媒体计算技术

李泽超

计算机科学与工程学院

大趋势

—Big Data

社交媒体、移动互联网、物联网的发展将我们带入大数据时代！

社会媒体数据集

数据源（Data Source）	数据类型（Data Type）	记录数（Records）
腾讯微博	图片	78957657
	Tweets	453265925
新浪微博	图片	100459981
	Tweets	274063154
Youtube	视频	92484
	评论	38226779
Flickr	图片	76437219
	评论	124222889
Amazon	图片	1933814
	评论	4615099
大众点评网	图片	2147772
	评论	5383189
Instagram	图片	955762

背景

数据爆炸式增长仍在继续！！！
大规模数据处理与知识挖掘： 重要！ 紧迫！

《国民经济和社会发展第十二个五年规划纲要》中指出：
“…，海量信息处理及知识挖掘的理论与方法，…”。

大数据带来了什么？？
为我所用！ 获得知识！

知识带来机遇

挖掘带来挑战

国内大数据

■**马云对未来的预测，是建立在对用户行文分析的基础上。**

“2008年初,阿里巴巴平台上整个买家询盘数急剧下滑，欧美对中国采购在下滑。海关是卖了货,出去以后再获得数据;而我们提前半年时间从询盘上推断出世界贸易发生变化了。”

■腾讯在天津投资建立亚洲最大的数据中心；百度也在投资建立大数据处理中心；

美国的大数据战略

- 2012年3月，美国奥巴马政府宣布投资2亿美元启动“大数据研发计划”，旨在提高和改进从海量和复杂数据中获取知识的能力，加速美国在科学和工程领域发明的步伐，增强国家安全。

- 这是继1993年美国宣布“信息高速公路”计划后的又一次重大科技发展部署，由美国国家科学基金会、能源部等6个联邦部门共同投资。

大数据时代的背景

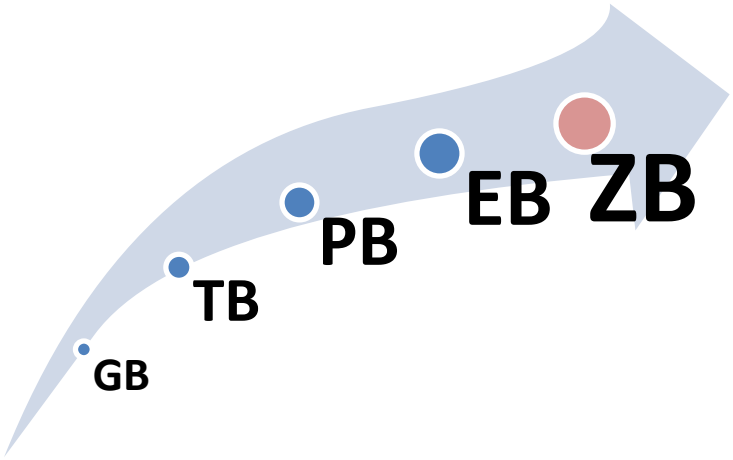
“大数据”的诞生：

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。信息爆炸的学科如天文学和基因学，创造出了“大数据”这个概念*。如今，这个概念几乎应用到了所有人类智力与发展的领域中。

21世纪是数据信息大发展的时代，移动互联、社交网络、电子商务等极大拓展了互联网的边界和应用范围，各种数据正在迅速膨胀并变大。

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。

数据大爆炸



地球上至今总共的数据量：

在**2006** 年，个人用户才刚刚迈进**TB**时代，全球一共新产生了约**180EB**的数据；

在**2011** 年，这个数字达到了**1.8ZB**。

而有市场研究机构预测：
到**2020** 年，整个世界的数据总量将会增长**44** 倍，达到**35.2ZB（1ZB=10** 亿**TB）！**

1PB = 2^50字节
1EB = 2^60字节
1ZB = 2^70字节

想驾驭这庞大的数据，我们必须了解大数据的特征。

大数据的4V特征

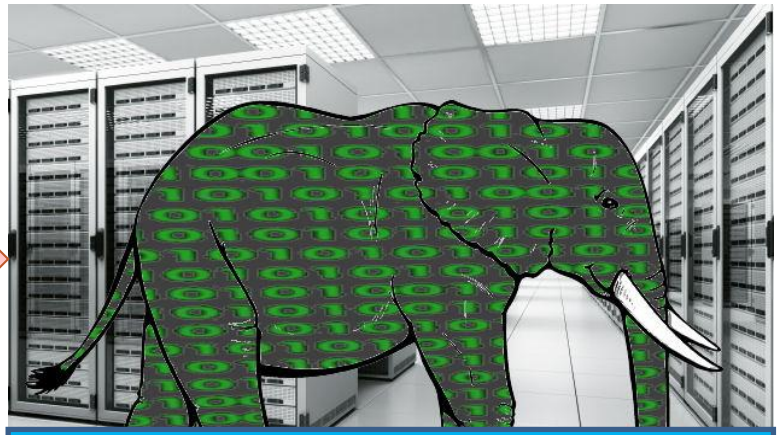
体量Volume	*非结构化数据*的超大规模和增长 总数据量的80~90% 比结构化数据增长快10倍到50倍 是传统数据仓库的10倍到50倍
多样性Variety	大数据的异构和多样性 很多不同形式（文本、图像、视频、机器数据） 无模式或者模式不明显 不连贯的语法或句义
价值密度Value	大量的不相关信息 对未来趋势与模式的可预测分析 深度复杂分析（机器学习、人工智能Vs传统商务智能(咨询、报告等）
速度Velocity	*实时分析*而非批量式分析 数据输入、处理与丢弃 立竿见影而非事后见效

