

2024-2025学年秋季学期

# 模 式 识 别

## 第一章：绪论

主讲人：张治国

[zhiguo.zhang@hit.edu.cn](mailto:zhiguo.zhang@hit.edu.cn)

# 本章内容

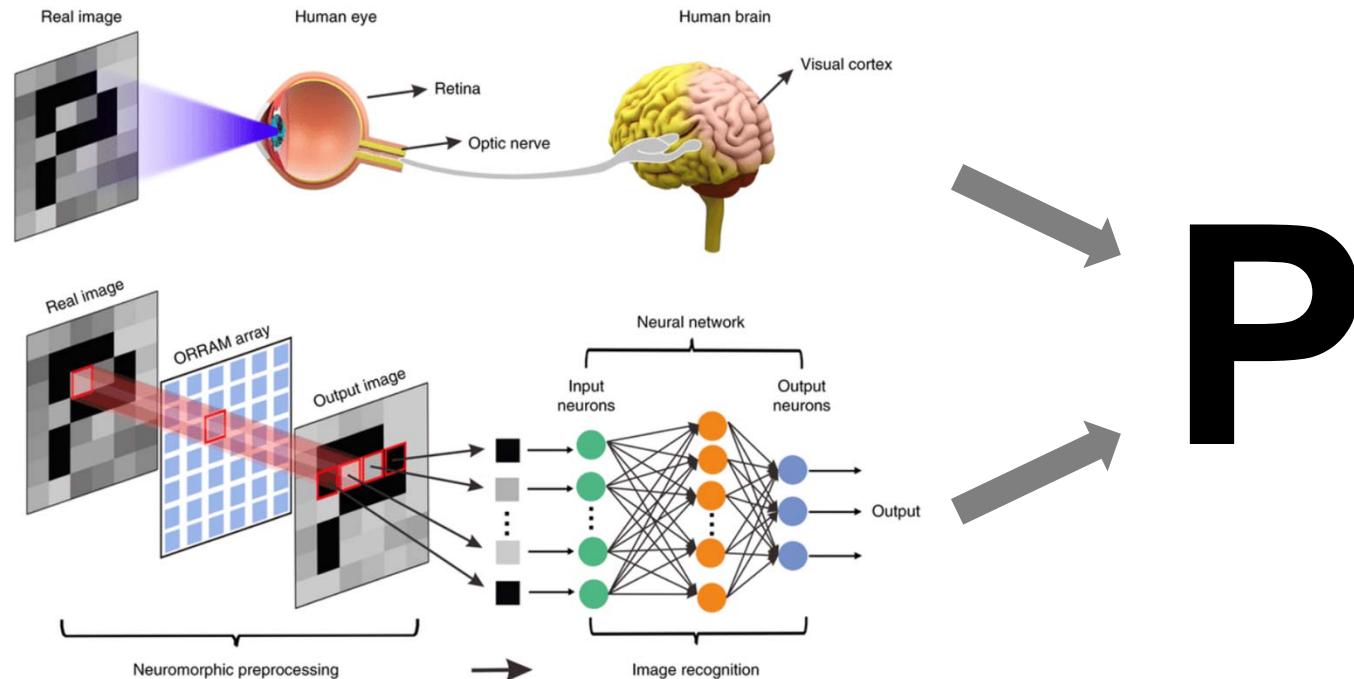
---

- ✓ • 模式识别基本概念
- ✓ • 模式识别系统
- 模式识别的应用
- 模式识别系统应用注意事项
  - 数据采集和样本集的构建
  - 预处理（处理缺失值和离群点，数据标准化）
  - 特征工程简介
  - 模式识别方法分类

# 模式识别基本概念

- 识别能力是人类最基本的智能行为。
  - 人是如何识别对象的？如何具有认知能力？
- 模式识别（Pattern Recognition）：**从工程的角度研究如何使计算机具有识别能力的理论和方法

○  
人类识别  
VS.  
机器识别



# 模式识别基本概念

- 模式 (Pattern) : 待识别对象的一组属性集合。



对象



属性 (特征) : {颜色、形状、外表、酸度、重量……}

模式

- 识别 (Recognition) : 根据模式判断不同的对象是否属于同类或属于哪种类别。



苹果梨子分类

特征：颜色、形状等



苹果好坏分类

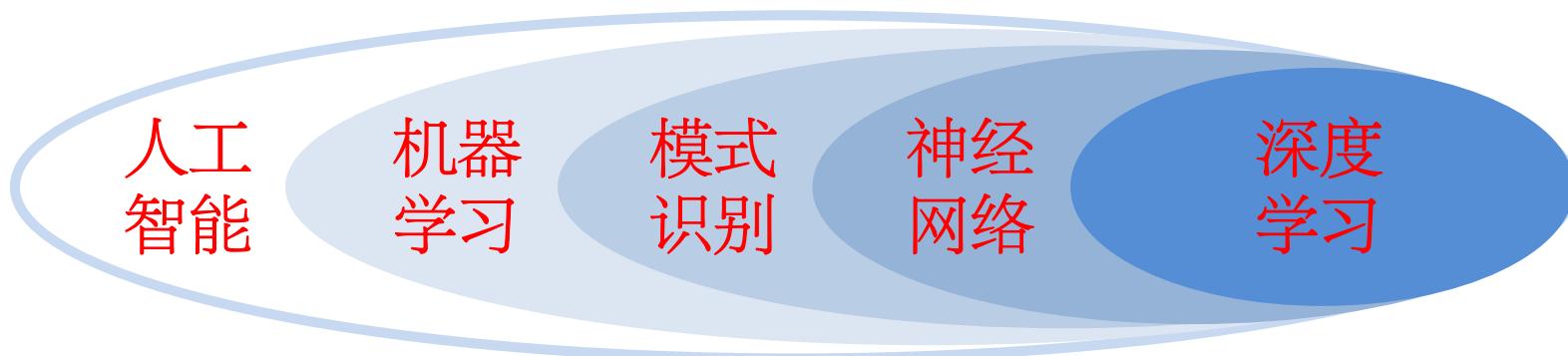
特征：颜色、外表、酸度等

针对具体特定的应用，选择有效的特征及分类方法。

# 模式识别基本概念

---

- 模式识别和以下名词有类似含义。
- 人工智能：通过计算机程序的手段实现类人智能
- 机器学习：使机器基于数据进行自主学习和预测
- 神经网络：模拟神经系统结构处理输入输出信息
- 深度学习：用多层次级联神经网络进行机器学习



# 模式识别的历史

---

- 模式识别诞生于20世纪20年代，随着40年代计算机的出现，50年代人工智能的兴起，模式识别在60年代初迅速发展成一门学科。
  - 1929年Tauschek发明阅读机，能够阅读0~9的数字。
  - 20世纪30年代Fisher提出统计分类理论。
  - 50年代Chomsky提出形式语言理论，付京孙提出句法结构识别。
  - 60年代Rosenblatt提出感知器，Zadeh提出模糊集理论。
  - 1977年IEEE计算机学会成立模式分析与机器智能（PAMI）委员会，召开模式识别与图像处理学术会议（CVPR前身）。
  - 80年代Hopfield提出神经元网络模型理论。
  - 90年代Vapnik等人提出支持向量机SVM。
  - 21世纪初Hinton等人提出深度信念网络（DBNs）。
  - 2010年之后进入深度学习时代。
  - 2020年之后进入大模型/基础模型时代。

# 模式识别应用

---

- **工业用途：**产品质量检验，设备故障检测，智能机器人的感知系统；
- **商业用途：**钱币的自动识伪，信函的自动分拣，电话信息查询，声控拨号；
- **医学用途：**对心电、脑电、磁共振影像等信号进行处理和识别，自动进行疾病的诊断；
- **安全领域：**生理特征鉴别，网上电子商务的身份确认，对公安对象的刑侦和鉴别；
- **军事领域：**巡航导弹的景物识别，战斗单元的敌我识别；
- **办公自动化：**文字识别技术和声音识别技术；

