清華大學
Tsinghua University

大数据与程序设计

陈文光

第5届全国高等学校计算机程序设计课程研讨会
厦门大学、清华大学出版社联合承办
2015年12月4日-6日 厦门大学

提纲

- 什么是大数据
- 大数据带来的思想变革
- 大数据带来的商业变革
- 大数据系统与大数据程序设计

什么是大数据？

- 大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合（维基百科定义）
- 大数据 = “海量数据”+“复杂类型的数据”
- 大数据的特性（ Volume，Variety，Velocity）
  - 数据量大：PB、TB、EB、ZB级别的数据量
  - 种类多：包括文档、视频、图片、音频、数据库、层次状数据等
  - 速度快：数据生产速度很快；要求对数据处理和I/O速度很快

什么是大数据？

- 当数据的规模和性能要求成为数据管理分析系统的重要设计和决定因素时，这样的数据就被称为大数据
  - 不是简单地以数据规模来界定大数据，要考虑数据查询与分析的复杂程度
- 以目前计算机硬件的发展水平看
  - 针对简单查询（如关键字搜索），数据量为TB至PB级时可称为大数据
  - 针对复杂查询（如数据挖掘），数据量为GB至TB级时即可称为大数据

大数据涉及诸多不同的领域


天文


气象


基因

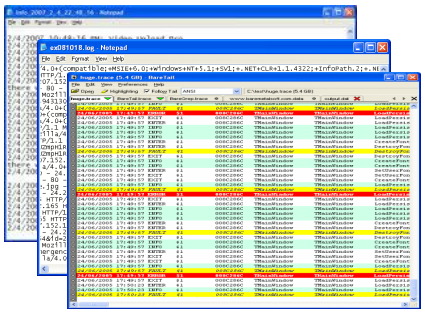

医学

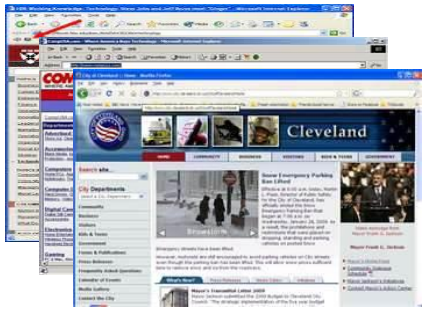

经济


物理

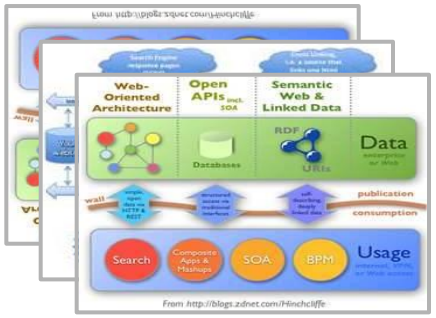

其他领域


用户生成数据


Deep Web数据

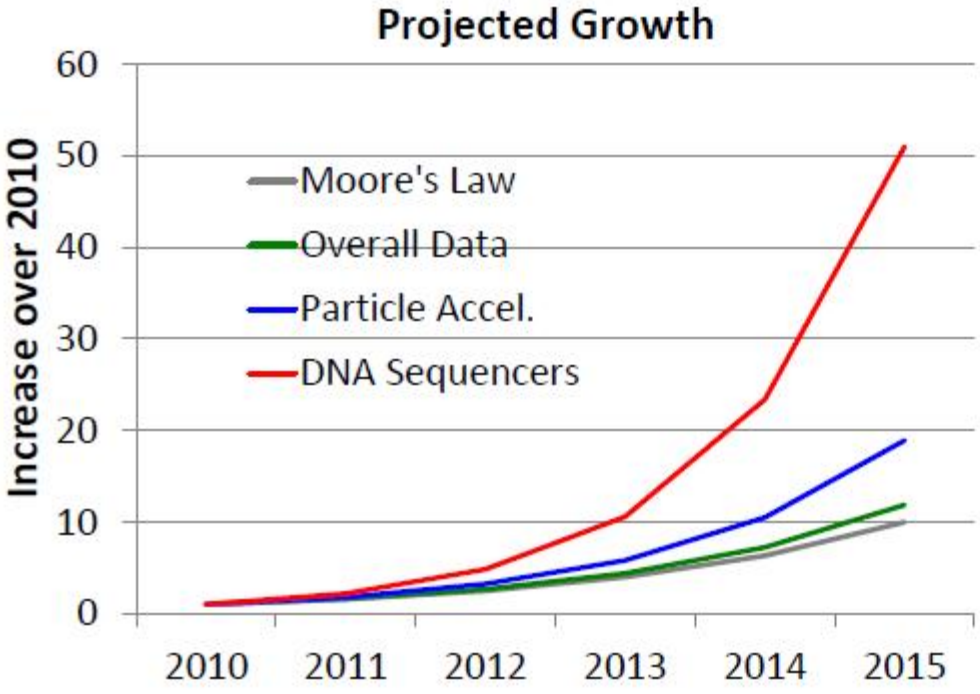


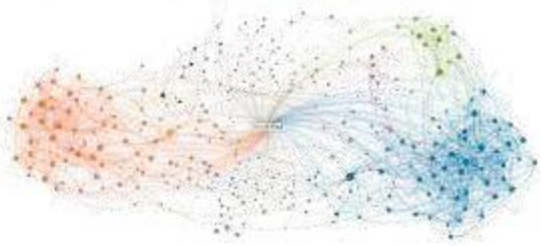
多模态内容数据



网络与关系数据

大数据总量增长态势



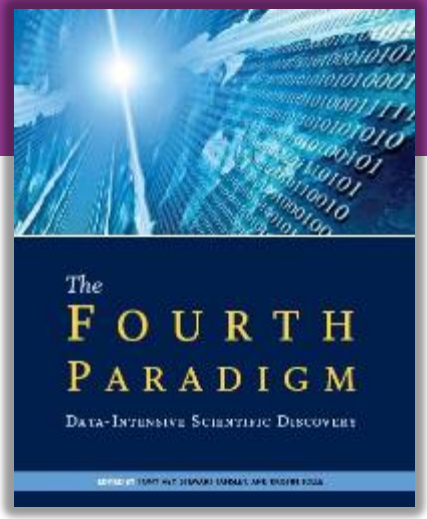


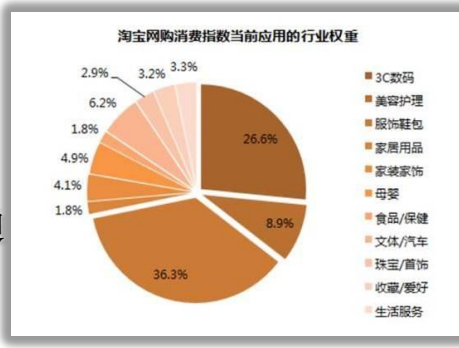
Data Grows faster than Moore's Law

[IDC report, Kathy Yelick, LBNL]

大数据的价值

- 科研价值
  - 1998 年图灵奖得主、数据库技术奠基人Jim Gray认为数据驱动的研究将是第四种科学研究范式
    - ”The Fourth Paradigm: Data-Intensive Scientific Discovery”
  - 大数据已为多个不同学科的研究工作提供了宝贵机遇
- 经济价值
  - 麦肯锡全球研究院：大数据可为世界经济创造巨大价值，提高企业和公共部门的生产率和竞争力，并为消费者创造巨大的经济利益
  - 著名Gartner公司：到2015年，采用大数据和海量信息管理的公司将在各项财务指标上，超过未做准备的竞争对手20%
- 工业价值
  - 分析使用：揭示隐藏其中的信息，对生产流程进行分析和改造，产品的故障检测和诊断
- 社会价值
  - 例如：2009年淘宝网推出淘宝CPI来反映网络购物的消费趋势和价格动态





社交网络数据-Volume

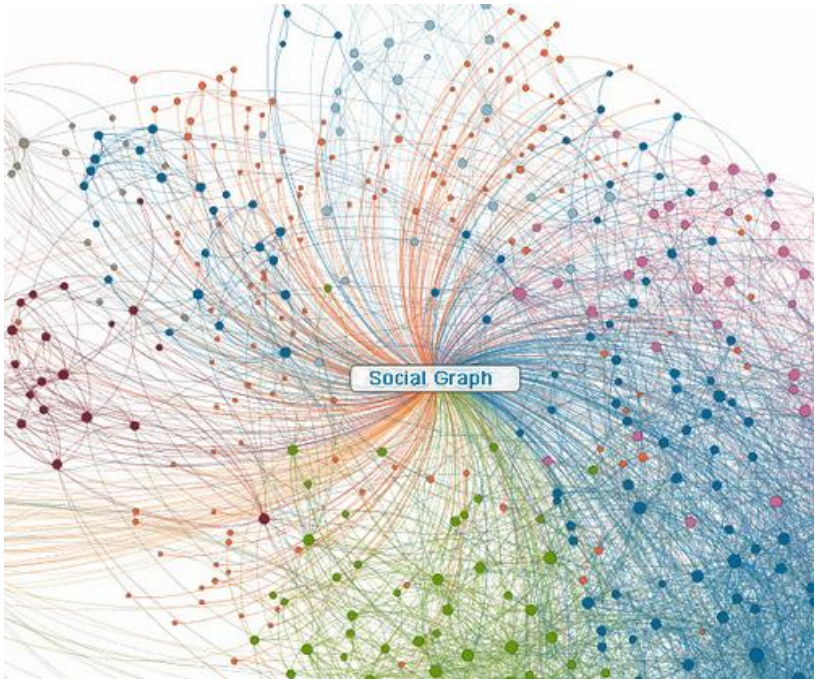
- 新浪1000万人每人（最多）1000条微博 – 5TB
  - 3亿用户 ~ 100TB
  - 还没包括评论和图片
- 用户Profile
  - 100GB量级
- 用户关系
  - 数亿用户，几百亿条边，100GB量级
  - 数十亿用户，几个TB量级

社交网络数据-Velocity

- 每天新浪发出上亿条微博
  - 512 * 10^8 Byte ~= 50GB
- 关注关系的演化
  - 结点的增加
    - 按半年增加8000万用户估算，每天平均新增40万
  - 关注关系的增加与取消

