected reward or utility signal received
- This is much like how you can train a dog to do tricks

Pacman has to “learn” by playing the game by himself



\+ Reward: eating the points

\- Reward: bumping into the ghosts

Important: Pacman does not know any of this at the beginning

Reinforcement learning

Training a self-driving car:

- The algorithm learns to make decisions about acceleration, braking, and steering based on the feedback it receives from the car's sensors.

Machine Learning Models

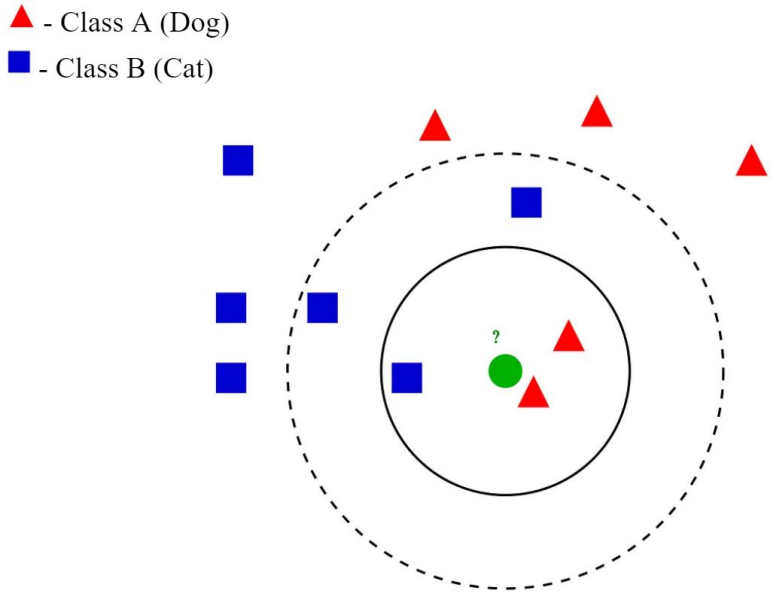
K-nearest neighbors (KNN)

Decision trees

Neural Networks

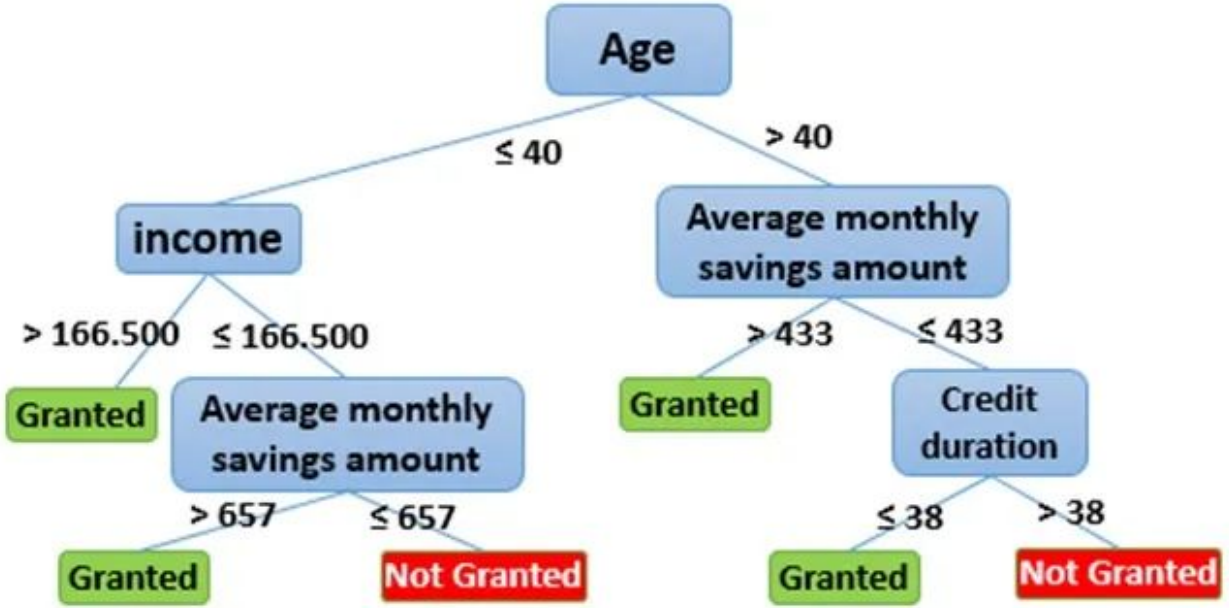
Deep Neural Networks

K-nearest neighbors (KNN)



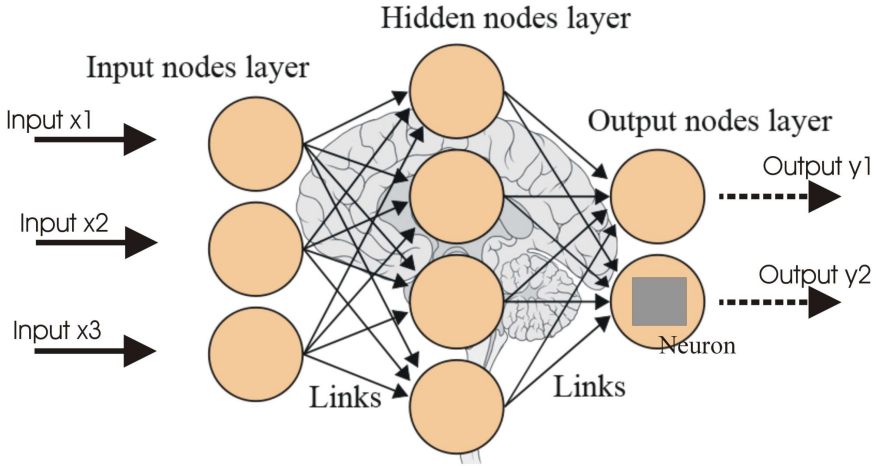
Based on closeness or similarity what is the class of the green circle?

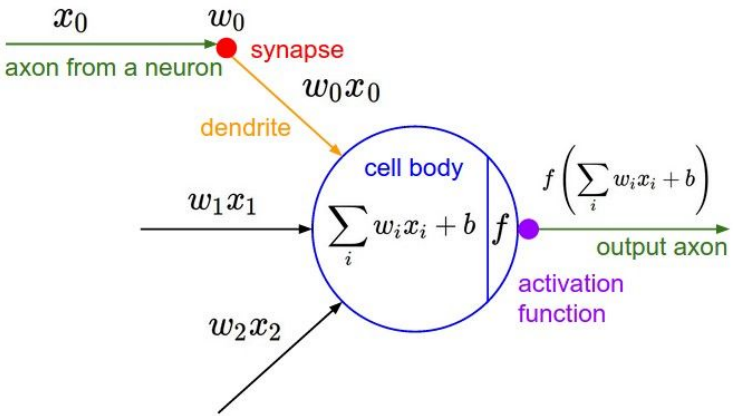
Decision Tree



Age	Income	Avg Monthly Savings Amt	Credit Duration	Decision
38	162000	650	37	?