.....  
模式识别几乎应用在所有需要人类识别甚至人类无法识别的场景中。

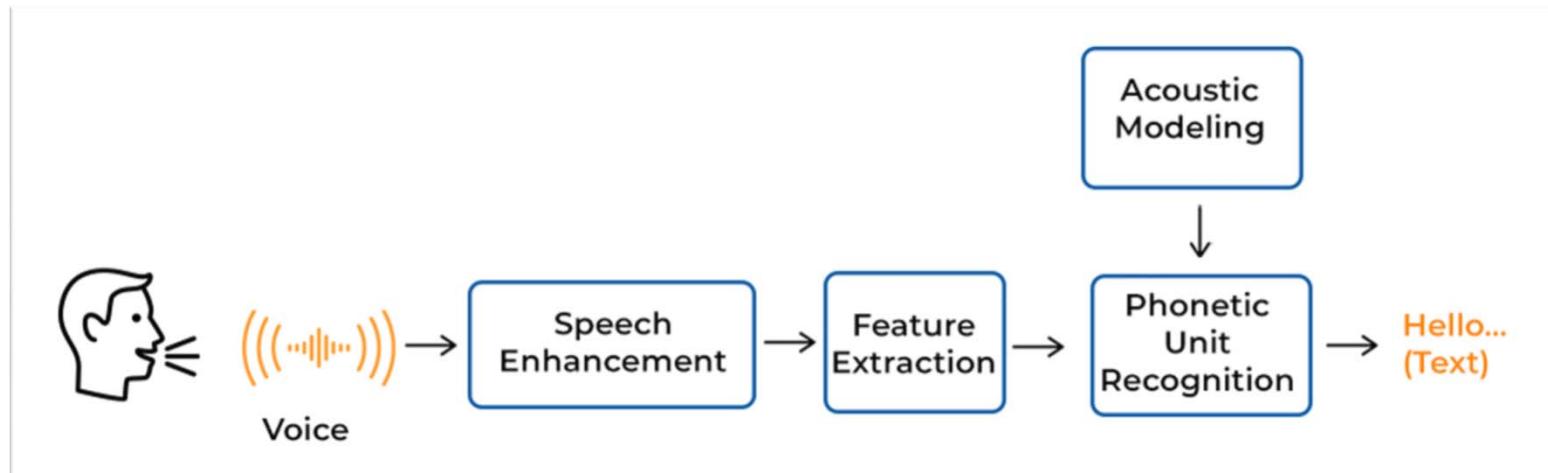
# 模式识别应用

- 例：人脸识别



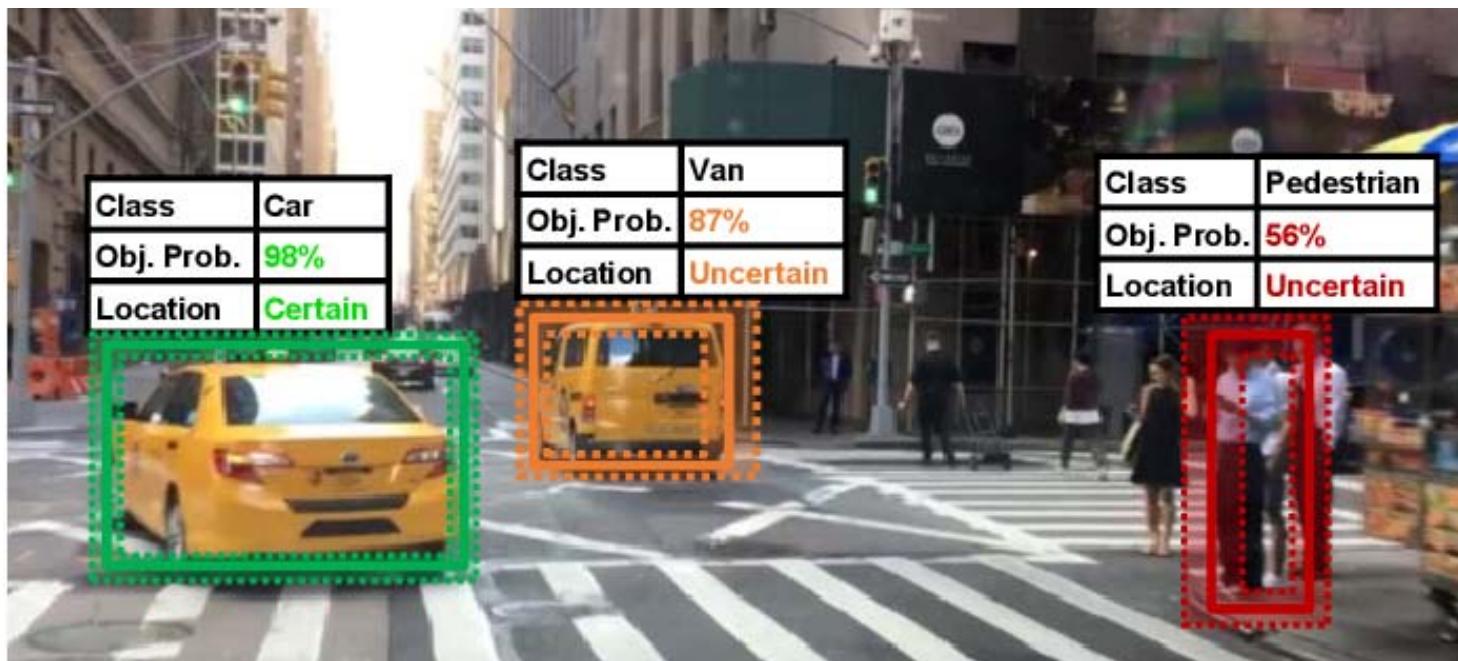
# 模式识别应用

- 例：语音识别



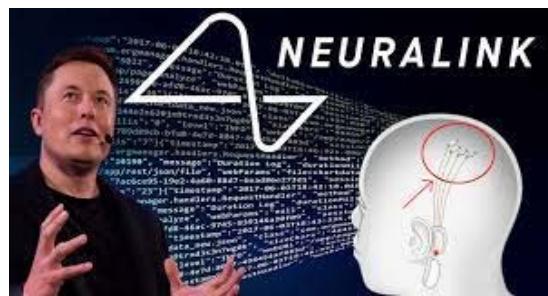
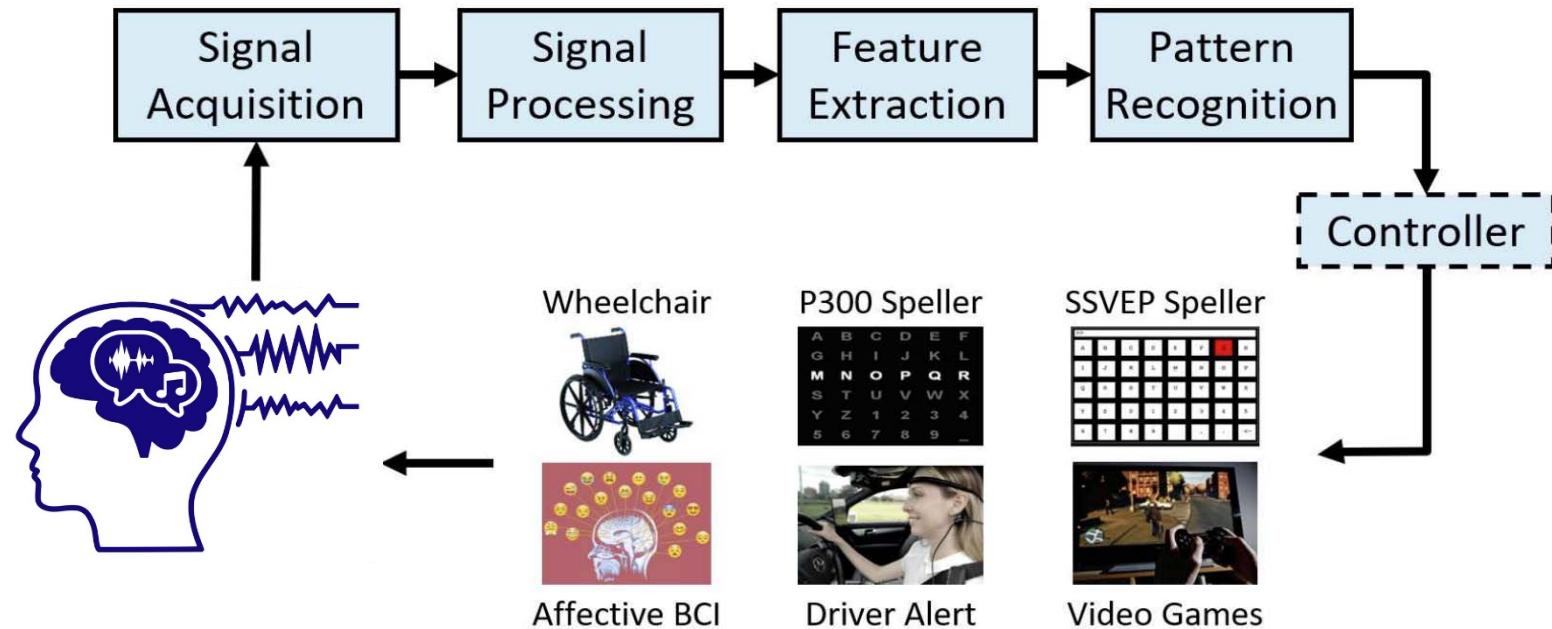
# 模式识别应用

- 例：自动驾驶



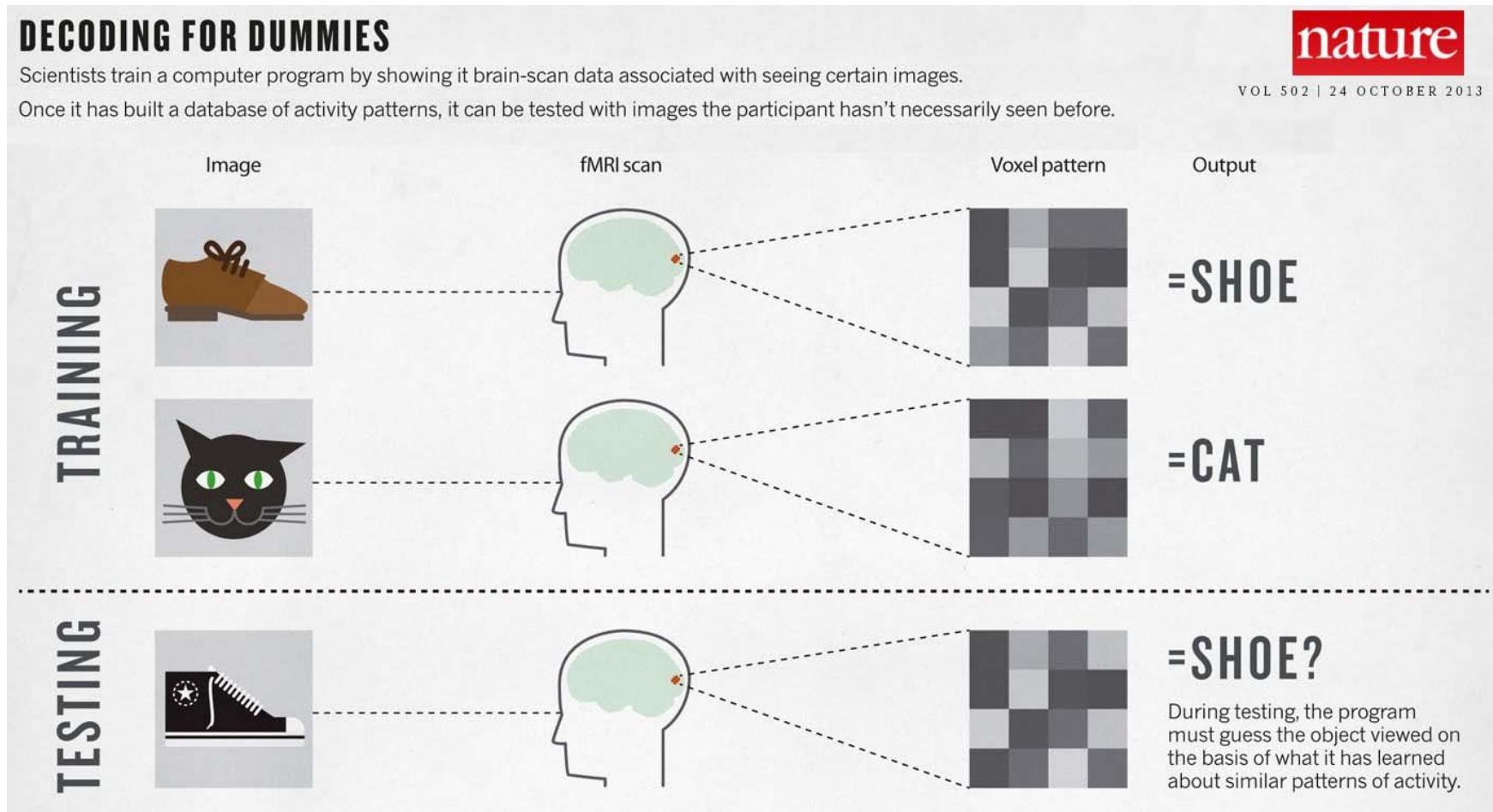
# 模式识别应用

- 例：脑机接口



# 模式识别应用

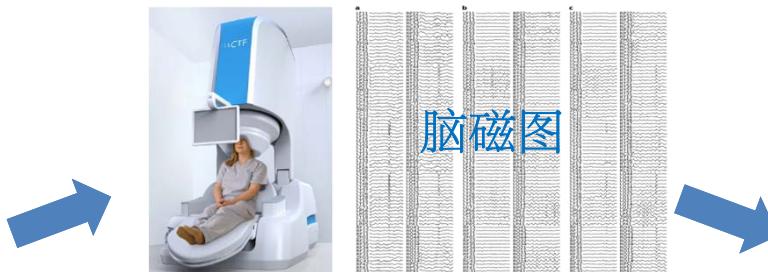
- 例：脑机接口



# 模式识别应用

- 例：脑机接口 - “读脑术”

<https://ai.meta.com/static-resource/image-decoding>



Viewed Image



Predicted Image

# 模式识别应用

- 例：脑机接口 - “读脑术”

2008

vs

2023

原始图像



重建图像



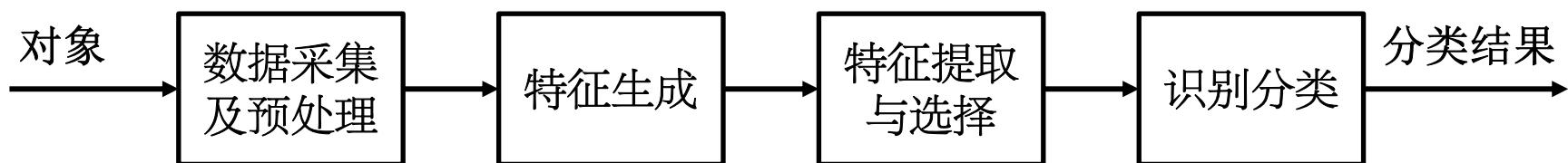
**Neuron**  
Visual Image Reconstruction from Human fMRI

**Image shown**  
(Viewed for one second)

**Decoded output**  
(Shown here at 1/4 speed)

# 模式识别系统

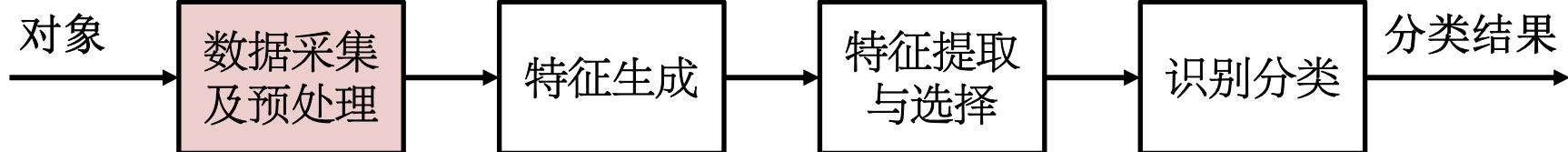
- 模式识别系统完成的工作是从外部世界获取一个所要识别对象的数据，经过分析处理后辨识其类别属性。
- 数据的采集、分析和处理一般包含以下步骤：
  - ✓ 数据采集及预处理
  - ✓ 特征生成
  - ✓ 特征提取与选择
  - ✓ 识别分类



# 模式识别系统

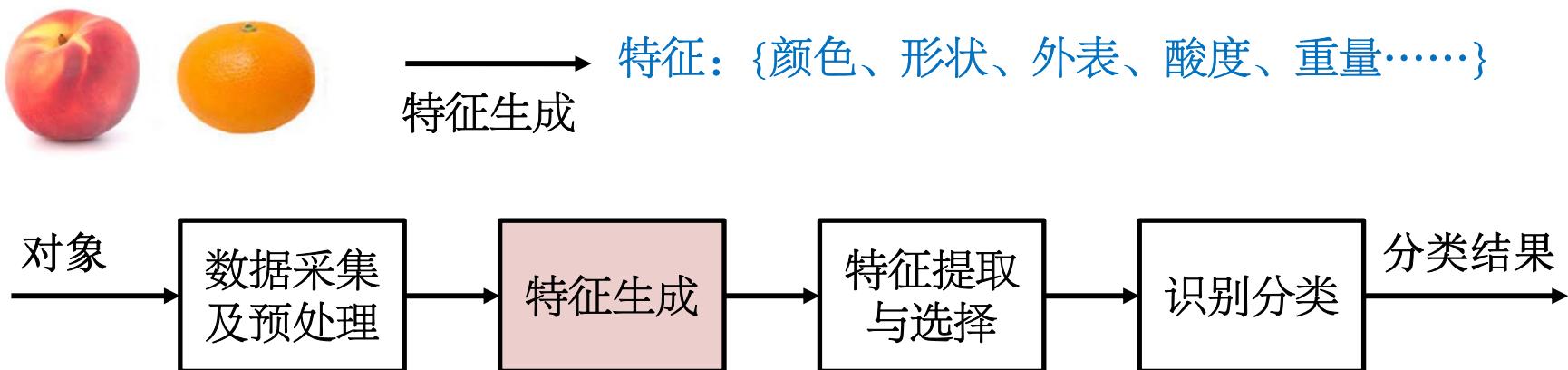
- 数据采集及预处理：

- **数据采集**是将外部世界需要识别的对象数字化为波形、图像、文本等计算机可处理的形式；
- **预处理**是将数据中混杂的与识别对象无关的噪声去除，将识别对象从背景中分离出来。



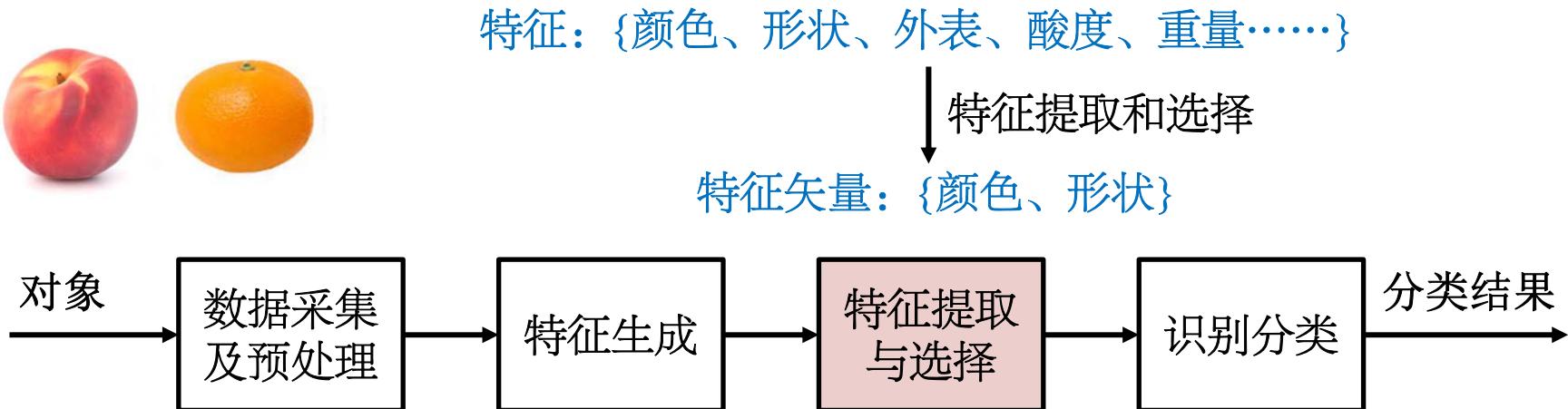
# 模式识别系统

- 特征生成：
  - **特征：**可以描述不同类别对象之间差异的易计算测量的属性；
  - 原始数据量大且复杂，难以直接识别，需要进一步处理找出特征，再将特征输入分类器以识别对象的类别。注：这是传统模式识别与深度学习的主要区别之一。



