机器学习导论 (2017 春季学期)

一、绪论

主讲教师：周志华

机器学习

智能化是信息科学技术发展的主流趋势，机器学习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 [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

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.)

2010 年度

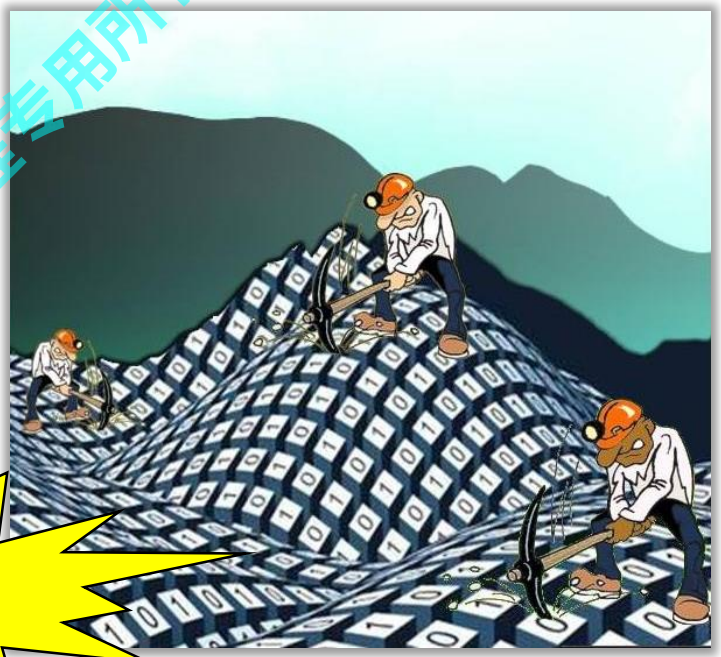
“计算学习理论”奠基人

Judea Pearl
(1936 -)
(UCLA)

2011 年度

“图模型学习方法”先驱

大数据时代



大数据 ≠ 大价值

机器学习

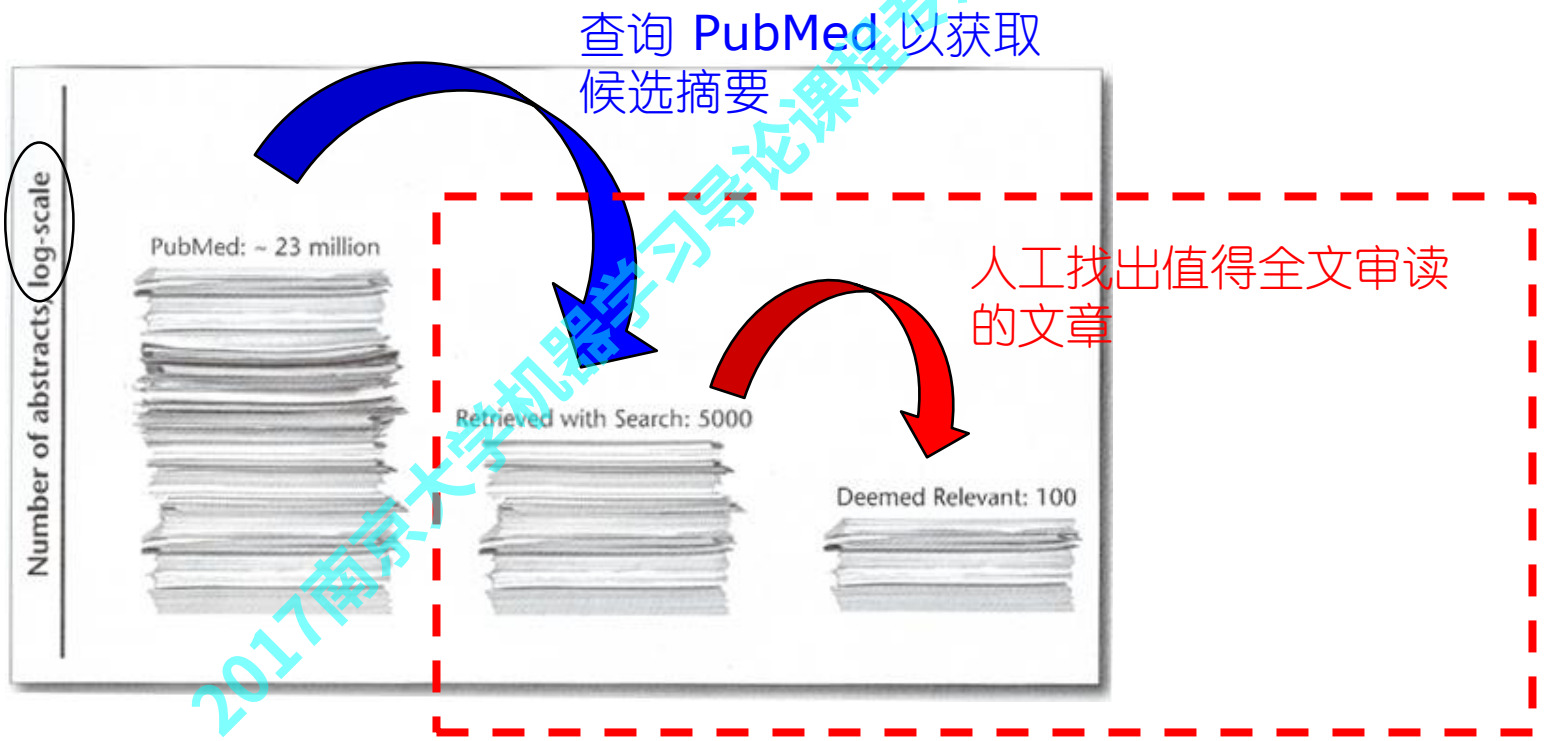
有效的数据分析

机器学习
(Machine Learning)

