

中国人工智能基础数据服务 行业白皮书

2019年



海量行研报告免费读



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。

“ 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。 ”

人工智能基础数据服务行业概述	1
----------------	---

人工智能基础数据服务市场现状	2
----------------	---

人工智能基础数据服务场景分析	3
----------------	---

人工智能基础数据服务需求分析	4
----------------	---

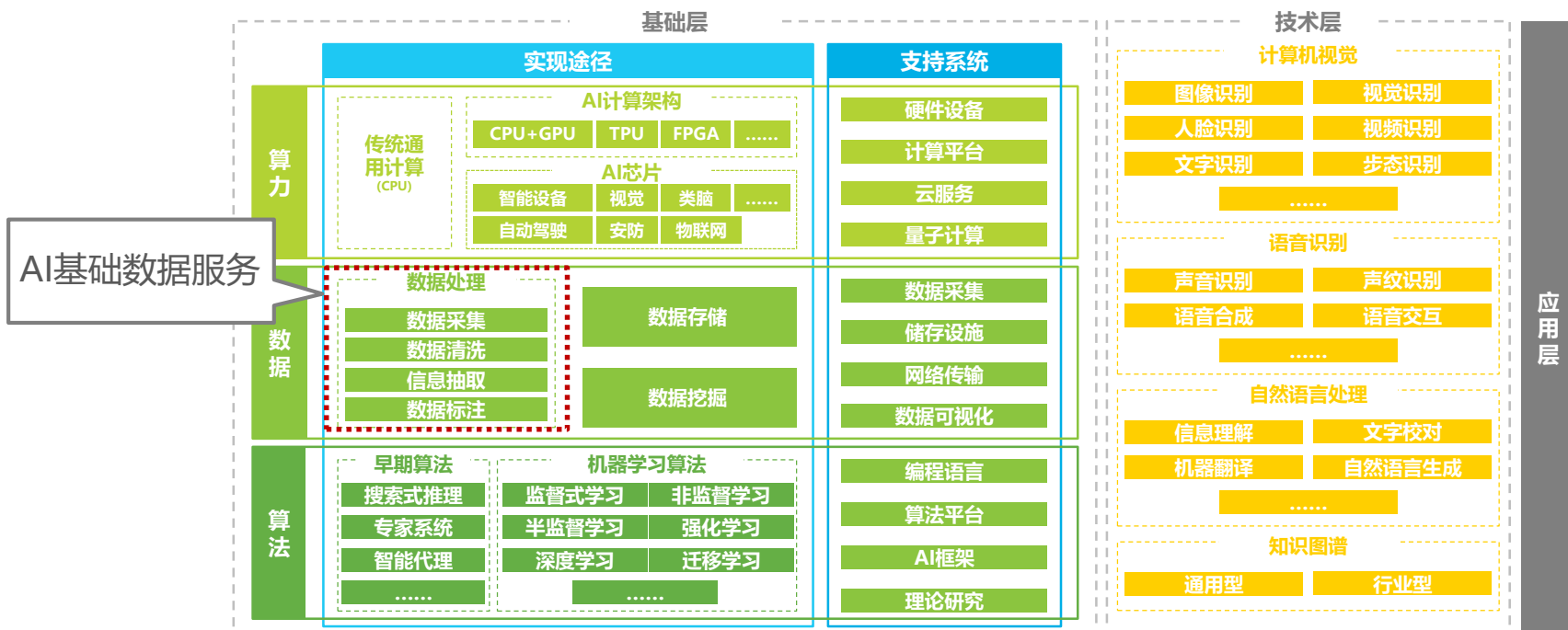
人工智能基础数据服务发展趋势与建议	5
-------------------	---

