973 计划信息领域 战略调研材料之三

大数据（Big Data）科学问题研究

李国杰

1、 前言

1.1 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）

用传统算法和数据库系统可以处理的海量数据不算“大数据”。

大数据 =“海量数据”+“复杂类型的数据”

大数据的特性包括 4 个“V”： Volume，Variety，Velocity，Value

- ●数据量大：目前一般认为 PB 级以上数据看成是大数据；
- ●种类多：包括文档、视频、图片、音频、数据库数据等；
- ●速度快：数据生产速度很快，要求数据处理和 I/O 速度很快；
- ●价值大：对国民经济和社会发展有重大影响。

1.2 目前大数据的规模

工业革命以后，以文字为载体的信息量大约每十年翻一番；1970 年以后，信息量大约每三年就翻一番；如今，全球信息总量每两年就可以翻一番。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。根据麦肯锡全球研究院（MGI）预测，到 2020 年，全球数据使用量预计将暴增 44 倍，达到 35ZB（1ZB=10^{21}Byte）。医疗卫生、地理信息、电子商务、影视娱乐、科学研究等行业，每天也都在创造着大量的数据。数据采集成本的下降推动了数据量的剧增，新的数据源和数据采集技术的出现大大增加了数据的类型，数据

类型的增加导致数据空间维度增加，极大地增加了大数据的复杂度。

1.3 大数据公司的现状：

- Google 公司通过大规模集群和 MapReduce 软件，每个月处理的数据量超过 400PB。
- 百度的数据量：数百 PB，每天大约要处理几十 PB 数据，大多要实时处理，如微博、团购、秒杀。
- Facebook：注册用户超过 8.5 亿，每月上传 10 亿照片，每天生成 300TB 日志数据
- 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。

1.4 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。

（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：网络上用户根据自己的需要和喜好发布、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。

2、 国家重大战略需求

数据已成为与自然资源、人力资源一样重要的战略资源，隐含巨大的价值，已引起科技界和和企业界的高度重视。如果我们能够有效地组织和使用大数据，人们将得到更多的机会发挥科学技术对社会发展的巨大推动作用，孕育着前所未有的机遇。O'Reilly 公司断言:“数据是下一个‘Intel Inside’，未来属于将数据转换成产品的公司和人们。”

过去几十年，我们一直大力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，战略需求正在发生重大转变：关注的重点落在数据（信息）上，计算机行业要转变为真正的信息行业，从追求计算速度转变为大数据处理能力，软件也从编程为主转变为以数据为中心。

实验发现、理论预测和计算机模拟是目前广泛采用三大科研范式。现在，数据密集型研究已成为科研的第四范式。不论是基因组学、蛋白组学研究，天体物理研究还是脑科学研究都是以数据为中心的研究。用电子显微镜重建大脑中所有的突触网络，$1mm^3$ 大脑的图像数据就超过 1PB。取之不尽的实验数据是科学新发现的源泉。

大数据分析技术不仅是促进基础科学发展的强大杠杆，也是许多行业技术进步和企业发展的推动力。大数据的真正意义并不在于大带宽和大存储，而在于对容量大且种类繁多的数据进行分析并从中萃取大价值。采用大数据处理方法，生物制药、新材料研制生产的流程会发生革命性的变化，可以通过数据处理能力极高的计算机并行处理，同时进行大批量的仿真比较和筛选，大大提高科研和生产效率。数据已成为矿物和化学元素一样的原始材料，未来可能形成“数据探矿”、“数据化学”等新学科和新工艺模式。大数据处理的兴起也将改变云计算的发展方向，云计算正在进入以 AaaS(分析即服务)为主要标志的 Cloud 2.0 时代。

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，将计算用于数据，而不是将数据用于计算。大数据也导致高可扩展性成为信息系统最本质的需求，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。

近十年来增长最快的是网络上传播的各种非结构化或半结构化的数据。网络数据的背后是相互联系的各种人群。网络大数据的处理能力直接关系到国家的信息空间安全和社会稳定。未来国家层面的竞争力将部分体现为一国拥有数据的规模、活性以及解释、运用数据的能力。国家的数字主权体现在对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间。从心理学、经济学、信息科学等不同学科领域共同探讨网络数据的产生、扩散、涌现的基本规律，是建立安全和谐的网络环境的重大战略需求，是促使国家长治久安的大事。

