50.007 Machine Learning

Lecture 1

Introduction

Who are we?

Instructors

Prof. Yixiao Wang
Instructor
Weeks 1-3

Prof. Berrak Sisman
Instructor
Weeks 4-6

Prof. Roy Lee
Instructor
Weeks 8-12

Teaching Assistants

- Mihir Chhiber, mihir_chhiber@mymail.sutd.edu.sg
- Ho Xin Yi Felice, felice_ho@mymail.sutd.edu.sg

Outline

- Administrative details
- What is machine learning?
- Types of machine learning
- A case study for supervised learning
- Linear Classification

Administrative details

- **Course material:**
  - No required textbook
  - Recommended reading (from books or research papers) posted on eDimension
  - Lecture notes and slides: posted on eDimension
  - **Lecture slides are provided to help better understand the lecture notes and by no means should be considered as a replacement.**
  - For a better understanding of the course, please go through the lecture notes prior to each lecture.

- **Pre-requisite:**
  - Linear Algebra
  - Probability/statistics
  - Knowledge of Algorithms
  - Python Programming

Evaluation

- Homework (45%)
  - Programming and theory
  - Honour Code
    - Form study groups to work on homework
    - You can discuss with other classmates
    - Write-up solutions on your own
    - List names of anyone you talked to
- Project (23%)
- Midterm Exam (15%)
- Final Exam (15%)
- Participation (2%)

Course Goals

- Curious to discover more
- Confident of doing it yourself
- Contemplative of the theory
- Aware of the limitation

Acknowledgements

- MIT 6.036 Introduction to Machine Learning
- SUTD 50.007 Machine Learning (Prof. Liang Zheng)
- Stanford CS229 Machine Learning
- McGill COMP-652 Machine Learning

Introduction

What is Machine Learning?

Hardcoded

Trained

- Giving computers the **ability to learn** without being explicitly programmed – Arthur Samuel (1959)

What is Machine Learning?

Task

Performance

Experience

- A computer program learns from **experience (E)** with respect to **task (T)** and some **performance measure (P),** if it’s performance on T, as measured by P, improves with their experience E – Tom Mitchell (1998)

What is Machine Learning?

- A branch of **artificial intelligence**, concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data.

- As intelligence requires knowledge, it is necessary for the computers to first **acquire knowledge through learning**.

- Machine learning is programming computers to optimize a **performance criterion** using example data or **past experience**.

Why study Machine Learning?

Engineering reasons:

- Easier to build a learning system than to hand-code a working program!
  - Robot that learns a map of the environment by exploring
  - Programs that learn to play games by playing against themselves
- Improving on existing programs,
  - Instruction scheduling and register allocation in compilers
  - Combinatorial optimization problems
- Solving tasks that require a system to be adaptive
  - Speech and handwriting recognition
  - “Intelligent” user interfaces

Why study Machine Learning?

Scientific reasons:

- Discover knowledge and patterns in high dimensional, complex data
  - Sky surveys
  - High-energy physics data
  - Sequence analysis in bioinformatics
  - Social network analysis
  - Ecosystem analysis
- Learning from animal and human learning
  - How do we learn language?
  - How do we recognize faces?
- Creating real AI!

  "If an expert system–brilliantly designed, engineered and implemented– cannot learn not to repeat its mistakes, it is not as intelligent as a worm or a sea anemone or a kitten." (Oliver Selfridge).

