安徽理工大学
ANHUI UNIVERSITY OF SCIENCE & TECHNOLOGY

• 人工智能专业 学科基础教育必修模块

Python与机器学习

Python and Machine Learning

Chapter 6：Machine Learning Overview

• Lecturer ： 孟亦凡

• E-mail：myf@aust.edu.cn

Chapter 6: Contents

6.1 The Concept of Machine Learning

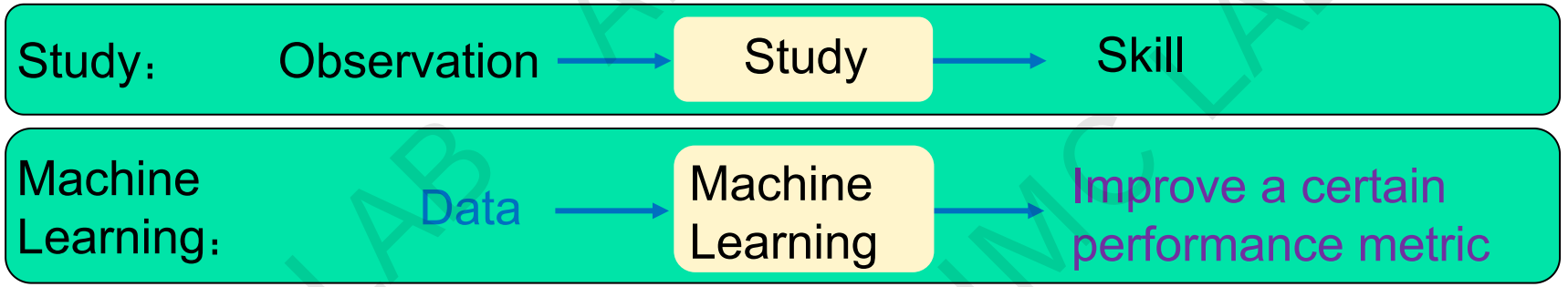
6.1.1 The Concept of Machine Learning

- Machine Learning is a science of artificial intelligence. The main research object in this field is artificial intelligence, especially how to improve the performance of specific algorithms through experiential learning. (Langley, 1996)
- Machine Learning is the study of computer algorithms that can improve automatically through experience. (Tom Mitchell, 1997)
- Machine Learning is programming a computer using sample data or past experience to optimize performance metrics. (Alpaydin, 2004)

6.1 The Concept of Machine Learning

6.1.1 The Concept（概念） of Machine Learning

Machine Learning is dedicated to studying how to use computational means and experience to improve the system's own performance, thereby generating a "model" from data on the computer, used to "make judgments on new situations".



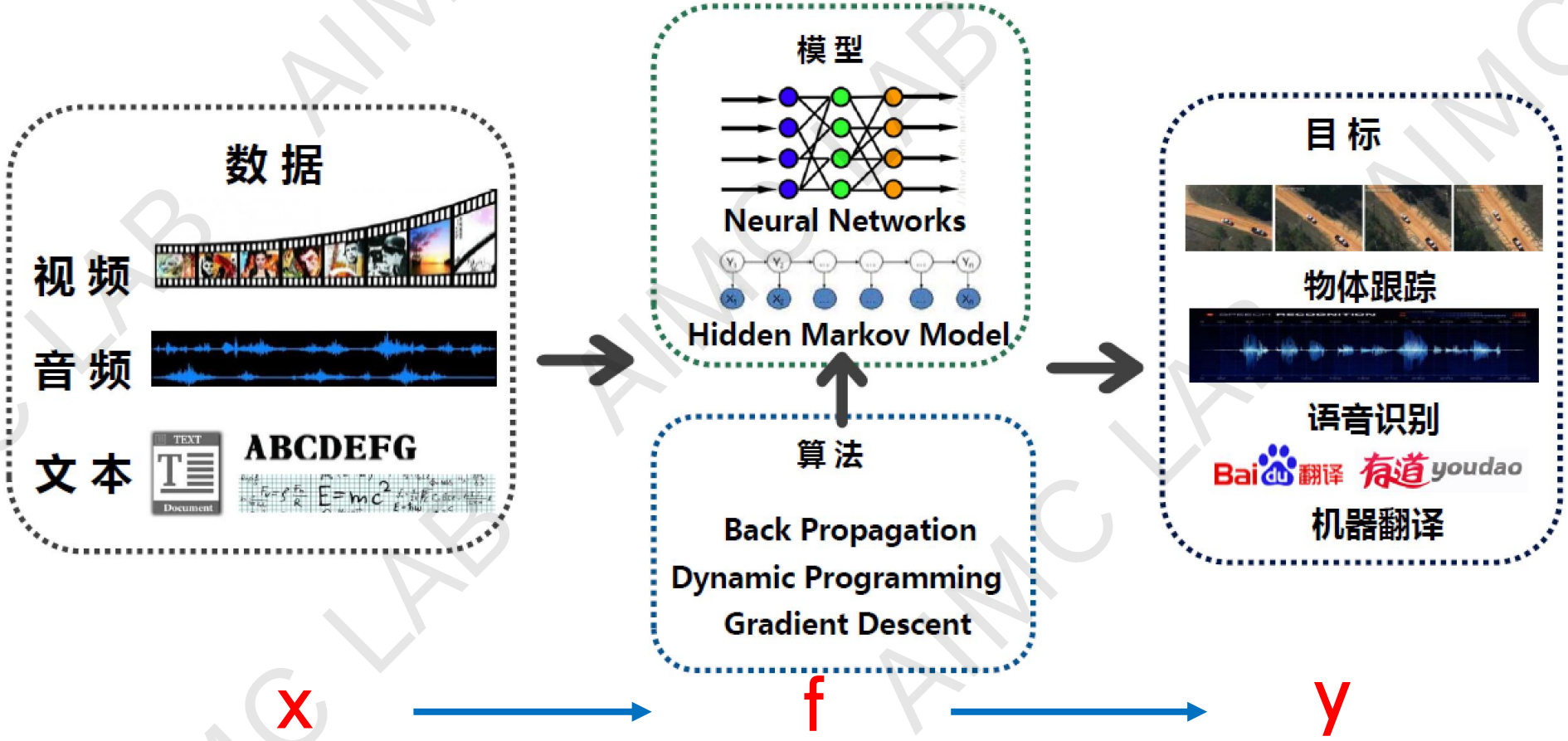
6.1 The Concept of Machine Learning

6.1.2 Analogy between Machine Learning and Human Learning

A baby learns to recognize objects through picture cards or real objects.