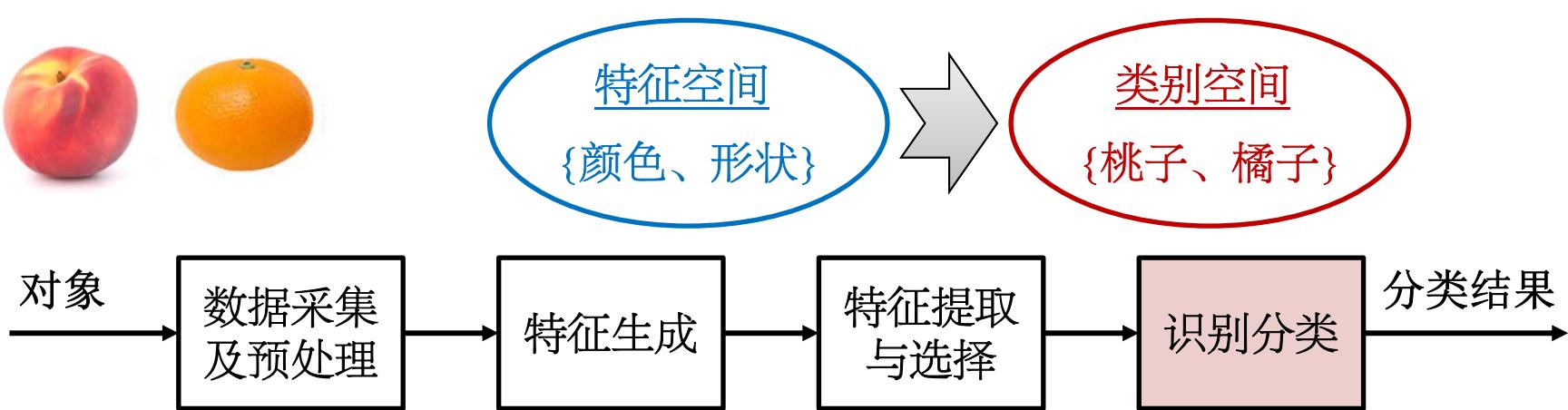
# 模式识别系统

- 特征提取与选择：
  - 选择什么特征进行识别是一个关键问题，选择的特征与分类问题密切相关。
  - 通常做法是生成尽可能多的特征，然后选择或组合出最有效的特征，这一过程称之为**特征提取和选择**。



# 模式识别系统

- 识别分类：
  - 特征选择后，每个对象被描述为一组特征，称为**特征矢量**，可看作**特征空间**的一个点。
  - 以特征矢量形式描述的对象称为**样本**。
  - 识别分类根据特征矢量判断样本类别，该过程是一个从**特征空间**到**类别空间**的映射。

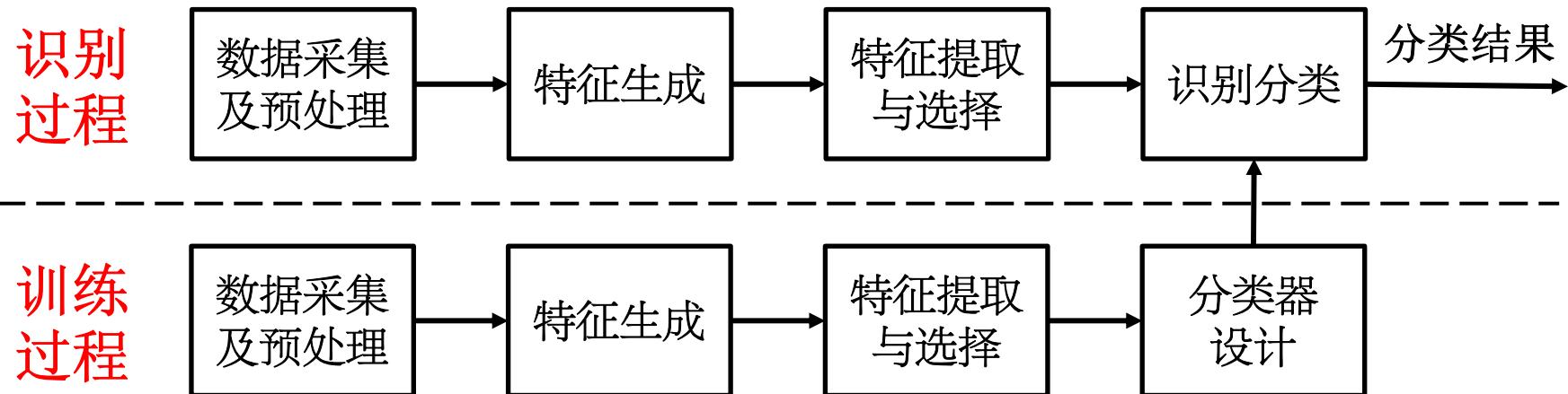


# 模式识别系统

- 问题：分类器如何设计？可否人工设计？
- 答案：人工设计低效且能力有限。
- 分类器一般需要一个**训练和学习**的过程。
- 训练过程中，设计者提供大量有代表性的识别对象的实例（称为**训练样本**），识别系统采用一定的训练和学习算法在训练样本基础上自动完成分类器的设计。

# 模式识别系统

- 完整的模式识别系统包括训练和识别两个过程。



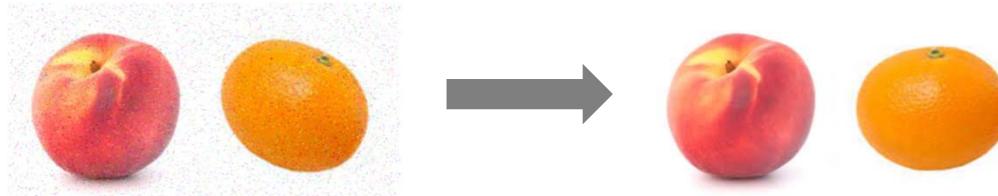
# 模式识别系统举例

---

水果识别：（教材《模式识别》3-5页）

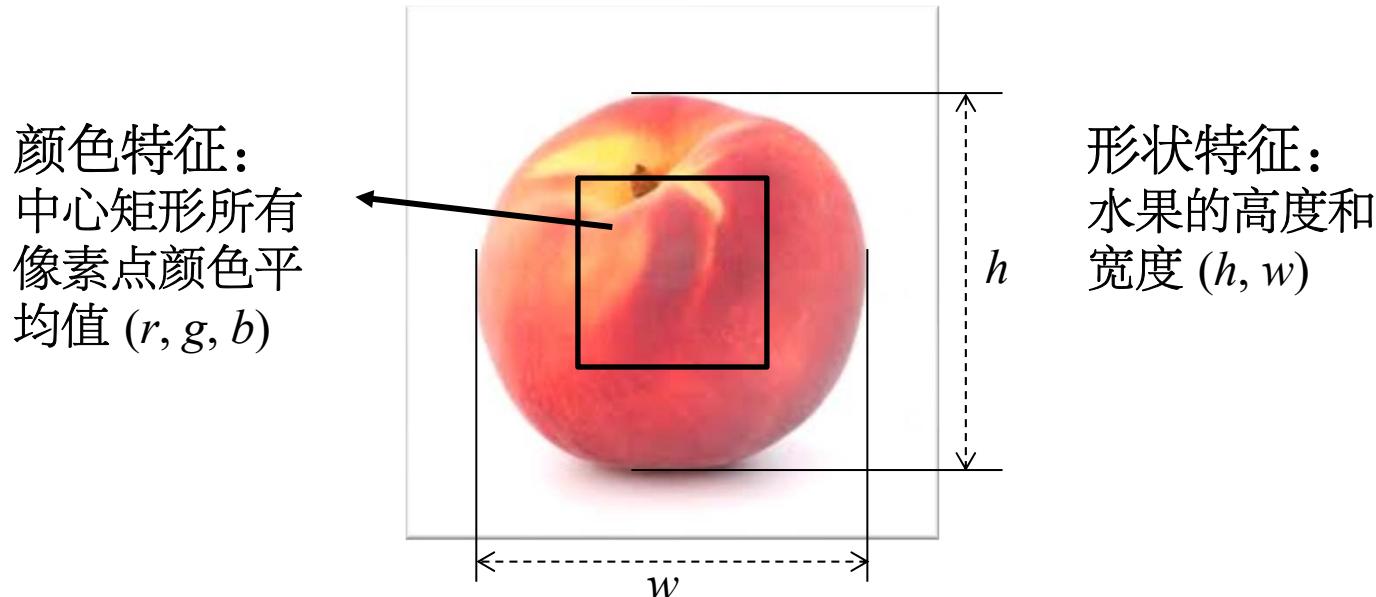
食品加工厂需要自动分开桃子和橘子两种水果。

- **数据采集：**在传送带特定位置安装摄像机，每一个水果到达镜头下方时自动拍摄，数字图像输入计算机。
- **预处理：**采用数字图像处理技术将水果从背景中分离，旋转摆正，去除噪声。



# 模式识别系统举例

- **特征生成：**桃子和橘子的颜色和形状有差别（桃子偏红，橘子偏橙；桃子略圆，橘子略扁），因此可以提取颜色和形状特征。



- 描述水果的特征矢量： $y = (r, g, b, h, w)^T$

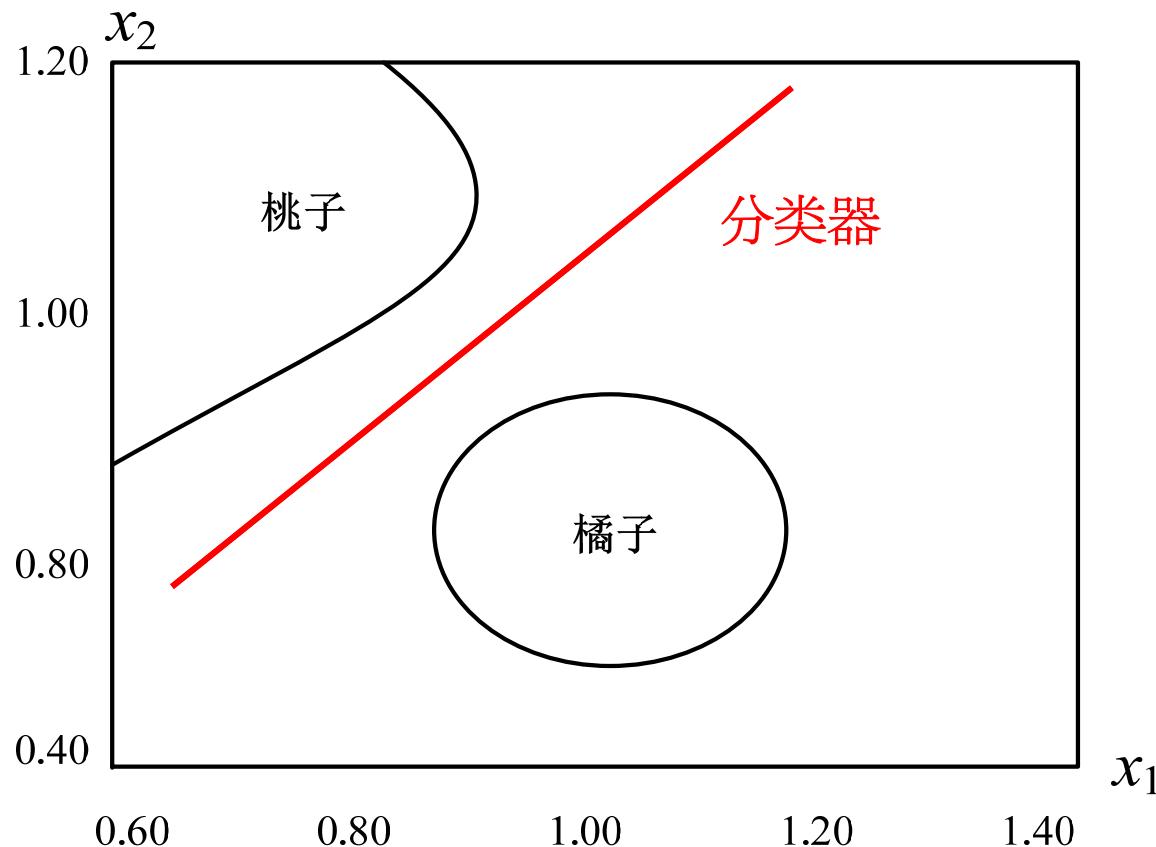
# 模式识别系统举例

---

- 特征提取和选择：
  - 区分桃子橘子的颜色分量主要是红色和绿色，因此可以从特征集合中去除蓝色。
  - 考虑到图片颜色会随着光照同步增强，因此可以定义新的颜色特征去除光照影响： $x_1 = g / r$ ；
  - 考虑到水果大小不同，可以构建新的形状特征避免水果大小的影响： $x_2 = h / w$ 。
- 选择后的描述水果的特征矢量： $x = (x_1, x_2)^T$

