



第 1 讲

大数据概述

了解大数据



最火IT词汇之一



了解大数据



“反正你接受它也来了，不接受它也来了，接受不接受大数据时代它都带着诚意扑面而来。”

引用自小品《功夫》

了解大数据



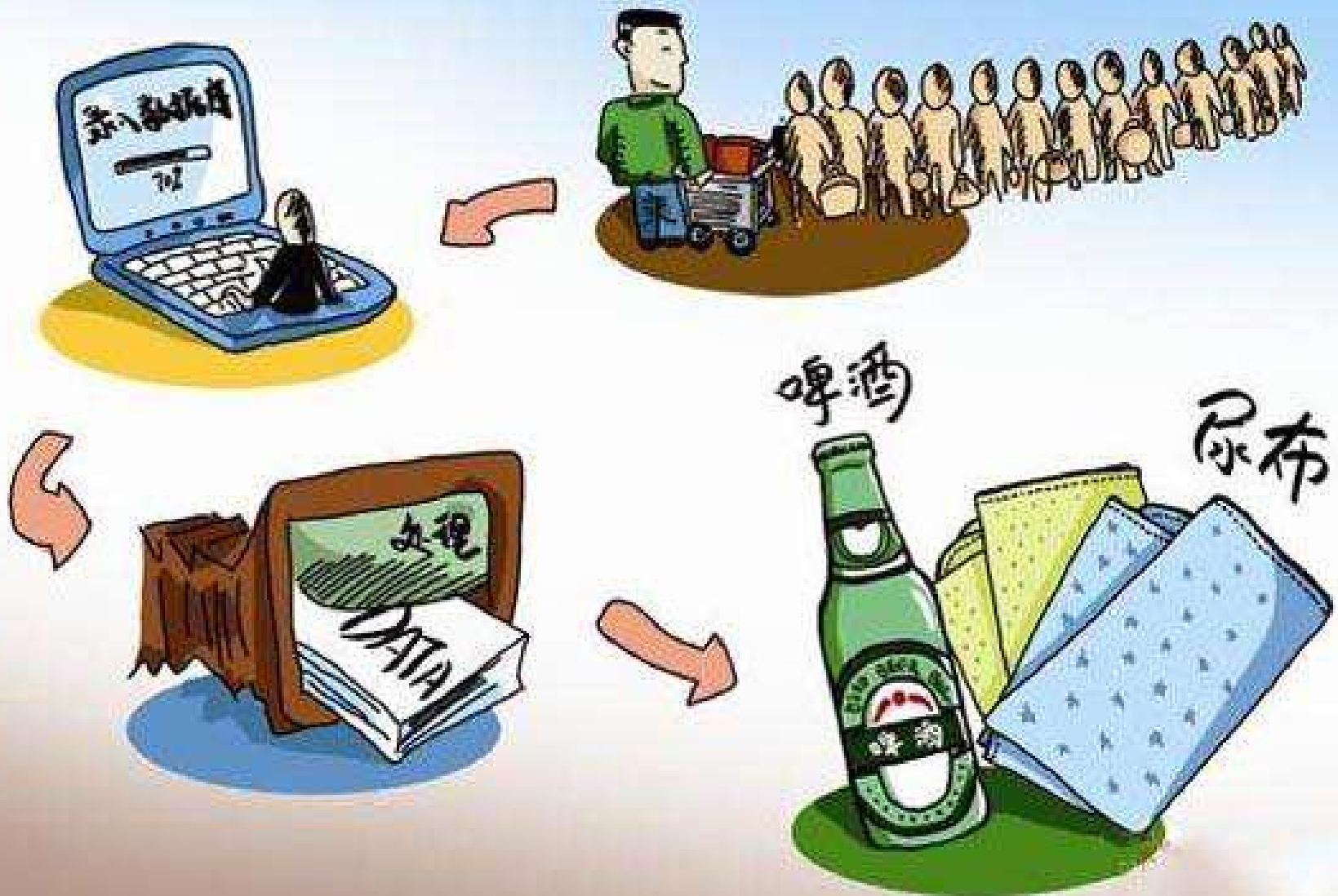
频频出现在《纽约时报》、《华尔街日报》的专栏封面，多次进入美国白宫官网的头条新闻。

了解大数据



“大数据”经典案例
啤酒与尿布

了解大数据



了解大数据



了解大数据



“大数据”经典案例
数据新闻让英国撤军

了解大数据



超市预知高中生顾客怀孕

1. 大数据让政府治理更精准透明



谷歌流感趋势

谷歌工程师认为，搜索流感信息的人数与实际患病人数之间存在密切关联。

设计人员编入流感关键词，如温度计、流感症状、肌肉疼痛、胸闷等

只要用户输入这些关键词，系统就会展开跟踪分析，创建地区流感图表和流感地图

预测出世界上不同国家和地区的流感传播情况

2009年，甲型H1N1流感暴发的几周前，“谷歌流感趋势”成功预测了流感在美国境内的传播，其分析结果甚至具体到特定的地区和州，并非常及时，令公共卫生官员备感震惊。而传统上，美国疾病控制中心要在流感暴发一两周之后才可以做到这些。

