



机器学习



讲师：曾江峰



E-mail: jfzeng@ccnu.edu.cn

A.I.



大数据，成就未来



机器学习概述

什么是机器学
习?



机器学习能做
什么?



机器学习的分
类?



机器学习如何
工作?



全民AI是必然趋势

- 第一，有智能的地方很大概率上会用到AI。
- 第二，AI带来了崭新的思维方式，如AI思维，数据思维。
- 第三，未来10-20年之内编程和AI即将会普及到每一个人，不管从事的是什么岗位。



AI的门槛变得越来越低

十几年前 ➡ 五年前 ➡ 现在 ➡ 10年后

大学教授

科学家

研究员

博士生

数据科学家

AI工程师

算法工程师

AI工程师

研发工程师

数据分析师

几乎所有的岗位



AI人才结构与趋势



人工智能背后的驱动力——机器学习

问题1



outline

1.1 机器学习的定义

1.2 机器学习的发展阶段

1.3 机器学习的种类

1.4 机器学习的三要素

1.5 机器学习的应用

1.6 常用第三方库介绍

1.7 机器学习的挑战



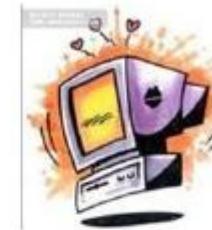
机器学习

机器学习是从人工智能中产生的一个重要学科分支，是实现智能化的关键。

经典定义：利用经验改善系统自身的性能



经验 → 数据



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

图灵奖连续授予在该方面取得突出成就的学者



Leslie Valiant
(1949 -)
(Harvard Univ.)

“计算学习理论”奠基人

2011
年度



Judea Pearl
(1936 -)
(UCLA)

“图模型学习方法”先驱

2012
年度

2018年图灵奖获得者

Yann LeCun、Geoffrey Hinton、Yoshua Bengio



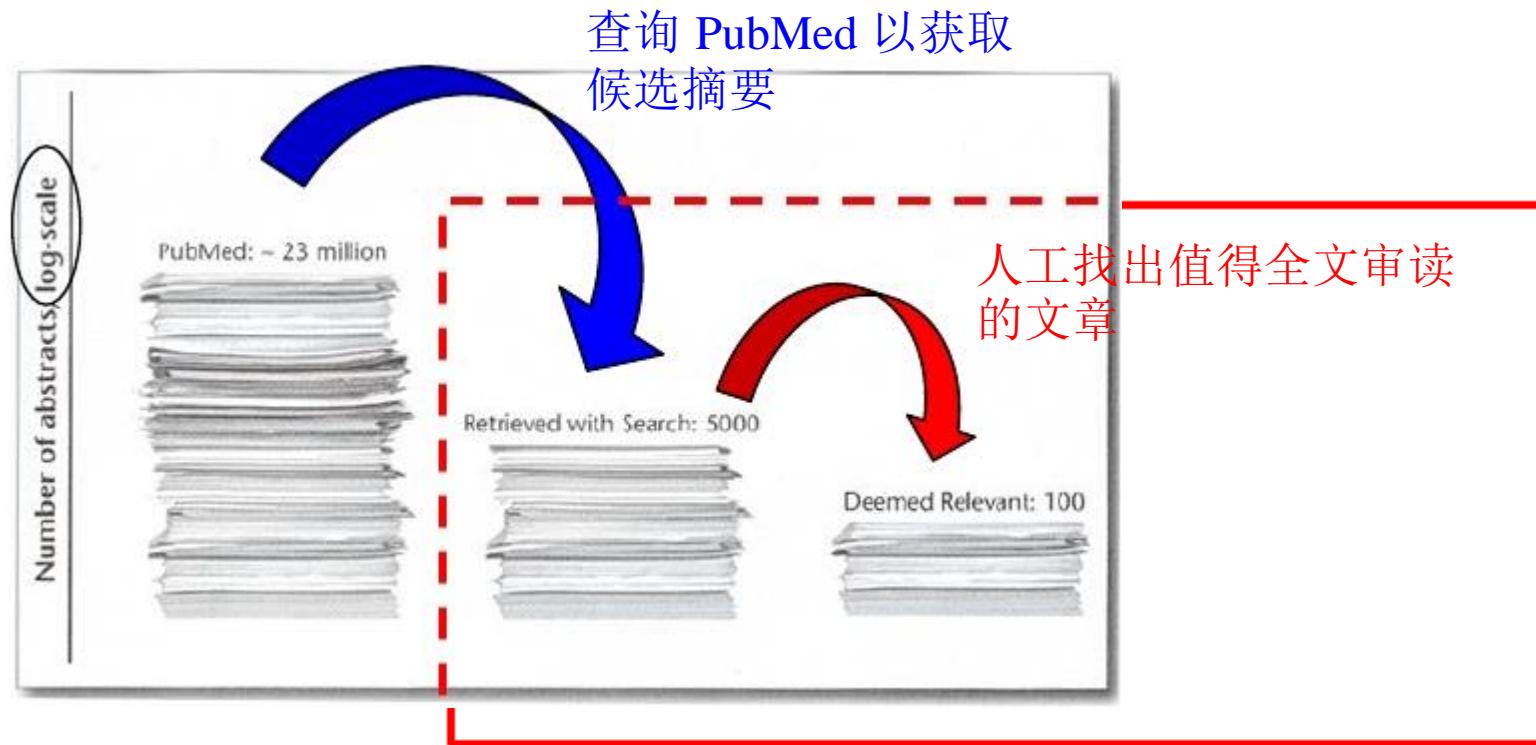
机器学习 (Machine Learning)



看个例子 ➔

“文献筛选”的故事

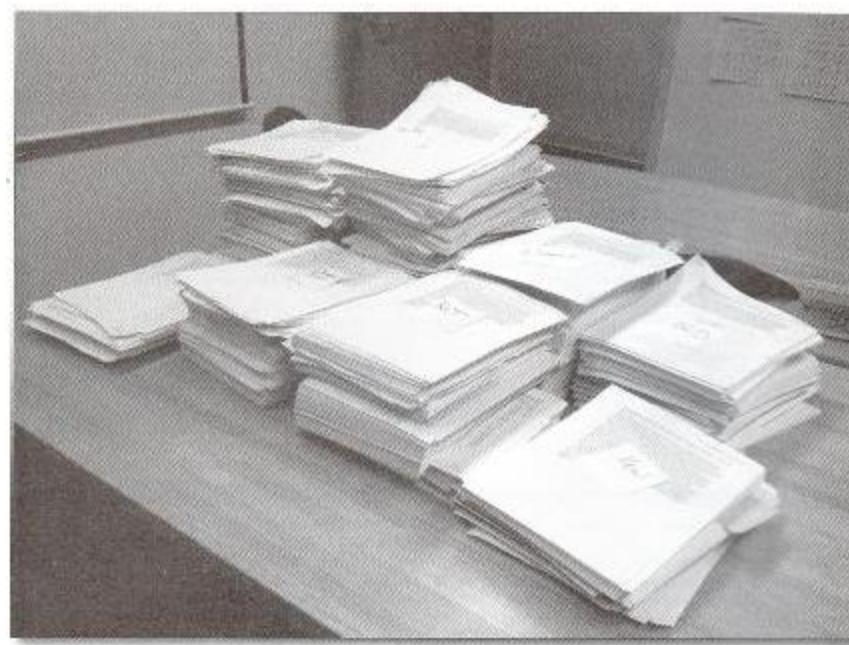
在“循证医学”(evidence-based medicine)中，针对特定的临床问题，先要对相关研究报告进行详尽评估



“文献筛选”的故事

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



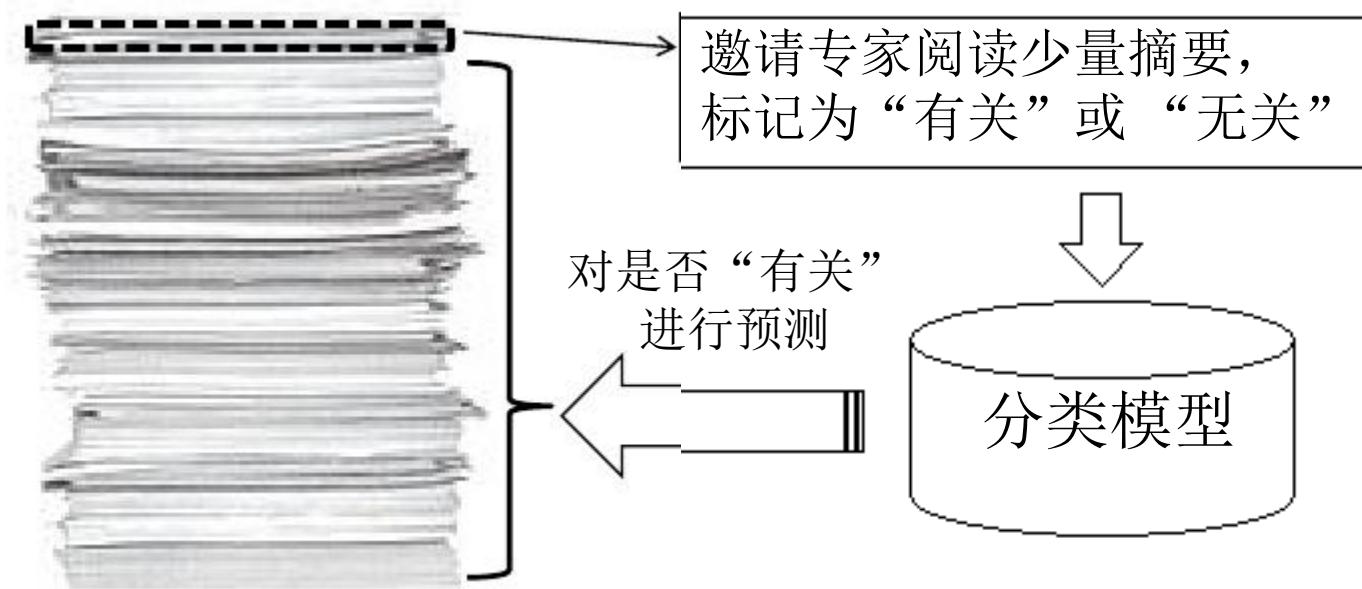
a portion of the 33,000 abstracts

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”的故事

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, 则系统的自动筛选敏感度达到 **95%**

(人类专家以前需阅读 **33,000** 篇摘要才能获得此效果)

机器学习的经典定义①

什么是机器学习

Field of study that gives computers
the ability to learn without being
explicitly programmed

机器学习的经典定义②

:: What's machine learning ?

一个计算机程序在完成任务T之后，获得经验E，其表现效果为P，如果任务T的性能表现P，随着E的增加而增加，可以称为学习。

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

- Mitchell, T. (1997). Machine Learning, McGraw Hill

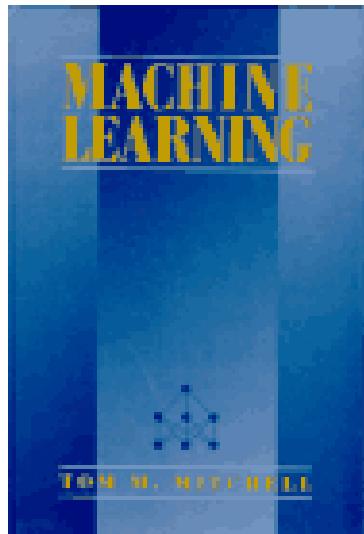
简易理解：计算机通过某项任务的经验数据提高了在该项任务上的能力。



What is machine learning

- **Machine Learning <T,P,E>:**
 - Computer automatically improves
 - at task **T**(任务)
 - according to performance metric **P**(性能)
 - through experience **E**(经验)

—— Tom Mitchell, 1997



汤姆·米切尔
(Tom M. Mitchell)

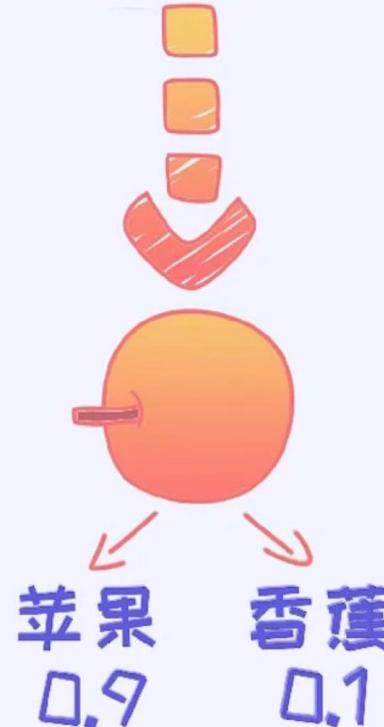
Examples of the Learning Tasks

- 下棋
 - T: 下棋
 - P: 比赛中击败对手的百分比
 - E: 与自己对弈的训练
- 手写体识别
 - T: 识别手写文字
 - P: 识别的正确率
 - E: 已经做好的具有代表性分类的手写体数据库
- 自动驾驶
 - T: 通过视觉传感器在高速路上自动驾驶
 - P: 平均无差错行驶里程
 - E: 在观察人的驾驶过程中记录的一系列图像和驾驶指令数据库

