



人工智能基础数据服务白皮书

2023/03

因我不同
成就不凡

始于1845

核心观点



整体市场：人工智能产业的快速增长带动了人工智能基础数据服务市场的蓬勃发展，自动驾驶是未来五年最重要的应用领域

- 结构化数据是人工智能算法开发迭代的重要基础，人工智能基础数据服务市场受人工智能核心产业发展带动仍将保持高速增长，预计**2027年市场规模有望达到130-160亿元**。
- **自动驾驶**是人工智能基础数据服务市场占比最大的下游应用，随着自动驾驶算法技术不断迭代与场景落地，**未来占比有望进一步提升**。



发展趋势：标注复杂化、自动化、全栈式服务需求以及愈加严格的数据合规需求是AI基础数据服务市场的四大趋势

- **标注复杂化：**随着算法迭代创新以及场景功能的持续扩展，数据标注元素和标注信息维度均将大幅增加，对于数据基础服务供应商提出了更高的要求；
- **自动化标注：**AI赋能的自动标注工具逐渐成为基础数据服务商和AI算法公司降本增效的利器，推高行业集中度；
- **全栈式服务：**下游算法应用方自研人工智能算法的趋势逐渐显现，需求方对于“基础数据服务+云资源+工具链”的全栈式服务需求提升（包括算法公司，但主要由应用方驱动），特别是对于工具链产品的需求将随着商业化场景的成熟由自动驾驶领域向各行各业拓展，适应未来的迭代需求；
- 从自动驾驶基础数据服务需求方的角度出发，**整车厂及Tier1自研需求不断提升**，同时技术迭代带来的更复杂、更专业的数据标注需求，这将推升整个自动驾驶行业的基础数据服务外包需求，**并进一步释放对工具链及全栈式服务的需求**。
- **数据合规性：**数据安全法律法规体系不断完善，基础数据服务商在数据脱敏、数据采集的测绘资质要求等环节的专业性价值会为其带来竞争优势。



竞争格局：传统的专业型基础数据服务商仍是行业重要组成，但科技巨头企业依托其科技实力和强大资源，逐渐占据竞争优势

- 科技巨头、专业型基础数据服务商以及科技初创企业是人工智能基础数据服务行业的主要参与者，其中专业型基础数据服务商布局早，服务经验积累深，在市场中仍占有较大份额，而科技巨头近两年发力明显，快速抢占市场；
- 自动化标注、专业数据采标及全栈式服务是人工智能基础数据的三大核心能力，其中领先的科技巨头在三个维度均有持续的积累，综合能力最强。以百度为代表的科技巨头依托其研发能力、产业链协同资源和对AI算法的理解、稳定和专业的标注团队，竞争优势显著，市场份额有望持续提升。

目录

章节	页码
第一章：人工智能，数据先行：AI基础数据服务持续快速发展	4
第二章：AI基础数据服务趋势：复杂化、自动化、全栈化及合规化	12
第三章：科技巨头已下场，强者优势愈发清晰	19



第一章

人工智能，数据先行：

AI基础数据服务持续快速发展



人工智能正在加速渗透应用到各行各业

中国人工智能产业处于高速增长期，正在加速向各行各业渗透，包括互联网娱乐、智能制造、智慧医疗、智能安防及自动驾驶等，而自动驾驶等应用场景的复杂性又反向推动了人工智能的迭代演进

人工智能主要应用场景



互联网娱乐



智能制造



智慧医疗



智能安防



自动驾驶



场景痛点

- 存在大量针对黄恐暴、抄袭等方面的内容审核需求，但人工审核效率低、成本高
- 传统客服也面临成本高昂的问题
- 人力工序过程失误率高，且难以追溯
- 部分工作环境存在高危性
- 国内医疗水平参差不齐，基层卫生医疗水平低下，有经验的医生资源稀缺
- 新药设计难度大、成本高且耗时
- 传统安防无法准确识别人、物与场景
- 犯罪、恐怖袭击等事件无法预知
- 人口红利消失，驾驶员成本高且资源短缺
- 超载及疲劳驾驶导致安全事故频发，造成生命财产损失



AI解决方案

- 采用语音识别、语义切割、图像识别等方式对内容数据进行识别分类，高效实现审核工作
- ChatGPT的诞生大大加快了人机交互的效率与应用
- 利用计算机视觉技术高效准确发现瑕疵品
- 机器人代替人在危险场所完成工作
- 智能影像识别可以通过自动读片快速进行疾病筛查，弥补医疗资源差异
- AI制药能够以更低成本高效发现药物靶点、筛选化合物，大幅提升新药研发效率
- 通过计算机视觉等技术实现人脸识别，从而发现嫌疑人行动轨迹
- 进出楼宇与园区时采用指纹或人脸识别提高识别精确度
- 自动驾驶通过传感器、计算机视觉等技术逐步解放驾驶员，实现车辆的自动驾驶



AI应用场景复杂性



AI精度要求



信息来源：德勤访谈、研究与分析

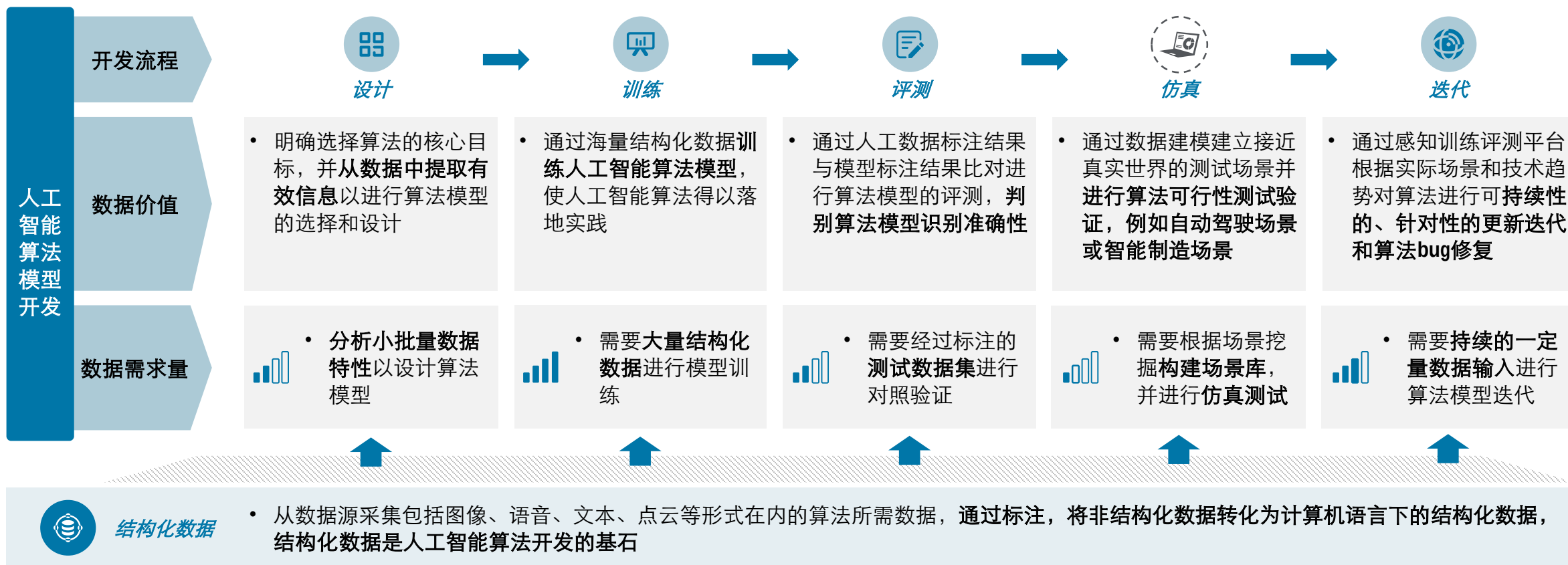
低 高



结构化数据是人工智能快速发展的基石

结构化数据是人工智能算法模型开发和迭代的基础，从设计、训练、评测、仿真到整个算法更新迭代的全生命周期都需要持续不断的结构化数据的输入作为支撑

人工智能算法模型开发流程



信息来源：德勤访谈、研究与分析

低 高



对结构化数据的需求催生基础数据服务产业

基于结构化数据的重要性，人工智能产业逐渐诞生了一批专业人工智能基础数据服务商—通过数据采集与数据标注，有效衔接数据源与具有算法开发需求的企业

人工智能基础数据服务商处于产业链中游，通过提供数据采集和标注服务，连接上游数据来源方和下游人工智能算法研发方

- 市场上现存的大量数据均为非结构化数据，无法直接应用于人工智能算法的研发与训练，需要通过数据的采集与标注将其转化为结构化数据，以供下游人工智能算法研发商使用。
- 这个采集与标注的过程逐渐形成了一项专项工作，主要由专业的基础数据服务商来提供，少量由算法研发企业的自有团队执行

