机器学习

Machine Learning

主讲人：张敏 清华大学长聘副教授

ML 机器学习

MACHINE LEARNING-MIN ZHANG

Unit.01

机器学习基础

*图片均来自网络或已发表刊物

课程信息

- 要点
  - 关键概念
  - 基础学习理论
  - 经典 / 重要算法
    - 问题定义
    - 基本思路
    - 算法设计与分析
    - 未来方向
  - 分析（问题、特征、结果）
  - **解决实际问题的能力**

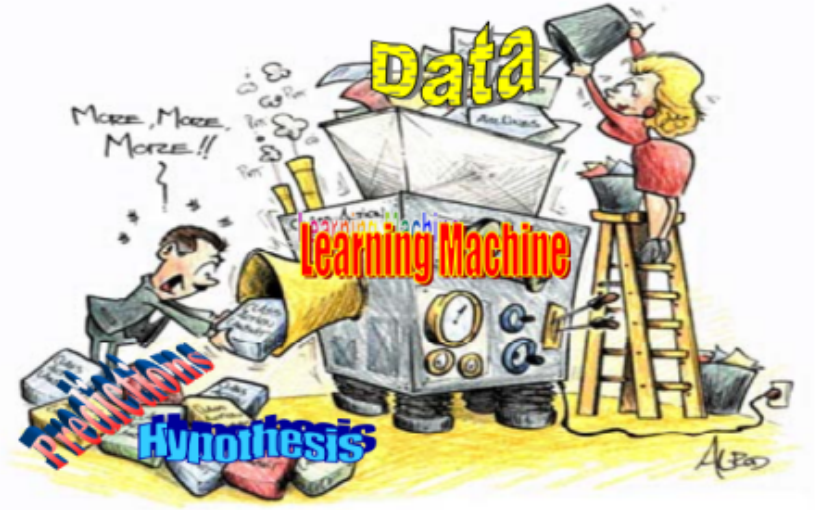
目录

Topic 1.1 概论

一、机器学习应用背景

机器学习的应用背景

- 数据挖掘
  - 使用历史数据改善决策
  - ……



应用背景1：数据挖掘（1）

- 商业智能
  - 例子：商品摆放位置（P&G, Walmart, …）




优惠券、重复消费

应用背景1：数据挖掘（2）

- 信用危机分析

顾客 103 （时间 = t0）	顾客 103 （时间 = t1）	顾客 103 （时间 = tn）
信用年限: 9	信用年限: 9	信用年限: 9
贷款余额: $2,400	贷款余额: $3,250	贷款余额: $4,500
收入: $52K	收入: ?	收入: ?
拥有房产: 是	拥有房产: 是	拥有房产: 是
违约次数: 2	违约次数: 2	违约次数: 3
最长延迟付费周期: 3	最长延迟付费周期: 4	最长延迟付费周期: 6
有价值客户?:	**有价值客户 ?:**	**有价值客户?:**
?	**?**	**否**

信用卡公司的客户确认电话

机器学习的应用背景2.1

- 数据挖掘
  - 使用的历史数据改善决策
- **个性化定制**
  - **邮件过滤**
  - 学习用户兴趣的应用（e.g. 信息流、论坛、社交网络 ...）
  - 电子商务中的推荐
  - ……

机器学习的应用背景2.2

- 数据挖掘
  - 使用的历史数据改善决策
- **个性化定制**
  - 邮件过滤
  - **学习用户兴趣的应用（e.g. 信息流、论坛、社交网络 ...）**
  - 电子商务中的推荐
  - ……

