浅谈大数据及其相关技术

马 帅

北京航空航天大學
BEIHANG UNIVERSITY

个人简介

- 北京航空航天大学计算机学院教授、博士生导师；数据库专业委员会委员，大数据专家委员会委员。
- 2011年作为海外优秀中青年人才加入北京航空航天大学计算机学院软件开发环境国家重点实验，并特聘为教授。
- 获得了北京大学(2004)和英国爱丁堡大学 (2011)的两个博士学位。英国爱丁堡大学博士后，并曾在美国贝尔实验室总部实习，在微软亚洲研究院访问。

Homepage: http://mashuai.buaa.edu.cn

Email: mashuai@buaa.edu.cn

Address: Room G1122,
New Main Building,
Beihang University

提纲

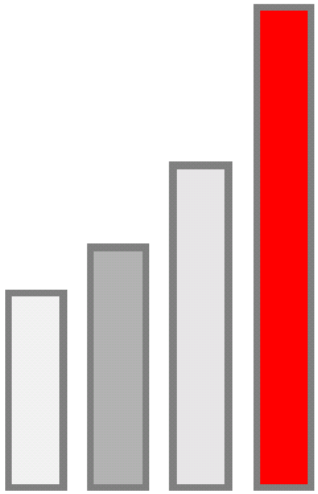
- 大数据定义
- 大数据溯源
- 大数据的应用
- 大数据相关技术

大数据(维基百科)

- [英文定义] **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- [中文定义]**大数据**或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模巨大到无法通过人工，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的信息[1][2]。
- 在总数据量相同的情况下，与个别分析独立的小型数据集相比，将各个小型数据集合并后进行分析可得出许多额外的信息和数据关系性，可用来察觉商业趋势、判定研究质量、避免疾病扩散、打击犯罪或测定实时交通路况等；这样的用途正是大型数据集盛行的原因[3][4][5]

- [1]Kusnetzky, Dan. What is "Big Data?". ZDNet.
- [2]Vance, Ashley. Start-Up Goes After Big Data With Hadoop Helper. New York Times Blog. 2010.
- [3]Data, data everywhere. The Economist. [2010-02-25].
- [4] Cat Casey and Alejandra Perez. E-Discovery Special Report: The Rising Tide of Nonlinear Review. Hudson Global. [1 July 2012]
- [5]What Technology-Assisted Electronic Discovery Teaches Us About The Role Of Humans In Technology — Re-Humanizing Technology-Assisted Review. Forbes. [1 July 2012]

“大数据”特征 – 4V

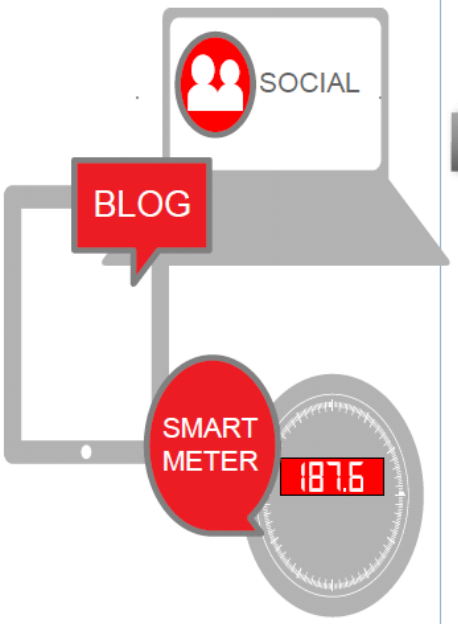


VOLUME

规模大

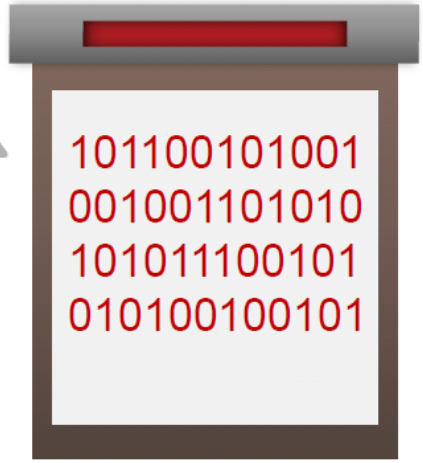
VELOCITY

变化快



VARIETY

种类杂



VALUE

价值密度低

- 时效性比正确性重要
- 技术创新比问题重要
- 大数据价值密度较低

提纲

- 大数据定义
- 大数据溯源
- 大数据的应用
- 大数据带来的挑战

“大数据”溯源

- 2008年9月4日《Nature》刊登了一个名为“Big Data”的专辑
  - Researchers need to adapt their institutions and practices in response to torrents of new data — and need to complement smart science with smart searching. http://www.nature.com/news/specials/bigdata/
- 2009年10月微软为纪念Jim Gray, 出版了“第四范式—数据密集的科学发现 (The Fourth Paradigm — Data Intensive Scientific Discovery)

“大数据”溯源

- 2011年2月11日：Science刊登了名为Dealing with Data的专辑，联合Science: Signaling、Science: Translational Medicine和Science Careers推出相关专题，讨论数据对科学研究的重要性
- 2012年3月29日，美国总统科技政策办公室OSTP（Office of Science and Technology Policy）宣布了每年投资两亿美元的“大数据研究计划”（Big Data R&D Initiative）
- 同天，我国科技部发布的“‘十二五’国家科技计划信息技术领域2013年度备选项目征集指南”把“大数据研究”列在首位

“大数据”溯源

- 2014年我国科技部关于发布国家重点基础研究发展计划：大数据计算的基础研究
- 面向网络信息空间大数据挖掘的需求，结合1-2种重要应用，研究多源异构大数据的表示、度量和语义理解方法，研究建模理论和计算模型，提出能效优化的分布存储和处理的硬件及软件系统架构，分析大数据的复杂性、可计算性与处理效率的关系，为建立大数据的科学体系提供理论依据。

“大数据”溯源

- 2015年我国科技部国家重点基础研究发展计划： 城市大数据的计算理论和方法
- 面向公共安全领域以及智能城市的实际需求，研究空间信息数据、社会网络数据等的协同表示，研究面向信息空间、物理世界和人类社会三元空间的协同感知与群智认知理论，提出视觉计算模型，建立深度计算模型，研究三元空间虚拟交互与智能控制新的模式，适应社会管理、智能城市和工业化生产等方面应用需求。

