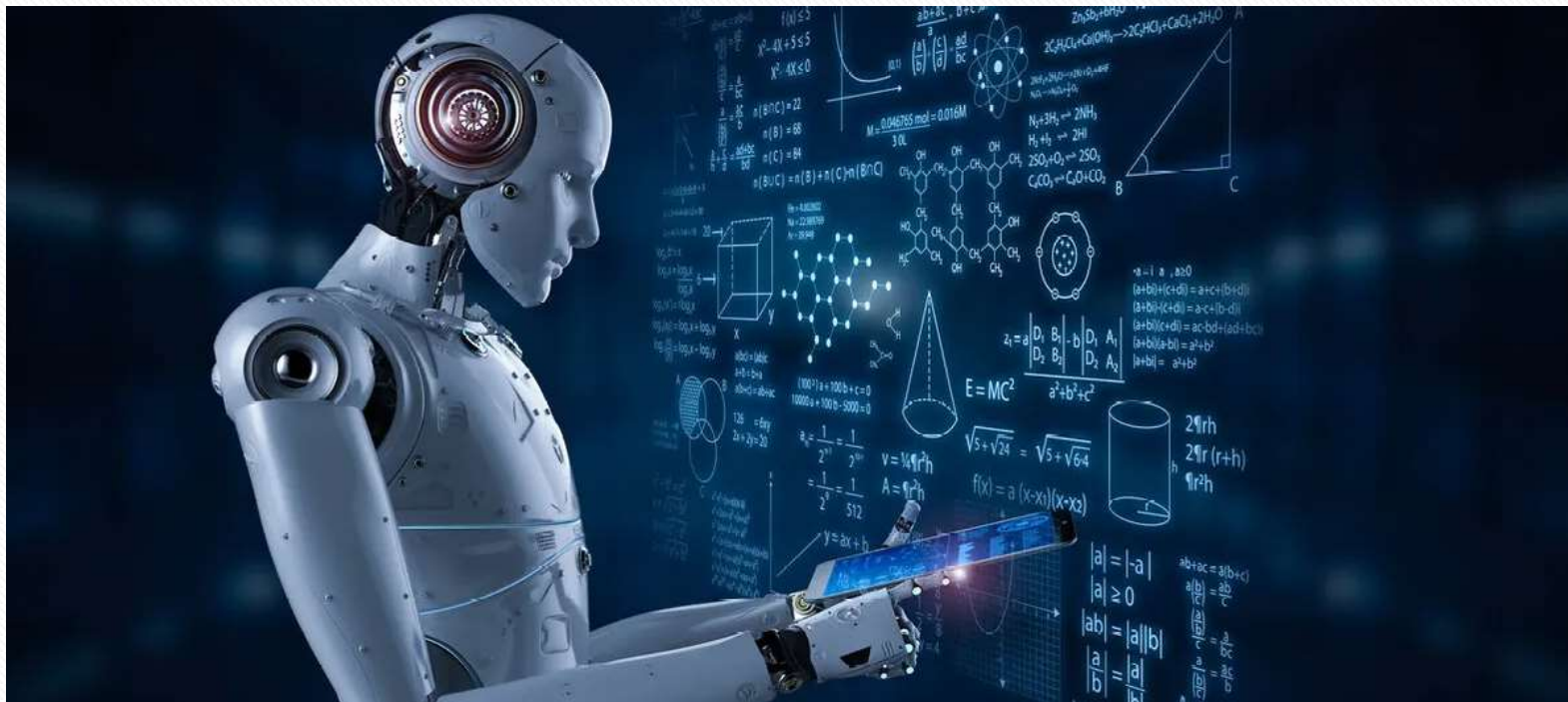


课程交流群



统计机器学习



教学参考信息

参考教材:

- (1) 李航. 机器学习方法, 北京: 清华大学出版社, 2022.
- (2) 周志华. 机器学习, 北京: 清华大学出版社, 2016.



网络参考资料:

斯坦福机器学习

<https://mooc.study.163.com/smartSpec/detail/1001319001.htm>

<https://www.coursera.org/learn/machine-learning> (吴恩达)

CMU 机器学习课程 (Eric Xing)

<http://www.cs.cmu.edu/~epxing/Class/10715>

<http://www.cs.cmu.edu/~epxing/Class/10701>

中国MOOC、B站

教学大纲



理论：22学时（平时成绩10%+期末60%）
实践：10学时（20%）

助教：高文博、吉保子坡

课程交流群

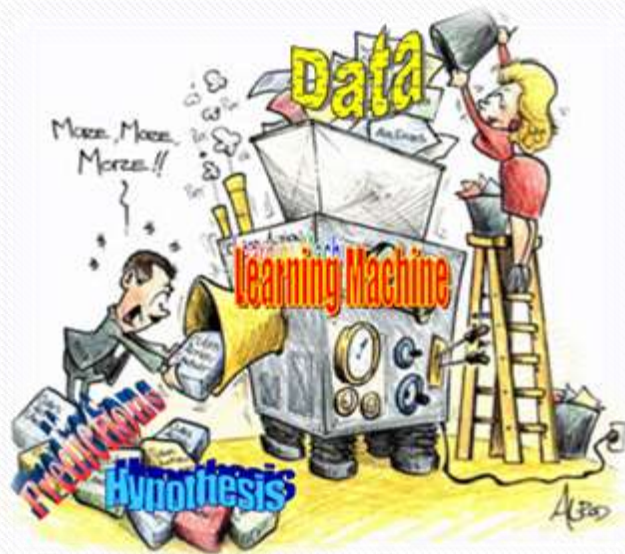


第一章：统计机器学习概论

Ch1: Basics of Statistical Machine Learning

什么是机器学习?

- 机器学习是人工智能的一个分支，主要是设计和开发一些算法让计算机从经验数据中**自动“学习”**。
- 机器学习是近40多年兴起的一门**多领域交叉学科**，涉及概率论、统计学、信息论、逼近论、最优化理论、计算机科学等多门学科。
- 机器学习算法中涉及了大量的**统计学理论**，机器学习与统计推断学联系尤为密切，也被称为**统计学习理论**。



赫尔伯特·西蒙 (Herbert A. Simon) 曾对“学习”给出以下定义：“如果一个系统能够通过执行某个过程改进它的性能，这就是学习。”按照这个观点，**统计学习就是计算机系统通过运用数据及统计方法提高系统性能的机器学习。**

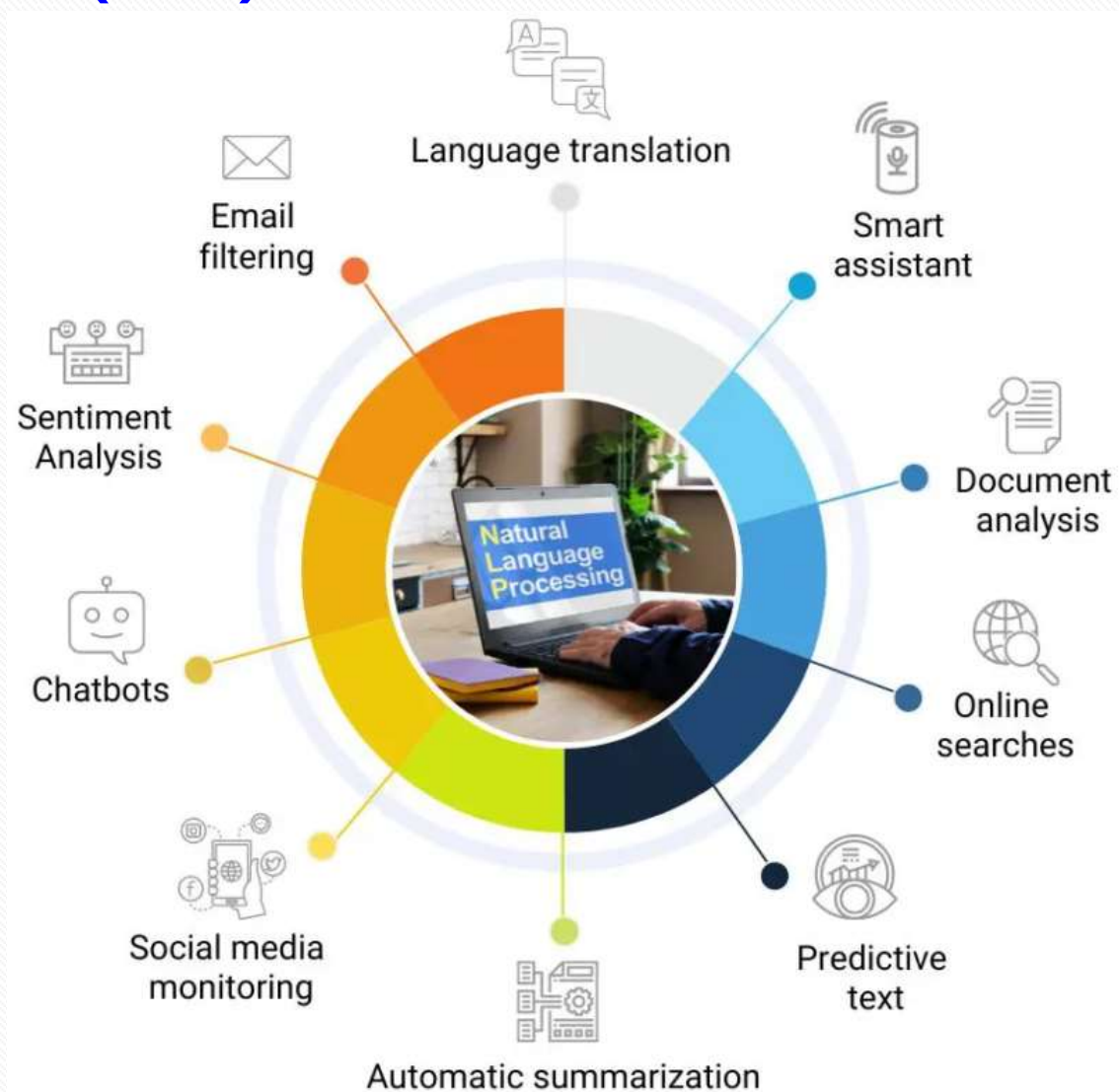
机器学习应用

- **自然语言处理**
- **计算机视觉**
- **语音和手写识别**
- **搜索引擎**
- **数据挖掘**
- **医学诊断**
- **证券市场分析**
- **机器人**
-



机器学习应用

自然语言处理(NLP)



机器学习应用

计算机视觉(CV)

图像识别



图像分类



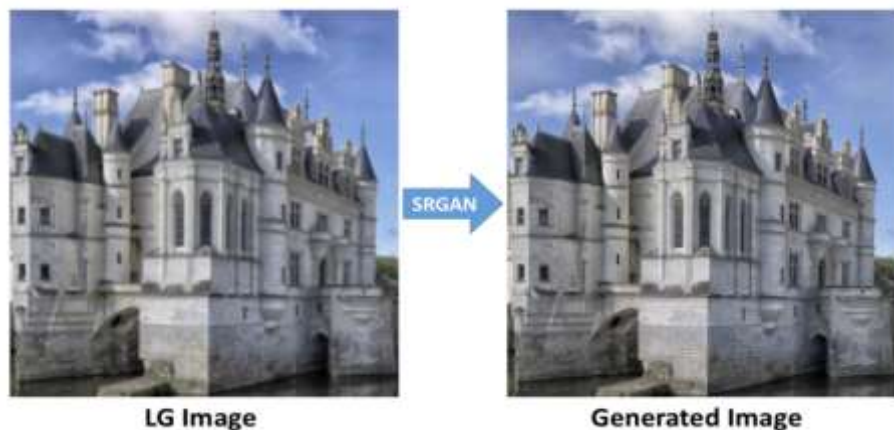
图像理解



机器学习应用

计算机视觉(CV)

