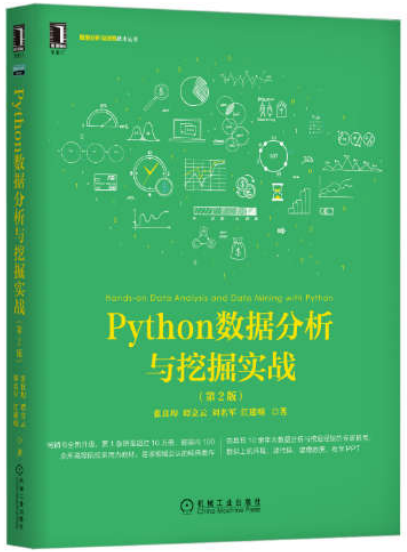
大数据分析

大数据概述

课程内容

- 大数据概述
- 数据预处理
- 分类分析
- 聚类分析
- 关联规则
- 离群点检测



Satellite Reveals New Views of China

VIIRS light data
by NASA

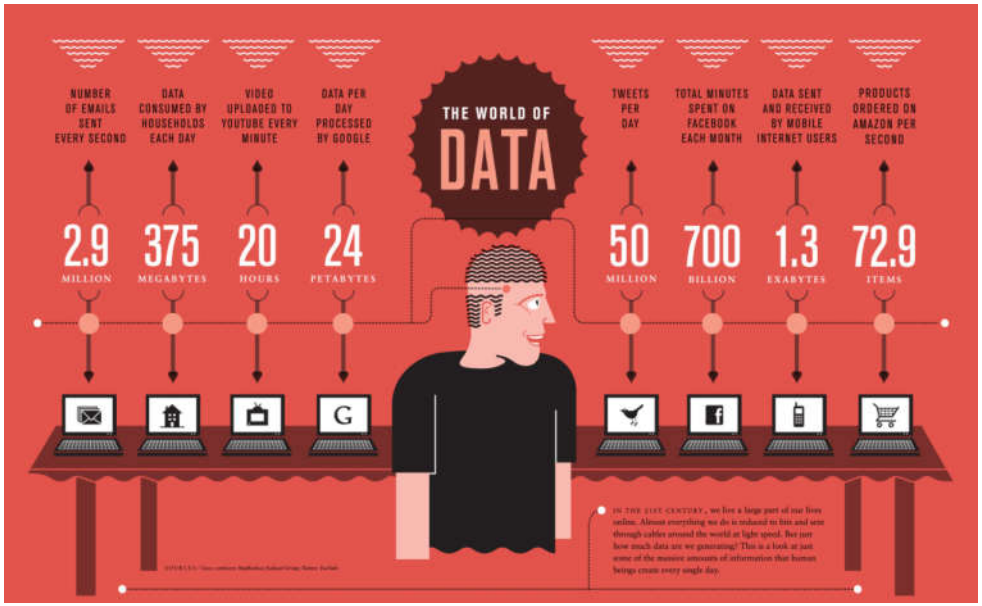
Mobile（摩拜单车）


mobike 摩拜单车
351192
总骑行距离（公里）
42563
总减排量（千克）
1917/2843
被骑行 / 总运营车辆
153/5163
实时 / 总骑行次数
昨天
今天
骑行次数
1500
1000
500
0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
性别比例
年龄比例
骑行比例
长宁区
徐汇区
闵行区
东华大学
上海植物园
华东理工大学
三林镇
华泾镇
乔里康
北殷家宅
复旦大学
附属儿科医院
S20
S2

大数据时代的背景

半个世纪以来，随着计算机技术全面融入社会生活，信息爆炸已经积累到了一个开始引发变革的程度。它不仅使世界充斥着比以往更多的信息，而且其增长速度也在加快。

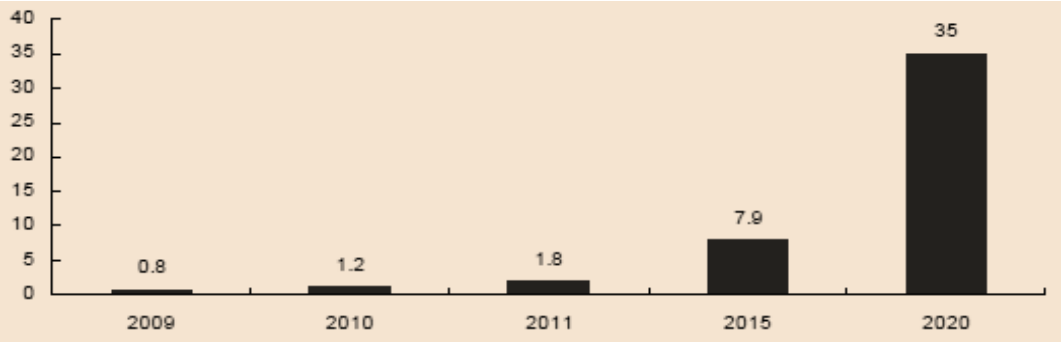
互联网（社交、搜索、电商）、移动互联网（微博）、物联网（传感器，智慧地球）、车联网、GPS、医学影像、安全监控、金融（银行、股市、保险）、电信（通话、短信）都在疯狂产生着数据。



- 全球每秒钟发送 **2.9 百万封**电子邮件，一分钟读一篇的话，足够一个人昼夜不息的读5.5 年…
- 每天会有 **2.88 万个小时**的视频上传到Youtube，足够一个人昼夜不息的观看3.3 年…
- 推特上每天发布 **5 千万条**消息，假设10 秒钟浏览一条信息，这些消息足够一个人昼夜不息的浏览16 年…
- 每天亚马逊上将产生 **6.3 百万笔**订单…
- 每个月网民在Facebook 上要花费**7 千亿分钟**，被移动互联网使用者发送和接收的数据高达**1.3EB**…
- Google 上每天需要处理**24PB** 的数据…

大数据时代正在来临...

数据量增加

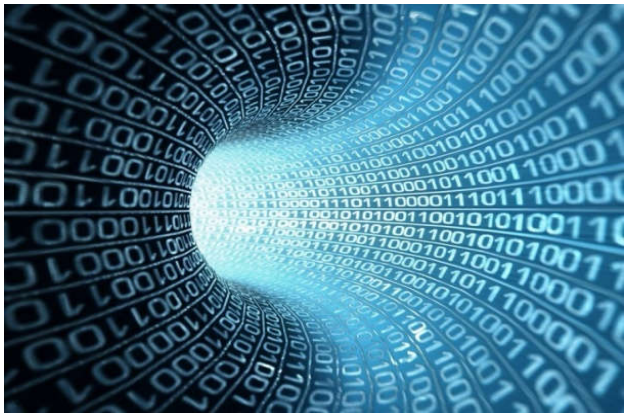


根据IDC 监测，人类产生的数据量正在呈指数级增长，大约每两年翻一番，这个速度在2020年之前会继续保持下去。这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量。

