

Python Machine Learning

Lecture 1: Giving Computers the Ability to Learn from Data

Chongshou Li

lics@swjtu.edu.cn

SWJTU-LEEDS Joint School

SOUTHWEST JIAOTONG UNIVERSITY

自我介绍



10年海（境）外经历
学习（5年）+工作（5年）

入选2022年四川省
“天府峨眉计划”
青年科技人才项目

省级特聘专家

香港城市大学

(2022 QS全球排名53)

管理科学系

研究助理

南洋理工大学

(2022 QS全球排名12)

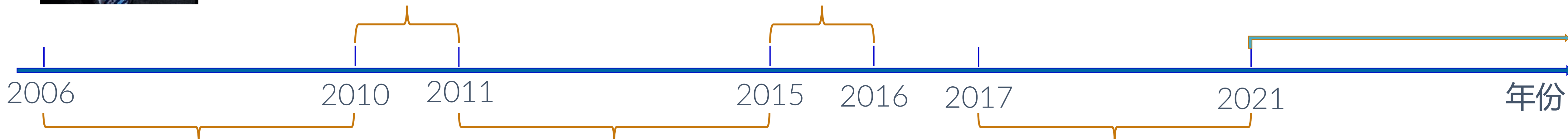
数学系

助理研究员

西南交通大学

计算机与人工智能学院

副教授



西北工业大学

(双一流A类、985)

计算机学院

计算机科学与技术专业

(国家重点学科、学科评估A+)

学士

香港城市大学

(2022 QS全球排名53)

商学院

管理科学专业

博士

新加坡国立大学

(2022 QS全球排名11)

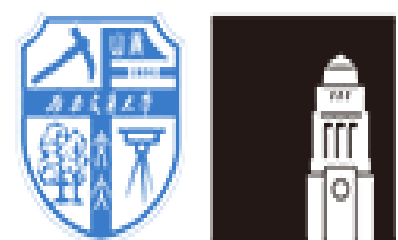
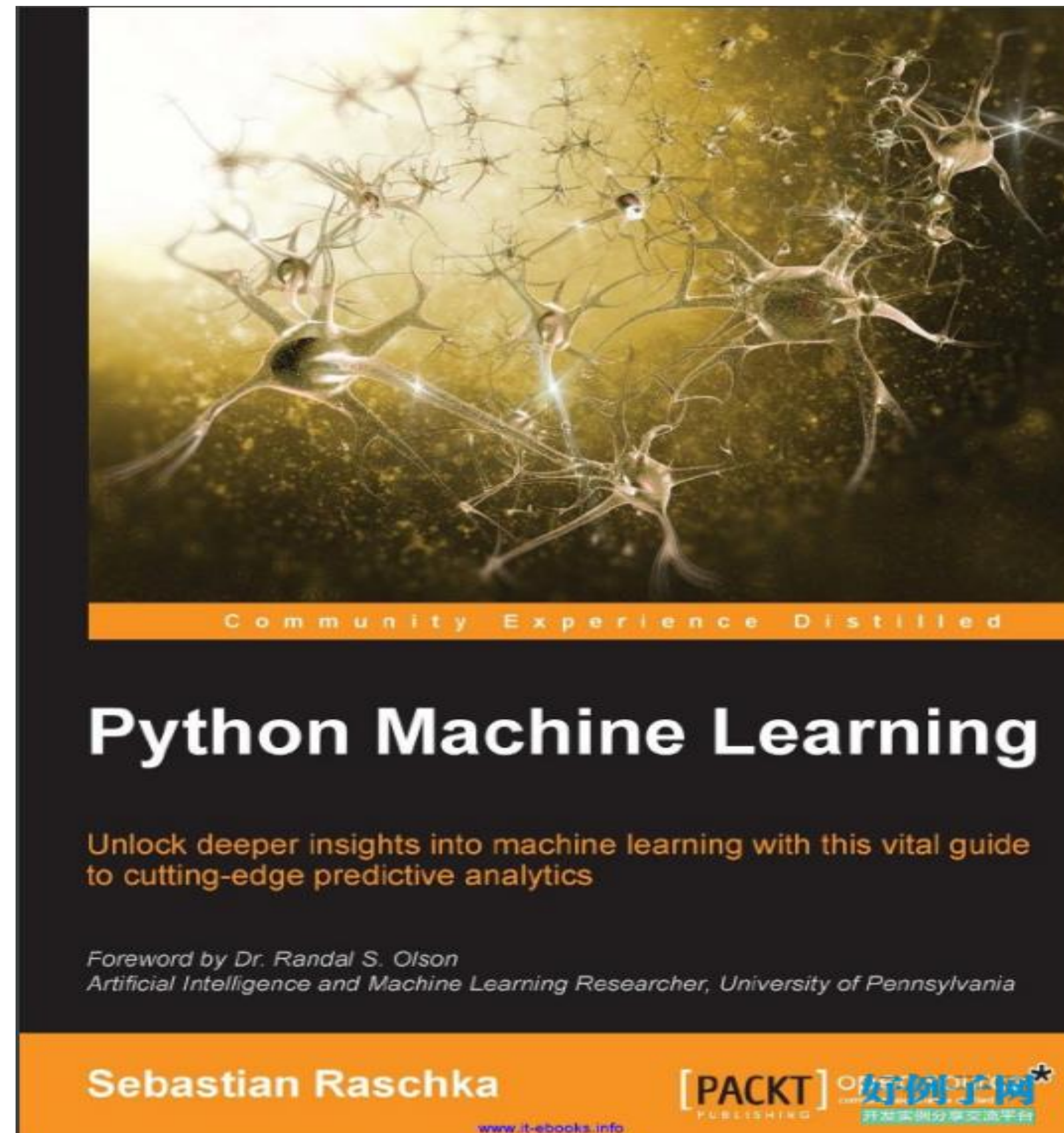
工程学院

工业系统工程与管理系

助理教授（研究型）



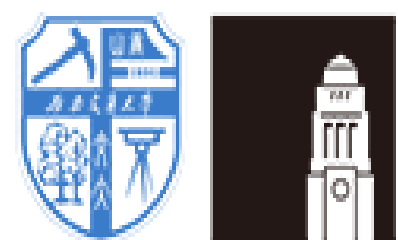
Textbook



Intended Learning Outcomes (OILs)

On the completion of this lecture, students are supposed to be able to:

- The three types of learning and basic terminology
 - supervised learning
 - reinforcement learning
 - unsupervised learning
- An introduction to the basic terminology and notations
- A roadmap for building machine learning systems
 - Preprocessing – getting data into shape
 - Training and selecting a predictive model
 - Evaluating models and predicting unseen data instances
- Installing and setting up Python for data analysis and machine learning

