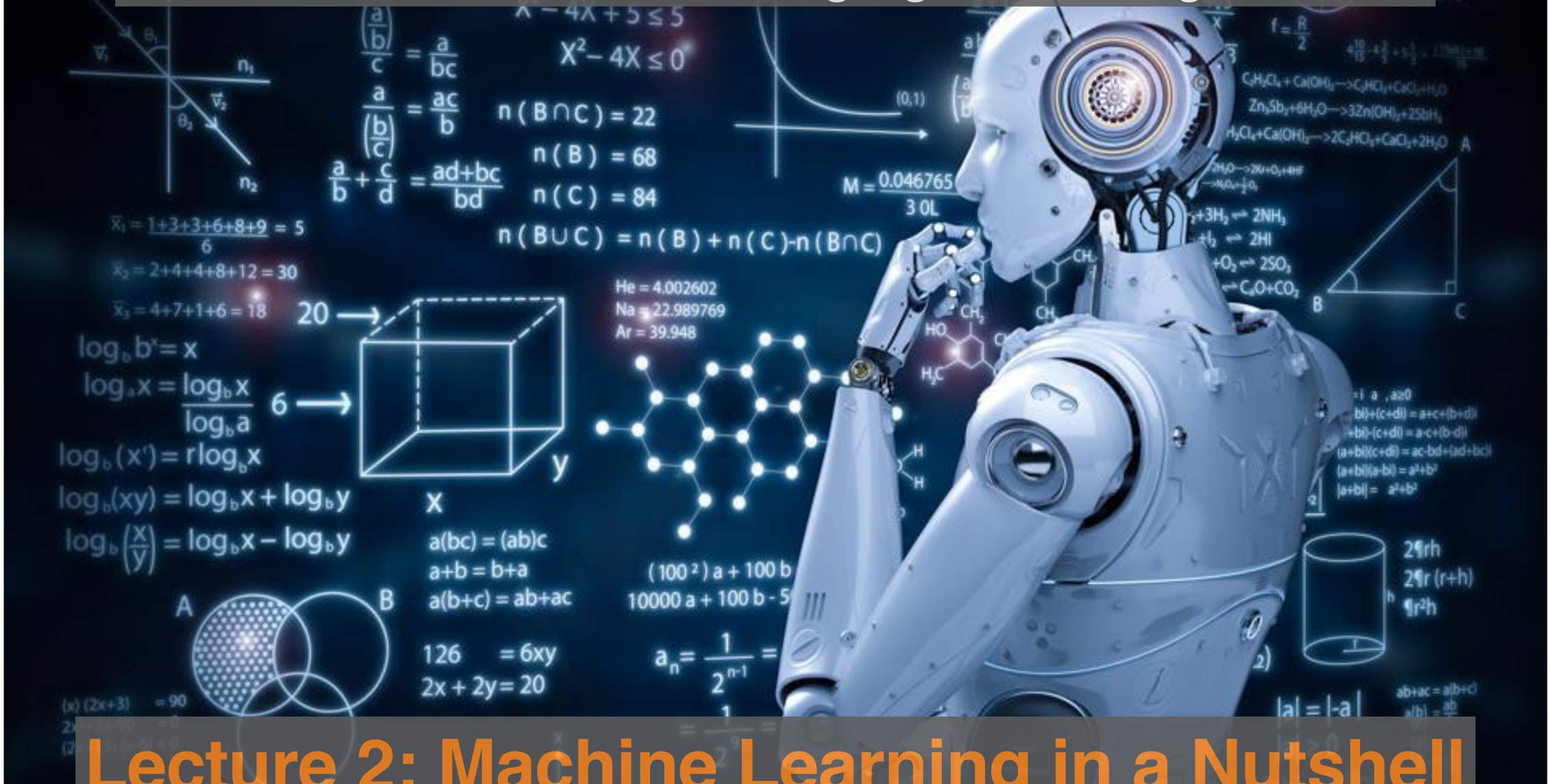
CSC3100 - Fundamentals of Speech and Language Processing
MDS6002 - Natural Language Processing

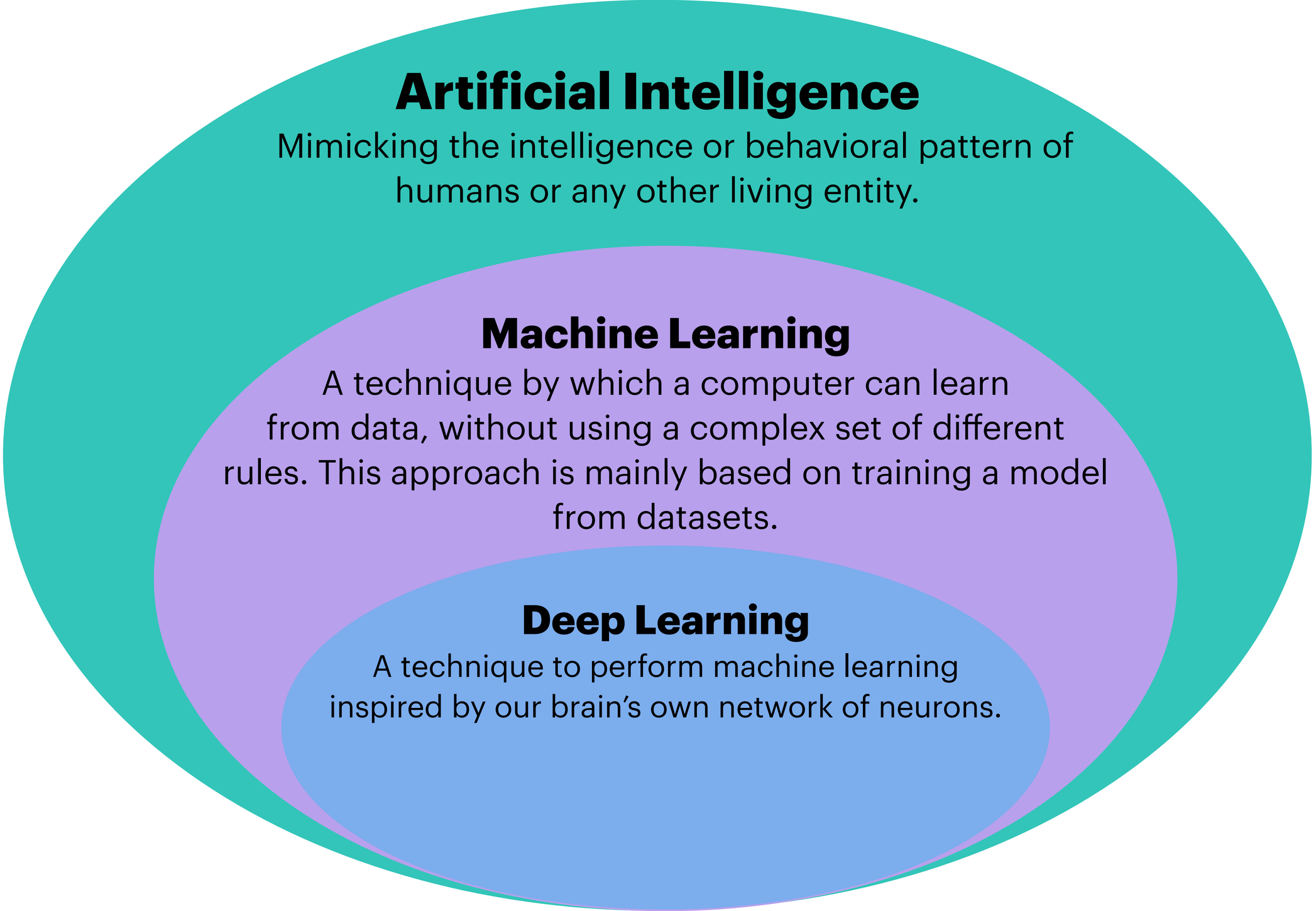


Lecture 2: Machine Learning in a Nutshell

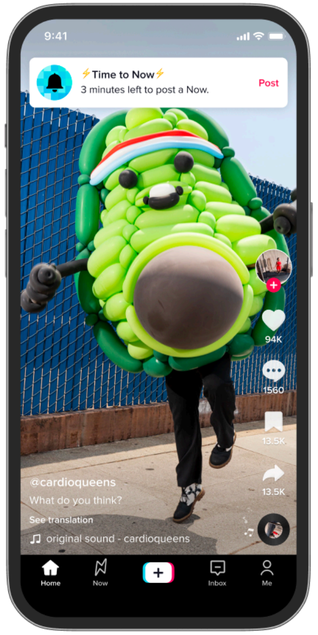
Zhizheng Wu

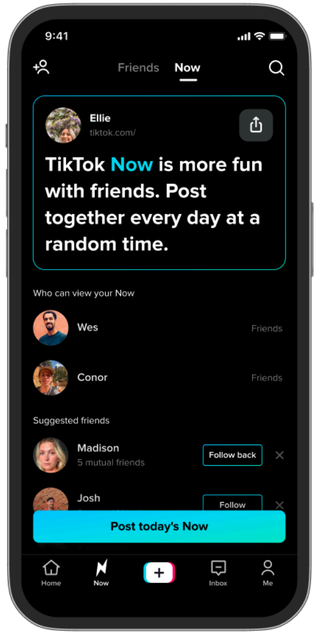
Outline

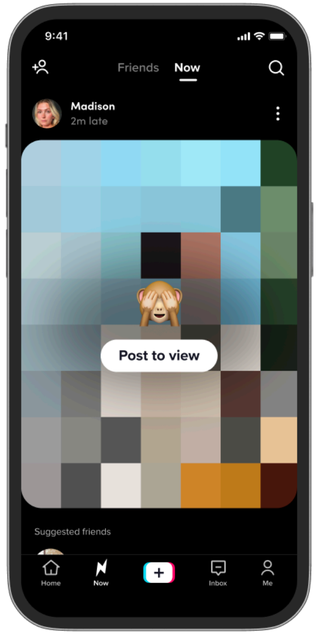
- Machine learning: An example
- Learning paradigms
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Deep learning models
- Loss function and evaluation metrics
- Data is the new oil
- ML in research vs in product

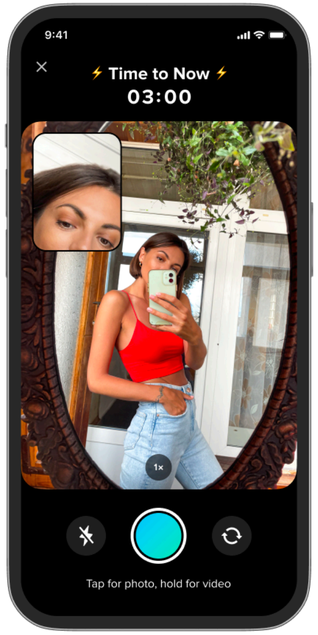

Artificial Intelligence
Mimicking the intelligence or behavioral pattern of humans or any other living entity.
Machine Learning
A technique by which a computer can learn from data, without using a complex set of different rules. This approach is mainly based on training a model from datasets.
Deep Learning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

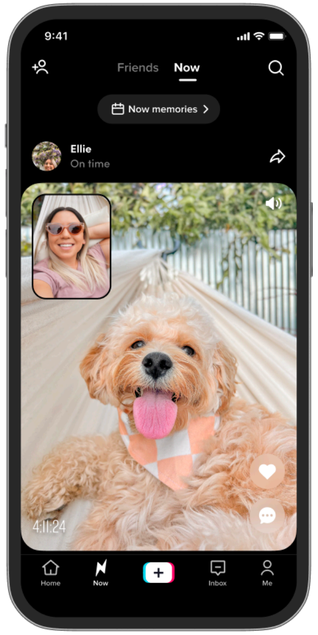
What does Lucas like?



