击破业务落地要害

中国面向人工智能的数据治理行业研究报告

摘要

前言-数据与数据治理：如今数据不再局限于传统数字形式的认知，由结构化数据延伸到半结构化、非结构化的数据范畴。数据治理越来越受到企业的普遍重视，在数据生命周期的各个阶段通过相应的工具与方法论，使数据发挥出更大的价值，是实现数据服务与应用必不可少的阶段。

主题-面向人工智能的数据治理：AI技术创新应用走向大规模落地，带动了大数据智能市场的蓬勃发展。2021年大数据智能市场规模约为553亿元。目前传统数据治理体系多停留在结构性数据化治理工作，尚难满足AI应用对数据的高质量要求。企业可吸收传统体系的智慧沉淀，以AI应用数据需求为核心，优化建设“面向人工智能的数据治理”体系，显著提升AI应用的规模化落地效果。

参与-行业规模与受益圈立足点：数据治理与AI应用产品开始交汇融合，厂商参与更加多元，咨询公司、数据服务提供商和人工智能产品服务商三方阵营构建行业竞合格局，而“智”，即AI应用，为面向人工智能的数据治理服务的核心立足点。2021年面向人工智能的数据治理市场规模约为40亿元，预计五年后规模将突破百亿。

实践-高频高价值应用及数据痛点：本篇报告选择金融、零售、医疗和工业四大典型行业为切入点，分析呈现各行业的信息化建设阶段与高频高价值的AI应用场景，并基于高频高价值AI应用引发的数据治理需求，对面向人工智能的数据治理体系搭建给到建设指导。

展望-治理陷阱与趋势洞察：1）企业需避免落入“数据埋点大而全”的治理陷阱；2）供需两侧需共同保证数据治理体系建设后的运营流转；3）企业需建立符合管理现状及发展需求的数据安全治理框架，确保数据全周期的安全与合规；4）联邦学习技术可带来数据安全合规线内的共同富裕；5）数据的“自治与自我进化”成为未来数据处理发展的必由之路，为企业打造“治理+AI”体系的良性循环。

来源：艾瑞研究院自主研究绘制。

前言：数据与数据治理	1
主题：面向人工智能的数据治理	2
参与：行业规模与受益圈立足点	3
实践：高频高价值应用及数据痛点	4
案例：标杆企业与新锐势力	5
展望：治理陷阱与趋势洞察	6

数据：范围界定

信息经济的“货币”，早已不限于数字形式

数据的价值被不断认可，“数据资产化”已经成为了企业发展的重要组成部分。长期以来，数据被理解为以数字形式存储的信息，而目前技术可以测量更多的事件和活动，人们可以收集、存储并分析这些不被视为传统数据的各类信息，如邮件、图片、音视频等。数据可根据其特性及治理方法差异划分为内部数据与外部数据，结构化数据、非结构化数据与半结构化数据，元数据与主数据等。

企业数据的主要类型

分类标准	数据类型	定义以及特征	举例
数据治理常用数据类型	元数据	是描述数据的数据（描述性标签），描述了数据（如数据元素、数据模型）、相关概念（如业务流程、应用系统、软件代码、技术架构）以及他们之间的联系	实体型组织、客户、人员基本配置
	主数据	描述企业核心实体的一组一致而统一的标识符和拓展属性，实体可包括现有或潜在客户、产品、服务、员工、供应商、提供商、层次结构和会计科目表等	数据标准、业务术语、指标定义
	实时数据	是在收集后立即传递的信息，所提供信息的及时性没有延迟	实时OLAP场景下的数据
按照数据格式分类	结构化数据	可以存储在传统的关系型数据库中，用二维表结构来表达实现的数据，可以用关系型数据库存储	Excel表格、 SQL数据库里的数据
	非结构化数据	形式相对不固定，不方便用数据库二维逻辑表来表现的数据，通常存储在非关系型数据库中，数据量通常较大	文本、图片、HTML、各类报表和音频、视频
	半结构化数据	介于结构化与非结构化之间，半结构化数据可以通过灵活的键值调整获取相应信息，且数据的格式不固定	日志文件、XML文档、JSON文档、Email等
按照数据来源分类	企业内部数据	在企业内部经营中产生的数据，在企业的业务流程中产生或在业务管理规定中定义，受企业经营影响	国家、币种、汇率
	企业外部数据	企业通过公共领域合规获得的数据，其产生、修改不受公司影响	合同、项目、组织

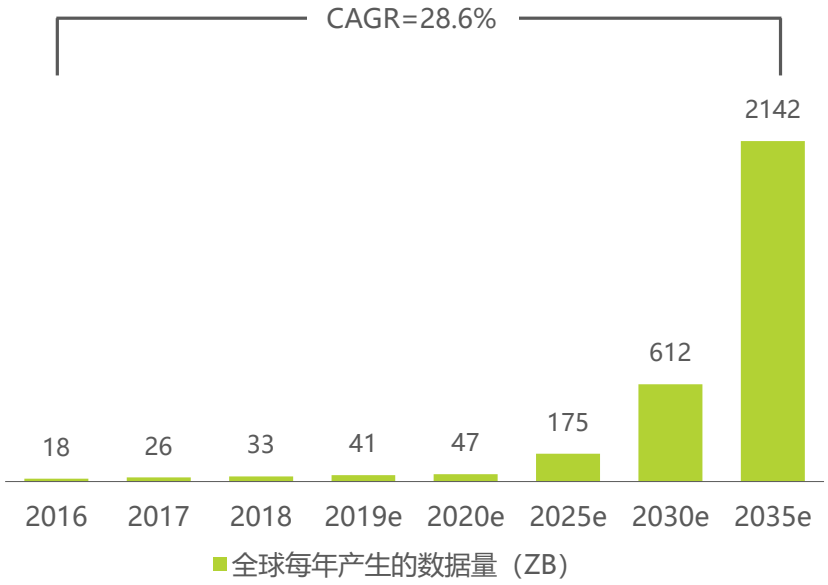
来源：艾瑞研究院自主研究绘制。

数据量：爆发式增长

基础设施“扩容”、IoT 广泛连接带来的数据量暴涨

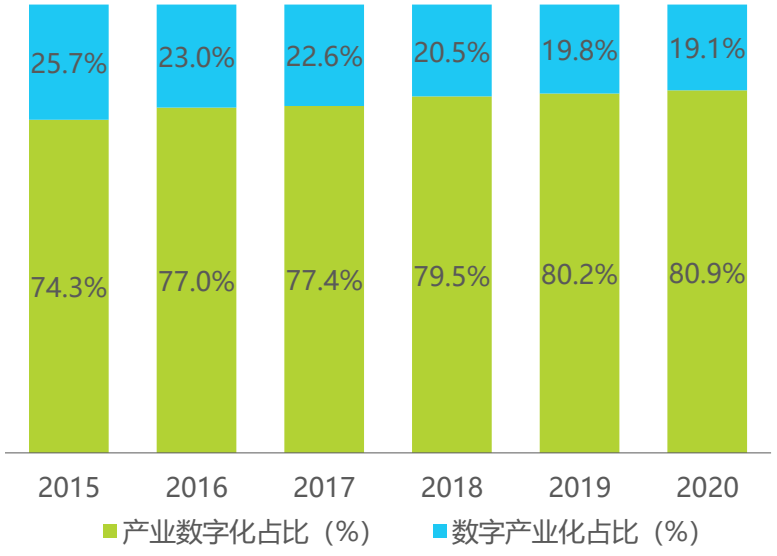
数据时代来临，数据量的暴涨为企业数字化提供了基础支撑，大量的业务数据能够被采集、存储并最终创造经济效益。数字化转型从头部企业的可选项，转变为更广泛企业的必选项。新变化为企业带来新机遇的同时，也带来了诸多挑战。很多企业在前期的信息化建设中，缺乏统筹规划，为解决当下业务问题而按照垂直的、个性化的业务逻辑独立采购与部署IT系统，导致企业内部形成多个数据孤岛。数据不规范、不一致、难以互联互通成为普遍问题，阻碍企业去充分发挥数据价值。这种先建设后治理的常态，使得数据治理越来越受到企业的普遍重视，另一方面，新兴技术与应用场景的快速落地，也带领数据治理需求在加速攀升。

2016-2035年全球产生的数据量



注释：1ZB = 1024^4 GB
来源：中国信通院，Statista（2020），艾瑞研究院自主研究绘制。

2015-2020年中国数字经济内部结构变化



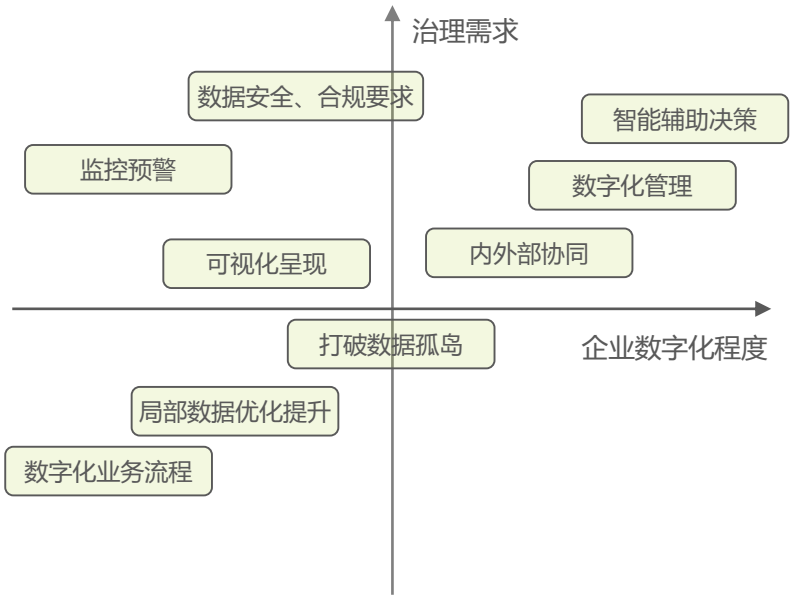
来源：中国信通院，艾瑞研究院根据专家访谈与公开资料研究绘制。

数据治理：需求释放

治理需求普遍存在，非结构化数据成为价值挖掘的重难点

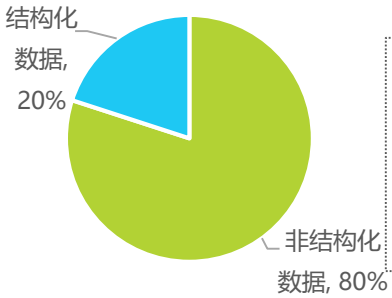
企业历经数字化转型不同阶段时，需通过数据治理解决数据在生产、管理和使用中的问题，而数据治理的需求与复杂度也会随着企业数字化程度提升而增加。从企业内部的数据类型来看，非结构化数据占企业内数据总量的80%，却仅占整体使用率的30%，长期以来其价值未得到充分有效利用。未来，随着非结构化数据的积累增加与AI应用的数据需求推动，企业对非结构化数据的价值化需求将加速释放，而多源异构数据基础下的数据治理模块也将获得进一步的关注与优化。

不同企业数字化程度下的主要数据治理需求



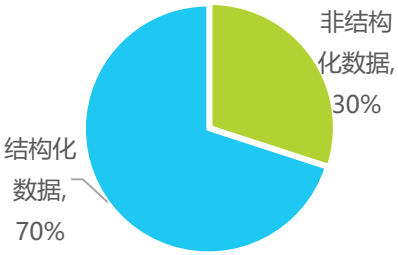
注释：仅列举代表性数据治理需求。
来源：艾瑞研究院自主研究绘制。

企业内结构化数据与非结构化数据占比情况



在企业的数据中，结构化数据仅占20%，其余80%都是以**文件、语音、图片等形式**存在的非结构化数据。且**非结构化数据的增速远远高于结构化数据**，随着时间的推移，非结构化数据所占的比例将会越来越高

企业内结构化数据与非结构化使用现状



企业长期以来，受技术影响，对结构化数据的利用率均高于对非结构化数据的利用率。但实际上，**非结构化数据的体量与其包含的信息量都更多**，是企业未得到充分利用的宝贵资产

来源：艾瑞研究院根据专家访谈自主研究绘制。

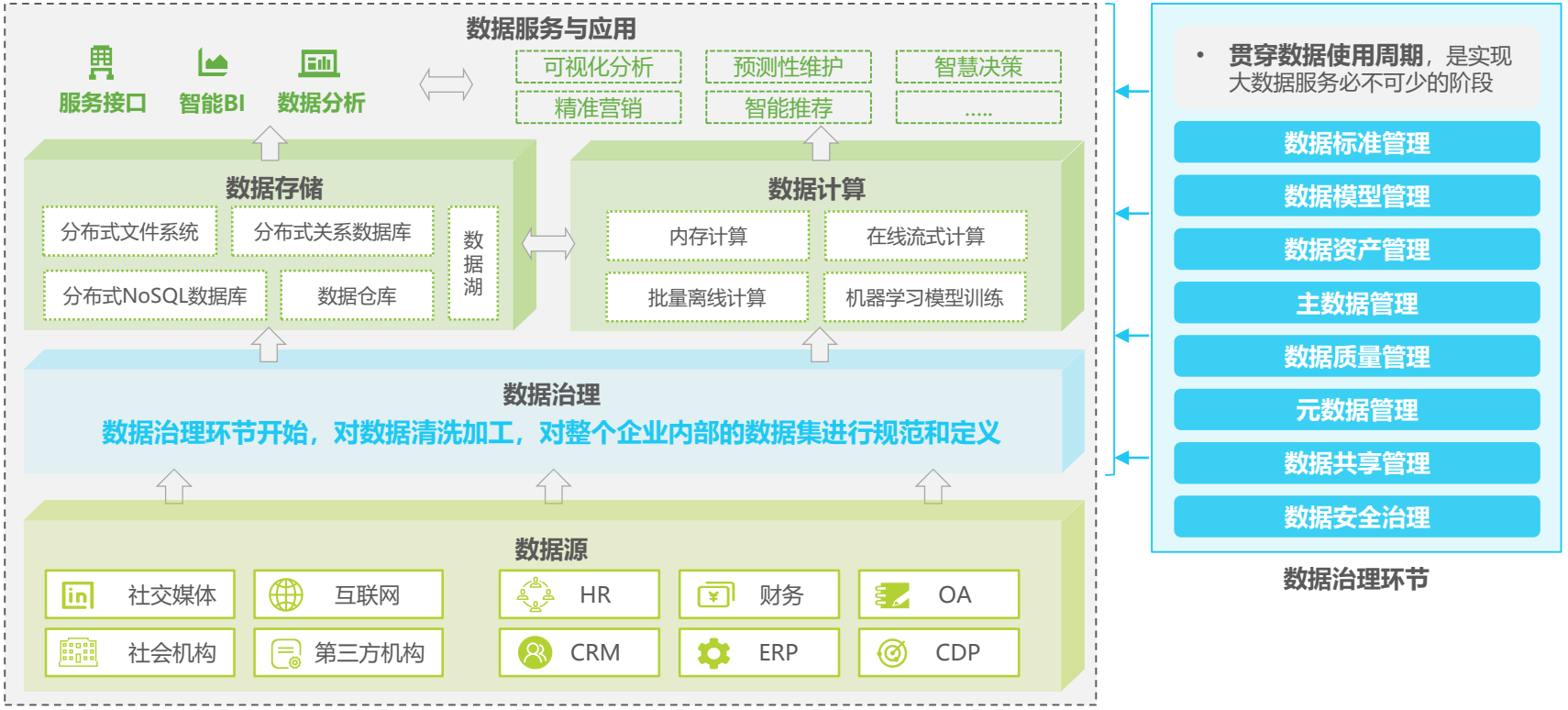
©2022.3 iResearch Inc.

数据治理：范围界定

数据治理为实现企业数据应用服务的重要环节

数据治理以数据源汇入为伊始，对数据进行清洗加工，并在数据存储、数据计算、数据服务应用等环节予以持续的治理服务，是企业实现数据服务与应用的重要环节。从数据层面来看，数据本身存在着从生产到消亡的生命周期，而数据治理会在数据生命周期的各阶段通过相应工具与方法论进行规范与定义，在企业内部构建出切实有效的数据闭环，使数据发挥出更大的价值。

数据治理在数据应用流程中的位置



来源：艾瑞研究院自主研究绘制。

数据治理：整体概述

让数据可知、可用、可管，成为业务发展与创新的基石

数据治理旨在消除数据的不一致性，建立规范的数据标准，提高组织的数据质量与实现数据广泛共享，最终将数据变为宝贵资产，应用于企业的经营、管理与决策中。当下，让数据可知、可用、可管，充分发挥数据资产的价值已成为企业共同的数据治理目标。数据治理的对象与范围则会根据企业需求差异而有所区别。在不断发展变化的外部环境与业务需求下，企业数据治理工作在对应阶段也会有各自不同的目标。

数据治理的对象、目的与范围概述

数据治理的对象

- **找到企业可变为的“数据资产”**：数据治理范围并非为企业全部数据，而是要在企业海量数据中找到“值得”治理的数据范围，将其变为可用宝贵的“数据资产”，为企业进一步发挥数据要素价值。

数据治理的目的

- **提升数据使用价值**：在实践中，企业发现原始数据或只经过初步处理的数据，与价值化数据之间存在巨大鸿沟，需经由数据治理做对应的清洗、规范及定义等，以提升数据使用价值。

数据治理的范围

- **贯穿数据生命周期**：数据治理是贯穿整个数据生命周期，复杂且需要长期建设的项目。对不同企业而言，业务需求千差万别，聚焦于核心数据问题、结合企业特点选取合适的数据范围，方能把控好治理方向。

- 从企业的数据使用现状来看，集中于对**结构化数据的开发与利用**，所以数据治理工作多围绕于结构化数据的治理，非结构化数据仅做入库、入湖等初步处理，利用率并不高。

- **缺乏技术手段、缺乏方法指导、缺乏保障机制、缺乏流程规范**的等是大多数企业无法解决数据价值化问题的主要原因。

- 大部分企业都有**明确的数据治理目的**，供应商仅需要围绕企业需求的模型及模型效果来确定需要治理的数据源，在其中，充分了解企业需求与现状是必要程序。

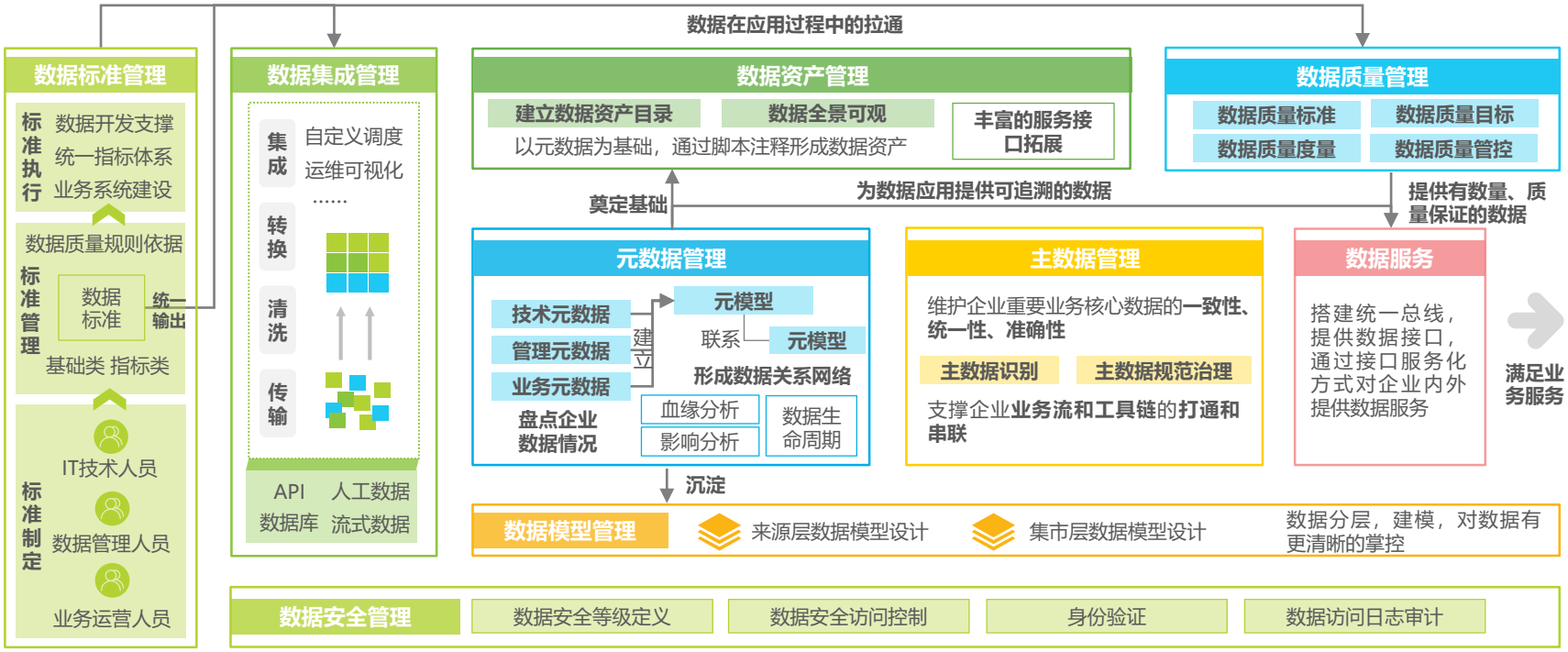
来源：艾瑞研究院自主研究绘制。

数据治理：体系架构

结合企业的特点及需求，设计符合企业要求的数据治理架构

虽然业界对数据治理的定义不尽相同，但涉及的数据架构模块大体一致，核心包括数据标准管理、数据集成管理、元数据管理、主数据管理、数据资产管理、数据质量管理、数据模型管理、数据服务与数据安全管理模块。依托于企业对数据治理的侧重点不同，数据治理体系与架构也会根据企业所在的行业特点、经营性质及信息化程度的不同而有所差异。在实际设计时，一方面，企业可参考先进体系框架与行业最佳实践，另一方面，企业也需从实际需求与发展需要出发，设计搭建适合自身情况的数据治理架构。

数据治理各模块内容以及相互之间的关系



来源：艾瑞研究院自主研究绘制。

数据治理：政策指引

推动各行业数据治理标准建设，为相关主体提供指引性文件

近年来，我国政府从战略规划、体系建设、标准制定和制度落地四个方面，全力推动数据治理的行业规范发展。一方面，国家通过立法构建数据安全保障、明确数据安全法律责任、完善监管体系；另一方面，各地方政府、行业主管部门、各行业组织、标准化机构积极规划制定数据规范文件与鼓励政策，推进数据治理考核、评估标准建立，为相关数据治理项目主体提供指引，共同促进数据治理行业的发展。

