大数据产业调研及分析报告

数据堂（北京）科技股份有限公司

大数据的出现，引发了全球范围内深刻的技术与商业变革，已经成为全球发展的趋势以及国家和企业间的竞争焦点，直接关系到国家安全、社会稳定、经济发展和民生幸福等诸多方面。对于中国而言，在经历了 30 多年的高速发展之后，各种严峻问题也伴随而生，包括产业升级、社会稳定、环境保护、医疗健康和食品安全等方面的挑战。因此，亟需将大数据的发展提升到战略高度，以此为契机，通过各种创新和探索，推动产业升级和创新、经济转型和民生建设。

本报告主要以启发性和独创性为主线，选择国内外最具典型意义的案例进行描述，尽可能地从数据源、分析方法和价值实现等角度体现出大数据的真正内涵，并对我国大数据产业的发展提出相应的建议。

（一） 全球及我国大数据产业链论述

当前，大数据在政府决策、交通、物流、金融、广告、电信、医疗、娱乐和农业等领域的应用蓬勃发展，据 IDC 预测，全球大数据市场规模年增长率达 40%，在 2017 年将达 530 亿美元。本报告以大数据产业链划分（彭博）为框架，对国内外大数据产业链条进行了全面梳理，其中收录了近 300 余家国内大数据企业和应用。

（二） 我国大数据产业现状分析

绝大部分拥有数据的企业都在分析挖掘的基础上对外提供服务；垂直领域内的数据链条在孕育和发展，但是在所有纯数据源企业或平台，只有不到 8%在开展数据的租售业务；政府/公共服务、农业和医疗健康领域的应用案例相对缺乏另一方面，在政府/公共服务、农业和医疗健康领域的应用案例相对缺乏，尤其是政府和公共服务事业单位沉淀的海量数据未能与广大传统行业的需求形成对接。

（三）我国发展大数据产业发展的建议

以大数据供需两端（数据源和应用环节）为抓手实现重点突破，大力推动全社会的数据开放，尤其是政府数据的开放，力争在短期内降低全社会的数据获取成本并起到显著的社会示范效应。

我国幅员辽阔、人口众多，交通、医疗、金融及农业等事关国计民生的领域汇集了海量的人口、个体行为和环境数据，通过人工智能技术的应用可以极大带动政府决策、公共服务和传统行业的发展，同时培育数据银行和众包平台等产业模式的创新。

目录

（一）大数据综述

1.1 大数据概念溯源

数据来自一切客观存在，包括宏观到微观的物理世界，各种生物体，人类社会活动，人类感知、认识和思维的结果。随着信息技术的发展，当前通常所说的数据是指经过数字化转换后的信息，是可被量化、分析和再利用的信息，包括数值、文字、符号、音频、视频等形态。

对数据的分析并非新鲜事物。交通规划、宏观经济分析、电力系统规划、气象预测、高能物理、航空航天、基因工程等大规模数据的分析和计算早已在人类生产生活中发挥着关键的作用。1970 年哈佛大学关于资源三角形的论述中，将材料、能源、信息看成是推动社会发展的三种基本资源。因此，传统的商业智能和数据库厂商得以出现并获得快速发展。

而大数据概念的出现，是以信息技术的发展和应用为主线的：

- 数据规模和类型的剧变。互联网和移动互联网的发展、传感技术的广泛应用，使得数据的规模和种类急剧增长。数据类型不仅包含关系型数据，还出现了大量的日志、文本、图片、音频和传感器非结构化和半结构化数据。数据呈指数级增长态势，据麦肯锡全球研究院（MGI）预测，2020 年产生的数据量将是 2009 年的 44 倍，接近 35ZB（1ZB=10^{21}Byte）。
- 数据存储成本下降。单位信息存储成本的下降，使得对海量数据

的分布式存储技术难度降低。30 年前，1TB 存储的成本大约是 16 亿美元，如今通过云存储服务所需不到 100 美元。

- 大规模数据处理成为可能。计算能力不断发展、对非结构化数据处理和分析方法的逐渐成熟、MapReduce 模型以及云计算模式的出现，使大规模数据处理的成本和技术门槛大为降低。
- 数据的采集更为密集和广泛。人类活动和自然环境的各类数据被广泛地采集和记录，其中蕴含的信息和知识可以极大推动人类社会的发展。据预测,2020 年物联网传感器的数量将达到 500 亿个。
- 数据分析应用的发展。Google（海量数据的分析利用）和沃尔玛公司（啤酒与尿布的关联销售）的数据分析经典案例给业界带来的冲击。

以上因素，使得学术界和企业界开始思考新时代下数据分析所能带来的巨大价值，所谓大数据的概念得以引爆并且逐渐为人所熟知：

- 《自然》杂志在 2008 年 9 月推出了名为“大数据”的封面专栏，讲述了数据在数学、物理、生物、工程及社会经济等多学科扮演的愈加重要的角色；
- 《科学》杂志 2011 年推出大数据专刊，将大数据深度分析看成未来研究的突破点；
- 2011 年 6 月，麦肯锡发布研究报告《大数据：下一个创新、竞争和生产力的前沿领域》，研究了当下全球数据的状态，并阐述了挖掘这些数据能够释放出的潜在价值。

对于大数据的概念，至今没有一个为业界所广泛接受的明确定义。各界纷纷给出了关于大数据概念的描述：

- 麦肯锡：大数据是指其大小超出了典型数据库软件的采集、储存、管理和分析等能力的数据集；
- 维基百科：无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的大量而复杂的数据集合；
- Gartner：体量大、快速和多样化的信息资产，需用高效率和创新型的信息技术加以处理，以提高发现洞察、做出决策和优化流程的能力。
- Forrester：大数据本质在于“数据存储、处理和访问的流程与业务目标的集成”。

在以上定义中，更具高度的是 Gartner 和 Forrester 的定义，不再局限于数据本身进行描述，而是将大数据概念提升到了数据价值实现的层面。数据价值的实现来自于对数据的分析，Teradata 和波士顿咨询（BCG）以此为主线，直观地阐明了传统数据分析向大数据时代演进的历程：

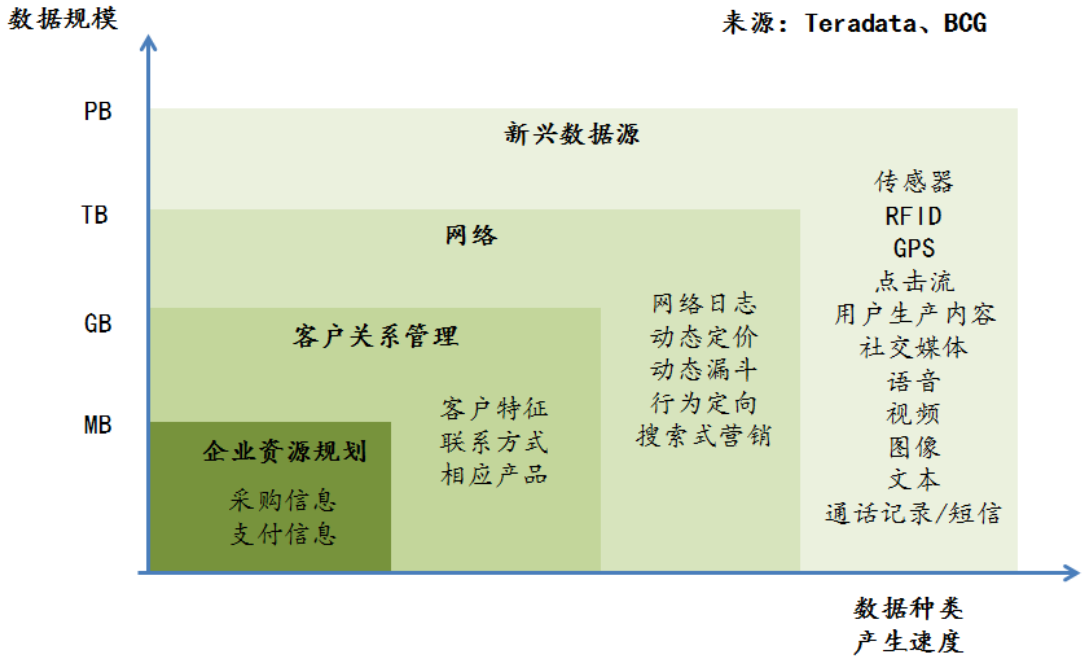


图.数据分析演进历程（Teradata）

作为一个超越具体技术和业务领域的词汇，追求对大数据的准确定义是没有意义的。更具现实意义的是理解大数据所带来的数据分析模式的演变，即引入更为丰富的数据源，对海量、异构数据进行快速的处理和分析，进而挖掘出传统数据分析方法所无法获得的信息和知识。在具体的分析思路上，大数据不再追求用尽可能少的数据来构建模型，而是尽可能追求数据的全面性，同时在不放弃因果性探索的基础上强调了对数据之间相关性的重视。

随着大数据分析在互联网和零售业等行业中的经典应用案例不断涌现，越来越多的政府、企业和研究机构开始意识到数据的资产属性，数据分析能力正在成为未来的核心竞争力。而大数据不但能够带动信息产业本身的发展，更能在人类生产生活的各个方面引发深刻的技术与商业变革，本报告将在后续部分展开对于大数据应用场景的论述。

1.2 大数据产业的战略意义

上个世纪信息科技的迅猛发展导致了人类生产生活模式的电子化和数字化，其主要作用在于效率的提升，而在大数据时代，关注的重点逐渐转移到数据本身，人类寄希望于从海量的各种数据中萃取具有真正价值的信息和知识，并形成对未来发展的准确的预测。大数据的出现，引发了全球范围内深刻的技术与商业变革，已经成为全球发展的趋势，国家和企业间的竞争焦点正从资本、土地、人口、能源转向数据资源。一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，直接影响到国家安全、社会稳定、经济发展和民生幸福等诸多方面。

大数据引发巨大社会和经济变革的潜力，得到了各国政府、全球学术界和工业界的高度关注和重视。美国、日本以及欧洲一些发达国家政府都从国家战略层面提出了一系列的大数据发展计划，以推动全社会对大数据技术和应用的探索和研究。对于中国而言，在经历了30 多年的高速发展之后，各种严峻问题也伴随而生，包括产业升级、社会稳定、环境保护、医疗健康和食品安全等方面的挑战。因此，亟需将大数据的发展提升到战略高度，以此为契机，通过各种创新和探索，推动产业升级和创新、经济转型和民生建设。

我国的大数据发展不仅是时代的需要，更有着得天独厚的基础优势。我国庞大的人口和经济规模为大数据发展提供了肥沃的土壤，也为理念、技术和模式的创新提供了无限的可能性。可见，大力发展大数据产业，主动掌握新一代信息技术产业发展的主动权，推动整个国

家和社会的良性、可持续发展，是以大数据为代表的第三次产业革命带给我国的历史契机。

（二）全球大数据产业分析

大数据涵盖数据从产生到最终被分析利用的各个环节，其中所涉及的相关技术都可以被称为大数据技术，而对数据施加影响的各方则共同构成了大数据产业链。

根据 IDC 的报告显示，全球大数据市场规模年增长率达 40%，在 2017 年将达 530 亿美元。其中，大数据技术及服务市场复合年增长率（CAGR）将达 31.7%，2016 年收入将达 238 亿美元，其增速约为信息通信技术（ICT）市场整体增速的七倍之多。

当前各界对大数据产业链的划分有诸多版本，其中逻辑相对清晰的刻画来自于彭博发布的研究报告，将大数据产业分为六大区块，包括数据源类、基础设施类、分析类、应用类、跨基础设施类和开源项目类。本报告依据此划分进行阐述，但所引述的大数据应用和探索案例并不限于彭博的报告内容，在地域上也不局限于北美地区。

大数据的定义没有明确的限定和边界，能够归入大数据范畴的案例数不胜数，本文主要以启发性和独创性为主线，选择最具典型意义的案例进行描述，希望尽可能地从数据源、分析方法和价值实现等角度体现出大数据的真正内涵。

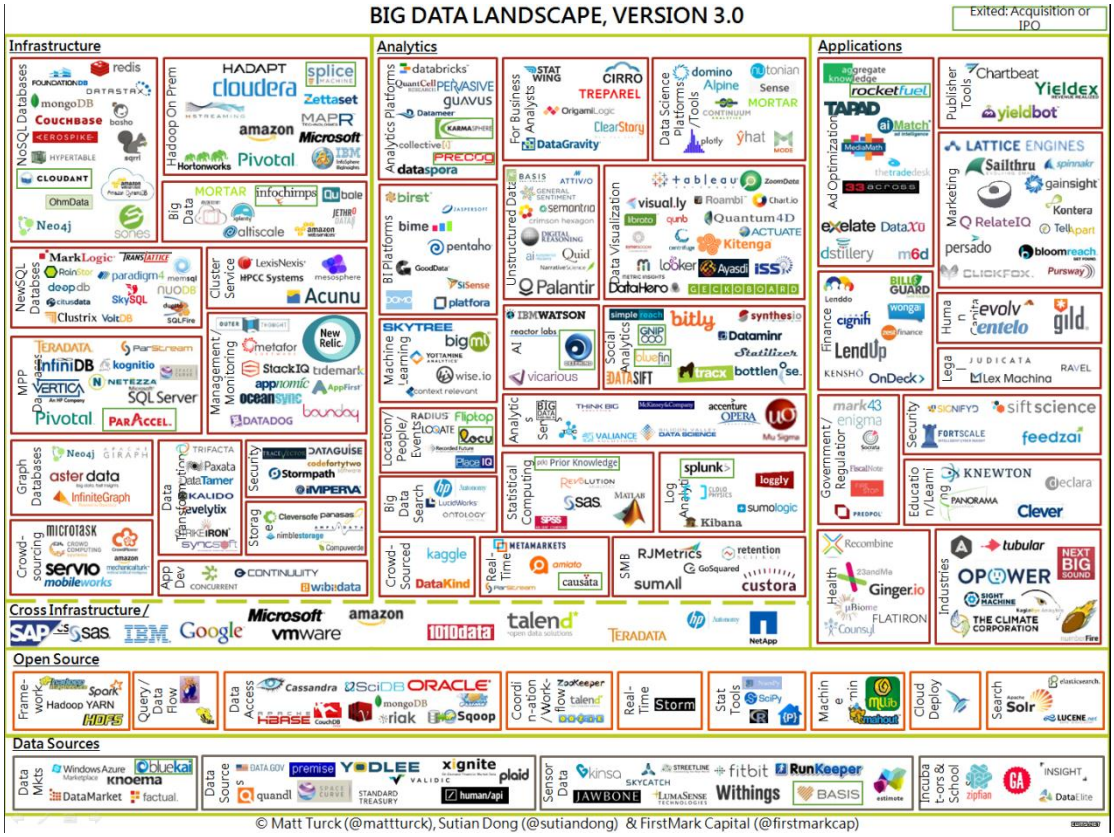


图. 大数据产业链分布（彭博）

2.1 数据源（Data sources）

本区块内的企业基于自身业务产生或采集了大量数据，并通过租售等方式直接对外交付数据，还包括纯粹提供数据交易平台的企业。判断企业是否属于数据源的关键在于，这类企业的客户还需要对所获得的数据进行分析和挖掘才能对决策形成真正的支持。

大数据与传统数据分析理念的一大区别就在于强调数据的外部性，即数据离开了其产生和消费的传统路径，为其他行业或领域所用。数据外部性的典型场景包括：电信运营商和政府合作，可以在交通运输、市政规划和人口统计等方面发挥作用；金融数据和电商数据结合，可以用于诸如小微贷款一类的金融产品和服务；物流数据和电商数据相

结合，可以勾勒出经济领域的宏观和微观运行情况；农业和气象数据应用到金融领域，可以为农业保险和理赔提供高价值的信息；遥感卫星数据与耕地抽样数据相结合，可以打破传统的统计路径，实现更为客观的粮食产量统计；电表数据可供房地产行业进行空置率的估算。

数据源类企业就是实现数据外部性的基础渠道，在对各类数据进行采集和整合之后，提供给各行各业进行目的和方法各不相同的分析和挖掘，使数据的价值得以充分实现。比如 Bluekai 公司收集和销售的用户数据包括：

数据类型	描述	分类维度
互联网	有消费倾向的客户	财务：贷款、投资、借贷 旅游：起始地点、停留日期、交通工具、酒店 零售：类型和品牌
B2B	企业	规模、行业
购物历史	持续购物的预期	产品和品牌
地理/人口统计特征	以位置或特征相近的人群	地理：按州分 人口特征：家庭收入、性别、教育程度、孩子数目
娱乐/偏好	有特定兴趣爱好的人群	产品类型 生活方式：高消费、旅游、运动 世代：战后婴儿潮
经济状况	客户的财务信息	收入、开销

图.Bluekai 公司售卖的数据

总体而言，数据源区块内的企业可分为数据交易、产生、采集和聚合几大类：

- 彭博社（Bloomberg）和路透社（Thomson Reuters）采集并整合金融相关数据，然后提供给金融机构。
- 安客诚（Acxiom）通过聚合超市、药店、专卖店等企业的客户数

据，经过加工之后转卖给所需的企业。

- BlueKai、Lotame、RapLeaf 等企业搜集并出售客户的上网行为数据，主要提供给广告业客户。
- AggData 和 Datafiniti 定位为数据的聚合者，将来自网络的不同来源的数据聚合在一起，并提供下载服务。
- Opera solutions，本身不拥有数据，而是通过购买或搜集用户的行为信息(如征信数据、医疗就诊记录等),再销售给所需的企业。
- Factual 定位于各类数据的交易平台，尤其是地理位置相关的数据集。
- InfoChimps 定位于各类数据的交易平台，尤其是地理位置、社交网络、网络信息等方面的数据。
- Datamarket 为客户提供国民经济与工业相关的数据集。
- Yodlee 聚合并提供私人银行财务数据。
- SureScripts 主要采集医院的处方数据。
- Sermo.com 为提供分享医疗见解的平台，通过收费模式允许医药公司访问数据。
- Moovit 通过众包方式采集公共交通信息，包括负载信息和公交车准点信息。
- 租车公司 Zipcar 通过车辆内置系统,采集乘车人和车辆本身的数据。
- 旅游网站 Tripadvisor，提供平台供用户发布自己对景点、饭店和酒店的评论，形成了一个高价值的旅游相关产业数据源。

- Truecaller，通过读取用户手机上的通讯簿，采集全球的电话号码，并与相应的社交媒体关联，为用户提供联系信息搜索服务。

2.1.1 模式创新

数据银行

在大数据时代，数据已经成为一种资产，企业、组织和个人开始普遍认知到自身所拥有数据的外部价值，数据价值挖潜的概念在全社会发酵。与金融资产类似，数据资产的供给和需求方分别对数据资产存在着管理和融资的需求，因此在大数据产业链的数据源区块，孕育着一种类似银行性质的产业形态，即数据银行。

从融资角度看，数据银行的本质在于实现数据供需的对接，为数据所有者找到数据资产价值变现的出口，同时使需求方获得自身业务发展所需的数据资源。与传统银行或交易平台不同的是，数据银行并不是简单地对供需进行对接和撮合：数据资产类型各异，没有统一的形式和度量，数据银行必须积极从多个源头引入资源，以专业的知识和技能建立原始数据和最终应用之间的通路，让数据需求方可以顺利以市场化的手段获得所需的数据资源。

从资产管理的角度看，数据银行的本质在于为数据资源的价值体现提供便利。传统银行所管理的金融资产，其形态和内容已经为资产供需双方所普遍理解，而数据资产尤其是原始数据的价值需要通过各种手段主动挖掘才能体现，且不同的需求方对于同一数据的价值认知也存在较大偏差。因此，在数据价值实现和数据资产流通的过程中必

须进行数据清洗、格式化以及必要的应用场景开发等工作，并且以相应的机制设计和技术实现，聚合数据的供给和需求，确保数据资产接入、发布和访问通道的顺畅。

数据的原始形态通常与应用方的需求有一定差距，因此更准确地说，数据资源必须经过有目的的发掘和特定的处理环节，才能真正实现资产化。相对于数据交易平台一类的形态，数据银行最大的差异点在于需要对数据的转化和规整。数据的质量直接与企业成本负相关，据 Gartner 公司估算，数据混杂的 CRM 系统将使企业收入下降 25%，而 Experian 公司则认为不准确的数据很可能导致公司的收入平均损失 12%。通过脱敏、去噪和去重处理，以及针对文本、图像和音视频等海量非结构化数据的标注和特征抽取，能够将数据转化为需求方易于理解和利用的形式，降低数据分析和挖掘的难度。这一环节是数据资产融资和管理的关键，是提取和挖掘数据所含高价值信息和知识的前提，是大数据技术发展和应用开拓的核心推动力。

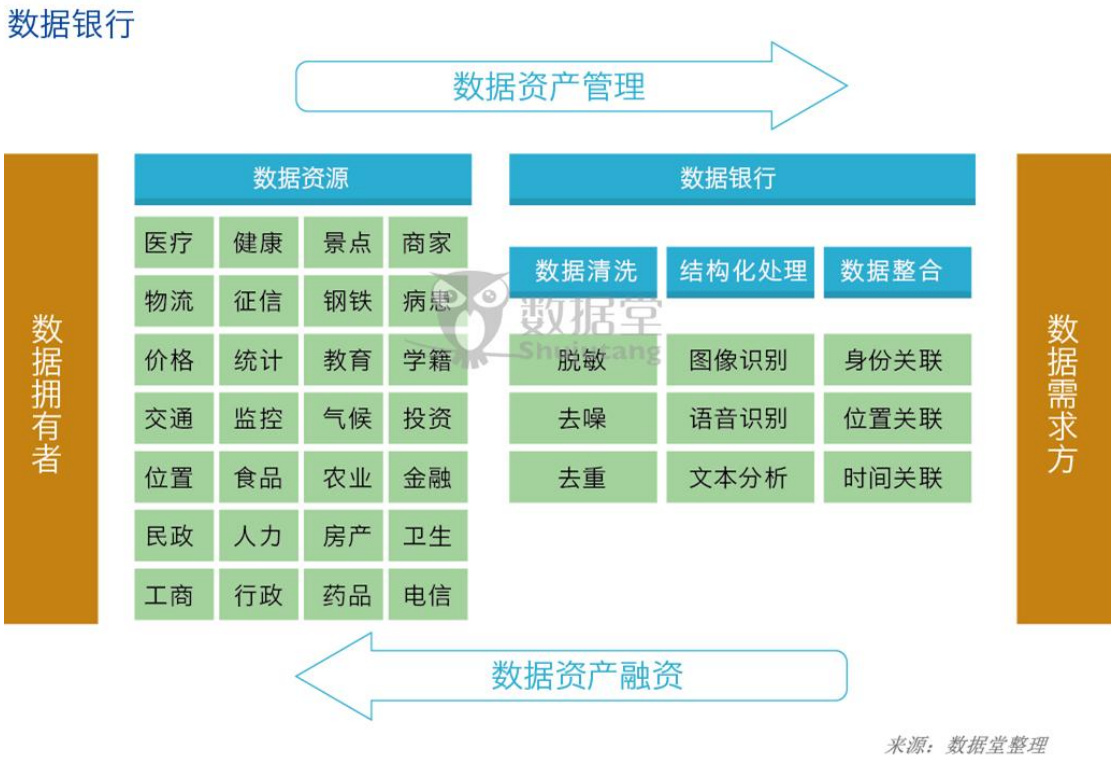


图. 数据银行模式

综上所述，数据银行的本质就在于实现了数据资源的商品化、标准化、资产化，承载着数据资产形成、管理和交易的职责。由于欧美先进国家在数据利用方面有较深传统，在多个行业内自发形成了数据流通的渠道，数据资源商品化、标准化和资产化的动作已渗透到大数据产业链的各个环节中，全社会对于数据银行或综合性数据交易平台的需求并不突出。但是，对于信息产业相对落后、数据利用意识比较淡薄的国家和地区，数据银行形态的培育非常重要，从而在市场供给和需求之间对数据资源进行深度资产化改造，使数据真正成为大数据产业以及其他行业和领域发展的助推器。

众包模式

随着大数据的应用扩展和深化，社会对数据的需求日益上升，现

有的数据获取渠道和方式已无法满足科学研究和产业发展的需求，更为便捷的获取方式、更为广泛的覆盖面和更为真实准确的数据已成为迫切的需要。随着互联网、移动互联网和智能设备的发展，每个人都成为一个潜在的数据采集点，导致众包（crowdsourcing）模式成为大数据时代一种极具现实意义的数据采集方式。

众包这一概念由美国《连线》杂志的记者杰夫•豪（JeffHowe）在2006 年 6 月明确提出，指把工作任务以自由、自愿的形式外包给大众的方法，通常用于完成那些耗费大量人力的繁重任务。众包的思路并非新鲜事物，18 世纪英国就通过向民间征集海洋经度的精确测量方法，解决了牛顿、惠更斯和哈雷等著名科学家未能解决的问题，获奖者是一位来自于乡村的木匠。此后的几个世纪中，类似的方法曾经有力地推动了航空、计算机等行业的发展。

随着互联网的出现，众包的覆盖范围和可参与度都大大提升，日益成为一种可行的商业模式与组织方式。以当前的技术发展情况而言，数据的采集、标注和清理等重复性工作还很难完全实现自动化，比如人体特征的采集、图片的标注和重复数据的剔除等，而这些工作所需耗用的人力随着数据量的剧增而成为企业或组织难以承受的重负。对于这些需要大量人力介入才能保证质量的工作，众包模式提供了一种成本可控、规模易伸缩的实现途径。比较典型的案例有：

- Twitter 使用亚马逊的众包平台 Mechanical Turk，来响应用户对热点话题的搜索查询。
- 《国家地理》曾发动近 2.8 万人在蒙古的卫星图像中搜寻成吉思

汗的墓地。

- 澳大利亚昆士兰的公交乘客用随身应用采集信息（比如公车到站时间等），提升市民的通勤效率，2011 年已经可以做到通知下一班车的到站时间。
- 个人手工艺术品网站 Etsy 发动用户来鉴定新发布的手工品是否存在版权侵权的问题。
- Foursquare(据报道,将被雅虎以9亿美元的价格收购)和Factual等公司让企业用户自己提交地理位置信息的做法，也属于众包采集的范畴。
- 亚马逊与移动打车应用Flywheel合作,呼叫小型配送中心附近的出租车来为用户递送包裹。

在大数据时代，由于需要采集海量的底层原始数据，在成本可接受的范围内，很多时候已无法基于现有采集设备来完成任务，因此众包模式在大数据产业中最重要的应用场景就是数据的采集。同时，海量数据的加工和标注等任务所需的人力和时间太高，使得众包模式在数据处理环节也具有较大的应用空间。

除了采集和加工等高人力和时间消耗的任务之外，通过众包模式也可以将需要高智力和技术水平的问题外包给大众，通过受众面的扩大来提高任务完成的效率。比如，将原始数据公布于众，让公众积极参与到对数据的分析挖掘和应用创新活动中，能够有效推动大数据技术和产业的发展。

2.2 基础架构（Infrastructure）

与传统 IT 基础架构相比，大数据基础架构必须应对空前规模的数据和各类音频、图像、视频和文本等非结构化信息；互联网、移动互联网和物联网数据的指数级增长，使得基础架构必须拥有高度的可扩展性；为了快速应对变化、响应市场，实时分析的需求日益强烈，基础架构必须具有强大的数据吞吐和计算能力。

基础架构区块中的企业主要提供大数据的存储和管理产品或服务，为后续的分析和挖掘提供支撑，包含各类新兴的 NoSQL、NewSQL、MPP 和图数据库，以及云服务、数据转换工具、管理/监控工具和存储设备等。

- Neo4j。图形数据库，将结构化数据以图结构进行存储，具备完全的事物特性。
- Aster data。MPP 数据库，起源于斯坦福大学，已被 Teradata 收购。
- Cloudera。基于 Hadoop 的产品与解决方案提供商。
- MapR。基于 Hadoop 的产品与解决方案提供商，用自身文件系统取代 HDFS，实现高速、镜像、快照等功能。
- Cleversafe。分布式存储产品，为提升系统吞吐率优化了 HDFS 的副本设置。
- VoltDB。内存数据库产品，NewSQL 的代表之一。同时满足关系型数据库的 ACID 原则以及 NoSQL 的可扩展性。
- StackIQ。Hadoop 系统管理工具。

- Greenplum。MPP 数据库的代表之一，具有高可扩展性的关系型并行数据库。
- 微软 Dryad。关系型数据库的并行实现，能够将 SQL 语句转化为基于 DAG 的多个操作。
- Box 和 Dropbox。提供大数据存储的云服务。

2.3 跨平台（Cross infrastructure）

本区块中多为提供计算、存储和分析平台或服务的大型厂商，提供对大数据分析进行支撑的软硬件一体化方案。

- IBM 在 DB2 中集成了 BLU 技术、列式优化和并行向量处理等技术，以内存计算大幅提升数据分析效率。在基础平台方面，为 Hadoop 平台提供支持，同时有针对性地对 GPFS 文件系统进行了改造。
- 微软推出了基于 Hadoop 的大数据处理的组件，实现了 SQL Server 与 Hadoop 的连接；推出 LINQ Pack、Project“Daytona”以及 Excel DataScope，让用户可以在 Windows Azure 云上进行大数据分析；2015 年初，微软收购 R 语言的商业版提供商 Revolution Analytics，加强数据分析方面的能力建设。
- SAS 通过与 Hadoop 的集成，为客户提供分布式的分析产品。
- 1010data 公司提供基于云计算平台的数据分析服务。
- Talend 公司针对数据集成提供专业的 ETL 工具和主数据管理云服务。
- 惠普推出了针对 Hadoop 平台优化的 AppSystem for Apache Hadoop，

提供包括底层硬件、Hadoop 和实时数据分析的一体式解决方案。

2.4 开源（Open source）

由企业、高校或科研机构所研发并开源的大数据产品，是当前大数据基础技术发展的最大推动力，通常集中在基础性平台和分析工具两大类。

- Hadoop。起源于雅虎公司，是当前主流的大数据存储和处理平台，实现了分布式的计算框架 MapReduce 和文件存储系统 HDFS。
- Spark。诞生于加州伯克利大学 AMP 实验室，是新一代大数据分布式处理框架，以高效的内存计算著称，逐渐成为大数据处理环节的主流平台。
- MongoDB。由 10gen 公司开发，著名的分布式 NoSQL 数据库，由于功能丰富，在使用方面最接近关系数据库。
- Storm。由推特开发的大数据流式分析解决方案，在接收数据的同时就进行计算和分析，具备一定的故障处理能力。
- Mahout。数据挖掘工具，起源于 Apache 基金会，实现了一个分布式机器学习算法的集合。
- Solr。起源于 Apache Lucene 项目的开源企业搜索平台，功能包括全文检索、命中标示和分面搜索等。

