百度智能云 数据众包

iResearch 艾瑞咨询

中国人工智能基础数据服务行业白皮书

2019年

摘要

在经历了一段时期的野蛮生长之后，人工智能基础数据服务行业进入成长期，行业格局逐渐清晰。人工智能基础数据服务方的上游是数据生产和外包提供者，下游是AI算法研发单位，人工智能基础数据服务方通过数据处理能力和项目管理能力为其提供整体的数据资源服务，不过AI算法研发单位和AI中台也可提供一些数据处理工具，产业上下游普遍存在交叉。

2018年中国人工智能基础数据服务市场规模为25.86亿元，其中数据资源定制服务占比86%，预计2025年市场规模将突破113亿元。市场供给方主要由人工智能基础数据服务供应商和算法研发单位自建或直接获取外包标注团队的形式组成，其中供应商是行业主要支撑力量。

数据安全、采标能力、数据质量、管理能力、服务能力等仍是需求方的痛点，需要人工智能基础服务商有明确具体的安全管理流程、能够深入理解算法标注需求、可提供精力集中且高质量的服务、能够积极配合、快速响应需求方的要求。

随着算法需求越来越旺盛，依赖人工标注不能满足市场需求，因此增强数据处理平台持续学习能力，由机器持续学习人工标注，提升预标注和自动标注能力对人工的替代率将成趋势。远期，越来越多的长尾、小概率事件所产生的数据需求增强，机器模拟或机器生成数据会是解决这一问题的良好途径，及早研发相应技术也将成为AI基础数据服务商未来的护城河。

来源：艾瑞自主研究绘制。

序言

算法、算力、数据是人工智能发展的三大要素，人工智能已经从讲技术教育市场的阶段，过渡到思考如何将技术与商业相结合落地的阶段，而数据作为AI算法的“燃料”，是实现这一能力的必要条件，因此，为机器学习算法训练提供数据采集、标注等服务的人工智能基础数据服务成为近年人工智能热潮中必不可少的一环。

2018年1月，由国务院办公厅发布的《科学数据管理办法》中，明确了科学数据的责任、安全使用和共享利用等行为规范，政策层面的关注表明，科学数据是国家科技创新发展和经济社会发展的重要基础性战略资源，科技创新越来越依赖于大量、系统、高可信度的科学数据。

当人工智能技术在更多场景尝试下沉时，AI基础数据服务将迎来挑战，新兴垂直场景数据既难以获取，又需要有经验、有专业素养的人员进行标注，考验着从业玩家的研发、管理、培训能力，但也同样伴随着机遇。

人工智能基础数据服务并非人们想象中的数据作坊，其发展依赖于基于技术的数据处理平台和工具，以及科学高效的管理。该赛道还是科技巨头早早布局的“逐鹿场”。以百度为代表的巨头企业纷纷建设数据采集与标注服务团队，在支撑自身人工智能技术研发的同时，对外输出数据采标能力，成为行业领先力量。随着高难度、前沿独特性需求渐成主流，数据服务行业早期鱼龙混杂的现象将改变，优势公司实力将逐渐凸显。

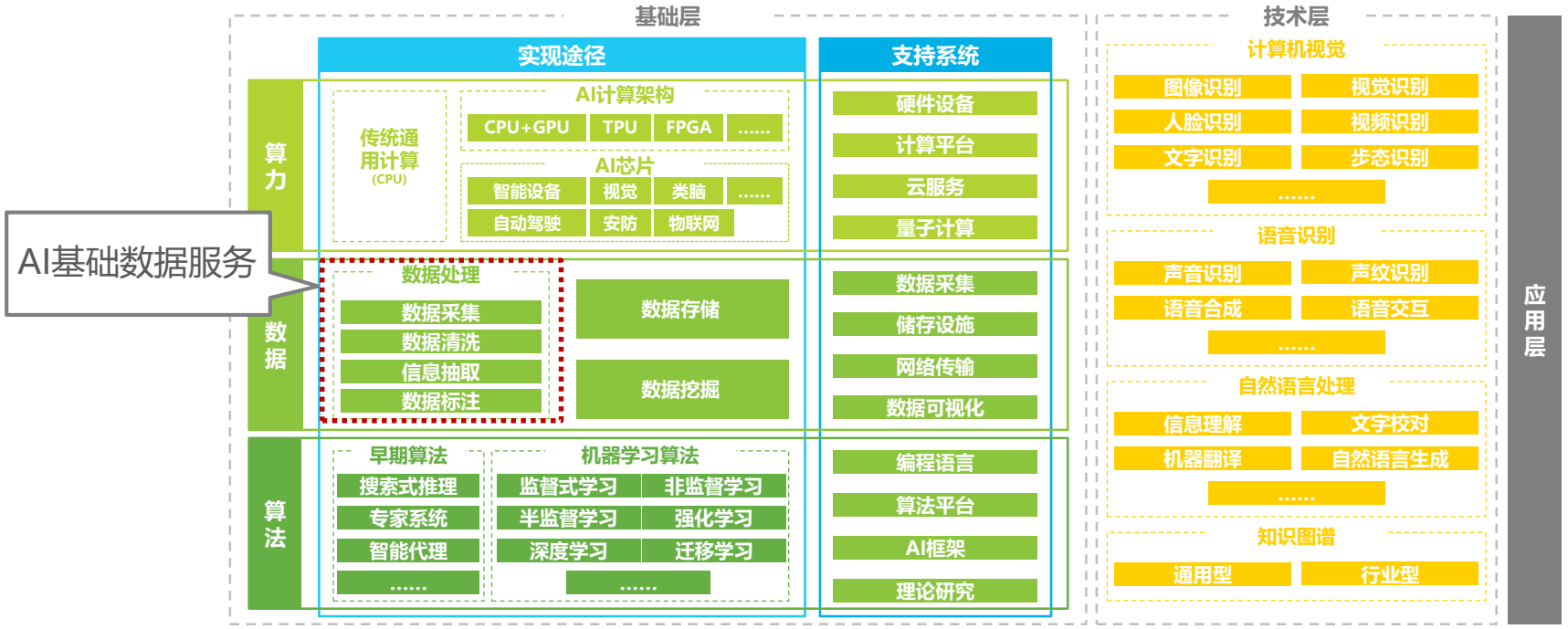
——艾瑞咨询研究院

人工智能基础数据服务定义

意指为AI算法训练及优化提供数据采集和标注等形式的服务

人工智能基础数据服务指为AI算法训练及优化提供的数据采集、清洗、信息抽取、标注等服务，以采集和标注为主。人工智能概念爆发伊始，算法、算力、数据就作为最重要的三要素被人们乐道，进入落地阶段，智能交互、人脸识别、无人驾驶等应用成为了最大的热门，AI公司开始比拼技术与产业的结合能力，而数据作为AI算法的“燃料”，是实现这一能力的必要条件，因此，为机器学习算法训练、优化提供数据采集、标注等服务的人工智能基础数据服务成为了这一人工智能热潮中必不可少的一环。如果说计算机工程师是AI的老师，那基础数据服务就是老师手中的教材。

人工智能技术框架



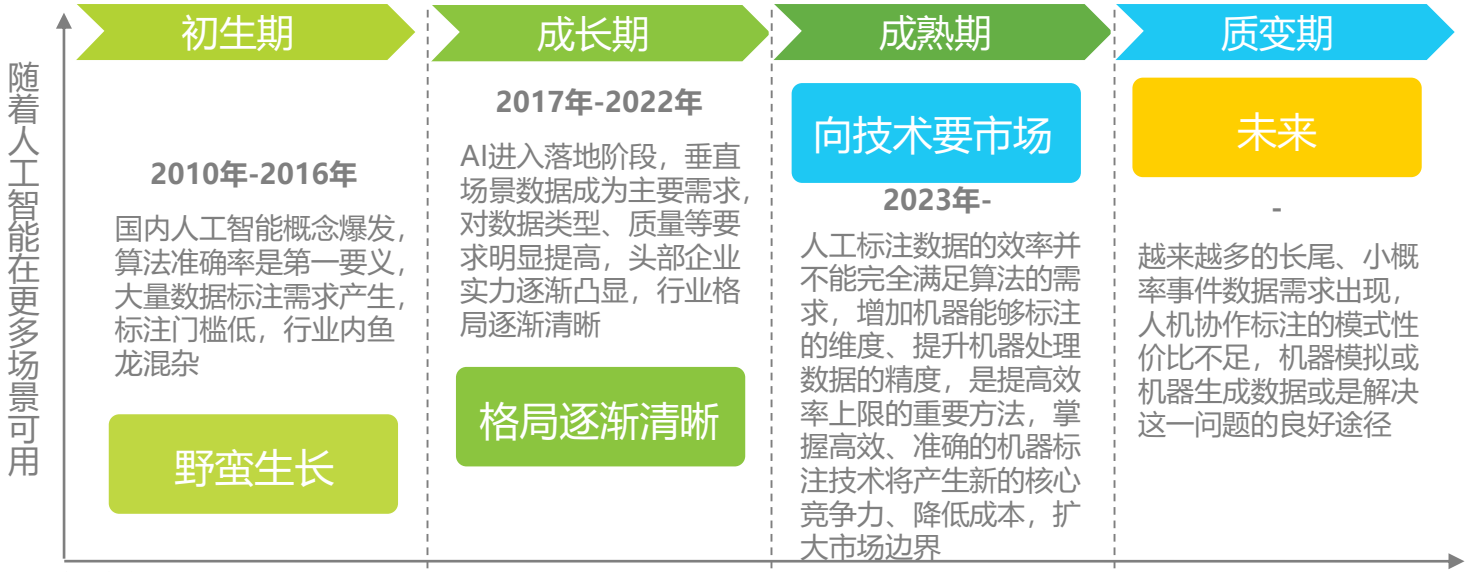
来源：艾瑞根据公开资料自主研究绘制。

人工智能基础数据服务发展历程

行业进入成长期，行业格局逐渐清晰

伴随国内人工智能热潮爆发，大量的AI公司拿到融资，为了不断提高算法精度，数据采标需求也空前爆发，一度催生了行业的繁荣。但早期的AI基础数据服务门槛较低，玩家鱼龙混杂，使行业标准模糊、服务质量参差不齐。随着竞争加快，AI公司对训练数据的质量要求也不断提高，并且当产业落地成为主旋律，需求方对垂直场景的定制化数据采标需求成为主流，众多小型AI基础数据服务公司从数据质量和采标能力上达不到要求，或被淘汰，或依附大平台，行业格局逐渐清晰，头部公司实力逐渐凸显。随着算法需求越来越旺盛，目前机器辅助标注、人工主要标注的手段需要改进提升，增强数据处理平台持续学习和自学习能力，增加机器能够标注维度、提升机器处理数据的精度，由机器承担主要标注工作将成为下一阶段的行业重心。未来，越来越多的长尾、小概率事件所产生的数据需求增强，人机协作标注的模式性价比不足，机器模拟或机器生成数据会是解决这一问题的良好途径，及早研发相应技术也将成为AI基础数据服务商未来的护城河。

