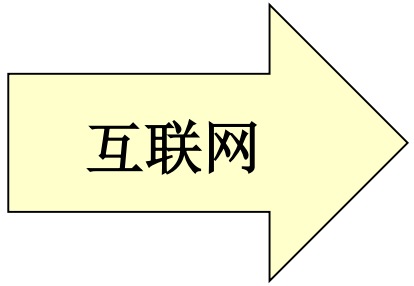
大数据分析基础

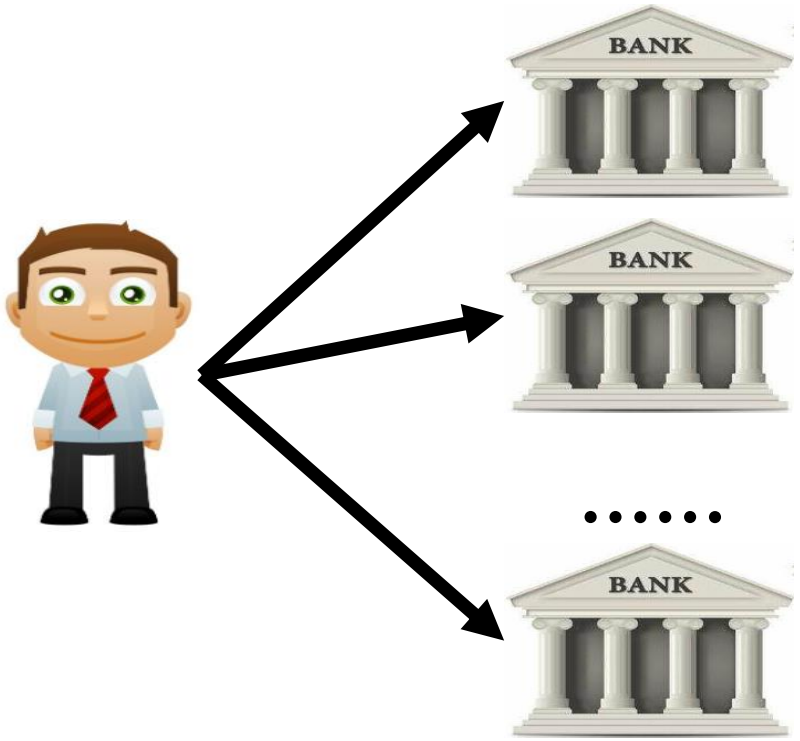
大数据分析简介

About me

- 方斌 (e-mail: fangbin@xmu.edu.cn)，Ph. D.
  - 嘉庚二609
  - 研究方向
    - 在线消费者决策
    - 在线交易市场
    - 社交媒体

互联网时代的商业贷款





- 用户会有大量的信贷企业可供选择以满足其贷款需求
- 借贷行业由以前的借款方占据主动的卖方市场逐渐转变成了贷款方占据主动的买方市场
- 信贷企业之间的竞争已经从单纯的风险控制能力演变成为了客户获取能力和风险控制能力的全面比拼