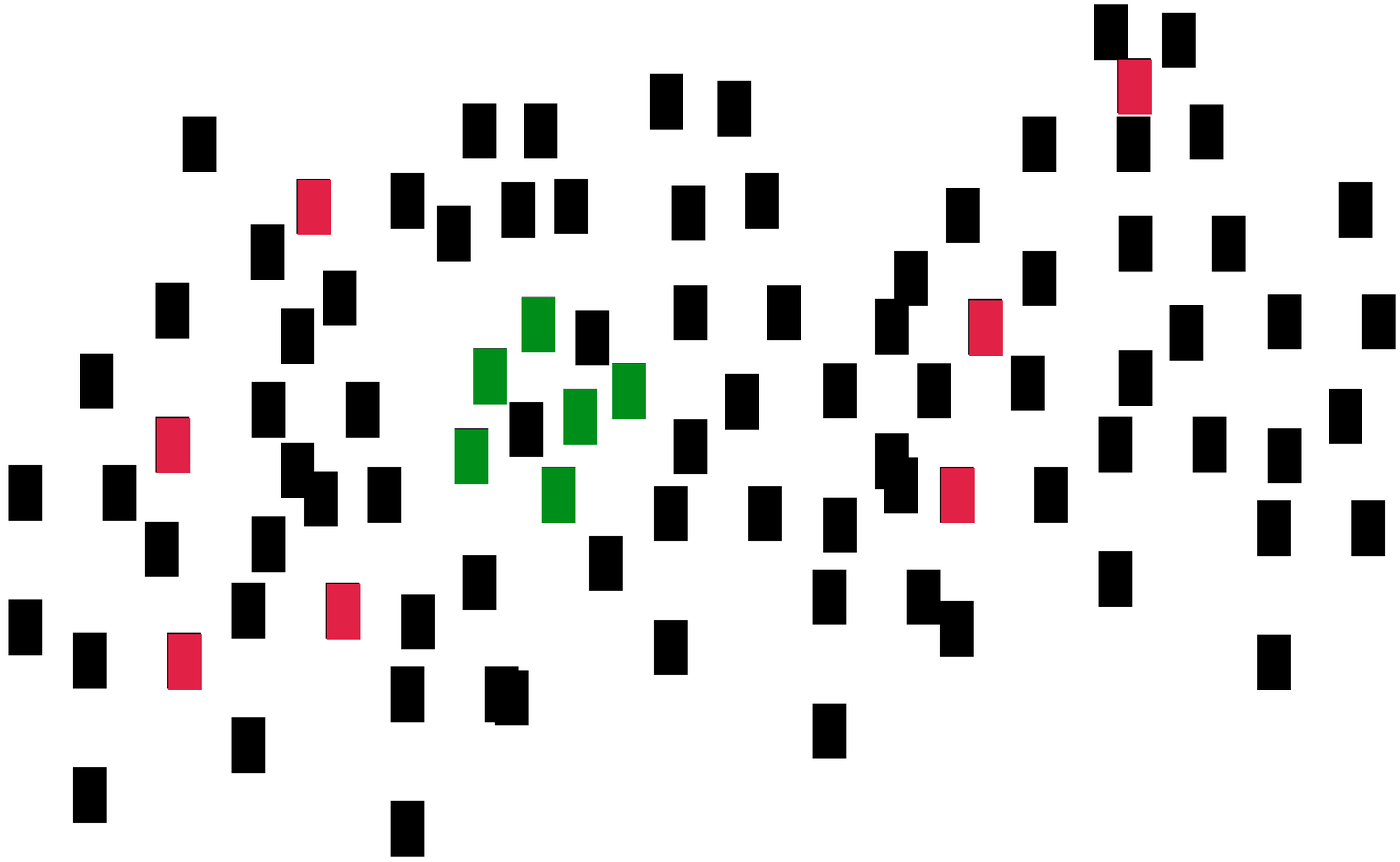






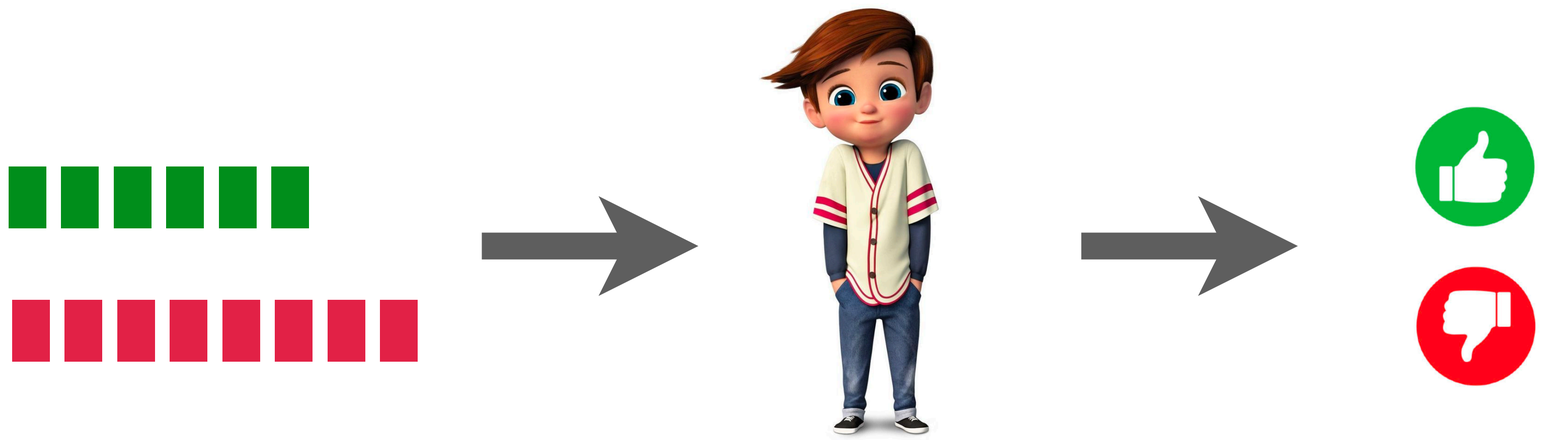


Lucas likes and dislikes



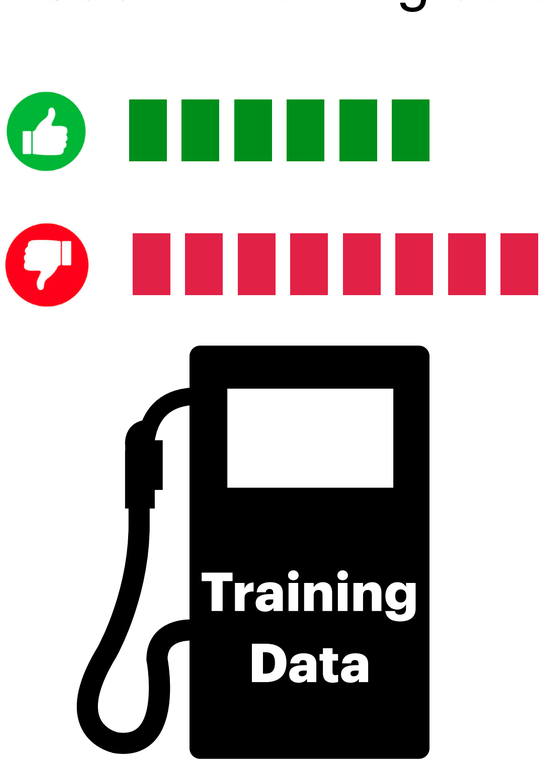
Machine learning to learn the behaviors

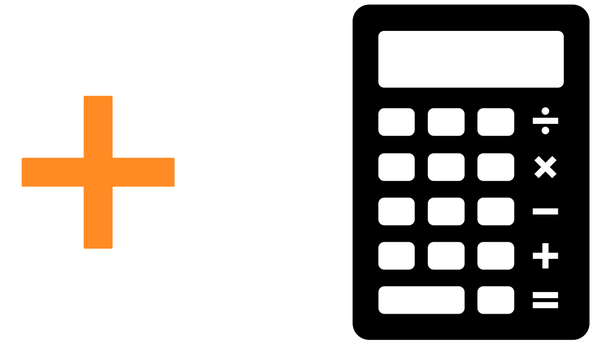
- Problem definition: Classify whether a video Lucas likes or dislikes