中国数据治理相关政策梳理与解读

战略规划

《关于构建更加完善的要素市场化配置体制机制的意见》
2020年4月10日 国务院

- 强调要加快培育数据要素市场，推进政府数据开放共享，提升社会数据资源价值。培育数字经济新产业、新业态和新模式，支持构建工业、安防等领域规范化数据开发利用的场景。加强数据资源整合和安全保护。**探索建立统一规范的数据管理制度，提高数据质量和规范性**，丰富数据产品

《促进大数据发展行动纲要》
2015年8月31日 国务院

- 建立标准规范体系，推进**关键共性标准的制定和实施**，开展标准验证和应用试点示范，**建立标准符合性评价体系**

地方政府、行业组织、标准化机构陆续发布数据规范文件与鼓励政策

2018年3月15日｜国家标准化管理委员会
国家标准《数据管理能力成熟度评估模型》（DCMM）

2019年1月1日｜国家标准化管理委员会
《信息技术数据质量评价指标》

2020年2月27日｜工业和信息化部办公厅
《工业数据分类分级指南（试行）》

2021年5月31日｜深圳市人大常委会办公厅
《深圳经济特区数据条例（征求意见稿）》

行业主管部门探索制定和出台数据治理相关要求、标准、框架与体系

2016年9月5日｜国务院
《政务信息资源共享管理暂行办法》

2018年5月21日｜中国银行业监督管理委员会
《银行业金融机构数据治理指引》

2018年5月25日｜民政部
《关于加强和完善民政统计工作 全面提高统计数据真实性的实施意见》

2019年9月29日｜中国银行保险监督管理委员会
《银行业金融机构监管数据标准化规范》

构建数据安全保障，明确数据安全法律责任，完善监管体系

2021年6月10日｜全国人大会常务委员会
《数据安全法》

2021年11月1日｜全国人大会常务委员会
《个人信息保护法》

2019年5月28日｜国家互联网信息办公室
《数据安全管理办法》

2020年3月6日｜信息安全标准化技术委员会
《个人信息安全规范》

来源：艾瑞研究院自主研究绘制。

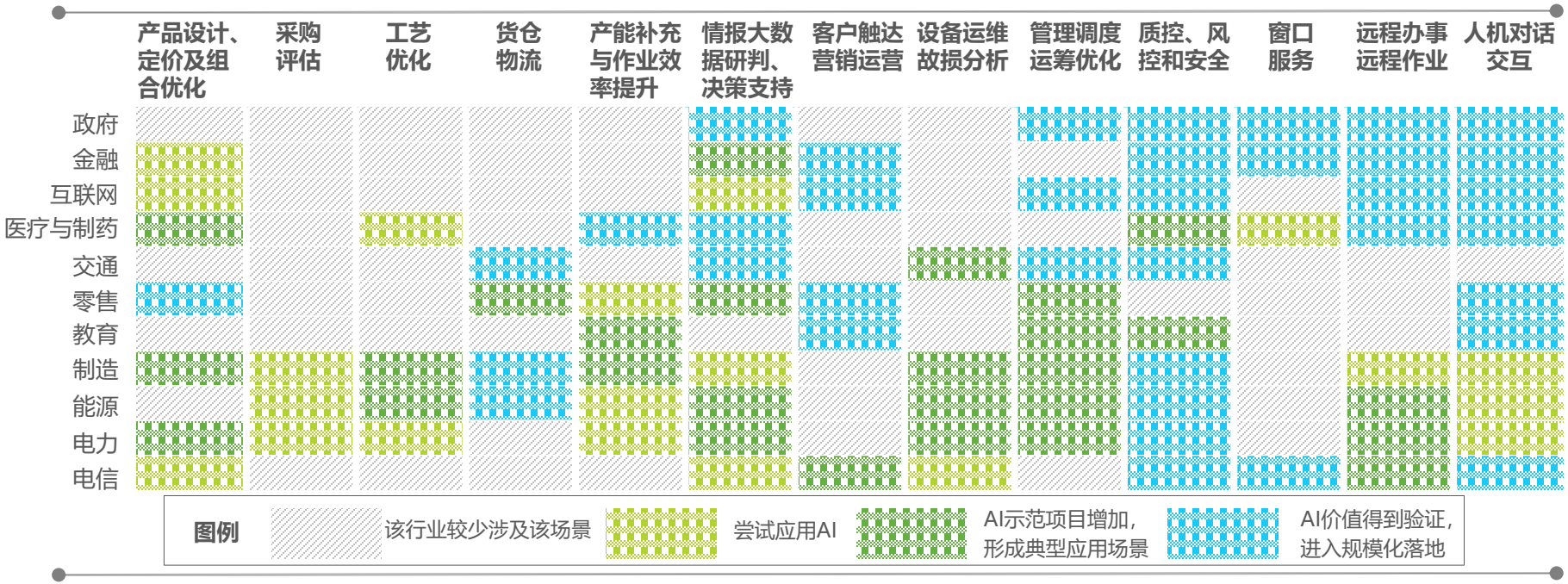
前言：数据与数据治理	1
主题：面向人工智能的数据治理	2
参与：行业规模与受益圈立足点	3
实践：高频高价值应用及数据痛点	4
案例：标杆企业与新锐势力	5
展望：治理陷阱与趋势洞察	6

AI应用规模化

AI技术创新应用大规模落地，带动大数据智能市场蓬勃发展

近年来，随着新技术模型出现、各行业应用场景价值打磨与海量数据积累下的产品效果提升，人工智能应用已从消费、互联网等泛C端领域，向制造、能源、电力等传统行业辐射。各行业企业在设计、采购、生产、管理、营销等经济生产活动主要环节的人工智能技术与应用成熟度在不断提升，加速人工智能在各环节的落地覆盖，逐渐将其与主营业务相结合，以实现产业地位提高或经营效益优化，进一步扩大自身优势。AI技术创新应用的大规模落地，带动了大数据智能市场的蓬勃发展，同样也为底层的数据治理服务注入了市场活力。

人工智能技术广泛渗透进经济生产活动主要环节



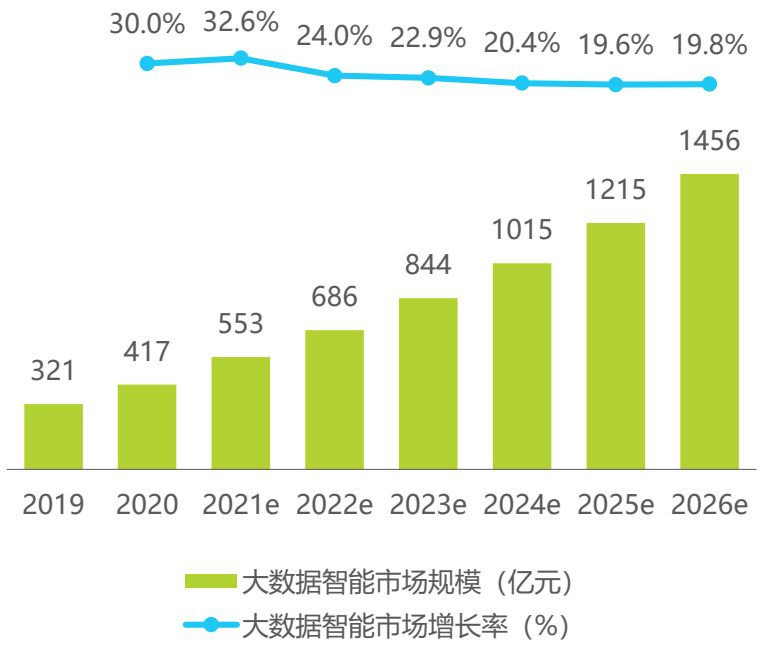
来源：《2021年中国人工智能产业研究报告（IV）》，艾瑞研究院自主研究绘制。

大数据智能市场的行业规模

2021年市场规模约为553亿元，金融数据率先得到价值释放

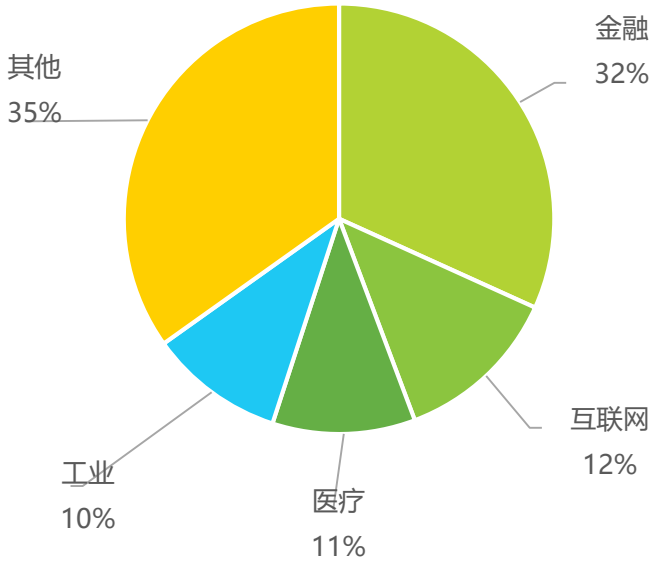
据艾瑞咨询统计测算，2021年涵盖大数据分析预测（机器学习/深度学习模型）、领域知识图谱及NLP应用的大数据智能市场规模约为553亿元，预计2026年市场规模将达到1456亿元，2021-2026 CAGR=21.3%。随着市场大数据基础的完善与数据需求的唤醒推动，大数据智能市场的规模将持续走高，但未来在行业理性建设与增量市场逐步完善的大背景下，大数据智能市场增速会出现下降趋势。从细分结构来看中，金融领域的数据价值率先得到释放，市场规模占比高达32%。

2019-2026年大数据智能市场规模



来源：《2021年人工智能产业研究报告（IV）》，艾瑞根据专家访谈、招投标项目统计推算而得

2021年大数据智能市场规模细分结构



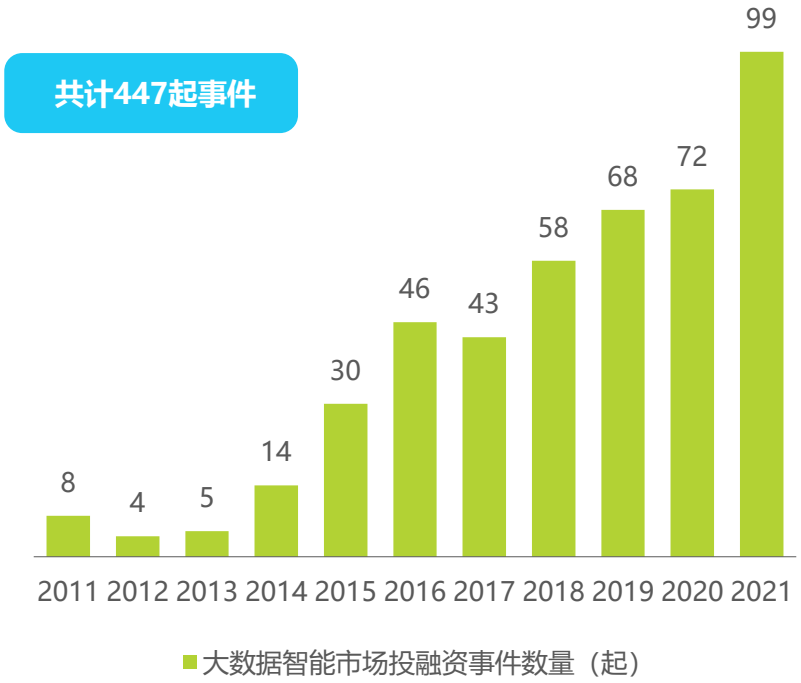
来源：《2021年人工智能产业研究报告（IV）》，艾瑞根据专家访谈、招投标项目统计推算而得

大数据智能市场的投融资热度

融资规模稳步提升，事件数量创历史新高

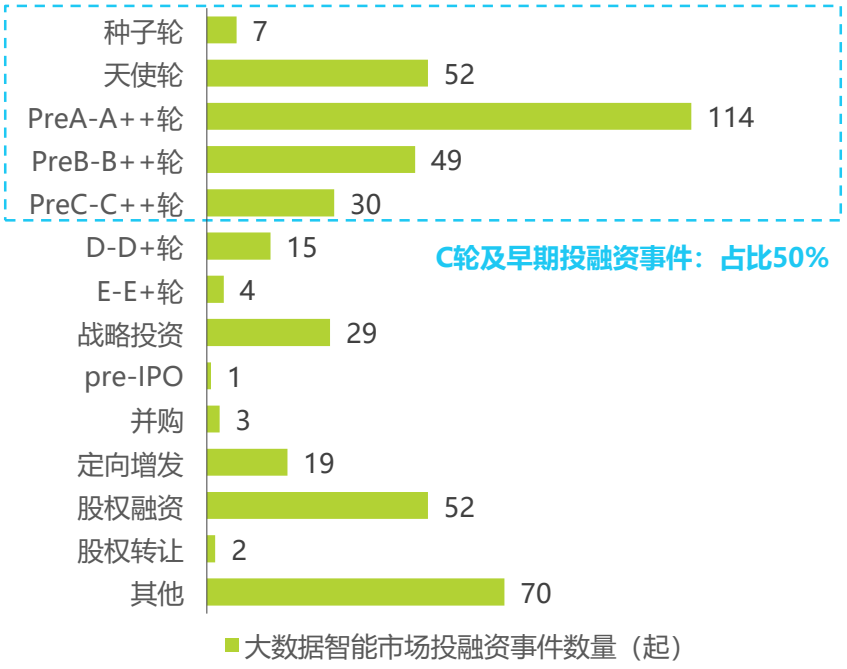
从2011-2021年的投资数量来看，资本市场对大数据智能市场的关注度不断提高，融资事件逐年攀升，2021年大数据智能市场单年投融资数量已高达99起；从2011-2021年的融资轮次来看，C轮及早期投融资事件占比达到50%。受政策的高度支持与技术的成熟推动，大数据智能应用在多行业的成功落地极大地增强了市场与投资者的信心，“大数据智能”标签已成为市场创业与投资的热点，具备市场想象空间与明确使用价值是企业早期吸引投资的关键。

2011-2021年大数据智能市场投融资事件数量



来源：艾瑞研究院根据融资网站数据调整与处理绘制

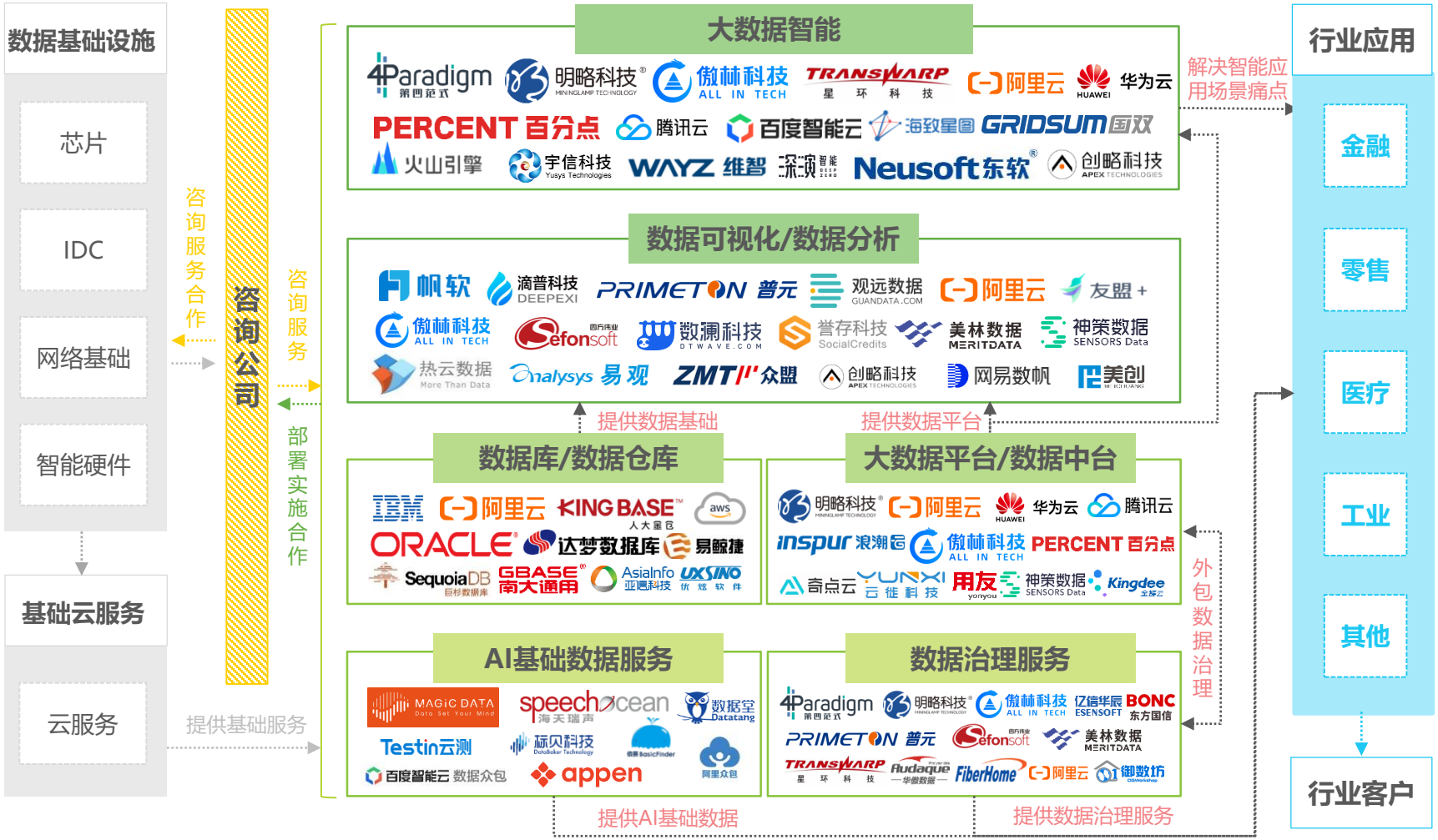
2011-2021年大数据智能市场投融资事件轮次情况



注释：其他包含IPO上市与基石投资轮。
来源：艾瑞研究院根据融资网站数据调整与处理绘制

大数据智能产业生态圈

大数据产业图谱与数据服务关系链



注释：以上厂商与行业为不完全列举，排名不分先后。
来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理：需求传导

人工智能应用引发的数据治理需求

企业在部署AI应用时，数据资源的优劣极大程度决定了AI应用的落地效果。因此，为推进AI应用的高质量落地，开展针对性的数据治理工作为首要且必要的环节。而对于企业本身已搭建的传统数据治理体系，目前多停留在对于结构性数据的治理优化，在数据质量、数据字段丰富度、数据分布和数据实时性等维度尚难满足AI应用对数据的高质量要求。为保证AI应用的高质效落地，企业仍需进行面向人工智能应用的二次数据治理工作。

AI应用对数据治理需求传导图

AI应用的数据要求

数据规模

传统的数据治理更多是以人为面向对象，基于有限数据容量进行聚合类信息展示，AI可以接纳的数据量远远大于人所接纳的数据量和信息量，且**可用高质量数据越多，模型质量和准确性越好。**

数据类型

AI应用，尤其是知识图谱的搭建，需要大量的半结构化和非结构化数据支持来开展工作。因此AI应用**在结构化数据的基础上，会将各类半结构化或非结构化数据纳入数据源并支持上层分析应用。**

数据质量

AI模型对数据高度敏感，其质量优劣极大程度影响AI模型的应用效果，因此AI数据源需极力规避“garbage in， garbage out”的问题发生，**多维度的质量检查成为必修课。**

数据实时性

AI模型对实时性要求高，大部分应用需基于实时数据实现分析、推荐和预警等目的，**支持AI应用的数据源更强调具备实时性接入能力。**

数据治理的需求传导

基于AI应用的数据治理需求

接入多源异构数据源

挖掘企业内外部信息，纳入结构化数据、半结构化数据和非结构化数据，提升与AI模型相关的数据积累。**数据训练规模扩张，数据类型异构，数据噪声指数级增加，对此建立针对性的数据治理体系**

数据融合&质量优化

1）多个数据源下的数据内容不一致等问题
2）缺失值、缺失字段；
3）错误值、异常样本；

特征工程

融合结构化数据、半结构化数据和非结构化数据，**进行以AI应用为目的的特征工程**

接入实时性数据

AI模型需纳入实时数据，构建**批流一体的数据聚合计算模式**

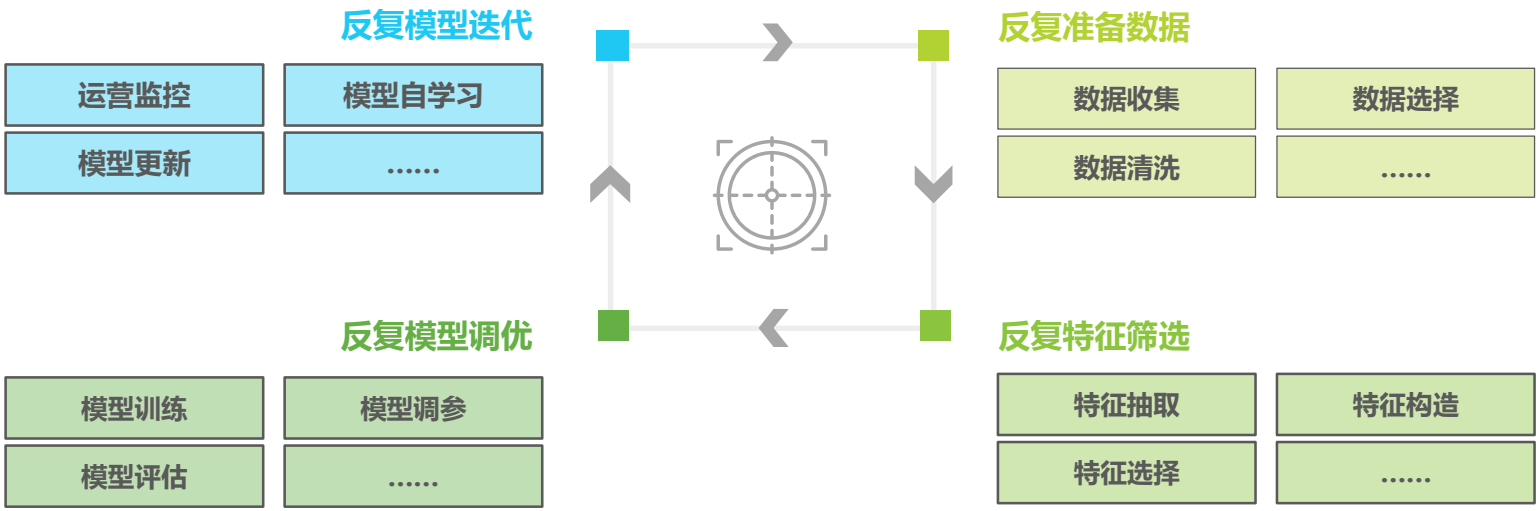
来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理：反复治理

面对反复的治理工作，搭建针对性体系解决重复性环节

数据治理在人工智能项目的实施中花费90%以上的精力，而面对企业的各人工智能项目，在AI数据层面多存在反复治理工作，极大拉低了AI应用的规模化落地效率。借助有效的方法论和实用的工具提高数据治理的效率，是企业管理数据资产与实现AI规模化应用的重要课题。搭建面向人工智能的数据治理体系，可将面向AI应用的数据治理环节流程化、标准化和体系化，降低数据反复准备、特征筛选、模型调优迭代的成本，缩短AI模型的开发构建全流程周期，最终显著提升AI应用的规模化落地效率。

搭建面向人工智能的数据治理体系 – 解决AI数据的重复性“治理”



来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理：体系搭建

吸收传统体系智慧沉淀，以AI应用数据需求为核心优化建设

面向人工智能的数据治理是传统数据治理体系在以AI应用落地为导向下的体系“升级”。从数据管理维度来看，在接入并处理分析半结构化数据、非结构化数据与流式数据的多源异构数据基础上，面向人工智能的数据治理体系仍会根据数据结构化流向、数据资产管理需要、数据安全需求等角度顺应搭建元数据管理、数据资产管理、主数据管理、数据生命周期管理和数据安全隐私管理等组件模块。而在数据治理过程中，则会更强调底层实现多源数据融合、数据采集频率、数据标准建立、数据质量管理，满足AI模型所需数据的规模、质量和时效，以AI应用的数据需求为核心，优化对应模块的体系建设。

面向人工智能的数据治理体系

来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理：数据准备

基于AI模型需求明确数据的特征准备、实时与否和闭环流通

从搭建流程来看，AI模型可大致分为离线训练和上线推理两个阶段。离线训练时，需基于AI模型运行目的确认数据采集来源，选择数据对应的时间间隔和时间节点，让AI能够在离线建模及上线运行后获取真实业务数据，模型训练效果能够保质保量落地。如果模型需要AI数据的实时接入，还需打造批流一体式的产品体系。基于实时数据处理、实时特征开发和实时应用开发等数据架构搭建批流一体的数据产品，将流式数据的接入实时反馈到模型运行输出，使模型结果更加及时准确。另外，AI模型上线后，需达到AI数据的闭环流通，通过打造数据采集和回馈分析的闭环式自学习体系，达到AI模型上线后的持续迭代优化。

AI模型的数据准备

确认AI模型所需数据的采集来源，对接企业IT系统获取数据源

时序性

AI数据的时间选择

时序性数据为按照时间维度索引的数据，描述了数据产生主体在某时间范围内的时点上的产生值。面对时序性数据，需准确记录每个数据的产生时间点，并**基于AI模型运行目的选择所需数据的时间间隔和时间节点。**比如，若想通过AI模型去预测分析某导演A电影的上映票房 ，则需选择该导演A电影上映以前的历史票房数据作为模型原料，而不能将A电影上映以后的电影票房数据计算在内，**避免AI“偷看”未来数据，给模型判断结果带来偏差。**

判断是否需要

AI数据的实时接入

在批式数据（全量）基础上，将流式数据（增量）纳入模型计算，打造**“批流一体”的数据产品，使模型分析结果更加实时准确，满足用户对动态变化的场景可即时作出反应的需求**

保证离线和在线的特征开发标准一致

实时特征开发 **实时AI应用开发**

实时交易反欺诈 / 实时推荐 / 实时订单监控....

