浅谈大数据


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、大数据科学与工程国际联合研究中心主任；国家杰青；中国计算机学会数据库专业委员会常委，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

国家重点基础研究发展计划

- 网络信息空间大数据计算的基础研究(2014-2018)
  - Chief Scientist: Prof. Jinpeng Huai.
  - 8 institutes involved
  - Focus on “computing theory and practice on Big Data”
  - http://cnbigdata.org/

北京航空航天大學 BEIHANG UNIVERSITY

國防科学技術大学 National University of Defense Technology

ISCAS

HUAWEI

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

CNCERT/CC 国家计算机网络应急技术处理协调中心

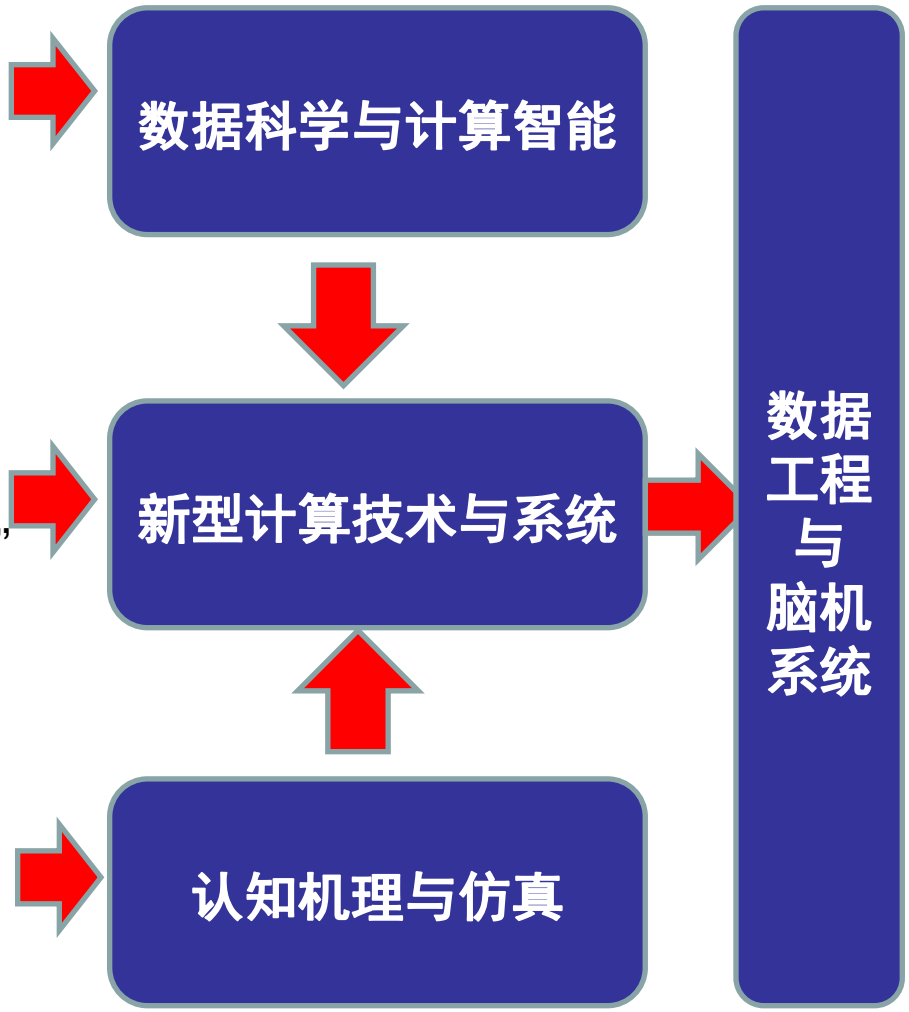
上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Baidu 百度

北京市大数据科学与脑机智能创新中心

- 瓶颈1：计算的有效性遇到障碍
  - 计算的有效性：
  - 认识数据的内在特征，复杂网络、数学（统计）方法
- 瓶颈2：能耗成为突出问题
  - 随着规模增大，调度复杂，计算系统功耗问题日益突出
  - 传统存算分离的结构，产生大量的数据搬移开销
  - 传统的计算和存储器件“功耗”不友好

