# 大数据支撑政府科学决策精准决策方面的研究

## 1 大数据时代背景

大数据概念最初起源于美国，是由思科、威睿、甲骨文、IBM等公司倡议发展起来的。而“大数据”的名称则出现更早，最早来自1980年著名未来学家托夫勒的著作《第三次浪潮》，他将“大数据”称颂为“第三次浪潮的华彩乐章”。《自然》杂志在2008年9月推出名为“大数据”的封面专栏，大数据研究开始迎来热潮。

2009年，“大数据”成为互联网信息技术行业的流行词汇，大数据产业逐渐兴起，是指建立在对互联网、物联网、云计算等渠道广泛、大量数据资源收集基础上的数据存储、价值提炼、智能处理和分发的信息服务业，大数据企业大多致力于让所有用户几乎能够从任何数据中获得可转换为业务执行的洞察力，包括之前隐藏在非结构化数据中的洞察力。

最早提出“大数据时代已经到来”的机构是全球知名咨询公司麦肯锡，在2011年，麦肯锡在题为《海量数据，创新、竞争和提高生产率的下一个新领域》的研究报告中指出，数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产因素，而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。自此，大数据开始跳出信息产业，逐渐受到各行各业的关注。

大数据是一个不断演变的概念，当前的兴起，是因为从IT技术到数据积累，都已经发生重大变化。仅仅数年时间，大数据就从互联网行业的热词演变成决定我们未来数字生活方式的重大技术命题。2012年，联合国发表大数据政务白皮书《大数据促发展：挑战与机遇》；EMC、IBM、Oracle等跨国IT巨头纷纷发布大数据战略及产品，几乎所有世界级的互联网企业，都将业务触角延伸至大数据产业，而美国政府投资2亿美元启动“大数据研究和发展计划”更是将大数据上升到国家战略层面。大数据已是一股时代热潮，将深刻的影响社会生活的方方面面。

## 2 大数据的内涵与外延

### 2.1 内涵

舍恩伯格和库克首先在《大数据时代》中给出大数据的定义：不用随机分析法（抽样调查）这样的捷径，而采用所有的数据进行分析处理。随着大数据渗透到各行各业，对于大数据的概念、内涵等存在着众多不同的理解。国内外至今对于大数据的概念还没有一个准确、权威的定义。本文尝试从大数据的**属性、来源、比较、架构**四个方面来阐述大数据的内涵：

**大数据的属性。**首先由Gartner机构的分析师提出了大数据的3V定义，即大数据是指需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的**数量大、高增长率**和**多样化**的信息资产。3V定义被广泛接受和采用，随之发展成为4V，即大数据代表了信息资产的特点：高容量、速度快、种类多、需要特定的技术和分析方法将其转化成价值。

**大数据的来源。**大数据的来源可分为三类：一是人们通过互联网进行信息传递和交互产生的文本、图片、声音和视频等数据；二是计算机信息系统内部产生的文件、数据库等自动生成；三是各类物联网，如传感器、家电、汽车、仪器仪表等设备产生的海量数字数据。根据来源主体的不同，大数据又可以分为三类：一是来自于人，即人的互联网活动所产生的各类数据；二是来自于机，即计算机信息系统自身产生的信息；三是来自于物，即物联网的各类数字设备产生的数据。

**大数据与传统数据的比较。**大数据数据量大，结构复杂，变化快，需要用更先进的算法来实现分析，相较于传统数据的特点和分析方法，有明显的差异，表1列出了两者的差异。

**表1 大数据与传统数据的对比**

|  |  |  |
| --- | --- | --- |
| 项 | 传统数据 | 大数据 |
| 容量单位 | GB | TB\PB\ZB |
| 产生速度 | 每小时，每天计算 | 实时 |
| 结构 | 结构化 | 半结构化或非结构化 |
| 数据源 | 集中 | 分散 |
| 数据整合 | 简单 | 困难 |
| 数据存储 | RDBMS | HDFS,NOSQL |
| 使用 | 交互式 | 批处理或者实时 |