需达成

AI数据的闭环流通

为了让AI模型的预测结果更加准确，可将模型运行后的结果数据更新反馈给AI模型，**利用实时闭环数据进行自学习，强化反馈回路以优化模型算法，防止模型效果因搭建时间久远而衰退。**比如，在智能推荐模型上线后，如果持续用静态模型预测动态数据，不更新闭环数据反馈的话，随着时间的流逝，模型将逐渐降低精准推荐效果，导致业务决策的效果越来越差。

来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理：数据质量

对应AI应用的高质量要求，唤醒沉睡数据，挖掘核心价值

多源异构数据的质量管理体系可从数据有效性、数据一致性、数据唯一性、数据时序性、数据完备性、数据完整性、数据合理性和数据准确性六个维度建立。其中，传统数据治理体系同样会高度关注数据的有效性、一致性和唯一性，但当数据治理范围扩大到多源异构数据时，需在数据融合过程中对这三个维度进行重新判断，例如非结构化数据在清洗处理后与结构化数据出现实体重复或内容不一致的情况；数据时序性是对数据时间维度的质量要求，从AI应用模型的需求出发，考虑数据接入的实时性和如何选择数据的时间间隔；数据完备性和数据完整性是对数据选取的评估维度，数据完备性要求数据需符合多维度字段特征以满足建模，数据完整性则对数据从历史到上线反馈的完整性接入以达到优质闭环；数据合理性和数据准确性则是对数据本身表达的更高质量要求。传统数据治理体系为做数据可视化和数据基本分析应用服务时，不会过多考虑到数据分布是否合理及表达内容是否准确等问题。然而在AI模型开发训练时，数据的合理分布和准确表达极大程度上决定了AI模型的分析决策效果，因此在面向人工智能的数据治理体系中，数据合理性和数据准确性的质量评估是体系需重点关注提升的维度模块。

多源异构数据的质量管理体系

数据有效性

即数据值与定义的**值域（有效值/有效参考范围/通过规定确定的值）一致**

数据一致性

即**数据属性表达一致**，数据一致性是数据标准化的基础，确保数据符合内容和形式规范

数据唯一性

即**数据集的实体不会重复出现。**对数据进行去重，底层实现数据一致性管理

数据时序性

一方面需根据数据更新频率和数据需求时效判断数据的**及时性**，一方面需根据AI模型的需求结果**判断数据选取的时间间隔**

数据完备性

即数据字段维度是否符合AI建模要求。尤其对于非结构化数据来说，需要有足够**完备的数据基础可提取到建模时所需字段特征**

数据完整性

AI 应用的算法模型**不仅需要业务历史数据训练，也需及时更新模型上线后的数据**，基于反馈对模型进行不断的迭代优化，打造优质数据闭环

数据合理性

即数据模式符合预期的程度。或通过基准数据比较，或基于过去相似数据集实例**判断数据的分布、变化和模式是否合理，是否出现异常值影响建模效果**

数据准确性

即数据正确表示“真实”实体的程度。数据准确性是基于数据有效性和数据合理性的进阶版，**需人或机器基于事实或规则判断数据是否准确**

来源：《DAMA数据管理知识体系指南第二版》，艾瑞研究院根据参考资料与专家访谈自主研究绘制。

面向人工智能的数据治理：数据标准

为AI模型开发提供“一致的数据语言”，实现数据复用共享

数据标准是数据治理工作的开展基础，为AI模型开发及应用提供“一致的数据语言”。在面向人工智能的数据治理体系中，数据标准的建立仍是数据实现共享流通、价值挖掘的核心环节。企业根据对应的国家标准、行业标准、地方标准等规范，结合自身情况和业务术语参考，以AI应用需求圈定的数据范围为治理导向，构建相关基础数据标准、指标数据标准和数据模型标准，形成全局统一的数据定义与价值体系。

多源异构数据的标准体系

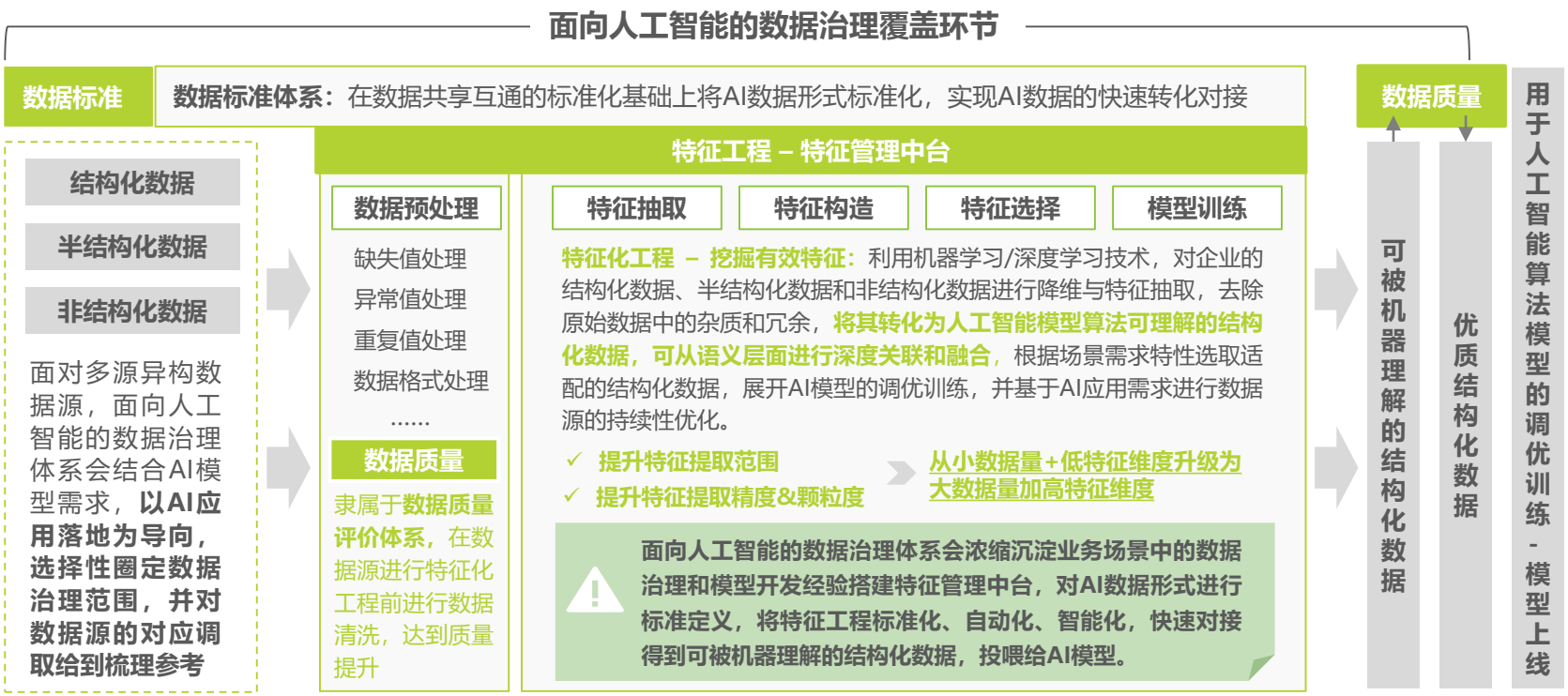
来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理：特征管理

将多源异构数据源转化为机器可理解的“结构化数据”

在圈定AI数据源范围并接入相应数据后，特征管理中台会对数据进行预处理，基于AI应用的数据要求处理缺失值、异常值、重复值和数据格式等问题，而后经过特征工程转化为人工智能模型可理解的结构化数据。在特征化工程环节中，面向人工智能的数据治理体系可浓缩沉淀业务场景中的数据治理和模型开发经验，对AI数据形式进行标准定义，搭建特征管理中台，将特征工程环节标准化、自动化、智能化，快速对接得到可被机器理解的优质结构化数据，投喂给AI模型。

让机器 “理解”多源异构数据的流程图

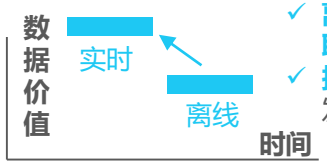


来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理：效果优化

显著提升AI应用的规模化落地效果

体系搭建-效果优化

关注环节	问题	优化	效果
数据采集准备	● 未考虑**数据时序性** ● **时效性差**，难支持数据实时接入	✓ 基于AI模型运行目的选择所需**数据的时间间隔和时间节点** ✓ 接入实时性数据，打造**“批流一体”的产品架构**	 ✓ **离线建模的时候获取真实业务数据** ✓ **接入实时性数据**，发挥数据时效价值
数据质量&标准	● 多源异构数据的**质量待优化** ● 数据标准不统一，**难以共享复用**	✓ 打造**多源异构数据的质量管理体系**，从六维度针对性评估提升数据质量 ✓ 构建**基础数据标准、指标数据标准和数据模型标准**，在数据共享流通基础上为模型开发提供“一致语言”	**数据质量管理体系** **数据标准体系** ✓ 为AI模型提供高质量数据原料，**提高模型拟合效果** ✓ 一致性语言**减少数据反复治理工作**
数据特征维度	● **重复性特征工程** ● **特征维度低**，模型欠拟合	✓ 沉淀AI项目的数据治理经验，构建**特征管理中台** ✓ 提升特征提取范围、精度和颗粒度，**从小数据量+低特征维度升级为大数据量加高特征维度**	**特征管理中台** ✓ **减少重复性特征工程的精力投入** ✓ 指数级提升数据的特征维度，**优化模型拟合效果**
模型迭代优化	● 模型上线后不迭代优化，**随时间流逝拟合效果越来越差**	✓ 打造数据采集和反馈分析的闭环体系，**强化反馈回路优化模型算法效果** ✓ 借助机器学习技术，使模型进行**自学习式迭代优化**	**自学习式闭环体系** ✓ 模型不过时，**基于数据变化实时更新迭代** ✓ 拟合效果**优化达到持续且自动化**

来源：艾瑞研究院自主研究绘制。

前言：数据与数据治理	1
主题：面向人工智能的数据治理	2
参与：行业规模与受益圈立足点	3
实践：高频高价值应用及数据痛点	4
案例：标杆企业与新锐势力	5
展望：治理陷阱与趋势洞察	6

AI数据：产业生态圈

中游厂商提供数据开发与治理服务，助力AI应用高效落地

AI基础数据服务与数据治理产业链结构为：1）上游：数据源与数据产能；2）中游：数据产品开发工具与管理服务；3）下游：人工智能应用。处于中游的两类服务商中，AI基础数据服务商使用数据采集与标注工具处理图片、语音、文本等非结构化数据，面向AI的数据治理服务商则负责使用数据治理的各组件管治多源异构数据，使其形成数据资产，从而提高数据质量。二者处理后的数据可直接为下游的AI训练所用，使AI应用落地能够省时省力。

面向AI的数据治理产业图谱

注释：图谱中所展示的公司logo顺序及大小并无实际意义，不涉及排名。

来源：《2021年中国人工智能基础层行业研究报告》，艾瑞研究院自主研究绘制。

数智融合产业带来多元厂商参与

数据治理与AI应用开展交汇融合，厂商参与更加丰富多元

依托于数据与AI模型的紧密关联，数据治理与AI应用产品已逐步开展交汇融合，展现“由数据治理到开发AI应用平台/产品”与“AI应用平台/产品开发到面向AI的数据治理”的两路发展方向：1）数据治理厂商在积累数据经验与AI模型理解后，为实现业务拓展而将领域从数据层延伸至AI应用及平台开发层；2）从事AI应用及平台开发的AI厂商，也会在数据治理经验不断丰富的背景下，着手向底层开展面向AI的数据治理业务，依托于自身AI技术与业务理解，让面向AI的数据源更加契合AI应用模型要求以提升模型拟合效果。因此，面向AI的数据治理从业者不仅仅为数据治理厂商，更包括众多AI企业，参与者更加丰富多元。

数据治理与AI应用产品的交汇融合

来源：艾瑞研究院自主研究绘制。

数智产业生态圈的受益节奏

三方阵营厂商构建行业竞合格局

AI应用的加速落地带来的大量数据治理需求，吸引众多厂商参与其中。从行业厂商类型来看，主要包括咨询公司、数据服务相关提供商和人工智能产品提供商三类。各类厂商根据自身业务特点和切入方式获得差异化的竞争优势，而由于面向人工智能的数据治理服务的参与立足点丰富，厂商之间可能基于同类业务展开竞争，同时在差异化领域进行合作，形成竞争与合作高度共存的行业格局。

面向人工智能的数据治理 - 行业厂商类型与竞合格局

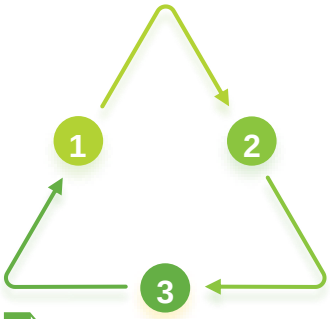
竞争领域：在方案设计层面，两者存在竞争关系
合作领域：咨询公司提供方案，数据平台服务商提供落地实施

咨询公司

数据治理规划咨询：为企业提供数据治理的科学规划和顶层设计，以方案形式交付，从调研诊断、方案设计、落地跟踪与应用指导等阶段持续跟进，为数据平台服务商和人工智能服务商的数据治理服务落地提供参考指引。

数据服务相关提供商

基于企业AI应用的数据需求提供相关治理服务：企业上层应用涉及诸多AI技术，基于数据平台（大数据平台/数据中台/数据仓库）项目或数据治理服务项目中，厂商会在传统数据治理体系基础上，基于AI应用所需的数据质量要求，提供面向AI的数据治理工作。



竞争领域：在方案设计层面，两者存在竞争关系

合作领域：咨询公司提供方案，人工智能服务商提供落地实施

人工智能产品提供商

为提供人工智能平台应用开展针对性数据治理工作：在机器学习、自然语言理解和知识图谱的建设项目中，基于AI模型训练所需的数据原料要求，提供结构化、特征化处理和数据质量优化的治理服务，是人工智能项目交付必不可少的工作构成。另外，AI厂商在沉淀项目经验和AI能力后搭建AI开发平台，也会将面向AI的数据治理能力标准化，从而纳入AI开发平台模块中。

竞争领域：两者均可基于AI应用需求提供面向AI的数据治理服务

合作领域：数据平台服务商搭建基础数据治理体系，人工智能厂商优化以AI模型为导向的数据治理服务

来源：艾瑞研究院自主研究绘制。

数智产业生态圈的参与立足点

“智”为面向人工智能的数据治理服务的核心立足点

面向人工智能的数据治理服务常包含于数据服务、平台能力和数据产品三类采购形式中。第一类，数据服务即以单独的数据治理产品形式出现，但由于面向人工智能的数据治理尚未发展出成熟独立的产品模式，因此以该类形式出现的业务涉及较少。市面上大多数据治理服务仍以传统数据治理的形式存在，对于已搭建AI中台或AI应用较为广泛的智能化转型先行企业，会在数据治理产品采购方案中添加对支持AI应用数据的治理需求；第二类，数据平台，主要包括大数据平台、数据中台、数据仓库和AI能力平台等项目。大体量大数据平台、数据中台和数仓项目多会包含AI应用体系建设，从而涵盖部分对应的AI数据治理服务。另外AI平台/中台可沉淀面向AI的数据治理能力，将其标准化后纳入平台模块和产品项目中；第三类，数据产品，范围限定在应用AI算法的数据产品，可划分为机器学习产品、自然语言理解产品和知识图谱三类AI产品。为保证AI算法模型的优质运行效果，更好地提供预测、决策、推荐和风控等产品功能，需要对算法模型的训练原料，即支持AI应用的底层数据，进行针对性优化治理。如今AI产品需求旺盛，AI开发平台陆续推进AI产品的规模化落地，且AI数据治理效果与最终平台产品交付效果紧密相连，AI应用驱动成为面向人工智能的数据治理服务的核心立足点。

面向人工智能的数据治理 – 受益节奏与参与立足点

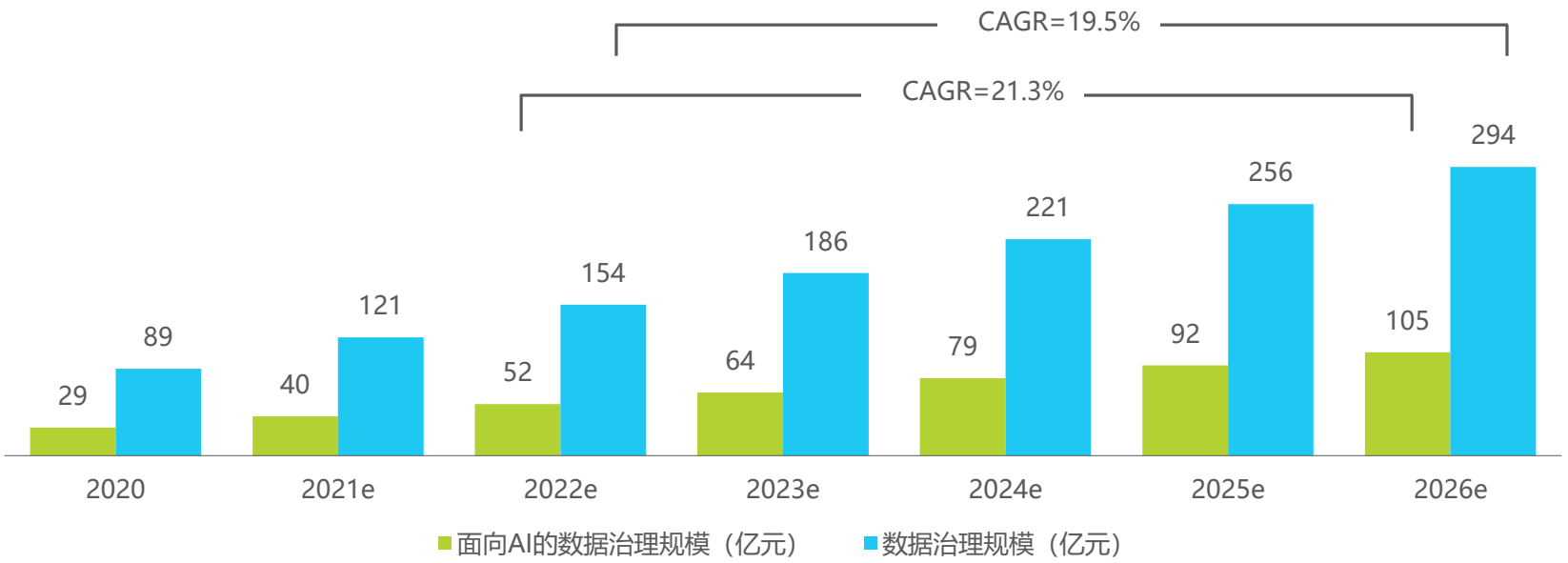
来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理：行业规模

2021年市场规模约为40亿元，预计五年后规模突破百亿

从数智产业圈的参与立足点出发，艾瑞提取测算了大数据平台、数据中台、AI应用与数据治理服务的项目中与AI应用相关的数据治理规模并加总而得，2021年中国面向人工智能的数据治理规模约为40亿元。受数据平台服务、数据治理服务和AI应用建设的需求推动影响，面向人工智能的数据治理规模将持续上升，2026年规模突破百亿达105亿元，2021-2026 CAGR=21.3%。2021年，中国的数据治理市场规模约为121亿元。作为数据服务的基础工作，数据治理规模将保持上扬态势，预计2026年市场规模达到294亿元，2021-2026 CAGR=19.5%。从发展曲线来看，数据治理与面向人工智能的数据治理规模增长均处于良性区间，共同巩固相关治理产业生态圈的向好形势。

