



互联网+

大数据思想、技术及应用



个人简介

曹正凤



统计学 博士

人大经济论坛大数据中心 总工程师

人大经济论坛 hadoop大数据分析培训负责人

博士论文：随机森林算法优化研究

出版专著：从零进阶！数据分析的统计基础

译著：智能大数据SMART准则：数据分析方法、案例和行动纲领

EI核心论文：Parameter Settings of Genetic Algorithm

Based on Multi-Factor Analysis of Variance

在研国家社科基金青年项目：基于大数据整合的空气质量测度方法

从零进阶！数据分析的统计学基础

◆2月份，当当新书热卖榜排名第2名

◆2015年2月上架，到9月底已经销售了过万册



商品评价

96% 好评度

好评(96%) 中评(2%) 差评(2%)

买家印象：
专业必备(30) 正版(29) 很实用(15) 脉络清晰(14)
理论基础(14) 帮助很大(10) 性价比高(7) 查阅方便(6)
实例经典(2) 科技前沿(2)

您可对已购商品进行评价
发评价拿京豆
前五名可获赠京东豆(规则)

全部评价(123) 好评(118) 中评(2) 差评(3) 有图片的评价(3)

评价心得	顾客满意度	购买信息	评论者
好书 推荐 我喜欢 2015-05-09 18:21 回复(0) 赞(0)	★★★★★	-	J***2 钻石会员 北京 2015-04-30 23:42 购买 来自京东iPhone客户端
对于一个从事很多年的数据分析师来说，此书欲可以再打基础 2015-05-08 19:50 专业必备 查阅方便 很实用 回复(0) 赞(0)	★★★★★	-	飞***京 金牌会员 上海 2015-05-08 08:15 购买
送货速度太给力啦，看了一章，不错 2015-05-06 20:11 回复(0) 赞(0)	★★★★★	-	JL_137796au 铜牌会员 2015-05-05 20:18 购买 来自京东iPad客户端

Broadview®
www.broadview.com.cn

WILEY

CDA数据分析师系列丛书

智能大数据 SMART准则

数据分析方法、案例和行动纲领

【美】Bernard Marr 著

秦磊 曹正凤 译

人大经济论坛 对外经济贸易大学大数据与风险管理研究中心 审校

台北医学大学管理学院暨大数据研究中心 谢邦昌教授 倾情作序

BIG DATA

Using SMART Big Data, Analytics and Metrics
To Make Better Decisions and Improve Performance

中国工信出版集团

电子工业出版社
WILEY

“数据和分析推动我们工作的方方面面。本书是数据领域的必看指南，真好书！”

Henrik von Scheel
谷歌咨询委员会成员

Broadview® 博文视点·IT出版旗舰品牌
www.broadview.com.cn 技术凝聚实力·专业创新出版

智能大数据 SMART 准则

数据分析方法、案例和行动纲领

拥有SMART数据分析方法，你可以将大数据愿景变为实现。

手头太多混杂的数据，很多人想知道它代表了什么？以及如何用好它？但你真正唯一要关心的事情是如何使用大数据得到清晰的、真实的商业结果，并将它作为提高绩效最主要的手段。

本书将告诉你如何实施领先公司已经使用的相同做法，以获得新的盈利能力。从清晰的解释和无数成功的案例中，你将学会如何成功地利用大或大数据使用SMART模型预测未来。

S: 制定智能战略

M: 度量指标和数据

A: 运用数据分析技术

R: 展示数据分析结果

T: 改变商业模式

大数据是智能革命的核心。大数据背后的基本思想是，人类一切行为都会留下数字痕迹（或数据），我们（或他人）可以对其加以利用，变得更加智慧。掌握数量日益增加的数据并利用技术能力将其转化成具有商业价值的想法，是推动新世界的主要力量。无疑大数据正在改变世界，我们的居住、择偶、治疗癌症、科研、提升绩效、管理城市、治理国家和管理企业的方式都因此而发生完全改变。

