北京交通大学
BEIJING JIAOTONG UNIVERSITY

大数据概论

大数据概念与应用

- 大数据的概念
- 大数据的来源
- 大数据的特征及意义
- 大数据的表现形态
- 大数据的应用场景

大数据的概念

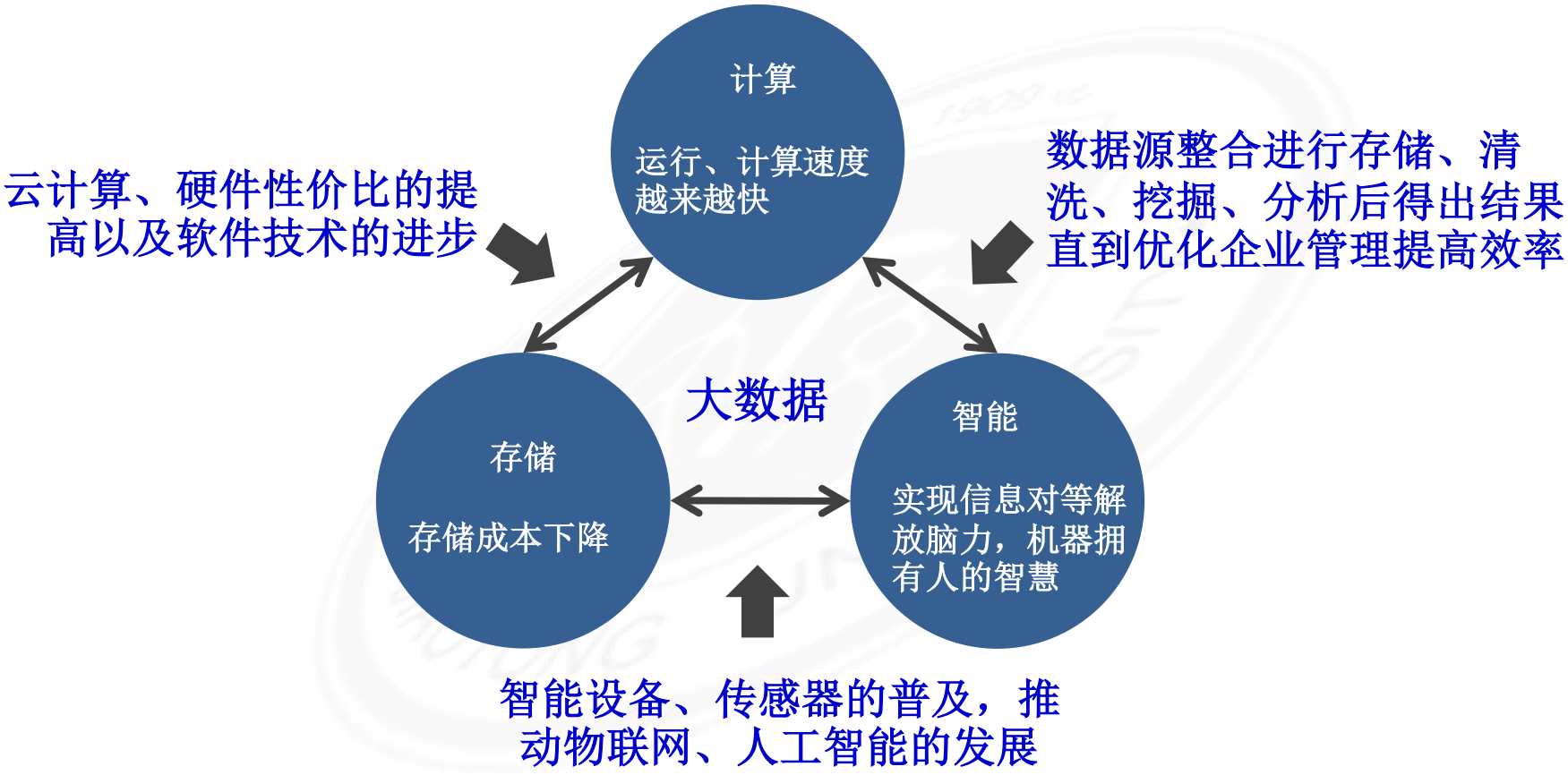
- 从“数据”到“大数据”
  - 时至今日，“数据”变身“大数据”，“开启了一次重大的时代转型”。
  - “大数据”这一概念的形成，有三个标志性事件：
    - 2008年9 月，美国《自然》（Nature）杂志专刊——The next google,第一次正式提出“大数据”概念。
    - 2011年2月1日，《科学》（Science）杂志专刊——Dealing with data，通过社会调查的方式，第一次综合分析了大数据对人们生活造成的影响，详细描述了人类面临的“数据困境”。
    - 2011年5月，麦肯锡研究院发布报告——Big data: The next frontier for innovation, competition, and productivity,第一次给大数据做出相对清晰的定义：“大数据是指其大小超出了常规数据库工具获取、储存、管理和分析能力的数据集。”