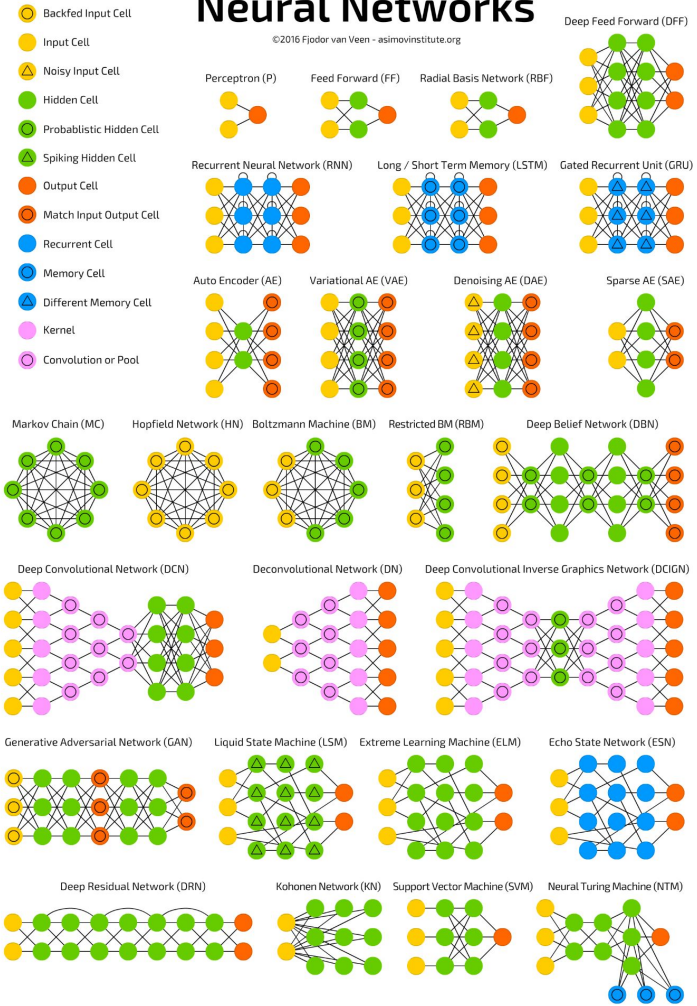
Neural Networks





A mostly complete chart of

Neural Networks



Deep Learning

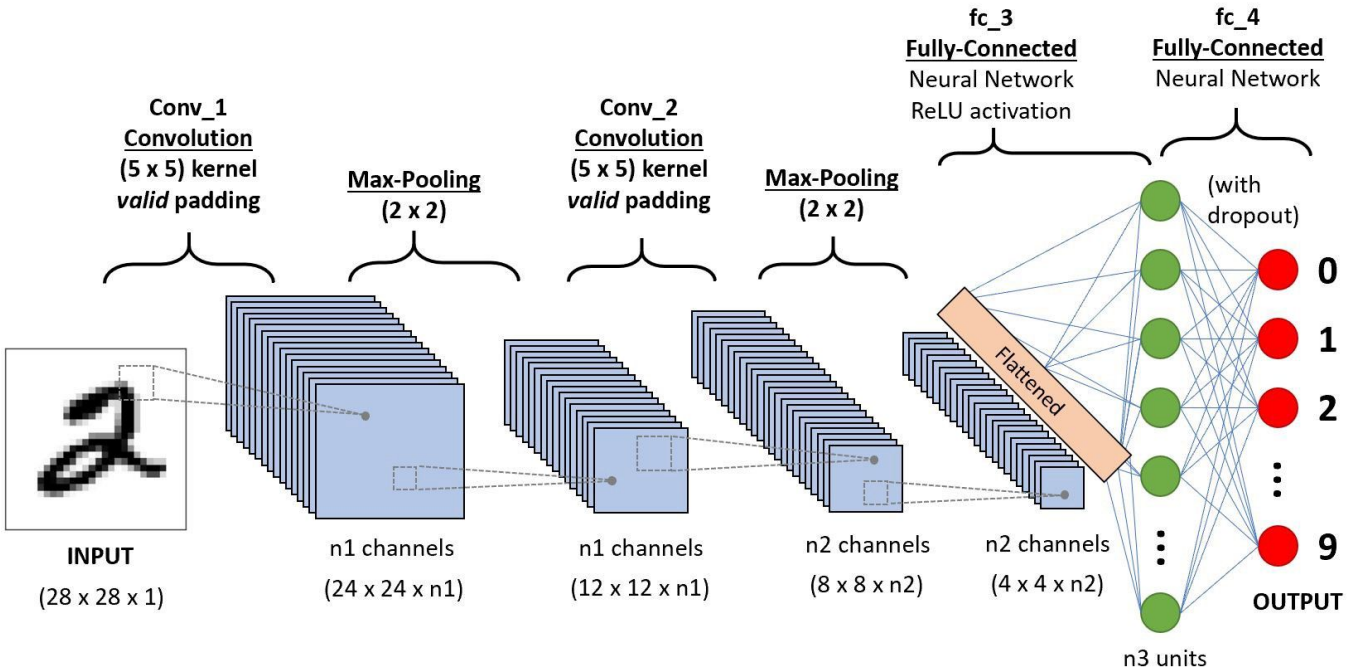
Deep neural networks have

many hidden layers or

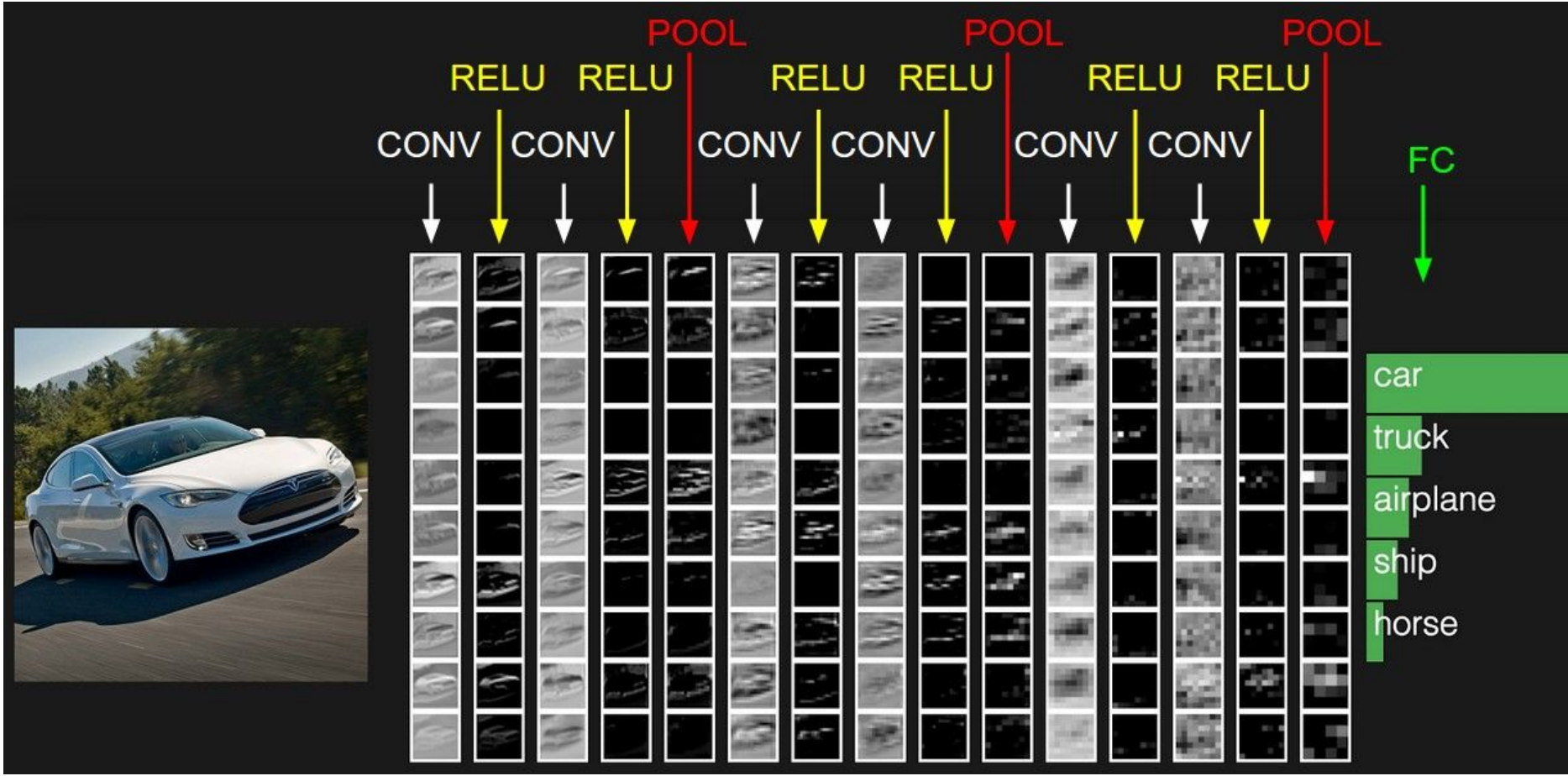
many many neuron connected in special ways