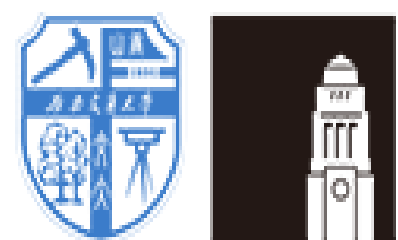


Today's Topics

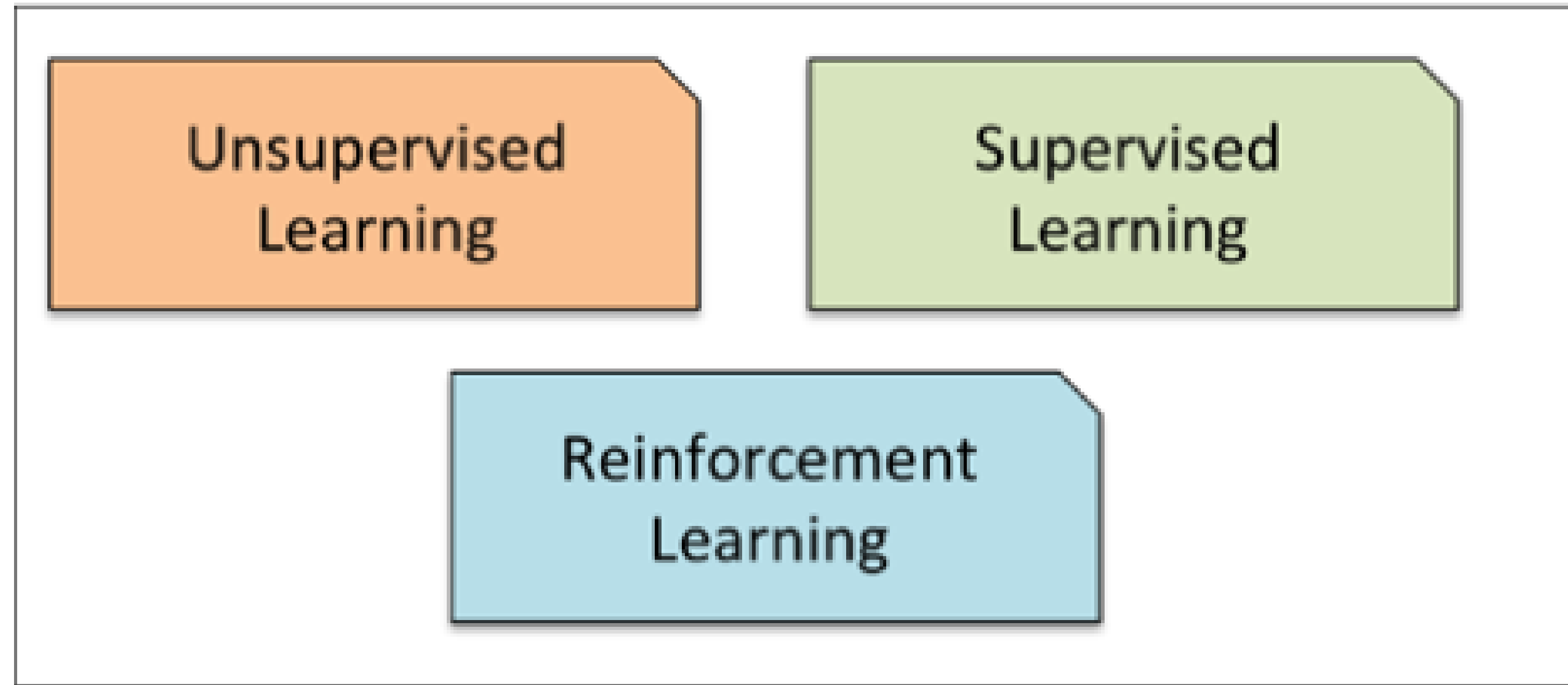
- The three different types of machine learning
- An introduction to the basic terminology and notations
- A roadmap for building machine learning systems
- Using Python for machine learning
- Summary

Today's Topics

- **The three different types of machine learning**
 - supervised learning
 - reinforcement learning
 - unsupervised learning
- An introduction to the basic terminology and notations
- A roadmap for building machine learning systems
- Using Python for machine learning
- Summary



The three different types of machine learning

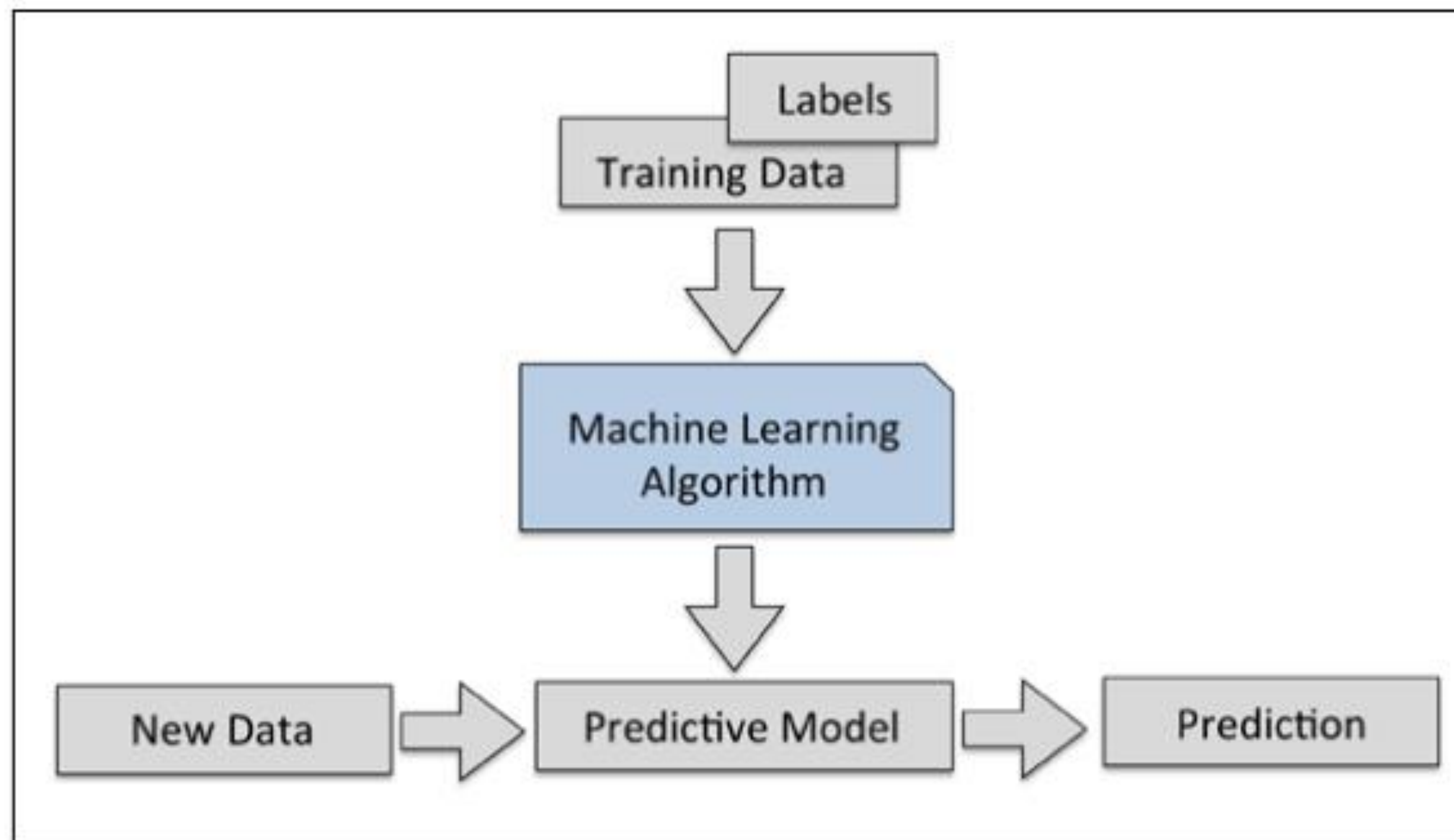


The three different types of machine learning

supervised learning

The main goal in supervised learning is to learn a model from labeled training data that allows us to make predictions about unseen or future data.

supervised refers to a set of samples where the desired output signals (labels) are already known.



- Classification for predicting class labels
- Regression for predicting continuous outcomes

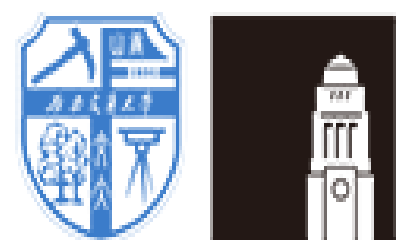
Classification for predicting class labels

Classification is a subcategory of supervised learning where the goal is to predict the categorical class labels of new instances based on past observations. **Those class labels are discrete, unordered values.**

The example of **e-mail-spam detection represents** a typical example of a binary classification task.

