Introduction to machine learning

- **Outline:**

  1. What is machine learning?

  2. Types of machine learning

  3. Data for machine learning

  4. Machine learning for classification

Introduction to machine learning

- **Outline:**

  1. What is machine learning?

  2. Types of machine learning

  3. Data for machine learning

  4. Machine learning for classification

What is machine learning?

- The process of solving a practical problem by:
  - Gathering a dataset
  - Based on that dataset building a statistical model which is assumed to be used somehow to solve the practical problem.

History

- In 1959, Arthur Samuel (American pioneer in computer gaming and AI) coined the term “machine learning” while at IBM
- In 1960, IBM used new cool term “machine learning” to attract clients and talented employees

Arthur Lee Samuel

Born	December 5, 1901 Emporia, Kansas
Died	July 29, 1990 (aged 88) Stanford, California

Arthur Lee Samuel *(1959)*

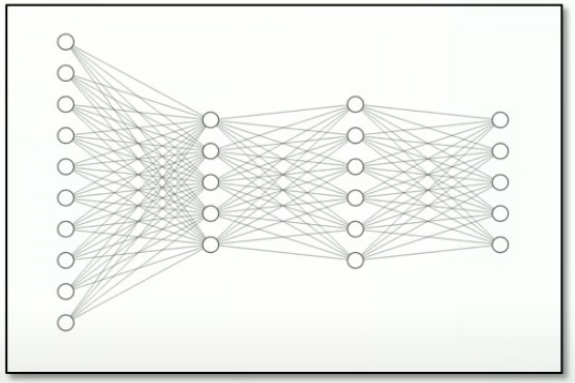
Machine Learning *the "field of study that gives computers the ability to learn without being explicitly programmed".*