ML for deep neural networks are called deep learning

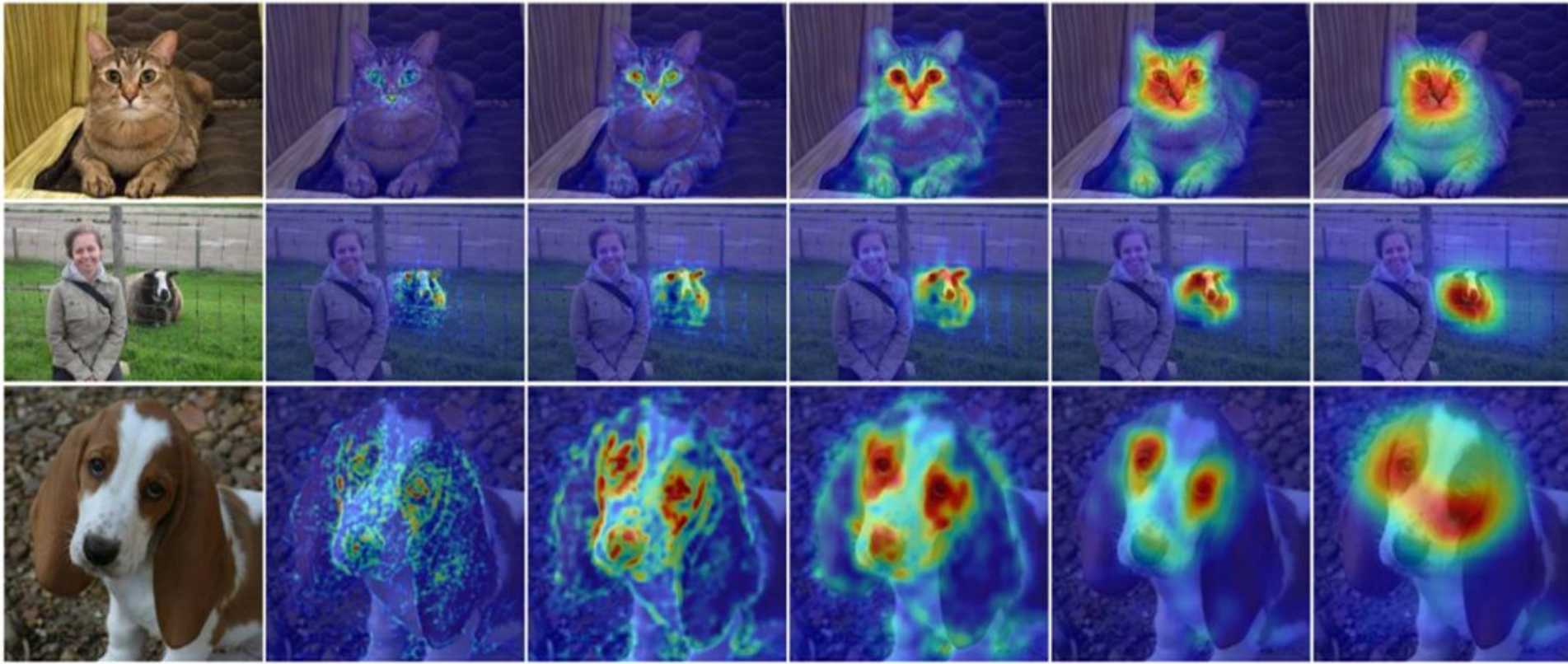
Basic CNN architecture



Automatic Feature Extraction - Convolutional NNs



Class Activation Maps



Encoder Decoder Architecture

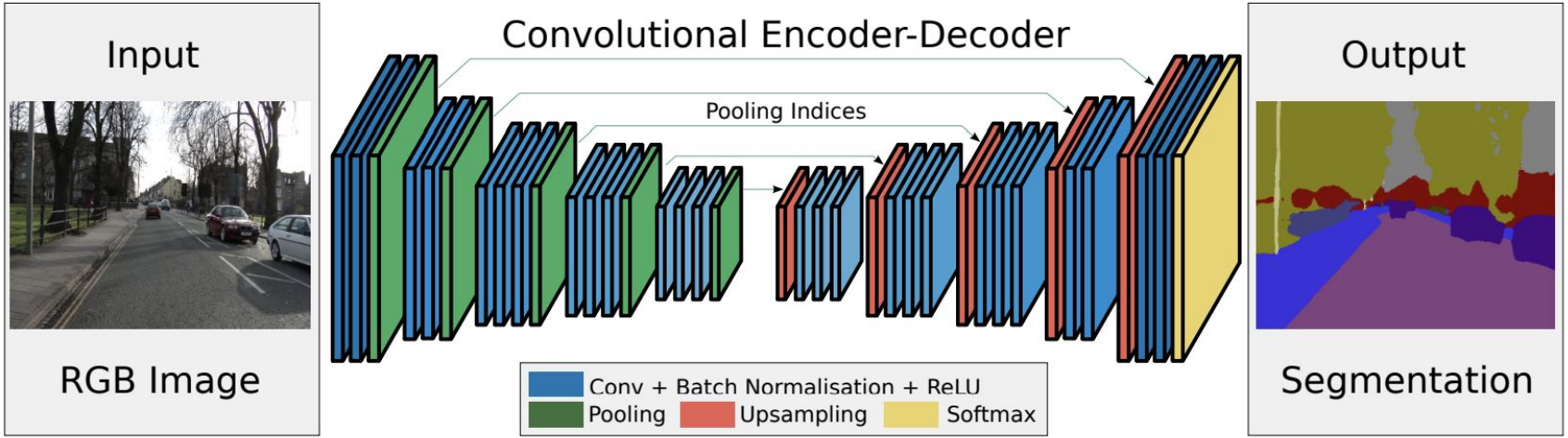
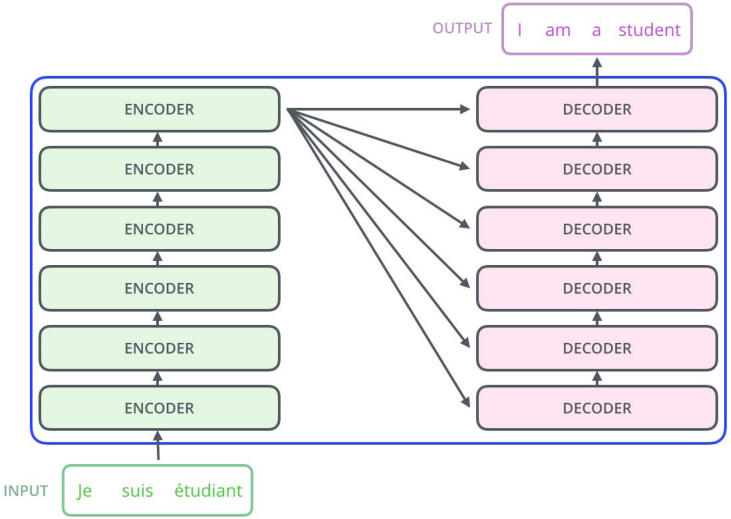
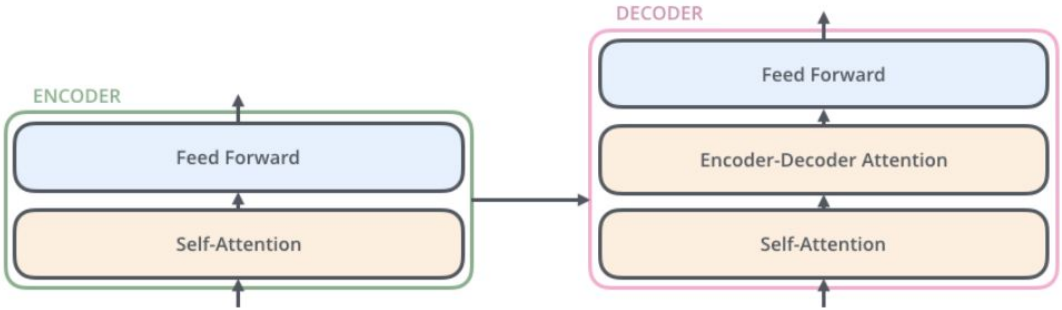