2019-2026年中国数据治理与面向人工智能的数据治理规模



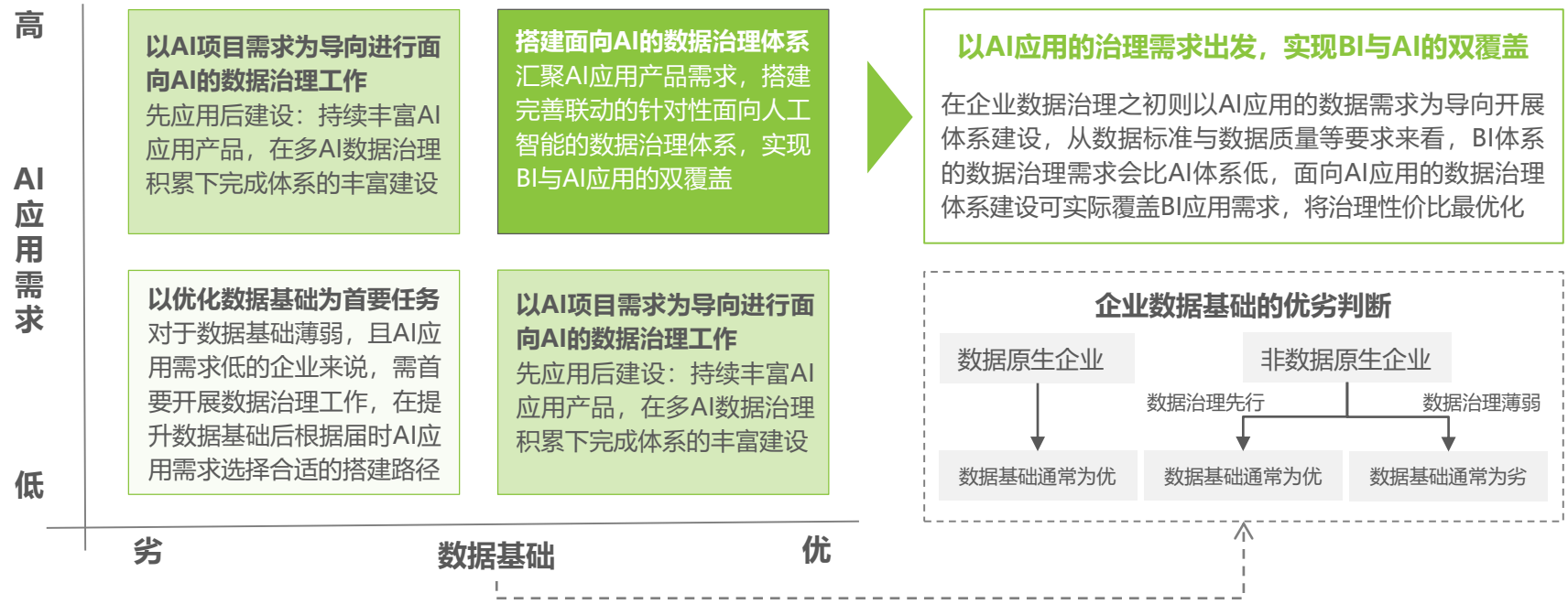
来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理：时机路径

契合客户的数据基础和AI应用需求的多元化选择

从数据基础的维度划分，可将企业分为数据原生企业与非数据原生企业。数据原生企业以互联网为代表的数据原生企业，在设立之初则以数字世界为中心构建，生成以软件和数据平台为核心的数字世界入口。该类企业往往不需要信息化、数字化转型，所要做的即为让数据共享流通的规范式管理。非数据原生企业成立之初以物理世界构建，围绕生产、流通、服务等具体的经济活动展开，天然缺乏以软件和数据平台为核心的数字世界入口。为了更好挖掘自身企业数据价值，非数字原生往往需要进行企业的数字化转型，需通过数字化转型程度与数据治理阶段判断非数据原生企业的数据基础优劣。最终，结合企业数据基础与AI应用需求为面向人工智能的数据治理的体系搭建提供契合路径，完成企业数据体系的进一步升级。

面向人工智能的数据治理 – 体系搭建评估框架



来源：《华为数据之道》；艾瑞研究院自主研究绘制。

前言：数据与数据治理 1

主题：面向人工智能的数据治理 2

参与：行业规模与受益圈立足点 3

实践：高频高价值应用及数据痛点 4

案例：标杆企业与新锐势力 5

展望：治理陷阱与趋势洞察 6

金融

—— *Finance*

金融信息化建设阶段：金融行业是我国信息化发展程度最高、信息技术应用最密集的行业之一，在政策引导与内生需求的双重推动下，经历了从基础建设、平台建设到AI应用建设的过程。银行金融机构中是对IT技术投入最高，在AI布局较早的主体。

高频高价值业务场景：近年来，AI与金融业务深度融合，促进了金融机构的业务流程、运营模式、风险管控等更流畅、更高效地运作。智能风控与智能营销的应用度最高，银行应用大数据及人工智能建模加强风控力度，同时建立精准客户画像与客户分层分类差异化定价体系，提高对客户的经营能力和营销管理水平。

面向人工智能的数据治理：在银行主要集中的高频高价值AI应用场景中，对技术融合、数据质量与业务理解的要求普遍较高，需要供应方具备充足的技术与项目积累，AI应用面临的数据质量问题日渐凸显，一些银行开始寻求构建面向人工智能的数据治理体系。对金融行业来说，构建面向人工智能的数据治理体系对支撑AI应用起着重要作用，高质量数据的持续输入可以更好地保障后续特征工程和模型训练的效果，并降低模型上线的成本和潜在数据问题的风险。

行业发展背景

AI应用蓬勃发展，银行是主要需求方

金融行业是我国信息化发展程度最高、信息技术应用最密集的行业之一。在政策引导与内生需求的双重推动下，金融行业经历了从基础建设、平台建设到AI应用建设的发展过程。随着信息化创新的不断深入，金融机构对IT系统的安全性、稳定性、灵活性、功能性、可扩展性等方面也随之提出更高的要求。在此过程中，银行持续加大IT投入，纷纷设立单独的科技子公司，在满足内生需求的同时技术外溢为金融行业提供解决方案。总体来看，银行金融机构中是对IT技术投入最高，并AI布局较早的主体。

中国金融行业科技发展历程

基础建设期

金融机构通过IT系统构建、对外服务渠道等基础设施建设实现办公和业务的电子化。数据仓库建设规模小，规模和范围都很有限。

平台建设期

随着信息化率的提升，数据量、数据维度与数据体系逐渐庞大，金融机构开始搭建大数据平台、数据中台等实现技术与核心业务的融合。

AI应用建设期

充分利用客户行为、交易、负债、收入等信息进行获客活客、额度增信、贷后预警等核心业务的模型建设和数据分析，切实指导业务决策。

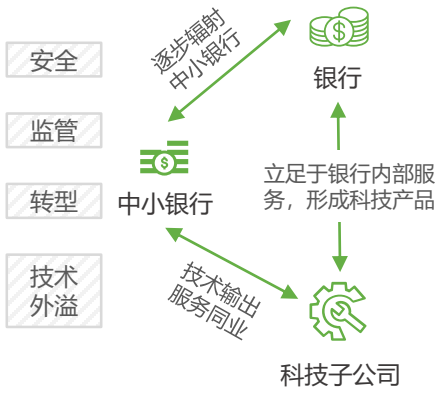
我国金融科技产业相关政策

政策驱动是金融机构不断扩大信息化支出的重要推动力，政策的变动会引发IT系统的**存量改造与增量建设**，针对金融行业，**数据治理**的相关要求、标准、框架与体系在逐渐建立。

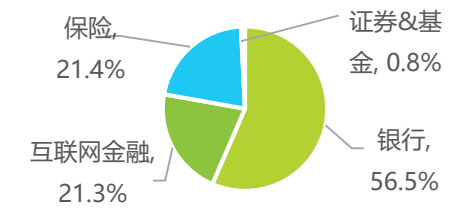
政策名称	发布时间	发布主体	主要内容
《银行业金融机构数据治理指引》	2018.05	银保监	对管理范围内的**原则、规范、技术**等领域都进行了更细化的要求，是金融机构合规运行的“教科书”
《金融科技发展规划（2019-2021年）》	2019.08	中国人民银行	金融科技应用先进可控，金融服务能力稳步增强，金融风控水平明显提高，金融监管效能持续提升，金融科技支撑不断完善。
《关于开展监管数据质量专项治理工作的通知》	2020.05	银保监	要求各金融机构开展监管数据质量专项治理

金融机构设立科技子公司

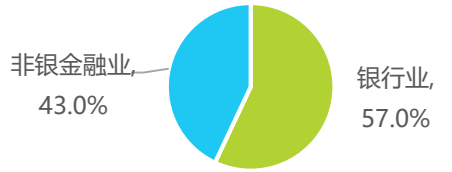
金融机构不断加大IT方面的**自研投入**，纷纷设立单独的**金融科技子公司**，辐射中小金融行业与更多行业受众。



2019年中国金融机构AI投入规模结构



预计2022年全球金融业支出结构



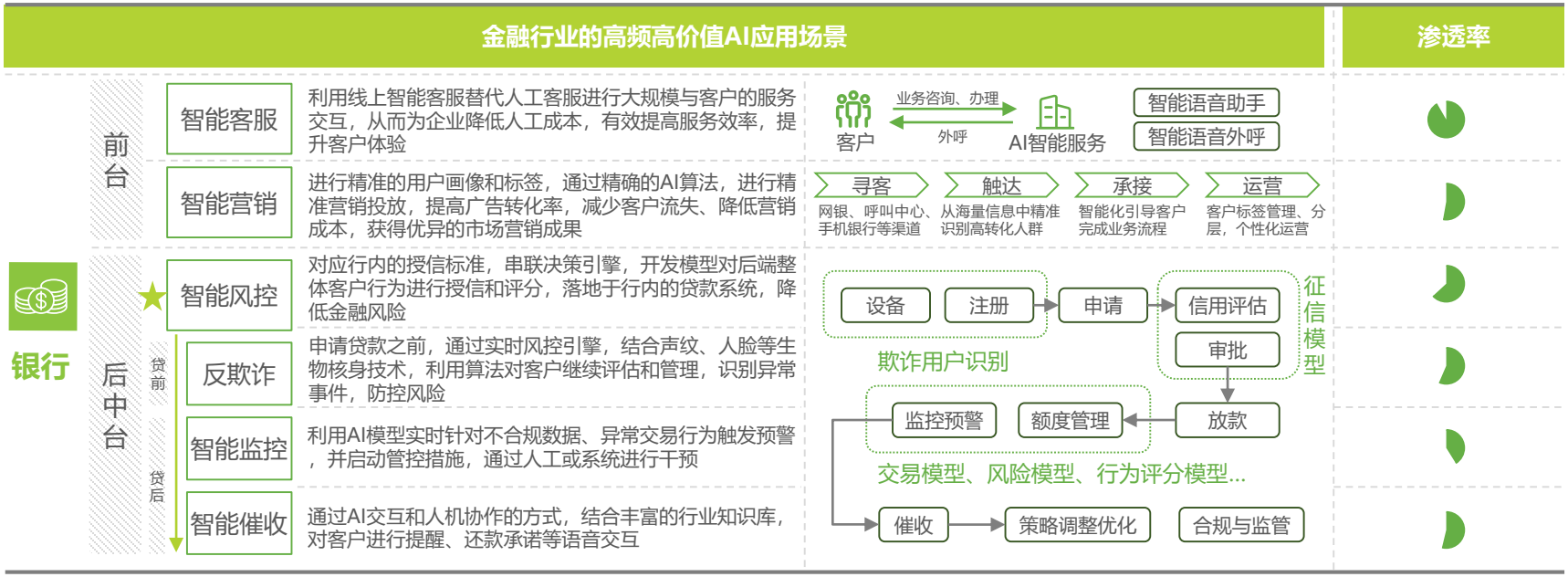
来源：《2020中国AI+金融行业发展研究报告》，艾瑞研究院自主研究绘制。

高频高价值场景（1/2）

AI应用业务伴生，由前台向后中台转化渗透

近年来，AI与金融业务深度融合，促进了金融机构的业务流程、运营模式、风险管控等更流畅、更高效地运作。银行AI应用的渗透率较高，覆盖日常运营与核心业务，其中智能风控与智能营销的应用度最高。银行以大数据及人工智能建模持续加强风控力度，同时建立了精准客户画像与客户分层分类的差异化定价体系，提高对客户的经营能力和营销管理水平。保险与证券的AI应用则集中在了风险评估与反欺诈场景。总体来看，金融行业在不断进行AI应用场景的加深拓宽与AI应用效果的落地推广，业务领域实现了由前台向后中台的转化渗透。

中国金融行业高频高价值AI应用场景



来源：艾瑞研究院自主研究绘制。

高频高价值场景（2/2）

场景多元化，对技术与业务理解的要求普遍较高

银行业的数据建设成熟、数据基础完整且标准化程度高，信息化程度普遍较高。而银行业AI应用的落地场景呈现多元化特征，高频高价值AI应用场景对技术融合、数据质量与业务理解的要求普遍较高，因此需要供应方具备充足的技术与项目积累。金融领域的AI应用多为业务导向型，即AI建设逻辑为应用落地先行，而AI应用面临的数据质量问题日渐凸显，一些银行开始寻求构建面向人工智能的数据治理体系的解决之道。

银行行业高频高价值应用场景具体特征

AI高频高价值应用场景			技术要求	数据来源	数据质量要求	内部管理	数据痛点	面向人工智能的数据治理
银行	前台	智能客服	✓ 中等 ✓ 高并发、语义理解	**自有数据**：客户信息、行为数据、交易数据、账务数据、日志文件等 **外部数据**：企业征信数据、交易票据信息、税务、工商等信息 **网络数据**：补充完善客户行为偏好、挖掘用户社交关系的信息 原始数据 加工后数据	✓ 较低	银行通过自身能力或第三方支持自建AI应用，其自身组织结构与内部管理需要适应整个风控体系做出调整，例如区分基础数据人员、大数据分析人员、建模人员等	✓ 数据孤岛	✓ 有一定数据积累，且AI基础数据服务市场较为成熟 ✓ 数据治理工具标准化程度、智能化程度高 ✓ 数据安全与监管非常重要
		智能营销	✓ 中等		✓ 中等		✓ 数据孤岛	
	后中台	★ 智能风控	✓ 较高 ✓ 授信交易时间要求		✓ 较高 ✓ 模型准确率要求高，通过不良率、预期率、监控模型反馈分析模型准确性		✓ 数据质量问题，例如数据同源、一致性问题 ✓ 需更多数据案例积累实现精准刻画	✓ 面对AI需求，对接入数据**二次治理比例约为60%**，比如数据量、频率、波动范围等 ✓ 业务部门提出相应需求，由大数据团队、模型团队进行二次加工处理，随着逐步适应业务续期，自主加工性更强
	贷前	反欺诈	✓ 较高					
	贷前	智能监控	✓ 较高					
	贷后	智能催收	✓ 较高					

需求特点：①场景**多元化**，**技术要求普遍较高，业务理解要求高**；②**AI集中于局部应用**，数据不直接为建模准备，而是**业务导向型**，根据业务需求case by case治理数据，**二次加工比例高**

来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理体系

AI需求作用于传统数据治理的运行逻辑

对金融行业来说，面向人工智能的数据治理体系对支撑AI应用起着重要作用，高质量数据的持续输入可以更好地保障后续特征工程和模型训练的效果，并降低模型上线的成本和潜在数据问题的风险。如前文所叙，AI应用多为业务导向型，需根据业务落地需求治理数据，因此金融机构多在传统数据治理体系之上，随着AI应用的丰富，逐渐区分、建立出面向人工智能的数据治理体系。不管金融机构作为被动的接受者还是主动的参与者，都需要以战略、机制、技术为支撑，从明确目标开始，理清数据治理的关键点，再从组织、人员、流程、数据四个方面的入手制定规划，执行指标建设与数据管理，最终保障数据治理的持续运营能力，形成由上至下做推进指导的多维数据治理体系。

金融行业面向人工智能的数据治理体系建设

Step1：单点业务，明确目标

企业实施面向AI的数据治理的第一步，就是根据建设AI应用而出现的数据质量需求，**明确数据治理的目标，理清数据治理的关键点**

Step2：数据准备，制定规划

分析数据管理和数据质量现状，确定面向AI应用的“不一致”、“不完整”、“不准确”待治理数据，**从组织、人员、流程、数据四个方面的入手，规划全方位可持久的数据治理体系**

Step3：指标建设、数据管理

根据AI建模需要，建设针对性数据集，涉及元数据管理、贴源层管理、应用层管理、数据权限管理等，**从业务视角对需求场景涉及的指标进行抽象、归类**

Step4：技术支撑、应用开发

确保高质量的数据输入，保证特征工程与模型训练效果，**不断优化数据治理方案快速应用于相关的AI场景**，保障持续学习与模型更新能力、提高复用率，减少重复治理

金融行业数据治理体系建设与维护水平较高，**多在现有数据治理体系之上建设运营面向AI的数据治理体系**，管理调优AI应用

①传统的数据治理体系可以部分支撑AI应用需求，但会经常遇到特征缺失等问题

传统数据治理体系

面向AI的数据治理体系

②不能满足AI需求的数据，在ODS层的基础上进行二次治理

源数据

③同时还需要从源数据层提取需要的数据进行治理

模型训练
特征工程

来源：艾瑞研究院自主研究绘制。

金融 | 零售 | 医疗 | 工业

产业图谱

iResearch 艾瑞咨询

金融行业大数据智能产业图谱

下游

银行AI应用层

智能营销：明略科技、4Paradigm 第四范式、腾银财智、深演智能、Convertlab、创略科技、美信数科、百融云创

智能客服：科大讯飞 iFLYTEK、EMOTIBOT 竹间智能科技、阿里云、追一科技、JDT 京东科技、UDESK、沃丰科技、智齿客服

智能风控：4Paradigm 第四范式、同盾科技、iSoftStone 软通动力、海致星图、宇信科技、钒钛智能 FANTAI DATA、金融壹账通 ONECONNECT、百融金服、数美、ICE KREDIT 冰鉴

反欺诈：4Paradigm 第四范式、明略科技、同盾科技、众安科技、邦盛科技、FICO、中证鹏元、百融云创、AHI 慧安金科

智能监控：硅基智能、MIOTECH、AHI 慧安金科、邦盛科技、云智慧、信也科技、PiNTEC 品钛、elens、知因智慧 KnowLeGene

智能催收：建信金科 CCB Fintech、信盟、互连众信、百融云创、金融壹账通 ONECONNECT、百应、ZHUIYI、AIRUDDER

保险：4Paradigm 第四范式、明略科技、平安科技 PING AN TECHNOLOGY、中科软科技 Sinosoft Co.,Ltd、iSoftStone 软通动力、钛铭科技

证券：4Paradigm 第四范式、明略科技、云从科技 CLOUDWALK、赢时胜软件 YSSTECH SOFTWARE、HUNDSUN、symbiont

AI技术服务

- 计算机视觉：MEGVII 旷视、商汤 sensetime、依图 YITU
- 自然语言处理：科大讯飞 iFLYTEK、达观数据 DATA GRAND、云从科技 CLOUDWALK
- 语音技术：云知声 Unisound、思必驰 AISPEECH、科大讯飞 iFLYTEK、阿里云
- 机器学习：4Paradigm 第四范式、京东数科、九章云极 DataCanvas
- 知识图谱：明略科技、GTCOM、海致星图、JDT 京东科技

数据平台层

内部

数据湖：阿里云、腾讯云、ORACLE、aws、e-Hualu 易华录、华为云 HUAWEI、GBASE 南大通用、Google Cloud、IBM、中电金信 GienTech

大数据平台/数据中台：华为云 HUAWEI、腾讯云、用友 yonyou、阿里云、云徙科技 YUNXI、inspur 浪潮、奇点云、Kingdee 金蝶云、DCITS 神州信息

数据库/数据仓库：IBM、SequoiaDB 巨杉数据库、阿里云、GBASE 南大通用、aws、ORACLE、达梦数据库、UXSINO 优炫软件、AsiaInfo 亚信科技

AI基础数据服务：MAGIC DATA、Testin 云测、Baidu 众测、speechocean 海天瑞声

数据治理服务：4Paradigm 第四范式、明略科技、TRANSWARP 星环科技、BONC 东方国信

外部

征信数据：鹏元征信有限公司 PENGYUAN CREDIT SERVICE CO.,LTD.、百行征信 Baihang Credit、上海资信 cis、百融云创、信联征信 CREDLINK

基础数据：百望 Baiwang、Aisino 航天信息、微众税银 VZOOM CREDIT、用友 yonyou、东方微银、世联评估 WORLDUNION APPRAISAL

基础层

场景包含面向AI的数据治理

注释：以上厂商与行业为不完全列举，排名不分先后。
来源：艾瑞研究院自主研究绘制。

零售

—— *Retail*

零售信息化建设阶段：面对消费者购物渠道的多元化和消费主力军的年轻化，零售企业可通过数字化创新应用赋能企业，另受疫情宅经济和社区团购等零售新势力影响，中国零售行业以及相关企业正加速数字化升级进程，向全渠道、数字化、智能化的新零售转型。

高频高价值业务场景：如今零售企业面临线上和线下渠道的双重压力，需积极通过大数据和人工智能等新技术，赋能高频高价值的核心业务场景，而营销运营和供应链管理为零售转型升级的核心业务场景。

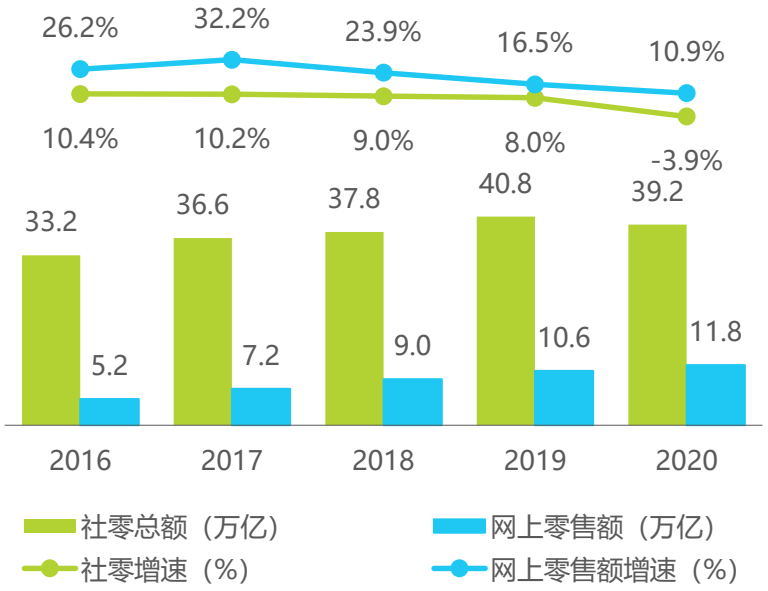
面向人工智能的数据治理：在营销运营和供应链管理的两大高频高价值的AI应用中，存在从数据采集、数据处理到数据应用的全维度数据痛点，极大影响AI模型的落地应用效果。零售企业大多为民营企业，上层管理者在进行项目决策时，会首要关注于开展此项目所带来的价值和产出。零售企业面向人工智能的数据治理体系建设多会通过AI应用驱动。以高频高价值的AI应用场景带来的价值产出打动管理者进行项目决策，并在AI应用的开发建设过程中，从数据层面进行针对性治理体系的建设和丰富。

