大数据价值期望探讨

王崇骏
南京大学软件新技术国家重点实验室，江苏 南京 210023

摘要
各边利益主体对大数据价值的共同期盼，引发了社会各界对大数据的普遍关注。不同利益主体的自有利益使然，各边的价值期望是不同的，但这些迥异的价值期望恰恰都是大数据价值实现的目标。尝试从大数据的多边定义和理解出发，梳理不同研究视角的相关研究以及不同利益角色的价值期望，介绍了相关研究及产业化现状，并给出了实践可行的方法、思路和策略。

关键词
大数据；数据科学；价值期望

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017045

Discussions of the value expectations of big data

WANG Chongjun
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Abstract
People from all social circles are concerned about the big data, because all of them think that big data is valuable. However, different people have different value expectation, all of which are the goals when implementing big data project. Multiple kinds of definitions and understandings of big data were attempted to indicate, and then different research perspectives and different value expectations from different people were introduced. Furthermore, some practical and feasible methods, ideas and strategies were given after briefly expressing the relevant research status and industrialization status.

Key words
big data, data science, value expectations

1 引言

“大数据”是近年来最为热门的技术名词之一，从其诞生之日起，就引起了“政、产、学、研、用”各界的普遍热议，也获得了哲学家、科学家、技术研究者、工程研发人员等的普遍关注[1]，其原因或许在于各行各业的人们对“大数据价值”都有共同的预判或期待。

由于价值具有较强的主观性，因而不同利益角色对大数据价值期望并不一样，这直接导致了在大数据价值实现过程中难以形成统一的建设思维和技术口径，从而影响大数据项目的价值彰显。本文对不同的大数据价值期望以及由此引发的对大数据彼此互异的理解进行梳理，尝试抽象出价值实现过程中必须响应的挑战和问题，并从已有的参考文献和工程实践方面给出一些具体的思路和提示。

2 大数据多边理解

2.1 大数据溯源

从宇宙大爆炸到地球诞生，再到简单生命体出现（38亿年前），地球上的物种开始不断地丰富、发展和进化。与其他动物仅通过遗传进化不同，人类在进化过程中发展和演化出了一种非遗传性的继承：通过独一无二且日益发达的文化媒介（语言、文字以及不同历史阶段的存储介质），将知识留给后代，这种文化传统使得人类以很高的速度进化，并最终成为这个星球的统治者。

这里所说的知识指的是人类在改造世界的实践中获得的认识和经验的总结归纳，可以指导解决实践问题的观点、经验、程序等信息。因此发现、传递和学习知识是人类文明进程中亘古不变的主题。参考文献[2]阐述了知识构建和管理的完整过程：从噪音中分拣出数据，转化为信息，升级为知识，升华为智慧，让信息从庞大无序到分类有序。

人类对数据的认知经历了漫长的发展历史，最早可溯源至“数觉”。所谓数觉，指的是在一个小的集合里，增加或减去一个元素的时候，尽管未曾直接知道元素增减，也能够辨认到其中有所变化。**图1**给出了人类从数觉到大数据的大致脉络。

- 人类从最原始的数觉，逐渐形成了数值的概念，然后发明了计数、算术、模拟计算和电子计算，并因为电子计算机的发明和广泛应用，人类从繁重的脑力劳动中解放出来，推动人类向信息社会迈进。
- 因为人类需求的不断膨胀以及软硬件技术的不断发展，人们的关注点在不断地演变：从最开始的数值，到后来的数据，再到目前的大数据。人们的理念也在不断地演变：在数据时代，计算是中心；在大数据时代，数据是中心。
- 从数觉到数据的发展，人类用了亿万年的时间，在这个时间段发明的一系列工具、算法、设备大多是在响应和满足人类最原始的需求；从数据到大数据的发展，人类仅仅用了50年左右的时间，在这个时间段发明的许多工具、算法、设备一方面是在响应用户的膨胀需求，另一方面是在响应因为许多新技术的使用而引发的一系列新问题。

2.2 大数据难题

数据是指存储在某种介质上能够识别的物理符号（数、字符或者其他）。这个定义暗含着数据获取、存储和使用的一般路

径，从计算机的角度来看：

- 数据获取意味着必须将物理信号转换成计算机可以存储的数据，这涉及传感、采样、模拟信号/数字信号转换（A/D转换）以及在bit基础上的字节化和数据化；
- 数据存储意味着将数据存储在什么介质上以及如何组织和管理这些数据，任何一个数据被记录、存储一定有其最原始的价值期望，而一旦原始价值被实现，数据事实就是以一种成本存在；
- 数据使用意味着需要针对某个具体的应用目标，使用计算机相关技术完成问题建模和求解。

围绕数据获取、存储及使用的相关技术涉及的基础学科的发展，使得数据在规模量级、数据精度（类型）、获得速度上都得到迅猛的发展。计算机技术的发展尚不能完全匹配基础学科迅猛发展，应对人类需求不断膨胀而引发的在数据层、计算层、应用层的难题和挑战，在这个情境下，大数据作为一个“难题”被提到人们的面前。

大数据的定义可以参见维基百科：所谓大数据，就是大到无法通过现有手段在合理时间内截取、管理、处理并整理成为人类所能解读的信息。4V（volume、variety、velocity、value）往往是伴随着类似定义的通常描述，而不同的利益角色又会根据不同视角给予更多的补充，比如精确性（veracity）、可视化（visualization）、有效性（validity）等。事实上，所有这些特征都是尝试从数据层、计算层和应用层进行大数据特征描述。总体而言，大数据暗含以下3个方面的属性。