ML model = Data + Algorithms

- ML model = Training data + Algorithms

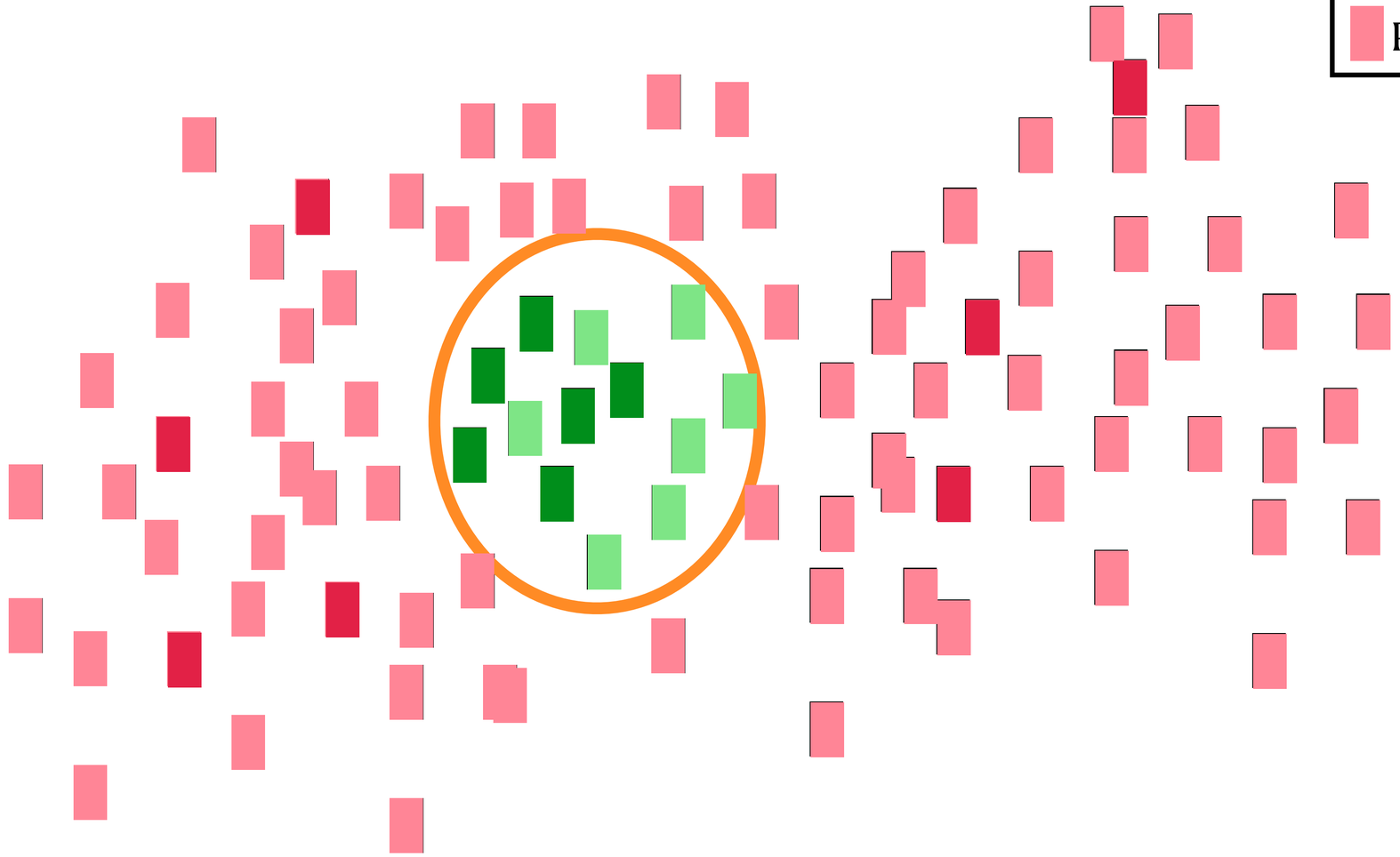




Algorithms

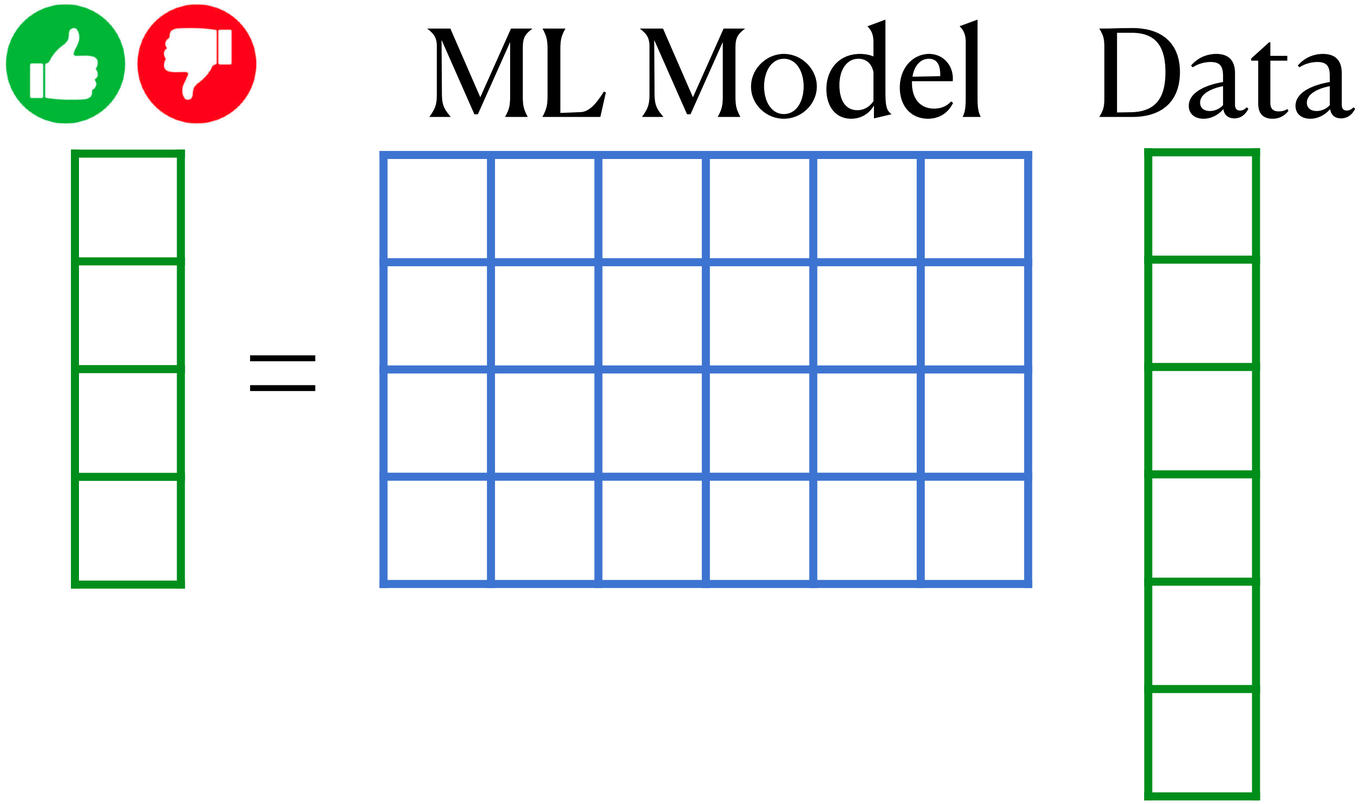
ML Model

ML prediction



Recommending videos that Lucas might like

ML Model ≅ a transformation function vs Linear algebra

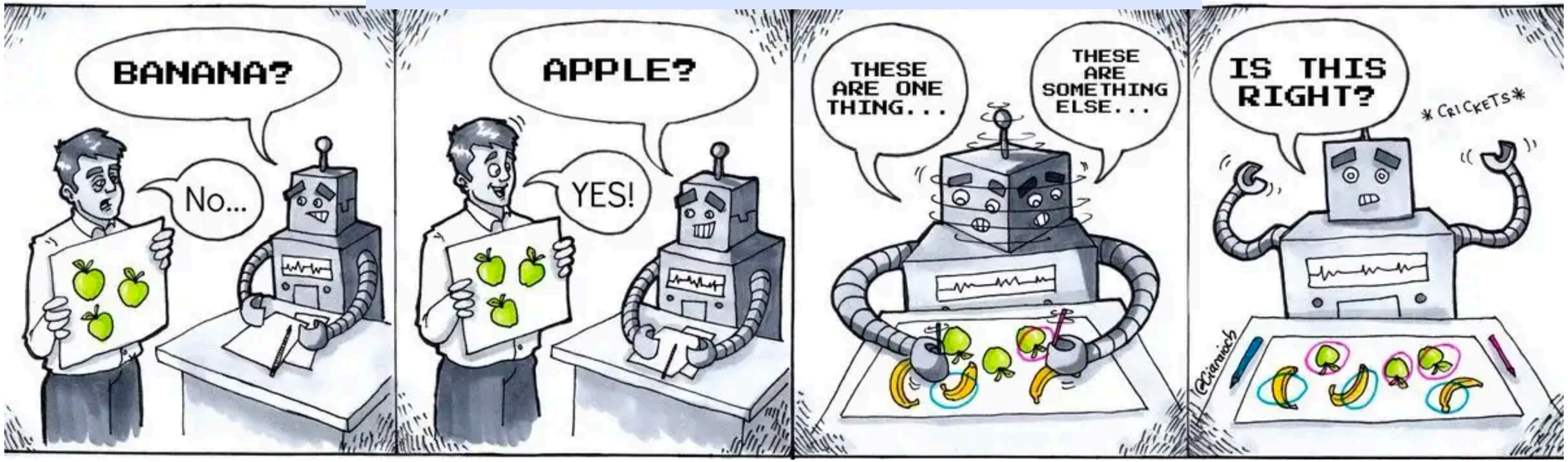


We need data and algorithms to learn the function

Learning paradigms

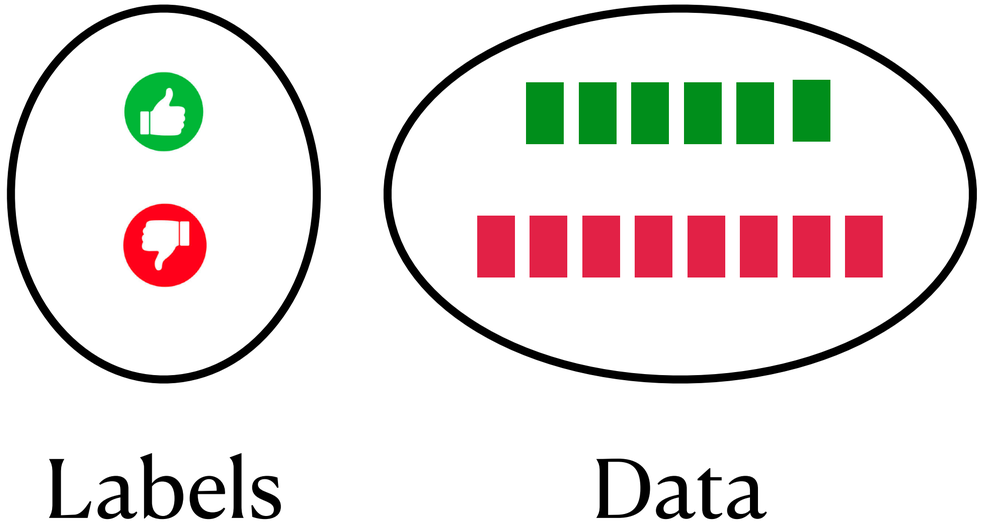
- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised vs Unsupervised learning



Supervised learning

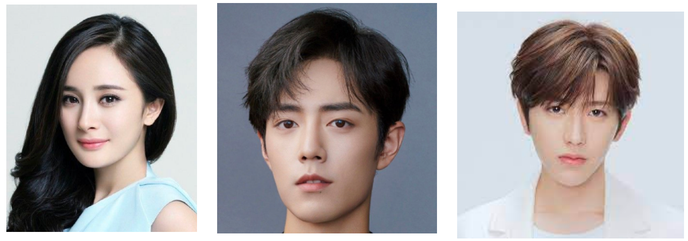
- Each data point consists of features and a label (or multiple labels)