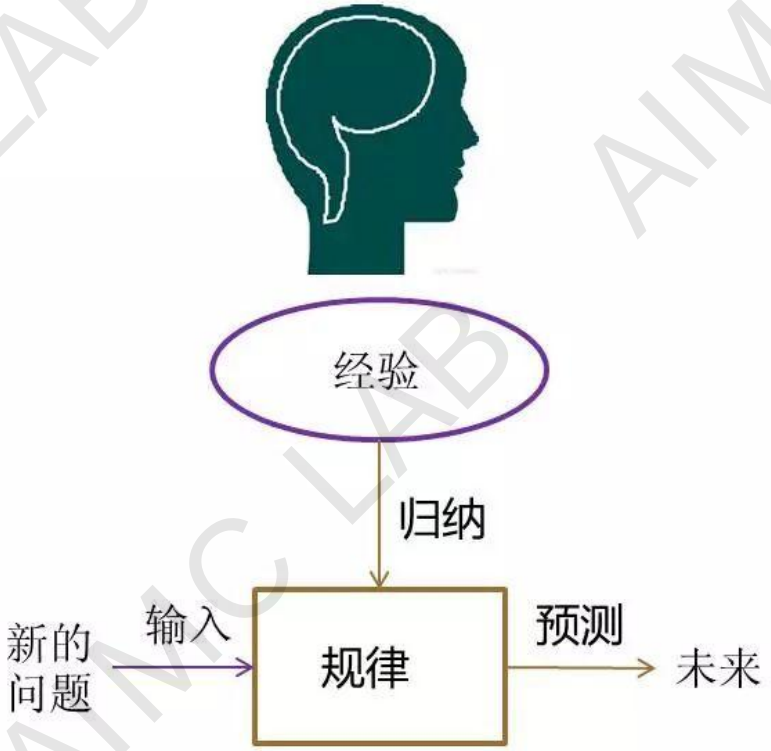
6.1 The Concept of Machine Learning

6.1.2 Analogy between Machine Learning and Human Learning



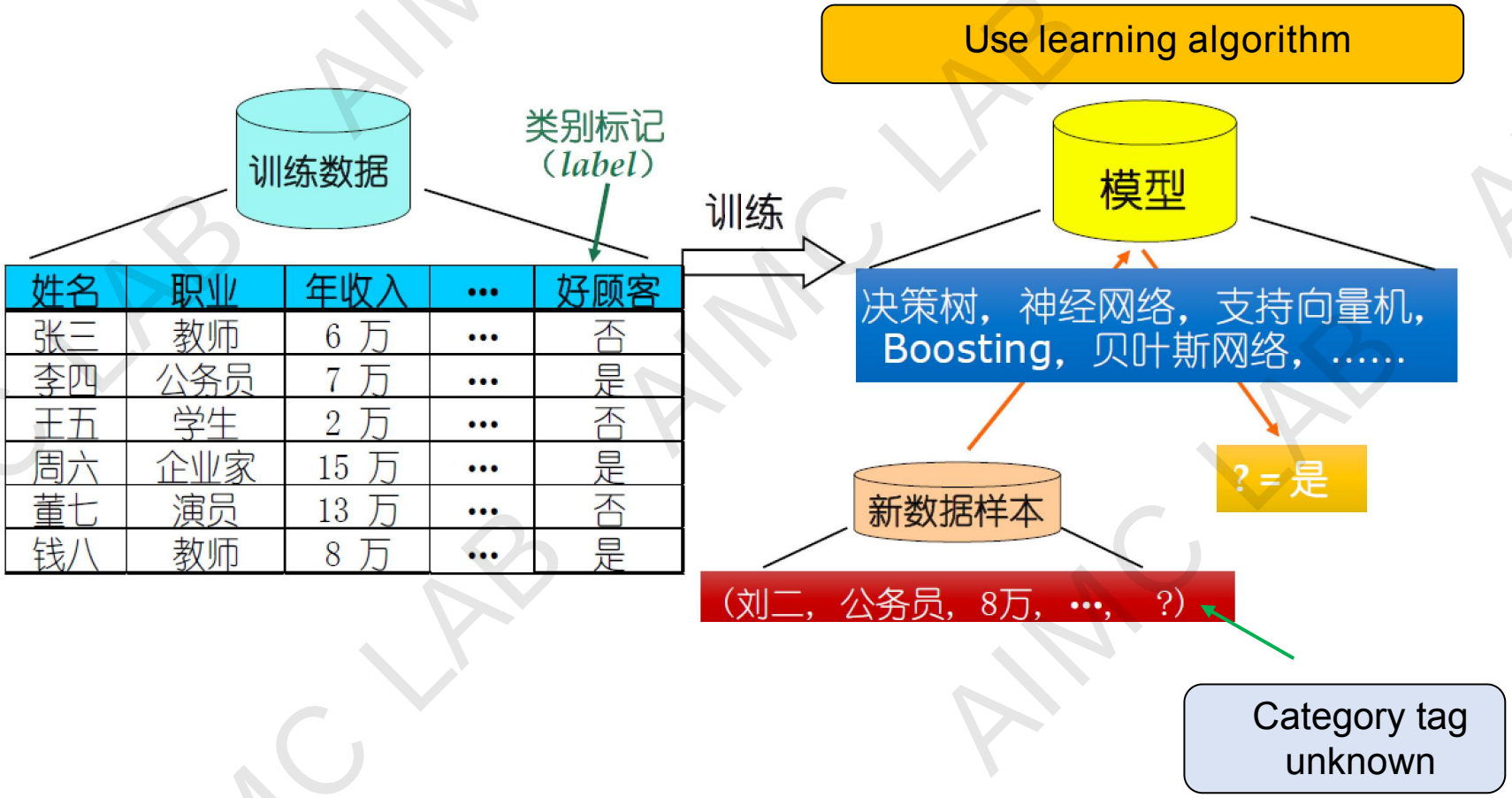
6.1 The Concept of Machine Learning

6.1.2 Analogy between Machine Learning and Human Learning



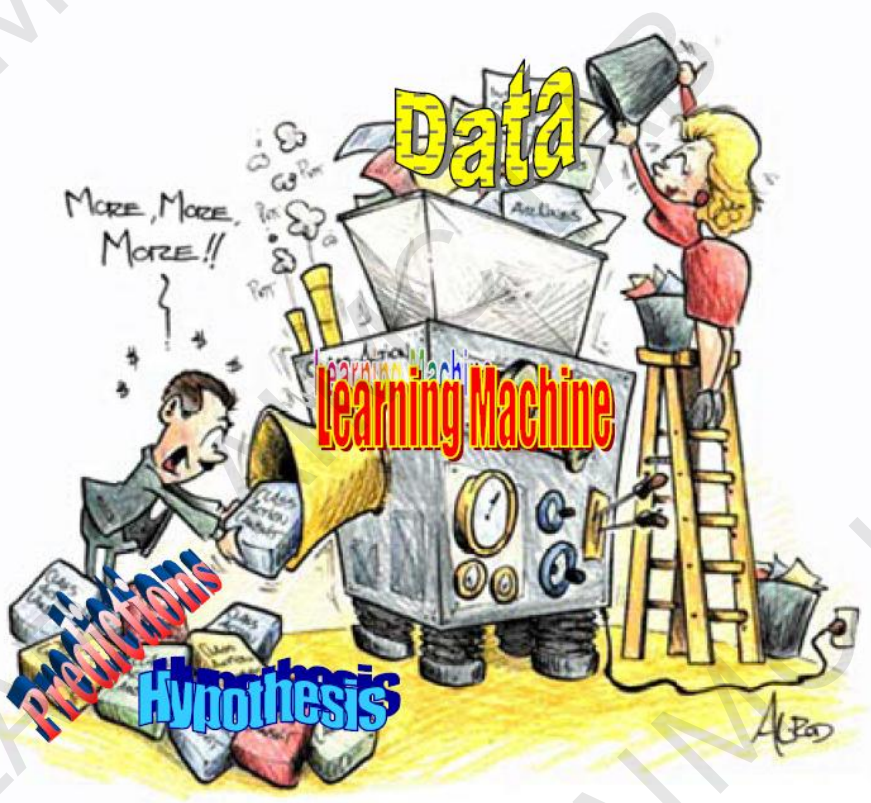
6.1 The Concept of Machine Learning

6.1.3 Typical Machine Learning Process



6.1 The Concept of Machine Learning

