AiCon
全球人工智能与机器学习技术大会

AI时代的数据解决方案

李鑫

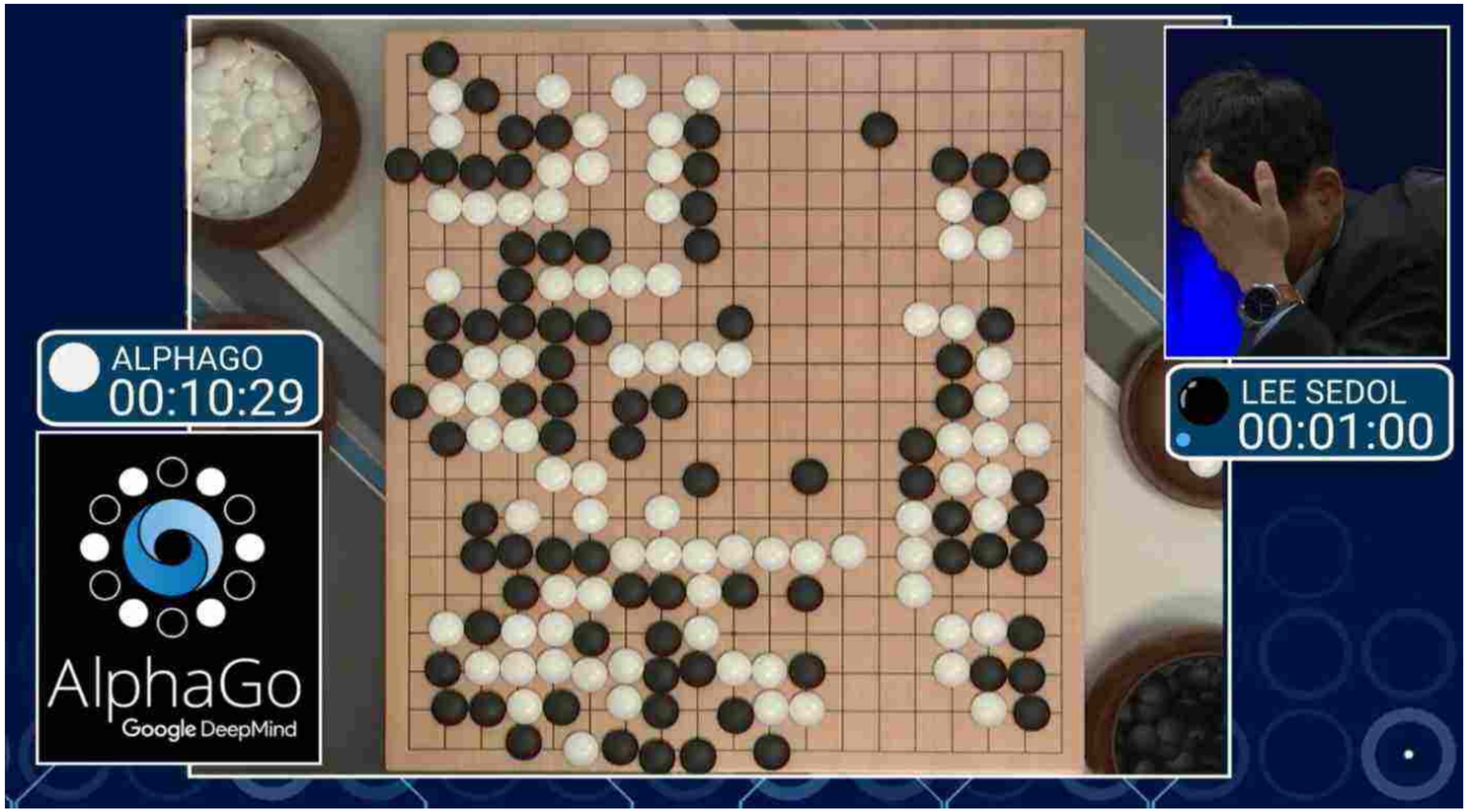
百度资深研发工程师

主办方 Geekbang. 极客邦科技 InfoQ

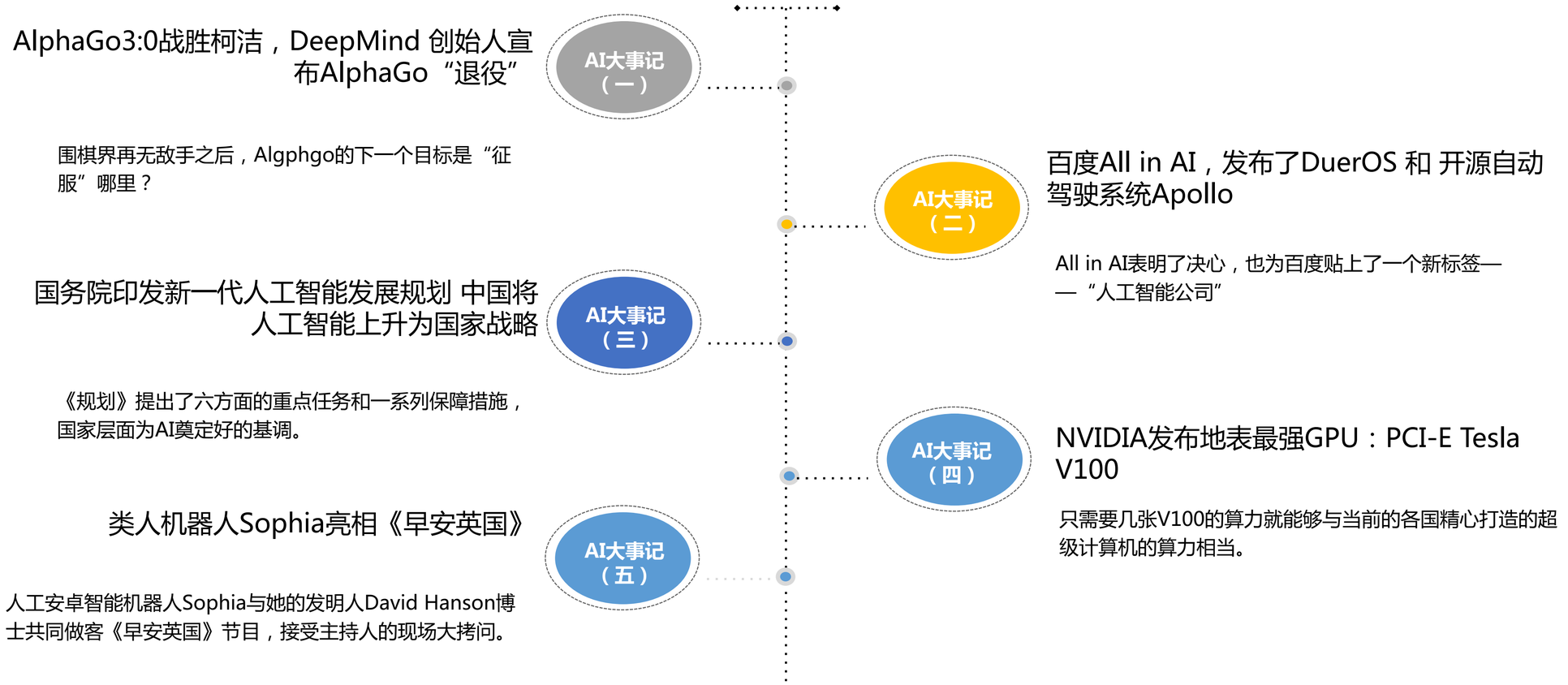
TABLE OF CONTENTES

人工智能进入公众视野

人工智能的强大能力已被证明

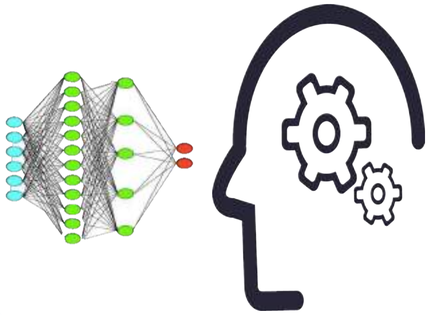


人工智能2017大事记

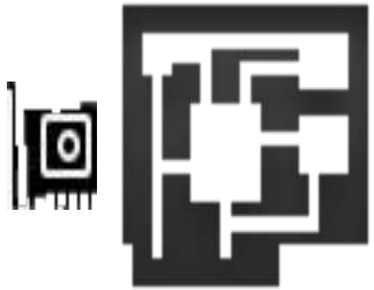


人工智能爆发的三大因素

深度学习



高性能运算



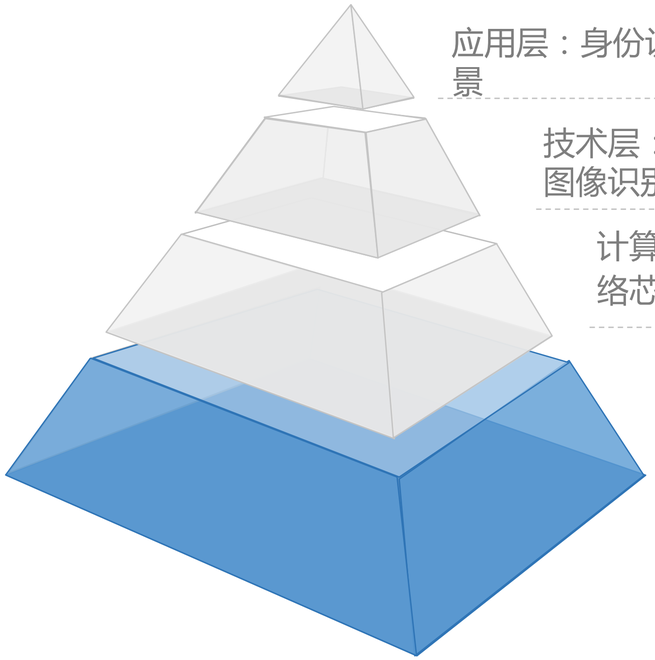