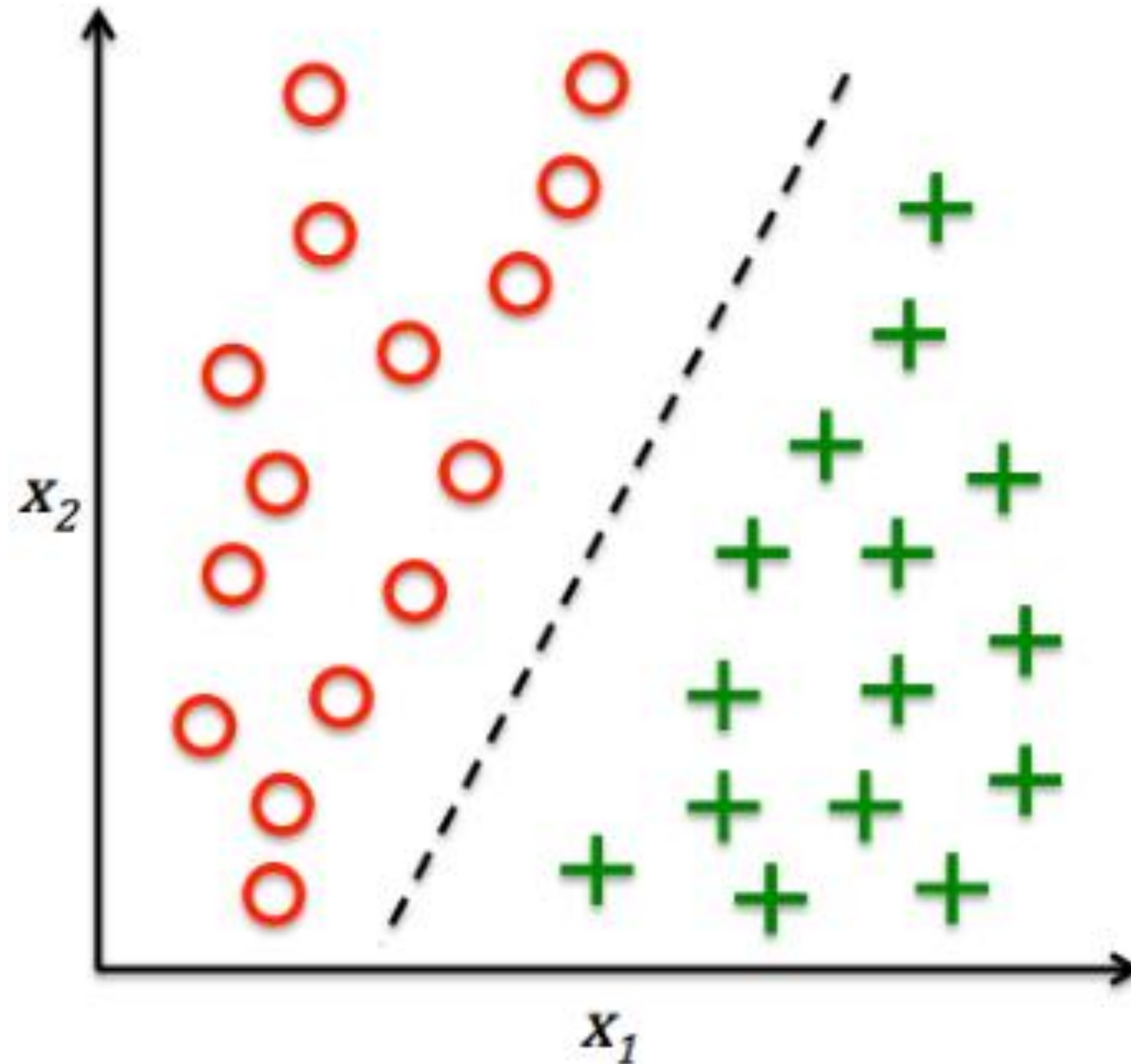
However, the set of class labels does not have to be of a binary nature. The predictive model learned by a supervised learning algorithm can assign any class label that was presented in the training dataset to a new, unlabeled instance.

A typical example of a multi-class classification task is **handwritten character recognition.**



Classification for predicting class labels

Classification:
use a supervised
machine learning
algorithm to learn a rule:
the decision boundary
represented as a black
dashed line—that can
separate those two
classes and classify new
data into each of those
two categories given its
 x_1 and x_2 values:



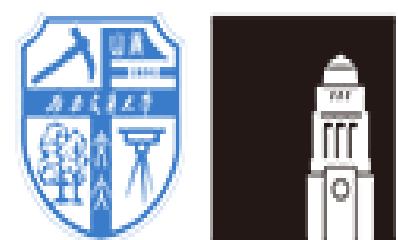
a binary classification task given 30 training samples

Regression for predicting continuous outcomes

In regression analysis, we are given a number of predictor (explanatory) variables and a continuous response variable (outcome), and we try to find a relationship between those variables that allows us to predict an outcome.

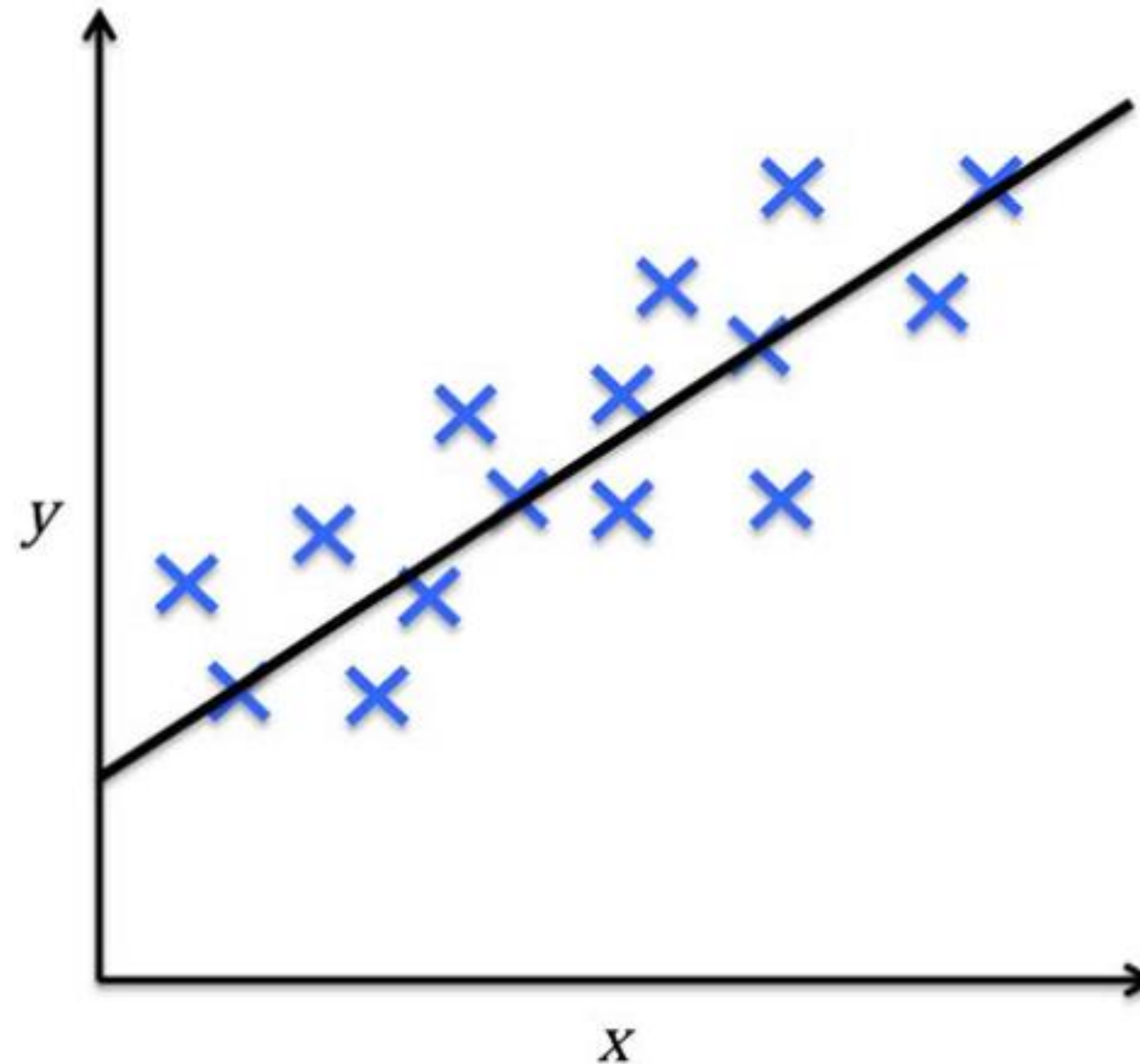
Those class labels are discrete, unordered values

For example, let's assume that we are interested in predicting the Math SAT scores of our students.



