

第1.1讲 机器学习绪论

主讲教师：欧新宇

February 21, 2020

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.)

2011
年度

“计算学习理论”奠基人



Judea Pearl
(1936 -)
(UCLA)

2012
年度

“图模型学习方法”先驱

机器学习 (Machine Learning)

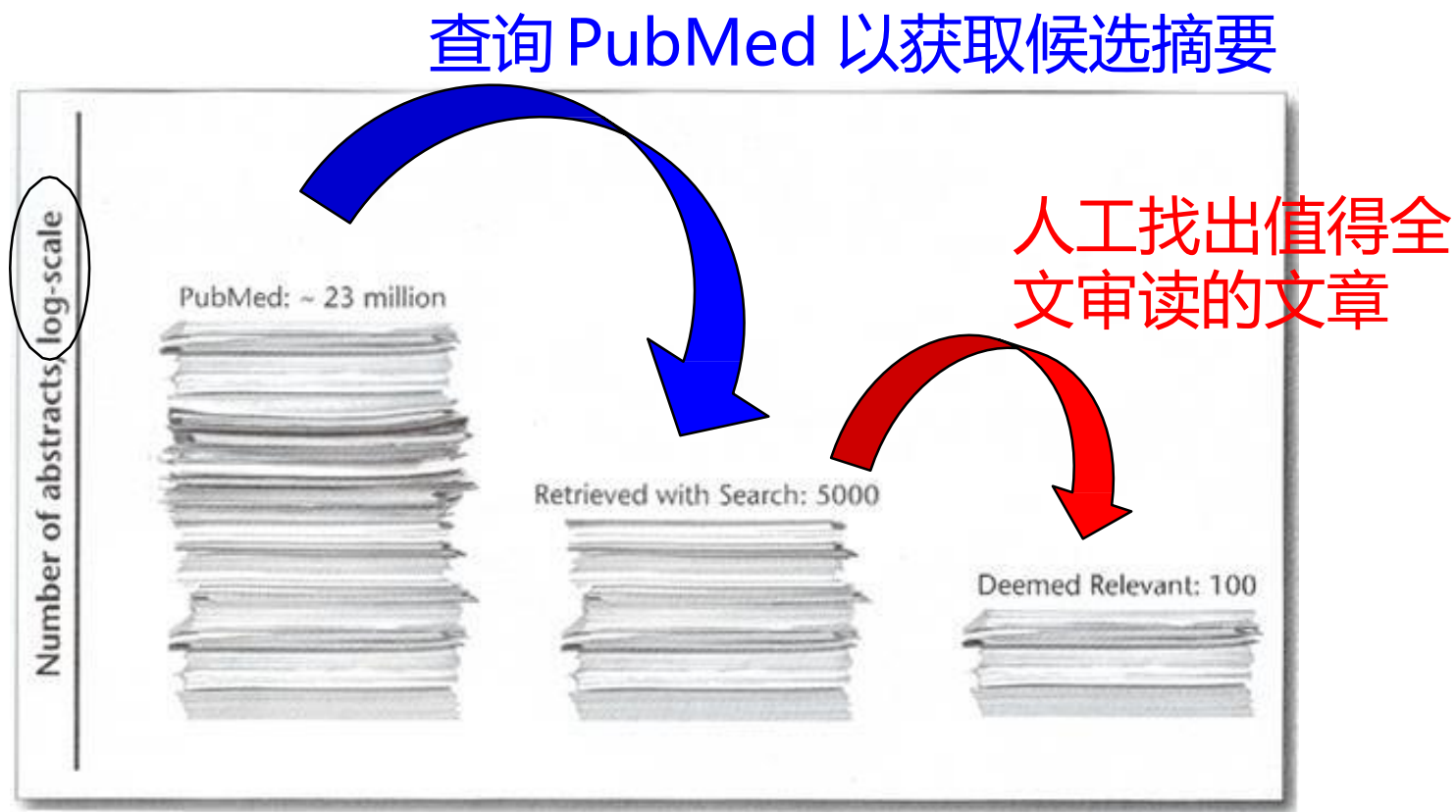
究竟是什么东东?



看个例子 ➡

“文献筛选”的故事

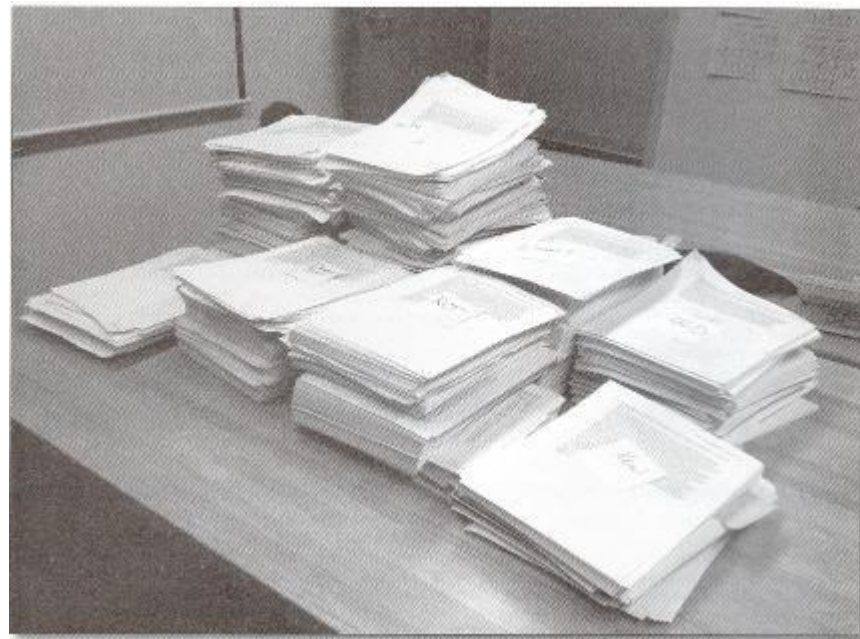
在“循证医学”（evidence-based medicine）中，针对特定的临床问题，先要对相关研究报告进行详尽评估。



