机器学习基础

主讲：王星
单位：中国人民大学统计学院
助教：陈志豪
电话：86-10-82500167
上课时间：周三上午
上课地点：0513
Email: wangxingwisdom@126.com
办公地点: 明德主楼 1019

个人简介

王星，中国人民大学统计学院教授，经济学博士。国家社科基金重点项目负责人，美国加州伯克利大学、密歇根大学、卡内基梅隆大学、台湾中央研究院和辅仁大学访问学者。教育部统计学重点学科研究基地项目负责人，教育部学位论文抽评匿名评审专家，国内外发表论文40余篇。代表作：《大数据分析：方法与应用》，《非参数统计》，《统计学习导论：基于R的应用》，《数据挖掘与商务分析：R语言》，《数据挖掘--客户关系管理的科学与艺术》，《人文社会科学文献网络知识模型与应用》，第十届全国统计科学教材二等奖，全国应用统计案例大赛一等奖指导教师。主要研究领域为复杂数据分析和统计建模，长期从事R语言、Python、SAS、SPSS等分析工具的教学与研发，复杂数据分析咨询，全国应用统计案例培训讲师，长期从事大数据分析项目分析和培训。

教材和参考文献

1.Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, An Introduction to Statistical Learning with Applications in R

2.周志华，机器学习，清华大学出版社，2016,01

3.Sebastian Raschka，Python机器学习，机械工业出版社，2017,03

4.赵卫东，董亮，机器学习基础，人民邮电出版社，2018.

5. Christopher bishop, PRML,2007

6.T. M. Mitchell, Machine Learning, McGraw Hill, 1997

7.R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000

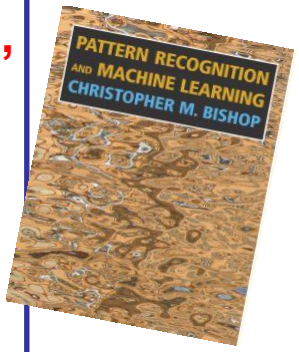
8. Hastie, Tibshirani and Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. The second Edition.

9.P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005

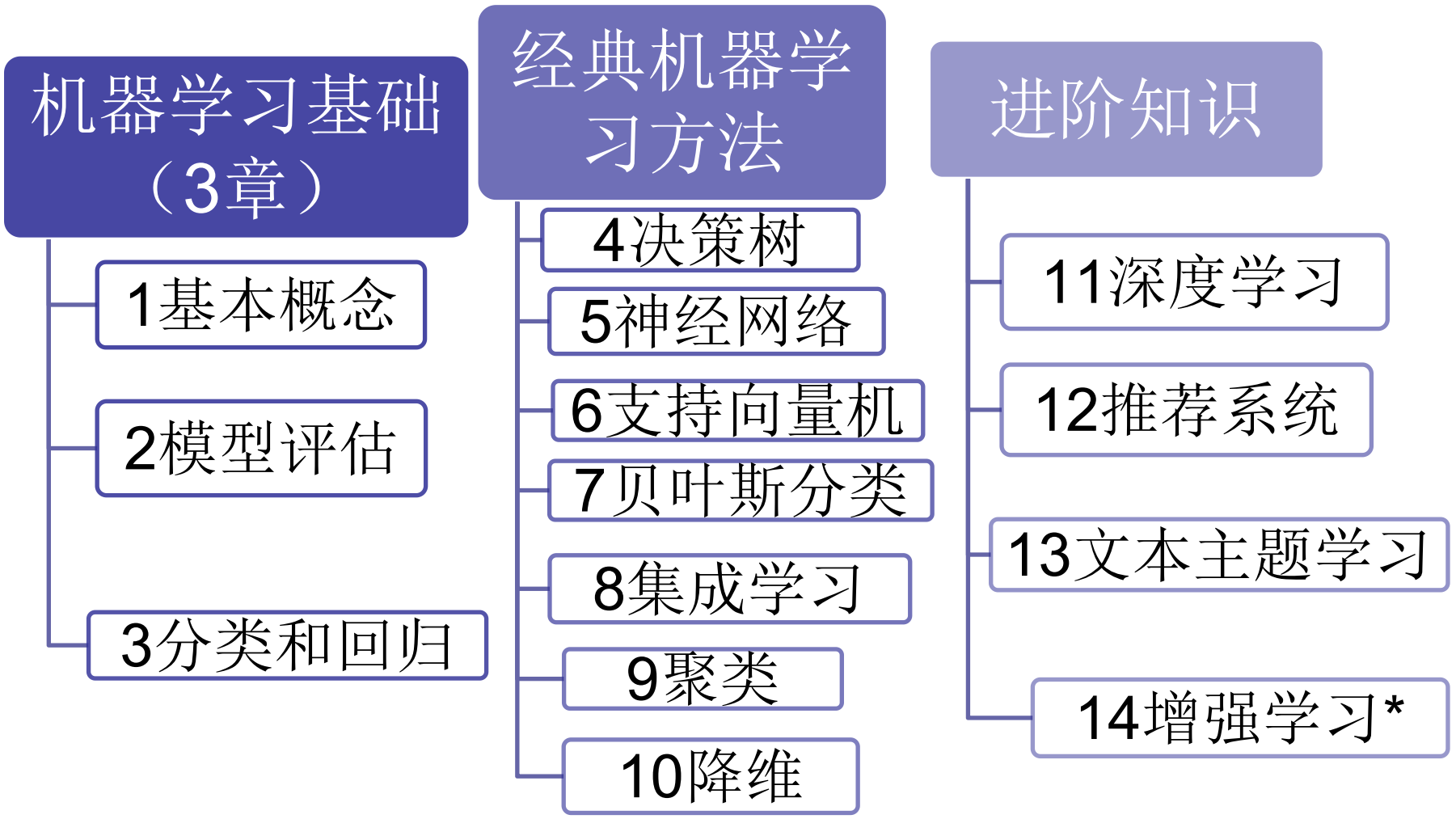
10.Han Jiawei, Data mining Concepts and techniques, 机械工业出版社。