大数据

算法是核心，**计算**、**数据**是基础

TABLE OF CONTENTES

海量、精准、高质量的数据是人工智能的根本

数据是一切人工智能技术和应用实现的基础保障和前提！



应用层：身份识别、语音助手、机器人、智能安防等场景

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

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**

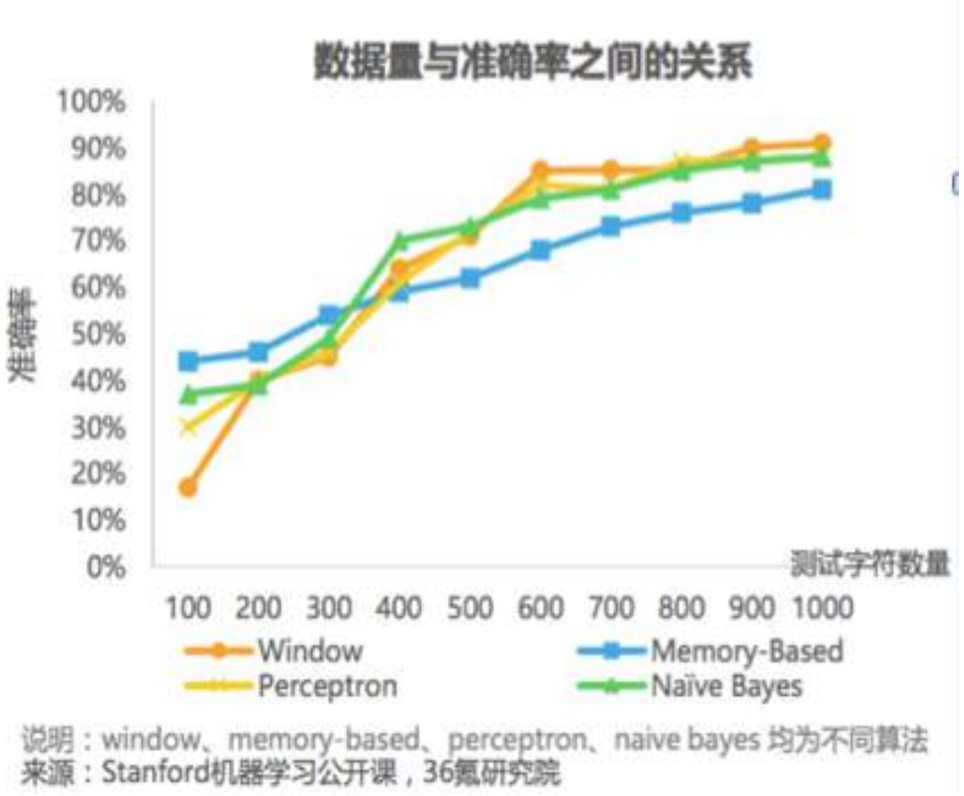
- **社交网络文本数据**
- **新闻媒体舆情数据**
- **社区、论坛数据**
- **多语种文本数据**
- **语义标注数据**