行业发展背景

受疫情宅经济和竞争新势力影响，行业加快数字化升级进程

面对消费者购物渠道的多元化和消费主力军的年轻化，零售企业可通过数字化创新应用赋能企业，线下渠道向数字化、智能化转型，线上渠道积极利用大数据和人工智能应用，以全渠道流程优化提升消费者体验和企业竞争力。另外受疫情宅经济和社区团购等零售新势力影响，中国零售行业以及相关企业正加速数字化升级进程，向全渠道、数字化、智能化的新零售转型。

2016-2020中国社会消费品零售总额及网上零售额



注释：社会消费品零售总额是指企业（单位）通过交易售给个人、社会集团，非生产、非经营用的实物商品金额，以及提供餐饮服务所取得的收入金额。而网上零售额是指通过公共网络交易平台（包括自建网站和第三方平台）实现的商品和服务零售额之和。网上零售额与社会消费品零售总额两者不是完全的包含与被包含关系。
来源：国家统计局；艾瑞研究院自主研究绘制。

在疫情时代冲击、新势力竞争压力下，相关零售企业加快数字化升级进程

互联网时代下，消费者购物渠道选择更加多元

互联网的渗透普及改变着人们生活的方方面面，电商的兴起和繁荣让**网络购物**逐渐成为人们重要的消费方式。

消费主力军变化，更具线上购物倾向

以90后、95后为主的**“后浪”开始成为消费主力军**，他们从小接触互联网，更加熟悉线上购物习惯

数字经济稳步驱动

疫情下的宅经济爆发进一步扩大线上消费群体

受疫情影响，**“宅经济”爆发**，线上消费、送货上门成为了重要的消费形式，且渗透人群从年轻人慢慢向中老年人渗透。

社区团购对存量零售市场的蚕食压力

社区团购开始对传统零售行业造成影响，目前主要打击超市业态，也会存在对其余业态的潜在威胁。**零售新势力的崛起**正在给传统零售行业带来巨大压力。

疫情和新势力影响下的进程加快

来源：艾瑞研究院自主研究绘制。

©2022.3 iResearch Inc. www.iresearch.com.cn

高频高价值业务场景

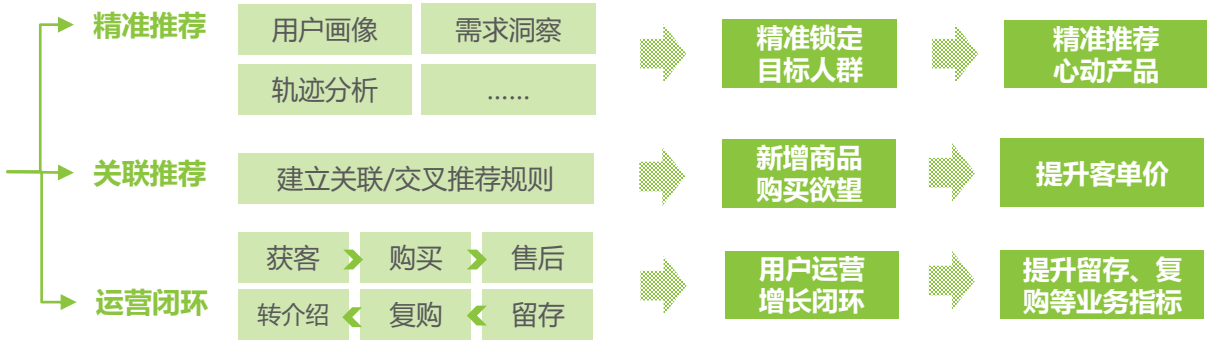
营销运营和供应链管理为零售转型升级的核心业务场景

如今零售企业面临线上和线下渠道的双重压力。线上来看，互联网红利逐步见顶，流量争夺日趋白热化，企业的获客成本持续攀升；线上来看，门店“坪效”天花板明显，亟待提升经营质效。对此，零售企业需积极通过大数据和人工智能等新技术，赋能高频高价值的核心业务场景，为企业在竞争日益激烈、产品日趋同质化的零售行业找到自身立足之地。

高频高价值业务需求

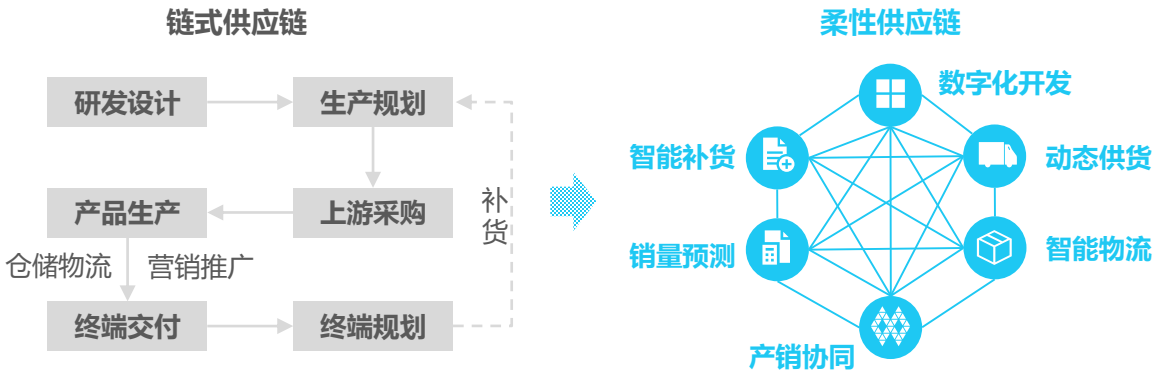
①营销运营

营销运营：通过分析用户行为数据等建立用户画像，洞察消费者需求，达到“千人千面”的个性化推荐效果，继续推荐关联产品以拉高客单价，打造从获客到复购、转介绍的运营闭环。**在提升消费者购买欲望的同时，优化消费者购物体验。**



②供应链管理

供应链管理：打通供应链数据并进行数字化、智能化管理。从供货、物流、销量预测和补货等环节进行全局优化，基于大数据和AI做动态分析，为企业降低仓储物流成本，提升供货调货效率。**有效解决人工情况下，销量预测不准确、物流成本高、补货调货效率低等情况。**



来源：艾瑞研究院自主研究绘制。

数据痛点与核心诉求

从数据采集、数据处理到数据应用的全维度治理需求

传统零售时代依靠人力经验做出商业决策，而在智慧新零售时代，大数下的AI应用可帮助零售企业能够快速知晓现况并做出最优反应，及时改进销售策略和调整产品，加深优化对消费者的服务质量与深度。在营销运营和供应链管理的两大高频高价值的AI应用中，存在着从数据采集、数据处理到数据应用的全维度数据痛点，极大影响AI模型的落地应用效果。零售企业可通过建立面向人工智能的数据治理体系针对性解决AI应用的数据问题，以提高AI应用的落地质量与效率。

零售应用中的数据痛点

数据来源多、类型繁多、管理复杂，数据质量标准亟需优化

传统线下零售企业的数字化转型带来数据孤岛，部分企业的业务系统数据林立，尚未联通。消费者的购物、行为数据等来源庞杂，多渠道且大规模的数据汇入给数据清洗带来困难。

数据采集丰富度有待提升

线下门店的消费者信息缺乏。消费者对商品的评估评价过程难以被收集量化，只能从销量数据的单一维度判断产品优劣，且线下购物的大量消费者行为数据缺失，无法掌握消费者群体属性。

多环节数据难以贯通，旨在通过协同应用加深数据服务深度

首先仍然存在数据打通问题，其次在数据打通情况下，由于零售供应链路过长，且影响因素众多，仍需进一步治理让生产、销售和物流等多环节数据得到贯通，提供智能调补货、产销协同等优化服务。

数据采集及时性有待提升

AI实时预测模型可实现分钟级别的交易量转换，实时基于产品销量改变产品曝光营销策略，从而促进整体的交易量，因此对数据源采集的实时性会有高要求。

数据基础劣 → 数据基础优

以线下实体零售企业起家： 多属于非数字原生企业，以物理世界为中心构建，围绕生产、流通和服务等环节展开，数字化转型过程中历经信息系统搭建、数据平台/中台整合和数据应用丰富等阶段，总体来说数字化基础较弱，且数字化转型进度根据企业规模和信息化投入不同会有较大差异。

以线上零售企业起家： 以线上电商零售企业起家，主体包括线上零售企业和电商自营平台。多属于数字原生企业，设立之初便以数字世界为中心构建，生成以软件和数据平台为核心的数字世界入口，可便捷获取和存储大量客户和交易等数据，数字化基础良好。

来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理体系

以AI场景应用为驱动力，行业Know how为重点厂商衡量

零售企业大多为民营企业，上层管理者在进行项目决策时，会首要关注于开展此项目所带来的价值和产出。因此，零售企业面向人工智能的数据治理体系建设多会通过AI应用驱动。以高频高价值的AI应用场景带来的价值产出打动管理者进行项目决策，并在AI应用的开发建设过程中，从数据层面进行针对性治理体系的建设和丰富。

面向人工智能的数据治理体系建设及供应商选型要素列举

来源：艾瑞研究院自主研究绘制。

产业图谱

零售行业大数据智能产业图谱

应用服务层

零售AI应用层

营销运营

4Paradigm 第四范式 | 明略科技 | JDT 京东科技 | 腾讯云 | 百度智能云 | 达观数据 DATA GRAND | ByteDance 字节跳动 | 讯飞AI营销云 | 阿里云 | PERCENT 百分点 | 云从科技 CLOUDWALK

供应链管理

阿里云 | 4Paradigm 第四范式 | HUAWEI 华为云 | inspur 浪潮 | 京东零售云 JD Retail Cloud

AI技术服务

- 计算机视觉：商汤 sensetime | MEGVII 旷视
- 自然语言处理：科大讯飞 iFLYTEK | 达观数据 DATA GRAND
- 语音识别：科大讯飞 iFLYTEK | 阿里云
- 机器学习：4Paradigm 第四范式 | 阿里云
- 知识图谱：明略科技 | 海致星图

数据平台层

企业数据源

- 内部数据
  - 用户数据
  - 商品数据
  - 渠道数据
  - 管理数据
  - ……

数据服务

场景包含面向AI的数据治理

AI基础数据服务

MAGIC DATA Data Set Your Mind | 百度智能云 数据众包 | 标贝科技 Databaker technology | speechocean 海天瑞声

零售大数据平台/数据中台

明略科技 | 阿里云 | 腾讯云 | 奇点云

零售智能云厂商

阿里云 | 百度智能云 | 腾讯云 | aws | 奇点云 | 苏宁易购 | 京东零售云 JD Retail Cloud

注释：以上厂商与行业为不完全列举，排名不分先后。
来源：艾瑞研究院自主研究绘制。

医疗

—— Medical

医疗信息化建设阶段：医疗行业的信息化发展历程可分为数字化阶段与智能化阶段，而其中医疗数字化建设又可分为医院管理信息化、医院临床信息化以及区域医疗信息化三个阶段，医院作为信息化的主体已经积累了大量医疗数据并具备了基础的处理分析能力。近年来，在国家政策与医院实际需求的推动下，大数据、人工智能等技术在医疗领域的应用探索成为市场关注重点，AI+医疗场景步入快速发展期。

高频高价值业务场景：AI医疗已经从医院管理类应用逐渐延伸至医院核心业务，能够有效提升医院管理水平与诊疗水平，医学影像与医保控费是院内AI应用最成熟的领域，基因检测与新药研发是院外AI应用最主要的应用领域。

面向人工智能的数据治理：医疗数据在流通、共享、存储、管理等环节尚未标准化，导致数据多源异构难汇集、数据标准体系不健全等问题始终存在，掣肘着AI应用乃至行业的发展。从院内院外整个医疗数据环境来看，除数据质量欠佳之外，数据量欠缺也制约了AI模型训练的效果。结合医疗行业在AI应用过程中的数据痛点，医疗行业数据治理体系的搭建应该以打好数据基础为主，其次，医院应以平台建设为主，在满足医院管理、运营的基础之上，以场景做切入，支撑AI业务单点的运营；最后，以建立持续迭代运营的数据治理体系为核心，搭建切实可行的执行通道，形成数据闭环。

行业发展背景

政策与需求双重推动，AI+医疗步入发展期

医疗行业的信息化发展历程可分为数字化阶段与智能化阶段，而其中医疗数字化建设又可分为医院管理信息化、医院临床信息化以及区域医疗信息化三个阶段，医院作为信息化的主体已经积累了大量医疗数据并具备了基础的处理分析能力。近年来，在国家政策与医院实际需求的推动下，大数据、人工智能等技术在医疗领域的应用探索成为市场关注的重点，AI+医疗场景步入快速发展期。同时，医疗服务范围也在逐渐向院外延伸，形成了以大健康产业为重点的医疗卫生服务体系。

中国医疗行业信息化发展历程

数字化阶段 → **智能化阶段**

★ 部分先进医疗机构已进入智能化发展阶段

数据采集存储

- 以优化流程为核心，逐渐深入核心业务，为医疗数据的积累、算法的搭建提供了基础。医疗信息化建设是发挥医疗数据价值的基础

区域医疗信息化

以医疗大数据为核心，通过联动院内院外平台、各子系统协调，实现各级医疗机构的信息共享与数据融合利用

医院临床信息化

以电子病历为核心的临床信息化系统建设，能够服务医生，对医院临床诊疗质量、患者服务能力与效率有改善和提升

医院管理信息化

医院内HIS等管理保障系统得到广泛应用，主要用作实现医院管理的规范化与电子化

数据治理分析

- 对各类医疗系统中积累的海量数据进行结构化、标准化治理，结合应用场景进行筛选处理以提升数据质量，并进行数据分析及可视化应用

电子病历 | 影像数据 | 检验数据 | 费用数据 | 体检数据 | 药品数据

医疗数据主要包括**医院数据、健康数据与基因数据**，其中**医院数据规模最大**，而基因数据和健康数据则是增速最快的数据类型，健康数据的主要来源为在线问诊平台和健康智能设备

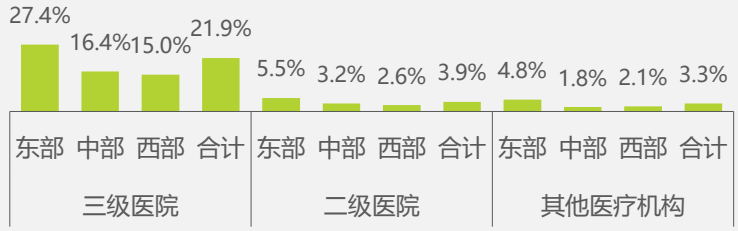
医院数据中**电子病历和影像数据**是最核心的数据资源

- 电子病历囊括的数据非常丰富，包含患者基本信息、检验数据、诊断数据、治疗数据等。其中非结构化数据占比较大，将其转化为适合计算机分析的结构化形式是挖掘其数据价值的基础
- 医学影像数据量庞大、增速快且标准化程度高，是AI医疗领域应用成熟的数据类型

数据智能应用

- 智能化将基于医疗信息化过程中积累的丰富数据、数据开发、应用场景等为基础，深刻地结合AI技术，在疾病的预防、诊断、监测及长期管理等各个环节更好地辅助医生诊断与治疗，大幅提升医院的服务能力和质量

2021年中国医院智能应用现状



- ✓ 中国正由AI医疗的**起步期步入发展期**，医疗数据互联互通建设进一步展开，认知智能技术迈向成熟，与感知智能协同探索、推进更多AI应用均衡互补发展
- ✓ 在各级医院中，**三级医院是智能化推进最快的主体类型**；从地域来看，**东部地区的医疗智能化应用渗透率更高**

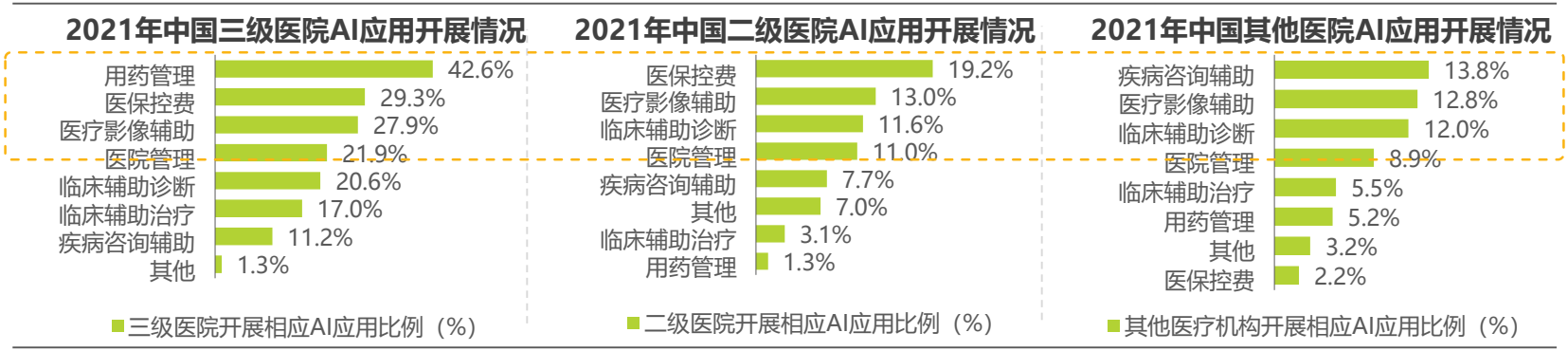
来源：国家卫生健康委统计信息中心《全民健康信息化调查报告》，艾瑞研究院自主研究绘制。

高频高价值场景（1/2）

医学影像与医保控费是院内AI应用最成熟的领域

AI医疗是挖掘医疗大数据的关键应用，已从医院管理类应用逐渐延伸至医院核心业务，可有效提升医院管理水平与诊疗水平。其中，医院管理信息化与辅助类诊疗应用是热门方向，针对影像、乳腺癌等专科诊断类产品已经率先落地。从AI应用在我国不同等级医院中的开展情况来看，用药管理在三级医院中渗透率最高，医保控费在公立医院有较高的普及率，AI辅助诊疗则在各类医院中均有较高应用，尤其下级医疗卫生机构，更加需要利用AI技术弥补各级医院之间能力差异，对建设落地我国分级诊疗体系有重大意义。

中国医院内高频高价值应用场景



院内

用药管理

实现**药品的全流程智能化管理**，能够节约资源、防范差错、促进信息互通，主要应用于医院门诊药房、静配中心，目前正在向院外零售药房等多个场景延申

- **智慧药房**：药房药品的自动化存储、调配、传送和发放
- **智能化静配中心**：智能化静脉输液药物的存储、配置、复核等

库存查看 实时动态　报表统计 数据挖掘　……

医保控费

通过**AI智能分析病历数据和费用数据**，对医院、药店、医生、参保人进行“线上+线下”、诊疗全流程与费用行为的**全方位审核**，确保基金使用合法、合规、合理

风险模型实时监控+医保智能审核

- 诊疗过程行为：智能提示与干预，促进诊疗行为合规；病历数据 病案数据
- 诊后分析：多维度费用智能审核；费用数据

风险场景特征识别　知识库

来源：国家卫生健康委统计信息中心《全民健康信息化调查报告》，艾瑞研究院自主研究绘制。

高频高价值场景（2/2）

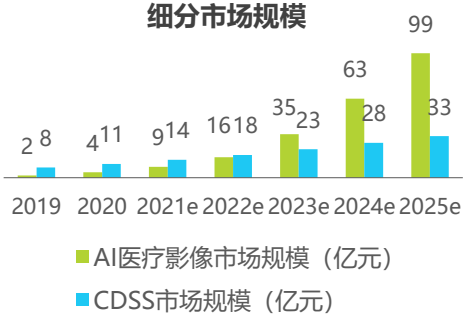
基因检测与新药研发是院外AI应用最主要的应用领域

随着医疗服务向整体大健康的范围拓展，AI在院外的应用场景同样极具潜力。基因测序能够通过对生物的DNA的检测，提前预知疾病的发生的概率，是AI疾病预测的重要应用场景，近年来我国基因测序市场增长迅速，逐渐被消费市场认可。AI制药是医疗领域另一个发展迅猛的应用场景，一方面由于传统制药成本增加，效率面临挑战，驱动行业在技术层面寻求突破方法，另一方面AI擅长的数据处理和分析技术与新药研发场景高度融合，共同促成了AI制药赛道的升温。

中国医院内院外高频高价值应用场景

院内

2019-2025年AI辅助诊断细分市场规模



医疗影像辅助

智能CT影像利用深度学习技术自动学习资料各层次抽象特征来**分析医学影像并给出辅助诊断结论**，协助医生完成病例筛查、分析诊断。其能够快速、准确、批量地处理大量影像资料，对不同疾病行准确诊断、精准分割、分类和预后，有效提高临床诊断的效率和准确性

图像识别 | 深度学习

医学原始图像 / 非结构化数据 → 基于二维、三维影像提取准确的结构信息 → 通过大量影响数据与诊断数据学习训练相应算法，给出可靠诊疗和治疗方案

临床辅助诊断

AI辅助临床诊断是融合了**自然语言处理、认知技术、自动推理、机器学习、信息检索等技术**，给予假设认知和大规模的证据搜集、分析、评价，提供专科到全科多病种智能诊疗的人工智能系统，能够有效减少医生误诊、漏诊的情况，提高医疗质量和安全，控制费用成本

文献类数据 / 实践类数据 / 临床类数据 / …… → 自然语言处理技术实现结构化数据 → **建立核心知识图谱** → 针对临床思路抽取成机器可理解和执行的诊疗逻辑，构建知识图谱，通过算法综合分析各类数据，生成诊断结论和治疗方案建议

院外

AI+基因检测

基因信息 → 定序、分析 → 资料库比对 → 机器学习

通过对公共数据库大量非结构化数据进行学习和整合，挖掘并计算其中的关联，更新突变位点和疾病的潜在联系

AI+基因检测是针对生物的血液、体液、细胞等通过AI等技术手段进行DNA检测，并分析该个体含有的基因是否含有潜在高风险疾病可能性及表达功能是否正常，最终明确被检测生物未来患病风险的技术，**NIPT、肿瘤检测、消费基因、预知疾病**等都是AI基因检测的重要应用场景