3、 国内外研究动向与基础

3.1 科研“第四范式”

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。

2007 年，已故的图灵奖得主吉姆•格雷（Jim Gray）在他最后一次演讲中描绘了数据密集型科研“第四范式”（the fourth paradigm）的愿景。2008 年 9 月《Nature》杂志出版了一期专刊—“Big Data”，2011 年 2 月，《Science》期刊联合其姊妹刊推出了一期关于数据处理的专刊—“Dealing with 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海量数据所带来的技术挑战。

将大数据科学从第三范式（计算机模拟）中分离出来单独作为一种科研范式，是因为其研究方式不同于基于数学模型的传统研究方式。Google 公司的研究部主任 Peter Norvig 的一句名言可以概括两者的区别： ″All models are wrong, and increasingly you can succeed without them″。 Petabyte 级的数据使我们可以做到没有模型和假设就可以分析数据。将数据丢进巨大的计算机机群中，只要有相互关系的数据，统计分析算法可以发现过去的科学方法发现不了的新模式、新知识甚至新规律。实际上，Google 的广告优化配置、战胜人类的 IBM 沃森问答系统都是这么实现的，这就是“第四范式”的魅力！

美国 Wired 杂志主编 Chris Anderson 2008 年曾发出“理论的终结（The End of Theory）”的惊人断言：“The Data Deluge Makes the Scientific Method Obsolete”。他指出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。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，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。

Chris Anderson 的极端看法并没有得到科学界的普遍认同，数据量的增加能否引起科研方法本质性的改变仍然是一个值得探讨的问题。对研究领域的深刻理解（如空气动力学方程用于风洞实验）和数据量的积累应该是一个迭代累进的过程。没有科学假设和模型就能发现新知识究竟有多大的普适性也需要实践来检验，我们需要思考：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？有些领域可能先用第四范式，等领域知识逐步丰富了在过渡到第三范式。

3.2 21 世纪的网络理论相当于 20 世纪的量子力学

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

论则反其道而行之，重新组装这些节点和链接，帮助我们重新看到整体。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,这些性质和参数也许能刻画大数据背后的网络的共性。

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，正在构建一个新的理论和算法的框架。

3.3 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，争取增加 100 倍的分析能力从各种语言的文本中抽取信息。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。在这个计划中，不同部门的侧重点并不一样。

3.3.1 国防部高级研究计划局(DARPA)项目举例：

- 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。

3.3.2 能源部 (DOE) 项目举例：

- 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。

3.3.3 国家人文基金会(NEH) 项目举例：

- 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。

3.3.4 美国国家科学基金会(NSF) 项目举例：

- 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）。

从以上项目简介中可以看出，美国政府的大数据计划目前最重视的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。但 NSF 的项目包含“统一的理论框架”和“形成一个独特的学科”等的科学目标。

4、重大科学问题

在讨论大数据带来的科学挑战问题之前，需要先阐述几句大数据面临的技术挑战问题，因为对大数据而言，技术走在科学前面。目前的局面是各个学科的科学家都以自己为主处理本领域的海量数据，信息领域的科学家只能起到助手的作用。也就是说，各领域的科学问题还掌握在各学科的科学家手里，计算机科学家并没有提炼出多少共性的大数据科学问题。技术上解决不了的问题越来越多，就会逐步凝练出共性的科学挑战问题。在条件还不成熟的时候，计算所科学家应虚心地甘当一段时期的“助手”。在网络大数据方面可能计算机学者的主动性会较早发挥出来。

4.1、 需要重视的一些技术挑战问题

4.1.1 高扩展性的数据分析技术

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。系统的高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。目前 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。

4.1.2 新的数据表示方法

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。

4.1.3 数据融合

大数据的挑战之一是对数据的整合，如果不整合则发挥不出大数据的大价值。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。

4.2 大数据提出的科学挑战问题

4.2.1 数据科学的重点是研究数据背后的关系网络

大数据科学面临的首要问题是“研究对象是什么”？许多学者说：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。寻找新算法是有目标的研究，但当前数据科学的目标还不很明确。人们常比喻数据科学是“大海捞针”，“大海捞针”的前提是先知道有一枚“针”在海里，而海量数据的挖掘往往不知道有没有“针”。因此有学者比喻大数据研究是“大海捕鱼”，捕到什么鱼算什么鱼。

