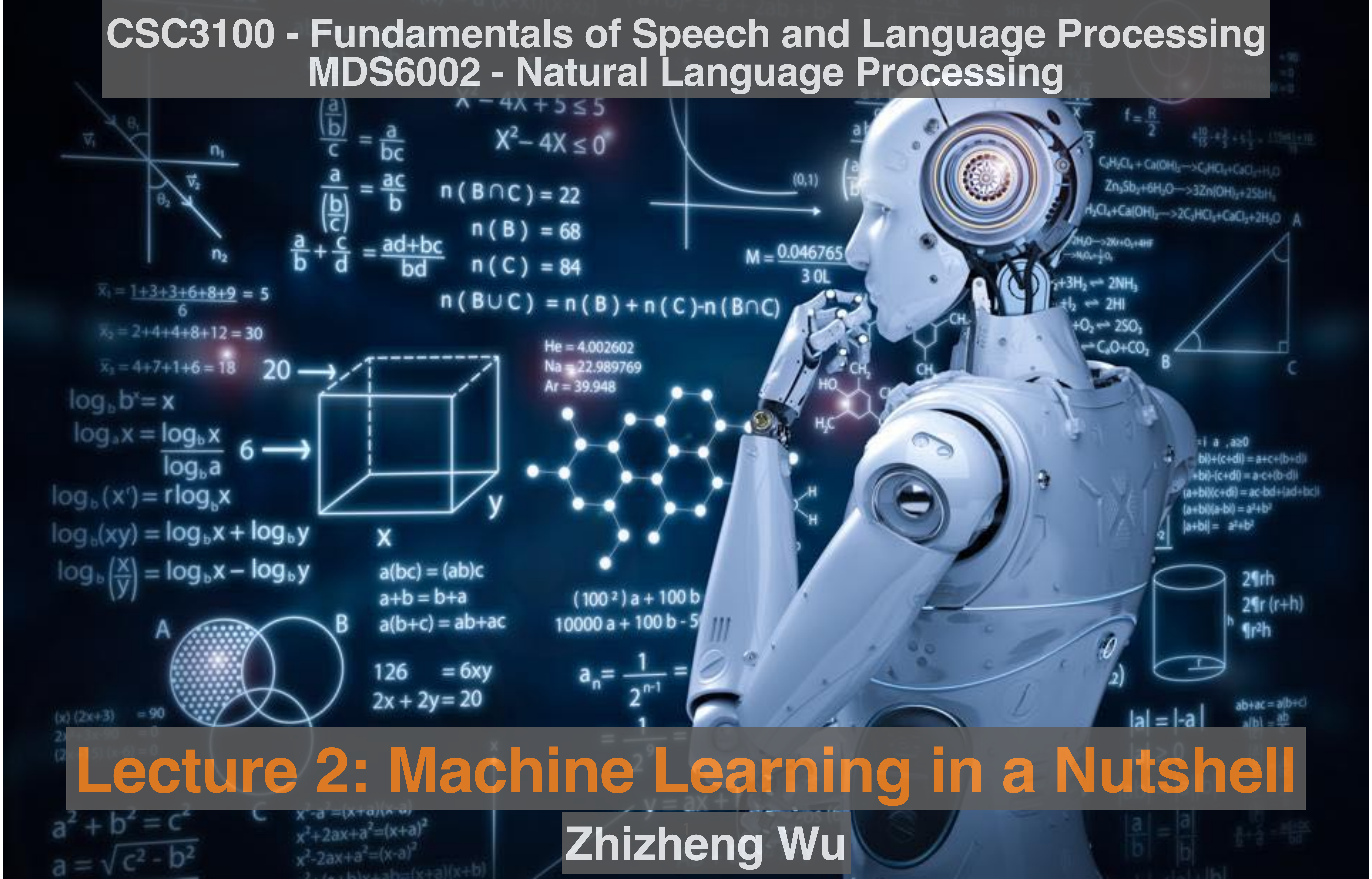
CSC3100 - Fundamentals of Speech and Language Processing
MDS6002 - Natural Language Processing

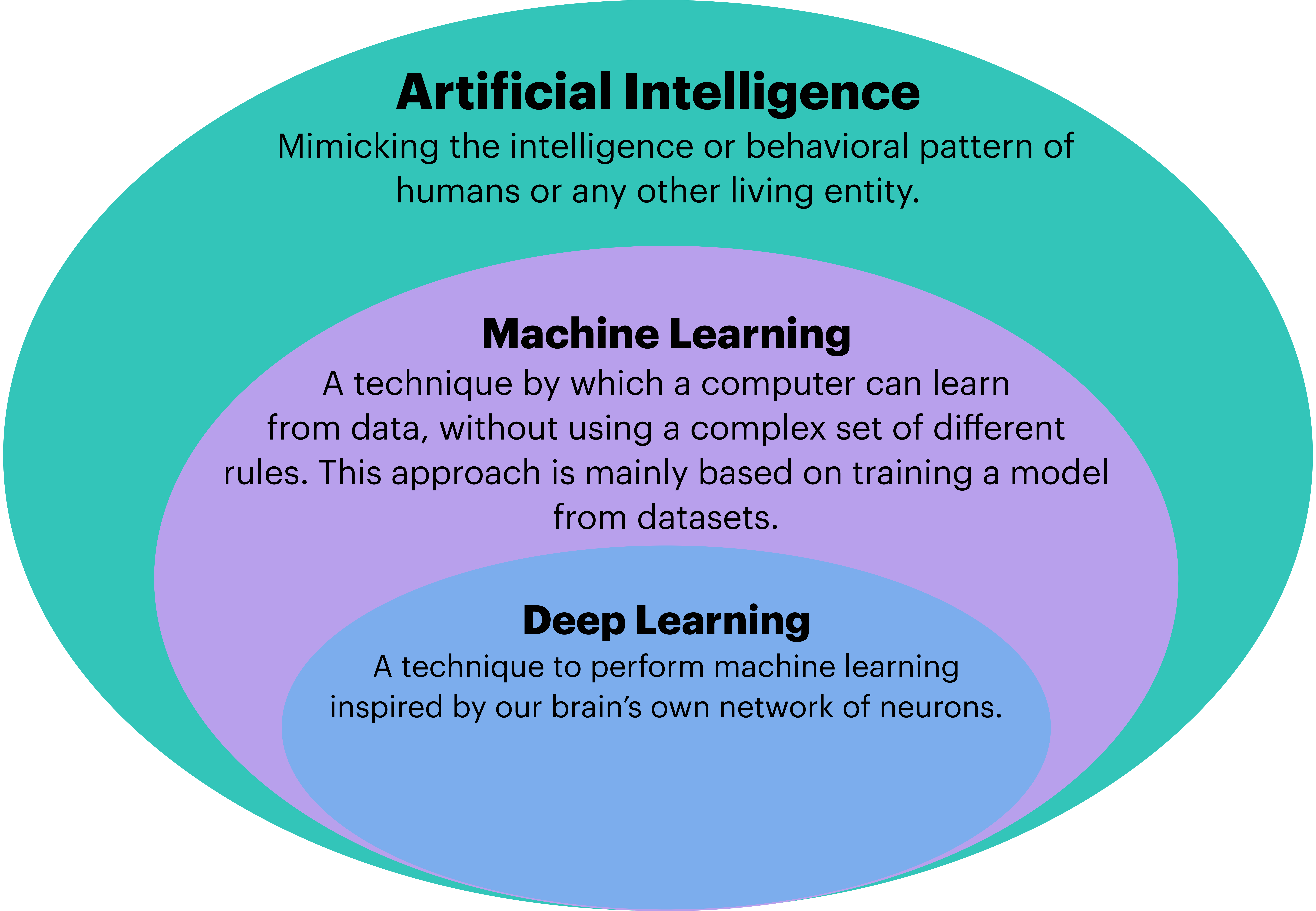


Lecture 2: Machine Learning in a Nutshell

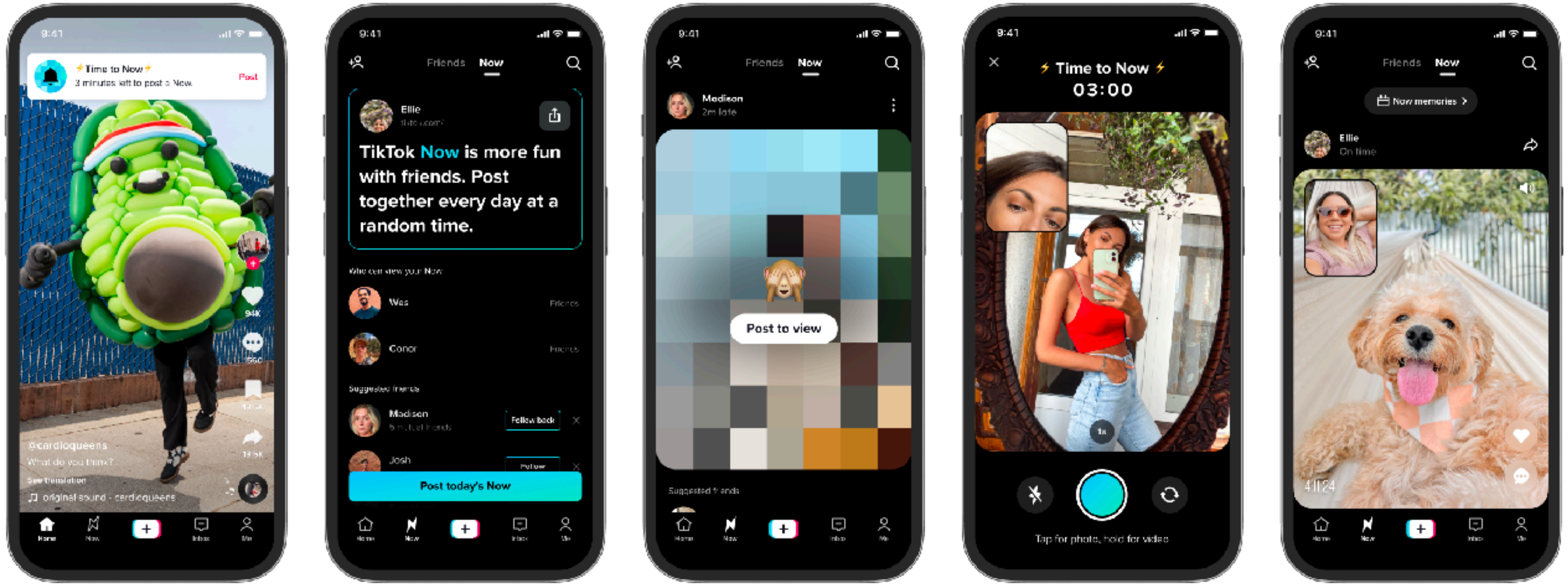
Zhizheng Wu

Outline

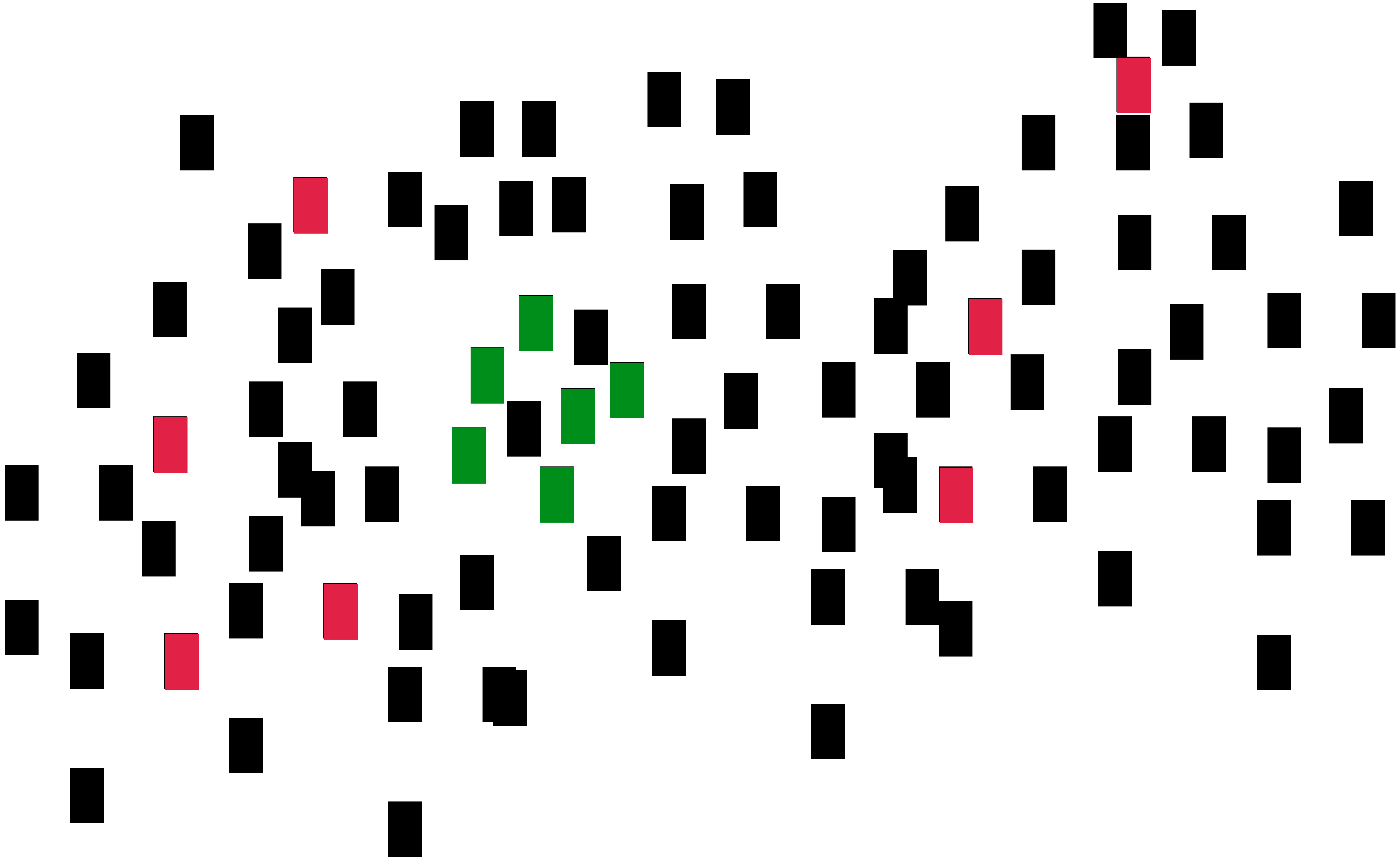
- Machine learning: An example
- Learning paradigms
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Deep learning models
- Loss function and evaluation metrics
- Data is the new oil
- ML in research vs in product


Artificial Intelligence
Mimicking the intelligence or behavioral pattern of humans or any other living entity.
Machine Learning
A technique by which a computer can learn from data, without using a complex set of different rules. This approach is mainly based on training a model from datasets.
Deep Learning
A technique to perform machine learning inspired by our brain’s own network of neurons.

What does Lucas like?