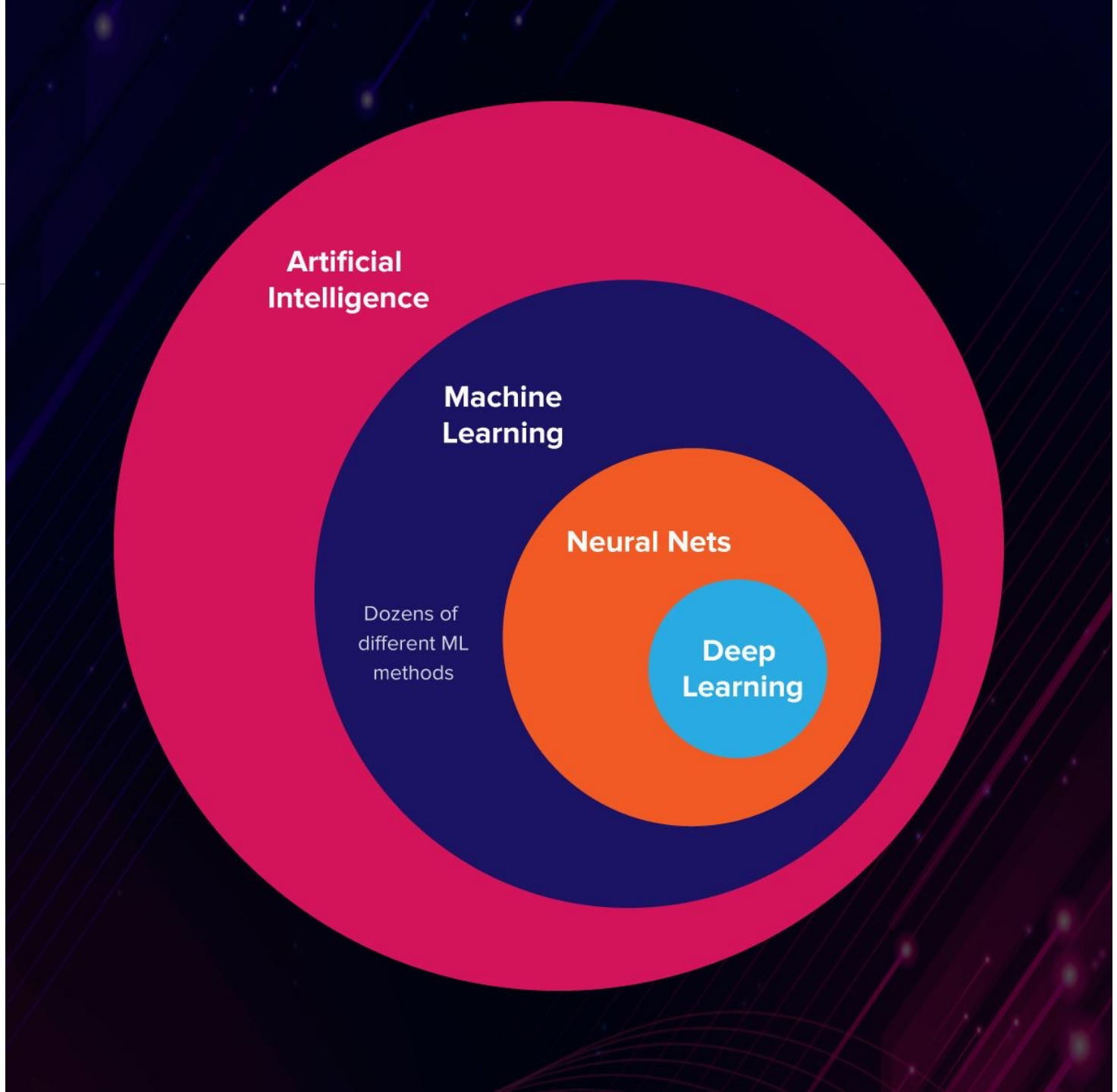
Three forces brought AI to life

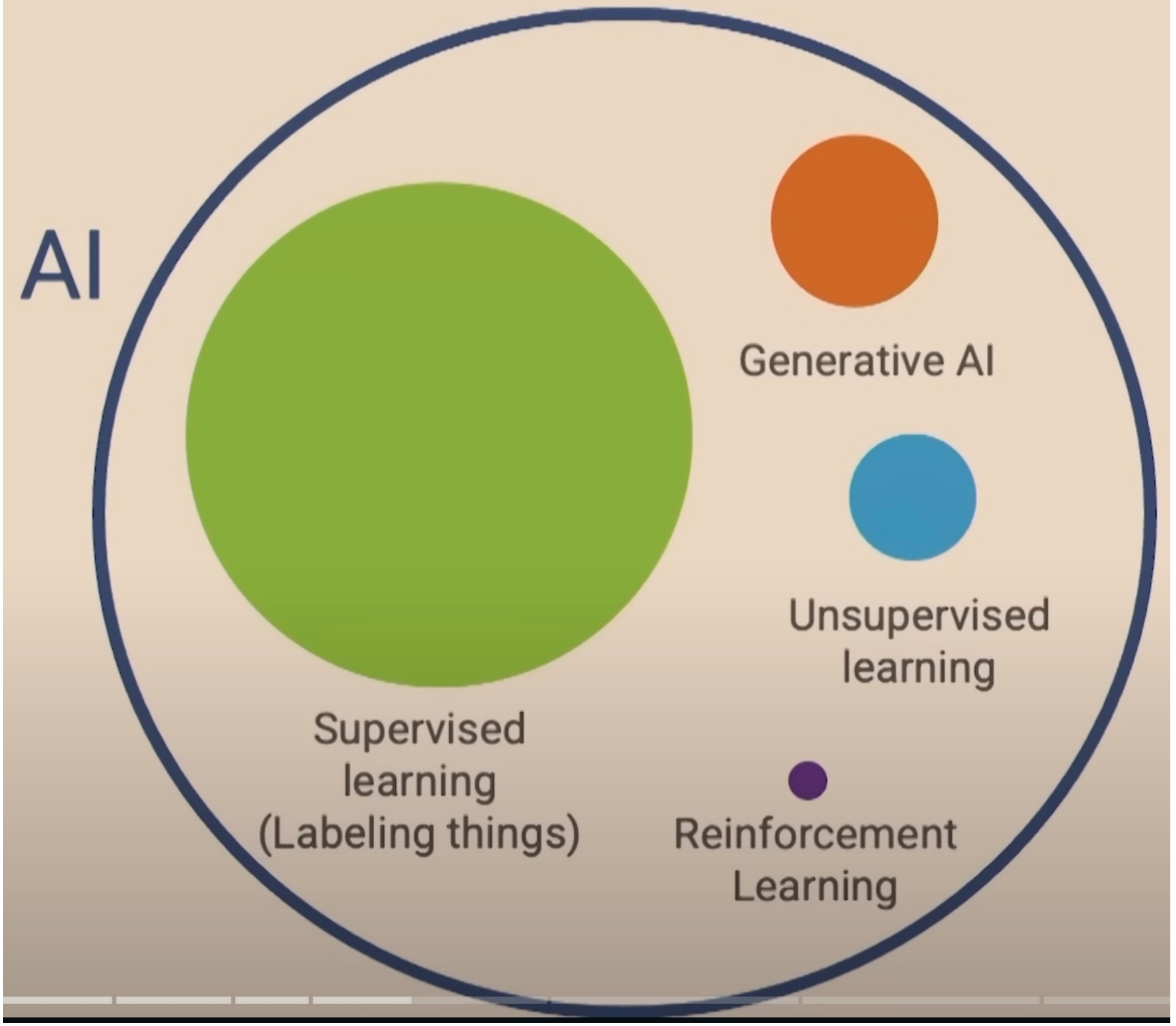
Big Data

Compute Power

Machine Learning Algorithms

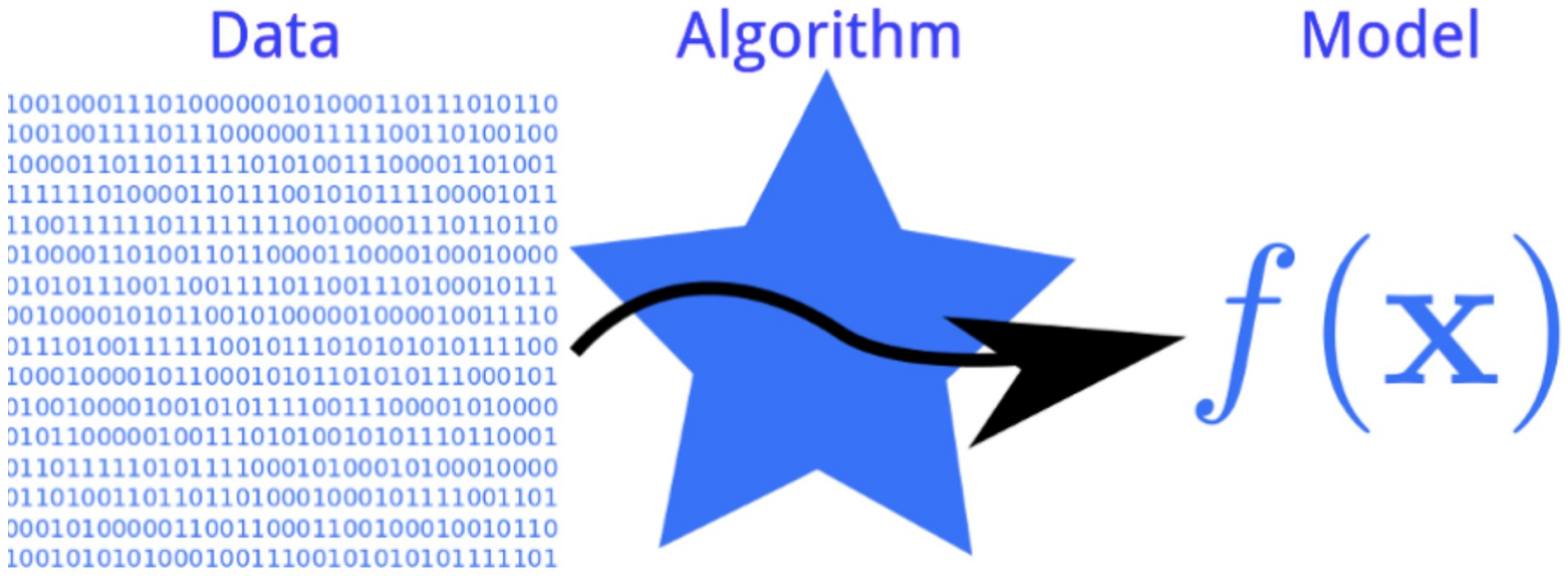



Artificial Intelligence
Machine Learning
Neural Nets
Deep Learning
Dozens of different ML methods


AI
Generative AI
Unsupervised learning
Supervised learning (Labeling things)
Reinforcement Learning

Machine learning algorithm

- Finding a mathematical formula based on a collection of inputs (i.e, “training data”)
- Applying formula to training inputs → produces the desired outputs.
- Applying formula to novel inputs → generates the correct outputs.
- New inputs come from the same or a similar statistical distribution.



