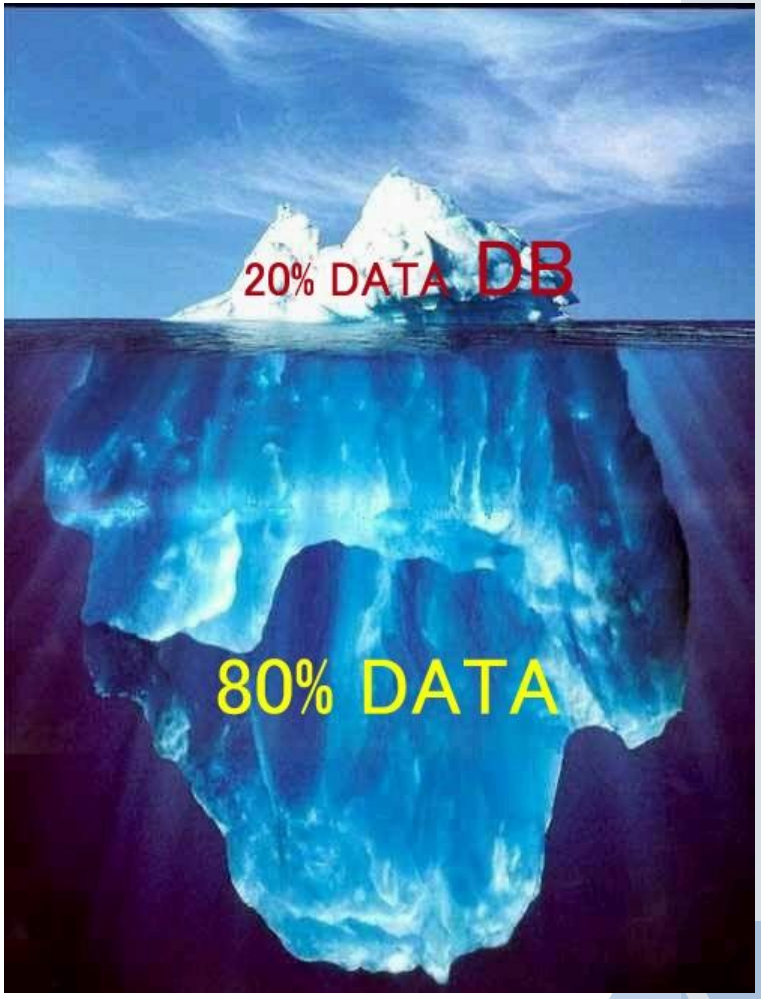
大数据管理与数据思维
Big Data and Data Thinking

孟小峰
中国人民大学信息学院

•http://idke.ruc.edu.cn

数据海洋

- 海量异构
  - **281EB in 2009**
  - 仍以指数级增长
- 内容多样
  - 半结构
  - 非结构
- 动态增长



Data Explosion

Big Data

MB = 10^6 bytes

a typical book in text format

GB = 10^9 bytes

a one hour video is about 1GB;
data produced by a biology experiment in one day

TB = 10^{12} bytes

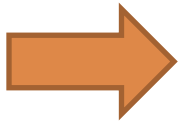
data produced in astronomy data in one night;
US Library of Congress has 1000 TB data;
search log of Bing is 20 TB per day

PB = 10^{15} bytes

Facebook uses more than 1 PB for its use
Google processes 20 PB data per day;
the entire written works of humankind, i

VLDB vs. Big Data

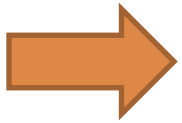
- **Very Large Database（VLDB）**
  - MB, 结构数据
  - 以数据为对象解决其存储和管理问题



数据工程

Data Engineering

- **Big Data**
  - >PB，非结构数据
  - 以数据为资源解决诸领域问题



数据思维

Data Thinking

More data vs. better algorithms

- **The Netflix challenge ($1M awarded Sep. 22, 09)**
  - A: sophisticated algorithm
  - B: simple algorithm + additional data (IMDB)
  - *B gets much better results*
- **Google PageRank:**
  - Rank pages by additional data: link + anchortext
- **Overture and Google AdWords:**
  - Overture ranks advertisers on their bids
  - Google ranks by bids * CTR
- **More data beats better algorithms**

The Big Data Makes the Scientific Method Obsolete

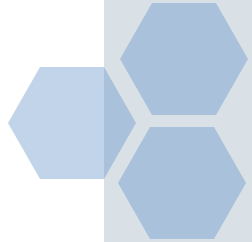
- by Chris Anderson, WIRED

"All models are wrong, but some are useful." --- George Box

"All models are wrong, and increasingly you can succeed without them."
--- Peter Norvig

Can you succeed without models?