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，国外不少学者认为数据科学的主要任务就是搞清楚数据背后的“关系网络”。 因此大数据面临的科学问题本质上可能就是网络科学问题，复杂网络分析是数据科学的重要基石。

目前，研究 Internet 网络数据的学者以复杂网络上的数据（信息）传播机理、搜索、聚类、同步和控制作为主要研究方向。最新的研究成果表明，随机的 scale-free 网络不是一般的“小世界”，而是“超小世界（ultrasmall world），规模为 N 的网络的最短路径的平均长度是 lnlnN（不是一般小世界的 lnN）。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。

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，也许将为研究复杂系统提供新的途径。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。

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。

4.2.2 数据界（Data Nature)的共性科学问题是什么？

数据科学试图把数据当成一个“自然体”来研究，即所谓“数据界（data nature）”，也就是尝试把计算机科学划归为自然科学。但脱离各个领域的“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？数据库理论是一个很好的例子，在经历了层次数据库、网状数据库多年实践以后，Codd 发现了数据库应用的共性规律，建立了有坚实理论基础的关系模型。在这之前人们也一直在问今天同样的问题。现在我要做的事就是提出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。

提炼“数据界”的共性科学问题还需要一段时间的实践积累，至少近五年内计算机界的学者还需要多花精力协助其他领域的学者解决大数据带来的技术挑战问题。通过分层次的不断抽象，大数据的共性科学问题才会逐步清晰明朗。

4.2.3 大数据研究作为一种研究方法的特点

目前，大数据研究主要是作为一种研究方法或一种发现新知识的工具，不是把数据本身当成研究目标。作为一种研究方法，它与数据

挖掘、统计分析、搜索等人工智能方法有密切联系。

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。大数据研究肯定要采用数据挖掘技术。但目前数据挖掘中急用先研的短期行为较多，多数是为某个具体问题研究应用技术，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 PB 级以上的大数据还需研究新的方法。网络数据科学强调与社会科学的深度交叉融合，需要揭示社会科学领域的深层次机制和规律，只用传统的数据挖掘技术难以到达目的。

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有价值的信息，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。单用统计方法往往有能力的极限，例如只用统计机器翻译方法，翻译质量的提高就有限度。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综合性模型。

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。

大数据研究不应该只是上述几种方法的集成，应该有不同于统计学和人工智能的本质内涵。大数据研究是一种交叉科学研究，如何体现其交叉学科的特点需要认真思考。

4.2.4 如何变“大数据”为“小数据”

获取大数据本身不是我们的目的，能用“小数据”解决的问题绝不要故意增大数据量。当年开普勒发现行星三大定律，牛顿发现力学

三大定律现在看来都是基于小数据。我们也应从通过“小数据”获取知识的案例中得到启发，比如人脑就是小样本学习的典型。

2-3 岁的小孩看少量图片就能正确区分马与狗、汽车与火车，似乎人类具有与生俱来的知识抽象能力。我们不能迷信大数据，从少量数据中如何高效抽取概念和知识是值得深入研究的方向。至少应明白解决某类问题，多大的数据量是合适的，不要盲目追求超额的数据。

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。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。

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，各种各样的大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。

5、可能的原始创新

现在来预测我国在大数据研究上可能取得的原始创新可能为时尚早。但可以大致判断一下哪些领域可能取得原始性的贡献。

5.1 基因组学和蛋白组学研究

中国的基因测序能力世界领先，已占到全世界的一半。中国也有不少独特的基因资源，为开展基因组学研究提供了有利条件。但是，在提出新的基因测序原理和方法上，我国学者的贡献还不大，现在用的设备和测序软件几乎都是进口的。如果组织计算机和生物领域的学者密切合作，有可能在信息生物学的大数据研究方面做出原始性创新贡献。

5.2 Web 网络大数据分析

Web 网拥有最大的数据量，而且增长很快，是大数据分析最主要的领域。我国拥有世界上最多的网民和最大的访问量，在网络大数据分析方面已经有较强的基础，有可能做出世界领先的原始创新成果，