AI基础数据服务行业发展历程及展望



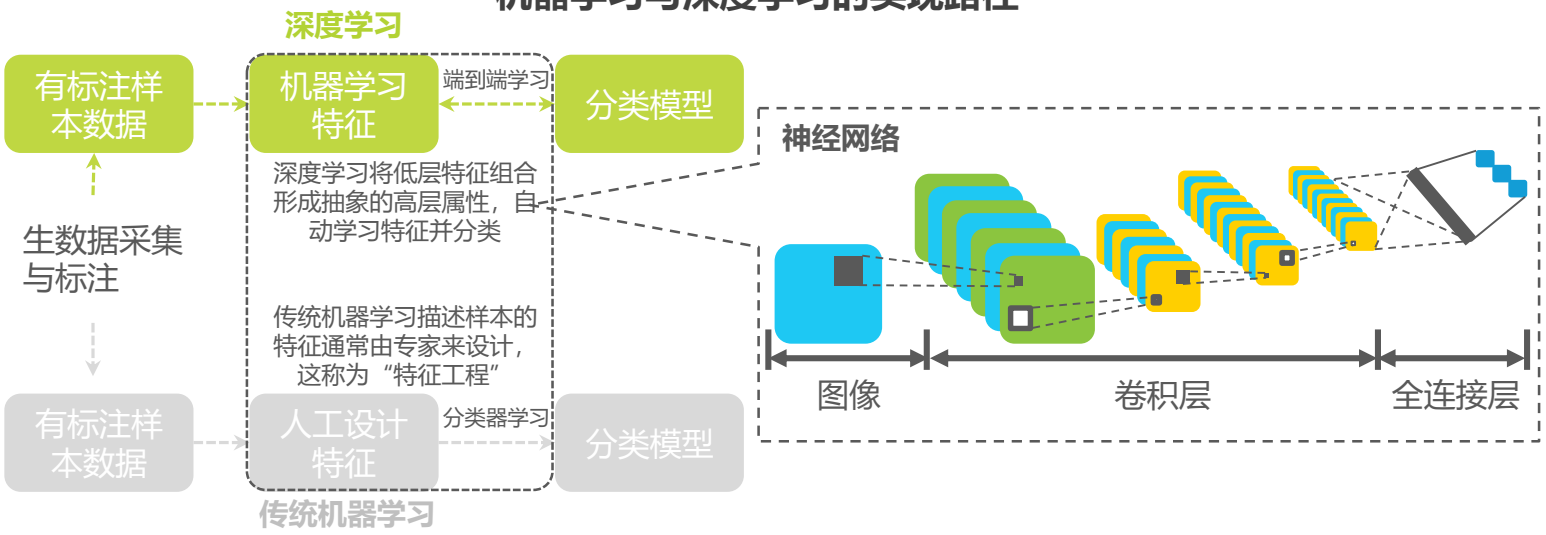
来源：艾瑞根据公开资料自主研究绘制。

人工智能基础数据服务的行业价值

目前有监督的深度学习是主流，标注数据是其学习根本

人工智能是研究如何通过机器来模拟人类认知能力的科学，机器学习是现阶段实现人工智能的主要手段。机器学习方法通常是从已知数据中学习规律或者判断规则，建立预测模型，其中，深度学习可以通过对低层特征的组合，形成更加抽象的高层属性类别，自动从信息中学习有效的特征并进行分类，而无需人为选取特征。凭借自动提取特征、神经网络结构、端到端学习等优势，深度学习在图像和语音领域学习效果最佳，是当今最热门的算法架构。在实际应用中，深度学习算法多采用有监督学习模式，即需要标注数据对学习结果进行反馈，在大量数据训练下，算法错误率能大大降低。现在的人脸识别、自动驾驶、语音交互等应用都采用这类方法训练，对于各类标注数据有着海量需求，可以说数据资源决定了当今人工智能的高度。由于应用有监督学习的AI算法对于标注数据的需求远大于现有的标注效率和投入预算，无监督或仅需要少量标注数据的弱监督学习、小样本学习成为了科学家探索的方向，但目前无论从学习效果和使用边界来看，均不能有效替代有监督学习，人工智能基础数据服务将持续释放其对于人工智能的基础支撑价值。

机器学习与深度学习的实现路径



来源：艾瑞根据公开资料自主研究绘制。

人工智能基础数据服务的主要产品形式

定制服务为主要服务形式，数据集产品集中于语音类赛道

目前，国内AI基础数据服务主要为数据集产品和数据资源定制服务，数据集产品往往是AI基础数据服务商根据自身积累产出的标准数据集，以语音数据集为主，主体偏普通话语音、英文语音、方言语音等；为保证算法优势，客户更多采用定制化服务，由客户提出具体需求，数据服务商或直接对客户提供的数据进行标注、或对数据进行采集并标注。大型的需求方，为保障数据的安全，往往提供Web形式的自有标注平台给执行方，以此对整体项目进行把控，也有一些AI基础数据服务商向客户提供私有化平台建设服务，或将自身平台与甲方系统兼容；除以上两种形式外，部分AI基础数据服务商还向算法服务进行拓展，提供算法训练、模型搭建等服务。

AI基础数据服务行业主要产品形式

形式一

数据集产品

分为开源数据集和收费的数据集产品，主体主要分为语音类数据集、图像类数据集、NLP类数据集等，其中以语音类数据集为主

更适合算法处于早期研究阶段的客户使用

形式二

数据资源定制服务

定制服务是AI基础数据服务行业最为主要的服务形式，涵盖采集和标注服务，数据内容以语音、图像、NLP、OCR等为主，根据需求方的具体要求设计方案，并执行

适合算法训练、优化等需求，对于业务类需求有较强的支撑效果

形式三

其他数据资源应用服务

部分AI基础数据服务商还向算法服务方向进行拓展，提供算法训练、模型搭建等服务

倾向于AI中台概念中的部分能力

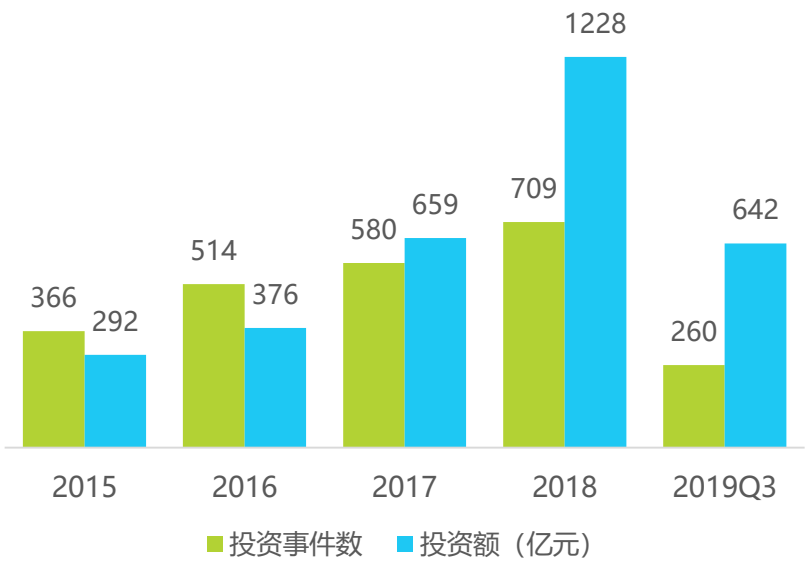
来源：艾瑞根据公开资料自主研究绘制。

人工智能基础数据服务的发展背景

人工智能经济崛起为基础数据服务提供长期向好的基本面

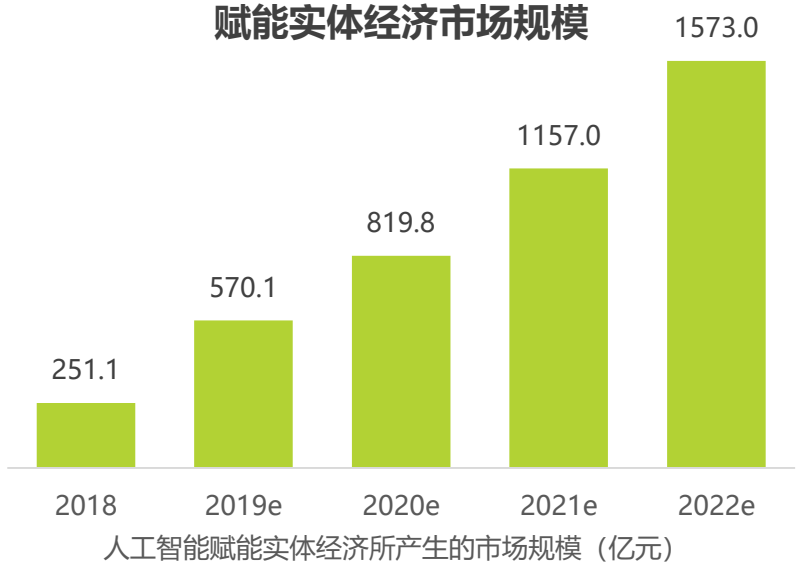
2010年语音识别和计算机视觉领域产生重大突破，国内开始萌生AI概念。到2015年，国内迎来人工智能创业热潮，独角兽不断涌现，融资记录被不断打破。2012年-2019年8月人工智能领域共发生2787件投融资事件，总融资额达4740亿元，人工智能成为最炙手可热的融资热点，百度、阿里、腾讯、京东、华为等科技企业也纷纷加注。2017年至今，产业落地成为AI行业的主流，人工智能赋能实体经济保持高速发展态势，涉及行业包括安防、金融、零售、交通、教育、医疗、营销、工业、农业、企服等众多领域。下游的爆发式增长为人工智能基础数据服务的发展提供了长期向好的基本面。

2015-2019年Q3中国AI领域投融资情况



来源：艾瑞《2019年中国人工智能产业研究报告》。

2018-2022e中国人工智能赋能实体经济市场规模



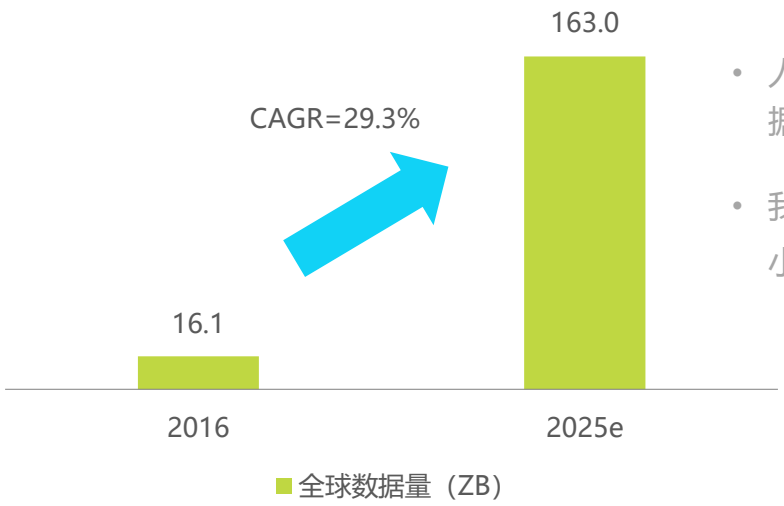
来源：艾瑞《2019年中国人工智能产业研究报告》。

人工智能基础数据服务的发展背景

数据量呈指数式增长，非结构化数据的应用依赖于清洗标注

PC、互联网、消费级移动设备的兴起宣告了数据时代的来临，物联网的发展更使线下业务产生的大量数据被采集起来，数据量呈指数式增长，据IDC统计，全球每年生产的数据量将从2016年的16.1ZB猛增至2025年的163ZB，其中80%-90%是非结构化数据。过去计算机主要处理结构化数据，人工智能模型却以处理非结构化数据见长，但 “玉不琢不成器”，数据经过清洗与标注才能被唤醒价值，这就产生了源源不断的清洗与标注需求。在我国，每年需要进行标注的语音数据超过200万小时，图片则有数亿张。

2016-2025年全球数据量的爆发式增长



- 人工智能模型以处理非结构化数据见长，但数据经过清洗与标注才能被唤醒价值
- 我国每年需要进行标注的语音数据超过**200万**小时，图片则达**数亿**张

来源：柱状图数据来自IDC，文字来自艾瑞自主研究。

人工智能基础数据服务的发展背景

人工智能政策向好，AI基础数据服务公司与政府积极配合

人工智能是中国大力发展的新一代信息技术中重要的组成部分，相关促进、指导意见连续三年出现在总理报告中，2019年“智能+”的概念又首次被写入到政府工作报告中，其发展意义已上升至国家竞争力层面。作为人工智能产业链中必不可少的一环，发展AI基础数据服务成为了各地方推进AI建设的重要方向之一，贵州、山西、重庆等地相继出台指导意见，引入科技公司，共建数据基地、数据交易中心，打造具有地方特色的人工智能产业园。以百度（山西）人工智能基础数据产业项目为例，是在山西省综合改革示范区支持下，由百度智能云数据众包团队筹建专业化、集中管理的AI数据标注基地。目前，基地拥有近1万平方米的办公场地，专业标注员和审核员达1500人，基地业务全方位覆盖了无人车、语音、人脸、图像、NLP、地图测绘等数据类型的标注和加工处理服务，是山西2019年重点推进项目。