*

密不可分的大数据与云计算

大数据是落地的云

商业模式驱动



应用需求驱动

云计算本身也是大数据的一种业务模式

- 云计算的模式是业务模式，本质是数据处理技术。
- 数据是资产，云为数据资产提供存储、访问和计算。
- 当前云计算更偏重海量存储和计算，以及提供的云服务，运行云应用，但是缺乏盘活数据资产的能力，挖掘价值性信息和预测性分析，为国家、企业、个人提供决策和服务，是大数据核心议题，也是云计算的最终方向。

大数据不仅仅是“大”

多大？
至少PB级

比大更重要的是数据的复杂性，有时甚至大数据中的小数据如一条微博就具有颠覆性的价值

大数据的应用不仅仅是精准营销

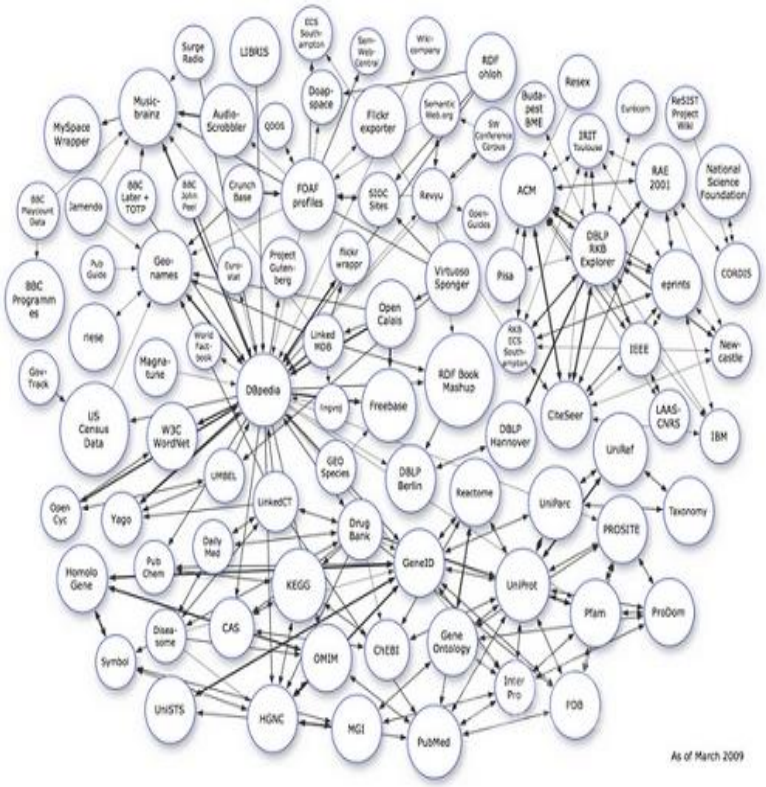
通过用户行为分析实现精准营销是大数据的典型应用，但是大数据在各行各业特别是公共服务领域具有广阔的应用前景

管理大数据“易”理解大数据“难”

虽然大数据是一个重大问题，真正的问题是让大数据更有意义

目前大数据管理多从架构和并行等方面考虑，解决高并发数据存取的性能要求及数据存储的横向扩展，但对非结构化数据的内容理解仍缺乏实质性的突破和进展，这是实现大数据资源化、知识化、普适化的核心

非结构化海量信息的智能化处理：自然语言理解、多媒体内容理解、机器学习等



一些相关技术

- **分析技术：**

- 数据处理：自然语言处理技术
- 统计和分析：A/B test; top N排行榜；地域占比；文本情感分析
- 数据挖掘：关联规则分析；分类；聚类
- 模型预测：预测模型；机器学习；建模仿真

- **存储**

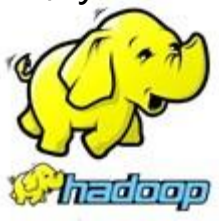
- 结构化数据：
  - 海量数据的查询、统计、更新等操作效率低
- 非结构化数据
  - 图片、视频、word、pdf、ppt等文件存储
  - 不利于检索、查询和存储
- 半结构化数据
  - 转换为结构化存储
  - 按照非结构化存储

