



# 机器学习导论

## 第一章 絮论

詹德川  
李宇峰



# 机器学习 (Machine Learning)

机器学习是人工智能的核心研究领域，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

2010  
年度



**Judea Pearl**  
**(1936 - )**  
**(UCLA)**

“图模型学习方法”先驱

2011  
年度

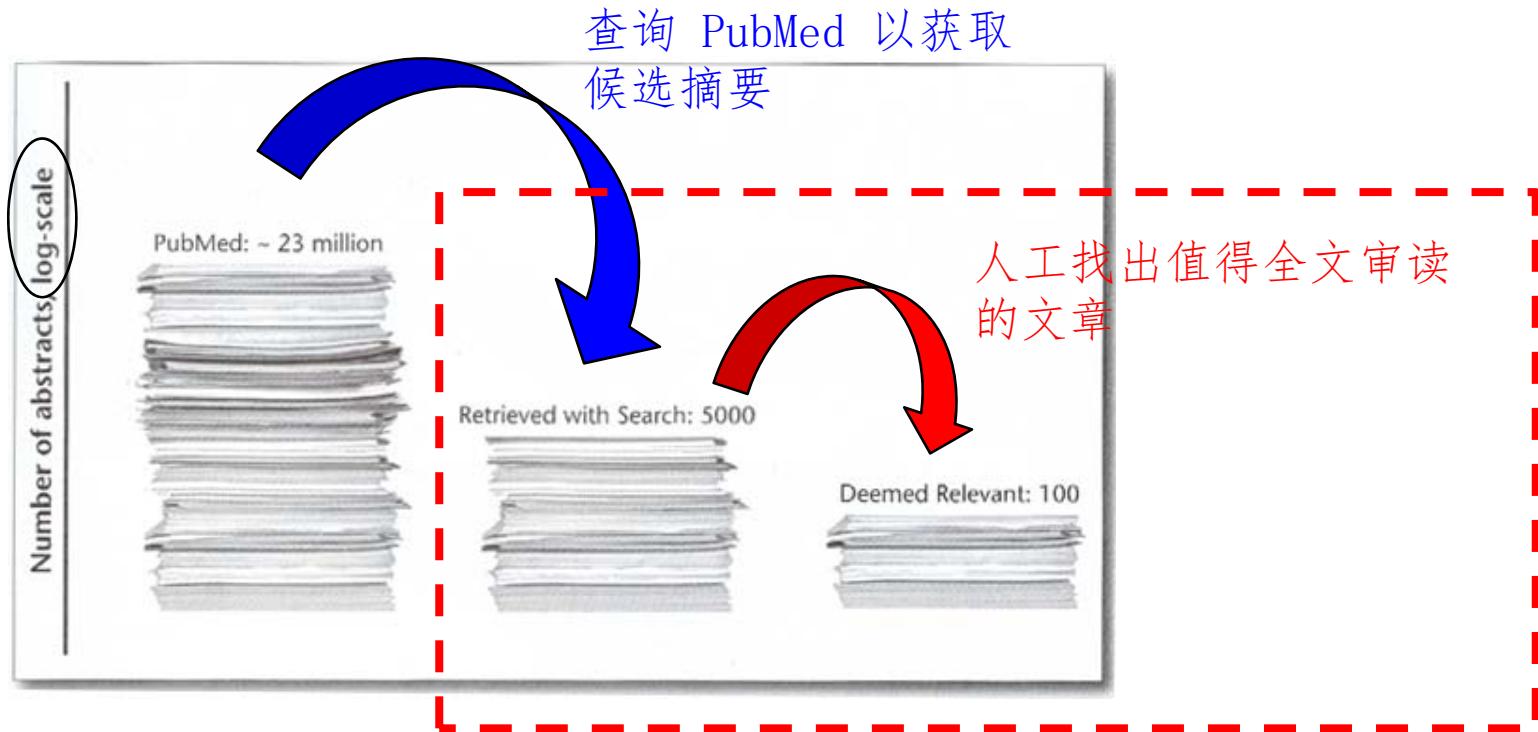
---

# 机器学习究竟是什么？

先看例子 ➔

# “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，  
每篇摘要筛选时间只需 **30** 秒钟，  
但该工作仍花费了 **250** 小时



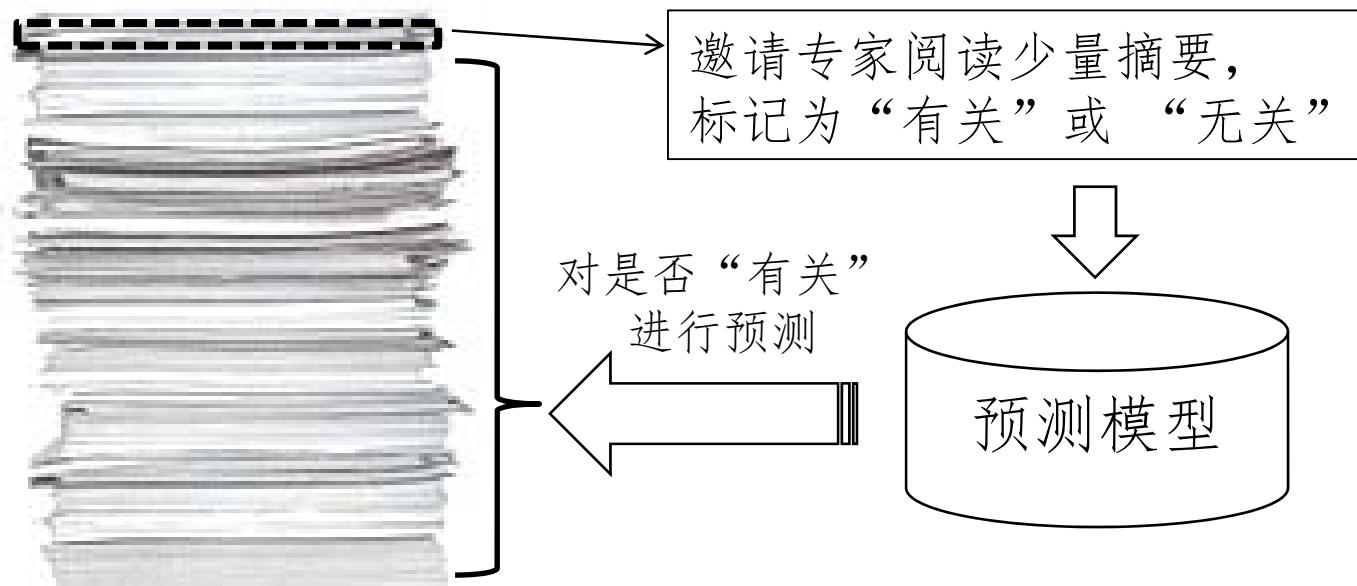
*A portion of the 33,000 abstracts*

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”