“大数据”溯源

- **国家自然科学基金委：**
  - **大数据技术和应用中的挑战性科学问题研究**
- 海量、异构和混杂大数据的广泛存在与爆炸式增长给当代信息传输、存储、计算以及面向各种应用的数据处理技术提出了前所未有的挑战。如何根据社会与国家发展需求，高效准确地传输、存储与计算各种大数据、并从已存在或动态变化的大数据中挖掘有价值的知识成为亟待解决的科学问题。
- 本重点项目群要求各申请团队：结合具体应用，突破传统研究方法的思维定式，研究和发展革命性的、可满足时代需求的大数据传输、存储、计算和处理的新方法和新技术；主要研究成果需在特定大数据集上得到验证。本重点项目群涉及如下研究方向：
- 面向大数据的知识表达、推理及在线学习理论与方法（F02）
- 基于认知计算的大数据分析方法（F02）
- 面向大数据的粒计算理论与方法（F02）
- 大数据环境下复杂多媒体内容分析、推送与展示（F02）
- 大数据管理系统评测基准的理论与方法（F02）
- 多层多域网络化大数据的高效传输理论与方法（F03）
- 大数据高效能存储与管理方法（F03）
- 大数据高时效计算体系结构与关键技术（F03）
- 大数据结构与关系的发现与简约计算方法（F03）
- 基于大数据的复杂系统行为预测与控制（F03）

提纲

- 大数据定义
- 大数据溯源
- 大数据的应用
- 大数据带来的挑战

人类 vs. 计算机 + 数据

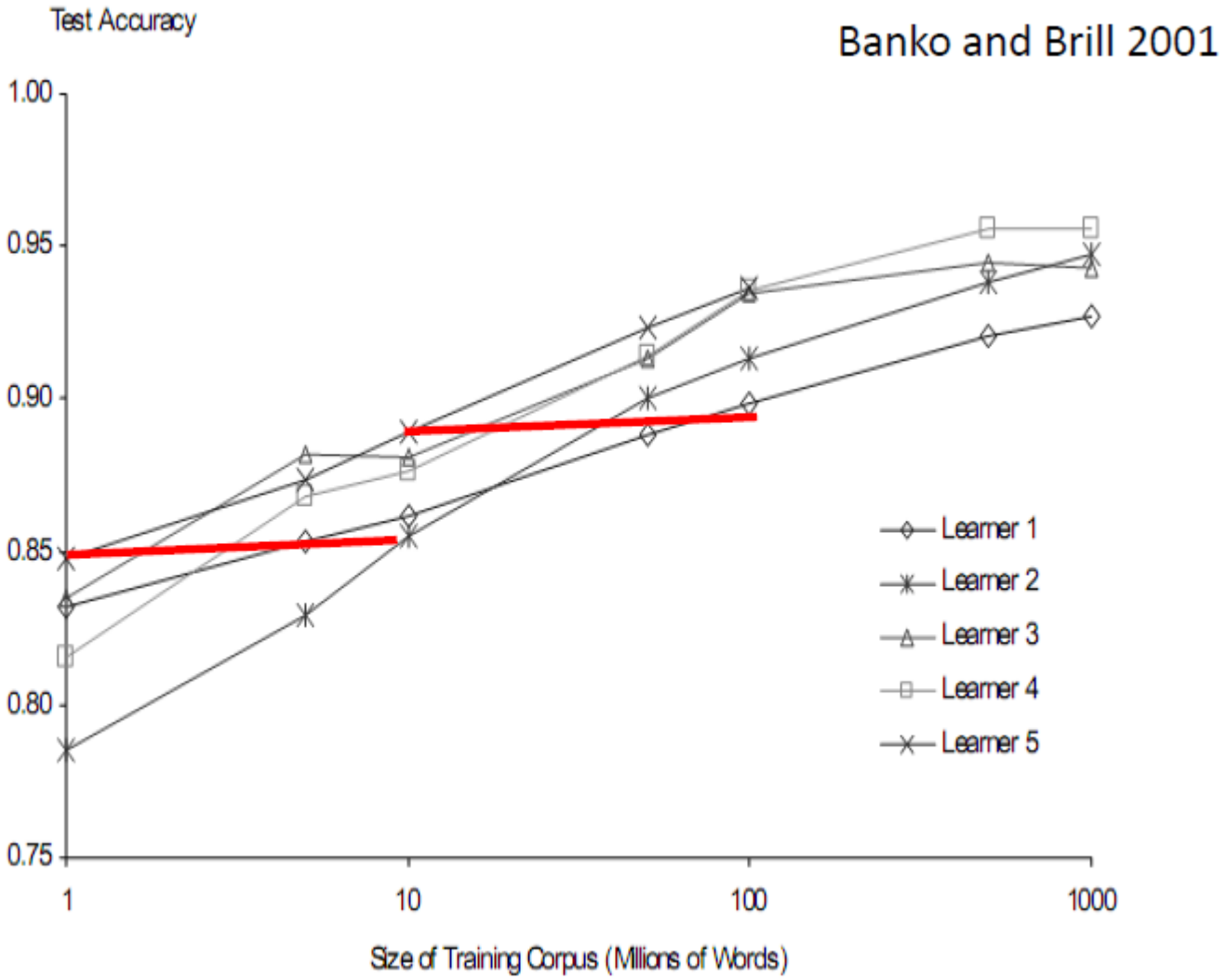
- 2011年2月11日，美国很受欢迎的智力竞答“危险边缘（Jeopardy）”电视节目
  - IBM的“沃森”系统以绝对优势战胜两名人类顶级选手
  - 和14年前的“深蓝”（战胜**加里·卡斯帕罗夫**）相比，“沃森”除具有超群的计算能力外，更拥有超大规模的数据以及数据处理能力

开普勒第三定律的发现

- 是以太阳为焦点的椭圆轨道运行的所有行星，其椭圆轨道半长轴的立方与周期的平方之比是一个常量。

Planet	Period (yr)	Ave. Dist. (au)	T^2/R^3 (yr^2/au^3)
Mercury	0.241	0.39	0.98
Venus	.615	0.72	1.01
Earth	1.00	1.00	1.00
Mars	1.88	1.52	1.01
Jupiter	11.8	5.20	0.99
Saturn	29.5	9.54	1.00
Uranus	84.0	19.18	1.00
Neptune	165	30.06	1.00
Pluto	248	39.44	1.00

更多的数据 vs. 更好的算法



互联网需求：大数据处理

- 大数据：规模大、变化快、种类杂

社交类应用

- **Facebook**：用户规模超过10亿，每天新增数据量10TB
- **四大微博(新浪，腾讯、搜狐和网易)**：用户8亿多，每天新增微博超过2亿条，图片2000万张