Traditional programming vs. ML

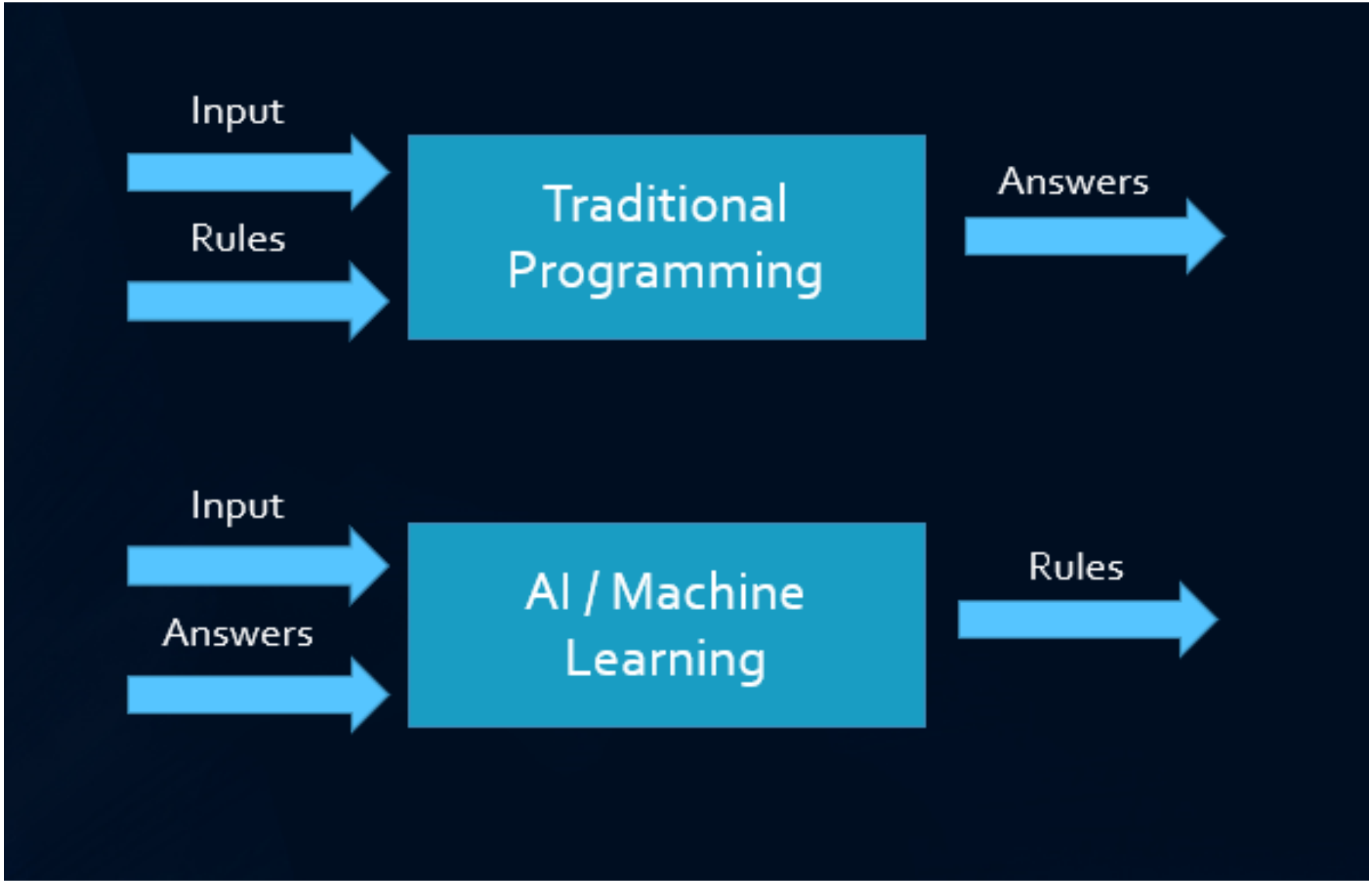
Traditional programming

- Feeding computer with rules
- Computer utilizes computing
- Coming up with answers

Machine learning

- Feeding computer with huge amount of data
- Computer processes the data
- Coming up with trained model that can solve the unseen problems of the real world

Traditional programming vs. ML



Example of Fizzbuzz game solving

1. If the number is a **multiple of 3**, **print "Fizz"** instead of the number.
2. If the number is a **multiple of 5**, **print "Buzz"** instead of the number.
3. If the number is a **multiple of 3** ***and*** **5**, **print "FizzBuzz"** instead of the number.

Example of Fizzbuzz game solving

Bài tập nhóm

- **Đề:** Cho $a \neq 0, b, c$ thực

$$ax^2 + bx + c = 0$$

Tìm x thoả mãn phương trình trên

- Hãy lập trình giải phương trình trên theo phương pháp truyền thống
- Tìm hiểu về phương pháp giải phương trình trên dùng ML

Gợi ý tìm kiếm: quadratic equation solving

Example of quadratic equation solving

- Given $a \neq 0, b, c$
- $ax^2 + bx + c = 0$
- Find x?

- **Solving:**

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Example of quadratic equation solving

- Given $y = ax^2 + bx + c$

when x=	0	y=	1.1
when x=	1	y=	5.9
when x=	2	y=	16.8
when x=	3	y=	33.9

- **Solve for** $x = 8$?
- Using ML → $y = 3.078x^2 + 1.701x + 1.106$

Expected 1.1,5.9,16.8,33.9
Got 1.106,5.884,16.817,33.906

- $x = 8 \longrightarrow y = ?$

When ML is used?

A real-world problem is a candidate for the application of machine learning if -

1. Historical data exists in a huge amount
2. A pattern exists in the data
3. Extremely hard to pin down a solution mathematically

Introduction to machine learning

- **Outline:**

  1. What is machine learning?
  2. **Types of machine learning**
  3. Data for machine learning
  4. Teachable machine

Types of ML

- **Supervised learning:** training data includes desired outputs (labels)
  - Classification
  - Regression
- **Unsupervised learning:** training data does not include desired outputs (labels)
  - Clustering
- **Reinforcement:** rewards from a sequence of actions

Supervised learning

- **Dataset:** set of labeled examples (labeled data)

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

Feature vector — Label

[height weight gender age] {normal, thin, fat}

- **Goal:** to produce a model that takes a **feature vector** as input and outputs the label for this feature vector.
- **Ex:** detection, classification

Classification

- An application of supervised learning
- Automatically assigning a **label** to an **unlabeled example**
- An classification learning algorithm takes a collection of **labeled examples** as inputs and produces a **model** that can take an unlabeled example as input and output a label (or a number label)
- **Ex:** Covid-19 detection, traffic light classification

Regression

- An application of supervised learning
- Automatically predicting a **real-valued label** (i.e., *target*) given an **unlabeled example**.
- A regression learning algorithm takes a collection of **labeled examples** as inputs and produces a **model** that can take an unlabeled example as input and output a target.
- **Ex**: estimating house price based on house features [area, # bedrooms, location, etc]

Unsupervised learning

- **Dataset:** set of unlabeled examples

$$\{\mathbf{x}_i\}_{i=1}^{N}$$

Feature vector

[height weight gender age]

- **Goal:** to produce a model that takes a feature vector as input and transforms it to another vector or to a value used to solve a practical problem.
- **Ex:** clustering, outlier detection

Clustering

- An application of unsupervised learning
- Automatically assigning a **label** to examples
- Dividing the examples into a number of **groups/clusters** such that examples in the same groups are more similar to other examples in the same group than those in other groups.

