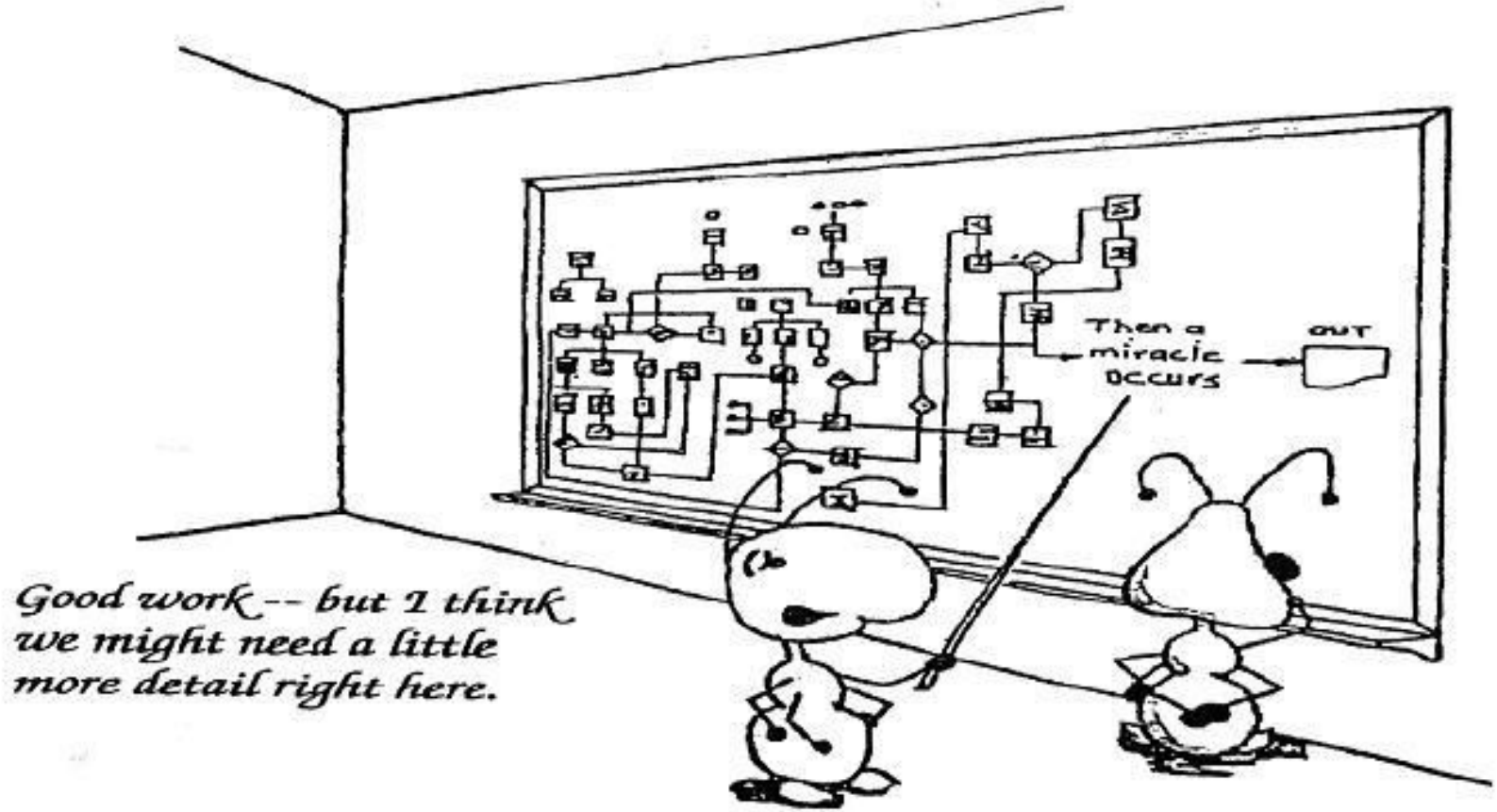
Machine Learning Basics and Usages in Bioinformatics



Burcin Buket OGUL
burcinbo@uio.no
Senior Engineer at USIT

What is ~~Artificial~~ Intelligence?

Refers to the ability to:

- acquire and apply knowledge,
- solve problems,
- learn from experience, and adapt to new situations.

Human intelligence encompasses a wide range of cognitive abilities, including reasoning, problem-solving, learning, perception, language understanding, and emotional intelligence.

What is Artificial Intelligence?

"AI is a field of science concerned with building computers and machines that can **reason**, **learn**, and **act** in such a way that would normally require human intelligence or that involves **data** whose scale exceeds what humans can analyze."

Google

"AI is the capability of a computer system to **mimic** human-like cognitive functions such as **learn**ing and **problem-solving**."

Microsoft

"AI leverages computers and machines to **mimic** the **problem-solving** and **decision-making** capabilities of the human mind"

IBM

"AI makes it possible for machines to **learn** from **experience**, adjust to new inputs and perform **human-like** tasks."

SAS

AI

- mimic human
- reasoning, problem solving, decision making
- **learn from data/experience**

ChatGPT

Siri

Shazam Again

Google Translate

What is ~~Machine~~ Learning?

- "the acquisition of knowledge or skills through experience, study, or by being taught."

What is Machine Learning?

- [Arthur Samuel, 1959]
  - Field of study that gives computers
  - the ability to learn without being explicitly programmed

- Instead of writing a program by hand, we collect lots of examples that specify the correct output for a given input

ARTIFICIAL INTELLIGENCE VS MACHINE LEARNING VS DEEP LEARNING

1 Artificial Intelligence

Development of smart systems and machines that can carry out tasks that typically require human intelligence

2 Machine Learning

Creates algorithms that can learn from data and make decisions based on patterns observed

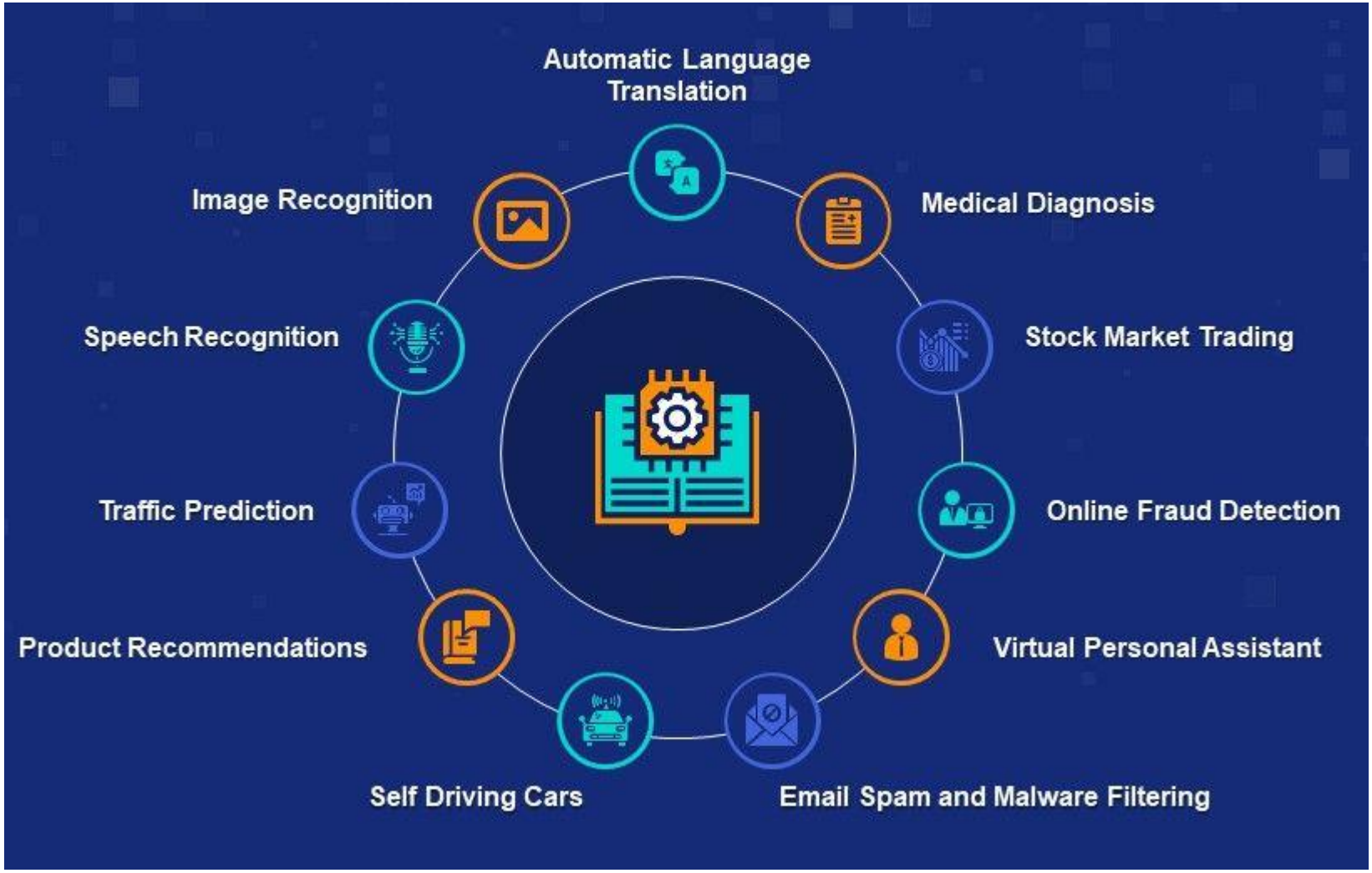
Require human intervention when decision is incorrect

3 Deep Learning

Uses an artificial neural network to reach accurate conclusions without human intervention

SINGAPORE COMPUTER SOCIETY

Applications of AI

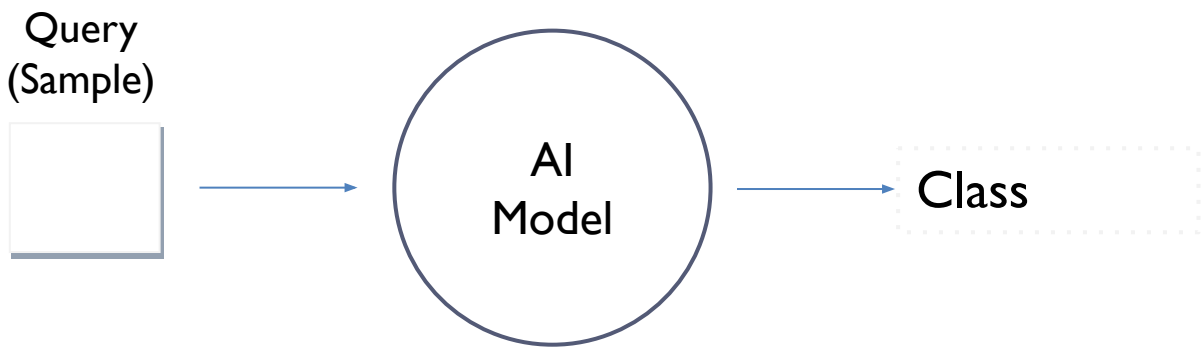


Two broad categories

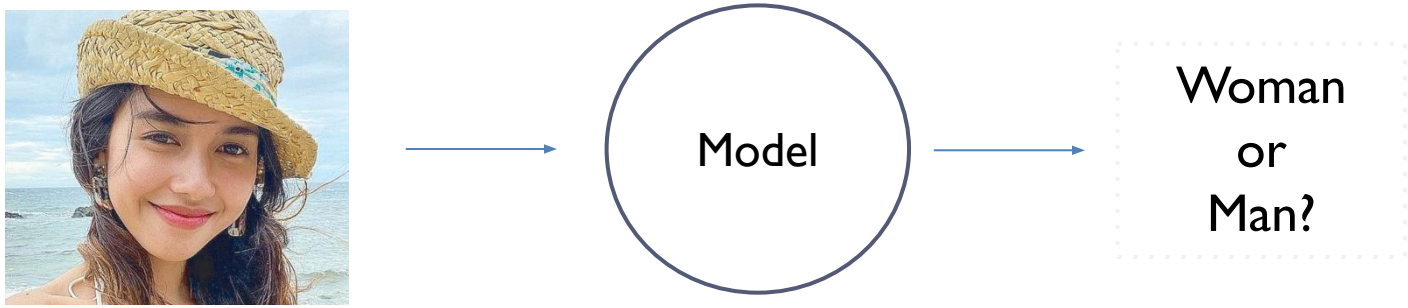
Discriminative (Predictive) AI – More traditional