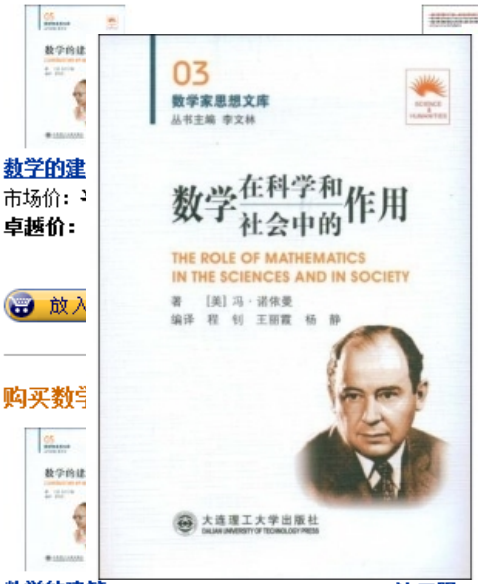
机器学习的应用背景2.3

- 数据挖掘
  - 使用的历史数据改善决策
- **个性化定制**
  - 邮件过滤
  - 学习用户兴趣的应用（e.g. 信息流、论坛、社交网络 ...）
  - **电子商务中的推荐**
  - ……

应用背景2.3：社交网络中的推荐

应用背景2.3：电子商务中的推荐

购买数学在科学和社会中的作用的顾客同时也购买了：



数学的建
市场价：
卓越价：
放入

魔法数学—世界上最简单的心算
¥20.00元
¥10.70元
购物车

一个数学家的辩白
市场价：¥16.50元
卓越价：¥13.30元
放入购物车

购买数学

浏览了：

数学的建筑
市场价：¥24.00元
卓越价：¥17.00元
放入购物车

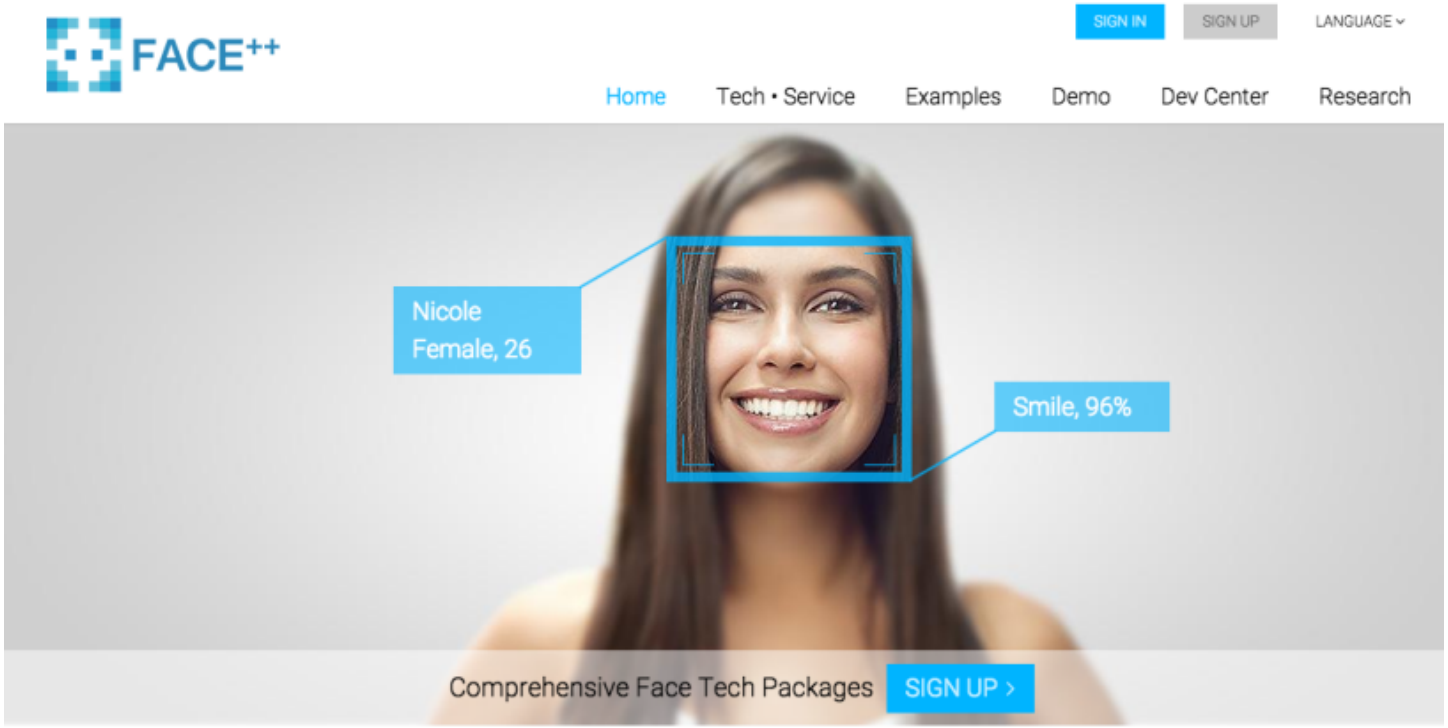
论无限—无限的数学与哲学
市场价：¥30.00元
卓越价：¥25.30元
放入购物车