- **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**

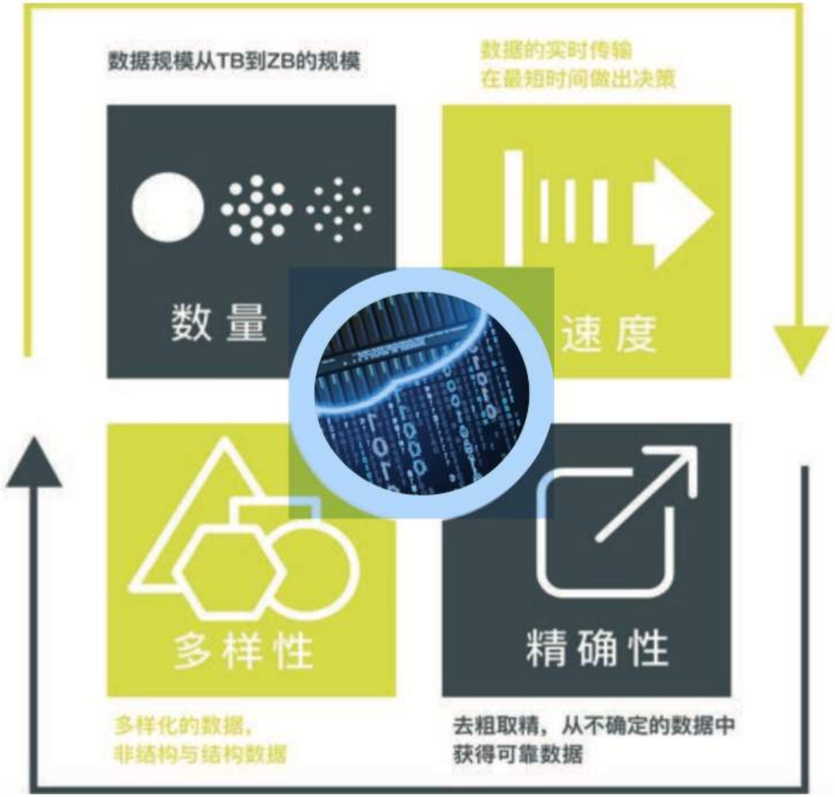
数据样本与算法模型

人工智能需要通过**大量的数据样本**来“训练”自己，才能不断提升输出结果的质量。

有时候，数据真的可以秒杀算法



说明：window、memory-based、perceptron、naive bayes 均为不同算法
来源：Stanford机器学习公开课，36氪研究院



快人一步抢占先机，数据竞赛“质&量”取胜

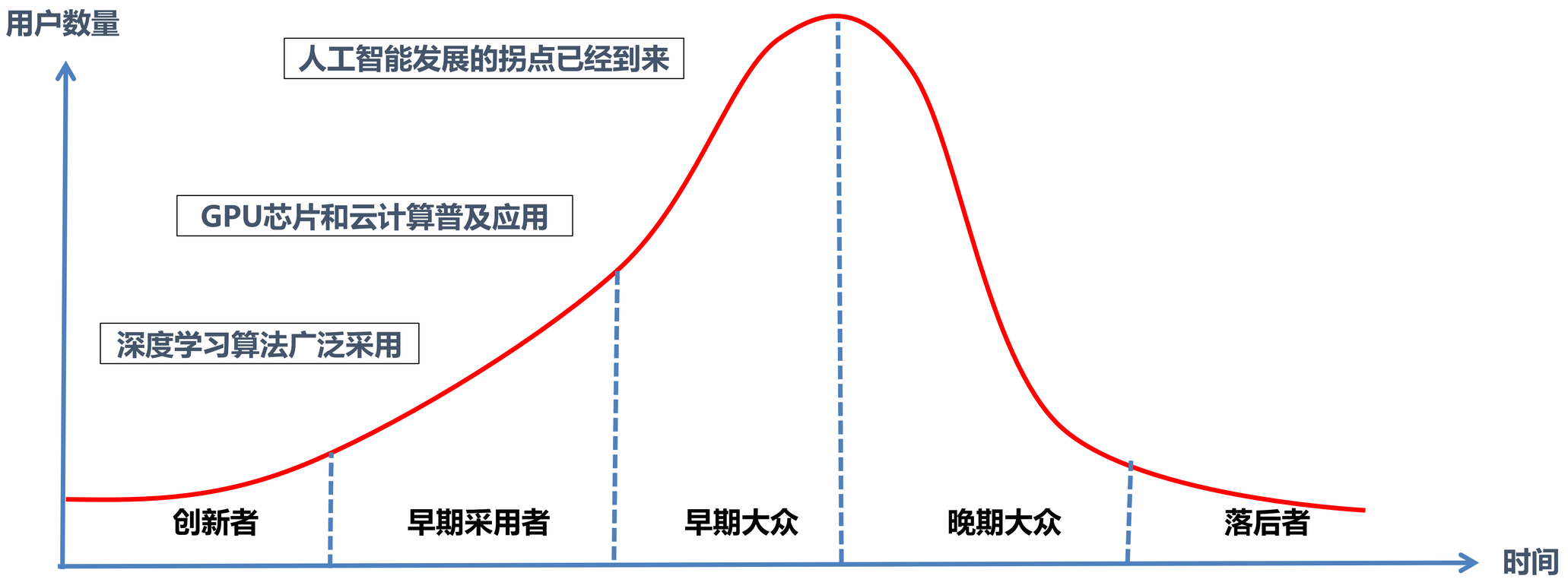
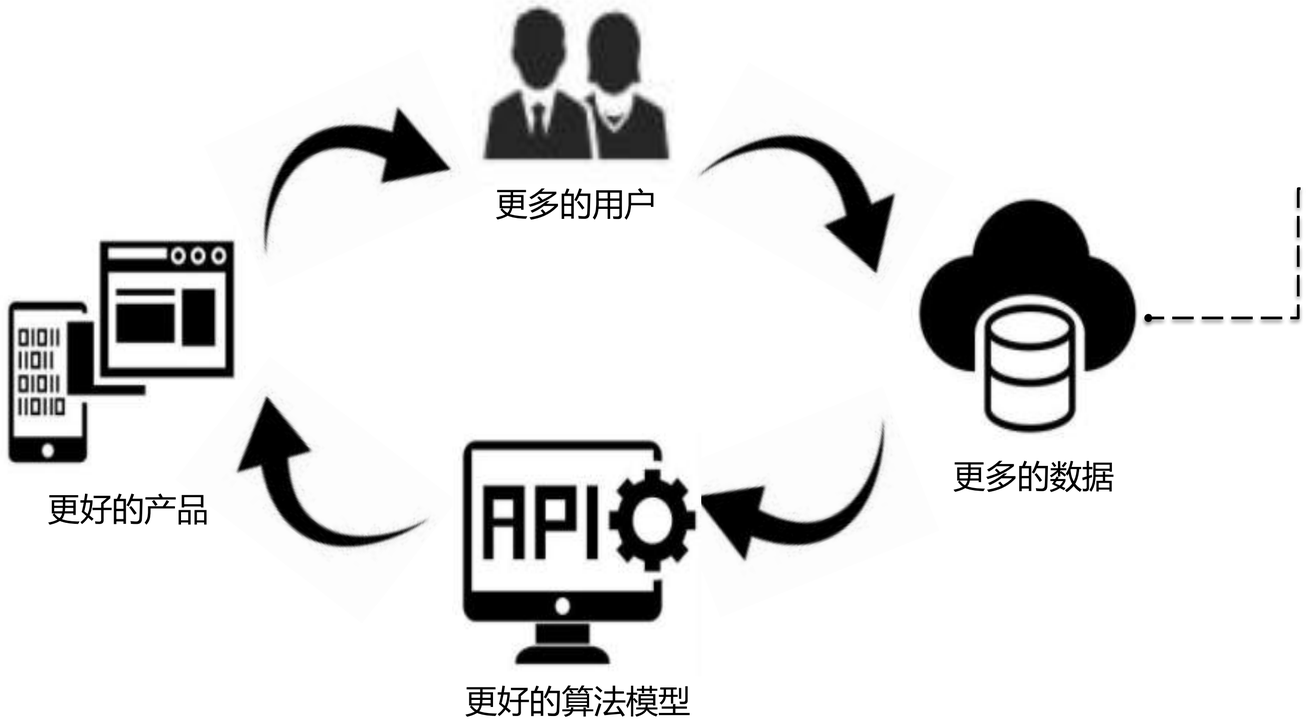


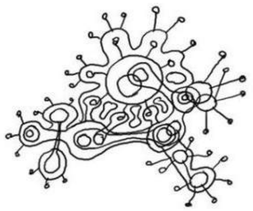
TABLE OF CONTENTES

项目“冷”启动的数据困扰



通常项目**启动门槛**会出现在这里。

获取和加工数据，AI基础数据的两大难题



混杂的原始数据

数据获取

网页采集
文本采集
图片采集
语音采集
视频采集

大量非结构化的数据

数据加工

机器可识别数据

如何采集：原属数据类型繁杂，没有统一的采集标准，同一批数据会出现多轮采集

谁来采集：线上数据可借助机器采集，线下数据需要纯人工采集，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

