南京大学
NANJING UNIVERSITY

人工智能导论

机器学习概览

郭兰哲

南京大学 智能科学与技术学院

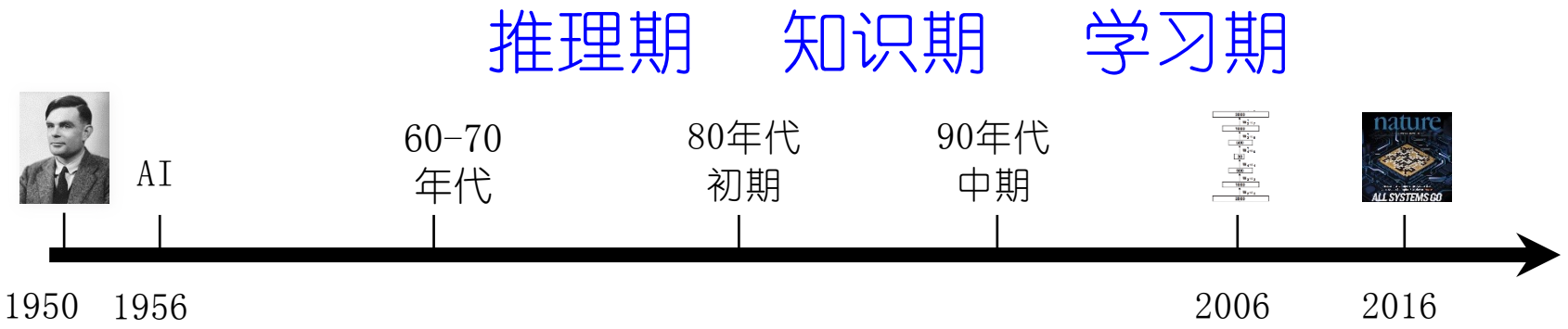
Homepage: www.lamda.nju.edu.cn/guolz

Email: guolz@nju.edu.cn

Announcements

- HW1 is due **Sep.30**, 24:00

机器学习从何而来


推理期
知识期
学习期
AI
60-70
年代
80年代
初期
90年代
中期
nature
ALL SYSTEMS GO
1950
1956
2006
2016

机器学习（Machine Learning）

机器

学习

机器学习 ??

机器学习（Machine Learning）

1）人类学习利用**经验**不断提高性能

2）机器善于处理**数据**不断提高性能

- 能否把“经验”变成数据，让机器可以“模仿”人类进行学习？

- 机器学习：机器利用数据学习人类经验，不断提高性能的过程

机器学习（Machine Learning）

经典定义：利用经验改善系统自身的性能 ［T. Mitchell 教科书， 1997］

经验 → 数据

随着该领域的发展，目前主要研究智能数据分析的理论和方法，并已成为智能数据分析技术的源泉之一

大数据时代

大数据 ≠ 大价值

智能
数据分析

机器
学习

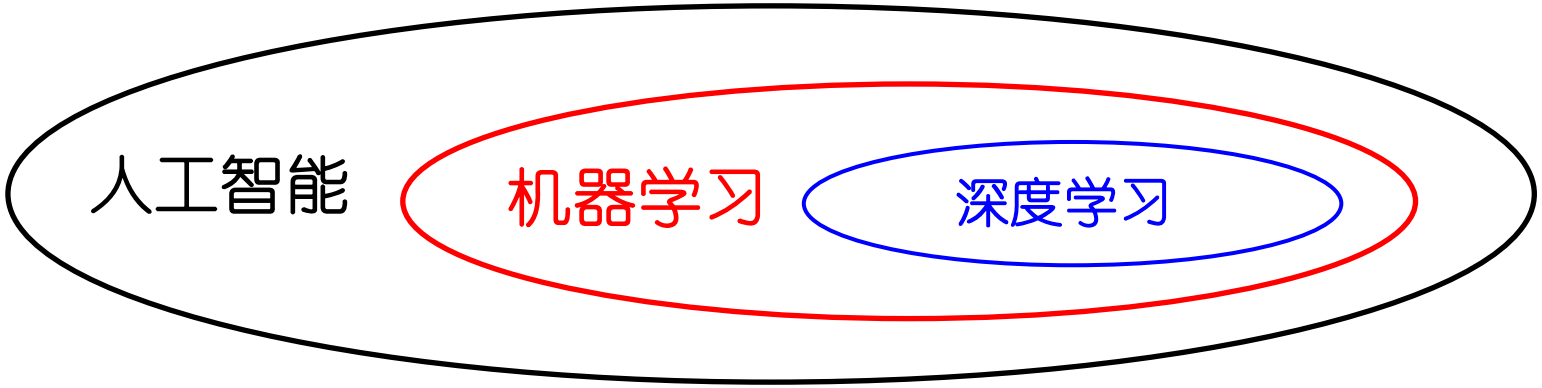
机器学习（Machine Learning）

机器学习是人工智能的核心研究领域（之一）

今天的“人工智能热潮”

正是由于机器学习、尤其深度学习技术取得了巨大进展

基于大数据、大算力发挥出巨大威力



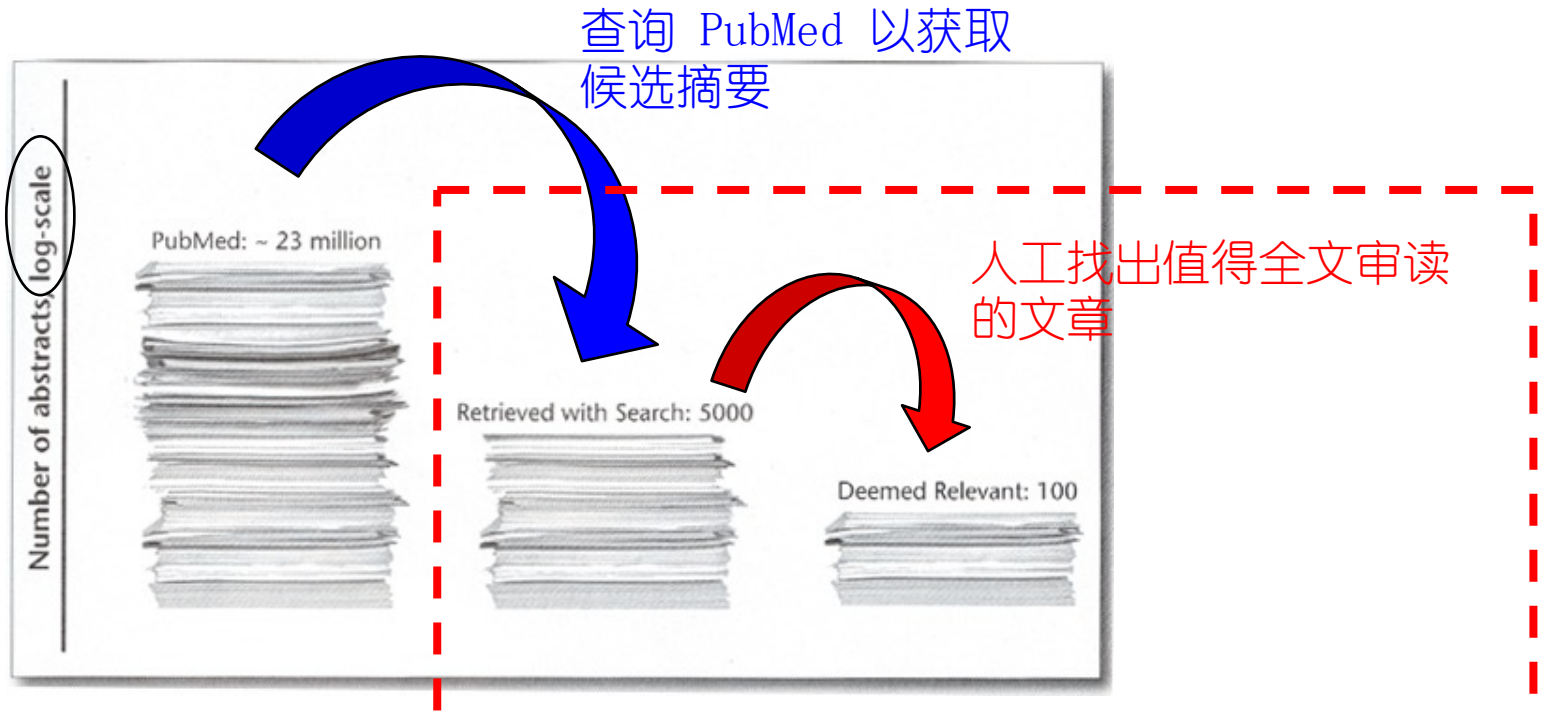
机器学习（Machine Learning）

机器学习
（Machine Learning）

看两个例子 ⇨

一个例子：“文献筛选”

在“循证医学”（evidence-based medicine）中，针对特定的临床问题，先要对相关研究报告进行详尽评估



出自 [C. Brodley et al., AI Magazine 2012]

“文献筛选”

在一项关于婴儿和儿童残疾的研究中，美国Tufts医学中心筛选了约33,000 篇摘要

尽管Tufts医学中心的专家效率很高，对每篇摘要只需 30 秒钟，但该工作仍花费了 250 小时

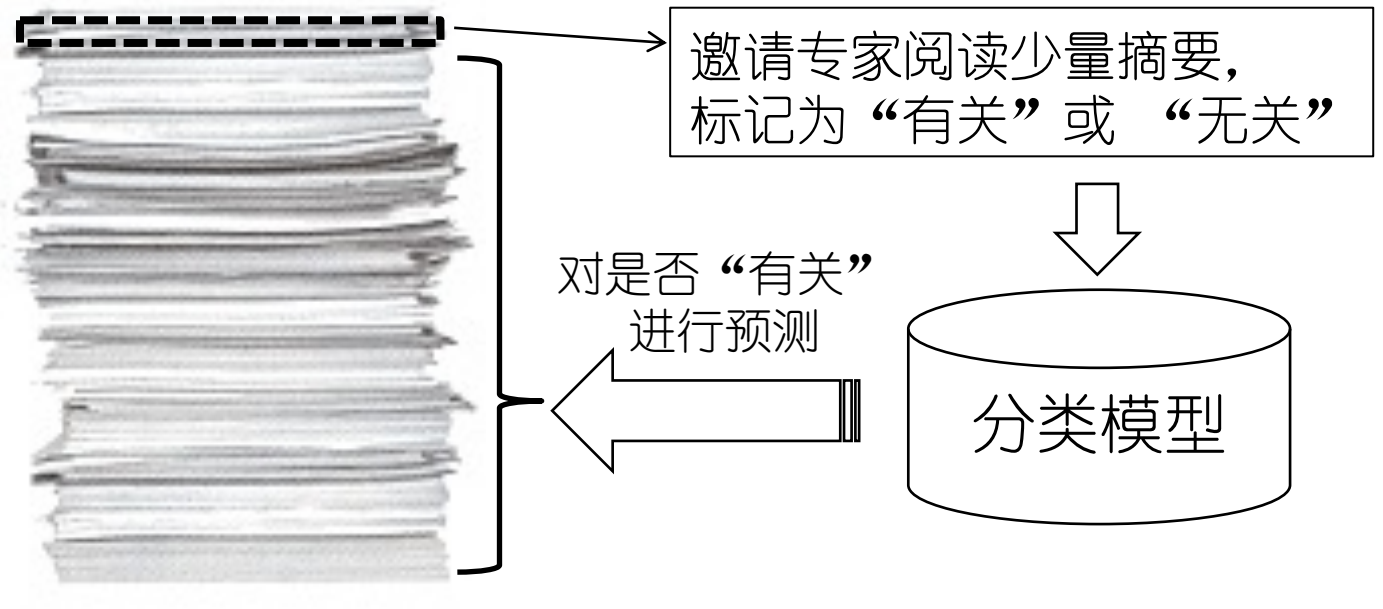
a portion of the 33,000 abstracts

每项新的研究都要重复这个麻烦的过程！

需筛选的文章数在不断显著增长！

“文献筛选”

为了降低昂贵的成本，Tufts医学中心引入了机器学习技术



人类专家只需阅读 50 篇摘要，系统的自动筛选精度就达到 93%

人类专家阅读 1,000 篇摘要，则系统的自动筛选精度度达到 95%

（人类专家以前需阅读 33,000 篇摘要才能获得此效果）

“画作鉴别”