社交网络数据-Variety

- 微博 – 自然语言
- Profile / Tags
- 用户关注关系 – 图
  - 非结构化数据
- 微博的转发与评论关系-图



提纲

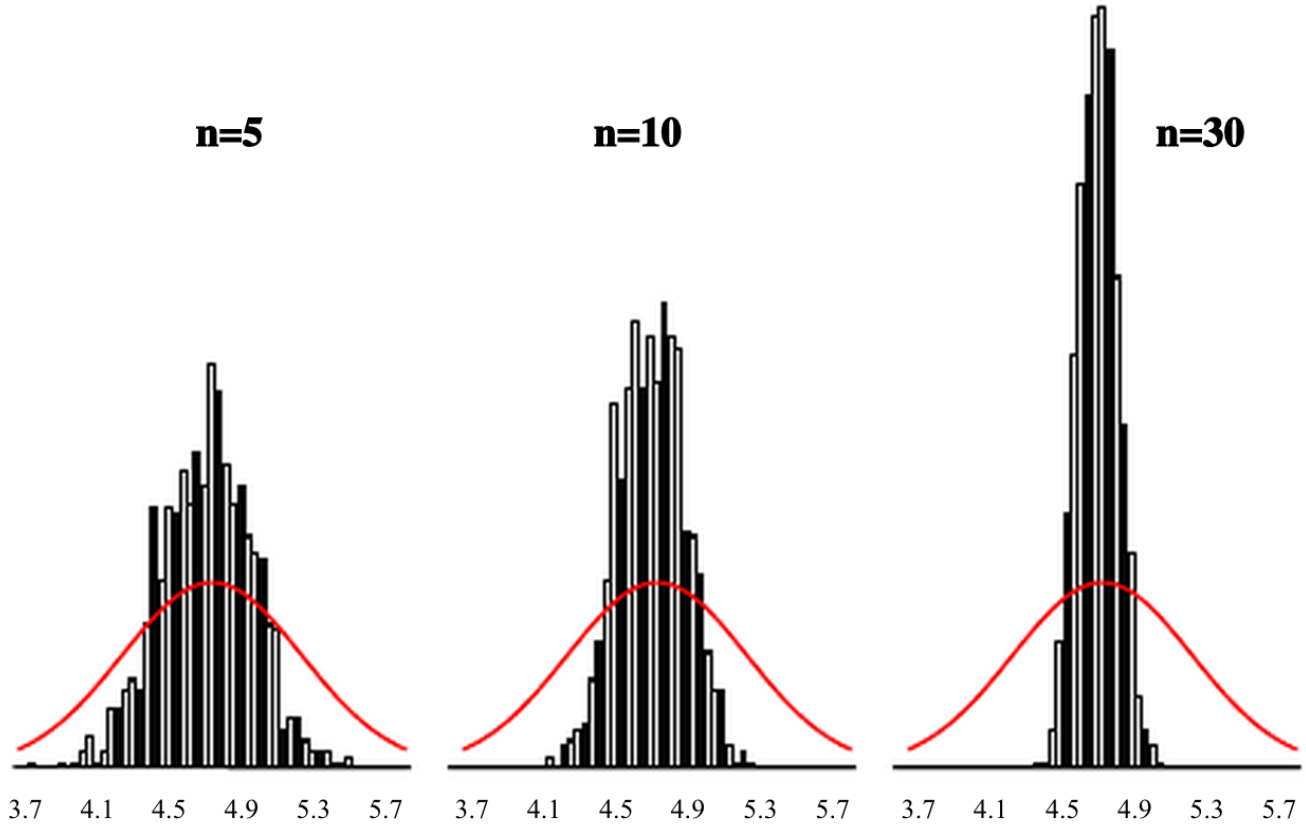
- 什么是大数据
- 大数据带来的思想变革
- 大数据带来的商业变革
- 大数据系统与大数据程序设计

大数据带来的思想变革

- 不是随机采样，而是所有数据
- 不是精确性，而是混杂性
- 不是因果性，而是相关性

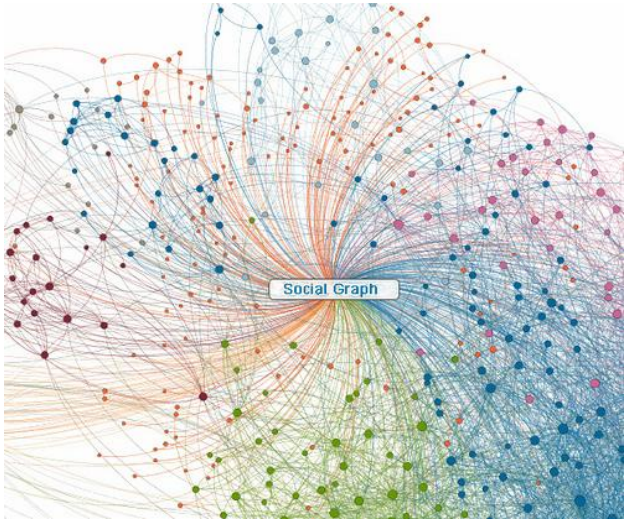
不是随机采样，而是所有数据

- 随机采样很有效，**样本数增加则估计更加准确**
- **样本的随机性比样本数量更重要**



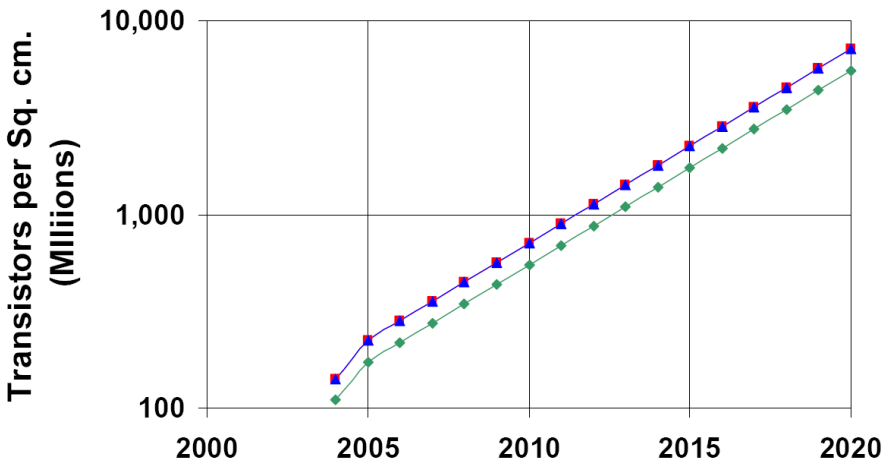
随机采样的问题

- 不准确的采样方法
  - 例如，使用随机选取固定电话号码进行采样？
    - 没有考虑到那些只使用手机的人
  - 美国2008年总统选举的抽样民调发现不考虑仅使用手机的人，民调结果误差会扩大2个百分点
- 采样数据不适合进行深入分析
  - 某种类型的人可能只有很少几个
- 缺乏随机采样方法的数据
  - 例如对社交网络如何进行随机采样？
    - 能保持图的三角形密度属性？
    - 两个结点的平均最短路径？



所有数据

- 信息技术的发展为记录所有数据提供了机遇
  - 传感器、摄像头、社交网络、电子化交易将各种行为数字化
  - 存储技术的发展，能够轻易存储TB、PB乃至ZB级的数据，价格随摩尔定律下降
  - 芯片处理能力随摩尔定律上升，分布式系统的成熟
  - 网络速度近似以摩尔定律上升



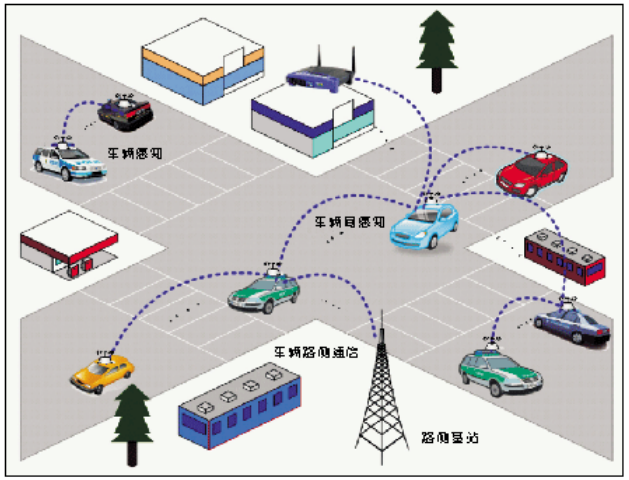


图1 车辆传感器网络通信模式示意图

不是随机采样，而是全部数据

- 真的是“全部”数据？
  - 微博上的民意是真正的民意吗？
    - 并非所有人都上微博
    - “沉默的大多数”意见如何？
- 某些分析仍需采样
  - 计算量很大，例如最短路径算法的复杂度是O(V*E)
  - 要很快得出结果，不能对所有数据进行计算

不是精确性，而是混杂性

- 数据来源增加，以智能交通为例
  - 种类增加
    - 视频监控，线圈，GPS，收费口
  - 测量频率增加
    - 每次通信记录
  - 测量点增加
    - 部署更多的摄像头，出租车、公交车带GPS
- 造成数据量和数据种类大大增加，但精确度下降
  - 损坏的传感器
  - 测量、传输过程中丢失
  - 多个来源数据的不一致

数据量和算法

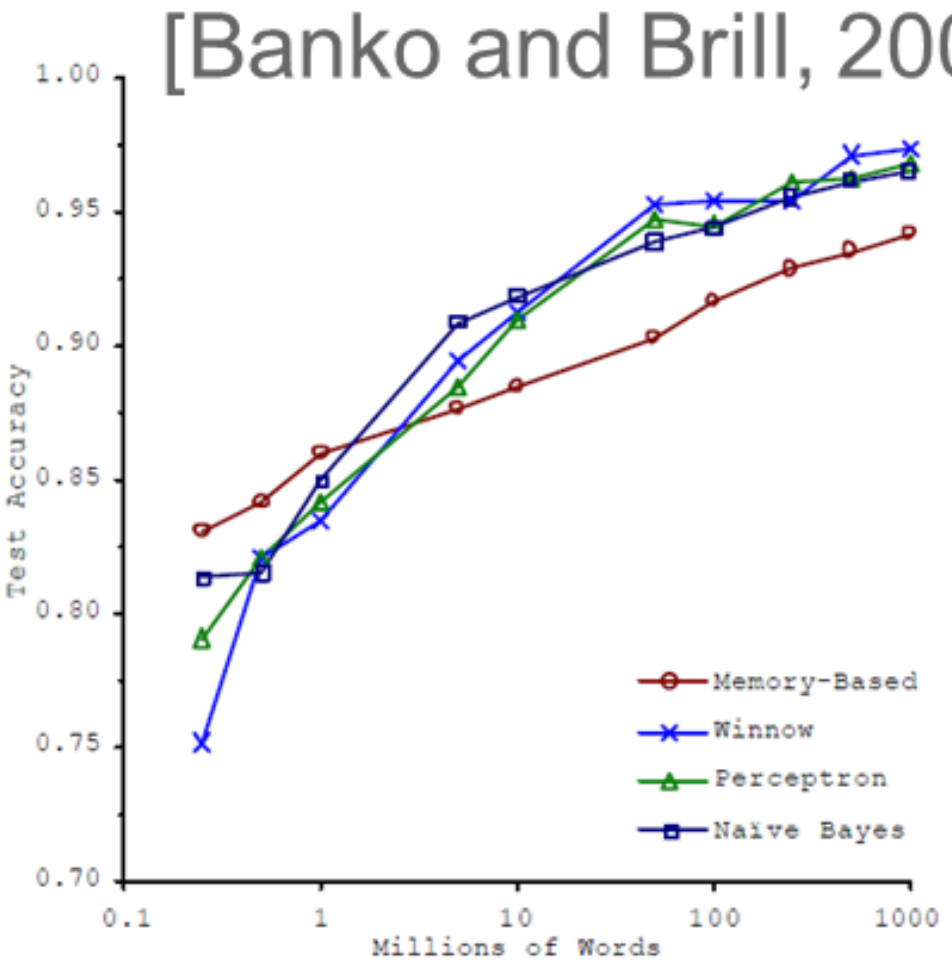
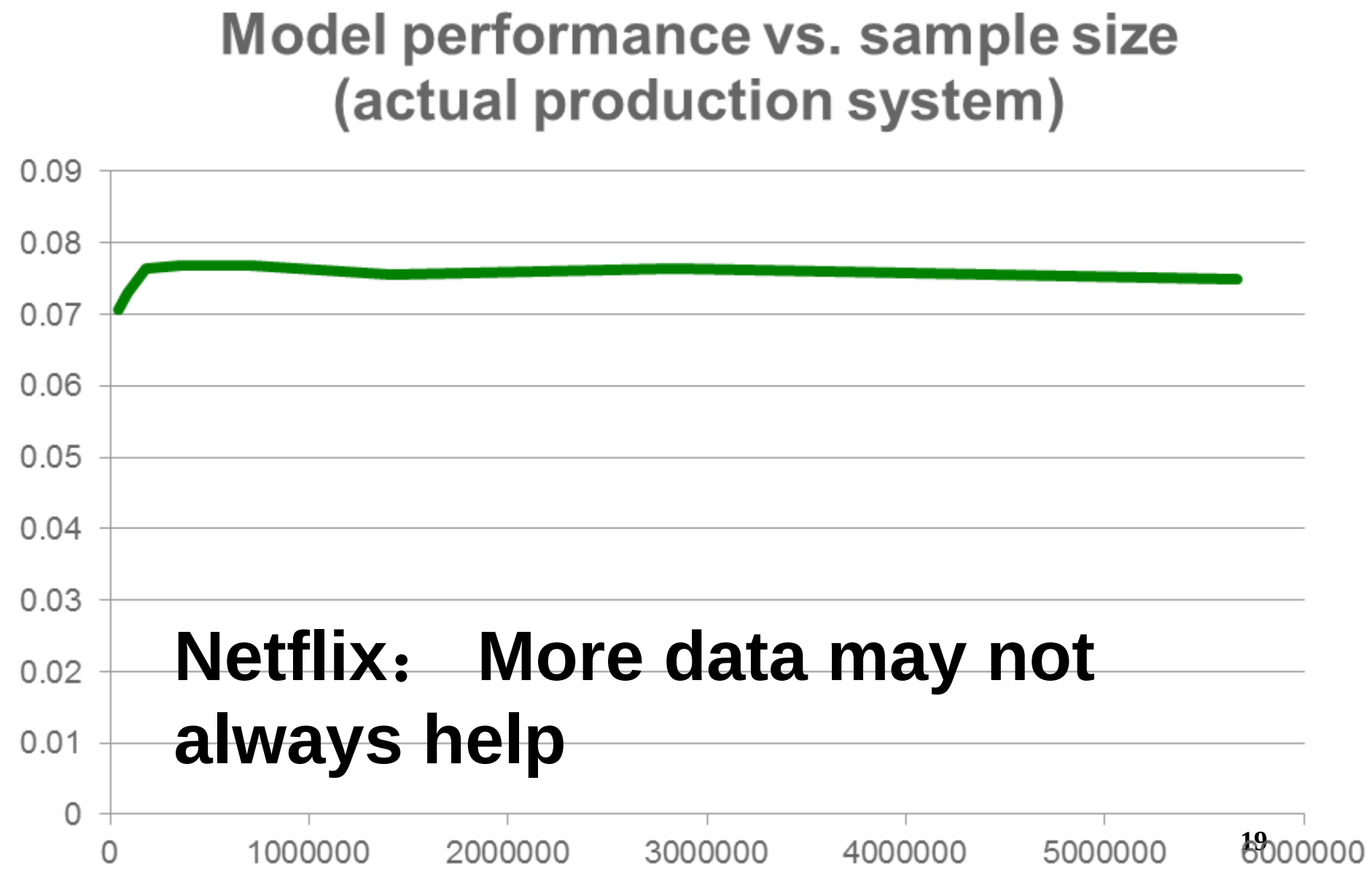