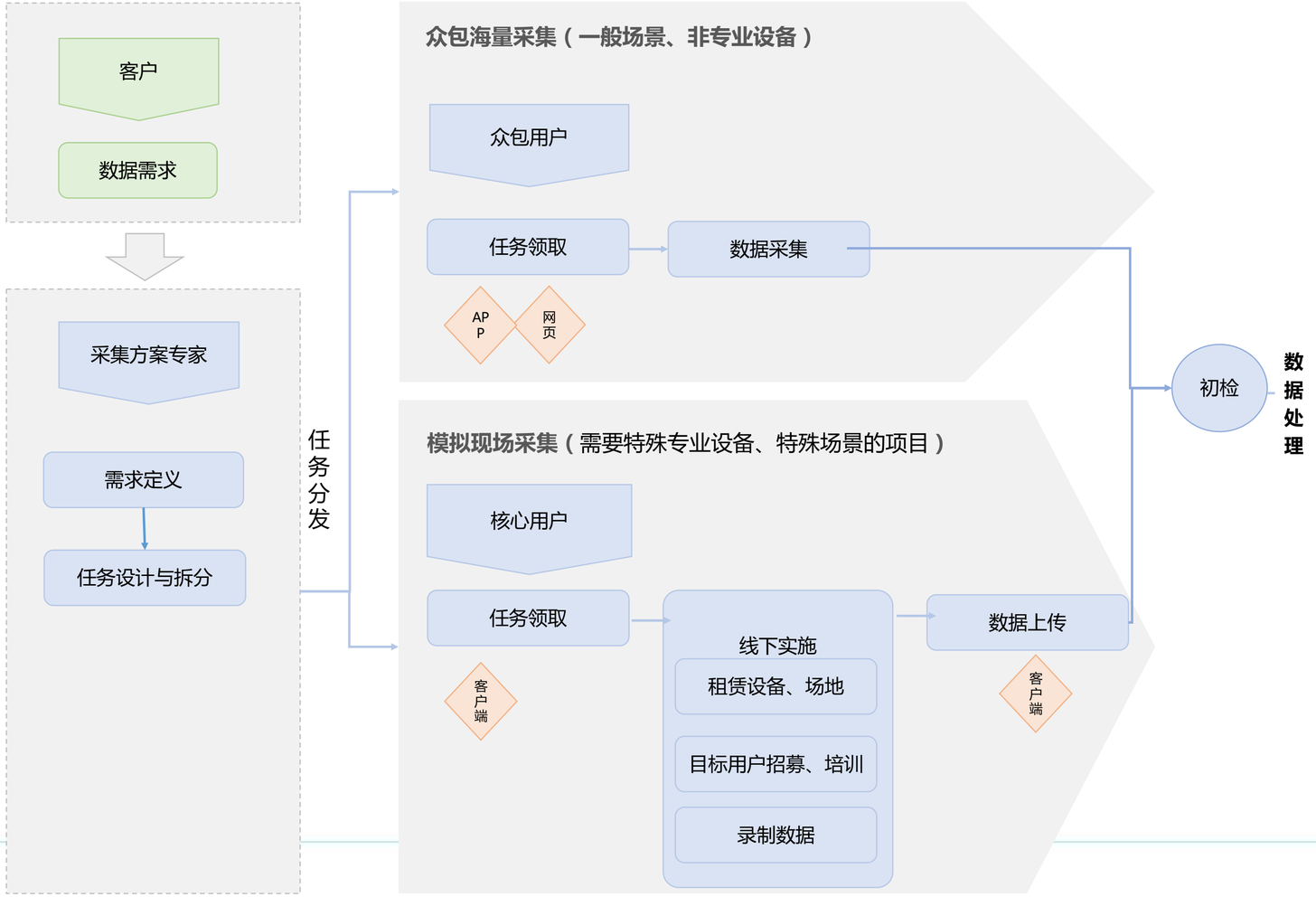
谁来加工：机器难以完全胜任；

人工处理花费大量人力、物力

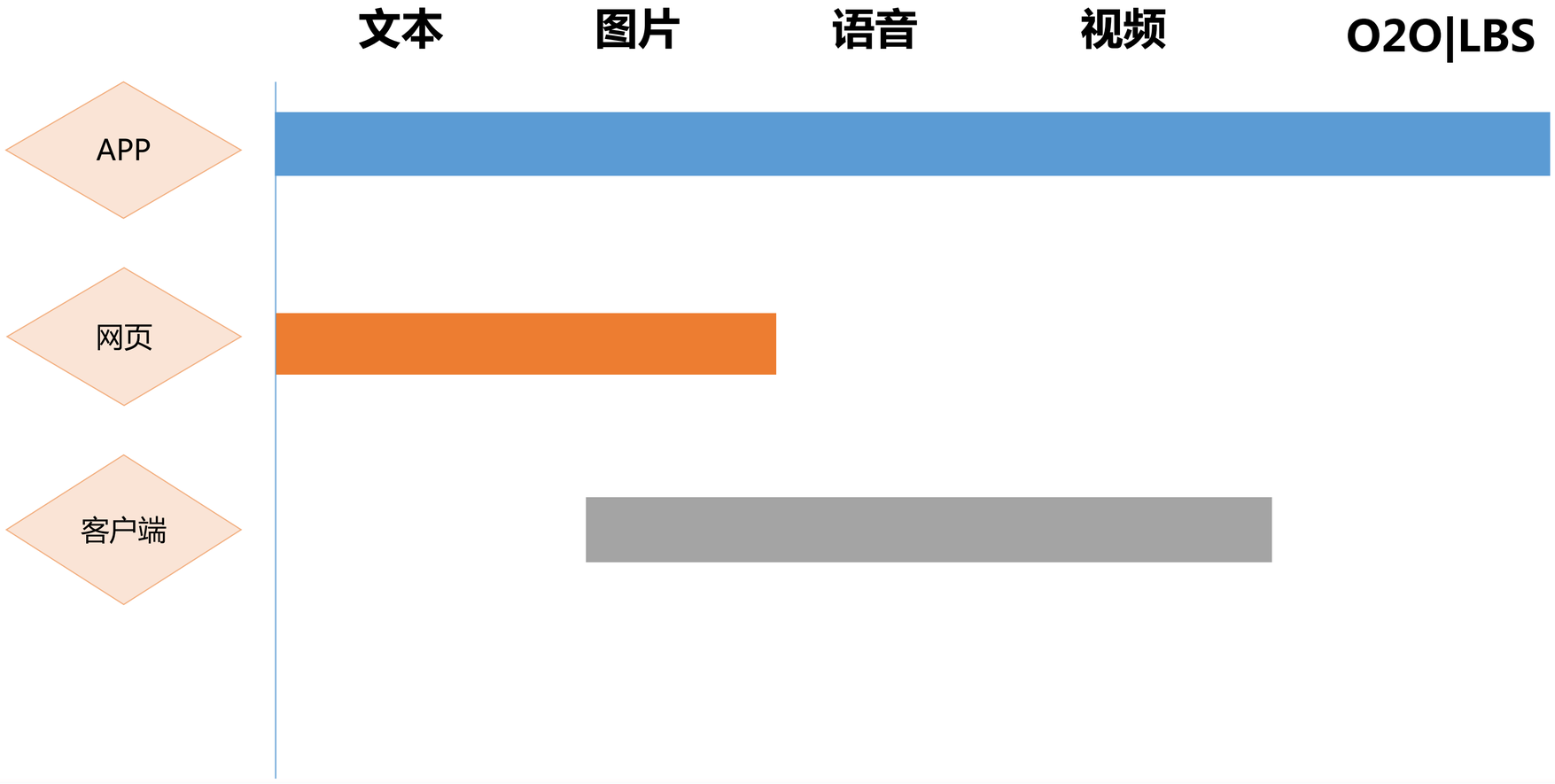
质量如何保障：人工抽检，覆盖面有限，准确率有瓶颈

TABLE OF CONTENTES

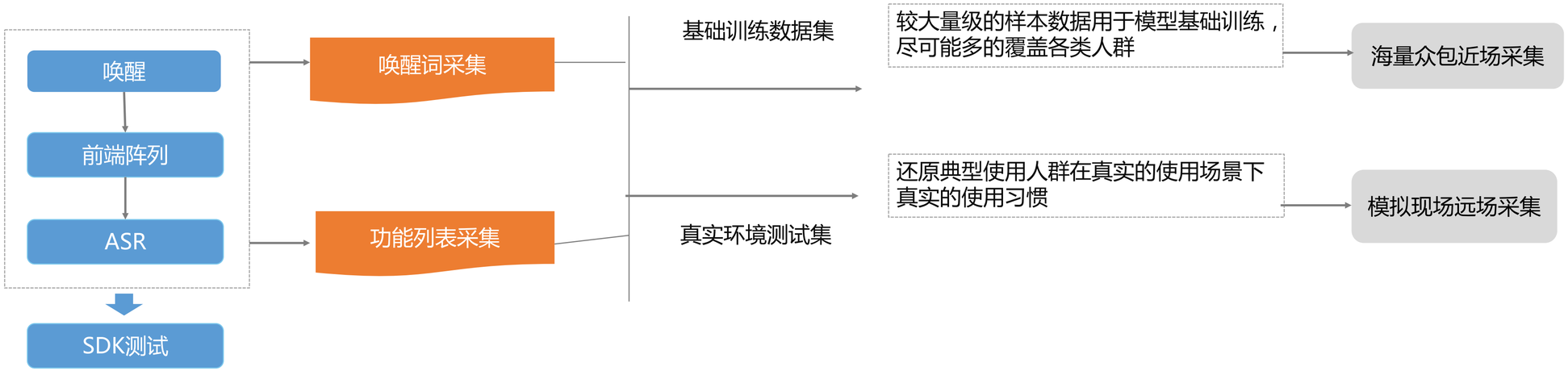
众包数据采集，为人工智能提供海量数据支持



采集场景工具化，全面覆盖各种数据类型



采集场景工具化，全面覆盖各种数据类型



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

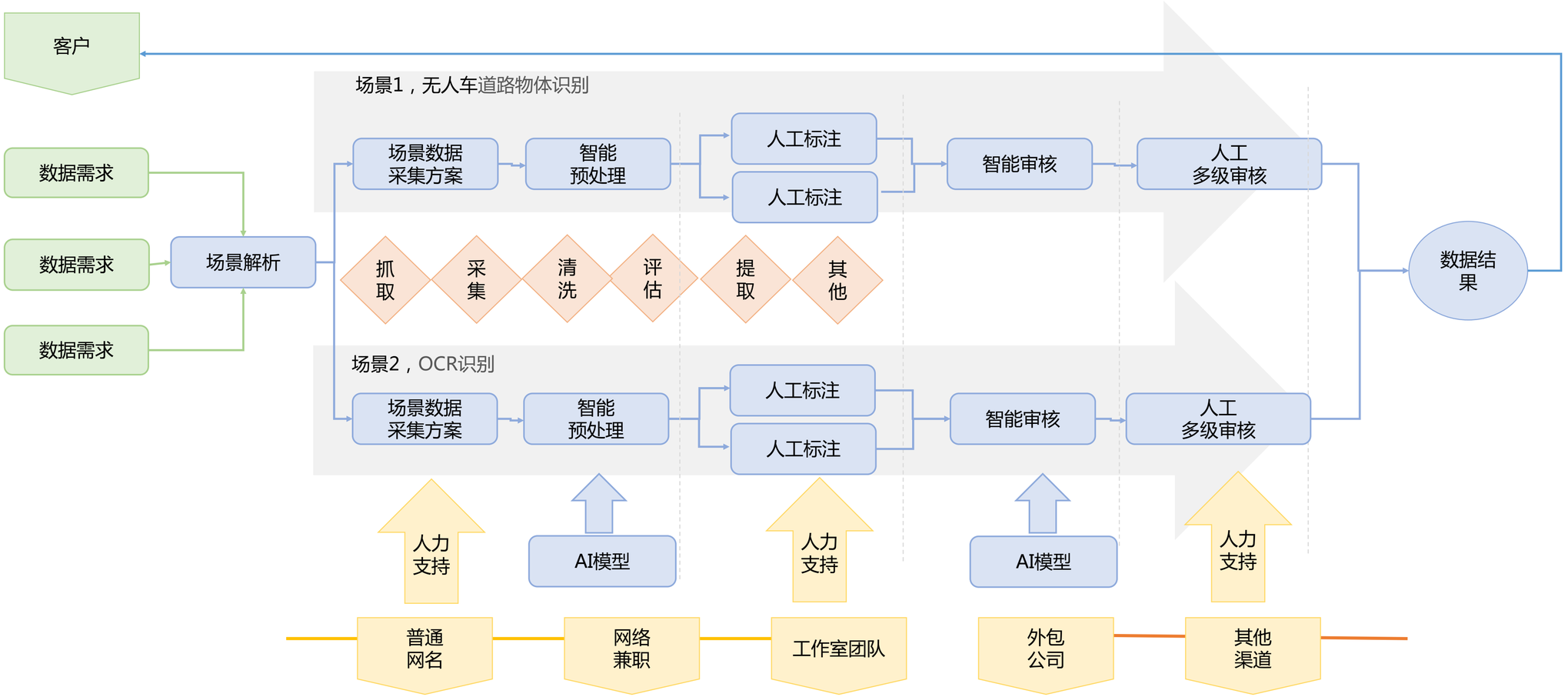
按照要求进行统一采集。

项目执行方案：

5种真实环境、3种真实距离

200人/天

链条化AI数据加工厂，为AI发展保驾护航



沉淀数据处理方法，建立数据处理规则

- 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

固化数据处理工具：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。

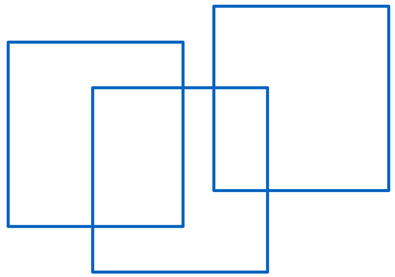
1. 多图 vs. 单图；
2. 图+参考文字/参考图/搜索页面/参考链接/预识别结果/特定内部参考页面；
3. 多题 vs. 单题；
4. 题目类型：单选/多选/多级菜单选择/填写