11.王星，大数据分析：方法与应用，清华大学出版社，201309。

12.李航，统计学习方法，清华大学出版社，2012,03



目录大纲 16章


机器学习基础（3章）
1基本概念
2模型评估
3分类和回归
经典机器学习方法
4决策树
5神经网络
6支持向量机
7贝叶斯分类
8集成学习
9聚类
10降维
进阶知识
11深度学习
12推荐系统
13文本主题学习
14增强学习*

要 求

- 每周都有上机作业，做好实验的准备
- 个人习题，期中报告和期末考试（开卷）；
- 成绩占比：个人作业20%，期中报告20%，期末成绩 60% .
- 迟交作业将在成绩上有惩罚.
- 有问题吗?

作业范例：Jupyter Notebook

https://nbviewer.org/github/JWarmenhoven/ISL-python/blob/master/Notebooks/Chapter%203.ipynb

https://github.com/JWarmenhoven/ISLR-python

https://github.com/shilpa9a/Introduction_to_statistical_learning_summary_python/blob/master/notebook/Chapter_2_Statistical_Learning.ipynb

jupyter nbviewer

ISL-python / Notebooks

Chapter 3 - Linear Regression

- Load Datasets
- 3.1 Simple Linear Regression
- 3.2 Multiple Linear Regression
- 3.3 Other Considerations in the Regression Model

In [39]:
```
# %load ../standard_import.txt
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import axes3d
import seaborn as sns

from sklearn.preprocessing import scale
import sklearn.linear_model as skl_lm
from sklearn.metrics import mean_squared_error, r2_score
import statsmodels.api as sm
import statsmodels.formula.api as smf

%matplotlib inline
plt.style.use('seaborn-white')
```

Load Datasets

Datasets available on https://www.statlearning.com/resources-first-edition

In [2]:
```
advertising = pd.read_csv('Data/Advertising.csv', usecols=[1,2,3,4])
advertising.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
TV          200 non-null float64
```