人工智能基础数据服务产业链¹



注释：1. 产业链图谱中代表厂商为不完全列举，排名不分先后
信息来源：德勤访谈、研究与分析

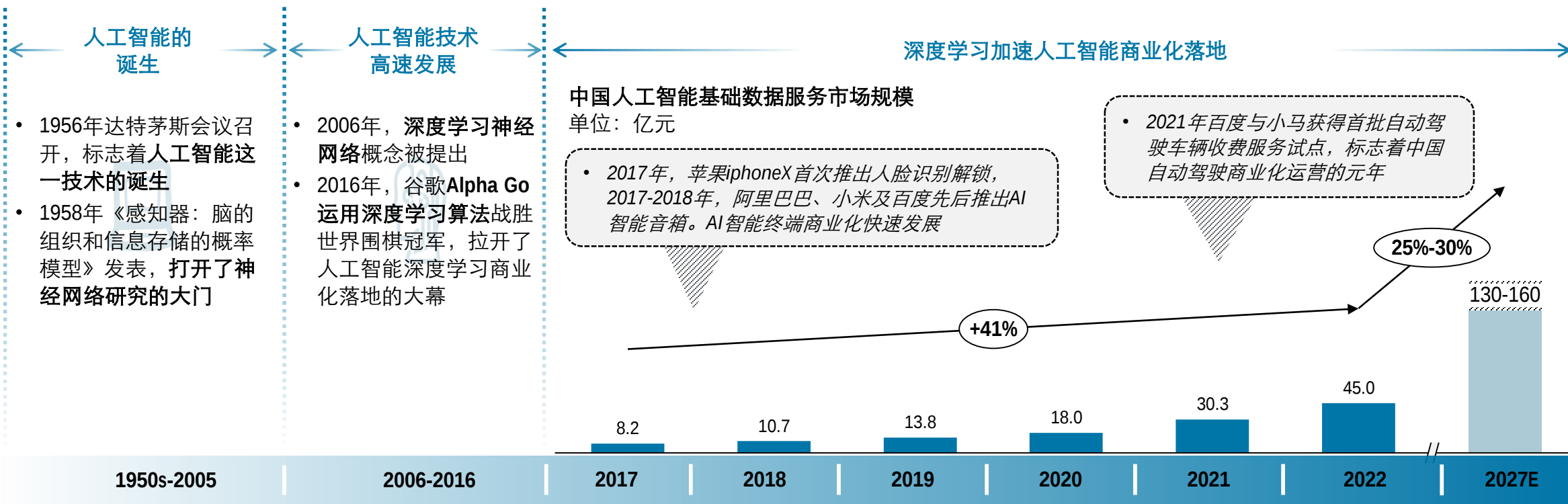


基础数据服务受AI商业化落地驱动高速增长

人工智能技术经历数十载的发展，近年来深度学习加速了人工智能技术的商业化落地，同时也带来了大量AI算法训练需求，推动基础数据服务市场的快速增长

人工智能技术已经经历了较长时间的发展，近年来由深度学习带来的人工智能技术商业化应用落地极大的推动了AI基础数据服务的需求

人工智能发展阶段及重要里程碑



信息来源：公开资料整理、IDC；德勤访谈、研究与分析

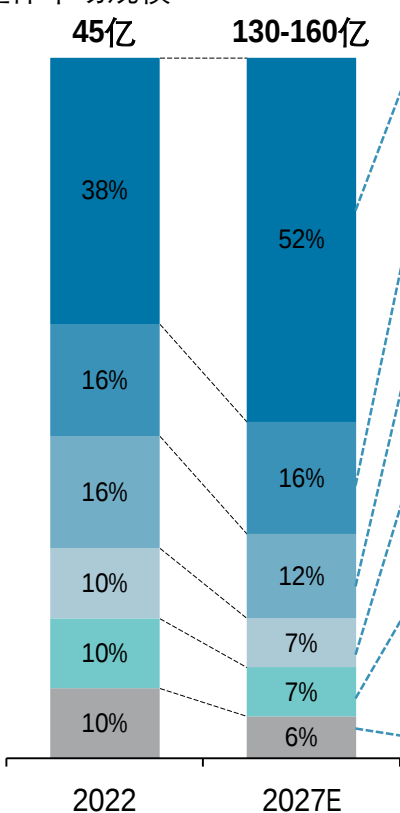


基础数据服务在不同场景的需求各不相同

人工智能基础数据服务应用于众多下游场景，但不同下游场景对数据采集类型以及数据标注对象有着各自的差异化需求，自动驾驶当前是人工智能基础数据服务最重要的应用领域，并将在未来继续维系这一地位