看个例子 ⇨

“文献筛选”的故事

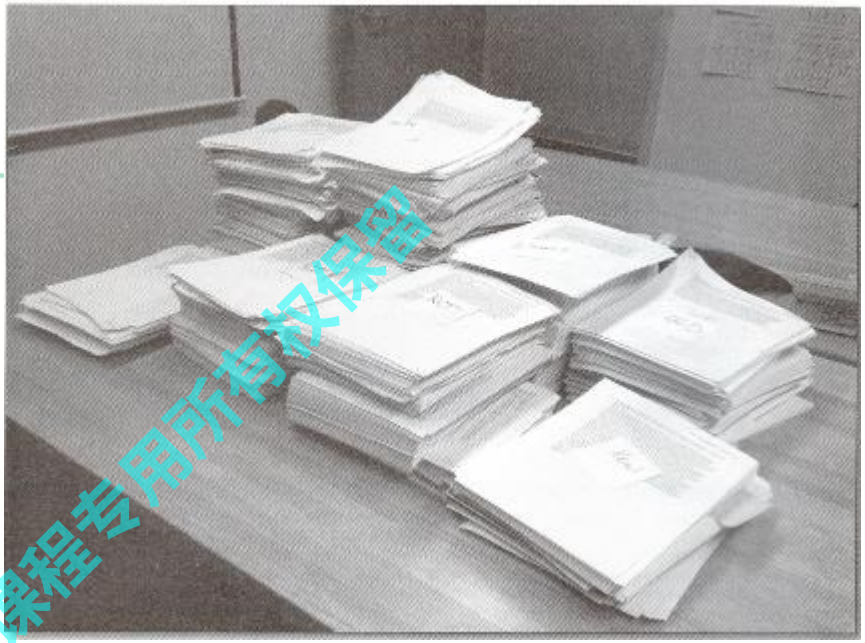
在“循证医学”（evidence-based medicine）中，针对特定的临床问题，先要对相关研究报告进行详尽评估



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

“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



a portion of the 33,000 abstracts

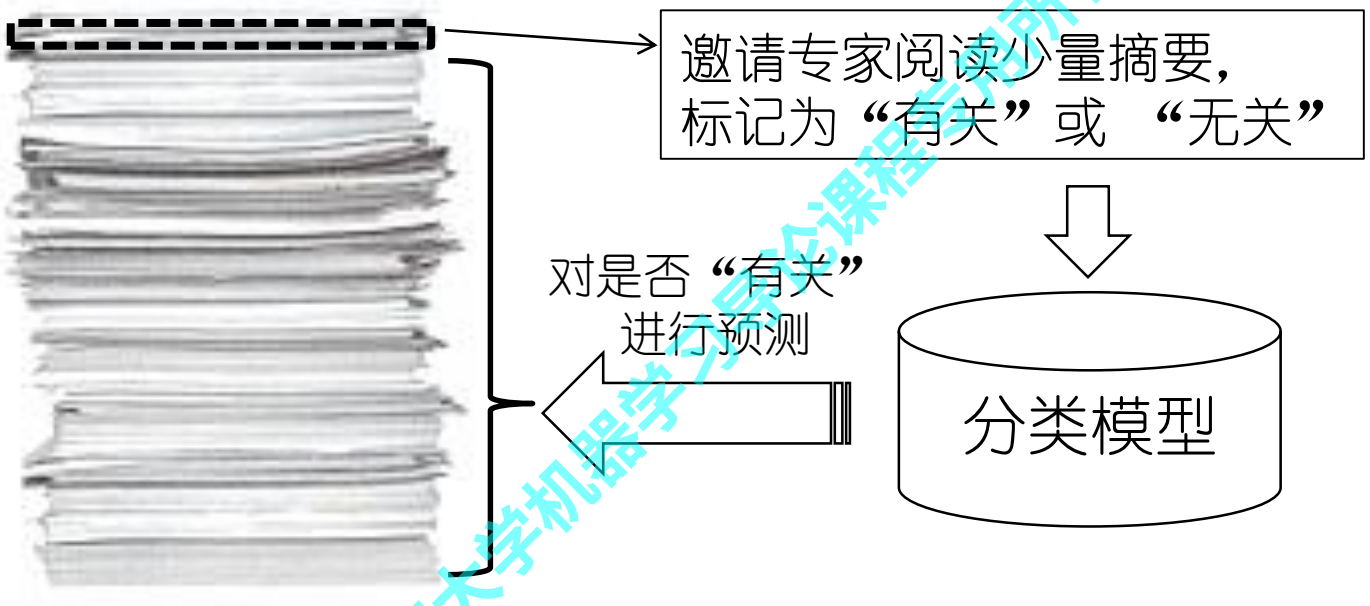
尽管Tufts医学中心的专家效率很高，对每篇摘要只需 30 秒钟，但该工作仍花费了 250 小时