- 规模属性：大数据在规模量级上无论是相对于以往还是绝对值都很大，数据层的大规模性以及数据本身具备的多模式性、多模态性和异构性给存取、算法、计算和应用带来了极大的挑战。
- 技术属性：大数据的价值实现依赖一系列技术合集，涉及数据层、算法层、计算层、应用开发层等多个方面。
- 价值属性：各边角色对大数据价值都有共识和期望，不同利益角色的个体（组织）对大数据的价值理解和关注点不同。

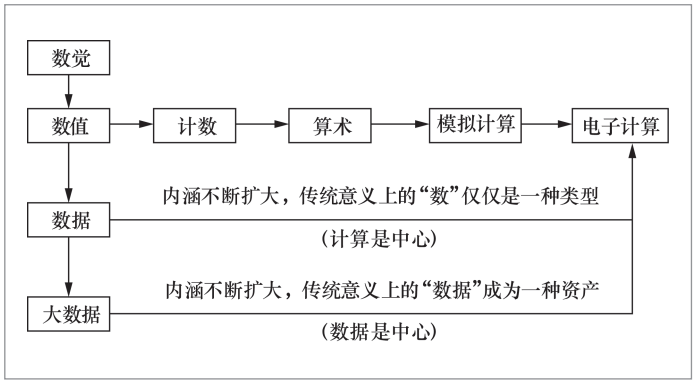

图1 从数觉到大数据

2.3 大数据流程

大数据是数据本身及面向数据价值挖掘的技术集合，能够实现应用驱动的多边商业价值，因此，数据、技术、思维是大数据价值实现的3个基本要素，其基本逻辑在于：富集和整合多数据源数据，然后开发数据处理与分析工具集，在此基础上开发目标应用系统。开展大数据项目建设，自底向上需要从4个方面进行，分别是：数据渠道层、数据基础实施层、应用基础实施层和应用层。

（1）数据渠道层

无论价值期望如何，“数据在哪”是摆在每一位大数据项目建设者面前的首要问题，因此数据渠道梳理、建设是大数据项目开展的第一步。总体而言，数据源可能为既有业务系统、第三方业务系统、政府类数据、社会面数据等。虽然每一类数据源都对应着不同的数据获取协议、方法加以响应，但数据源的梳理和渠道建设往往与技术无关，在更多的情形下往往需要某种商务模式的支撑。值得注意的是，政府类

数据的逐步开放为大数据开放共享提供了一个示范之外，事实上也降低了大数据项目建设在大部分公共数据获取的难度和门槛，特别有助于中小型初创企业的产品研发。

（2）数据基础实施层

在数据源给定并且商务上已经确认的基础上，数据基础实施层的功能目标在于将数据源数据获取、连接、整合和融合，为潜在的应用提供高性能的数据服务。因此在这个层次上，几个基本的技术要点在于：情境知识库的建设、数据采集工具的设计与实现、数据预处理、数据集成、语义融合等。

（3）应用基础实施层

该层主要为未来的潜在应用开发提供基础的计算服务，具体而言几个基本的技术要点在于：通用数据接口、一体化网络服务、数据分析服务、安全服务等。

（4）应用层

该层次主要面向具体的应用需求定制开发垂直应用系统（平台），一般而言，该层次需要考虑的问题有：新开发的系统（功能）与原业务系统的关系、如何为创新应用提供服务支撑、如何与第三方有效无缝连接等。

3 大数据研究视角

3.1 哲学层面

从哲学层面理解大数据包括但不限于以下3个方面。

● 从哲学层面研究大数据给人们带来了什么以及如何理解大数据引发的各类新生事物，往往哲学研究者会关注于此。几个朴素的问题包括：大数据是什么、大数据从哪里来、大数据到哪里去、大数据的本质是什么、是否真的需要大数据、应该如何拥抱大数据。

● 将大数据作为一种思维方式，如何利用大数据解决一些实际问题，“第四范式”[3,4]就是基于这个视角提出的一个新兴概念。

● 作为科学研究的“第四范式”，大数据技术已经成为了科学研究中的一个强大的推动力，目前“科学大数据”已经在地理遥感、高能物理、气候气象、药物化学、社会人文科学等领域得到了广泛应用[5-7]；另一方面，在进行大数据项目开发时，利用数据平台为领域用户进行领域研究和后评估，是大数据项目落地时必须考虑的隐含需求。

3.2 科学层面

从科学的层面来看，大数据背后隐含着若干关键科学问题。

计算机科学关注的是可计算问题，而传统的计算问题可以归结为算法问题，即如果此算法非图灵机可计算，即可判定该问题（计算）无法由计算机处理。而传统的计算机科学关于计算的研究是专注于计算本身，而不在意数据。或者说传统的计算是基于“数据完备、不变”的假设。大数据场景下，数据是持续变化和更新的，传统的计算理论在大数据场景下是否还有效？或者说，在大数据场景下，针对数据海量且不断增量的事实，这个问题是否可计算？进而，在可计算问题之后还可延伸的若干问题，比如计算可信问题、资源（数据、计算、能源等）管理问题等。