人工智能基础数据服务下游应用占比（2022-2027E）

整体市场规模：



注释：1. 2022-2027E CAGR
信息来源：德勤访谈、研究与分析

数据采集标注类型

未来增长潜力

未来增速¹



自动驾驶

- 采集大量真实、覆盖不同道路天气小概率事件的道路视频图像以及激光点云图像，标注视频图像以及点云数据中的道路可行驶区域、车辆、行人等各类元素

- 自动驾驶行业对数据的需求处于起步阶段，未来技术与功能迭代、场景拓展将带动**数据需求量几何级增长**

32-37%



智慧工业

- 采集产品图像、生产环境画面、设备运行状态画面等数据，标注各类生产状况及产品图像及其状态，如钢铁表面瑕疵或裂纹

- 工业视觉是行业增长主要驱动力，伴随国家对工业领域数字化智能化的重投入，未来行业**需求量有望放量提升**

24-29%



智能安防

- 采集各类公共场所、居民住宅楼及商用楼的监控摄像头数据，标注视频图像中的人脸骨骼点、车辆、动作行为等元素

- 人脸识别精确度的可提升空间有限，但事件感知识别等新场景需求为智能安防基础数据服务**需求带来一定增长空间**

17-22%



AI+互联网

- 采集用户生成的文章、搜索、直播、视频、图像等内容素材，标注文本中的敏感字眼以及视频图像中人的行为、手势、嘴型等动作元素

- 行业快速技术迭代驱动数据迭代需求增长，但由于技术路径正向无监督训练倾斜，**未来长期看数据标注的需求量或将先增后减**

15-20%



智慧医疗

- 采集医疗影像、手术工具、处方、设备控制、病例等数据，标注医疗影像中的人体拉框、骨骼点以及处方病例中的文本等

- 我国老龄化明显，医疗行业AI应用发展旺盛，带动基础数据服务**需求呈现一定增长**

15-20%



其他

- 智能终端：各国人像、小语种、方言等数据
- 智慧金融：票据单据、保险标的、人脸、对话语音等非结构化数据以及风控数据等结构化数据

- 智能终端、智慧金融等场景已较为成熟固化，未来**增长潜力稍低，将趋于稳定**
- 但其他潜在应用领域例如元宇宙板块随着发展成熟或存在市场增长爆发的机遇

12-16%

份额占比：增加 ▲ 持平 ► 降低 ▼

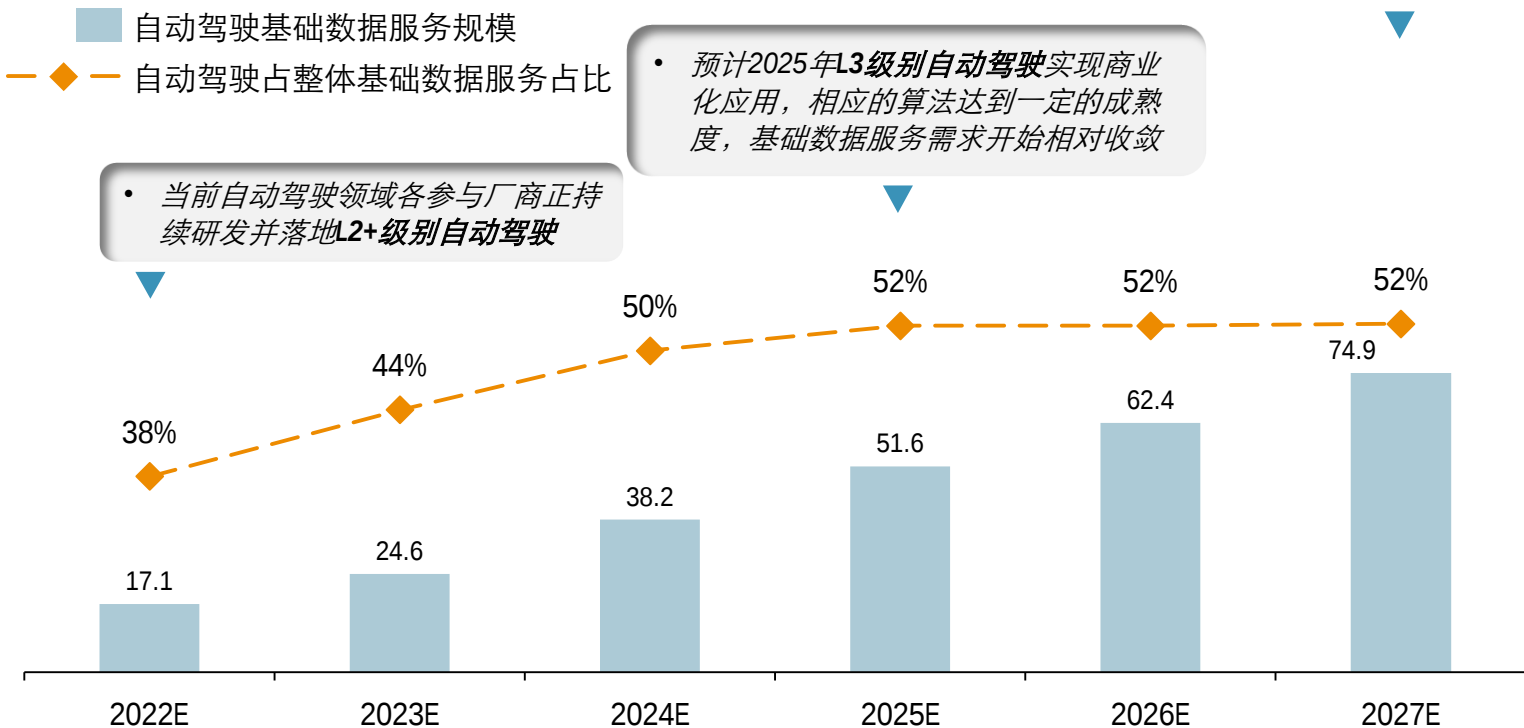


自动驾驶将在未来持续释放数据基础服务需求

自动驾驶AI算法的升级迭代及模型训练数据量的指数级增长，将持续拉动人工智能基础数据服务需求

自动驾驶人工智能基础数据服务市场规模及整体占比

单位：亿元；%



核心驱动因素



持续稳定的“小确幸”

大量整车厂与Tier1开始自研自动驾驶人工智能算法带来新数据需求，同时搭载自动驾驶车型渗透率不断攀升，算法模型的跨车型搭载带来适配需求

自动驾驶技术从目前低级别向高级别的迭代将带来人工智能模型训练数据量指数级的需求增长

技术迭代带来的需求“大爆发”



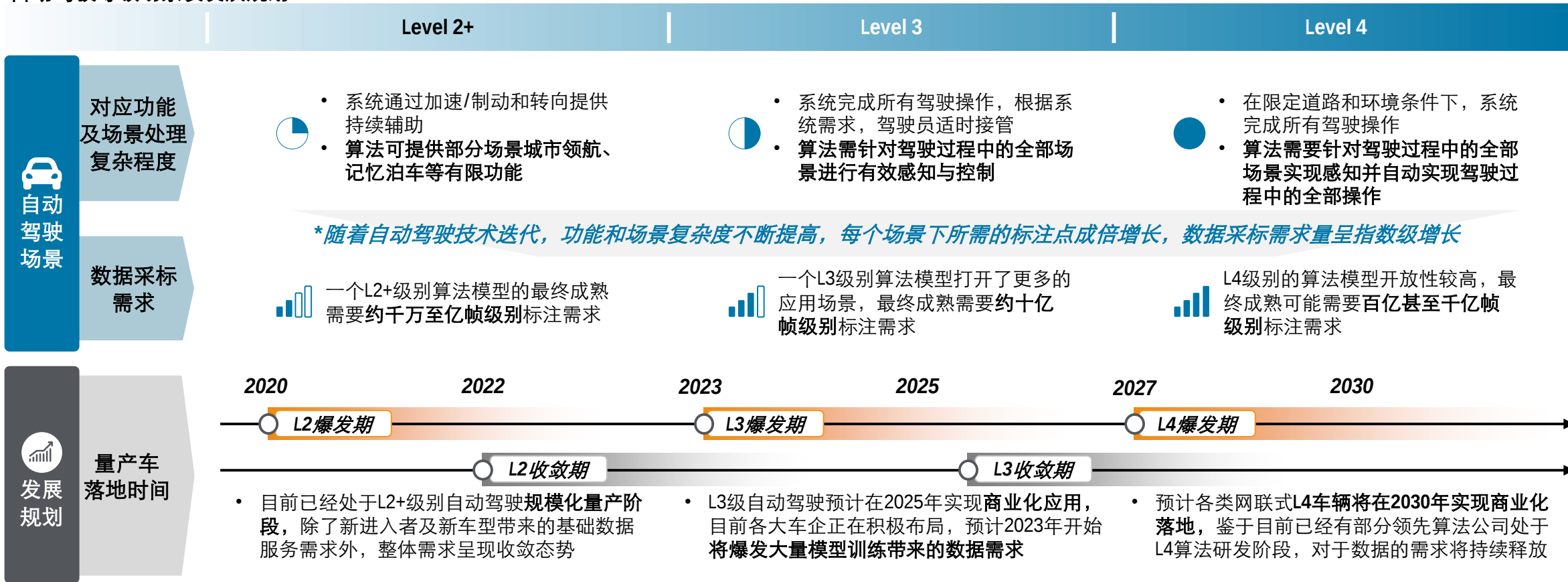
信息来源：德勤访谈、研究与分析



自动驾驶将在未来持续释放数据基础服务需求

目前自动驾驶主要聚焦于L2+级别开发和应用，随着算法趋于成熟，算法开发对于数据的需求量呈周期性收敛趋势，而高级别L3和L4自动驾驶技术场景更为复杂，算法模型训练所需的数据量将逐步呈现指数级上升