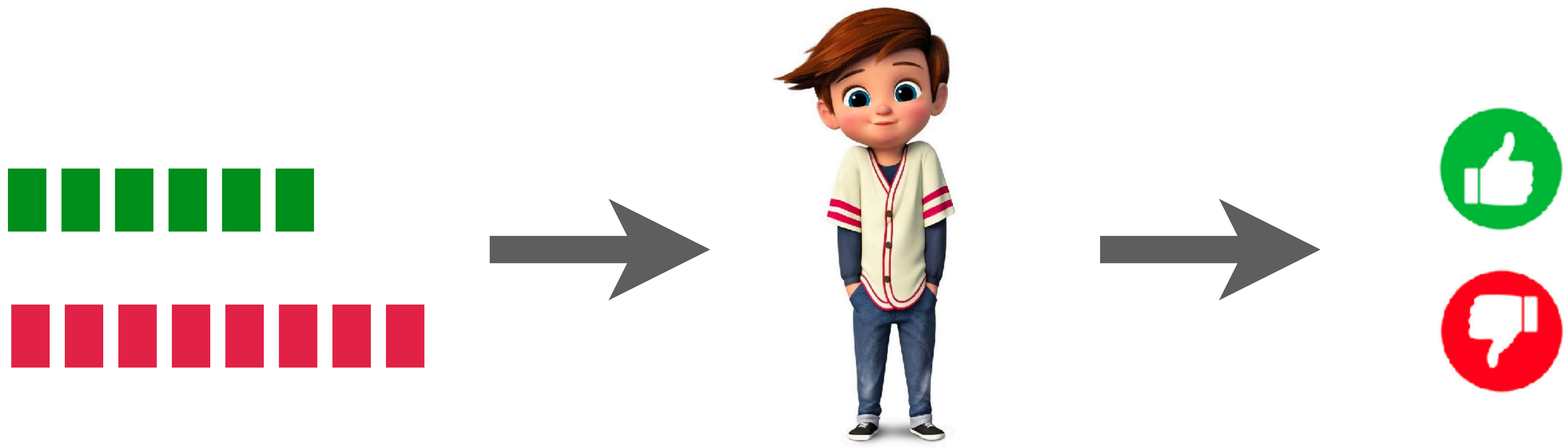


Lucas likes and dislikes



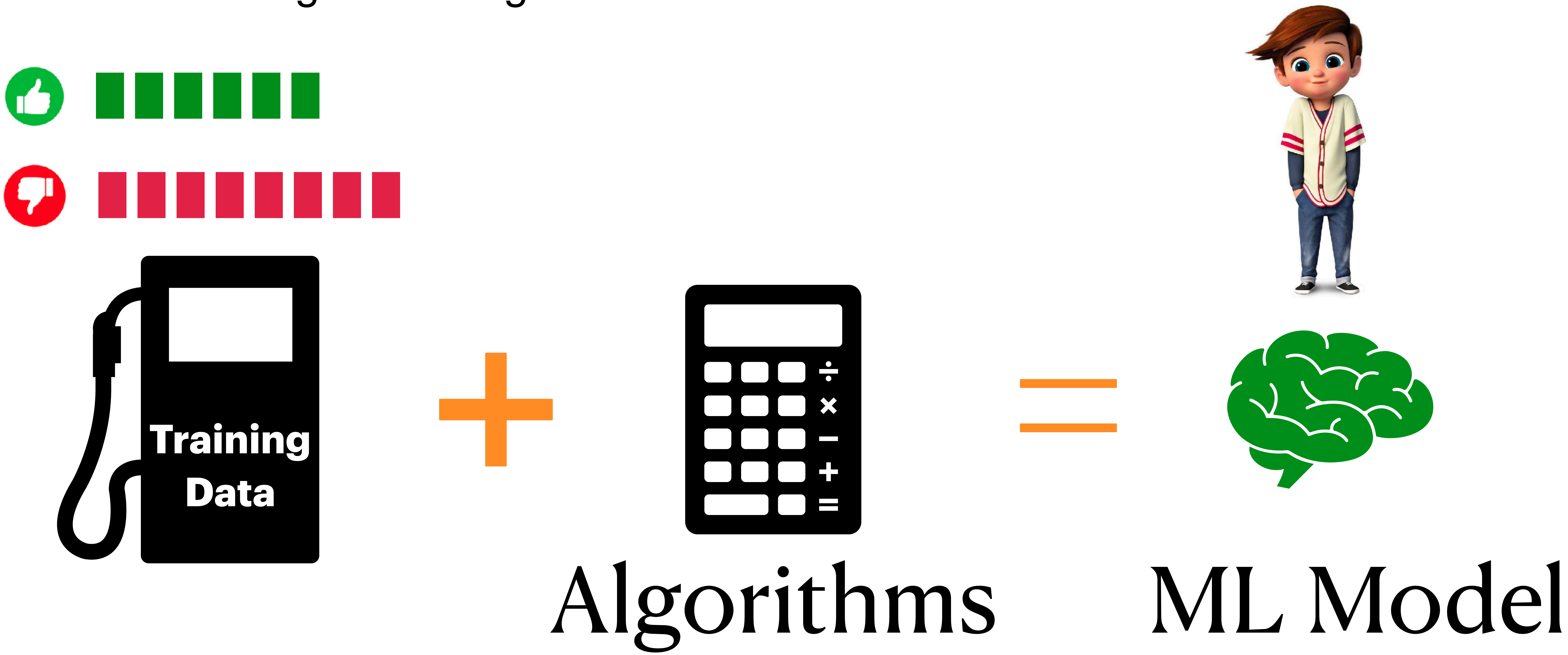
Machine learning to learn the behaviors

- Problem definition: Classify whether a video Lucas likes or dislikes

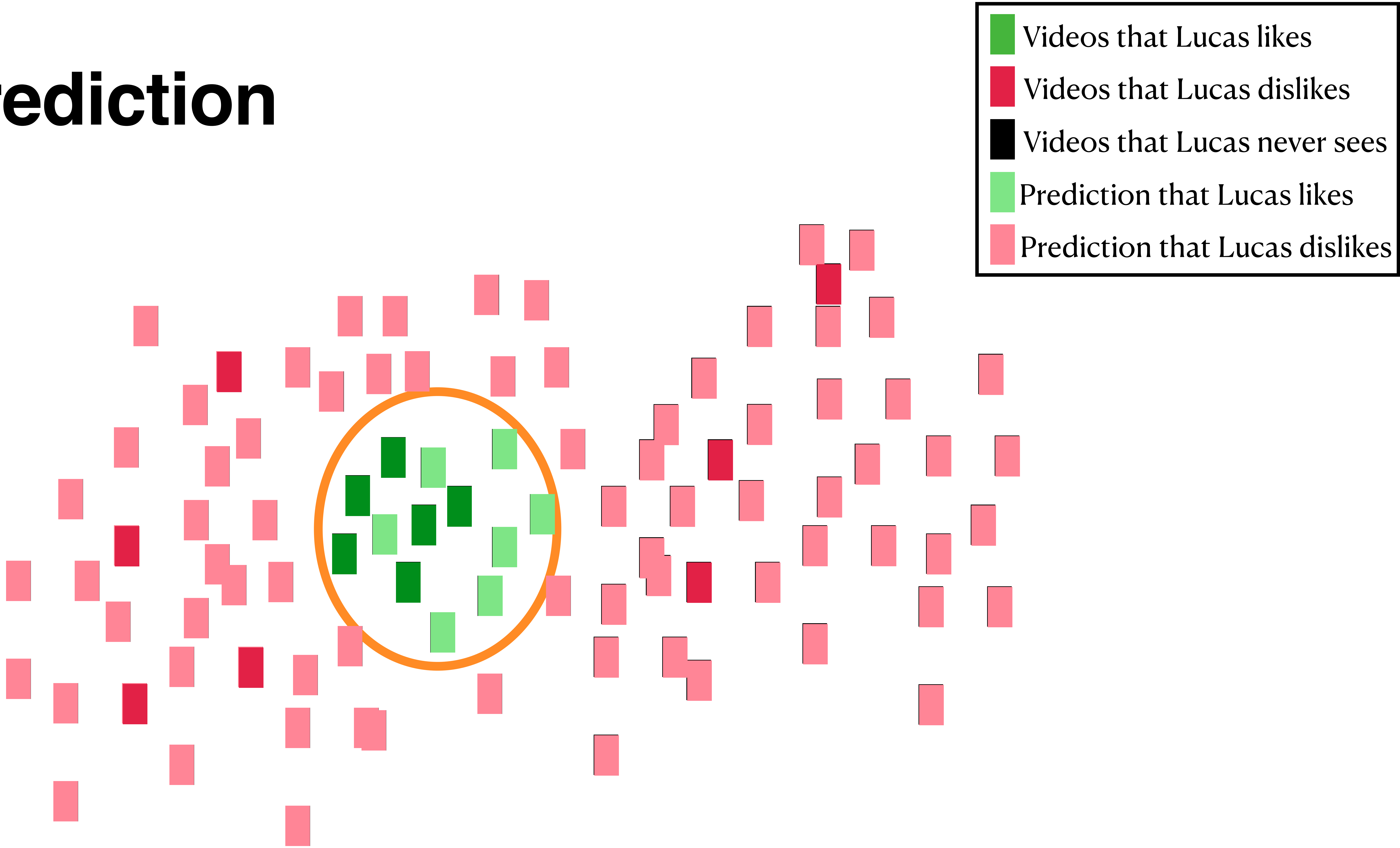


ML model = Data + Algorithms

- ML model = Training data + Algorithms

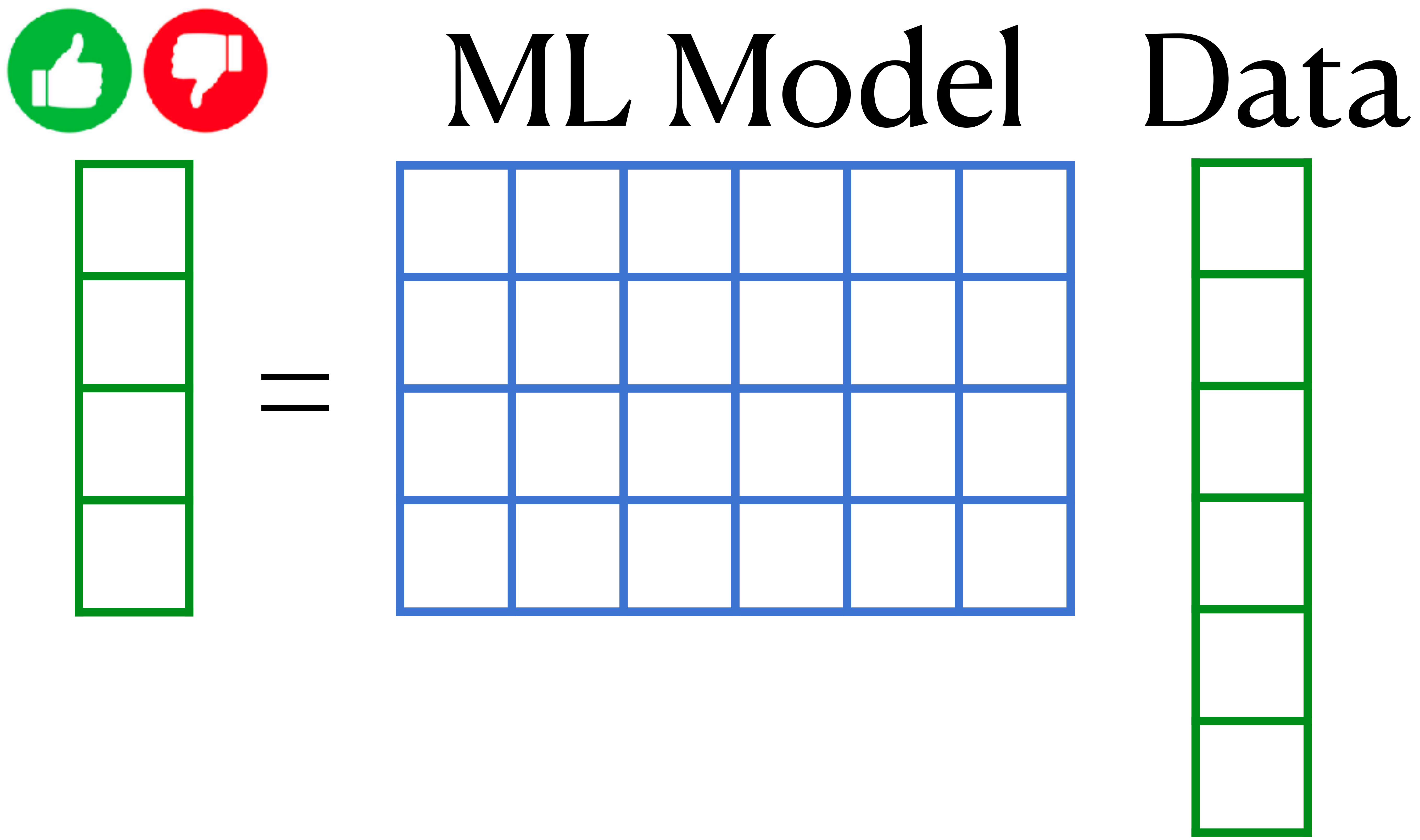


ML prediction



Recommending videos that Lucas might like

ML Model ≅ a transformation function vs Linear algebra

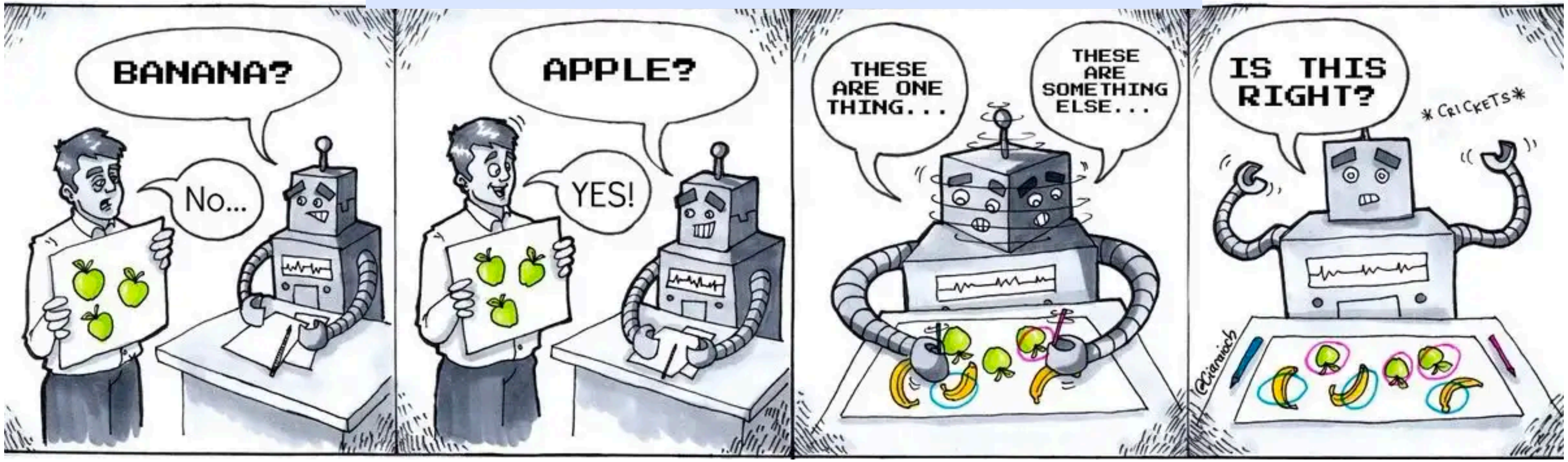


We need data and algorithms to learn the function

Learning paradigms

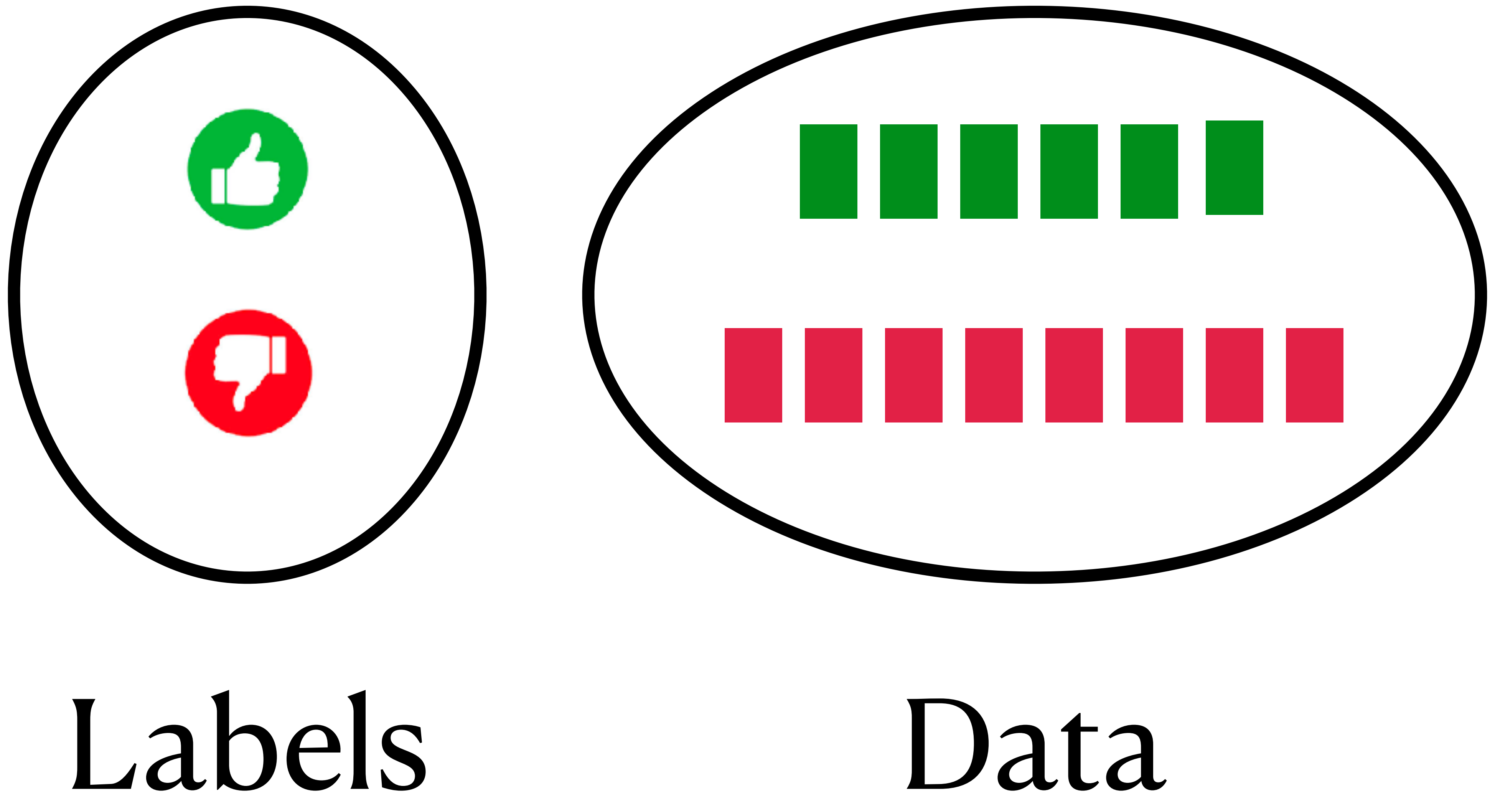
- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised vs Unsupervised learning



Supervised learning

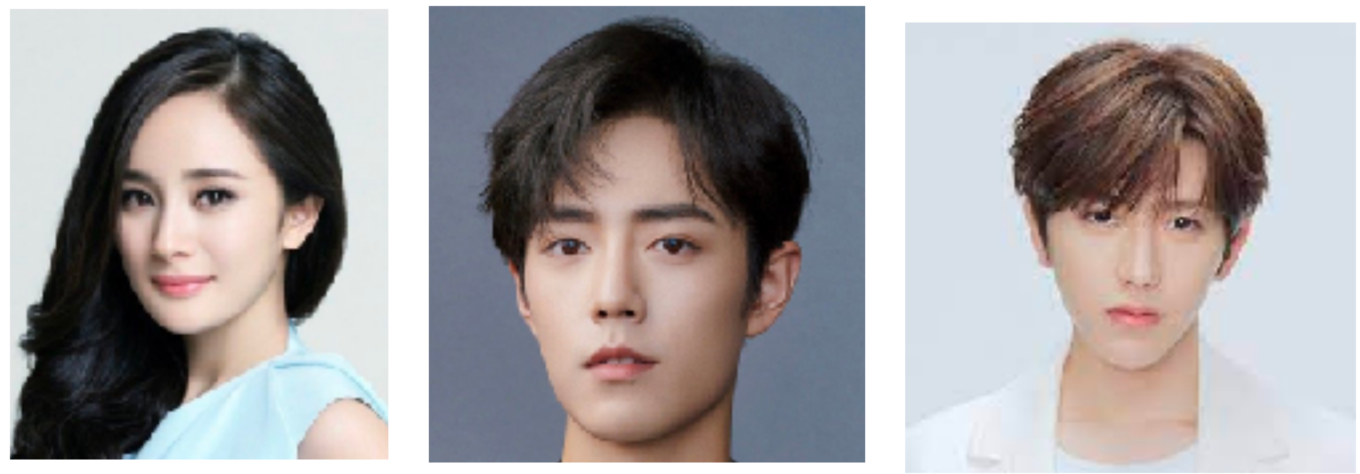
- Each data point consists of features and a label (or multiple labels)



Supervised learning: Label spaces

- Binary classification