搜索类应用

- **百度**：每天新增日志数据量近1PB，数据总量近1000PB
- **Google**：每天新处理数据总量已超过20PB

图灵奖得主Jim Gray和IDC报告

- 数据每18月翻一番，过去数据是确定的，当前伴随人机物融合，**网络信息空间大数据呈现多样性和异构性**
- **IDC报告：全球数据2009年0.8 ZB，2012年2.7 ZB，预计2020年达35ZB (2012年的13倍)**

大数据：价值取向？

互联网改变
交流方式

大数据处理改变
经济和社会生活

Google：2007年，通过2万亿单词 训练语言模型，发现简单算法在大数据集时产生更好效果

2008年，通过庞大搜索数据训练4.5亿个数学模型，提前几周预测出H1N1流感的爆发和传播

阿里巴巴：2008年，提前8-9个月预测出金融危机

百度：通过4亿用户分析提供个性化搜索服务

Baidu 百度
Google 谷歌

熟悉用户
浏览行为

淘宝网
Taobao.com
亚马逊
amazon.com

熟悉用户
购物习惯

新浪微博
weibo.com
twitter

了解用户
思维习惯及
社会认知

数据：应用价值？

Twitter：日本海啸、地震信息提前传播，协助紧急事件的应急处理；
微博7.21北京暴雨900万条（受灾分布）、钓鱼岛4000万条（民众情绪）

Google：
2008年在甲型H1N1流感爆发几周前，提前预测冬季流感的传播
阿里巴巴：
提前8-9个月预测08年金融危机；
淘宝网：根据你的消费与浏览商品，判断你可能购买什么。
百度：通过4亿用户分析提供个性化搜索

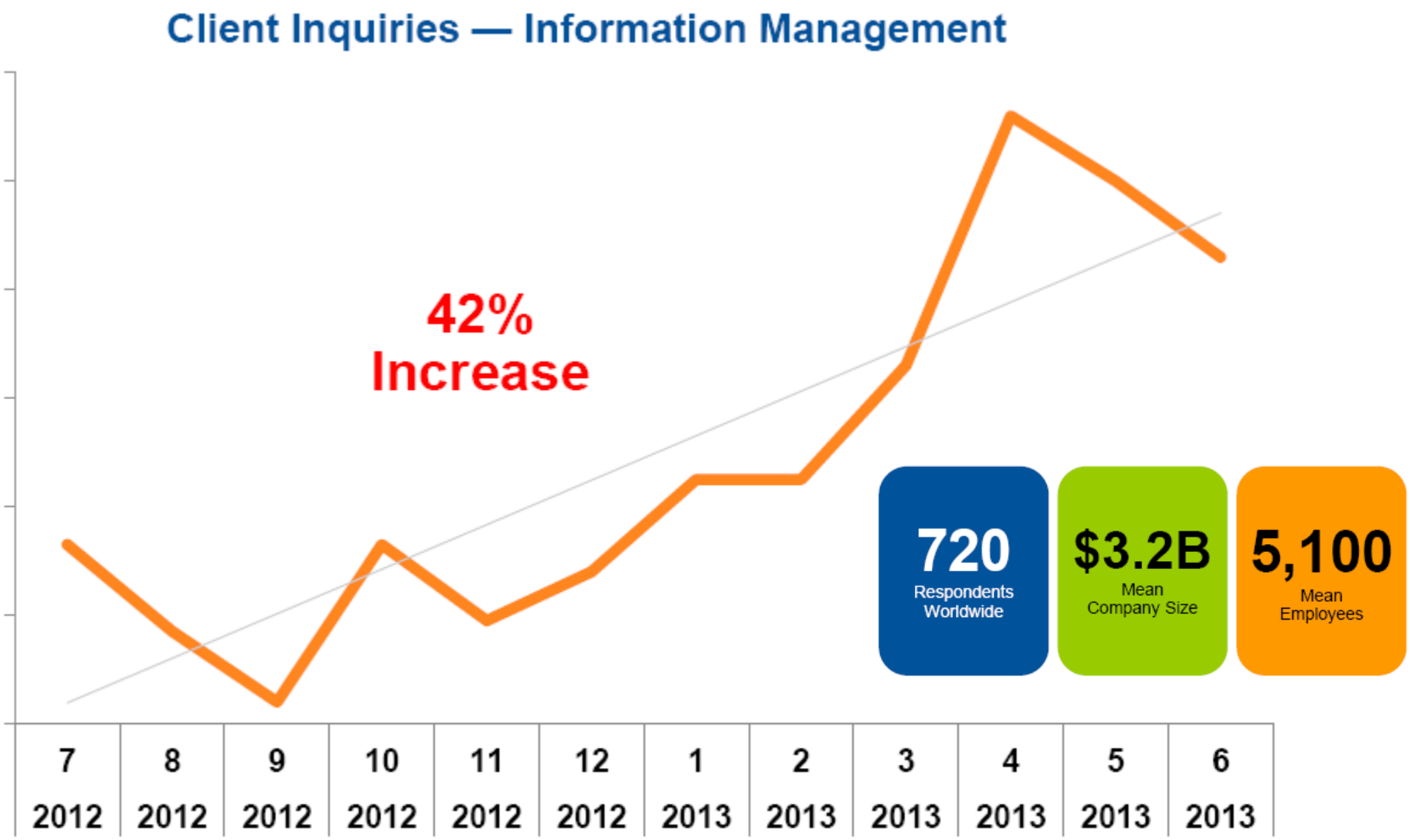
1、根据民众情绪抛售股票。对冲基金从购物网站的顾客评论，分析企业产品销售状况；
2、投资机构搜集分析上市企业声明，寻找破产的蛛丝马迹
3、**根据你关注了哪些人来判断你还可能对哪些人感兴趣。**
美国总统奥巴马的竞选团队依据选民的微博，实时分析选民对总统竞选人的喜好