画作鉴别(painting authentication)：确定作品的真伪


勃鲁盖尔（1525-1569）的作品？



梵高（1853-1890）的作品？

该工作对专业知识要求极高

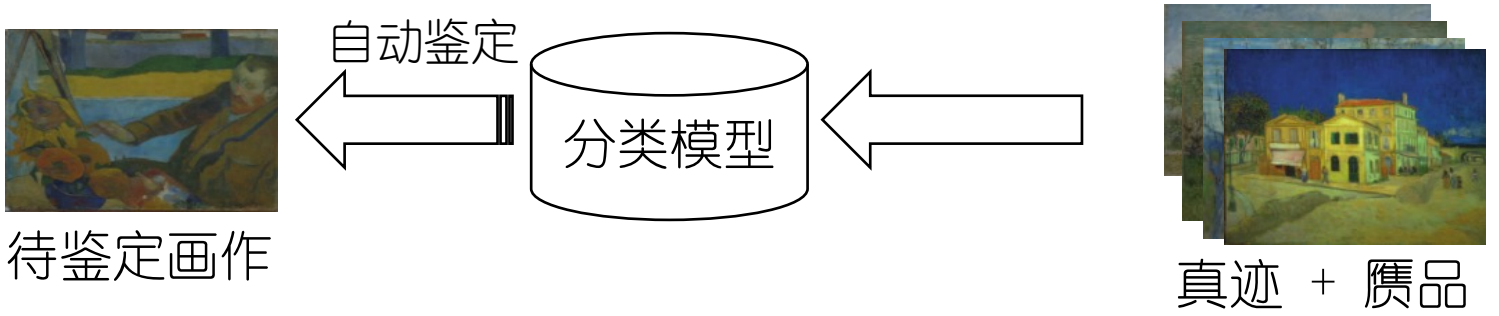
- 具有较高的绘画艺术修养
- 掌握画家的特定绘画习惯

只有少数专家花费很大精力才能完成分析工作！

很难同时掌握不同时期、不同流派多位画家的绘画风格！

“画作鉴别”

为了降低分析成本，**机器学习**技术被引入



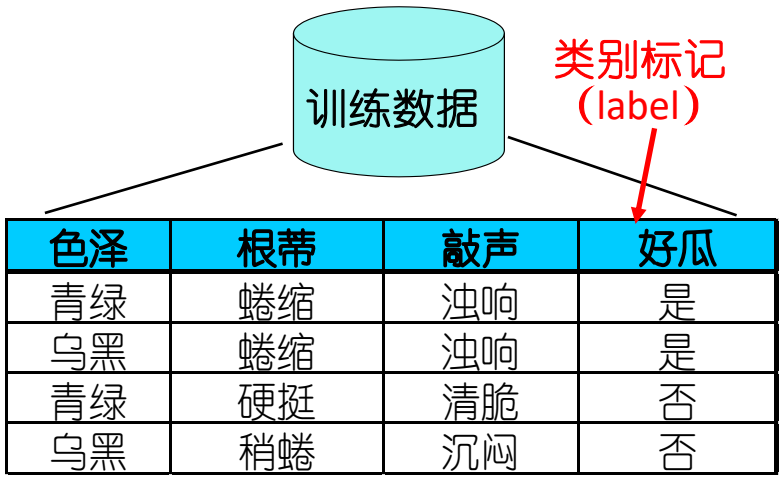
Kröller Müller美术馆与Cornell等大学的学者对82幅梵高真迹和6幅赝品进行分析，自动鉴别精度达 **95%** [C. Johnson et al., 2008]

Dartmouth学院、巴黎高师的学者对8幅勃鲁盖尔真迹和5幅赝品进行分析，自动鉴别精度达 **100%** [J. Hughes et al., 2009][J. Mairal et al., 2012]

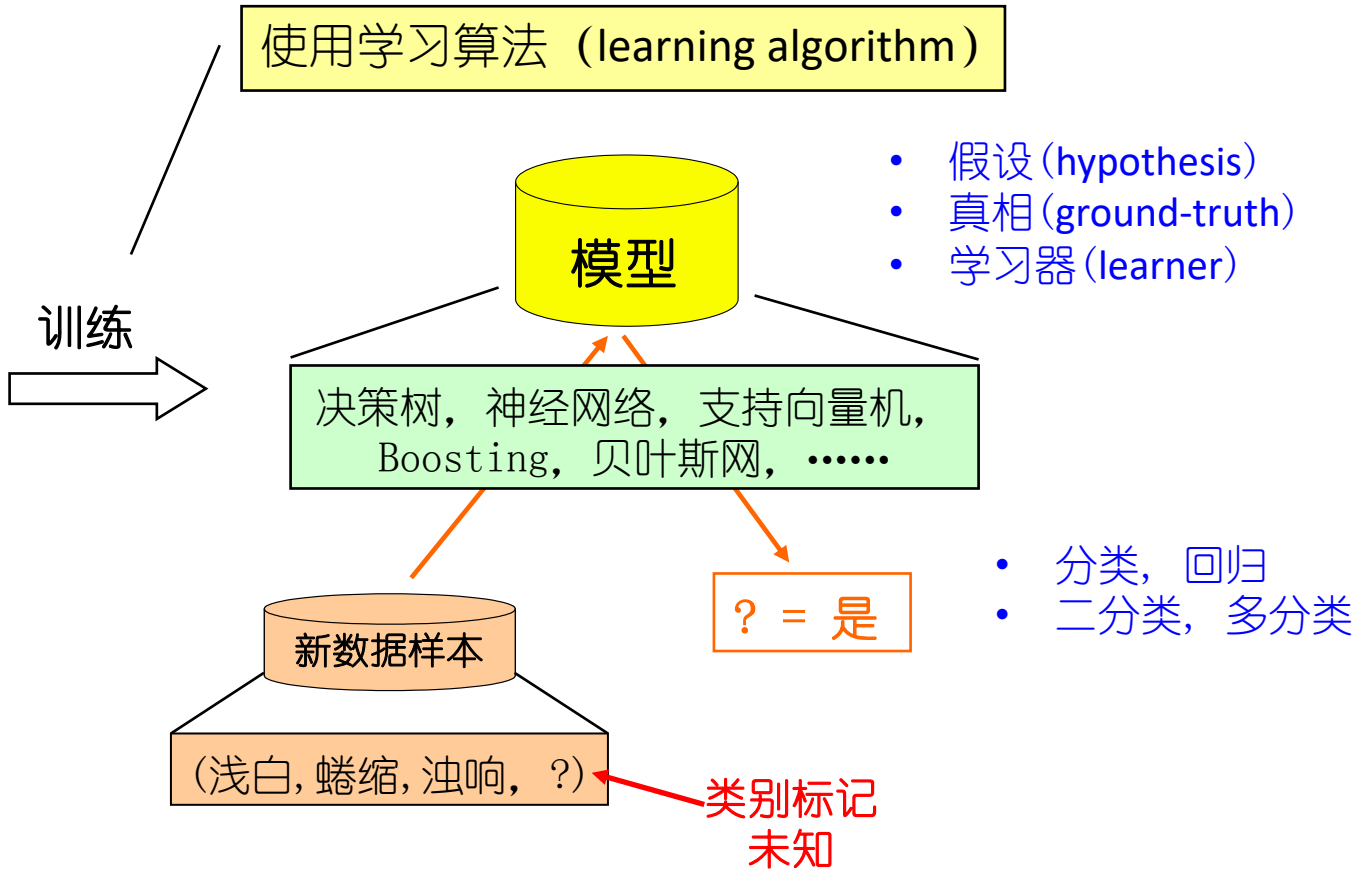
（对用户要求低、准确高效、适用范围广）

典型的机器学习过程

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)
- 强化学习(reinforcement learning)



- 数据集：训练集、测试集
- 示例(instance), 样例(example), 样本(sample)
- 属性(attribute), 特征(feature)
- 属性值
- 属性空间, 样本空间, 输入空间
- 特征向量(feature vector)
- 标记空间, 输出空间



- 假设(hypothesis)
- 真相(ground-truth)
- 学习器(learner)

- 分类，回归
- 二分类，多分类

潜在意义

横：输入（数据） -> 输出（标记）

纵：
历史（数据）
->
未来（数据）

特征　　标记

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否
1	青绿	蜷缩	沉闷	?

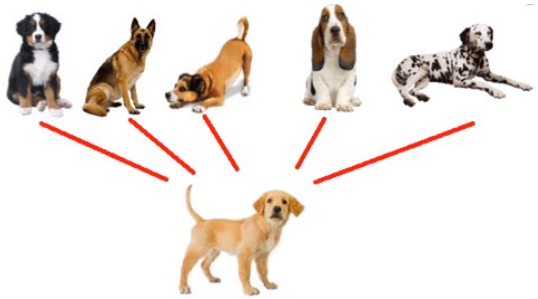
训练集

测试集

机器学习：面向未来的技术

学习的目标

机器学习技术的根本目标就是



模型具有泛化能力！

“简单理解”：应对未见样本的预测能力

未来不可知，依靠“合理假设”，利用历史数据估计模型泛化能力

如：历史和未来数据来自于相同的分布
(I.I.D. 假设)

任务 数据 目标 算法

Machine Learning = task + data + objective + algorithm

--Tom Mitchell

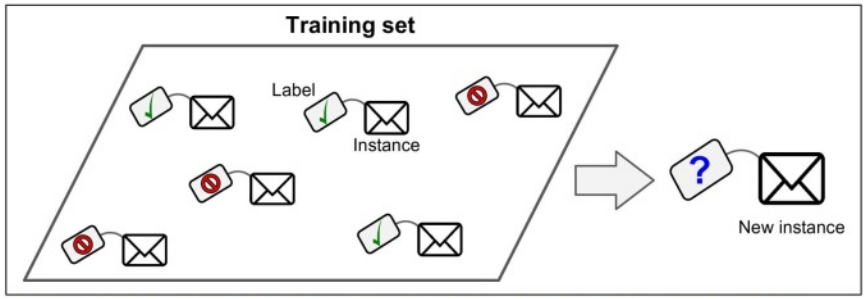
Learning = **Task** + Data + objective + Algorithm

- 任务通常可以指学习一个从特征空间到类别空间的映射 $f: X \to Y$
- 以西瓜任务为例：
  - X：西瓜的特征，如颜色、根蒂的蜷缩程度、敲击的声音等
  - Y：西瓜是好还是坏
- f 通常来自一个约定好的空间$\mathcal{F}$，即$f \in \mathcal{F}$

色泽	根蒂	敲声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否

Learning = **Task** + Data + objective + Algorithm

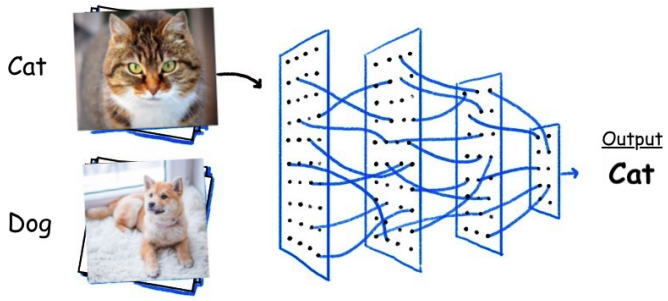
分类任务： Y包含若干离散的属性值
二分类： {0, 1}， K分类{0, 1, … K}



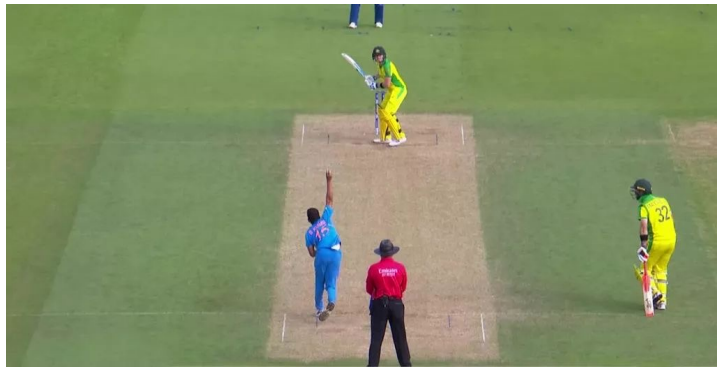
垃圾邮件分类

号码类型	主叫通话次数	主叫率	主叫外地联系人个数	主叫外地联系地个数	主叫通话频率	主叫通话时长	回拨率	联系人/通话次数比例
正常号码	少，平均每天主叫通话6次	适中	适中，平均每天与4人联系	极少	极低	正常，通话时长在50s左右	正常	正常
诈骗/骚扰电话	多，平均每天主叫通话54次	极高，主叫比例接近100%	极多，平均每天与51人联系	较高，每天约与9个不同的城市联系	高，约1h通话11次	未受骗用户识别诈骗需要一定时间，通话时长在25s左右	几乎为0，说明该类号码基本不会被回拨	几乎为1，说明该类号码对多个号码进行单次拨号