  - Yes/No
  - Positive/Negative
- Applications
  - Spam filtering
  - Medical testing
  - etc

- Multi-class classification
  - K labels (K > 2)
- Applications
  - Face recognition
  - Sentiment classification
  - etc



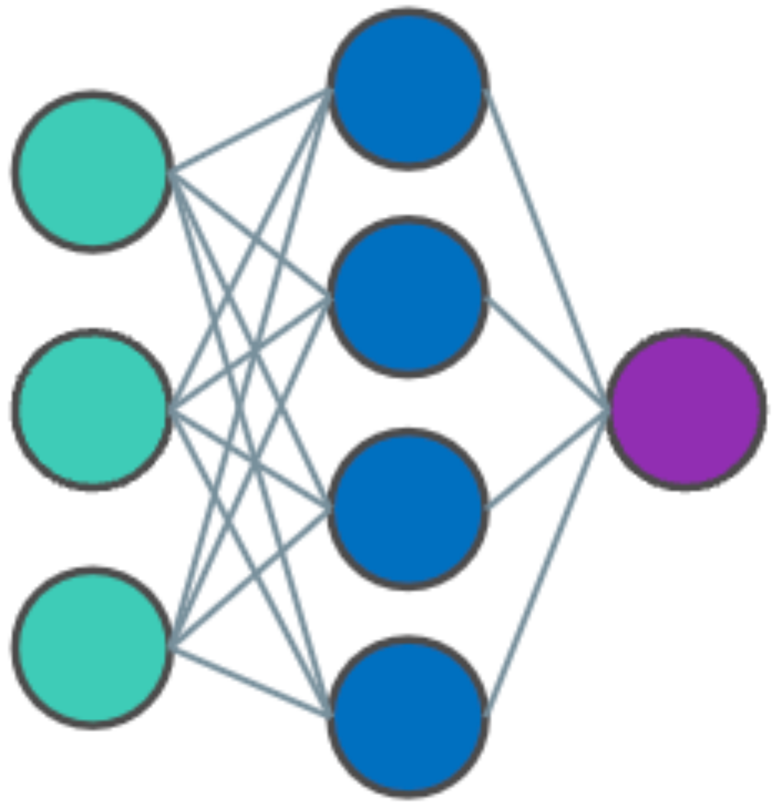
- Regression
  - Continuous real values (e.g. temperature)
- Applications
  - Voice generation
  - Image generation
  - etc

Generated from stable diffusion

Some typical supervised ML models

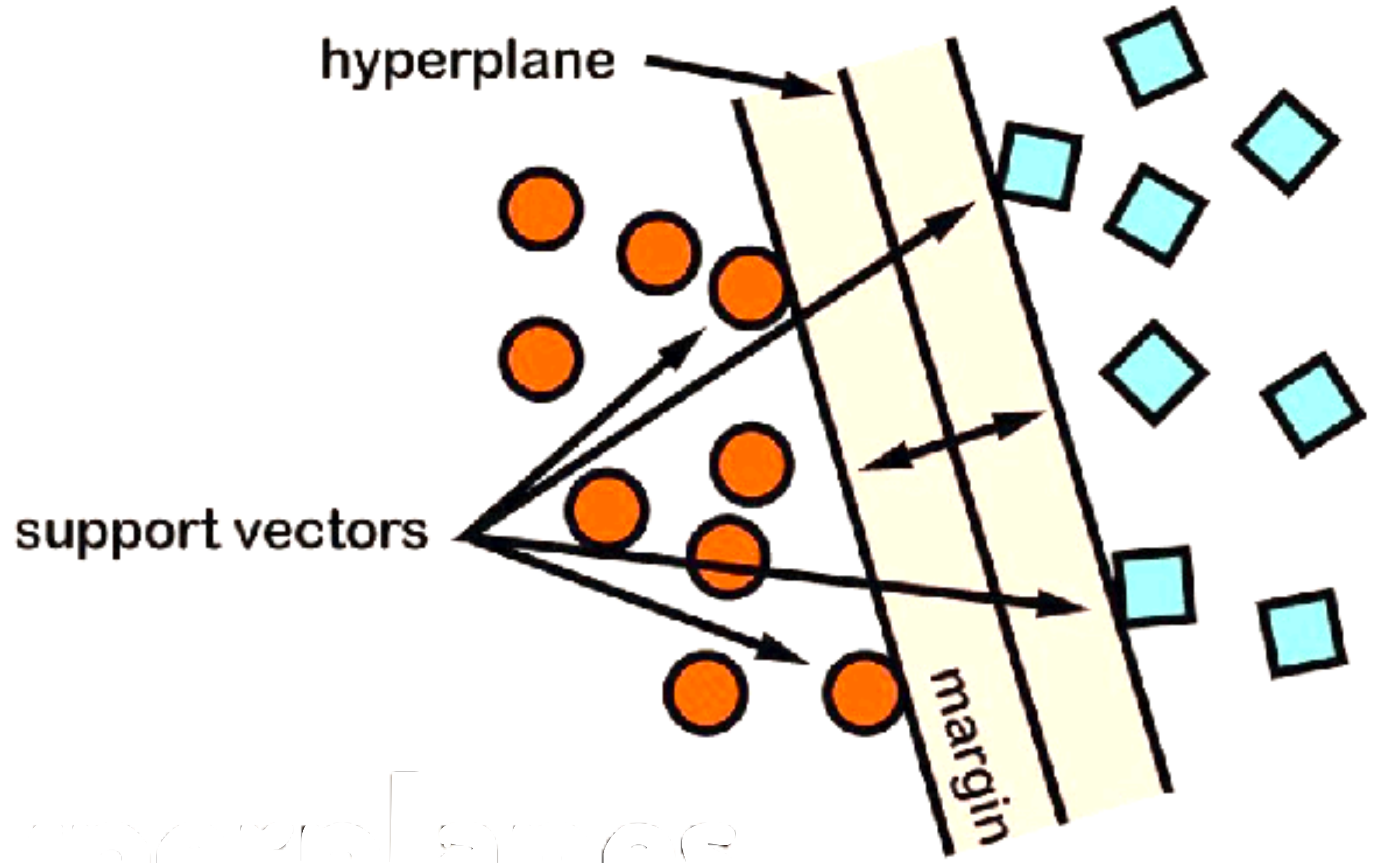
Neural Networks

a type of machine learning model inspired by the structure and function of the human brain



Support Vector Machines

maps training examples to points in a high-dimensional space in order to maximize the distance between the two categories.

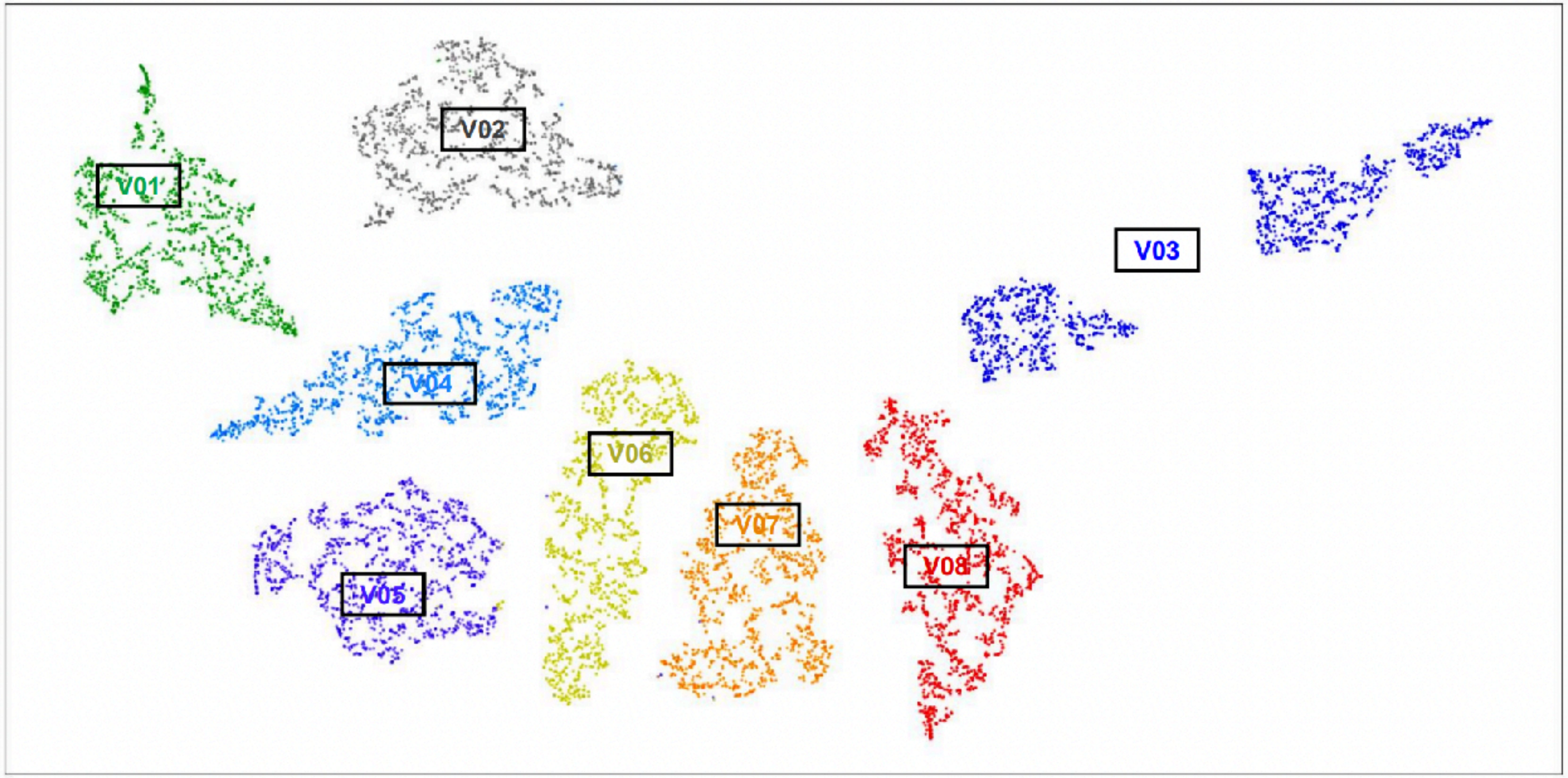


Random Forests

a machine learning method for classification, regression and other tasks that builds multiple decision trees during training