标注工具：目标框选类

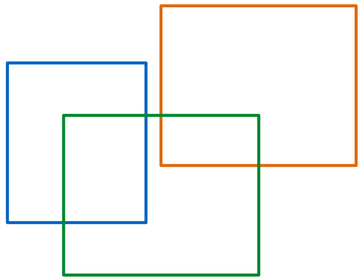
框选类能力涵盖：

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

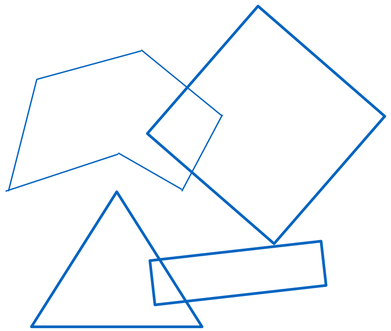
普通矩形
框



分类矩形

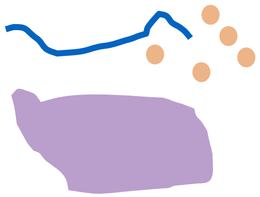


普通多边形



区域填色|多级属性多边形|Parsing

点+线+区域复合



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

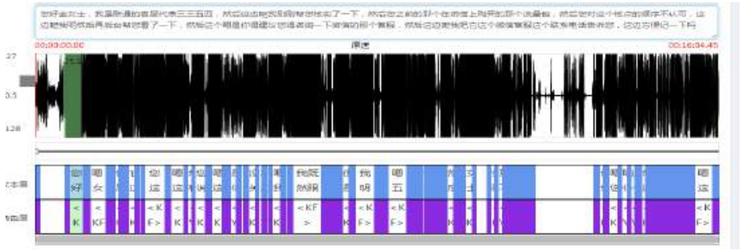
要素提取

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。

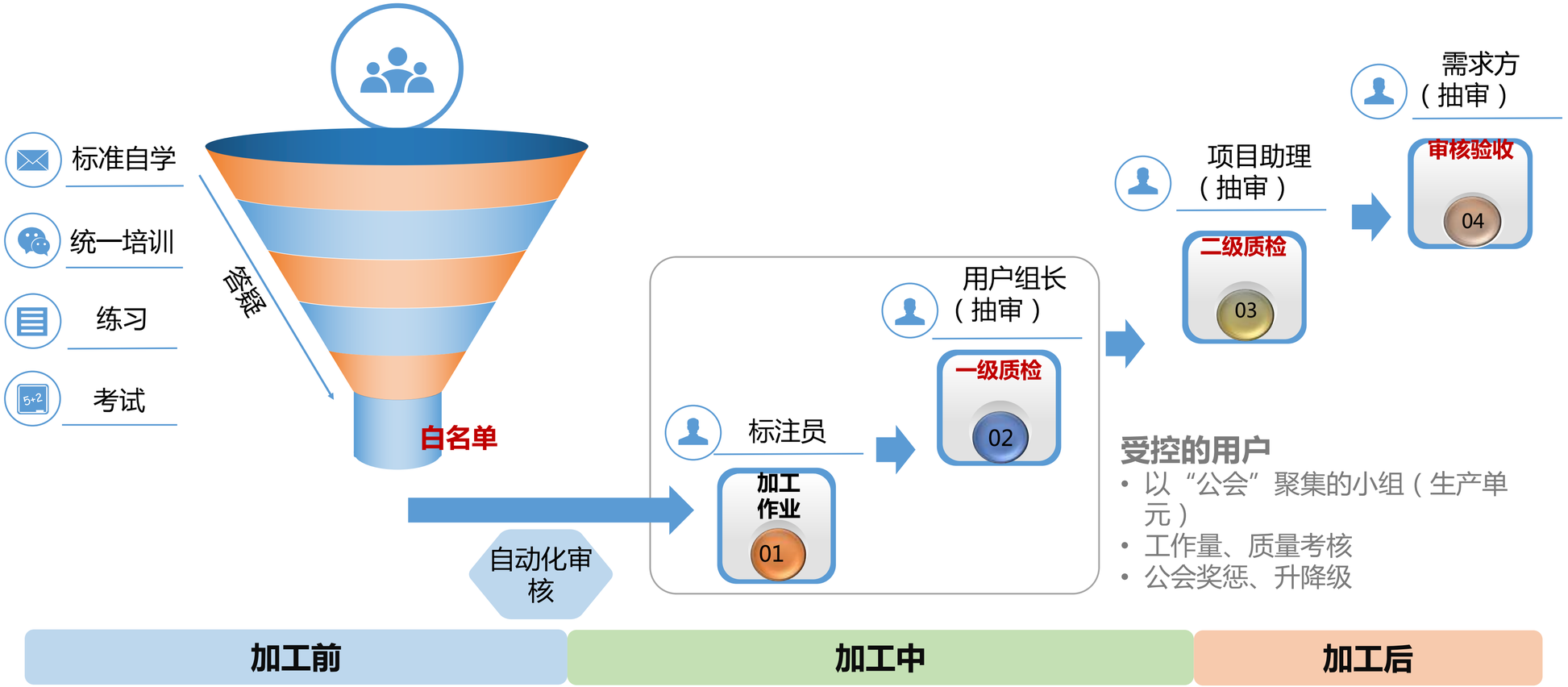
标注工具：图片&语音转写

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



多级质量管控，突破准确率瓶颈



智能标注：提升标注效率&降低标注成本

成本 = 题目数× 拟合人数 × 标注单价 + 审核成本

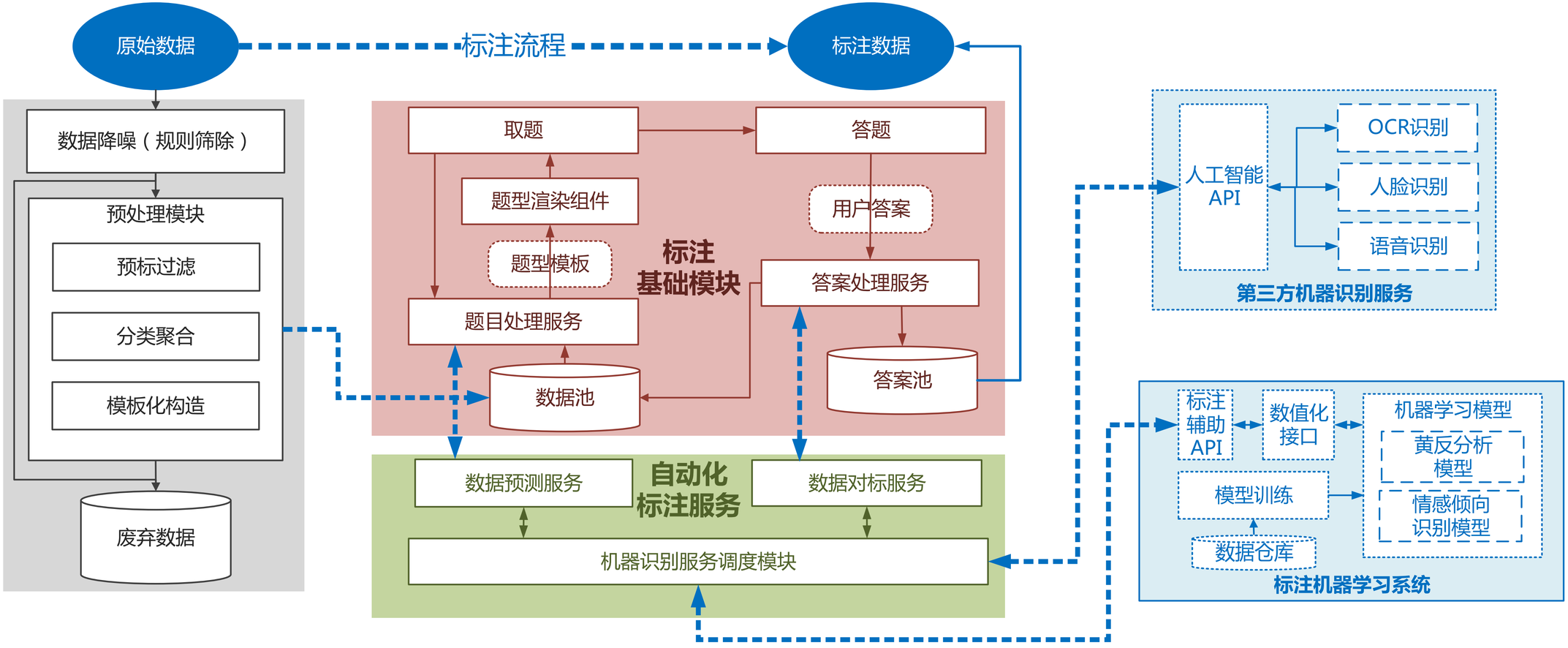
标注前：数据预处理
- 按标注难度区分任务
- 分类原始数据
- 去除无效的空白数据