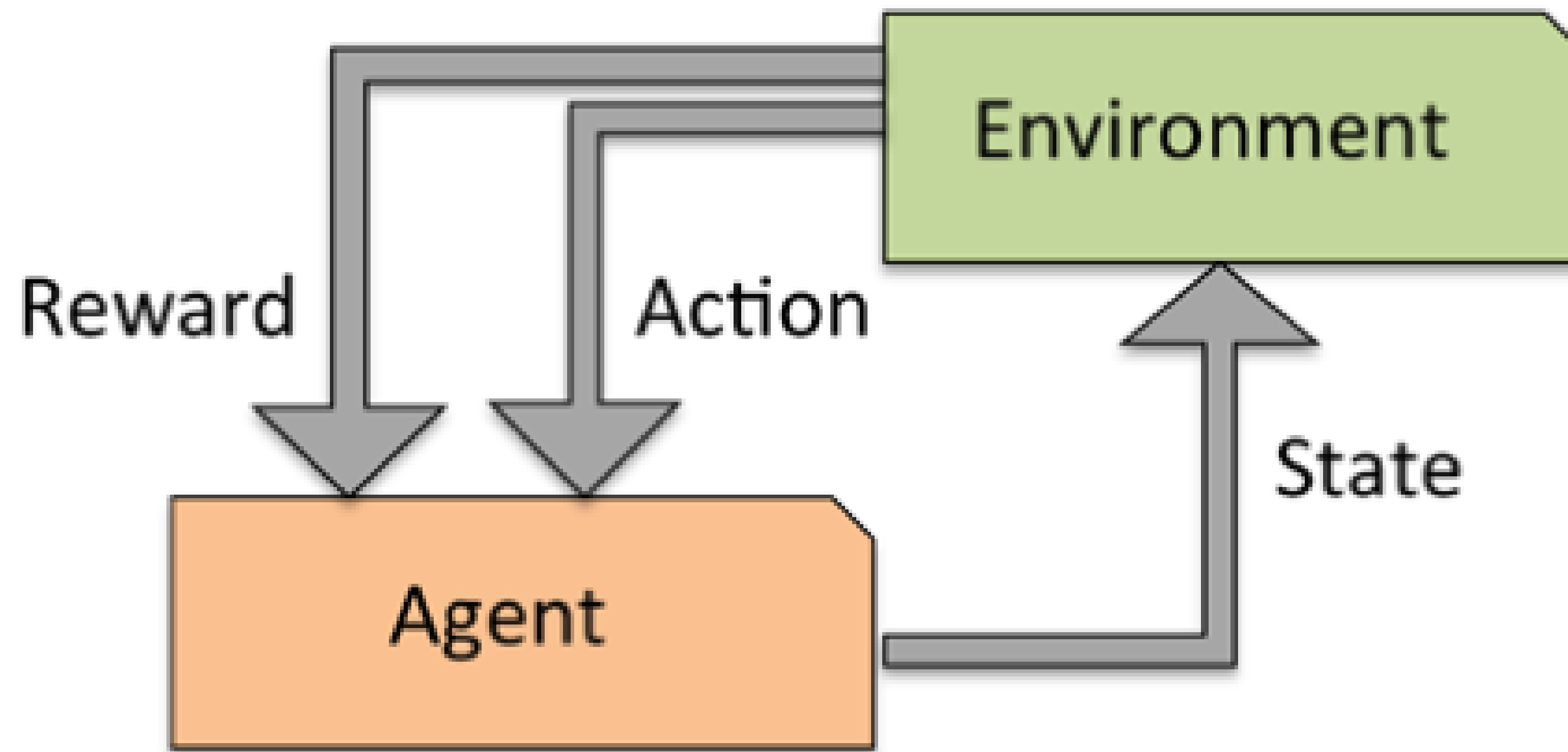
supervised learning

Regression:
Given a predictor variable x and a response variable y , we fit a straight line to this data that minimizes the distance



