第一篇 中国大数据开放共享发展报告

本篇主要论述数据的开放共享在大数据应用中的重要作用，回顾了大数据开放共享国内外现状，分析了大数据开放共享的风险与对策，列举了大数据开放共享的典型案例，并对大数据的开放共享提出展望。

第1章 引言

当前主要发达国家运用大数据推动经济发展、完善社会治理、提升政府服务和监管能力正成为趋势。鉴于大数据潜在的巨大影响，欧美日等发达国家都将大数据视作战略资源，纷纷出台了大数据开放共享战略。截至2014年4月，全球已有63个国家制定了开放政府数据计划。目前，我国大力推动“信息化与工业化深度融合”，全面实施“中国制造2025”规划，互联网、移动互联网应用规模居全球首位，拥有丰富的数据资源和应用市场优势，大数据部分关键技术研发取得突破，涌现出一批互联网创新企业和创新应用，一些地方政府已启动大数据相关工作。坚持创新驱动发展，加快大数据部署，促进大数据开放共享与深度应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。国务院在2015年8月31日印发了《促进大数据发展行动纲要》（简称《大数据行动纲要》），为我国发展大数据相关产业提供了重要的政策支持和方向选择。

中国现行政策及管理格局，使得政府掌握全国80%的社会信息资源，而《大数据行动纲要》最基本的要求就是开放政府数据，打造以人为本的国家精准治理新模式，并以数据推动高端智能产业发展的新生态。这是一个宏伟的目标，为了达到这个目标，数据的有效开放共享是关键之一。

大数据开放共享将成为推动经济转型发展的新动力。大数据推动社会生产要素的网络化共享、集约化整合、协作化开发和高效化利用，改变了传统的生产方式和经济运行机制，可显著提升经济运行水平和效率。大数据开放共享将持续激发商业模式创新，不断催生新业态，将成为互联网等新兴领域促进业务创新增值、提升企业核心价值的重要驱动力。

大数据开放共享将成为重塑国家竞争优势的新机遇。在全球信息化快速发展的大背景下，大数据已成为国家重要的基础性战略资源，正引领新一轮科技创新。充分利用我国的数据规模优势，推进大数据开放共享，进而实现数据规模、质量和应用水平同步提升，发掘和释放数据资源的潜在价值，将有利于更好发挥数据资源的战略作用，增强网络空间数据主权保护能力，维护国家安全，有效提升国家竞争力。在国家“一带一路”的战略背景下，我国企业

将对外广泛开展“基建产能输出+资源输入”的合作，并着眼于发挥数据的战略作用，提高国家的竞争实力，完成从“工业制造”到“工业智造”的转变，进而实现“商贸文化互通，区域共同繁荣”。

大数据开放共享将成为提升政府治理能力的新途径。推动政府数据开放共享，促进社会事业数据融合和资源整合，将极大提升政府整体数据分析能力，为有效处理复杂社会问题提供新的手段。建立“用数据说话、用数据决策、用数据管理、用数据创新”的管理机制，实现基于数据的科学决策，将推动政府管理理念和社会治理模式进步，加快建设与社会主义市场经济体制和中国特色社会主义事业发展相适应的法治政府、创新政府、廉洁政府和服务型政府，逐步实现政府治理能力现代化。为完成我国从“工业制造”到“工业智造”的转变，我们认为“高质量数据的开放、共享和流通，是亟待解决的基础问题”。

因此，如何实现《大数据行动纲要》中提出数据开放共享，对社会的发展至关重要。但数据开放共享的路途艰难，正如《大数据行动纲要》中提到的，“政府数据开放共享不足、产业基础薄弱、缺乏顶层设计和统筹规划、法律法规建设滞后、创新应用领域不广等问题亟待解决”。怎样借鉴历史，借鉴其他国家的先进经验，结合实际，在数据开放共享的基础上，对数据有效地开发使用，是摆在我们面前的一个重要的任务。

除《大数据行动纲要》中重点提出的政府数据开放、共享外，行业数据、科学数据的开放、共享和流通也是需要解决的问题。因此，我们认为依托政府数据统一共享交换平台，大力推进我国各类国家、行业和科研的基础数据资源，完成跨部门、跨行业、跨区域的开放、共享和流通是大数据应用的基础。大数据开放共享，可以从以下三个方面，齐头并进，大力推动。

1.1 政府－加快数据开放共享

政府层面的数据开放共享是一种态度和能力。一些重大基础数据的开放，可以构成社会各行业的数据基础。其质量大大高于来自行业、互联网的数据质量。这些高质量数据的引入，将大大加快各领域、各行业数据融合的速度；其开放（或部分开放）也将会产生类似核聚变一样的价值发现效应。

我国已充分认识到数据开放、共享的重要性。为此，在《大数据行动纲要》中的主要任务中，数据的开放共享被放在第一个重要任务的位置上。虽然一些政府职能部门已进行过一些尝试（如国家统计局曾联合百度、阿里巴巴进行了一些探索性的尝试），但这些尝试仍无法有效解决“数据割据，拥数自重”的普遍现象。如气象观测数据对研究大气变化、气候

演变、农业指导，包括特殊天气情况下的资源调配、物资运输等具有非常重要的意义。但就政府数据的开放共享现状而言，数据的应用范围仍有很大提升空间。同时，政府需要在顶层设计、统筹规划、法律法规建设等方面加快步伐。在提升数据安全性的同时，逐步消除部门数据割据，建立公开、透明、共享的数据公共平台。

1.2 行业－互融互通

在大数据时代背景下，传统行业需要重新审视自己的发展战略，行业内部（上下游之间）、行业之间的数据需要更好的融合与利用。

大数据在诸多传统行业得到良好应用。给企业家带来冲击的不仅是大数据本身，而是一些新兴企业不可思议的跨界能力。行业之间的界限也随之变得模糊，这些新兴公司大规模采集数据，并采用新模式、新技术，快速形成预测，指导行业发展。譬如乐视网以影视数据的运营和服务为核心业务，已拓展到电视销售、电影拍摄；小米公司以移动互联网终端数据为核心，逐步涉及白家电以及智能家居行业；滴滴、快滴等以出租车营运数据为核心业务，应用快速普及，大有颠覆传统出租车行业之势。从大数据的视角来看，任何产业中，数据资产都将成为最核心的竞争力。从产业发展的视角来看，“拥数自重”的发展模式，并不会带来企业的长久繁荣。企业在数据开放和共享上的封闭自守，只能导致企业留在过去，退出未来的舞台。

未来，各行业中有两类企业将得以更好发展：一类是平台型企业，一类是依据平台且具有特色应用的企业。平台型企业拥有行业内部，以及相关行业的大量数据，但最终一个行业可能只有少量具有核心竞争力的企业可以存活；应用型企业将以特色经营和快速发展见长，它们依托平台的海量数据，提供简洁快速的服务，迅速发展壮大。

1.3 科教－数据共享，促进科学研究领域的快速发展

科研活动在大数据时代具有更为广阔的舞台。近年来，随着互联网的快速发展，科研活动逐渐摆脱“不了解国际最新动态”的情况；特别是各种搜索引擎的广泛使用，使得信息检索的能力大幅增强。随着国家科教方面的大力投入，在人才培养方面，我国取得了巨大的进步。据测算，截止2011年，我国各类科研人员达到6300万人，总量占到世界总量的25.3%，超过美国科研人员总量（美国科研人员占世界总量17%），居世界第一（据清华大学技术创新研究中心近日发布的《国家创新蓝皮书》）。虽然科研领域在研究方法等方面取得了长足进步，但科技方面的人均产出效率远落后于发达国家。其重要原因之一是受限于科学和应用研究领域的基础数据缺失。科学研究、工程应用等缺乏有效的数据支持，深刻影响了工程应用

领域的研究进展。

当前已进入大数据时代，许多学科的研究人员除发布其研究成果，还在发布该领域研究的基础数据。有了这些开放和共享的数据，大量的科研人员得以广泛开展研究工作。可以说，科学数据的开放和共享为进一步研究和应用注入了新的活力。

通过建立大数据共享实验平台和国家级的大数据研究实验室，搭建产业界和学术界的桥梁，为学术界优秀的方法提供广阔的实验平台，为产业界提供破解难题的机会，从而间接推动数据科学领域学科建设与人才培养的任务。同时，大量开放的科学数据还使得我国的科学教育有据可依、有数可循。

国际数据公司（IDC）的研究结果表明，2008年全球产生的数据量为0.49ZB，2009年的数据量为0.8ZB，2010年增长为1.2ZB，2011年的数量更是高达1.82ZB，相当于全球每人产生200GB以上的数据。而到2012年为止，人类生产的所有印刷材料的数据量是200PB，全人类历史上说过的所有话的数据量大约是5EB。

虽然全球所产生的数据量每年以约50%的速度增长，但开放共享的数据量却增长有限。就数据的开放比例而言，逐年下降。完善上述大数据开放共享的推进工作，将在实质意义上解放大数据所带来的新的机遇，将在今后的若干年乃至几十年给予社会发展提供一个极大的推进力。

第2章 数据开放共享是社会发展的驱动力

国务院印发的《大数据行动纲要》指出了数据开放共享的必要性，及其对社会发展的迫切性。大数据的开放共享，将在国家治理，各行业的发展，以及科学研究创新方面，起到根本性、革命性的推动作用。

2.1 大数据助力国家治理现代化

大数据不仅是一场技术革命，一场经济变革，也是一场国家治理的变革。世界上越来越多的国家将数据管理上升到了战略层面，大数据思维和应用已逐渐渗透到公共管理和政府治理范畴内，对政府治理理念、治理范式、治理内容、治理手段等产生不可忽视的影响。

大数据正有力地推动着国家治理体系和治理能力走向现代化，正日益成为社会管理的驱动力、政府治理的“幕僚高参”。习近平在参观腾讯公司时提到：“互联网在社会管理方面有较大作用，我们怎么去适应它？”，“我看到你们做的工作都是很重要的，比如在这样的海量信息中，你们占有了最充分的数据，然后可以做出最客观、精准的分析。这方面对政府提供的建议是很有价值的”。李克强在考察北京•贵阳大数据应用展示中心时说：“把执法权力关进‘数据铁笼’，让失信市场行为无处遁形，权力运行处处留痕，为政府决策提供第一手科学依据，实现‘人在干、云在算’。”李克强考察山东浪潮集团时指出：“不管是推动政府的简政放权，放管结合，还是推进新型工业化、城镇化、农业现代化，都要依靠大数据、云计算。所以，它应该是大势所趋，是一个潮流。”大数据使得政府决策的基础从少量的“样本数据”转变为海量的“全体数据”。政府树立大数据意识，促进相关数据完全共享，更多地依赖数据进行决策，可以实现从以有限个案为基础向“用数据说话”转变的全新决策。

政府治理理论的核心是主张通过合作、协商、伙伴关系，确定共同的目标等途径，实现对公共事务的管理，即权力多中心化以及由此引发主体多元化、结构网络化、过程互动化和方式协调化的诉求。如何增强社会公众对政府的信任，在公共治理中建立起政府与社会、公众的紧密合作关系，不断提高治理能力和治理绩效，实现政府与社会公众之间的良性互动是政府治理能力的重要体现。大数据的社会属性与“治理”理论在多中心、回应性、协同化等诸多方面高度一致。因此，将大数据应用到政府治理中将加速政府治理的创新，可以产生“倍增”效应。（1）政府需要及时有效地回应公众诉求，并反馈收集到的意见、处置措施和政策对策、最终的实际整改成效等，这就要求政府进行绩效管理和评估时，既要通过对政府内部业务系统进行分析，同时也要参考外部因素，如政府网站中公众的参与度、点击率和评价等，甚至包括新浪、腾讯等非政府媒体中与政府有关的评价；（2）让社会公众对政府部门

进行评价，不仅体现了政府的服务导向，也因为大数据方法和技术的应用，大大增强了绩效管理的问题发现功能。一方面，政府通过加深在网络反腐、舆情监控等公共领域对数据的应用，实现由事后决策转变为事前预警，将数据转化为科学决策，提升政府决策力；另一方面，通过大数据方法和技术，政府把低价值度的数据转变成有效有用的绩效信息，从而提升了政府治理能力。

2.2 行业数据的流通对产业发展的驱动

大数据的开放流通对许多行业的发展都有着强力的推动作用，尤其是在社会发展、人民福祉相关的行业，数据的使用可以很大地推动对现状的了解及研究，进而做出预测及防控。这可以从医疗、制造、交通、科学研究几个方面可见一斑。

2.2.1 医疗健康大数据

随着国民经济水平的不断提高，民众对医疗健康的要求也越来越高。在互联网时代，通过对医疗健康领域数据的开放共享，可以有效提高国民健康管理水平、提高医疗机构服务质量和效率、提升我国医疗健康水平，打造“健康”产业生态圈，从而有效缓解“看病难、看病贵”等问题，做到“人人健康，健康人人”。

（1）提高国民健康管理水平

在传统医疗环境下，多数患者出现明显病症才就医，但此时往往已经错过最佳治疗时机，不仅增加医疗费用，更重要的是要忍受更多痛苦，甚至错失最佳救治时机。这种被动就医的健康管理方式可以通过互联网和大数据等信息技术手段，变被动治疗为主动疾病预防，实现战略前移，防患于未然，而医疗服务的重心也将从短期急性病医疗向着慢性病治疗和预防性的健康保健转变，改变国民健康管理方式。

目前，我国慢性病人口多，数字显示，我国高血压、糖尿病患者分别达到了2.7亿和9200万人。患者可以通过使用可穿戴设备，随时随地进行自我健康管理，并将正常医疗流程无法获取的数据转换为实时数据流，为及时预防疾病奠定基础；医生可根据患者发送的健康数据，及时采取干预措施或提出诊疗建议；同时，可穿戴设备收集身体各项体征数据后，借助云计算、大数据技术，通过数据开放共享、挖掘寻找日常生活行为与疾病的发生联系，使患者可以及时享受专业医护人员的各种健康咨询、筛查、预防、监护和干预服务。

医疗健康数据的开放、共享和分析，也有助于实现自身个性化、精准化的健康管理。一方面在患者身体从健康状态到亚健康状态的转变中，能够得到及时的提醒，并得到个性化、有针对性的指导建议，防患于未然。既减少了患者的病痛，又减轻了其经济压力；另一方面，

根据患者具体的身体情况，进行个性化的治疗方案设计，真正做到医疗服务的个性化和精准化，实现传统环境下难以完成之事。在医疗大数据的支撑下，可以个性化、精准化地为患者提供服务。

（2）提高医疗机构服务质量和效率

医疗领域的服务模式在互联网时代正在发生翻天覆地的变化，医疗机构在长期经营过程中积累了大量的数据，包括患者的体征数据、疗效数据、分析诊疗操作与绩效数据以及费用数据等，通过对这些医疗数据的挖掘、开放和共享，将产生巨大的价值，重构就医方式、改善就医体验、重构购药方式、重构医患生态，提高医疗服务效率，降低医疗费用，使患者享受安全、便利、优质的诊疗服务。

数据的开放共享数据可以有效缓解医疗资源匮乏的现状，实现医院管理法制化、科学化、规范化、精细化、信息化。目前，我国医疗基础设施不健全，医疗资源配置不合理并且医疗资源匮乏，导致患者预约挂号难、医院拥挤不堪、医疗效率低下、医疗服务质量低等问题频发。首先，不同医疗机构的数据开放共享，医生可以快速调取病人此前在其他医院就诊的病历和处方信息，减少了对病人的重复检查，有效缓解患者“看病难、看病贵”等问题；再次，通过诊疗和决策系统，将可有效扩展临床医生的知识，减少人为疏忽，帮助医生确定最有效和最具有成本效益的治疗方法，提高医生工作效率和诊疗质量；最后，在线问诊和远程医疗借助互联网实现有限医疗资源的跨时空配置，提高患者、医疗服务机构和医生彼此之间的沟通能力。

（3）提高我国医疗健康水平

首先，随着电子病历系统在医疗机构的迅速普及，大量医疗相关的重要信息以电子形式存储于医疗信息系统中，例如：病人的主诉、检测结果、诊断信息、服用药物等。经过不断积累，各种形式的电子化医疗系统将产生体量庞大的医疗大数据，通过这些数据的开放共享，将更好为企业提供决策和支持，促进我国医疗企业健康快速发展。

例如，对于药物研发企业来说，通过分析临床试验数据和电子病历，辅助药物效用分析与合理用药，降低耐药性、药物相互作用等带来的影响。通过分析疾病患病率与发展趋势，模拟市场需求与费用，预测新药研发的临床结果，帮助确定新药研发投资策略和资源配置；对于医药电子商务企业来说，通过对海量医疗数据中患者的分析，大力整合医药供应链资源，占据患者流量、品牌黏性等优势，明确未来发展方向。

其次，医疗健康大数据的开放共享，是对新形势下医改的重要探索，也是实现智慧医疗的重要手段，可以推动医疗体制改革。

最后，医疗健康大数据的开放共享可以推动医疗卫生信息化发展。“大数据”的数据资源分散在六大方面上，分别是医疗服务的EHRs数据，医院与医保的结算与费用数据，医学研究的学术、社会、政府数据，医疗厂商的医药、医械、临床实验数据，居民的行为与健康管理数据以及政府的人口与公共卫生数据。利用物联网、大数据、云计算等技术，将不同层次的医疗机构连接起来，实现数据的互联共享，可以有效节省资源的投入，推动医疗卫生信息化发展。

（4）打造“健康”生态圈

随着互联网时代的发展，利用大数据，推动医疗健康产业上下游数据的共享，打造全面满足患者、社会保险、商业保险机构作为医疗付费者和药企、医生、医院作为医疗提供者需求的“健康”生态圈，例如险企和公立医院的医疗数据共享、阿里云的“云上医院”等，可以推动我国医疗健康产业发展及创新，满足国民众多层次、多样化的健康服务需求，同时促进经济转型升级和社会快速发展。

2.2.2 工业大数据

工业是整个国家实体经济中的主要组成部分，同样的根据麦肯锡大数据研究报告显示，涉及工业的数据总量占整个社会的大数据中的绝大部分，但是由于数据一直以来都存放在工业的企业内部没有得到共享，因此数据价值没能得到有效的体现。我国由于工业发展水平参差不齐，管理方式粗放，管理思想落后，大数据的应用水平还十分落后，尤其是与互联网数据相比，工业大数据的充分利用仍然存在十分巨大的挑战。但是，总结历史发展的趋势工业大数据共享是历史的必然，原因如下：

首先，泛在互联网的技术发展会不断加速数据共享的过程。从技术的角度来看，网络基础已经从局域网为基础走向互联网和移动互联网为基础的架构，企业内部局域网和广域网之间已经形成相互渗透的局面，在这种情况下数据共享的技术门槛已经完全消失。

其次，信息爆炸使得企业不可能将所有数据归为己有。从规模角度来看，随着社交应用、电子商务、视频应用等一大批网络应用的普及，目前的大量数据已经从个人的本地硬盘转移到互联网上成为可共享数据，而对于工业企业而言，如果将所有和企业有关数据纳入内部进行自行管理是不可能的，随着时间的推移，数据在受保护的情况下实现共享将是必然趋势。

最后，新一代劳动力主体将营造数据的共享的氛围。从变化角度来看，根据中国互联网信息中心的统计显示，截至2015年6月，我国网民规模已达6.68亿，其中10岁~30岁的人群为绝对主力，这意味着：上网已经成为未来劳动力主体的行为习惯，而共享意识必然成为他

们工作时的核心诉求。

对行业来说，数据共享同样具有非常重要的现实意义，可以在企业升级、调整企业结构、催生新型产业、提高宏观决策水平等方面得到促进。

首先，数据共享加速企业转型升级。当今的工业企业大都采用科层制的组织架构，随着组织结构的不断加大，科层之间信息的横向传递速度非常缓慢，生产力被严重束缚，数据的共享，将使企业僵化的管理制度受到巨大的冲击，并促进企业把原来像金字塔一样的科层制变成一个网络性的组织，进而加速企业转型升级的过程。

其次，数据共享可以促进产业结构调整。数据共享后的另一个重大意义是信息透明，而信息透明带来市场的双向选择，一方面，压缩了由于地方保护壁垒导致行业落后产能的生存空间，另一方面，提升了用户的辨别能力，从而实现“市场配置资源的决定性作用”，最终达到加速落后产能的自然淘汰，调整产业结构的目的。

数据共享也将有助于催生新型业态。通过数据共享不仅对行业以及行业内部的企业具有非常重要的意义，随着工业大数据共享，将其与社交数据、地理数据、金融数据、产业链上下游其他行业的数据进行交叉融合，必将引起跨界经营的革命以及行业整合的机会，由此带动的新型业态也会推动社会整体经济发展。

最后，数据共享推动宏观决策水平。工业直接关系国计民生的方方面面，这些点点滴滴的数据无不反映着全社会经济运转的情况，对这些数据进行分析将使得政府的宏观经济政策治理水平提升到一个更高的水平。

利用共享数据对企业自身的意义，通过海尔集团的一个案例便可见一斑：2015年11月5日，海尔集团董事局主席、首席执行官张瑞敏作为唯一受邀的中国企业家，在维也纳出席了第七届彼得.德鲁克全球论坛，并在这一全球知名的管理论坛上分享了海尔平台化互联网转型利用共享数据的一个案例：2013年，海尔发现在百度百科上有1500万条用户提出“洗衣机内桶如何清洗”问题的词条，随即通过创意大赛征集解决洗衣机内筒脏这一问题的创意方案，吸引了990多万用户参与，15万用户在平台上进行交互，活动结束时总共收到846个创意方案，最后通过用户投票筛选了10个方案。在吸引用户参与交互的过程中，全球一流资源也参与了进来，来自全球的21个专家团队尝试了120多种材质、300多种尺寸、1000多种形状，并进行了近20万次的模拟测试、5000个周期的寿命试验，最终形成了第一代“免清洗”洗衣机的解决方案。在这个案例中，通过数据分析和数据共享不仅发现了创新的机会而且聚集了优质的研发资源。因此，张瑞敏提出了互联网+改造计划，企业向网络化组织转型后，就没有上下级，只有两类人：第一类人叫做平台主，他（她）不是领导，是做创业平台的，为大家能够

在这个平台上面成功创业而提供服务；第二类人叫小微主，小微是一个微型的创业团队，一个小微最好不超过8个人，小微创业团队在平台上充分地创业。

红领集团的案例是利用数据共享，进行工业智造的案例。集团将顾客需求标准化，然后根据标准化后的需求进行柔性化定制，完成传统服装制造业转型的第一步。红领正进行互联网+的平台建设，将其它中小型服装制造商的生产系统和客户系统对接到该平台，期望将分布在全国各地的中小型服装企业产能利用起来，并为各自的客户服务。这样的网络式开发、分布式制造以及大规模定制，使得原有产业运行方式被打破。利用数据的开放和共享，原有的企业边界被打破，产生了新的制造业模式。

宏观决策和跨界经营可以从三一重工的实例来看。三一重工是国际工程机械装备制造的龙头企业，虽然其是工程机械制造商，但是这家企业在一开始就敏锐地意识到数据的重要性，2008年开始组建智能研究院自主研发工程机械智能控制器件和传感器，并研究工程机械的状态数据采集整套解决方案，从2011年开始，三一联合清华大学软件学院研究面向工程机械工况数据管理的工业大数据平台和分析方法。截止笔者撰稿时间为止，三一实时接入的工程机械设备已达到20万台以上，高峰时段瞬时活跃的设备达到1.2万台，每日录入的工况数据达1.2亿条，累计车辆的位置数据、开工数据、报警数据达到千亿条以上，可谓名副其实的“大数据”。在这些数据的支持下，一旦主机发生故障，三一企业控制中心ECC的二线服务工程师可随时调集最近的服务车赶往现场，同时将主机开工数据发给维修人员，在业界实现了接单“两小时到底，24小时完工”的服务承诺，达成了服务大数据的初步价值转化。然而，在服务数据的利用方面，三一已经跨出狭义的维修服务界限，开始探索基于服务大数据的延伸应用，包括租赁服务及宏观决策应用。

提供融资租赁服务，延伸经营范围。工程机械产品具有总量大、分布散、价值高的特点，无论是主机采购还是主机租赁均需要大量资金投入，这些特点催生了广阔的工程机械金融服务市场。传统银行的信贷体系和业务运作方式无法实现对客户的有效评估，因此无法为工程机械用户提供优质的金融服务。而作为主机厂商的三一具备大量客户资源，同时又积累了海量开工数据，可以形成具有天然的行业特征的融资租赁的基础信用数据，在服务大数据的支持下对客户和设备进行信用评估可以进而形成工程行业互联网+金融的典型应用。

深化数据分析，支持宏观决策。三一重工在平台建设过程中，积累了全国各地车辆的大量开工数据，这些开工数据将间接反映各地的基础设施建设和固定资产投资的情况，这些信息不仅可以帮助企业自身提前实现战略决策的优化，同时也可以对政府、科研和普通民众提供了解宏观经济形势的依据。例如：三一重工利用车辆开工热度指数作为房地产与基础设

施建设的重要宏观经济指标每月按要求呈报国务院就是典型案例，目前三一正在和清华合作以此为基础打造一套完整的区域宏观经济指数体系，计划在北京、湖南等地区开展试点，建立工程机械行业支持宏观决策的典型应用。

2.2.3 交通大数据

智能交通系统 （Intelligent Transport System，简称ITS）将先进的信息技术、数据通讯传输技术、电子传感技术、电子控制技术以及计算机处理技术等有效地集成运用于交通运输管理体系，以有效的信息采集、处理、分析、发布、利用为手段，建立起的一种在大范围内、全方位发挥作用的交通与运输综合管理系统，为交通参与者和管理者提供多样化、智能化的服务。交通数据共享对智能交通领域发展具有重大的推进作用。

智能交通系统作为一种实时、高效、准确的新型交通运输系统具有重大的社会意义和经济价值，目前在欧美等发达国家正得到广泛应用。有关数据显示，应用智能交通系统后，可有效提高交通运输效益，使交通拥挤降低20%，延误损失减少10～25%，车祸降低50～80%，油料消耗减少30%，废气排放减少。国外的研究表明，智能公路系统可极大地提高公路的通行能力和服务水平，缩短行车时间35%-50%，是解决交通问题的关键技术。此外，智能公路系统可以大大提高公路交通的安全性。智能公路系统降低并排除了人为错误、驾驶员心理对交通安全的消极影响，使预防和避免交通事故成为可能。从理论上讲，智能公路系统可以减少事故31－85%。因此，ITS系统不仅为解决交通拥堵、改善交通服务、监控道路环境、缓解能源紧缺等社会问题提供了新的机遇，而且对认知人们的社会活动、优化公共资源配置有着特殊意义。世界各国都在大力发展智能交通，将智能交通提升到了国家发展战略层面。我国政府和企业也在逐年加大对智能交通系统研发的投入，这些系统一定程度缓解了交通拥堵、环境污染等问题。

随着各类交通大数据被采集并保存，以数据为驱动的交通管理智能化已经成为ITS系统的核心。数据被视为同资本、能源同等重要的生产要素。高质量的交通数据为解决交通拥堵、改善交通服务、监控道路环境等问题提供了新的机遇，成为各地政府与企业的重要财富而受到广泛重视。例如，大规模GPS轨迹数据中蕴含了群体对象的泛在移动模式与规律，有助于理解交通演化的内在机理；手机的蜂窝定位数据中包含了人群的分布、移动和相关行为等信息；通过交通卡数据可以分析公共交通流量等特征，引导公交线路、班次的优化；通过位置服务网站和社交媒体数据分析，我们可以了解用户的出行意图，实现合理的出行推荐；大型活动、事故会对交通造成影响，可以通过相关数据分析评估它们所造成的影响及其演化；同时，结合气象数据的交通分析使我们能够了解不同天气条件下的交通规律，实现更加精准的

预测与导航。不难看出，交通数据具有多样、海量等特征，具有重要分析和应用价值。在ITS系统构建中，人们期望能够汇总这些数据并使之高效地处理，从中挖掘交通模式、车流规律、拥堵演化等关键知识，从而优化车辆导航、行程推荐、城市规划等业务应用。

当前，交通大数据的开放与共享已经成为推动智能交通领域发展的核心问题。交通大数据之“大”，在于通过不同来源、类型的海量低价值密度数据的叠加、整合与综合分析，产生丰富而全面的交通知识，创造新的价值。现有的交通数据已经具备大幅提升ITS系统的能力。但是由于各企业与部门只拥有较为有限的数据，数据在现实中以大量数据孤岛的形态存在，意味着这些数据背后的潜在价值未能被充分挖掘与利用。例如，某地图搜索或导航服务公司未有气象、事故、大型活动等数据，也未有不同情景模式下的出行和路况规律等，因此无法对移动对象GPS轨迹数据进行深入理解；如果能够综合分析区域内的实时蜂窝定位、GPS轨迹等数据，可以进一步对踩踏等非常规突发事件实施及时的预警、监控和主动干预，将相关信息在道路LED显示屏、智能移动终端等设备上推送发布，保证人民的生命财产安全。此外，各企业、机构所拥有的轨迹数据、路网数据、兴趣点数据等通常存在时空不匹配等问题（即对应区域、时间版本不一致），致使海量数据无法利用。这些现象极大制约了智能交通领域发展，亟需加速推动交通领域的数据开放与共享，使得数据能够有效转化为领域知识并驱动应用创新。具体来说，交通数据的开放与共享可以为以下部门或企业提升交通服务：

（1）对地图与导航服务商，可以从交通大数据中提取不同天气、事故条件下的泛在移动模式，挖掘不同环境下的高质量路径，识别施工道路，更加深入地理解用户查询，根据实际情况匹配多样化的“智慧”路径，提升交通导航服务的效率；

（2）对交通管理相关的部门和企业，通过交通大数据挖掘理解交通演化机理，综合运用时空、活动、天气等多维相关信息，构建数据驱动的交通指挥体系，做到先知先觉的指挥决策；分析大规模手机的实时位置信息，监控高密度人群，对事故进行预警，避免上海外滩踩踏等事故。

（3）对城市规划部门，交通大数据分析可以帮助理解城市交通网络，了解各区域的功能特点，不同区域之间的车辆流动特征，指导城市交通网络的开发、建设和管理，更好地协调城市的空间布局；

（4）对出租服务管理平台，交通大数据共享可以促使服务商更好地理解用户的行为特征，准确预测不同区域的叫车请求量，合理匹配出租车与用户，引导车辆只能运行以减少空驶，为用户推荐合理的打车地点以避免等待；

（5）对搜索引擎服务商，可以根据相似用户的签到、轨迹、评论等信息，结合实时路

况、开放时间等信息，为用户推荐能够满足搜索意图的服务商、路径和旅行路线。

交通数据的开放与共享已经受到了广泛的关注，一些政府部门和企业机构开展了数据共享工作，取得了喜人的初步成果。在政府层面，多个城市的政府部门在交通数据开放方面进行了尝试。上海市经信委、市交通委联合主办了SODA开放数据应用创新大赛，开放了一系列交通数据，包括交通路网、城市道路交通指数、公交一卡通、出租车轨迹、气象资料、事故数据等。这个比赛吸引了大量社会力量参与，为改善上海市的城市交通、市民出行、商业模式、充电柱选址提供解决方案。原本封闭数据的开放激发了超乎想象的创新活力和价值，共有三千多参赛团队针对交通管理、智能导航、出行规划、公交运营优化等不同“交通难”问题提供了宝贵思路。在企业层面，数据堂免费开放了一些交通数据，并提供定制化交通数据的收集、处理有偿服务，这为业务分析、智能交通系统开发提供了极大便利。此外，微软亚洲研究院公开了一些经过预处理的高质量GPS轨迹等数据，这些数据促使学术界开展了大量的智能交通领域研究，提出了一系列交通数据的存储与计算框架，设计了热点路径挖掘、交通拥堵模式学习、城市功能区划分、服务商合理选址、公交线路优化、道路设计缺陷识别等算法，使得这些方法可以被验证、比较，最终服务于高水平的ITS系统研发工作。

2.3 科学数据的共享对科技发展的驱动

科学数据是人们在科技活动中所产生数据的总称，它包括原始的观测数据、实验数据、调查数据、统计数据、研究数据等。作为一种特殊的数据资源，科学数据体量巨大，且经过特定的处理后，可转化出巨大的价值：它不仅是科技创新、经济发展和国家安全的重要战略资源，也是政府部门制定政策、进行科学决策的重要依据。

科学数据的共享、开放和流通，对于增强人类社会的科技发展，提高科技创新能力，提升全社会的科技整体水平，促进经济发展具有十分重要的意义。近年来，科学数据被视为人类社会可持续发展的一种重要资源，很多国家和组织已开展了科学数据共享、开放方面的尝试。科学数据的开放、共享和流通受到日益广泛的关注，其原因有三：

1）科学数据的共享和流通是国家战略的需要。科学数据是现代科学可持续发展的重要资源。一个国家和企业的发展在很大程度上取决于其科学创新和技术创新的水平，而对科学数据进行系统的综合利用和分析是实现其科学技术长期快速发展的重要手段。无论是资源的综合开发利用，还是高新技术产业化，无不是在科学数据的积累与支持下，实现理论与技术创新的结果。特别是在以知识为基础的经济中，使越来越多的知识产品以各种方式驱动着经济的快速增长，以知识、信息和数据应用为主要目的的信息管理、加工与发布成为迅速发展