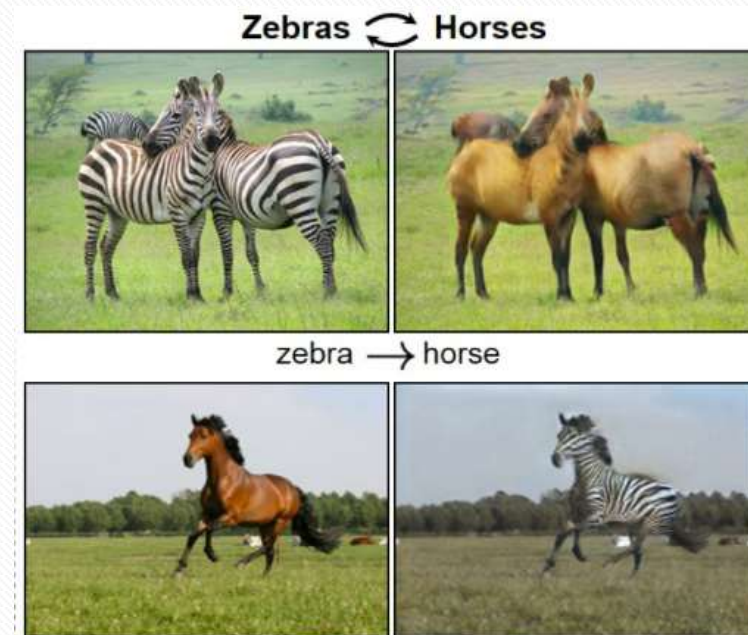
图片修复



<https://github.com/tensorlayer/srgan>



图像编辑



<https://junyanz.github.io/CycleGAN/>

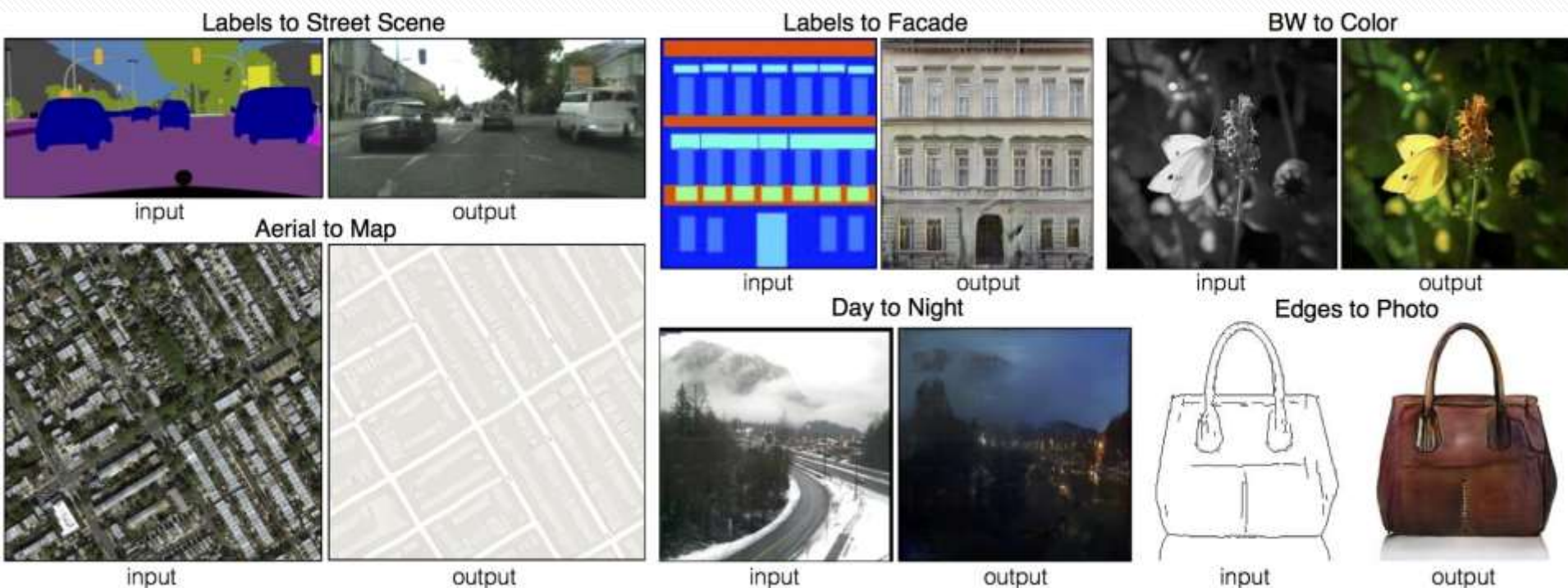
图像分割

<https://github.com/divamgupta/image-segmentation-keras>

机器学习应用

计算机视觉(CV)


Pix2Pix风格转换



<https://phillipi.github.io/pix2pix/>

机器学习应用

计算机视觉(CV)



These are not real people

https://github.com/tkarras/progressive_growing_of_gans

机器学习应用

计算机视觉(CV)

Liquid Warping GAN with Attention: A Unified Framework for Human Image Synthesis

Wen Liu, Zhao Piao, Zhi Tu, Wei-han Luo, Lin Ma, and Shenghua Gao

Abstract—We tackle human image synthesis, including human motion imitation, appearance transfer, and novel view synthesis, within a unified framework. It means that the model, once being trained, can be used to handle all these tasks. The existing task-specific methods mainly use 2D keypoints (poses) to estimate the human body structure. However, they only express the position information with no abilities to characterize the personalized shape of the person and model the limb rotations. In this paper, we propose to use a 3D body mesh recovery module to disentangle the pose and shape. It can not only model the joint location and rotation but also characterize the personalized body shape. To preserve the source information, such as texture, style, color, and face identity, we propose an Attentional Liquid Warping GAN with Attentional Liquid Warping Block (ALWB) that propagates the source information in both image and feature spaces to the synthesized reference. Specifically, the source features are extracted by a denoising convolutional auto-encoder for characterizing the source identity well. Furthermore, our proposed method can support a more flexible warping from multiple sources. To further improve the generalization ability of the unseen source images, a one-view-shot adversarial learning is applied. In detail, it firstly trains a model in an extensive training set. Then, it finetunes the model by one-view-shot unseen image(s) in a self-supervised way to generate high-resolution (512×512 and 1024×1024) results. Also, we build a new dataset, namely Imperatorator (iPER) dataset, for the evaluation of human motion imitation, appearance transfer, and novel view synthesis. Extensive experiments demonstrate the effectiveness of our methods in terms of preserving face identity, shape consistency, and clothes details. All codes and dataset are available on <https://imperatorator.github.io/iPERCore/> (pick plus icon).

Index Terms—Human Image Synthesis, Motion Imitation, Appearance Transfer, Novel View Synthesis, Generative Adversarial Network, and One-View-Shot Learning.

1 INTRODUCTION

HUMAN image synthesis aims to make believable and photo-realistic images of humans, including motion imitation [1], [2], [3], appearance transfer [4], [5] and novel view synthesis [6]. It has vast potential applications in character animation, recruitment, virtual clothes try-on, movie or game making, etc. Given a source human image and a human reference image, (i) the goal of motion imitation is to generate an image with the features from source human and pose from reference human, as depicted in the top row of Fig. 1; (ii) human novel view synthesis aims to synthesize new images of the human body, captured from different viewpoints, as illustrated in the middle row of Fig. 1; (iii) the goal of appearance transfer is to generate a human image preserving the source face identity while wearing the clothes of the reference, as shown in the bottom row of Fig. 1 where each garment (upper-clothes or pants) might come from different people.

Taking human motion imitation as an example, existing methods can be roughly categorized into an image-to-image translation-based [7], [8], [9], [10] pipeline and a warping-based pipeline [11], [12], [13]. The image-to-image translation-based pipeline trains a person-specific mapping function from the human conditions, characterized by a skeleton, dense pose, and parsing results, to the image from a video with paired sequences of conditions and images. Thus, everybody needs to train their model from scratch, and



Fig. 1: Illustration of human motion imitation, novel view synthesis and appearance transfer. The 1st row is the source image and the 2nd row is reference condition, such as image or novel viewpoint of camera. The 3rd row is the synthesized results.

• Wen Liu is with School of Information Science and Technology, Shanghai Jiao Tong University and Chinese Academy of Sciences, Shanghai Institute of Microsystem and Information Technology, and University of Chinese Academy of Sciences.

<https://github.com/iPERDance/iPERCore>

机器学习应用

计算机视觉 (Text2Image) : Midjourney



AI油画—太空歌剧院



CEO山姆·奥特曼乘坐UFO访华



马斯克在苏联工厂



中国情侣



AI水墨画—行到水穷处，坐看云起时



Fake news—特朗普被捕

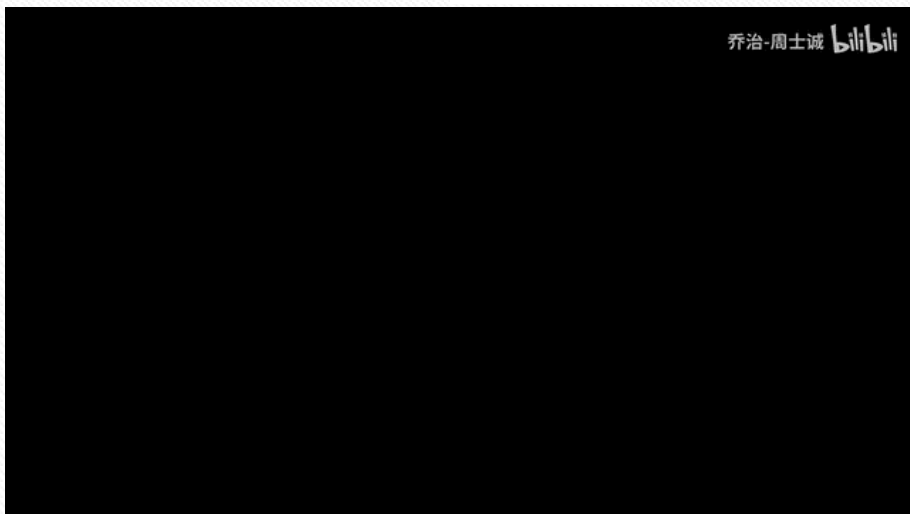
机器学习应用

计算机视觉 (Text2Video)

OpenAI Sora



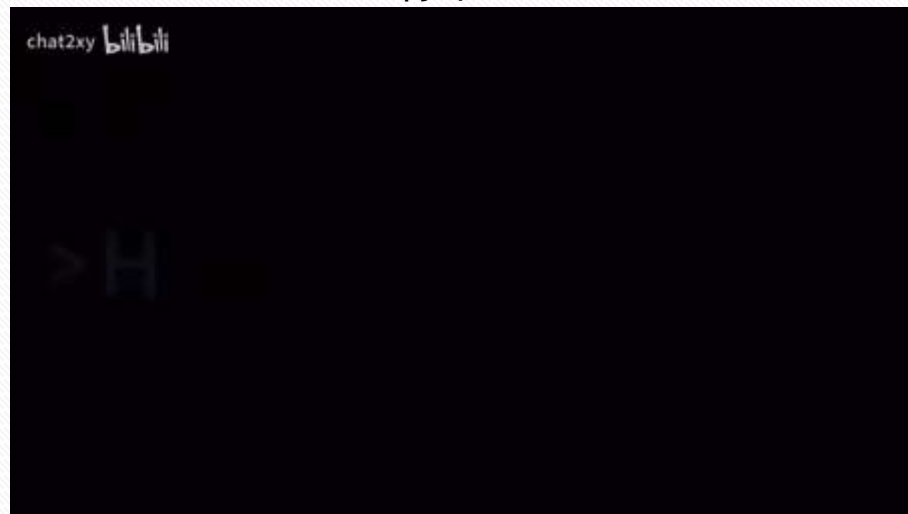
字节跳动Dreamina



OpenAI Sora



清华 Vidu



机器学习应用

语音识别

语音识别



语义理解



语音合成



https://google.github.io/tacotron/publications/speaker_adaptation

机器学习应用

决策系统

游戏AI



自动化



量化投资



穿越牛熊 旨在绝对收益



机器学习应用

大数据应用

推荐



金融



零售



机器学习应用

机器人应用



波士顿动力公司机器人Atlas



连接GPT-4的机器人Figure1

Law I: A robot may not injure a human being or, through inaction, allow a human being to come to harm.