**大数据的架构。**大数据可以被进一步分为，大数据科学和大数据架构。大数据科学是指“学习获得，修复，调高大数据价值的技术”，大数据架构是指“学习各种软件及其相关算法，使得分布式处理和分析大数据的跨越计算机单元的聚集问题”。

### 2.2 外延

**大数据思维。**大数据不但从技术上改变人们的生活方式，也要求人们转变思维方式，以应对大数据时代的到来。大数据思维可以从十个方面展开：

1. 数据核心原理，从“流程”核心转变“数据”核心。
2. 数据价值原理，由“功能是价值”转变为“数据是价值”。
3. 全样本原理，从抽样转变为需要全部数据样本。
4. 关注效率原理，由关注精确度变为关注效率。
5. 关注相关性原理，由因果关系转变为关注相关性。
6. 预测原理，从不能预测转变为可以预测。
7. 信息找人原理，从人找信息，转变为信息找人。
8. 机器懂人原理，由人懂机器转为机器更懂人。
9. 电子商务智能原理，大数据改变了电子商务模式，让电子商务更智能。
10. 定制产品原理，由企业生产产品转为由客户定制产品。

**大数据分析。**大数据已经不简简单单是数据大的事实了，而最重要的现实是对大数据进行分析，只有通过分析才能获取很多智能的，深入的，有价值的信息。那么越来越多的应用涉及到大数据，而这些大数据的属性，包括数量，速度，多样性等等都是呈现了大数据不断增长的复杂性，所以大数据的分析方法在大数据领域就显得尤为重要，可以说是决定最终信息是否有价值的决定性因素。大数据分析可分为如下四个方面：

（1）可视化分析。大数据分析的使用者有大数据分析专家，同时还有普通用户，但是他们二者对于大数据分析最基本的要求就是可视化分析，因为可视化分析能够直观的呈现大数据特点，同时能够非常容易被读者所接受，就如同看图说话一样简单明了。

（2）数据挖掘算法。大数据分析的理论核心就是数据挖掘算法，各种数据挖掘的算法基于不同的数据类型和格式才能更加科学的呈现出数据本身具备的特点，也正是因为这些被全世界统计学家所公认的各种统计方法（可以称之为真理）才能深入数据内部，挖掘出公认的价值。另外一个方面也是因为有这些数据挖掘的算法才能更快速的处理大数据，如果一个算法得花上好几年才能得出结论，那大数据的价值也就无从说起了。

（3）预测性分析能力。大数据分析最终要的应用领域之一就是预测性分析，从大数据中挖掘出特点，通过科学的建立模型，之后便可以通过模型带入新的数据，从而预测未来的数据。

（4）数据质量和数据管理。大数据分析离不开数据质量和数据管理，高质量的数据和有效的数据管理，无论是在学术研究还是在商业应用领域，都能够保证分析结果的真实和有价值。

**大数据技术。**从数据的生命周期角度，从如下八个方面阐述大数据的技术：

（1）数据采集：ETL工具负责将分布的、异构数据源中的数据如关系数据、平面数据文件等抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。

（2）数据存取：关系数据库、NOSQL、SQL等。

（3）基础架构：云存储、分布式文件存储等。

（4）数据处理：自然语言处理(NLP，NaturalLanguageProcessing)是研究人与计算机交互的语言问题的一门学科。处理自然语言的关键是要让计算机"理解"自然语言，所以自然语言处理又叫做自然语言理解(NLU，NaturalLanguage Understanding)，也称为计算语言学(Computational Linguistics。一方面它是语言信息处理的一个分支，另一方面它是人工智能(AI, Artificial Intelligence)的核心课题之一。