寻找前世之旅前传·阴阳师物语
市场价：¥24.80元
卓越价：¥12.30元
放入购物车

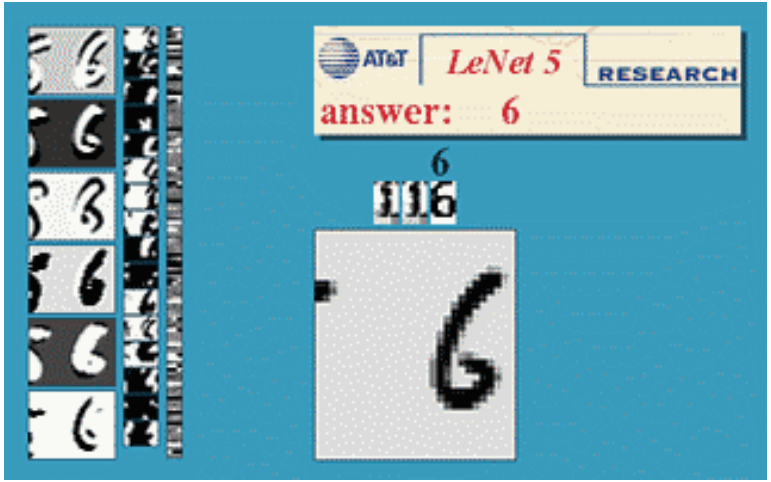
机器学习的应用背景3.1

- 数据挖掘
  - 使用的历史数据改善决策
- 个性化定制
  - 邮件过滤
  - 学习用户兴趣的应用（e.g. 信息流、论坛、社交网络 ...）
  - 电子商务中的推荐
  - ……
- **替代人力的软件应用**
  - **模式识别：人脸识别，语音识别，手势识别，OCR, ……**
  - 自动驾驶
  - 信息检索（如搜索引擎）
  - ……

3.1 人脸识别


FACE++
SIGN IN
SIGN UP
LANGUAGE
Home
Tech • Service
Examples
Demo
Dev Center
Research
Nicole
Female, 26
Smile, 96%
Comprehensive Face Tech Packages
SIGN UP ›