Very brief history

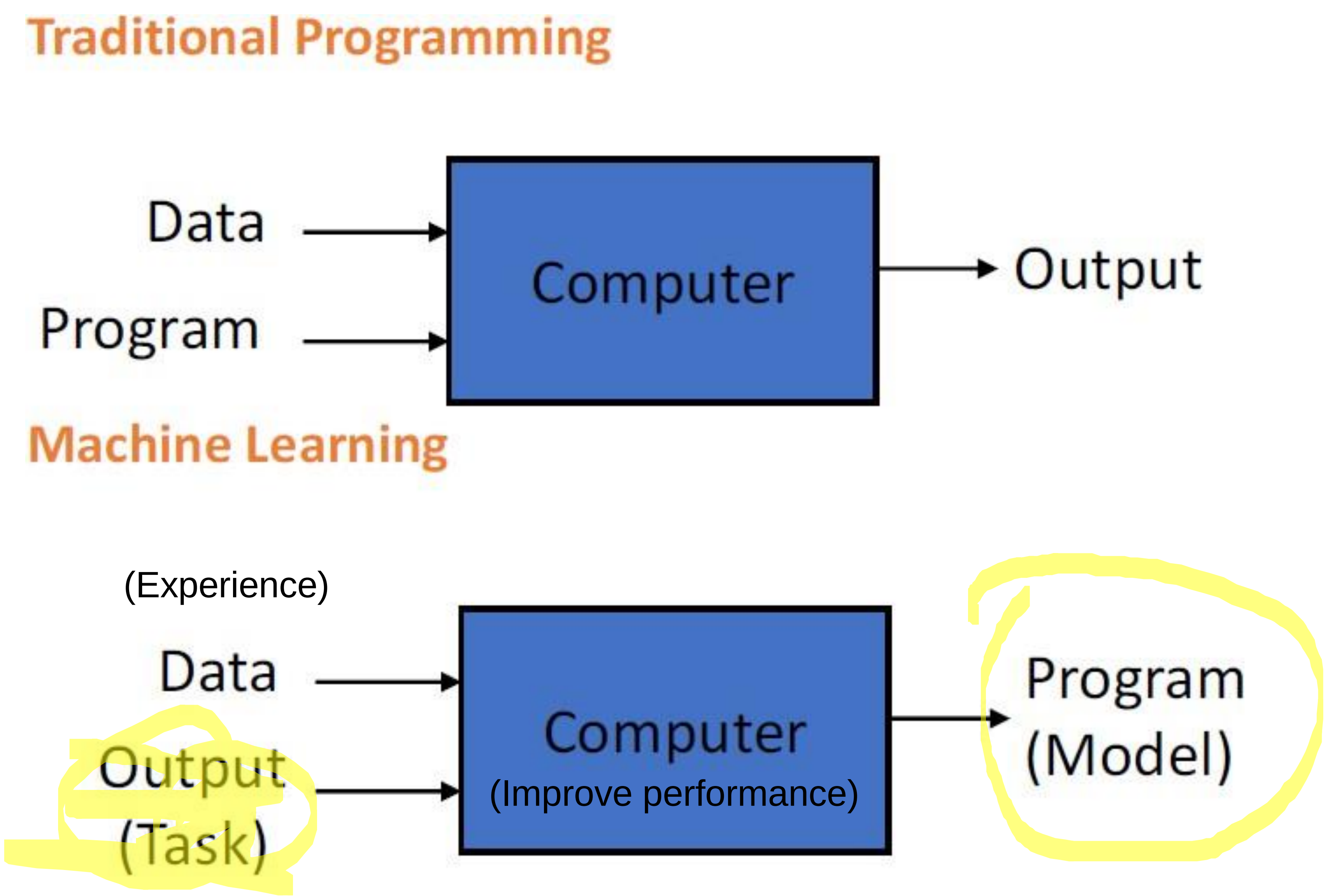
- Studied ever since computers were invented (e.g. Samuel's checkers player)
- Very active in 1960s (neural networks)
- Died down in the 1970s
- Revival in early 1980s (decision trees, backpropagation) - coined as "machine learning"
- Exploded since the 1990s
- Now: very active research field, several yearly conferences (e.g., ICML, NIPS), major journals (e.g., Machine Learning, Journal of Machine Learning Research), rapidly growing number of researchers
- The time is right to study in the field!
    - Lots of recent progress in algorithms and theory
    - Flood of data to be analyzed
    - Computational power is available
    - Growing demand for industrial applications

What are good Machine Learning tasks?

