综合实训：
数据挖掘与社会网络技术及应用

郑子彬　教授

中山大学 数据科学与计算机学院

http://www.inpluslab.com

2019年

课程信息

- 教学目标：
  - 掌握数据挖掘领域的常用挖掘工具与算法实现
  - 通过参加数据挖掘比赛，基于真实的业界数据来建模及设计出独立思考的方案

- 两个数据挖掘比赛
  - 2018年甜橙金融杯大数据建模大赛（案例教学）

    http://www.dcjingsai.com/common/cmpt/2018年甜橙金融杯大数据建模大赛_竞赛信息.html

  - 待定（实际比赛）

授课老师信息

授课老师：郑子彬 教授、博导 (百人计划)

- 中山大学 计算机系 学士、硕士；
- 香港中文大学 博士
- 软件工程系主任
- 区块链与智能金融研究中心主任

助教信息

- 助教：马蒙蒙(mmma@inpluslab.com)
- 助教：陈志豪(zhhchen2@inpluslab.com)

- Office：东校区南实验楼D203

课程安排

第一次：知识准备

- 课程总体介绍
- 大数据挖掘简介
- 甜橙金融杯项目大数据建模大赛 总体介绍

课程安排

第二次：案例挖掘入门

- 甜橙金融杯项目大数据建模大赛 项目技术介绍
- 数据预处理
- 特征选择方法
- 常用模型介绍

第三次：案例挖掘总结

- 实战比赛总体介绍
- 甜橙金融杯项目大数据建模大赛 比赛总结
- 数据挖掘竞赛经验交流

课程安排

第四次：新案例挖掘

- 数据挖掘竞赛经验交流
- 新案例问题答疑

第五次：实训总结

- 比赛总结与展示
- 上交实训报告

成绩评定

项目	比例
比赛排名	40%
比赛代码及report	30%
考勤	20%
周报（两周一次）	10%

以个人形式完成比赛，同学间可以相互讨论，但严禁抄袭！！！

大数据挖掘

郑子彬

中山大学 数据科学与计算机学院

http://www.inpluslab.com

2019年

例子

- 如何找出美国人最喜欢的派?

例子

- 根据销售记录，30寸的派中，苹果派卖的最好

- 对于11寸的派，苹果派只能排4到5名，为什么呢?

例子

- 30寸的派：整个家庭需要都能够接受的口味，苹果派是大家都能接受的，但是不一定是最喜欢的，妥协的结果

- 11寸的派，自己一个人吃，选择自己最喜欢的口味

- **大数据 → 获得片面数据无法获得的信息**

什么叫大数据

某必胜客店的电话铃响了，客服人员拿起电话。

- 客服：必胜客。您好，请问有什么需要我为您服务？
- 顾客：你好，我想要一份……
- 客服：先生，烦请先把您的会员卡号告诉我。
- 顾客：16846146***。
- 客服：陈先生，您好！您是住在泉州路一号12楼1205室，您家电话是2646****，您公司电话是4666****，您的手机是1391234****。请问您想用哪一个电话付费？
- 顾客：你为什么知道我所有的电话号码？
- 客服：陈先生，因为我们联机到CRM系统。
- 顾客：我想要一个海鲜比萨……
- 客服：陈先生，海鲜比萨不适合您。
- 顾客：为什么？
- 客服：根据您的医疗记录，你的血压和胆固醇都偏高。

什么叫大数据

- 顾客：那你们有什么可以推荐的？
- 客服：您可以试试我们的低脂健康比萨。
- 顾客：你怎么知道我会喜欢吃这种的？
- 客服：您上星期一在国家图书馆借了一本《低脂健康食谱》。
- 顾客：好。那我要一个家庭特大号比萨，要付多少钱？
- 客服：99元，这个足够您一家六口吃了。但您母亲应该少吃，她上个月刚刚做了心脏搭桥手术，还处在恢复期。
- 顾客：那可以刷卡吗？
- 客服：陈先生，对不起。请您付现款，因为您的信用卡已经刷爆了，您现在还欠银行4807元，而且还不包括房贷利息。
- 顾客：那我先去附近的提款机提款。
- 客服：陈先生，根据您的记录，您已经超过今日提款限额。