AI基础数据服务基地代表案例

案例：贵州惠水百鸟河数字小镇

- 为惠水产业转型，建设新兴工业化路线而建设的新型园区
- 总规划面积18平方公里，起步区百鸟河核心区域5平方公里
- 园区自营超1500+席位的数据工场，提供了数以万计的数据标注和采集服务

贵州大数据产业发展较早，已形成一定的区域优势。2018年，省内软件和信息技术服务收入环比增长达到18%以上、电子信息制造增加值增长10%左右。2019年贵州打造10个省级、60个市州级试点项目，积极拓展新一代信息技术能力，实现产业转型与升级

成都一直是我国大数据产业发展较强的地区，拥有大数据相关企业400余家，涉及数据采集、数据存储、数据可视化、大数据应用等大数据全产业链。2018年中国大数据企业50强中，有超过70%入榜企业在成都设有分支机构或有关联企业，当地政府高度重视数据产业发展，未来将持续保持优势

案例：《成都市促进大数据发展工作方案》

- 2020年，重点培育3至5个大数据产业集聚区，推进政府数据开放数据集1000个以上
- 大数据从业人员规模达到6万人以上；大数据核心产业产值突破800亿元

案例：百度（山西）人工智能基础数据产业项目

- ✓ 近1万平方米的办公产地
- ✓ 1500名专业标注员和审核员
- ✓ 基地业务涵盖了无人车、语音、人脸、图像、NLP、地图测绘等数据类型的标注和加工处理服务

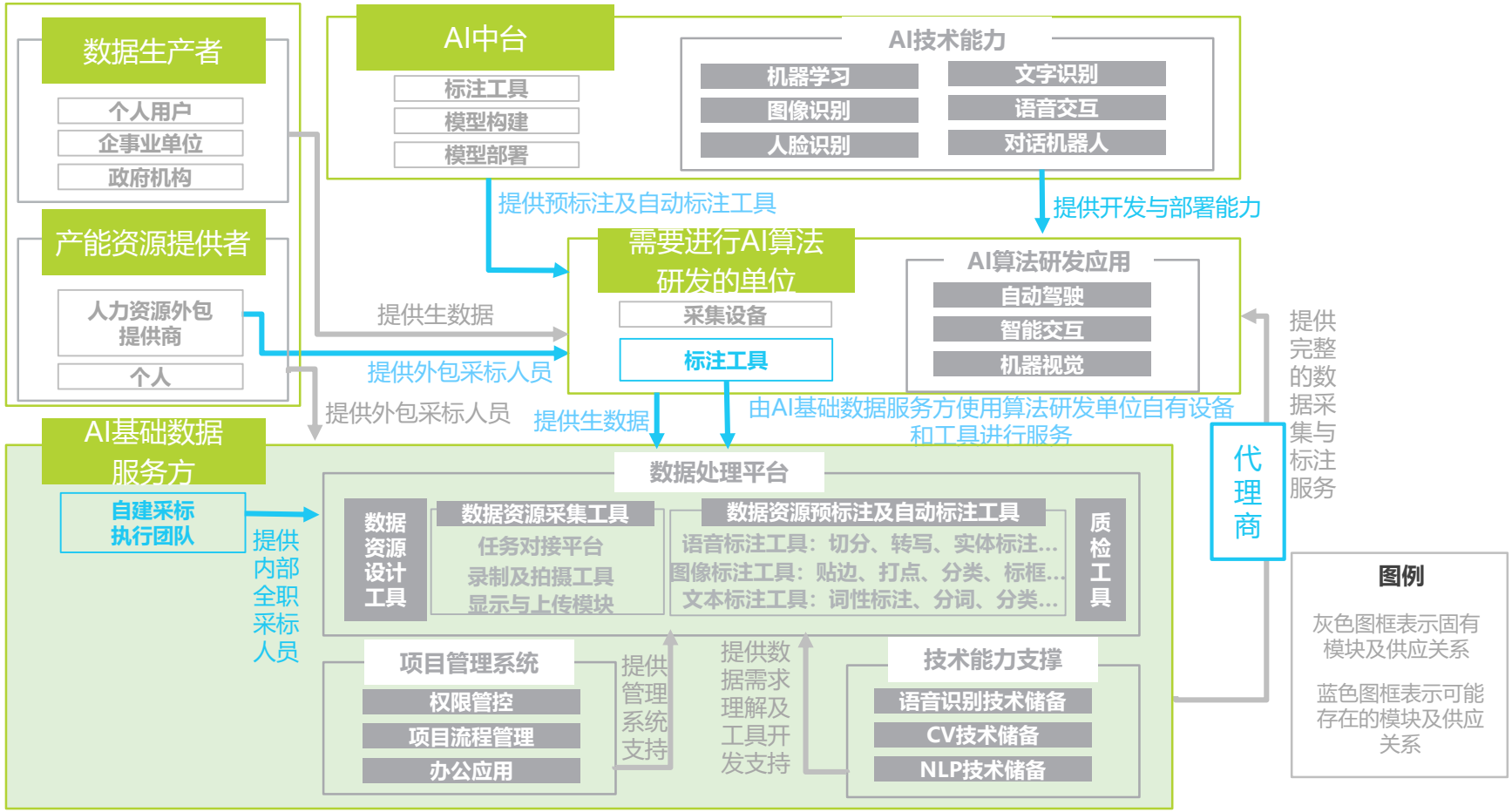
山西正处于由传统产业向科技型产业发展的转型期，数据标注行业是其重要的抓手。目前省内本土科技型公司和人才储备尚显不足，与巨头企业紧密合作带动整体发展，成为了切实可行的策略。山西省计划到2022年初步形成集数据采集、清洗、标注、交易、应用为一体的基础数据服务产业体系

来源：艾瑞根据公开资料自主研究绘制。

人工智能基础数据服务产业链

AI基础数据服务方是行业核心环节

2018年人工智能基础数据服务产业链



来源：根据公开资料和专家访谈，艾瑞自主研究绘制。

人工智能基础数据服务产业图谱

产业上下游普遍存在交叉

AI基础数据服务方的上游是数据生产和外包提供者，下游是AI算法研发单位，AI基础数据服务方通过数据处理能力和项目管理能力为其提供整体的数据资源服务。 AI基础数据服务方整体有两大类，一种是具备自有的标注基地或全职标注团队，这类企业也参与产业上游部分直接提供产能资源，另一种是依靠众包或外包模式，专注于数据产品的开发与项目执行。下游部分AI公司拥有自己的标注工具，也可通过AI中台获取一些通用标注工具，同时一些数据需求大的企业还孵化了自己的数据服务团队。整体而言，产业上下游普遍存在交叉关系。

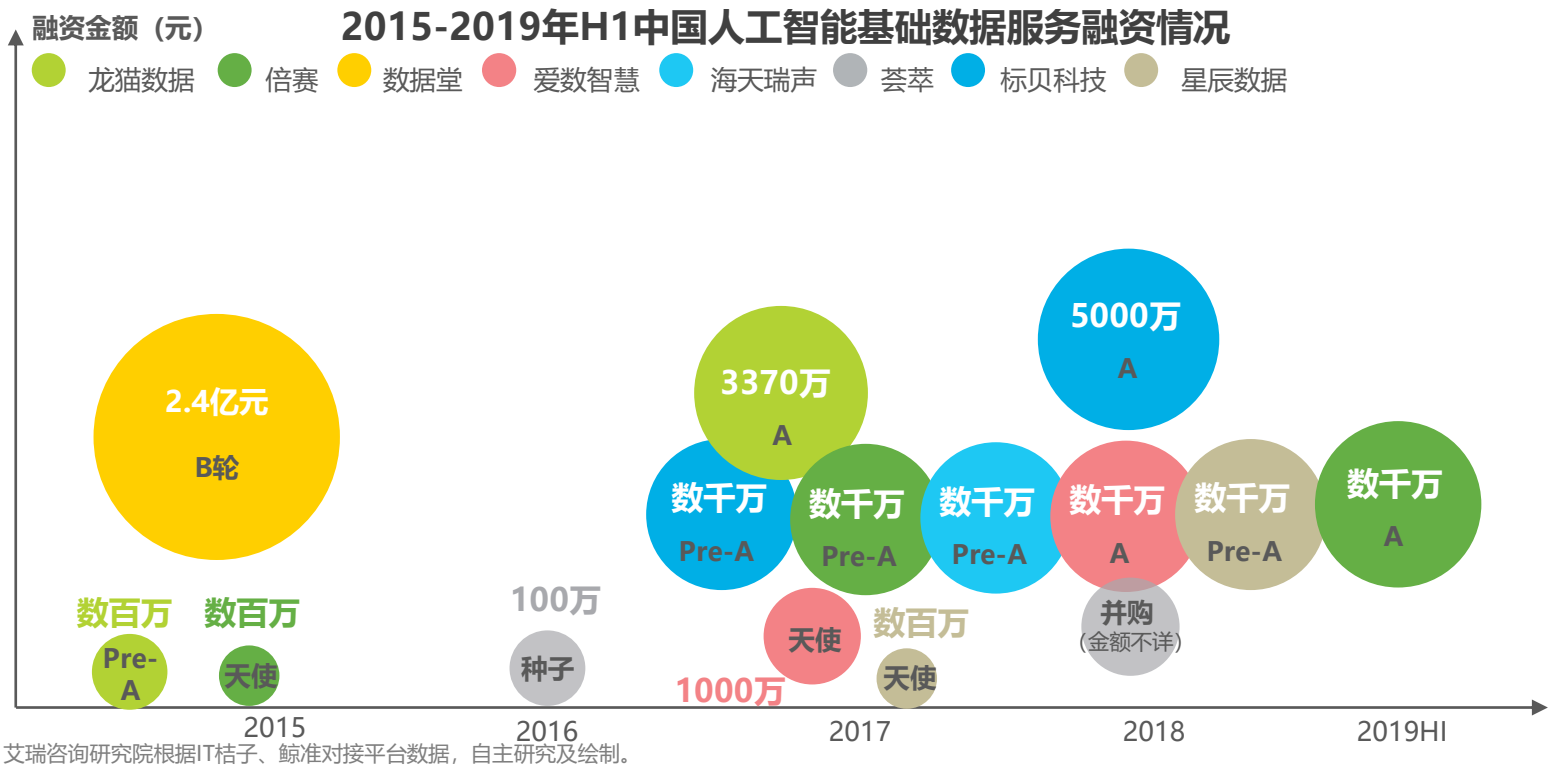
2018年人工智能基础数据服务产业图谱

来源：根据公开资料和专家访谈，艾瑞自主研究绘制。

人工智能基础数据服务行业投融资

融资规模集中于千万量级，早期融资项目居多

从融资规模来看，人工智能基础数据服务市场的融资多集中在千万级别。从时间维度来看，2015年人工智能基础数据服务商获得的融资金额相对较高，标志着行业初露头角，受到资本的认可。从获得融资的企业数量来看，目前获得融资的玩家并不多，资本市场表现的活跃度不高。从融资轮次来看，大部分融资仍然集中于早期融资，目前上市的企业仅挂牌新三板的数据堂一家（不考虑科技公司内部孵化的基础数据服务商）。人工智能基础数据服务毛利率普遍较高，但为保持与人工智能市场前沿算法的匹配，需要投入大量研发成本进行数据处理平台与工具的研发升级，因此对融资仍有较强依赖。