第一定律：机器人不得伤害人类个体，或者目睹人类个体将遭受危险而袖手不管。

—阿西莫夫的《我，机器人》

机器学习应用

思考：算法可以凌驾于法律之上吗？

互联网信息服务算法推荐管理规定

2022年01月04日 10:02 来源：中国网信网  【打印】 【纠错】

国家互联网信息办公室
中华人民共和国工业和信息化部
中华人民共和国公安部
国家市场监督管理总局

令
第9号

《互联网信息服务算法推荐管理规定》已经2021年11月16日国家互联网信息办公室2021年第20次室务会议审议通过，并经工业和信息化部、公安部、国家市场监督管理总局同意，现予公布，自2022年3月1日起施行。

国家互联网信息办公室主任 庄荣文
工业和信息化部部长 肖亚庆
公安部部长 赵克志
国家市场监督管理总局局长 张 工

2021年12月31日

https://www.sohu.com/a/515442721_416839

机器学习应用

网站首页 > 网站导航 > 正文

申请入驻 >

AI抗疫！美国科学家正借助机器学习、大数据寻找新冠肺炎最佳疗法

清华大学利用人工智能优化新冠抗体，获得高效广谱中和活性！

2022
03/15

张林琦课题组 / 清华大学医学院

A+ A-

与传统方法相比，深度学习具有更大的搜索空间，并且可同时针对不同新冠进行突变株抗体优化，由此理论上可获得更高效更广谱的中和抗体。

MIT News

ON CAMPUS AND AROUND THE WORLD

Deep learning helps predict new drug combinations to fight Covid-19

Neural network identifies synergistic drug blends for treating viruses like SARS-CoV-2.

Rachel Gordon | MIT CSAIL

September 24, 2021

首页 > 央广网财经 > 聚焦 > 财经数评

中国工程院院士张亚勤：人工智能加快新冠疫苗研发进程

2021-12-01 14:05:07 来源：央广网

我国学者在人工智能辅助新冠肺炎影像诊断方面取得进展

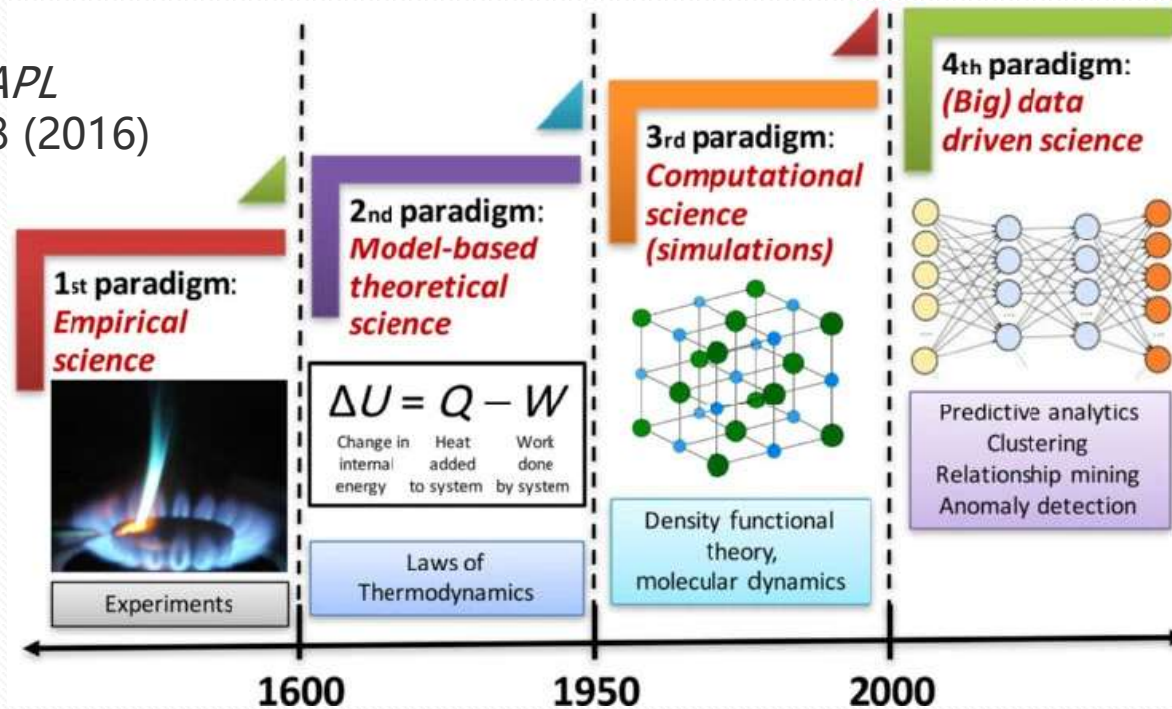
2021/11/02 19:01:04

导读：广东工业大学智能信息处理团队基于注意力机制、可变形卷积理论和方法，提出了“面向新冠肺炎快速影像检测”应注意力深度学习网络AANet。

《**抗击新冠肺炎疫情的中国行动**》白皮书指出中国充分利用**大数据、人工智能**等新技术，进行疫情趋势研判，开展流行病学调查，努力找到每一个感染者、穷尽式地追踪密切接触者并进行隔离。

机器学习应用

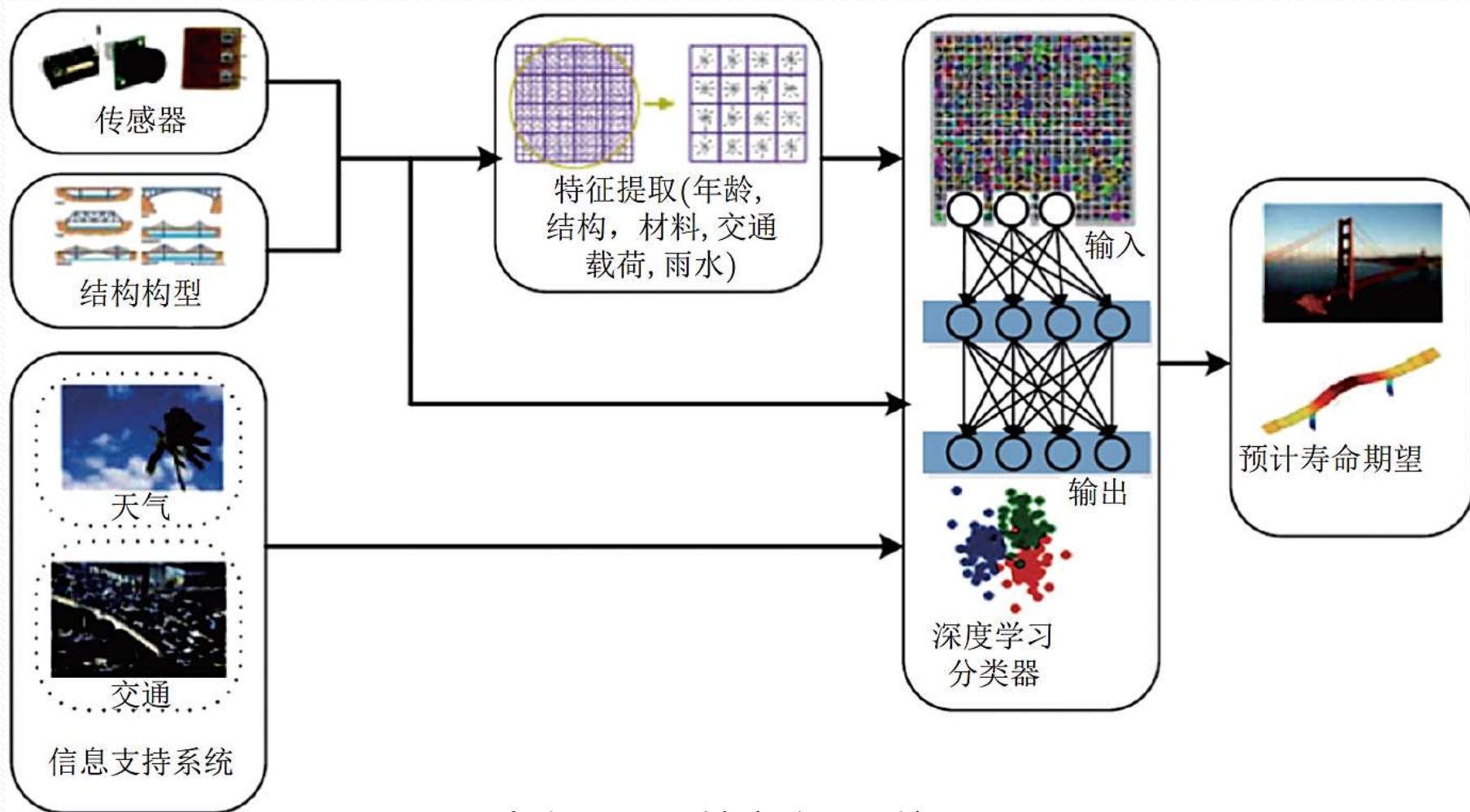
Agrawal et al., *APL Mater.* **4**, 053208 (2016)



最初，只有实验科学，后来出现了理论科学，如开普勒定律、牛顿运动定律、麦克斯韦方程等。然而，对于许多问题，理论模型变得过于复杂，以至于无法用分析法解决，人们不得不开始模拟。这些模拟让我们度过了上个世纪后半叶的大部分时间。到目前为止，这些模拟产生了大量的数据，再加上实验科学数据的急剧增加。现在人们不再真正通过望远镜观察。相反，他们通过大规模、复杂的仪器“观察”，这些仪器将数据传送到数据中心，然后他们才在电脑上查看这些信息。

