

# 大数据分析

程学旗 靳小龙 刘盛华

## 授课团队

### ■ 主讲教师

#### □ 程学旗

- 中科院计算所研究员，博导，副所长
- Email: [cxq@ict.ac.cn](mailto:cxq@ict.ac.cn)

#### □ 靳小龙

- 中科院计算所研究员，博导
- Email: [jinxiaolong@ict.ac.cn](mailto:jinxiaolong@ict.ac.cn)

#### □ 刘盛华

- 中科院计算所副研究员，硕导
- Email: [liushenghua@ict.ac.cn](mailto:liushenghua@ict.ac.cn)



### ■ 助教

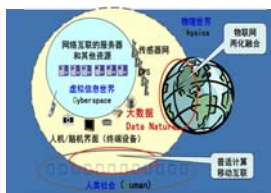
#### □ 赵凯琳


- Email: [zhaokailin17z@ict.ac.cn](mailto:zhaokailin17z@ict.ac.cn)

□ .....



## 大数据分析是大数据价值落地的关键环节



信息技术革命与人机物三元世界的交融  大数据  
(数量巨大、种类繁多、增长极快、价值稀疏的复杂数据)



大数据  $\xrightarrow{\text{分析处理能力}}$  大价值

- 美国政府大数据计划和Google 等大公司目前最重视的都是数据价值，着力于**大数据分析技术和系统的应用**；

## 我国当前痛点：大数据→小价值

据IDC统计数据显示，中国目前拥有的数据量占全球的14%，但数据利用率不到0.4%



**突破大数据分析技术瓶颈，推动大数据价值落地成为大数据领域的当务之急**

# 大数据分析书籍概览

书名	著作者	出版社	时间
大数据分析	王星等编著	清华大学出版社	2013年09月
大数据分析	张重生编著	机械工业出版社	2016年12月
大数据分析	[美]Simon Tatham著	南京大学出版社	2017年11月
大数据分析	[美]Michael Minelli, Michele Chambers, Ambiga Dhiraj著	人民邮电出版社	2014年08月
大数据分析	[美]Thomas H. Davenport编	机械工业出版社	2015年03月
大数据分析	[荷]Johan Bosman著	机械工业出版社	2016年03月
大数据分析的道与术	毕然编著	电子工业出版社	2016年04月
大数据分析方法	[美]Michele Chambers, Thomas W. Dinsmore	机械工业出版社	2016年08月
大数据分析计算机基础	张廷松 王成军 徐天晟	中国人民大学出版社	2016年07月
大数据分析原理与实践	王宏志	机械工业出版社	2017年07月
大数据分析	[美]Lawrence S. Maisel, Gary Cokins等著	人民邮电出版社	2014年11月
数据科学与大数据分析	[美]EMC教育服务团队	人民邮电出版社	2016年07月
Spark大数据分析	经管之家	电子工业出版社	2017年07月
Spark大数据分析	[美]Mohammed Guller	机械工业出版社	2017年05月
Spark大数据分析实战	高彦杰 倪亚宇	机械工业出版社	2016年01月
Python金融大数据分析	[德]Yves Hilpsch著	人民邮电出版社	2015年12月
群体智能与大数据分析技术	陶乾等	暨南大学出版社	2018年04月
社交媒体大数据分析	[美]Lutz Finger, Soumitra Dutta	人民邮电出版社	2016年10月

# 大数据分析书籍的副标题

书名	副标题
大数据分析	方法与应用
大数据分析	决胜互联网金融时代
大数据预测分析	决策优化与绩效提升
大数据分析	数据驱动的企业绩效优化、过程管理
大数据分析的道与术	
大数据探索性分析	
数据科学与大数据分析	数据的发现、分析、可视化与表示
大数据分析计算机基础	
大数据分析方法	用分析驱动商业价值
大数据分析	数据挖掘必备算法示例详解
大数据分析方法	
大数据分析原理与实践	
大数据分析	R语言实现
大数据分析	创造价值 做聪明的市场决策
Python金融大数据分析	
Spark大数据分析实战	
社交媒体大数据分析	理解并影响消费者行为
Spark大数据分析	核心概念、技术及实践
Spark大数据分析	技术与实战
.....	

16本利用副标题对内容作了进一步的阐释和限定，说明对大数据分析的内涵、内容设置、定位等有多种不同的理解

# 大数据分析教材

书名	著作者	出版社	内容简介	读者对象
大数据分析原理与实践	王宏志	机械工业出版社 (2017)	介绍了大数据预处理，数据仓库，以及分类、聚类、关联分析、结构分析和文本分析模型；大数据的并行、流式与图分析平台等	计算机科学专业的本科生或者研究生，也可以作为从事大数据相关工作的工程技术人员的参考用书
大数据分析：方法与应用	王星等编著	清华大学出版社 (2013)	主要介绍数据挖掘、统计学习和模式识别中与大数据分析相关的理论、方法及工具	统计学、管理学、计算机科学等专业的高年级本科生或研究生
大数据探索性分析	吴翌琳, 房祥忠	中国人民大学出版社 (2016)	介绍大数据的预处理、采样、大数据探索性分析案例、数据可视化等	有统计学基础的硕士研究生，统计专业高年级本科生

## 课程安排

40个课时；14次课；2个学分

	时间	内容
1	09. 03	大数据与数据科学简介 大规模数据计算简介
2	09. 10	大规模机器学习
3	09. 17	大图挖掘 I
4	09. 24	大图挖掘 II 流式数据、时间序列分析
	国庆假期	
5	10. 08	文本数据处理
6	10. 15	知识计算 I
7	10. 22	知识计算 II
8	10. 29	期中复习和评估



高等教育出版社  
程学旗 主编

## 课程安排

	时间	内容
9	11. 05	社交媒体分析
10	11. 12	跨媒体数据分析
11	11. 19	大数据分析系统
12	11. 26	大数据分析应用、总结与展望
13	12. 03	大作业：分组Poster报告、Q&A
14	12. 10	闭卷考试



高等教育出版社  
程学旗 主编

# 第一讲：大数据与数据科学简介

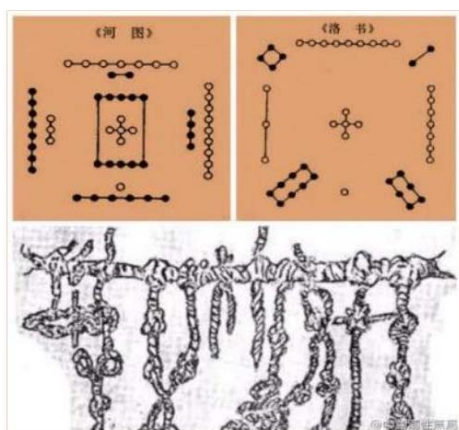
## 目录

- 1.1 大数据的来源、特征与影响
- 1.2 国际大数据技术发展现状
- 1.3 我国的大数据发展战略
- 1.4 我国的大数据技术与产业现状
- 1.5 我国大数据发展面临的问题
- 1.6 数据科学简介

# 1.1 大数据的来源、特征与影响

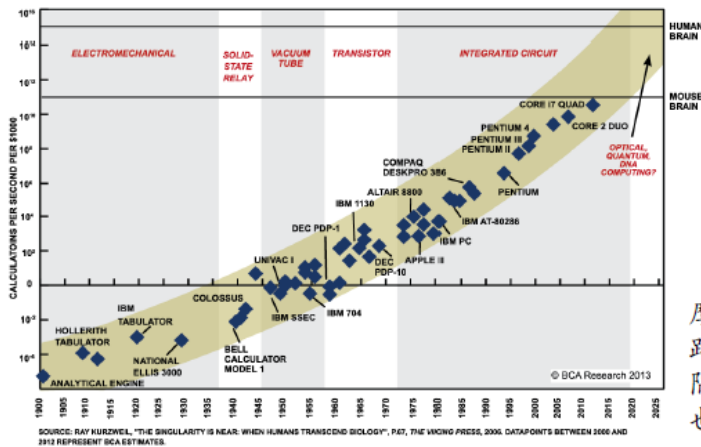
## 数据(信息)的重要性

- 数据的获取、处理与应用在人类社会发展一直扮演着重要角色
  - 文明之初的“结绳记事”、文字发明后的“文以载道”，到近现代科学的“数据建模”，承载了人类基于数据(信息)认识世界的努力和巨大进步
- 信息技术为数据处理提供了自动的方法和手段，推动数据(信息)成为继物质、能源之后的第三大战略资源





# 信息技术的指数发展模式



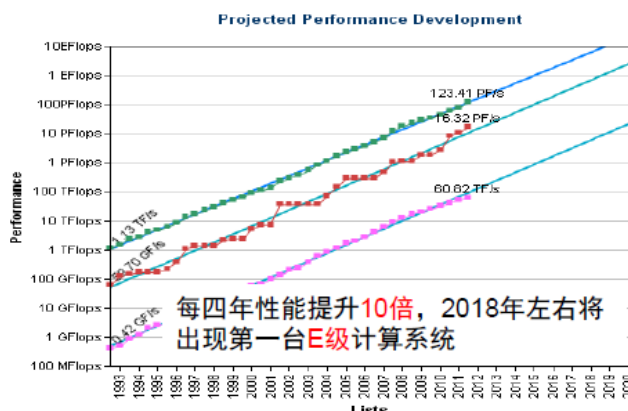
美国芯片厂商英特尔（Intel）创始人之一戈登·摩尔（Gordon Moore）于1965年提出摩尔定律，该定律已经在过去半个多世纪，到目前为止仍与实践相符。

摩尔定律

摩尔定律：当价格不变时，集成电路上可容纳的元器件数目，约每隔18-24个月便会增加一倍，性能也将提升一倍。

- 现代信息技术自创立以来，一直遵循摩尔定理呈指数发展模式，高速发展
- 信息技术及其应用（信息化）广泛并深刻地影响和改变了人类社会，而且这种作用正在加强！无处不在的信息技术深度应用甚至将重构人类社会！

## 计算能力的提升：超级计算机



1996年，日本Hitachi SR2201

- 2048个处理器
- 速度峰值：浮点运算6亿次/秒



2017年6月，神威·太湖之光

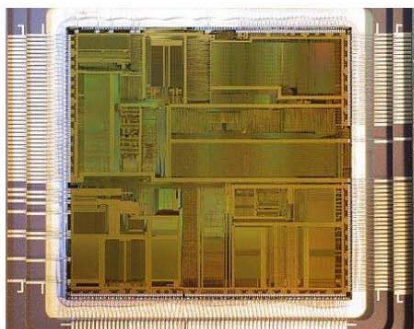
- 40960个中国自主研发的“申威26010”众核处理器
- 双精浮点峰值：12.5亿亿次/秒
- 持续性能为9.3亿亿次/秒

20多年，全球速度最快的超级计算机速度提升2.08亿倍

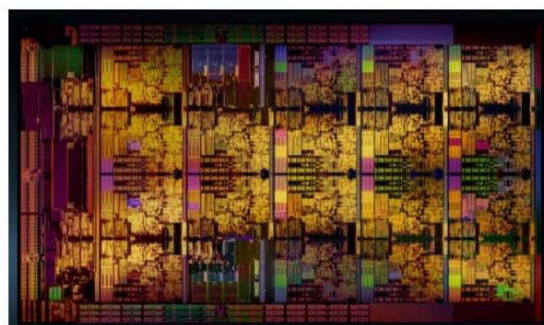
“神威·太湖之光”一分钟的计算能力相当于全球72亿人口同时用计算器连续不间断计算32年；这是全球第一台运行速度超过10亿亿次/秒的超级计算机

# 计算能力的提升：个人计算机

## 主流个人计算机



1993年 Intel Pentium  
(奔腾) CPU  
共集成310万个晶体管  
线宽0.8μm



2017年 AMD Ryzen (锐龙)  
ThreadRipper CPU (16核32线程)  
共集成92亿个晶体管  
线宽14nm  
主频3.4GHz

集成晶体管数量提升2968倍，  
主频提升57倍

人类头发直径0.05毫米，将其径向平均剖成2500份，一份就是AMD RYZEN CPU中相邻两条线路之间的宽度

# 计算能力的提升：个人计算机

## 移动智能终端对比



1994年IBM推出的人类历史上首部智能手机，IBM Simon，除打电话外，还可以收发传真和电子邮件。



智能终端体积迅速减小，价格迅速下降



IBM Simon (1994)

重量：510g  
CPU: Vadem 16MHz  
内存：1M字节  
网速：Modem 2.4Kbps  
待机：1小时  
功能：电话、传真、电邮



华为 Mate 10 (2017)