来源：艾瑞咨询研究院根据IT桔子、鲸准对接平台数据，自主研究及绘制。

人工智能基础数据服务行业商业模式

生产、获客、部署合力驱动发展

人工智能基础数据服务行业是典型的To B型业务，商业模式较为稳定。在生产方面，主要通过自建标注基地或标注团队、搭建众包平台、采购供应商外包服务（BPO）等模式实现生产运营，大多企业主要采取众包与外包模式，百度数据众包、倍赛等企业自建标注基地或全职标注团队，对于培训较高素质工作人员、完善团队管理有积极意义；在获客方面，主要通过口碑传播、学术会议与展会及代理渠道等模式进入市场，对销售人员熟悉市场趋势、客户需求的要求较高；在实施交付方面，有私有化部署和公有部署两类，能够较为灵活地应对客户对数据安全、交付周期与成本的个性化需求。

2018年中国人工智能基础数据服务商业模式

生产模式

自建标注基地或标注团队
拥有专业标注人员，通过完善的管理制度和培训，提升产能质量与效率

搭建众包平台
利用大众力量及资源，低成本、高效率地采集和制作专业数据

采购供应商外包服务（BPO）
增强生产能力弹性，由供应商承担生数据采集和标注等基础操作，优化企业运营

获客模式

口碑
通过提供优质服务，进入客户的供应商名录，是一种非标准化的获客模式

学术会议、展会
通过专业性学术会议与行业展会，取得客户关注

代理模式
通过代理合作拓展下游客户

实施模式

私有化部署
在数据生产者愈加重视数据隐私与安全的背景下，基础数据服务可以实现私有化离线部署，驻场标注，数据存储在客户本地

公有部署
数据接入在公有云服务器，可降低项目实施成本，通过数据接口加密、定期巡查、反爬虫机制保证数据安全

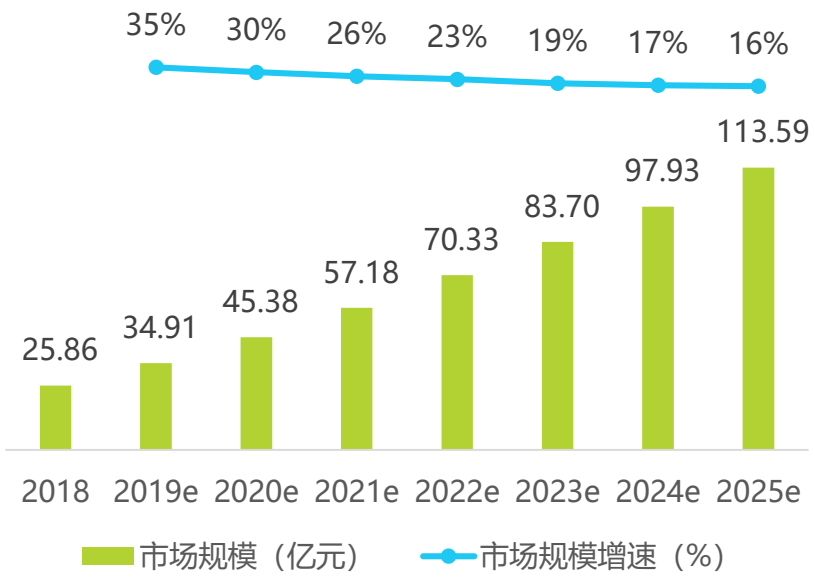
来源：根据公开资料，艾瑞自主研究绘制。

人工智能基础数据服务市场规模

2025年市场规模将突破百亿，行业年复合增长率为23.5%

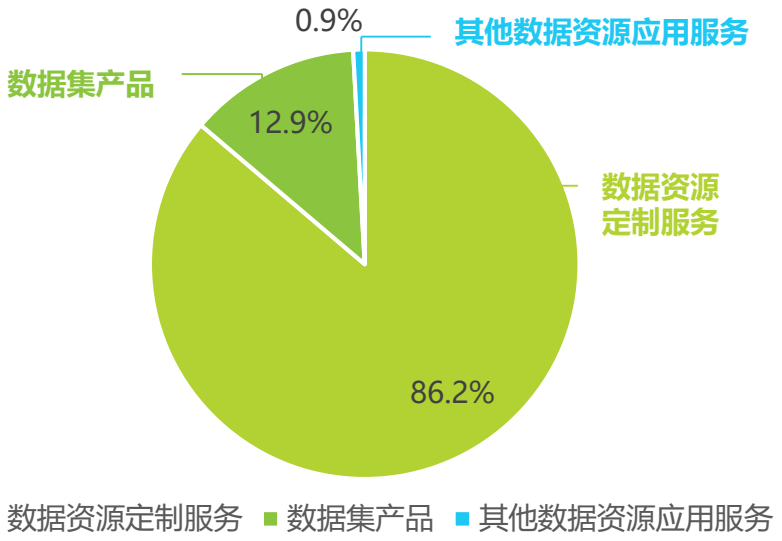
2018年中国人工智能基础数据服务市场规模为25.86亿元，其中数据资源定制服务占比86.2%，数据集产品占比12.9%，其他数据资源应用服务占比0.9%；行业年复合增长率为23.5%，预计2025年市场规模将突破110亿元。从整体增速来看，行业发展较为稳健，下游人工智能行业持续发力将形成长期利好。

2018年-2025e中国人工智能基础数据服务市场规模



来源：根据专家访谈与模型推算，艾瑞自主研究绘制。

2018年中国人工智能基础数据服务市场细分结构



来源：根据专家访谈与模型推算，艾瑞自主研究绘制。

人工智能基础数据服务细分结构

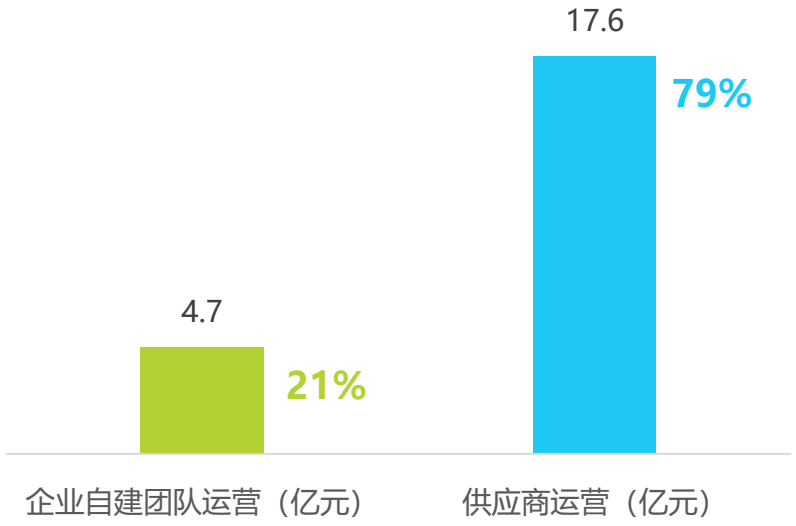
纯标注服务为主体，由供应商提供服务占79%

2018年中国人工智能基础数据服务市场以语音、视觉、NLP领域的标注服务为主，同时提供采集与标注服务占比较少，这是由于生数据由需求方提供的情况较多，但这并不意味着市场中数据采集需求弱，相反，人工智能技术落地后产生了大量新兴垂直领域的数据需求，然而这些数据采集难度大，能够提供相关采集工具和服务的供应商将获取竞争优势。市场供给方主要由企业自建或直接获取外包团队的形式以及供应商组成，又以供应商为行业主要支撑力量，占比79%。

2018年人工智能数据资源定制服务市场细分结构

来源：根据专家访谈与模型推算，艾瑞自主研究绘制。

2018年人工智能数据资源定制服务市场细分结构



注释：企业自建团队运营数据统计指企业在内部形成独立的团队/品牌或直接通过人力外包机构获取团队来负责数据采集与标注，不含由企业内部各岗位人员兼职地、分散地、非标地承担标注工作发生的成本，也不含内部孵化标注团队对外提供服务的收入。
来源：根据行业专家及需求方访谈与模型推算，艾瑞自主研究绘制。

人工智能基础数据服务市场格局

自建标注团队增加，但未对行业产生挤出效应

出于对数据安全性、成本和整体布局的考虑，人工智能和科技型互联网领域的头部公司开始组建自有标注团队，大部分采用聘用项目经理，执行团队外包的形式运营，所处理项目从少量较为简单基础的标注需求，逐渐向大量复杂任务发展，但由于数据需求总量大，未对市场产生明显挤出效应。其中AI公司的数据标注团队主要承担自身研发需求，而科技型互联网巨头组建的标注团队开始依靠集团优势，对外输出AI基础数据服务能力，形成了行业中较强的一方阵营。

AI基础数据服务自建团队

自建标注团队

百度、阿里、腾讯、京东等科技公司和科大讯飞、商汤科技等AI公司均开始自建标注团队

任务量级和复杂性提升

自建团队早期更多承担公司内部的算法研发和业务需求，往往标注内容较为简单，但随着团队经验的累计，任务量和复杂性明显推升

代表性公司团队对外输出能力

以百度数据众包为代表的数据标注团队成立较早，拥有大量活跃用户的众包平台，标注能力在集团中得到充足的锻炼，对外输出能力时也形成了较强的竞争力

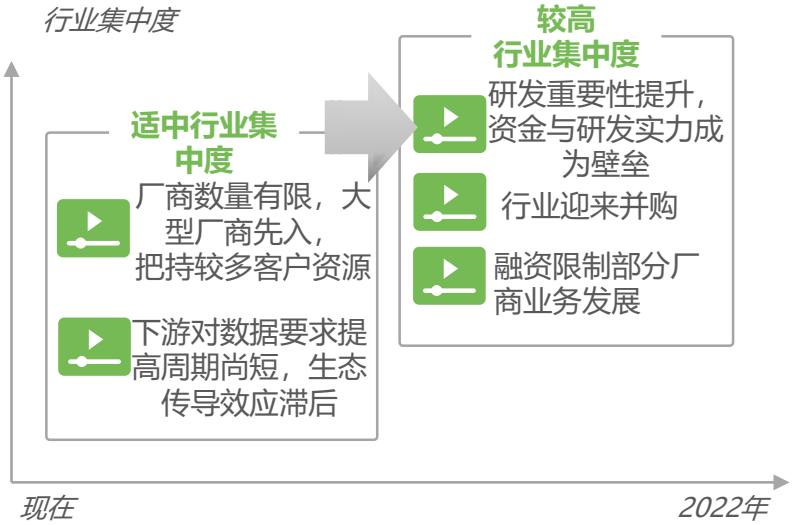
来源：艾瑞根据公开资料自主研究绘制。

人工智能基础数据服务市场格局

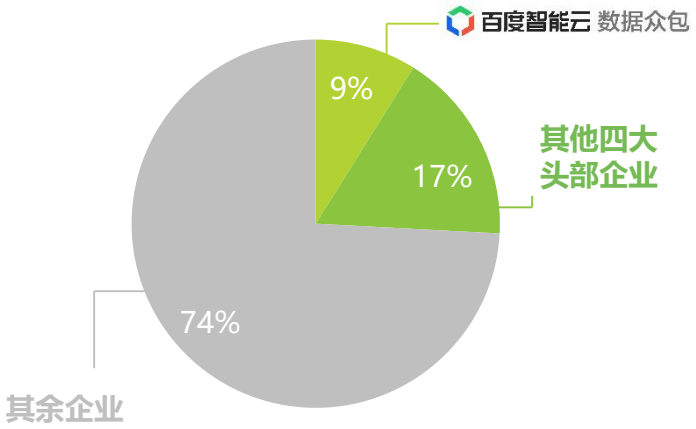
行业将提升至较高集中度，CR5占26%市场份额

目前人工智能基础数据服务行业CR5占26%市场份额，行业集中度较为适中，既非寡占型市场也非充分竞争市场，这一方面是由于百度数据众包、海天瑞声、数据堂等企业进入市场较早，积累了较多客户资源，另一方面则是由于下游企业之前多采用公开数据集训练模型，对数据的高精度要求由来尚短，受生态传导效应滞后影响，市场门槛还不显著，资金与研发实力较为薄弱的中小企业还有较强的发展土壤。然而未来，随着下游企业发展壮大，直接使用外包团队成本低廉、数据安全可控性强，一些基础性需求将由下游企业自给自足，外部的数据服务商现有的存量市场面临下降，因此必须承担高难度、前沿独特性任务，这就要求其自身投入高精度、专业化数据处理工具的研发和人工智能算法基础研究，以把握客户需求，开拓增量市场，因此资金与研发实力成为较高行业门槛，同时受近年资本市场冷却影响，一批中小型厂商面临业务收缩，再者部分厂商如倍赛开始在业内并购，参考海外数据服务市场发展情况（海外行业巨头Appen多次并购其他企业），并购也将成为市场趋势，多种因素叠加影响下，行业集中度将提升。

