人工智能领域数据处理解决方案

打造高品质的数据深度加工链

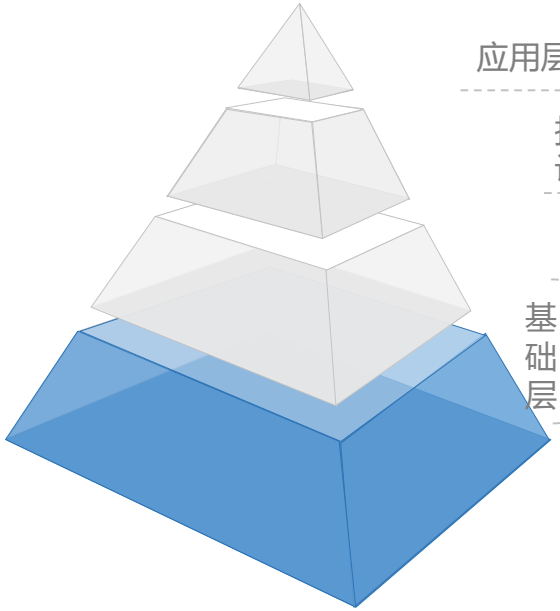
Baidu 百度 | 数据众包

百度数据众包 - AI基础数据提供商

数据之于人工智能

海量、精准、高质量的数据为训练人工智能提供了原材料

数据为人工智能技术的实现和人工智能应用的落地提供基础的后台保障！



应用层：身份识别、无人车、机器人等场景应用

技术层：机器学习、深度学习、语音识别、图像识别、人脸识别、NLP等。

计算能力：大数据、云计算、神经网络芯片等计算能力提供商

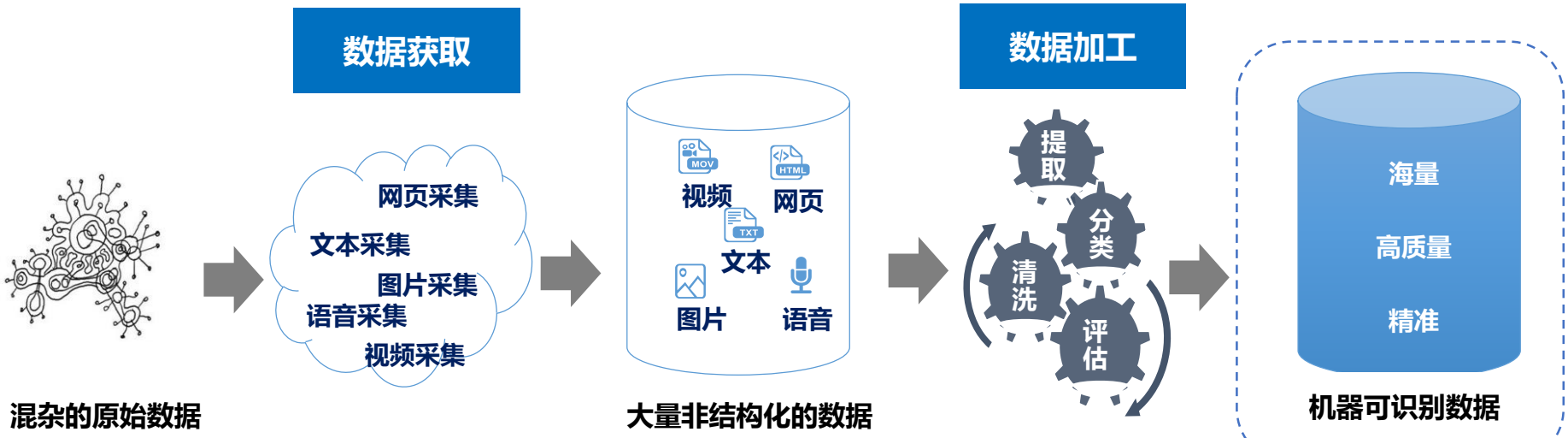
基础层

数据：身份信息、医疗、购物、交通出行等各行业、各场景的一手数据。

- 方言语音数据
- 场景语音数据
- 语音文本数据
- 百科、音乐、游戏、电影等不同领域的词汇、属性及关系数据
- 社交网络文本数据
- 新闻媒体舆情数据
- 社区、论坛知识数据
- 多语种文本数据
- 细粒度语义标注数据
- 人脸图像数据
- 字符图像数据
- 物体图像数据
- 物体属性数据
- 物体行为数据

