大数据技术和应用中的挑战性科学问题

第 89 期双清论坛论证报告

大数据是人类进入信息化时代的产物和必然结果。“大数据发展的核心动力来源于人类测量、记录和分析世界的渴望”，而这种渴望又源于人类努力改善自身生存和生活状况的无尽追求。

在人类社会发展进程中，人们观测自然现象、揭示和把握自然规律并进而用于改善自身生存和生活状况的活动从来都没有停止过。人类揭示和运用自然规律是从观测和记录自然现象开始的，而这种观测和记录的结果要么就是数据，要么可以通过某种方法转化为数据。人类把握和运用自然规律的能力越强，社会经济和科学技术就越发展；社会经济和科学技术越发展，人类揭示和运用自然规律的愿望和需求就越强烈，结果是获取和存储的观测数据就会越来越多。伴随着近代传感器、无线通信、计算机与互联网等技术的迅猛发展及在各个领域的广泛应用，人类获取数据的手段和途径越来越多，成本越来越低，速度越来越快，所获数据的种类、层次和尺度也越来越多样化，这就在广度、速度和深度三个方面催生了大数据时代的到来。

一、开展大数据技术和应用研究的意义

粗略地讲，大数据是指在可容忍的时间内无法用现有的信息技术和软硬件工具对其进行传输、存储、计算与应用等的数据集合。与传统意义上的数据概念相比，大数据具有如下几个显著特征：（1）**数据**

规模（Volume）不断扩大，数据量已从 GB（10^9）、TB（10^{12}）再到 PB（10^{15}）字节，甚至已开始以 EB（10^{18}）和 ZB（10^{21}）字节来计量。“到 2013 年,世界上存储的数据预计能达到 1.2ZB 字节。如果把这些数据全部记录在书中，这些书可以覆盖整个美国 52 次；如果将之存储在只读光盘上，这些光盘可以堆成 5 堆，每一堆都可以伸到月球上。”（2）**数据类型（Variety）繁多**，包括结构化、半结构化和非结构化数据，甚至包括非完整和错误数据。现代互联网上半结构化和非结构化数据所占比例已达 95%以上。(3)**产生和增长速度（Velocity）快**。美国国际数据公司（IDC）的研究报告称，到 2020 年全球的数据获取能力将增加 50 倍，用于数据存储的服务器将增加 10 倍。当今世界，各种数据采集和存储设备每时每刻都在获取和存储大量新的数据。这些数据有时以高密度流的形式快速演变，具有很强的时效性，只有快速适时处理才可有效利用。（4）**数据价值（Value）大，且可整合与多次利用**。对于某一特定的、仅需少量数据的应用而言，大数据呈现出价值密度低的特点，但对于众多潜在的应用而言，大数据整体往往蕴藏着巨大的价值。

大数据时代的到来，撼动了世界的方方面面，从商业、科技、医疗卫生到政府、教育以及社会的其他各个领域。大数据技术和应用一方面对社会、经济和科技的发展带来了重要机遇，另一方面也对数据获取、存储、传输、计算以及应用提出了全新的挑战。开展大数据技术与应用研究，是时代发展的必然要求，具有无可估量的社会经济价值和巨大的科学意义。

开展大数据技术与应用研究的意义可主要概括为如下三个方：

（一）大数据已渗透到每一个行业和业务职能领域，已成为继物质和人力资源之后的另一种重要资源，将在社会经济发展过程中发挥不可替代的作用。大数据将逐渐成为现代社会基础设施的重要组成部分，就像公路、铁路、港口、水电和通信网络一样不可或缺。资源、环境、经济、医疗卫生和国防建设等各种各样的大数据已经和物质资源、人力资源一样成为一个国家的重要战略资源，直接影响着国家和社会的安全、稳定与发展。大数据时代国家层面的竞争力将部分地体现为一个国家拥有的数据规模、活性以及解译和运用数据的能力。