Figure 3.1 - Least squares fit

In [7]:
```
sns.regplot(advertising.TV, advertising.Sales, order=1, ci=None, scatter_kws={'color':'r', 's':9})
plt.xlim(-10,310)
plt.ylim(ymin=0);
```

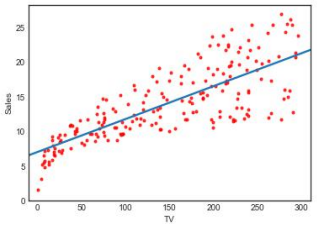


Figure 3.2 - Regression coefficients - RSS

Note that the text in the book describes the coefficients based on uncentered data, whereas the plot shows the model based on centered data. The latter is visually more appealing for explaining the concept of a minimum RSS. I think that, in order not to confuse the reader, the values on the axis of the B0 coefficients have been changed to correspond with the text. The axes on the plots below are unaltered.

In [8]:
```
# Regression coefficients (Ordinary Least Squares)
regr = skl_lm.LinearRegression()

X = scale(advertising.TV, with_mean=True, with_std=False).reshape(-1,1)
y = advertising.Sales

regr.fit(X,y)
print(regr.intercept_)
print(regr.coef_)
```

```
14.0225
[ 0.04753664]
```

课程目标

- 掌握机器学习、数据挖掘和知识发现里的基本概念、算法和模型: 包括统计概念,数据可视化,分类,回归,聚类,关联、降维等主题.
- 熟悉机器学习的产生与发展
- 机器学习基本流程
- 会使用Python语言对数据做机器学习算法和案例研究

1.机器学习简史

1. 大数据雨数据科学
2. 数据挖掘的概念
3. 数据挖掘的产生与发展
4. 数据挖掘应用
5. 数据挖掘的基本流程
6. 数据挖掘的应用案例

一、什么是机器学习？

- 计算机中，“经验”通常是以“数据”形式储存下来，机器学习所研究的主要内容是关于设计算法，如何从观测数据中不断学习和总结经验，从数据中产生可计算的“模型”（model）帮助计算机做出准确判断的自动化技术，这种技术称为算法，也称为“learning algorithm”.
- 两个注释：模型是算法的结果，通过经验提升自身的性能。

机器学习

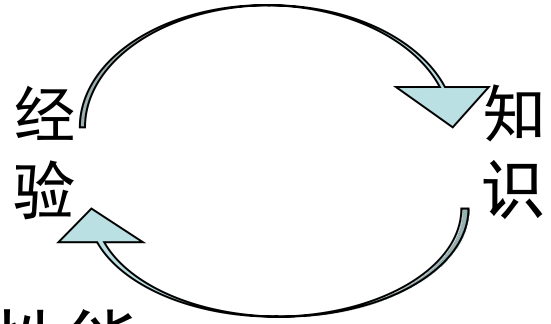
定义：如果一个计算机程序针对某个任务T，用P作为性能的衡量标准，根据经验E自我完善，就称这个计算机程序是从经验E中学习任务T，衡量性能为P。

- 例如：

  T：识别和分类图象中的手写文字。

  P：分类的正确率；

  E：已知分类的手写文字数据库。



– 利用经验自动改善计算机系统性能

the computer systems that automatically improv
experience

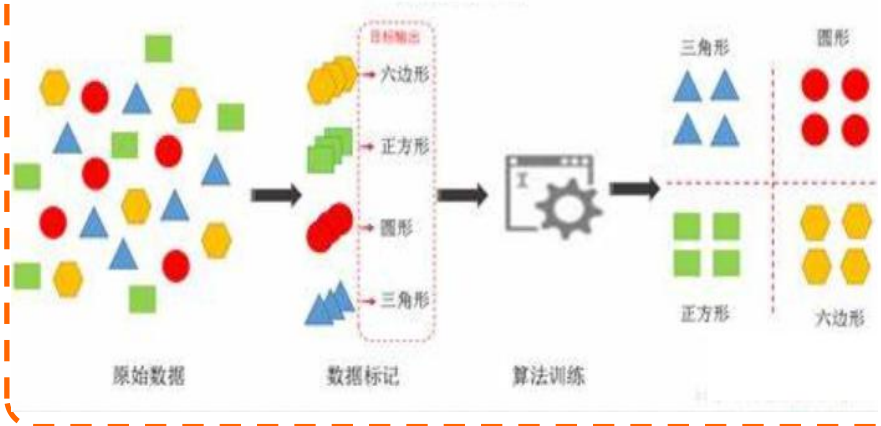
- 把握存在于学习之旅中的基本规律

the fundamental laws that govern all learning process.