数据处理的困境

如何获取和加工数据，人工智能基础数据的两大难题



如何采集：原属数据类型繁杂，没有统一的采集标准

谁来采集：线下数据数量多、类型广，需要外包人工采集，时间、经济成本大

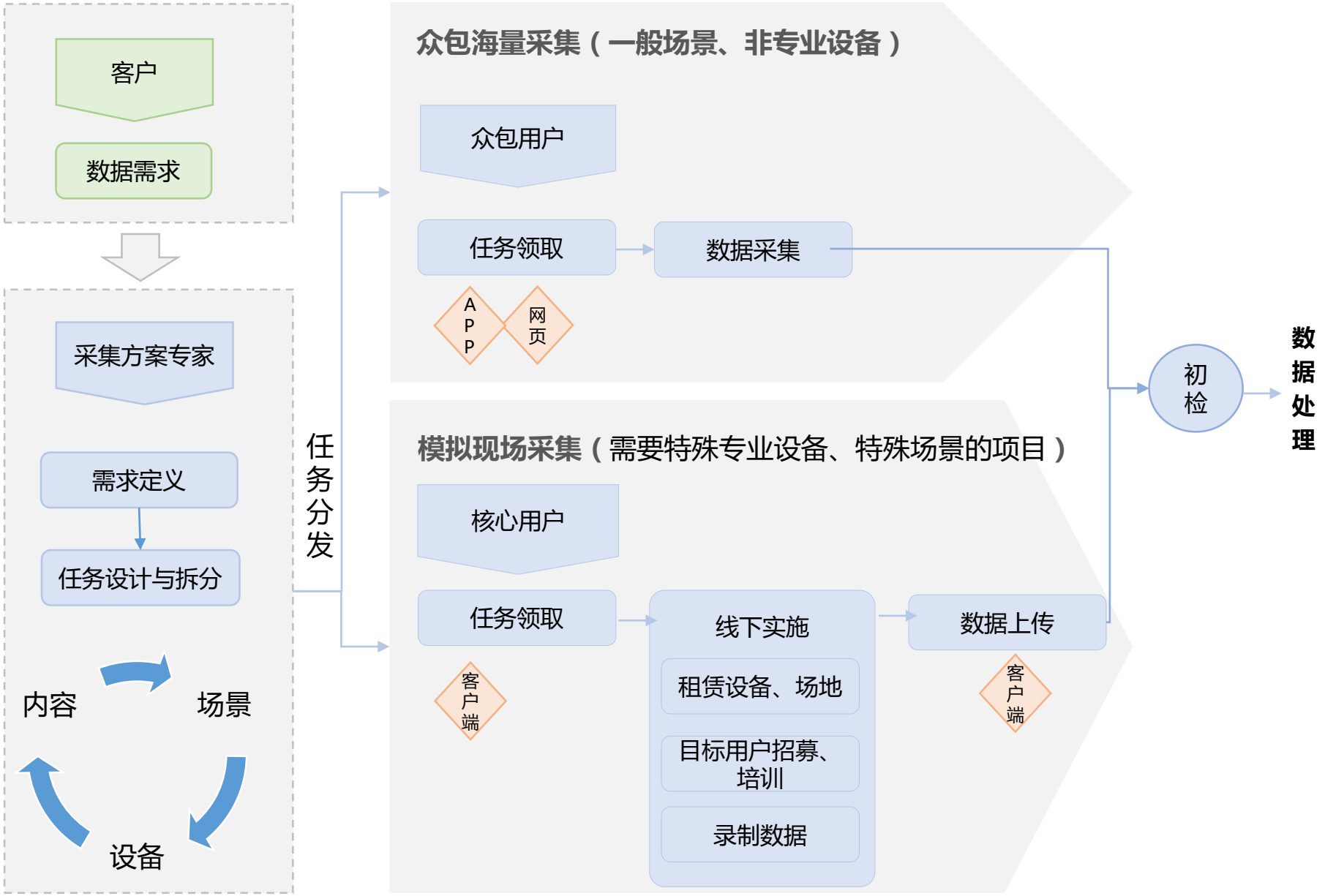
如何加工：行业缺乏统一标准，方法不一致，重复人力投入

谁来加工：机器难以完全胜任；人工处理花费大量人力、物力

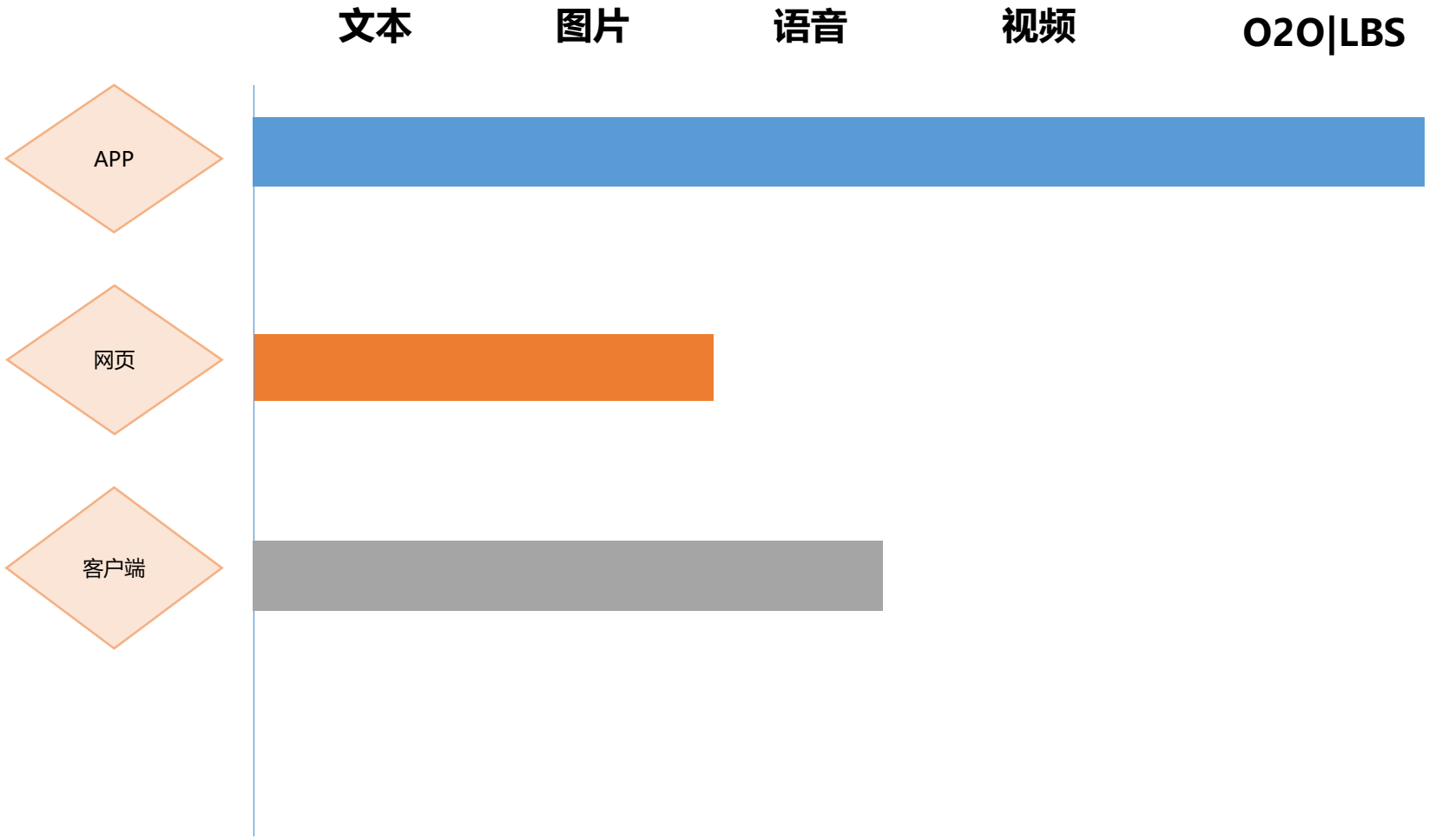
质量如何保障：人工抽检，覆盖面有限，准确率有瓶颈

如何获取数据？

