

大数据的发展与应用

何宝宏



（工业和信息化部电信研究院）

2016年05月26日 贵阳

厚德實學 興業致遠

主要内容

- 互联网技术趋势
- 大数据应用概述
- 大数据的思考

移动

从蒸汽机到互联网



交通出行



医疗健康



金融理财

交通业：又一个移动互联网



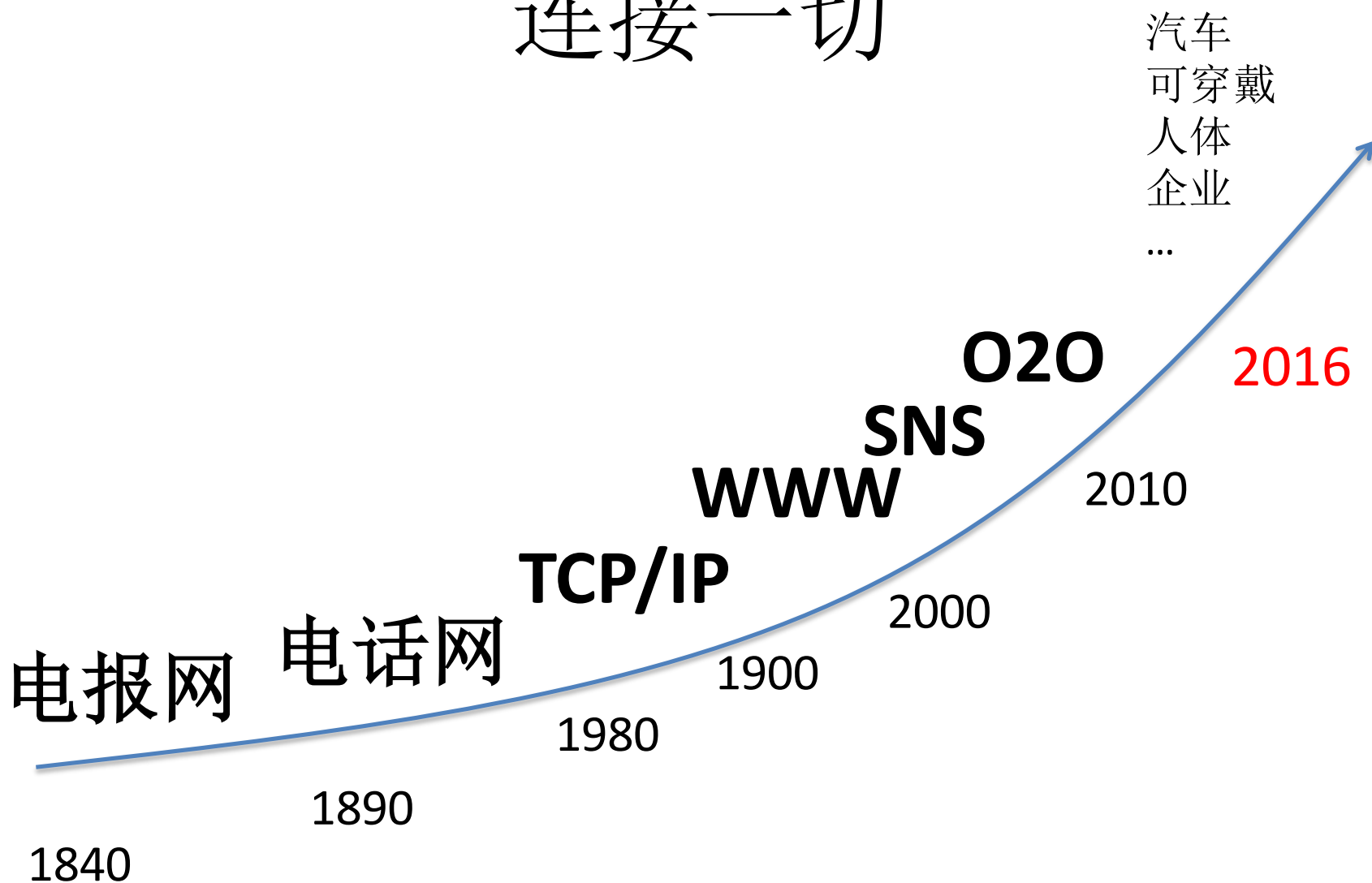
车联网

“硬件时代”
的汽车业

重新定义
道路？

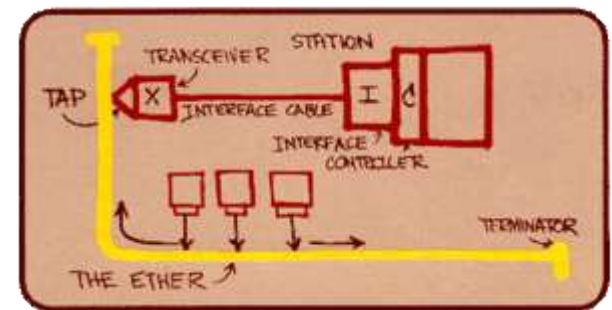
万物互联

连接一切

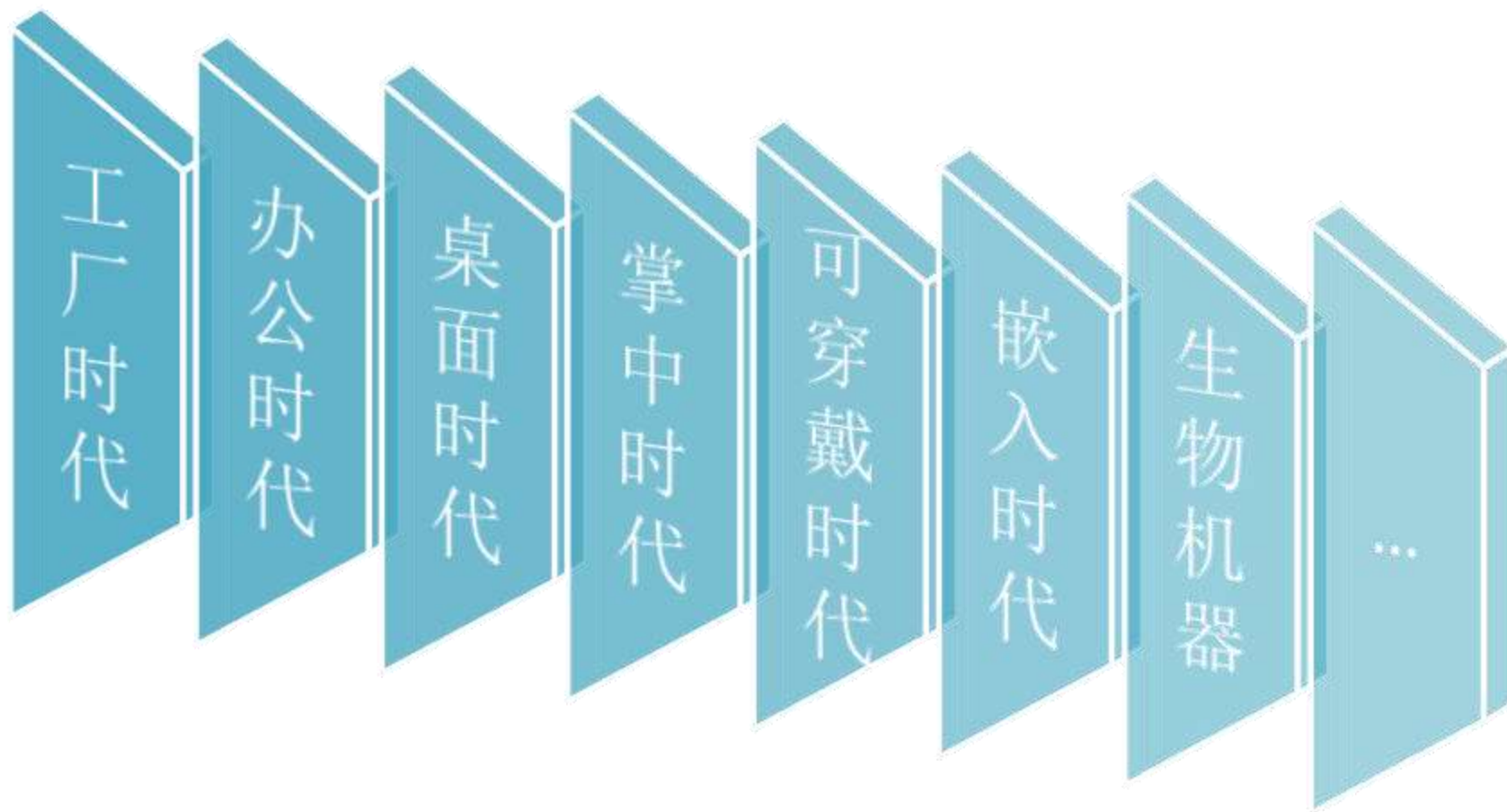


无连接成就全连接

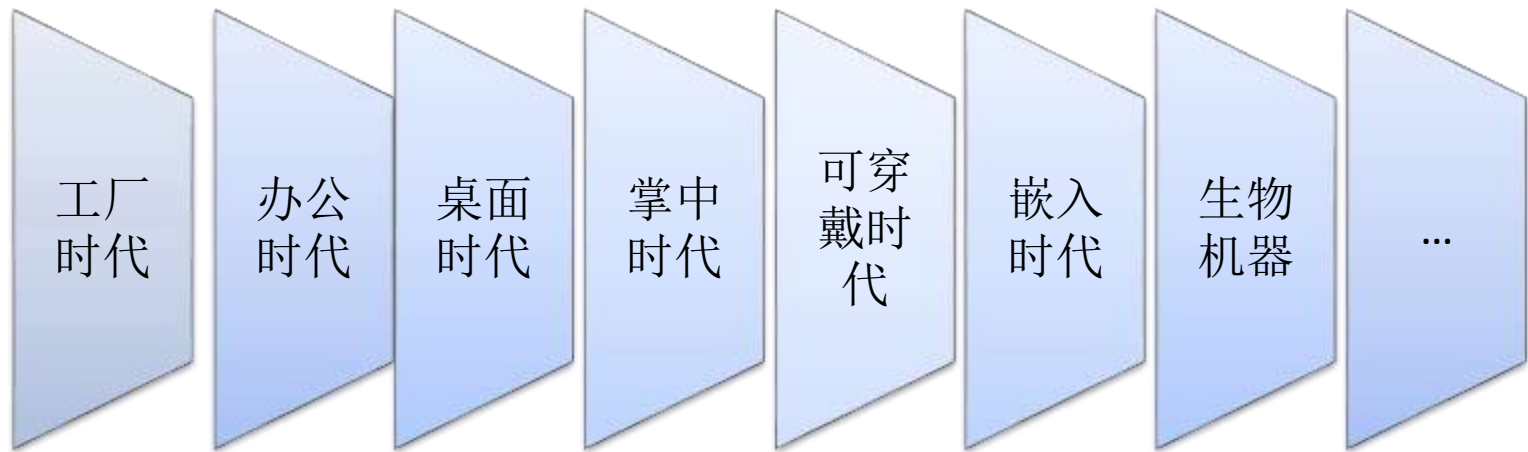
- Who, Where, How
- Keep it Simple and Stupid
- No QoS/安全保证
- Running Code
- P2P/开源/IDC/云计算
- 共享经济
- ...



终端的变迁

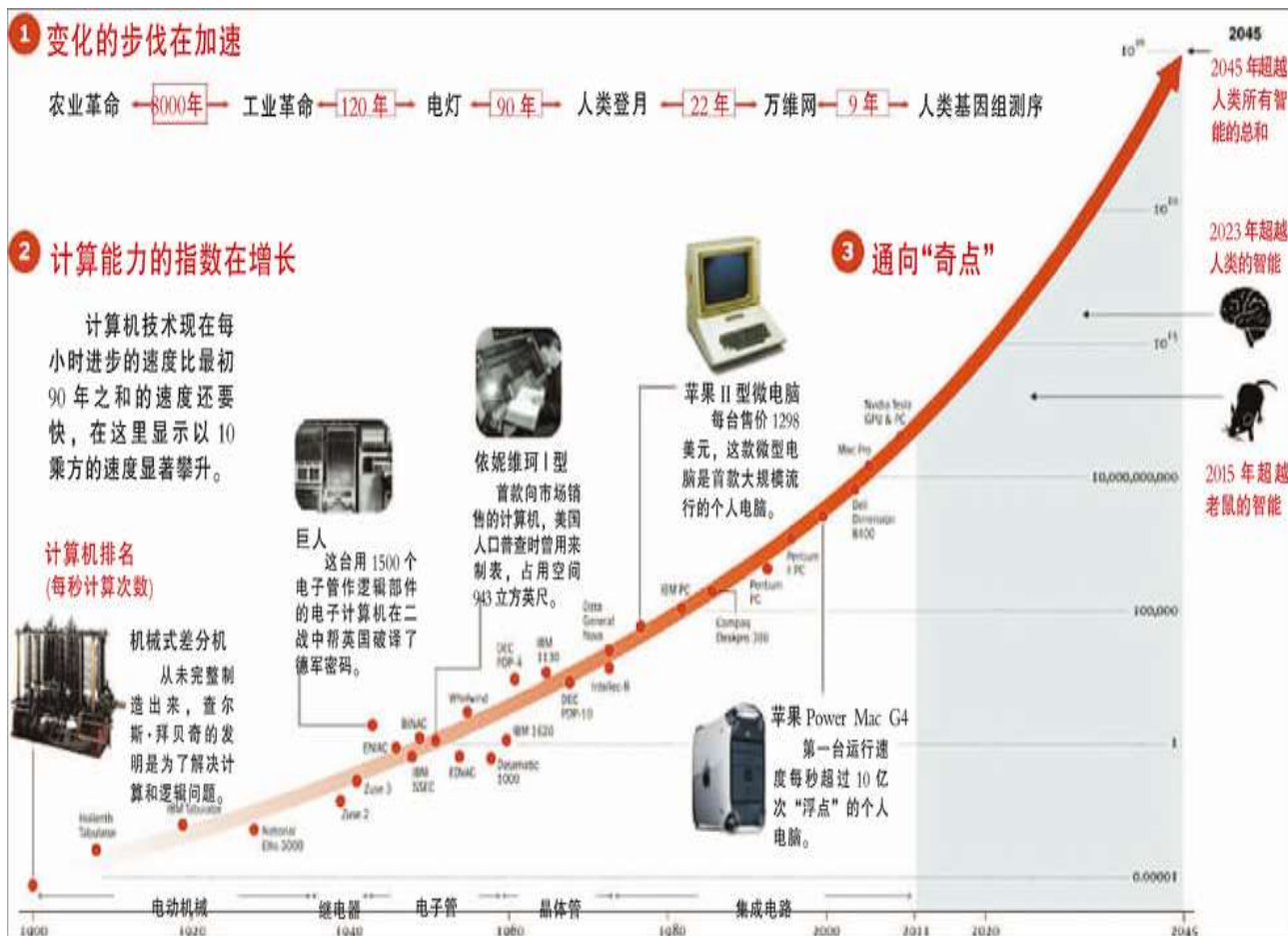