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

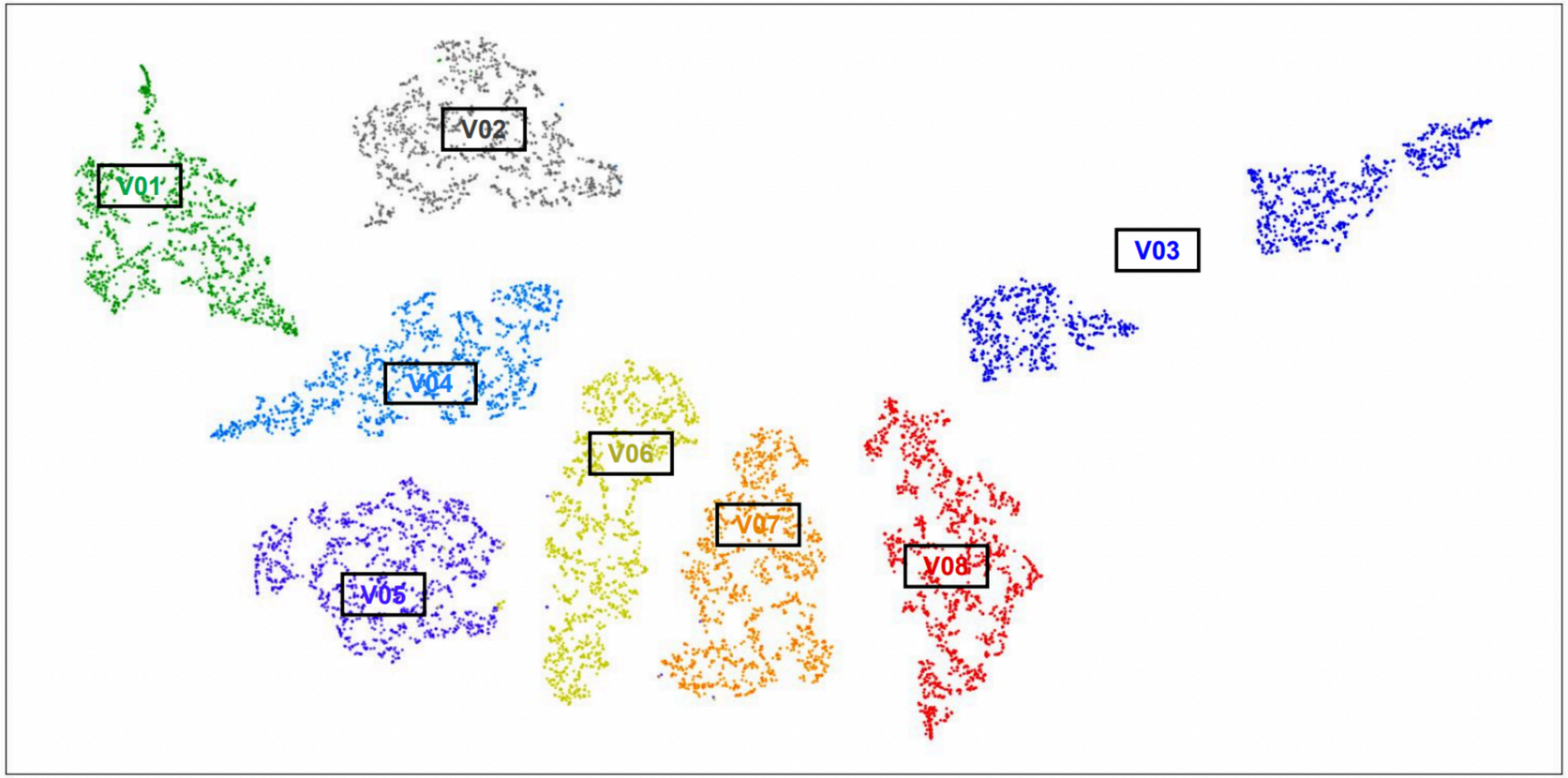
Generated from stable diffusion

Some typical supervised ML models

- Logistic Regression
- Support Vector Machines
- Random Forest

Unsupervised learning

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



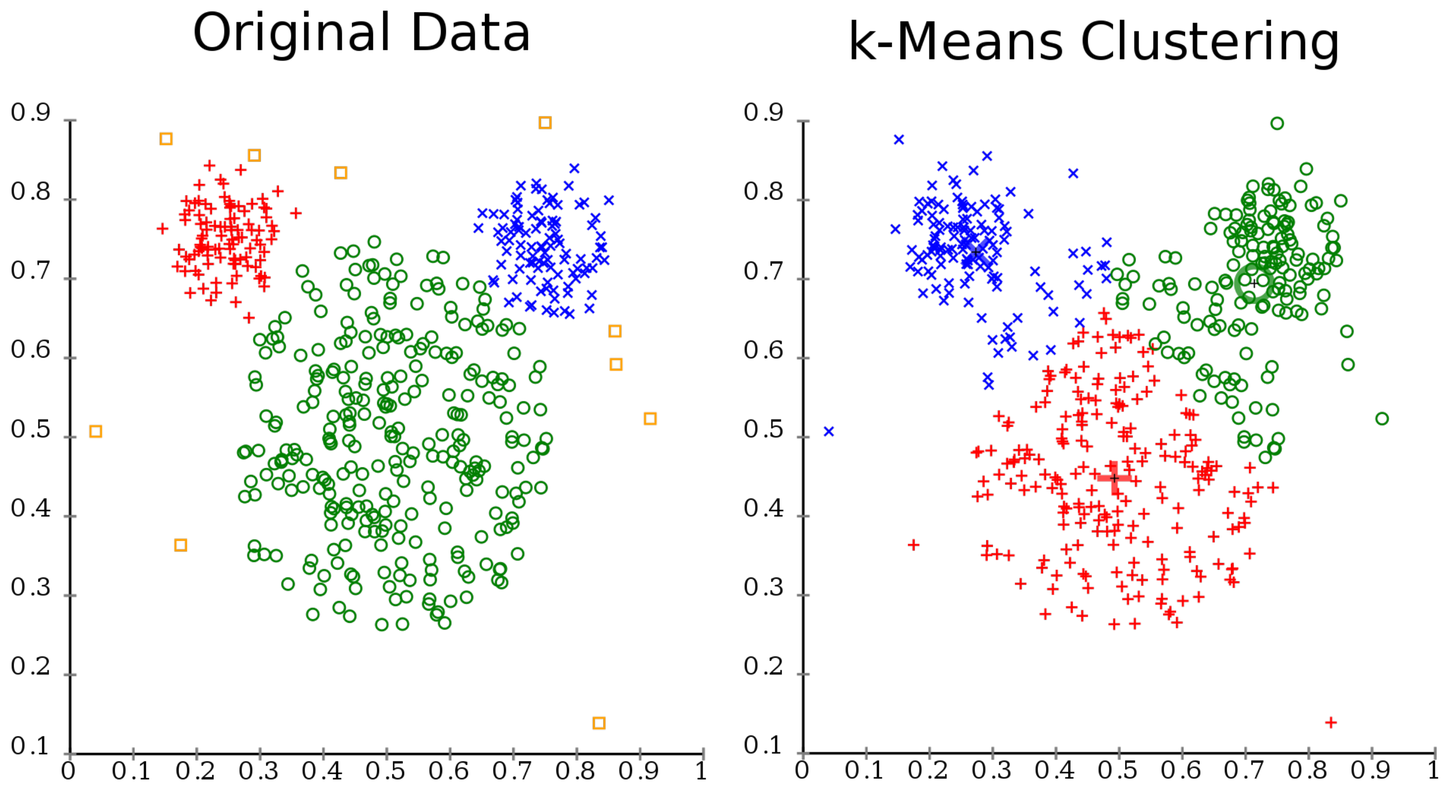
Typical supervised ML models

- 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

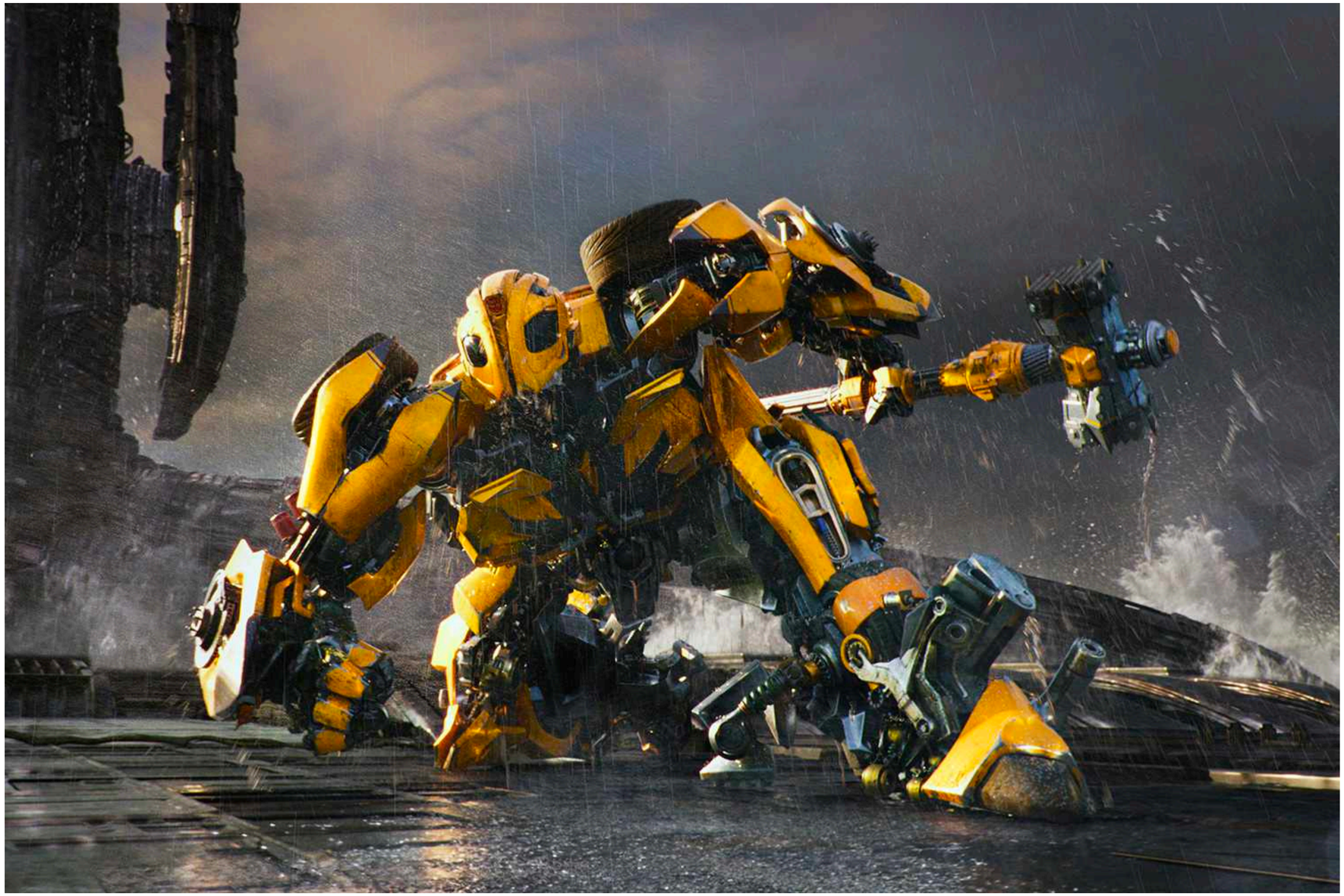


$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{\mathbf{x},\mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|^2$$