的产业，在现代信息技术引领下正在拉动“数字经济”，也就是现在正在发生的“大数据经济”。长期以来，我国已经积累了较为丰富的科学数据资源，但大多数仍存于资料堆和档案柜中，没有经过有效的整理，数字化程度较低，很多数据库往往局限于本部门、本单位使用，甚至个人使用，造成了科技资源的巨大浪费。所以打破科学数据壁垒，实施科学数据开放、共享和流通，是国家发展战略的必然要求。2012年3月29日，奥巴马政府宣布启动《大数据研究和发展计划》，同时组建“大数据高级指导小组”，涉及美国国家科学基金、国家卫生研究院、能源部、国防部等联邦政府部门，宣布将启动2亿美元的投资计划，提高从大量数据中访问、组织、收集发现信息的工具和技术水平。这使得美国成为全球首个将大数据从商业行为上升到国家意志和国家战略的国家。我国政府发布的《大数据行动纲要》将数据的开放共享，推动资源整合，提升治理能力提升至国家的战略层面。作为一种特殊的大数据资源，避免科学数据束之以阁，使之合理开放和共享，以产生巨大的价值，是一个国家保持国家科技长期可持续发展的战略需要。

2）科学数据的共享是科学研究的需要。当前的科学是多学科交叉的科学，是围绕数据展开的全球性研究，且在一些科学研究领域中，对数据的依赖愈演愈烈。科学研究的本身就是科学数据的生产过程，一些科学数据就是及其重要的研究成果。科学数据资源既是研究的成果与积累，又是支持更为复杂的创新研究所不可替代的资源存量。尤其在大数据时代，科学数据的数据量激增，科学研究越来越依赖于系统的、高可信度的基础科学数据分析。21世纪以来全球科技活动不断增强，一系列复杂科学问题研究的提出、大型科学研究计划的产生、重大科学工程的兴起，导致了前所未有的国际合作局面的形成，也导致了全球范围内对科技信息资源交流、互通的客观需求。因此，实现科学数据的共享，科学家就可以不再受限于数据的来源、格式以及国界，可在全球海量的科学数据中发掘创新的潜力。

3）科学数据的开放、共享也是公众的需要。大数据环境下，科学数据的需求不仅仅局限于政府、科研单位以及企业，社会公众也越来越需要科学数据。随着科学研究与人们的日常生活结合得越来越紧密，在科学研究中用到的很多数据就是来自于人们的日常生活，这些数据不仅对研究有用，对于社会公众提高自身知识水平和科学素养，推动万众创新具有重要的作用。如今个人电脑、智能手机及其他掌上智能设备的普及，互联网的应用和发展，使得公众对这些基本的科学数据获取的需求更为强烈。例如随着智能手机的普及，许多驾驶员使用手机装载的定位系统确定行车路线。和传统的定位系统不同，这些通过智能手机定位的信息都传递和保存在大数据库中。这些海量数据不仅能像传统的交通信息一样让人们了解某一个时段一条路上的车流量，还能明晰的标示出这条路上每个时段的每一辆车从何处来、往何

处去，并记录每辆车的停车情况。同时，现有技术也能够支撑信息的反馈，即可以向车辆驾驶者和乘客发布拥堵预警、拥堵状况和停车场分布和占用情况等信息。同样地，在医疗健康领域，大规模复杂数据已经变得很普遍，通过对大量病人的各类数据进行挖掘分析，有助于更有效地找出疾病成因，进而提供有针对性的预防、诊断和治疗措施。尽管社会公众大多数是非专业人士，但可见在大数据时代，公众对科学数据的质量要求是越来越高，对科学数据的发布渠道、发布频率、发布质量和表现形式等的要求也会越来越高。

第3章　国内外数据开放共享现状

3.1 国外数据开放共享现状

近些年来，全球各国纷纷将数据开放纳入到国家发展战略。截至2014年，全球已有63个国家加入开放政府联盟，并制定了开放政府数据的纲领。如：2011年，巴西、印度尼西亚等八个国家就联合签署了《开放数据声明》，成为开放政府合作伙伴；2013年，八国集团签署了《开放数据宪章》；2013年6月，欧盟颁布对2003年《公共部门信息再利用指令》的修订指令；2013年2月，美国颁布《增加联邦资助的科研超过访问的政策》，2013年5月，奥巴马签署《政府信息公开和机器可读行政命令》；2013年6月，日本颁布《日本再兴战略》，提出开放数据；2013年8月，澳大利亚政府信息管理办公室（AGIMO）发布《公共服务大数据战略》，以六条“大数据原则”为支撑，旨在推动公共行业利用大数据分析进行服务改革，制定更好的公共政策，保护公民隐私，使澳大利亚在该领域跻身全球领先水平。

从世界各国的实践来看，建立统一的公共信息资源开放共享网站，集中开放可加工的数据集已经成为了一个通行做法。如美国的data.gov网站、新加坡的data.gov.sg网站、印度的data.gov.in网站、西班牙的datos.gob网站等。

3.1.1 美国

美国政府最先对大数据革命做出战略反应的。2009年，美国联邦政府发布《开放政府指令》，作为大数据的前奏推出了“data.gov”公共数据开放网站。2012年3月，美国联邦政府发布了《大数据研究和发展计划》，正式启动了“大数据发展计划”，宣布将投入超过2亿美元在大数据研究上。同年5月，联邦政府发布《数字政府战略》（Digital Government Strategy），致力于为公众提供更好的“数字化”服务，围绕数据进行的一系列措施在美国政府全面推进，大数据对美国政府的影响逐步显现。

美国通过立法赋予社会公众数据获有权，即“公民对于政府数据有获取权”是公民的基本权利，是受美国宪法保护的权利。美国通过立法、战略等措施，设计政府数据的开放原则，比方说政府数据必须以公开为原则，以不公开为例外，政府数据面前人人平等，政府拒绝提供信息的时候必须有举证责任，司法有重新审定政府数据开放事实的权利。美国采取了类似于负面清单的方式，规定了数据开放的范围，列举了除国防、外交、内部人士消息等九类信息之外其他数据都必须开放。

2013年5月9日，奥巴马总统签署第13642号总统行政令，对联邦大数据管理工作提出了新的准则，提出在保护好隐私安全性与机密性的同时，将数据公开化以及可读写化纳入政府

的义务范围。2014年5月1日，美国总统行政办公室向奥巴马提交了一份名为《大数据：把握机遇，维护价值》的报告，阐述了大数据带来的机遇与挑战。报告认为，大数据技术为美国经济、人民的健康和教育、能源利用率以及包括信息安全在内的国家安全等提供了难得的机遇。同时，报告也指出了大数据为美国隐私保护、信息安全和社会发展带来了新的挑战。在这些战略框架中，基本都考虑了大数据对既有法律制度的挑战和相应对策。

白宫科技政策办公室领导下的开放数据行动（Open Data Initiative）很大一部分内容便是跨部门整合以针对某个问题整合相关数据，以美国政府内被认为较为成功经验的健康数据为例，美国国家卫生研究院和美国卫生与人口服务部于 2010 年共同发起的“健康数据计划”，是政府与私营部门的一次典型协作。美国政府专门制作了“健康数据网站”，使企业家、研究人员、政策制定者们可以更加方便可及地获取高价值的健康数据。“健康数据计划”向公众开放了美国疾病防治中心、食品药品管理局、国家卫生研究所等机构的诸多数据，开放的信息包括临床医疗服务提供者的服务质量信息、全国医疗卫生服务提供者名录、最新的医疗和科学知识数据库、医用消费品数据、社区卫生服务状况信息、政府相关开支等。除了公布最新的数据，政府也致力于使现有数据更可及、更容易下载和转化使用，与此同时也努力保障个人隐私和机密信息。

从美国的实践来看，国家将信息视为资产，鼓励社会各界发掘其中的经济价值，发挥了公共信息资源的作用。通过《开放数据政策——将信息作为资产进行管理》战略、《大数据研发倡议》，美国政府以政策战略为动力，驱动数据进入生产流通环节。在此基础上，政府鼓励社会各界从日益增长的数据资产中进行经济价值挖掘。如汇聚了美国顶尖大学各类优质资源的大型开放式网络课程（MOOC）面向更多学生提供系统学习，降低了优质教育资源的扩散与传播成本，已经对提升美国智力自由、推动美国知识创新产生了积极而巨大的潜在影响。

2008年，美国总统行政办公室和科技政策办公室联名向农业部、商业部、能源部、教育部、国土部、内政部、交通部、自然科学基金委、美国国家卫生院等部门发出《科学研究成果发布原则》。该原则根据《美国竞争法》第1009的规定，同时，也是对科学研究成果的提交的方向性指导意见。同时，美国国会也指示“这些原则是为了确保科学研究的成果和数据可以有效开放、共享和流通”。《原则》从核心原则和配套原则两个方面进行了描述。在具体配套原则中，规定了科研机构在与新闻媒体交流所提供的数据，以及科学数据、结果开放共享的基本准则。

A. 与新闻媒体的交流。机构应该根据其科学活动及其结果而传播切实可行及合适的客

观信息；机构应该就其雇员如何与出版社及大众沟通制定相应的管理政策，并在必要时修订这些政策；当雇员以机构之名传达官方信息时，这些政策应该保证以下几点：a）提供的科学内容是准确的，并用文本或者其他尽可能完整和及时的方式提供给受众；b）在不违反现存的法律法规及分类限制时，机构雇员可以免费公开地和大众讨论其基于官方工作的科技思想、方法、发现及结论；c）所有联邦雇员都应该将其个人看法从工作岗位上区分开。并且应该有相应的程序来确保这种区分被完全地执行；d）机构所指派的政策制定者有责任来决定机构定位，并且负责将这些机构政策传达给公众。科研机构与媒体交流的政策应当包括解决争议的明确程序，并且保证所有雇员就科学信息交流相关事宜都有上诉的渠道；机构应该致力于保证组织内科学家、工程师及公共事务员工之间的合作及协调，并且在机构内部的指导方针上增加这些事宜及相关的联邦法律法规。

B. 联邦科学家的研究数据和成果的开放交流。在不违反现存的联邦法律法规、总统签令及相关研究领域的现存惯例时，联邦机构的科学家所产生的研究数据应该最大化地公开：a）机构应当就如何分享联邦科学家所产生的研究数据及结果制定明确的指导方针，并且在毕业时修订这些方针。这些方针应当与《信息质量指导》保持一致；b）制定指导方针时，机构应当就如何保存及获取这些公开数据的事宜制定明确的政策；c）机构应当时刻注意并遵循这些指南，并且保证对大众公开的数据是准确的、完整的，及时的。同时，同行的评论可以为科技研究企业以后验证研究数据及结果的可信性提供重要作用，所以经常会被一起发布。机构应该采取措施保证同行评论是以不违反相关研究领域现存惯例的方式获得的。

美国国立卫生研究院（NIH）在定义自己的数据共享政策时，将“数据”定义为“无论是以可记录的形式还是以媒介作为记录信息，包括文字、电影、录音、图片复制品、图纸、设计、或其他图形表示、程序手册、表格、图表、工作流程图、设备描述、数据文件、数据处理或计算机程序（软件）、统计记录等”。

美国国家科学基金委希望研究者与其他研究人员共享一些在不超过边际成本或在合理的时间内可以获得的数据、样品、实体馆藏和在工作过程中其他辅助材料。它还鼓励获奖者共享软件和发明或以其他方式采取行动，以使它们体现广泛的、有用的和可用的创新。同时允许一些调整和必要的措施，以保障个人和主体的权利、结果的有效性、集合的完整性，或保证调查的合法权益。

同时，一些美国部门还成立了科研诚信办公室，负责监督对不当行为的指控。如美国卫生与人口服务部的网站上将ORI的任务定义为：“研究诚信办公室（ORI）在全球拥有约4000机构来支持生物医学和行为学研究的完整性。ORI监测科研不端行为来帮助机构调查，并通

过教育，预防和监管活动来促进研究的规范责任行为。”

为此，美国国家自然基金委员会（National Science Foundation，NSF）在其《奖励和管理指南》（Award and Administration Guide）第四章专门就研究成果的传播和共享问题进行阐述。2010年，NSF声明更改其数据共享政策，要求自2011年7月18日起，所有申请NSF资助的项目计划要以两页补充文件形式提交研究项目的数据管理计划。类似地，美国国家医学卫生研究院（NIH）专门就科学数据共享问题发布一系列政策。

3.1.2 欧盟

2010年3月，欧盟委员会公布了《2020 战略》，认为数据是最好的创新资源，开放数据将成为新的就业和经济增长的重要工具；2011年11月，欧盟数字议程采纳欧盟通信委员会《开放数据：创新、增长和透明治理的引擎》的报告，开始推进开放数据战略，该战略从三方面对原有法律、政策进行修订与补充：（1）建立适应信息再利用的法律框架，对公共部门信息再利用指令修订的决定；（2）动用金融工具，以支持开放数据和行动作为建立欧洲经济数据门户的部署；（3）促进各成员国之间的协调与经验交流，为开放数据与共享提供平台。计划于2012年春建立欧洲开放数据门户网站，提供委员会和欧盟其他机构的数据访问，2013 年春建立泛欧洲的数据门户网站，允许访问整个欧盟自2011年起所有成员国的数据，保证公众可以自由获取这些创新资源。

2012年10月，欧洲委员会提出《云计算发展战略及三大关键行动建议》，三大关键行动为：（1）规范和简化的云计算标准；（2）云计算安全和公平的合同条款及条件；（3）建设欧盟云计算伙伴关系，驱动创新和增长。其他的具体行动举措还包括：数据保护、网络安全、信任举措、云计算互操作性、宽带部署、在线服务、公共行业首先参与云计算和国际对话与合作等。欧盟这些战略部署成为之后欧盟及其成员国数据立法的基本路线图。

欧盟专门在2014年发布了《数据驱动经济战略》，有望近期内成为欧盟经济单列行业，为欧盟恢复经济增长和扩大就业，做出巨大贡献。欧盟在大数据方面的活动主要涉及两方面内容：（1）研究数据价值链战略计划；（2）资助“大数据”和“开放数据”领域的研究和创新活动。

数据价值链战略计划包括开放数据、云计算、高性能计算和科学知识开放获取四大战略，主要原则是：高质量数据的广泛获得性，包括公共资助数据的免费获得；作为数字化单一市场一部分，欧盟内数据的自由流动；寻求个人潜在隐私问题与其数据再利用潜力之间的适当平衡，同时赋予公民以其希望形式使用自己数据的权利。

3.1.3 日本

日本2004年制定《行政信息电子公开基本方针》，日本各部委基于此方针将各自持有的部分公共数据在网站上公开，但是由于未涉及重要数据且形式不统一，公开的数据未在民间得到有效利用。2012年7月，日本的《电子行政开放数据战略》指出需以便于二次利用的数据形式公开数据，同时兼顾商业利用，消除公共数据在商业利用中的障碍。由内阁官房和经济产业省、总务省主导战略实施，内阁官房负责数据标准化方面的工作并提供使用公共数据所需的工具。日本经济产业省已先于其他部委实施公共数据对外公开，并提出“DATA METI”构想。在实验性地进行数据开放后，经济产业省还将与其他部委、地方政府共享数据开放的经验，开设网站公开各种统计信息、政策数据，并提供相应的工具和应用程序。总务省将构筑融合医疗、农业、行政等不同领域数据的信息合作平台，以提供新的数据利用服务，外界可以利用应用程序获取各种数据，而且还将面向交通和灾害领域提供各种数据应用服务。

除政府部门外，民间团体和地方政府也在积极推进公共数据的灵活利用。2012年7月，51个企业、团体和6个地方政府共同组建开放数据流通促进联盟，旨在通过产学官合作推进公共数据的有效利用。联盟设置三个委员会，负责拟定公共数据利用所需技术、许可以及公共数据普及等事宜，对于已经公开的公共数据，需建立可二次利用的机制和许可体系，尽快建立可自由使用的公共数据环境。

2012年7月日本非盈利性机构Open Knowledge Foundation（OKF）诞生，旨在促进公共数据的有效利用，其开发的开源软件CKAN对于构筑地方政府的公共数据门户网站起到重要支撑作用。

3.1.4 新加坡

新加坡数据开放网站（data.gov.sg）是世界上发展最为完善的公共信息资源开放共享网站之一，目前已经汇集了来自68个政府部门和机构的8600多个数据集，实现了全国范围内的整合。该网站具有几个鲜明的特点，一是网站具有指导性的开放共享原则，即数据简易访问原则、可共同创造原则、及时发布原则、可机读格式原则、原始数据原则。二是网站具有简洁清晰的页面，整个页面仅有菜单栏、数据查询区块，以及常用数据概览区块组成。三是网站具有庞大的数据体系，包括39类、8600个数据集，111个应用程序，以及不同机构的地图相关API、陆路交通管理局的交通相关API、图书馆相关的数据资料和国家图书馆管理局Web服务三类开发者资源。四是网站具有多样的数据查询方法，包括通过搜索过滤选项进行查询，按政府机构进行查询，以及通过关键字搜索框搜索查询三种方法。五是网站具有查询式的服务指南，包括门户功能服务指南和技术查询服务指南。

新加坡数据开放网站于2011年6月启用。其中，OneMap是一个地理空间数据共享平台，目前有60种不同的地图主题。利用这些开放数据企业和部门已经开发了100多项应用，涉及停车信息、公共厕所、野猫管理等。2013年8月中旬，新加坡政府宣布将在2013年底之前开放更多数据，并使得数据支持OneMap平台的机读要求，以促进社会创新与协作。2012年，新加坡政府还公布了《个人资料保护法》（PDPA），旨在防范对国内数据以及源于境外的个人资料的滥用行为。新加坡政府积极推进数据公开，新加坡土地管理局（Singapore Land Authority）为基于位置的服务（LBS）的企业提供了开放数据平台，新加坡陆路交通管理局通开放交通数据，鼓励企业或个人开发提升公共交通效率的应用软件。

3.2 我国大数据开放共享现状

随着数据治理理念的影响逐步渗透，我国公共数据开放共享进程开始逐渐加快。2011-2013年陆续上线的国家数据（http://data.stats.gov.cn/）、北京市政务数据资源网（www.bjdata.gov.cn）和上海市政府数据资源网（www.datashanghai.gov.cn）。然而，总体而言与发达国家还有非常大的差距，据“开放知识基金会”发布的《2013年开放政府数据普查》结果，在被普查的全球70个国家和地区政府中，我国综合排名第35位，与我国经济大国和数据大国的身份极不匹配。近年来，我国在大数据开放方面也做了大量的工作，国家制定了一系列的政策，各地方政府、社会各行业积极响应，推动大数据的开放与共享。以下将从国家政策层面、地方政策与措施层面，以及行业层面介绍我国大数据开放共享的现状。

3.2.1 国家政策

2013 年《国务院关于促进信息消费扩大内需的若干意见》中对“制定公共信息资源开放共享管理办法”、“加快启动政务信息共享国家示范省市建设”做出了工作部署，要求促进公共信息资源共享和开发利用，推动市政公用企事业单位、公共服务事业单位等机构开放信息资源。在中央各部门及地方政府的推动下，我国公共信息资源开放共享步伐正在加快。

2014 年 8 月，国务院发布了《企业信息公示暂行条例》，以促进工商部门、其他政府部门、企业的信息公示。该条例一方面可强化对企业的信用约束，另一方面也将有利于数据开发者对企业信息的再利用。被要求公示的企业信息包括企业从事生产经营活动过程中形成的信息，以及政府部门在履行职责过程中产生的能够反映企业状况的信息。

从 2015 年开始，中国政府对互联网、高科技和大数据产业的空前重视，并且明确表态要开放大数据。2015 年两会期间，李克强总理明确表态，政府应该尽量公开非涉密的数据，以便利用这些数据更好地服务社会，也为政府决策和监管服务。这是中国政府首次正式公开

表态支持数据开放。2015 年 5 月，国务院发布《中国制造 2025》，它是我国实施制造强国战略第一个十年的行动纲领，提出“建设重点领域制造业工程数据中心，为企业提供创新知识和工程数据的开放共享服务”。

2015 年 8 月 31 日，国务院印发的《大数据行动纲要》强调要大力推动政府部门数据共享，稳步推动公共数据资源开放，统筹规划大数据基础设施建设，支持宏观调控科学化，推动政府治理精准化，推进商事服务便捷化，促进安全保障高效化，加快民生服务普惠化，明确了大数据领域的十大工程建设。《大数据行动纲要》显示，在未来 10-15 年逐步实现以下目标：2017 年底前形成跨部门数据资源共享共用格局；2018 年底前建成国家政府数据统一开放平台。提出“加快政府数据开放共享，推动资源整合，提升治理能力。2015 年 10 月，中国共产党第十八届中央委员会第五次全体会议通过了《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》。《建议》明确提出“实施国家大数据战略，推进数据资源开放共享”。

当前，我国的数据开放政策仍然处于起步阶段，北京、上海、贵州等部分地方政府先行进行政府数据开放的积极探索。

3.2.2 地方政策与措施

（1）上海

上海对大量政府数据资源的梳理和开放，始自 2012 年，当时，上海在全国率先启动政务数据资源目录编制和开放试点，包括上海市公安局等 9 家单位纳入试点，并以实有人口、法人和空间地理三大基础数据库为基础，开启数据资源开放共享。2012 年 6 月，上海市在公安、工商、商委等 9 家单位开展政府数据资源开放试点，创立国内首个“上海市政府数据服务网”（www.datashanghai.gov.cn），揭开了国内政府数据开放的序幕。

上海数据资源丰富，已经积累并将继续产生庞大的数据资源，在众多领域的重要作用越来越凸显。例如，上海拥有世界最大的医联数据共享系统，有 4800 万张交通卡、每天 30GB 交通流量信息数据，亚洲第二的证券交易额，世界第一的货物和集装箱吞吐量等。但是，《上海推进大数据研究与发展三年行动计划（2013-2015 年）》也提出，数据资源的利用不充分，大量信息系统中的历史数据长期闲置，即使不涉及秘密，许多数据资源拥有单位公开和共享动力不足，这给跨行业数据汇聚整合造成困难，影响了大数据资源的形成。而在产业方面，近年来，上海在数据资源整合、数据技术开发、数据应用服务等数据产业环节涌现出一批机构和企业，已经成为或正在成为推动上海数据产业发展的中坚力量，数据产业初显轮廓。

2014 年起，上海市政府数据资源向社会开放工作进入全面推进阶段，由试点单位拓展

至 44 家市级政府部门，进一步扩大开放范围。2014 年上半年，上海市政府开放了上海市政府数据服务网，集中开放公共数据，以方便企业、个人用户开发利用。上海市政府确定了总计 190 项数据内容作为 2014 年重点开放领域，涉及 28 个市级政府部门，涵盖公共安全、公共服务、交通服务、教育科技、产业发展、金融服务、能源环境、健康卫生、文化娱乐等 11 个领域。截至 2015 年 5 月底，上海政府数据服务网开放内容已基本覆盖各部门主要业务领域，涵盖了经济建设、资源环境、教育科技、道路交通、社会发展、公共安全、文化休闲、卫生健康、民生服务、机构团体、城市建设等 11 个重点领域，累计开放数据资源逾 480 项。目前围绕该网站开放的数据已经有 55 项数据应用和 24 项移动应用。基于这些大数据，相关的数据产业即可开始挖掘新的机会。比如，由于政府数据资源向腾讯等第三方市场主体开放，目前上海市民可以通过微信的“城市服务”来支付生活账单、预约办理护照等事宜。作为全国首个政府数据服务网站，上海市政府数据服务网上的部分数据已被信息服务企业调取利用，数据的经济价值初步显现。

按照《上海市政务数据资源共享和开放 2015 年度工作计划》，相关部门需要结合各领域、各行业实际情况和社会需求，根据年度计划明确的重点开放领域，制定 2015 年度政务数据资源开放清单。其中，以应用程序接口方式开放的数据资源原则上比例不低于年度总数的 30%。同时也提出，要鼓励拓展多元化的政务数据资源开放渠道，相关数据资源在数据服务网开放的同时，可探索与市场化渠道如市民云、微信等的积极合作，拓宽提供服务的方式。

上海市经信委正在研究成立大数据局，成立后将推进上海政府层面的数据公开和信息共享，以解决政府信息资源家底不清、认识不够以及部门间的数据共享不充分等在数据资源管理和运用上存在的问题，其目的是“协调一个城市的信息化建设和做到各部门的数据共享。期待通过深化改革和完善法制，推动政府的数据开放和智慧城市的健康发展”。

（2）北京

北京市经信委也组织各政务部门建设了“北京市政务数据资源网”（www.bjdata.gov.cn），以汇集北京市各政务部门可开放的、有经济和社会价值的数据资源，为企业和个人提供各类实时与非实时数据的下载与服务，促进基于政务数据资源的信息服务产品开发和创新，满足社会公众的信息服务需求。目前，网站已整合了 36 个政府部门，为社会提供了土地使用、教育、旅游、交通、文化、医疗等 306 类数据。通过原始数据（WPS、CSV）下载、带有地理坐标信息的空间数据（SHAPE）下载和在线调用 API 三种形式提供数据开放共享服务。

（3）浙江

浙江省经济信息中心建设有政务数据开发共享平台，其目标是建设全省最综合、最全面、

最权威的人口基础数据库、法人基础数据库和公共数据库、共享空间地理基础数据库、开发数据交换等一系列数据工具和技术成果，围绕应用领域支撑一批各具特色的专题数据库和专题应用系统，建设满足决策支持和公众服务的信息发布查询系统，形成行之有效的运行机制，努力建成全国一流的省级政务数据开发共享平台。

（4）广东

2014 年 2 月，广东省政府出台《广东省经济和信息化委员会主要职责内设机构和人员编制规定》设立广东省经济和信息化委员会 21 个内设机构，其中包括成立广东省大数据管理局，统筹推进政府部门的信息采集、整理、共享和应用，打破信息壁垒，建立公开数据开放机制，逐步公开民生各项数据。其具体职责是：研究拟订并组织实施大数据战略、规划和政策措施，引导和推动大数据研究和应用工作；组织制定大数据收集、管理、开放、应用等标准规范；推动形成全社会大数据形成机制的建立和开发应用；承担企业情况综合工作，负责企业数据收集和存储；组织编制电子政务建设规划并组织实施；组织协调政务信息资源共享；组织协调省级重大电子政务项目建设，组织协调网上办事大厅等电子政务一站式服务建设；负责统筹政务信息网络系统、政务数据中心的建设、管理；统筹协调信息安全保障体系建设；承担信息安全等级保护、应急协调和数字认证相关工作。此后，当年 5 月，广东佛山南海区挂牌成立数据统筹局；随后，广东清远在其经济与信息化局的“三定方案”（定机构、定职权、定人员）中要求设置大数据管理科。

IDA 国际有限公司（新加坡资讯通信发展管理局全资子公司）编制了《智慧珠海 2015 行动计划》。根据计划，珠海拟于 2015 年建立大数据局，负责智慧城市基础性项目的建设和各跨部门、跨领域应用项目的建设工作和运营管理，同时建立首席信息官人才库，推动智慧城市的建设。

（5）沈阳

2015 年 6 月，沈阳市大数据管理局揭牌。该局下设大数据产业处、标准与应用处和数据资源处。其主要职责是负责组织制定智慧沈阳的总体规划和实施方案；研究制定大数据战略、规划和相关政策；组织制定大数据的标准体系和考核体系，统筹推动全社会大数据库建设，组织制定大数据采集、管理、开放、交易、应用等标准规范；指导大数据产业发展；研究制定全市电子政务建设的总体规划、实施方案并组织实施；组织协调政务信息资源共享；统筹协调信息安全保障体系建设等工作。

（6）成都

2015 年 9 月底，成都市大数据管理局正式成立，将创新大数据的应用、挖掘大数据的价值、集聚大数据的成果，推动政府治理能力提升和经济转型升级，这是继广州、沈阳后，

全国第三个设立大数据管理局的城市。该局主要职能包括：负责拟定全市大数据战略、规划和政策措施并组织实施；推动信息数据收集、管理、开放、应用等标准规范，推动信息数据资源和基础设施建设的互联互通、资源共享;制定全市电子政务建设的总体规划并组织实施，牵头组织电子政务项目审核工作；推进电子政务外网现有信息系统整合，组织协调全市信息安全保障体系建设；承担市信息化工作领导小组办公室的日常工作。

目前，成都大数据管理局已按照国务院的《大数据行动纲要》，提出了五个目标：加快推进政务数据资源的共享与开放，加快政府信息平台的整合，创新大数据的应用，挖掘大数据的价值，集聚大数据的成果，推动政府治理能力提升和经济转型升级。

（7）贵州

2014 年 12 月 31 日，贵州省批准成立国内第一家大数据交易所。2015 年初，贵阳市发布《贵阳大数据产业行动计划》。该纲领性计划对贵州的大数据产业发展构建了高层发展设想。2015 年 2 月，国家级“贵阳 • 贵安大数据产业发展集聚区”授牌仪式在贵阳进行。2015 年 9 月，贵阳市交通大数据孵化器正式开通，以国内首创的免费提供计算资源和数据资源方式开放交通大数据。这一系列活动拉开了贵州省数据开放的序幕。

3.2.3 行业政策与措施

（1）健康、医疗行业

国务院发布的《大数据行动纲要》指出在公共服务大数据工程中要构建医疗健康服务大数据。科学推进医疗健康大数据的应用，依法有序地推动医疗健康大数据开放融合及大数据库的建立。医疗健康大数据的地位和作用越来越得到政府和社会的关注，并将作为“十三五”人口健康信息化建设的核心。医疗健康大数据的开放共享，是对新形势下医改的重要探索，也是实现智慧医疗的重要手段。目前，国内一些医疗机构及相关的企业已经迈开了数据分析共享的探索步伐。2014 年，百度与北京市政府联合推出北京健康云平台；2015 年 9 月，广东移动在清远市清城区建立卫生信息化平台，实现了电子病历档案的共享；阿里云将合作组建“云上安心”联盟等。

“国家人口与健康科学数据共享平台”整合了全国人口健康领域科学数据资源，包括基础医学、临床医学、公共卫生、中医、药学、人口与生殖健康和地方医学七大类数据资源。该平台已整合数据资源总量为 18393GB，包括 238 个共享数据集（库）。数据来源以科技计划课题产生的科学数据为主，包括 973、863 和科技支撑计划国家科技计划课题。“十一五”期间人口健康领域科技计划课题总数为 2260 余项，其中基础医学 906 项，临床医学 597 项、药学 332 项、中医药 227 项、公共卫生 197 项、人口与计生 6 项。数据内容包括基础研究、

疾病诊疗、传染病预防、慢性病控制、中医、药学、人口普查、人群调查、营养和生殖健康等有关人口与健康的数据资源。根据用户需求，通过数据加工和数据产品制作，打造精品数据库。如中国人生理常数数据库、传染病疫源地监测数据库、中药化学成分数据库、药品不良反应、药物靶点数据库、ECRI 文库、中国计划生育生殖健康远程教学知识库、农村三级医疗服务等。

（2）交通行业

2014 年交通运输部提出要深化改革，务实创新，加快发展“四个交通”，初步建成综合交通出行信息服务平台，向社会及时发布出行信息，解决出行信息不畅等问题。在上述背景下，依托交通运输部重点科技项目“基于云平台的开放式公共出行信息服务研究与示范”，由百度公司与交通运输部公路科学研究院、国家智能交通系统工程技术研究中心共同打造“百度智慧交通服务合作平台”，旨在激活现有数据，建立部省数据信息资源共享交换机制，促进政企间出行服务信息共享应用，为更广大百姓、交通行业相关部门提供更优质、广泛的智慧交通服务，探索形成综合交通信息服务产业的健康生态环境。

上海市交通委员会在《关于加强智慧交通体系建设的指导意见》中提出“强化大数据管理，推进交通信息资源向社会开放”，“不断扩大交通信息资源向社会开放，建立标准接口或数据平台，优化公共信息资源配置，支持中小企业创新、创业”。2015 年 8 月，上海将首次开放十大领域、总容量达上千 GB 的交通大数据，包括城市道路交通指数、地铁运行数据、一卡通乘客刷卡数据、浦东公交车实时数据、强生出租车行车数据、空气质量状况、气象数据、道路事故数据等，并通过由上海市经信委、市交通委主办的“上海开放数据创新应用大赛”，面向全球征集改善城市交通、便利市民出行、创新商业模式的应用程序和解决方案。优秀项目将在赛后获得投资对接和孵化落地支持，实际运用于解决上海的交通拥堵等难题。

2015 年 9 月，贵阳市交通大数据孵化器正式开通，以国内首创的免费提供计算资源和数据资源方式开放交通大数据，拉开了贵阳市政府数据开放的序幕。结合成熟的科技孵化体系，贵阳市搭建集数据、计算、商务为一体的新型创业孵化平台，着力打造贵阳市交通大数据产业发展聚集区。目前，贵阳市交管局已建立起云计算中心、交通管控平台和交通大数据中心，前端集成了视频监控、信号控制、交通诱导、信息采集与统计分析、可视化警员定位和交通仿真等六大子系统共计 172 个功能模块。现阶段，贵阳市已有超 1 万路高清视频监控点位、205 个路口行为监测、100 个交通流量检测点、154 个拥堵检测点及 146 块信息发布屏，每天增加过车记录近 1600 万条，视频图像记录 495TB，车辆流量记录 2500 万条，视频

数据已突破 50PB，图片存储数据已突破 10PB，结构化流量数据已突破 150TB，其他非结构化数据增加近 600TB。此次，贵阳市交管局将开放近 2000GB 数据，贵阳市交通大数据孵化器还将全面整合公交、物流、客货运输等交通领域内的相关数据进行开放，为创业者研发交通、停车、物流类软件等产品提供数据支持。贵阳市交通大数据孵化器目前搭建了计算资源、数据管理和孵化交易三大平台。计算资源平台可在线调用交通数据、提供云计算服务的综合基础信息环境；数据管理平台，将交通相关数据融合在一起，是一个大交通数据资源库；孵化交易平台提供至少一年免费的优质数据和云计算资源，而孵化的优质产品在市场上获得收益后，通过收益分成的方式实现利益共享，最终形成互惠互利，相互成长的生态体系。

