

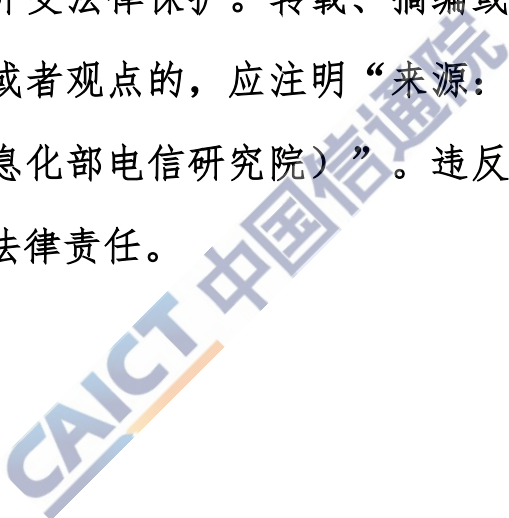
大 数 据 白 皮 书

(2016年)

中国信息通信研究院
2016年12月

版权声明

本白皮书版权属于中国信息通信研究院（工业和信息化部电信研究院），并受法律保护。转载、摘编或利用其它方式使用本白皮书文字或者观点的，应注明“来源：中国信息通信研究院（工业和信息化部电信研究院）”。违反上述声明者，本院将追究其相关法律责任。



前 言

大数据是国家基础性战略资源，是 21 世纪的“钻石矿”。党中央、国务院高度重视大数据在经济社会发展中的作用，提出“实施国家大数据战略”，出台《促进大数据发展行动纲要》，全面推进大数据发展，加快建设“数据强国”。

“十三五”时期是我国全面建成小康社会的决胜阶段，是新旧产业和发展动能转换接续的关键时期，全球新一代信息技术产业正处于加速变革期，国内市场应用需求处于爆发期，我国大数据产业发展面临重要的发展机遇。

本白皮书是继《大数据白皮书（2014）》之后我院第二次发布大数据白皮书。本白皮书首先回顾和阐述了大数据的内涵及产业界定，并以大数据产业几个关键要素为核心，重点从大数据技术发展、数据资源开放共享、大数据在重点行业的应用、大数据政策法规等四个方面分析了最新进展，力求反映我国大数据产业发展状况的概貌。最后结合我国大数据发展最新状况及问题，提出了进一步促进大数据发展的相关策略建议。

目 录

一、大数据产业发展概述.....	1
（一）大数据再认识	1
（二）大数据产业界定	2
（三）大数据关键问题	5
二、大数据技术发展趋势.....	6
（一）社交网络和物联网技术拓展了数据采集技术渠道	6
（二）分布式存储和计算技术夯实了大数据处理的技术基础	9
（三）深度神经网络等新兴技术开辟大数据分析技术的新时代	11
三、大数据资源开放与共享.....	15
（一）数据资源总量评估	15
（二）政府数据共享	16
（三）政府数据开放	19
（四）数据交易流通	20
四、重点行业大数据应用.....	27
（一）大数据应用整体情况	28
（二）各领域应用进展情况	28
（三）大数据应用发展趋势	39
五、大数据政策法规.....	40
（一）政府数据开放与信息公开	40
（二）个人数据保护	43
（三）跨境数据流动	46
（四）数据权属问题	48
六、结论与建议.....	50
（一）避免盲目跟风，大数据热潮还需冷思考	51
（二）推动开放共享，倒逼信息化建设升级	53
（三）强调供需对接，拉动技术产业跨越发展	55
（四）完善法律制度，切实保障数据安全	56
（五）突出地方特色，形成差异化的区域产业布局	58

一、大数据产业发展概述

（一）大数据再认识

大数据是新资源、新技术和新理念的混合体。从资源视角来看，大数据是新资源，体现了一种全新的资源观。1990 年以来，在摩尔定律的推动下，计算存储和传输数据的能力在以指数速度增长，每 GB 存储器的价格每年下降 40%。2000 年以来，以 Hadoop 为代表的分布式存储和计算技术迅猛发展，极大的提升了互联网企业数据管理能力，互联网企业对“数据废气”（Data Exhaust）的挖掘利用大获成功，引发全社会开始重新审视“数据”的价值，开始把数据当作一种独特的战略资源对待。大数据的所谓 3V 特征（体量大、结构多样、产生处理速度快）主要是从这个角度描述的。

从技术视角看，大数据代表了新一代数据管理与分析技术。传统的数据管理与分析技术以结构化数据为管理对象、在小数据集上进行分析、以集中式架构为主，成本高昂。与“贵族化”的数据分析技术相比，源于互联网的，面向多源异构数据、在超大规模数据集（PB 量级）上进行分析、以分布式架构为主的新一代数据管理技术，与开源软件潮流叠加，在大幅提高处理效率的同时（数据分析从 T+1 到 T+0 甚至实时），成百倍的降低了数据应用成本。

从理念的视角看，大数据打开了一种全新的思维角度。大数据的应用，赋予了“实事求是”新的内涵，其一是“数据驱动”，即经营管理决策可以自下而上地由数据来驱动，甚至像量化股票交易、实时竞价广告等场景中那样，可以由机器根据数据直接决策；其二是“数

据闭环”，观察互联网行业大数据案例，它们往往能够构造起包括数据采集、建模分析、效果评估到反馈修正各个环节在内的完整“数据闭环”，从而能够不断地自我升级，螺旋上升。目前很多“大数据应用”，要么数据量不够大，要么并非必须使用新一代技术，但体现了数据驱动和数据闭环的思维，改进了生产管理效率，这是大数据思维理念应用的体现。

（二）大数据产业界定

大数据本身既能形成新兴产业，也能推动其他产业发展。当前，国内外缺乏对大数据产业的公认界定。我们认为，大数据产业可以从狭义和广义两个层次界定。

从狭义看，当前全球围绕大数据采集、存储、管理和挖掘，正在逐渐形成了一个“小生态”，即大数据核心产业。大数据核心产业为全社会大数据应用提供数据资源、产品工具和应用服务，支撑各个领域的大数据应用，是大数据在各个领域应用的基石。应该注意到，狭义大数据产业仍然围绕信息的采集加工构建，属于信息产业的一部分。

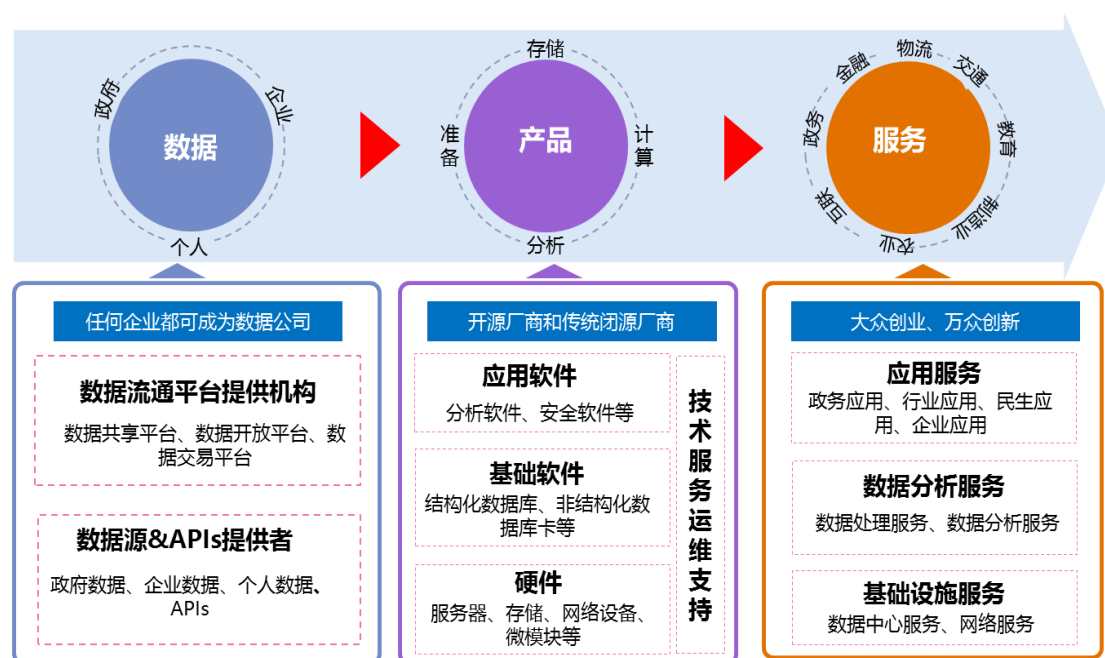


图1 大数据核心产业构成

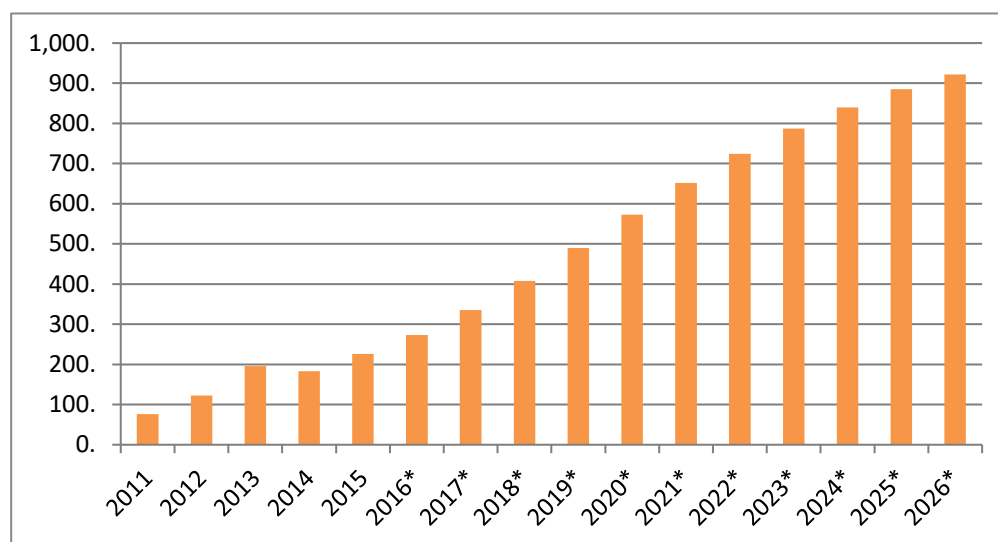
数据资源部分负责原始数据的供给和交换，根据数据来源的不同，可以细分为数据资源提供者和数据交易平台两种角色。

数据基础能力部分负责与数据生产加工相关的基础设施和技术要素供应，根据数据加工和价值提升的生产流程，数据基础能力部分主要包括数据存储、数据处理和数据库（数据管理）等多个角色。

数据分析/可视化部分负责数据隐含价值的挖掘、数据关联分析和可视化展现等，既包括传统意义上的 BI、可视化和通用数据分析工具，也包括面向非结构化数据提供的语音、图像等媒体识别服务。

数据应用部分根据数据分析和加工的结果，面向电商、金融、交通、气象、安全等细分行业提供精准营销、信用评估、出行引导、信息防护等企业或公众服务。

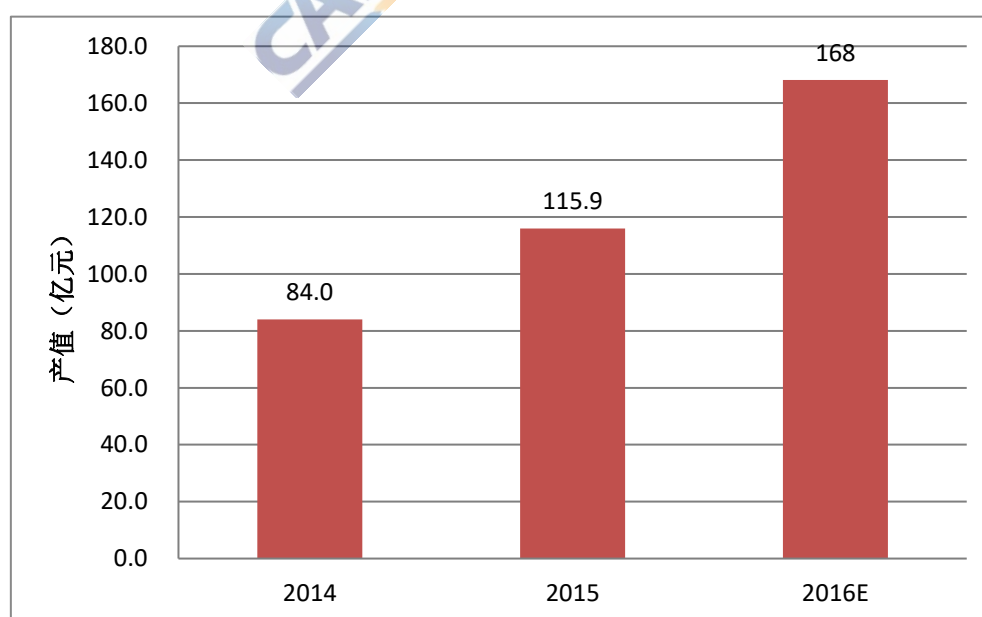
根据 IDC、Wikibon 等咨询机构预测，2016 年，全球的大数据核心产业规模约为 300 亿美元。