- **PageRank: Rank pages without knowing why one page is better than another**
- **Statistical NLP: translate languages without knowing the languages**
- **Spell Checker: you can write a reasonably good one in 24 lines of code using big data**



Model

“All models are wrong, and increasingly you can suc Norvig

“数据思维”

nore
s that
kinds of

- In the era of big data, more isn't just more. More is different.

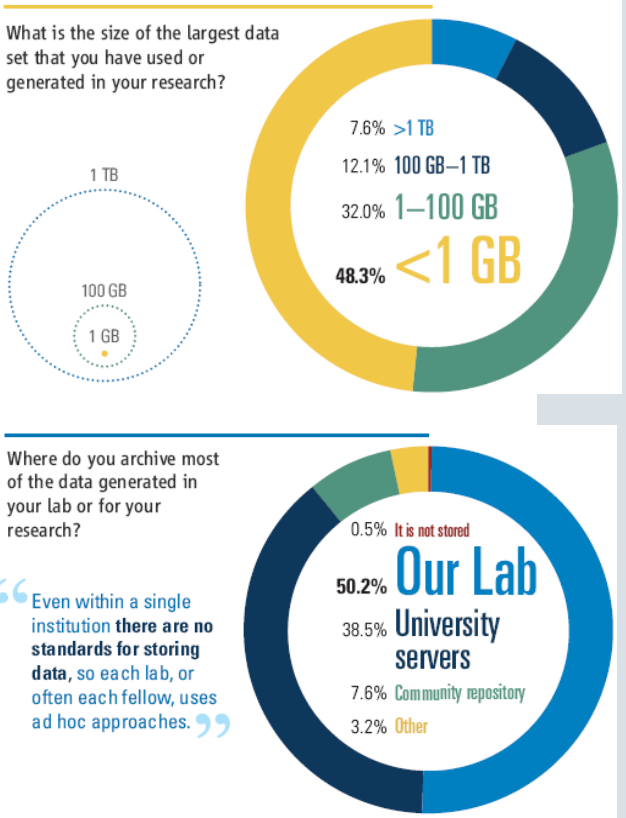
大数据管理面临的机遇与挑战

- 对许多学科而言，海量数据意味着更严峻的挑战；
- 若能更好地组织和使用这些数据，会有助我们将巨大机遇变为现实。

Science：数据处理专题

数据处理中的机遇与挑战

- **针对海量海量数据管理的挑战性问题，《科学》该刊的1700位审稿人和评论家进行了调查。**
  - 大约20%人经常使用或分析的数据集超过100GB；
  - 7%的人所用的数据集超过1TB；
  - 约有50%人仅在他们的实验室中存储数据。
  - **这些调查充分说明了人们在抱怨由于缺乏通用海量数据管理方法，而阻碍了数据的存储、使用和有效访问。**

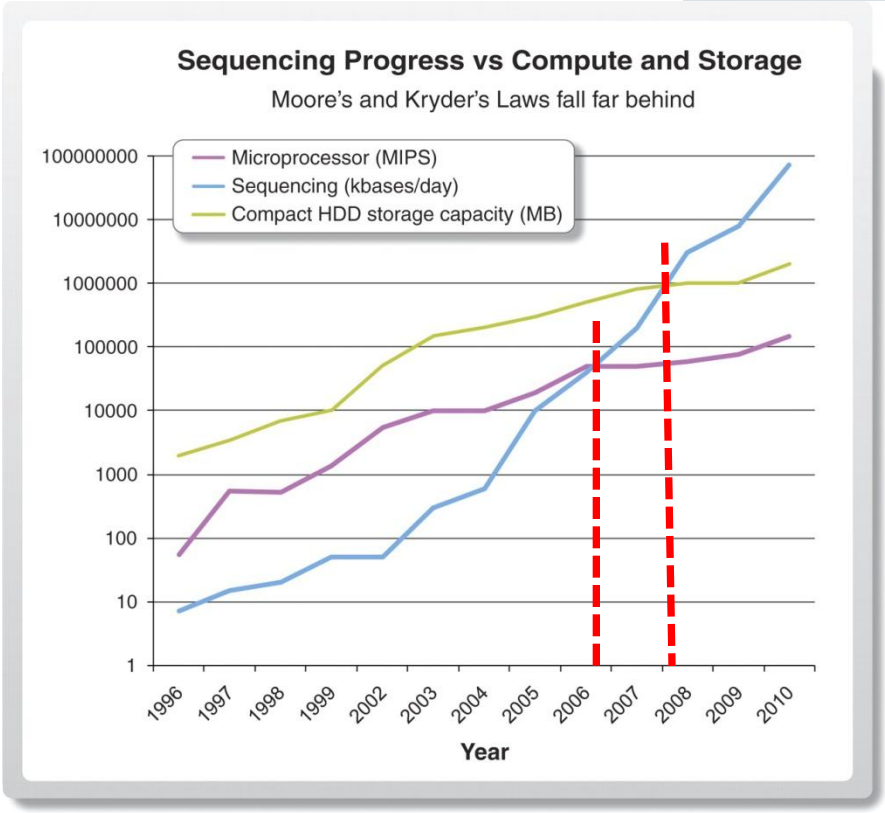


特定领域中的大数据

- **这些特定领域主要包括：**
  - 基因组学中的海量数据；
  - 神经科学领域中的海量数据；
  - 海量数据的可视化技术；
  - 社会科学领域中的数据；
  - 气候数据；
  - 海量生态数据；