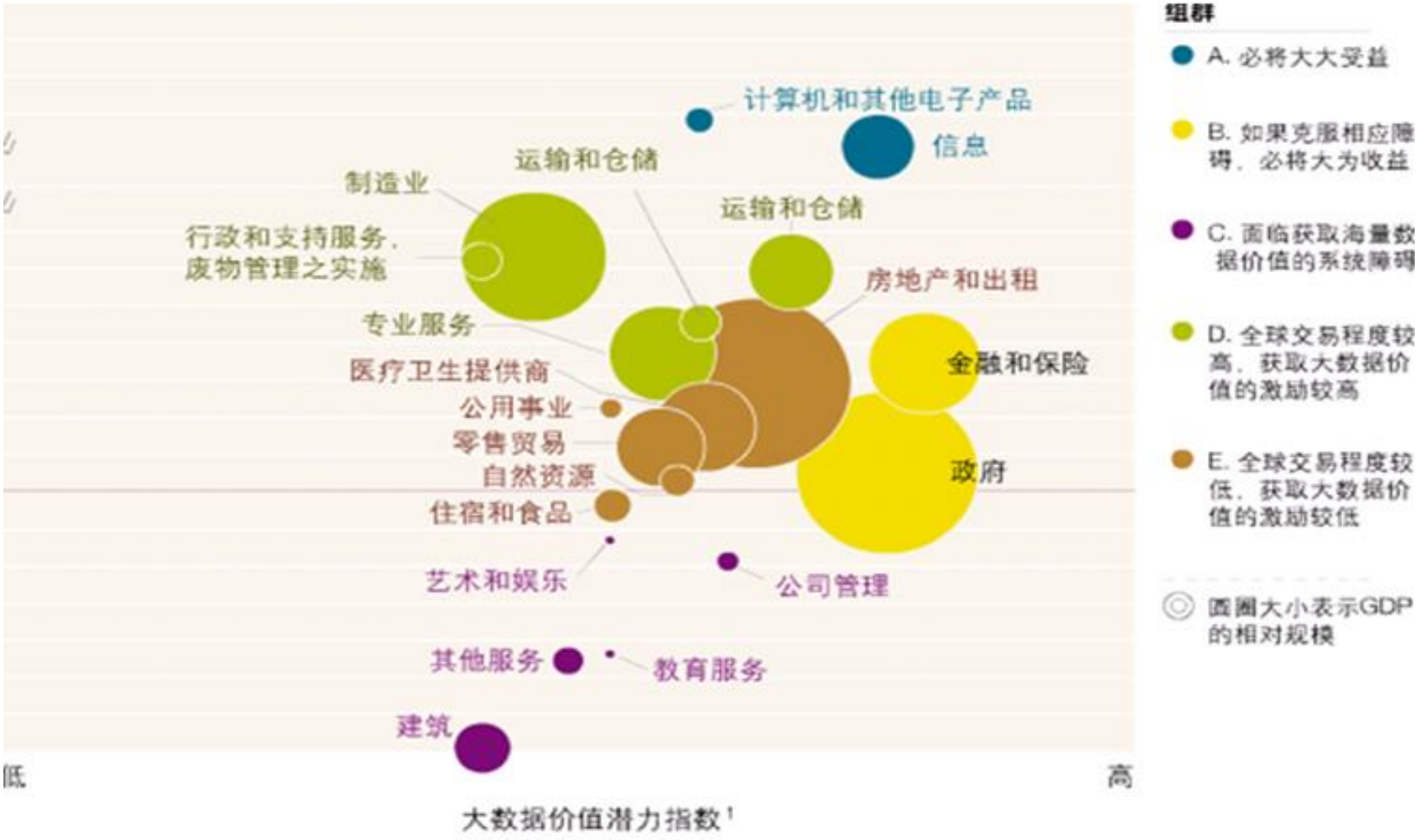
- **大数据技术：**

- 数据采集：ETL工具
- 数据存取：关系数据库；NoSQL；SQL等
- 基础架构支持：云存储；分布式文件系统等
- 计算结果展现：云计算；标签云；关系图等

- **解决方案：**

- Hadoop（MapReduce技术）
- 流计算（twitter的storm和yahoo！的S4）





- **政府、金融、电信等行业投资建立大数据的处理分析手段，实现综合治理、业务开拓等目标；应用到制造等更多行业。**

大数据的应用

——未来，改变一切

未来，企业会依靠洞悉数据中的信息更加了解自己，也更加了解客户。

数据的再利用：

由于在信息价值链中的特殊位置，有些公司可能会收集到大量的数据，但他们并不急需使用也不擅长再次利用这些数据。例如，移动电话运营商手机用户的位置信息来传输电话信号，这对以他们来说，数据只有狭窄的技术用途。但当它被一些发布个性化位置广告服务和促销活动的公司再次利用时，则变得更有价值。

大数据价值链的3大构成：数据本身、技能与思维

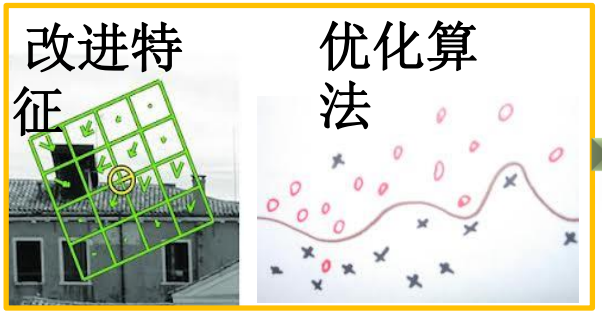
其中三者兼具的又谷歌公司，谷歌在刚开始收集数据的时候就已经有多次使用数据的想法。比方说，它的街景采集车手机全球定位系统数据不光是为了创建谷歌地图，也是为了制成全自动汽车以及谷歌眼镜等与实景交汇的产品。