186g  
海思8核，2.36GHz  
128G字节  
4G网 100M~150M bps  
2~3天  
游戏、办公、摄像、定位……



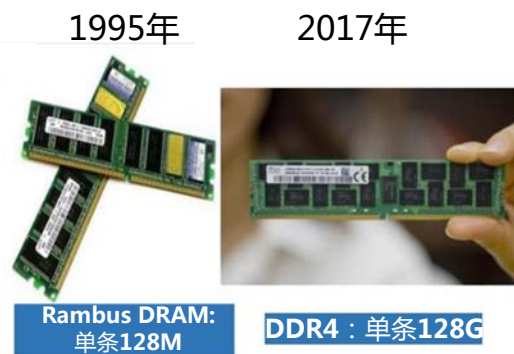
1948年ENIAC计算机  
每秒5000条指令  
50万美元



2014年智能手机  
每秒10亿条指令  
约数千元人民币



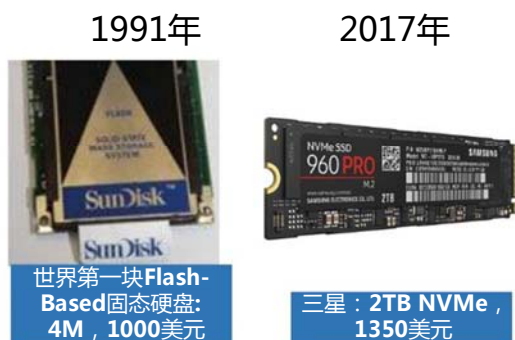
# 存储容量的提升



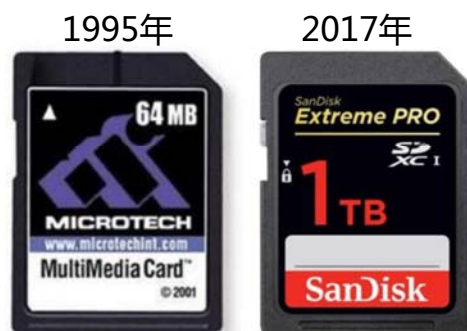
内存：提升**1000**倍



机械硬盘：提升**12000**倍



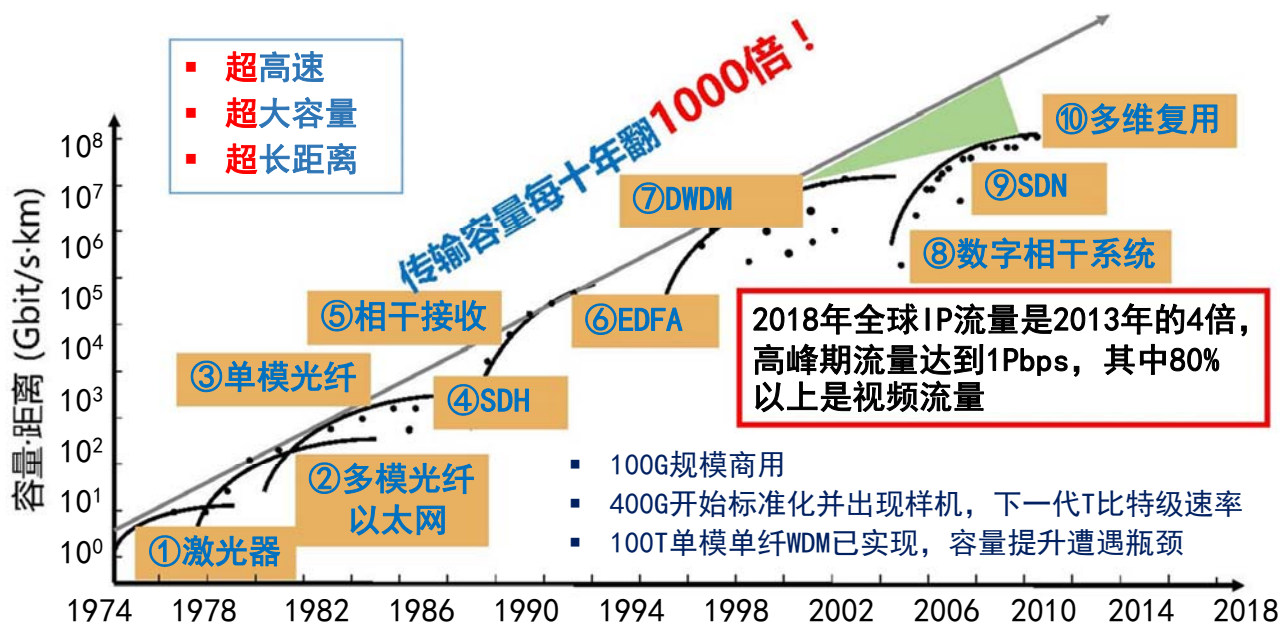
固态硬盘：提升**50万**倍



闪存卡：提升**16000**倍

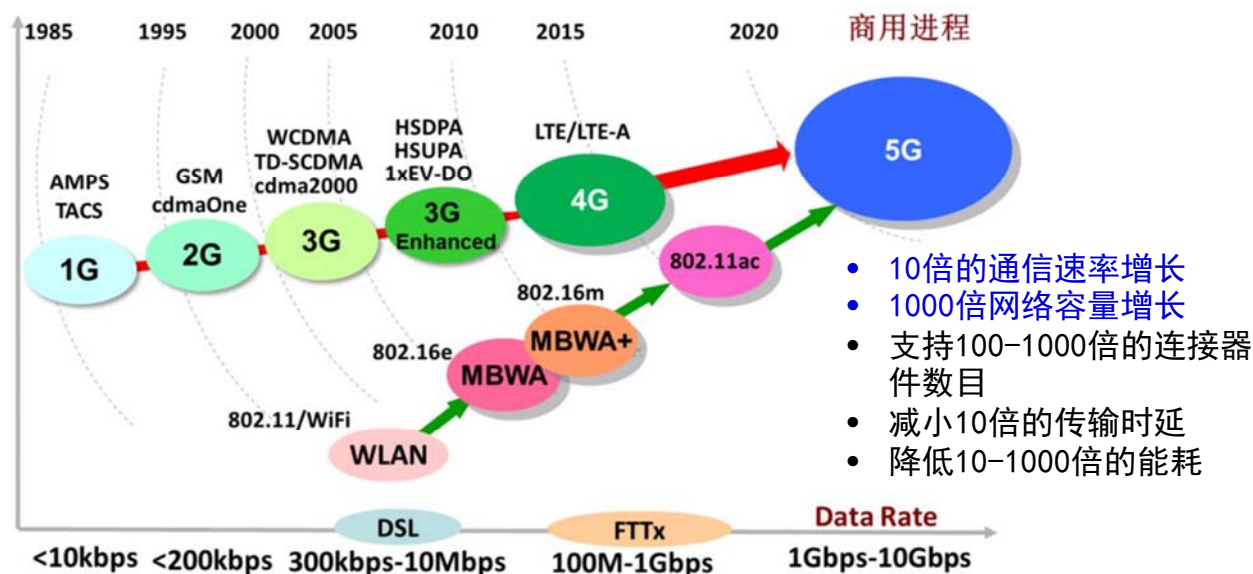
# 通信带宽的提升

- 宽带通信技术的核心器件和光网络创新推动**传输速率、容量和距离**的持续提升，数字媒体流量的猛增需要Pb级传输



# 无线通信速率的提升

- 无线通信每隔5年一次换代，速率提升10倍



# 无线通信速率的提升

- 移动通信网络速率与用户体验

	2G(1992~)	3G(2003~)	4G(2013~)	5G(2020~)
理论速度上限	1M bps	22M bps	128M bps	≥1Gbps
实测速度 (约)	10~20K bps	4M bps	60~80M bps	-

下载一部电影2G  
字节



2G: 9.25天  
3G: 66分钟  
4G: 200秒  
5G: 约30秒

下载一首歌曲  
5M字节



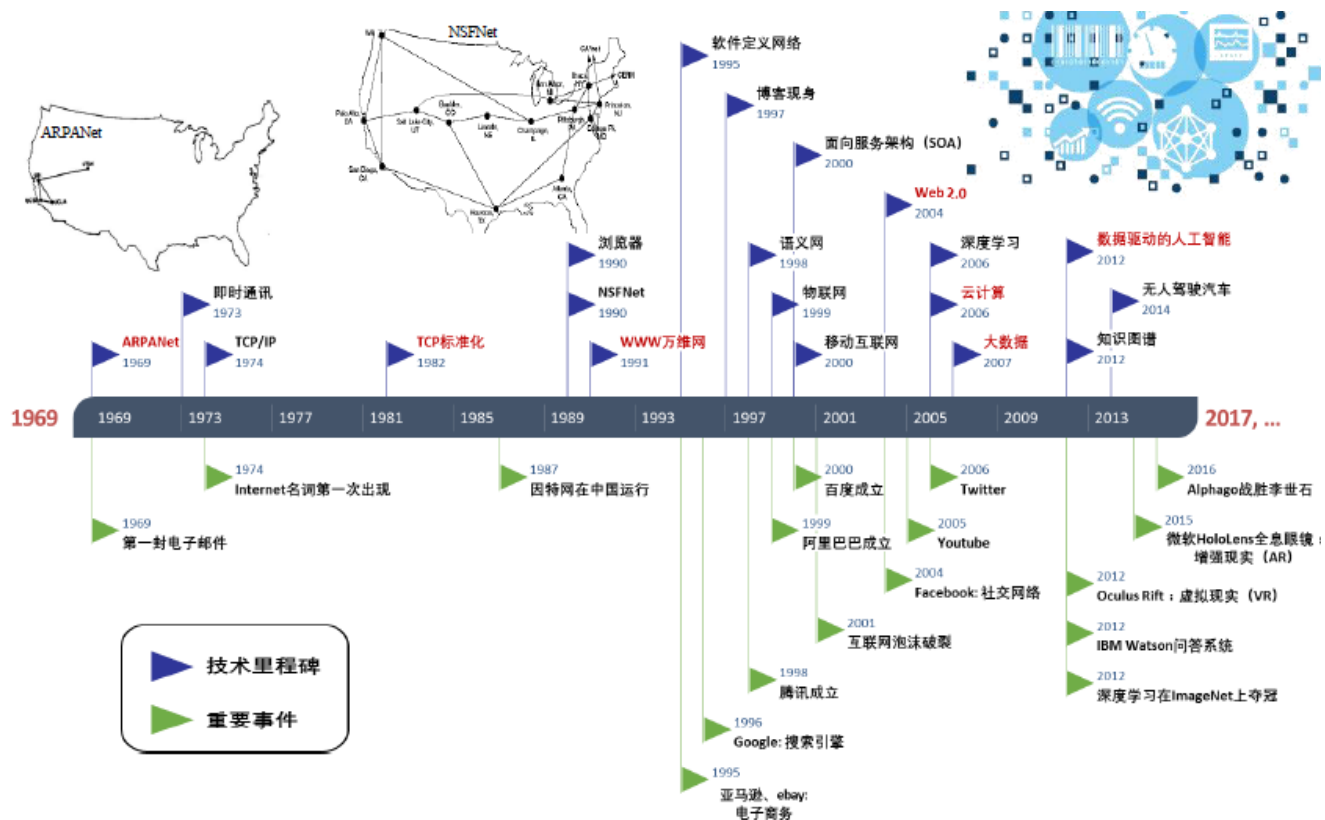
2G: 33分钟  
3G: 10秒  
4G: 0.5秒  
5G: 约0.075秒

接收一封邮件  
100K字节



2G: 40秒  
3G: 0.2秒  
4G: 0.01秒  
5G: 约0.0015秒

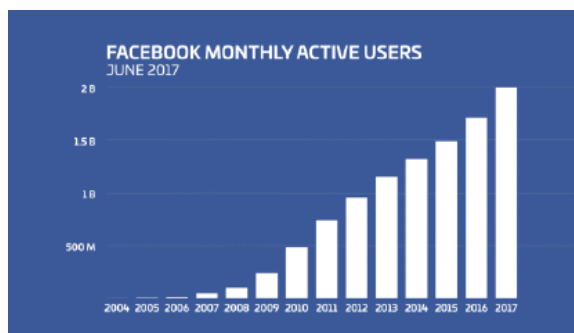
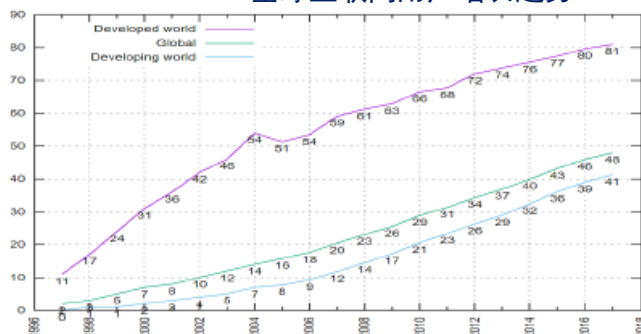
# 互联网及其发展



## 互联网的快速发展与普及率的提升

- 1997年全球访问互联网人口比例2%，2017年上升到48%，达到36亿

1997-2017全球互联网用户增长趋势



- Facebook 2017年月活跃用户达20亿，是1995年整个互联网用户数的44倍
- 若将Facebook看成一个国家，2017年已经接近中、美、欧盟人口之和