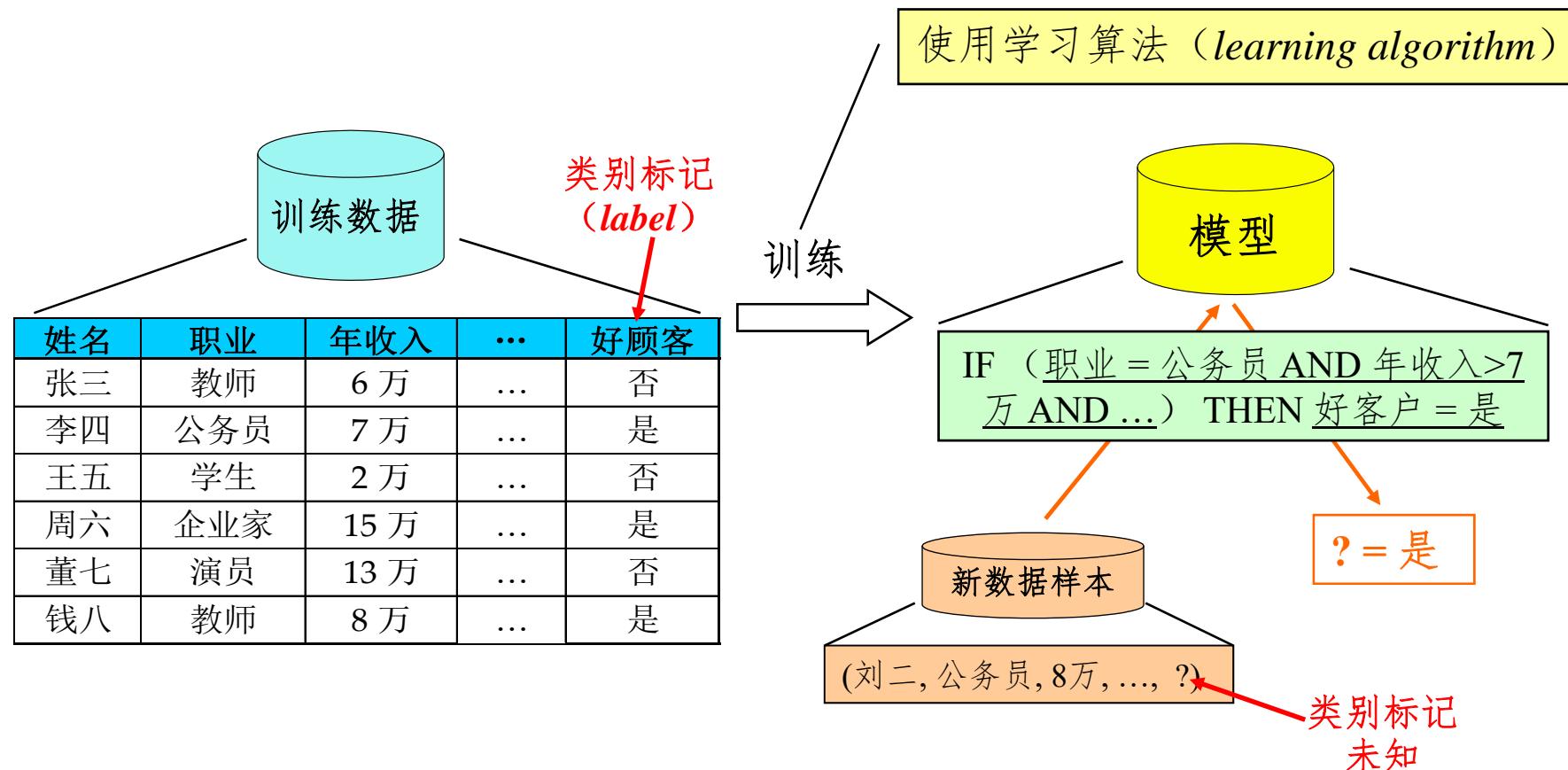
为了降低昂贵的成本，Tufts医学中心引入了机器学习技术



- 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%**

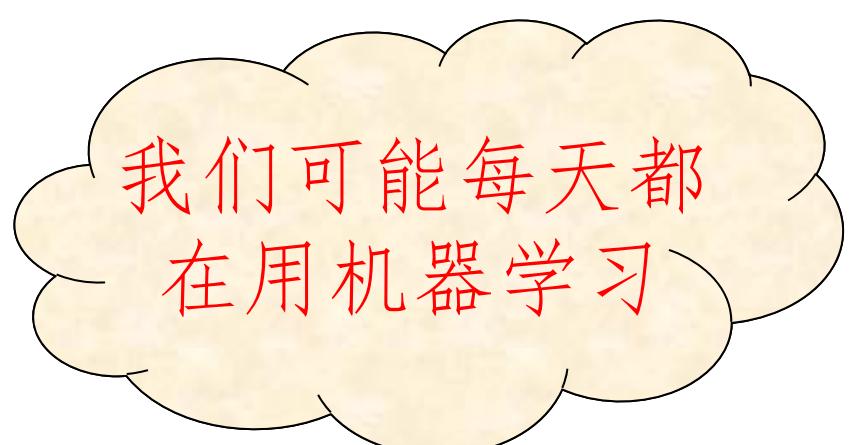
人类专家以前需阅读 33,000 篇摘要才能获得此效果

# 典型的机器学习过程



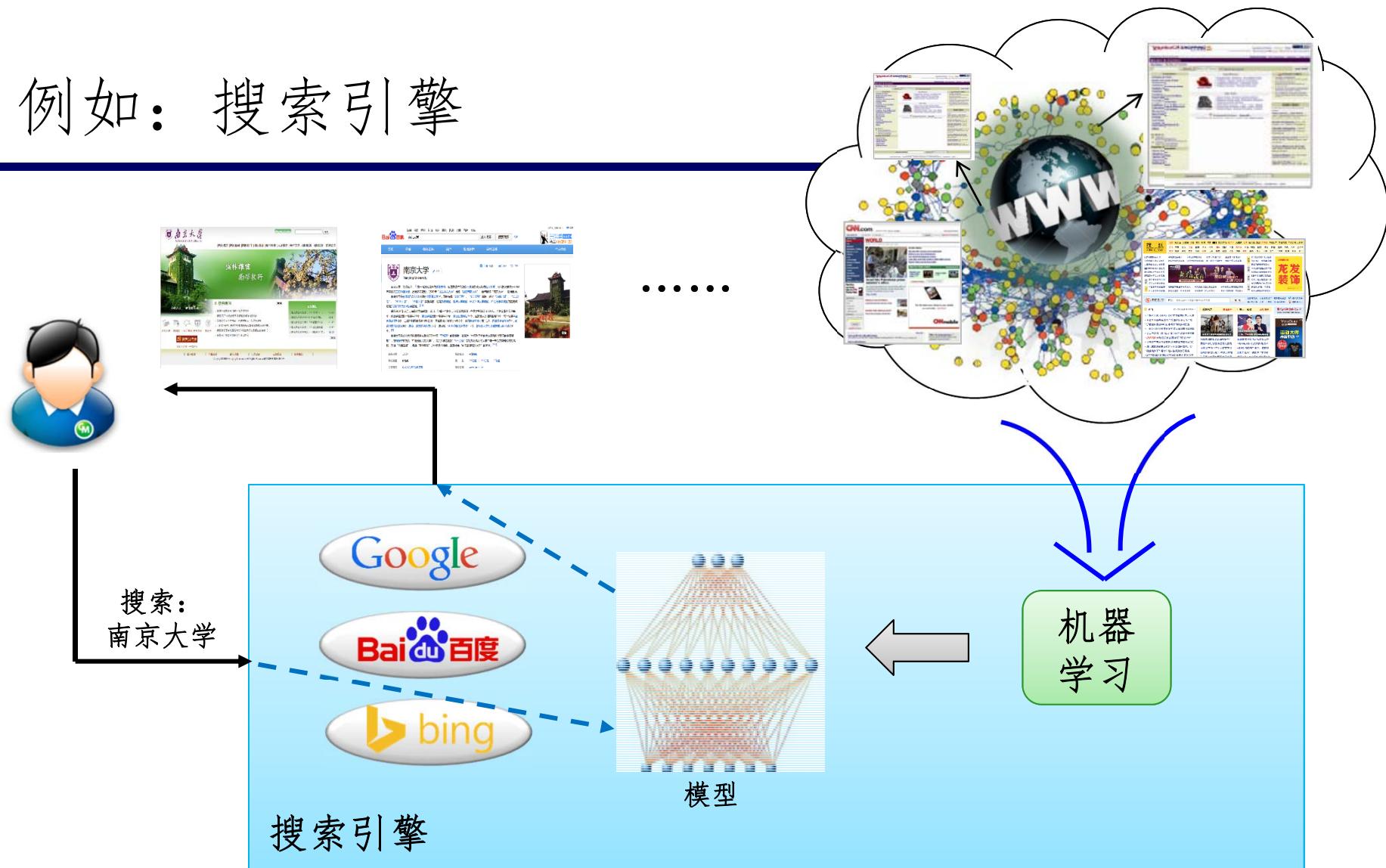
---

# 机器学习能做什么？



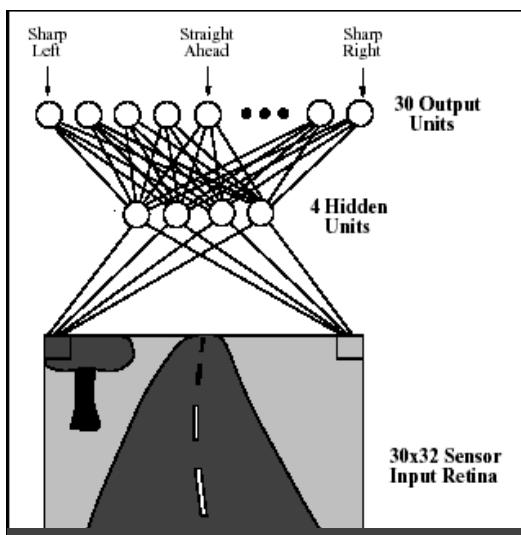
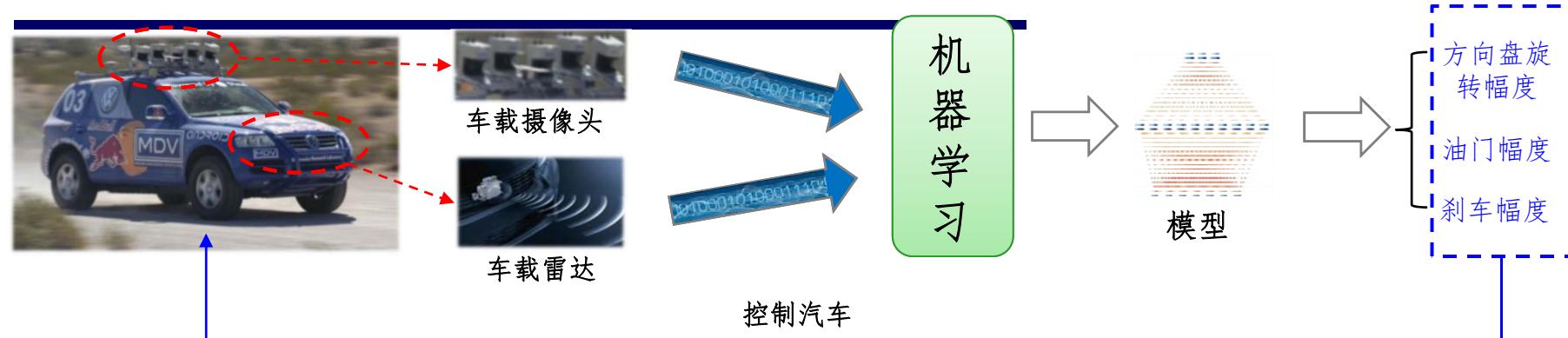
我们可能每天都  
在用机器学习

例如：搜索引擎

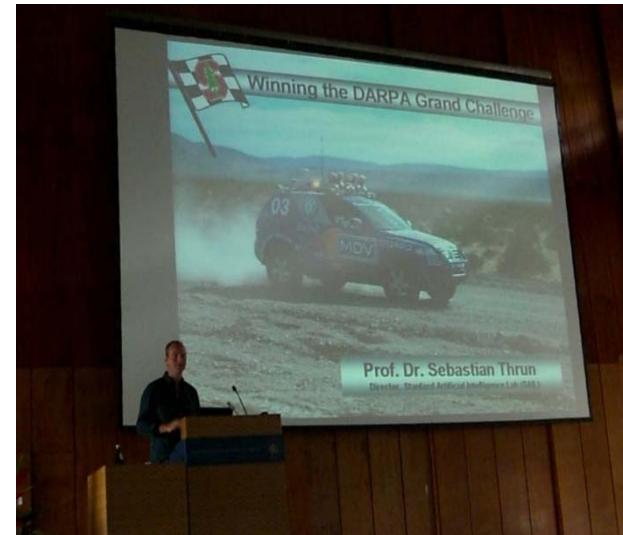


机器学习技术正在支撑着各种搜索引擎

例如：自动驾驶汽车