Figure 1. Learning Curves for Confusion Set Disambiguation

“We don’t have better algorithms. We just have more data.”
– Peter Norvig, Google

数据量和算法



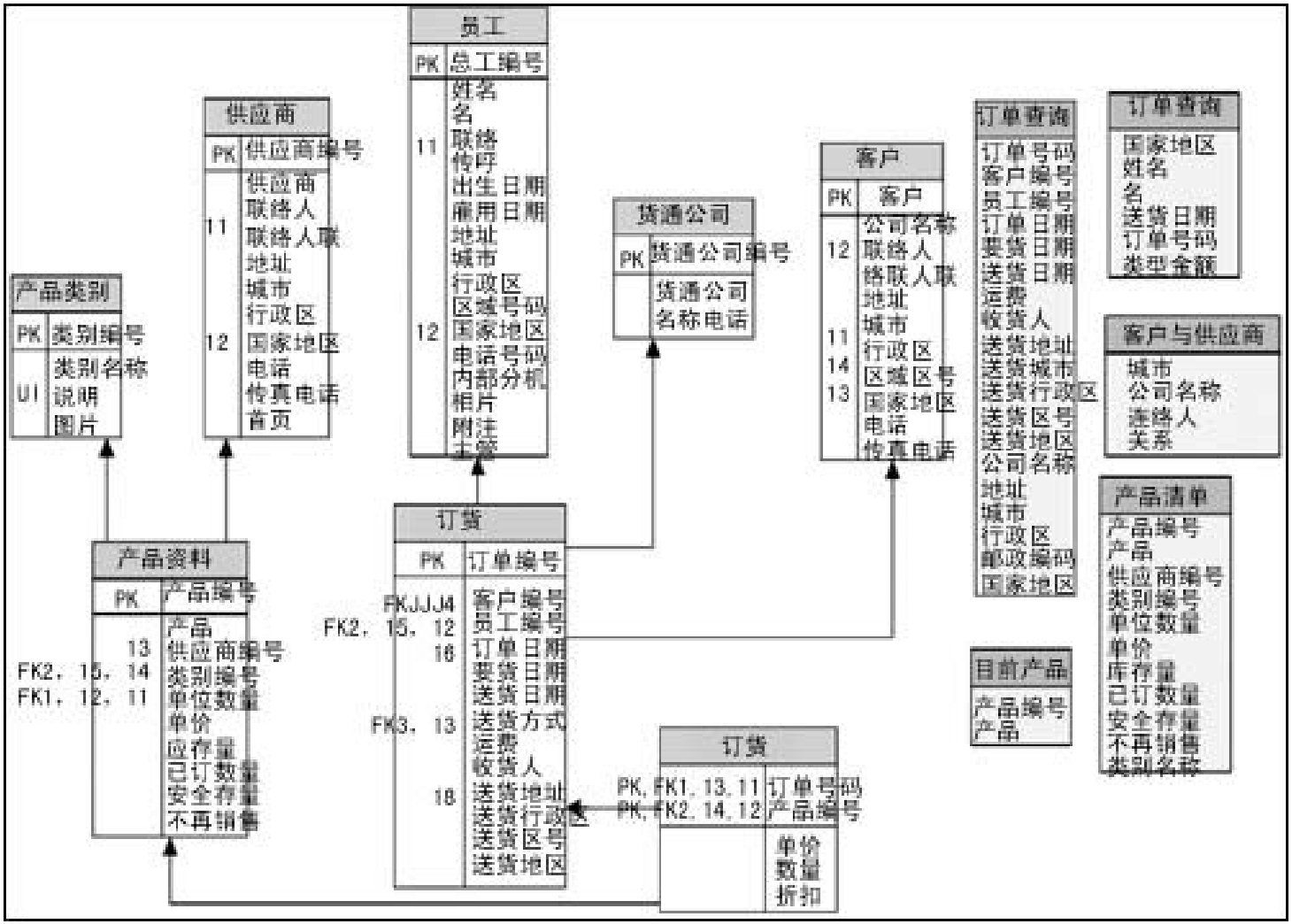
Netflix： More data may not always help

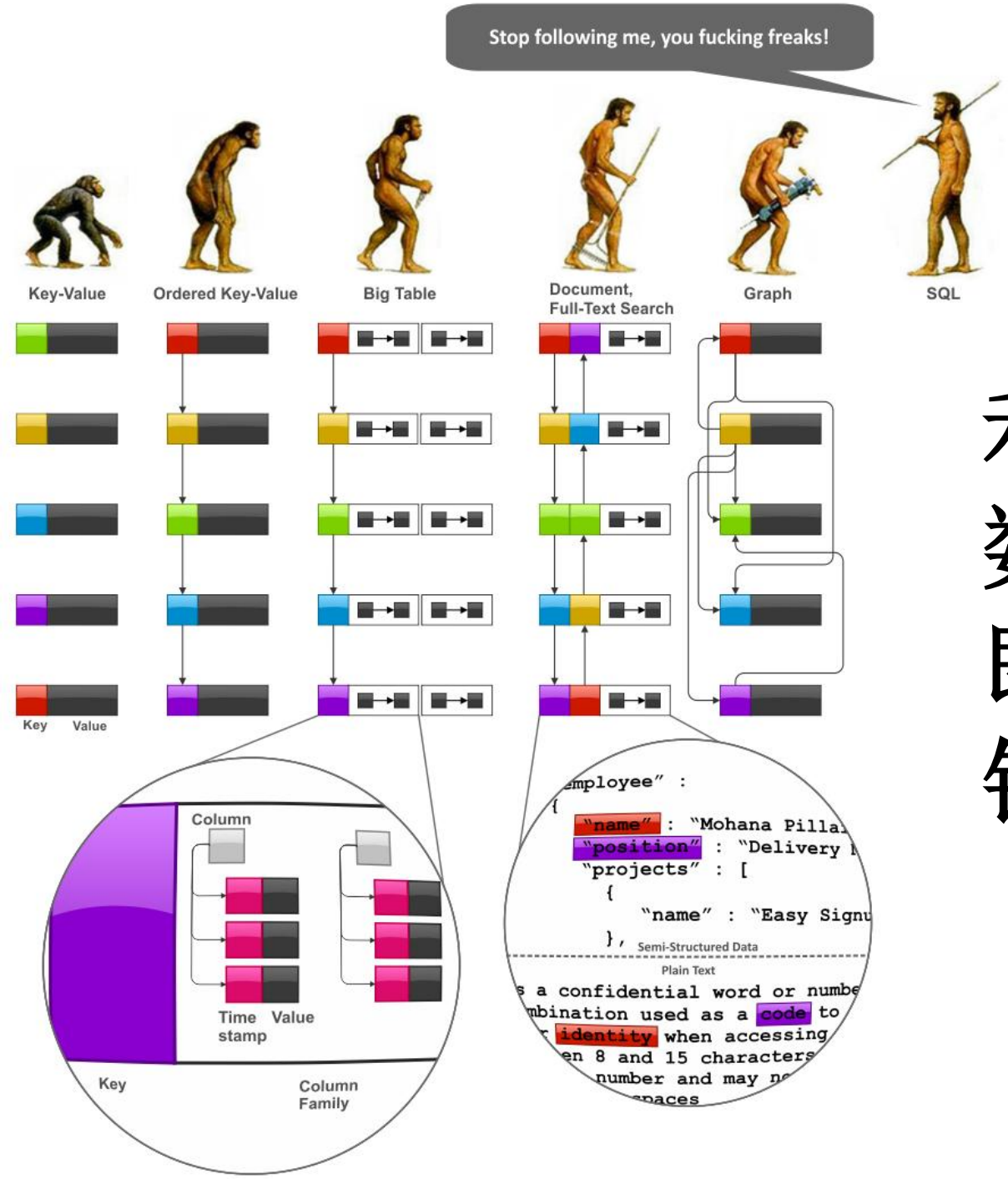
数据量和算法

- 大数据量能否提高性能仍然同模型（算法）有关
  - 模型足够复杂而数据量不足时，增加数据量能提高性能
  - 模型过于简单而不足以描述实际情况时，增加数据量并不能提高性能