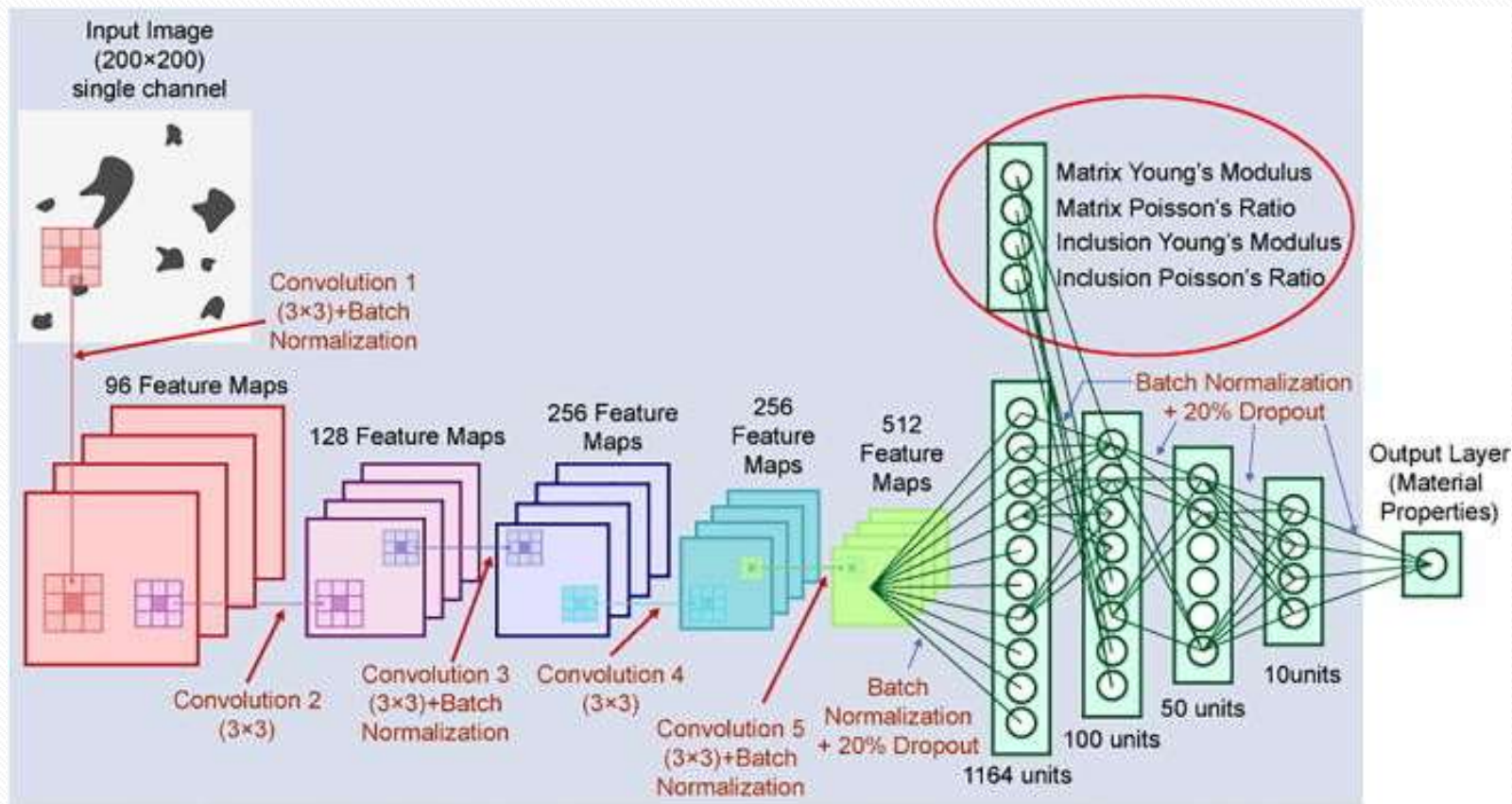
In 2007 Jim Gray, Turing Award winner (1998)