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



荒野中的无人车竞赛



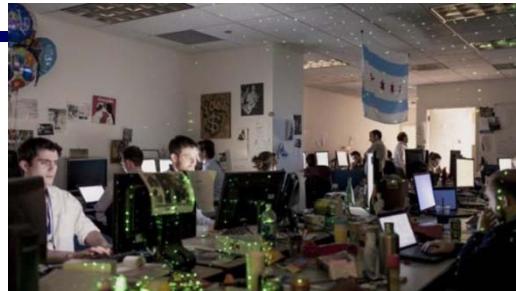
# 例如：帮助奥巴马胜选（政治）

How Obama's data crunchers helped him win

TIME

By Michael Scherer  
November 8, 2012 — Updated 1p

《时代》周刊



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





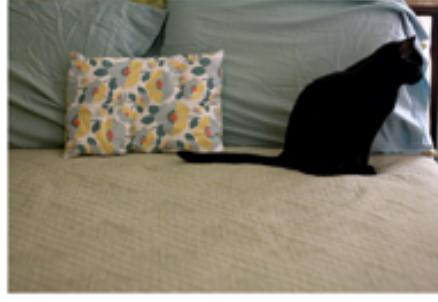


**LAMDA**

Learning And Mining from Data

<http://lamda.nju.edu.cn>

<http://lamda.nju.edu.cn>

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.