大数据助力信贷公司获取优质客户

- 利用大数据实现用更低的成本更加效率地获取更优质的客户

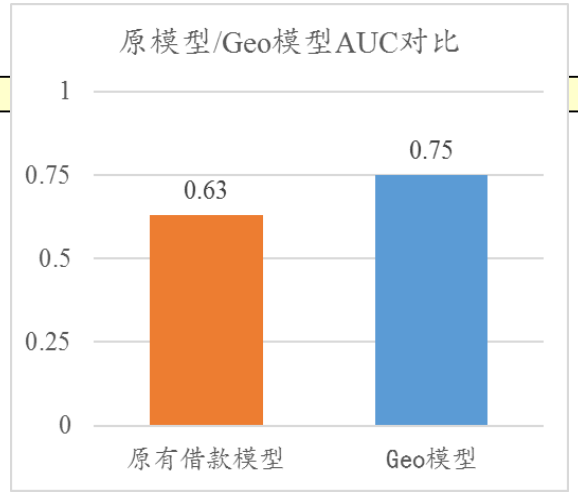
Geo的自有数据和信贷企业的用户数据进行匹配，取共同的用户作为分析对象。

根据Geo和信贷企业双方的数据来对用户进行客户画像

根据客户的需求和价值，筛选出高价值、高需求的目标客户，帮助企业通过线下短信或者外呼准确触达客户。

通过接触，了解客户的实际意向

- 分析每个客户的借贷需求和客户价值。
- 用于分析的数据包括客户的基本属性、网页浏览、关键词、习惯偏好、消费能力等通过各个渠道收集得到的数据。

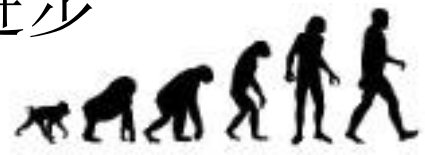


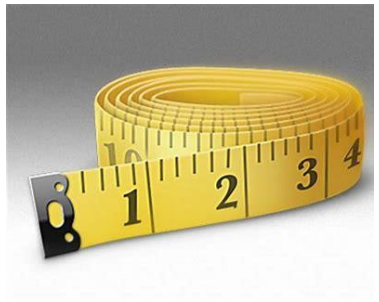
- 预约率为其它项目的**2.5**倍左右
- 每单运营成本下降**39%**

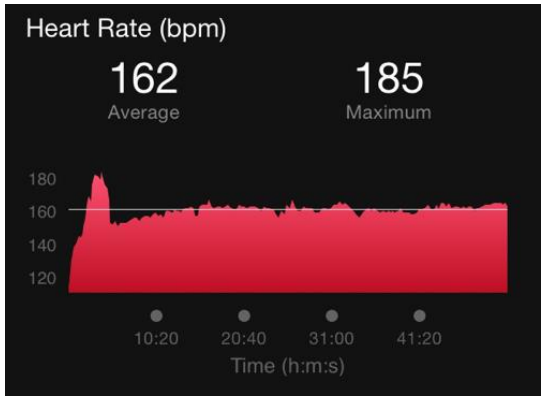
- 商品推荐——你“决定”了你能看到什么
  - 淘宝个性化的首页
  - 你的消费水平决定了淘宝给你看什么：同样的关键词，不同的搜索结果
- 广告投放——抖音广告视频的位置、内容
- 充电宝租赁——不同地区不同价格
- 保险保费千人千面
  - 平安医疗保险运动越多打折越多
  - 基于使用行为的车险（Usage-Based Insurance）
- 供应链管理
  - 亚马逊推出“一小时快递”服务