- 瓶颈3：学习效率和灵活性
  - 学习效率：需要大量的输入数据及标定数据，学习效率低
  - 灵活性：普遍缺乏“类比、联想”等学习功能



http://www.bdbc.org.cn/

BIG DATA
BRAIN COMPUTING
大数据科学与脑机智能高精尖创新中心

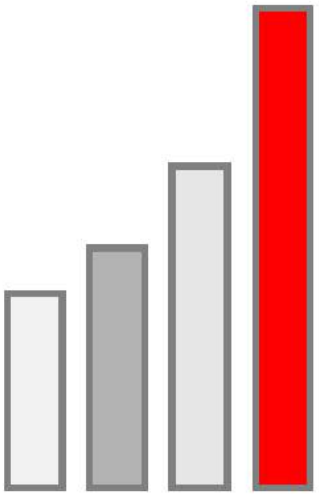
大数据简介

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

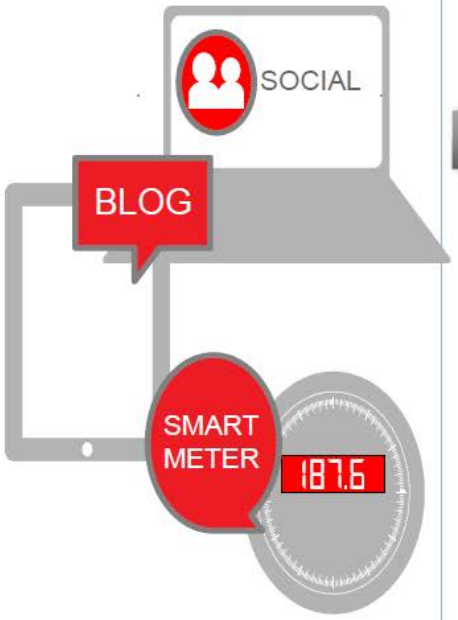


VOLUME

规模大

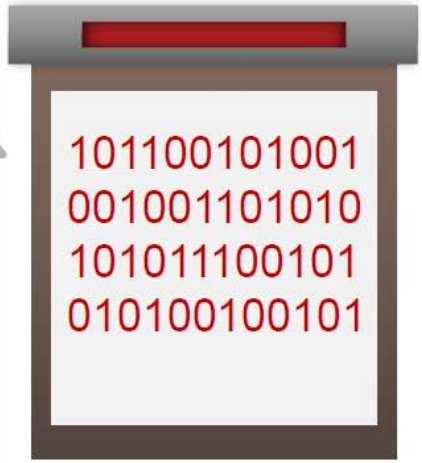
VELOCITY

变化快



VARIETY

种类杂



VALUE

价值密度低

“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)
- 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

“大数据”溯源

- 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

- 美国2016年5月发布《联邦大数据研究与开发战略计划》
- 其目标是对联邦机构的大数据相关项目和投资进行指导,主要围绕代表大数据研发关键领域的七个战略进行，包括促进人类对科学、医学和安全所有分支的认识;确保美国在研发领域继续发挥领导作用；通过研发来提高美国和世界解决紧迫社会和环境问题的能力。

“数据是一项有价值的国家资本，应对公众开放，而不是把其禁锢在政府体制。”
——美国联邦政府

http://www.thebigdata.cn/YeJieDongTai/30500.html

大数据成功案例

大数据的研究与应用：取得重大突破

- 过去8年大数据的研究，已经产生了重大突破，并在部分领域取得良好的应用
  - 计算基础：大规模云计算、大规模深度学习
  - 感知处理的角度：大规模深度学习，imageNet
  - 知识组织与管理角度：大规模知识图谱
- 基于数据产生知识的问答系统与个人辅助系统
  - Watson DeepQA：智能搜索→知识引擎
  - Apple Siri & Wolfram Alpha

大数据的研究与应用1：IBM Watson

大数据的研究与应用1： IBM Watson

- 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

大数据的研究与应用2：Google AlphaGo

2016年3月的发生在韩国首尔的一场人机围棋比赛，Google AlphaGo对战世界顶级棋手李世石九段。

- 2016年3月9日、10日和12日的三局对战均为AlphaGo获胜，而13日的对战则为李世石获胜，15日的最终局则又是AlphaGo获胜。因此对弈结果为AlphaGo 4:1战胜了李世石
- AlphaGo使用谷歌位于美国的云计算服务器，并通过光缆网络连接到韩国。