# 人工智能

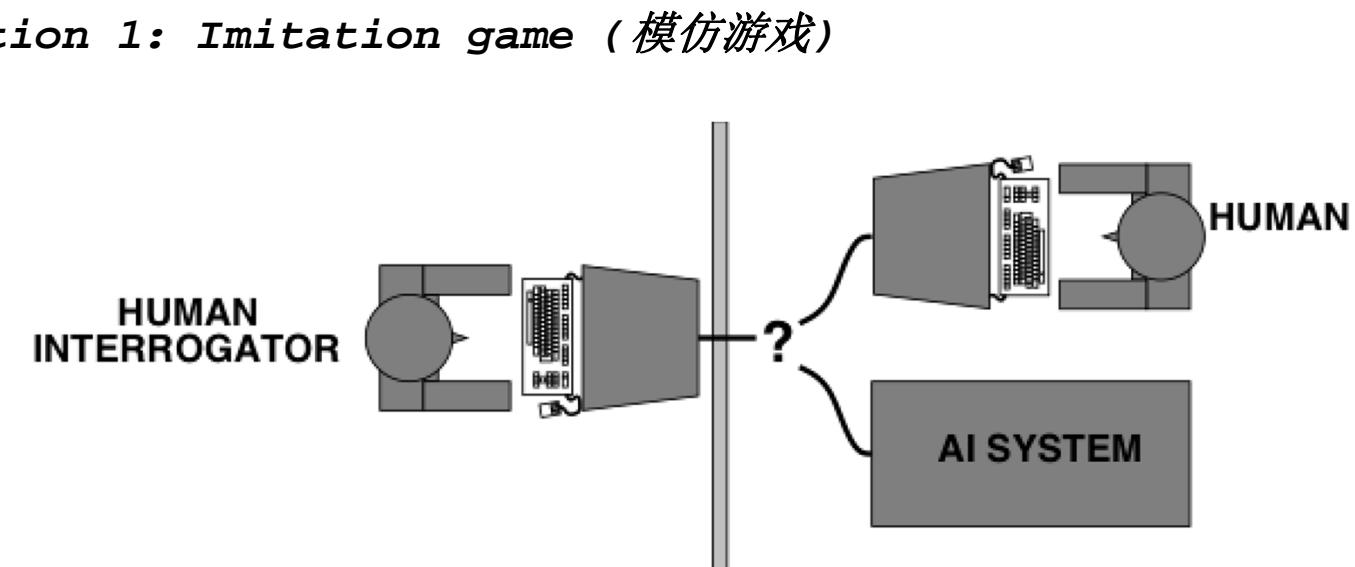


1950

[*Computing machinery and intelligence. Mind* 49: 433-460, 1950.]



Alan Turing  
1912-1954



# 人工智能



AI

1950 1956

*1956 Dartmouth 会议, 命名 “Artificial Intelligence”*



# 人工智能



AI

60-70年  
代

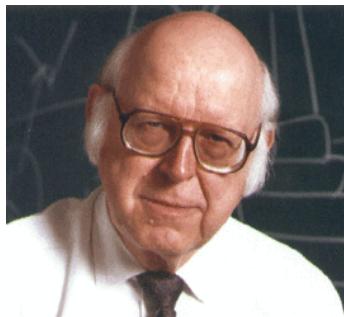
1950 1956

LAMDA

Learning And Mining from Data

<http://lamda.nju.edu.cn>

逻辑学家



*Allen Newell*



*Herbert A. Simon*

PRINCIPIA  
MATHEMATICA

TO \*56

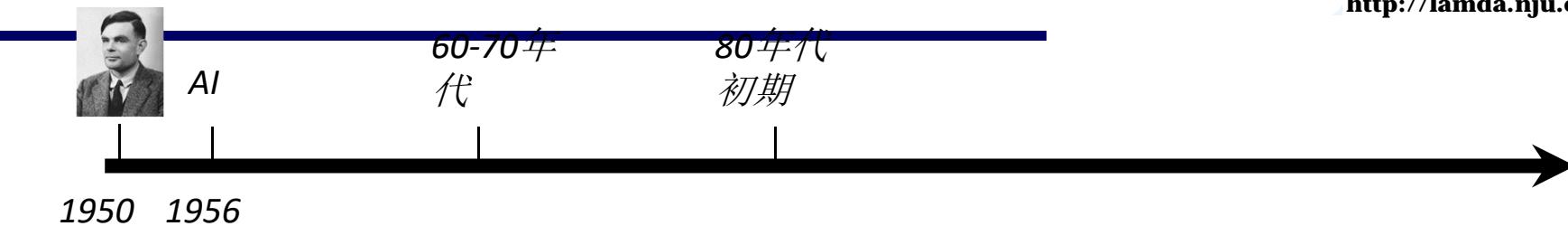
BY  
ALFRED NORTH WHITEHEAD  
AND  
BERTRAND RUSSELL, F.R.S.



CAMBRIDGE  
AT THE UNIVERSITY PRESS

[u.cn](#)

# 人工智能

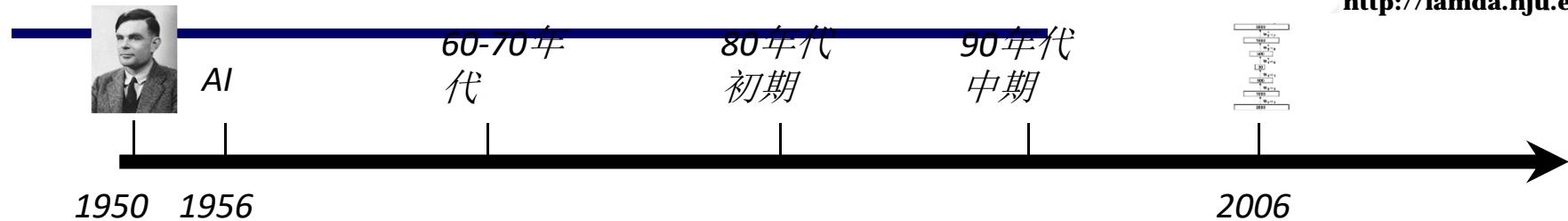


*Edward Albert Feigenbaum*

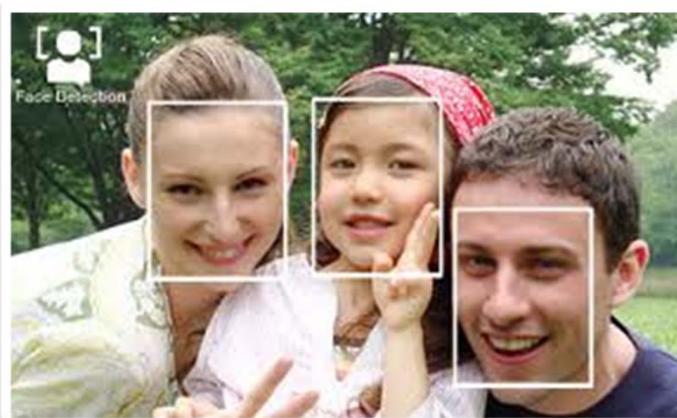
## 专家系统

A screenshot of the Agricultural Expert System (农业专家系统) software interface. The top navigation bar shows the user information "当前用户: 李教授" and the date "2011年6月30日". The main window displays a search interface for plant diseases. On the left, there are tabs for "根部症状" (Root symptoms), "茎部症状" (Stem symptoms), "叶部症状" (Leaf symptoms), and "花果症状" (Flower and fruit symptoms). A search bar at the top right contains the text "已选择的症状". Below it, a table lists "已确诊的结论" (Conclusions confirmed) and "其他可能的结论" (Other possible conclusions). The "已确诊的结论" section includes a row for "番茄猝倒病" (Tomato猝倒病) with a confidence level of 100%, showing symptoms like "茎基部水渍状变褐" (Stem base water-soaked turning brown) and "茎基部茎维只去支撑" (Stem base stem維 only support). The "其他可能的结论" section includes rows for "番茄根腐病" (Tomato root rot) with 80% confidence and "番茄萎蔫" (Tomato wilting) with 75% confidence. The interface also features a sidebar with various icons and a footer with copyright information.