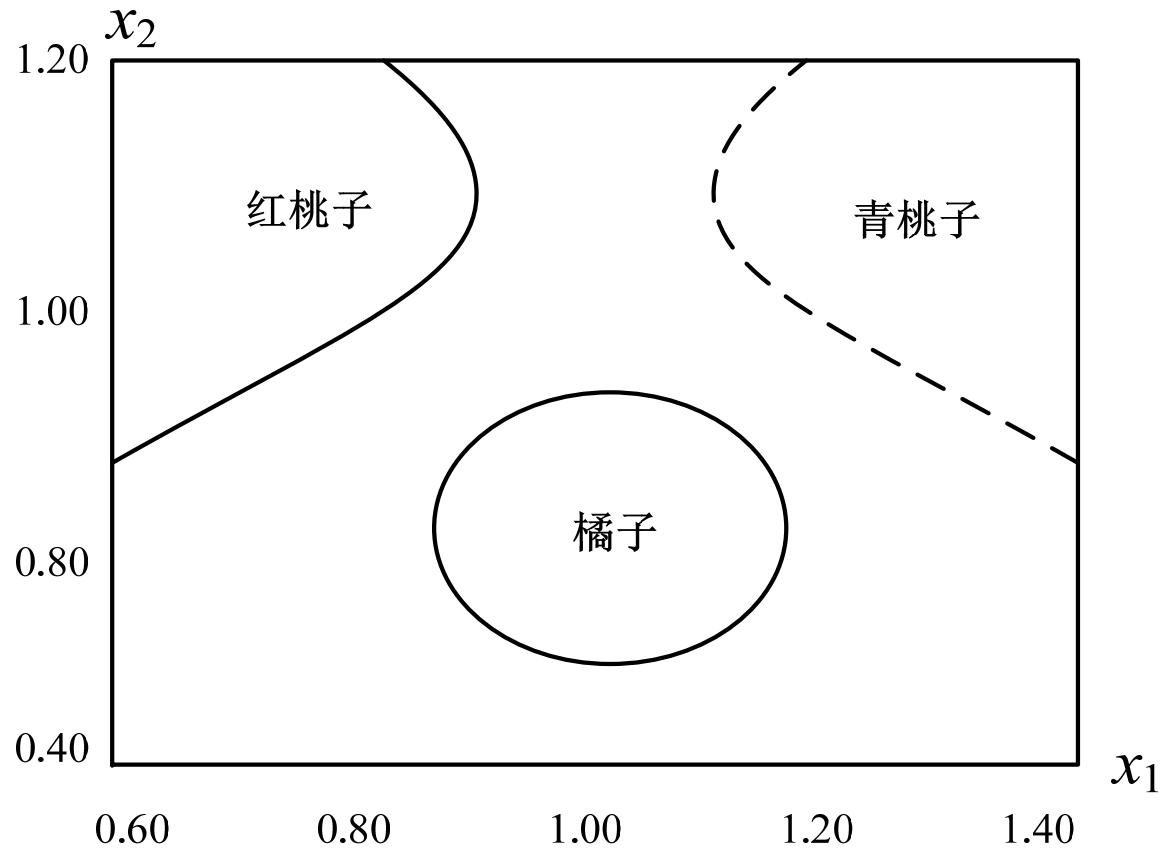
# 模式识别系统举例

- 特征空间中桃子和橘子的大致分布：



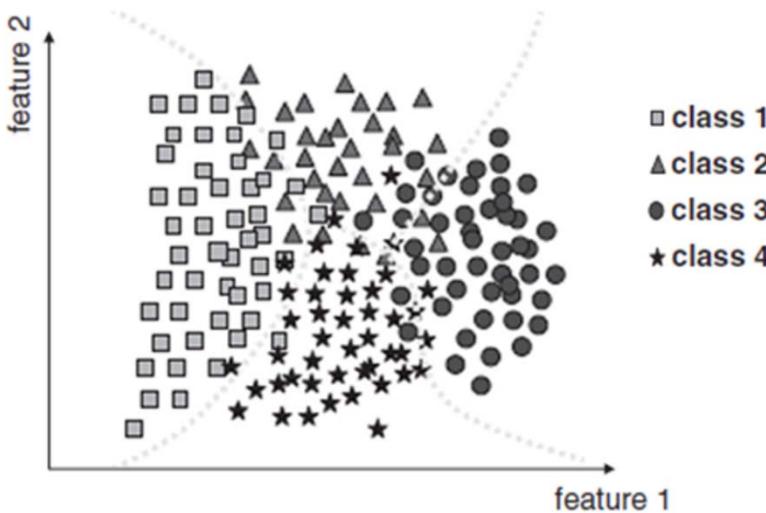
# 模式识别系统举例

- 思考：如果桃子再分为红桃子和青桃子怎么办？



# 模式识别系统举例

- 对于复杂的分类问题（多类别，多特征），难以人工设计分类器。

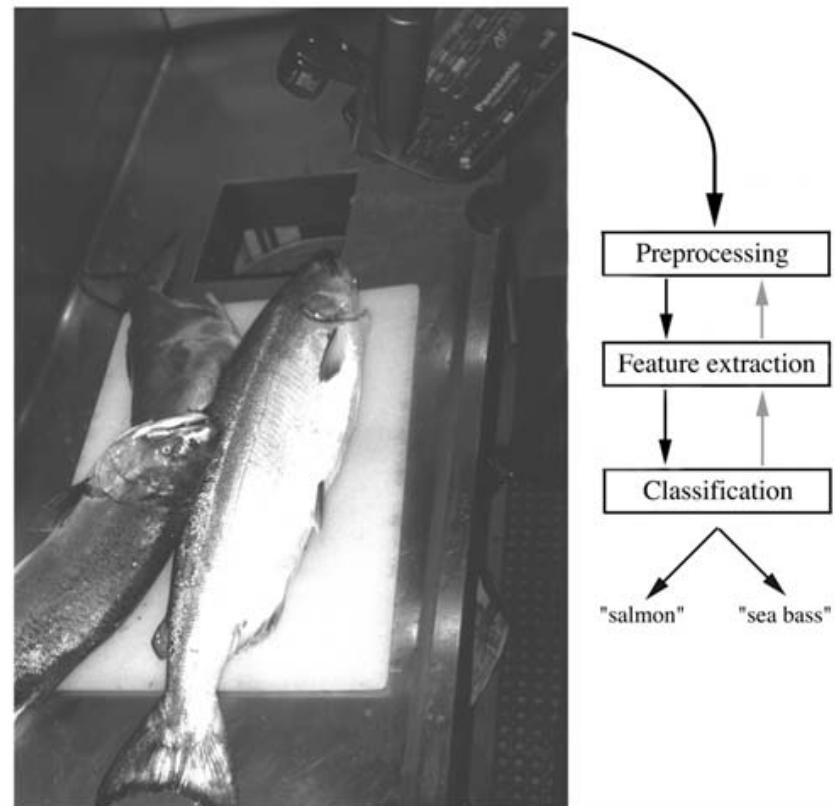


- 通常需要计算机根据训练样本集（样本有已知的类别信息）自动完成分类器的设计。
- 如何让计算机自动设计分类器，是本门课程的主要内容。

# 模式识别系统举例

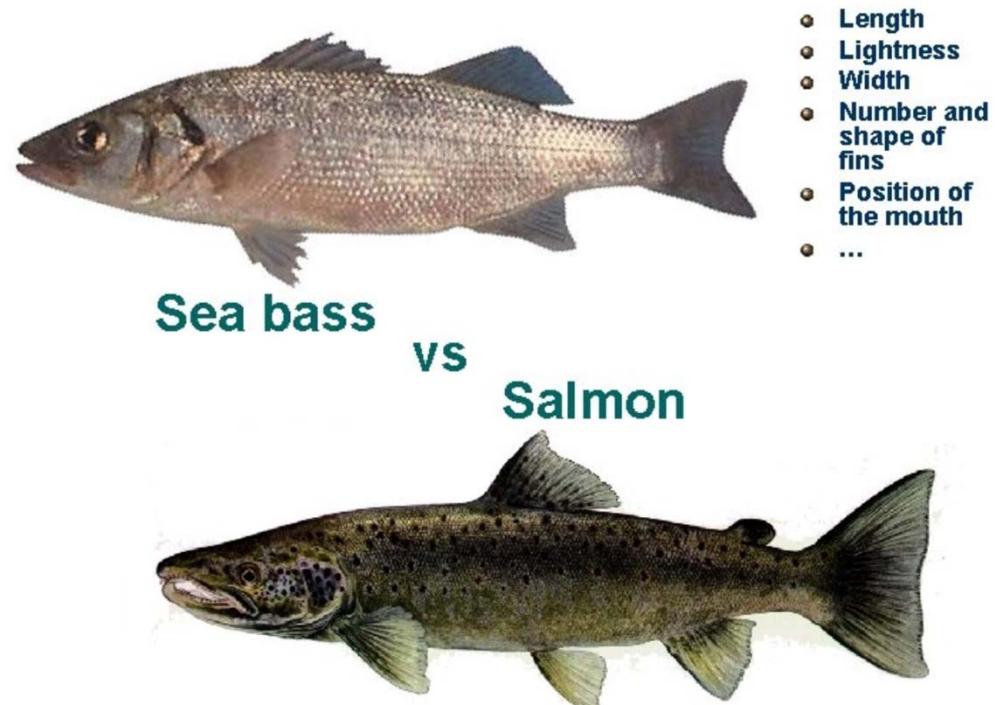
鱼类识别：（参考书《模式分类》1-5页）

- 一个鱼类加工厂需要自动分开鲑鱼和海鲈鱼。
- 数据采集和预处理类似前例。



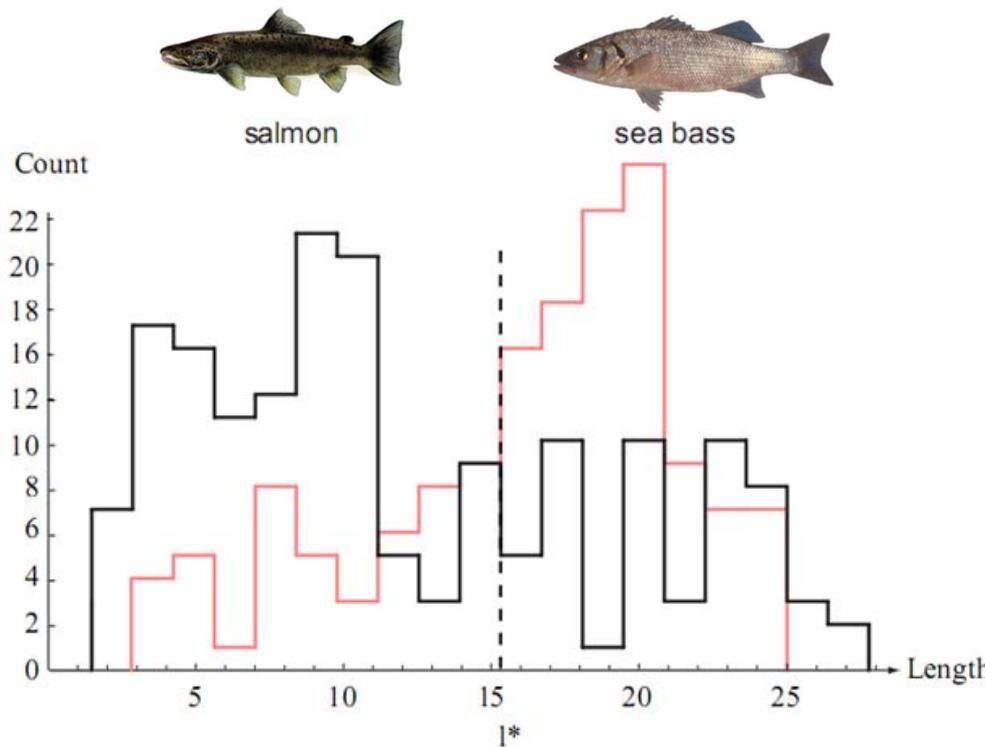
# 模式识别系统举例

- 问题：鲑鱼和海鲈鱼有什么差别？即，有什么可以区分两类鱼的特征？



# 模式识别系统举例

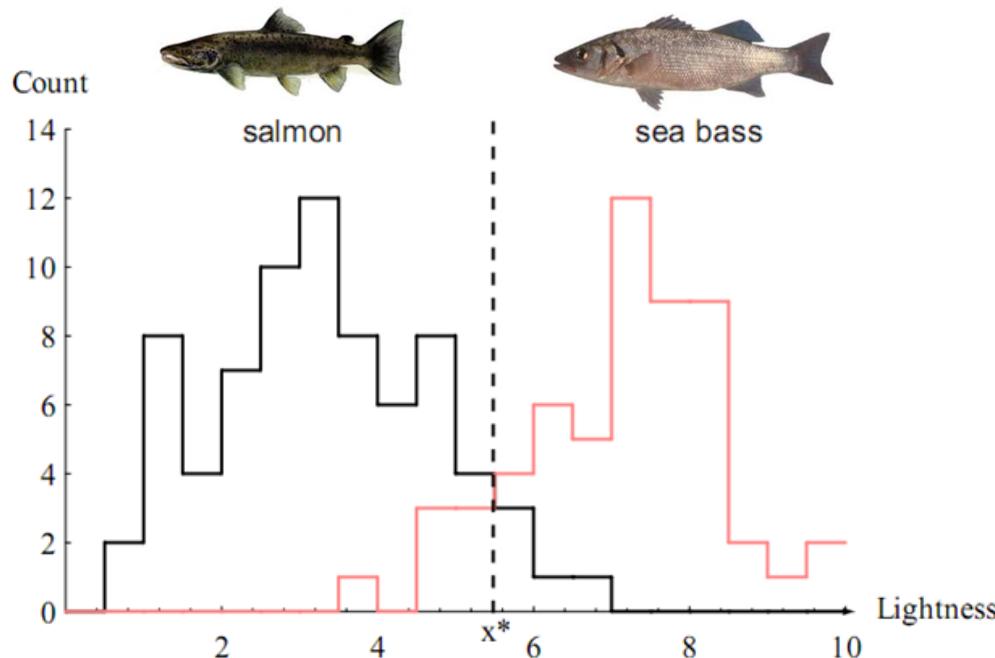
- 两种鱼长度的平均直方图



- 不存在一个长度阈值可以将两类鱼完美分开；图中虚线为最佳阈值，分类的平均误差率最小。

# 模式识别系统举例

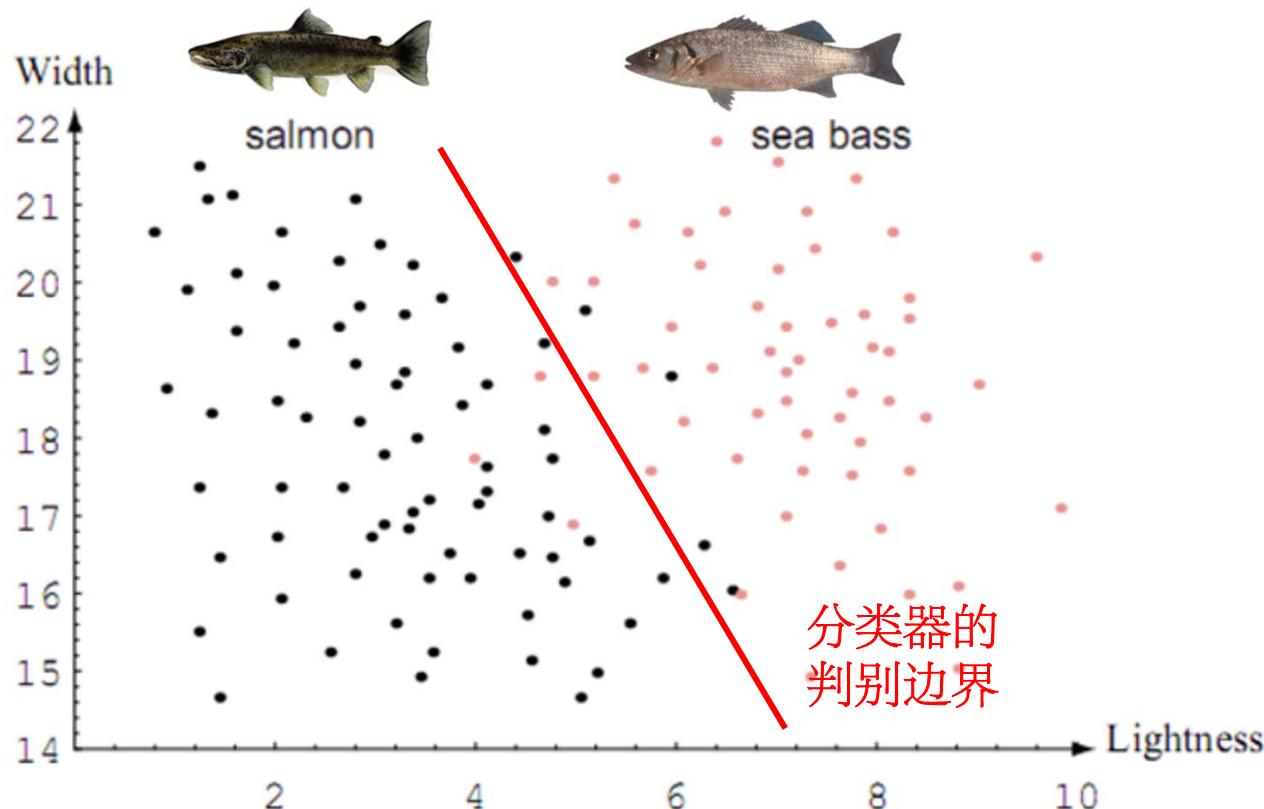
- 两种鱼光泽度的平均直方图



- 不存在一个光泽度阈值可以将两类鱼完美分开；图中虚线为最佳阈值，分类的平均误差率最小（小于用长度分类的误差）。

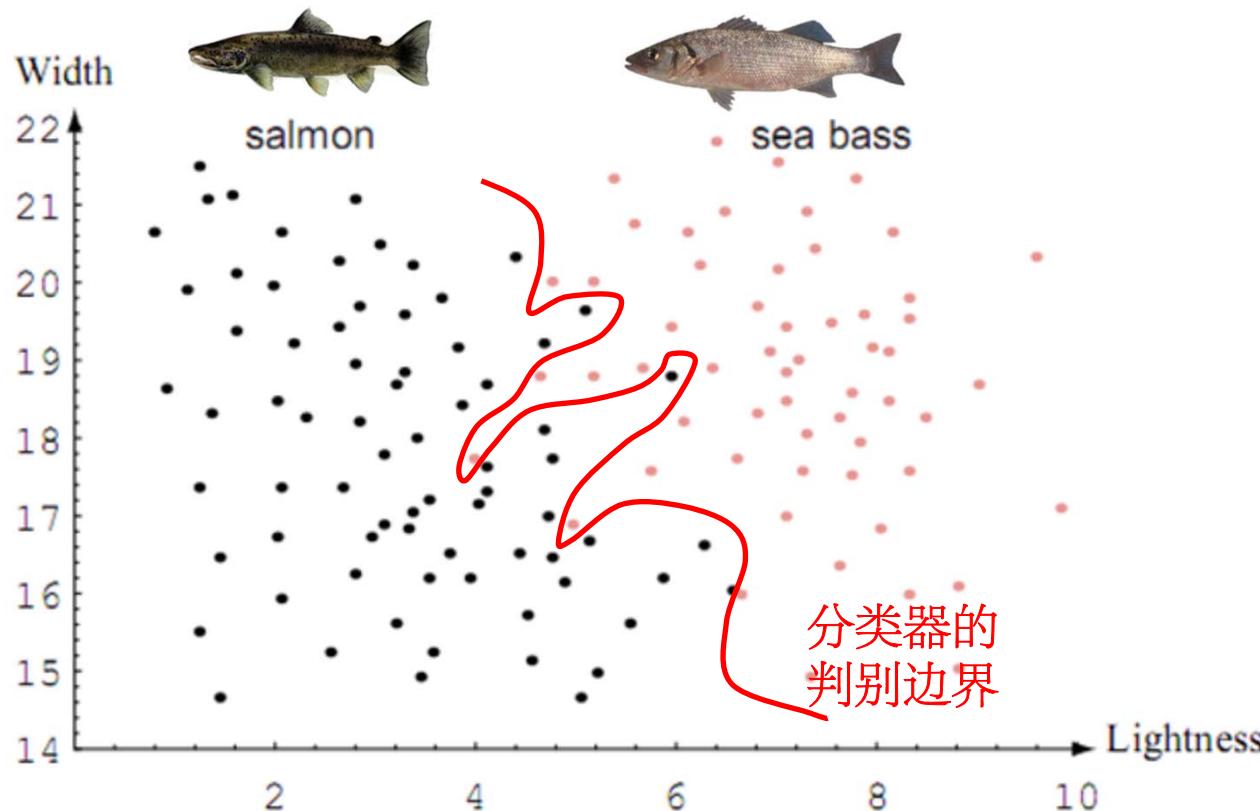
# 模式识别系统举例

- 使用光泽度和宽度（海鲈鱼一般比鲑鱼宽）两个特征可较好分类两种鱼（但仍有分类错误）。



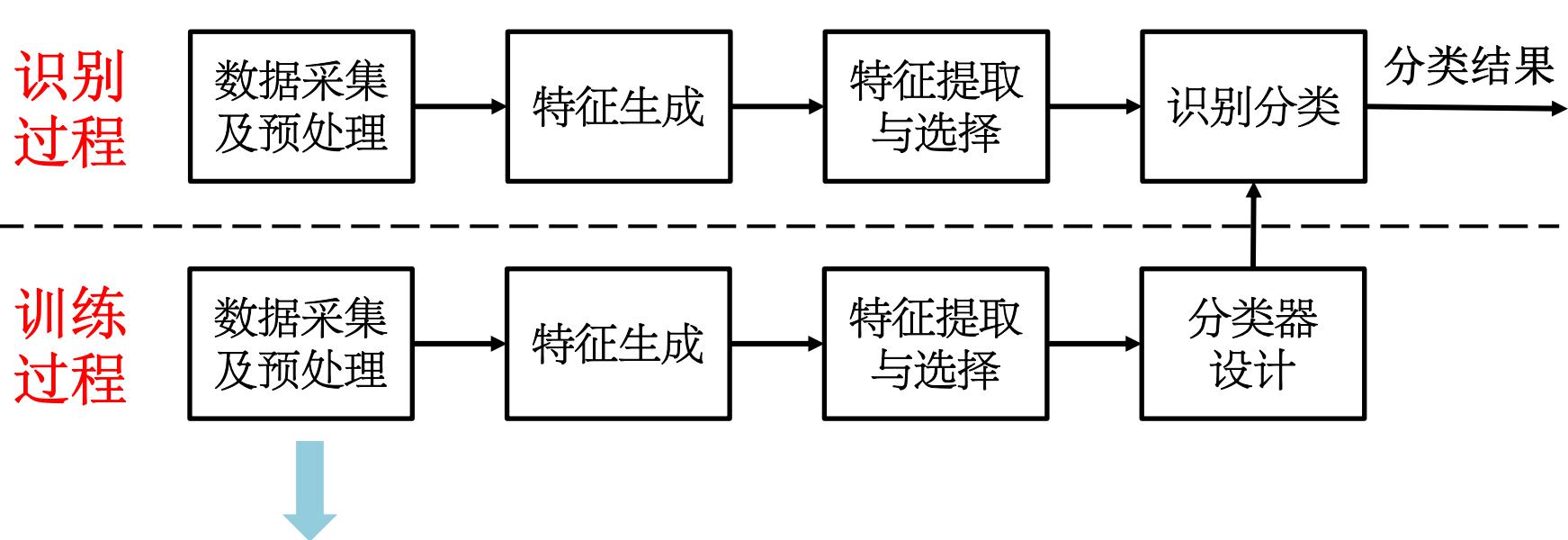
# 模式识别系统举例

- 思考：如下图中更复杂的判别边界（即更复杂的分类器）是不是一个更好的分类器？



# 模式识别系统

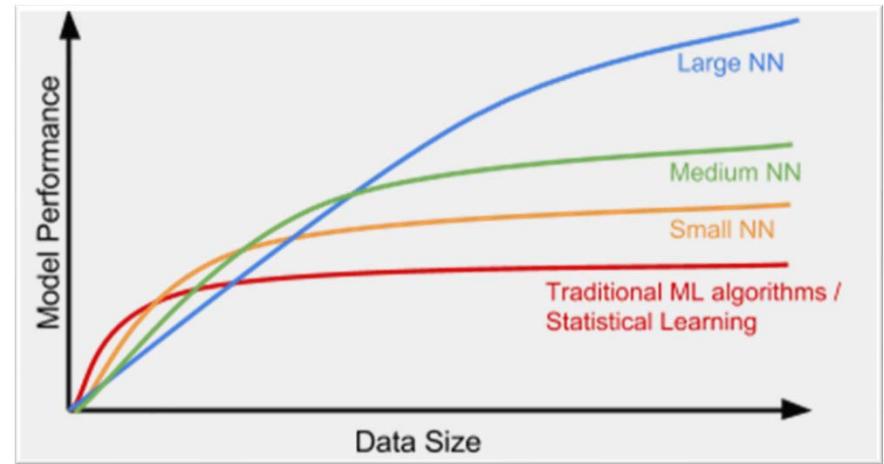