自动驾驶等级场景及发展规划



信息来源：《智能网联汽车技术路线图2.0》；德勤访谈、研究与分析

场景处理复杂程度：低   高 数据采集需求：低       高

第二章 AI基础数据服务趋势： 复杂化、自动化、全栈化及合规化

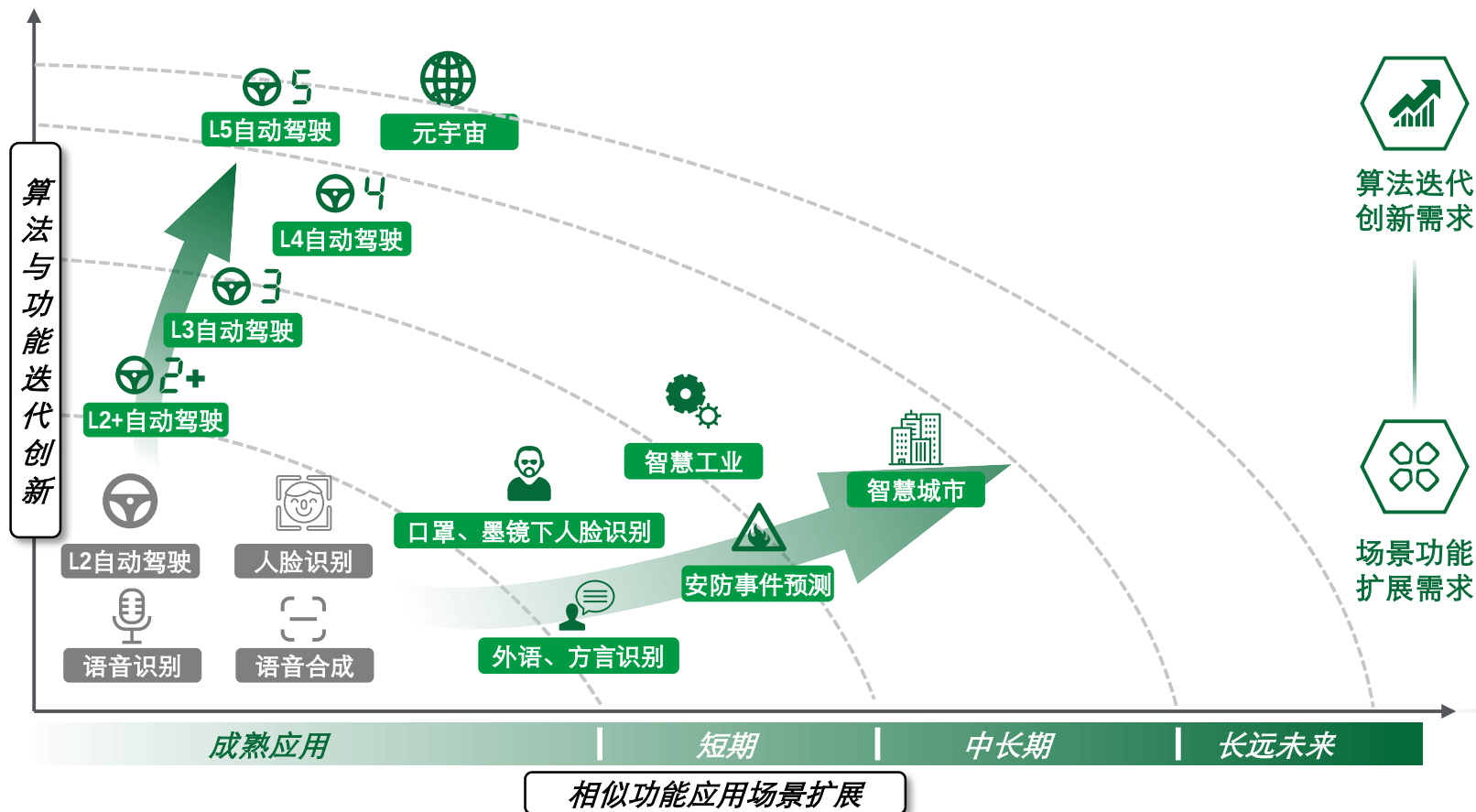




数据基础服务复杂度不断提升

人工智能算法仍处于快速动态演进阶段，随着算法与功能的迭代创新，场景功能的持续扩展，数据标注元素和标注信息维度均将大幅增加，对于数据基础服务供应商提出了更高要求

人工智能发展趋势示意图



算法迭代
创新需求

随着不同场景下的功能不断拓展完善，算法存在迭代创新的需求。以自动驾驶领域为例，随着L2至L4自动驾驶技术的迭代发展需要，相应的算法对于功能性要求愈发提高，对于数据采集与标注的需求也将愈发庞大复杂，需要感知训练评测平台加持模型迭代的效率与精确度。



场景功能
扩展需求

利用AI算法，实现对于同一种或者相似度较高的功能（例如人脸识别）需求不断拓展，对算法进行挖掘需求提炼、规则定义、工具制作、数据处理等工作，深度挖掘高价值数据标签。

信息来源：德勤访谈、研究与分析

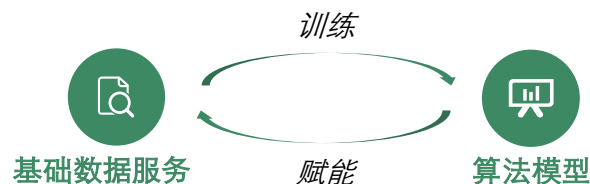


AI自动标注能力推高行业竞争门槛

随着标注量的增大，纯人工标注在成本上不再具有优势，AI赋能的自动标注能力与相关工具逐渐成为基础数据服务商和AI算法公司降本增效的利器，也推高了行业门槛，未来市场集中度有望提升

AI自动标注通过算法模型反哺，基于基础图像识别能力演化而来

- AI自动标注基于通过大量数据训练的算法模型，能够实现对原始数据中需要标注的元素的自动识别、检测以及标注



基础数据服务训练算法的同时算法也能赋能自动标注

借助AI自动标注可大幅提升标注效率

AI自动标注的不同功能模块

AI辅助工具

自动化程度：较低

- 以人工标注为主，标注过程中利用AI能力形成辅助工具帮助实现自动贴边、自动分割等功能，从而提高人工标注效率

AI自动预标注

自动化程度：较高

- 通过AI算法初步生成标注结果，AI标注完成后再通过人工进一步核对和验证

AI自动标注的主要作用

标注效率提升

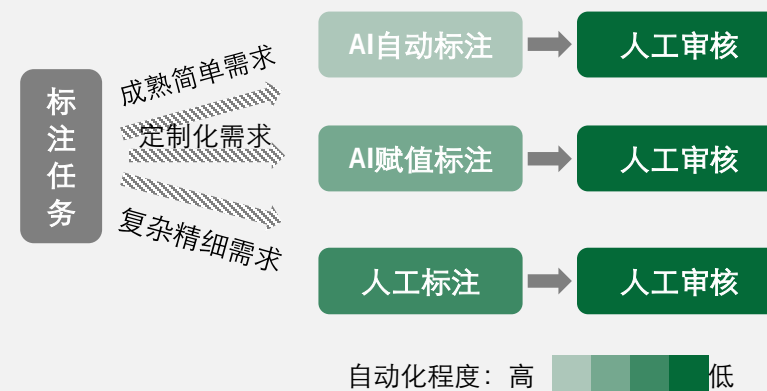
- 使标注更简单高效，帮助提升数据标注速度

标注成本降低

- 提高效率的同时实现人工的部分替代，节省人力成本

AI自动标注无法完全替代人工标注

- AI自动标注仍需要人工审核，且复杂度和精细度较高的需求仍然依赖人工



随着人工智能算法功能的不断迭代演进，自动标注需要持续的训练以及更强的算法能力支撑，行业门槛提高，具备较强法能力和稳定训练资源的数据基础服务供应商以及算法公司将更具优势