2.5 分析（Analytics）

除了存储，大数据管理的另一项大的挑战是数据分析，只有通过

分析才能获取智能、深入、有价值的信息。数据分析大致可以分为以下几类：数据挖掘，大数据分析的理论核心，基于不同的数据类型和格式呈现出数据的各种特性，挖掘其中蕴涵的价值；预测性分析，大数据分析最重要的应用领域之一，通过训练数据建立模型，并以此为基础预测未来的趋势和走向；非结构化分析，针对海量的音频、图像、视频和文本数据，结合人工智能技术抽取和提炼，使之能够用于后续的分析的挖掘；可视化分析，直观的呈现数据统计分布特性，使普通用户能够对数据形成大致的理解。

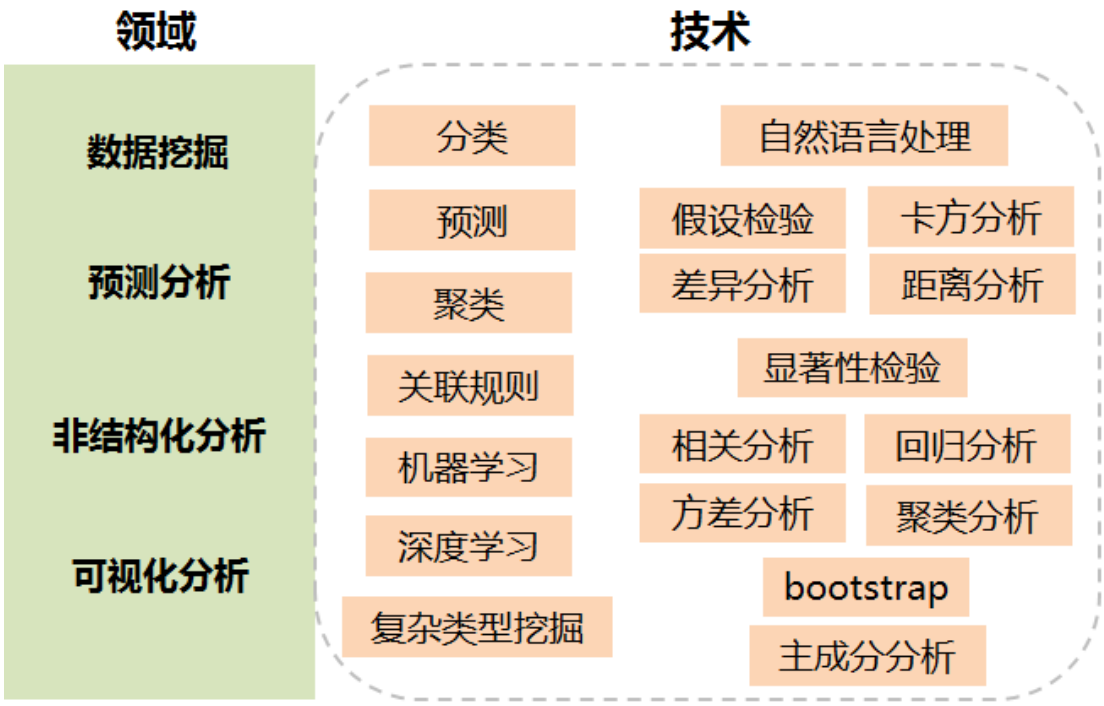


图. 分析领域及技术

在大数据海量、多源、异构特性和实时处理的需求推动下，专门针对大数据的分析工具蓬勃发展，尤其是对于非结构化数据以及对未来进行预测的分析。从创业和获得投资的情况来看，本区块是整个大数据产业链中最为活跃的部分，其中典型的企业或产品如下所述：

- 甲骨文宣布收购 Endeca Technologies，为企业用户提供非结构化数据的搜索和管理服务。
- SAP 推出了 Hana 平台，能够对非结构化数据进行高速分析，是大数据内存计算的代表性技术之一。
- Google 推出企业级大数据分析云服务 BigQuery，用来在云端处理大数据，帮助企业在云平台上分析数据、构建应用和分享服务。
- Dataminr 公司通过对社会化媒体进行分析，为金融行业与政府部门提供服务，曾在股市下跌前三分钟，预知黑莓股票将被抛售。
- Splunk。专注于日志分析，首家上市的新兴大数据公司，首个交易日市值即突破 30 亿美元。
- Palantir Technologies。数据分析工具提供商，旗下 Palantir Gotham 系统主要用于反恐，协助美国中央情报局发现本·拉登行踪；Palantir Metropolis 则主要为对冲基金、银行等提供服务。
- 为行业尤其是政府、金融等行业提供解决方案，2014 年底以 150 亿美元估值完成融资。
- Clear story 为客户提供数据整合工具，可以整合包括 Facebook 在内的多种数据源。
- Affectiva 专注于人脸表情识别，商业媒体评为发展最快的创业公司之一。2012 年美国总统竞选期间，Affectiva 追踪人们观看奥巴马和罗姆尼辩论片段的表情，结果以 73%正确率判断出了选民投票结果。

2.5.1 可视化

图形是直观呈现数据的直接方法，数据可视化就是研究如何利用图形，展现数据中隐含的信息，发掘数据中所包含的规律。数据可视化所需的专业知识横跨计算机、统计和心理学。随着大数据的发展，海量的数据需要以直观、便捷的方式展示给技术和业务人员。

大数据可视化分析领域的典型公司包括：

- Risk Management Solutions 用热图来直观标示自然灾害的风险和类别，包括地震、龙卷风、飓风、暴风、森林大火和火山爆发等，进而为保险公司提供自然灾难风险模型，供客户估算理赔风险。
- Compuware 公司每天采集 80 亿个数据点，对外提供 Web 服务故障热图，监测全球 1500 个 Web 服务。
- RetailNext 基于店内的摄像头、Wi-Fi 和其他探测设备所采集的数据，用热图显示顾客在商店内的实际行走模式，超市或零售店家可以据此来摆放货物或评估促销活动的实际效果。
- DOMO 公司为企业整合多源头数据并以高度可视化的形式呈现出来，为管理人员的决策提供支持，估值高达 20 亿美元。

2.6 应用（Application）

掌握数据资产的企业群是大数据的首批和直接受益者，可以方便地对大数据进行加工、消化、利用。而随着应用价值的逐步体现及大数据产业的发展，应用将必然扩张到传统产业的方方面面，不断创造

新的应用场景。

在全球范围内，大数据的应用已经具备了初步的实践基础，在政府决策、公共服务、影视娱乐、交通物流、医疗健康、金融、电信、人力资源、零售、广告营销、农业、能源等领域得到了较为深入的应用，下面就是当前全球大数据应用的各主要场景。

2.6.1 影视/娱乐

传统上,影视娱乐行业所积累的数据主要集中在票房、收视记录、演职人员基本资料等。基于这些数据，相关人员已经将简单的数据分析应用于市场预测和主角人员选定等方面。随着互联网、移动互联网和有线电视网的发展，以及数据采集、存储和处理技术的进步，影视娱乐业产生了大量文本、图片或UGC数据，其中蕴含了海量的用户对于影视剧作品的偏好信息和观看数据,结合了其他来源的用户数据之后，能够进一步加深用户洞察，提升用户黏性。

基于更为准确和具有时效性的影视剧收视和票房数据，能够分析各地区、人群对不同题材的偏好，了解观影的时段分布，从而支持市场预测、播出平台和投放时间的选择和发行方案制定等。同时，收集主创班底参与过的影视作品基本资料及市场表现，可以构建导演、演员、编剧等主创人员的评估模型，优化团队结构，降低投资风险。

当前，大数据在影视娱乐业的典型应用案例有:

- 视频服务提供商Netflix基于其广大的影视租赁用户群数据，通过偏好分析，搭建了《纸牌屋》的主创班底，成为大数据应用的

早期经典案例。

- 调查公司 Rentrak 基于机顶盒数据，监测各种屏幕上的媒体消费情况，为影视制作公司和广告公司提供咨询服务。
- UnitedTalentAgency 通过 Twitter、YouTube、Tumblr、Facebook、Instagram 和电影类博客等渠道获取数据，评估电影受欢迎的程度，为 20 世纪福克斯公司和索尼影业等巨头提供咨询服务。
- Pandora、Rithm、Spotify 等通过对客户的音乐偏好分析，为消费者提供个性化推荐服务。
- 博彩业巨头凯撒娱乐（Caesars Entertainment）分析客户的网页点击和老虎机游戏记录，提升客户营销和服务的实时性。

2.6.2 交通/物流

随着交通系统信息化程度的加深，以及各种路测和车载智能传感器的普及，大量包含道路、公交、轨道交通、出租汽车、航空、铁路、航运等信息的数据得以产生并被存储下来，可在构建实时、准确、高效的综合交通运输管理系统方面发挥巨大作用。交通基础设施建设和运营涉及大量工程和多个环节，而大数据技术能够对海量信息进行分析，有助于提升交通效率，降低物流成本。在实时监控交通动态的基础上，利用大数据预测模型，可及时预警拥堵情况，协调交通路线。通过传感器感知交通工具的轨迹和周边环境的变化，并通过大数据技术进行数据整合和分析，可以快速建模评估交通情景的安全性，降低交通事故发生的概率。

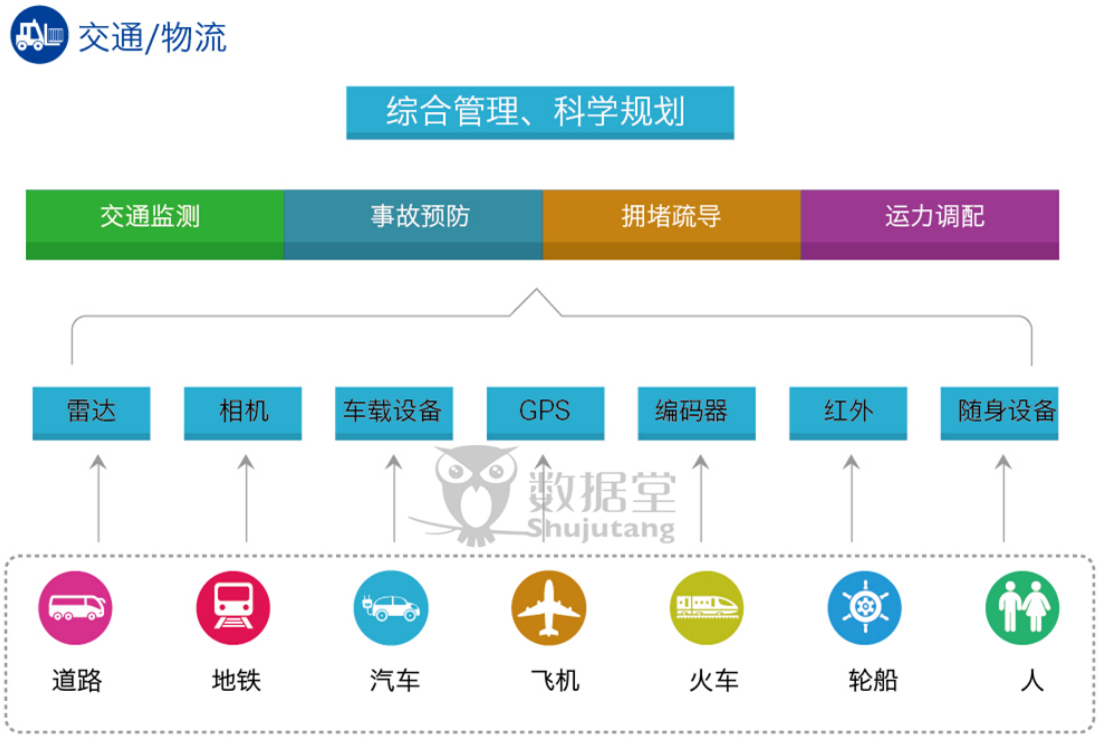


图. 交通/物流领域大数据应用

交通作为人类行为的重要组成和重要条件之一，其相关数据的内在价值极高，对其他行业来说也是大有裨益。最明显的就是与交通运输密切相关的物流行业。物流产业需要对各种物流要素进行优化组合和合理配置，从而提升物流活动效率并降低物流成本。其中物流要素的组合优化必须以全面的数据采集、整合和分析为基础，其最基础的数据源就是交通领域的数据。此外，物流业和大型电商的仓储配送，以及以 Uber 为代表的出租车市场的颠覆者，也都引入了大数据的特性。

当前，大数据在交通和物流领域的典型应用案例有：

- UPS 基于超过 46000 货车上的传感器数据来规划运输线路。数据包括速度、方向、刹车等。同时基于在线地图数据，实时规划每辆车的取货和送货。2011 年，减少了 8500 万英里的行车路程，

从而节省了840万加仑的燃料。早在2000年，UPS就利用这种基于大数据的预测性分析系统来检测全美60000辆车的实时车况，以便及时地进行主动修理。

- 法国电信公司Orange承建了一个法国高速公路数据监测项目，每天产生500万条记录，对这些记录进行大数据分析，为行驶于高速公路上的车辆提供准确及时的信息，有效提高道路通畅率。
- 航班延误时间的分析系统（Flyontime.us），由民间程序员基于美国交通部开放的全美航班起飞、到达、延误数据而开发，向全美国社会免费开放使用，成为很多人乘机、候机的参考信息。
- 亚马逊尝试名为“预判发货”的服务。根据订单、搜索记录、愿望清单、购物车、鼠标在某件商品上的悬留时间等数据，预判用户消费行为，在用户下单之前就将商品发货出库，以此提升配送的速度。
- Uber 根据用户需求的波动开发溢价算法，动态调整出租车价格，缓解特定时段的打车供需矛盾。

2.6.3 医疗健康

医疗行业正处于重要转折点。据麦肯锡预测，2050 年 60 岁以上老龄人口将占全球人口比例的20%，当前医疗费用在全球GDP中的占比逐渐升高，医疗服务市场规模的增长以及日益强烈的个性化医疗服务需求，使得医疗健康行业必须朝着更加精细化和科学化的方向发展，在提供个性化服务、支持临床决策、识别医疗服务相关的欺诈行为以

及分析并应对各类新型疾病等方面，都需要以数据分析为基础。据预测，大数据的应用，能够为全球每人每年带来 1000 美元的费用削减。

医药行业是数据密集的行业，包括药企研发、科研进展、医生诊疗记录、患者病历、检测和用药记录、患者身体状况和保险赔偿记录等各类数据都被持续地记录和存储，作为技术研发、业务决策和服务交付的基础。当前，医疗健康数据呈现下列主要特征：

- 来源广泛。包括制药业、临床治疗信息、医疗费用、病患体征及日常生活记录等。
- 规模膨胀。由于传感器、影像、病检、设备和基因数据的迅速增长，据麦肯锡分析，2020 年医疗数据将增长到 35ZB，相当于 2009 年的 44 倍。
- 类型多样。医疗健康数据即包含结构化的病患档案，又包括大量口述或手写数据、图片和影像等非结构化数据。
- 实时分析。比如，在治疗过程中，需要实时整合、处理和分析不断流入的各类最新信息。

海量医疗数据的积累，开辟了大数据在疫情监测、疾病防控、临床研究、医疗诊断、资源调度和远程医疗顾问等方面广阔的应用空间，而实时分析和图像分析一类的技术需求则进一步提升了大数据进入医疗健康行业的必要性。可见，大数据的引入已成为医疗健康行业进一步发展的必要条件和助推器。据麦肯锡 2010 年测算，大数据将给医疗健康产业带来 3330 亿美金的增值。当前，大数据在医疗健康领

域的主要应用场景包括:

- 医疗数据的结构化。全球医疗数据仅有五分之一为适于计算机处理的结构化数据,其余五分之四为非结构化数据,包括手写病历、各类文档、音视频文件等，其增长速度是结构化数据的 15 倍。
- 优化运营。医疗机构通过对医疗档案数据的转化、整合、统计和分析，实现对管理和监管等环节的优化。
- 新药研制。通过大数据技术，加大临床数据采集力度，运用基因序列分析等先进技术，提升疾病发现和新药研发的效率。通过对产品上市后用药人群分析，检测其疗效和副作用，从而达到提高研发成功率的目的。
- 个性化医疗。在患者就诊时实时整合其体征数据、临床记录和日常生活信息，提供具有针对性的高效医疗服务。

当前，医疗和健康行业的典型大数据企业有:

- HealthFidelity、Explorys、PracticeFusion、athenahealth Inc. 和 Humedica 等,采用自然语言处理技术实现非结构化数据到结构化数据的转化。
- Flatiron 被称为“癌症治疗的基础设施”，通过对临床数据收集整理方法的创新，为医生提供全面而详尽的数据，从而在整体上加速征服疾病的进度。
- Foundation Medicine 采集和分析患者基因组数据，通过特定算法进行突变分析和解读,以临床建议形式辅助医生设定治疗方案。

- Ginger.IO 记录患者的行为和位置移动，帮助护士远程监控诸如糖尿病等类患者的实时情况，以便提醒其停止不利于治疗的行为。
- 美国北卡罗来纳医疗体系（Carolinas HealthCare System）采集 200 多万客户的消费数据，识别其中高风险的患者，比如经常购买酒精饮料的人可能有抑郁症隐患等。
- DNAnexus 公司为医疗行业客户提供 DNA 数据的管理和分析平台。
- Bina Technology、23andMe 以及 Spiral Genetics 则专注于基因测序技术本身的研究。
- 罗氏制药收购基因测序公司 Signature Diagnostics，加速靶向药物的开发。
- IBM 基于超级计算机 Watson，与 WellPoint 合作进行恶性肿瘤的临床诊断，与 BlueCross 合作进行医保数据分析，与 Sloan 癌症中心合作进行癌症研究。

2.6.4 金融

金融业在信息技术和人才等方面相较其他产业具有明显的优势，在开展业务的过程中积累了海量的高价值数据，不但属于数据密集型的行业，而且具有巨大的数据价值变现潜力。比如，据波士顿咨询(BCG)统计，银行业每 100 万美元收入所产生和使用的数据大概是 820GB，远多于其它行业。相应地，数据分析在金融领域的运用历史悠久，比如经济学家很早就利用计量经济学知识和金融市场数据建模预测金融市场产品收益同风险波动的关系。

由于数据种类和规模的增长以及相关技术的发展，大数据在银行、保险公司和券商等金融机构的应用逐渐深入和拓宽，被全面运用于客户画像、精准营销、风险管控、运营优化和市场预测等方面。

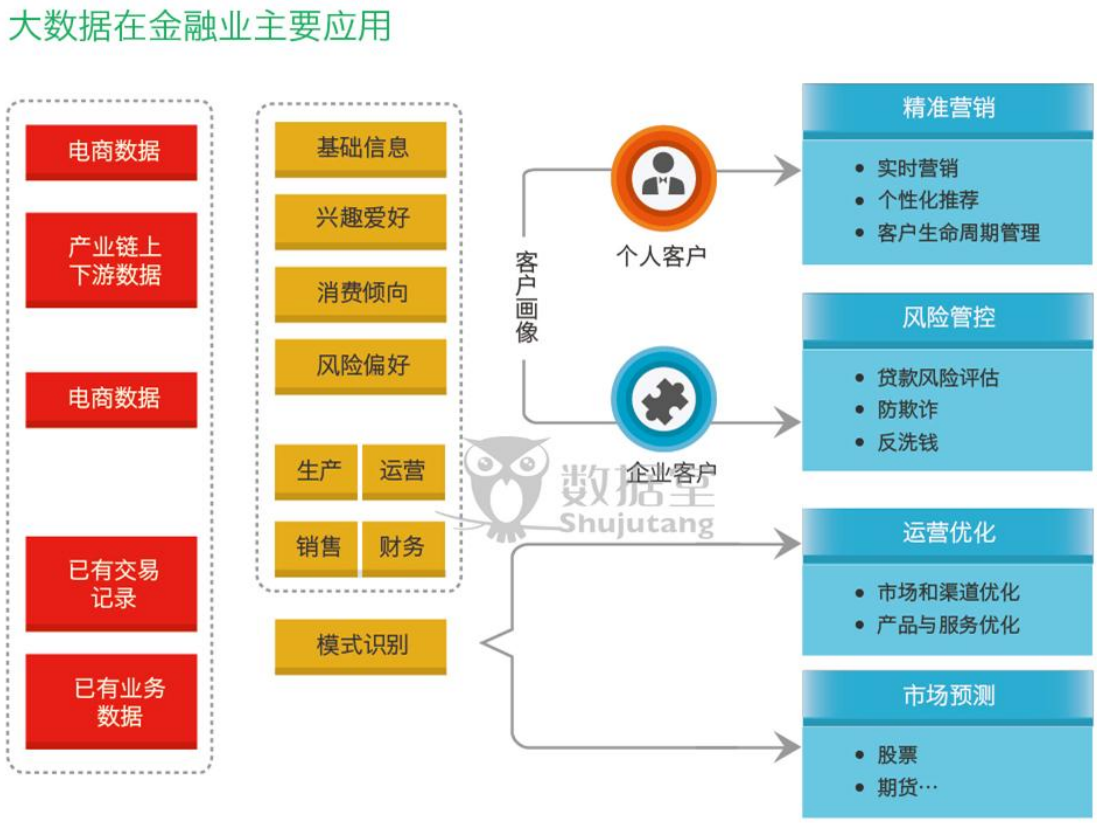


图. 金融大数据应用场景

客户画像。客户画像主要包括对用户的基础信息、社交数据、兴趣爱好、生活习惯、消费习惯等、风险偏好等进行分析，以及对企业用户的生产、运营、财务和销售等情况进行分析。为了更全面的了解客户，除了自身业务所采集到的数据之外，需要整合更多的外部数据，比如客户在社交媒体上的行为数据、客户在电商网站的交易数据、企业客户的产业链上下游数据等。

精准营销。在运用大数据理念和技术对客户全面画像之后，可以通过多种创新提升营销的精度和准度：

（1）实时营销。获取用户的当前状况来提升营销活动的时效性，比如根据客户当前的地点、最近的消费记录和其他信息（工作变化、迁移等）来推送具有针对性的产品或服务；

（2）个性化推荐。根据客户的年龄、资产规模、理财偏好等进行产品或服务的推荐，挖掘出用户潜在的金融服务需求；

（3）客户挽留。根据用户近期的动态构建流失预警模型，对高流失概率的客户提供相应的优惠或个性服务。

风险管控。包括中小企业贷款风险的评估和欺诈识别等：

（1）中小企业贷款风险评估。银行可通过企业的产、流通、销售、财务等相关信息结合大数据挖掘方法进行贷款风险分析，量化企业的信用额度，更有效的开展中小企业贷款。

（2）欺诈交易实时甄别、反洗钱分析、防止恶意透支。银行可以利用持卡人基本信息、卡基本信息、交易历史、客户历史行为模式、正在发生行为模式（如转账）等，结合智能规则引擎（如从一个不经常出现的国家为一个特有用户转账或从一个不熟悉的位置进行在线交易）进行实时的交易反欺诈分析。

运营优化。包括金融机构本身的运营效率优化和产品服务提升：

（1）市场和渠道分析优化。通过大数据，银行可以监控不同市场推广渠道尤其是网络渠道推广的质量，从而进行合作渠道的调整和优化。同时，也可以分析哪些渠道更适合推广哪类银行产品或者服务，从而进行渠道推广策略的优化。

（2）产品和服务优化：银行可以将客户行为转化为信息流，并从

中分析客户的个性特征和风险偏好，更深层次地理解客户的习惯，智能化分析和预测客户需求，从而进行产品创新和服务优化。如兴业银行目前对大数据进行初步分析，通过对还款数据的挖掘比较，区分优质客户，根据客户还款数额的差别，提供差异化的金融产品和服务方式。

市场预测。金融市场价格走势很大程度上受市场情绪的左右，通过对社交媒体类数据的抽取和分析，能够挖掘出其中蕴含的大量市场情绪信息，从而以较高的准确度预测未来金融市场的走向。在金融业与信息业结合最紧密的美国，这方面的研究和实践已经成为热点。比如，华尔街部分交易公司就是通过采集互联网上的政经新闻来预测市场走向；麻省理工学院的学者，根据情绪词将推特内容标定为正面或负面情绪，发现某些负面词汇在推特的占比，都与道琼斯指数、标准普尔500指数、纳斯达克指数的下跌有较强的关联性；美国印第安纳大学利用谷歌公司提供的心情分析工具对道琼斯工业指数变化的预测；美国佩斯大学追踪星巴克、可口可乐和耐克三家公司在社交媒体上的受欢迎程度，发现 Facebook 的粉丝数、推特上的听众数和YouTube上的观看人数都和公司股价密切相关。

当前，全球金融大数据应用的典型案例有：

- 挪威DNB银行将电子渠道所采集的数据与原有数据进行整合，引入Teradata等大数据技术服务提供商构建分析预测模型，以此提升客户满意度并实现对销售活动全天候的追踪。

- 万事达（Mastercard）公司通过大量的数据清洗工作，整合了全球 19 亿张信用卡和 3200 万商家客户信息，基于 Mu Sigma 公司的技术进行欺诈识别和客户洞察分析。
- Zestfinace 突破了传统征信的 FICO 征信模型，主要是将用户的搬家、电话、联系、水电等线下信息纳入到征信模型中，描述每个用户的变量可达 1000 个以上。
- 新加坡星展银行（DBS）与商铺进行合作，当消费者路经店铺时，向其手机推送相应的优惠提示，比如在 20 分钟内使用星展银行信用卡可享受 10%的折扣。
- Paypal 基于 Hadoop、Cassandra 和 Luster 文件系统构建大数据技术栈，并应用于欺诈识别，第一年即挽回多达 7 亿美元的损失。
- 美国 KeyBank 银行通过对网点客户行为的实时监测，定时调配员工并关掉某些非必要的网点，每年因此节省 3500 万美元的运营费用。
- 波士顿咨询（BCG）结合银行业内部数据（现有的网点分布和业绩状况等）和外部数据（各个地区的人口数量、人口结构、收入水平等），帮助澳大利亚某家银行优化网点布局。
- 高盛基于图分析(Graph Analytics)技术，对社交数据进行挖掘，协助合规审查和欺诈行为的识别。2013 年高盛对大数据公司 Applied Predictive Technologies 投资 1 亿美元，以期获得后者所拥有的大量客户数据以及相应的分析模型。
- 摩根大通利用大数据技术追踪盗取客户账号或侵入自动柜员机系

统的罪犯。

- 保险公司 United Healthcare 对客户拨打服务电话的音频数据进行文字化处理，并采用自然语言处理技术识别重点客户和自身业务短板。
- MarketPsych 公司基于路透社的新闻，评估 119 个国家的 18000 多个独立指数，如每分钟的心情状态（乐观、忧郁、快乐、恐惧和生气等），为金融机构提供第三方服务。
- 华尔街的德温特资本市场公司对 3.4 亿社交媒体用户的留言进行情感分析，以大众情绪为指引来决定股票买卖时机。
- 英国对冲基金 DerwentCapital Markets 专门建立了一支对冲基金，通过分析 Twitter 的数据内容来感知市场情绪，指导投资策略。在首月的交易中以 1.85%的收益率超过 0.76%的市场平均业绩。
- Estimize 公司通过众包模式采集大众对于上市公司下一季度每股收益和收入的预测，准确率往往超过专业机构。
- Cignifi 针对手机预付费用户开发风险评估模型，基于付费、通话、上网及其他使用情况预测贷款人的还款意愿和能力。
- Kabbage 使用来自于亚马逊,UPS 和 Intuit 的信用评分模型数据，以及企业销量和客户反馈，评估中小企业的风险等级。
- 美国一家创业公司用电梯数据和黄页数据帮助银行进行风险预警。如果某家公司的电梯数据突然发生异常变化，某种程度上反映了该公司的经营状况，比如电梯停靠次数突然减少可能意味着员工的减少或者客户拜访次数的减少。

互联网金融

除了各类应用场景之外，大数据在金融领域的应用还催生了所谓互联网金融的新兴服务模式。以互联网、云计算和大数据为代表的现代信息科技对金融行业产生了巨大的影响，再加上全球化浪潮的冲击，金融领域内原有的信息壁垒被打破，导致传统的中介开始消解，新型融资模式和平台开始涌现，即互联网金融的概念。

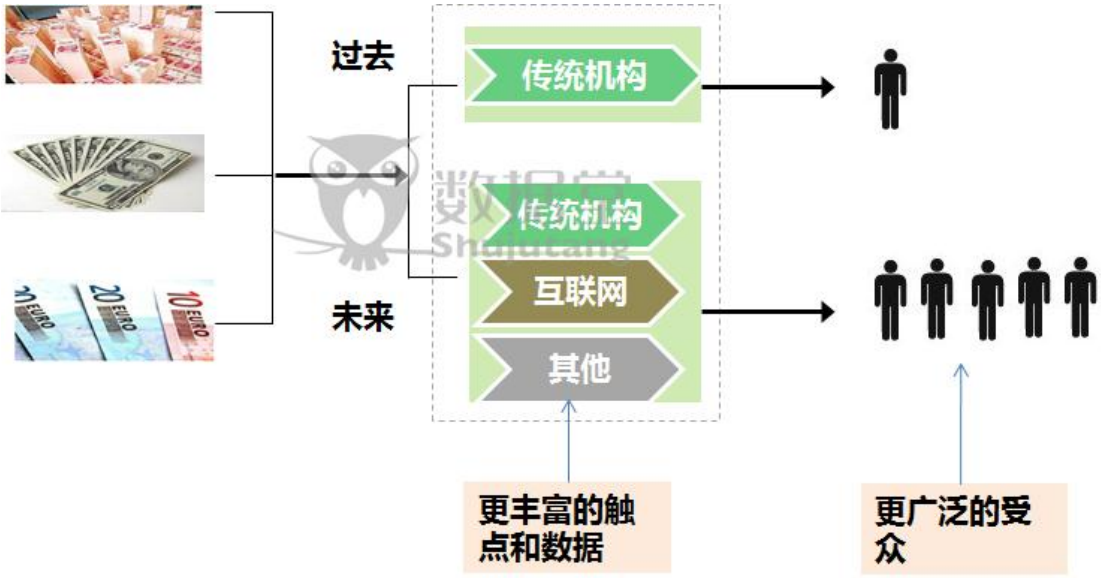


图. 互联网金融

互联网金融的称谓本身从字面上看并不准确，其本质不是将金融产品和服务的交付界面移到互联网或移动应用上，而主要在于以下两点：

- 更全面和准确的企业和个人画像。引入各类网络数据，在很大程度上解决市场信息不对称的问题。网络数据尤其是社交媒体信息蕴含了丰富的企业、组织或个人之间接触、关联和群聚的信息，能够展现财产、经营、消费习惯和商业道德等各个方面的历史和当前状况。在目前的现实规则和客观条件限制下，上述信息中很

大一部分属于个人和机构没有披露义务的范畴，同时采用传统技术手段也难以采集。随着互联网和大数据技术的发展，上述数据的采集和分析成为可能，使得决策的数据支撑面得以扩展，进而极大提升了信用评级和风险控制的效率和准确率，大幅降低全社会的交易成本。