人工智能基础数据服务集中度趋势



2018年人工智能基础数据服务市场份额

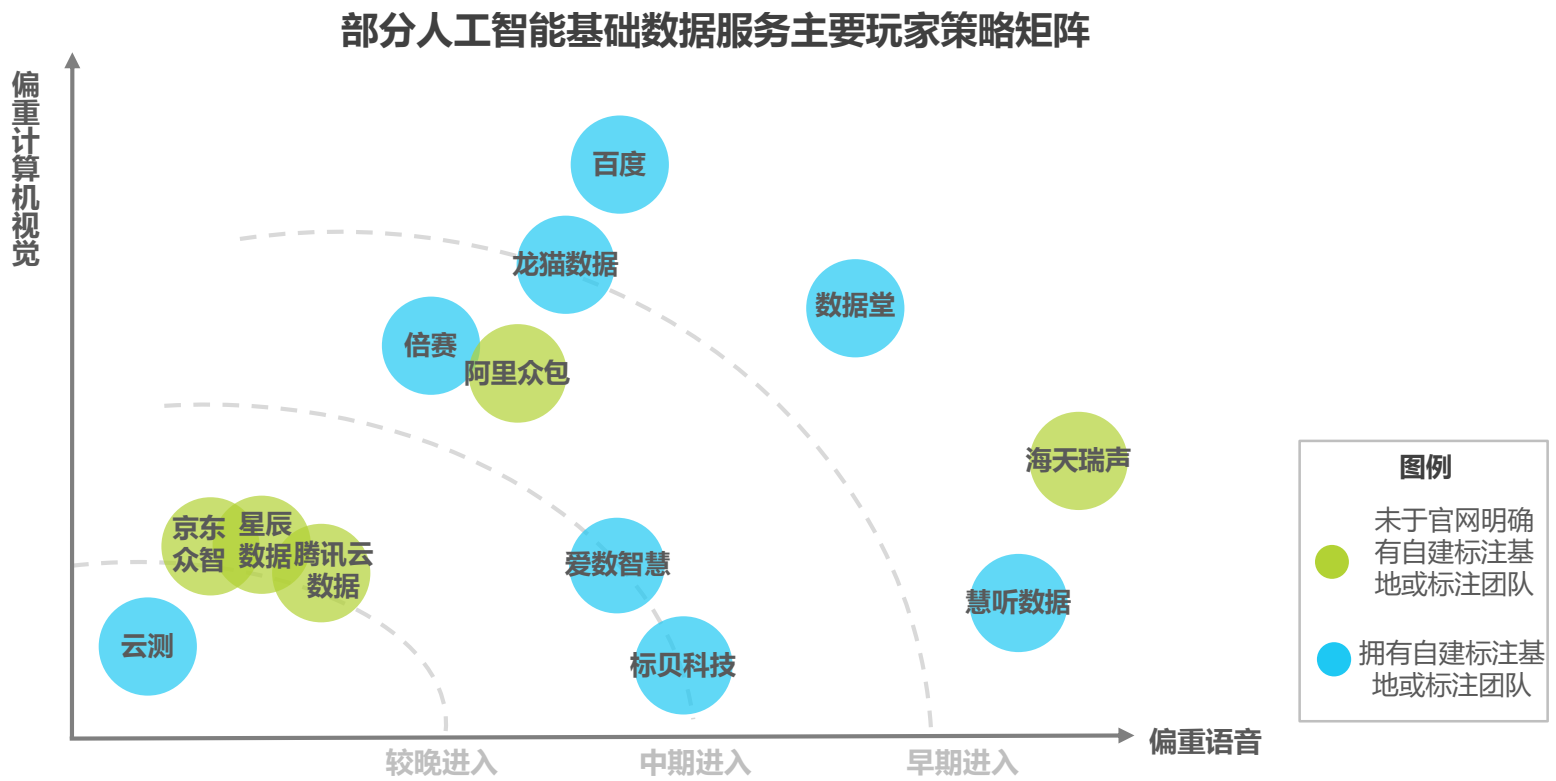


来源：根据公开资料和专家访谈，结合统计模型推算，艾瑞自主研究绘制。

人工智能基础数据服务市场格局

部分人工智能基础数据服务主要玩家策略矩阵

业内玩家按照业务方向和进入市场的时机可做粗略划分，包括早期进入玩家、中晚期进入玩家、偏重视觉类业务玩家、偏重语音类业务玩家等。其中，业务更偏重语音类数据的玩家，通常拥有较多的自有知识产权数据集；拥有自建标注基地或全职标注团队的则多为偏重视觉类的玩家。



来源：根据企业官网公开资料，艾瑞自主研究绘制。

人工智能基础数据服务竞争力要素

优质人工智能基础数据服务供应商要素模型

优质人工智能基础数据服务供应商的基本发展态势可从技术、产能、商务、数据资产、管理等五个方面判断。技术主要关注数据处理工具、平台和人工智能基础技术研究，产能主要关注产能的充足性和调度能力，商务主要关注市场覆盖率和续单率，数据资产主要关注数据资产合规性、复用率，管理主要关注资金、项目管理平台质量与安全管控度、有经验人员保有率等。

优质人工智能基础数据服务供应商要素模型

技术

- 开发与时俱进的数据处理工具，以应对高精细化、高细粒度的数据处理需求，并保证针对差异化需求架构较快完成定制开发
- 增强数据处理平台终身学习能力，由机器学习人工标注，提升预标注和自动标注的能力
- 介入人工智能基础技术研究，紧跟下游需求变化

产能

- 厂商应具备充足产能，一方面拓展与上游供应商的合作关系，另一方面为众包平台引流。同时，下游客户对标注人员的素养和信誉度愈发看重，自建一部分专职标注团队承接高精度任务将带来竞争优势
- 产能调度方面，除项目经理调度外，尽早研发需求与供给规模分布模型，依据需求的时间与质量要求和产能人员的经验、信誉、负载等维度实现智能化任务分配，将优化产能提供效率、降低项目风险，获取竞争优势

数据资产

- 对公司拥有知识产权的数据，确保数据授权的完备，避免合规风险，并合理配置，针对复用率高的数据类型建立数据资产库

商务

- 销售团队对市场渠道的覆盖增强，针对不同类型的客户资源个性化维护，提升续单率
- 售前售中售后体系的完善程度

管理

- 确保资金周转能持续为企业回血，股东与企业管理层维持良好的沟通和业务协同机制
- 完善项目管理平台，建立全面的质量管理和人员培训机制，以降低管理成本、优化口碑
- 保持大量订单吞吐，建立激励机制，以降低有经验的标注人员流失率
- 强调数据安全性，通过私有部署、管理流程全程多层把控、自建标注基地或全职团队等方式实现对数据安全的有效管理

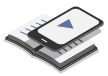
来源：根据公开资料，艾瑞自主研究绘制。

视图基础数据服务分类及应用场景

主要对视图数据检测、框选、分割，新型产品需求不断攀升

视图基础数据服务主要为计算机视觉算法模型提供场景对应的算法训练采集所需的视图数据，针对已采集数据进行框选、关键点标注、属性标注等标注工作，现主要应用于智慧城市、智慧零售、手机拍照、智能质检与预测性维护、商业地产、医学影像AI等领域。由于智慧城市等CV主赛道算法逐渐走向成熟，当前计算机视觉厂商对新赋能领域以及目前已进入领域中较为长尾化的细分场景需求开始增强，及时拓展相应新的精细化数据产品在视图基础数据服务市场中至关重要。

视图基础数据服务分类及应用场景

用于图片分类处理		● 场景化图片数据服务 采集数据后，对图片进行描点、划线、框选、目标检测、关键点标注、目标分割、属性标注等，可用于智慧零售、工业质检、人体识别、动物识别与监测等各类场景
用于内容提取比对		● 人脸人像数据服务 采集阶段提供不同姿态、不同年龄段、肤色的人像数据，标注阶段提供在图像中检测和跟踪人脸、人脸关键点标注、人脸特征标注等服务
		● OCR数据服务 对含有文本的图片做框选标注，包括手写内容、卡片、票据等
用于视频处理		● 视频数据标注服务 对视频主体分类、进行人物及物体属性标记、主体行踪轨迹分析、主体朝向标记、画面起始点标记等

典型案例 x 百度智能云 数据众包

- ✓ 人像采集能力完备，可实现汉族、少数民族、白人、黑人、印第安人、中东人、中亚人、南亚人、东南亚人等多种人像采集
- ✓ 拥有人脸打点、物品分类、自动贴边等标注工具，人像标注准确率达到98%，单张人脸支持150点的精细标注
- ✓ 拥有复杂条件采集能力，可在不同光线、道具、表情、背景采集数据

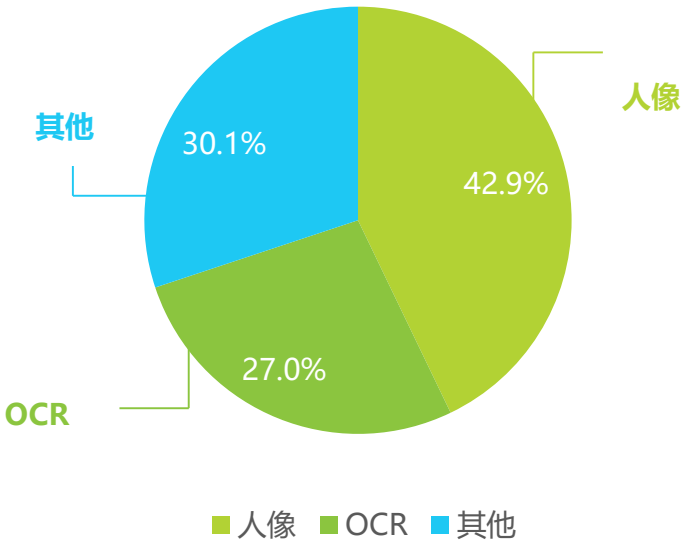
来源：根据公开资料和专家访谈，艾瑞自主研究绘制。
注：本章将自动驾驶相关的数据服务单独阐述，因此本章视图基础数据服务均指除自动驾驶以外的视图基础数据服务。

视图基础数据服务市场现状

人像与OCR数据是视图基础数据服务的主流

在不考虑自动驾驶的前提下，2018年视图基础数据服务市场达到6.6亿元，人像与OCR数据是视图基础数据服务的主流，尤其人像数据占市场的42.9%。OCR占27%，其他的人体识别数据、商品识别数据、工业质检数据、医学影像数据及其他新场景数据等较为分散，合计占市场30.1%。