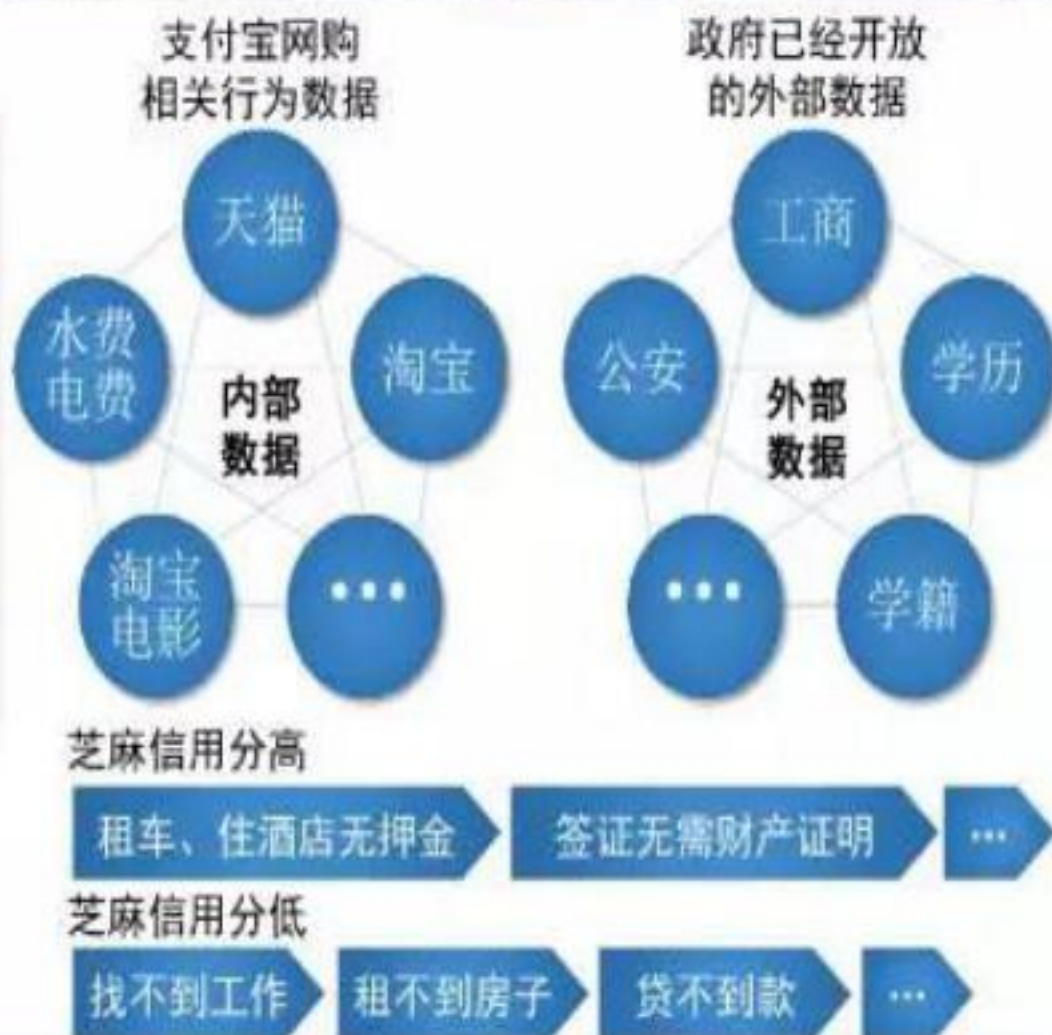


谷歌一周疫情报告

2. 大数据让经济治理更有效



支付宝“芝麻信用”——
“芝麻信用分”，授权开通后，每个支付宝用户都可以看到自己的芝麻信用分。分数越高代表信用程度越好，违约可能性越低。



芝麻信用分	有机会做的事				
高于 600 分且无不良记录	免押金租用永安城市自行车	阿里旅行多间酒店享受信用住	阿里旅行深圳华侨城先旅游后付费	相寓租房减免押金	享受花呗额度
高于 650 分且无不良记录	优拜单车免押金用车、神州租车、一嗨租车免押金租车	来分期申请线上极速贷款			
高于 700 分且无不良记录	方便申请新加坡签证	可以使用支付宝 A pp扫一扫小蓝单车上的二维码，即可下载免押金骑行。			

3. 大数据让公共服务更智慧

高德公司基于位置服务大数据的能力，与乌镇、古北水镇两家旅游公司合作，上线了全国首个“智慧景区”服务，解决游客在景区容易遇到的迷路、拥堵、排队等问题。



高德导航

地图
渲染

高德在地图上增加
游览车、游船的
线路地址，增加了
重要景点的渲染图

分类
筛选

商店、卫生间
餐厅、灯景区
重要地点信息
一目了然

导游
语音

当游客走到某个
对应景点附近，
导游语音会自动
播放

智慧
景区

引入热力图，游客
可以看到该处游客
人数的多少，合理
安排游览时间

4. 大数据让商业创新更迅猛



北美最大的付费订阅视频网站——Netflix

2012年，Netflix准备推出自制剧。不过在决定拍什么、怎么拍上，Netflix推出了自己的秘密武器——大数据。

用大数据拍自制剧

01 收集

收集该网站上用户每天产生的行为，如收藏、推荐、回放、暂停等，还包括用户的搜索请求等。

02 预测

分析出凯文·史派西、大卫·芬奇和“BBC出品”这三种元素结合在一起的电视剧产品将会大火。

03 拍摄

融合三者拍摄了一部《纸牌屋》，结果大获成功，成为了2013年全球最火的美剧。

了解大数据



2014年3月“大数据”
首次写入《政府工作报告》
2017年3月第四次写入

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

了解大数据

排名	文件名称	发文单位	发文日期	关注度
1	《大数据产业发展规划（2016-2020年）》	工信部	2016年12月30日	92.11
2	《关于促进和规范健康医疗大数据应用发展的指导意见》	国务院办公厅	2016年6月24日	82.94
3	《农业农村大数据试点方案》	农业部	2016年10月14日	78.34
4	《关于推进交通运输行业数据资源开放共享的实施意见》	交通部	2016年9月2日	69.68
5	《关于加快中国林业大数据发展的指导意见》	林业局	2016年7月13日	53.89
6	国家林业局落实《促进大数据发展行动纲要》的三年工作方案	林业局	2016年2月24日	50.64
7	《生态环境大数据建设总体方案》	环保部	2016年3月8日	43.65
8	《促进大数据发展三年工作方案（2016-2018）》	国家发改委等部委	2016年4月13日	40.93
9	《促进国土资源大数据应用发展实施意见》	国土资源部	2016年7月4日	32.11
10	《关于推进全国发展改革系统大数据工作的指导意见》	国家发改委	2016年9月9日	30.56

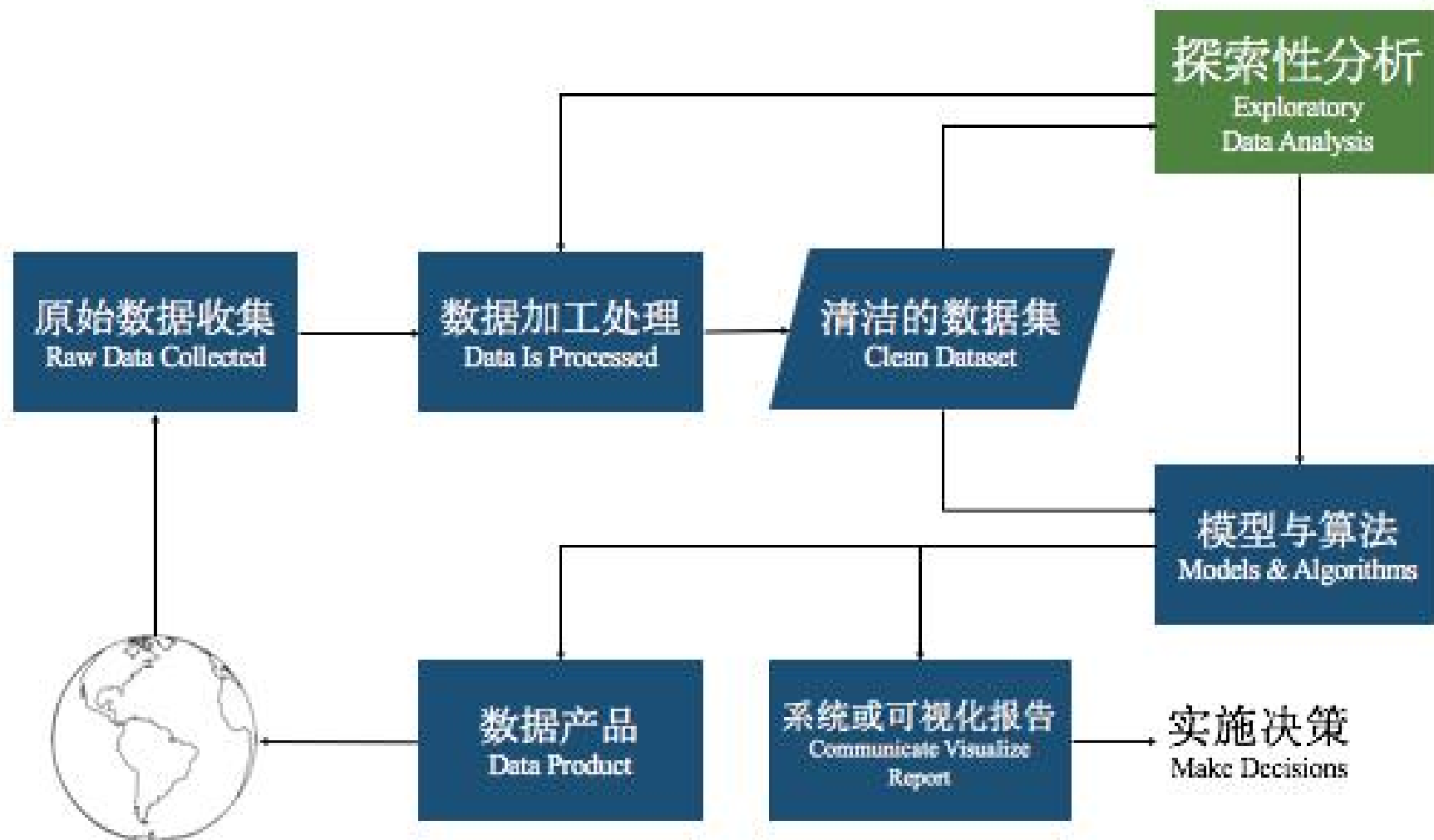
2016年中央及部委大数据领域最受关注的十大政策



马云说“我们已从IT时代进入了DT时代，未来我们的汽车、电灯泡、电视机、电冰箱等将全部装上操作系统，并进行数据集成，数据将会让机器更“聪明”。DT时代，数据将成为主要的能源，离开了数据，任何组织的创新都基本上是空壳。”