TB ⇨ PB ⇨ EB ⇨ ZB

数据结构日趋复杂

大量新数据源的出现则导致了非结构化、半结构化数据爆发式的增长



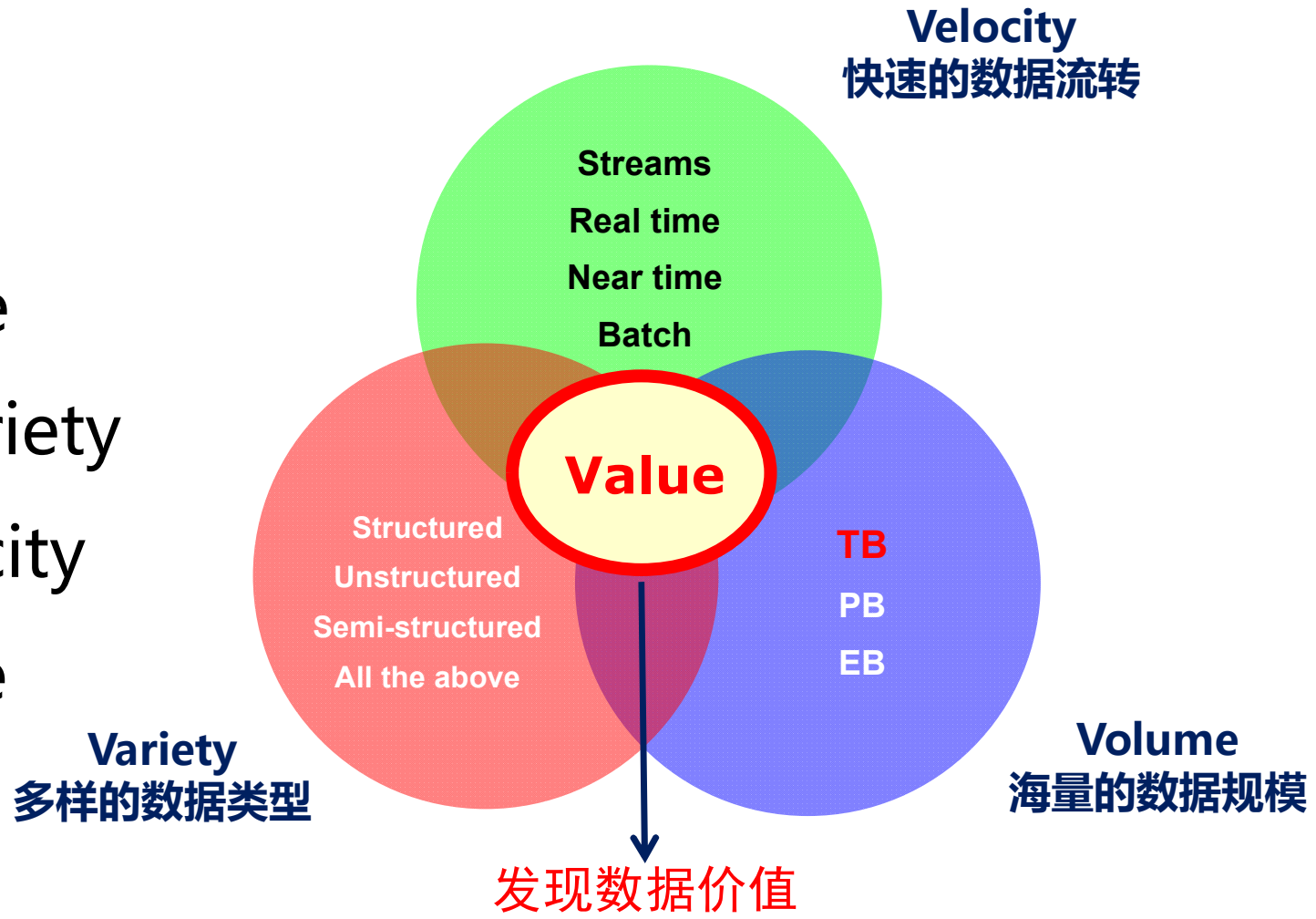
大数据的定义

- **Big data** is a term used to refer to data sets that are too large or complex for traditional data-processing application software to adequately deal with.

-http://en.Wikipedia.org/wiki/Big_data

大数据特点

- 数据量大Volume
- 数据类型繁多Variety
- 处理速度快Velocity
- 价值密度低Value



数据量大VOLUME

- 大数据摩尔定律：人类社会产生的数据一直都在以每年50%的速度增长，也就是说，每两年就增加一倍。
- 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
- 预计到2020年，全球将总共拥有35ZB的数据量，相较于2010年，数据量将增长近30倍

1Byte = 8 Bit
1KB = 1,024 Bytes
1MB = 1,024 KB = 1,048,576 Bytes
1GB = 1,024 MB = 1,048,576 KB = 1,073,741,824 Bytes
1TB = 1,024 GB = 1,048,576 MB = 1,099,511,627,776 Bytes
1PB = 1,024 TB = 1,048,576 GB =1,125,899,906,842,624 Bytes
1EB = 1,024 PB = 1,048,576 TB = 1,152,921,504,606,846,976 Bytes
1ZB = 1,024 EB = 1,180,591,620,717,411,303,424 Bytes
1YB = 1,024 ZB = 1,208,925,819,614,629,174,706,176 Bytes

数据类型繁多VARIETY

- 结构化数据：

- 半结构化数据：

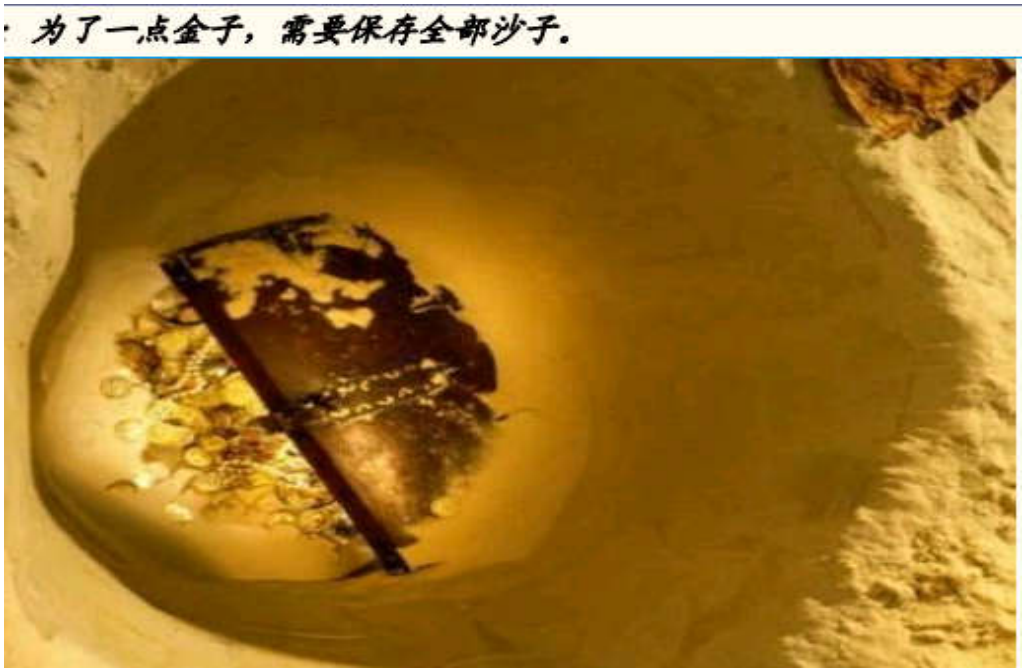
- 非结构化数据： 文本、音频、视频 ...

