# 大数据时代的机遇与挑战

大数据泛指巨量的数据集，因可从中挖掘出有价值的信息而受到重视。《华尔街日报》将大数据时代、智能化生产和无线网络革命称为引领未来繁荣的三大技术变革。麦肯锡公司的报告指出数据是一种生产资料，大数据是下一个创新、竞争、生产力提高的前沿。世界经济论坛的报告认定大数据为新财富，价值堪比石油。因此，发达国家纷纷将开发利用大数据作为夺取新一轮竞争制高点的重要抓手。

**大数据时代的来临**

互联网特别是移动互联网的发展，加快了信息化向社会经济各方面、大众日常生活的渗透。有资料显示，1998 年全球网民平均每月使用流量是 1MB（兆字节）,2000 年是 10MB ,2003 年是 100MB ,2008 年是 1GB( 1GB 等于 1024MB)，2014 年将是 10GB。全网流量累计达到 1EB（即 10 亿 GB 或 1000PB）的时间在 2001 年是一年，在 2004 年是一个月，在 2007 年是一周，而 2013 年仅需一天，即一天产生的信息量可刻满 1.88 亿张 DVD 光盘。我国网民数居世界之首，每天产生的数据量也位于世界前列。淘宝网站每天有超过数千万笔交易，单日数据产生量超过 50TB（1TB 等于 1000GB），存储量 40PB(1PB 等于 1000TB)。百度公司目前数据总量接近 1000PB，存储网页数量接近 1 万亿页，每天大约要处理 60 亿次搜索请求，几十 PB 数据。一个 8Mbps（兆比特每秒）的摄像头一小时能产生 3.6GB 数据，一个城市若安装几十万个交通和安防摄像头，每月产生的数据量将达几十 PB。医院也是数据产生集中的地方。现在，一个病人的 CT 影像数据量达几十 GB，而全国每年门诊人数以数十亿计，并且他们的信息需要长时间保存。总之，大数据存在于各行各业，一个大数据时代正在到来。

信息爆炸不自今日起 ,但近年来人们更加感受到大数据的来势迅猛。一方面，网民数量不断增加 ,另一方面 ,以物联网和家电为代表的联网设备数量增长更快。2007 年全球有 5 亿个设备联网，人均 0.1 个；2013 年全球将有 500 亿个设备联

网，人均 70 个。随着宽带化的发展，人均网络接入带宽和流量也迅速提升。全球新产生数据年增 40%，即信息总量每两年就可以翻番，这一趋势还将持续。目前，单一数据集容量超过几十 TB 甚至数 PB 已不罕见，其规模大到无法在容许的时间内用常规软件工具对其内容进行抓取、管理和处理。

数据规模越大，处理的难度也越大，但对其进行挖掘可能得到的价值更大，这就是大数据热的原因。首先，大数据反映舆情和民意。网民在网上产生的海量数据，记录着他们的思想、行为乃至情感，这是信息时代现实社会与网络空间深度融合的产物，蕴含着丰富的内涵和很多规律性信息。根据中国互联网络信息中心统计，2012 年底我国网民数为 5.64 亿，手机网民为 4.2 亿，通过分析相关数据，可以了解大众需求、诉求和意见。其次，企业和政府的信息系统每天源源不断产生大量数据。根据赛门铁克公司的调研报告，全球企业的信息存储总量已达 2.2ZB（1ZB 等于 1000EB），年增 67%。医院、学校和银行等也都会收集和存储大量信息。政府可以部署传感器等感知单元，收集环境和社会管理所需的信息。2011 年，英国《自然》杂志曾出版专刊指出，倘若能够更有效地组织和使用大数据，人类将得到更多的机会发挥科学技术对社会发展的巨大推动作用。

**大数据应用的领域**

大数据技术可运用到各行各业。宏观经济方面，IBM 日本公司建立经济指标预测系统，从互联网新闻中搜索影响制造业的 480 项经济数据，计算采购经理人指数的预测值。印第安纳大学利用谷歌公司提供的心情分析工具，从近千万条网民留言中归纳出六种心情，进而对道琼斯工业指数的变化进行预测，准确率达到 87%。制造业方面，华尔街对冲基金依据购物网站的顾客评论，分析企业产品销售状况；一些企业利用大数据分析实现对采购和合理库存量的管理，通过分析网上数据了解客户需求、掌握市场动向。有资料显示，全球零售商因盲目进货导致的销售损失每年达 1000 亿美元，这方面的数据分析大有作为。

在农业领域，硅谷有个气候公司，从美国气象局等数据库中获得几十年的天气数据，将各地降雨、气温、土壤状况与历年农作物产量的相关度做成精密图表，预测农场来年产量，向农户出售个性化保险。在商业领域，沃尔玛公司通过分析

销售数据，了解顾客购物习惯，得出适合搭配在一起出售的商品，还可从中细分顾客群体，提供个性化服务。在金融领域，华尔街“德温特资本市场”公司分析3.4 亿微博账户留言，判断民众情绪，依据人们高兴时买股票、焦虑时抛售股票的规律，决定公司股票的买入或卖出。阿里公司根据在淘宝网上中小企业的交易状况筛选出财务健康和讲究诚信的企业，对他们发放无需担保的贷款。目前已放贷 300 多亿元，坏账率仅 0.3%。