什么叫大数据

- 顾客：算了，你们直接把比萨送我家吧，家里有现金。你们多久会送到？
- 客服：大约30分钟。如果您不想等，可以自己骑车来。
- 顾客：为什么?
- 客服：根据我们CRM全球定位系统的车辆行驶自动跟踪系统记录。您登记有一辆车号为SB-748的摩托车，而目前您正在解放路东段华联商场右侧骑着这辆摩托车。
- 顾客:当即晕倒……

大数据元年

2012年2月《纽约时报》的一篇专栏中所称，“大数据”时代已经降临，在商业、经济及其他领域中

2012年3月份美国奥巴马政府发布了“大数据研究和发展倡议”

2012年5月，联合国发表名为《大数据促发展：挑战与机遇》的政务白皮书

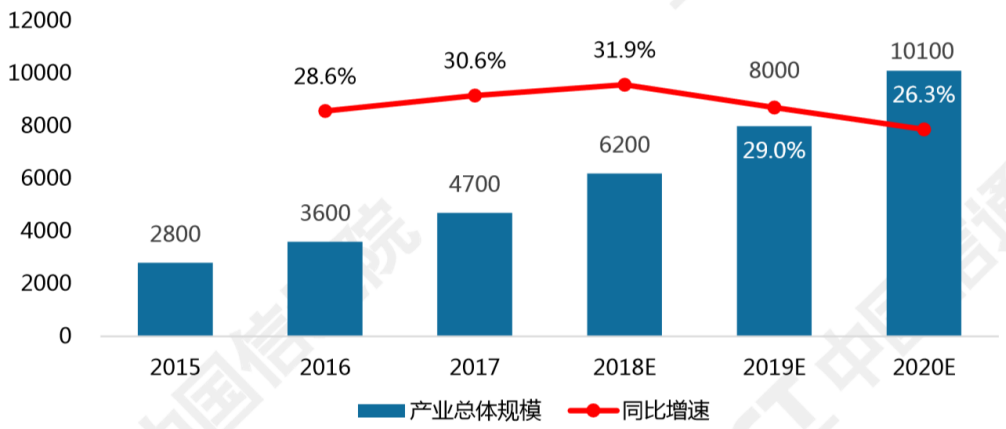
2012年12月13日被命名为首个“中关村大数据日”

随着一系列标志性事件的发生和建立，人们越发感觉到大数据时代的力量。因此2013年被许多国外媒体和专家称为“大数据元年”。

大数据产业规模

- 2017 年我国大数据产业规模为 4700 亿元人民币，同比增长 30%

中国大数据产业总体规模及增速（单位：亿元）

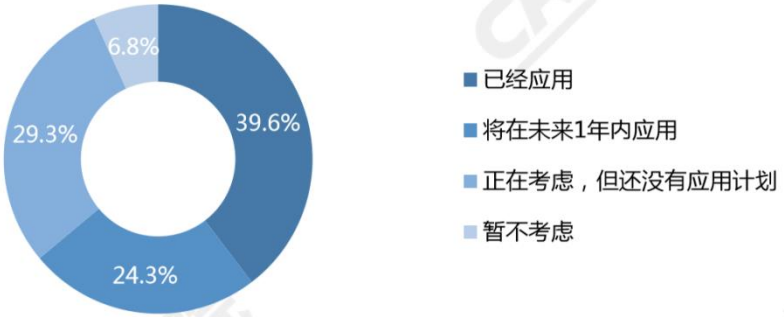


数据来源:中国信息通信研究院

大数据应用现状

- 在接受调查的 1572 家企业中，已经应用大数据的企业有 623 家，占比为 39.6%
- 垂直行业中如金融等领域大数据应用增加趋势较为明显
- 24.3%的企业表示未来一年内将应用大数据