处理速度快VELOCITY

- 数据处理和分析的速度通常要达到秒级响应，否则处理的结果就是过时和无效的。
- 实时处理的要求，是区别大数据应用和传统数据库技术的关键差别之一。

价值密度低VALUE

- 挖掘大数据的价值类似沙里淘金,从海量数据中挖掘稀疏但珍贵的信息.



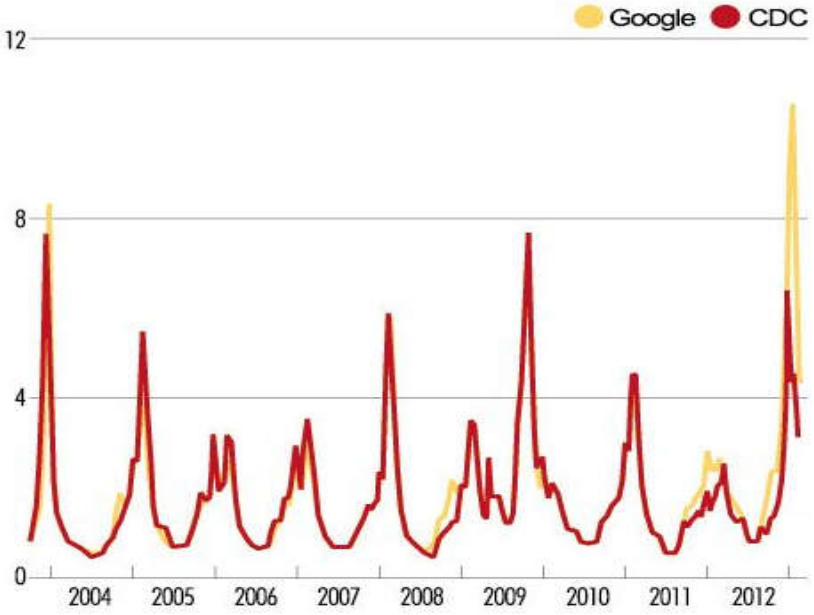
大数据应用-互联网


您最近查看的商品和相关推荐
根据您的浏览历史记录推荐商品
第 1 页，共 10 页 第一页
beyerdynamic...
(134)
¥ 980.00
LEGO 乐高 创意百变组...
(98)
¥ 212.00
JVC 杰伟世 头戴耳机 S400
(70)
¥ 178.23
Sennheiser...
(209)
¥ 299.00
Doc d'O 迪克步...
(27)
¥ 118.00
基于你的个人音乐库
今日推荐歌单
基于歌曲 Space Oddi...
VH1 Storytellers
David Bowie
基于歌曲 Franz Liszt...
感動のショパン・ラ...
李云迪
基于歌曲 Paganini: C...
The Master Violinist...
Jascha Heifetz
基于歌曲 Romance...
Jewel Box
Govi