传统行业最终都会转变为大数据行业，无论是金融服务业、医药还是制造业。

《大数据时代》

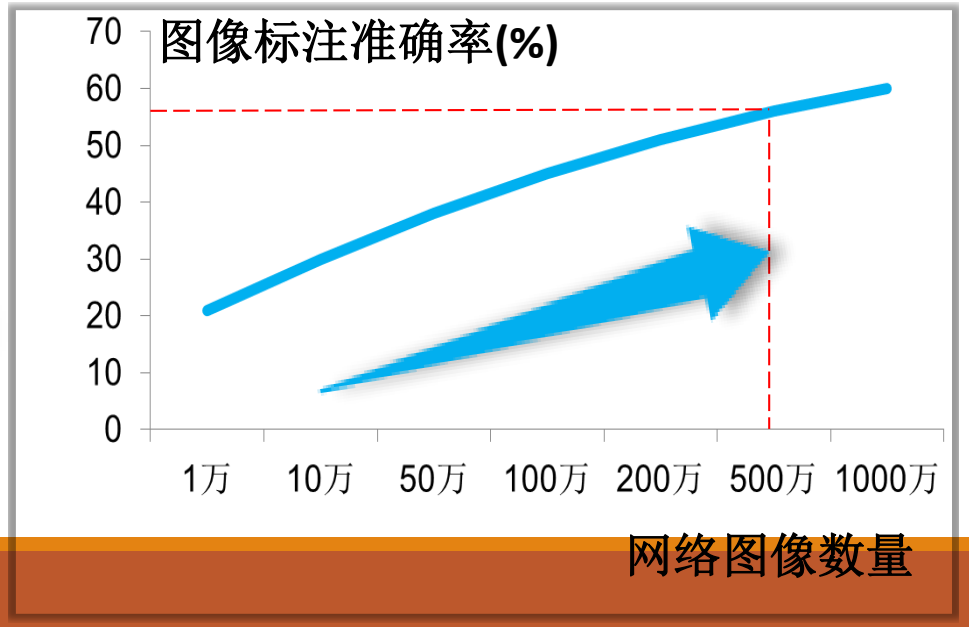
机遇： 媒体数据语义分析

语义鸿沟



20%

大规模
异质数据

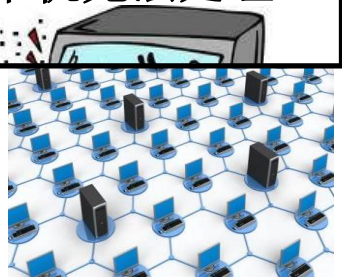


项目背景－挑战

大

TB量级数

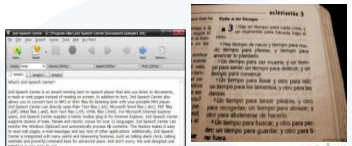
单机无法处理！



杂

多源

异质



文字

图像

多种形式

声音

视频

实例2-个性化服务与健康

XXX比萨店。您好，请问有什么需要我为您服务？请先告知会员号。

16846146***

X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？

你为什么知道我所有的电话号码？

X先生，因为我们联机到CRM系统

我想要一个海鲜比萨

……

X先生，海鲜比萨不适合您？

为什么？

根据您的医疗记录，你的血压和胆固醇都偏高。

那你们有什么可以推荐的？

您可以试试我们的低脂健康比萨。

实例2-个性化服务与健康

您上星期一在中央图书馆借了一本《低脂健康食谱》。

你怎么知道我会喜欢吃这种的？

好。那我要一个家庭特大号比萨，要付多少钱？

99元，这个足够您一家六口吃了。但您母亲应该少吃，她上个月刚刚做了心脏搭桥手术，还处在恢复期。

那可以刷卡吗？

X先生，对不起。请您付现款，因为您的信用卡已经刷爆了，您现在还欠银行4807元，而且还不包括房贷利息。

那我先去附近的提款机提款

陈先生，根据您的记录，您已经超过今日提款限额。

你们直接把比萨送我家吧，家里有现金。多久会送到？

大约30分钟。如果您不想等，可以自己骑车来

为什么？

根据我们CRM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。

实例3-政府管理

关注民生，提高服务水平——药品食品安全

问题与困难

社会困境：

各种食品药品问题层出不穷：恒天然奶粉、铬大米、有毒疫苗、H7N9，处理不当容易快速演化、影响公众对食品药品的信任

政府关注：

哪些事件可能演化为食品药品热点，事件间有何关系，出现事件怎样处理，社交网络如何引导，效果如何评估、经验如何总结

决策建议

处置：

挖掘出与铬中毒症状相关，政府在社交网络进行引导，防止恐慌，通过下级管理机构推动线下的铬中毒预防与防治

。。。

事件发现

事例：

某地区网民在社交论坛一段时间内多人次咨询皮肤损害、胃肠道疾病症状

。。。

食品药品监督管理应用系统

政府数据 

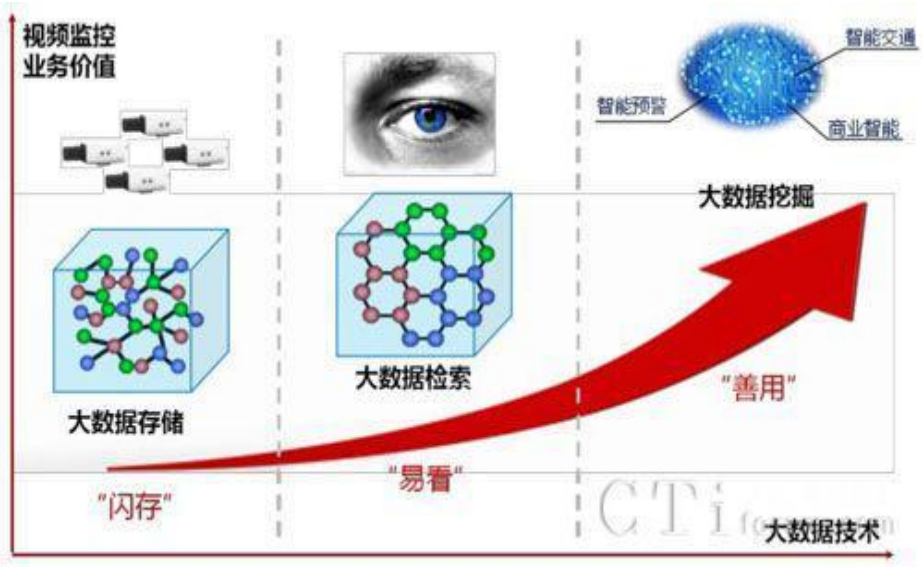
互联网数据 

行业数据 

实例4-平安城市（视频监控）

基于视频监控大数据，分析与整合，理解以下场景

“红衣人，走下小汽车，走进银行， 5分钟后走出来”