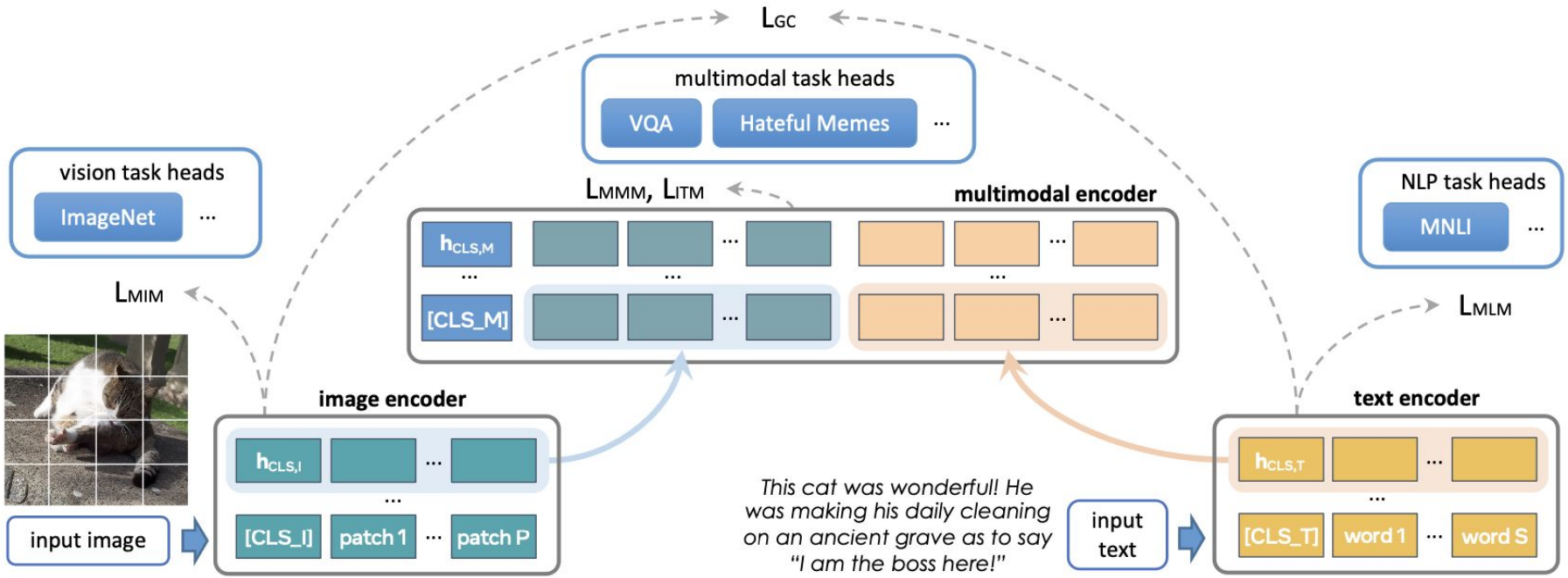
Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

Transformers



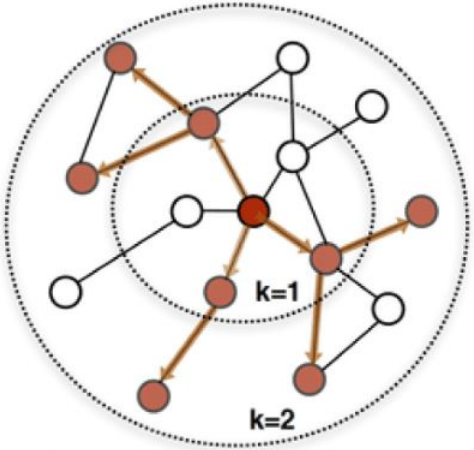


Multimodal Models

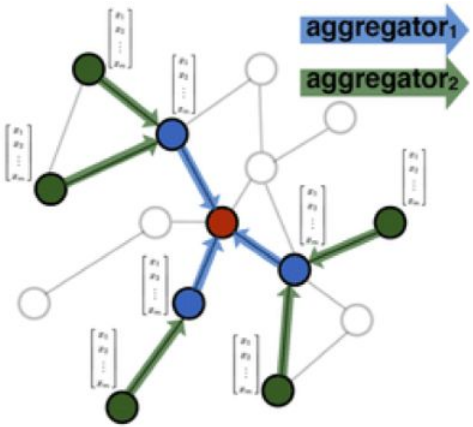


Source: [Scaling Multimodal Foundation Models in TorchMultimodal with Pytorch Distributed | PyTorch]