WILEY



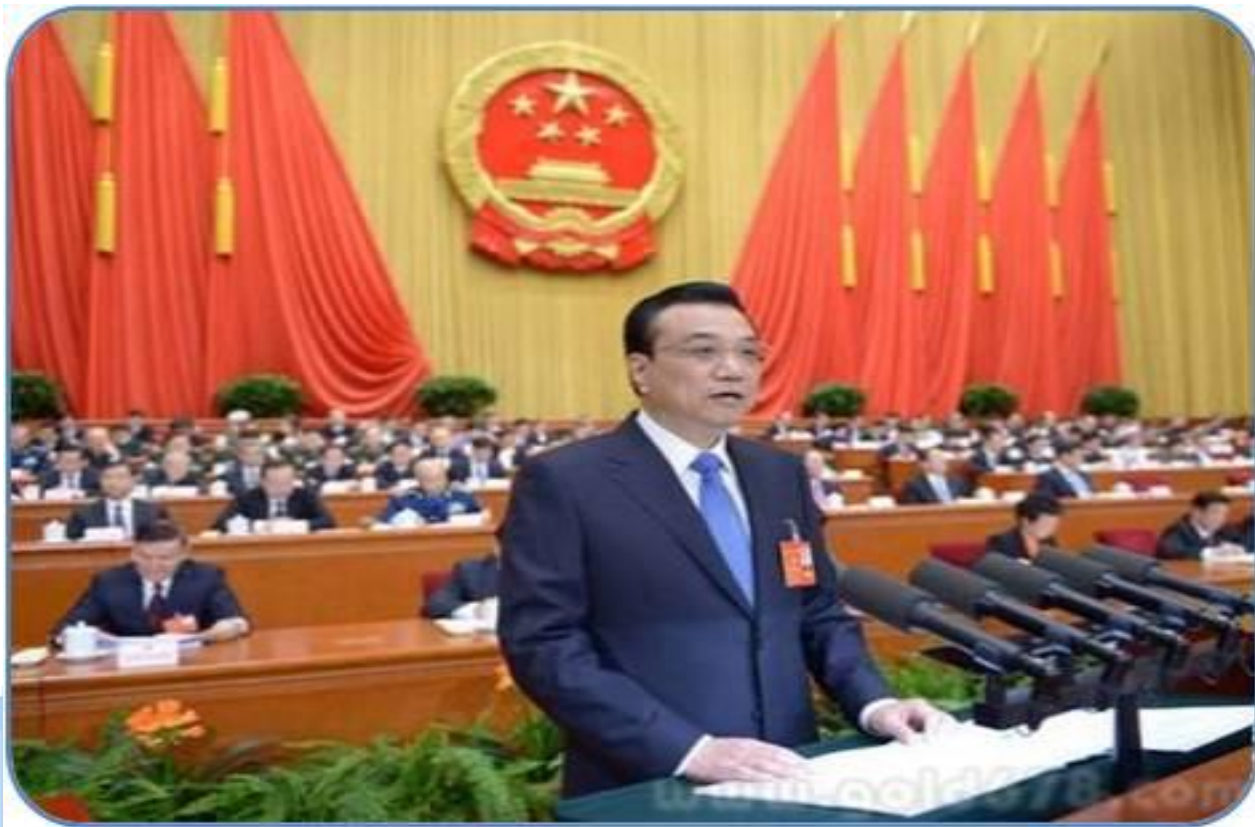
策划编辑：张慧敏
责任编辑：李利健
封面设计：李玲

上架建议：大数据



定价：49.00元

大数据已经上升为国家战略



2015/10/19

今年3月李克强总理的政府工作报告

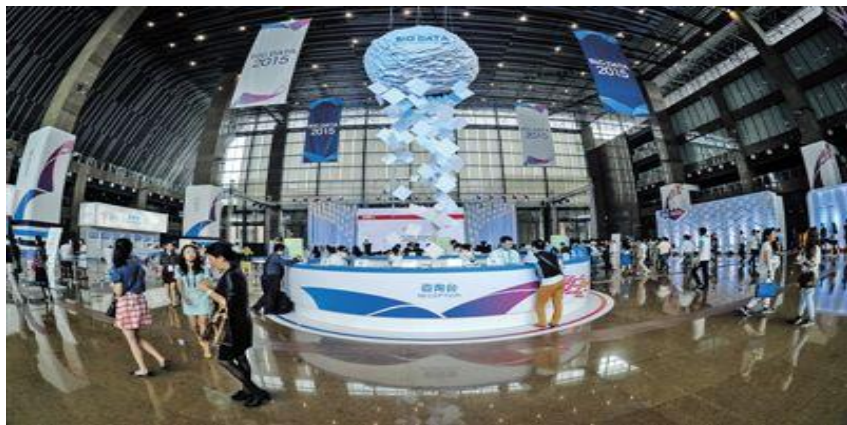
4月成立中国大数据交易所

- 2015年4月14日，全国首个大数据交易所——贵阳大数据交易所正式挂牌运营并完成首批大数据交易。
- 贵阳大数据交易所预计在未来3年至5年，交易所日交易额将突破100亿元，预计将诞生一个万亿元级别的交易市场。