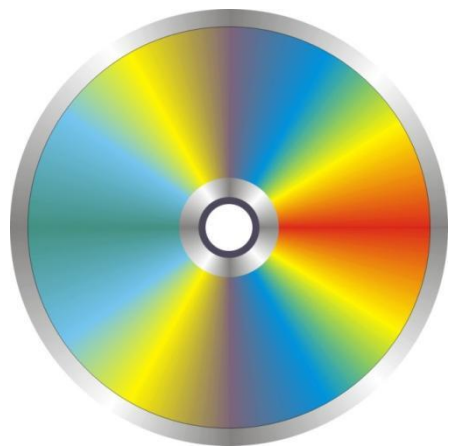
数据，是未来的一切！

数据科学的方法和流程



大数据有多大？

互联网每天产生的流量信息
可以装满 1.68 亿张碟



大数据有多大？

每天发出2940 亿封邮件，
每天的社区论坛上发出200万个帖子，
每天世界各地有 1.72 亿人访问 Facebook，
每天2 000 万人访问Google+，
每天在Facebook 上耗费的时间总计 47 亿分钟，
每天在Netflix观看2 200 万小时的电视电影节目，
每天人们将 86.4万小时视频上传到YouTube，
每天卖出 37.8 万台手机

.....





13000+个
App下载



13.6万条
垃圾短信拦截



8万
净收入



152台
PC售出



iOS

25部



176部



13.9万张
照片上传



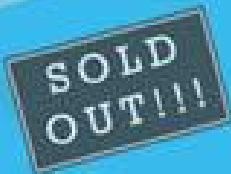
8.5万元
收入



7.3万笔
交易数
(2012双11)



486笔
订单



1042笔
订单
(2012双11)