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

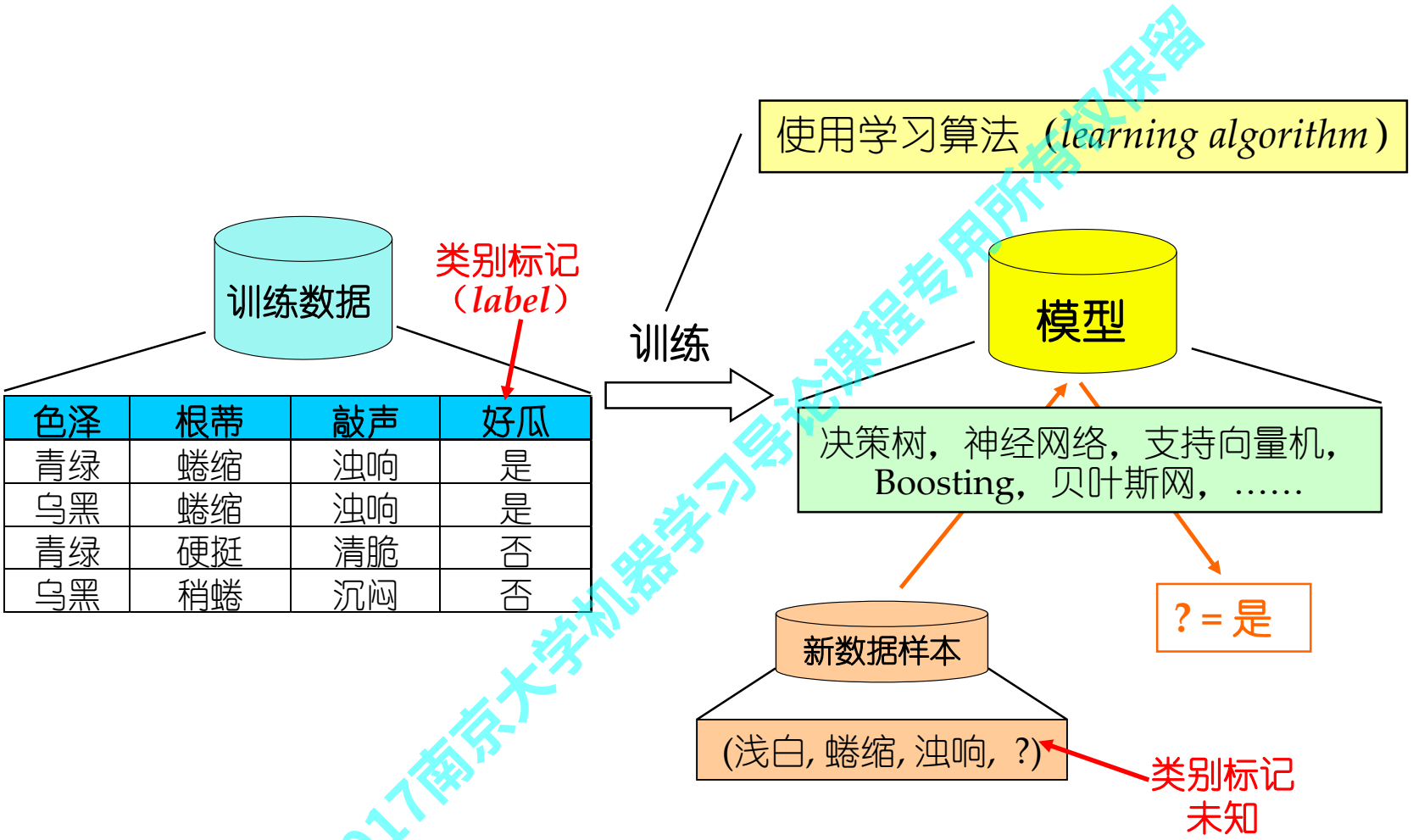


人类专家只需阅读 **50** 篇摘要，系统的自动筛选精度就达到 **93%**

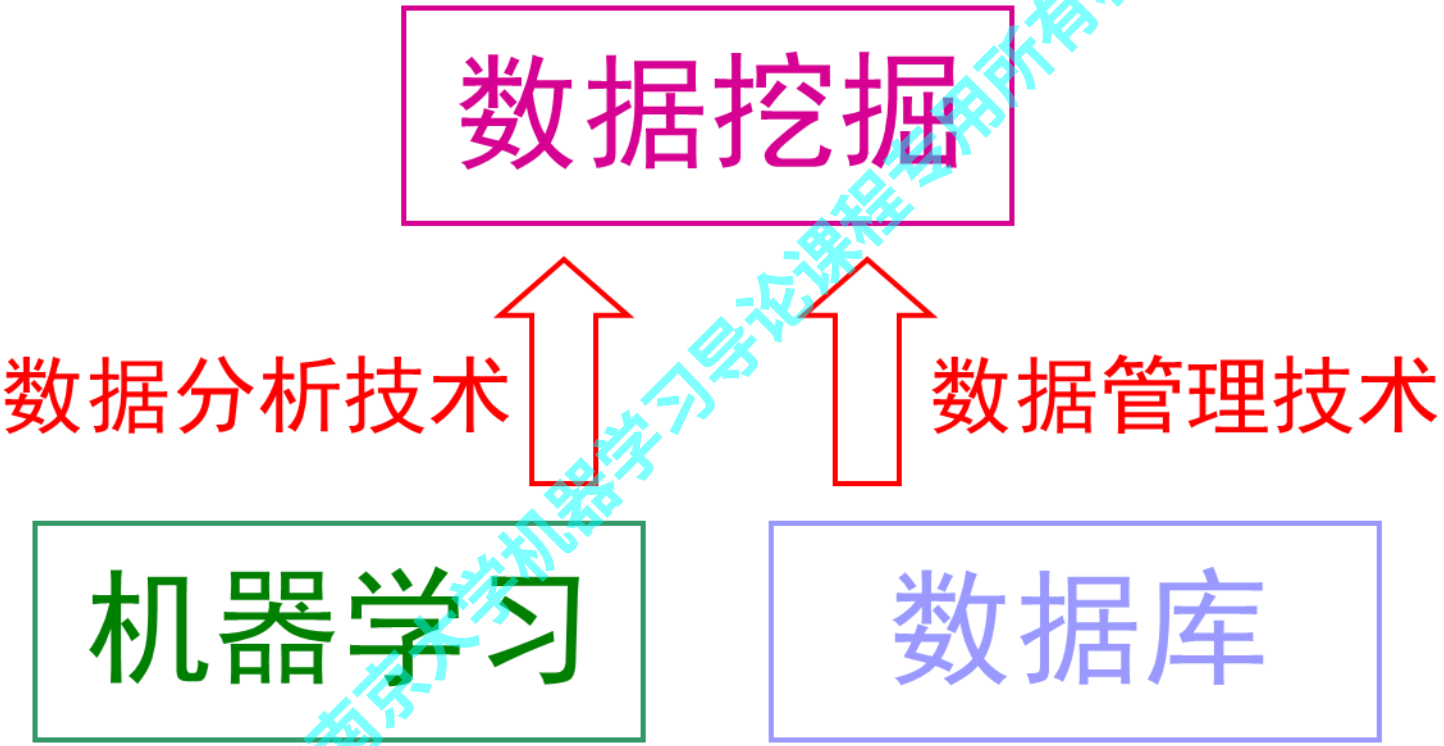
人类专家阅读 **1,000** 篇摘要，则系统的自动筛选敏感度达到 **95%**