正是由于洞察到大数据无可估量的资源价值，美欧日等发达国家纷纷将大数据技术和应用提升为国家发展战略，旨在抢占大数据时代的战略制高点。2012 年 3 月美国发布《大数据研究和发展倡议》，旨在利用大量复杂数据获取知识和提升洞见能力。2012 年 7 月，日本推出《新 ICT 战略研究计划》，重点关注大数据应用，旨在提升日本竞争力。我国拥有众多的大数据资源，整合与利用的前景极为广阔，尽快将大数据技术和应用提升为国家发展战略具有更为重大的战略意义。

（二）大数据的出现将部分地使科学研究从过去的假设驱动型转化为数据驱动型，从而将为科学技术的发展开辟一条新的途径。有相当数量的科研活动是按如下两条路径展开的：（1）假设事物各组成部分及其相互关系遵从某些规律，然后通过实验或数理逻辑的方法得到该事物的整体规律；（2）假设所研究的事物集合具有某种同质性且各

事物在行为演化过程中互不影响（对应统计学上的独立同分布），随机地选择该集合中的少量事物进行观测并获取相关数据，然后进行数据处理和分析，进而得出该事物集合整体上所遵循的统计规律。第一种路径在没有已知规律可循或事物各组成部分之间的关系过于复杂而难于建立模型时失效；第二种路径在独立同分布假设不成立或采样的随机性得不到保证时失效。需要说明的是有相当多的事物（如人口普查）集合不满足独立同分布假设，且很难做到随机采样。“一旦采样过程中存在任何偏见，分析结果就会相去甚远。” 继第三种科研范式—“计算机模拟仿真”之后，已故图灵奖得主吉姆 •格雷（Jim Gray）在 2007 年的最后一次演讲中将基于数据密集型的科学研究描绘为“第四范式”，并指出面对各种最棘手的全球性挑战，在传统的理论方法因过于复杂而难以解决这些问题时，数据驱动的“第四范式”可能是最有希望解决这些难题的方法。

目前，各学科的发展已越来越离不开数据。除传统的模式识别、数据挖掘和机器学习外，基于数据的建模、预测、反演、决策与控制等已逐渐成为新的研究领域。大数据正在部分地改变着现有的科研模式，也在逐渐地改变着人们的思维定式。因此，面向复杂对象开展大数据处理方法及其应用研究具有重要的科学意义。

（三）大数据及相关处理技术可转化为巨大的社会经济价值，被誉为“未来的新石油”。美英等发达国家在大数据应用方面已有许多成功的案例，例如：利用医疗卫生数据监视医疗体制的运行状况和民众健康的变化趋势，评估不同的医疗技术和治疗方案，并帮助政府选

择和制定恰当的医疗改革方案；利用能源数据推动各相关部门实行节能减排方案；利用交通运输数据疏解交通拥堵；利用网络数据提供信息服务，分析舆情和保障国家安全等。据麦肯锡全球研究所预测，单就医疗卫生一个行业，有效的数据处理和利用每年可带来 3000 亿美元的经济价值。

大数据已被广泛地认为是创造新价值的利器和引领下一轮经济增长的助推剂。开展大数据技术与应用研究具有巨大的经济价值和社会意义。

二、国内外研究现状

在科学研究领域，大数据的规模已经并正在迅速增长的事实同样震撼人心：斯坦福大学已经存储了 350TB 字节的物理实验数据，且每年增长 10PB 字节，预计到 2015 年将达到 EB 级字节；欧洲原子能研究机构的高能物理粒子加速器 LHC 的 CMS 检测器每秒产生 320TB 字节的检测数据；欧洲空间卫星中心每年获取 30 PB 字节的空间信息数据；英国 Sanger 中心 2002 年就已拥有 20TB 字节的基因数据，之后每年以 4 倍的速度增长，至今已达数十 PB 字节。面对极大的数据规模、复杂繁多的数据类型和某些因时效性约束需要快速处理的大数据集合，传统的数据管理、特别是数据处理和分析技术已远远不能满足各种应用需求。方法与技术上的局限性常常使人们处于数据到处“泛滥”而所获知识和价值甚少的困境。大数据具有大价值与其价值利用率低的现实引起了世界科技界的广泛关注和各发达国家政府的高度重视。

