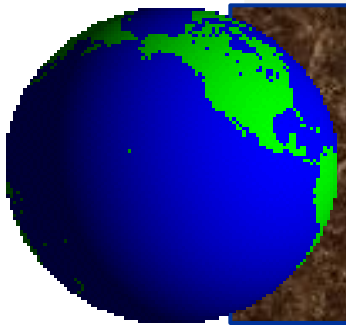
INSTITUTE OF COMPUTING TECHNOLOGY

网络大数据应用提出的挑战性问题



李国杰
中国科学院计算技术研究所
2012.5.22，香山科学会议

网络大数据的广泛应用

信息社会的发展变化

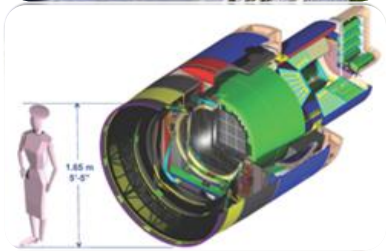
- 60年前, 数字计算机使得信息可读；20年前，Internet使得信息可获得；10年前，搜索引擎爬虫将互联网变成一个数据库；现在，Google 及类似公司处理海量语料库如同一个人类社会实验室。
- 数据量的指数级增长不但改变了人们的生活方式、企业的运营模式，而且改变了科研范式。
- 过去几十年，我们经常讲发展信息科学技术和产业，但主要的工作是电子化和数字化。现在，数据为王的大数据时代已经到来，我们需要完成观念上的重大转变：将关注的重点真正落在数据（信息）上，计算机行业要转变为真正的信息行业。计算机要从追求计算速度转变为大数据处理能力，软件要从编程为主转变为数据优先。

什么是大数据？

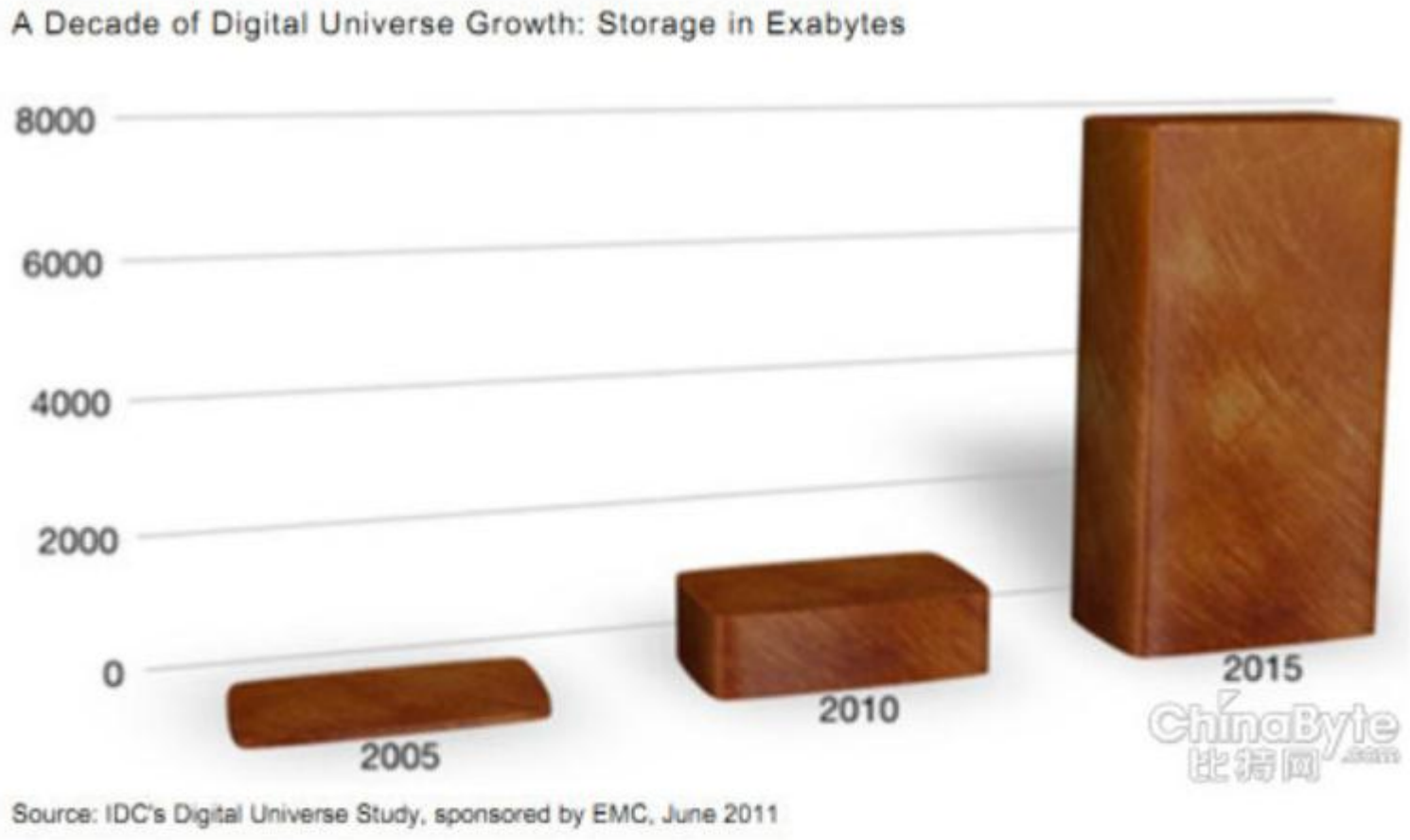
- 大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合（维基百科定义）
- 大数据 = “海量数据” + “复杂类型的数据”
- 大数据的特性（Volume，Variety，Velocity）
  - 数据量大：PB、TB、EB、ZB级别的数据量
  - 种类多：包括文档、视频、图片、音频、数据库、层次状数据等
  - 速度快：数据生产速度很快；对数据处理和I/O速度很快
- 涉及多个领域
  - 包括天文、气象、基因、医学、经济、物理、互联网等
  - 本次会议重点讨论与人类社会活动有关的网络数据