# 人工智能



## 机器学习



This screenshot shows a product page from an e-commerce website. At the top, it displays '店铺28天服务情况' (Shop service status over 28 days) with metrics like return rate and completion time. Below that is a '店铺动态评分' (Dynamic rating) section for book sales, showing a 4.89 rating and a distribution of 5-star reviews (92.47%). Further down, there's a '卖家承诺' (Seller commitment) section stating that the store supports returns and refunds. The main area shows a grid of recommended books, each with its title, price, and a small thumbnail image.

店铺28天服务情况

店铺动态评分: (所属行业: 书籍音像)

商品与描述相符	4.8分	比同行业平均水平 低 0.86%
商家的服务态度	4.7分	比同行业平均水平 低 0.92%
商家发货的速度	4.7分	比同行业平均水平 低 1.09%

卖家承诺：凡使用支付宝服务付款购买本店商品，若存在质量问题或与描述不符，本店支持退换货服务并承担来回邮费！

给我推荐

Head First C# (中文版)	FUNDAMENTALS AND ALGORITHMS OF FACE RECOGNITION	数学之美
¥98.6	¥65.9	¥35

智能车辆导航技术	视觉机器学习 20讲	剑指Offer
¥51.8	¥42	¥41

统计学习方法	程序员的数学① 概率论	深度学习 方法及应用
¥28.5	¥153.2	¥35.9

u.cn

# 人工智能



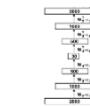
AI

60-70年代

80年代初期

90年代中期

1950 1956



LAMDA

Learning And Mining from Data

<http://lamda.nju.edu.cn>



nature

ALL SYSTEMS GO

2006

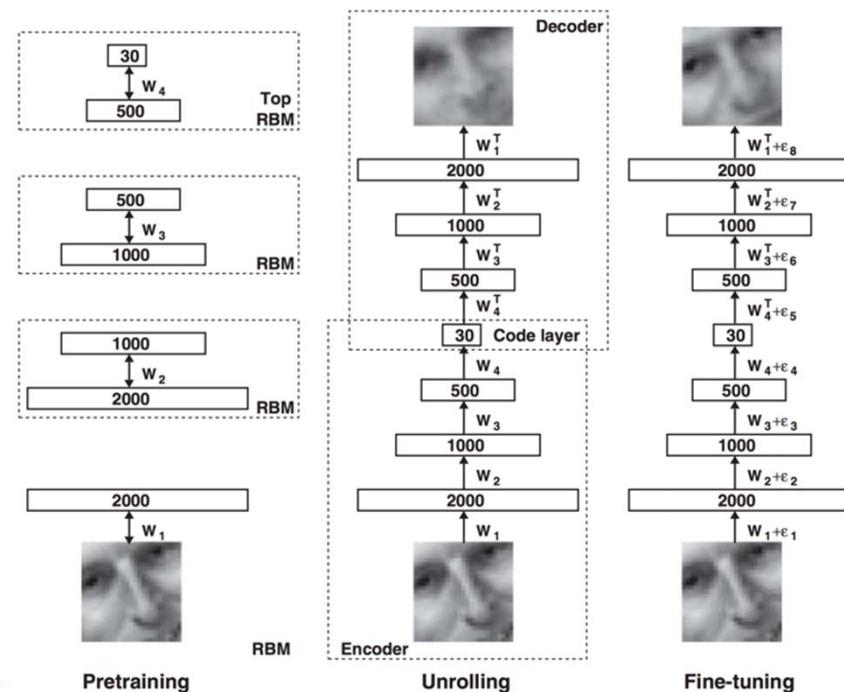
2016

机器学习

深度学习

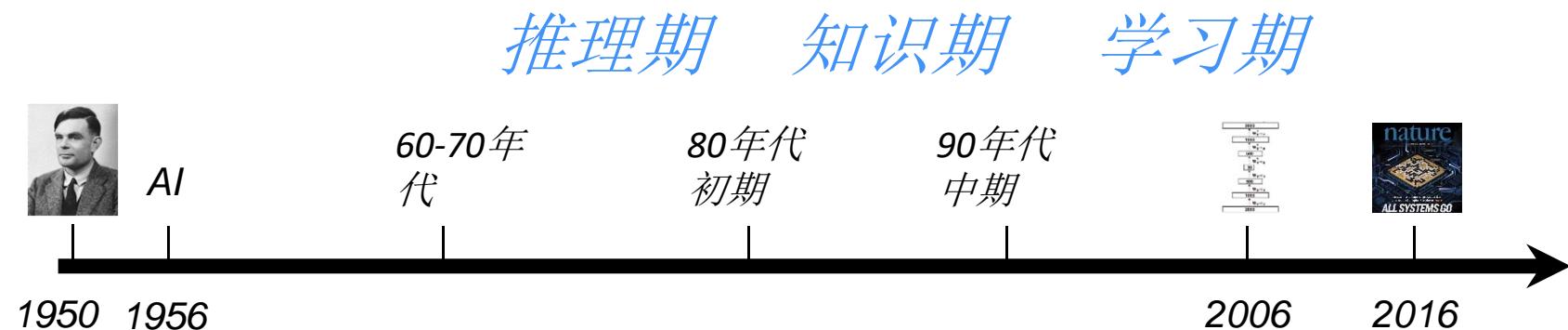


Geoff Hinton



<http://lamda.nju.edu.cn>

# 人工智能



# 2015年

---

《Nature》2015年2月统计学习先驱B. Schölkopf发文评论了基于学习的人工智能

《Nature》2015年5月发表7篇文章的专栏聚焦机器智能  
深度学习、强化学习、概率机器学习、小型自主无人机、机器人、  
演化计算

《Science》2015年7月发表人工智能专辑  
机器学习、自然语言处理、计算理性、数据隐私

互联网巨头纷纷开源机器学习 / 深度学习系统  
FBCUNN、TensorFlow、SystemML、VELES



专用于机器学习等计算  
任务的通用GPU



# 2015年

---

香港科技大学计算机系主任杨强教授（AAAI Fellow、IEEE Fellow、AAAS Fellow）：“现在我国人工智能的水平和国际几乎没有差距”

2015年国际人工智能联合会（IJCAI 2015），中国投稿和录用论文数首次超过美国：欧盟、中国、美国

2015年MLA参会人数超过1400人

微软亚洲研究院2015年11月开源了“分布式机器学习工具箱”  
百度硅谷研究院也开源了其大规模深度学习系统

但是.....