5月召开贵阳国际大数据产业博览会

- 5月26-29日，以“互联网+时代的数据安全与发展”为主题的2015国际大数据产业博览会暨全球大数据时代贵阳峰会在贵阳举行。
- 马云、马化腾、阿南德、郭台铭、许罗德、周鸿祎等行业巨头围绕“‘互联网+’时代的数据安全与发展”发表精彩演讲。



国务院总理李克强发来贺信

中华人民共和国国务院

贺 信

值此 2015 贵阳国际大数据产业博览会暨全球大数据时代贵阳峰会开幕之际，我谨代表中国政府表示热烈祝贺！

当今世界，新一轮科技和产业革命正在蓬勃兴起，数据是基础性资源，也是重要生产力。大数据与云计算、物联网等新技术相结合，正在迅猛并将日益深刻地改变人们生产生活方式，“互联网+”对提升产业乃至国家综合竞争力将发挥关键作用。

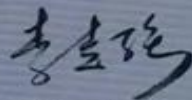
中国是人口大国和信息应用大国，拥有海量数据资源，发展大数据产业空间无限。中国正在研究制定“互联网+”行动计划，推动各行各业依托大数据创新商业模式，实现融合发展，推动提升政府科学决策和管理水平，用新的思路和工具解决交通、医疗、教育等公共问题，助力大众创业、万众创新，促进中国经济保持中高速增长，迈向中高端水平。

中华人民共和国国务院

互联网缩短了时空距离，大数据产业给不同国家和地区发展带来了机遇，相信大家围绕“‘互联网+’时代的数据安全与发展”这个主题交流互鉴，分享成果，深化合作，会进一步汇聚新动能，推动实现更高效、更绿色、更惠民的发展。