AI+制药

研究开发 → 临床试验 → 药品生产

- 研究开发：靶点筛选、药物挖掘、药物优化
- 临床试验：药物依从性
- 药品生产：生产质控

80%的AI应用集中于研发阶段

医院/药企/CRO/实验室/公开资料等各类数据集 → AI/物理模型 → 模型训练

AI制药以医药大数据为学习研究土壤，运用**NLP、CV、知识图谱、机器学习、深度学习等AI技术**参与制药过程，去计算、预测、挑选合适的化合物、潜在药物分子并观察药物临床效果，**帮助药企缩短研发周期、节约研发成本、降低风险、提高制药成功率**

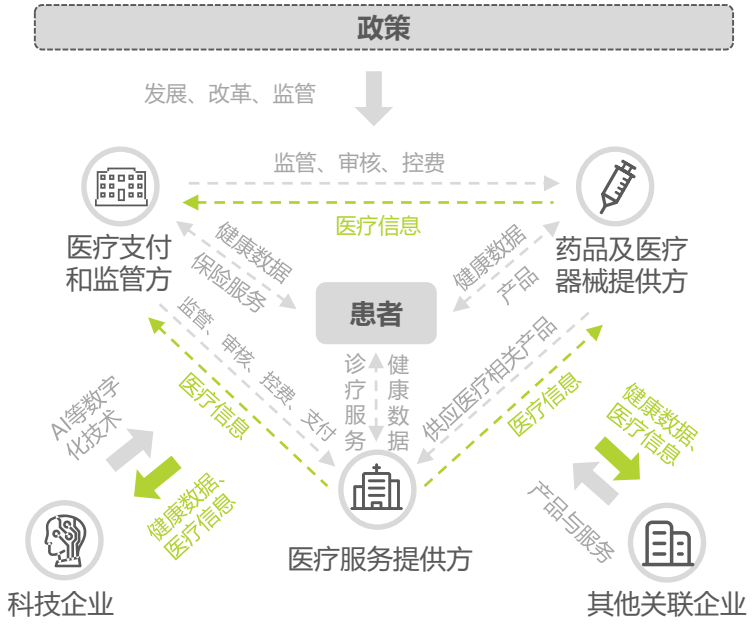
来源：《2021年中国人工智能+医疗与生命科学行业研究报告》，艾瑞研究院自主研究绘制。

数据痛点与核心诉求

数据质量是掣肘AI医疗应用发展的核心痛点

AI医疗发展的驱动力在于满足医疗健康服务发展的刚性需求，通过数字化建设以人为本的整合型医疗卫生服务体系，但受医疗资源分配不均等历史遗留因素的影响，目前制约AI医疗发展的原因有很多：从医院层面看，医疗信息化建设支持了医疗数据的爆炸式增长，但是医疗数据在流通、共享、存储、管理等环节尚未标准化，导致数据多源异构难汇集、数据标准体系不健全等问题始终存在，掣肘着AI应用乃至行业的发展。从院内院外整个医疗数据环境来看，除数据质量欠佳之外，数据量的欠缺也制约了AI模型训练的效果。从政策层面来看，我国没有明确的法律规定数据归属问题，各方在使用医疗数据时需考虑数据安全与潜在风险，数据开放受限，数据的价值难以得到体现。

医疗行业在AI应用过程中的核心痛点



数据量与数据质量

- **数据质量**：医院原有的信息系统支持了医疗数据的爆炸式增长，但大部分信息资源仍在信息库中“沉睡”，且整体质量偏低、非结构化数据占比高、数据孤岛问题普遍存在，难以直接支持AI建设
- **数据量/数据开放受限**：现有健康数据主要来源于院中就诊数据，是以医疗行为为中心的诊疗数据，缺乏以人为中心的全场景医疗数据整合，包括体检数据等日常健康数据。在院外的医疗外延场景下，大部分数据来源于文献、实验与合作方，数据量不高且结构化难度高，高质量数据的缺乏大大影响了AI效果

数据安全

- **数据安全及隐私**：健康数据包含了换这个人的健康信息，如何合理利用数据提升医疗水平的同时，保证用户隐私性，需要进一步明确数据归属、权限等问题

来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理

以数据治理为抓手，夯实数据基础，打造持续运营体系

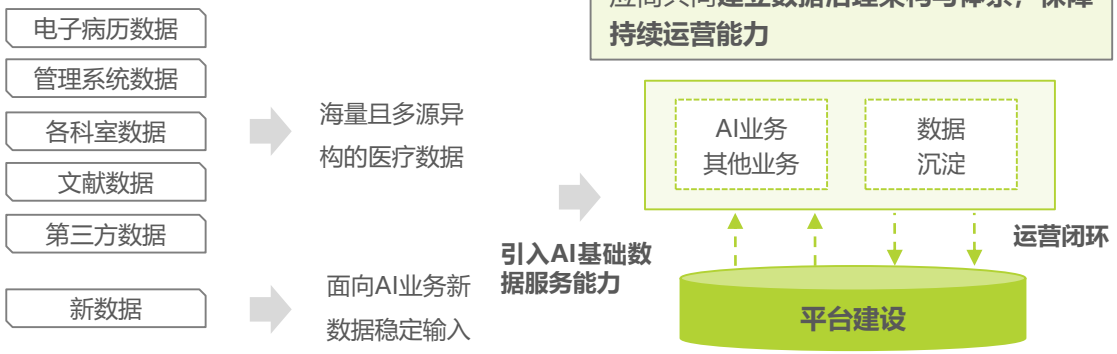
医疗行业数据治理的市场起步较晚，大部分主体并未建设专门的数据治理平台及体系，仅仅只是单点的产品应用。结合医疗行业在AI应用过程中的数据痛点，医疗行业数据治理体系的搭建应该以打好数据基础为主，具体而言，首先需要促进全域数据的集成，并且引入AI基础数据服务能力对提升数据质量，保障数据稳定、持续治理。其次，医院应以平台建设为主，在满足医院管理、运营的基础之上，以场景做切入，支撑AI业务单点的运营。数据管理规范与安全保障需要在此过程中确立，最后，以建立持续迭代运营的数据治理体系为核心，搭建切实可行的执行通道，形成数据闭环。

医疗行业面向人工智能的数据治理体系搭建

Step1：夯实基础 ★
在现有数据布局下提升数据质量是AI发挥作用的基础，此外，尽可能以数据多样性为指导，引入全域数据

Step2：选择场景、共建运营能力
医疗行业的IT能力普遍相对较弱，AI建设多为**单点建设**。在确定AI应用场景之后，应**以平台建设为主**，引入治理工具，与供应商共同**建立数据治理架构与体系，保障持续运营能力**

Step3：数据管理规范与安全保障
在数据治理过程中必须要保障患者数据安全，数据的共享、接入、标准化、管理等过程中，需要**制定严格的管理规范**，数据治理与使用需在保证**数据持有者隐私、数据安全和保障医疗质量**的前提下进行



Step4：建立执行通道、迭代推广
医疗行业有大量历史数据中的有价值信息等待挖掘，医院在搭建了完整的数据治理体系之后，能否将能力持续运营是关系，最终达到以患者为中心的数据建设

核心建议：

- **通过数据上报标准、对接协议统一管理、构建完备的数据质量体系、构建数据血缘和业务血缘等措施，提升数据质量**
- **多方共同推进数据共享，采集与引入全业务、多终端、多形态的数据**

来源：艾瑞研究院自主研究绘制。

产业图谱

iResearch 艾瑞集团

医疗行业大数据智能产业图谱

公立民营疗机构 | 体检中心 | 卫健委 | 药企 | CRO | 医疗流通公司 | 康复中心 | 基因检测厂商

应用服务层

医院AI应用层

用药管理

健麾信息 GENERAL HEALTHY | swisslog | 深圳瑞驰 | BD

医保控费

国新健康 CRHMS | Neusoft | 万达信息 | 卫宁健康 WINNING HEALTH

医疗影像辅助

数坤科技 SHUKUN | 推想医疗 | UNITED 联影 IMAGING | 依图 YITU | 兰丁股份 LANDING MED | 深睿医疗 deepwise | ZYING 智影医疗 | 汇医慧影 huiyihuiying.com | 腾讯觅影 | Airdoc

临床辅助诊断

灵医智惠 | 卫宁健康 WINNING HEALTH | Neusoft | 惠每医疗 HUIMEI HEALTHCARE | 嘉和美康 GOODWILL | 医渡云 YIDUCLOUD | 森亿智能 SYNYI·AI | BaseBit.ai 翼方健数 | 颐圣智能 DOCTOR AI

基因检测

华大BGI | BerryGenomics 贝瑞基因 | 志诺维思 GENOME WISDOM | SINOMD 鑫诺基因

新药研发

药明康德 WuXi AppTec | 先声药业 Simcere | 豪森药业 HANSOH PHARMA | iCarbonX

AI技术服务

知识图谱：阿里云 | 灵医智惠 | 惠每医疗 HUIMEI HEALTHCARE | 医渡云 YIDUCLOUD

机器视觉：商汤 sensetime | 数坤科技 SHUKUN | 依图 YITU | 汇医慧影 huiyihuiying.com

自然语言处理：科大讯飞 IFLYTEK | 森亿智能 SYNYI·AI | 惠每医疗 HUIMEI HEALTHCARE | 零氪科技 LinkDoo

机器学习：4Paradigm 第四范式 | 阿里云 | 灵医智惠 | EMOTIBOT 竹间智能科技

数据平台层

医疗大数据

森亿智能 SYNYI·AI | 零氪科技 LinkDoo | China Unicom 中国联通 | 思派 Medbanks | 医渡云 YIDUCLOUD | 中国电信 CHINA TELECOM | 神州医疗

医院信息系统

HIS：Neusoft | 卫宁健康 WINNING HEALTH | B-Soft 创业慧康

CIS：嘉和美康 GOODWILL | 和仁科技 Heren Health | EWELL 思创医惠

其他：AliHealth 阿里健康 | 平安好医生 | JDH 京东健康 | 医联

外部数据源

AI基础数据服务

MAGIC DATA | 数据堂 Shujutang | speechocean 海天瑞声

医疗数据服务

第三方厂商数据 | 医疗知识库 | ...

场景包含面向AI的数据治理

注释：以上厂商与行业为不完全列举，排名不分先后。
来源：艾瑞研究院自主研究绘制。

工业

—— Industry

工业信息化建设阶段：历经蒸汽时代、电气时代和信息时代后，工业革命正逐步走入以大数据、人工智能、机器人等新型技术为代表的工业新时代。新一代信息技术逐步与工业制造业深度融合，引发影响深远的产业变革，中国工业已然迎来大数据驱动，与行业机理、知识经验相结合的智能化发展。

高频高价值业务场景：目前AI应用部署以浅层点状应用为主。计算机视觉应用率先展开落地，为企业提供工业机器人的视觉引导和工业产品的视觉质检等功能，助力企业提高生产制造效率与产品质检准确率。基于工业企业数据平台的逐步成熟和应用痛点的需求驱动，目前与经营运维相关的AI应用已热度渐起，为工业企业的AI应用体系建设“添砖加瓦”。

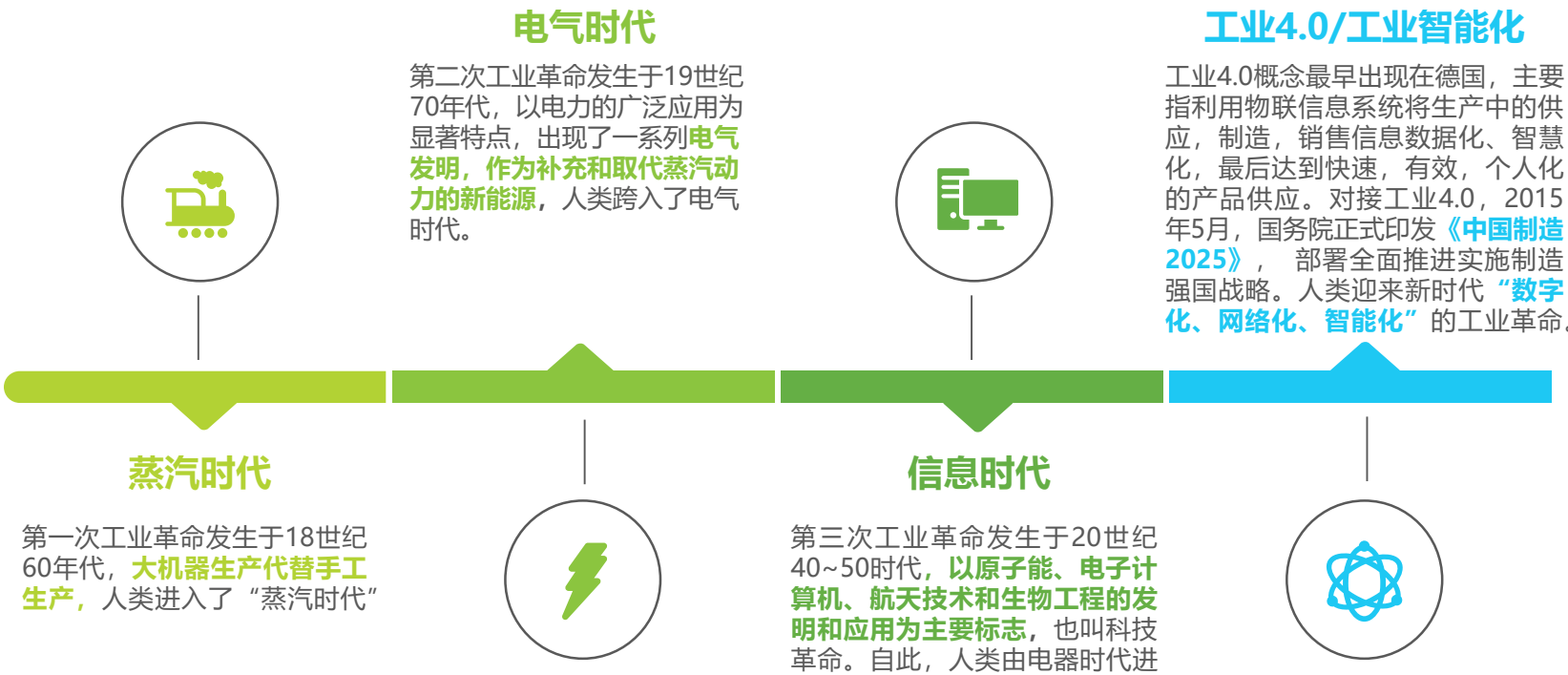
面向人工智能的数据治理：工业企业数据在原本庞大的体量基础上仍在飞速增长。但在支持上层应用时，工业大数据在来源多样性、数据时序性和机器复杂性上面临诸多痛点，工业企业需要建立完善的数据治理体系，在优质数据基础上发挥出工业数据智能应用的价值潜能。目前中国工业整体数据治理水平处于滞后状态。对于数字化转型现行、资金实力雄厚、IT支持力度强的工业企业来说，不断加强自身数字化转型程度，在AI平台搭建和算子丰富的同时，也需从平台级能力考量，搭建整体面向AI的数据治理体系；而对于数据基础薄弱、业务需求不清晰、IT支持力度弱的工业企业来说，寻求自身典型场景的AI应用落地并开展针对性范围的AI数据治理工作，为目前数字化智能化升级的首要任务。

行业发展背景

大数据驱动的，与行业机理、知识经验相结合的智能化发展

历经蒸汽时代、电气时代和信息时代后，工业革命正逐步走入以大数据、人工智能、机器人等新型技术为代表的工业新时代。新一代信息技术逐步与工业制造业深度融合，引发影响深远的产业变革。2015年5月，国务院印发《中国制造2025》，中国开始打造自己的“工业４.０”蓝图，全面推进“制造强国”战略，中国工业已然迎来大数据驱动，与行业机理、知识经验相结合的智能化发展。

工业革命的演变和发展趋势



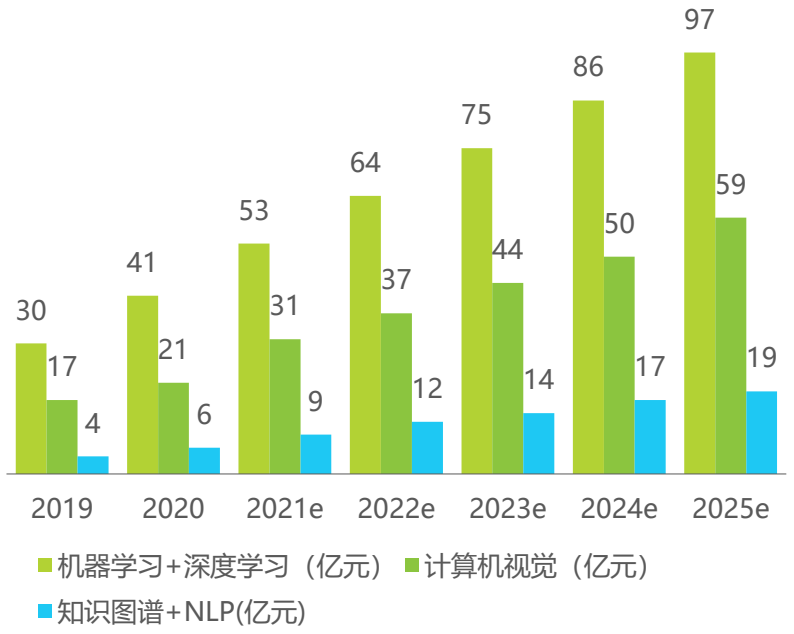
来源：艾瑞研究院自主研究绘制。

高频高价值业务场景

部署以浅层点状应用为主，视觉先行，经营运维热度渐起

2021年，以机器学习与深度学习、知识图谱、NLP、计算机视觉为技术主导的中国工业智能应用核心产业规模为93亿元。其中，计算机视觉应用率先展开落地，为企业提供工业机器人的视觉引导和工业产品的视觉质检等功能，助力企业提高生产制造效率与产品质检准确率。而在经营管理、运维服务领域，AI应用从赋能环节和带动价值角度均存在广阔发展空间。但同样经营运维AI应用在落地时的部署难度较大，需同时保证工业数据和工业机理的成熟度。基于工业企业数据平台的逐步成熟和应用痛点的需求驱动，目前与经营运维相关的AI应用已热度渐起，为工业企业的AI应用体系建设“添砖加瓦”。

2019-2025年中国工业智能核心产业规模



来源：《2021年“新基建”背景下中国工业互联网与工业智能研究报告》，艾瑞研究院自主研究绘制。

AI+工业高频高价值应用场景分布

来源：艾瑞研究院自主研究绘制。

数据痛点与核心诉求

工业数据具备多样、时序与复杂性特征，海量线下待挖掘

如今，在工业企业数字化转型升级、工业互联网飞速发展的时代背景下，工业企业数据在原本庞大的体量基础上仍在飞速增长。但在支持上层应用时，工业大数据在来源多样性、数据时序性和机器复杂性上面临诸多痛点，工业企业需要建立完善的数据治理体系，在优质数据基础上发挥出工业数据智能应用的价值潜能。

工业高频高价值应用下的数据痛点

来源多样性

设计、加工、测试、维护... 客户、供应商、营销...

产品数据 | 市场数据

生产数据 | 外部数据

业务管理、生产、采购... 行业数据、友商数据...

来源多 **设备种类多、接口复杂** **海量**

工业数据来源众多，不仅有企业经营管理数据，还会涉及到工业的生产制造数据及传感器设备的海量数据等，且数据采集设备种类多、接口复杂。**如何打通数据，让数据汇聚并在治理后得到流转应用是工业大数据的起步核心痛点。**

数据时序性

设备传感器 时序数据

"时间戳、时效性"

海量设备测点 | 数据采集频率

高并发高吞吐 | 支持实时应用

规模级的写入、存储和计算能力 **流式计算**

在产品的设计加工和生产制造流程中，工业企业需对时序数据的采集、存储、查询、处理和分析帮助企业实时监控企业正常的生产经营过程。工厂在数字化转型过程中将产生巨大量级的时序数据，因此**对时序数据的测点范围选取、采集频率考量、高并发高吞吐能力、能否支持云边协同及实时应用的数据需求带来了技术挑战与核心诉求。**

机理复杂性

业务联动规律 | 专业机理模型

步骤精细复杂 | 长产业链条

复杂工业机理的融合应用 **数据链条联动**

工业产业链条长及工业机理复杂，**需了解大量工业技术原理、行业知识、基础工艺等，十分强调赋能者对行业背景的理解。**

海量线下数据：工业企业属于典型的非数字原生企业，产业链条长且多业态并存，在工业生产制造等各个流程中沉淀着大量的复杂数据。而目前工业的数字化进程不一，仍有大量中小企业停留在原始线下生产阶段，海量线下数据尚未被收集汇聚，价值有待挖掘。

来源：艾瑞研究院自主研究绘制。

面向人工智能的数据治理体系

数据基础不一，按需判断是否匹配平台级数据治理能力

综合来看，工业AI应用可从内部经营管理角度降低产品生产运营成本，从外部市场分析角度提升产品附加值和市场竞争力。但工业数据体量庞大，对应的数据治理投入也随之提升，目前中国工业整体数据治理水平处于滞后状态。对于数字化转型现行、资金实力雄厚、IT支持力度强的工业企业来说，不断加强自身数字化转型程度，在AI平台搭建和算子丰富的同时，也需从平台级能力考量，搭建整体面向AI的数据治理体系；而对于数据基础薄弱、业务需求不清晰、IT支持力度弱的工业企业来说，寻求自身典型场景的AI应用落地并开展针对性范围的AI数据治理工作，为目前数字化智能化升级的首要任务。

针对性的数据治理工作

企业画像

- 数字化转型先行
- 资金实力雄厚
- IT支持力度强

数字化转型进程与数据基础

典型工业企业类型一

大型能源企业，如电力企业、油气企业、智能制造企业...

数字化先行企业，数据基础逐步优化，具备较强AI能力建设，通常偏向于采购平台或算子，结合产业需求、行业场景搭建自己的智能应用服务。该类企业在进行AI模型与算法训练采购与AI应用落地搭建时，**对数据整合能力、数据治理与清洗能力、模型训练能力，优化迭代能力及灵活部署能力有同步强烈需求**，会在数字化转型过程中**同步进行平台能力搭建和数据治理体系完善。**

企业画像

- 数据基础薄弱
- 业务需求不清晰
- IT支持力度弱

数字化转型进程与数据基础

典型工业企业类型二

中小型工业企业...