- 更广阔的金融服务市场。网络数据的采集和网络接口的构建，其本质在于通过更丰富多样的数据和客户触点，扩大了金融服务的受众面。据统计，全球范围约有 50%的成年人还没有被金融服务所覆盖，其原因有二：征信方式的落后导致很大一部分民众并未被纳入到信贷业务的覆盖范围内；对于银行记录较少、学生或新移民等人群，征信数据的缺失导致无法对其进行信用评估。而社交媒体、电商消费记录等数据的引入则能够极大地扩大信用评估模型所适用人群的覆盖面，从而实现诸如信贷一类金融服务市场的增容。

互联网金融代表了一种理念和发展趋势，而不是一种企业归类的标准。当前，诸多传统金融机构引入了网络数据并建设了互联网交付的渠道，而众多所谓互联网金融企业也同样需要依赖线下数据。可见，评定企业是否属于互联网金融公司是毫无意义的，真正值得关注的是是否通过新数据、新技术和新触媒提升了风险管理水平、降低了融资成本、改善了用户体验、简化了交易手段。

2.6.5 电信业

运营商所拥有的海量高价值数据，是大数据在电信业应用的主要立足点。电信运营商作为社会信息的基础收录者，比任何行业（也许银行业除外）都更可能了解客户。全世界的 60 亿部手机无时无刻不在产生着各类地理位置、商业活动、搜索历史和社交网络信息，根据思科的预测，2017 年全球移动数据流量将比 2012 年提升 13 倍。而且，随着移动支付的普及，商业活动的信息也能够被移动设备所采集和记录。可见，运营商所处的数据交换中心地位，是 Google 和 Facebook 等巨头无法企及的：

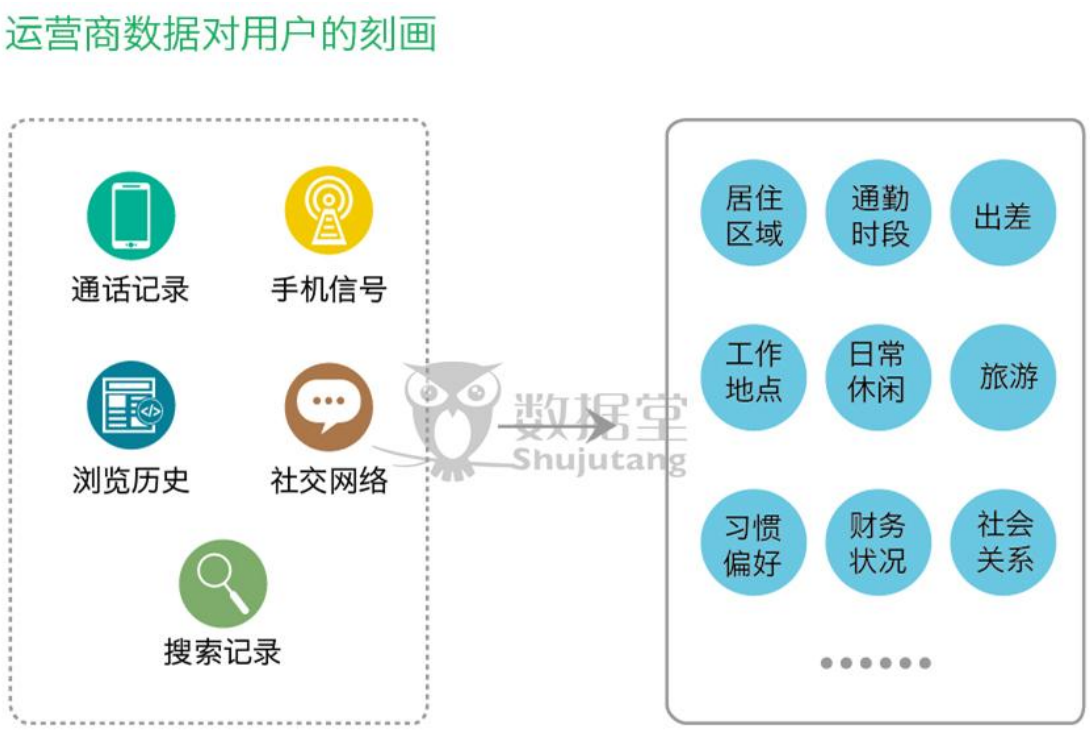


图. 运营商数据对客户的刻画

电信运营商数据的优势主要体现在：

（1）维度丰富。运营商数据涵盖了用户的行为、位置、消费和各类人口统计学特征，是用户画像的物理基础；

（2）群体性强。全球各个国家的电信市场基本都由为数不多的运营商所占据，海量用户的累积导致运营商数据可以充分反映群体客户的特征；

（3）连续性好。由于用户一般不会频繁的切换运营商，数据累积时间相对较长，使得预测模型的建立具有较高的可能性；

（4）网络行为全覆盖。运营商掌握的用户全网行为数据，是确保分析准确性的基础。

（5）关联性强。更进一步，由于丰富维度和全网覆盖性质，运营商数据具备了更为深远的意义，即不同领域数据集之间的跨域关联分析。比如，社交媒体数据、手机通讯记录和网上购物记录等可以通过同一个用户关联起来，进而催生更具想象空间的分析和应用场景。

基于上述优势，除了提升和优化自身业务之外，运营商开展大数据应用的主要方向让自身数据与外部数据产生碰撞，创新业务模式并助力其他行业的发展。对于企业界来说，使用运营商的数据对客户精准画像，可以在客户关系管理、经营决策指导和个性化推荐等方面发挥巨大作用。对于社会管理和公共服务，运营商数据可以广泛用于交通、市政、治安等方面。

当前，全球运营商开展大数据的典型案例包括：

- Sprint 公司将设备位置信息提供给数据集成与分析商 Locately，由后者在汇总分析之后将结果提供给市场营销公司 HAVAS 和 Mobext，最终 Whole Foods、Sears、Target 和沃尔玛等超市连锁获得相应的市场调研报告及咨询建议。

- Verizon 成立 Precision Market Insight 部门，专门为媒体和广告行业、各种活动或赛事的举办者以及零售商提供分析服务。
- SAP 公司从运营商处收集智能手机使用信息和位置数据，并销售给市场营销机构。
- 美国定位服务提供商 AirSage 采集来自运营商基站的实时数据，在经过匿名处理和加密之后，对外提供特定地区内的人群特征分析服务。
- 西班牙电信成立了 Dynamic Insights 部门，基于完全匿名和聚合的移动网络数据，对影响某个时段、某个地点人流量的关键因素进行分析，并将分析结果作为商品出售。
- 日本 NTT 公司利用所掌握的用户信息向金融、电子商务、物联网等周边产业进行扩张。
- 法国运营商 SFR 通过 Intersec 公司的软件方案，分析球赛后观众主要通过哪些地铁站疏散，而后交通部门则根据分析结果在热点地区提供短程的摆渡服务。
- 麻省理工基于意大利罗马的电信运营商数据，分析工作日和休息日的人群流动模式和交通拥堵状况。
- 哈佛大学的流行病学家 Caroline Buckee 基于肯尼亚的 1500 万部手机数据，分析传染病感染人群的流动方向，其成果发表在《科学》杂志上。

2.6.6 人力资源

企业人力资源部门或者人力公司，拥有大量的市场劳动力信息，

员工工作表现信息、企业内部组织和层级信息。在上述数据的基础上，补充以互联网上散落的大量企业和个人碎片信息，能够被人力资源管理者用来对企业或个人进行更为全面的画像，尤其是了解和掌握企业内部运作和个人工作生活相关的不易被搜集的信息，进而更精准的定位人才或对接人力市场供需两端。在具体方法上，生物计量和游戏化等元素的引入，可以极大丰富数据来源，为更加全面和准确的评估奠定基础。大数据在人力资源方面的主要应用场景有：

首先，在人力招聘和职位供需对接方面，通过大数据技术，对前述信息进行分析和挖掘，提升对人才信息的收集、分析和搜索能力，使人力测评由主观性强的人为判断向基于大数据的建模评估方向转变。同时，业界也越来越多地将社交媒体等类信息引入到分析模型中，进一步提升人才评估和职位匹配的准确性。

其次，除了人才招聘和管理范畴之外，人力数据中所蕴含的信息还能够反映企业的运营状况和业务需求。人力资源部门可以基于自身数据的分析挖掘，为管理层提供相应的决策支持。而人力公司则可以通过数据分析，为各行各业提供人力资源管理解决方案和行业宏观分析报告。

当前，大数据在人力资源领域的各类典型案例有：

- LinkedIn 公司基于大数据分析，构建了用户的身份识别系统，比如哪些人在企业里属于关键决策人等。
- Kiran Analytics 公司为美国富国银行提供基于生物计量的员工评测方案，对员工进行分析并预测离职可能性，并对表现优异的

员工进行群体画像，从而指导未来的招聘方案。

- Glassdoor、Simply hired 和 ResumUp 等企业通过各种方式收集企业信息，为求职者提供企业评价、薪水范围、面试问题、招聘启事等信息，达成供需双方的信息对等。
- Pymetrics 通过游戏化的测试方案，对应聘者的性格特质进行分析评估，供招聘企业进行筛选。
- Wanted Analytics 和 Forensic JobStats 为企业提供招聘广告的投放方案，帮助企业快速定位适合的人力资源。
- Identified 公司提供基于 Facebook 的职业搜索引擎，通过工作经历、教育背景和社交网络数据等评估应聘者，类似于 Google PageRank 的人力版。
- Evolv 和 Knack 公司为企业提供招聘，培训和人才数据分析的大数据服务。
- IBM 开发 Professional Marketplace 数据库，包含员工技能、薪资以及近期工作安排等信息，进而通过分析为项目实现最佳的资源配置。2012 年，以 13 亿美元收购人力服务公司 Kenexa，基于后者对 4000 万员工和管理人员的调查问卷，分析特定岗位所需的品质特性。

2.6.7 零售

零售业属于较早引入大数据应用的行业之一。传统上，零售业的数据分析基础是 POS 机数据和竞争对手信息等。随着电脑和移动设备

的购物模式的普及，零售商可以通过网络收集、存储和分析更为广泛的用户信息，从而形成对用户更加深刻的洞察。此外，天气、公共事件信息的引入、客户电话信息的转化，也逐渐成为大数据在零售业应用的基础。

总体而言，目前零售业的大数据应用场景包括：

- 购物体验的个性化。主要是收集和分析 POS、在线交易、社交媒体、客服电话等信息，提升客户画像、个性化推荐以及服务水平；
- 商品布局的优化。主要是收集并分析竞争对手、天气、公共事件等信息，实时调整价格、店面商品布局和物流计划；
- 运营效率提升。根据各种交通、天气、政治和需求信息，形成整个零售网络的全局管理体系，更全面地进行财务决策以及提升防欺诈水平。

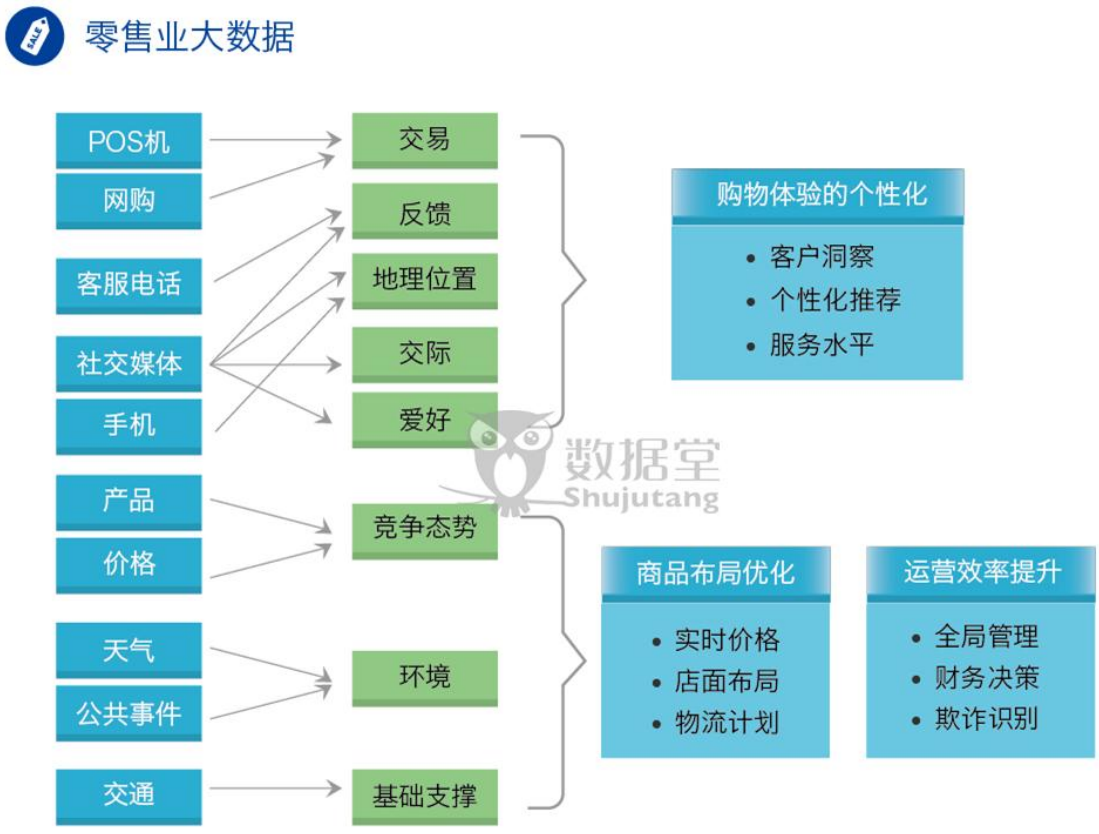


图.零售业大数据应用

大数据在零售业的典型应用案例包括:

- 沃尔玛利用全球各分店产生的海量数据，结合气象信息、经济和人口数据等，对货架、定价、库存和促销进行优化。例如，通过数据分析及时指导库存调整，将一些店面的业绩提升了 40%；同时，其40%以上交易是靠个性推荐转化而成的。沃尔玛在2013年6月收购了大数据预测公司 Inkiru，以此获得所需的分析人才、技术和平台。
- 梅西百货（Macys）基于Hadoop平台，综合运用R、Impala、SAS、Vertica 和 Tableau 等各类分析工具，开发机器学习算法，对企业数据进行分析，提升客户认知水平和个性化推荐的精度。
- 西尔斯百货（Sears）引入Kafka和Storm，对业务数据进行实时处理，将复杂营销活动的准备期从8个礼拜减少到1个礼拜。
- 塔吉特百货（Target）专门成立了 Guest Marketing Analytics 部门，针对每个用户专门建档进行分析，包括与Target相关的信息（如购物记录、刷卡信息、服务电话拨打记录）以及从外部引入的信息（如婚姻、家庭、种族、职业、教育情况、住址、开车到Target所需时间、薪酬、上网记录、阅读喜好和政治观点等）。
- Bloomreach。通过互联网用户分析，预测用户的购物偏好，实现电子商务站点的网页优化。
- RelateIQ。大数据智能关系管理公司，从 Gmail、Facebook 和 LinkedIn 抓取并整合数据，为企业提供更加详细的客户信息。

2.6.8 广告

大数据在广告方面的应用首先是提升广告可见度(viewability),即评估广告是否精准地被送达目标群体。广告可见度的重要性正逐渐被市场认识，根据 ClickZ 公司的统计，有 77%的广告甚至都没有被潜在目标群体看到。

目前，互联网作为新兴媒介承载了巨大的广告投放量，在全球范围内其广告投放金额已超越平面广告,成为仅次于电视媒体的第二大信息传播平台。基于用户的网络浏览和点击数据，结合其他能够反映用户特性的数据，通过大数据分析，专业公司可以对广告位的可见度进行评估，具体而言就是在投放平台、受众群体和投放时间的选择上更加精准。其中，最典型的代表就是自 2010 年以来兴起的实时竞价（Real Time Bid）广告模式，根据海量数据将当前受众与广告精确匹配，实时提升广告的投放精度和成本收益率。与之相对应，一门被称之为计算广告学的分支学科正在兴起，旨在结合文本分析、信息获取、统计学、机器学习、优化以及微观经济学等，研究用户和广告之间的匹配问题。

除了网络数据之外，广告业还可以基于电信运营商数据对线下广告的可见度和效果进行评估。比如对某广告牌附近的日常人流进行分析，结合广告主在广告投放一段时期内的销售数据或网站访问数据，可以评估广告对企业所产生的实际影响。此外，基于有线电视网络收集的数据也能够用于对电视广告的可见度进行评估。

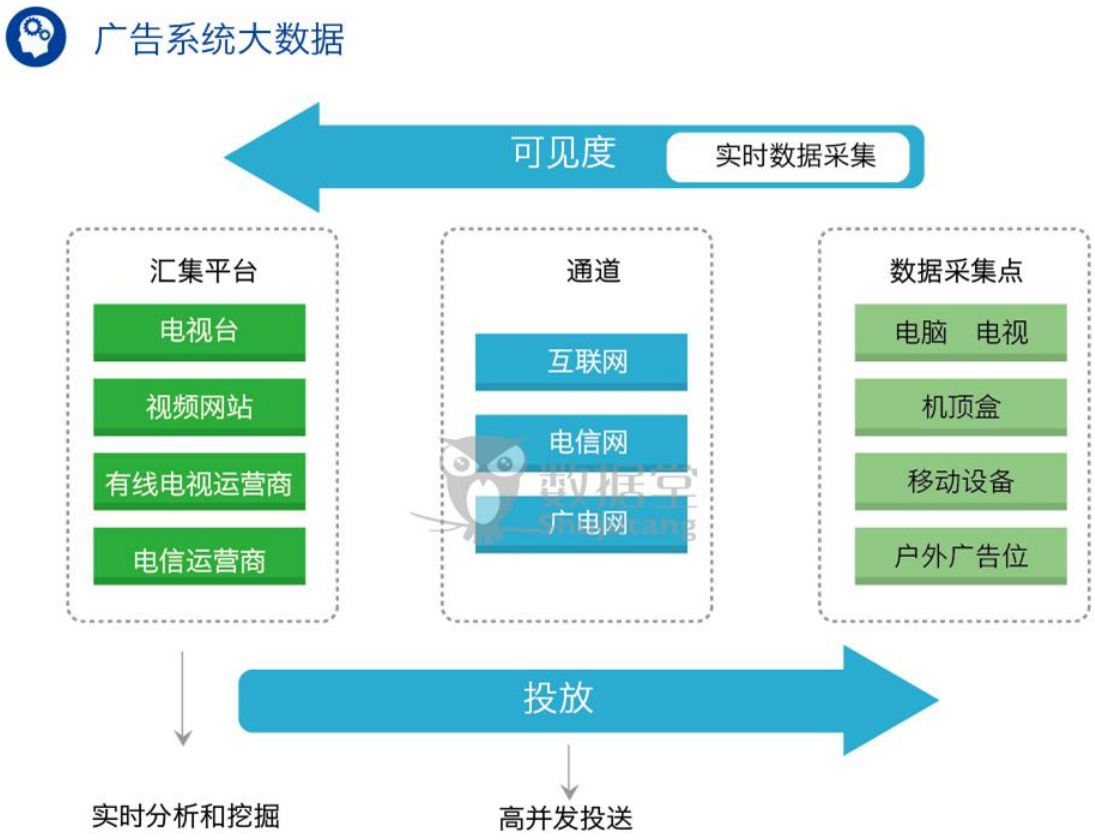


图.广告业大数据应用

目前，将大数据技术应用于广告行业的典型案例有：

- eBay 通过数据挖掘精确计算出广告中的每一个关键字带来的回报。通过对广告投放的优化，2007 年以来 eBay 产品销售的广告费降低了 99%。
- Simpli.fi 公司基于非结构数据，为互联网广告实时竞价提供支撑。
- Integral Ad Science（原 AdSafe）公司基于机器学习算法，评估广告位的可见度。
- Exelate 运用独特的建模算法，为实时竞价广告行业提供数据和分析服务，让广告商和 DSP 可以深入地了解受众属性。
- Route 利用 GPS、眼球追踪软件以及流量模式分析等工具，确定长

椅以及公交车两侧广告的可见度。

2.6.9 农业

大数据在农业领域具有广阔的应用前景，这主要是由农业本身的特点以及当前全球在农业方面面临的各种挑战所决定的：

（1）农业生产周期长、影响因子复杂，导致农业数据涵盖面广、数据源复杂，大数据理念、技术和方法具有极大的应用空间，有助于解决农业领域数据的采集、存储、计算与应用问题。

（2）随着全球人口的增加、气候极端化和能源价格波动的加剧，农业领域对于提升风险管理水平和运营效率的需求日益强烈。传感器、物联网、云计算和大数据等技术的成熟为应对挑战提供了契机，能够推动农业向集约化、精准化和智能化的方向转变。

大数据在农业领域主要有以下应用场景：

- 科学化管理。通过对农田环境数据采集和遥感监测，通过大数据分析，为土质管理、产量评估、病虫害防控和化肥药剂管理提供支持。
- 市场监控。通过对气候、农产品价格、道路交通信息和终端消费等数据的整合与分析，实时监测并评估市场需求和价格变动等情况。
- 精细化耕种。整合有关土壤、水资源、动植物和气候的数据并进行分析，实现耕种模式和方法的精细化。

- 食品安全。实时采集生产、运输和消费等各个环节数据，结合条形码、RFID 等技术，实现对农产品全产业链条的监控和溯源。

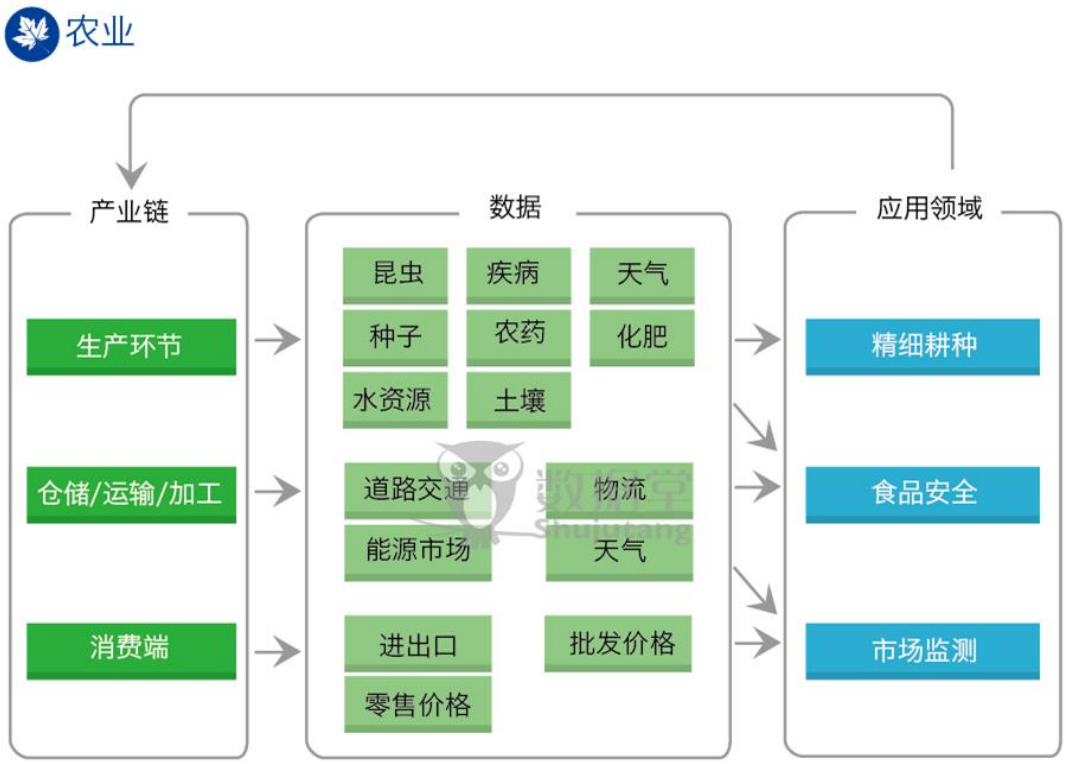


图. 农业大数据示意

当前，在农业领域，最具有大数据内涵和现实意义的应用场景是基于环境数据提升生产、管理、运输等环节的效率。据 IBM 统计，全球 90%以上的农作物损失是由于天气原因造成的，50%的粮食损耗发生在运输环节。准确的天气建模预测可以将农作物损失减少 25%，提升农产品生产和运输的效率。

当前，大数据在全球农业领域的典型应用案例有：

- 孟山都公司用历史的降水和土壤品质数据，帮助农民预测产出和管理风险。2012 年，孟山都花 2.5 亿美元收购的种植技术公司 Precision Planting，实现每块农田上耕种模式的差异化。

- Climate 公司（已被孟山都以 9.3 亿美元收购）每天从 250 万个采集点获取天气数据，并结合大量的天气模拟、海量的植物根部构造和土质分析等信息对意外天气风险做出综合判断，向农民提供农作物保险。
- 芝加哥天气交易所（Chicago Weather Exchange）基于气象数据的采集和分析，提供温度和降雪的期货合约。
- Farmeron 公司为农民提供类似于 Google Analytics 的数据跟踪和分析服务。通过将饲料库存、消耗和花费，畜牧的出生、死亡、产奶等信息进行整合，以分析报告的形式监测农场整体状况。
- Solum 公司通过高效、精准的土壤抽样分析，帮助种植者在正确的时间和地点精确施肥，帮助农民提高产出、降低成本。
- 文莱政府与 IBM 合作，研发了当地天气预报的数据模型，目标是在 2015 年末将本国之前 3%稻米自给率提高到 60%。同时，通过气象学和水文地理学模型预测降雨量、降雨时间和地表水流变化趋势，合理安排排涝、施肥和喷洒杀虫药等生产活动。

2.6.10 企业应用

企业应用主要指采购、库存、生产线、物流、办公、团队协作和客户管理等基础性横向应用，用于协调管理者、员工与部门之间的关系，为企业运营、管理及业务提供全面的综合支持。

超越行业的界限，当前全球企业界面临的共同的挑战和机遇。在金融危机的威胁尚未完全消除的背景下，市场竞争加剧、利润下滑、

企业增长放缓等是企业界必须应对的挑战。而在互联网发展的带动下，新市场、新渠道、新需求和新交付模式的不断涌现，也为各行业带来巨大的市场空间，对企业来说是宝贵的历史机遇。

为了迎接挑战、把握机遇，企业必须贯彻大数据的理念，以数据驱动业务，引入相应的大数据技术，全面了解企业运营状况和客户需求，持续优化运营，提升产品和服务水平，发掘新的业务增长点。

大数据在企业应用领域的切入点主要是实时采集、整合和分析企业内外部的数据(包括供应商信息、采购数据、库存信息、订单信息、销售数据和客户反馈等)，从而形成对企业运营的全面、深刻洞察并对经营决策进行支撑。与企业应用领域传统的数据分析相比，大数据在企业应用中最鲜明的特点是引入了对音频、视频、文档、工作流、日志、社交媒体等非结构化数据的整合和挖掘能力。

不局限于具体垂直行业划分，大数据在企业界的通用性案例有：

- Gainsight 公司整合包括 Salesforce 的多种数据源，对销售日志进行分析，以 SaaS 的模式为企业提供客户挽留与流失预测服务。
- Fractal Analytics 公司基于客户的交易记录和社交媒体内容分析客户特征，帮助企业对客户进行全面画像。
- Aspect 公司开发了语音数据分析平台 Aspect Analytics for Speech and Text，专门针对呼叫中心的非结构化语音数据进行分析，深刻洞察客户的需求。
- Elasticsearch 公司基于开源的 Apache Lucene 系统，为企业提

供内容管理和搜索方案。

- WGBH 电视台采用 RedPoint 公司的技术来实现多渠道用户数据的清洗和整合。
- TideMark 公司为企业客户提供绩效管理和预测分析云服务。

2.6.11 能源

能源行业是国民经济和社会发展的基础，具有多环节、多地域特色，需要对长期负荷以及环境变化的监控、分析和预测，这是大数据在能源行业应用的主要切入点。基于各类传感器所采集的海量数据，能够对能源系统实时的监测和分析，提升运维和服务的效率，优化能耗负载。同时，基于气象数据的分析建模来减少极端天气给基础设施造成的损失。下列是几种能源行业的大数据典型应用场景：

- 在能源供应链上实现信息链的全覆盖，及时掌握上下游的行为和变化，实现能源生产、分配以及消耗的优化，支持能源网络的安全检测与控制（包括灾难预警、调度决策和用量预测）、客户行为分析和精细化运营管理等多方面。
- 基于诸如智能电表等能耗采集设备，能源供应公司可以快速采集分析能源用量，根据能耗高峰和低谷时段的制定不同的价格策略，在平衡了系统负载的同时，降低用户的能耗成本；
- 避免全球气候极端化给能源基础设施造成的损失。据 IBM 推算，截至 2030 年，气候相关的损害将给能源设施带来每年 180 亿美元的损失。因此，准确、高效的天气预测模型将对能源行业带来巨

大的收益。

当前，大数据在能源行业的典型应用案例如下：

- GE 公司基于 Hadoop 和 Amazon Web Services 构建监控平台，用于设备维护、故障预警、节能和流程优化。
- 维斯塔斯风力系统公司（Vestas Wind Systems）利用大数据技术对气象数据进行分析，找出安装风力涡轮机和整个风电场最佳的地点，使以往需要数周的分析工作在不足 1 小时内便可完成。
- The California ISO 公司采集天气、传感器和电表数据，实现美国加州电网的优化调配，预测停电等事故的发生概率。
- Arad 公司在 IBM 的协助下监测自来水管道和水表的实时状况，实现水资源的优化配置。
- Energy Hub 公司与传感器网络厂商 Earth Networks 合作，利用气象数据提高能源供给设施的效率。
- TXU Energy 公司利用智能电表实时采集的数据，通过价格杠杆来平抑用电高峰和低谷的波动幅度，鼓励用户避开用电高峰。
- Grid Navigator 公司为楼宇业主提供能够更好控制能源使用的软件系统。
- 德国部分地区通过智能电网终端每隔五分钟或十分钟收集一次数据，以此预测客户的用电习惯，推断出在未来数月时间里的整体电力需求，进而售卖电力期货。

2.6.12 政府决策与公共服务

大数据分析在经济发展、社会管理、卫生防疫等方面对各国政府都具有极其重要的意义。扩展数据源，引入大数据相关的新兴技术，可以推动政府和公共服务机构从全新的视角审视自身工作，实现公共管理和服务方式方法的创新，进而提升决策水平、社会管理水平和公共服务水平。

大数据在政府和公共服务领域典型应用是统计及后续政策的制定工作。政策制定离不开全面、准确的统计，比如宏观经济运行情况、人口普查、工农业产值和流通环节的价格指数等。相对于传统的方法，以大数据理念采集和分析可以使统计结果更具时效性、更加全面。比如，麻省理工通过在线商品价格预测 CPI 数据，IBM 日本通过搜索引擎热词统计来评估美国的 ISM 制造业指数。