欺诈检测



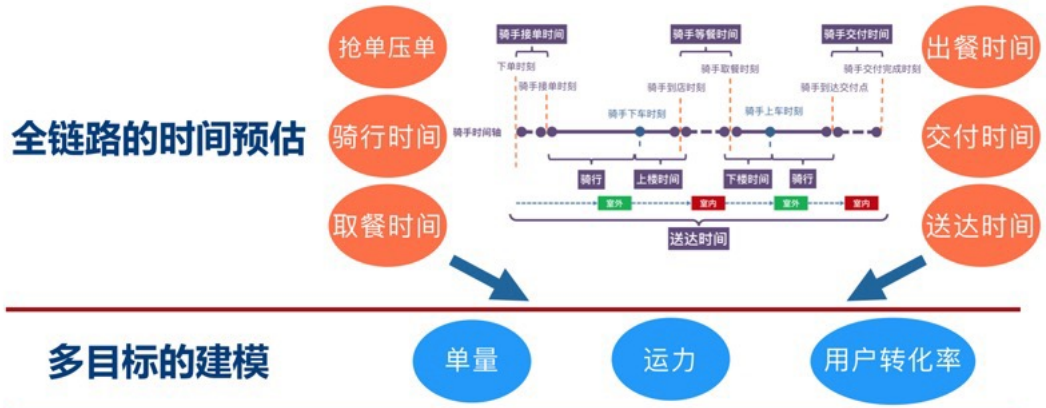
动物识别



动作分类

Learning = **Task** + Data + objective + Algorithm

回归任务： Y通常是实数值



外卖送达的时间

方向盘旋转幅度

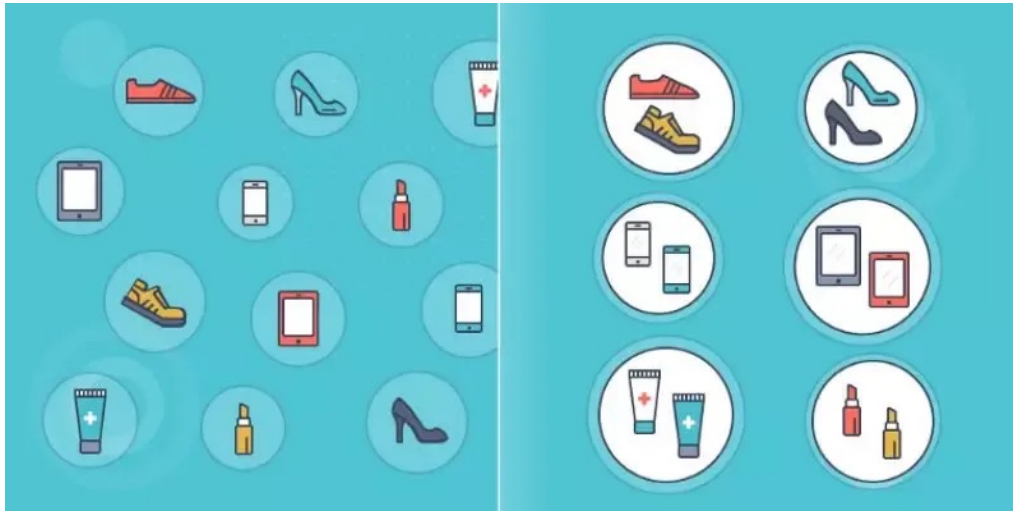
油门幅度

刹车幅度

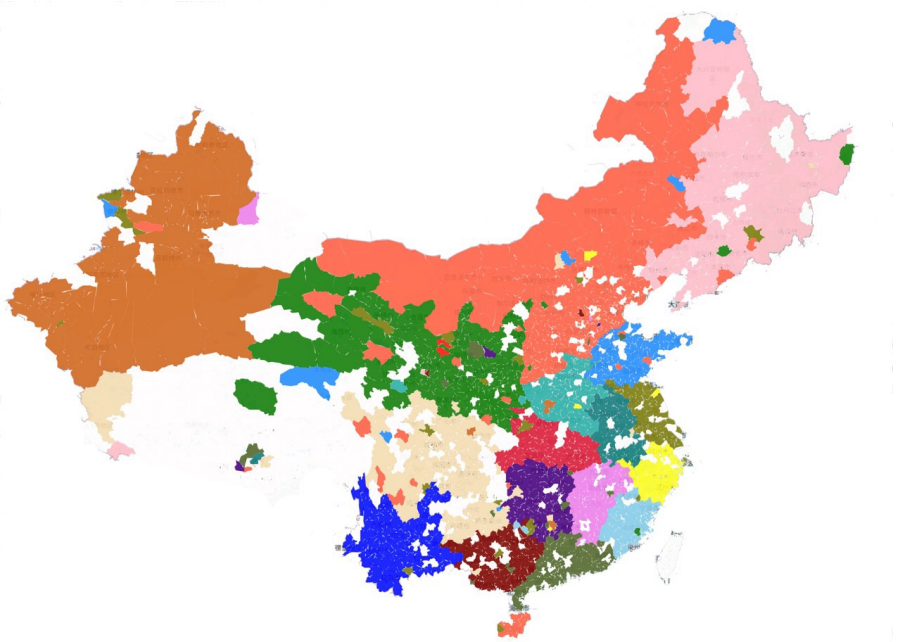
自动驾驶：方向盘旋转的幅度、油门幅度、刹车幅度

Learning = **Task** + Data + objective + Algorithm

聚类任务：把数据集中的样本划分为若干个子集



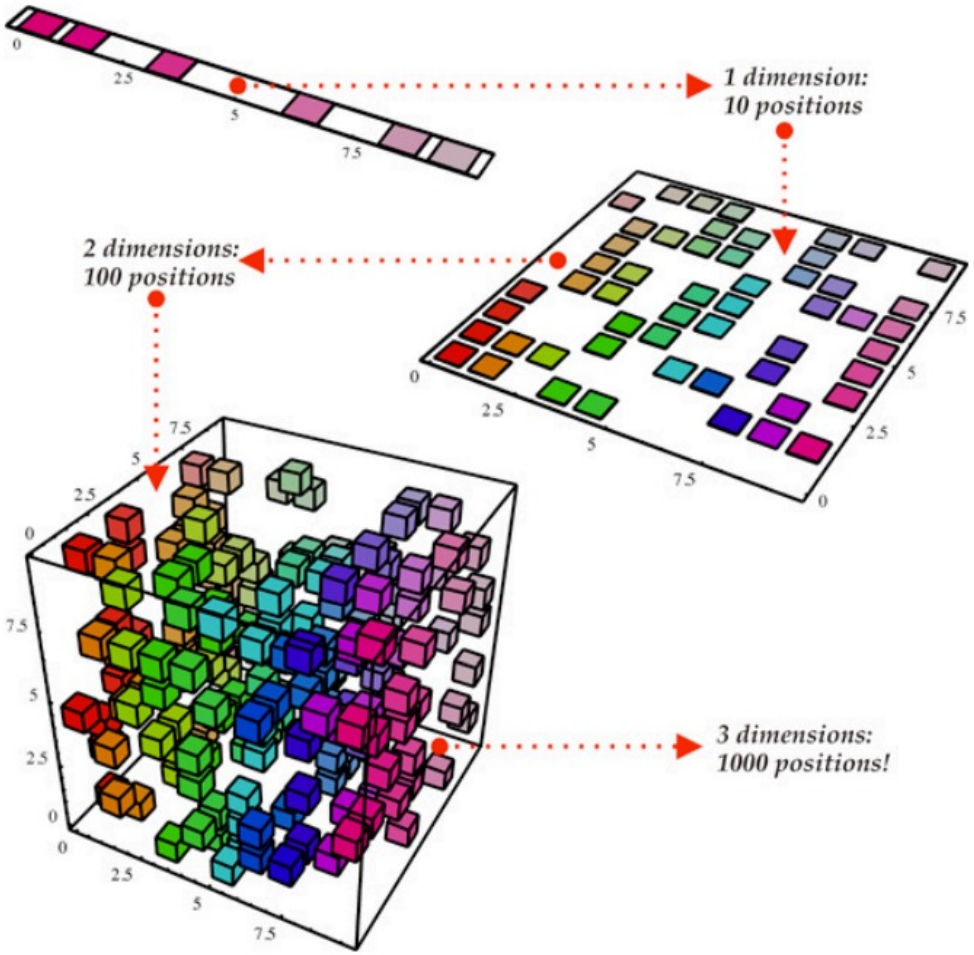
推荐系统：商品、用户聚类



区域（县）投资网络社区检测

Learning = **Task** + Data + objective + Algorithm

降维任务：降低特征维度，用更少的特征表示数据

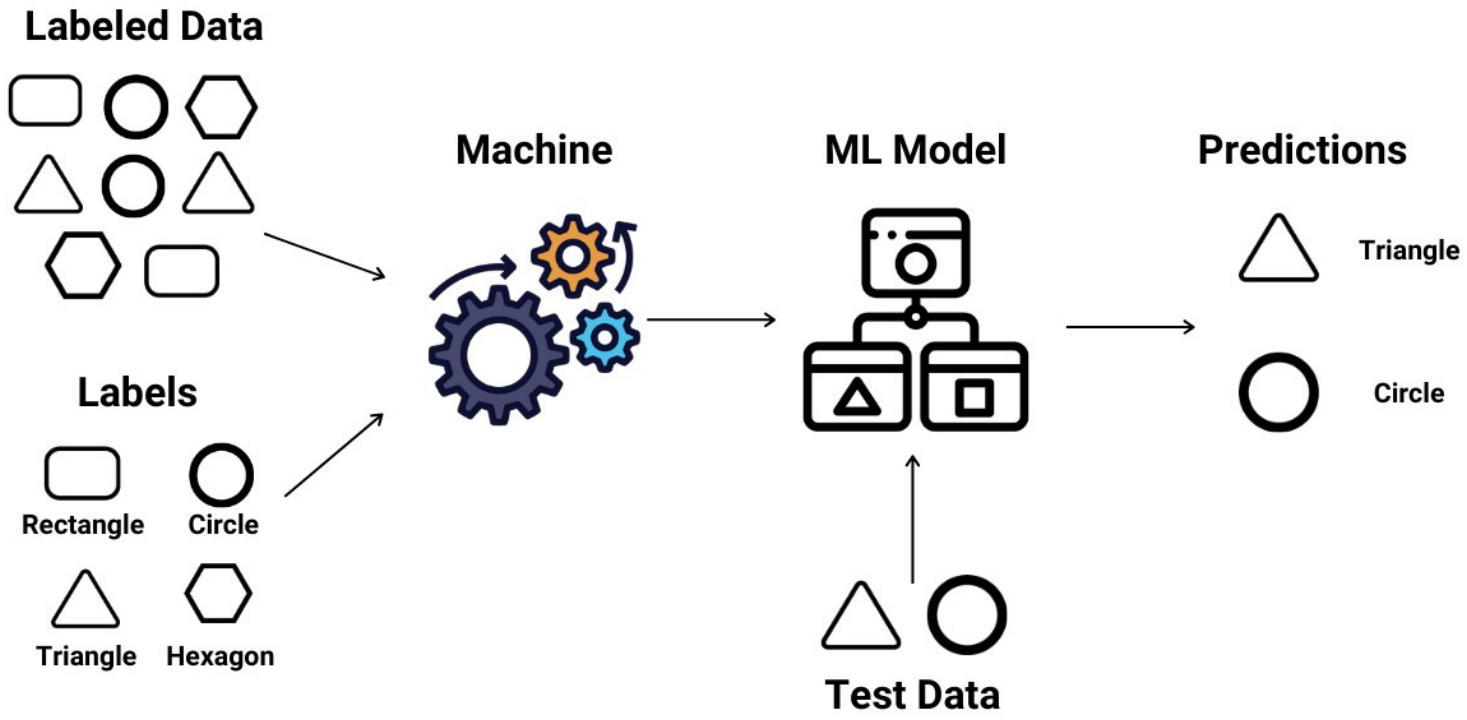


Learning = Task + **Data** + objective + Algorithm

- **训练数据**：训练机器学习模型的基础资源
  - 监督学习：$D_{tr} = \{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\}$
  - 无监督学习：$D_{tr} = \{x_1, x_2, \cdots x_n)\}$
  - 半监督学习：$D_{tr} = \{(x_1, y_1), (x_2, y_2), \cdots (x_t, y_t), x_{t+1}, \cdots x_n\}$
- **测试数据**：评估模型的性能，在训练过程中无法见到
  - $D_{te} = \{(x_1, y_1), (x_2, y_2), \cdots (x_m, y_m)\}$
- **验证数据**：用于训练过程中辅助评估模型的性能

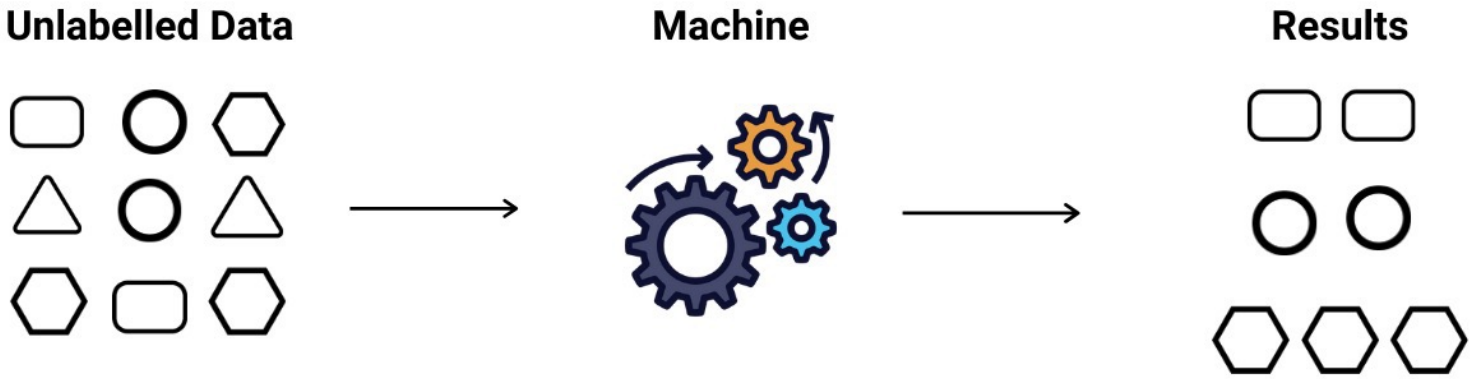
Learning = Task + **Data** + objective + Algorithm