2017年企业对大数据的应用状况（N=1,572）



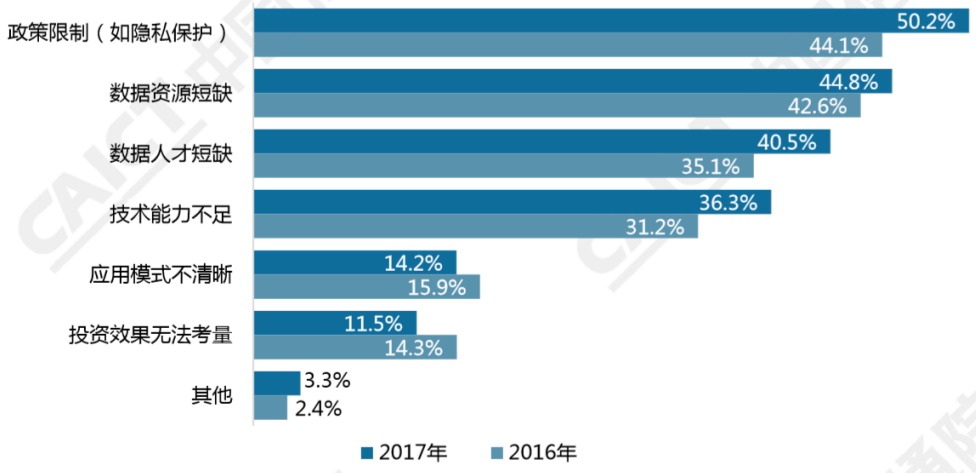
数据来源:中国信息通信研究院

大数据应用现状

- 大数据应用的主要障碍

政策限制、数据资源短缺和数据人才短缺是限制企业大数据发展最主要的三个因素

制约企业大数据发展主要因素（N=1,572）



数据来源:中国信息通信研究院

大数据与实体经济

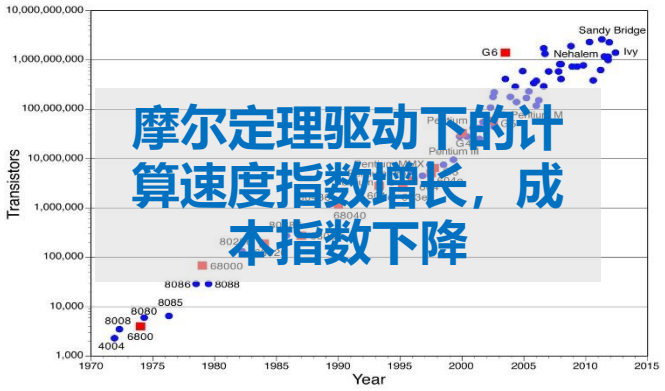
- 《2018年中国大数据产业发展水平评估报告》

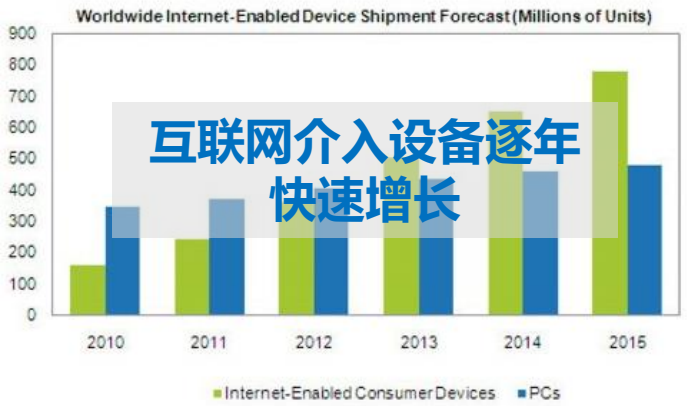
  --工业和信息化部赛迪智库

- 2017年十大行业的大数据发展水平:
  - 金融、电信、政务、交通、商贸、医疗、工业、教育、旅游、农业
  - 金融、电信、政务大数据发展指数分别为45.35、41.69和39.44，超过行业指数平均值30.51
  - 工业领域2017年的指数为24.28，相较2016年的15.41显著提高
  - 而农业的指数最低，仅为8.4