“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。



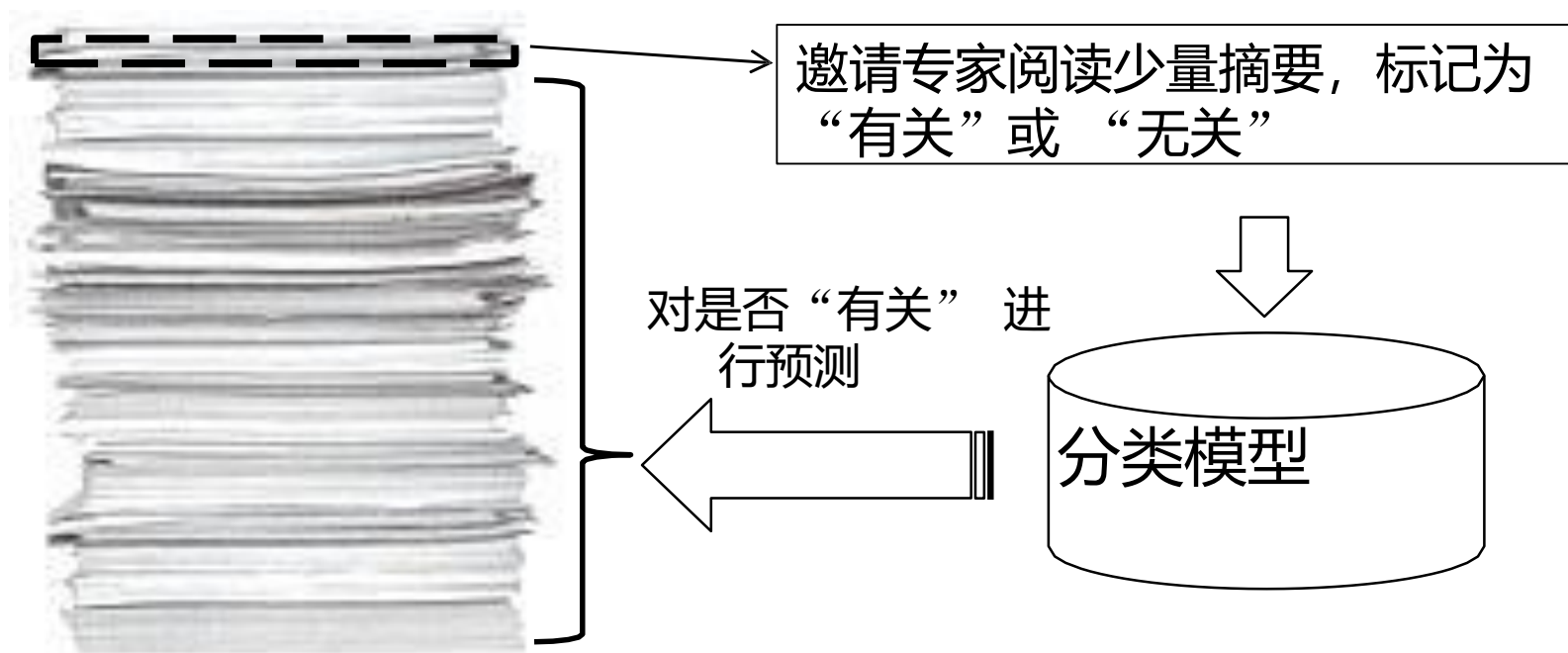
a portion of the 33,000 abstracts

**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！