迅猛增长的网络数据背后是相互联系的各种人群，结合心理学、经济学、信息科学等不同学科共同探索网络数据产生、扩散和涌现的基本规律，揭示其后所隐藏的社会动向和矛盾隐患，能够为社会稳定和国家安全提供及时而充分的决策依据。

此外，大数据在提升突发事件反应速度和处理方式上也有较大的发挥空间。政府部门和公共服务机构沉淀了大量的宝贵数据资源，这些数据是整个社会经济活动的数字化记录，运用大数据的理念和技术可以极大地提升工作效率。比如，通过图像识别技术，对海量的监控视频数据进行分析和挖掘，可以及时发现治安隐患。

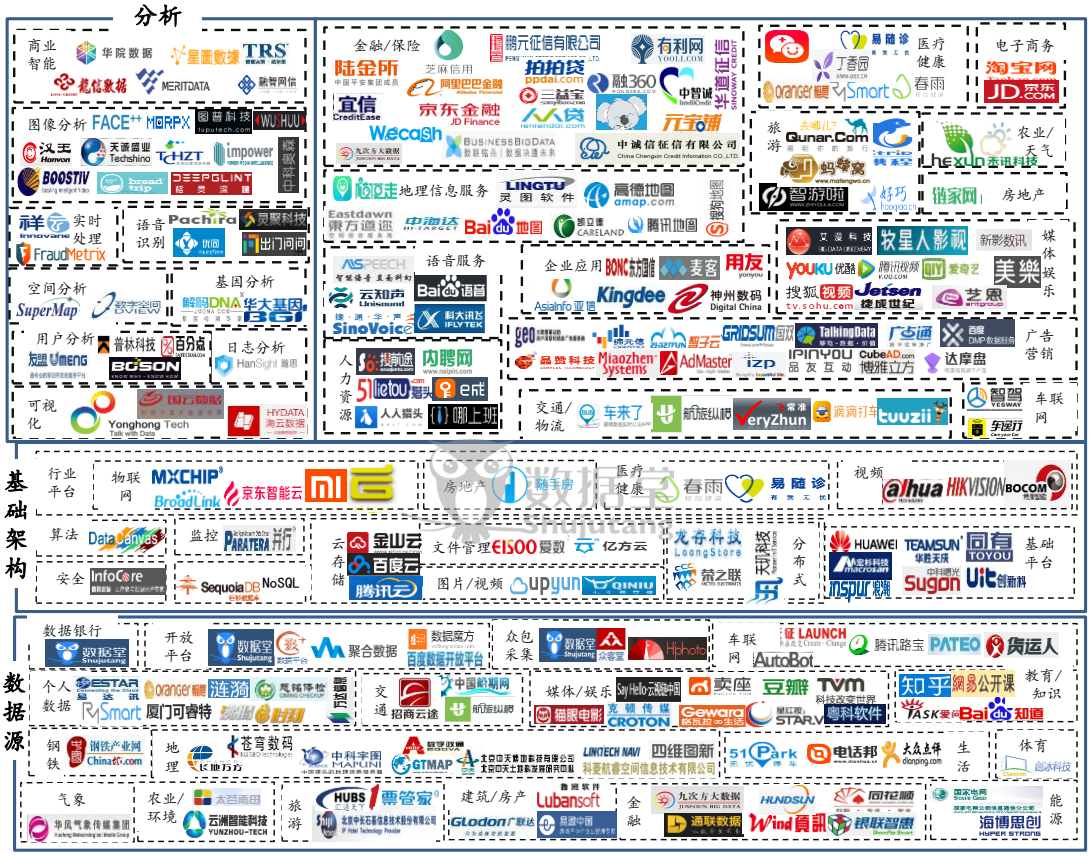
目前，各国在政府决策和公共服务领域的大数据应用典型案例有：

- 英国国家统计局通过大数据技术进行人口普查，在节省约 5 亿英镑经费的同时还提升了人口数据的实时性。
- 纽约州能源研究和发展管理局运用大数据技术来评估气候变化的影响，并为农业、公共卫生、能源和交通运输等领域提供应对气候变化的策略。
- 美国拉斯维加斯市通过传感器技术和数据可视化建模，构建市政基础设施实景图，提升事故处置的效率。
- 德国某些州政府建立了覆盖人口分布、地理数据、矿藏信息等领域的数据库，通过数据分析让决策更加科学化。
- 富士通公司基于用户手机访问社交媒体的数据，帮助东京市政府建立预测犯罪或暴力事件的热点图，探索治安事件的高效应对机制。
- PrePol 公司与美国洛杉矶警方合作，将地震预测算法运用于精确到 500 平方英尺范围内的治安案件预测，使盗窃罪和暴力犯罪率分别下降了 33%和 21%。
- 联合国启动的“全球脉动计划（UN Global Pulse)”，采集 2009 年到 2011 年之间来自于美国和爱尔兰的社交媒体数据并按住房、交通、酒精、教育、医疗等主题进行分类，通过文本标注和情感计算发现了交通通勤类话题高峰与失业率之间的强相关性。
- Shotspotter 公司建立了全美的枪声监测平台，为警方提供实时的枪击事件信息。

- WideNoise 公司通过手机应用感知并采集周围噪音，分析出整个城市的噪音分布，从而对城市规划起到参考作用。

（三）我国大数据产业分析

在我国的大数据产业链条上，由于尚未出现微软、IBM、惠普一类称得上跨平台基础架构的厂商，同时也没有出现具有较大影响力的大数据开源项目。因此，国内的产业区块暂不计入跨平台基础架构和开源两个部分，整体大数据产业链布局图如下所示：



来源：数据堂整理

图. 我国大数据产业链布局

关于企业在产业链上的定位，由于我国的大数据产业发展还未完全成型，各企业的业务模式都还处于探索期间，本报告站在大数据产业的角度，对相关企业、产品和服务进行梳理，确定其在大数据产业链中应处的环节，比如：

- 大型电子商务平台尽管拥有海量数据，但不太可能将其高价值数据大范围开放，因而不计入数据源区块；
- 某些云服务提供商没有直接通过所拥有的数据获得收入，但由于其数据可能对其他行业具有较高利用价值，因而也归入数据源区块；
- 某些企业拥有较强的数据分析技术，在大数据产业中最大的价值体现在于为其他企业提供分析技术，尽管其当前没有直接向市场交付分析服务或方案，仍然归入分析区块。

3.1 数据源

与国外的数据源区块相比，我国大数据产业对于数据源有着自身的特点。由于数据流通在全社会还未形成规模，国内数据源区块中的综合性平台比例较为明显，尤其是培育出了集采集、加工和流通功能于一体的数据银行典型案例。同时，在国内互联网向其他行业迅速渗透的趋势下，各类线下数据的采集和整合现象也较为突出（尽管未必直接通过数据租赁或销售获取利益）。

根据数据所涉及的行业和企业服务形态，国内的数据源类企业可分为以下几类：

- 数据银行。最为典型的是数据堂（北京）科技股份有限公司。具有鲜明的数据市场中介功能，成为数据资源供需对接的桥梁：主动采集数据资源满足各类需求；履行数据资产管理职能，通过数据加工和转化服务，实现数据资源的变现和增值。

- 综合性数据流通平台。中关村数海大数据交易平台是国内首个面向数据交易的产业组织，通过开放 API 进行数据录入、检索、调用，为政府机构、科研单位、企业乃至个人提供数据交易和使用的场所。
- 开放平台。数据堂基于自身的各类海量数据，对外开放通用数据 API；淘宝和百度基于自身的业务数据，分别推出了数据魔方和数据开放平台；聚合数据也通过 API 接口为应用开发者提供各类数据。
- 众包采集平台。数据堂基于广大的用户群体，众包采集各类语音和图片数据；成都夏陌科技推出的嗨图应用，通过众包方式对图片数据进行标注。
- 建筑类数。广联达通过销售建筑材料价格信息赚取利润，为客户提供反映经济形势和供求关系变化的价格数据；鲁班软件构建了建筑业的数据汇集平台，积累了大量建材相关数据。
- 金融类。万得信息专注于金融行业，致力于为国内投资者提供金融信息和软件服务；九次方金融和通联数据掌握海量的上市企业各类数据；同花顺为金融机构提供财经资讯及全球金融市场行情等信息；恒生电子（阿里系以 33 亿元入股）主营金融 IT 产品与服务，掌握各大金融机构重要数据，全面覆盖客户的各类交易记录；银联智惠，掌握所有银联卡用户的刷卡数据。在保险行业，2014 年中国保险保障基金有限责任公司出资 20 亿元人民币成立中国保险信息技术管理有限责任公司，旨在打造中国保险行业的

数据共享平台。

- 气象类。华风气象直属于国家气象局，是最具权威性的气象数据源。
- 房地产类。易遨中国通过为房产中介行业开发 ERP 系统，积累了大量房源、经纪人和买家数据。
- 影视/娱乐类。克顿公司（已被华策公司以 16 亿元的价格收购）建立了影视剧行业数据库，收录自 97 年以来近万部电视剧的收视评估情况、国内主要制作班底以及近万主创人员的信息；猫眼电影基于在线售票和在线选座业务，积累了观众的性别、年龄、地域分布和观影时间等各类数据，以及影片的真实上座数据；星红桉公司通过机顶盒和有线电视网，积累了大量关于电视节目收看信息和电视用户数据；粤科软件（2015 年 4 月被阿里影业以 8.3 亿元收购）作为我国影院市场的主要系统供应商，掌握最为底层的票房数据，并为各类在线选座服务提供支持；无锡天脉聚源建有全国最大的视频节目加工中心和数据库。
- 旅游类。汇通天下公司为酒店提供在线中央预订、分销、管理和交易系统，北京中长石基公司（阿里巴巴集团 28 亿入股）在国内五星级酒店信息管理系统市场占 90%份额，都掌握有大量的酒店业务和客户信息；票管家为景区提供电子票务解决方案，掌握大量景区人流数据。
- 地理/环境类。长地万方、凯立德、北京城际高科等企业通过测绘收集各类地理数据；中科宇图主要采集环境监测数据；科菱航睿

（腾讯入股）为奇瑞、汇众、华泰等汽车厂商提供地图数据。

- 生活类。电话邦为客户提供电话号码数据，获小米参与的千万美元 A 轮融资；无忧停车采集停车场和停车位信息，为百度地图提供数据支持。
- 能源类。海博思创开发智能电网系统，掌握大量用电数据；国家电网下属的国网信通则由于特殊的行业背景，掌握海量的电力行业数据。
- 交通类。中航信是我国最主要的航空数据源，拥有最准确的航班实时信息；中国船期网采集全球班轮数据。
- 车联网类。腾讯路宝、元征科技、Autobot 和博泰电子都基于自身的车载设备，掌握大量的里程、耗油、急刹车等行车数据；北京汇通天下采集大量的汽车发动机数据，为物流行业提供支持。
- 个人类。涟漪采集医师的职称、论文、同行口碑等大量文本信息；微糖和橙意家人分别采集糖尿病和鼾症患者身体数据；易达讯负责建设全国人口库和法人库，拥有全面的法人数据；上海万达信息建设了全国性的医疗健康服务平台和社会保障系统，覆盖上亿人口；厦门可睿特通过专业仪器采集人体足型、体形等数据，为电商和鞋类品牌提供服务；爱康国宾和慈铭体检拥有海量且较为全面的个人健康数据。
- 教育/知识类。同方知网拥有海量的科技文献资源；知乎、百度知道、新浪爱问等通过众包式的问题解答模式，采集了海量的知识信息，百度基于百度知道推出了教育类应用作业帮。

- 农业/环境类。北京太谷雨田，因承建农业部和各省"金农"工程、"三电合一"信息服务工程、12316 综合信息服务工程，掌握大量农业生产经营相关的信息；珠海云洲智能科技有限公司，无人船制造商，采集水质、水文和辐射数据。
- 体育类。上海创冰通过图像可视化加人工辅助统计的方式，在一场比赛中收集超过 6000 项数据，并实时进行统计和可视化处理。

3.2 基础架构

在基础架构区块，国内目前最主要的形态还局限于数据的存储和简单管理上，缺少面向大数据的计算和网络系统，参与者多为正在试图转型的传统 IT 厂商。

- 云存储。金山、百度和腾讯等互联网公司都推出了面向全社会的通用云存储服务；爱数和亿方云推出了侧重于大数据管理的云存储服务；七牛和 upyun 的云存储服务主要针对于图片和视频数据。
- 基础平台。华为、华胜天成、浪潮、曙光等公司推出大数据分析平台级的方案；同有科技主要致力于大数据数据的存储、保护和容灾系统的研发；用友基于大数据技术，提供营销管理、供应链、项目管理等企业云服务。
- 算法。九章云极科技为企业提供大数据分析所需的基础环境和常见算法库。
- 监控。并行科技开发了针对大数据基础设施的监控和性能分析工具。

- 非关系型数据库。巨彬公司推出了分布式 NoSQL 数据库 sequoiadb，已获千万美元级的投资。
- 分布式存储。龙存科技研发了具有自主知识产权的分布式存储方案，并以在石油、广电和互联网等行业得到了普遍应用。
- 行业数据平台。京东、小米、Broadlink 和庆科致力于做物联网行业的基础设施，成为各类智能设备的数据收集和管理平台；随手房供中介记录诸如房屋类型、地址面积、客户信息等数据；春雨医生和易随诊为医生提供病患资料的统一存储和管理平台；博康智能与海康威视等专注于各类视频监控数据的采集、存储和管理。

3.3 分析

在大数据分析区块，从分析技术的角度进行区分，国内的企业大致可分为以下几类：

- 商业智能类。直接为企业提供决策支持，侧重的行业各有不同，但基本都是针对业务、运维和客户进行分析，典型的厂商有华院数据、美林数据、龙信数据、星图数据等。专注于金融行业大数据分析的有融智网信，拓尔思专注于非结构化数据处理的软件研发。
- 图像分析。Face++主要提供人脸识别的技术方案；汉王科技侧重于文字和人脸的识别；格林深瞳专注于计算机视觉方面的研究，获红杉 3000 万美元投资；图谱科技基于用户提供的标签进行建模，

实现图片的识别和分类；面包旅行针对海量的风景区图片进行结构化处理和识别；天创征腾针对金融行业提供票据的识别技术；南京智搜智能专注于流媒体的自动化识别和搜索；杭州摩图科技致力于图像识别引擎的开发；中科奥森基于图像识别技术探索人、车、物、事件的自动识别和检索；重庆中科雲從专注于动态人脸识别、大规模人群监测、车辆多属性深度分析、警用图侦等领域；上海银晨智能识别科技有限公司的人脸识别技术广泛用于公安、金融、司法、民航等领域，支持了上海世博会安保工作。

- 语音识别。北京羽扇智公司开发的出门问问应用，专注于中文语音的识别技术研发；广州灵聚信息致力于以语音领域的中文人工智能交互引擎开发；普强信息（北京）专注于中文的智能语音识别和自然语言处理技术。其他诸如科大讯飞等语音识别的领先企业由于大多直接进入了面向消费者的服务领域，因此归类到应用区块中。
- 实时处理。深圳祥云信息科技专注于复杂事务处理、CUDA 和神经网络等技术的融合，面向股票交易进行实时分析；杭州同盾科技针对网络交易进行实时分析，识别欺诈现象。
- 空间分析。基于地理信息系统(GIS)基础软件对外提供地理空间信息技术服务，包括超图软件和数字空间等公司；中科九度（北京）空间信息技术有限责任公司专注于遥感图像处理和空间信息分析。
- 基因分析。华大基因和解码 DNA 公司致力于基因的检测和分析。
- 日志分析。翰思（Hansight）公司基于日志分析，提供企业安全

解决方案。

- 个体分析。百分点通过网络采集大量的消费者偏好信息，为企业提供业务优化方案；友盟专注于移动互联网用户的分析，为应用开发者提供决策支持；北京至信普林科技有限公司基于自然语言处理和深度学习技术，为企业提供全面的客户画像服务。
- 可视化。永洪科技、海云数据和苏州国云数据专注于大数据的可视化分析。

3.4 应用

3.4.1 医疗/健康

正如全球大数据大数据产业链部分的分析，医疗健康行业的大数据应用主要在于医疗档案整合和分析、病患实时监控和新药研制等环节。而由于非技术因素的影响，我国医疗数据的鸿沟仍然客观存在，导致当前的医疗健康大数据应用不得不从数据源切入，透过基础架构和分析区块，最终才抵达应用环节。医疗相关仪器设备以及信息系统的各自为政，也导致了数据标准化方面的巨大需求，比如同一指标的化验结果，可能在不同医院以不同的格式存在，给数据价值的充分实现造成了极大的障碍。以春雨医生、易随诊、华大基因、丁香园、微糖等为例，都是从健康数据的采集环节入手，在数据统一存储、清理和标准化的基础上，直接面对消费市场提供服务，形成一个完整的商业闭环。

国内逐渐兴起的体检行业同样是医疗大数据的典型应用场景。根

据 Frost & Sullivan 公司的调查，2014 年-2019 年中国体检市场份额的复合增长率将达到 22.5%,中国健康体检的市场规模将会在 2020 年达到 3000 亿元。庞大的市场规模和发展空间，将使体检机构成为我国个人健康数据的主要源头之一,诸如爱康国宾和慈铭体检等大型体检连锁都开始从自身掌握的海量体检数据为入口,对客户健康状况加以解读和判断,将业务扩展到后续的医疗健康服务环节。但是，由于目前大数据分析的技术色彩仍显单薄,这类体检机构被归入到数据源区块中。

同时,行业外的大型企业也开始以类似的模式涉足医疗健康行业,比如百度、平安、阿里和腾讯等。其中，以百度的举措最具大数据特质：接入北京市的卫生信息系统,通过移动医疗健康平台和智能穿戴设备记录人们的健康数据，依托百度知道专家资源和病例问答内容,以及好大夫在线、39 健康网、寻医问药网、有问必答网、育儿网、中国育婴网、宝宝树等医疗健康类网站的数据,上线了百度医前智能问诊平台。

在医院环节，以东软、金蝶等大型厂商为代表的企业也开始在我国医院信息化建设中涉足医疗档案整合和分析。

3.4.2 电子商务

我国幅员辽阔、市场庞大和线下成本过高等因素，导致以淘宝、京东等为代表的网络零售交易平台和电子商务网站得以蓬勃发展。基于所掌握的海量消费者和商家的数据,电商可以将大数据应用在下列

三个方面：

- 精准营销。对用户消费全过程数据（包括浏览、交易、客服、配送和物流等）进行分析，掌握用户基本属性、购买能力、行为特征、社交特征、心理特征和兴趣偏好等多方面信息，为其提供具有高度针对性的服务；
- 商家和供应商决策支持。提供具有高度时效性的行业平均数据、市场需求变化、产业上下游动态等市场信息，帮助商家和供应商分析运营状态，预测销售和用户趋势，并提供针对性的运营优化策略；
- 自身平台运营优化。通过大数据分析为管理层以及各级运营管理人员提供数据分析和决策支持服务。

需要指出的是，Google、亚马逊和Facebook等互联网巨头对大数据概念的推广起到了重要的作用，对于社交媒体和网购记录等数据的分析一直是大数据研究的热点，其中基本的理念、算法和模型都具有较高的普及度。因此，无论规模大小以及是否局限于某垂直行业，大多数电子商务网站都具备了一定的大数据特征。但是，从数据规模、覆盖面、视角丰富度和实时性等角度考虑，只有足够大的电子商务平台才能为大数据提供真正的应用场景，比如亚马逊由于几乎涵盖了全美所有生活必需品并掌握了海量消费者的原始数据，分析和预测的准确性才有足够的保证。有鉴于此，本文只将淘宝和京东等大型平台作为大数据在我国电子商务领域的典型应用案例。

3.4.3 语音服务

我国互联网和移动互联网具有庞大的市场规模和潜力，促使业界在网络服务上不断追求创新和用户体验的提升。对于移动设备和诸如车辆驾驶之类的特定情景，操控上的局限性更使得便捷的服务交付和用户互动方式成为各类厂商的研发热点。由于语音交流是人类最基本的沟通方式，基于语音的网络服务成为必然的发展趋势。

鉴于我国语音服务领域的独有特点和蓬勃态势，本文将其单独作为一个子类进行论述。目前，语音服务多运用在客服中心、电子导航、智能家电、可穿戴设备上，今后将进一步扩展到工业、家电、通信、汽车电子、医疗、家庭服务、消费电子产品等更多的领域。语音服务的主要用途有：

- 情绪识别。通过对基音频率、音量、持续时间和语速等指标的分析，对语音数据中所包含的情绪进行识别，可用于客服；
- 语义分析。基于语义理解技术，准确识别语音内容，尤其是常用词、专有名词和术语的识别。
- 声波分析。语音数据中的声音波纹与指纹一样具有个人指向的作用，基于大数据分析能够识别说话人的身份。

基于上述用途，当前我国语音服务的主要场景是客服或呼叫中心和互联网语音服务。对于客户或呼叫中心，通常是对富含客户身份信息、客户偏好信息、服务质量信息、市场动态信息和竞争对手信息的语音数据进行分析，获取客户群体特征、客户流失原因、业务热点趋势，进而提升运营效率，及时响应市场需求。对于互联网语音服务，

主要是用于人机交互方面，比如百度为用户提供基于语音的搜索服务；腾讯通过微信语音开放平台为开发者提供语音识别服务。

我国提供语音识别专业服务的典型企业有科大讯飞、云知声、思必驰、捷通华声等。另外，各行业巨头也开始涉足语音业务，包括：京东与科大讯飞合作研发智能家居领域的语音技术和服务；阿里与云知声共同针对可穿戴设备等智能硬件产品提供语音服务；交通银行和中金数据基于海量客户电话录音提取客户信息和市场动态。值得一提的是，由于我国方言众多、地区间口音差异较大，语音识别技术的研发对各种方言和口音数据有着巨大的需求，进而推动了产业生态链的成型，比如数据堂等数据源企业就向上述应用厂商大量提供各类经过加工处理后的语音数据。

3.4.4 广告营销

我国新媒体迅猛发展，随着互联网、移动互联网、数字电视、智能设备的发展、普及与功能延伸，广告相关的数据采集、投放通道和分析方法也发生了巨大的变化。据统计，2010 年，我国广告市场总体额度为 7000 亿左右，超过日本成为仅次于美国的全球第二大广告市场；2014 年，我国广告市场传统广告市场首次出现下降，微降 1.7%。电视广告费用首次呈现停滞状态，而新媒体类广告多数保持了高增长的态势，在总体市场的占比为 17%，较 2013 年加了 3 个百分点，其中网络广告营收超过 1500 亿元，同比增长 40%；预计到 2018 年，我国网络广告市场规模将达到 3900 亿元。

随着广告投放重心的逐渐迁移，业界更多地依赖于网络数据和相应的分析技术，达成更加智能的广告匹配以及更加高效的广告资源配置，导致本子类的企业大多集中于互联网广告领域。同时，自 2012 年开始，以品友互动为代表的我国网络广告实时竞价企业也开始迅速发展，而腾讯、阿里和百度也分布推出了自己的 DMP 系统：腾讯广点通、阿里妈妈（达摩盘）和百度 DMP 数据服务。

目前，我国广告营销领域的大数据应用典型案例有：

- 亿赞普。通过与运营商的数据合作，对用户进行行为偏好分析，为企业营销提供支持。
- 智子云、秒针系统、品友互动和精硕科技。专注于互联网广告的实时监测和分析，为广告投放提供咨询服务。
- 集奥聚合和缔元信。采集网络用户信息，提供用户洞察，为广告主提供决策支持。

3.4.5 金融

在宏观层面，我国经过十几年改革，金融业以空前未有的速度和规模在成长。但是，由于潜在的系统金融风险不断积累，直接融资比例过低，银行、证券和保险业的发展不足难以适应多元化经济主体的投资需求。因此，产品和服务的创新成为金融市场完善和长远发展的必然需求。优秀的数据分析能力是当今金融市场创新的关键，资本管理、交易执行、安全和反欺诈等相关的数据洞察力，成为金融企业运营和发展的核心竞争力。

在具体业务层面，由于市场本身以及互联网的迅猛发展，我国金融行业对大数据应用的需求也日益凸显：

- 海量历史数据的挖掘。横向来看，相对于我国其他传统行业，金融业属于信息密集型，累积了海量数据，其巨大价值的挖掘需要引入大数据理念和技术；
- 数据规模的迅猛增长。比如，随着证券市场的发展，上市公司标的正在向海量发展，导致相关机构必须依赖大数据分析指导业务；
- 异构数据的整合需求。随着互联网和移动互联网的发展，多种新型服务渠道的涌现，必须对各类多源、异构的数据进行处理和整合；
- 分析预测准确性的需求。由于市场竞争加剧、风险升高，诸如保险或征信业需要进一步丰富数据维度，并基于大数据分析的新兴技术建立分析和预测模型。

在上述客观需求的推动下，中国金融行业正在步入大数据时代的初级阶段，诸多证券、保险、银行等金融机构都开始运用大数据。目前，我国金融市场中最具创新意义的大数据应用来自于互联网金融所涉及的各类 P2P 和征信企业，自 2012 年以来，中国 P2P 网贷规模增长了近 13 倍，在 2014 年达 410 亿美元的规模。随着市场的迅速增长，我国在金融领域的大数据创新全面覆盖了数据采集、处理和分析的各个环节：

- 闪银（Wecash）。利用大数据技术分析用户的社交数据，完成个人

授信，获得 IDG4000 万元投资。

- 宜信。获取征信对象在各电商平台及社交媒体上的信息，纳入到信用模型评估体系中。
- 拍拍贷、陆金所、人人贷等都利用企业或个人相关的各类数据来完成贷款风险的评估。
- 芝麻征信（阿里）、腾讯征信和京小贷（京东）都凭借自身所积累的海量用户数据来进行风险评估。
- 元宝铺。采集各个电商平台上的卖家数据，为贷款决策提供支持。
- 融 360（百度背景）。提供贷款、理财和信用卡产品的一站式搜索服务，融资总额达 1 亿美元。
- 数联铭品，注重引入各类非受控的外部数据源，对外提供大数据征信服务。
- 九次方，基于所掌握的海量金融数据，构建全国性的企业征信大数据平台。

除上述新兴公司之外，建行、广发（与百度合作）、中信和光大等传统大中型银行也基于大数据技术在客户洞察等方面进行了探索，但是从总体上来说仍偏重于传统的商业智能领域，而且在主动引入外部数据源进行创新等方面力度仍有待提升。

3.4.6 影视/娱乐

大数据在我国影视娱乐业的应用，首先体现在各类视频和音乐网

站对于自身的音视频数据资源的加工上。通过分段和标签等手段，形成更为丰富的分类维度和检索体系，提升了用户体验。这类典型企业包括腾讯视频、爱奇艺、优酷&土豆、搜狐视频、虾米音乐、网易云音乐、豆瓣 FM、QQ 音乐等。同时，通过对访问数据的分析，比如点击率和访问评论，评估音乐和影视作品的市场反应，并为用户提供个性化推荐服务。

除了上述场景，大数据在我国影视娱乐业最具特色的应用在于将新数据源引入到业务分析中。由于互联网/移动互联网与影视娱乐业的充分融合，在线售票和选座业务迅猛发展，前所未有地积累了海量的线下观影用户数据，为影视业的发展提供了坚实基础，这在数据源区块中已经提及。此外，类同于 Rentrak 公司，各有线电视服务商和盒子厂商采集了大量的电视节目观看数据，在结合了其他渠道获取的用户信息之后，能够形成对电视用户的更为深入的洞察，比如特定节目观众群的年龄分布、地区分布、消费能力、文化水平和观看习惯等。新数据源的引入是对传统票房数据和收视率的校正，创新式的数据采集手段将使行业陋规和陈旧的调查问卷方式成为历史。

在夯实数据基础的前提下，针对影视行业的第三方数据分析服务随之出现。除了在数据源区块提及的猫眼电影和星红桉之外，比较典型的企业还有：

- 艺恩世纪国际信息咨询（北京）有限公司。整合多屏终端消费行为数据，为影视业提供决策支持。
- 新影数讯。基于社交数据，为影视业提供商业智能方案。

- 艾漫科技。抓取全网的娱乐相关信息，为文娱产业提供决策支持。
- 牧星人影视。采集演员档期、性别、外形、社交关系、口碑以及剧组预算等数据，通过分析为剧组推荐演员。

3.4.7 在线教育

传统教育模式受限于地点和时间的限制，再加上课程单一等缺陷，使得用户对于方式灵活的个性化教育需求日益高涨。而随着互联网和移动互联网技术的发展，在线教育行业发展迅速，成为我国教育信息化发展最快的领域。2014 年，我国在线教育市场规模达到 1334 亿元，预计到 2017 年时市场规模将达到 2864 亿元。

由于在线教育的迅速发展，海量的课程、教材和用户信息得以采集和汇聚，对于大数据技术和理念的需求成为客观现实。大数据应用于在线教育领域的切入点包括模式创新、用户体验和个性化教育等：通过多点触摸、增强现实、眼动和体感等方式，构建新型的教育模式和渠道，提升学习效率；通过图像识别技术和游戏化手段，改善用户的学习体验；通过人工智能、数据挖掘、推荐引擎等技术，为用户提供量身定制的学习环境和课程。

当前，我国大数据在教育领域的典型应用有：

- 作业通和学大教育。通过教材和试题的收集、组织和搜索以及对用户的特性分析，实现个性化的教育服务。
- 作业帮（百度）。基于百度知道平台所积累的海量知识，百度推出了作业帮应用，除海量数据的管理和检索之外，还涉及了图像识

别等大数据技术。

- 学霸君。通过图像识别技术提供试题答案的查询检索服务。

3.4.8 人力资源

社会经济的发展，推动了我国对于各类劳动力和专业人才的需求，从而带动了人力资源产业的发展，政府、企业和组织都已认识到人力资源的重要作用。据统计，我国的人力资源从业者已达 300 万人以上。但是，从整体上来看，我国人力资源行业仍处于传统的以行政性和事务性劳动为主的阶段，管理和控制仍偏静态，缺乏适应时代发展特征的理念和手段，最基本的表现是缺乏对人才与职位的科学分析，没有严谨的数据体系和分析方法，同时无法对相关法律法规和市场环境的变化做出及时的调整。大数据时代的一大特点就在于对组织机构和人的画像，而互联网则逐渐在消解招聘和求职者之间存在的信息不对称现象，集二者于一身的在线招聘行业为我国人力招聘市场带来巨大的想象空间。

从之前的论述可见，大数据分析在人力资源行业的应用已经相对成熟，随着国内市场的发展，我国人力资源管理与招聘行业也会沿着相同的路径，告别过分依靠市场投入驱动增长的模式，进入以数据为驱动的时代。

当前，我国人力资源行业比较典型的大数据应用案例如下所述：

- 搜前途。通过大数据分析实现简历和职位的精准匹配。
- 哪上班。基于大数据算法进行人才分析和职位匹配。

- e 成招聘。通过机器学习算法帮助企业进行人才筛选。
- 望才招聘。基于社交媒体内容对候选人进行分析画像。
- 内聘网。基于文本分析，实现简历和职位描述的格式化和自动匹配。
- 人人猎头。基于熟人推荐的模式，用众包方式进行人才的搜索。