Generative AI – like ChatGPT

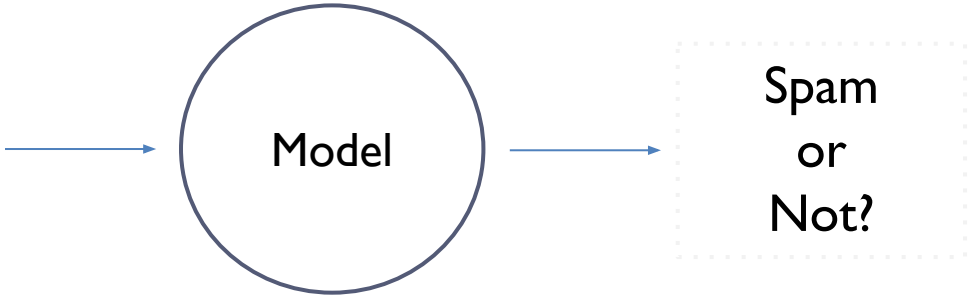
Discriminative/Predictive AI
Main Problem: Classification



Example 1:
Gender Classification from Image



Example 2:
Spam Classification from E-mail

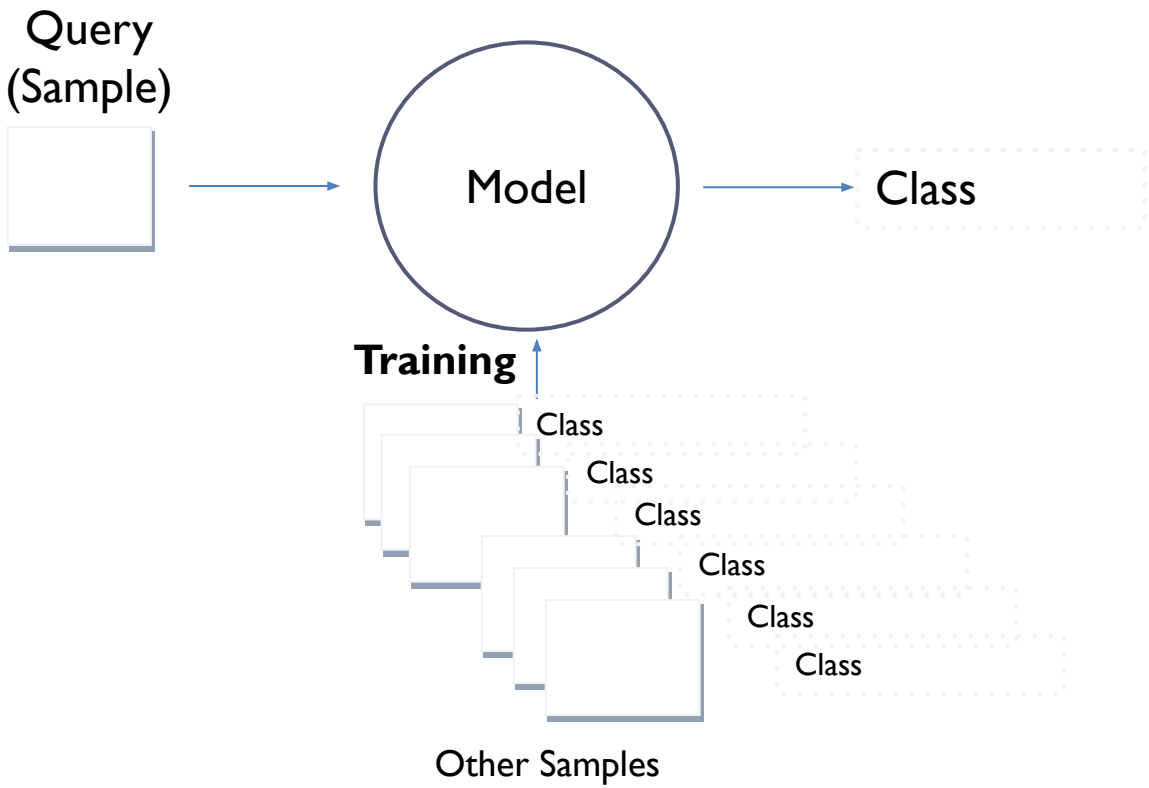


Example 3:
Music Classification from Song

Model

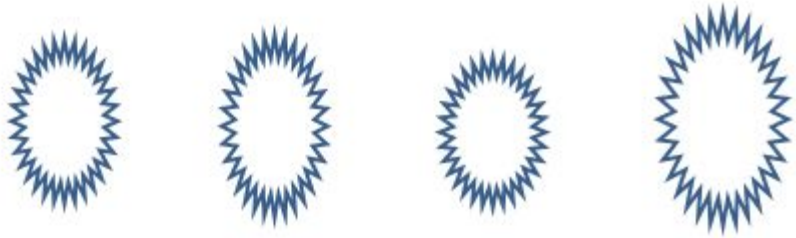
Pop
or
Rock?

General Learning Model



An easier example: Avocado or Mango?

Avocados



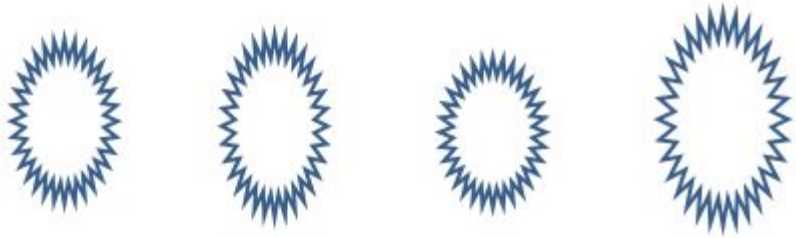
Height (cm): 9 10 8 12

Mangos

Height (cm): 3.5 4 2 3

An easier example: Avocado or Mango?

Avocados

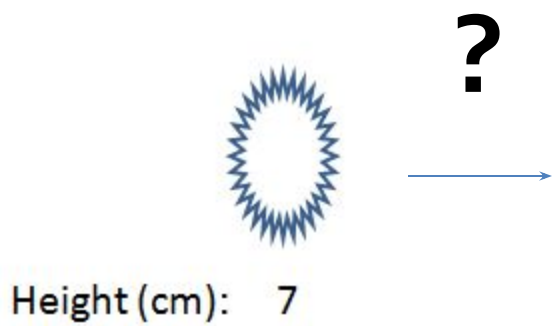


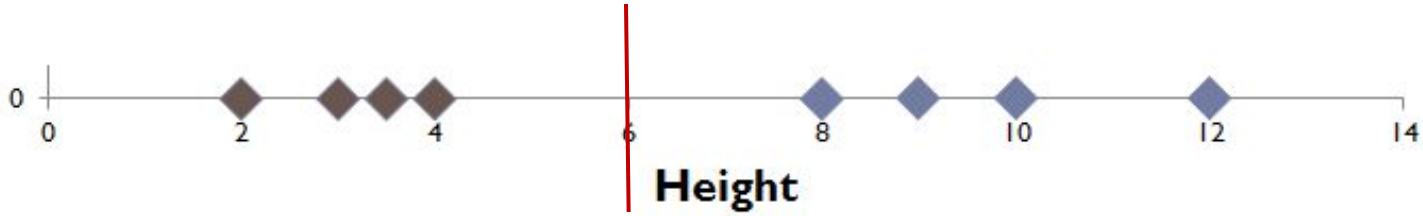
Height (cm): 9 10 8 12

Mangos

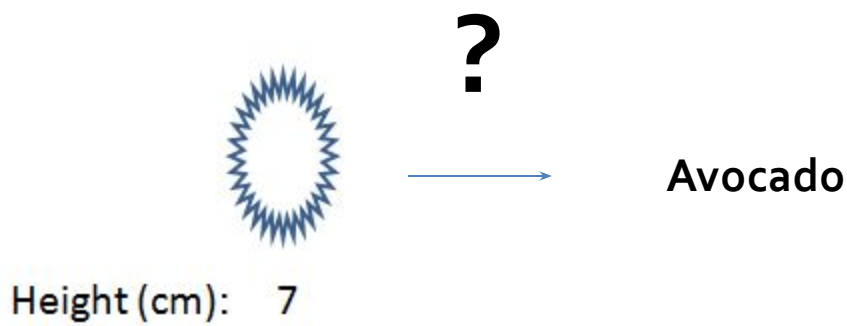
Height (cm): 3.5 4 2 3

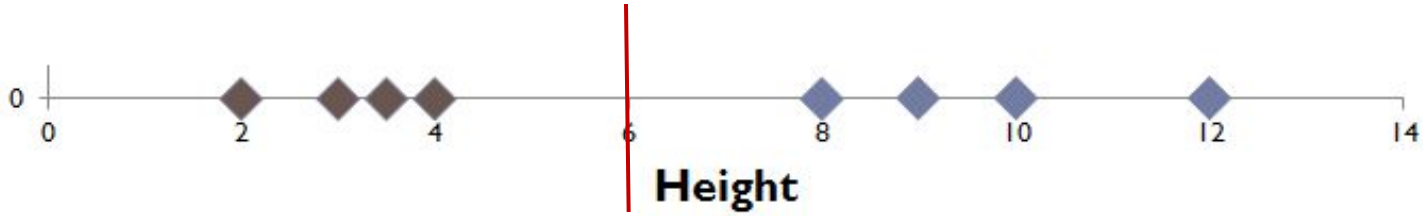
An easier example: Avocado or Mango?





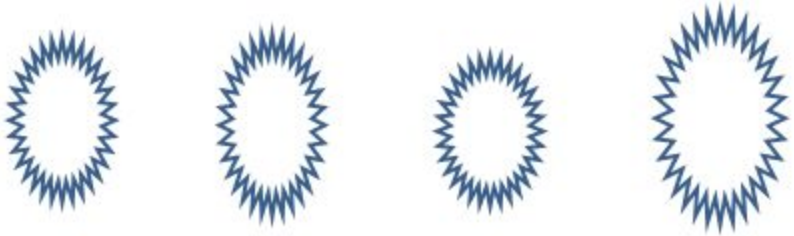
An easier example: Avocado or Mango?





An easier example: Avocado or Mango?

Avocados



Height (cm):	9	10	8	12
Roughness:	4	6	5	4

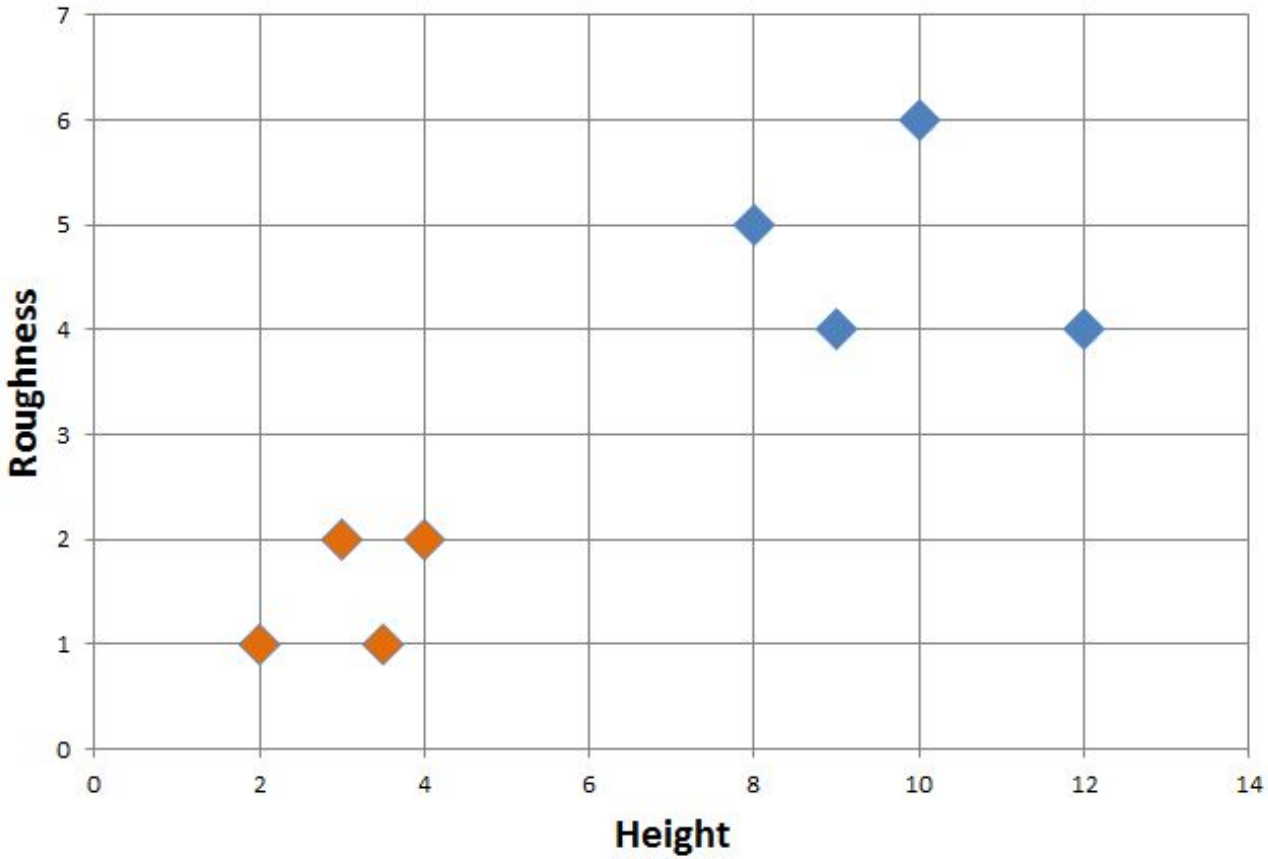
Mangos

Height (cm):	3.5	4	2	3
Roughness:	1	2	1	2

An easier example: Avocado or Mango?

