

大数据管理

Big Data Management



张海腾

htzhang@ecust.edu.cn

课程概述



- 课程名称： 大数据管理
- 课程性质： 选修课
- 理论学时： 28学时
- 实验学时： 8学时
- 课程教材： 大数据技术-原理与应用（第3版），林子雨，人民邮电出版社
- 考核方式： 期末成绩占70%，
平时成绩（包括课堂表现、作业和实验）占30%

课程内容



□ 第一篇 大数据基础

- 第一章 大数据概述
- 第二章 大数据处理架构Hadoop

□ 第二篇 大数据存储与管理

- 第三章 分布式文件系统HDFS
- 第四章 分布式数据库HBase
- 第五章 Nosql数据库

□ 第三篇 大数据处理与分析

- 第七章 MapReduce

第1章 大数据概述



- 1.1 大数据时代
- 1.2 大数据概念
- 1.3 大数据的影响
- 1.4 大数据的应用
- 1.5 大数据关键技术
- 1.6 大数据计算模式
- 1.7 大数据产业
- 1.8 大数据与云计算、物联网的关系

1.1 大数据时代



□ 第三次信息化浪潮涌动，大数据时代全面到来



□ 信息科技的发展为大数据时代提供技术支撑



□ 数据产生方式的变革是促成大数据时代到来的重要因素

第三次信息化浪潮

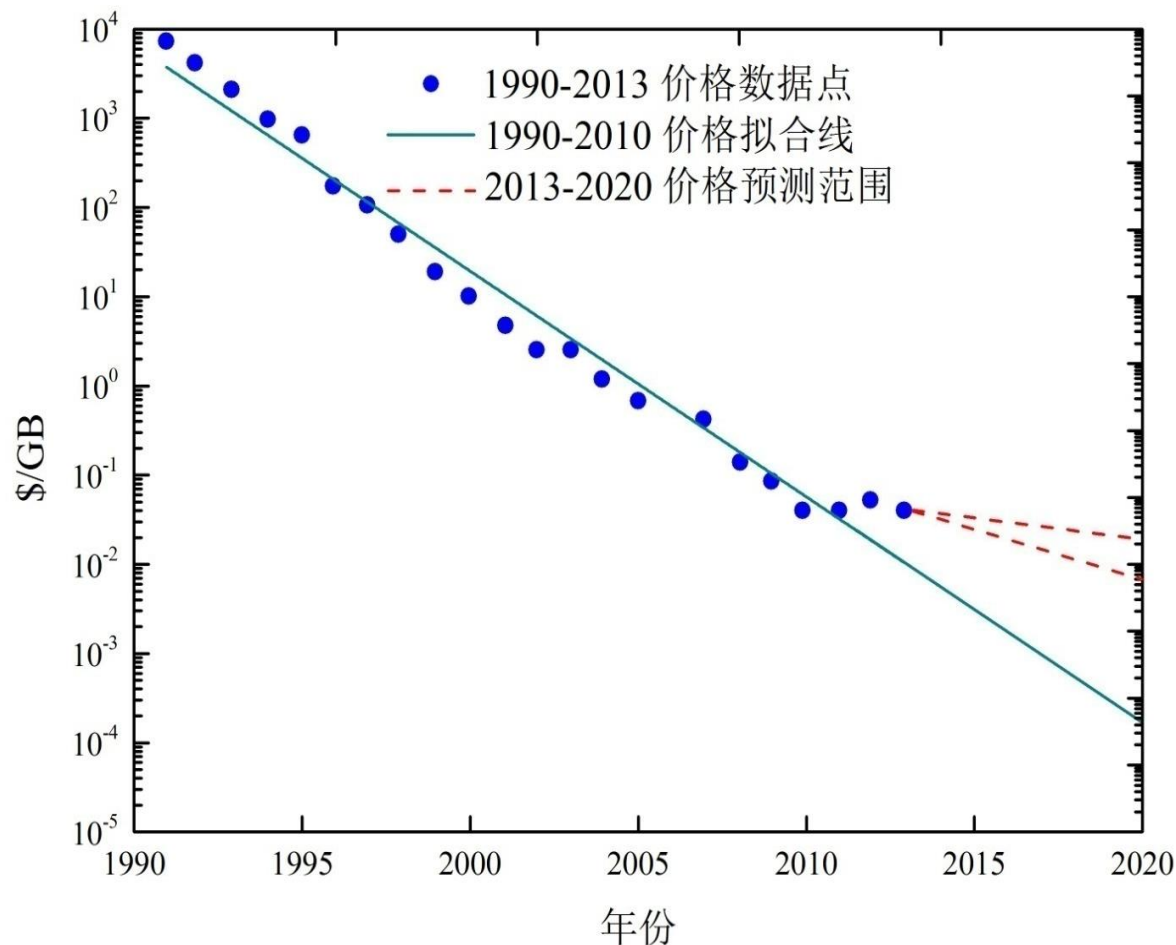
- 根据IBM前首席执行官郭士纳的观点，IT领域**每隔十五年**就会迎来一次重大变革。

信息化浪潮	发生时间	标志	解决问题	代表企业
第一次浪潮	1980年前后	个人计算机	信息处理	Intel、AMD、IBM、苹果、微软、联想、戴尔、惠普等
第二次浪潮	1995年前后	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次浪潮	2010年前后	物联网、云计算和大数据	信息爆炸	Amazon、Google、IBM、VMWare、Cloudera、Hortonworks、阿里云

信息科技为大数据时代提供技术支撑

□ 存储设备容量不断增加

- 存储设备的制造工艺不断升级，容量大幅增加，读写速度不断提升，价格却在不断下降。
- 随着单位存储空间价格的不断下降，人们开始倾向于把更多的数据保存起来，以期在未来某个时刻可以用更先进的数据分析工具从中挖掘价值。



存储价格随时间变化情况

信息科技为大数据时代提供技术支撑

□ CPU处理能力大幅度提升

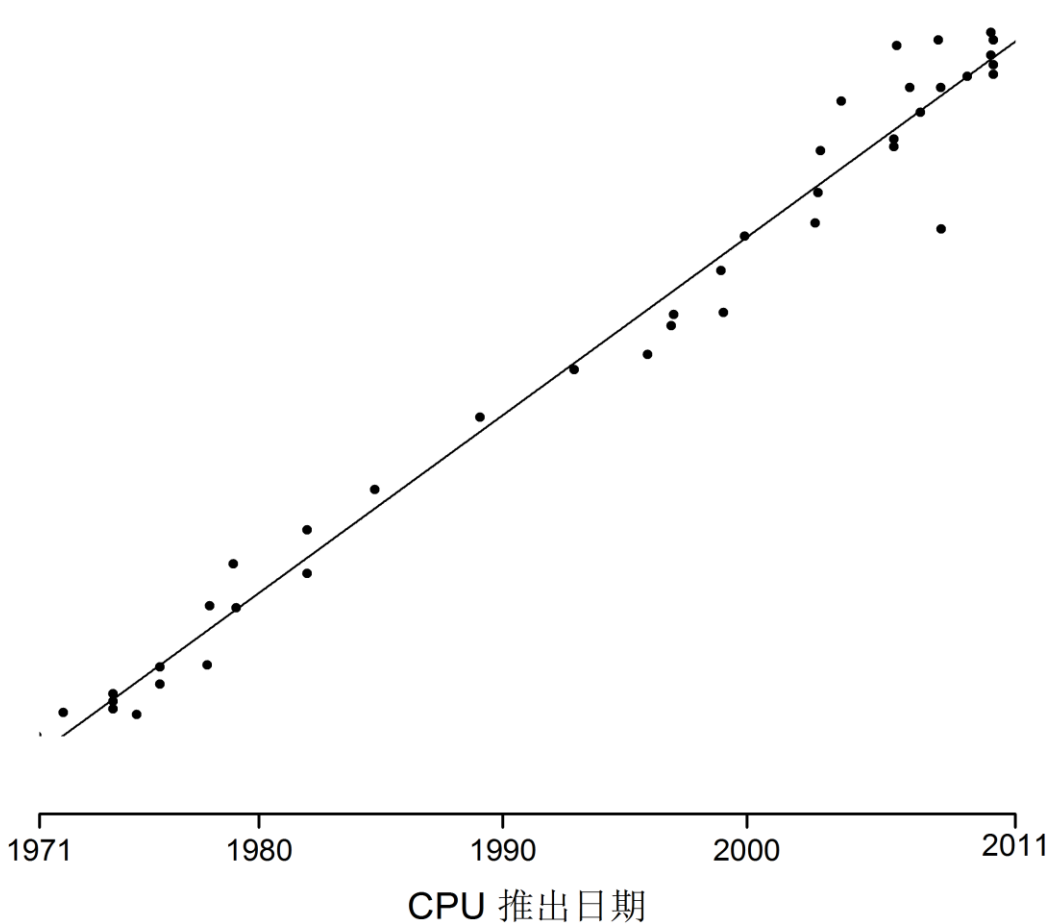
➤ “摩尔定律”：

1、集成电路芯片上所集成的晶体管的数目，每隔18个月就翻一番。

2、微处理器的性能每隔18个月提高一倍，而价格下降一半。

3、用一个美元所能买到的电脑性能，每隔18个月翻两番。

CPU 晶体管数目



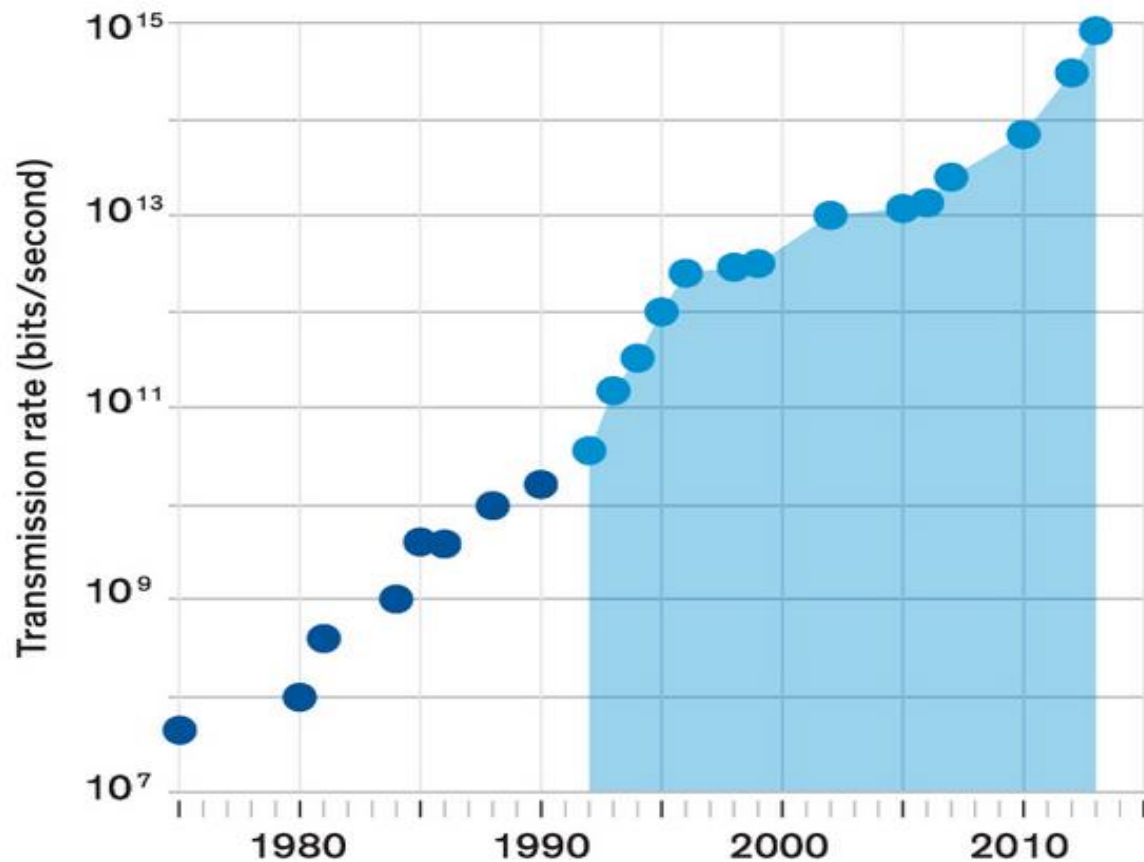
年份	CPU 晶体管数目 (估算)
1971	2,300
1974	6,000
1976	29,000
1978	290,000
1982	2.9 million
1985	29 million
1989	290 million
1993	2.9 billion
1997	29 billion
2000	290 million
2003	2.9 billion
2006	29 billion
2009	290 million
2011	2.9 billion