监督学习：所有训练样本均有对应的标注



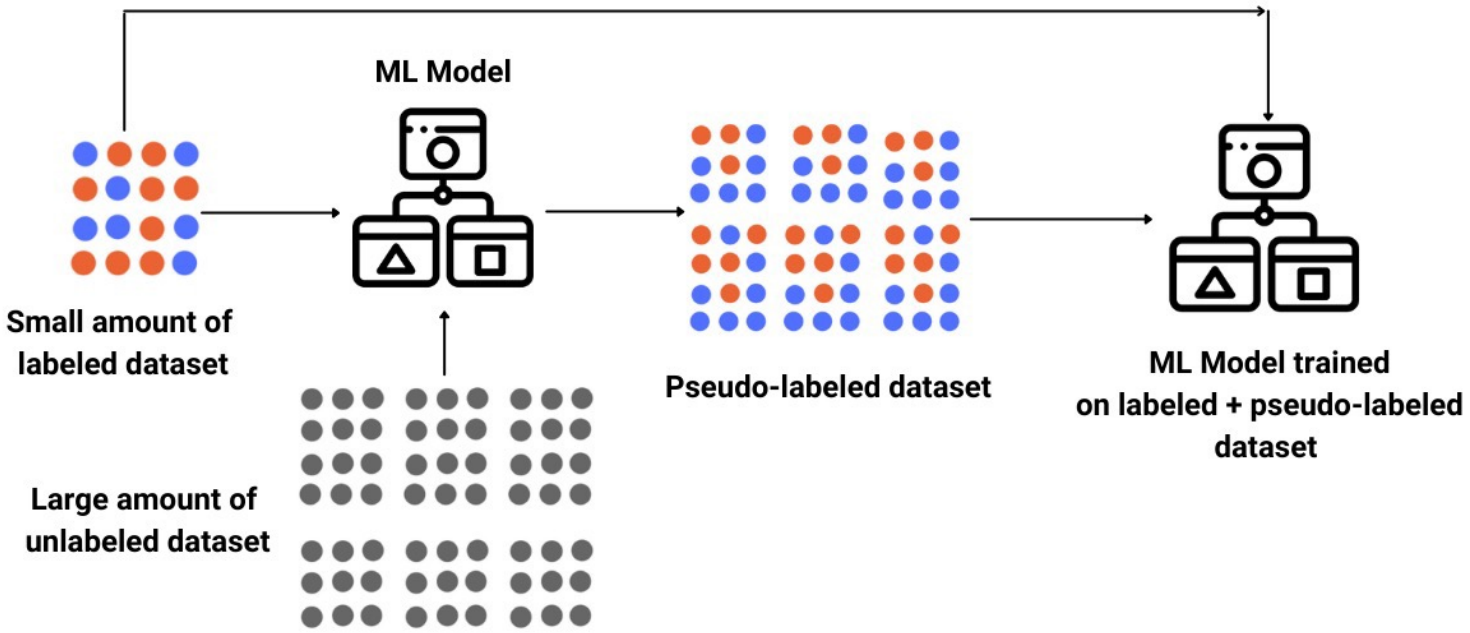
Learning = Task + **Data** + objective + Algorithm

无监督学习：所有训练样本均没有标注



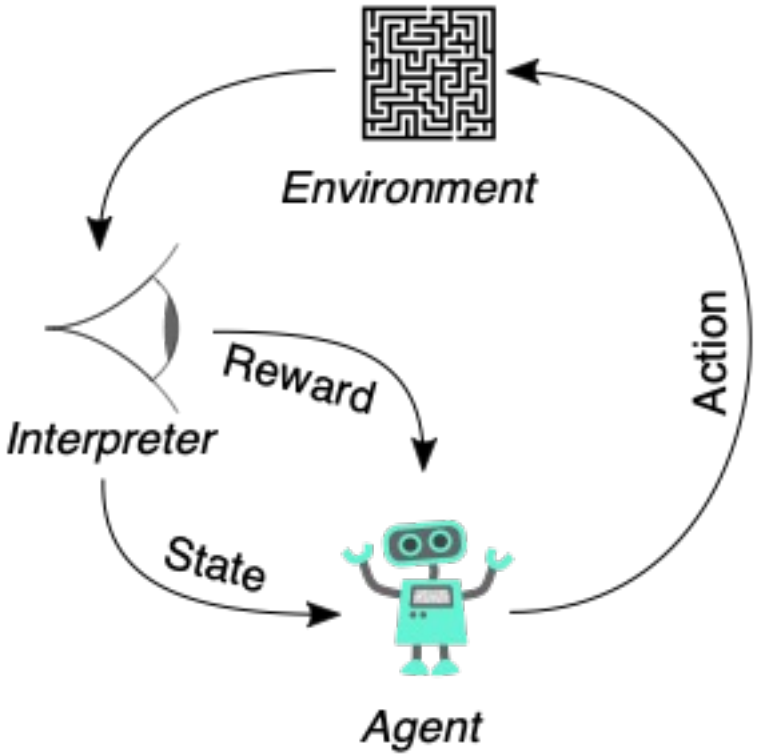
Learning = Task + **Data** + objective + Algorithm

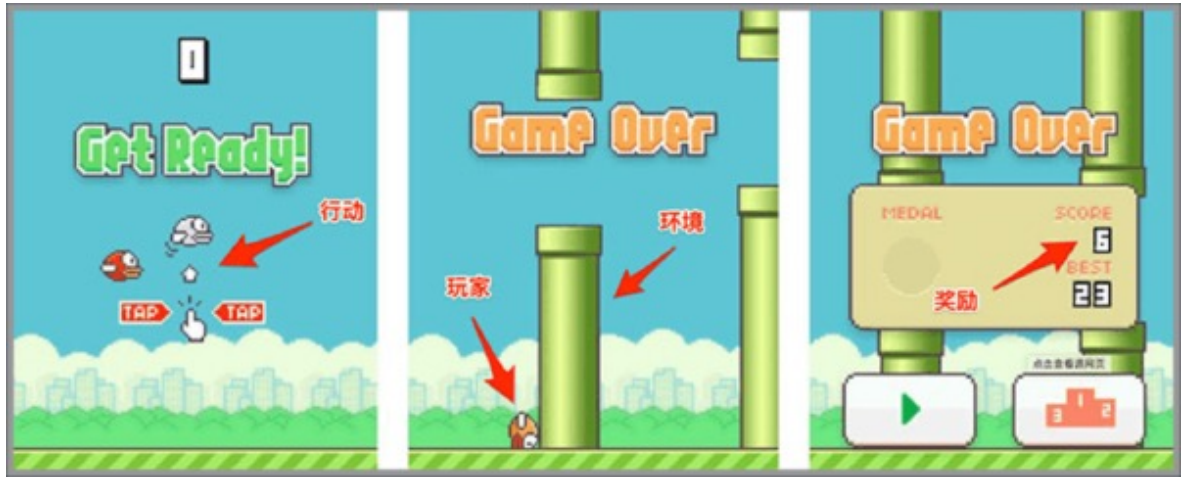
半监督学习：部分数据有标注，部分数据无标注



Learning = Task + **Data** + objective + Algorithm

强化学习：标记信息延迟的监督学习问题





Learning = Task + Data + **objective** + Algorithm

- **评价指标/损失函数**：反映了模型f的性能好坏，用于指导模型训练
- 分类问题：
  - 正确率：模型预测正确的概率

$$P(f(x) = y) \qquad \sum_{i=1}^{m} \mathbb{I}(f(x_i) = y_i)$$

- 回归问题：
  - 均方误差（Mean Squared Error）

$$\sum_{i=1}^{m} (f(x) - y)^2$$

Learning = Task + Data + objective + **Algorithm**

机器学习算法$\mathcal{A}$

- 输入：训练数据集D_{tr}，评价指标$M(f)$/损失函数$loss(f)$，f的函数空间$\mathcal{F}$
- 输出：学得的模型f

$$\mathcal{A}: \mathcal{F} \times \mathcal{M} \times \mathcal{S} \rightarrow f$$

学习算法运行的过程称为模型的训练过程

即，在所有可能的f组成的空间中进行**搜索**的过程

经验风险最小化（empirical risk minimization）

学习目标：在空间$\mathcal{F}$中寻找能够在整个数据分布上表现最好的模型 f

$$\min_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim D}[loss(f(x),y)]$$ 泛化风险

现实任务中，无法得知完整的数据分布，只能获取训练数据

假设所有训练样本都是独立地从这个分布中采样而得

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n} loss(f(x_i),y_i)$$ 经验风险

例：房价预测

任务：训练机器学习模型，能够根据房子的面积预测出房价

训练数据

面积	房价
100	300
110	330
180	540

测试数据

面积	房价
140	?

- 假设f是线性函数

$$f = w * x$$

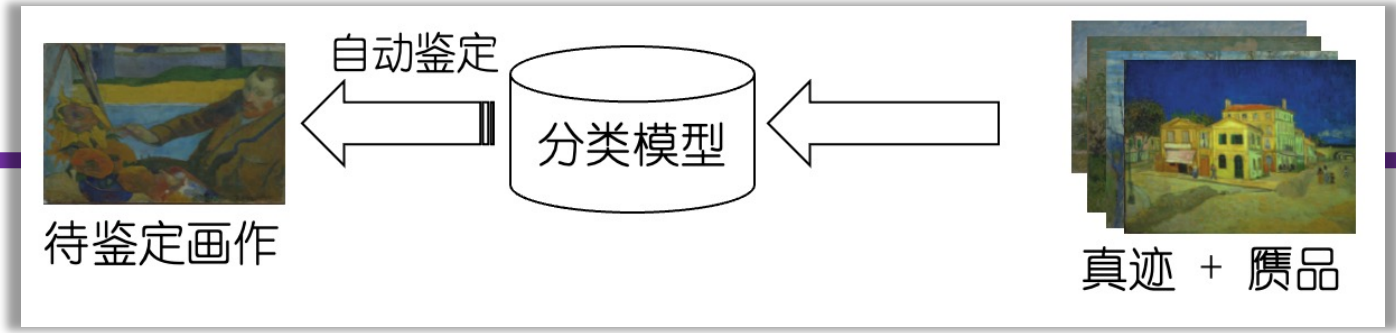
- Objective: 损失函数为均方误差

$$loss(f(x), y) = (y - f(x)^2$$

$$\min_w (300 - w * 100)^2 + (330 - w * 110)^2 + (540 - w * 180)^2$$

求得：$w = 3$

例：画作鉴别



假设f是线性函数：$f = w^{\mathrm{T}}X + b$

将像素转换为向量

0.2	-0.5	0.1	2.0
1.5	1.3	2.1	0.0
0	0.25	0.2	-0.3

w

56
231
24
2

x

1.1
3.2
-1.2

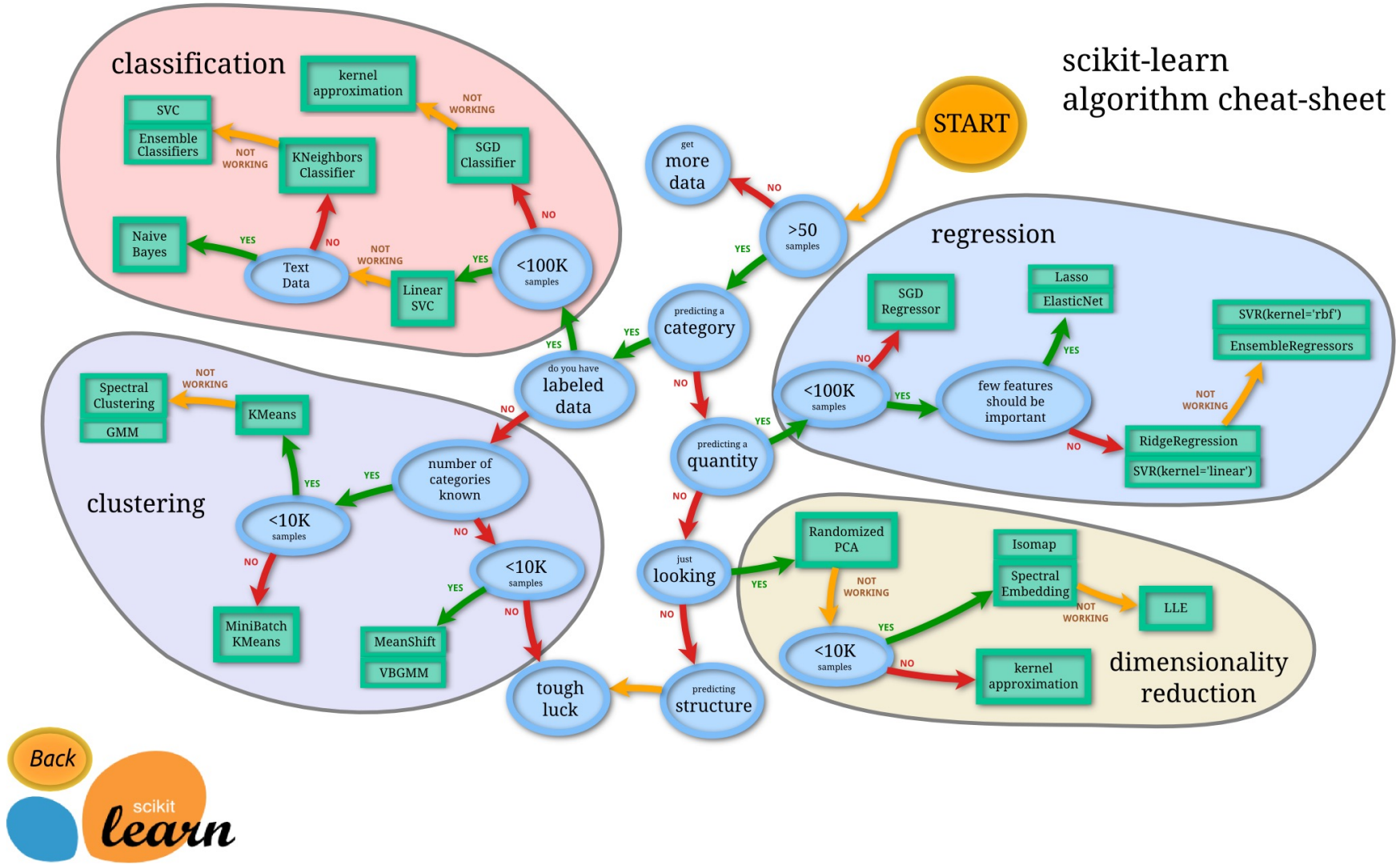
b

-96.8	梵高
437.9	勃鲁盖尔
61.95	其他人

机器学习常用算法

- 线性回归算法 (Linear Regression)
- 逻辑回归算法 (Logistic Regression)
- 支持向量机算法 (Support Vector Machine, SVM)
- k-近邻算法 (K-Nearest Neighbors, KNN)
- k-Means算法
- 决策树算法 (Decision Tree)
- 随机森林算法 (Random Forest)
- 朴素贝叶斯算法 (Naive Bayes)
- 神经网络 (Neural Network)
- …