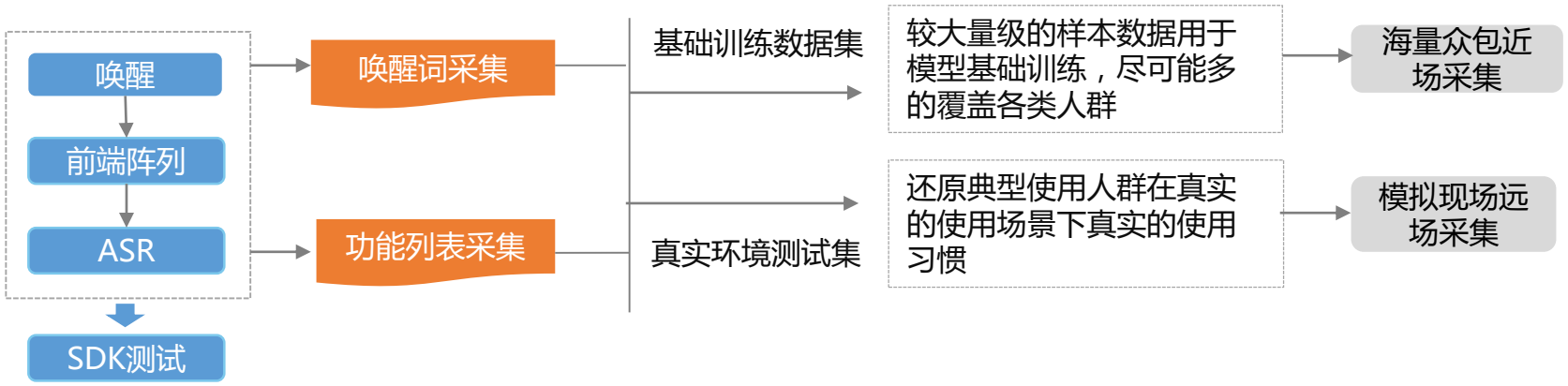
众包数据采集，为人工智能提供海量数据支持



采集场景工具化，全面覆盖各种数据类型



智能家居语料采集方案



近场数据：

众包用户 手机app自助采集。

采集能力：

累计完成超过5000小时，覆盖10w人的20种唤醒词数据。

近场数据：

核心用户辅助百度官方运营实施。

在真实使用场地，模拟真实使用情况，寻找设备真实受众人群的代表按照要求进行统一采集。

。

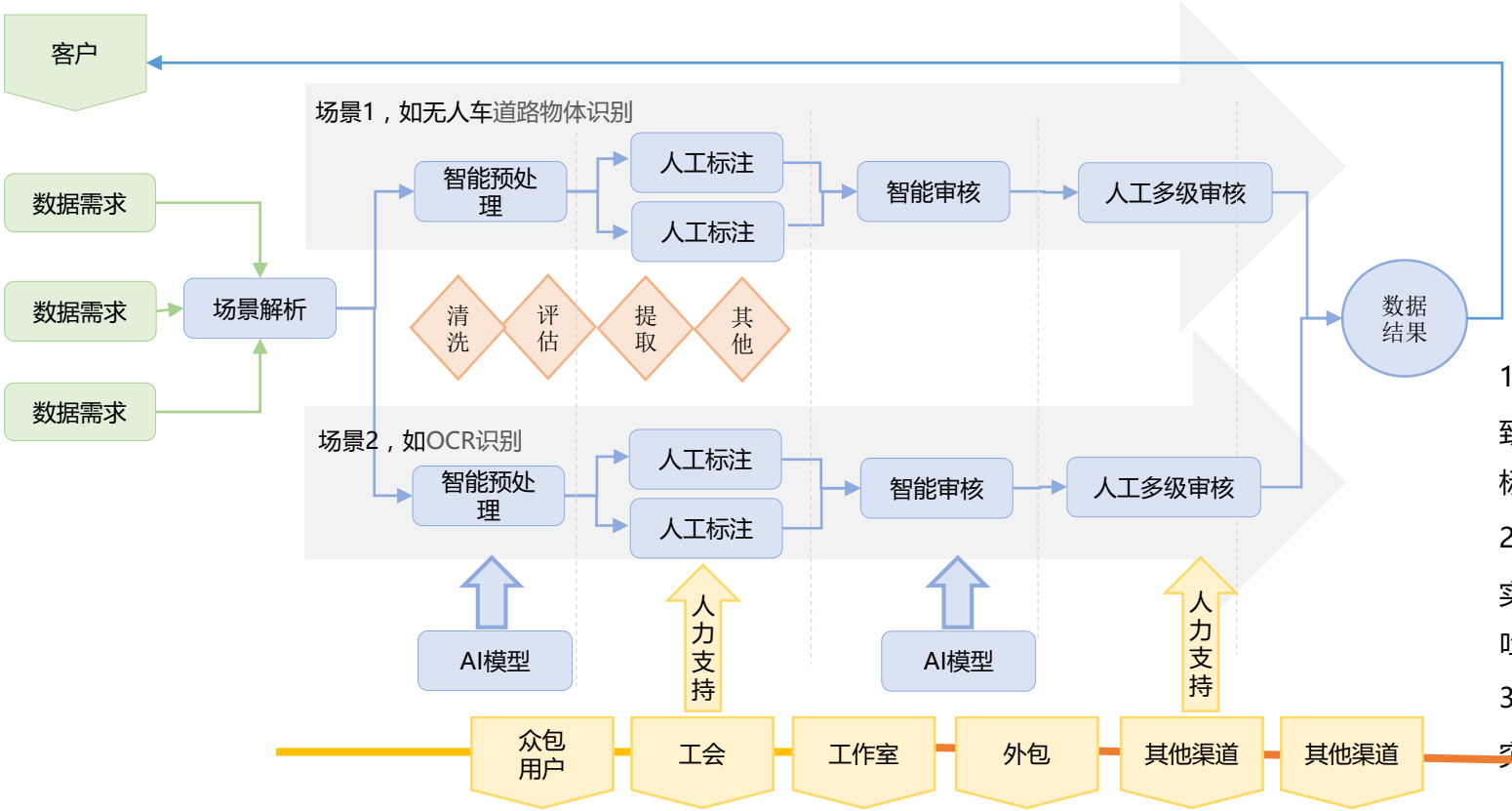