CPU 推出日期

CPU晶体管数目随时间变化情况

信息科技为大数据时代提供技术支撑

□网络带宽不断增加



网络带宽随时间变化情况

固定宽带：

截至2020年底，我国互联网宽带接入端口数量达到9.46亿个，其中光纤接入端口达到8.8亿个；

信息科技为大数据时代提供技术支撑

□网络带宽不断增加



移动通信宽带:

- 5G-目标: 让终端用户始终处于联网状态
- 2022年1月, 工业和信息化部发布《2021年通信业统计公报》, 显示截至2021年底, 我国累计建成并开通5G基站142.5万个, 总量占全球60%以上, 每万人拥有5G基站数达到10.1个。

数据产生方式的变革促成大数据时代的来临



大数据的发展历程

阶段	时间	内容
第一阶段： 萌芽期	上世纪90年代至本世纪初	随着数据挖掘理论和数据库技术的逐步成熟， 一批商业智能工具和知识管理技术开始被应用 ，如数据仓库、专家系统、知识管理系统等。
第二阶段： 成熟期	本世纪前十年	Web2.0应用迅猛发展，非结构化数据大量产生，传统处理方法难以应对，带动了大数据技术的快速突破，大数据解决方案逐渐走向成熟， 形成了并行计算与分布式系统两大核心技术 ，谷歌的GFS和MapReduce等大数据技术受到追捧，Hadoop平台开始大行其道
第三阶段： 大规模应用期	2010年以后	大数据 应用渗透各行各业，数据驱动决策 ，信息社会智能化程度大幅提高

1.2 大数据的概念

□ **大数据是指数据量超过一定的大小，导致常规软件无法在一个可接受的时间范围内完成对其进行抓取、管理和处理工作的数据。**

□ **例如：**

- **互联网上的网页数据**
- **社交网站上的用户交互数据**
- **物联网中产生的活动数据**
- **电信网络中的话单数据**

1.2 大数据的概念

大数据 = 海量数据（交易数据、交互数据）+ 针对海量数据处理的解决方案



想驾驭这庞大的数据，我们必须了解大数据的特征。

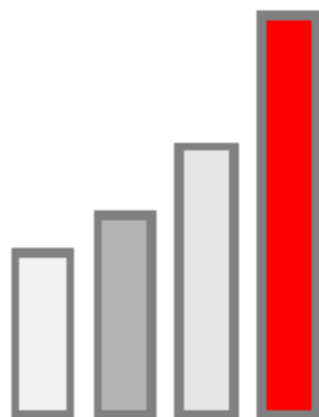
海量交易数据： 企业内部的经营交易信息主要包括联机交易数据和联机分析数据，是结构化的、通过关系数据库进行管理和访问的静态、历史数据。通过这些数据，我们能了解过去发生了什么。

海量交互数据： 源于Facebook、Twitter、微博、及其他来源的社交媒体数据构成。它包括了呼叫详细记录CDR、设备和传感器信息、GPS和地理定位映射数据、通过管理文件传输 Manage File Transfer协议传送的海量图像文件、Web 文本和点击流数据、科学信息、电子邮件等等。可以告诉我们未来会发生什么。

海量数据处理： 大数据的涌现已经催生出了设计用于数据密集型处理的架构。例如具有开放源码、在商品硬件群中运行的 Apache Hadoop。

注：大数据 不仅仅指的是数据量庞大，更为重要的是数据类型复杂

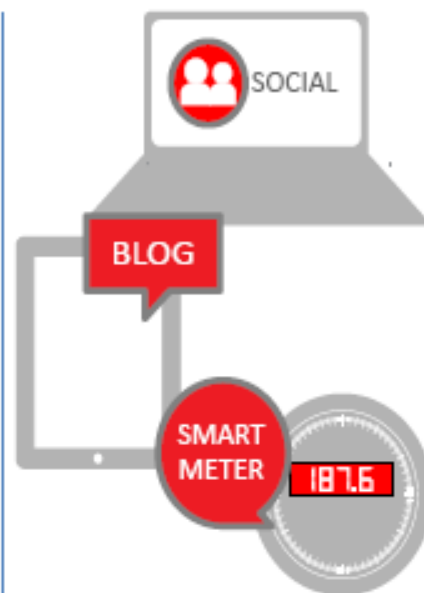
大数据的“4V”特征



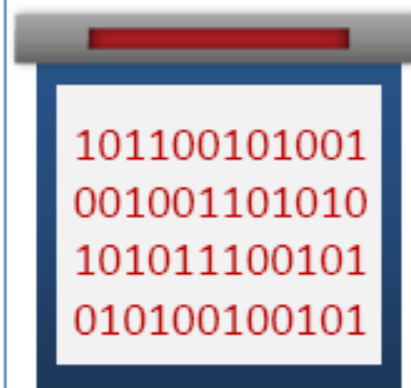
VOLUME
大量化



VELOCITY
快速化



VARIETY
多样化



VALUE

大数据不仅仅是数据的“大量化”，而是包含“快速化”、“多样化”和“价值化”等多重属性。

Volume—数据量大

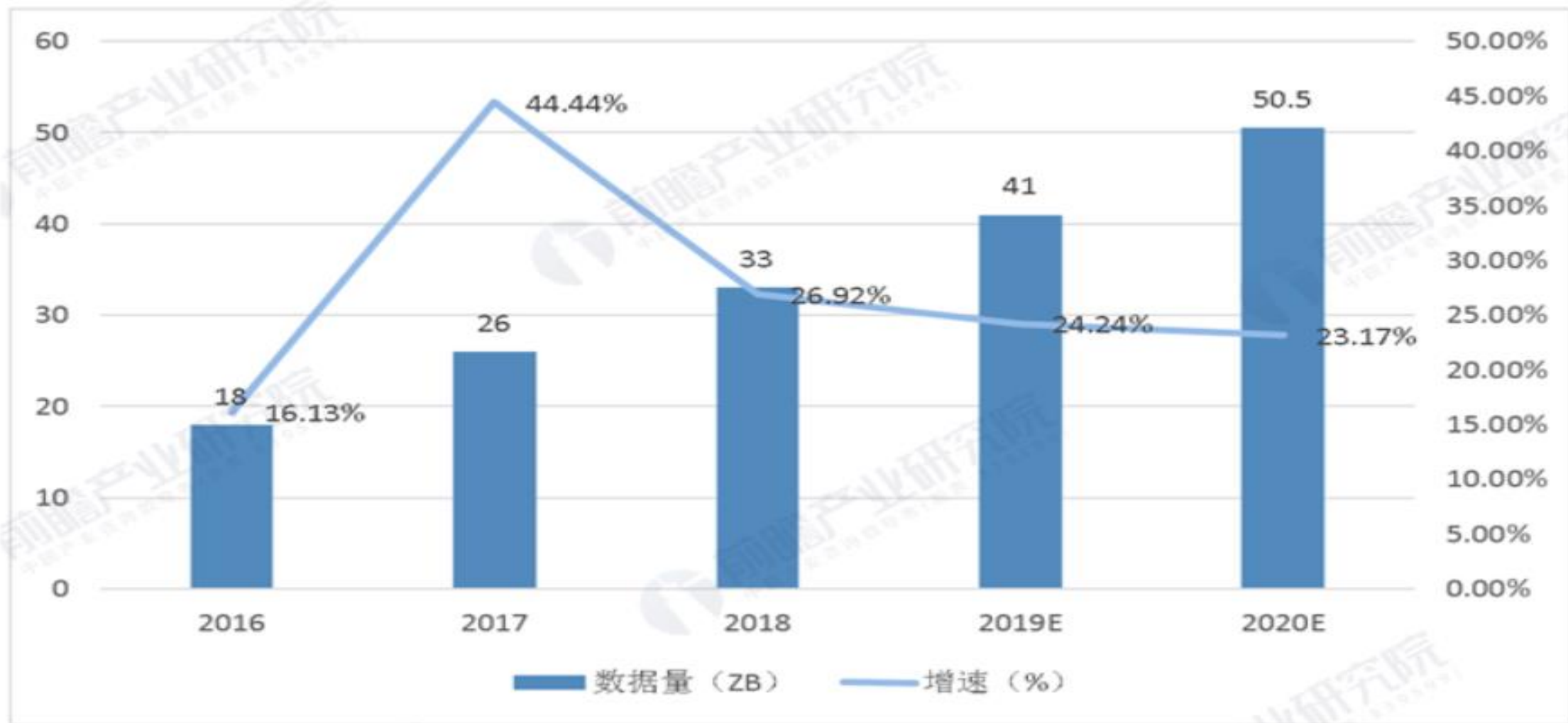
- 根据IDC作出的估测，数据一直都在以**每年50%的速度增长**，也就是说每两年就增长一倍（大数据摩尔定律）
- 人类在最近两年产生的数据量**相当于之前产生的全部数据量**
- 预计到2025年，全球数据总量将**比2016年的16.1ZB增加十倍**，达到163ZB(1ZB=1024EB,1EB=1024PB,1PB=1024TB)
- 中国的数据产生量约**占全球数据产生量的23%**，美国的数据产生量占比约为21%，EMEA(欧洲、中东、非洲)的数据产生量占比约为30%，APJXC(日本和亚太)数据产生量占比约为18%，全球其他地区数据产生量占比约为8%。

Volume —数据量大

TB	10的12次方	一块1TB硬盘		20000照片或mp3歌曲
PB	10的15次方	两个数据中心机柜		16个Blackblaze pod存储单元
EB	10的18次方	2000个机柜		占据一个街区的4层数据中心
ZB	10的21次方	1000个数据中心		纽约曼哈顿的1/5区域
YB	10的24次方	一百万个数据中心		特拉华州和罗德岛州

Volume —数据量大

图表1：2016-2020年全球每年产生数据量(单位：ZB，%)