2018年视图基础数据服务市场细分结构
（不含自动驾驶）



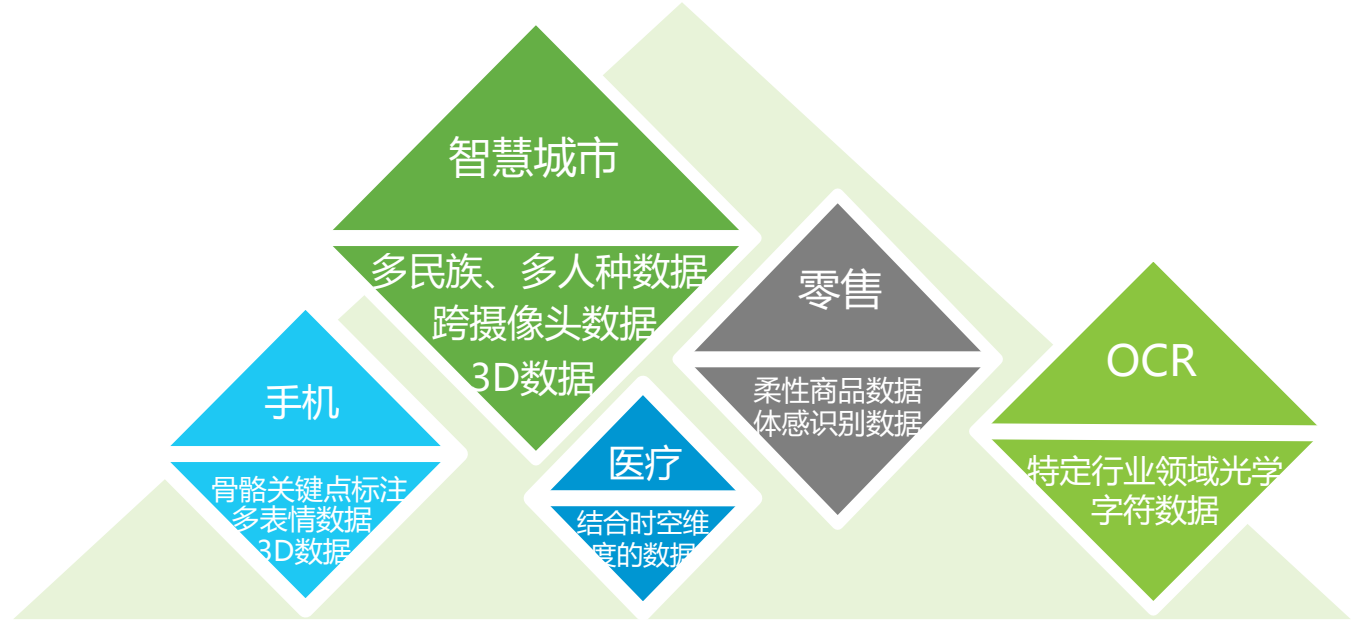
来源：根据公开资料和专家访谈，结合统计模型推算，艾瑞自主研究绘制。

视图基础数据服务技术趋势

针对算法研发方向判断数据需求，挖掘增量市场

按照数据使用方向，可以划分为新算法模型搭建与研发、在已有算法基础上增加新模块、解决方案交付过程中定制优化等三类，其中新算法模型搭建与研发和在已有算法基础上增加新模块类型的数据需求是可以根据相应机器视觉算法的前沿研发方向来判断预测的。例如，就智慧城市场景而言，针对汉族的人脸识别和视频结构化已较为成熟，在实际应用场景中还需针对少数民族和其他人种进行优化以提升整体算法准确率，此外，跨镜追踪成为场景研发热点，相应的跨摄像头数据如何标注对算法训练也会产生较大影响，再及，深度相机可以帮计算机读懂三维立体的监控视频，还能够较好地解决复杂光照条件下视图数据采集的问题，也将在未来成为重要的研发方向，综上，多民族、多人种数据、跨摄像头数据、3D数据的采集与标注服务将为视图基础数据服务市场的发展带来增量空间，OCR、手机、零售等其他领域也同理可针对算法研发方向挖掘增量市场。

机器视觉研究方向与数据使用趋势



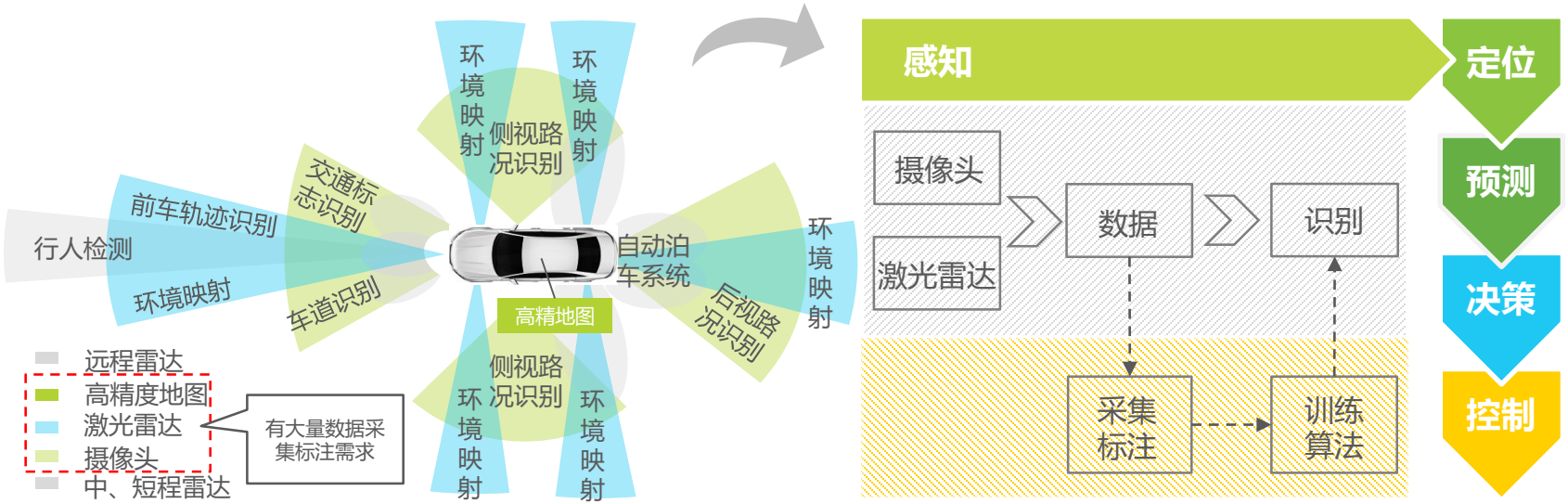
来源：根据公开资料和专家访谈，结合统计模型推算，艾瑞自主研究绘制。

自动驾驶基础数据服务应用场景

算法尚未成熟，对数据有长期需求，且缺口仍在

L3级别以上的自动驾驶系统主要有感知、定位、预测、决策和控制五部分，其对于计算机视觉技术的需求度远高于ADAS，系统需要对雷达、摄像头等传感器采集的点云和图像数据进行抽取、处理和融合，构建车辆行驶环境，为预测和决策做依据，这对于算法的准确性和实时性有极大考验。目前自动驾驶的视觉技术主要应用有监督的深度学习，是基于已知变量和因变量推导函数关系的算法模型，需要大量的标注数据对模型进行训练和调优。在世界级无人驾驶大赛中，主办方往往提供近亿张图片、数十万张标注图片供参赛团队训练使用；在路测或真实道路驾驶时，如人车混杂、分布稠密、行为多变等复杂环境问题更需要海量的真实路况数据不断对算法进行优化，才能保障无人驾驶车辆正常可用。如今国内自动驾驶飞速发展，AI公司、科技公司、高精地图厂商、车厂等参与者众多，该领域的数据采集和标注需求已经成为AI基础数据服务的主要项目之一，且自动驾驶算法应用仍待优化，数据需求缺口仍在，市场远未饱和。

自动驾驶场景中AI基础数据服务的价值



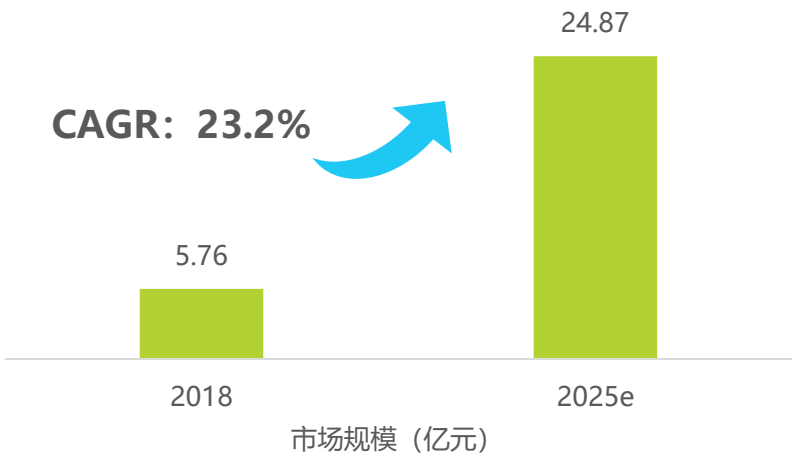
来源：艾瑞根据公开资料自主研究绘制。

自动驾驶基础数据服务市场现状

2025年采标规模将超24亿，科技公司和车厂是主要需求方

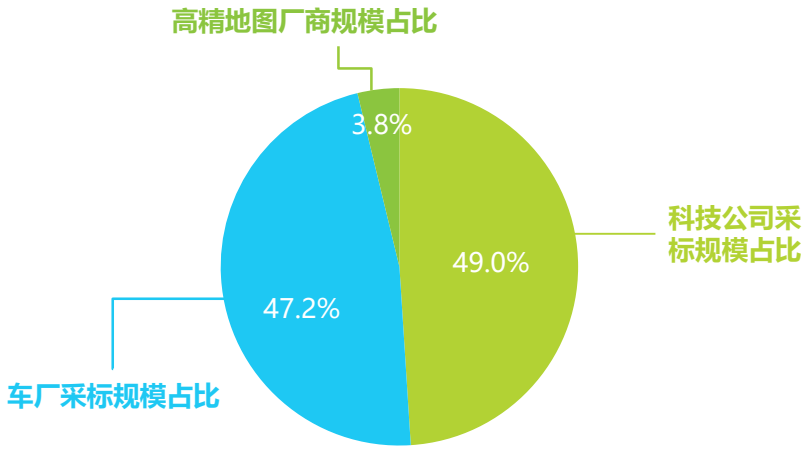
自动驾驶基础数据主要是道路交通图像、障碍物图像、车辆行驶环境图像等，需求方以科技公司、汽车厂商和高精地图厂商为主，2018年自动驾驶行业基础数据服务规模为5.76亿元，预计2025年将超24亿元，三方规模占比分别为49%、47.2%和3.8%，行业数据总任务量超一亿张，2D图像标注与3D点云标注任务量基本为2:1。其中高精地图厂商算法较为成熟，数据自动化标注程度可达90%左右，外包需求较少；以百度、图森未来为代表的自动驾驶科技公司一直是该领域基础数据服务的主要买方，平均各家算法训练图像数据累积需求在千万级以上，随着落地项目进程加快，将会有更多细分场景的需求产生；近几年，汽车厂商在ADAS和自动驾驶方向的投入明显，上汽、吉利等厂商年投入均可达数亿元，对于数据的采集和标注需求也逐年增加，预计未来3年中，汽车厂商将成为需求主力。

2018年-2025e中国自动驾驶AI基础数据服务规模



来源：通过对需求方和供应商项目结构的研究，利用模型测算2018年规模；根据需求方研发计划和供应商执行上线测算增速及未来规模。

2018年自动驾驶AI基础数据服务规模占比



来源：通过对需求方和供应商项目结构的研究，得到当年的规模占比。

自动驾驶基础数据服务技术趋势

2D图像标注项目较多，3D点云数据采标能力门槛较高

自动驾驶领域的视觉数据可分为车载摄像头采集的2D图像数据和激光雷达采集的3D点云数据，从项目来看目前以2D图像的标注居多，数据标注公司通过标注工具，利用语义分割、障碍物识别等技术对图片内目标元素进行框选标注，一期项目往往交付数万或数十万张标注图片，准确率要求高于95%以上；相比于2D图像，3D点云数据的采集和标注更为复杂，激光雷达通过发射激光束并接收物体反射信号，来探测车辆周围物体的距离和速度，是目前自动驾驶车辆不可或缺的传感器，系统除知道空间数据外，还要对探测到的物体进行识别，这就需要对3D点云数据进行标注以训练算法。从项目来看，车厂对于采集和标注都有需求，而科技公司自身测试车会采集大量数据，标注需求强于采集需求。3D点云标注对于标注工具的能力和标注员经验都有较高要求，项目的成本也高于其他图像标注，又因为道路数据雷达车采集有资质要求，所以国内能同时对3D点云数据进行采集和标注的公司更具有竞争力。