10笔
订单

14.8万
独立访客



460万
网页打开



1.1万GB
文件下载

搜狐IT
it.sohu.com



9.5万条
微博发送



347万次
搜索

1.25万
浏览人数



265万张
订单



1.4万
净收入



24.6万
净收入

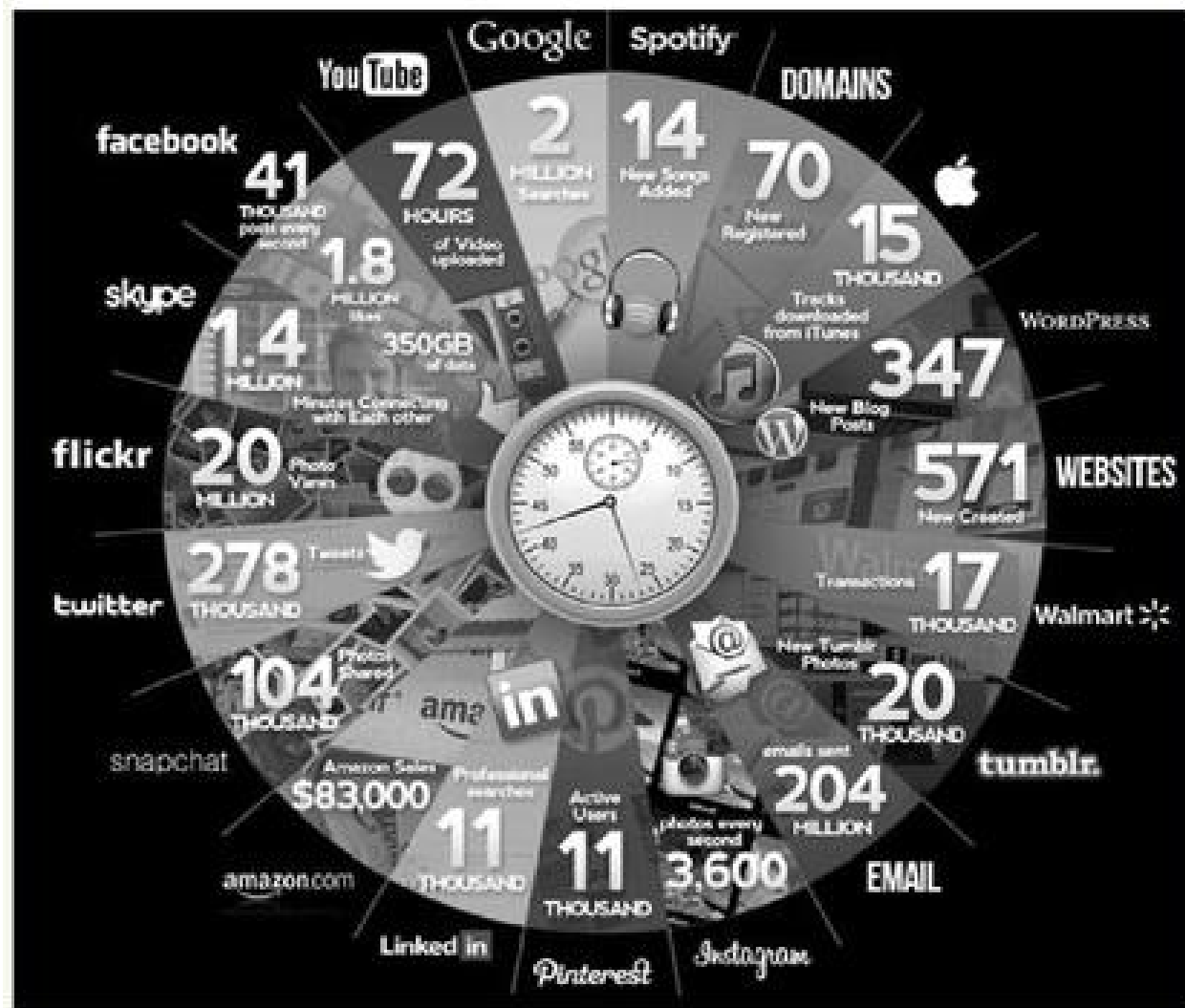


2.8万
净收入



搜狐IT
it.sohu.com
@iPhone一姐

大数据有多大？



(b) 国外互联网

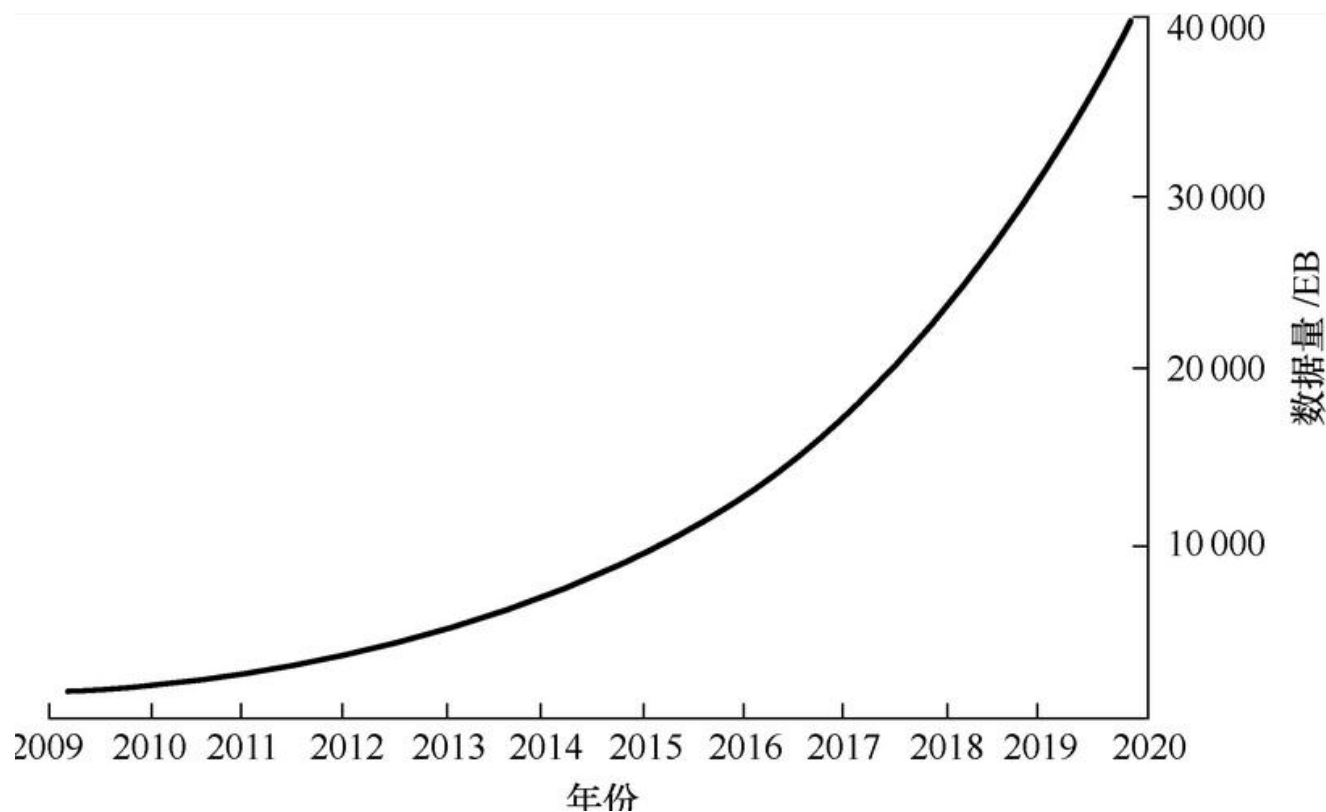
大数据有多大？

存储单位	换算关系	含义与实例
bit（比特或位）		1bit 是指一个二进制数（1 或 0），这是信息存储的基本逻辑单元
B（字节）	1 B=8 bit	这是信息存储的基本物理单位，在计算机内存储 1 个汉字，其大小是 2B
KB（千字节）	1 KB=1024 B= 2^{10} B	1 KB 相当于 512 个汉字
MB（兆字节）	1 MB=1024 KB= 2^{20} B	由崔子格演唱的 MP3 格式的《老婆最大》约为 4 MB
GB（吉字节）	1 GB=1024 MB= 2^{30} B	由郭敬明导演的高清电影《小时代 I》约为 1 GB
TB（太字节）	1 TB=1024 GB= 2^{40} B	eBay 每天产生的数据量约为 50 TB
PB（拍字节）	1 PB=1024 TB= 2^{50} B	Google 每月需要处理的数据量高达 600 PB
EB（艾字节）	1 EB=1024 PB= 2^{60} B	美国的医疗数据量约为 150 EB
ZB（泽字节）	1 ZB=1024 EB= 2^{70} B	2013 年全球数据量预计达到 4 ZB
YB（尧字节）	1 YB=1024 ZB= 2^{80} B	2029 年全球数据量预计达到 1 YB

大数据有多大？

《2020年的数字宇宙》研究报告

来自IDC (International Data Corporation , 国际数据公司)



大数据从哪里来？

赵老师，我不想知道大数据是怎么没的，我就想知道大数据是怎么来的！

你算来着了！咱们今天就学习大数据是怎么来的。



大数据从哪里来？

1、泛互联网

使信息和服务通过当下可能的技术和手段在计算设备、通信设备、机器、人之间传递和交付的网络，包括物联网、移动互联网和车联网等。

物联网：通过射频识别（Radio Frequency Identification，RFID）、红外感应器、全球定位系统、激光扫描器等信息传感设备，按约定的协议，把任何物体与互联网连接起来，进行信息交换和通信，以实现智能化识别、定位、跟踪、监控和管理的一种网络。

大数据从哪里来？

1、泛互联网



物与物、物与人的连接

大数据从哪里来？

1、泛互联网

物联网：农牧业-射频识别耳标



大数据从哪里来？

1、泛互联网

移动互联网：

移动接入已经将世界带到互联网的下一站——移动互联网。



大数据从哪里来？

1、泛互联网 移动互联网

中国互联网络信息中心（CNNIC）在北京发布第40次《中国互联网络发展状况统计报告》

中国网民规模和互联网普及率



大数据从哪里来？

1、泛互联网 车联网

由车辆位置、速度和路线等信息构成的巨大交互网络。

大数据从哪里来？

1、泛互联网 车联网

谷歌无人驾驶汽车每秒产生约1 GB 数据。

≈用计算机上传100 张高清数码相片



大数据从哪里来？

2、工业互联网

一种开放的全球化网络，它将人、数据和机器连接起来，目标是升级关键的工业领域。

如：在机器上安装传感器，机器启动，传感器开始采集数据，并将机器的快慢、歪斜等状态一五一十地形成数据，传到云端进行存储、分析与决策。

大数据从哪里来？

2、工业互联网

医疗保健：

大型医疗设备的联网，患者可以在社区附近医院做 CT，由“医疗工业互联网”按照距离和水平来分配空闲的医生。

病人可以根据医生的长相、性别、年龄、从医经验、毕业院校等条件自动挂号，还可以与医生通过聊天软件进行沟通。

住院部没有病房也不要紧，患者可以通过云端预订其他医院的病房，即拿着A医院的片子、B医院的诊断书，到C医院照方抓药，在D医院安心疗养。

大数据从哪里来？

2、工业互联网

医疗保健：上海 “健康云”

可利用信息化平台实现远程诊断监控

基于云计算技术的健康档案数据中心

全面覆盖居民在区辖医疗机构的就诊信息和公共卫生服务信息



大数据从哪里来？

2、工业互联网

租用 远程生命体征仪



使用远程生命体征仪测心电图、血压、胎心、血氧含量等健康数据，所测数据可以通过手机、无线网络等多种方式传送至长宁区“健康云”平台。

专业的健康监测团队对这些数据及时分析和监测，一旦发现居民的健康可能出现问题，就会立即通知家庭医生，由他为居民实施进一步诊疗。

大数据从哪里来？

3、行业/企业信息系统

IBM 《分析：大数据在现实世界中的应用》：

“ 超过一半的受访者把内部数据视为 “大数据” 的主要来源 ”。

大数据从哪里来？

4、社交网络 SNS (Social Networking Services)

对比	排名	APP	领域	开发商	活跃人数(万) ▼
☆	1	 微信	社交网络	深圳市腾讯计算机系统有限公司	78,696.20 ↑
☆	2	 QQ	社交网络	深圳市腾讯计算机系统有限公司	57,944.80 ↓
☆	3	 支付宝	支付	阿里巴巴网络技术有限公司	33,712.70 ↓
☆	4	 淘宝	综合电商	阿里巴巴网络技术有限公司	29,740.20 ↑
☆	5	 爱奇艺视频	综合视频	北京爱奇艺科技有限公司	24,561.10 ↑
☆	6	 QQ浏览器	浏览器	深圳市腾讯计算机系统有限公司	24,128.90 ↓
☆	7	 腾讯视频	综合视频	深圳市腾讯计算机系统有限公司	23,919.60 ↑
☆	8	 手机百度	搜索	百度在线网络技术（北京）有限公司	23,638.70 ↑
☆	9	 酷狗音乐	移动音乐	广州酷狗计算机科技有限公司	23,079.10 ↓
☆	10	 WiFi万能钥匙	无线管理/WIFI管理	上海连尚网络科技有限公司	22,832.50 ↑
☆	11	 腾讯手机管家	安全管理	深圳市腾讯计算机系统有限公司	19,169.60 ↓
☆	12	 应用宝	应用商店	深圳市腾讯计算机系统有限公司	18,979.50 ↑
☆	13	 微博	社交网络	新浪网技术(中国)有限公司	17,053.20 ↑

大数据从哪里来？

4、社交网络 SNS (Social Networking Services)

社交网络和移动互联网的发展催生出大量的非结构化数据。

常见的有常见的图像、视频、音乐、办公文档、Web页面等。

大数据的幕后推手

在数字世界里的任何踪迹都变成了数据的一部分，如：每条评论，吐槽，点赞，购买记录等。

“大数据是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。”