资料来源：IDC、Seagate、Statista estimates 前瞻产业研究院整理

Variety—数据种类繁多

□大数据是由**结构化和非结构化数据**组成的

➤10%的结构化数据，存储在数据库中

➤90%的非结构化数据，它们与人类信息密切相关

- 科学研究（基因组，LHC 加速器，地球与空间探测）
- 企业应用（EMAIL、文档、文件，应用日志，交易记录）
- WEB 1.0数据（文本，图像，视频）
- WEB 2.0数据（查询日志，点击流，TWITTER/ BLOG / SNS-Social Networking Services, WIKI）

Velocity—处理速度快

- 从数据的生成到消耗，**时间窗口非常小**，可用于生成决策的时间非常少
- 1秒定律**：数据处理和分析的速度通常要达到**秒级响应**，这一点也是和传统的数据挖掘技术有着本质的不同



Value—价值密度低

□ 价值**密度低**，商业**价值高**

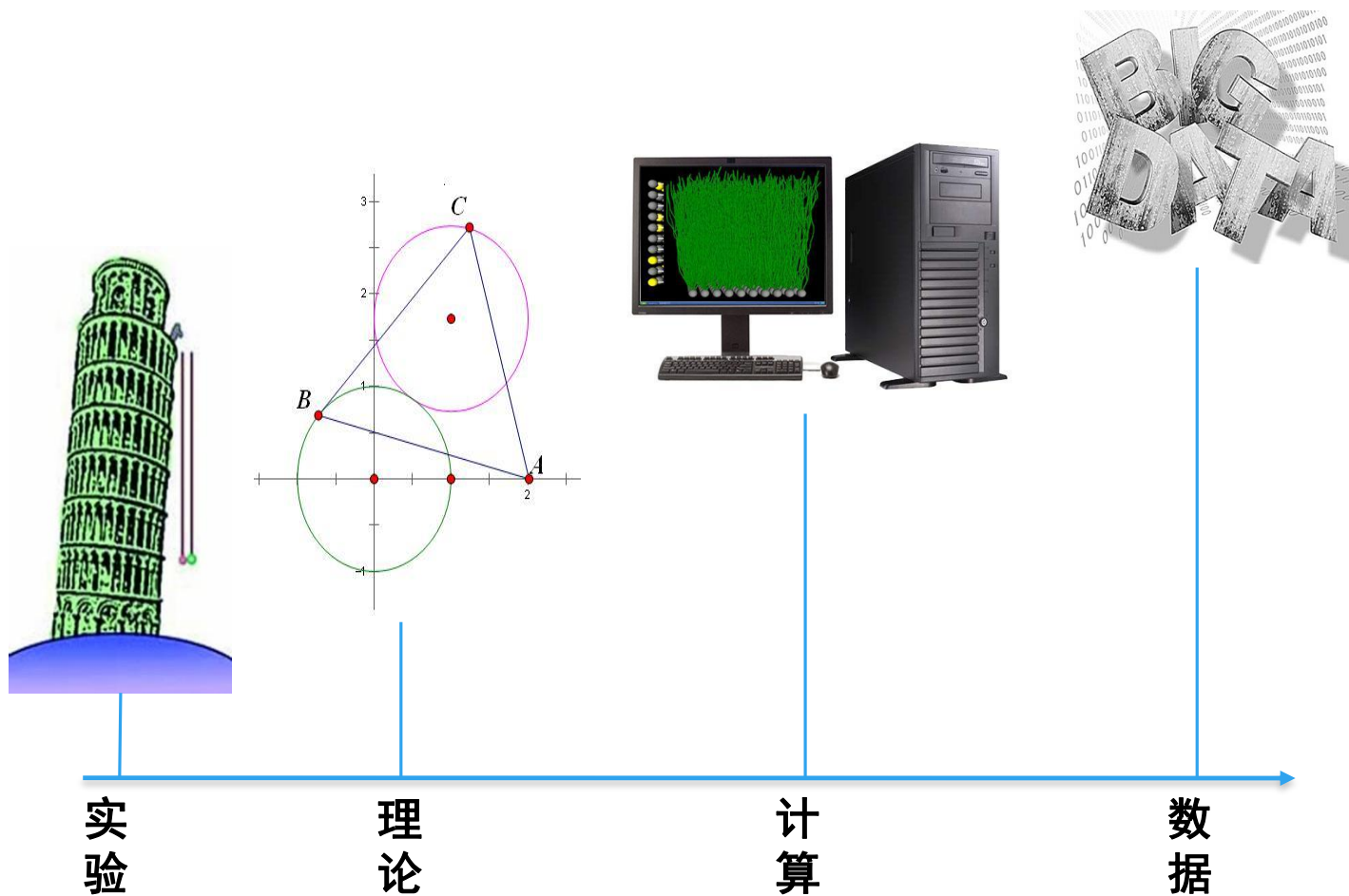
□ 以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒，但是具有很高的商业价值



1.3 大数据的影响

□大数据对科学研究的影响

➤图灵奖获得者、著名数据库专家Jim Gray 博士观察并总结人类自古以来，在科学研究上，先后历经了**实验科学**、**理论科学**、**计算科学**和**数据密集型科学**四种范式。



1.3 大数据的影响

□科学研究四个范式

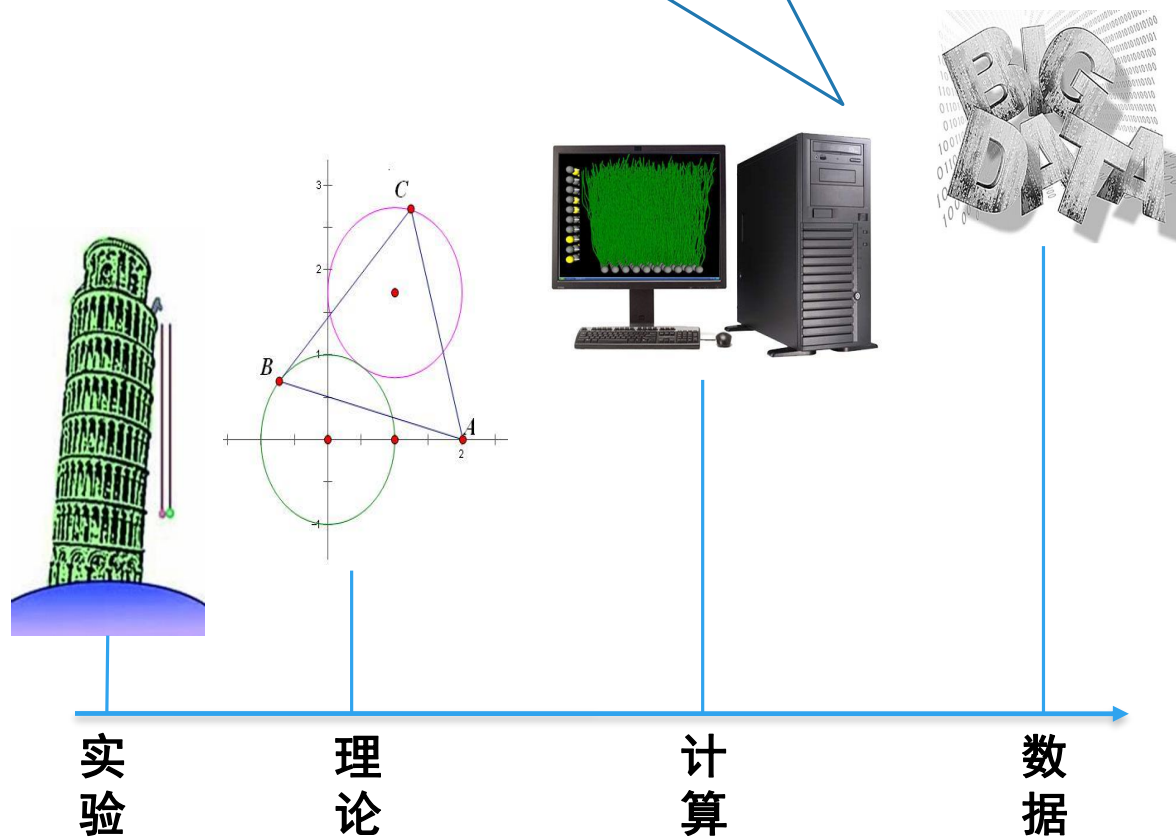
➤ **实验科学：** 人类采用实验来解决一些科学问题。

➤ **理论科学：** 人类开始采用数学、几何、物理等理论，构建问题模型和寻找解决方案。

➤ **计算科学：** 对各个科学进行计算模拟和其他形式的计算

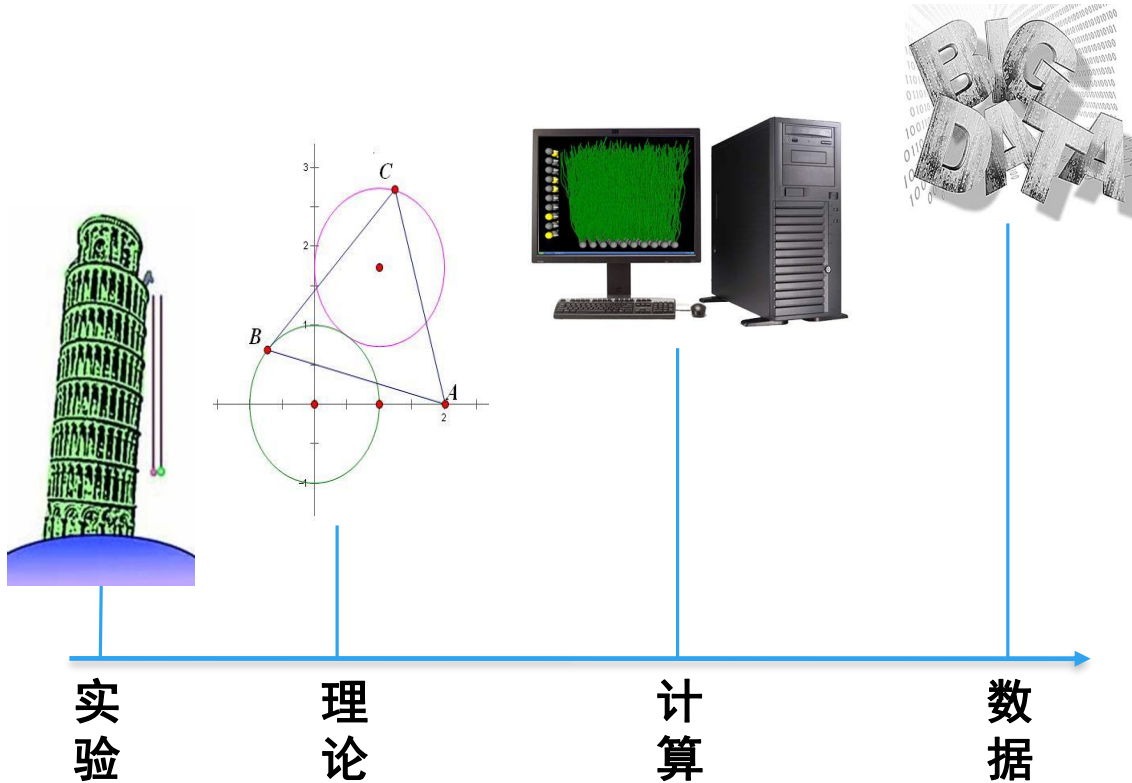
➤ **数据科学：** 以数据为中心，从数据中发现问题，解决问题，真正体现数据的价值。