- There is no human expert
  E.g., DNA analysis
- Humans can perform the task but cannot explain how
  E.g., character recognition
- Desired function changes frequently
  E.g., predicting stock prices based on recent trading data
- Each user needs a customized function
  E.g., news filtering

Important application areas

- **Bioinformatics:** sequence alignment, analyzing microarray data, information integration, ...
- **Computer vision:** object recognition, tracking, segmentation, …
- **Robotics:** state estimation, map building, decision making, user movement prediction, personalized robot, etc.
- **Graphics:** building realistic simulations, compressing a picture, etc.
- **Speech:** recognition (cocktail party effect), speaker identification
- **E-commerce:** automated trading agents, data mining, spam filtering, ...
- **Healthcare:** diagnosis, treatment, social assistance, ...
- **Computer games:** building adaptive opponents
- **Education / Working:** learning analytics, hybrid intelligence, …


Traditional Programming
Data
Program
Computer
Output
Machine Learning
(Experience)
Data
Output
(Task)
Computer
(Improve performance)
Program
(Model)

Machine Learning model

- We have a model which we can use to predict new data
- E.g. Image classification

Input

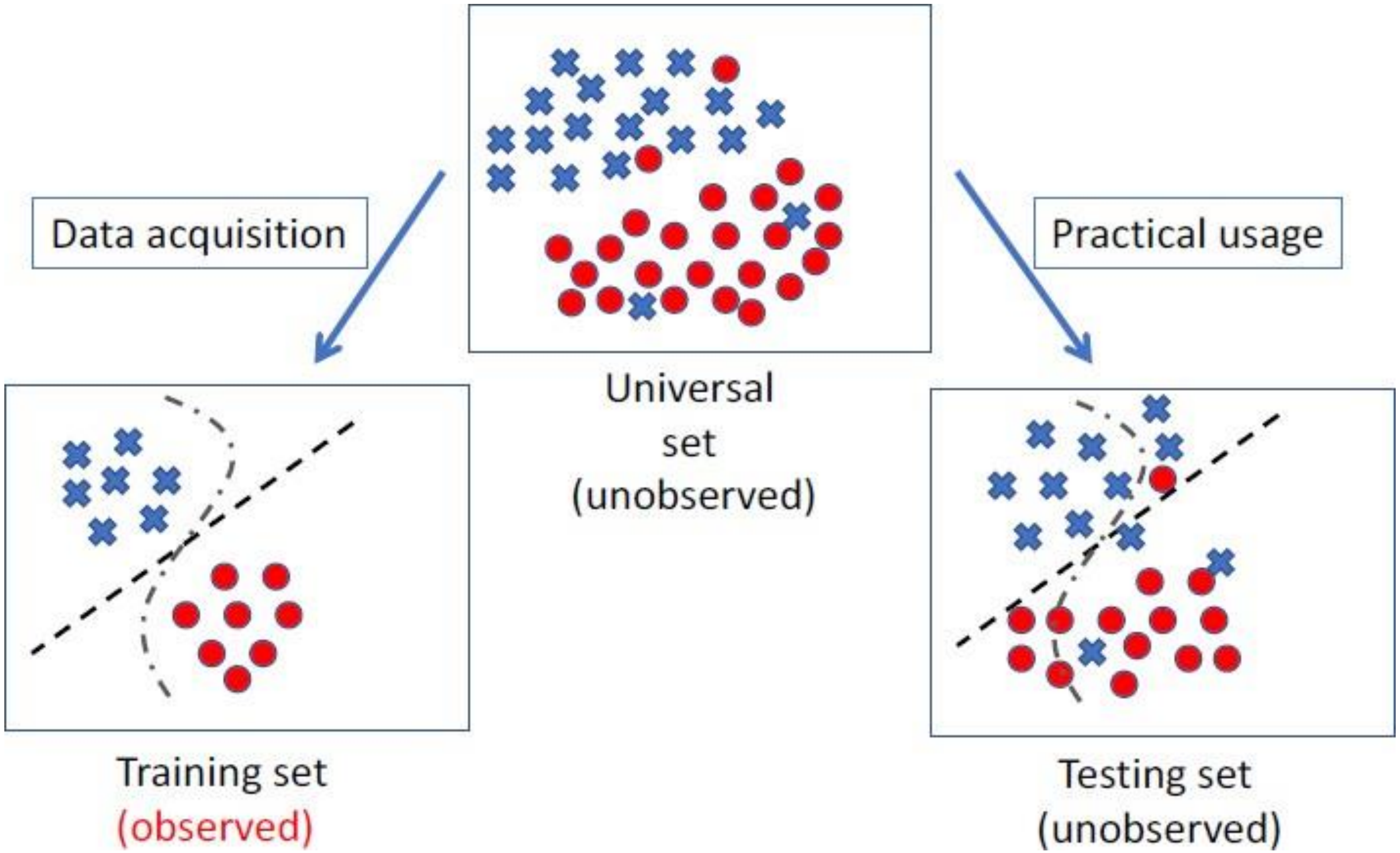
model →

Dog
Building
Cat ✓
Human
Car