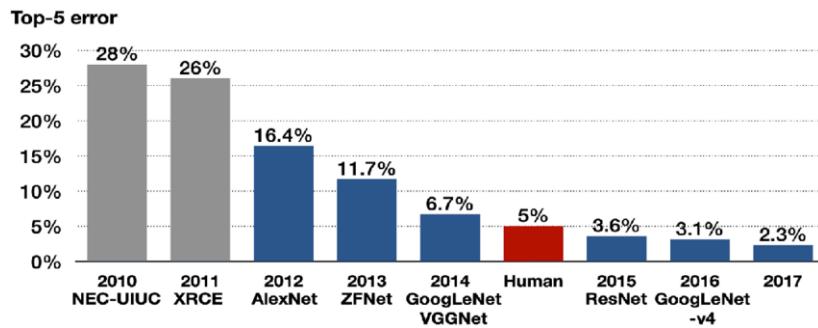
- 一个完整的模式识别系统：



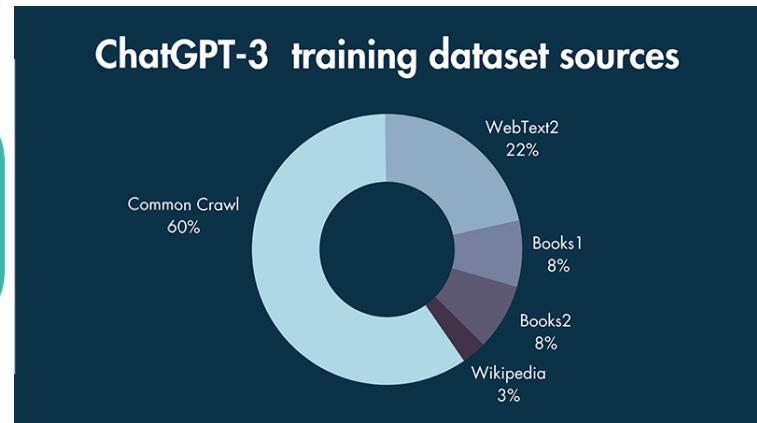
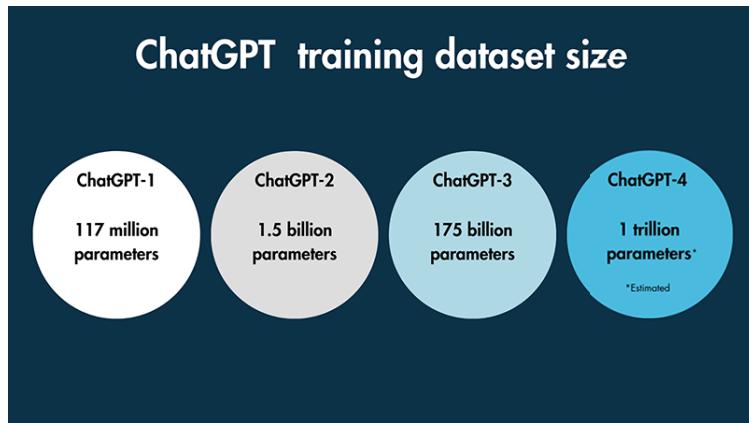
- 数据采集和样本集的构建
- 预处理（处理缺失值和离群点，数据标准化）

# 数据采集

- 模式识别是数据驱动的科学；良好的数据集是有效模式识别系统的基础。
- 优秀的数据集可以推动模型算法的高速发展。



# 数据采集



**ChatGPT-3's dataset comprised textual data from 5 sources, each with a different proportional weighting. (Source: OpenAI)**

- **60%** of ChatGPT-3's dataset was based on a filtered version of what is known as 'common crawl' data, which consists of web page data, metadata extracts and text extracts from over 8 years of web crawling.
- **22%** of ChatGPT-3's dataset came from 'WebText2', which consists of Reddit posts that have three or more upvotes.
- **16%** of ChatGPT-3's dataset come from two Internet-based book collections. These books included fiction, non-fiction and also a wide range of academic articles.
- **3%** of ChatGPT-3's dataset comes from the English-language version of Wikipedia.
- **93%** of ChatGPT-3's data set was in English.