2015 年 11 月，百度公司与江苏省交通运输厅签署《战略合作框架协议》，以便捷公众出行为旨归，双方将整合共享数据资源，应用云计算平台、大数据分析等产品和技术，在交通运输信息化服务领域展开深度合作。这是国内省级交通运输主管部门首次向互联网企业开放交通出行的大数据。此次战略合作中，江苏省交通运输厅将逐步向百度提供包括实时公交、实时路况、出租、交通公共设施信息、道路管养等丰富、权威的交通出行大数据，百度也将利用大数据和云计算技术、地图产品以及开放平台优势，向社会公众提供江苏出行信息服务，为江苏交通十三五发展提供大数据分析服务。

（3）天气预报行业

2015 年 9 月，中国气象局《基本气象资料和产品共享目录》（以下简称《目录》）正式实施，《目录》所列的 5 类 17 种基本气象资料和产品也将正式提供共享服务，公众可以免费获取和使用《目录》所列的气象资料和产品。此次实施的《目录》涵盖地面、高空、气象卫星、天气雷达、数值模式天气预报等 5 类基本气象资料和产品，主要内容包括地面气象观测站的基本气象要素实时观测资料和气候标准值数据集、高空气象观测站的基本气象要素实时观测资料和气候标准值数据集，天气雷达站的实时图像产品，风云气象卫星的实时云图产品、定量产品和历史数据产品，以及中国气象局收集的国外地面、高空、卫星观测资料，全球数值模式天气预报（T639）产品，全国区域数值模式天气预报（GRAPES）产品。

与过去相比，此次《目录》中最明显的变化就是在原有的地面和高空两类资料的基础上，丰富了气象卫星、天气雷达和数值模式天气预报等 3 类资料和产品。地面气象观测资料的更新频率也从过去的每日更新提高到每小时更新，天气雷达图像产品实现了每 6 分钟更新。此次共享的所有资料和产品都将通过中国气象数据网和风云卫星遥感数据服务网向用户在线提供，中国气象局官方门户网站和中国天气网也会提供上述两个数据共享网站的访问链接。

中国天气网（www.weather.com.cn）是中国气象局面向社会和公众、以公益性为基础的气象服务门户网站，由中国气象局公共气象服务中心主办并进行具体开发、运行及维护，提供气象数据开放平台，在发布气象信息、服务防灾减灾等方面始终保持国内业界的领先地位。

2015 年 8 月，深圳市气象局发布《深圳市气象数据和产品共享开放目录》（以下简称《目录》），并同期推出“深圳气象数据网”，向社会开放气象数据。开放的气象数据和产品共 19 类 75 种，涵盖了气象观测、雷达与卫星、预警预报、气候资料，数据更新频次最高达到 6 分钟/次，数据总量达到了 946 万组，其中，开放的深圳国家基本气象站数据时段长达 35 年。所开放气象数据的种类、要素、时间尺度、起止时间、更新频率、数据类型以及数据共享方式等信息均在《目录》中有详细介绍。凡列入《目录》的气象数据和产品，任何公民、法人或其他组织注册后可直接在“深圳市气象局气象数据网”（http：//data.szmb.gov.cn）上查询和下载。

（4）金融行业

2015 年 6 月，通联数据旗下全新的数据开放平台（open.datayes.com）成功上线，针对应用开发者提供丰富的数据接口，免费使用。开发者可以随时调用多达上千 G 的经济金融数据指标，包括股票、期货、债券的交易信息、行情数据，宏观经济、行业经济数据以及电商数据这类实体经济数据，可以据此开发出各类理财、股票、经济信息查询播报等 PC 端或移动端的开发应用。通联数据开放平台是由通联资深数据团队精心打造，该团队此前推出的数据商城于 2015 年 1 月上线，是一个聚合了多种经济、金融数据的丰富数据库。其中不仅有通联数据自身的数据，还有来自汤森路透、巨灵、聚源、九次方大数据等众多国际国内知名数据商的数据，在国内金融数据领域，首先将开放共享的理念变为实践。

京东万象将帮助数据的提供方与需求方进行数据对接，解决企业之间的数据缺失问题，完善数据价值，提升企业效率。平台本身会对接多维度的丰富数据，保证数据的安全性与接入效率，是企业数据输出与流入的最佳渠道。与此同时，京东万象还将京东内部的企业总线服务云化并对外提供，帮助企业实现内部各应用系统之间的数据互联互通，解决企业内部数据孤岛以及多系统之间的数据整合问题。京东万象自 2015 年 3 月开展内测以来，已与超过 30 家权威数据提供方建立了合作关系，由数据提供方提供权威数据。任成元强调，“平台的愿景是做最权威最值得信赖的大数据开放平台，为全行业提供权威数据支持，打造全行业数据开放的优质数据生态圈”。京东万象目前主推的是金融行业的相关数据，现已覆盖了包括个人和企业征信报告、黑名单数据、失信数据等金融数据，此类数据给互联网金融创新企业

带来巨大的数据共享价值，同时也提高了金融相关数据提供方的变现能力和价值体现。

2015 年 4 月，蚂蚁金服宣布推出内部代号“维他命”的金融信息服务平台，是一个面向金融行业的统一开放平台和数据共创平台，向金融机构开放自身的大数据平台和实验环境。它意味着蚂蚁金服在数据、技术、渠道等方面向金融机构的全面开放，能够协助金融机构完成大数据金融时代的转型。

（5）科学研究领域

“十一五”以来，国家有关部门贯彻“整合、共享、完善、提高”的方针，组织开展了国家科技基础条件平台建设工作。初步建成了以研究实验基地和大型科学仪器设备、自然科技资源、科学数据、科技文献等六大领域为基本框架的国家科技基础条件平台建设体系；同时，各地方结合本地科技经济发展的具体需求和自身优势，因地制宜地建成了一批各具特色的地方科技平台。建设有包括气象、测绘、地震、水文水资源、农业、林业、医药卫生、海洋、国土资源 9 个领域的国家科学数据共享中心和地球系统科学、医药卫生、基础科学 3 大科学数据共享网。

下面仅以地球系统科学数据共享平台和气象科学数据共享平台为例进行介绍。

（1）地球系统科学数据共享平台

地球系统科学数据共享平台属于国家科技基础条件平台下的科学数据共享平台。该平台早在 2002 年就作为我国科学数据共享工程的首批 9 个试点之一启动建设，于 2004 年度纳入国家科技基础条件平台。它属于科学数据共享工程规划中的“基础科学与前沿研究”领域，主要是为地球系统科学的基础研究和学科前沿创新提供科学数据支撑和数据服务，是目前科学数据共享中唯一以整合、集成科研院所、高等院校和科学家个人通过科研活动所产生的分散科学数据为重点的平台。

地球系统科学数据共享平台承担单位是中国科学院地理科学与资源研究所。中科院资源、环境领域的研究所，国内地学领域的知名高校共 40 多家单位、世界数据中心（WDC）和国际山地中心（ICIMOD）、美国马里兰大学等国际组织和机构参与本平台建设与运行服务。

地球系统科学数据共享平台的总体目标是整合集成分布在国内外数据中心群、高等院校、科研院所和野外监测台站以及科学家个人手中历史的、现状的和未来的科学研究产生的数据资源，接收国家重大科研项目产生的数据成果及引进国际数据资源，加工、生产满足人地系统及地球系统各圈层相互关系研究的专题数据集。建立健全运行机制，形成一个非盈利的“以各运行服务中心”为构架的分布式地球系统科学前沿研究与全球变化研究数据支撑平台。2010 年前，重点整合满足资源、环境与人地关系等重大前沿及社会经济发展和国家重大战

略研究所需的数据资源，并为全面建设地球系统科学数据共享平台和长期稳定运行奠定基础。

近年来，平台以专题服务为牵引，突出资源的整合集成与深度挖掘。截止到 2014 年底，已经构建了“全球-全国-典型区域”三个层面的 11 个专题库，涵盖 5 大圈层，18 个学科，筛选翻译了 1500 多个国际数据资源网站，建立了 5 个国际数据资源镜像站点，数据总量达到 54.66TB，占应整合数据资源量的 66.8%。整合集成的数据资源全部经过规范化处理，同时，开展了数据资源的深度加工和数据产品的生产，形成了多要素、长时间系列的特色数据产品。通过数据产品，引领和驱动地球系统科学的发展。全部数据资源已经向社会公布并对外提供了 91.53TB 的数据服务量，数据资源利用率达到 167.4%。

（2）气象科学数据共享平台

气象科学数据共享平台是由一个国家级主节点、31 个省级分节点以及若干个专题节点组成的覆盖全国的分布式气象数据共享服务网络体系。该共享平台以满足国家和社会发展对气象科学数据的共享需求为目的，重点围绕标准规范体系建立、数据资源整合、共享平台建设和公益性数据共享服务等四个方面开展工作。气象科学数据共享中心研制的近 600 个数据集产品涵盖了大气科学领域主要数据种类，既包括地面、高空、海洋等常规气象要素，也包括了卫星、酸雨、雷达等非常规探测手段获取的气象科学数据；既有来源于气象部门观测获得的资料，也有来源于其他部门观测获得的气象资料；既有中国范围的气象资料，也有通过国际合作获得的全球数据产品。截止到 2011 年底，通过资源整合集成、历史资料数字化、数据分析研究和国外数据资源引进，气象科学数据共享平台共开发了 599 个基本覆盖大气科学领域的数据集产品，数据量达到 116TB，可在线共享服务的数据量超过了 50TB。

3.3 科学数据开放共享历程与经验

作为现代科学可持续发展的一种重要资源，科学数据的产生与科技创新密不可分。为促进科学数据资源的最大化使用，各方人员已经走过 50 年的历程。这个历程值得我们学习和研究，对我国科学数据的开放共享之路提供有益的经验。

3.3.1 开放共享历程

作为现代科学可持续发展的一种重要资源，科学数据的产生与科技创新密不可分。为促进科学数据资源的共享和交换，全世界许多国家和国际组织都开展了一系列的基于计算机网络的科学数据共享的研究和实践，目的是将长期积累的科学数据为本国以及全球的可持续发展提供科研数据和科研成果的支撑服务。例如世界数据中心（WDC：World Data Center）等国际组织的成立；很多国家都建立了国家级科学数据中心群和数据共享服务网络，如 NASA

主持的全球变化数据和信息系统等。

20 世纪 60 年代，美国进行了“数据图书馆”的尝试，以支持科研人员在科研活动中数据的存储和使用。稍后，英国于 1983 年在爱丁堡大学，如今在伦敦经济学院和牛津大学设有“数据图书馆服务”。如今图书馆界越来越多地参与到数据的管理和共享活动中。图书情报专业排名第一的伊利诺伊大学图书情报学院将科学数据管理作为图书情报硕士学位的方向之一。

进入 21 世纪后，科学数据的开放共享受到日益广泛的重视。2008 年，为了确保科学研究的成果和数据可以有效开放、共享和流通，美国总统行政办公室和科技政策办公室联名向农业部、商业部、能源部、教育部、国土部、内政部、交通部、科学基金委、美国国家卫生院等部门发出《科学研究成果发布原则》。为此，美国的 NSF 和 NIH 专门就科学数据共享问题发布了一些政策。2011 年，新西兰健康研究委员会（The Health Research Council of New Zealand，HRC）等 17 家健康机构共同签署共享科学数据联合声明，希望通过建立数据管理和共享的框架和标准，促进医疗卫生的更快发展、发挥更好的资金价值以及实现科学更高质量的发展。2012 年 3 月，美国政府宣布启动“大数据研究和发展计划”，同时组建“大数据高级指导小组”，涉及美国国家科学基金、国家卫生研究院、能源部、国防部等联邦政府部门，宣布将启动 2 亿美元的投资计划，提高从大量数据中访问、组织、收集发现信息以及信息共享的工具和技术水平。

我国自上世纪80年代起就开始在多个层面上推动科学数据的共享工作。1982年，中国科学院就已提出“科学数据库及其信息系统”建设项目，经过20多年的发展已经成为综合性的科学信息服务系统；在上世纪80年代初，腐蚀试验站网得到了国家自然科学基金的持续资助与支持；1989年，中国科学院联合有关部门和科研机构，组建了世界数据中心的中国分中心（WDC-D）；1999年，科技部在科技基础性工作专项中陆续启动了一批数据资源的建设项目，同时还就数据开放和共享中的若干技术问题委托WDC-D开展专门的研究工作。

进入21世纪后，科学数据的开放和共享得到日益重视，国家层面的数据开放和共享工程陆续实施。2001年，在科技部的主持下，调研报告《实施科学数据共享工程，增强国家科技创新能力》出炉，该报告对我国科学数据开放和共享中的主要问题以及可能的解决办法等一系列问题进行了详细地调查研究。随后，科技部和中国气象局联合召开新闻发布会，宣布气象数据共享试点正式启动；2002年，科技部向国务院提出了关于启动科技基础条件平台建设的建议，把建立科学数据开放和共享机制作为增强科技创新能力的重要环节；2003年，科学数据共享工程3个数据网（可持续发展科学数据共享网、地球系统科学数据共享网、医药

卫生科学数据共享网）和6个数据中心（气象科学数据中心、测绘科学数据中心、林业科学数据中心、地震科学数据中心、水文水资源科学数据中心、农业科学数据中心）等试点工作全面启动。在2009年，腐蚀站网被纳入科技部国家科技基础条件平台建设项目，发展成为国家材料环境腐蚀平台。该平台的建设，旨在解决“材料腐蚀问题给人类社会带来巨大的经济损失”。该平台上，有研究者多年数据积累与数据库建设、建模、模拟仿真、共享和工程应用系列化的工作。

近年来，随着大数据浪潮的来临，数据的开放、共享和流通得到进一步关注。在《大数据行动纲要》中，将“加快政府数据开放共享，推动资源整合，提升治理能力”作为第一条主要任务，确立了数据开放、共享和流通的国家战略地位。

3.3.2 挑战和存在的主要问题

若干年以来，各国政府和组织开展了大量科学数据开放和共享的探索，全球范围的科学数据共享工作取得了一系列的成果，科学数据共享理念逐渐普及，人们已经意识到科学数据开放和共享的重要性；整合集成了一批分散的数据资源，特别是抢救了一批珍贵的数据资源。科学数据的开放和共享，为人类社会的科研活动、高等教育和生产应用提供了很好的数据支撑等。在我国，科学数据的开放共享虽然取得了令人瞩目的成绩，但普遍存在一些突出的共性问题：

（1）共享理念尚不十分普及。虽然，全球范围内的科学数据开放共享已取得一些阶段性的成果，但受限于意识形态、保护主义等原因，前进的道路依然困难重重。而在我国，科学数据的开放共享主要是政府行为，大部分的数据共享活动是通过政府投资、项目驱动的形式进行，各部门经常“各自为战”，科学研究项目经常“各项目组为战”的情况。对于数据共享的重要性认识不充分，主动开放共享科学数据的研究单位和个人还比较少。

（2）共享机制尚不十分健全。虽然目前有些行业和部门已经出台有关数据共享的政策、规定和条例，鼓励和推动行业或部门数据的共享。然而，这些政策和条例都有这样或那样的限制，很多数据库只能限于部门和行业内部使用。同时，国家层面完善的共享机制并未形成。由于科学数据开放共享的机制尚不十分健全，尚未形成高效的数据开放、共享和流通局面。

（3）共享平台和技术规范发展慢：数据共享的技术标准与国外主流平台兼容性差，数据交换和整合存在障碍；平台功能与用户群体需求不匹配，造成一些亟需的科学数据资源依然不能依赖互联网方便获取。

（4）数据共享服务效果不明显：长期以来各单位数据资源本身并不规范，短时间内对大量科学数据进行规整合尚有困难，且用户能够直接使用的数据产品并不多。另外，由于缺

乏配套的数据使用文档、数据来源及处理说明，以及提供的数据开放和共享服务力度不够，用户将这些数据利用起来。这些都导致了数据开放和共享服务的效果并不十分理想。

3.3.3 开放共享模式

（1）政府驱动模式

在科学数据的开放和共享模式方面，美国是典型的国家政策驱动模式。政府引导下的科学数据开放共享有很多成功的尝试。1991年，美国总统事务办公厅就发布了“全球变化研究数据管理政策”，该政策的核心就是实行“彻底与开放”的科学数据开放共享。美国政府在科学数据共享方面根据投资来源的不同，严格区分不同数据的开放和共享机制：政府拥有、生产和政府资助生产的数据纳入到“彻底与开放”的共享机制下，即除涉及危害国家安全、影响政府政务和公务员个人隐私的数据外，其他都必须公开；私人所投资的公司生产的数据纳入到“平等竞争”市场化共享和流通机制下。在这两种截然不同的开放和共享机制中，美国联邦政府均起到主导的作用，所不同的是采取的管理方式有所不同。这两种机制相互补充，促进美国社会对科学数据的获取、共享和广泛应用。

政府驱动模式的核心是：由国家统筹规划数据共享机制与体系，提供数据共享工作预算和保障，以及相关政策法规的制定、完善和监察。“彻底与开放”的科学数据开放共享政策，使一度曾各自为政的混乱的数据管理走向了有序运作的轨道，科学家从得不到数据的抱怨走向数据的全面应用，科学数据的开发水平和开发能力逐步提高，并惠及了众多学科，也极大地促进了各国经济的健康发展。

（2）企业驱动模式

现代企业的发展离不开科学数据的支撑；反之，企业的发展也能促进了科学数据的传播与分享。一个企业的发展需要科学的发展方案，也需要科学数据来做产品的进一步研发以及科技成果的转化。大数据环境中数据积累量、数据分析能力、数据驱动业务而非流程驱动业务的能力将是保证企业长期高速发展的关键。事实上，数据的重要性使企业必将收集和分析海量的各种类型的数据，并快速获取影响未来的信息。在这一过程中，企业就会做出益于科学数据共享的决策与措施，例如由企业出资的科学合作项目的开发，有企业参与的科学资源共享平台的构建，以及企业自建商业性的科学数据库。只有学术和产业价值融合，才能真正发挥科学数据的应用价值。虽然学术界和产业界关注的价值点并不完全一致，但仍存在一些共性。发现和利用其中的共性，对解决科学数据共享中出现的问题很重要。跨界合作是积极且有意义的尝试，学术界可以致力于基础技术的研究，盈利模式的分析则由企业去完成。同时，学术界和产业界在某些交叉领域形成竞争也是一种良性模式。一些大企业会对前沿技

术和数据积累追踪最新的学术成果，甚至自己做学术研究，学术界也在积极推进产业化思考。

（3）部门之间的数据开放共享

科学数据的共享首先应该从生产科学数据的部门共享开始。为避免重复生产，科研单位内部之间以及各科研单位之间的科学数据，在不侵犯知识产权的情况下，要努力做到共享的第一步。以数据和信息为基础的经济、社会和科学发展中，一般情况下没有哪一个部门能够总是拥有某项科研活动需要的所有数据产品，尤其对于广大的科学社区，其研究内容广泛，对开放共享有着强烈的需求，研究过程中往往需要来自多个数据生产部门的不同区域、不同时期、不同标准、不同学科的数据资源。因此部门之间的数据交换就显得急需和迫切。例如，地震工作部门各单位收集并存档的各种地震科学数据，其他部门或单位为保障重大工程的地震安全而专门建设和管理的专用地震监测台网和强震动监测设施所收集并存档的地震科学数据，均属于共享范围。

（4）国际组织参与模式

随着人们对科学数据共享意识的提高，越来越多的国际组织参与进来，进行国际间的交流与合作，满足国际社会对科学数据共享的需求。在国际科学联合会（ICSU）的组织下，1957 年成立了世界数据中心（World Data Center），开展地球科学、空间科学和环境科学领域数据的收集、整理、系统化、标准化及交流服务等活动。世界数据中心不仅在地球科学、空间科学和环境科学领域积极推进了数据管理和共享，还积极参与许多重大的国际科学计划，为人类科学事业的发展作出了贡献。国际科技数据委员会（Committee on Data for Science and Technology）成立于1966 年，其宗旨是提高科学数据的质量，推动对科学数据的收集、交换、服务和共享。CODATA 致力于提高对整个科技领域有重要变化的数据的质量、可靠性、管理与可访问性，向科学家和工程师提供对国际数据活动的访问，促进直接合作，并利用互联网初步构建了全球范围内的科学数据交换体系。CODATA通过建立标准格式促进数据交换、共享，并协调各国数据项目，定期召开国际数据学术会议，扩大国际对科学数据共享的认识和深入探讨数据共享等方面的问题。

目前，很多由政府部门主导产生的科学数据在某种意义下可部分开放，且可共享或有条件可共享的科学数据越来越多。在遥感领域，美国共享了一些中分辨度的遥感数据，如美国国家航空航天局（NASA）的中分辨率成像MODIS（http://modis.gsfc.nasa.gov）、美国国家海洋大气局的第三代实用气象观测卫星NOAA（http://www.class.ngdc.noaa.gov/saa/products/catSearch）。在科学数据的共享方面，我国也提供部分无条件可共享的气象卫星数据（风云系列，http://satellite.cma.gov.cn/portalsite/default.aspx）和有条件共享的中巴地球资

源卫星系列（http://www.cresda.com/n16/n1130/index.html）。在这方面，欧洲空间局数据的共享是有条件的。

3.3.4 开放共享服务

科学数据由于具有较强的专业性，所以提供服务是科学数据开放共享取得较好效果的基础。通常，无条件免费获取的数据没有专业的机构对其进行技术支持，使用该数据的人只能通过简单的数据说明并根据自己的需求自行开发使用，所以效果难以得到保障。然而，欧空局的数据有非常系统的培训体系，除了详细介绍哪些应用适合使用哪个类型的数据（数据附有应用示范），还有针对性地开发出一系列分析该数据的软件，并定期提供免费的培训课程，同时，对应的材料和软件都是可无条件共享的。例如，遥感中的合成孔径雷达图像分析往往需要比较高的门槛，但通过其提供的技术支持服务，使得对该数据的使用变得轻松，不仅增加了该数据的用户群，同时也促进了该学科和相关行业的发展。因而，科学数据的共享需要专业的机构对其使用提供一定的技术支持，使其能更好被科研和应用部门使用应用。

科学数据共享作为国家的科学基础设施，其发展不是一项短期行为。其目标应该是实现科学数据资源的开放与共用，需要科学界广大工作者长时间共同努力才能实现。

第4章　数据开放共享的风险与对策

在大数据时代，数据被公认为堪比石油的重要资源，数据分析已经成为提升各类应用系统水平的重要驱动力，数据的有效共享与流动已经成为充分发挥数据价值的关键。但另一方面，数据开放给我们重大机遇的同时，也带来了一系列的风险和挑战。数据完备性、部门利益、隐私保护、国家安全、法律法规、国家政策、行业自律等问题，都是数据开放所面临的风险及困难。

4.1 数据完备性及质量风险与对策

在数据开放的环境下，数据具有海量异构、完备性低、不一致等特点，数据质量成为制约数据利用的首要挑战。近年来，数据的生产速度呈现爆炸式增长，政府部门、各行业以及互联网等领域已积累数以TB、PB乃至EB计的大数据。这些数据可以造福人类社会，是现代社会的宝贵财富。现今，人们已充分认识到“大数据可以为人们更深入地感知、认识和控制物理世界提供前所未有的丰富信息”。

“大数据分析能为社会带来巨大价值”已成为大数据领域的一项共识，在高质量数据上进行分析是有效发挥大数据作用的一个前提。《大数据资产：聪明的企业怎样致胜于数据治理》一书的作者Tony Fisher曾提到，如果基本数据不可靠，大多数企业的大数据计划要么会失败，要么效果会低于预期。为了更好地利用这些数据造福人类，使之成为现代社会的重要财富，“大数据的有效开放、共享和流通，并加以高效整合、使用”，如何解决数据的完备性及质量问题是成为人们在进行大数据分析过程中需要解决的一个基础问题和紧迫的难点问题。

数据的完备性及质量问题由来已久。以往，由于数据规模有限，面向业务系统的统计分析通常进行数据的清洗。在大数据时代，这种比较“昂贵”的数据清洗方式举步维艰。而低质量数据的开放、共享和流通也会对大数据分析结果的质量产生重大影响，进而影响政府决策和行业发展。例如，我国的地方GDP之和连年超越全国总量，除核算制度等原因，还有数据不一致性导致的重复计算等问题。国家统计局从近年来全国统计执法检查的情况表明，虚报、瞒报、伪造、篡改统计资料的违法行为约占全部统计违法行为的60%；有研究团队对我国某国有大型企业信息中心的数据抽样检验，发现10%的信息存在各种类型的错误；在美国银行业，由于数据不一致性问题而失察的信用卡欺诈在2006年就造成48亿美元的损失；据国外权威机构的分析表明，美国企业信息中1%--30%的数据存在各种错误和误差。

数据不完备主要表现在元数据（即描述数据的数据）不齐，造成数据语意不清，比如丈

量尺度的计量单位不明，数据使用者面临使用错误的风险。另外一个不完备的情况是数据采集不完全、采样不够随机等情况，导致分析结果可用性降低，甚至误导的情况。

究其根源，数据完备性缺乏及质量低下的主要原因有以下几方面：

（1）基础数据的完备性及质量有待提高。造成上述质量低下的关键原因在于，数据生命周期之中流入了不一致、不准确、不可靠的数据，大数据时代这个问题更加突出。与传统企业内部数据分析不同，大数据的采集往往与其使用脱钩，比如数据采集的使用目标不清晰、数据采集人员的专业训练不明确、数据采集的责权利难把握等。在数据管理和分析领域流行的一个说法是“更好的数据意味着更好的决策”，在大数据时代甚至更为真切，但往往“更好的数据”的来源并不能有所保证。

（2）缺乏更好的低质数据检测和修复方法。数据质量的检测和修复方法往往建立在传统的关系数据库系统以及一些质量相对较高的系统内部。事实上，适用于大数据的一致性、完整性、精确性、时效性错误的检测和修复方法还比较少，且比较零散，需要进一步有效整合；另外，平台性的工作还有待加强。

（3）缺少更好的方法以完成对低质数据的计算和分析。传统的计算和分析流程中，对低质数据的清理通常需要占用30%-80%的开发时间和预算，然后在清理后的高质量的数据上进行计算和分析。由于数据量的爆炸式增长，且计算和分析任务的紧迫性，这种集中式的数据清理工作已力不从心。

根据以上分析，数据质量及完备性可以通过以下方法解决：

（1）高质量大数据的获取

大数据的来源具有多样化的特点。一般来说，大数据的来源可以分为4类：政府数据、行业数据、科学实验和观测数据，以及Web数据。政府数据来自政府各职能部门在日常事务办理时所积累下来的数据；行业数据是通过行业内部的企业收集而来；科学实验数据、观测数据以及行业内的一些监测数据往往通过传感器或观测设备来获取；Web数据来自互联网。虽然Web上存有的丰富数据源，但究其质量，以低质数据为主。在大数据分析中，人们往往经常需要从多个Web数据源获取数据，并将其整合为自己需要的数据集合（这个过程通常被称为Web数据集成）。在Web数据集成中，数据源的质量会极大地影响集成数据的可用性。因此，需要判定和选择高质量数据源，使其成为数据获取的源泉，是获得高质量集成数据的途径之一。针对观测数据，需要利用领域知识，建立适用于该领域的计算模型，修正数据的精度或过滤掉低质数据。

（2）数据错误的自动检测和修复

目前，大数据的数据错误检测和修复手段主要包括以下几类：1）利用语义规则（主要是来自现实世界的依赖规则）进行数据错误的检测和修复；2）引入数据完整性的评价；3）利用数据的时效性对数据进行检测，以更正过时的信息；4）数据实体的同一性判定。在检测方面，主要通过技术手段自动检测数据的一致性错误和实体同一性错误；在修复方面，主要通过事先定义的语义规则和基于统计的方法对一致性错误和同一性错误进行修复。考虑到一次性修复的不可逆转特性，建议采用一系列修复动作替代一次性修复的思想，避免修复引入新的错误。目前，针对数据错误检测和修复的工具等方面尚有很大空间。

（3）开发低质数据上的计算和分析方法

低质数据是指包含一定错误，质量不高的数据。低质数据上的计算和分析任务，可以考虑利用约束条件形成新的计算任务，改写已有的计算任务，使改写后的任务可以适应数据的不一致性错误，在不一致数据上求解计算结果。而针对分析任务，需要采用部分不确定数据上的分析方法对数据加以分析。目前，直接在低质数据上直接进行计算和分析的思想已有提出，该方面的理论及技术还处于不断发展的阶段，仅针对一些特定的问题提出了相应的技术，针对大数据及更多可用性问题的深入研究还有待开展。

4.2 政府部门壁垒风险与对策

复旦大学国际关系与公共事务学院数字与移动治理实验室采用开放数据的通用评估框架（Common Assessment Framework）研究了北京、上海、贵州等 8 个省市的政府数据开放情况，结论很不乐观，存在六个方面主要问题：数据量少、价值低、可机读比例低；开放的多为静态数据；数据授权协议条款含糊；缺乏便捷的数据获取渠道；缺乏高质量的数据应用；缺乏便捷、及时、有效、公开的互动交流等。

造成这一状况的政府数据公开的主要壁垒有三个方面：一是不敢开放，由于缺乏相应的法律法规与技术标准，政府部门不知道哪些数据能够开放、哪些数据不能开放，特别是担心数据可能涉密更加不敢开放；二是不肯开放，对数据资源的控制是政府部门事权体现，开放数据会影响部门的权力，政府数据资源的采集与处理往往委托第三方承担，数据的共享开放直接会影响干系人的利益；三是不情愿地开放，挡不住开放的趋势而必需开放，则通过私有数据格式、公开处理过的数据、公开过时的或部分的数据、收费开放、需采用专门的软件（甚至收费的商业软件）来读取等来设置重重障碍。

行政分割也导致数据无法共享。我国政府数据资源多按地域或部门进行分割管理。不同地域和部门为了自身利益，形成人为数据共享壁垒，加大了政府大数据开发难度。由于政府

部门业务管理信息系统开发和建设的“部门化”，政府信息系统出现“系统林立”和分裂状态，政府公共信息资源重复采集现象严重，信息摩擦和治理成本偏高。

开放政府工作组（Open Government Working Group）提出的政府数据开放 8 项原则，值得我们参考与借鉴：

（1）完整性：所有的政府数据都应该开放，除非涉及隐私、安全和特别限制；

（2）一手性：数据从源头采集到，保持尽可能细的粒度，未经整合过或者修改过；

（3）及时性：以尽可能快的速度发布数据以保证数据的价值；

（4）可获取性：数据是可获得的，可提供给最广泛的用户、可提供最广泛的用途；

（5）可机读性：数据拥有合理的结构，可由机器自动处理；

（6）非歧视性：数据对所有人可用，不需要专门注册登记；

（7）非私有性：数据的格式是开放的，没有任何实体有独占控制权；

（8）无需授权使用：数据不受版权、专利、商标或贸易保密规则限制，除非涉及隐私、安全和特别限制；

开放政府工作组还提出了额外的 7 条原则：

（1）在线和免费：如果数据不在互联网上、不是免费的就不是真正意义上的公开，此外，数据还应该容易被找到；

（2）永久性：数据应该在固定的互联网的网站上，并且在尽可能长的时间内保持稳定的数据格式；

（3）可信的：应该采用数字签名保证数据发布的时间、数据的完整性和真实性；

（4）默认开放：数据的默认开放要通过法律法规来保证，通过数据目录等工具来实施；

（5）文件化：将数据格式和数据的含义文件化将促进数据的使用；

（6）安全性：开放的数据中不应该包括可执行代码，防止恶意代码（如计算机病毒、木马）传播；

（7）采纳公众意见：公众最明白哪种信息技术最适合他们自己的应用，采纳公众意见对开放数据的设计非常重要。

国务院《大数据行动纲要》中提出，要明确各部门数据共享的范围边界和使用方式、制订数据共享开放的资源清单和目录、厘清各部门数据管理及共享的义务和权利、建立政府数据统一共享交换平台等措施都是符合上述原则的具体举措。政府数据公开的基础是政府自身的公开透明，我们相信随着我国进一步深化政治体制改革，政府管理理念向开放的政府转变，政府的数据开放程度将与政府治理能力现代化同步提升。

随着数据治理理念的影响渗透，我国公共数据开放共享进程开始逐步加快。2013 年，国务院发布了《关于促进信息消费扩大内需的若干意见》，要求促进公共信息资源共享和开发利用，推动市政公用企事业单位、公共服务事业单位等机构开放信息资源。随着《大数据行动纲要》的出台，对促进大数据开放共享做出具体规定，将会适时解决数据壁垒问题，推进大数据驱动下的社会发展。

首先，加强顶层设计和统筹规划，明确各部门数据共享的范围边界和使用方式，厘清各部门数据管理及共享的义务和权利，依托政府数据统一共享交换平台，大力推进国家人口基础信息库、法人单位信息资源库、自然资源和空间地理基础信息库等国家基础数据资源，以及金税、金关、金财、金审、金盾、金宏、金保、金土、金农、金水、金质等信息系统跨部门、跨区域共享。加快各地区、各部门、各有关企事业单位及社会组织信用信息系统的互联互通和信息共享，丰富面向公众的信用信息服务，提高政府服务和监管水平。结合信息惠民工程实施和智慧城市建设，推动中央部门与地方政府条块结合、联合试点，实现公共服务的多方数据共享、制度对接和协同配合。

其次，稳步推动公共数据资源开放。在依法加强安全保障和隐私保护的前提下，稳步推动公共数据资源开放。推动建立政府部门和事业单位等公共机构数据资源清单，按照“增量先行”的方式，加强对政府部门数据的国家统筹管理，加快建设国家政府数据统一开放平台。制定公共机构数据开放计划，落实数据开放和维护责任，推进公共机构数据资源统一汇聚和集中向社会开放，提升政府数据开放共享标准化程度，优先推动信用、交通、医疗、卫生、就业、社保、地理、文化、教育、科技、资源、农业、环境、安监、金融、质量、统计、气象、海洋、企业登记监管等民生保障服务相关领域的政府数据向社会开放。建立政府和社会互动的大数据采集形成机制，制定政府数据共享开放目录。通过政务数据公开共享，引导企业、行业协会、科研机构、社会组织等主动采集并开放数据。