数据中心就是计算机



万物智能

摩尔定律的世界

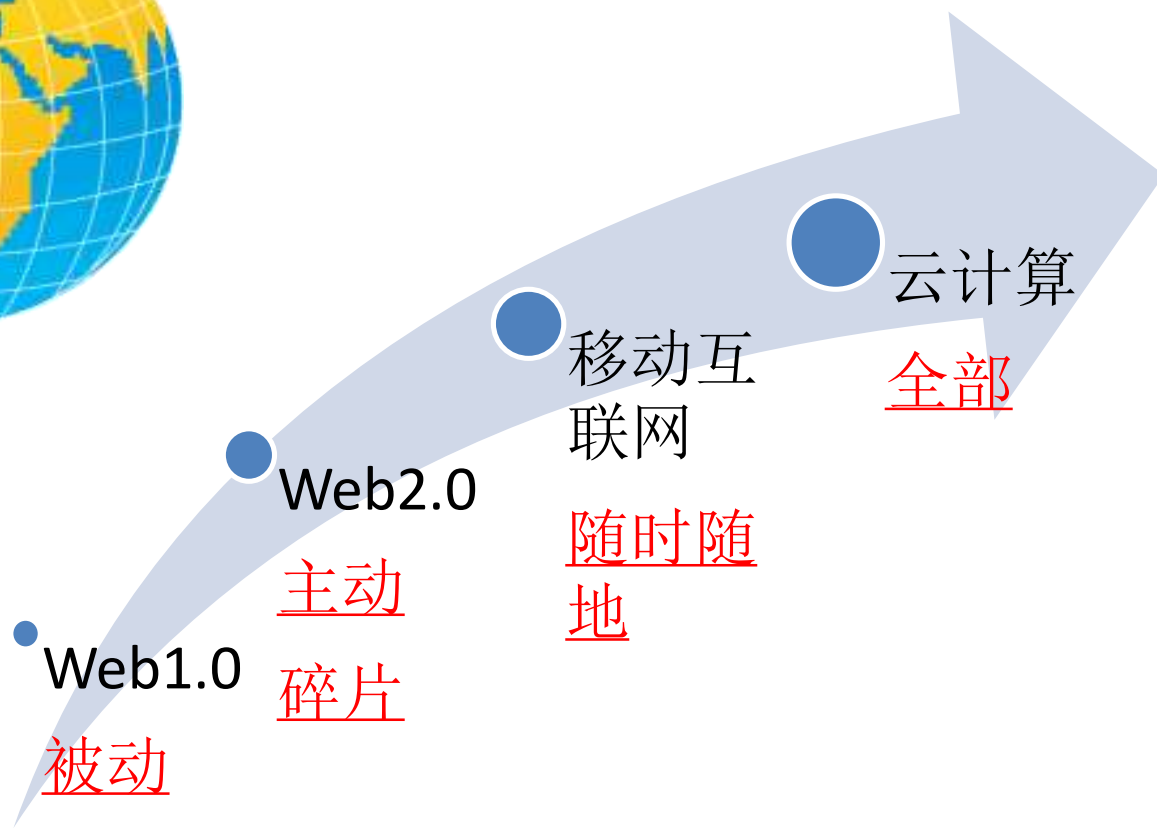


从编程机器到训练机器

- 编程成为元技能，一种掌握世界秘密的语言
- 就算程序员不掌控世界，它们也掌控那些掌控世界的东西
- 生物细胞都不过是些**DNA**软件驱动的生物机器
- 一个由深度学习机器主宰的世界，需要另一种劳动者
- 未来将是一小撮人改变世界
- “看不见的力量”将终结我们的信仰

数据时代

从IT到DT



软件定义世界

软件统治世界

全球最成功的软件企业

- 书店：Amazon
 - 视频商：Netflix
 - 音乐公司：iTunes
 - 图片公司：Flickr
 - 游戏公司：Zynga
 - 猎头公司：LinkedIn
 - 话音通信：Skype
 - ...
- 软件定义电子产品
 - 软件定义终端
 - 软件定义服务器
 - 软件定义网络
 - 软件定义数据中心
 - 软件定义的安全
 - 软件定义的存储
 - 软件定义的无线电(SDR)
 - 软件定义的收发器
 - ...