（5）统计分析：假设检验、显著性检验、差异分析、相关分析、T检验、方差分析、卡方分析、偏相关分析、距离分析、回归分析、简单回归分析、多元回归分析、逐步回归、回归预测与残差分析、岭回归、logistic回归分析、曲线估计、因子分析、聚类分析、主成分分析、因子分析、快速聚类法与聚类法、判别分析、对应分析、多元对应分析（最优尺度分析）、bootstrap技术等等。

（6）数据挖掘：分类（Classification）、估计（Estimation）、预测（Prediction）、相关性分组或关联规则（Affinity grouping orassociation rules）、聚类（Clustering）、描述和可视化、Description and Visualization）、复杂数据类型挖掘(Text,Web ,图形图像，视频，音频等)

（7）模型预测：预测模型、机器学习、建模仿真。

（8）结果呈现：云计算、标签云、关系图等。

### 2.3 意义

毫无疑问，大数据正以极速方式嵌入经济社会生活的方方面面。数据驱动发展已成为时代主题，大数据正在开启一个崭新时代。它不仅为新一代信息技术产业提供了核心支撑，正成为经济增长和社会发展的高速引擎、科技创新的新动力和新源泉。

**大数据正重塑企业未来发展的方向。**大数据作为未来的“石油”、隐形的“金矿”、经济的“燃料”，正和企业固定资产、人力资本等生产要素一起，被纳入企业最宝贵的生产要素进行管理；大数据正与企业品牌、人才、发展战略等形成企业巨大的无形资产，作为全球经济低迷环境下的产业亮点，成为是战略新兴产业最活跃的部分。谁能顺应数字驱动发展这一主题，谁就能顺势谋变、乘势而起，必将成为新时代的新领军人物；谁墨守成规而无动于衷，谁就会使企业丧失市场竞争的活力和动力。

**大数据正掀起国家和社会治理革命。**利用大数据及其分析技术，可以使政府及其各社会治理主体快速获取有价值的信息，提高公共决策和社会治理能力，及时破解治理中存在的顽疾和痛点，深化阳光政府、责任政府和智慧政府建设，提升国家和社会治理效率、重构治理模式；它可使精准医疗、个性化教育、社会监管、舆情监测、安全预警等以往无法实现的环节变得更加简单、可操作；它可以更好地维护数据主权、开放数据资产、保持在数字空间领域的国家竞争力，进而更好地维护国家网络空间主权。

**大数据正成为地方和城市发展的新引擎。**迄今为止，越来越多的地方和城市，已深刻认识到大数据的巨大价值及潜力，将大力发展大数据产业摆上重要日程，纷纷推出相关战略及行动计划，着力推动大数据与“工业化、信息化、城市化、农业现代化”的互动、协同、整合发展，以确保在大数据发展浪潮中抢占一席之地，进而提升地区和城市的竞争力和综合实力。

## 3 大数据带来的政府决策变革

### 3.1 决策变革

**政府决策主体：从“精英”转向“大众”。**传统政府决策主体主要是政府管理层、企业精英、专家学者作为智囊团辅助政府人员进行决策，但普通大众很少参与。随着大数据思维的冲击，政府决策主体必将趋向于多元化、大众化。大数据处理平台会对各种数据进行采集、整理、分析，使普通的大众数据能成为政府决策基础的高质量信息，这就为政府和大众的信息交互提供了便利，普通大众也能有序协同参与政府决策过程，使政府决策更具广泛代表性、民主性。

**政府决策结构：从“等级制”转向“扁平化”。**现行的政府决策大多采用自上而下的决策模式，决策结构是和金字塔式的组织形式相对应的，容易产生“信息孤岛”，部门间信息共享受阻，导致决策滞后、不协调。随着大数据技术的不断更新，政府决策将会形成一种基于数据决策支持系统的网络状新型政府决策模式，趋于扁平化，使得各部门在决策系统中协同合作，提高决策效率。