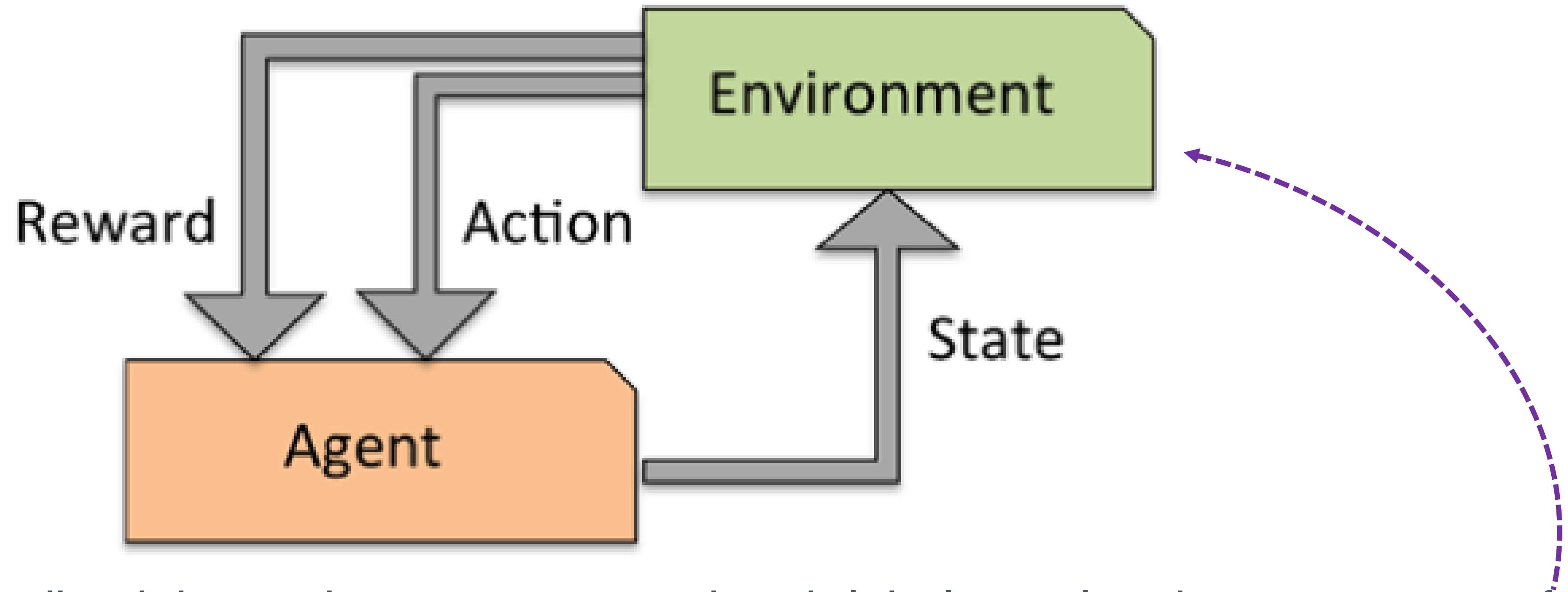
linear regression

reinforcement learning



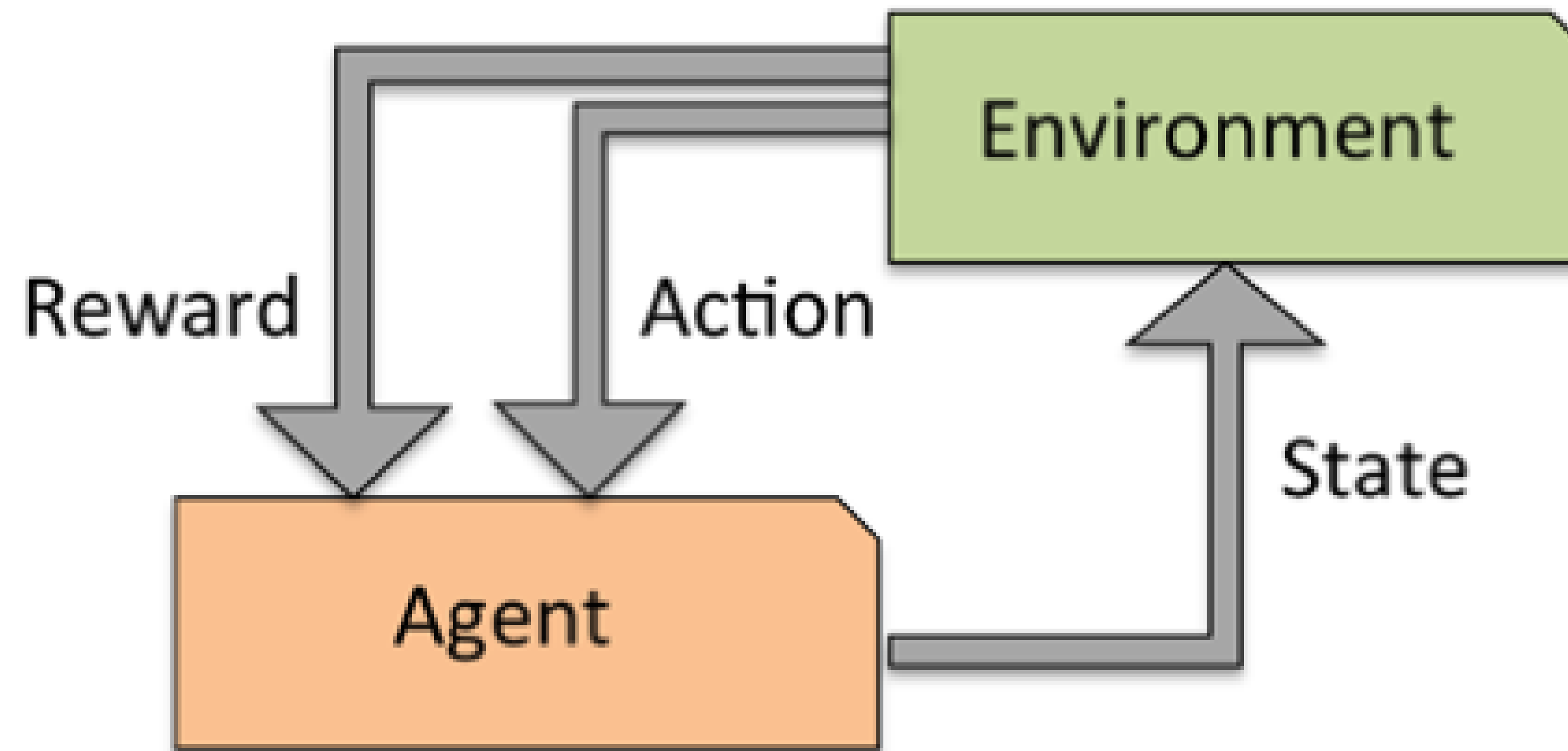
In reinforcement learning, the goal is to develop a system (**agent**) that improves its performance based on interactions with the environment. Since the information about the current state of the environment typically also includes a so-called reward signal, we can think of reinforcement learning as a field related to supervised learning.

reinforcement learning



However, this feedback is not the correct ground truth label or value, but a measure of how well the **action** was measured by a **reward** function. Through the interaction with the environment, an agent can then use reinforcement learning to learn a series of actions that **maximizes this reward** via an exploratory trial-and-error approach or deliberative planning.

reinforcement learning



chess engine :

the agent decides upon a series of moves depending on the state of the board (the environment), and the reward can be defined as win or lose at the end of the game:

unsupervised learning

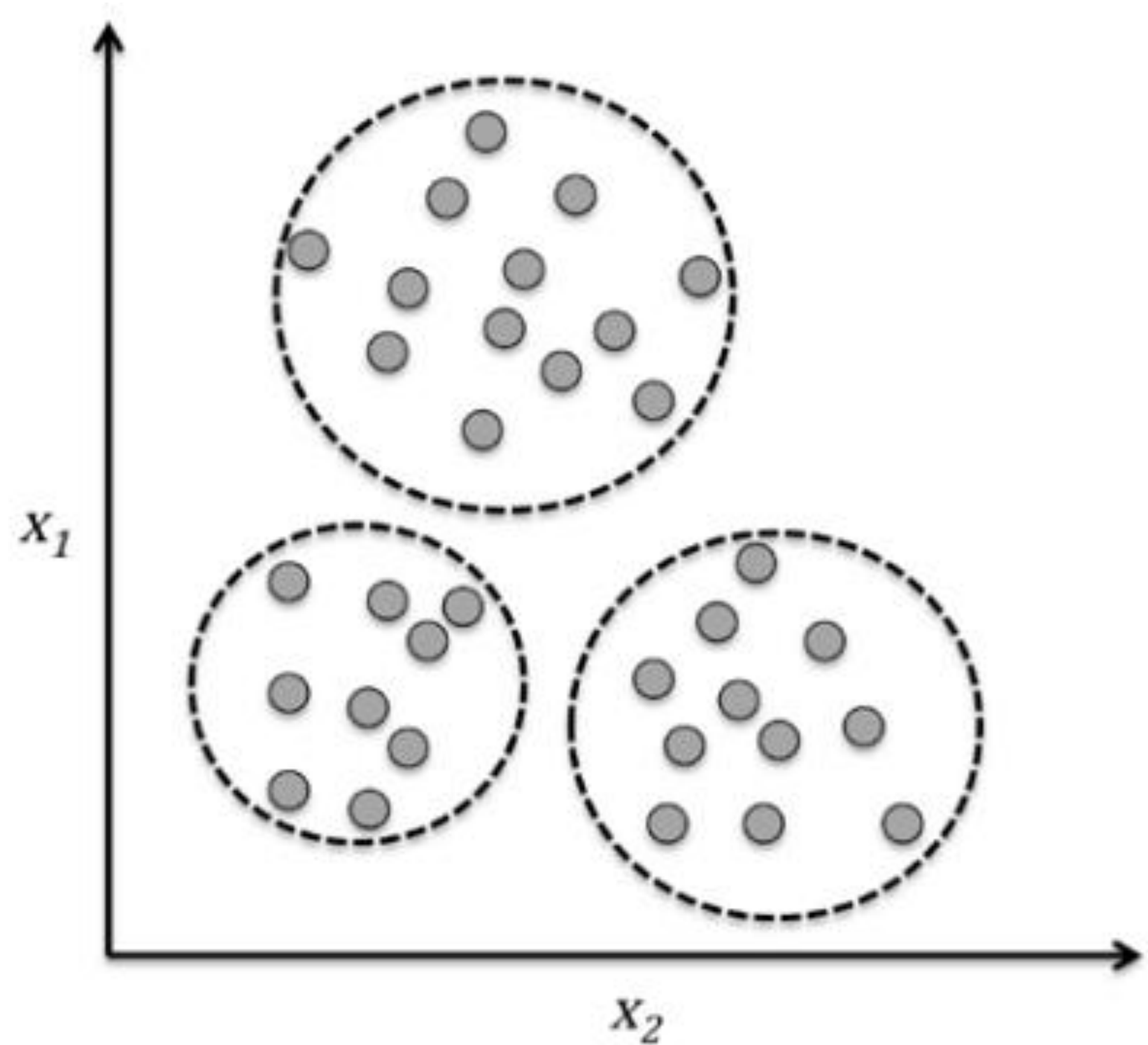
In unsupervised learning, dealing with unlabeled data or data of unknown structure.

