CCF大专委2016年大数据发展趋势预测——解读和行动建议

Developing trend forecasting of big data in 2016 from CCF TFBD: interpretation and proposals

潘柱廷（1969–），男，启明星辰首席战略官。教授级高级工程师，长期从事信息安全技术和战略研究工作。中国计算机学会（CCF）常务理事，CCF大数据专家委员会委员兼副秘书长，CCF计算机安全专家委员会常务委员，中国互联网协会常务理事，云安全联盟CSA中国区理事。网络空间安全协会——网络安全人才与教育工委（筹）召集人。

程学旗（1971–），男，中国科学院计算技术研究所研究员、所长助理、副总工程师，中国科学院网络数据科学与技术重点实验室主任，CCF大数据专家委员会秘书长，国家杰出青年科学基金获得者。先后主持并完成了10余项国家自然科学基金、国家“973”计划、国家“863”计划、国家信息安全重大专项等科研任务。两次获得国家科技进步奖二等奖，获得第十二届中国青年科技奖、中国计算机学会青年科学家奖、中国科学院青年科学家奖等荣誉。

袁晓如（1975–），男，北京大学“百人计划”研究员，北京大学信息科学技术学院博士生导师。主要研究方向包括：高动态范围视频、图像和可视化；大规模数据的高性能绘制和可视化；非真实性绘制及插图式可视化；新颖可视化界面与人机交互研究；高维数据可视化。关于高动态范围可视化的工作获得2005年IEEE Visualization大会的最佳论文奖。

周涛（1979–），男，博士，北京启明星辰信息安全技术有限公司教授级高工，主要研究方向为大数据安全分析、事件关联分析、入侵检测等。

靳小龙（1976–），男，中国科学院计算技术研究所副研究员，博士生导师，中国科学院网络数据科学与技术重点实验室网络数据科学研究部负责人，CCF大数据专家委员会委员。主要研究兴趣包括社会计算、社会网络、网络数据分析、多智能体系统等。

中图分类号：TP399　　文献标识码：A
doi: 10.11959/j.issn.2096-0271.2016012

1 引言

2015中国大数据技术大会（BDTC）于2015年12月10日在北京召开，会上中国计算机学会（CCF）大数据专家委员会（task force on big data，TFBD，以下简称大专委）发布了中国大数据技术与产业发展报告（2015），并对2016年大数据发展趋势进行了展望。

自2012年10月CCF大专委成立，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。从预测2013年到预测2016年，现在已经是第4次年度预测。每次预测都是基于对大专委专家委员观点的收集整理、投票、汇总、解读，最终形成年度预测，此预测是大专委群体智慧的结晶。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联盟成员参加了投票和汇总，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，参加投票的大专委专家和产业联盟成员是116位。根据这116位专家投票结果，汇总形成了对2016年大数据发展十大趋势的预测，下面对这十大发展趋势进行详细介绍。

2 2016年大数据发展十大趋势

2.1 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大众人心，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。

2.2 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。

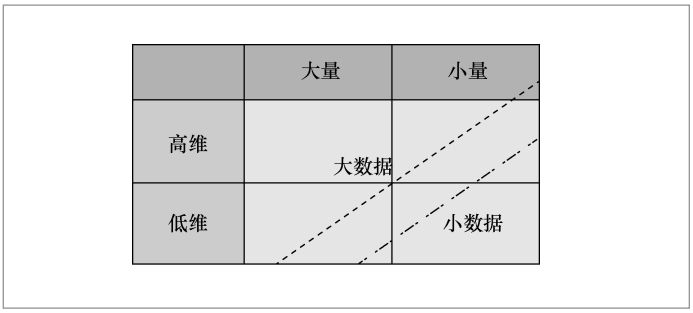

图1 大数据与小数据

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。

大专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。

2.3 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

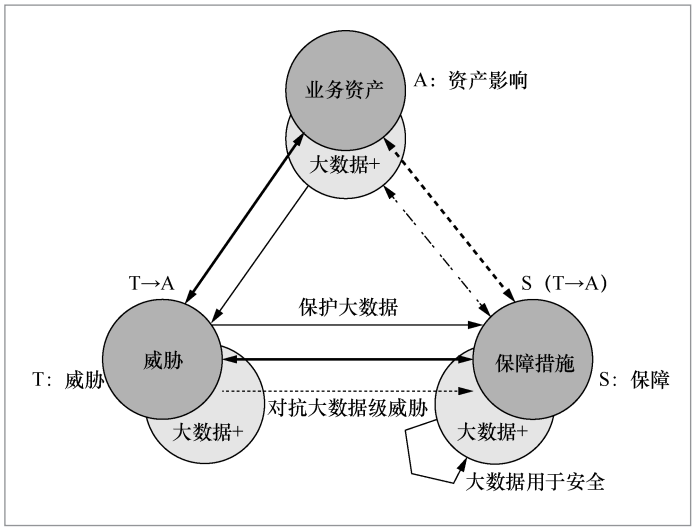

图 2 大数据技术作用于业务、威胁、保障措施之上

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。

2.4 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。大专委的专家们认为，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。

2.5 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。

2.6 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。

2.7 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。

2.8 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、

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。

表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

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。

2.9 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1和表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1和表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。

2.10 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。

2016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3。

3 大数据发展的单项调研结果

3.1 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。大专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。

表 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

3.2 我国大数据发展最主要的推动者

表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。

3.3 数据资源流转并不乐观

在大专委即将发布的第三本《中国大数据技术与产业发展年度报告》中，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。

表 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	

表 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%

3.4 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5为对大数据发展阶段的判断结果。大专委的专家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。

3.5 群体智慧和“黑天鹅”

上述是对大专委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。 □