数据来源：Wikibon，2016 年 3 月，单位：亿美元

图 2 全球大数据产业规模（2011-2026）

目前大数据产业的统计口径尚未建立。对于我国大数据产业的规模，各个研究机构均采取间接方法估算。中国信息通信研究院结合对大数据相关企业的调研测算，2015 年我国大数据核心产业的市场规模达到 115.9 亿元，增速达 38%，预计 2016 年将达到 168 亿元，2017-2018 年还将维持 40%左右的高速增长。



数据来源：中国信息通信研究院，2016 年 8 月，单位：亿人民币

图 3 中国大数据产业规模估计

从广义看，大数据具有通用技术的属性，能够提升运作效率，提高决策水平，从而形成由数据驱动经济发展的“大生态”，即广义大数据产业。广义大数据产业包含了大数据在各个领域的应用，已经超出了信息产业的范畴。据华沙经济研究所测算，欧盟 27 国因大数据的引进，至 2020 年将获得 1.9% 的额外 GDP 增长。美国麦肯锡预计，到 2020 年美国大数据应用带来的增加值将占 2020 年 GDP 的 2%-4%。中国信息通信研究院预计，到 2020 年大数据将带动中国 GDP 2.8-4.2%。

（三）大数据关键问题

我国大数据产业发展已具备一定基础，但要实现从“数据大国”向“数据强国”转变，还面临诸多挑战。

一是对数据资源及其价值的认识不足。全社会尚未形成对大数据客观、科学的认识，对数据资源及其在人类生产、生活和社会管理方面的价值利用认识不足，存在盲目追逐硬件设施投资、轻视数据资源积累和价值挖掘利用等现象。

二是技术创新与支撑能力不够。大数据需要从底层芯片到基础软件再到应用分析软件等信息产业全产业链的支撑，无论是新型计算平台、分布式计算架构，还是大数据处理、分析和呈现方面与国外均存在较大差距，对开源技术和相关生态系统的影响力仍然较弱，总体上难以满足各行各业大数据应用需求。

三是数据资源建设和应用水平不高。用户普遍不重视数据资源的建设，即使有数据意识的机构也大多只重视数据的简单存储，很少针对后续应用需求进行加工整理。数据资源普遍存在质量差，标准规范

缺乏，管理能力弱等现象。跨部门、跨行业的数据共享仍不顺畅，有价值的公共信息资源和商业数据开放程度低。数据价值难以被有效挖掘利用，大数据应用整体上处于起步阶段，潜力远未释放。

四是信息安全和数据管理体系尚未建立。数据所有权、隐私权等相关法律法规和信息安全、开放共享等标准规范缺乏，技术安全防范和管理能力不够，尚未建立起兼顾安全与发展的数据开放、管理和信息安全保障体系。

五是人才队伍建设亟需加强。综合掌握数学、统计学、计算机等相关学科及应用领域知识的综合性数据科学人才缺乏，远不能满足发展需要，尤其是缺乏既熟悉行业业务需求，又掌握大数据技术与管理综合型人才。

二、大数据技术发展趋势

（一）社交网络和物联网技术拓展了数据采集技术渠道

经过行业信息化建设，医疗、交通、金融等领域已经积累了许多内部数据，构成大数据资源的“存量”；而移动互联网和物联网的发展，大大丰富了大数据的采集渠道，来自外部社交网络、可穿戴设备、车联网、物联网及政府公开信息平台的数据将成为大数据增量数据资源的主体。

当前，移动互联网的深度普及，为大数据应用提供了丰富的数据源。根据中国互联网络信息中心（CNNIC）第 38 次《中国互联网络发展状况统计报告》，截至 2016 年 6 月，我国网民规模达 7.1 亿，互

联网普及率达到 51.7%，超过全球平均水平 3.1 个百分点。其中，我国手机网民规模达 6.65 亿。网民中使用手机上网的人群占比提升至 92.5%。线下企业通过与互联网企业的合作，或者利用开放的应用编程接口（API, Application Programming Interface）或网络爬虫¹，可以采集到丰富的网络数据，可以作为内容数据的有效补充。

另外，快速发展的物联网，也将成为越来越重要的大数据资源提供者。相对于现有互联网数据杂乱无章和价值密度低的特点，通过可穿戴、车联网等多种数据采集终端，定向采集的数据资源更具利用价值。例如，智能化的可穿戴设备经过几年的发展，智能手环、腕带、手表等可穿戴正在走向成熟，智能钥匙扣、自行车、筷子等设备层出不穷，国外 Intel、Google、Facebook，国内百度、京东、小米等有所布局。根据 IDC 公司预计，到 2016 年底，全球可穿戴设备的出货量将达到 1.019 亿台，较 2015 年增长 29.0%。到 2020 年之前，可穿戴设备市场的年复合增长率将为 20.3%，而 2020 年将达到 2.136 亿台²。可穿戴设备可以 7×24 小时不间断地收集个人健康数据，在医疗保健领域有广阔的应用前景，一旦技术成熟，设备测量精度达到医用要求，电池续航能力也有显著增强，就很可能进入大规模应用阶段，从而成为重要的大数据来源。再如，车联网已经进入快速成长期。据 StrategyAnalytics 公司预计，2016 年前装车联网市场渗透率将达到 19%，在未来 5 年内迎来发展黄金期，2020 年将达到 49%³。

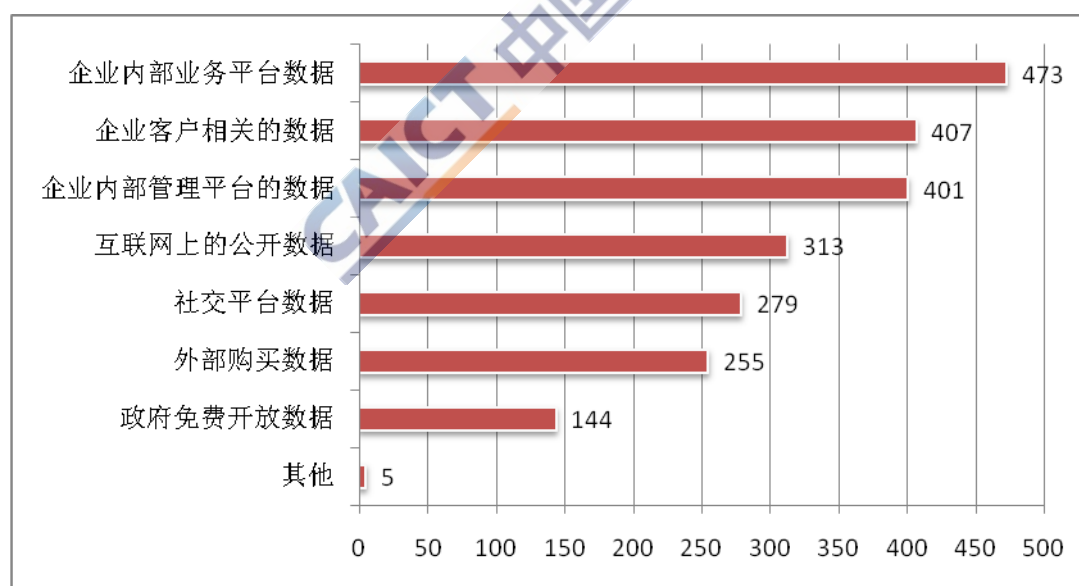
不过，值得注意的是，即便外部数据越来越丰富，但可获取性还

¹注释：网络爬虫（Web crawler），是一种按照一定的规则自动抓取互联网网页信息的计算机程序。

²<http://www.idc.com/getdoc.jsp?containerId=prUS41530816>

³<http://www.askci.com/news/dxf/20160727/15510447326.shtml>

不够高，一方面受目前技术水平所限，车联网、可穿戴设备等数据采集精度、数据清洗技术和数据质量还达不到实用要求；另一方面，由于体制机制原因，导致行业和区域上的条块分割，数据割据和孤岛普遍存在，跨企业跨行业数据资源的融合仍然面临诸多障碍。根据中国信息通信研究院 2015 年对国内 800 多家企业的调研来看，有 50% 以上的企业把内部业务平台数据、客户数据和管理平台数据作为大数据应用最主要的数据来源。企业内部数据仍是大数据主要来源，但对外部数据的需求日益强烈。当前，有 32% 的企业通过外部购买所获得的数据；只有 18% 的企业使用政府开放数据。如何促进大数据资源建设，提高数据质量，推动跨界融合流通，是推动大数据应用进一步发展的关键问题之一。



数据来源：中国信息通信研究院，2015 年 5 月

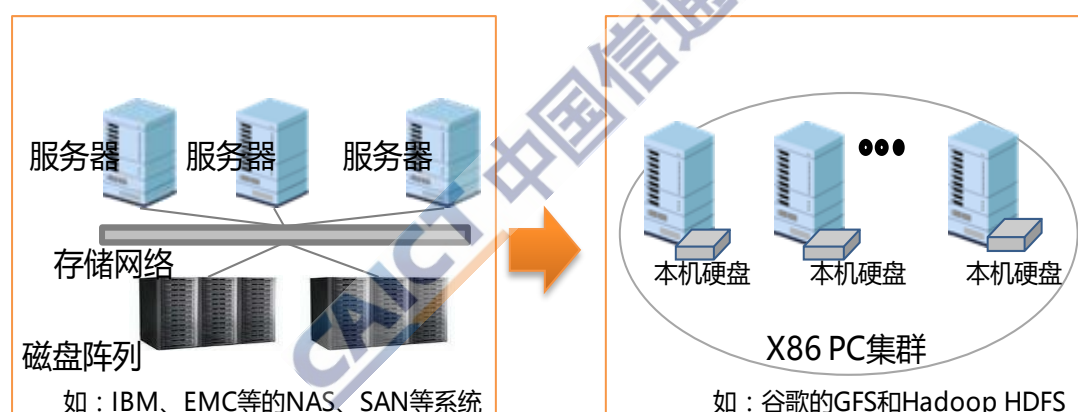
图 4 企业大数据来源情况（企业数量，n=809）

总体来看，各行业都在致力于在用好存量资源的基础之上，积极拓展新兴数据收集的技术渠道，开发增量资源。社交媒体、物联网等大大丰富了数据采集的潜在渠道，理论上，数据获取将变得越来越容

易。

（二）分布式存储和计算技术夯实了大数据处理的技术基础

大数据存储和计算技术是整个大数据系统的基础。在存储方面，2000 年左右谷歌等提出的文件系统（GFS）、以及随后的 Hadoop 的分布式文件系统 HDFS（Hadoop Distributed File System）奠定了大数据存储技术的基础。与传统系统相比，GFS/HDFS 将计算和存储节点在物理上结合在一起，从而避免在数据密集计算中易形成的 I/O 吞吐量的制约，同时这类分布式存储系统的文件系统也采用了分布式架构，能达到较高的并发访问能力。存储架构的变化如图 5 所示。



数据来源：中国信息通信研究院，2014 年

图 5 大数据存储架构的变化

在计算方面，谷歌在 2004 年公开的 MapReduce 分布式并行计算技术，是新型分布式计算技术的代表。一个 MapReduce 系统由廉价的通用服务器构成，通过添加服务器节点可线性扩展系统的总处理能力（Scale Out），在成本和可扩展性上都有巨大的优势。谷歌的 MapReduce 是其内部网页索引、广告等核心系统的基础。之后出现的 Apache Hadoop MapReduce 是谷歌 MapReduce 的开源实现，目前已经

成为应用最广泛的大数据计算软件平台。

MapReduce 架构能够满足“先存储后处理”的离线批量计算(batch processing)需求,但也存在局限性,最大的问题是时延过长,难以适用于机器学习迭代、流处理等实时计算任务,也不适合针对大规模图数据等特定数据结构的快速运算。为此,业界在 MapReduce 基础上,提出了多种不同的并行计算技术路线。如 Yahoo 提出的 S4 系统、Twitter 的 Storm 系统是针对“边到达边计算”的实时流计算(Real time streaming process)框架,可在一个时间窗口上对数据流进行在线实时分析,已经在实时广告、微博等系统中得到应用。谷歌 2010 年公布的 Dremel 系统,是一种交互分析(Interactive Analysis)引擎,几秒钟就可完成 PB 级数据查询操作。此外,还出现了将 MapReduce 内存化以提高实时性的 Spark 框架、针对大规模图数据进行了优化的 Pregel 系统等等。

以 Hadoop 为代表的开源软件大幅度降低数据的存储与计算的成本。传统数据存储和分析的成本约为 3 万美元/TB,而采用 Hadoop 技术,成本可以降到 300-1000 美元/TB。新一代计算平台 Spark 进一步把 Hadoop 性能提升了 30 多倍,性能越来越高,技术门槛越来越低。目前,开源 Hadoop 和 Spark 已经形成了比较成熟的产品供应体系,基本上可以满足大部分企业建设大数据存储和分析平台的需求,为企业提供了低成本解决方案。