应加大网络大数据分析方面的研究力度。

5.3 大数据平台的创新

大数据研究需要的处理平台不同于高性能计算机，需要在体系结构和系统软件上进行原始性创新。我国的高性能计算机研制能力已进入世界三强（美、日、中），有能力在数据密集型计算机方面做出国际领先的原始创新。

5.2 中医和经络的大数据研究

中医中药，特别是经络学说是中华文化的宝贵遗产，但在经络原理的研究方面有落后于韩国的危险。能不能将中医包括经络研究数字化，将几千年的传统医学文献和大量的中医实践记录变成可用计算机分析的大数据，也许能走出一条新路，做出令全世界为之一震的原始创新成果，为中华文化争光。

6、对开展该方向研究的建议

6.1 研究革命性的算法和处理平台结构

大数据研究不是简单地建一个数据中心，也不是使用传统方法在超级计算机上处理生物信息、脑科学、天文物理、遥感、气象等领域的海量数据，即使找到线性复杂性的算法也对付不了 Peta 级以上的数据（如用传统方法备份 PB 级数据就需要数月时间）。必须研究革命性的大数据处理系统结构和革命性的算法和软件，以应对数据指数级增长的挑战。

6.2 选择“预言性数据分析问题”做研究

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）。采用这种研究模式即使问题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解决问题，这类问题的基本的物理本质还不清楚，增加计算规模也无济于事。（c）变革式计算，只要计算能力足够强大，就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

研究社会或其他问题，应考虑首先选择“预言性数据分析问题”，即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，实现有关科学的“变革式”进步。

6.3 研究大数据的测量与感知理论，

大数据不是采集得越多越好，要在不明显增加采集成本的条件下尽可能提高数据的质量。要研究如何科学合理地抽样采集数据，减少不必要的数据采集。当前数据跨领域跨行业的拉通和共享仍存在大量壁垒，海量数据的收集，特别是关联领域的同时收集和处理存在很大挑战。只有跨领域的数据分析才更有可能形成真正的知识和智能，从而产生更大的价值。

6.4 研究数据的去冗余和高效率低成本的数据存储

大数据中有大量的冗余，消除冗余是降低开销的重要途径。大数据的存储方式不仅影响效率也影响成本，需要研究高效率低成本的数据存储方式。需要研究多源多模态数据高质量获取与整合的理论技术、错误自动检测与修复的理论技术和低质量数据上的近似计算的理论和算法

6.5 研究适合不同行业的大数据挖掘分析工具和开发环境

不同行业需要不同的大数据分析工具和开发环境，应鼓励计算机算法研究人员与各领域的科研人员密切合作，在分析工具和开发环境上创新。对于网络上大数据，需要研究互联网信息价值及其社会效应的可计算性以及计算结果的社会性解释。

6.6 研究大幅度降低数据处理、存储和通信能耗的新技术

大数据的处理、存储和通信都将消耗大量的能源，研究创新的节能技术是重要的基础研究方向。

6.7 逐步深入地开展以数据内在共性为研究对象的数据科学研究

目前的数据挖掘主要依赖先进的工具，是工具依赖而不是数据依赖，需要研究隐藏在数据本身中规律和知识，当积累足够多的技术挑战和实践知识后，应适时开展有关 data-nature 方面的理论研究，争取总结出类似关系代数的大数据基础理论。还需要研究海量数据计算

的复杂性理论、海量数据计算的算法设计方法学、海数据管理的理论和算法等。

6.8 大力开展交叉科学研究

大数据研究是跨学科的研究，可以发展为一门新型交叉学科。这项研究不仅与自然科学有关，还涉及心理学、经济学、社会学等社会科学。探讨网络数据的产生、扩散的基本机制，就需要从社会、经济和技术层面探讨网络数据涌现的规律与价值度量方法。应积极鼓励开展交叉科学研究

6.9 改变科研的组织结构和合作形式

开展数据密集型研究需要改变科研的组织结构和合作形式，形成有利于协作创新的“知识生态系统”，强调个人在单学科领域学术成就的“个人化科研范式”不再适合大数据研究，行会文化和过分细分的专业化教育是推广大数据研究的阻力。