上述这些都是典型的计算机科学问题，在大数据场景下，大家更愿意用数据科学来进行描述[3, 8-10]。

3.3 技术层面

在技术层次，大数据项目涉及的技术

流程如**图2**所示，即：在具体计算环境中，从多个数据源中采集数据，然后汇聚在一起，通过对数据进行处理和分析，为具体的应用提供服务[1,11,12]。

下文简要介绍**图2**中涉及的4个技术要点。

(1)数据采集与汇聚

功能上，通过不同的数据获取协议从不同的数据源中获得数据，并将这些数据以某种形式进行集成和连接，有以下3个难点。

● 大数据源自数据层的普适“多源、异构、跨时空”的典型特征，使得在数据采集技术层次上必须基于不同的数据协议进行数据的提取和交换。但是在实际情况下，原始系统开发团队缺位导致的文档缺失、数据库封闭使得数据交换协议缺失；另一方面，由于不同的数据往往是存放在不同的利益主体的服务器上，如果没有持续、匹配的商务合作支撑，数据获取几乎不可实现。

● 任何一个数据源数据的存在都有其最原始的价值期望，每一个数据源表示的物理对象并不一致，加之每个数据源的数据建设依托于不同的IT实施思路和建设水平，这都给有效的数据集成带来了障碍。

● 如何对这些多源、异构、跨时空数据进行有效特征提取、语义理解和融合是重中之重，但也是难题。

(2)数据存储与管理

从不同数据源采集来的数据以及进行各种预处理后的数据以何种方式存取也是一个在技术选型阶段尤其需要考虑的问题。在大数据的应用环境下，随着持续增长的种类繁多的非结构化的海量数据，传统的关系型数据库已经不能满足更加迥异的数据格式了。非关系数据库等新兴的数据存储技术应运而生，如基于Hadoop平台的HBase、Cassandra、MongoDB、Neo4J和Riak等。

在实际应用过程中，数据存储方案的选型往往需要综合考虑源自数据层、应用层和实际应用场景及部署实施的条件约束等多种因素。

(3)数据处理与分析

通过对数据进行有效处理和分析，达到为应用目标服务的目的，这里包含3条路径。

● 在领域知识已经丰富完备的前提下，以逻辑为基础，利用领域知识对数据进行加工处理，然后直接为应用服务。

● 以机器学习与数据挖掘为基础，通过对历史数据进行建模获得知识，然后利用此知识对数据进行加工处理，然后直接为应用服务。在大数据应用环境下，机器学习与数

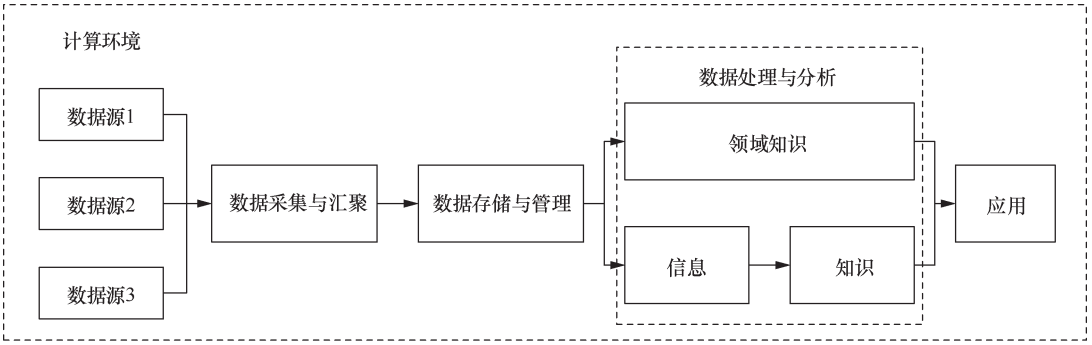

图 2 大数据技术逻辑

据挖掘面临的挑战在于数据及标签的动态变化性、专家标注样本成本过高等，参考文献[13-21]都是围绕这些挑战展开的。

● 将上述基于领域知识和数据驱动的知识发现有效地结合在一起，从而达到双边互补，同时更好地为应用服务。以AAAI 2017的最佳论文[22]为例，该论文利用了先前的领域知识，从而将输出空间约束到一个特定的学习结构，而不是简单地从输入到输出的映射，这种做法使得在进行神经网络训练的时候不需要大量有标签数据，却能够学习得到更先进的结构。

（4）计算环境

大数据的复杂性及规模性给大数据分析带来的挑战至少有两点：如何响应数据类型的复杂性给数据的理解、建模带来的挑战；如何用更快的计算效率响应数据的海量、并行及快速更新的特性。前者的挑战需要研发新型的理论、算法、技术，而后者需要所有的算法、技术（改进）必须依赖合适的高性能计算架构，目前用于高性能计算的策略有以下3类。

● 充分提升和挖掘单个计算节点的计算性能，比如通过对计算主机进行中央处理器（central processing unit，CPU）、内存、硬盘等的扩容尝试，增加单个计算节点的计算性能，显然，这已不是纯粹的技术层次的问题。