计算科学和数据密集型科学这两种范式的本质区别？



1.3 大数据的影响

□ 大数据对科学研究的影响



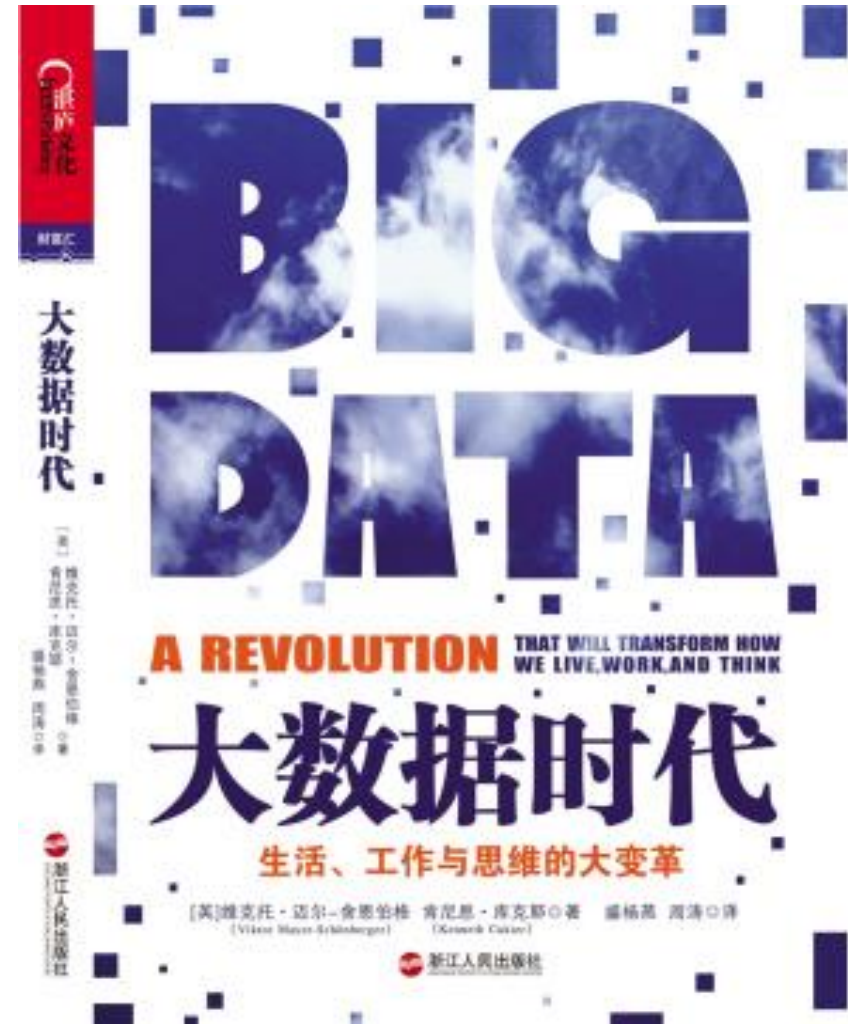
- 计算科学和数据密集型科学这两种范式都是利用计算机进行计算，但是二者有本质的区别：
 - 计算科学范式一般是先提出可能的理论，再搜集数据，然后通过计算来验证；
 - 数据密集型科学范式是先有了大量已知的数据，然后通过计算得出之前未知的理论。

1.3 大数据的影响

□大数据对思维方式的影响--大数据完全颠覆了传统的思维方式：

- 全样而非抽样
- 效率而非精确
- 相关而非因果

https://www.bilibili.com/video/BV1kW411y7vd/?spm_id_from=333.788.videocard.17



1.3 大数据的影响

1、全样而非抽样：

过去受到数据存储和处理能力的限制，在科学分析中，通常采用抽样的方法，通过对样本数据的分析来推断全集数据的总体特征；

大数据时代，有了大数据技术的支持，学科分析完全可以直接针对全集数据而不是抽样数据，并在短时间内得到分析结果。

1.3 大数据的影响

2、效率而非精确：

过去，在科学分析中采用抽样分析方法，必须追求分析方法的精确性，因为抽样分析只是针对部分样本的分析，其分析结果被应用到全集数据以后，误差会被放大，因此必须确保抽样分析结果的精确性；

大数据时代采用全样分析而不是抽样分析，分析结果不存在误差被放大的问题，数据分析的效率成为关注的核心。

1.3 大数据的影响

3、相关而非因果：

过去，数据分析的目的有两个方面：（1）解释事物背后的发展机理（2）预测未来可能发生的事件；

大数据时代，因果关系不再那么重要，人们转而追求“相关性”而非“因果性”。

1.3 大数据的影响

□大数据对社会发展的影响

- 在社会发展方面，**大数据决策逐渐成为一种新的决策方式**，大数据应用有力促进了信息技术与各行业的深度融合，**大数据开发大大推动了新技术和新应用的不断涌现**

□大数据对就业市场的影响

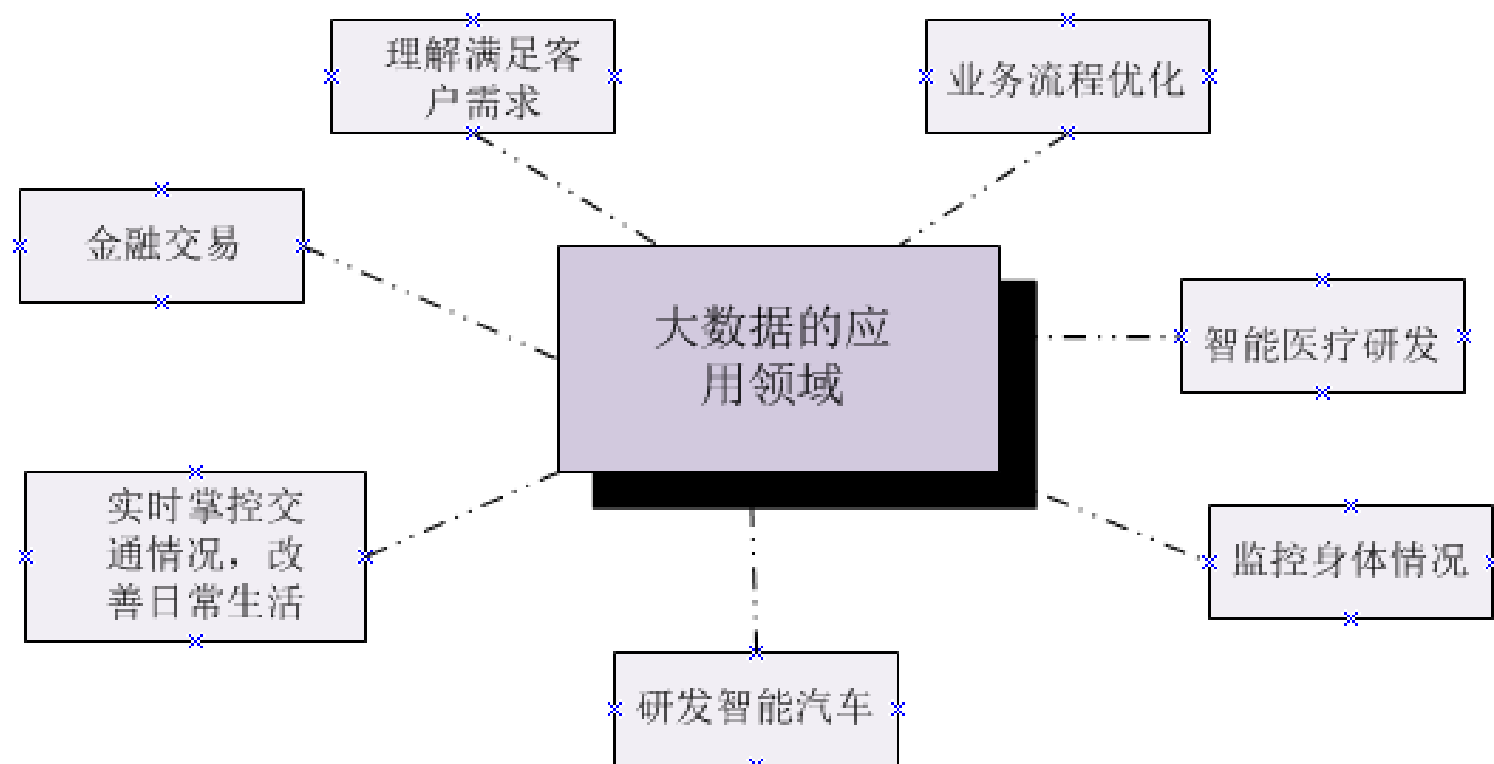
- 在就业市场方面，大数据的兴起使得**数据科学家成为热门职业**

□大数据对人才培养的影响

- 在人才培养方面，大数据的兴起，**将在很大程度上改变中国高校信息技术相关专业的现有教学和科研体制**

1.4 大数据的应用

□大数据无处不在，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的社会各行各业都已经融入了大数据的印迹



典型的大数据应用实例一

- 《纸牌屋》由视频网站NETFLIX投资并制作，电视剧的导演和男主角都是被“算”出来的。NETFLIX在美国有接近**2700万的订阅用户**，这些人每天在NETFLIX上产生**3000多万个网络点击行为**，例如收藏、推荐、暂停、回放、快进或者停止，并且用户每天还会给出**400万个评分**以及**300万次搜索请求**.....
- 根据数据，点击率非常高的鬼才导演大卫·芬奇和男演员凯文·史派西，成为主创的选择。
- 《纸牌屋》被誉为电视剧行业**通过互联网挖掘用户行为数据分析结果**的第一次战略运用。

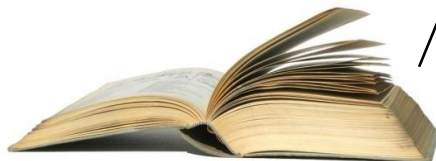
典型的大数据应用实例一



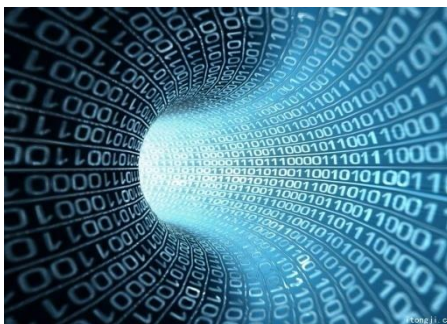
Kevin Spacey



David Fincher



英国同名小说《纸牌屋》



大数据分析

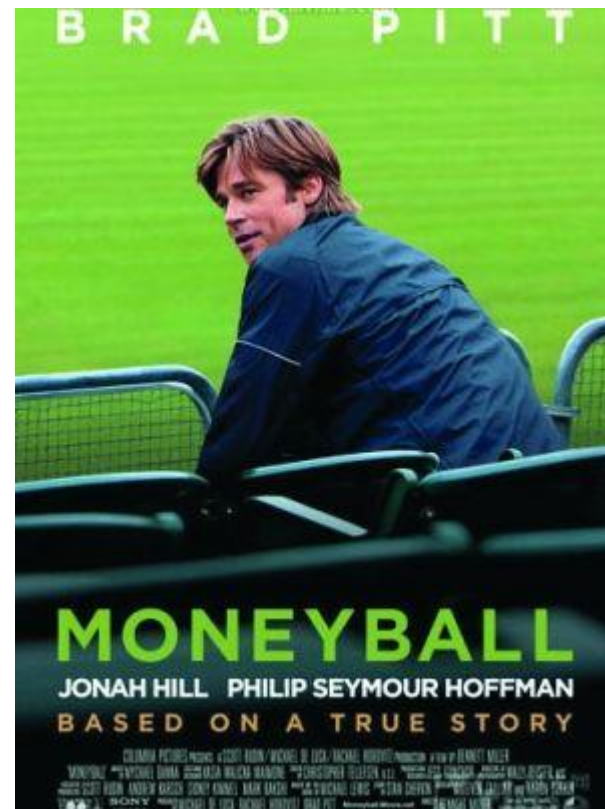


风靡全球的美剧《纸牌屋》

典型的大数据应用实例二

□美国奥斯卡获奖影片《点球成金》（布拉德·皮特主演）讲述的是棒球队总经理利用计算机数据分析，对球队进行了翻天覆地的改造，让一家不起眼的小球队能都取得巨大的成功。

- **基于历史数据**，利用**数据建模定量分析**不同球员的特点，合理搭配，重新组队。
- 打破传统思维，通过**分析比赛数据**，寻找“性价比”最高的球员，运用数据取得成功。