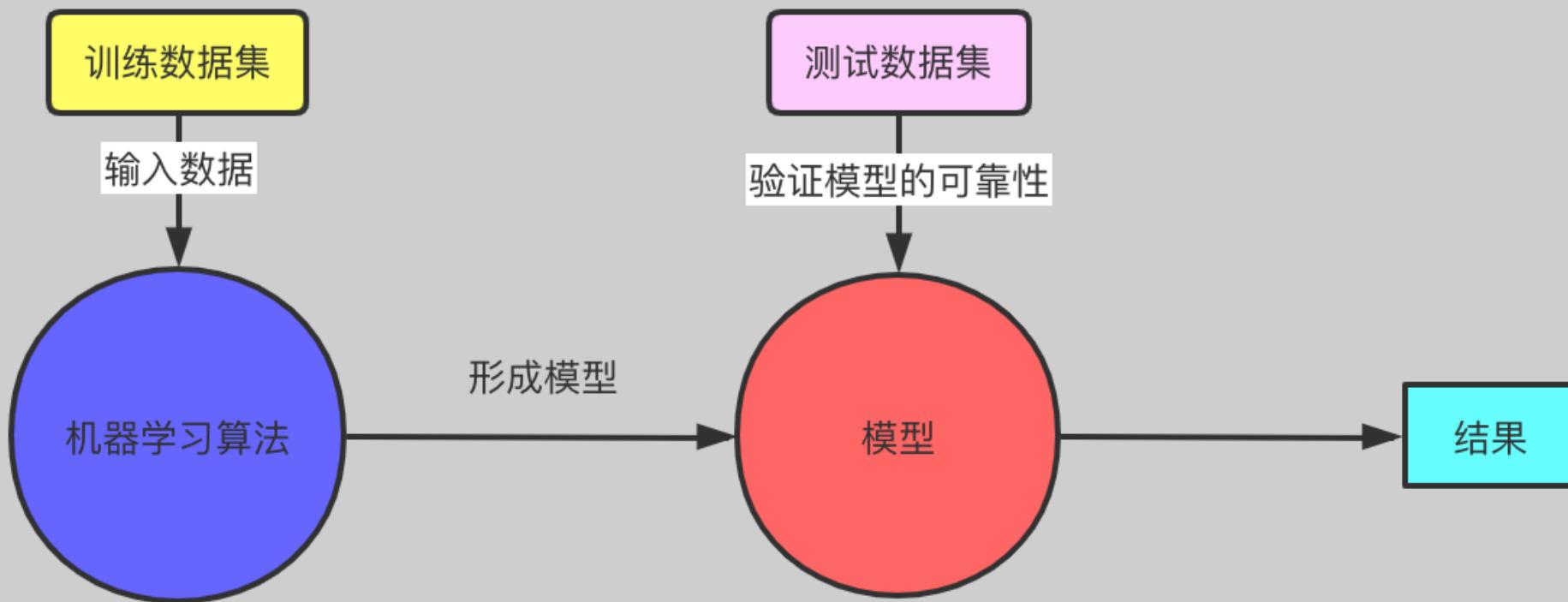
从数据中学出规律



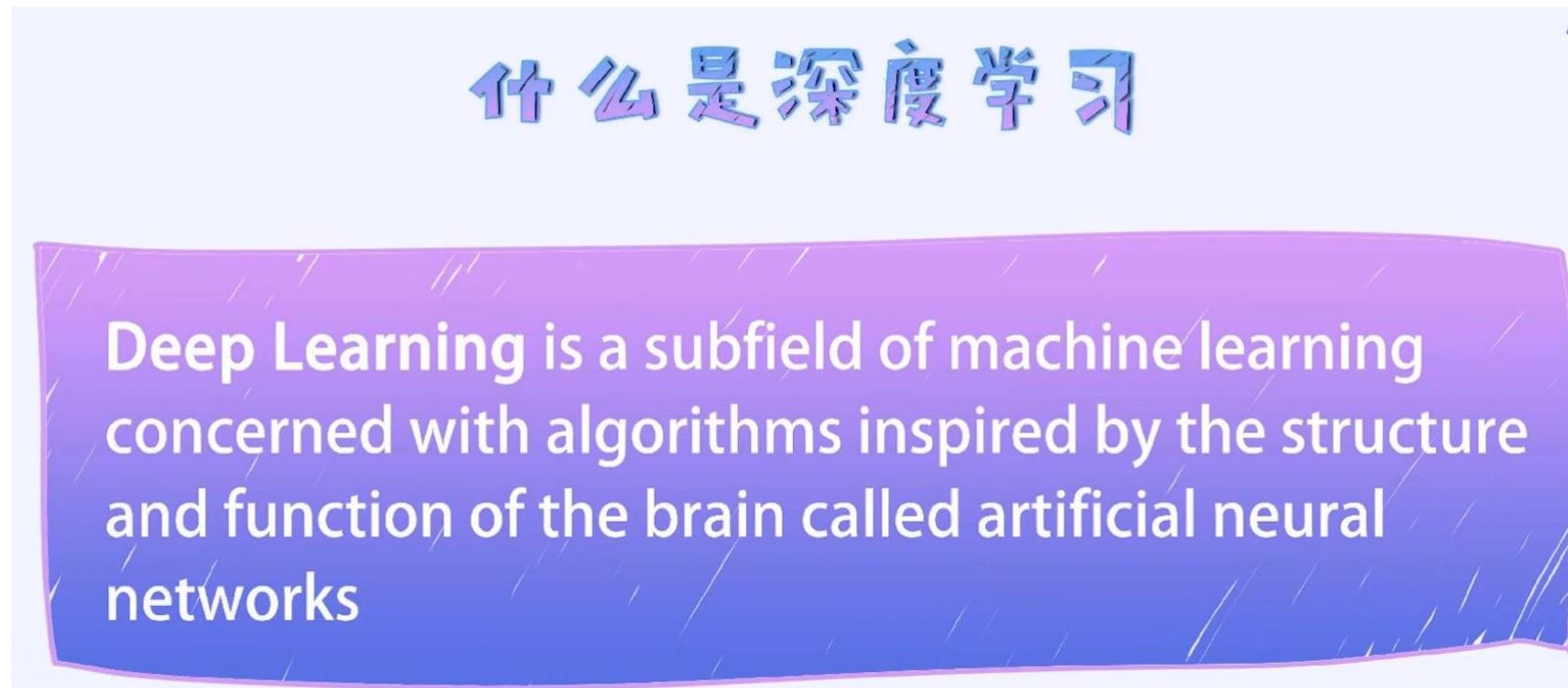
机器学习



机器学习的一般过程

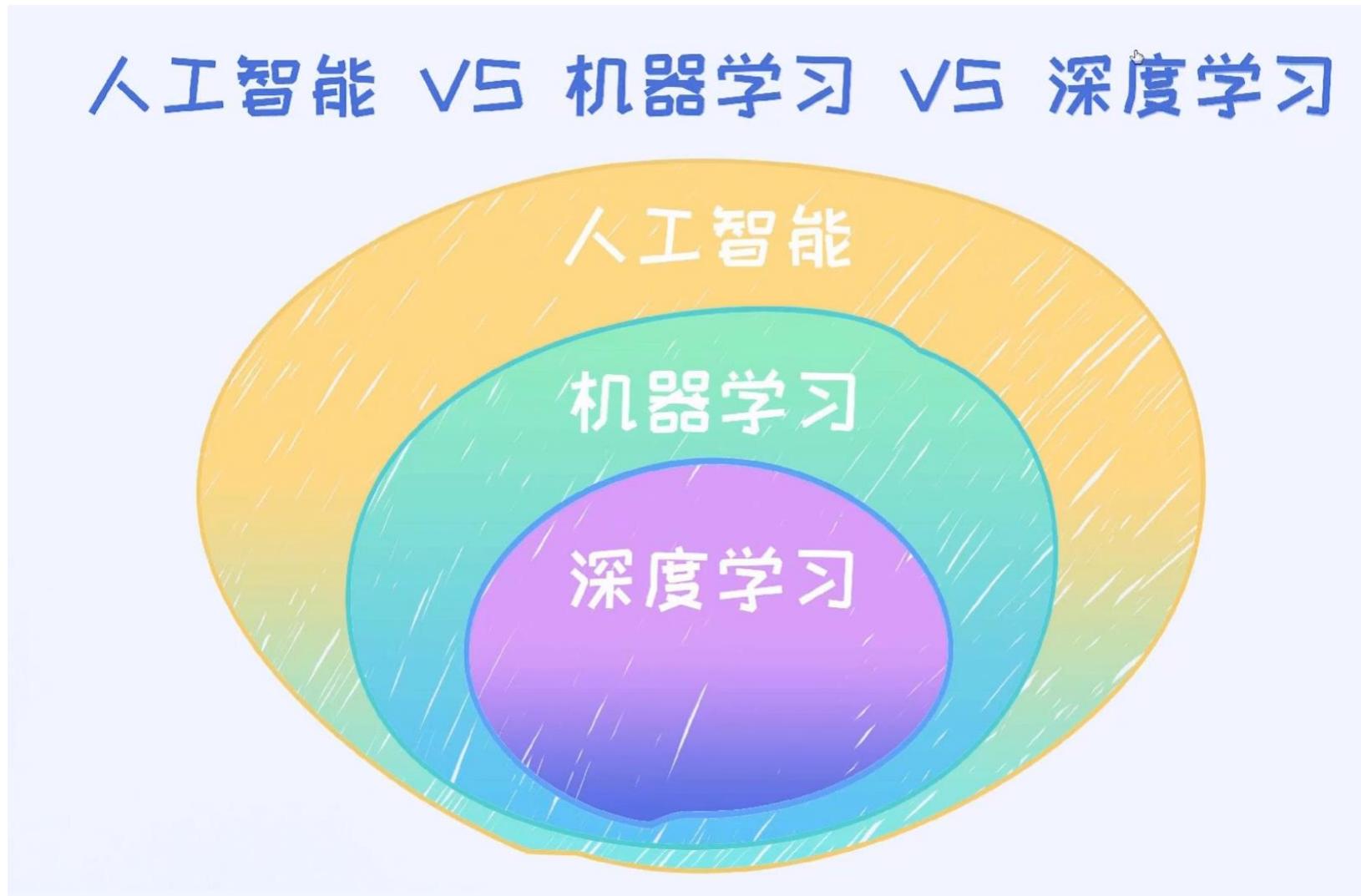


什么是深度学习

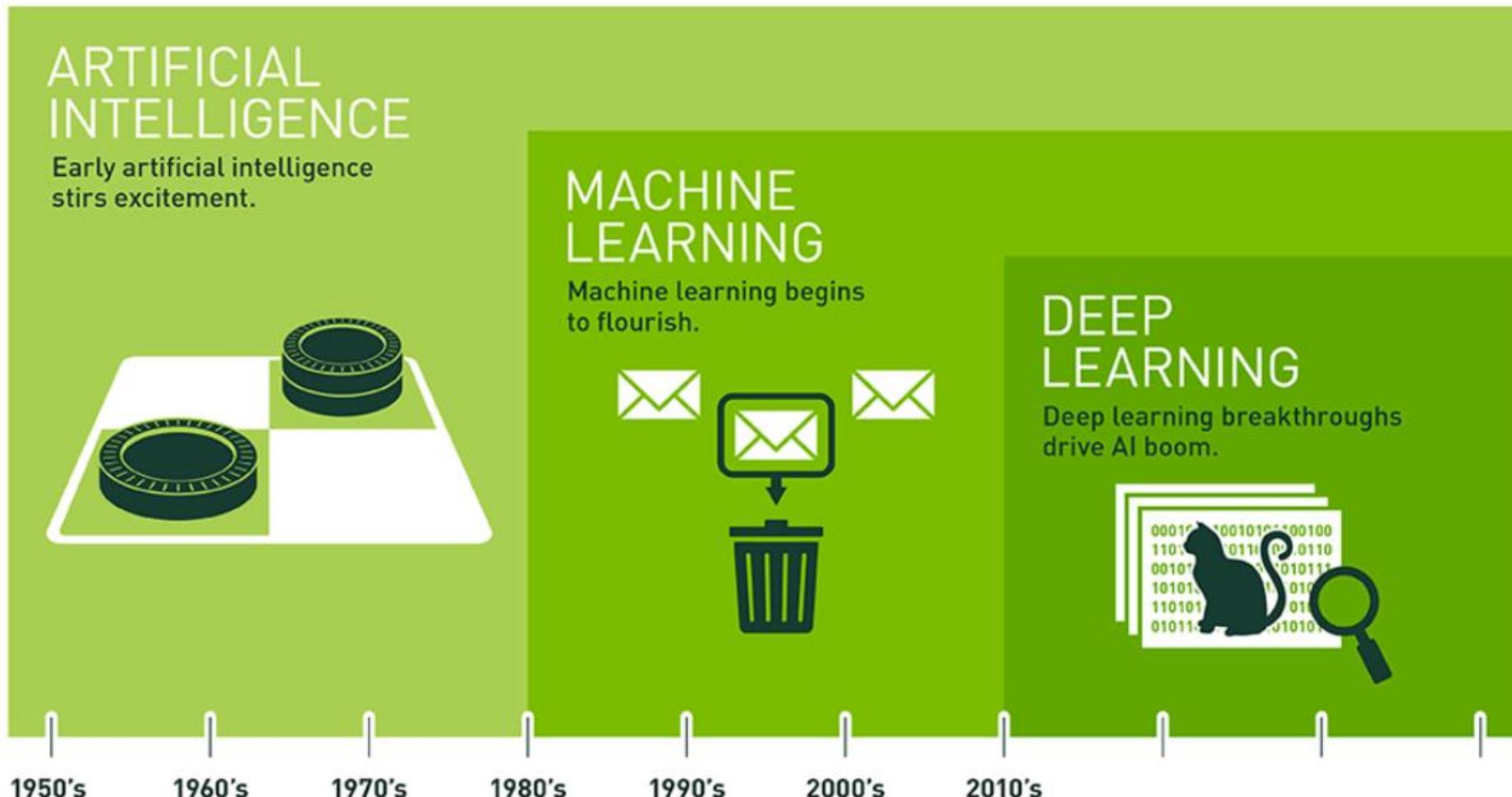


深度学习算法是一类包含多层非线性变换的神经网络，通过逐层特征变换，将样本在原始空间中的特征表示变换到一个新的特征空间，从而更准确地预测结果。与传统机器学习方法的最大不同在于，深度学习方法能够从数据中自动学习出刻画数据本质的特征表示，摒弃了复杂的人工特征提取过程。

人工智能、机器学习、深度学习之间的关系



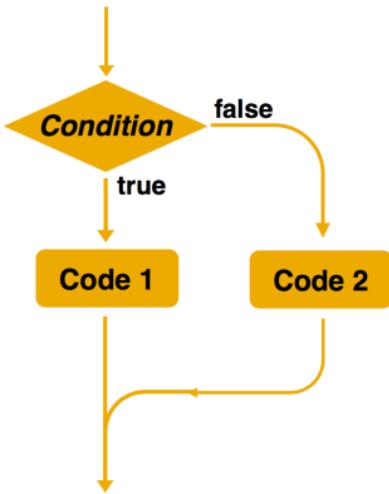
人工智能的演化进程



Since an early flush optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions

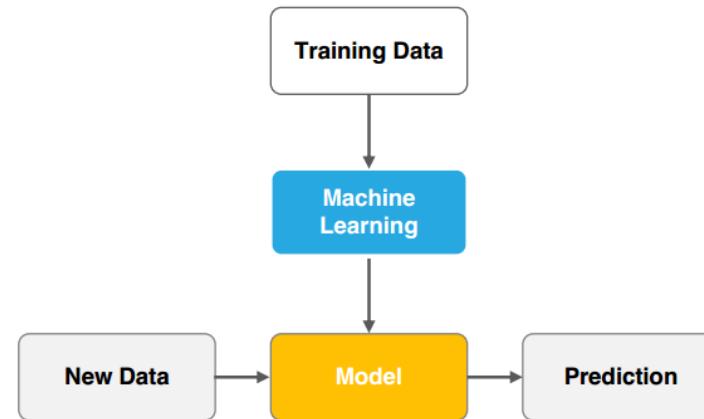
:: The difference between traditional approach and Machine Learning

Rule-based approach



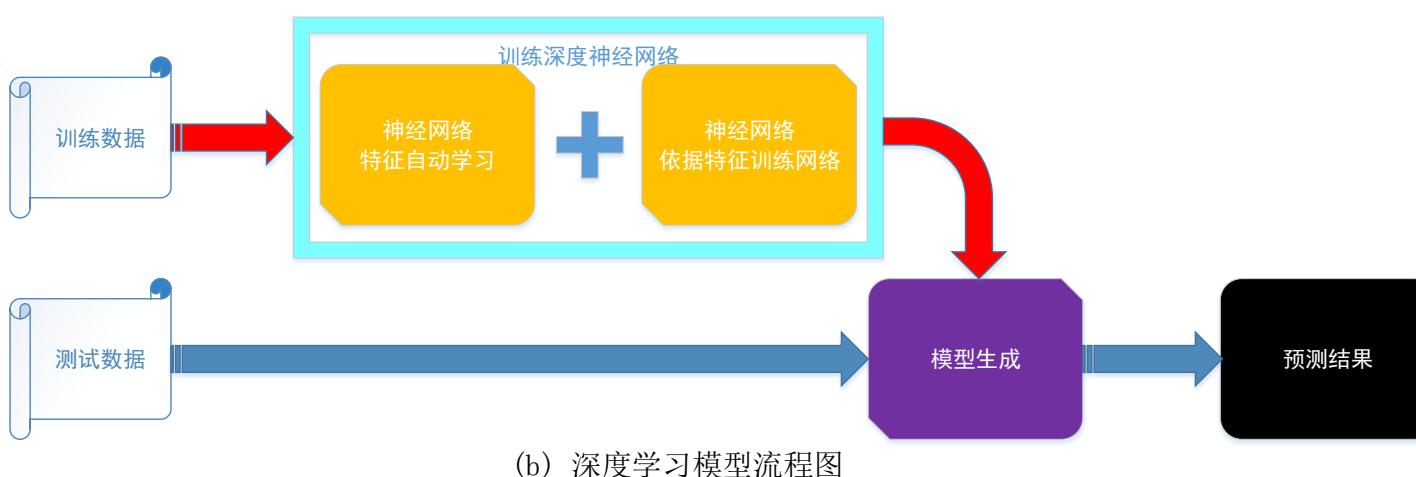
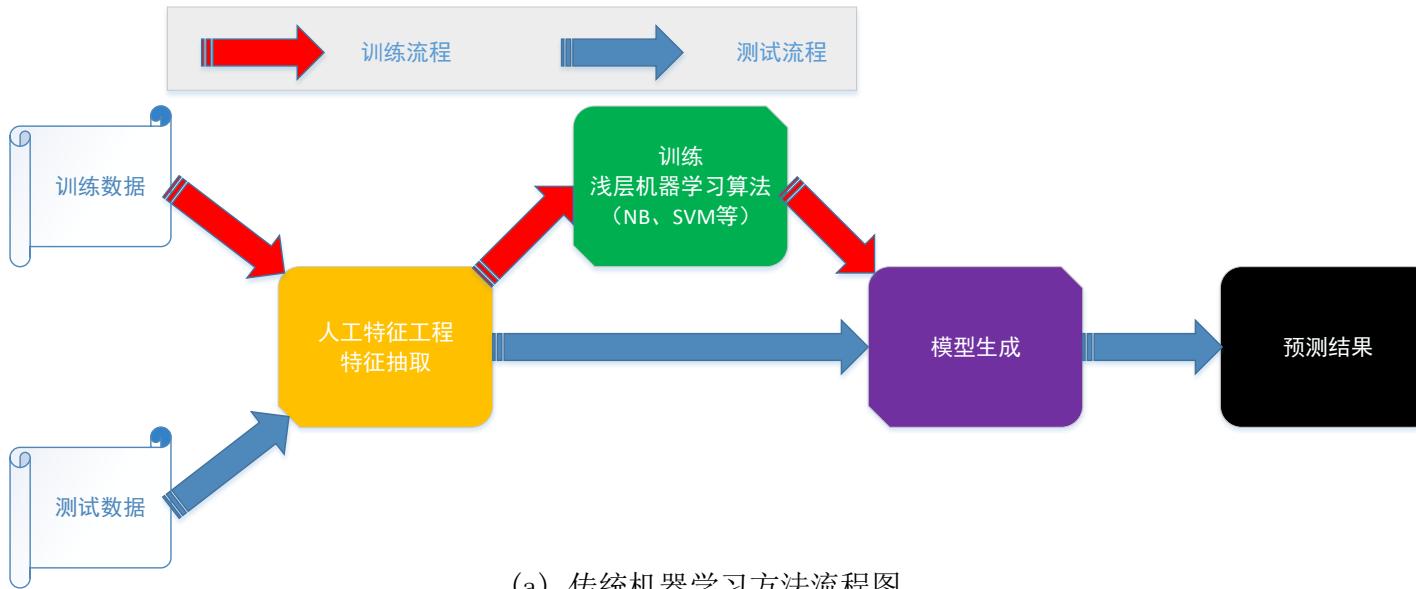
- Explicitly programmed to solve problem
- Decision rules are clearly defined by humans

Machine learning



- Trained from examples
- Decision rules complex or fuzzy
- Rules are not defined by humans but learned by the machine from data