实质上，通过数据的归类与分析，进行预测，例如出现某种行为的人还很有可能出现另种行为。

提纲

- 大数据定义
- 大数据溯源
- 大数据的应用
- 大数据带来的挑战

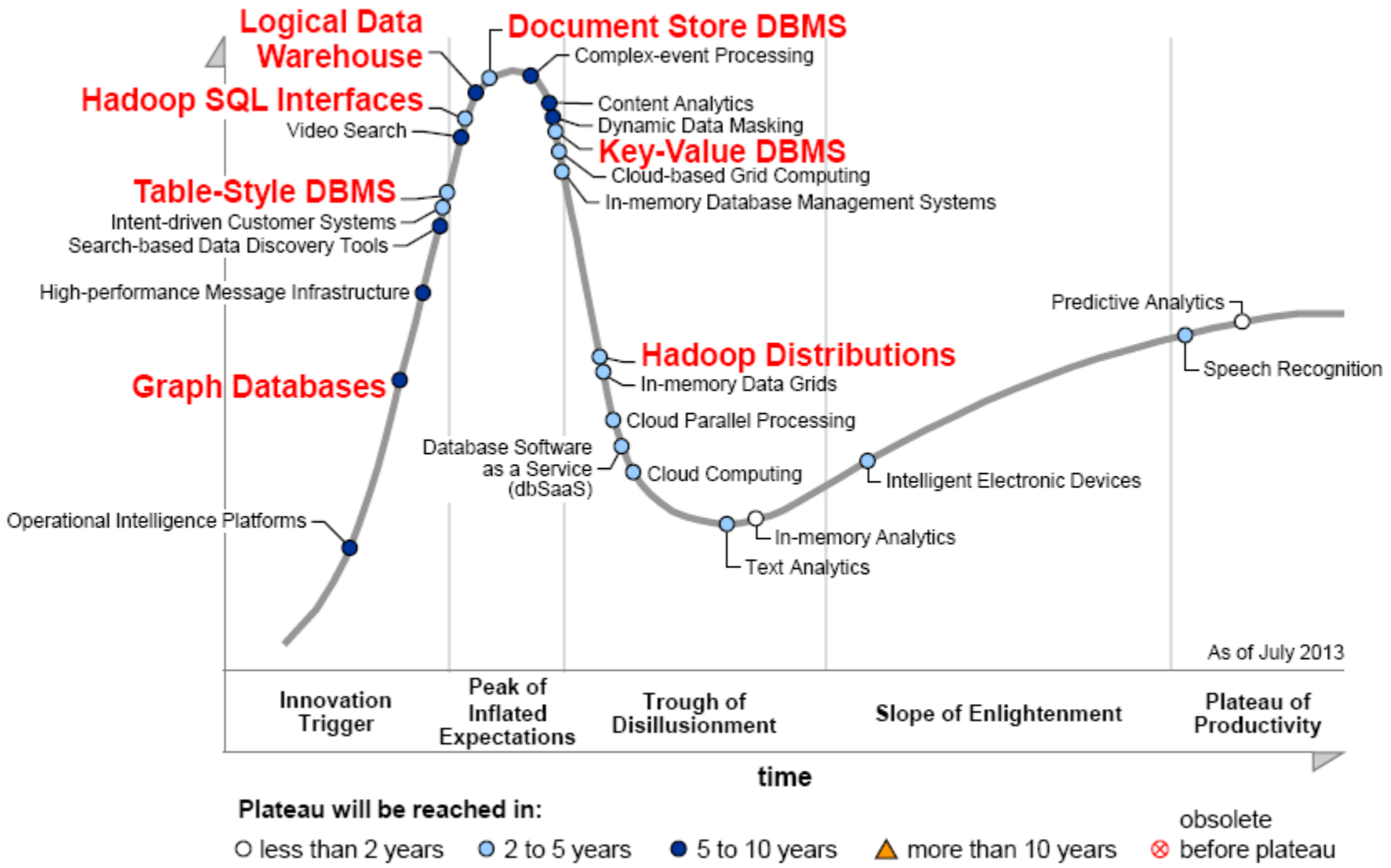
Gartner关于业界对Big Data兴趣的分析



Source: Information Management Team Inquiry Data, July 2012-June 2013

Gartner

Gartner关于Big Data处理技术的分析

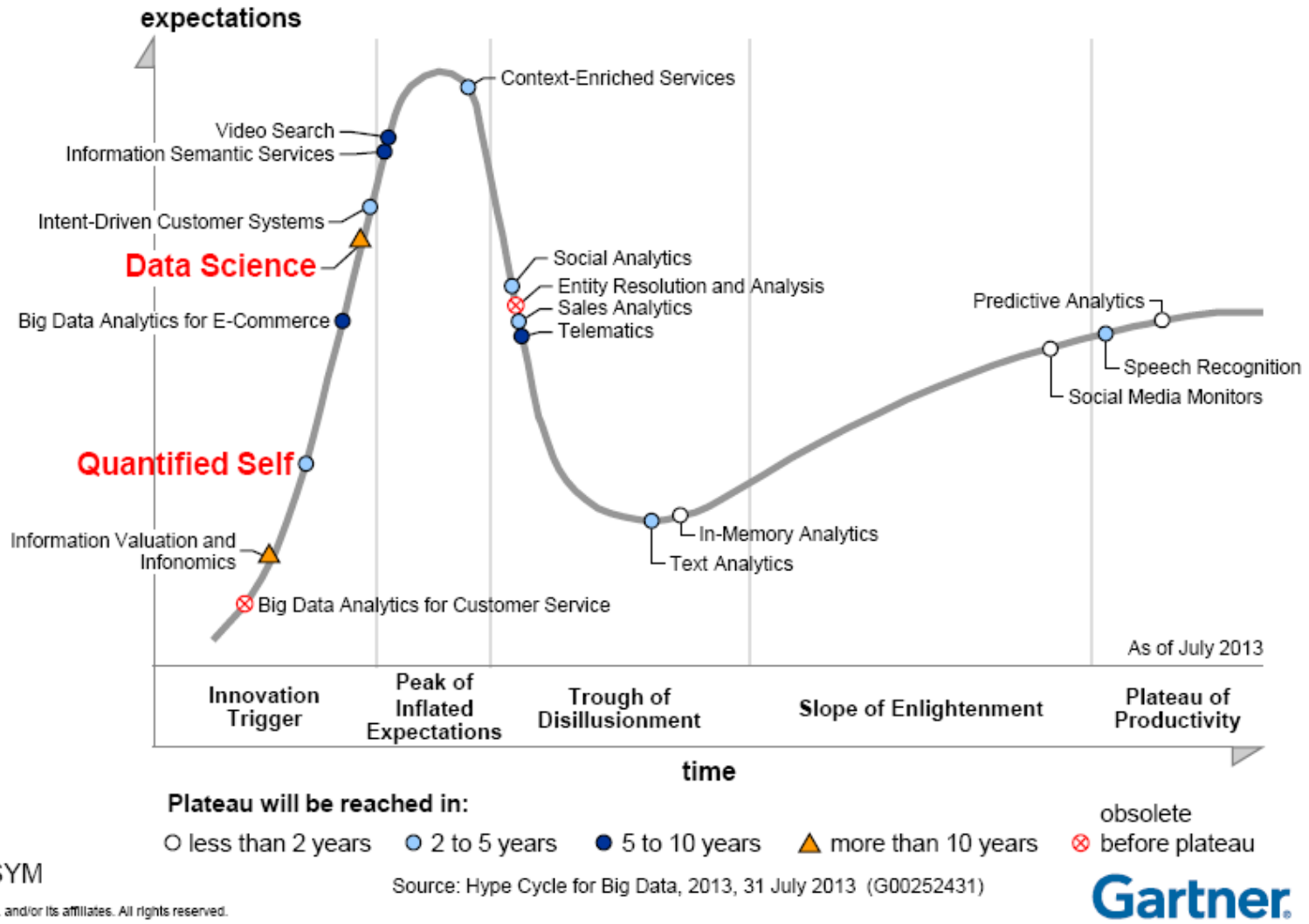


Source: Hype Cycle for Big Data, 2013, 31 July 2013 (G00252431)