目前大数据的规模

- IDC公司发布的数字宇宙研究报告称：全球信息总量每两年就会增长一倍，2011年全球被创建和被复制的数据总量为1.8ZB (10^{21}),其中 75%来自于个人。
- IDC认为，到下一个十年(2020年)，全球所有IT部门拥有服务器的总量将会比现在多出10倍，所管理的数据将会比现在多出50倍。预计到2020年，全球将总共拥有35ZB的数据量
- 2011年企业创造、采集、管理和储存信息的成本已经下降到2005年的1/6，而同期企业关于数据的总投资自2005年以来却反而上升了50%。
- 数据成本的下降助推了数据量的增长，而新的数据源和数据采集技术的出现则大大增加了未来数据的类型，数据类型的增加导致现有数据空间维度增加，极大地增加了未来大数据的复杂度。



大数据总量增长态势
（数据摩尔定律：两年翻一番）



数字宇宙研究报告：存储这十年

大数据公司的现状

- Google 公司通过大规模集群和MapReduce 软件,每天处理超过20PB 的数据,每个月处理的数据量超过400PB。
- Gartner公司对未来五年的预测：到2015年，85%的世界五百强企业如果不采取大数据的策略将失去竞争力！
- 淘宝网：有3.7亿会员，在线商品8.8亿，每天交易数千万，产生约20TB数据。
- Yahoo!的数据量：Hadoop云计算平台有34个集群，超过3万台机器，总存储容量超过100PB。（按照欧盟的 规定，不能存储超过一年的用户数据）。

海量数据创造的巨大价值

- Data is the next “Intel Inside”.
- The future belongs to the companies and people that turn data into products.

----著名出版公司O‘Reilly的创始人Tim O‘Reilly

美国政府启动“Big Data”计划

- 2012年3月29日，美国政府启动“Big Data Research and Development Initiative”计划，6个部门拨款2亿美元。

  - transform our ability to use Big Data for scientific discovery, environmental and biomedical research, education, and national security.
  - prepare the next generation of data scientists and engineers
  - seeking a 100-fold increase in the ability of analysts to extract information from texts in any language

- 这是一个标致性事件，说明继集成电路和互联网之后，大数据已成为信息科技关注的重点。

网络大数据的特点

（1）多源异构：描述同一主题的数据由不同的用户、不同的网站产生。网络数据有多种不同的呈现形式，如音视频、图片、文本等，导致网络数据格式上的异构性。

（2）交互性：不同于测量和传感获取的大规模科学数据，微博等社交网络兴起导至大量网络数据具有很强的交互性。（从交易到交互）

（3）时效性：在网络平台上，每时每刻都有大量新的网络数据发布，网络信息内容不断变化，导致了信息传播的时序相关性。

（4）社会性：网络上用户既可以根据需要发布信息，也可以根据自己的喜好回复或转发信息，因而网络数据成了对社会状态的直接反映。

（5）突发性：有些信息在传播过程中会在短时间内引起大量新的网络数据与信息的产生，并使相关的网络用户形成网络群体，体现出网络大数据以及网络群体的突发特性。

（6）高噪声：网络数据来自于众多不同的网络用户，具有很高的噪声。

从企业智能（BI）到个人消费智能

- 建立数据仓库的主要目的是为大型企业的业务人员提供智能。现在，一种新型消费者正在兴起，许多人热衷于自己动手使用技术工具，利用数据来制定个人决策。移动设备的普及和消费行为的变革催生了市场对消费智能的需求，消费者希望直接访问数据，制定相应决策。
- 网络大数据的处理不仅仅局限于数据中心和大企业，中小企业和个人消费者都可能需要进行大数据处理。因此，在简易的设备和系统上处理大数据成为值得关注的科学技术问题，大数据分析算法和软件的易用性也成为新的需求。Mapreduce 和Hadoop 的广泛流行值得深思。

对本次香山科学会议的期望

- 本次会议将对海量网络数据研究的背景、需求和现状进行全方位深入而广泛地讨论，尽可能形成对网络数据处理潜在问题、发展方向和面临挑战的共识，为促进形成网络数据科学（一门新型交叉学科）奠定基础。

- 探讨网络数据科学的学科基础及理论边界，讨论其独立成为一门新型学科的可行性；

- 从社会科学、心理学、经济学、信息科学等学科领域探讨网络数据的产生、扩散、涌现及其影响力评价的基本机制，从社会、经济和技术层面提出网络数据涌现规律与价值的度量手段；

- 探讨海量网络数据存储、管理、计算的系统体系架构，分析适用于海量网络数据处理的新模型、新型计算范式以及网络化算法设计与算法优化的基础理论。

网络大数据带来的技术挑战

重点是应对大数据带来的技术挑战

- 美国政府的大数据计划和Google 等大公司目前最重视的都是数据工程而不是数据科学，主要考虑大数据分析算法和系统的效率。我们也应把主要精力放在应对大数据工程的技术挑战上。
- 面对大数据应用，技术走在科学前面，技术上解决不了的问题就构成科学挑战问题。本次会议的重点不是讨论数据挖掘等技术问题，而是讨论大数据带来的科学挑战。
- 企业中80%的数据是非结构化或半结构化数据，（只有20%的数据是结构化的）。当今世界结构化数据增长率大概是32%，而非结构化数据增长则是63%，至2012年，非结构化数据占有比例将达到互联网整个数据量的75%以上。大数据的技术挑战主要是指非结构化数据。