# 互联网的快速发展与普及率的提升



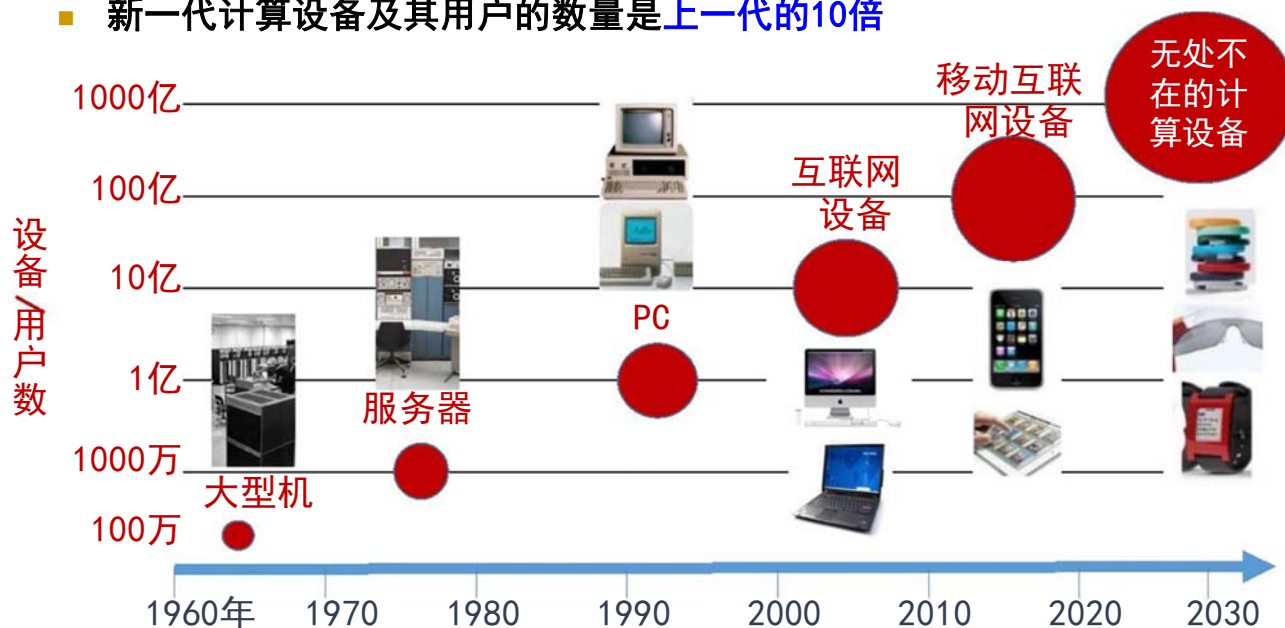
- 2017年8月，全球移动互联网用户超过**35亿**，较2007年增长**7.8倍**，智能手机用户超过**23亿**，人口渗透率为**32.3%**。**中国移动互联网人口渗透率**已达**47.2%**

- 根据彭博社的报道，美国成年人每天平均花费**2小时57分钟**访问手机。人们花费在手机上的时间已经超过电视



## 计算设备数量的快速增长

- 每过10-15年左右，产生新一代计算技术与系统
- 新一代计算设备及其用户的数量是**上一代的10倍**



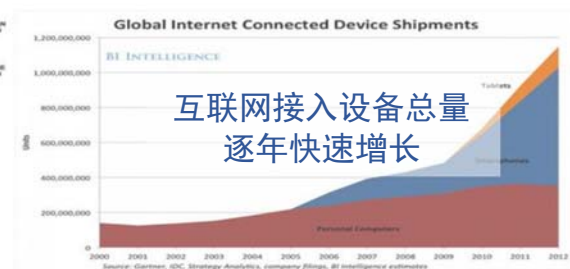
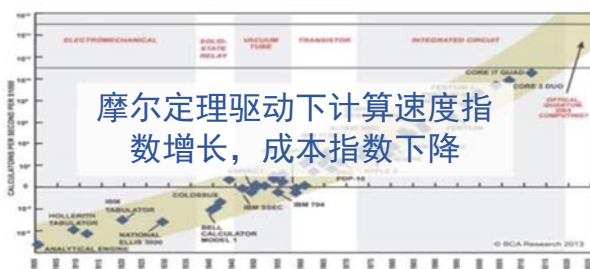
\*PC在1993年达到1亿台；互联网设备在2005年达到10亿；移动互联网设备预计2016年达到100亿；物联网设备预计2020年达到300-500亿——ITU, Cisco, Erricson, ABI...



# 互联网及其延伸导致大数据现象

大数据源于**信息技术的不断廉价化**与互联网及其延伸所带来的**无处不在的信息技术应用**。四个驱动：

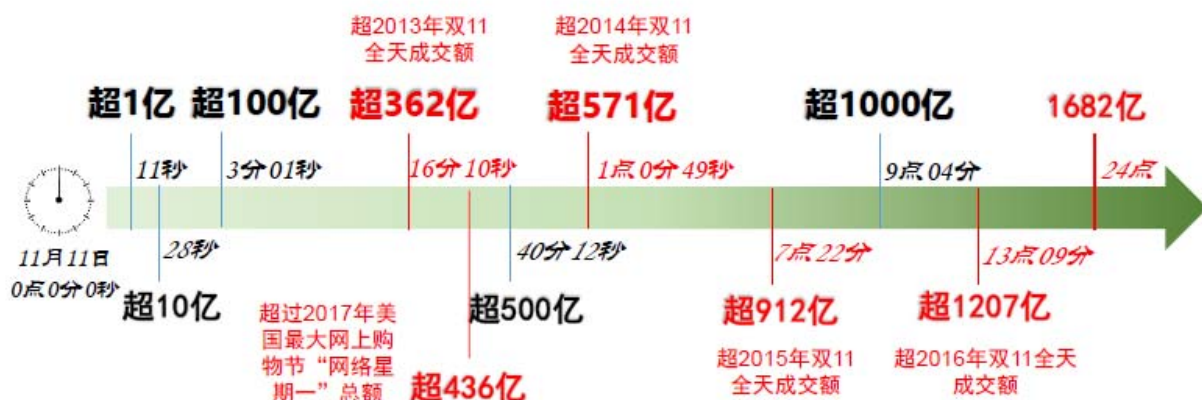
- 摩尔定律驱动的指数增长模式
- 技术低成本化驱动的万物数字化
- 宽带移动泛在互联驱动的人机物广泛联接
- 云计算模式驱动的数据大规模汇聚



## 大数据规模汇聚一例



- 2017天猫双11成交额**1682亿元** 支付达**14.8亿笔**
- 开场5分钟22秒，支付宝的支付峰值达到**25.6万笔/秒**
- 数据库处理峰值：**4200万次/秒**

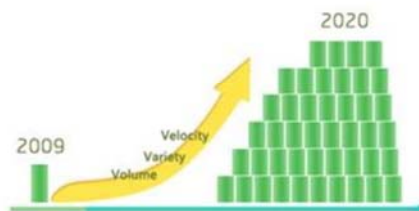
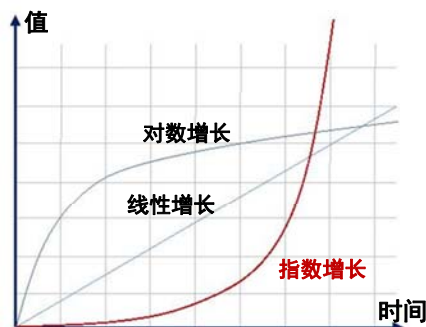




# 数据的指数增长

## 全球数据总量统计（IDC）

- 2003年全球产生数据仅 5百万 TB
- 2009年全球产生数据约 0.8 ZB
- 2012年全球产生数据约 2.8 ZB
- 2020年全球产生数据约 44 ZB
- 2030年全球产生数据约 2500 ZB



1KB = 1024 B  
1MB = 1024 KB  
1GB = 1024 MB  
1TB = 1024 GB  
1PB = 1024 TB  
1EB = 1024 PB  
1ZB = 1024 EB

## 1ZB有多大？

假设一部时长为2小时的电影存储成蓝光DVD，大小为2GB，

那么1ZB的数据存储空间可存储蓝光DVD电影5500亿部，如果全部看一遍，需要1.3亿年

# 何为大数据？

## 技术能力的视角

- 大数据指的是**规模超过现有数据库工具获取、存储、管理和分析能力**的数据集。并不是超过某个特定数量级的数据集才是大数据。

——麦肯锡《大数据：创新、竞争和生产力的下一个前沿领域》

## 大数据内涵的视角

- 大数据是具备**海量、高速、多样、可变**等特征的多维数据集，需要通过**可伸缩的体系结构**实现高效的存储、处理和分析。

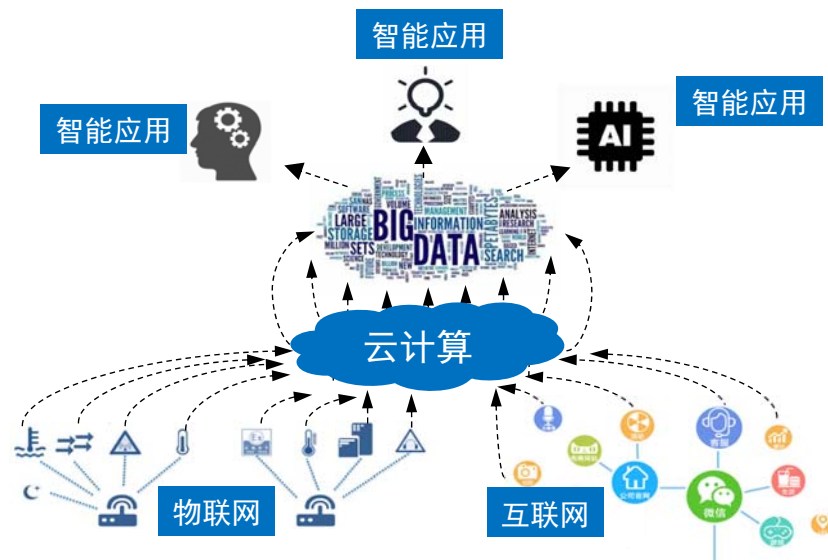
——NIST《大数据白皮书》

# 大数据的基本特征



## 大数据与物联网、互联网、云计算、人工智能

- 物联网和互联网产生的数据成为大数据的主要来源
- 云计算为大数据运算提供核心硬件和软件支撑
- 人工智能是驱动创新应用的重要引擎
- 大数据为人工智能引擎提供必要的燃料，二者缺一不可



# 大数据的认知误区

## 大数据 $\neq$ 数据中心

数据中心(IDC)是对互联网业务资源进行集中式处理和分发的物理环境。在大数据产业的传输层,是大数据应用的网络基础设施

## 大数据 $\neq$ 云计算

云计算是互联网业务的系统平台,实现海量数据的高效存储和利用。在大数据产业的物理层,是大数据应用的系统基础设施

## 大数据 $\neq$ 数字化信息

数字化信息是大数据的组成部分,但不是所有的数字化信息都能产生大数据。大数据是数字化信息被生产、消费的过程的记录

## 大数据 $\neq$ 海量数据

海量是大数据的特征之一,但如前所述,大数据并不简单地指海量的数据,而是具有nV特性的海量数据

# 大数据催生了大量新的互联网应用

- 马云说阿里巴巴实质上是一家“数据”公司,未来发展的是DT(数据科技)



**互联网+金融:** 大数据征信既基于传统信贷数据,又基于社交、电商数据等互联网大数据



**互联网+餐饮:** 根据搜索数据、浏览数据、交易数据以及评论数据,呈现了全国消费热图



**互联网+交通:** Uber数据显示,乘客平均等车时间,目前北京为5.7分钟,上海为5.1分钟,美国旧金山为2.3分钟

# 大数据给国家安全保障带来了新的挑战

为攻打叙利亚，充分利用网络媒体炒作，一步步制造对其有利的国际舆论环境



美政府欲对叙动武  
其实际动机：

1. 能源问题
2. 国内危机
3. 地缘政治
4. 大国博弈



## 化武事件

2013年8月21日发生在叙利亚大马士革东部郊区Ghouta的化学武器攻击事件  
BBC报道，三名反对派人士于8月21日凌晨在Facebook首先发布了化武攻击的消息



美国抢先发表报告：化武致死1429人

美国政府8月30日公开一份解密报告，以“大量独立消息来源”为依据，声称有“充分信心”认定叙利亚当局21日对首都大马士革郊区一些地区动用化学武器，造成大量人员伤亡



白宫公布叙利亚化武袭击新视频

9月7日美国参议院情报委员会传出13部化武视频，并声称这些视频都是来自美国视频网站YouTube视频在社交媒体、主流媒体上广泛传播，观看次数均在千万次以上。对全球民众的心理产生了重要影响

同类情况在伊拉克战争、中东颜色革命中屡见不鲜

美国通过网络空间和大数据手段，宣传其价值观、国家意志，形成强大的影响力，牢牢掌握了其在国际社会上的政治、经济和军事上的话语权。如颜色革命、伊拉克战争、中东地缘事件等活动中，均产生了重要作用

# 大数据给政府的社会治理带来新的机遇

## 政府1.0

### MIS

- MIS系统、OA系统为代表的信息技术的广泛应用，实现了政府办公自动化，提升办公效率

## 政府2.0

### e-Gov

- 互联网技术，使政府高质量数据得以汇聚和积累
- 推动了政府信息公开、政务公开，提升服务水平

## 政府3.0

### smart-Gov

- 大数据挖掘、处理和智能分析能力将大大重构政府工作流程
- 透明政府、责任政府、智慧政府
- 政府数据是最有价值的“金矿”