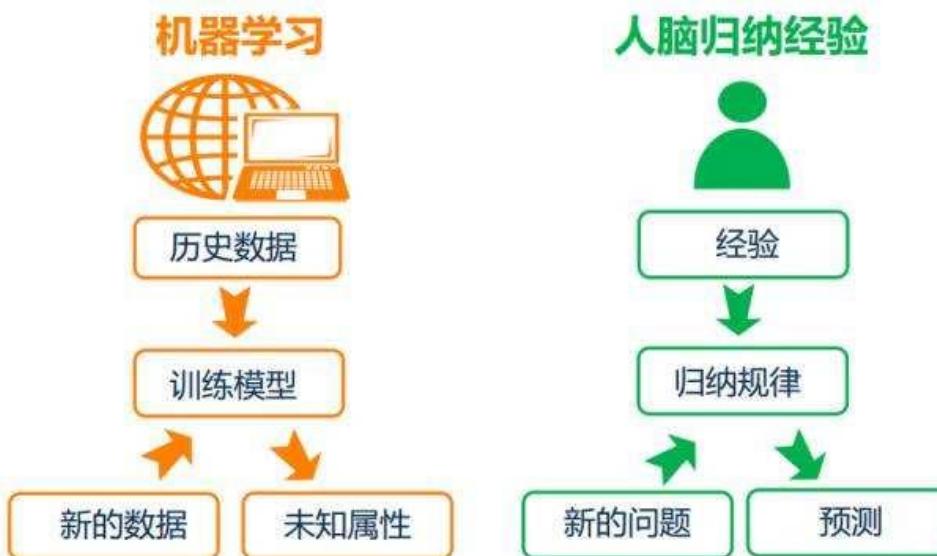
传统机器学习 VS 深度学习



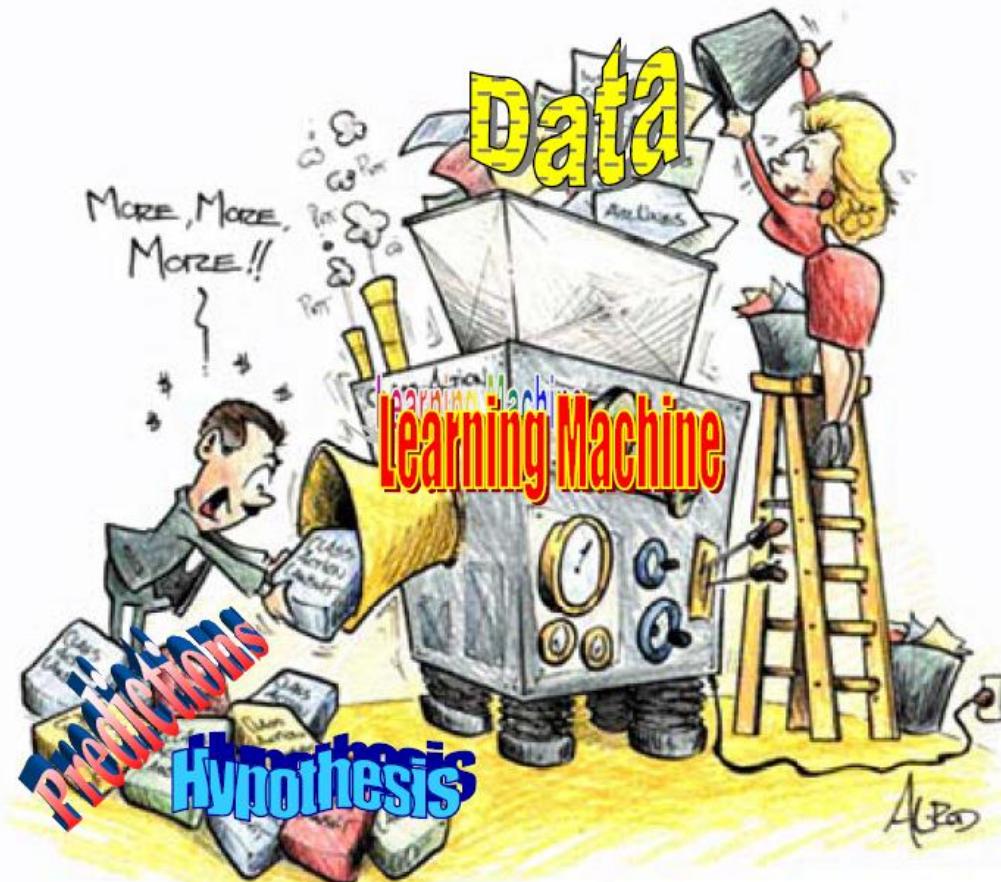
(b) 深度学习模型流程图

什么是机器学习

- 不直接编程却能赋予计算机提高能力的方法——人工智能领域的先驱 Arthur Samuel
- 系统通过计算手段利用经验来改善自身性能的效果——美国工程院院士 Tom Mitchell
- 简而言之，机器学习是让机器学会算法的算法。
- 机器学习是一门通过分析和计算数据来归纳出数据中普遍规律的学科。

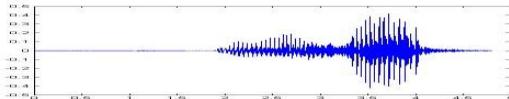


机器学习的一个形象描述



Machine Learning ≈ Looking for Function

- Speech Recognition

$$f($$

$$) = \text{“How are you”}$$

- Image Recognition

$$f($$

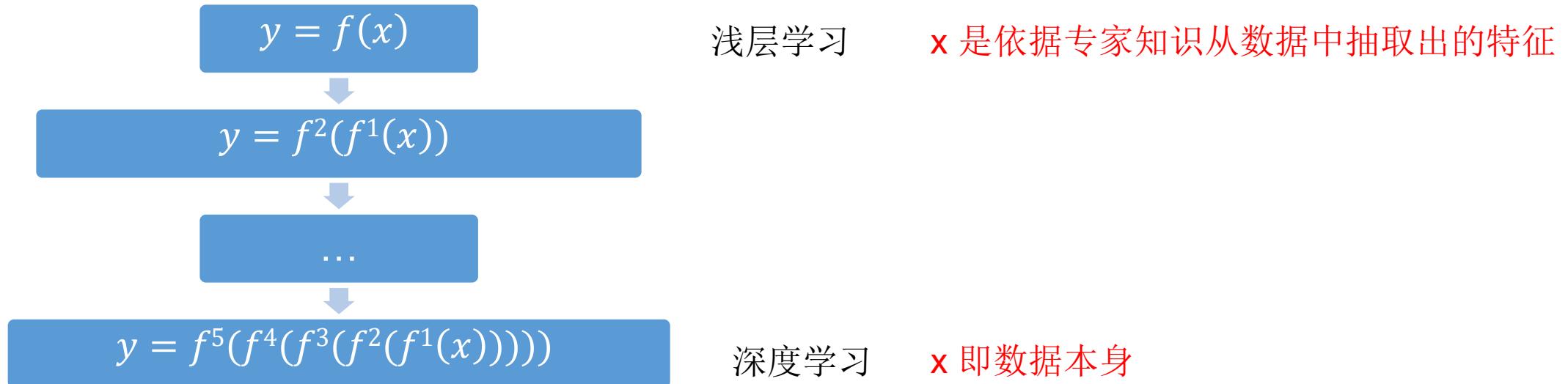
$$) = \text{“Cat”}$$

- Playing Go

$$f($$

$$) = \text{“5-5”}_{\text{(next move)}}$$

Machine Learning ≈ Looking for Function



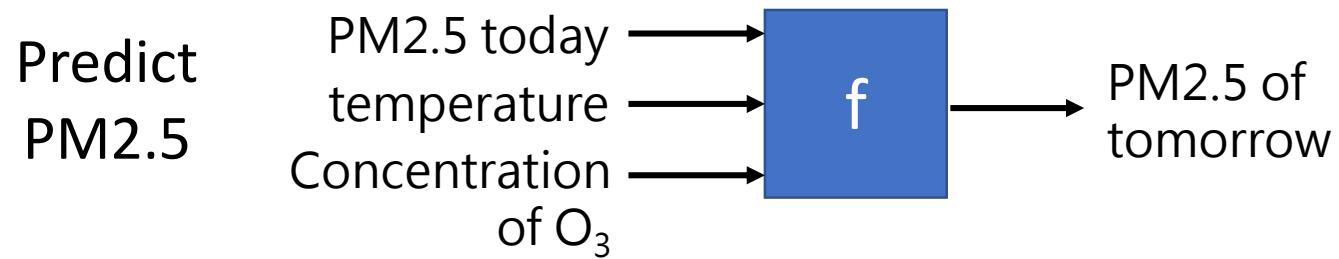
$f($  $) = \text{“你好”}$

$f($  $) = \text{“9”}$

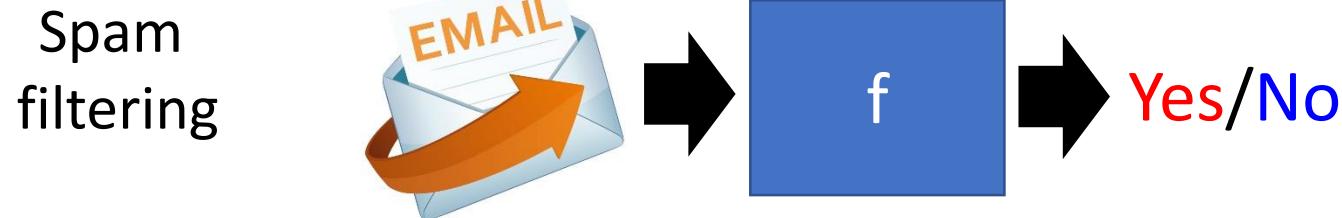
$f($  $) = \text{“Hello!”}$

Different types of Functions

Regression: The function outputs a scalar.

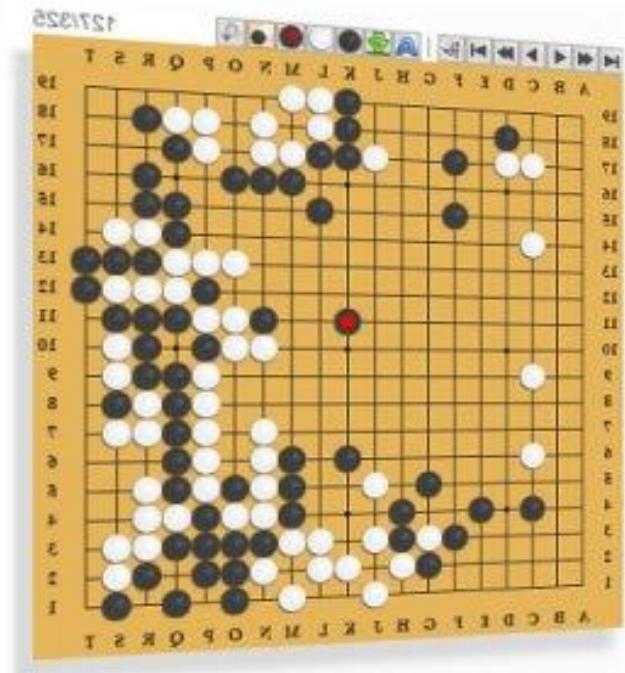


Classification: Given options (**classes**), the function outputs the correct one.



Different types of Functions

Classification: Given options (**classes**), the function outputs the correct one.



Playing GO



Each position
is a class
(19×19 classes)



a position on
the board

Next move

outline

1.1 机器学习的定义

1.2 机器学习的发展阶段

1.3 机器学习的种类

1.4 机器学习的三要素

1.5 机器学习的应用

1.6 常用第三方库介绍

1.7 机器学习的挑战



机器学习源自“人工智能”

Artificial Intelligence (AI), 1956 -

1956年夏 美国达特茅斯学院

J. McCarthy, M. Minsky, N. Rochester, C. E. Shannon,
H.A. Simon, A. Newell, A. L. Samuel 等10余人



约翰·麦卡锡
(1927-2011)
“人工智能之父”
1971年图灵奖

达特茅斯会议标志着人工智能这一学科的诞生

John McCarthy (1927 - 2011):

1971年获图灵奖，1985年获IJCAI终身成就奖。人工智能之父。他提出了“人工智能”的概念，设计出函数型程序设计语言Lisp，发展了递归的概念，提出常识推理和情境演算。出生于共产党家庭，从小阅读《10万个为什么》，中学时自修CalTech的数学课程，17岁进入CalTech时免修两年数学，22岁在Princeton获博士学位，37岁担任Stanford大学AI实验室主任。

第一阶段：推理期



1956-1960s: Logic Reasoning

- ◆ 出发点：“数学家真聪明！”
- ◆ 主要成就：自动定理证明系统（例如，西蒙与纽厄尔的“Logic Theorist”系统）

渐渐地，研究者们意识到，仅有逻辑推理能力是不够的 ...

赫伯特 西蒙
(1916-2001)
1975年图灵奖



阿伦 纽厄尔
(1927-1992)
1975年图灵奖

第二阶段：知识期



1970s -1980s: Knowledge Engineering

- ◆ 出发点：“知识就是力量！”
- ◆ 主要成就：专家系统（例如，费根鲍姆等人的“DENDRAL”系统）

爱德华 费根鲍姆
(1936-)
1994年图灵奖

渐渐地，研究者们发现，要总结出知识再“教”给系统，实在太难了 ...

第三阶段：学习期

1990s -now: Machine Learning

- ◆ 出发点：“让系统自己学！”
- ◆ 主要成就:

机器学习是作为“突破知识工程瓶颈”
之利器而出现的



恰好在20世纪90年代中后期，人类发现自己淹没在数据的汪洋中，对自动数据分析技术——机器学习的需求日益迫切

- 机器学习进入新阶段的表现
 - 机器学习已成为新的**边缘学科**并在高校形成课程。
 - 机器学习与人工智能问题的**统一性观点**正在形成。
 - 各种学习方法的**应用范围**不断扩大。
 - **数据挖掘和知识发现**的研究已形成热潮。
 - 与机器学习有关的**学术活动**空前活跃。

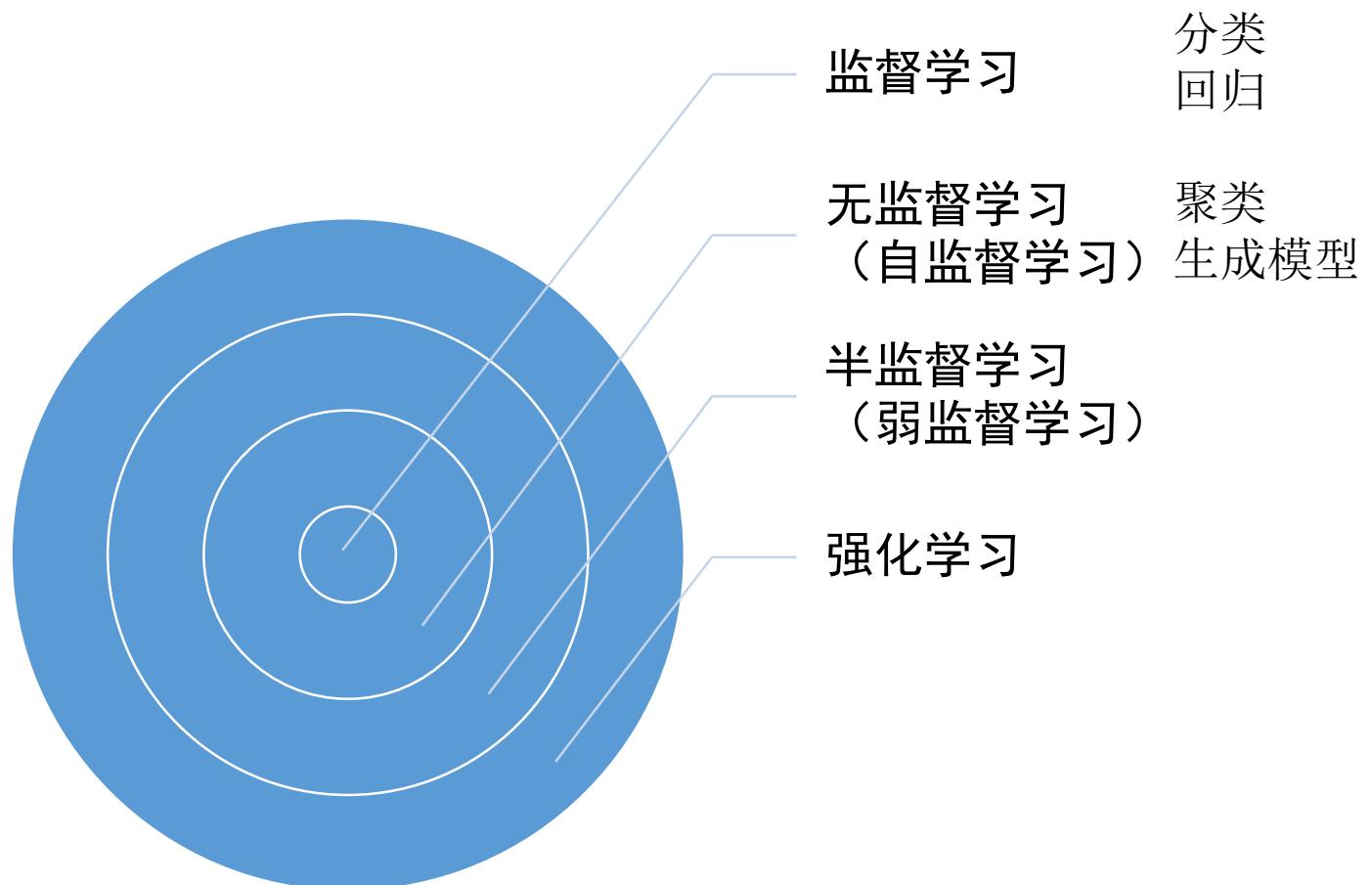
可能应该把机器学习真正当成一种支持技术（手段而非目的），考虑不同领域甚至不同学科对机器学习的需求，找出其中具有共性的、必须解决的问题，并进而着手研究。