预祝峰会取得圆满成功！

中华人民共和国国务院总理



2015年5月17日

8月国务院印发大数据行动纲要

- 2015.8.31国务院《关于印发促进大数据发展行动纲要的通知》发布，**大数据已上升为国家战略**。

国务院关于印发促进大数据发展 行动纲要的通知

国发〔2015〕50号

各省、自治区、直辖市人民政府，国务院各部委、各直属机构：

现将《促进大数据发展行动纲要》印发给你们，请认真贯彻落实。

国务院

2015年8月31日



演讲提纲



大数据的概念与技术



大数据的思维变革

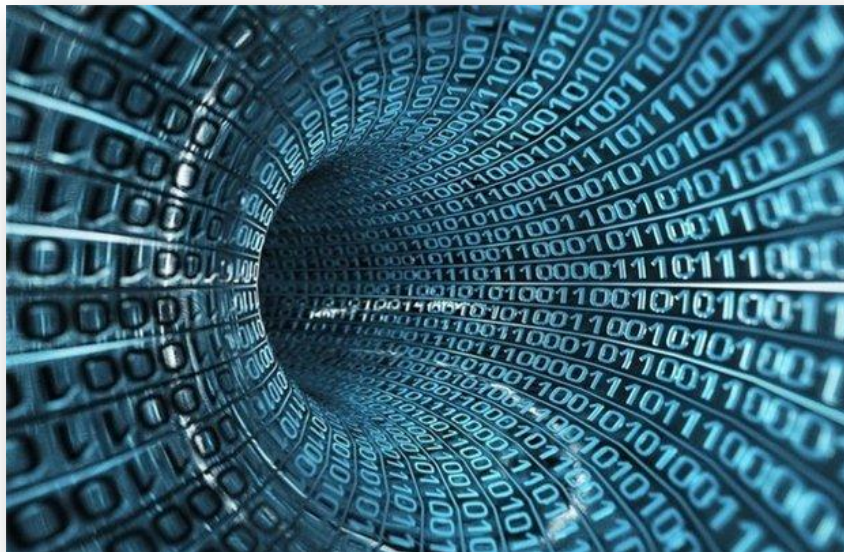


大数据在企业中的应用



“大数据”的诞生

2008年9月4日《自然》杂志社，推出的名为“大数据”的专刊，创造出了“大数据”这个概念。



“谷歌流感趋势”把大数据推上风口浪尖

搜索流感信息的人数

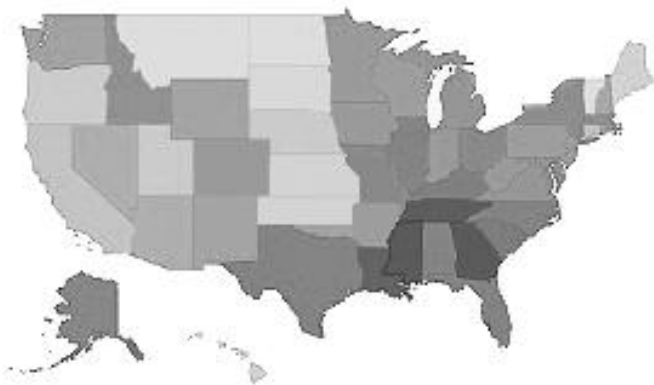
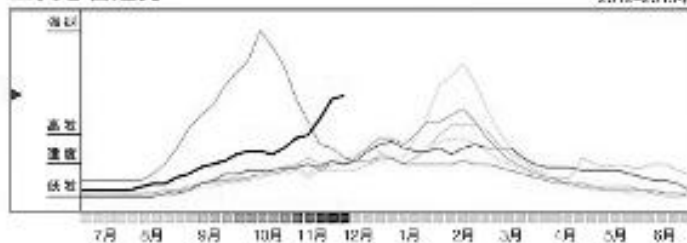
实际患病人数

流感爆发

- 美国疾病控制中心要在流感爆发**两周**后才知道
- 谷歌的大数据预测只需要**一天**

全美感冒趋势

2012-2013年

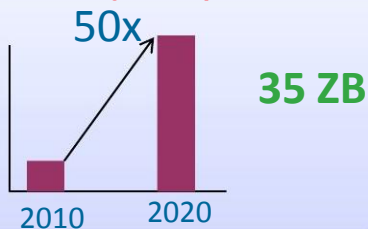


大数据的特征与趋势

PB是大数据的临界点

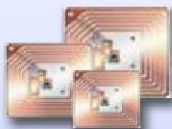
数据间要有**关联性**

Volume (大量)



能够反应不断且更快速到达的数据

Velocity (快速)



超过300亿

RFID 感应装置

整合性收集与分析更多元的数据

Variety (种类多)



全球80%

数据为非结构



建立大数据来源的不确定与不准确的数据

的可信性

VALUE

Veracity(真实性)

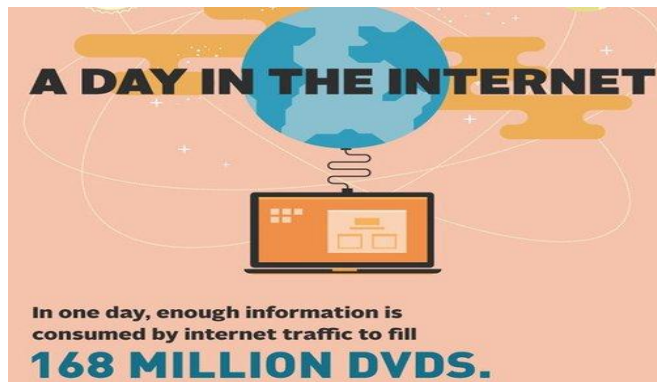
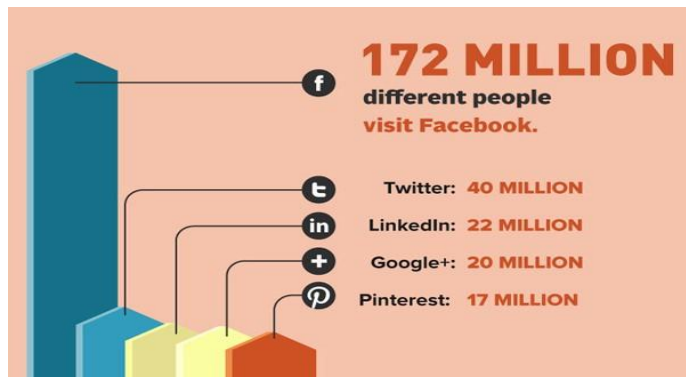
1/3 企业领导者不信任用来作为业决策