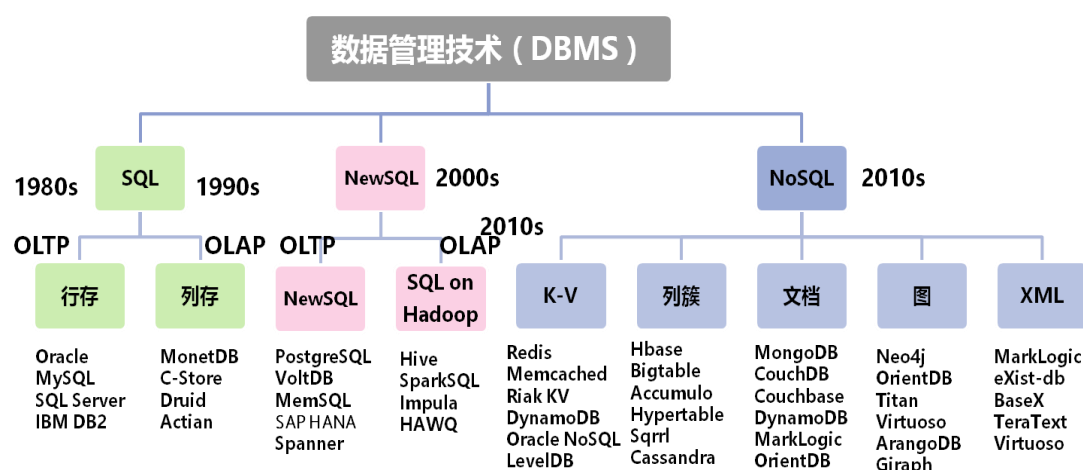


图6 数据管理技术图谱

（三）深度神经网络等新兴技术开辟大数据分析技术的新时代

大数据数据分析技术，一般分为联机分析处理（OLAP，Online Analytical Processing）和数据挖掘（Data Mining）两大类。OLAP 技术，一般基于用户的一系列假设，在多维数据集上进行交互式的数据集查询、关联等操作（一般使用 SQL 语句）来验证这些假设，代表了演绎推理的思想方法。

数据挖掘技术，一般是在海量数据中主动寻找模型，自动发展隐藏在数据中的模式（Pattern），代表了归纳的思想方法。传统的数据挖掘算法主要有：（1）聚类，又称群分析，是研究（样品或指标）分类问题的一种统计分析方法，针对数据的相似性和差异性将一组数据分为几个类别。属于同一类别的数据间的相似性很大，但不同类别之间数据的相似性很小，跨类的数据关联性很低。企业通过使用聚类分析算法可以进行客户分群，在不明确客户群行为特征的情况下对客户数据从不同维度进行分群，再对分群客户进行特征提取和分析，从而抓住客户特点推荐相应的产品和服务。（2）分类，类似于聚类，

但是目的不同，分类可以使用聚类预先生成的模型，也可以通过经验数据找出一组数据对象的共同点，将数据划分成不同的类，其目的是通过分类模型将数据项映射到某个给定的类别中，代表算法是 CART（分类与回归树）。企业可以将用户、产品、服务等各业务数据进行分类，构建分类模型，再对新的数据进行预测分析，使之归于已有类中。分类算法比较成熟，分类准确率也比较高，对于客户的精准定位、营销和服务有着非常好的预测能力，帮助企业进行决策。（3）回归，反映了数据的属性值的特征，通过函数表达数据映射的关系来发现属性值之间的一览关系。它可以应用到对数据序列的预测和相关关系的研究中。企业可以利用回归模型对市场销售情况进行分析和预测，及时作出对应策略调整。在风险防范、反欺诈等方面也可以通过回归模型进行预警。

传统的数据方法，不管是传统的 OLAP 技术还是数据挖掘技术，都难以应付大数据的挑战。首先是执行效率低。传统数据挖掘技术都是基于集中式的底层软件架构开发，难以并行化，因而在处理 TB 级以上数据的效率低。其次是数据分析精度难以随着数据量提升而得到改进，特别是难以应对非结构化数据。在人类全部数字化数据中，仅有非常小的一部分（约占总数据量的 1%）数值型数据得到了深入分析和挖掘（如回归、分类、聚类），大型互联网企业对网页索引、社交数据等半结构化数据进行了浅层分析（如排序），占总量近 60% 的语音、图片、视频等非结构化数据还难以进行有效的分析。

所以，大数据分析技术的发展需要在两个方面取得突破，一是对

体量庞大的结构化和半结构化数据进行高效率的深度分析，挖掘隐性知识，如从自然语言构成的文本网页中理解和识别语义、情感、意图等；二是对非结构化数据进行分析，将海量复杂多源的语音、图像和视频数据转化为机器可识别的、具有明确语义的信息，进而从中提取有用的知识。目前来看，以深度神经网络等新兴技术为代表的大数据分析技术已经得到一定发展。

神经网络是一种先进的人工智能技术，具有自身自行处理、分布存储和高度容错等特性，非常适合处理非线性的以及那些以模糊、不完整、不严密的知识或数据，十分适合解决大数据挖掘的问题。典型的神经网络模型主要分为三大类：第一类是以用于分类预测和模式识别的前馈式神经网络模型，其主要代表为函数型网络、感知机；第二类是用于联想记忆和优化算法的反馈式神经网络模型，以 Hopfield 的离散模型和连续模型为代表。第三类是用于聚类的自组织映射方法，以 ART 模型为代表。不过，虽然神经网络有多种模型及算法，但在特定领域的数据挖掘中使用何种模型及算法并没有统一的规则，而且人们很难理解网络的学习及决策过程。

深度学习是近年来机器学习领域最令人瞩目的方向。自 2006 年深度学习界泰斗 Geoffrey Hinton 在《Science》杂志上发表 *Deep Belief Networks* 的论文后，激活了神经网络的研究，开启了深度神经网络的新时代。学术界和工业界对深度学习热情高涨，并逐渐在语音识别、图像识别、自然语言处理等领域获得突破性进展，深度学习在语音识别领域获得 20%到 30%的准确率提升，突破了近十年的瓶颈。

2012 年图像识别领域在 ImageNet 图像分类竞赛中取得了 85% 的 top5 准确率，相比前一年 74% 的准确率有里程碑式的提升，并进一步在 2013 年将准确率提高到 89%。目前 Google、Facebook、Microsoft、IBM 等国际巨头，以及国内百度、阿里巴巴、腾讯等互联网巨头争相布局深度学习。由于神经网络算法的结构和流程特性，非常适合于大数据分布式处理平台进行计算，通过神经网络领域的各种分析算法的实现和应用，公司可以实现对多样化的分析，并在产品创新、客户服务、营销等方面取得创新性进展。

随着互联网与传统行业融合程度日益加深，对于 web 数据的挖掘和分析成为了需求分析和市场预测的重要手段。Web 数据挖掘是一项综合性的技术，可以从文档结构和使用集合中发现隐藏的输入到输出的映射过程。目前研究和应用比较多的是 PageRank 算法。PageRank 是 Google 算法的重要内容，于 2001 年 9 月被授予美国专利，以 Google 创始人之一拉里·佩奇（Larry Page）命名。PageRank 根据网站的外部链接和内部链接的数量和质量衡量网站的价值。这个概念的灵感，来自于学术研究中的这样一种现象，即一篇论文的被引述的频度越多，一般会判断这篇论文的权威性和质量越高。在互联网场景中，每个到页面的链接都是对该页面的一次投票，被链接的越多，就意味着被其他网站投票越多。这就是所谓的链接流行度，可以衡量多少人愿意将他们的网站和你的网站挂钩。让机器自动学习和理解人类语言中的近百万种语义、并从海量用户行为数据汇总归纳用户兴趣是一个已经持续 20 多年的研究方向。腾讯效果广告平台部研发的 Peacock 大规模

主题模型机器学习系统，通过并行计算可以高效的对 10 亿*1 亿的大规模矩阵进行分解，从海量样本数据中学习 10 万到 100 万两级的隐含语义。这对于挖掘用户兴趣、相似用户扩展，精准推荐具有重大意义。

需要指出的是，数据挖掘与分析的行业与企业特点强，除了一些最基本的数据分析工具（如 SAS）外，目前还缺少针对性的、一般化的建模与分析工具。各个行业与企业需要根据自身业务构建特定数据模型。数据分析模型构建的能力强弱，成为不同企业在大数据竞争中取胜的关键。

三、大数据资源开放与共享

（一）数据资源总量评估

未来五年，全球数据量呈指数级增长。据国际数据公司（IDC）统计，2014 年全球数据总量为 8ZB，预计 2020 年达到 44ZB。同期，我国数据总量为 909EB，占全球数据总量的 13%。其中，媒体、互联网数据量占比为 1/3，政府部门、电信企业数据量占比为 1/3，其他的金融、教育、制造、服务业等数据量占比为 1/3。预计到 2020 年我国数据量将达到 8060EB，占全球数据总量的 18%⁴。

我国具有天然的大数据规模优势。信息技术与经济社会的交汇融合引发了数据迅猛增长，数据成为物理世界在网络空间的客观映射，如同工业时代的钢铁、石油，已成为新的生产要素和战略资源。我国

⁴<https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014-china.pdf>

巨大的人口基数以及经济规模，具有形成大规模数据的天然优势。截至 2016 年 6 月，我国网民规模已达 7.10 亿，互联网普及率达到 51.7%，网站数量为 454 万个。丰富的数据资源，构成了我国推进大数据应用的资源基础。

（二）政府数据共享

推进政府数据资源开放共享是实施大数据战略的关键，也就是着力解决“不愿开放共享”、“不敢开放共享”、“不会开放共享”问题，打破部门分割和行业壁垒，促进互联互通、数据开放、信息共享和业务协同，切实以数据流引领技术流、物资流、资金流、人才流，强化统筹衔接和条块结合，实现跨部门、跨区域、跨层级、跨系统的数据交换与共享，构建全流程、全覆盖、全模式、全响应的信息化管理与服务体系。从“十五”计划起，跨部门信息共享一直被各部门和各级政府列为重要课题，但几大难题始终未能解决。

基础信息库总体进展缓慢。2002 年《国家信息化领导小组关于我国电子政务建设指导意见》提出规划和开发重要政务信息资源，并启动建设政务信息化四大基础数据库，即人口基础信息库、法人单位基础信息库、自然资源和空间地理基础信息库、宏观经济数据库。从全国范围来看，四个数据库建设进度不同，除自然资源和空间地理基础数据库已基本建成外，大部分地方政府的人口基础数据库和法人单位基础数据库建设进程缓慢，而宏观经济基础数据库几乎处于搁置状态。

金字工程信息孤岛严重。以“十二金工程”为代表的电子政务重点工程项目为政府核心业务提供了信息化支撑，但各个部委、各级政府分散建设的信息系统形成的信息化壁垒很高，信息孤岛、信息烟囱现象严重。国务院《促进大数据发展行动纲要》提出到 2018 年，中央政府层面实现金税、金关、金财、金审、金盾、金宏、金保、金土、金农、金水、金质等信息系统通过统一平台进行数据共享和交换。

信息共享和业务协同尚未取得根本突破。中央和部分省市在综合治税、人口管理、应急管理等方面积极推进信息共享和业务协同，共享内容和范围不断扩大，业务协同能力不断增强，取得了一定成效。但从全国总体来看，跨部门、跨地区的共享协同尚未取得根本突破。数据显示，区域部门间基本实现共享的省级地方仅占 13%，区域部门间少量实现共享的地市和区县仅占 32% 和 28%，信息共享和业务协同在地市和区县进展缓慢，信息共享成为制约部门业务协同的重要因素。

在当前以简政放权为核心，加快转变政府职能的改革背景下，政府数据共享需求迫切、意义重大。建设一体化政务服务平台，打通后台数据流动环节，“让数据多跑路、百姓少跑腿”。“证明我妈是我妈”、“老年证丢失找派出所开证明”、异地办理准生证跑断腿儿……这些让人“添堵”的证明或将成为历史。

为加快推动信息共享工作，国家发改委按照“统一平台、互联互通，存量共享、增量共建，物理分散、逻辑集中”的原则，以开放数据交换接口的方式，推动政府部门间的信息共享，已取得初步成效。目前，全国统一的国家电子政务外网已初步建成，横向连接了 118 个

中央单位和 14.4 万个地方单位，纵向基本覆盖了中央、省、市、县四级，承载了 47 个全国性业务系统和 5000 余项地方业务系统。依托国家电子政务外网搭建的全国统一的国家数据共享交换平台基本建成，13 个行业领域的跨部门共享交换业务已通过或拟通过国家数据共享交换平台实现，涉及部门超过 100 个。国务院《促进大数据发展行动纲要》进一步提出要在 2017 年底前形成跨部门数据资源共享共用格局；在 2018 年底前建成国家政府数据统一开放平台。