(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



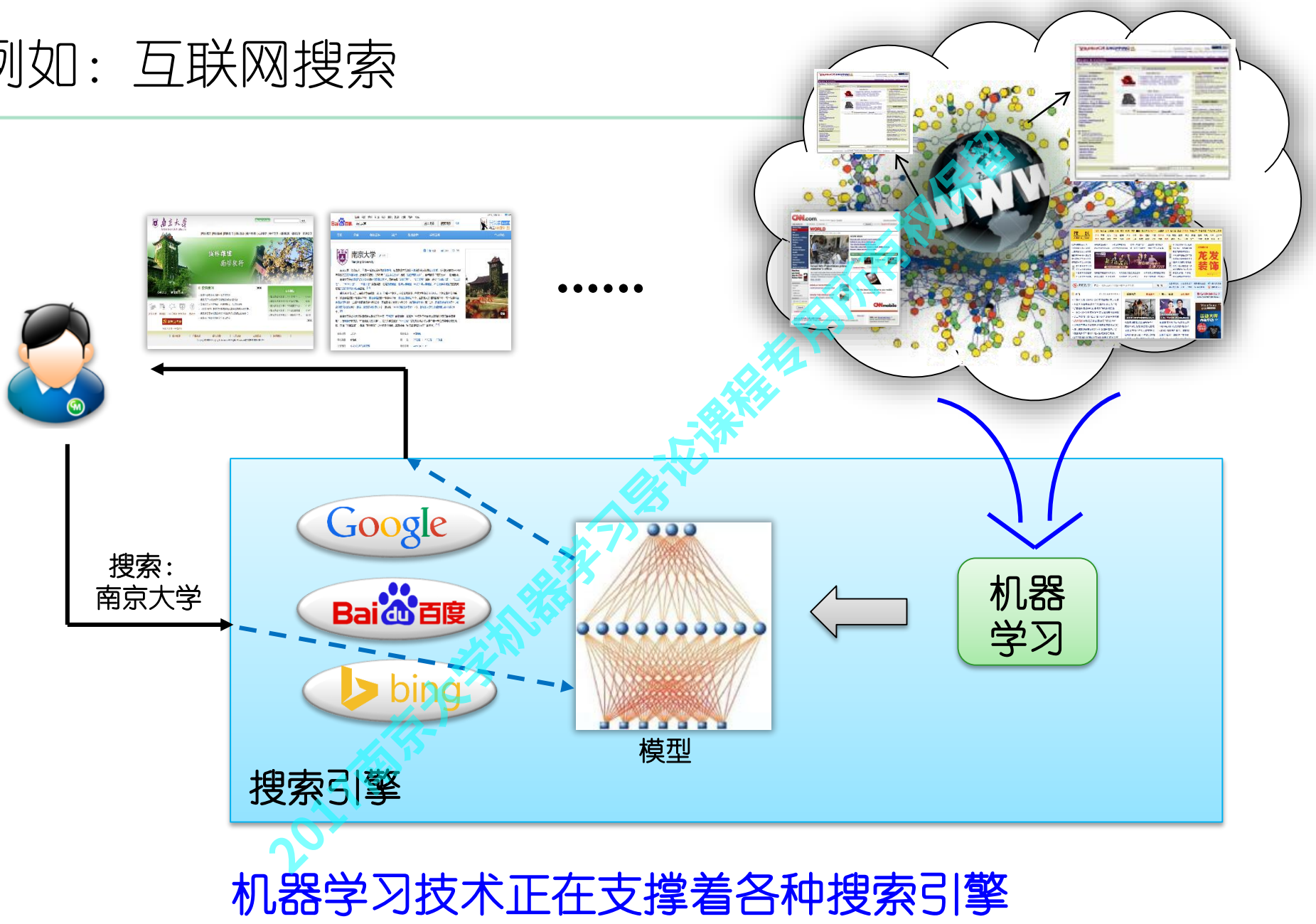
机器学习与数据挖掘



机器学习能做什么？

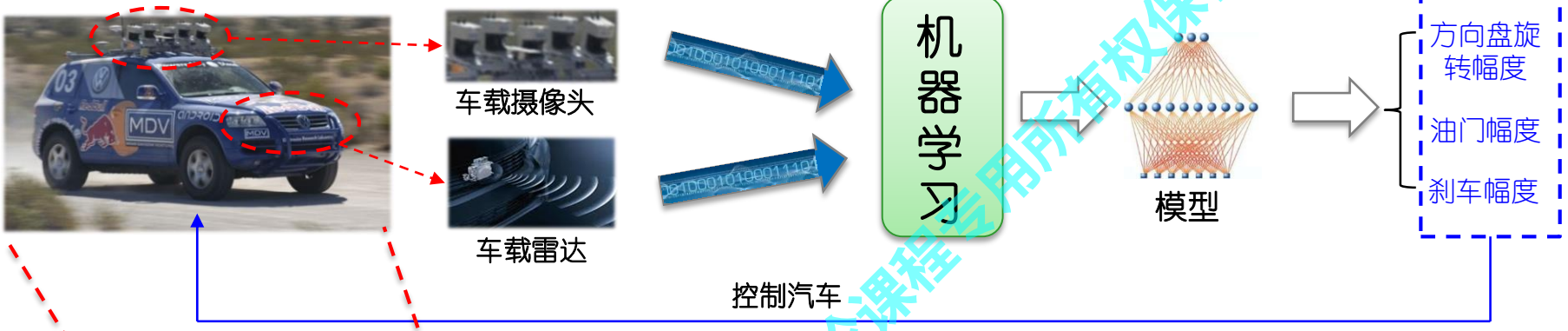
我们可能每天都
在用机器学习

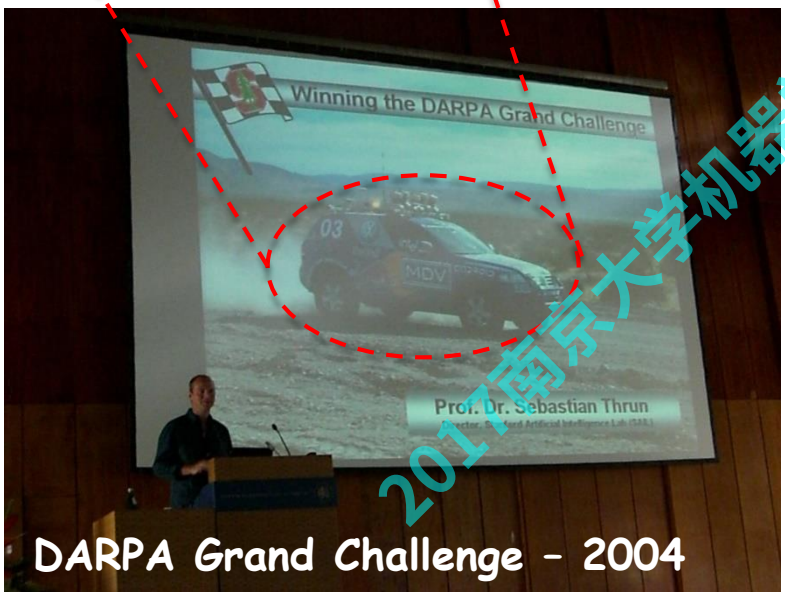
例如：互联网搜索

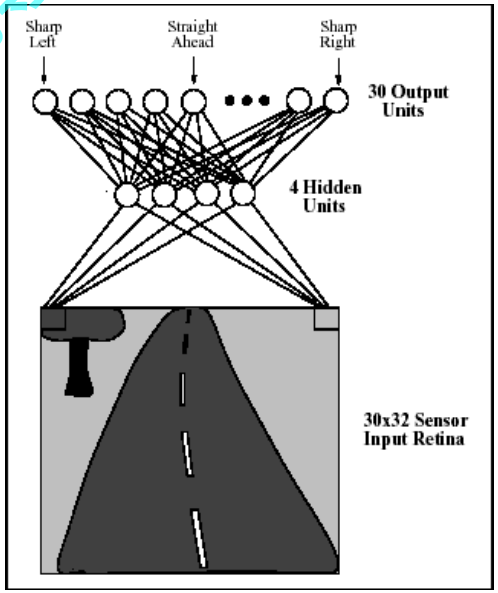


机器学习技术正在支撑着各种搜索引擎

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