再次，统筹规划大数据基础设施建设。结合国家政务信息化工程建设规划，统筹政务数据资源和社会数据资源，布局国家大数据平台、数据中心等基础设施。加快完善国家人口基础信息库、法人单位信息资源库、自然资源和空间地理基础信息库等基础信息资源和健康、就业、社保、能源、信用、统计、质量、国土、农业、城乡建设、企业登记监管等重要领域信息资源，加强与社会大数据的汇聚整合和关联分析。推动国民经济动员大数据应用，加强军民信息资源共享。充分利用现有企业、政府等数据资源和平台设施，注重对现有数据中心及服务器资源的改造和利用，建设绿色环保、低成本、高效率、基于云计算的大数据基础设施和区域性、行业性数据汇聚平台，避免盲目建设和重复投资。加强对互联网重要数据资源

的备份及保护。

4.3 隐私保护需求的风险及对策

数据中蕴含了大量的用户隐私信息，保护这些敏感信息、用户行为习惯等不被非法获取、利用是信息系统的基本要求。对此，人们采取了一些手段来保护隐私信息，例如通过加密技术实现对身份等敏感信息的保护；通过信息混淆、匿名化等方法来管理位置等隐私，为数据拥有着提供了一定程度隐私保证，同时使得数据能够以相对低效的方式参与查询和计算；通过 DAC、MAC 等访问控制策略保证用户只能访问授权数据。数据保护与数据开放流通在一定的意义上是相悖的，怎样在防范隐私泄露风险的情况下做到最大化数据开放与流通，就成为一个紧迫的命题。

在大量数据开放与共享的环境下，不法分子有机会获得丰富的数据资源，企业和用户的隐私保护问题因此面临极大风险，特别是所开放或共享的数据未经过妥善的加密、匿名化等处理。通过移动对象轨迹数据，可以推断出用户的工作单位、家庭地址，掌握用户的位置信息甚至出行规律；通过社交网络数据，掌握用户的交友和社会关系，进而根据文本信息推断真实姓名和喜好；医疗数据的开放可能暴露用户的病患等隐私；此外，不法分子还可以利用一些查询规则来规避信息混淆和匿名化策略，收集得到的各种业务数据得到客户的手机、邮箱等敏感信息，甚至分析出企业的财务、规划等敏感信息。这显然，将对用户的隐私保护造成极大的挑战。

需要特别指出，由于不法分子可以通过关联分析等手段窃取敏感信息，大量数据的开放与共享还面临一些非常规的隐私保护问题。例如，用户在社交网站、电商、地图服务平台分别留有自己的交友、购物、出行等相关的数据。但是，一旦攻击者同时获取这些公开数据，就可以通过简单的关联分析将网络上的虚拟信息直接关联到同一用户，得到用户的住址、工作地址、收入水平、健康状况、社交关系等敏感信息的完整链条。不法分子甚至可以进一步准确地刻画用户的身份、性格、偏好，这些隐私信息的泄露对用户安全的威胁不言而喻。

在大数据开放的前提下解决的隐私问题可以从以下几方面着手。从数据开放方式和管理的角度来说，需要加强对数据开放的审核工作，针对必要的字段进行加密处理，审核后的用户才能提供隐私弱保护（例如匿名化策略）处理的数据，并采取策略防止关联分析。从数据隐私保护的技术层面来说，研究适用于大数据开放与共享环境下的隐私保护机制，例如改进同态加密技术，以及基于关联分析的隐私安全评价体系，使得一方面数据隐私能够被有效保护，同时多源异构数据能够在一定性能保证的前提下参与应用处理与计算。

4.4 国家安全风险及对策

随着云计算、大数据、物联网、移动互联网技术的发展，移动智能设备的普及，数据呈现爆炸式增长。数据正在改变各国综合国力、重塑未来国际战略格局，也为产业升级与改革创新奠定了基础。数据主权成为各种利益集团争夺的战场，握有政治权力的国家、占据资本优势的商业机构、自恃技术高超的个体，围绕数据展开了激烈角逐。大数据成为“未来的新石油”，继陆、海、空之后的又一项“关键性核心资产”。随着大数据的开放与共享，网络安全问题凸显，国家利益和安全面临严峻挑战。“大数据安全”已上升为国家安全。

目前，来自网络安全的威胁可分为四大类型：黑客入侵、组织犯罪、网络恐怖主义以及国家参与的网络攻击。为表达不满而未造成破坏性的黑客入侵行为最为常见，一般不构成对国家安全的严重威胁，但所造成的个体、组织侵害不可忽视；那些影响到国家、社会利益，为达一定目的、有组织、蓄意实施的破坏性黑客行动对国家安全威胁程度最高，或直接表现为军事威胁和打击，或对金融、交通等要害造成冲击、破坏，或实施网络恐怖主义，挑起网络战争，成为各国网络安全防范的重要目标。

2014 年 3 月，美国非营利性组织开放安全基金会和威胁情报咨询公司 RBS 合作发布的报告称，2013 年堪称数据丢失大年，全球共发生 2164 起数据泄露事件，超过 8.22 亿条记录被曝光，几乎是 2011 年的两倍。全球范围内，受害方在每起数据泄露事件中遭受的平均损失为 145 美元，比 2012 年增加 9%。此外，据斯诺登爆料，美国信息技术领域龙头企业如谷歌、雅虎、微软、苹果等巨头，无一不参与了间谍活动，他们向美国国家安全局开放服务器、为美国中央情报局预留软件后门等等。与政府的合作渗透于大数据管理的各个层面。

据国家信息中心等部门相关报告显示，2013 年中国 7.6 万多个网站被境外通过植入后门实施控制，其中政府网站 2452 个。中国境内 1.5 万台主机被 APT 木马控制，关键基础设施和重要信息系统安全遭受严重威胁。2014 年初，爆出美国大规模入侵华为服务器的消息，针对无线路由器等上网设备所造成的重大安全漏洞，可导致用户被“终身监视”的严重后果。

大数据开放共享带来的数据安全威胁随时都有可能发生。各种国家信息基础设施和重要机构所承载着的庞大数据信息，如由信息网络系统所控制的石油和天然气管道、水、电力、交通、银行、金融、商业和军事等，都有可能成为被攻击的目标，这一切加剧了人们对大数据安全的忧惧，需要在数据开放过程中防止数据泄露而造成的国家安全问题。

伴随着大数据的采集及应用的不断增长，国家安全面临异常严峻的风险与挑战，主要体现在下列几个方面：

（1）我国的大数据战略顶层设计有待细化

从主要发达国家大数据发展经验看，美国等国持续强化国家战略的顶层设计，重点关注大数据对创新能力、国家安全能力、产业竞争力等国家竞争优势的重构。在顶层设计的基础上，持续推出大数据国家战略规划，各部门协同形成合力以保护国家安全。在大数据时代，国家安全防御是系统性、全局性的战略问题，需要有全面推动大数据战略实施的权力部门和核心决策机构对国家安全防御问题进行整体布局和规划，包括对政府和重要行业数据开放的评估与决策。目前，我国虽然已推出《大数据行动纲领》，但政府各部门对大数据产业的整体布局和规划尚不细致。

（2）传统国家安全防御思维和体制在开放大数据时代出现明显不适，并引发全新难题

大数据正在重构政府、市场、社会三者之间关系模式，然而，现有国家安全防御思维和体制已经明显不适应这种大数据时代数据开放新趋势的变化。如果国家安全防御思维和体制不能有效跟进，很可能将大数据技术带来的国家治理契机转化为威胁国家安全的一大挑战，可能引发新的问题。

（3）法治建设滞后，维护“国家安全”的法律法规规标准及配套政策严重缺失

目前，我国大数据法治建设明显滞后，用于规范、界定“数据主权”的相关法律缺失，缺乏有效的大数据思维和法律框架。一是对于政府、商业组织和社会机构的数据开放、信息公开的相关法律法规尚待进一步完善。缺乏企业和应用程序关于搜集、存储、分析、应用数据的相关法规。二是没有对保护本国数据、限制数据跨境流通等做出明确规定。金融、证券、保险等重要行业在华开展业务的外国企业将大量敏感数据传输、存储至其国外的数据中心，存在不可控风险。

（4）全球大数据战略博弈升级，我国面临较大数据安全与数据防御风险

借助大数据技术，发达国家对全球数据的监控能力逐步提升，我国国家安全受到严重威胁。例如 IBM 服务器、英特尔电脑设备、思科的通讯设备产品、微软的操作系统等对我国数据安全乃至国家安全提出了严峻的考验。

中国共产党十八届三中全会公报指出：将设立国家安全委员会，完善国家安全体制和国家安全战略，确保国家安全。打破数据孤岛，开放和共享各方数据，是未来信息社会发展的必然要求。但是所有拥有海量数据的政府机构和企业，应在不危及国家安全和个人隐私等原则的基础上，以开放的姿态、积极的行动促进大数据的深度应用，并通过立法保障各方在大数据应用中的共享共赢。针对上述挑战，一些可行的对策主要包括：

（1）全面实施大数据国家战略

我国需要加快实施大数据国家战略，把数据主权纳入国家核心利益的范畴，加快大数据立法，推动安全、合理的数据开放与流动。规划重点领域的大数据研究计划，布局关键技术研发方向，做好体制机制、资金、法规标准等方面的保障，全方位保证国家安全。

（2）全面完善数据安全等级制度

国家需要对提供公共通信、广播电视传输等服务的基础信息网络，能源、交通、水利、金融等重要行业和供电、供水、医疗卫生等公共服务领域的重要信息系统，特别是针对重要的国家网络基础设施系统，如党政系统、金融系统（银行、保险、证券）、财税系统（财政、税务、工商）、经贸系统（贸易、海关）、交通运输系统（航空、航天、铁路、公路、水运、海运）、能源系统（电力、电气、燃气、煤炭、油料）、社会应急系统（医疗、消防、救援）、科研教育系统、国防系统等实行严格的国家安全等级保护制度。

（3）加大国产安全可靠软硬件的推广力度

随着我国信息化程度的提高，软、硬件国产化已经成为国家安全的重要组成部分。此前的“棱镜门”事件为国家信息安全保护敲响了警钟，在关系国计民生的国家级重点项目知识产权、公安、安全等领域采用国产软、硬件，对于国家安全有着重要的意义。

（4）加强国际交流合作，积极参与国际规则制定，维护国家网络空间安全

在各国争占大数据安全制高点的博弈中，中国作为世界网络大国之一，要进一步开展国际交流、合作，倡导合理利用数据，叫停数据强权争夺。要积极参与和促进平等、公平、公正原则下国际数据规范的制定和出台，捍卫国家主权、守护国家利益、争取合法权益，维护世界网络空间的良好秩序。

4.5 使用法律法规规避风险

当今世界，各国竞争愈演愈烈，数据资源成为重要的竞争要素。近年来，数据开放的经济、社会价值备受关注。数据开放上升为国家竞争力的重要组成部分。一些国家共同签署了数据开放的协议，旨在团结协作共同推进数据开放进程。各国也纷纷发布一系列法律法规，针对各国国情现状对数据开放进行了详细地阐述和规定，迎接数据开放带来的效益和挑战。

面对数据开放的潮流，各国逐步建立符合国家现状的法律法规，将数据开放上升到法律层面，强制执行，同时兼顾国家安全和公民隐私。比如美国 2009 年 1 月，奥巴马签署《透明与公开政府备忘录》，提出透明、共享与协作的政府公开工作原则。12 月，总统执行办公室签署《公开政府命令》，下达详细的开放命令，要求在互联网上开放政府数据，制度化开放政府文化，并且提出扫除政府数据开放的政策屏障。2010 年 11 月，美国颁布《13556 号

总统令》，强调政府实践的开放性和统一性，缩小政府数据保密范围，开放非涉密的信息，不对大众进行过度隐瞒，使政府工作更加公开透明，公民更好地了解政府工作，利用相关数据资源。2011 年 1 月，美国颁布《13563 号总统令》，提出建立数据开放的交流环境，允许州、地区和部落官员的开放信息交流，允许公民参与制度草案的修改，允许公民提建议，开放相关科学和技术发现，使之容易搜索和下载。2013 年 5 月，美国发布《实现政府信息公开化和机器可读取化总统行政命令》，提出以开放化、机器可读化作为政府数据的基本形态，保证公民可以随时随地，以任何设备查询、获取信息。2000 年 11 月，英国通过《信息公开法》，提出公民享有数据权，设立信息专员与专门委员会，同时制定信息公开的具体范围和例外。2001 年 12 月，英国大法官宣布该法从 2002 年起分步实施，2005 年 1 月 1 日起全面生效。德国 2000 年 9 月发布《2005 年联邦政府在线计划》，提出联邦政府到 2005 年在线提供所有可在网上提供的服务给公众，方便公众获取数据资源和服务，同时还起草了一个标准框架并使其对管理系统的所有层面可用。英国于 2000 年 11 月通过《信息公开法》，提出公民享有数据权，设立信息专员与专门委员会，同时制定信息公开的具体范围和例外。

我国的数据开放相关的法律法规体系建设起步较晚。现有的法律法规仍然存在阻碍数据开放的要素。比如 2007 年出台的《政府信息公开条例》，规定了行政机关、各级人民政府的信息公开范围。但是公民、法人或者其他组织申请获取未公开信息只能采用书面形式（包括数据电文形式），对于行政机关不能当场答复的，自收到申请之日起 15 个工作日内予以答复。途径过于单一，流程复杂，局限性大，成为公民自由获取数据资源的一道屏障。并且涉及民生内容的开放数据类别不够全面，条目不够细致。这些都妨碍了数据开放。

2011 年 8 月，中共中央办公厅及国务院办公厅发布《关于深化政务公开加强政务服务的意见》，提出深化政务公开的总体要求，各级政府政务公开的重点内容。提到逐步建立统一规范的公共资源交易平台，有条件的地方可探索公共资源交易平台与服务中心合并的一体化管理模式。公共资源平台的建立缺乏具体的计划，没有硬性的规定，而且仅仅是有条件的地方可探索一体化管理模式。公共资源平台的缺乏一定程度上阻碍了数据开放。

2015 年 8 月，国务院发布《大数据行动纲要》，将 2018 年底前建成国家政府数据统一开放平台作为目标，将加快政府数据开放共享，推动资源整合，提升治理能力作为主要任务之一。文件提出优先推动民生保障服务相关领域的政府数据向社会开放（2020 年底前逐步实现），制定政府数据共享开放目录。文件在以往法律法规的基础上，实现了梳理和完善，更加具体地阐述了数据开放共享的目标、任务、实现期限。

针对上述问题，借鉴国际做法针对中国数据开放法律法规的建设，提出如下几点建议：

（1）发布法律法规，指导建设数据统一开放平台，建立、推广地区数据开放门户网站。

（2）建立数据开放优先级机制，详细列出各优先级包含的数据资源。涉及民生的数据赋予最高优先级，以人为本；科研数据放在较高的优先级，促进国家科技竞争力。

（3）扩大数据开放范围，全面细致地列出各个领域的开放数据类别及具体条目，明确工作细节，增强法律约束效力。

（4）注重数据开放的质量，开放优质的数据资源，鼓励大众利用数据资源，充分发挥数据开放的价值。

（5）增加申请获取未公开信息的途径，简化申请程序和时间周期。

（6）在法规中注重平衡数据开放和安全隐私问题，数据开放的同时注重维护。

4.6 信息保密需求的风险及对策

发展安全是国家安全的核心，与发展要求相适应的各类大数据作用日益凸显的同时，大数据保护也成为了国家安全保密部门的重要职责。经济、教育、医疗、能源、交通、国土安全等数据都会对国家安全和社会稳定产生重大影响，因此，信息保密条例和信息系统安保等级政策对大数据的开放共享同样适用。自 20 世纪 90 年代以来，我国发布了多项有关信息系统安全保密的政策，在维护我国信息安全和隐私的同时，对大数据产业的开放共享也造成了一定的制约。

1994 年，国务院颁布《中华人民共和国计算机信息系统安全保护条例》，首次明确了计算机信息系统实行信息系统安全等级保护。

1997 年，《中共中央关于加强新形势下保密工作的决定》发布，文件指出了新形势下保密工作的指导思想和基本任务，提出要建立与《保密法》相配套的保密法规体系和执法体系，建立现代化的保密技术防范体系。

1999 年，我国发布 GB17859《计算机信息系统安全保护等级划分准则》，并于 2001 年 1 月 1 日开始实施，这标志着我国计算机信息系统安全保护的等级制度正式在全国范围内施行。

2003 年，中办、国办发布《国家信息化领导小组关于加强信息安全保障工作的意见》（中办发〔2003〕27 号），提出实行信息安全等级保护，建立国家信息安全保障体系的要求。意见明确指出：“要重点保护基础信息网络和关系国家安全、经济命脉、社会稳定等方面的重要信息系统，抓紧建立信息系统安全等级保护制度，制定信息系统安全等级保护的管理办法和技术指南”。

2004 年 9 月 17 日，公安部、国家保密局、国家密码管理委员会办公室和国务院信息办再次下发《关于信息安全等级保护工作的实施意见》，意见明确了信息系统安全等级保护制度是国家在国民经济和社会信息化的发展过程中，提高信息安全保障能力和水平，维护国家安全、社会稳定和公共利益，保障和促进信息化建设健康发展的一项基本制度，并对信息等级保护的基本内容、工作职责分工、要求和实施计划进行了详细阐述。

中央保密委员会于 2004 年 12 月 23 日下发了《关于加强信息安全保障工作中保密管理若干意见》，意见明确提出要建立健全涉密信息系统分级保护制度。2005 年 12 月 28 日，国家保密局下发了《涉及国家秘密的信息系统分级保护管理办法》，进一步明确了信息系统分级保护工作中的流程和步骤。

2006 年 1 月 17 日，公安部等四部门下发《信息安全等级保护管理办法（试行）》，进一步明确了涉及国家秘密的信息系统应当依据国家信息安全等级保护的基本要求，按照国家保密工作部门涉密信息系统分级保护的管理规定和技术标准，结合系统实际情况进行保护。同时，《保密法》修订草案也增加了网络安全保密管理的条款。随着《保密法》的贯彻实施，我国已经基本形成了完善的保密法规体系。

目前，正在执行的两个分级保护的国家保密标准是 BMB17《涉及国家秘密的信息系统分级保护技术要求》和 BMB20《涉及国家秘密的信息系统分级保护管理规范》，分别从物理安全、信息安全、运行安全和安全保密管理等方面，对不同级别的涉密信息系统制定了明确的分级保护措施，并从技术要求和管理标准两个层面规范了涉密信息系统的分级保护问题。

上述一系列政策的发布，加之信息系统在运行及维护过程的不可控性以及随意性，直接导致了大数据的开放和共享举步维艰。无论从信息安全角度，还是从隐私角度，数据在不同载体间的流动都可能会导致一系列的法律和社会问题，从而阻碍了大数据产业规模化、集群化的发展。

当前，我们国家民主与法制建设进程的不断推进，保密的范围和事项正在逐步减少，致使一些涉密人员保密意识和敌情观念淡化，对保密工作的必要性和重要性认识不足。政府部门掌握着大量重要甚至核心的机密，已成为各种窃密活动的重点目标。以数据开放为名义来窃取我国党政机关和军工单位机密的事件有可能成为未来我国窃密与反窃密，渗透与反渗透的主战场。因此，严格按照涉密信息系统分级保护的要求，在大数据产业的发展过程中加强涉密信息系统的建设意义重大。

结合国际经验与我国面临的实际问题，我国政府为推动大数据的发展，应当加快建设国家级的大数据标准化体系，将关于信息安全、信息等级保护等政策协同规划，从而在政府层

面为实现大数据平台之间的互联互通和开放共享保驾护航，标准化应当涉及标准主体、切入点、运营、建设等方面。通过加强安全等级保护、数据安全和用户隐私保护等方面的制度制定，稳步推动大数据应用、关键技术研发扩散、产业培育、数据开放、数据保护、市场监管、法律法规等关键布局的统筹规划。

在政府关键数据开放上，建议完善配套制度，分类分批推动政府数据开放，推进政府和公用事业领域数据资源的普查工作，界定数据权属，理顺利益机制。同时，按照相关法规制定政府和公共数据开放中的安全和隐私保护检查表，对可能涉及国家安全和公民隐私的风险点进行严格控制。在此基础上，按敏感性对政府和公共数据进行分类，确定开放优先级，制定分步骤的数据开放路线图。

需要健全大数据安全保障体系。加强大数据环境下的网络安全问题研究和基于大数据的网络安全技术研究，落实信息安全等级保护、风险评估等网络安全制度，建立健全大数据安全保障体系。建立大数据安全评估体系。切实加强关键信息基础设施安全防护，做好大数据平台及服务商的可靠性及安全性评测、应用安全评测、监测预警和风险评估。明确数据采集、传输、存储、使用、开放等各环节保障网络安全的范围边界、责任主体和具体要求，切实加强对涉及国家利益、公共安全、商业秘密、个人隐私、军工科研生产等信息的保护。妥善处理发展创新与保障安全的关系，审慎监管，保护创新，探索完善安全保密管理规范措施，切实保障数据安全。

强化安全支撑措施。采用安全可信产品和服务，提升基础设施关键设备安全可靠水平。建设国家网络安全信息汇聚共享和关联分析平台，促进网络安全相关数据融合和资源合理分配，提升重大网络安全事件应急处理能力；深化网络安全防护体系和态势感知能力建设，增强网络空间安全防护和安全事件识别能力。开展安全监测和预警通报工作，加强大数据（特别是数据开放环境下）的防攻击、防泄露、防窃取的监测、预警、控制和应急处置能力建设。

积极研究数据开放、保护等方面制度，实现对数据资源采集、传输、存储、利用、开放的规范管理，促进政府数据在风险可控原则下最大程度开放，明确政府统筹利用市场主体大数据的权限及范围。制定政府信息资源管理办法，建立政府部门数据资源统筹管理和共享复用制度。研究推动网上个人信息保护立法工作，界定个人信息采集应用的范围和方式，明确相关主体的权利、责任和义务，加强对数据滥用、侵犯个人隐私等行为的管理和惩戒。推动出台相关法律法规，加强对基础信息网络和关键行业领域重要信息系统的安全保护，保障网络数据安全，研究推动数据资源权益相关立法工作。

4.7 行业自律控制风险

随着社会的发展，隐私权客体范围在不断扩展，以一种尺度便能试用所有情形为原则的立法方法不够精细，过于严格的立法势必影响大数据产业的发展，采取行业自律模式能够激发相关企业积极性。因此，通过产业联盟、行业协会以保护数据隐私权是法律法规的有益补充，在大数据的发展过程中具有积极的意义。

2014 年 6 月 19 日，中关村大数据交易产业联盟专家顾问委员会在北京成立，会议发布了我国首个大数据交易行业规范，规范旨在推动行业自律，打造完善、健康、有序的大数据交易产业链条，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，对交易市场内的在线数据交易、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 1 月，中国电信、中国联通、世纪互联、太平洋电信等国内数据中心行业骨干企业共同签署了由中国数据中心产业发展联盟发起的国内首部《中国数据中心行业自律公约》。公约为规范我国数据中心行业从业者行为，促进和保障数据中心行业健康发展制定的首部自律性公约，对我国大数据行业未来健康发展，以及提高大数据行业在国民经济的地位均具有重要的现实意义。

大数据的行业自律应当鼓励、支持企业开展合法、公平、有序的行业竞争，反对采用不正当手段进行行业竞争；自觉维护大数据用户的合法权益，保守用户信息秘密，承诺不利用用户提供的信息从事任何与向用户作出的承诺无关的活动，不利用技术或其他优势侵犯用户的合法权益；鼓励企业、科研、教育机构等单位和个人大力开发具有自主知识产权的大数据系统和基础设施产品等，为我国大数据行业的进一步发展提供有力支持。

大数据行业的发展无疑需要行业自律。另一方面，行业自律存在的缺陷也很明显。首先，行业联盟本身毕竟只是一个联合体，并不能直接监督企业，也无法对企业违反指南的行为进行制裁。其次，行业联盟的建立，又往往夹杂着物质利益。因此，各联盟往往不愿意推动联盟之间的资源共享，反而力求保护数据的封闭性和联盟的边际，为此，也阻碍了数据公开和分享。

我国有着不同于世界其他国家的独特文化背景与历史背景。因此，在个人数据隐私权保护模式的选择上，应该结合我国基本国情，吸取世界上先进国家的经验，平衡行业各参与方利益，在保护大数据隐私权的同时，促进信息产业的平稳发展。

行业自律不仅是大数据时代对发展网络经济的优先考虑，也是基于对隐私权保护的道德约束。我国长期以来受传统法律文化的影响，对隐私权的保护重视不够；要实现意识上的转

变，必然要经历循序渐进的过程。我国信息行业起步较晚，近年来虽然呈爆发态势，但总体上还较为落后。推进和普及大数据发展,让更多企业分享大数据时代的果实仍是当前及今后一段时期的重要任务。单纯运用法律规制模式，以严格立法的形式对个人数据隐私权进行保护，很可能会打击了相关行业发展的积极性，使其受到严格限制，失去发展良机；而过度依赖行业自律，又将造成大数据产业的离散化和分裂化发展，从而造成利益团体的各自为站，对大数据行业也极为不利。

因此，在完善相关法律规范基础上，国家应当召集具有专业素养的法律、经济界人士，建立个人数据隐私权保护的第三方监督机制。依托专家和智库资源，建立起行业自律机制，对相关企业经营行为进行规范，发挥企业在保护数据信息方面的自主性，增强企业内部员工保护信息数据的自觉性。通过接受投诉，处理数据隐私权纠纷，责令数据隐私权侵权者停止侵害、赔偿损失、消除影响，协助公检法机关调查取证等机制，完善侵权行为的事后救济，形成维护大数据良性发展的顺畅机制，从而更好地推动大数据产业的协同发展。

同时，推进大数据产业标准体系建设，加快建立政府部门、事业单位等公共机构的数据标准和统计标准体系，推进数据采集、政府数据开放、指标口径、分类目录、交换接口、访问接口、数据质量、数据交易、技术产品、安全保密等关键共性标准的制定和实施。加快建立大数据市场交易标准体系。开展标准验证和应用试点示范，建立标准符合性评估体系，充分发挥标准在培育服务市场、提升服务能力、支撑行业管理等方面的作用，积极参与相关国际标准制定工作。

第5章　促进中国大数据开放共享的探索

正如国务院《大数据行动纲要》中指出的，数据的充分开放及共享将为社会发展起到极大的推动作用。包括政府和企业在内的社会各方，对数据的开放共享的机理机制进行了各种探索，从技术措施到行政措施方面进行了各类有益的实验。

5.1 大数据局

大数据管理局的提法最早出现在2014年1月的中共广州市委十届五次全会上，成立这个“局”的目的是为统筹推进政府部门的信息采集、整理、共享和应用，消除信息孤岛，建立公共数据开放机制。一年多来，部分地方政府在这方面做了大量探索，如广东省、贵州省、上海市、沈阳市、成都市、珠海市等，出台了一些推进政府大数据管理的政策和制度。一些地方政府还成立了专门的部门，研究拟订并组织实施大数据战略、规划和政策措施，引导和推动大数据研究和应用工作；组织制定大数据收集、管理、开放、应用等标准规范等相关工作。成立大数据管理局，旨在消除过去的部门数据壁垒，从全局上规划跨部门、跨行业的大数据整合、分析和综合使用，真正形成大数据的综合使用效果，进而探索新的大数据应用模式。各地的政策力度不一，其中沈阳市政府赋予了大数据管理局很高的数据管理“行政权限”。这些探索在一定的程度上证明，高效的数据开放、共享和流通，真正消除数据共享中的部门壁垒，需要各级政府的强力保证，而非流于形式。只有这样，才能推动大数据产业的持续发展。

个别城市成立大数据局是一种管理上的探索，但不宜在全国普遍推广。信息化是一个相当长时期的社会转型，除了充分利用数据资源之外，还涉及传统产业的转型、信息消费水平的提高、政府治理的改进等方方面面的工作。信息化的各项工作之间有千丝万缕的联系，发展大数据与信息化管理部门过去分管的电子政务、信息化推广、信息安全等工作很难分割，与目前正在大力推进的“互联网+”也很难划清界限。发展大数据技术和推动大数据应用是信息化工作的一部分，不能以发展大数据代替全部信息化工作。

信息技术发展很快，新技术层出不穷。短短十多年间，移动互联网、云计算、物联网、大数据等技术相继成为热点。相对于移动互联网和云计算，大数据技术更不成熟。根据Gartner公司发布的新兴技术成熟度曲线预测，大数据技术还要5-10年才会成为市场主流技术。各地政府需要关注信息技术的发展趋势，但不能追着技术热点跑，更不宜因为一项信息技术的兴起单独设立一个行政部门。

大数据技术的兴起不是对以往信息化建设的否定或摈弃，而是信息化进程的一个新阶段。过去信息化推进过程中的问题不会因出现大数据而自动消失，反而会带来与大数据相关的新问题。在实际工作中难以找到专门针对大数据问题的解决方案，因为原来的问题不解决，大数据的发展就会受到牵连。信息化工作必须保持连续性。

长期以来，我国信息化管理机构的设置比较混乱。全国省级信息化有十多个不同类型的部门分头管理，有些省市甚至安排在党委、人大的机构之中，难以有效协调，地方反映强烈。随着信息技术的发展，各地政府先后成立了“信息化处”、“电子政务中心”、“云计算办公室”、“物联网中心”、“智慧城市办公室”等机构，各种信息化管理机构的职权划分不清晰。如果再成立“大数据局（处）”，可能让政府的信息化管理工作处于更加割裂的状态。党的十八大以来，国务院在大力推进政府精简机构、简政放权，信息化管理也要贯彻“简政放权”的精神，避免出现开展一项新工作就增加一个新机构的恶性循环。国家应做好顶层设计，规范全国的信息化管理机构设置，在原有机构的基础上重构，建立有指挥和协调能力的信息化管理体制。

5.2 块数据

贵阳市在大数据的探索实践中，率先提出来“块数据”的概念。所谓“块数据”，就是一个物理空间或者行政区域内形成的涉及人、事、物的各类数据的总和。打个比方，“块数据”就像一个计算机的主板，建立起了一个开放、共享、连接的数据基地，各个行业和部门的“条数据”就如同一个个可插拔的板卡，它们只有融合并集成到主板上，才能有效发挥数据资产的功效。

块数据为什么在贵阳发生？第一，贵阳市有电力、气候、劳动力便宜、地质结构稳定等优势，贵阳市搞数据中心有先发的优势；第二，由于数据发展的灵活性以及数据安全的敏感性问题，不适宜在大城市这样的敏感区域先行先试，贵阳市城市规模适中，适宜作为试验场。

在块数据的实践中，贵阳市围绕“数据从哪里来、数据放在哪里、数据谁来应用、数据如何使用”四个问题来具体展开。

一是搭建系统平台，着力解决数据从哪里来的问题。目前，大量有价值的数据在政府手里，政府的数据又分散于各部门，部门间“数据壁垒”和“信息孤岛”现象普遍。为解决数据从哪里来的问题，贵阳市以全域公共免费WiFi城市项目为基础，搭建块上集聚的大数据公共平台，实现政府、企业和社会数据的汇聚，并将各领域、各行业的“条数据”关联为符合问题导向的、多维度的、价值更高的“块数据”。目前，全域公共免费WiFi城市项目已开始

二期建设，政务数据共享交换平台、中小企业云已启动建设。

二是改善基础设施，着力解决数据放在哪里的问题。贵阳市像抓交通、水利建设一样抓信息基础设施建设，加快构建宽带、融合、安全、泛在的信息基础设施体系。中国三大电信运营商集团数据基地齐聚贵州，规划建设的数据中心服务器规模超过200万台，贵州已成为中国南方数据中心和长江经济带数据中心。

三是推广数据应用，着力解决数据谁来应用、如何使用的问题。贵阳市注重以大数据提升政府治理能力、以大数据服务社会民生、以大数据引领产业转型升级。在提升政府治理能力方面，贵阳市已经在12家政府部门实施了“数据铁笼”计划，用大数据编制制约权力的“笼子”，权力运行全程电子化、处处留“痕迹”，实现“人在干、云在算、天在看”。再如，贵阳市正在实施的大数据综合治税工程，就是依托政务数据共享交换平台，将全市36个单位和部门涉税信息全部集中到社会综合治税信息平台，形成数据采集、处理、挖掘、展示为一体的 “大数据治税”雏形。在服务社会民生方面，贵阳市正在实施“民生云”、“社会和（谐）云”、“医疗健康云”，推进教育、医疗、社区等公共服务领域的应用示范，汇聚各类民生数据，促进民生数据的开发和利用。中国首家互联网医院——贵阳朗玛互联网医院已于2015年5月26日在贵阳市正式开通。在引领产业转型升级方面，数据集聚人气，创新促发展。惠普、微软、戴尔、富士康、阿里巴巴、腾讯、百度、京东、浪潮等一大批国内外知名企业纷至沓来，寻求合作机会。大数据与传统产业的融合发展，在贵阳市已经催生了数据交易、数据加工、众包服务、众包制造、众筹金融、网络新媒体、互联网医院等一批新模式、新业态、新产业。截至目前，今年贵阳市新增大数据关联企业达1270户（注册资金100万以上）。

5.3 数据花园

数据花园是**大数据专委会**对我国近年来承办重大会议、活动等所采用模式的一种形象比喻，其寓意为“扎起篱笆，让篱笆内的数据能够真正开放共享，有效整合”。多年以来，各级政府的数据开放情况不容乐观，其原因有三：一是不敢开放（即不清楚数据开放的范围和粒度）；二是不肯开放（数据开放会影响部门的事权）；三是不情愿地开放（设置开放障碍等）。这使数据整合的有效进行步履维艰。回顾我国曾举办过的重大国际赛事、博览会，如 2008 年北京奥运会和 2010 年上海世博会。这些重大事件中，交通、商业、气象、警卫、卫生医疗等行业和部门在信息上可以有效整合，推动事件的有序进行。