An example of clustering

Reinforcement

- The machine is capable of perceiving the state of the environment around as a feature vector.
- The machine can execute actions in every state.
- Different action brings different rewards → move the machine to the other state
- **Goal:** to learn a policy (function f ~ model in supervised learning) that takes the feature vector of a state as input and outputs an optimal action (=action maximizes the expected average reward)
- **Ex:** game playing, robotics, logistics.

Introduction to machine learning

- **Outline:**
  1. What is machine learning?
  2. Types of machine learning
  3. **Data for machine learning**
  4. Machine learning for classification

Importance of data

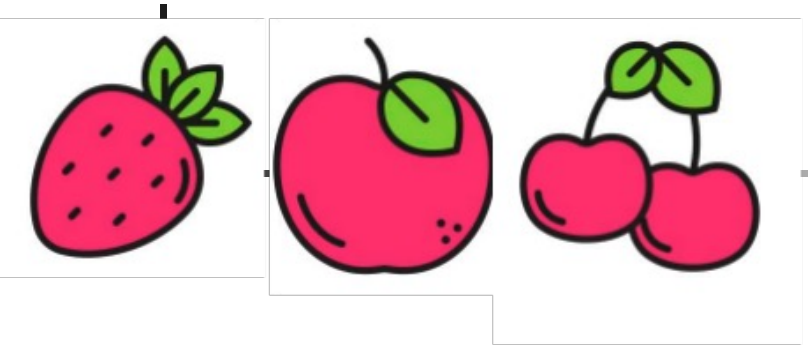
- ML depends heavily on data.
- In every ML/AI projects, data preparation takes most of time
- Data in unorganized format is not useful for machines to ingest the useful information.

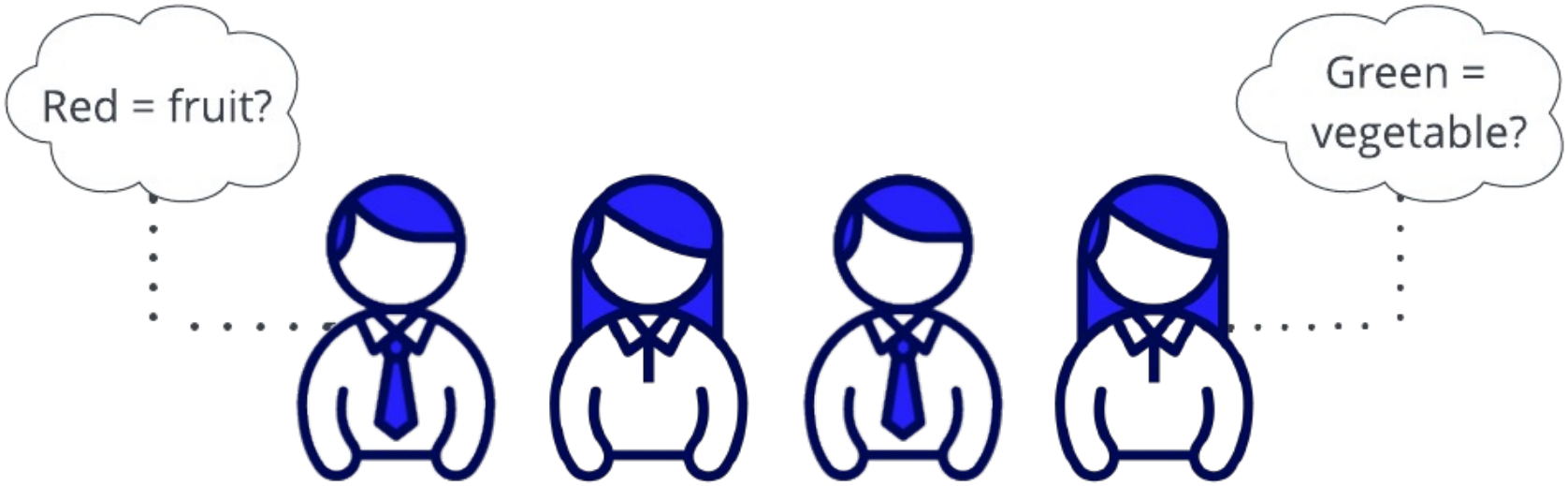
Ex: self-driving car crash 2017 in Florida, Amazon's AI recruiting tool "learnt" gender bias

- Flawed data can make a ML system harmful.