Unsupervised learning

- Analyze and cluster unlabeled datasets to discover hidden patterns or data groupings without the need for human intervention



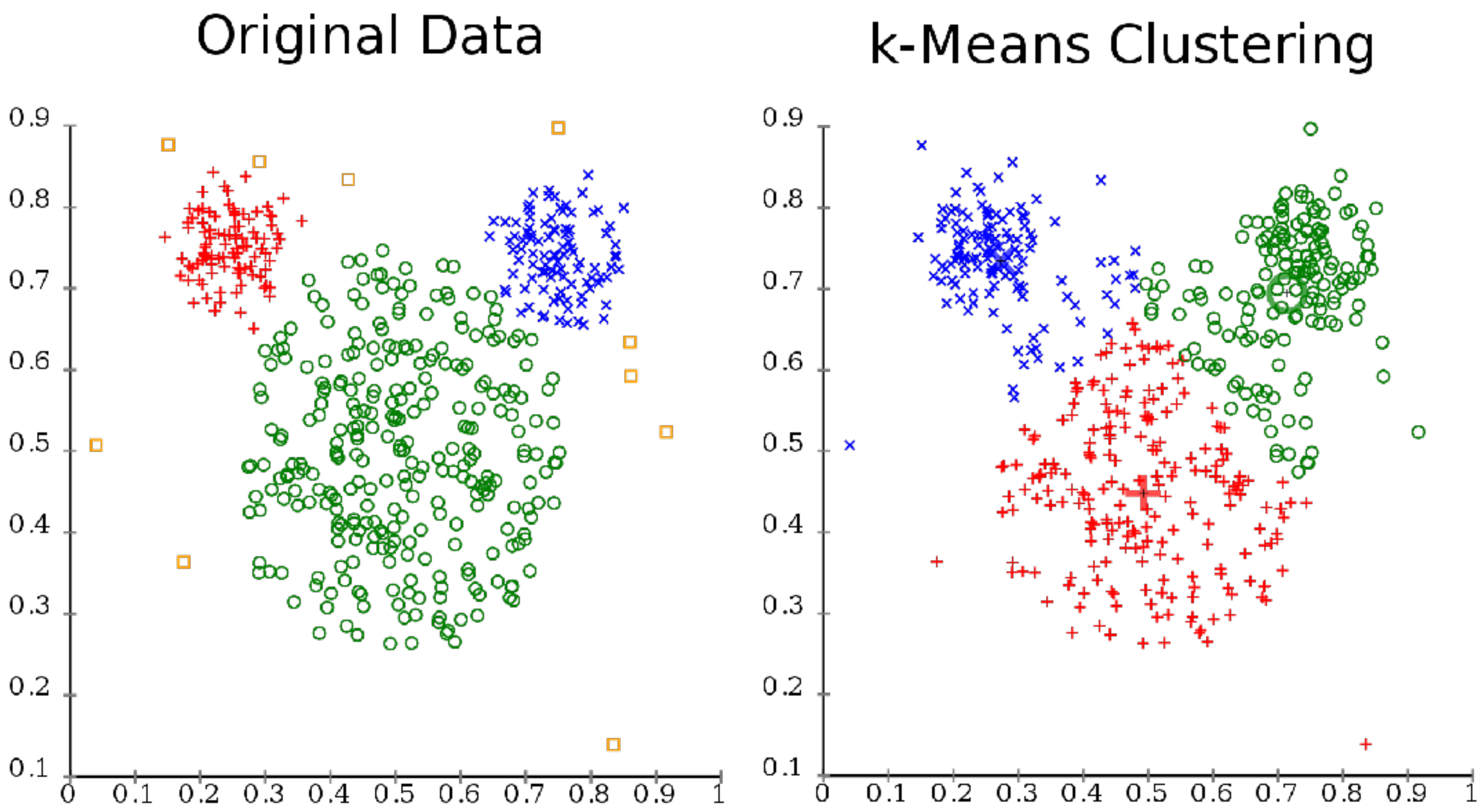
Typical unsupervised ML models

- K-means
  - The K-Means algorithm finds similarities between objects and groups them into K different clusters

- Hierarchical Clustering
  - Hierarchical clustering builds a tree of nested clusters without having to specify the number of clusters

Unsupervised learning: k-means clustering

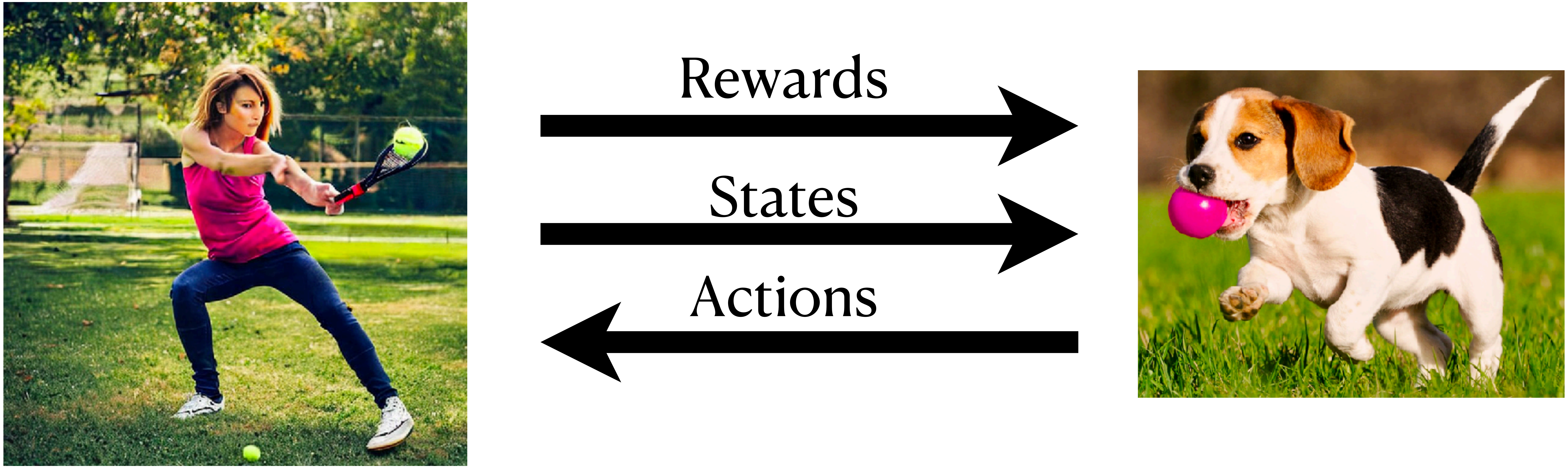
- k-means clustering: group data samples into *k* classes



$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{\mathbf{x},\mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|^2$$