——来自哈佛大学社会学教授加里·金



大数据的幕后推手

摩尔定律：铸造数据滋生的利器

表述：当价格不变时，集成电路上可容纳的晶体管数目，约每隔 18 个月便会增加一倍，性能也将提升一倍。

大数据的幕后推手

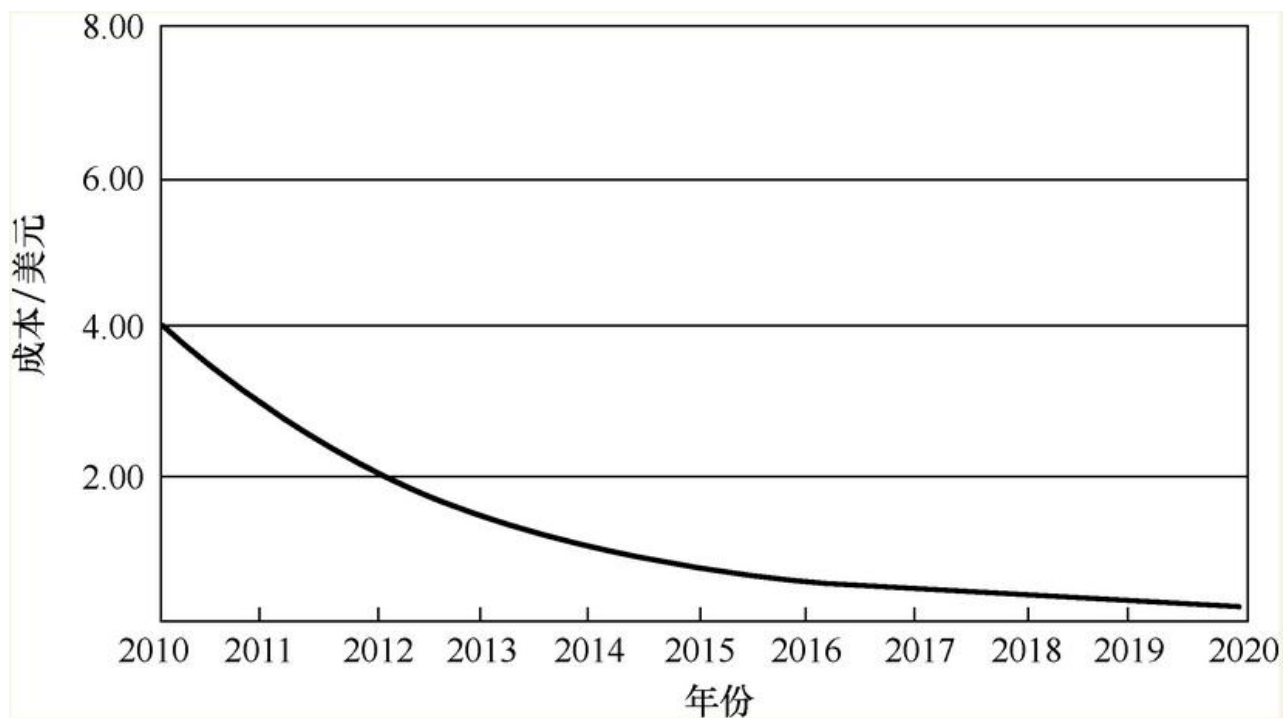
摩尔定律：铸造数据滋生的利器

数据的存储成本正在下降！

大数据的幕后推手

摩尔定律：铸造数据滋生的利器

数据的存储成本发展趋势



大数据的幕后推手

吉尔德定律：大带宽支撑大数据

大带宽是处理极端高速关键数据的基本要求，
是实现快速高效消化和处理大型数据集的基础。

大数据头号天敌：带宽



大数据的幕后推手

吉尔德定律：大带宽支撑大数据

大数据头号天敌：带宽



大数据的幕后推手

吉尔德定律：大带宽支撑大数据

占用高带宽应用：以视频图像为主的监控业务

北京：3万多辆公交车，每辆车上装 4 个摄像头，则数据总量预计达到约 180 GB，且对图像的连续性和实时性有较高要求，其传输带宽需求绝不是目前3G甚至4G可以轻松承载的。