的信息真实性

关键 - 数据的可信性

数据来源: IBM

互联网每天产生的数据



• 每天有：

- 1.72亿人登陆Facebook
- 4000万人登陆Twitter
- 2200万人登陆LinkedIn
- 2000万人登陆Google+

互联网一天产生的内容足够刻满1.68亿张DVD光碟

百度每天的关键词搜索量**50亿**

大数据存放在哪？如何分析？

Hadoop是基于Google有关大数据的论文实现的开源项目，最初的框架由Doug Cutting在2005年提出，目前是由Apache 维护的开源项目。从初创到现在，Hadoop体系在7年中开发完成了一系列重要的子项目，已经形成了一个涵盖数据存储、管理和分析功能的较为完整的大数据生态系统，成为大数据存储与处理领域地位最重要、应用最广泛的开源框架。



核心组件

MapReduce

- Hadoop的分布式并行处理框架
- 实现对HDFS上海量数据的批量分析

HDFS

- Hadoop的一个分布式文件系统
- 高容错性，部署在低廉商业硬件
- 提供高吞吐量,适合批量处理

Hadoop是运行在大量通用计算单位上提供海量数据存储与并行计算的平台框架

- ❑ 基于x86集群水平可扩展
- ❑ 基于MapReduce的并行计算能力
- ❑ 设计规模：PB级的数据量，数千台计算节点

应用最广泛的大数据框架

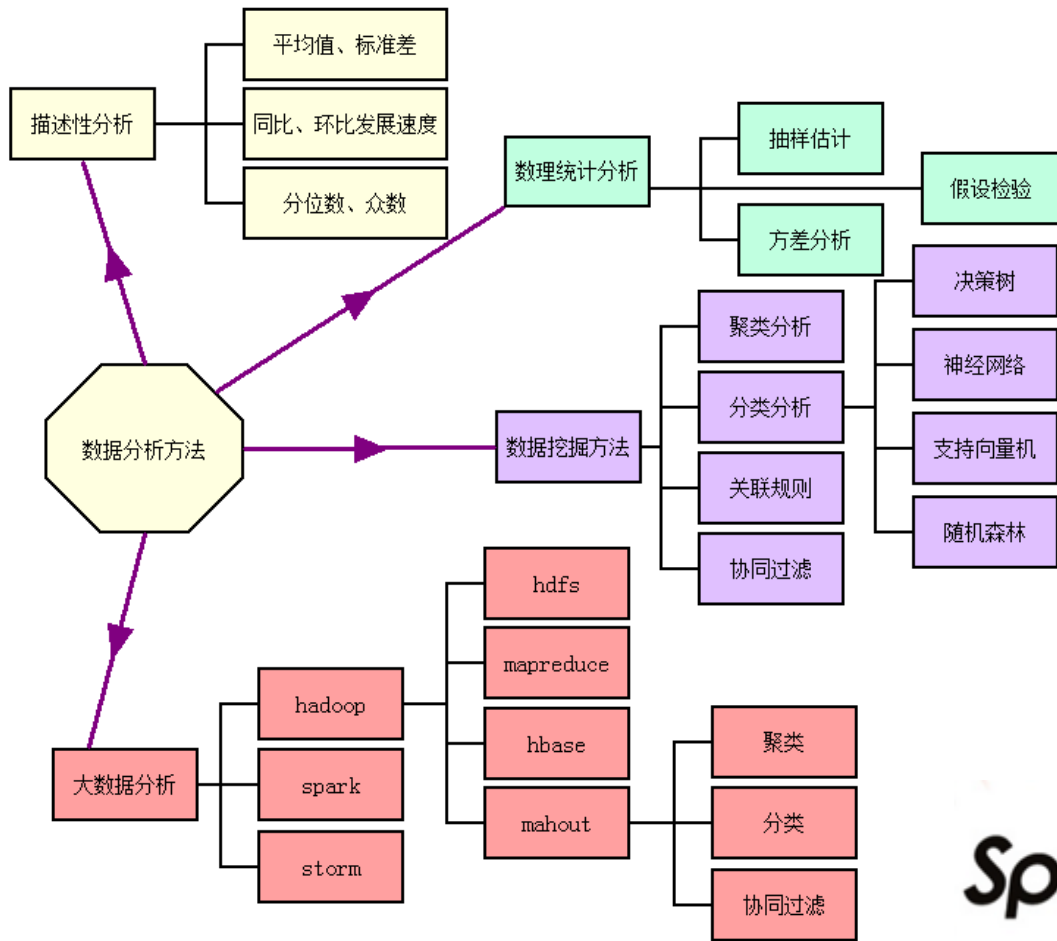


Hadoop大数据分析生态环境简介

Hadoop Ecosystem Map



大数据分析的位置



演讲提纲



大数据的概念与技术



大数据的思维变革



大数据在企业中的应用



“更多” ——不是随机样本，而是全体数据

当数据处理技术已经发生翻天覆地的变化时，在大数据时代进行抽样分析就像在汽车时代骑马一样。一切都改变了，我们需要的是所有的数据，“样本 = 总体”。

1. **小数据时代的随机采样，最少的数据获得最多的信息**
2. **全数据模式，样本 = 总体**
3. **大数据的简单算法比小数据的复杂算法更有效**

大数据时代的思维变革—更杂