6.1.3 Typical Machine Learning Process



A vivid description of machine learning

6.1 The Concept of Machine Learning

Data Collection → Data Cleaning → Feature Engineering → Data Modeling

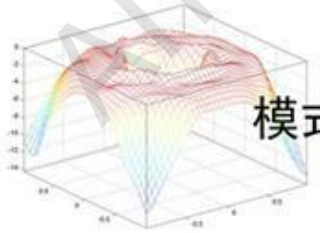






6.1 The Concept of Machine Learning

6.1.4 Connection between Machine Learning and Other Courses



模式识别

计算机视觉

数据挖掘

机器学习

语音识别

统计学习

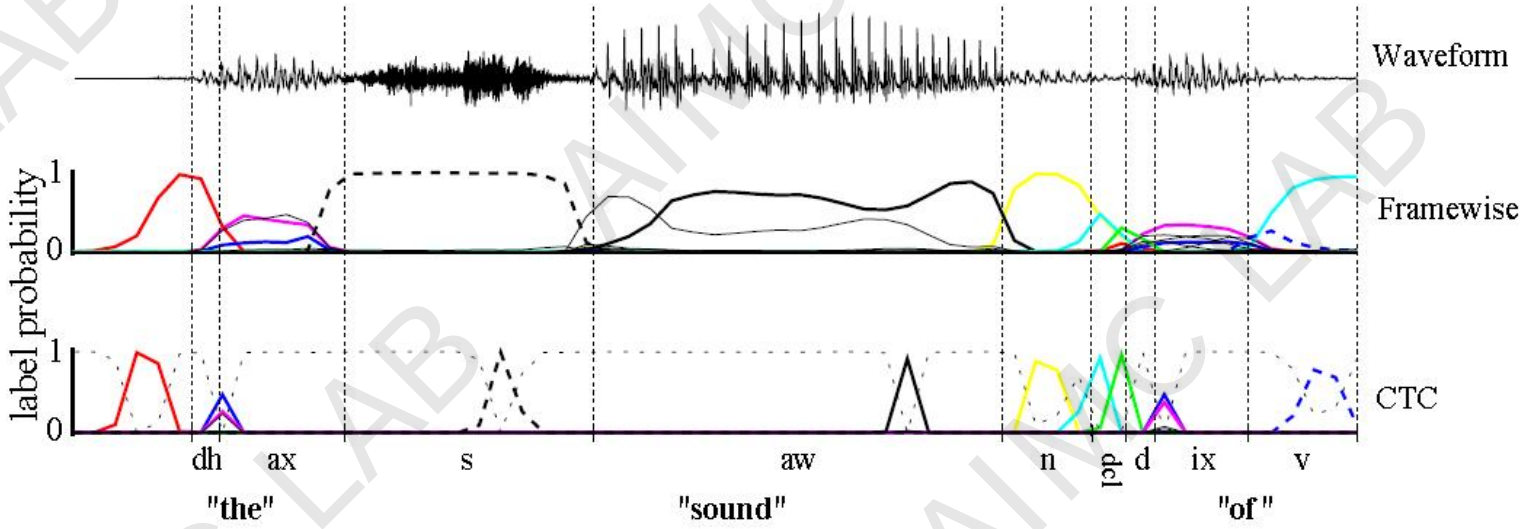