大数据的幕后推手

吉尔德定律：大带宽支撑大数据

内容：

在未来25年，主干网的带宽每6个月增长一倍，其增长速度是摩尔定律预测的CPU增长速度的3倍，且将来上网会免费。

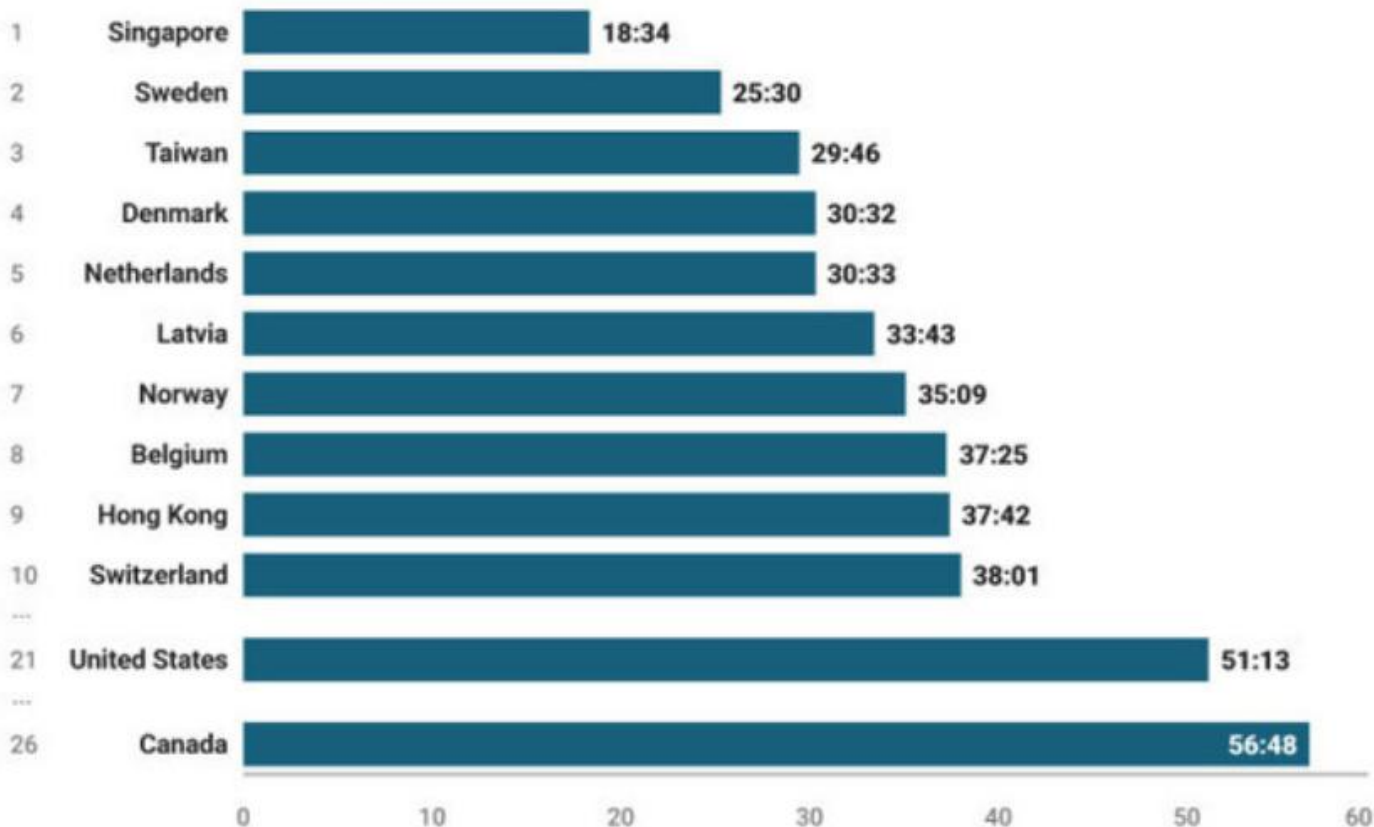


大数据的草尸堆干

TECH CHART OF THE DAY

BUFFERING: US STILL LAGS BEHIND IN DOWNLOAD SPEED

Average time to download an HD movie (7.5GB) in 2017, mm:ss



SOURCE: Cable.co.uk

statista

BUSINESS INSIDER

全球宽带网速大排名

吉尔德定律

新加坡：55.13

也门：0.34

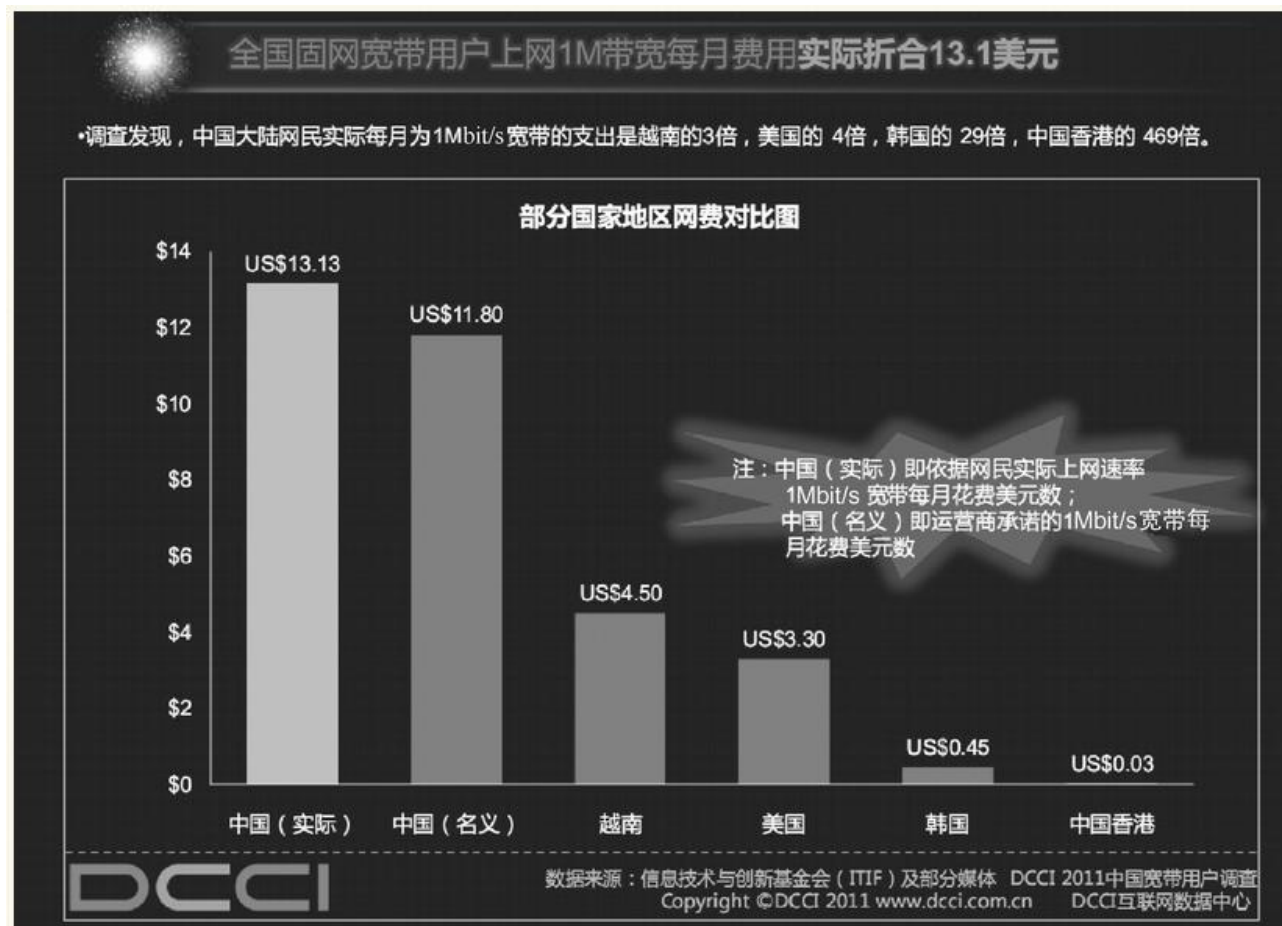
大陆：1.55

平均带宽

下载网速

大数据的幕后推手

吉尔德定律



部分国家地区网费对比

大数据的幕后推手

麦特卡夫定律：大数据价值是用户创造的

大数据时代，数据就是能源！

奥巴马：数据是“未来的石油”



大数据的幕后推手

麦特卡夫定律：大数据价值是用户创造的



网络价值以用户数量平方的速度增长，即网络价值等于网络节点数的平方，即 $V=n^2$ （ V 为网络总价值， n 为用户数或节点数）。

大数据的幕后推手

麦特卡夫定律：大数据价值是用户创造的

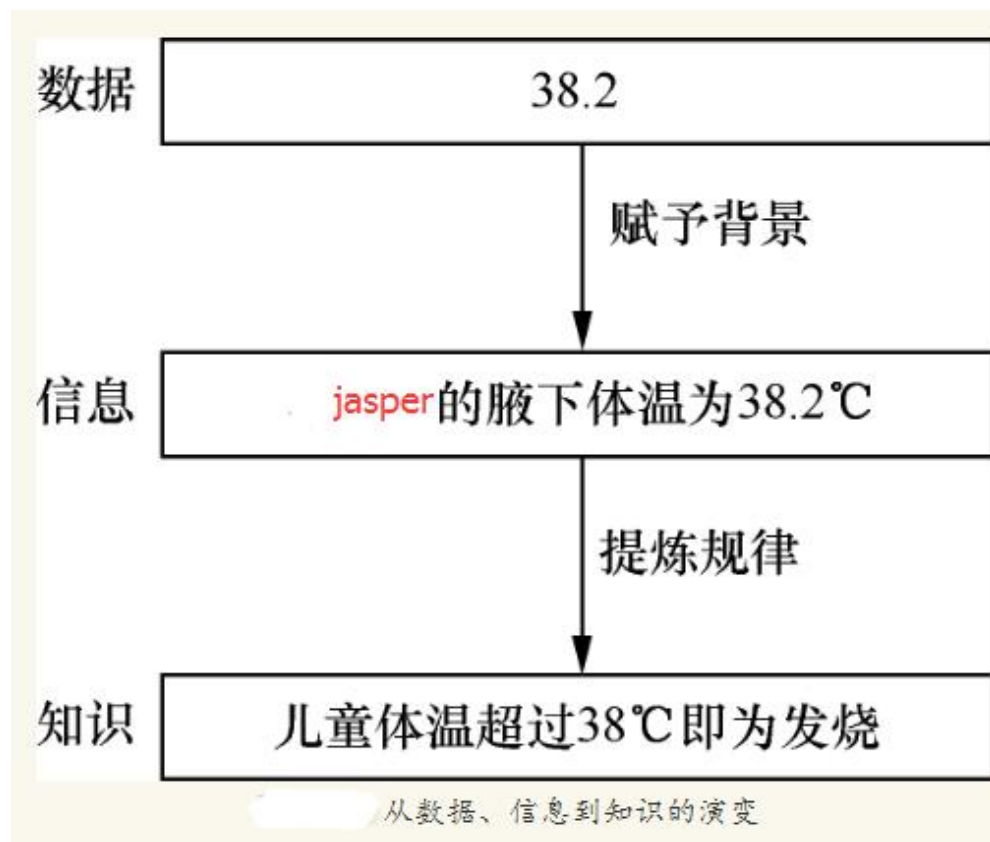


最终网络计算发展方向：

(Wireless Broadband , Online Always) 无线宽带，永远在线

什么是大数据

数据、信息、知识 “三级跳”