大数据—互联网及其延伸导致的“自然现象”

- **大数据源于信息技术的不断廉价化与互联网及其延伸所带来的无处不在的信息技术应用，四个驱动：**
  - 摩尔定律驱动的指数增长模式 （硬件）
  - 技术低成本化驱动的万物数字化 （技术）
  - 宽带移动泛在互联驱动的人机物广泛联接 （联接）
  - 云计算模式驱动的数据大规模汇聚 （平台）





信息化3.0—大数据开启信息化的第三波浪潮

全球数据总量统计
(数据来源：IDG)

- 2003年：5百万TB
- 2009年：约8亿TB
- 2012年：约27亿TB
- 2020年：预计440亿TB

信息化1.0

—1995
以单机应用为特征的数字化阶段

信息化2.0

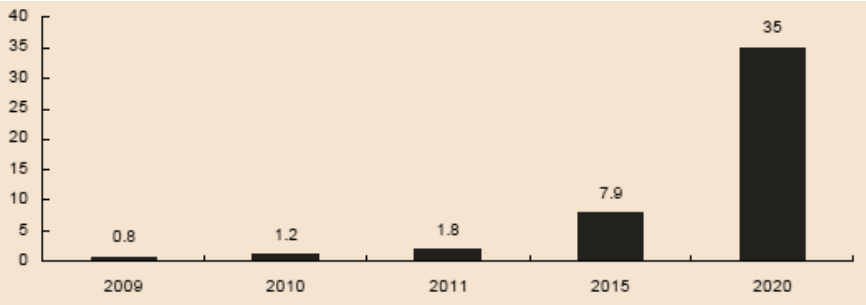
—2010
以联网应用为特征的一个网络阶段

信息化3.0

2010—
以数据的深度挖掘和融合应用为特征的一智慧化阶段

Big Data时代到来

数据量增加



根据IDC 监测，人类产生的数据量正在呈指数级增长，大约每两年翻一番，这个速度在2020 年之前会继续保持下去。这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量

TB ⇨ PB ⇨ EB ⇨ ZB

从数据库到大数据

“池塘捕鱼”VS“大海捕鱼”　　“鱼”是待处理的数据

- 人类产生的数据早已经远远超越了目前人力所能处理的范畴
- 量级的提升带来的挑战，类比：建筑、管理、系统开发

Big Data时代到来

何为大数据？（两个视角定义）

技术能力视角

大数据指的是规模超过现有数据库工具获取、存储、管理和分析能力的数据集，并同时强调并不是超过某个特定数量级的数据集才是大数据。

—麦肯锡《大数据 下一个创新、竞争和生产力的前沿》

大数据内涵视角

大数据是具备海量、高速、多样、可变等特征的多维数据集，需要通过可伸缩的体系结构实现高效的存储、处理和分析。

—MIST《大数据白皮书》

大数据的特点

1. Volume

数据量巨大

全球在2010 年正式进入ZB 时代，IDC预计到2020 年，全球将总共拥有35ZB 的数据量

2. Variety

结构化数据、半结构化数据和非结构化数据

如今的数据类型早已不是单一的文本形式，订单、日志、音频，能力提出了更高的要求

3. Value

沙里淘金，价值密度低

以视频为例，一部一小时的视频，在连续不间断监控过程中，可能有用的数据仅仅只有一两秒。如何通过强大的机器算法更迅速地完成数据的价值“提纯”是目前大数据汹涌背景下亟待解决的难题

4. Velocity

实时获取需要的信息

大数据区分于传统数据最显著的特征。如今已是ZB时代，在如此海量的数据面前，处理数据的效率就是企业的生命

大数据应用的三个层次

- **描述**：关注到底当前发生了什么，把发展的态势描述出来，呈现发展的历程
- **预测**：在分析的基础之上，预测它未来可能会发生什么，呈现事物发展的趋势。比如流感预测，奥斯卡预测等
- **指导性**：指导性的就当前的态势，如果你做一个动作，会产生什么后果，便于根据当前态势做出决策，不仅预测未来，而是做一个动作以后，做一个决策以后，会不会影响未来的结果