outline

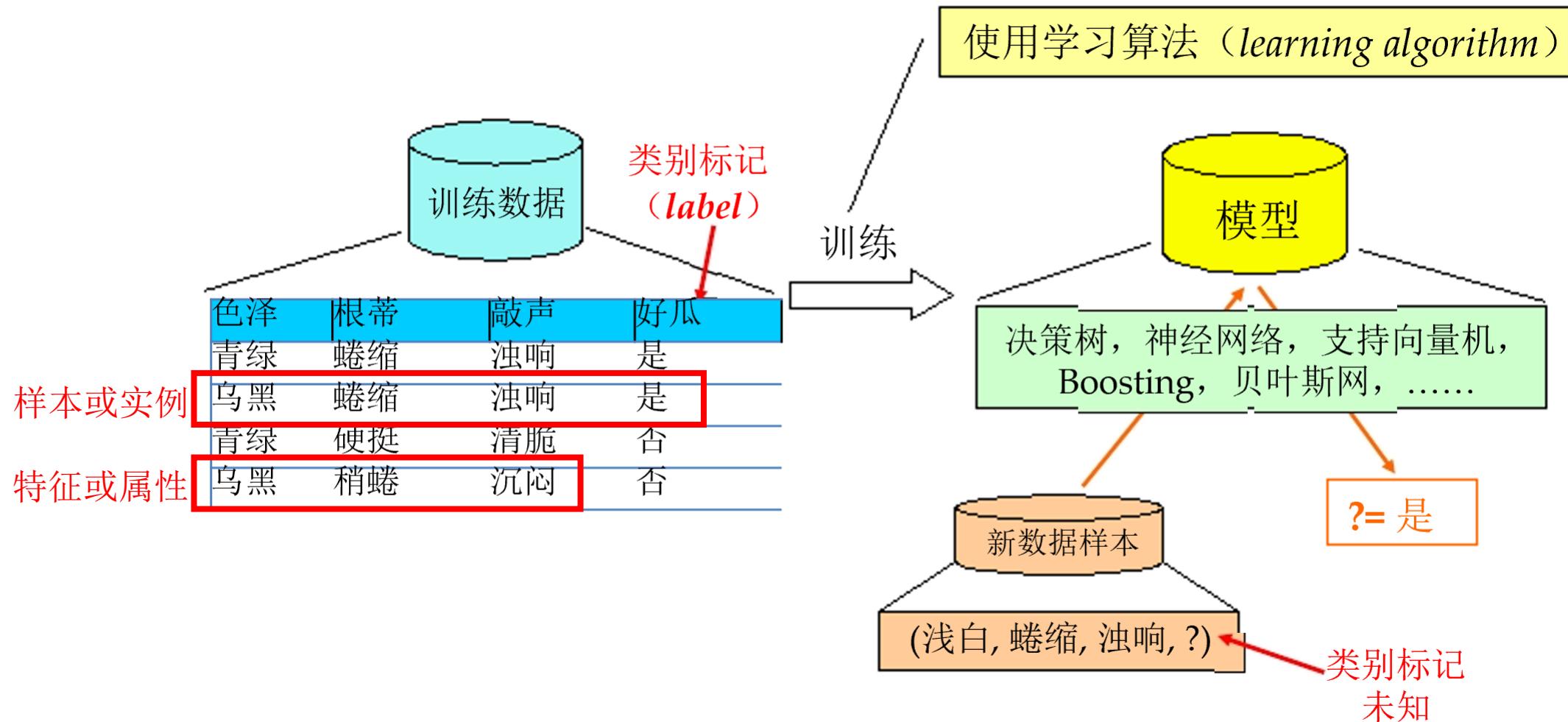
- 
- 1.1 机器学习的定义
 - 1.2 机器学习的发展阶段
 - 1.3 机器学习的种类
 - 1.4 机器学习的三要素
 - 1.5 机器学习的应用
 - 1.6 常用第三方库介绍
 - 1.7 机器学习的挑战



机器学习的分类



关键术语



监督学习与无监督学习

监督学习

无监督学习

$$D = (X, y) ^ +$$

$$D = (X)$$

学习 $X \rightarrow y$ 的关系

寻找 x 中的特征
或者规律

有老师的学习 vs 无老师的学习

监督学习的例子



无监督学习的例子



常见的机器学习算法

常见的机器学习算法

监督学习

- 线性回归
- 逻辑回归
- 朴素贝叶斯
- 决策树
- 随机森林
- SVM
- 神经网络

无监督学习

- PCA
- K-means
- GMM
- LDA
-

回归 vs 分类

回归与分类问题

回归问题

输出是连续性数值，
比如温度，身高，
气温等

分类问题

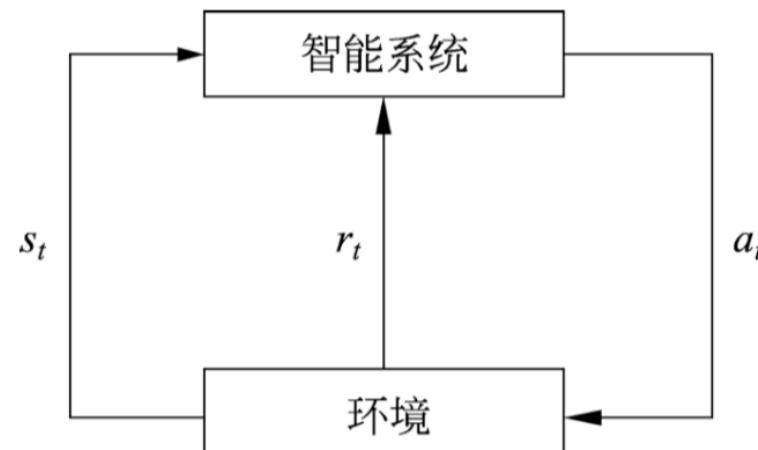
输出是定性输出
比如阴或者晴，
好或者坏



半监督学习

- 少量标注数据，大量未标注数据
- 利用未标注数据的信息，辅助标注数据，进行监督学习
- 较低成本

强化学习



outline

1.1 机器学习的定义

1.2 机器学习的发展阶段

1.3 机器学习的种类

1.4 机器学习的三要素

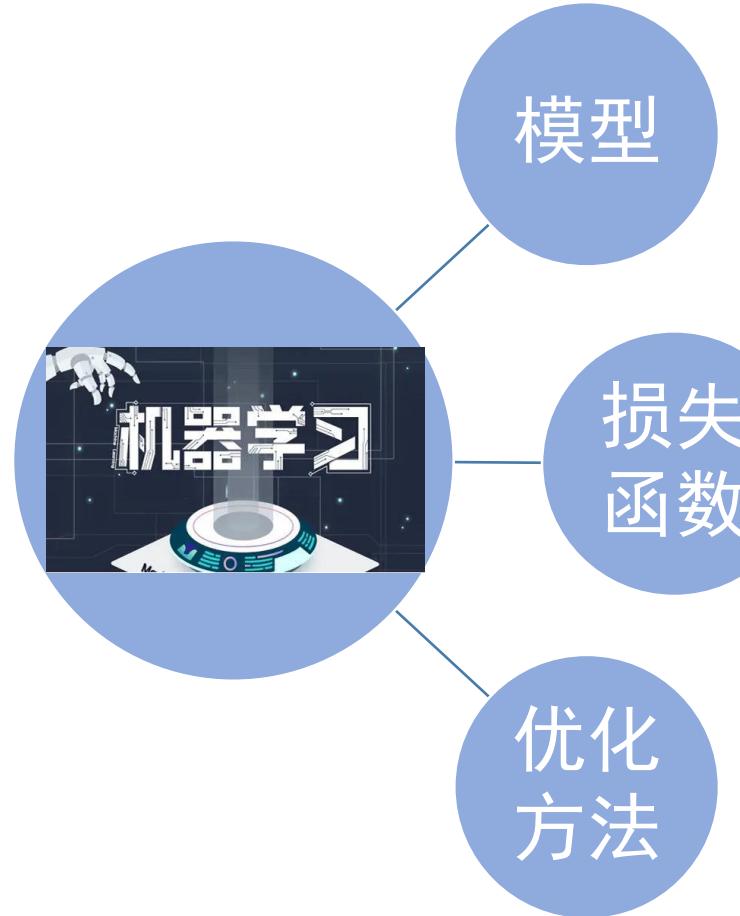
1.5 机器学习的应用

1.6 常用第三方库介绍

1.7 机器学习的挑战



机器学习三要素



要学习的决策函数或者条件概率分布，一般用假设空间 \mathcal{F} 来描述所有可能的决策函数或条件概率分布。

$$\mathcal{F} = \{f \mid Y = f(X)\} \text{ 或 } \mathcal{F} = \{P \mid P(Y|X)\}$$

损失函数，评估训练过程中模型的最优参数，即按照什么标准来选择最有模型。

给定模型，预测值 $f(X)$ 与真实标签 Y 之间的误差可以用一个损失函数 $L(Y, F(x))$ 来度量。

当机器学习的模型和损失函数确定后，机器学习就可以具体地形式化为一个最优化问题，可以通过常用的优化算法，如梯度下降、牛顿法等进行模型参数的优化求解。



监督机器学习的核心

- 该公式可谓是机器学习中最核心最关键最能概述监督学习的核心思想的公式：

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

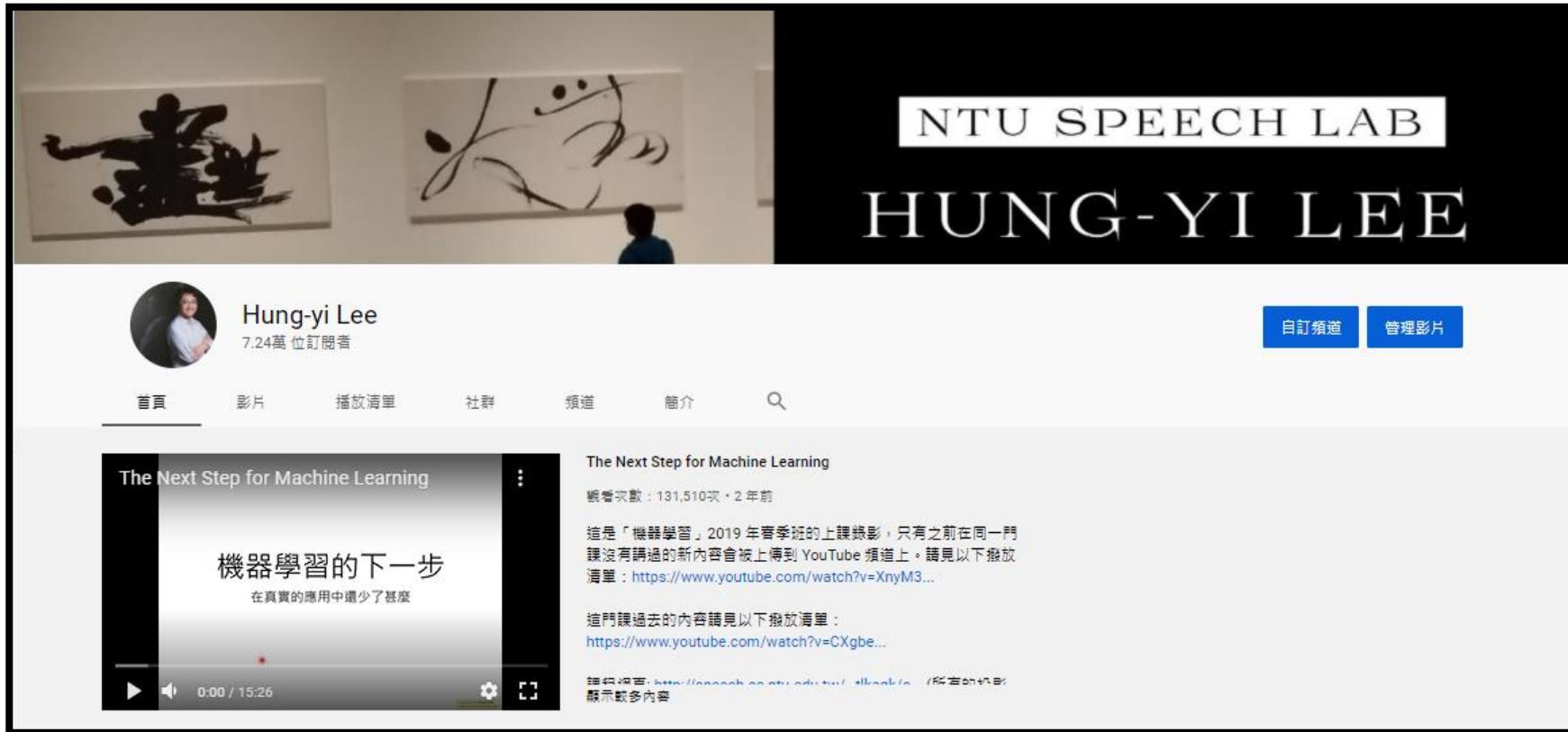
- 所有的有监督机器学习，无非就是正则化参数的同时最小化经验误差函数。最小化经验误差是为了极大程度的拟合训练数据，正则化参数是为了防止过分的拟合训练数据。



How to find a function?

A Case Study

YouTube Channel



<https://www.youtube.com/c/HungyiLeeNTU>

The function we want to find ...

$y = f($
no. of
views on
2/26

| 日期 | 喜歡的人數 | 訂閱人數 | 觀看次數 |
|------------|---------|---------|------------|
| 2021年1月26日 | 54 4.9% | 69 5.5% | 6,788 5.2% |
| 2021年1月27日 | 60 5.4% | 71 5.6% | 6,242 4.7% |
| 2021年1月28日 | 36 3.2% | 63 5.0% | 5,868 4.5% |
| 2021年1月29日 | 27 2.4% | 40 3.2% | 4,413 3.4% |
| 2021年1月30日 | 40 3.6% | 40 3.2% | 4,372 3.3% |
| 2021年1月31日 | 47 4.2% | 51 4.0% | 5,135 3.9% |
| 2021年2月1日 | 61 5.5% | 29 2.3% | 5,527 4.2% |
| 2021年2月2日 | 49 4.4% | 43 3.4% | 5,911 4.5% |
| 2021年2月3日 | 26 2.3% | 44 3.5% | 5,248 4.0% |
| 2021年2月4日 | 43 3.9% | 33 2.6% | 4,771 3.6% |
| 2021年2月5日 | 45 4.0% | 49 3.9% | 3,850 2.9% |
| 2021年2月6日 | 29 2.6% | 42 3.3% | 3,828 2.9% |
| 2021年2月7日 | 26 2.3% | 46 3.6% | 4,559 3.5% |
| 2021年2月8日 | 38 3.4% | 26 2.1% | 4,772 3.6% |
| 2021年2月9日 | 29 2.6% | 25 2.0% | 3,847 2.9% |
| 2021年2月10日 | 31 2.8% | 35 2.8% | 3,382 2.6% |

)

1. Function with Unknown Parameters

$$y = f($$





| 日期 | 新增的影片數 | 獨家的人數 | 獨家的訂閱人數 | 總次數 | 總觀次數 | 總觀時間 (小時) | 平均觀看時間長度 |
|------------|--------|---------|----------|------------|------------|------------|----------|
| 總計 | 199 | 17,022 | 26,011 | 27,602,732 | 2,066,634 | 266,778.0 | 7:48 |
| 2020年1月1日 | — | 16 0.1% | 52 0.5% | 57,093 | 3,977 0.2% | 565.6 0.2% | 8:32 |
| 2020年1月2日 | — | 38 0.2% | 58 0.2% | 56,204 | 4,214 0.2% | 589.8 0.2% | 8:23 |
| 2020年1月3日 | — | 24 0.1% | 89 0.3% | 53,321 | 3,288 0.2% | 457.4 0.2% | 8:20 |
| 2020年1月4日 | 1 0.5% | 27 0.2% | 66 0.3% | 53,599 | 3,559 0.2% | 483.5 0.2% | 8:09 |
| 2020年1月5日 | — | 35 0.2% | 85 0.3% | 63,001 | 4,677 0.2% | 596.4 0.2% | 7:39 |
| 2020年1月6日 | — | 31 0.2% | 69 0.3% | 60,175 | 4,682 0.2% | 642.0 0.2% | 8:13 |
| 2020年1月7日 | — | 40 0.2% | 70 0.3% | 63,638 | 4,695 0.2% | 618.4 0.2% | 7:54 |
| 2020年1月8日 | — | 39 0.2% | 59 0.2% | 59,900 | 4,785 0.2% | 646.7 0.2% | 8:06 |
| 2020年1月9日 | — | 28 0.2% | 64 0.3% | 54,988 | 4,911 0.2% | 670.9 0.3% | 8:11 |
| 2020年1月10日 | — | 17 0.1% | 51 0.2% | 40,631 | 3,069 0.2% | 372.0 0.1% | 7:16 |
| 2020年1月11日 | — | 12 0.1% | 54 0.2% | 36,168 | 2,898 0.1% | 369.3 0.1% | 7:38 |
| 2020年1月12日 | — | 40 0.2% | 169 0.7% | 53,964 | 4,477 0.2% | 572.9 0.2% | 7:40 |
| 2020年1月13日 | — | 29 0.2% | 75 0.3% | 61,043 | 5,017 0.2% | 661.4 0.3% | 7:54 |
| 2020年1月14日 | — | 32 0.2% | 83 0.3% | 64,968 | 5,186 0.3% | 618.3 0.2% | 7:09 |