#GartnerSYM

Gartner

Gartner关于Big Data处理技术的分析



#GartnerSYM

Source: Hype Cycle for Big Data, 2013, 31 July 2013 (G00252431)

Gartner

数据的处理流程

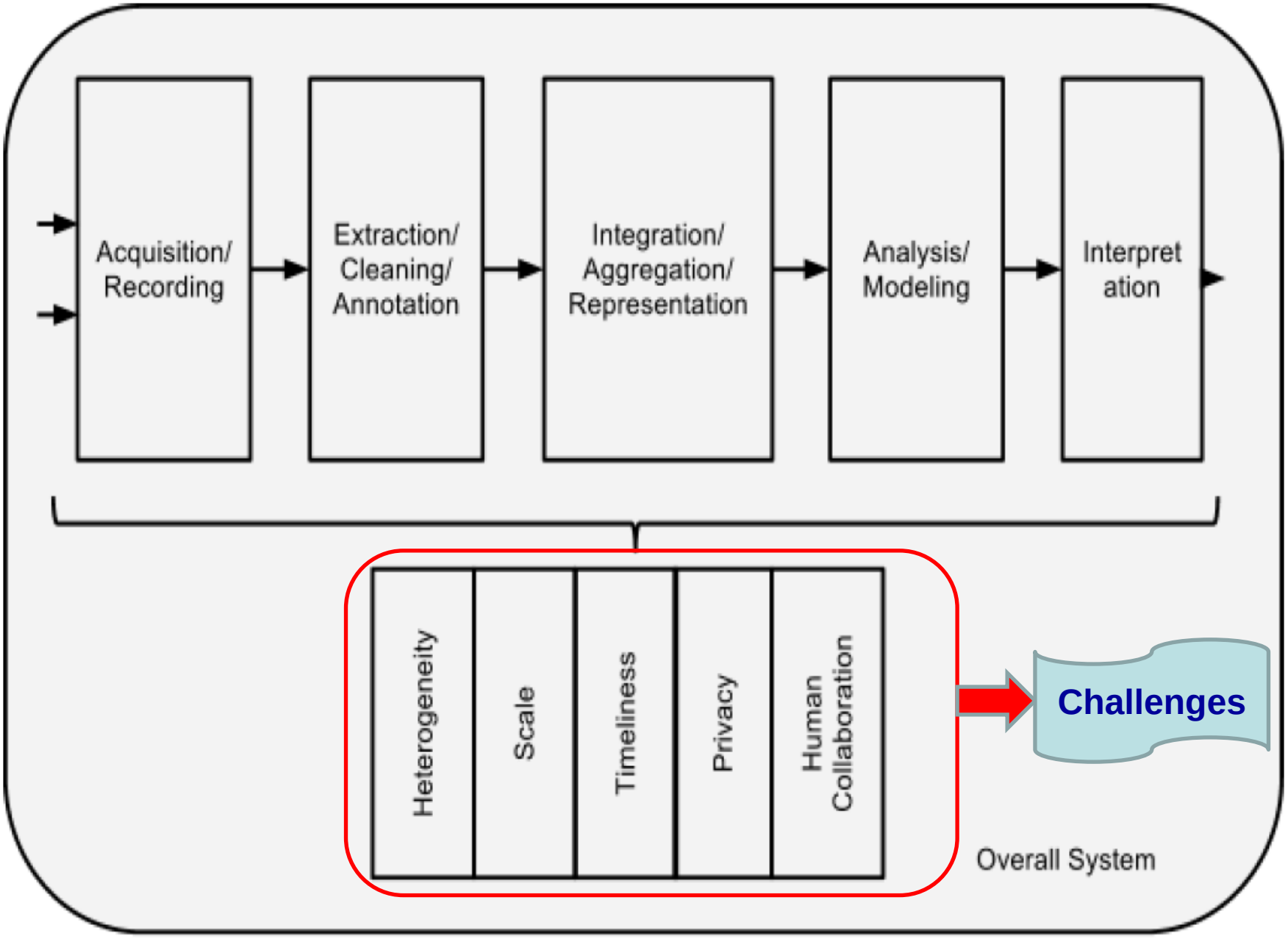
Challenges and Opportunities with Big Data

- **A community white paper** developed by leading researchers across US

Divyakant Agrawal, UC Santa Barbara
Philip Bernstein, Microsoft
Elisa Bertino, Purdue Univ.
Susan Davidson, Univ. of Pennsylvania
Umeshwar Dayal, HP
Michael Franklin, UC Berkeley
Johannes Gehrke, Cornell Univ.
Laura Haas, IBM
Alon Halevy, Google
Jiawei Han, UIUC
Alexandros Labrinidis, Univ. of Pittsburgh

Sam Madden, MIT
Yannis Papakonstantinou, UC San Diego
Jignesh M. Patel, Univ. of Wisconsin
Raghu Ramakrishnan, Yahoo!
Kenneth Ross, Columbia Univ.
Cyrus Shahabi, Univ. of Southern California
Dan Suciu, Univ. of Washington
Shiv Vaithyanathan, IBM
Jennifer Widom, Stanford Univ

A result of remote conversation lasted about 3 months (Nov. 2011 ~ Feb. 2012)