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

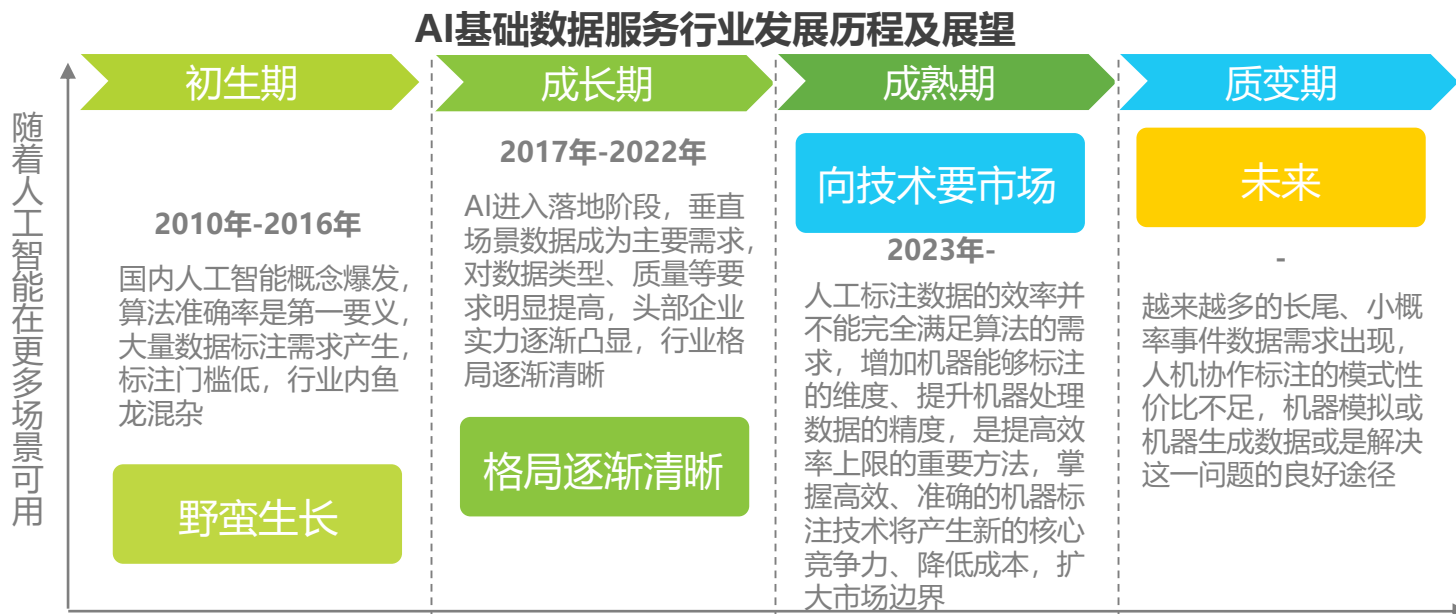


来源：艾瑞根据公开资料自主研究绘制。

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。



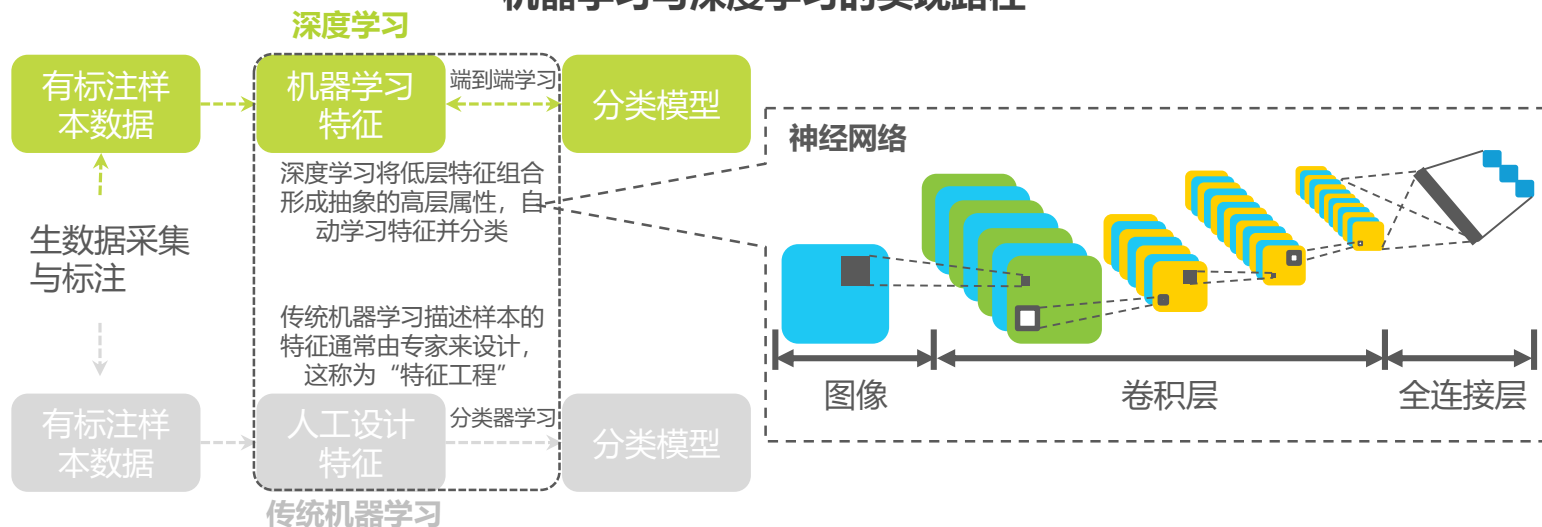
来源：艾瑞根据公开资料自主研究绘制。

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 iResearch

艾 瑞 咨 询

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

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

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

预览已结束，完整报告链接和二维码如下：

https://www.yunbaogao.cn/report/index/report?reportId=1_21061