项目执行方案：

5种真实环境、3种真实距离

200人/天

如何加工数据？

打造链条化智能化数据深度加工厂，为人工智能发展保驾护航



1、沉淀处理方法论：

致力于数据处理规则的标准化与工具化

2、智能加众包标注：

实现日均百万的数据吞吐

3、多级质量管控：

突破准确率瓶颈

沉淀处理方法论，致力于数据处理规则的标准化与工具化

- 不完整数据
- 错误数据
- 冗余数据
- 数据标签化
- 垂类数据

- 相关性评估
- 时效性评估
- 竞品评估
- 互联网，社交网络舆情
- 电子商务评论

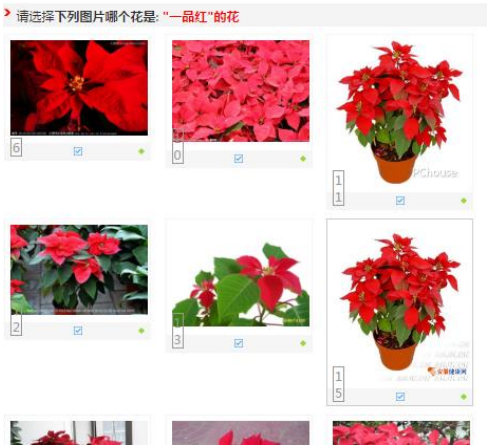
- 关键词提取
- 网页内容提取
- 图片内容提取（OCR识别，人脸识别，物体识别等）

- 地图信息制作
- 语音转写
- 其他数据标注

标注工具——通用图片检测

通用图片检测类型涵盖商品、动物、植物、菜品、服装搭配、黄反、暴恐、 建筑、素材等多种垂类。

1. 多图 vs. 单图；
2. 图+参考文字/参考图/搜索页面/参考链接/预识别结果/特定内部参考页面；
3. 多题 vs. 单题；
4. 题目类型：单选/多选/多级菜单选择/填写