自然语言处理

6.1 The Concept of Machine Learning

6.1.5 Main Application Areas of Machine Learning

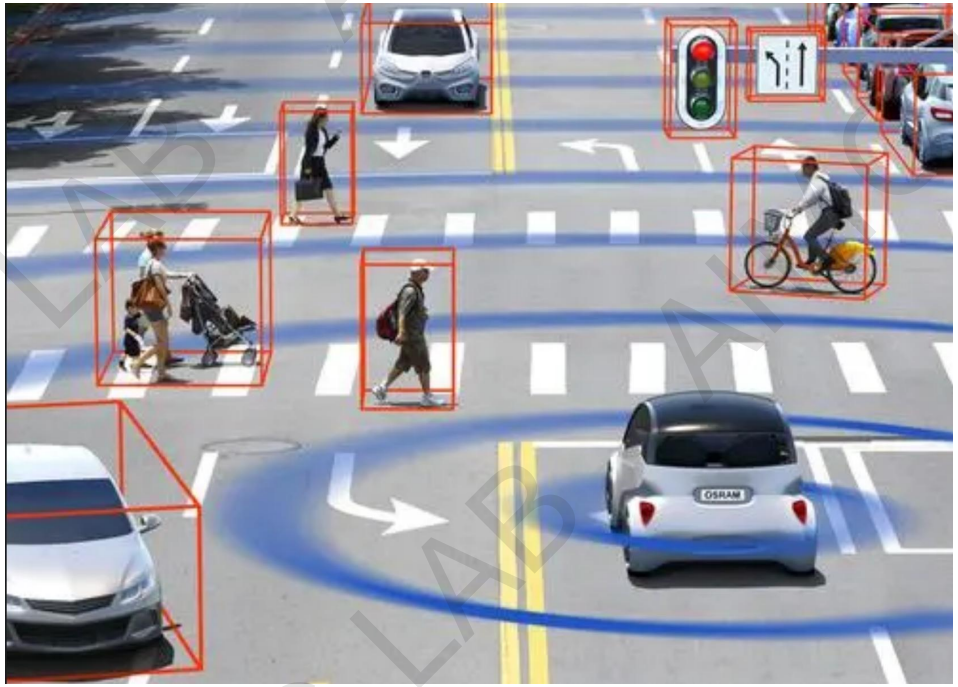
Text to speech and speech recognition

6.1 The Concept of Machine Learning



Speech Recognition

6.1 The Concept of Machine Learning



- **Autonomous Driving**

6.1 The Concept of Machine Learning

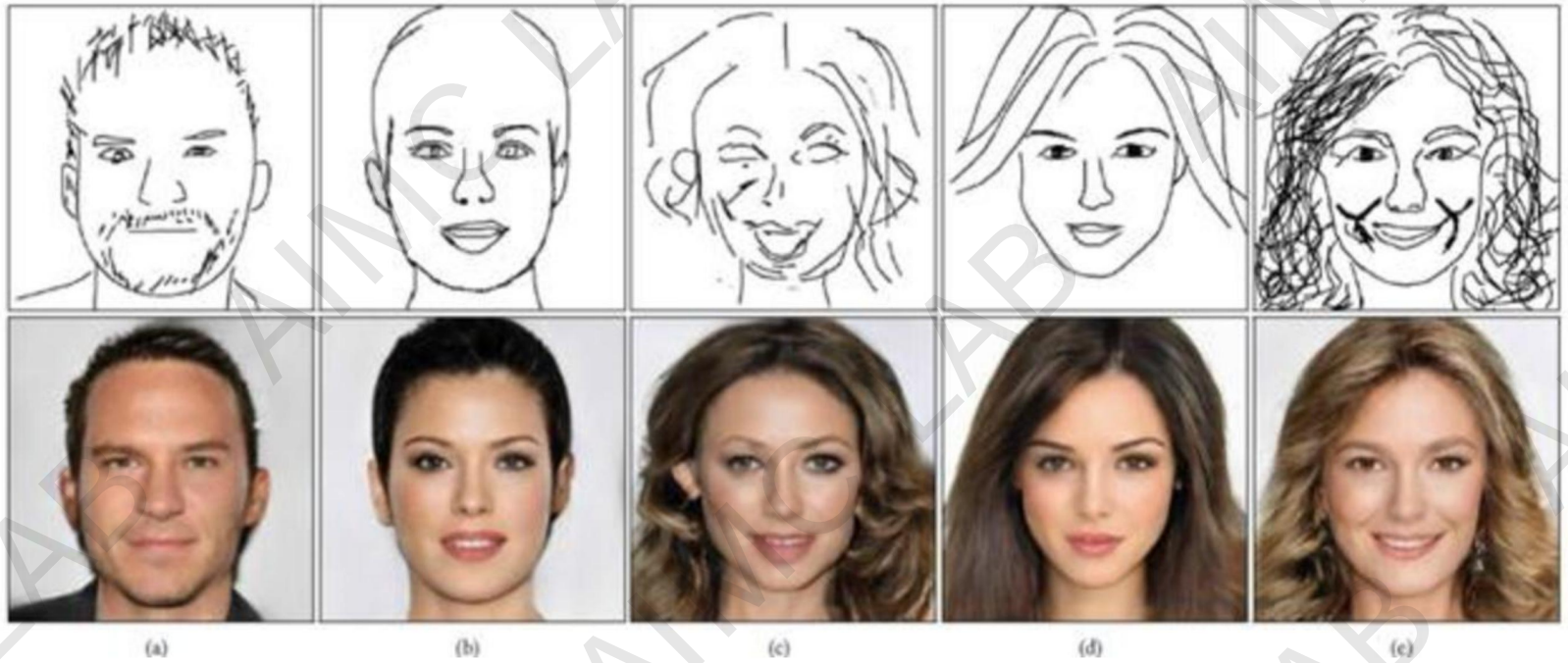
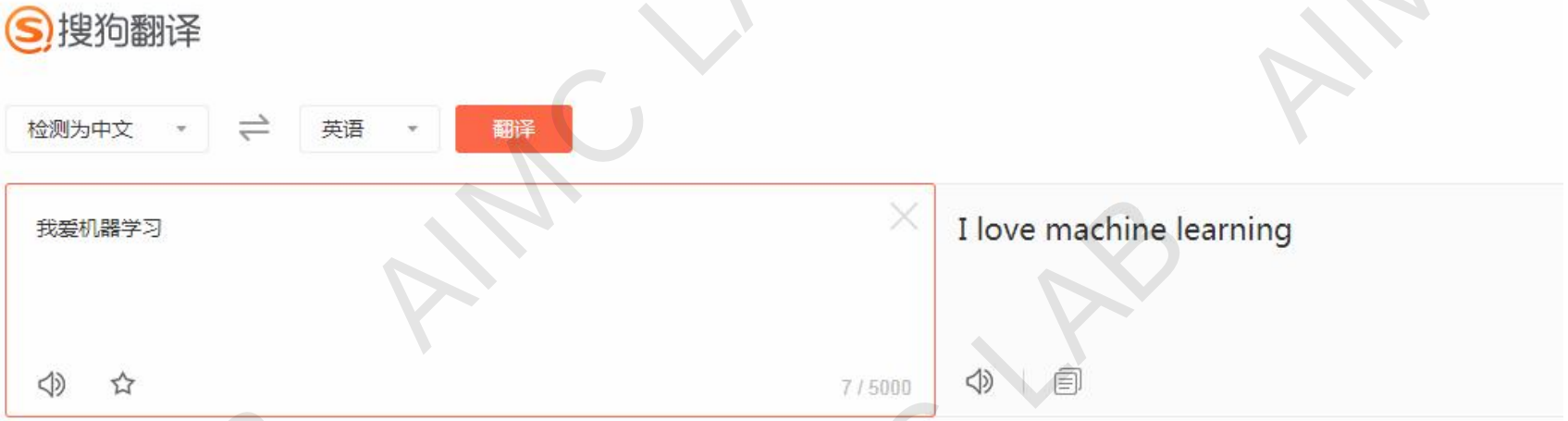