随着大数据在世界各个领域的快速渗透和发展，2008 年《Nature》出版了“Big Data”专刊，从互联网、经济、超级计算、环境科学和生物医药等多个方面介绍了海量数据带来的一系列技术问题和挑战。自此，“大数据”开始进入学术界，逐渐成为备受关注的前沿研究课题。2011 年，《Science》推出了数据处理专刊《Dealing with data》。该专刊的核心观点是：有效组织和利用数据将能够进一步发挥科学技术对社会发展的巨大推动作用。2012 年 4 月，欧洲信息学与数学研究协会会刊《ERCIM News》出版专刊《Big Data》，重点讨论了大数据时代的数据管理与处理技术方面的关键问题。IEEE 计算机学会决定，从 2013 年开始，每年举办一次 IEEE Big Data 国际学术会议。Springer 等科技出版社也于近年来相继创刊了大数据方面的国际杂志。上述情况表明，大数据已成为一门新兴科学并已受到科技界广泛重视。

发达国家政府对大数据技术与应用研究给予了高度的重视和关注。美国于 2012 年 3 月发布了《大数据研究和发展倡议》，旨在提高人们从海量数据中提取知识的能力，加快科学发现与工程研发的步伐。2013 年 4 月，美国众议院科学、空间和技术委员会以大数据为专题举行了听证会；多名资深教授和国家科学基金会的高官就如何促进海量数据的分析和利用、如何利用大数据技术激励创新、美国在大数据技术领域的创新能力和研究现状等问题在听证会上发言。2013 年 9 月，美国国立卫生研究院（NIH）宣布，今后 4 年每年提供 2400 万美元，资助 6 至 8 个“从大数据到知识发现的卓越中心”（简称大数据卓越中心），以开发和推广大数据共享、集成、分析与管理的创新

方法、软件和工具，从而帮助研究人员提升利用大规模复杂数据集的能力。这表明美国已把大数据技术和应用研究上升为国家战略，视为推动经济复苏的关键所在。欧盟专门设立了大数据研究征集项目（FP7 Call 8），旨在以大数据基础设施为先导，大幅度提高大数据分析算法和处理系统的效率。日本也推出了《活力 ICT 日本计划》，把大数据研究和应用技术视为国家发展战略。

我国科技界及与信息技术密切相关的产业领域对大数据技术和应用的关注程度正在逐渐增强，并引起了政府相关部门的重视。中国科学院先后于 2012 年 5 月和 2013 年 5 月组织召开了题为“大数据科学与工程”和“数据科学与大数据的科学原理及发展前景”香山会议。国家自然科学基金委员会于 2013 年 3 月在上海召开了题为“大数据技术和应用中的挑战性科学问题”双清论坛，并将“大数据技术和应用中的挑战性科学问题”列入 2014 年的项目指南中，拟以重点项目群的方式支持和推动相关领域的基础研究。国家发展改革委员会与地方政府主导的“智慧城市”计划已开始实施，部分省份已经建成或正在建设一批大数据中心。科技部已经部署了若干个大数据及与大数据密切相关的“973”计划和专项研究计划。

近几年来，美欧将大数据研究的焦点主要集中在面向互联网的信息服务、产品推荐、舆情分析、国家安全以及公共卫生等领域的技术和应用层面。研究思路是摈弃随机采样的传统观点，采用“样本等于总体”的策略对整个大数据集合进行计算和分析；研究目的在于发现隐藏在数据中的事物之间的相关关系而不是因果关系；研究方法仍是

现有的数据挖掘、机器学习和模式识别等方法，并无显著的创新性突破；研究手段主要是分布式并行计算和云计算；研究成果主要体现在数据获取、集成、管理和系统构建、特别是成功应用并产生重大社会影响和经济效益上。需要特别指出的是：大数据技术和应用研究能有今天的局面和影响，美欧 IT 巨头企业及与 IT 密切相关的其他企业发挥了极其重要的作用。就当今世界而言，大数据科学研究的时间还很短，从给定的大数据集中发现事物间的因果关系或其它规律的研究成果还鲜有报道，高效、普适与系统性的大数据处理和分析方法及相关理论还没有形成。这既是当前大数据技术和应用研究的缺憾，也是今后大数据科学研究的努力方向。