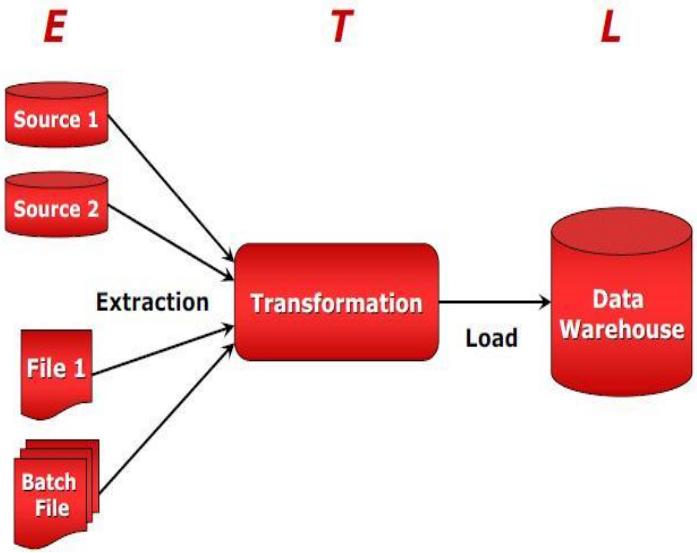

Acquisition/
Recording
Extraction/
Cleaning/
Annotation
Integration/
Aggregation/
Representation
Analysis/
Modeling
Interpret
ation
Heterogeneity
Scale
Timeliness
Privacy
Human
Collaboration
Challenges
Overall System

大数据处理技术分析

- 数据采集
  - ETL工具、爬虫、传感器
- 数据存储
  - 文件系统、关系数据库、图数据库；NoSQL（hadoop）；
- 数据分析
  - NLP、统计、数据挖掘、机器学习、数据库
- 数据展现
- 数据类别
  - 类型（结构、）
  - 行业（医疗、社交）

数据采集–ETL

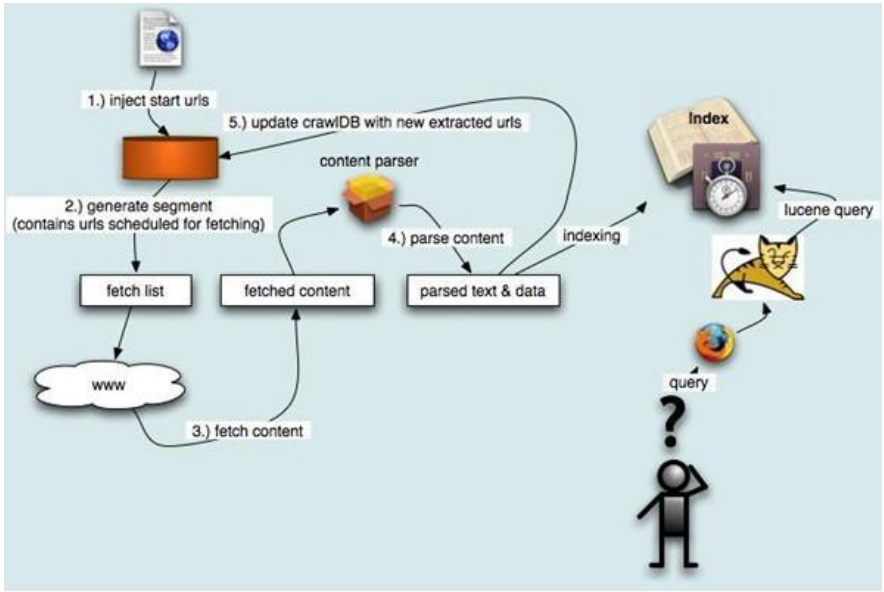
- Extract, Transform and Load (ETL)
  - ETL按照统一的规则集成并提高数据的价值，是负责完成数据从数据源向目标数据仓库转化的过程。



图片来源：http://igorportela.com/extract-transform-and-load-etl/

数据采集-爬虫

- 网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列，直到满足系统的一定停止条件。



图片来源：http://blog.csdn.net/pipi521520/article/details/5599919

数据采集–传感器

- 数据采集是指从传感器和其它待测设备等模拟和数字被测单元中自动采非电量或者电量信号,送到上位机中进行分析，处理。

图片来源http://www.acurite.com/sensor-based-forecasting

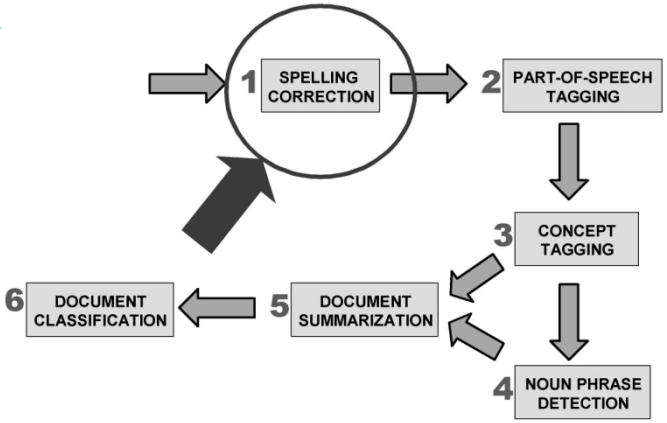
数据存储

- 文件系统
  - 文件数据库又叫嵌入式数据库，将整个数据库的内容保存在单个索引文件中，以便于数据库的发布。
- 关系数据库
  - 关系数据库，是建立在关系模型基础上的数据库，借助于集合代数等数学概念和方法来处理数据库中的数据
- 图数据库
  - 图数据库的基本含义是以“图”这种数据结构存储和查询数据.
- NoSQL（hadoop）
  - 非关系型数据库以键值对存储（key-value），它的结构不固定，每一个元组可以有不一样的字段，每个元组可以根据需要增加一些自己的键值对，这样就不会局限于固定的结构，可以减少一些时间和空间的开销。

数据处理与分析

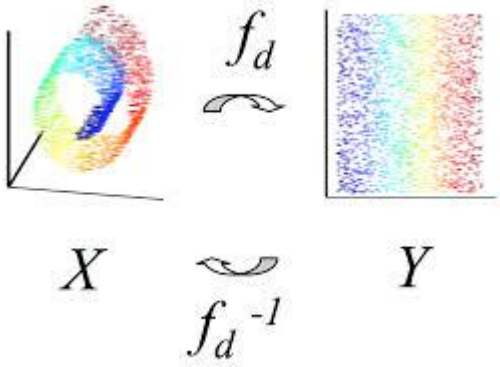
- 数据处理：
  - 自然语言处理技术

实现人与计算机之间用自然语言进行有效通信的各种理论和方法



  - 数据降维技术

将样本点从输入空间通过线性或非线性变换映射到一个低维空间，从而获得一个关于原数据集紧致的低维表示



  - 数据清理技术

发现并纠正*数据*文件中可识别的错误，包括检查*数据*一致性，处理无效值和缺失值等

数据仓库与联机分析处理

- 1988年IBM两位研究人员（Barry Devlin和Paul Murphy）创造性地提出了一个新的术语：数据仓库（Data Warehouse）

- 1992年比尔.恩门出版专著《Building the Data Warehouse》，真正拉开了数据仓库走向大规模应用的序幕，被誉为“数据仓库之父”