实例5-金融

应用：客户管理、营销管理、风险管理

我们处在一个“数字宇宙”中

如果将2008年全球数据总量印成书的话，将它们排列起来，其长度是地球到银河系的10倍。这些数据中间，70%、80%都是数字媒体，促进了传统产业的升级改型。

图像视频在其中起到了很大的作用，包括监视视频和在线下载音视频等等。根据2012年IDC的调查显示：监控视频在大数据中占的比例，2010年大概有一半，2015年占三分之二，预测到2020年大概占40%左右。

“信息海洋”带来的三大挑战

面临三大挑战：

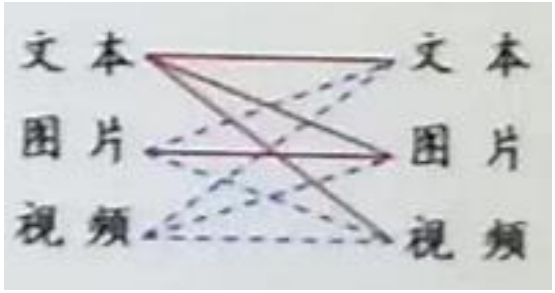
- 存不起
- 查不准
- 管不住

数字媒体获取——存不起

- 据IDC公司统计，人类有史以来所有印刷材料的数据总量仅200PB（10^9MB）
- 2014年全球新增加的数据高达4.1ZB （10^{15}MB），其中超过75%来自个人（多媒体数据：图像、视频和音乐）
- 如果用1TB的硬盘来存储，需要12300亿元左右来买这些硬盘。

数字媒体使用——查不准

- Google、百度等商业搜索引擎，支持文本检索，以图搜图
- 跨媒体能力有限



- 以图搜图离预期尚有差距，“语义鸿沟”
- 比如我们拿“毕福剑事件”中这张视频图检索出来的相关结果，差距还是比较远。

媒体内容监管——管不住

- 数字媒体传播速度快、信息量大、内容丰富、互动性、影响力大

- 从“毕福剑事件”来看，社会发展监管存在管不住的问题。首先，现在的传播速度非常快，一个热点事件及照片、视频出来以后，马上就是全国人民基本上就都知道了。这也导致一些危及到社会治安和国家安全的东西很难监管。

发展趋势

- 真三维
- 建立在云端的媒体内容分析与服务
- 个性化与强互动结合是媒体消费的新模式。

发展趋势

真三维是未来的发展趋势

- 三维产业逐渐成为信息科技升级和自主创新的主要驱动力
- 高速高精度高质量的场景刻画、建模与测量是研究热点
- 先进制造、3D打印：高精度实时采集与建模速获取
- 数字城市：大规模场景的三维模型快速获取

- 真三维的视频帮助人们对物质世界实现高逼真度的感知。另外从研究的角度，二维跳到三维很多东西很难实现的，比如阴影怎么去避免等等。从全球来看，真三维视频是未来科技竞争的一个焦点，也是未来的发展趋势。

发展趋势

建立在“云端”的媒体内容分析与服务

- 数字媒体数据量大、内容分析关联复杂，越来越多的分析和服务需要到“云端”
- 云计算将大量媒体资源统一管理和调度，构成一个计算平台向用户按需服务
- 提高媒体内容搜索、推荐、舆情监测、个性化分析等服务的精准度，降低数据处理的重复工作
- 媒体内容的分析关联是十分的复杂，应用端要做复杂的处理是很困难的，而且要做深度的分析理解，我们只能把它建立在云这一级来进行处理、分析。包括现在的电视处理也都是这样的。

发展趋势

个性化与强互动结合是媒体消费的新模式

- 广播电视：单向服务，不能提供个性化的内容；频道多，难以快速找到需要的内容
- 网络电视：视频点播、内容搜索、更灵活机动
- 社交网络：兴趣相同的群体进行互动社交活动
- 发展趋势：个性化+社交化

数字媒体应用的机会和难点

新一代电视媒体技术

互动电视技术

- 促进传统电视产业的升级转型：交互性更强、业务形式丰丰富
- 经济效益

真三维视频技术

- 支持高清裸眼三维电视产业革命

数字媒体应用的机会和难点

媒体内容的理解分析

- 海量数字媒体内容分析与理解是解决媒体内容管理与应用的关键技术，对基于语义的媒体监管、搜索与服务等相关产业的发展具有职称作用

数字媒体应用的机会和难点

面向移动设备媒体内容的分析与挖掘

- 移动互联网用户增长趋势很快，这里边使用的相当多的内容是数字媒体内容。这面临一些新的问题：显示屏小、传输带宽处理能力较弱点等等

云计算

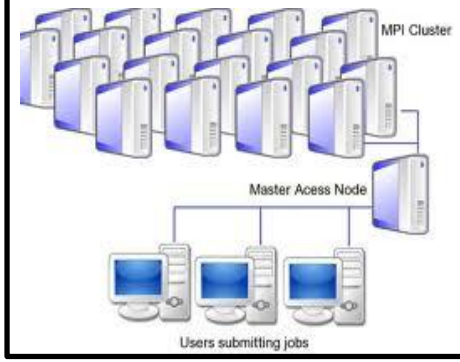
大

杂

多源

异质

并行平台，云计算



大数据

转化

知识

云

IT不再重要

互联网大转换的制高点—云计算

【美】尼古拉斯·卡尔 著

“卡尔的观点的确令人惊恐！但在我们如火如荼地发展计算机产业的时候，思考一下它的另一面还是个值得肯定的想法。”

——《商业周刊》

在各大企业拼力角逐持续变革的数字化时代时，卡尔对IT产业的“挑衅”将深刻影响各大企业的董事会、CEO们以及支持他们的投资者们。

——《华尔街日报》

在本书中，卡尔做了深入的分析与研究，他的观点非常有说服力，也很有权威。他的推断有条不紊，结论大胆合理。是一部极其出色的作品。

——《科技世界》

“不论这场大论战最后的答案是什么，整个经济的风貌都将因此而改变”、“这是IT历史上最重大的一场论战！”