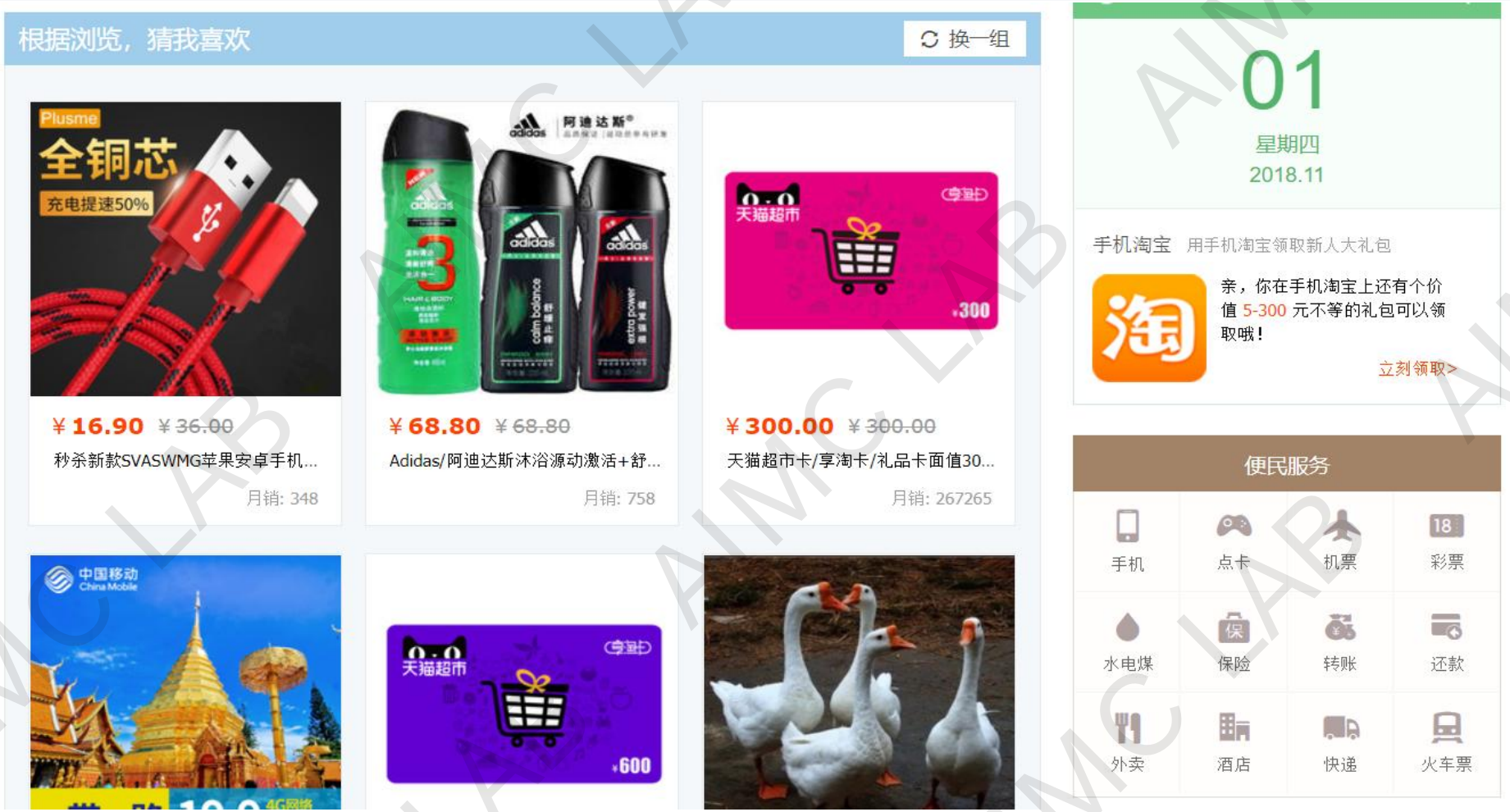
Image Generation

6.1 The Concept of Machine Learning



Machine Translation

6.1 The Concept of Machine Learning



- **Recommendation Systems**

Chapter 6: Contents

6.2 Basic Terminology of Machine Learning

- Feature（特征）
- Attribute（属性）
- Feature Vector（特征向量）
- Supervised Learning（监督式学习）
- Label（标签）
- Example（样本）
- Instance（实例）
- Regression（回归）
- Classification（分类）
- Tagging（标注）

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

6.2 Basic Terminology of Machine Learning

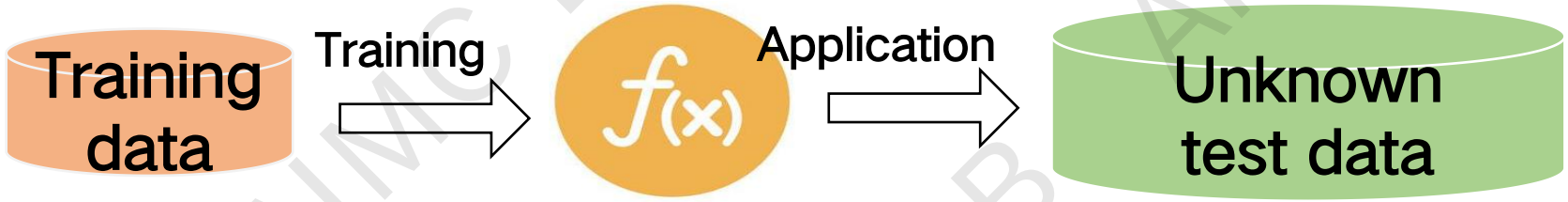
6.1.2 Basic Terminology

- Data Set（数据集）
- Training Set（训练集）
- Test Set（测试集）
- Independently and Identically Distributed，i.i.d（独立同分布）
- Model（模型）
- Training（训练）
- Hypothesis（假设）
- Learner（学习器）
- Input Space（输入空间）
- Output Space（输出空间）
- Hypothesis Space（假设空间）

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Chapter 6: Contents

6.3 Three Basic Elements of Machine Learning



Model

Classification

Regression

Clustering

Strategy

Loss Function(损失函数)

Selection

Model Selection

Algorithm

Determine parameter $\boldsymbol{\theta}$

Optimization(最优化):
Gradient descent,
least squares, etc.

6.3 Three Basic Elements of Machine Learning

6.3.1 Model

$$h(\boldsymbol{x};\boldsymbol{\theta})$$

- **Model**
  - **Probability(概率) Model** （Conditional Probability Distribution）
  - **Non-probabilistic models** (decision functions)
    - **Linear model** **线性模型**
    - **Nonlinear model** **非线性模型**

6.3 Three Basic Elements of Machine Learning

6.3.2 Learning Criterion (Strategy)

- **Loss Function（损失函数）：** Measures the error degree of a single sample prediction. The smaller the loss function value, the better the model.
- **Cost Function（代价函数）：** Cost Function: Measures the average error of the entire sample set. Common cost functions include mean squared error, root mean squared error, mean absolute error, etc.
- **Objective Function（目标函数）：** The final function to be optimized, including the cost function and regularization function.