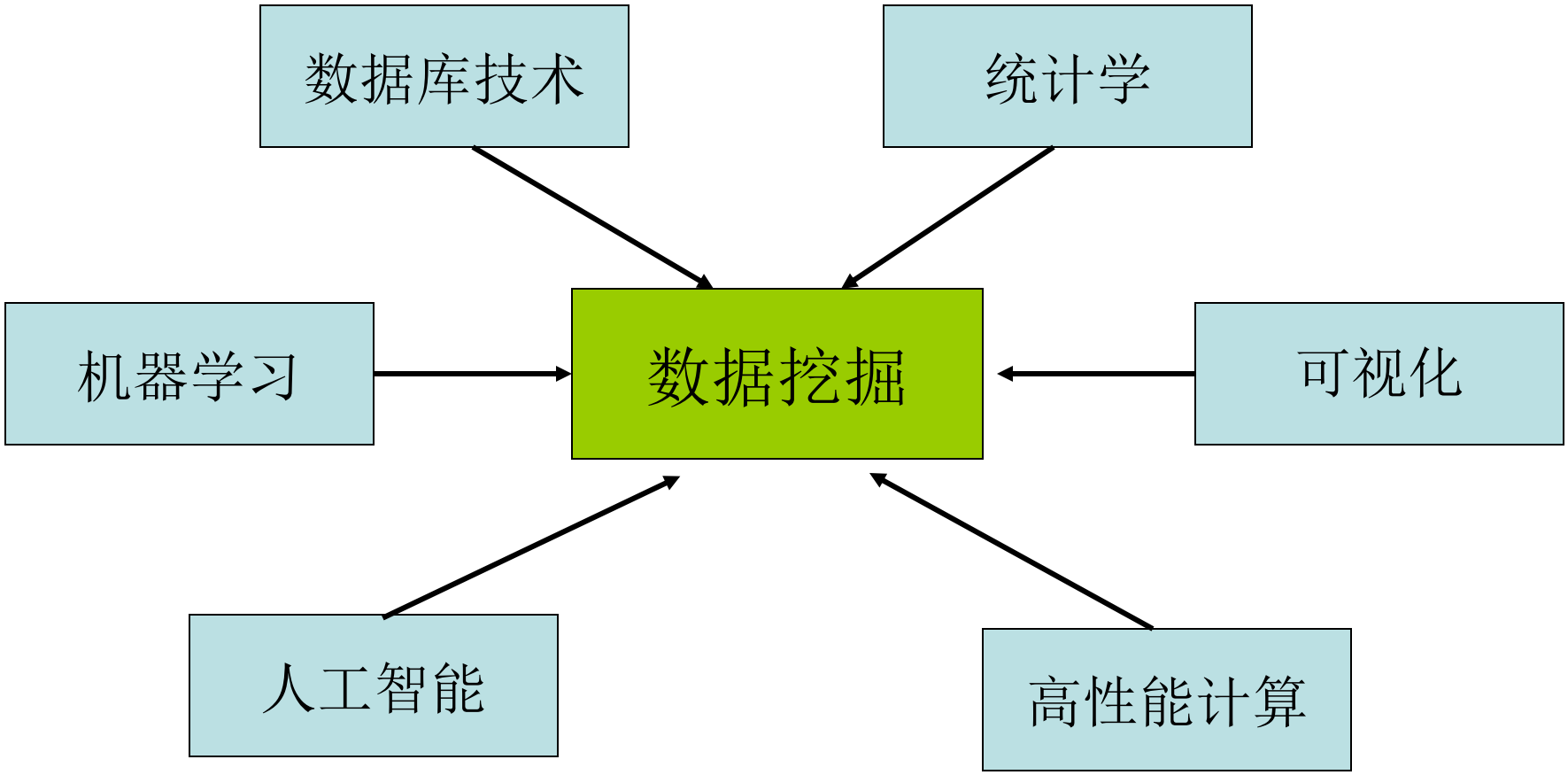
“数据仓库是一个面向主题的、集成的、相对稳定、反映历史变化的数据集合，用于支持管理中的决策制定”

- 数据仓库与数据库的主要区别：
  - 数据仓库以数据分析、决策支持为目的来组织存储数据
  - 数据库的主要目的是为系统保存、查询数据

数据挖掘的定义

- **数据挖掘是从大量数据中提取或“挖掘”知识。**
- 技术上的定义：数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。
- 商业角度定义：数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。
- 所谓**基于数据库的知识发现**（KDD）是指从大量数据中提取有效的、新颖的、潜在有用的、最终可被理解的模式的非平凡过程。

数据挖掘是多学科的产物



数据挖掘

- **数据挖掘算法按挖掘目的分为：**
  - 关联规则分析
  - 分类与预测
    - ✓ 信息自动分类，信息过滤，图像识别等
  - 聚类分析
  - 异常分析
    - ✓ 入侵检测，金融安全等
  - 趋势、演化分析
    - ✓ 回归，序列模式挖掘

数据挖掘：在你的数据中搜索知识（有趣的模式）

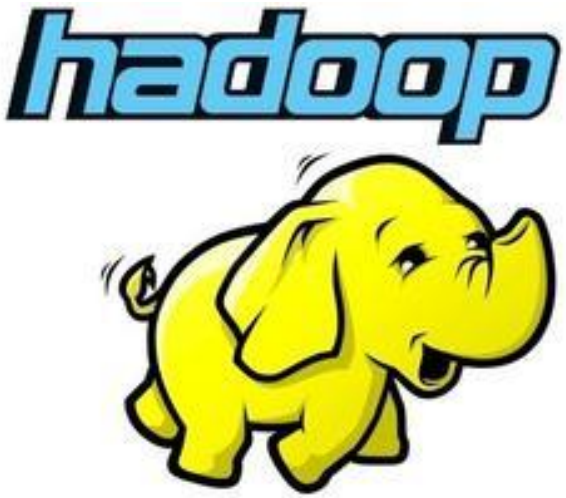
大数据的应用—决策支持

- 1947年，赫伯特. 西蒙在著作《行政组织的决策过程》中指出“人类的理性是有限的，因此所有的决策都是基于有限理论（bounded rationality）的结果”，并指出“如果能利用存储在计算机里的信息来辅助决策，人类理性的范围将会扩大，决策的质量就能提高”
- 预测“在后工业时代，也就是信息时代，人类社会面临的中心问题将从如何提高生产庇转变为如何更好地利用信息来辅助决策”