# 大数据引发科学研究思维与方法的变革



# 大数据给产业经济带来巨大动能

- 大数据正在引发深刻的技术与产业变革，为我国加速**产业经济的转型和结构调整**带来了**巨大动能**



贵州大数据发展的三步走战略

- 2016年贵州GDP达到11734亿元, 增速为10.5%;
- 贵州的GDP增速全国前三;
- 其中, **贵州以大数据为引领的“大数据+”产业增加值比2015年增长66.6%, 成为贵州GDP增长的大功臣**



# 大数据给家庭生活带来个性化的变化

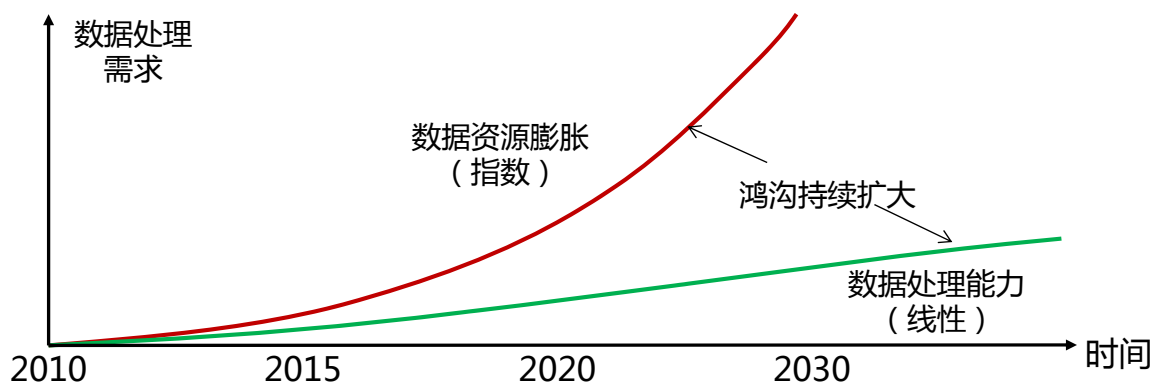
- 万物互联时代的智能应用背后，是对人们生活习惯、作息时间等大数据分析后形成的**智慧服务**模式

□ 如：灯光的柔和度、室内的温度、适时的音乐等



# 大数据给信息技术带来了挑战

信息技术能力提升落后于数据体量增长，大数据现象将长期存在



当前全球企业存储的数据中，**52%**的数据为产生后被存储，却从未被处理和利用，价值尚不明确的“**暗数据**”

——存储管理软件供应商 Veritas 2016年《数据冰山报告》

## 1.2 国际大数据技术发展现状

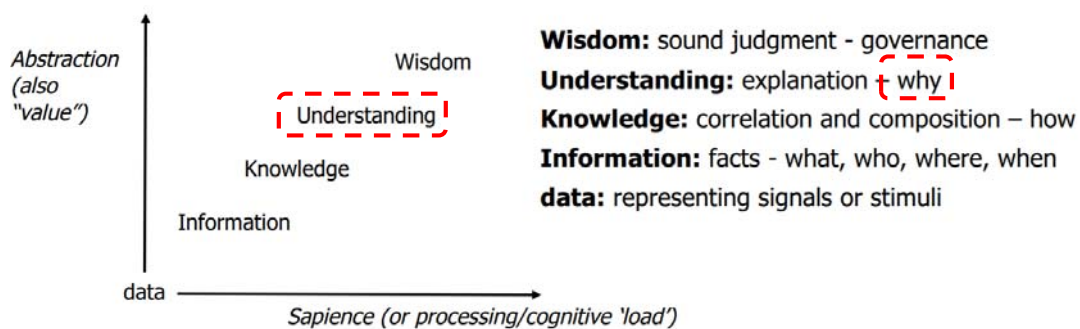
### 国际大数据技术发展现状：美国

- 美国是大数据技术的先驱
  - 军 方：DARPA驱动大数据军事应用
  - 工业界：谷歌开启大数据技术创新
  - 学术界：奠定了大数据基础理论



# 国际大数据技术发展现状：美国

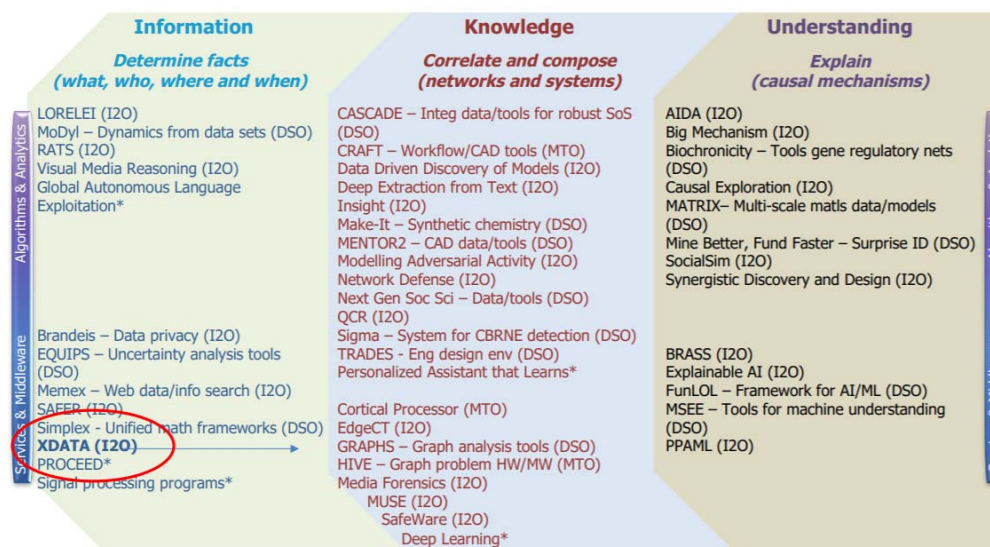
- 美国军方——敢于尝试大数据新技术
  - 由DARPA主导，清晰的数据创新思路



DARPA的数据分析路线图

# 国际大数据技术发展现状：美国

- 美国军方——敢于尝试大数据新技术
  - DARPA围绕数据科学布局了大量项目

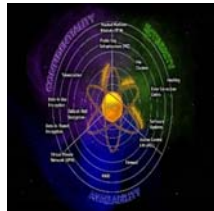


# 国际大数据技术发展现状：美国

- 美国军方——敢于尝试大数据新技术
  - DARPA自2012年起支持10项大数据课题



Anomaly Detection at Multiple Scales



Cyber-Insider Threat



Insight



Machine Reading



Mind Eye



Mission-oriented Resilient Clouds



Computation on Programming Encrypted Data



Video and Image Retrieval and Analysis Tool



X-DATA



Data to Decisions

# 国际大数据技术发展现状：美国

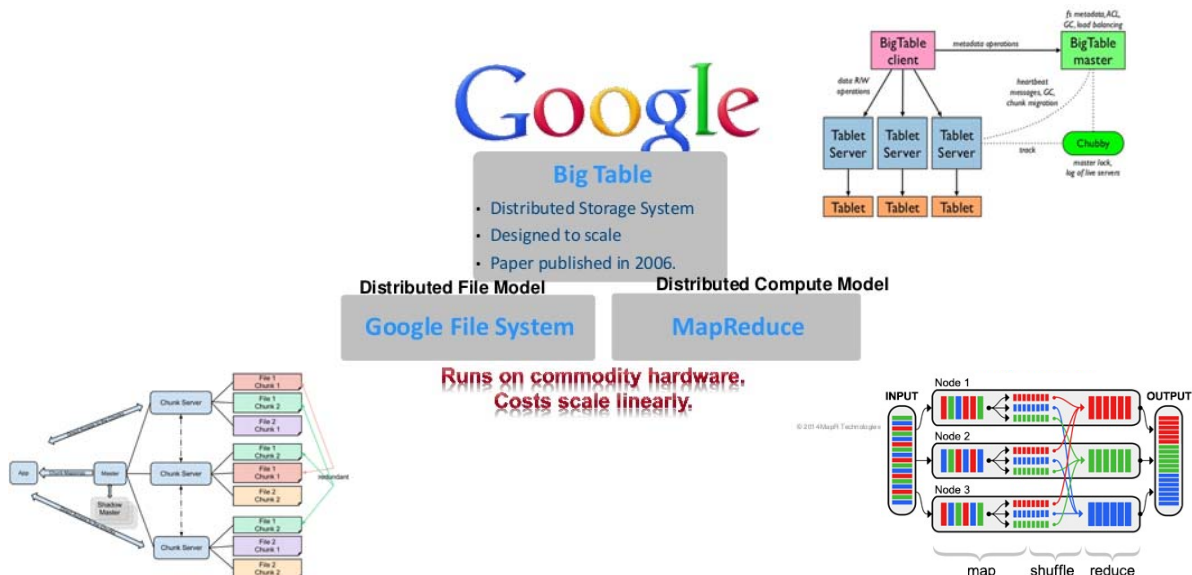
- 美国军方——敢于尝试大数据新技术
  - X-DATA是大数据技术的幕后推手

技术领域	技术名称
编程工具	Python, SciPy, SciDB, Julia
大数据平台	Spark, Mesos, Blaze, BayesDB
机器学习工具	Skylark, SNAP
自然语言处理	MITIE, Topic
可视化	Tangelo, D3, Vega

# 国际大数据技术发展现状：美国

## ■ 美国工业界——引领大数据技术创新

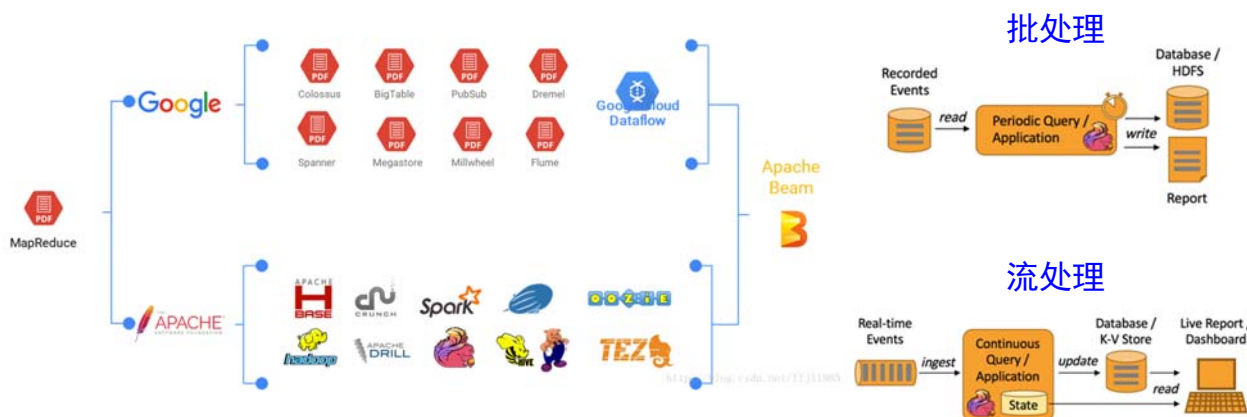
- 谷歌的三驾马车，大数据技术的奠基之作
- GFS (2003)，MapReduce (2004)，BigTable (2006)