大数据的概念

- 什么是“大数据”？
  - 大数据是指以多元形式，自许多来源搜集而来的庞大数据组，往往具有实时性。在企业对企业销售的情况下，这些数据可能得自社交网络、电子商务网站、顾客来访纪录，还有许多其他来源。这些数据，并非公司顾客关系管理数据库的常态数据组。

大数据的概念

- 大数据的技术支撑



大数据的概念

- **存储：存储成本的下降**

云计算出现之前	云计算出现之后
在云计算出现之前，数据存储的成本是非常高的。 例如，公司要建设网站，需要购置和部署服务器，安排技术人员维护服务器，保证数据存储的安全性和数据传输的畅通性，还会定期清理数据，腾出空间以便存储新的数据，机房整体的人力和管理成本都很高。	云计算出现后，数据存储服务衍生出了新的商业模式，数据中心的出现降低了公司的计算和存储成本。 例如，公司现在要建设网站，不需要去购买服务器，不需要去雇用技术人员维护服务器，可以通过租用硬件设备的方式解决问题。

  - **存储成本的下降，也改变了大家对数据的看法，更加愿意把1年、2年甚至更久远的历史数据保存下来，有了历史数据的沉淀，才可以通过对比，发现数据之间的关联和价值。正是由于存储成本的下降，才能为大数据搭建最好的基础设施。**

大数据的概念

- 计算：计算速度越来越快
  - 海量数据从原始数据源到产生价值，期间会经过存储、清洗、挖掘、分析等多个环节，如果计算速度不够快，很多事情是无法实现的。所以，在大数据的发展过程中，计算速度是非常关键的因素。
    - » 分布式系统基础架构Hadoop的出现，为大数据带来了新的曙光；
    - » HDFS为海量的数据提供了存储；
    - » MapReduce则为海量的数据提供了并行计算，从而大大提高了计算效率；
    - » Spark、Storm、Impala等各种各样的技术进入人们的视野。

大数据的概念

- 智能：机器拥有理解数据的能力
  - 大数据带来的最大价值就是“智慧”，大数据让机器变得有智慧，同时人工智能进一步提升了处理和理解数据的能力。例如：
    - » 谷歌AlphaGo大胜世界围棋冠军李世石
    - » 阿里云小Ai成功预测出《我是歌手》的总决赛歌王
    - » iPhone上智能化语音机器人Siri
    - » 微信上与大家聊天的微软小冰

大数据的概念

- 大数据的意义

美国著名管理学家爱德华·戴明所言：“我们信靠上帝。除了上帝，任何人都必须用数据来说话。”

– 有数据可说
  » 在大数据时代，“万物皆数”，“量化一切”，“一切都将被数据化”。人类生活在一个海量、动态、多样的数据世界中，数据无处不在、无时不有、无人不用，数据就像阳光、空气、水分一样常见，好比放大镜、望远镜、显微镜那般重要。
– 说数据可靠
  » 大数据中的“数据”真实可靠，它实质上是表征事物现象的一种符号语言和逻辑关系，其可靠性的数理哲学基础是世界同构原理。世界具有物质统一性，统一的世界中的一切事物都存在着时空一致性的同构关系。这意味着任何事物的属性和规律，只要通过适当编码，均可以通过统一的数字信号表达出来。