从政策推动和自身收益角度出发，中小工业企业也迫切希望通过数字化转型升级提升生产效率和提高产品质量，但普遍面临人才不足、基础薄弱、经费短缺的问题。另外**大多企业没有数据中台/数据平台基础，对AI应用需求上不清晰**，需供给侧和需求侧共同摸索，供给侧做好场景落地引导，从可解决企业痛点难点和驱动经济效益的**高频高价值业务场景需求出发，开展AI应用落地试点和对应范围的数据治理工作。**

来源：艾瑞研究院自主研究绘制。

产业图谱

iResearch 艾瑞咨询

工业行业大数据智能产业图谱

下游

工业智能应用

- 研发设计：SIEMENS、DASSAULT SYSTEMES、ZWCAD中望、安世亚太 PERA CHINA
- 生产制造：SIEMENS、傲林科技 ALL IN TECH、宝信软件 BAOSIGHT、SIE、4Paradigm 第四范式
- 经营管理：用友 yonyou、傲林科技 ALL IN TECH、4Paradigm 第四范式、明略科技 MININGLAMP TECHNOLOGY、DHC 东华软件
- 运维服务：ABB、树根互联 ROOTCLOUD、傲林科技 ALL IN TECH、4Paradigm 第四范式、明略科技 MININGLAMP TECHNOLOGY、HUAWEI

AI技术服务

- 计算机视觉：MEGVII 旷视、阿里云
- 自然语言处理：科大讯飞 iFLYTEK、达观数据 DATA GRAND
- 语音技术：科大讯飞 iFLYTEK、阿里云
- 机器学习：4Paradigm 第四范式、百度智能云
- 知识图谱：明略科技 MININGLAMP TECHNOLOGY、海致星图

平台层

工业互联网平台/工业大数据平台

HUAWEI 华为云、百度智能云、腾讯云、阿里云、美林数据 MERITDATA、GRIDSUM 国双、傲林科技 ALL IN TECH、树根互联、卡奥斯 COSMOPlat、航天云网 CASICloud、徐工信息、智通云联 SMARTECH

工业数据治理服务

傲林科技 ALL IN TECH、4Paradigm 第四范式、明略科技 MININGLAMP TECHNOLOGY、美林数据 MERITDATA、PRIMETON 普元、亿信华辰 ESENSOFT

工业企业内部数据源

工业设备

- 智能终端：传感器、智能网关、芯片、机器视觉
- 生产设备：数控机床、智能设备、工业机器人
- 嵌入式软件
- 工业IDC

企业管理系统

ERP、MES、财务、OA、CRM、HR

工业企业外部数据源

互联网数据

市场数据

……

场景包含面向AI的数据治理

注释：以上厂商与行业为不完全列举，排名不分先后。
来源：艾瑞研究院自主研究绘制。

前言：数据与数据治理	1
主题：面向人工智能的数据治理	2
参与：行业规模与受益圈立足点	3
实践：高频高价值应用及数据痛点	4
案例：标杆企业与新锐势力	5
展望：治理陷阱与趋势洞察	6

第四范式

提出AI数据的“3C标准”，让已落地的数据治理可实践复用

第四范式致力于解决企业智能化转型中面临的效率、成本、价值问题，帮助提升企业的决策水平。据第四范式经验，数据治理是企业AI转型的一大阻力，在企业落地AI的过程中占据了高达95%的时间。企业要获得高质量AI数据，需要将业务、IT、AI三类know-how均纳入考虑，这使得数据治理工作的效率、效果和复用性受限。由此，在指定业务场景下包含数据从采集到使用的完整数据治理过程中，第四范式构建管理出一套“3C”数据形式标准，基于数据形式提供面向AI场景的数据，标准化规范化数据准备过程，全面提升AI数据的质量、效果和准备效率。

AI数据痛点

数据难以一致

- 线上线下数据不一致，导致模型效果差
- 采用人工方式进行比对，人力消耗巨大

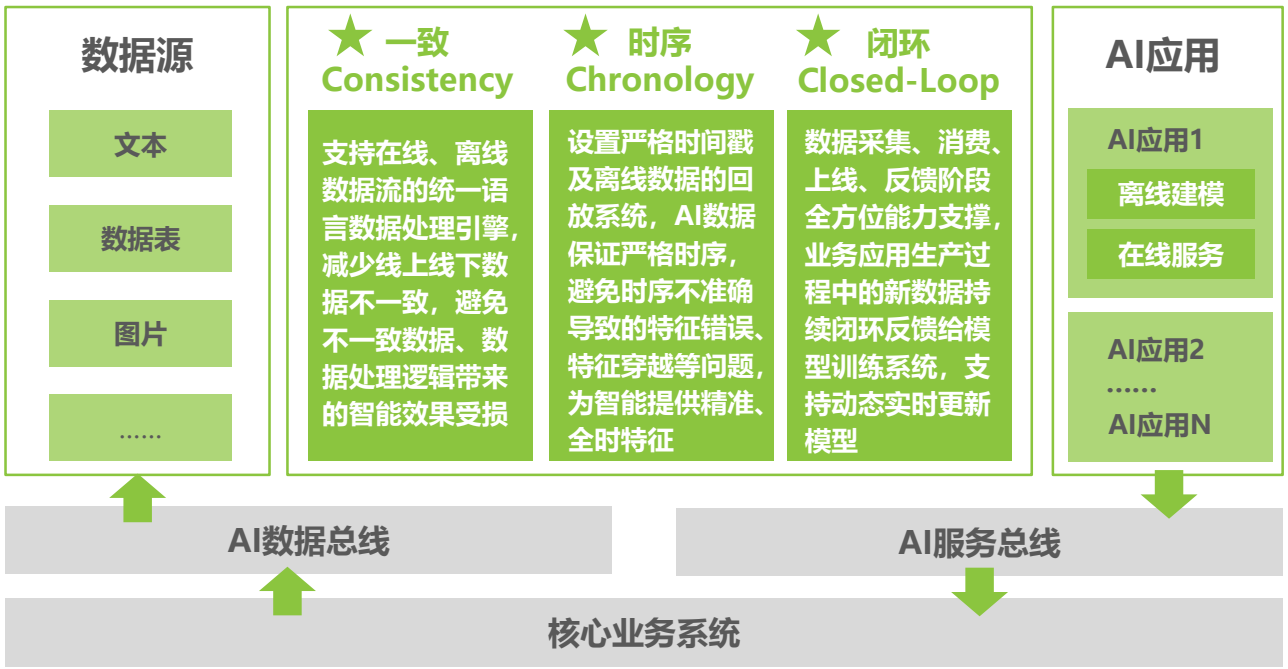
无时间属性

- 数据缺乏时序特征维度，特征计算不具备严格的时间戳信息，离线数据不支持回放，线上效果差，穿越隐患难查

缺乏数据闭环

- 建模过程无闭环反馈机制，无法支持模型实时迭代更新
- 模型随时间推移衰减

第四范式-AI数据的“3C”标准



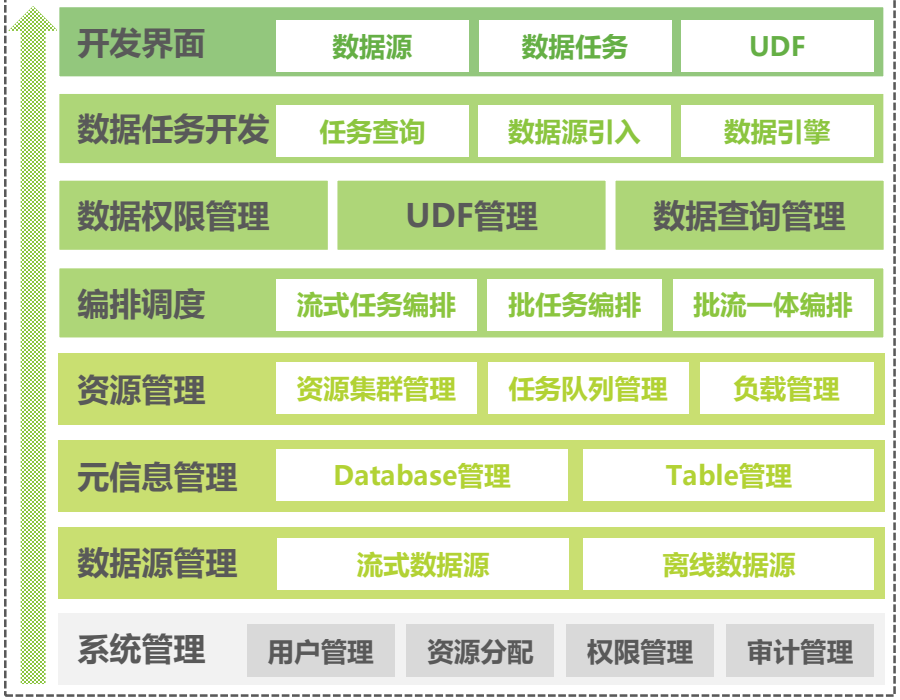
来源：艾瑞研究院自主研究绘制。

第四范式

构建AI数据治理平台，批流一体架构支持多元场景落地

对应AI数据的3C标准，第四范式构建出批流一体架构的AI数据治理平台，支持AI落地全过程。第四范式旨在通过产品形式降低AI上线过程中的数据成本，避免AI应用上线过程中存在的数据问题，以服务业务场景这一终极目标出发，为企业梳理并搭建数据治理体系。AI数据治理平台不仅支持批流数据的对接和编排，还会基于数据权限管理、UDF管理和数据查询管理模块提供数据使用和复用的保障。如今，AI数据治理平台已成功在零售餐饮、金融银行和工业制造业等多行业场景落地实践，未来将沉淀更加丰富的数据形式，承载越来越多的 AI 应用，持续赋能企业智能化转型变革。

批流一体产品架构



批流一体产品业务实践

某餐饮连锁巨头-推荐中台

- ✓ **实时特征开发处理：**实现推荐场景下物料库存、优惠的实时计算，作为实时特征进行模型预估；
- ✓ **实施数据准备监控：**实现数据处理流程的实时监控，对异常数据业务发起监控告警；
- ✓ **实时数据治理管理：**实现对特征质量的管理，对缺失率高于阈值的数据处理出发监控告警；

某股份制银行-实时交易反欺诈

- ✓ **用户近实时交易数据计算：**聚合用户近几笔交易的时间、地点，作为实时特征参与模型预估；
- ✓ **用户特征表数据回流：**经过特征工程处理的特征表通过实时数据回流，实现模型自学习数据留存；

某制造业巨头-实时工业信号处理

- ✓ 对设备发生的工业信号**实时进行数据收集与回流**，并实现**初步计算统计**，实现工业信号数据的**数据化分析；**

来源：艾瑞研究院自主研究绘制。

明略科技

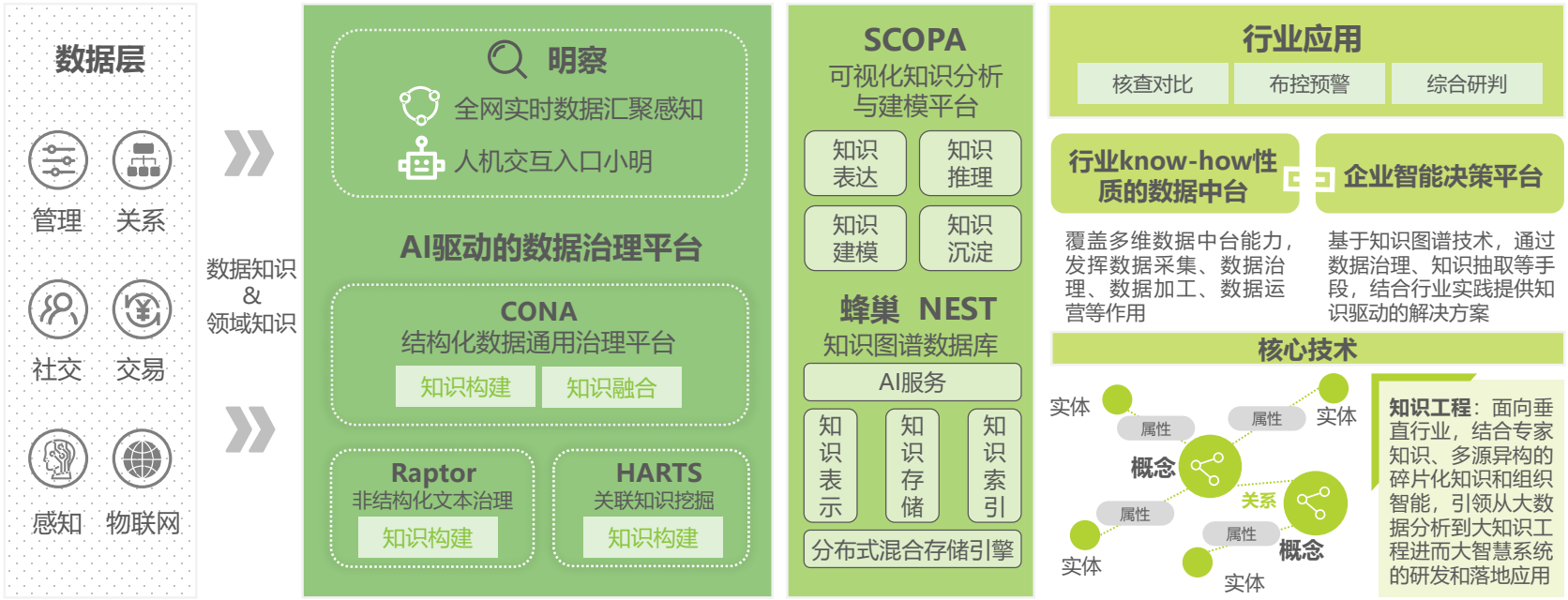
明略科技®
MININGLAMP TECHNOLOGY

iResearch
艾 瑞 咨 询

国内领先的企业数据智能应用软件供应商

明略科技是国内领先的企业数据智能应用软件供应商，通过大数据分析、挖掘和认知智能技术助力企业数字化转型。明略科技拥有数据中台和知识图谱等核心产品及解决方案，结合明略科技在大量实践中积累的行业“know-how”能力，能够帮助企业加速创新、打造核心竞争力，赋能数字化转型。目前，明略科技已向政府、银行、保险、证券、轨交、电力、制造、融媒体等领域100+行业的标杆客户展现了服务落地实力，通过行业知识与数据智能的结合，释放数据的业务价值。

明略科技核心产品架构与技术概览



来源：艾瑞研究院自主研究绘制。

明略科技

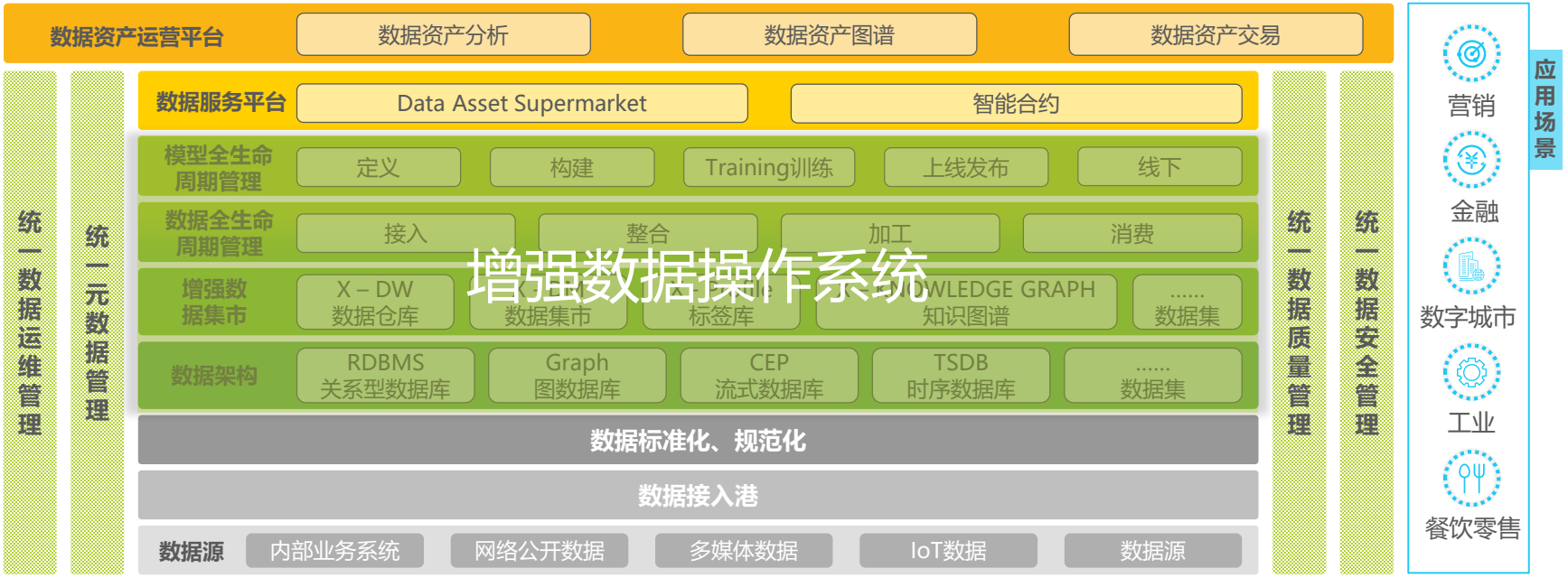
明略科技®
MININGLAMP TECHNOLOGY

iResearch
艾 瑞 咨 询

沉淀行业知识经验，助力企业人工智能落地

面对越来越复杂的数据来源与场景需求，数据治理在企业信息化转型过程中的重要性不断提升，越来越多的企业开始重视底层的数据治理工作。明略科技在知识提取层经过多年的行业模型积累与能力复用，打磨出一套自研的AI驱动数据治理平台，能够实现多维数据的采集以及符号化过程，并且具备对多源异构数据的实时处理能力，满足客户对实时性、高并发等场景的要求。不仅如此，明略科技还能为企业抽象出标杆型核心流程、定义相关标准，创新性地为企业开发降本增效场景，并按时、高效地实施落地，填补企业能力的空白、提升企业核心生产力。最后，明略科技在产品的工具性质基础上，将行业“know-how”逐步沉淀进基础工具中，变成标准模型，再基于知识图谱数据库完成数据和知识的汇聚、融合、推理及复杂运算，深入垂直领域为客户打造打通感知和认知层面的行业人工智能大脑。

明略科技数据治理平台架构



来源：艾瑞研究院自主研究绘制。

傲林科技

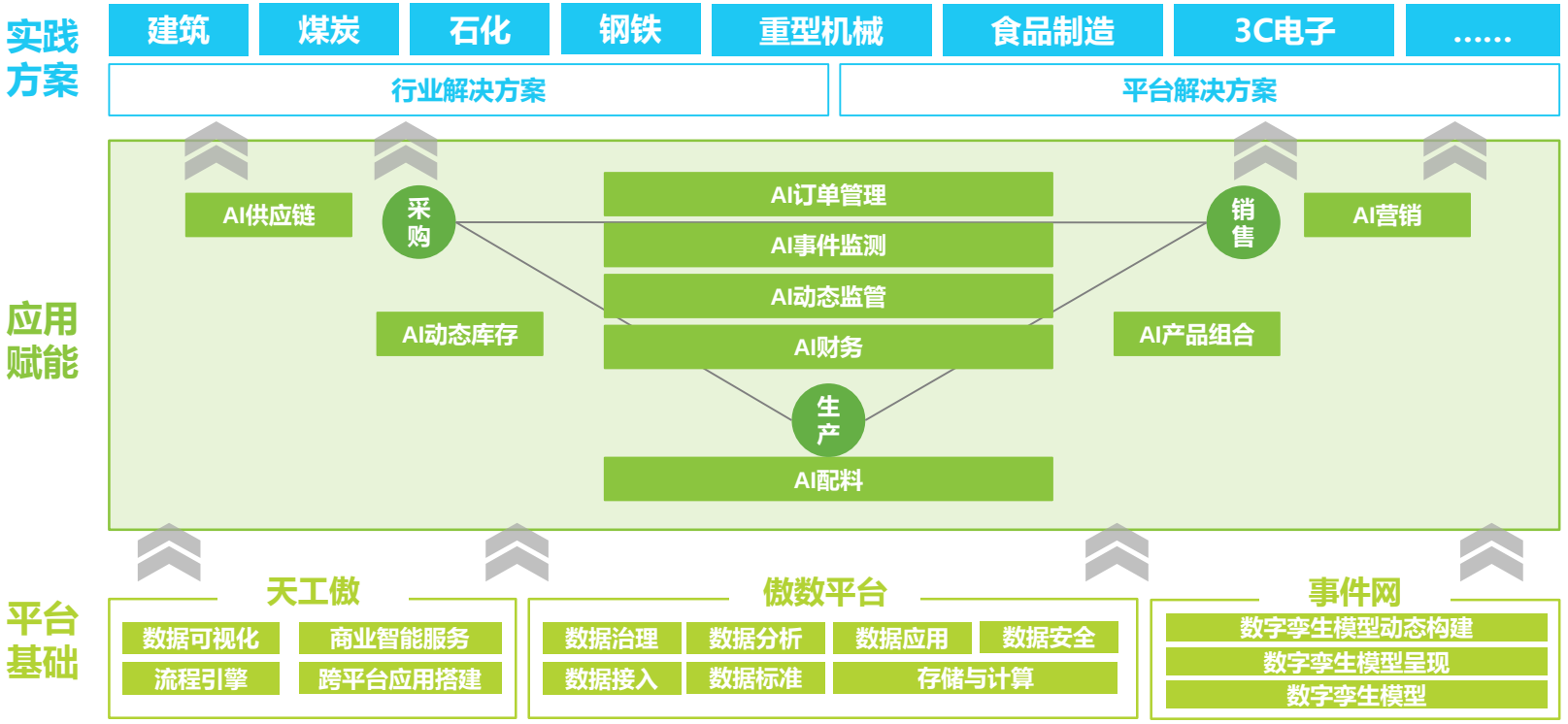
傲林科技 ALL IN TECH

iResearch 艾瑞咨询

打造“企业级数字孪生”，赋能企业数智化转型

傲林科技成立于2019年，是一家专注于利用大数据和人工智能技术，帮助企业实现数字化、智能化转型的高新技术厂商。傲林科技从采购、生产、销售“经营铁三角”入手，构建企业级数字孪生，全面打通企业传统信息化系统（MES、ERP、CRM、SCM等），打造企业整体生产经营智能优化解决方案，助力工业企业提升数字化竞争力，更好融入“双循环”新发展格局。目前傲林科技已拥有近400项服务于工业企业场景的算法模型，服务于钢铁、石化、消费品、汽车等行业，为企业客户带来数亿元效益。