3.4.9 旅游

多年以来，我国的旅游业保持了 7%年均增长率，带动了相关产业的全面发展，已经成为国民经济新的经济增长点和支柱性产业之一。相应地，旅游相关信息呈爆炸性增长态势，不止是景区和景点所掌握的大量人流数据，各类在线旅游网站也聚集了大量的有关游客、景区、酒店等的数据。

对于旅游服务提供商，充分挖掘这些海量的旅游原始数据，可以掌握游客感兴趣的旅游目标和信息，分析客户的行为、兴趣、爱好等，针对不同类型客户，在线路、交通、住宿、用餐、娱乐、观光项目等方面提供个性化的旅游方案。对于管理部门和景区，针对我国旅游高峰期集中的特点，可以根据历史和实时的游客人流数据，及时在交通、食宿和流量管控等环节制定应对方案，提升用户体验并确保旅游安全。

当前我国大数据在旅游业的应用主要集中在各类在线旅游网站，典型案例有：

- 去哪儿。积累了大量的游客、航班、酒店信息，为用户提供出行方案的检索比对，同时也为相关行业提供决策支持。

- 好巧网与蚂蜂窝。抓取国外酒店各种用户评论和标签数据，提供检索和个性化推荐服务。
- 智游啦。抓取酒店和景点的评论和标签数据，提供个性化的旅游产品推荐服务。
- 携程。基于客户及订单信息，实现个性化推荐服务，并为景区管理提供决策辅助。旗下的慧评网则主要针对酒店业务提供数据挖掘方案。

3.4.10 地理信息服务

地理信息是人类社会的基础性信息，据统计，在与人类生活息息相关的所有信息中,与位置有关的信息大约占 80%。经过多年发展，地理信息服务已成为信息服务业的重要组成部分,其应用覆盖面极广，其中具有大数据特点的主要应用场景有：

（1)分析用户偏好,将基础地理信息与用户当前位置信息相结合，主动推送适当的信息、产品和服务，或在适当的位置投放广告、建设门店。

（2）响应用户需求，为其提供准确的位置信息。这一类场景目前在我国应用的最为广泛,即向公众提供与之衣食住行密切相关的各类地理信息查询和导航服务,如商场、景点、交通、娱乐、餐饮、医院、学校等。

以诺基亚 Here 地图业务（占全球汽车导航市场 80%以上的份额）为例，其收购历程清晰地反映出了对上述两类业务的覆盖，以及支撑

地理信息服务所需的大数据技术要素：

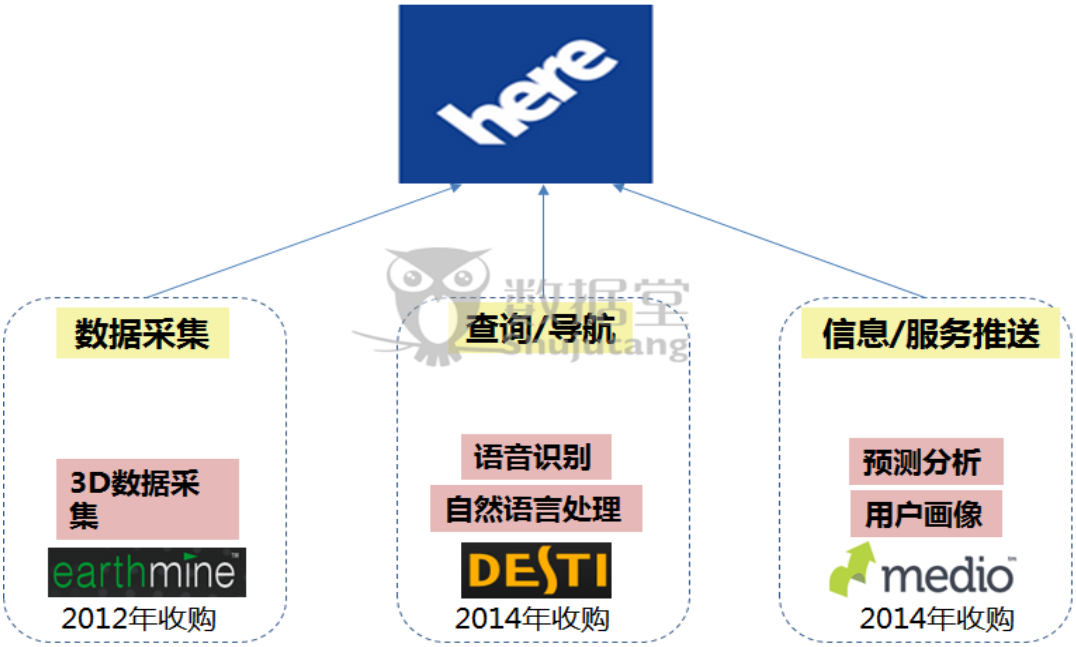


图. 地理信息服务概况

此外，以底层的地理信息数据为基础，政府和企业也可以通过地理信息服务获取基础设施、交通信息、投资环境、行业分布、房地产和人口分布等方面的宏观统计信息，为相关决策提供支持。

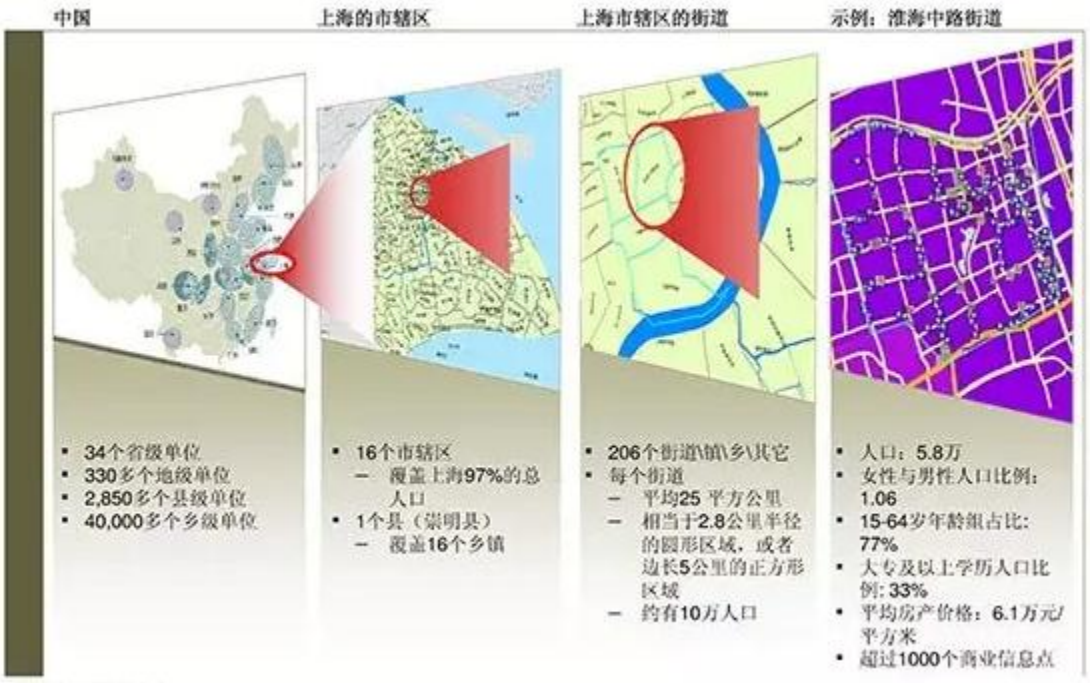


图.地理信息宏观统计（麦肯锡）

当前我国地理信息服务产业发展迅速，主要原因在于互联网企业线上线下融合的大力推动。中国互联网信息中心（CNNIC）数据显示，2012 年中国手机地图用户，使用路线导航和地点查找比例分别为62.7%和45.3%。使用周边生活信息等热点查询比例为29.2%，签到或位置信息分享比例为10.4%。旺盛的需求对整个产业链造成了极大的影响，地理信息数据、系统和服务提供商快速成长，从数据采集、加工和交付的地理信息完整产业链已经形成。

鉴于我国目前地理信息服务的发展态势，本报告单独将其列为大数据应用区块的一个子类。目前，我国比较典型的地理信息服务企业有：

- 百度地图、搜狗地图和腾讯地图等。直接针对用户提供地图服务。
- 高德地图。除为大众提供地图服务外，还为产业链上其他企业提供地理数据服务。
- 中海达。提供测绘、卫星定位和街景等数据服务。
- 东方道迩。提供卫星数据加工和地理信息服务。
- 灵图软件。地理信息服务提供商。

3.4.11 交通/物流

我国在交通网络上投入巨大，积累了大量实时、底层的数据。比如，重庆高速公路视频监控数据每天达到 50TB，我国每天的船舶航速数据达5500万条,广州每日新增城市交通运营数据记录12亿以上、

数据规模在150GB到300GB之间。尽管规模庞大，目前我国各类交通类数据被局限在垂直应用系统中，相互之间缺乏共享和流动。借助大数据理念和技术,有助于整合不同范围、不同区域、不同领域的数据，构建综合性的交通信息体系和一体化的智能交通系统。在管理方面，我国的交通管理过于依靠人工方式。基于大数据技术，能够全面统筹与协调交通资源，准确评估相关建设的科学性，有效提升交通决策水平。

物流产业与交通领域紧密相关。作为现代物流的核心，我国在物流信息化建设还处于相对落后的阶段。尤其是在整个行业的供应链中，上下游之间的信息流没有打通,从而在很大程度上导致了物流成本居高不下的局面。通过大数据技术对物流数据进行全面整合，可以极大提升物流产业的运营效率。同时，基于对物流信息的实时分析，能够挖掘出反映物流规律的信息，进而持续优化物流流程。

汽车保有量的大幅增长（2014年国内汽车保有量将近1.4亿）为车联网的发展提供了巨大的空间。通过车联网，车险业可以提供差异化服务（防盗、事故后的数据分析理赔、车险的差异化定制等等），交通业可以获取实时的路况信息。

当前,大数据在我国交通/物流相关行业的应用场景主要体现在交通工具信息的实时采集和分析，用于信息查询和出行服务等方面。

- 航旅纵横。由中航信基于自身所掌握的数据开发，面向社会提供航班的实时信息。
- 飞常准。基于中航信、空管局、机场和航空公司的数据，提供航

班查询服务。

- 途志。收集国际航班的各种底层数据，为用户提供出行方案优化服务。
- 车来了。通过车辆上的 GPS 获取数据，提供公交实时查询服务。
- 北京汇通天下物联科技。通过车载设备 G7 采集汽车行驶实时数据，为物流公司提供运营决策支持，客户基本涵盖了我国中上规模的物流公司。
- 快的打车。采集用户和出租司机信息，分析其基本信息、信用、行为模式，在用户下单后通过局部地理范围内的人车匹配，提升叫车效率和服务质量。
- 美的空调。查交通违章数据，优化对自身运输车队的管理。
- 快逸行。通过车辆和用户数据的采集，面向车主用户提供车辆健康、安全行驶等服务。
- 九五智驾。基于自身的 OBD 盒子采集行驶相关数据，提供车联网相关的各类服务。

需要说明的是，诸如顺丰等大型物流企业已经开始利用自身系统中累积的海量数据进行业务优化和决策支撑，但是在数据源的丰富性和分析的时效性等方面仍然带有较为浓厚的传统数据分析色彩。

3.4.12 农业

以“金农工程”为代表的一批农村信息化建设工作，已初步构建

起了我国农村信息化基础架构，有效地推动了我国农业产业化和现代化进程。但是，农村信息资源分散在多个规模较小且内容经常重复的系统中，缺乏有效的信息整合且数据的实时性较差。基于大数据技术，整合数据资源、规范数据标准、统一标识和规范协议等，是打通数据流动通道、推动大数据在我国农业领域应用深化的关键所在。这方面的典型案例包括：

- 蒙牛通过奶牛的“智能耳环”、“云端牧场”应用、质检工序监测和社交媒体营销等策略，打造从生产到消费的闭环，构建集质量追溯、生产管理和市场开拓于一体的大数据架构。
- 软通动力在河北廊坊的农田里安装内置摄像头的传感器，采集诸如气温、湿度、雨量等农作物生长环境的数据，并将数据汇聚到云端进行实时监测、分析和管理。

在基于大数据技术的预测分析方面，我国农业领域的应用屈指可数，其中最为典型的是禾讯科技，利用卫星数据，评估农作物长势，建模预测农业产量，与之相较的是 Planet Labs 用卫星数据评估地区发展水平。

3.4.13 房地产

过去的十多年中，中国房地产行业在整体上呈现粗放式的增长模式。但是，近年来在整体调控、供应过剩、融资成本攀高、市场需求趋理性化和个性化的背景下，房地产业开始向精细化运营和深入挖潜的方向转变。房地产行业的决策始终围绕着土地、房屋和消费者三者

展开，而大数据的应用，使房地产商和中介行业更加深刻地洞察土地和房屋的价值以及消费者的真实需求成为可能。

大数据在房地产行业的应用场景大致可分为以下三类：

- 引入大数据理念，通过用电数据和电信运营商数据等新数据源，以创新方法评估城市经济发展状况、消费力、人口组成、区域内购房人群特征、日常通勤人流量等指标，支持投资决策。
- 与互联网企业或电信运营商等掌握海量用户数据的企业合作，结合销售过程中掌握的用户信息，对主要客户群进行群体画像，提升项目建设、产品营销等环节的针对性，同时探索新的盈利模式和空间。
- 在大型综合性购物休闲中心的运营中，引入室内定位技术，分析商场人流轨迹，结合外部数据渠道，对日常客流进行群体分析，进而优化商场布局，帮助商家制定具有针对性的销售策略。

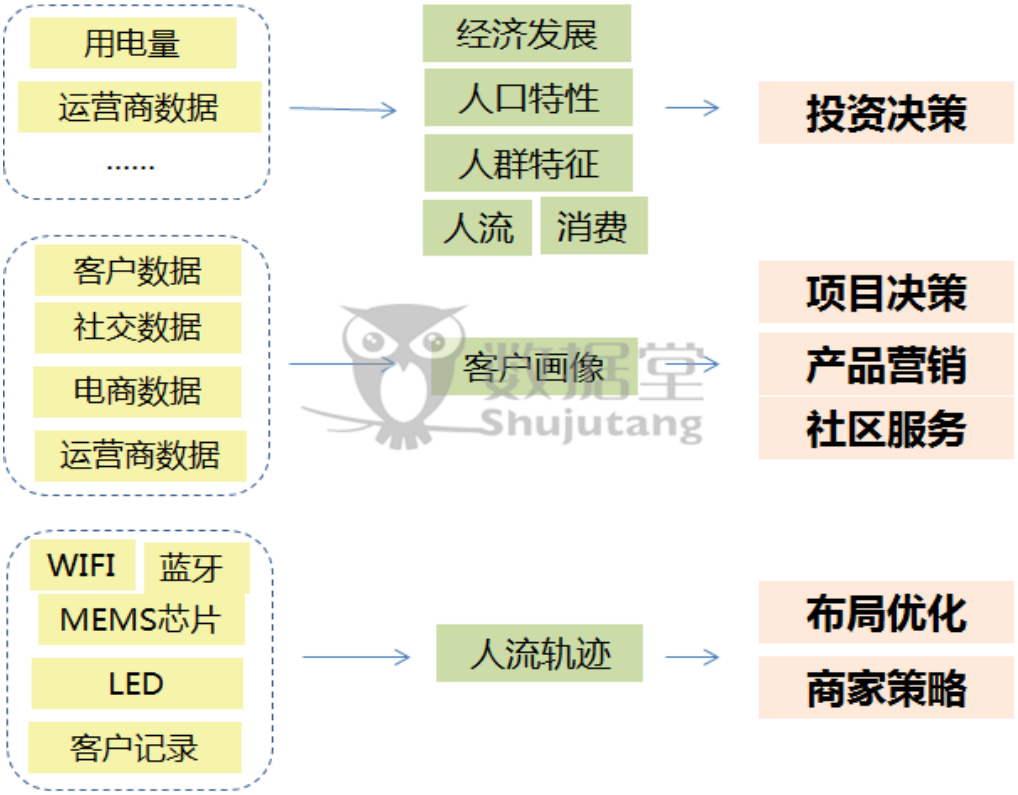


图. 房地产大数据应用场景

需要指出的是，在评估城市和地块的开发前景的过程中，如果仅以传统途径获取宏观经济、市场热点、市场活跃度、供需对比、常住及户籍人口数、房价收入比、地价房价比、人口年龄结构、周边楼盘价格、商业和生活配套等数据，则属于传统的数据分析思路，不应归属为大数据的创新应用。

当前，我国房地产行业的大数据典型应用案例如下：

- 万科地产。与移动、联通、电信三大手机运营商合作，测算北京地区的实际人口规模，并结合当年北京的新房供应量、存量房数量、房屋均价、购房人群年龄结构等数据，支持土地投资决策。
- 链家地产。基于自身的房源和客户数据进行关联分析，打造包括生活支付、社区服务、智能家居等方面的生活服务平台

- 易遨中国。基于房产中介业 ERP 系统的数据积累，开发了美丽屋应用，实现房主和中介的对接。
- 万达地产和万科地产（与百度合作）等。通过室内定位技术分析商场内人流模式，或引入新数据源对用户全面画像，以此优化商场布局，帮助商家提升销售业绩。

3.4.14 企业应用

本类别中的企业很少有初创公司，主要是传统的 IT 厂商，普遍还处于技术跟随的阶段，典型企业有东方国信、亚信、金蝶、用友、神州数码等。

（四）我国大数据产业发展策略

4.1 现状分析

根据对国内大数据产业的调研，可以得出以下主要结论：

（一）互联网企业在引领大数据应用

国内互联网企业由于在拥有了海量的用户数据之后开始着手开展各类分析工作，用以支撑自身的电子商务、定向广告和影音娱乐等业务。同时，在互联网产业020的趋势下，互联网企业逐渐将业务延伸到金融、保险、生活、旅游、健康、教育等多个行业，极大丰富了数据来源，促进了分析技术的发展，拓展了大数据分析在诸多传统行业的应用场景。

这种现象仅从阿里、百度和腾讯三大互联网巨头的情况即可得知：

- 阿里相关的有淘宝、庆科（物联网）、芝麻信用、蚂蚁金服、中长石基、数据魔方、高德地图、虾米音乐、阿里旅游等；
- 腾讯相关的有腾讯广点通、面包旅行、大众点评、腾讯视频、腾讯路宝、四维图新、科菱航赛、腾讯征信、腾讯云、中海达、丁香园、腾讯健康云等；
- 百度相关的有百度云、爱奇艺、百度视频、百度迁移、百度精算、百度语音、百度知道、作业帮、元征科技、无忧停车、百度数据开放平台、百度地图、长地万方、融360、去哪儿等。

（二）产业分布过于偏重应用环节

应用类的企业或产品的占比达到了39%,产业链分工还不够精细。

除了国内企业在基础架构和分析技术上多处于跟随状态，缺乏自主创新之外，一个很重要的原因是绝大部分拥有数据的企业都在分析挖掘的基础上对外提供服务，比如阿里巴巴开始做金融行业的数据分析应用，百度基于自身数据涉足在线教育领域，中航信通过航旅纵横提供航班信息服务，春雨医生和丁香园之类也没有直接通过数据的租售获利。相比之下，国外案例中的 Sermo.com 则在累积了大量医疗数据之后直接销售给医药公司。

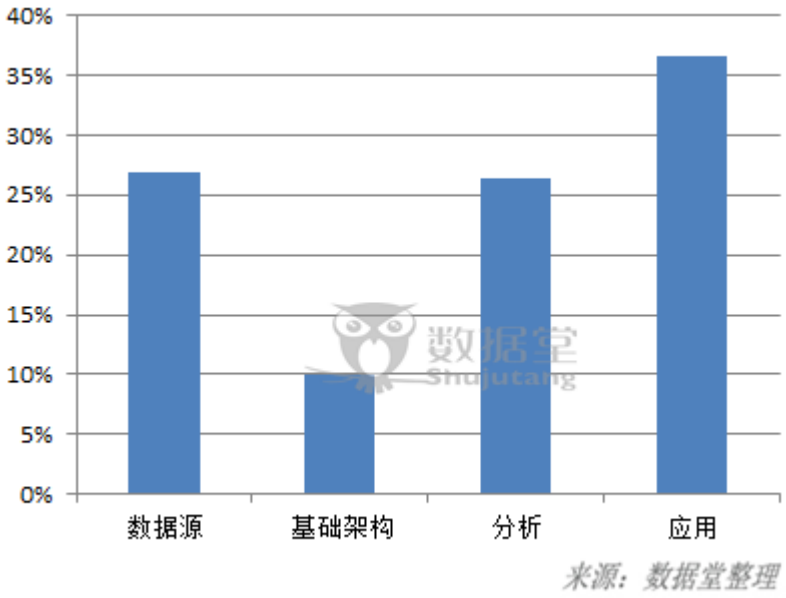


图. 国内大数据产业分布

（三）缺乏综合性的数据聚合流通平台

由于欧美日等发达国家和地区较早跨入了信息化时代，对数据的整合和利用也已成为常态，初步培育形成了各垂直领域内的数据采集和租售的产业生态。而我国由于细分行业的数据链条还处于起步阶段，对于综合性的数据汇集和流通平台有着客观的需求。

国内企业由于主观认知的不足或客观条件的限制，很多数据拥有者仍主要关注于将数据用于自身的业务。这种局面造成的一个后果是，

对数据的采集也只是局限在多个行业细类之内。在所有纯数据源企业或平台，只有不到 8%在开展数据的租售业务，其中面向各类数据的平台只有数据堂和聚合数据等少数几家。

（四）基础架构和分析环节比较薄弱

我国互联网企业快速将国际上先进的开源大数据技术整合到自身系统中，并构建了较大规模的系统，在国内保持领先。但总体上仍缺乏原创的平台和分析技术，对国际主流开源社区的贡献程度也不高，学界和产业界在全球大数据技术发展进程中的话语权不够。

与技术创新不足直接相关的一个问题是，我国科研机构和高校在大数据技术研究方面缺乏建树。国外的很多大数据技术或产品出自于高校，比如起源于加州伯克利大学的实时大数据平台 Spark 和起源于麻省理工学院的人脸识别公司 Affectiva 等。

（五）应用领域的行业分布仍不够全面

全球的大数据应用主要集中在金融、保险、电信、媒体、政府、零售、交通、公共服务、医疗健康等。而由于国情的不同，我国当前的大数据应用热点主要集中在，金融/保险、医疗健康、娱乐、广告、教育等领域，在有所交叉的同时体现出了很大的差异性。

值得一提的是，可能与直观的印象不符，本文并未计入大量的电子商务类企业。这是因为大数据在电子商务领域的应用模式相对单一，基本都是基于社交网络和网购行为数据，对潜在用户的挖掘并进行个性化推荐，对行业和技术的发展都没有显著的意义。因此，只有如淘宝和京东等自身拥有海量数据的厂商才能算作大数据产业链的组成

部分。

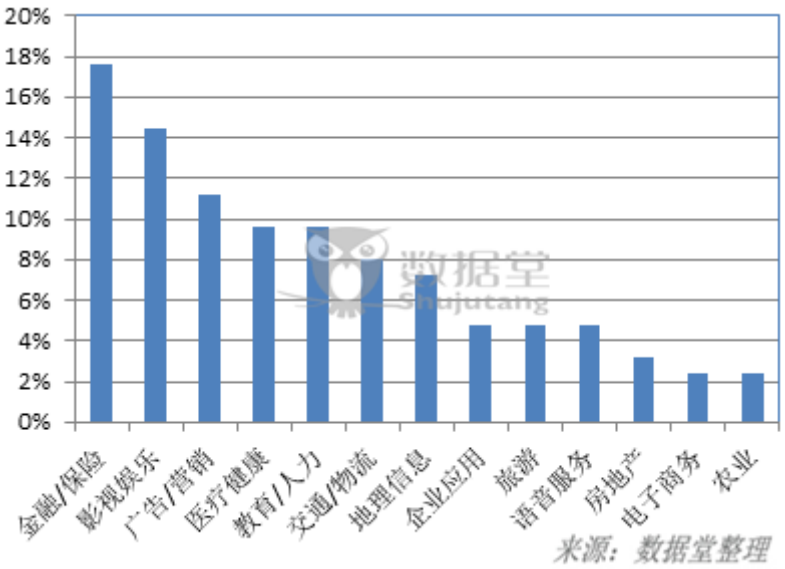


图. 大数据应用热点

我国当前在大数据应用环节的主要问题是：

- 政府/公共服务类应用基本缺位。我国正处于新型工业化、信息化、城镇化、农业现代化同步发展的新时期，大数据技术在经济发展、社会管理和公共服务方面应该大有作为。但是，较为典型的应用案例仅有粮食库存监控和利用卫星数据估算耕地面积等少数几个。
- 农业类应用基本缺位。与北美等农业先进地区相比，我国农业产业链条 IT 化的程度还不够充分，在短时间内确实难有较大的突破。目前，应该着力于数据采集网络的建设，并进一步推动气象、卫星等相关数据的开放，为大数据应用奠定基本的物质基础。
- 电信业和银行业应用相对滞后。首先，我国电信业和银行业在数据分析的对象、方法和目的上已经有了长足的进步，但是思路上创新不够，多为已有案例的移植和模仿；其次，从全球的大数据案例中可以看到，电信业的数据已经开始与外部数据产生碰撞，

或为电信业自身所用，或为其他行业提供决策支持，而我国运营商数据的流通仍存在较大的困难。但是，随着征信行业的开放和互联网金融的兴起，央行征信系统逐渐成为一个真实可用的数据源。

- 医疗健康领域的应用还不够深入。当前，我国医疗健康行业的大数据应用多由各类健康类 APP 推动，主要集中在数据的收集和分享环节，很少有数据转换、整合和分析等挖掘数据潜在价值的企业或产品，也没有将大数据与医院业务系统紧密结合的典型案例。

（六）大数据应用的思路较为单一

当前，我国大数据应用的模式和方向上较为单一，倾向于抄袭或模仿国外现成的案例，其中最为典型的就是金融行业的应用对比。

随着国外如 Zestfinance 等公司超越行业惯用的 FICO 模型，引入征信对象更为全面（并非只是线上）的信息并开发新的模型评估个人信用等级，国内 P2P 网贷企业也以此为参照，引入各类网络信息，试图以自动化方式取代耗时耗力的针对贷款企业的实地调查，在一定程度上忽略了数据覆盖面、技术积累和征信对象类型等种种不同。

更进一步，国外金融行业的案例包括了引入网络数据、清洗现有数据、基于先进架构提升欺诈识别效率、实时采集网点数据、转化客服音频数据等各种不同的切入角度和实现思路，都是基于自身的业务痛点而做出的。反观我国金融行业，几乎都集中在信用评估、业务分析和客户画像上，思路较为单一，缺乏创新。下图以中外银行业为例，说明了在大数据应用上的差异：

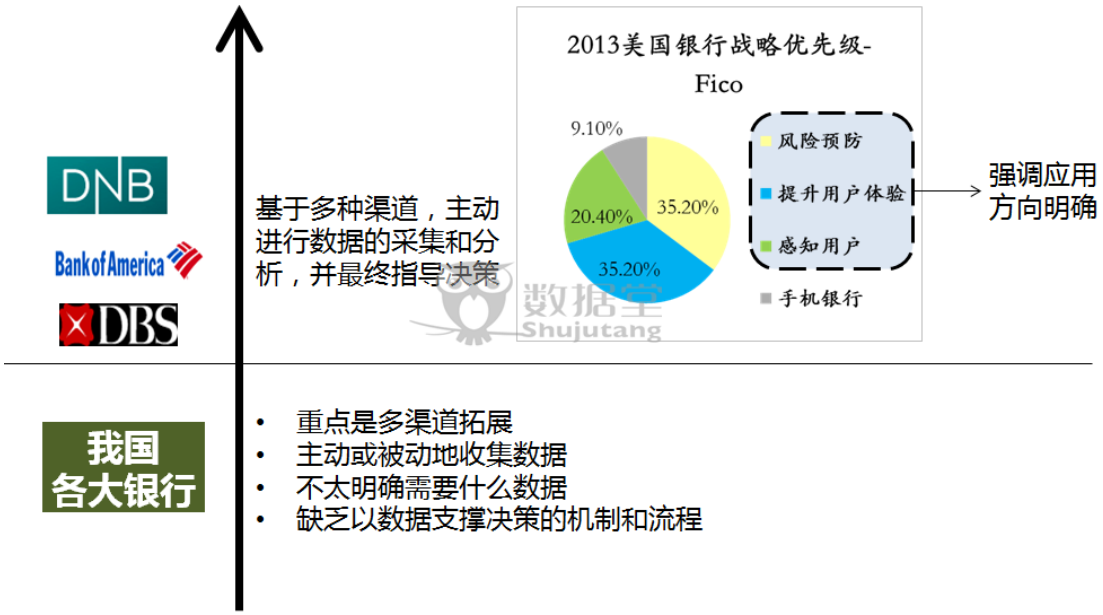


图. 国内外银行业大数据应用对比

4.2 趋势分析

结合大数据近年来的发展态势，以及大数据所涉及的技术本质，我国未来大数据产业的发展将呈现以下趋势。

（一）数据的流通以综合性的流通和交易平台为主

在用户需求的推动下，综合性的数据交易平台将成为行业主流，形成少数几个全国性的数据流通集市，主要是因为：

1. 数据的跨域、关联分析已成为普通的共识，业务决策支持可能会同时接入多方的数据，因此市场需要综合性的流通平台来降低数据获取的成本。

2. 我国目前缺乏对各类数据的统一标准规范以及大规模的自动化处理手段，相关工作的推进需要如果数据流通平台局限在特定行业或领域内

3. 不同用户对于数据价值的认知有较大差异，为了客观、真实的

反映数据的内在价值，必须通过综合性交易平台来撮合供需多方，形成对特定数据价值的普遍共识。

（二）产业链的分工将日益清晰和细化

随着数据价值认知的深化，用户对于数据的质量要求越来越高，由于标签的准确性、无效或错误值的比例、数据检索的效率和数据关联的精准度等指标对产业链上各类产品或服务价格的影响日渐加大。同时，在综合性的大型数据交易平台带动下，围绕数据所产生的各类需求能够得到最大程度的凸显，必然会促进大数据产业链的划分逐渐清晰和细化，诸如数据采集、汇聚、加工、交易和分析等环节在内的细分产业都将得到极大的推动。

此外，企业在大数据产业链中的定位将会更加明确和聚焦。当前很多跨越了多个环节的企业，将会根据自身的优势和特点来重新定位在大数据产业链中的角色，可以预见的是某些数据拥有者将专注于对外提供数据服务，而目前横跨分析和应用环节的企业将会有很大一部分成为专业的分析技术提供商。