因此，有部分业内人士提出“数据花园”的模式，即希望在花园内部可以充分进行数据的开放、共享，实现内部的数据无障碍使用。消除数据不敢开放、不肯开放和不愿意开放的

现状。同时，对进入花园使用数据的部门、企业或个人进行法律、法规、制度上的约束；并用各种信息技术手段保障。我们也认识到，虽然数据花园模式为我国的大数据应用提供了一个思维上的探索，但在实践过程中仍需不断技术探索和经验积累。

5.4 数据交易

数据交易是将数据作为一种资产进行交易的模式，它与软件交易、保险产品交易、金融产品交易等商品交易相同，是数据提供方、数据购买方及数据代理人对原始或经处理后的数字化信息进行交易的活动。传统的咨询公司可以视为数据交易在商业上的一种探索。宏观上，数据交易可以分为两类：一是基础数据交易，二是大数据交易。基础数据交易在国内外其实并不是一个新概念。之前各种咨询公司已经在或明或暗的情况下进行着一些数据交易活动，大部分交易是以原始数据或者分析报告的形式进行的。如北京中电经纬咨询有限公司（简称中电经纬）是一家专注于电力行业的咨询公司，其依赖于自身拥有电力能源行业数据以及采购的第三方数据，形成了相对权威的电力行业分析报告产品，主要交易对象是银行以及投资机构，也有部分能源行业企业。与基础数据交易不同，大数据交易是建立在大量的行业行为数据之上，通过收集数据、加工处理数据、分析数据，然后结合数据的商业应用场景对数据再处理、再分析得出可视化结果，最后通过交易规则将可视化结果出售给数据需求者。如罗计物流软件专注于提升物流行业运力与货物的匹配效率，通过产品免费的方式收集了大量的行业基础数据、运力行为数据、发货行为数据等，这为其核心战略目标“提升物流行业的整体效率”奠定了数据基础，成立不到一年已经顺利通过B轮融资，金额达到1.46亿美元。

贵阳市进行了数据交易所模式的探索。目前，很多企业拥有自己的大数据分析平台，也有了大量的数据资源，但是想要找到其他多种维度的数据进行补充很难，特别是政府方面的数据资源。其关键问题仍然是推动政府数据公开，从而推动行业数据共享，联合各个行业的领军企业基于大数据技术来解决各行业的难点问题。

5.5 数据的前置机模式

数据的前置机模式是政府部门和企业将数据开放给特定部门、企业等的一种特殊模式，其使用模式在银行等传统业务向外拓展时被普遍采用（如 ATM、POS、IC 卡等），它实现的主要功能有网络通信、交易数据格式转换、交易数据统计等。在大数据时代，这种前置机模式分为两类：一是限定使用场地的数据开放共享模式；二是限定使用数据的前置机模式。采用第一种模式的部门和企业，往往介意自身的数据外流。在需要开放数据访问时，它们通常会在特定的场地开放数据；所有对该部门或企业数据的访问被受限于该场地。目前，阿里、

百度等企业开放了该类型的数据访问。采用第二种模式的部门和企业，通常在意开放数据的范围，它们介意开放哪些数据。为此，它们通常希望数据使用方通过特定的前置机和前置接口来访问被共享的数据。在我国，政府部门之间往往通过这种方式来开放和共享数据。

前置机的使用模式，在使用上，结构比较复杂，维护工作繁复，运行效率也会降低。系统投资也相应增加，这些投资的主要来源为前置机服务器、操作系统、数据库、应用软件、通信设备和网络设备等的重复购置。事实上，前置机的模式只是大数据开放、共享和流通过程中的一种权宜方式。

5.6 异业联盟

异业联盟的原意为“水平结合”，是指产业间（而非上下游的垂直关系），而是双方具有共同营销互利互惠的水平式合作关系。凭借着彼此的品牌形象与影响力，来吸引更多意向群体的客源，借此来创造出双赢的市场利益。联盟的商业主体之间共谋发展、合作共赢，是异业联盟各商业主体的共同目标。

随着商业大数据时代的到来，各行业的市场竞争越来越残酷，竞争逐渐白热化。部分大品牌、大商家逐渐形成占领市场资源的格局。各个行业都是如此。大品牌为了主宰整个行业的发展，吸引大部分的意向客源。大量的中小企业和品牌的生存和发展受到巨大威胁。为打破这种局面，中小企业和品牌联合起来，积众为强，以跨行业的合作联盟方式寻求更大的发展空间，以产业联合、数据共享、利润共赢的方式抢占市场份额，制造规模效应、把握先机。异业联盟应运而生。

一般而言，联盟具有以下几个特点：a）联盟主体间具有差异性，即没有利益冲突；b）联盟主体具有非竞争性，即合作者不存在竞争性或是非直接利益冲突；c）联盟主体具有互补性，即合作将产生共赢的效果；d）联盟运作数据化，即联盟的主体间拥有的数据资源可以有机地结合起来，发挥其团体效益。在异业联盟范围内，任何参与的主体都会找到与其他参与主体的利益关联。如中国联通多次和商业实体合作，来拓宽各自的会员服务。

异业联盟的目的是借助第三方资源提升自己服务，谋求1+1>2的营销效果，但能否实现可持续的快速发展最终将取决于数据资源的更好整合和利用。

5.7 美国 NSF/NIH 模式

虽非我国在大数据方面的探索行为，但美国政府部门，尤其是NSF和NIH的模式依然值得我们借鉴。我们将其纳入“促进中国大数据开放共享的探索”，希望其他国家的有益探索

能在促进我国大数据开放共享方面提供宝贵的有益经验。

2008年，美国总统行政办公室和科技政策办公室联名向多个美国政府部门发出《科学研究成果发布原则》，为科学研究成果的提交提出方向性指导意见。同时，美国国会也指示“这些原则是为了确保科学研究的成果和数据可以有效开放、共享和流通”。该《原则》规定了科学数据及成果开放共享，以及科研机构在与新闻媒体交流所提供数据的基本准则。该原则规定，在不违反现存的联邦法律法规、总统签令及相关研究领域的现存惯例时，联邦机构的科学家所产生的研究数据应该最大化地公开：a）机构应当就如何分享联邦科学家所产生的研究数据及结果制定明确的指导方针；b）制定指导方针时，机构应当就如何保存及获取这些公开数据的事宜制定明确的政策；c）机构应当时刻注意并遵循这些指南，并且保证对大众公开的数据是准确的，是完整的，是及时的。同时，同行的评论可以为科技研究企业以后验证研究数据及结果的可信性提供重要作用，所以经常会被一起发布。机构应该采取措施保证同行评论是以不违反相关研究领域现存惯例的方式获得的。同时，一些美国部门还成立了科研诚信办公室，负责监督对不当行为的指控。以监测科研不端行为来帮助机构调查，并通过教育、预防和监管活动来促进研究的规范责任行为。”

为此，美国国家科学基金委（NSF）特地在其《奖励和管理指南》中专门就研究成果的开放和共享问题进行阐述。2010年，NSF声明更改其数据共享政策，要求自2011年7月起，所有申请NSF资助的项目计划必须要以两页补充文件形式提交研究项目的数据管理计划，该数据管理计划作为项目计划的一部分被评阅。

美国NSF的模式，使得科研单位在进行项目申请的同时，认真思考其数据开放共享的具体计划。该方式值得我们思考和借鉴。

5.8 众包数据采集

数据共享使得人们可以更加充分地利用已有数据资源，减少资料收集、数据采集等重复劳动和相应费用，实现深入全面的数据分析和智能化、个性化的应用系统开发。但是，由于共享数据具有多源异构等特点，其数据格式、内容、表示方式千差万别，使得传统的数据共享在实际应用中面对诸多问题，例如数据覆盖低、数据质量差、数据安全弱等。而基于众包的数据采集和共享是一种新型的数据采集与共享模式。即，将数据采集、计算、识别等的工作任务分配给非特定的大众处理。这种基于众包的数据共享通常由一定的奖励机制驱动，具有灵活性高、覆盖广、成本低等特点。

近年来，众包数据采集已经成为常见商业模式，在社会各行各业得到广泛应用。在互联

网行业，UGC（即用户生成内容）造就了Wikipedia、Foursquare、Youtube等公司。我国也出现了一大批典型的UGC企业，包括百度百科、优酷、微博等。360、百度等企业也通过众包数据采集的方式提供了多种服务，例如手机、邮箱的黑名单等。在一些基于地图的众包数据平台，用户可以选择周边的各种任务差事，比如回答食堂排队人多不多、看一段品牌视频回答几个问题、参与用户调研、免费尝试新品并提交反馈等，并获得相应的现金奖励。一些众包采集平台提供了采集声音、采集图片、采集内容、采集视频的完整解决方案。

随着众包模式理念的日益成熟和完善，众包数据采集和共享可以得到更为广泛的应用。例如在环境监控等应用中，需要通过数据共享来实现对环境的有效监控。但是环境监控采集设备有限，即使对北京这样的城市也无法实现细粒度的采样覆盖，无法满足不同区域用户的需要。一种可行的办法是通过众包采集，鼓励居民通过个人的传感设备动态感知环境数据，以有偿的方式发送给服务提供商。根据数据在时空上的缺失度，服务商可以对不同区域的采集信息给予差别化的奖励机制，增加对监控缺失区域的采集数量。这样的方法可以有效解决环境监控问题，促进数据的共享。众包模式的数据采集还可以用于社会化制造等应用领域，即将设计工作分为一系列识别和评价任务，基于一定的激励机制选择适当的人来处理，最后汇总结果得到“群智”方案。这种方式改造了传统的设计模式，在电影制作、服装设计等领域已有很多成功案例。

第6章　结论与展望

大数据技术的兴起正完成对各传统领域的颠覆。全球范围内，运用大数据推动经济发展、完善社会治理、提升政府服务和监管能力正成为趋势。各国已相继制定实施大数据战略性文件，大力推动大数据发展和应用。为了发挥大数据的价值，高质量数据的开放和共享决定了大数据应用的深度和广度。

从全球大数据发展的趋势来看，大数据产业推动社会生产要素的网络化共享、集约化整合、协作开发和高效利用，改变了传统的生产方式和经济运行机制，可显著提升经济运行水平和效率。大数据开放共享将持续激发商业模式创新，不断催生新业态，将成为互联网等新兴领域促进业务创新增值、提升企业核心价值的重要驱动力。

中国是数据生产大国。目前，我国互联网、移动互联网用户规模居全球第一，拥有丰富的数据资源和应用市场优势。如果能在大数据管理和分析技术的研发与应用方面取得突破，可持续推动互联网创新企业和创新应用的高速成长。我国各级政府已启动若干大数据相关工作，并坚持创新驱动发展，以加快大数据的部署。随着数据治理理念的影响渗透，我国的公共数据开放共享进程逐步加快。2013年，《关于促进信息消费扩大内需的若干意见》中提出“促进公共信息资源共享和开发利用，推动市政公用企事业单位、公共服务事业单位等机构开放信息资源”；2015年，《大数据行动纲要》将促进大数据的开放共享作为第一个重要任务。虽然，我国在大数据的开放共享方面加大了探索的力度，但总体而言，我国在大数据的开放共享方面仍有很长的道路要走，主要面临以下关键挑战：

第一，政策、法律、法规有待健全。为了实现对数据资源采集、传输、存储、利用、开放的规范管理，促进政府数据在风险可控原则下最大程度开放，明确政府统筹利用市场主体大数据的权限及范围。政府信息资源管理方面，仍需建立政府部门数据资源统筹管理和共享复用制度。研究推动网上个人信息保护立法工作，界定个人信息采集应用的范围和方式，明确相关主体的权利、责任和义务，加强对数据滥用、侵犯个人隐私等行为的管理和惩戒。推动出台相关法律法规，加强对基础信息网络和关键行业领域重要信息系统的安全保护，保障网络数据安全。研究推动数据资源权益相关立法工作。

第二，数据开放共享的壁垒有待消除。目前，我国行政的分割导致数据无法有效共享，其原因是我国政府数据资源多按地域或部门进行分割管理。不同地域和部门为了自身利益，形成人为数据共享壁垒，加大了政府大数据开发难度。由于政府部门业务管理信息系统开发和建设的“部门化”，政府信息系统出现“系统林立”和分裂状态，政府公共信息资源重复

采集现象严重，信息摩擦和治理成本偏高。

第三，数据使用和消费的模式有待探索。在全球信息化快速发展的大背景下，大数据已成为国家重要的基础性战略资源，数据的有效开放共享必将引领新一轮科技创新和数据使用模式创新。持续激发的商业模式创新，不断催生新业态，也将成为加快大数据开放共享的催化剂。充分利用我国的数据规模优势，推进大数据开放共享，实现数据规模、质量和应用水平同步提升，发掘和释放数据资源的潜在价值，有利于更好发挥数据资源的战略作用，增强网络空间数据主权保护能力，维护国家安全，有效提升国家竞争力。

第四，数据质量的控制方面。基础数据的质量决定了分析结果的可用程度。加快数据开放共享步伐的同时，必须提高基础数据的质量，并研究如何提高低质数据上计算和分析结果的准确性。提高数据的质量，有助于提高大数据分析结果的可用性，进而增强政府管理的现代化，加快工业从制造向“智造”的升级，促进商业模式的不断创新。

综上，加快大数据的开放共享是《大数据行动纲要》中提出的第一项重要任务，是大数据应用持续繁荣的基础。我们认为，加快大数据开放共享的步伐，有效保障数据的开放、共享和流通，是提高政府治理现代化的重要举措。

第二篇 中国工业大数据发展报告

本篇主要论述工业大数据与新一轮工业革命的关系，论述工业大数据对“中国制造 2025”的支撑作用，阐述工业大数据技术架构与发展趋势，反映国内外工业大数据发展现状和典型案例，明确未来中国工业大数据发展方向。

第7章 工业大数据之新一轮工业变革

7.1 国际工业发展趋势

新工业革命浪潮正席卷全球。随着全球化加剧各国生产力扁平竞争、老龄化带来工作人口下降和劳动力成本上升、自然环境的恶化和自然资源的不断减少以及更加多元化的用户需求，工业企业面临巨大的产业升级压力，如何向更加高质量、更快速响应市场需求、更个性化产品、更高效绿色的方向演进，是摆在每个工业企业面前的挑战。另一方面，自动化和信息化技术近年来得到了长足的发展，特别是信息技术随着消费互联网的发展迅速进步，在全面改造第三产业的同时，也进一步与自动化技术产生融合而进入到工业，为传统制造业的升级创造了技术基础。新工业革命以信息物理系统为载体，以创新商务模式为引领，以数字化、网络化、智能化为特征，其核心是将以云计算、物联网、大数据为代表的新一代信息技术与现代制造业、生产性服务业深度融合，以推动产业转型升级。在新一代信息技术支撑下，制造业产品全生命周期各环节的业务模式都将发生质的改变：在产品设计阶段，客户与合作伙伴将能广泛参与价值创造过程，基于众包、众筹、众智的创新设计转变，将使社会上广泛的非专业人员能够在云端共同参与产品系统的设计；在生产阶段，生产要素将能被高度灵活配置，企业可以突破传统企业边界连接并高效利用全球设计资源、制造资源，实现智慧化生产；在产品使用和服务阶段，产品使用方式、运维模式将会发生巨大变化，实现产权模式多元化、人机交互自然化、使用过程智能化、使用环境感知化，运营模式更加多样化。

新工业革命以美国的工业互联网、德国的工业 4.0 为代表，根据各自国家制造业发展优势的不同又各具特点。美国制造业大量外包生产环节，比如波音公司的飞机部件是在全球多个国家进行制造，但是美国制造企业牢牢占据高知识产权和高附加值的产品设计和服务环节，同时把控整个生态链的上下游为其服务，因此美国提出的工业互联网关注的核心不在实体工厂和制造本身，而在如何利用互联网技术的资源优化配置作用提升整个产业生态链的效率、提升服务和创造新的用户价值。德国的情况则正相反，德国的制造业优势在于实体制造，在于有精良的生产设备和工艺手段，因此德国提出的工业 4.0 重点在于智能工厂本身的建设，

并以智能制造为核心带动上下游业务如物流和服务等的发展。

虽然不同国家针对自己国情而制定的应对新工业革命的对策各有不同，但共识是新工业革命的主要特征是从自动化、信息化时代进入到网络化、智能化时代，大数据是新工业革命的关键技术要素。在德国的《工业 4.0 十大挑战与机遇》报告当中指出，数据的整合分析与使用是实现工业 4.0 的关键能力。

图 7.1 德国工业企业调查：工业 4.0 的机遇与挑战

（资料来源：“Opportunities and Challenges of Industrial Internet”, TNS Emnid, Germany）

工业 4.0 有两大关键支撑技术，一个是信息物理系统（Cyber-Physical System，CPS），主要特征是用更加智能的基础设施来降低车间复杂度和提高灵活性；另一个是数字化企业平台：跨生产“shop floor”和经营“top floor”、贯穿 CAD/PLM/MES/ERP 等生产、经营信息系统的信息集成与数据融合贯通，建立伴随产品制造过程的完整数据流，并基于对这些数据的分析，使企业可以全面深入把握和优化提升产品质量、生产效率、资源利用率。美国通用电气公司的《工业互联网白皮书》中指出工业互联网实现的三大要素是智能联网的机器、人与机器协同工作及先进的数据分析能力。工业互联网的核心是通过智能联网的机器感知机器本身状况、周边环境以及用户操作行为，并通过这些数据的深入分析来提供诸如资产性能优

化等制造服务。

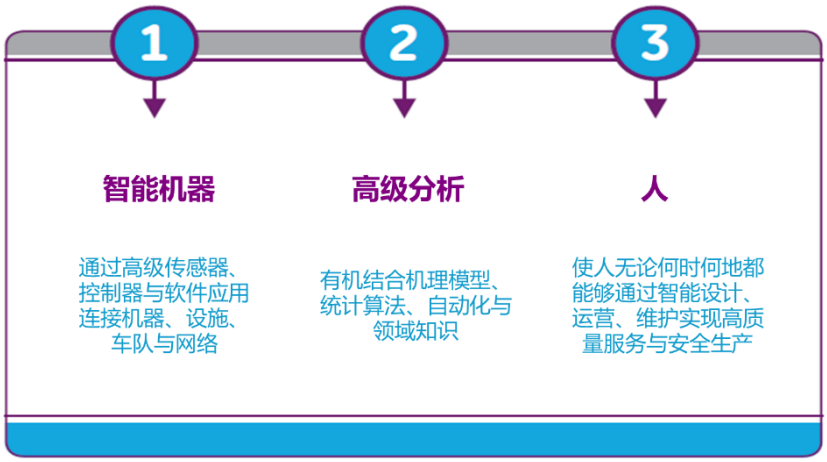

图 7.2 GE 工业互联网白皮书：工业互联网的三大要素

（资料来源："Industrial Internet: Pushing the Boundaries of Minds and Machines"，GE）

7.2 大数据对工业变革的支撑

大数据是制造业实现从要素驱动向创新驱动转型的有力手段。大数据可以帮助企业更全面、深入、及时了解市场发展趋势、用户潜在需求、竞争对手态势，以推出更有竞争力的创新产品。大数据可以支持企业利用众包众智的手段利用企业外部力量进行产品研发工作。同时，新产品研制过程中产生的海量实验数据利用大数据技术来管理分析也将大大加速产品试制迭代过程。大数据也是提升产品质量的有效手段。通过建立包括产品生产过程工艺数据、在线监测数据、使用过程数据等在内的产品全生命周期质量数据体系，可以有效追溯质量问题的产生原因，并持续改进生产过程的质量保障能力。通过关联企业内外部多数据源的大数据分析，可以挖掘发现复杂成因品质问题的根本原因。同时，大数据是提升生产效率、降低能耗，转变高耗能、低效率、劳动密集的粗放型生产面貌的必要手段。结合数控机床、工业机器人等自动生产设备的使用，并建立从经营到生产系统贯通融合的数据流，做到数据全打通和数据流通不落地，可以提升企业整体生产效率，降低劳动力投入，有效管理并优化各种资源的流转与消耗。此外，大数据也是实现工业企业从制造向服务转型的关键支撑技术。工业领域智能服务的本质就是智能产品加上感知控制能力和大数据分析，通过对产品使用过程中的自身工作状况、周边环境、用户操作行为等数据的采集和分析，可以提供在线健康检测、故障诊断预警等服务，以及支持在线租用、按使用付费等新的服务模型。

鉴于大数据技术对于新工业革命的要素支撑作用，在新工业革命的世界竞争中，工业大数据必将是各国信息技术企业竞争的焦点。要迎接新工业革命的挑战，必须发展工业大数据。

第8章 《中国制造 2025》与工业大数据

8.1 《中国制造 2025》的战略背景

中国已成为名副其实的全球“制造大国”。制造业贡献了国内生产总值的 40%以上。根据世界银行数据，2012 年我国制造业增加值在世界占比达到 20.8%，并超越美国成为全球第一制造大国。根据美国 Trading Economics 网站公开数据，2013 年 7 月份至 2015 年 6 月，中国制造业生产总量同比月平均增长率为 9.12%。

但是，我国制造业整体利润率低下。2014 年中国制造业企业 500 强的利润率仅为 2.7%，远低于世界制造业的利润率，创下 2009 年以来的最低水平。制造业正在承受产业“双向转移”的压力，一方面，劳动密集型的以出口或代工为主的中小制造企业正在向越南、缅甸、印度和印尼等劳动力和资源等更低廉的新兴发展中国家转移。另一方面，部分高端制造业在美国、欧洲等发达国家“再工业化”战略的引导下回流。如不能快速实现制造业的转型升级，在高端制造业产品尚未具备竞争力的条件下，中低制造业产品的竞争力也将被削弱，我国制造业“产业空心化”的风险不断增加。

不仅如此，麦肯锡全球研究院通过分析近两百年世界经济发展发现这样一个规律：一个国家从低等收入到中等收入，再到高等收入发展的过程中，制造业产值在 GDP 中占比呈现倒 U 字形，也就是大家常说的“苦笑曲线”。也就是说，在低收入的时候，一个国家的制造业在 GDP 中的占比是比较低的，但是随着整个国家的国民收入水平的提高，制造业在 GDP 当中的占比不断上升，当到了中等收入国家的时候，制造业在 GDP 当中的占比达到了顶峰。当它走向高收入国家的时候，制造业在 GDP 当中的占比就会下降。中国经济正在迈入中等收入国家的行列，也必须应对“苦笑曲线”带来的挑战。

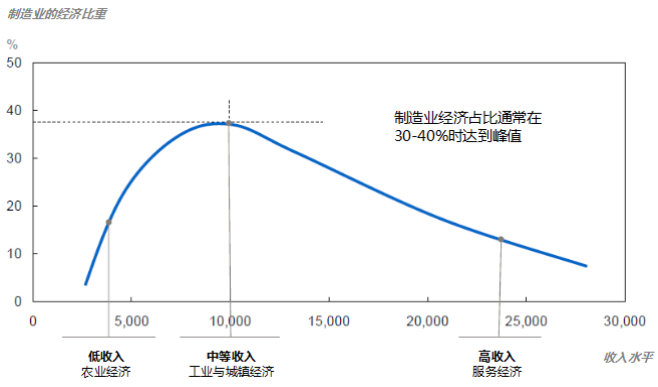

图 8.1 制造业在 GDP 占比的苦笑曲线

（资料来源：” Manufacturing the future, The next era of global growth and innovation”, Mckinsey Global Institute ）

当前，新一轮科技革命和产业变革与我国加快转变经济发展方式形成历史性交汇，国际产业分工格局正在重塑，我国制造业必须紧紧抓住这一重大历史机遇。因此，我国推出了“制造强国战略”，按照“四个全面”战略布局要求，加强统筹规划和前瞻部署，力争通过三个十年的努力，到新中国成立一百年时，把我国建设成为引领世界制造业发展的制造强国，为实现中华民族伟大复兴的中国梦打下坚实基础。

《中国制造 2025》是“制造强国战略”的第一阶段实施计划，为中国制造业未来 10 年设计了顶层规划和路线图，通过努力实现中国制造向中国创造、中国速度向中国质量、中国产品向中国品牌三大转变，推动中国到 2025 年基本实现工业化，迈入制造强国行列。《中国制造 2025》的核心思路可以总结为“一条主线，四个转变”：“一条主线”，即以体现信息技术与制造技术深度融合的数字化、网络化、智能化制造为主线。“四个转变”，即要素驱动向创新驱动转变，低成本竞争优势向质量效益竞争优势转变，资源消耗大、污染物排放多的粗放制造向绿色制造转变，生产型制造向服务型制造转变。

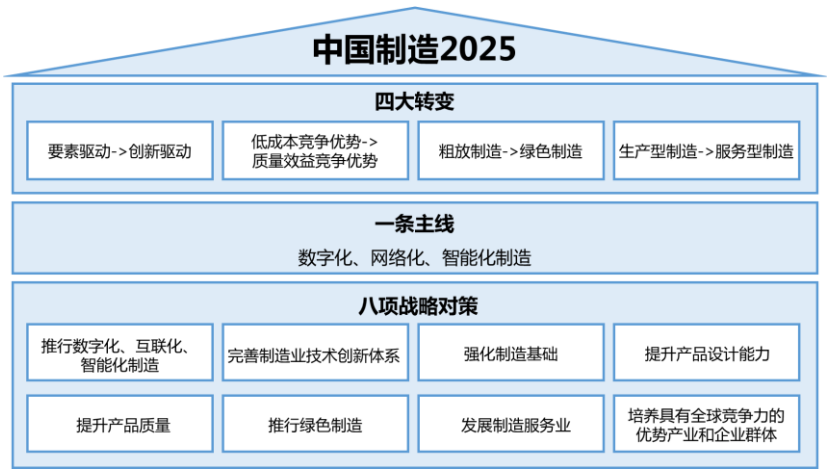

图 8.2 《中国制造 2025》顶层设计

8.2 工业大数据与《中国制造 2025》

《中国制造 2025》规划中明确指出，工业大数据是制造业转型升级的重要战略资源。如何有效利用工业大数据推动工业升级，需要针对我国工业自己的特点：一方面，我国是世界工厂，实体制造比重大，同时技术含量低、劳动密集、高资源消耗制造的比重也大，实体

工厂和实体制造升级迫在眉睫；另一方面，我国互联网产业发展具有领先优势，过去十多年消费互联网的高速发展使互联网技术得到长足发展，互联网思维深入人心，我们需要充分发挥这一优势并将之和制造业紧密结合，促进制造业升级和生产性服务业的发展。因此，我国在推进工业大数据的应用过程中，应当兼顾智能制造和制造服务，用数据驱动制造全生命周期从设计、制造到交付、服务、回收各个环节的智能化升级，推动制造全产业链智能协同，优化生产要素配置和资源利用，消除低效中间环节，整体提升中国制造业发展水平和世界竞争力。

另一方面，工业转型升级的迫切需要同时也为我国大数据技术的发展提供了最好的市场基础。纵观现代工业格局，发达国家在引领航空、航天、汽车、造船等高端制造业发展的同时，始终牢牢把控工业软件核心技术，主导着国际工业竞争的话语权。我国制造业领域的高端工业软件，无论是在基础软件领域的高端工业实时操作系统、实时数据库，还是大型应用软件领域的三维辅助设计、制造执行系统、产品生命周期管理系统、制造资源管理系统等基本由国外公司所控制。

面对新工业革命，工业数据已成为国际产业竞争和国家工业安全的基础要素，工业大数据软件也必将是各国信息技术企业竞争的焦点。我国大数据技术研究人员以及数据软件企业需要牢牢把握这一历史性的机遇，突破工业大数据关键核心技术，研发世界一流的工业大数据软件产品，培育工业大数据技术创新与应用服务生态体系。同时，积极把握中国工业转型升级带来的市场需求，推广自主工业大数据软件在装备制造、电力、冶金、石化、航空航天等工业领域的应用，把工业数据留在中国，保障我国产业经济安全，助力“中国制造”迈向“中国智造”。

第9章　工业大数据发展历程

9.1 什么是工业大数据

2012 年，GE 公司发布《工业互联网：突破机器与智慧的界限》研究报告，率先明确了“工业大数据”的概念。工业大数据的来源是在工业领域相关自动化与信息化应用中所产生的海量数据。在 GE 给出的工业大数据软件架构中，图 9.1 所示，这里的“相关应用”不仅包括企业内和产业链，还包括外部来自市场、用户和气象等数据。总结来说，工业大数据是由以下三方面数据汇聚而成的，来自于企业制造业信息系统的数据、来自自动化生产线和智能产品的物联网机器数据、来自互联网世界的相关外部数据。

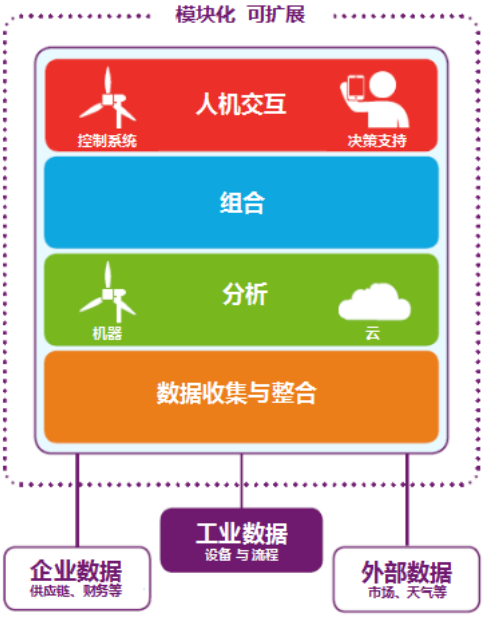

图 9.1 GE 工业大数据平台软件架构

（资料来源：“The Case for an Industrial Big Data Platform”，GE Software）

尽管相对于传统数据管理与分析技术，大数据本身缺乏一个形式化的定义，但是首先需要回答的一个问题是为何能够将工业领域的数据称之为大数据。在本节中我们从大数据较为公认的 4V 特点对工业数据分别加以分析，即 Volume（数据规模大）、Velocity（处理速度快）、Variety（数据多样化）、Value（数据价值密度低）。

与 GE 发布的报告同年，麦肯锡的一份大数据报告中给出了一个有趣的事实：那就是在虚拟经济占主导地位的美国，其工业界蕴含的数据总量反而是最大的。

图 9.2 美国各行业数据规模

（资料来源："Big data: The next frontier for innovation, competition, and productivity"，McKinsey & Company）

事实上这个结论并不令人意外，工业数据的主体，即由机器设备所产生的数据量是远超过其他行业以人为主产生的数据。以风力发电机为例，在终端上正常状态下每秒会产生一个数据包，包含 500 个左右的测点数据。如果全部数据需要处理与存储，1000 台风机会产生高达 50 万数据点每秒的数据写入吞吐。而无论是大型的风电场运营企业还是风电设备制造商，其需要监控的风机都会达到数千甚至上万的规模。在流程制造业中经常使用到的如鼓风机等高速旋转设备出于监控并通过频谱分析手段进行故障诊断的需要，其在每台机组上对振动量的采集频率可以高达上万赫兹。与金融、电信等传统服务业可以区分忙时与闲时不同，大多数工业设备的运转都具有长时间连续性的特点。另一方面，出于对历史数据分析以及安全生产审计的要求，数据通常会要求保存 5 年以上，甚至永久保存。因而其数据量大的特点可以总结为吞吐高（每秒可达上百万数据点）、不间断（24*7 工作）、总量大（数百 TB 到 PB 级的存储量）。

从处理速度来看，由于源数据的持续高吞吐，大数据处理平台必须能够高速地对数据进行实时解包、协议解析、格式转换等基本处理。而在越来越多的智能化应用中，需要能够进行实时的数据分析并完成相应操作。特别是在控制系统中，针对安全生产的实时故障检测要求从数据收集到完成数据分析能够实现秒级、甚至毫秒级的事前预警或者事后报警停机，以避免事故的发生或者对设备本身造成更大的连锁损害。

工业数据不仅仅是机器设备产生时间序列、时空数据、高维矩阵等数据，同时来自于如 ERP 等信息化管理系统的关系型数据、设计研发环节的产品图纸、工艺文档、加工代码等非结构化数据，以及来自与外部互联网上的半结构化（如 JSON、XML 等）与非结构化数据（如文本等）构成了一个典型的多样化数据体系。

由于大量的工业设备与智能产品在绝大部分时间内工作于正常的工况条件下，因而在工业大数据分析的典型场景中，以生产运营优化为目的的应用只是需要使用聚合后的数据，而以故障分析为目标的应用针对的数据仅为少量非正常的工况，因此相对于传统企业信息化数据而言，工业数据的低价值密度特点相对较低。

因此，工业数据作为一种典型的大数据，在具有广阔的应用前景的同时，对于传统的数据管理技术与数据分析技术也提出了很大的挑战。

9.2 从自动化与信息化到网络化与智能化

在工业 2.0 时代，制造业企业大多完成了生产制造过程自动化的改造，而 3.0 主要是制造业信息化，即传统制造业企业信息化的“四大件”，广义 PLM 系统（包括 CAX）主要支持产品开发、ERP 系统负责“人财物、产供销”、SCM 系统协调供应链，CRM 系统关照企业客户和用户。相比于工业 2.0 是自动化的应用，工业 3.0 是信息化应用，工业 4.0 有独特的特征是网络化与智能化应用，而数据恰好就是实现网络化与支撑智能化的载体。

在进入工业 4.0 所说的网络化与智能化时代，碰到的第一个挑战是 2.0 与 3.0 的基础是否打好了，这是我们所说的两化融合问题。信息化概念相对较为广泛，中国制造业企业的信息化起步于设计研发与经营管理的信息化，而自动化生产线上的信息化发展滞后，大多数企业的 SCADA、PLC、DCS 等自动化控制系统与信息化系统之间并没有形成完整闭环。而跳出生产过程来看，产品全生命周期包括三个阶段：开发制造阶段（即 Beginning of Life），使用维护阶段（即 Middle of Life）和回收利用阶段（即 End of Life），如图 9.3 所示。而现在企业的主要着力点是开发制造（BOL）阶段的信息化，MOL 阶段的主要使用的 CRM 系统与 MRO 系统往往得不到重视，再制造过程的信息化更是缺乏。