**政府决策方式：从“经验型”转向“数据驱动型”。**传统情况下，政府人员常常依据工作经验进行决策，常因“拍脑袋决策”而产生决策失误问题。在大数据时代，数据将会渗透到政府决策的各个方面，不管是经济管理方面，还是公共服务领域，政府必将在数据驱动的基础上进行决策，而数据分析是决定决策质量高低的关键性因素。数据驱动型决策是一个将原始数据转换成信息，并最终通过一系列的认知技术和处理流程转变为知识的连续的决策过程。

**政府决策过程：从“事后解决”转向“事先预测”。**当今社会是风险社会，各种矛盾不断激发。政府的决策行为往往滞后于事态的发展，是一种被动决策，无法及时解决社会问题。全球复杂网络的权威研究者巴拉巴西认为“93％的人类行为都是可预测的” 大数据的核心就是预测，通过将人类的行为量化，分析数据潜在规律，预测事态的发展趋势，有助于政府事先阻止矛盾激发，尤其是在公共安全治理方面。例如，面对交通安全突发事件，传感器或其他移动设备会成为交通状况的数据源，位置、速度和其他事件信息会实时发回大数据中心，并通过地图标记显示各路段的运行状况，以预测交通状况和可能产生的风险，实现突发事件事前控制。

**政府决策结果：从“预报”转向“实报”。**联合国在2009年最先提出“数据脉动”计划时，就已追求政府决策从“预报”转向“实报”，着重强调数据分析、结果预测的实时性以及报告的精准性。例如，美国联邦政府国家气象局就利用传感器技术获取气象数据，每10秒钟采集一次，每天要采集超过10万次的数据，这些海量、精度极高的实时数据意味着天气预报将逐渐从“预报”走向“实报”“精报”。

**政府决策信息：从“静态公开”转向“动态公开”。**目前，我国政府已初步公开决策信息，但都是以表格、文字报告的形式展现出来，信息更新迟缓。大数据技术将推动政府信息公开从“静态”转向“动态”。通过高速搜集数据，利用流计算、图计算等技术及时处理信息，以可视化图像形式将决策信息实时、动态地展示出来，使公众清楚了解政府动态的决策过程。

### 3.2 面临的挑战

**政府决策的主体缺乏大数据观念，数据挖掘能力不足。**在管理和技术层面上，各级政府部门的数据还缺乏统一标准，在决策时应用大数据决策的观念仍然不高，政府决策过程中的对大数据的挖掘能力还有待改进。

**政府决策过程中制度性障碍，导致数据碎片化、孤岛化，影响数据关联性和数据共享。**目前政府决策过程中，传统政府决策体制仍然发挥着重要作用，其基本特征就是政府决策过程实行垂直化领导，各部门之间各自为政，且上下级之间权责关系分明，遇事逐级请示汇报，在此基础上进行相应的决策部署，这种政府决策体制在推进社会经济有序发展中起到了积极的作用，但也存在许多弊端，如机构庞大、决策时间长、决策成本高，缺乏协同交流，回应社会与公众的诉求反应迟缓。条块分割的管理体制还造成政府决策系统内部数据往往自成体系，相互割裂，不能做到数据互通有无和资源共享，导致目前政府掌握的数据大多处于割裂和休眠碎片化状态，阻断了政府决策过程中相关数据的共享和交流，极大浪费了丰富的数据资源。

**专业人才短缺成为政府发展大数据决策掣肘。**目前政府决策过程仍然以传统的关系数据库系统为工具处理公务。大数据是海量、多模式和低价值密度的信息资产，需要新的处理模式促成更强的决策力和优化处理能力。需要从业者具备高级分析领域的技能和经验，目前非常缺乏，给政府大数据决策带来了特殊的挑战。