软件定义“互联网+”？

开源即标准

程序猿：世界上最大的手工艺群体

软件业：从卖许可证到卖服务

全球开源相关产值已经超过了IT服务**1/3**

一切皆服务

电信运营商

App Store

云计算：

从话音服务到IP服务

手机即服务

从软件产品到IT服务

云服务增速是IT支出的4倍，在全球IT支出中的占比将从2013年的3.6%提高到2018年的6.6%

案例*：微软：Office软件 -41%，云服务 +106%

7年之痒



产业互联网
2015—2022?

移动时代
2008—2015

社会的
2001—2008

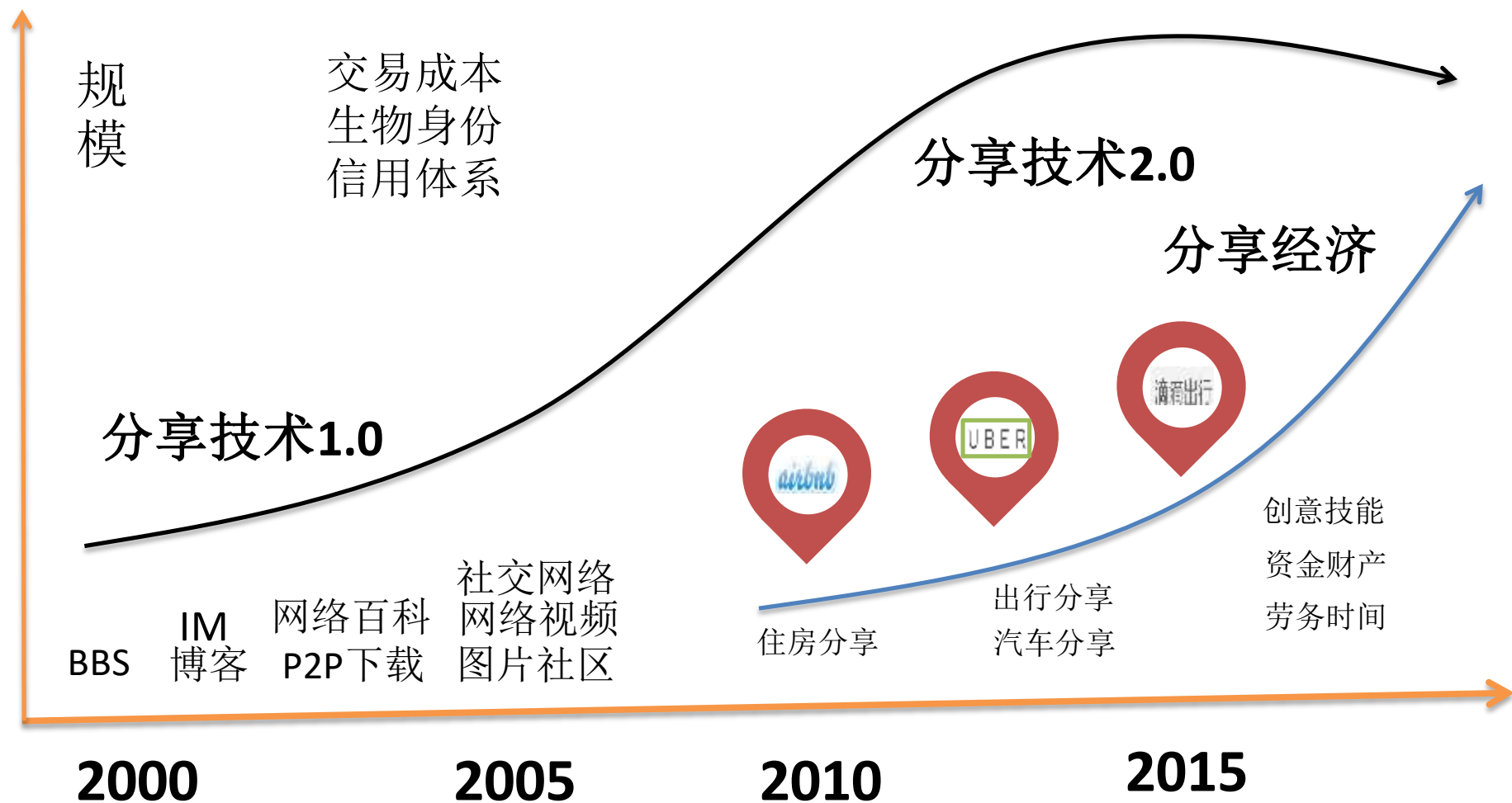
资本盛宴
1994—2001

一统江湖
1984—1991

网络春秋
1969—1983

包交换理论
1962—1969

分享经济：P2P+O2O？



忽悠还是未来？

区块链

让互联网可以直接传递价值！

主要内容

- 互联网技术趋势
- 大数据应用概述
- 大数据的思考

史上“大数据”革命



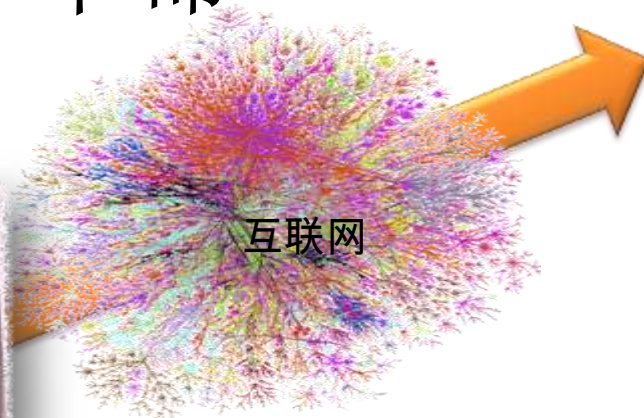
口耳相传
语言



保存与传承
文字



大范围传播与
持久保存
印刷术



互联网

硅介质
数字化
网络化

距今10万年前

距今5000年前

公元1000年

公元2000年

古登堡印刷术发明后，50年书籍增加了800万册，增长1倍，是之前西方数据量的总和

史上大数据：政府大数据

约前2200年大禹“平水土分九州数
万民”