- 大数据分析为何变得如此普及？
- 技术进步带来了数据爆炸
  - 采集技术的进步
  - 存储（硬件）技术的进步
  - 其它硬件技术的进步

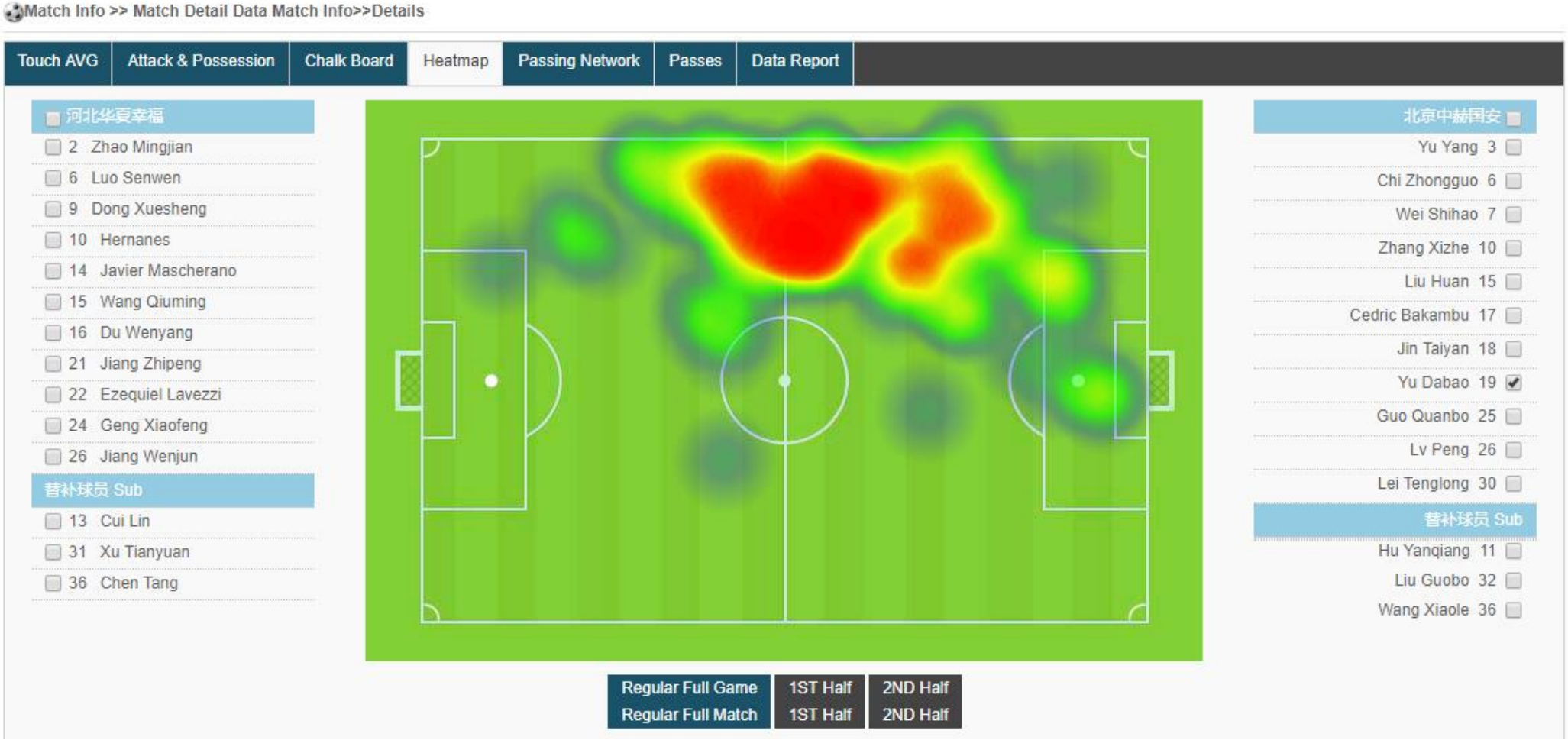
- 数据采集技术的进步
  - 跑步的进化史







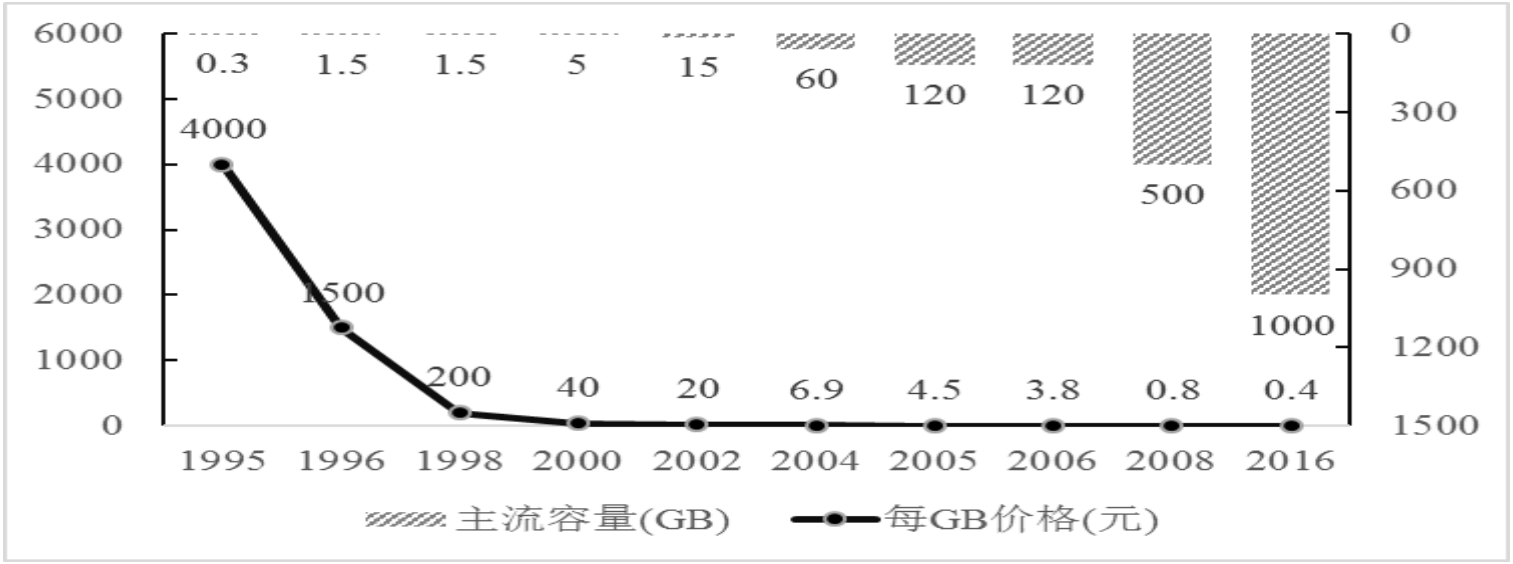
- 数据采集技术的进步


Match Info >> Match Detail Data Match Info>>Details
Touch AVG
Attack & Possession
Chalk Board
Heatmap
Passing Network
Passes
Data Report
河北华夏幸福
2 Zhao Mingjian
6 Luo Senwen
9 Dong Xuesheng
10 Hernanes
14 Javier Mascherano
15 Wang Qiuming
16 Du Wenyang
21 Jiang Zhipeng
22 Ezequiel Lavezzi
24 Geng Xiaofeng
26 Jiang Wenjun
替补球员 Sub
13 Cui Lin
31 Xu Tianyuan
36 Chen Tang
北京中赫国安
Yu Yang 3
Chi Zhongguo 6
Wei Shihao 7
Zhang Xizhe 10
Liu Huan 15
Cedric Bakambu 17
Jin Taiyan 18
Yu Dabao 19
Guo Quanbo 25
Lv Peng 26
Lei Tenglong 30
替补球员 Sub
Hu Yanqiang 11
Liu Guobo 32
Wang Xiaole 36
Regular Full Game
Regular Full Match
1ST Half
1ST Half
2ND Half
2ND Half