美国政府“Big Data计划”部分内容

- **国防部高级研究计划局(DARPA)**
  - 多尺度异常检测项目解决大规模数据集的异常检测和特征化。
  - 网络内部威胁计划通过分析图像和非图像的传感器信息和其他来源的信息，进行网络威胁的自动识别和非常规的战争行为。
  - Machine Reading 项目旨在实现人工智能的应用和发展学习系统，对自然文本进行知识插入。
  - Mind‘s Eye 项目旨在建立一个更完整的视觉智能。
- **能源部 (DOE)**
  - 从庞大的科学数据集中提取信息，发现其主要特征，并理解其间的关系。研究领域包括机器学习，数据流的实时分析，非线性随机的数据缩减技术和可扩展的统计分析技术。
  - 生物和环境研究计划，大气辐射测量气候研究设施
  - 系统生物学知识库对微生物，植物和环境条件下的生物群落功能的数据驱动的预测。

美国政府“Big Data计划”部分内容

- **国家人文基金会(NEH)**
  - 分析大数据的变化对人文社会科学的影响，如数字化的书籍和报纸数据库，从网络搜索，传感器和手机记录交易数据。
- **美国国家科学基金会(NSF)**
  - 推进大数据科学与工程的核心技术，旨在促进从大量、多样、分散、异构的数据集中提取有用信息的核心技术。
  - 深入整合算法，机器和人，以解决大数据的研究挑战。
  - 开发一种以统一的理论框架为原则的统计方法，可伸缩的网络模型算法，以区别适合随机性网络的方法
  - 形成一个独特的学科包括数学、统计基础和计算机算法。
  - 开放科学网格(OSG)，使得全世界超过8000名的科学家合作进行发现，包括寻找希格斯玻色子（“上帝粒子”，宇宙中所有物质的质量之源）。

变“大数据”为“小数据”

- 数据无处不在，但许多数据是重复的或者没有价值，未来的任务主要不是获取越来来越多的数据，而是数据的去冗分类、去粗取精，从数据中挖掘知识。
- 数据量大到一定程度，数据压缩就必不可少。去重、压缩和归档技术是大数据处理技术中不可或缺的组成部分。
- “大数据”有简单和复杂之分。个体间联系很少，只是个体数量庞大的“大数据”问题并不难解决；组合爆炸的困难产生于个体之间的联系，社会网络的复杂性来源于社会联系。“小世界”也会产生“大数据”。
- 几百年来，科学研究一直在做“从薄到厚”的事情，把“小数据”变成“大数据”，现在要做的事情是“从厚到薄”，要把大数据变成小数据。

大数据分析的误区

- **样本缺乏代表性**

  统计结论依赖于样本的代表性。要确保样本数据代表研究总体，否则分析结论就缺乏坚实的基础。

- **事物是变化的**

  不能只进行一次分析，要持续验证之前的结论。

- **理解数据方式有多种方式**

  一组数据可以提供多种类型的信息。需要找到不同的解释方式，并加以分析。

- **错误和偏差**

  不能只使用一种方法，要用事实来检验假设是否奏效。

数据管理的挑战性问题

- 大数据分析是否有价值，关键在于数据本身的的“质量”，数据量的多少不一定是决定因数。要在获得“好”的数据上下功夫。
- 数据敏感性分级问题：不同数据产生的价值是不同的，不同时期产生的价值也不一样，必须要定义哪些数据有价值，哪些没有价值，需要定义价值的时间期限。
- 热点数据问题：热点数据在不断变化。根据热点的程度和时间调整访问权限。
- 数据质量管理：数据保真度、数据的相关性、数据的有效性、数据的有效期限等。

需要高扩展高可用的数据分析技术

- 传统的关系数据库无法胜任大数据分析的任务，因为并行关系数据库系统的出发点是追求高度的数据一致性和容错性。根据CAP理论(Consistency, Availability, tolerance to network Partitions),在分布式系统中,一致性、可用性、分区容错性三者不可兼得, 因而并行关系数据库必然无法获得较强的扩展性和良好的系统可用性。系统的扩展性是大数据分析最重要的需求，必须寻找高扩展性的数据分析技术。
- 以MapReduce 和Hadoop为代表的非关系数据分析技术，以其适合大规模并行处理、简单易用等突出优势,在互联网信息搜索和其他大数据分析领域取得重大进展，已成为目前大数据分析的主流技术。
- MapReduce 和Hadoop在一些应用的性能上还比不过关系数据库，还需要研究开发更有效、更实用的大数据分析和管理技术，需要发展像关系数据库这样的理论来指导海量非结构化Web数据的处理。

必须研究数据表示方法

- 我们目前表示数据的方法，不一定能直观地展现出数据本身的意义。要想有效利用数据并挖掘其中的知识，必须找到最合适的数据表示方法。
- 我们在一种不合适的数据表示中寻找大数据的固定模式、因果关系和关联时，可能已落入固有的偏见之中。
- 数据表示方法和最初的数据填写者有着密切关系。如果原始数据有必要的标识，就会大大减轻事后数据识别和分类的困难。但为标识数据给用户增添麻烦往往得不到用户认可。研究既有效又简易的数据表示方法是处理网络大数据必须解决的技术难题之一。