Machine Learning model

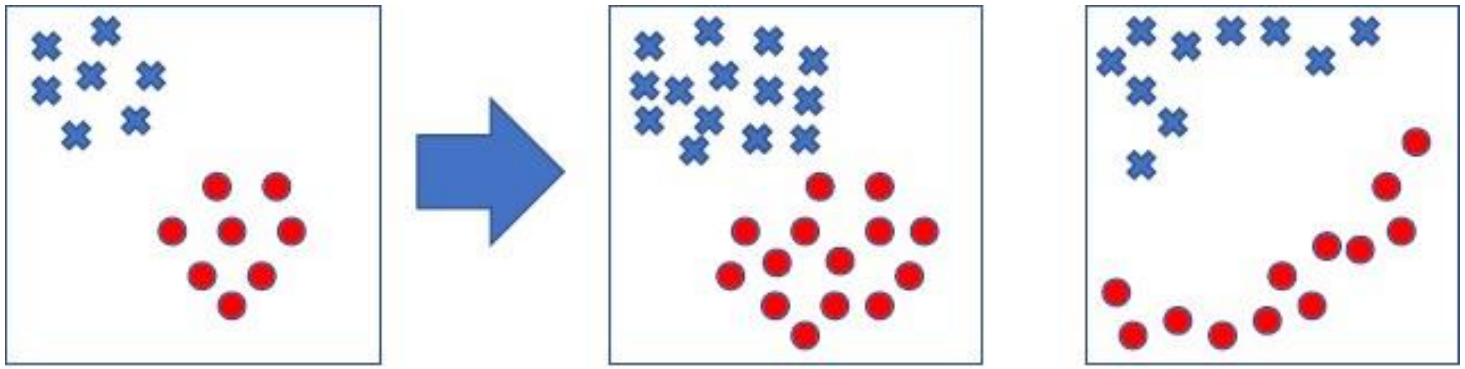
- Generalizability+Robustness: Learning general models from a data set.
- Data is cheap (?) and abundant (data warehouses, data mars); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behaviour:
  - *People who bought "Da Vinci Code" also bought "The Five People You Meet in Heaven" (www.amazon.com)*
- Build a model that is *a good and useful approximation* to the data.

Training and Testing



Training and Testing

- Training is the process of making system able to learn.
- No free lunch rule:
  - Training set and testing set come from the same distribution
  - Need to make some assumption or bias



Training data

Testing data

Performance

- There are several factors affecting the performance:
  - **Quality of training data** provided
  - The form and extent of any initial **background knowledge**
  - The **type of feedback** provided
  - The **learning algorithm** used

- Two important factors:
  - Modelling
  - Optimization

Algorithms

- The success of machine learning system also depends on the algorithms.

- The learning algorithms should **extract** useful information from training examples (e.g., find patterns from your training samples).

- The algorithms control the **search** to find and build the knowledge structures (e.g., find the optimized parameters of different variables).