# 国际大数据技术发展现状：美国

## ■ 美国工业界——引领大数据技术创新

- 谷歌最新的Beam架构，引领大数据技术变革



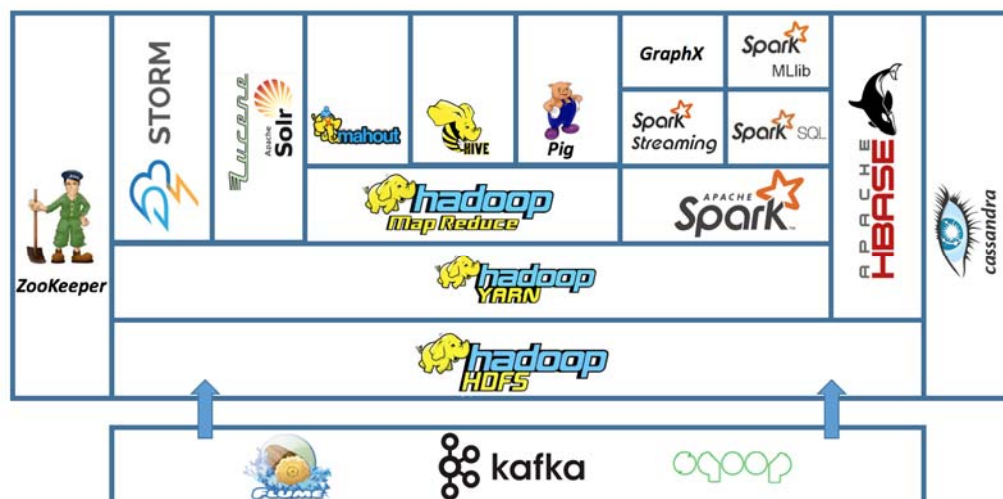
统一了数据批处理和流处理编程范式，能在任何执行引擎上运行



# 国际大数据技术发展现状：美国

## ■ 美国工业界——引领大数据技术创新

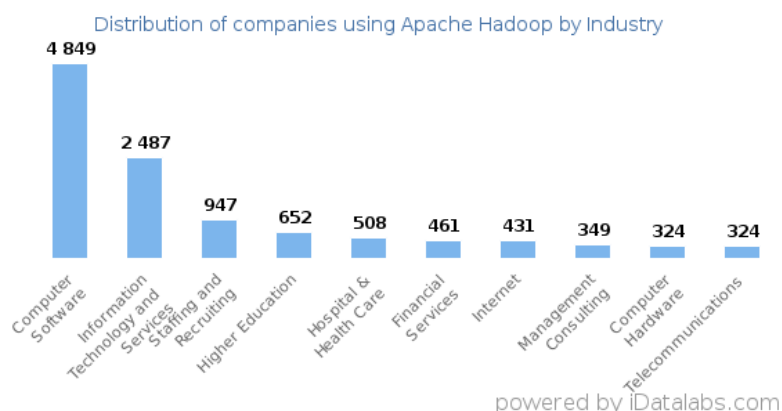
- Apache 基金会开创大数据技术开源时代



# 国际大数据技术发展现状：美国

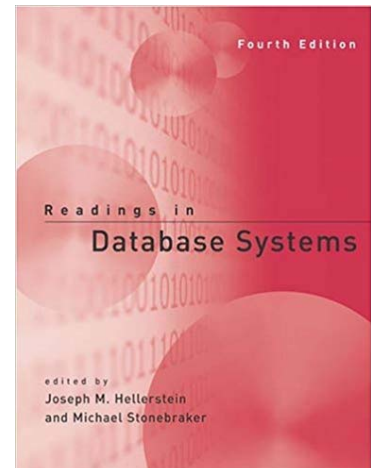
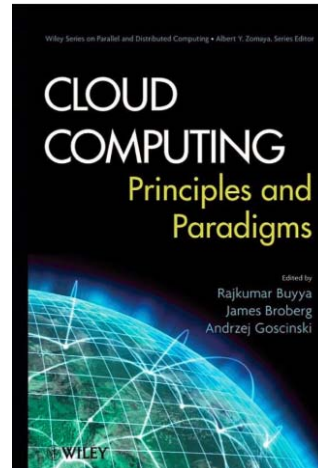
## ■ 美国工业界——引领大数据技术创新

- 全球应用Apache开源大数据技术的公司统计
- 国内的BAT、京东、小米都是其重度用户



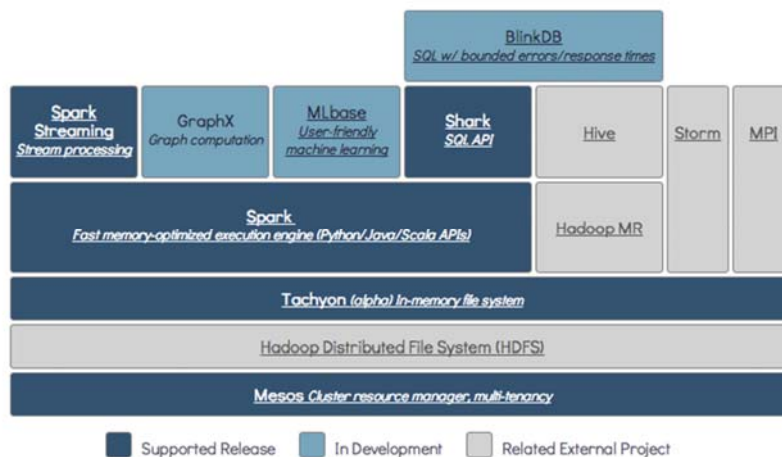
# 国际大数据技术发展现状：美国

- 美国学术界——定义大数据基础理论
  - 麻省理工、斯坦福、卡内基梅隆等科研机构
  - 定义分布式计算、数据库系统等一系列基础理论



# 国际大数据技术发展现状：美国

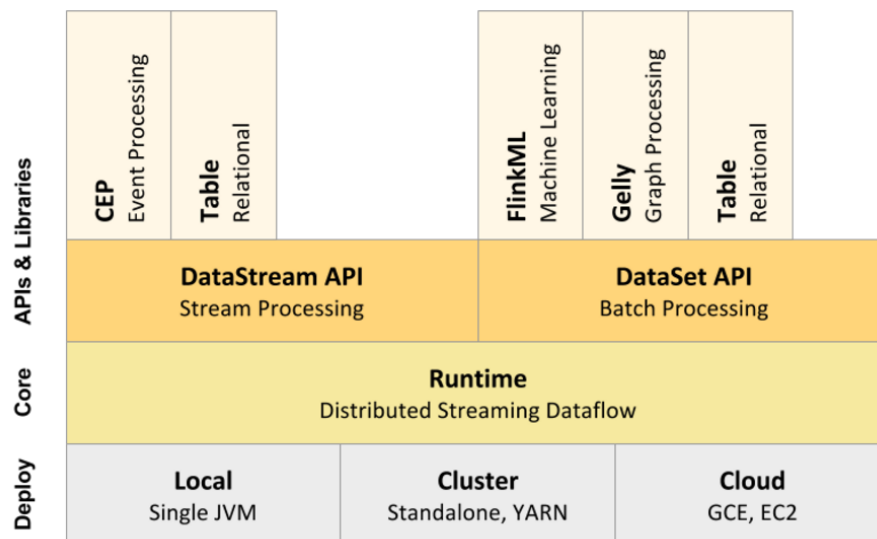
- 美国学术界——定义大数据基础理论
  - University of California, Berkeley
  - 内存计算理论架构



# 国际大数据技术发展现状：欧洲

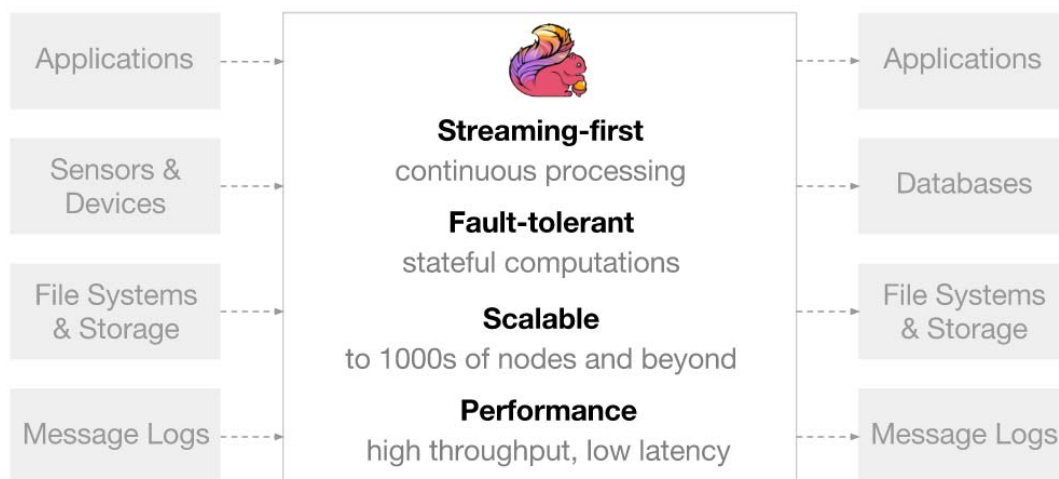
## ■ 欧洲——开创大规模流计算新范式

### □ Apache Flink



# 国际大数据技术发展现状：欧洲

## ■ 欧洲——开创大规模流计算新范式



# 国际大数据技术发展现状：欧洲

- 欧洲——定义大数据新语言
  - ❑ Scala语言由洛桑理工大学开发
  - ❑ 大数据时代的新语言
  - ❑ 被Spark、Kafka等广泛采用



# 国际大数据技术发展现状：欧洲

- 欧洲——定义大数据新语言
  - ❑ Scala语言具有简洁、高效的特点
  - ❑ 特别适合大规模数据处理和挖掘分析

```
1 import scala.collection.mutable.HashMap
2 import org.apache.spark.SparkConf
3 import org.apache.spark.SparkContext
4 import org.apache.spark.rdd.RDD
5 import org.apache.spark.sql.SparkSession
6 import org.apache.spark.sql.functions._
7 import org.apache.spark.sql.functions._
8 import org.apache.spark.sql.functions._
9 import org.apache.spark.sql.functions._
10 import org.apache.spark.sql.functions._
11 import scala.collection.mutable.HashMap
12
13 object WordCount {
14   def main(args: Array[String]) {
15     SparkConf conf = new SparkConf().setMaster("local[*]").setAppName("wc")
16     SparkContext sc = new SparkContext(conf)
17     //read a text file
18     val text = sc.textFile("hdfs://vagrant:spark@192.168.1.100:8020/words.txt")
19     //flatMap
20     val words = text.flatMap(new FlatMapFunction[String, String]() {
21       override def apply(line: String): Iterable[String] = {
22         line.split(" ")
23       }
24     })
25     //groupByKey
26     val wordGroups = words.groupByKey()
27     //mapValues
28     val wordCounts = wordGroups.mapValues {
29       case Iterable(word) => word
30     }
31     //reduceByKey
32     val wordCounts = wordCounts.reduceByKey {
33       case (word, count1, count2) => count1 + count2
34     }
35     //print
36     wordCounts.foreach {
37       case (word, count) => println(word + " " + count)
38     }
39   }
40 }
```



```
1 import org.apache.spark.SparkContext, SparkConf
2
3 object test {
4   def main(args: Array[String]) {
5     val sparkConf = new SparkConf().setMaster("local[*]").setAppName("MyWordCounts")
6     val sc = new SparkContext(sparkConf)
7     sc.textFile("hdfs://vagrant:spark@192.168.1.100:8020/words.txt").flatMap(_ => _.split(" ")).groupByKey().foreach {
8       case (word, count) => println(word + " " + count)
9     }
10  }
```

Scala

Java

## 国际大数据技术发展现状：俄罗斯

## ■ 俄罗斯——大数据存储技术的领导者

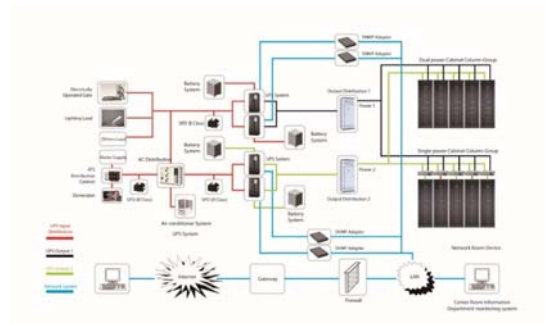
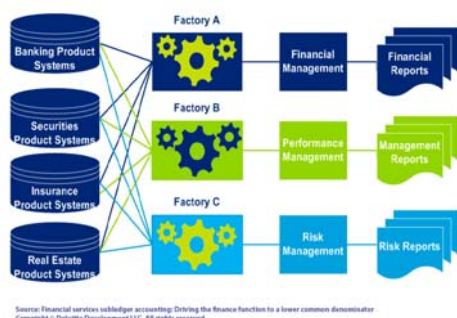
- ❑ 存储硬件方面
- ❑ 研究出将数据存储逾100万年的永久光盘
- ❑ 高密度、低磨损



## 国际大数据技术发展现状：俄罗斯

## ■ 俄罗斯——大数据存储技术的领导者

- ❑ 存储软件方面
- ❑ 安全高效的**银行数据交易系统**
- ❑ 在大数据存储开源界非常活跃





# 国际大数据技术发展现状

- 全球大数据百强公司数量对比
  - 美国遥遥领先，中日数量接近
  - 中美日占78%



## 1.3 我国的大数据发展战略



# 大数据：新时代的国家战略

- 2015年09月05日，国务院印发促进大数据发展行动纲要

“为贯彻落实党中央、国务院决策部署，全面推进我国大数据发展和应用，加快建设数据强国。”

- 2015年10月26日，十八届五中全会公报

将实施“国家大数据战略”写入党的全会决议，标志着大数据战略正式上升为国家战略。

- 2017年10月18日，党的十九大报告

“推动互联网、大数据、人工智能和实体经济深度融合”，建设“数字中国”

- 2017年12月8日，在中共中央政治局第二次集体学习时的讲话

“审时度势精心谋划超前布局力争主动实施国家大数据战略加快建设数字中国”



## 对习近平总书记在中共中央政治局第二次集体学习上重要讲话的理解

### 一个重要论断：

大数据是信息化发展的新阶段。

### 五项工作部署：

- 推动大数据技术产业创新发展。
- 构建以数据为关键要素的数字经济。
- 运用大数据提升国家治理现代化水平。
- 运用大数据促进保障和改善民生。
- 切实保障国家数据安全。