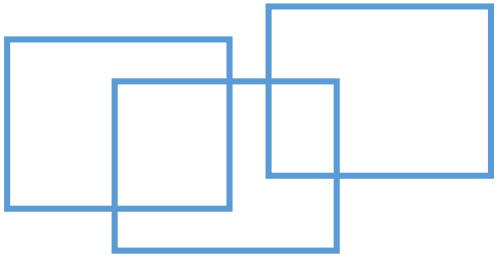


标注工具——目标框选类

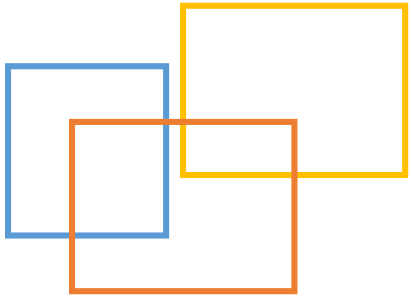
框选类能力涵盖：

普通矩形、分类矩形、普通多边形、分类多边形、区域填色、多级属性多边形、Parsing、点+线+区域复合检测

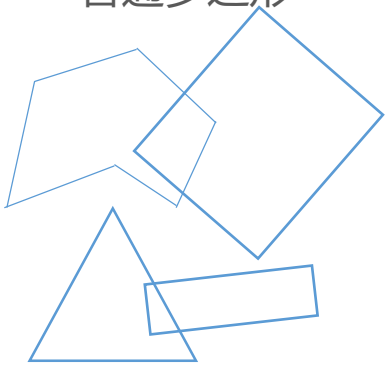
普通矩形框



分类矩形

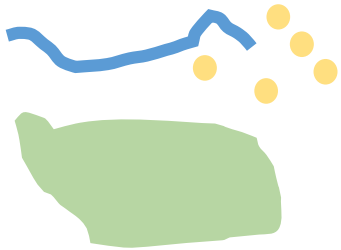


普通多边形

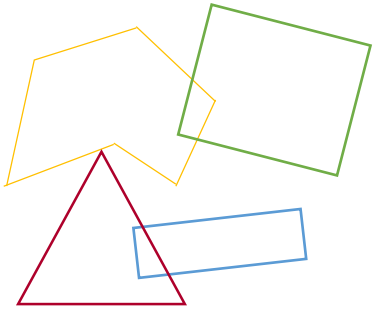


区域填色|多级属性多边形|Parsing

点+线+区域复合



分类多边形



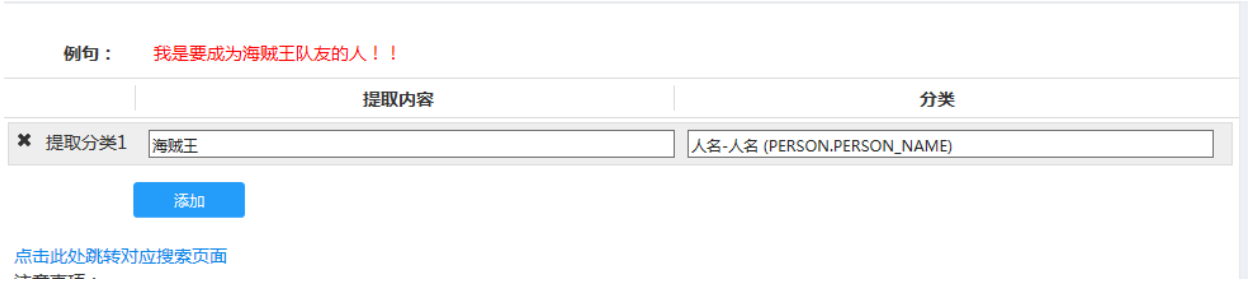
标注工具——内容评估

◆ 用户行为画像

对“兴趣偏好“属性进行策略优化，通过第三方人工标注，通过用户人工贡献评价，评估策略优化后的标签准确率

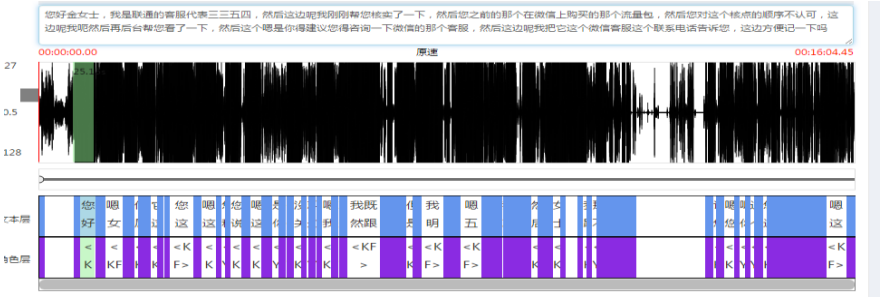
◆ 要素提取

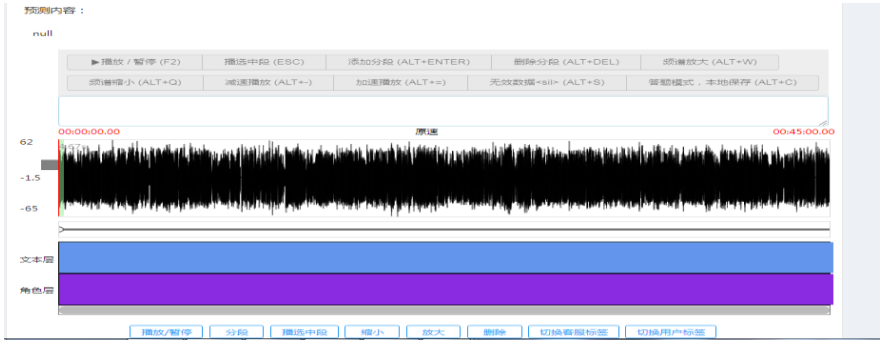