公元2年，汉书·地理志：全国103个
郡国，人口是59,594,978人

1909年，清政府颁布调查户口章程



古罗马，每5年一次

税收！



1086年，英国《末日审判书》
人口、土地、财产登记

美国宪法规定10年一次人口普查
1880年统计用了8年，预计1890年
要13年
穿孔卡片制表机的发明，要1年

史上大数据：天文大数据



第谷
人肉的“大数据”资源



开普勒
突破性的算法

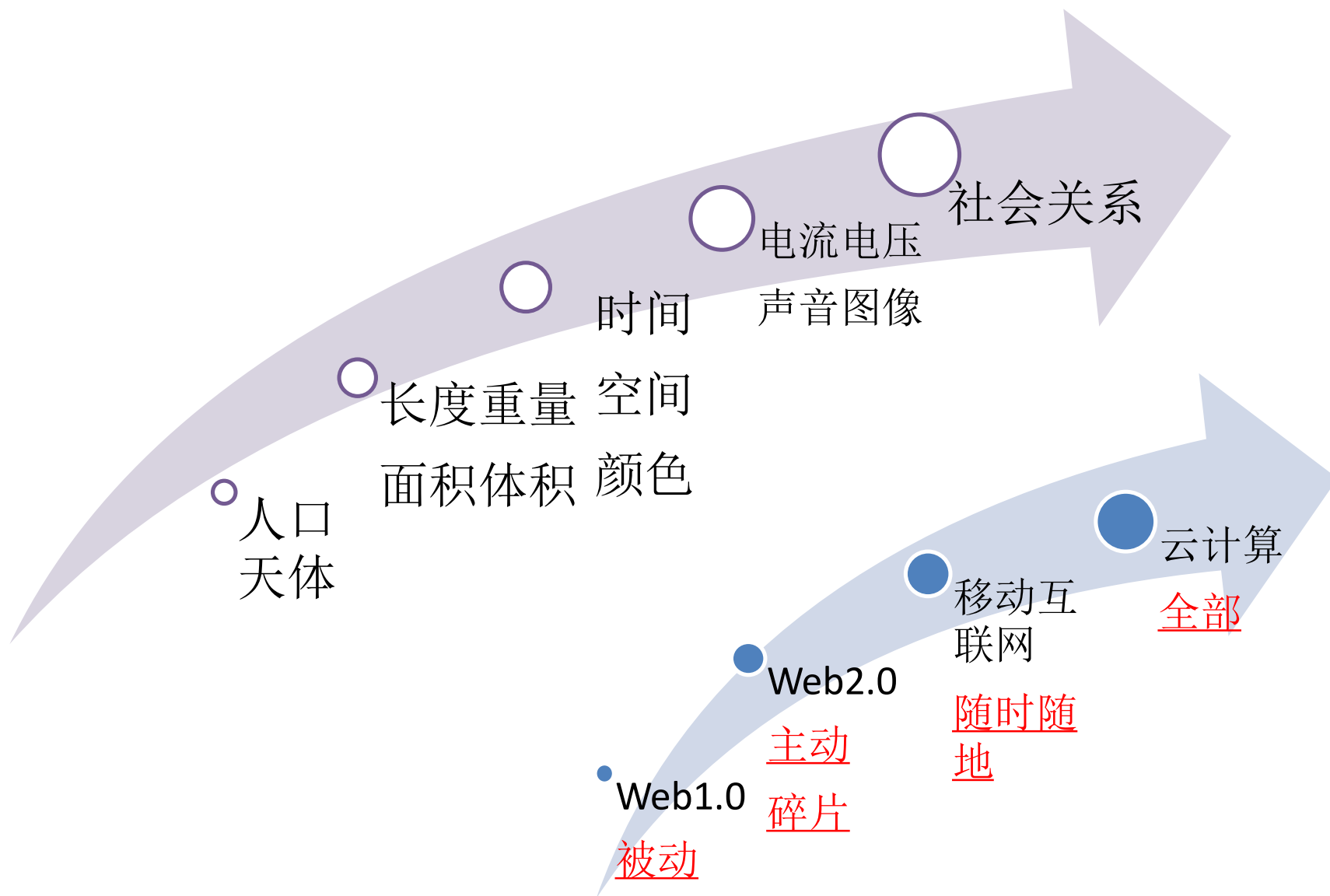


伽利略
工具的“大数据”资源



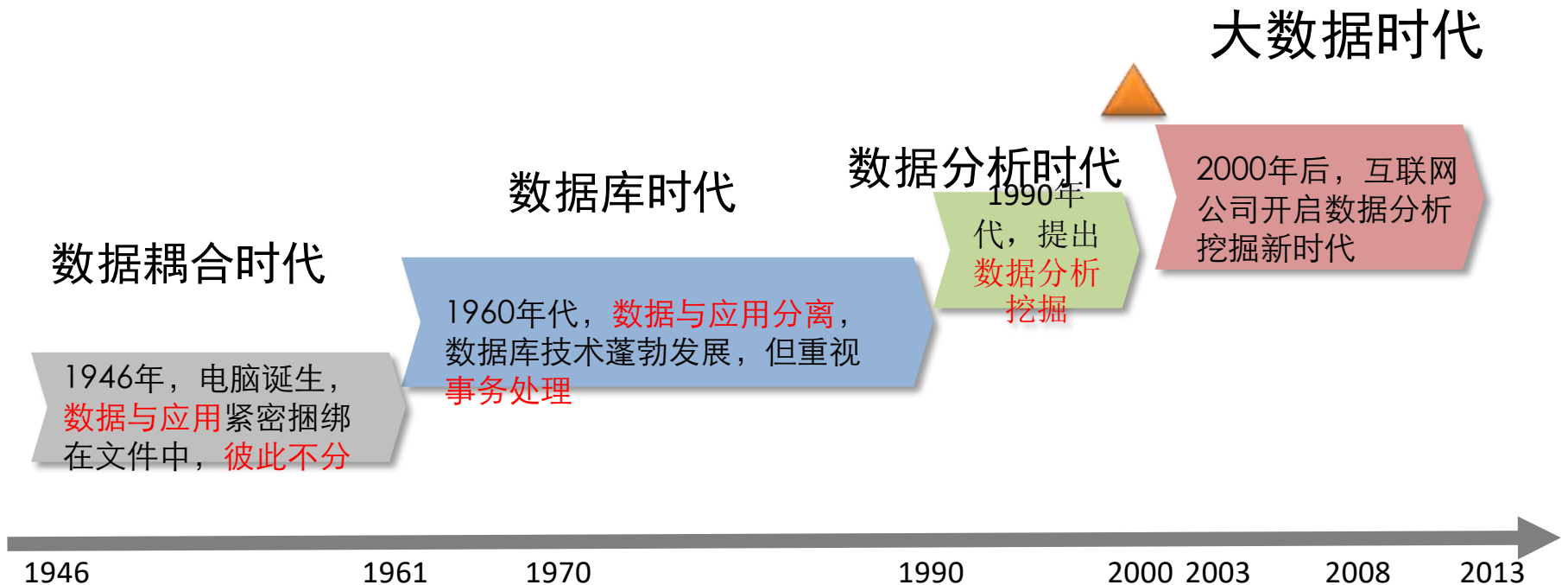
牛顿
划时代先进的理论

不断量化的世界



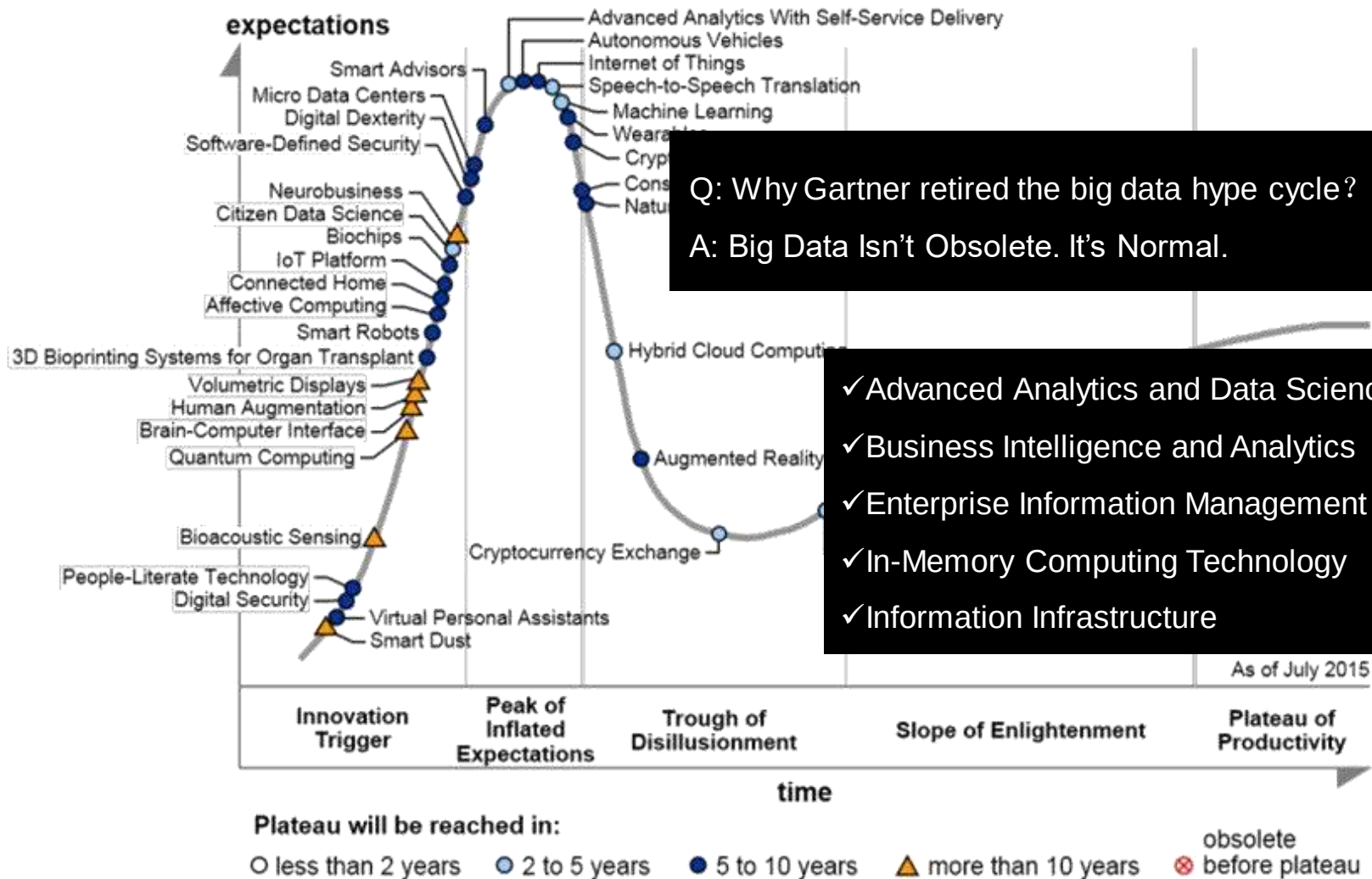
引导此轮大数据的原因

- 1) 数据更加丰富，ROI增长了3.5%
- 2) 分析工具更加强大，成本够低，MapReduce
- 3) 互联网商业上的成功，引起重视，麦肯锡报告



全球

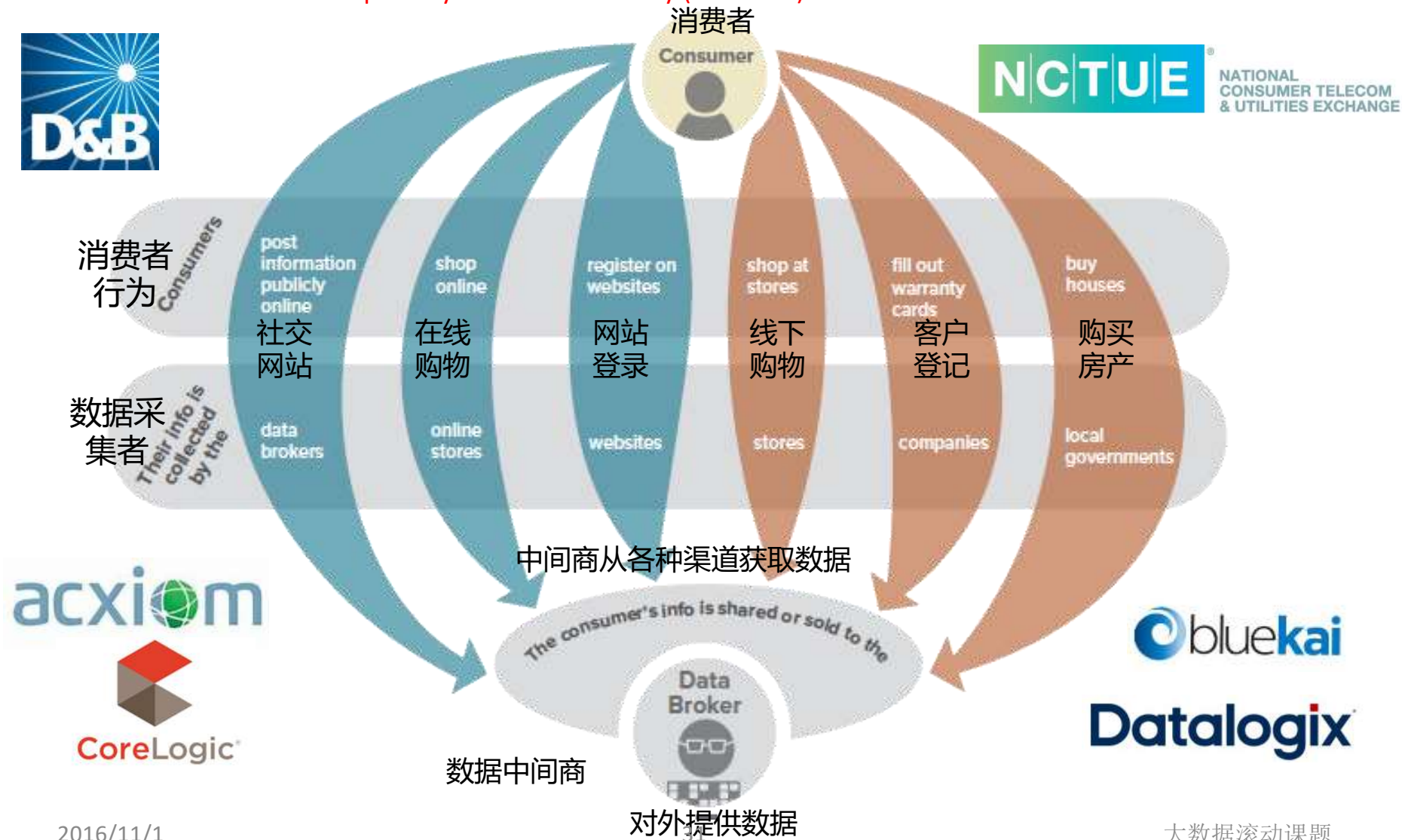
Gartner: 大数据已经成为“新常态”