https://en.wikipedia.org/wiki/K-means_clustering

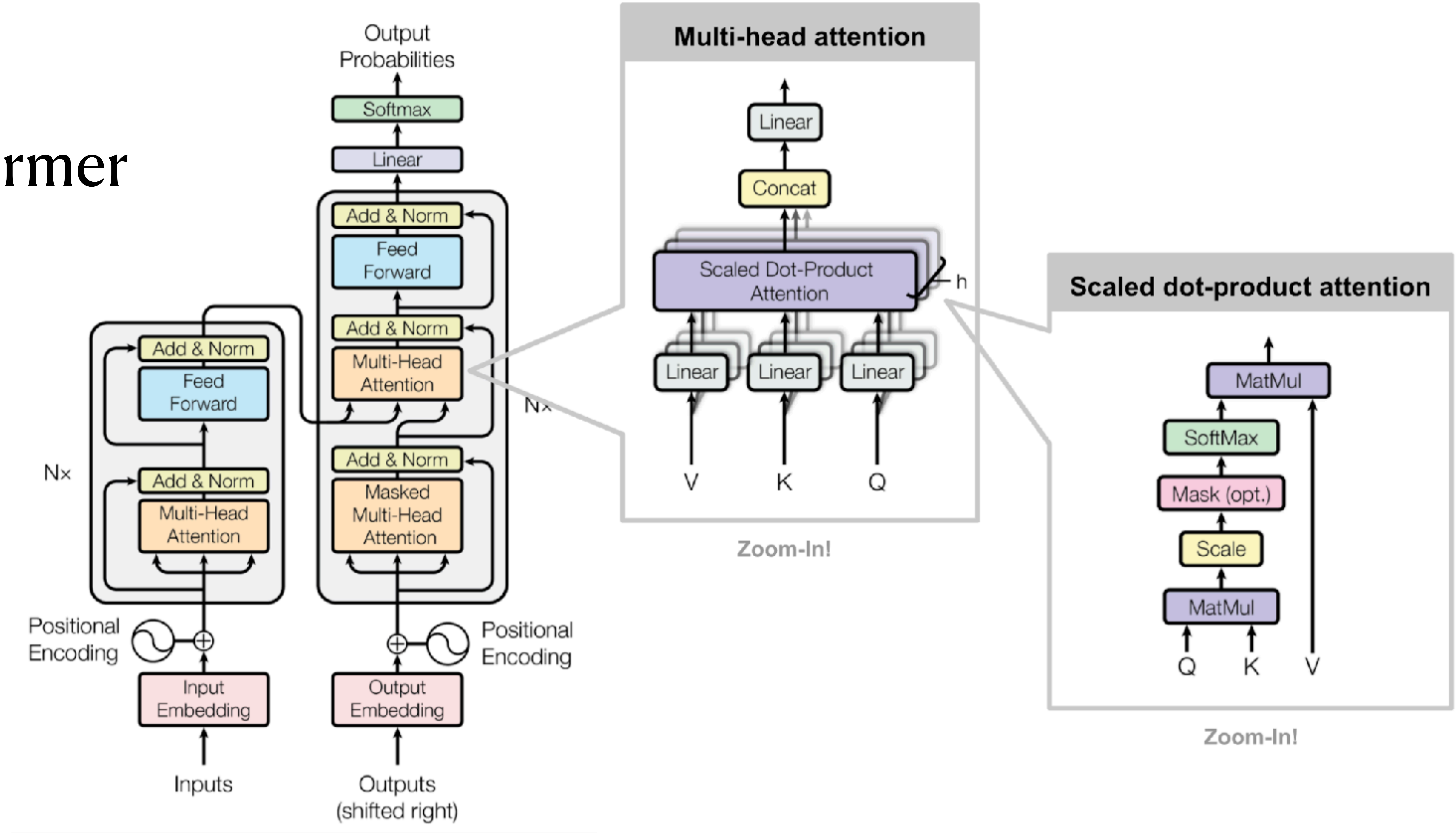
Reinforcement learning



Deep learning models

Deep learning models

Transformer



https://lilianweng.github.io/posts/2020-04-07-the-transformer-family/

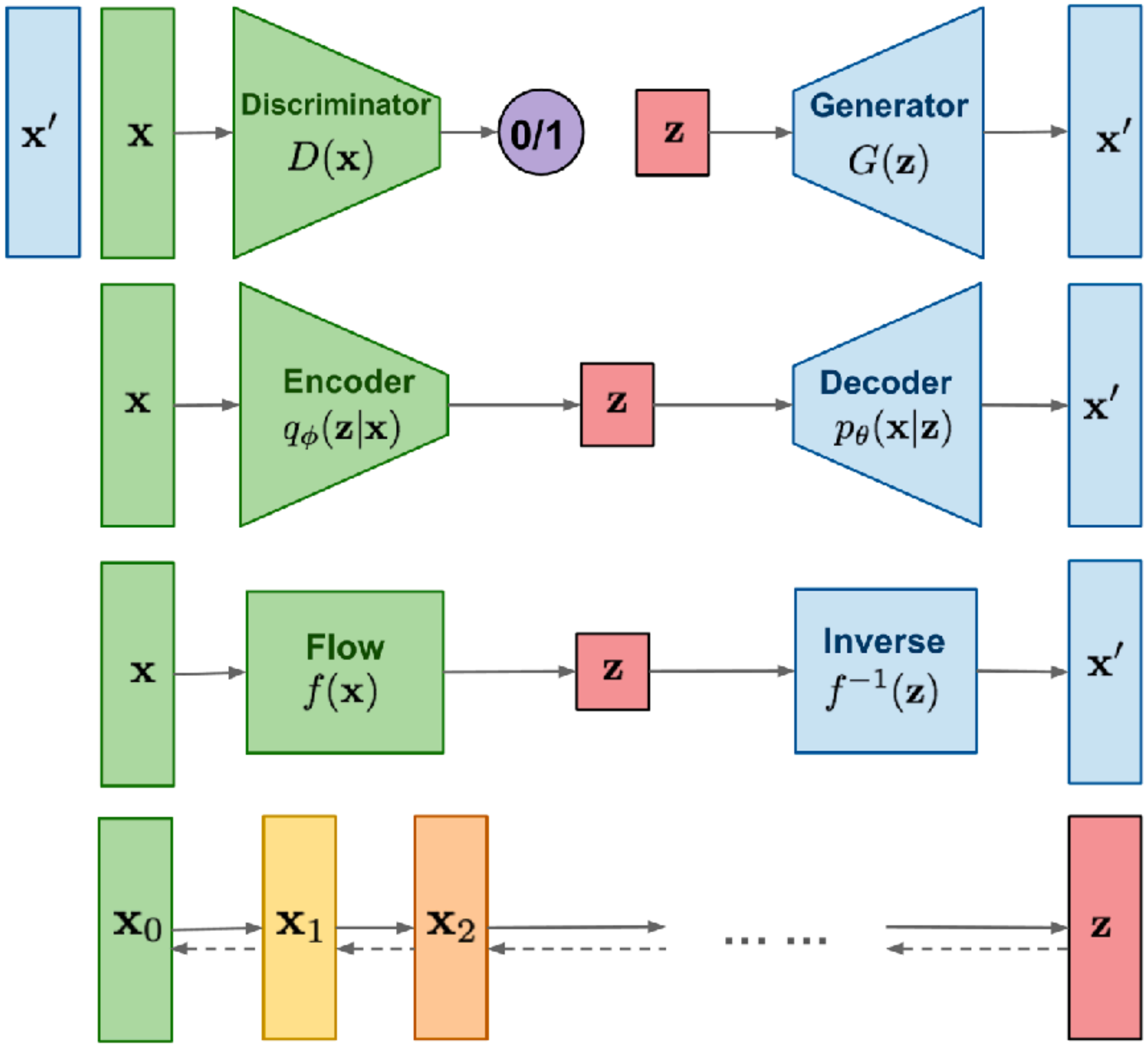
Deep learning generative models

GAN: Adversarial training

VAE: maximize variational lower bound

Flow-based models: Invertible transform of distributions

Diffusion models: Gradually add Gaussian noise and then reverse

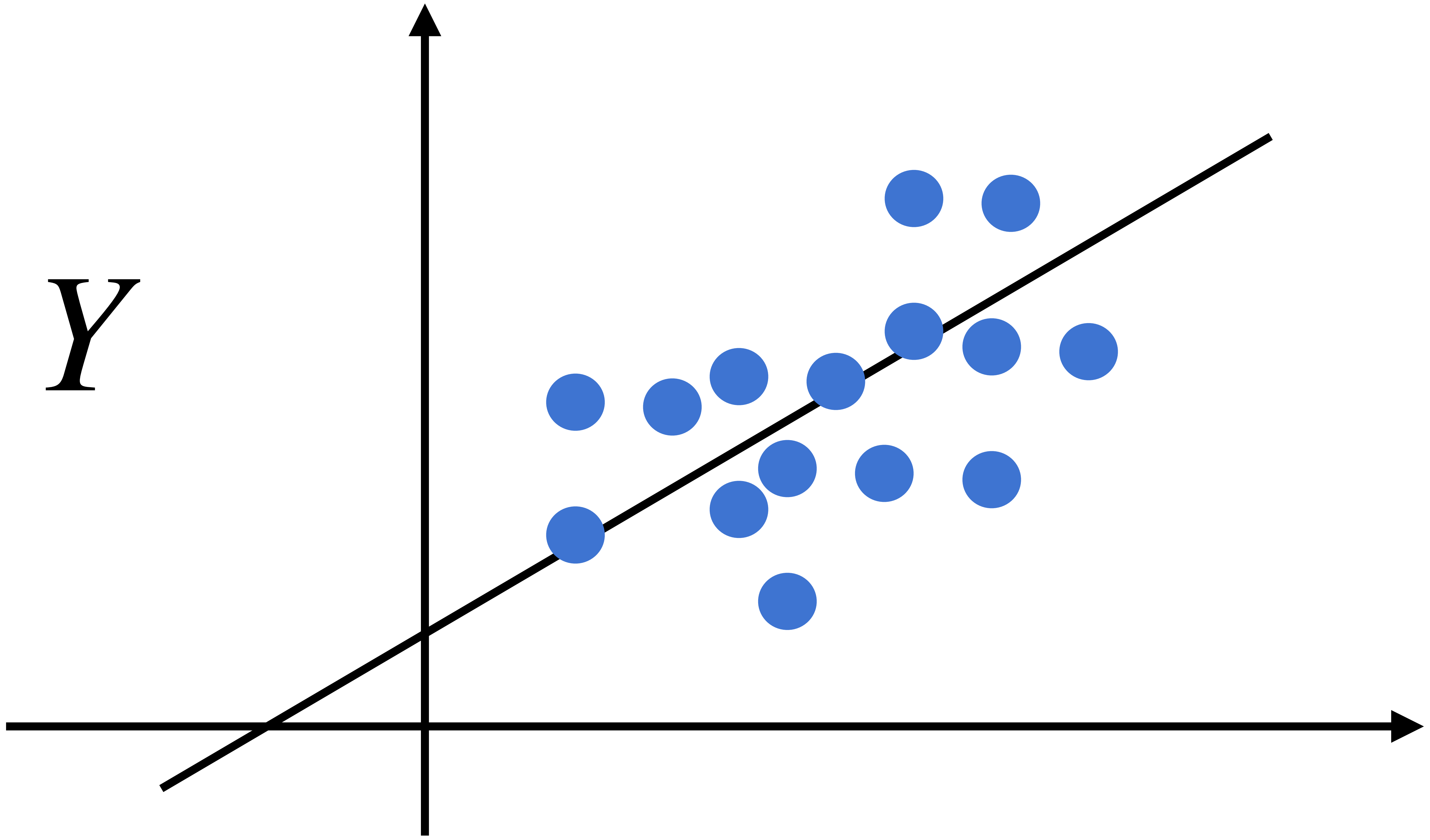


https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Loss function

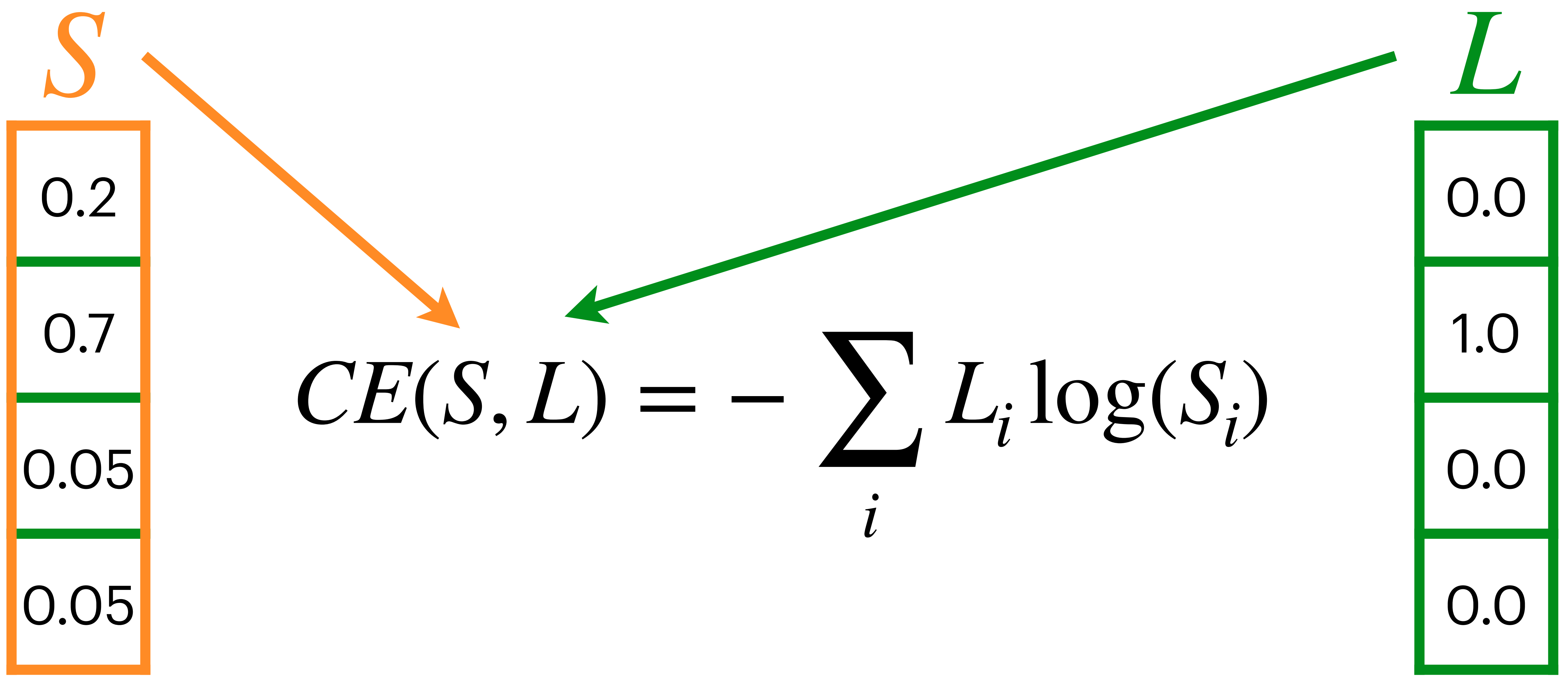
- A method of evaluating how well your algorithm fits/models your dataset

$$\hat{Y} = f(X) \rightarrow Y$$



Loss function

- Cross-entropy loss
  - Usually used in classification tasks



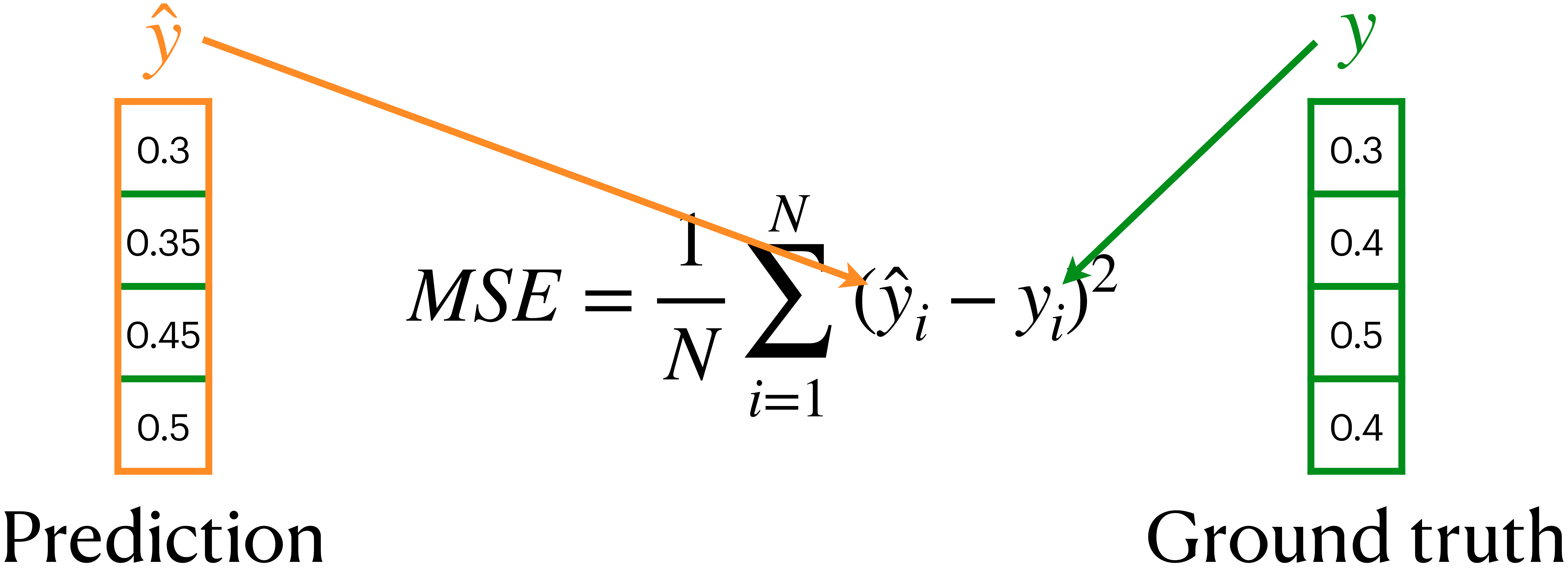
$$CE(S, L) = -\sum_i L_i \log(S_i)$$

Prediction

Ground truth

Loss function: Mean-squared loss

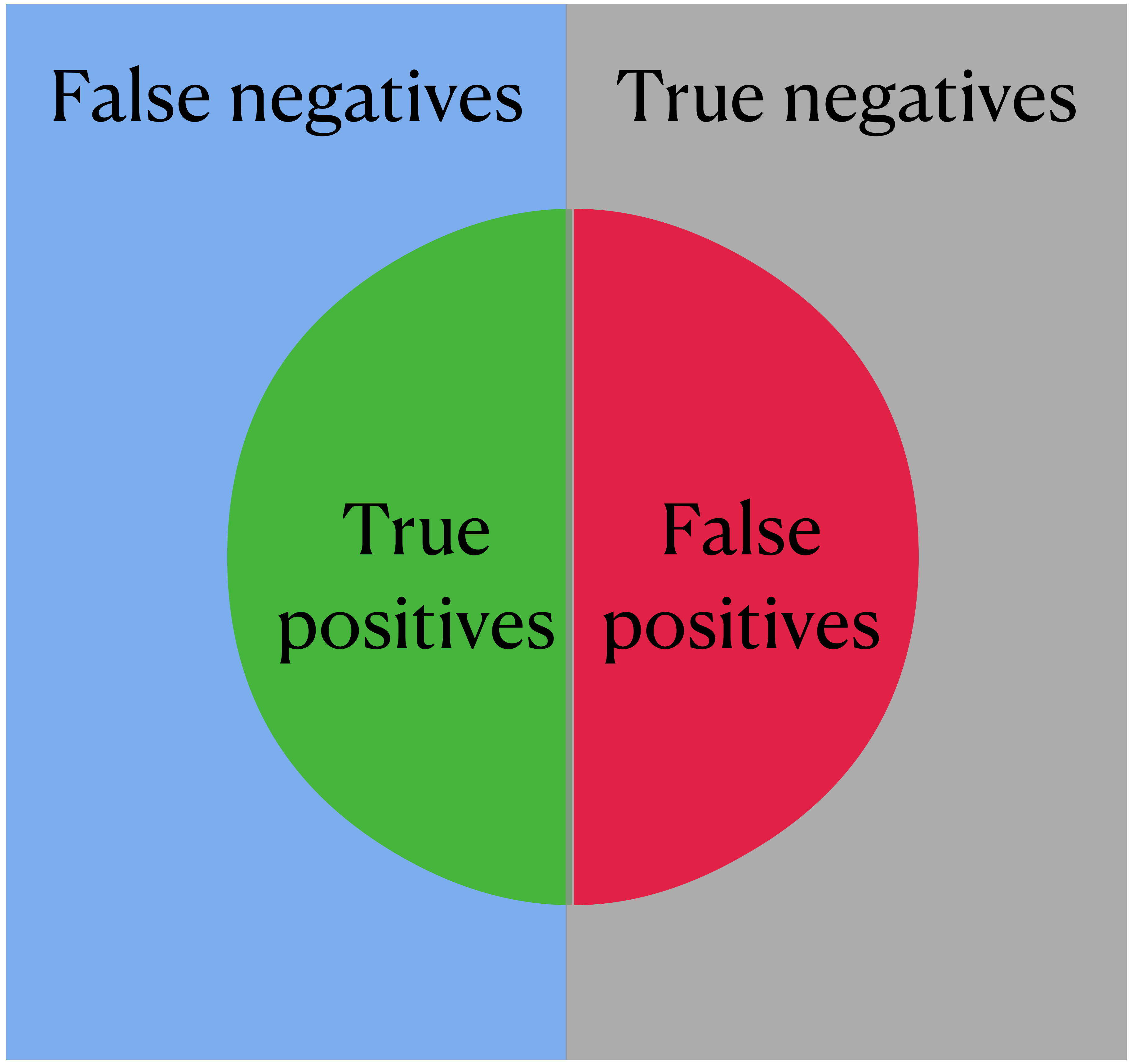
- Mean-squared distance between ground truth and prediction
  - Usually used in regression tasks