Fruits

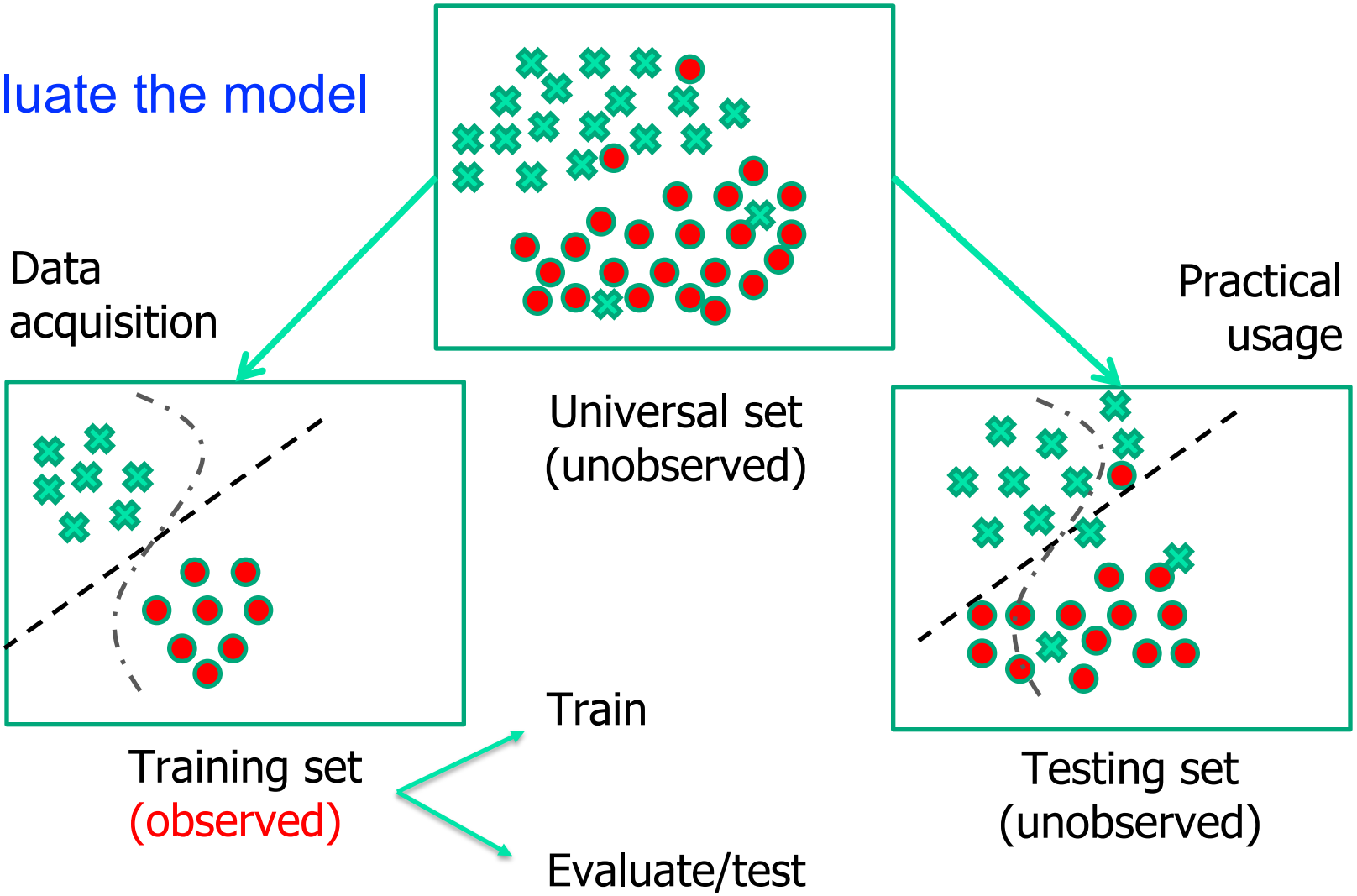
Vegetables





Data is used for…

- Train the model
- Evaluate the model



What factors make a good dataset?

- The right quantity
- The approach to split data
- The past history
- Domain expertise (Two key qualities: independence and identical distribution)
- The right kind of data transformation