Avocados

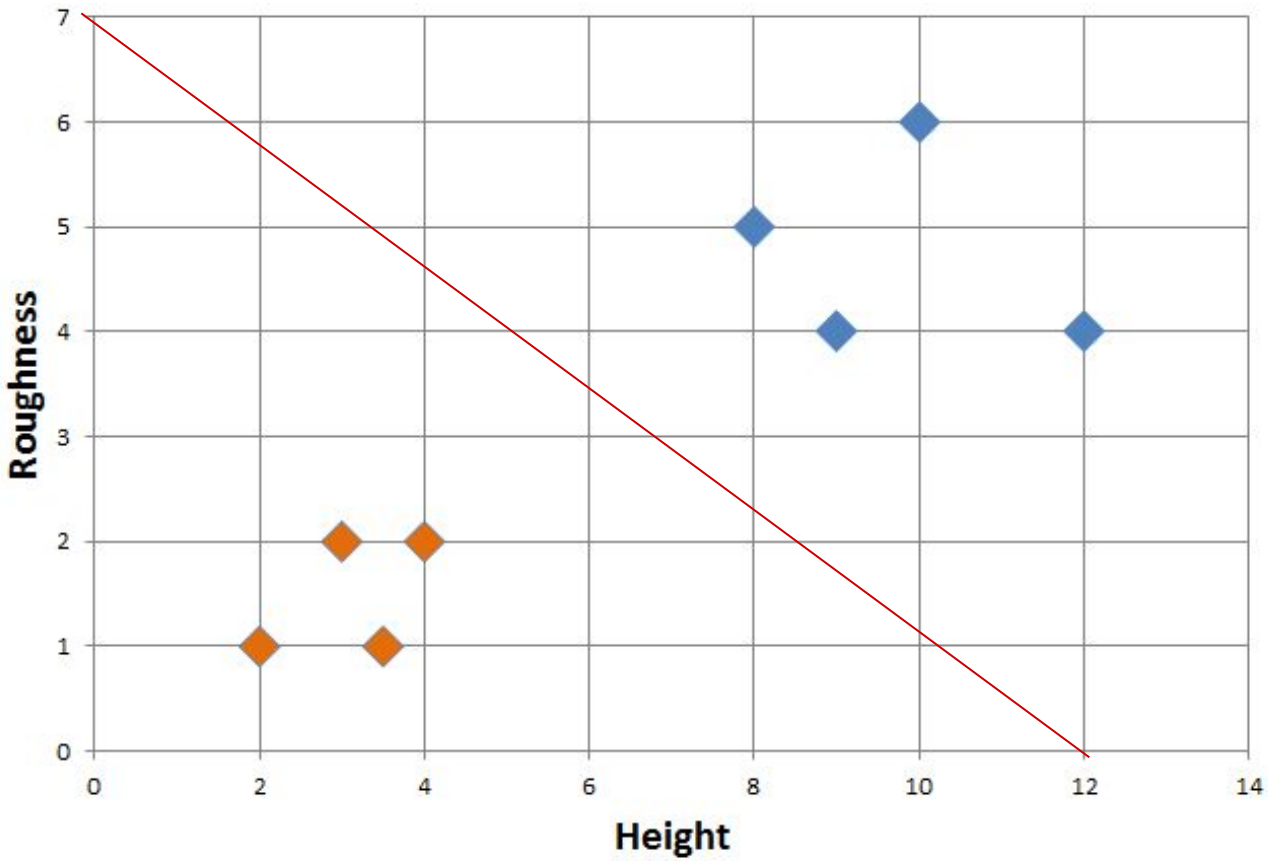
Mangos



An easier example: Avocado or Mango?

Avocados

Mangos

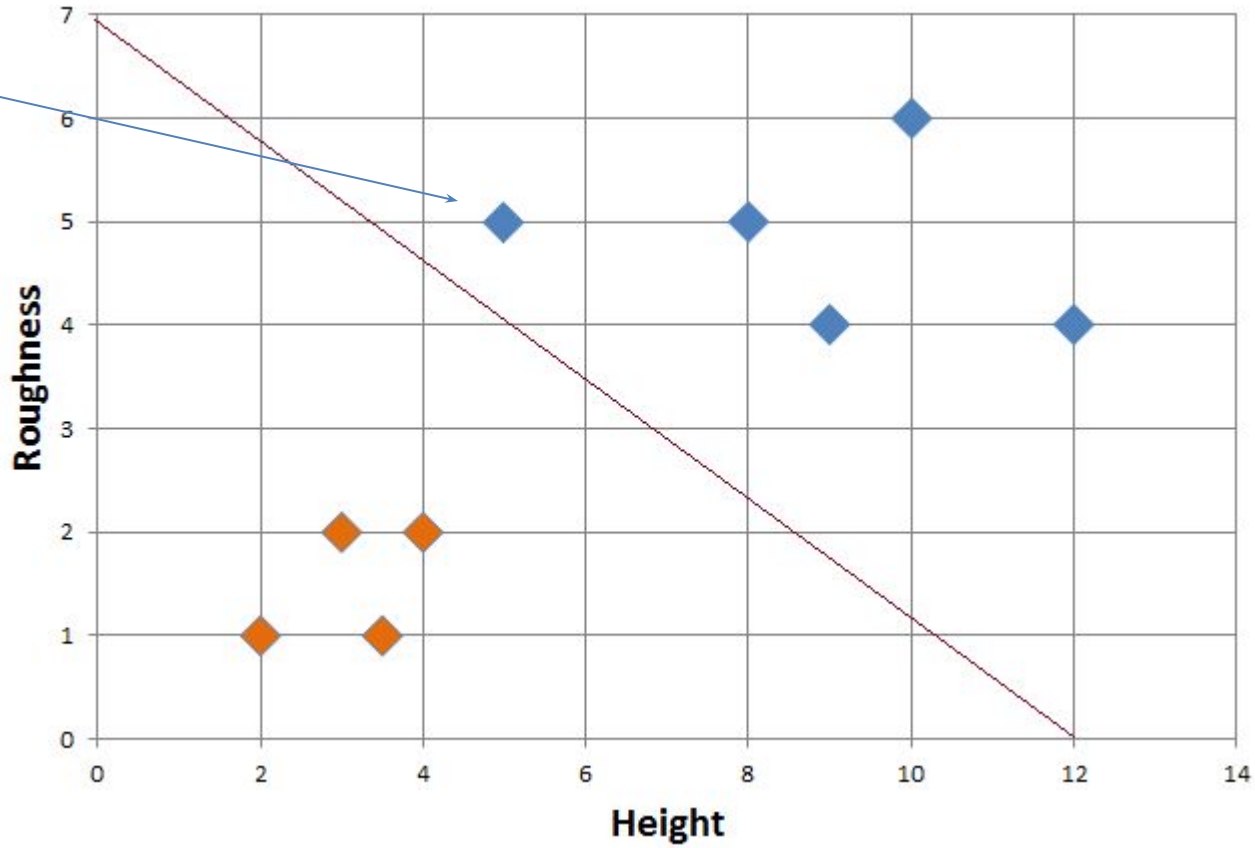


An easier example: Avocado or Mango?

?

Height (cm): 5

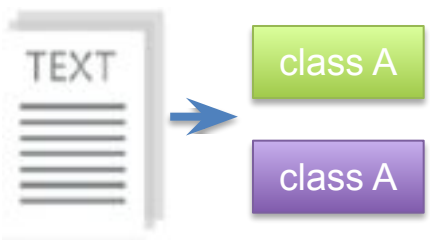
Roughness: 5



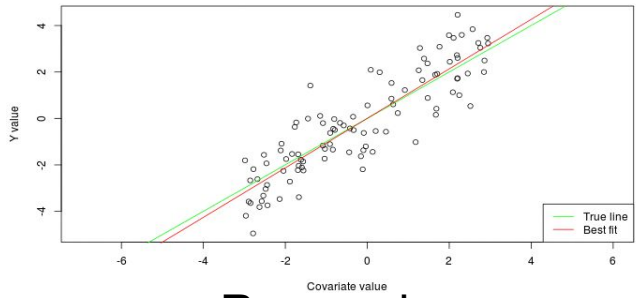
Types of Learning

Supervised: Learning with a **labeled training** set
Example: email *classification* with already labeled emails

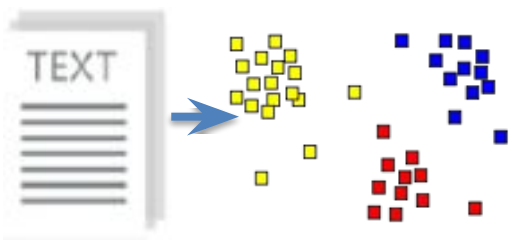
Unsupervised: Discover **patterns** in **unlabeled** data
Example: *cluster* similar documents based on text