——《纽约时报》

云计算概念

云计算概念（Cloud Computing）是最早由Google提出的，微软和其他公司又提出了自己云计算架构。

云计算是指服务的交付和使用模式，指通过网络以按需、易扩展的方式获得所需的服务。这种服务可以是IT和软件、互联网相关的，也可以是任意其他的服务，它具有超大规模、虚拟化、可靠安全等独特功效。

云计算是一种**服务模式**。

简单的类比

说您刚刚搬到一个新的城市，要找个地方住……

您可以……

建一座房子……

或

租一座房子……


如果选择建一座房子，您必须作出一
系列的抉择……

物业费

雇员 装饰设计师

电费 水费

管道工

电工

空调

清洁工人

假设您所在的城市提供这样一种房屋：

一个拥有 大量 公寓单元的房屋

每个单元可以 *容易地* 被转换成

2,3,4 或更多的单元

只需付出 租金

对应使用功能！

云计算：21世纪的全球化信息电厂

和中央电厂一样，云计算的核心是让计算服务和运营能力像电一样能够长距离传输，随需索取。 这将会对IT业以及每家公司的结构造成深刻变革。 100年来的故事或许将在未来重演……

电的历史

1860年代	西门子发明了可实用的发电机，商业化发电机开始应用到生产中
1870年代	爱迪生发明了可实用的电灯，并同时开设社区电厂，电开始影响人类的生活
1890年代	长距离输电网络技术逐渐成熟，并开始商用
1900年代	第一个城市中央电厂在芝加哥出现，并迅速得到了广泛的应用
20世纪	电的应用主宰了人类社会，各种电器出现，催生了史上最大的家用电器行业，人类社会开始有了现在的家庭生活

IT的历史

1960年代	集成电路计算机出现，计算机在商业中迅速普及得到了广泛应用
1970年代	微型计算机出现，并在其后20年间开始影响人类的生活
1990年代	Internet网络技术在全球迅速普及，广泛应用
2000年代	Google、Amazon 等开创了云计算的商业化应用， 实现了IT能力的远距离传输
21世纪	云计算的应用是否会主宰人类社会？ 人类社会将有怎样崭新的生活？

如果你理解了中央电厂的来由和它为人类社会带来的历史性变化，那么，你就会明白云计算将会开创多么激动人心的未来。

云计算能够做什么

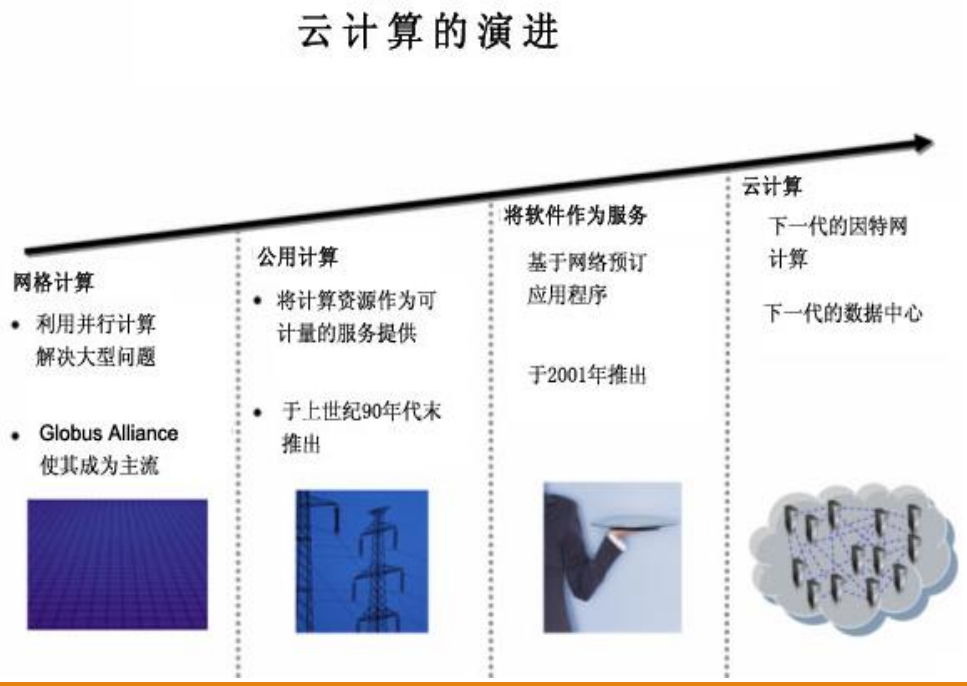
“基于互联网的商业计算模型”，利用高速互联网的传输能力，将数据的处理过程从个人计算机或服务器移到互联网上的服务器集群中。这些服务器由一个大型的数据处理中心管理着，数据中心按客户的需要分配计算资源，达到与超级计算机同样的效果。