Model $y = b + wx_1$ based on domain knowledge **feature**

y : no. of views on 2/26, x_1 : no. of views on 2/25

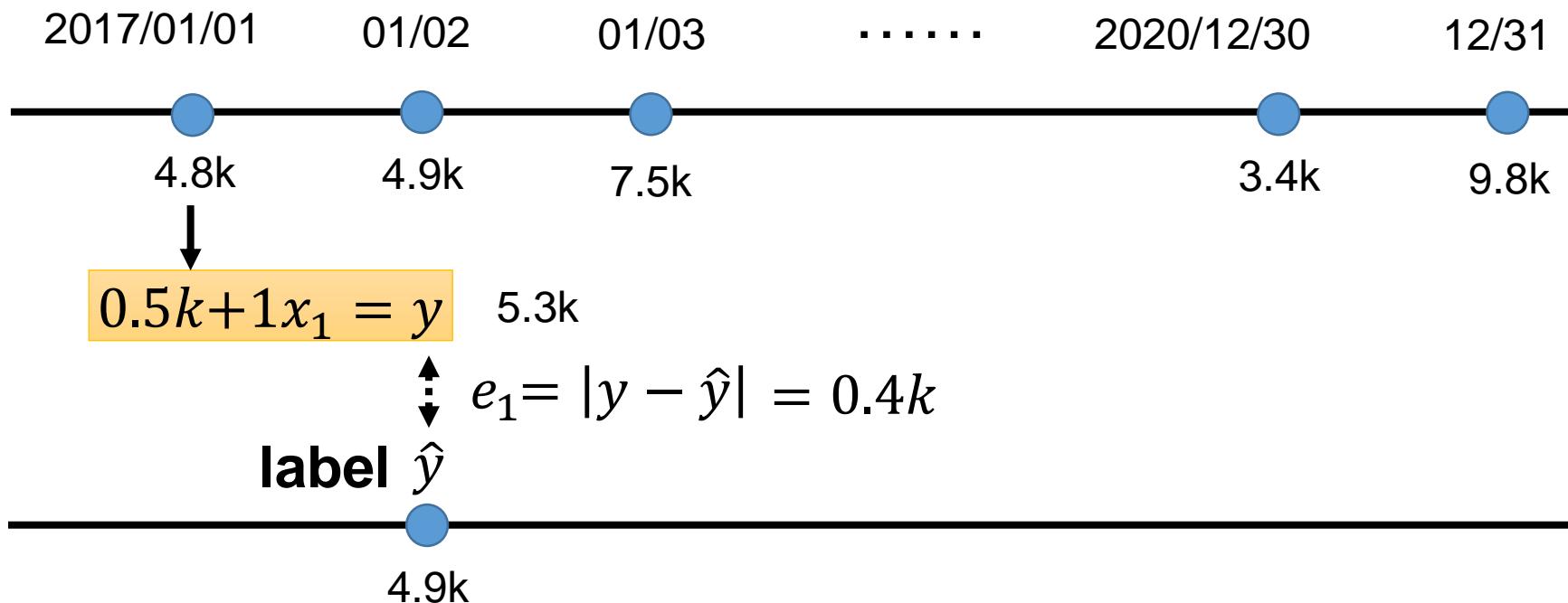
w and b are unknown parameters (learned from data) **weight** **bias**

2. Define **Loss** from Training Data

- Loss is a function of parameters $L(b, w)$
- Loss: how good a set of values is.

$$L(0.5k, 1) \quad y = b + wx_1 \rightarrow y = 0.5k + 1x_1 \quad \text{How good it is?}$$

Data from 2017/01/01 – 2020/12/31

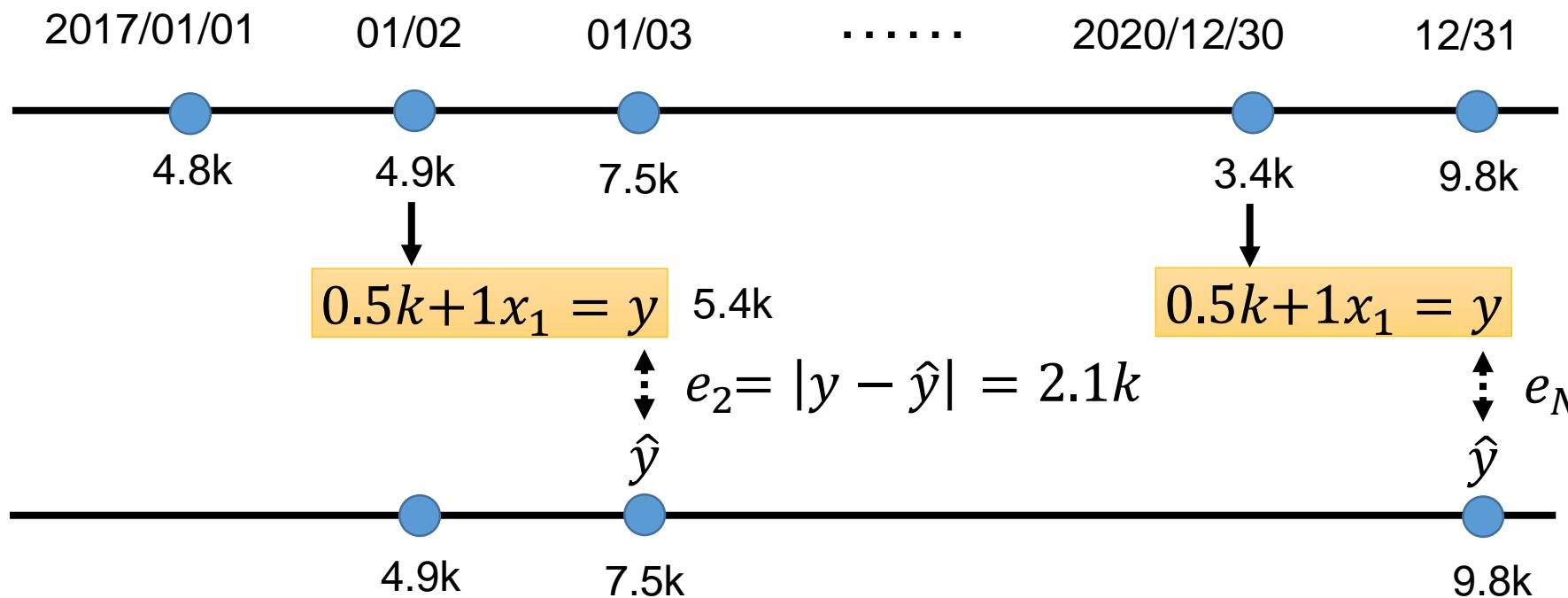


2. Define **Loss** from Training Data

- Loss is a function of parameters $L(b, w)$
- Loss: how good a set of values is.

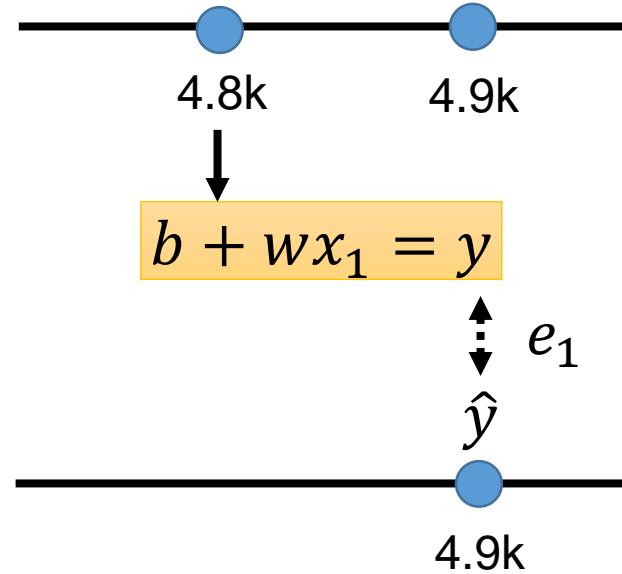
$L(0.5k, 1) \quad y = b + wx_1 \rightarrow y = 0.5k + 1x_1$ How good it is?

Data from 2017/01/01 – 2020/12/31



2. Define Loss from Training Data

- Loss is a function of parameters $L(b, w)$
- Loss: how good a set of values is.



$$\text{Loss: } L = \frac{1}{N} \sum_n e_n$$

$e = |y - \hat{y}|$ L is mean absolute error (MAE)

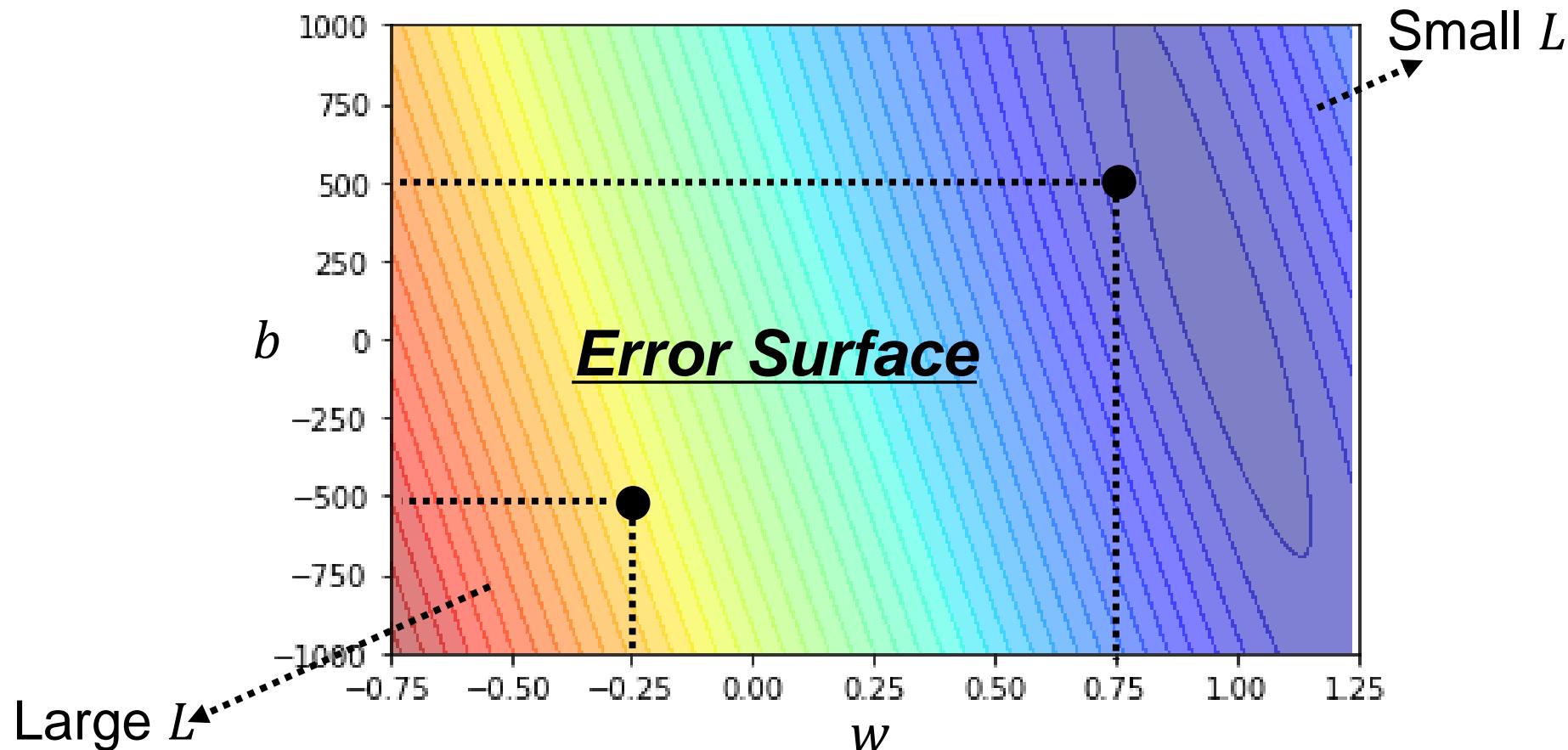
$e = (y - \hat{y})^2$ L is mean square error (MSE)

If y and \hat{y} are both probability distributions → Cross-entropy

2. Define Loss from Training Data

Model $y = b + wx_1$

- Loss is a function of parameters $L(b, w)$
- Loss: how good a set of values is.

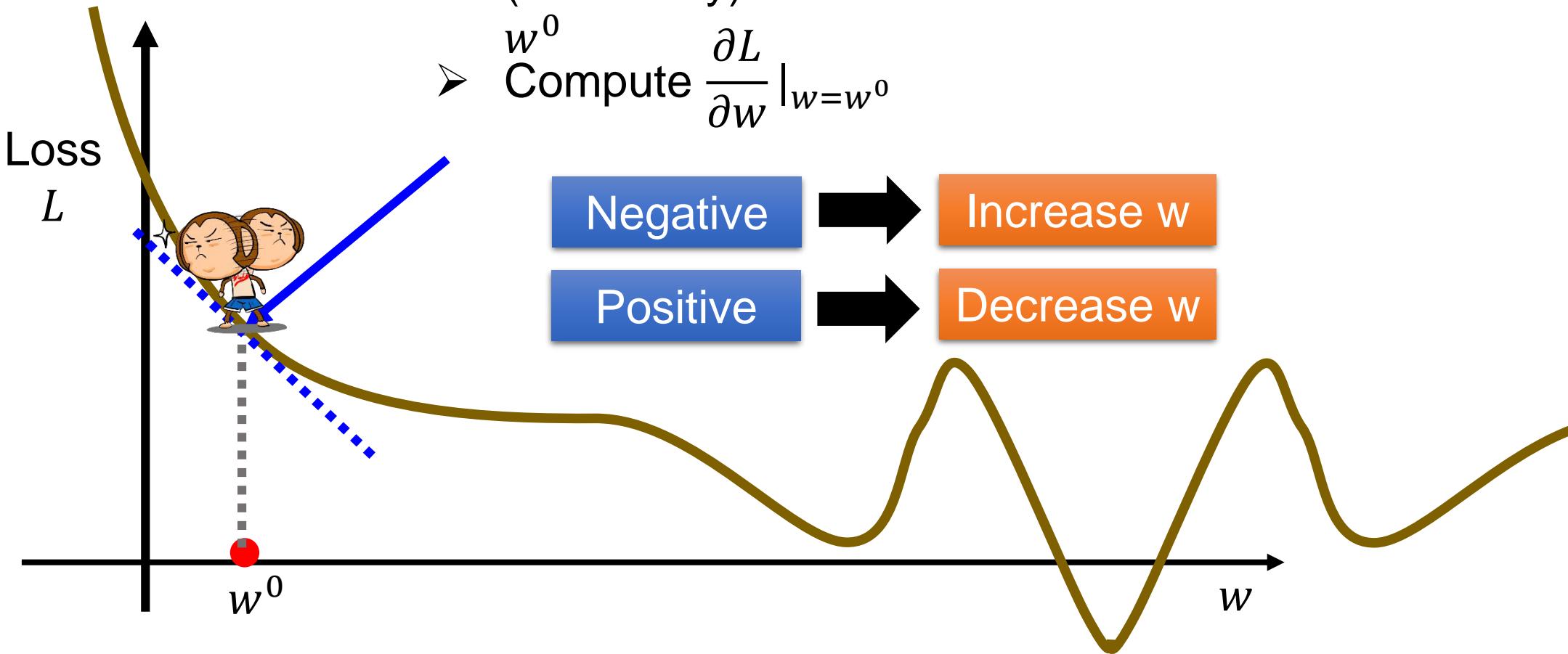


3. Optimization

$$w^* = \arg \min_w L$$

Gradient Descent

- (Randomly) Pick an initial value w^0
- Compute $\frac{\partial L}{\partial w} |_{w=w^0}$