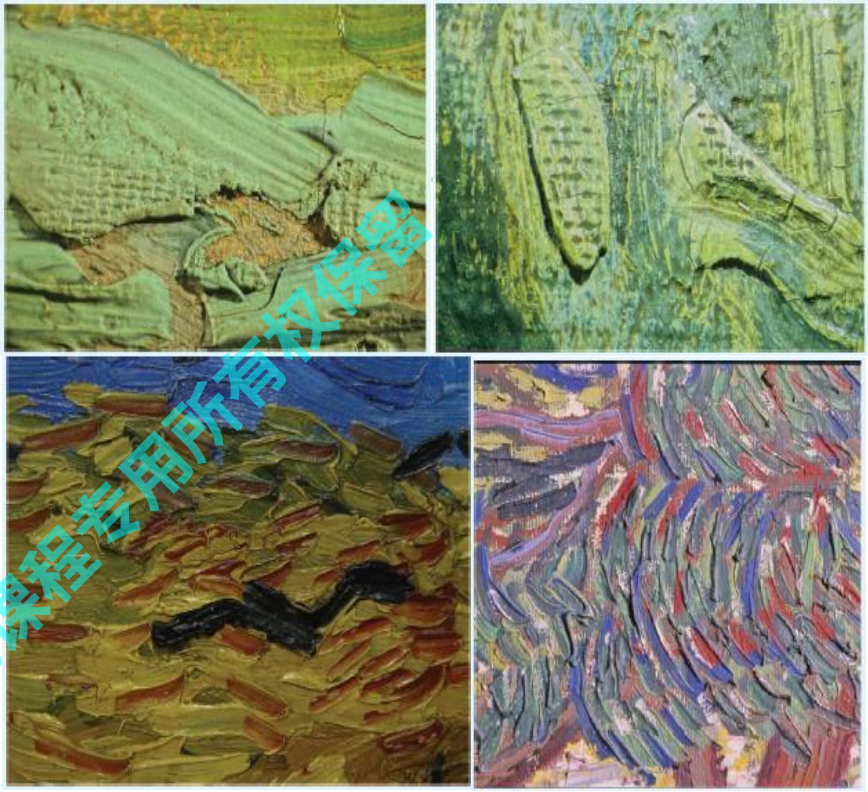
出自 [J. Hughes et al., PNAS 2009]

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）



Portions of van Gogh paintings

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

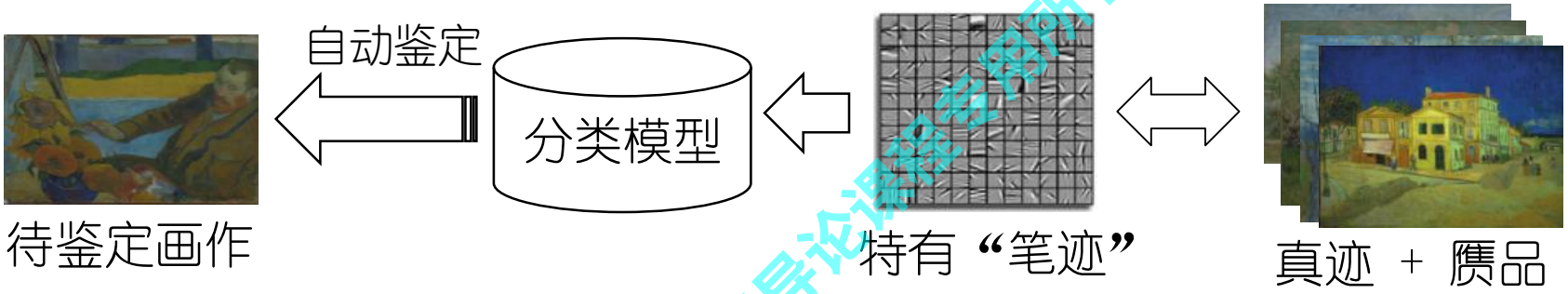
只有少数专家花费很大精力才能完成分析工作！

很难同时掌握不同时期、不同流派多位画家的绘画风格！

[C. Johnson et al., IEEE-SP, 2008]