“更杂”——不是精确性，而是混杂性

执迷于精确性是信息缺乏时代和模拟时代的产物。只有5%的数据是有框架且能适用于传统数据库的。如果不能接受混乱，剩下95%的非框架数据都无法被利用，只有接受不精确性，我们才能打开一扇从未涉足的世界的窗户。

- 允许**不**精确
- 纷繁的数据越多越好
- 混杂性，不是竭力避免，而是标准途径

大数据的性质——不是精确性，而是混杂性

目标：测量一个葡萄园的温度

广度：获得更广泛的数据而牺牲了精确性

1个温度测量仪--->精度

100个温度测量仪：数据可能会是错误的，可能会更加混乱，但众多的读数合起来就可以提供一个更加准确的结果。

深度：为了高频率而放弃了精确性

每分钟测量一次，测量结果按照时间有序排列。

每分钟测量十次甚至百次的话，不仅读数可能出错，连时间先后都可能搞混掉。

在很多情况下，与致力于避免错误相比，对错误的包容会带给我们更多好处。



“更好” ——不是因果关系，而是相关关系

知道“是什么”就够了，没必要知道“为什么”。
在大数据时代，我们不必非得知道现象背后的原因，
而是要让数据自己“发声”。

- **关联物，预测的关键**
- **“是什么”，而不是“为什么”**
- **改变，从操作方式开始**



大数据的概念及技术



大数据的思维变革



大数据在企业中的应用



大数据在企业中的应用之一

预测



大数据提升预测准确性

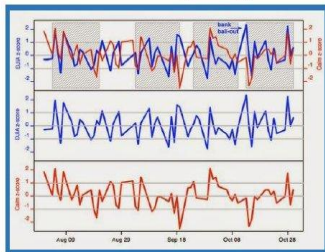
天氣



过去60年天气信息
820亿次分析
即时天气比对

成功预测
未来**40天**气象

股價



上亿条社群推荐/讨论
语意与情感分析

准确率达**87.6%**
15%投资报酬率

健康



和美国疾病控制及预防中心合作，以关键字
搜寻次数掌握流感

提前**1天**
掌握流感爆发关键

大數據整合與分析

世界杯大数据预测火了百度

	1/4决赛准确率	1/8决赛准确率	小组赛准确率
	100%	100%	58.33%
 Microsoft	100%	100%	56.25%
	100%	100%	37.5%
	75%	100%	/