机器学习常用算法


scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
<100K samples
Text Data
Naive Bayes
Linear SVC
SVC
Ensemble Classifiers
KNeighbors Classifier
kernel approximation
SGD Classifier
NOT WORKING
YES
NO
classification
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
few features should be important
RidgeRegression
SVR(kernel='linear')
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift
VBGMM
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
kernel approximation
Back
scikit learn

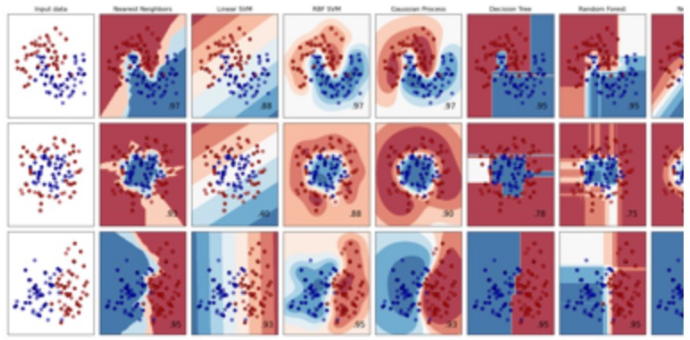
Sklearn

https://scikit-learn.org/

Classification

Identifying which category an object belongs to.

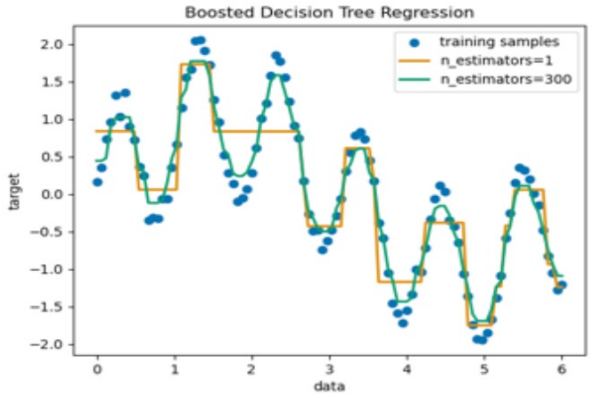
Applications: Spam detection, image recognition.
Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more...



Regression

Predicting a continuous-valued attribute associated with an object.

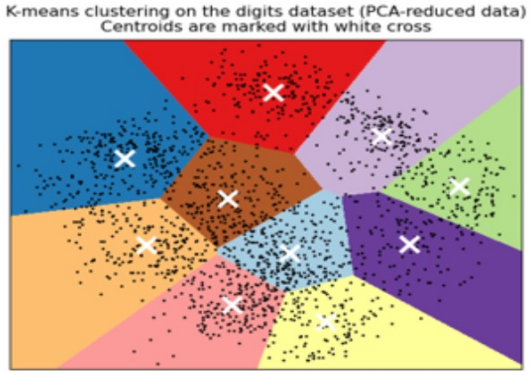
Applications: Drug response, Stock prices.
Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...



Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...



示例

训练阶段：

```
>>> import numpy as np
>>> fromsklearn.linear_model import LinearRegression
>>> X = np.array([100], [110], [180]])
>>> y = np.array([300], [330], [540]])
>>> reg = LinearRegression().fit(X, y)
```

测试阶段：

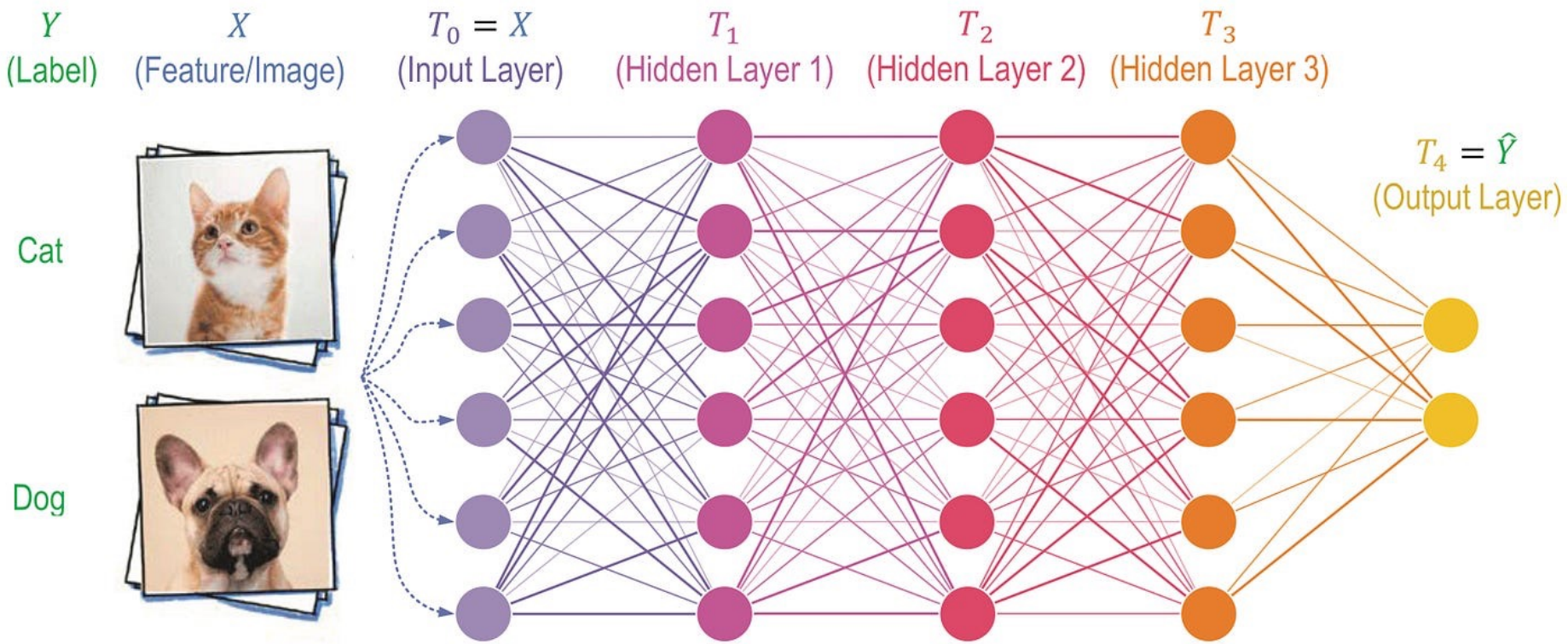
```
>>> reg.predict(np.array([[140]]))
```

示例

```
In [12]: import numpy as np

In [13]: from sklearn.linear_model import LinearRegression

In [14]: X = np.array([[100], [110], [180]])

In [15]: y = np.array([[300], [330], [540]])

In [16]: reg = LinearRegression().fit(X,y)

In [17]: reg.predict(np.array([[140]]))
Out[17]: array([[420.]])

In [18]:
```

深度学习

深度学习：通常是指以深度神经网络为模型的一类机器学习方法



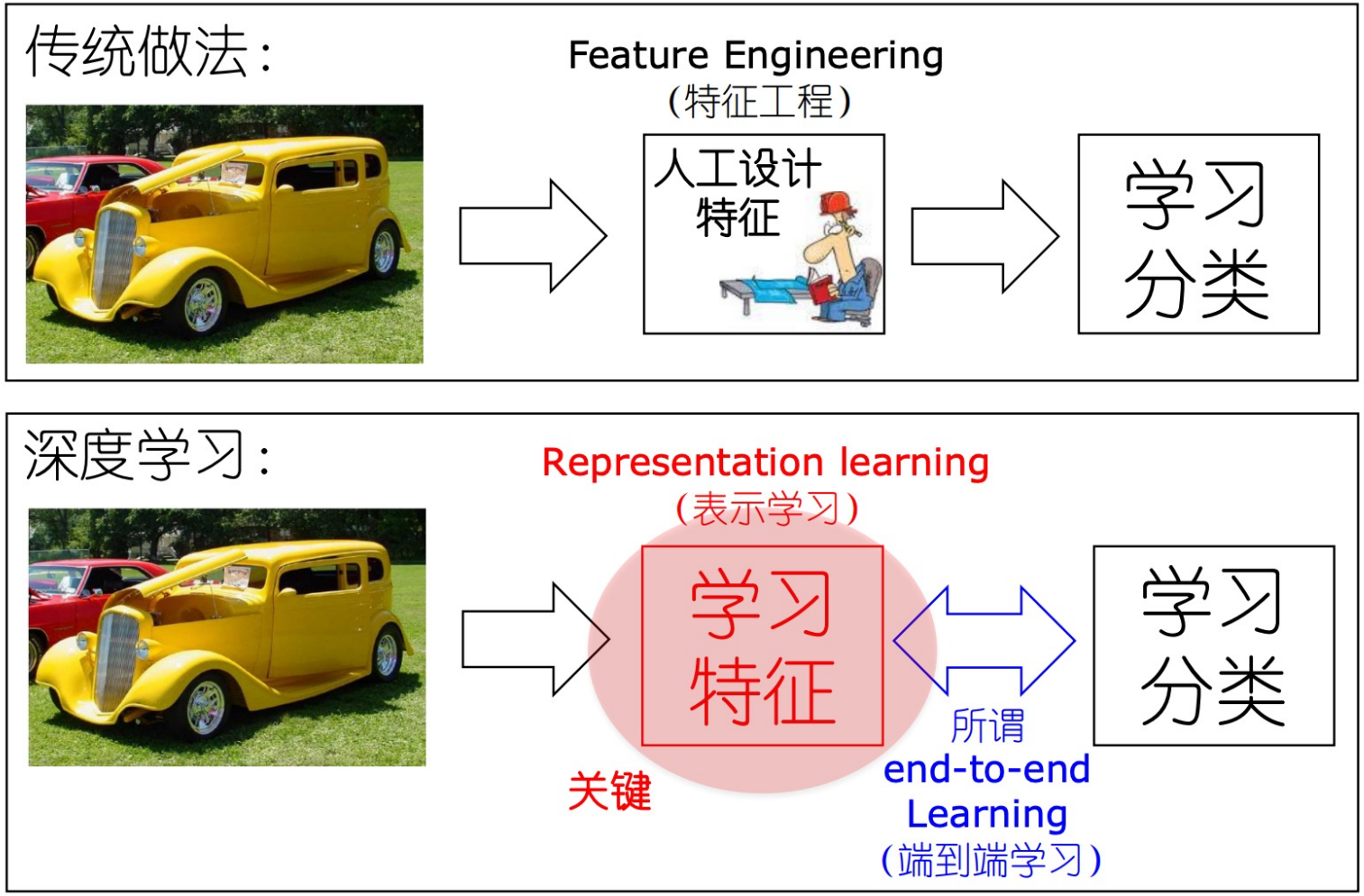
深度学习

- 2006年， Hinton发表了深度学习的science文章
- 2012年, Hinton 组参加ImageNet 竞赛, 使用 CNN 模型以超过 第二名10个百分点的成绩夺得当年竞赛的冠军
- 伴随着数据量的增加，和各种高性能计算设备的出现，深度学习在许多领域都取得了较好的效果