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？







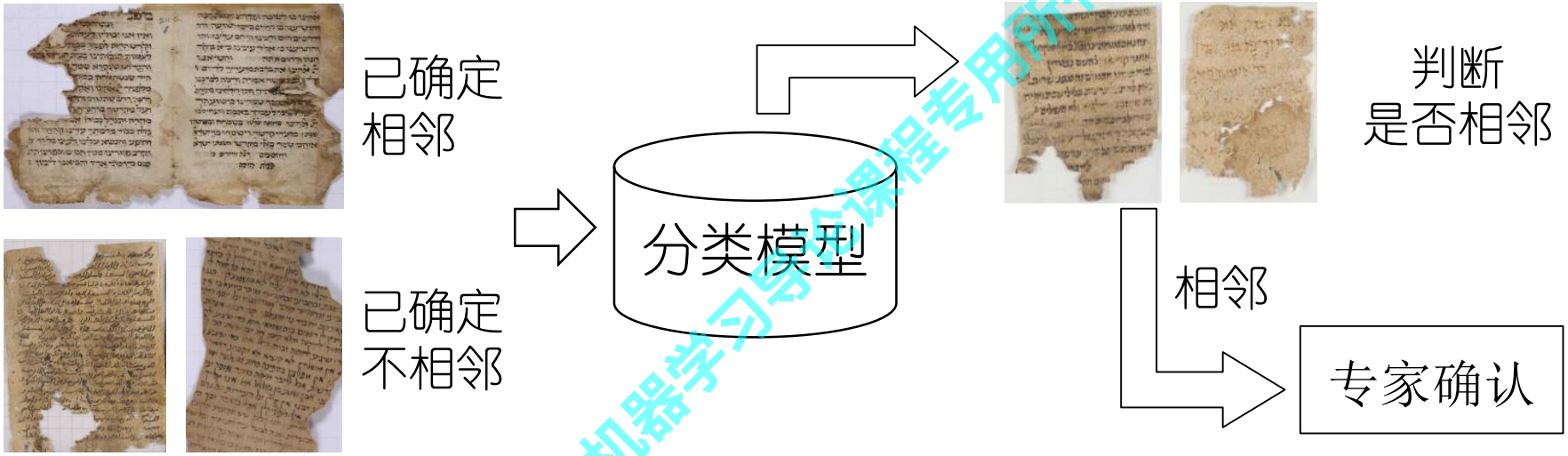
人工完成书页拼接十分困难

- 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

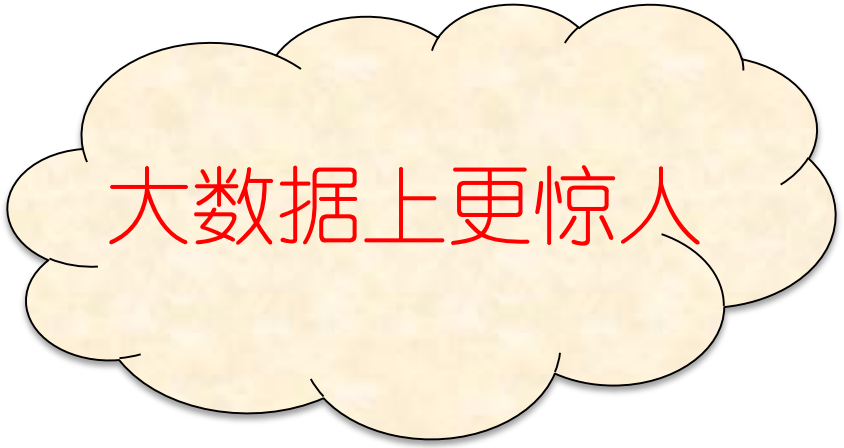


在Cairo Genizah测试数据上，系统的自动判断精度超过 **93%**

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）

机器学习能做什么？



例如：美国总统选举 （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：美国总统选举 （政治）

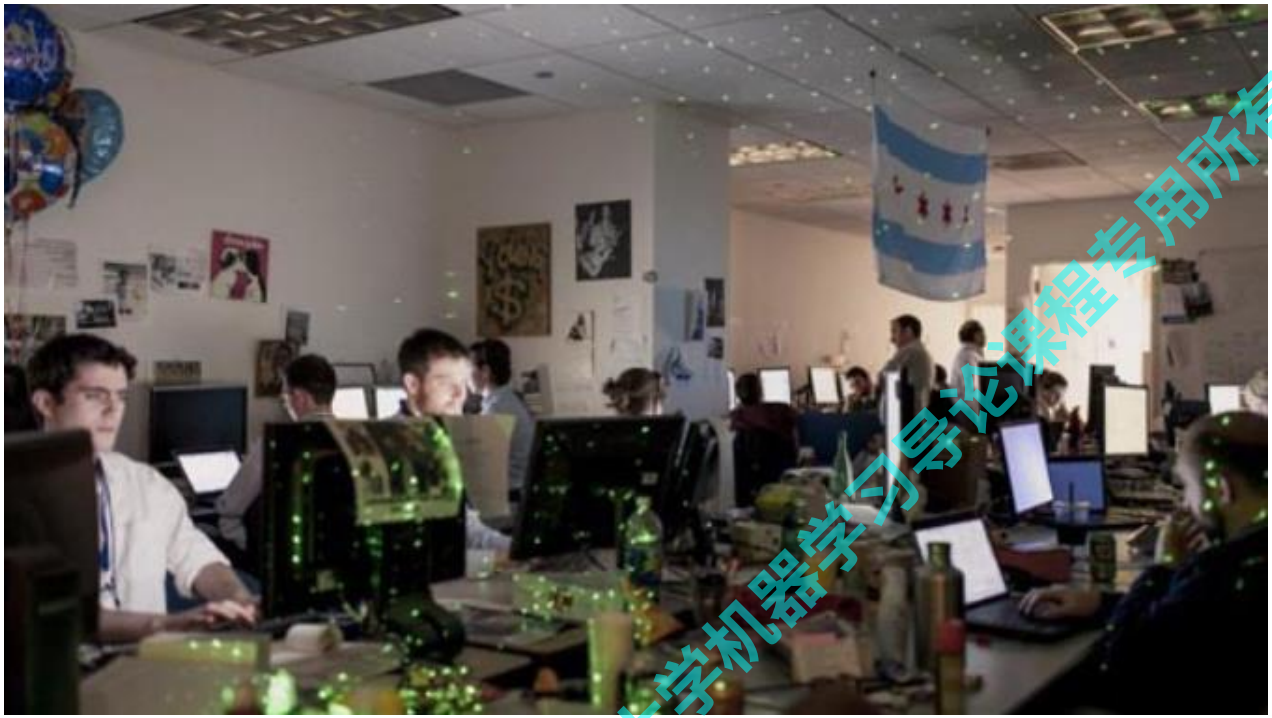
通过机器学习模型：

- 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）对于某地区某有闲有钱特定人群颇具吸引力，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：美国总统选举 （政治）



负责人：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”）

AlphaGo 战胜人类顶尖围棋手

ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

对规则明确的棋类游戏，机器最终一定能超越人类

多种机器学习技术：

- 深度学习
- 强化学习
- 蒙特卡洛树搜索
- …

使用了职业6段-9段人类对弈的160,000局共计29,400,000个盘面，自我对弈的30,000,000个盘面进行学习

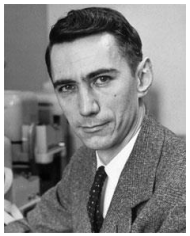
机器学习源自“人工智能”

Artificial Intelligence (AI), 1956 -

1956年夏 美国达特茅斯学院

J. McCarthy
“人工智能之父”
图灵奖(1971)

M. Minsky
图灵奖(1969)



C. Shannon
“信息论之父”

H. A. Simon
图灵奖(1975)
诺贝尔经济学奖(1978)

A. Newell
图灵奖(1975)

……

……

达特茅斯会议标志着人工智能这一学科的诞生

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

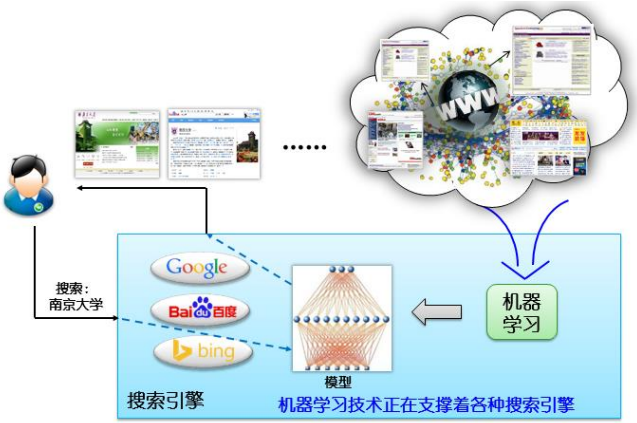
1990s -now: Machine Learning

- 出发点：　“让系统自己学！”
- 主要成就：……

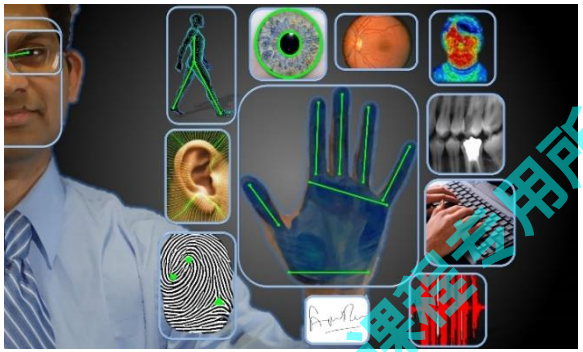
机器学习是作为“突破知识工程瓶颈”之利器而出现的

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

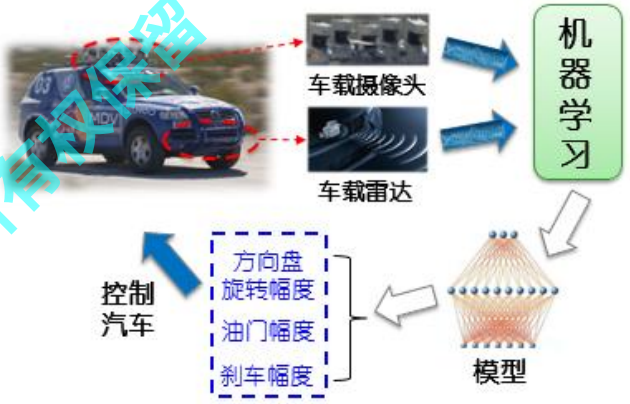
机器学习已经“无处不在”