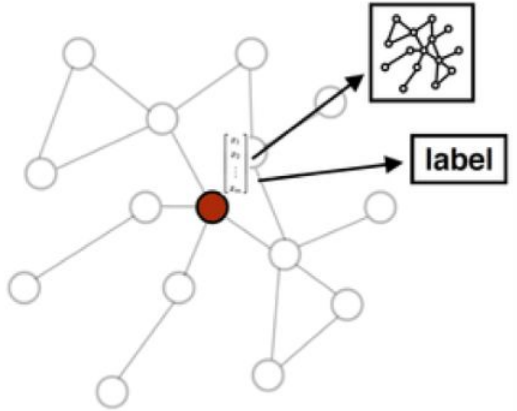
Graph Neural Networks



1. Sample neighborhood

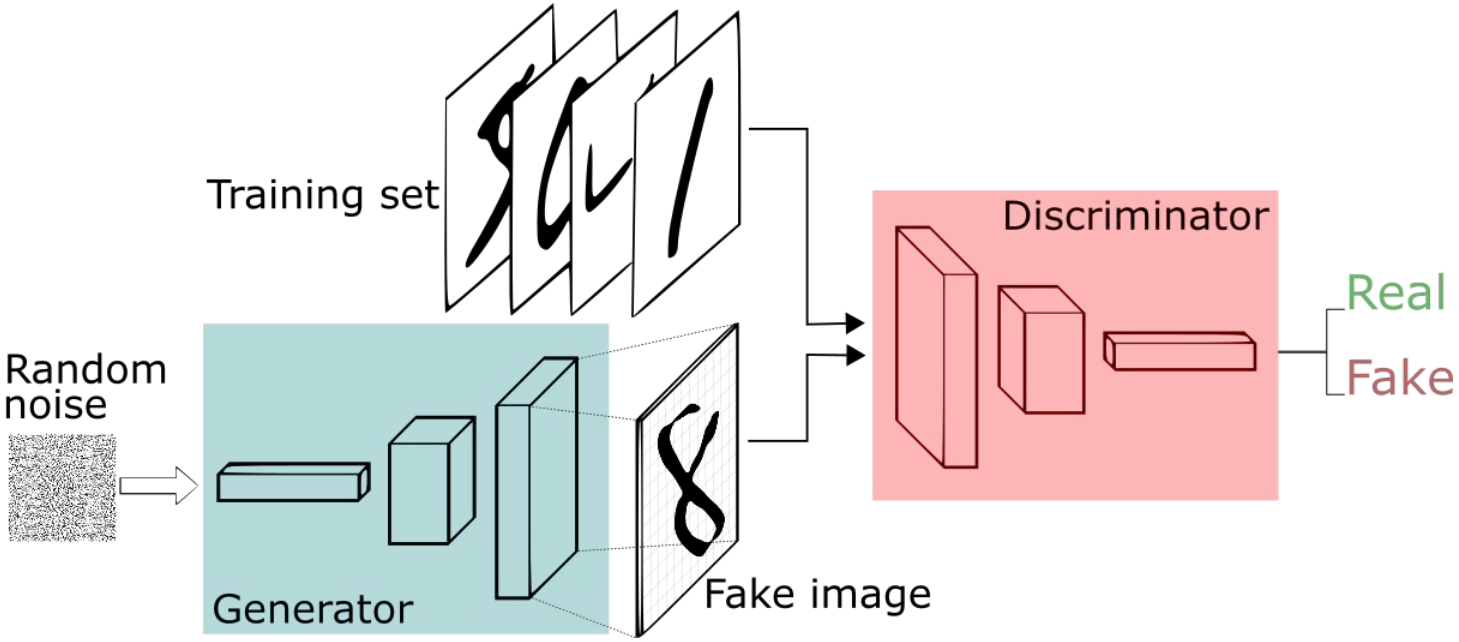


2. Aggregate feature information from neighbors



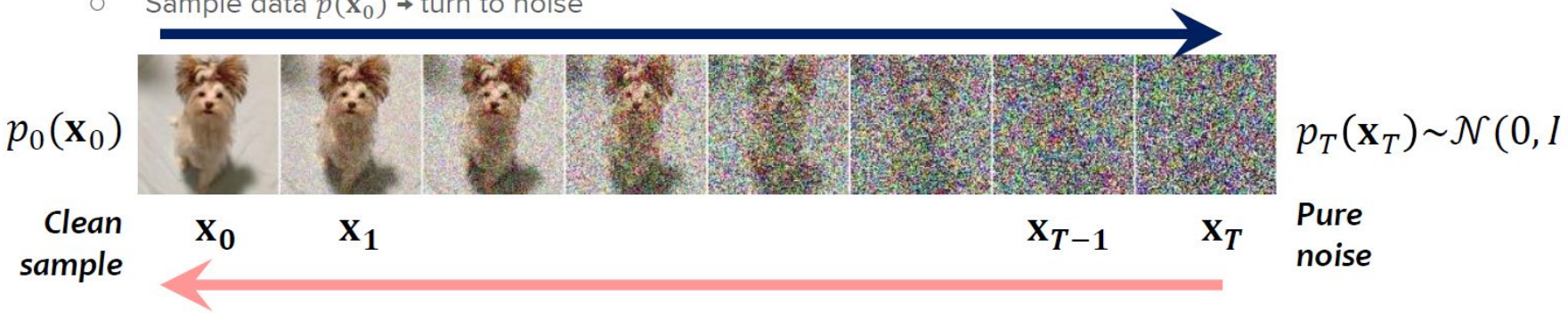
3. Predict graph context and label using aggregated information

Generative Adversarial Network (GAN)



Denoising diffusion models

- **Forward / noising process**
  - Sample data $p(\mathbf{x}_0)$ → turn to noise

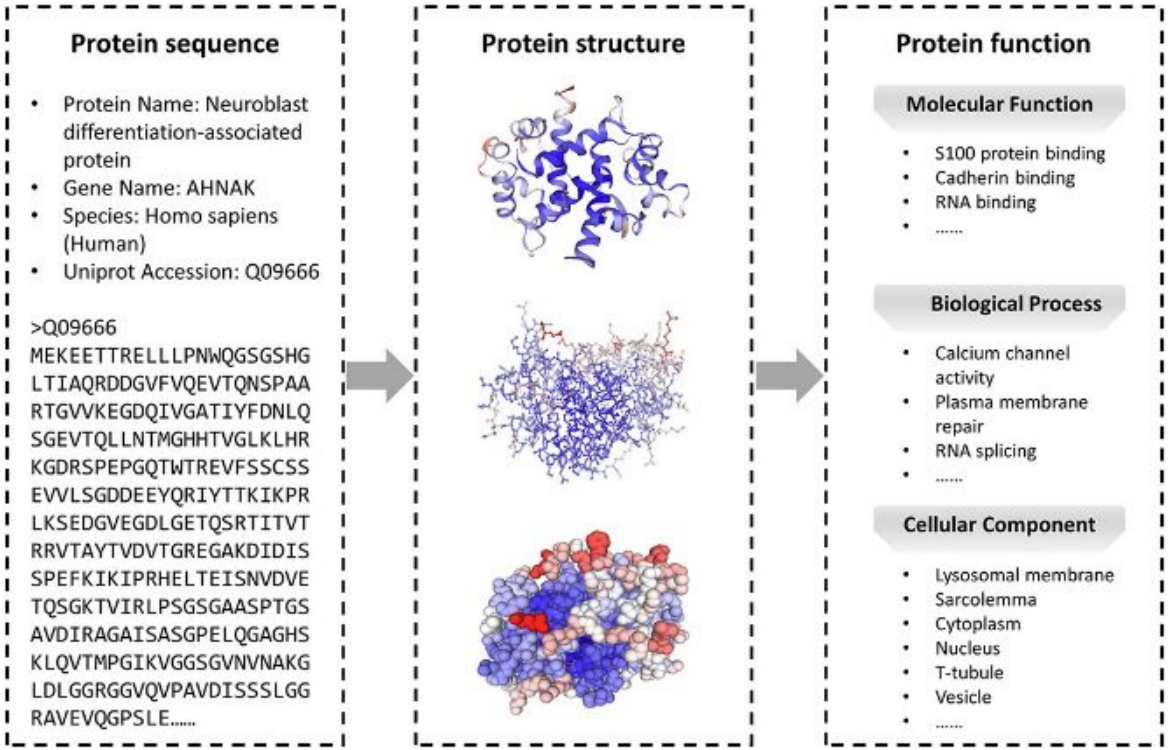


- **Reverse / denoising process**
  - Sample noise $p_T(\mathbf{x}_T)$ → turn into data