---

70年代，对人工智能发展阶段的认识：

- “解决了神秘的心/身问题，解释了物质构成的系统如何获得心灵的性质。”
- “十年之内，数字计算机将成为国际象棋世界冠军。”
- “二十年内，机器将能完成人能做到的一切工作。”
- “一代之内……创造‘人工智能’的问题将获得实质上的解决。”
- “在三到八年的时间里我们将得到一台具有人类平均智能的机器。”

*--By Simon, Newell and Minsky*

英国政府对无方向的AI研究逐渐停止了资助

美国国家科学委员会在拨款二千万美元后停止资助

1973年Lighthill针对英国AI研究状况的报告批评了AI在实现其“宏伟目标”上的完全失败，并导致了英国AI研究的低潮

DARPA则对CMU的语音理解研究项目深感失望，从而取消了每年三百万美元的资助

到了1974年已经很难再找到对AI项目的资助。。

# 人工智能

---

90年代初，第二次AI之冬

- AI硬件市场需求下跌
- 专家系统维护成本高昂
- 日本五代机失败
- DARPA大幅缩减AI项目资助

低估智能的复杂性

脱离现实问题

但是。 . .

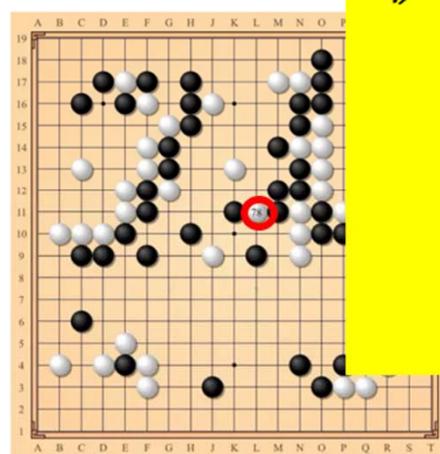
Machine Learning (Turing Award Lecture)

### AlphaGo 并非“解决之道”

AlphaGo is not the solution to AI

Tags: AI, Machine Learning, Reinforcement — jli@ 4:46 pm

Congratulations are in order for the folks at Google Deepmind who have created AlphaGo. However, some of the discussion around this seems like giddy overstatement. Wired says Machines have conquered the last frontier of computer science. I don't think we need any big new breakthroughs to get there.



3月13日李世石九段的  
“神之一手”

John Langford

国际机器学习大会  
ICML'12的程序主席



人类犯错：水平从九段降到八段  
机器犯错：水平从九段降到业余

离“超越人类棋手”还远

“鲁棒性”是关键！

Mistake was on move 9, but #AlphaGo only came to that realisation on around move 87



350

163

AlphaGo以为自己做得很好，但第87手迷惑了。人们有麻烦了

错误出现在第79手犯了错误，但AlphaGo在第87手才发现

# 国际上对AI发展的探讨



**AAAI “主席报告”  
("Presidential Address")  
2016.02.14**



## STEPS TOWARD ROBUST ARTIFICIAL INTELLIGENCE

走向鲁棒的人工智能

Tom Dietterich  
President, Association for the Advancement of Artificial  
Intelligence

**Tom Dietterich**

**AAAI/AAAS/ACM Fellow**

**AAAI 现任主席**

**国际机器学习学会创始主席 (2001-2008)**

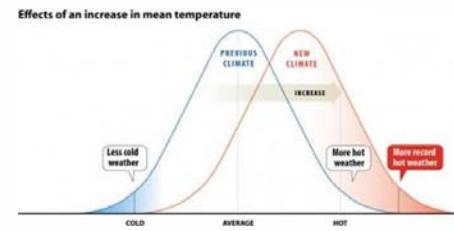
# 国际上对AI发展的探讨

T. Dietterich强调：随着人工智能技术的发展，越来越多地面临“高风险应用”

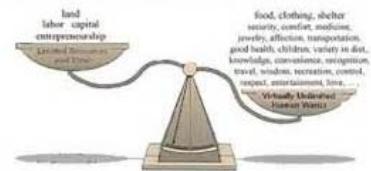
因此，必须要有“鲁棒的AI”

- 对人类用户错误鲁棒
- 对网络攻击鲁棒
- 对错误目标鲁棒
- 对不正确模型鲁棒
- 对未建模现象鲁棒





Interesting  
special cases      Common  
cases



难以获得充足样本

难以适应环境变化

*What's future?*



难以了解模型能力



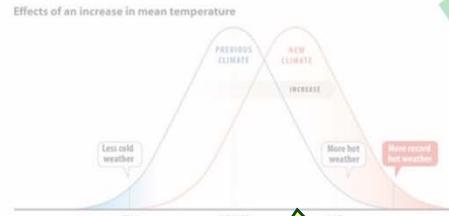
V.S.



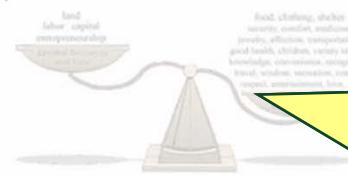
难以获得专家级结果



难以避免数据泄露



Interesting  
special cases



Common  
cases

## 学件将催生一个新兴的产业 (Learnware Industry)

样本不充足



V.S.



结果差异大

有差

可用

模型难理解



隐私难保护

Learner

BlackDOX

Output

# 机器学习与数据挖掘研究团队

http://lamda.nju.edu.cn/CH.MainPage.ashx

主页 - LAMDA

文件(F) 编辑(E) 查看(V) 收藏夹(A) 工具(T) 帮助(H)

LAMDA Learning And Mining from Data

南京大学 [English]

機器學習與數據挖掘研究所

**主页**

LAMDA隶属于计算机软件新技术国家重点实验室和南京大学计算机科学与技术系。LAMDA位于南京大学仙林校区计算机科学技术楼，总部在910室，负责人是周志华教授。

"LAMDA" 的含义是 "Learning And Mining from Data". LAMDA 的主要研究兴趣包括机器学习、数据挖掘、模式识别、信息检索、演化计算、神经计算，以及相关的其他领域。目前的主要研究内容包括：集成学习、半监督与主动学习、多示例与多标记学习、代价敏感和类别不平衡学习、度量学习、降维与特征选择、结构学习与聚类、演化计算的理论基础、增强可理解性、基于内容的图像检索、Web 搜索与挖掘、人脸识别、计算机辅助医疗诊断、生物信息学等。LAMDA 的成员包括教师、工作人员、学生、访问学者和进修教师。若您希望申请国内访问学者或进修教师，请与南京大学人事处联系；如果您希望申请博士后，请与周志华教授联系；如果您希望在 LAMDA 攻读博士或硕士学位，请在前一年的五月左右阅读这个网页并根据要求提供必要的申请材料，请注意，所有申请者都要经过提前面试（面试程序相同，最终结果取决于您与导师的双向选择）；如果您希望在 LAMDA 做本科毕业论文（限南京大学本校生），请通过 Email 与具体的 LAMDA 全职 教师成员 联系。

**主办会议**

2013-05-15 到 2013-05-17 The 11th International Conference on Multiple Classifier Systems (MCS 2013)



## *Machine Learning & Data Mining*

- *Ensemble Learning*
- *Semi-supervised and active learning*
- *Multi-instance and multi-label learning*
- *Cost-sensitive / class-imbalance learning*
- *Metric Learning, Dimensionality reduction / feature selection*
- *Structure learning and clustering*

- *Information Retrieval*
  - *Image retrieval*
  - *Web search and mining*
  - ... ...
- *Evolutionary Computation*
  - *Theoretical issues*
  - ... ...
- *Bioinformatics*
- *Pattern Recognition*
  - *Face recognition*
  - *Computer-aided medical diagnosis*
  - ... ...
- *Neural Computation*
  - *Improve comprehensibility*
  - ... ...