数据转换和统一数据编码

- 网上数据尤其是流媒体数据的泛滥与数据格式太多有关。每个大企业都有自己不同数据格式，用户为了摆脱大企业的“绑定”，需要不断地做格式转换。格式繁多也给海量数据分析增加了许多工作量。
- 大数据面临的一个重要问题是个人、企业和跨部门的政府机构的各种数据和信息能否方便的融合。
- 如同人类有许多种自然语言一样，作为Cyberspace中唯一客观存在的数据难免有多种格式。但为了扫清网络大数据处理的障碍，应研究推广不与平台绑定的数据格式。
- 图像、语音、文字都有不同的数据格式，在大数据存储和处理中这三者的融合已成为一种趋势，有必要研究囊括各种数据的统一格式，简化大数据处理。

大数据应用要求整个IT架构进行革命性的重构

- 现有的数据中心技术很难满足大数据的需求，需要考虑对整个IT架构进行革命性的重构。
- 存储能力的增长远远赶不上数据的增长，设计最合理的分层存储架构已成为信息系统的关键。
- 数据的移动已成为信息系统最大的开销，信息系统需要从数据围着处理器转改变为处理能力围着数据转。
- 提高可扩展性成为信息系统最本质需求，并发执行（同时执行的线程）的规模要从现在的千万量级提高10亿级以上
- 大数据已成为联系人类社会、物理世界和赛博空间（Cyberspace）的纽带，需要构建融合人、机、物三元世界的统一的信息系统。

从高性能计算机到高通量计算机

- 大数据处理不同于科学计算的超级计算机，不是追求尽量缩短单个任务的计算时间，而是在允许的时间范围内处理尽可能多的任务（数据或线程），体系结构需要根本性的变革。
- 下一代数据中心的服务器
  - 基于数据中心，提供高并发数据处理服务的高扩展、低成本的大型计算机软硬件系统
- 特征
  - 尽量提高并发线程数
  - 尽量提高每瓦线程数
  - 适当控制每线程的功率

“Little’s Law”： $\lambda = L / W$

New observations:

$\lambda = L \times (E/W) \times (1/E)$

Throughput =

Volume × Watts per thread

× Threads per Joule

数据处理器（DPU）

- 海量并发的处理单元结构、支持新的执行模型，支持海量的并发线程（单芯片千线程）。
  - 需要保存海量线程的状态
  - 从体系结构考虑减少每个线程的状态
  - 研究有效的资源共享机制
- 设计目标是追求数据处理能力
  - 尽可能地提高各个部件的利用率
  - 减少浮点部件和缓存，提高内存数据存取速度
  - 不优先考虑单个数据处理的速度
  - 而是考虑单位时间内数据处理的数量

面向大数据的存储和系统网络

- 低成本EB级存储
  - 基于闪存与磁盘的混合存储系统
  - 提高存储在系统的地位
  - 缩短存储与处理器之间的距离
  - 扩大存储与处理器之间的通路
- 高带宽的系统网络
  - 研究更简化的通信协议
  - 简化数据中心内部的连接
  - 提高系统网络与处理器之间耦合度
  - 提高系统网络与存储之间耦合度

数据中心网络

- 研究适合大规模数据中心的网络结构
  - 支持节点动态加入、迁移和退出
  - 故障的自动检测、处理和屏蔽
  - 有效利用网络传输带宽
  - 有效降低网络系统的功耗
- 关键技术
  - 大规模全系统DCN模拟器
  - 在协议、网卡、交换等层次增加对应用的直接支持
  - 新型网卡:协议精简和加速, 降低端到端延迟
  - 新型交换/路由器：数据流动感知
  - 全局资源控制器：资源调度、虚拟化支持，故障管理、功耗优化

网络大数据提出的科学挑战

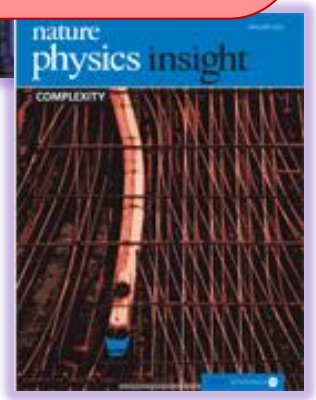
国际学术界高度重视大数据研究

- 2008年，Nature出版专刊“Big Data”，
  - 从互联网技术、互联网经济学、超级计算、环境科学、生物医药等多个方面介绍了大数据所带来的科学与技术挑战

尽管学术界已注意到大数据带来的机遇和挑战，但对大数据提出的科学挑战问题还没有形成共识。

巨大推动作用

- 2012年，Nature Physics上出版专刊“Complexity”
  - 特别指出大数据为科学研究，特别是复杂性科学的研究提供了史无前例的机遇



本报告要讨论的科学挑战问题

- “网络数据科学” 研究的对象是什么？
- 数据界（Data Nature)有共性问题吗？
- “网络数据科学”就是“数据挖掘”吗？
- “网络数据科学”就是“统计学”吗？
- “网络数据科学”与人工智能是什么关系？
- 应选择何种社会问题做大数据研究？
- 为什么人脑学习不需要大数据？
- 所谓“第四科研范式”的本质是什么？
- “大数据研究”真的不需要假设和模型吗？
- 研究“网络数据”还是研究“数据关系网络”？

“网络数据科学”研究的对象是什么？

- 计算机科学的关于算法的科学，数据科学是关于数据的科学。找算法是有目标的研究，数据科学的问题是没有目标(数据性质不是唯一的)。
- 人们常比喻数据科学是“大海捞针”，但“大海捞针”的前提是事先知道有一枚“针”在海里，数据挖掘或一般的数据分析往往不知道有没有“针”。
- 现在处理big data, 究竟相当60年代水平还是70年水平？