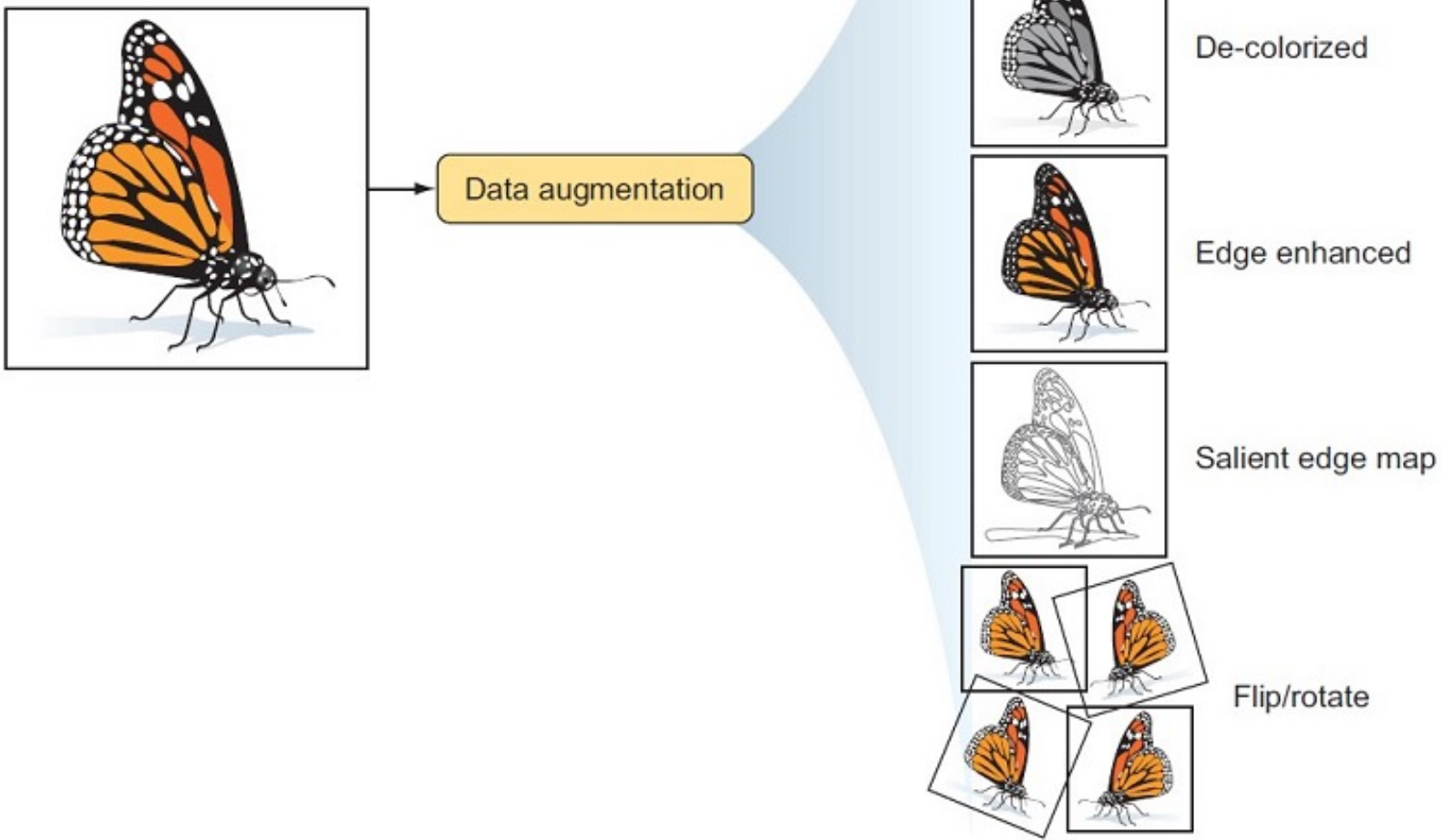
https://www.promptcloud.com/blog/what-to-look-for-in-training-dataset/

Dataset structure

- Dataset comprises data and labels:
  - Data: array [m, k] stores the k-D feature vectors of m objects
  - Labels: contain the m object labels
- Label types:
  - Integer numbers
  - String (class name)
  - Soft: real numbers in interval [0,1]
  - Target: numeric values in interval $(-\infty, +\infty)$

How to build dataset?

- Start small and reduce the complexity of the data.
- Articulate the problem early (i.e., classification, detection, ranking,…)
- Establish data collection mechanisms
- Check the data quality (human errors, technical problems, missing features, adequate?, imbalanced?)
- Format data
- Clean data
- Segmentation
- Complete **feature engineering**


Data augmentation
Original image
De-texturized
De-colorized
Edge enhanced
Salient edge map
Flip/rotate

An example

Iris dataset (cơ sở dữ liệu hoa diên vĩ)

Iris dataset

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	3505150

- Perhaps the best known database
- The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant.
- Inputs: sepal length in cm, sepal width in cm, petal length in cm, petal width in cm
- Outputs: Iris Setosa, Iris Versicolour, Iris Virginica

Introduction to machine learning

- **Outline:**
  1. What is machine learning?
  2. Types of machine learning
  3. Data for machine learning
  4. **Machine learning for classification**

Basic terms of classification

- **Classification in pattern recognition system:** to statistically categorize the feature vectors extracted from input patterns into a given number of classes
- **Classifier:** algorithm that maps each feature vector to a specific class (image-based)
- **Classification model:** draws some conclusions from the input values given for training → predict the class label for the new data

Types of classification

- **Binary classification:** two possible outcomes
  - Ex: spam email detection
  - **Multi-class classification:** more than two classes, each sample is assigned to one and only one target label
  - Ex: a flower can be rose or daisy but not both
- **Multi-label classification:** each sample is mapped to a set of target labels
  - Ex: a song can be about person, homeland, and love