The two subfield of unsupervised learning:

- Finding subgroups with clustering
- Dimensionality reduction for data compression

unsupervised learning

- Finding subgroups with clustering



an exploratory data analysis technique:

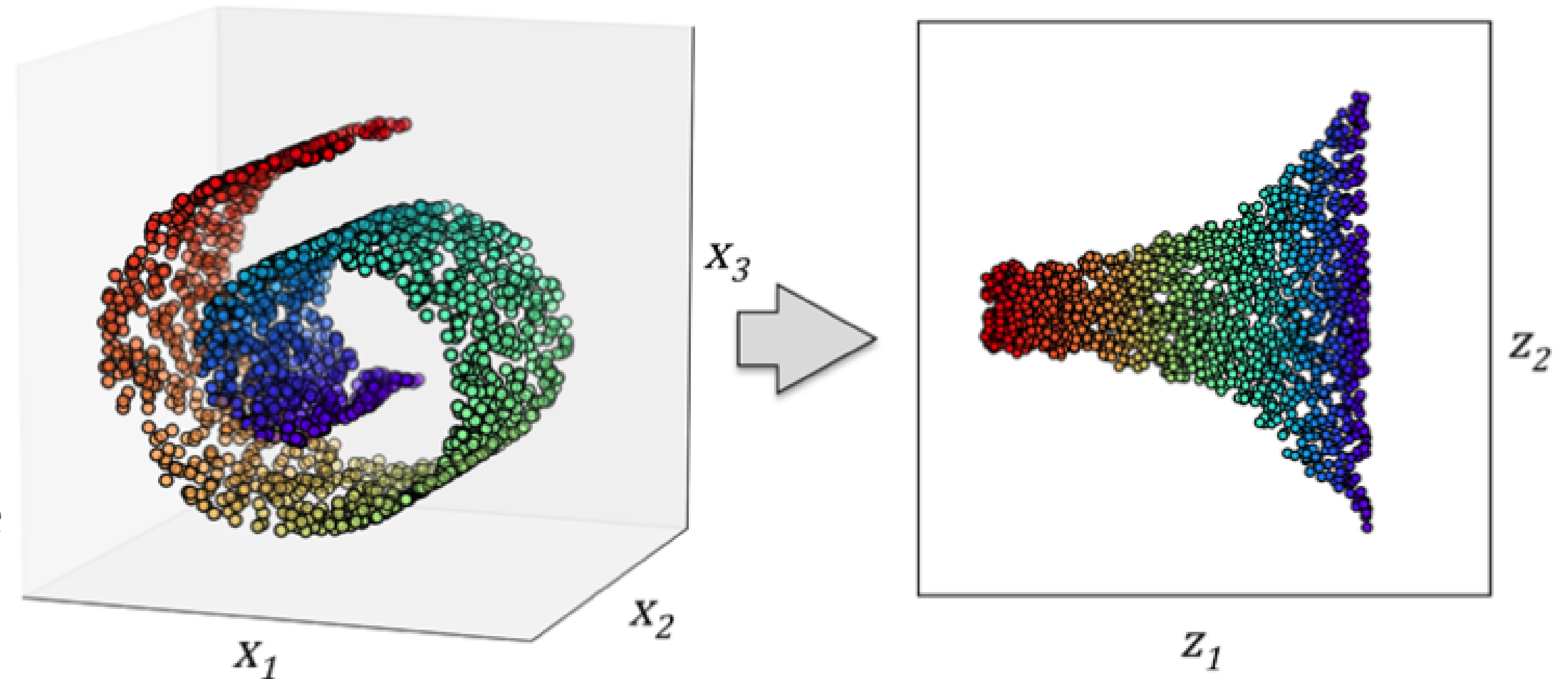
organize a pile of information into meaningful subgroups (clusters) without having any prior knowledge of their group memberships

Eg: allows marketers to discover customer groups based on their interests in order to develop distinct marketing programs.

unsupervised learning

- Dimensionality reduction for data compression

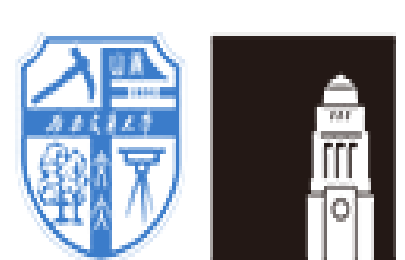
Unsupervised dimensionality reduction is a commonly used approach in feature preprocessing to remove noise from data, which can also degrade the predictive performance of certain algorithms, and compress the data onto a smaller dimensional subspace while retaining most of the relevant information.



Eg: compress a 3D Swiss Roll onto a new 2D feature subspace

Today's Topics

- The three different types of machine learning
- An introduction to the basic terminology and notations
- A roadmap for building machine learning systems
- Using Python for machine learning
- Summary



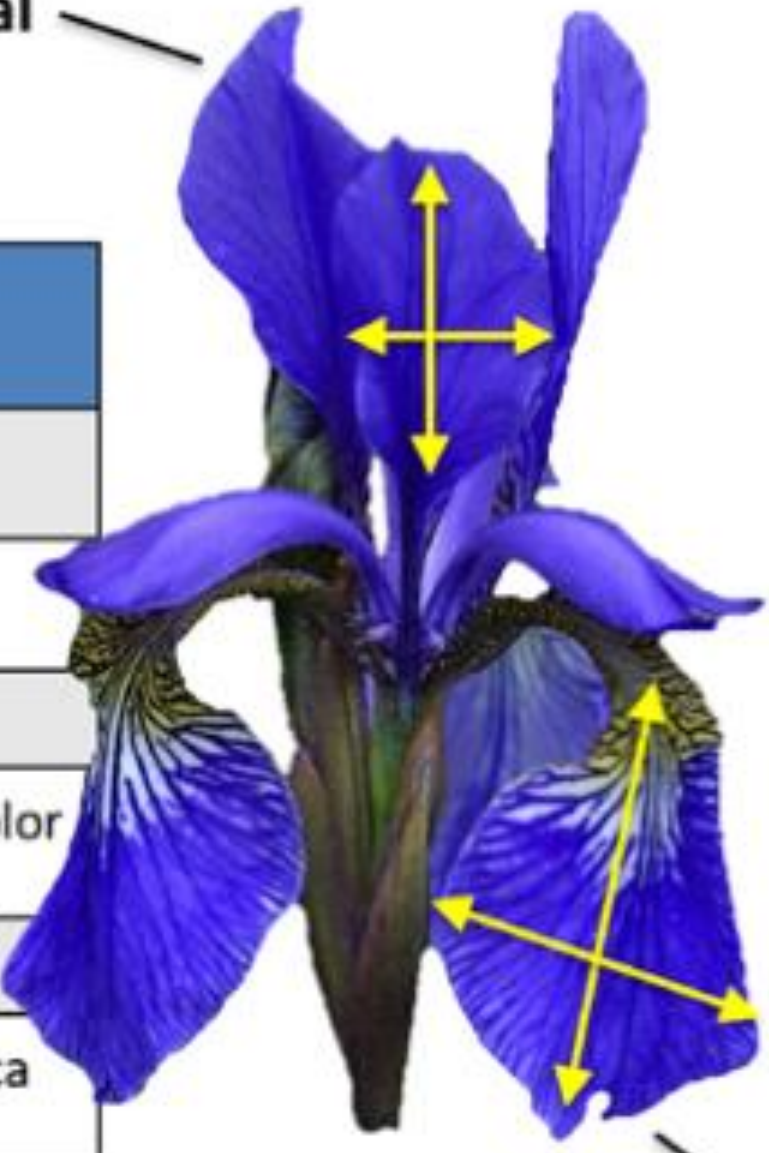
basic terminology and notations

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes, measurements, dimensions)

Class labels
(targets)



The Iris dataset contains the measurements of 150 iris flowers from three different species: **Setosa, Versicolor, and Virginica**.