在医疗保健领域，“谷歌流感趋势”项目依据网民搜索内容分析全球范围内流感等病疫传播状况，与美国疾病控制和预防中心提供的报告对比，追踪疾病的精确率达到 97%。社交网络为许多慢性病患者提供临床症状交流和诊治经验分享平台，医生借此可获得在医院通常得不到的临床效果统计数据。基于对人体基因的大数据分析，可以实现对症下药的个性化治疗。在社会安全管理领域，通过对手机数据的挖掘，可以分析实时动态的流动人口来源、出行，实时交通客流信息及拥堵情况。利用短信、微博、微信和搜索引擎，可以收集热点事件，挖掘舆情，还可以追踪造谣信息的源头。美国麻省理工学院通过对十万多人手机的通话、短信和空间位置等信息进行处理，提取人们行为的时空规律性，进行犯罪预测。在科学研究领域，基于密集数据分析的科学发现成为继实验科学、理论科学和计算科学之后的第四个范例，基于大数据分析的材料基因组学和合成生物学等正在兴起。

麦肯锡公司 2011 年报告推测，如果把大数据用于美国的医疗保健，一年产生潜在价值 3000 亿美元，用于欧洲的公共管理可获得年度潜在价值 2500 亿欧元；服务提供商利用个人位置数据可获得潜在的消费者年度盈余 6000 亿美元；利用大数据分析，零售商可增加运营利润 60%，制造业设备装配成本会减少 50%。

**大数据技术的挑战和启示**

目前，大数据技术的运用仍存在一些困难与挑战，体现在大数据挖掘的四个环节中。首先在数据收集方面。要对来自网络包括物联网和机构信息系统的数据附上时空标志，去伪存真，尽可能收集异源甚至是异构的数据，必要时还可与历史数据对照，多角度验证数据的全面性和可信性。其次是数据存储。要达到低成

本、低能耗、高可靠性目标，通常要用到冗余配置、分布化和云计算技术，在存储时要按照一定规则对数据进行分类，通过过滤和去重，减少存储量，同时加入便于日后检索的标签。第三是数据处理。有些行业的数据涉及上百个参数，其复杂性不仅体现在数据样本本身，更体现在多源异构、多实体和多空间之间的交互动态性，难以用传统的方法描述与度量，处理的复杂度很大，需要将高维图像等多媒体数据降维后度量与处理，利用上下文关联进行语义分析，从大量动态而且可能是模棱两可的数据中综合信息，并导出可理解的内容。第四是结果的可视化呈现，使结果更直观以便于洞察。目前，尽管计算机智能化有了很大进步，但还只能针对小规模、有结构或类结构的数据进行分析，谈不上深层次的数据挖掘，现有的数据挖掘算法在不同行业中难以通用。

大数据技术的运用前景是十分光明的。当前，我国正处在全面建成小康社会征程中，工业化、信息化、城镇化、农业现代化任务很重，建设下一代信息基础设施，发展现代信息技术产业体系，健全信息安全保障体系，推进信息网络技术广泛运用，是实现四化同步发展的保证。大数据分析对我们深刻领会世情和国情，把握规律，实现科学发展，做出科学决策具有重要意义，我们必须重新认识数据的重要价值。

为了开发大数据这一金矿，我们要做的工作还很多。首先，大数据分析需要有大数据的技术与产品支持。发达国家一些信息技术（IT）企业已提前发力，通过加大开发力度和兼并等多种手段，努力向成为大数据解决方案提供商转型。国外一些企业打出免费承接大数据分析的招牌，既是为了练兵，也是为了获取情报。过分依赖国外的大数据分析技术与平台，难以回避信息泄密风险。有些日常生活信息看似无关紧要，其实从中也可摸到国家经济和社会脉搏。因此，我们需要有自主可控的大数据技术与产品。美国政府 2012 年 3 月发布《大数据研究与发展倡议》，这是继 1993 年宣布“信息高速公路”之后又一重大科技部署，联邦政府和一些部委已安排资金用于大数据开发。我们与发达国家有不少差距，更需要国家政策支持。

中国人口居世界首位，将会成为产生数据量最多的国家，但我们对数据保存不够重视，对存储数据的利用率也不高。此外，我国一些部门和机构拥有大量数据却不愿与其他部门共享，导致信息不完整或重复投资。政府应通过体制机制改革打破数据割据与封锁，应注重公开信息，应重视数据挖掘。美国联邦政府建立统一数据开放门户网站，为社会提供信息服务并鼓励挖掘与利用。例如，提供各地天气与航班延误的关系，推动航空公司提升正点率。

大数据的挖掘与利用应当有法可依。去年底全国人大通过的加强网络信息保护的决定是一个好的开始，当前要尽快制定“信息公开法”以适应大数据时代的到来。现在很多机构和企业拥有大量客户信息。应当既鼓励面向群体、服务社会的数据挖掘，又要防止侵犯个体隐私；既提倡数据共享，又要防止数据被滥用。此外，还需要界定数据挖掘、利用的权限和范围。大数据系统本身的安全性也是值得特别关注的，要注意技术安全性和管理制度安全性并重，防止信息被损坏、篡改、泄露或被窃，保护公民和国家的信息安全。

大数据时代呼唤创新型人才。盖特纳咨询公司预测大数据将为全球带来 440 万个 IT 新岗位和上千万个非 IT 岗位。麦肯锡公司预测美国到 2018 年需要深度数据分析人才 44 万—49 万，缺口 14 万—19 万人；需要既熟悉本单位需求又了解大数据技术与应用的管理者 150 万，这方面的人才缺口更大。中国是人才大国，但能理解与应用大数据的创新人才更是稀缺资源。

大数据是新一代信息技术的集中反映，是一个应用驱动性很强的服务领域，是具有无穷潜力的新兴产业领域；目前，其标准和产业格局尚未形成，这是我国实现跨越式发展的宝贵机会。我们要从战略上重视大数据的开发利用，将它作为转变经济增长方式的有效抓手，但要注意科学规划，切忌一哄而上。