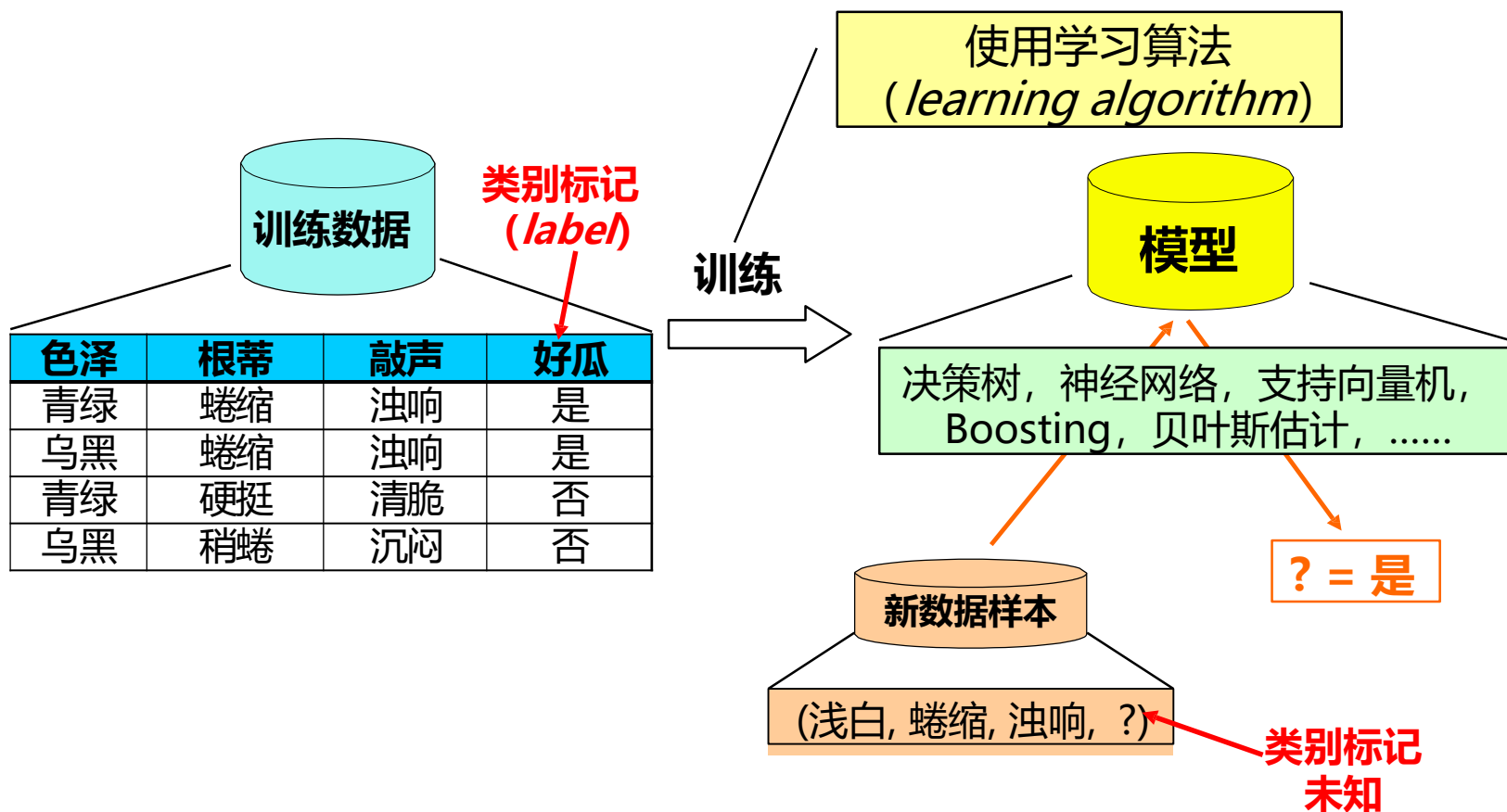
“文献筛选”的故事

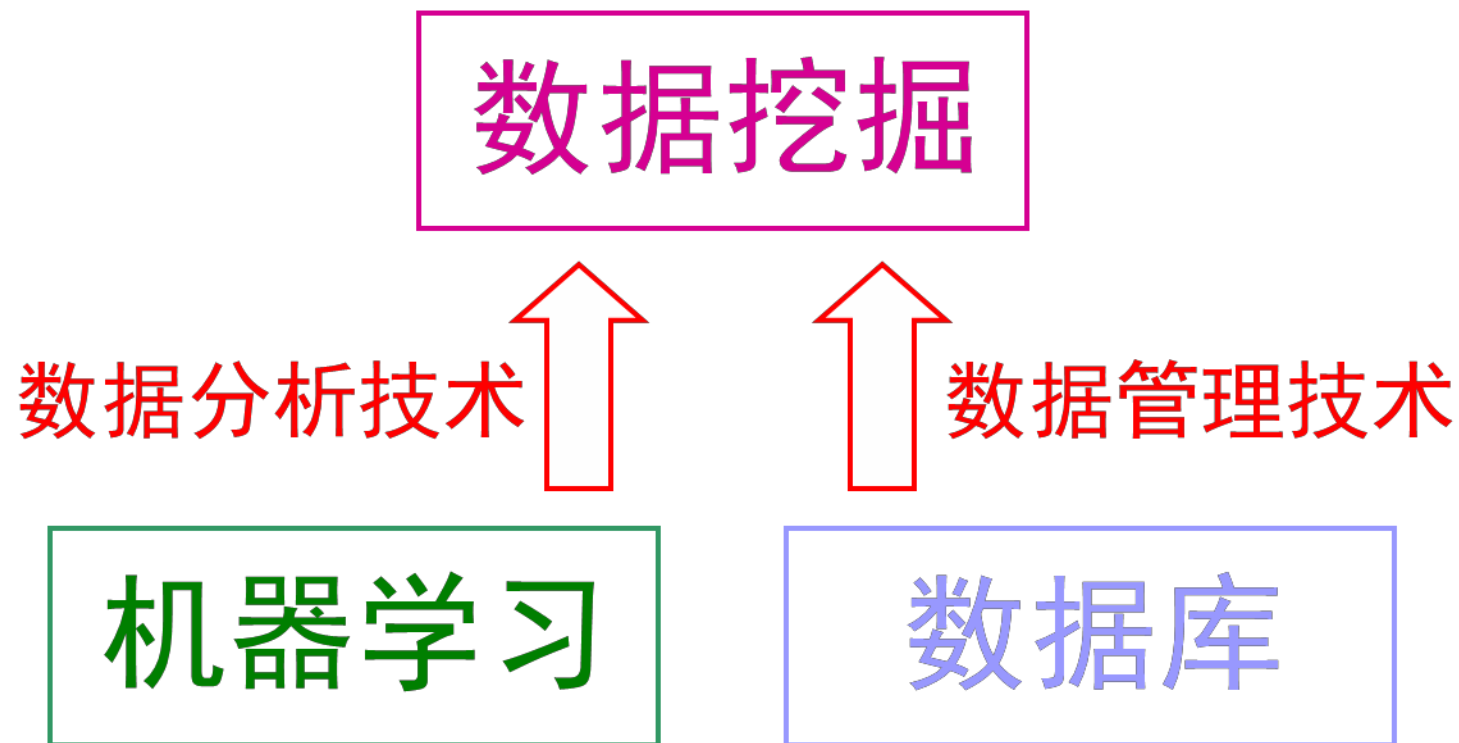
为了降低昂贵的成本, Tufts医学中心引入了机器学习技术。



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)