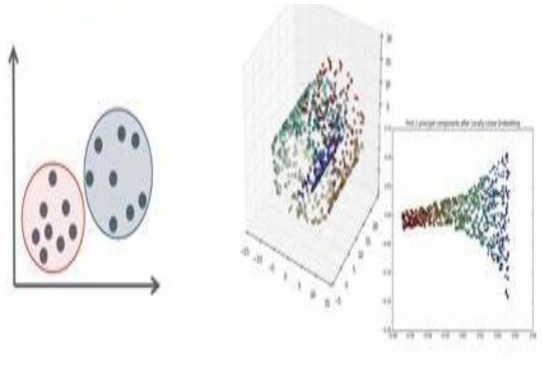
----------Tom Mitchell June2006

学习的分类

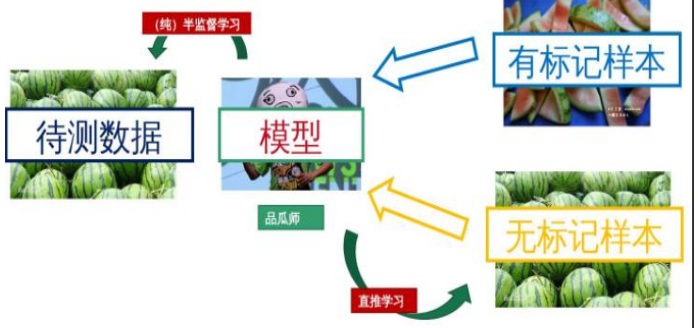
有指导学习



无指导学习

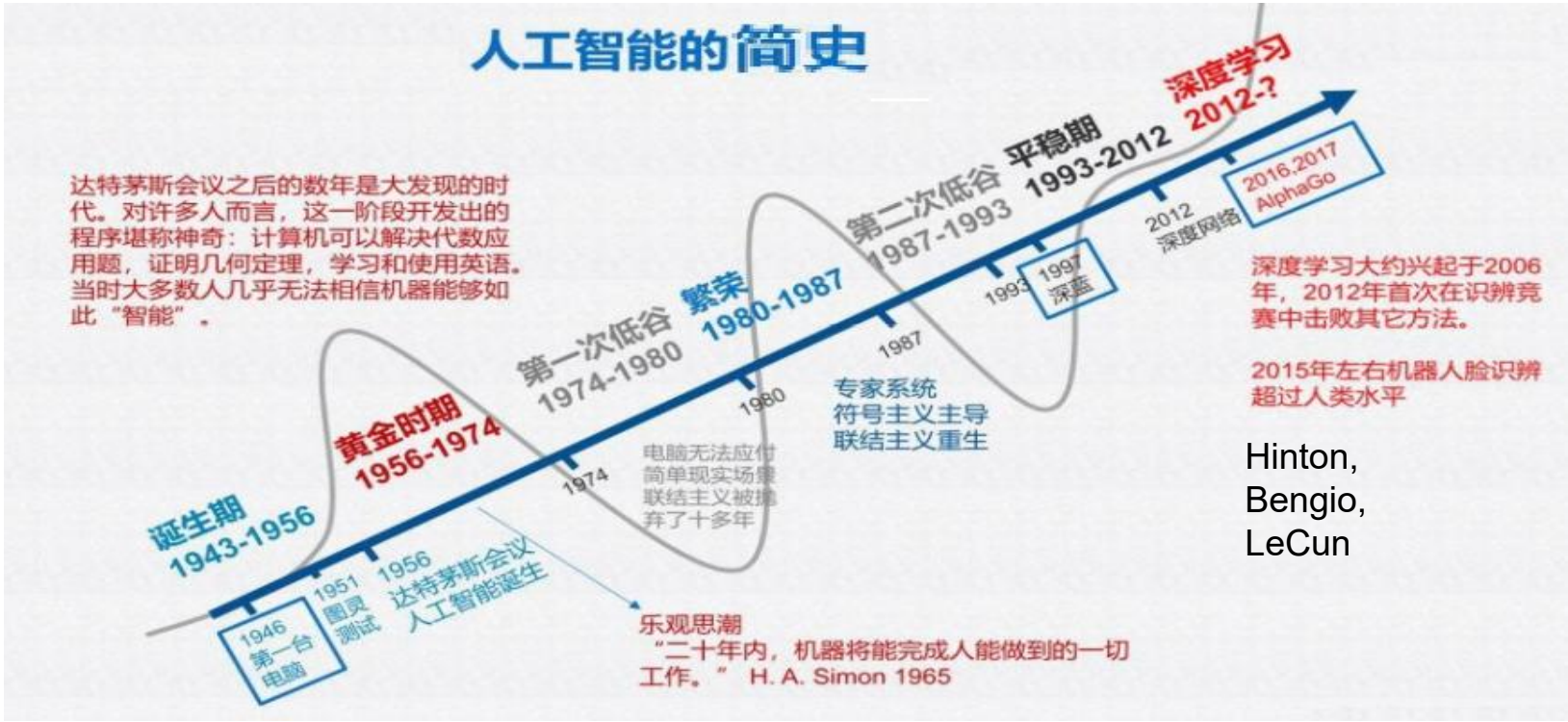


半指导学习

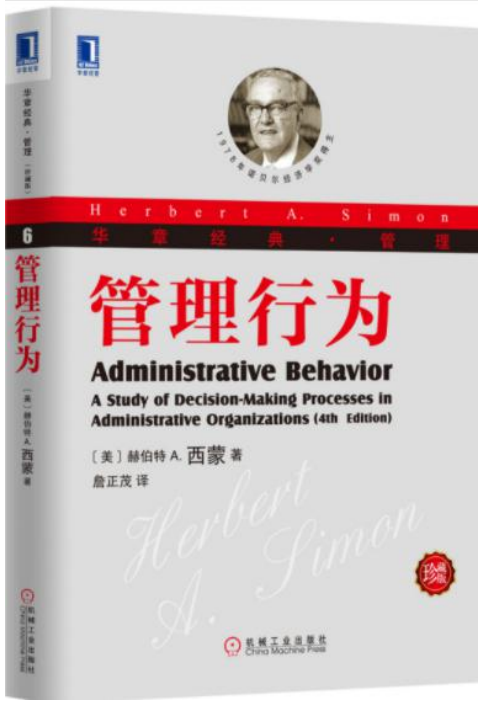


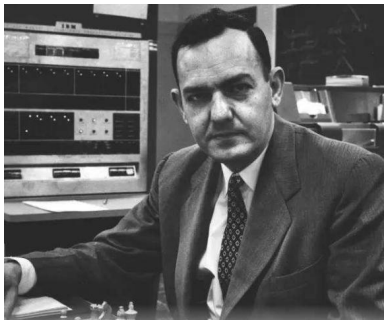
增强学习

人工智能简史



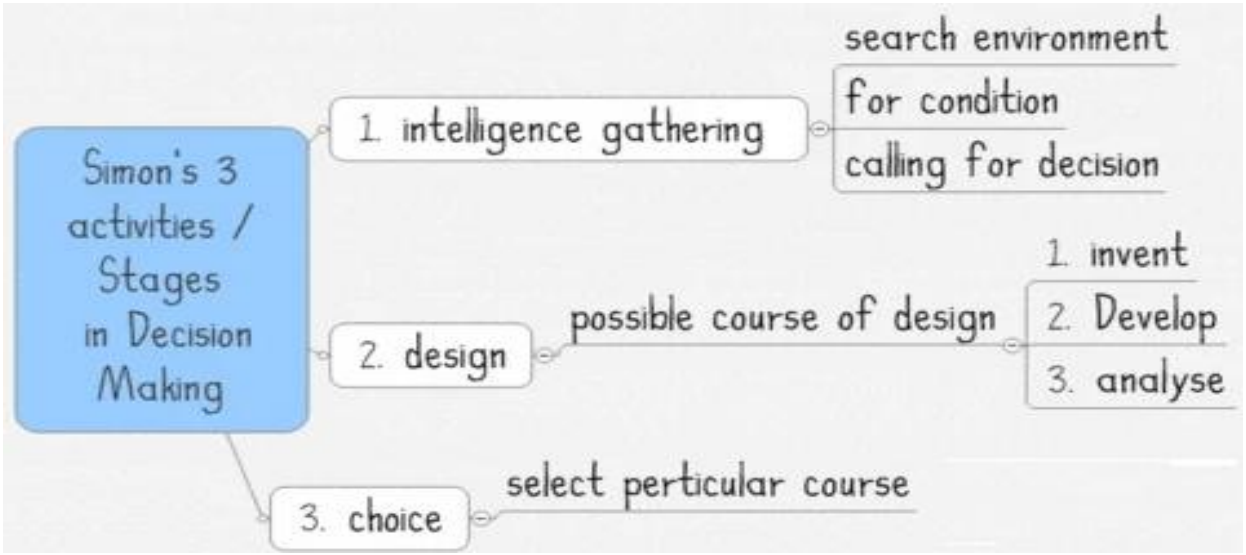
知识推理期	知识工程期	浅层学习	深度学习
1936自动机模型：图灵测试；1943MP模型； 1951符号演算（冯诺依曼）； 1956人工智能（约	1958LISP； 1943感知器收敛理论； 1951通用问题求解GPS； 1975框架知识表示	1984决策树CART 1986决策树ID3算法； 1988Boosting算法； 1993C4.5； 1995 