“云”会替用户做存储和计算的工作

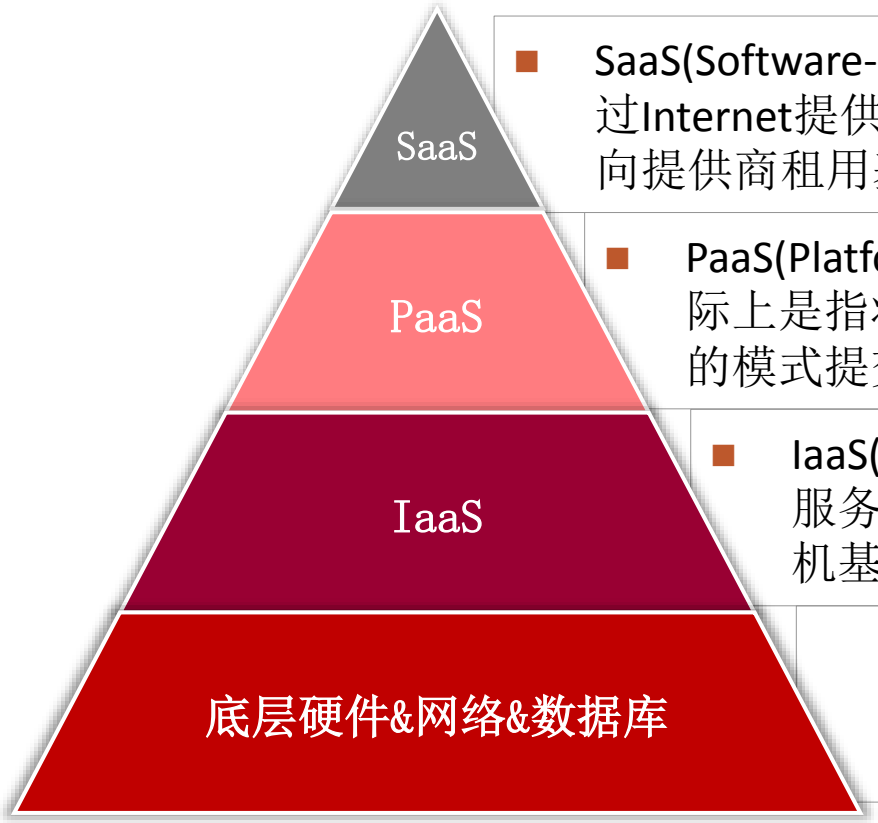
- 只需要一台能上网的电脑，不需关心存储或计算发生在哪朵“云”上，但一旦有需要，我们可以在任何地点用任何设备，如电脑、手机等，通过网络服务快速地计算和找到资料，甚至实现超级计算这样的任务。
- 用户再也不用担心资料丢失
- 从这个角度而言，最终用户才是云计算的真正拥有者

云计算发展演变

云计算是由分布式计算（Distributed Computing）、并行处理（Parallel Computing）、网格计算（Grid Computing）发展来的，是一种新兴的商业计算模型。

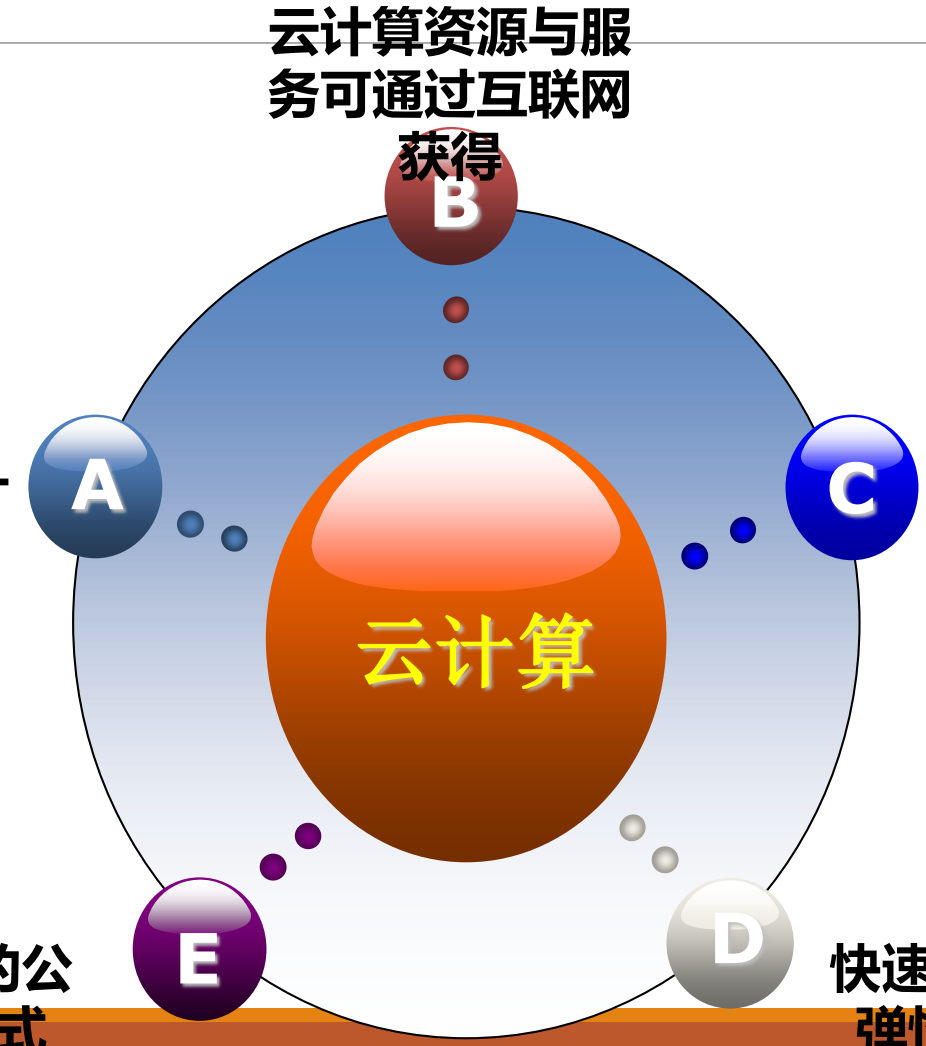


云计算服务的分类



- SaaS(Software-as-a- Service)：软件即服务。它是一种通过Internet提供软件的模式，用户无需购买软件，而是向提供商租用基于Web的软件，来管理企业经营活动。
- PaaS(Platform-as-a- Service)：平台即服务。PaaS实际上是指将软件研发的平台作为一种服务，以SaaS的模式提交给用户。
- IaaS(Infrastructure-as-a- Service)：基础设施即服务。消费者通过Internet可以从完善的计算机基础设施获得服务。