除了以发改委为代表的中央政府，数据资源整合需求方和实施的另一个重要主体是城市。目前全国650多个城市中近有 $\frac{2}{3}$ 的城市提出了智慧城市的计划，智慧城市建设和发展的核心就是基于城市信息资源的整合和利用。智慧城市将推动政务数据在内的城市公共信息共享，形成城市数据交换共享平台、GIS平台和信息资源目录库，实现不同职能部门之间的业务协同和信息共享、信息资源社会化开放与利用，有利于创新社会治理模式，推动形成“用数据说话、用数据决策、用数据管理、用数据创新”的城市管理新方式。北京市从2006年开始，61个市政部门通过开展流程和协同工作清、网上服务清、信息资源清、实现路径清，统一平台、统一网络的“四清两统一”工作，进行业务梳理和资源梳理，编制信息资源目录。基于这项工作，北京市建成了统一的信息化基础设施共享交换平台，各委办局在该平台上每天进行大量的数据交换。另外，有先见的地方政府已经看到大数据带动地方经济的发展机遇，试图通过政府带头，打造大数据基础设施与共享平台，拉动相关产业发展，带动传统产业升级。

贵州省建

设“云上贵州”平台，成为全国第一个实现省级政府、企业和事业单位数据整合和互通共享的云服务平台，致力打造成为全国的大数据运算中心和交易中心。

（三）政府数据开放

政府数据资源是大数据资源的重要组成部分。近年来，随着互联网与各领域的深度融合以及数据资源战略价值的日益凸显，国际社会高度重视数据资源的开放与利用，将其视作促进互联网产业创新，支撑新兴业态发展的必备要素。政府数据资源可以与社会数据资源互为补充，服务于新兴业态的发展。政府数据资源基于公共事务管理和公共服务采集和产生，具有较强的公信力，甚至可能是唯一的数据来源，能够促进简单或片面的数据资源进行深度挖掘利用。政府数据资源采集和产生已经付出了财政成本，在政府利用之余“一次投入，全民利用”，能够降低全社会的数据资源利用成本，促进企业产品产出和社会福利提升。

做优存量、做大增量是数据资源开发利用的基本和基础。在近年政府数据资源开放大趋势下，一些部门、地区也开始了自发探索，气象、统计、环境、交通等十余类数据不同程度面向社会开放，北京、上海、青岛、无锡等地开放数据得到不同程度开发利用。2015 年，

《国务院关于积极推进“互联网+”行动的指导意见》和《促进大数据发展行动纲要》等文件从国家层面明确要求推进政府数据资源开放，带动社会数据资源开放，做大做强国家大数据资源，夯实培育新兴业态的数据资源基础。

政府数据资源开放拓展了信息服务企业的数据来源，发挥了数据开发利用倍增效应，在我国最早推动政府数据资源开放的北京和上海，数据资源培育新业态发展的趋势已经显现。截止 2016 年 6 月，北京市政府数据开放网站共开放数据集 321 个，涉及 39 个政府部门和公共机构。同期，上海共开放数据集 735 个，涉及 40 个政府部门和公共机构。企业、个人利用开放数据成功开发了近百个互联网、移动互联网应用，典型代表包括口口安全（食品安全应用）、阿拉自来水（水质监控查询应用）、上海空气质量（空气监控查询应用）、上海防汛（防汛查询应用）、交通英雄（交通出行应用）等，涉及到公众的安全、出行、入学、就业、医疗等各方面信息服务。

目前数据资源开放在开放范围、开发利用方式、开放模式和标准等方面仍存在局限性和不足。一是数据资源开放具有行业和区域的局限性，开放集中在经济发达东部地区和信息化基础较好的行业，数据资源开放依赖地方和部门自行推动，数据资源开放地域、行业局限性强。二是已经开放的数据资源开发利用方式单一，综合利用困难，受到数据开放局限性的影响，数据资源开发利用内容和方式比较单一，不具备综合开发利用的基础，仍需加快开放经济价值高、社会需求大的数据资源，不断扩大数据资源开放的范围。

（四）数据交易流通

1、国内外大数据交易现状

数据作为一种资源，有着重要的价值。随着数据的资源价值逐渐

得到认可，数据交易的需求不断增加。2015 年国务院印发的《促进大数据发展行动纲要》中明确指出，“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，鼓励产业链各环节的市场主体进行数据交换和交易，促进数据资源流通，建立健全数据资源交易机制和定价机制，规范交易行为等一系列健全市场发展机制的思路与举措”。

国际上，数据交易大致始于 2008 年左右，一些具有前瞻性的企业开始加大对数据业务的投入，“数据市场”、“数据银行”、“数据交易公约”等数据应用新业态已初见端倪。国外的数据交易形式主要为数据中介公司通过政府、公开和商业渠道，从数据源头处获取各类信息，进而向用户直接交付数据产品或服务。其中，数据源头、数据中介和最终用户构成了数据流通和交易的主体。例如，Twitter 公司将自身数据授权给公司 Gnip、DataSift 和 NTT DATA 进行售卖；Acxiom 等公司通过各种手段收集、汇聚关于企业和个人的信息；Sermo.com 和 Inrix 等公司则通过网络和传感器直接从公众采集数据，获得了传统上单个企业难以采集的海量、实时数据。

近年来，国际上各种数据相关平台都选择了自己有所侧重的数据类型，不再以“综合性”为主要策略。例如，Datamarket 公司以国民经济与工业相关的数据集为主；InfoChimps 公司在地理位置、社交网络、网络信息等方面的数据更为突出，且逐渐转型为 PaaS 平台；Factual 公司从提供全范围的数据交易平台转为专注于提供地理位置相关的数据集。

从国内来看，中国信息消费市场规模量级巨大，增长迅速。中国潜在的大数据资源非常丰富，从电信、金融、社保、房地产、医疗、政务、交通、物流、征信体系等部门，到电力、石化、气象、教育、制造等传统行业，再到电子商务平台、社交网站等，覆盖广泛。如果数据交易行业可以得到充分、健康发展，必将对国民经济各个方面起到积极的影响。然而，尽管中国当前大数据存储和挖掘技术已经逐步成熟，但“数据孤岛”的大量存在，制约了数据的流通和变现。在大数据时代要实现商业价值变现，需要实时对接数据市场的多样化需求，而平台化运营成为满足这一产业需求的必要条件。唯有将数据进行合理定价，出现数据交易市场、交易指数，才能真正带动大数据产业的繁荣。大数据实现交易，将打破行业信息壁垒，优化提高生产效率，深度推进产业创新。这正是大数据交易平台最核心的价值和意义所在。

中国已经有地方政府相继试水大数据应用和建立交易市场，众多企业也在积极投资布局。2015 年 4 月 14 日，全国首个大数据交易所——贵阳大数据交易所正式挂牌运营并完成首批大数据交易。由上海经济和信息化委员会指导的上海大数据交易中心也将于 2016 年 4 月 1 日挂牌成立。此外，北京数海科技、数据堂、TalkingData、中关村大数据产业联盟等企业和产业联盟在数据交易流通也走在了行业前列。

大数据应用也给云计算带来落地的途径，使得基于云计算的业务创新和服务创新成为现实。大数据的应用不仅仅停留在 IT 领域，在医药、科学、制造以及气象等行业，都出现海量的数据应用，如果能

合理的利用这些资源，对行业将带来巨大的推动，但目前来看，大数据流通率仍然较低，数据交易潜力有待进一步挖掘。未来几年，大数据也将走向更多应用实践并拓宽到更多行业。

2. 隐私保护与行业自律

数据交易无法回避的问题就是隐私保护问题。如何在交易过程中保障用户隐私，成为决定行业发展最为关键的问题之一。

从国际来看，许多发达国家已经形成了非常发达的个人数据保护行业自律组织体系。例如，美国电子隐私信息中心（EPIC）是从事数据隐私权保护活动的行业自律组织；美国在线隐私联盟（OPA, Online Privacy Alliance）旨在为通过互联网直接收集他人个人数据提供广为接受的规范指引；美国 TRUSTe 组织已发展为美国著名的隐私权保护第三方认证机构之一。这些组织对于数据流通等环节均建立了较为规范的自律准则，以督促企业在进行数据交易时确保对个人隐私权的保护。欧盟《通用数据保护条例》（General Data Protection Regulation, GDPR，以下简称《条例》）将在 2018 年 5 月 25 日正式生效。相对于 1995 年欧盟颁布的《个人数据保护指令》，《条例》规范了各成员国的法律标准，进一步保护了用户的数据安全。例如，对于认定“用户同意”的标准，《条例》规范的更加严格（必须是具体的、清晰的，是用户在充分知情的前提下自由做出的）。

尽管国外有着较为成熟的数据开放、隐私保护等方面的法律体系，但在数据交易方面国外的法律体系也仍亟需健全。尤其是数据交易面临的一些现实问题与现有隐私保护的法律存在冲突，也都需要逐渐加

以完善。

从国内来看，目前我国没有专门的隐私保护法律，可适用的法律法规主要有《宪法》、《关于加强网络信息保护的決定》、《刑法》、《治安管理处罚法》、《未成年人保护法》、《居民身份证法》、《商业银行法》、《邮政法》、《档案法》、《电信和互联网用户个人信息保护规定》等，这些法律对用户的个人隐私和通信自由作了零散规定。不可否认，我国已经开始高度重视隐私保护，自 2012 年以来连续三年通过了隐私保护的相关法规，尤其是 2014 年 10 月颁布的《关于审理利用信息网络侵害人身权益民事纠纷案件适用法律若干问题的规定》涉及到对敏感数据的保护，是立法的飞跃性突破。但我国仍然存在效力层级低、法律法规协调性弱、保护内容片面等立法不足，有待于相关部门更深层次的研究和改善。由于数据行业发展日新月异，且往往实践领先于监管，因而长期以来行业诸多领域存在“无门槛、无标准、无监管”的三无状态。由于法律法规的特性，一项行业标准成为国家标准，或行业监管措施以法律的形式确定下来，需要经过长期的调查、研究、讨论、试行等步骤，需要较长的时间。在相关法律法规完善前，行业的自律管理就显得尤为重要。

目前，随着我国数据流通行业的发展，数据交易平台已经出台了自身的数据交易规则或自律准则。2014 年 6 月，中国第一份大数据交易规则——《中关村数海大数据交易平台规则（征求意见稿）》在中关村大数据交易产业联盟专家顾问委员会宣布成立当天同步推出。《规则》从交易平台、交易主体、交易对象三个方面规范交易市场行

为，并对在线数据交易、离线数据交易、托管数据交易等三种数据交易模式进行规范。2015 年 5 月，贵阳大数据交易所发布《贵阳大数据交易所 702 公约》。规则规定，大数据交易平台要为数据交易提供可靠运行平台和安全保障措施；从事数据交易的主体包括数据提供方、数据购买方及数据代理人；交易数据的种类包括政府、医疗、金融、电商、能源、交易、交通、商品、消费、信用卡、教育、社交、社会等各方面；数据交易平台要求交易数据合法获取、权利清晰，禁止对法律不允许交易的数据进行交易；对于争议解决，采取自行协商解决、平台解决或起诉至法院等方式。

可以说，目前我国建立广泛的数据流通行业自律公约的时机已经相对成熟，行业内部各企业对数据交易自律性协议的需求呼之欲出。基于以上原因，中国信息通信研究院牵头，以数据中心联盟为依托，在 2016 年 4 月起草并推出了“数据流通行业自律公约”（1.0 版本），并于 2016 年 7 月推出了经过修订的“数据流通行业自律公约”（2.0 版本）。“数据流通行业自律公约”（以下简称“公约”）的推出，将作为一个有力的平台，帮助参与单位共同维护良好的数据流通生态环境，共同推动大数据产业发展。

3、我国大数据交易面临的问题

在《促进大数据发展行动纲要》等一系列文件政策的推动下，我国大数据交易呈现出百花齐放、蓬勃发展之势。但总体来看，我国大数据交易还处于起步阶段，仍面临着一些问题。

一是数据的产品化困难。作为一种新的交易品种，数据交易目前

还在产品化上存在一些障碍。首先，在数据标准化方面。交易所产品的重要特点就是交易产品的标准化。而大数据由于数据种类繁多，格式多样，难以形成一种普适的标准化方法，直接影响到其成为一种集中化、大规模交易的产品。其次，在数据处理方式的适当性方面。大多数企业在数据交易中，并不倾向于选择原始数据，而是选择经过一定处理后的数据。数据需求方在拿到经过处理后的数据时，缺乏对数据的信度进行判断的手段。在这个过程中，一旦出售方的数据处理方式存在问题，则对数据需求方就可能产生误导。而在现实应用中，不同的业务场景往往需要不同的数据处理方式，数据的出售方若不熟悉需求方的业务场景，则可能选择并不适当的数据处理方式。最后，数据的时效性与敏感性存在矛盾。作为数据交易的需求方，这些企业往往是需要数据进行产品研发或市场研判，这些用途都对数据的时效性有一定的要求。而作为数据交易的供给方，这些企业出售的数据往往来自于自身的经营活动，如果过早地将数据对外出售，则很可能泄露自身的商业秘密。需求方对数据时效性的要求，和供给方对数据敏感性的保护，是一组天然的矛盾。数据的时效性导致数据极易贬值。如上文所述，数据的需求方往往对数据的时效性有一定要求。不同于股权和一般商品，数据的价值会较为确定地随时间而贬值。