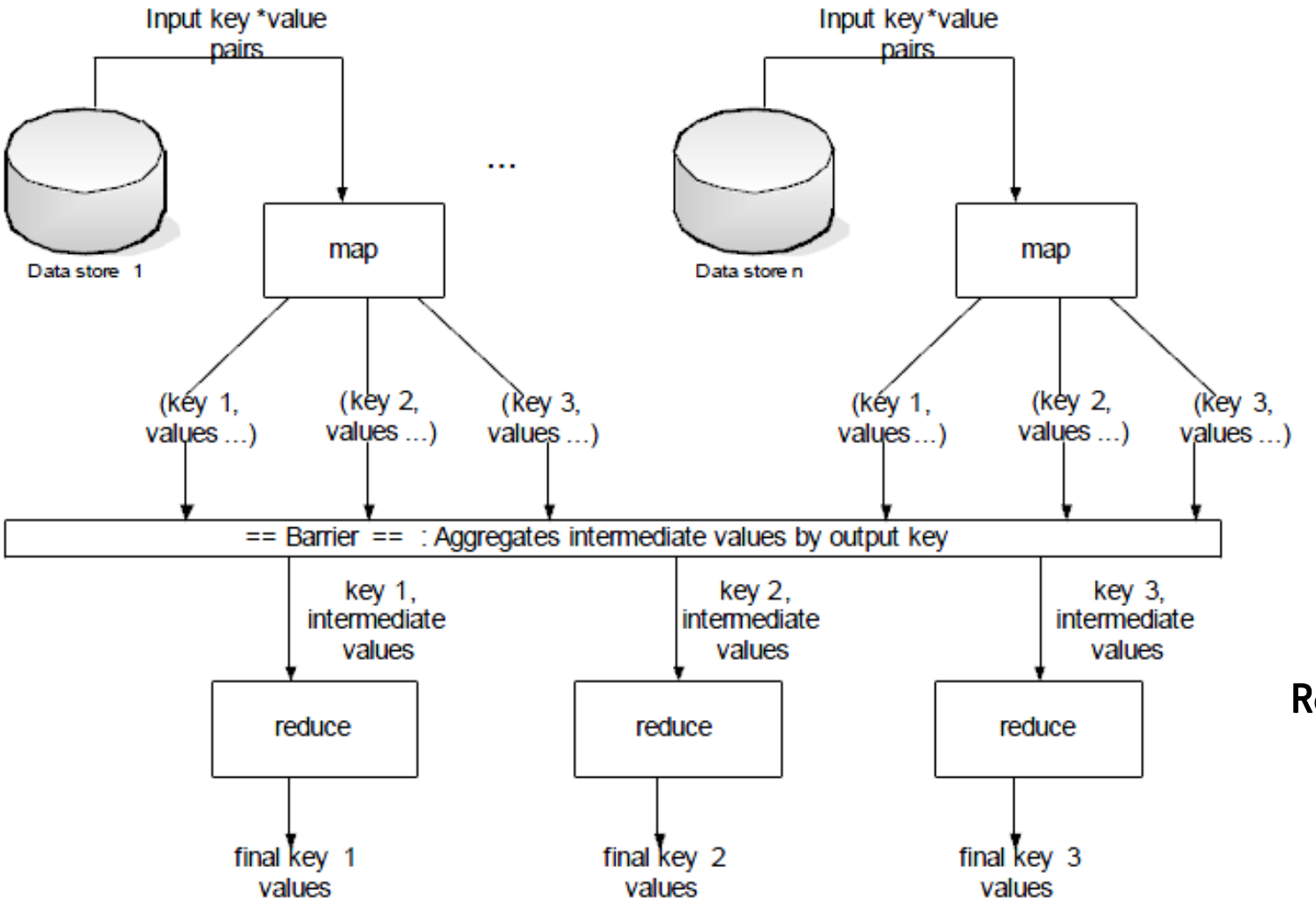
1975年图灵奖
1978年诺贝尔经济学奖
1993年美国心理协会终身成就奖

MapReduce/Hadoop and Beyond

- 由Google提出的一个用于大数据处理的系统
  - Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, OSDI 2004.
- Apache开源社会项目： Hadoop
- 主要的思想来自于functional programming



MapReduce/Hadoop and Beyond



Map阶段

Reduce阶段

MapReduce/Hadoop and Beyond

- MapReduce/Hadoop的局限性
  - 比较底层的编程模型
  - 对实时处理和递归处理支持不够
  - 适合处理具有“局部性”的数理

- Beyond MapReduce
  - 高层编程语言：Hive (Facebook), Pig (Yahoo!)等…
  - 流式计算：S4 (Yahoo!), Storm (Twitter), Spark (UC Berkeley AMP lab)
  - 支持递归的系统：Google Pregel
  - 其他技术。。。

大数据可视化

- 数据可视化
  - 主要旨在借助于图形化手段，清晰有效地传达与沟通信息。
  - 美学形式与功能齐头并进；通过直观地传达关键的方面与特征，实现对于相当稀疏而又复杂的数据集的深入洞察。
- 数据可视化的分类 （ Frits H. Post, Gregory M. Nielson and Georges-Pierre Bonneau (2002). *Data Visualization: The State of the Art*. ）
  - 可视化算法与技术方法
  - 立体可视化
  - 信息可视化
  - 多分辨率方法
  - 建模技术方法
  - 交互技术方法与体系架构

核医学成像

螺旋星云可见光图像

大数据类别

- 数据类型
  - 结构化数据：
    - 关系数据等：数据的查询、统计、更新等操作效率低。
  - 半结构化数据：
    - XML、图数据等：转换为结构化存储或者按照非结构化存储。
  - 非结构化数据：
    - 图片、视频、word、pdf、ppt等：不利于检索、查询和存储
- 行业数据
  - 大规模的电子商务数据
  - 社会数据（社会网络，互联网等），是一类重要的图数据
  - 移动数据(呼叫详细记录、 RFID、传感器网络)
  - 医疗数据
  - 天文学，大气科学，基因组学，生物地球化学，生物和其他复杂和/或跨学科的科研数据

小结

计算模式变迁成就了时代智者

- 大数据是产业＋资源＋科学，其发展环境支撑互联网科技创新、产业应用发展和政策保障，形成“生态链”
- 需要智者：多学科融合
  - 教育思考＋科研人员＋产业领袖
- 需要实践者：真实的数据和计算平台
  - 开放的数据服务与共享平台
  - 产学研紧密合作
- 需要政策：支撑和政府支持


谢谢