6.3 Three Basic Elements of Machine Learning

- **Loss Function（损失函数）**
  - **0-1 Loss Function**

$$L(y_i, h(\boldsymbol{x}_i;\boldsymbol{\theta})) = \begin{cases} 0, h(\boldsymbol{x}_i;\boldsymbol{\theta}) = y_i \\ 1, h(\boldsymbol{x}_i;\boldsymbol{\theta}) \neq y_i \end{cases}$$

  - **Squared Loss Function**

$$L(y_i, h(\boldsymbol{x}_i;\boldsymbol{\theta})) = [y_i - h(\boldsymbol{x}_i;\boldsymbol{\theta})]^2$$

  - **Absolute Loss Function**

$$L(y_i, h(\boldsymbol{x}_i;\boldsymbol{\theta})) = |y_i - h(\boldsymbol{x}_i;\boldsymbol{\theta})|$$

6.3 Three Basic Elements of Machine Learning

- **Loss Function（损失函数）**
  - **Log Loss Function**

$$L(y_i, h(\boldsymbol{x}_i;\boldsymbol{\theta})) = -\log P(y_i \mid \boldsymbol{x}_i)$$

  - **Cross-Entropy Loss Function**

$$L(y_i, h(\boldsymbol{x}_i;\boldsymbol{\theta})) = -[y_i \log h(\boldsymbol{x}_i;\boldsymbol{\theta}) + (1-y_i)\log(1-h(\boldsymbol{x}_i;\boldsymbol{\theta}))]$$

  - **Hinge Loss Function合页损失函数(Hinge损失)**

$$L(y_i, h(\boldsymbol{x}_i;\boldsymbol{\theta})) = \max\{0, 1-y_i h(\boldsymbol{x}_i;\boldsymbol{\theta})\}$$

6.3 Three Basic Elements of Machine Learning

6.3.3 Optimization Algorithm

- 梯度下降法（Gradient Descent，GD）
- 随机梯度下降（Stochastic Gradient Descent，SGD）
- 小批量梯度下降 （Mini-Batch Gradient Descent，MBGD）
- 牛顿法
- 拟牛顿法

Chapter 6: Contents

6.4 Classification of Machine Learning Models

6.4.1 Classification by Task Type

- **回归**（Regression）
- **分类**（Classification）
- **聚类**（Clustering）
- **维数约简**（Dimensionality Reduction）

6.4 Classification of Machine Learning Models

6.4.1 Classification by Task Type

- **回归（Regression）**
- **案例：预测明天上海的温度是多少度?**

类型：回归问题

特点：输出是连续数值

典型算法：线性回归、决策树回归

6.4 Classification of Machine Learning Models

6.4.1 Classification by Task Type

- **分类（Classification）**
- **案例：判断一封邮件是否是垃圾邮件?**

类型：分类问题

特点：输出是离散类别

典型算法：逻辑回归、支持向量机

6.4 Classification of Machine Learning Models

6.4.1 Classification by Task Type

- **聚类（Clustering）**
- **案例：电商平台有100万用户，如何将他们分组?**

类型：聚类问题

特点：没有预设标签，让数据自己‘说话’

典型算法：K-means、层次聚类

6.4 Classification of Machine Learning Models

6.4.1 Classification by Task Type

- **维数约简** (Dimensionality Reduction)
- **案例：高维度描述：身高、体重、发色、瞳孔颜色、鞋子尺码、衣服品牌、早餐吃什么...（100个特征）**
- **维数约简后：①外貌吸引力分数、②健康指数、③时尚程度**

类型：降维问题

特点：减少特征数量，保留主要信息

典型算法：PCA（主成分分析）、t-SNE

6.4 Classification of Machine Learning Models

6.4.1 Classification by Task Type

方面	回归	分类	聚类	维数约简
任务类型	预测连续值	预测离散类别	发现数据分组	减少特征数量
监督性	监督学习	监督学习	无监督学习	通常无监督
输出形式	数值（如价格）	类别标签（如猫/狗）	分组标签（如群1/群2）	低维表示（如2D坐标）
典型问题	"多少？"	"是什么？"	"怎么分组？"	"怎么简化？"
评估方式	MSE, R^2	准确率, F1	轮廓系数, SSE	重建误差, 方差解释率
常用算法	线性回归	逻辑回归, SVM	K-means, DBSCAN	PCA, t-SNE

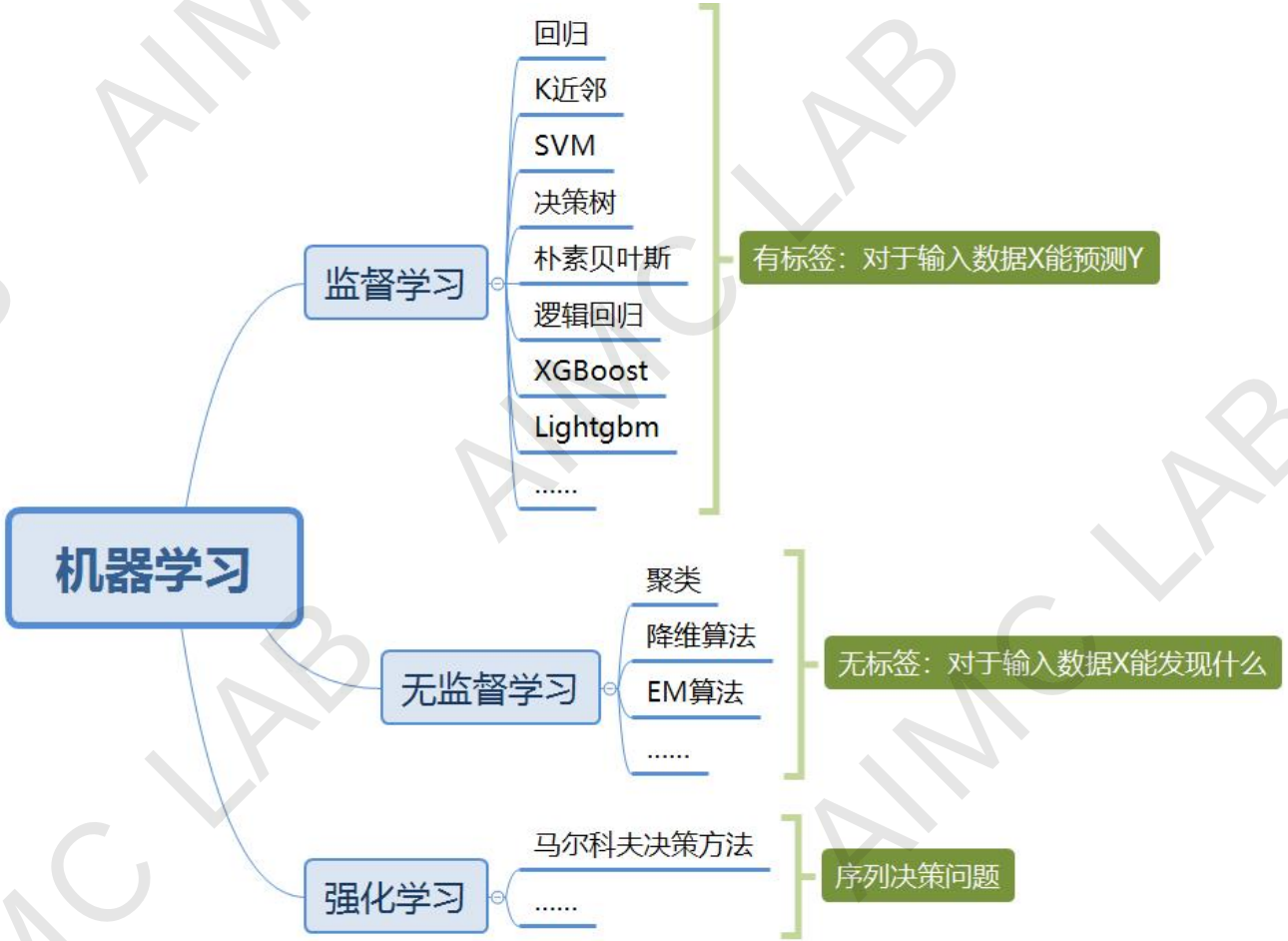
6.4 Classification of Machine Learning Models