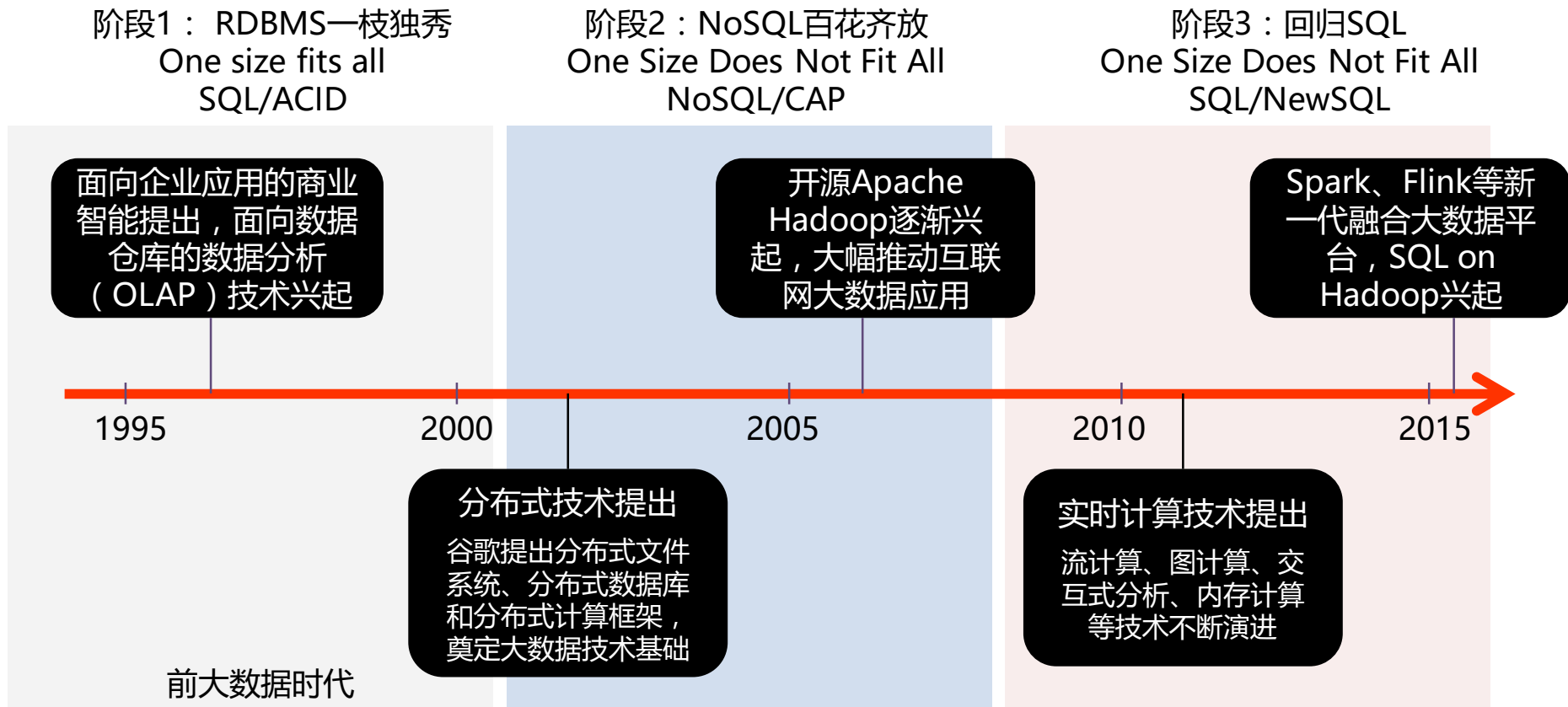
Source: Gartner (July 2015)

数据资源：开放政府数据外，商业交换需求迸发

Data broker: A Call for Transparency and Accountability (FTC 2014)



技术创新：技术快速演进，底层技术趋成熟



底层分布式技术日趋成熟，释放上层数据分析技术的创新活力：
实时流分析、交互分析、机器学习

应用创新：按技术和数据两条路径不断深化



产业生态：大数据生态系统不断分化繁荣

2016年

2012年

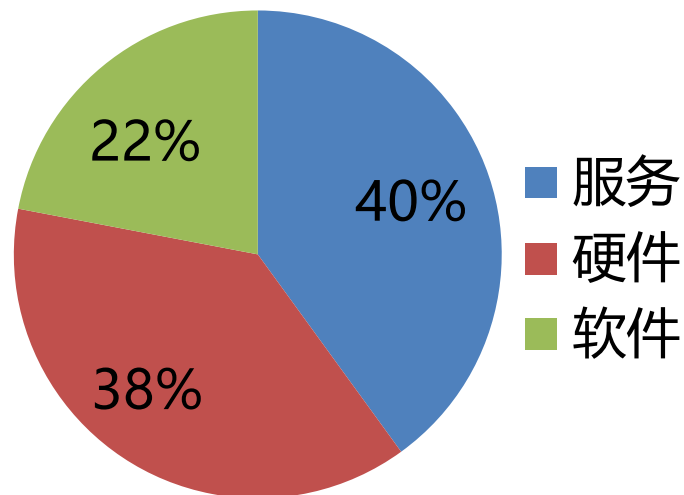
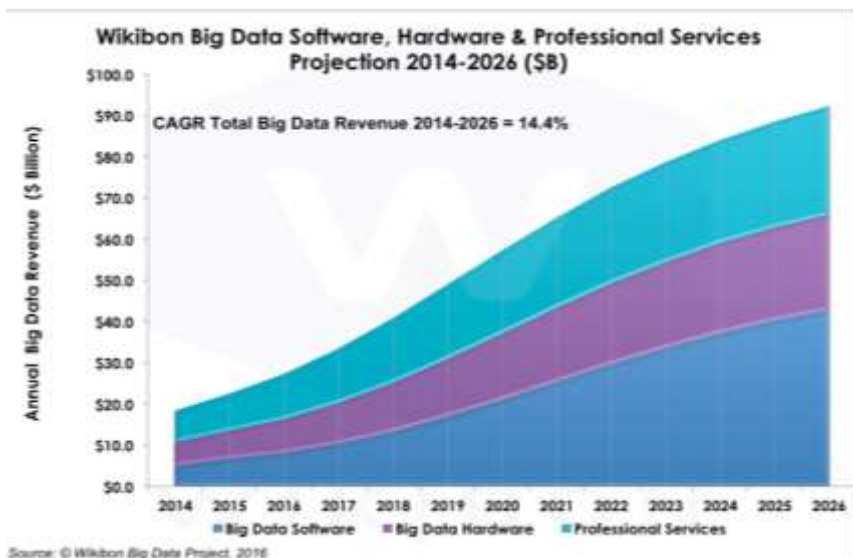