ADABoost,SVM 2001 RF	2006深度信念网络； 2012谷歌大脑（吴恩达）； 2014生成对抗网络(Ian Goodfellow)





希尔伯特·西蒙，人工智能奠基人；
符号主义学派创始人；
1975年图灵奖获得者；
1978年用1年时间里用计算机、管理学和心理学杂交出来了一个理论获得诺贝尔经济学奖。

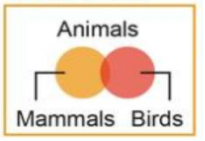
1976年，现在的世界与过去显著不同，**基于过去情况推测未来需要持谨慎态度**。现在的世界是一个信息及其丰富的世界。**信息并不稀缺**。

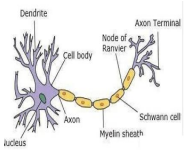


机器学习的五大流派
推荐看林宙辰短片--机器学习简史

五大流派	起源	算法	核心思想
符号主义 代表人物：Tom Mitchell、Steve Muggleton、Ross Quinlan	逻辑学、哲学	逆演绎算法 （Inverse deduction）	1980年代，知识是信息的符号的表示，符号输入计算机进行模拟推理。从而实现人工智能，知识工程知识图谱
联结主义	神经科学	反向传播算法 （Backpropagation）	2010年早期到中期，起源于神经科学，主要算法是神经网络，神经元以一定结构组成，通过误差进行修正，2010年代末，联结主义加符号主义
进化主义 代表人物：John Koda、John Holland、Hod Lipson	进化生物学	基因编程 （Genetic programming）	进化过程是基因交叉、突变的过程，典型的产品是进化算法
贝叶斯派 代表人物：David Heckerman、Judea Pearl、Michael Jordan	统计学	概率推理 （Probabilistic inference）	1990-2000主观概率估计、发生概率修正、最优决策，自然语言的情感分类、自动驾驶和垃圾邮件过滤等问题
行为类比主义	心理学	核机器 （Kernel machines）	根据约束条件优化函数，行为类推注意者通过类比推理获得知识和理论，将未知情况与已知情况建立对应关系，新旧知识间的相似性，推荐系统

Symbolists





Evolutionaries

Bayesians

Analogizers

一个学习系统的基本架构