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

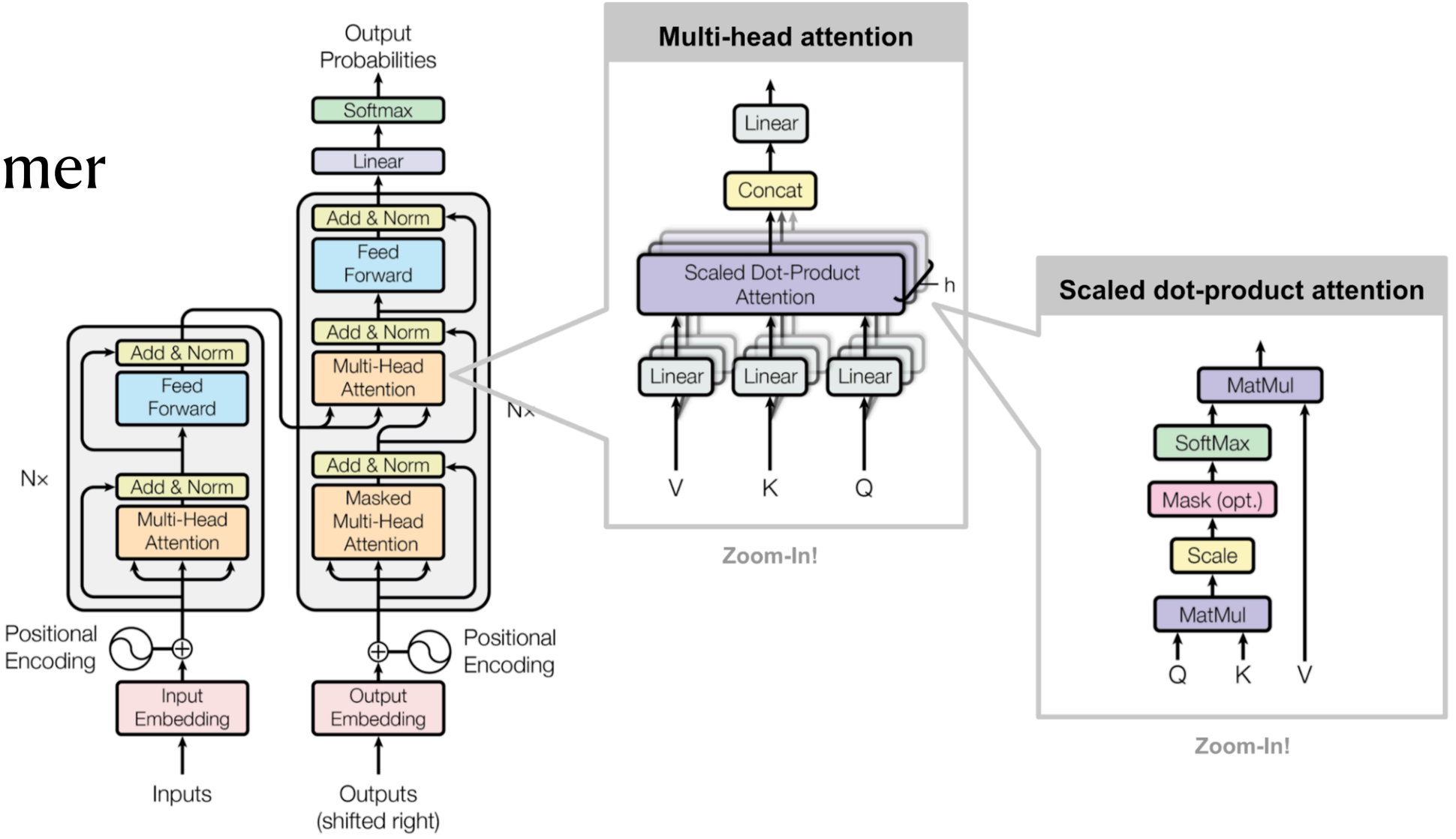
Reinforcement learning

Deep learning models



Deep learning models

Transformer



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

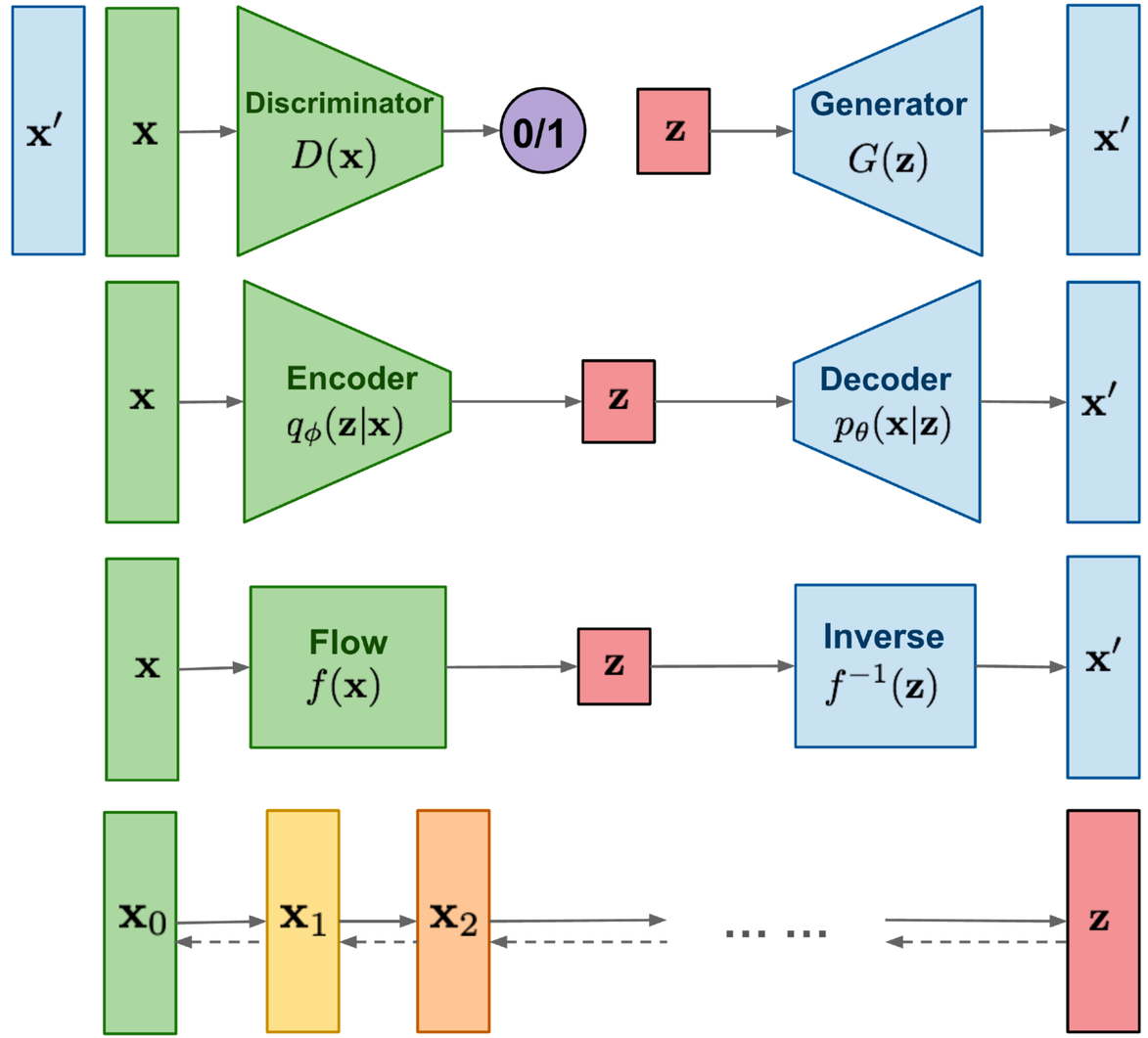
Deep learning generative models

GAN: Adversarial training

VAE: maximize variational lower bound

Flow-based models: Invertible transform of distributions

Diffusion models: Gradually add Gaussian noise and then reverse

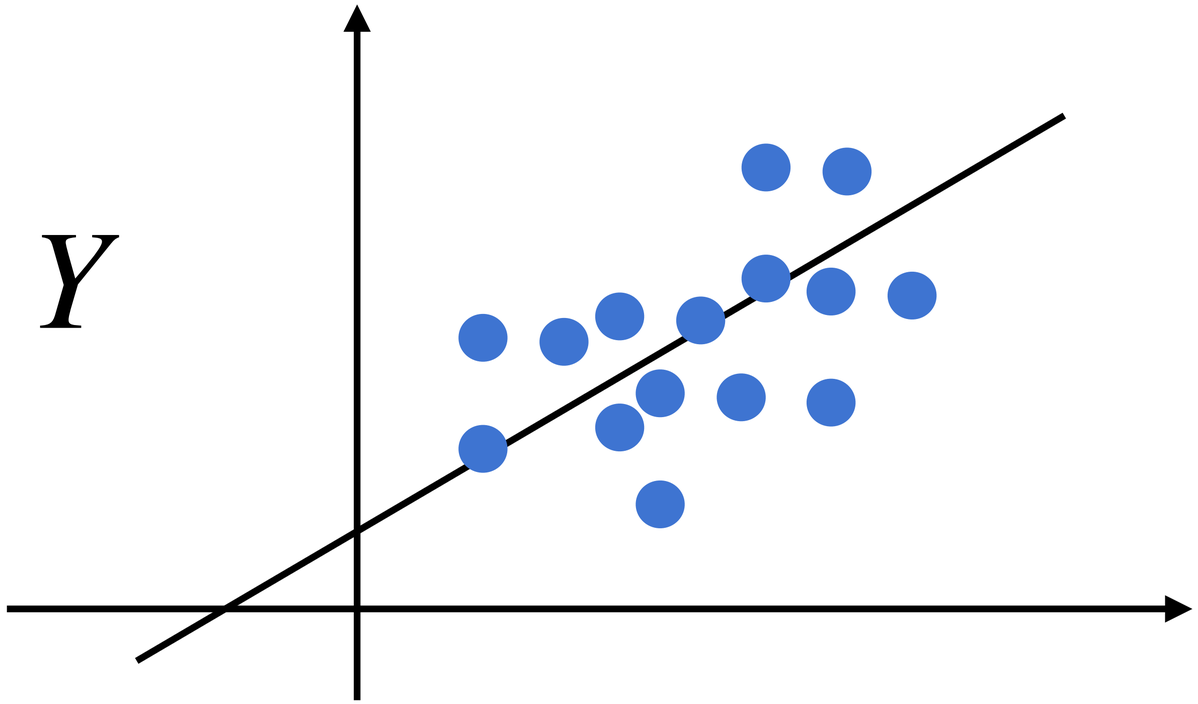


https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Loss function

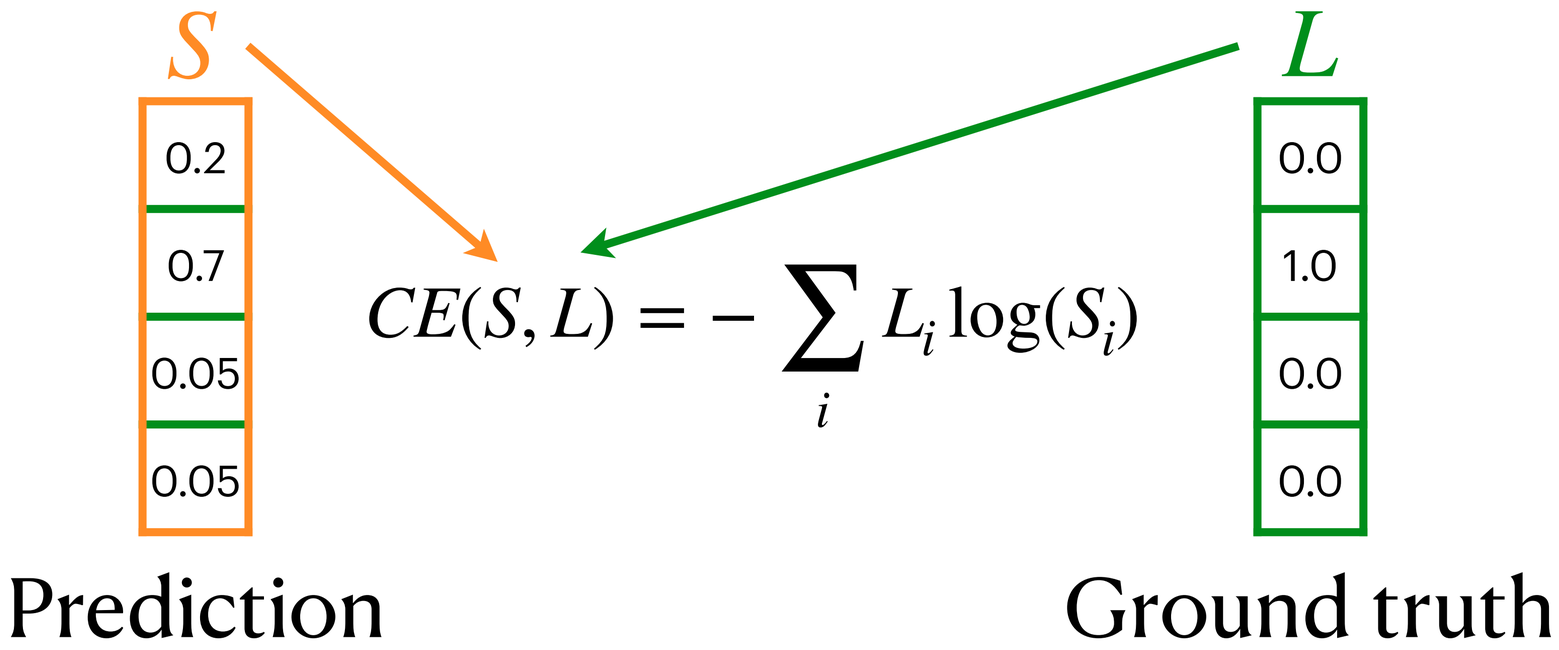
- A method of evaluating how well your algorithm fits/models your dataset