全栈式服务需求逐渐显现

随着人工智能技术应用深化，下游AI算法应用方自研人工智能算法的趋势逐渐显现，他们相对算法公司而言，更需要“基础数据服务+云资源+工具链”的全栈式服务

下游应用厂商向上游布局趋势显著

- 人工智能技术应用的逐渐深化促使部分下游人工智能应用厂商在产业链上进行延伸，逐步开始向上游人工智能算法领域布局

示例



智能安防

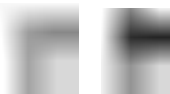


大华技术

- 2015年海康威视成立人工智能研究院
- 2017年大华成立聚焦AI等技术的先进技术研究院



自动驾驶



- 2018年，上汽宣布建立人工智能实验室——上汽ai lab，大举进入自动驾驶领域，加大算法研发投入



智慧金融



- 2016年，招商银行正式提出加快推进金融科技战略，通过旗下招银网络科技加大人工智能的创新应用

新格局下的全栈式服务需求

- 随着下游应用方的下场参与，他们算法能力积累较少，而同时面临算法能力快速部署与迭代的需求，进而催生出了对包括云资源、算法工具链等全栈式工具服务的需求

- 传统算法公司对于基础数据服务的需求聚焦于标注效率、标注质量以及标注成本等基础型需求

- 算法应用方对于数据存储以及算法开发和运行所依赖的云算力资源提出了相应的新需求

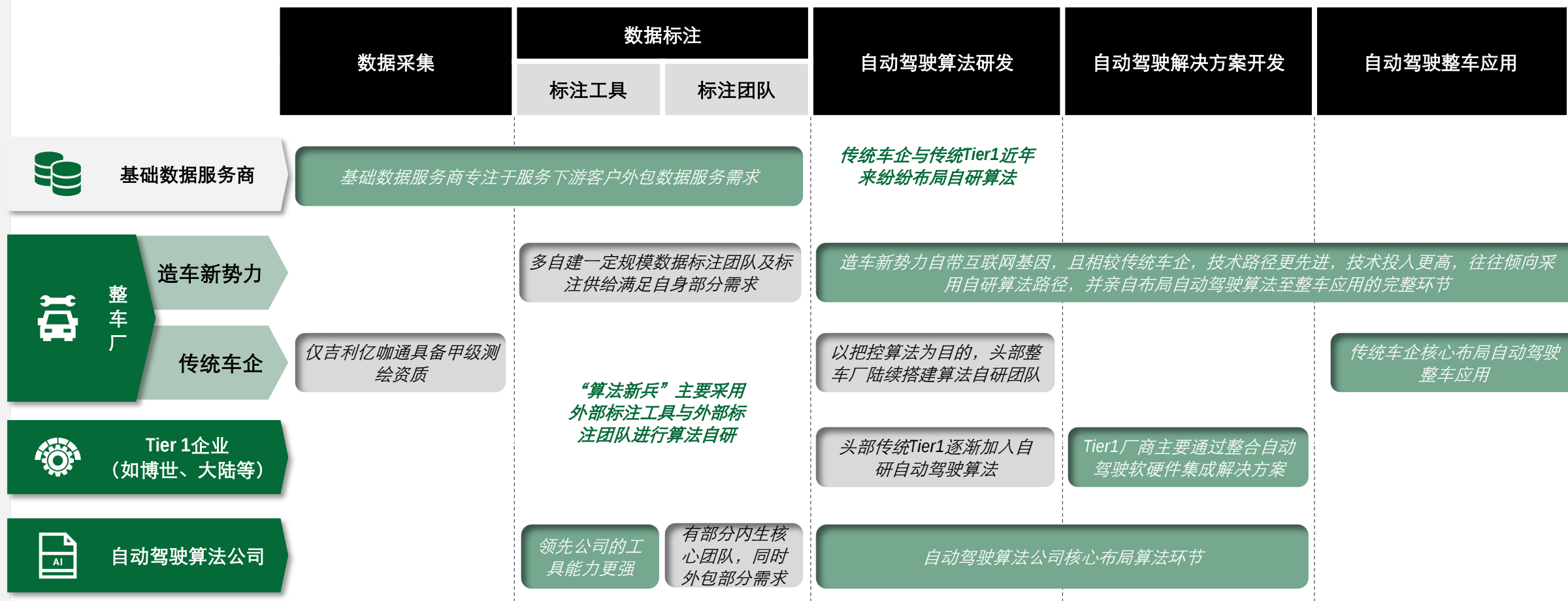
- 需借助数据管理中台、智能标注平台、感知训练平台、仿真平台等成熟的算法工具链产品实现快速部署与持续的快速迭代
- 目前工具链主要应用在自动驾驶领域，未来需求将随着人工智能商业化场景的成熟扩展至各行各业





自动驾驶算法成为“兵家必争之地”¹

自动驾驶板块表现十分明显，算法研发过去由专业算法公司把控，近年来越来越多的整车厂与头部Tier1也开始构建自有算法，尝试掌握自动驾驶的核心环节，由此成为行业的“算法新兵” ...