### 一个基本功要求：

- 领导干部善于获取、分析、运用数据。

# 部委与地方政府的大数据规划

## 科研投入

科技部、基金委前期通过973/863计划、国家科技重大专项等部署了大数据相关研究，“十三五”期间实施了“云计算与大数据”重点研发专项

## 基地建设

发改委组织建设11个大数据国家工程实验室

## 地方战略

17个省市出台大数据相关战略  
发改委、工信部、中央网信办联合批复贵州、上海、京津冀、珠三角等8个综合试验区，并正在加快建设

## 1.4 我国大数据技术与产业现状

# 我国大数据技术现状

## ■ 技术发展现状

- 一批世界一流的大数据技术公司
- 世界百强中，占据13席



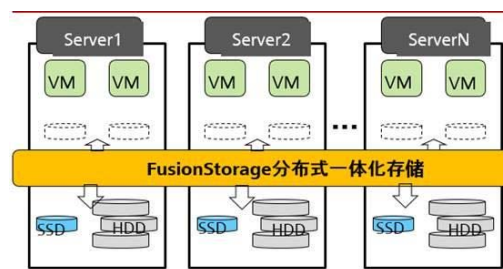
# 我国大数据技术现状

## ■ 技术发展现状 大数据计算和存储

- 自主知识产权的大数据技术，世界知名
- 代表公司：阿里巴巴，华为



阿里巴巴飞天分布式计算平台



华为FusionStorage存储平台



# 我国大数据技术现状

- 技术发展现状 大数据挖掘算法
  - 工业领域创造巨大价值
  - 代表公司：阿里巴巴，腾讯



阿里巴巴大数据运营平台



腾讯广点通数据分析平台

# 我国大数据技术现状

- 技术发展现状 大数据挖掘算法
  - 国防安全领域支撑融合决策
  - 代表单位：中科院计算所、中科天玑



中科院计算所“海上天网”融合平台



中科天玑舆情分析预警平台



# 中美大数据技术对比

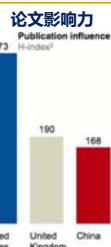
- 美国大数据技术创新水平遥遥领先，呈3个梯队

## 第一梯队：原始创新

Google

2000+学术论文

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



## 第三梯队：系统平台

IBM Microsoft  
ORACLE EMC<sup>2</sup>

软件平台的性能优化  
与传统数据系统整合  
开发特定领域的工具

## 第二梯队：工程实现

YAHOO!

147个项目@github

facebook

103个项目@github

twitter

102个项目@github

Apache Hadoop  
TensorFlow  
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)  
...

相比而言，我国大数据技术水平不高、技术扩散不畅

- 我国互联网企业快速将国际上先进的开源大数据技术整合到自身系统中，并构建了较大的系统，在国内保持领先；
- 总体上仍缺乏平台级的原创技术，对国际主流开源社区的贡献程度也不高，技术影响力较弱；
- 国内产业界在大数据技术路线发展中的话语权微弱；

阿里巴巴 Alibaba.com Bai 百度 Tencent 腾讯

# 中美大数据技术对比

## 中美大数据应用领域对比分析（投融资）



美国众多大型企业引领大数据应用，  
主要侧重于大数据技术  
(Google、IBM、Facebook、Oracle等)

- 我国侧重于大数据行业领域应用，具有广阔的市场空间
- 在数据处理分析、语音识别、视频识别、商业智能软件、数据中心建设和维护、IT咨询等领域都有代表性企业——形成获取、存储、处理、应用的大数据产业链

阿里巴巴 Alibaba.com Bai 百度 Tencent 腾讯

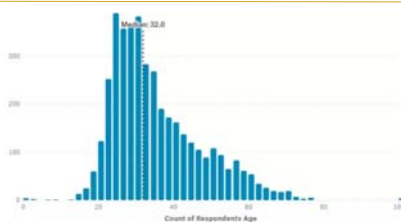
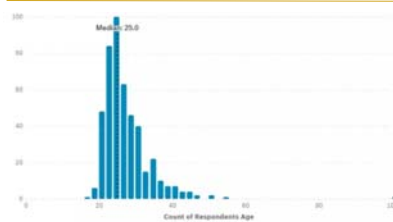


# 中美大数据技术对比

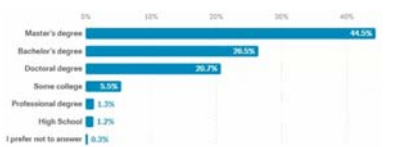
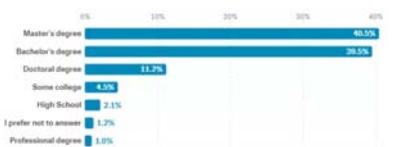
中国从业者