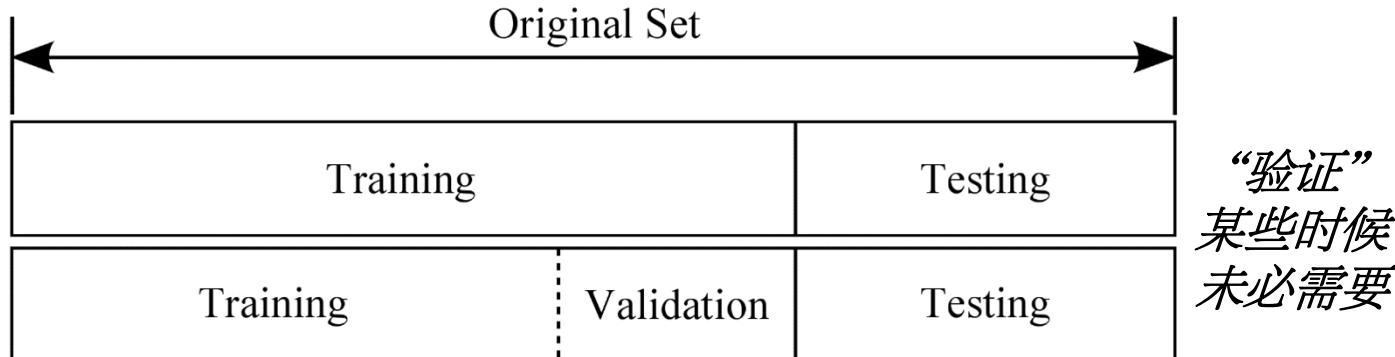
# 数据采集

---

- 数据集基本要求：噪声干扰小，标注清晰准确。
- 数据集需要保证模式识别系统具有充足、均衡、同质、且有代表性的训练和测试数据集
  - 充足：确保每个类别都存在足够多的数据来学习模型
  - 均衡：确保各个类别都有相似的数据量
  - 同质：确保不同样本有共同的属性
  - 有代表性：确保系统可能用到的测试样本中有意义的信息都可以从训练数据中得到

# 数据采集

- 数据样本集的划分
  - **训练集** ( training set ) : 训练模型的**参数** ( 模型中可被训练出来的变量 )
  - **验证集** ( validation set ) : 检验模型在新数据上的表现, 调整选择模型**超参数** ( 预先定义的模型变量 )
  - **测试集** ( test set ) : 评价模型的最终性能

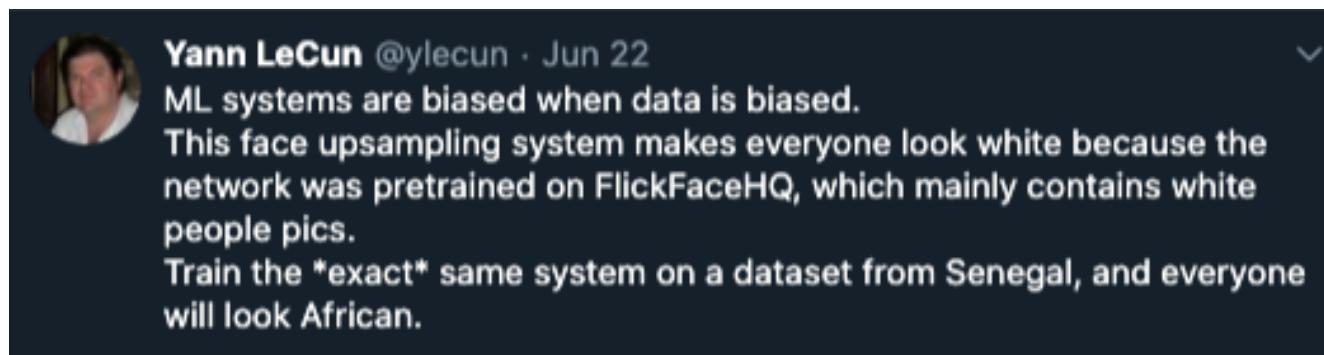


# 数据采集

- 质量不佳、有偏差的数据集会带来无用的模型甚至恶劣的后果。



2016年，MIT非裔美国研究员Joy Buolamwini在工作中发现，人们利用数据库来培训人工智能，而数据库中的大多数照片都是白人男性的照片。由此导致人工智能在识别黑人妇女或亚洲男子时的表现，要比在识别白人男性时差得多。



延伸阅读：<https://www.51cto.com/article/619830.html>

# 数据预处理

---

- 预处理：对获取的数据进行调整，尽可能消除各种来源的噪声，加强数据一致性。
- 常用的数据预处理方法
  - 缺失值处理
  - 去除噪声
  - 删删除离群点
  - 数据标准化

# 缺失值处理

---

- 数据缺失很常见（特别在人工搜集数据时）。多数算法无法处理有缺失数据，需要填补数据。
- 如果某样本缺失较多数据，或某数据缺失较多样本，可以直接删除该样本或该数据。

Examples of missing data in EHR

	<b>Gender</b>	<b>Glucose</b>	<b>AST</b>	<b>Age</b>
Patient 1	?	120	?	?
Patient 2	M	105	?	68
Patient 3	F	203	45	63
Patient 4	M	145	?	42
Patient 5	M	89	?	80

# 缺失值处理

- 缺失值补全方法：均值插补、同类均值插补、建模预测、高维映射、多重插补、极大似然估计和矩阵补全等。
- 数据补全方法应根据缺失值的类型和需要选择。

The diagram illustrates the process of imputing missing values (NaN) using the `mean()` function. It consists of two tables, each with columns labeled `Feature` and `col1`, `col2`, `col3`, `col4`, and `col5`. The first table, labeled "Sample", contains three rows of data. The second table, also labeled "Sample", shows the result after applying the `mean()` function to each column. A blue arrow points from the first table to the second, indicating the transformation.

	Feature				
	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.6	NaN
1	9.0	NaN	9.0	0.0	7.0
2	19.0	17.0	NaN	9.0	NaN

`mean()`

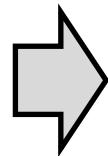
→

	Feature				
	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

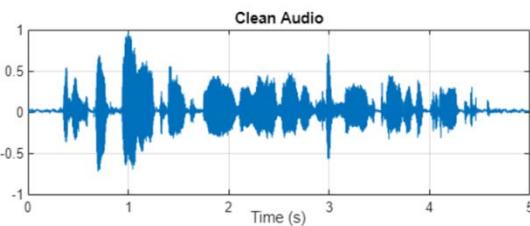
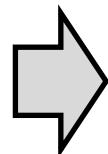
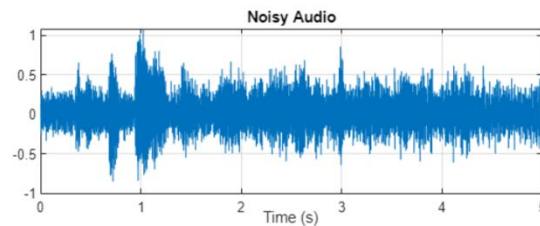
# 去除噪声

- 噪声是观测值和真实值之间的随机误差；不同数据类型的噪声种类不同，需要用特定的方式进行去噪。

图像去噪

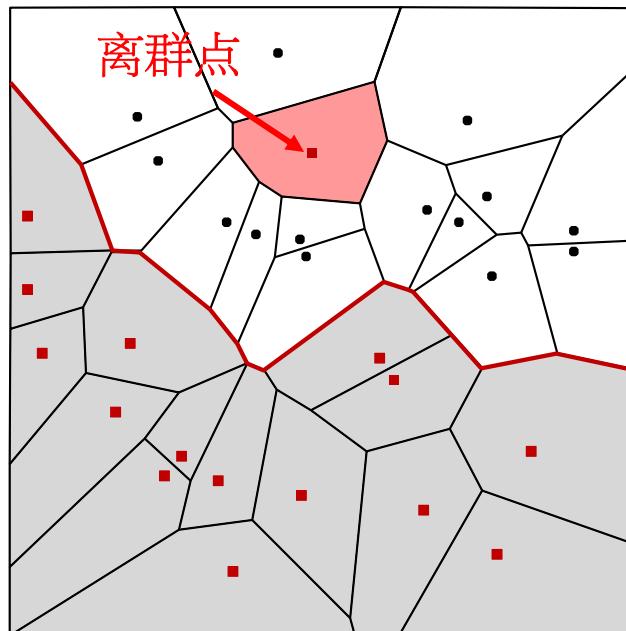


音频去噪



# 删除离群点

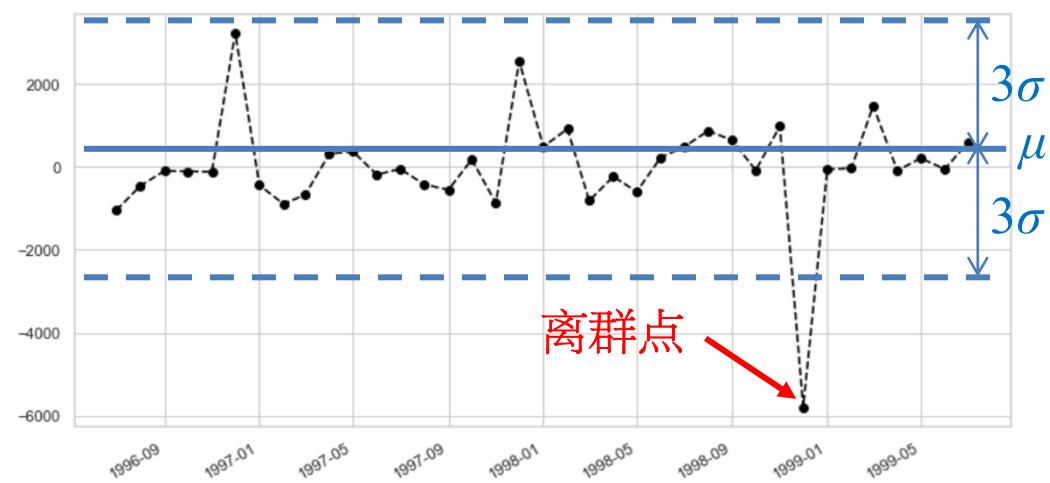
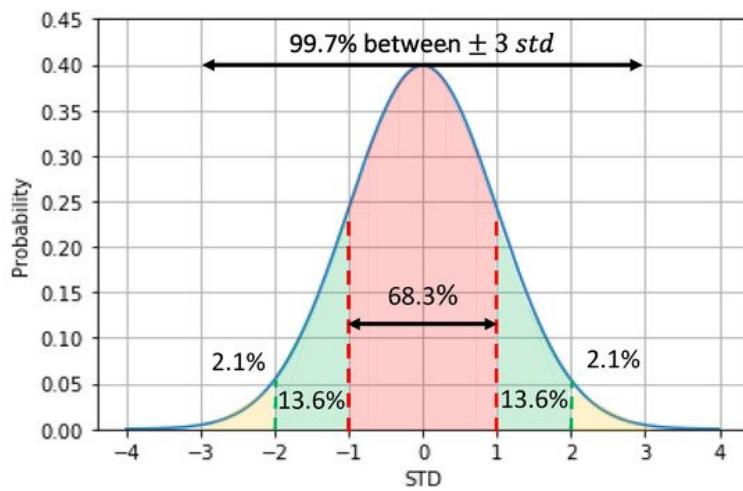
- 离群点 (outlier)，又称奇异值，是与绝大多数其他数据差别极大的少量数据。
- 离群点极大地影响正确分类器的训练。



例：最近邻方法中，一个离群点样本导致一片错误分类区域。

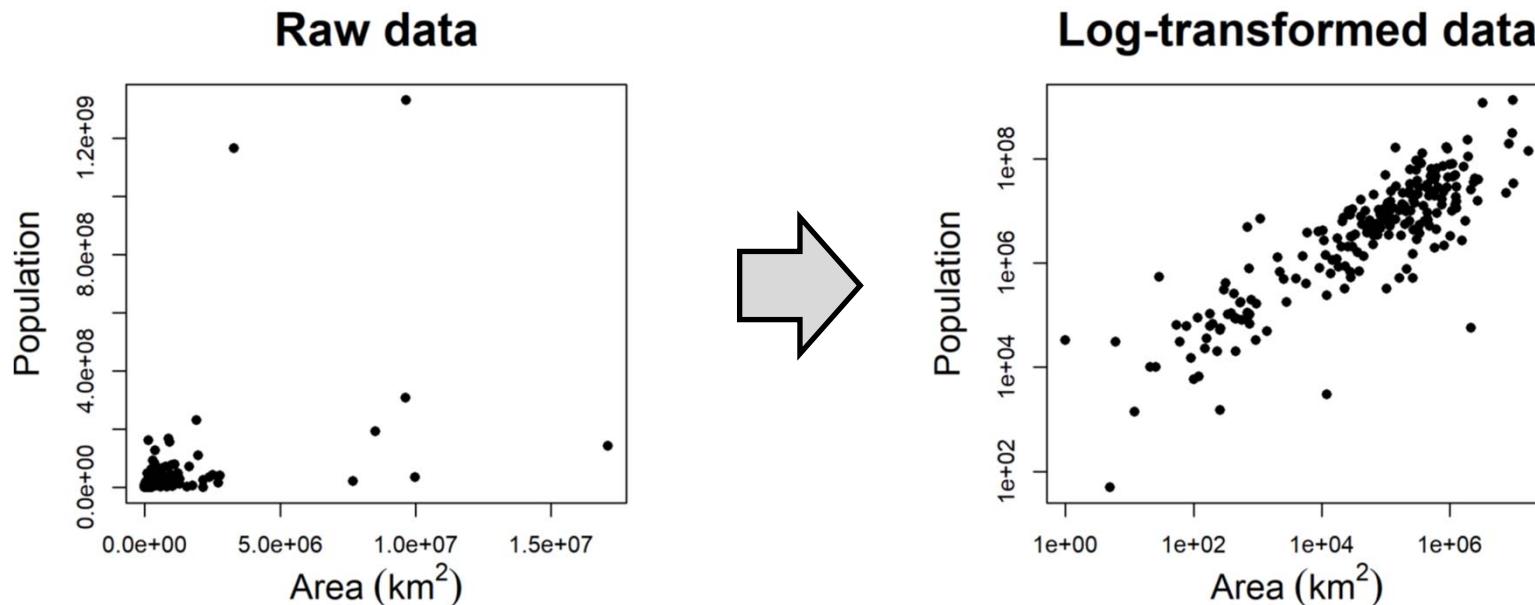
# 删除离群点

- 一般地，可以检测出离群点并删除相应的样本。
- 最常用的检测离群点方法： $3\sigma$ 法。如果样本值与均值的差值绝对值大于标准差  $\sigma$  的三倍，则可以将该样本视为离群点。



# 删除离群点

- 样本或有多类数据，但仅有一类数据含离群点，则可删除该点后填补数据，避免删除整个样本。
- 离群点未必要删除。若对数据做对数变换（可看作**数据标准化**的一种形式），有可能消除离群点且不损失信息。