3.1 语音识别与手写识别



http://yann.lecun.com/exdb/lenet/index.html

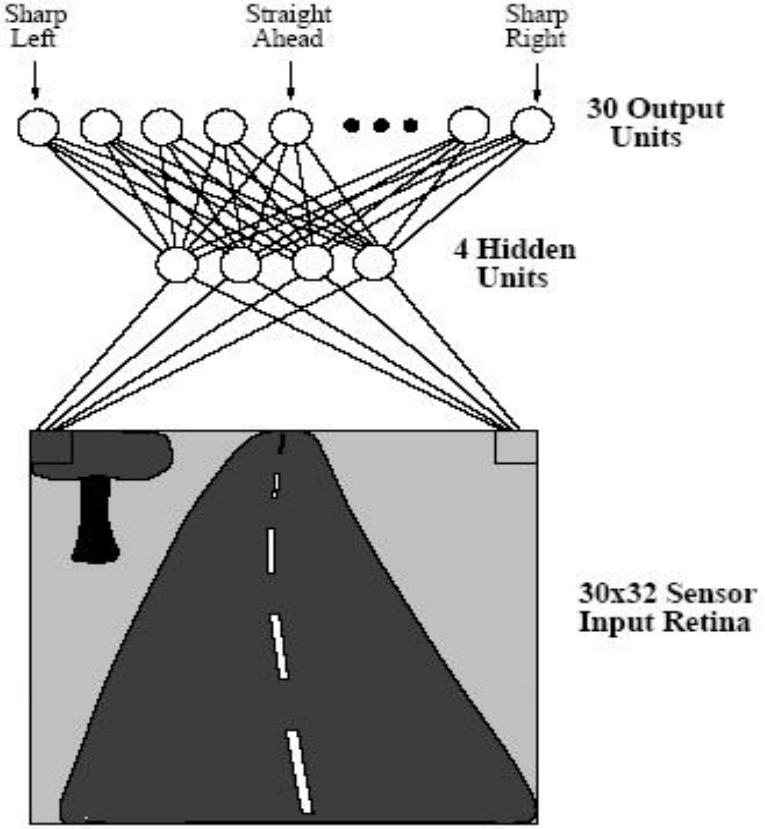
机器学习的应用背景3.2

- 数据挖掘
  - 使用的历史数据改善决策
- 个性化定制
  - 邮件（机构）过滤
  - 学习用户兴趣的应用（e.g. 信息流、论坛、社交网络 ...)
  - 电子商务中的推荐
  - ……
- **替代人力的软件应用**
  - 模式识别：人脸识别，语音识别，手势识别，OCR, ……
  - **自动驾驶**
  - 信息检索（如搜索引擎）
  - ……

例：自动驾驶（CMU）

- ALVINN (1989~1996, CMU)