三、亟待解决的挑战性科学问题

大数据体量大、结构多样、增长速度快、整体价值大而部分价值稀疏等特点，对数据的实时获取、存储、传输、计算与应用等诸多方面提出了全新挑战。传统的面向小数据的信息技术已很难满足大数据时代下的种种需求。因此，突破传统的思维定式和技术局限，面向具体领域的实际应用，深入研究和发展革命性的、可满足时代需求的大数据获取、存储、传输、处理与计算的新方法和新技术，从大数据中萃取大价值，并将大数据技术向更多应用领域推广，将成为今后信息科学技术及相关领域一项紧迫而重大的科学任务和奋斗目标。

针对大数据技术和应用发展面临的重大机遇和挑战，本期双清论坛的 70 多位与会专家经过认真的研讨和交流，归纳和总结出了如下八个重要科学问题。

1. 高效压缩感知与选择性感知方法

除技术上的可行性和成本之外，过去人们关注的焦点主要是获取数据的质量（客观性、准确性、完备性、分辨率等）和速度，这在小数据时代无疑是正确的，因为小数据在存储、传输、处理和计算上所花的时间很少会影响到应用上的时效性。就某些时效性要求很高的应用而言，如果相应的数据因其大而无法得到适时处理和计算，那么再大再好的数据也将失去其应有的意义。在大数据时代，人们不仅要关注获取数据的质量和速度，更要关心大数据给存储、传输、处理和计算能力以及成本上带来的压力。因此，面对种类繁多、灵活多变的大数据应用，“够用即可”不仅应该成为数据获取的一种指导思想，还应该成为数据获取方法与技术的一个追求目标。

数据获取是数据存储、传输、处理、计算与应用的源头。为了提高大数据从获取到应用整个过程的效率，就必需在存储或传输之前，在满足精度要求的前提下，尽可能大地减小数据的规模。一个现阶段可以想到的方案是从如下三个方面开展研究：（1）进一步改进和发展压缩感知方法和技术，大幅度提高数据的无损压缩比例；（2）借鉴人或其他动物在信息获取时的选择性注意机制，研发“按需获取”且适应能力强的数据获取新方法及相关实现技术（这对图像数据尤为重要），大幅度去除与目标或应用无关的数据；（3）进一步研发将某些数据处理和分析功能（如数据校正、清洗、去噪、特征提取等）提前到数据获取阶段的方法和技术，尽可能减少无用或有害信息，为大幅度提高后续的数据处理和计算效率做准备。

2. 大数据高效存储与管理方法

数据存储是为数据处理和计算、特别是为数据应用服务的。大数据时代下的数据存储和管理必需解决如下挑战性问题：（1）数据增长速度远远超过存储空间增长速度；（2）现有数据存储、管理和调度方法不能适应多源海量异构数据在多种存储设备之间频繁密集流动、以及不同应用对灵活性、便捷性和快速性等的要求。为了解决第一个问题，必需研究高效的去重去冗机制和方法、高效的压缩浓缩机制和方法、高效的遗忘与删除机制和方法，将那些重复的、冗余的、无用的和过时的数据及时地从数据存储设备中清理掉，尽可能大地提升存储空间利用率。为了解决第二个问题，必需协同优化和配置各种数据存取资源，研发高效的数据存储模型、存取技术与交换算法，尽可能大地提升数据存取的速度、效率以及存储管理的灵活性和适应性。

3. 多层多域网络化大数据的高效传输方法

大数据传输的核心和要害问题是传输的“时效性”和“完整性”。不同的应用对时效性和完整性的要求也不尽相同：有的苛求前者而适度放宽后者，如预测、决策和控制等；有的苛求后者而适度放宽前者，如与数据相关的科学研究等；有的对前后两者均有高要求，如救援救灾和军事侦察等。

网络技术的快速发展和应用需求的强大拉动一方面导致网络的规模和异构性急剧增大，另一方面使大数据在跨层、跨域之间实现完整与实时传输变得更加困难。除一部分大数据仍然要求实时传输之外，越来越多的大数据对在传输过程中保持数据本身的完整性提出更为