（三）形成多个细分的垂直行业生态

随着大数据应用在不同行业的落地和深入，数据分析终将成为企业日常运作的基础性工作。应用环节对于数据、平台和分析的需求会逐渐细化，并连锁式地、逐一反向作用于产业链上游的各个环节，进而形成具有行业特色的大数据垂直生态系统。这一趋势将对基础架构和分析环节产生较大影响。

大数据分析所涉及的理论和技术因数据类型、数据规模和应用场

景而异，最终无法收敛于统一的数学模型，因此必然会在算法层面产生各种分支，衍生出侧重于不同行业的分析技术供应商。

随着分析技术的差异化，基础架构也必将出现各种面向应用的定制和优化,从而削弱当前MapReduce模型和Hadoop平台的强势地位，内存计算模型的提出和Spark平台的出现就是证明。更进一步，在体系架构方面也将会突破由通用硬件搭建分布式系统的模式。

综上所述，围绕大型数据流通平台，产业链在横向和纵向上都将逐渐细化，形成大数据产业的网格状布局。

4.3 各国推动大数据发展的案例

全球主要国家、各类国际组织和国内多地政府，都将大数据的发展提升到了战略层面，并陆续出台了一系列的政策加以推动。

（一）国外现状

美国政府将大数据从商业行为上升到国家战略层面，推动大数据在经济社会各个层面、各个领域的应用深化。2012年3月29日，奥巴马政府宣布“大数据的研究和发展计划”，由美国科学基金会、卫生福利部/国家卫生研究所、能源部、国防部等多个联邦政府部门共同推进。

2011年9月，巴西、印度尼西亚、墨西哥、挪威、菲律宾、南非、英国、美国等八个国家联合签署《开放数据声明》，成立开放政府合作伙伴组织，目前全球已有60多个国家加入。

日本政府于2012年7月宣布《面向2020年的ICT综合战略》，

力图“通过大数据应用促进社会发展和经济增长”，并提出活力数据战略。文部科学省同时指出将大数据收集、存储、分析、可视化、建模、信息整合作为研究的重点。

英国在2013年1月宣布将大数据作为未来八大关键技术领域，计划在两年内投入1.89亿英镑。同时，在2013年11月承诺所有的数据都通过数据门户data.gov.uk向社会公开，并为此专门构建一个国家级的信息基础设施。

2013年6月18日，八国集团首脑在北爱峰会上签署《开放数据宪章》，要求各成员国率先开放公司信息、犯罪与司法、地球观测、教育、能源与环境、医疗健康、科学研究、统计、社会福利、交通运输与基础设施等数据。

联合国发布了《大数据促发展：挑战与机遇》的白皮书，指出大数据时代已经到来，大数据对于联合国和各国政府都是一次历史性的机遇。报告讨论了如何利用大量丰富的数据资源帮助政府更好地响应社会需求，指导经济运行。

（二）国内动态

2013年8月，国务院发布了《关于促进信息消费扩大内需的若干意见》（国发〔2013〕32 号），文件提出的“促进公共信息资源共享和开发利用”和“提升民生领域信息服务水平”等方针都明确指向了数据资源的开放。

在中央层面，中央网络安全和信息化领导小组办公室已经于2014年初开始进行国家公共信息资源开放的规划工作，充分借鉴各国政府

的数据开放工作成果，旨在形成覆盖全国、贯穿各级政府和公共服务单位的数据采集、存储和社会化服务体系，提升全社会对公共信息资源的再利用水平。

在地方层面，各地政府已逐步推动数据开放的工作。北京市政务数据资源网目前已汇集了 35 个政府部门的 269 项、共计约 36 万余条原始数据，涵盖旅游、教育、交通、医疗等多个领域。上海市公共信用信息服务平台于 2014 年 4 月正式开通，已发布交通、公共服务、经济统计、行政管理等 200 多个数据产品。武汉各市直部门已公开数据类别 500 多个，涉及政务、警务、环保、医疗、农业、交通、物流等 30 多个领域。此外，青岛、贵阳、广州、厦门等城市也在积极着手准备相关工作。

4.4 我国大数据产业发展建议

4.4.1 从数据源和应用环节入手

大数据涵盖的范围较广，需要一定程度的聚焦才能让大数据概念和应用尽快落地，形成具有示范意义的典型案例，实现质的突破。对于当前阶段大数据发展的关注重点，需要着眼于产业链全局进行考虑。任何一个产业链的成型和发展的核心本质都在于达成供需两端的平衡，而供需两端的发展和完善也是一个产业链最终走向成熟的必要条件。大数据产业链的供需两端分别是数据源和应用环节，从前述世界各国以及我国各地政府对大数据的推动举措来看，也多从数据和应用领域入手。由于我国大数据产业链尚处于孕育期，数据供给和应用需

求更是当前工作的重中之重。

对于基础架构和分析环节，由于应用场景的不同，必然会呈现多种形态并存的局面。当前，基础架构的发展明显未能满足应用场景优化的需求，较为单一的架构导致了大量存储、能耗、机房空间和管理成本的浪费，而分析技术则必然与应用场景紧密挂钩，不可能出现一种特定算法未经修正即有效运用于各类业务的情况。因此，在大数据发展初期，基础架构和分析算法的发展应主要以科研机构和企业界自发行动为主，无须在战略或政策层面进行规划，否则极有可能导致大量投资和人力的浪费。而只有打破现有的数据藩篱，加大全社会数据资源的供给，促进数据资源共享和流通体系的建设，同时以应用需求为导向，为目前涵盖面过广的大数据领域指出具有真正实用意义的前进路径，才能够给基础架构和分析技术的发展注入长久的驱动力。

4.4.2 积极推动数据开放

数据开放的意义

数据是继土地、劳动力、资金之后的第四种生产资料，是大数据发展的核心所在。数据在人类的生产生活过程中不断产生，为人类的各种决策提供着事实依据，推动社会向前发展。由于云计算、大数据和物联网的发展使得各种信息被更加详细的记录下来。尽管数据规模已足够庞大，要真正实现大数据的价值，首先要面对数据开放和流通的挑战，只有结合多源头数据的跨域分析才能提炼出更完整的知识和更深刻的洞察，才能真正达成社会管理、公共服务、金融保险、科研

教育、医疗卫生、零售消费、文化娱乐及制造业的跨越式发展。

由于数据所蕴含的巨大价值，数据开放和流通的价值已成为一种普遍的认知。根据麦肯锡预测，开放数据在全球的教育、交通运输、消费市场、电力、石油/天然气、医疗健康、消费金融（包括银行、保险和房地产）等 7 个领域可以撬动 3.2 万亿到 5.4 万亿美元的经济价值；根据美国参议院商务、科学与运输委员会发布的报告，全美数据中介市场 2012 年的总规模已达 1500 亿美元，相当于当年美国情报总预算的两倍。可见，数据资源日益成为人类社会的生产要素和战略资产，而数据的开放和流通是其价值体现的前提和基础。

以数据开放为切入点，也是符合大数据发展的客观规律。大数据的应用，主要的瓶颈分别是技术、数据和人才，技术的研发和人才的培养在很大程度上取决于企业界、学术界和教育界的共同努力，需要时间较长。而且，对技术的需求是来自于数据的驱动，如果在数据规模和多样性没有达到一定的程度之前就急于对技术进行大量投入，是一种本末倒置的行为。

相反，数据获取的成本则可以通过政府的推动而得以降低，并在短期内起到立竿见影的社会示范效应。因此，着眼于大数据发展的客观规律，应该首先从数据开放做起。比如，以地理信息为例，目前在中国，地理信息数据的可获取性、准确性和全面性仍然制约着中国地理信息产业的发展。能够提供商业价值较高的街道及以下层次（如街道、邮编区域、居委会乃至小区）边界的地图供应商极为稀少，与之相配套的数据，如人口、收入、消费、住房房价和商业楼盘的租金，

也不易获取。通过全社会数据的开放和流通，以及在采集方面的创新，可以有效改善这种现状。

综上所述，数据的开放共享和流通是大数据产业链后续环节的基础所在，应该着力加以推动。

政府数据开放的意义

在全社会的范围内，由于担负经济发展、社会管理和公共服务职能，政府以及相关机构所拥有最高价值的数据，比如统计、税收、治安、土地、就业、环境、交通等各类数据，无不隐含着人类社会和自然环境的最真实和最及时的信息。可见，从政府数据的开放和共享入手，能够释放出巨大的经济和社会价值。

此外，大数据最终的价值体现在各个行业领域内的应用，单靠企业界和学术界难免会局限于特定行业，无法体现出大数据跨领域的真正特色。因此，由政府牵头建立数据开放共享的典范，将对大数据产业链的发展大有裨益。

目前，由于庞大的国土、人口和经济规模，我国已成为仅次于美国的数据大国。预计到 2020 年，我国的数据量将突破 8.5ZB，占全球数据总量的 21%。其中，据麦肯锡分析，我国三分之一的数据属于政府及提供公共服务的机构与企业。比如，“智慧城市”建设大约一个季度就能产生 200PB 数据，其他农业、气象、环境、工业制造和人口流动等数据也规模庞大。除了规模庞大之外，政府所拥有的数据价值极高，这是由政府的社会管理职能所决定的，其中比如统计、税收、

预算、土地、就业、空气质量、治安、公共设施、交通等类数据都含有极高的应用价值。因此，从宏观层面看，由政府带动的数据开放和流通，对经济发展、产业升级、社会管理和科技创新等方面都具有极其重要的意义。

（一）提升社会管理水平

政府掌握有事关社会、经济、环境和民生等方面的各类宏观数据，有巨大的信息共享与数据分析的需求。通过政府数据的开放，促进政府各部门之间的信息交流，已成为政府决策科学化的基本保障。同时，以适当的方式实现政府数据向全社会的开放，可以充分调动各方力量，探索在卫生医疗、交通、文化教育、环境和资源保护等方面的发展机遇，形成提升政府决策和公共服务水平的巨大动力。

（二）推动产业升级和创新

作为 2015 年中央政府工作报告中拉动经济发展的两架马车，创业创新和公共产品及服务都能在很大程度上受益于全社会的数据开放。

首先，数据已逐渐成为企业的生产要素和战略资产，其价值的挖掘是企业发展的基础所在。通过数据的开放共享，引入新的数据和分析思路，可以为教育、医疗、零售业、物流业、制造业及互联网等行业创造巨大的发展空间，其中典型的代表就是由金融信息、地理信息和气象信息的开放共享所催生的各类新兴服务业态，比如 2006 年中国人民银行上海总部实施信息公开之后，催生了大批金融信息咨询服务公司。在当前我国经济转型升级的大背景下，这一点具有特殊的时代意义。

其次，数据的开放和流通将推动产业的创新。在信息时代，数据正日益成为科研和生产活动中不可或缺的要素，但是创新和创业活动面临着数据获取成本较高的问题，尤其对于中小企业来说更是如此。通过数据开放共享，可以让技术企业和科研单位专注于技术的开发和业务的发展，全力探索新的应用领域和产业机会。

（三）推动科技创新

数据是科研成果最直接的体现。但是长期以来由于条块分割，未能得到充分的挖掘利用，导致我国科研活动与市场实际需求严重脱钩。通过科研数据的开放共享，可以形成科研活动和实际应用之间的良性互动，为科技成果的转化提供新的通路。同时，通过各高校和科研机构之间的数据共享，可以促成最新科技成果的交流，推动相同领域内各科研主体的协同合作，提升交叉学科的研究水平。

（四）促进环境监督和保护

近年来，大气污染，固体垃圾排放和水污染已经对社会可持续发展的一大阻力，成为影响我国居民健康和环境安全的重要因素。基于收集到的大量环境质量相关数据，通过开放共享，打破政府内部的条块分割，引导和鼓励全社会积极参与到对环境的监控和保护中，可以有效提升环境保护工作的广度和深度。

综上所述，数据的开放和流通已成为全球的潮流和趋势，由政府为主导的数据开放是目前世界各国的普遍经验。在以大数据和云计算为标志的新一次 IT 浪潮兴起的同时，大力推进数据开放平台的建设

将在产业升级、经济发展、民生建设和公共服务等方面对我国社会的协调、可持续发展起到巨大的推动作用，创造可观的社会效益和经济效益。

4.4.3 注重应用和模式的创新

在应用环节的推动上，政府的推动不仅能够促进大数据产业链的发展和完善，更能够在事关国计民生的诸多领域，尤其是在基础设施建设、公共服务和新兴行业方面发挥巨大的作用。

在上述领域的应用探索，对于大数据基础架构和分析技术的发展也具有特殊意义。社会管理、农业、交通、能源等领域所包含的数据规模庞大，采集难度高，通常需要进行实时的处理和分析，这些特性对于基础架构和分析环节都是极具现实意义的课题。

此外，对于数据银行和众包等创新模式，也需要政府从全局角度进行规划、设计和推动。

社会治理

在经过了 30 多年的改革开放之后，我国社会治理面临着种种新的问题和挑战，比如人口问题、环境问题、群体冲突、社会治安、公共危机处理等。为此，十八届三中全会将推进国家治理体系和治理能力的现代化纳入到全面深化改革的总目标中。

国家治理体系和治理能力现代化，要求治理要更加科学，因而必须准确掌握治理对象的状况及其外部环境信息。现阶段，我国正处在

社会转型期，需要对包括人、财、物、事等在内的庞大而复杂的信息进行采集、管理和分析，这与大数据的发展不期而遇。基于大数据技术对海量数据进行收集，大数据中呈现的宏观趋势将会越来越准确而清晰，揭示一些潜在的隐含模式，例如经济形势、风险异常区域、整体灾情等。

科学的决策和管理需要以客观事实为基础，即支撑决策的数据必须足够准确。而且，为了防止相关政策和措施的滞后性，必须尽可能收集最及时的信息。但是，当前我国在诸如经济总量、GDP 和 CPI 一类宏观指标的统计上，基层的数据采集方式和方法对专职的人力有较重的依赖，而且在层层上报的过程中不可避免地会造成数据的偏差，已经很难适应结构复杂、高速变化和高流动性的经济和社会体系。此外，我国幅员辽阔，各地区发展程度不一，信息化建设或交通落后地区难免在数据收集和上报过程中落后于先进地区，在一定程度上也延缓了整体性决策的形成。

可见，关于经济运行和社会治安等类信息的及时和准确收集在我国尤其具有急迫性，这也正是大数据思维和方式在社会管理方面的典型应用场景。以视频监控等方案和云计算实现数据采集和存储，结合众包等创新模式，可以在传统的信息收集手段之外实现更具时效性、覆盖面更广、更为底层的数据采样，缩短数据汇报所需经过的路径，从而大幅提升统计的准确性，为政策制定的合理性和科学性奠定基础。在社会运转相关的底层数据采集方面，已有成功案例出现。2012 年，苏州警方通过各类流动警力实时采集治安隐患信息，每天达 700 多万

条，所累积的海量数据为破案提供了极大便利。

高效的社会治理需要能够正确识别出企业、组织和个人等社会治理工作的基本要素，并对这些要素的社会活动和时空环境进行分析。首先，基于现有的国家基础人口库和法人库，结合诸如互联网企业等数据源，实现网络身份与现实世界的映射；通过电信运营商和视频监控等数据，通过大数据技术进行清洗和挖掘，可以更为准确地掌握治理要素的活动轨迹；通过网购数据和银行交易数据，可以更加准确的把握企业、组织和个人的社会和经济活动情况。基于上述手段，能够加强对社会风险的控制，提高政府的预警能力以及对社情民意和紧急事件的响应能力，有助于进一步加强和完善社会公共安全体系和社会应急管理体制。

以社会治理为典型探索大数据在政府领域的应用，更重要的意义在于普及数据治国、科学管理的意识，为政务领域各方面的高效和精细化管理奠定基础，提升政府运作效率和决策的科学性，最终构筑国家整体实力方面的竞争优势。

综上所述，在当前的时代背景下，政府应积极引入大数据理念和技术手段，推进全社会基础信息的采集和管理工作，并鼓励高校、科研机构和相关行业的力量参与到政府和公共服务事业单位的数据价值深度挖掘中，不断提升我国社会治理体系和治理能力的现代化水平，为构建和谐社会、促进社会发展创造既有秩序又有活力的基础运行条件和社会环境。

智能交通/物流

交通和物流行业的存在和发展是人口迁移和商品货物流通的基本条件，据研究表明，交通运输与国家和地区的经济增长的相关系数在0.9 以上，是衡量一个国家现代化程度和综合国力的重要标志之一。我国幅员辽阔，人口众多，随着改革开放以来经济社会的不断发展，交通运输和物流行业成长迅速。同时，交通和物流业的发展已成为经济结构调整的重要力量，全国社会物流总额在2004 年到2012 年之间的年复合增长率达到21.07%,有力推动了我国电子商务市场的发展。

但是，随着我国城镇化和工业化进程的深入、地区间人口流动的日益频繁、机动车数量的激增和交通基础设施建设的快速发展，运输效率和大气污染等问题亟需解决。据统计，因交通堵塞造成的损失占到了 GDP 的 1.5%至 4%，相应的燃料损失及环境污染整治费用也高达千亿级别。鉴于这种情况，我国必须大力推动智能交通和智能物流建设。

智能交通的立足点在于交通运输工具（汽车、船舶和飞机等）与信息化的全面结合，通过对交通信息的实时感知，及时发现拥堵，调控交通流量，预警安全隐患，从而达成对交通系统的全方位、立体式管控和优化。智能物流涉及物联网、网络通信和云计算与物流基础设施的结合，通过对货物实时位置监控和信息分析，形成对物流全过程的感知、反馈和控制，优化成本并提供差异化的物流服务。

当前，我国的智能交通和智能物流建设尽管已有长足进步，但仍存在着种种问题。比如，对交通信息的感知和收集广度和深度不够；

对存在于各个管理系统中的海量的数据无法共享运用；对交通态势缺乏预测能力，未能充分满足公众的交通信息服务需求；各类交通和物流数据的潜在价值未能得到有效分析和挖掘。通过引入大数据理念和技术，有针对性地改善或解决上述问题，是智能交通和智能物流发展的必由之路。

智能电网

随着我国经济进入新常态，国家经济政策调控逐渐偏向于结构优化、增长质量、节能降耗、环境保护和民生改善等，电力需求出现趋势性拐点，进入了低速（相对于 GDP 增速）增长的常态。而且，随着第三产业用电比重的日益提升，以及第二产业逐渐向中西部转移的趋势，我国整体的用电结构也发生了巨大变化。尤其是大量随城市兴起的工业园区及相关数据中心的建设，使得我国电力需求重点在地理分布上逐渐扩散、趋向均匀化。与上述背景相对应的是，我国总装机容量已超过美国跃居世界第一，但是至少 30%的装机发电能力处于闲置状态，而长距离输变电过程中造成的能耗损失依然无法避免。

鉴于上述情况，电网结构需要从以少量集中的大主力电源为主进行远距离、大容量输送电的方式，转变为以大量、分散的小型发电系统为主、就近生产和消化的模式，即向着分布式能源体系的方向转变。这种转变将推动电网设备和用电设备的小型化和智能化，即向着智能电网的方向发展。智能电网的建设是大势所趋，据统计，2013 年全球与智能电网配套使用的智能电表安装数量已超过 7.6 亿只，到 2020

年智能电网预计将覆盖全世界80%的人口。

智能电网导致的一个必然结果是电网运行控制信息的爆发式增长，由此催生的对海量数据采集、管理和分析的需求使得大数据在智能电网建设和运营过程中的应用成为必然。因此，必须充分认识到大数据在整个智能电网发展过程中的的基础性作用，顺应我国在新发展阶段电力供需的变化趋势，大力推动大数据技术在智能电网规划、设计、建设、运行和维护等各个环节的应用。

智慧医疗

医疗体制的改革是我国社会和经济改革探索的重中之重。随着人们生活水平的不断提高及人口老龄化加速到来，我国的医疗服务需求正在稳步增加，我国 2014 年的健康医疗支出占总支出 23%，预计到2020 年，健康医疗支出占总支出上升到 32%。在医疗服务产业快速发展的同时，我国医疗服务体系仍然存在诸多严重的问题，包括医疗资源在城乡之间和地区之间配置失衡、总体医疗卫生成本过高等。从上世纪 80 年代到 2005 年的 25 年间，我国卫生总费用增长了 52 倍，其中居民个人支付费用增长了 133 倍，两项指标都远远超过了经济总量的增速。

为了解决上述医疗问题，一个根本的思路就是实现患者与医务人员、医疗机构、医疗设备之间的互动，构建医疗健康行业的智能化管控和决策体系，实现资源的合理配置和动态平衡，解决或减少由于医疗资源缺乏所导致的看病难、医患关系紧张、事故频发等现象，从而

全面提升国民医疗服务质量。

从医疗资源和服务接口均衡配置的角度出发，必然会导致医疗相关信息在整个体系不同系统中的流动,涉及一系列的数据采集、转化、标准化和整合工作。下列领域是智能医疗体系建设的重点，也是大数据发挥巨大作用的场景：

- 移动和远程医疗系统的建设。通过各类移动应用，使医护人员能够在远程进行诊断并提供治疗方案。根据 IDC 统计，截至 2013 年 9 月全国范围内 17.5%的三级医院已经使用了移动医疗系统,包括移动查房、移动输液、移动诊断、患者统计、用药统计和移动挂号等。移动和远程医疗模式的核心在于采集患者体征数据和治疗信息，并基于此开展健康咨询类的服务。
- 区域卫生信息化建设。区域卫生信息化建设最基本的需求是让医生信息和患者健康档案能够在不同医疗机构之间实现共享，涉及医生和患者信息的标准化工作。
- 专业科室的信息化建设。专业科室的信息化建设需要将某些以往难以数字化的特有疾病信息记录下来，并实现数据的转化和高效存储，为后期的抽取、统计和挖掘提供便利。
- 数据中心的建设。随着医疗信息化进程的深入，有大量来自多个源头的数据需要实现统一的存储和管理，尤其是新增临床和管理类数据。

此外,引入基因序列分析等大数据技术能够加速新药的研发速度，

以及更有针对性的进行临床开发，降低研发中的风险。这对于我国已进入糖尿病、癌症和心脑血管疾病高发期的现状具有特殊的现实意义。而且，由于人口众多、基因组数据资源丰富，使得我国在相关领域的研究上具有独特的优势。

互联网金融

当前，我国正处于加快转变经济发展方式的关键时期，深化金融体制改革，完善金融监管，推进金融创新，维护金融稳定，成为中国经济发展的整体需要。但是，我国传统金融行业在一定程度上普遍存在着机制僵化的问题，现有的金融服务无法满足大量中小微企业以及个人客户的基本和差异化金融服务需求。

自 2007 年以来，我国互联网产业迅猛发展，逐渐渗透到社会运转的各个领域。其中，涉及到广义金融的互联网应用，被统称互联网金融，包括但不限于为第三方支付、在线理财产品、信用评价审核、金融中介、金融电子商务等。由于增量市场空间巨大，加之传统金融机构在服务质量和服务受众群方面的局限，我国互联网金融市场的发展速度和规模远远领先于发达国家和地区。比如，仅就支付一项来说，2013 年有 153.38 亿笔业务通过互联网完成，金额总计达到 9.22 万亿元。

针对我国金融市场存在的问题，我国互联网金融的发展具有非常现实的意义：

- 以 P2P 网贷为代表的互联网金融模式有助于发展普惠金融，能够

在一定程度上填补传统金融覆盖面的空白，与传统金融形成相互促进、良性竞争和共同发展的局面；

- 有利于发挥民间资本作用，为数额庞大的民间资本提供高效、合理的投资方式和渠道，有力促进实体经济的发展；
- 满足电子商务相关的创业融资、周转融资需求和客户消费融资需求，扩大社会消费；
- 提供有别于传统银行和证券市场的新融资渠道，有助于降低成本，提升资金配置效率和金融服务质量；
- 改善传统金融的信息不对称问题，提升风险控制能力，推出个性化金融产品，满足客户的多样化需求。
- 从制度创新和机制探索角度出发，互联网金融有助于支持市场自律组织履行职能，完善资本市场诚信监管制度，强化守信激励、失信惩戒等机制。

尽管发展迅速，在技术创新上，国内互联网金融行业仍有较大的提升空间。当前，相关企业过于偏重网上数据的收集，模型开发成果较单薄，在以技术为基础的应用创新不够。互联网金融的市场定位主要在“小微”层面，具有“海量交易笔数，小微单笔金额”的特征，需要整合海量的企业、商户及个人的消费、交往、贸易、税务等信息，洞察资本供需两端，评估客户的资信状况。鉴于我国互联网金融的广阔前景和巨大影响力，有必要积极促进大数据理念和技术的应用，推动相关企业积极融合金融数据、社交数据、电子商务交易记录和各类

线下数据，基于大数据分析技术，深刻了解企业、组织或个人之间的关联信息，准确掌握财产、经营、消费习惯和商业道德等各个方面的情况，消除对客户信息的垄断，为中小型企业指明筹资方向,同时为社会提供低成本、高回报的投资渠道。

智慧农业

农业是我国国民经济的基础，也是经济发展、社会安定、国家安全的基础，对于实现我国经济社会长期稳定发展有重大战略意义。当前，我国农业取得了举世瞩目的成就，粮食产量连续 11 年增长，农民收入也实现连续 11 年增长。但是，在农业领域，我国面临的形势不容乐观。

在生产环境上，我国农业发展面临着土壤、水资源、气候等诸多严重问题：我国人均耕地面积不到世界平均水平的一半，2030 年作为我国重点粮食调出区域的东北地区将接近农业需水极限,极端气候发生频率由上世纪 50 年代的不足 20 次发展到 2010 年的 100 多次。由经济起飞拉动的市场需求无法自给,粮食安全问题日益严重，2010 年起我国成为粮食的净进口国，粮、棉、油、糖、肉、奶六大农产品无法完全自给且进口量呈增长态势。食品安全形势严峻，近年来镉大米、瘦肉精、奶粉等食品安全事件频发，根据有关部门统计，每年我国消费者因食物残留农药和化学添加剂中毒的人数超过 10 万人。

针对我国农业发展所面临的种种问题，中央政治局在分析研究 2015 年经济工作的会议中指出，要加快转变农业发展方式，从主要

追求产量增长和拼资源、拼消耗的粗放经营，向数量质量效益并重、注重提高竞争力、注重可持续的集约发展转变。在这个大的指导思想下，2015 年一号文件将农业信息化作为农业现代化的突破口，而大数据、物联网和云计算等技术则是实现农业信息化的基础所在。

大数据对于我国农业发展的推动作用主要表现在以下方面：

1．基于大数据技术对整个农业产业链进行全面、实时的监控，结合诸如天气报告、土壤条件、地图、水资源、市场动态等数据，可以形成对农业整体情况的准确把握和有效的规划。

2．对农业生产过程进行监控和预测，可以提高运营管理和生产效率，有助于农业生产的精准化、标准化和规模化。

3．通过大数据采集和分析流通环节的库存、价格和物流数据，引入农产品期货交易信息，可以及时掌握真实库存，预测市场波动，主动调控生产过程和生产布局。

4. 通过传感器、条形码和 RFID 等采集和识别手段，运用大数据和云计算技术建立农产品信息管理平台，构建覆盖产地、品种、土壤、水质、病虫害、农药、化肥、储藏、加工、运输、销售等环节的农产品安全追溯体系。

可见，采用大数据研究手段，在搜集、存储气象、水利、农资、农业科研成果、动物和植物生产发展情况、农业机械、病虫害防治、农产品加工等诸多环节大数据的基础上，通过专业化处理，对海量数据进行快速分析挖掘，能够为政府、企业和农户的决策提供支持，对

保障我国农业安全、提升农业生产水平、促进农产品市场健康发展等具有重大意义。

人工智能技术商业化

大数据需要对多源、海量数据进行自顶向下的挖掘和关联，其中相当部分的研究领域与人工智能技术相关。严格的人工智能概念起源于用计算机来解释人类思考过程的想法，在上世纪 50 到 70 年代间，西方发达国家政府和企业界投入了大量的资金来资助人工智能领域的研究。尽管如此，直至上世纪末期，人工智能领域仍偏重于学术研究的性质，在实用方面取得的进步相对有限。

随着互联网的兴起，机器与人之间的连接和互动日益紧密，各类产品和服务需要更为清晰地理解人的意图，并且更好的满足用户需求，使得人工智能获得了广阔的实践土壤和应用空间。其中，作为人类信息表达最主要的三种方式，通过计算机来模拟人类大脑对文本、音频和图像的分析和识别，是人工智能研究的热点所在，也成为大数据发展的基础性支撑技术之一。

基于语音、图像和文字的识别是大数据领域的基本研究内容和各类成功应用的基础，具有巨大的产业前景，据预测，未来五年，基于语音和图像的搜索将达到全球搜索份额的 50%以上。人工智能技术的快速发展，将极大推动社会管理、智能交通、智慧医疗的发展，而上述领域也正是我国大数据应用的重点所在。

传统上，对文本、音频和图像的分析都需要研究人员在具备一定

专业知识的前提下，耗费大量时间对各种规则进行手动编程。而随着数据量的增长，人工智能领域的研究思路发生了巨大的转变，深度学习等新方法可以让计算机对海量样本数据进行自动的学习和建模。Google 研究部主任 Peter Norvig 对此的描述是：“All models are wrong, and increasingly you can succeed without them”。比如，Google 用 1000 台电脑组成的神经网络，花费了三天时间来分析约一千万张静态图片，最终能够自动识别人脸、身体和猫。计算准确性和效率的大幅提升，使得人工智能技术的大规模应用成为可能。

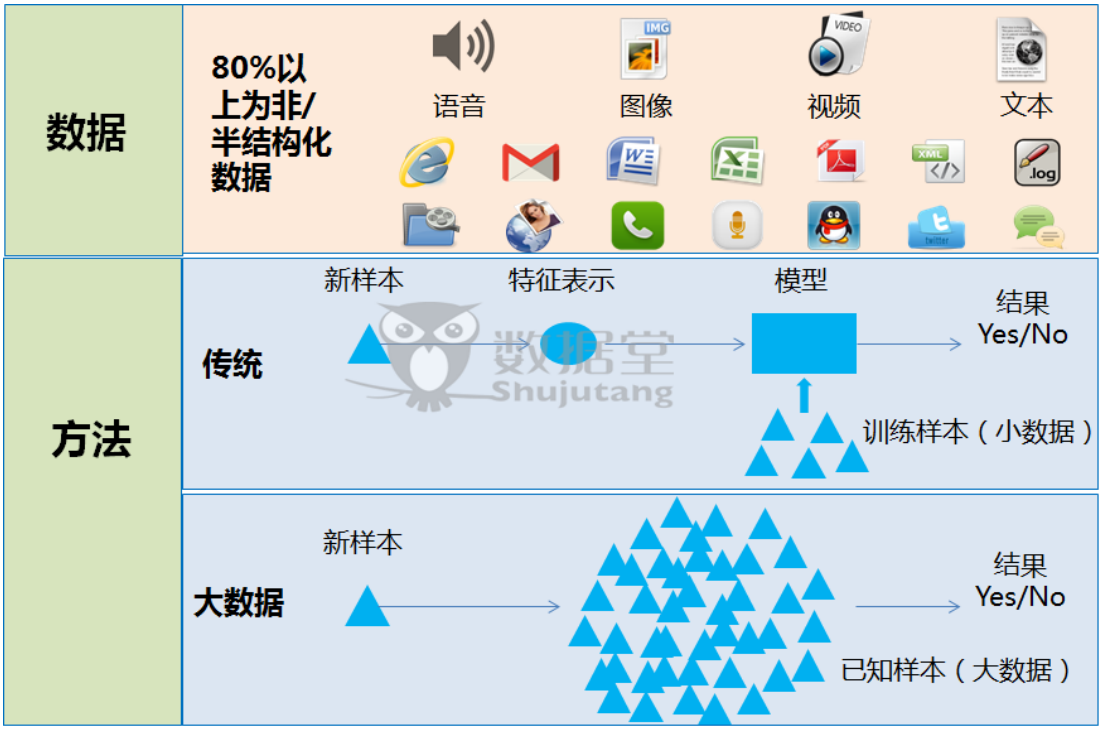


图. 大数据时代的人工智能研究

人工智能技术的发展具有高度的本土化色彩，而且图像、语音、文本数据的处理和分析严重依赖于训练数据的规模和质量，导致我国在人工智能商业化方面具有天然的优势。传统的人工智能技术研究偏重于模型方面，而随着互联网和移动互联网的发展，产生了海量的图