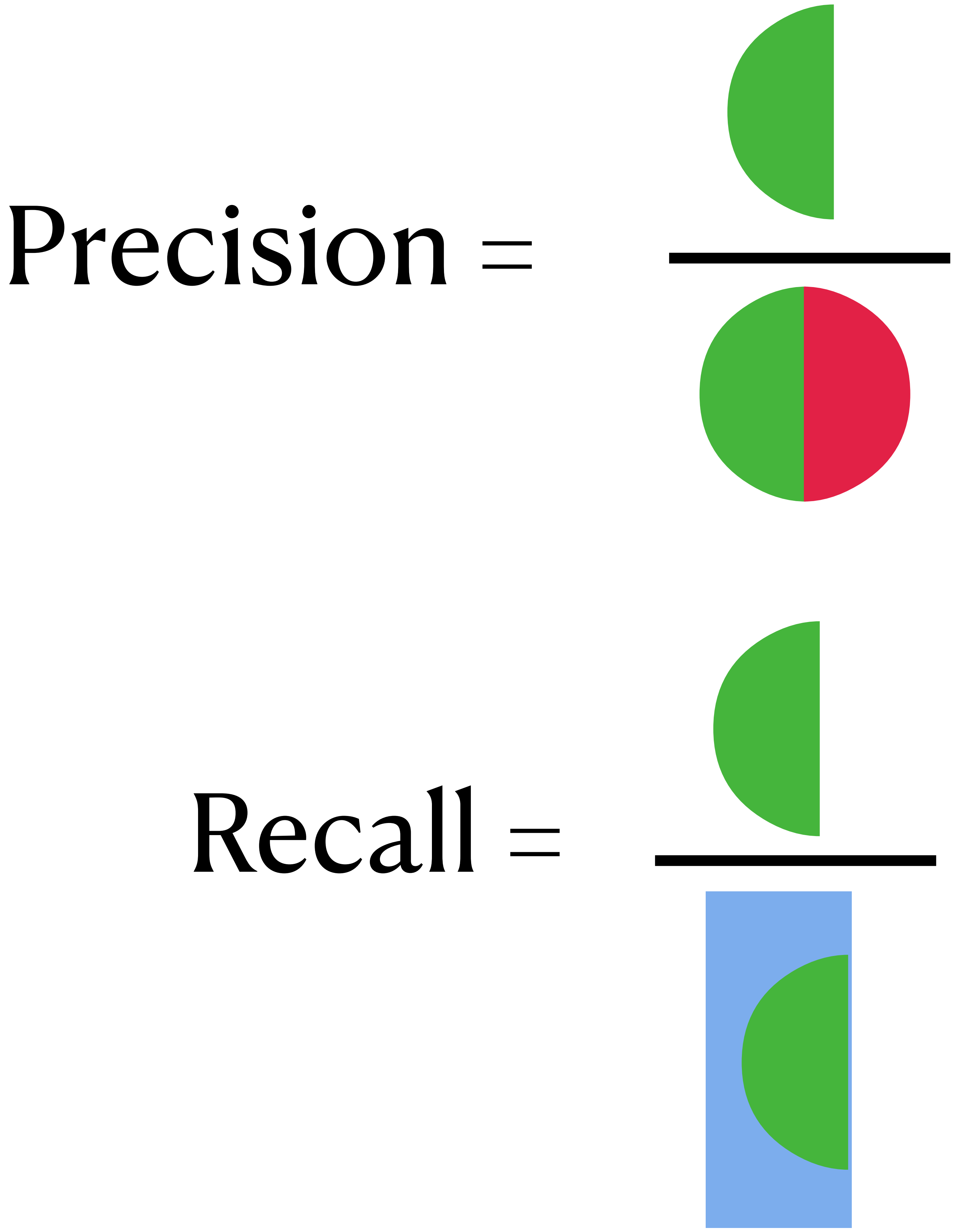


$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Evaluation metrics

- Precision and recall





Evaluation metrics

- *F*-score
  - The harmonic mean of precision and recall
  - F_1 gives equal importance to precision and recall

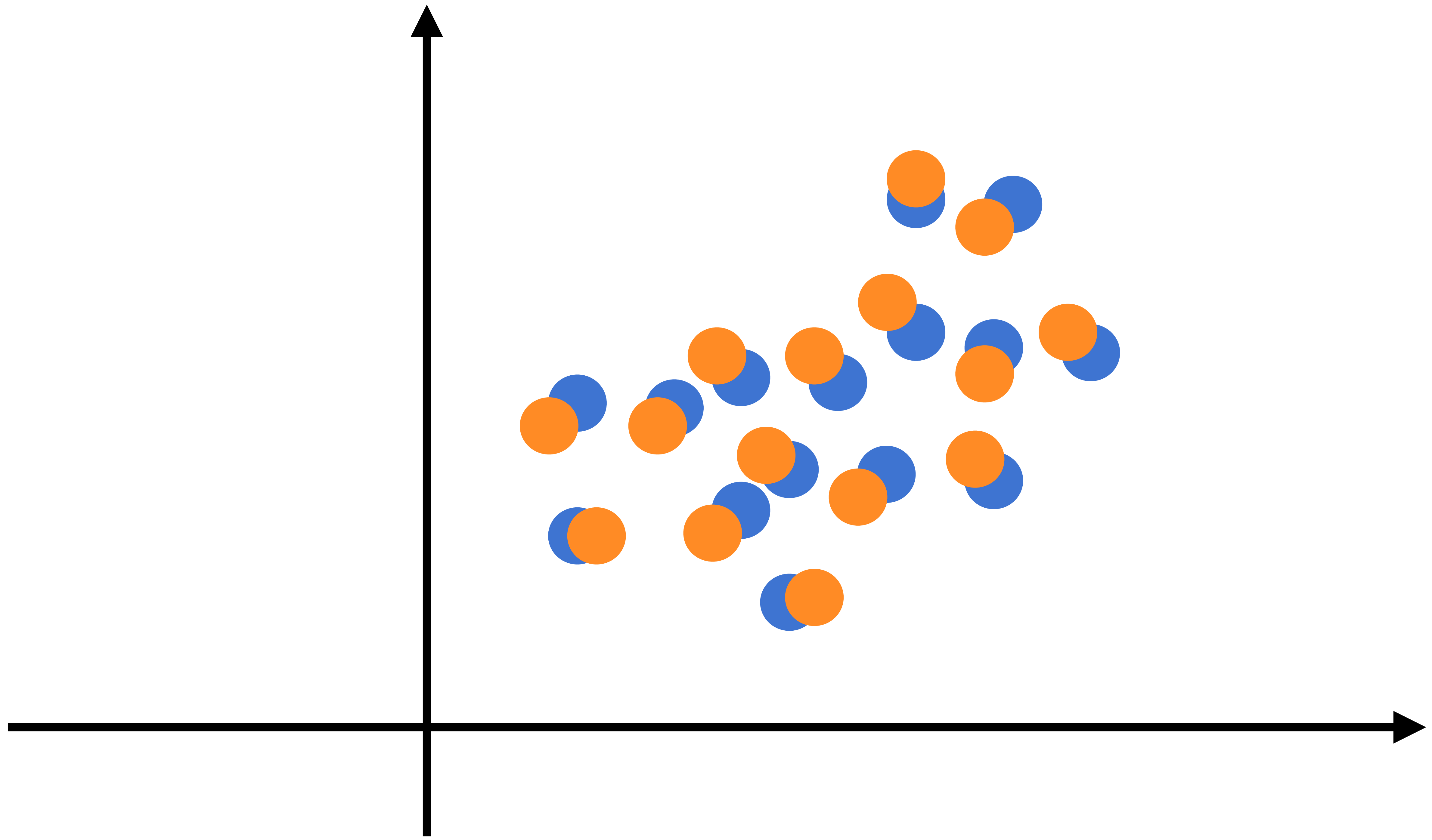
$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

- Accuracy
  - Binary classification $\text{Accuracy} = \dfrac{TP + TN}{TP + TN + FP + FN}$
  - Multi-class classification $\text{Accuracy} = \dfrac{\text{Correct classifications}}{\text{All classification}}$

TP = True positive; FP = False positive; TN = True negative; FN = False negative

Evaluation metrics

- Root Mean Squared Error (RMSE)
  - Usually used for regression tasks

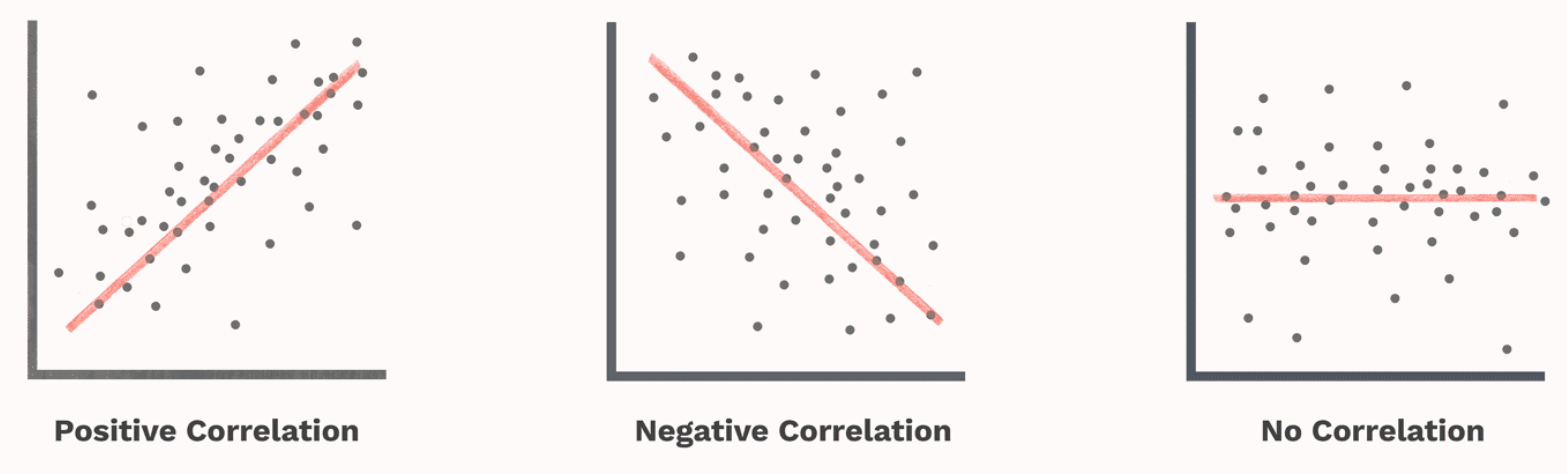


$$RMSE = \sqrt{\frac{\sum_i^N (y_i - \hat{y}_i)^2}{N}}$$

Evaluation metrics

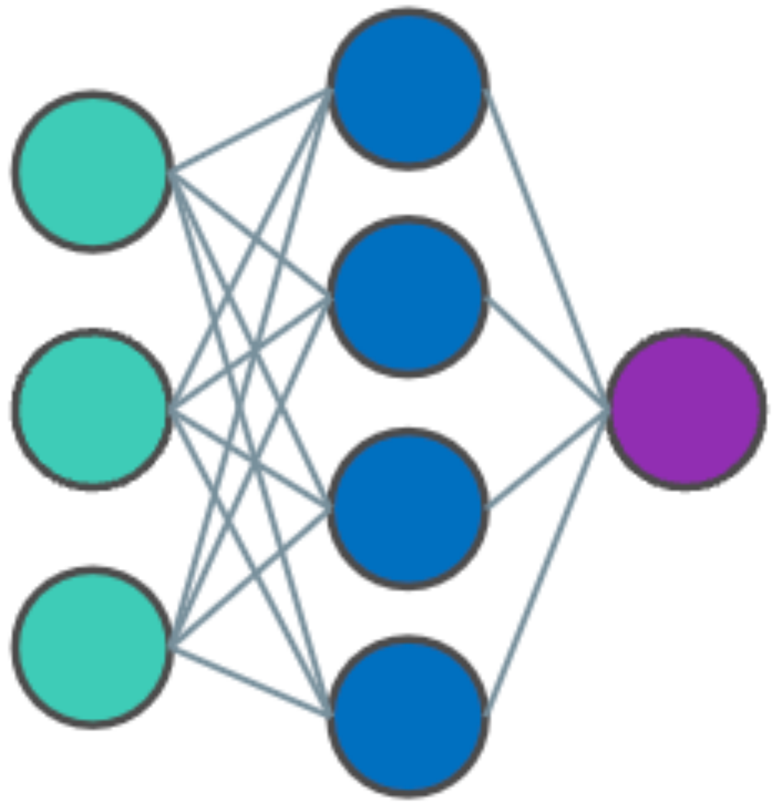
- Pearson correlation coefficient
  - a measure of linear correlation between two sets of data

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



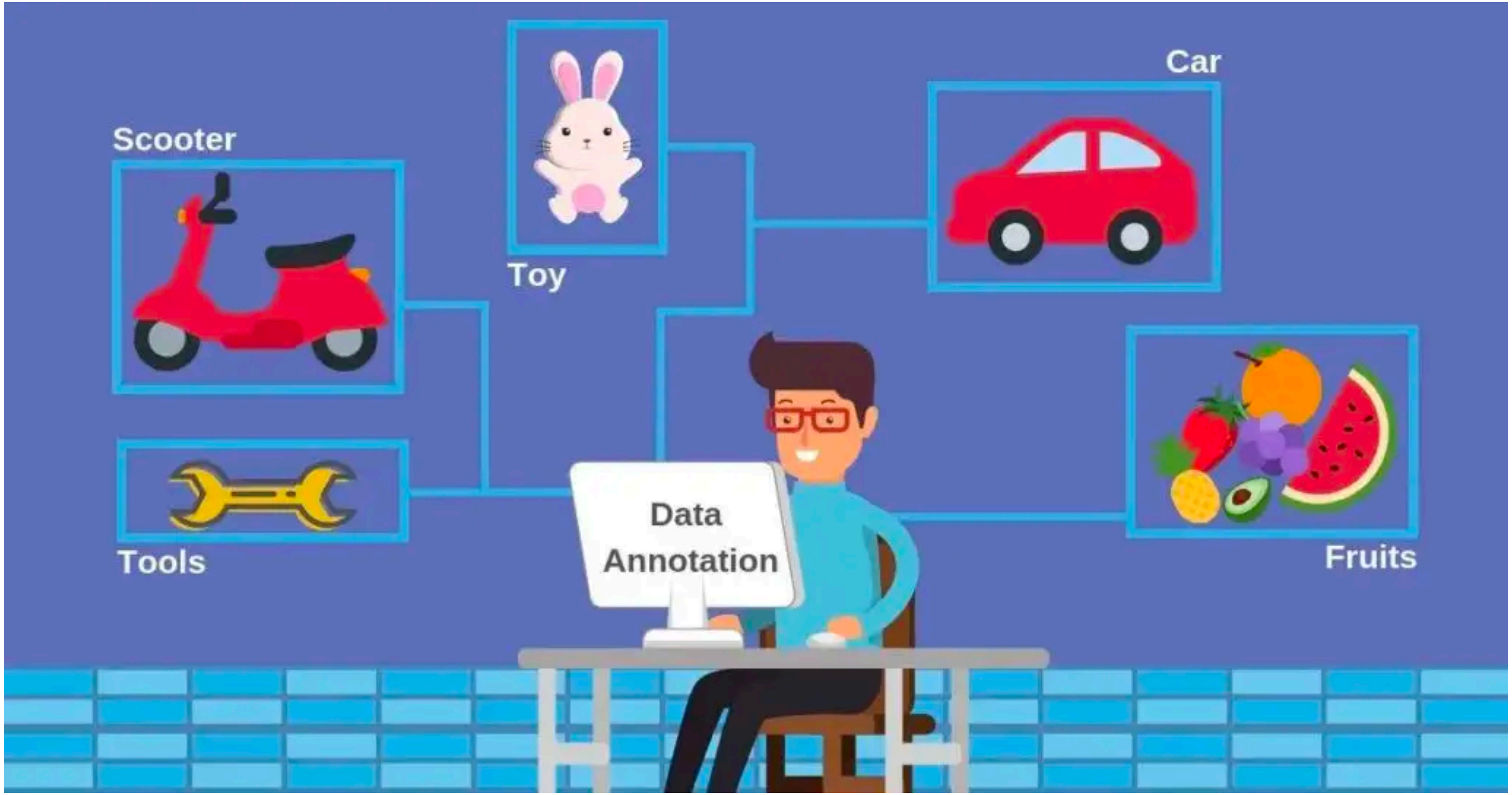
Data is the new oil

+



= ML Model

Labeling data



https://becominghuman.ai/what-is-the-difference-between-data-annotation-and-labeling-in-ai-ml-bcocabeob9d6