互联网搜索



生物特征识别



汽车自动驾驶

火星机器人

美国总统选举

军事决策助手 (DARPA)

今天的“机器学习”已经是一个广袤的学科领域

例如，这是第33届国际机器学习大会(ICML 2016)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 Learning Theory
- Computational Social Sciences
- Computer Vision
- Cost-Sensitive Learning
- Digital Humanities
- Economics and Finance
- Ensemble Methods
- Feature Selection and Dimensionality Reduction
- Gaussian Processes
- Graphical Models
- Graphs and Social Networks
- Health Care
- Inductive Logic Programming and Relational Learning
- Information Retrieval
- Information Theory
- Kernel Methods
- Large Scale Learning and Big Data
- Latent Variable Models
- Learning and Game Theory
- Learning and Mechanism Design
- Learning for Games
- Network and Graph Analysis
- Neural Networks and Deep Learning
- Neuroscience
- Other Models and Methods
- Parallel and Distributed Learning
- Planning and Control
- Privacy, Anonymity, and Security
- Probabilistic Programming
- Ranking and Preference Learning
- Recommender Systems
- Reinforcement Learning
- Representation Learning
- Resource Efficient Learning
- Robotics
- Rule and Decision Tree Learning
- Semi-Supervised Learning
- Sparsity and Compressed Sensing
- Spectral Methods
- Speech Recognition
- Statistical Learning Theory
- Statistical Relational Learning
- Structured Prediction
- Supervised Learning

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:

- Encouraging research universities to develop interdisciplinary graduate programs to prepare the next ...
- Fu... $10 million Expeditions in Computing project based at the Uni... 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.

整合三大关键技术

大数据时代，机器学习必不可少

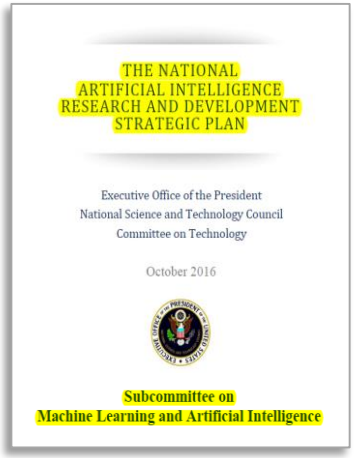
收集、传输、存储大数据的目的，

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

人工智能的核心技术



美国政府2016年5月设立“机器学习与人工智能”国家科学技术分委会，并于10月发布以机器学习为核心的“国家人工智能研究与发展战略规划”

Google

Microsoft

IBM

amazon

各大公司投入巨资，以满足公司对机器学习技术的迫切需求

LOCKHEED MARTIN

美国军工重镇洛克希德·马丁公司将机器学习作为新一代电子战致胜的关键技术进行研究应用

机器学习已经事关国家战略安全和经济社会发展

机器学习很强大，但是……

并非“一切皆可学”

- 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

机器学习有坚实的理论基础

计算学习理论

Computational learning theory

最重要的理论模型：

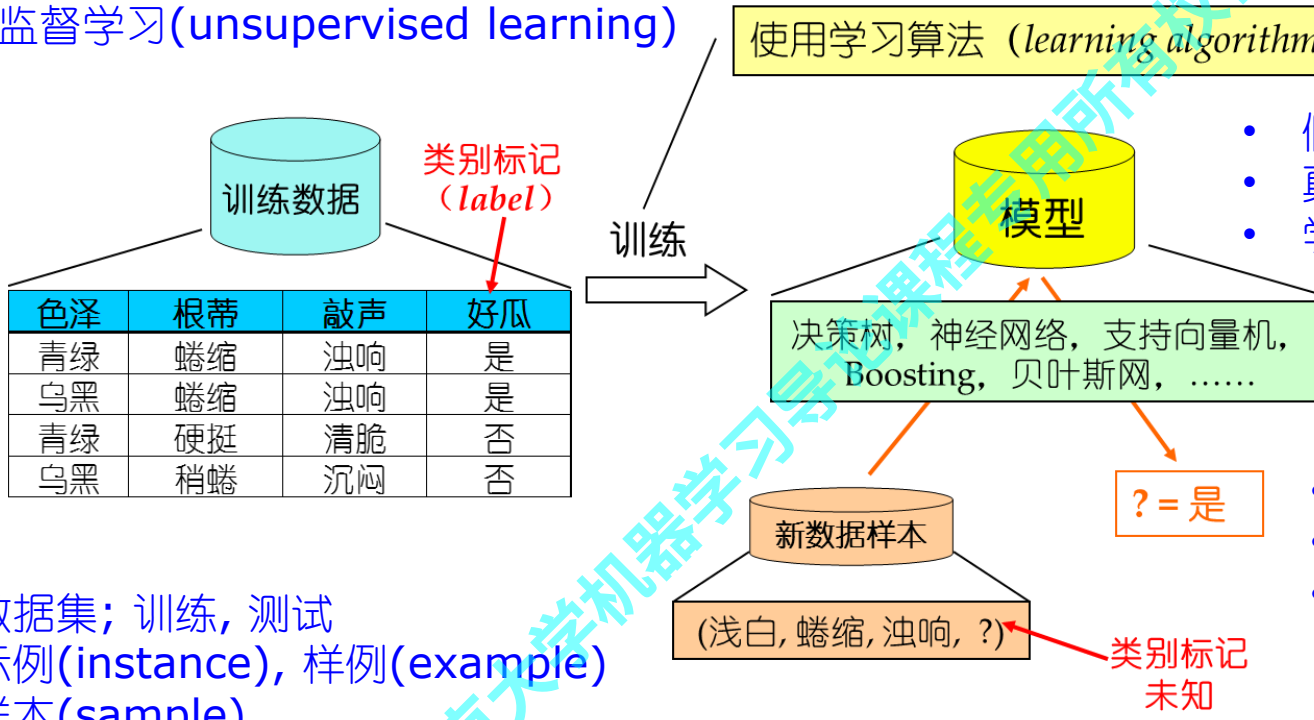
PAC (Probably Approximately Correct, 概率近似正确) learning model [Valiant, 1984]

$$P(|f(\boldsymbol{x}) - y| \leq \epsilon) \geq 1 - \delta$$

Leslie Valiant
(莱斯利•维利昂特)
(1949-)
2010年图灵奖

基本术语

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)