什么是大数据

大数据是什么

一人一个说法，一家一种解释，且公说公有理，婆说婆有理，如同“一千个读者心中有一千个哈姆雷特”。

什么是大数据

大数据是什么

世界著名咨询机构麦肯锡公司发布（201105）

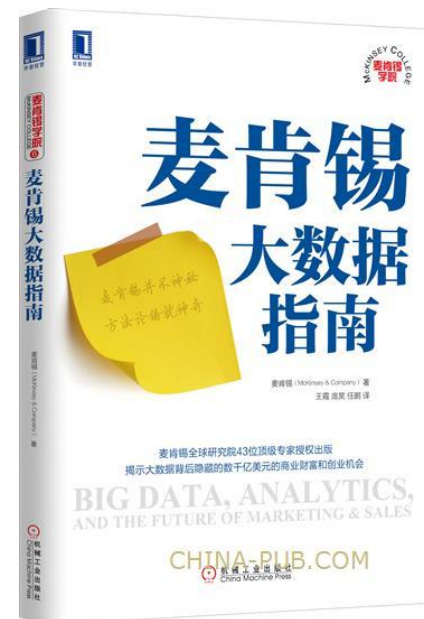
《大数据：下一个创新、竞争和生产力的前沿》研究报告

大数据是指其大小超出了典型数据库软件的采集、存储、管理和分析等能力的数据集。

一是符合大数据标准的数据集大小是变化的，会随着时间推移、技术进步而增长；

二是不同部门符合大数据标准的数据集大小会存在差别

大数据的一般范围是从几个TB到数个PB（数千TB）

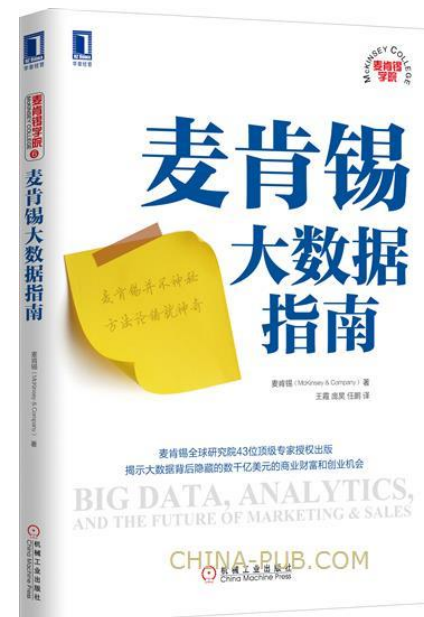


什么是大数据

大数据是什么

大数据的一般范围是从几个TB到数个PB（数千TB）

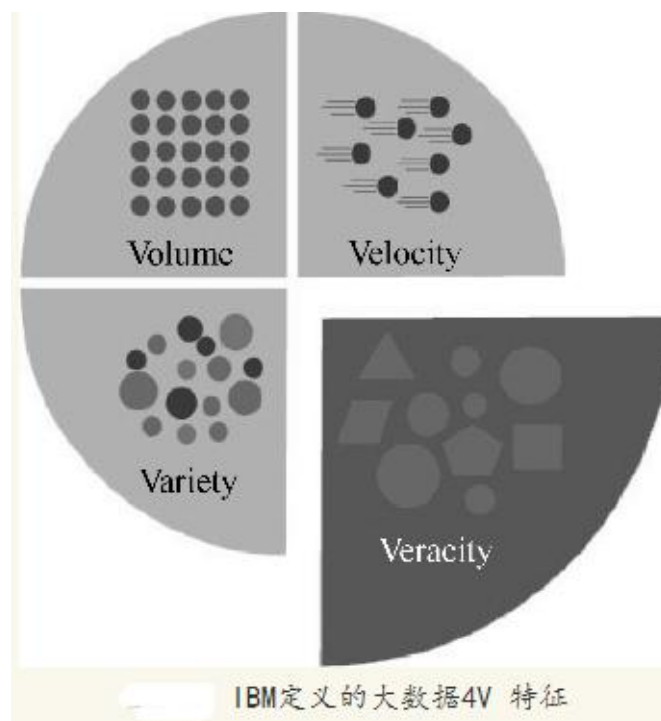
1KB (Kilobyte 千字节)=1024B ,
1MB (Megabyte 兆字节 简称“兆”)=1024KB ,
1GB (Gigabyte 吉字节 又称“千兆”)=1024MB ,
1TB (Trillionbyte 万亿字节 太字节)=1024GB ,
1PB(Petabyte 千万亿字节 拍字节)=1024TB ,
1EB(Exabyte 百亿亿字节 艾字节)=1024PB ,
1ZB (Zettabyte 十万亿亿字节 泽字节)= 1024 EB,
1YB (Yottabyte 一亿亿亿字节 尧字节)= 1024 ZB,
1BB (Brontobyte 一千亿亿亿字节)= 1024 YB.



什么是大数据

大数据是什么

IBM 《分析：大数据在现实世界中的应用》（201303）



什么是大数据

大数据是什么

EMC研究报告《大数据：创造商业价值的大机会》

报告指出：“大数据并不是一个准确的术语；相反，它是**对各种数据（其中大多数是非结构化的）永不休止积聚的一种表征**。它用以描述那些呈指数级增长，并且因太大、太原始或非结构化程度太高而**无法使用关系数据库方法进行分析的数据集**。不论是数 TB 还是数 PB，**数据的精确数量不如最终结果及数据如何使用重要。**”

什么是大数据

大数据是什么

EMC研究报告《大数据：创造商业价值的大机会》

强调大数据中的价值（Value），特别是商业价值。

大数据的特征

4V定律

Volume (容量)

Variety (多样性)

Velocity (速度)

Value (价值)



大数据的特征

4V定律

Volume (体量)

Variety (多样性)

Velocity (速度)

Value (价值)



前3条由道格提出。

第4条IDC发布《**从混沌中提取价值**》:大数据代表了新一代的技术和架构,通过使用高速(Velocity)采集、计算、处理和分析,可用于从超大容量(Volume)的多样化(Variety)数据中经济地提取价值(Value),即在大数
据传统3V特征的基础上增加了一个新特征: Value (价
值), 形成4V特征。

大数据的特征

4V定律

1、Volume（容量）

麦肯锡在2011年5月发布研究报告《大数据：下一个创新、竞争和生产力的前沿》指出：不同机构的研究成果都表明，**未来数年全球数据总量将会呈现指数级增长。**

大数据的特征

4V定律

1、Volume（容量）

2013年整个人类社会总共拍摄超过**3.5万亿**张照片，

如今人们每两分钟拍摄的照片数就比整个拍摄照片总数还要多。



大数据的特征

4V定律

2、Velocity (速度)

两个方面：

一是数据在不断更新，增长速度非常快。

二是对数据存储、传输、处理等速度要求非常快。



天下武功，唯快不破



大数据的特征

4V定律

2、Velocity (速度)

快的原因

一是时间就是金钱

时间越小，单位价值就越大



天下武功，唯快不破

大数据的特征

4V定律

2、Velocity (速度)

快的原因

二是数据的价值会折旧

过去一天的数据，可能比过去一个月的数据都更有价值



天下武功，唯快不破

大数据的特征

4V定律

2、Velocity (速度)

快的原因

三是数据具有时效性



天下武功，唯快不破



NOAA 气象数据采集船

数据的时效性**要求较高**，数据分析的结果会直接影响下一步的决策

大数据的特征

4V定律

3、Variety（多样性）

多样性体现：

一是数据来源多

大数据的特征

4V定律

3、Variety（多样性）

一是数据来源多

大数据通常可分为4类：

泛互联网

车联网

企业信息系统

社交网络

大数据的特征

4V定律

3、Variety（多样性）

二是数据类型多，且以非结构化数据为主



大数据的特征

4V定律

3、Variety（多样性）

二是数据类型多，且以非结构化数据为主

结构化数据是指存储在数据库当中、有统一结构和格式的数据，这种数据比较容易进行分析和处理，它包括数据库、数据表和指定格式文件。

大数据的特征

4V定律

3、Variety（多样性）

二是数据类型多，且以非结构化数据为主

非结构化数据是指无法用数字或统一的结构来表示，字段长度可变，且每个字段的记录又可以由可重复或不可重复的子字段构成的数据。

大数据的特征

4V定律

3、Variety（多样性）

二是数据类型多，且以非结构化数据为主

半结构化数据包含结构化的数据，但是结构变化很大，包括邮件、HTML（Hyper Text Mark-up Language，超文本标记语言）、报表、资源库、网络日志等。