信息来源：德勤访谈、研究与分析

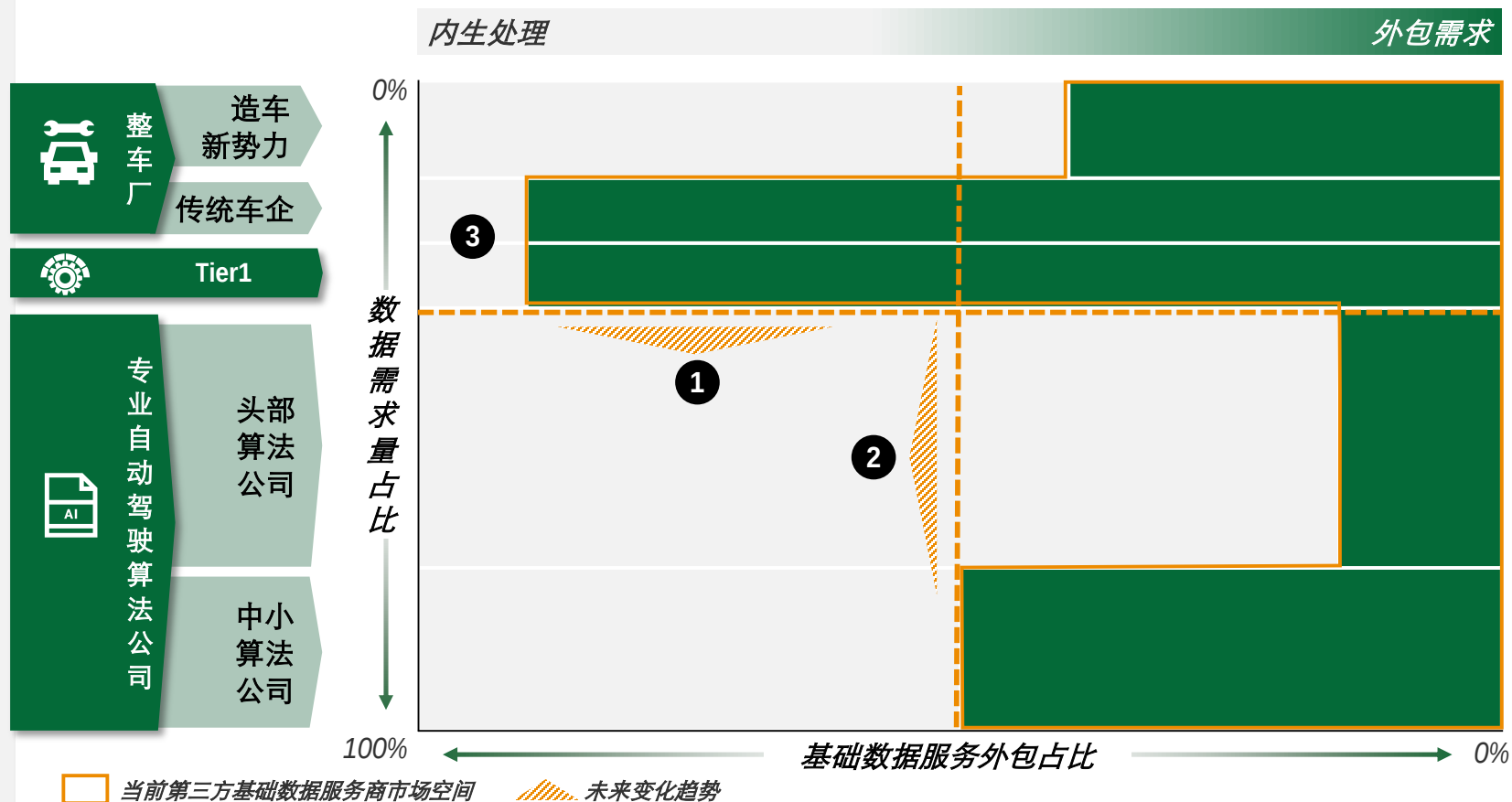
核心布局环节 延伸布局环节



“算法新兵” 释放全栈式需求

“算法新兵” 将在未来释放出大量基础数据服务需求的外包需求，同时随着整车厂及Tier1供应商对于自动驾驶算法自主研发能力的深化，也将释放更多工具链使用需求

自动驾驶基础数据服务不同下游客户数据处理需求量占比示意图



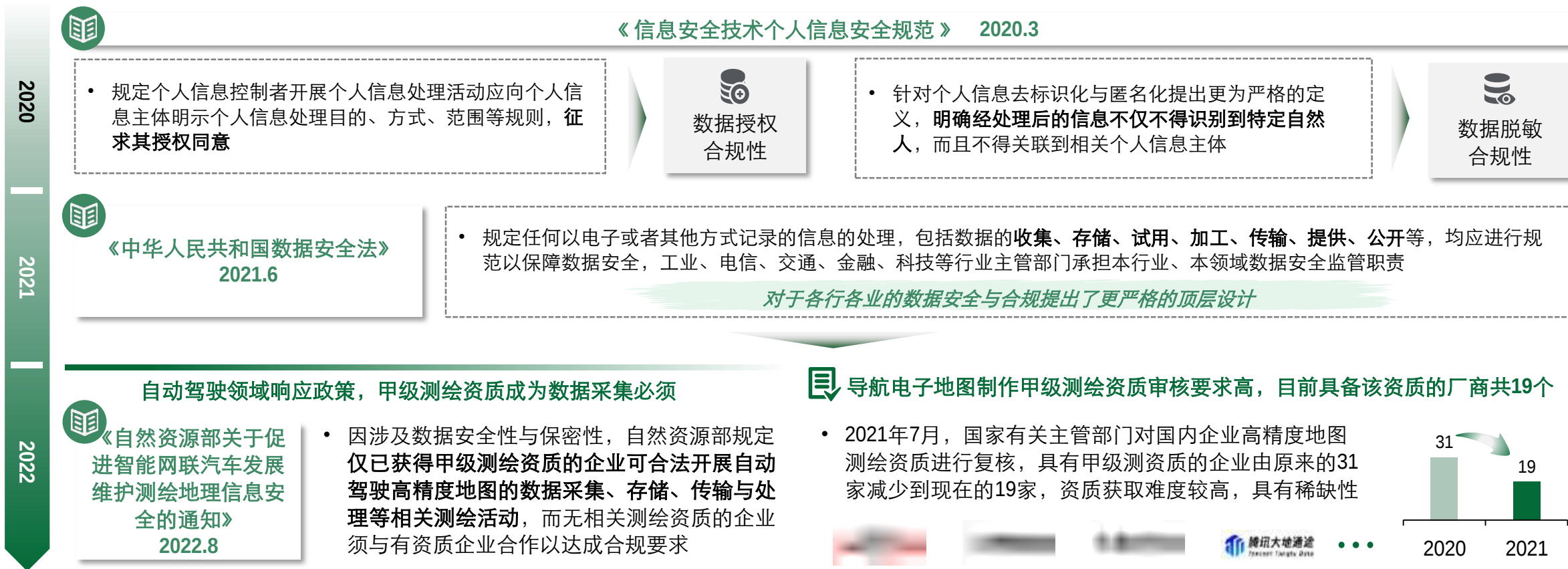
信息来源：德勤访谈、研究与分析



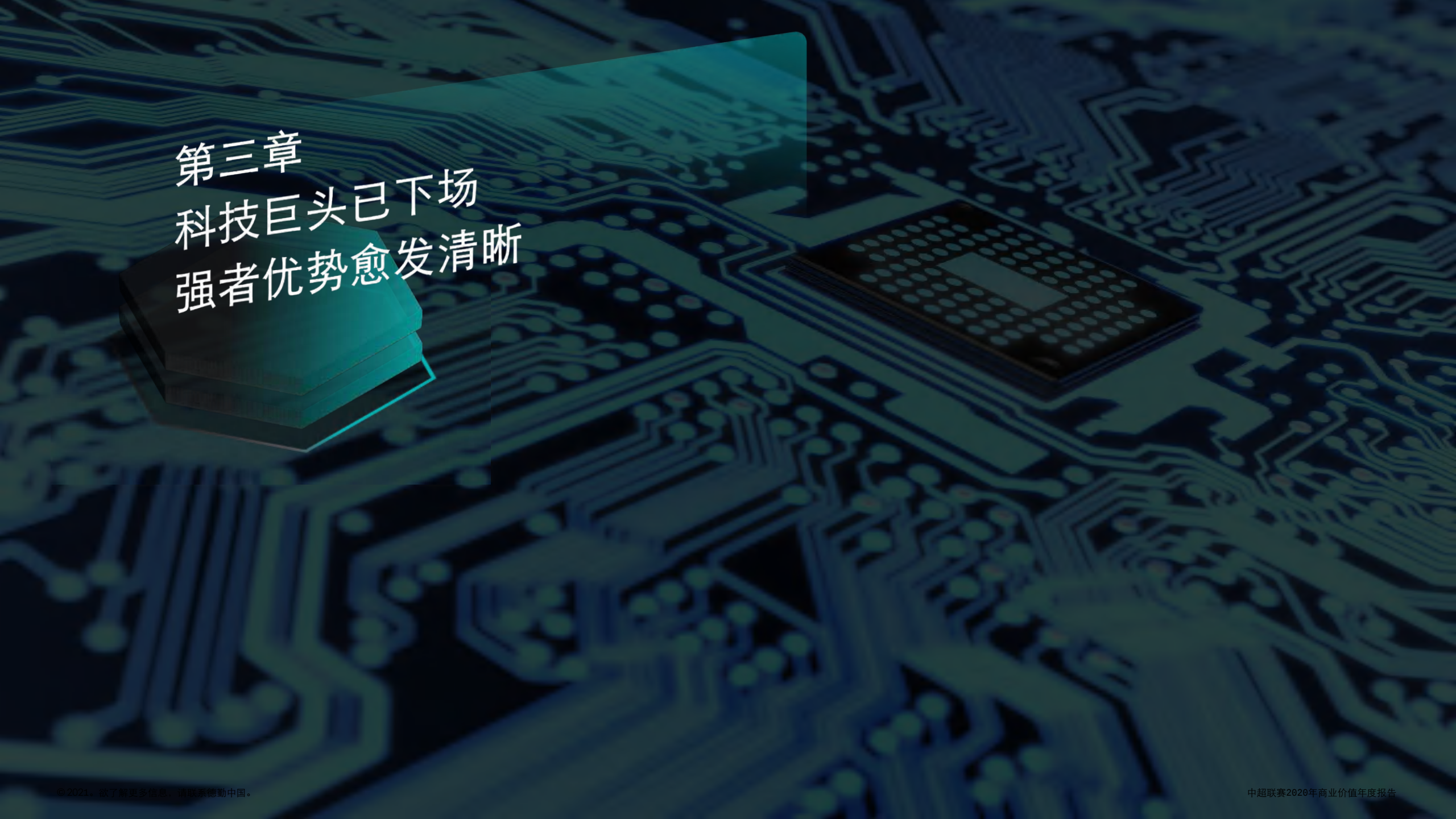
数据合规趋严，专业性价值凸显¹

随着数据安全相关法律法规体系的完善，数据合规要求愈发严格，基础数据服务商在数据脱敏处理、数据采集的测绘资质要求等环节的专业性价值优势凸显

数据合规相关法律法规及标准体系



信息来源：《民法典》、《信息安全技术个人信息安全规范》、《中华人民共和国数据安全法》、《自然资源部关于促进智能网联汽车发展维护测绘地理信息安全的通知》；德勤访谈、研究与分析