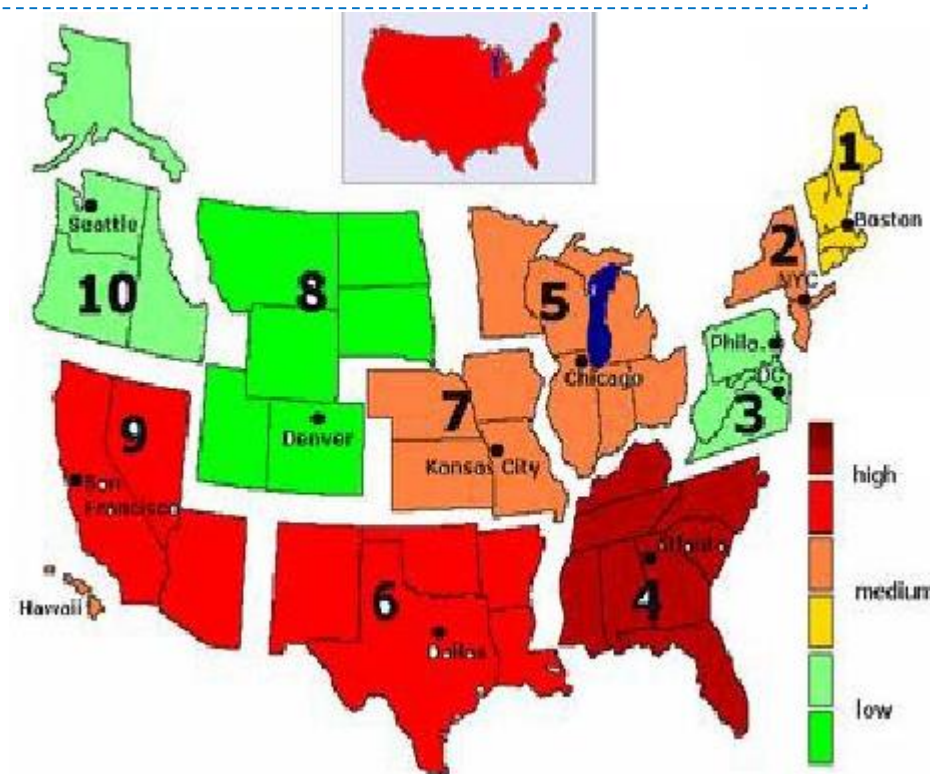
典型的大数据应用实例三



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

2009年，Google通过**分析**5000万条美国人最频繁检索的词汇，把它们和美国疾病中心在2003~2008年间季节性流感传播时期的数据**进行比较**，并建立一个**特定的数学模型**，最终**成功预测了**2009冬季流感的传播甚至可以具体到特定的地区和州。

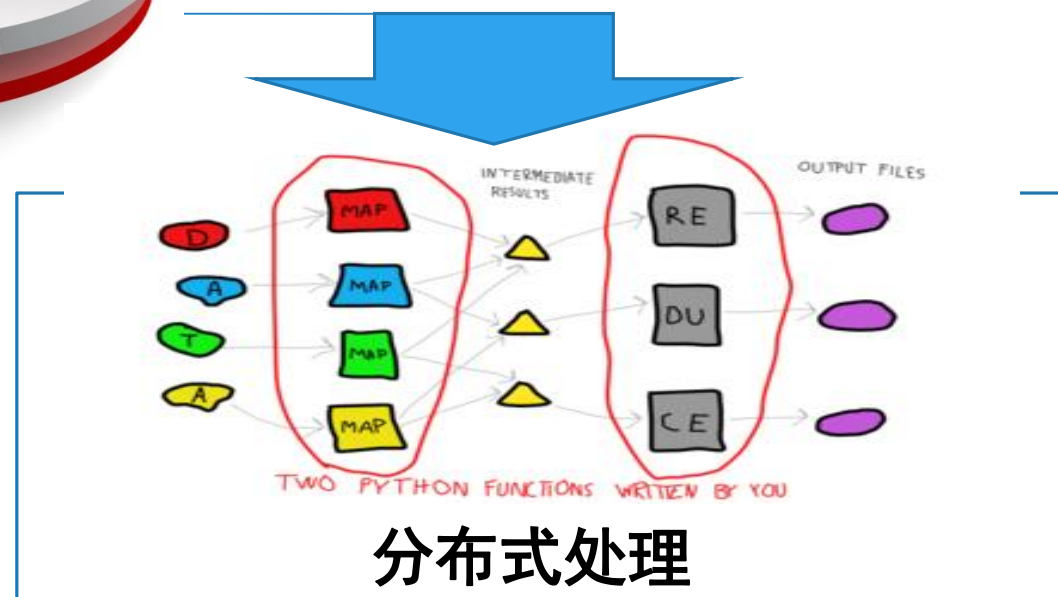
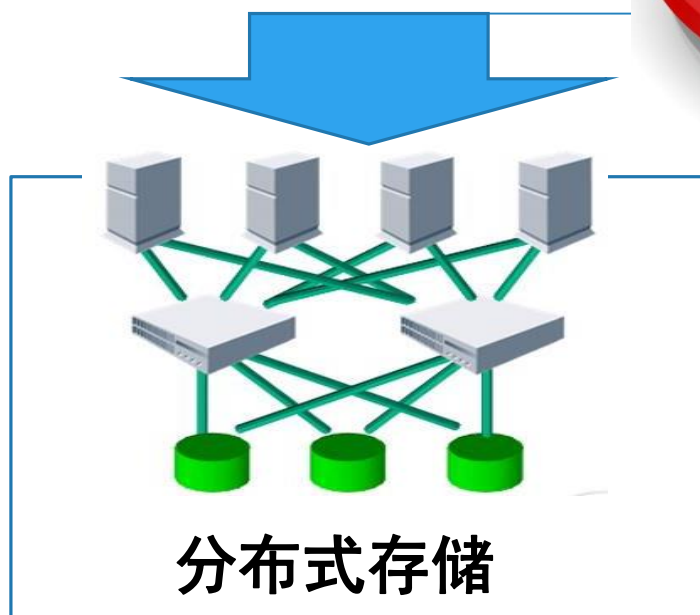


1.5 大数据关键技术

技术层面	功能
数据采集	利用ETL工具将分布的、异构数据源中的数据如关系数据、平面数据文件等，抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集中，成为联机分析处理、数据挖掘的基础；或者也可以把实时采集的数据作为流计算系统的输入，进行实时处理分析
数据存储和管理	利用分布式文件系统、数据仓库、关系数据库、NoSQL数据库、云数据库等，实现对结构化、半结构化和非结构化海量数据的存储和管理
数据处理、分析与展示	利用分布式并行编程模型和计算框架，结合机器学习和数据挖掘算法，实现对海量数据的处理和分析；对分析结果进行可视化呈现，帮助人们更好地理解数据、分析数据
数据隐私和安全	在从大数据中挖掘潜在的巨大商业价值和学术价值的同时，构建隐私数据保护体系和数据安全体系，有效保护个人隐私和数据安全

1.5 大数据关键技术

两大核心技术



GFS\HDFS

BigTable\HBase

NoSQL（键值、列族、图形、文档数据库）

NewSQL（如：SQL Azure）

MapReduce

1.6 大数据计算模式

大数据计算模式	解决问题	代表产品
批处理计算	针对大规模数据的批量处理	MapReduce、Spark等
流计算	针对流数据的实时计算	Storm、S4、Flume、Streams、Puma、DStream、Super Mario、银河流数据处理平台等
图计算	针对大规模图结构数据的处理	Pregel、GraphX、Giraph、PowerGraph、Hama、GoldenOrb等
查询分析计算	大规模数据的存储管理和查询分析	Dremel、Hive、Cassandra、Impala等

1.7 大数据产业

□大数据产业是指一切与**支撑大数据组织管理**和**价值发现**相关的企业经济活动的集合

产业链环节	包含内容
IT基础设施层	包括提供硬件、软件、网络等基础设施以及提供咨询、规划和系统集成服务的企业，比如，提供数据中心解决方案的IBM、惠普和戴尔等，提供存储解决方案的EMC，提供虚拟化管理软件微软、思杰、SUN、Redhat等
数据源层	大数据生态圈里的数据提供者，是生物大数据、交通大数据、医疗大数据、政务大数据、电商大数据、社交网络大数据、搜索引擎大数据等各种数据的来源

1.7 大数据产业

产业链环节	包含内容
数据管理层	包括数据抽取、转换、存储和管理等服务的各类企业或产品，比如分布式文件系统（如Hadoop的HDFS和谷歌的GFS）、ETL工具（Informatica、Datastage、Kettle等）、数据库和数据仓库（Oracle、MySQL、SQL Server、HBase、GreenPlum等）
数据分析层	包括提供分布式计算、数据挖掘、统计分析等服务的各类企业或产品，比如，分布式计算框架MapReduce、统计分析软件SPSS和SAS、数据挖掘工具Weka、数据可视化工具Tableau、BI工具（MicroStrategy、Cognos、BO）等等
数据平台层	包括提供数据分享平台、数据分析平台、数据租售平台等服务的企业或产品，比如阿里巴巴、谷歌、中国电信、百度等
数据应用层	提供智能交通、智慧医疗、智能物流、智能电网等行业应用的企业、机构或政府部门，比如交通主管部门、各大医疗机构、菜鸟网络、国家电网等