基因组学中的海量数

- 仅在一个基因序列定序器的数据增长速度：
  - 10M/day-->40G/day
- 全球有20多个基因序列研究中心，每个中心设置多于10个基因序列定序器。
- 从右图可知，数据增长速度早已超越摩尔定律

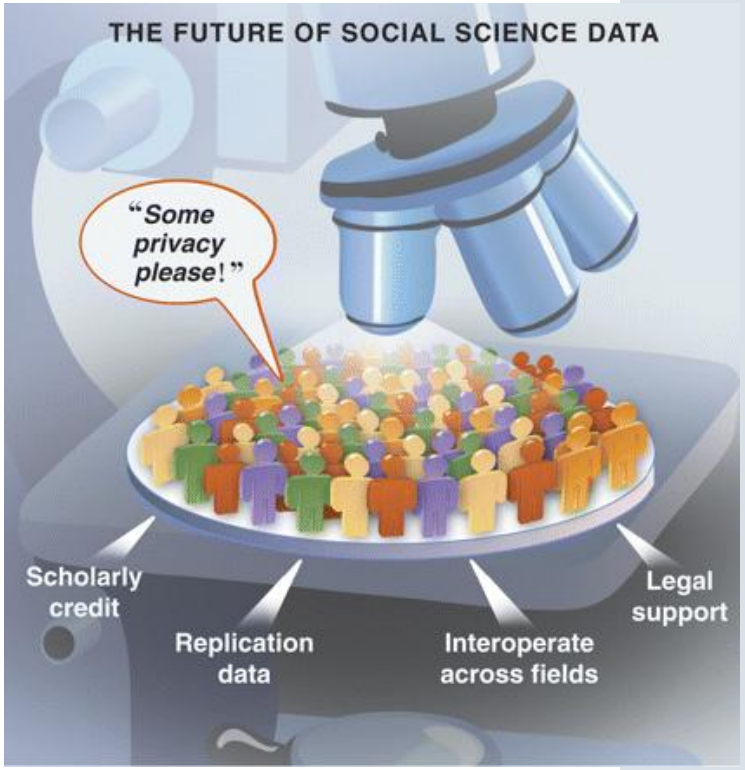


神经科学领域中的海量数据

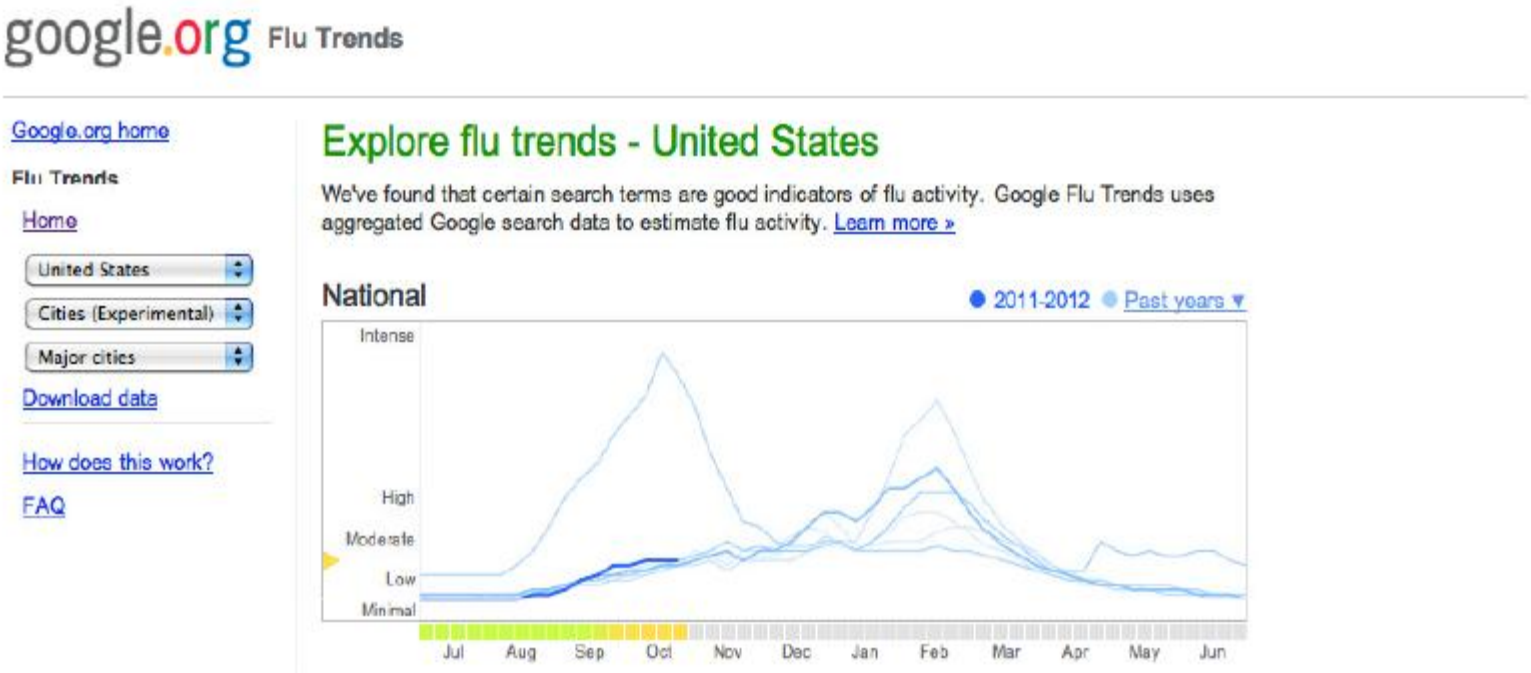
- 人的大脑大约有800亿个神经元，150万亿个突触。
- 神经科学信息框架涵盖了所有神经科学的资源，从而促进了已有知识和各种类型数据库的集成。
- 如何对这些海量数据进行组织和时空多维挖掘，存在很大的挑战。

社会科学领域中的数据

- 在社会科学领域中，同样存在着海量的信息化社会科学数据。
- 未来的社会科学数据包括信用卡交易、拷贝数据、电子医疗数据等。
- 这类数据对统计方法、隐私保护方法以及伦理学的研究提出了严峻的挑战。



Some Applications

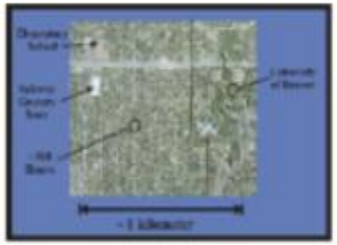


http://www.google.org/flutrends/

Some Applications

Google

A Sampling of Other Use Cases



- Disease Early Warning: Remote surveillance of disease and prediction of epidemics.
- Population Census: Supplements traditional census and mapping in developing regions.
- Humanitarian Crisis Mapping: Can detect and monitor a growing range of crisis types.
- Water Resources: Monitor water quality and availability and alleviate water shortage problems.
- Food Security: Famine early warning, rainfall and water requirements estimations, agr production estimates and irrigation and fertilizer supply & demand.
- Global Education Programs