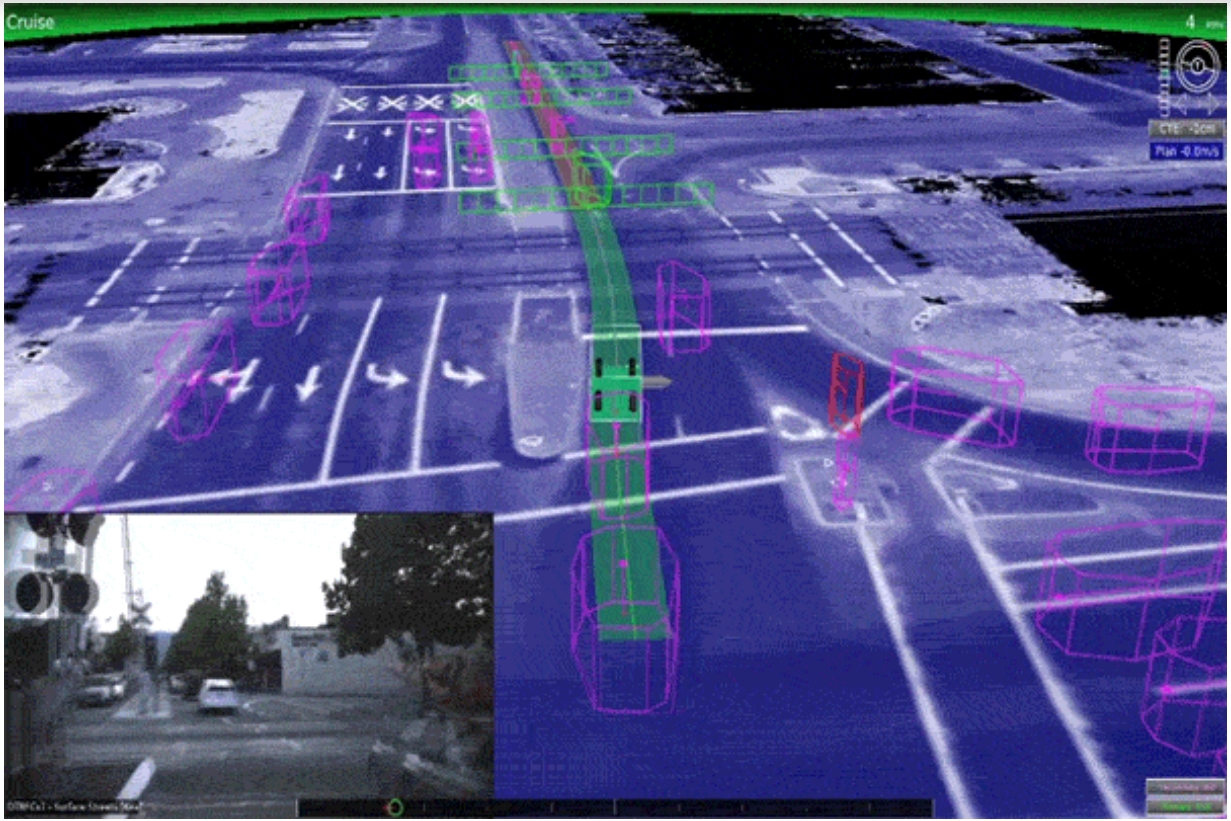
  70mph, highway



例：自动驾驶: Google Driverless Car

- 2009 Google 自动驾驶项目
- Dec. 2016 Google 将项目转移到一家新公司 *Waymo*

WAYMO

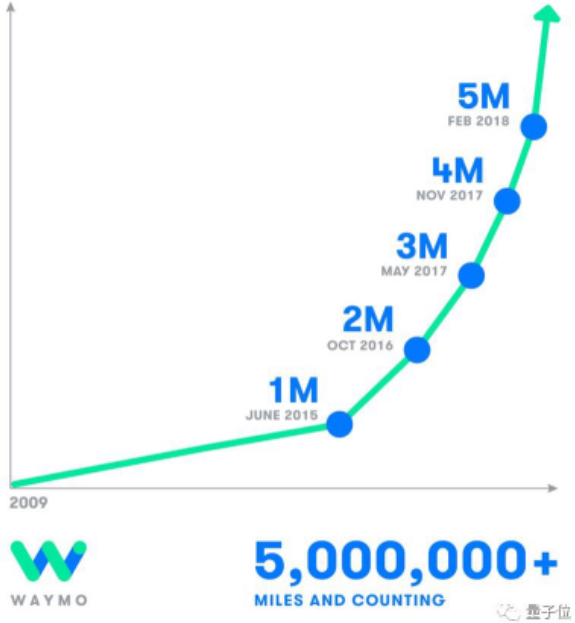


http://www.extremetech.com/extreme/189486-how-googles-self-driving-cars-detect-and-avoid-obstacles

Waymo: 超过 500 万公里的测试 (Feb. 28 2018)

- 超过 25 个美国城市: 雪地、山地、沙漠、城市
- 路程: 500 万公里
- 虚拟: 27 亿公里

1000+ Chrysler Pacifica by the end of 2018.



https://techcrunch.com/2018/02/28/waymo-360-degree-video-shows-how-autonomous-vehicles-work/

Uber 无人驾驶汽车事故 Sunday, Mar. 18 2018

- 2018 年 3 月 18 日，一辆 Uber 无人驾驶汽车发生事故导致 49 岁的 Elaine Herzberg 死亡。
- 一位 Uber 工程师在汽车驾驶座上，但发生事故时正处于自动驾驶模式。这被认为是第一起有关自动驾驶的严重事故。
  - Toyota 停止了自动驾驶技术在公共道路上的测试

机器学习的应用背景 3.3

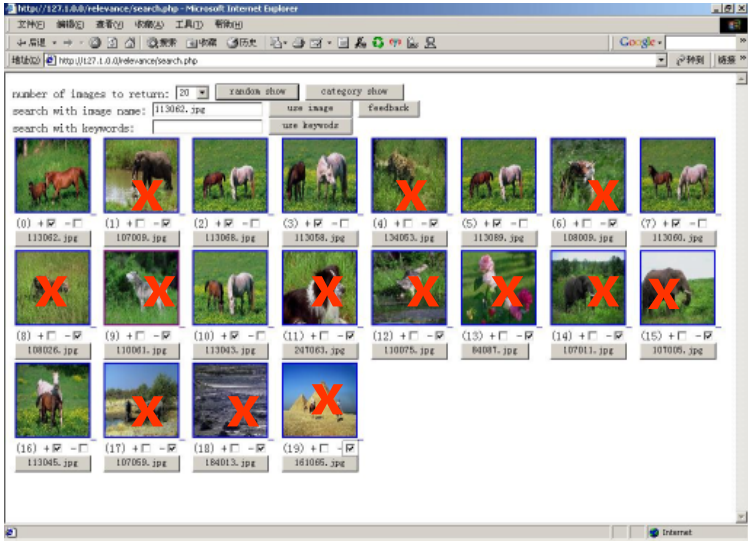
- 数据挖掘
  - 使用的历史数据改善决策
- 个性化定制
  - 邮件（机构）过滤
  - 学习用户兴趣的应用（e.g. 信息流、论坛、社交网络 ...)
  - 电子商务中的推荐
  - ……
- **替代人力的软件应用**
  - 模式识别：人脸识别，语音识别，手势识别，OCR, ……
  - 自动驾驶
  - **信息检索（如搜索引擎）**
  - ……