  是关注数据的property（60年代的计算机科学），

  还是关注处理数据的efficiency（70年代计算机科学）？
- 在讨论网络数据科学能不能成为一门新的交叉学科之前，首先要搞清楚“网络数据科学”研究的对象究竟是什么？

----引自在澳门大学与赵伟教授、华云生教授的讨论

从数据的查询和传播看“网络数据科学”

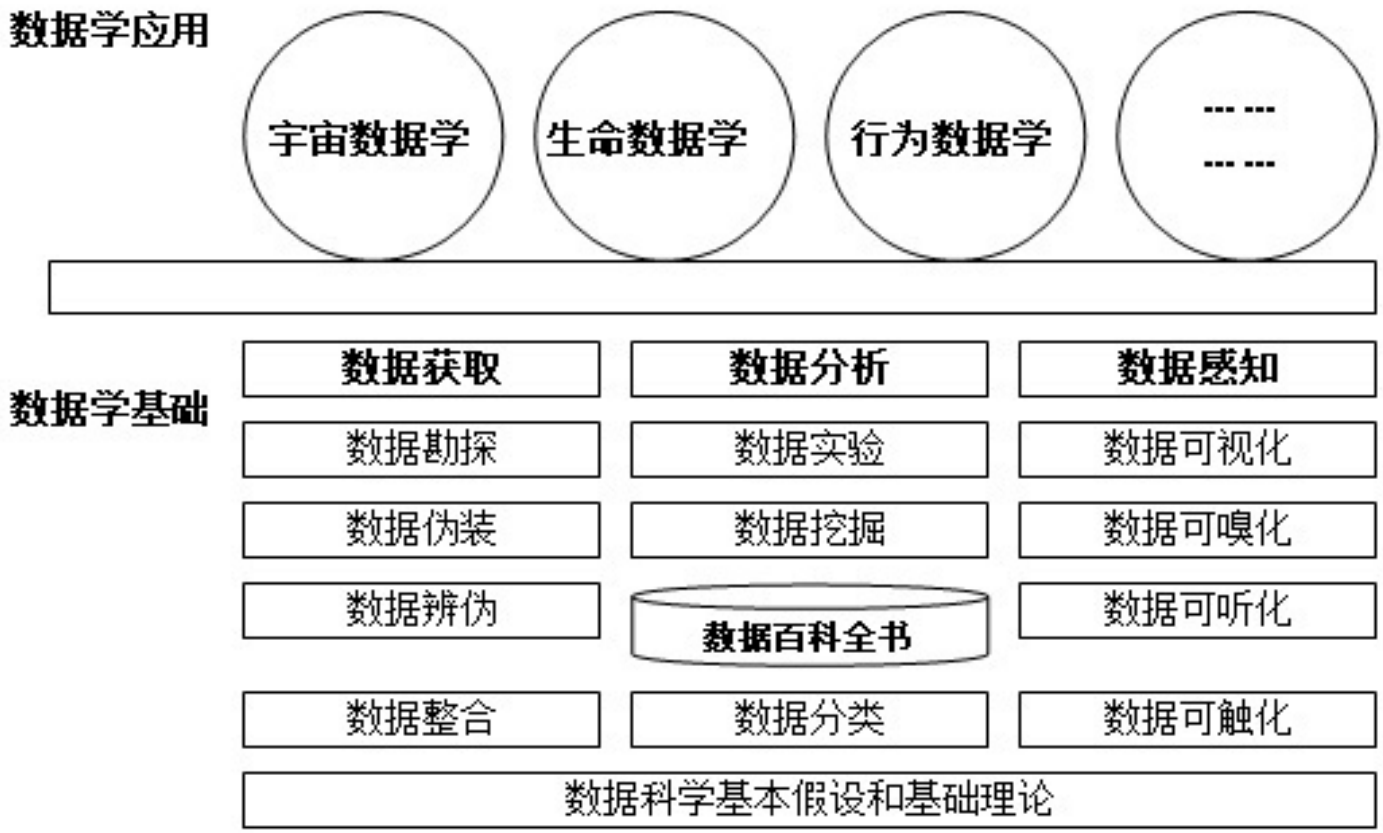
- 复杂网络上的数据（信息）的传播机理、搜索、聚类、同步和控制等应该是网络信息科学的主要研究内容。
- 有些人文进程也可以看成是一种聚类计算的过程
- 与能量的传递类似，数据（信息）的传播也有传导、对流与辐射，广播是一种数据（信息）辐射。
- **从寻找lnN长度的短路径到寻找lnlnN超短长度路径**

（Random scale-free networks are ultrasmall worlds. The average length of the shortest paths in networks of size N scales as lnlnN）

- 模式识别、机器学习、生物信息学等许多问题度可以看成的数据的搜索问题。

“网络数据科学”研究的对象是什么？

当作工程技术的“数据学”



“网络数据科学”研究的对象是什么？

从社会科学角度看“网络数据科学”？

- 网络数据科学应不同于传统的计算机科学，不只是局限于研究高效率的算法。
- 网络数据科学应发现网络数据与信息产生、传播、影响背后的社会学、心理学、经济学和信息科学的机理以及网络信息涌现的内在机制，同时利用这些机理研究互联网对政治、经济、文化、教育、科研的影响。
- 尽管人们已经充分认识并见证了互联网及其承载的网络大数据的影响作用，但这种影响如何形成，其内部机理为何迄今依然是未知数。而网络大数据的存在为综合研究并最终厘清这些机理机制提供了良好的基础。