像、语音和文本数据，为相关学术研究和技术发展带来了新的思路。基于海量的训练数据，配合以日益提升的计算能力和大数据处理技术，能够在可接受的时间内让计算机自动实现对模型的识别，不仅提升了分析的准确性，而且大大加速了技术研究和发展的进程。在这种以数据驱动的发展模式下，加之我国庞大的人口基础和网民数量所导致的数据规模，为相关领域的研究和发展提供了坚实的基础，同时构成了人工智能技术商业化的广阔空间。

（一）语音识别

语音识别的基本原理是以人类语音为研究对象，通过信号处理和模式识别让机器自动识别和理解人类语言。语音识别涉及声学、语音学、语言学、信息理论、模式识别理论以及神经生物学等学科，正逐步成为计算机信息处理技术中的关键技术。

从近年互联网和移动互联网的发展趋势看，服务接口的便利性成为企业获得竞争优势的根本所在，极高的沟通效率使基于语音技术的互动方式必将成为未来人机对话的主要模式。同时，通过在海量音频数据中捕捉和提取客户的意向和情绪，还能够实现业务流程、座席绩效和客户体验等的优化。语音技术在智能设备、车联网、在线教育、电信、零售、医疗、公共事业、运输/物流、旅游、传媒、能源等多个行业具有广阔前景，根据 Research and Markets 公司的预测，全球语音分析市场预计将从 2014 年的 4.56 亿美元增长至 2019 年的 13.3 亿美元，年均复合增长率高达 23.9%。

我国互联网和移动互联网的迅速发展为语音技术提供了广阔的应

用空间，而诸如科大讯飞、云知声等国内厂商在语音解决方案上的基本成熟，也使得语音技术在各个领域的应用呈爆发态势。据《2014中国智能语音产业发展白皮书》显示,随着在移动互联网、呼叫中心、智能家居、车联网和教育等领域的应用逐步深入，我国智能语音产业保持了持续、快速的增长，在未来将进一步带动工业、家电、通信、医疗、家庭服务、消费电子产品等众多领域的发展。

除了市场规模庞大以外,我国语音还有着独特的市场需求。首先，与英语等语言相比，中文语序表灵活、省略现象严重，在断句、词性判定、语序规整和词汇组合等方面需要进行专门的研究；其次，我国是个多民族国家，幅员辽阔，各民族语言和各地方言的差异给语音服务市场带来了多样化的需求。独特的技术挑战与细分的市场需求为我国语音服务产业提供了宝贵的市场空间和发展机遇。

（二）图像/视频理解

在日常生活中，人们感知的外界信息有 80%以上来自视觉系统。随着社会信息化程度的提高，对图像信息的自动化分析、理解的需求也变得越来越迫切。作为人工智能研究领域的一个重要分支，图像识别技术的目的在于用计算机实现对图像信息或图像模式的处理和分析，进而描述、识别和解释其中的物体对象或行为。

互联网和移动互联网的发展产生了海量的图像信息。据 2014 年的统计，Instagram 每天上传的图片量为 6000 万张，Whatsapp 每天的图片发送量为 5 亿张，国内的微信、微博和淘宝等电子商务平台也拥有了庞大的图像数据资源。数据资源的丰富引发了图像识别技术的巨

大进步，国内外互联网巨头如 Google 和百度等都在这方面取得长足进步。在应用环节，图像识别技术开始在互联网等领域得到大规模的应用,并逐渐渗透到其他行业，成为一个基础性的支撑服务。

图像识别技术在社会治安、智能交通、金融、工业、食品检测等诸多领域具有广泛的应用前景。目前，已有的主要应用场景有：

- 图像搜索。图像搜索未来将成为互联网和移动互联网的主要入口之一。通过将图像与其关联信息的实时、动态整合，能够实现极高的商业价值。比如对个人名片、杂志封面、电影海报、商品实物、店铺标志、衣物饰品的识别，能够实时与评论文章、门店地址、厂家信息、营销活动等关联起来。据百度统计，目前图像搜索应用中最为旺盛的是生活类搜索服务，在整个图像搜索中占比达 35.5%。
- 身份识别。高效可靠的身份认证技术在社会安全中起着至关重要的作用。近年来，以图像识别为核心的生物特征认证技术逐渐被应用于监控摄像、刑侦识别和金融支付等领域，包括对指纹、虹膜、人脸、掌纹、手形和耳形等人体生理物理特征的识别。比较典型的应用如阿里旗下的蚂蚁金服和腾讯的微众银行都基于人脸识别技术进行用户的
- 智能交通。智能交通系统需要捕获周边环境和交通工具状态的图像，并进行实时的处理和识别。比如车辆牌照自动识别系统，需要实时、自动地对含有车牌的图像进行分析处理，从而确定牌照在图像中的位置，并进一步提取和识别出文本字符。而对于无人

驾驶汽车来说，需要实时感知并识别出车辆周围环境，并结合道路、位置和障碍物信息来控制车辆的转向和速度。

- 工业应用。在工业生产领域，图像识别技术已成为自动化生产控制系统的核心技术之一，被广泛应用于质量检测与评估、快速测量、自动分拣以及智能工业机器人的视觉定位与环境感知等方面，极大提升了电子、汽车、纺织、印刷以及制造加工等行业的生产效率。
- 医疗健康。医疗健康行业拥有大量反映病患身体内部解剖学或生理功能信息的图像数据，医疗图像具有规模庞大、分辨率高和图像特征表达复杂等特点，使得图像识别技术在医疗领域具有极大的实用价值，可应用于医疗诊断、组织容积定量分析、病变组织定位、解剖结构学习、治疗规划、功能成像数据局部体效应校正和术后监测等各个环节。

（三）文本分析

文本挖掘是指对无结构的原始文本进行科学抽象和模型构建，转化为结构化的、计算机可以识别处理的信息，进而使计算机能够基于已有模型识别文本，并对散布在文本中知识进行提取和组织。例如，由LexisNexis公司开发的HPCC系统,通过整合来自不同系统的数据，抽取人名、地名、公司名以及其他重要信息;安全公司OpenDNS公司，基于自然语言处理理论（Natural Language Processing），提前识别出刻意模仿著名站点名称的恶意钓鱼网站。

由于互联网的迅速发展，以社交媒体为代表的非结构化文本信息

呈爆炸式增长态势，推动了文本分析领域的快速发展。当前，文本数据主要包括博客、微博、微信、设备日志与客服对话记录等，基本都以人类语言的形式呈现，使得文本分析的核心逐渐转向人工智能研究中的自然语言理解领域，包含词法分析、依存分析、句法分析和机器翻译等。

以自然语言处理为核心的文本分析技术，属于大数据分析中最为基础的部分。自然语言处理是研究人与计算机交互的语言问题的一门学科，是语言信息处理的一个分支，也是是人工智能领域的核心课题之一。由于文本分析的应用支撑面极为广泛，在此只描述几类最具代表性的应用场景：

- 互联网服务。诸如百度、淘宝等大型互联网平台的信息搜索系统，能够直接回答用户提问的知识引擎，各类基于机器翻译技术开发的在线词典等。此外，在刨除音频特征之后，语音识别及其应用服务也是以自然语言处理为基础。
- 企业营销。对用户的社交媒体内容进行分析、掌握用户的性格、年龄阶段、星座、性别、偏好等。在对用户全面刻画的基础上，向用户推送相应的折扣、优惠和最新产品信息。
- 金融业务。金融信息中的的绝大部分数据均是以文本形式存在，如交易信息、金融论坛、研究报告、财经新闻和社交媒体等，通过文本分析可以用于市场洞察、信用评估和风险管理等方面。比如前文案例所述，有公司基于社交媒体预测市场走势，进而对股票操作进行指导。

- 社情民意。当前，网络论坛和社交媒体中存在大量以非结构化数据形式出现的舆情信息，其中蕴含真实而广泛的对某种社会现象或社会问题的看法，分析提供了方法和技术支持，通过分析可以及时掌握民众所关心的热点、难点和舆情动态，为合理决策和突发事件预防提供重要依据。
- 医疗。医疗档案是病患在医疗机构就诊过程中产生的完整、详细的临床信息资源。医疗档案中包含大量的非结构化文本信息，例如以自然语言记录的临床表现等医疗记录，运用相应的文本分析技术，可以有效提升医疗服务的质量。

与语音识别类似，我国语言类型多样，包含汉语和各类少数民族语言，拥有足够细分的子领域和应用场景，为国内厂商提供了巨大的发展机遇。

数据银行

针对我国信息化建设和大数据发展较先进国家和地区仍相对落后的现状，有必要从全局推动数据银行一类实现全社会数据资源供需的产业形态的发展。

首先，我国大多数政府数据和企业数据仍然处在沉睡状态。对数据外部性认知不足，大部分数据拥有者无法意识到自身数据的资产属性，缺乏足够的动力将自己的数据公开。由于缺乏足够的利益驱动，企业对数据资源的垄断意识仍较强烈，尤其是一些大型企业往往不愿

意把自己的数据资源向自己业务圈外的市场提供。

其次，即便有主观意愿通过交易来实现数据增值和业务成长，企业仍面临着成本消耗过高的问题。对于数据拥有者，必须经历陡峭的学习曲线去探知自身数据在其他领域的应用价值，而数据需求方在寻找所需数据时可能会耗费大量时间和人力成本。

最后，由于对数据利用认知的不足以及技术手段上的局限，我国在数据采集和处理方面仍相对落后，各类数据源在质量和准确性方面缺乏一致的标准。低质量、混乱的数据会导致错误的分析结果，进而对用户的决策造成负面影响。因此，必须要有对数据质量进行规整的产业环节，将数据资源转换成易于为市场所理解和使用的形态，提升数据资源商品化、标准化和资产化的水平，从而盘活数据资产，带动资源的优化配置，有效推动大数据产业以及其他行业的发展。

可见，积极发展数据银行一类的产业形态，能够深化全社会对于数据外部价值的认识，通过汇集各类数据供需方并提供必要的数据商品加工手段，降低实现和利用数据价值的成本，打造我国大数据产业快速成长所需的开放、透明、资源高度聚集的市场环境。

众包模式

数据的准确性、实时性和覆盖面等质量指标是关乎大数据产业发展的关键所在，针对我国数据采集基础较弱的情况，众包模式在某些领域具有巨大的应用价值。

随着我国城市化进程的发展，城市资源和环境的限制日益明显，

交通方面的挑战尤为严峻，对于数据的准确和及时程度有较高的要求。当前，我国在这方面的数据采集大多通过雷达、摄像头、传感器和实地观测等方式，耗时耗力且难于维护。在环保领域，环境监测数据是预测、预报环境质量状况的重要基础，关系到能否对环境质量、生态环境现状及变化趋势进行实时、准确的监测。目前，由于过分依赖环境监测仪器、测试手段和数据传输方式不够完善等问题，使得监测得到的数据过于稀疏、数据量过小且实时性不够。

通过众包模式，比如基于民众随身的移动设备来进行采集交通或城市环境数据，能够有效提升所需数据的真实性、密度和实时程度。正如前文中所述，当前国外已有通过民众的智能手机来实时采集公交信息的案例，而 2013 年全球电信日也将基于运营商数据来改善交通状况作为主题之一。

在自然环境监测和灾害预警方面，由于我国自然环境日益恶化、地质灾害频发，众包模式也具有极大的应用价值。根据研究显示，地震发生时如果能提前 10 秒预警，生存率可以增加 12%，提前 30 秒，生存率能增加到 40%。而日益普及的智能设备为环境和灾害信息的采集和分发提供了高效通道，能够大幅提升信息采集的覆盖面和预警的实时性。这方面的研究和探索已有先例。美国地质勘探局和航天局尝试利用众包 GPS 及其他数据，监测地震发生时的地面移动情况并快速预警。在环境监测领域，WeatherSignal 应用基于用户手机中内置的气压计、湿度计、温度计和照度计等传感器，实时采集天气数据。

我国具有基于众包模式开展数据采集的天然优势。我国拥有庞大

的互联网和移动互联网用户群体，据统计，截止到 2014 年 12 月，我国网民规模达 6.49 亿，手机上网用户 5.57 亿，微信用户 3.5 亿、智能连接设备近 8 亿，构成了巨大的信息采集和发布网络，几乎覆盖全部国土空间，可以更加实时和广泛地汇集各类自然和社会信息。基于这种得天独厚的大规模感知系统，可以全方位监测人口移动、经济运转、交通运输和自然环境等各个方面的实时状况。

目前，众包数据采集在我国已有成功实践。比如，数据堂通过数十万众客采集语音和图像数据，为我国在语音识别和人脸识别等领域的发展提供了坚实的基础。而百度、腾讯等地图服务的上游数据供应商也逐渐通过众包模式采集传统方法难以应对的各种数据，比如海量街景图片。

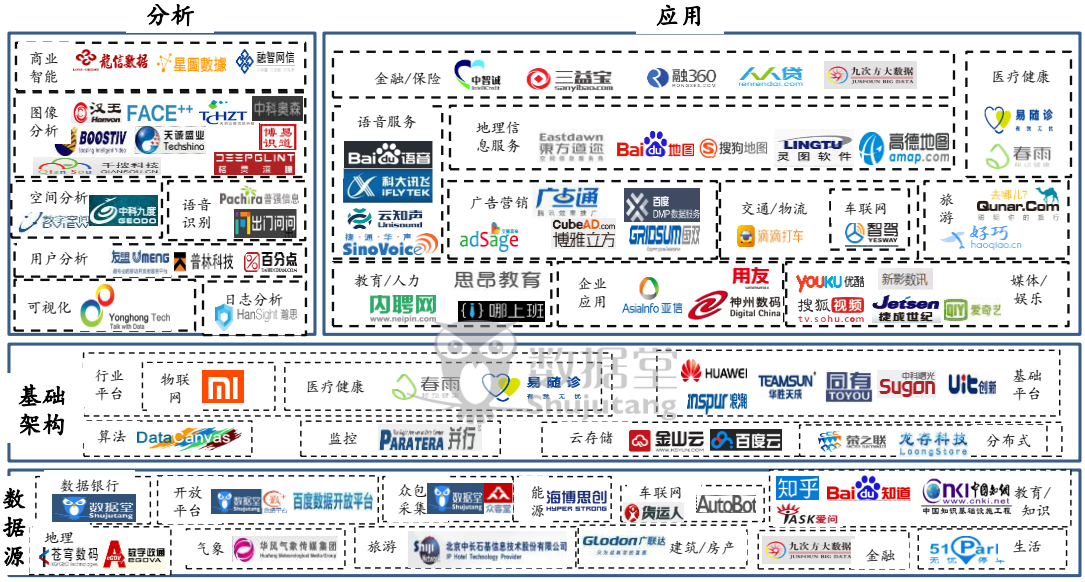
综上所述，应该大力提倡和推动众包模式在数据采集环节的运用，这不仅是实现我国大数据产业快速发展的有效途径，更是提升政府和企业决策科学性、降低灾害损失、确保社会可持续发展的有力手段。除了采集之外，在数据分析和价值挖掘环节，众包模式同样能够发挥巨大作用。针对我国数据利用极不充分的现状，充分调动高校、科研机构和企业研发力量，进行技术和应用的探索，能够大力推动我国大数据产业以及相关行业的发展。

4.5 海淀区大数据产业发展策略

根据各地实际情况的不同，对于大数据发展的策略也应该因地制宜、有所侧重。以海淀区为例，作为我国科技领域和 IT 产业的高地，

在大数据产业发展现状上体现出了鲜明的特色。相应地，也应有针对性地制定大数据长远发展的规划。

4.5.1 海淀区大数据产业现状



来源：数据堂整理

图. 海淀区大数据产业分布

海淀区是我国 IT 产业的主要发源地，在大数据产业发展上也在全国处于绝对的领先地位。已统计的大数据企业中，北京、上海、广东（主要是深圳）和浙江（主要是杭州）的占比达 92%，其中北京处于遥遥领先的地位，全国占比接近 60%。

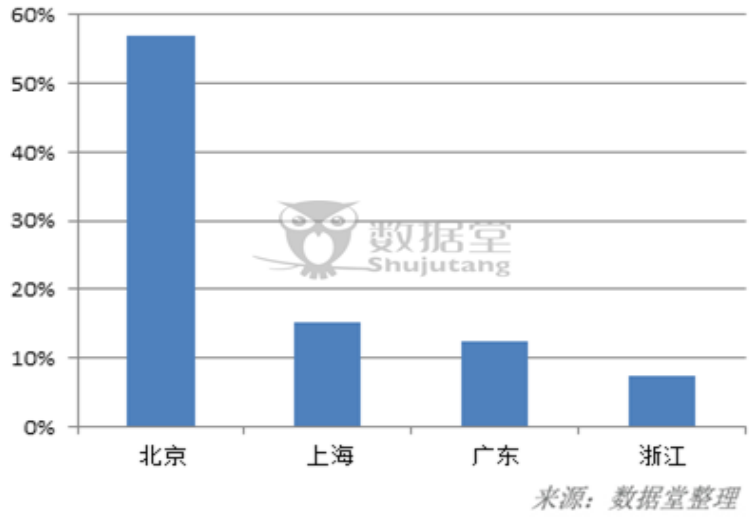


图. 大数据企业地区分布

在北京的大数据企业或产品中，海淀区又占有绝对的优势地位，占北京大数据企业的 63%，在全国来看占比接近三分之一。

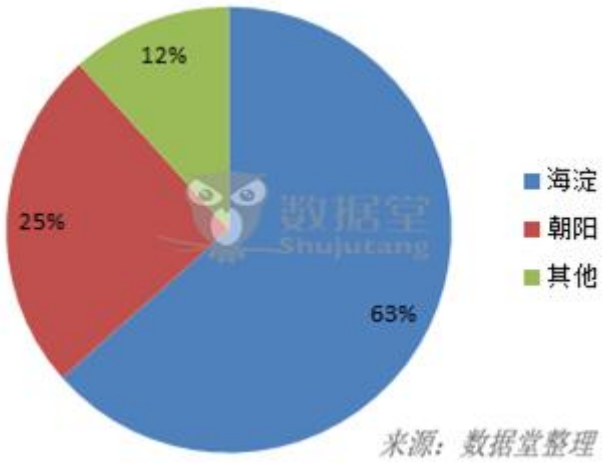


图. 北京大数据企业按区分布

如果细化到产业链的各个环节，可以看出海淀区大数据产业分布的特点所在：

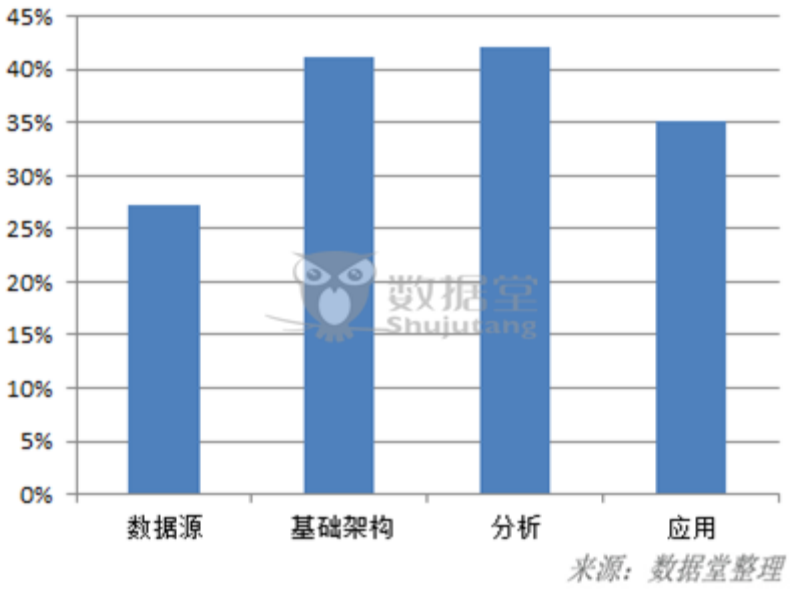


图. 海淀区大数据企业全国分类占比

如上图所示，在基础架构和分析类海淀区企业的占比明显高出数据源和应用类。而且，在基础架构和分析这两大区块中，海淀区大数据企业基本实现了对子类别的全覆盖，尤其是在数据库、分布式系统、管理工具、可视化、机器学习、图像识别和语音识别等方面，国内仅有的少数企业基本都是起源于海淀区。

在应用方面，以地理信息、人力/教育、金融/保险、影视/娱乐、企业应用和语音服务最多；在数据源方面，主要偏重于地理信息和互联网用户信息，影视娱乐、个人健康、金融、能源和交通数据都比较缺乏。可以看出，基本都是由互联网企业推动为主，传统产业较少涉及，这也是和海淀区本身的产业特点决定的。

对于海淀来说，在大数据产业发展方面，可以从数据流通和技术创新两个方面入手。

4.5.2 海淀大数据产业发展建议

推动数据开放流通

大数据产业的发展涉及诸多方面，但其中最为关键的一点就是数据的开放和流通，只有极大丰富的各类数据能够为产业链其他环节所见所用，大数据才能在各行各业发挥出最大的价值。结合各国政府的大数据战略举措，构建数据流通生态最为有效的方法就是由政府推动数据的开放。

政府在具体操作和风险规避上具有较大优势。数据开放涉及数据采集、存储、管理、分析、交付及隐私保护等方面，只有政府能够切实推动上述各个环节的实践，为大数据产业的良性发展坚实的基础。同时，金融、能源、电信等公共服务部门富含大量高价值数据，政府在推动上述领域数据的开放上比企业具有无可比拟的优势。

政府数据本身富含极高的价值，可以改变目前大数据在我国社会管理、农业发展和公共服务等领域应用薄弱的局面。目前，由于庞大的国土、人口和经济规模，我国已成为仅次于美国的数据大国，其中三分之一的数据属于政府及提供公共服务的机构与企业。大数据之所以在我国政府和农业等领域应用案例稀少，数据的封闭是首要原因。

为了推动数据开放和流通的工作，可以考虑由政府牵头进行数据开放平台的建设或对接已有的数据流通平台，以数据的聚合为牵引，向全社会进行推广，进而逐步推动数据流通所涉及的各项工作走向深入。

孵化大数据技术创新

大数据科学作为一个新兴的交叉学科方向，其研究和实践必须由不同的学科领域协作完成，比如计算机科学、统计学、人工智能、社会科学等。海淀在科研力量方面优势明显，在技术色彩最浓的基础架构和分析环节属于国内领头羊的地位，完全具备成为我国大数据技术创新基地的条件。

在具体措施上，可以通过产业园区规划和创业创新大赛等措施形成大数据产业链在海淀的聚集，尤其是注意引入数据源类企业，挖掘现实的数据存储、管理和分析需求，让海淀广大科研院所和企业研发力量能够切实把握大数据基础架构和分析技术方面的未来方向，孵化出一批小而美的专业技术型企业。

值得一提的是，目前已经商用的大数据管理系统或工具，还没有全面响应大数据系统的需求，而且大多由开源项目发展而成，这非常有利于我国在大数据时代打破国外厂商把持信息产业基础软硬件系统的局面，对于具有大量信息技术领域科研力量的海淀来说更是难得的历史机遇。

附录：大数据企业名录

下表罗列了具有一定代表性的国内大数据企业名录和大型企业名下的大数据相关产品或服务。

注册地信息:“海淀”为北京工商行政管理局海淀分局,“北京”为北京工商行政管理局,其他地区企业则只具体到省/直辖市一级(广东省情况特殊，广州和深圳单列)。

企业地点信息：指当前企业总部或主要机构所在地 -- 本报告中的地域相关分析以此指标为准。

	公司	简介	地点	分类	注册时间	注册地点
1	北京九章云极科技有限公司（Data Canvas）	面向企业,提供大数据基础环境和常见算法库	海淀	基础架构;分析	2013.2.6	海淀
2	广联达软件	售卖建筑材料价格信息	海淀	数据源	1998.8.13	北京
3	上海鲁班软件有限公司	建筑业数据汇集平台	上海	数据源	2005.8.16	海淀
4	北京四维图新科技股份有限公司（腾讯入股）	地图和地理数据提供商	海淀	数据源	2012.12.03	北京
5	高德信息技术有限公司（阿里入股）	地图和地理数据提供商	海淀	数据源;应用	2006.8.24	海淀
6	北京灵图软件技术有限公司	地理信息服务提供商	海淀	数据源;应用	1999.4.1	海淀
7	北京长地万方科技有限公司（百度）	导航电子地图测绘,为百度提供地理数据	石景山	数据源	2003.2.19	海淀
8	深圳市凯立德科技股份有限公司	地理数据提供商	深圳	数据源	1997.12.12	深圳
9	北京城际高科信息技术有限公司	地理数据提供商	西城	数据源	2000.1.27	海淀
10	科菱航睿空间信息技术有限公司（腾讯收购）	为奇瑞、和华泰等汽车厂商提供地图数据	朝阳	数据源	2005.2	北京
11	亿赞普（中国）网络技术有限公司	基于运营商数据,提供广告及商业智能服务	朝阳	应用	2010.3.1	北京
12	星图数据（Syntun）	为企业提供竞争对手及自身运营方面的分析	海淀	分析	成立于美国，中国有技术团队	

13	华院数据技术（上海）有限公司	提供营销、客户关系管理和决策支持的分析服务	上海	分析	2002.4.23	上海
14	北京闪银奇异科技有限公司（wecash）	中国 Zestfinace，基于社交数据，完成个人授信。IDG 投资 4000 万	朝阳	应用	2014.4.23	石景山
15	宜信	通过征信对象在各电商平台及社交媒体上的信息，进行信用评估	朝阳	应用	2006	北京
16	上海拍拍贷金融信息服务有限公司	P2P 网贷	上海	应用	2011.1.18	上海
17	上海陆家嘴国际金融资产交易市场股份有限公司（陆金所）	平安集团旗下，P2P 业务	上海	应用	2011.9	上海
18	有利网（北京弘合柏基金融信息服务有限责任公司）	P2P，信用模型来自于美国 FICO	海淀	应用	2012.5.31	北京
19	人人贷金融信息服务（北京）有限公司	P2P 业务	海淀	应用	2014.2.11	海淀
20	三平伟业(北京)投资管理有限公司（三益宝）	P2P。搜集企业收入、固定资产、债务等数据，结合银行的征信数据，评估信用等级	海淀	应用	2012.10.30	海淀
21	北京华胜天成科技股份有限公司	承建各类大数据平台，开始涉足互联网金融	海淀	基础架构	1998.11.30	北京
22	深圳前海征信中心股份有限公司	征信，隶属于平安保险	深圳	应用	2013.8.8	深圳
23	鹏元征信有限公司	征信，人行与深圳市政府背景	深圳	应用	2005.4.8	深圳
24	中诚信征信有限公司	我国第一家从事全国性信用评级和信息服务的非银行类金融机构	东城	应用	2005.3.23	北京
25	中智诚征信有限公司	征信	海淀	应用	2013.9.11	北京
26	考拉征信服务有限公司	暨原来的拉卡拉（北京）信用管理有限公司。属于联想系统，拓尔思入股	海淀	应用	2015.1.29	海淀
27	北京华道征信有限公司	征信	朝阳	应用	2013.12.23	北京
28	浙江电融数据技术有限公司（元宝铺）	第三方信贷平台，以电商卖家数据为授信依据的短期纯信用贷款	浙江	应用	2014.3.28	浙江

29	股票赢家（上海财新信息科技有限公司）	提供实时的股票交易信息;未来计划参与到征信产业链中	上海	应用	2013.9.25	上海
30	九次方财富资讯（北京）有限责任公司	九次方大数据平台,据称掌握 900 万家公司数据	海淀	数据源；应用	2010.8.5	朝阳
31	浙江核新同花顺网络信息股份有限公司	为客户提供全球金融市场行情数据	西城	数据源;应用	2001.8.24	浙江
32	上海万得信息技术股份有限公司	即万得（wind）咨询,提供类似彭博终端的产品	西城	数据源	2005.4.4	上海
33	通联数据股份公司	掌握大量企业信息,类似九次方大数据	上海	数据源	2013.12.28	上海
34	北京融世纪信息技术有限公司（融 360,百度背景）	提供贷款、理财和信用卡产品的搜索服务。融资总额 1 亿美元	海淀	分析;应用	2011.11.10	海淀
35	深圳祥云信息科技有限公司	中科院计算所有关。将复杂事务处理、CUDA、神经网络等应用于股票交易	深圳	分析;应用	2011.8.29	深圳
36	融智网信（北京）管理咨询有限公司	金融行业数据技术解决方案	海淀	分析	2010.9.26	朝阳
37	恒生电子股份有限公司（马云以 33 亿元入股）	主营金融 IT 产品与服务,掌握各大金融机构重要数据,全面覆盖客户的各类交易记录	浙江	数据源	2000.12.13	浙江
38	杭州又拍云科技有限公司（upyun）	云存储,主要针对图片和视频	浙江	基础架构	2014.6.17	浙江
39	上海七牛信息技术有限公司（七牛云）	云存储,擅长文件管理	上海	基础架构	2011.8.3	上海
40	上海庆科信息技术有限公司	同 broadlink 类似,实现更为底层,从芯片入手。与阿里云合作	上海	基础架构	2010.1.27	上海
41	杭州古北电子科技有限公司（broadlink）	物联网基础模块,实现数据的采集、传输和存储	浙江	基础架构	2013.7.30	浙江
42	机智云（广州杰升信息科技有限公司）	为智能硬件提供云平台,沉淀各类硬件设备的数据并进行统计分析	广州	基础架构;分析	2006.9.18	广州
43	亿方云（杭州亿方云网络科技有限公司）	文件管理云平台	浙江	基础架构	2013.12.6	浙江
44	北京同有飞骥科技股份有限公司	存储系统研发	海淀	基础架构	1998.11.3	北京