第三章 科技巨头已下场 强者优势愈发清晰

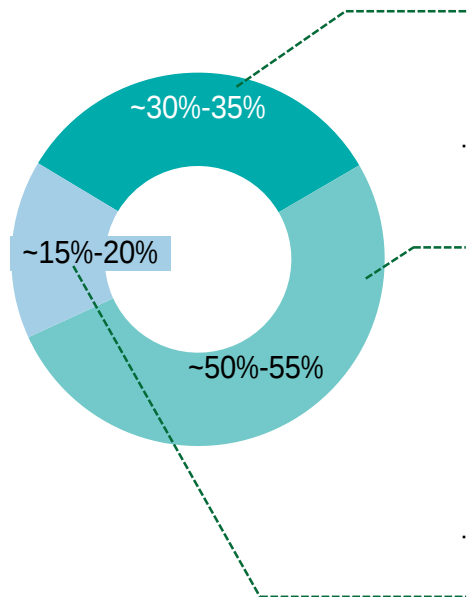


科技巨头下场，行业竞争格局正在重塑

人工智能基础数据服务行业有三类参与企业：以百度为代表的科技巨头，核心优势在于资源整合和研发能力；以数据堂为代表的专业型服务商，众包或人力外包模式起家；以曼孚科技为代表的科创公司，基于算法研发能力，以自动标注等标注工具切入市场。从市场份额看，百度位居行业第一，市占率达到17-18%

人工智能基础数据服务商市场份额占比及核心能力分析

基础数据服务整体格局（2022）



科技巨头公司

百度智能云数据众包

阿里众包

京东众智

腾讯数据厨房



专业型基础数据服务商



品牌化专业服务商

数据堂

海天瑞声

澳鹏

龙猫数据



中小型人力外包企业

以山东、山西、河南、贵州等地为主，人力资源充沛，人力成本较低，具有大量数据基础服务人力外包企业



科技初创企业

MindFlow

Stardust

BodenAI

SurfingTech

信息来源：德勤访谈、研究与分析



顺应行业“四化”趋势，头部企业竞争优势愈加明显¹

顺应标注复杂化、标注自动化、全栈式服务化以及数据合规化的“四化”趋势，科技巨头依托长期的技术和资源积累，综合能力更强，竞争优势突出

	百度智能云数据众包	阿里众包	京东众智	澳鹏	海天瑞声	数据堂
 自动化标注能力	强大的AI算法能力支撑其自动标注以及场景筛选等辅助标注能力	自动标注算法能力较强，主要部署在云端，一般通过打包云服务方式销售	具备一定的AI预标注能力，但整体AI算法能力相对百度和阿里稍弱	缺乏强AI大模型能力支撑，但基于海外市场20余年行业经验，具备一定的AI自动化标注能力	缺乏强算法能力支撑，自动标注能力较弱，但基于自有标注平台的研发与迭代，在语音、基础图像领域针对相对成熟的应用、具备一定的AI辅助标注能力	
 专业资质及合规性	拥有自动驾驶甲级测绘资质	集团内拥有自动驾驶甲级测绘资质	不具备甲级测绘资质，但也都高度注重数据合规性及数据安全			
 专业化基础数据服务能力	- 背靠百度，强大的产研团队定制标注工具，对算法行业理解深刻 - 全面的法务合规团队 - 与政府合作共建大规模、稳定的数据标注基地	- 对互联网行业具备深刻的数据需求理解 - 主要采取灵活的众包和人力外包模式，暂无自建标注团队	- 对互联网行业具备深刻的数据需求理解 - 千人规模的自建标注团队可保障交付质量	- 具有多年专业化基础数据服务经验，具备较强的服务响应能力 - 拥有无锡与大连两处自建标注中心团队，具备一定规模的内部交付团队	语音领域具有深厚资源，能够提供标注数据集产品	具备专业化标注基地，承接业务种类丰富
 全栈式服务能力	- 基于百度AI基础数据服务能力、云资源、算法能力提供强大的全栈式服务能力 - 拥有相对完善工具链产品，在自动驾驶行业积累丰富，帮助企业自建自研	能够提供基于阿里的包括云资源、工具链等功能集成的全栈式服务	云资源相对稍弱，全栈式服务能力不及百度和阿里	缺乏大型科技企业技术能力资源背书，主要专注于数据采集标注服务，全栈式服务能力弱，但具备一定的工具能力，可以协助客户进行数据标注工具的开发		

信息来源：德勤访谈、研究与分析



标杆案例：百度数据众包



百度数据众包拥有行业最大的自建标注团队，同时依托百度深耕人工智能多年的行业理解与经验，实现了数据标注能力与全流程业务能力的相互反哺，构建数据闭环...

庞大的自建标注团队为数据服务能力提供强力保障

- AI数据处理具有高并发的需求特征，另外随着算法迭代升级和场景功能的拓展，稳定的自建标注基地带来的专业性将得以凸显，为高效、高质量处理客户需求提供保障
- 百度数据众包在山东济南、山西临汾、重庆奉节、四川达州、甘肃酒泉、江西新余等10个地区有自建标注基地，可**全力保障数据标注服务的高质量交付**



山东



山西



甘肃



四川



江西

...