大数据的研究与应用： Google AlphaGo

大数据的研究与应用： Google AlphaGo

大数据的研究与应用3：开普勒第三定律

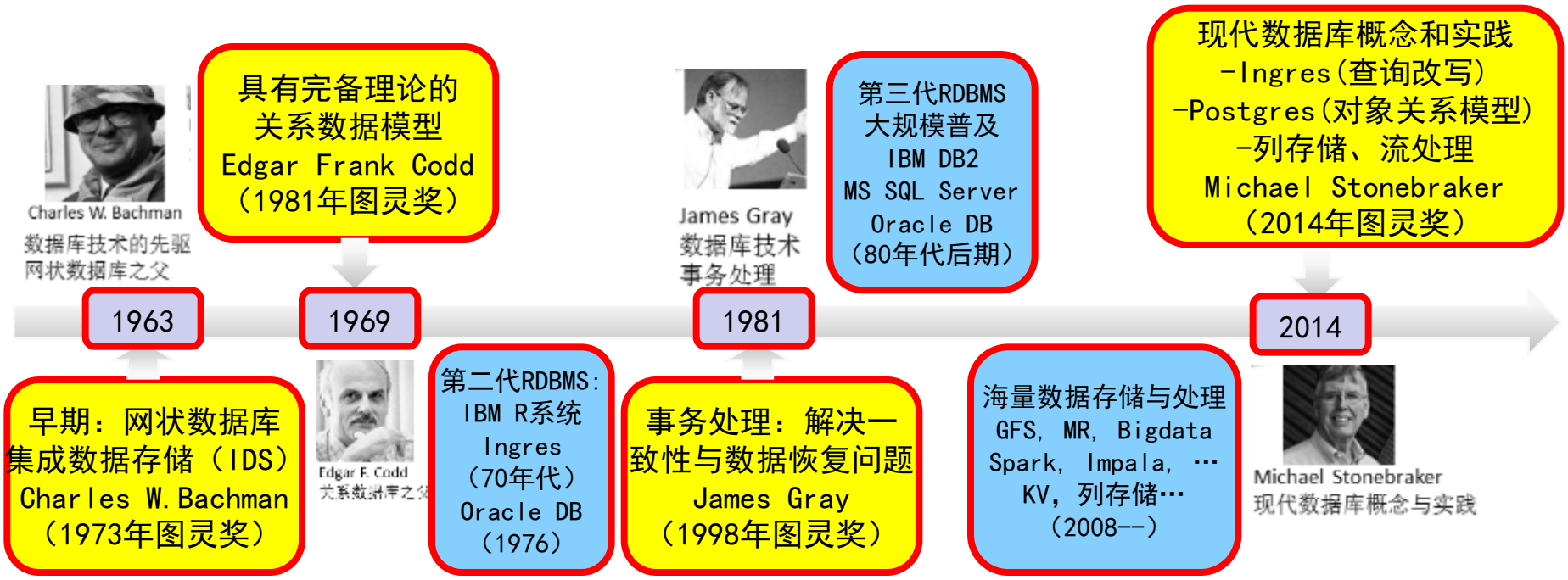
- 是以太阳为焦点的椭圆轨道运行的所有行星，其椭圆轨道半长轴的立方与周期的平方之比是一个常量。