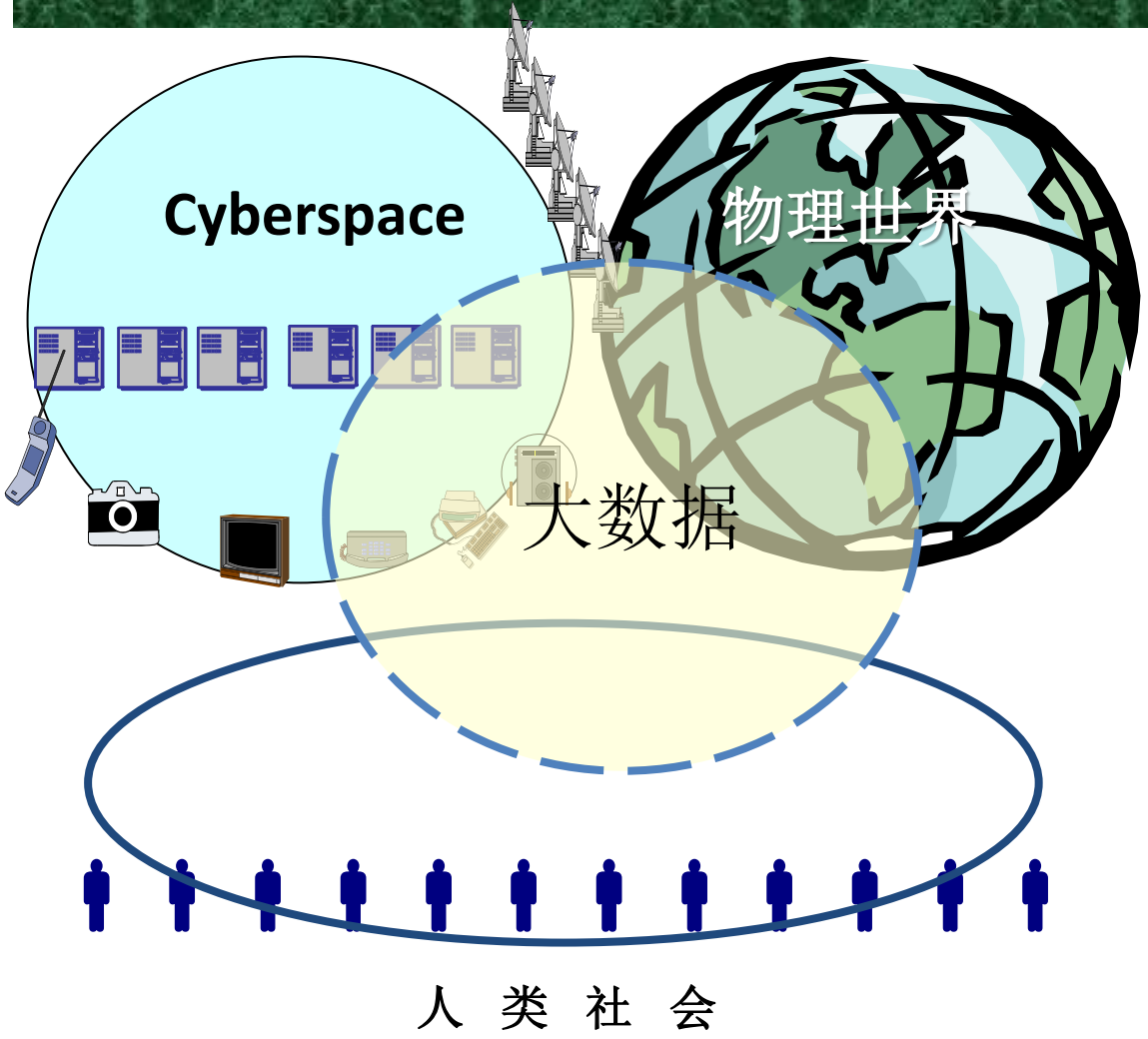
“网络数据科学”研究的对象是什么？

从还原论的对立面看“网络数据科学”

- 过去几个世纪主宰科学研究的方法一直是“还原论”（Reductionism），将世界万物不断分解到最小的单元。作为一种科研范式已经快走到尽头。
- 对单个人、单个基因、单个原子等了解越多，我们对整个社会、整个生命系统、物质系统的理解并没有增加很多，有时可能离理解系统的真谛更远。
- 基于大数据对复杂社会系统进行整体性的研究，也许为研究复杂系统提供了新的途径。从这种意义上看，“网络数据科学”是从整体上研究复杂系统（社会）的一门科学。这门交叉学科如其称为“网络数据科学”，不如称为“数据网络科学”，因为其重点在研究数据背后的网络。

“网络数据科学”研究的对象是什么？

从三元融合世界（人、机、物）看“网络数据科学”



- 云计算、物联网等信息技术的发展使得物理世界、信息世界和人类社会已融合成一个三元世界 （the ternary human-cyber-physical universe)
- 大数据是形成统一的三元世界的纽带
- 数据背后是网络，网络背后是人。研究数据网络实际上是研究人组成的社会网络。

数据界（Data Nature)有共性问题吗？

- 数据科学要把数据当成一个“自然体“（nature）来研究，也就是把computer science 正式划归为”自然科学。
- 脱离各个领域的“物理世界”，作为客观事物间接存在形式的“数据界”究竟有什么共性问题还不清楚。
- 物理世界在Cyberspace中有其数据映像，研究数据界的规律其实就是研究物理世界的规律（还需要在物理世界中测试验证），除去各个领域（天文、物理、生物、社会等）的规律，还有“数据界”共同的规律吗？
- 数据的分类可能是大数据研究的基本科学问题，如同分类在生物学的地位一样，网络数据如何按不同性质分类需要认真研究，分类清楚了，数据标识问题也就解决了，许多数据分析问题也会迎刃而解。