1.8 大数据与云计算、物联网

□ 云计算、大数据和物联网代表了IT领域最新的技术发展趋势，
三者相辅相成，既有联系又有区别



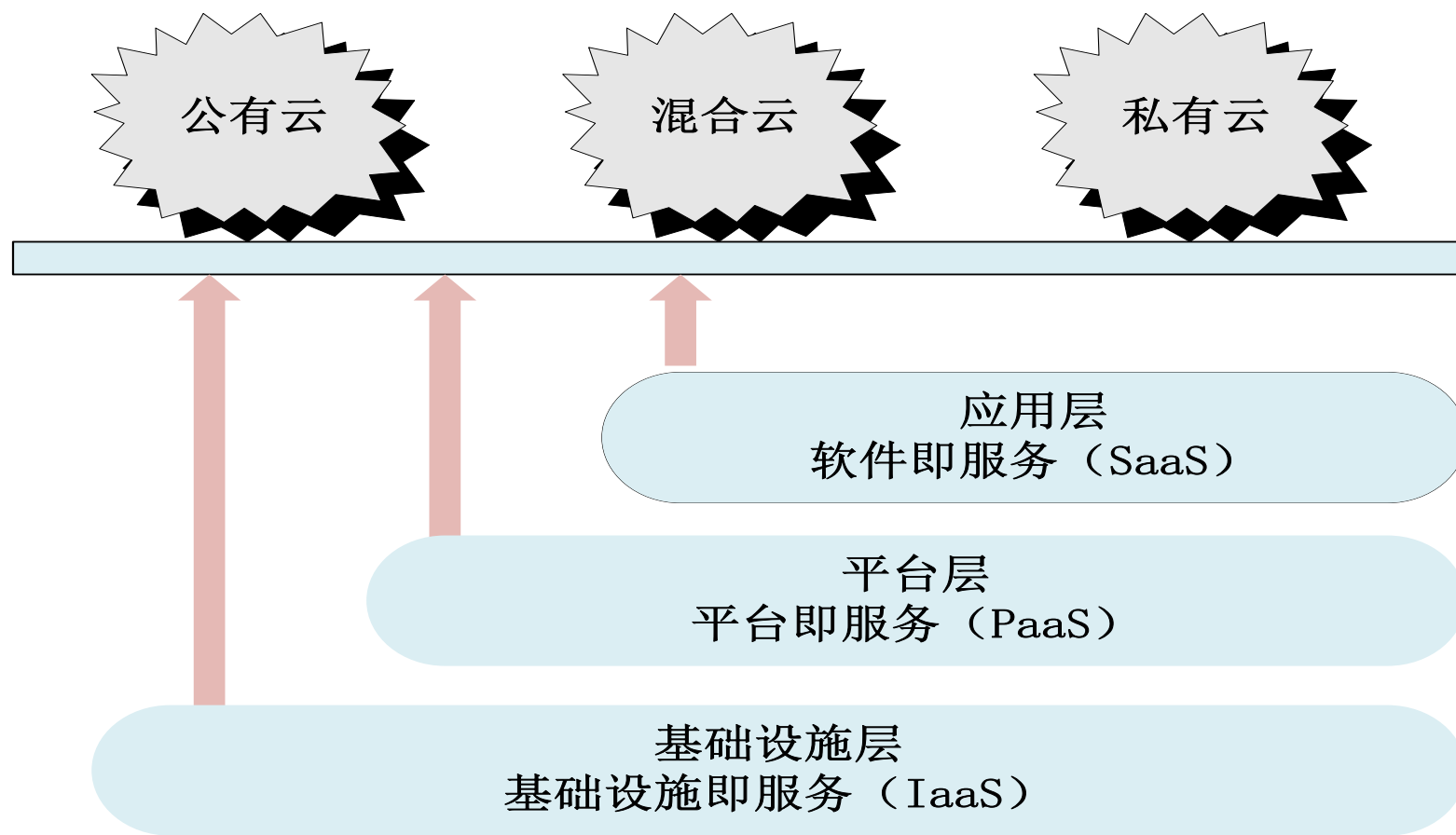
云
彻底改变IT

大数据
彻底改变业务



1.8.1 云计算—云计算的概念

□云计算实现了通过网络提供可伸缩的、廉价的分布式计算能力，用户只需要在具备网络接入条件的地方，就可以随时随地获得所需的各种IT资源。



云计算的服务模式和类型

1.8.1 云计算—云计算的概念

SaaS Software as a Service

- 从一个集中的系统部署软件，使之在一台本地计算机上(或从云中远程地)运行的一个模型。由于是计量服务，SaaS 允许出租一个应用程序，并计时收费
- Google Apps, Microsoft “Software+Services”

PaaS Platform as a Service

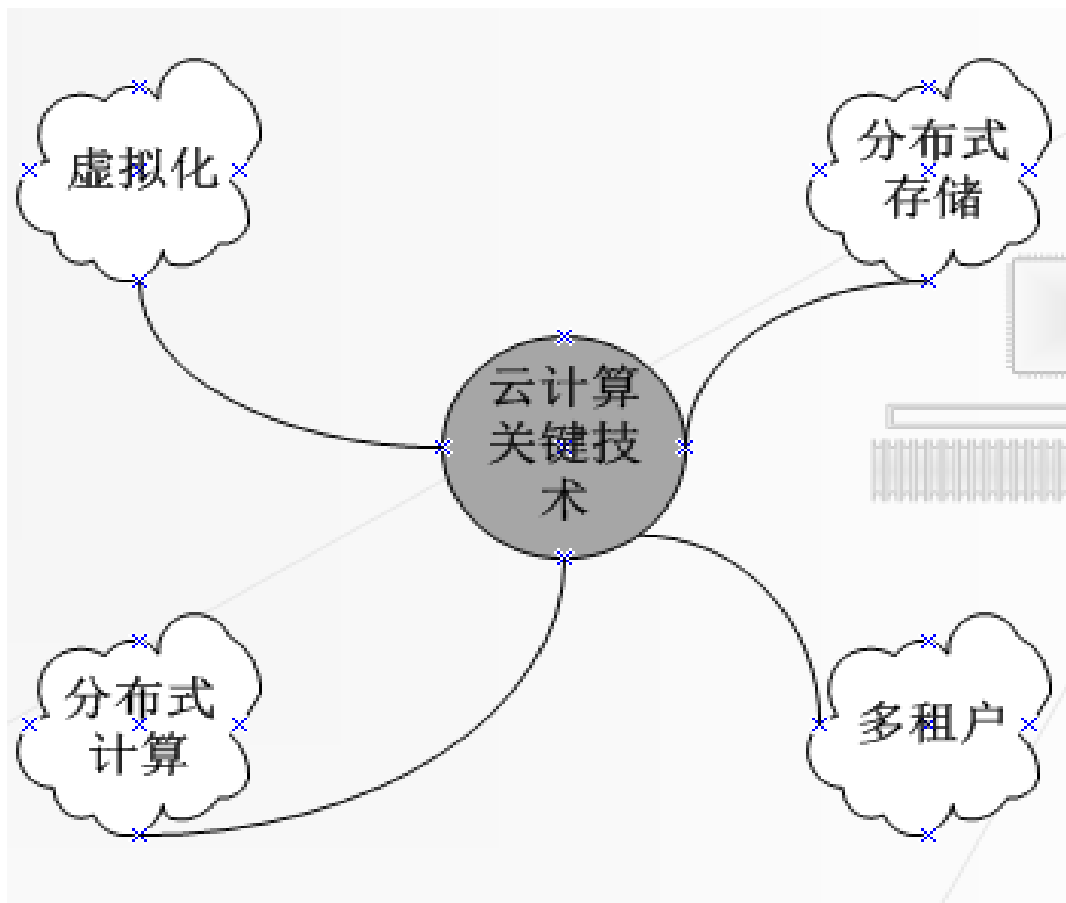
- 类似于 IaaS，但是它包括操作系统和围绕特定应用的必需的服务
- IBM IT factory, Google App Engine, Force.com

IaaS Infrastructure as a Service

- 将基础设施（计算资源和存储）作为服务出租
- Amazon EC2, IBM Blue Cloud, Sun Grid

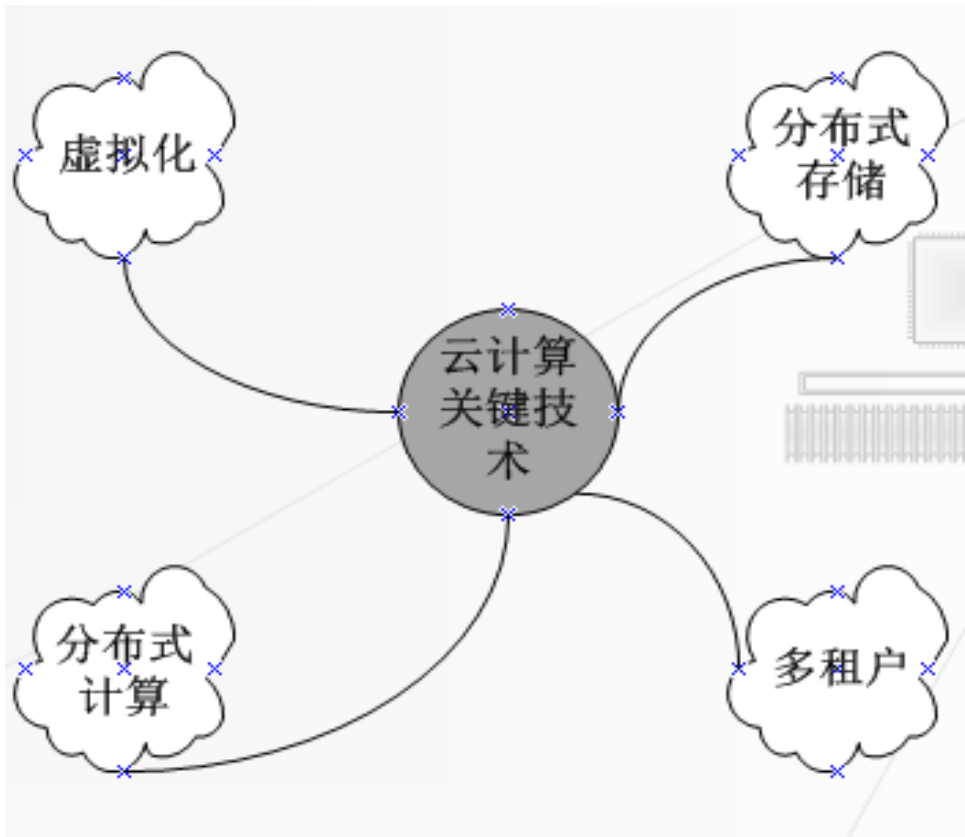
1.8.1 云计算—云计算的关键技术

□云计算关键技术包括：虚拟化、分布式存储、分布式计算、多租户等



- 1.虚拟化：**将一台计算机**虚拟成多台计算机**，多个应用程序在相互独立的空间运行，显著提高计算机工作效率。
- 2.分布式存储：**集中式存储无法满足海量数据的需求，而分布式存储可以在廉价PC服务器上搭建起**大规模存储集群**。

1.8.1 云计算—云计算的关键技术



3.分布式计算：在多个机器上并行处理数据，极大地提高了数据处理速度，可以满足对海量数据的批量处理需求。

4.多租户：多租户是指软件架构支持一个实例服务多个用户（Customer），每一个用户被称之为租户（tenant），软件给予租户可以对系统进行部分定制的能力，如用户界面颜色或业务规则，但是他们不能定制修改软件的代码。