● 通过图形处理器（graphics processing unit，GPU）技术的引入达到大幅提升单台计算设备的计算性能。相对而言，CPU的灵活性最大，可以高效运行各种计算任务，但局限是一次只能处理相对很少量的任务；GPU不像CPU那样灵活，处理的任务范围较小，但其强大之处在于能够同时执行许多任务。谷歌公司在ISCA 2017 公布的张量处理器（tensor processing unit，TPU）论文代表了谷歌公司为其人工智能服务设计专用硬件迈出的第一步[23]。TPU是专门用于TensorFlow（谷歌公司开源的一个机器学习软件库）中执行特性功能的专用集成电路（application specific integrated circuit，ASIC），相对而言几乎没有灵活性，但性能极佳。

● 将复杂的任务“分而治之”，引入分布式计算架构以提升计算性能，分布式计算的基本出发点在于通过更多的计算能力不是很强的计算节点，利用某种合适的策略达到整体计算性能的极大提升，利用不同的分布式策略和目标，达到高性能计算的目的。目前主流的分布式计算架构有Hadoop、Spark、Storm等。当前许多工作[24-34]是针对不同的应用场景、不同的数据特点，围绕计算架构的改良展开相关研究，也有许多工作是在既定的计算架构基础上进行算法改良研究。以参考文献[35]为例，作者在Spark平台上实现了一个可扩展复合优化学习方法——SCOPE，实验结果表明，其优于Spark上的其他分布式学习方法。

4 大数据价值期望

4.1 作为一种战略资源

对于政府而言，大数据被认为是提高综合国力、增强本国竞争力的重要来源，因此，从国家层面出台各类政策、指南，借此引导企业和组织进行顺势的发展，就成为各国政府角力的法门。

美国政府在2012年3月29日发布全球首个国家层面的大数据战略，2016年5月23日发布“联邦大数据研究与开发战略计划”，以加速其2012年提出的“大数据研发行动”进程。此外，联合国的“数据脉动”计划、英国的“数据权”运动、日本的“面向2020年的ICT综合战略”、韩国

的大数据中心战略都是政府从战略层面进行的国家意志层次的规划和部署。

我国政府在2014年政府两会工作报告中就提及大数据计划，2016年11月29日，国务院印发《关于印发“十三五”国家战略性新兴产业发展规划的通知》，提出实施大数据战略。

在各国政府实施大数据战略的时候，政府数据开放是其中的一个共同行动。美国政府从2009年开始就进行了政府数据开放的行动，并于2010年实施以个人为中心的数据富集、整合和开放服务的“我的大数据计划”。欧盟在2010年开始提出“欧洲数字化议程”，旨在建立一个统一的“数字市场”，并在2012年制定了大数据战略，强调了公共数据安全及挖掘公共机构数据的价值潜力。英国是最早推进大数据规划的欧洲国家，2004年，英国设立了水平扫描中心（horizon scanning center，HSC）项目，以提升政府处理跨部门和多学科挑战的能力；2011年，HSC项目启动气候变化的未来国际影响计划，通过对多数据源进行深度分析，研究解决气候变化对食品和水的可获得性。

2008年5月1日，我国正式施行《中华人民共和国政府信息公开条例》，全面推行政府信息公开。2015年4月22日，国务院办公厅印发《2015年政府信息公开工作要点》，列出了推进行政权力清单、国有企业、财政资金、社会组织和中介机构、环境保护、食品药品安全等9个重点领域的信息公开；2015年8月31日，国务院印发《促进大数据发展行动纲要》，明确提出要建立国家大数据，统筹规划大数据基础设施建设，为政府治理大数据的建设服务。

优政、惠民、兴业是政府数据开放的一个核心价值观，而如何充分利用政府大数据的开放共享是值得产业界和工业界关注的一个课题。

4.2 作为一种竞争资本

社会各界对大数据都抱以极大的热情，认为引入大数据能够提高自身的竞争力。人们会有这样价值期望的一个基本动机在于：人们认为通过大数据处理与分析能够洞悉客户、友商、产品、渠道在各个维度的信息情报和知识洞见，借此为创新应用模式及商业模式的设计提供研判线索和技术基础。以芝麻信用为例，其从身份特质、行为偏好、人脉关系、信用历史、履约能力等多个角度对一个自然人的相关数据进行搜集和汇聚，在此基础上对个人进行信用研判，根据信用评级就可以进一步进行信用骑行、便利交通、基础通信、信用借还、信用回收等一系列产品的设计和运维。

因此，如何在多源数据汇聚的基础上实现对实体（客户、友商、产品、渠道等）的精准画像，为潜在应用提供多层面的实体语义服务就是重中之重，其中涉及的技术要点主要有以下几方面。

（1）数据连接及实体结构化

由于大数据汇聚的数据源数据都是建立在不同IT建设思路、不同的数据原始价值期望，以“烟囱式”软件开发方式生成并存储的，因此如何将这种格式、表示均异构的数据以实体为中心连接起来，最终形成以实体为中心的数据表示是实体画像的基础。涉及的相关技术包括：数据结构化、数据预处理、去重、歧义消解、实体及实体关系抽取（针对非结构化数据）等。

（2）实体特征化及语义化

在以实体为中心的数据汇聚的基础上，以实体为中心，针对不同视角和采集渠道的数据进行特征提取，并在此基础上实现标签化提取，从而最终达到实体特征表示和实体语义标签化的目的（该标签可作为后续进一步应用或者分析的特征

表示）。由于数据的类型不一样，因此特征表示和提取的方式也不一样。另一方面，语义标签的提取往往是通过监督学习完成，这意味着需要通过人机交互配置训练集，而出于不同的商业目标或者源自领域专家的主观性，这些语义标签是动态变化的，如何动态维护标签池以及如何响应标签的动态变化也是大数据应用的一个挑战。