Types of Machine Learning

Based on information available

- **Supervised learning ($\{x_n \in R^d, y_n \in R\}_{n=1}^N$)**
  - Classification (discrete labels)
  - Regression (real values)
- **Unsupervised learning ($\{x_n \in R^d\}_{n=1}^N$)**
  - Clustering
  - Probability distribution estimation
  - Finding association (in features)
  - Dimension reduction
- **Semi-supervised learning**
- **Reinforcement learning**
  - Decision making (robotics, board games)

Based on learner's role

- **Passive learning**
- **Active learning**

Types of Machine Learning

Based on learner's role

- Traditionally, learning algorithms have been **passive learners**, which take a given batch of data and process it to produce a hypothesis or model.

  Data → Learner → Model

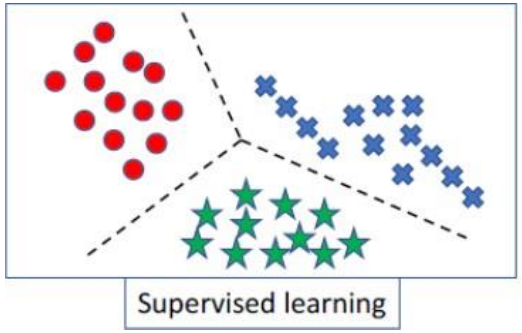
- **Active learners** are instead allowed to query the environment