二是产品定价困难。因为数据价值因人而异，因此以经济效益最大化角度考虑，理应在不同行业采取歧视定价；但是，这样则很难形成统一的市场价格。另一方面，若以竞价的形式进行数据拍卖，则会阻碍数据的广泛应用，与数据交易所初衷背道而驰。

三是交易平台的交易机制缺乏。由于数据易复制、易传播、估值困难等因素，数据的交易机制并不能照搬金融交易所和商品交易所的模式。从买卖方来说，传统交易所的连续竞价，是多对多关系，而数据交易是天然的一对一或一对多交易。从世界范围来看，目前尚未形成成熟的、可大规模商用的数据集中撮合交易模式，按交易所模式组织的大数据交易机制仍待探索。

四是隐私及版权保护、信息安全面临挑战。在大数据交易中的许多标的都是基于以个人为粒度的数据。即使这些数据经过了一定的清洗，但也很难保证个人的隐私不被泄露。更为令人担忧的是，将此类数据在用户不知情的情况下出售给第三方，可能引来知识产权的纠纷，同时也需要防止买方未经授权地转售数据资产。另外，黑客攻击、病毒危害、恶意篡改等信息安全问题也成为困扰大数据交易各方的隐患。

四、重点行业大数据应用

传统的数据应用主要集中在对业务数据的统计分析，作为系统或企业的辅助支撑，应用范围以系统内部或企业内部为主，例如各类统计报表、展示图表等。伴随着各种随身设备、物联网和云计算、云存储等技术的发展，数据内容和数据格式多样化，数据颗粒度也愈来愈细，随之出现了分布式存储、分布式计算、流处理等大数据技术，各行业基于多种甚至跨行业的数据源相互关联探索更多的应用场景，同时更注重面向个体的决策和应用的时效性。因此，大数据的数据形态、处理技术、应用形式构成了区别于传统数据应用的大数据应用。

（一）大数据应用整体情况

大数据在各个领域的应用持续升温。据 Gartner 公司 2015 年的调研，全球范围内已经或未来 2 年计划投资大数据应用的企业比例达到 76%，比 2014 年增长 3%。中国信息通信研究院 2015 年的调查显示中国地区的受访企业中有 32% 的企业已经实现了大数据应用，另有 24% 的企业正在部署大数据平台。

另一方面，大数据的效益尚未充分验证。大多数的大数据系统尚处于早期部署阶段，因此它们的投资回报还未得到充分验证，比如 Wikibon 公司 2014 年的统计显示，美国企业的高层管理人员期望大数据能够带来总计 3.5 倍的投资回报，但实际回报当时只能达到 55%。

总体来看，大数据应用尚处发展前期阶段，应用快速部署，效益有待检验。大数据前景很美好，同时也可能存在“忽悠”出来的“泡沫”成分。

（二）各领域应用进展情况

整体来看，大数据应用尚处于从热点行业领域向传统领域渗透的阶段。中国信息通信研究院的调查显示大数据应用水平较高的行业主要分布在互联网、电信、金融行业，一些传统行业的大数据应用发展较为缓慢，批发零售业甚至有超过 80% 的企业并没有大数据应用计划，远低于整体平均水平。

1. 电信领域

电信行业掌握体量巨大的数据资源，单个运营商其手机用户每天

产生的话单记录、信令数据、上网日志等数据就可达到 PB 级规模。电信行业利用 IT 技术采集数据改善网络运营、提供客户服务已有数十年的历史，而传统处理技术下运营商实际上只能用到其中百分之一左右的数据。

大数据对于电信运营商而言，首先意味着利用廉价便捷的大数据技术提升其传统的数据处理能力，聚合更多的数据提升洞察能力。比如法国电信、T-Mobile 借助大数据加快了诊断网络潜在问题的效率，改善服务水平，为客户提供了更好的体验，获得了更多的客户以及更高的业务增长。中国移动、德国电信、沃达丰利用大数据技术加大对历史数据的分析，动态优化调整网络资源配置，大幅提高无线网络的运行效率。T-Mobile 通过集成数据综合分析客户流失原因，在一个季度内将客户流失率减半。SK 电讯成立 SK Planet 公司专门处理与大数据相关的业务，通过分析客户的使用行为防止客户流失。中国联通利用大数据技术对其全国 3G/4G 用户进行精准画像，形成大量有价值的标签数据，为客户服务和市场营销提供了有力支持。中国移动通过对消费、通话、位置、浏览、使用和交往圈等数据的分析，利用各种联系记录发现各种圈子，分析影响力及关键人员，用来进行家庭客户、政企客户和关键客户的识别，以实现主动营销和客户维系。

二是提高数据意识，寻求合适的商业模式，尝试数据价值的外部变现。主要有数据即服务（DaaS）和分析即服务（AaaS）两种模式，数据即服务（DaaS）模式往往通过开放数据或开放 API 的方式直接向外出售脱敏后的数据；分析即服务（AaaS）模式往往与第三方公司合

作，利用脱敏后的（自身或整合外部）数据资源为政府、企业或行业客户提供通用信息、数据建模、策略分析等多种形式的信息和服务，以创造外部收益，实现数据资源变现。

数据即服务方面，AT&T 将客户在 WiFi 网络中的地理位置、网络浏览历史记录以及使用的应用等数据销售给广告公司可以获取客观收益；AT&T 同时还提供 Alert 业务，当客户距离商家很近时，就有可能收到该商家提供的折扣很大的电子优惠券；英国电信基于安全数据分析服务 Assure Analytics，帮助企业收集、管理和评估大数据集，将这些数据通过可视化的方式呈现给企业，帮助企业改进决策；德国电信和沃达丰主要尝试通过开放 API，向数据挖掘公司等合作方提供部分用户匿名地理位置数据，以掌握人群出行规律，有效的与一些 LBS 应用服务对接。限于国内对数据交易流通方面缺乏明确规定，国内运营商很少尝试数据即服务（DaaS）模式。

分析即服务方面，西班牙电信成立动态洞察部门 Dynamic Insights 开展大数据业务，与市场研究机构 Gfk 进行合作，在英国、巴西推出名为智慧足迹（Smart Steps）的创新产品，该产品基于完全匿名和聚合的移动网络数据，可对某个时段、某个地点人流量的关键影响因素进行分析，并将洞察结果面向政企客户提供；Verizon 成立精准营销部门 Precision Marketing Division，提供精准营销洞察、精准营销、移动商务等服务，包括联合第三方机构对其用户群进行大数据分析，再将有价值的信息提供给政府或企业获取额外价值；中国电信在大数据 RTB 精准广告业务（根据客户行为和位置分析进行商

铺选址和实施营销）、景区流动人口监测业务、基于客户行为的中小微企业通用信用评价等方面均有尝试，且成效显著，借助对不同行业、不同类型企业的行为数据分析，中国电信的“贷 189”平台，一个月吸引中小企业 580 家，金融机构 24 家，订单成交 3368 万元。中国移动和中国联通也与第三方合作，开展智慧旅游、智能交通、智慧城市等项目，探索数据外部变现的新型商业模式，寻找新的业务增长点。

2. 金融领域

金融行业是信息产业之外大数据的又一重要应用领域，大数据在金融三大业务——银行、保险和证券中均具有较为广阔的应用前景。总体说来，金融行业的主要业务应用包括企业内外部的风险管理、信用评估、借贷、保险、理财、证券分析等，都可以通过获取、关联和分析更多维度、更深层次的数据，并通过不断发展的大数据处理技术得以更好、更快、更准确的实现，从而使得原来不可担保的信贷可以担保，不可保险的风险可以保险，不可预测的证券行情可以预测。

利用大数据可以提升金融企业内部数据分析能力。中信银行信用卡中心从 2010 年开始引入大数据分析解决方案，为企业中心提供了统一的客户视图。借助客户统一视图，可以从交易、服务、风险、权益等多个层面获取和分析数据，对客户按照低、中、高价值来进行分类，根据银行整体经营策略积极地提供相应的个性化服务，在降低成本的同时大幅提升精准营销能力。

更多的金融企业利用大数据技术整合来自互联网等渠道的更大的外部数据。

淘宝网的“阿里小贷”依托阿里巴巴（B2B）、淘宝、支付宝等平台数据，海量的交易数据在阿里的平台上运行，阿里通过对商户最近 100 天的数据分析，准确把握商户可能存在的资金问题。美国的 Lending Club 通过获取 ebay 等公司的网店店主的销售、信用记录、顾客流量、评论、商品价格和存货等信息，以及他们在 Facebook 和 Twitter 上与客户的互动信息，借助数据挖掘技术，把这些店主分成不同的风险等级，以此来确定提供贷款金额数量与贷款利率水平。宜信的互联网金融产品就是以互联网为获客主要渠道，除了借贷信用记录，还结合大数据分析技术，捕捉来自大众点评、豆瓣等社交网络上的有用信息，帮助信用审核人员多维度分析借款客户的信用状况。

众安保险依托阿里云服务，包括存储、处理和分析（ODPS）服务，同时不断改进其数据分析模型和挖掘手段，构建了强大的大数据能力，推出了针对高频小额事件的运费险。国内一款互联网车险产品利用手机获取车主驾驶行为的数据，结合车型因子、违章历史数据、个人信用数据等维度信息，对车主安全行为画像，从而进行风险定价。

IBM 使用大数据信息技术成功开发了“经济指标预测系统”，可通过统计分析新闻中出现的单词等信息来预测股价等走势。另外英美甚至国内都有基于社交网络的证券投资的探索，根据从 Twitter、微博等社交网络数据内容感知的市场情绪来进行投资。

3. 政务领域

大数据政务应用获得世界各国政府日益重视。美国 2012 年启动“大数据研究和发展计划”，联合国 2012 年推出“数据脉动”计划，

日本 2013 年正式公布以大数据为核心的新 IT 国家战略。英国政府通过高效的使用公共大数据的技术每年可以节省 330 亿英镑，相当于英国人每人每年节省 500 英镑（约每人每年节约 5000 人民币）。我国政府也非常重视利用大数据提升国家治理能力。《国务院关于印发促进大数据发展行动纲要的通知》（国发〔2015〕50 号）提出“大数据成为提升政府治理能力的新途径”，要“打造精准治理、多方协作的社会治理新模式”。

首先，大数据有助于提升政府提供的公共产品和服务。一方面，基于政务数据共享互通，实现政务服务一号认证（身份认证号）、一窗申请（政务服务大厅）、一网办事（联网办事），大大简化办事手续。另一方面，通过建设医疗、社保、教育、交通等民生事业大数据平台，有助于提升民生服务，同时引导鼓励企业和社会机构开展创新应用研究，深入发掘公共服务数据，有助于激发社会活力、促进大数据应用市场化服务。

其次，大数据支持宏观调控科学化。政府通过对各部门、社会企业的经济相关数据进行关联分析和融合利用，可以提高宏观调控的科学性、预见性和有效性。比如电商交易、人流、物流、金融等各类信息的融合交汇可以绘出国家经济发展的气象云图，帮助人们了解未来经济走向，提前预知通货膨胀或经济危机。

第三，大数据有助于政府加强事中事后监管和服务，提高监管和服务的针对性、有效性。《国务院办公厅关于运用大数据加强对市场主体服务和监管的若干意见》（国办发〔2015〕51 号）提出四项主

要目标：一是提高政府运用大数据能力，增强政府服务和监管的有效性；二是推动简政放权和政府职能转变，促进市场主体依法诚信经营；三是提高政府服务水平和监管效率，降低服务和监管成本；四是实现政府监管和社会监督有机结合，构建全方位的市场监管体系。“大数据综合治税”、“大数据信用体系”等以大数据融合加强企业事中事后监管的新模式的探索正在全国各地展开。

最后，大数据有助于推动权利管控精准化。借助大数据实现政府负面清单、权利清单和责任清单的透明化管理，完善大数据监督和技术反腐体系，促进政府依法行政。李克强 2014 年 2 月考察北京·贵阳大数据应用展示中心，了解贵阳利用执法记录仪和大数据云平台监督执法权力情况时说，要把执法权力关进“数据铁笼”，权力运行处处留痕，实现“人在干、云在算”。

大数据超越了传统行政思维模式，推动政府从“经验治理”转向“科学治理”。随着国家大数据战略渐次明细，各方实践逐步展开，大数据在政府领域的应用将迎来高速发展。

4. 交通领域

交通数据资源丰富、具有实时性特征。在交通领域，数据主要包括各类交通运行监控、服务和应用数据，如公路、航道、客运场站和港口等视频监控数据，城市和高速公路、干线公路的各类流量、气象检测数据，城市公交、出租车和客运车辆卫星定位数据，以及公路和航道收费数据等，这些交通数据类型繁多，而且体积巨大。此外，交通领域的数据采集和应用服务均对实时性要求较高。目前，大数据技