45	杭州宏杉科技有限公司	存储系统研发	浙江	基础架构	2010.5.27	浙江
46	曙光信息产业股份有限公司	计算和存储设备,大数据系统平台研究	海淀	基础架构	1995.6	北京
47	浪潮集团有限公司	计算与存储设备、大数据系统平台研究	海淀	基础架构	1989.2.3	北京
48	上海爱数软件有限公司	数据管理,备份一体机	上海	基础架构	2006.9.18	上海
49	杭州信核数据科技有限公司	数据安全方案提供商	浙江	基础架构	2006.7.11	浙江
50	创新科存储技术有限公司	存储方案提供商	海淀	基础架构	2005.11.24	海淀
51	北京龙存科技有限责任公司	分布式存储产品研发	海淀	基础架构	2007.7.23	海淀
52	神州数码信息系统有限公司	为行业客户提供各类IT 解决方案与服务供应商	海淀	应用	1998.12.25	海淀
53	深圳金蝶中间件有限公司	为行业客户提供 IT解决方案与服务供应商,在医疗行业比较突出	深圳	应用	2000.8.29	深圳
54	东软集团股份有限公司	为行业客户提供各类IT 解决方案与服务供应商	海淀	应用	1991.6.17	辽宁
55	亚信集团股份有限公司	为行业客户提供各类IT 解决方案与服务供应商,主要针对运营商	海淀	应用	2009.12.6	海淀
56	北京用友科技有限公司	为行业客户提供ERP、CRM、人力资源管理、商业分析等系统方案	海淀	应用	1999.4.15	海淀
57	北京永洪商智科技有限公司	各种可视化产品:仪表盘、报表、即席、OLAP	海淀	分析	2012.2.17	海淀
58	龙信数据	提供数据管理和分析方案，偏重政府行业	海淀	分析	2010.10.21	海淀
59	北京荣之联科技股份有限公司	提供数据中心解决方案，有基于Greenplum的方案	海淀	基础架构	2001.3.12	北京
60	上海天玑科技股份有限公司	类同荣之联,使用了高速网络 infiniband	上海	基础架构	2001.10.24	上海
61	苏州思必驰信息科技有限公司	智能语音服务提供商	江苏	分析;应用	2007.10.26	苏州
62	北京捷通华声语音技术有限公司	智能语音服务提供商	海淀	分析;应用	2000.10.28	海淀
63	北京云知声信息技术有限公司	提供音识别服务中间件和语音识别服务云平台	海淀	分析;应用	2012.6.29	北京

64	北京中科大讯飞信息科技有限公司（科大讯飞股份有限公司）	智能语音服务提供商	海淀	分析;应用	2004.7.6	海淀
65	厦门市美亚柏科信息股份有限公司	公安数据管理和分析	福建	分析	1999.9.22	厦门
66	北京东方国信科技股份有限公司	主要为电信业提供 BI 产品	朝阳	应用	1997.7.28	北京
67	西安美林数据技术股份有限公司	各行业的数据挖掘方案	陕西	分析	1998.3.16	西安
68	北京集奥聚合科技有限公司	以非 cookie 技术为基础，提供用户洞察、实时广告和私有 DMP 方案	东城	分析;应用	2012.8.10	海淀
69	北京缔元信互联网数据技术有限公司	类似集奥聚合,采集网络用户的行为数据,为广告业服务	东城	分析;应用	2007.2.28	东城
70	易达讯网络科技（北京）有限公司	建设全国人口库和法人库,拥有海量个人和企业相关的数据	朝阳	数据源	2000.6.5	海淀
71	北京融信汇智科技有限公司	基于运营商数据进行分析,应用于旅游业和智慧城市项目	海淀	分析	2013.11.18	海淀
72	北京并行科技有限公司	系统管理和性能监测	海淀	基础架构	2007.2.15	海淀
73	北京超图软件股份有限公司	地理信息系统基础平台研发,为政府和企业提供相关技术的咨询服务	朝阳	分析	1997.6.18	北京
74	北京中天博地科技有限公司	土地规划、国土资源数据采集及后端支撑系统	朝阳	数据源;分析	2006.7.10	朝阳
75	南京国图信息产业股份有限公司（GTMAP）	土地规划、国土资源数据采集及后端系统建设	江苏	数据源;分析	2001.3.16	南京
76	北京苍穹数码测绘有限公司	国土资源数据采集、管理、分析	海淀	数据源;分析	2001.5.25	北京
77	北京数字空间科技有限公司	地理信息分析,起源中科院地理所	海淀	分析	2000.8.2	海淀
78	广州中海达卫星导航技术股份有限公司	采集大量测绘、卫星和街景数据	广州	数据源;应用	2006.6.21	广州
79	随便走 APP(深圳市感知网络有限公司)	基于真实图片,实现最后一公里的导航	深圳	应用	2013.7.4	深圳
80	北京东方道迩信息技术股份有限公司	拥有多颗国际卫星数据,提供卫星数据加工和地理信息服务	海淀	数据源;应用	2001.11.22	海淀

81	武汉禾讯农业信息科技有限公司	利用卫星数据,判断农作物长势,估算农业产量	湖北	数据源;应用	2009.6.17	武汉
82	中科宇图天下科技有限公司（遥感所背景）	采集环境数据、提供地理信息服务	朝阳	数据源;应用	2001.11.07	朝阳
83	杭州海康威视数字技术股份有限公司	视频监控,面向安防领域	浙江	基础架构	2001.11.30	浙江
84	浙江大华技术股份有限公司	视频监控方案	浙江	基础架构	2001.3.12	浙江
85	杭州中威电子股份有限公司	视频监控,面向安防领域	浙江	基础架构	2000.3.14	浙江
86	博康智能网络科技股份有限公司	视频监控方案,偏重交通领域,有智能交通产品	上海	基础架构;应用	2008.1.15	上海
87	北京百分点信息科技有限公司	采集消费者偏好信息,为企业提供 BI 优化方案	海淀	分析	2009.7.1	海淀
88	上海晶赞科技发展有限公司	数字广告技术及数据服务商,主要是做受众分析	上海	分析;应用	2013.7.12	上海
89	精硕世纪科技（北京）有限公司	提供互联网广告分析、监测和定向投放的支持	东城	分析;应用	2010.8.13	北京
90	北京学之途网络科技有限公司（秒针系统）	广告监测,帮助广告主评估和优化数字广告效果	朝阳	应用	2005.11.30	海淀
91	北京艾德思奇科技有限公司	互联网广告定向投放技术研发	海淀	应用	2007.3.19	海淀
92	北京国双科技有限公司	提供基于数据分析的在线业务优化解决方案	海淀	应用	2005.12.15	海淀
93	上海智子信息科技有限公司（智子云）	中国的“Criteo”，互联网广告效果评测	上海	分析;应用	2012.8.8	上海
94	博雅立方	主要基于社交媒体数据，提供营销方案	海淀	应用	2008.11.26	海淀
95	时云医疗科技（上海）有限公司	健康数据采集和分析:设备+后端系统+APP。主要的意义在于采集	上海	数据源;应用	2012.11.27	上海
96	涟漪	根据职称、论文、口碑等信息为患者推荐医师	朝阳	数据源		
97	深圳华大基因科技有限公司	基因测序巨头	深圳	分析;应用	2008.8.12	深圳

98	解码(上海)生物医药科技有限公司	基因检测及健康服务	上海	分析;应用	2011.8.12	上海
99	丁香园（观澜网络（杭州）有限公司）	腾讯 7000 万美元投资。面向医疗行业从业者,提供专业知识的交流平台	浙江	应用	2010.1.8	杭州
100	春雨医生（北京春雨天下软件有限公司）	提供病患健康数据采集和管理平台,供医生参考	海淀	基础架构;应用	2011.7.21	海淀
101	易随诊 APP（西部天使（北京）健康科技有限公司）	供病患和医生使用;对病历进行统一管理和检索	海淀	基础架构;应用	2003.7.1	海淀
102	沸腾时刻 APP(深圳市沸腾时刻信息技术有限公司）	采集用户身体数据和运动成绩,汇集健身教练资源,提供个性化健身指导	深圳	数据源;应用	2014.2.17	深圳
103	橙意家人科技（天津）有限公司	通过鼾症监测仪采集患者身体数据,结合医患互动的 APP 形成监测、治疗、服务的闭环产品	天津	数据源;应用	2014.3.6	天津
104	微糖 APP(上海格平信息科技有限公司）	针对糖尿病患者，，聚合医生资源并对接到患者。将来想做数据平台,包含患者和医生两端的数据	上海	数据源;应用	2012.4.26	上海
105	HUBS1 汇通天下(汇通百达网络科技（上海）有限公司）	为酒店提供在线预订、分销、管理和交易系统	上海	数据源	2008.8.19	上海
106	北京中长石基信息技术股份有限公司（阿里 28 亿入股）	国内五星级酒店信息管理系统市场占 90%份额,掌握海量酒店数据	海淀	数据源	1998.2.6	北京
107	去哪儿（北京趣拿信息技术有限公司）	机票、酒店信息的汇集和相关服务	海淀	数据源;应用	2006.3.17	海淀
108	好巧网（好巧科技有限公司）	团队来自腾讯和百度。抓取国外酒店各种 UGC	海淀	分析;应用	2013.11.13	北京
109	蚂蜂窝（北京蚂蜂窝网络科技有限公司）	类同好巧网,更侧重于旅游攻略	朝阳	分析;应用	2007.11.29	朝阳
110	票管家（上海时域电子商务有限公司）	为景区提供电子票务解决方案,掌握大量景区人流数据	上海	数据源	2012.5.16	上海

111	智游啦（香港远译国际有限公司）	港科大背景。数据：酒店景点等评论、用户标签。 利用:个性化的旅游产品推荐,降低用户面对海量产品的选择成本	香港	应用	香港公司	
112	面包旅行（北京道玺优讯科技有限公司，腾讯入股）	基于社交关系进行旅游推荐,拥有海量的图片数据且已结构化。核心技术是图片识别	朝阳	分析;应用	2012.3.14	朝阳
113	携程（携程旅游网络技术（上海）有限公司）	数据：客户及订单信息。应用:个性化推荐、为景区提供决策辅助	上海	应用	2005.4.14	上海
114	新影数讯网络科技（北京）有限公司	基于社交数据,做影业BI	海淀	分析;应用	2012、6、29	海淀
115	猫眼电影（美团网，北京三快科技有限公司）	线上购票选座业务,分析用户与票房的关联	朝阳	数据源;应用	2007.4.10	海淀
116	大众点评（上海汉涛信息咨询有限公司）	产生了大量用户的评价信息，以及各类POI的位置等信息	上海	数据源	2003.9.23	上海
117	北京捷成世纪科技股份有限公司	覆盖广电行业全产品线，视频识别处理	海淀	应用	2006.8.23	北京
118	卖座网（深圳市华宇讯科技有限公司）	属于深圳市华宇讯科技有限公司,华谊入股51%。	深圳	数据源	2004.12.16	深圳
119	艾漫科技	抓取全网娱乐相关信息，提供决策依据	朝阳	应用	2012.3.5	海淀
120	牧星人影视策划有限公司	从预算、档期、性别、外形、社交关系、口碑等角度为剧组推荐演员	朝阳	分析;应用	2011	
121	上海星红桉数据科技有限公司	拥有海量电视节目收视数据。收购了AC尼尔森中国团队	上海	数据源;应用	2013.12.30	上海
122	美乐网（爱美乐(北京)科技发展有限公司）	音乐个性化推荐	海淀	分析;应用	2012.2.20	海淀
123	虾米网（ 杭州缪斯客网络科技有限公司）	音乐个性化推荐	浙江	分析;应用	2006.12.1	浙江
124	浙江华策影视（以16亿元收购克顿传媒）	影视剧行业数据库,收录自97年以来近万部电视剧的收视数据、国内大量影视业人员信息。近期经典案例《何以笙箫默》	浙江	数据源	2005.10.25	浙江

125	搜前途（北京搜前途科技有限公司）	基于大数据的在线招聘创业公司,它通过大数据和特殊算法来实现简历和职位的精准匹配	朝阳	应用	2012.1.4	朝阳
126	哪上班（贝维优（北京）科技有限公司）	基于算法,做人才数据的分析和匹配。CTO 是 CMU 的人工智能博士	海淀	应用	2013.11.13	海淀
127	e 成招聘(上海逸橙信息科技有限公司）	总部在上海。用机器学习帮助 HR 进行筛选	上海	应用	2012.7.30	上海
128	望才招聘	特色:基于社交媒体内容对候选人进行画像	上海	应用		
129	内聘网（北京亿联宏谦科技有限公司）	基于文本分析，自动+人工方式实现简历和职位描述的格式化,然后进行匹配	海淀	应用	2013.12.30	海淀
130	人人猎头（上海众聘信息科技有限公司）	基于熟人推荐,用众包方式寻找人才	上海	应用	2012.10.15	上海
131	途志（北京途志优旅信息科技有限公司）	“中国版 ITA”。收集底层航班数据,提供多种国际航班的选择	朝阳	应用	2011.2.22	朝阳
132	航旅纵横（中国民航信息网络股份有限公司）	最主要的航空数据源,典型的数据拥有者介入应用领域的案例	东城	数据源;应用	2000.10.18	北京
133	飞常准（合肥飞友网络科技有限公司），携程入股，	航班实时查询。数据:向中航信、空管局、机场和航空公司购买	安徽	应用	2005.10.28	合肥
134	中国船期网（厦门鑫炬信息科技有限公司）	全球班轮数据提供商	福建	数据源;应用	2011.12.29	厦门
135	滴滴&快的（北京小桔科技有限公司）	用户画像、司机画像、位置匹配	海淀	分析;应用	2012.7.10	海淀
136	车来了（武汉元光科技有限公司）	实时公交查询。数据:通过在公交上装 GPS 获得	武汉	分析;应用	2010.2.4	武汉
137	重庆云途交通科技有限公司（招商局旗下）	提供各种车载设备和智能交通信息技术服务	重庆	数据源	2013.12.17	重庆
138	深圳市元征科技股份有限公司	车联网企业,与百度合作，推出 golo 盒子	深圳	数据源	1993.7.27	深圳
139	北京九五智驾信息技术股份有限公司	车联网服务商,也推出了 OBD 盒子	海淀	应用	2007.1.8	海淀
140	上海博泰悦臻电子设备制造有限公司	车联网设备制造商,推出 iVoka MiniX 盒子	上海	数据源	2009.10.20	上海

141	上海快逸行信息科技有限公司	车联网服务商,推出车逸行终端	上海	数据源;应用	2009.11.27	上海
142	autobot（北京微格互动科技有限公司）	车载智能设备制造商。通过OBD接口,获取里程、耗油、急刹车等行车数据	海淀	数据源	2011.5.26	海淀
143	聚合数据（苏州新科兰德科技有限公司）	通过 api 为开发者提供各类数据	江苏	数据源	2010.2.25	苏州
144	作业通（长沙拓欣菁优网络科技有限公司）	基础教育试题收集、组织和搜索	湖南	数据源;应用	2014.8.7	长沙
145	学大教育（学大教育科技（北京）有限公司）	课程及教材积累、用户数据分析、个性化教学。纽交所上市	朝阳	数据源;应用	2001.9.10	北京
146	知乎（北京智者天下科技有限公司）	大型问答平台,积累各行业领域海量知识	海淀	数据源	2011.6.8	海淀
147	学霸君（上海谦问万答吧云计算科技有限公司）	通过图像识别技术,识别试题并返回结果	上海	分析;应用	2013.12.31	上海
148	北京天创征腾信息科技	主要针对金融行业,票据账面识别	海淀	分析	2006.10.31	海淀
149	嗨图（成都夏陌科技有限公司）	图片众包标注平台	四川	数据源	2014.7.7	成都
150	汉王科技股份有限公司	人脸识别、文字识别	海淀	分析	1998.9.11	北京
151	face++（北京旷视科技有限公司）	融合机器视觉、机器学习、大数据挖掘技术,提供人脸识别服务	海淀	分析	2011.10.8	海淀
152	云视链（上海极链网络科技有限公司开发）	海量视频标签化和搜索	上海	数据源	2014.10.3	上海
153	格灵深瞳信息技术有限公司	计算机视觉产品研发。红杉数千万美元投资	海淀	分析	2013.8.16	北京
154	北京天诚盛业科技有限公司	研发图像识别和指纹识别等技术,提供数据安全解决方案	海淀	分析	2005.4.18	海淀
155	北京博思廷科技有限公司	视频分析,目前主要用于安防领域	海淀	分析	2007.9.26	海淀
156	北京吉祥海云数据科技有限公司（hydata）	主要提供数据可视化服务	西城	分析	2013.1	朝阳
157	广州图普网络科技有限公司	用户上传图片并打标签，后台自动构建模型、案例:迅雷用其来完成黄色图片的识别。收益:积累了大量图片数据和分析模型	广州	分析	2014.4.1	广州

158	随手房（北京悦商行知信息技术有限公司）	“房产经纪的evernote”，供房产中介随手记录客户及房屋信息	深圳	基础架构；应用	2011.9.28	北京
159	天津易邀在线科技有限公司（美丽屋APP）	数据:为房产中介开发ERP,积累了大量房源、经纪人和买家数据	天津	数据源；应用	2014.6.12	天津
160	无忧停车（北京紫光百会科技有限公司）	采集停车场信息,包括停车场名称、位置、车位数量、出入口 POI 信息、营业时间、收费标准和照片等。为百度地图提供数据	海淀	数据源；应用	2005.9.16	海淀
161	电话邦（北京羽乐创新科技有限公司）	电话号码数据服务商，获小米等千万美元投资	朝阳	数据源；应用	2012.1.16	昌平
162	杭州同盾科技有限公司	针对网络交易的欺诈识别	浙江	分析	2012.10.10	杭州
163	北京瀚思安信科技有限公司（HanSight ）	基于日志分析,提供企业安全解决方案。类似splunk	海淀	分析	2014.1.6	海淀
164	Talkingdata（北京腾云天下科技有限公司）	移动应用统计分析平台。北京腾云天下科技有限公司	东城	数据源；分析	2011.7.19	海淀
165	友盟（友盟同欣（北京）科技有限公司）	移动互联网用户分析，为开发者提供决策支持	海淀	数据源；分析	2011.10.21	海淀
166	厦门可睿特信息科技有限公司	脚型数据采集 --KRT-Foot in 3D 扫描仪。为电商和鞋类品牌提供服务	福建	数据源	2010.5.13	厦门
167	sequoiadb（广州巨杉软件开发有限公司）	分布式文档型 NoSQL 数据库,支持事务处理和 SQL。已获启明创投千万美元级的 A 轮	广州	基础架构	2012.10.11	广州
168	北京海博思创科技有限公司	开发智能电网系统,掌握大量用电数据	海淀	数据源	2011.11.4	海淀
169	国网信息通信有限公司	国家电网下属,电力数据源头	宣武	数据源	1994.9.28	北京
170	北京拓尔思信息技术股份有限公司	非结构化信息检索和分析	朝阳	分析	1993.2.18	北京
171	麦客（北京易多客信息技术有限公司）	mikecrm 产品，帮助企业做调查和联系人管理。获红杉 400 万美元	成都	应用	2013.7.4	石景山

172	北京银瀑技术有限公司	类似 RetailNext，为多媒体设备提供高效智能视频分析算法	朝阳	分析;应用	2009.11.26	朝阳
173	出门问问（北京羽扇智信息科技有限公司）	中文语音分析	海淀	分析	2014.3.6	海淀
174	墨迹天气(墨迹风云(北京）软件科技发展有限公司）	数据：可能来自气象局。不产生数据也没有分析,但是处于数据交付链条上的一环	朝阳	应用	2010.3.8	朝阳
175	彩云天气（北京彩彻区明科技有限公司）	数据:来自中国气象科学数据共享服务网和气象雷达。通过机器学习算法,对外提供的未来短时间内的降雨预报	海淀	分析;应用	2014.4.14	海淀
176	优酷土豆（合一信息技术（北京）有限公司）	视频应用和网络平台，分析类型:用户分析和视频数据结构化	海淀	数据源;应用	2006.2.24	海淀
177	北京爱奇艺科技有限公司	视频应用和网络平台，分析类型:用户分析和视频数据结构化	海淀	数据源;应用	2007.3.27	海淀
178	芝麻信用管理有限公司（阿里旗下）	结合阿里所掌握的数据，提供征信服务	浙江	应用	2015.1.8	浙江
179	金柚网（杭州今元标矩科技有限公司）	人力资源服务 SaaS 平台，针对中小企业，在社保管理方面口碑优良。如果发展顺利，将沉淀大量的中小企业和职工信息	杭州	数据源	2014.3.31	杭州
180	银联智惠信息服务（上海）有限公司	银联旗下子公司,掌握全国银联卡用户的刷卡记录	上海	数据源	2012.12.7	上海
181	华风气象传媒集团有限责任公司	国家气象局直属企业，掌握最权威的气象数据	海淀	数据源	2002.9.27	海淀
182	北京太谷雨田信息科技有限责任公司	承建农业部和各省"金农"工程、"三电合一"信息服务工程、12316综合信息服务工程。掌握大量农业产业链相关信息	东城	数据源	2011.4.1	通州
183	深圳市车音网科技有限公司	专注车载系统的中文语音识别技术的研发	深圳	分析;应用	2008.11.3	深圳

184	唐山市唐宋企业管理咨询有限公司（钢铁产业网）	拥有最准确和全面的钢铁行业信息	河北	数据源	2005.1.25	河北
185	豆瓣（北京豆瓣互动科技有限公司）	生产大量文娱类的评论数据和受众群信息，并进行个性化推荐	朝阳	数据源	2006.8.21	北京
186	成都数联铭品科技有限公司	引入各类非受控的外部数据，为金融、法律、商业和财务机构提供决策的数据支持服务	四川	应用	2013.7.30	四川
187	广州灵聚信息科技有限公司	中文人工智能交互引擎，主要偏向语音方面的交互	广州	分析	2013.6	广州
188	杭州摩图科技有限公司	前 Google 员工创办，专注于图像识别引擎的开发，2015 年 1 月完成 A 轮融资	浙江	分析	2013.7.12	浙江
189	南京智搜智能科技有限公司	专注于流媒体的自动化处理、识别和搜索（WUSHUU 智能视频分析系统）	江苏	分析	2013.5.30	江苏
190	上海优同科技有限公司	从事自然语言、语音和人脸动画等先进人机交互技术研发与应用	上海	分析	2009.5.13	上海
191	无锡天脉聚源传媒科技有限公司	全国最大的视频节目加工中心和数据库	江苏	数据源	2008.11.4	江苏
192	北京中科奥森科技有限公司	中科院自动化所背景。基于图像识别技术，实现人、车、物、事件的自动识别与检索	海淀	分析	2005.12.29	北京
193	苏州国云数据科技有限公司	旗下魔镜平台致力于各类数据的可视化展现	江苏	分析	2013.8.15	江苏
194	广东粤科软件工程有限公司	我国影院市场的主要系统供应商，掌握最为底层的票房数据，并为各类在线选座服务提供支持	广东	数据源	1997.10.6	广东
195	上海创冰信息科技有限公司	致力于足球及篮球赛事分析系统的研发及服务，拥有海量赛事数据	上海	数据源	2014.12.26	上海

196	北京汇通天下物联科技公司	采集大量的汽车发动机数据，主要为物流行业提供支持	海淀	数据源;应用	2011.2.25	北京
197	中国保险信息技术管理有限责任公司（由中国保险保障基金有限责任公司出资20亿元人民币成立）	采集保险行业的经营管理数据及相关外部数据，为保险公司、监管部门和消费者提供信息服务	石景山	数据源	2014.3.21	北京
198	北京至信普林科技有限公司	致力于自然语言处理、深度学习等大数据技术研发，引入运营商数据，为企业提供精准的客户画像和风险分析服务	海淀	分析	2014.5.27	北京
199	爱康国宾健康体检管理集团有限公司	从自身掌握的海量体检数据为入口，对客户健康状况加以解读和判断	朝阳	数据源		朝阳
200	慈铭健康体检管理集团股份有限公司	从自身掌握的海量体检数据为入口，对客户健康状况加以解读和判断	朝阳	数据源	2004.9.27	朝阳
201	珠海云洲智能科技有限公司	无人船制造商，用于水质监测、水文测绘、核辐射监测和水文研究等	广东	数据源	2010.4.15	广东
202	思昂教育（北京凌声芯语音科技有限公司）	专注于英语口语评测、语音识别等领域，将语音技术运用于教学、培训和考试等	海淀	分析；应用	2005.3.4	海淀
203	重庆中科雲從科技有限公司（广州云从信息科技有限公司和中国科学院重庆绿色智能技术研究院合资）	专注于人脸和车辆识别、警用图侦等领域。首席专家：黄煦涛	重庆	分析	2015.5.4	重庆
204	上海骏聿数码科技有限公司	图像识别及人体生物特征识别核心技术研究，包括：人脸识别、人体车体检测、行为识别分析、视频检索等	上海	分析	2010.11.22	上海
205	北京数字政通科技股份有限公司	街景影像采集	海淀	数据源	2001.11.6	北京

206	北京易道博识科技有限公司	发票、人脸、版面识别	海淀	分析	2013.3.27	海淀
207	深圳市赛为智能股份有限公司	将视频分析用于智慧交通、智慧建筑、智慧水利等行业	深圳	分析	1997.2.27	深圳
208	深圳市飞瑞斯科技有限公司	专注于人脸识别和智能视频分析等技术的开发	深圳	分析	2007.8.28	深圳
209	江苏清大维森科技有限责任公司	人脸智能识别和后台分析比对系统研发	江苏	分析	2011.7.28	江苏
210	广东铂亚信息技术有限公司（欧比特5.25亿元收购）	从事生物特征识别核心技术研究，正在建设人脸数据库	广州	数据源；分析	1999.8.28	广州
211	上海银晨智能识别科技有限公司	人脸识别技术研发，用于公安、金融、司法、民航等领域，曾参与世博会安保	上海	分析	2001.12.29	上海
212	北京千搜科技有限公司	专注于人脸检测、识别、分析和重建等技术领域	海淀	分析	2013.8.21	海淀
213	亮风台（上海）信息科技有限公司	专注于智能图像识别与视觉交互技术，与三星、乐视、美图秀秀等有合作	上海	分析	2012.11.21	上海
214	北京致生联发信息技术股份有限公司	图像数据整合与分析，承接多个平安城市项目	朝阳	分析	1997.3.24	朝阳
215	江苏视图网络科技有限公司	专注于图像识别和图像相似搜索技术	江苏	分析	2011.12.27	江苏
216	普强信息技术（北京）有限公司	智能语音识别和自然语言处理技术，提供以中文为主的智能语音产品	海淀	分析	2010.9.27	海淀
217	同方知网(北京)技术有限公司	拥有海量文献资源，将自然语言处理运用于文本挖掘和信息检索	海淀	数据源；分析	2004.11.18	北京
218	上海玻森数据科技有限公司	专注于中文语义分析，对外提供语义分析API，涵盖情感计算、实体、分类、聚类等技术领域	上海	分析	2012.4.20	上海

219	万达信息股份有限公司	其医疗健康服务平台覆盖了全国 3.6 亿人口，社会保障系统覆盖 1.3 亿人口	上海	数据源	1995.11.9	上海
220	中科九度（北京）空间信息技术有限责任公司	中科院电子所背景，遥感图像处理和空间信息分析	海淀	分析	2010.10.26	海淀
221	艺恩世纪国际信息咨询（北京）有限公司	整合多屏终端消费行为数据，为影视业提供决策支持	朝阳	数据源；应用	2009.3.9	朝阳
222	上海语天信息技术有限公司	专注于多语种自然语言处理技术，曾参与上海世博会	上海	分析	2008.9.12	上海

国内大型企业大数据相关产品或服务

1.	百度地图	地图服务提供	海淀	应用
2.	搜狗地图	地图服务提供	海淀	应用
3.	腾讯地图/街景	地图服务提供	海淀	应用
4.	新浪地图	地图服务提供	海淀	应用
5.	腾讯云	云存储服务	深圳	基础架构
6.	金山云	云存储服务	北京	基础架构
7.	百度云	云存储服务	海淀	基础架构
8.	阿里金融	基于电商业务积累的海量数据，通过线上提供各类金融服务	浙江	应用
9.	万达地产	通过室内定位、人流分析、用户画像等，达成商场整体布局的优化	朝阳	应用
10.	阿里巴巴（数据魔方）	为企业决策提供数据支持	浙江	数据源
11.	京东	为企业决策提供数据支持	朝阳	数据源；应用
12.	百度数据开放平台	类同数据魔方	海淀	数据源
13.	链家	基于自身的房源和客户数据，打造包括生活支付、社区服务、智能家居等方面的平台	朝阳	数据源；应用
14.	爱问知识人（新浪）	知识平台，类同于百度知道	海淀	数据源
15.	百度知道&作业帮	基于海量知识积累，推出了作业帮，直接介入分析和应用环节。	海淀	数据源;分析;应用
16.	网易公开课	积累大量国内外名校课程资料	广州	数据源
17.	腾讯路宝	车载设备，通过 OBD 接口读取数据	深圳	数据源
18.	网易云音乐	网络音乐平台，基于用户分析做个性化推荐	广州	分析；应用
19.	QQ 音乐	网络音乐平台，基于用户分析做个性化推荐	深圳	分析；应用
20.	百度视频	视频数据搜索	海淀	数据源；分析
21.	腾讯视频	视频应用和网络平台，分析类型：用户分析和视频数据结构化	深圳	数据源；应用
22.	搜狐视频	视频应用和网络平台，分析类型：用户分析和视频数据结构化	海淀	数据源；应用

23.	百度问诊	根据用户的病症搜索关键词,筛选和推送相关的个性化信息	海淀	应用
24.	平安好医生(平安健康管家)	涉及从健康数据采集到跟踪、诊疗、管理方案及奖励、评估保险、付费和药品等环节	深圳	应用
25.	百度语音	语音服务	海淀	分析；应用
26.	腾讯征信有限公司	征信服务	深圳	应用
27.	小米智能家居业务	构建物联网和智能设备基础平台	海淀	基础架构
28.	京东金融	基于自身掌握的用户和企业信息，全面介入各类金融服务，相当于阿里金融的counterpart	朝阳	应用
29.	京东智能云	选择智能家居和健康两个领域，向多家合作伙伴提供软硬件两个层面上的技术支持	朝阳	基础架构
30.	腾讯广点通	基于腾讯大社交网络体系的数据,实现更加智能的广告匹配和高效的广告资源利用	海淀	广告营销
31.	阿里妈妈达摩盘	大数据管理平台 DMP，通过数据流通构建了一个以 DMP 为核心的数据生态圈,构建全网营销能力	浙江	广告营销
32.	百度 DMP 数据服务	基于自身数据优化跨渠道的广告投放方案，帮助广告主管理用户数据	海淀	广告营销