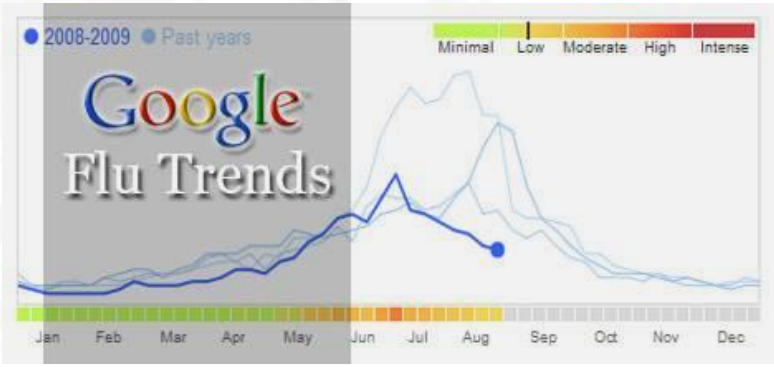
因此，“用数据说话”、“让数据发声”，已成为人类认知世界的一种全新方法。

大数据的概念

- 风马牛可相及
  - 在大数据背景下，因海量无限、包罗万象的数据存在，让许多看似毫不相干的现象之间发生一定的关联，使人们能够更简捷、更清晰地认知事物和把握局势。大数据的巨大潜能与作用现在难以进行估量，但揭示事物的相关关系无疑是其真正的价值所在。
    - » 经典案例：