ML in Computational Biology

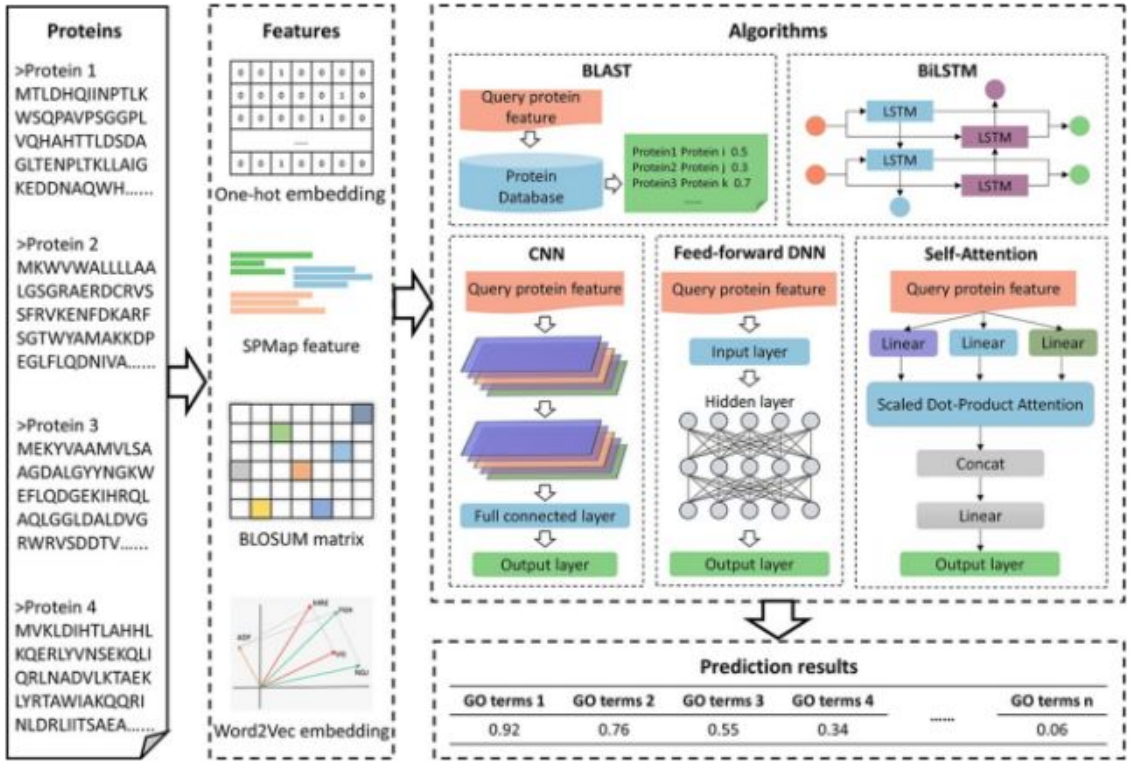
Protein Function Prediction:
“Sequence -> Structure -> Function” Paradigm

Figure 1 A schematic view of the 'sequence-structure-function' paradigm. Sequence refers to the arrangement of amino ...



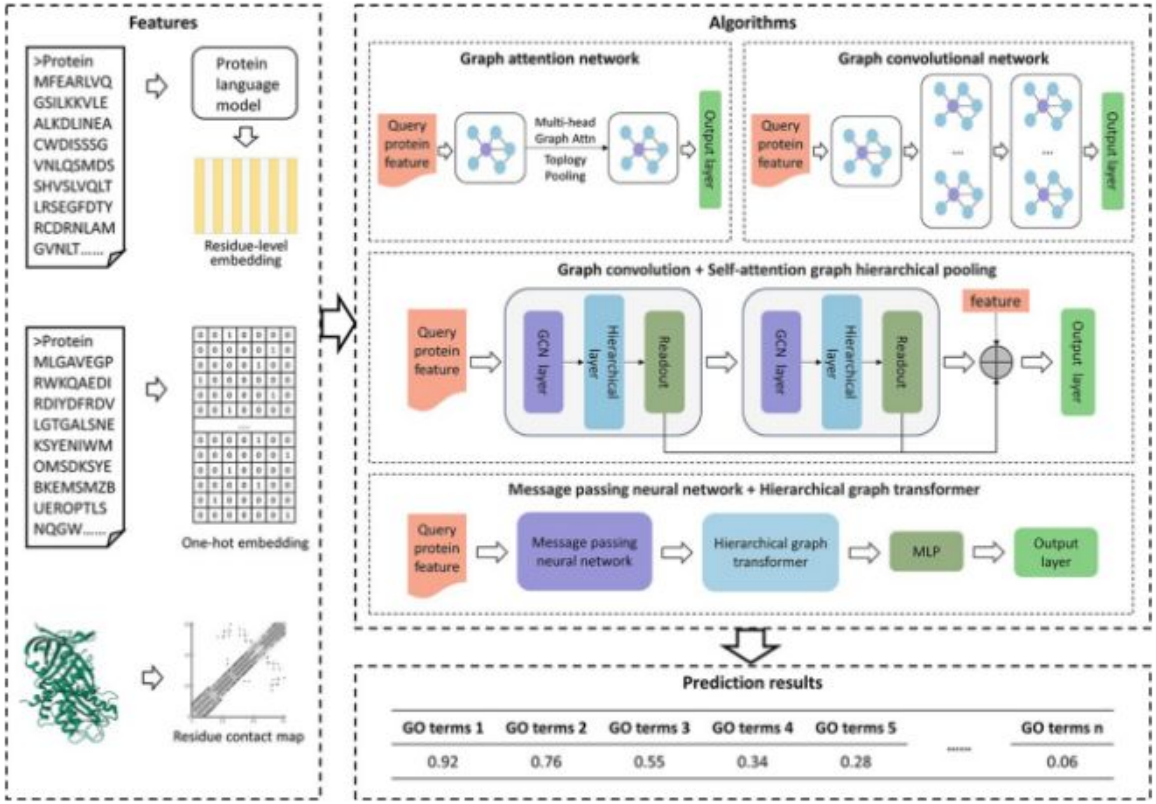
Brief Bioinform, Volume 25, Issue 4, July 2024, bbae289, https://doi.org/10.1093/bib/bbae289

Figure 2 The flowchart of sequence-based methods. First, features are extracted from the sequence by different ...