# 相关项目

---

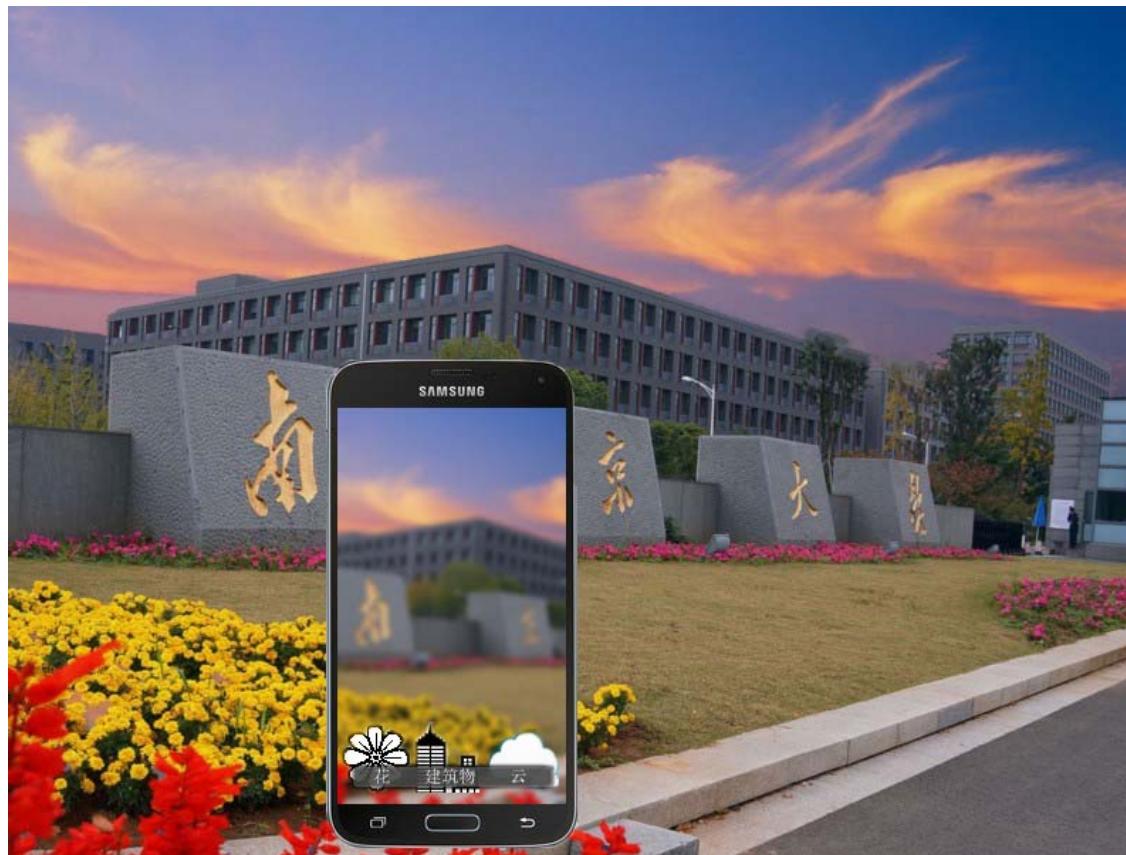
在使用手机进行视频对话时隐去（替换）背景



# 相关项目

---

将手机拍摄照片或存储的照片按类别进行分类



# 相关项目

---

陈庆国，邹晓川. 第一届中国大数据技术创新与创业大赛，冠军，2013.12

## 任务描述：

本题目从百度的行业分类的体系里，选取33个人为定义好的类别，提供大量（约1000万）查询词样本和少量（约100万）的标注信息，以及多角度的辅助信息。要求参赛选手根据需要选取自己觉得有用的信息来设计学习算法，完成对样本数据中未标注样本的分类。



# 相关项目

---

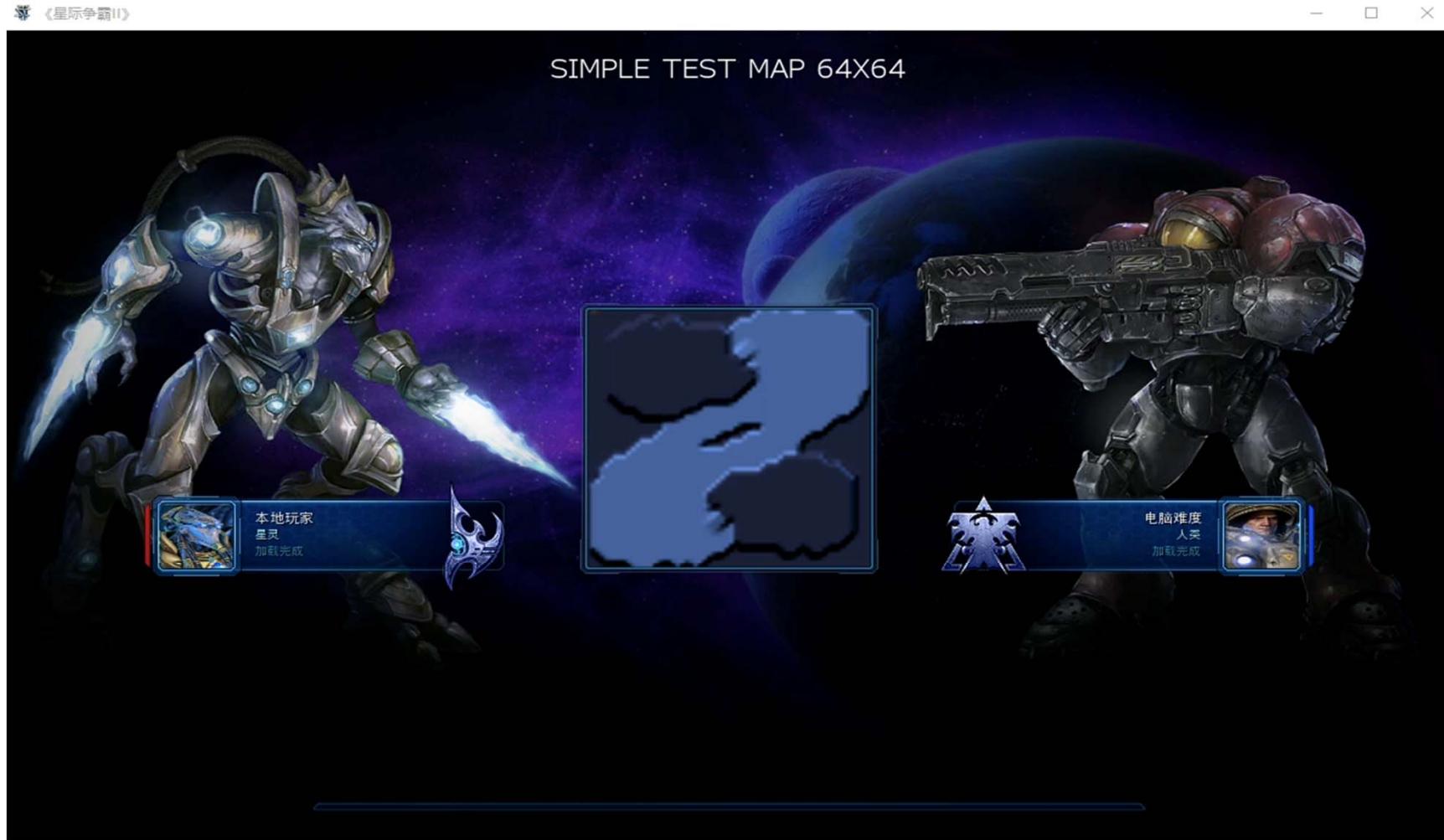


- 多因子选股
- 音乐多标签推荐
- 国家电网故障分析
- 多模态、多语义数据分析
- .....