云计算的特征


云计算资源与服务可通过互联网获得
B
可按需获得云计算资源与服务
A
C
绿色、安全的规模化数据中心提供的共享资源池与服务池
云计算
E
按需付费的公用服务模式
D
快速伸缩并且弹性易扩展

云计算特点（1）

超大规模

- Google云计算已经拥有100多万台服务器，Amazon、IBM、微软、Yahoo等的“云”均拥有几十万台服务器。企业私有云一般拥有数百上千台服务器。

虚拟化

- 云计算支持用户在任意位置、使用各种终端获取应用服务。所请求的资源来自“云”，而不是固定的有形的实体。应用在“云”中某处运行，但实际上用户无需了解、也不用担心应用运行的具体位置。只需要一台笔记本或者一个手机，就可以通过网络服务来实现我们需要的一切，甚至包括超级计算这样的任务。

云计算特点（2）

高可靠性

- “云”使用了数据多副本容错、计算节点同构可互换等措施来保障服务的高可靠性，使用云计算比使用本地计算机可靠。

高可扩展性

- “云”的规模可以动态伸缩，满足应用和用户规模增长需要。

通用性

- 云计算不针对特定的应用，在“云”的支撑下可以构造出千变万化的应用，同一个“云”可以同时支撑不同的应用运行。

云计算特点（3）

按需服务

- “云”是一个庞大的资源池，按需购买；云可以象自来水，电，煤气那样计费。

极其廉价

- 由于“云”的特殊容错措施可以采用极其廉价的节点来构成云，“云”的自动化集中式管理使大量企业无需负担日益高昂的数据中心管理成本，“云”的通用性使资源的利用率较之传统系统大幅提升，因此用户可以充分享受“云”的低成本优势，经常只要花费几百美元、几天时间就能完成以前需要数万美元、数月时间才能完成的任务。

我国云计算发展现状

2008年5月10日，IBM在中国无锡太湖新城科教产业园建立的中国第一个云计算中心投入运营

2008年6月24日，IBM在北京IBM中国创新中心成立了第二家中国的云计算中心——IBM大中华区云计算中心

2008年11月28日，广东电子工业研究院与东莞松山湖科技产业园管委会签约，广东电子工业研究院将在东莞松山湖投资2亿元建立云计算平台

2008年12月30日，阿里巴巴集团旗下子公司阿里软件与江苏省南京市政府正式签订了2009年战略合作框架协议，计划于2009年初在南京建立国内首个“电子商务云计算中心”，首期投资额将达上亿元人民币

世纪互联推出CloudEx产品线，包括完整的互联网主机服务"CloudEx Computing Service"，基于在线存储虚拟化的"CloudEx Storage Service"，供个人及企业进行互联网云端备份的数据保全服务等系列互联网云计算服务

中国移动研究院做云计算的探索起步较早，已经完成了云计算中心

2008年11月25日，中国电子学会专门成立云计算专家委员会

2009年5月22日，中国电子学会在北京中国大饭店举办首届中国云计算大会

云计算离我们有多远？

云计算中还有很多问题没有解决

- 安全风险
  - 根据Gartner的说法，云计算充满了安全风险。云计算的独特属性要求在诸如数据的完整性、数据恢复和隐私保护等方面予以风险评估，同时还要求在一些司法领域如电子证据、法规遵从和审计等方面进行评估。
  - 1 特权用户的访问权限
    - 在企业之外处理敏感数据会带来内在的风险，因为外包服务一般都会绕过企业IT部门对内部所施加的“物理的、逻辑的和人员控制”
  - 2法规遵从
    - 客户对于他们自己数据的安全和完整性是要负最终责任的，即便是把数据交由外部的服务供应商托管也是如此

云计算离我们有多远？

云计算中还有很多问题没有解决

- 安全风险
  - 3 数据的存放位置
    - 客户在使用云计算时，可能无法确切地知道你的数据到底被托管在什么地方，以及服务商是否会按照客户的要求通过合同约定以遵守当地的隐私保护条例
  - 4. 数据隔离
    - 数据在云中通常是在一个和其他客户的数据共享的环境中。加密虽然是有效的，但任何加密上的意外，都有可能导致数据完全无法使用，即便正常的加密也可能使数据的可用性变得相当复杂
  - 5. 数据恢复
    - 任何无法提供异地数据与应用基础设施复制的服务都很容易完全失效 。另外，考虑到云计算应用程序具有大规模分布式的特性，要明白出现了哪些种类的故障、出现在何处也许并非易事

云计算离我们有多远？

云计算中还有很多问题没有解决

- 网络制约
  - 云计算依托于互联网，互联网的状况也决定着云计算的发展程度：
  - a. 对企业带宽的要求
    - 如果应用云计算，那么就必须将目前在局域网中完成的工作放到互联网上去。企业可以很容易的构建一个100M带宽的局域网，却很难拥有一个100M带宽的互联网。云计算应用得越多，对带宽的要求就越高
  - b. 性能价格比问题
    - 云计算供应商有能力构建一个快速的“云”，但无法帮助用户修建一条通往“云”的高速路。数据在提供商与用户之间能跑多快，取决于这条路有多宽多平整。目前国内网络状况不理想，宽带费用也不低，企业想要应用云计算还需要额外支出一笔不菲的费用。