信息化的两大关键分别是流程和数据。如果从数据的角度来分析中国制造业的现状，我们会发现实现 BOL 阶段两化融合的基础之一，就是来自于企业制造车间的机器数据与信息化系统的业务数据之间的整合。而实现 MOL 阶段的信息化提升的关键，是产品工况数据、用户使用行为以及市场反馈数据与 BOL 阶段数据的集成打通。对于离散制造业，特别是以生产智能产品为主的装备制造业而言，最重要的用户使用行为数据就是来自智能设备的传感器产生的实时机器数据。EOL 阶段目前就是要充分利用 BOL 和 MOL 阶段的数据，并将 EOL 阶段的数据反馈给前两个阶段，构建服务于绿色制造的数据闭环体系。

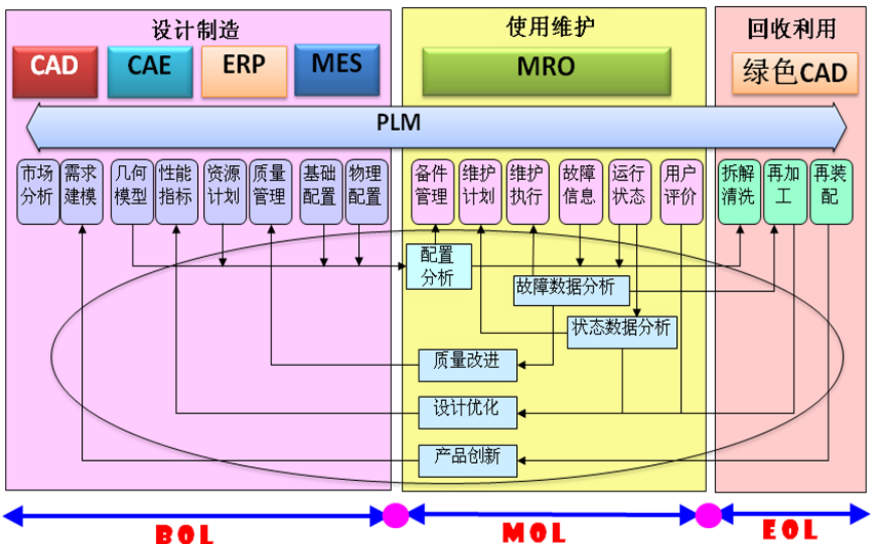

图 9.3 产品全生命周期

以上提到的传统企业信息化数据、机器数据和外部数据恰好就是工业大数据的三个组成部分，同时打通各个环节的数据流与业务流也是实现自动化与信息化融合的基础，因而**工业大数据的首要目标就是要实现产品全生命周期数据采集、传输、集成、存储与一体化的整合。**

工业大数据来源于产品生命周期的各个环节，包括市场、设计、制造、服务、再利用各个环节，“全”生命周期每个环节都会有大数据。但是企业外、产业链外的“跨界”数据也是工业大数据不可忽视的重要来源。工业革命以来，大部分技术和创新发生在车间，随着新一轮工业革命的到来，互联网已经渗透到经济社会的每一个角落，企业可以在整个产业链进行创新，以及进行跨界创新，跟其他产业融合。工业大数据可以把产业链上下游各个企业主体连接起来，促进创新，形成独特价值生态，这是**工业大数据的第二个目标，即实现跨界数据整合，推进网络化与自动化和信息化的融合，进而实现生产车间以外的跨界产业变革。**

制造业企业内、外数据的整合与流通无疑可以有效的提升了企业自身与产业链的效率，但在此过程中数据与数据之间依然是一个物理作用的过程。当我们使数据之间产生化学反应时，把数据提炼为知识，并把知识应用于产品全生命周期每个环节的自动化与信息化系统，实现设计、研发、制造环节的智能化，生产管理的智能化，服务的智能化以及产品本身的智能化，这就实现了整体效率的再次提升与价值再创造。这是**工业大数据的第三个目标：以大数据分析为手段，为实现智能化夯实基础。**

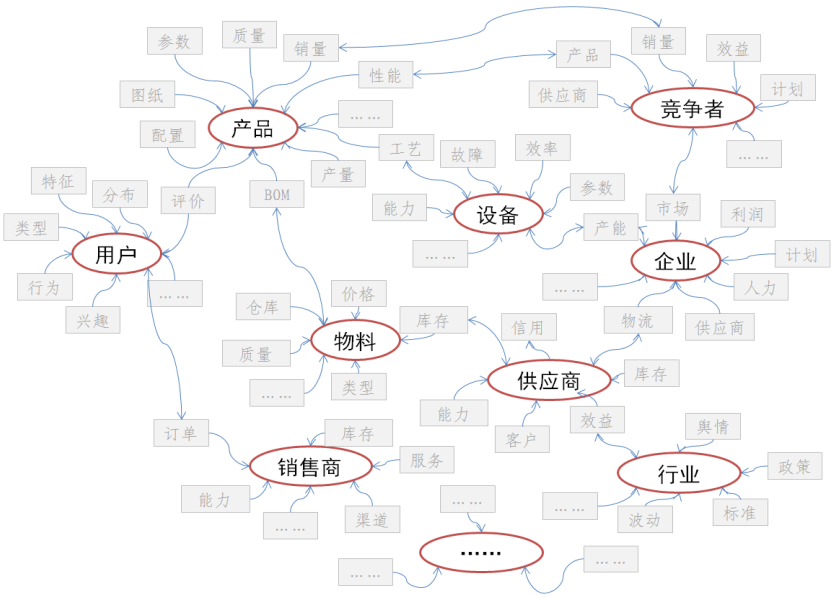

图 9.4 制造业生态系统与分析对象的数据关系网络示意图

相比于图9.3从产品生命周期的时间维度展示了从自动化与信息化到达网络化与智能化的路径，图 9.4 从空间维度展示了网络化后的制造业产业链生态体系，以及若干在每个点、边上的潜在智能化应用。以“用户”分析为出发点遍历其他节点，我们可以发现一条从需求逐步倒推到产品创新以及工艺优化、供应链协同、个性化服务等应用层面，最终通过从服务过程对用户行为以及变化趋势的精准把握，再次驱动产品的升级与创新的闭环路径。

9.3 国内外发展现状

应对工业大数据的巨大战略机会，国际知名工业软件公司开展了大量的创新产品和系统的研发。GE 联合 Pivotal 基于 GE Historian 实时数据库与 Hadoop 大数据平台推出了整合的工业大数据平台，并提出了工业数据湖的概念，在其中实现海量工业数据的整合与处理。同时，GE 开放了工业互联网云平台 Predix。Predix 是一个云操作系统，负责将各种工业资产设备和供应商相互连接并接入云端，实现工业数据管理与工业数据分析，并提供资产性能管理（APM）和运营优化服务。SAP 开发了面向物联网应用和实时数据处理的 HANA 大数据平台，并利用其在传统企业信息化 ERP 系统上的优势，推动 HANA 与信息化系统的集成。PTC 则收购了物联网云平台公司 Axeda，以此为基础打造智能互联产品 M2M 云平台，并在此之上针对制造业提供了相应的智能化解决方案。

我国工业大数据的发展相对较为缺乏体系，市场主体以传统实时数据库厂商、提供基于开源 Hadoop 平台的信息化厂商以及专注于制造业企业信息化应用的解决方案提供商为主。

由于缺乏系统性的规划与整合，一方面欠缺专门针对工业大数据特点的拳头平台软件产品，另一方面大数据分析的应用较为零散化且与业务结合的深度有所欠缺，因此总体而言其产品与国际厂商相比仍有较大差距。

我国制造业领域的高端工业软件 80%的市场份额长期被国外厂商所占领，信息化领域的关系型数据市场更是如此，但工业大数据软件不能再重蹈覆辙。对于智能化变革中的企业而言，数据是最重要的资产之一，而特别是对于航空航天、军事装备、能源、交通等国计民生息息相关的基础制造业而言，数据更是事关国家安全与主权，因此我们必须有中国自己的工业大数据架构体系。同时，我们也注意到国外厂商在工业大数据软件的发展方面也才刚刚起步，如果积极把握全球大数据软件开源化趋势，是非常有希望实现国产化工业大数据软件技术的弯道超车与应用的自主可控。

9.4 工业大数据平台的技术体系与挑战

区别于图 9.1 所展示的 GE 工业大数据软件架构，清华大学工业大数据研究中心通过系统性的研究与分析，提出了如图 9.3 所示的工业大数据功能架构体系，共分三大流程与九个功能模块。

在流程 1 中，重点实现工业大数据的主体，即机器数据的实时采集、处理与存储。流程 2 的主要功能为数据治理，完成对数据模型的规范与统一管理、数据质量分析、信息系统业务数据与外部数据、机器数据的集成。流程 3 主要针对数据分析应用开发的数据探索、分析与实验反馈过程。限于篇幅，本书无法对每个模块的功能与技术研究挑战展开详细分析，仅选取工业数据存储一个功能进行简述。工业大数据分析将在 9.5 节进行阐述。

工业数据存储平台主要需要完成对工业大数据的存储、查询与分布式分析计算功能。我们首先来探讨市场现有三类技术所面临的挑战。

传统的关系型数据库例如 Oracle、DB2 等多被应用于存储低频、少量设备的机器数据场景。现有基于行存储的关系数据库系统基于关系模型存储，对于机器数据的格式没有原生的支持，数据写入性能距机器大数据系统的要求有一到两个数量级的差距。同时在数据存储方面的额外负担较大，一般数据经数据库存储建立相关索引后，数据大小膨胀为原大小的数倍。列式数据库主要是针对数据仓库的查询场景进行优化，与关系数据库有类似的问题。

图 9.5 工业大数据功能体系架构

NoSQL 数据库突破了传统关系型数据库的诸多限制，也有实际应用使用例如 HBase、Cassendra、MongoDB、CounchBase 来进行机器数据的存储。由于抛弃了事务与锁机制的处理而且底层直接通过文件对数据进行存储，其对于写入操作的处理速度可以达到关系型数据库的数倍。但就其本质依然不是为时间序列或时空数据的原生结构而设计，通过其存储的时间序列数据在存储效率与丰富的查询能力等方面依然有所欠缺。

PI Server、GE Historian 等实时数据库系统作为工业企业多年来采用的机器数据存储方案，从存储结构到系统架构都是为海量时间序列度身订造，特别是将数据的写入速度达到了百万数据点/秒。但其主要问题在于机器数据分析时仍需大量关联静态数据，例如传感器的部署信息等。而此类操作通常需要将时间序列数据于关系型数据库中存储的信息进行跨库的连接操作。另一方面，由于底层基于平面文件存储的方式仅能按时间维度对数据进行排序存储，对于数据多条件复杂查询以及分析性查询支撑不足。因此，在实践中实时数据库仍被作为在线数据的存储平台，而数据分析等工作需要配合在其他系统完成，例如 GE 的工业大数据平台采用实时数据库加 Hadoop 平台的混合解决方案。

理想的工业大数据存储平台仍有待相关前沿技术的研究和突破。从系统架构角度，与传统实时数据库与关系型数据库的小型机加存储的架构不同，工业大数据存储平台应采用大数据系统流行的基于 PC 服务器加内置盘的全分布式架构，数据通过多副本备份的方式提高系统的数据安全性和高可用性。从数据写入速度的角度，工业大数据存储平台实时写入或从实时数据库系统同步数据的速度需要超过现有的实时数据库系统，即单节点达到数十万到一百万数据点/秒，而且可以通过扩展服务器达到近线性的扩展能力。从数据查询、分析的角度，工业大数据存储平台需要支持多维度的高速数据检索。同时基于时间序列的查询分析，提供

比实时数据库和关系型数据库更为丰富的查询语义，比如时间序列切片查询、同比环比查询、模式片段查询等。此外，能够基于完全分布式架构支持 Map Reduce、Spark 编程框架直接对实时数据进行访问，可以编程实现更加复杂的数据分析语义，从而提供分析与查询一体化的数据存储平台。从存储空间的角度，由于需要存储海量的历史数据，系统需要针对不同历史周期的数据采取不同的压缩算法（或参数），实现多层次的压缩，进一步节省数据存储空间。因此，与工业大数据存储类似，工业大数据的管理技术仍需要我们的科技工作者认真分析工业数据的负载和应用特点，通过不懈努力实现技术突破。

9.5 工业大数据分析技术与挑战

当前很多流行的大数据理念来自于互联网和商务领域，不少分析技术也是针对商业大数据。工业大数据与商业大数据在很多地方存在比较大的差别，本文将从如表 9.1 所示的四个角度来分析工业大数据对分析技术带来新的挑战。

研究对象不同：工业领域以物理系统（物理实体或环境）为中心，研究动态过程的规律和因果关系，而商业大数据以人造系统（人或流程）为研究对象，试图理解其中的行为模式。

现有基础不同：在工业领域，人们对生产过程的研究一般比较深入，形成了很多系统化的中观、微观机理模型，领域知识也比较丰富。客观来讲，对物理系统本身的突破性知识发现难度很大。工业数据中体现出来的物理系统间的规律难以突破现有生产技术人员的认知范围。与之相比，商业领域中仅存在一些宏观理念，定性描述人的行为偏好和经济活动规律，给大数据分析留有广阔的提升空间。

新的驱动力不同：感知技术的发展和普及是工业大数据的驱动力，现有的工控技术很难处理大数据量的挑战，大量的监测数据也为大数据分析带来与业务数据融合分析的机会。而互联网的发展为企业带来与客户交互的新渠道，极大促进了商业大数据分析的发展。工业领域的大数据大多是具有时空信息的结构化数据，且背后有明确的物理结构（如系统动力学、网络拓扑关系等），对时间序列、时空模式、序列模式等结构模式挖掘非常重要。而商业大数据分析大多集中在结构化的数据仓库表或非结构化数据（如文本、视频），数据间除了实体关系和部分时空信息外，结构性关系较弱。

对分析技术的要求不同：工业系统的实时性高，动态性强，对分析结果的精度要求高，很难接受概率性预测，而商业应用常遵循大数原则，概率性的分析就可以为运营提供很大的帮助。

表 9.1　工业大数据与商业大数据的区别

	工业大数据	商业大数据
研究对象	以物理实体与环境为中心（Cyber-Physical-People）	以互联网支撑的交互（Cyber-Cyber-People）
现有基础	中/微观机理模型与定量领域知识，在当前基础上前进“半”步都很困难	宏观理念与定性认识，存在广阔的提升空间
新驱动力	新的感知技术 产品的服务化转型	新的交互渠道（如社交媒体）
对分析的期望	因果关系才有用 模型的高可靠性（很难接受概率性的预测）	相关性关系就非常有帮助 大数原则

鉴于上述区别，在业务场景选择上，工业大数据分析应集中在：1）物理过程和业务过程的融合。能将物理量与经营过程量（如产品质量、生产效率、设备可靠性等）的关系定量化，突破现有生产技术人员的知识盲点，实现过程痕迹的可视化；2）对于物理过程环节，重视知识的“自动化”，而不是知识的“发现”。将领域知识进行系统化管理，通过大数据分析进行检索和更新优化；对于相对明确的专家知识，借助大数据建模工具提供的典型时空模式描述与识别技术，进行形式化建模，在海量历史数据上进行验证和优化，不断萃取专家知识；3）“软”测量。在工业应用中，不同过程量监测的技术可行性、精度、频度、成本差别较大，通过大数据分析，建立指标间的关联关系模型，通过易测的过程量去推断难测的过程量，提升生产过程的整体可观可控。

在分析软件上，工业大数据分析应解决如下技术挑战：1）建模效率。针对时间序列、时空等结构化数据，应提供丰富的特征模板库，方便对典型物理事件（如风速平稳时段、发电机转速快速下降、环境温度逐渐上升等）的描述；另外，还应提供丰富的时间序列、时空模式、序列模式的深度挖掘算法库，提升工业数据分析的建模效率；2）分析软件工具的运行效率。工业大数据分析必须能够满足大规模、分散控制和动态改变的要求。在实时处理上，传统的商业数据分析系统尚未能够支持面向大规模数据状态下的低等待时间复杂事件检测。在离线分析上，前台分析建模应与后台的工业大数据平台应有很好的整合，支持大数据上的挖掘；3）现有分析工具资产的有效利用。针对重要的应用需求，工业企业通常有一定的分析工具和科学计算软件积累，然而这些工具通常没有考虑大数据架构。如何有效重用这些分析工具，是工业大数据分析软件应该解决的问题；4）专家知识库的建设，将领域专家知识进行有效的沉淀、萃取、自动化，并融入到工业大数据分析中来。

第10章　离散工业大数据

飞机、高铁、火箭、武器装备、船舶、电子设备、机床、汽车等制造业，都属于离散制造型企业。离散制造的产品往往由多个零件经过一系列并不连续的加工工序，最终装配而成。生产此类产品的企业即称为离散制造型企业。

由于离散型制造企业核心业务是把自己加工或者采购的物料组装成产品（广义的物料），进行销售和服务，因此其运作和管理是围绕物料及其变化过程展开的。可以说离散制造型企业是以物料为主要操作对象，这些物料有基本信息。而其变化过程（形态变化、位置变化、价值增值等）衍生了其他信息，管理系统正是通过对物料的信息处理、分析来决策对实体物流实施何种动作。

离散型制造企业其生命周期包括市场、研发、设计、制造、服务、再利用各个环节，每个环节都有大数据的沉淀。但从应用角度来看，每个环节的智能化应用对数据的需求是全方位的，并不局限于本阶段所产生的数据。本篇将重点介绍典型产品生命周期中研发和服务两个阶段的工业大数据。

10.1 研发大数据

10.1.1 研发大数据的定义

研发大数据的定义是：产品研发所需要的和产生的大数据。这里的数据是广义的，包括信息和知识。

研发工作是以研发人员为中心。研发大数据主要用于提高研发人员的研发创新能力、研发效率和质量，支持协同研发。

10.1.2 研发大数据的主要内容和需求

1）用户需求的大数据

用户需求是研发灵感的重要源泉，满足用户需求也是研发的目标。

（1）基于网络的用户需求大数据

在当今的网络时代，用户的大量需求出现在网络空间中，如博客、网络社区、电子商务网站等。这些需求数据非常分散，混杂在大量的无用的、重复的“数据垃圾”中，对此需要进行过滤。目前有些企业建立了网站、网络情报搜集系统等搜集用户需求，但绝大部分企业还没有开展这方面的工作。

（2）来自销售线的用户需求大数据

现在的企业越来越重视建立自己的销售渠道和网点，因为不仅有助于建立自己的品牌，

还可以在销售中直接接触用户，了解用户需求。这对于一些快速消费品，尤为重要。如世界上最好的服装企业之一——ZARA，其销售门店的每一位销售人员见到顾客的时尚服装，听到顾客的意见，都通过手机、Ipad 等及时反馈到企业总部，帮助研发人员快速开发出受顾客欢迎的时尚服装。其每天汇总的数据量巨大。

（3）来自服务线的用户需求大数据

用户在不同的环境中使用产品的过程中，会发现许多问题，提出许多新的需求。这些都往往是设计人员所不曾想到的。这些需求正是研发的重要源泉。如海尔集团十分注重来自服务过程中发现的用户需求，一款“洗地瓜”的洗衣机的需求就来自服务过程。服务过程中的问题和需求的数量很大，构成了研发所需要的大数据。

2）研发知识大数据

研发一般都需要用到以往的设计数据、经验、模型、知识等，这些统称为研发知识大数据。研发知识可分为显性和隐性两部分。隐性知识存在于员工的头脑里，并被逐渐地遗忘和流失，只有很少一部分被记录下来，这就是显性知识。但显性知识大多分散在不同设计人员的计算机和笔记本中，很难得到共享。

（1）显性知识：显性知识包括两大部分：①外部的公开知识：如专利、标准、期刊论文、网络文章和随想等。这些知识数量多，许多知识是非结构化的，需要对这些知识进行过滤、有序化；②内部的知识：如研究报告、测试数据、研发经验、失败教训等，不仅需要将已经记录的研发知识进行共享，还要将分散在员工的头脑里的研发知识及时整理出来，以便共享和重用。这需要知识管理系统和产品生命周期管理系统的支持，并需要有一套能够激励员工贡献自己知识的方法、制度和文化，支持“集思广益”。显性研发知识往往很多、很乱，需要进行有序化整理，这样才能提高研发知识的利用效率。

（2）隐性知识：许多隐性知识是难以显性化的。因此需要将隐性知识定位到具体的研发人员，以便在需要的时候可以快速找到掌握某些隐性知识的研发人员，这对于充分发挥研发人员的聪明才智具有重要的作用。这里需要通过对研发人员在进行研发和知识共享时的各种活动所积累的大数据的集成和分析，了解研发人员的“所知”。

3）产品重用大数据

研发中重用以往的产品零部件，可以降低成本，缩短交货期，提高质量。需要对以往的产品零部件的大数据进行优化，即模块化、序列化和标准化，使杂乱的大数据变成有序的小数据，提高产品中重用的零部件比例，支持大批量定制。网络零件库，如国外最大的零件库（transparts.com）已经有 1 亿多种 3D 零件模型，这些零件模型是零件专业生产企业提供的，

他们面向全球用户，形成较大批量。研发中重用这些零件模型，不仅提高研发速度，而且成本低、质量好。

4）研发协同大数据

研发协同大数据包括：

（1）企业间的研发协同大数据：现代产品越来越复杂，对研发的实效性和成本控制要求越来越高，企业需要专注自己的核心能力，将大量研发任务外包。大范围的协同研发遇到的最大问题是企业的诚信问题、知识产权保护问题。需要对协同研发全过程进行监控，需要对协同研发中的各个环节进行评价，由此会形成大数据，利用这些大数据，可以帮助管控企业诚信、研发质量、知识产权保护等，最终形成高效协同研发的环境。

（2）企业与用户间的研发协同大数据：互联网和 Web2.0 的发展，使得用户可以方便地参与产品研发，如小米手机的 1/3 的功能是用户在网上参与设计的。这里被采用的设计方案是“百里挑一”出来的。又如海尔有几十万用户参与家电的设计，这里也存在研发协同大数据。

（3）产品全生命周期各个环节之间的研发协同大数据：研发需要利用产品生命周期管理（PLM）系统，集成 CAD/CAE/CAPP/CAM/CAT/PDM 系统和相应的大数据，在此基础上开展并行工程，组织与产品全生命周期有关的设计、制造、装配、使用和维护人员协同研发，提高研发效率，提高产品开发设计的一次成功率。

（4）多学科协同优化大数据：复杂产品研发需要进行多学科优化，这里有大量的数据进行交互，需要依靠数据进行产品性能、物理、装配、制造、使用、维护等仿真。

5）产品制造过程大数据

产品制造过程中的质量、成本、时间、环境影响等大数据都是研发中所要考虑的因素。随着 ERP（企业资源计划）系统和 MES（制造执行系统）的深入和广泛的应用，质量、成本、时间等大数据的获取将变得越来越容易。

6）产品使用状态大数据

产品制造企业需要了解一些重要产品的使用状态，如能耗、“三废”排放、工况等，不仅帮助企业进行产品的远程诊断和监控，还有助于企业改进设计，提高产品质量，还可以支持企业开展租赁服务，例如，现在已经大量出现的大型工程机械的租赁服务和飞机发动机的时间（功能）租赁服务。每一种租赁产品的实时信息都将记录并被集成，形成大数据，有效支持产品以及产品功能的租赁服务。

产品使用状态的监控往往需要许多传感器，并且往往是高密度采集，产品生命周期又很

长，因此所采集的数据量十分庞大，例如，劳斯莱斯公司对全世界数以万计的飞机引擎进行实时监控，每年传送 PB 数量级的数据。这里需要指出，产品使用状态数据在大多数情况下是光滑曲线，即正常工况，对后续的数据分析和知识挖掘意义不大。

7）产品维修大数据

有些产品一生中要经过多次维修，例如工业汽轮机。这些维修数据都要记录下来，不仅有助于后面的维修工作的开展，对于产品设计的改进也有好处。维修数据也是十分庞大的。

8）产品环境足迹大数据

减少产品全生命周期对环境的影响已是研发的主要目标之一，这需要产品环境足迹大数据，包括：能耗、“三废”排放、零部件再制造、回收重用等方面的大数据。产品环境足迹的获取需要有关部门的法规、标准和制度的强制措施，尽管这些数据与企业利润关系不大，企业往往对此视而不见，但是这些数据承载着企业的社会责任。

10.1.3 研发大数据的特点

（1）研发以人为中心，大数据支持人的创新

研发主要靠人的创造性工作，因此大数据主要是辅助研发人员的工作。

（2）从小数据到大数据，再从大数据到有序化的小数据

研发大数据来自产品生命周期各个环节，跨产品和跨行业，种类繁多。研发大数据最终要研发人员使用，需要对大数据进行有序化和简化，这样才能提高大数据的利用效率。

大数据有序化和简化的内容包括：去掉没有价值的数据；从大数据中自动分析得到有价值的信息；产品序列化、模块化和标准化；确定知识价值和关系；……。

（3）从有序化的小数据到研发智能化

集成大量的由实验、现场经验、理论分析得到知识，并将这些知识标准化，嵌入到专业产品研发软件系统中去，提高系统的智能化程度，使一般的研发人员也可以进行复杂产品的研发。

（4）大数据的自组织优化

通过研发人员使用、评价和应用研发大数据的行为，使基于大数据的研发信息管理系统，如知识共享平台、网络零件库等，越使用越聪明、越完善，大数据也得到自组织优化。

10.1.4 研发大数据的研究与创新

（1）基于大数据的研发知识共享方法和平台：不仅帮助研发人员共享知识，更重要的是通过对研发人员在知识共享平台中的表现，如知识发布、评价、整理、应用等活动的大数据，对研发人员的知识领域和水平、知识贡献度等进行评价，并与研发人员的奖励、工资、

升职、荣誉称号等挂钩，激励研发人员积极共享知识，不仅使自己的知识为企业创造更多的价值，使自己获得更多的回报；而且形成有序化的知识网络，建立起企业研发创新的基础设施，开展协同学习和协同研发，减少重复学习和研究；最终促进企业形成知识共享、协同创新的文化，提高企业的竞争力。知识共享与知识保密是一对矛盾，如何平衡，需要研究。

（2）基于大数据的网络零件库方法和平台：网络零件库对开展大范围的企业专业化协同具有重要价值。利用网络零件库中的 3D 零件模型建立、使用、评价等大数据，可以提高网络零件库的使用效率，可以帮助进行零件模块化、序列化和标准化，可以对模型的供应商进行评价和监督，帮助实现大批量定制。

（3）基于大数据的产品生命周期管理方法和系统：不仅帮助获取产品生命周期管理各个环节，了解各个环节的问题和需求，支持研发人员的创新，而且帮助进行产品生命周期的状态的有效管理，为用户提供精准的服务。

（4）基于大数据的产品研发智能化技术和系统：开展研发大数据分析和知识发现，将知识嵌入到产品研发软件中，形成研发智能系统，支持快速研发。

10.2 服务大数据

随着宏观经济形势下行通道的出现，人口红利的逐步消失和互联网企业的跨界竞争，中国的传统离散制造业格局正悄然变化。各种新观念、新模式、新技术向传统企业奔涌而来，一时间泥石俱下、摧枯拉朽般将企业多年来苦心经营获得的规模优势、渠道优势、成本优势和品牌优势纷纷瓦解。今天，无论是消费品制造企业还是装备制造企业无不陷入转型的焦虑和迷茫之中，而企业在变局中谋求重生的过程中，服务大数据将扮演十分关键的角色。在详细讨论服务大数据的定义与作用之前，先来看看三个不同行业利用服务大数据的案例：

案例 1：服务大数据驱动产品创新

当今，传统制造业企业受到互联网冲击最大的莫过于直接接触的大众消费品制造业，因为互联网的聚合效应直接影响的对象就是消费者。面临这种压力时，制造业企业并非无所作为，下面是传统消费品制造业利用大数据实现突破的一个典型案例：

2014 年 8 月 15 日，百度鹰眼宣布与 MTK（联发科）合作，针对 LBS（基于位置的服务）的应用推出更新的室内外导航方案，用于可穿戴智能硬件产品。通过 MTK 的算法结合百度 LBS 应用，可以快速实现更准确的室内定位。据称，通过 WIFI 定位，精度已经可以控制在 20 米左右。随着蓝牙 beacon 定位应用的普及及算法的优化，之后室内定位的准确度会更高。

在全新的百度鹰眼发布后，361 度运动鞋厂家随即推出了应用百度鹰眼服务的儿童防丢鞋。为了增强鞋子的安全性，361 度儿童鞋把防丢模块内置到了鞋垫下面。模块的 GSM 天线内置于鞋底，由于经过了特别设计，孩子穿鞋后不会影响到 GPS、WIFI 等天线的正常工作。鞋子的外观与普通鞋子并无二致，这样的设计是为了尽可能增加孩子的安全性，特殊的鞋子有可能会引起不法人员的注意。鞋子采用了可更换模块设计。如果鞋码不合适了或要换其他款式的鞋子，只要将防丢模块取出，换到 361 度同款新鞋子里就可以继续使用。这个防丢鞋还配置了无线充电底板，在回家换鞋时将鞋子放到底板上即可自动充电。

这款产品发布以后其最大的买点没有体现在鞋子本身，而是体现在透过产品回收到的儿童位置数据上，利用这些数据 361 度联合百度地图将数据对数据进行综合分析，提供了防丢、报警、健康、社交以及一系列服务，并将这些结果以 APP 和微信的方式将信息推送给关心小孩的用户，由此突破了传统的国产制鞋业一直在低端徘徊的格局。更重要的是利用这些数据 361 度可以精确把握小孩的行为特征，从而为后续的改进设计提供了依据，为后续的个性化定制打下了牢固的基础。

案例 2：服务大数据驱动制造转型

如果消费品制造仅仅代表了轻工业数据的话，下面这个定制家具行业的案例则反映了服务大数据广阔的内涵，因为数据不是在产品卖出后才有的，而是在产品制造前就开始产生了。

家具行业一直以来是一个古老的劳动力密集型工业，近年来受到宏观经济政策的影响，家具行业的刚需下降，资金回笼速度变慢，成品家具厂商的库存压力日益明显。广州尚品宅配家具股份有限公司，却通过服务大数据实现了家具行业的转型。这家公司成立之初是一家以家具 CAD 为产品的软件公司，公司初创人员都是 IT 技术人员，2000 年其产品推出以后，没有 3 年销量即开始下降，因为当时的家具企业购买 CAD 软件知识是为了摆门面，买来以后将软件搁置一边，继续按照旧有的模式生产成品家具。

最初，为了带动软件销售，公司成立了一个家具定制服务的网站：新居网。经过反复试错，不断改版，公司逐步将新居网发展成为一个集装修、家具咨询设计服务于一体的平台。几年下来，网站的后台数据库积累了全国 48 万个楼盘的 159 万个户型数据，并且形成了 463 万个方案，通过这些服务，尚品宅配形成了家具行业的三大核心数据：房型数据、方案数据和家具数据。现在，该网站每天接受全国的订单量超过 1 千笔，聚集了不同水平的 1 万多名设计师队伍。

完成数据的原始积累之后，尚品宅配开始摒弃使用中小家具企业代工生产的方式，自建

工厂，通过对三大核心数据反复研究，总结出来一套定制家具的方法，将控制点放到家具的板件上，实现多个订单同批次混合生产，改变了传统的按订单生产模式转为按零件模式生产。2007 年尚品宅配投入巨资打造的基于数字条形码管理的生产流程控制系统，“秒”级的加工控制评估，其思想和技术堪称世界前茅，为我国定制家具生产行业的示范企业和领跑者。

案例 3：服务大数据驱动跨界经营

如果说服务数据对生活资料的制造企业有价值，那么对于面向机构客户的重资产装备制造业企业有价值吗？答案是肯定的。下面以三一重工为例介绍服务大数据在其跨界转型过程中的体现的价值：

三一重工是国际工程机械装备制造的龙头企业，虽然其是工程机械制造商，但是这家企业在一开始就敏锐地意识到服务的重要性，将产品、研发和服务视为三一的三大核心竞争力，要求下属的员工将这三方面工作做到“无以复加”的地步。也就是在这一思路的指导下，三一从 2008 年开始组建智能研究院自主研发工程机械智能控制器件和传感器，并研究工程机械的状态数据采集整套解决方案。从 2011 年开始，三一联合清华大学软件学院研究面向工程机械工况数据管理的工业大数据平台和分析方法。

目前为止，三一实时接入的工程机械设备已达到 20 万台以上，高峰时段瞬时活跃的设备达到 1.2 万台，每日录入的工况数据达 1.2 亿条，累计车辆的位置数据、开工数据、报警数据达到千亿条以上，可谓名副其实的“大数据”。在这些数据的支持下，一旦主机发生故障，三一企业控制中心 ECC 的二线服务工程师可随时调集附近的服务车赶往现场，同时将主机开工数据发给维修人员，在业界实现了接单“两小时到底，24 小时完工”的服务承诺，达成了服务大数据的初步价值转化。

然而，在服务数据的利用方面，三一已经跨出狭义的维修服务界限，开始探索基于服务大数据的延伸应用：

深化数据分析，支持宏观决策。三一重工在平台建设过程中，积累了全国各地车辆的大量开工数据，这些开工数据将间接反映各地的基础设施建设和固定资产投资的情况，这些信息不仅可以帮助企业自身提前实现战略决策的优化，同时也可以对政府、科研和普通民众提供了解宏观经济形势的依据。当前，三一重工利用车辆开工热度指数作为房地产与基础设施建设的重要宏观经济指标，每月都按要求呈报国务院相关部门。