# 演 示 例 举







To be continue

# 基本术语-数据

编号	特征			标记
	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

训练集 ←

测试集 ← 1	青绿	蜷缩	沉闷	?
---------	----	----	----	---

# 基本术语-任务

---

## □ 预测目标:

- 分类:离散值
  - 二分类:好瓜;坏瓜
  - 多分类:冬瓜;南瓜;西瓜
- 回归:连续值
  - 瓜的成熟度
- 聚类:无标记信息

# 基本术语-任务

---

- 有无标记信息
  - 监督学习：分类、回归
  - 无监督学习：聚类
  - 半监督学习：两者结合

# 假设空间

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

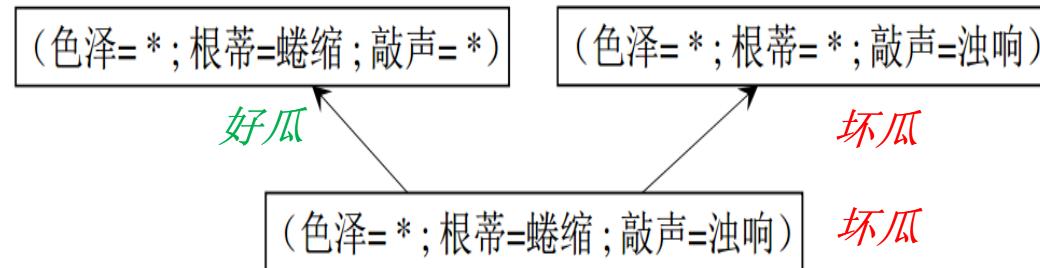
(色泽=? )  $\wedge$  (根蒂=? )  $\wedge$  (敲声=? )  $\leftrightarrow$  好瓜

在模型空间中搜索不违背训练集的假设

假设空间大小:  $3 \times 3 \times 4 + 1 = 37$

# 归纳偏好

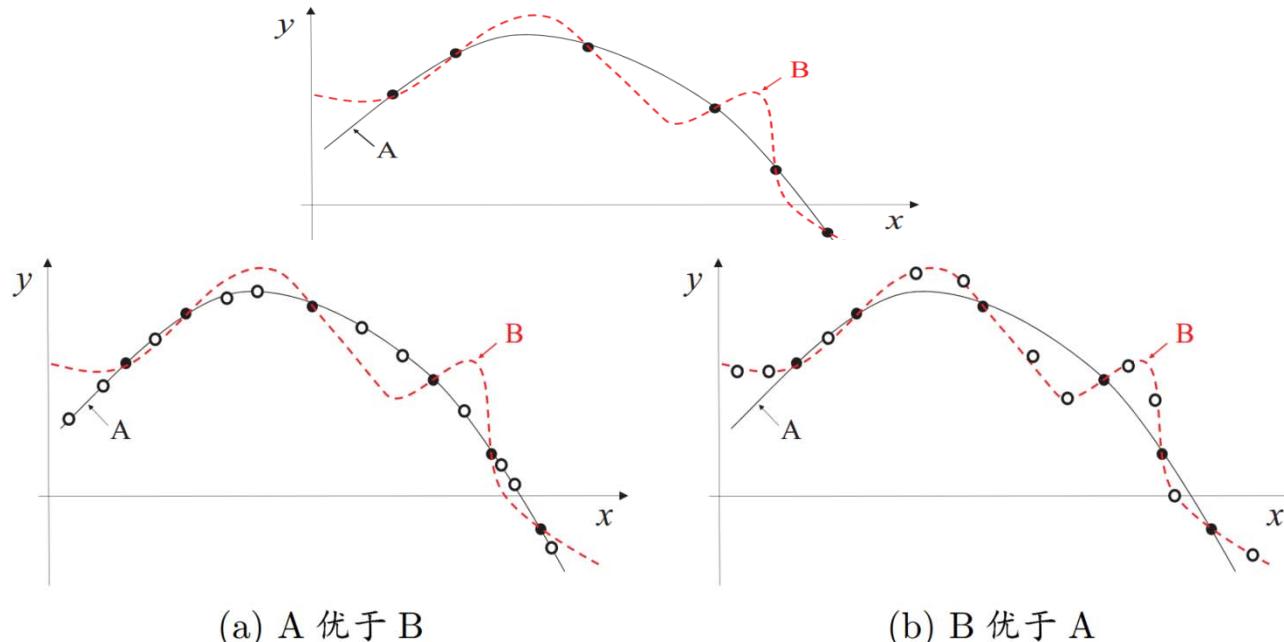
假设空间中有三个与训练集一致的假设，但他们对(色泽=青绿；根蒂=蜷缩；敲声=沉闷)的瓜会预测出不同的结果：



选取哪个假设作为学习模型？

# 归纳偏好

学习过程中对某种类型假设的偏好称作归纳偏好



# 归纳偏好

归纳偏好可看作学习算法自身在一个可能很庞大的假设空间中对假设进行选择的启发式或“价值观”.

“奥卡姆剃刀”是一种常用的、自然科学研究中最基本的原则，即“若有个假设与观察一致，选最简单的那个”.

具体的现实问题中，学习算法本身所做的假设是否成立，也即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.

# NoFreeLunch

---

一个算法 $\xi_a$ 如果在某些问题上比另一个算法 $\xi_b$ 好，必然存在另一些问题， $\xi_b$ 比 $\xi_a$ 好，也即没有免费的午餐定理。

简单起见，假设样本空间  $\mathcal{X}$  和假设空间  $\mathcal{H}$  离散，令  $P(h|X, \mathcal{L}_a)$  代表算法  $\mathcal{L}_a$  基于训练数据  $X$  产生假设  $h$  的概率，在令  $f$  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

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

$\mathbb{I}(\cdot)$  为指示函数，若  $\cdot$  为真取值 1，否则取值 0

# NoFreeLunch

---

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  $f$  按均匀分布对误差求和，有：

$$\begin{aligned}\sum_f E_{ote}(\mathfrak{L}_a | X, f) &= \sum_f \sum_h \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) P(h | X, \mathfrak{L}_a) \\ &= \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathfrak{L}_a) \sum_f \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) \\ &= \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathfrak{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\ &= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathfrak{L}_a) \\ &= 2^{|\mathcal{X}|-1} \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \cdot 1 . \quad \text{总误差与学习算法无关!}\end{aligned}$$

实际问题中，并非所有问题出现的可能性都相同  
脱离具体问题，空谈“什么学习算法更好”毫无意义