（3）应用场景的挖掘

“场景”指的是在某时（when）、某地（where）、周围出现了某物（with what）时，特定类型的用户（who）萌发了某种欲望（desire），会想到通过某种手段（method）来满足欲望。因而应用场景挖掘最终归结为“用户是谁（who）”“需求在哪（desire and method）”“应用情境是什么（when, where and with what）”，几个可能的实现途径有：从利益角色的职能定位、可用性需求、用户体验需求角度出发，研判以用户为中心的“痛点”和利益增长点；从既有业务功能、任务或者功能驱动的角度出发，研判以既有业务为中心的可拓展功能和新增功能；从数据本身出发，通过对数据表示的实体内涵以及从该数据中发现的实体关系等找出相应的应用场景，往往是数据、思维和想象力驱动的。

为了实现大数据价值，政（见第4.1节）、产（面向大数据应用及产品开发提供各类服务的产业）、学、研各边也都从各自的角度开展了相应行动，同时提高自己的竞争优势。

（1）产业界

在大数据产业链中，逐渐形成了3种典型的公司形态①[36,37]：数据型公司，如Kaggle、LinkedIn、Twitter、Facebook等；技术型公司，如Amazon、Microsoft、Apple、Intel、Oracle等；思维型公司，如DeepMind、Salesforce、Tableau、SAS等。在两化融合战略的引领下，中国企业本身也自发形成了利用大数据改良既有生产、制造、商务流程的新动力。

（2）学术界

作为关键技术预研的担当，学术界围绕大数据涉及的各类科学理论和关键技术展开研究，积淀形成越来越丰富的科研成果，为大数据应用提供了坚实的理论技术支撑。特别值得一提的是，基于开源精神的开源社区在大数据环境下得到极大的发展，这进一步促进了大数据技术发展的同时，也大范围降低了大数据应用的门槛。开源使新项目直接使用大数据技术成为可能，包括Spark、Apache NiFi、Kafka等在内的开源大数据技术为大数据应用提供了极大的便利。大数据方面的开源工具主要分为4个方面：数据存储（如HBase）、开发平台（如Hadoop）、开发工具和集成（如Apache Mahout）以及分析和报告工具（如Talend）等[38-43]。

（3）教育界

从人才培养的角度而言，为社会培养大数据人才，并借此提高影响力也成为教育界关注的事，体现在两个方面：一是出现了越来越多的大数据技术培训班；二是教育部从发展战略角度，也有意部署大数据专业建设战略，比如教育部于2015年、2016年分两批批复包括北京大学、中国人民大学等35所大学建设“数据科学与大数据技术专业”。

4.3 作为一种商品形式

作为一种商品，数据可以买卖，可以增值，这也是大数据时代的一个基本特征。

国外数据交易大致开始于2008年，一些前瞻性的企业开始加大对数据业务的投入，初见端倪的数据应用新业态包括“数据市场”“数据银行”“数据交易公约”等，知名数据服务商有Microsoft

① http://mattturck.com/2016/02/01/big-data-landscape

数据市场、Amazon公共数据集、Oracle在线数据交易等。国内数据交易起步于2010年左右，2015年9月，我国发布的《促进大数据发展行动纲要》中明确提出要引导培育大数据交易市场，开展面向应用的数据交易市场试点，探索开展大数据衍生产品交易，建立健全数据资源交易机制和定价机制。

参考文献[44]综述了我国数据交易的发展历史，并将我国目前的数据交易归纳为如下4种主要形式。

● 基于大数据交易所的大数据交易：如贵阳大数据交易所、长江大数据交易所、武汉东湖大数据交易平台、上海数据交易中心等。

● 基于行业特定领域数据的大数据交易：针对具体的行业，数据标准较易实现数据的统一采集、统一评估、统一管理、统一交易，从而实现特定行业数据的交易。

● 数据资源企业推动的大数据交易，如数据堂、美林数据、爱数据等。

● 互联网企业“派生”出的大数据交易。

作为一个新生事物，许多来自应用层的基本问题也需要考虑，比如：数据交易的卖方是数据的拥有方还是纯粹的交易平台、卖的是原始数据还是加工过后的数据、买家是个人（组织）还是机器、交易目标是社会福利最大化还是数据拥有方或者平台利润最大化、是封闭拍卖还是开放拍卖。

另一方面，数据交易显然还需在若干法律、技术及管理都匹配的情况下才能得以实现。法律方面需要考虑的是：哪些数据可以买卖、数据的所有权和使用权归属谁、是否所有人（组织）都具有买卖资格。在监管方面：数据交易秩序如何监管、数据质量如何保障、定价机制如何形成以及如何监管、交易监管的目标是什么。而在技术层次需要考虑的问题更多，比如：数据质量评估、定价机制设计、交易机制设计、交易行为分析、程序化交易如何实现等。

作为一种技术响应，若干既有工作或许可以作为大数据交易的有益参考。

● 大数据交易的交付物“数据”本身是一个典型的数字产品，因此参考文献[45-47]探讨的关于“数字产品拍卖中如何收益最大化”的研究成果和思路，对于大数据交易具有极大参考价值。