典型的机器学习过程

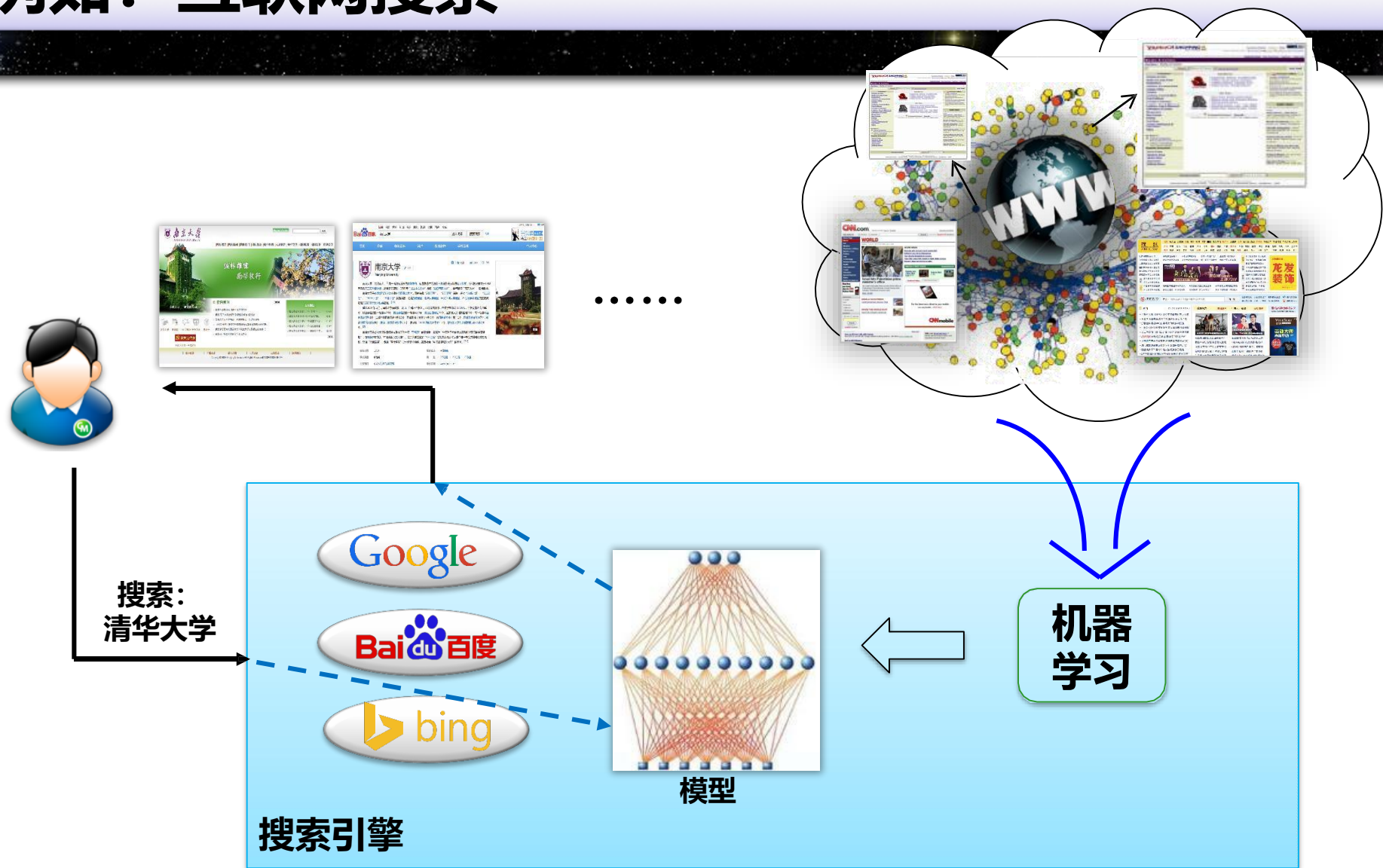




机器学习能做什么？

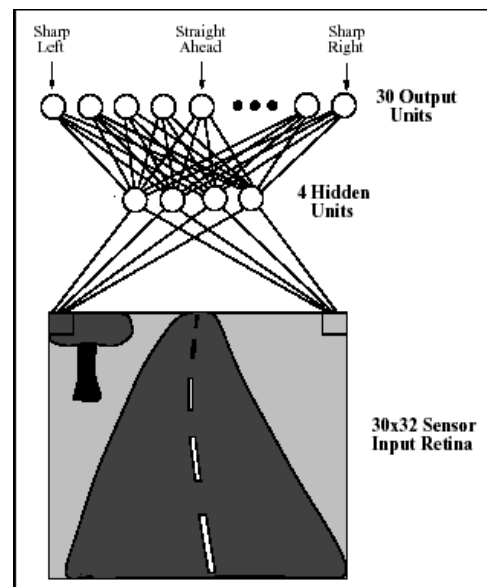
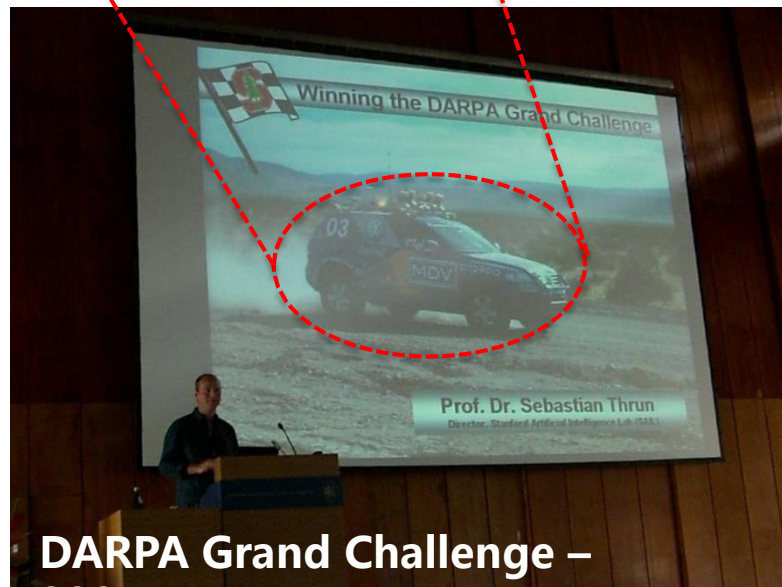
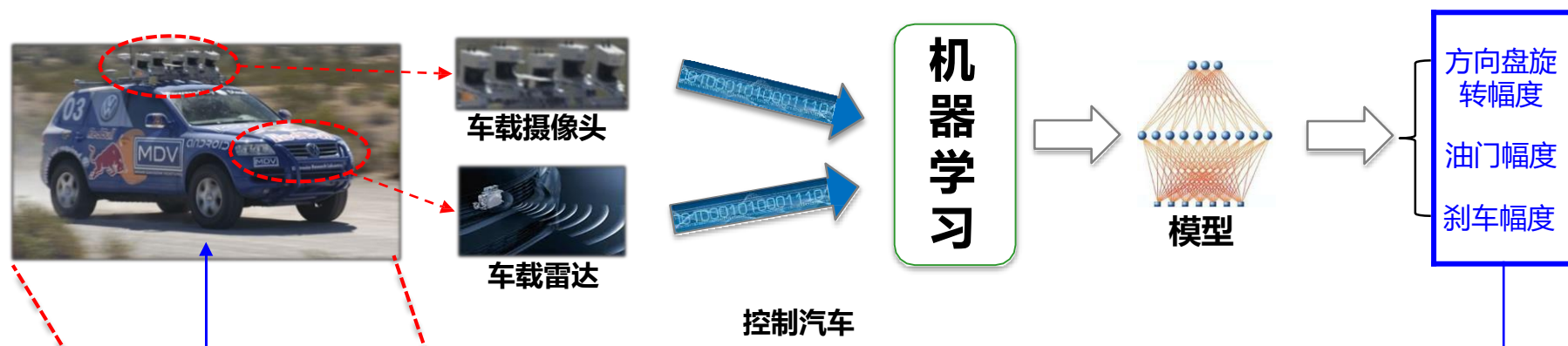
我们可能每天都在
在用机器学习