苛刻的要求。目前的互联网、移动和光网络无法满足大数据传输在实时性和完整性（受数据丢包、乱序和误码等影响）方面的要求，急需研发新的网络体系结构、传输交换机理、通信协议以及高效数据流和网络资源调度方法等，以满足不同应用对大数据流跨层、跨域、实时、完整以及灵活传输的需求。

4. 大数据的多粒度表示与知识提取方法

数据是客观事物某些属性的测量结果和量化表示，具有体现客观事物多层次性和多尺度性等的多粒度属性（大小、结构、分布等）。正确发现（合理划分）与恰当利用数据的粒度属性不仅可以提高计算效率，还有利于发现数据背后隐藏的规律。为了应对多源异构大数据对快速处理和分析提出的挑战，必需研究高效的大数据多粒度表示方法、大数据多粒度智能分析与处理方法、以及跨粒度知识学习与提取方法，进而实现对大数据的深度分析和充分利用。

5. 大数据结构与关系的发现与简约计算方法

传统的简约计算方法主要包括随机采样、主元分析、向量降维或子空间表示、核函数以及数据压缩等方法，而应用广泛的随机采样方法因其建立在坚实的概率统计学之上并有独立同分布之假设，使其能够方便地从“小样本”中发现“大规律”。多源异构大数据不满足独立同分布假设，其处理和计算不得不采用“样本等于总体”的策略。因此，发现或构造面向大数据的、具有坚实数学理论支撑的约简算法既十分困难又十分必要。

大数据内部隐藏着相关事物之间的复杂关系，从而具有高度复杂

的结构。有效挖掘和利用大数据的内在结构和关联关系，并在此基础上实现对大数据的简约计算，对提高大数据计算的性能和效率具有重要意义。为了快速准确地理解和把握大数据内部隐藏着的事物之间的复杂关系，急需研发高效的大数据结构与关联关系的表达、发现、分布、度量、分析与简约计算方法，以满足未来社会对大数据进行在线优化处理的需要。

如何依据领域知识（包括理论和工程上的知识）做到：

以少算多　如随机取样，数据压缩等方法；

以小算大　如主元分析，向量降维等方法；

化整为零　如各种拆分和分布式算法。

6. 大数据高效计算系统结构与方法

大数据计算的核心、关键与要害问题是效率和成本。如果效率和成本不能满足需求，则我们仍将处在大数据环境下的小数据时代。提升计算效率和降低计算成本的两条主要措施是研发新的更高效率的计算系统结构和构造面向大数据的“易计算性”算法。现行的计算系统结构和算法在时效性和成本上不能满足动态、异构与关联性强的大数据计算和分析要求，迫切需要研发新的可扩展并行计算机系统结构，实时高效的内存计算技术，高效存储与计算耦合技术，高效并行的分布式计算方法及相关的基础理论，以满足实时、高效、低能耗与低成本的大数据分析与计算需求。

7. 空间媒体大数据的计算理论与方法

空间媒体数据是一类重要而典型的大数据。空间媒体大数据的高

效计算理论与方法在国家安全、应急救灾与环境监控等领域具有重大的战略意义。空间媒体大数据的应用面临一系列的挑战，如面向全局的集约表达与组织、复杂时空关系的联合计算、以及多尺度复杂行为事件的深度理解等。迫切需要研究空间媒体大数据的高效表达与组织机理、多源异构时空融合、多维关联与协同计算、以及模式发现与价值提炼等核心科学问题，并通过典型应用建立大数据计算、分析与验证平台，促进我国空间媒体大数据关键技术与应用研究快速健康发展。

8. 基于大数据的系统决策、控制和故障诊断方法

面向大规模复杂系统所具有的特征多变、难以准确描述与预测的动态行为以及来源广域、形式混杂、层次多样和持续涌现的系统大数据，急需研究面向系统决策、控制和故障诊断的高速、高精度和低成本的大数据处理、融合与知识获取方法，复杂大系统行为描述、建模、预测与评估方法，高效、安全与高可信的复杂大系统决策、控制和故障诊断新方法及实现技术，为优化系统结构、提高系统运行效益和安全性指标、进而为推动社会经济快速发展提供崭新的理论依据和技术保障。