傲林科技产品体系

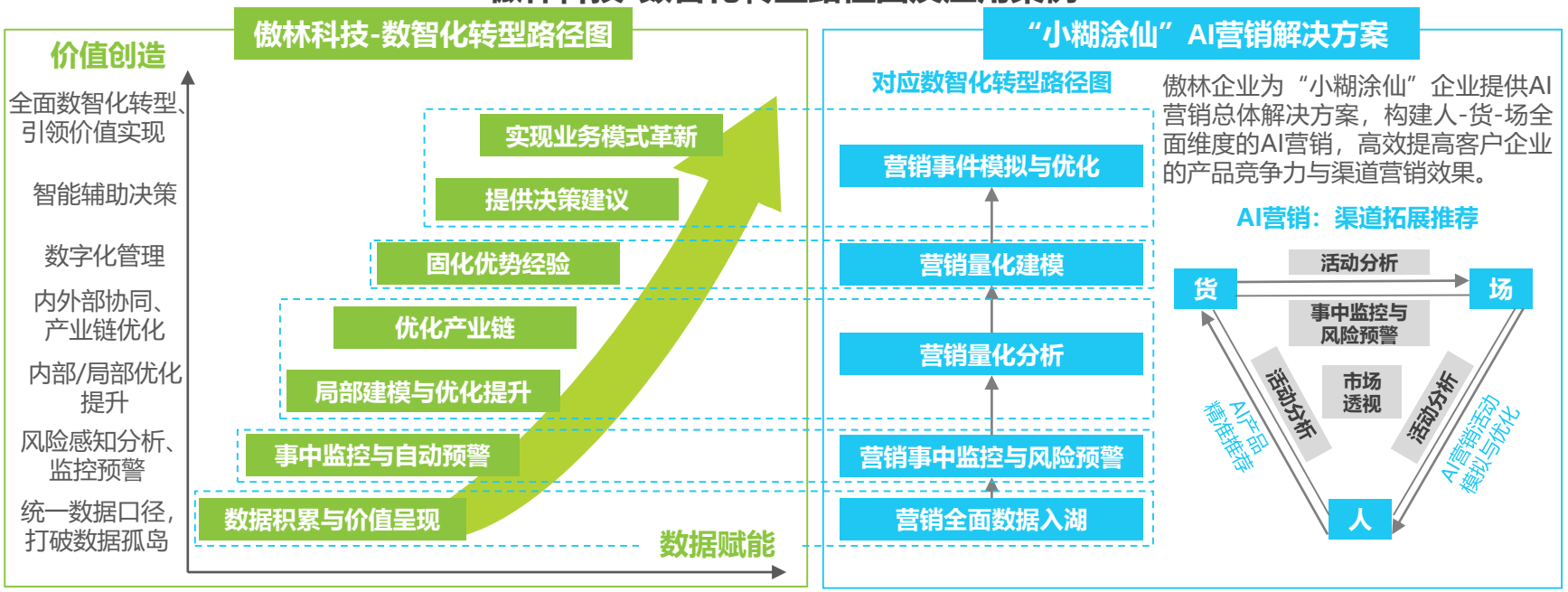


来源：艾瑞研究院自主研究绘制。

“自上而下”数字化转型方法论，让数据创造真正的价值

基于行业Know-How理解与项目经验积累，傲林科技总结出企业数智化转型路径图，为企业在数字化转型的道路上提供理论参考。傲林科技依靠数据＋AI算法，以企业管理层的困惑、痛点、需求为问题源点，以自上而下方法论为基础，“有的放矢式”而非“搭积木式”梳理重要的经营管理KPI指标，为企业提供全局化数字化转型解决方案。如今傲林科技已沉淀梳理出泛工业领域（钢铁、石化、煤炭、重型机械等）的行业综合指标体系，助力工业企业实现生产原料成本降低、采购成本降低、成品销售利润增加和库存资金占用减少等经营效果优化。另外，面对消费领域对于研发及营销等局部优化需求，傲林科技也可基于大数据和AI技术实现数据互通互联，为企业提供研发优化、渠道优化、精准营销和监控预警功能，帮助企业实现研发与营销数字化的转型升级。

傲林科技-数智化转型路径图及应用案例



来源：艾瑞研究院自主研究绘制。

Magic Data

全球领先的AI数据解决方案提供商

Magic Data是一家AI数据解决方案提供商，为从事自然语言理解、语音识别、语音合成、计算机视觉等人工智能领域研发与应用的企业、科研机构提供专业的AI数据解决方案。Magic Data AI数据解决方案覆盖智慧出行、智慧金融、智能社交、智能家居、智能终端等五大行业，面向AI机器学习提供三大核心产品，Annotator智能化标注平台、MD数据集、数据采集与标注服务，迄今已为国内外近两百家合作伙伴提供AI数据原油，业务涉及全球五大洲数十种语言。Annotator智能化标注平台(支持私有化、SaaS、云镜像)、MD数据集、与数据采标服务是Magic Data三大业务板块。三大业务基石相互辅助，为AI从业者提供多模态机器学习训练数据，帮助企业更好地落地AI应用，实现降本增效，并最终促进AI行业的快速发展。

- 通过专业的数据采集和标注服务，将多源异构的数据转化成可识别的AI训练数据
- 面向AI模型的数据采标
- 多模态多场景的AI数据解决方案

Magic Data三大核心产品覆盖五大行业

自有数据集

语音数据集 | 文本数据集 | 图像数据集

- ✓ Magic Data拥有超过20万小时语音数据集，覆盖60+语言，30+方言，专注研发自发对话式AI数据，拥有14万小时的自然对话式数据集，为机器学习赋能
- ✓ MD数据集有多维度标签、覆盖多场景、数据时效性、内容高精度和安全合规等特点

Annotator®智能化标注平台

- ✓ 拥有四大核心功能

多模态标注 | 任务可拆分 | 可视化管理 | 智能化辅助

数据采集和标注服务

数据采集 | 数据标注

- ✓ 平台支持私有化部署，**充分保障数据安全**的同时为企业提供高效可定制的数据标注服务
- ✓ 根据项目需求，提供**定制化的数据采集服务，加快项目进程**
- ✓ 提供**精细化的标注服务**，释放数据的价值

智慧出行：智能座舱 | 自动驾驶 | 智能营销

智慧金融：智能客服 | 智能营销 | 智能会议网点多模态交互

智能社交：内容审核 | 实时字幕翻译 | 智能推荐 | 语音转文本

智能家居：唤醒模式 | 远场交互 | 异常监控声纹识别

智能终端：虚拟助手 | 机器翻译 | 语音输入法

来源：艾瑞咨询研究院自主研究绘制。

Magic Data

以数据重塑生产力，从数据采标处理流程确保安全隐私合规

Magic Data为客户提供一站式的数据解决方案，包括制定方案/计划、数据采集、数据标注/处理、私有化部署等全链服务，将多源异构的数据转化成可识别的AI训练数据，为AI模型提供标准化数据集产品，提升模型的开发效率与应用质量。Magic Data采用人机协同等方式开发的数据产品，在确保数据质量核心能力的同时，也具备数据安全、数据多样化等显著优势。Magic Data从数据管理-团队管理-流程管理-工具模块搭建数据处理体系，提供完善标准的数据采标处理流程服务，并在数据处理过程中严格遵循GDPR法规与数据安全法等要求，加密与监控数据的整个数据生产流程，服务和产品通过ISO/IEC27701：2019标准认证，并由国际四大会计师事务所提供数据安全保护合规服务。在数据安全合规的治理趋向下，Magic Data严格从采标处理的数据源层面确保AI数据的安全隐私合规。

确保数据安全隐私合规

- 成为首批通过 ISO/IEC 27701：2019标准认证企业
- 由国际四大会计师事务所，提供数据安全保护合规服务
- 数据处理过程遵循GDPR法规
- 数据处理过程遵循网络安全法

加密与监控——数据生产流程

采集 → 清洗 → 标注 → 质检 → 交付

对数据进行实时审核

人工进行比例抽样

AI抽取：语音模型、NLP模型、机器学习

数据采标处理流程

工具模块	音频	文本	图像	视频
流程管理	标注	质检	验收	
团队管理	数据管理	团队管理员/验收经理	标注员/质检员/验收员	
	创建项目	管理团队成员	标注	
	质量监控	管理团队任务	质检	
	项目管理	团队绩效把控	验收	
数据管理	输入数据	数据安全处理/存储	输出结果	

来源：艾瑞咨询研究院自主研究绘制。

前言：数据与数据治理	1
主题：面向人工智能的数据治理	2
参与：行业规模与受益圈立足点	3
实践：高频高价值应用及数据痛点	4
案例：标杆企业与新锐势力	5
展望：治理陷阱与趋势洞察	6

数据埋点的大而全陷阱

抓大放小，从核心数据着手

数据埋点是指针对特定用户行为或事件进行捕获，处理和发送的相关技术及其实施过程，是数据治理中范围圈定的一环。出于对投资回报的考虑，客户往往倾向于做一个覆盖全业务和技术域的、大而全的数据治理项目，将每个数据都纳入到数据治理的范围中，这就导致进行数据埋点时放纵提需，埋点需求爆炸，给后续的数据治理和数据分析带来隐患。为避免数据埋点的大而全陷阱，企业应该做到抓大放小，谨记2/8原则——80%的问题产生于20%的系统和数据——从最核心的系统、最重要的数据、最容易产生问题的地方开始着手做数据治理。

数据埋点的大而全陷阱

什么是数据埋点

数据埋点是指针对特定用户行为或事件进行捕获，处理和发送的相关技术及其实施过程，是数据治理中范围圈定的一环。针对不同行为的埋点采集，从埋点在应用中的位置也可以区分成前端埋点、后端埋点等；从实现手段上划分，可分为：代码埋点、可视化埋点、全埋点等。

数据埋点的大而全陷阱

出于对**投资回报**的考虑，客户往往倾向于做一个覆盖全业务和技术域的、大而全的数据治理项目,每个数据他们希望都能被纳入到数据治理的范围中。这就需要尽可能捕获、处理和发送产品的大量事件的相对全面的属性，而这就挟持了厂商进行“大而全”的数据埋点，但“大而全”的数据埋点会随之带来许多问题。

潜在危害

大而全的数据埋点会占用大量空间和流量，数据治理后得到数据质量低，后续数据分析困难，从而沟通协作和整体项目推进出现问题。

- **实施难度高**
- **占用空间大**
- **数据质量低**

预防措施

埋点需求和设计需要有明确的提需规范和把控。

a.需求侧： 确立好自身AI应用的需求，以及需求对应下的数据范围。

b.供给侧： 充分了解客户的公司架构与数据结构，引导客户选定可行性高、实施性高的应用落地与数据框架

来源：艾瑞研究院根据专家访谈、公开资料研究绘制

数据治理体系的流转运营

沟通、组织、聚焦、文化

为能充分发挥数据治理的价值、避免一次性数据治理，供需两侧要齐心协力，共同、持续、优质地运营数据治理体系。数据治理是系统性工程，是由上至下指导，由下而上推进的体系工作。因此，供给侧企业与需求侧厂商，在体系运营和建设方面需形成共识，具备明确的目标、合理的组织、严格的监管、完善的系统，这样才能使数据治理工作得到保障，达到体系的流转运营。

供需两侧的流转运营

供给侧-数据治理厂商

事前事中建管一体： 供应侧在建设数据治理体系时需要做好与客户的沟通，对数据治理体系的建设和运营的全流程进行精细化管理，持续地将数据生产和数据管理做好对接，在数据治理的清洗、规范、化分和关联过程中做好流程控制，以最大限度的发挥数据治理的价值

事后质量监控： 供应侧可以通过建立一套切实可行的数据质量监控体系，监测流入的数据是否符合数据标准，以便于提早发现问题并对数据治理的规则、定义等作出相应的调整。目前只有部分头部厂商有类似的数据治理监控模块

上层目标明确： 需求侧要在企业发展战略框架下，建立数据治理的战略文化，意识到数据治理是一个长期的过程，明确数据治理的目标，使治理体系与企业发展战略和业务相匹配

治理组织合理： 建立合理的数据治理组织是保证企业数据治理体系健康运营的关键。需求侧在明确战略目标之后需要建设体系化的组织架构和合理的组织层次，并完善体系实施制度，确定中下层员工的职责分工，从上到下的指导数据治理运营工作的开展

需求侧-企业端

来源：艾瑞研究院自主研究绘制。

关注数据治理中的安全合规性

完善数据安全治理框架，确保数据安全合规

数据泄露事件在大数据时代层出不穷，随着行业新网络形态、新技术以及新应用场景的发展，新的数据类型、数据生产方式、数据处理方式和终端形式不断涌现，数据安全挑战也随之加剧。国家已出台各级各行业的法律法规及配套文件，不断加大数据安全与隐私保护的监管力度。对此，企业需建立符合企业管理现状及发展需求的数据安全治理框架，数据在采集、存储、传输、处理上均有对应的执行管理依据，做到挖掘数据资产、发挥数据价值的同时，确保数据全周期的安全与合规。

数据安全隐私相关的法律法规及配套文件

法律	《民法典》《网络安全法》《电子商务法》《数据安全法》《消费者权益保护法》《全国人大常委会关于加强网络信息保护的决定》《个人信息保护法》（草案）
行政法规	《征信业管理条例》《关键信息基础设施安全保护条例（征求意见稿）》
部门规章	《电信与互联网用户个人信息保护规定》《网络安全审查办法》《网络交易监督管理办法》《个人信息处境安全评估办法（征求意见稿）》
规范性文件	《2020信息化和网络安全工作要点》的通知
司法解释	《最高人民法院管理审理利用信息网络侵害人身权益民事纠纷案件适用法律若干问题的规定》
国家推荐性标准	GB/T 35273《信息安全技术个人信息安全规范》（2020）《信息安全技术个人信息安全影响评估指南》（送审稿）

数据安全治理参考框架

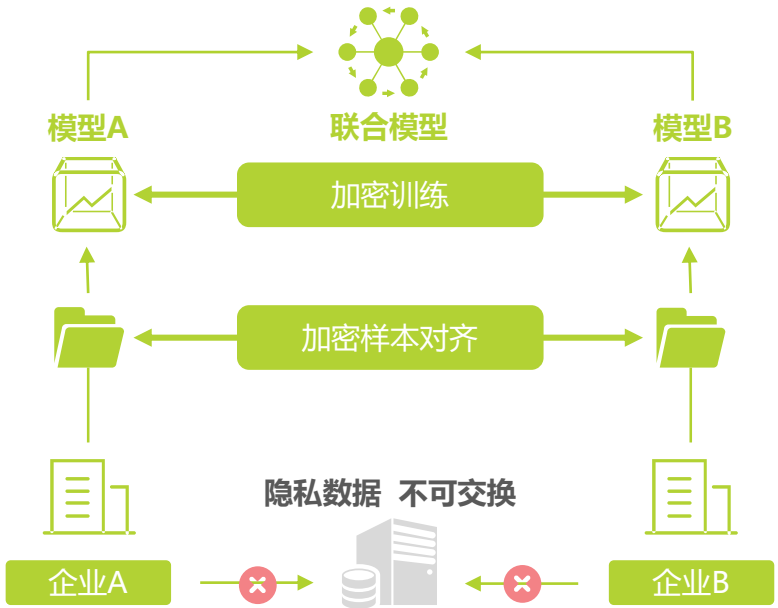
来源：《数据安全治理白皮书实践》，信通院，艾瑞研究院自主研究绘制。

联邦学习带来数据治理升华

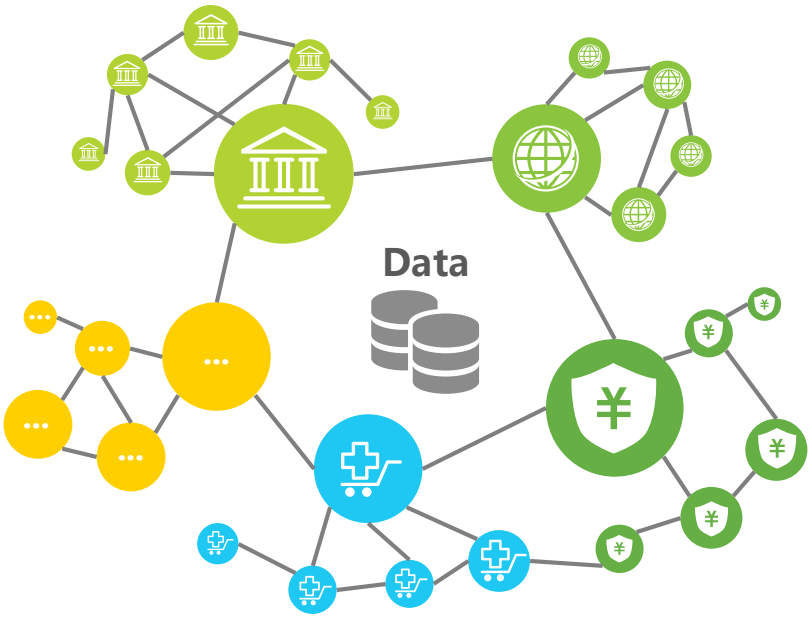
治理升华，数据安全合规线内的共同富裕

在数据治理及准备过程中，企业一方面需要尽可能全面的获取数据以扩充训练样本规模，另一方面出于隐私与安全的相关要求不能随意收集、融合和使用数据进行AI处理。为解决以上难题，联邦学习技术应运而生。联邦学习的建模原理为基于分布在多个设备上的数据集构建机器学习模型，通过安全多方计算、差别隐私、同态加密等技术为模型提供隐私保证以防数据泄露。因此，联邦学习可有效打通企业间的数据孤岛，并将数据可用而不可见，在满足数据安全合规的基础上，通过连通协同发挥出数据的更高价值。目前，联邦学习技术已成为大数据智能厂商的核心开拓方向，率先在金融、医疗和政务等领域展开应用。

联邦学习-建模原理



联邦学习-数据联邦



来源：艾瑞研究院自主研究绘制。

数据的“自治与自我进化”

将数据治理流程化、自动化、智能化

数据规模的指数级增长给数据治理工作带来巨大压力，传统人工方式做数据的清洗、分辨与调优使治理工作耗时冗长，带来高昂的人力成本，且愈发难以满足智能应用对数据在规模量与质量的高要求，传统的人工数据治理工作已变得捉襟见肘。如今，人工智能和RPA等技术手段已被逐渐应用于数据治理的模型管理、质量管理、资产管理、元数据管理等模块，最终实现数据系统的“自治与自我进化”。总体来看，前沿技术手段应用可以让数据治理工作趋于流程化、自动化与智能化，同时让数据变得可扩展、更负责可溯、更可信，已然成为未来数据管理发展的必由之路。

人工智能+RPA技术加载于数据治理全流程

注释：RPA，RPA是Robotic Process Automation（机器人流程自动化）的简称，是指可以模拟人类在计算机等数字化设备中的操作，并利用和融合现有各项技术减少人为重复、繁琐、大批量的工作任务，实现业务流程自动化的机器人软件。

来源：艾瑞研究院自主研究绘制。

打造“治理+AI”体系的良性循环

相互关联，互为依托，共同促进人工智能应用的内外发展

面向人工智能的数据治理充分利用机器学习技术，将数据治理环节自动化、智能化，可极大提升数据治理工作效率，同时基于自然语言理解和知识图谱挖掘关联非结构化数据的应用价值，解决数据质量管理的传统难题，使治理后的数据更加契合AI应用的要求，从效率和质量双侧推进AI模型的落地应用。另一方面，AI应用落地效果的显著优化也会给企业带来更多智能化转型信心，让其加大相关AI项目的预算投入，进一步推进了相关治理体系建设，打造“治理+AI”的良性循环。

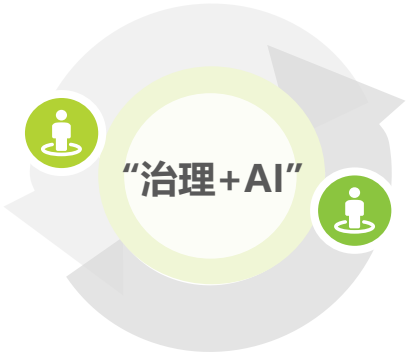
面向人工智能的数据治理-体系价值循环图

治理后的数据更契合AI应用要求，从效率和质量双侧推进AI模型的落地应用

AI技术”带动”面向人工智能的数据治理体系运作

机器学习+自然语言理解+知识图谱

- **机器学习：**机器学习模型可将数据治理环节自动化、智能化，减少数据治理的人力投入，提升数据治理的工作效率
- **机器学习+特征识别 → 自然语言理解：**除原有结构化数据的治理外，将非结构化数据纳入公司数据资产范围内，基于自然语言理解技术实现非结构化数据的理解、治理和价值挖掘
- **知识图谱：**知识图谱是企业以实现上层应用为目的构建，但在实际数据治理过程中，构建的知识图谱可帮助机器识别错误数据、挖掘关联数据，解决数据质量管理的传统难题



面向人工智能的数据治理体系”加速”AI应用落地

AI应用：机器学习产品+自然语言理解产品+知识图谱产品

- **AI应用开发过程：**统一开发标准，减少数据反复清洗的成本
- **AI应用知识沉淀：**数据治理所形成的业务发现沉淀到知识图谱里，助力AI应用构建
- **AI应用模型效果：**纳入多源异构和实时性数据，从AI模型的数据需求角度进行数据的质量优化和特征工程，提升数据与模型的契合度，优化应用落地效果

AI应用的优质效果带来AI预算投入追加，推进面向人工智能的数据治理体系建设

来源：艾瑞研究院自主研究绘制。

艾瑞新经济产业研究解决方案

iResearch 艾瑞咨询

行业咨询

- **市场进入**：为企业提供市场进入机会扫描，可行性分析及路径规划
- **竞争策略**：为企业提供竞争策略制定，帮助企业构建长期竞争壁垒

投资研究

- **IPO行业顾问**：为企业提供上市招股书编撰及相关工作流程中的行业顾问服务
- **募投**：为企业提供融资、上市中的募投报告撰写及咨询服务
- **商业尽职调查**：为投资机构提供拟投标的所在行业的基本面研究、标的项目的机会收益风险等方面的深度调查
- **投后战略咨询**：为投资机构提供投后项目的跟踪评估，包括盈利能力、风险情况、行业竞对表现、未来战略等方向。协助投资机构为投后项目公司的长期经营增长提供咨询服务

关于艾瑞

艾瑞咨询是中国新经济与产业数字化洞察研究咨询服务领域的领导品牌，为客户提供专业的行业分析、数据洞察、市场研究、战略咨询及数字化解决方案，助力客户提升认知水平、盈利能力和综合竞争力。

自2002年成立至今，累计发布超过3000份行业研究报告，在互联网、新经济领域的研究覆盖能力处于行业领先水平。

如今，艾瑞咨询一直致力于通过科技与数据手段，并结合外部数据、客户反馈数据、内部运营数据等全域数据的收集与分析，提升客户的商业决策效率。并通过系统的数字产业、产业数据化研究及全面的供应商选择，帮助客户制定数字化战略以及落地数字化解决方案，提升客户运营效率。

未来，艾瑞咨询将持续深耕商业决策服务领域，致力于成为解决商业决策问题的顶级服务机构。

联系我们 Contact Us

400 - 026 - 2099

ask@iresearch.com.cn

企 业 微 信

微 信 公 众 号

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，部分文字和数据采集于公开信息，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，艾瑞咨询对该等信息的准确性、完整性或可靠性作尽最大努力的追求，但不作任何保证。在任何情况下，本报告中的信息或所表述的观点均不构成任何建议。

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