- 存储（硬件）技术的进步

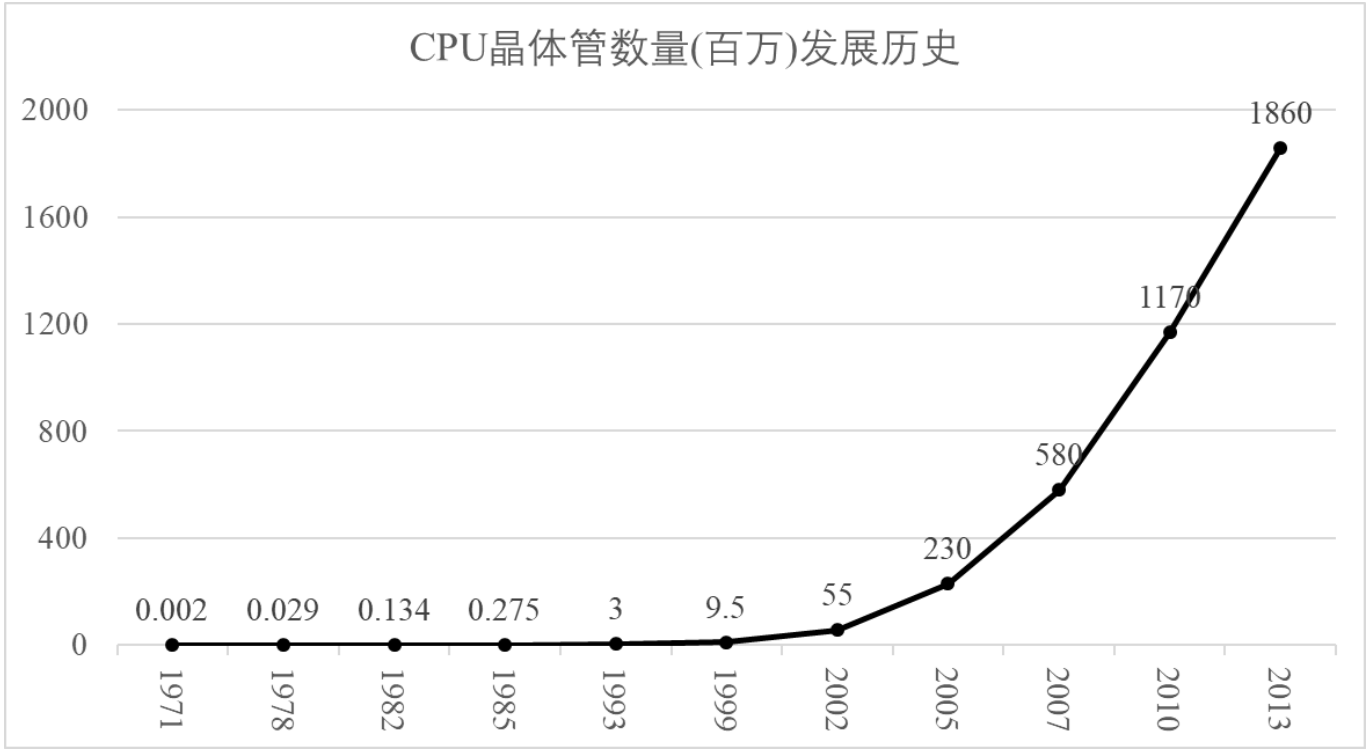

Cloneshit
Podcast
75
2
SM


SONY
SONY
formatted for IBM

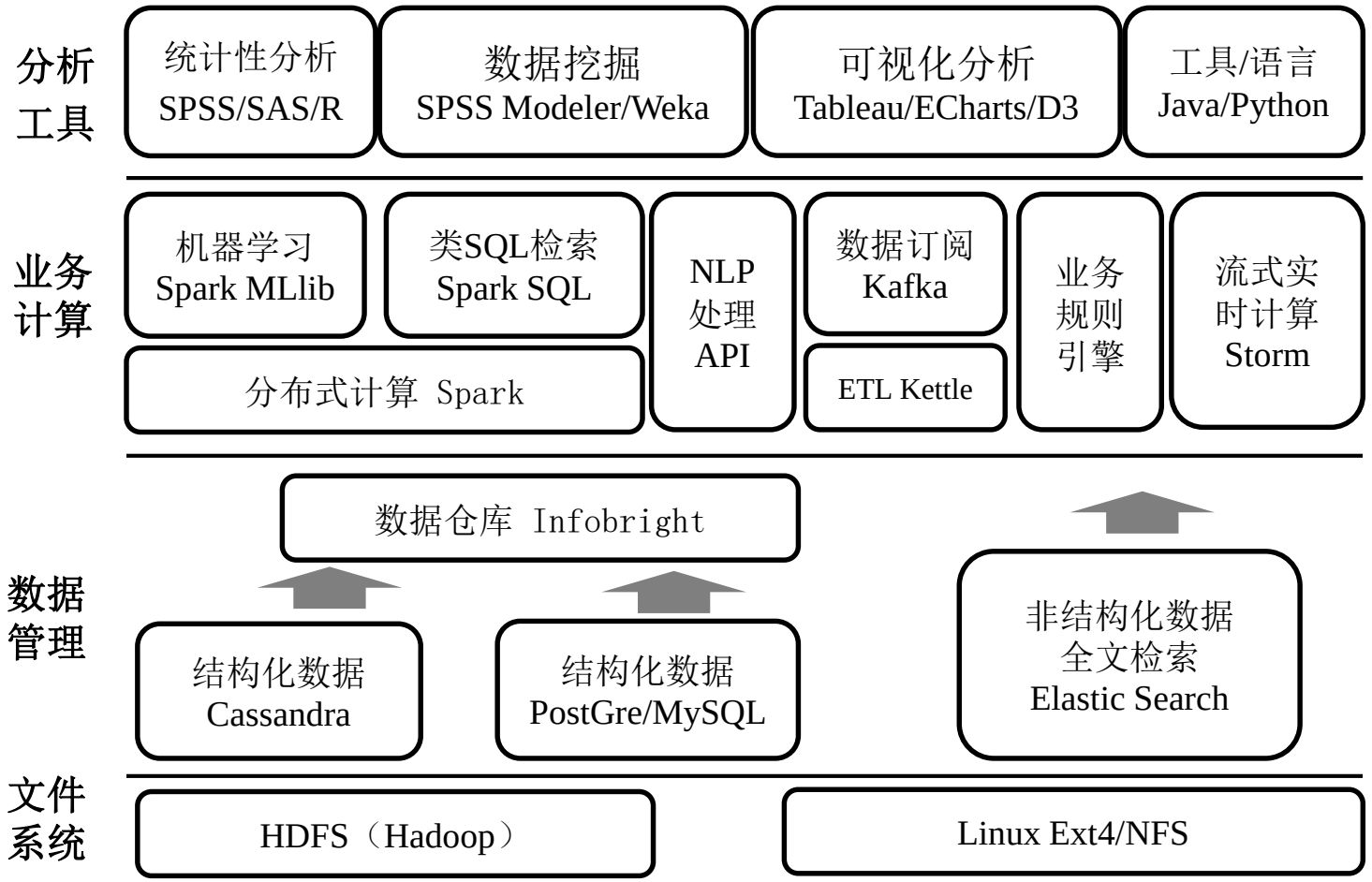

6000
5000
4000
3000
2000
1000
0
0
300
600
900
1200
1500
0.3
1.5
1.5
5
15
60
120
120
500
1000
4000
1500
200
40
20
6.9
4.5
3.8
0.8
0.4
1995
1996
1998
2000
2002
2004
2005
2006
2008
2016
主流容量(GB)
每GB价格(元)