机器学习应用



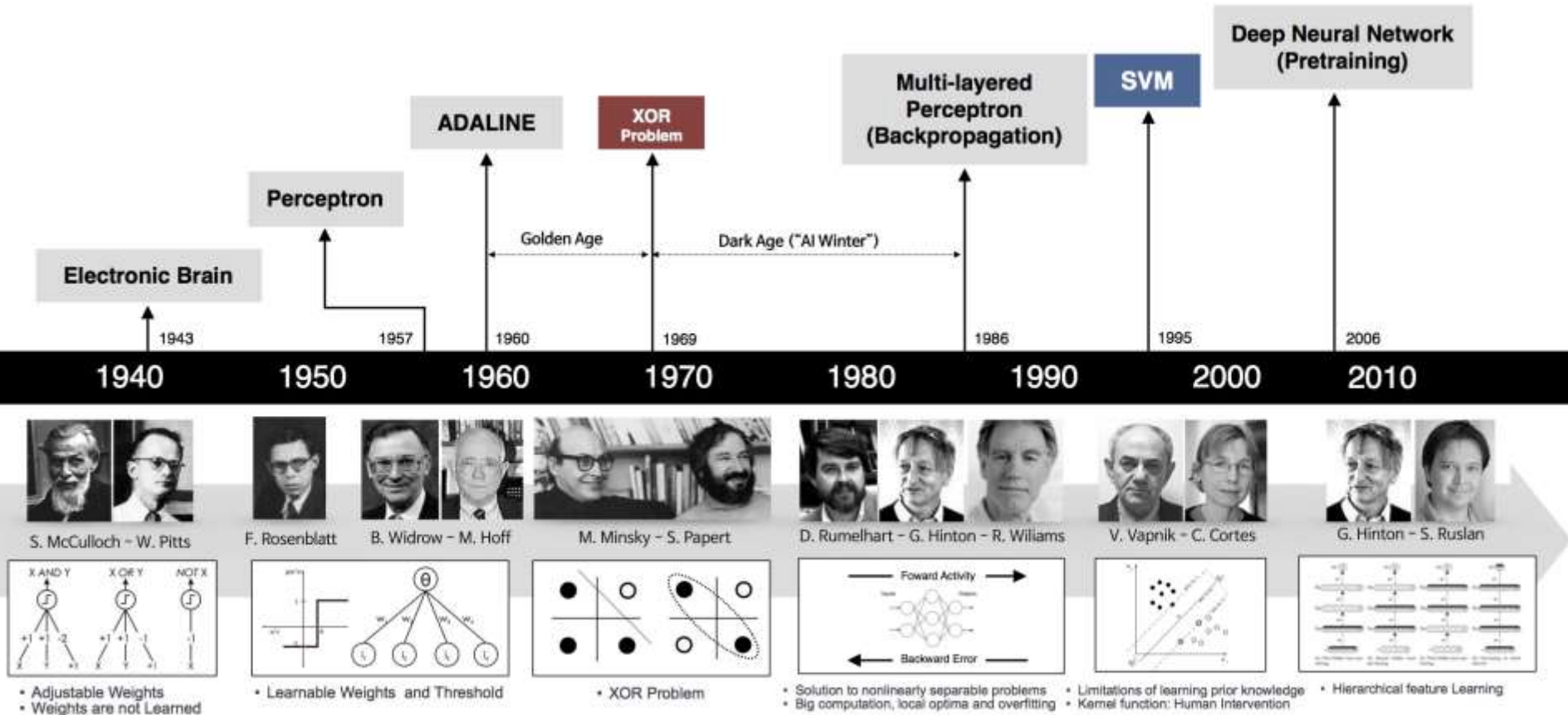
桥梁 SHM 健康监测系统

机器学习应用

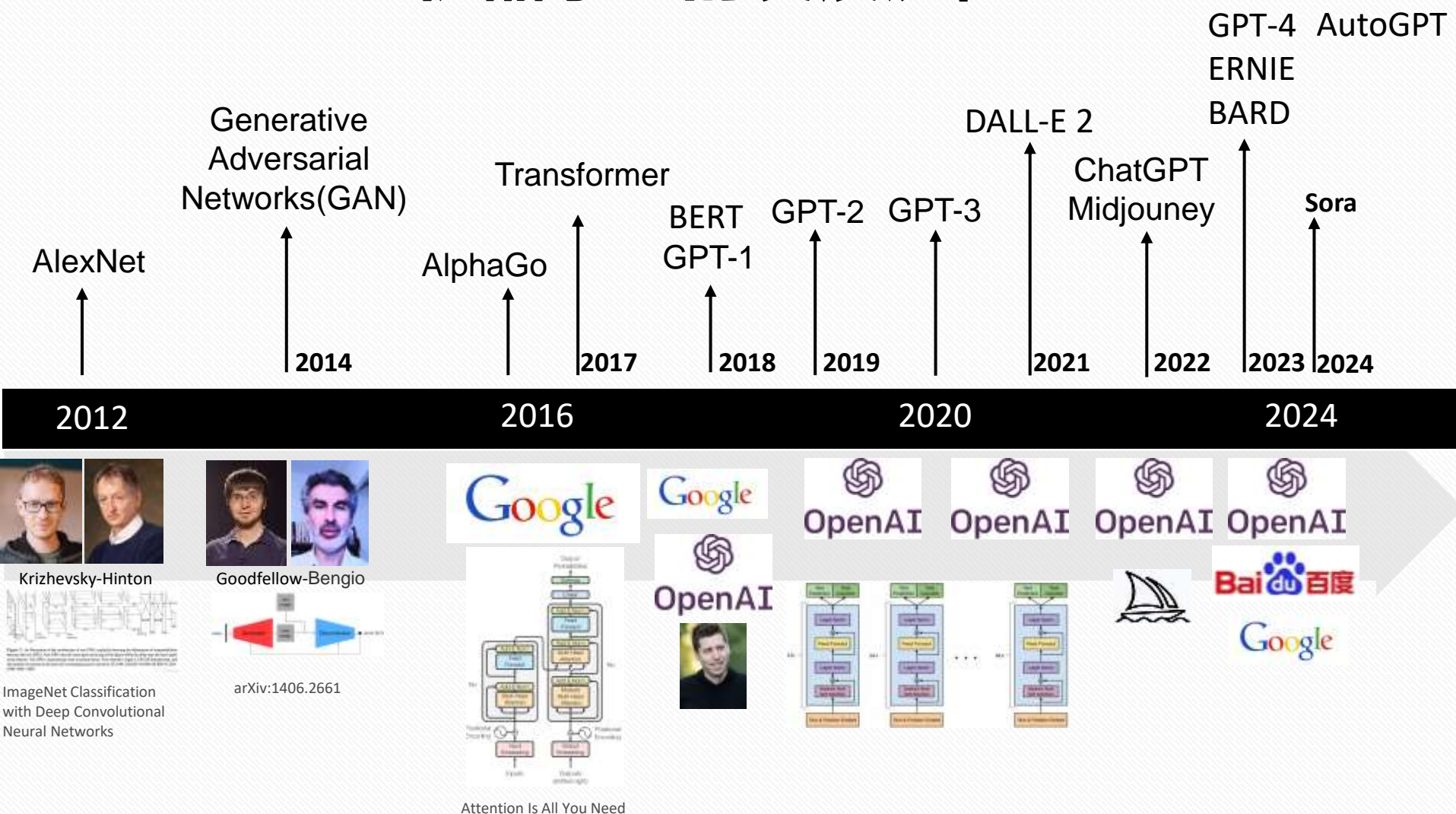


信息力学

机器学习的发展历程

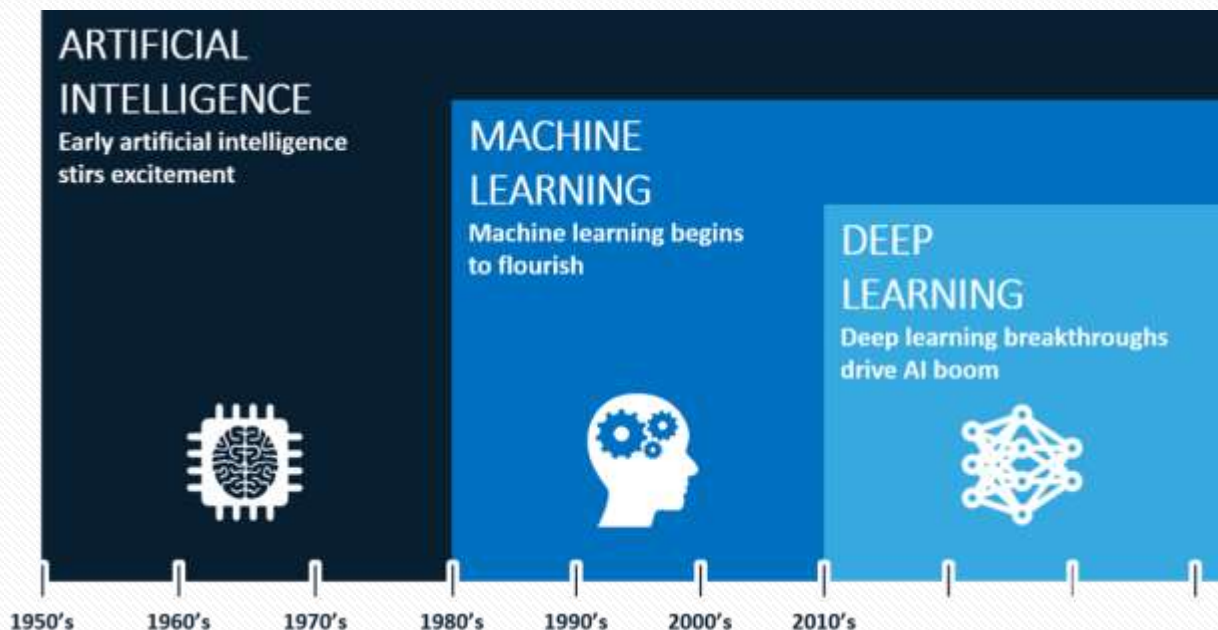


机器学习的发展历程



头脑风暴：马斯克、苹果联合创始人、图灵奖得主联名上书叫停GPT-5。

机器学习的发展历程



人工智能 Artificial Intelligence

AI 的初衷及最终理想是通过逻辑、推理、演绎来实现智慧。当今的AI实际上是大数据加深度学习。

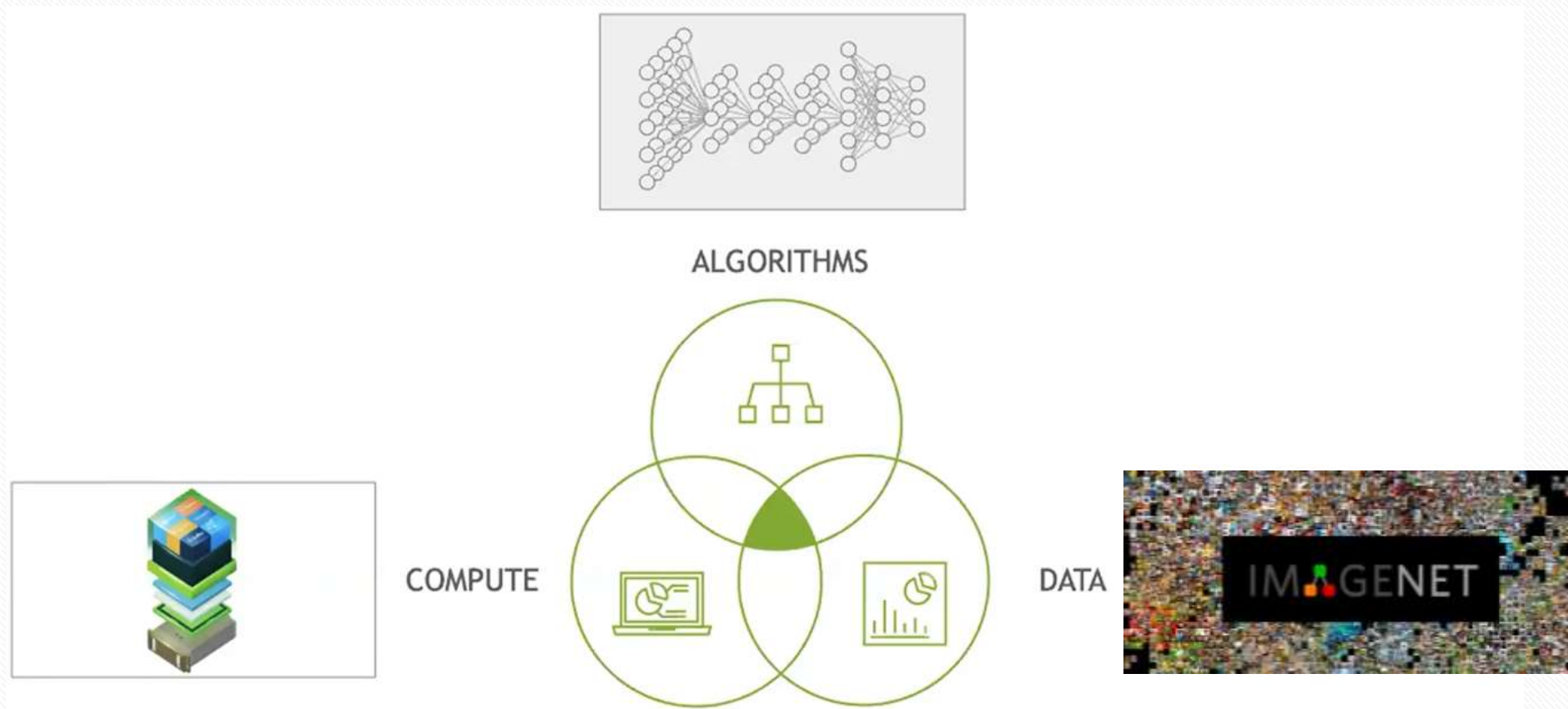
机器学习 Machine Learning

最早是使用经验学习研究人工智能的方法，后来纳入很多统计学的思想和方法，并在计算机算法方面取得了很大进展。

深度学习 Deep Learning

机器学习中的一个分支，使用神经网络技术，在大数据的基础上，结合GPU计算实现深层次的网络结构。

机器学习的发展历程



More than 14 million images have been hand-annotated by the project

As of Mar 11 2021

<https://www.image-net.org/>

机器学习的发展历程

新的方向：

- **集成学习** (Ensemble Learning) 通过构建并结合多个学习器来完成学习任务，有时也被称为多分类器系统、基于委员会的学习等。
- **强化学习** (Reinforcement Learning)，又称再励学习、评价学习或增强学习，是机器学习的范式和方法论之一，用于描述和解决**智能体** (agent) 在与**环境**的交互过程中通过学习策略以达成**回报最大化**或实现特定目标的问题。
- **迁移学习** (Transfer Learning) 顾名思义就是把已训练好的模型（预训练模型）参数迁移到新的模型来帮助新模型训练。
- **深度学习** (Deep Learning) 是用于建立、模拟人脑进行分析学习的神经网络，并模仿人脑的机制来解释数据的一种机器学习技术。

统计学习

统计学习的主要特点：

- 以**计算机及网络**为平台，是建立在计算机及网络之上的；
- 统计学习以数据为研究对象，是**数据驱动**的学科；
- 其目的是对数据进行**预测与分析**；
- 它以**方法**为中心，统计学习方法构建模型并应用模型进行预测与分析；
- 统计学习是概率论、统计学、信息论、计算理论、最优化理论及计算机科学等多个领域的**交叉学科**（*e.g.*, **CMU的机器学习系**），并且已经逐步形成**独自的理论体系与方法论**。

统计学习

统计学习的对象：

- **数据 (Data)**：计算机及互联网上的各种数字、文字、图像、视频、音频数据以及它们的组合。
- 数据的基本假设是同类数据具有一定的**统计规律性**。

统计学习的目的：

- 用于对数据（特别是未知数据）进行**预测**和**分析**。
- 使计算机更加智能化，或者说使计算机的某些性能得到提高。

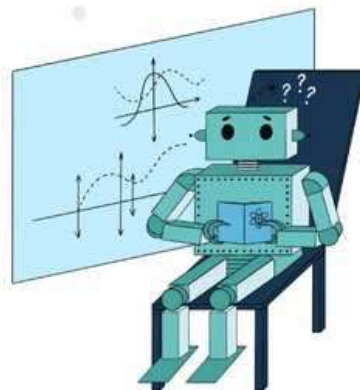
统计学习

统计学习方法分类：

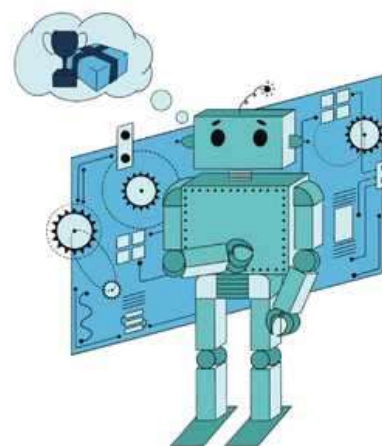
- **监督学习：**有特征、有标签（即有标准答案）。例如，回归和分类。
- **非监督学习：**有特征、无标签（即无标准答案）。例如，聚类、关联规则。
- **强化学习：**智能体与环境的交互。例如Alpha go。



监督学习
Supervised learning



非监督学习
Unsupervised learning



强化学习
Reinforcement learning

监督学习

定义： 是
的统计规律

从输入到输出

输入空间：

空间)

输出空间：

空间)

实例： 每

$(x^{(i)}, \dots, x^{(n)})^T$

特征空间：

训练集和

$T = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})\}$

为第*i*个实例

应用场景：

回归问题： 输入输出变量均为连续变量

分类问题： 输出变量为离散变量

标注问题： 输入输出变量均为变量序列



监督学习

联合概率分布（基本假设）

统计学习假设数据存在一定的统计规律

假设输入与输出的随机变量 X 和 Y 遵循联合概率分布 $P(X, Y)$

$P(X, Y)$ 为分布函数（**离散型**随机变量）或分布密度函数（**连续性**随机变量）

对于学习系统来说，联合概率分布是未知的

训练数据和测试数据被看作是依联合概率分布 $P(X, Y)$ **独立同分布(IID)**产生的

假设空间

监督学习目的是学习一个由**输入到输出的映射**，称为模型

模型的集合就是假设空间（hypothesis space）（学习范围的指定）

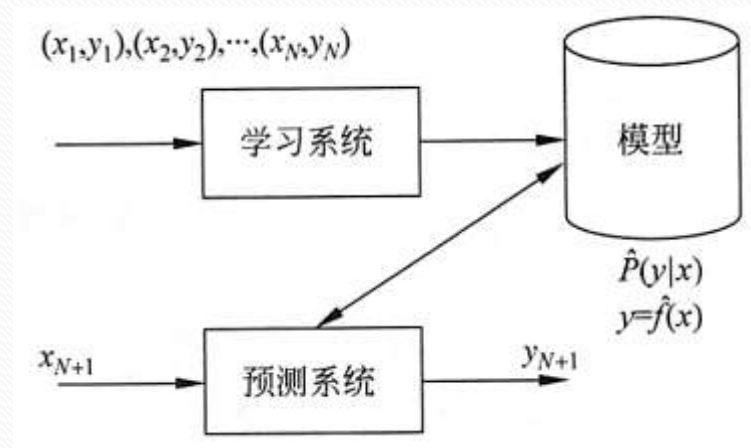
非概率模型：决策函数 $Y=f(X)$

概率模型：条件概率分布 $P(Y|X)$

监督学习

问题的形式化（统计学习的一般流程）

- (1) 得到一个有限的**训练数据集**；
- (2) 确定包含所有可能的模型的**假设空间**，即学习模型的集合；
- (3) 确定模型选取的准则，即学习的**策略**；
- (4) 实现求解最优模型的算法，即学习的**算法**；
- (5) 通过学习方法选取**最优模型**；
- (6) 利用学习的最优模型对新数据进行**预测分析**。