例：信息检索

- 排序：1000+ 参数
- Learning to rank（机器学习 + 信息检索）

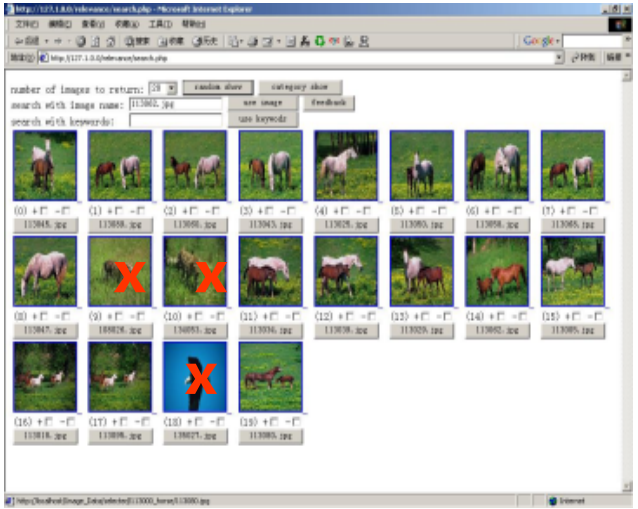
例：信息检索（续）

- 相关性反馈
  - e.g. 图像检索、视频检索 ……



查询: 一张关于马的图片
反馈前的检索结果

有 1 个相关性反馈后的检索结果



机器学习的应用背景

- 数据挖掘
  - 使用的历史数据改善决策
- 个性化定制
  - 邮件过滤
  - 学习用户兴趣的应用（e.g. 信息流、论坛、社交网络 ...)
  - 电子商务中的推荐
  - ……
- **替代人力的软件应用**
  - 模式识别：人脸识别，语音识别，手势识别，OCR, ……
  - 自动驾驶
  - **信息检索（如搜索引擎）**
  - ……

现在你应该已经大致了解了机器学习和以下这些概念的区别:

- 数据挖掘 / 模式识别 ...
- 神经网络 / 深度学习 ...
- ...

二、什么是机器学习？

什么是机器学习（1）

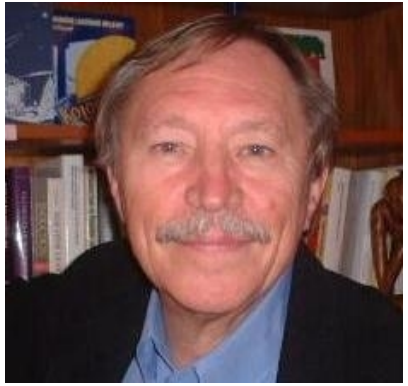
- “学习是要表示出系统中的变化 …

  使得系统在下次进行同样的任务时变得更有效”

  -- Herbert Simon

- Herbert Simon (1916 – 2001)
  - 1956，达特茅斯会议，“人工智能之父”
  - 1975 年获得图灵奖
  - 1978年获的诺贝尔经济学奖
  - 1986年获得国家科学奖章
  - 1993年由于其心理学方面的杰出贡献被授予美国心理学会奖
  - 1994年他成为一名外籍中科院科学家

什么是机器学习（2）



- “学习是对经验的表示方法的构造或修改”

  --Ryszard S. Michalski

- Ryszard S. Michalski (1937-2007)

  http://www.mli.gmu.edu/michalski/

  - Michalski, Ryszard S. and Kodratoff, Y. Machine Learning, an AI approach （《机器学习：一种人工智能方法》） 1990
  - 共同创始人: Machine learning research field
  - 共同创始人: Machine Learning (Journal)
  - 共同创始人: ICML