研究和使用能够利用大数据量的模型和算法

从结构化数据到非结构化数据

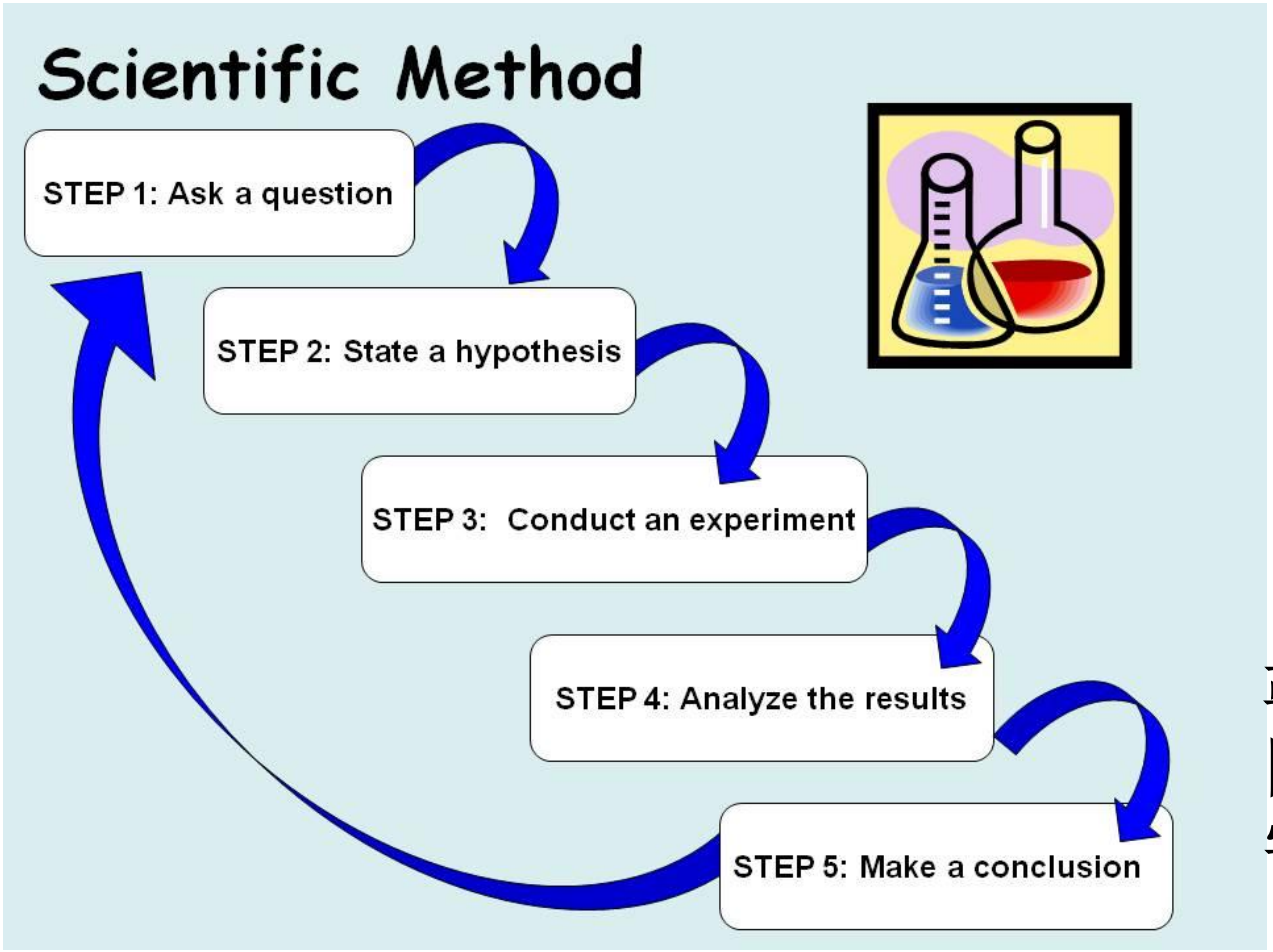


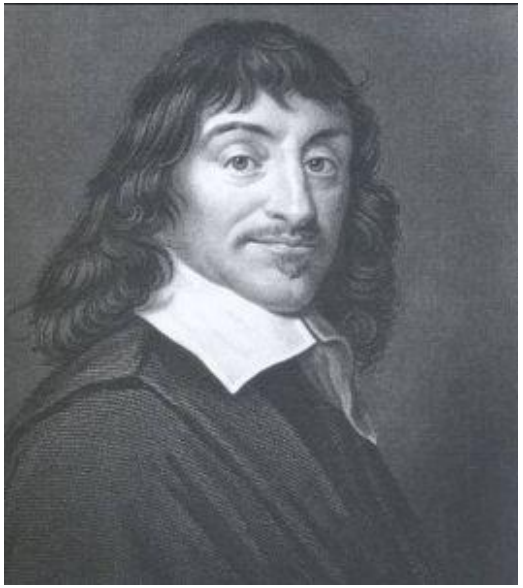


利用非结构化数据
即使其中存在错误和缺失

不是因果性，而是相关性

- 因果性和科学方法





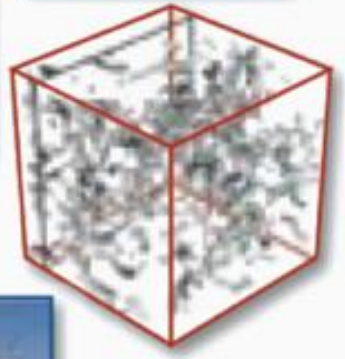
勒内·笛卡尔，著名法国哲学家、科学家和数学家，**1596-1650**

科学研究的三个范式和第四范式

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$



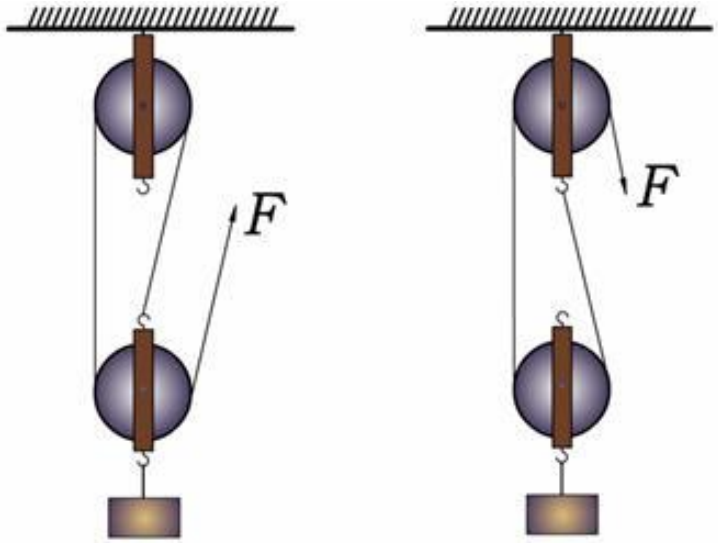
因果性和相关性

- 因果性
  - 在小规模实践中发现F与物体重量成正比，与承担滑轮重量的绳子数成反比
  - 改变绳子的段数和物体重量，进行实验验证
- 相关性
  - 有大量拉起物体的数据记录

<物体重量，绳子段数，提起的重量>

从中总结出：F = W / N

  - 有时是概率的相关性
    - 橙色二手车有质量问题的可能性是其它颜色车的一半



相关性为什么重要

- 有时相关性就够用了
  - 有某种基因的人患某种癌症的可能性较大
    - 这种基因是如何调控蛋白质，从而提高癌症概率的目前并不清楚
  - 蛋挞和飓风用品放在一起
- 因果性并不绝对
  - 因果性和我们认为正确的科学理论仅仅是我们还未能通过实验证伪它们，并非绝对真理
  - 依据这些因果性进行规划比依据其它方法更可能成功

相关性的局限性

- 数据的有限性
  - 用来分析的数据并非全部数据
    - 微博人群不是全部人群
- 相关性成立的条件
  - 用夏天的数据预测冬天的行为？
  - 美国的发展规律预测中国的？
- 只能发现已出现的规律

关于相关性的一个著名例子

财经网V：【荐读·网络如何救赎百年《纽约时报》】纽约时报发表了“公司如何知道你的秘密”的文章，说美国大型超市如何分析用户消费数据，但仅获60个“赞”。福布斯一名网编改了个标题，以“Target如何比未成年少女的父亲更早知道她女儿怀孕了”为题，获得了超250万次浏览，60万个“赞”。http://t.cn/RvbOKv1

6月23日16:30 来自微博 weibo.com (59) | 转发(186) | 收藏 | 评论(30)

回顾：大数据带来的思想变革

- 不是随机采样，而是所有数据
  - 随机很好，但很难
- 不是精确性，而是混杂性
  - 数据越大越好？不总是
- 不是因果性，而是相关性
  - 相关性有用途，也有局限

提纲

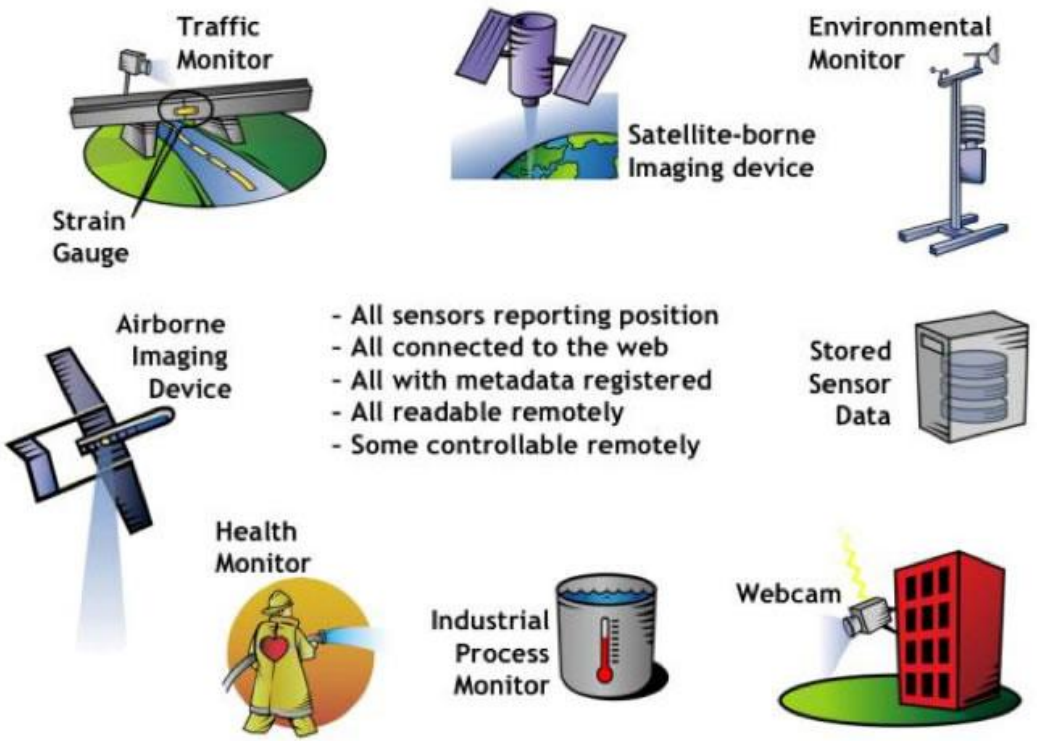
- 什么是大数据
- 大数据带来的思想变革
- 大数据带来的商业变革
- 大数据系统与大数据程序设计