提供融资租赁服务，延伸经营范围。工程机械产品具有总量大，分布散，价值高的特点，无论是主机采购还是主机租赁均需要大量资金投入，这些特点催生了广阔的工程机械金

融服务市场。传统银行的信贷体系和业务运作方式无法实现对客户的有效评估，因此无法为工程机械用户提供优质的金融服务。而作为主机厂商的三一具备大量客户资源，同时又积累了海量开工数据，可以形成具有天然的行业特征的融资租赁的基础信用数据，在服务大数据的支持下对客户和设备进行信用评估作为融资的基础，并将业务拓展到保险等相关金融业务。

10.2.1 服务大数据的内容

透过上面的三个案例可以看到：当大家认真地审视服务的时候，会发现服务大数据和工业产品的全生命周期的各个阶段都有关系，那么如何理解和管理服务大数据的组成呢？本节试图站在企业的立场，从两个角度来讨论服务大数据这个命题：第一个角度，从产品全生命周期的角度来看待服务数据，重点讨论研究与服务有关的数据；第二个角度，则从发展的角度，站在当下看待这些数据过往的状态和未来的发展。

1）设计、制造环节数据

当今中国规模以上的制造业企业均在不同程度的开展了信息化实施工作，并且部署了一大批信息系统，包括：SCM、CAX、PLM、CAPP、ERP、MES、DCS、OA 以及财务等，由于大数据技术在普及之前这些系统就已经全面应用，因此其数据仍存放传统的关系数据库中，在本书前面描述的章节中具有介绍，此处不予赘述。对于产品生命周期的服务而言，这些初期数据虽然在容量相对较小，但是其价值密度极高，绝大部分都将与服务阶段产生关联，而如何建立服务阶段数据与设计制造阶段的数据仍是企业面临的重要挑战。以层次化的物料表（BOM）为核心，通过中性 BOM 进行关联是一种可行的方案，如图 10.1 所示。

中性 BOM 通过制造 BOM 关联模型，可以将生命初期设计与制造的产品需求、概念设计、制造工艺、包装运输等 15 大类业务数据正向传递到生命中期阶段；反之，中性 BOM 通过实例 BOM 关联模型和实例运行追溯模型，将生命中期服务保障的生命特征、运行状态、维修计划、服务评价等 14 大类数据反馈到生命初期阶段。该框架解决了不同生命周期阶段 BOM 结构失配所带来的双向信息集成难题，支持复杂装备全生命周期的闭环管理。

2）服务过程数据

进入产品生命中期之后，产品的维护和保修等服务则是产生服务数据的主要来源，这个过程中各个环节产生的数据将是所有服务数据的核心。对于原厂商而言产品服务不仅承担着与客户接触保证产品发挥价值的责任，同时还担负着将用户的需求和现场情况反馈回上游设计、工艺等环节改进质量的任务，因此需要综合考虑全生命周期上下游之间的关系。

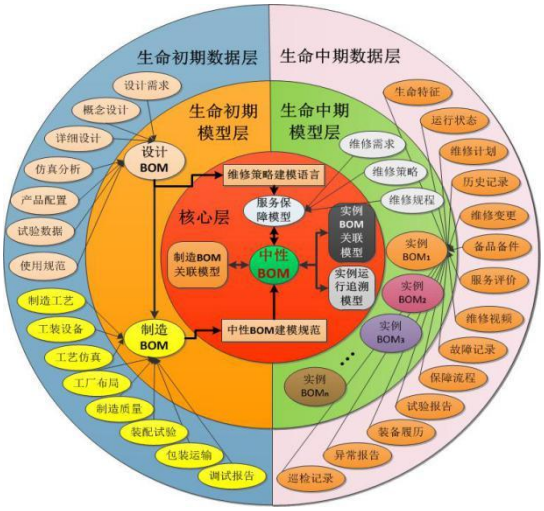

图 10.1 以中性 BOM 为核心的服务保障信息管理框架

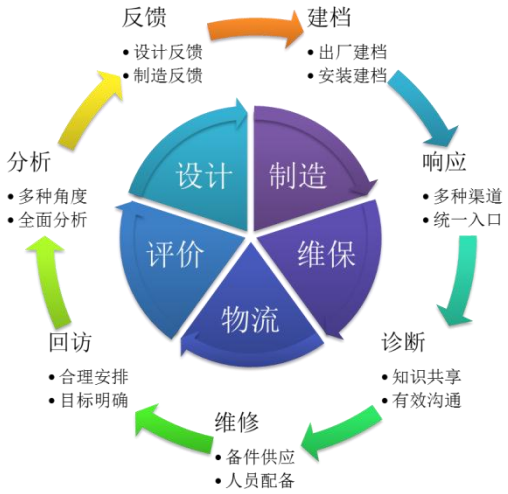

图 10.2 服务过程中的 7 个核心环节

围绕着产品整个服务流程可以划分为 7 个关键环节，如图 10.2 所示，首先，从建档环节开始，在产品投入使用前形成唯一的产品档案；产品投入使用后需要及时响应客户需求，获得客户的需求响应后需要判断产品的状态并且根据客户返回的服务请求对产品进行诊断，在确定产品的状态后确定维修计划并派出维修人员并保证备件等资源的供应，完成服务以后需要对服务情况进行有针对性的回访以便了解服务质量，更重要的是需要通过服务分析产品在设计、工艺、生产等环节存在的问题进而确定改进方案，最后将改进方案反馈到上游环节形成面向全生命周期环节的闭环管理。而不同的行业，不同的产品则可以通过对 7 项核心环

节分析的基础上建立服务过程数据。

3）产品运行数据

近年来，互联网和移动互联网技术的快速普及，大家的“连接”意识全面觉醒，在未来的几年里面，几乎我们目力所及的东西都可能会产生可上网的版本，而透过产品产生的数据聚集起来将成几何级数的增长，这些数据将成为服务大数据价值发掘的主要来源。其主要包括：

时序、时空数据：时序数据是随着机器自身传感器产生的数据，以时间序列和位置序列为主，此外随着产品自身的控制器越来越智能，其自身会产生大量与时间有关的控制变量数据。例如：温度、压力、震动、转速等数据。

非结构化数据：除了时序、时空数据之外，很多智能产品会加装摄像头、麦克风等传感器，其会产生大量的图片数据、音频数据、视频数据，此类数据统称为非结构化数据。例如：ATM 机上产生的取款人的照片等。

事件数据：在传感器获取得到数据的同时，产品在运行过程中也会产生大量以单个时间点为基础的事件数据。例如：云服务器硬件自身产生的开关机日志文件等。

4） 企业外部数据

上述服务阶段的数据均为企业自身可以控制，或者通过努力可以得到的数据，对企业而言其具有独占性，也是企业展开业务必不可少的数据，然而，服务大数据的魅力就在于其不仅与企业自身有关，而且也和跨出企业的外部的被服务对象密切相关，将这些数据纳入考虑的范围，并加以利用会产生数据间的化学反应。这些数据包括：

外部互联网数据：互联网数据无疑是所有外部数据中最有价值而又颇具争议的数据，前面描述的 361 度运动鞋的案例中，如果仅仅是鞋子采集到的传感器数据，那么数据自身可以发挥的价值很少，一旦和互联网地图数据、社交数据交叉融合之后即产生了大量的增值服务，这个案例其背后隐忧是：服务价值全部集中在百度的数据平台之上。所以在使用互联网外部数据时企业需要慎重考虑数据的所有权归属问题。

外部企业的数据：除去服务的过程数据之外，从整个产业链的角度来看，上下游企业也会收集自身数据，这些数据也可以成为服务大数据的重要组成部分，例如：前面描述的三一重工的案例中，上游供应商提供的数据是其制定服务政策的重要依据。在外部企业数据的利用上目前必须突破的技术难点在于数据的版权保护问题。

外部环境数据：气象、水文等数据和宏观政策的数据，这些数据对提高服务的品质至关重要。例如对于风机的发电量而言，气象数据的准确性就是非常重要的，尤其是大型装备在服务的过程中受到外部环境影响也很明显。

10.2.2 如何利用服务大数据

一切创新的源头源自观念和意识的变化，长期以来，中国制造业的关注点放在制造本身，实施粗放的管理，导致的结果是数据是碎片化的没有建立联系。对于服务阶段数据一般没有认真的考虑。

因此，首先需要树立对数据积累和开放的意识，所谓积累意识，是通过积累数据建立起壁垒，形成独特的竞争优势，在上文中介绍的尚品宅配的案例就是通过积累服务数据最终为他们改变制造模式的典型案例。而数据的开放，也不是说将辛苦得到的数据一股脑让别人拷走就是开放了，而是要将数据的价值与其他的数据勾连在一起考虑，达到左右逢源发现价值获得增量业务的思路，从而摆脱一味的在存量业务上不断重复的局限性。

下面介绍挖掘服务大数据的三条思路：

思路 1：从服务自身挖掘数据价值

当没有任何数据基础的时候，可从服务数据自身挖掘数据价值，可以从用户角度、产品角度、地域角度、时间季节角度、维护体系角度以及供应商角度来建立指标体系进行数据分析，具体的分析目标如图 10.3 所示。利用现有的数据仓库技术即可完成上述分析，获得的分析结果可以优化服务体系，改善配件供应，识别优质客户、制定个性化服务政策等。

图 10.3 多维度挖掘服务数据价值

思路 2：从全生命周期挖掘数据价值

第二条思路是从产品全生命周期的角度挖掘数据价值，在服务数据积累的同时还可以考虑，从产品的销售、设计、工艺、制造以及再制造环节对服务数据加以利用形成闭环反馈。在上文中提到的案例中，三一重工服务部通过定期分析由于产品故障引起的服务订单，找到关键部位的重大品质问题，然后质保部门配合制造、研发等部门建立重大品质问题小组，提出指定改进策略，最后反馈服务的整个过程就是典型例子，如图 10.4 所示。

图 10.4 数据驱动的重大品质问题闭环管理

思路 3：从产业生态价值链挖掘数据价值

第三条思路，是从价值链的角度来考虑利用服务数据，因为工业产品在投入运行的过程中一定会参与到一个新的制造循环，那么，在这个循环的过程中产品会产出新的产成品，也会消耗即有的生产资料（如：配件、燃料、原材料等），将这些消耗来源和产出目的作为研究对象时就可以发现服务数据的价值，尤其是数据量达到一定程度的规模的时候，数据对价值链下游的实体的价值会通过其他的形式体现出来。例如：当所有的搜索偏好数据被收集到百度的大数据平台上时，它的对产品制造商价值就会通过广告体现出来。

今天，传统企业在讨论互联网+、转移支付、跨界经营这些热门话题的时候总觉得“瞧不明白”，实际上如果从数据的角度来看这些现象的时候，道理就会变得非常容易理解：只要将数据聚合到自己手里形成规模，价值就会从价值链上下游企业身上发掘出来，然后通过合适的渠道进行转化就可以形成诸如“羊毛出在猪身上，狗来付钱”的机会。

10.2.3 服务大数据利用过程中的挑战

通过前面的论述，服务大数据对于工业而言的价值已经非常明显，但是，冰冻三尺非一日之寒，要很好的利用服务大数据，企业还面临非常巨大的挑战。下面仅从数据的角度，结

合在调研过程中发现的问题，介绍服务大数据利用过程中可能会遇到的挑战，以帮助企业明确发展服务大数据的努力方向，提前做好准备。

打通数据间的联系的挑战

首先，利用服务大数据并不是服务阶段孤立，当我们在考虑服务大数据时，往往需要将用户数据、物料数据、工艺数据和产品数据联系起来，才能发现价值。然而，这些数据在今天的企业里面仍然独立的分布在各自的信息系统之中，互相之间甚至连编码都未能建立联系，如何将这些数据之间的联系打通将是利用服务大数据过程中面临的第一个挑战。面对这个挑战，企业可以从两个中心入手进行数据梳理：一个是以用户为中心的数据梳理，即将售前和售后的用户数据联系，形成统一的用户档案。另一个则是以产品为中心的数据梳理，将产品出厂前和出厂后的数据联系起来，形成可追溯的产品档案。

建设数据采集体系的挑战

数据采集的渠道建设的挑战，如前文所述产品产生的数据是服务阶段最有价值的数据，而且又是体量最大的数据。为了将这些数据采集集中起来，企业必须想方设法建立起一套合理的数据收集渠道，由于设备生产方与设备使用方之间对数据产权界定不明确，数据传输通信限制等原因，数据采集渠道的建设充满挑战。同时，如果采集的数据质量无法保证，那么服务质量必然大打折扣。然而，对于服务大数据而言数据质量问题仍普遍存在。因此，在后续开展服务大数据的应用过程中，需要建立质量管控体系，并且将质量问题消除在萌芽阶段。目前对于企业而言仍然缺乏行之有效的手段和方法。

第11章　流程工业大数据

流程工业是所有涉及物质转化过程的工业领域，是能源、材料、工业品和日用品的制造工业，如炼油、化工、冶金、建材工业等。我国早已成为这些领域的世界制造业大国，如钢铁产量占世界 50%，炼油能力占世界 10%，水泥产量占世界 46%等等。

流程工业是指被加工对像不间断地通过生产设备，通过一系列的加工装置使原材料进行化学或物理变化，最终得到产品。由于流程制造中物料的变动性强，工艺流程的制约变量多，造成了其在生产、物流管理上与离散行业的显著差异。流程工业大数据主要来源于四个层次，即仪表、调节阀等生产过程（0 层），生产过程控制 PCS（1 层），生产执行系统 MES（2 层）和企业资源规划 ERP（3 层）。下文从两个最典型的流程工业：钢铁与石化行业分别介绍其工业大数据的应用。

11.1 钢铁行业大数据

11.1.1 行业发展现状总述

钢铁工业作为国民经济的重要基础产业，支撑了国民经济和下游用钢行业的快速发展。钢铁工业自身的快速发展取得了举世瞩目的成就，我国钢产量占世界钢产量的一半以上，钢材品种、质量、性能不断提高，节能环保水平不断提升。

但随着钢铁工业的快速发展而出现的一些问题也日益突出，如产能过剩、行业利润率不高、环保压力大。钢铁工业要满足可持续协调发展的要求，必须在两化深度融合的基础上，充分发挥大数据技术作用，加快实现智能化、绿色化制造进程。

钢铁企业信息系统由基础自动化、过程控制系统、制造执行系统、企业计划管理系统、决策分析与商业智能系统等多层次构成。以数字控制、现场总线、工业以以太网结合的网络在钢铁工业已经普及，无线通信应用逐步推广，钢铁生产的各道工序已经普遍采用基于 PLC、DCS、工业 PC 实现数字化控制。常规检测仪表的配备比较齐全，基于数学模型的计算机过程控制覆盖采选、炼铁、炼钢、轧钢等主要工艺过程。MES（制造执行系统）在重点钢铁企业已基本普及，近百家钢铁企业建立能源中心。ERP（企业资源计划）、CRM（客户关系管理）和 SCM（供应链管理）等取得成功应用，在更好地满足客户需求、精细控制生产成本等方面发挥了作用。一些重点企业在聚集了海量的企业生产经营管理信息资源的基础上，建立了数据仓库、联机数据分析、决策支持和预测预警系统，着手进行数据挖掘、商业智能等深度开发。

11.1.2 大数据在钢铁行业向智能化转型过程中的作用

钢铁行业，特别是在升级转型过程中的大数据作用可以分为以下四个方面：

- 提高钢铁工业市场适应能力
  - 通过钢铁工业供应链全局优化，提升对市场发展趋势的分析预判能力，更好适应市场变化；
  - 通过制造执行产品质量一贯制管理和工艺闭环控制，提升产品质量，实现批量个性化制造，避免同质化竞争。
- 提升钢铁企业盈利能力
  - 通过供应链协调优化，提升对原燃料供应的控制能力，降低物流成本；
  - 通过生产计划与制造执行一体化，缩短产品交货期，降低库存成本；
  - 通过可循环流程设计和物质流能量流优化，保障流程动态有序运行，降低原料、能源成本；
  - 通过智能冶炼系统和智能轧制系统，提升生产效率和能源效率，节能降耗。
- 满足国家环保要求
  - 通过物质流能量流优化、智能冶炼系统和智能轧制系统，减少能耗和污染物源头排放；
  - 通过全流程污染物排放在线监测，加大对钢铁企业环保监管力度。
- 推动钢铁工业管理模式改变
  - 推动电子商务和供应链协同发展，支撑实时精细管理模式，促进企业知识管理。

11.1.3 行业大数据应用现状

钢铁企业信息化系统为大数据应用提供了较好的数据基础。国内很多钢铁企业都开始了大数据应用工作，目前钢铁企业大数据应用主要围绕两个层次展开，一是搭建大数据信息采集、整合、储存和智能化分析等平台，二是围绕钢铁企业产品质量分析、供应链管理、精细化能源管理、产品研发和设计等开展大数据应用。

应用案例1：宝钢股份工序一贯质量分析应用系统

钢材生产需要经过焦化、烧结、炼铁、炼钢、精炼、连铸、热轧、冷轧、涂镀、包装等多道工序，钢水、钢坯、钢材长度、厚度、缺陷位置以及排序等信息也随之发生多次变化。如果每一工序质量都只采用一个简单的平均值来表征，发生质量追溯时，不仅过程漫长，而且很难实现对缺陷产生位置和原因的精准定位，严重影响到快速响应用户需求的服务质量。为此，宝钢股份坚持以用户为中心，自主研发“工序一贯质量分析应用系统”，将现场产线表面缺陷检测仪、多功能缺陷检测仪等设备采集到的实时数据，通过信息技术在线应用和离

线分析，实现产品全流程表面质量智能监视、跟踪和追溯。

该项目一期上线覆盖热轧及后道工序的 13 条生产线，实现了影响产品缺陷绝对位置、相对位置变化数据采集及传递，关键质量高频数据的采集及传递，并运用智能化技术，将采集到的工序中关键的实时、有效数据经过精确计算，第一时间传输给制造管理部，为其对下道工序作出快速、准确的生产制造指令提供依据，这为提升产品表面质量和消缺起到积极作用。目前，该项目已全面覆盖热镀锌家电板产线，可实时记录工序数据，并增加质量异常情况下的声、光实时报警装置，有利于现场及时采取补救措施，有效降低工序缺陷产品的产生。2015 年成功上线运行半年来，现场缺陷的检出和分析效率明显提高，产品质量缺陷得到有效控制，用户满意度逐步提升，并为宝钢股份“智能制造”大数据收集和贯通奠定了基础。

应用案例 2：沙钢基于大数据的能源优化

能源成本占钢铁生产成本的 30%左右，钢铁企业非常重视节能工作，国内近 100 家钢铁企业建立了能源中心，实现了能源远程监控、集中调配，以及能源计划、能源质量、能源设备、成本综合管理等功能。沙钢应用大数据技术，在现有能源中心基础上进行升级改造，实现了能源分时管理，取得了显著的经济效益。

能源中心升级改造采用了冶金自动化研究设计院基于钢铁产耗信息模型 NMP 技术的大数据平台，以流的方式建立企业产耗关系模型，用属性对原子化的数据进行再组织。将数据与应用解耦，将数据在统一的平台进行采集、处理、存档。大数据平台能够表达企业的组织结构、能源结构、产供关系，组织计量、质量、计划、指标等数据，对多时间尺度（秒、小时、班、日、月）的海量数据进行高效处理和动态属性计算，利用原始数据计算出各种指标（峰平谷占比、单耗、单位成本、工序能耗等）。动态表格技术可以配置出各种报表，以满足各种用户的数据查询和分析。面向公司领导、业务部门、总厂、车间 提供班、日、月及单耗报表，使用 60 张模板，生成 400 张报表。

此次改造，沙钢在大数据平台的支撑下，利用峰、平、谷电价差，通过精细化能源管理产生了上亿元的经济效益。包括煤气柜移谷填峰发电，通过煤气柜谷、平段储气，峰段发电产生利润；间峰平谷用电管理，根据每小时用电量统计出尖峰平谷各时段用电量、各时段用电量占比、车间平均电价，利用峰平谷电价差，要求车间避峰生产；车间分时补水，根据每小时水泵补水量，统计出车间尖峰平谷各时段补水量、各时段补水量占比；原料车间皮带分时运行，根据皮带机启停信号，统计出每条皮带尖峰平谷各时段运行时间、各时段用电量占比，利用峰平谷电价差，实现原料烧结 22 条皮带避峰运行。

11.1.4 行业大数据发展规划

工信部在2015年发布的《原材料工业两化深度融合推进计划（2015-2018年）》明确指出要促进工业大数据集成应用：支持冶金工业大数据平台建设，促进信息共享和数据开放，加强行业经济运行监测，推动大数据在钢铁等企业经营决策中的应用，实现产品、市场和效益的动态监控、预测预警，提高行业管理水平和企业决策科学水平。鼓励骨干企业在工业生产经营过程中应用商业智能系统（BI）和产品生命周期管理（PLM），提升生产制造、产品研发、供应链管理、营销及服务环节的资源优化配置能力和智能决策水平。

围绕钢铁工业强国战略，推动钢铁产业智能化、绿色化可持续发展，大数据技术是关键使能技术，今后需要充分发挥大数据作用，开展以下工作：

（1）建设钢铁工业大数据平台。

数据是钢铁工业数字化、智能化制造的基础，这里的数据是广义数据，既有来自企业现场检测仪表、RFID、质量分析仪表、过程控制系统的各种连续变量，也有声音图像信号、现场离散事件记录、物流能流空间信息（GPS）、调度操作指令等非结构化数据，还包括设备规格、设计图纸、产品规格、工艺规程、电子商务等文档型资料。需要整合现场传感网、物联网、工业控制以太网、内部外部互联网、社会无线通信网，构成钢铁工业数字化智能化制造的工业互联网，实现数据在不同业务间的互操作集成和共享。在此基础上，实现多源数据融合，包括物联网和工业互联网构建、不同业务数据互操作集成、多源大数据融合分析。进而实现数据智能分析处理，包括多业务数据仓库、多源数据可视化、数据挖掘和知识发现等。

（2）钢铁供应链全局优化

钢铁供应链全局优化是面向钢厂原燃料采购及运输、钢材生产加工、产品销售及物流等供应链全过程，综合应用现代传感技术、网络技术、自动化技术、智能化技术和管理技术等先进技术，实现优化资源配置、动态响应市场变化、整体效益最大化。包括：

1）优化上游资源选择与配料：跟踪原料市场变化，预测分析市场趋势，围绕供应链最终产品，优化原料选择和运输。强调原料的优化配置和综合利用。

2）加强与下游客户供应链深度协同：建立电子商务和供应链协同信息EDI规范，迅速响应客户需求，及时提供合格产品，减少库存、中间环节和储运费用。

3）生产计划与制造执行一体化协同：订单产品规格自动匹配，前后工序协调一致，后

一工序及时获取前一工序的生产数据并按照生产指令进行最优生产。

4）全供应链物流跟踪：覆盖原燃料、在制品、产品、废弃物资源化利用的物流跟踪，通过准确、直观地反映物流资源分布动态、计划执行情况和库存变化趋势，为优化资源调配提供依据。

（3）钢铁制造精细化管理

钢铁智能工厂是面向钢铁生产的运行环节，综合应用现代传感技术、网络技术、自动化技术、智能化技术和管理技术等先进技术，并与现有钢铁生产过程的工艺和设备运行技术高度集成的新型现代化钢铁工厂，通过企业资源计划管理层、生产执行管理层和过程控制层互联，实现物质流、能源流和信息流的三流合一，实现钢铁企业安稳运行、节能减排、降本增效、优化生产、质量升级等业务目标。包括：

1）生产管控：实现对综合生产指标→全流程的运行指标→过程运行控制指标→控制系统设定值过程的自适应的分解与调整，满足市场需求和生产工况的频繁变化，提升生产管控的协同优化能力，实现生产计划的闭环管理和持续优化。

2）能源管控：通过能量流的全流程、多能源介质综合动态调控，形成能源生产、余热余能回收利用和能源使用全局优化模式，实现能源利用的精细规范。

3）环境监控：建立全流程污染源排放在线实时监测系统，实时采集相关信息，并进行趋势分析判断，确保生产满足国家环保要求。

4）设备管控：实现设备的全面监控与故障诊断，通过预测维护降低运营成本，实现资产的全生命周期管理。

5）工艺和质量管控：实现工艺、质量标准的科学决策，实现有效的技术管理，提高工艺管理水平，提高产品质量。

6）物质流与能源流协同优化：研究钢铁生产物质流与能量流的特征和信息模型，分析物质流与能源流动态涨落和相互耦合影响。综合考虑效率最大化、耗散最小化、环境友好性，实现多目标协同优化。

（4）钢铁生产过程智能控制

钢铁生产制造全流程是由多个生产过程有机连接而成的，其具有多变量、变量类型混杂、变量之间强非线性强耦合的特点，受到原料成分、运行工况、设备状态等多种不确定因素的

干扰，其特性随生产条件变化而变化。实现上述的发展要求，面向钢铁复杂生产过程智能控制系统的关键需求主要包括以下几方面：

1）智能冶炼系统。实现冶炼工位闭环控制，包括机理和数据混合模型，工艺设定点实时优化，钢水质量自动闭环控制。实现冶炼工序协调优化控制，包括冶炼工序集成协调优化模型，各工序设定点动态协调优化。

2）智能轧钢系统。实现轧制工位闭环控制，包括产品性能预报模型，工艺设定点实时优化，钢材质量自动闭环控制。实现轧钢工序协调优化控制，包括控轧控冷模型，轧制工序动态协调优化，高端产品质量自动闭环控制。

11.2 石化行业大数据

石油化学工业是国民经济的重要的支柱性行业，工业总产值在 2010 年超过美国，跃居世界第一，三大合成材料、轮胎、氮肥、磷肥、纯碱、烧碱、硫酸、甲醇、电石等产品产量也占据世界第一的位置。乙烯、原油加工等位居世界第二。2014 年，全行业销售收入 14.06 万亿元，利润为 7911.1 亿元，但是高附加值产品较少，利润率较低，仅为 5.6%，企业效益下滑。一方面大宗原材料产能严重过剩，市场持续低迷，另一方面每年仍需要进口高附加值化工产品 4313 亿美元来满足国内市场的需要。伴随着石化行业的高速增长，石化行业的重大安全生产事故给人民生命安全、环境保护甚至社会稳定造成的负面影响也越来越大，全行业的绿色发展、安全发展距离国家和人民的期望仍存在较大差距。因此，如何利用大数据技术促进全行业的创新发展和转型升级具有重要的战略意义。

以石化行业为代表的流程行业产品加工持续进行，并伴随着复杂的物理和化学变化，同时，产品制造的工艺路线相对固定，但工艺复杂多变，且生产自动化水平较高，生产控制技术要求实时性。随着石化行业生产装置向大型化、复杂化发展，工厂的采集点越来越多、采集频率越来越高，生产运行、监测、管理和优化产生的数据呈指数级增长，生产和运营的智能化需求也不断提升。因此，数据的真正价值在于其为石化产业链提供有价值的服务，提升产品的附加值和风险防控能力。数据是资产、数据是金矿、数据是财富的观念已经深入石化行业。

11.2.1 石化行业大数据发展现状

石化业务数据大致分为 3 个过程域：一是从采购、制造、销售到配送的企业供应链全流程数据，如石化企业包括从原油资源采购、原油配送（管道、船运、车运等）、炼油加工、化工生产、成品油配送（一次配送、二次配送）、化工产品配送及销售的供应链全过程数据；

二是从设计到运营的工厂全生命周期数据，如石化整个工程项目设计（可行性研究、初步设计、详细设计）、工程（建设准备、建设实施、中间交工、单试联试、试生产、竣工验收）、运营（交付使用、运营、改造维修）涉及的各类数据；三是从经营管理、生产管理到自动化控制的企业管控一体化数据，以 ERP/MES/DCS 三层结构的企业信息系统为核心，覆盖工厂中多个专业管理领域，包括物流、能流、工艺、质量、设备、安全环保、成本等专业化数据。

石化行业大数据具有以下特点：

1. 数据体量大。区别于离散制造行业，石化行业大数据是持续实时产生的数据，产生于生产运营和自动化的各个环节，数据量以几何级增长的速度在增长。如石化行业中，生产监控及应用中产生的过程数据量巨大（如腐蚀监控、设备监控、环境监控、火灾监控、DCS、实时数据库等）。以一个典型的炼化一体化企业为例，拥有 30000 个采样点，现场采样率达到 100 次/秒，每年约产生 495TB 数据（495TB = 12 字节/次×100 次/s×30000 点×86400s/天×365 天），数据量达到 PB 级。

2. 数据类型多。石化行业生产过程数据种类众多，包括实时数据、历史数据、工程和技术文档、工业视频、音频、图像、装置和设备三维模型、GIS 等各类结构化、半结构化数据以及非结构化数据。流程企业中 80% 的数据都是非结构化数据，特别是工业技术文档（安全评价报告、环境影响评价报告、工艺规程、操作规程、应急预案、事故分析报告、工程设计图纸等）、工业视频数据（危险操作点监测视频等）、检测图像（裂解炉温度红外检测图像、环境指标探测图像）、时间序列数据（化工装置传感器历史记录数据）等，这些数据每年都按指数增长。

3. 数据处理的时效性强。石化行业生产运行是以工业现场大量的实时传感数据为基础的，处理时限要求高。各类传感终端产生实时、连续的事件流，数据流处理系统必须快速对其进行响应并即时输出结果。针对时效性场合，一般可以分为实时、准实时、离线三类。石化行业工业现场操作和控制类一般是实时要求，是秒级的响应；生产调度内原料切换、异常工况等处理一般是准实时要求，是分钟级的响应；经营决策一般是离线要求，是小时级或天级的响应。

4. 显性和隐性知识混杂。石化行业中大数据应用的核心是知识库，其包含两类知识。一类是能用启发式规则、数学模型等表达出来显性知识，如利用石化的工程原理（动力学模型、质量守恒、能量守恒、物理化学反应等）建立的机理模型；另一类是海量的现场传感数据中蕴涵着大量宝贵的信息，包含了丰富的反映运行规律和运行参数之间关系的隐性知识

（专家经验）。

11.2.2 石化行业大数据应用现状

1. 国外石化行业大数据应用现状

国外石化企业正与 IT 公司一起，将众多技术应用于战略决策、科技研发、生产经营和安全环保等领域，从大数据资源中不断挖掘更多的财富和价值。

据埃森哲和微软对 200 位业界专家的调查，国际大石油公司中 75%的投入与信息技术领域有关，各大企业广泛建设生产指挥中心，对生产信息、设备运行、能源消耗、原料和产品市场变化等内容进行全面分析，并基于智能化的数据挖掘和预测模型进行优化决策。

在壳牌上游业务中，壳牌通过对地理信息等数据的实时采集、分析，从而提高油井开采的成功率；在油罐管理上，壳牌使用实时数据分析，从而减少潜在事故数量，及早发现问题，大幅减少泄漏事件发生；同时，可减少误报所引起的损失。该项技术带给每个油站平均每年 4000 美元的成本节约。在下游燃油、润滑油销售上，壳牌通过移动设备、车联网的客户数据分析向客户推送定制的服务消息。同时，壳牌与阿里巴巴合作，对网上交易及社交数据的分析，精确定位潜在客户，实现高达 70%的客户转化率。另一方面，壳牌与银联的合作，通过对相关银行卡交易数据进行分析，从而明晰壳牌的市场份额。

英国石油公司将数据分析应用于设备预测，根据经验模型在故障出现明显迹象前进行预测分析，在发现零件出错后进行振动监测，可监测到 90%设备故障。在采油厂安装无线感应器，通过全网式的数据采集，发现有些种类的原油比其他种类更有腐蚀性，这个发现可以在设备和管线的使用上加强防范，使生产更安全。

2. 国内石化行业大数据应用现状

随着经济的发展和不断改革，我国石化企业面临着巨大的压力，其行业结构和当前的经济发展速度相矛盾，在石化生产中由于生产工艺和装备落后、资金结构布置不合理，对于安全环保的重视程度不够使得石化生产难以满足现代的生产要求。因此，利用大数据技术，在战略决策、科技研发、生产经营和安全环保等领域进行创新和变革，支撑和引领公司转型发展将成为一个重要的技术支撑手段。

中国石化茂名石化将大数据分析技术引入催化重整装置，对关键指标进行参数相关性分析，据此调整产品分布，使汽油收率明显提高；对原料性质参数建模分析，实时预测关键指标，及时发现问题，快速做出处理决策；对关键报警点的根原因进行分析，寻找问题诱因，

对症调整生产因素；同时，健全了茂名石化的智能生产平台，有助于提高企业经济效益。

中国石化九江石化、石化盈科和清华大学化工系协同合作，应用大数据技术对催化裂化装置进行了报警合理化和预警辅助决策，通过对关键报警位点根原因分析，构建基于人工免疫系统的异常工况预警模型，在关键报警发生前数分钟提前向生产一线的操作人员发出预警信息，对该异常工况进行根原因诊断，并提供排除异常工况的建议措施，对于预防重大事故或非计划停车起到关键作用。以再生滑阀降压为例，利用大数据发现了仅凭以往专家经验发现不了的根原因“待生斜管滑阀阀位”，并经过工业验证，证明分析结果正确。其次，在 DCS 系统报警发出前实现了提前预警，预警模型依据“回炼油流量”的变化，提前 2 分钟诊断出了催化装置“二再密相温度”的异常，为技术人员及时采取措施争取到了宝贵时间。

中国石化金陵石化通过技术和操作培训，将大数据用于生产实践，实现了焦化装置的生产优化；通过对比数据，增加了装置处理量，降低了混合汽油的硫含量，让更多汽油进入苏Ⅴ调和系统，达到了增产汽油的效果。

中国石化青岛安全工程研究院加强大数据应用的研究，在石化生产全生命周期的风险评估技术、异常工况监测预警技术及过程安全管理评估技术等研究及应用方面承担了多项国家级课题，针对石化装置开发了异常工况监测预警、泄漏监测智能预警、防爆 EXPAD 及移动应用平台等一系列技术和相关产品。