依据客户要求对文字内容或槽位进行提取并定位具体属性。



标注工具——图片&语音转写

1. 进行多种语言OCR文字转写
2. 进行多种口音的语音文字转写





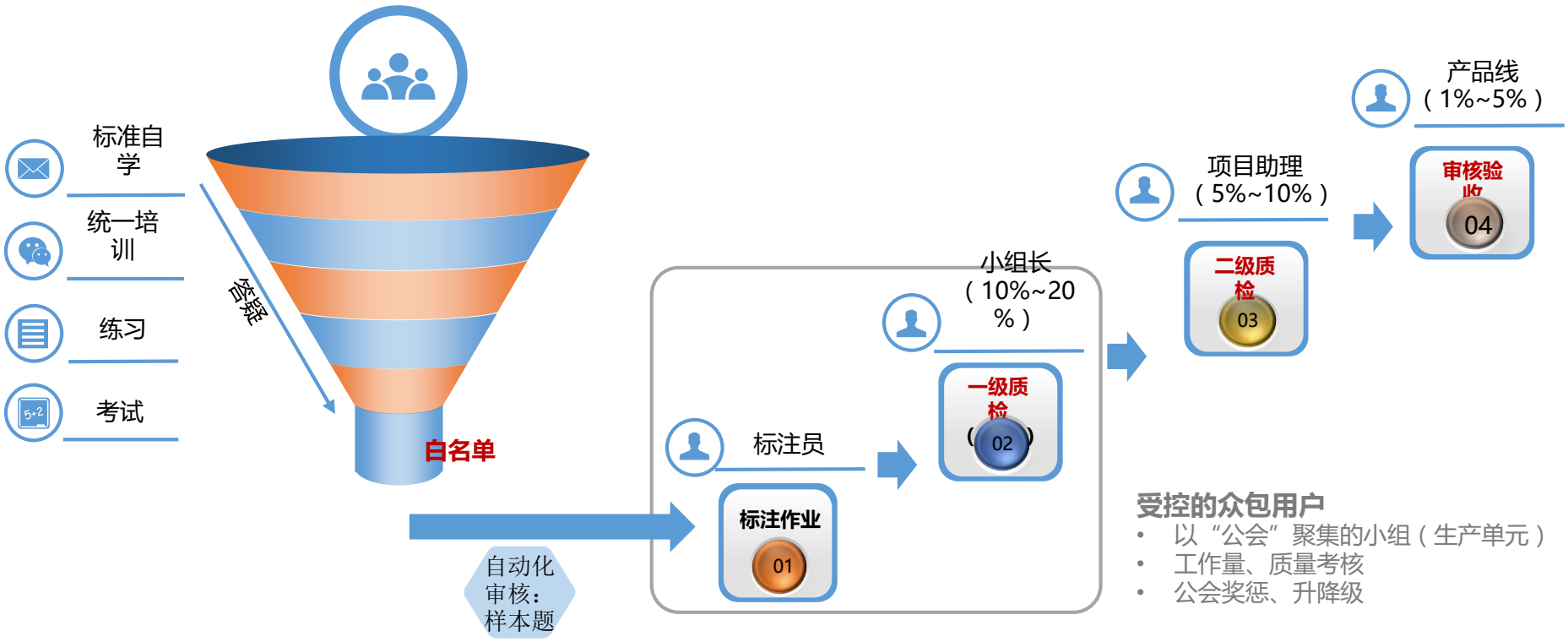
智能加众包标注，实现日均百万的数据吞吐

分类
删除

自动化生成预标注结果
辅助标注

众包用户
核心用户团队

多级质量管控，突破准确率瓶颈



经典案例：人脸识别基础数据服务

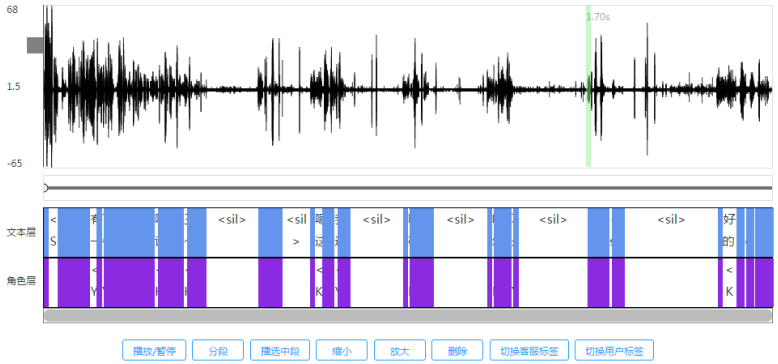
使用场景：
身份识别，摄像监视系统，支付系统，门禁系统

- 采
  - 一人多照人脸图片：多表情、多姿态
  - 跨年龄段人脸图片：70-00后全年龄段
  - 多光照条件、
  - 多遮挡条件
- 标
  - 人脸检测标注：人脸位置框选
  - 人脸关键点标注：人脸5点-72点标注
  - 验收通过率100%

经典案例：语音识别基础数据服务

使用场景：
语音识别，智能机器人

- 采
  - 采集指定地区的汉语方言数据
  - 通过手机自带麦克录制
  - 四川话 / 上海话 / 湖南话等8种方言
  - 安静 / 吵闹环境录制
- 标
  - 语音数据转写
    - 中文方言、普通话
    - 转写准确率98%，业内第一