3. Optimization

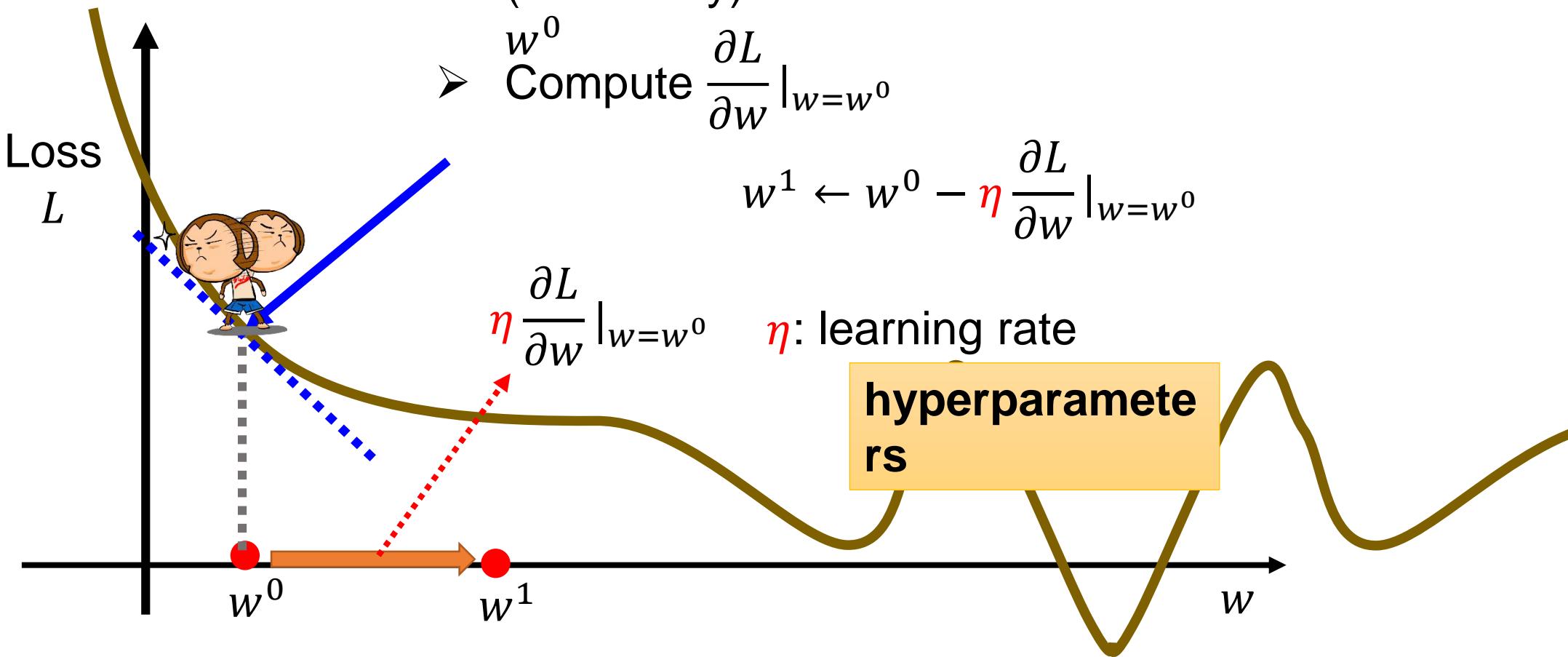
$$w^* = \arg \min_w L$$

Gradient Descent

➤ (Randomly) Pick an initial value

➤ Compute $\frac{\partial L}{\partial w} \Big|_{w=w^0}$

$$w^1 \leftarrow w^0 - \eta \frac{\partial L}{\partial w} \Big|_{w=w^0}$$



3. Optimization

$$w^* = \arg \min_w L$$

Gradient Descent

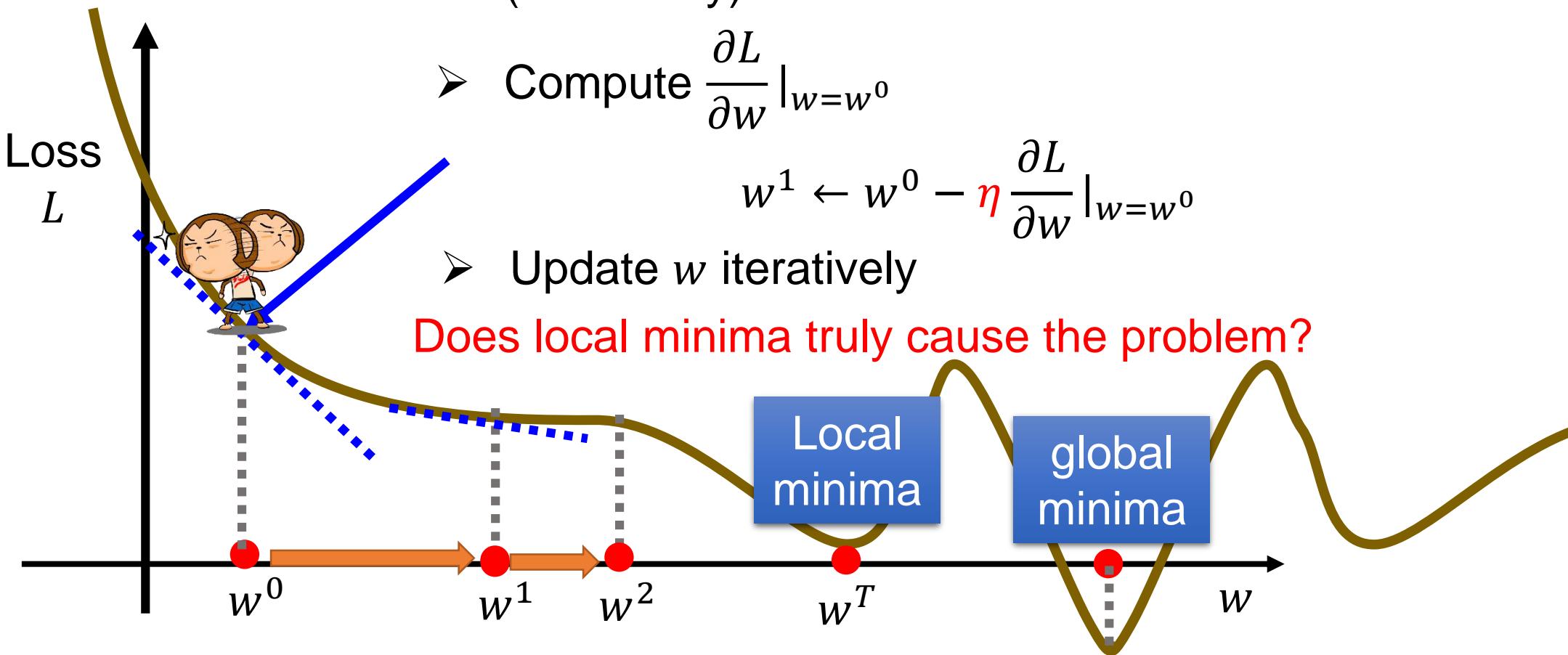
➤ (Randomly) Pick an initial value w^0

➤ Compute $\frac{\partial L}{\partial w} |_{w=w^0}$

$$w^1 \leftarrow w^0 - \eta \frac{\partial L}{\partial w} |_{w=w^0}$$

➤ Update w iteratively

Does local minima truly cause the problem?



3. Optimization

$$w^*, b^* = \arg \min_{w,b} L$$

- (Randomly) Pick initial values w^0, b^0
- Compute

$$\begin{array}{l} \frac{\partial L}{\partial w} \Big|_{w=w^0, b=b^0} \\ \frac{\partial L}{\partial b} \Big|_{w=w^0, b=b^0} \end{array}$$

$$w^1 \leftarrow w^0 - \eta \frac{\partial L}{\partial w} \Big|_{w=w^0, b=b^0}$$

$$b^1 \leftarrow b^0 - \eta \frac{\partial L}{\partial b} \Big|_{w=w^0, b=b^0}$$

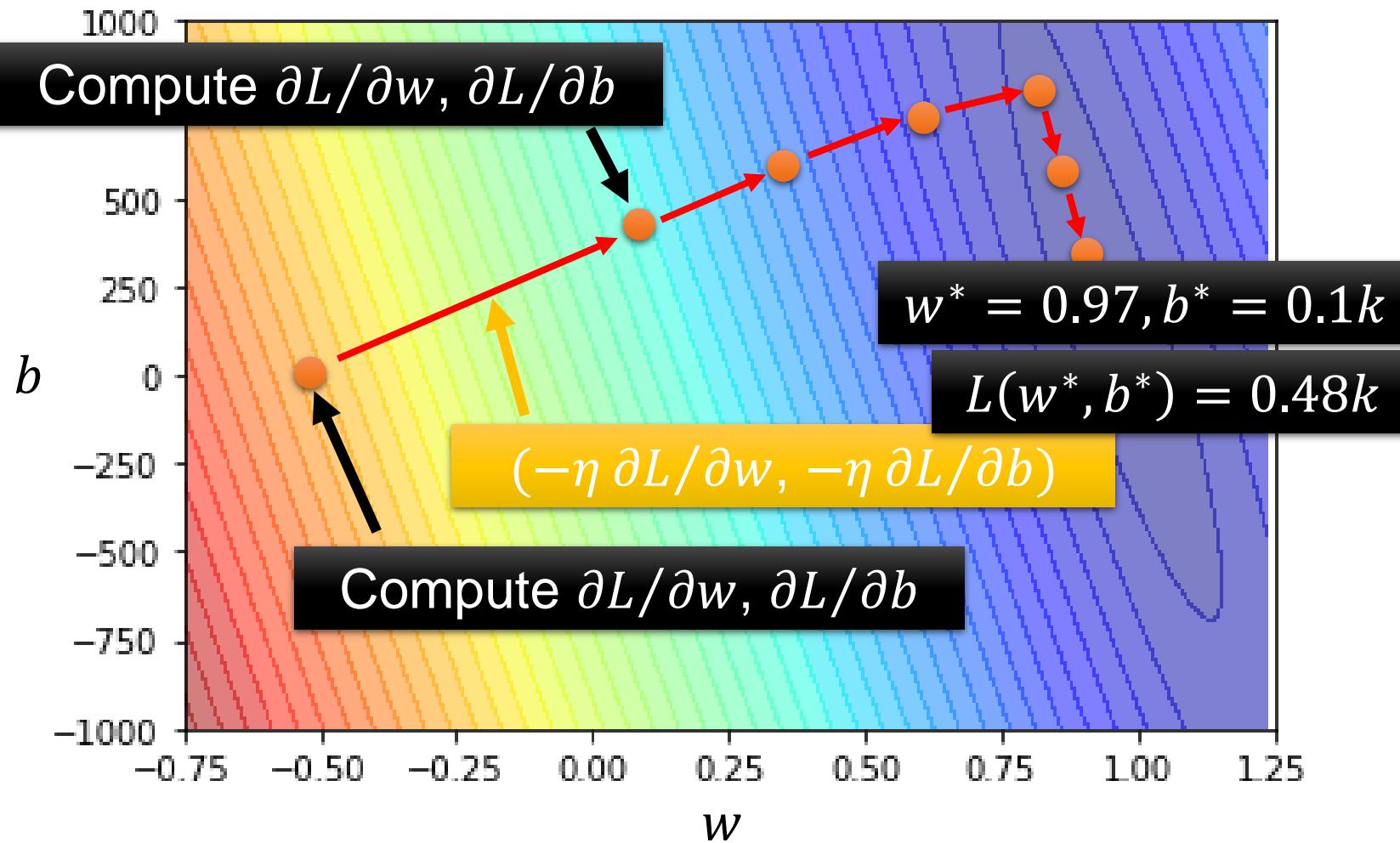
Can be done in one line in most deep learning frameworks

- Update w and b interatively

Model $y = b + wx_1$

3. Optimization

$$w^*, b^* = \arg \min_{w,b} L$$



Machine Learning is so simple

$$y = b + wx_1$$

Step 1:
function with
unknown

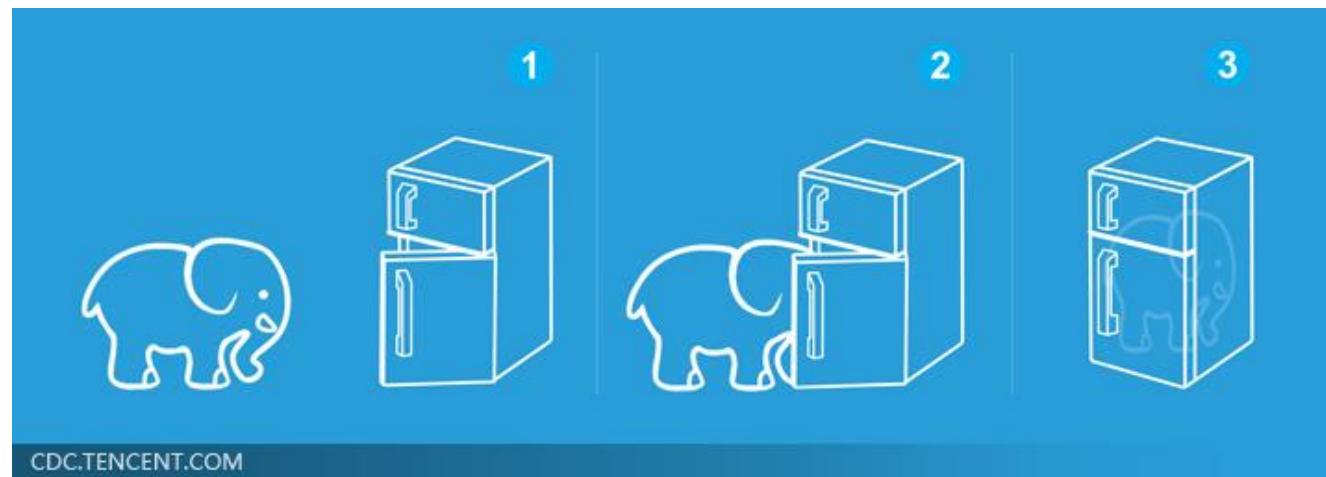


Step 2:
define loss
from training
data

$$w^* = 0.97, b^* = 0.1k$$

$$L(w^*, b^*) = 0.48k$$

Step 3:
optimization



Machine Learning is so simple



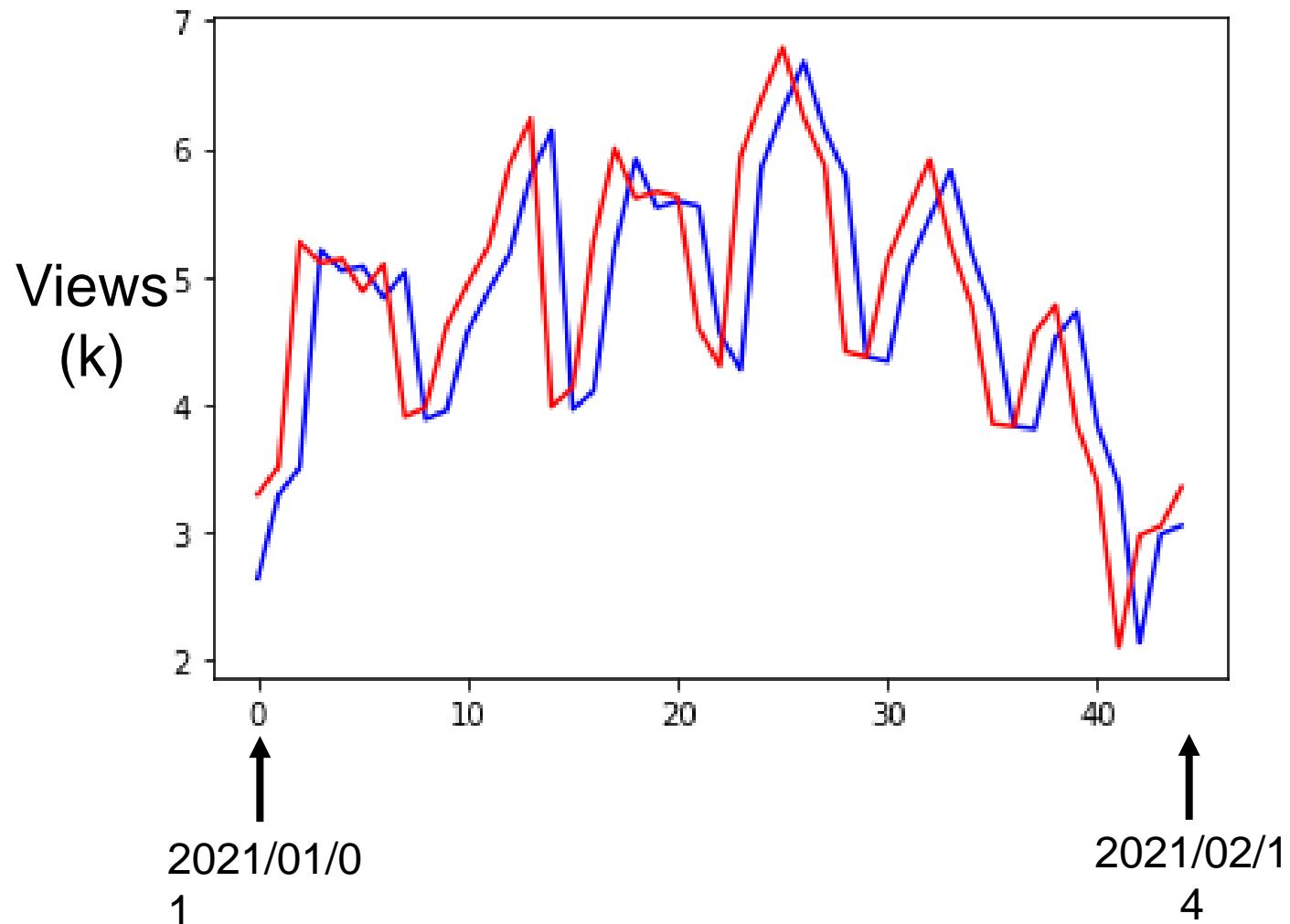
$y = 0.1k + 0.97x_1$ achieves the smallest loss $L = 0.48k$
on data of 2017 – 2020 (**training data**)

How about data of 2021 (**unseen during
training**)?

$$L' = 0.58k$$

$$y = 0.1k + 0.97x_1$$

Red: real no. of views
blue: estimated no. of views



$$y = b + w x_1$$

2017 - 2020

$$L = 0.48k$$

2021

$$L' = 0.58k$$

$$y = b + \sum_{j=1}^7 w_j x_j$$

2017 - 2020

$$L = 0.38k$$

2021

$$L' = 0.49k$$

| b | w_1^* | w_2^* | w_3^* | w_4^* | w_5^* | w_6^* | w_7^* |
|-------|---------|---------|---------|---------|---------|---------|---------|
| 0.05k | 0.79 | -0.31 | 0.12 | -0.01 | -0.10 | 0.30 | 0.18 |

$$y = b + \sum_{j=1}^{28} w_j x_j$$

2017 - 2020

$$L = 0.33k$$

2021

$$L' = 0.46k$$

$$y = b + \sum_{j=1}^{56} w_j x_j$$

2017 - 2020

$$L = 0.32k$$

2021

$$L' = 0.46k$$

Linear models

outline

1.1 机器学习的定义

1.2 机器学习的发展阶段

1.3 机器学习的种类

1.4 机器学习的三要素

1.5 机器学习的应用

1.6 常用第三方库介绍

1.7 机器学习的挑战



大数据时代，机器学习必不可少

收集、传输、存储大数据的目的，

是为了“利用”大数据

没有机器学习技术分析大数据，

“利用”无从谈起

机器学习能做什么？



小数据上就已经
很有用

例如：画作鉴别（艺术）

画作鉴别(painting authentication): 确定作品的真伪



出自 [J. Hughes et al., PNAS 2009]