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

- 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)**

假设空间

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

(色泽=?)∧(根蒂=?)∧(敲声=?)↔好瓜

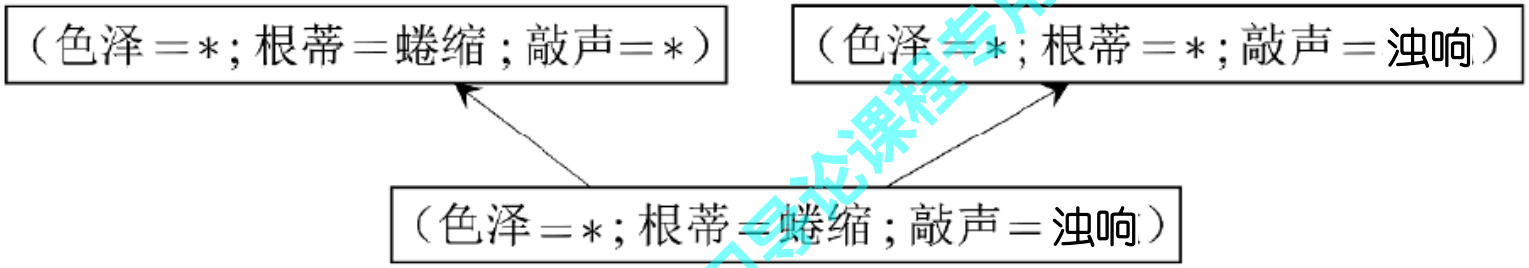
学习过程 → 在所有假设(hypothesis)组成的空间中进行**搜索**的过程

目标：找到与训练集“匹配”(fit)的假设

假设空间的大小：n1 x n2 x n3 + 1

版本空间

版本空间(version space)：与训练集一致的假设集合



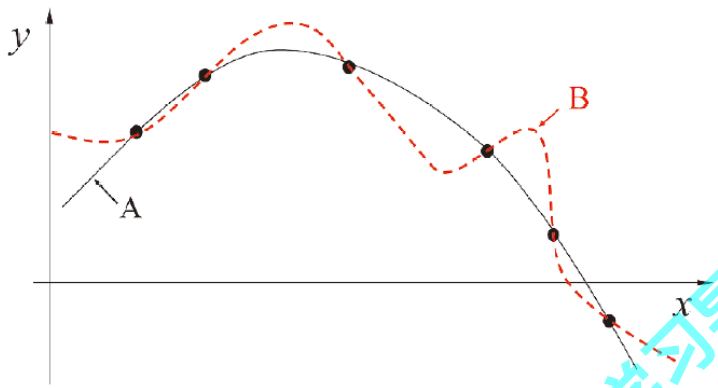
在面临新样本时，会产生不同的输出

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？

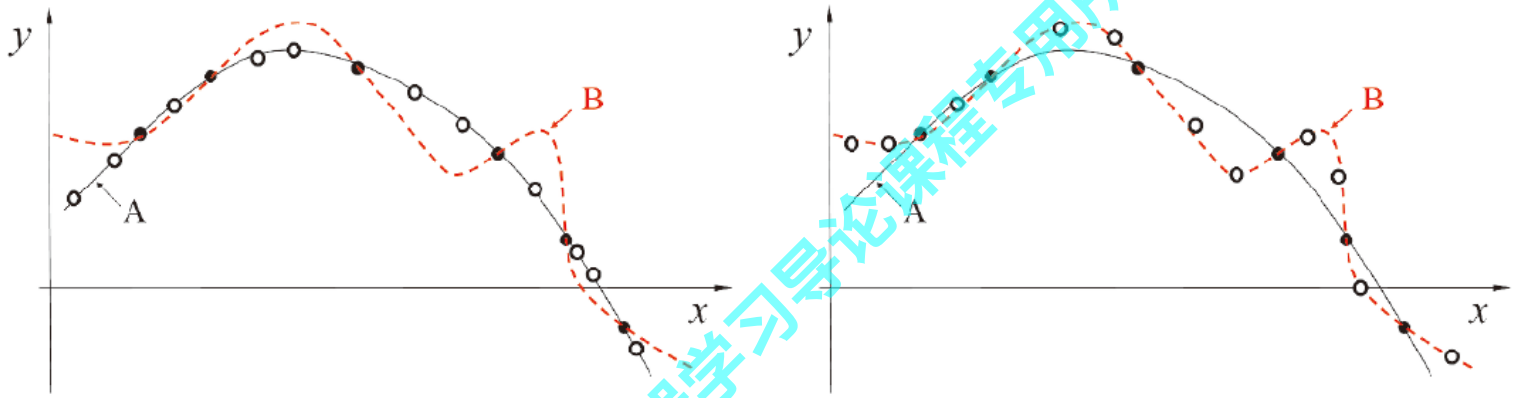
一般原则：
奥卡姆剃刀
(Ocam's razor)

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

NFL定理

简单起见，假设样本空间 $\mathcal{X}$ 和假设空间 $\mathcal{H}$ 离散，令 $P(h|X,\mathfrak{L}_a)$ 代表算法 $\mathfrak{L}_a$ 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，$\mathfrak{L}_a$ 在训练集之外所有样本上的总误差为

$$E_{ote}(\mathfrak{L}_a|X,f)=\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}\left(h\left(\boldsymbol{x}\right)\neq f\left(\boldsymbol{x}\right)\right)P\left(h\mid X,\mathfrak{L}_a\right)$$

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

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

NFL定理

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

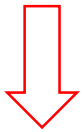
$$= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h\mid X,\mathfrak{L}_a) \sum_f \mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))$$

$$= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h\mid X,\mathfrak{L}_a) \frac{1}{2} 2^{|\mathcal{X}|}$$

$$= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h\mid X,\mathfrak{L}_a)$$

$$= 2^{|\mathcal{X}|-1} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\cdot 1$$

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！

具体问题，具体分析！

现实机器学习应用中 ……

把机器学习的“十八般兵器”都弄熟，
逐个试一遍，是不是就OK了？

NO！

机器学习不是“十八般兵器” 的堆积

在现实任务中，很少能“照搬”兵器取得好结果

按需设计、度身定制

前往第二站……