什么是机器学习（3）

- 学习 = 在某种任务上基于经验不断进步
- Tom M. Mitchell (CMU)
  - 1973 MIT S.B. ; 1979 Ph.D. Stanford Uni.
  - 共同创始人: Machine Learning (Journal)
  - 共同创始人: ICML
  - IJCAI Computers and Thought Award, 1983
- T （Task 任务）
- E （Experience 经验）
- P （Performance 性能）

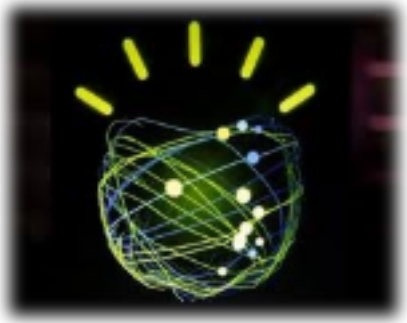
学习：
变化 / 构造或修改 / 进步

什么是机器学习（例）

- 学习如何下国际跳棋
  - T: 下国际跳棋
  - P：获胜率
  - E：e.g. 和自己下棋

- 手写识别
  - T: 识别字符
  - P: 识别精度
  - E: 已知类别的字符集

三、机器学习系统举例



- IBM Watson DeepQA

Jeopardy:

一个美国的电视节目。

需要参赛者分辨出笑话、双关、反讽、字谜等语句中的微妙之处

IBM Watson @ Jeopardy

- February 14, 15, and 16, 2011
  - *Jeopardy* 的两个著名冠军
  - Brad Rutter（右）：
    - 赢得 *Jeopardy 史上最多的奖金*（325 万美金）
    - Johns Hopkins 大学辍学生
  - Ken Jennings（左）：
    - *Jeopardy* 连胜纪录保持者（2004年连续获胜 74 场）
    - 拥有 Brigham Young 大学的计算机和英语学位以及Seoul Foreign 的学士学位

IBM Watson 在 Jeopardy 中获胜

体现了在问答领域的出色成果

最终结果:

$77,147

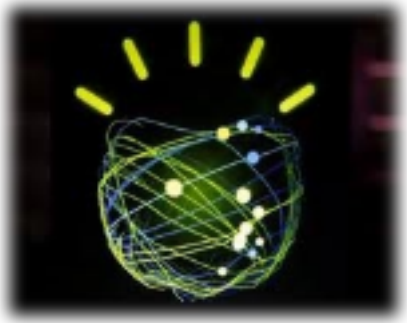
(5,000 + 35,734 + 41,413)

vs.

$21,600 &

$24,000.

IBM Watson



- 开发 4 年
- 90 台 Power 7 服务器 （每台有4 个 8 核 power 7 处理器）
- 基于大规模知识库而不是互联网的检索(没有联网）
  - 3 秒内在上百亿的页面进行检索
- 用之前节目的题目进行训练
  - Jeopardy 参与者: 77 (2009) + 55 (2010, 优胜者)
  - 缺乏实时学习的能力

Category: US City

Q: "Its largest airport was named for a World War II hero; its second largest, for a World War II battle."

A: "What is Toronto?" (Chicago)

技术需求

- 回答任意话题的问题
  - 自然科学、地理、流行文化 …
- 准确度：不只是一个答案，还需要高置信度
- 速度：3 秒内甚至更快

- 语言理解
  - 解析复杂句子，理解笑话、双关、反讽等
- 问题的实时分析
- 从错误中学习
- 应对意料之外的情况 …

相关技术 -- DeepQA

- 一种大规模的基于概率和实例的问答架构
  - 不基于数据库
  - 深度文本分析
    - NLP 以及基于统计的 NLP
  - 确定多种相似可能性的置信度
    - 投票、问题解释…
  - 搜索
  - 风险评估
  - Hadoop、UIMA
- 现实应用场景中的挑战 / 问题

下节课：

Topic1.2 通用机器学习系统的设计