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：自动汽车驾驶（即将改变人类生活）



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

机器学习还能做什么？



看看它在
小数据上的应用

例如：画作鉴别 (艺术)

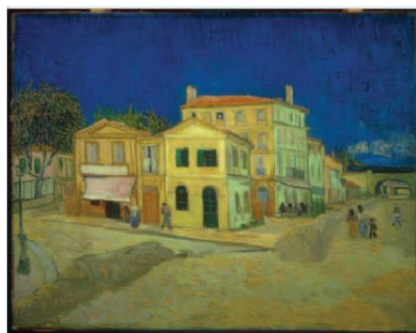
画作鉴别(painting authentication): 确定作品的真伪



勃鲁盖尔 (1525–1569)
的作品?

出自 [J. Hughes et al., PNAS 2009]

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



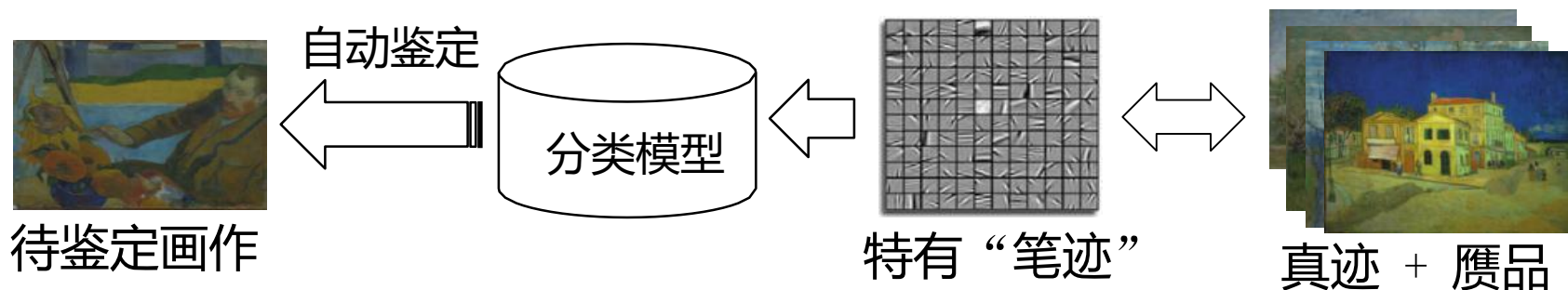
Portions of van Gogh paintings

**只有少数专家花费很大精力
才能完成分析工作！**

很难同时掌握不同时期、不同流派多位画家的绘画风格！

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**

[L. Wolf et al., IJCV 2011]

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？



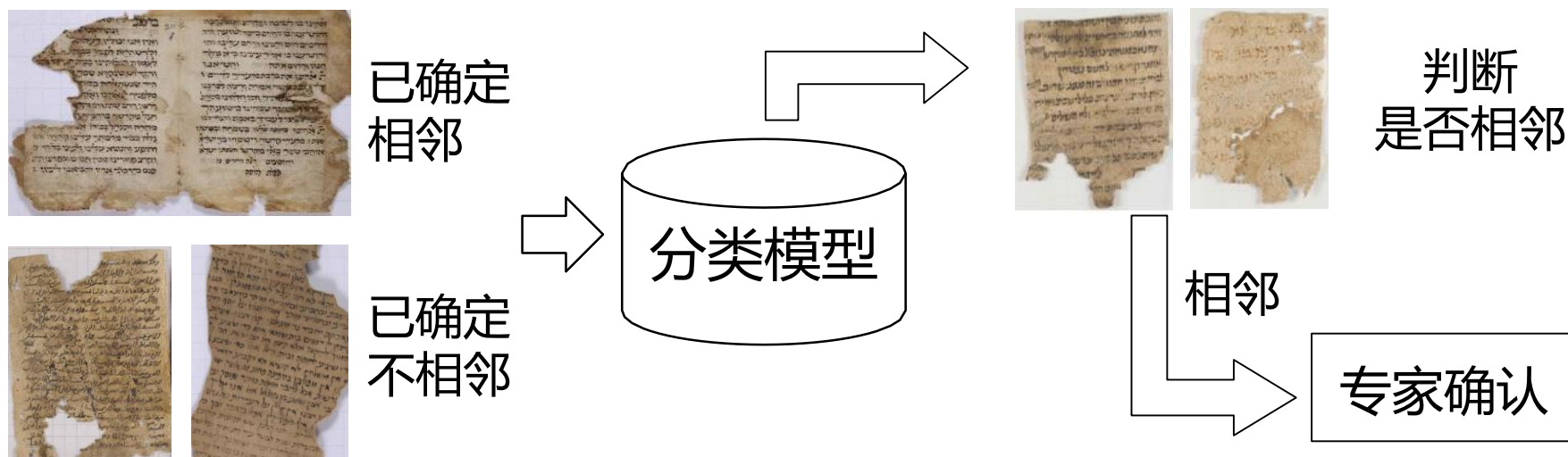
人工完成书页拼接十分困难

- 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）

机器学习还能做什么？

再看看它在大数
据上的惊人表现

例如：帮助奥巴马胜选（政治）

How Obama's data crunchers helped him win

《时代》周刊

TIME

By Michael Scherer

November 8, 2012 — Updated 1645 GMT (0045 HKT) | Filed under: [Web](#)