大数据应用-生物医学

从谷歌流感趋势看大数据的应用价值

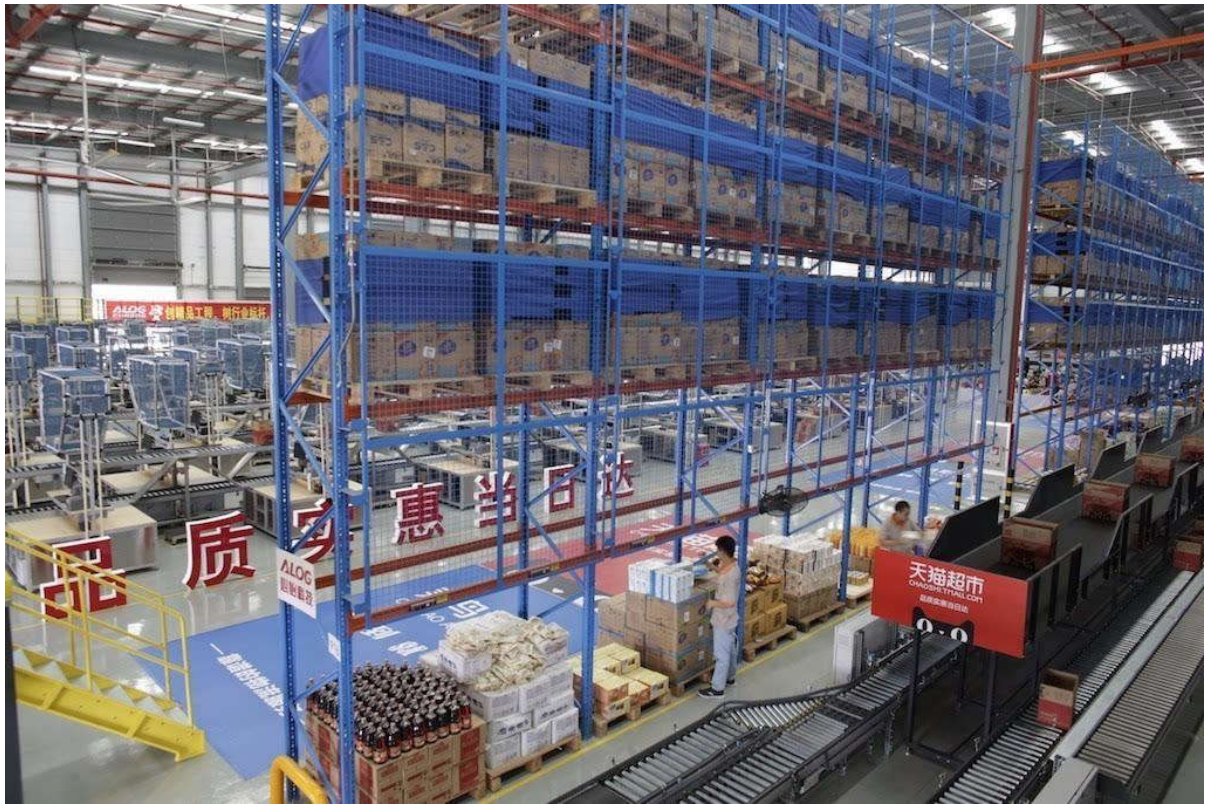
“谷歌流感趋势”，通过跟踪搜索词相关数据来判断全美地区的流感情况



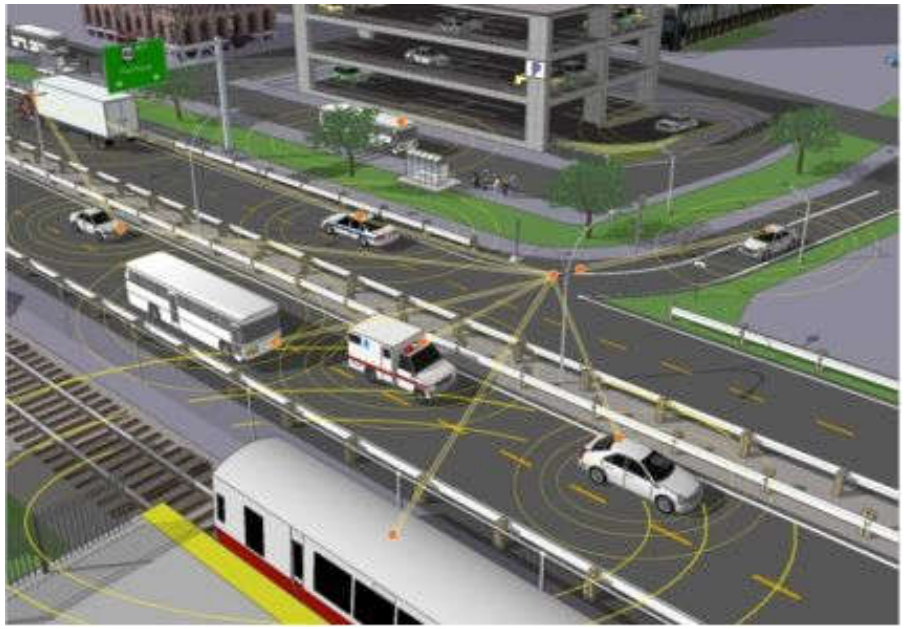
大数据应用-零售行业

大数据应用-智慧物流


ZTO
申通
申通
STO申通快递
圆通
YTO
银泰集团
阿里巴巴
SF
顺丰
韵达
YUNDA
韵达快运
富春集团
复星集团

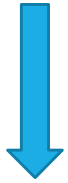

天猫超市

大数据应用...



大数据分析

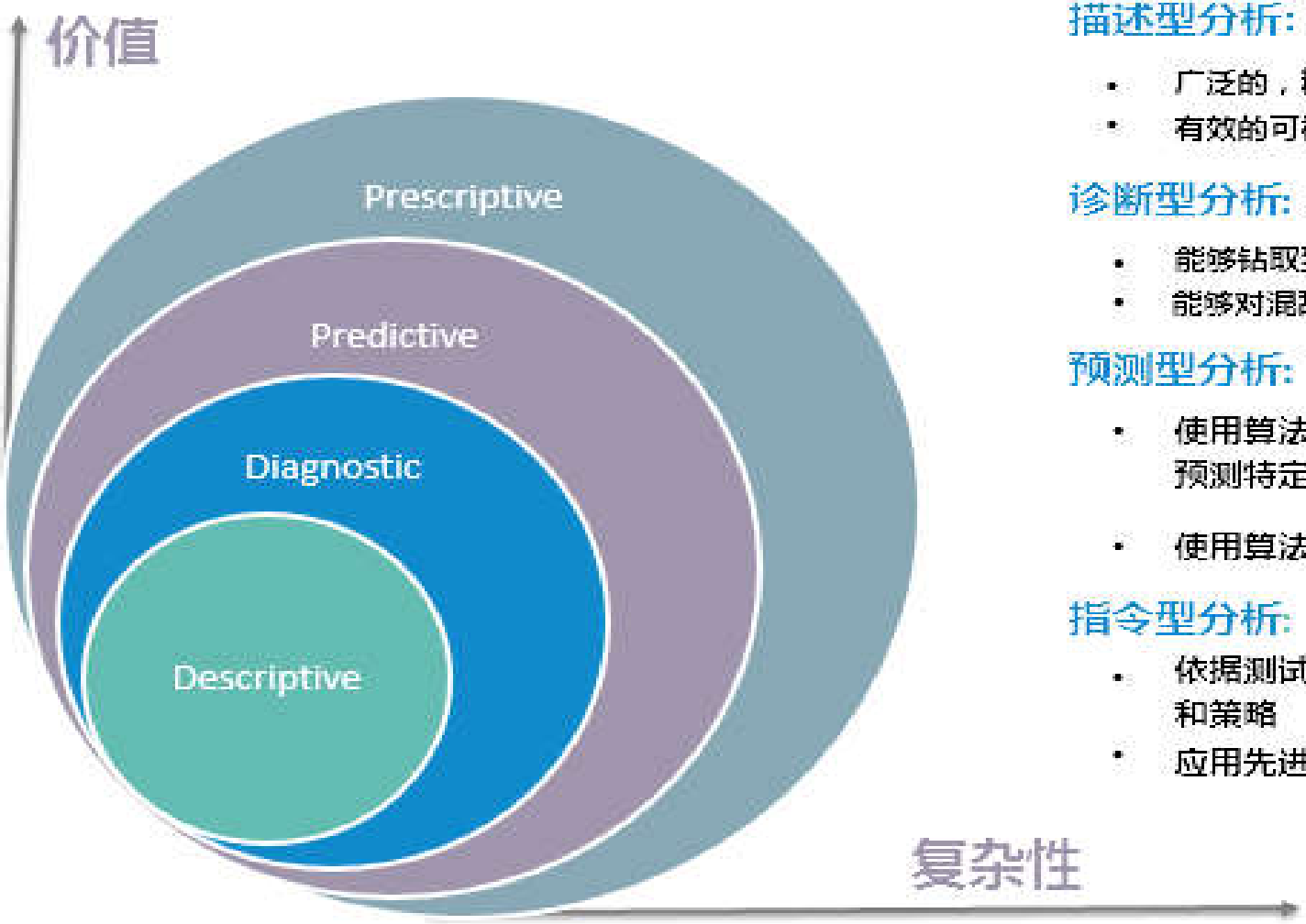
- 数据分析是指用适当的统计分析方法对收集来的数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。
- 数据分析的目的是把隐藏在一大批看来杂乱无章的数据中的信息集中和提炼出来，从而找出所研究对象的内在规律。