資料來源：Forbes，2012/6



大数据产值：直接产值持续增长，间接效益备受瞩目

狭义大数据产值口径：专门用于大数据的软件、硬件及服务，300亿美元



数据来源：Wikibon估计，2016年

大数据应用于各个行业，产生的外部价值受到各国重视

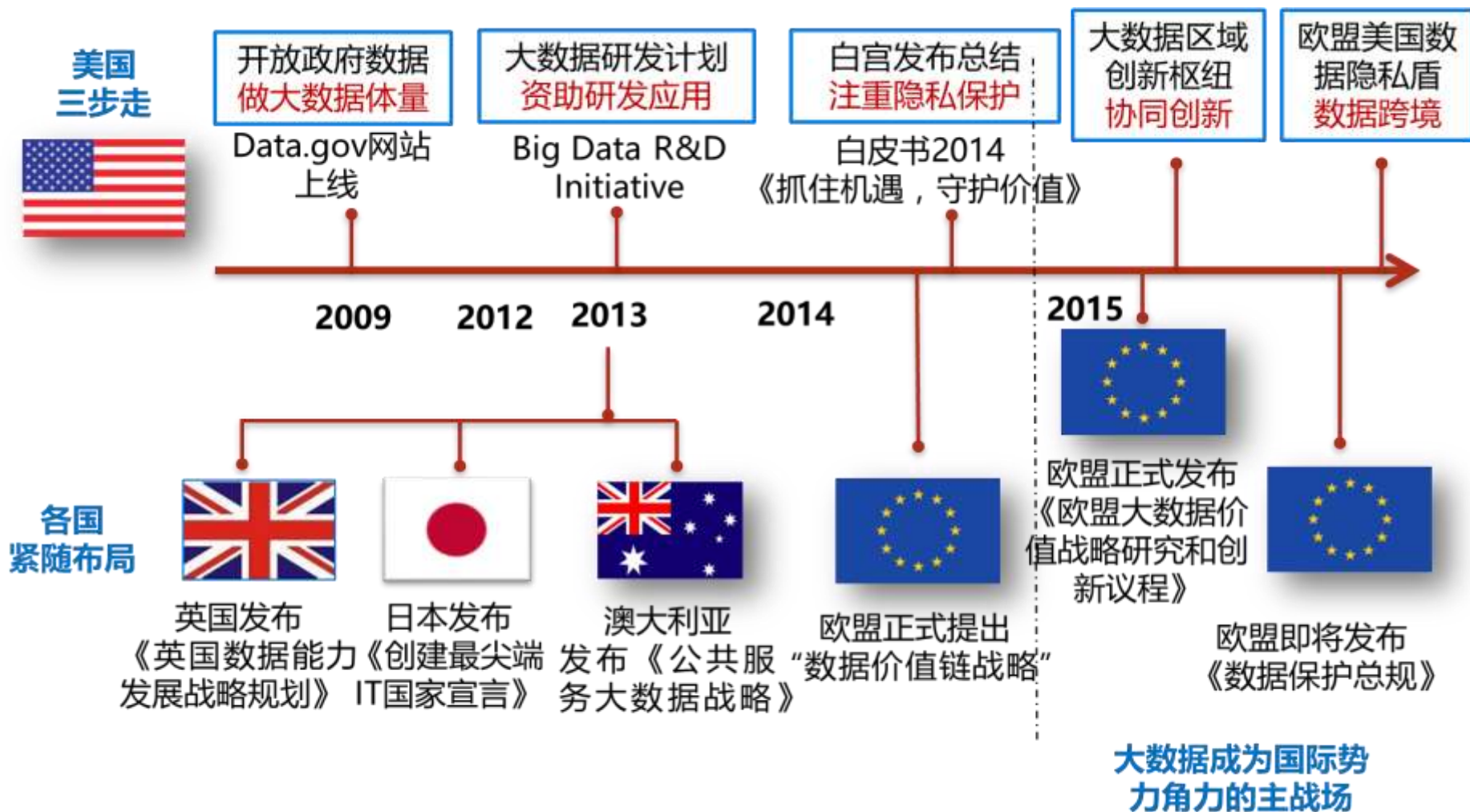


WARSAW INSTITUTE
FOR ECONOMIC STUDIES

The EU economy
will grow by an additional
1.9% by 2020
thanks to big
and open data

区域	测算机构	带动GDP增
欧盟	华沙经济研究所	1.9% (2020)
美国	麦肯锡	2%-4% (2020)
英国	CEBR	2.3% (2017)
中国	CAICT	0.6%-0.85% (2013)

各国政策：战略向纵深发展，关注落地具体问题



我国

国内大数据核心产业基本形成



数据资源——国内数据开放全球滞后

我国公共数据资源开放刚起步：2015年开放知识基金会数据开放评价中，我国第93（共122国）位

2015年 权威机构英国开放知识基金会 13个关键性数据开放指数评价



国际排名落后原因

- 评估国家增多，其中不乏后发实力的国家赶超
- 评估指标增加，由10个增加至13个，其中在国家统计数据、天气、地图、水质等关键领域，我国数据开放的整体水平低

数据资源——数据开放问题与挑战

我国公共数据资源开放处于起步阶段，与发达国家相比，**面临制度、规范、平台、数据可用性等多个方面的问题与挑战。**

指标	国际实践	国内问题与挑战
战略制定	发布了国家战略文件 修改了相关法律法规	目前《政府信息公开条例》在信息范围、深度、可用性等方面要求尚浅，缺乏强制性规定，缺少信息安全、隐私保护等方面的立法保障 目前，相关部委正在研究制定国家指导意见
开放范围	美英等正在加强数据清单的建设 目前通过行政命令逐渐扩大开放范围	受制于开放意识不足，许多部门缺乏开放数据的动力，需要具体的、可操作的清单指导落实工作。 目前，国家正在计划制定政府数据共享开放目录
标准规范	形成一系列的标准规范	尚未建立标准规范体系
使用管理	英国等保留公共信息资源的“皇家版权” 通过网站“开放政府授权”公众使用 不当使用数据的情况下自动终止授权	尚未进行使用管理
平台建设	美国、法国、德国、日本等国家 普遍通过全国统一平台集中开放数据	各地方分别建设数据开放平台，建设国家政府数据统一开放平台已列入《十三五规划纲要》
数据可用性	美国Data.gov采用XML、CSV、XLS等多种机器可读格式，同时提高数据质量和时效性，方便公众开发利用	数据的可机读性较差（以武汉开放网站为例，可机读数据仅占半数）；开放数据以静态数据居多；交互性整体较差

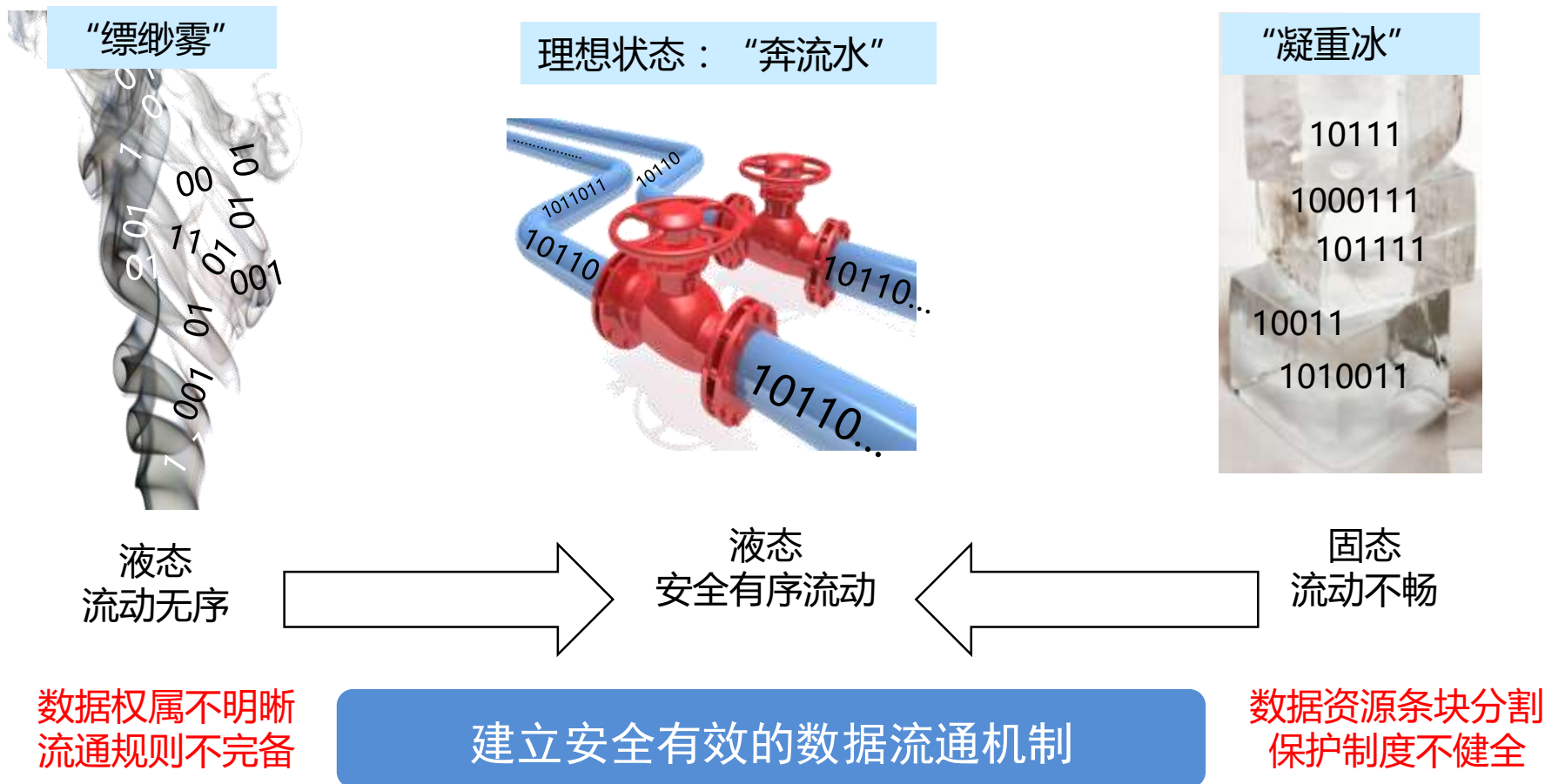
数据资源——数据交易平台建设如火如荼

当前我国数据交易市场较为活跃，交易活动主要依托第三方数据交易平台。全国范围内已建成、及正在筹建的地方数据交易机构十余家。

序号	单位名称	进展情况	备注
1	北京数据交易所	北京市金融工作局牵头筹建，近期将向北京市政府汇报筹建方案。	划拨现有闲置交易所改制
2	上海数据交易所	1、经营范围：商业公司互联网数据产品（一期）。 2、业务模式：使用权交易；多源数据聚合；实时在线交易。	2015年开始筹建，2016年正式成立
3	广州数据交易所	广州市大数据管理局牵头筹建。	一期成立广州数据交易中心，省政府报批流程中。
4	深圳金融数据资产交易所	经营范围：大数据交易、互联网金融。	拥有现货和权益类牌照。
5	贵阳大数据交易所	1、业务指导单位：贵阳市大数据办、贵阳市经信委； 2、经营范围：政府数据、金融数据、征信数据； 3、业务模式：产权交易；线下交易为主。	1、存在问题：政府数据占到总量的60%以上；直接参与数据加工与定价；数据产权争议；个人隐私保护争议。 2、组建形式：贵州阳光产权交易所参股。
6	长江大数据交易所	1、业务指导单位：武汉市经信委； 2、经营范围：数据产品、模型算法。 3、业务模式：倡导使用权交易。	组建形式：武汉光谷知识产权交易所参股。
7	陕西大数据交易所	1、业务指导单位：陕西省工信厅、西安西咸新区； 2、经营范围：政府数据、工业数据、个人数据等。	
8	华中大数据交易所	1、经湖北省政府批准成立 2、经营范围：互联网数据、政府数据、交通旅游数据等	2015年底，由北京东华软件股份公司等3家北京IT企业注资1亿元成立，湖北省国有和民营资本相继入股
-	其它正在筹建的数据交易机构	8、徐州大数据交易所（已成立）；9、江苏省大数据交易中心（即将成立）；10、重庆大数据交易所（重庆市经信委）；11、沈阳大数据交易中心（沈阳市大数据管理局）；12、哈尔滨大数据交易中心；13、青岛航运数据交易所；14、河南数据交易机构；15、其它省市和地区（如济宁、东莞、新疆中亚商品交易中心等）	