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

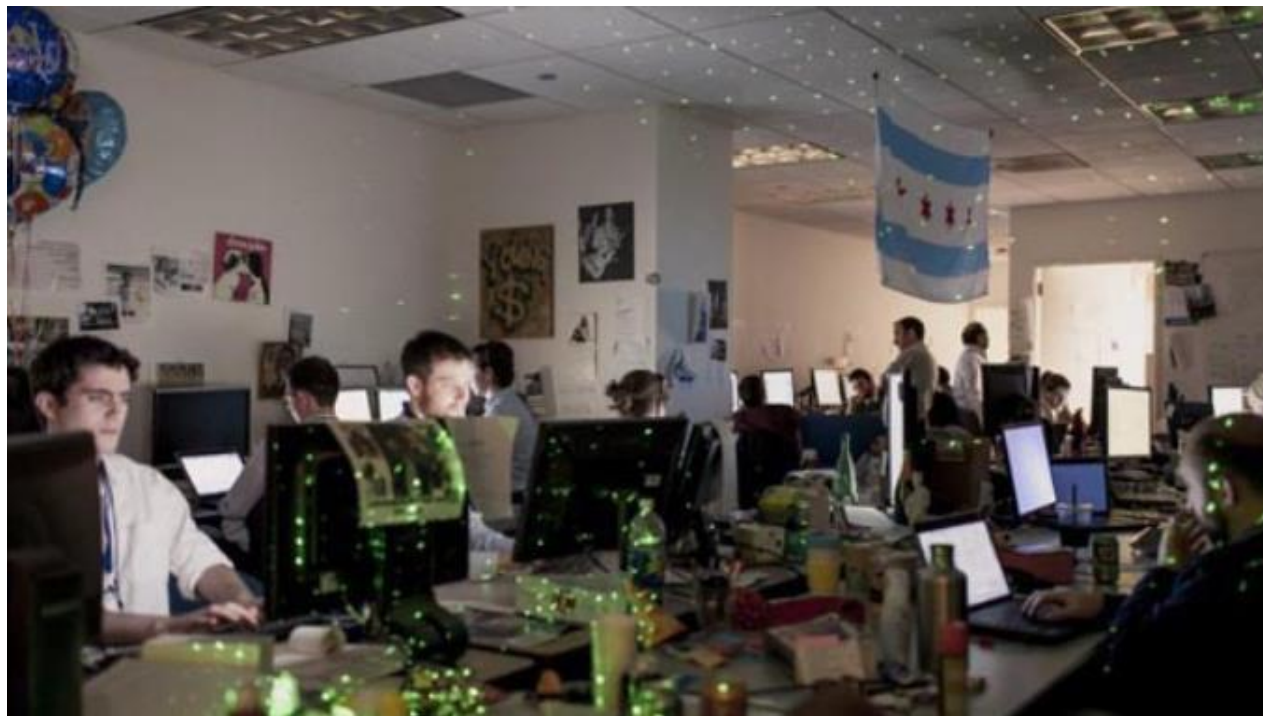
例如：利用模型分析出，明星乔治克鲁尼（George Clooney）对于年龄在40-49岁的美西地区女性颇具吸引力，而她们恰是最愿意为和克鲁尼/奥巴马共进晚餐而掏钱的人.....

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，定期向奥巴马报送各种预测结果；
被奥巴马公开称为总统竞选的“核武器按钮”
(“They are our nuclear codes”)

机器学习源自“人工智能”



约翰 麦卡锡
(1927-2011)
“人工智能之父”
1971年图灵奖

Artificial Intelligence (AI), 1956 -

1956年夏 美国达特茅斯学院

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

达特茅斯会议标志着人工智能这一学科的诞生

John McCarthy (1927 - 2011):

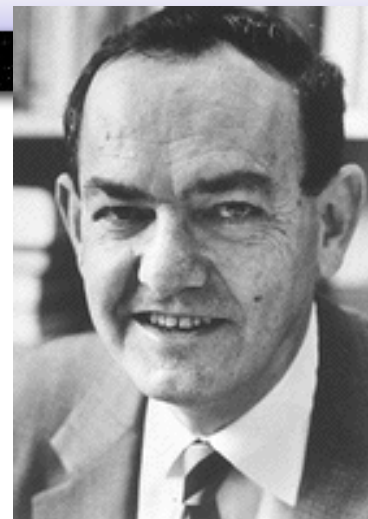
1971年获图灵奖，1985年获IJCAI终身成就奖。人工智能之父。他提出了“人工智能”的概念，设计出函数型程序设计语言Lisp，发展了递归的概念，提出常识推理和情境演算。出生于共产党家庭，从小阅读《10万个为什么》，中学时自修CalTech的数学课程，17岁进入CalTech时免修两年数学，22岁在Princeton获博士学位，37岁担任Stanford大学AI实验室主任。

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



爱德华 费根鲍姆
(1936-)
1994年图灵奖

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）

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

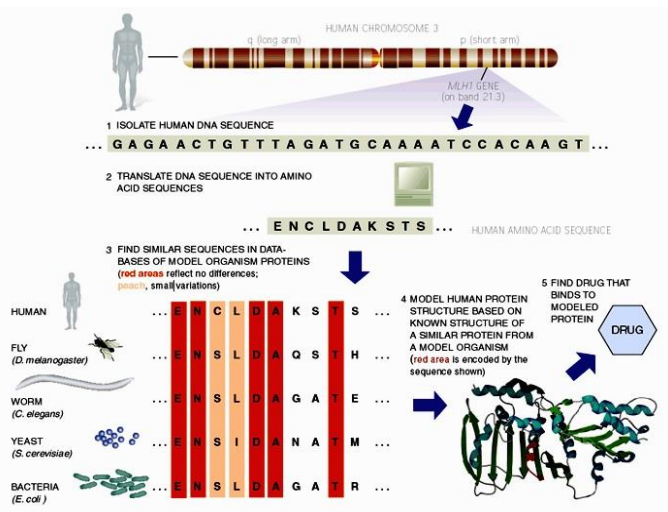
- ◆ 出发点：“让系统自己学！”
- ◆ 主要成就：……

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

机器学习已经“无处不在”