● 大数据交易往往在线实时进行，即买家随时可能到来，因此参考文献[48-51]关于在线机制设计的研究成果和思路对于大数据交易也具有极大的参考价值。

● 如何在拍卖中实现收入最大化，这在机制设计领域是一个重要的开放问题，也是目前最为活跃的一个研究领域，参考文献[52,53]对于以拍卖方式进行大数据交易而言，具有重要的参考价值。

● 其他方面：大数据交易中很可能会出现异常或者恶意扰乱市场的行为，这意味着在线和实时的自动侦测与预警是大数据交易监管中的一个重要课题，参考文献[54,55]对于数据交易中的异常侦测有较大参考价值；在数据交易的产权方面，可能会遇到产权不明、授权复杂、定价困难、交易缺乏透明性、隐私泄露等问题，参考文献[56,57]或许可以作为有益参考。

5 结束语

“value”是大数据的一个重要特征，或许也正是因为“有价值”，才引发了“政、产、学、研、用”各界对大数据的普遍关注。

不同利益主体迥异的价值期望都是大数据价值实现的目标。这意味着，在大数据项目建设中，必须通过具体的策略、思路和方法确保这些价值期望的实现。事实

上，一个成熟的大数据应用必然涉及理论研究、技术研发、产品开发、部署运维等多边的合作与协作，这其实也暗含大数据价值的最大体现：推进多边资源的融合，并刺激集体智慧的涌现。

本文尝试从大数据的多边定义和理解出发，针对不同的价值期望，从战略资源、竞争资本和商品形式3个角度介绍了实践可行的系列方法、思路和策略。应当看到，本文提及的具体思路和策略仅针对整个大数据部署、实施的普适性特点展开叙述。这意味着，针对特定的领域细分应用，也有其独特的价值实现路径特点，需要在具体的应用进程中加以把握和关注。

参考文献:

[1] 王崇骏. 大数据思维与应用攻略[M]. 北京: 机械工业出版社, 2016.
WANG C J. Big data thinking and application raiders[M]. Beijing: China Machine Press, 2016.

[2] SCARDAMALIA M, BEREITE C. Computer support for knowledge-building communities [J]. Journal of the Learning Sciences, 1994, 3(3): 265-283.

[3] 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域[J]. 中国科学院院刊, 2012(6): 647-657.
LI G J, CHENG X Q. Research status and scientific thinking of big data[J]. Bulletin of the Chinese Academy of Sciences, 2012(6): 647-657.

[4] 维克托·迈尔·舍恩伯, 肯尼思·库克耶. 大数据时代: 生活、工作与思维的大变革[M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013.
MAYER-SCHÖNBERGER V, CUKIER K. Big data: a revolution that will transform how we live, work and think[M]. Translated by SHENG Y Y, ZHOU T. Hangzhou: Zhejiang People's Publishing House, 2013.

[5] 郭华东, 王力哲, 陈方, 等. 科学大数据与数字地球[J]. 科学通报, 2014(12): 1047-1054.
GUO H D, WANG L Z, CHEN F, et al. Scientific big data and digital earth[J]. Chinese Science Bulletin, 2014(12): 1047-1054.

[6] 陈刚. 科学研究大数据挑战[J]. 科学通报, 2015(5): 439-444.
CHEN G. Challenges of big data in science researches[J]. Chinese Science Bulletin, 2015(5): 439-444.

[7] 刘言, 蔡文生, 邵学广. 大数据与化学数据挖掘[J]. 科学通报, 2015(8): 694-703.
LIU Y, CAI W S, SHAO X G. Big data and chemical data mining[J]. Chinese Science Bulletin, 2015(8): 694-703.

[8] KAISLER S, ARMOUR F, ESPINOSA J A, et al. Big data: issues and challenges moving forward[C]// 46th Hawaii International Conference on System Sciences (HICSS), Jan 7-10, 2013, Wailea, Maui, HI, USA. New Jersey: IEEE Press, 2013: 995-1004.

[9] JIN X, WAH B W, CHENG X, et al. Significance and challenges of big data research[J]. Big Data Research, 2015, 2(2): 59-64.

[10] WANG H, XU Z, FUJITA H, et al. Towards felicitous decision making: an overview on challenges and trends of big data[J]. Information Sciences, 2016(s367): 747-765.

[11] 李学龙, 龚海刚. 大数据系统综述[J]. 中国科学: 信息科学, 2015, 45(1): 1-44.
LI X L, GONG H G. Summary on big data system[J]. SCIENTIA SINICA Informationis, 2015, 45(1): 1-44.

[12] 程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(9): 1889-1908.
CHENG X Q, JIN X L, WANG Y Z, et al. Survey on big data system and analytic technology[J]. Journal of Software, 2014, 25(9): 1889-1908.