6.4.2 Classification by Learning Method

- **Supervised Learning**（监督式学习、监督学习）
  - **Uses a set of labeled samples to adjust model parameters and improve model performance.**
- **Unsupervised Learning** （非监督式学习、无监督学习）
  - Analyzes data by comparing some relationship between samples.
  - The key characteristic is that the input to the learning algorithm is unlabeled samples.
- **Reinforcement Learning**（强化学习）
  - **A machine learning approach that adjusts machine behavior based on feedback to achieve automated decision-making.**

6.4 Classification of Machine Learning Models

6.4.2 Classification by Learning Method



6.4 Classification of Machine Learning Models

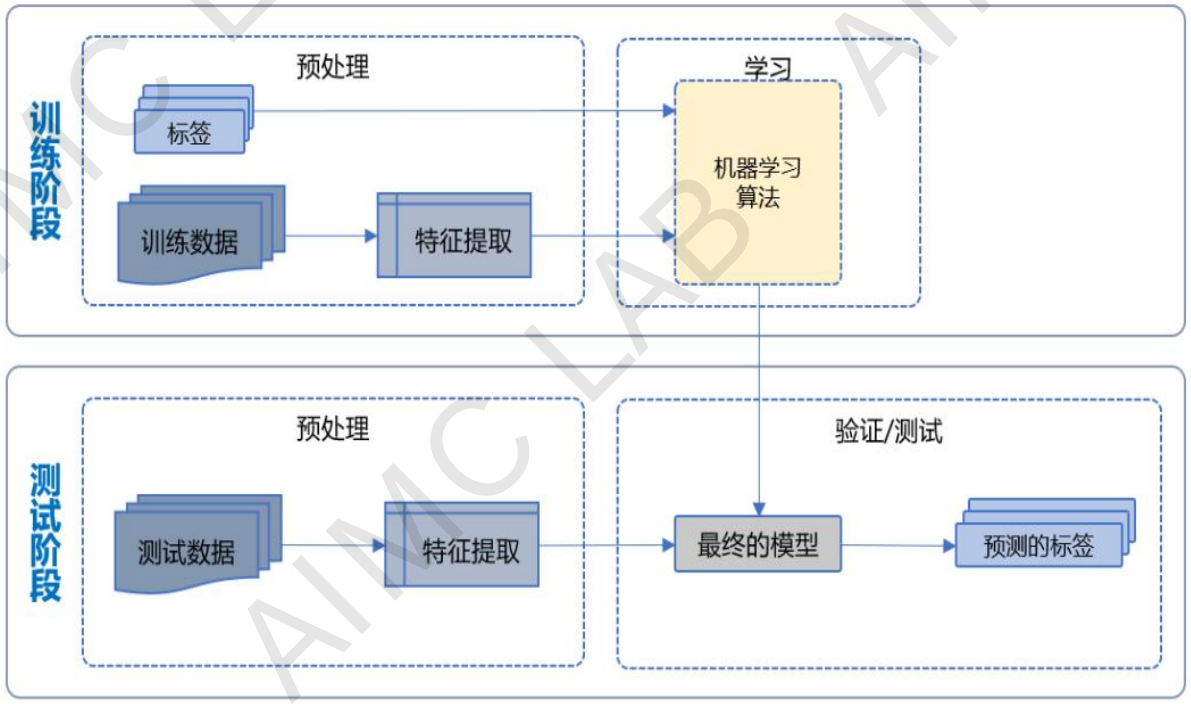
6.4.2 Classification by Learning Method

- **Supervised Learning** （监督式学习）

✓ Classification（分类）

✓ Is a man 1.65m tall and 100kg obese?

✓ Determine benign or malignant based on tumor volume and patient age?