四、存在的问题及相关政策建议

目前，我国大数据技术与应用研究主要存在如下几个方面的问题。

在大数据发展战略和相关政策层面：尽管政府相关部门对大数据技术与应用越来越重视，然而与美国等发达国家相比，我国目前仍缺乏国家层面的发展计划和相应的经费投入。我国目前明确规划大数据

发展战略的地区和部门还太少，更多是学术界的研讨和呼吁，国家层面的大数据发展战略、计划与相关政策尚未出台，亟需紧紧抓住大数据时代到来的重大机遇期，加快制订国家大数据发展战略和相关政策，推动大数据技术研发和应用在我国快速深入发展。

在大数据资源利用和共享层面：我国是世界第一人口大国，占据着全球最大的互联网市场，拥有大量与种类繁多的数据资源，这为大数据技术研发和应用奠定了广泛而深厚的基础。然而与发达国家相比，由于受发展程度、经济利益、文化传统和其它因素的影响，我国数据资源的“孤岛化”、“荒漠化”与“不完整化”现象比较严重，数据交换、利用和共享率低下，这在一定程度上制约着我国大数据技术研发和应用的发展，急需从技术、市场、政策和法规层面开展深入研究并逐步加以解决。

在大数据科学研究和技术开发层面：与美国等发达国家相比，我国的大数据科学研究和技术开发起步稍晚。在科学研究层面，由于发展时间尚短且在理论和方法上没有大的突破，我国与发达国家的差距不大，但我国科技界因缺乏经费和数据资源的支持而表现出主动性与积极性不足的现状却令人担忧。在技术开发层面，我国与发达国家有较大的差距，这主要由于我国企业界和相关部门对大数据技术研发缺乏足够的认识、坚定的信念、以及科技人才不足等原因所致。

在大数据应用层面：与发达国家相比，我国在大数据应用层面的差距更大。其主要表现是大数据技术在我国的应用和推广发展缓慢，应用领域还很少，所取得的社会经济效益十分有限。其主要原因可概

括为：我国企业界、相关部门和科技界缺乏对“大数据具有大价值”的深刻体会和认识；数据资源开放和利用率低下，数据应用市场不够成熟；大数据处理和分析技术遇到瓶颈性难题，初期投入和应用成本偏高；缺乏具有自主知识产权的核心技术和具有开拓进取精神的创新型人才；安全与隐私保护措施不到位等。

针对上述问题，本期双清论坛的与会专家建议：

（一）加强对大数据发展战略的规划和部署

建议政府主管部门进一步提高对大数据技术与应用的重视，牢牢抓住大数据发展带来的大机遇，在国家层面上规划大数据发展战略并尽快部署和实施，全方位推进我国大数据技术与应用的发展，占领大数据时代国际竞争力的制高点。

（二）加大对大数据基础研究的支持力度

与云计算有所区别，大数据技术和应用涉及数据获取、存储、传输、处理、计算和应用全过程，其重点和难点在于优先研发快速、高效和低成本的大数据处理和分析方法，而这需要加大对基础研究的支持力度。为此建议国家自然科学基金委员会，尽快论证和启动大数据基础理论和关键技术方面的重大研究计划，力争取得一批重大原创性的研究成果，培养一批高水平的研究人才和队伍，为提升我国大数据研发和应用水平提供必要的理论、技术和人才支撑。

（三）进一步完善数据利用、共享、安全与隐私保护措施

虽然我国拥有丰富的数据资源，但数据利用率低和共享程度差的

状况仍然没有得到明显的改善，数据公开共享与数据安全及隐私保护之间的矛盾十分突出。建议政府相关部门进一步完善数据利用、共享、安全与隐私保护方面的法律法规和相关政策措施，尽快改善大数据技术研发和应用的社会环境，提高大数据资源利用率，为国民经济转型升级和社会快速发展做出贡献。

第 89 期双清论坛秘书组

2013 年 10 月 10 日