深度学习

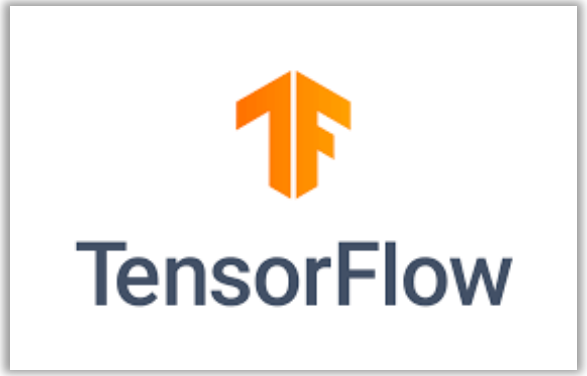
“深度学习三驾马车”
获2018年度图灵奖

深度学习最重要的作用：表示学习



深度学习常用框架


PyTorch


TensorFlow


scikit
learn
Machine Learning with Scikit-Learn


PaddlePaddle


MindSpore

泛化风险 vs. 经验风险

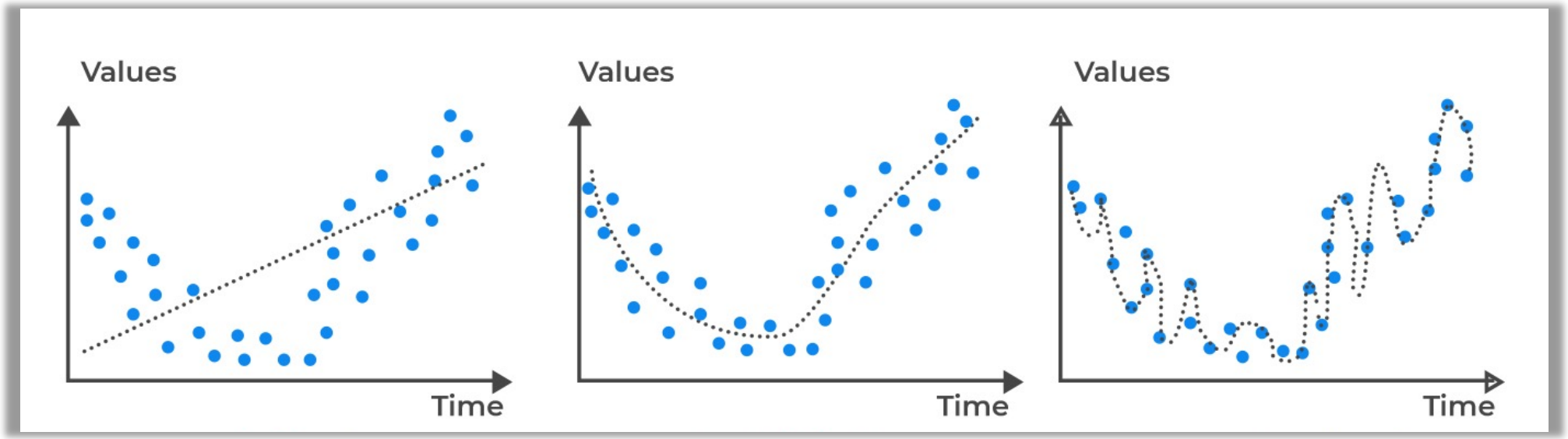
泛化误差：在“未来”样本上的误差

经验误差：在训练集上的误差，亦称“训练误差”

- 泛化误差越小越好
- 经验误差是否越小越好？

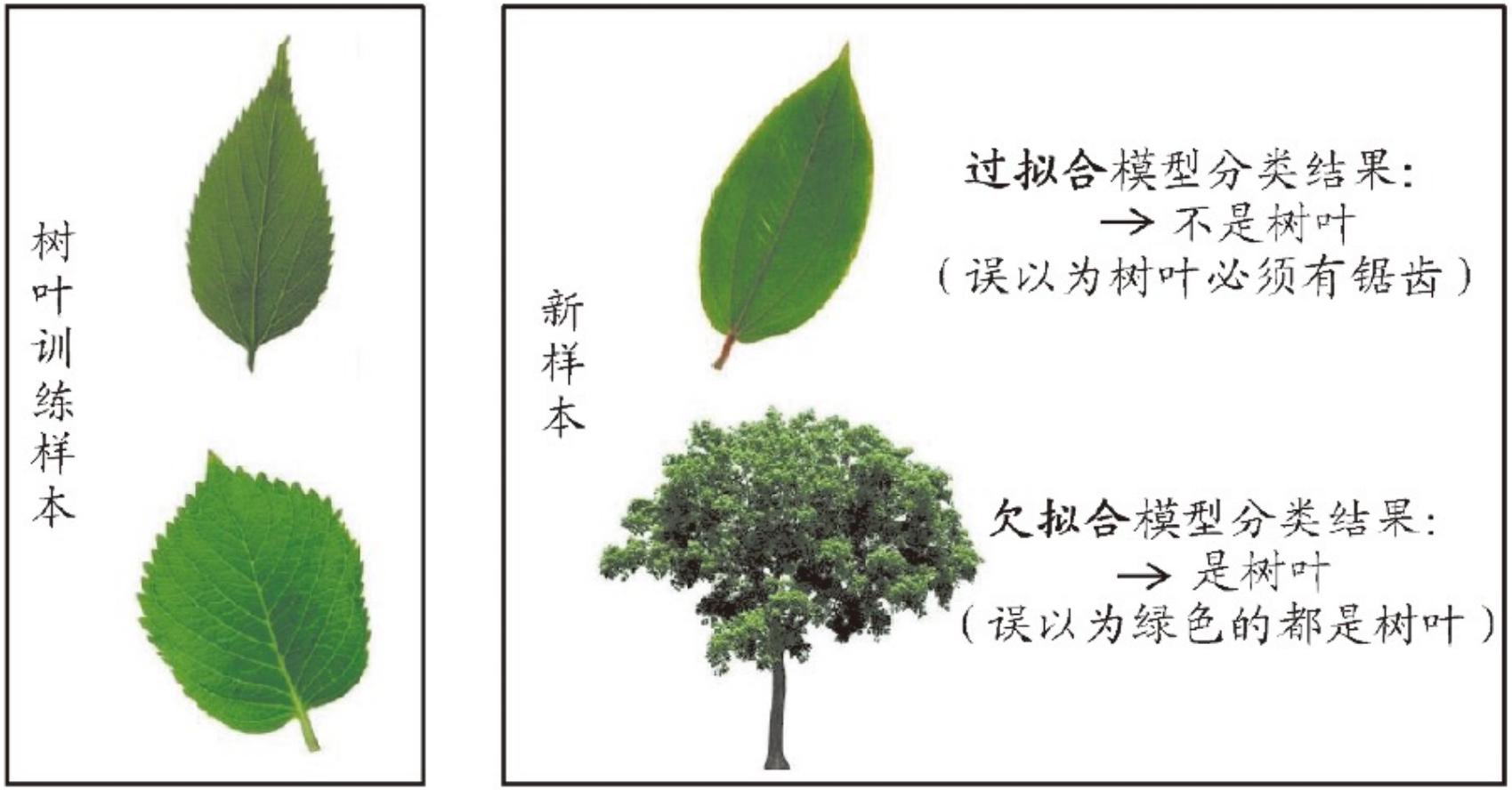
NO! 因为会出现“过拟合”(overfitting)

过拟合(overfitting) vs. 欠拟合(underfitting)



underfitting　　Good fit　　overfitting

过拟合(overfitting) vs 欠拟合(underfitting)



一般而言，训练样本越少，模型越复杂，越容易过拟合

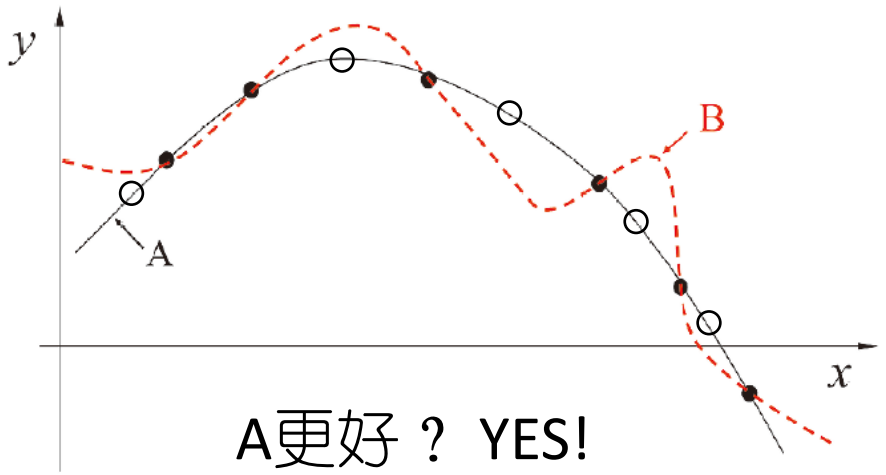
机器学习是无所不能的吗？

并非“一切皆可学”，例如：

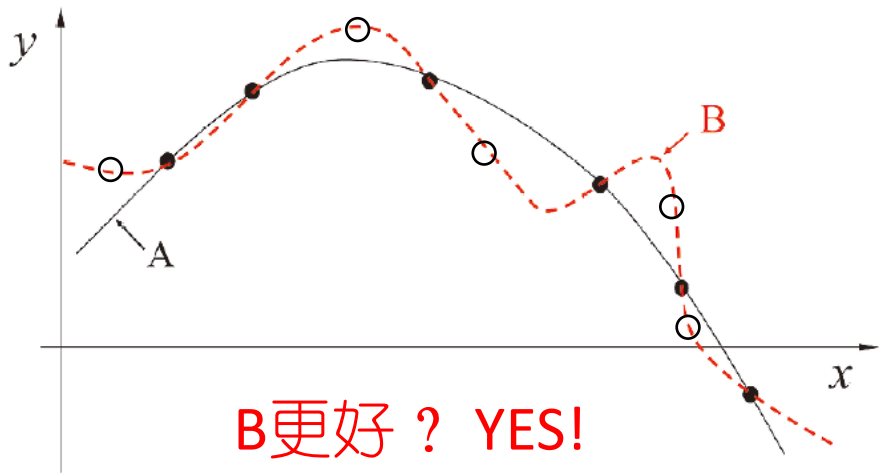
- 特征信息不充分
  - 例如，重要特征信息没有获得
- 样本信息不充分
  - 例如，仅有很少的数据样本

哪个算法更好？

黑点：训练样本；白点：测试样本



A更好？ YES!



B更好？ YES!

没有免费的午餐！

No Free Lunch 定理：一个算法$\mathcal{A}_1$若在某些问题上比另一个算法$\mathcal{A}_2$好，必存在另一些问题，$\mathcal{A}_2$比$\mathcal{A}_1$好。