1.8.1 云计算—云计算数据中心

- **云计算数据中心是一整套复杂的设施**，包括刀片服务器、宽带网络连接、环境控制设备、监控设备以及各种安全装置等
- **数据中心是云计算的重要载体**，为云计算提供计算、存储、带宽等各种硬件资源，为各种平台和应用提供运行支撑环境
- **全国各地推进数据中心建设**

<https://www.bilibili.com/video/av430278173/>



阿里巴巴张北数据中心



- 位于河北省西北部的张北县。
重点考虑了风能和太阳能。
- 该云计算基地为“**一点三中心**”部署，即三个相互备份的数据中心园区以及一个示范展示点。
- 一号园区和二号园区分别投资为60亿元、占地200亩、建设容量**10万台服务器，实现机械制冷和自然风冷之间的精确控制调换。**
- 预计全年只有二周需要开启传统的压缩机空调制冷，**仅制冷能耗就能降低近60%。**

1.8.1 云计算—云计算应用

- **政务云**上可以部署公共安全管理、容灾备份、城市管理、应急管理、智能交通、社会保障等应用，通过集约化建设、管理和运行，可以实现信息资源整合和政务资源共享，**推动政务管理创新，加快向服务型政府转型。**
- **教育云**可以有效整合幼儿教育、中小学教育、高等教育以及继续教育等优质教育资源，**逐步实现教育信息共享、教育资源共享及教育资源深度挖掘等目标**
- **中小企业云**能够让企业以低廉的成本建立财务、供应链、客户关系等管理应用系统，大大降低企业信息化门槛，**迅速提升企业信息化水平，增强企业市场竞争力**
- **医疗云**可以推动医院与医院、医院与社区、医院与急救中心、医院与家庭之间的服务共享，**并形成一套全新的医疗健康服务系统，从而有效地提高医疗保健的质量**

1.8.1 云计算—云计算产业

□云计算产业作为**战略性新兴产业**，近些年得到了迅速发展，形成了成熟的产业链结构。

□**产业涵盖**硬件与设备制造、基础设施运营、软件与解决方案供应商、基础设施即服务（IAAS）、平台即服务（PAAS）、软件即服务（SAAS）、终端设备、云安全、云计算交付/咨询/认证等环节。



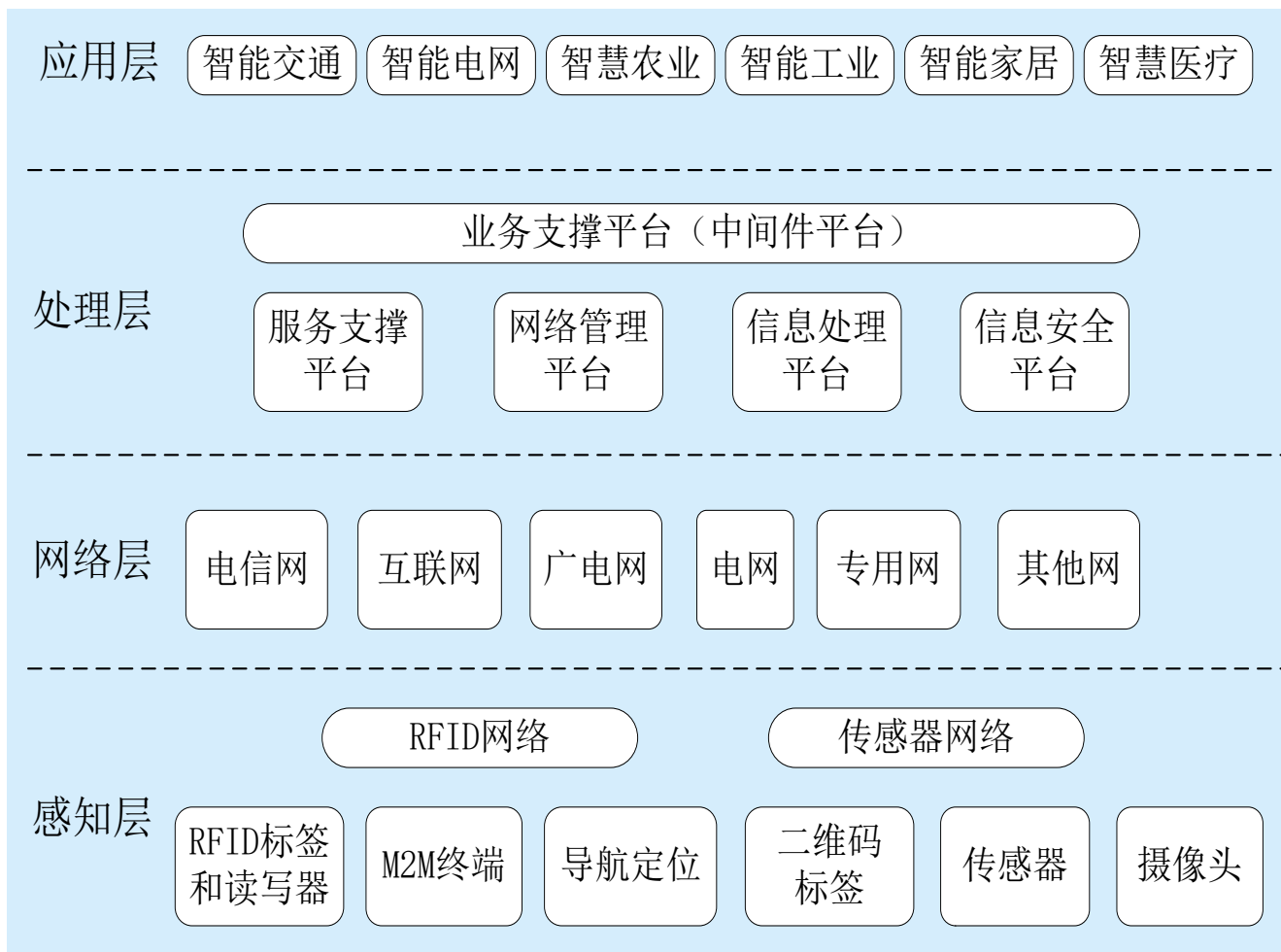
1.8.2 物联网—物联网的概念

□ 物联网（INTERNET OF THINGS, IOT）是物物相连的互联网，是互联网的延伸。

□ 它利用局部网络或互联网等通信技术把传感器、控制器、机器、人员和物等通过新的方式联在一起，形成人与物、物与物相连，实现信息化和远程管理控制。



1.8.2 物联网——物联网的概念



物联网体系架构

(1) 感知层：如果把物联网系统比喻为一个人体，感知层就好比人体的神经末梢，用来感知物理世界，**采集来自物理世界的各种信息。该层包含了大量的传感器。**

(2) 网络层：相当于人体的神经中枢，**起到信息传输的作用。**网络层包含各种类型的网络，如互联网、移动通信网络、卫星通信网络等。

1.8.2 物联网——物联网的概念



物联网体系架构

(3) 处理层：相当于人体的大脑，起到存储和处理的作用，**包括数据存储、管理和分析平台。**

(4) 应用层：直接面向用户，**满足各种应用需求**，如智能交通、智慧农业、智慧医疗、智能工业等。

1.8.2 物联网—物联网关键技术

物联网中的关键技术包括识别和**感知技术**（二维码、RFID、传感器等）、**网络与通信技术**、**数据挖掘与融合技术**等

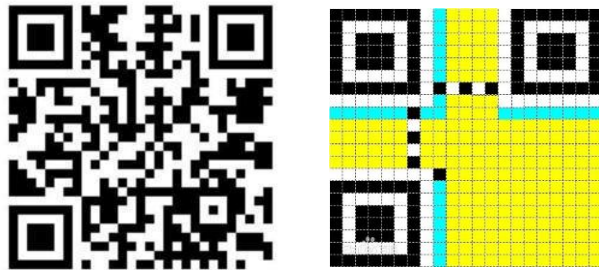


图1 矩阵式二维码



图2 采用RFID芯片的公交卡



(a)温湿度传感器



(b)压力传感器



(c)烟雾传感器

图3 不同类型的传感器

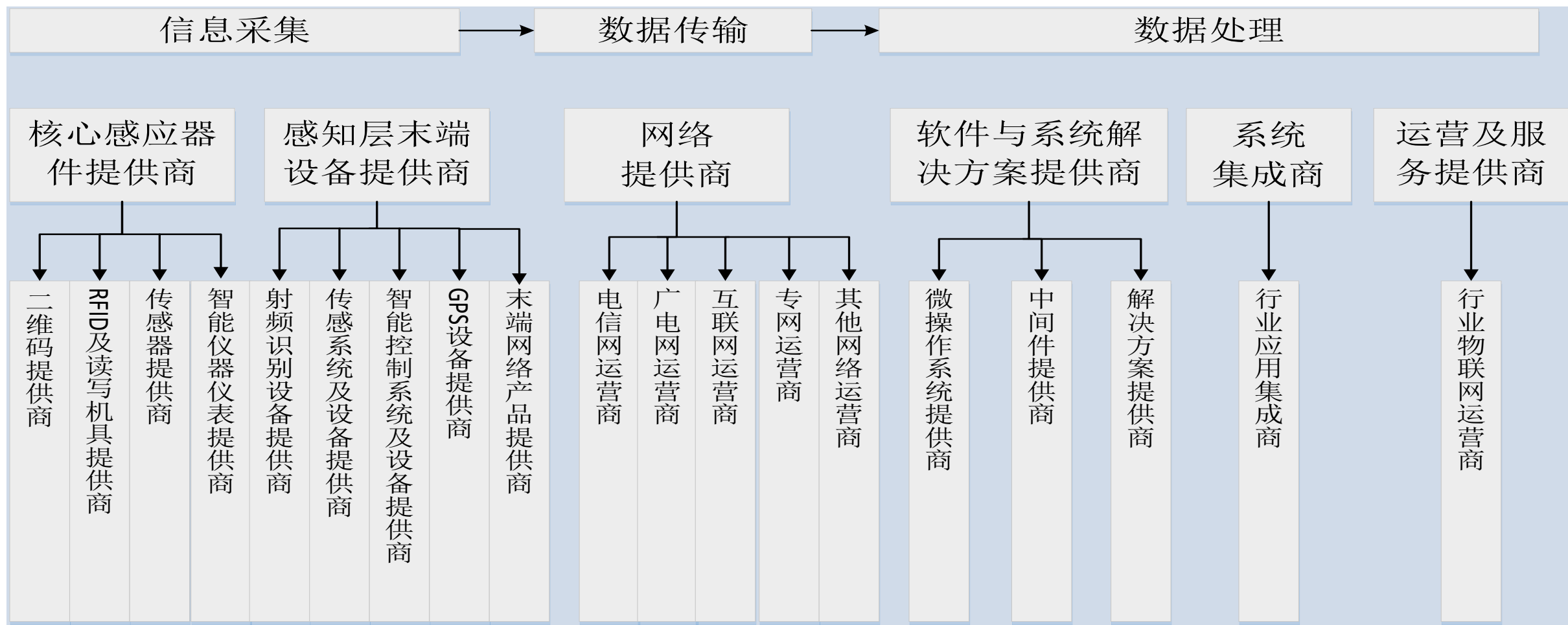
1.8.2 物联网—物联网应用

□物联网已经广泛应用于智能交通、智慧医疗、智能家居、环保监测、智能安防、智能物流、智能电网、智慧农业、智能工业等领域，**对国民经济与社会发展起到了重要的推动作用**