big data

大数据分析是指对规模巨大的数据进行分析。

大数据分析的分类



数据会告诉我们什么？

描述型分析：发生了什么？

- 广泛的，精确的实时数据
- 有效的可视化

诊断型分析：为什么会发生？

- 能够钻取到数据的核心
- 能够对混乱的信息进行分离

预测型分析：可能发生什么？

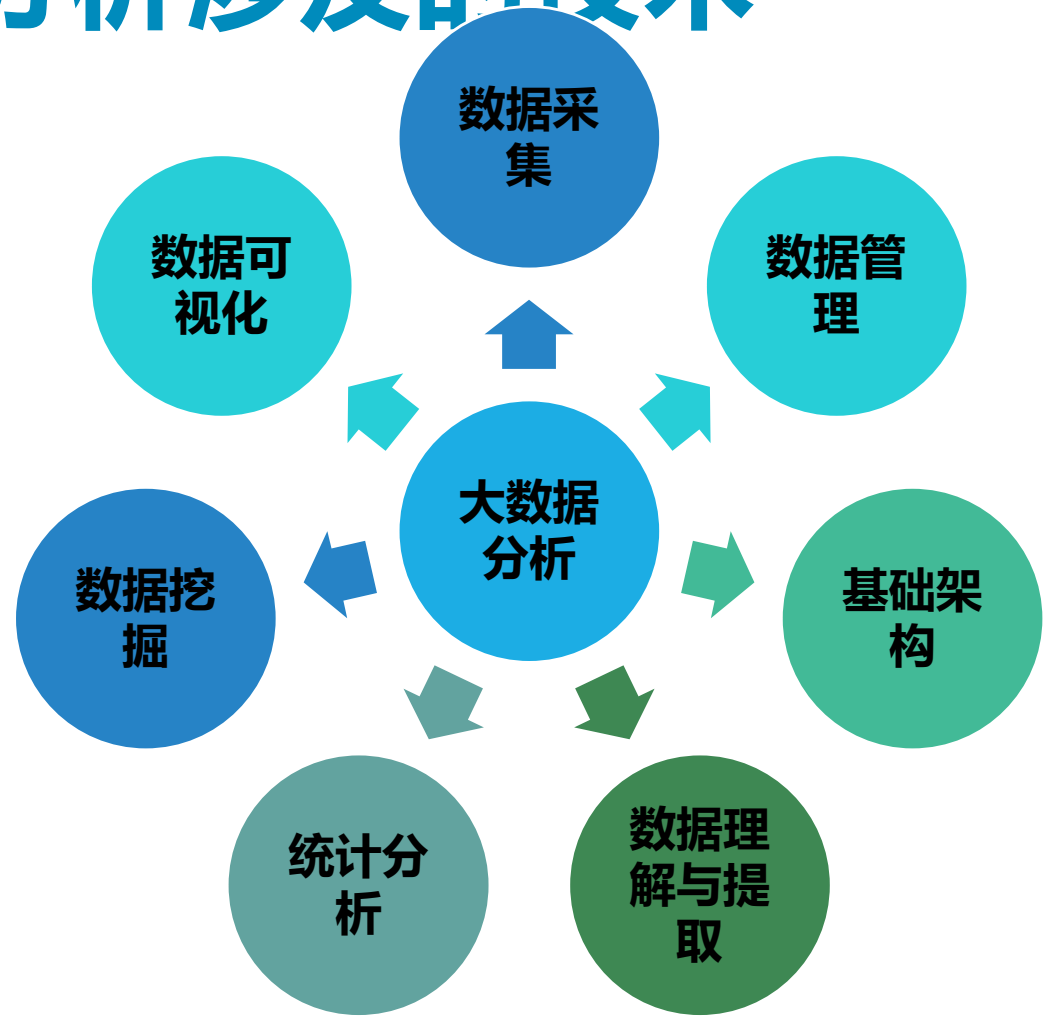
- 使用算法确保历史模型能够用于预测特定的结果
- 使用算法和技术确保自动生成决定