术在交通运行管理优化、面向车辆和出行者的智能化服务，以及交通应急和安全保障等方面都有着重大发展。

在出行方面，面向公众出行信息需求，整合交通出行服务信息，在公共交通、出租汽车、道路交通、公共停车，以及公路客运等领域扩大信息服务覆盖面，使公众出行更便捷。可以提供综合性、多层次信息服务，包括交通资讯、实时路况、公交车辆动态信息、停车动态信息、水上客运、航班和铁路等动态信息服务以及出行路径规划、出租召车等信息交互服务。例如，滴滴、Uber 打车软件提供出租车、快车、专车、顺风车服务，同时接入地图、路线查询、实时路况、在线支付等相关服务。智能停车软件也进入市场，如停简单、好停车、PP 停车等，实现停车行业与动态交通的有效衔接。

在物流方面，物流数据可以为物流市场预测、物流中心选址、优化配送线路、仓库储位优化等提供支撑，甚至能够提供交通路况、车辆运行、社会经济发展动态的信息。对于跨境物流，整合集口岸监管、物流运输、航运信息，可以实现物流产业链的业务单据、车辆船舶动态、通关状态等要素信息的跨行业、跨区域贯通，提高物流效率。

在管理方面，利用交通行业数据，支撑交通管理与决策。利用数据挖掘技术可以深入研究交通网优化，为行业发展趋势研判、政策制定及效果评估等提供支撑保障。此外，交通与公安、建管、环保等相关职能部门的大数据平台对接，可以提高跨领域管理能力。

在运营方面，整合行业数据，形成地面公交、出租汽车、轨道交通、路网建设、汽车服务、港口、航空等领域的一体化智能管理。通

过车载、运营数据的精确、实时采集，可以实现公交调度、行车安全监控、公交场站管理，支持公交安全、服务、成本管控的全过程管理和交互。通过打通出租汽车电调平台与互联网召车平台之间的信息渠道，可以提供多渠道便捷的召车服务，实现对出租汽车服务质量的动态跟踪、评估和管理。对轨道交通线网基础设施、运行状况、运营数据、服务质量、隐患治理、安全保护区等进行监测，可以实现安全管理和应急协同。

5. 医疗领域

医疗卫生领域每年都会产生海量的数据，一般的医疗机构每年会产生 1TB-20TB 的相关数据，个别大规模医院的年医疗数据甚至达到了 PB 级别。从数据种类上来看，医疗机构的数据不仅涉及服务结算数据和行政管理数据，还涉及大量复杂的门诊数据，包括门诊记录、住院记录、影像学记录、用药记录、手术记录、医保数据等，作为医疗患者的医疗档案，颗粒度极为细致。所以医疗数据无论从体量还是种类上来说都符合大数据特征，基于这些数据，可以有效辅助临床决策有效支撑临床方案。同时通过对疾病的流行病学分析，还可以对疾病危险进行分析和预警。

临床中遇到的疑难杂症，有时即便专家也缺乏经验，做出正确的诊断和治疗更加困难。临床决策支持系统可以通过海量文献的学习和不断的错误修正，给出最适宜诊断和最佳治疗。大数据分析技术将使临床决策支持系统更智能，这得益于对非结构化数据的分析能力的日益加强。比如可以使用图像分析和识别技术，识别医疗影像（X 光、

CT、MRI)数据，或者挖掘医疗文献数据建立医疗专家数据库，从而给医生提出诊疗建议。此外，临床决策支持系统还可以使医疗流程中大部分的工作流流向护理人员和助理医生，使医生从耗时过长的简单咨询工作中解脱出来，从而提高治疗效率。以 IBM Watson 为代表的临床决策系统在开发之初只是用来进行分诊的工作。而如今，通过建立医疗文献及专家数据库，Watson 已经可以依据与疗效相关的临床、病理及基因等特征，为医生提出规范化临床路径及个体化治疗建议，不仅可以提高工作效率和诊疗质量，也可以减少不良反应和治疗差错。在美国 Metropolitan 儿科重症病房的研究中，临床决策支持系统就避免了 40% 的药品不良反应事件。世界各地的很多医疗机构（如英国的 NICE，德国 IQWiG 等）已经开始了比较效果研究（CER）项目并取得了初步成功。

大量的基因数据、临床实验数据、环境数据以及居民的行为与健康数据形成了“大数据”，同时随着人类对疾病与基因之间映射关系的认识加深，基因测序成本的下降，可穿戴设备的普及，监控设备的微型化，移动连接和网络覆盖范围的扩大和大数据处理能力的大幅提升，针对患者个体的精准医疗和远程医疗成为可能。通过收集和分析数据，医生可以更好地判断病人病情，可实现计算机远程监护，对慢性病进行管理。通过对远程监控系统产生的数据的分析，可以减少病人住院时间，减少急诊量，实现提高家庭护理比例和门诊医生预约量的目标。

公共卫生部门可以通过覆盖全国的患者电子病历数据库，快速检

测传染病，进行全面的疫情监测，并通过集成疾病监测和响应程序，快速进行响应。百度通过对全国各地用户产生的搜索日志的分析，提供全国 331 个地级市，2870 个区县的疾病态势。百度还准备将社交媒体数据、问答社区数据、甚至是各地区天气变化、各地疾病人群迁徙等特征数据融合到预测里，进一步提高预测的准确性。很多研究者试图利用其他渠道（比如社交网站）的数据来预测流感。纽约罗切斯特大学的一个数据挖掘团队就曾利用 Twitter 的数据进行了尝试，研究者在一个月内收集了 60 余万人的 440 万条 Twitter 信息，挖掘其中的身体状态信息。分析结果表明，研究人员可以提前 8 天预报流感对个体的侵袭状况，而且准确率高达 90%。

基因测序研究一直是大数据应用的重点领域，随着大数据处理能力的不断提升，该领域的研究也进展显著。随着计算能力和基因测序能力逐步增加，美国哈佛医学院个人基因组项目负责人詹森·鲍比认为，2015 年会有 5000 万人拥有个人基因图谱，而一个基因组序列文件大小约为 750MB。成立于 2011 年的初创公司 Bina Technology 主要从事的工作就是利用大数据来分析人类的基因序列，他们的分析成果将为研究机构、临床医师等下游医疗服务行业提供最基础的研究素材。在同斯坦福大学研究者进行的试点研究结果表明，Bina Technology 平台利用大数据处理技术在 5 个小时内可完成几百人的基因序列分析，按照传统的分析方法，需要花费一周时间来完成。

以上我们从电信、金融、政府、交通和医疗健康等 5 个行业，分析行业大数据应用的典型模式、发展状况。大数据的应用其实是无所

不在的，其他行业如工业、零售业、农业的应用场景也非常多。但是总体来说，大数据应用尚处于初步阶段，受制于数据获得、数据质量、体制机制、法律法规、社会伦理、技术成本等多方面因素制约，实际成果还需要时间检验。

（三）大数据应用发展趋势

大数据行业应用的发展，是沿袭数据分析应用而来的渐变的过程。观察大数据应用的发展演变，可以从技术强度、数据广度和应用深度三个视角切入。从以上的应用来看，大数据区别于传统的数据分析有以下特征。数据方面，逐步从单一内部的小数据，向多源内外交融的大数据方向发展，数据多样性、体量逐渐增加。技术方面，从过去的报表等简单的描述性分析为主，向关联性、预测性分析演进，最终向决策性分析技术阶段发展。应用方面，传统数据分析以辅助决策为主，大数据应用中，数据分析已经成为核心业务系统的有机组成部分，最终生产、科研、行政等各类经济社会活动将普遍基于数据的决策，组织转型成为真正的数据驱动型组织。

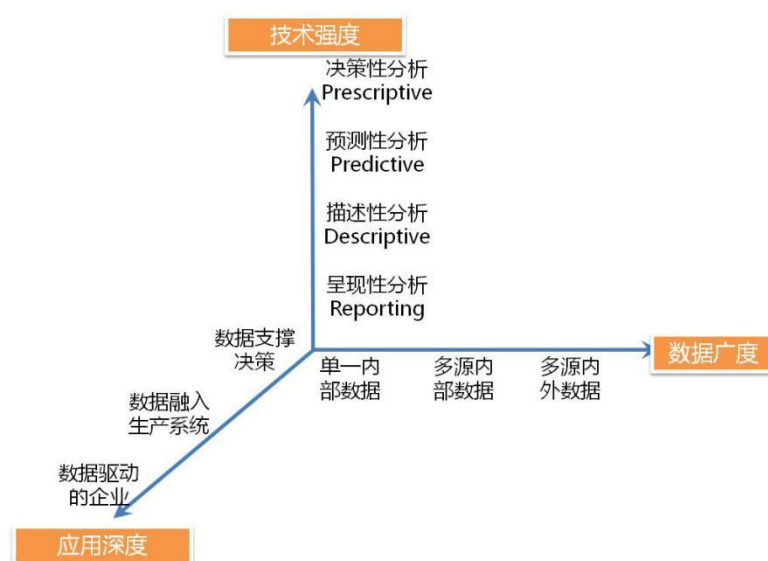


图 7 大数据应用的演进趋势

五、大数据政策法规

大数据的发展以及与传统产业的深度融合影响着社会、经济、生活的方方面面，也必然会对原有制度提出新的要求和挑战。为应对大数据发展带来的各种问题和需求，各国政府动作频频，通过修改原有法律法规、制定新的法律政策等方式，不断完善政府数据开放与信息公开、个人信息保护、数据跨境流动以及数据权属等方面的制度规定，为大数据的持续健康发展提供法律上的保障。

（一）政府数据开放与信息公开

大数据的发展推动了政府数据的爆发式增长，政府及公共部门汇聚了海量数据，成为最重要的数据资源库。在 2000 年左右，对政府信息公开工作的关注仍集中于提高透明度方面，但是随着 ICT 技术发展和大数据技术的出现，人们开始认识到政府信息公开不仅可增强民主和公众参与，作为电子数据形式的政府数据同时也是一种拥有重大

经济社会价值潜能的重要国家资源，对于提升政府服务效率、促进经济增长、增加就业机会和促进创新有重大作用。如何推动政府数据对外开放以及政府信息公开成为当前大数据发展中各方的重要关切。

1. 国际趋势

2015 年经济合作与发展组织 OECD 发布的《开放公共部门信息的政府措施评估》报告显示，制定政府数据开放战略成为各国普遍趋势。主要经济体相继发布国家战略或政策积极推动政府数据开放，并跟随数据开放的形势和要求不断修正和调整相关政策以适应新形势的发展。2009 年 1 月，美国总统奥巴马发布备忘录——《开放和透明的政府指令》，要求建立更加开放、透明、参与合作的政府。同年，奥巴马先后签署了信息自由法案备案录，明确提出了新的数据开放的原则，即：“面对存疑的数据是否开放的问题，以开放为导向的原则”。⁵2010 年 5 月，英国首相卡梅伦发布《致政府部门开放数据的一封信》，提出了各中央政府部门和本地政府部门初步开放政府数据的工作要求。2012 年，英国修改了《信息自由法案》，在新法案中将公共机构所持有的信息作为数据集，在数据开放的事物中，政府公共机构应当公布被申请要求公开的数据集和任何有更新的数据集。⁶

2011 年 9 月，美国、巴西、南非、英国等 8 国成立开放政府合作伙伴组织，并于 2013 年发布了《G8 集团开放数据章程》，承诺逐步开放高价值数据。2013 年 7 月欧盟对原《公共部门信息再利用指令》做出修正，以适应新形势下的政府数据开放。2015 年，澳大利

⁵See Ginsberg R W. The Obama Administration's Open Government Initiative: Issues for Congress.http://digital.library.unt.edu/ark:/67531/metadc31471/m1/1/high_res_d/R41361_2011Jan28.pdf

⁶参见迪莉娅：《大数据环境下政府数据开放研究》，知识产权出版社，2014 年版。

亚发布新的《公共数据政策声明》，对各政府部门提出新的开放数据要求。⁷

2. 我国法律政策现状

从国内立法政策来看，在早期阶段，我国政府数据开放主要是以政府信息公开为主。2007 年国务院制定并发布了《中华人民共和国政府信息公开条例》（以下简称《条例》），《条例》规定了政府信息公开的基本原则、公开的范围、公开的方式和程序以及监督和保障措施，初步奠定了我国政府信息公开的基础。同时，为正确审理政府信息公开行政案件，最高人民法院于 2011 年发布了《关于审理政府信息公开行政案件若干问题的规定》的司法解释，进一步明确了政府信息公开中的相关操作问题。

2015 年以来，随着大数据技术发展和产业推动，我国也从政府信息公开转向强调政府数据开放。在党中央、国务院发布的多个文件都对推进政府数据开放进行了不同程度的部署，包括：《关于促进云计算创新发展培育信息产业新业态的意见》、《关于大力推进大众创业万众创新若干政策措施的意见》、《国务院办公厅关于运用大数据加强对市场主体服务和监管的若干意见》、《国务院关于积极推进“互联网+”行动的指导意见》、《促进大数据发展行动纲要》、《关于加快构建大众创业万众创新支撑平台的指导意见》、《关于全面推进政务公开工作的意见》、《2016 年推进简政放权放管结合优化服务改革工作要点》以及《国家信息化发展战略纲要》等。其中，国务院