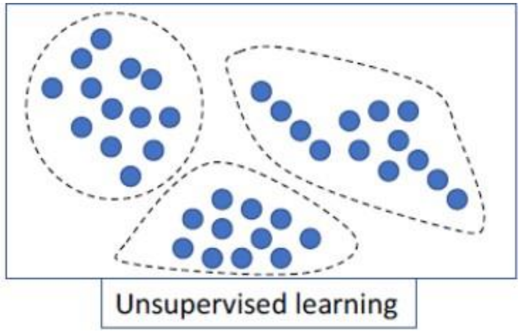
– Make queries

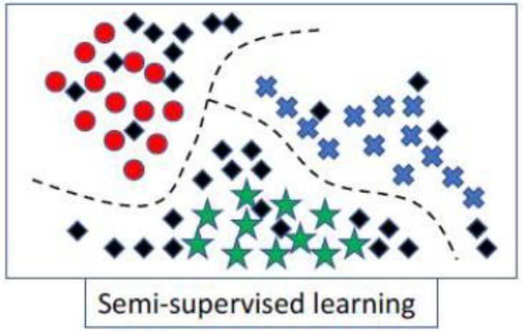
– Perform experiments

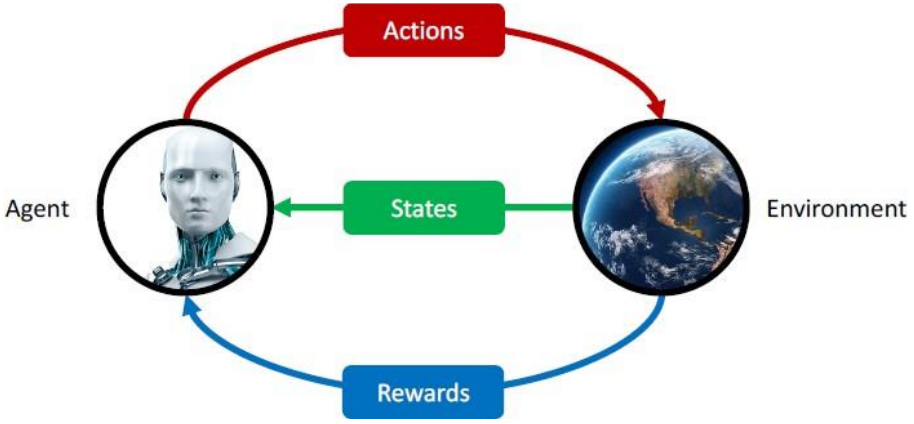
Types of Machine Learning

Based on information available









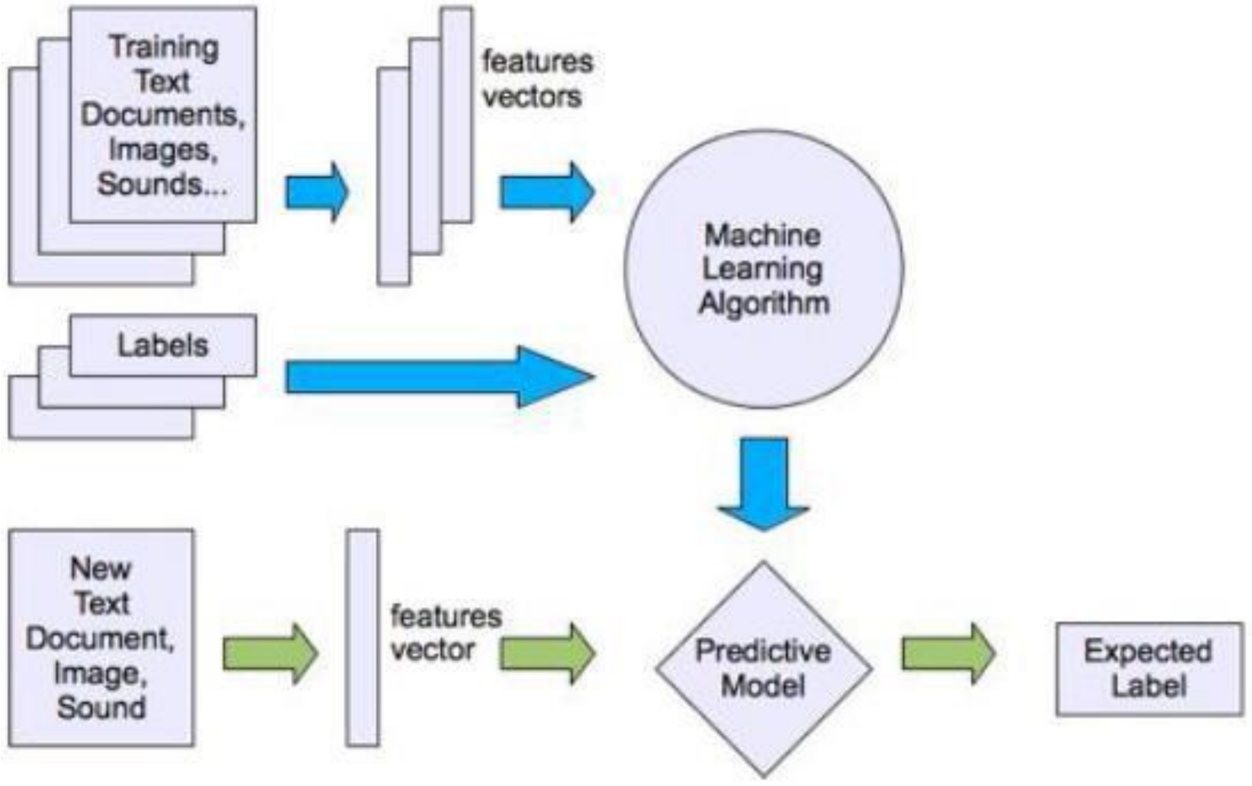
Reinforcement learning

Types of Machine Learning