- 机器实时对标人工标注结果，减少需要拟合的人工标注数量
- 机器标注结果与普通用户标注结果一起参与审核和验收

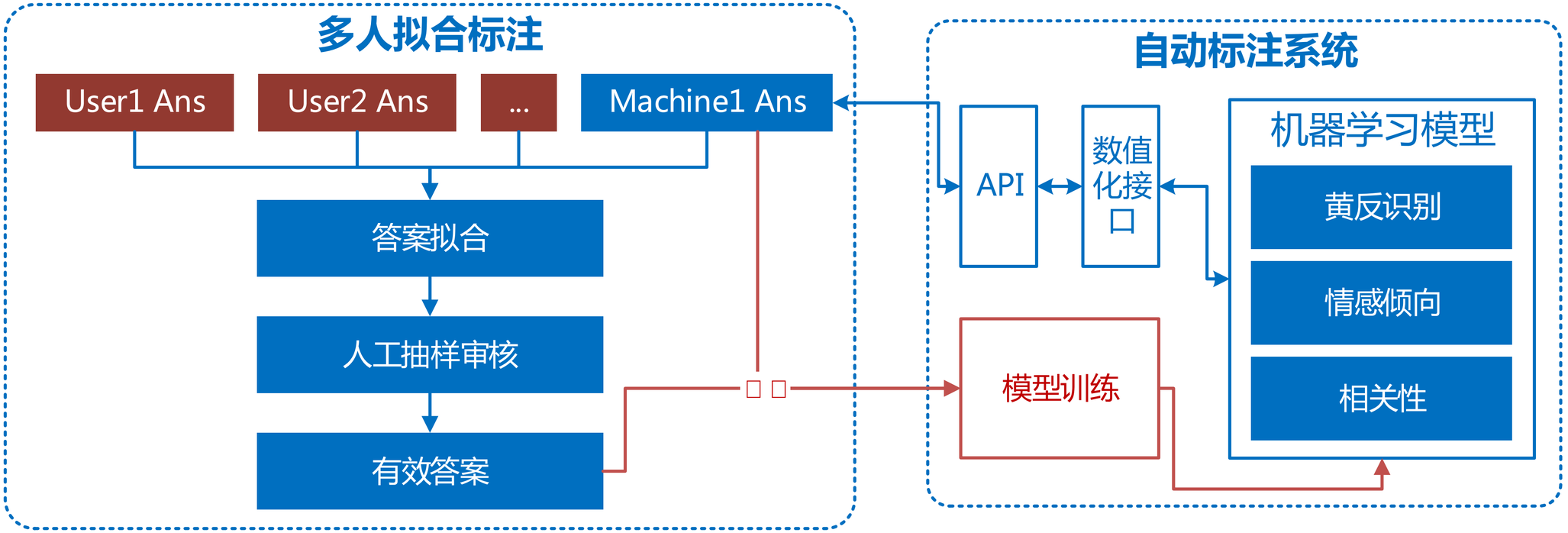
标注中：辅助标注
- 自动化生成预标注结果
- 辅助用户标注，简化操作步骤
- 尽量进行内容修改，降低标注难度

- 通过自动化审核规则&算法，抽样概率较大的badcase，降低审核成本

智能标注：架构总体设计


原始数据
标注流程
标注数据
数据降噪（规则筛除）
预处理模块
预标过滤
分类聚合
模板化构造
废弃数据
取题
答题
题型渲染组件
用户答案
题型模板
标注
基础模块
答案处理服务
题目处理服务
答案池
数据池
数据预测服务
自动化
标注服务
数据对标服务
机器识别服务调度模块
人工智能
API
OCR识别
人脸识别
语音识别
第三方机器识别服务
标注
辅助
API
数值化
接口
机器学习模型
黄反分析
模型
模型训练
情感倾向
识别模型
数据仓库
标注机器学习系统

智能标注：标注中（人工+机器对标）



智能标注：标注中（辅助标注）

标注特点

- OCR模式识别技术已经比较成熟，进一步优化模型需要更多的bad case，因此需要更好地获取算法与人工标注存在较大diff的数据反馈到模型训练中
- OCR图片识别产品化，可直接调用接口服务对图片进行预识别