Data labeling

Expensive

The cost can be high, especially when specialized subject matter expertise is required

Privacy concern

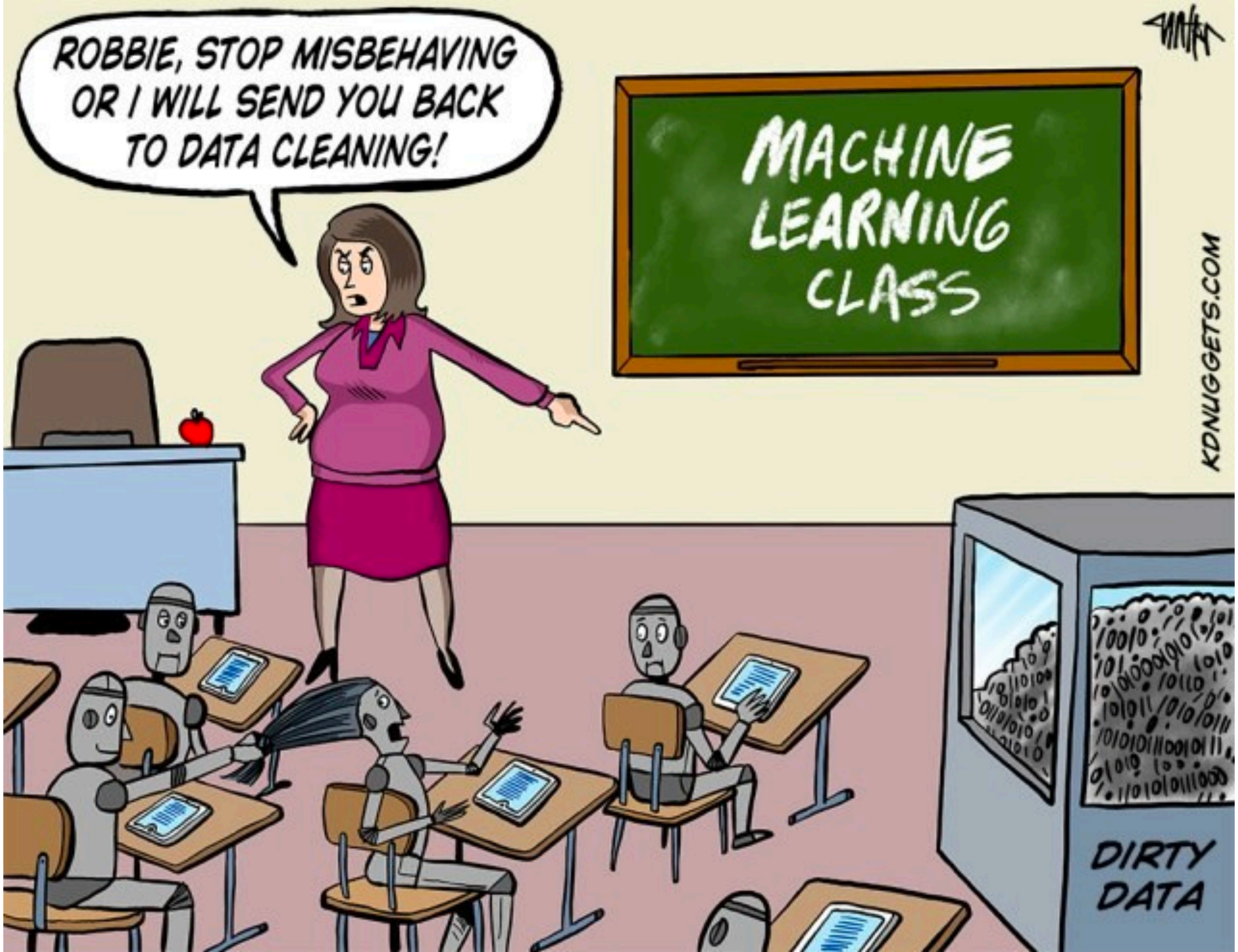
The process is not private because data needs to be shipped to human annotators

Non-adaptive

Any changes to the guidelines necessitate re-labeling the entire dataset, making the process inflexible

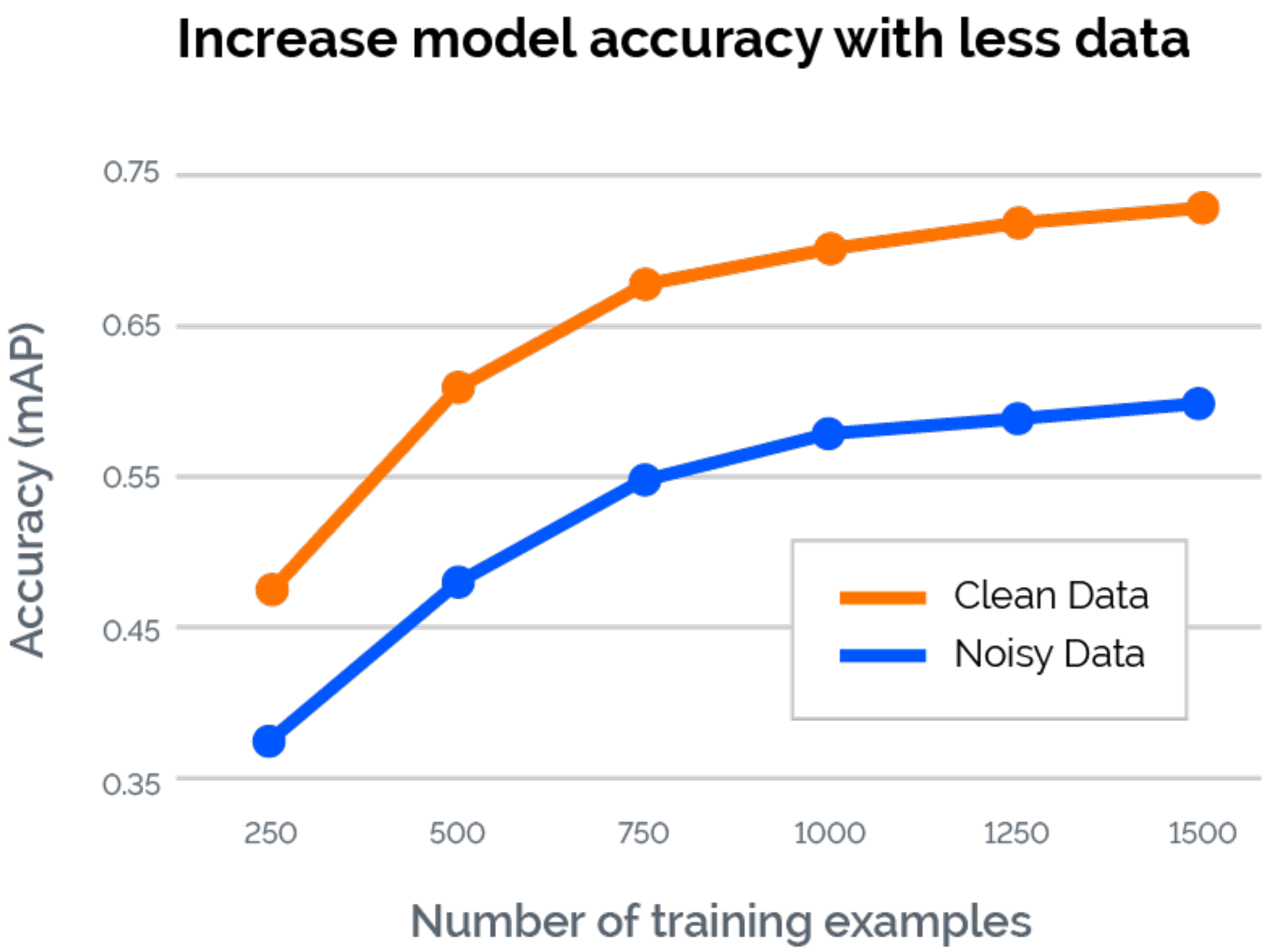
Scalability

The time needed to complete the task scales linearly with the number of labels required, making it difficult to handle large datasets


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
KDNUGGETS.COM
DIRTY DATA

Focusing on high-quality data that is consistently labeled would unlock the value of AI for sectors such as health care, government technology, and manufacturing

- Andrew Ng



https://mitsloan.mit.edu/ideas-made-to-matter/why-its-time-data-centric-artificial-intelligence

Machine learning in research

ML is cool

Everyone is working on ML! Me too!

Public data

Someone has already prepared data for me!

Fancy model

Read some papers, a crazy idea, train a **fancy** model

Paper

Results beat baseline, submit a paper, move on!

News

CSC3160/ MDS6002 students **saved the world**!