# 数据标准化

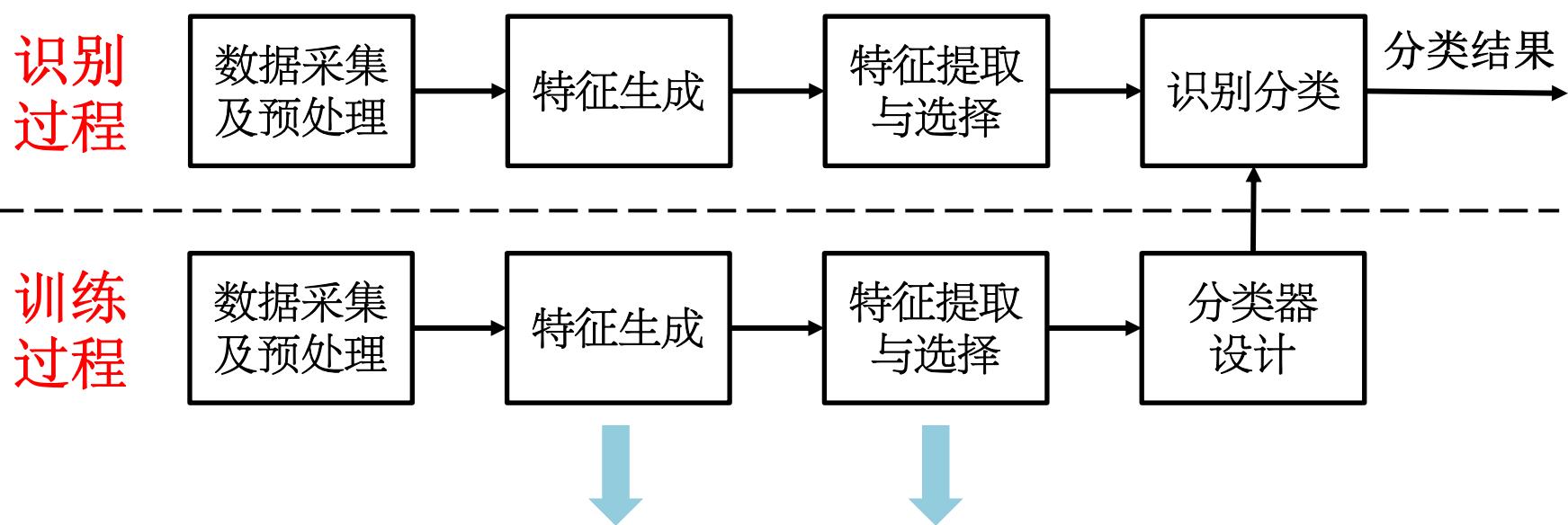
---

- 不同类型数据的取值范围变化很大，有不同的量纲；如果不对数据进行标准化，某些分类器无法正常工作。
- 均匀缩放（min-max规范化/归一化）：**假设每一维数据都服从均匀分布，将每一维数据平移和缩放到  $[0, 1]$  内。
- 高斯缩放（标准化/正态化/Z-score）：**假设每一维数据都符合高斯分布，将每一维数据平移和缩放为标准高斯分布。

见第2讲：“样本规格化”

# 模式识别系统

- 一个完整的模式识别系统：



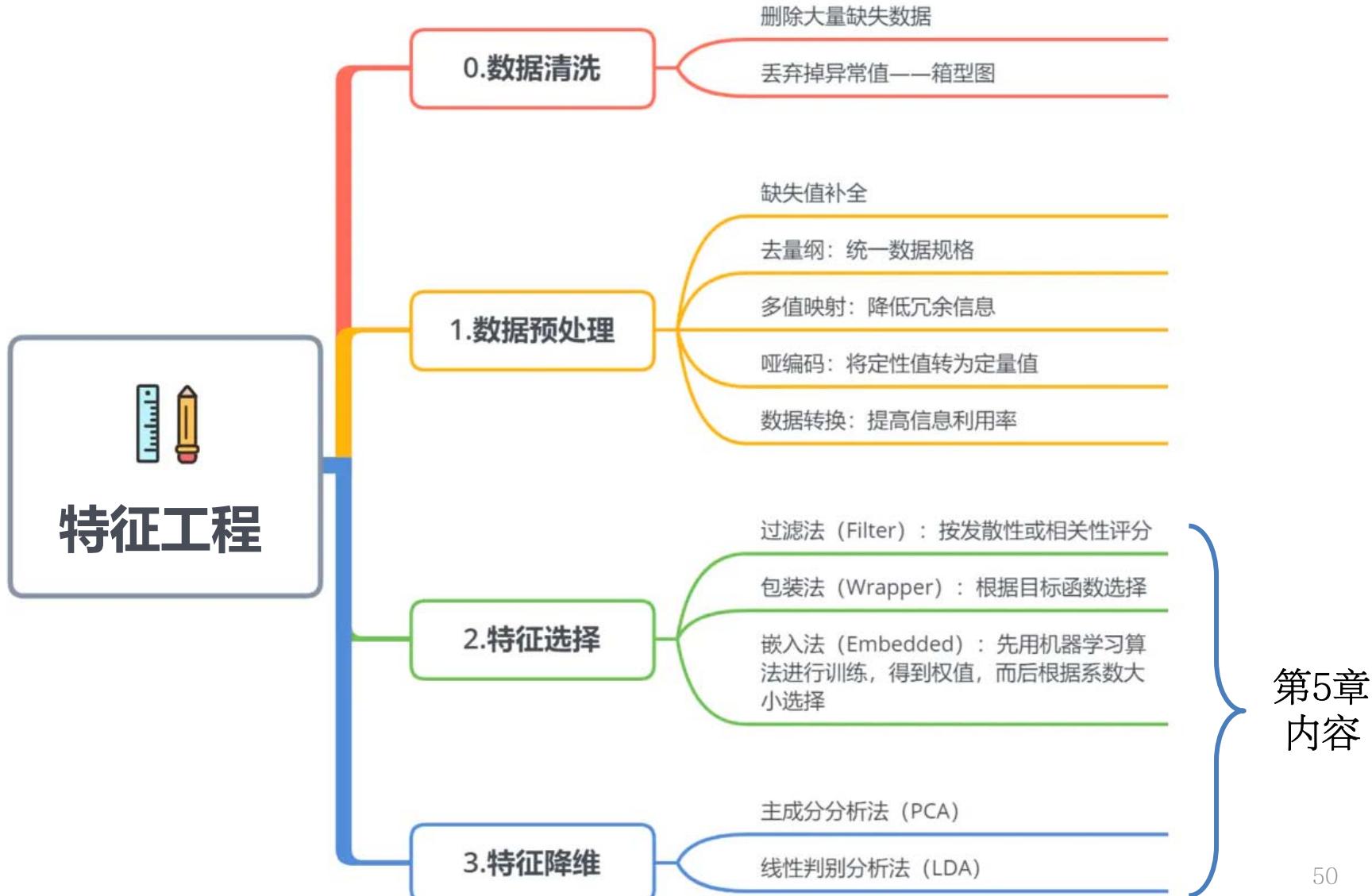
– 特征工程简介

# 特征工程

- **特征工程**: 将原始数据转换为（可以更好地表示预测模型潜在问题的）特征的过程，目的是提高模型对未知数据的预测准确性。
- **特征生成**: 从原始数据中可以描述不同类别对象之间差异的易计算测量的属性。
- **特征提取和选择**: 选择最具有判别以及预测能力的特征集合。

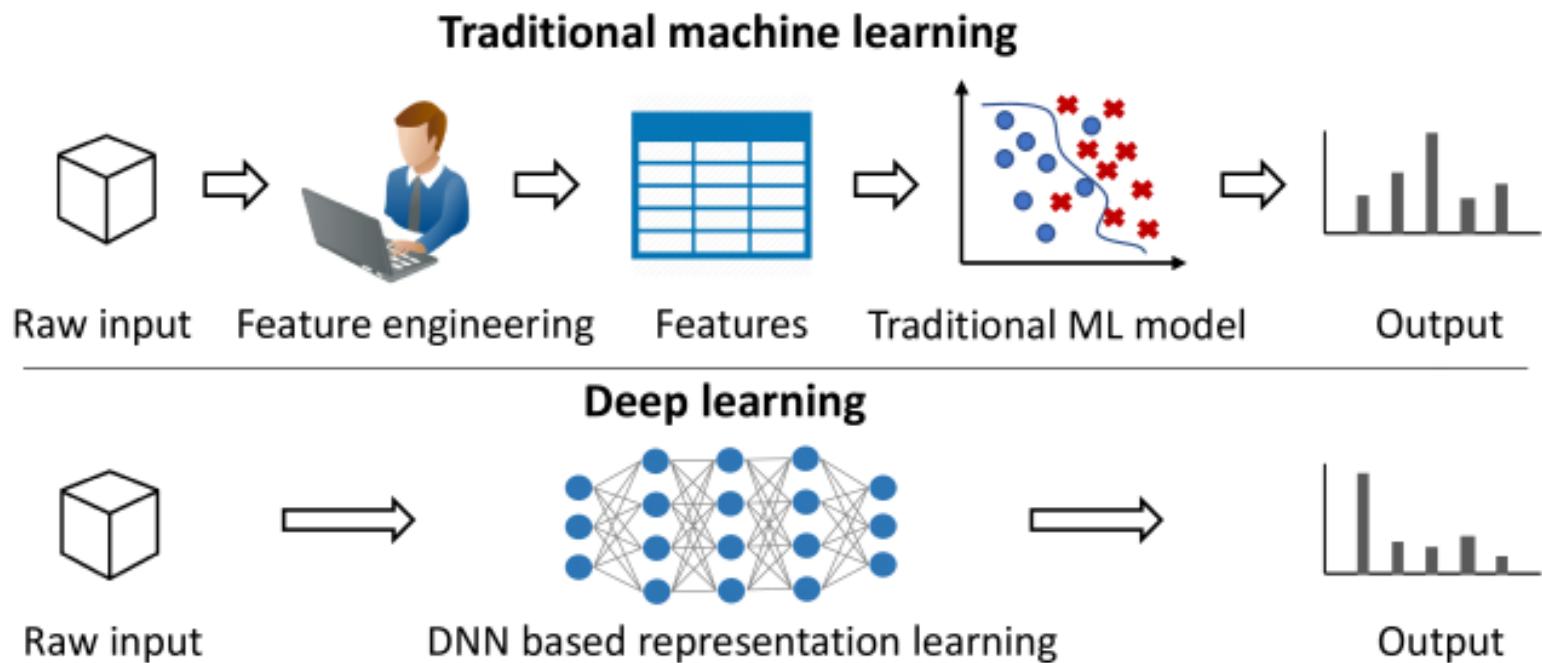
数据和特征决定了模式识别的上限，  
而模型和算法只是逼近这个上限而已。

# 特征工程



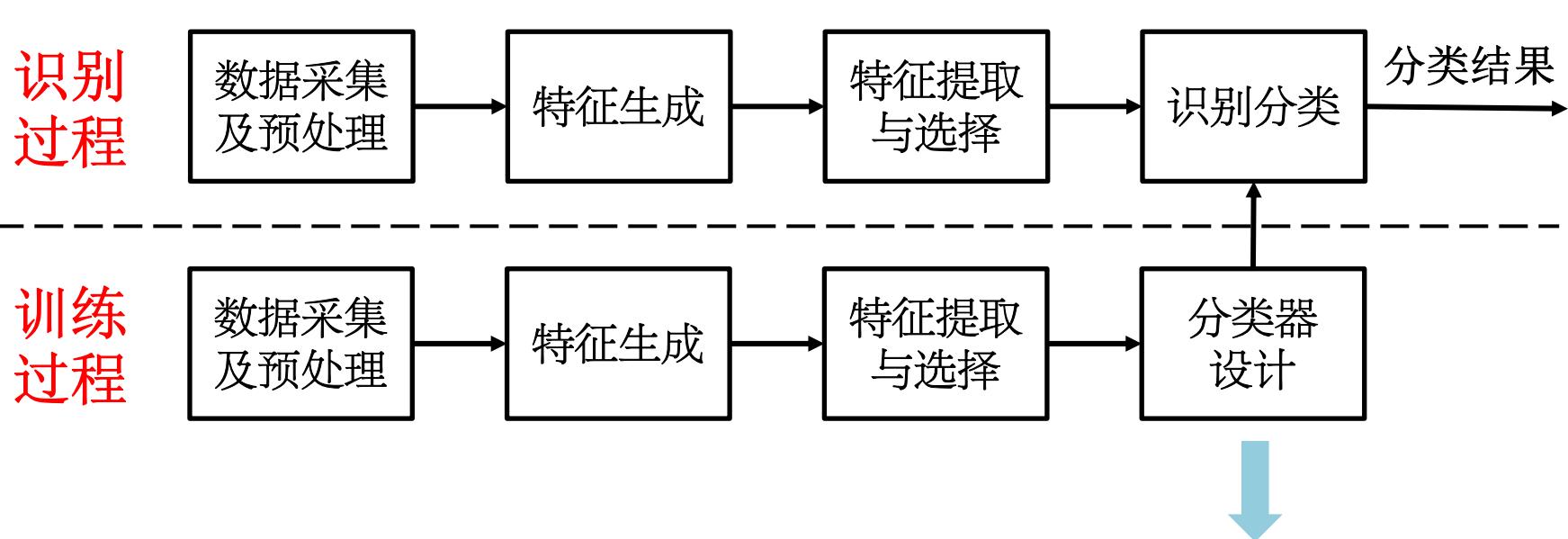
# 特征工程

- 特征工程：适用于数据量少但有较多先验知识和领域知识的应用，模型可解释性和可调节性好。
- 深度学习：适用于数据量大但先验特征知识少的应用（“黑盒”，“炼丹”……）。