⁷See Declaration of Open Government:
<http://www.finance.gov.au/blog/2010/07/16/declaration-open-government/>

《促进大数据发展行动纲要》明确提出 2017 年底前形成跨部门数据资源共享共用格局，在 2018 年底前建成国家政府数据统一开放平台；《国家信息化发展战略纲要》要求建立公共信息资源开放目录，构建统一规范、互联互通、安全可控的国家数据开放体系，积极稳妥推进公共信息资源开放共享。

3. 立法展望

总体来看，我国政府数据开放与信息公开立法基础相对还比较薄弱，现行法律法规仅《政府信息公开条例》和《最高人民法院关于审理政府信息公开行政案件若干问题的规定》的司法解释有所涉及。建议加快制定《政府数据开放法》，从法律层面将政府数据开放制度化，推动从政府信息公开到政府数据开放，充分发挥和利用政府数据的价值。目前虽然有《政府信息公开条例》，但因为条例的立法层级较低，且《条例》的规定主要涉及政府信息公开，并未对数据开放问题作具体规定。此外，需要加快制定数据开放行动计划，建立统一的政府数据开放平台，国务院可出台《政府数据开放行动计划纲要》，从国家层面指导和推动政府数据开放。

（二）个人数据保护

大数据技术的广泛应用为个人数据保护带来新的挑战，在数据收集、数据分析、数据流转等环节的风险加大，个人信息泄露越来越严重，给现有的个人信息保护法律制度带来了新的挑战。为应对大数据发展带来的这些风险，各国纷纷修改和完善国内立法政策，以适应新形势下个人数据保护的要求。

1. 国际趋势

个人数据保护是国际社会重点关注的内容，近年来各国立法动作频频。韩国于 2011 年颁布《个人信息保护法》，废除了 1999 年《公共机关个人信息保护法》，新法适用范围涵盖公共与私人部门管理的所有个人数据信息。法国个人信息保护机构——国家信息与自由委员会发布《云计算数据保护指南》，对云计算服务协议应当包含的因素和云计算的安全管理提出了建议。2015 年 9 月，日本颁布了《个人信息保护修正法》，增加了对个人数据跨境转移的限制，将该法的适用范围扩大到境外，扩大了个人信息的定义并“敏感信息”进行了区分，此外，该法还建立了个人信息保护委员会。欧盟也对个人数据保护进行了大刀阔斧的改革，2016 年 4 月，欧盟的《一般个人数据保护条例》（以下简称《条例》）正式出台，《条例》将取代 1995 年出台的《数据保护指令》并在 2018 年正式生效实施。《条例》继续坚守保护公民基本权利理念，全面提升了个人数据保护力度，并特别针对大数据背景下的数据分析、画像活动，予以严格规制。

2. 我国立法现状

我国个人数据保护立法起步较晚。2012 年底，全国人大常委会通过了《关于加强网络信息保护的决定》（以下简称《人大决定》），首次以法律的形式明确规定保护公民个人及法人信息安全，奠定了我国个人信息保护的立法基础。在此基础上，相关部门也针对各行业个人信息保护做出了规定。2013 年 7 月，工业和信息化部出台了《电信和互联网用户个人信息保护规定》，明确了电信业务经营者、互联

网信息服务提供者收集、使用用户个人信息的规则和信息安全保障措施等。2016 年 11 月 7 日，全国人大常委会通过了《网络安全法》，并将个人信息保护纳入网络安全保护的范畴，《网络安全法》第四章“网络信息安全”也被称为“个人信息保护专章”。《网络安全法》统一了“个人信息”的定义和范围，确立了个人信息收集使用的基本原则，规定了相关主体的个人信息保护义务。更为重要的是，《网络安全法》明确了违反个人信息保护的法律责任，弥补了《人大决定》中没有罚则的不足。

《人大决定》、工信部《规定》和《网络安全法》的相继发布，标志着我国个人信息保护工作取得了重大进展。《人大决定》以最高立法层级的形式对个人信息保护的要求做了明确规定，为各项工作的开展提供了上位法依据；工信部《规定》从落实的角度对个人信息保护相关工作进行了细化，明确了责任主体、法定义务、罚则等。《网络安全法》则总结了我国个人信息保护立法经验，针对实践中存在的突出问题，将近年来一些成熟的做法作为制度确定下来。

此外，我国在《刑法》、《消费者权益保护法》、《身份证法》、《征信业管理条例》等法律法规中也对个人信息保护做了相关规定，进一步补充健全了我国的个人信息保护法律体系。

3. 立法展望

大数据的发展使得个人信息保护面临的形势更加复杂，个人信息泄露事件引发各界高度关注，全球个人信息保护立法活动持续升温。首先，我国应加快制定专门的个人信息保护法，明确公民个人对其信

息享有的基本权利，规范企业收集和使用个人信息的行为。其次，重点关注新业务新技术领域的个人信息保护。当前互联网技术不断发展，新技术新业务层出不穷，我们在确定个人信息保护基本规则的同时也要关注新技术新业务的特点。此外，还应加强国际间的合作与对话，积极参与个人信息保护的国际或区域规则制定，加强个人信息和隐私保护的跨境执法合作。

（三）跨境数据流动

大数据推动全球进入数字经济时代。在数字经济的驱动下，跨境数据流动日益频繁，如何应对跨境数据流动带来的数据安全风险成为当前国际社会争论最为激烈的话题。

1. 国际趋势

跨境数据流动的概念最早出现在个人隐私和数据保护条款中。从国际组织以及部分国家对跨境数据流动的管理制度来看，对跨境数据流动主要有两种理解：一种是数据被跨越国界传输和处理；另一种是数据本身虽未被传输出境，但能被别国的主体访问。“棱镜门”事件前，网络数据开放逐年深化，针对跨境流动等的国际合作不断推进，“注重开放”成为了国际网络空间数据使用的主流态度；而“后棱镜门”时代，各国开始明确并不断强化网络数据安全保护，加强网络数据安全安全管理。

2014 年，俄罗斯通过《关于信息、信息技术和信息保护法》、《俄罗斯联邦个人数据法》的修改确立了数据跨境流动的本地存储规则，要求商业机构有义务确保俄公民个人数据的处理活动均应使用俄

联邦境内的服务器。2015 年 10 月，澳大利亚通过《电信（监控和接入）修正（数据留存）提案》，对数据留存做出强制性法律规定，要求电信运营商对电话、互联网、电子邮件的用户数据留存两年。2015 年 10 月 6 日，欧洲法院认定欧盟委员会 2000 年通过的关于认可美欧安全港框架的决定无效，使得美欧之间最重要的跨境数据传输方式丧失了合法性基础。经过紧急谈判协商，2016 年，欧盟和美国商务部达成了新的“隐私盾”协议，对美方的企业规定了更为严格的个人数据保护义务和更高的要求。

2. 我国法律政策现状

我国现有部分法律法规已经对跨境数据流动管理做了相关规定。例如，《保守国家秘密法》要求防止含有国家秘密的数据流出中国；《征信管理条例》规定征信机构对在中国境内采集的信息的整理、保存和加工，应当在中国境内进行；《地图管理条例》规定互联网地图服务单位应当将存放地图数据的服务器设在中华人民共和国境内，并制定互联网地图数据安全管理制度和保障措施。此外，中国人民银行对个人金融信息数据⁸、国家卫计委对涉及人口健康信息数据⁹以及网络出版¹⁰、网络约车¹¹等都要求在中国境内存储。管理实践中，我国在相关政策标准中提出过基础设施本地化的管理要求。《关于大力推进信息化发展和切实保障信息安全的若干意见》中规定，为政府机关提供服务的数据中心、云计算平台等要设在境内；《信息安全技术云计

⁸参见《关于银行业金融机构做好个人金融信息保护工作的通知》（银发〔2011〕17 号）。

⁹参见《人口健康信息管理办法（试行）》（国卫规划发〔2014〕24 号）。

¹⁰参见《网络出版服务管理规定》第八条。

¹¹参见《网络预约出租汽车经营服务管理暂行办法》第二十七条。

算服务安全能力要求》（GB/T）提出，云服务商应确保云计算服务器及运行关键业务和数据的物理设备位于中国境内。

为应对日益严峻的国际跨境数据流动风险，我国《网络安全法》也对涉及关键信息基础设施的数据做了相关要求。其第三十七条规定：关键信息基础设施的运营者在中华人民共和国境内运营中收集和产生的个人信息和重要数据应当在境内存储。因业务需要，确需向境外提供的，应当按照国家网信部门会同国务院有关部门制定的办法进行安全评估；法律、行政法规另有规定的，依照其规定。

3. 立法展望

全球数字经济的发展对跨境数据流动提出了一定要求，完全禁止数据的跨境流动也不符合我国大数据产业发展的现实需要，平衡数据跨境流动与确保安全成为当前和未来我国立法的一个重要方向。一方面，应建立数据分级分类管理制度，对涉及国家秘密、社会安全以及经济安全的数据严格禁止流出；对政府和公共部门掌握的其他数据实施有条件限制的跨境流动管理；对普通的个人数据通过合同监管落实数据控制主体的安全责任以及实施保护。另一方面，强化相关主体数据安全保护责任，政府部门创新监管方式，完善执法手段；云服务提供商等企业应当采取多种措施确保数据存储安全，严格执行国家关于跨境数据流动的管理规定。

（四）数据权属问题

随着大数据产业的持续推进，在数据采集、使用、交易、流转等环节中，数据产权不明晰的问题日益凸显，严重制约着大数据产业的

健康持续发展。厘清数据权属被视为解决数据采集、使用、流通、转移等环节中的权利关系，保障数据交易合法性、规范大数据应用秩序等的先决条件，成为时下大数据产业和信息经济、共享经济发展的紧迫需求。

1. 国际趋势

从全球的范围来看，对于数据权属的界定仍然是一个相对模糊的状态，特别是对于汇集起来的大数据权属问题，在法律上各国并无明确规定，也成为困扰国内外学界与实务界的难题。在规定个人和企业对于数据的权利时，美欧纷纷回避了对数据权属的界定。一般观点认为，企业对匿名化的数据集享有所有权，但目前也呈逐渐限制的趋势。例如，美国联邦贸易委员会（FTC）2014 年针对数据经纪行业发布的报告《数据经纪行业，呼唤透明与问责》中就表达了 FTC 对于数据交易缺乏透明性的关切，并建议国会应当专门针对数据经纪行业制定立法，通过立法要求开展数据交易活动的企业对用户提供透明度。

2. 国内法律政策现状

整体来看，目前我国并无对数据所有权的专门立法，对于数据本身的属性到底是财产性权利还是人身性权利也存在理论上的争议。但从市场实践来看，大数据的商品化充分说明了其具有财产性权利的性质。从这个角度出发，目前我国的《物权法》、《民法通则》、《合同法》等都可以适用。根据我国《物权法》第四十五条之规定，所有权人对自己的不动产或者动产，依照法律规定享有占有、使用、收益和处分的权利。由于缺乏立法上的明确规定，部分企业开始探索通过

行业公约的形式来解决数据权利的归属问题。2015 年 7 月，阿里巴巴旗下云计算子公司阿里云发布了“数据保护协议”，承诺云计算平台客户对自己的数据拥有绝对所有权，有权在任何时间访问、分享、交换、转移或删除其数据。

3. 立法展望

解决数据所有权问题，一方面要着眼于明确产权归属，为数据交易的顺利开展提供稳定的法律基础；另一方面，仍要关切数据交易带来的隐私与信息安全风险，对数据交易活动做出相关限制性要求，特别提出透明性方面的要求。首先需明确的是，依据现有法律规定，我国禁止公民个人信息的出售行为，但从未来来看，个人信息的交易合法化也并非完全没有可能，因此在法律上并不能完全排除个人将其个人信息出卖而获益的正当性。其次，对于在用户数据基础上，做出充分匿名化处理的数据集，可以规定企业（即数据控制者）对其享有有限定的所有权。在法律上赋予企业对匿名化数据集的所有权，增加数据交易的法律稳定性与可预期性，为企业利用数据创造财富提供积累机制。

六、结论与建议

去年以来，《促进大数据发展行动纲要》出台，十八届五中全会进一步提出要在“十三五”期间实施国家大数据战略。当前社会各界对大数据的期待上升到了前所未有的高度。如何推动大数据战略落地成为未来几年的政策重点。

（一）避免盲目跟风，大数据热潮还需冷思考

身处大数据热潮中，既要充分认识大数据的潜力，积极把握技术进步带来的机遇，也要认清大数据的局限性，警惕大数据万能论。一些被广泛传播的经典案例现在被证明是子虚乌有，比如，啤酒与尿布的故事实际上是 Teradata 公司的工程师 Thomas Blischok 在 1992 年杜撰的，从来没发生过；而 Netflix 号称用大数据分析帮助自制剧《纸牌屋》取得成功，而实际上是把大数据作为公关活动的噱头。当前，以下几点值得思考：