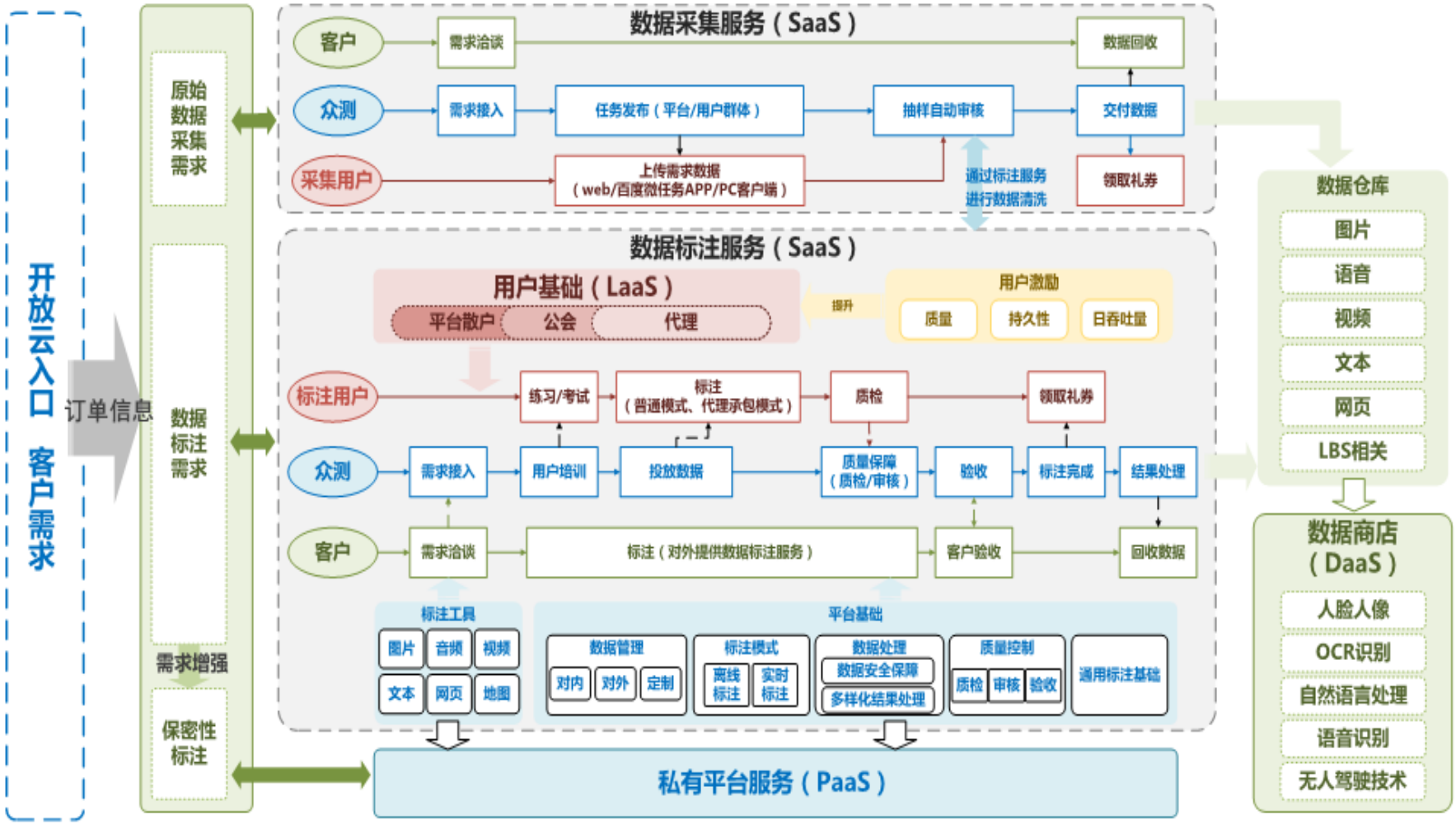


根据数据需求类型，覆盖更多实际应用场景

人像识别	语音识别	OCR识别	无人驾驶	模式识别
多角度自拍 跨年龄段 暗光人脸 亲子全家福 人脸打点	唤醒词语料 客服语音 普通话文本转录 中英文混读 方言 - 粤语 方言 - 四川话	驾驶证图片 名片图片 商标LOGO 彩票图片 医疗单据图片	红绿灯图片 道路障碍物 交通行驶区域 道路分界线 交通路面边界 泰国车牌	手部图片 时尚服装 推荐菜品图片 汽车外观图片 动物图像 花卉图像

百度数据产品

百度数据产品矩阵


开放云入口 客户需求
订单信息
原始数据采集需求
数据标注需求
需求增强
保密性标注
数据采集服务（SaaS）
客户
需求洽谈
数据回收
众测
需求接入
任务发布（平台/用户群体）
抽样自动审核
交付数据
采集用户
上传需求数据
（web/百度微任务APP/PC客户端）
通过标注服务进行数据清洗
领取礼券
数据标注服务（SaaS）
用户基础（LaaS）
平台散户
公会
代理
提升
用户激励
质量
持久性
日吞吐量
标注用户
练习/考试
标注
（普通模式、代理承包模式）
质检
领取礼券
众测
需求接入
用户培训
投放数据
质量保障
（质检/审核）
验收
标注完成
结果处理
客户
需求洽谈
标注（对外提供数据标注服务）
客户验收
回收数据
标注工具
图片
音频
视频
文本
网页
地图
平台基础
数据管理
对内
对外
定制
标注模式
离线标注
实时标注
数据处理
数据安全保障
多样化结果处理
质量控制
质检
审核
验收
通用标注基础
私有平台服务（PaaS）
数据仓库
图片
语音
视频
文本
网页
LBS相关
数据商店
（DaaS）
人脸人像
OCR识别
自然语言处理
语音识别
无人驾驶技术

Thanks

合作联系

zhongbao.baidu.com