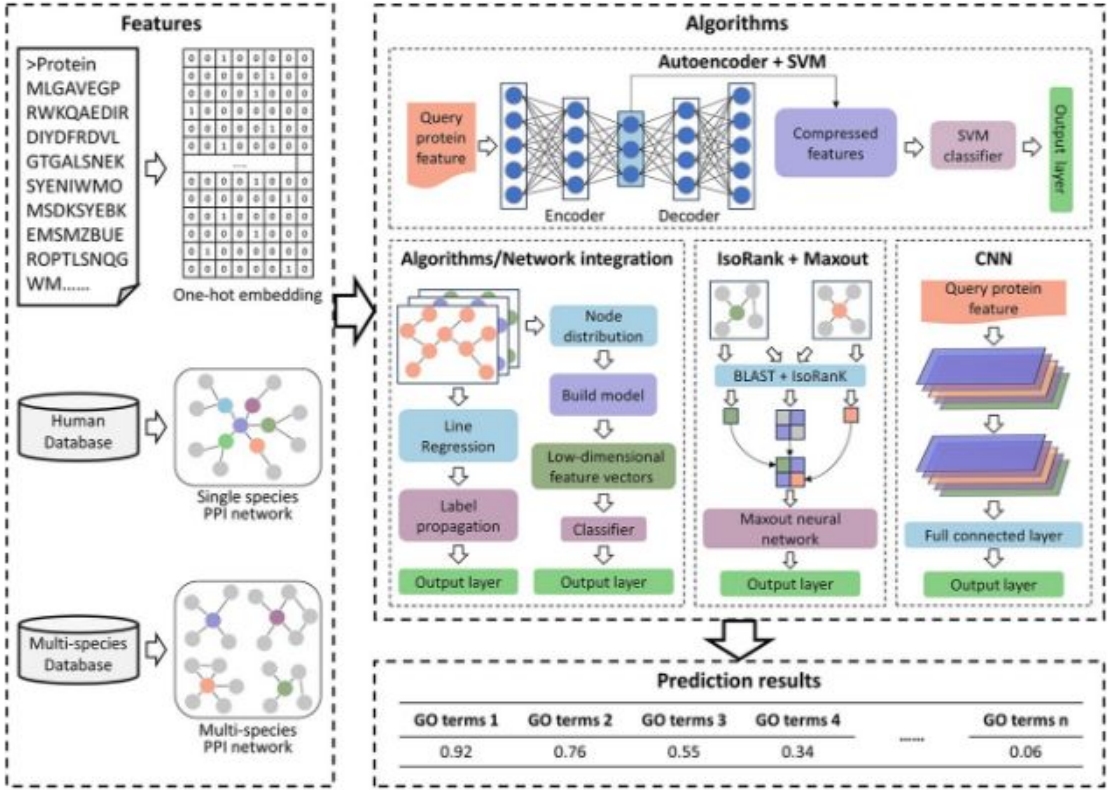
Brief Bioinform, Volume 25, Issue 4, July 2024, bbae289, https://doi.org/10.1093/bib/bbae289

Figure 3 The flowchart of 3D structure-based methods. First, the feature extraction methods can be categorized into ...



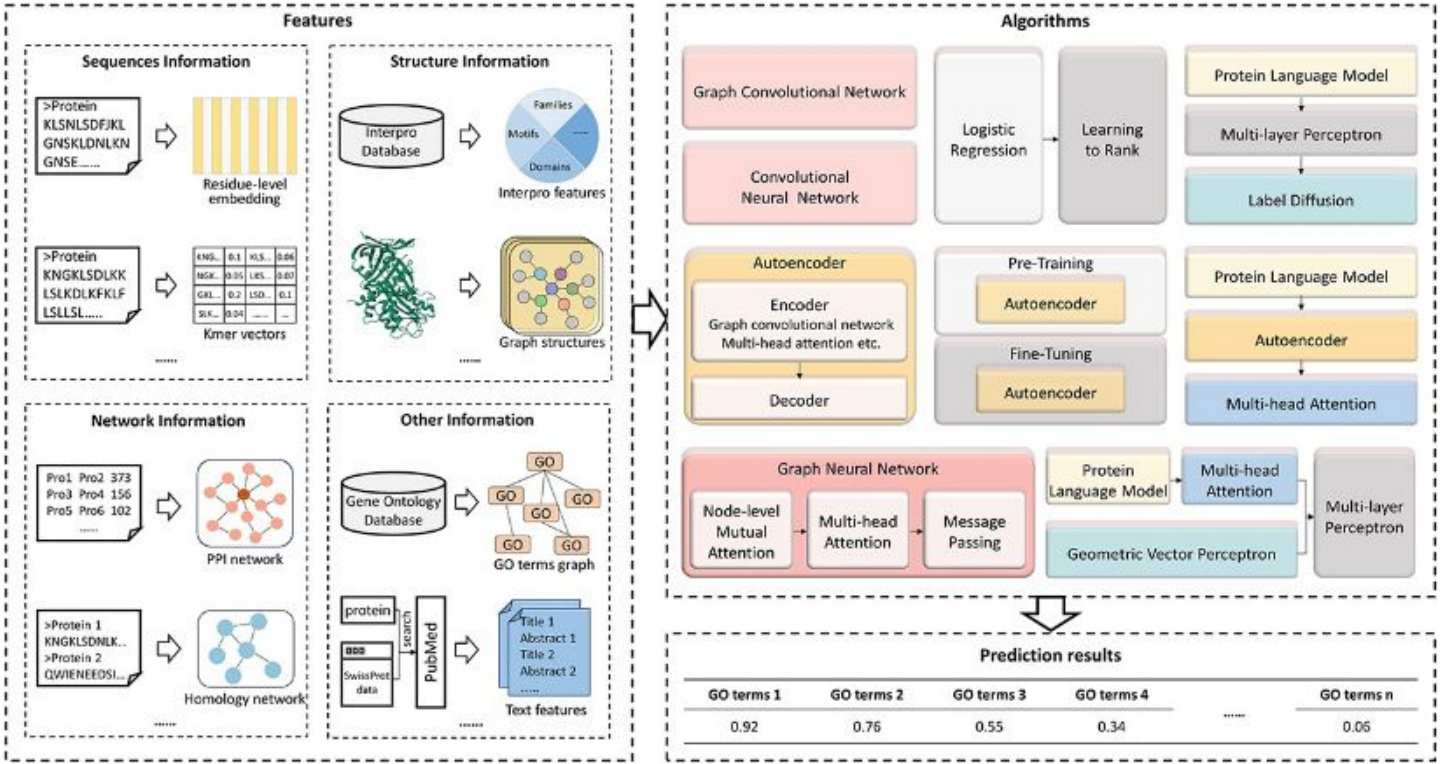
Brief Bioinform, Volume 25, Issue 4, July 2024, bbae289, https://doi.org/10.1093/bib/bbae289

Figure 4 The flowchart of PPI network-based methods. First, the feature extraction methods can be categorized into ...



Brief Bioinform, Volume 25, Issue 4, July 2024, bbae289, https://doi.org/10.1093/bib/bbae289

Figure 5 The flowchart of hybrid information-based methods. First, features can be extracted from four different types ...



Brief Bioinform, Volume 25, Issue 4, July 2024, bbae289, https://doi.org/10.1093/bib/bbae289