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



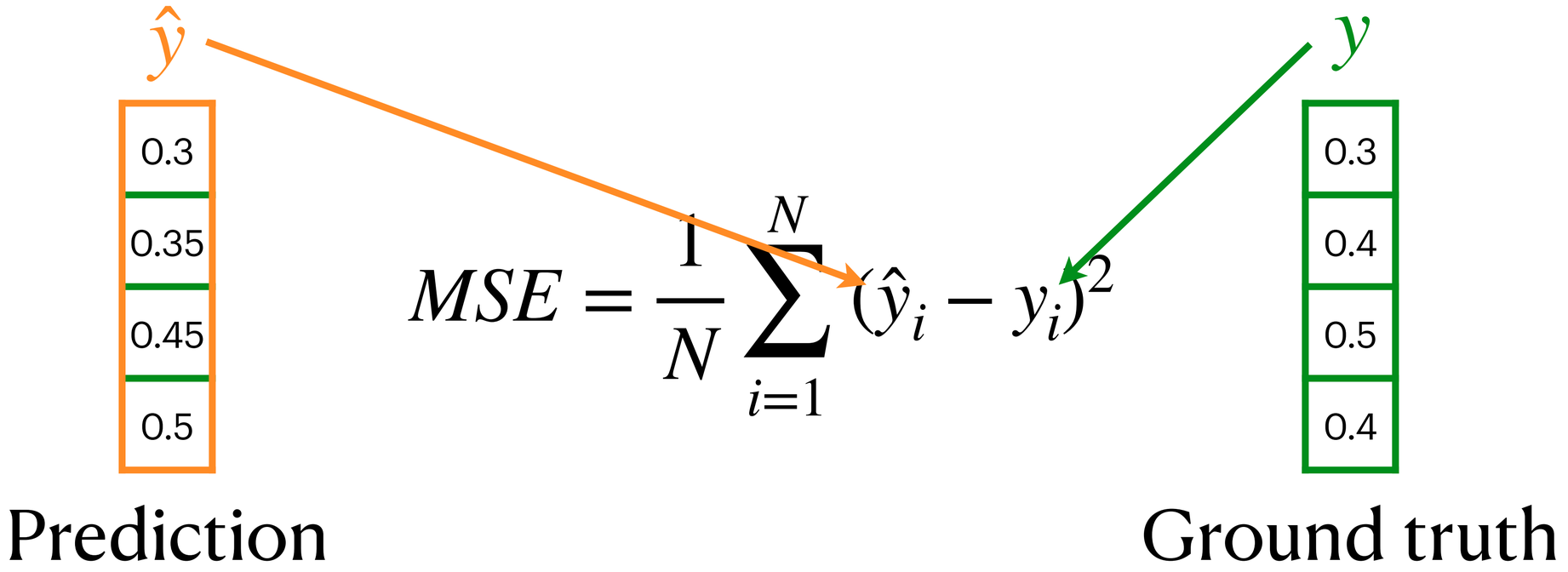
Loss function

- Cross-entropy loss
  - Usually used in classification tasks



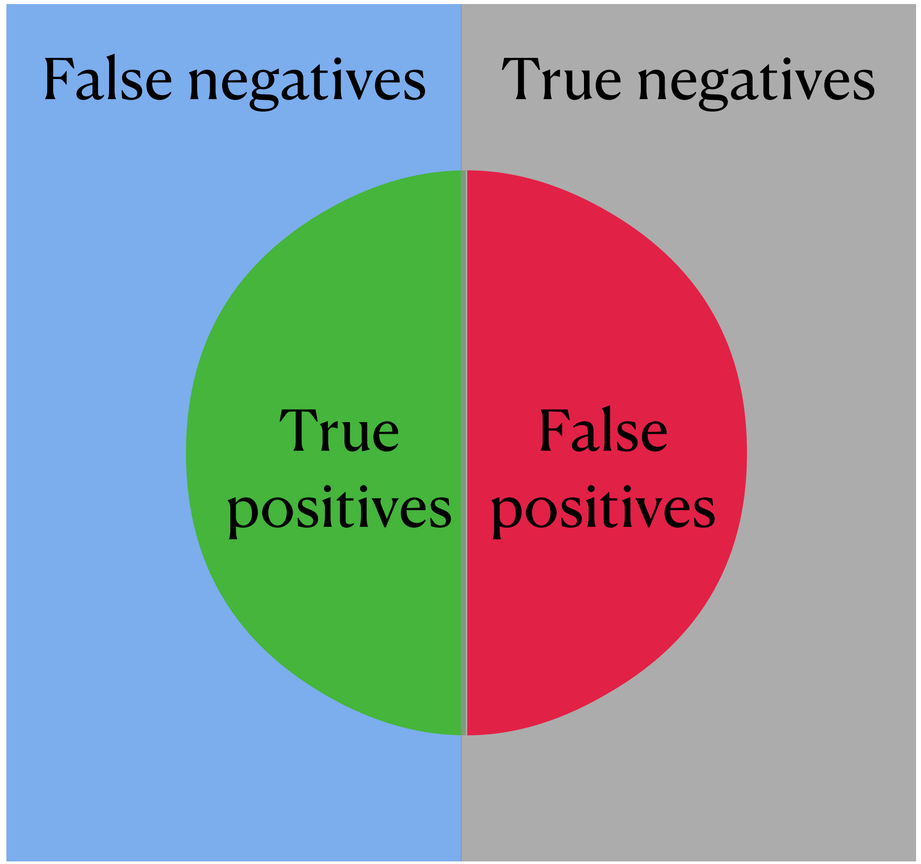
Loss function: Mean-squared loss

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



Evaluation metrics

- Precision and recall



$$\text{Prediction} = \frac{\text{●}}{\text{●●}}$$

$$\text{Recall} = \frac{\text{●}}{\text{▮●}}$$

Evaluation metrics

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

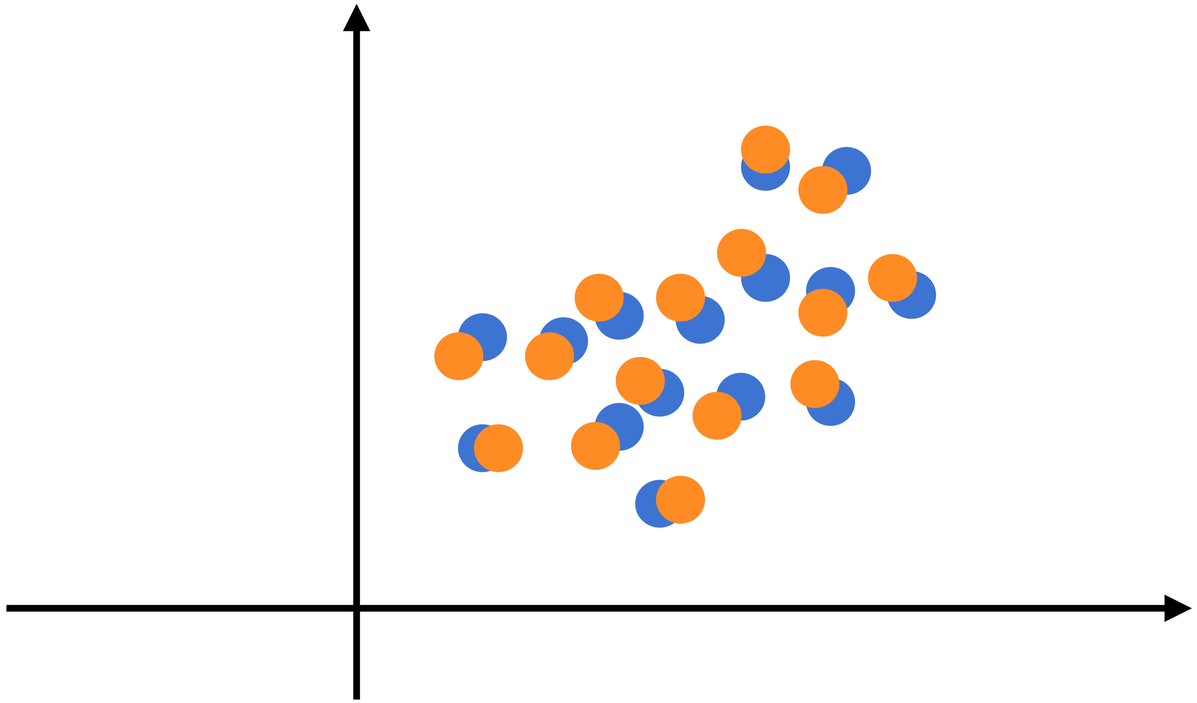
$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