- **Supervised Learning**

Types of Machine Learning

- **Supervised Learning**

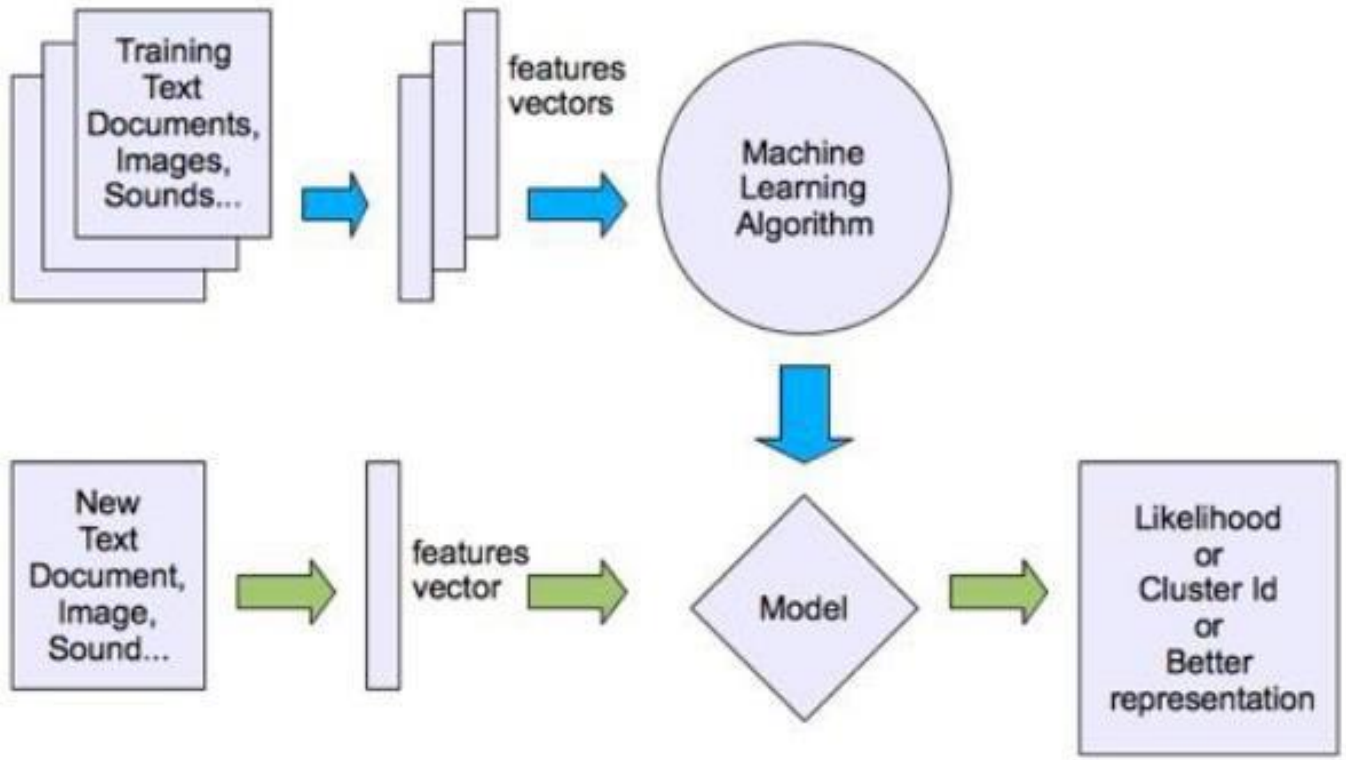


Types of Machine Learning

- **Unsupervised Learning**

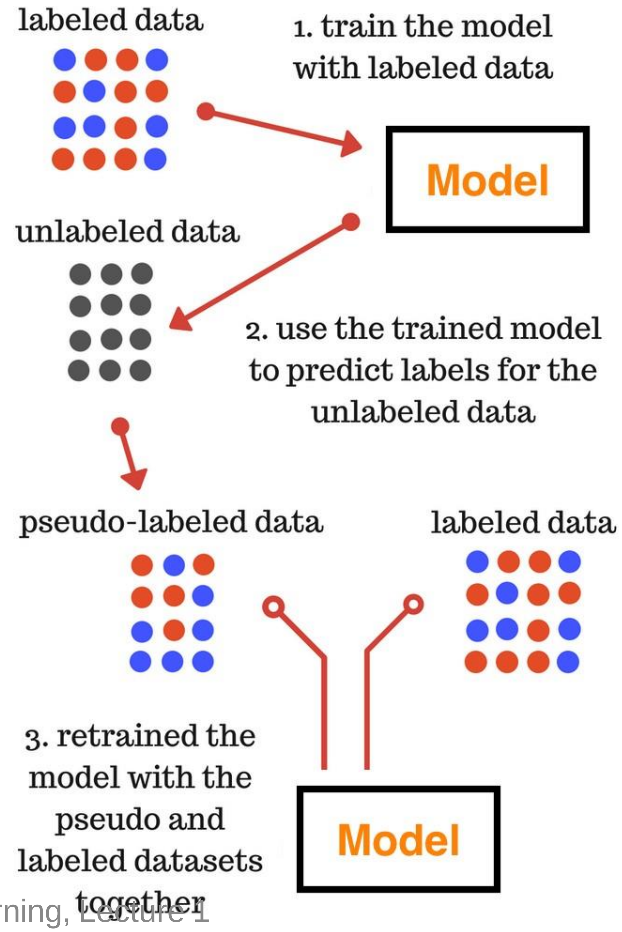
Types of Machine Learning

- **Unsupervised Learning**



Types of Machine Learning

- **Semi-supervised Learning**