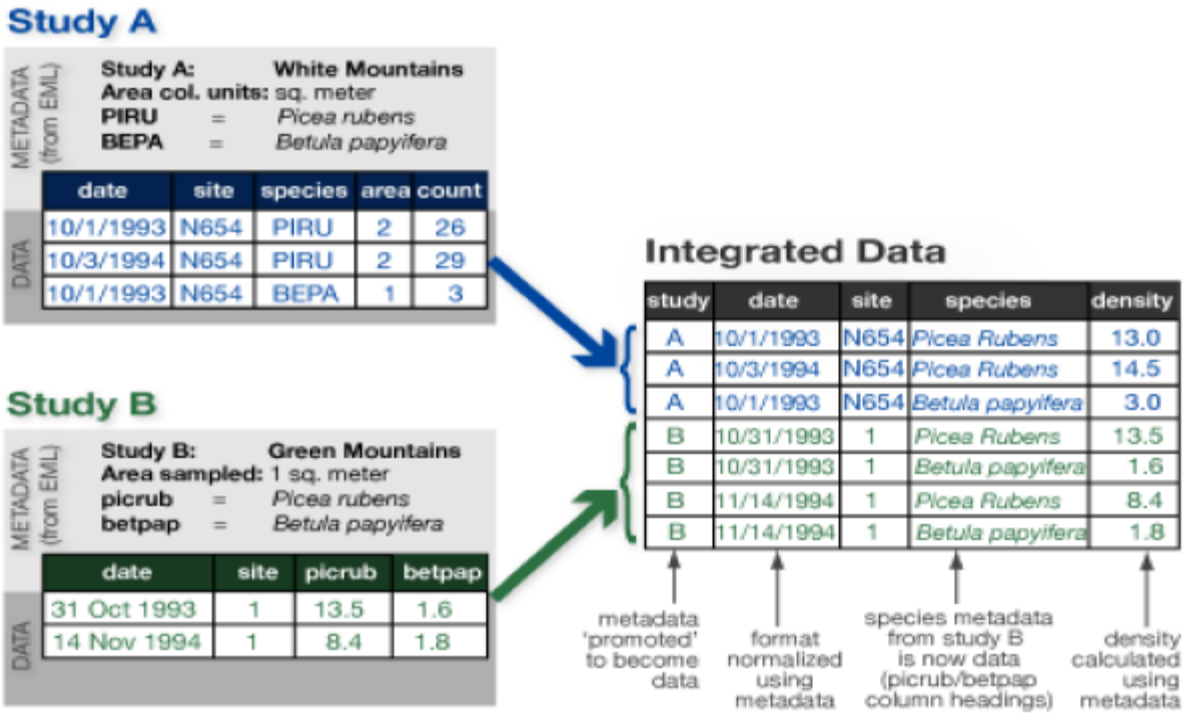
大数据管理框架

- **大数据特征**
  - 多源异构：存在较大的异质性
  - 分布广泛：分布在各个区域
  - 动态增长：增长快，更新快
  - 数据-模式：先有数据后有模式

如何高效管理海量数据？

例：多源异构

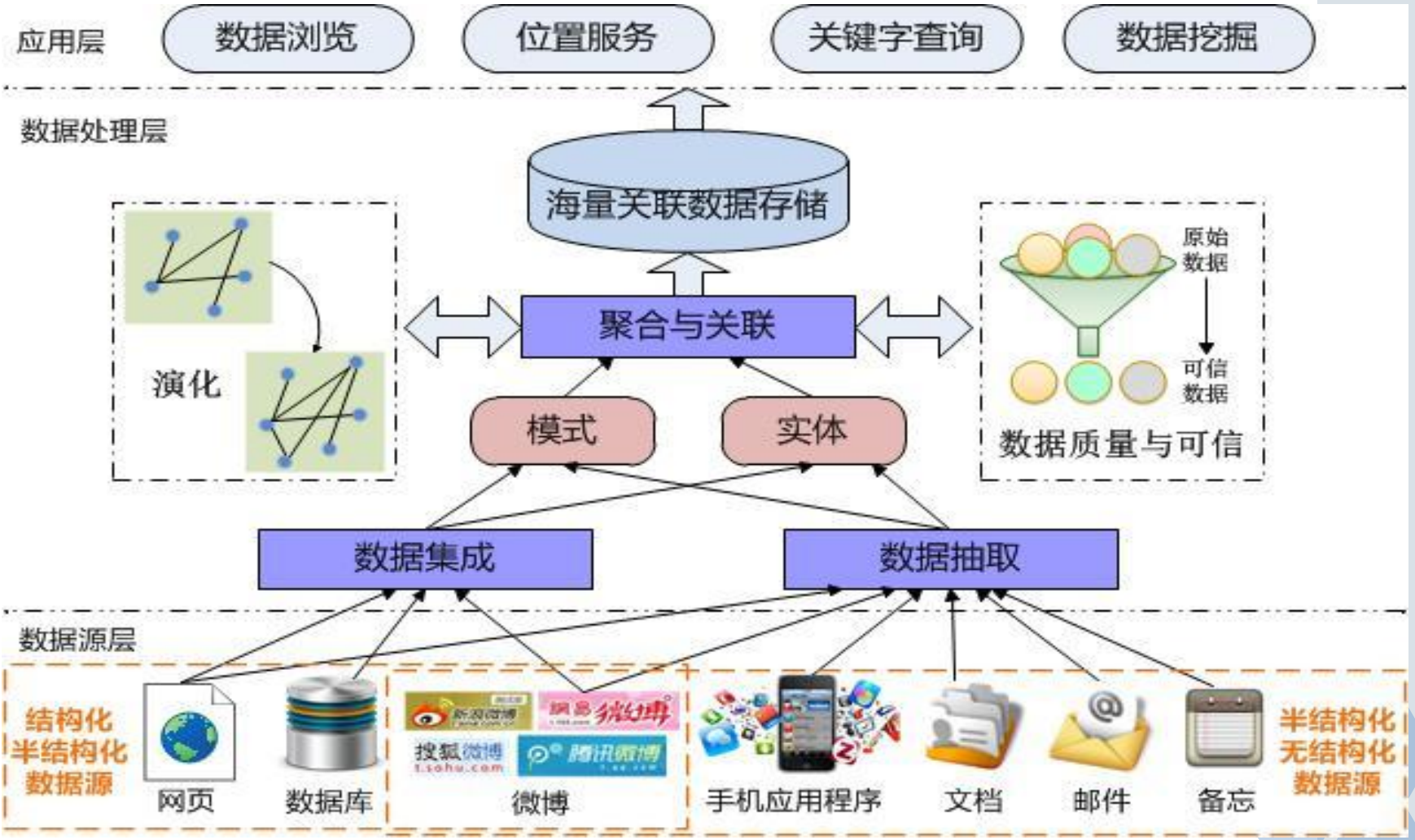
- Syntax
  - (format)
- Schema
  - (model)
- Semantics
  - (meaning)