数据挖掘

- 山西挖矿
  - 前提是有矿，包括煤矿的储藏量，储藏深度，煤的成色
  - 之后是挖矿，要把这些埋在地下的矿挖出来，需要挖矿工，挖矿机，运输机
  - 之后是加工，洗煤，炼丹，等等
  - 最后才是转化为银子
- 数据挖掘
  - 前提是有数据，包括数据储藏量，储藏深度，数据的成色
  - 之后是数据挖掘，要把这些埋藏的数据挖掘出来
  - 之后是把数据可视化输出，指导分析、商业实践
  - 直到这一步，才创造了价值

大数据：一座正在形成的巨型矿山！

相关领域

- 人工智能
- 机器学习
- 模式识别
- 统计学
- 数据库
- .....

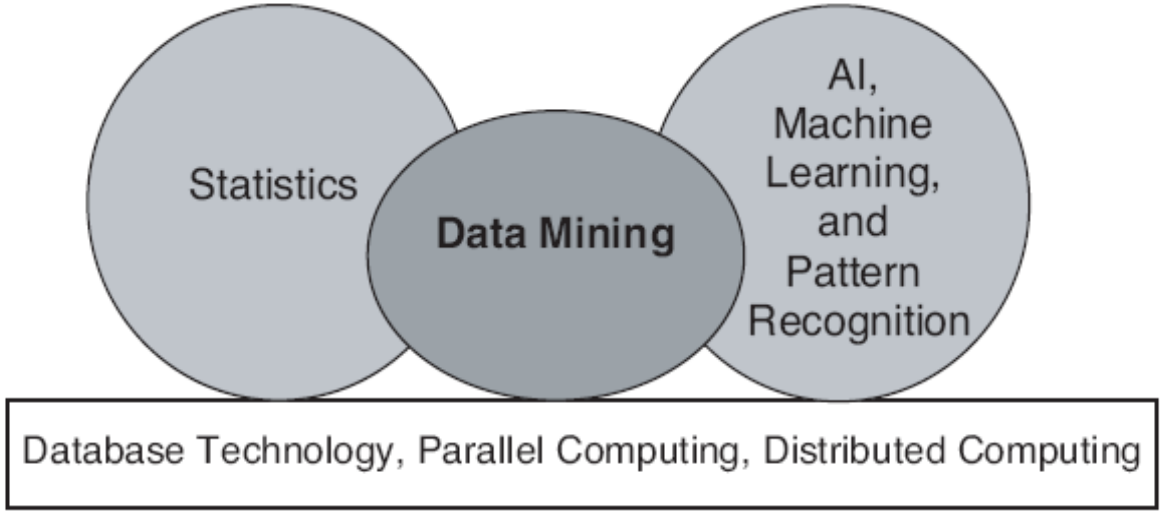
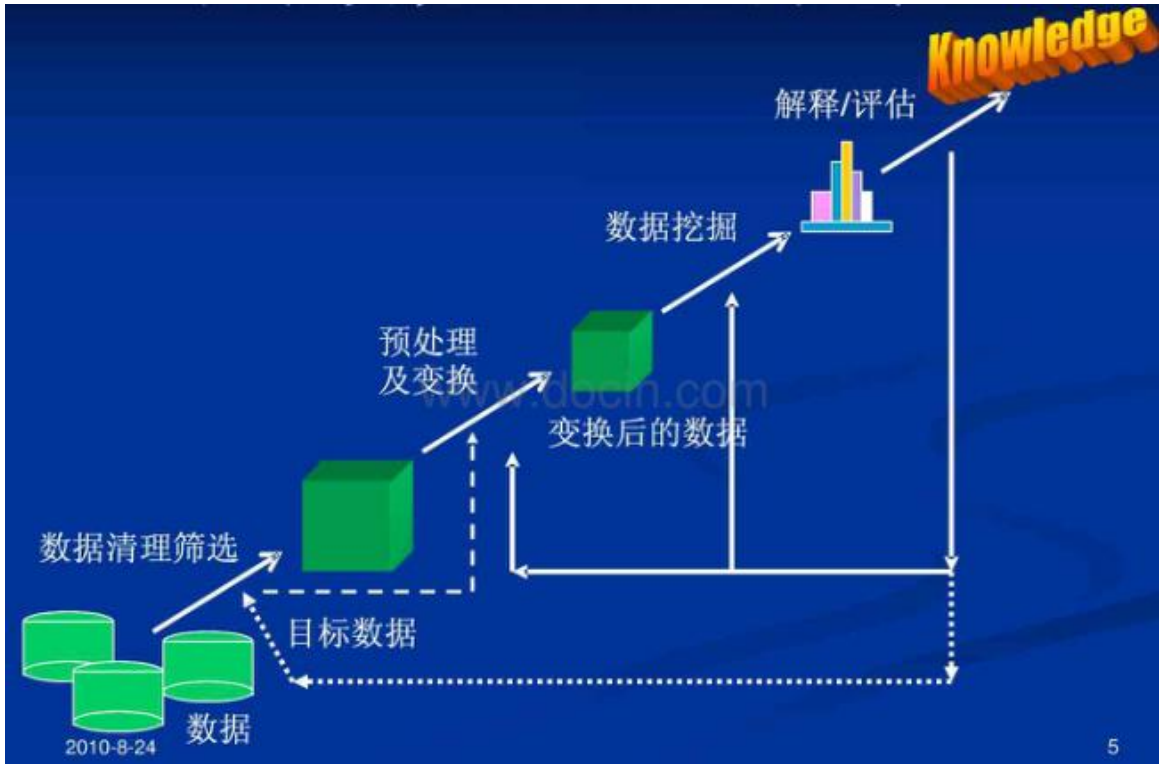