指令型分析：应该采取什么措施？

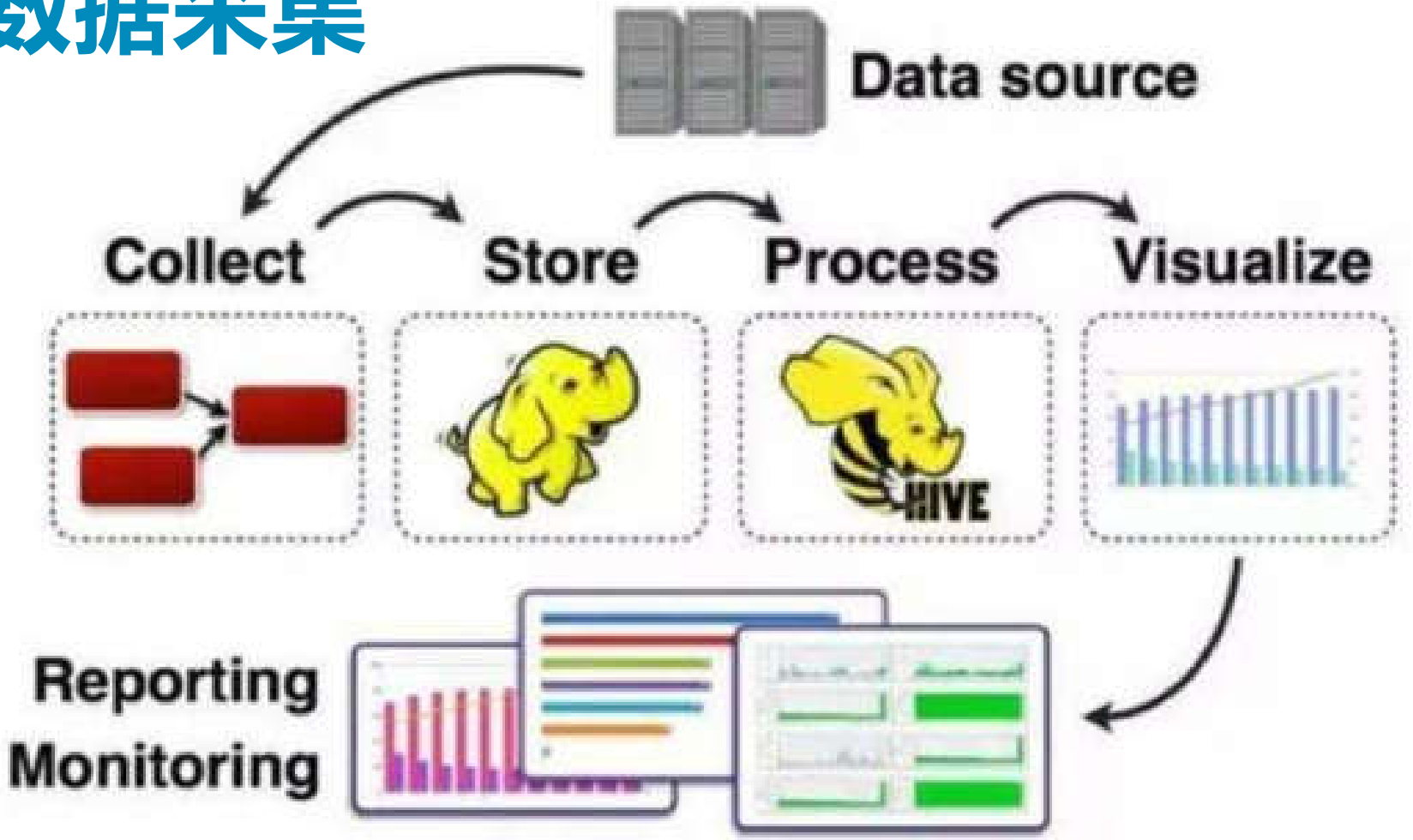
- 依据测试结果来选定最佳的行为和策略
- 应用先进的分析技术帮助做出决策

大数据分析流程

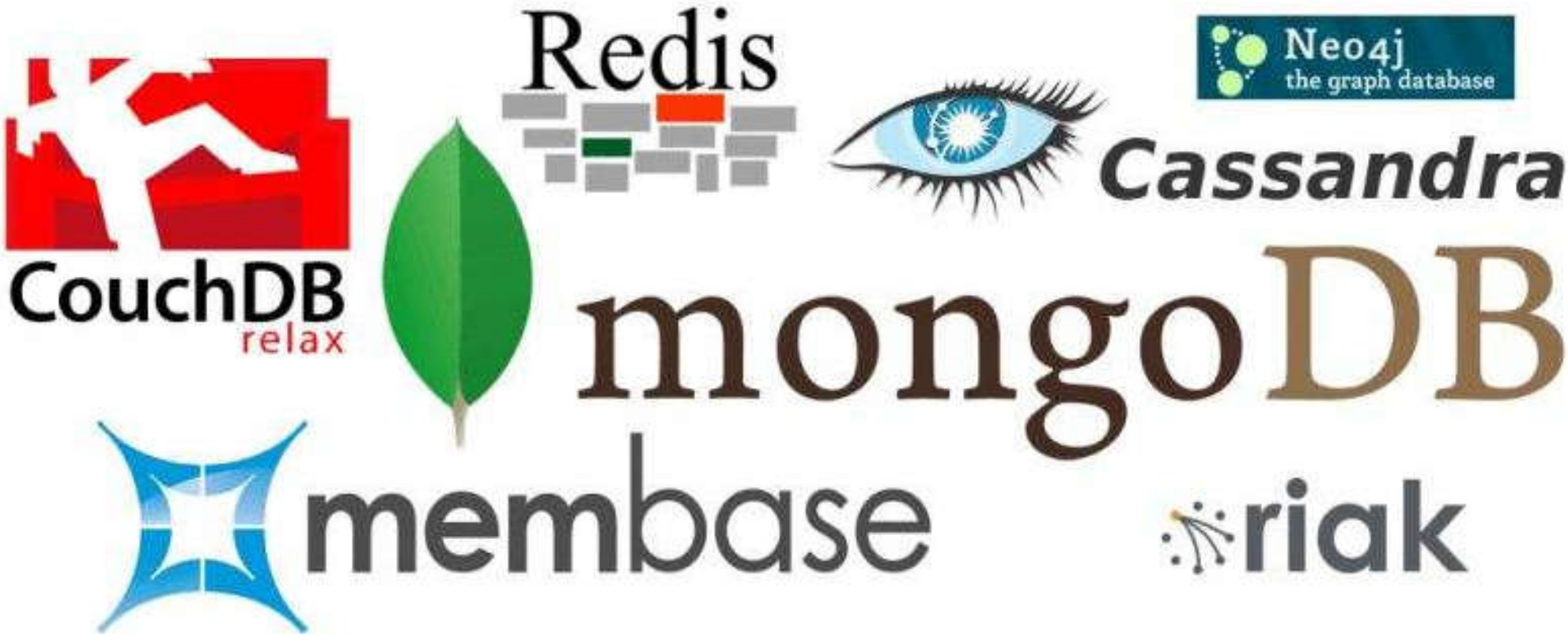
大数据分析涉及的技术



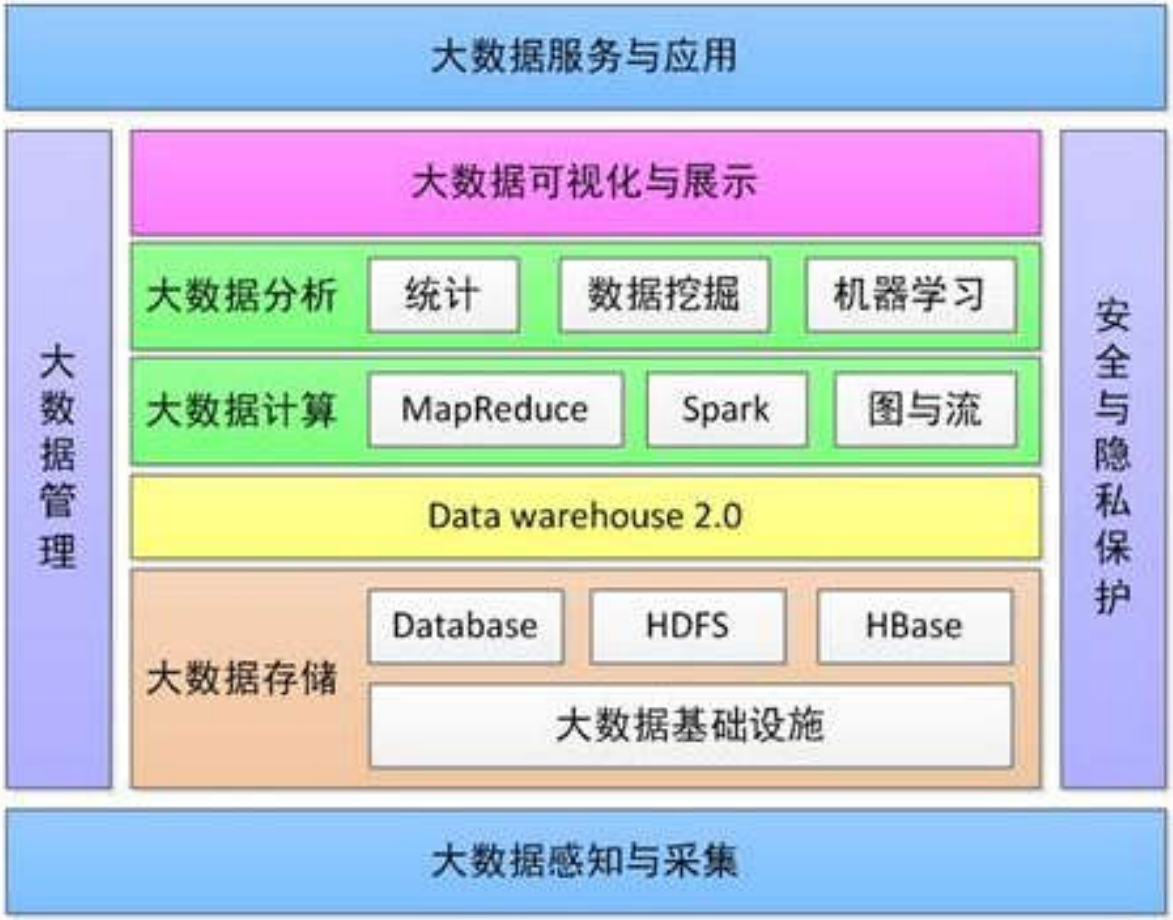
数据采集


Data source
Collect
Store
Process
Visualize
HIVE
Reporting
Monitoring

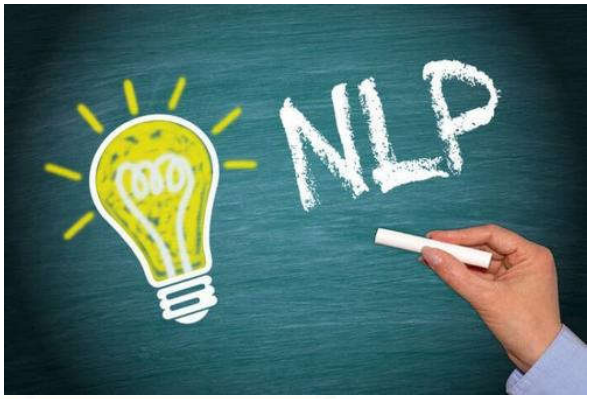
数据管理


Redis
Neo4j
the graph database
Cassandra
CouchDB
relax
mongoDB
membase
riak

基础架构


大数据服务与应用
大数据管理
大数据可视化与展示
大数据分析
统计
数据挖掘
机器学习
大数据计算
MapReduce
Spark
图与流
Data warehouse 2.0
大数据存储
Database
HDFS
HBase
大数据基础设施
安全与隐私保护
大数据感知与采集