Classification



Regression



Clustering

Anomaly Detection
Sequence labeling
…

http://mbjoseph.github.io/2013/11/27/measure.html

Tasks

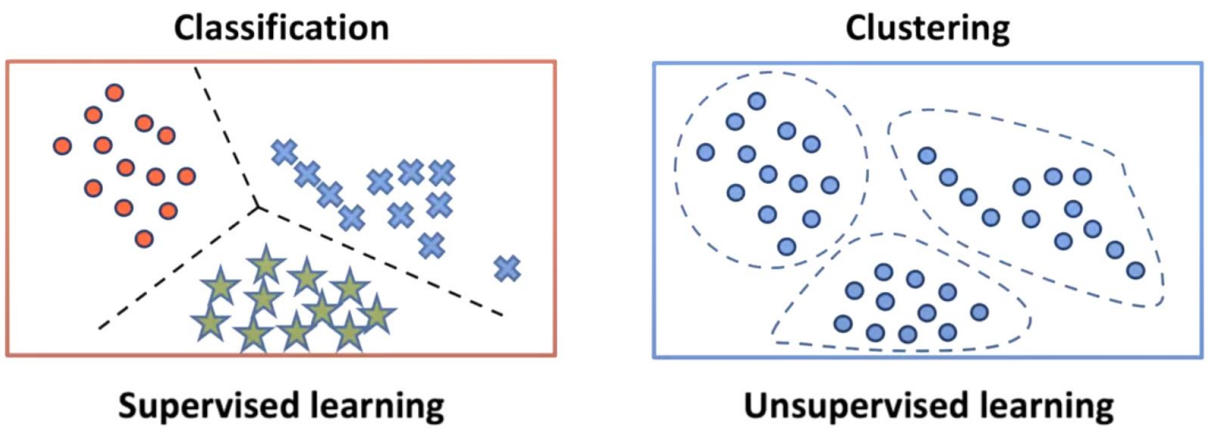
Supervised Learning

x → Classification → y Discrete

x → Regression → y Continuous

Unsupervised Learning

x → Clustering → y Discrete ID



Pattern Classification

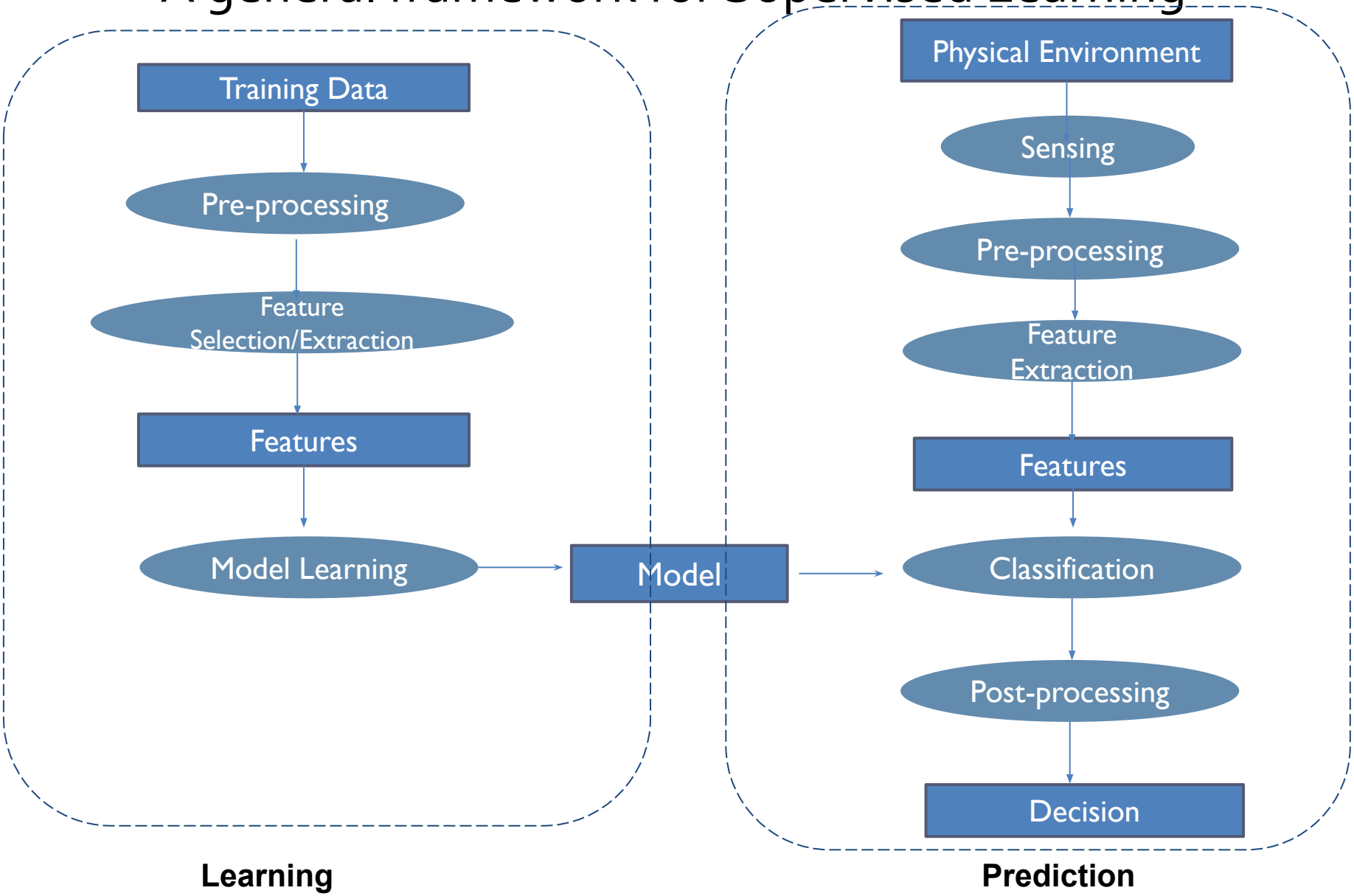
A **pattern** is an entity, vaguely defined, that could be given a name, e.g.

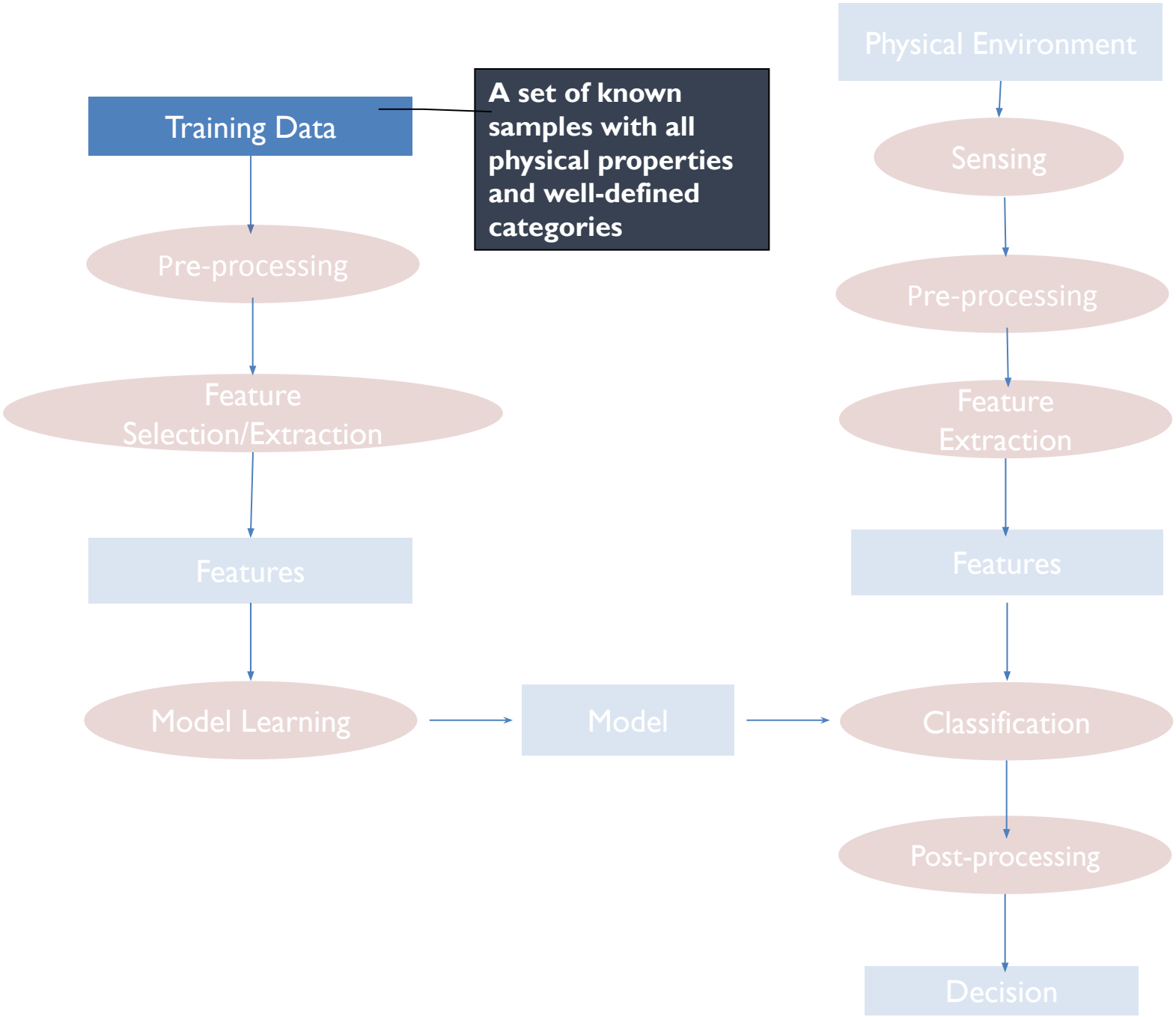
- Fingerprint image
- Handwritten word
- Face in a picture
- Speech signal
- E-mail text
- DNA sequence
- Protein sequence
- Gene expression data

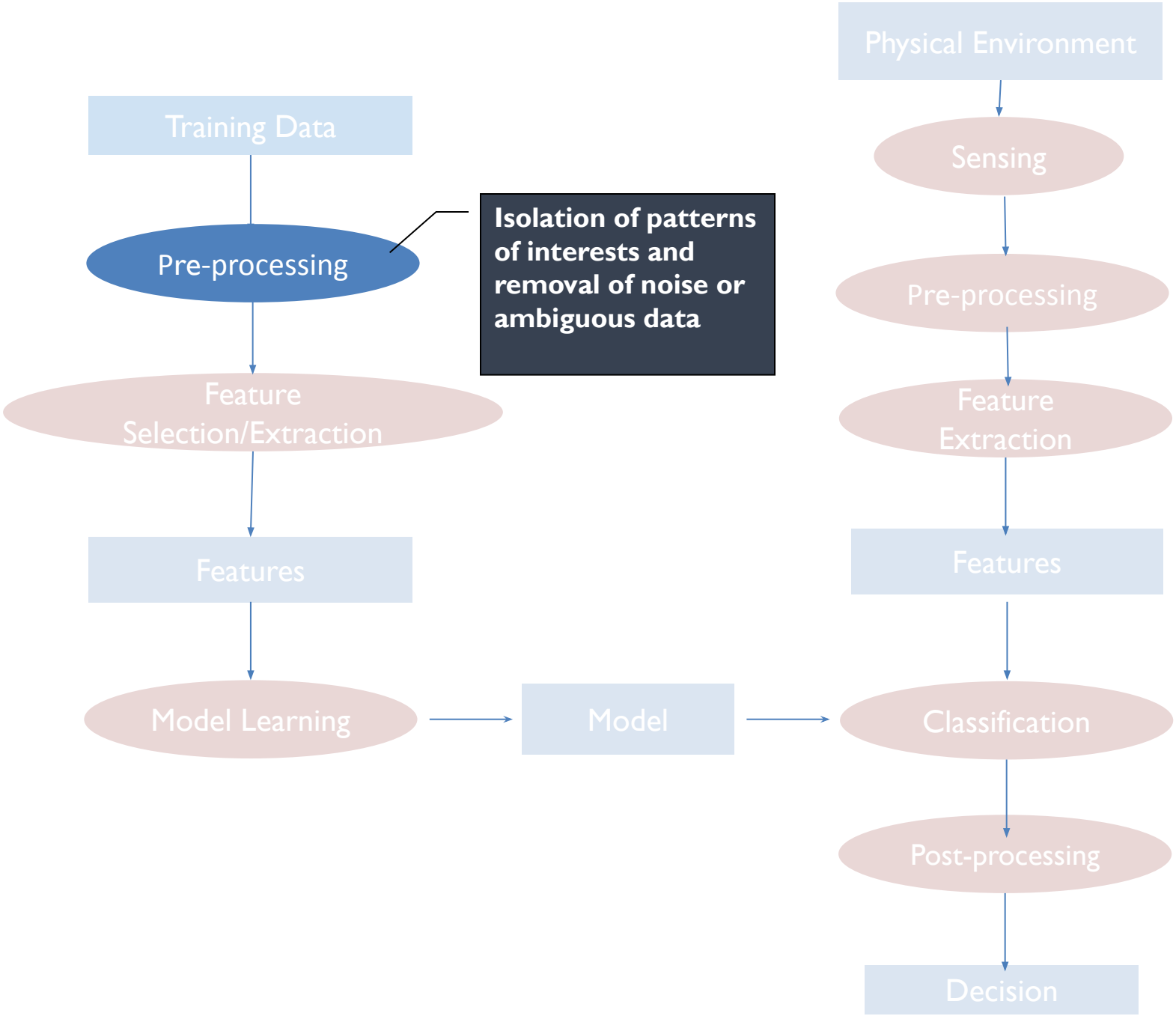
Pattern classification (recognition) is the study of how machines can