啤酒与尿布

谷歌与流感

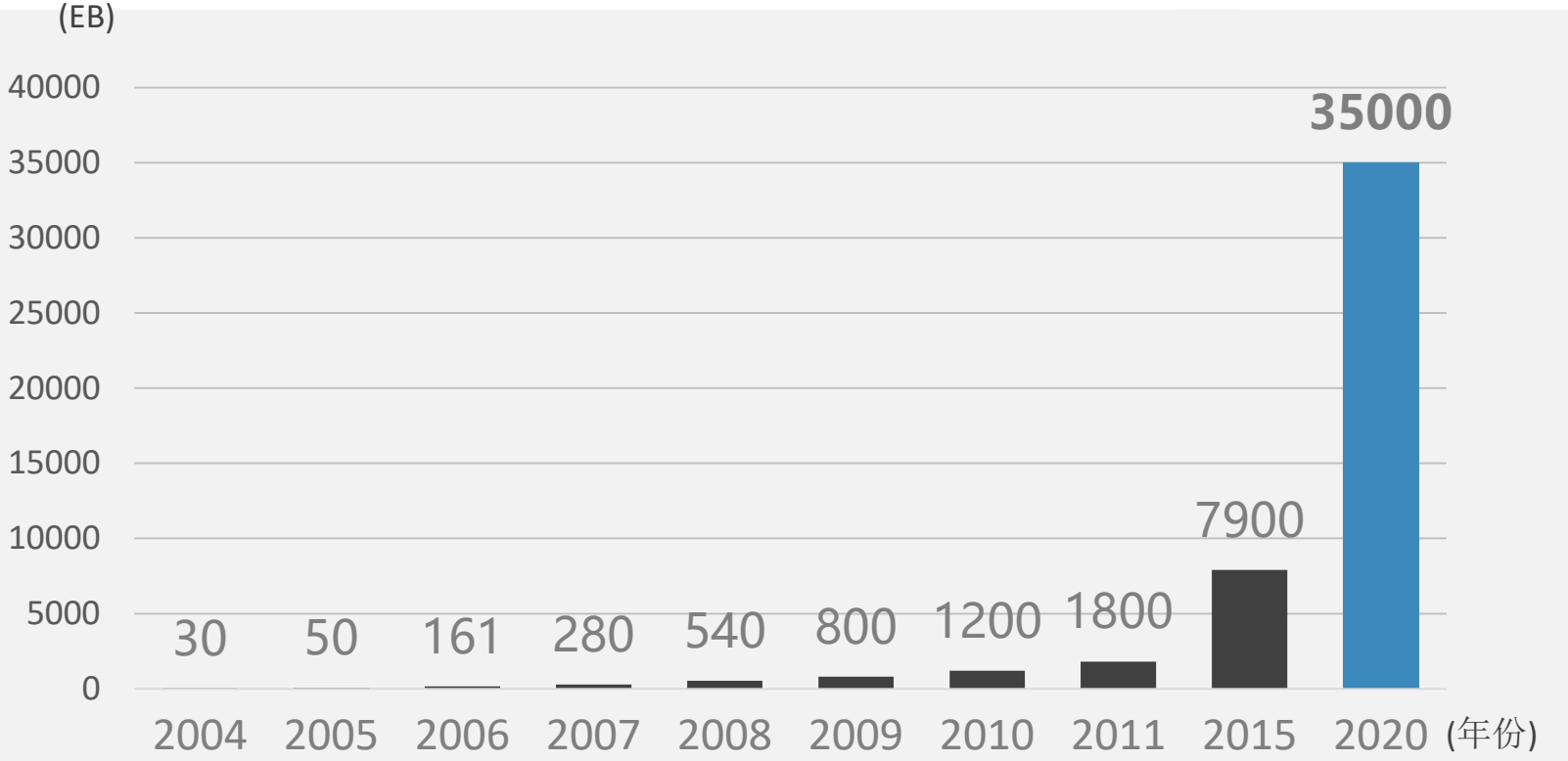


大数据概念与应用

- 大数据的概念
- 大数据的来源
- 大数据的特征及意义
- 大数据的表现形态
- 大数据的应用场景

大数据的来源

- 杰姆·格雷（Jim Gray）提出著名的“新摩尔定律”，即人类有史以来的数据总量，每过18个月就会翻一番。



全球数据总量图

为什么全球数据量

增长如此之快？

大数据的来源


大数据的主要来源
电脑
Home Automation
Patently Apple
智能家电
手机
智能手表
iPad

大数据的来源

互联网每天产生的全部内容可以刻满6.4亿张DVD

全球每秒发送290万封电子邮件，一分钟读一篇的话，足够一个人昼夜不停地读5.5年

Google每天需要处理24PB的数据

每天会有2.88万个小时的视频上传到YouTube，足够一个人昼夜不停地观看3.3年

网民每天在Facebook上要花费234亿分钟，被移动互联网使用者发送和接收的数据高达44PB

Twitter上每天发布5000万条消息，假设10秒就浏览一条消息，足够一个人昼夜不停地浏览16年

大数据到底有多大？ 以上一组互联网数据

大数据的来源

- 海量数据的产生

智能终端拍照、拍视频

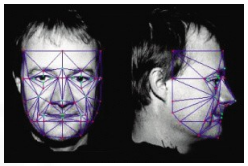
发微博、发微信

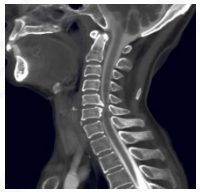
淘宝网 Taobao.com

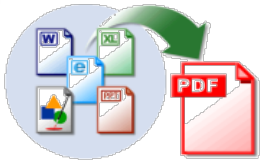
其他互联网数据

来自“大人群”泛互联网数据

来自大量传感器的机器数据







科学研究及行业多结构专业数据

随着人类活动的进一步扩展，数据规模会急剧膨胀，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的各行业累积的数据量越来越大，数据类型也越来越多、越来越复杂，已经超越了传统数据管理系统、处理模式的能力范围，于是“大数据”这样一个概念才会应运而生。

大数据的来源

- 按产生数据的主体划分
  - 少量企业应用产生的数据
    - 如关系型数据库中的数据和数据仓库中的数据等。
  - 大量人产生的数据
    - 如推特、微博、通信软件、移动通信数据、电子商务在线交易日志数据、企业应用的相关评论数据等。
  - 巨量机器产生的数据
    - 如应用服务器日志、各类传感器数据、图像和视频监控数据、二维码和条形码（条码）扫描数据等。

大数据的来源

- 按数据来源的行业划分
  - 以BAT为代表的互联网公司
    - 百度公司数据总量超过了千PB级别，阿里巴巴公司保存的数据量超过了百PB级别，拥有90%以上的电商数据，腾讯公司总存储数据量经压缩处理以后仍然超过了百PB级别，数据量月增加达到10%。
  - 电信、金融、保险、电力、石化系统
    - 电信行业数据年度用户数据增长超过10%，金融每年产生的数据超过数十PB，保险系统的数据量也超过了PB级别，电力与石化方面，仅国家电网采集获得的数据总量就达到了数十PB，石油化工领域每年产生和保存下来的数据量也将近百PB级别。

大数据的来源

- 公共安全、医疗、交通领域
  - 一个中、大型城市，一个月的交通卡口记录数可以达到3亿条；整个医疗卫生行业一年能够保存下来的数据就可达到数百PB级别；航班往返一次产生的数据就达到TB级别；列车、水陆路运输产生的各种视频、文本类数据，每年保存下来的也达到数十PB。
- 气象、地理、政务等领域
  - 中国气象局保存的数据将近10PB，每年约增数百TB；各种地图和地理位置信息每年约数十PB；政务数据则涵盖了旅游、教育、交通、医疗等多个门类，且多为结构化数据。