生物信息学



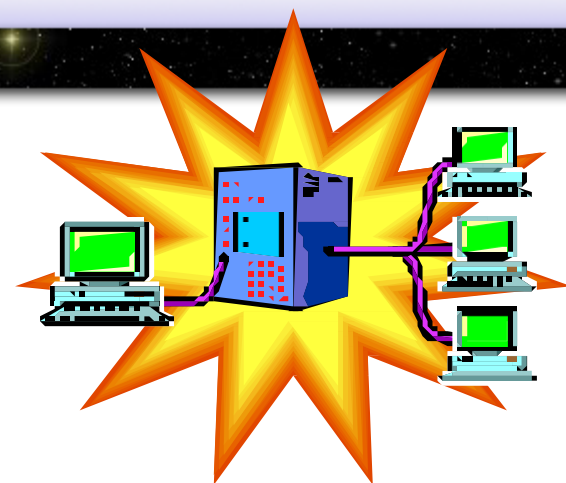
汽车自动驾驶
(DARPA Grand Challenge)



Web搜索



火星机器人 (JPL)



决策助手(DARPA)

机器学习已经“无处不在”

今天的“机器学习”已经是一个广袤的学科领域

例如，这是第32届国际机器学习大会的“主题领域”

2006年，美国CMU
(卡内基梅隆大学)
成立“机器学习系”

- ☐ Active Learning
- ☐ Approximate Inference
- ☐ Bayesian Nonparametric Methods
- ☐ Bioinformatics
- ☐ Causal Inference
- ☐ Clustering
- ☐ Computational Social Sciences
- ☐ Cost-Sensitive Learning
- ☐ Digital Humanities
- ☐ Ensemble Methods
- ☐ Feature Selection and Dimensionality Reduction
- ☐ Finance
- ☐ Gaussian Processes
- ☐ Graphical Models
- ☐ Inductive Logic Programming and Relational Learning
- ☐ Information Retrieval
- ☐ Kernel Methods
- ☐ Large-Scale Machine Learning
- ☐ Latent Variable Models
- ☐ Learning for Games
- ☐ Learning Theory
- ☐ Manifold Learning
- ☐ Network and Graph Analysis
- ☐ Neural Networks and Deep Learning
- ☐ Planning and Control
- ☐ Privacy, Anonymity, and Security
- ☐ Ranking and Preference Learning
- ☐ Recommender Systems
- ☐ Reinforcement Learning
- ☐ Robotics
- ☐ Rule and Decision Tree Learning
- ☐ Semi-Supervised Learning
- ☐ Sparsity and Compressed Sensing
- ☐ Spectral Methods
- ☐ Speech Recognition
- ☐ Statistical Relational Learning
- ☐ Structured Output Prediction
- ☐ Supervised Learning
- ☐ Sustainability, Climate, and Environment
- ☐ Time-Series Analysis

经常被谈到的“深度学习”
(Deep Learning)仅是
机器学习中的一个小分支

大数据时代的关键技术



奥巴马提出“大数据计划”后，美国NSF进一步加强资助UC Berkeley研究如何整合将“数据”转变为“信息”的三大关键技术——机器学习、云计算、众包(crowd sourcing)

National Science Foundation: In addition to funding the Big Data solicitation, and keeping with its focus on basic research, NSF is implementing a comprehensive, long-term strategy that includes new methods to derive knowledge from data; infrastructure to manage, curate, and serve data to communities; and new approaches to education and workforce development. Specifically, NSF is:

整合三大关键技术

- Encouraging research universities to develop interdisciplinary graduate programs to prepare the next generation
- Funding a million Expeditions in Computing project based at the University of California, Berkeley, that will integrate three powerful approaches for turning data into information - machine learning, cloud computing, and crowd sourcing;
- Providing the first round of grants to support “EarthCube” – a system that will allow geoscientists to access, analyze and share information about our planet;
- Issuing a \$2 million award for a research training group to support training for undergraduates to use graphical and visualization techniques for complex data.
- Providing \$1.4 million in support for a focused research group of statisticians and biologists to determine protein structures and biological pathways.
- Convening researchers across disciplines to determine how Big Data can transform teaching and learning.

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，

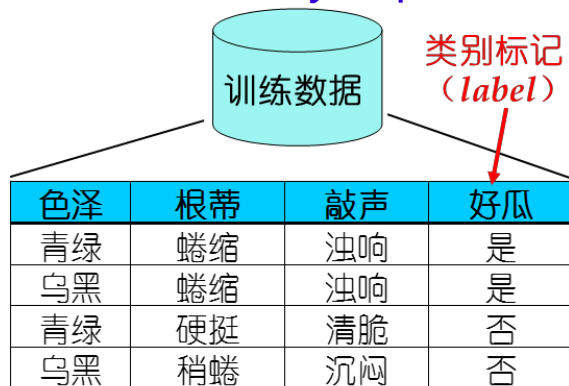
是为了“利用”大数据

没有机器学习技术分析大数据，

“利用”无从谈起

基本术语

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)
- 半监督学习(semi-supervised learning)
- 弱监督学习(Weakly Supervised Learning)



训练

使用学习算法 (learning algorithm)



决策树, 神经网络, 支持向量机,
Boosting, 贝叶斯网,

新数据样本

(浅白, 蜷缩, 浊响, ?)

? = 是

类别标记
未知

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

- 分类, 回归
- 二分类, 多分类
- 正类, 反类

- 数据集; 训练, 测试
- 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

- 未见样本(unseen instance)
- 未知“分布”
- 独立同分布(i.i.d.)
- **泛化(generalization)**

表 1.1 西瓜数据集

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

$$(\text{色泽}=\text{?}) \wedge (\text{根蒂}=\text{?}) \wedge (\text{敲声}=\text{?}) \leftrightarrow \text{好瓜}$$