Google Scale in Hardware and Storage

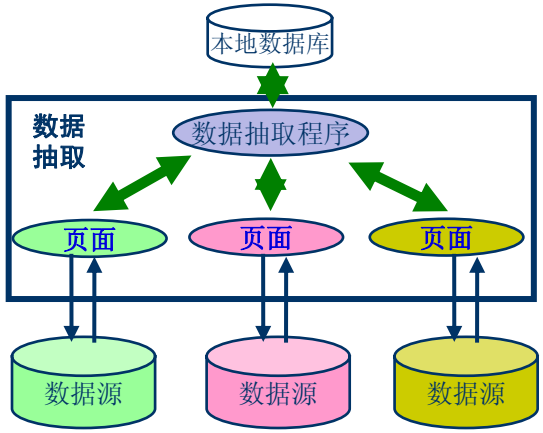
- Giga 10^9, Tera 10^{12}, Peta 10^{15}, Exa 10^{18}, Zetta10^{21}
- Publicized: Bigtable of 70 petabytes, 10M ops/sec.
- Some representative numbers:
  - Storage: 10^{18} -> $10^{20\text{-}21}$
  - Users: 10^9 -> 10^{10}
  - Network: 10^{20}, now, ->$10^{21/yr}$ (32 KB/sec. for 1B people)
- Warehouse computing possibilities

大数据管理框架


应用层
数据浏览
位置服务
关键字查询
数据挖掘
数据处理层
海量关联数据存储
聚合与关联
演化
模式
实体
原始数据
可信数据
数据质量与可信
数据集成
数据抽取
数据源层
结构化
半结构化
数据源
网页
数据库
微博
手机应用程序
文档
邮件
备忘
半结构化
无结构化
数据源