大数据带来的商业变革

- 无处不在的数字化
  - 收集各种数据
- 数据利用方法
  - 数据有潜在价值
- 大数据价值链
  - 数据、技术与思维

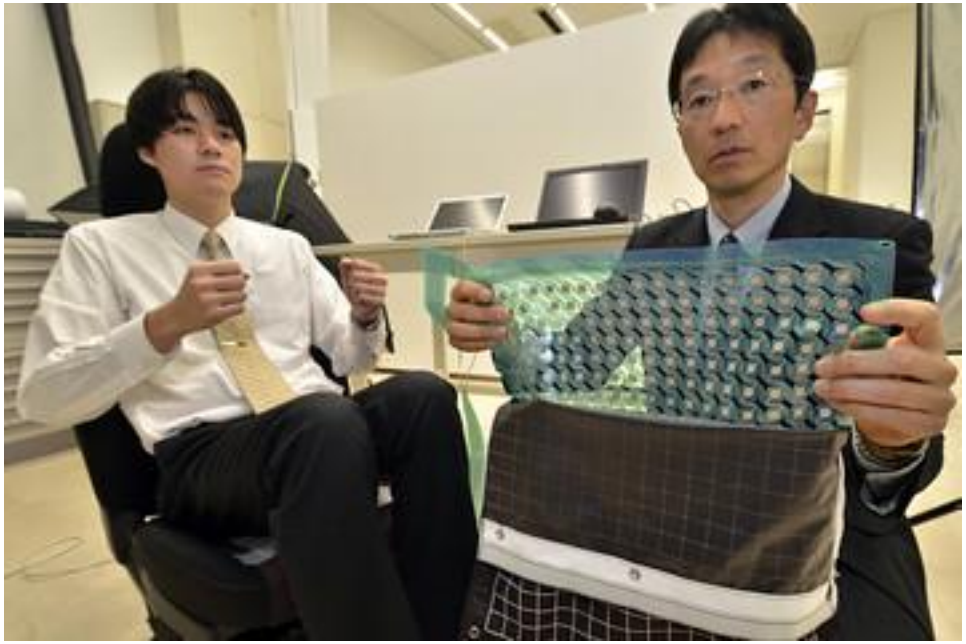
量化一切

- 传感器
  - 温度、湿度、压力、速度、加速度成为数据
- 数字图书馆
  - 文字、知识成为数据
- GPS
  - 位置成为数据
- 社交网络，电信网络
  - 沟通成为数据

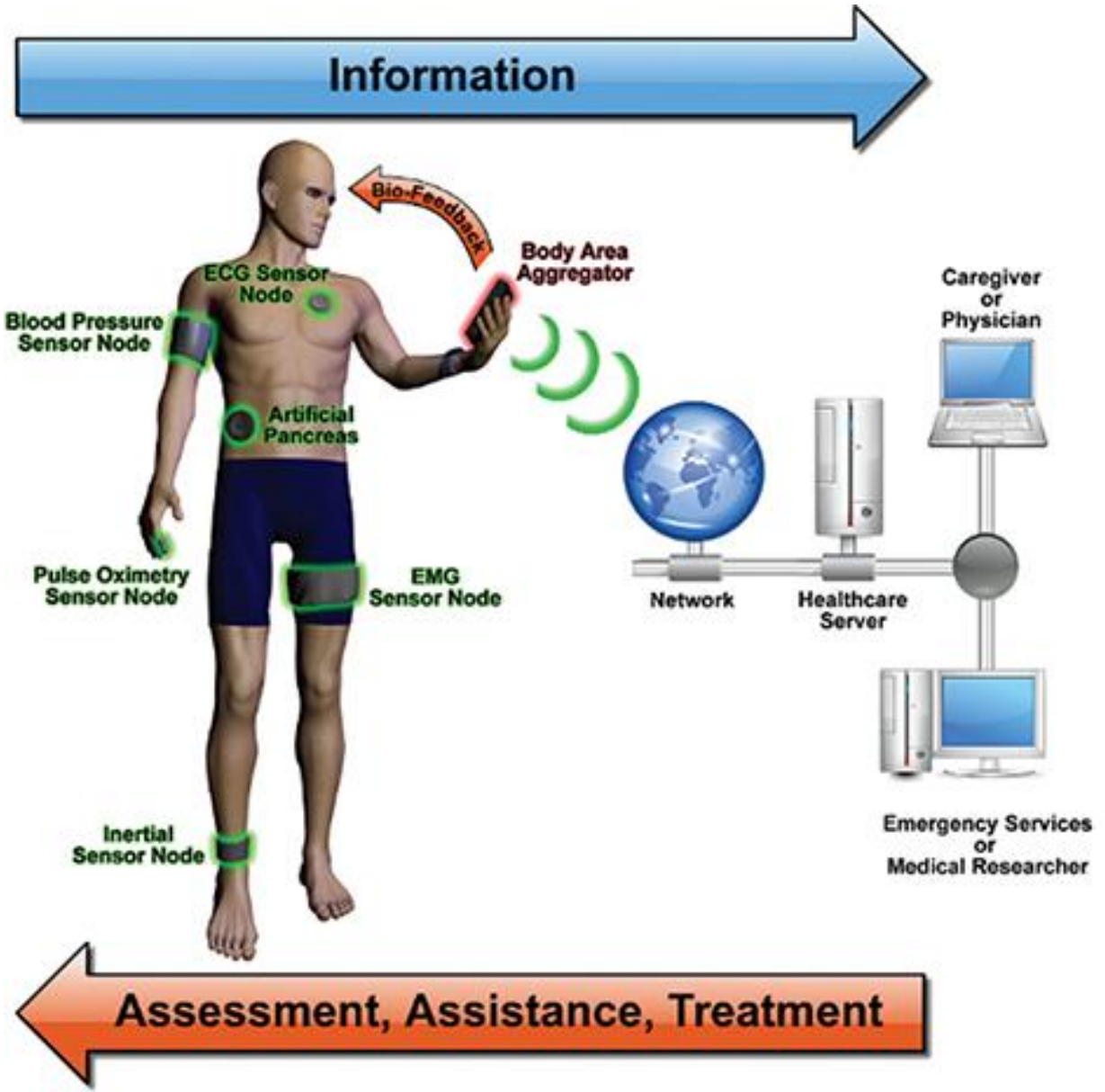


数据的利用 – 传感器

- 利用坐姿识别驾驶者身份



数据的利用 – 传感器

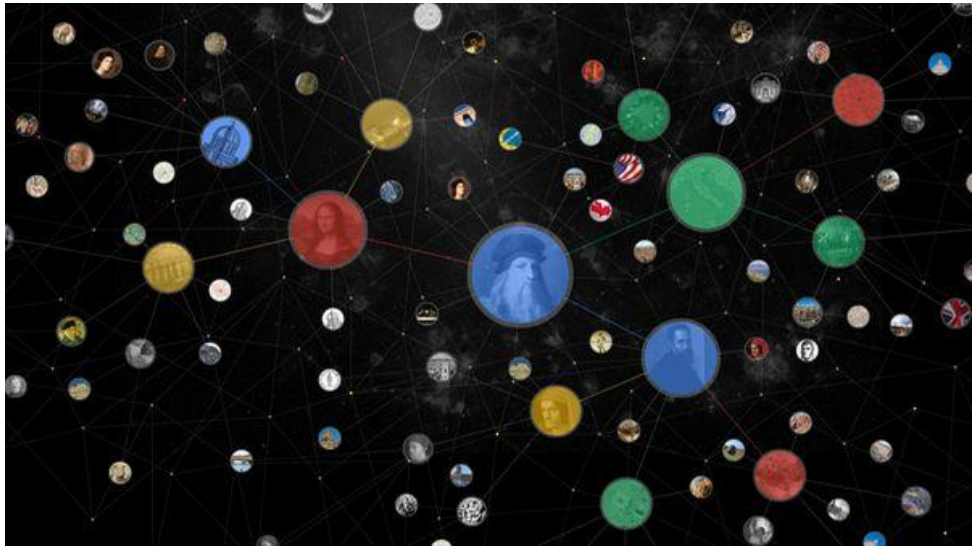

Information
Bio-Feedback
Body Area Aggregator
ECG Sensor Node
Blood Pressure Sensor Node
Artificial Pancreas
Pulse Oximetry Sensor Node
EMG Sensor Node
Inertial Sensor Node
Network
Healthcare Server
Caregiver or Physician
Emergency Services or Medical Researcher
Assessment, Assistance, Treatment

数据的利用 – 个人基因组

23andMe

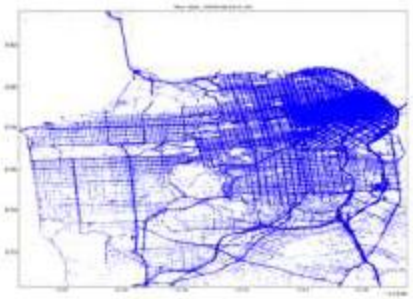
数据的利用 – 数字图书馆

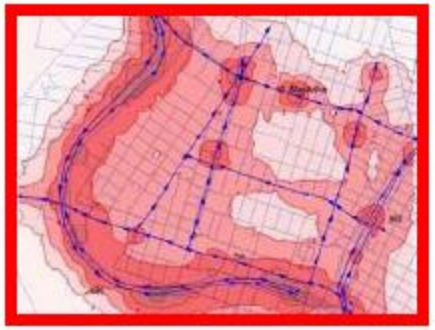
- Google Digital Library
- Google Knowledge Graph
- Google Translation



数据的利用 – 智能交通

- 加州大学伯克利分校 Mobile Millennium





Crowdsourcing + physical modeling + sensing + data assimilation

to produce:

数据的利用 – 智能交通

学习出租车司机找到快速的路

使用了北京**33,000**辆出租车**3**个月的数据

Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, and Yan Huang, T-Drive: Driving Directions Based on Taxi Trajectories, in ACM SIGSPATIAL GIS 2010, Association for Computing Machinery, Inc., 1 November 2010

数据的利用：交易数据

这些产品是根据你已经有的产品 和一些其它数据推荐的。

查看: 所有 | 新品 | 即将上架

1.

Jabra 捷波朗 easygo+ 易行+ 蓝牙耳机(白),全新版本,蓝牙3.0,支持音乐!

品牌: Jabra 捷波朗 (2012年6月24日)

用户评分: ★★★★½ (27)

现在有货

市场价: ~~¥ 399.00~~

价格: ¥ 239.00

全新品 17 售价从 ¥ 239.00 起

由 北京同升科技 提供

加入购物车 | 加入心愿单

我已经有了 | 不感兴趣 | ☆☆☆☆☆ 为该商品评分

我们提供这个推荐是因为您已购买 **Jabra 捷波朗 easygo+ 易行 蓝牙耳机(全新版本 蓝牙3.0 支持音乐 黑色)** (修改)

您最近浏览过的商品 (这是什么?)