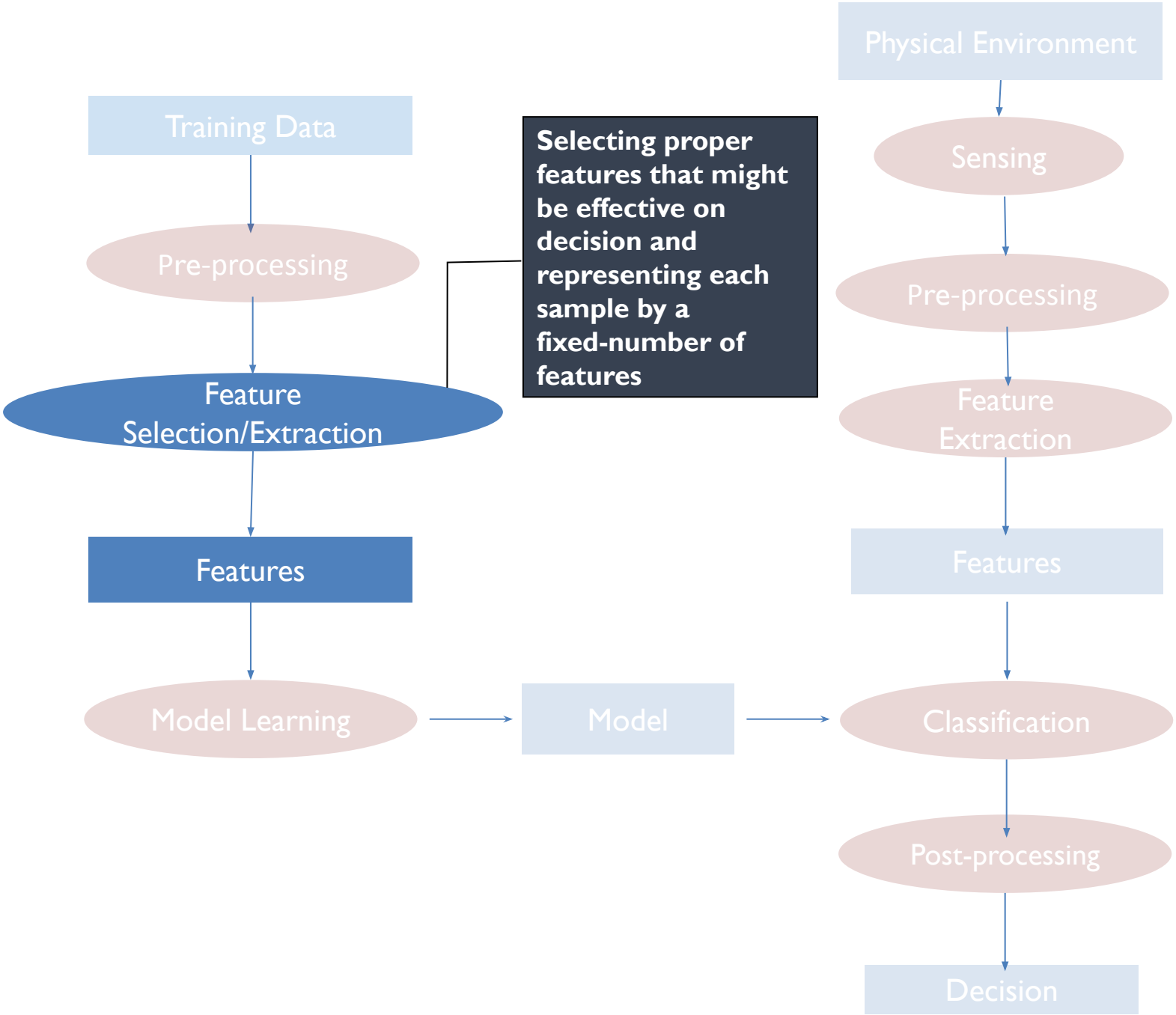
- Learn how to distinguish patterns of interest
- Make reasonable decisions about the categories of the patterns

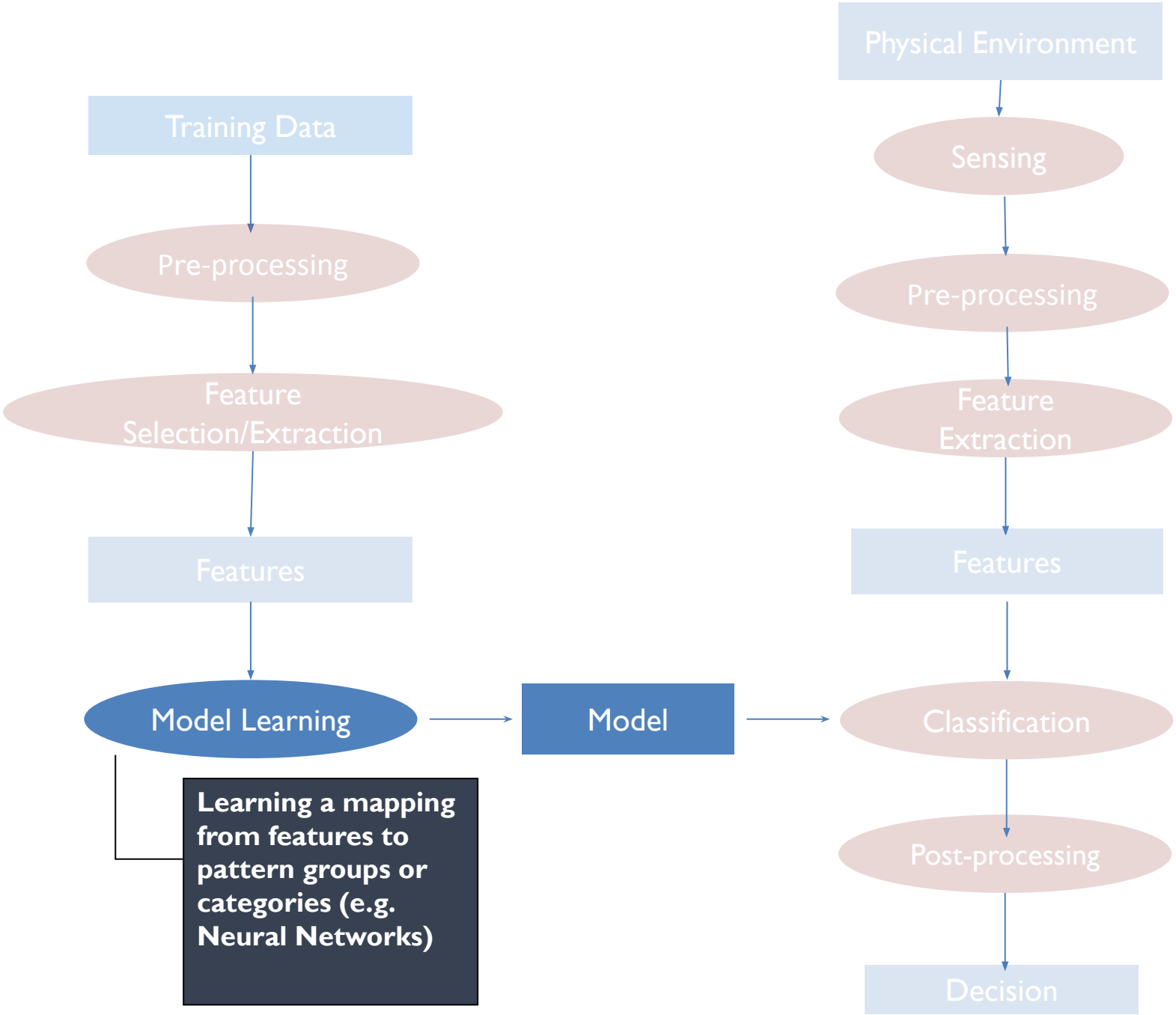
Machine Learning:
A general framework for Supervised Learning

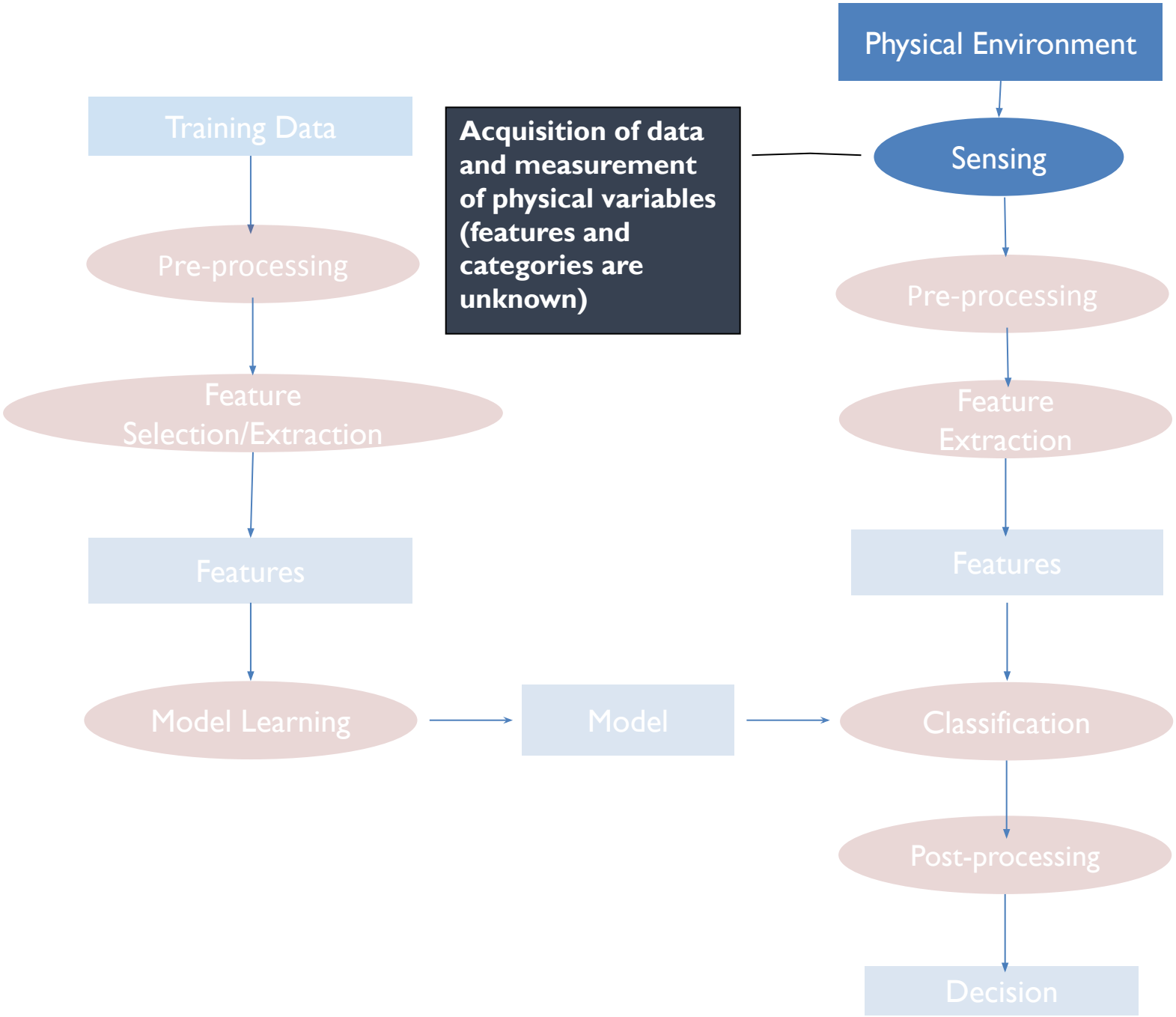


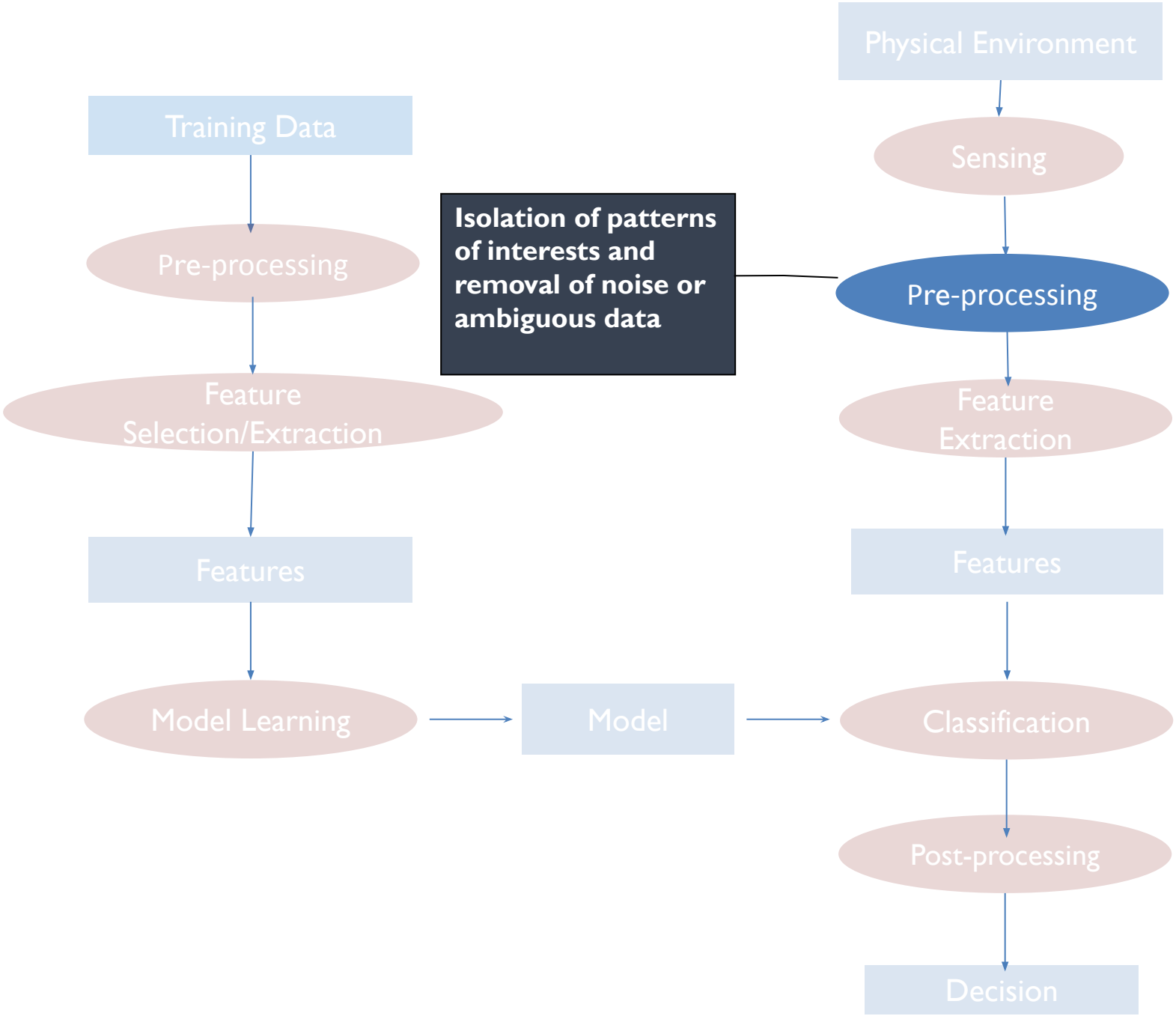

Training Data
A set of known samples with all physical properties and well-defined categories
Pre-processing
Feature Selection/Extraction
Features
Model Learning
Model
Physical Environment
Sensing
Pre-processing
Feature Extraction
Features
Classification
Post-processing
Decision