大数据的来源

- 制造业和其他传统行业
  - 制造业的大数据类型以产品设计数据、企业生产环节的业务数据和生产监控数据为主。其中产品设计数据以文件为主，非结构化，共享要求较高，保存时间较长；企业生产环节的业务数据主要是数据库结构化数据，而生产监控数据则数据量非常大。在其他传统行业，虽然线下商业销售、农林牧渔业、线下餐饮、食品、科研、物流运输等行业数据量剧增，但是数据量还处于积累期，整体体量都不算大，多则达到PB级别，少则数十TB或数百TB级别。

大数据的来源

- 按数据存储的形式划分
  - 大数据不仅仅体现在数据量大，还体现在数据类型多。如此海量的数据中，仅有20%左右属于结构化的数据，80%的数据属于广泛存在于社交网络、物联网、电子商务等领域的非结构化数据。
    - 结构化数据简单来说就是数据库，如企业ERP、财务系统、医疗HIS数据库、教育一卡通、政府行政审批、其他核心数据库等数据。
    - 非结构化数据包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频、视频信息等数据。

大数据的来源

- 常用的大数据获取途径
  - 系统日志采集
    - 可以使用海量数据采集工具，用于系统日志采集，如Hadoop的Chukwa、Cloudera的Flume、Facebook的Scribe等，这些工具均采用分布式架构，能满足大数据的日志数据采集和传输需求。
  - 互联网数据采集
    - 通过网络爬虫或网站公开API等方式从网站上获取数据信息，该方法可以数据从网页中抽取出来，将其存储为统一的本地数据文件，它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。除了网站中包含的内容之外，还可以使用DPI或DFI等带宽管理技术实现对网络流量的采集。
  - APP移动端数据采集
    - APP是获取用户移动端数据的一种有效方法，APP中的SDK插件可以将用户使用APP的信息汇总给指定服务器，即便用户在没有访问时，也能获知用户终端的相关信息，包括安装应用的数量和类型等。单个APP用户规模有限，数据量有限；但数十万APP用户，获取的用户终端数据和部分行为数据也会达到数亿的量级。
  - 与数据服务机构进行合作
    - 数据服务机构通常具备规范的数据共享和交易渠道，人们可以在平台上快速、明确地获取自己所需要的数据。而对于企业生产经营数据或学科研究数据等保密性要求较高的数据，也可以通过与企业或研究机构合作，使用特定系统接口等相关方式采集数据。

大数据概念与应用

- 大数据的概念
- 大数据的来源
- 大数据的特征及意义
- 大数据的表现形态
- 大数据的应用场景

大数据的特征及意义

- 大数据的3S
  - 大数据是数据分析的前沿技术。从各种各样类型的数据中，快速高效获得有价值信息的能力，就是大数据技术。在IT业界有的学者使用3S来描述大数据，还有的学者使用3I来描述大数据。
    - Size：数据的大小
    - Speed：数据的处理速度
    - Structur：数据的结构化

大数据的特征及意义

- 从技术上看“大数据”
  - 从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式计算架构。它的特色在于对海量数据的挖掘，但它必须依托云计算的分布式处理、分布式数据库、云存储和/或虚拟化技术。(在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法）
  - 大数据的4V特征：Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值）。

大数据的特征及意义

- **大数据的4V特征**

价值高（Value）

大数据有巨大的潜在价值，但同其呈几何指数爆发式增长相比，某一对象或模块数据的价值密度较低，这无疑给我们开发海量数据增加了难度和成本。

体量大（Volume）

从2013年至2020年，人类的数据规模将扩大50倍，每年产生的数据量将增长到44万亿GB，相当于美国国家图书馆数据量的数百万倍，且每18个月翻一番。

速度快（Velocity）

随着现代感测、互联网、计算机技术的发展，数据生成、储存、分析、处理的速度远远超出人们的想象力，这是大数据区别于传统数据或小数据的显著特征。

种类多（Variety）

大数据与传统数据相比，数据来源广、维度多、类型杂，各种机器仪表在自动产生数据的同时，人自身的生活行为也在不断创造数据；不仅有企业组织内部的业务数据，还有海量相关的外部数据。

大数据概念与应用

- 大数据的概念
- 大数据的来源
- 大数据的特征及意义
- 大数据的表现形态
- 大数据的应用场景

大数据的表现形态

- 大数据的表现形态

大数据在当今社会非常时髦，大数据的信息量是海量的，这个海量并不是某个时间端点的量级总结，而是持续更新，持续增量。由于大数据产生的过程中诸多的不确定性，使得大数据的表现形态多种多样。