[13] HUANG S J, CHEN S, ZHOU Z H. Multi-label active learning: query type matters[C]// 24th International Conference on Artificial Intelligence, July 25-31, 2015, Buenos Aires, Argentina. New

York: ACM Press, 2015: 946-952.
[14] ZHU Y, GAO W, ZHOU Z H. One-pass multi-view learning[C]// 7th Asian Conference on Machine Learning, November 20-22, 2015, Hong Kong, China. New York: ACM Press, 2015: 407-422.
[15] HUANG S J, CHEN S, ZHOU Z H. Multi-label active learning: query type matters[C]// 24th International Conference on Artificial Intelligence, July 25-31, 2015, Kyoto, Japan. New York: ACM Press, 2015: 946-952.
[16] ZHOU Z H. Ensemble learning[M]. Beijing: Tsinghua University Press, 2015: 411-416.
[17] ZHANG M L, ZHOU Z H. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837.
[18] NGUYEN C T, WANG X L, LIU J, et al. Labeling complicated objects: multi-view multi-instance multi-label learning[C]// 28th AAAI Conference on Artificial Intelligence, July 27-31, 2014, Québec City, Québec, Canada. New York: ACM Press, 2014: 2013-2019.
[19] WEI X S, WU J, ZHOU Z H. Scalable algorithms for multi-instance learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(4): 975-987.
[20] ZHU Y, TING K M, ZHOU Z H. Discover multiple novel labels in multi-instance multi-label learning[C]// 31st AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, USA. [S.l.:s.n.], 2017: 2977-2984
[21] WEI X S, WU J, ZHOU Z H. Scalable multi-instance learning[C]// IEEE International Conference on Data Mining, Dec 14-17, 2014, Québec, Canada. New Jersey: IEEE Press, 2014: 1037-1042.
[22] STEWART R, ERMON S. Label-free supervision of neural networks with physics and domain knowledge[C]// 31st AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, USA. [S.l.:s.n.], 2017: 2576-2582.
[23] JOUPPI N P, YOUNG C, PATIL N, et al. In-datacenter performance analysis of a tensor processing unit[C]// 44th Annual International Symposium on Computer Architecture, June 24 - 28, 2017, Toronto, Canada. New York: ACM Press, 2017: 1-12.
[24] SAYAR A. Hadoop optimization for massive image processing: case study face detection[J]. International Journal of Computers Communications & Control, 2014, 9(6): 664-671.
[25] GU R, WANG S, WANG F, et al. Cichlid: efficient large scale RDFS/OWL reasoning with spark[C]// 2015 IEEE International Parallel and Distributed Processing Symposium, May 25 - 29, 2015, Orlando, USA. Washington, DC: IEEE Computer Society, 2015: 700-709.
[26] GU R, YANG X, YAN J, et al. SHadoop: improving MapReduce performance by optimizing job execution mechanism in Hadoop clusters[J]. Journal of Parallel and Distributed Computing, 2014, 74(3): 2166-2179.
[27] 顾荣, 严金双, 杨晓亮, 等. Hadoop MapReduce短作业执行性能优化[J]. 计算机研究与发展, 2014, 51(6): 1270-1280.
GU R, YAN J S, YANG X L, et al. Performance optimization for short job execution in Hadoop MapReduce[J]. Journal of Computer Research and Development, 2014, 51(6): 1270-1280.
[28] YANG M, MA R T B. Smooth task migration in Apache storm[C]// 2015 ACM SIGMOD International Conference on Management of Data, May 31-June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 2067-2068.
[29] ARMBRUST M, DAS T, DAVIDSON A, et al. Scaling spark in the real world: performance and usability[J]. Proceedings of the VLDB Endowment, 2015, 8(12): 1840-1843.
[30] ARMBRUST M, XIN R S, LIAN C, et al. Spark sql: relational data processing in spark[C]// 2015 ACM SIGMOD International Conference on Management of Data, May 31-June 4, 2015,

Melbourne, Australia. New York: ACM Press, 2015: 1383-1394.

[31] VAN DER VEEN J S, VAN DER WAAIJ B, LAZOVIK E, et al. Dynamically scaling apache storm for the analysis of streaming data[C]// 1st International Conference on Big Data Computing Service and Applications, March 30-April 2, 2015, San Francisco, USA. New Jersey: IEEE Press, 2015: 154-161.

[32] SCHAEFER C, MANOJ P M. Enabling privacy mechanisms in apache storm[C]// 1st International Conference on Big Data Computing Service and Applications, March 30-April 2, 2015, San Francisco, USA. New Jersey: IEEE Press, 2015: 102-109.

[33] BOSAGH ZADEH R, MENG X, ULANOV A, et al. Matrix computations and optimization in apache spark[C]// 2016 ACM SIGKDD Knowledge Discovery and Data Mining (SIGKDD-16), August 13-17, 2016, San Francisco, USA. New York: ACM Press, 2016: 31-38.

[34] ARMBRUST M, XIN R S, LIAN C, et al. Spark sql: Relational data processing in spark[C]// 2015 ACM SIGMOD International Conference on Management of Data, May 31-June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 1383-1394.

[35] ZHAO S Y, XIANG R, SHI Y H, et al. SCOPE: scalable composite optimization for learning on spark[C]// 31st AAAI Conference on Artificial Intelligence(AAAI-17), February 4-9, 2017, San Francisco, USA. [S.l.:s.n.], 2017: 2928-2934.

[36] 中国电子信息产业发展研究院. 2016中国大数据产业生态地图[Z]. 北京: 中国电子信息产业发展研究院, 2016.
China Center for Information Industry Development. 2016 Chinese big data industrial ecology map[Z]. Beijing: China Center for Information Industry Development, 2016.

[37] 王叁寿. 大数据商业应用场景[M]. 北京: 机械工业出版社, 2016.
WANG S S. Big data commercial application scenarios[M]. Beijing: China Machine Press, 2016.

[38] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[J]. Communications, 2008, 51(1): 107-113.