Each flower sample represents one row in our data set, and the flower measurements in centimeters are stored as columns, which we also call **the features** of the dataset.

represent each sample as separate row in a feature matrix X , where each feature is stored as a separate column.

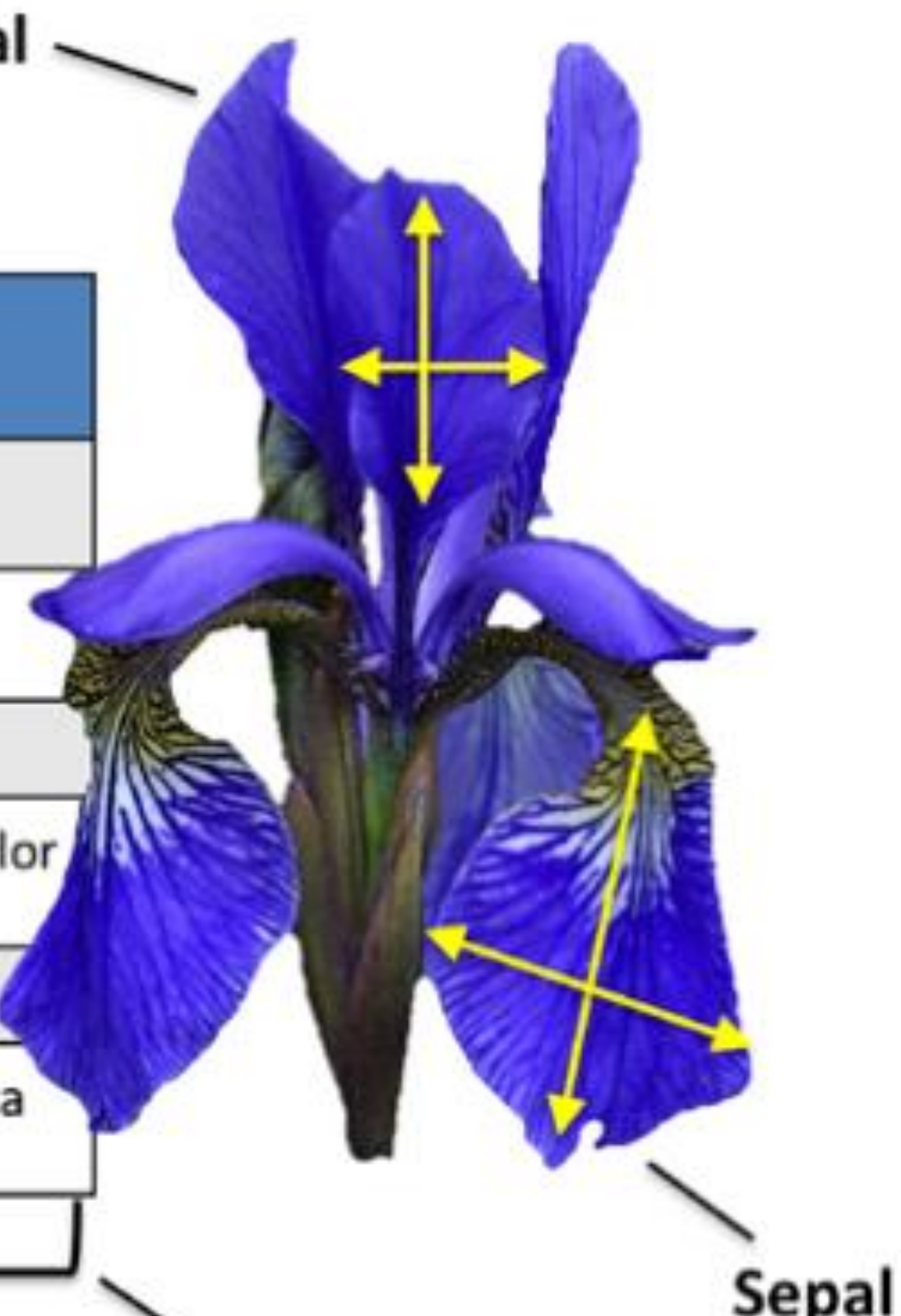
basic terminology and notations

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes, measurements, dimensions)

Class labels
(targets)



Petal

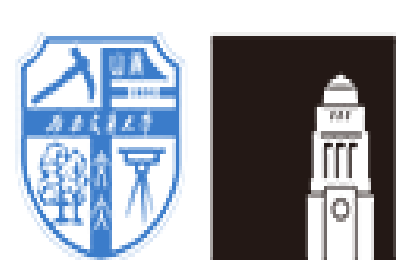
Sepal

The Iris dataset, consisting of 150 samples and 4 features
150 x 4 matrix :

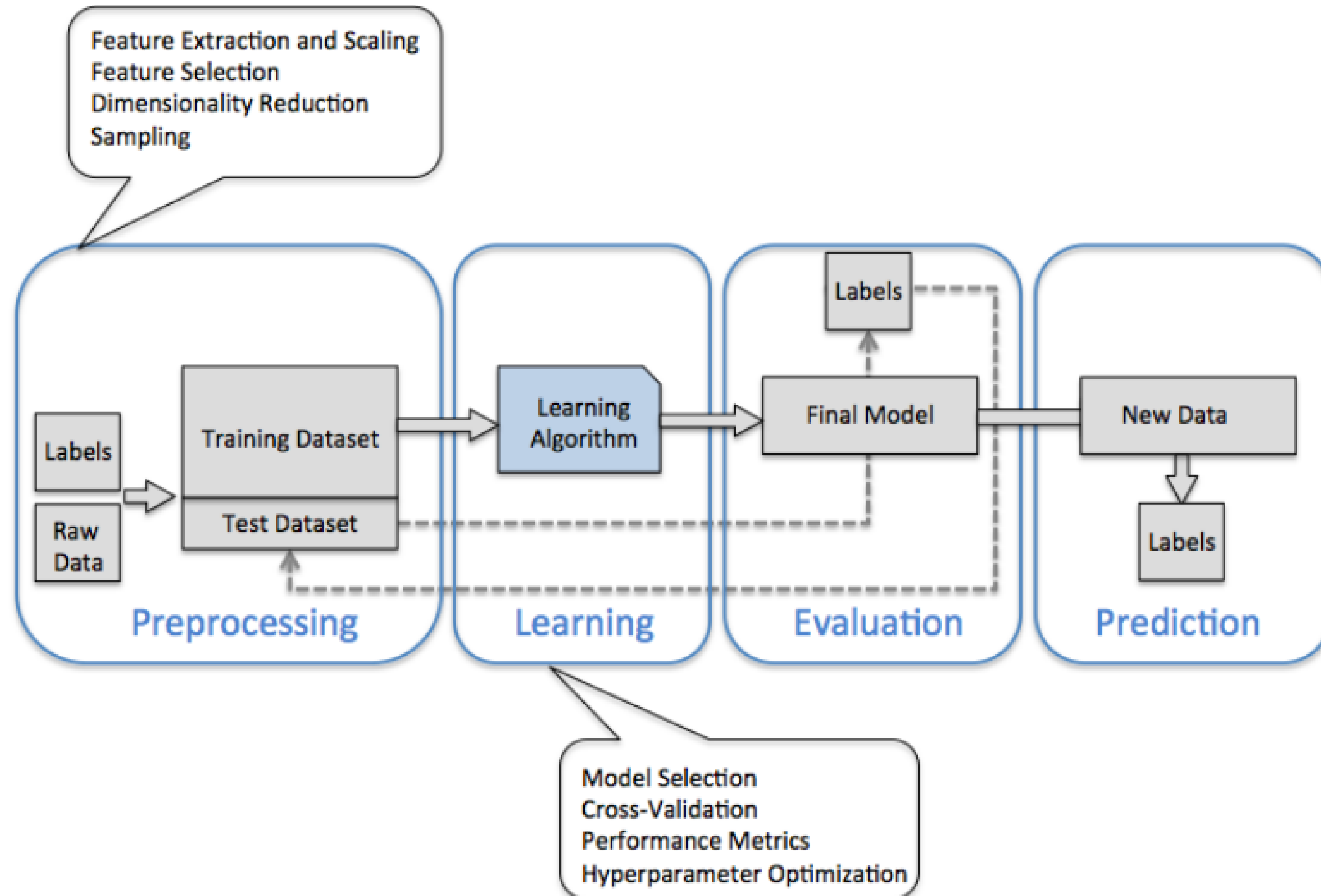
$$\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & x_4^{(150)} \end{bmatrix}$$