Figure 1.2. Data mining as a confluence of many disciplines.

数据挖掘 vs 知识发现（KDD）

- 数据挖掘是KDD中利用算法处理数据的步骤
- 逐渐演变成KDD(**K**nowledge **D**iscovery in **D**atabase)的同义词



数据挖掘 vs 统计学

- 数据挖掘很多工作由统计方法完成
- 目标相似，许多算法源于数理统计
- 部分统计学家认为数据挖掘是统计学的分支
- 大部分数据挖掘研究人员不这么认为

数据挖掘 vs 传统数据分析方法

- 数据源
  - 数据是海量的
  - 数据有噪声
  - 数据可能非结构化，异构多源
- 传统数据分析方法：假设驱动
  - 给出一个假设，然后通过数据验证
- 数据挖掘：发现驱动
  - 模式从数据中自动提取出来
  - 发现不能靠直觉发现的信息或知识
  - 挖掘出的信息越出乎意料，可能越有价值

例子

- 美国沃尔玛：啤酒跟尿布经常被一起购买
- 打败“康师傅”不是“统一”
  - 是外卖
- 影响“美团外卖”的不是“饿了么”
  - 是共享单车
- 打败口香糖不是益达
  - 而是微信、王者荣耀

例子

伯樵

42分钟前

在开行业会议，某航母级互联网影业的发言人说：“通过大数据挖掘，我们发现不同观众的相关卖品偏好。比如《芳华》的观众比《战狼2》消费了更多的热饮。这些都是以前我们不知道的，也无法预测的。”——