# 模式识别系统

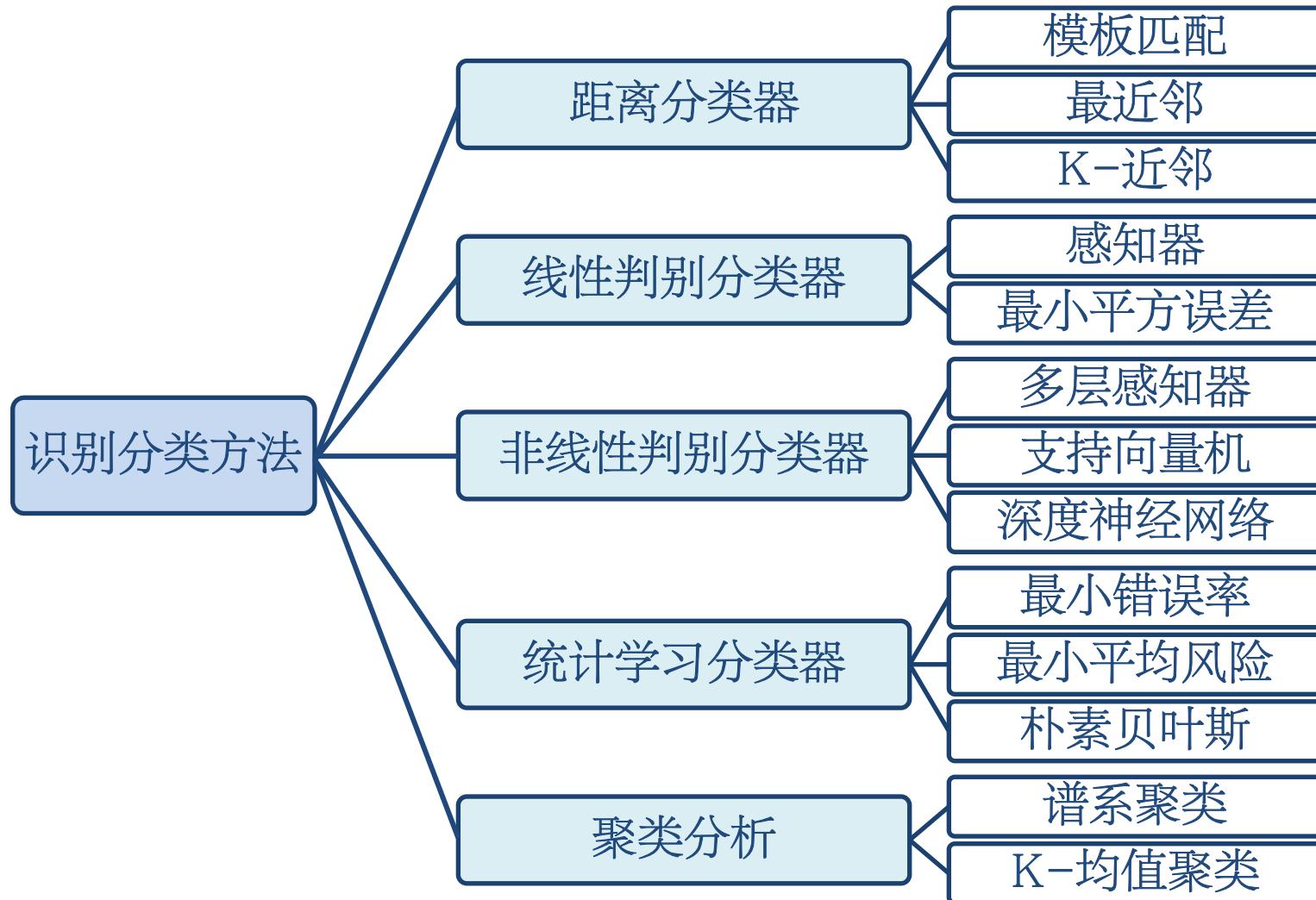
- 一个完整的模式识别系统：



– 模式识别方法分类

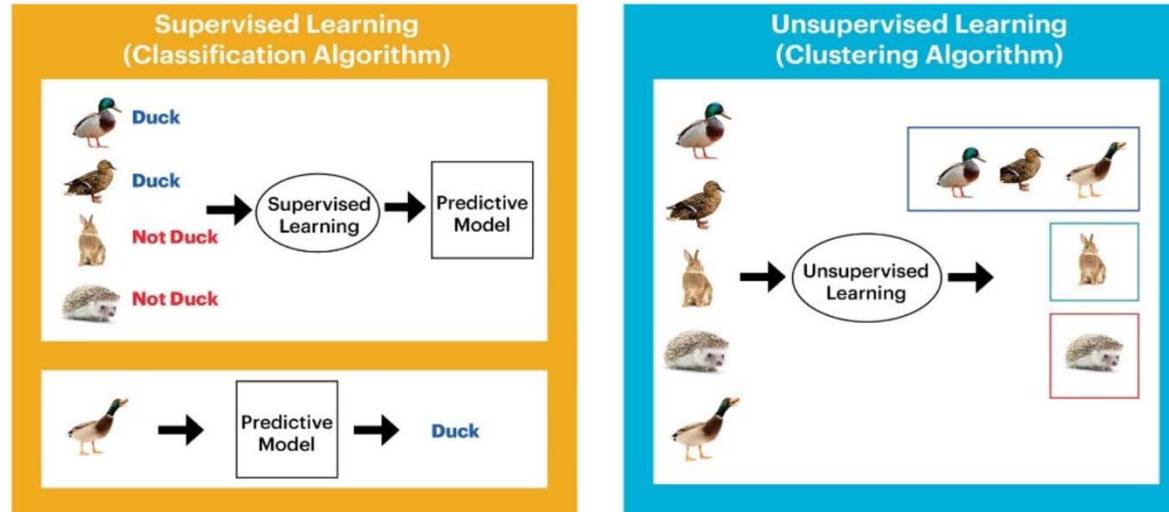
# 识别分类

- 本课程最主要的内容是介绍各种识别分类方法。



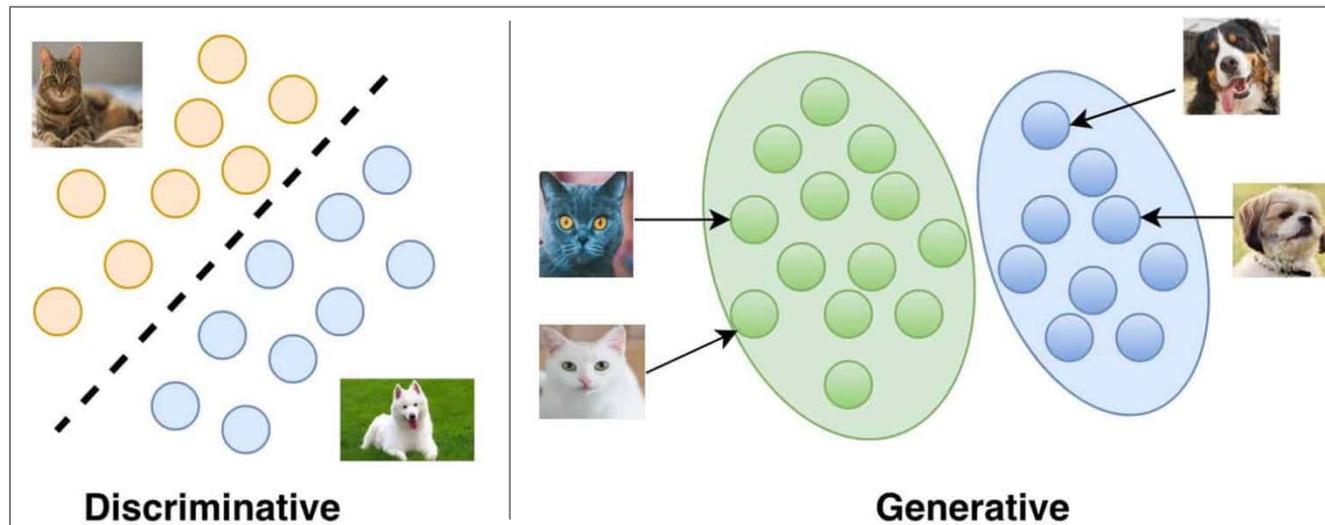
# 识别分类方法的类别

- 根据训练样本有无类别标号，模式识别方法分为有监督学习（分类）和无监督学习（聚类）。
- 有监督（supervised）学习：**预先已知训练样本集中每个样本的类别标号。
- 无监督（unsupervised）学习：**预先不知道训练集中样本的类别标号，甚至不知道类别的数量。



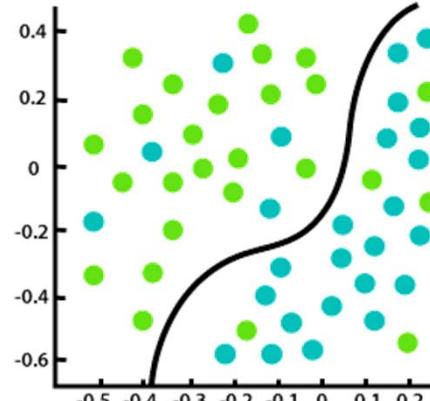
# 识别分类方法的类别

- 监督学习可以分为鉴别模型（判别模型）和产生式模型（生成模型）两大类。
- **鉴别模型（discriminative）**：样本模式确定地处在特征空间不同区域，通过训练得到类别边界。
- **产生式模型（generative）**：样本模式是特征空间的随机变量，估计概率密度以确定类别属性。

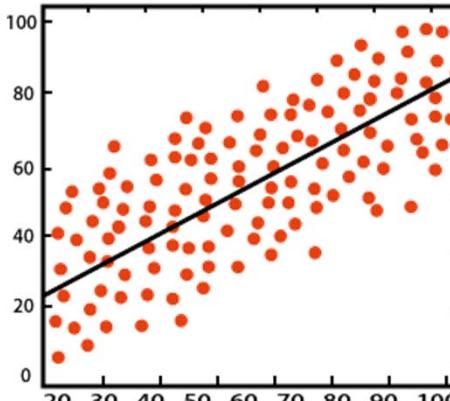


# 识别分类方法的类别

- 根据类别标号的性质，监督学习可以分为分类和回归两大类。
- 分类 (classification)**：类别标号是有限的离散值，模型寻找最优决策边界。
- 回归 (regression)**：类别标号出是连续值，模型寻找对样本的最优拟合。



Classification



Regression

# 多分类

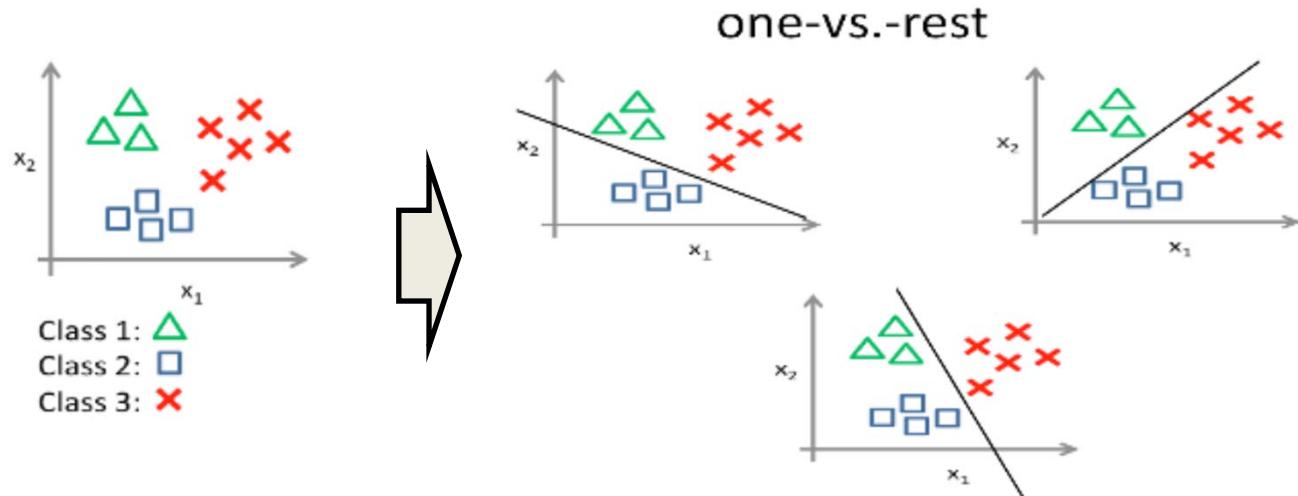
---

- 多分类：将样本分为  $L$  类，其中  $L > 2$
- 虽然一些分类器允许使用两个以上的类，但多数分类器为二分类器（输出两个类别）。
- 通常来说，多分类是利用常用的二分类器通过不同的策略来实现的：
  - 一对多 (One vs. Rest)
  - 一对一 (One vs. One)

# 多分类

- 一对多 (One vs. Rest)

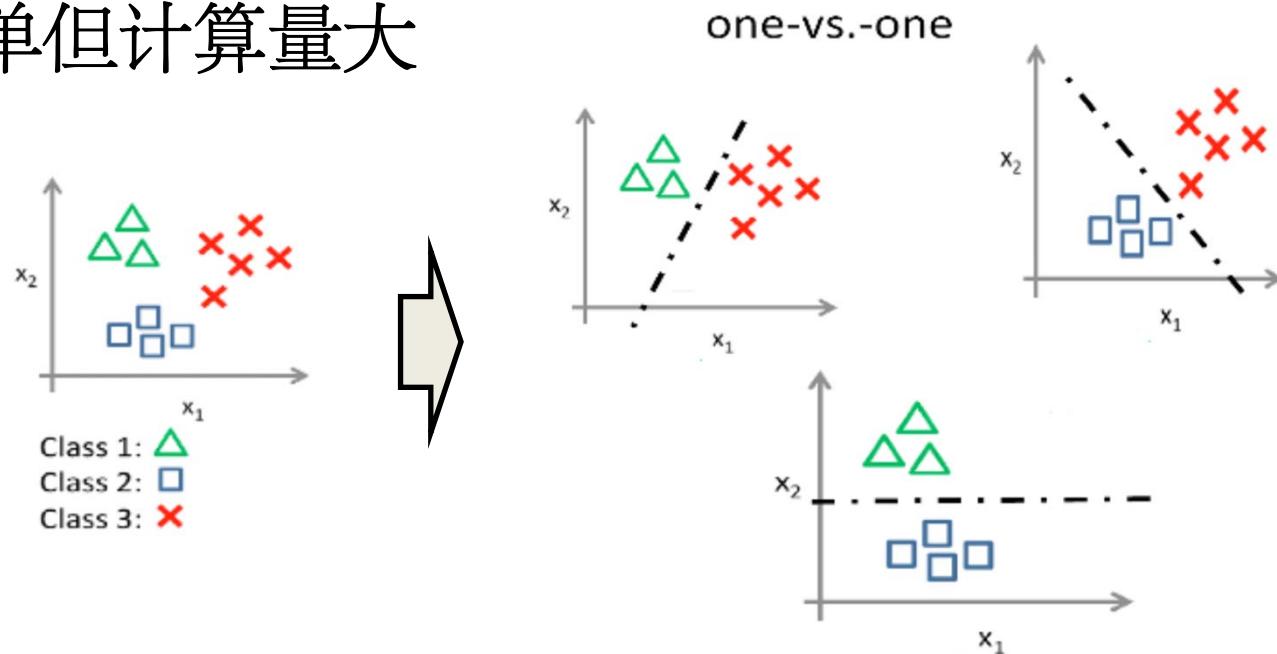
- 对每一个类，训练一个分类器将该类与其它类分开，因此共有  $L$  个分类器
- 返回获得最高置信度或得分（如朴素贝叶斯分类器的最大后验概率）的分类器得到的类别
- 比一对一策略的计算量少，但样本不均衡



# 多分类

- 一对 (One vs. One)

- 给类别两两配对，对每一对类分别训练一个二分类器，因此共有  $L(L - 1) / 2$  个二分类器
- 根据投票把预测最多的类别作为分类结果
- 简单但计算量大



# 本章小结

---

- 介绍了模式识别的一系列基本概念
- 介绍了模式识别系统的组成
- 介绍了两个模式识别应用实例（水果/鱼类识别）
- 介绍了构建数据集的原则
- 介绍了常见的预处理方法，主要是如何处理缺失值和离群点，如何标准化数据
- 介绍了模式识别方法的分类（有监督和无监督，鉴别和产生式模型，分类和回归）