### 3.3 对策建议

**树立大数据思维，提升大数据在政府决策中的战略地位。**在政府决策过程中融入大数据思维。大数据思维就是改变传统的思维方式，以大数据为基础，以大数据分析进行思考和分析问题的思维方法。大数据思维能够直观呈现数据背后的人类行为模式。运用大数据提升政府决策能力，需要把这种大数据思维融入政府决策理念，改变政府决策的理念维度、决策方式和决策思路。在大数据视角下，人们面对的是呈现出混杂特征的全体数据，这些数据相互之间不再是简单的因果关系，而是多重混合交叉的相关关系。传统的政府决策理念往往是根据一定技术方法选择有限社会样本数据的抽象分析方法，是用对社会少部分人的需求来推断出社会大多数人的需求和行为模式，必须变革这种不适应大数据时代要求的决策理念，这种变革就是要把大数据思维的价。值判断和大数据分析方法论融入到现代政府决策理念中，将对基于有限样本数据的决策，转向覆盖更全面、涉及多数人的大数据分析，从数据收集者转向数据分析者，从大数据关联分析中预测社会需求，预判社会问题，探索政府决策的多元主体、实时分布式及价值关联特征，提升政府决策能力。

**完善制度和机制，形成有利于政府大数据决策的环境。**在大数据环境下政府决策面临的种种不适和挑战，迫切需要完善政府决策相关制度和机制，为提升政府科学决策奠定坚实的基础。

（1）建立首席数据官制度。其主要职责是根据政府机构的政策制定规划和某一阶段性任务需要，建立相关决策数据库，选择决策数据的提取、转换和分析等工具，进行大数据的数据挖掘、处理和相关性分析，并对决策提供相应的建议。要让首席数据官进入决策层，使其成为政府大数据决策的中坚力量。

（2）建立政府数据开放和共享机制。按照开放、保障安全、价值导向、质量保障和权责利统一的原则，建立数据开放和共享的相关法规和制度,明确政府数据共享机制的责任主体，打破科层制模式下不必要的数据壁垒，为提升决策实效提供帮助。

（3）建构大数据在政府决策领域的安全机制。大数据时代个人隐私的控制与风险已经成为一个核心问题，对隐私信息进行保护是政府数据管理的职责，要建立数据监管机构，建立数据归集与使用的规范制度。在数据监管上，着重于创立数据使用者为其行为承担责任的新型隐私保护模式，设立大数据审计员制度，在使用预测分析时考虑自我监管评估数据源的选择，数据预测分析工具的选取，包括运算模型和法则，以及运算结果的解读是否正确合理，以确保政府决策过程的大数据应用安全无隐患。

**以大数据为核心资源，构建政府的现代决策模式。**政府科学决策必须做到与时俱进，大数据时代的到来，必将意味着大数据技术及其应用都将是政府科学决策的必然趋势。现代政府如何清晰把握大数据时代特征和演化规律是现代政府科学管理决策所必须应对的，因此建构以大数据为核心资源的现代决策模式是大数据环境下政府决策的内在要求。 要清醒地认识到这一点，达到以大数据为核心资源的政府科学决策转变，就要着力这两个层面：第一，要进一步强化科学决策体系综合数据系统。也就是说，加强各类非结构化数据、包括非数值型数据存储系统建设力度从而对政府决策统合数据系统进行有效管理，确保大数据下政府对决策进程地把握，同时积极开发政府决策层面的数据挖掘系统，严谨良好的数据治理方式。进一步提升涉及国家重大安全领域的数据采集、分析、处理系统。加强对计算机基础设施建设的投入力度，积极高效地建设政府科学决策生态系统网络尤其是云计算平台硬件系统，从提高对数据的处理效率，为政府决策能力提升提供大数据支撑服务。第二，提高政府决策科学性还要有数据创新，以大数据为基础的决策流程的不断创新和优化，通过挖掘整理生态系统中相互合作的信息数据、决策的反馈数据及社会的参与行为数据等关键资源，有针对性地修补欠缺和改进决策流程，从而确保政府科学决策能力的不断提高。