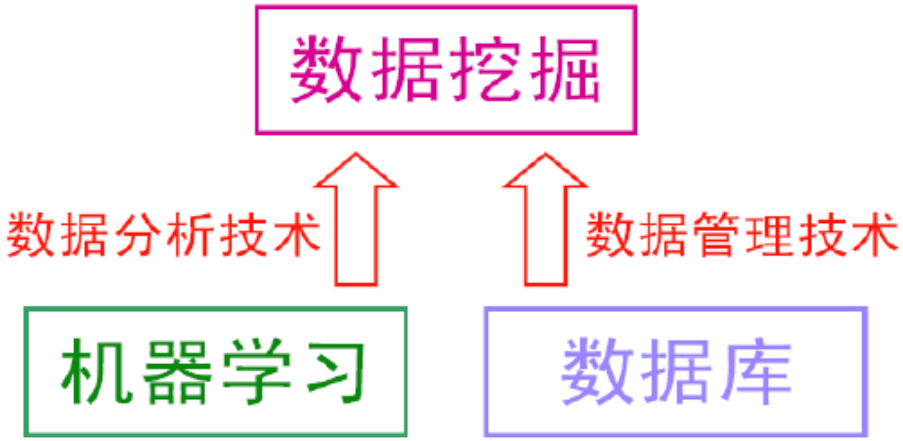
《战狼2》7月底盛夏上映，《芳华》12月15日冬日上映...互联网这大数据挖的...

赞

78

数据挖掘 vs 机器学习

- 机器学习：利用经验来改善计算机系统自身的性能
- 数据挖掘(知识发现)：从海量数据中找出有用的知识
  - 利用机器学习界提供的技术来分析海量数据
  - 利用数据库界提供的技术来管理海量数据



算法与程序员

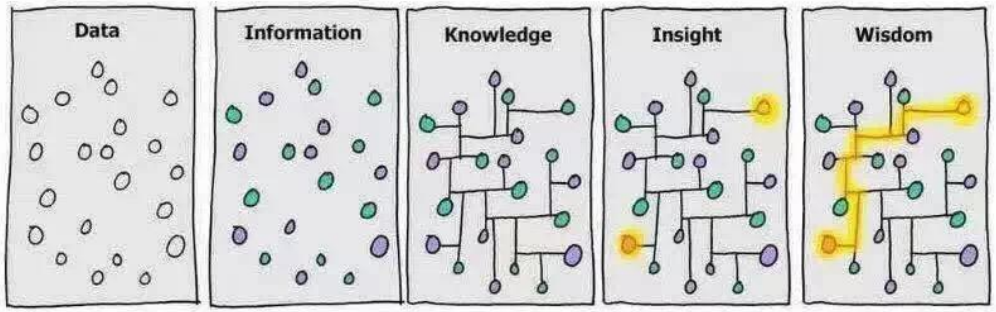
- 算法是一套严格的标准
  - 人们常说，你没法真正了解某样东西，知道你能用一种算法把它表达出来
- 程序员：创造算法并将其编码的人
  - 能任意创造不同的世界
  - 编程语言是他创造世界的工具
  - 天敌：复杂性（时间、空间复杂性，算法复杂性...）

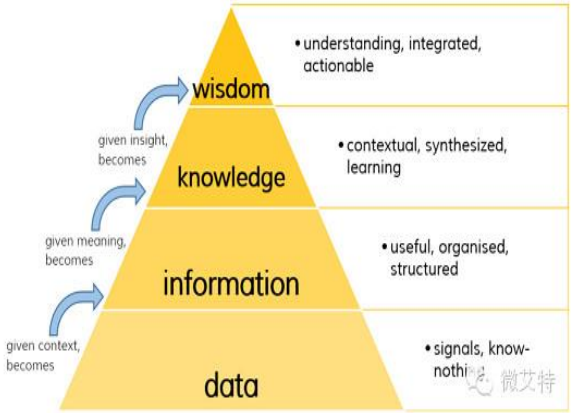
算法 vs 机器学习

- 算法：输入为数据，输出为结果，算法负责中间的处理
- 机器学习：输入为数据，及想要的结果，输出的是算法（把数据转换成结果的算法）
  - 计算机会自己编写程序
  - 比如：手写数字识别、自动驾驶汽车
  - 数据越多，学的越多（大数据时代提供足够的数据）
  - 制定规则 vs 自动学习规则
  - 不同领域类似的学习框架，极大降低复杂性