数据资源——开放流通的权责利不平衡

数据流通四大问题：数据分布条块分割、保护制度不健全、数据权属不清晰、流通规则不完备



技术创新：处于第三梯队格局难以改变

国际格局

阵营一
理论原创地



输出：原创论文

分布式系统+人工智能为代表的系列论文

领域	论文数
人工智能和机器学习	347
算法与理论	319
人机交互与可视化	237
自然语言处理	216
数据管理与挖掘	212
机器识别	202
信息检索和Web	158
软件系统	119
分布式系统与并行计算	110
安全与隐私	157

影响

阵营二
技术制高点



输出：开源+技术服务

Hadoop、Hive、Storm、Spark、Flink等大数据开源技术

Apache Hadoop
Cassandra (NoSQL)
Apache Thrift(跨语言框架)
Apache Hive(Hadoop分析)
RocksDB(Flash DB)
Presto (SQL on Hadoop)
Open Compute(数据中心)
Apache Storm(实时计算)
Open Academy(培训项目)
Apache Spark
Apache Flink...

影响

阵营三
产业先锋队



输出：商用产品

Hadoop/Spark两大商用技术平台和一体机

传统数据库数据仓库软件

Hadoop/Spark发行版的混搭解决方案

我国

?

?

我国大部分企业处于该阵营的跟随者，但比较分散，尚未形成实力

应用创新：行业应用尚未大规模启动

互联网企业热



——大互联网企业纷纷把大数据作为战略：中国互联网前十名全部有“大数据战略”：百度大脑、阿里“从IT到DT”、腾讯“大数据连接的未来”、京东大脑、小米“大数据不赚钱就破产”……

——大批创业公司也以互联网大数据为创投方向：秒针系统、百分点、闪银奇异、宜信、上海拍拍贷、北京云知声……

政府大数据产业



布局拉开序幕

政府热

- 23个省出台了74个与大数据相关的政策和规划。上海、贵州、重庆、天津等省市出台大数据发展规划，
- 全国已建大数据产业园区超过10个。
- 呈现中央高度重视、地方积极部署的良好局面。