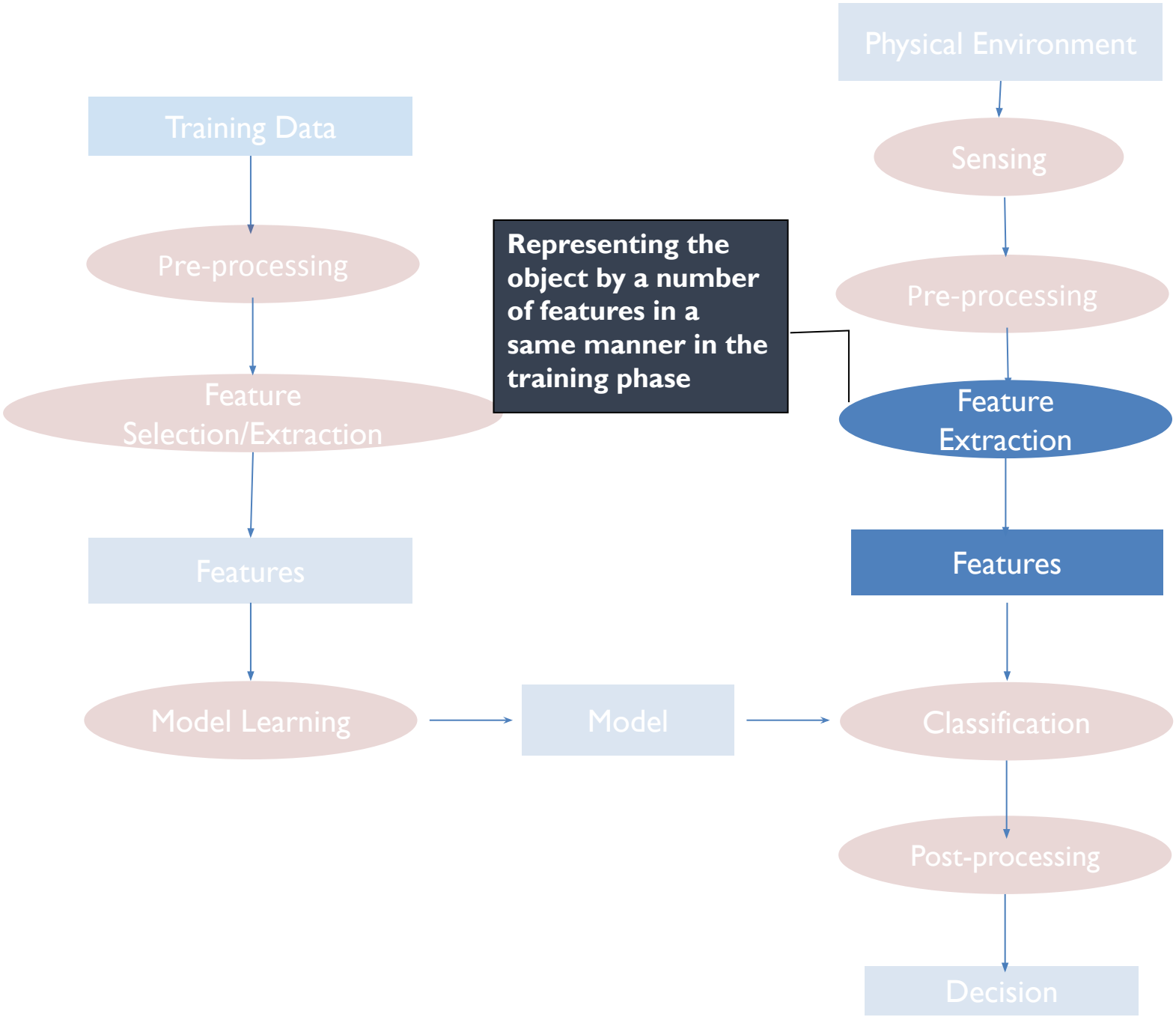

Training Data
Pre-processing
Isolation of patterns of interests and removal of noise or ambiguous data
Feature Selection/Extraction
Features
Model Learning
Model
Physical Environment
Sensing
Pre-processing
Feature Extraction
Features
Classification
Post-processing
Decision

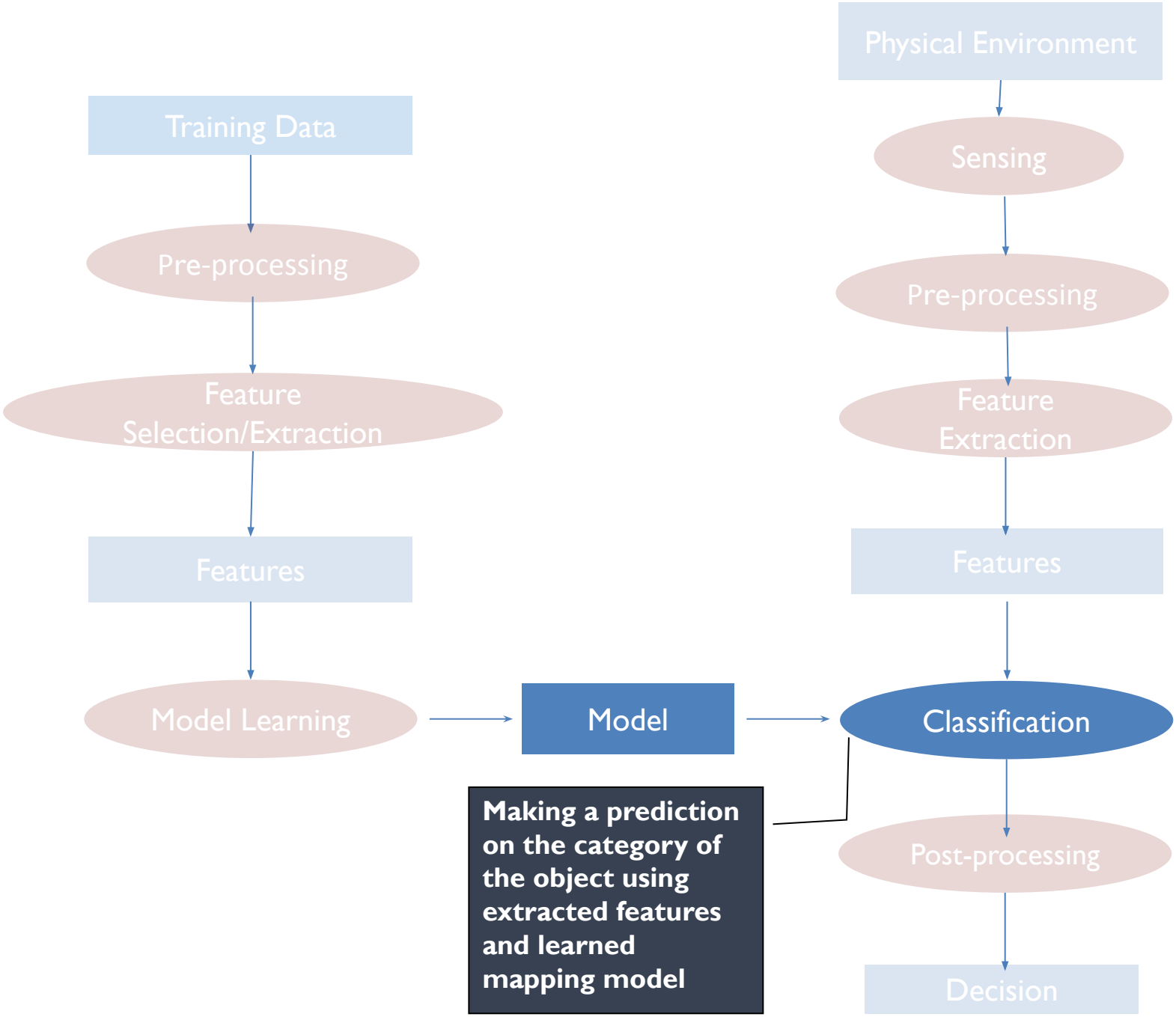

Training Data
Pre-processing
Feature Selection/Extraction
Features
Model Learning
Model
Selecting proper features that might be effective on decision and representing each sample by a fixed-number of features
Physical Environment
Sensing
Pre-processing
Feature Extraction
Features
Classification
Post-processing
Decision

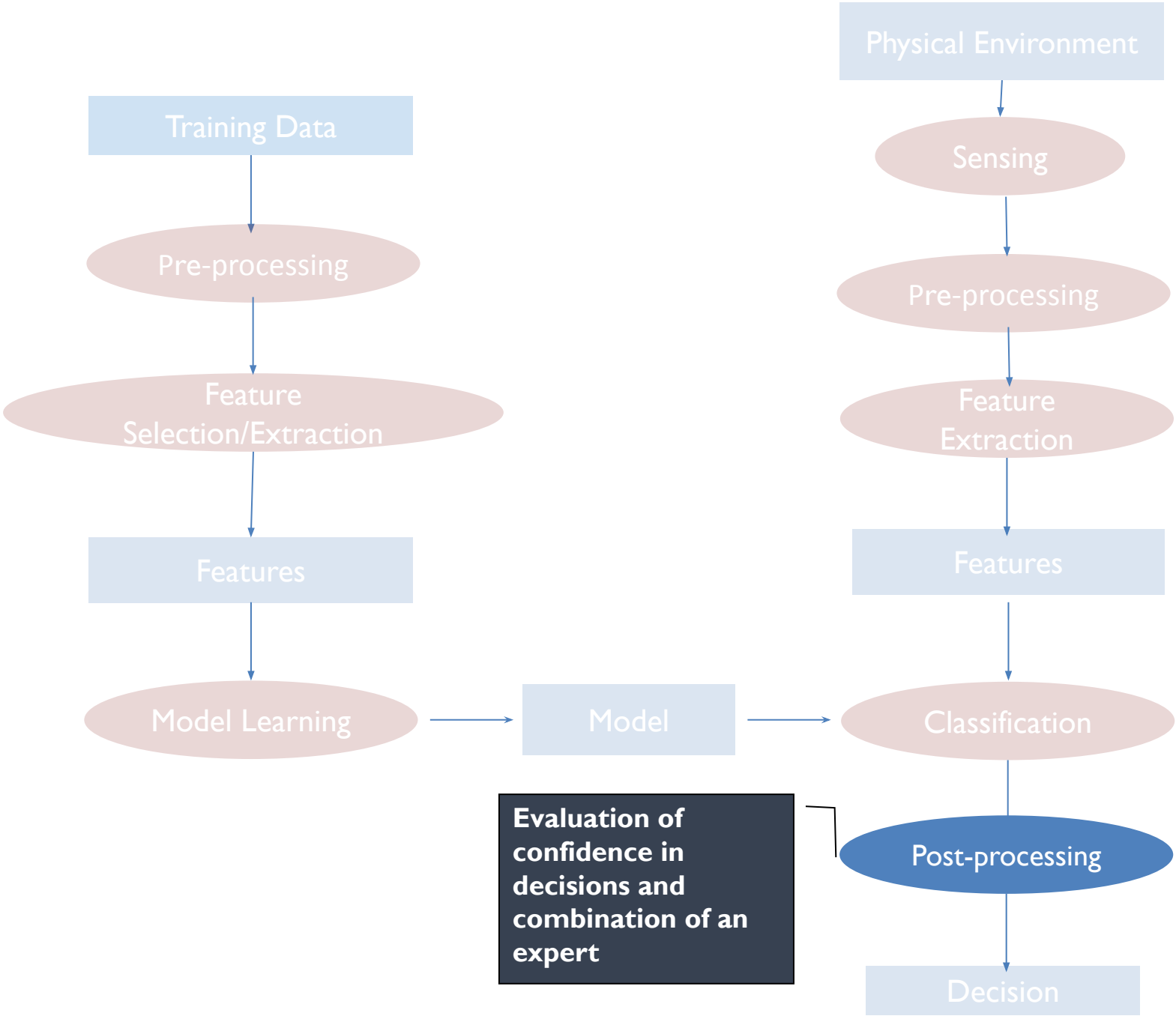

Training Data
Pre-processing
Feature Selection/Extraction
Features
Model Learning
Learning a mapping from features to pattern groups or categories (e.g. Neural Networks)
Model
Physical Environment
Sensing
Pre-processing
Feature Extraction
Features
Classification
Post-processing
Decision