数据理解与提取



基础技术

- 词法分析
- 句法分析
- 实体识别
- 语义分析
- 篇章分析
- 语言模型

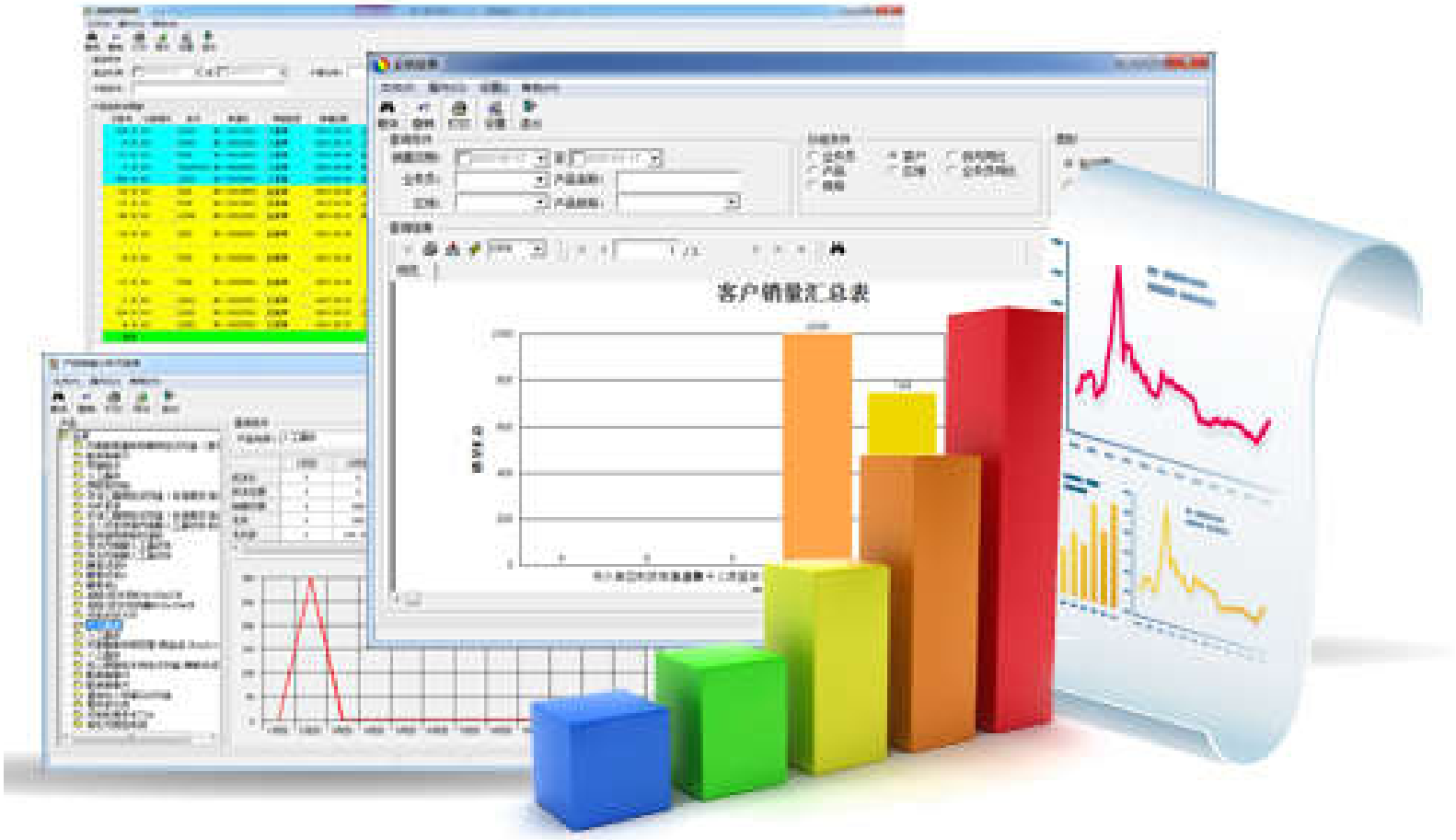
核心技术

- 机器翻译
- 自动问答
- 情感分析
- 信息抽取
- 文本摘要
- 文本蕴涵

应用

- 智能客服
- 搜索引擎
- 个人助理
- 推荐系统
- 舆情分析
- 知识图谱

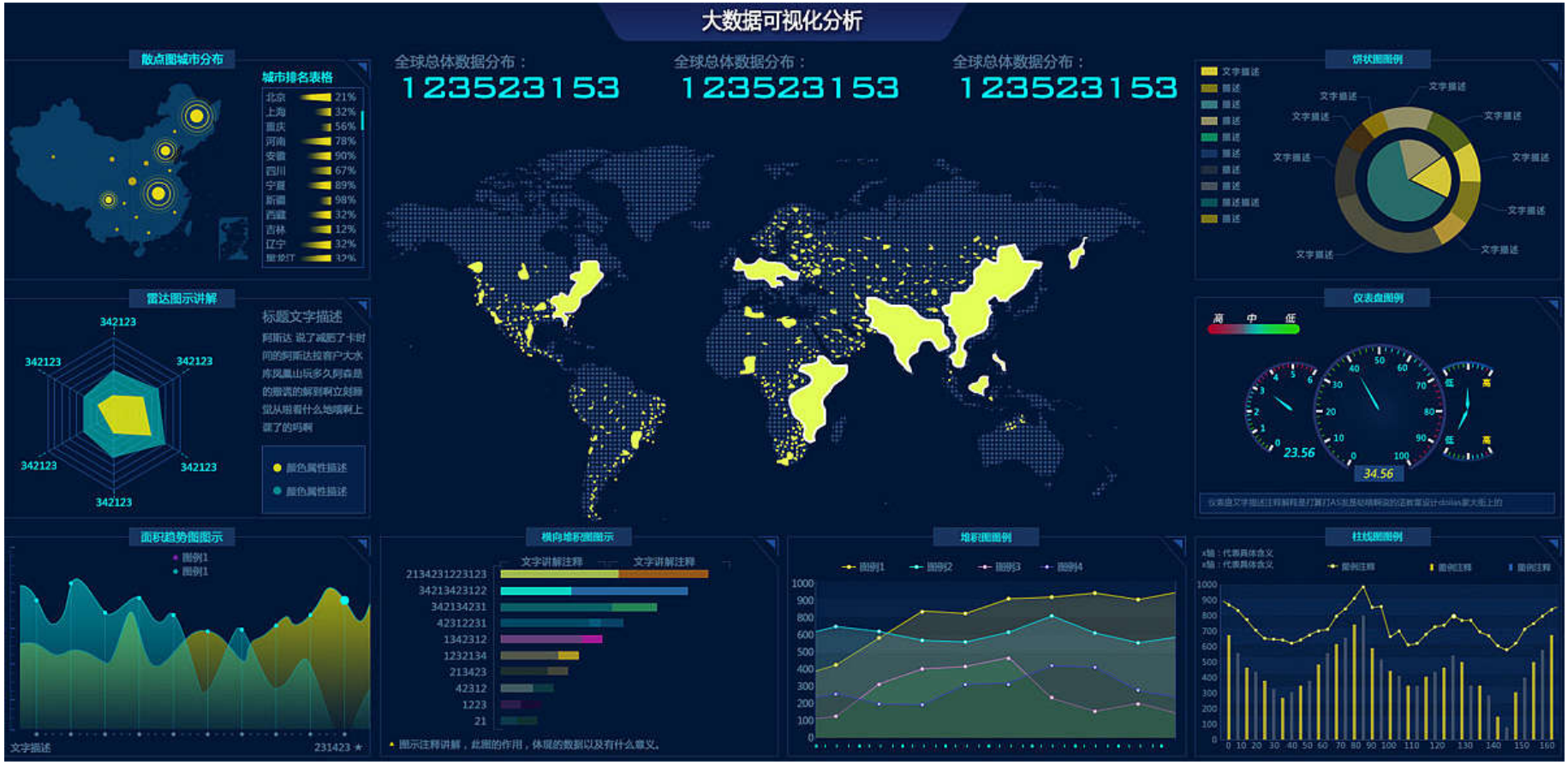
统计分析



数据挖掘


Data
Business problem
What you know
挖掘工具
Insight
Better data mining results

数据可视化


大数据可视化分析
散点图城市分布
城市排名表格
全球总体数据分布：
123523153
全球总体数据分布：
123523153
全球总体数据分布：
123523153
饼状图图例
雷达图示讲解
342123
342123
342123
342123
342123
标题文字描述
面积趋势图图示
横向堆积图图示
堆积图图例
柱线图图例
仪表盘图例
高
中
低
23.56
34.56

餐饮行业实例

- 国内某餐饮连锁有限公司（简称T餐饮）主要经营粤菜，兼顾湘菜、川菜、中餐等综合菜系。至今已经发展成为在国内具有一定知名度的大型餐饮连锁企业。
- 近年来餐饮行业面临较为复杂的市场环境，与其他行业一样餐饮企业都遇到了原材料成本升高、人力成本升高、房租成本升高等问题，这也使得整个行业的利润率急剧下降。如何在保持产品质量同时提高企业效率，成为了T餐饮急需面对的问题。从2000年开始，T餐饮通过加强信息化管理来提高效率，目前已上线的管理系统包括：

1. 客户关系管理系统
2. 前厅管理系统
3. 后厨管理系统
4. 财务管理系统
5. 物资管理系统

餐饮行业数据挖掘

各类菜品销量、成本单价、会员消费、促销活动等内部数据

天气、节假日、竞争对手以及周边商业氛围等外部数据

菜品智能推荐、促销效果分析、客户价值分析、新店选点优化、热销/滞销菜品分析和销量趋势预测

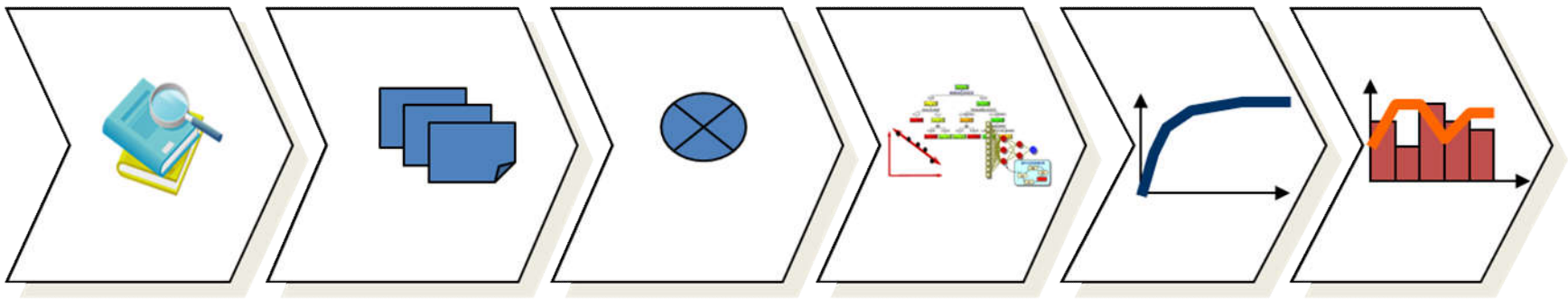