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

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

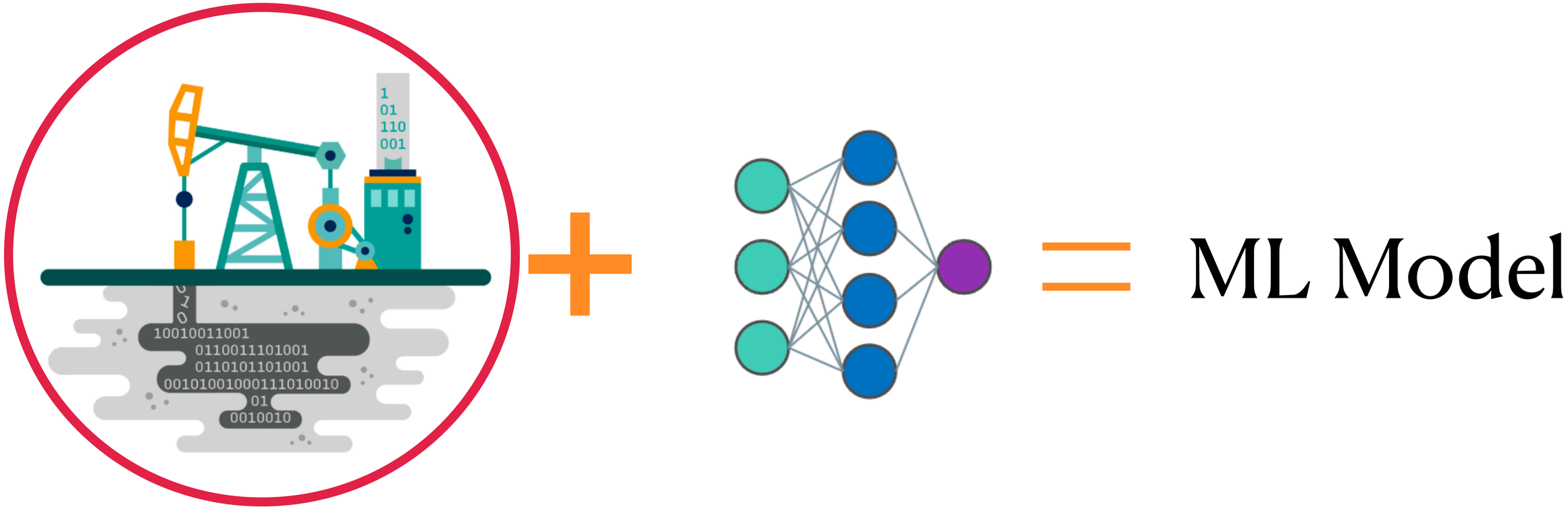
Evaluation metrics

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

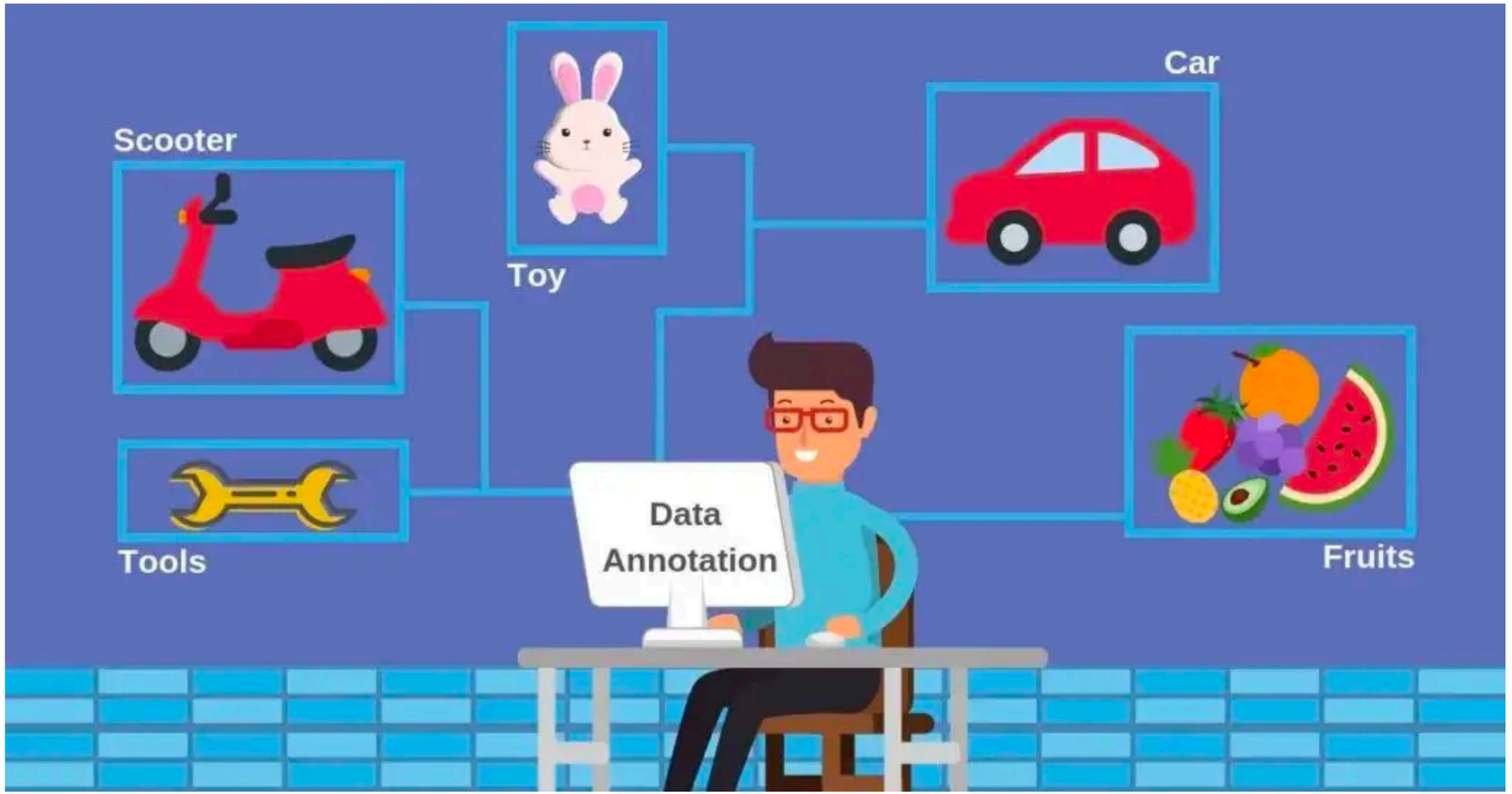


$$RMSE = \sqrt{\frac{\sum_i^N (y_i - \hat{y}_i)^2}{N}}$$

Data is the new oil



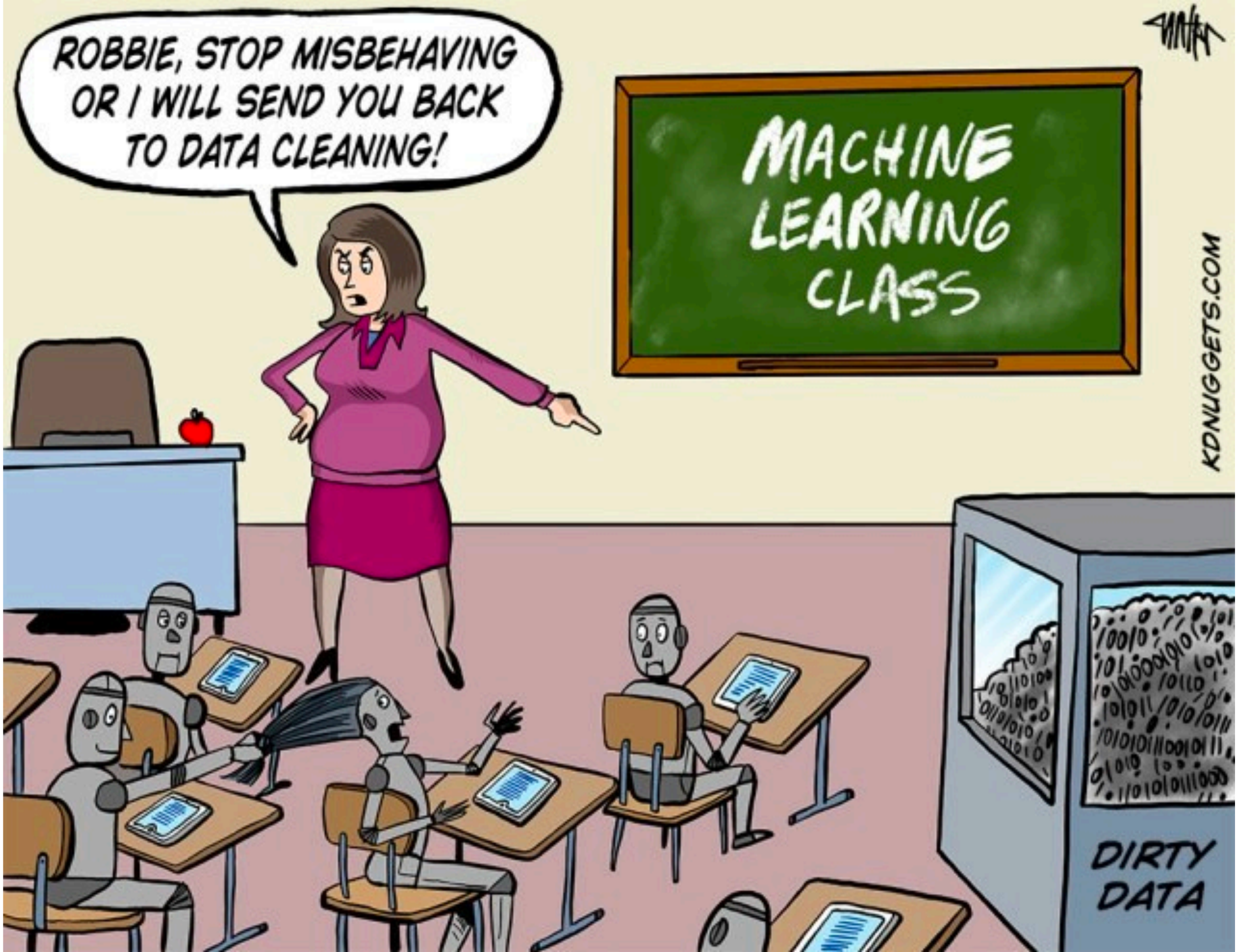
Labeling data



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

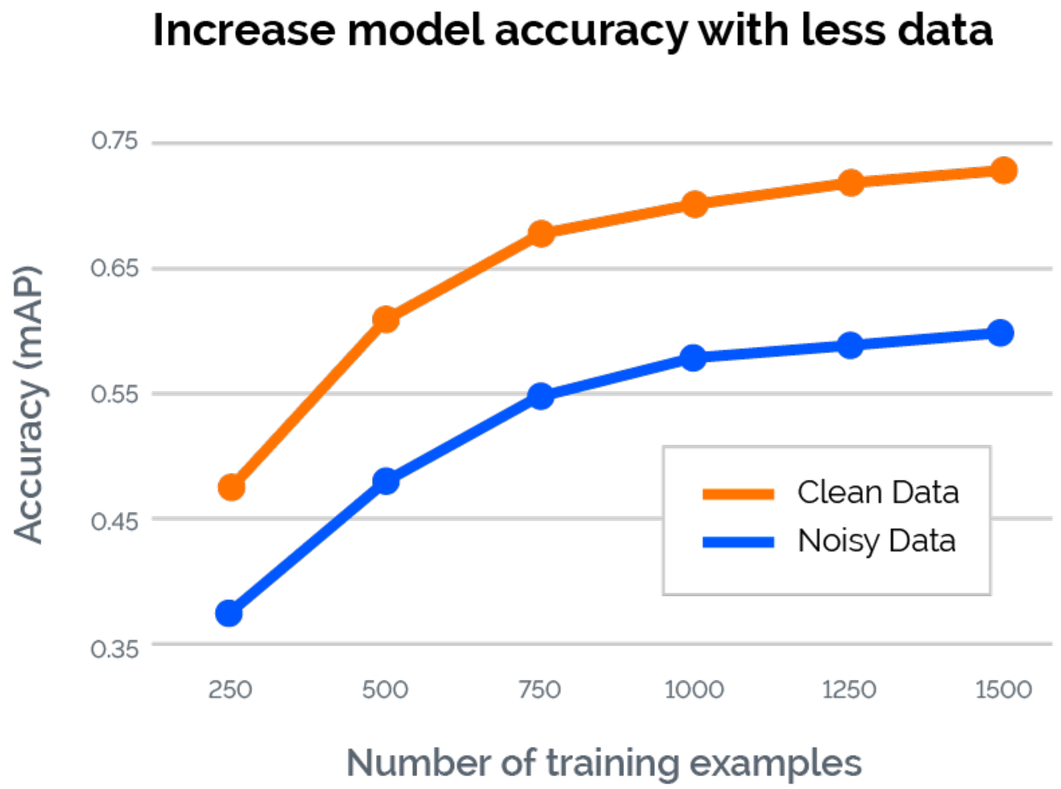
Data labeling

Expensive: Especially when subject matter expertise is required
Non-private: Need to ship data to human annotators
Scalability: Time required is linearly correlated with # labels needed
Non-adaptive: Every guideline change requires re-labeling the dataset


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

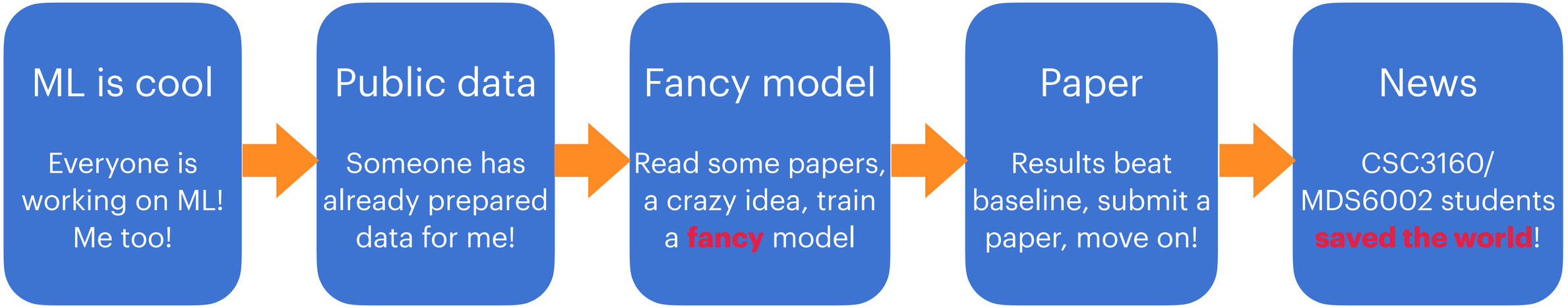
Focusing on high-quality data that is consistently labeled would unlock the value of AI for sectors such as health care, government technology, and manufacturing

- Andrew Ng

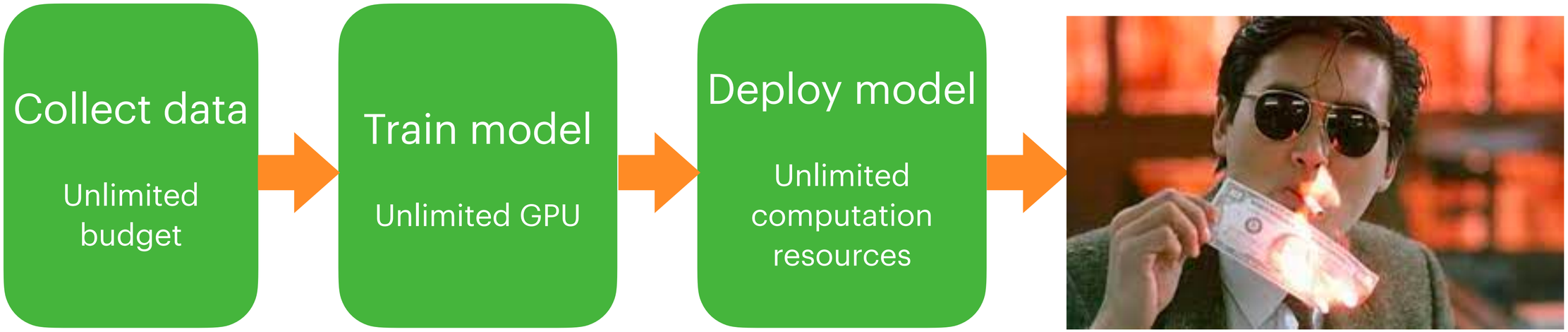


https://mitsloan.mit.edu/ideas-made-to-matter/why-its-time-data-centric-artificial-intelligence