典型案例——百度AI基础数据服务中自动驾驶2D、3D标注能力

百度是为数不多同时具备甲级测绘资质及数据采标能力的企业自建采集车队，有15辆一体化采集车，可采集2D/3D道路数据

- 激光雷达：64/128线
- 工业摄像头：15Hz，环绕车身一周，不同方向焦距不同
- 毫米波雷达：正前+正后

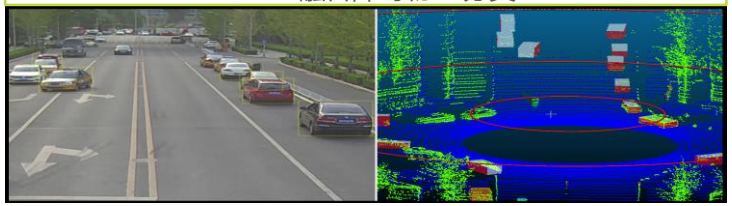
来源：艾瑞根据公开资料自主研究绘制。

自动驾驶数据采集

百度智能云 数据众包

自动驾驶数据标注

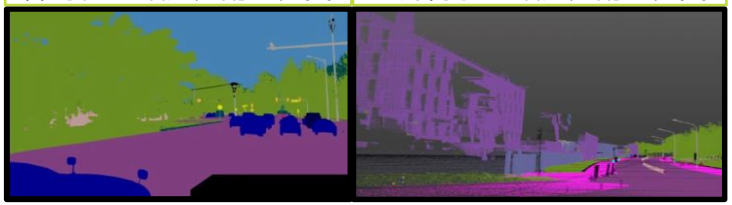
2D-3D融合障碍物16分类



2D框选能力
2400个框/人/天；4万张图/天；正确率99.9%

3D框选能力
800个框/人/天；10万个框/天；正确率99%

障碍物2D语义分割17分类 | 室外街景3D语义分割19分类



2D分割能力
500区域/人/天；5万区域/天；正确率98%+

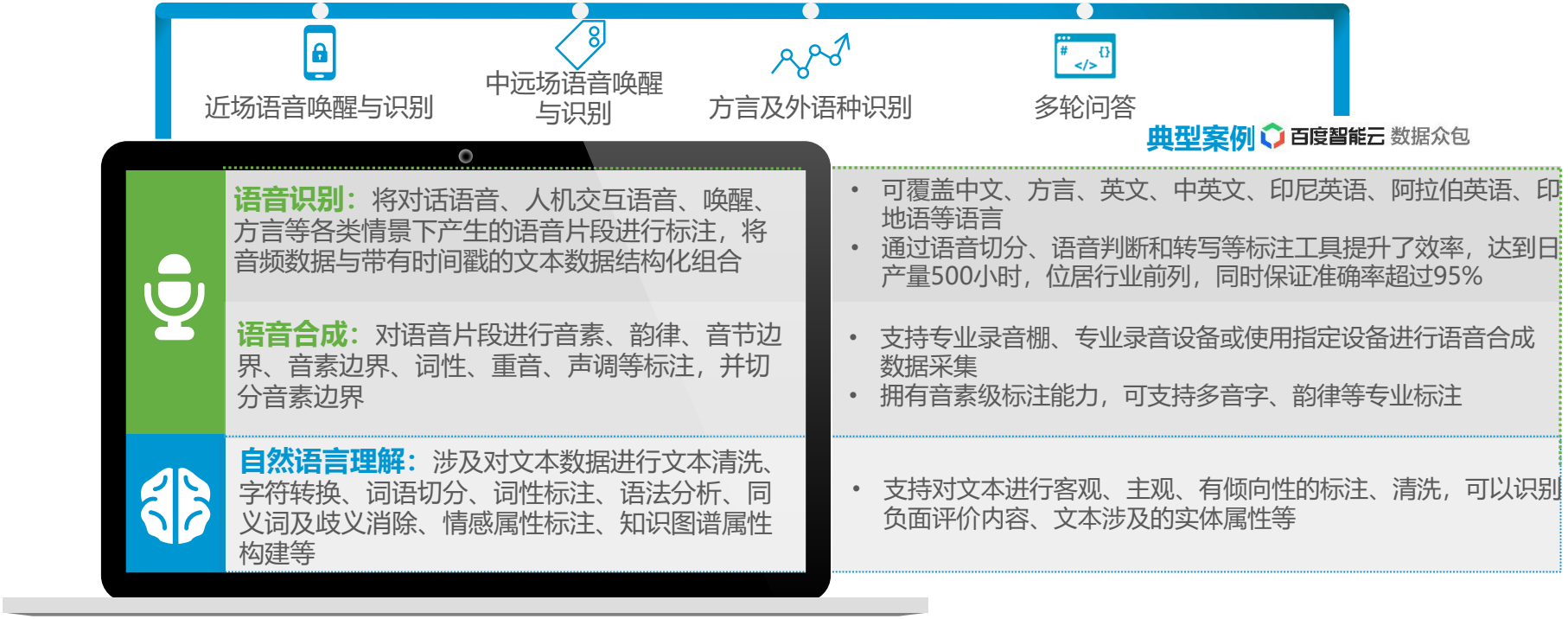
3D分割能力
8帧/人/天；500帧/天；正确率98%+

智能交互基础数据服务分类及应用场景

包括ASR，TTS和NLP数据需求，呈复杂化、高精度趋势

智能交互基础数据服务主要涉及语音识别数据、语音合成数据与自然语言理解数据的采集与标注，应用于近场语音唤醒与识别、中远场唤醒与识别、方言及外语种识别、多轮问答等场景。近年来，智能交互技术对基础数据服务精准度要求越来越高，同时要求基础数据服务与场景的贴合度提升，需要实现歧义消除、在噪音及远场等因素干扰下完成等较高难度的任务。

智能交互基础数据服务分类及应用场景



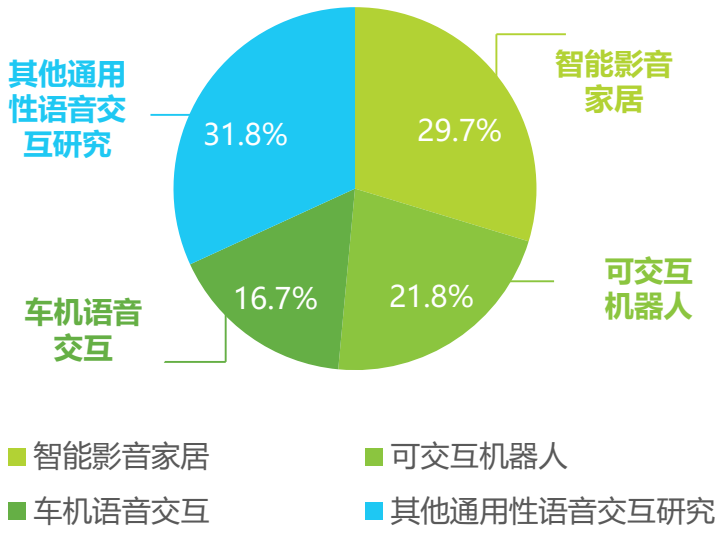
来源：根据公开资料和专家访谈，艾瑞自主研究绘制。

智能交互基础数据服务市场现状

远场语音交互成为主流需求，中文类数据仍占据市场核心

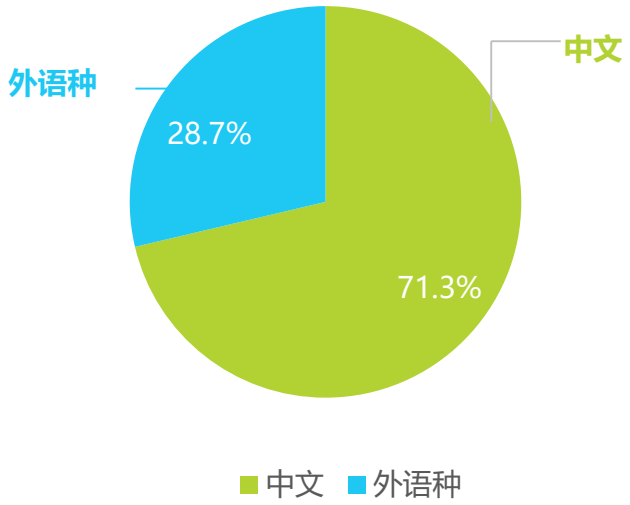
2018年语音交互相关数据服务市场规模达到13.5亿元。语音交互主要分为近场交互、中场交互和远场交互，以智能影音家居、可交互机器人和车机为代表的中远场交互类数据服务需求合计占到智能交互基础数据服务的68%，成为当前智能交互基础数据服务的主流需求，因此针对远场语音交互的低噪声环境服务具有较强发展潜力和议价能力。在服务语种上，中文（含方言）服务占据71%的市场份额，外语种资源相对稀缺，采集和标注难度较大，成本相对更高，目前占29%的市场份额。

2018年智能交互基础数据服务应用领域分布



来源：根据公开资料和需求方访谈，结合模型推算，艾瑞自主研究绘制。

2018年智能交互基础数据服务市场语种分布



来源：根据公开资料和行业专家访谈，结合模型推算，艾瑞自主研究绘制。

智能交互基础数据服务技术趋势

实现跨语音识别、语义理解的复合数据标注

目前企业在智能交互系统的建设中，对单纯的语音识别或合成方面技术能力相对较完善，而在上下文理解、多轮对话、情绪识别、模糊语义识别、意图判断等方面的研发痛点更强，根据智能交互系统算法的发展，迭代并设计符合算法需求的NLP数据产品，有助于从数据层面推动智能交互系统的发展。特别的，对话系统的效果对标注数据的质量和规模依赖性很强，但目前受标注数据和模型能力的双重制约，对话流程还无法对语音、语义整个交互流程打通，而实现跨语音识别、语义理解的复合数据标注可以帮助减轻语音信息与文本信息之间的信息误传导，对整个对话流程效果增强能够产生积极影响，将增加智能交互基础数据服务探索的可能性。

智能交互系统研究方向与数据使用趋势

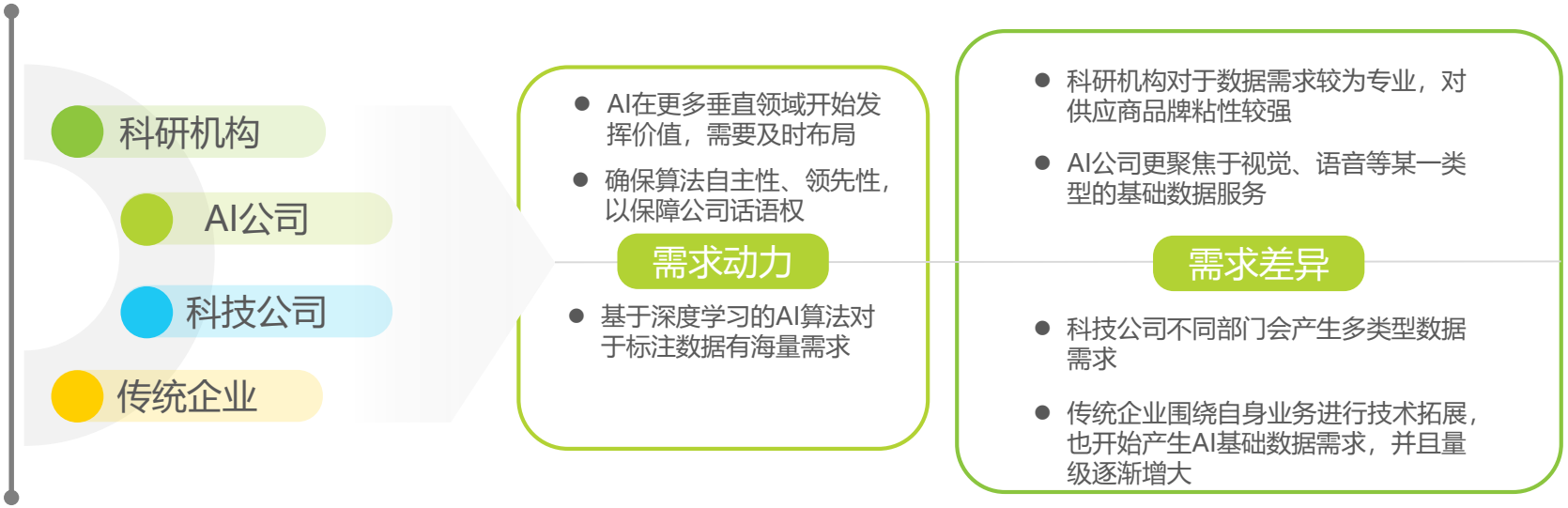
来源：根据公开资料和专家访谈，艾瑞自主研究绘制。

人工智能基础数据服务客户定位

客户分为AI公司、科技公司、科研机构、行业企业四类

从需求方来看，AI公司和科技公司占主要份额，AI公司更聚焦于视觉、语音等某一类型的基础数据服务，而科技公司结合集团优势，向人工智能整体发力，不同部门会产生多类型数据需求，科研机构需求占比较小。此外传统意义上的行业企业，如汽车厂商、手机品牌商、安防厂商等传统企业围绕自身业务进行技术拓展，也开始产生AI基础数据需求，并且量级逐渐增大，未来将释放更多市场空间。