6.4 Classification of Machine Learning Models

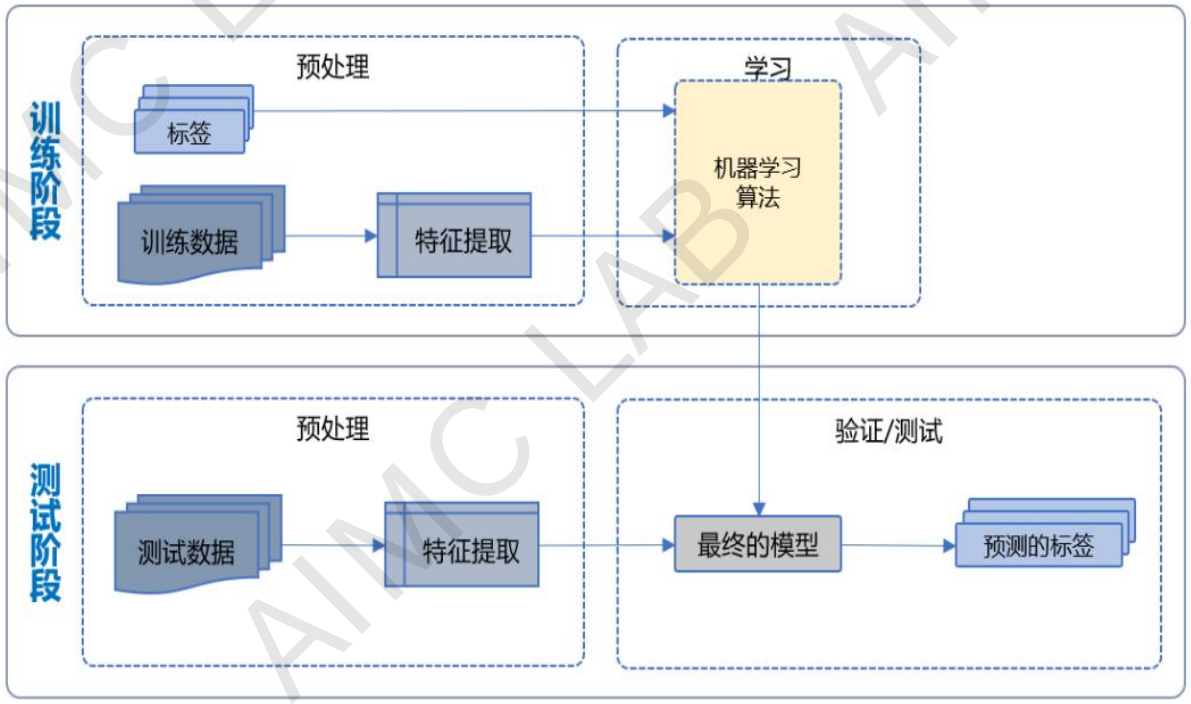
6.4.2 Classification by Learning Method

- **Supervised Learning** （监督式学习）

✓ Regression、Prediction（回归）

✓ How to predict housing prices in Shanghai Pudong?

✓ Future stock market trends?



6.4 Classification of Machine Learning Models

6.4.2 Classification by Learning Method

- **Unsupervised Learning （无监督式学习）**

  ✓ Clustering（聚类）

  ✓ How to divide students in a classroom into 5 categories based on hobbies and height?

  ✓ Dimensionality Reduction （降维）

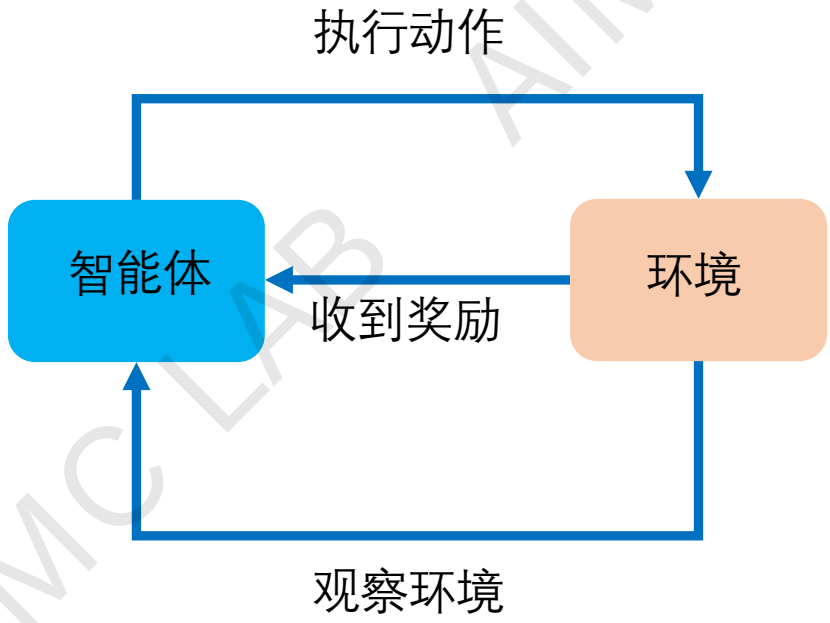
  ✓ How to map data points from the original high-dimensional space to a low-dimensional space?

6.4 Classification of Machine Learning Models

6.4.2 Classification by Learning Method

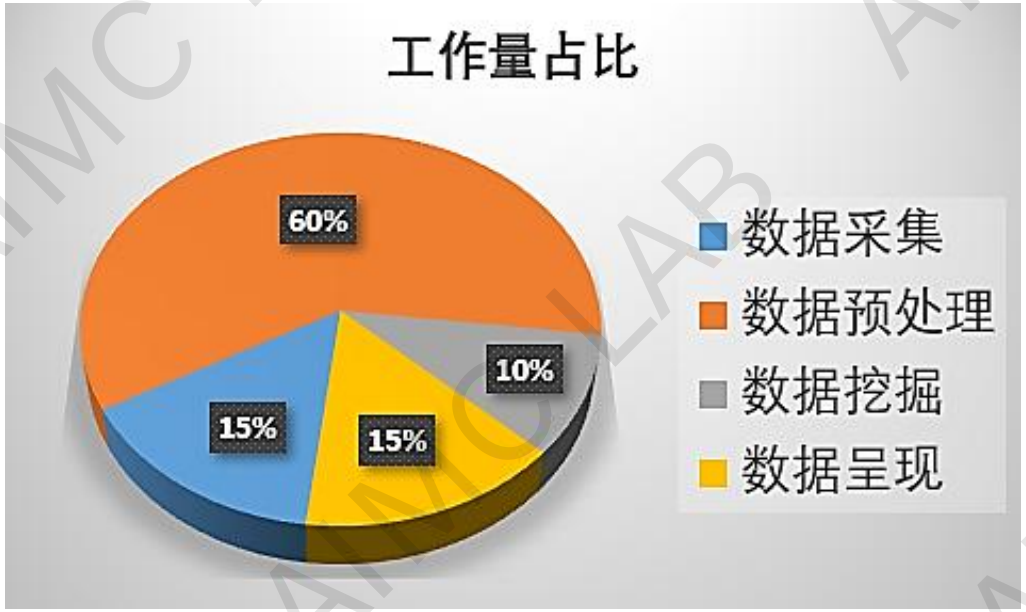
- **Reinforcement Learning**（强化学习）

✓ Describes and addresses problems where intelligent agents learn strategies through interaction with their environment to maximize rewards or achieve specific objectives.



Chapter 6: Contents

6.5 Data Preprocessing



- **The quality of the input data determines the quality of the output results of the data mining model, i.e., the data determines the upper limit of the model. Without high-quality data, there is no high-quality mining result!**