餐饮行业数据挖掘建模过程

餐饮、前厅、后厨等部门访谈

原材采购、菜品销量客户及外部数据收集

菜品智能推荐、新店选址优化、客户价值分析和销量预测等模型构建

提供菜品智能推荐、新店选址优化销量预测等服务



目标定义
- 任务理解
- 指标确定

数据采集
- 建模抽样
- 质量把控
- 实时采集

数据整理
- 数据探索
- 数据清洗
- 数据变换

构建模型
- 模式发现
- 构建模型
- 验证模型

模型评价
- 设定评价标准
- 多模型对比
- 模型优化

模型发布
- 模型部署
- 模型重构

第1步：定义挖掘目标

1.实现动态菜品智能推荐，帮助顾客快速发现自己感兴趣的菜品，同时确保推荐给顾客的菜品也是餐饮企业所期望的，实现餐饮消费者和餐饮企业的双赢；

2.对餐饮客户进行细分，了解不同客户的贡献度和消费特征，分析哪些客户是最有价值的，哪些是最需要关注的，对不同价值的客户采取不同的营销策略，将有限的资源投放到最有价值的客户身上，实现精准化营销；

3.基于菜品历史销售情况，综合考虑节假日、气候和竞争对手等影响因素，对菜品销量进行趋势预测，方便餐饮企业准备原材料；

4.基于餐饮大数据，优化新店选址，并对新店位置的潜在顾客口味偏好进行分析，以便及时进行菜式调整。

第2步：数据取样

1.餐饮企业信息：名称、位置、规模、联系方式；部门、人员、角色等；

2.餐饮客户信息：姓名、联系方式、消费时间、消费金额等；

3.餐饮企业菜品信息：菜品名称、菜品单价、菜品成本、所属部门等；

4.菜品销量数据：菜品名称、销售日期、销售金额、销售份数；

5.原材料供应商资料及商品数据：供应商姓名、联系方式、商品名称；客户评价信息；

6.促销活动数据：促销日期、促销内容、促销描述；

7.外部数据，如天气、节假日、竞争对手以及周边商业氛围等数据。

第3步：数据预处理

- 针对采集的餐饮数据，数据预处理主要包括：数据筛选、数据变量转换、缺失值处理、数据标准化、主成分分析、属性选择、数据规约等。

第4步：挖掘建模

- 针对餐饮行业的数据挖掘应用，挖掘建模主要包括基于关联规则算法的动态菜品智能推荐、基于聚类算法的餐饮客户价值分析、基于分类与预测算法的菜品销量预测、基于整体优化的新店选址。