“网络数据科学”就是“数据挖掘”吗？

- 数据挖掘是目前数据分析的热门技术，金融、零售等企业已广泛采用数据挖掘技术分析用户的可信度和购物偏好等。网络数据科学研究肯定要采用数据挖掘技术。
- 目前数据挖掘中急用先研的短期行为较多，多数是为单个问题研究应用技术，尚无统一的理论。
- 传统的数据挖掘技术，在数据维度和规模增大时，所需资源指数级地增加，应对超大数据还需研究新的方法。
- 网络数据科学更强调与社会科学的深度交叉融合，需要揭示找社会科学问题的深层次的机制和规律，只用传统的数据挖掘技术不一定能到达目的。

“网络数据科学”就是“统计学”吗？

- 统计学是收集、分析、表述和解释数据的科学，从字面上看，似乎与大数据的研究范围一致。
- 统计学的目标是从各种类型的数据中提取科学的和有意义的信息，给人后见之明 (hindsight)或预见 (foresight)，但一般不强调对事物的洞察力 (insight)。
- 统计方法强烈依赖与结论有关的应用类型，网络数据常呈现重尾分布，使得方差等标准方法无效，长相依和不平稳性往往超出经典时间序列的基本假设。
- 一种可能的途径是把其他方法和统计方法结合起来，采用多元化的方法来建立处理社会问题的综合性模型。如同只用统计机器翻译方法，翻译质量提高有限度。

网络数据科学与AI是什么关系？

- 传统AI（如机器学习）先通过在较小的数据样本集学习，验证分类、判定等“假设”和“模型”的适合性，再应用推广(Generalization)到更大的数据集。一般NlogN、N^2 级的学习算法复杂度可以接受。
- 面对P级以上的海量网络数据， NlogN、N^2 级的学习算法难以接受，处理大数据需要更简单的人工智能算法和新的问题求解方法。
- 大数据处理和智能处理的核心都是降维（从n维降到1维）。样本数量将随着维数的增加而指数增长就出现维数灾难
- 大数据相当于学习集的样本非常大，应能提高结果的正确性

数据增长对智能应用的推动

- 利用网络上的数据重构三维城市

*摘自“RECONSTRUCTING ROME”，IEEE Computer, June, 2010

应选择何种社会问题做大数据研究？

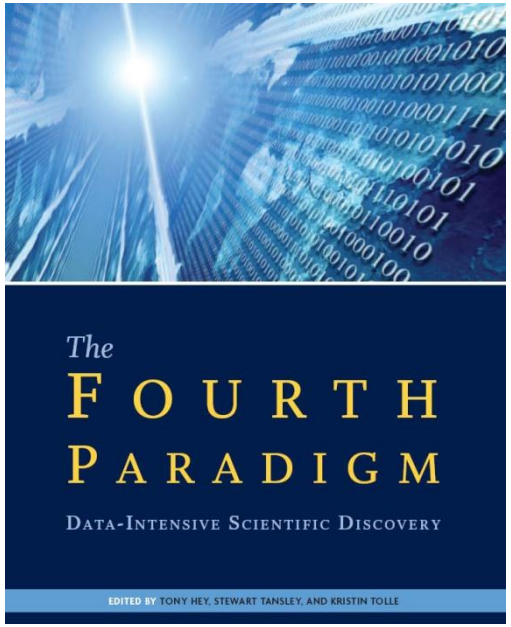
- 科学工程计算可分成三类：（a）基于唯象假设的增量式进步（计算规模大一点，结果就好一些）。这种计算规模再大也不可能变革一个学科。（b）无底洞式的计算—无论多大的计算能力都不可能彻底解决问题（基本的物理本质还不清楚）。（c）变革式计算，只要计算能力足够强大，可以彻底解决以前解决不了的问题。
- 大数据研究可能与科学工程计算有类似的分类，应用大数据方法研究社会问题，应考虑首先选择“预言性数据分析问题”，当数据规模大到一定程度，就可以解决以前解决不了的问题，实现社会科学的“变革式”进步。

为什么人脑学习不需要大数据？

- 网络数据科学不只是研究大数据，也需要从“小数据”的案例获得启发，比如人脑就是小样本学习的典型。
- 2岁小孩看少量图片就能正确使用新的单词“马”或“梳子”，似乎人类的抽象知识不是从经验中学到的而是与生俱来的
  - “How to Grow a Mind: Statistics , Structure, and Abstraction”， 11 March 2011，Vol. 331，SCIENCE
  - Bayesian models can explain how people learn with abstract knowledge
- 我们不能迷信大数据，从少量网络数据中如何高效抽取概念和知识也是值得深入研究的方向。

Emergence of a Fourth Research Paradigm

- Thousand years ago
  - – Experimental Science, Description of natural phenomena
- Last few hundred years
  - – Theoretical Science, Newton's Laws, Maxwell's Equations…
- Last few decades
  - – Computational Science , Simulation of complex phenomena
- Today
  - – Data Intensive Science, Scientists overwhelmed with data set



Tony Hey
Corporate Vice President
Microsoft External Research

所谓“第四科研范式”的本质是什么？

- 网络数据科学的另一个目标是为自然科学和社会科学提供基于大数据分析的科学研究方法，科学界称之为“科研的第四范式”。
- 图灵奖得主、已故科学家吉姆•格雷（Jim Gray）提出了科研的“第四范式”（the fourth paradigm）。吉姆认为：人类需要用强大的新工具去分析、呈现、挖掘和处理科学数据。要解决我们面临的某些最棘手的全球性挑战，“第四范式”可能是唯一具有系统性的方法。
- 所谓科研第四范式与计算机模拟（第三范式）不同，后者建立在对研究领域有深刻理解的基础上（如空气动力学方程用于风洞实验）
- 所谓“第四范式”的本质就是用计算机做统计分析吗？数据量的增加会引起科研上质的改变吗？