$$y_{N+1} = \arg \max_y \hat{P}(y|x_{N+1})$$

$$y_{N+1} = \hat{f}(x_{N+1})$$

经典的监督学习模型：支持向量机、线性判别、决策树、朴素贝叶斯。

监督学习



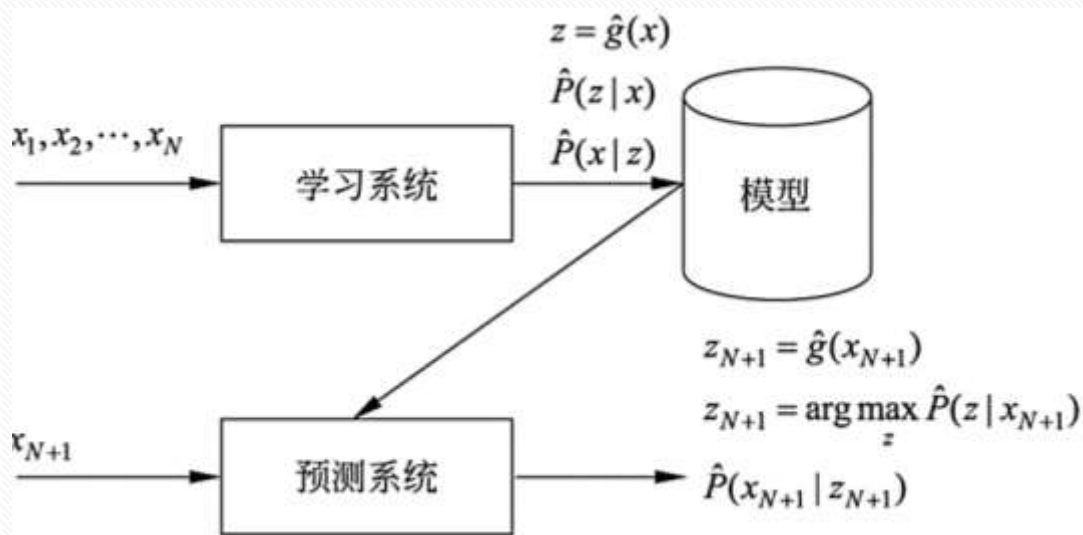
无监督学习

定义：是指从**无标注**的数据中学习预测模型的机器学习问题。无监督学习的本质是学习数据中的规律和潜在结构。

训练集： $U = \{x_1, x_2, \dots, x_N\}$

模型函数： $z = g(x)$

条件概率分布： $P(z|x)$
 $P(x|z)$



应用： 类聚、数据降维和标记处理

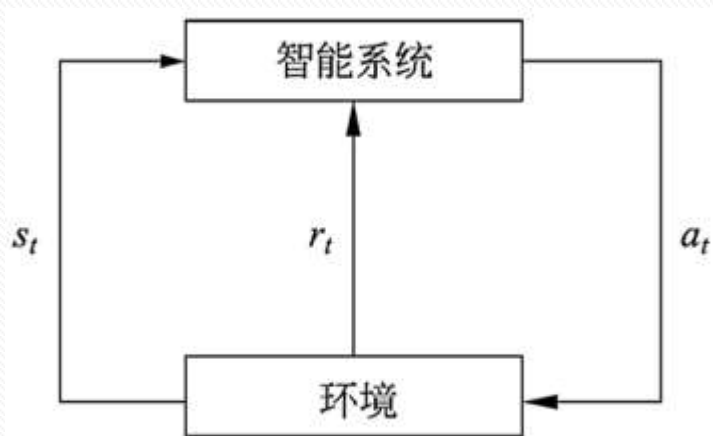


无监督学习



强化学习

定义：是指智能系统与环境的连续**互动中学习**最优化行为策略的机器学习问题。智能系统的目标不是短期奖励的最大化，而是**长期累积奖励的最大化**，强化学习过程中，系统不断地试错（trial and error），以达到学习最优策略的目的。强化学习的本质是**学习最优序贯决策**。

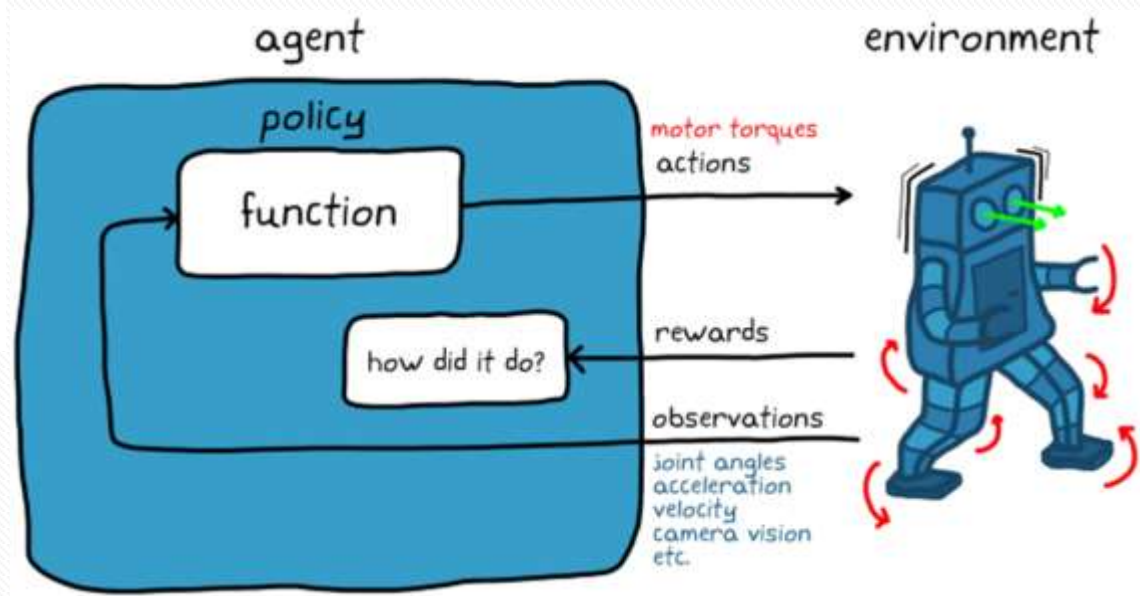


s_t -状态 (State)

r_t -奖励 (Reward)

a_t -动作 (Action)

例子：机器人走路

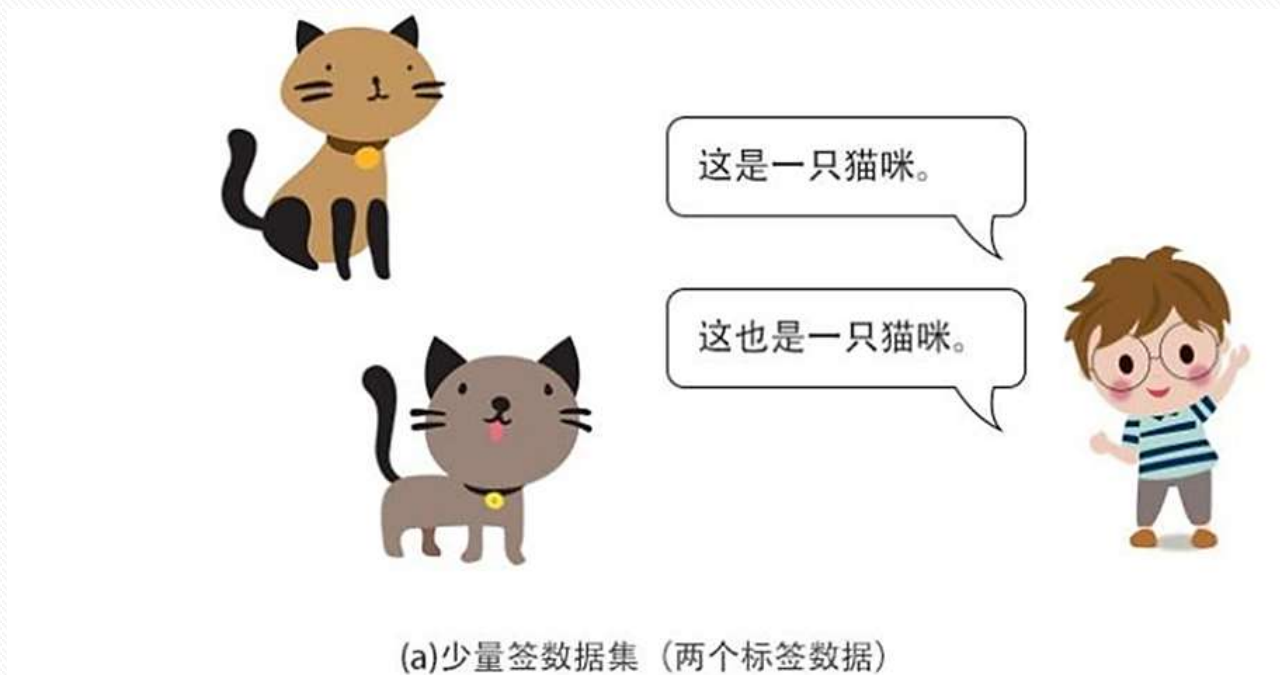


半监督学习

半监督学习 (semi-supervised learning) 是指利用**标注**数据和**未标注**数据学习预测模型的机器学习问题。通常有少量标注数据、大量未标注数据，因为标注数据的构建往往需要人工，成本较高，未标注数据的收集不需太多成本。半监督学习旨在利用未标注数据中的信息，辅助标注数据，进行监督学习，以较低的成本达到较好的学习效果。