勃鲁盖尔 (1525-1569)
的作品？



梵高 (1853-1890)
的作品？

出自 [C. Johnson et al., IEEE-SP, 2008]

例如：画作鉴别（艺术）

除专用技术手段外，**笔触分析**(brushstroke analysis)是画作鉴定的重要工具；它旨在从视觉上判断画作中是否具有艺术家的特有“笔迹”。

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

很难同时掌握不同时期、不同流派多位画家的绘画风格！

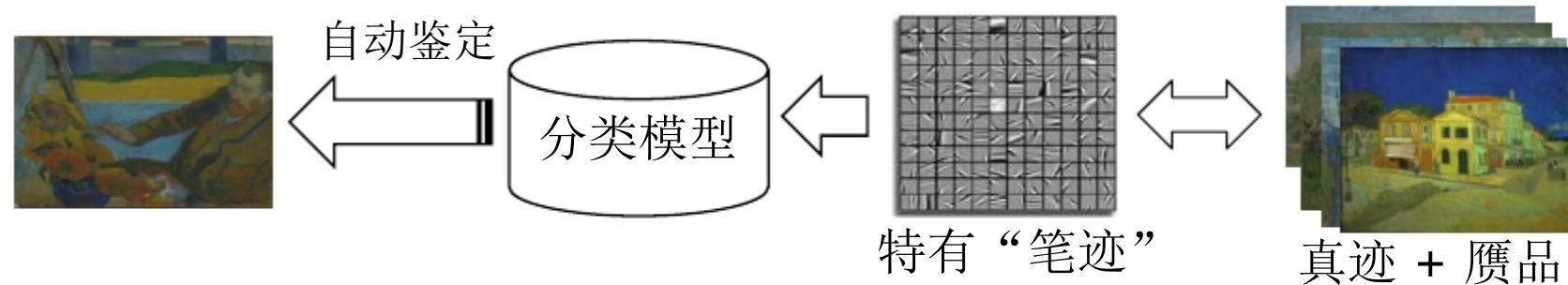


Portions of van Gogh paintings

只有少数专家花费很大精力
才能完成分析工作！

例如：画作鉴别（艺术）

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., IEEE-SP, 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., PNAS 2009][J. Mairal et al., PAMI'12]

(对用户要求低、准确高效、适用范围广)

例如：古文献修复（文化）

古文献是进行历史研究的重要素材，但是其中很多损毁严重

Dead Sea Scrolls (死海古卷)

- 1947年出土
- 超过30,000个羊皮纸片段



Cairo Genizah

- 19世纪末被发现
- 超过300,000个片段
- 散布于全球多家博物馆



高水平专家的大量精力
被用于古文献修复

例如：古文献修复（文化）

一个重要问题：

原书籍已经变成分散且混杂的多个书页，如何拼接相邻的书页？



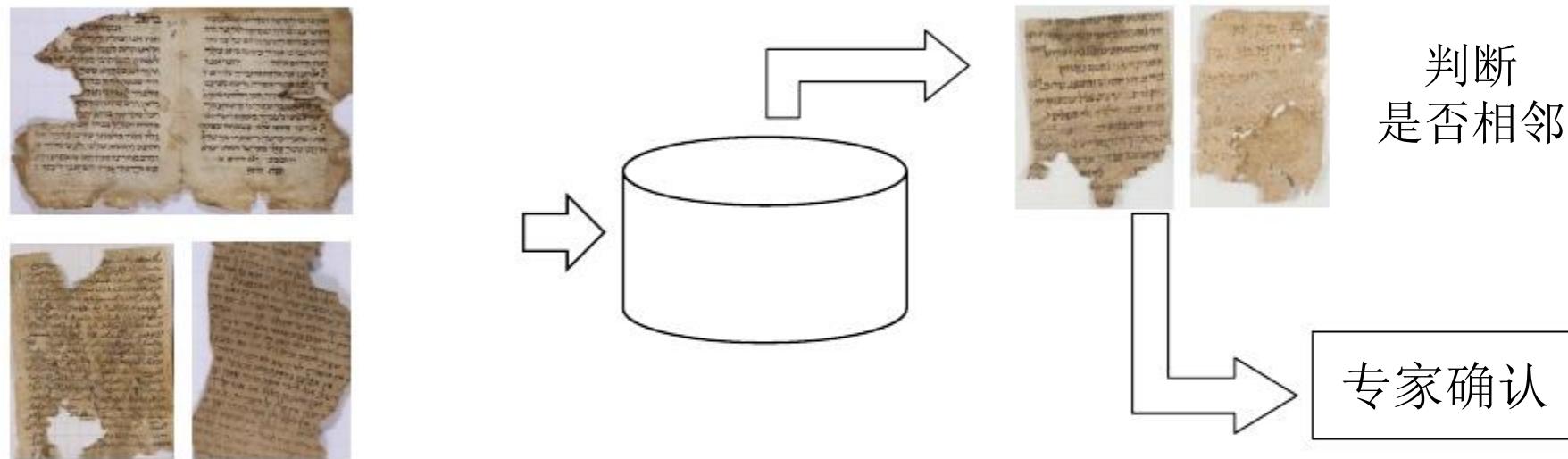
人工完成书页拼接十分困难

- 书页数量大，且分布在多处
- 部分损毁较严重，字迹模糊
- 需要大量掌握古文字的专业人才

近年来，古文献的数字化浪潮给自动文学修复提供了机会

例如：古文献修复（文化）

以色列特拉维夫大学的学者将机器学习用于自动的书页拼接



在Cairo Genizah测试数据上，系统的自动判断精度超过 **93%**

新完成约 1,000 篇Cairo Genizah文章的拼接

(对比：过去整个世纪，数百人类专家只完成了几千篇文章拼接)

机器学习能做什么？



大数据上更惊人

例如：帮助奥巴马胜选（政治）

How Obama's data crunchers helped him win

TIME

By Michael Scherer

November 8, 2012 – Updated 1645 GMT (0045 HKT) | Filed under: Web

《时代》周刊



例如：帮助奥巴马胜选（政治）

通过机器学习模型：

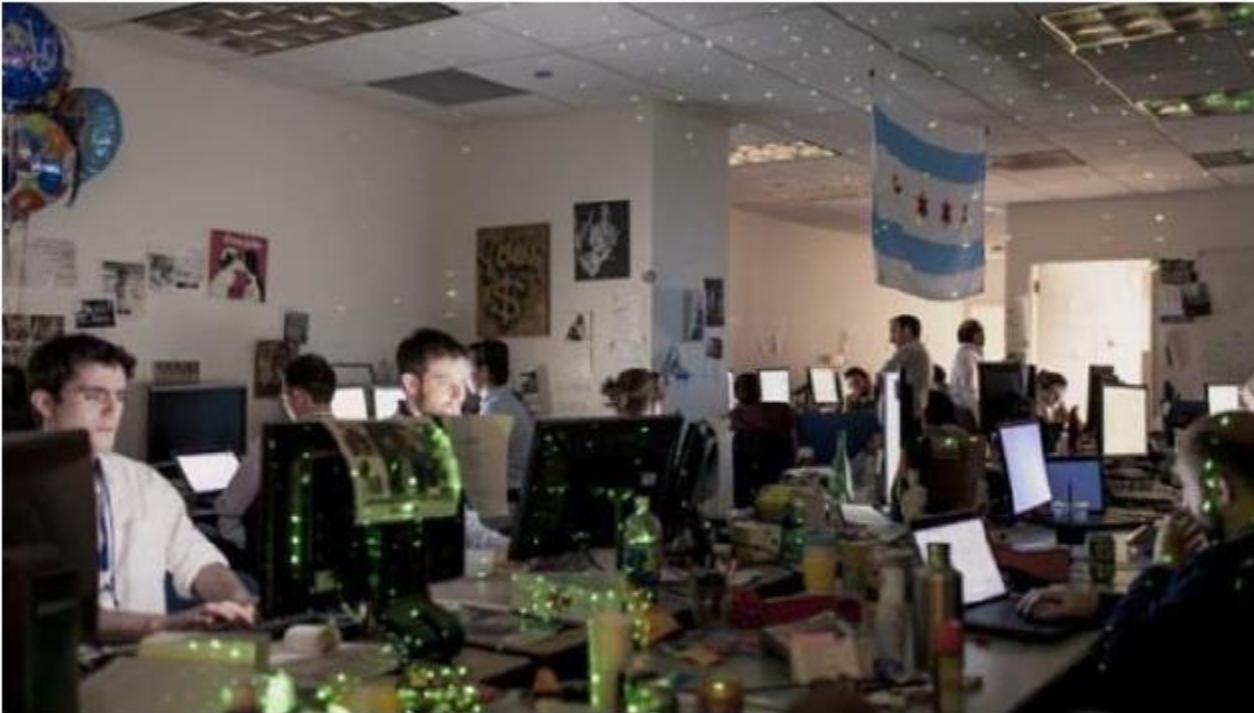
- ◆ 在总统候选人第一次辩论后，分析出哪些选民将倒戈，为每位选民找出一个最能说服他的理由
- ◆ 精准定位不同选民群体，建议购买冷门广告时段，广告资金效率比2008年提高14%
- ◆ 向奥巴马推荐，竞选后期应当在什么地方展开活动——那里有很多争取对象
- ◆ 借助模型帮助奥巴马筹集到创纪录的10亿美元

例如：利用模型分析出，明星乔治克鲁尼（George Clooney）对于年龄在40-49岁的美西地区女性颇具吸引力，而她们恰是最愿意为和克鲁尼/奥巴马共进晚餐而掏钱的人……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



- ◆

例如：帮助奥巴马胜选 (政治)

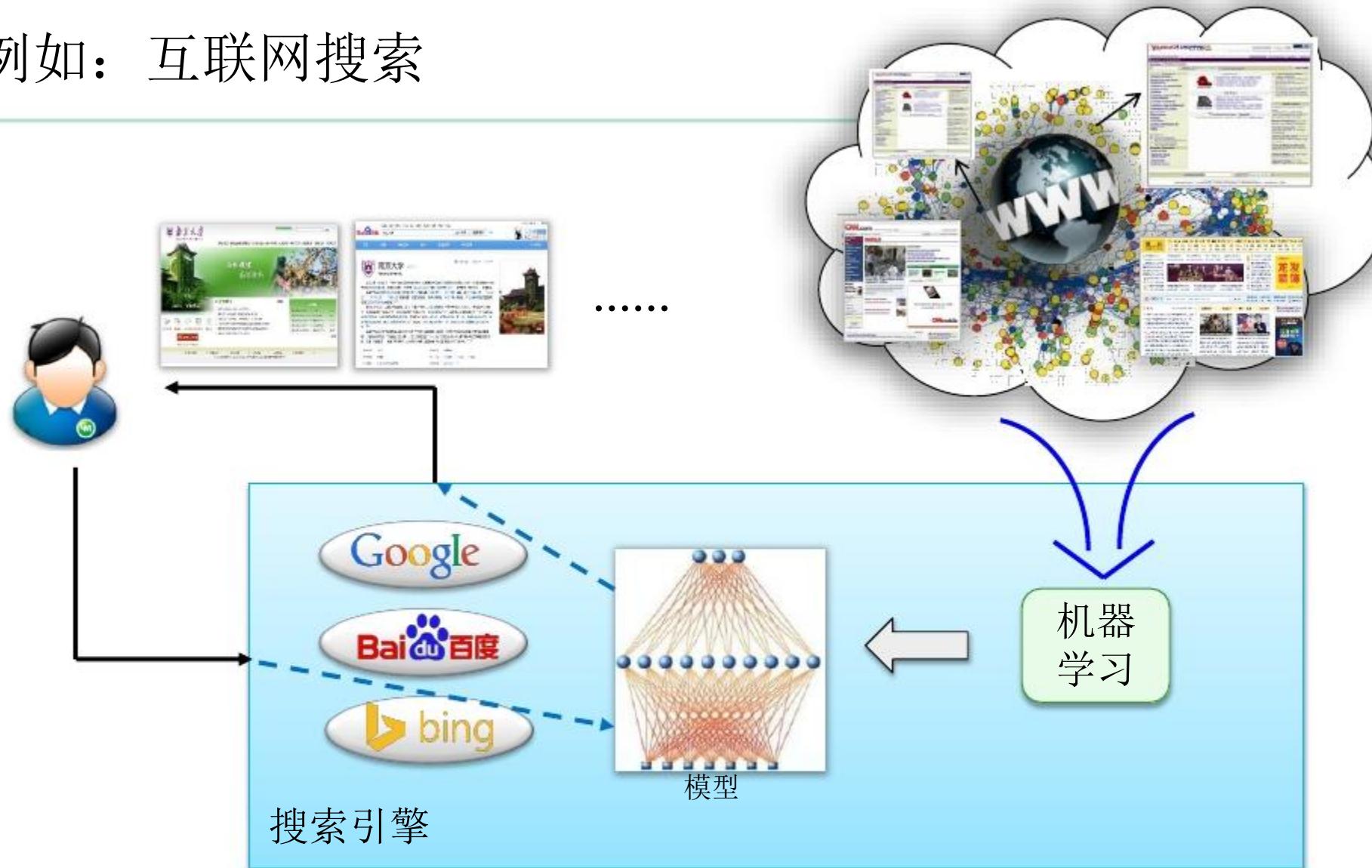


队长：Rayid Ghani

卡内基梅隆大学机器学习系
首任系主任Tom Mitchell
教授的博士生

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”)

例如：互联网搜索



机器学习技术正在支撑着各种搜索引擎

例如：AI+零售（经济）

“人工智能+零售”：消费场景全覆盖



得益于零售行业的数字化转型，人工智能已渗透到零售各个价值链环节，实现了消费场景流程的全覆盖。在新零售的商业图谱下，人工智能助力零售商强化与消费者的互动并提供个性化商品与服务；同时，通过消费者数据优化货架布局，提升坪效以节约成本。