行业：金融、电信基础好，起点高。其他行业渗透程度不一

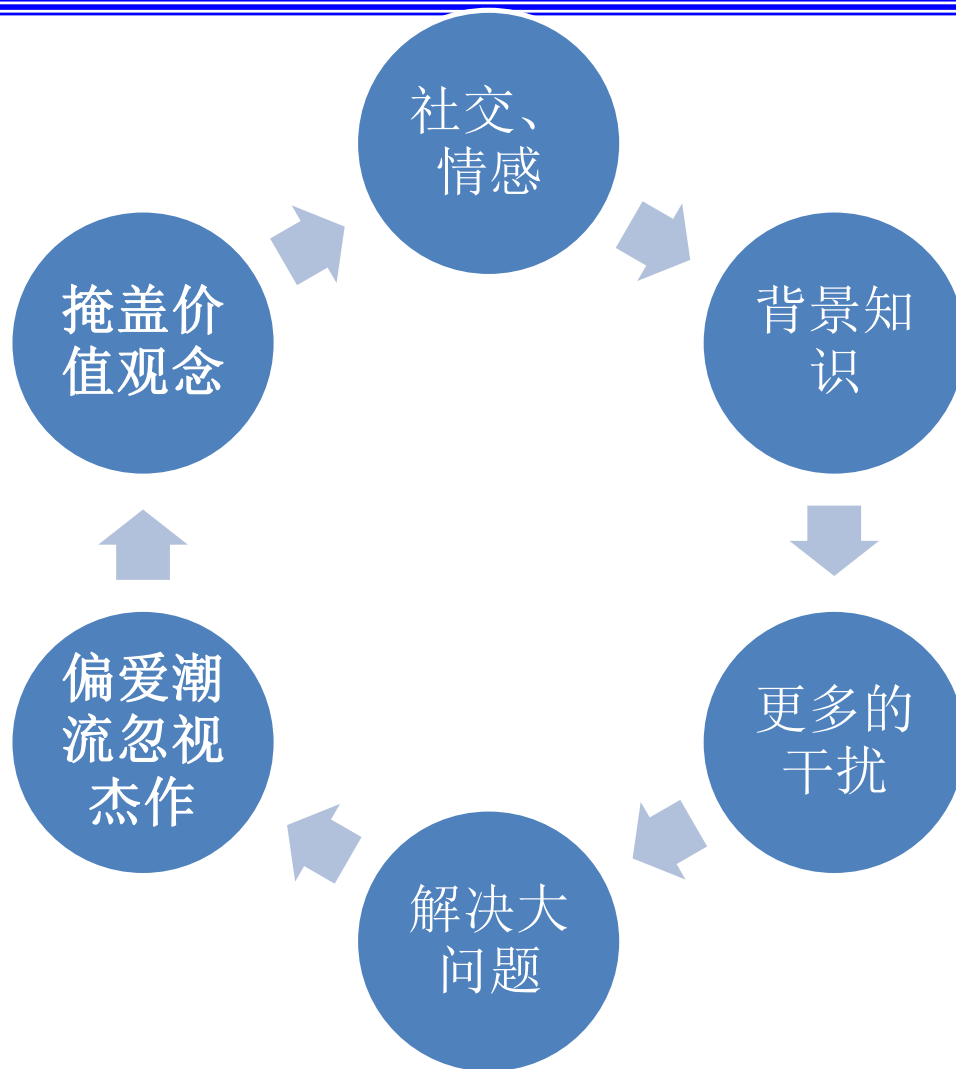
主要内容

- 互联网技术趋势
- 大数据应用概述
- 大数据的思考

大数据的相对性

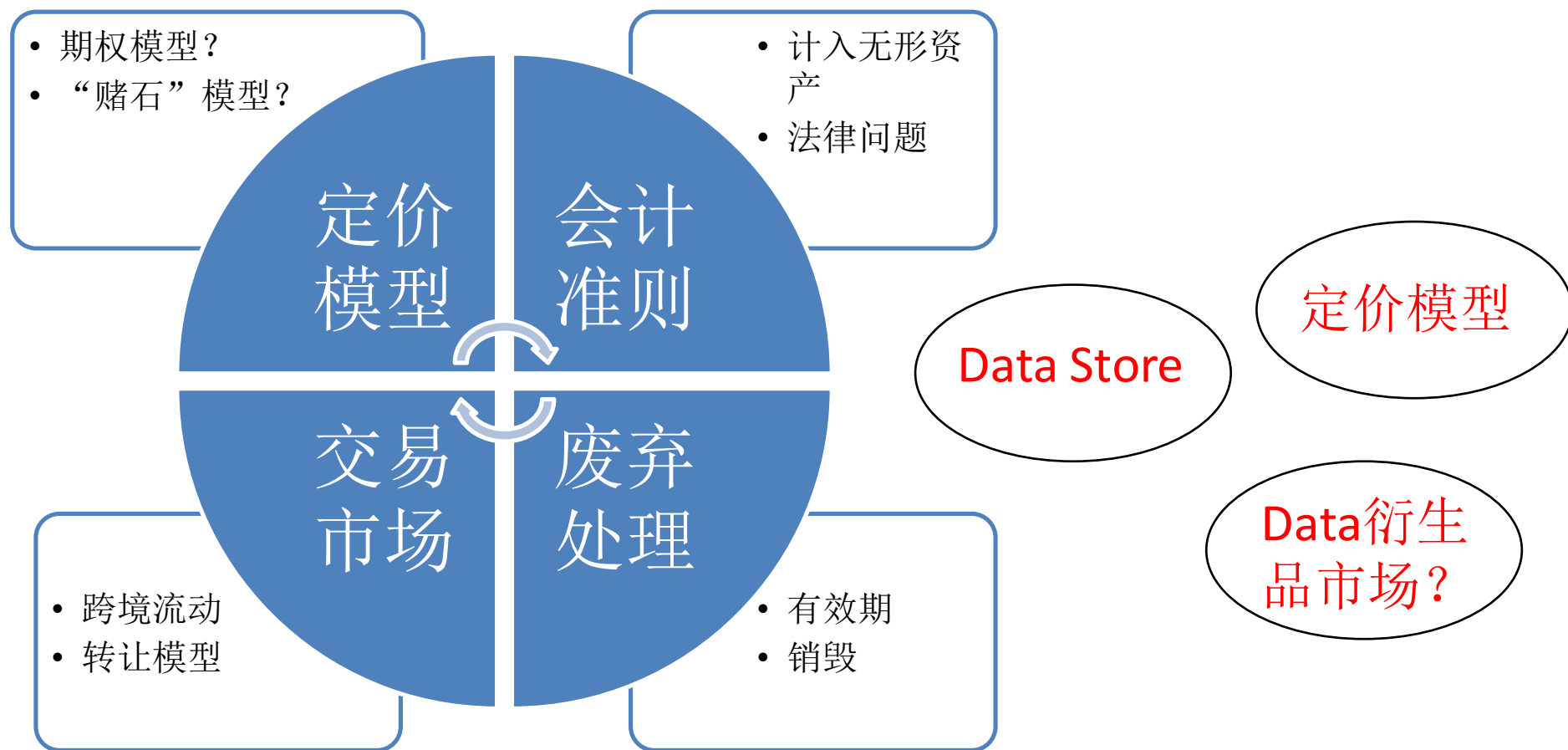


大数据做不到的



纽约时报

如何让大数据成为资产？



从精确到效率

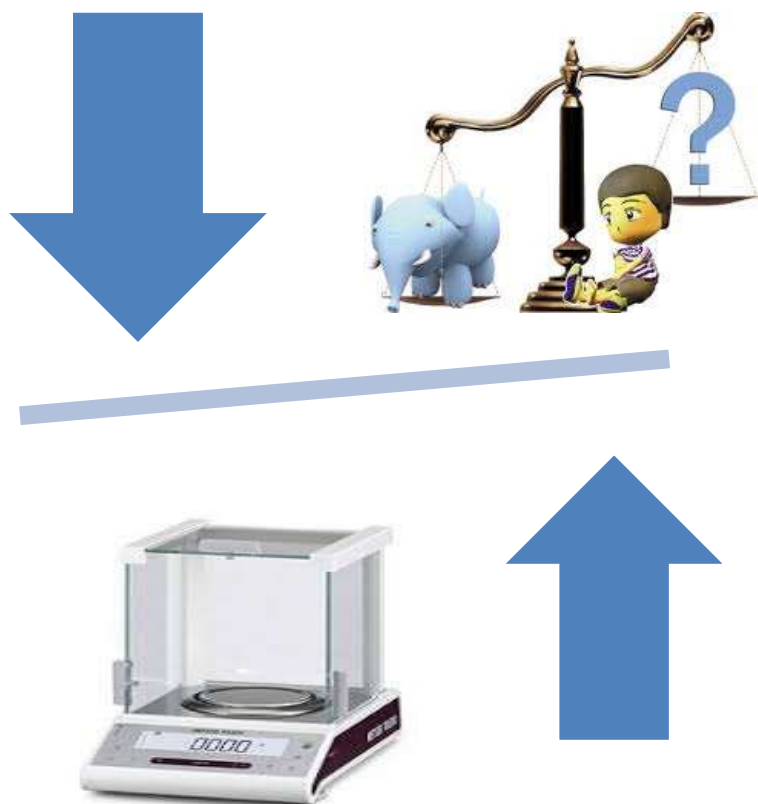
- 大量信息与精确性之间存在矛盾
- 算法的进步快于芯片，大数据优于算法
 - 用简单算法替代小数据的精确算法
 - 简单算法+海量数据？

“精而不确”
掷硬币的概率



图1-2 我们的数据丰富，但信息贫乏

从短缺到富裕



从因果到关联

因果关系是对世界的简化
因果关系存在吗？

机器会告诉是什么，没有为什么

- “丁蟹效应”

- 凡是播出由郑少秋主演的电视剧，恒生指数或A股已有32个明显下跌。女儿郑欣宜的电视剧，也有类似现象

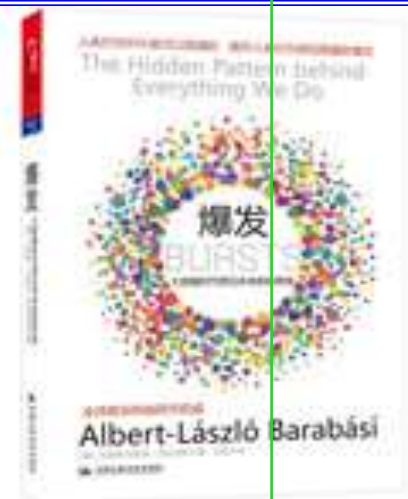


发改委会打飞机？

基金老鼠仓

大数据会颠覆社会学？

- 人类工作娱乐等活动具有间歇性，总是在长时间休息后突然爆发，然后沉寂
- 人类93%的行为是可预测的
- 自然界人类世界，都符合幂律



人类的行为是随机的，不可预测的
人类的行为具有传染性
人类需要的是自主性和随机智慧

沙滩上的大数据革命...



壁画
1.1万年前



乌龟壳
前14—11世纪



磁盘
3—4年



磁带
30年

如何保存
如何打开
如何检索
如何甄别



《妙法莲花经》
南北朝



羊皮圣经
4世纪

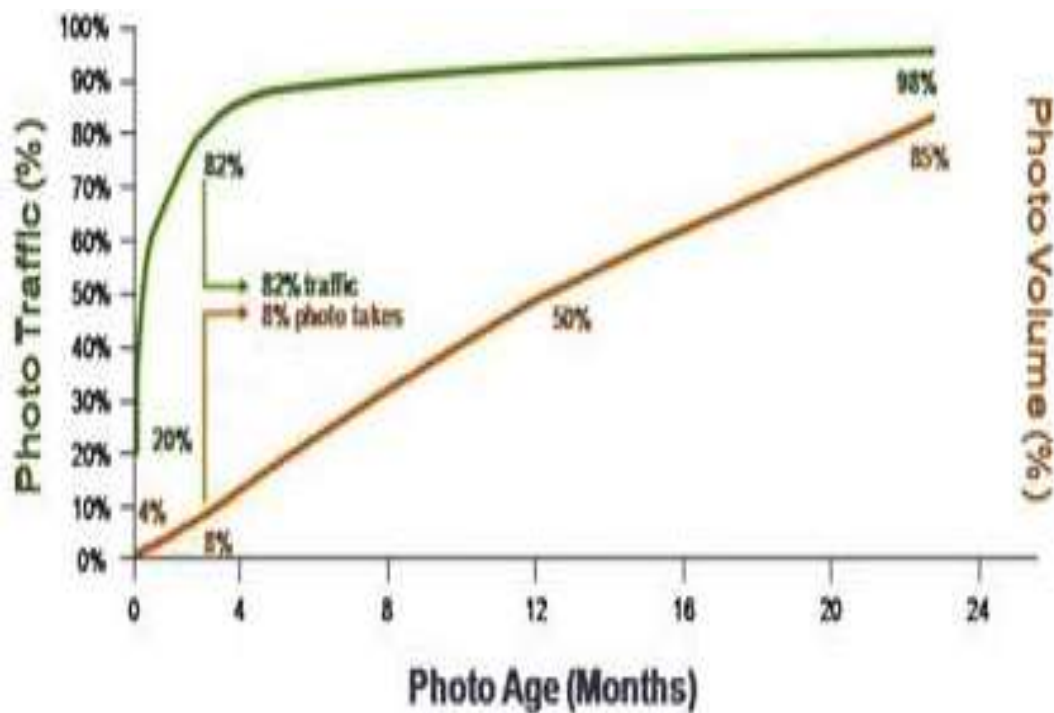


普通CD
20-50年

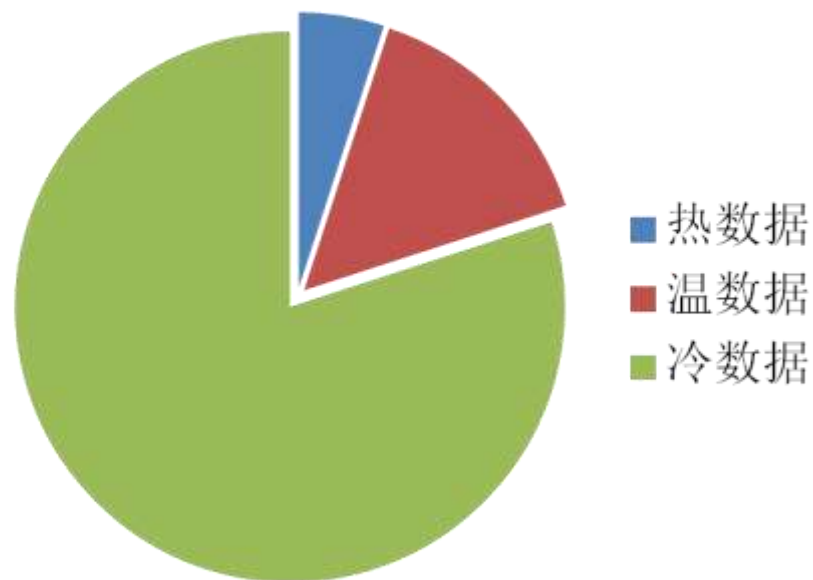


U盘
8—10年

数据的三种温度



占比



你没有隐私，忘记这事吧



大数据让匿名，在数学上已经不可能

20年前，“性别+邮编+出生年月日”，识别出87%的人

大数据时代，通过分析用户4个曾经到过的位置点，就可以识别出95%的用户

遗忘权成为基本人权

数据恒久远，一存永流传

世界经济论坛

— 没有坏数据，只有对数据的不合理使用

真实的谎言...

自从**2011年8月**以来，**Google**流感趋势在**108周**的时间里有**100周**出了错

去医院看流感的人，**80—90%**并非流感患者

波士顿：“颠簸的街道”和“爆炸案”

《纸牌屋》之后，Netflix自制的《铁杉树丛》《女子监狱》等其他剧集没有大热

Taget预测怀孕：没有怀孕的妇女也收到了

真正的现状

0 大数据将成为信息时代下一个引爆点



让数据发声!

谢谢

微信订阅号：何所思