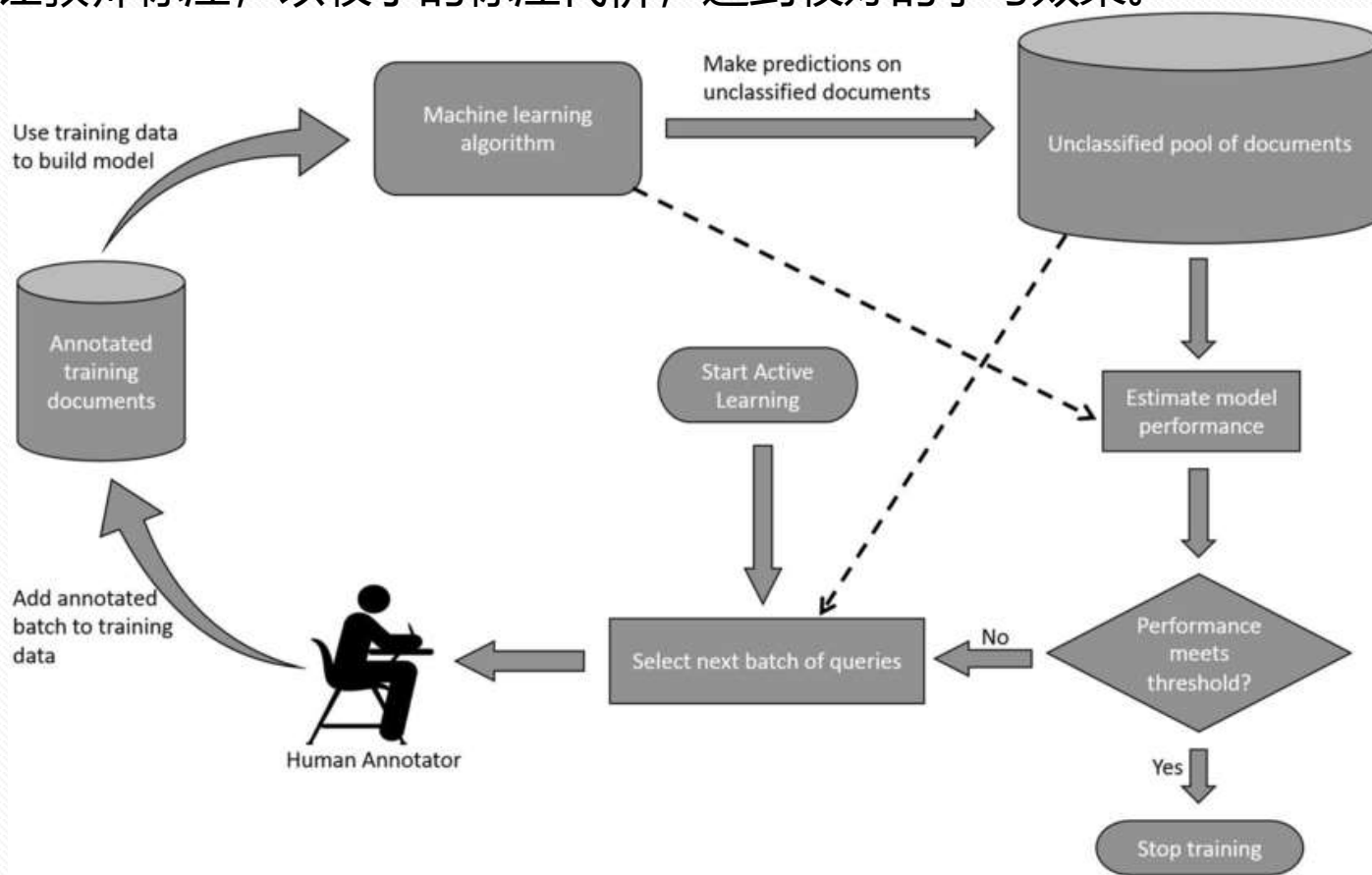
在大数据时代，半监督学习的现实需求非常强烈。因为有标签数据的收集和标记需要消耗大量的人力物力，而海量的非标签数据却触手可及，**“半监督学习”将成为大数据时代的发展趋势。**

半监督学习



主动学习

主动学习（active learning）是指机器不断主动给出实例让教师进行标注，然后利用标注数据学习预测模型的机器学习问题。通常的监督学习使用给定的标注数据，往往是随机得到的，可以看作是“被动学习”，主动学习的目标是找出对学习最有帮助的实例让教师标注，以较小的标注代价，达到较好的学习效果。



统计机器学习三要素

如何评价假设空间的
每一个假设？

策略

给定数据集和
学习任务，如
何选择模型
(假设空间)？

模型

算法

如何以最快的搜索
并发现最优假设？

统计
学习
方法

方法=模型+策略+算法

统计学习三要素

在监督学习过程中，**模型**就是要学习的**条件概率分布**或**决策函数**。模型的假设空间包含所有可能的条件概率分布或决策函数

模型：

决策函数的集合： $\mathcal{F} = \{f | Y = f(X)\}$

参数空间： $\mathcal{F} = \{f | Y = f_{\theta}(X), \theta \in \mathbf{R}^n\}$

θ 为参数向量

条件概率的集合： $\mathcal{F} = \{P | P(Y | X)\}$

参数空间： $\mathcal{F} = \{P | P_{\theta}(Y | X), \theta \in \mathbf{R}^n\}$

方法=**模型**+策略+算法

统计学习三要素

策略：

统计学习的目标在于从假设空间中选取**最优模型**。为了评价模型的优劣，引入损失函数与风险函数的概念。

损失函数：一次预测的好坏

风险函数：平均意义下模型预测的好坏

0-1损失函数 0-1 loss function $L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$ (分类问题)

平方损失函数 quadratic loss function $L(Y, f(X)) = (Y - f(X))^2$ (回归问题)

绝对损失函数 absolute loss function $L(Y, f(X)) = |Y - f(X)|$ (回归问题)

对数损失函数 logarithmic loss function

对数似然损失函数 loglikelihood loss function

$$L(Y, P(Y|X)) = -\log P(Y|X)$$

(逻辑回归)

方法=模型+策略+算法

为什么log? 为什么负号?

统计学习三要素

策略:

P(x, y) 存在, 但是未知!

损失函数的期望 $R_{\text{exp}}(f) = E_P[L(Y, f(X))] = \int_{\mathcal{X} \times \mathcal{Y}} L(y, f(x)) P(x, y) dx dy$

是平均意义下的损失, 称为**风险函数** risk function / **期望损失** expected loss。

模型关于训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 的平均损失称为**经验风险** (empirical risk) 或**经验损失** (empirical loss), 记作:

$$R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))$$

根据大数定律(law of large numbers), 当样本容量**N趋于无穷**时, **经验风险** empirical risk \approx **期望风险** empirical loss。所以很自然的想法就是用经验风险估计期望风险。

方法=模型+策略+算法

统计学习三要素

策略：经验风险最小化 or 结构风险最小化？

现实中，训练样本数量有限，甚至很小，经验风险最小化不理想。

经验风险最小化 empirical risk minimization 最优模型： $\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))$

当样本容量很小时，经验风险最小化学习的效果未必很好，会产生“**过拟合 (over-fitting)**”

结构风险最小化 structure risk minimization，为防止过拟合提出的策略，等价于**正则化** (regularization)，加入正则化项regularizer，或罚项 penalty term：

$$R_{srm}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f)$$

求最优模型就是求解**最优化问题**： $\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f)$

方法=模型+策略+算法

统计学习三要素

算法：

算法是指学习模型的具体计算方法。统计学习基于训练数据集，根据学习策略，从假设空间中选择最优模型，最后考虑用何种计算方法求解最优模型。这时，统计学习问题归结为最优化问题。**统计学习的算法即求解最优化问题的算法。**

如果最优化问题有**显式的解析式**，算法比较简单，但通常解析式不存在，就需要**数值计算**的方法：

- 最小二乘法：针对线性模型
- 梯度下降、上升法（批梯度、增量梯度）：任意模型

方法=模型+策略+算法

模型评估与模型选择

训练误差，训练数据集的平均损失： $R_{\text{emp}}(\hat{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$

测试误差，测试数据集的平均损失： $e_{\text{test}} = \frac{1}{N'} \sum_{i=1}^{N'} L(y_i, \hat{f}(x_i))$

损失函数是**0-1**损失时： $e_{\text{test}} = \frac{1}{N'} \sum_{i=1}^{N'} I(y_i \neq \hat{f}(x_i))$

测试数据集的**准确率**： $r_{\text{test}} = \frac{1}{N'} \sum_{i=1}^{N'} I(y_i = \hat{f}(x_i))$

显然， $r_{\text{test}} + e_{\text{test}} = 1$

过拟合与模型选择

如果一味追求提高对训练数据的预测能力，所选模型的复杂度则往往会比真模型更高。这种现象称为过拟合（over-fitting, **考试例子**）。过拟合是指学习时选择的模型所包含的参数过多，以至出现这一模型对已知数据预测得很好，但对未知数据预测得很差的现象。

模型选择旨在避免过拟合并提高模型的预测能力。

过拟合与模型选择

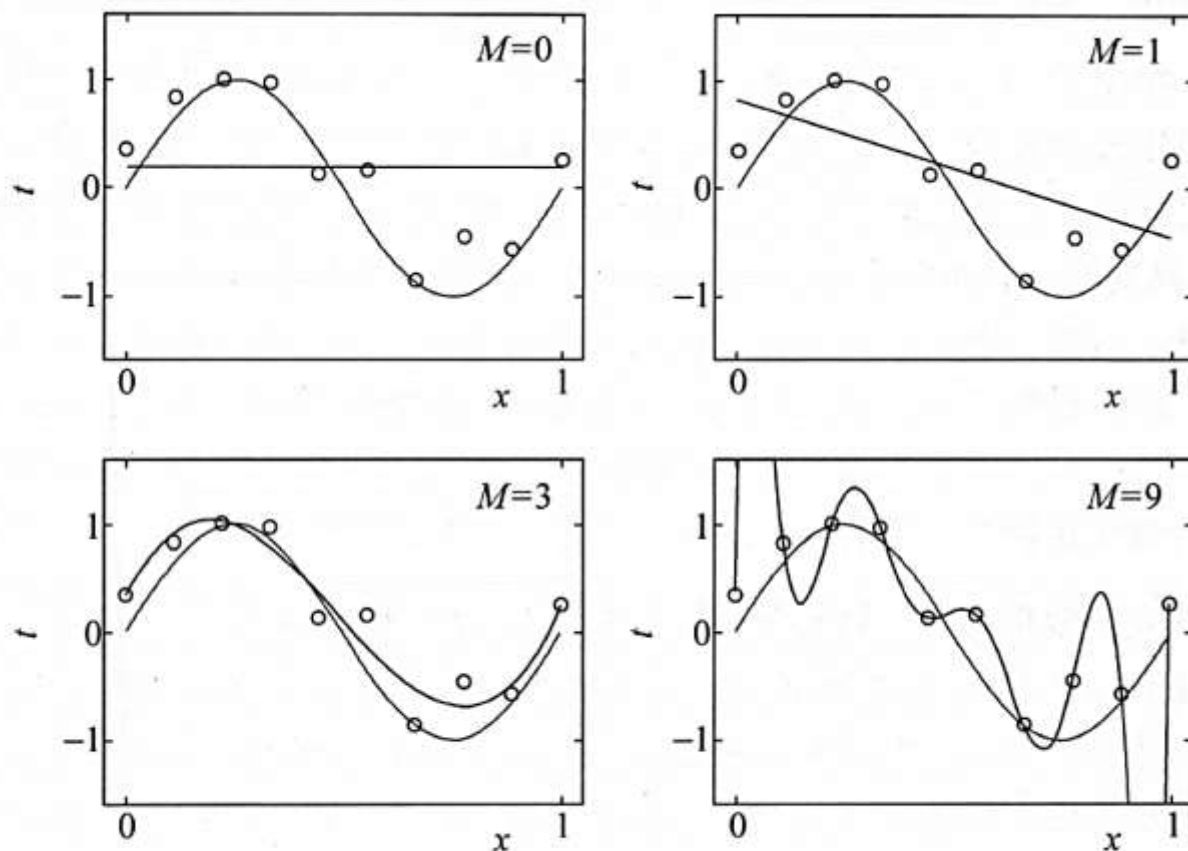
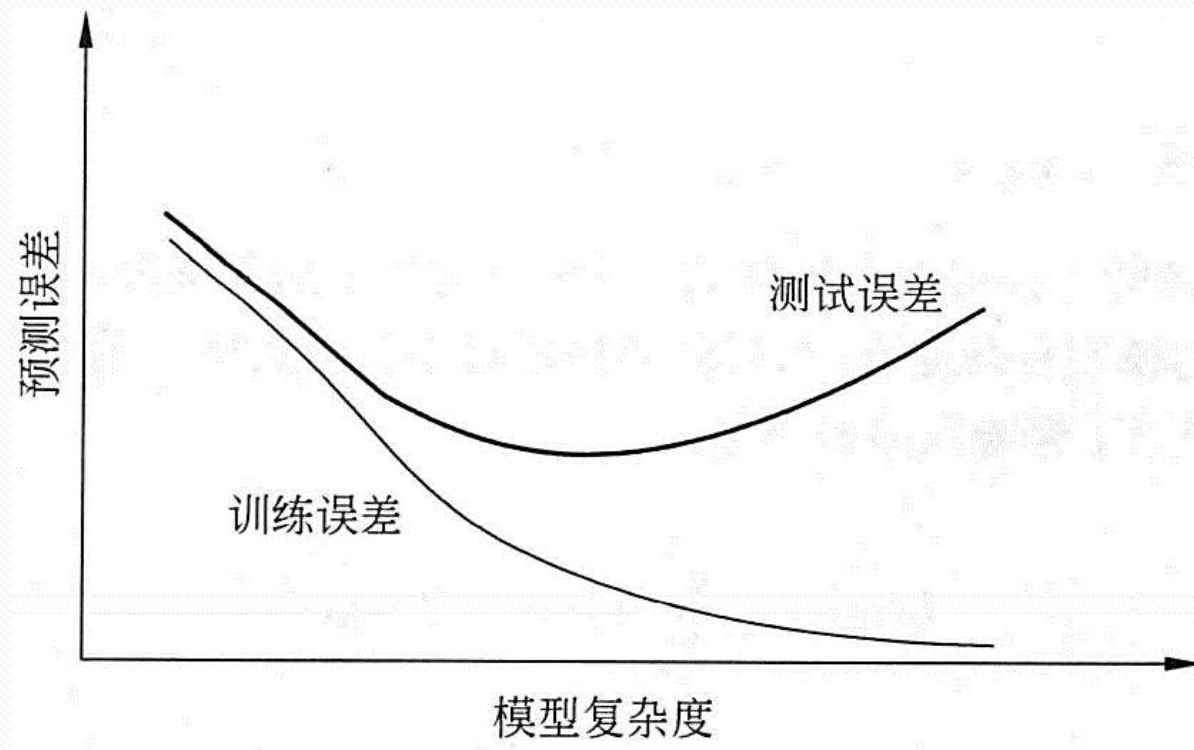