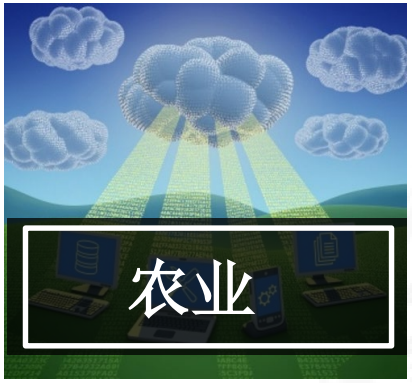
- 多源性：大数据来源的复杂性。网络技术的迅猛发展使得数据产生的途径多样化。大数据结构的复杂性。非结构化数据的格式多样化，而这些非结构化数据中可能蕴藏着非常有价值的信息。
- 实时性：大数据的实时性，体现在数据更新的实时性。如何及时、有效、全面的捕获到互联网、物联网、云计算上产生的大量的不同来源的数据是会直接影响数据价值体现的关键因素。
- 不确定性：体现的是数据的不确定性。原始数据的不准确以及数据采集处理粒度、应用需求与数据集成和展示等因素使得数据在不同尺度、不同维度上都有不同程度的不确定性。

大数据概念与应用

- 大数据的概念
- 大数据的来源
- 大数据的特征及意义
- 大数据的表现形态
- 大数据的应用场景

大数据的应用场景

大数据7个应用场景



大数据的应用场景

零售行业

零售行业大数据应用有两个层面，一个层面是零售行业可以了解客户的消费喜好和趋势，进行商品的精准营销，降低营销成本。另一个层面是依据客户购买的产品，为客户提供可能购买的其他产品，扩大销售额，也属于精准营销范畴。

未来考验零售企业的是如何挖掘消费者需求，以及高效整合供应链满足其需求的能力，因此，信息技术水平的高低成为获得竞争优势的关键要素。

金融行业

银行数据应用场景

利用数据挖掘来分析出一些交易数据背后的商业价值。

保险数据应用场景

用数据来提升保险产品的精算水平，提高利润水平和投资收益。

证券数据应用场景

对客户交易习惯和行为分析可以帮助证券公司获得更多的收益。

大数据的应用场景

医疗行业

医疗行业拥有大量的病例、病理报告、治愈方案、药物报告等，通过对这些数据进行整理和分析将会极大地辅助医生提出治疗方案，帮助病人早日康复。可以构建大数据平台来收集不同病例和治疗方案，以及病人的基本特征，建立针对疾病特点的数据库，帮助医生进行疾病诊断。

医疗行业的大数据应用一直在进行，但是数据并没有完全打通，基本都是孤岛数据，没办法进行大规模的应用。未来可以将这些数据统一采集起来，纳入统一的大数据平台，为人类健康造福。

教育行业

信息技术已在教育领域有了越来越广泛的应用，教学、考试、师生互动、校园安全、家校关系等，只要技术达到的地方，各个环节都被数据包裹。

通过大数据的分析来优化教育机制，也可以作出更科学的决策，这将带来潜在的教育革命，在不久的将来，个性化学习终端将会更多地融入学习资源云平台，根据每个学生的不同兴趣爱好和特长，推送相关领域的前沿技术、资讯、资源乃至未来职业发展方向。

大数据的应用场景

农业行业

借助于大数据提供的消费能力和趋势报告，政府可为农业生产进行合理引导，依据需求进行生产，避免产能过剩造成不必要的资源和社会财富浪费。

通过大数据的分析将会更精确地预测未来的天气，帮助农民做好自然灾害的预防工作，帮助政府实现农业的精细化管理和科学决策。

环境行业

借助于大数据技术，天气预报的准确性和实效性将会大大提高，预报的及时性将会大大提升，同时对于重大自然灾害如龙卷风，通过大数据计算平台，人们将会更加精确地了解其运动轨迹和危害的等级，有利于帮助人们提高应对自然灾害的能力。

智慧城市

大数据技术可以了解经济发展情况、各产业发展情况、消费支出和产品销售情况等，依据分析结果，科学地制定宏观政策，平衡各产业发展，避免产能过剩，有效利用自然资源和社会资源，提高社会生产效率。大数据技术也能帮助政府进行支出管理，透明合理的财政支出将有利于提高公信力和监督财政支出。