Implementation of classification

- Template matching (~ grid-by-grid comparison)
- Machine learning

Implementation of classification

- **Template matching:**

- Implemented by comparing the testing feature vectors to training feature to determine the similarity
- The similarity between two data points is measured by distance

For a point $(x_1, x_2, ..., x_n)$ and a point $(y_1, y_2, ..., y_n)$,

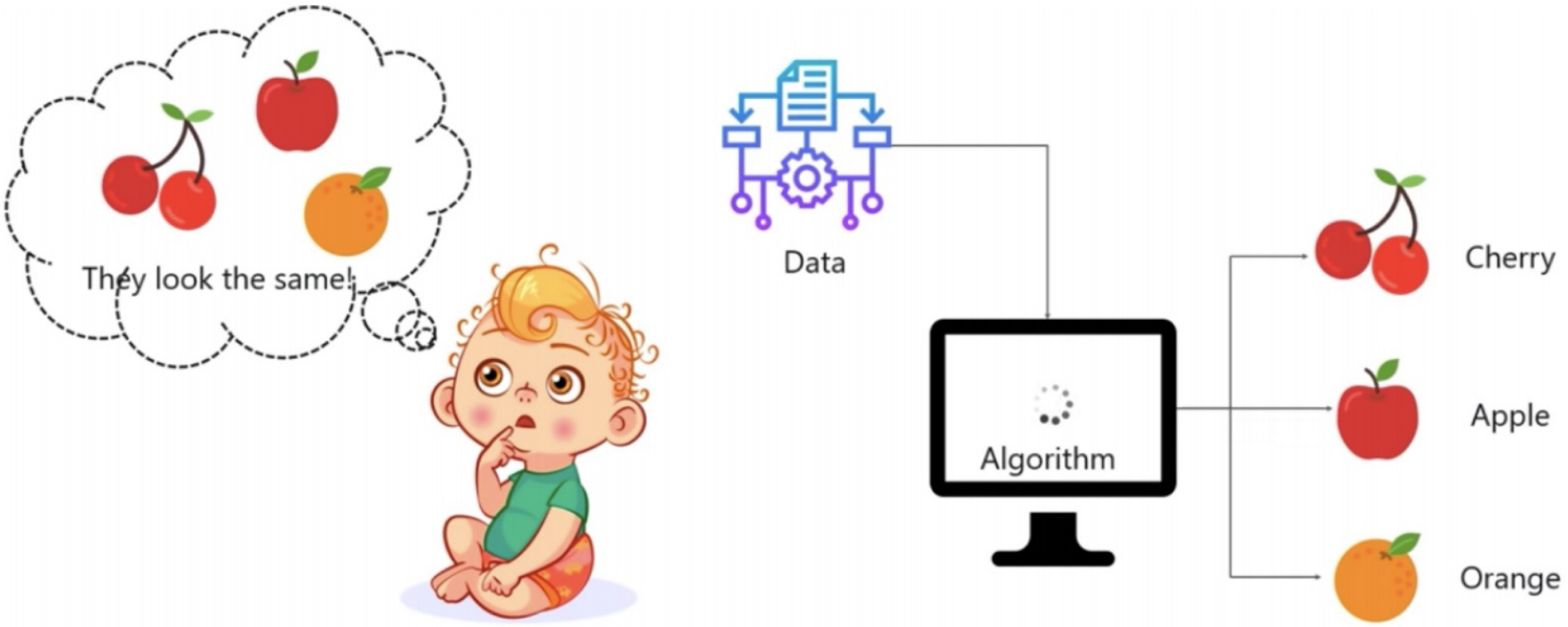
Manhattan distance = 1-norm distance $= \sum_{i=1}^{n} |x_i - y_i|$

Euclidean distance = 2-norm distance $= \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2}$

Implementation of classification

- **Machine learning:**

➢ Implemented by applying machine learning algorithms



Bài tập

- Cho biết vector đặc trưng rút trích từ một bức ảnh chụp quả cam là {2.7887, 6.5063, 9.4425, 9.8402, -19.5930} và ảnh chụp quả táo là: {2.6743, 5.7745, 9.9031, 11.0016, -21.4722}.
- Dùng phương pháp so khớp mẫu, với hai kiểu tính khoảng cách là norm1 và norm2, em hãy phân loại bức ảnh chụp trái cây có vector đặc trưng là {2.6588, 5.7358, 9.6682, 10.7427, -20.9914} là ảnh chụp loại quả nào? Giả sử bức ảnh này chỉ chụp một loại quả cam hoặc táo.

Bài tập mở rộng

- **Đề:** Phân loại rượu vang Ý bằng phương pháp template matching
- **Cơ sở dữ liệu:** https://github.com/MukeshTirupathi/Wine-Classifier-Italy?tab=readme-ov-file

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

- **Phân chia dữ liệu:** 1 test, toàn bộ còn lại là train