百度如何做到的

数据来源

5年内全世界987支球队的3.7万场比赛数据

469家博彩公司的赔率数据

赛事预测市场的数据

通过爬虫等方法取得

团队实力

主场效应

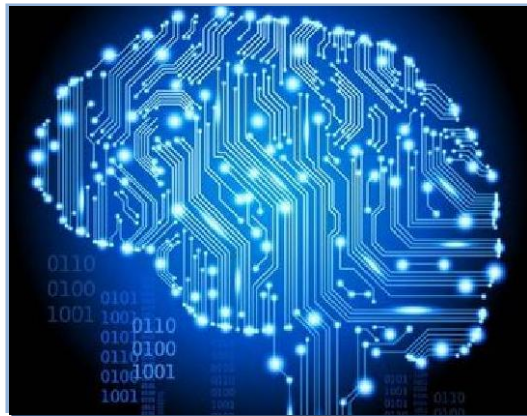
最近表现

大赛能力

博彩数据

多源异构数据

机器学习



做出预测结果

百度预测：<http://trends.baidu.com/>

Baidu 预测

联系

世界杯预测

巴西世界杯疯狂来袭，谁会成为最终的王者？
百度预测再现章鱼保罗的神话。

点击进入

淘汰赛
预测准确率

100%

2014
FIFA WORLD CUP
BRAZIL



经济指数预测



景点预测



疾病预测



城市预测



欧洲赛事预测



世界杯预测



高考预测



电影票房预测



预测开放平台

大数据在企业中的应用之二

营销：精准营销、整合营销、联合营销



电商巨头阿里大数据生态圈已经建立

数据分析为商业核心驱动力，打造以消费者为导向的盈利模式



单日RMB **600亿** 营业额
2014光棍节创下世界纪录

每分卖 **4.8万** 件商品
总商品数超过8亿件

超过 **5亿** 会员数
每日访客数为台湾人口三倍

传统制造企业**耐克公司**大数据战略

大数据采集

- 耐克凭借一种名为Nike+的新产品变身为大数据营销的创新公司。所谓**Nike+**，是一种以“**Nike**跑鞋或腕带+传感器”的产品，只要运动者穿着**Nike+**的跑鞋运动，iPod就可以存储并显示运动日期，时间、距离、热量消耗值等数据。用户上传数据到耐克社区，就能和同好分享讨论。
- 凭借运动者上传的数据，耐克公司已经成功建立了全球最大的运动网上社区，超过**1000万**活跃的用户，每天不停地上传数据，耐克借此与消费者建立前所未有的牢固关系。