图 1.8 M 次多项式函数拟合问题的例子

$$f_M(x, w) = w_0 + w_1x + w_2x^2 + \cdots + w_Mx^M = \sum_{j=0}^M w_j x^j$$
$$L(w) = \frac{1}{2} \sum_{i=1}^N \left(\sum_{j=0}^M w_j x_i^j - y_i \right)^2$$

过拟合与模型选择



正则化

正则化是结构风险最小化 structure risk minimization的实现，是在经验风险上加一个正则化项regularizer，或罚项 penalty term：

$$\min_{f \in \mathcal{F}} \underbrace{\frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))}_{\text{经验风险}} + \underbrace{\lambda J(f)}_{\text{正则项}} \quad \lambda \geq 0,$$

$J(f)$ 表示模型复杂度

回归问题中：

$$L(w) = \frac{1}{N} \sum_{i=1}^N (f(x_i; w) - y_i)^2 + \frac{\lambda}{2} \|w\|^2 \quad \|w\| - L_2 \text{范数}$$
$$L(w) = \frac{1}{N} \sum_{i=1}^N (f(x_i; w) - y_i)^2 + \lambda \|w\|_1 \quad \|w\|_1 - L_1 \text{范数}$$

正则化符合奥卡姆剃刀 (Occam's razor) 原理：能够很好地解释已知数据并且十分简单才是最好的模型，也就是应该选择的模型。

防止过拟合方法

- 交叉验证
- 正则化
- 特征选择
- 数据增强
- 增加数据集
- Early stopping
- 增加噪声
- 简化网络结构
- Dropout
- 贝叶斯方法
- 集成学习

交叉验证

如果给定的样本数据充足，进行模型选择的种简单方法是随机地将数据集切分成部分，分别为训练集（training set）、验证集（validation set）和测试集（test set）。训练集用来训练模型，验证集用于模型的选择，而测试集用于最终对学习方法的评估。在学习到的不同复杂度的模型中，选择对验证集有最小预测误差的模型。由于验证集有足够多的数据，用它对模型进行选择也是有效的。

训练集 Training set:

用于训练模型

验证集 Validation set:

用于模型选择

测试集 Test set:

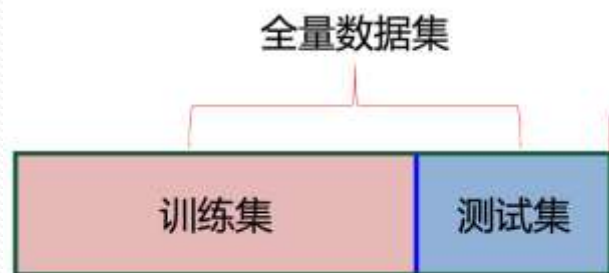
用于最终对学习方法的评估

} “互斥”

交叉验证

在许多实际应用中数据是不充足的。为了选择好的模型，可以采用交叉验证方法。交叉验证的基本想法是重复地使用数据；把给定的数据进行切分，将切分的数据集组合为训练集与测试集，在此基础上反复地进行训练、测试以及模型选择。

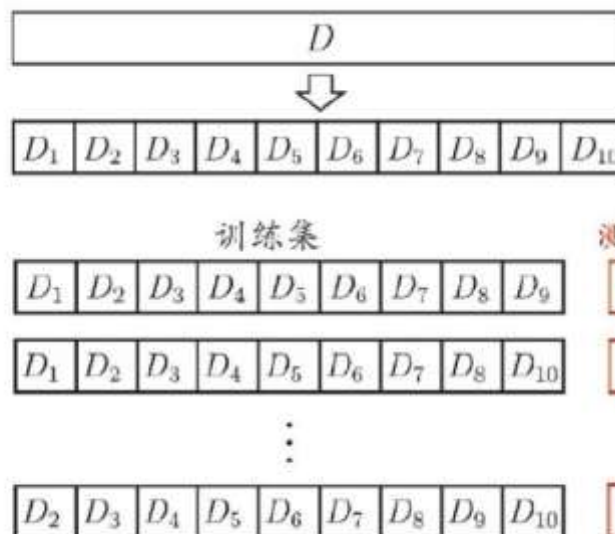
交叉验证方法：简单交叉验证、 k -折交叉验证、留一法



简单交叉验证

典型的
10折交叉验证

k -折交叉验证



留一交叉验证

若 $k = m$ ，则得到“留一法”
(leave-one-out, LOO)



泛化能力

泛化能力：泛化能力（generalization ability）是指由该方法学习到的模型对未知数据的预测能力，是学习方法本质上重要的性质。

泛化误差：学习到的模型对于未知数据预测的误差

$$\begin{aligned} R_{\text{exp}}(\hat{f}) &= E_P[L(Y, \hat{f}(X))] \\ &= \int_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{f}(x)) P(x, y) dx dy \end{aligned}$$

泛化误差上界：

性质：样本容量增加，泛化误差趋于0，假设空间容量越大，泛化误差越大

$$R(f) \leq \hat{R}(f) + \sqrt{\frac{1}{2N} \left(\log d + \log \frac{1}{\delta} \right)} \quad (\text{二分类问题})$$

训练误差上界

样本容量

假设空间容量

监督学习应用

分类问题

在监督学习中，当输出变量取有限个离散值时，预测问题便成为分类问题。这时，输入变量可以是离散的，也可以是连续的。监督学习从数据中学习一个分类模型或分类决策函数，称为分类器(classifier)，可以是 $P(Y|X)$ 或者 $f(X)$ 。

如何评价预测精度？

Confusion Matrix

<u>Actual</u>	0	1
<u>Predict</u>	0	1
0	TN	FN
1	FP	TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

二分类问题

TP — True Positive	真正类——	将正类预测为正类数；
FN — False Negative	假负类——	将正类预测为负类数；
FP — False Positive	假正类——	将负类预测为正类数；
TN — True Negative	真负类——	将负类预测为负类数。

监督学习应用

标注问题

标注 (Tagging) 问题的输入是一个观测序列，输出是一个标记序列或状态序列。标注问题的目标在于学习一个模型，使它能够对观测序列给出标记序列作为预测。

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

输入观测序列 $x_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)})^T, i = 1, 2, \dots, N$

输出标记序列 $y_i = (y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(n)})^T$

学习模型: $P(Y^{(1)}, Y^{(2)}, \dots, Y^{(n)} | X^{(1)}, X^{(2)}, \dots, X^{(n)})$

标记表示名词短语的“开始”、“结束”或“其他”（分别以B, E, O表示）

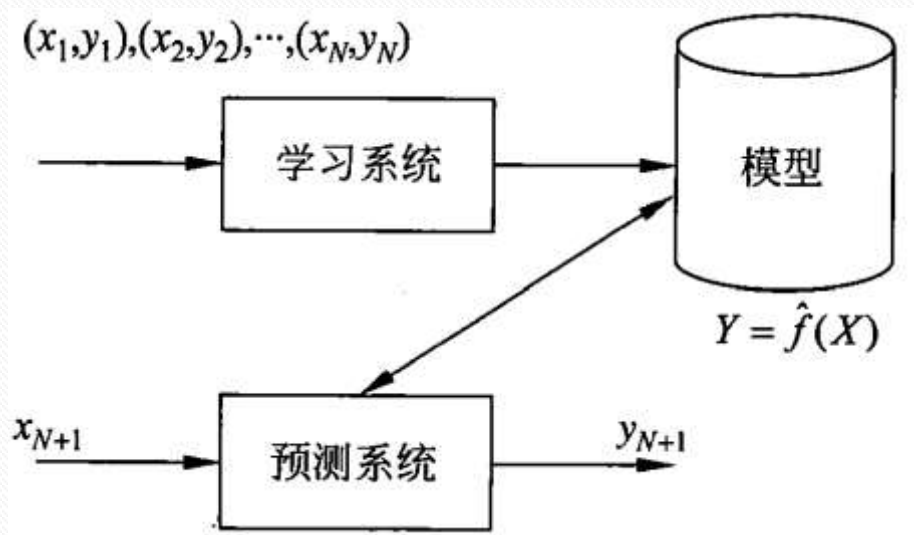
输入：At Microsoft Research, we have an insatiable curiosity and the desire to create new technology that will help define the computing experience.

输出：At/O Microsoft/B Research/E, we/O have/O an/O insatiable/B curiosity/E and/O the/O desire/BE to/O create/O new/B technology/E that/O will/O help/O define/O the/O computing/B experience/E.

监督学习应用

回归问题

回归（Regression）模型正是表示从输入变量到输出变量之间映射的函数。回归问题的学习等价于函数拟合：选择一条函数曲线使其很好地拟合已知数据且很好地预测未知数据。



变量的个数：一元回归和多元回归；
模型的类型：线性回归和非线性回归。

应用场景：股价预测？