美国 Wired杂志主编Chris Anderson 的断言：
——The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

- 统计学家George Box 30年前宣告:
  “**All models are wrong, but some are useful**.”
- Peter Norvig, Google's research director, offered an update to George Box's maxim: "**All models are wrong, and increasingly you can succeed without them**."
- 面对海量数据，有假设、模型、和检验构成的科学方法已经过时
- Petabytes 让我们说：相互关系已经足够（**Correlation is enough**.）我们可以停止寻找模型 . 我们无需假设就可以分析数据。

美国 Wired杂志主编Chris Anderson 的断言：——The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

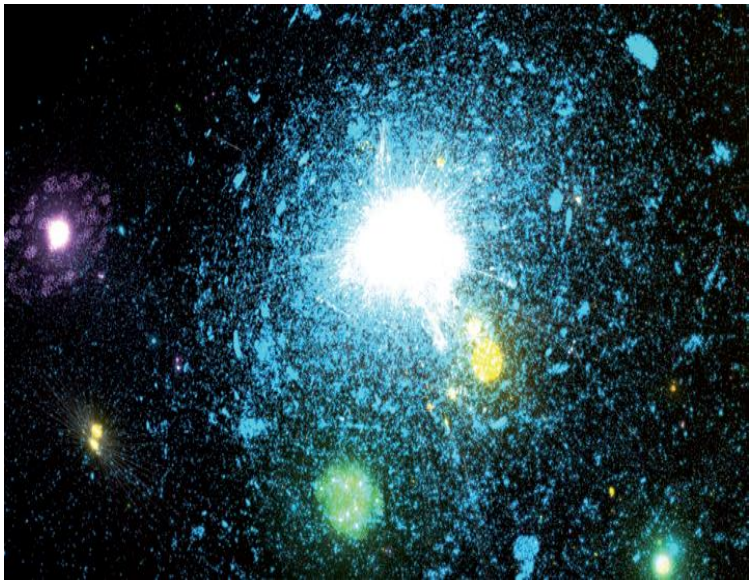
- 获得海量数据和处理这些数据的统计工具的可能性提供了理解世界的一条完整的新途径。
- 相互关系取代了因果关系，没有具有一致性的模型、统一的理论和任何机械式的说明，科学也可以进步
- 只要有足够的数据，数字将为自己辩护 ---With enough data, the numbers speak for themselves.
- 没有任何理由坚持旧的道路，现在是提出这样问题的时候了：科学可以向Google学习些什么

“大数据研究”真不需要假设和模型吗？

- 某些从事大数据研究的学者认为：与传统的统计及人工智能的区别是，大数据处理不需要事先给出“假设”和“模型”，可以直接从数据的相互关系中求解问题。
- 这类问题有多大的普遍性？这种优势是数据量特别大带来的还是问题本身有这种特性？
- 只知道相互关系不知道因果关系会不会“知其然不知其所以然”。所谓从数据中获取知识要不要人的参与，人在机器自动学习和运行中应该扮演什么角色？
- 大数据研究方法是“理论的终结（The End of Theory)”还是科学方法的补充。

究竟是研究“网络数据”还是研究“数据关系网络”？

●上 图是太空中一种星云分布

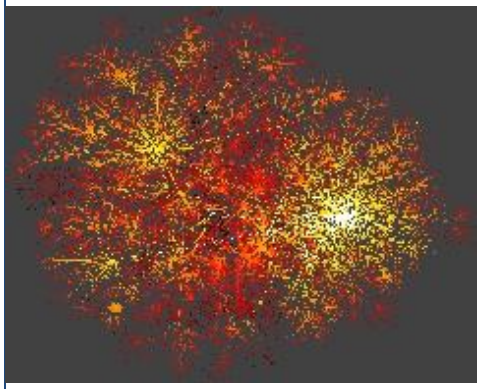


●上 图是1999年画出的万维网分布（由此导出Scale free网），

- 观察互联网、万维网、基因网和社会系统，你会发现尽管在节点本性及其相互作用方面有许多不同，但大多数复杂系统背后的网络都是被一系列确定和限制其行为的基本规律所控制。

究竟是研究“网络数据”还是研究“数据关系网络”？

- 发现Scale-Free网络的Albert-László Barabási教授在2012年1月的NATURE PHYSICS 上发表一篇重要文章： The network takeover，文章认为：
- 20世纪是量子力学的世纪，从电子学到天文物理学，从核能到量子计算，都离不开量子力学。而到了21世纪，网络理论正在成为量子力学的可尊敬的后继 ，它在构建一个理论和算法的框架，从许多研究领域吸取能量，后面紧紧跟随一大批企业
- 还原论解构复杂系统，带给我们单个节点和链接的理论。网络理论煞费苦心地重新组装这些节点和链接，帮助我们重新看到整体



究竟是研究“网络数据”还是研究“数据关系网络？

- 回到本报告中质问的“数据界的共性问题”，很可能数据的共性存在于数据背后的“网络”之中。
- 网络有不少参数和性质，如聚集系数、核数等，这些性质和参数能否刻画大数据背后的网络的共性。
- 所谓大数据的分类问题本质上是不是数据关系网络的的分类问题
- 观察各种复杂系统得到的大数据，直接反映的往往是个体和个别链接的特性，反映相互关系的网络的整体特征隐藏在大数据中，也许网络数据科学的主要任务就是搞清楚数据背后的“网络”。

谢谢！