系统成果

- 基于数据抽取的数据集成方法



- 学术领域：计算机中文文献集成系统C-DBLP

集成近6万作者的信息，日访问量6000次，半年累计访问超过100万次

- 求职领域：工作通数据集成系统

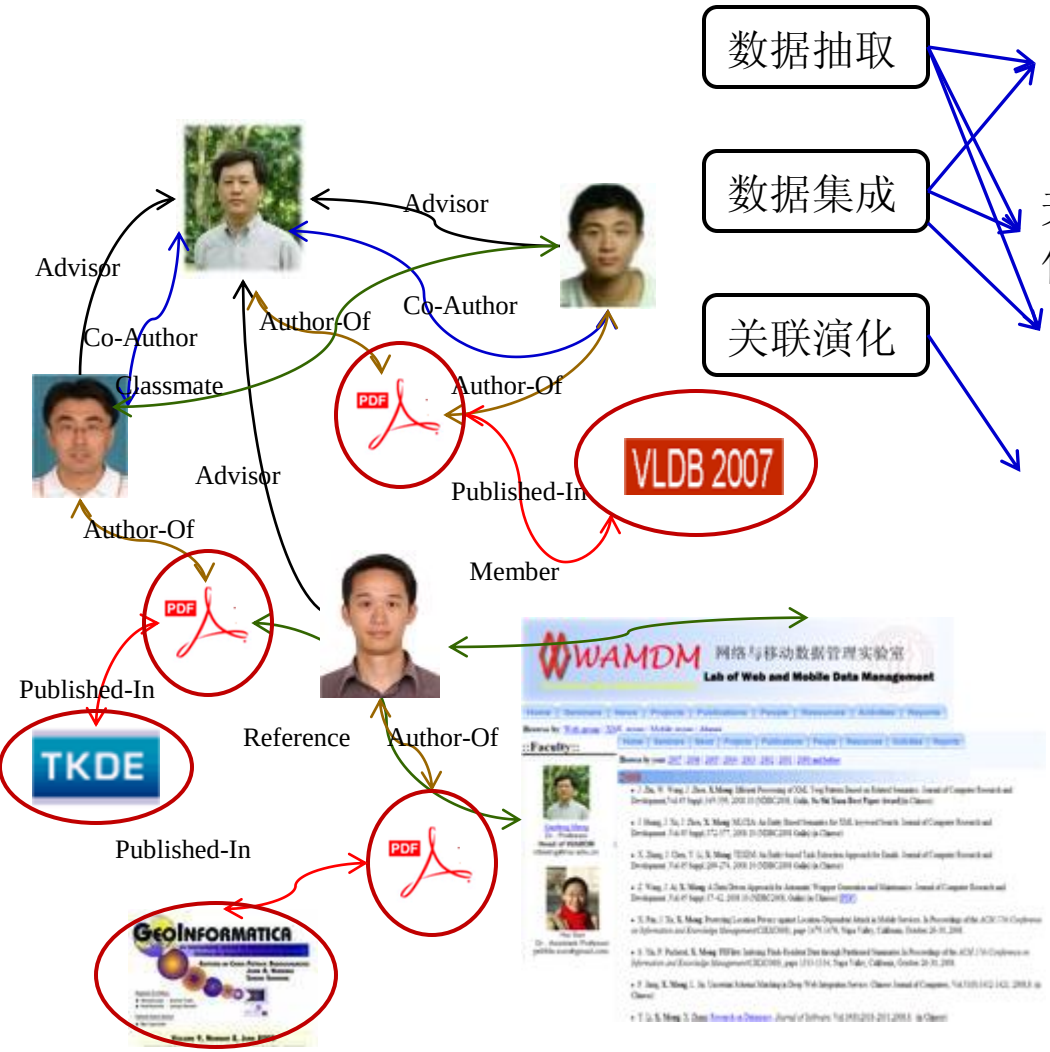
集成近100个数据源，数据量超过300万条

- 新闻领域：舆情监控系统

集成5个代表性论坛，1000多新闻媒体，十万个网上博客

计算机中文文献集成系统C-DBLP(1

- 文献集成系统SearchScholar