[39] GHEMAWAT S, GOBIOFF H, LEUNG S T. The Google file system[C]// 19th ACM Symposium on Operating Systems Principles(SOSP-03), October 19-22, 2003, Bolton Landing, USA. New York: ACM Press, 2003, 37(5): 29-43.

[40] CHANG F, DEAN J, GHEMAWAT S, et al. Bigtable: a distributed storage system for structured data[J]. Transactions on Computer Systems, 2006, 26(2): 4.

[41] SHVACHKO K, KUANG H, RADIA S, et al. The Hadoop distributed file system[C]// IEEE 26th Symposium on Mass Storage Systems and Technologies(MSST-10), May 3-7, 2010, Nevada, USA. New Jersey: IEEE Press, 2010: 1-10.

[42] BORTHAKUR D. The Hadoop distributed file system: architecture and design[J]. Hadoop Project Website, 2007, 11(11): 1-10.

[43] 雷军, 叶航军, 武泽胜, 等. 基于开源生态系统的大数据平台研究[J]. 计算机研究与发展, 2017, 54(1): 80-93.
LEI J, YE H J, WU Z S, et al. Big data platform based on open source ecosystem[J]. Journal of Computer Research and Development, 2017, 54(1): 80-93.

[44] 唐斯斯, 刘叶婷. 我国大数据交易亟待突破[J]. 中国发展观察, 2016(13): 19-21.
TANG S S, LIU Y T. Chinese big data transactions need to break through[J]. China Development Observation, 2016(13): 19-21.

[45] GOLDBERG A V, HARTLINE J D, WRIGHT A. Competitive auctions and digital goods[C]// ACM-SIAM Symposium on Discrete Algorithms(SODA-10), January 17-19, 2010, Austin, USA. [S.l.:s.n.], 2001: 735-744.

[46] GOLDBERG A V, HARTLINE J D, KARLIN A R, et al. Competitive auctions[J]. Games and Economic Behavior, 2006, 55(2): 242-269.
[47] CHEN N, GRAVIN N, LU P. Optimal competitive auctions[C]// 46th Annual ACM Symposium on Theory of Computing(STOC-14), May 31-June 3, 2014, New York, USA. [S.l.:s.n.], 2014: 253-262.
[48] LAVI R, NISAN N. Competitive analysis of incentive compatible on-line auctions[J]. Theoretical Computer Science, 2000, 310 (1): 159-180.
[49] LAVI R, NISAN N. Online ascending auctions for gradually expiring items[C]// 6th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-05), January 23-25, 2005, Vancouver, Canada. Philadelphia: Society for Industrial and Applied Mathematics, 2005: 1146-1155.
[50] FRIEDMAN E J, PARKES D C. Pricing wifi at starbucks: issues in online mechanism design[C]// 4th ACM Conference on Electronic Commerce(EC-03), June 9 - 12, 2003, San Diego, CA, USA. New York: ACM Press, 2003: 240-241.
[51] MASHAYEKHY L, NEJAD M M, GROSU D, et al. An online mechanism for resource allocation and pricing in clouds[J]. IEEE Transactions on Computers, 2016, 65(4): 1172-1184.
[52] MYERSON R B. Optimal auction design[J]. Mathematics of Operations Research, 1981, 6(1): 58-73.
[53] 唐平中. 计算经济学与最优机制设计问题[J]. 中国计算机学会通讯, 2013, 9(10): 18-23.
TANG P Z. Computational economics and optimal mechanism design[J]. Communications of the CCF, 2013, 9(10): 18-23.
[54] SUN J, QU H, CHAKRABARTI D, et al. Neighborhood formation and anomaly detection in bipartite graphs[C]// 5th IEEE International Conference on Data Mining(ICDM-05), Nov 27-30, 2005, Houston, TX, USA. New Jersey: IEEE Press, 2005: 1-8.
[55] RAZ O, KOOPMAN P, SHAW M. Semantic anomaly detection in online data sources[C]// 24rd International Conference on Software Engineering(ICSE-02), May 25, 2002, Orlando, USA. New Jersey: IEEE Press, 2002: 302-312.
[56] 汤琪. 大数据交易中的产权问题研究[J]. 图书与情报, 2016(4): 38-45.
TANG Q. Study on the property right issues in big data trade[J]. Library and Information, 2016(4): 38-45.
[57] REICHMAN J H, SAMUELSON P. Intellectual property rights in data[J]. Vanderbilt Law Review, 1997, 50(4): 337-348.

作者简介

王崇骏(1975-),男,博士,南京大学计算机科学与技术系及软件新技术国家重点实验室教授、博士生导师,主要研究方向为自主Agent及多Agent系统、复杂网络理论及应用、大数据分析及智能系统。截至2016年底,主持和参与包括“973”项目、国家发展和改革委员会专项、工业和信息化部产业化基金、国家自然科学基金、国家社会科学基金、省自然科学基金及支撑计划在内的国家及省部级基金与企事业资助项目50余项。在教育医疗类惠民行业、优政兴业类政府领域、互联网新经济领域有30余项科研成果获得产品化和商品化推广。

收稿日期: 2017-04-21
基金项目:国家自然科学基金资助项目(No.61375069, No.61403156);科技部重点研发基金资助项目(No.2016YFB1001102)
Foundation Items: The National Natural Science Foundation of China(No.61375069, No.61403156), The National Key Research and Development Program of China(No.2016YFB1001102)