美国从业者

大数据应用领域人才对比



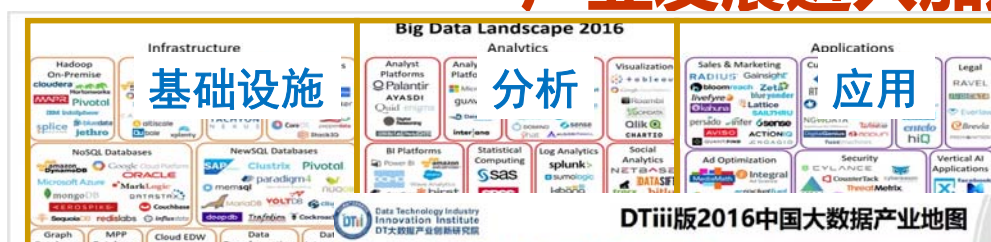
中国年龄中位数是25岁  
美国年龄中位数是32岁



美国博士学位高达20.7%，从占比上来看，接近中国的两倍（中国为11.2%）

- 人才短缺，经验不足：美国超过一半的数据科学家有10多年的工作经验，而我国经验不足五年的研究人员高达40%

## 大数据产业生态初步形成， 产业发展进入加速期



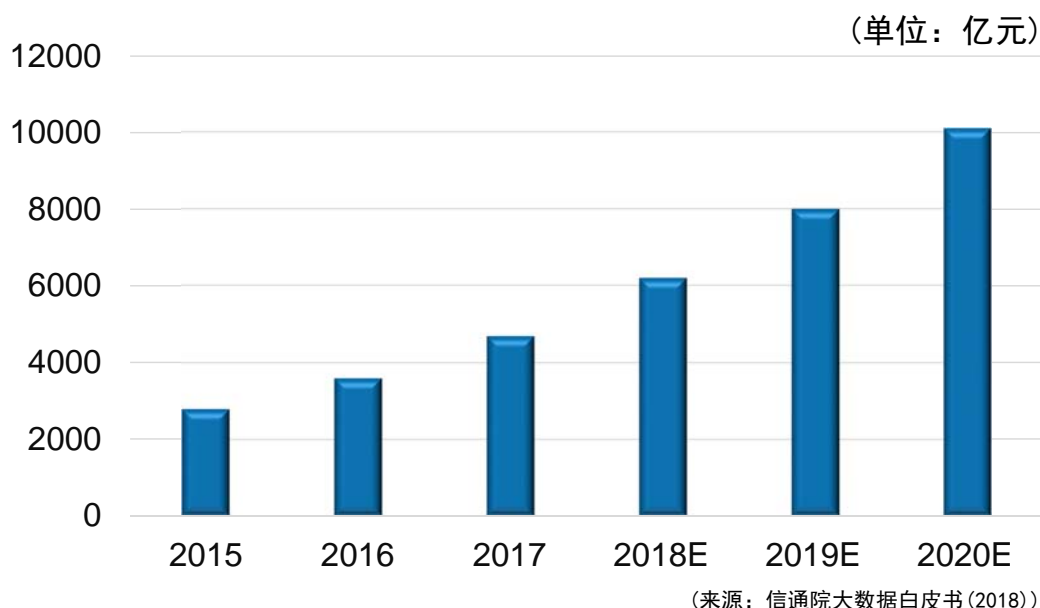
DTiii版2016中国大数据产业地图

- 中国企业在现有国际产业地图中极少出现，国际影响力不足
- 中国从事大数据应用的企业较多，掌握共性关键技术企业偏少



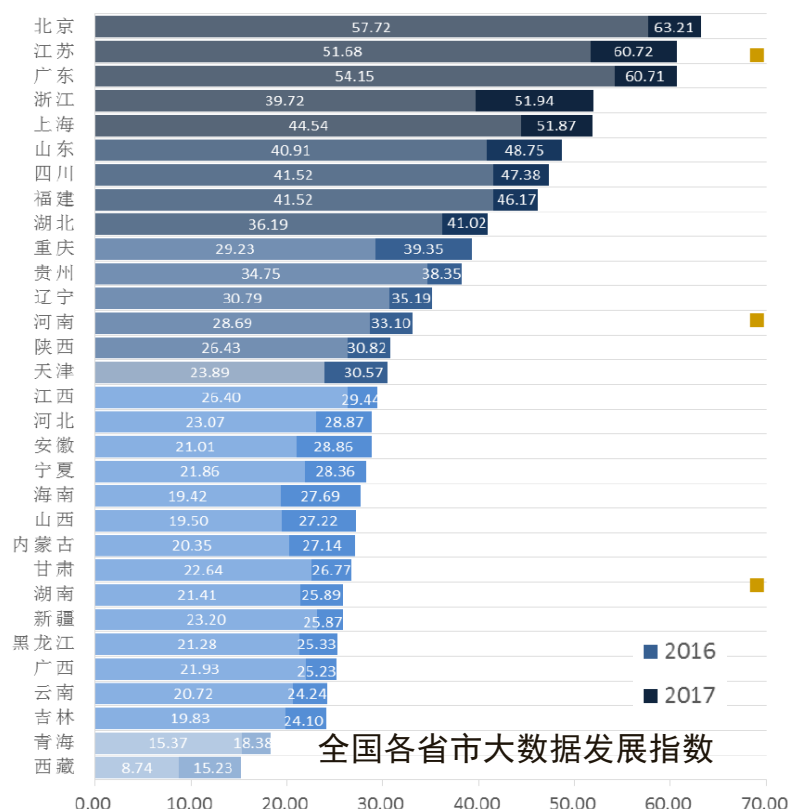
中国大数据产业地图

# 我国的大数据市场产值



- 2017年我国大数据产业规模为4700亿元人民币，同比增长30%
- 其中，大数据软硬件产品的值约为234亿元人民币，同比增长39%

## 全国大数据发展水平

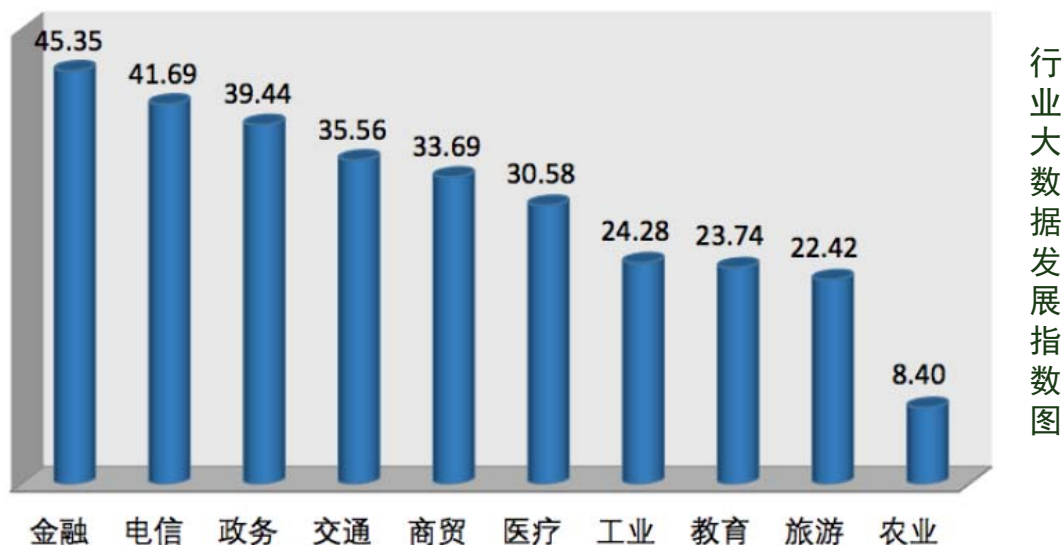


■ 各省市大数据产业发展水平差距依然较大，北京、江苏、广东、浙江、上海位列大数据产业发展第一梯队

■ 全国大数据发展已形成了以8个国家大数据综合试验区为引领，多区域集聚发展、第一梯队领先优势明显的格局

■ 东部地区的整体发展水平最高，增幅远超全国平均水平

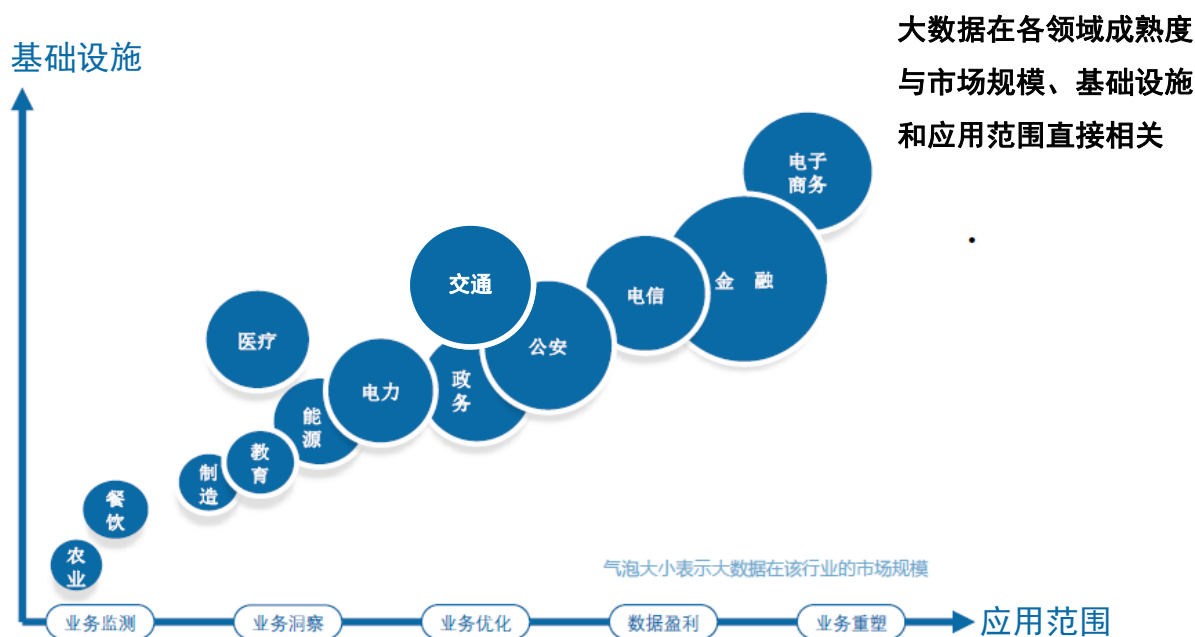
# 行业大数据发展水平



来源：中国大数据产业发展水平评估报告（2018年）

- 各行业大数据发展水平受**基础环境**、**数据汇聚**、**行业应用**等因素影响，**各行业大数据发展水平整体呈现差异化态势**，由高到低分别是金融、电信、政务、交通等
- 为金融、电信、政务三个行业提供大数据产品和解决方案的**企业**最多，分别占比63%、57%、47%

# 我国各行业的大数据成熟度



来源：Baidu & iFenxi

# 大数据行业应用的关键因素

## ■ 建立一体化的大数据平台

- 通过一体化大数据平台，数据的**汇聚和共享**得以实现，从而提升数据价值

## ■ 形成良好的数据管理体系

- 通过标准化的数据管控体系，**数据的质量与安全**得以保证

## ■ 形成平民化的数据应用

- **人人产生数据、人人使用数据**

## ■ 组建立强有力的数据管理部门

- 将数据的监管职责赋予**数据管理部门**，由数据管理部门集中管理监控数据

# 我国大数据行业应用的现状与问题

## ■ 利好：

- 近年来，在全球经济数字化浪潮的带动下，我国**大数据与实体经济**的融合应用不断拓展
- 利用大数据可以促进**实体经济行业**的**市场需求分析**、**生产流程优化**、**供应链与物流管理**、**能源管理**、**客户服务**等，这不但大大拓展了数据企业的**目标市场**，更成为众多大数据企业**技术进步**的重要推动力
- 随着融合深度增强和市场潜能不断被挖掘，融合发展给大数据企业带来的**益处和价值正在日显现**

## ■ 问题：

- 目前我国在大数据与实体经济融合领域**整体上还处于发展初期**
- 相对发达国家，在融合行业数量、融合应用深度、融合业务规模、发展均衡性等方面**还有一定差距**



# 我国大数据行业应用的特点(1)

## 三个不均衡：

### ■ 业务类型不均衡

- 大数据融合应用主要集中在**外围业务**上，而在**核心业务**方面的渗透程度还有待提高
- **营销分析、客户和内部运营管理**是应用最广泛的三个领域
  - 61.7%的企业将大数据用于营销分析
  - 50.2%的企业将大数据用于客户分析
  - 50.0%的企业将大数据用于内部运营管理
- 相比之下**大数据分析**在**产品设计、产品生产、企业供应链管理**等核心业务的应用比例还有待提升，大规模应用尚未展开

# 我国大数据行业应用的特点(2)

### ■ 地域发展不均衡

- 大数据融合应用在地区之间发展不均衡，各地大数据应用发展程度差距较大
- 受经济发达程度、人才聚集程度和技术发展水平影响，大数据应用的产学研力量仍主要分布在北京、上海、广东、浙江等东部发达地区，中西部地区的大数据应用虽然市场需求较大，但发展水平仍低

### ■ 行业分布不均衡

- 大数据与**金融、政务、电信**等行业的融合效果较好，而在其它众多行业的融合效果则有待深化

## 1.5 我国大数据发展面临的问题

### 我国大数据发展面临的问题

- 产业链条缺乏健壮性和完整性，未支持形成稳定的价值链
  - 目前，我国大数据产业价值链发展水平不高，价值链未完整建立，存在“重建设，轻应用，重硬件，轻软件”等产能过剩与产能不足并存的问题；
  - 大数据产业链包括数据采集、存储、交易、分析、应用各部分均有部署，但从产业链到产生价值链转化效果不好，大数据盈利模式模糊；
  - 大数据存储得到高度重视，使得大数据基础设施领域建设过热，扎堆建设大数据中心和大数据平台问题突出，导致基础设施资源闲置未充分利用；
  - 大数据交易发展火爆，但交易模式存在问题和隐患；
  - 相对大数据的采集、存储与管理，大数据的分析挖掘工具成熟度较低，大数据的深度分析挖掘在我国还处于比较初级的阶段，缺乏相应的软硬件工具与系统；
  - 大数据应用认识不到位、应用领域不广、应用程度不深、应用水平不高，服务对象不清晰；

# 我国大数据发展面临的问题

- 政府数据开放共享进展滞后，大数据资源应有的活力、红利并未得到充分释放
  - 我国数据开放程度在世界范围内排名比较靠后，与我国在世界经济发展中的表现不匹配。主要表现在：
    - 数据开放共享意识薄弱，在数据资源开放进程中进度缓慢，数据开放共享程度低；
    - 数据开放共享缺乏制度保障，顶层规划和实施细则未有效衔接，数据开放管理能力弱；
    - 数据开放共享缺乏技术支撑，数据资源流通不畅，数据质量不高；
  - 总体上，数据开放水平不高，导致数据价值难以被有效挖掘利用，大数据资源未能高效转变为大数据资产，大数据红利未充分释放。

# 我国大数据发展面临的问题

- 地方政府实践路径不清晰，需求认知模糊、推进不力
  - 从当前看，虽然各地方政府对于大数据发展的关注度非常高，但对大数据如何促进当地经济与社会发展的认知比较模糊，对于大数据在地方能够解决什么问题、促进什么发展、起到什么杠杆作用没有形成全面可行的方案；
  - 地方政府存在盲追概念的现象，未能突出地方特色、因地制宜制定差异化的大数据发展战略，未能结合大数据可以改变生产生活方式、促进经济发展、降低成本的本质主方向制定政策、推进落地，大数据对提升政府治理能力现代化水平不高；
  - 政务大数据发展缓慢，大数据产业的开发和建设难以找到合适的方向，对大数据应用不足存在的问题缺乏引导，对政府数据资源整合缺乏有力抓手，对大数据人才培养缺乏激励等；

# 我国大数据发展面临的问题

## ■ 核心技术尚未突破，应用仍处于低位水平

- 我国大数据核心技术发展仍处在跟跑国外先进技术阶段；
- 虽然大数据的应用技术在部分领域处于国际同等水平，但在核心器件、基础平台、基础软件等方面缺乏自主可控，在大数据原理性突破和颠覆性创新方面重视程度不够、投入不足；
- 核心技术未取得突破根源在于技术创新能力不足，存在关键技术壁垒：
  - 在新型计算平台、分布式计算架构、大数据处理、分析和呈现方面与国外仍存在较大差距；
  - 统计分析、计算复杂性、大规模异构数据融合、集群资源调度、分布式文件系统等大数据基础技术亟需突破；
  - 深度学习、类脑计算、认知计算、区块链、虚拟现实等前沿技术缺乏创新；
  - 面向大数据的新型计算、存储、传感、通信等芯片及融合架构发展未取得优势；

# 我国大数据发展面临的问题

## ■ 安全管理与隐私保护存在漏洞

- 大数据安全 and 数据隐私保护工作与大数据发展速度不匹配，安全隐患较大。数据安全是大数据发展中的突出问题，数据管理环节漏洞较多是大数据发展面临的首要问题，也是引发运营成本过高、资源利用率低、应用部署过于复杂和扩展性差等难点；
- 目前我国信息安全和数据安全的管理体系仍未健全，还未形成安全可控的大数据安全产品体系；
- 大数据应用的前提就是数据收集与挖掘等，但对立面是个人数据隐私保护，这两方面还未找到很好的平衡点；

# 我国大数据发展面临的问题

## ■ 制度建设尚不完善，市场活跃度堪忧

- 我国大数据制度建设不能满足发展建设需求，市场活力未充分激发；
- 中央和地方密集出台了许多规划、政策，但实施细则未及时出台，相关法律法规建设没有及时跟进，很大程度上限制了大数据产业链的构建与市场的发展；
- 数据所有权、隐私权等相关法律法规和信息安全、开放共享等标准规范不健全；
- 数据资产所有权与使用权还处在模糊地带，相关资产与交易行为未得到规范，缺乏较为完善的行业监管机制；
- 发展是引领，制度是支撑，在标准化方面，数据开放共享、交易、安全、系统级产品、管理以及评估类的标准较为缺乏，整体规划都需要完善；

# 我国大数据发展面临的问题

## ■ 人才供需不平衡

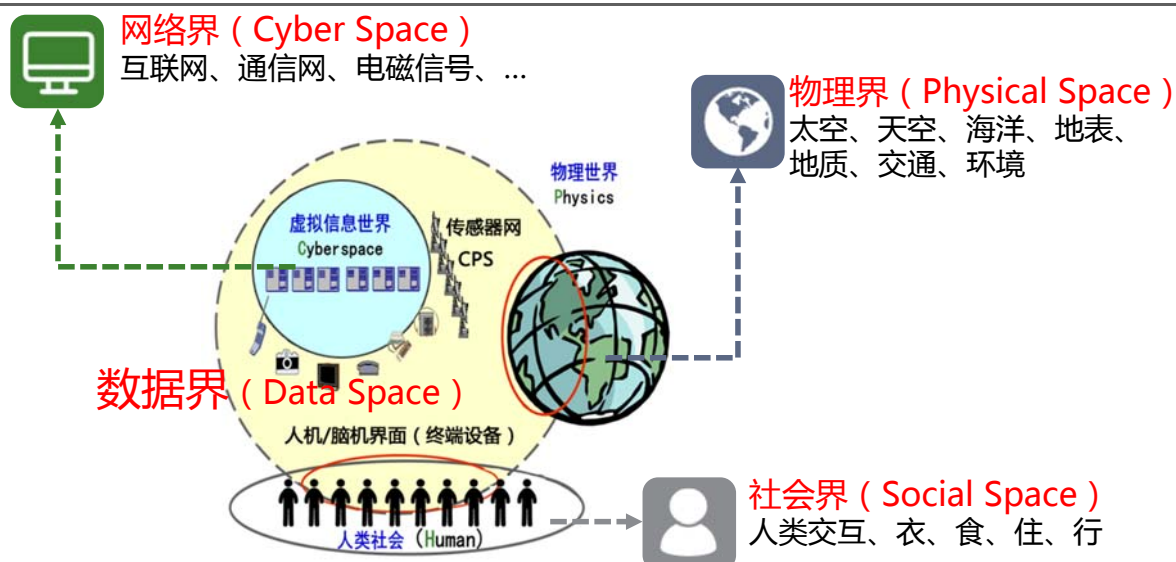
- 我国大数据人才供需不平衡，**关键技术人才缺乏**；
- 据相关测算，2016年我国大数据人才数量在46万人左右，而此后的3-5年内大数据人才需求将在150万人上下，这意味着中国大数据市场未来将**面临100万人左右的人才缺口**；
- 据国家信息中心统计显示，大数据人才供需存在三方面问题：
  - 一、**大数据人才学历层次错位明显**，主要表现为低学历（大专以下）的招聘需求高于求职数量占比，而对高学历（硕士以上）的需求则相反，大数据就业市场存在“高学低就”的现象；
  - 二、**地域供需不均衡**，受到社会大环境影响，北上广深等地区人才供给过多，但贵阳、合肥、天津等大数据市场活跃的地区人才供应不足；
  - 三、大数据领域数据分析、系统开发、数据管理、数据处理、数据采集等**技术类岗位“供不应求”**；