Kingston
60GB

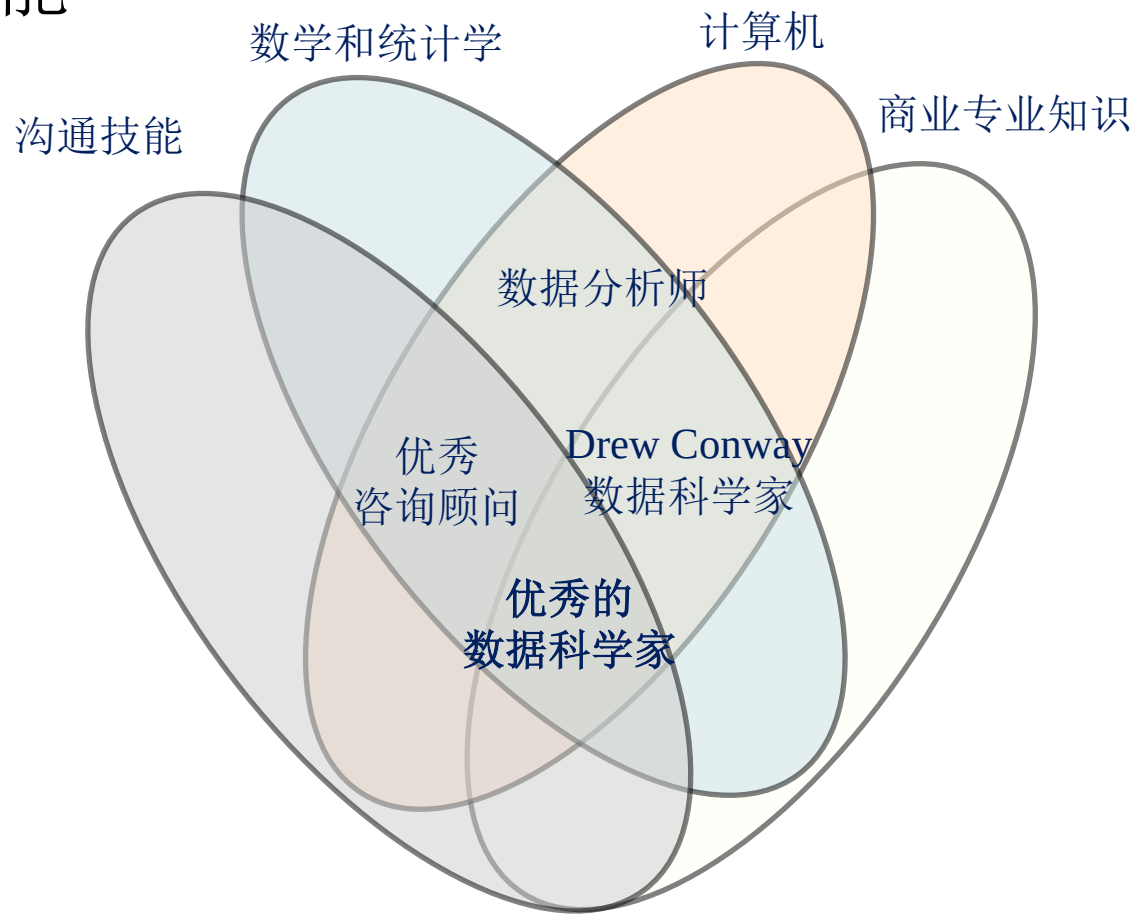
- 其它硬件技术的进步
  - CPU
  - 内存
    - 64KB -> 8GB
  - 网速
    - 64KB/s -> 100MB/s



- 1998年John Mashey在题为《Big Data and the Next Wave of InfraStress Problems, Solutions, Opportunities》的演讲中明确指出了飞速增长的数量和巨大的运算压力之间的尖锐矛盾，并认为这是一个充满了机会和挑战的研究领域。
  - 数据量和硬件计算能力的矛盾
  - 数据量和软件计算工具的矛盾
  - 数据量和数据分析算法的矛盾


分析工具
统计性分析
SPSS/SAS/R
数据挖掘
SPSS Modeler/Weka
可视化分析
Tableau/ECharts/D3
工具/语言
Java/Python
业务计算
机器学习
Spark MLlib
类SQL检索
Spark SQL
NLP处理API
数据订阅
Kafka
业务规则引擎
流式实时计算
Storm
分布式计算 Spark
ETL Kettle
数据管理
数据仓库 Infobright
结构化数据
Cassandra
结构化数据
PostGre/MySQL
非结构化数据
全文检索
Elastic Search
文件系统
HDFS（Hadoop）
Linux Ext4/NFS