我最近的浏览记录

诺基亚Lumia 710全新WP系统 个性彩壳 时尚智能3G手机
Nokia 诺基亚

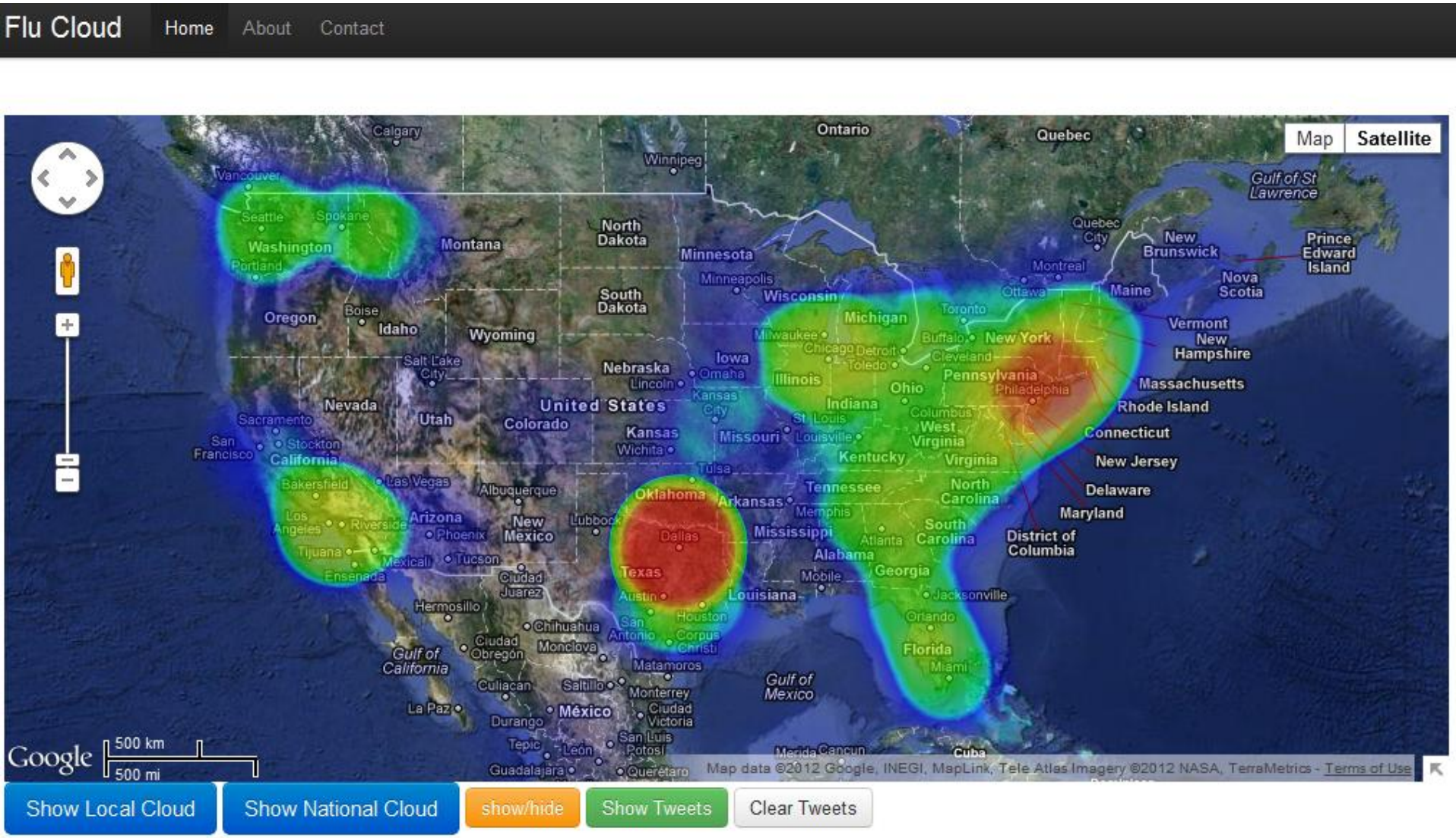
继续购买: 购买了您最近浏览过的商品的顾客同时购买了

Nokia 诺基亚 BP-3L...
★★★★½ (20)
¥ 65.00
修改这个商品推荐

诺基亚授权产品 香港VIKEN...
★★★★½ (20)
¥ 39.00
修改这个商品推荐

诺基亚Lumia...
★★★★☆ (128)
¥ 946.50
修改这个商品推荐

数据的利用 – 社交网络


Flu Cloud
Home
About
Contact
Map
Satellite
Show Local Cloud
Show National Cloud
show/hide
Show Tweets
Clear Tweets
Google
500 km
500 mi
Map data ©2012 Google, INEGI, MapLink, Tele Atlas Imagery ©2012 NASA, TerraMetrics - Terms of Use

数据的利用 – 社交网络

根据第三方票房统计机构@电影票房 统计结果： 《明日边缘》最终总票房4.05亿元！预测误差率：10% @刘挺 @丁效SCIR @景东Jingle @陈浩辰GTmacchc

@八维社会时空

本周上映电影票房预测，《明日边缘》：4.5亿。更多预测结果请关注哈工大社会计算与信息检索研究中心电影票房预测官方网站：http://t.cn/8sEEq3w @刘挺 @丁效SCIR @景东Jingle @陈浩辰GTmacchc

6月7日 11:59 来自微博 weibo.com

(1) | 转发(1) | 评论

数据的利用 – 社交（通信）网络

数据的利用 – 社交（通信）网络

大数据价值链

数据

- 数据的拥有者或收集者，价值链上最重要的一环
  - 不一定有从数据中提取价值的能力
  - 例如政府，Twitter，VISA，中航信
  - 开放或授权第三方使用数据

思维

- 先人一步发现大数据的机遇
  - 不需要拥有数据，也不需要具备专业技能
  - 只思考可能，不考虑可行性
  - 是大数据发展初期最重要的人才
  - flightCaster.com根据航班历史信息和天气信息，预测航班是否晚点
  - FlightOnTime.us也做这个预测，并慢慢削弱了flightCaster的优势

技术

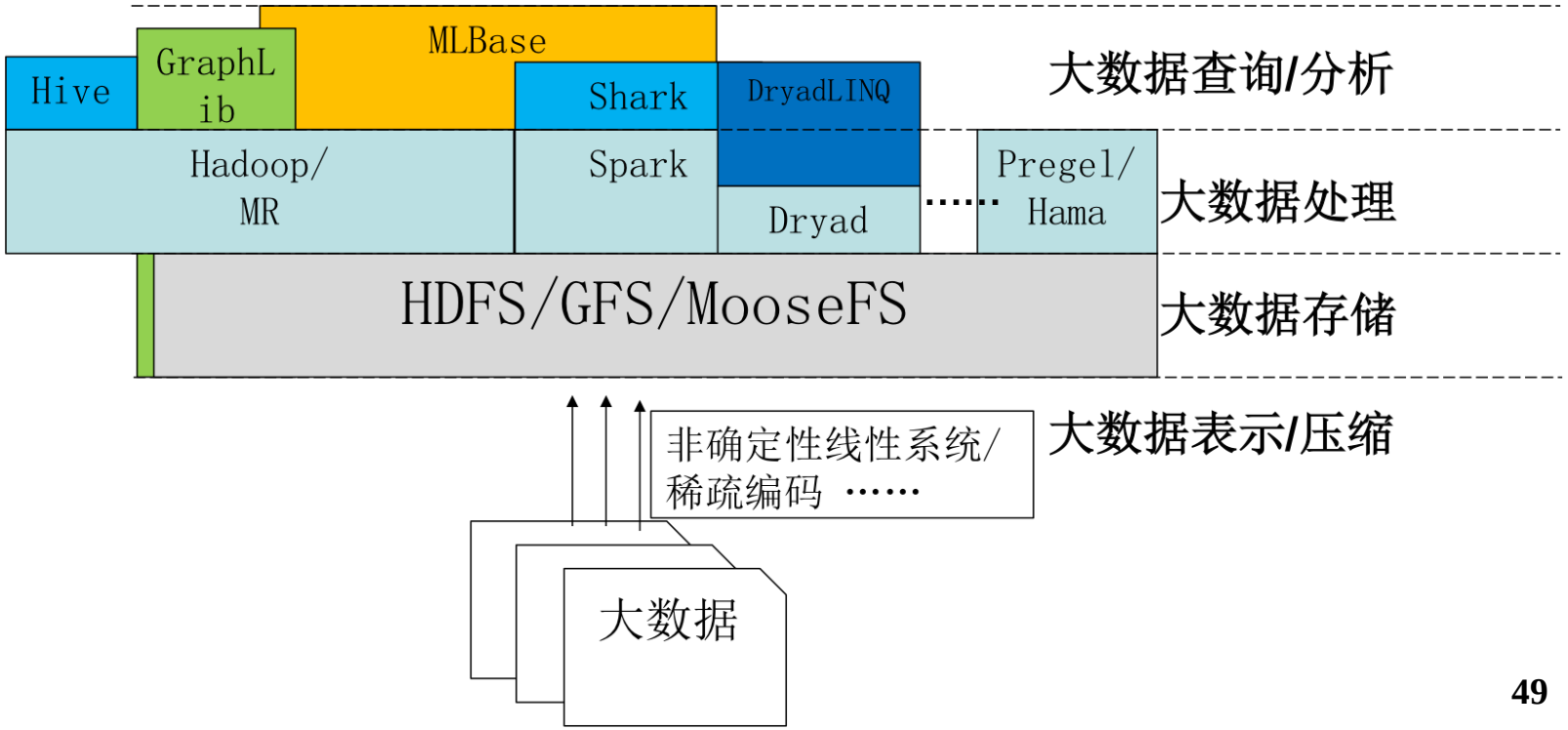
- 帮助数据拥有者完成大数据分析的任务
  - 统计方法，数据科学家
  - 利用或开发大数据分析工具
  - 例如埃森哲和美国圣路易斯市分析了公交车何时会抛锚以及最佳维修时机
  - 微软和华盛顿医院分析了如何减少感染率和再入院率 – 诊断中有“压抑”的再入院率高

提纲

- 什么是大数据
- 大数据带来的思想变革
- 大数据带来的商业变革
- 大数据系统与大数据程序设计

关键问题及技术

- 大数据的获取/表示及压缩
- **大数据存储**
- **大数据查询与处理**



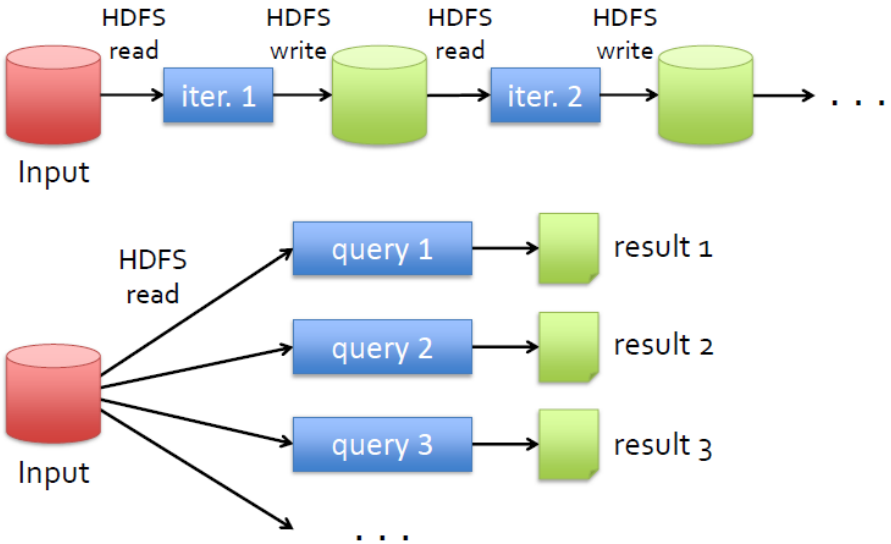
大数据存储

- 要求
  - 高可用 - 数据能够随时访问，不丢失
  - 成本低 - 对磁盘容量要求低
  - 性能 - 访问速度要快
  - 低开销 - 对CPU，网络资源占用少
- 优先级
  - 由于I/O的速度远低于CPU和网络，优先级应为
    - 高可用>成本低>性能>低开销