西安陕鼓智能信息科技有限公司把远程在线监测及故障诊断系统升级为远程工业智能服务平台，将硝酸、煤化工企业的动设备、静设备、仪表、备件等参数、振动、工艺信号等数据全部纳入，应用大数据关联分析技术，预测检修、状态检修，保证不发生事故、少发生事故，提高装置的在线率。

大数据在一个行业上应用能促进生产优化、效益提升，但是很多数据涉及企业的生产经营机密，很多企业不愿意公开或者交流，特别是安全评价、环境影响评价、安全事故的数据更不愿意提供。在数据处理上，由于大数据中大部分都是非结构化的数据，现有的软件和工具主要适用于以结构化数据为主的传统数据，要想及时捕捉、存储、聚合和管理这些大数据，以及对数据进行深度分析和挖掘，需要新的技术和能力。而我国数据存储、处理技术基础薄弱，总体上以跟随为主，难以满足大数据大规模应用的需求。因而石化行业在大数据领域的应用，要如何更好地开展数据搜集、数据存储、数据处理，让“沉睡”的数据创造价值，这是当前我国石化企业应用大数据的挑战。

11.2.3 大数据在石化行业向智能化转型过程中的作用

石化生产过程是复杂的物理化学变化过程，影响过程的因素多，机理十分复杂，难以用线性手段描述和精确方程表达，但是随着传感器等技术的发展，数据采集变得越来越容易。虽然传统石化行业有重视数据的传统，例如，在企业的控制、执行和管理三个层面上分别设置数据收集、储存和处理的机制和相应设备，许多企业正在或已经建立了 LIMS、MES 和 ERP 系统。但是，目前数据处理和应用多数情况下限于统计和查询，数据中蕴含的潜在价值远远没有被挖掘出来。随着数据量的不断增长，数据的大容量储存、管理和数据安全性问题，以及随着石化行业在更大范围的布局而产生的分散分布与数据集成之间的矛盾等日益凸显，如何从各种类型的数据中采用新处理模式快速获得有价值的信息，从而实现深度理解、快速洞察与精准决策等，是石化行业智能化发展面临的新挑战。大数据对石化行业智能化转型过程产生的影响可以总结为：

1. 在石化行业生产管理层，很多设备都安装了一个或多个微处理器采集生产数据。这些无处不在的传感器和微处理器，形成了极为庞大的数据来源，常规的数据库技术已难以完成捕捉、存储、管理和分析这种大规模的数据集合。而利用大数据技术，则能清晰而有逻辑地对这些数据进行有目的的分析，还可以从中发现某些异常或事故，对安全生产具有重大意义。大数据技术实现石化生产的全流程可视化，给石化行业带来的益处包括优化生产流程、降低成本、提高运营效率以及增强生产安全等。

2. 在石化行业分析决策层，大数据技术的战略意义不仅在于掌握庞大的数据信息，更在于对数据的“加工能力”——对大量的数据进行专业化的处理，使之转化成为对企业分析决策有用的信息。石化企业如果能够在工业环境中建立起大数据平台，提高工厂对不同设备收集海量信息的梳理能力，提高企业信息系统的计算能力和数据消化能力，实现对企业的产品数据、运营数据、销售数据、客户数据的实时而有针对性的分析，并用其指导下一轮的研发、生产、销售和服务等产业链的各个环节。这将会使得企业能够在低成本运营的同时，有效实现按需生产，智能化分析决策，从而实现向智能化转型的目的。

3. 对石化行业经营管理层，从大数据中发现有价值的结果，使石化行业突破传统的思维，产生更多的增值服务，催生新的管理模式，拓展新的业务领域，转变原有的发展方式，引发技术创新和产业革新。首先是引发了企业生产方式的变革，柔性化生产方式正在悄然取代原来的流水线生产模式；其次是引发了企业合作方式的变革，中介正在被大数据管理取代；再次是引发了商业模式的变革，制造业服务化、生产设备租赁化、个性定制化正在成为服务的主流。

11.2.4 石化行业大数据发展规划

大数据时代，石化企业应实现各业务层面数据共享，以“数据驱动”为核心，围绕石化产业链推动各领域大数据应用，形成融合创新的强大推动力，为推动大数据在石化企业研发设计、生产制造、经营管理、市场营销、售后服务等产品全生命周期、产业链全流程各环节的应用在以下 5 个方面进行规划：

1. 上游生产大数据应用：重点聚焦于勘探开发综合研究、油气生产运行各个环节的成本优化、运行效率监测、安全管控等。主要建设领域包括：勘探开发分析、油井完整性分析、运行效率分析、生产过程监测、设备预测性检修等。例如，利用大数据技术进行油气藏动态监控，通过实时的动态指标检测，集成原油的动态监测系统，建立全面的油气藏（油藏-井组-单井）监控指标体系和算法，自动计算指标数据反映油气藏实际开发状况，实现开发动态的全方位掌控。

2. 生产营运大数据应用：重点聚焦于跨板块协同及生产运行环节的动态监测/控、预测、平衡和优化等。主要建设领域包括：原油需求预测、采购成本优化、生产运行状态优化、供应链风险预测、供应链优化等。例如：利用市场数据、经营管理数据和生产实时监测数据，可以对供应链的各个生产环节或协同环节进行优化，支撑实时、有效的分析决策。

3. 炼化生产大数据应用：重点聚焦于新材料新产品的研发以及工艺流程、生产运行、设备维修、HSE 管理等环节的优化、监测和分析等。主要建设领域包括：全流程 HSE 风险预测、设备预测性维修、产品研发分析、生产运行状态优化、需求预测分析等。例如，在设备预测性维修方面，通过实施采集各地炼厂运营及资产数据，利用预知性算法进行大数据分析，提前发现资产漏洞，发出预警并更新维护计划，以避免事故发生，提升石化行业安全生产的风险防控能力。

4. 经营管理大数据应用：重点聚焦于提升传统企业级 BI 提供的决策支持分析能力，强化预测性、趋势性、实时性的分析应用等。主要建设领域包括：企业绩效分析、经营风险分析、投资分析、项目分析、财务分析等。例如，利用大数据分析技术，采集结构化和非结构化数据，然后进行预知性分析和关联，发现潜在的违规行为，从而提高风险预警水平。

5. 销售服务大数据应用：重点聚焦于对站级销售管理、营销活动的分析，以及对客户行为和偏好的洞察分析，以提供更加个性化、精准的营销服务，同时规避风险等。主要的建设领域包括：油品损益分析、站级销售分析、营销活动分析、定时定价分析、客户细节分析等。例如，利用大数据技术采集客户数据，建立预测模型，预测哪些客户流失风险高，哪些

客户值得挽留，客户对哪些营销产品感兴趣等，进而把市场需求信息及时反馈到产品的规划、研发、设计和生产加工等化学品生命周期的前端环节，实现客户化定制研发和敏捷制造，在提高附加值的同时，有助于避免全行业的产能过剩，促进节能减排降耗，进而推动传统石油化工行业尽快实现转型升级。

第12章　结论与展望

新兴信息技术正在深刻变革传统制造业，“以机械为核心的工业”正在向“以软件为核心的工业”转变，数字化、网络化、智能化是两化深度融合的主线。因此，工业大数据已成为智能制造的基础原料，是提升工业生产力、竞争力、创新力的关键要素。

世界两百年工业发展规律和中国经济社会发展进程决定了，处于制造业微笑曲线两端的“升级”和“转型”，是中国制造由大到强的必由之路。我国中低端产品制造产能过剩，而高端产品制造能力短缺，“升级”意味着产品创新，传统机电装备必须进化为智慧联网装备，成为新型“网络终端”。“十二五末”我国第三产业已经超过第二产业，“转型”意味着制造业的服务创新，以智慧联网装备为载体，发展第一、第二和第三产业相互融合的“互联网+”产业生态。

中国是装备使用大国，在船舶、飞机、机车等领域，有着其他国家无法比拟的使用数量，如果能从使用端投入分析力量，则不失为反向突破中国制造的有效途径。拥有大数据不是目的，发掘其价值才是关键。由企业信息化数据、装备物联网数据和外部互联网数据汇聚而成的工业大数据，蕴藏着巨大价值。例如，通过分析用户使用数据改进产品，通过分析现场测量数据提高工件加工水平，通过工况数据进行产品健康管理等等。

加快建设制造强国，实施《中国制造 2025》，是《中共中央十三五规划建议》中构建产业新体系的重要内容，是我国全面建成小康社会的重要举措。因此，我们必须抓住新兴信息技术驱动的新一轮工业革命给我国制造业跨越发展带来的历史机遇，以工业大数据为抓手进一步推动“两化融合，四化并举”方针的实现。

同时我们必须看到，工业大数据是一个正在发展的学科领域，其内涵外延、模型理论、技术方法及其实施策略等还有待发展与创新。唯有结合中国国情认真实践，才能走出中国工业大数据自主之路，实现制造强国的战略目标。

第三篇 中国大数据发展趋势与建议

自 2012 年 10 月中国计算机学会大数据专家委员会成立以来，在每年 12 月的大数据技术大会上都会发布对第二年大数据发展趋势的预测。从 2013 年到 2016 年，已经是第四次年度预测了。每次预测都是基于大数据专家委委员的观点收集、整理、投票、汇总以及解读，最终形成年度预测，此预测是专家委员会群体智慧的结晶。在 2015 年和 2016 年的两次预测中，还邀请了中关村大数据产业联盟的成员参加投票，也部分反映了产业联盟的趋势判断。2015 年底做出的 2016 年预测，参加投票的大数据专家委委员和产业联盟成员共 116 位。根据这 116 位专家投票结果，汇总形成如下十大趋势预测。

第13章 大数据技术发展趋势

13.1 2016 年大数据发展十大趋势

13.1.1 趋势一：可视化推动大数据平民化

“可视化”虽然已是连续第三次入选大数据发展十大趋势，但今年能占据第一位，实在是意料之外的意料之中。这几年，大数据这一概念迅速深入人心，大众直接看到的大数据更多是以可视化的方式体现。可视化实际上已经极大拉近了大数据和普通民众的距离，即使对 IT 技术不了解的普通民众和非技术专业的常规决策者也能够更好地理解大数据及其分析的效果和价值，从而可以从国计、民生两方面都充分发挥大数据的价值。

可视化是通过把复杂的数据转化为可以交互的图形，帮助用户更好地理解分析数据对象，发现、洞察其内在规律。数据是人类对于客观事物的抽象。人类对于数据的理解和掌握是需要经过学习训练才能达到的。理解更为复杂的数据，必须要越过更高的认知壁垒，才能对客观数据对象建立相应的心理图像，完成认知理解过程。好的可视化就能够极大地降低这个认知壁垒，将复杂未知数据的交互探索变得可行。可视化技术的进步和广泛应用对大数据走向平民来说，意义是双向的。一方面，可视化作为人和数据之间的界面，结合其他数据分析处理技术，为广大使用者提供了强大的理解、分析数据的能力。可视化使得大数据能够被更多人理解、使用。可视化使得大数据的使用者从少数专家扩展到更广泛的大众。另一方面，可视化也为大众提供了方便的工具，可以主动分析处理与个人工作、生活、环境有关的数据。大约在 10 年前，可视化研究界已经开始讨论为大众服务的可视化。在今天的大数据背景下，可视化将进一步推动大数据平民化。在这一过程中，急需更方便且适合大众使用需要的可视化方法和工具。可视化也将进一步和个人使用的移动通信设备（手机）结合。在这一过程中，将有更多面向大众的大数据可视化公司涌现出来。

建议在大数据相关的研究、开发和应用中，保持相应的比例用于可视化和可视分析。尤其建议利用产业生态中的已有成果。

13.1.2 趋势二：多学科融合与数据科学的兴起

很多与数据相关的专门实验室、专项研究院所相继出现，《数据学》等专门著作也纷纷出版，大家认为数据科学的雏形已经出现。

如图 13.1 所示，大数据并不是简单的“大的数据”。在近年对大数据的阐述中，至少有两种典型的对应提法：一种是点出“小数据”的重要性；另一种是去掉“大”字而强调“数据”本身，强调数据科学、数据技术、数据治理、数据产业等。

大数据技术是多学科多技术领域的融合，数学和统计学、计算机类技术、管理类等都有涉及，大数据应用更是与多领域产生交叉。这种多学科之间的交叉融合，呼唤并催生了专门的基础性学科——数据学科。基础性学科的夯实，将让学科的交叉融合更趋完美。

在大数据领域，许多相关学科从表面上看，研究的方向大不相同，但是从数据的视角看，其实是相通的。随着社会的数字化程度逐步加深，越来越多的学科在数据层面趋于一致，可以采用相似的思想进行统一研究。从事大数据研究的人不仅仅是计算机领域的科学家，也包括数学等方面的科学家。

大数据专家委希望业界对于大数据的边界采取一个更宽泛、更包容的姿态，包容所谓的“小数据”，甚至将领域的边界泛化到“数据科学”所对应的整个数据领域和数据产业。

建议共同支持“数据科学”的基础研究，并努力将基础研究的成果导入技术研究和应用的范畴中。

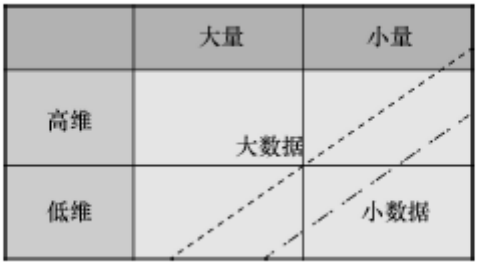

图 13.1 大数据与小数据

13.1.3 趋势三：大数据安全与隐私令人忧虑

安全和隐私每次调研都会出现在十大趋势中，这表示大家对于大数据所带来问题的深刻忧虑，这样的忧虑至少包括以下 3 个方面。

第一，大数据所受到的威胁也就是常说的安全问题。这里并不是指利用大数据进行安全分析的“安全大数据”应用，而是指当大数据技术、系统和应用聚集了大量价值的时候，必将成为被攻击的目标。虽然，现在影响巨大的针对大数据的攻击还没有出现，但是可以预见这样的攻击必将发生。

第二，大数据的过度滥用所带来的问题和副作用，比较典型的就是个人隐私泄露。在传统采集分析模式下，很多可以保护的隐私在大数据分析能力下变成了裸奔。类似的问题还包括大数据分析能力带来的商业秘密泄露和国家机密泄露。

第三，心智和意识上的安全问题。这包括两个极端：一个极端是忽视安全问题的盲目乐观；另一个极端是过度担忧所带来的对于大数据应用发展的掣肘。比如，大数据分析对于隐私保护的副作用，促使大家必须对于隐私保护的接受程度有一个新的认识和调整。

对大数据的威胁、大数据的副作用、对大数据的极端心智都会阻碍和破坏大数据的发展。

如图 13.2 所示，大数据技术分别作用在业务、威胁、保障措施 3 个要素之上，带来保护大数据、对抗大数据级威胁、大数据用于安全 3 方面的安全发展空间。

建议在大数据相关的研究和开发中，必须保持一个基础的比例用于相对应的安全研究，而让安全方面产生实质性进步的驱动力可能是对于大数据的攻击和滥用的“负面”研究。

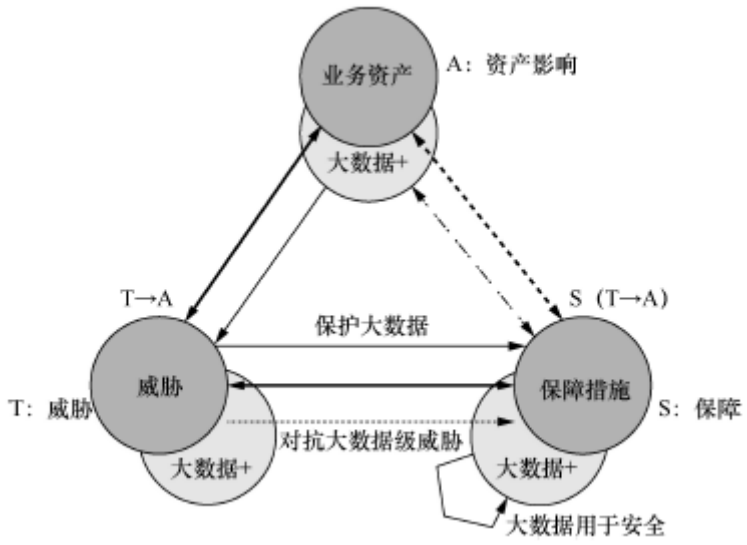

图 13.2 大数据技术作用于业务、威胁、保障措施之上

13.1.4 趋势四：新热点融入大数据多样化处理模式

大数据的处理模式更加多样化，Hadoop 不再成为构建大数据平台的必然选择。在应用模式上，大数据处理模式持续丰富，批量处理、流式计算、交互式计算等技术面向不同的需求场景，将持续丰富和发展；在实现技术上，内存计算将继续成为提高大数据处理性能的主要手段，相对传统的硬盘处理方式，在性能上有了显著提升。特别是开源项目 Spark，目前已经被大规模应用于实际业务环境中，并发展成为大数据领域最大的开源社区。Spark 拥有流计算、交互查询、机器学习、图计算等多种计算框架，支持 Java、Scala、Python、R 等语言接口，使得数据使用效率大大提高，吸引了众多开发者和应用厂商的关注。值得说明的是，Spark 系统可以基于 Hadoop 平台构建，也可以不依赖 Hadoop 平台独立运行。

很多新的技术热点持续地融入大数据的多样化模式中，目前不会有一个一统天下的唯一模式。从 2015 年中国大数据技术大会众多技术论坛的安排也可以看到这样的多样化态势。技术各有千秋，形成一个更加多样、平衡的发展路径，也满足大数据的多样化需求。大数据专家委的专家们认为，这样的态势还会持续下去。

建议将自己机构的大数据研究和开发，有意识地链接和融入大数据技术生态中，或者利用技术生态的成果，或者回馈技术生态。

13.1.5 趋势五：大数据提升社会治理和民生领域应用

基于大数据的社会治理成为业界关注热点，涉及智慧城市、应急、税收、反恐、农业等多个民生领域。

大数据从来都是应用驱动，技术发力。在最易获得大数据应用成果的互联网环境之后，大数据走进国计民生成为必然。而在 2016 年，与民生有关的应用将成为热点。国计与民生并不互斥，涉及民生的国计将是快速发展热点中的热点。比如，反恐、医疗健康等都与老百姓密切相关，同时也是国家大计。

由于更易获得关注并对接真实需求，建议优先投入社会治理和民生方面的大数据工作。

13.1.6 趋势六：《促进大数据发展行动纲要》驱动产业生态

国务院在 2015 年 8 月 31 日印发了《促进大数据发展行动纲要》。纲要明确指出了大数据的重要意义，大数据成为推动经济转型发展的新动力、重塑国家竞争优势的新机遇、提升政府治理能力的新途径。纲要还清晰地提出了大数据发展的主要任务：加快政府数据开放共享，推动资源整合，提升治理能力；推动产业创新发展，培育新兴业态，助力经济转型；强

化安全保障，提高管理水平，促进健康发展。纲要还提出了组织、法规、市场、标准、财政、人才、国际交流等几方面的政策机制要求。

纲要将对大数据的发展起到重大的推动作用，成为一个产业快速发展的催化剂和政策标杆。而各个地方政府一定会出台类似配套的政策。在中央和地方的政策推动下，政府的大数据专项扶植政策和一些相关政策（如大众创业、万众创新的双创政策）集中出台。

政府牵引产业生态，带动数据共享交换。政府带动的数据共享将成为数据流转的源动力，让数据开放共享、交换交易成为产业生态的新态势，政策让数据流转动起来。国有和民间资本的集中注入，大数据相关的基础设施建设的采购和投入，使政策和市场双重发力，让资金流转动起来。政府牵引的产业生态发展成为大数据发展历程在 2016 年的突出特点。

建议应及时关注和跟踪大数据相关的政策。有实力的机构应投入一定的北向[1]资源，主动影响和引导各级政府的政策和落实细则。

13.1.7 趋势七：深度分析推动大数据智能应用

在学术技术方面，深度分析会继续成为一个代表，推动整个大数据智能的应用。这里谈到的智能，尤其强调是涉及人的相关能力延伸，比如决策预测、精准推荐等。这些涉及人的思维、影响、理解的延展，都将成为大数据深度分析的关键应用方向。

相比于传统机器学习算法，深度学习提出了一种让计算机自动学习产生特征的方法，并将特征学习融入建立模型的过程中，从而减少了人为设计特征引发的不完备。深度学习借助深层次神经网络模型，能够更加智能地提取数据不同层次的特征，对数据进行更加准确、有效的表达。而且训练样本数量越大，深度学习算法相对传统机器学习算法就越有优势。

目前，深度学习已经在容易积累训练样本数据的领域，如图像分类、语音识别、问答系统等应用中获得了重大突破，并取得了成功的商业应用。预测随着越来越多的行业和领域逐步完善数据的采集和存储，深度学习的应用会更加广泛。当然，在分析领域，也并不会是深度学习一统天下的局面。由于大数据应用的复杂性，多种方法的融合将是一个持续的常态。

建议保持对于智能技术发展的持续关注。在各自的分析领域（如在策划阶段、技术层面、实践环节等）尝试一下深度学习还是值得的。

13.1.8 趋势八：数据权属与数据主权备受关注

数据权属与数据主权被高度关注，在个人和一般机构看是数据权属问题，从国家层面看

[1]北向：指战略、政策法规、产业环境、规划、架构、治理和管理等方面的非工程、非基础技术方面的工作和领域。

是数据主权问题。

大数据凸显了数据的巨大价值。而数据的权属问题并不是传统的财产权、知识产权等可以涵盖的权属问题。数据成为国家之间争夺的资源，数据主权成为网络空间主权的重要形态。

数据成为重要的战略资源。人口红利、地大物博、经济实力、文化优势等都纷纷体现为数据资源储备和数据服务影响力。

而数据资源化、价值化是数据权属问题和数据主权问题的根源。

过度关注数据权属，并仿照财产权或知识产权模式对数据增加过多的限制，不利于大数据的发展。在商业层面和科研层面，现阶段应当看淡一些数据权属问题。而在国家层面，应当积极推行数据主权认识，并且鼓励数据进口，适当限制数据出口。

13.1.9 趋势九：互联网、金融、健康保持热度，智慧城市、企业数据化、工业大数据是新增长点

我国大数据应用领域最早获得成果的就是互联网应用（包括电商等），而持续受到高度关注的应用领域还包括金融和健康，互联网、金融、健康可称为大数据应用领域的老三样。而智慧城市、企业数据化、工业大数据则成为新的增长点，这新三样就是城市、企业、工业的数据化，或者说是城市生活、企业贸易和管理、工业生产过程的数据化和大数据应用。新三样是一种更广泛的应用领域覆盖。

表 13.1 和表 13.2 分别为 2013-2016 年最令人瞩目的应用领域投票结果和 2015-2016 年将取得应用和技术突破的数据类型投票结果。

从表 13.1 和表 13.2 可以看出，“最令人瞩目的应用领域”和“将取得应用和技术突破的数据类型”这两项调研投票的结果印证了老三样和新三样的判断。

建议顺应潮流，这样更易获得资源支持。

13.1.10 趋势十：开源、测评、大赛催生良性人才与技术生态

大数据是应用驱动，技术发力，技术与应用一样至关重要。决定技术的是人才及其技术生产方式。开源系统将成为大数据领域的主流技术和系统选择。以 Hadoop 为代表的开源技术拉开了大数据技术的序幕，大数据应用的发展又促进了开源技术的进一步发展。开源技术的发展降低了数据处理的成本，引领了大数据生态系统的蓬勃发展，同时也给传统数据库厂商带来了挑战。新的替代性技术，都是新技术生态对于旧技术生态的侵蚀、拓展和进化。

对数据处理的能力、性能等进行测试、评估、标杆比对的第三方形态出现，并逐步成为热点。相对公正的技术评价有利于优秀技术占领市场，驱动优秀技术的研发生态。各类创业创新大赛纷纷举办，大赛为人才的培养和选拔提供了新模式。各类创业创新大赛完善人才生态。大数据技术生态是一个复杂环境。在 2016 年，“开源”会一如既往占据主流，而测评和大赛将形成突破性发展。建议不要闭门搞大数据技术和系统，要开门融入世界性的技术生态中。

表 13.1 2013-2016 年最令人瞩目的应用领域投票结果

（按照票数多少从上到下排序）

年份	2013年	2014年	2015年	2016年
应用领域	医疗	互联网；电子商务	互联网；电子商务	互联网；电子商务
	金融	金融	金融	金融
	电子商务	健康医疗	健康医疗	健康医疗
	城市管理	舆情分析；情报分析	城镇化；智慧城市	城镇化；智慧城市
			社会安全；犯罪侦查	舆情分析；情报分析

表 13.2 将取得应用和技术突破的数据类型投票结果

（按照票数多少从上到下排序）

年份	2015年	2016年
数据类型	社会化媒体数据	城市数据
	视频数据	互联网交易相关数据
	互联网日志与电商交易数据	企业数据
	语音数据、图形图像	
	设备测量和控制数据	视频数据
	图形图像数据	
	人体数据、宏观经济	人体数据

2 016 年大数据产业技术发展的十大趋势预测可以简单解读为 4 个关键词：一是“民生”，在众多的大数据相关应用中，相对来说，与民生相关的大数据可能会得到更快的发展，比如：健康医疗、社会治安、环境保护等；二是“多样性和融合性”，包括技术模式融合、产业融

合等各方面的融合；三是“政策拉动”；四是“生态”，产业生态、技术生态等生态的构建是发展的大环境。2013-2016 年对大数据发展的十大趋势预测结果见表 13.3。

表 13.3 2013-2016 年对大数据发展的十大趋势预测

年份	2013年	2014年	2015年	2016年
十大发展趋势预测	• 数据的资源化； • 大数据的隐私问题突出； • 大数据与云计算等深度融合； • 基于大数据的智能的出现； • 大数据分析的革命性方法； • 大数据安全； • 数据科学兴起； • 数据共享联盟； • 大数据新职业； • 更大的数据	• 大数据从“概念”走向“价值”； • 大数据架构的多样化模式并存； • 大数据安全与隐私； • 大数据分析与可视化； • 大数据产业成为战略性产业； • 数据商品化与数据共享联盟化； • 基于大数据的推荐与预测流行； • 深度学习与大数据智能成为支撑； • 数据科学的兴起； • 大数据生态环境逐步完善	• 大数据分析成为数据价值化的热点； • 数据科学带动学科融合，但自身尚未成体系； • 与各行业结合，跨领域应用； • “物云移社”融合，产生综合价值； • 平台架构与基础设施； • 大数据的安全与隐私保护； • 计算模式：深度学习、众包计算； • 可视化分析与可视化呈现； • 大数据人才与教育； • 开源系统将成为主流选择	• 可视化推动大数据平民化； • 多学科融合与数据科学的兴起； • 大数据安全与隐私令人忧虑； • 新热点融入大数据多样化处理模式； • 大数据提升社会治理和民生领域应用； • 《促进大数据发展行动纲要》驱动产业生态； • 深度分析推动大数据智能应用； • 数据权属与数据主权备受关注； • 互联网、金融、健康保持热度，智慧城市、企业数据化、工业大数据是新增长点； • 开源、测评、大赛催生良性人才与技术生态
关键词	最初的结构性认识	大数据从“概念”走向“价值”	跨界融合、基础突破	民生、多样、政策、生态

13.2 大数据发展的单项调研结果

13.2.1 与大数据最匹配的概念

大数据本身具有很强的概念性，不可否认大数据有它的泡沫（甚至炒作的成分），但是不能因为啤酒上面有泡沫放弃底下香浓的啤酒。大数据专家委针对时下流行的重大概念进行调研，在众多流行的概念中，专家们认为和大数据最匹配的概念是“互联网+、云计算和智慧城市”，而其他选项（物联网、移动互联网、大众创业万众创新、工业互联网（工业 4 .0）、智能生活设备、一带一路）则具有数量级的落差。

建议让自己的大数据工作，同时再挂上 1~2 个业界热点概念。这是有益而无害的，只要不仅仅停留在概念炒作。

13.2.2 我国大数据发展最主要的推动者

表 13.4 为 2015-2016 年我国大数据发展最主要推动者的调研结果，可以看出，目前最主要的推动者是大型互联网公司、政府机构和创业公司。

从表 13.4 可以看出大型互联网公司的惯性优势，2016 年以纲要为代表的政策性支持、双创政策对于创业激情的拉动，将是大数据发展的主要推动力，而科研和公共服务的影响则

相对弱化了。

建议让自己的机构变成推动者或者与这 3 类推动者建立合作。

表 13.4 2015-2016 年我国大数据发展最主要推动者的调研结果

年份	2015年	2016年
推动者	大型互联网公司	大型互联网公司
	政府机构	政府机构
	国内大学和科研院所	创业企业
	公共服务机构	
	创业企业	

13.3 数据资源流转并不乐观

在第一篇里，重点阐述了大数据开放共享的问题。今年的趋势调研也专门设立了这样一项调研：2016 年，100 多位专家和他的机构对数据的态度是什么，对数据流转的态度是什么。从调研结果中看到，大家都想自己收集数据，希望能够利用收集的数据进行数据服务，希望能够买到数据集，而准备卖数据集的机构非常少。整个数据流转上是需求大于供给的状态，数据确实奇货可居。而考虑数据国际交换和卖数据集的投票者更是屈指可数。整个数据流转的态势不容乐观。希望通过政府开放共享拉动数据交流和交换。

在现有的生态环境下，想要免费或者低价获得高品质的数据是有困难的，要降低这种期望值。在数据需求大于供给的大环境下，数据采集和储藏是一个很合算的投入方向，如果再结合轻度的数据冶炼，可以让自己的机构进入抢手的数据提供者行列。

13.4 对大数据发展阶段的判断体现出对于成长性的极为乐观

表 13.5 为对大数据发展阶段的判断结果。大数据专家委的专家们对当前中国大数据所处的阶段进行选择（单选）。从 2015 年和 2016 年的调研结果对比可以看出，专家们具有明显的乐观态度，2016 年预测上升的人数增加，而预测下降的人数屈指可数。而且选择“极为初级”和“即将快速扩张”两个阶段的专家超过 70%，也就是认为大数据的峰顶还远没有看到，是极为乐观的发展预期。在政策、市场、技术的多重推动下，大数据将有非常美好

的前景。

建议投入、投入、投入！投入资源到大数据领域，赢的概率很大。

表 13.5 对大数据发展阶段的判断

发展阶段	2015年	2016年
极为初级	17%	33%
即将快速扩张	31%	40%
爆发增长中	10%	9%
达到一个顶峰上升乏力	18%	4%
达到一个顶峰将下降和幻灭	5%	0%
稳步成长中	20%	14%

13.5 群体智慧和“黑天鹅”

上述是对大数据专家委的专家们观点的统计性结果和解读分析，难以涵盖专家们的独特观点和“黑天鹅判断”。不过，这样的群体性预测，仍具有很高的参考价值。2016 年大数据领域是否会出现重大“黑天鹅事件”的投票结果显示，42 %的专家认为会出现，而 58%的专家认为不会。

大数据领域的“黑天鹅”绝对是机遇大于威胁。积极地为“黑天鹅”做好准备，也就是让自己的机构有能力根据突发的“黑天鹅”而调动（或者撬动）10%以上的资源。

参考文献

[1]《促进大数据发展行动纲要》.
http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.
[2] NIH Data Sharing Policy.
http://grants.nih.gov/grants/policy/data_sharing/index.htm.
[3] NSF Policy. http://www.nsf.gov/pubs/2001/gc101/gc101rev1.pdf
[4] Treloar, A. (2014), "The Research Data Alliance: globally coordinated action against barriers to data publishing and sharing", Learned Publishing: special issue to Volume 27, pp 9-13(5), September 1.
[5] Kalil, Tom. "Big Data is a Big Deal". White House. Retrieved 26 September 2012.

[6] Reichman O.J., Jones M.B., Schildhauer M.P. (2011). "Challenges and Opportunities of Open Data in Ecology". Science 331 (6018): 703–705.
[7] Vogeli C, Yucel R, Bendavid E; et al. (February 2006). "Data withholding and the next generation of scientists: results of a national survey". Acad Med 81 (2): 128–36.
[8]Sharing and publishing data for public benefit.
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/223931/130717_Data_sharing_condoc_final_v2.pdf
[9]《上海推进大数据研究与发展三年行动计划（2013-2015 年）》.
http://www.stcsm.gov.cn/gk/ghjh/333008.htm
[10]《广东省经济和信息化委员会主要职责内设机构和人员编制规定》.
http://zwgk.gd.gov.cn/006939748/201402/t20140226_480387.html
[11]《智慧珠海 2015 行动计划》.
http://www.zhkgmx.gov.cn/gksxx/xxhtjk/tzgg/201405/P020140526385120547259.pdf.
[12]《贵阳大数据产业行动计划》.
http://zgcgy.gov.cn/art/2015/1/30/art_22803_699382.html.
[13]《关于加强智慧交通体系建设的指导意见》.
http://www.cac.gov.cn/2015-08/11/c_1116214892.html.
[14] Volkmar Koch et al, PWC, "Industry 4.0: Opportunities and challenges of the industrial internet", 2015.
[15] Peter C. Evans and Marco Annunziata, GE Corporation, "Industrial Internet: Pushing the Boundaries of Minds and Machines", 2012.
[16] Mckinsey Global Institute, "Manufacturing the future, The next era of global growth and innovation", 2012.
[17] 国务院,《中国制造 2025 规划纲要》, 2015 年 3 月 25 日.
[18] GE Software, "The Case for an Industrial Big Data Platform", 2012
[19] McKinsey & Company, "Big data: The next frontier for innovation, competition, and productivity", 2012.
[20] 搜狐, "借助百度鹰眼 361 度推儿童防丢鞋".
http://it.sohu.com/20150201/n408302924.shtml, 2015 年 2 月 1 日.