展示



大数据能为Nike带来什么



精准
营销



改进
产品



联合
营销

Nike+：硬件、软件、社区的大平台

大数据带给Nike的是利润



尽管耐克的使命在卖出更多球鞋
但它还在跟你谈生活方式



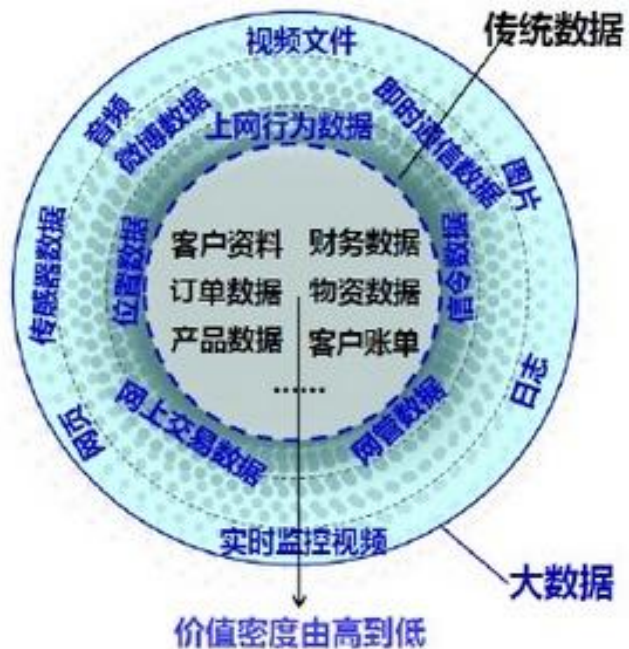
消费者与品牌的黏性

大数据在企业中的应用之三

中国移动大数据



中国移动的大数据

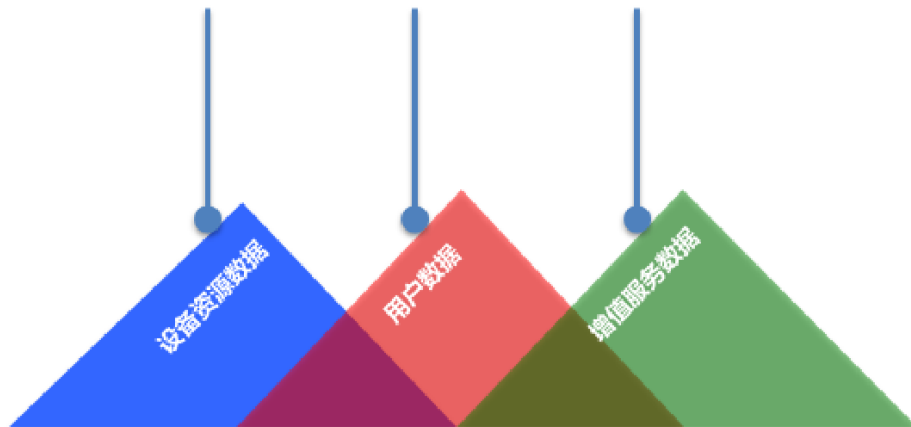


电信3类数据

主要为信道层面网络与运维数据，基本与用户无关

用户身份的账号数据、用户行为数据、信令数据等

流媒体内容数据、视频监控数据、网页数据等



内部的应用之精准营销

分析

- 根据历史使用过的终端和交往圈中人用的终端,分析用户的终端偏好和消费能力;

时机

- 根据终端的生命周期,合约机也有到期时间,确定换机时机;

推送

最后就是捕捉最近的特征事件然后通过短信、外呼、营业厅等渠道推送到用户手中。

中国移动的大数据对内部的应用



10086热线进行语义分析，
可以自动分析来话内容，
进行归类，

并识别其中的热点问题，
如果是网络、资费等可能
造成批量投诉的情况，还
可及时地预警。



分析话单和信令中用户的流量
在时间周期和位置特征方面
的分布

实现4G基站和WLAN热点的精
确选址

CBD白天配备多一些无线资源，
三里屯晚上配置多一些

中国移动的大数据对外部的应用

- 虽然大数据的外部性应用更加有趣，能发展新的商业模式，但是有数据所有权、隐私、体制等诸多因素，所以国内似乎目前只有看到电信在将固网的一些数据用来做RTB的互联网广告，除此之外看到的所有对外的商业应用基本都来自国际运营商。

对外部的应用

景点舒适指数预测

- 根据位置信令来分析景区用户数量，帮助旅游景区了解游客来源、分布等信息

客流量分析

- 帮助一些大的零售商分析顾客来源和各商铺、展位的人流情况。

北京市旅游局景点舒适度预报

北京旅游网 >> 北京市各景区游览人数及舒适度指数信息

景区当前游览舒适度指数

景区名称	指数
国家体育场	5
国家游泳中心	5
奥林匹克森林公园	5
颐和园	4
八达岭	4
十三陵	5
慕田峪长城	5
龙潭湖公园	5
中山公园	5
什刹海	4
北海公园	4
动物园	4
北京海洋馆	4
景山公园	5
陶然亭景区	5
北京欢乐谷	5
朝阳公园	5

故宫

数据更新时间：2015-07-06 15:45 (每15分钟自动更新)

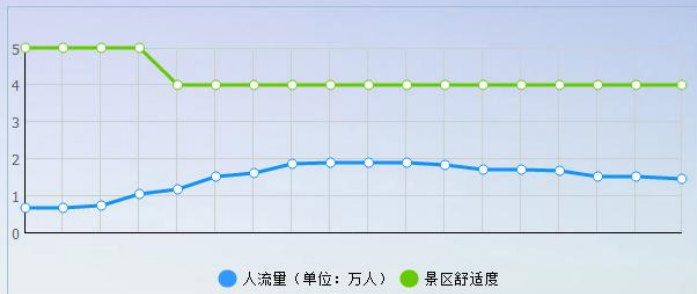
当前舒适度指数

4

较为舒适，适合参观游览。

当前客流量：1.46万人

趋势图：



一则笑话--读懂未来的大数据应用

- 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。



一则笑话--读懂未来的大数据应用

- 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。
- 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》。



一则笑话--读懂未来的大数据应用

- 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。
- 顾客：你们把比萨送我家吧。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.....



结语

数据越用越值钱

未来是数据驱动的时代

谁拥有数据，谁就是王者

但没有数据分析师，王者也要摆地摊





THANKS