高可用和低成本是两个主要目标

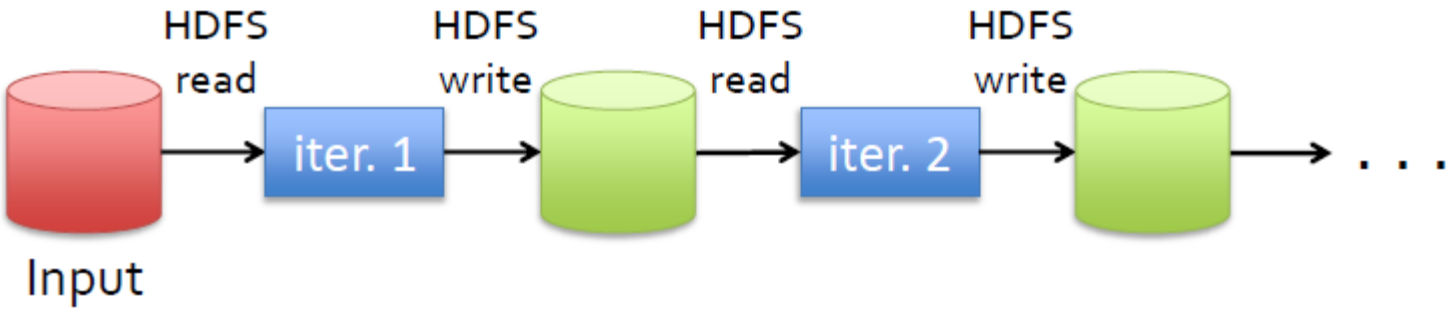
大数据的查询与分析处理

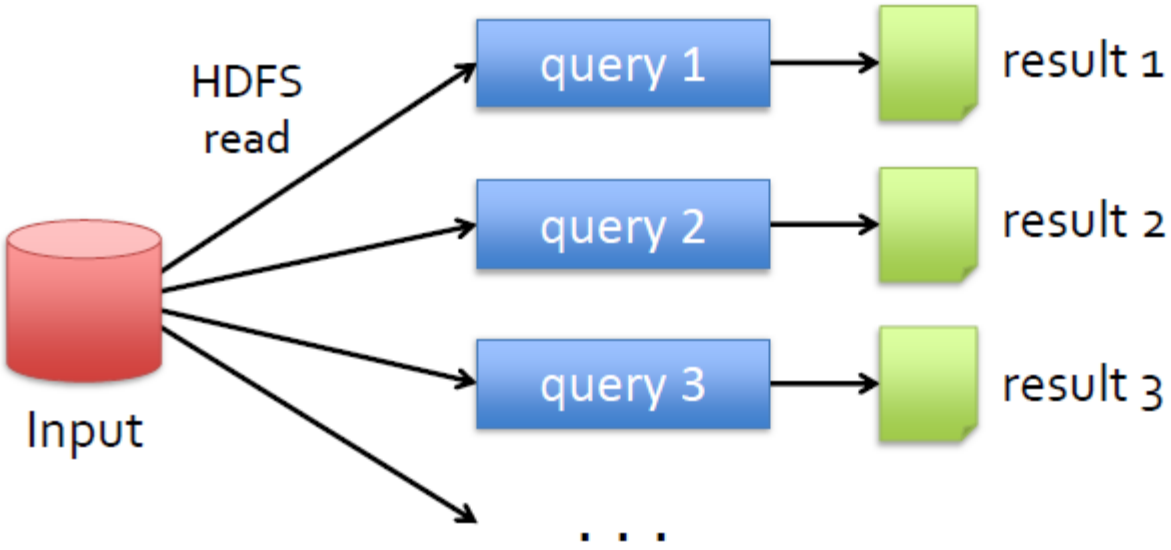
- 离线大数据分析
  - 数据分析结果没有实时性要求，例如计算网页的PageRank
- 在线大数据分析
  - 需要很快得出分析结果，例如判断金融诈骗
- 流式大数据分析
  - 存储数据有困难，分析算法可以增量化，例如网络的log处理，可以作为后续分析的过滤器



大数据的查询与处理

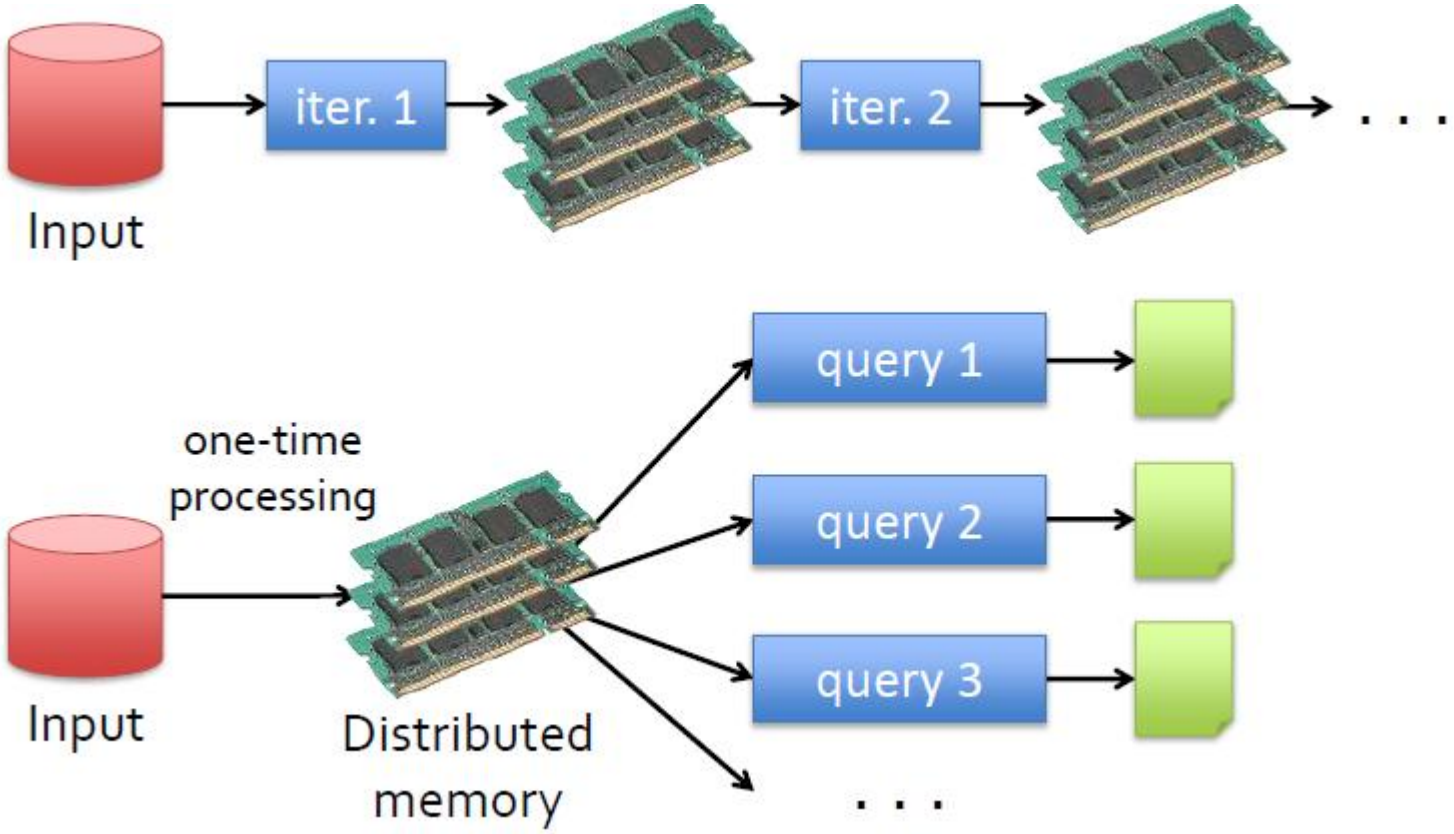
- Hadoop中的大数据处理，经常需要迭代
- 实时性较差，主要用于离线大数据的分析





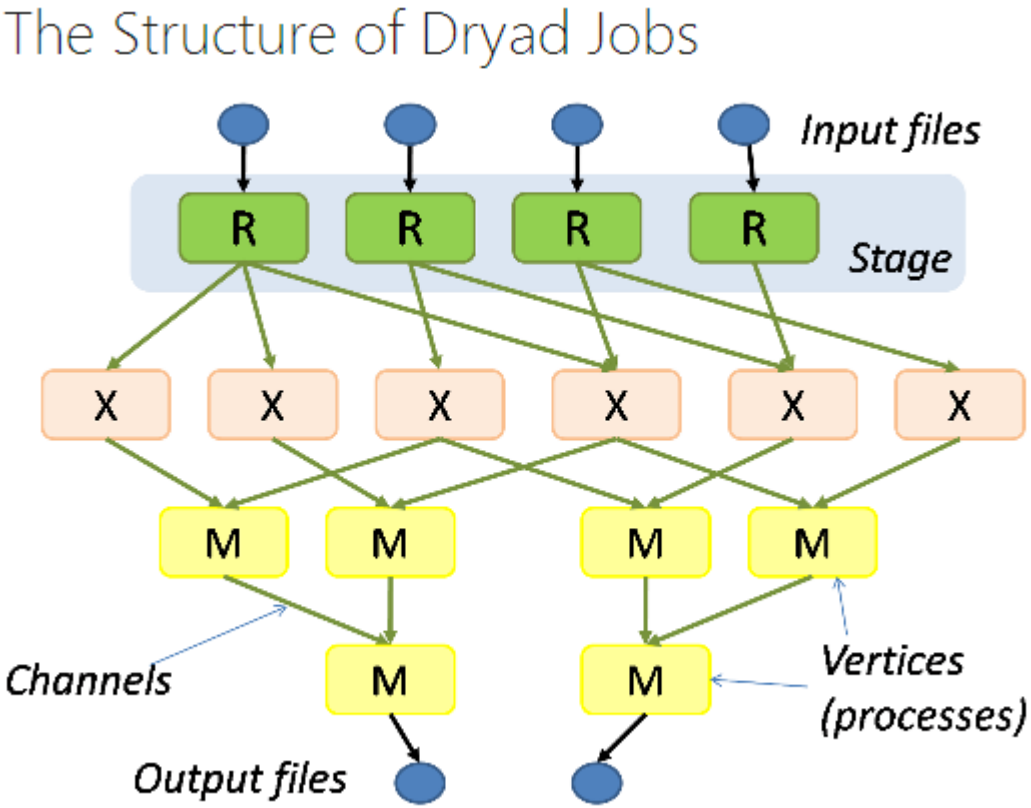
大数据处理

- Spark：对多轮迭代做优化，数据保存在内存中提高效率
- 可用于实时大数据处理



大数据处理

- Dryad：任务的DAG（有向无环图）表示，提供了更灵活的编程语义
- 可以用于流式大数据处理



大数据程序设计–基础语言

- 基础语言的选取，动态语言还是静态语言
  - 动态语言有很大的灵活性
    - 描述简洁
    - 性能可能较差
  - 静态语言性能高
    - 代码长
    - 灵活性差