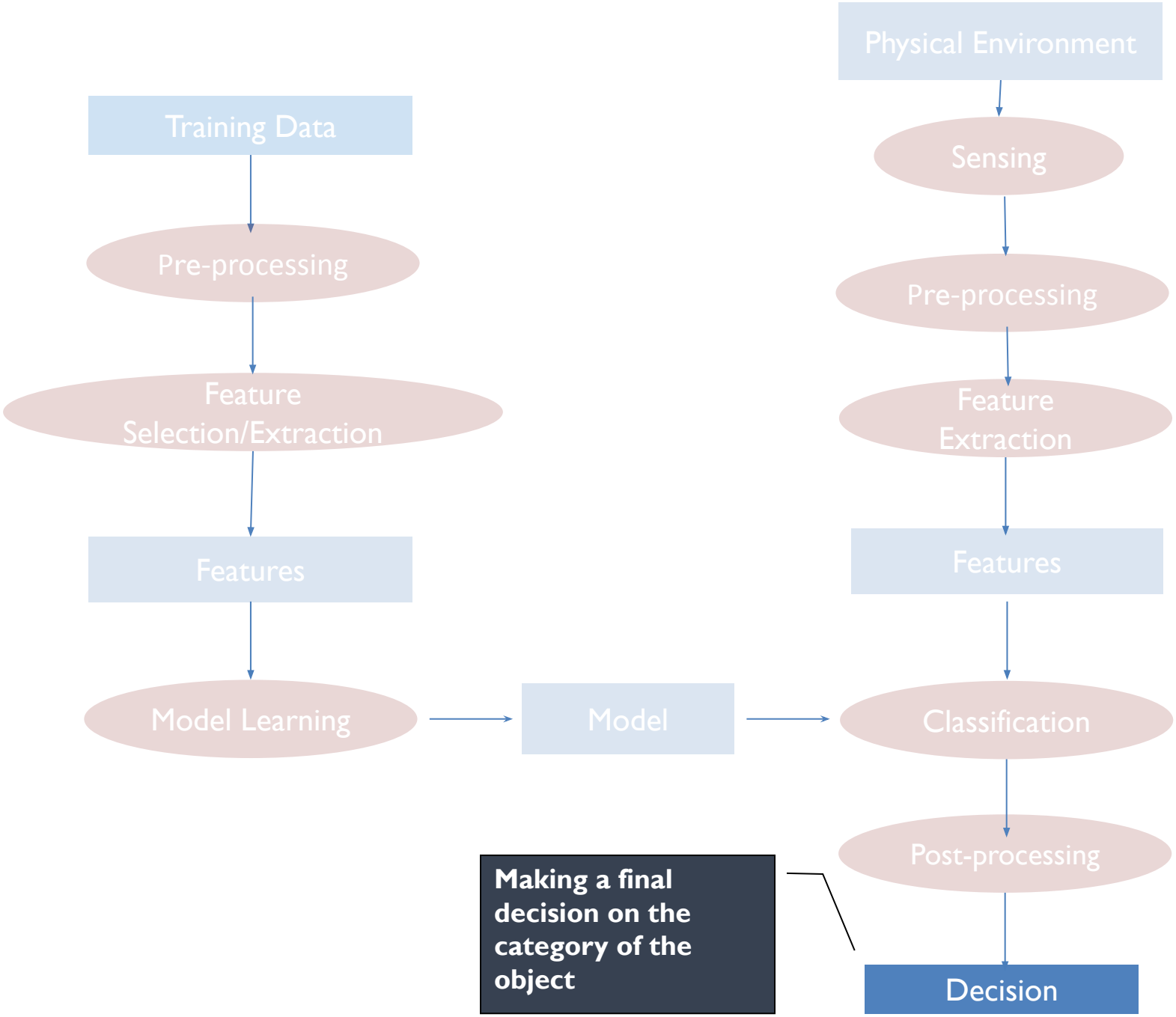

Physical Environment
Training Data
Acquisition of data and measurement of physical variables (features and categories are unknown)
Sensing
Pre-processing
Pre-processing
Feature Selection/Extraction
Feature Extraction
Features
Features
Model Learning
Model
Classification
Post-processing
Decision


Physical Environment
Training Data
Sensing
Isolation of patterns of interests and removal of noise or ambiguous data
Pre-processing
Pre-processing
Feature Selection/Extraction
Feature Extraction
Features
Features
Model Learning
Model
Classification
Post-processing
Decision


Training Data
Pre-processing
Feature Selection/Extraction
Features
Model Learning
Model
Physical Environment
Sensing
Pre-processing
Feature Extraction
Features
Classification
Post-processing
Decision
Representing the object by a number of features in a same manner in the training phase


Training Data
Pre-processing
Feature Selection/Extraction
Features
Model Learning
Model
Physical Environment
Sensing
Pre-processing
Feature Extraction
Features
Classification
Post-processing
Decision
Making a prediction on the category of the object using extracted features and learned mapping model


Training Data
Pre-processing
Feature Selection/Extraction
Features
Model Learning
Model
Physical Environment
Sensing
Pre-processing
Feature Extraction
Features
Classification
Post-processing
Decision
Evaluation of confidence in decisions and combination of an expert


Training Data
Pre-processing
Feature Selection/Extraction
Features
Model Learning
Model
Physical Environment
Sensing
Pre-processing
Feature Extraction
Features
Classification
Post-processing
Decision
Making a final decision on the category of the object

An example study

Subcellular Localization Prediction

Bioinformatics

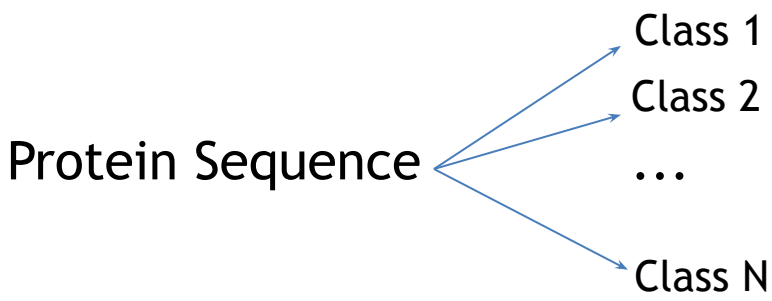
- Analyzing current data to infer new biological knowledge using computational techniques

- **DATA?**
  - Sequence: DNA, RNA, Protein…
  - Structure: Protein structure, RNA structures
  - Interaction: Protein-protein interaction, TF binding, microRNA targets
  - Experiment result: Gene expression, metabolomics…

Sequence Classification

- Whole sequence classification
- Sub-sequence classification
- Residue classification

Whole Sequence Classification



e.g.

- Structural classes: all-α, all-β, α+β,…
- Subcellular localizations: mito, cyst, ext,…
- Folds: globin-like, barstar-like, ferritin-like,…

An example study

Subcellular Localization Prediction

Motivation

Subcellular localization is a key property in functional annotation of proteins.

Automated categorization of proteins into their localizations based on sequence is an important challenge.

Subcellular Localization Prediction

Training Data

2427 annotated eukaryotic proteins

4 known locations

- Nuclear
- Cytoplasmic
- Mitochondrial
- Extracellular

Subcellular Localization Prediction

Pre-processing

Proteins that have ambiguous locations are removed

Subcellular Localization Prediction

Feature Selection

Amino acid composition

Dipeptide composition

n-peptide compositions

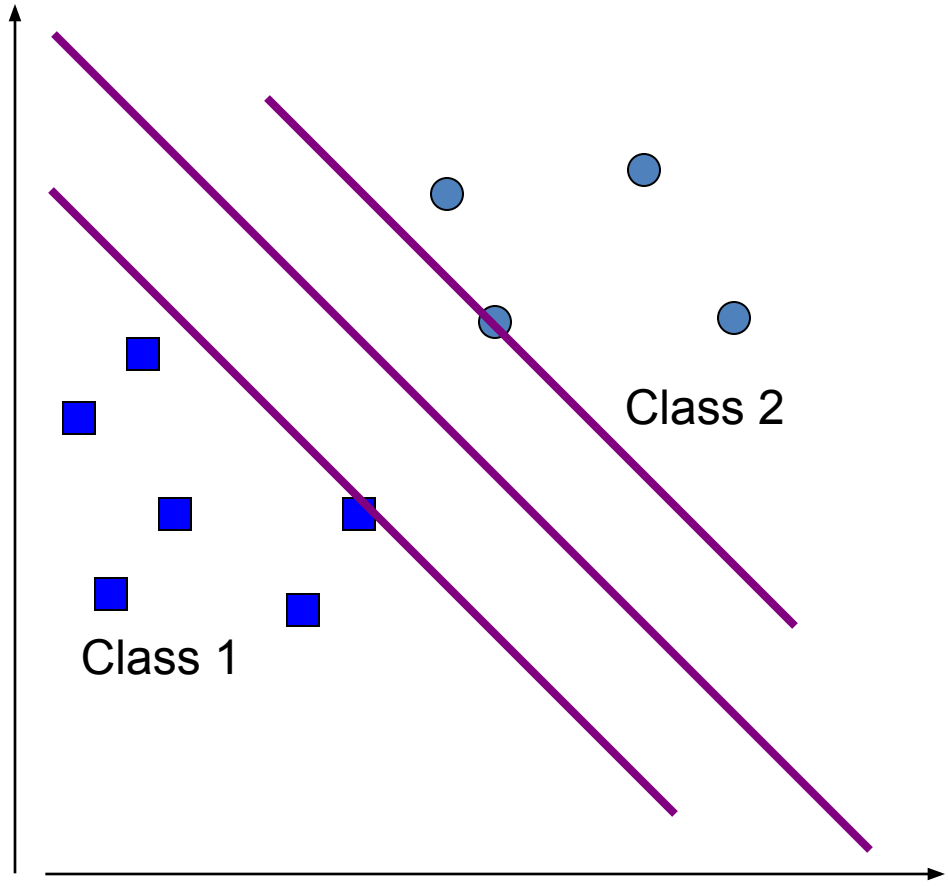
Physicochemical properties

Sequence similarity scores with other known proteins

Sub-sequence similarity scores (first k residue at N-terminal)

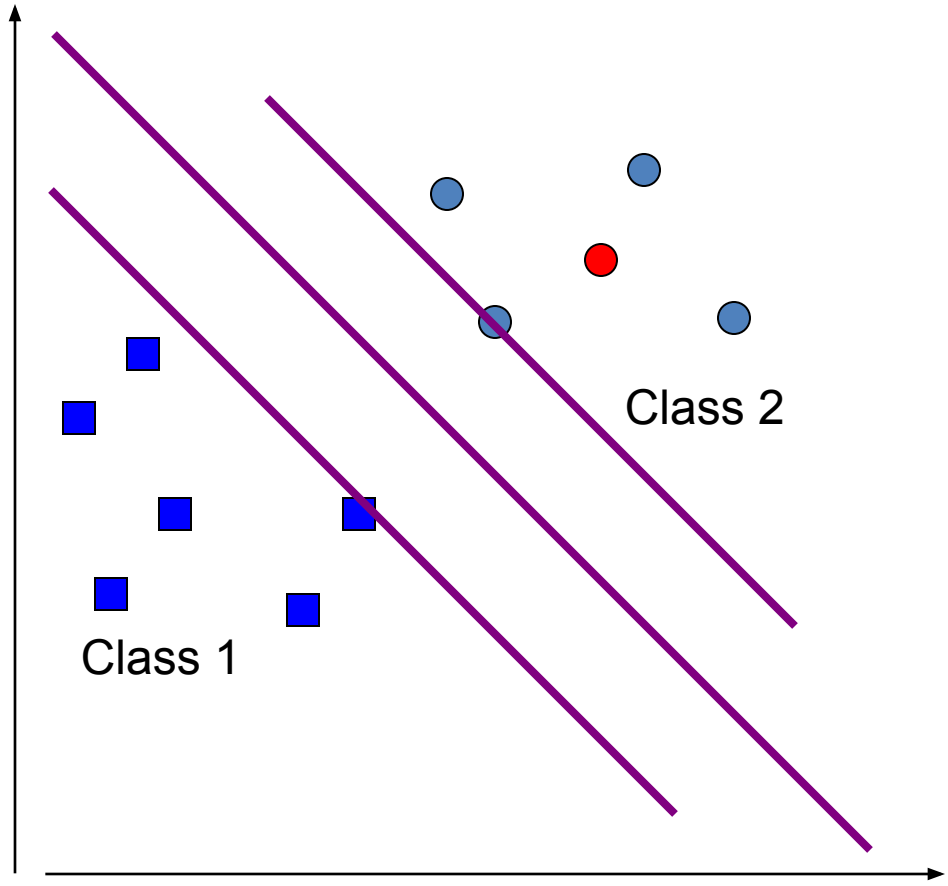
Subcellular Localization Prediction

Model Learning



Subcellular Localization Prediction

Classification



Subcellular Localization Prediction

Post-processing

Evaluate the outputs of the classifiers based on different feature sets

Subcellular Localization Prediction

Decision

Select the most confident output

Subcellular Localization Prediction

Results

Location	Amino acid composition (a)	Dipeptide composition (b)	Biochemical properties (c)	Combination of a, b and c	*n*-peptide compositions
Nuclear	86.1	92.7	85.6	93.2	94.3
Cytoplasmic	76.9	80.2	74.6	80.6	84.5
Mitochondrial	55.5	58.8	59.2	65.1	66.4
Extracellular	76.0	79.0	76.6	83.4	88.9
Overall	78.1	82.9	78.8	84.6	87.1