Today's Topics

- The three different types of machine learning
- An introduction to the basic terminology and notations
- **A roadmap for building machine learning systems**
 - Preprocessing – getting data into shape
 - Training and selecting a predictive model
 - Evaluating models and predicting unseen data instances
- Using Python for machine learning
- Summary



A roadmap for building machine learning systems



A roadmap for building machine learning systems

step1

- Preprocessing – getting data into shape
 - Raw data rarely comes in the form and shape that is necessary for the optimal performance of a learning algorithm. Thus, the preprocessing of the data is one of the most crucial steps in any machine learning application.

A roadmap for building machine learning systems

step2

- Training and selecting a predictive model
 - each classification algorithm has its inherent biases, and no single classification model enjoys superiority if we don't make any assumptions about the task. In practice, it is therefore essential to compare at least a handful of different algorithms in order to train and select the best performing model.

A roadmap for building machine learning systems

step3

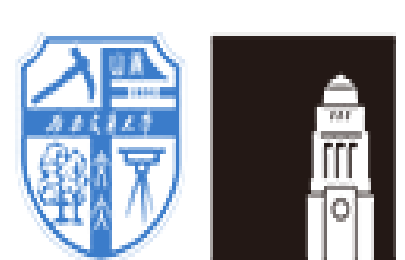
- Evaluating models
 - use the test dataset to estimate how well it performs on this unseen data to estimate the generalization error.

step4

- predicting unseen data instances

Today's Topics

- The three different types of machine learning
- An introduction to the basic terminology and notations
- A roadmap for building machine learning systems
- Using Python for machine learning
- Summary



Using Python for machine learning

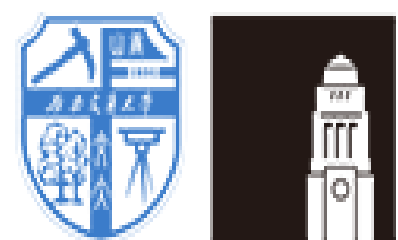
- Installing anaconda

- the official Python website: <https://www.python.org>.
- More information about pip can be found at <https://docs.python.org/3/installing/index.html>
- The Anaconda installer can be downloaded at <http://continuum.io/downloads#py34>

- Installing Python packages

- install additional Python packages: **conda install SomePackage**
- Already installed packages can be updated via the --upgrade flag:

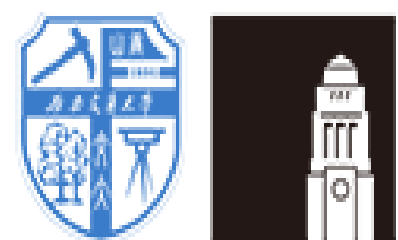
Conda update SomePackage



Using Python for machine learning

- NumPy 1.9.1
- SciPy 0.14.0
- scikit-learn 0.15.2
- matplotlib 1.4.0
- pandas 0.15.2

Make sure that the version numbers of your installed packages are **equal to, or greater than**, those version numbers to ensure the code examples run correctly:



Today's Topics

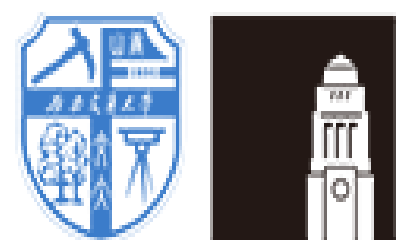
- The three different types of machine learning
- An introduction to the basic terminology and notations
- A roadmap for building machine learning systems
- Using Python for machine learning

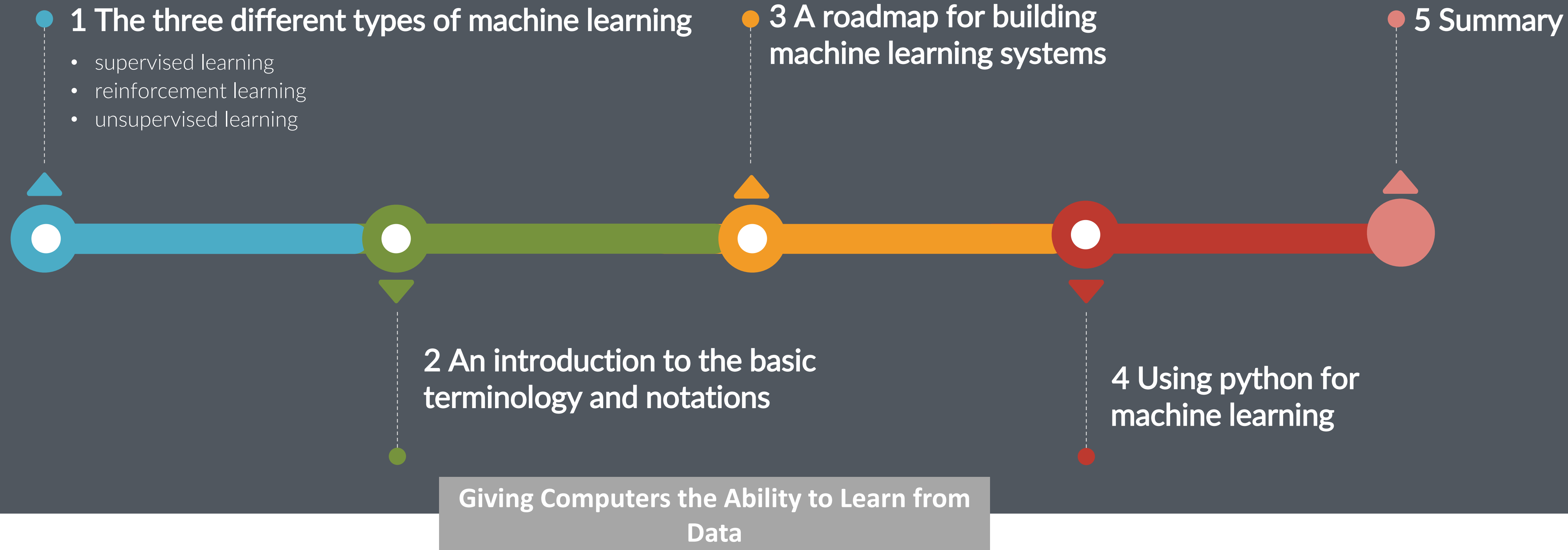
- **Summary**



Summary

- The three different types of machine learning
 - supervised learning
 - reinforcement learning
 - unsupervised learning
- An introduction to the basic terminology and notations
- A roadmap for building machine learning systems
- Using Python for machine learning
 - NumPy 1.9.1
 - SciPy 0.14.0
 - scikit-learn 0.15.2
 - matplotlib 1.4.0
 - pandas 0.15.2





“Thank You”

Chongshou LI
Southwest Jiaotong University