NFL定理的寓意

NFL定理的重要前提：

所有“问题”出现的机会相同、或所有问题同等重要

实际情形并非如此；我们通常只关注自己正在试图解决的问题

脱离具体问题，空泛地谈论“什么学习算法更好”
毫无意义 ！

具体问题，具体分析！

现实机器学习应用

把机器学习的“十大算法”“二十大算法”都弄熟，

逐个试一遍，是否就“止于至善” 了？

NO ！

机器学习并非“十大套路”“二十大招数” 的简单堆积

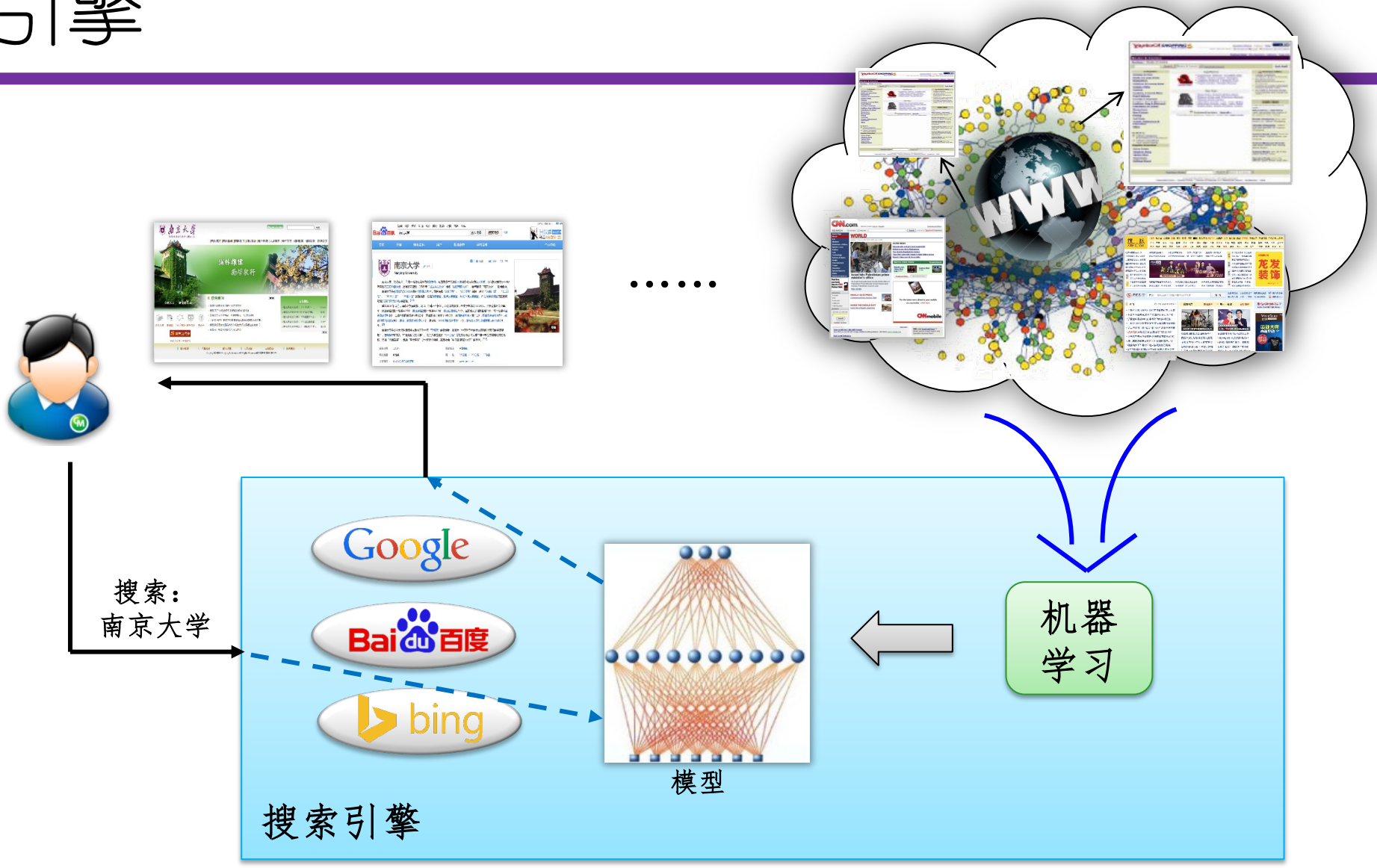
现实任务千变万化，

以有限的“套路”应对无限的“问题” ，焉有不败？

最优方案往往来自：**按需设计、度身定制**

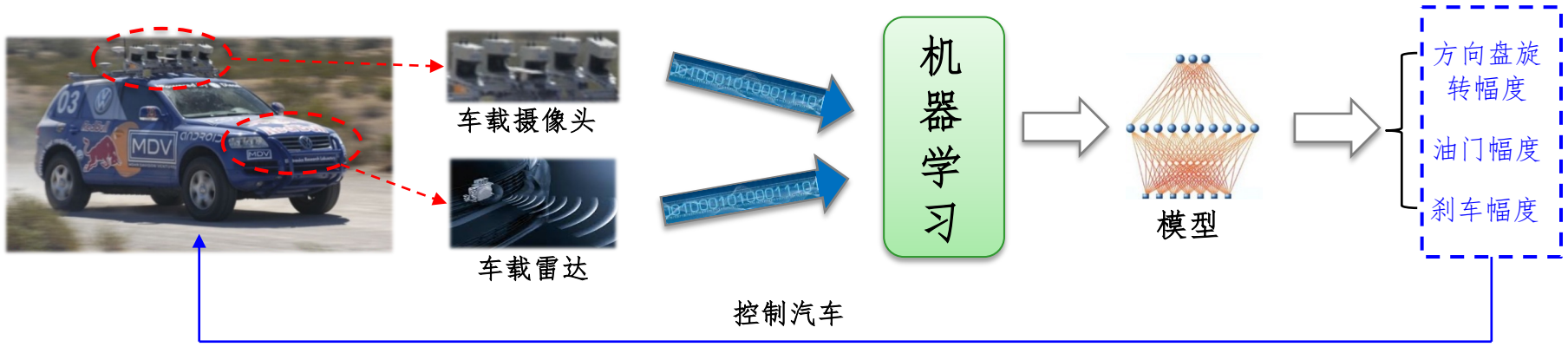
机器学习有哪些应用？

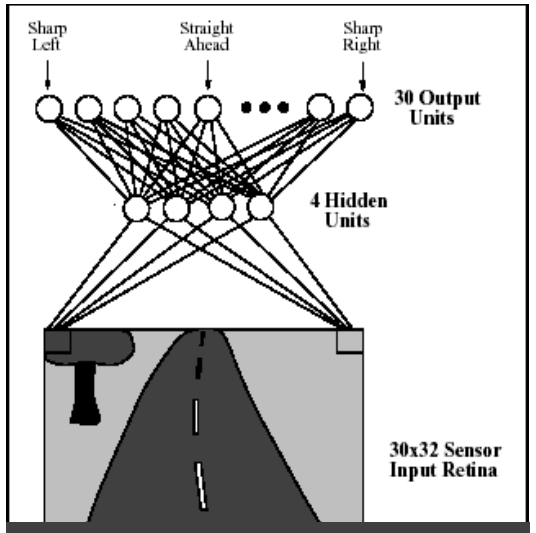
搜索引擎



机器学习技术正在支撑着各种搜索引擎

自动驾驶





美国在20世纪80年代就开始研究基于机器学习的汽车自动驾驶技术

DARPA Grand Challenge - 2004

荒野中的无人车竞赛

Google 无人驾驶汽车 - 2016

新加坡无人驾驶出租车 - 2016

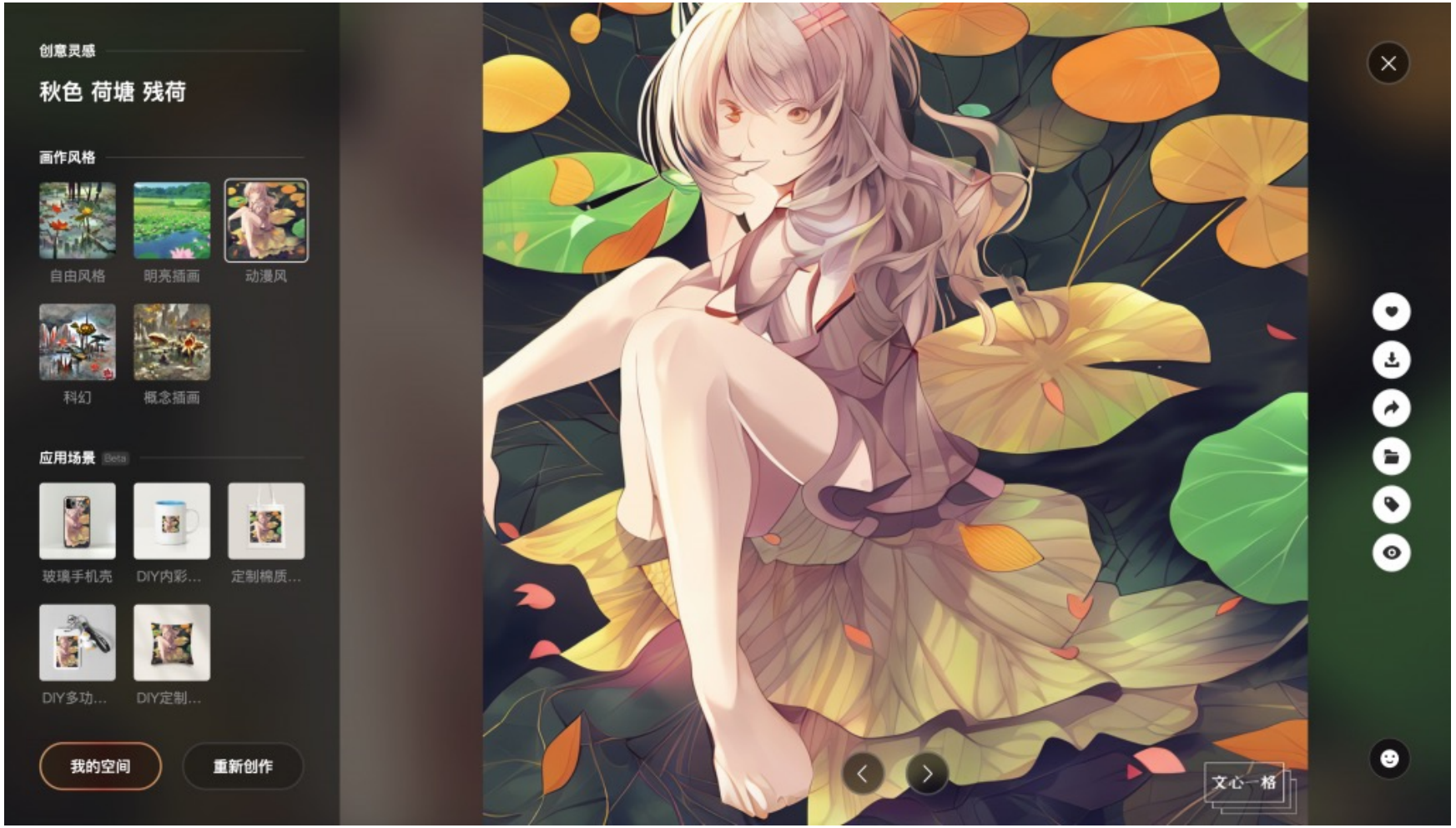
视频理解

计算机可初步理解视频内容

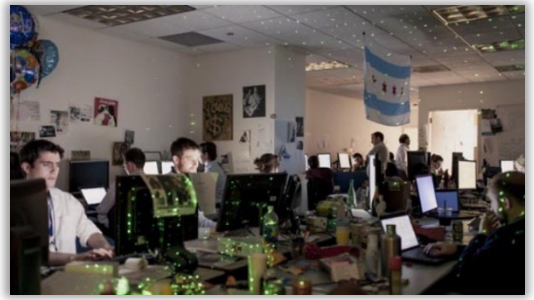
Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image
 A person riding a motorcycle on a dirt road.	 Two dogs play in the grass.	 A skateboarder does a trick on a ramp.	 A dog is jumping to catch a frisbee.
 A group of young people playing a game of frisbee.	 Two hockey players are fighting over the puck.	 A little girl in a pink hat is blowing bubbles.	 A refrigerator filled with lots of food and drinks.
 A herd of elephants walking across a dry grass field.	 A close up of a cat laying on a couch.	 A red motorcycle parked on the side of the road.	 A yellow school bus parked in a parking lot.

图像生成

根据文字描述生成相应图像



帮助奥马巴竞选



这个团队行动保密，定期向奥巴马报送结果；被奥巴马公开称为总统竞选的“核武器按钮”（“They are our nuclear codes”）

通过机器学习模型

◆ 个性化宣传

喜欢宠物？

奥巴马也有宠物！

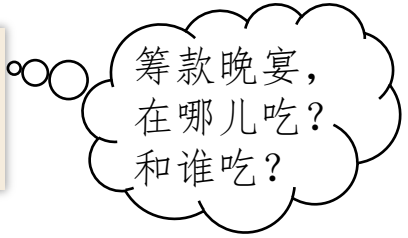
喜欢篮球？

奥巴马也是篮球迷！

◆ 广告购买

精准定位不同选民群体，建议购买冷门广告时段，广告资金效率比2008年提高14%

◆ 筹款



和乔治克鲁尼/奥巴马共进晚餐对于年龄在40-49岁的美西地区女性颇具吸引力…… 乔治克鲁尼为奥巴马举办的竞选筹资晚宴成功募集到1500万美元

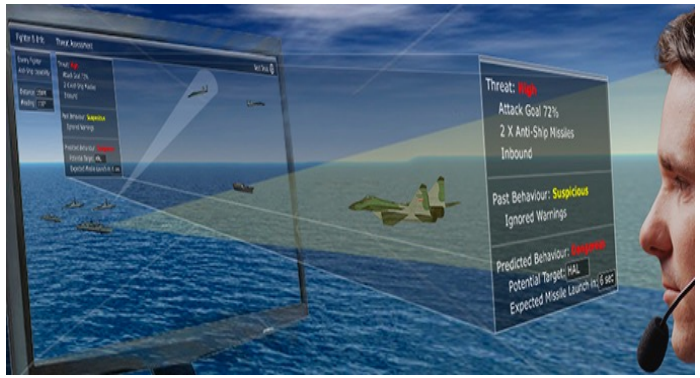
博弈AI

计算/预测出较高胜率的走法？

大量棋谱如何生成/利用

公开的计算难题，意义重大熟知的日常游戏，影响深远

战场战术（美）



眼镜蛇系统：

Coastal Battlefield Reconnaissance and Analysis (COBRA)

用于频海战斗舰，执行无人空中战术侦察。在两栖攻击之前，于海浪区和海滩区探测和定位雷区和障碍物

http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=1237&ct=2

http://www.navysbir.com/n15_1/N151-049.htm

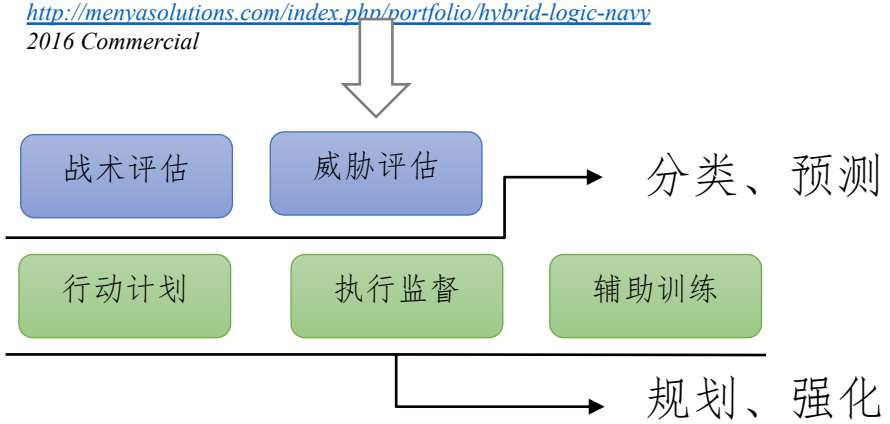
2015 US Navy Official

HybridLogic Navy:

一套自动的基于机器学习的代理，帮助人类和无人机理解战术状况，及时做出最佳决策，以对付海军作战中的威胁

http://menyasolutions.com/index.php/portfolio/hybrid-logic-navy

2016 Commercial



United States Navy
Fact File

AN/DVS-1 COASTAL BATTLEFIELD RECONNAISSANCE AND ANALYSIS - (COBRA)

The mission of the AN/DVS-1 Coastal Battlefield Reconnaissance and Analysis (COBRA) system is to conduct unmanned aerial tactical reconnaissance in the littoral battlespace for detection and localization of minefields and obstacles in the surf zone and beach zone prior to an amphibious assault. The COBRA airborne payload will be carried on the MQ-8 Fire Scout unmanned air system. This allows operators and other personnel to remain at a safe distance from the mine and obstacle belts and enemy direct and indirect fire. COBRA will be embarked in the Littoral Combat Ship (LCS) as part of the Mine Countermeasures (MCM) Mission Package (MP).