## 1.6 数据科学简介

### 数据界：大数据是关键的战略资源

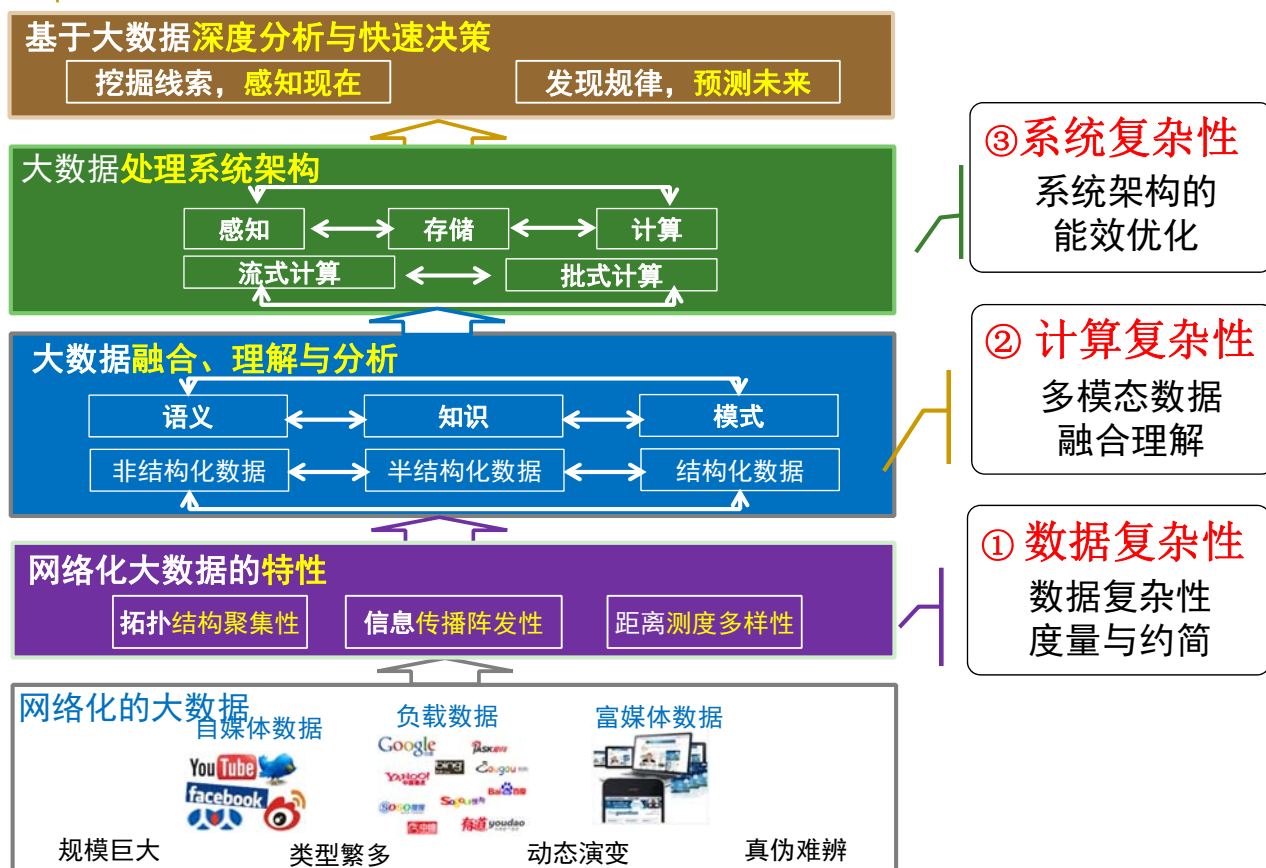
- 数据是三元世界融合的新物质 (CPSS的核心)
- 数据是数字经济的关键赋能资源与基础物质供给
- 离开数据支撑、智能将成无米之炊



# 什么是数据科学？

- **数据科学(Data Science)**：是研究大数据的感知、收集、传输、管理、分析与应用的交叉性学科，旨在揭示数据的内在规律，探索数据计算理论，实现**从数据到知识的转化**，为大数据的科学计算以及在重要应用领域的预测、决策与应用提供基础。
- 数据科学的**两大内涵**：
  - **数据内在规律**：大数据在多元数据空间中存在的特征、关联与演变规律等
  - **数据计算理论**：包括大数据计算的基础理论与方法，面向大数据的计算模式与体系架构等

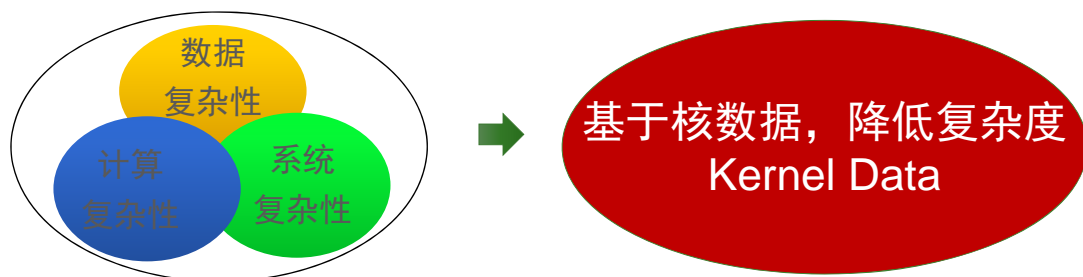
## 大数据处理面临的科学问题与挑战



# 解决数据复杂性的总体研究思路

大数据悖论：大数据小价值 vs. 小数据大价值

可获得（共享+交易）、可计算（数据+模型）、好计算（算力+加速）

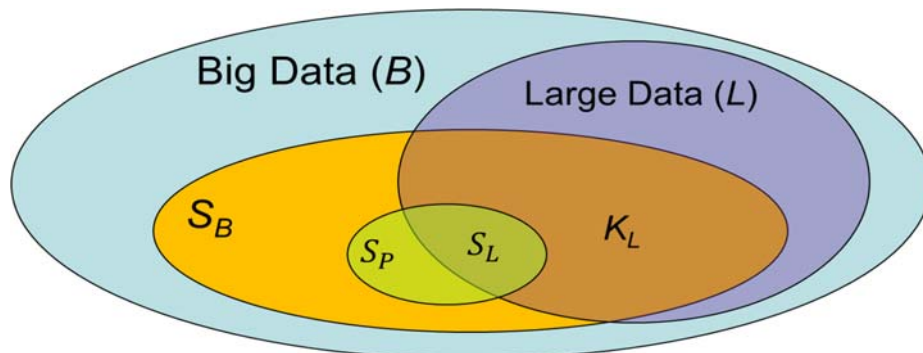


基本假设：通过领域知识和流程学习，**寻找解决问题的核数据**，实现对大数据分析应用复杂性的约简

## 核数据概念

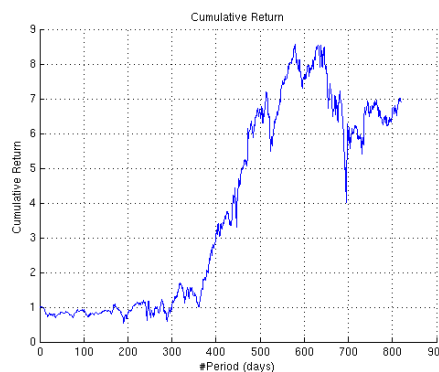
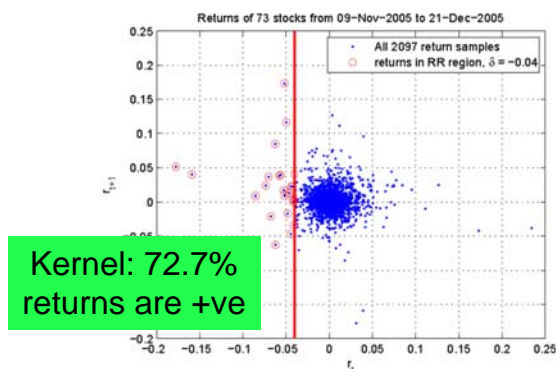
### ■ 从“大数据”到“核数据”

- 大数据(B)是无法获取的，我们拥有的是大数据的一个较大子集(L)
- 问题：从大数据中找到应用问题P的部分解，非全部解
- 难点挑战：问题解 $S_P$ 在数据空间中的分布是不均匀的
- **核数据**：非均匀分布下、解密度高的数据子集( $K_L$ )
  - **核数据是降低复杂性的关键**



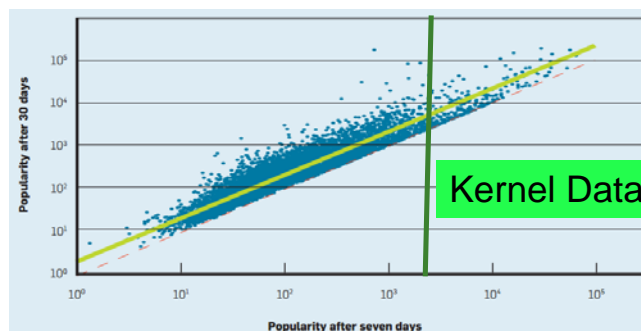
## 核数据举例 (1) : 金融交易预测中的核数据

- 应用问题P: 预测第二天具有正收益的股票
- 大数据B: 历史上至今所有的股票交易数据
- 可处理的大数据L: 当前所有的股票交易数据
- 领域知识 $\mathcal{K}$ : 均值回归定理 (Mean Reversion Property)
  - If a stock performs worse than others, it tends to perform better than others in the next trading day (1991)
- 核数据K: 当天交易结束时收盘价格低于MR的所有股票数据, 计算复杂度从指数级降为多项式级



## 核数据举例 (2) : 传播预测中的核数据

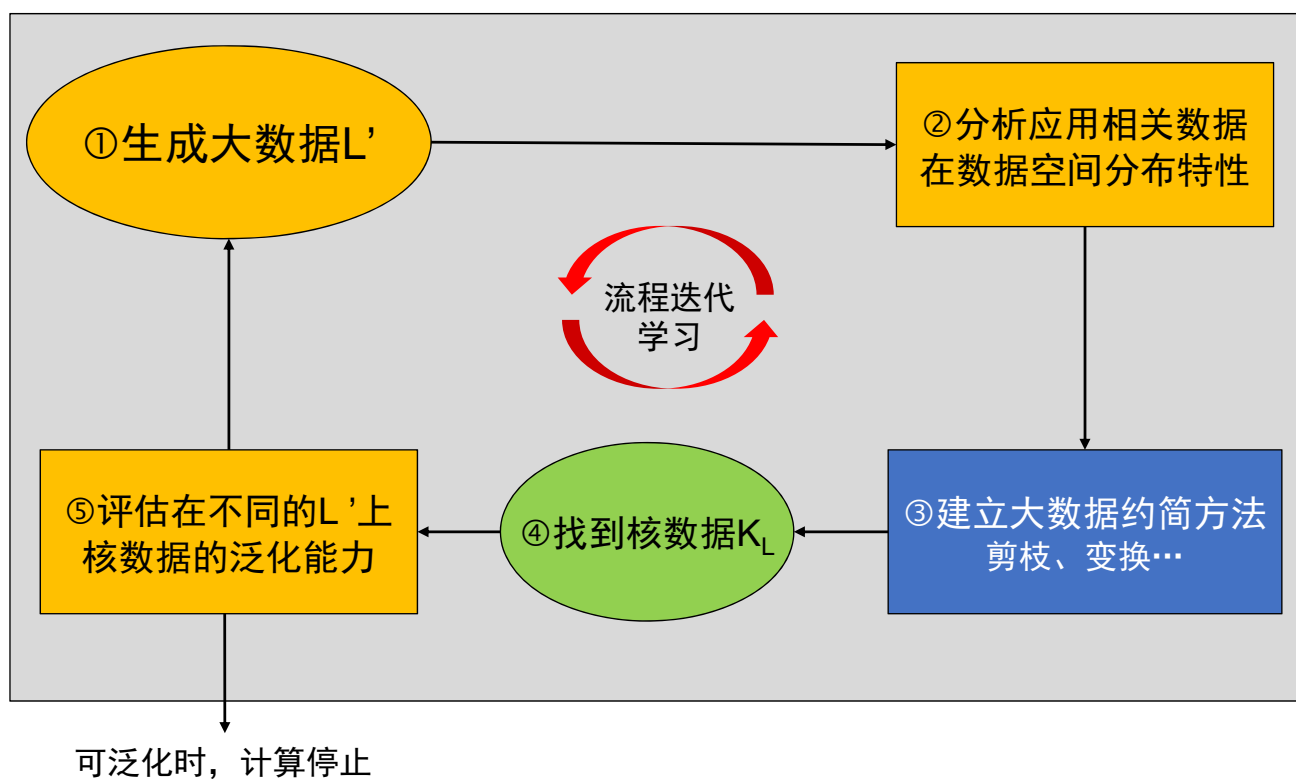
- 应用问题P: 社交媒体热点信息预测
- 大数据B: 历史上所有的社交媒体消息
- 可处理的大数据L: 当天发布的消息
- 领域知识 $\mathcal{K}$ : 消息流行度的对数时序相关性
  - Early patterns of access indicate long-term popularity of message. (2010)
- 核数据K: 发出一段时间后流行度超过某个阈值的消息, 将数据复杂度下降2-3个数量级



流行度预测:

$$\ln N(t_r) = \ln N(t_0) + \sum_{\tau=t_0}^{t_r} \eta(\tau)$$

# 处理大数据复杂性的基本框架



谢谢聆听！