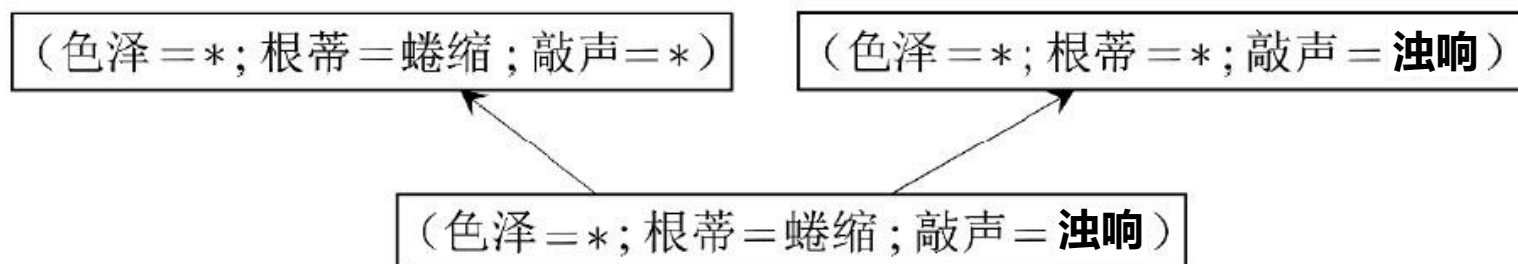
学习过程 →

在所有假设(hypothesis)组成的空间中进行搜索的过程

目标: 找到与训练集 “匹配” (fit)的假设

假设空间的大小: $n_1 \times n_2 \times n_3 + 1$

版本空间(version space): 与训练集一致的假设集合



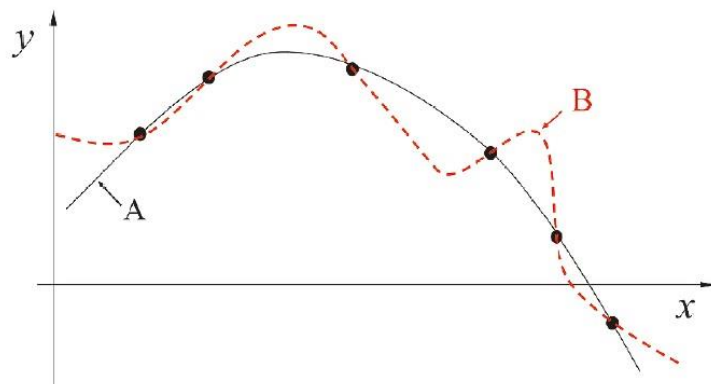
在面临新样本时, 会产生不同的输出

例如: (青绿; 蜷缩; 沉闷)

**应该采用哪一个
模型(假设)?**

归纳偏好 (inductive bias)

机器学习算法在学习过程中对某种类型假设的偏好



A更好?
B更好?

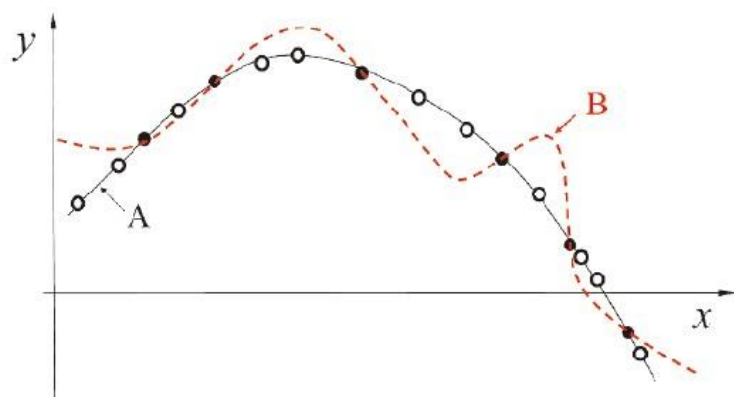
一般原则:
奥卡姆剃刀
(Ocam's razor)

任何一个有效的机器学习算法必有其偏好

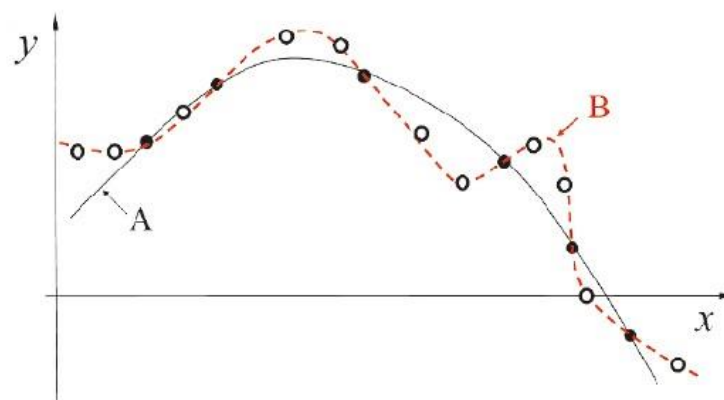
学习算法的归纳偏好是否与问题本身匹配,
大多数时候直接决定了算法能否取得好的性能!

哪个算法更好?

没有免费的午餐!



(a) A 优于 B



(b) B 优于 A

图 1.4 没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)

NFL定理: 一个算法 \mathcal{L}_a 若在某些问题上比另一个算法 \mathcal{L}_b 好, 必存在另一些问题, \mathcal{L}_b 比 \mathcal{L}_a 好。

简单起见, 假设样本空间 \mathcal{X} 和假设空间 \mathcal{H} 离散, 令 $P(h|X, \mathcal{L}_a)$ 代表算法 \mathcal{L}_a 基于训练数据 X 产生假设 h 的概率, f 代表要学的目标函数, \mathcal{L}_a 在训练集之外所有样本上的总误差为

$$E_{ote}(\mathcal{L}_a|X, f) = \sum_h \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) P(h | X, \mathcal{L}_a)$$

考虑二分类问题, 目标函数可以为任何函数 $\mathcal{X} \mapsto \{0, 1\}$, 函数空间为 $\{0, 1\}^{|\mathcal{X}|}$, 对所有可能的 f 按均匀分布对误差求和, 有

$$\sum_f E_{ote}(\mathcal{L}_a|X, f) = \sum_f \sum_h \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) P(h | X, \mathcal{L}_a)$$

NFL定理

$$\begin{aligned}\sum_f E_{ote}(\mathcal{L}_a | X, f) &= \sum_f \sum_h \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) P(h | X, \mathcal{L}_a) \\&= \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathcal{L}_a) \sum_f \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) \\&= \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathcal{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\&= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathcal{L}_a) \\&= 2^{|\mathcal{X}|-1} \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \cdot 1\end{aligned}$$

总误差与学习算法无关!



所有算法一样好!

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！

欧老师的联系方式

读万卷书 行万里路 只为最好的修炼

QQ: 14777591 (宇宙骑士)

Email: ouxinyu@alumni.hust.edu.cn

Tel: 18687840023