AI基础数据服务市场需求情况



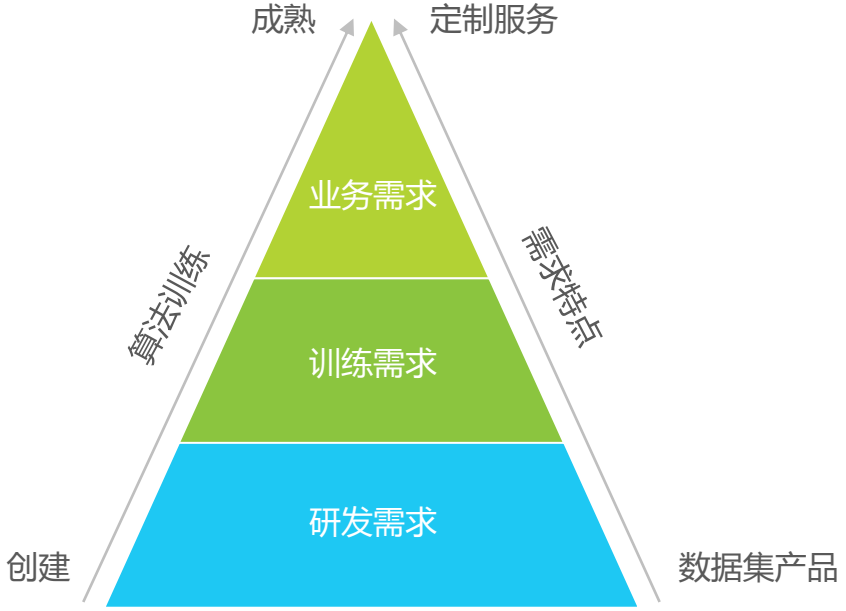
来源：艾瑞根据公开资料自主研究绘制。

人工智能基础数据服务核心需求类型

AI应用三大阶段，对基础数据服务产生差异化需求

企业应用人工智能算法要经历研发、训练和落地三个阶段，不同阶段对于AI基础数据服务也有差异化需求。研发需求是新算法研发拓展时产生的数据需求，一般量级较大，初期多采用标准数据集产品训练，中后期则需要专业的数据定制采标服务；训练需求是通过标注数据对已有算法的准确率、鲁棒性等能力进行优化，是市场中的主要需求，以定制化服务为主，对算法的准确性有较高要求；落地场景的业务需求中算法较为成熟，涉及的数据采集和标注更贴合具体业务，如飞机保养中的涂料识别数据等，对于标注能力和供应商主动提出优化意见的服务意识有较强要求。

AI基础数据服务需求类型和数据采标要求



业务需求

业务需求一般为算法较成熟的核心场景，对服务意识有较高要求

- 需要私有化部署的标注平台，或较强的数据安全管理流程
- 根据业务特点，对采标的数据内容有特殊指向，采标难度大
- 对执行方稳定性和效率有较高要求

训练需求

训练需求一般是对算法的准确性和鲁棒性进行打磨，对数据的准确性要求较高

- 通常需要私有化部署的标注平台，或较强的数据安全管理流程
- 对数据标注的内容需求较为丰富，准确率一般要求95%以上，对无效数据的处理有较高要求

研发需求

研发需求是对新拓展领域或新建算法的训练，一般量级较大

- 新建算法一般使用开源数据集或数据集产品训练，新拓展领域往往采用迁移学习训练，保密项目对数据安全要求高
- 数据标注需求量大，较上两种需求，标注内容倾向于标准化

来源：艾瑞根据公开资料自主研究绘制。

人工智能基础数据服务需求痛点

五大需求痛点决定AI基础数据服务商的服务标准

目前需求方在选择数据服务时往往会遇到数据安全、采标能力、数据质量、管理能力、服务能力等痛点。对于数据安全，需求方希望基础数据服务商有明确具体的安全管理流程，对数据传输、存储，以及结项后的数据销毁等环节比较重视。在采标能力方面，需求方算法越来越贴近业务，希望数据服务商对于自动驾驶、工业等有一定门槛的领域有采集能力，并且能理解客户意图，配合标注，甚至可以提出标注建议；根据市场反应，大多数数据服务公司首次交付项目时，数据的准确率普遍偏低，都需要一到两次的返工，故需求方对无效数据少、准确率高的公司更加青睐。对于执行效率，一般AI基础数据服务商都能在项目周期内完成，但管理能力较弱的公司很难在兼顾多个项目时做到精力集中、高质量地服务客户，同时执行团队的素养与信誉也是重要影响因素。服务意识是一项软实力，需要AI基础数据服务商能够积极配合、快速响应需求方要求。

AI基础数据服务需求痛点

来源：艾瑞根据客户访谈和公开资料自主研究绘制。

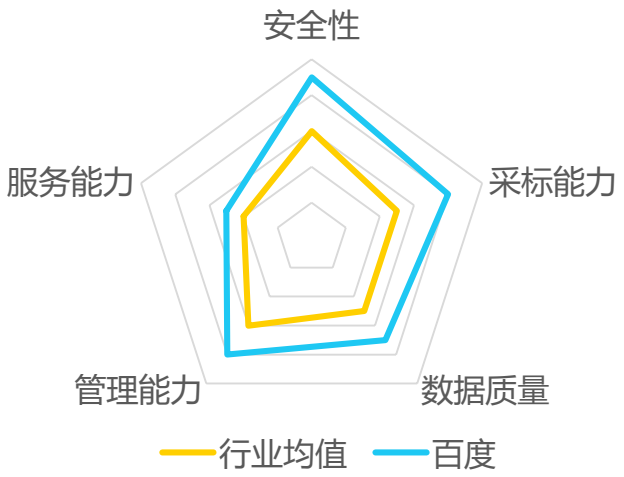
人工智能基础数据服务标准模型

人工智能基础数据服务标准模型

人工智能基础数据服务标准模型将五大需求痛点量化，对需求方而言，安全性、采标能力、数据质量、管理能力、服务能力等五要素重要程度相当，相对而言，采标能力是一项比较硬性的门槛，其重要度更高，服务能力更多影响商务沟通效率，重要度略低。目前，行业平均水平普遍与需求方的要求还有一定差异，特别是在采标能力与数据质量方面。百度数据众包作为行业头部玩家，在上述五要素方面都做出大量投入，如安全性上，架设专线将百度机房直连自建标注基地机房，保证数据传输安全，室内设有监控视频，可随时查看现场情况；采标能力上，具备可采集2D/3D道路数据的一体化采集车和人脸标注、语音标注、无人驾驶等多场景标注工具；管理能力上，聚合众包平台、标注基地和大量优质供应商，可随客户需求匹配人力，并建立了较为完备的考评、激励、培训机制；数据质量上，通过技术自动审核、实时校验，并设置了自检、质检团队、项目经理等三轮检验机制，项目管理平台还可根据正确率高低定制科学的抽审规则，以保障高准确率。

需求方理想型数据服务公司模型

代表品牌：百度数据众包



注释：评价指标如下——安全性：标注平台私有化部署、安全管理流程等情况；采标能力：采标场景覆盖完整度、采标工具开发能力等；数据质量：首次交付数据准确率；管理能力：项目管理系统、运营机制等；服务能力：配合度、响应度、前瞻思考能力、人工智能基础研究积累等。
来源：艾瑞自主研究绘制。

人工智能基础数据服务发展建议

企业由被动执行向主动服务的意识跃迁

单纯依据客户各个项目的诉求进行数据采集和标注属于被动执行，主观能动性低、行业边界有限，各家公司的产品和服务趋于同质化、竞争呈胶着状态，制约着AI基础数据服务的发展。通过对需求方的研究，发现除安全性、质量、效率等核心关注点之外，越来越多的需求方对数据服务公司产生了主动服务的需求，希望数据公司能够更懂算法技术、更懂需求场景，甚至能参与到算法的研发中来，给出数据采标方面的优化建议，这也为数据服务商形成差异化竞争带来了契机，尤其是在AI落地阶段，在垂直场景中能够形成一套集调研、咨询、设计、采集、标注为一体的AI基础数据整体解决办法，将在收入和业务边界上实现突破。

由被动执行向主动服务的意识跃迁

来源：艾瑞根据公开资料自主研究绘制。

人工智能基础数据服务发展建议

政府出台引导政策建设数据基地

在各地积极引导人工智能产业发展的大背景下，人工智能基础数据服务在政策层面的关注度仍有待提高。一方面，我国人工智能企业多集中在北、上、深、杭等城市，各地人工智能的发展需要因地制宜，数据作为人工智能产业发展的重要基石，优先发展人工智能基础数据服务基地对增强地区人工智能发展基础有着重要作用，另一方面，人工智能数据采集与标注的具体执行环节对人员专业背景要求门槛较低，建设数据基地有利于促进就业与社会经济发展。同时对于人工智能基础数据服务商而言，拥有一部分全职采标团队对于承接高精度任务有较强的积极影响，可以借助经验较为丰富、管理较为规范的专职团队高效完成任务，使有经验的标注人员流失率降低，以控制管理成本与风险，因此与各地政府合作引导数据基地建设将对企业发展有重要价值。

出台引导政策建设数据基地的意义

来源：艾瑞自主研究绘制。

关于艾瑞

在艾瑞　我们相信数据的力量，专注驱动大数据洞察为企业赋能。

在艾瑞　我们提供专业的数据、信息和咨询服务，让您更容易、更快捷的洞察市场、预见未来。

在艾瑞　我们重视人才培养，Keep Learning，坚信只有专业的团队，才能更好地为您服务。

在艾瑞　我们专注创新和变革，打破行业边界，探索更多可能。

在艾瑞　我们秉承汇聚智慧、成就价值理念为您赋能。

- 我们是艾瑞，我们致敬匠心　始终坚信“工匠精神，持之以恒”，致力于成为您专属的商业决策智囊。

扫描二维码
读懂全行业

海量的数据　专业的报告

400-026-2099　ask@iresearch.com.cn

法律声明

版权声明

免责条款

本报告中行业数据及相关市场预测主要为公司研究员采用桌面研究、行业访谈、市场调查及其他研究方法，并且结合艾瑞监测产品数据，通过艾瑞统计预测模型估算获得；企业数据主要为访谈获得，仅供参考。本报告中发布的调研数据采用样本调研方法，其数据结果受到样本的影响。由于调研方法及样本的限制，调查资料收集范围的限制，该数据仅代表调研时间和人群的基本状况，仅服务于当前的调研目的，为市场和客户提供基本参考。受研究方法和数据获取资源的限制，本报告只提供给用户作为市场参考资料，本公司对该报告的数据和观点不承担法律责任。

为商业决策赋能

EMPOWER BUSINESS DECISIONS

iResearch

艾 瑞 咨 询