数据抽取

数据集成

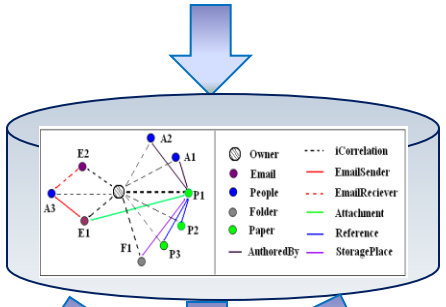
关联演化

实体:
作者, 论文, 期刊, 会议, 研究机构, …

关联:
作者关系, 论文发表关系,合作者关系,

隶属关系, 导师关系， 参考文献关系...

关联发现、删除、更新



浏览
基于任务

查询
多种形式

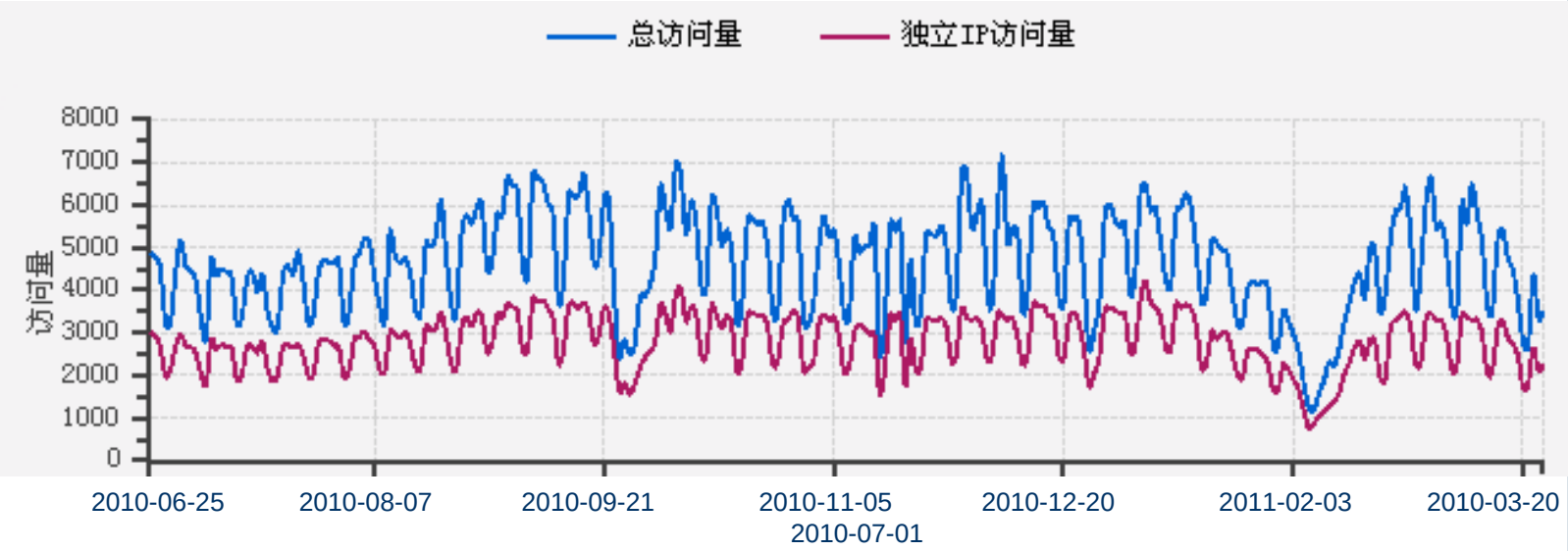
分析
丰富多样

计算机中文文献集成系统C-DBLP(2)