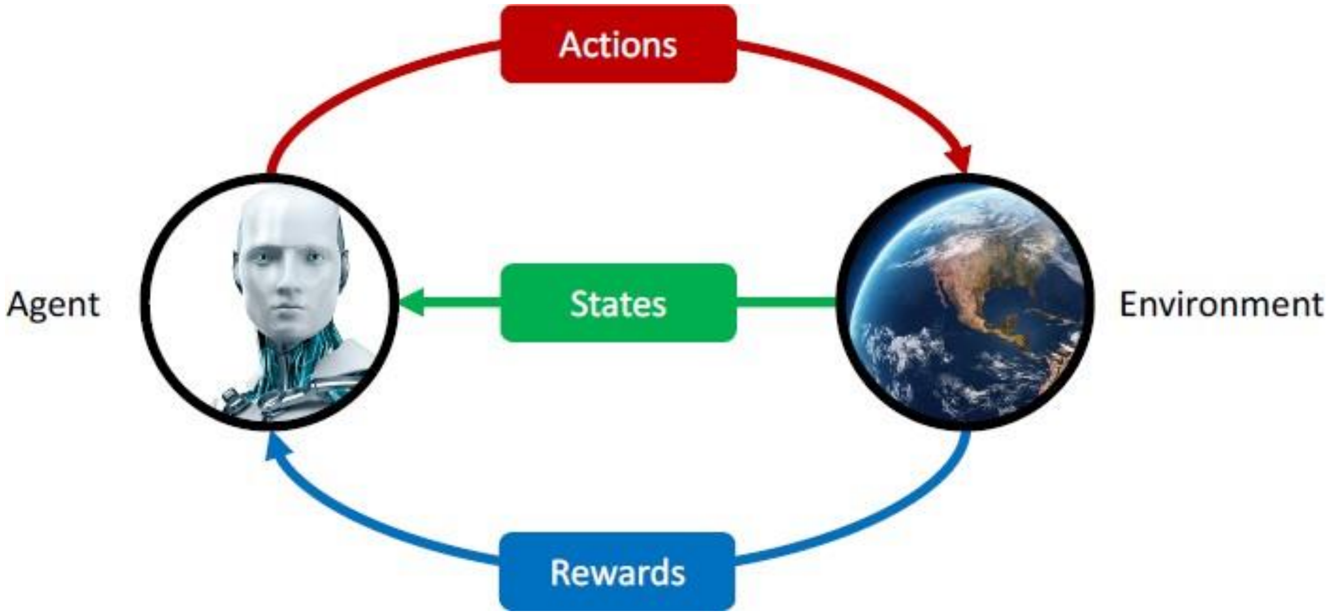


Types of Machine Learning

- **Reinforcement Learning:** Rewards from a sequence of actions

Types of Machine Learning

- **Reinforcement Learning**



The state space can be discrete or continuous. In case of continuous states, you would use a function approximator to represent your state.