数据科学家所需技能


数学和统计学
计算机
沟通技能
商业专业知识
数据分析师
优秀
咨询顾问
Drew Conway
数据科学家
优秀的
数据科学家

技能列表

编程	数学和统计学	商业知识和软技能	沟通能力和可视化技能
计算机科学基础	机器学习	对于商业的热情	与高级管理人员共同工作的能力
脚本语言（如Python）	**统计建模**	对数据的好奇	叙述故事的能力
统计计算软件（如R）	实验设计	不受权威影响	将数据驱动得到的启示转变为决策和行动
数据库：包括SQL和NoSQL	贝叶斯推理	黑客心态	可视化艺术设计
关系代数	监督学习：决策树、随机森林、logistic回归	问题解决者思维	**掌握可视化工具**（如Tableau、D3.js等）
并行数据库和并行查询处理	非监督学习：聚类、降维	具有策略性、主动性、创造性、创新性和合作性	
MapReduce概念	优化：梯度下降等		
Hadoop和Hive/Pig			
定制缩减器（reducer）			
使用AWS等SaaS的经验			

学习目标

- 了解并掌握利用Python进行数据处理和分析
- 了解并掌握数据库进行数据存储
- 建立大数据分析的思维方式
- 学习基础问卷

- 课程内容安排
  - 大数据（商务）分析简介
  - Python入门
    - Python基础知识
    - 利用Python进行可视化
  - 数据库入门
    - 数据库基础
    - Python + MySQL
    - 大数据时代的数据库发展
  - 统计分析入门
    - Python统计分析
  - 小组作业

- 大数据的3V（META集团的分析师Doug Laney，2001）
  - 容量（volume）：数据的容量越来越大
  - 速度（velocity）：数据的增长速度和所需的数据处理速度越来越快
  - 多样（variety）：数据的格式越来越多样化
- 2012年时包括IBM、高德纳（Gartner）等在内的组织将3V扩展到了4V
  - 新加入的属性是真实性（veracity）
  - 真实性强调的是数据的准确性
- 大数据的定义也被认为是“无法用现有的软件工具提取、存储、搜索、共享、分析和处理的海量的、复杂的数据集合。”

- 最重要的V应当是Value
  - 大数据及大数据分析能产生的价值应当是巨大的，只有具有大价值的数据才能叫真正的大数据
  - 海量数据进行无用输入无用输出（Garbage-In Garbage-Out）的处理和分析是没有意义的，也不应当被看作是大数据分析
- 价值的判断应当从两个方面来衡量
  - 结果本身是否对用户而言具有意义
    - 不同的用户都有各自的目标，目标达成那就是有意义的一次数据分析
  - 是否有必要进行大数据分析
    - 是否可以通过合理的抽样将大数据分析转化为传统的数据分析
    - 大数据处理分析的模型和工具是否是对得到期望的结果是不可或缺的

- 数据容量大就是大数据
  - 大数据除了容量大以外还有数据变化速度快、数据类型丰富这两个特点
  - 容量大不等于覆盖全面
    - 美国总统富兰克林·罗斯福在1932年他第一次当总统的时候，美国和许多国家正在遭受经济危机，罗斯福面临的压力也很大。因此到了1936年罗斯福想竞选自己的第二任总统的时候，美国许多人预测罗斯福很难连任。罗斯福的主要竞选对手是兰登。
    - 《文学文摘》在杂志里面夹上关于总统选举的调查问卷，然后收集反馈，最后收回来的有效问卷是240万份。根据这个调查结果，文学文摘宣布他们预测兰登将战胜罗斯福赢得大选。
    - 当时还有一个机构，准确地说是一个年轻人，叫盖洛普，他的预测结果跟文学文摘的预测正好相反。他只调查了5000个人，根据这5000人的调查结果，盖洛普预测罗斯福当选。
    - 罗斯福果然成功连任总统，盖洛普的预测胜利了。文学文摘因为这个事情后来就关门倒闭了。
    - 为什么会失败？文学文摘虽然号称调查了240万人之多，但是它调查的主要群体，是当时美国国内相对而言有钱的那部分人。