人员
规模

- 2-3万稳定签约标注人员



标注
能力

- 图像框选100万框/天
- 图像分割8000帧/天
- 点云框选70万框/天

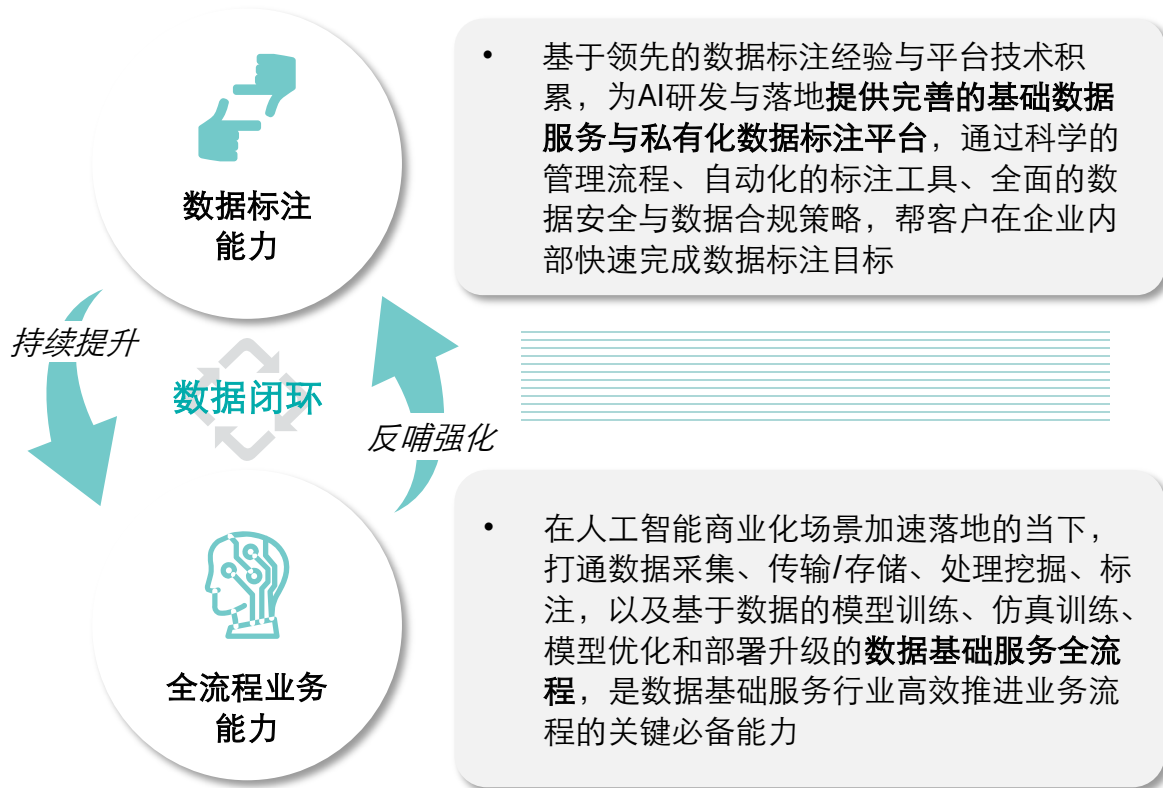


质量
保障

- 为基地标注人员提供综合标注技能培训
- 具有完善的人员考核机制和标注质检流程

百度强大的数据标注能力与全流程业务能力助力实现数据闭环

不断强化的数据闭环正循环



信息来源：公司官网、公开资料整理；德勤访谈、研究与分析

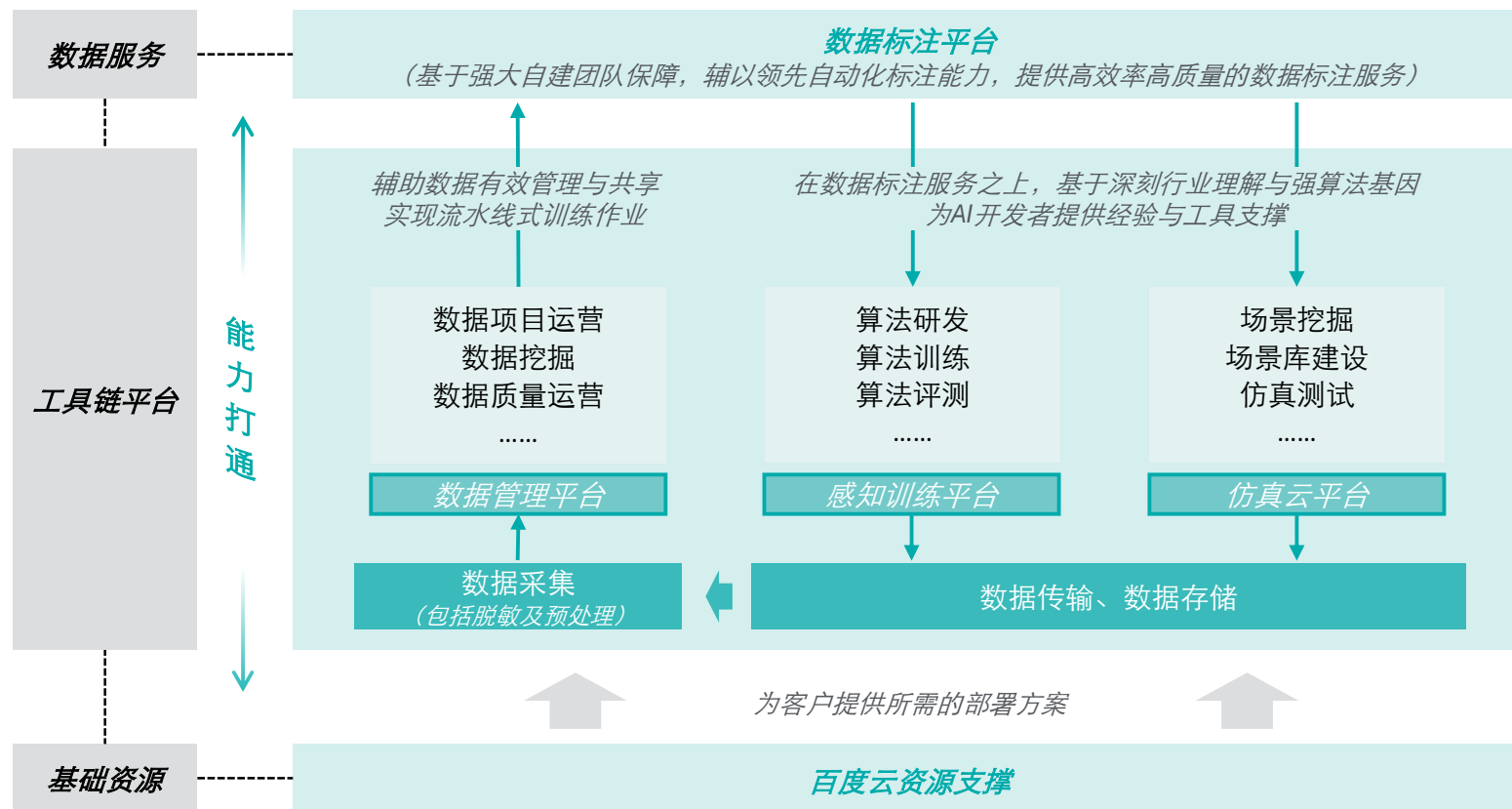


标杆案例：百度数据众包



...百度数据众包依托自身较强的技术能力及规模化的专业标注团队，形成了强大的基础数据全栈式服务能力，提供数据管理、智能标注、感知训练评测、仿真测试服务的全面赋能，即可提供整套解决方案也可解耦使用

百度强大的资源及能力支撑造就全栈式综合服务能力



- 百度数据众包基于其丰富的数据处理与管理经验，将数据采标服务、AI算法开发者数据平台以及底层的云资源打通，通过数据管理、智能标注、感知训练评测、仿真测试服务的全面赋能，百度数据众包将帮助AI开发者低成本高效构建人工智能算法技术研发、运营、商业化落地的全场景闭环，不仅能够基于自身平台交付高质量的数据采集与标注服务，同时更能依托百度自身在AI算法开发方面的丰富经验为使用方的AI算法快速落地提供支持

信息来源：公司官网；德勤访谈、研究与分析

联系我们

白皮书编写团队

林国恩
德勤中国
科技、传媒和电信行业主管合伙人
电子邮件: talam@deloitte.com.cn

廉勋晓
德勤中国
科技行业主管合伙人
电子邮件: mlian@deloitte.com.cn

濮清璐
德勤中国
科技、传媒和电信行业华东区主管合伙人
电子邮件: qlpu@deloitte.com.cn

濮清璐
德勤中国商业战略与研究
合伙人

胡佳男
德勤中国商业战略与研究
经理

潘乐晨
德勤中国商业战略与研究
咨询顾问

王小璐
德勤中国商业战略与研究
总监

李嘉程
德勤中国商业战略与研究
高级咨询顾问

黄鹭
德勤中国商业战略与研究
咨询顾问



关于德勤中国

德勤中国是一家立足本土、连接全球的综合性专业服务机构，由德勤中国的合伙人共同拥有，始终服务于中国改革开放和经济建设的前沿。我们的办公室遍布中国30个城市，现有超过2万名专业人士，向客户提供审计及鉴证、管理咨询、财务咨询、风险咨询、税务与商务咨询等全球领先的一站式专业服务。

我们诚信为本，坚守质量，勇于创新，以卓越的专业能力、丰富的行业洞察和智慧的技术解决方案，助力各行各业的客户与合作伙伴把握机遇，应对挑战，实现世界一流的高质量发展目标。

德勤品牌始于1845年，其中文名称“德勤”于1978年起用，寓意“敬德修业，业精于勤”。德勤专业网络的成员机构遍布150多个国家或地区，以“因我不同，成就不凡”为宗旨，为资本市场增强公众信任，为客户转型升级赋能，为更繁荣的经济、更公平的社会和可持续的世界而开拓前行。

关于德勤

Deloitte（“德勤”）泛指一家或多家德勤有限公司，以及其全球成员所网络和它们的关联机构（统称为“德勤组织”）。德勤有限公司（又称“德勤全球”）及其每一家成员所和它们的关联机构均为具有独立法律地位的法律实体，相互之间不因第三方而承担任何责任或约束对方。德勤有限公司及其每一家成员所和它们的关联机构仅对自身行为承担责任，而对相互的行为不承担任何法律责任。德勤有限公司并不向客户提供服务。

德勤亚太有限公司（即一家担保有限公司）是德勤有限公司的成员所。德勤亚太有限公司的每一家成员及其关联机构均为具有独立法律地位的法律实体，在亚太地区超过100座城市提供专业服务。

请参阅 <http://www.deloitte.com/cn/about> 了解更多信息。

免责声明

本通讯中所含内容乃一般性信息，任何德勤有限公司、其全球成员所网络或它们的关联机构（统称为“德勤组织”）并不因此构成提供任何专业建议或服务。在作出任何可能影响您的财务或业务的决策或采取任何相关行动前，您应咨询合格的专业顾问。

我们并未对本通讯所含信息的准确性或完整性作出任何（明示或暗示）陈述、保证或承诺。任何德勤有限公司、其成员所、关联机构、员工或代理方均不对任何方因使用本通讯而直接或间接导致的任何损失或损害承担责任。德勤有限公司及其每一家成员所和它们的关联机构均为具有独立法律地位的法律实体。

© 2023。欲了解更多信息，请联系德勤中国。