**把握信息制高点，强化大数据应用的自主能力。**大数据时代要实现政府决策能力提升，就要建构以大数据为核心资源的现代决策模式。迎接大数据时代的挑战，其首要任务就是要把握信息制高点，强化大数据应用下政府的自主把控能力。斯诺登泄密事件爆发，足以警醒我们大数据下的安全是多么重要。大数据时代到来，既为政府建设决策能力带来了契机，也同时告诫我们安全警钟要时时敲响。 “科学技术是第一生产力”，大数据是一个互联网、云计算，物联网等诸多因素在内的数据分析和处理能力模式，其涉及的技术层面非常多，例如集成学习、关联规则的挖掘。为确保国家安全和个体安全不受影响，“把重要的数据存储在自家的仓库里”，当下必须把大数据技术研发的支持力度摆在重要位置，政府要高度重视，力推核心技术掌握和自主创新，努力强化政府科学决策过程中大数据应用的自主技术含量，确保政府决策的大数据管理和应用真正建在自主技术的根基上，从而把握信息制高点。基于此还要培育大量的数据科学家和数据架构师，提高人才质量，并将其选送到政府科学决策工作岗位上来，组成一支高素质的政府科学决策大数据分析队伍，通过严格的考核造就一批政治过硬，又熟悉大数据关键技术的复合型人才，使其在政府决策岗位充分发挥作用，为政府科学决策能力提升作出积极贡献。

## 4 大数据支撑宁波市科研管理的实践

### 4.1基于大数据的知识产权资源布局信息与可视化平台概况

平台利用数据的关联化、层次化、图谱化技术，对全市专利数据、产业资源数据、教育数据、核心企业信息、地理信息等进行采集、整理，构建以知识产权为核心的新经济科技咨询数据库，研发知识产权资源布局信息与可视化平台。通过厘清知识产权资源、科教资源和产业资源的匹配关系，加强创新活动中信息数据和配套政策的协调，帮助政府科学决策和企业精准治理，最终实现创新资源集聚，统筹推进知识产权与经济社会发展深度融合。

### 4.2平台架构

平台将解决数据采集、基础数据库、布局地图和决策分析四个关键问题，实现分析结果的动态可视化。平台的整体架构将使用典型多层多阶架构模型，设计为数据资源层、检索服务层、分析应用层和业务表现层四个相互独立又相互联系的层次。同时，提供非功能性保障和运维管理功能。

一是对各类数据进行格式规范，包括专利、教育、科技、产业、人才和政策等基础数据库；二是在具备完善功能的同时，开发可拓展性，各个子模块特点突出，功能既要相对独立，又要能够根据分析需求与其他模块实现数据的关联和分析的协同；三是要以需求为导向，运用百度Echarts、阿里DataV等大数据可视化工具，对各个基础数据库的分析结果进行动态展示，满足用户的检索、分析和可视化需求。



**图1 知识产权资源布局信息与可视化平台研发思路图**

### 4.3意义和预期成效

针对我市科技服务业发展模式创新不足、科技创新支撑不足等突出问题，项目通过搭建知识产权资源布局信息与可视化平台，实现知识产权、人才、科技、教育、成果、企业等数据的动态联动，对全市重点产业领域、县市区之间的知识产权资源与创新资源进行匹配，进一步提升宁波市的科技服务业的创新能力。可扩展性和可视化分析方法是平台的主要特色，面向用户增加调整的接口，使用户可以根据需求建立子数据库，开展检索、分析和可视化工作。

预期形成以知识产权为核心的新经济科技咨询数据库，覆盖不少于10个前沿技术领域，数据量不少于10万条；形成一套基于大数据的科技咨询工具，建成基于大数据的知识产权资源布局信息与可视化平台，形成符合大数据特点的咨询交互模式，提供特色化的大数据咨询服务产品，服务和应用覆盖5个以上行业领域，累计服务10000项次以上，通过知识产权领域的公共服务便捷化和政府治理精细化，完善城市创新生态，支撑城市产业发展。