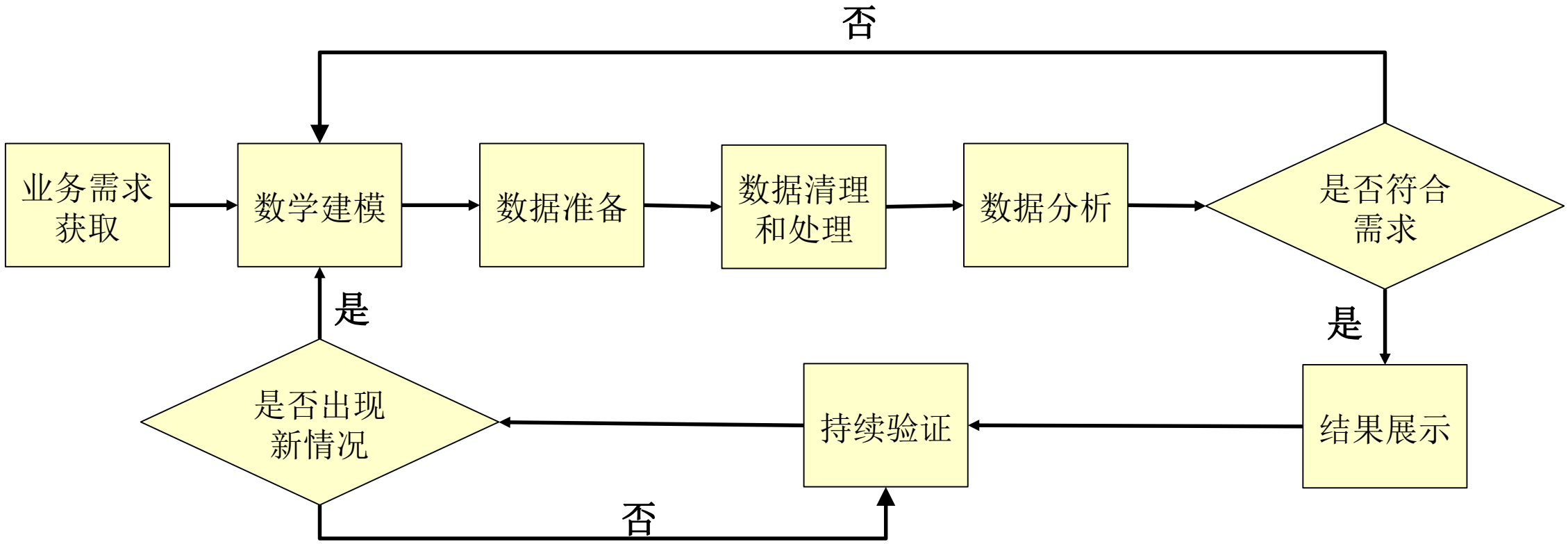
- 大数据只关注相关性不需要找出因果关系
  - “大数据分析只需要相关，不需要因果”（Viktor Mayer-Schönberger，《Big Data:A Revolution That Will Transform How We Live, Work, and Think》）
  - 橙色车出现质量问题的概率仅为其它颜色车的一半。汽车的质量和颜色相关?
  - Google搜索指数真的能预测流感?
  - FICO在2011年提出了“遵从医嘱评分”——他们通过数据分析发现是否有私家车与是否按时吃药和是否使用抗生素存在着相关性。背后的原因是什么?
    - 一个可能的原因就是一个人如果连私家车都无法负担，那么他的生活水平和收入状况很可能非常糟糕，也就无法负担起医药的费用，因此就会出现不遵照医嘱吃药和使用抗生素的情况。

- 为了大数据而大数据
  - 不是所有的企业都适合使用或者需要使用大数据技术
- 只有大数据才能拯救世界
  - 大数据目前的技术和应用都是在数据分析、数据仓库等方面，主要针对OLAP
  - 没有在OLTP方面做出革命性的改变
- 所有的数据分析都称为大数据分析
  - 比如对用户进行了人口统计学特征分析、简单的描述性行为分析等
  - 数据分析通过**新的处理模式**对大数据进行处理的一个过程
    - 新的模式是指新的数据存储、处理、分析的方法

Analysis v.s. Analytics

- Business analysis和business analytics都翻译为商业分析，但是两者有很大区别。
- Business analysis以定性的分析为主
  - 比如SWOT分析、PEST分析等，同时辅以简单的定量分析，比如描述性统计分析等。
- Business analytics的核心是用定量的分析方法对公司运营数据和外界商业数据的深度分析和挖掘
  - 通常都会采用数据挖掘、统计学等学科的技术、模型和算法
  - 得到business analysis所无法得到的隐藏在数据中的管理启示
  - 大数据商务分析就是在business analytics中应用了大数据分析技术

大数据分析基本流程


否
业务需求获取
数学建模
数据准备
数据清理和处理
数据分析
是否符合需求
是
是
是否出现新情况
持续验证
结果展示
否

- 考核方式
  - 平时
    - 出勤
    - 课堂表现
    - 个人作业
  - 小组展示
    - 小组得分
    - 小组组员互评
  - 期末考试