ML in product: Expectation

Collect data

Unlimited budget

Train model

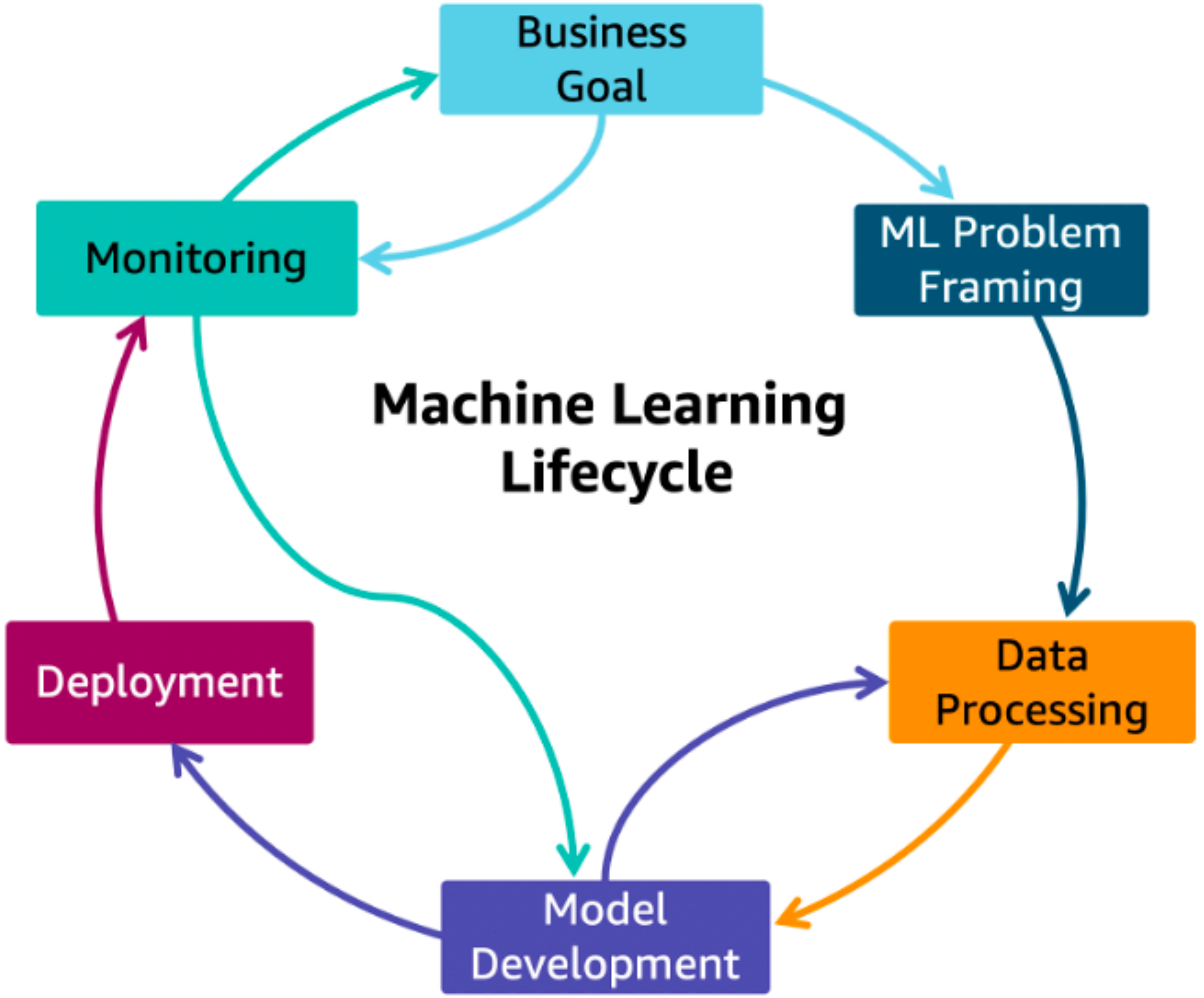
Unlimited GPU

Deploy model

Unlimited computation resources

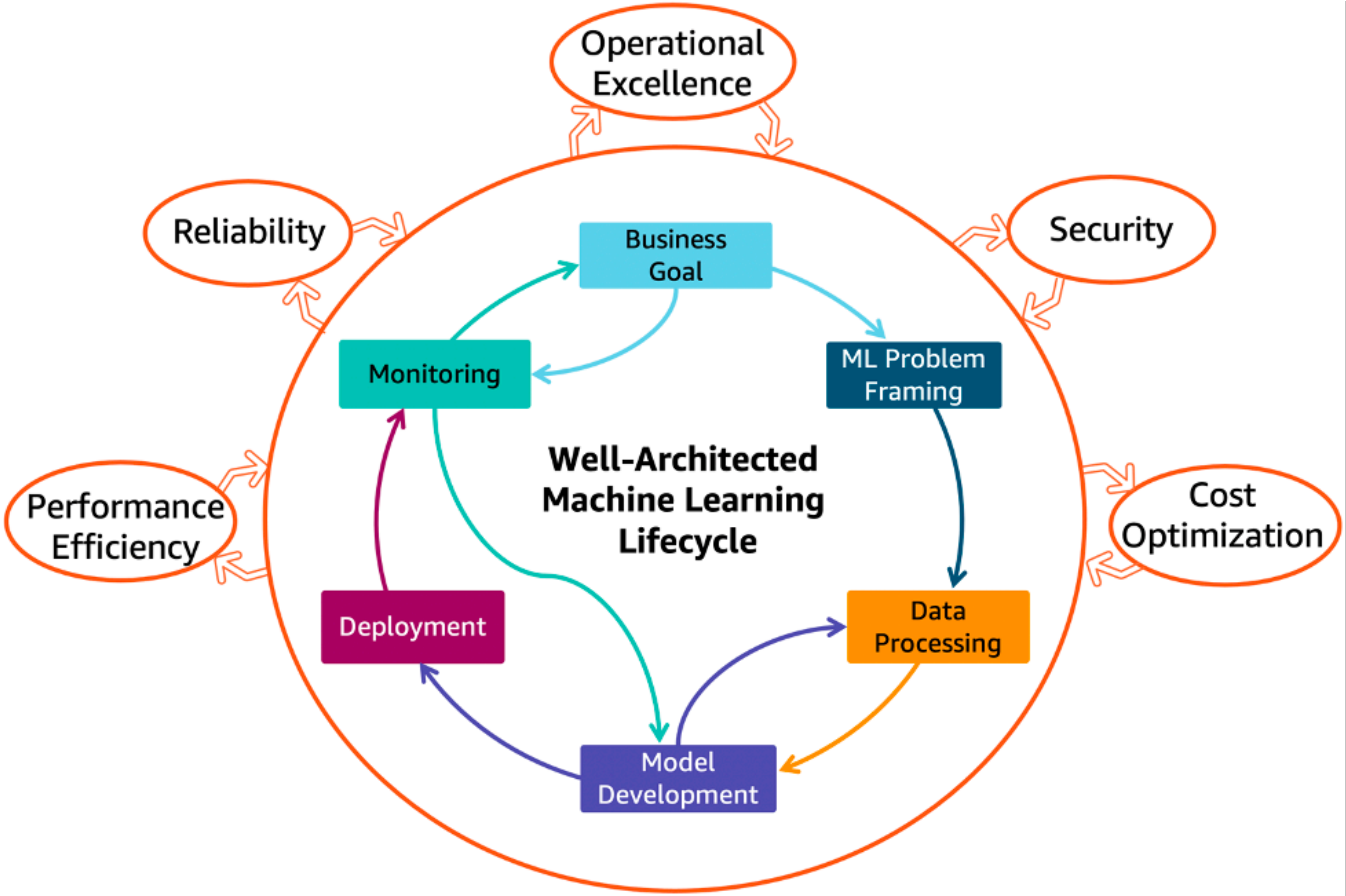
Machine learning in production: Reality

Machine learning lifecycle



https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/well-architected-machine-learning-lifecycle.html

Machine learning lifecycle



https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/well-architected-machine-learning-lifecycle.html

ML in product: Stakeholders

ML team

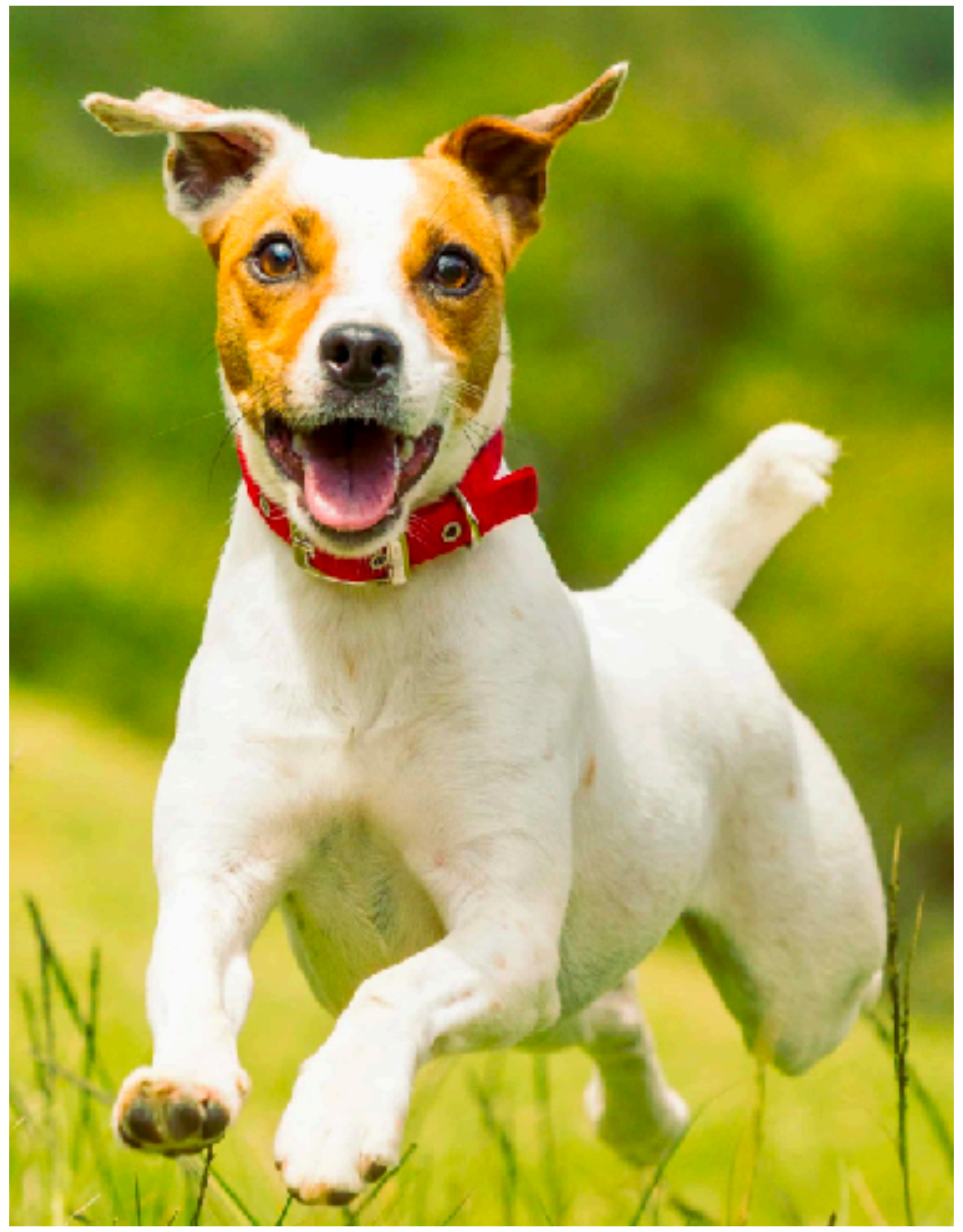
Fancy model
Highest accuracy

Sales

More clients
More revenue

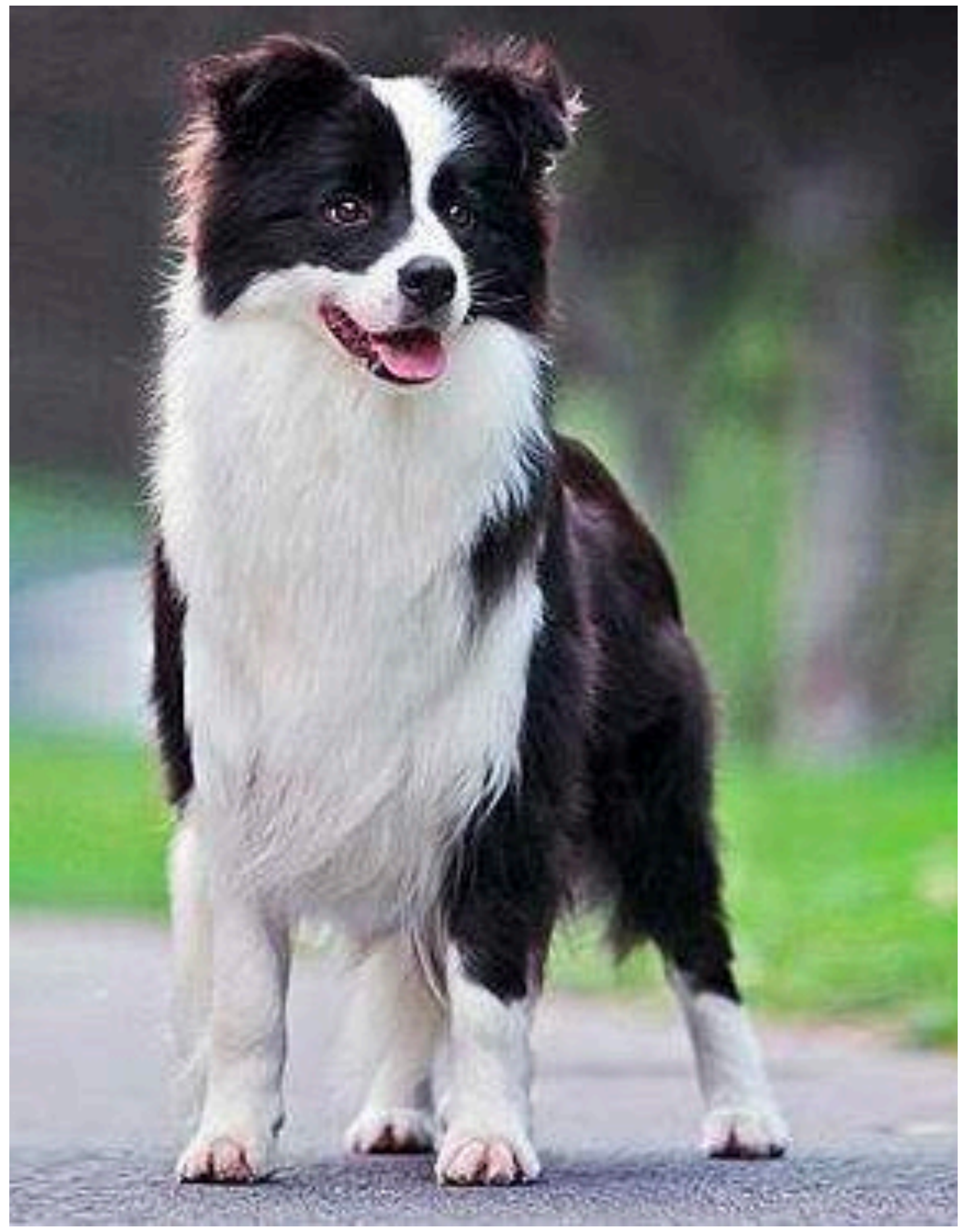
Product

Fastest inference
Reliability, interpretability



Management

Low cost, more profit!
= laying off ML team



Useful resources

- Open-source projects
  - PyTorch
  - TensorFlow

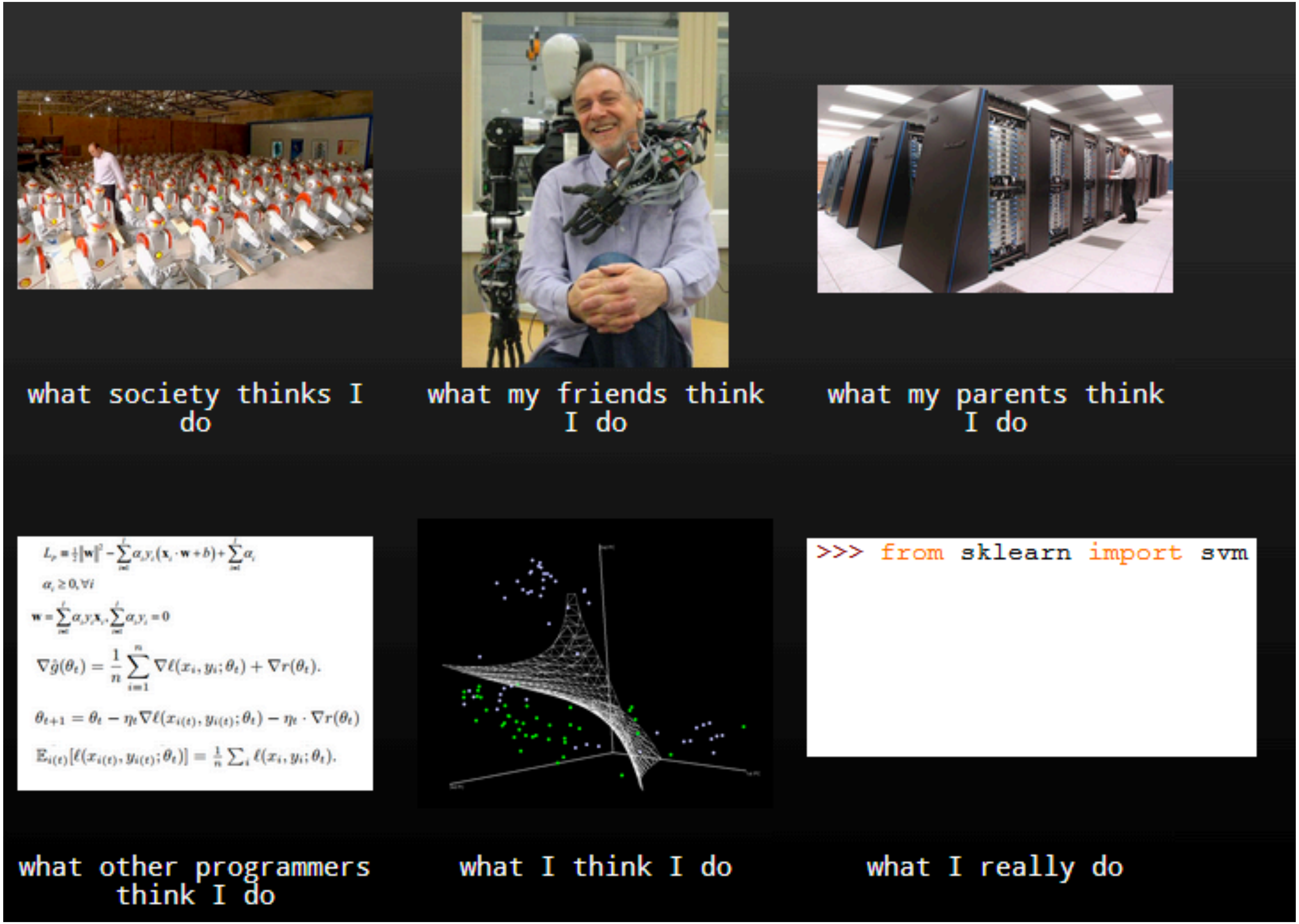
- Platforms
  - Colab: https://colab.research.google.com/
  - HuggingFace: https://huggingface.co/

- Cool demos
  - ChatGPT: https://chat.openai.com/
  - Whisper: https://openai.com/blog/whisper/

Suggestions by ChatGPT

Here are a few suggestions for students who want to learn machine learning:

1. Start by learning the basics of programming and math. Machine learning requires a strong foundation in programming and math, especially linear algebra and calculus.
2. Choose a programming language and start coding. There are many programming languages that are popular for machine learning, such as Python, R, and Java. Pick one and start writing code.
3. Take an online course or get a degree in machine learning. There are many online courses and degree programs that can teach you the concepts and techniques of machine learning.
4. Practice, practice, practice. The best way to learn machine learning is to apply it to real-world problems. Look for open-source datasets and try to build your own machine learning models.
5. Stay up to date with the latest developments in the field. Machine learning is a rapidly evolving field, with new techniques and tools being developed all the time. Stay current by reading papers and blogs, attending conferences, and following experts in the field.



Credit:
Harrison Kinsley

Thanks