Machine learning in research

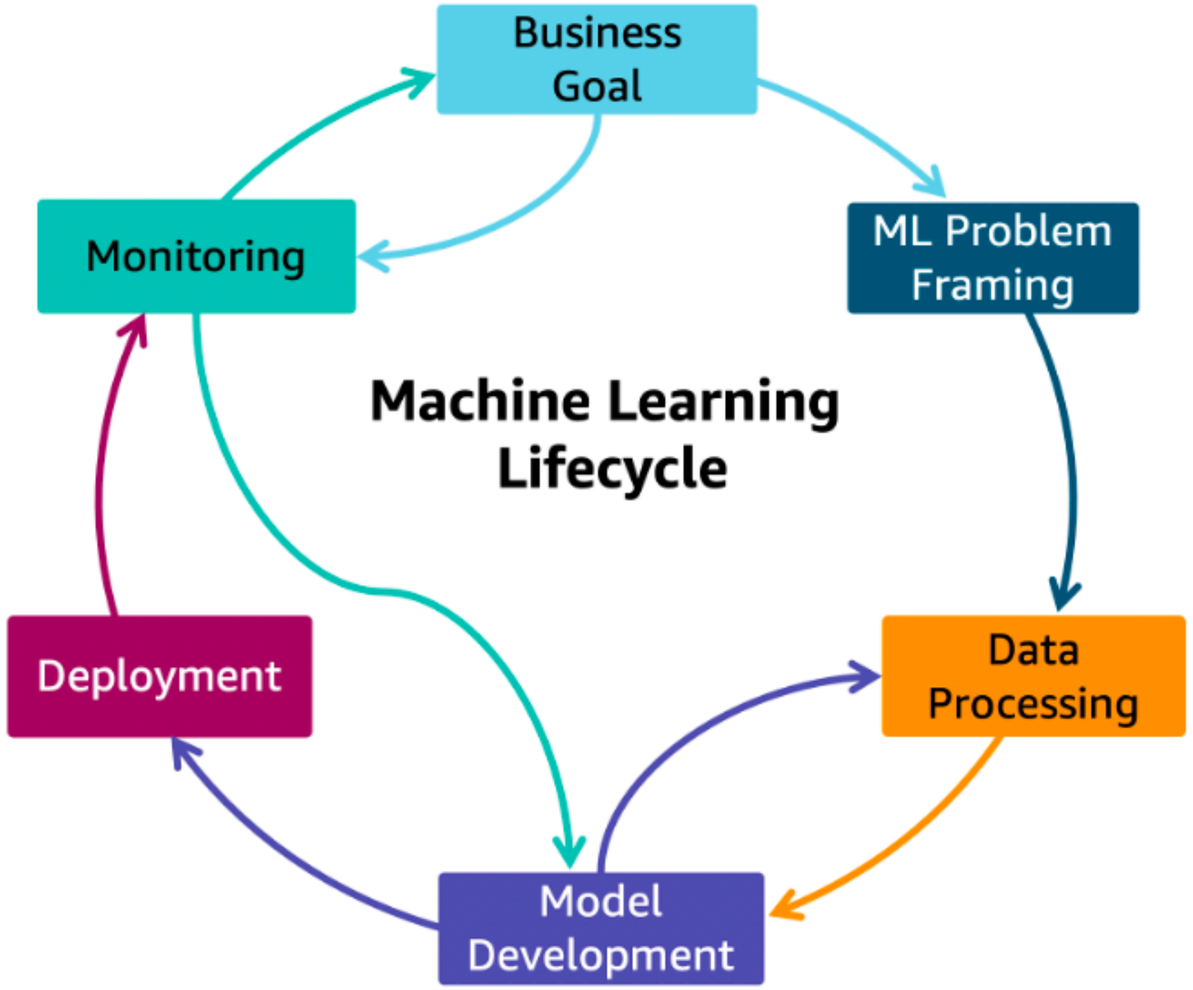


ML in product: Expectation



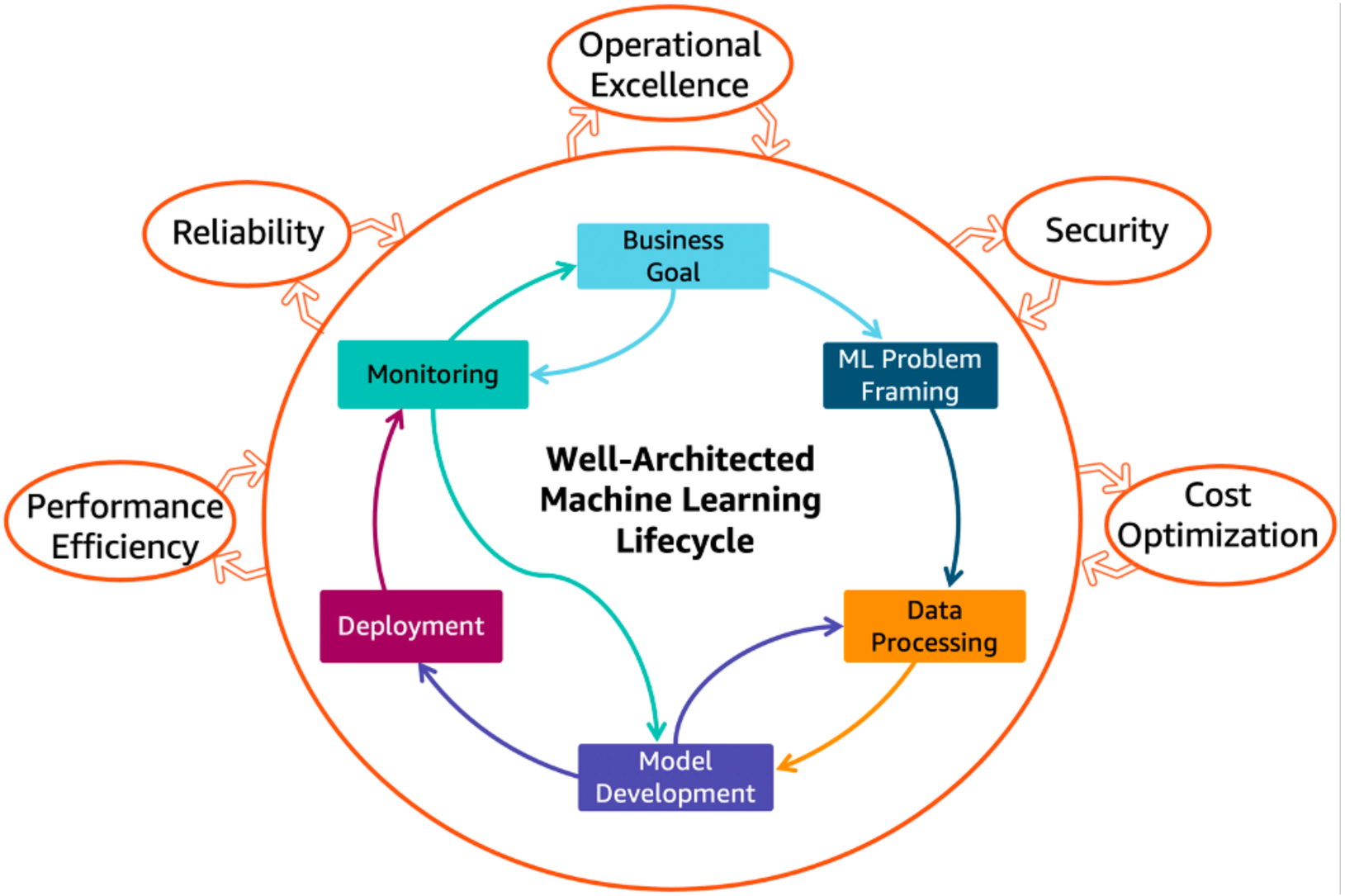
Machine learning in production: Reality

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

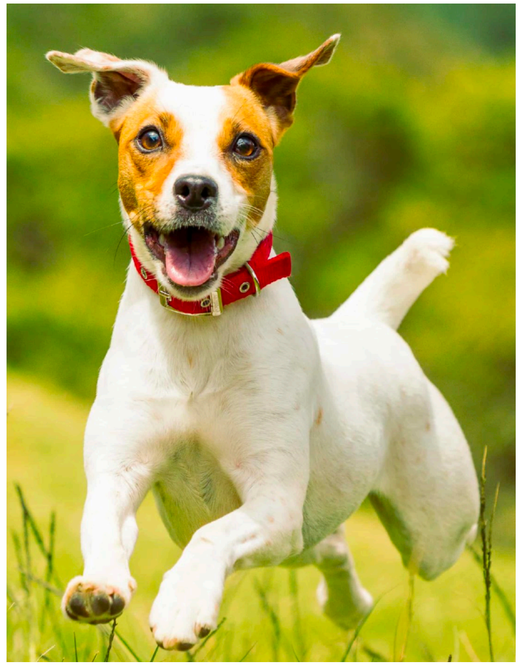
Fancy model
Highest accuracy

Sales

More clients
More revenue

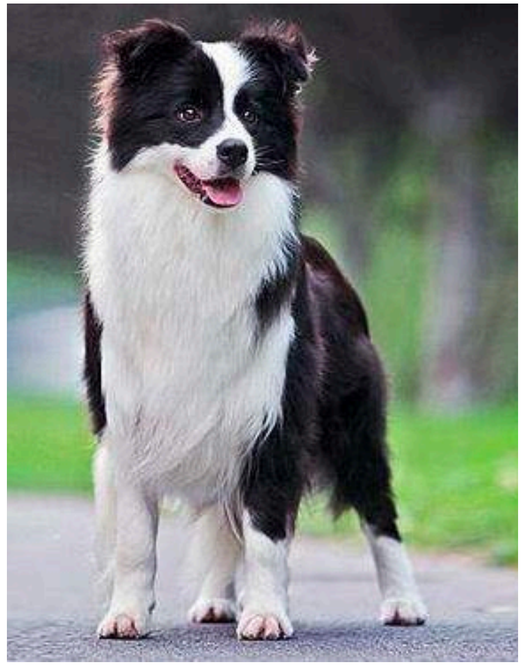
Product

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/

what society thinks I do

what my friends think I do

what my parents think I do

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i$$

$$\alpha_i \geq 0, \forall i$$

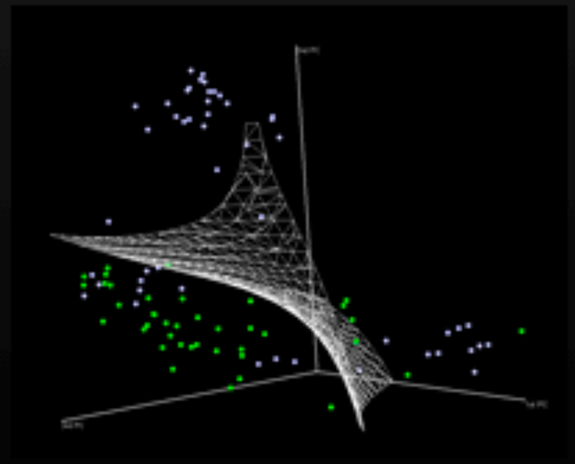
$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\nabla \hat{g}(\theta_t) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t).$$

$$\theta_{t+1} = \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t)$$

$$\mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] = \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t).$$

what other programmers think I do



what I think I do

```
>>> from sklearn import svm
```

what I really do

Credit:
Harrison Kinsley