1.8.2 物联网—物联网产业

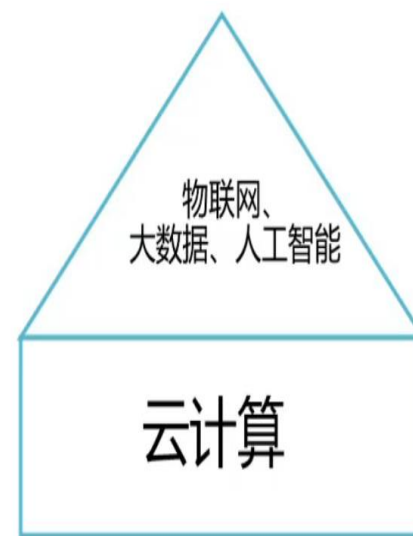
完整的物联网产业链主要包括**核心感应器件提供商、感知层末端设备提供商、网络提供商、软件与系统解决方案提供商、系统集成商、运营及服务提供商**等六大环节



1.8.3 大数据与云计算、物联网的关系

□云计算、大数据和物联网代表了IT领域最新的技术发展趋势，三者既有区别又有联系。

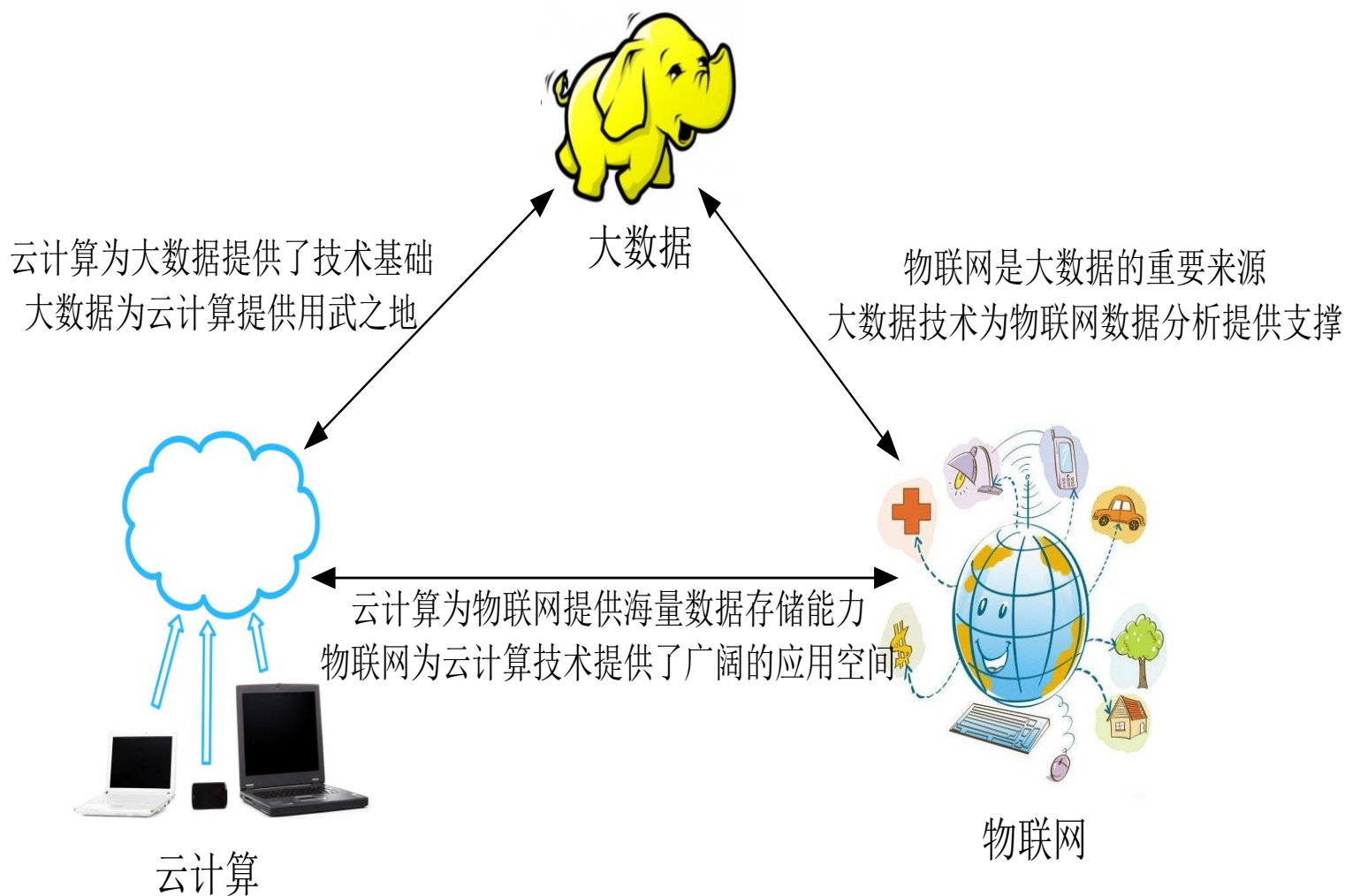
- **大数据**侧重于海量数据的存储、处理与分析，从海量数据中发现价值，服务于生产和生活；
- **云计算**本质上旨在整合和优化各种IT资源，并通过网络以服务的方式廉价提供给用户；
- **物联网**的发展目标是实现物物相连，应用创新是物联网发展的核心。



云计算是其他领域的基石，其他领域推动云计算的进一步发展

1.8.3 大数据与云计算、物联网的关系

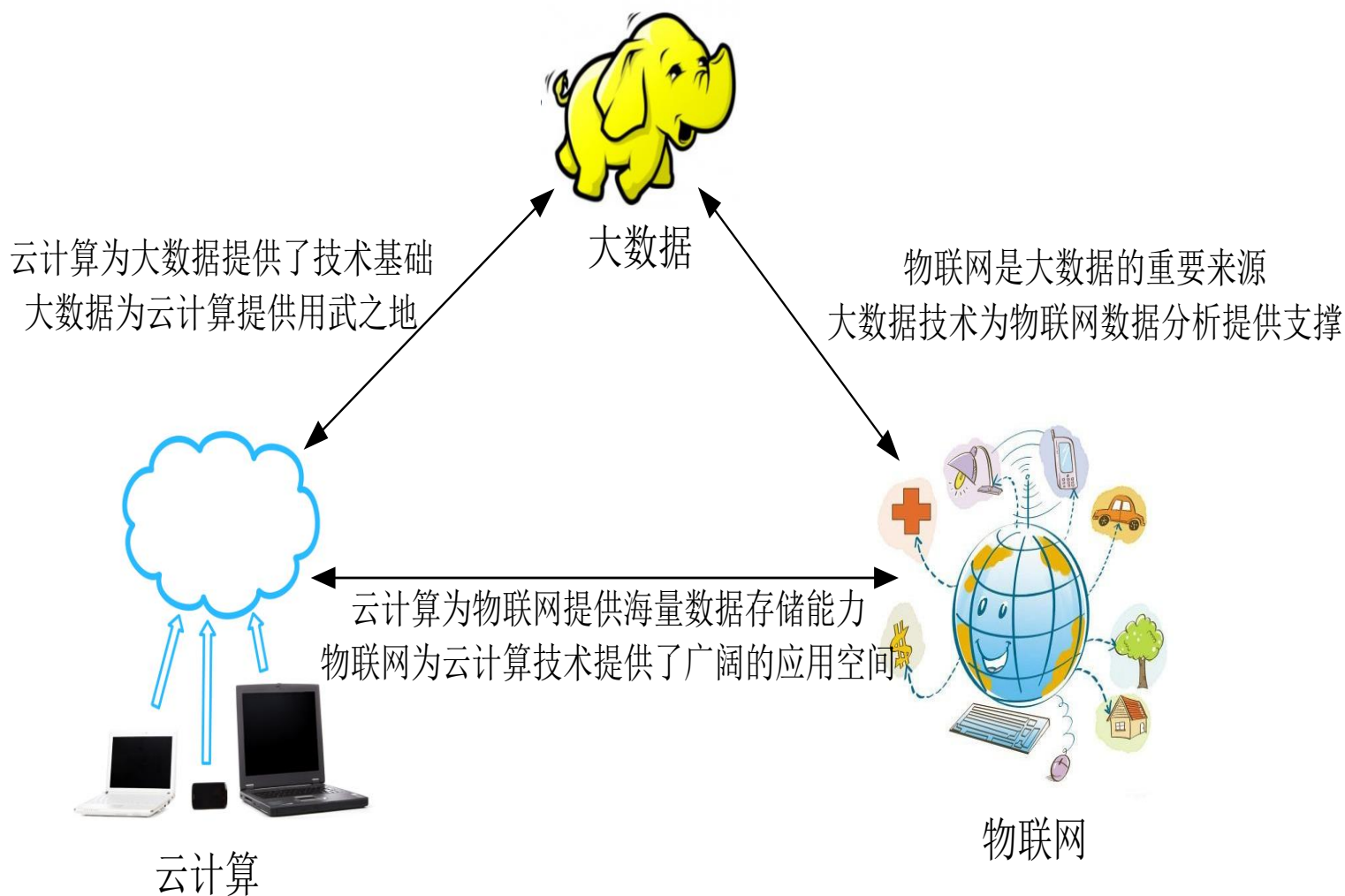
- **大数据根植于云计算，大数据分析的很多技术都来自于云计算。**云计算的分布式数据存储和管理系统提供了海量数据的存储和管理能力，分布式并行处理框架MapReduce提供了海量数据分析能力。
- **反之，大数据为云计算提供了“用武之地”，**没有大数据这个“练兵场”，云计算技术再先进，也不能发挥它的应用价值。



1.8.3 大数据与云计算、物联网的关系

□ 物联网的传感器源源不断产生的大量数据，构成了大数据的重要来源，没有物联网的飞速发展，就不会带来数据产生方式的变革，即由人工产生阶段向自动产生阶段，大数据时代也不会这么快就到来。

□ 同时，物联网需要借助于云计算和大数据技术、实现物联网大数据的存储、分析和处理。



本章小结

- 本章介绍了大数据技术的发展历程，并指出信息科技的不断进步为大数据时代提供了技术支撑，数据产生方式的变革促成了大数据时代的来临
- 大数据具有数据量大、数据类型繁多、处理速度快、价值密度低等特点，统称“4V”。大数据对科学研究、思维方式、社会发展、就业市场和人才培养等方面，都产生了重要的影响，深刻理解大数据的这些影响，有助于我们更好把握学习和应用大数据的方向
- 大数据在金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的社会各行各业都得到了日益广泛的应用，深刻地改变着我们的社会生产和日常生活

本章小结

- 大数据并非单一的数据或技术，而是数据和大数据技术的综合体。大数据技术主要包括数据采集、数据存储和管理、数据处理分析与展示、数据安全和隐私保护等几个层面的内容
- 大数据产业包括IT基础设施层、数据源层、数据管理层、数据分析层、数据平台层和数据应用层，在不同层面，都已经形成了一批引领市场的技术和企业
- 本章最后介绍了云计算和物联网的概念和关键技术，并阐述了大数据、云计算和物联网三者之间的区别与联系

Thank you all