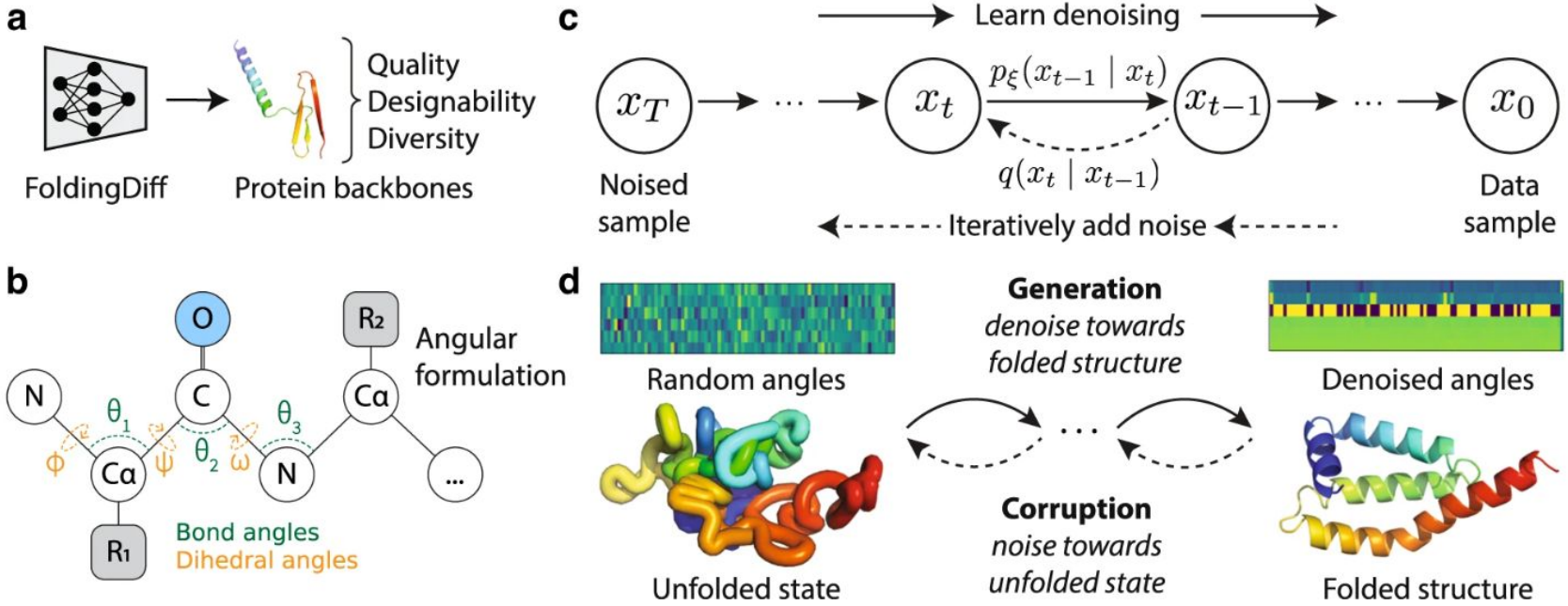
Is this enough?

- Function also depends on the molecular interactions of a group of proteins
  - These are not considered in any of the above models
- Combining Molecular Dynamics Simulation data can be useful
  - Given the positions of all of the atoms in a biomolecular system (e.g., a protein surrounded by water and perhaps a lipid bilayer),
  - calculate the force exerted on each atom by all of the other atoms. One can thus use Newton's laws of motion to predict the spatial position of each atom as a function of time.
  - The resulting trajectory is a three-dimensional movie that describes the atomic-level configuration of the system at every point during the simulated time interval.
- However, to ensure numerical stability, the time steps in an MD simulation must be short, typically only a few femtoseconds (10^{-15} s) each.
- But most of the events of biochemical interest—for example, functionally important structural changes in proteins—take place on timescales of nanoseconds, microseconds, or longer.
- Newer (ML based) methods and GPU hardware can help to do MD simulations at longer timescales

Diffusion Models

Fig. 1: Overview of FoldingDiff.

From: Protein structure generation via folding diffusion



Source: Protein Structure Generation via Folding Diffusion, Nature Communication, Feb 2024

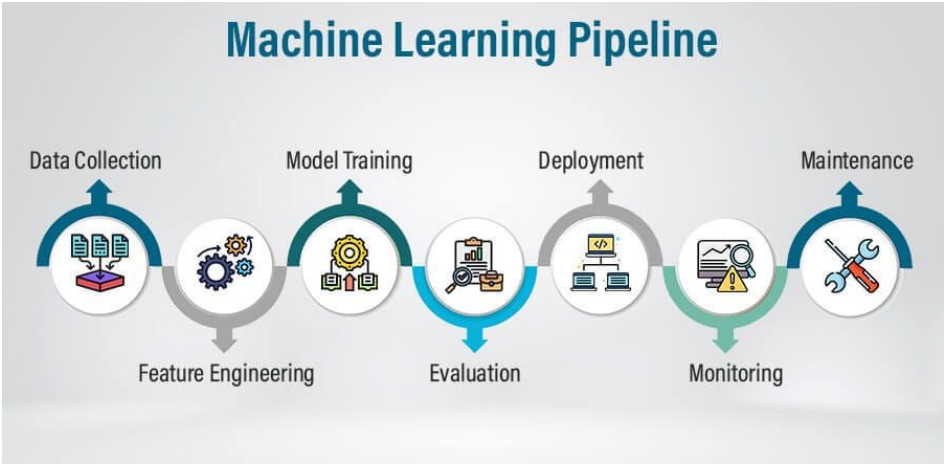
Machine Learning Development Workflow

Overview: ML involves building systems that learn from data to make predictions or decisions.

Key Phases:

1. Problem Definition
2. Data Preparation
3. Model Selection
4. Training and Evaluation

Objective: Solve real-world problems by leveraging data-driven insights.



From Problem to Data

Phase 1: Problem Definition

- Define goals and constraints.
- Identify success metrics (e.g., accuracy, precision).

Phase 2: Data Preparation

- Collect and clean data
- Split dataset into training and testing data.
- Engineer features .
- Ensure data quality for reliable model performance.

Modeling and Evaluation

Phase 3: Model Selection

- Choose algorithms based on the task.
- Leverage pre-trained models for complex problems.

Phase 4: Training and Evaluation

- Train models using train data.
- Evaluate the model on the test data
- Use metrics like accuracy, F1-score, MSE/RMSE etc
- Refine models through iterative tuning.

Key Takeaway: ML development is an iterative, end-to-end process aimed at

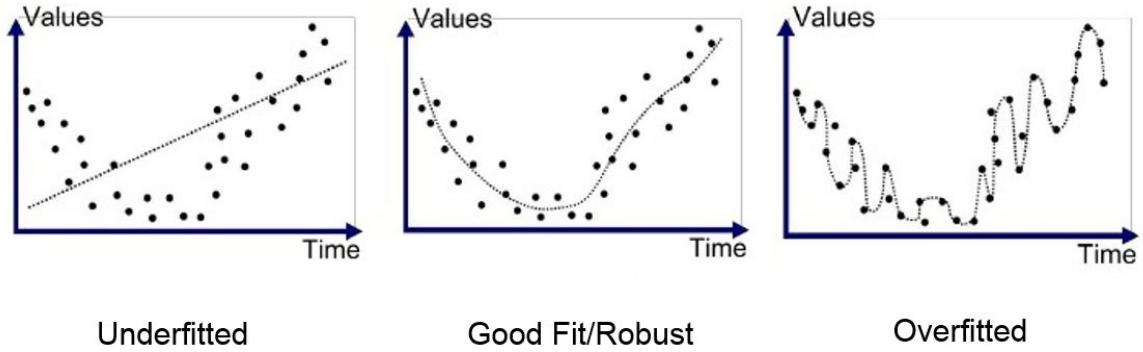
How do we evaluate regression models?

Split data into Training and Testing parts

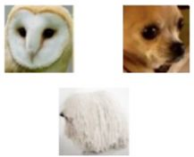
For **regression** a common metric is Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values



How to evaluate classification models

		Predicted: Animal	Predicted: Not animal
Actual	Animal		
	Not animal		

True Positives	2
True Negatives	3
False Positives	0
False Negatives	1

Accuracy	83%	$\frac{3+2}{3+2+0+1}$
Precision	75%	$\frac{3}{3+1}$
Recall	100%	$\frac{3}{3+0}$
F1 score	86%	$2 \cdot \frac{0.75 \cdot 1}{0.75+1}$

How to evaluate classification models

		Ground Truth			
		Interest	Confusion	Boredom	Precision
Prediction	Interest	**17**	1	3	81.0%
	Confusion	2	**17**	2	81.0%
	Boredom	1	1	**11**	84.6%
	None	0	1	4	Accuracy
	Recall	85.0%	85.0%	55.0%	**75%**

Overfitting?