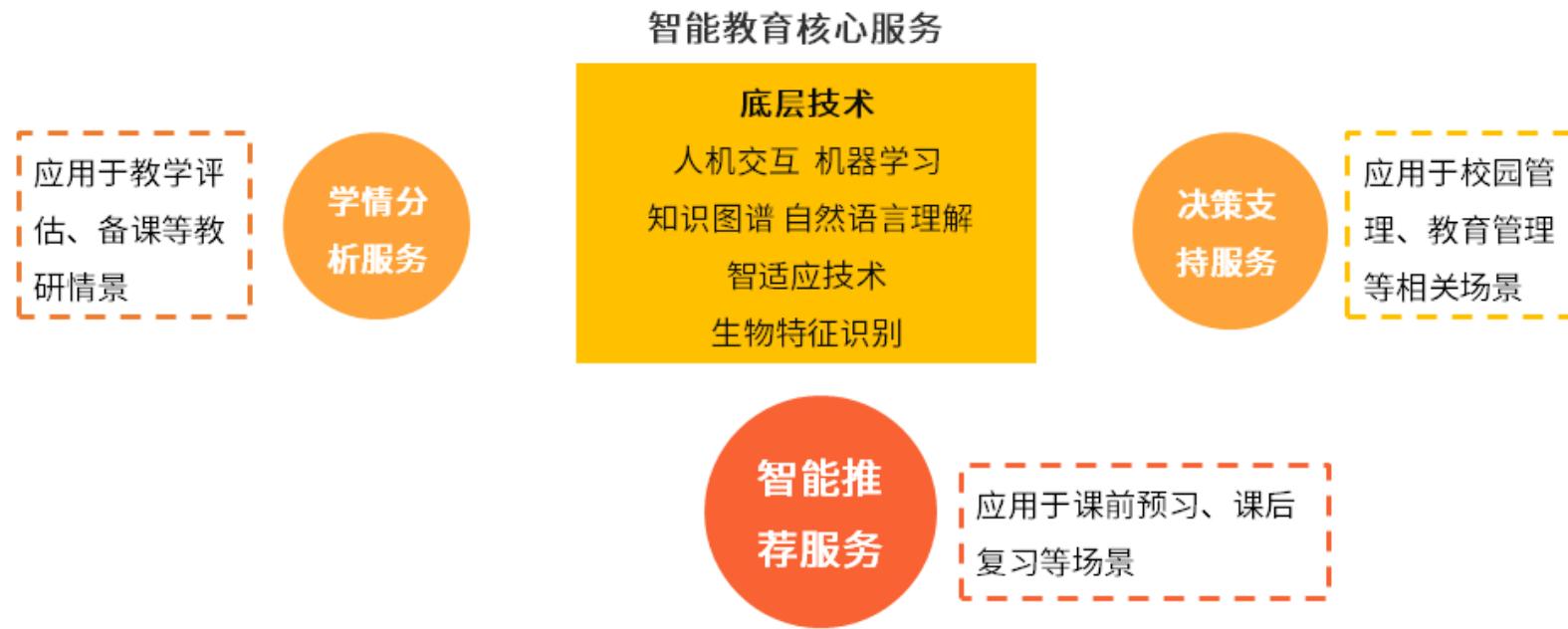
例如：AI+零售（经济）

“人工智能+零售”典型案例：云从科技+汽车零售



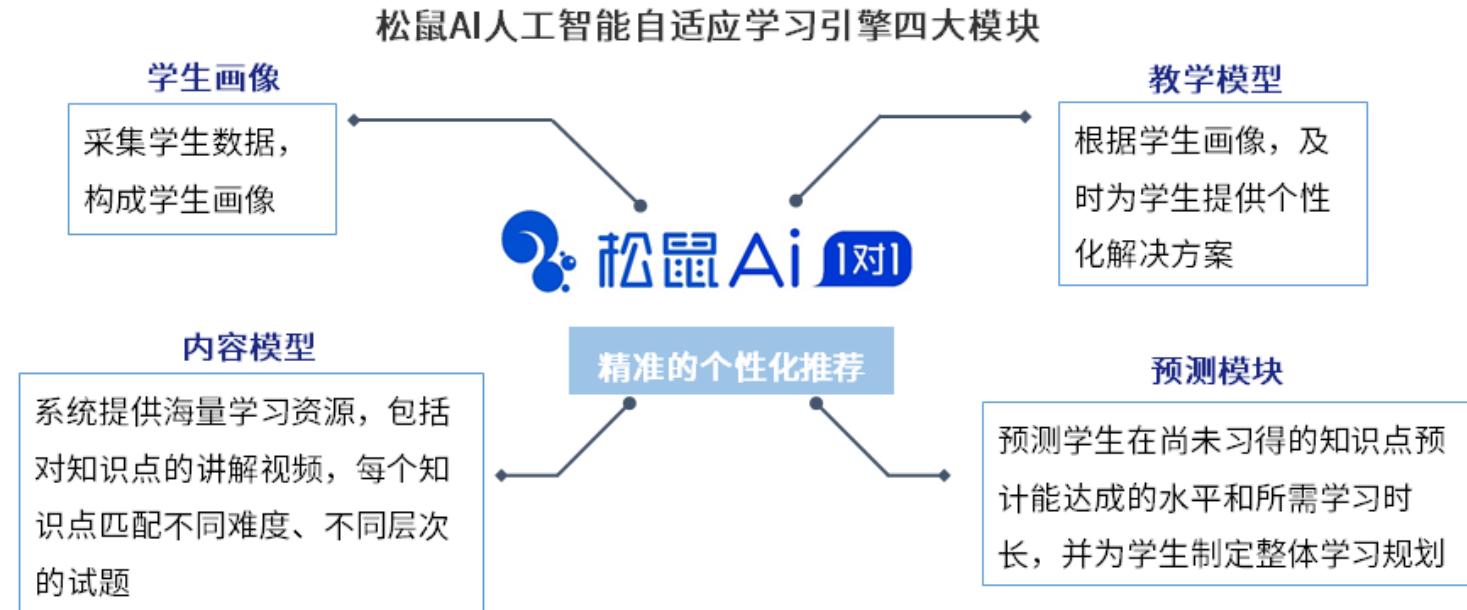
例如：AI+教育（教育）

“人工智能+教育”：AI赋能教育行业

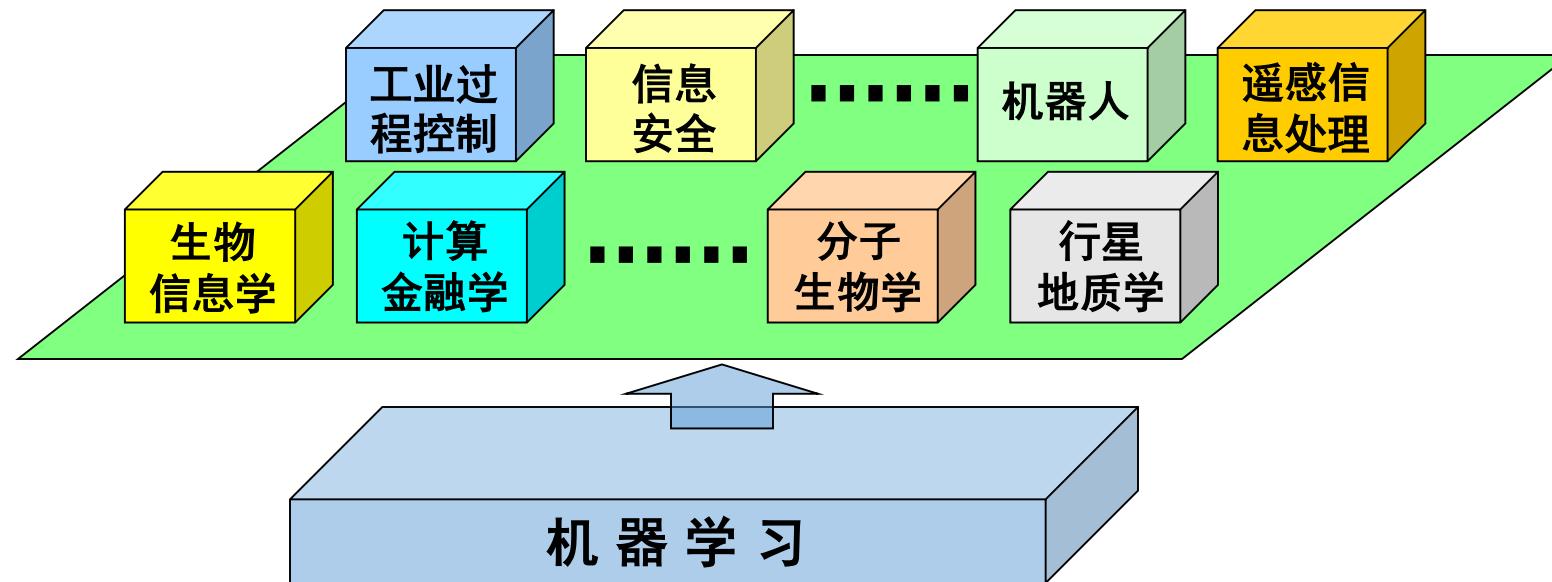


例如：AI+教育（教育）

“人工智能+中小学K12教育”典型案例：松鼠AI 1对1



机器学习的应用及意义



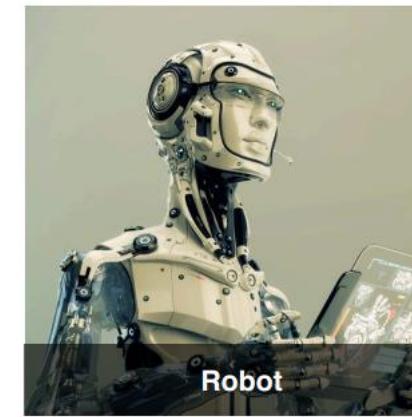
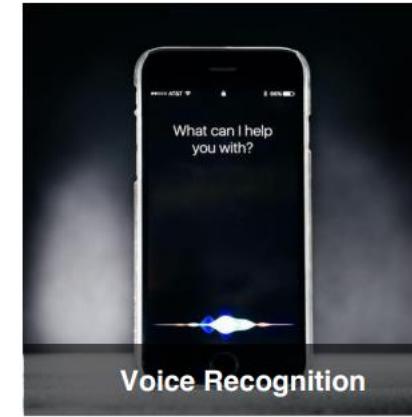
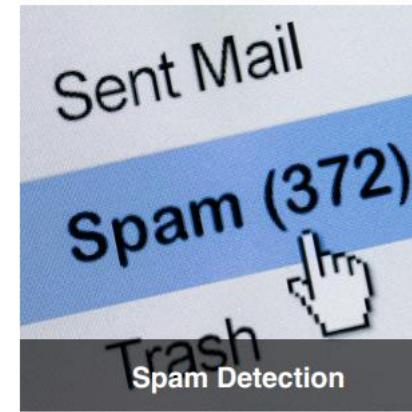
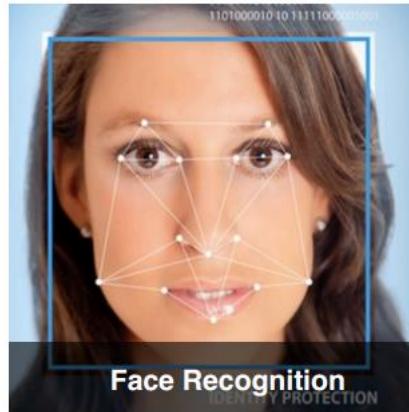
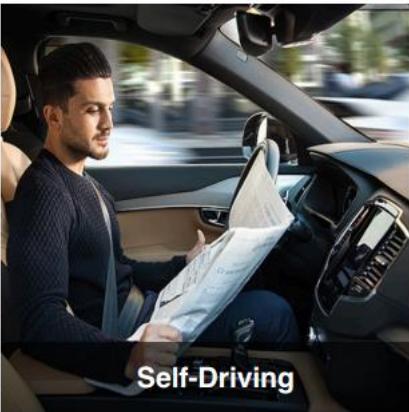
美国航空航天局JPL实验室的科学家在《Science》上撰文指出：机器学习对科学的研究的整个过程正起到越来越大的支持作用，……，该领域在今后的若干年内将取得稳定而快速的发展

问题2

你们在现实生活中还
遇到过哪些机器学习
应用？



机器学习已经“无处不在”



机器学习的应用场景



AI赋能各行各业



AI赋能各行各业



AI赋能各行各业



军事应用



百家号/奇点使者

outline

1.1 机器学习的定义

1.2 机器学习的发展阶段

1.3 机器学习的种类

1.4 机器学习的三要素

1.5 机器学习的应用

1.6 常用第三方库介绍

1.7 机器学习的挑战



了解ML常用类库

1. NumPy(Numerical Python)—— Python 科学计算的基础包

- 快速高效的多维数组对象 `ndarray`。
- 对数组执行元素级的计算以及直接对数组执行数学运算的函数。
- 读写硬盘上基于数组的数据集的工具。
- 线性代数运算、傅里叶变换，以及随机数生成的功能。
- 将 `C`、`C++`、`Fortran` 代码集成到 `Python` 的工具。



2. Pandas——数据分析核心库

- 提供了一系列能够快速、便捷地处理结构化数据的数据结构和函数。
- 高性能的数组计算功能以及电子表格和关系型数据库（如 SQL）灵活的数据处理功能。
- 复杂精细的索引功能，以便便捷地完成重塑、切片和切块、聚合及选取数据子集等操作。



3. Matplotlib——绘制数据图表的 Python 库

- Python的2D绘图库，非常适合创建出版物上用的图表。
- 操作比较容易，只需几行代码即可生成直方图、功率谱图、条形图、错误图和散点图等图形。
- 提供了pylab的模块，其中包括了NumPy和pyplot中许多常用的函数，方便用户快速进行计算和绘图。
- 交互式的数据绘图环境，绘制的图表也是交互式的。

4. scikit-learn——数据挖掘和数据分析工具

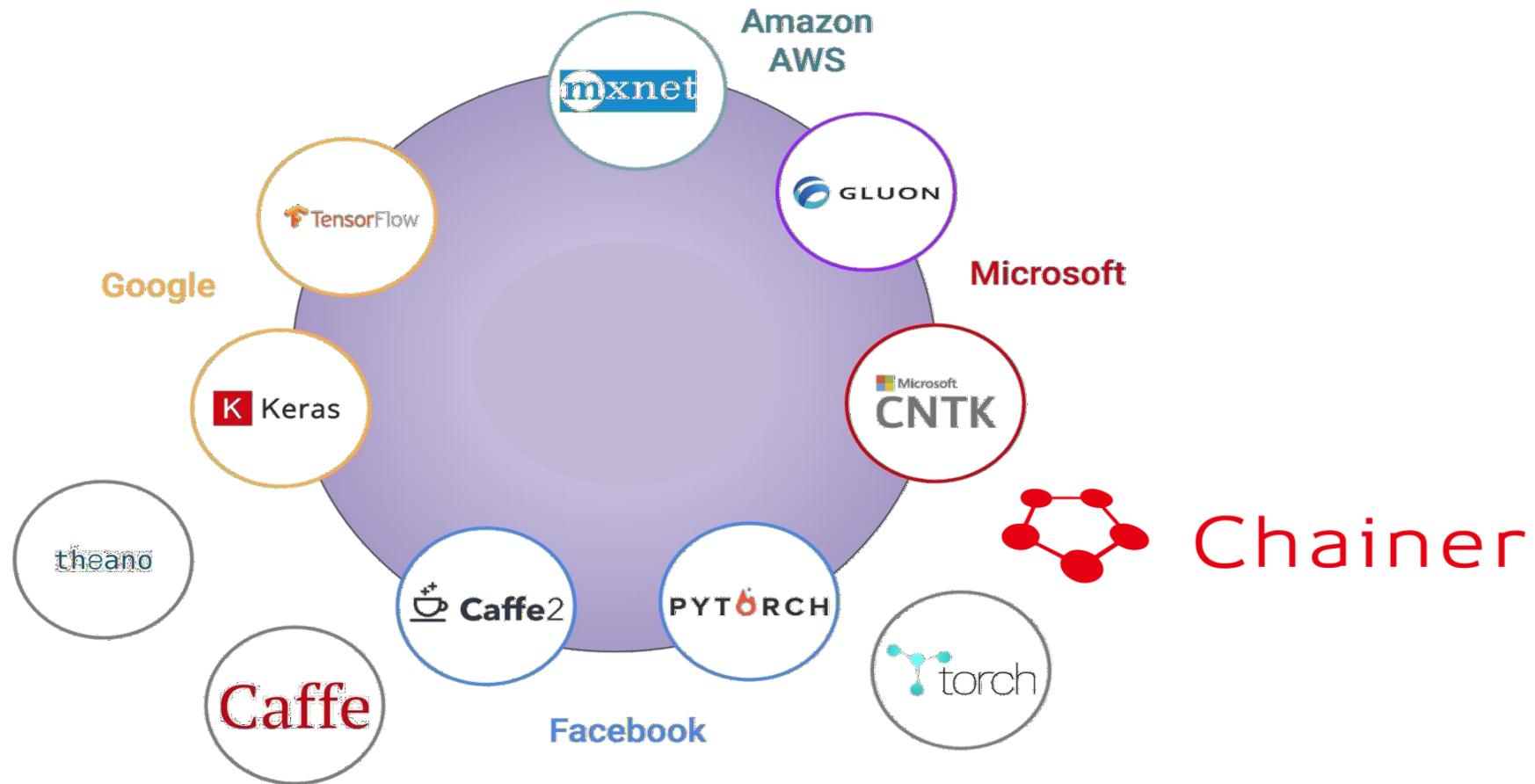
- 简单有效，可以供用户在各种环境下重复使用。
- 封装了一些常用的算法方法。
- 基本模块主要有数据预处理、模型选择、分类、聚类、数据降维和回归 6 个，在数据量不大的情况下，scikit-learn可以解决大部分问题。

参考网址：

<https://scikit-learn.org/stable/>
<https://www.scikitlearn.com.cn/>



5. pytorch、tensorflow、keras、caffe等——深度学习框架



1.1 机器学习的定义

1.2 机器学习的发展阶段

1.3 机器学习的种类

1.4 机器学习的三要素

1.5 机器学习的应用

1.6 常用第三方库介绍

1.7 机器学习的挑战



机器学习的五个挑战问题

挑战问题(1): 泛化能力

- 共性问题:
 - 几乎所有的领域，都希望越准越好
 - 提高泛化能力是永远的追求
 - 目前泛化能力最强的技术：
支持向量机 (SVM) 产生途径：理论→实践
集成学习 (ensemble learning) 产生途径：实践→理论
 - 第一个挑战问题：
 - 今后10年能否更“准”？
 - 如果能，会从哪儿来？

挑战问题(2): 速度

- 共性问题:
 - 几乎所有的领域，都希望越快越好
 - 加快速度也是永远的追求
- “训练速度” vs. “测试速度”
 - 训练速度快的往往测试速度慢: k近邻
 - 测试速度快的往往训练速度慢: 神经网络
- 第二个挑战问题:
 - 今后10年能否更“快”？
 - 能做到“训练快”、“测试也快”吗？
 - 如果能，如何做？

挑战问题(3): 可理解性

- 共性问题:
 - 绝大多数领域都希望有“可理解性”
 - 例: 医疗诊断、地震预测
 - 目前强大的技术几乎都是(或基本上是)“黑盒子”
神经网络、支持向量机、集成学习
- “黑盒子”能满足需要吗?
- 第三个挑战问题:
 - 今后10年能否产生“白盒子”?
 - 是和“黑盒子”完全不同的东西,还是从“黑盒子”变出来?

挑战问题(4)：数据利用能力

- 传统的机器学习技术
 - 对有标记数据进行学习（“标记”——事件所对应的结果）
- 共性问题：
 - 随着数据收集能力飞速提高、Internet的出现，在大多数领域中都可以很容易地获得大量未标记数据
- 没有标记的数据是没用的吗？
- 共性问题：
 - 在绝大多数领域中都会遇到“坏”数据，有时甚至只有“坏”数据（大量噪音、属性缺失、不一致、……）
 - 传统的“坏”数据处理方式：“扔掉”
- “坏”数据一点用也没有吗？
- 第四个挑战问题：
 - 今后10年能否“数据通吃”？
 - 如何“吃”？

挑战问题(5): 代价敏感

- 目前的机器学习技术
 - 降低错误率
 - “错误”是没有区别的吗?
 - 把“好”当成“坏”
 - 把“坏”当成“好”
- 共性问题:
 - 大多数领域中的错误代价都不一样的
- 第五个挑战问题:
 - 今后10年能否“趋利避害”?
 - 在达到较低的总错误率的基础上，如何“趋”、如何“避”?

A.I.



大数据，成就未来



Thank you!
感谢您！