DESCRIPTION: The Coastal Battlefield and Reconnaissance (COBRA) program (Ref 1) is interested in technologies that facilitate automated target recognition (ATR) capabilities in aerial multi-spectral images for previously unseen environments and target types. Targets of interest include minefields and obstacles in various land and marine environments. Typically, ATR algorithms are developed offline (post-mission) using previously acquired test data sets. These algorithms often rely on supervised learning methods (Ref 2) that incorporate data from a limited set of test fields. When data is acquired from new environments, the algorithms often must be re-optimized to have good performance in that environment, as well as maintain performance in previously seen environments. The process for performing this offline re-optimization is often costly since it requires the efforts of expert analysts to assimilate data sets, determine target truth, analyze target features, train the ATR classifiers and evaluate performance.

There is a need for innovative methods that can 1) incorporate information from new data sets into the ATR system as they are acquired, and 2) re-optimize ATR algorithms quickly across all known environments, including those of newly acquired data. Online Machine Learning (OML) algorithms (Ref 3-5) can potentially be used to "learn" in the field based on operator-provided results without affecting prior performance. The information collected online can be used to refine the prediction hypothesis (classifier) used in the ATR algorithms. In addition, the information may provide input for automated methods of optimizing ATR performance across all known data sets.

The proposed effort will develop innovative OML algorithms for ATR that can incorporate human operator decisions to optimize probability of detection and probability of false alarm performance in new environments and for new target types. These algorithms will be integrated into mission and post-mission analysis systems in which operators review acquired images. The algorithms will be implemented as object-oriented C++ code for insertion into the operator systems. Development of the online learning algorithms must be combined with identification of how the operator will interact with them to provide updated decision information. Robust optimization of the ATR algorithms may be performed post-mission, which will require the development of separate software tools for processing historical data sets. The OML algorithms and optimization tools developed in this effort will reduce program costs by minimizing the time required for optimizing ATR algorithms to perform well in unseen operational environments.

自动目标识别、监督学习以及在线学习技术被作为核心技术并多次提及

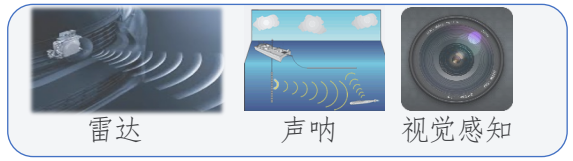
战场战术（英）

无人侦察快艇：

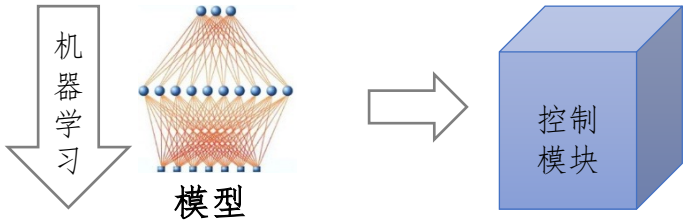
无人控制的情况下以50公里时速追踪快速目标并自动避障，进行跟踪、监视和间谍活动，或者用于海岸巡逻

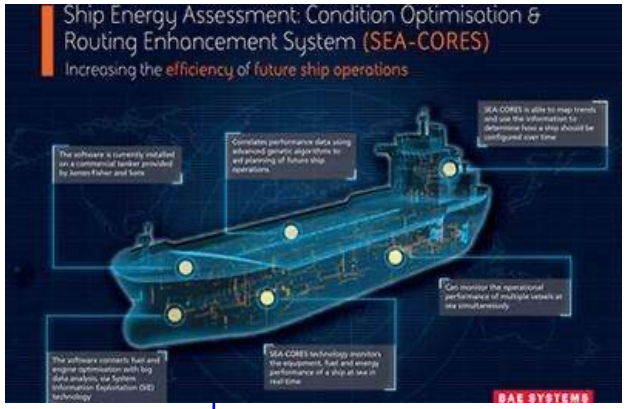
http://www.telegraph.co.uk/news/2016/09/05/navy-unveils-robot-spy-speedboat/ *2016 Royal Navy Official /Commercial*

无人自动驾驶



多源感知





应对现代军舰日益复杂系统结构、针对其系统产生的海量数据而开发，能够有机组织军舰各个子系统，最终优化全舰效能

https://phys.org/news/2016-09-software-ship-maintenance.html

2016 Royal Navy Official / Commercial

船舶能源评估-条件优化和路由增强系统

Software to transform ship maintenance

September 21, 2016

SEA-CORES. Credit: University of Southampton

Researchers from the University of Southampton are to develop software that can monitor the equipment, fuel and energy performance of a ship at sea.

The University is part of the Ship Energy Assessment – Condition Optimisation & Routing Enhancement System (SEA-CORES) consortium, which provides a live model of ship performance on global operations. The development of the software is led by BAE Systems and is sponsored by Innovate UK.

SEA-CORES is able to correlate variables that could affect a ship's performance, such as energy consumption and different weather conditions. Using genetic algorithms to track and capture the live data, SEA-CORES provides those on board with a greater understanding of the vessel's capabilities across a wide range of operations.

Researchers from Electronics and Computer Science at the University of Southampton will work on monitoring loads on the ship and applying novel machine learning techniques to a domain that has largely been data poor.

Dr Sarvapali Ramchurn, who is leading the Southampton research group, said: "Unleashing such technologies on the marine sector is likely to have a huge impact. The work we are doing at Southampton in terms of autonomous systems and machine learning will help improve the efficiency of ships and detect potential issues before they cause major damage."

BAE Systems is developing and testing SEA-CORES on a commercial tanker provided by James Fisher Marine Services. The trial will analyse the vibration and trim performance of the vessel, its hull state and monitor the integrity of the ship's superstructure.

Chris Courtaux, Head of Engineering and Energy Services at BAE Systems, said: "SEA-CORES is able to consider all of the important components which affect the performance of a vessel during deployment.

"For instance, reducing speed may save fuel but increase the wear to the engine if below its optimum performance. This could in turn increase the maintenance requirements for these vessels and reduce their availability. It is crucial that we continue to analyse what more can be done to maintain these vessels in an efficient manner and increase the number of ships available for the Royal Navy fleet."

The software connects technologies in delivering fuel and engine optimisation through the use of the BAE Systems' Ship Energy Assessment System (SEAS), together with big data analysis by using System Information Exploitation (SIE) technology.

SEA-CORES has been developed in response to the increasing complexities of modern warships and the amount of data their systems produce. The technology could transform how the Royal Navy and BAE Systems maintain and support warships in the future by using the genetic algorithms to identify the relationships between a ship's systems, calculate their different permutations and ultimately recommend a strategy to optimise the vessel's performance.

遗传算法以及其他一些机器学习方法用于获取追踪数据和确定舰船子系统关联的任务中

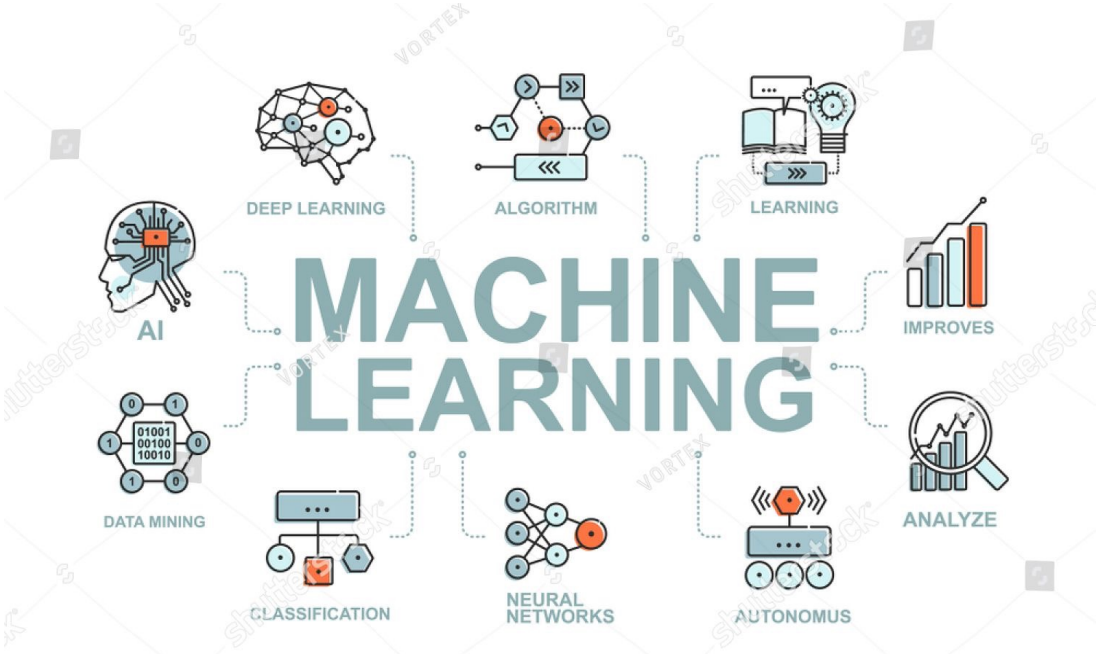
机器学习与其他领域的关系

机器学习

机器学习：以数据为经验的载体，利用经验数据不断提高性能的计算机系统/程序/算法

广袤的交叉学科

内容非常丰富



机器学习与其他领域的关系

- 机器学习与数据挖掘
  - 机器学习和数据挖掘都涉及数据分析
  - 机器学习、数据库、统计学是数据挖掘的关键支撑技术
  - 机器学习更偏技术方法，数据挖掘更偏应用些

- 机器学习与数据科学
  - 机器学习是数据科学实现智能化的关键步骤——数据智能分析
  - 数据科学还包括很多内容，例如收集、存储、传输、管理大数据
  - 大数据研究目的是为了利用大数据，没有机器学习提供数据分析技术，大数据利用无从谈起

机器学习与其他领域的关系

- 机器学习和计算机视觉
  - 机器学习是计算机视觉的核心技术
  - 计算机视觉是机器学习的重要应用
- 机器学习和自然语言处理、模式识别
  - 参考计算机视觉
- 机器学习和统计学
  - 统计学是机器学习的重要理论基础
- 机器学习和神经科学
  - 机器学习发展过程中，经常受到神经科学思想的启发
  - 神经科学发展比较缓慢，不够成熟，机器学习通常借鉴其思想，不会借鉴它的技术基础（例如，借鉴鸟儿造飞机，但是飞机的原理技术跟鸟没关系，主要靠物理、机械、材料等科学）

参考资料

- 学pyhton：

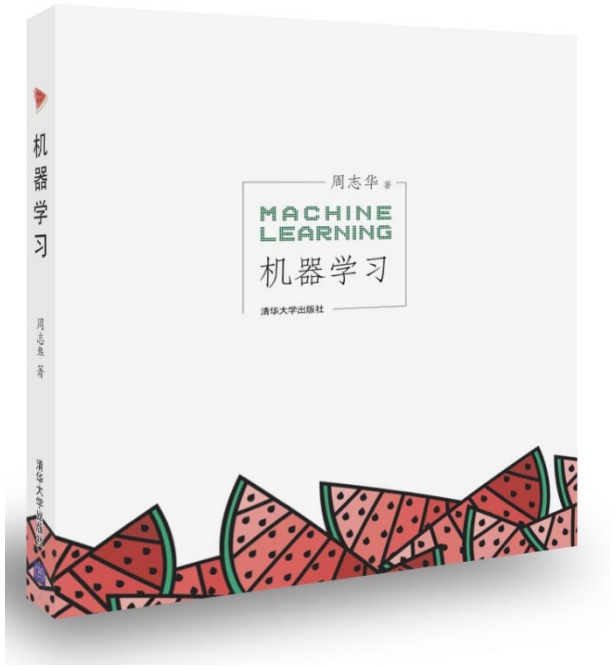
  https://www.w3schools.com/python/default.asp

- 学numpy：

https://www.numpy.org.cn/article/basics/an_introduction_to_scientific_python_numpy.html

- 学pytorch：
  - 60分钟速成：https://pytorch.org/tutorials/beginner/deep_learning_60min_blitz.html
  - 动手学深度学习（pytorch版）：https://zh.d2l.ai/

参考书籍



ISBN: 978-7-302-206853-6

2016年1月第1次印刷

2020年11月第35次印刷

周志华 著. 机器学习,

北京: 清华大学出版社,

2016年1月.

425页，62.6万字

16 章，3 附录

祝大家假期愉快！