由邮件系统、Web集群、教学资源库、数据挖掘系统、档案系统等应用产生的。

大数据的特征

4V定律

3、Variety（多样性）

三是数据之间关联性强，频繁交互

在海量、种类繁多的数据间发现其内在关联

大数据的特征

4V定律

Value (价值)

数据将是未来世界经济的“原油”

大数据的特征

4V定律

Value (价值)

未来的十年将是一个以“数据”为引领的智慧科技时代



大数据的特征

4V定律

4、Value（价值）

从某种意义上说，数据将成为企业的核心资产



大数据的特征

4V定律

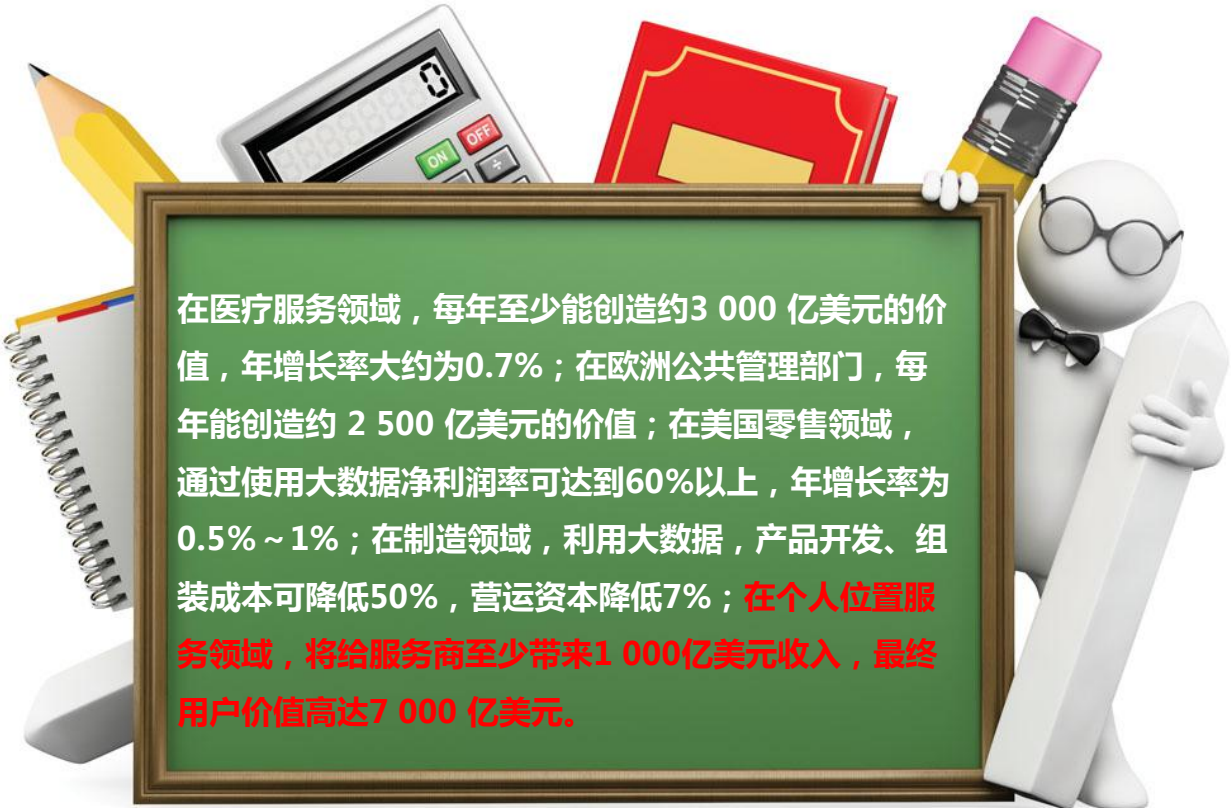
Value (价值)

三层含义：

一是大数据有大价值

麦肯锡发布：

《大数据：下一个创新、
竞争和生产力的前沿》



在医疗服务领域，每年至少能创造约3 000 亿美元的价值，年增长率大约为0.7%；在欧洲公共管理部门，每年能创造约 2 500 亿美元的价值；在美国零售领域，通过使用大数据净利润率可达到60%以上，年增长率为0.5%~1%；在制造领域，利用大数据，产品开发、组装成本可降低50%，营运资本降低7%；**在个人位置服务领域，将给服务商至少带来1 000亿美元收入，最终用户价值高达7 000 亿美元。**

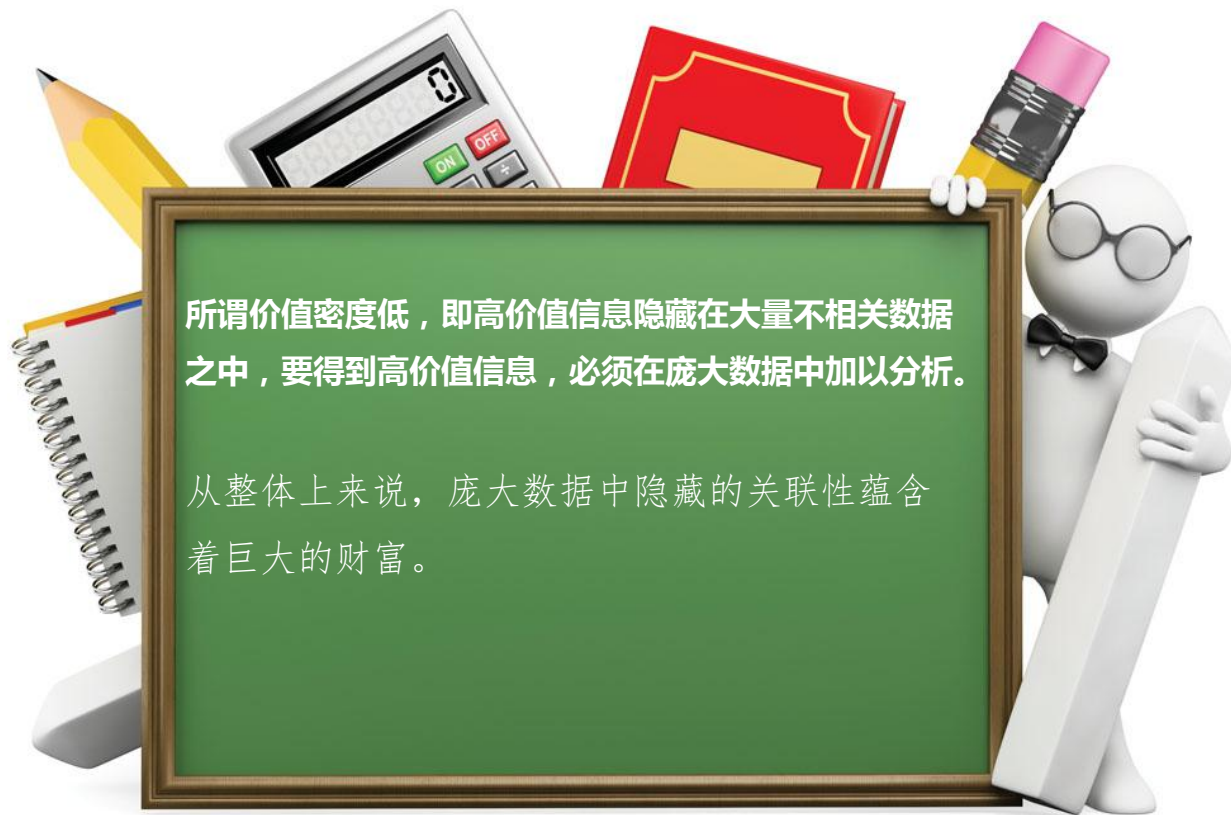
大数据的特征

4V定律

Value (价值)

三层含义：

二是价值密度低



所谓价值密度低，即高价值信息隐藏在大量不相关数据之中，要得到高价值信息，必须在庞大数据中加以分析。

从整体上来说，庞大数据中隐藏的关联性蕴含着巨大的财富。