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

大数据挑战示例

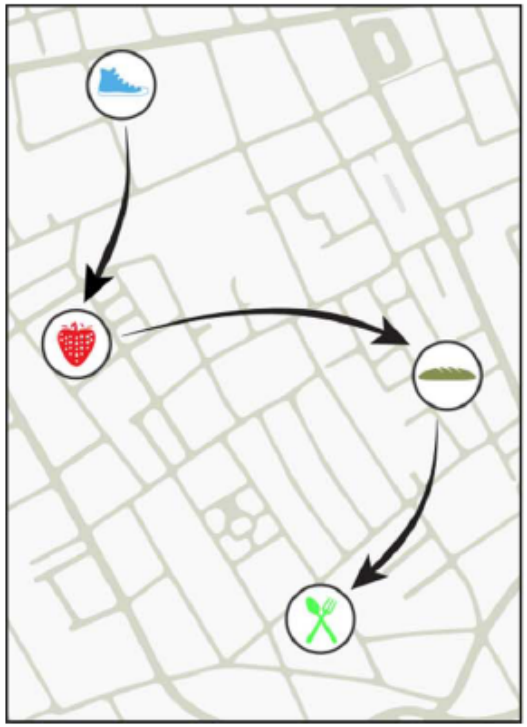
挑战：大数据的管理

- 回顾数据库领域的发展历程
  - 1960s前后，外部存储设备问世，催生数据管理需求，数据库从文件系统中分离出来
  - 1969年，具有完备理论的“关系数据模型”
  - 1981年，事务处理，解决数据一致性问题
  - 2000年以来，数据量持续增大带来的挑战



挑战：隐私保护

- Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, Alex Pentland, Unique in the shopping mall: On the reidentifiability of credit card metadata,Science, Vol. 347, Issue 6221, pp. 536-539, 2015 (report).

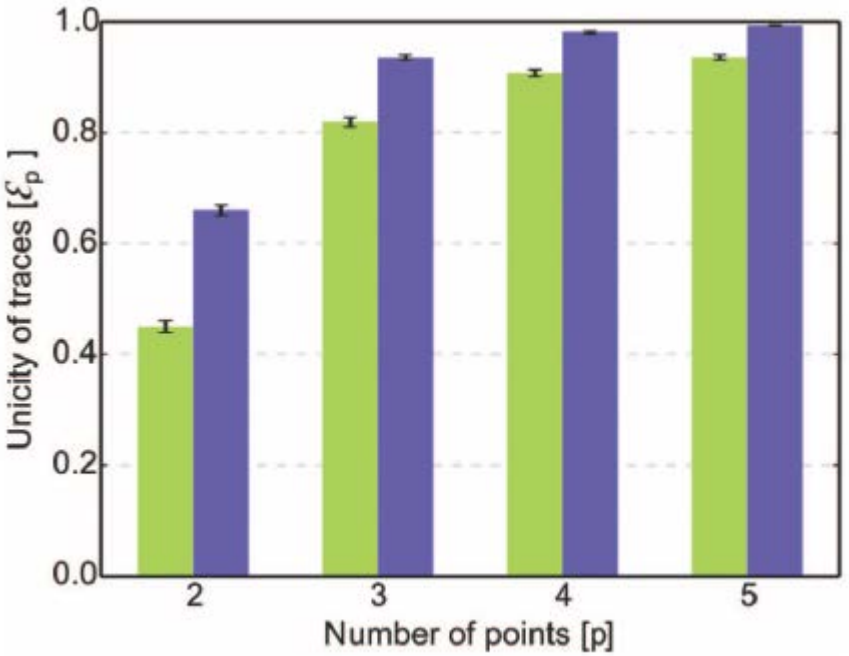


shop	user_id	time	price	price_bin
	7abc1a23	09/23	$97.30	$49 – $146
	7abc1a23	09/23	$15.13	$5 – $16
	3092fc10	09/23	$43.78	$16 – $49
	7abc1a23	09/23	$4.33	$2 – $5
	4c7af72a	09/23	$12.29	$5 – $16
	89c0829c	09/24	$3.66	$2 – $5
	7abc1a23	09/24	$35.81	$16 – $49

•**Financial traces** in a simply anonymized data set such as the one we use for this work.
•Arrows represent the **temporal sequence** of transactions for **user 7abc1a23.**.

挑战：隐私保护

- Yves-Alexandre de Montjoye, Laura Radaelli Vivek Kumar Singh, Alex Pentland, Unique in the shopping mall: On the reidentifiability of credit card metadata, Science, Vol. 347, Issue 6221, pp. 536-539, 2015 (report).



• The unicity of the credit card data set given p points.
• The green bars represent unicity when spatio-temporal tuples are known.
• The blue bars represent unicity when using spatial-temporal-price triples

• Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities, or perform research. Metadata, however, contain sensitive information.
• We study 3 months of credit card records for 1.1 million people and show that four spatiotemporal points are enough to uniquely reidentify 90% of individuals.

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

Email: mashuai@buaa.edu.cn

Address: Room G1122,
New Main Building,
Beihang University
VBeijing, China