机器学习

- 在信息处理的生态中，机器学习算法是顶级掠食者
  - 数据是草
  - 网络爬虫、数据库等是食草动物
  - 统计及分析算法是食肉动物，将数据变成信息
  - 机器学习算法把信息吞下、消化、将其变成知识



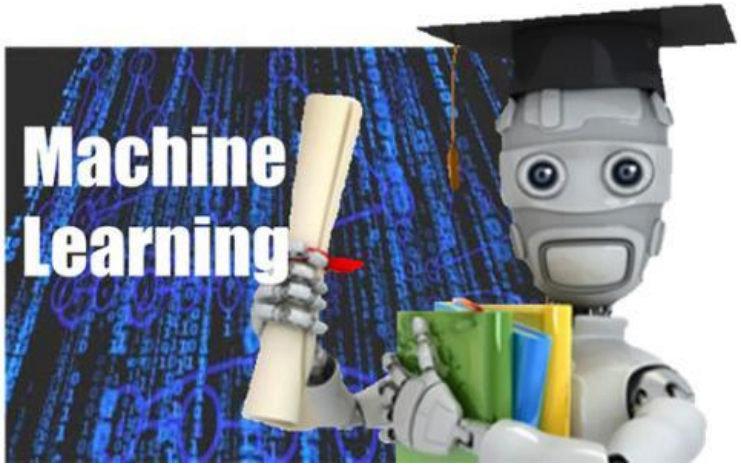


数据的价值

- “数据是新型石油”
  - 提炼石油是一笔大生意
  - 把数据看作战略资产：我有什么数据而竞争对手没有？如何利用好这些数据？
  - 亚马逊：专家写了上千条规则，预测用户喜好
  - 谷歌：算法学习了数十亿条规则
- 商业界拥护机器学习的原因：
  - 潮流新技术？
  - 别无选择

机器学习的未来

- 数据科学家是硅谷最热门职业
- 麦肯锡全球研究院估计，截至2018年，仅美国就需要在培养14万-19万机器学习专家才够用，另外还需要150万有数据头脑的经理
- 机器学习的应用爆发的太快，教育无法跟上其步伐



机器学习应用场景

- 80年代：金融领域，股票预测
- 90年代：挖掘企业数据库
- 网络及电子商务 （个性化）
- 网页搜索及广告投放
- 911之后，打击恐怖主义
- 网络2.0：社交网络（Facebook、微博、微信...）
- 各个领域：分子生物学、天文学
- 2011年：大数据概念流行起来，ML被明确归入全球经济未来的中心
- 各行各业

数据挖掘的主要内容

- 数据及数据预处理
- 分类
- 关联规则
- 聚类
- 协同过滤
- 图挖掘
- 应用案例
- ……

相关学术会议

- SIGIR, KDD, ICDM, SDM, CIKM, PAKDD
- WWW, WSDM
- AAAI, IJCAI
- VLDB, SIGMOD, ICDE
- BigData
- ICML, NIPS
- …

相关学术期刊

- IEEE Transactions on Knowledge and Data Engineering(TKDE)
- ACM Transactions on Knowledge Discovery from Data (TKDD)
- ACM Transactions on Intelligent Systems and Technology (TIST)
- ACM Transactions on Information Systems(TOIS)
- IEEE Transactions on Systems, Man, and Cybernetics, Part B
- IEEE Transactions on Neural Network (TNN)
- Knowledge and Information Systems (KAIS)
- Pattern Recognition (PR)

相关比赛

- 阿里天池比赛: http://tianchi.aliyun.com/
- CCF大数据与计算智能大赛http://www.wid.org.cn/
- Kaggle: https://www.kaggle.com/
- DataCastle: http://www.pkbigdata.com/
- ImageNet: http://image-net.org/challenges/LSVRC/2015/index
- KDD Cup: https://www.kddcup2015.com/information.html
- IJCAI: http://ijcai15.org/index.php/repeat-buyers-prediction-competition

http://www.inpluslab.com

移动互联网与金融大数据实验室