输入学者名字，可以查询出其发表的文章，参加的科研项目、研究兴趣、科研成果分析、合作者情况等大量信息。

计算机中文文献集成系统C-DBLP(4)

- 文献集成系统SearchScholar
  - **12个期刊(1960.01-2011.02), 1 个会议(NDBC 2000-2011), 8万多篇论文, 6.8万多个作者**
  - **日访问量超过5000次，累计访问超过350万次**



成果意义

- 建立了一种将数据结构化管理的途径，为解决特定领域的大数据集成问题奠定了基础
- 进而为大数据管理提供一种新的解决思路

日本海啸

数据海洋中的海啸

- 数据海啸一：Web网面海量数据
- 数据海啸二：微博海量数据
- 数据海啸三：移动App海量数据

数据海啸：微博海量数据


新浪微博
t.sina.com.cn
腾讯微博
t.qq.com
Follow5
12seconds
brightkite
hellotxt
PLURK
tumblr.
friendfeed
人民微博
t.people.com.cn
同学网
tongxue.com
soup
yammer
jaiku
plerb
和讯 财经微博
youare
shout'em
凤凰网 微博
ifeng.com
Pownce
Second Brain
renjian.com
Rejaw
numpa
Yonkly
網易微博
t.163.com
搜狐微博
t.sohu.com
meemi
utterli
koornk
CNFOL.COM 中金在线 微博
microblogr
Bloggino
hictu!
139社区
kwippy
beemood
ZOL微博
T.ZOL.COM.CN
9911
www.9911.com
鲜果联播
体坛微博
t.titan24.com

社会的数字化与数字的社会化

- **社会的数字化：数据足迹（data print）**
  - 在数字化时代，各色人等有意无意留下的数据足迹越来越丰富
  - 数据足迹是有社会意义（social meaning）的，蕴含着社会结构
- **数字的社会化：**
  - 数据足迹及其结构本身就是社会结构和过程的一个环节，不断塑造着新的社会秩序和关系

数据思维：计算社会科学

- 一切社会解释、监控、预测与规划都离不开对数据足迹的收集、整理和分析
- 计算社会科学方法：
  - 基于特定社会需要，在特定社会理论指导下，收集、整理和分析数据足迹，以便进行社会解释、监控、预测与规划的过程和活动

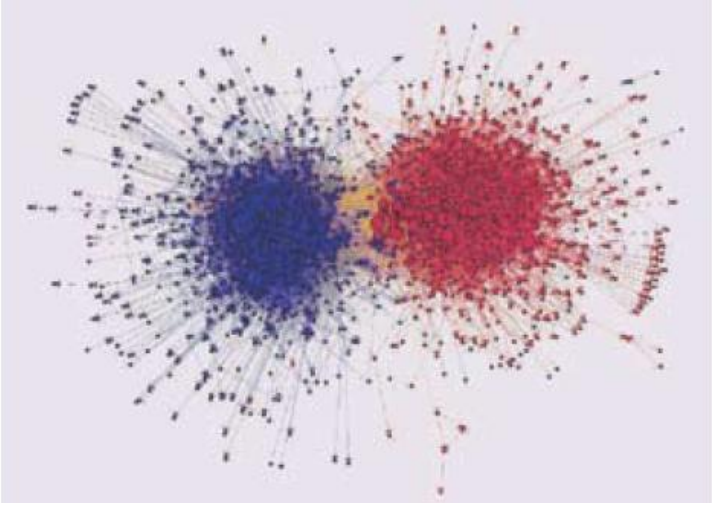


图1 来自博客（blogosphere）空间的数据：上图是一个政治博客群的链接结构（从2004年开始），红色节点代表保守派，蓝色节点代表自由派。橙色链接从自由派博客指向保守派，紫色链接反之。每个博客节点的大小反映了指向该博客的其他博客的数量。复制自文献[8]。

总结和展望

总结

- 积累数据财富应该成为我们的习惯，它或许也会成为国力的标志
- 研究“数据思维”的方法，或许会有意义，是下一个十年我们面临的机遇期

谢　谢！

未来每18 个月产生的数据量
等于有史以来的数据量之和

--Jim Gray
1998图灵奖获奖演说