Some remarks

Selection of

- training data
- feature representations
- learning models

"Deep Learning doesn't do different things,
it does things differently"

Supervised Learning

- Traditional pattern recognition models work with hand crafted features and relatively simple trainable classifiers.

Extract Hand Crafted Features → **Trainable Classifier** (e.g. SVM, Random Forrest) → **Output** (e.g. Outdoor Yes or No)

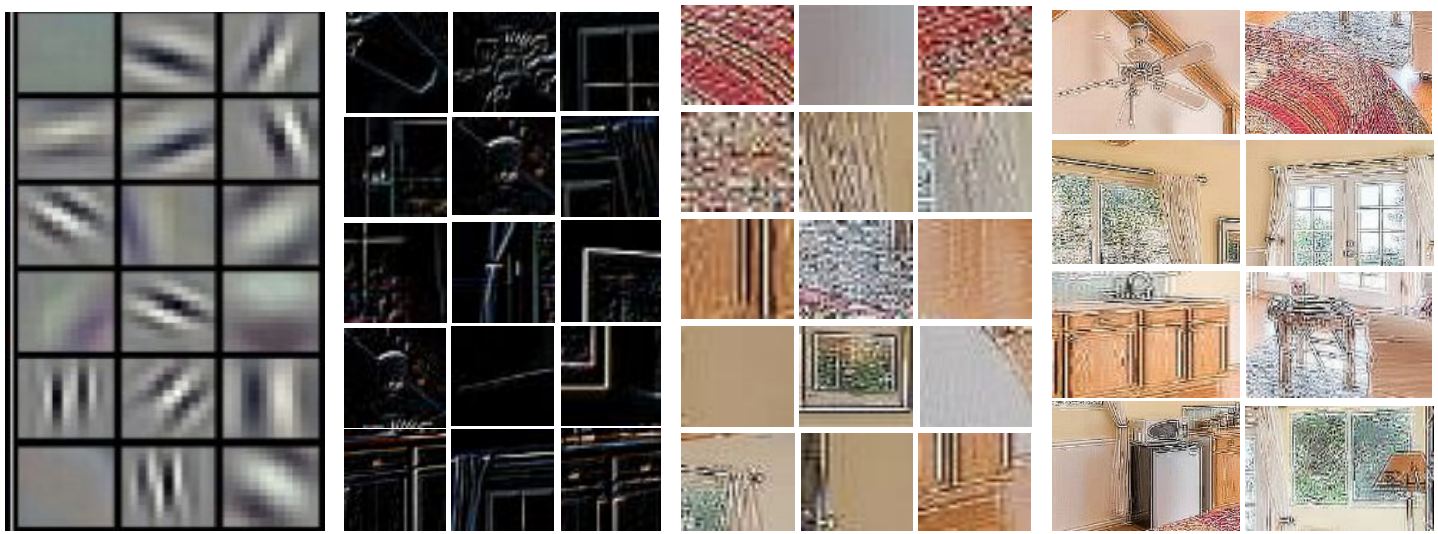
Limitations

- Very tedious and costly to develop hand crafted features.
- The hand-crafted features are usually highly dependents on one application.

Deep Learning

- Deep learning has an **inbuilt automatic multi stage feature learning process** that learns rich hierarchical representations (i.e. features).

Low Level Features → Mid Level Features → High Level Features → Trainable Classifier → **Output (e.g. outdoor, indoor)**

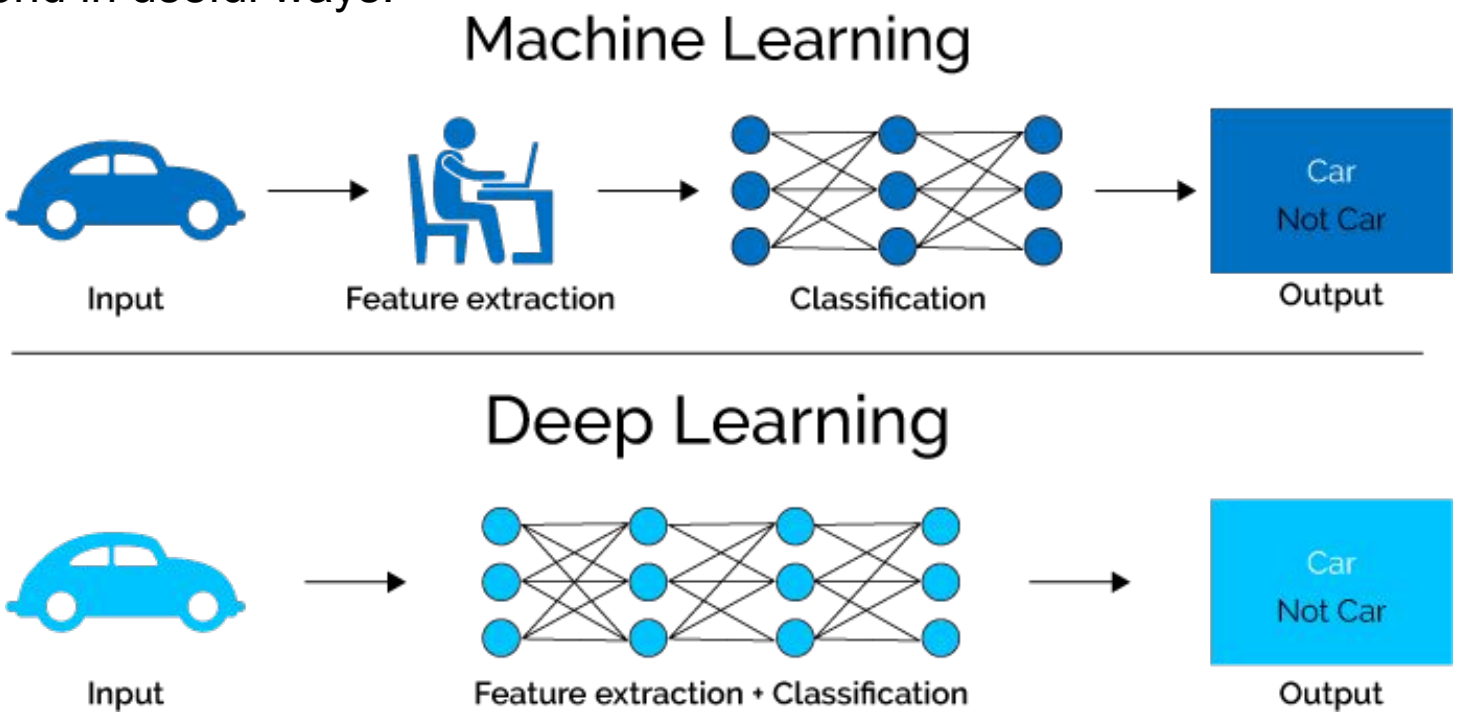


What is Deep Learning (DL) ?

A machine learning subfield of learning **representations** of data. Exceptional effective at **learning patterns**.

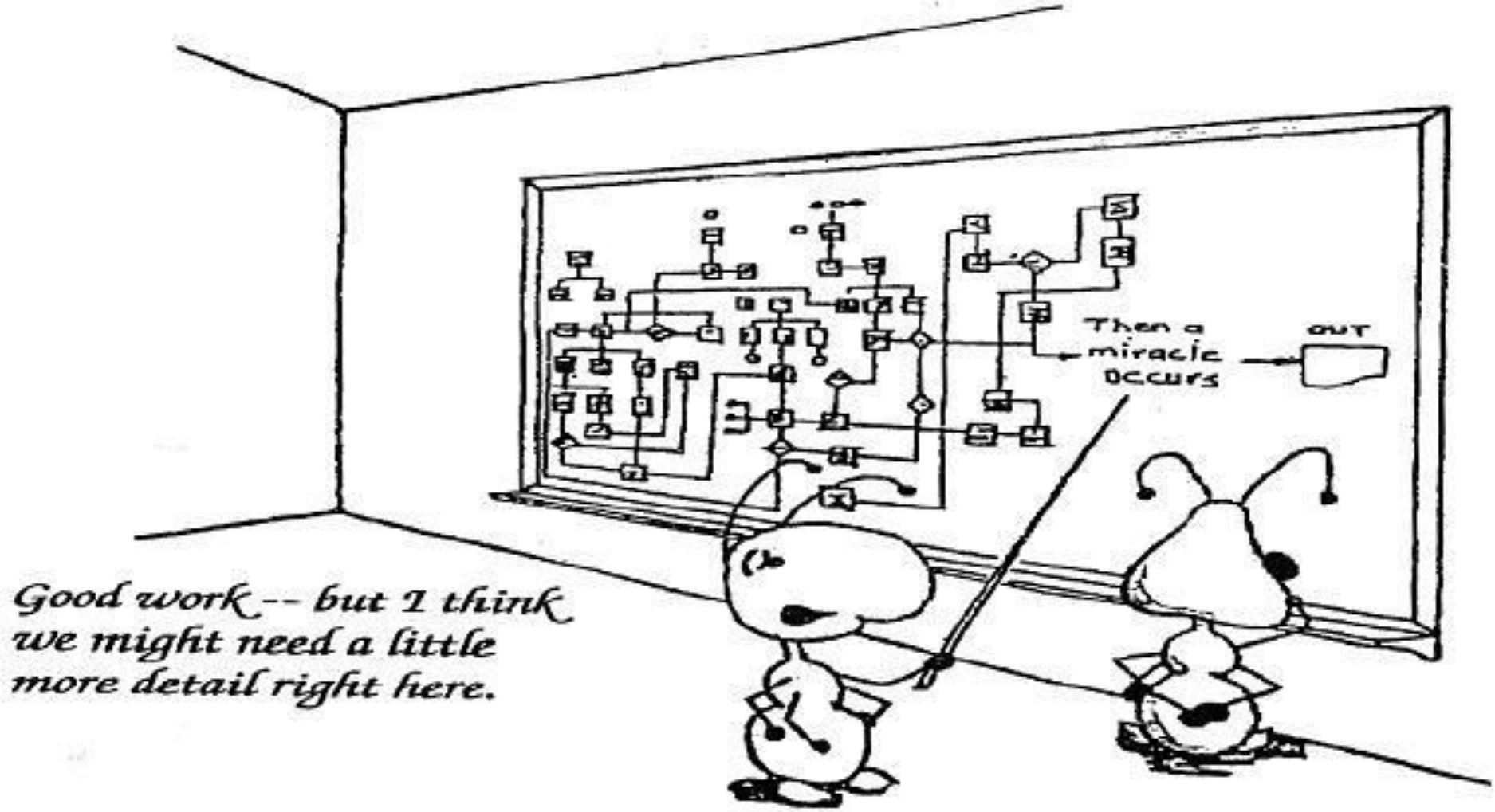
Deep learning algorithms attempt to learn (multiple levels of) representation by using a **hierarchy of multiple layers**

If you provide the system **tons of information**, it begins to understand it and respond in useful ways.



https://www.xenonstack.com/blog/static/public/uploads/media/machine-learning-vs-deep-learning.png

Yes it works, but how?



Thank you…