第一，大数据尚难对人的行为做出精确预测。在大数据是否能准确预测人类行为的问题上，还存在重大分歧。《黑天鹅》指出人类的行为不可预测，而《爆发》一书则根据对以往历史经验的总结，指出人类行为 93% 可预测。麻省理工学院教授罗伯特·莱格伯恩（Roberto Rigobon）称，虽然华尔街一直重视数据分析，但基于海量数据分析的对冲基金在全球都是失败的。“对于人和事件，如果放到越大的空间和时间范围，则是越可以精确预测的。如果放到越小的空间和时间范围，则是越不可以精确预测的。例如，我们几乎可以在 100% 的程度上预测一个人在 24 小时的时间范围内会吃饭；但若精确到某一分钟，则几乎不可能预测准确。”大数据无法预测人类行为，归根结底还是因为人具有“自由意志”，人会根据预测结果（如下个月的股票价格、明天的交通拥堵情况）改变自身行为，从而使得预测失效。

第二，大数据相关关系不能替代因果关系。舍恩伯格在《大数据时代》中说：“我们没有必要非得知道现象背后的原因，而是要让数

据自己发声”，“相关关系能够帮助我们更好地了解这个世界”。追寻相关关系和因果关系，是人类思维的两种重要方式，而用大数据进行预测往往依靠相关性，也就是说，很多情况下知道“是什么”即可，不必知道“为什么”。相关关系的运用在互联网推荐、精准广告等方面得到了实际应用。然而，在很多时候，如疾病诊断、工厂故障分析等场景下，需要根据确定的（或置信度非常高的）结论来决策，仅凭相关关系是远远不够的。换言之，大数据中的相关关系应用，需要区分场景，有时候数据无法自己说话，需要追本溯源。

第三，大数据来源不均衡会让数据“说谎”。有人说数据不会撒谎。实际上，如果忽视数据来源的不均衡性，数据分析结果就会“骗人”。中国互联网络信息中心 2015 年的统计数据显示，我国网民城乡分布严重不均，农村网民虽然迅猛增长，但仍不及城市新增网民数量的 1/10。社交网络的用户的性别分布也同样有很严重的倾斜，腾讯公司 2015 年年初的报告显示，微信用户的男女比例为 1.8:1，男性用户约占了 64.3%，而女性用户则只有 35.7%。如果利用网络大数据进行民意调查，却不把样本分布的不均衡性考虑进去，就可能使得某些群体未得到充分代表，而某些群体因使用率高，其意见或特征被过分放大。这种不均匀的数据来源会导致分析结果存在偏见和盲区。

第四，大数据无法消灭信息不对称现象。有人说，大数据有助于消灭信息不对称。虽然从全社会看，大数据的全面采集和融合应用有望在局部缓和信息不对称程度，但是在互联网世界中，马太效应很显著，拥有大数据资源和掌握大数据分析能力的企业，往往会在大数据

时代占据更加有利的地位、占有更多数据，从而更容易形成一批数据寡头，产生新的不平等，造成新的信息不对称。因此，大数据无法消灭信息不对称，反而更有可能助推数据寡头的出现。如果这种数据垄断地位被企业滥用，将会威胁个人、企业甚至国家利益。因此，在大数据时代，如何进一步弥合数据鸿沟、防止数据“霸权”的滥用，将会成为一个重要新的课题。热潮之下，对大数据的反思，还需要不断深入，才能让我们保持清醒的头脑。

（二）推动开放共享，倒逼信息化建设升级

以上从理论层面做了探讨。从大数据产业实际发展来看，我们国家还存在数据开放、技术创新、制度建设、区域协同等多方面的瓶颈需要突破。

开放政府数据，并带头用好大数据技术，是政府部门支持大数据发展最直接举措。经过多年发展，我国政府信息化建设取得了举世瞩目的成就。自 1993 年启动金桥工程、金关工程和金卡工程以来，“两网一站四库十二金”相继建成，政务信息化水平不断提升；面向公众服务的政府网站群也已经具有较大规模，截至 2015 上半年，全国各级政府网站总数达到 8.6 万个，其中地方 8.3 万个，国务院部门 3000 多个。政务履职和公众服务过程中积累了丰富的数据资源，是十分宝贵的资源。

数据开放共享一直是政务信息化建设的理想目标。以前，系统建设烟囱式的建设模式，加上数据权责利的管理制度没有建立起来，导致横向来看在政府内部数据孤岛普遍存在，纵向来看数据对外开放更

是缺乏技术与制度基础。现在，在全社会推进大数据的应用，数据的多源融合是先决条件，政府数据的共享开放已经成为不得不做的事情。恰好在最近几年，云计算不断成熟，为统一的政务信息平台建设提供了新工具，为数据共享融合提供了技术便利。

李克强总理强调，“首先要把政府大数据的建设事情办好，给社会一个好的示范。”用政府大数据的应用倒逼政务信息化升级，推动政务信息化建设从烟囱式、封闭式、集中式的模式，转向平台式、开放式、分布式的模式。国务院《促进大数据发展行动纲要》中，把这项工作放在首位，提出了统筹基础设施、整合应用平台、推动数据共享、推进数据开放等基础性工作，还提出要基于融合的数据，加强宏观调控科学化、政府治理精准化、商事服务便捷化等应用创新。从自身做起，体现了政府推动大数据的决心。

然而从各地推进情况看，政府数据的开放共享在实际操作中的阻力不小，动力不足。改变目前政府部门不愿开放、不敢开放、不能开放的现状，长远之计，是要自上而下，中央建立一套完善的数据开放共享机制，明确开放共享的数据目录、技术标准，以及平台建设思路，部委和地方去落实。短期来看，还需要结合渐进路线，逐步推进。例如在政府数据开放方面，可先从已经开放的数据如何便利化应用入手。

我国很多政府网站都已经开放了比较丰富的数据资源。很多政府部门已经在网站上公开了资质审核、行业统计、项目审批、产品信息、标准规范等信息。但政府已开放数据大多存在以下三方面问题，一是不好找，现有数据较为分散，检索缺乏统一入口。二是数据不好看，

大多以表格或文字综述报告形式呈现，直观性不强，公众理解起来比较困难。三是数据不好用，数据格式标准不统一，绝大部分不支持机器可读。这些都增加了我部数据社会化应用的技术门槛和成本。

解决上述问题，可以从技术上入手，统一标准，建立平台，首先让政府网站上本已开放的数据更好找、更好看、更好用，成熟后逐步扩大开放范围，将是务实可行的第一步。

（三）强调供需对接，拉动技术产业跨越发展

大数据资源与技术，就好比工业时代的燃料与引擎，不仅自成产业，还能够驱动其他产业更好发展。当前，开源模式迅猛发展，技术“民主化”潮流势不可挡，数据技术的轨道正在从集中式向分布式切换，传统产业的格局有望重塑。在这样的大变轨时期，一方面我国领先的互联网企业的 IT 制造企业与国际先进水平的差距不断缩小，甚至在一些方向上达到了前所未有的接近程度。另一方面，我国正在实施《中国制造 2025》战略，农业和服务业的正在加速转型，有数不清的问题等待着用大数据去解决，对大数据技术产品的需求空间也十分巨大。

技术产业加速变轨、国内产业快速崛起和庞大的市场内生需求三者叠加，使得我国具备在大数据领域实现跨越发展的条件。之前的信息化几次信息化浪潮，国内产业没能实现弯道超车，天时、地利、人和没有同时具备。然而在当前的时点上，我们同时具备了产业支撑能力和巨大应用空间的优势，如果能够将两方面优势结合起来，形成良性互动格局，就能够实现跨越发展。《促进大数据发展行动纲要》提

出“推动产业创新发展，培育新兴业态，助力经济转型”的任务，体现了谋划跨越发展的前瞻性。

我国大数据产业发展的一个重要目标是打造自主可控的产业体系。当前，从大数据技术与产品的供给侧看，我国虽然在局部技术实现了单点突破，但大数据领域系统性、平台级技术创新仍不多见，供应商面临着紧跟技术趋势、精准对接用户需求的压力。从大数据技术与产品的需求侧看，金融、电信、工业、医疗、政府等行业用户来说，正面临着如何规划技术路线、如何选择商用产品、如何构建和运维大数据平台等问题。

为此，下一步着力点应该按照中央提出的供给侧改革思路，发挥产业联盟等平台作用，深入挖掘业务需求，促进供需精准对接，把国内优势技术力量凝聚起来形成合力，突破关键技术，推出满足关键行业重大需求的大数据技术产品体系，并以产业实践为基础，逐步形成接地气的大数据标准体系和知识产权体系，逐渐向技术和产业前沿和高端跃升。

（四）完善法律制度，切实保障数据安全

大数据技术的发展和应用将数据作为一种新的资源推向了台前。当数据这种新的资源越来越受重视时，与数据相关的权利义务界定也就显得越来越重要。小到个人，大到企业 and 国家，都是大数据的利益相关方，法律制度需要随着大数据的发展而进行动态调整，在促进产业发展的同时切实保障数据安全。

对个人来说，大数据的应用对隐私保护提出了巨大挑战，技术面

前个人越来越渺小和脆弱。要严格保护大数据应用中的个人信息，就需要探索形成大数据环境下数据收集、开放、交换、应用等环节的规则，明确大数据应用相关各方的个人信息保护义务和责任。

对企业来说，数据资产的所有权、使用权还是个模糊地带，急需建立数据产权保护制度，明确各类市场主体所积累的信息资产所有权归属，建立规范化管理和使用机制，保护信息所有者、信息主体及公众合法权益。

对国家来说，数据空间成为主权的新领域。需要研究跨境数据流动分级、分类管理制度，涉及国家秘密、国家安全以及经济安全的数据进行管理，积极主张国家数据主权，确保大数据时代的国家安全。

李克强总理指出，“政府既要‘扶持’，为大数据产业创造一个健康发展的环境，又要‘引导’、‘规范’，保障信息安全”，并提出“要完善产业标准体系，依法依规打击数据滥用、侵犯隐私等行为，让各类市场主体公平分享大数据带来的技术、制度和创新活力”。《促进大数据发展行动纲要》明确了“强化安全保障，提高管理水平，促进健康发展”的任务，以及“加快法规制度建设”的措施，从法律法规、管理制度和技术手段等多层次保障大数据安全。

完善大数据发展的制度环境是一个长期过程。长远来看，要提升大数据治理水平，深入研究数据权益、数据管理、数据交易、数据安全等关键问题，推动建立数据流通和使用的行业自律机制，逐步完善出台大数据相关法律体系，推进法治化进程。建立大规模个人信息泄露报告制度，完善网络数据和用户信息的安全防护措施及管理机制，

健全网络数据保护制度。短期来看，在法律法规尚未出台之前，要充分发挥行业组织作用，构建大数据交易流通与合规应用的行业自律机制，推动行业自律，建立基于实践的大数据安全管控技术标准体系，开展大数据平台产品及服务商的可靠性及安全性评测工作、应用安全评测、监测预警和风险评估。

（五）突出地方特色，形成差异化的区域产业布局

国务院《促进大数据发展行动纲要》中明确提出，要“加强中央与地方协调，引导地方各级政府结合自身条件合理定位、科学谋划，将大数据发展纳入本地区经济社会和城镇化发展规划，制定出台促进大数据产业发展的政策措施，突出区域特色和分工，抓好措施落实，实现科学有序发展。”4月13日召开的促进大数据发展部际联席会议第一次会议，进一步明确了地方大数据发展的重点方向，加快综合试验区建设，鼓励地方开展制度创新探索，推动数据创新应用，破解大数据发展难题。

国务院《促进大数据发展行动纲要》出台后，各地抢抓发展机遇，谋划大数据发展蓝图，不少地方已经在顶层设计、体制机制创新、业态探索和基础设施建设等方面取得了明显进展。据中国信息通信研究院统计，目前已经有23个省市出台了74个大数据相关的指导意见或规划，广东省、贵州省、辽宁沈阳市、四川成都市等地方政府还成立了大数据相关的专门机构，北京、贵州、陕西、湖北、河北、上海、浙江等地成立大数据交易所或交易中心，内蒙古、贵州等资源禀赋较好地区的超大规模绿色数据中心建设相继建成，形成了良好的发展局

面。

需要特别关注的是，大部分省市的大数据规划都有大手笔的数据中心建设计划。需要注意的是，大数据绝不等于“大数据中心”，大数据发展并不一定需要大面积的产业园区。因此，地方发展大数据的重点，不是建产业园、建数据中心，而是要充分依托已有设施资源，把现有的利用好，坚决杜绝盲目新建数据中心，避免造成资源空间的浪费。地方需要差异化发展，应该把大数据的发展重心放在因地制宜的促进应用创新上，放在打造完善的发展环境上，让市场在大数据发展要素配置上起决定作用。

CAICT 中国信通院

CAICT 中国信通院

CAICT 中国信通院

中国信息通信研究院

地 址：北京市海淀区花园北路 52 号

邮政编码：100191

联系电话：010-62304839

传 真：010-62304980

网 址：www.caict.ac.cn