数据的解析 – Python

```
40920   8.326976      0.953952      largeDoses
14488   7.153469      1.673904      smallDoses
26052   1.441871      0.805124      didntLike
75136   13.147394      0.428964     didntLike
38344   1.669788      0.134296      didntLike

for line in fr.readlines():
        line = line.strip()
        listFromLine = line.split('\t')
        returnMat[index,:] = listFromLine[0:3]
        classLabelVector.append(int(listFromLine[-1]))
        index += 1
    return returnMat,classLabelVector
```

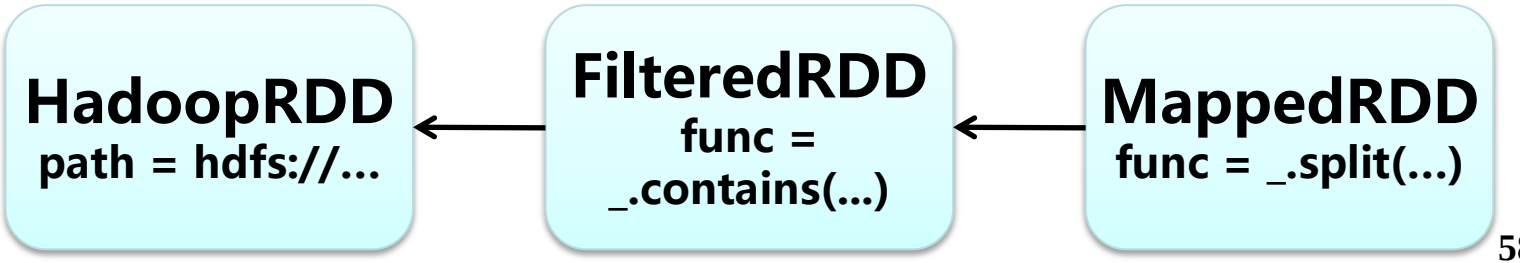
Spark的WordCount

```
val spark = new SparkContext(master, appName, [sparkHome], [jars])
val file = spark.textFile("hdfs://...")
val counts = file.flatMap(line => line.split(" "))
                 .map(word => (word, 1))
                 .reduceByKey(_ + _)
counts.saveAsTextFile("hdfs://...")
```

大数据程序设计-数据对象

- 数据对象是否可改写
- Immutable数据对象有利于容错，对性能有很大负面影响
- 数据一旦是确定性的产生，并且产生后不会变化
  - 就可以通过“重复计算”的方法来恢复数据
  - 只要记住rdd的生成过程就可以了，这样一次log可以用于很多数据，在不出错的时候几乎没有开销

```
messages = textFile(...).filter(_.contains(“error”))
                        .map(_.split(‘\t’)(2))
```

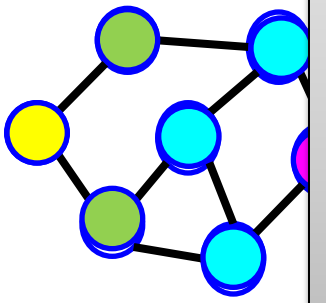


只读数据集的局限性

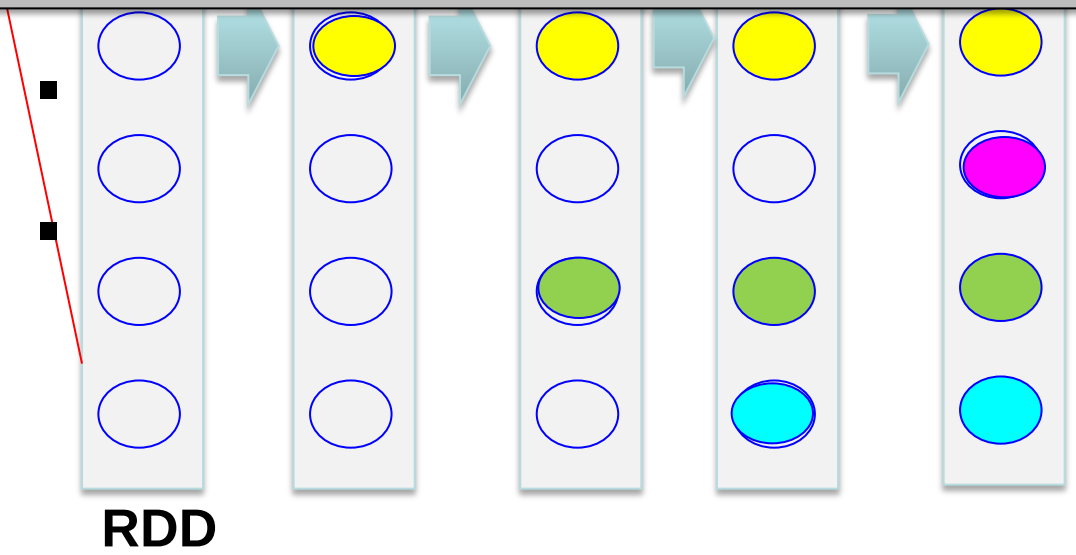
大数据应用:

部分数据更新

图遍历（BFS）



每次细粒度的数据更新，由于spark基于粗粒度RDD只读的数据对象模型，需要RDD变换，即有大量数据的复制，导致处理效率不高。



大数据程序设计 – 容错还是性能？

性能、扩展性、容错开销
应该哪个优先？

现有分布式大数据系统设计理念

MPI，OpenMP
- 可读写的数据
- 容错困难
- 不支持自动负载平衡

GraphLab
- 可读写的数据
- 容错困难
- 一定程度的自动负载平衡

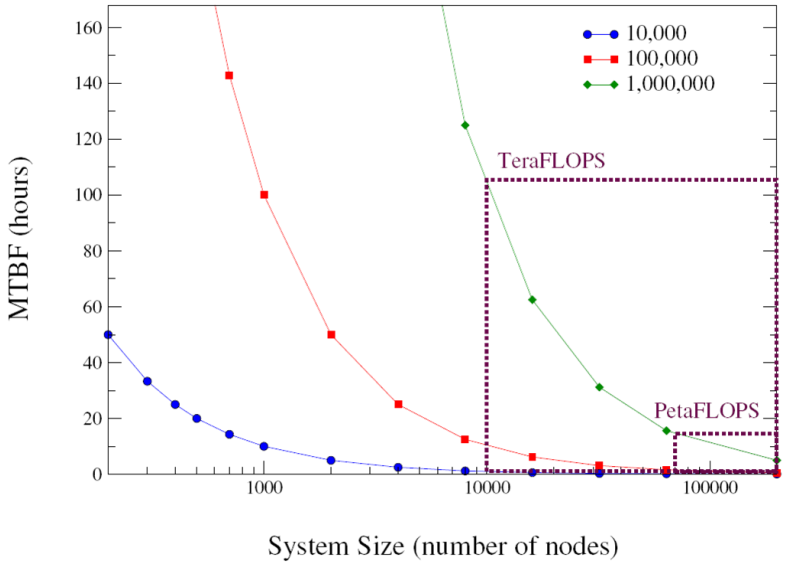
MapReduc，Spark
- 只读数据集
- 容错方便，扩展性好
- 自动负载平衡

性能

扩展性

性能的重要性

- 为什么容错是重要的?
  - 很多结点, 长运行时间
- 如果性能可以显著提高
  - 更少的结点 (可能是1个结点), 更短的运行时间
  - 容错的重要性可以大大下降



* **Fabrizio Petrini and Kei Davis and José Carlos Sancho. System-Level Fault-Tolerance in Large-Scale Parallel Machines with Buffered Coscheduling. FTPDS04, 2004.**

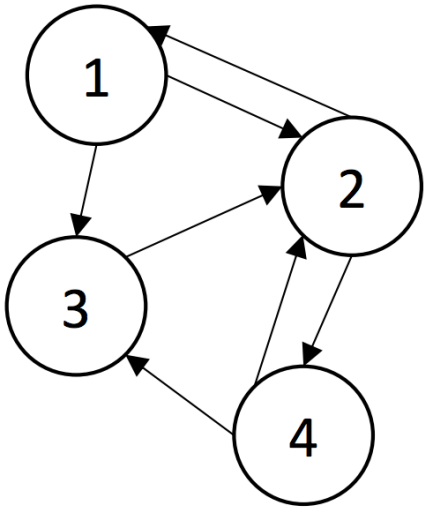
许多大数据问题的规模有限

- 人口数
  - 100亿，社交网络的大小约10TB
- 产品数
  - 100万
- 摩尔定律在驱动计算能力、内存大小和I/O带宽以指数速度增长

今天的规模有限的大数据问题
会成为明天的小数据问题

性能优先的大数据系统 - GridGraph

数据模型：可读写数据
数据结构：基于混洗的数据结构
编程抽象：基于点和边的集合
执行平台：单机内存 → Out of core
基础编程语言：C++



(1, 2) (2, 1)	(1, 3) (2, 4)
(3, 2) (4, 2)	(4, 3)

GridGraph与单机图计算系统GraphChi, X-Stream的性能对比

- 测试环境
  - 1 AWS i2.xlarge 主机
    - 4 (hypert.) vCPU cores
    - 30.5 GB memory
      - 只使用了8 GB
    - 800GB SSD

Runti

	BFS	WCC	SpMV	PageR.
LiveJournal				
GraphChi	22.05	17.28	10.12	52.08
X-Stream	6.54	14.65	6.63	18.22
GridGraph	2.11	2.53	1.96	10.54
Twitter				
GraphChi	411.3	439.6	254.0	1225
X-Stream	435.9	1199	143.9	1779
GridGraph	51.34	190.3	43.78	461.4
UK				
GraphChi	3776	2527	407.2	3307
X-Stream	8081	12057	383.7	4374
GridGraph	979.0	1264	106.3	1285
Yahoo				
GraphChi	-	-	1540	13416
X-Stream	-	-	1076	9957
GridGraph	11935	3694	379.0	3923

比现有单机图计算系统性能高3-5倍

与PowerGraph和GraphX的性能对比

- 测试环境
  - PowerGraph, GraphX
    - 16 AWS m2.4xlarge GridGraph
    - 1 AWS i2.4xlarge (4 SSDs)

System	Twitter WCC	Twitter PR	UK WCC	UK PR	Cost per hr.
PowerGraph	244	249	714	833	15.68
GraphX	251	419	647	462	15.68
GridGrpah	64	132	471	314	3.41

- **单结点性能是16结点分布式系统性能的2-3倍**
- **性价比高出一个数量级**

性能优先的大数据系统 - GridGraph

数据模型：区分只读数据和可读写数据
数据结构：基于混洗的数据结构
编程抽象：基于点和边的集合，编译与运行时优化
执行平台：单机内存 → Out of core → 分布式
CPU → GPU/APU/FPGA

编程系统	数据模型	容错能力	性能	自动负载平衡
MPI	可读写数据集	弱	高	无
MapReduce	只读数据集	强	很低	有
Spark	只读数据集	强	低	有
GraphLab	可读写数据集	弱	较高	有
GridGraph	部分只读，部分可读写	较强	高	有

总结

- 大数据是一场由信息技术推动的思想和科技、经济上的变革
  - 不是随机采样，而是所有数据
  - 不是精确性，而是混杂性
  - 不是因果性，而是相关性

- 大数据作为重要的应用领域，对程序设计的教学和时间都会产生影响
  - 大数据应用和系统都还处于发展的初期，大有可为
  - 增加对Python类的动态语言的教学
    - 灵活，自带基础数据结构
    - 如何提高Python性能是个难题

清華大學
Tsinghua University

谢谢！

第5届全国高等学校计算机程序设计课程研讨会

扫一扫访问大会官网

http://dblab.xmu.edu.cn/post/5120/