实时预测校正标注

- OCR识别接口的预标注准确率较高，用户只需要对少量bad case进行调整
- 单价构成：P_m*N_m
- 记录预测数据修改，大量修改的数据可作为训练数据

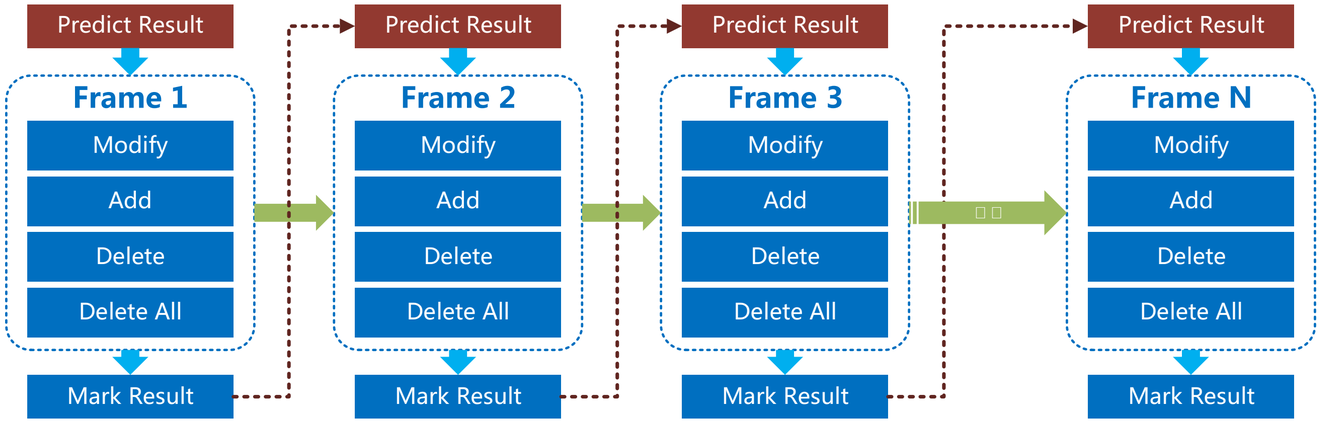
智能标注：标注中（辅助标注）

标注特点

- 投放的标注图片一般为连续帧，前后相邻帧中需要标注的语义类型、位置、大小等属性基本一致
- 通过算法进行预测数据生成准确性低，且无法达到实时获取

串行连续预测标注

- 连续帧图片相似度达到80%以上



智能标注：标注后（自动校验）

图片连续帧自动化审核

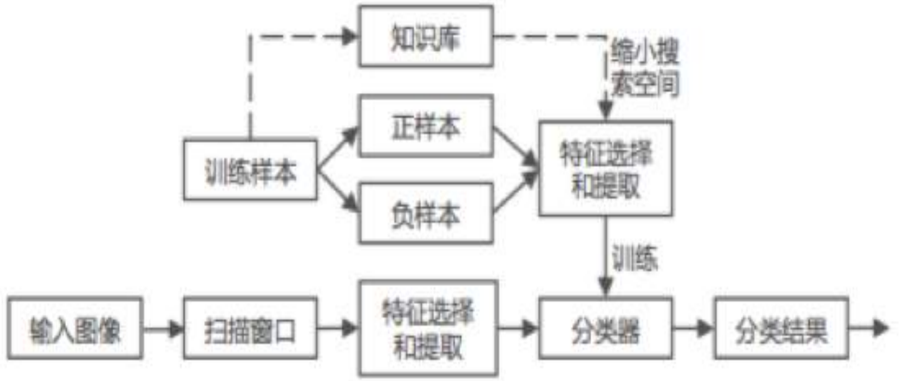
- 结果跳变检测：图片连续帧来说，每一帧的数据结果之间应该不存在跳变，如在第三帧的结果中突然少了很多第二帧的车辆，这种应该就需要被标记为异常结果，进而需要着重来进行查看

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

TABLE OF CONTENTES

计算机视觉数据解决方案



数据采集

根据实际计算机识别模型的要求，采集相应的图片、视频内容。

数据加工

将采集内容加工处理：标注关键点定位、提取特征信息打标签。

模型训练

将原始数据和特征标签数据提交到学习平台进行训练，提高识别精度

识别反馈

进行多次的迭代训练，最终计算机给予相应的识别反馈信息。

计算机视觉应用下的数据方案

泰国车辆号牌数据

泰国车牌图像采集

- 采集泰国车辆正面号牌图像数据
- 通过专业单反相机拍摄
- 真实泰国车牌超过2万张

车辆号牌框选标注

- 十字坐标对车牌进行定位框选
- 十字坐标对车牌字符进行框选
- 对车牌字符按照泰文进行标注

计算机视觉应用下的数据方案

特殊场景人脸图像数据

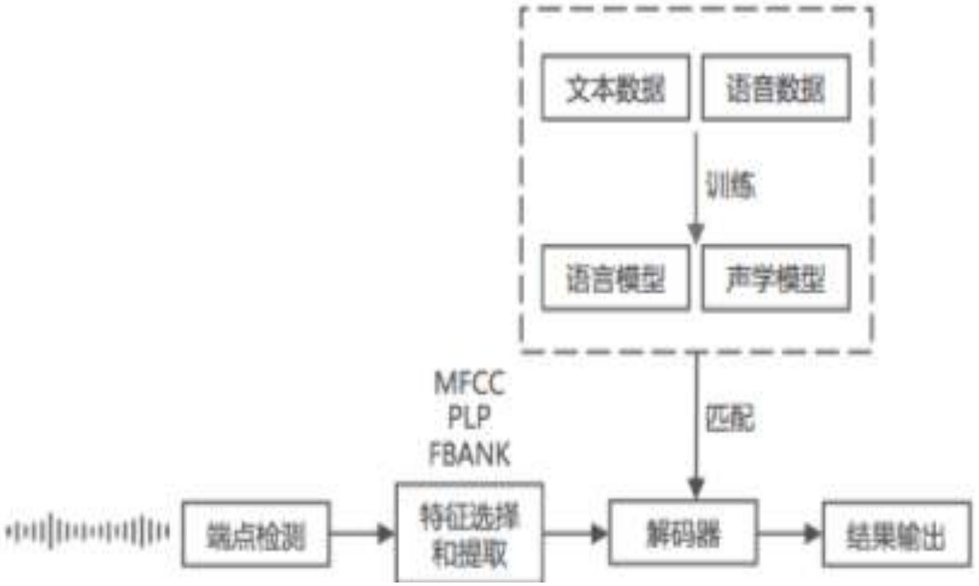
特殊要求人脸图像采集

- 采集指定条件下的人脸图像照片
- 通过手机自带相机拍摄
- 正常、暗光、微光多条件拍摄
- 口罩、墨镜、帽子多遮挡条件拍摄

人脸图像标注

- 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

语音识别数据解决方案



唤醒词、中英文语料、
方言语音识别。

语音识别

语义理解

多轮对话：上下文可随时打断,加入语境分析功能

机器翻译、实时同声传译

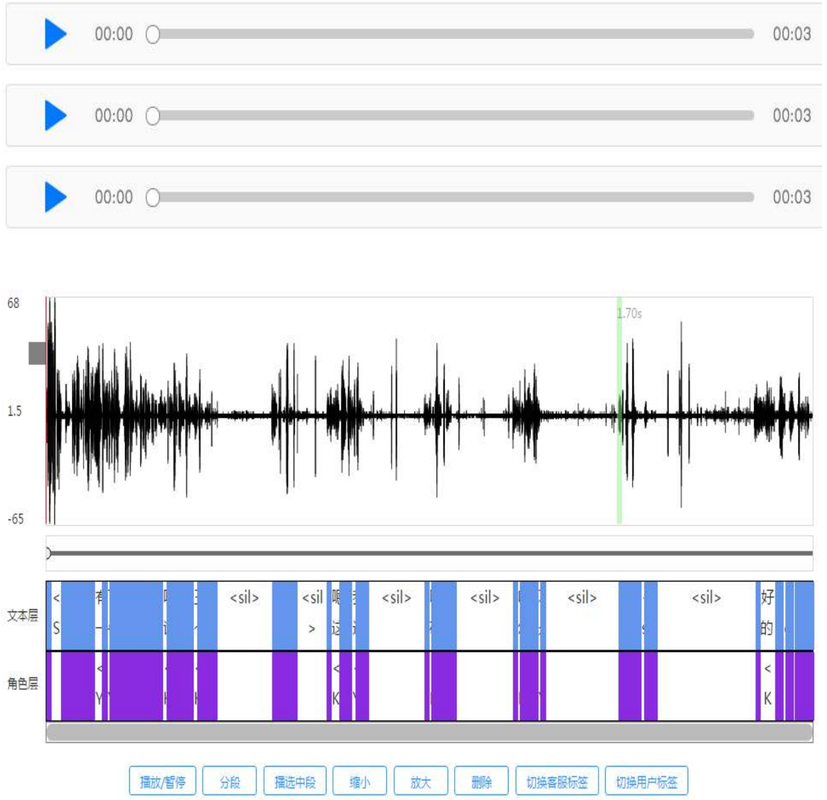
自然语言生成

语音合成

中文语音合成、中英文混合语音合成

语音识别数据解决方案

汉语方言语音数据



汉语方言语音数据采集

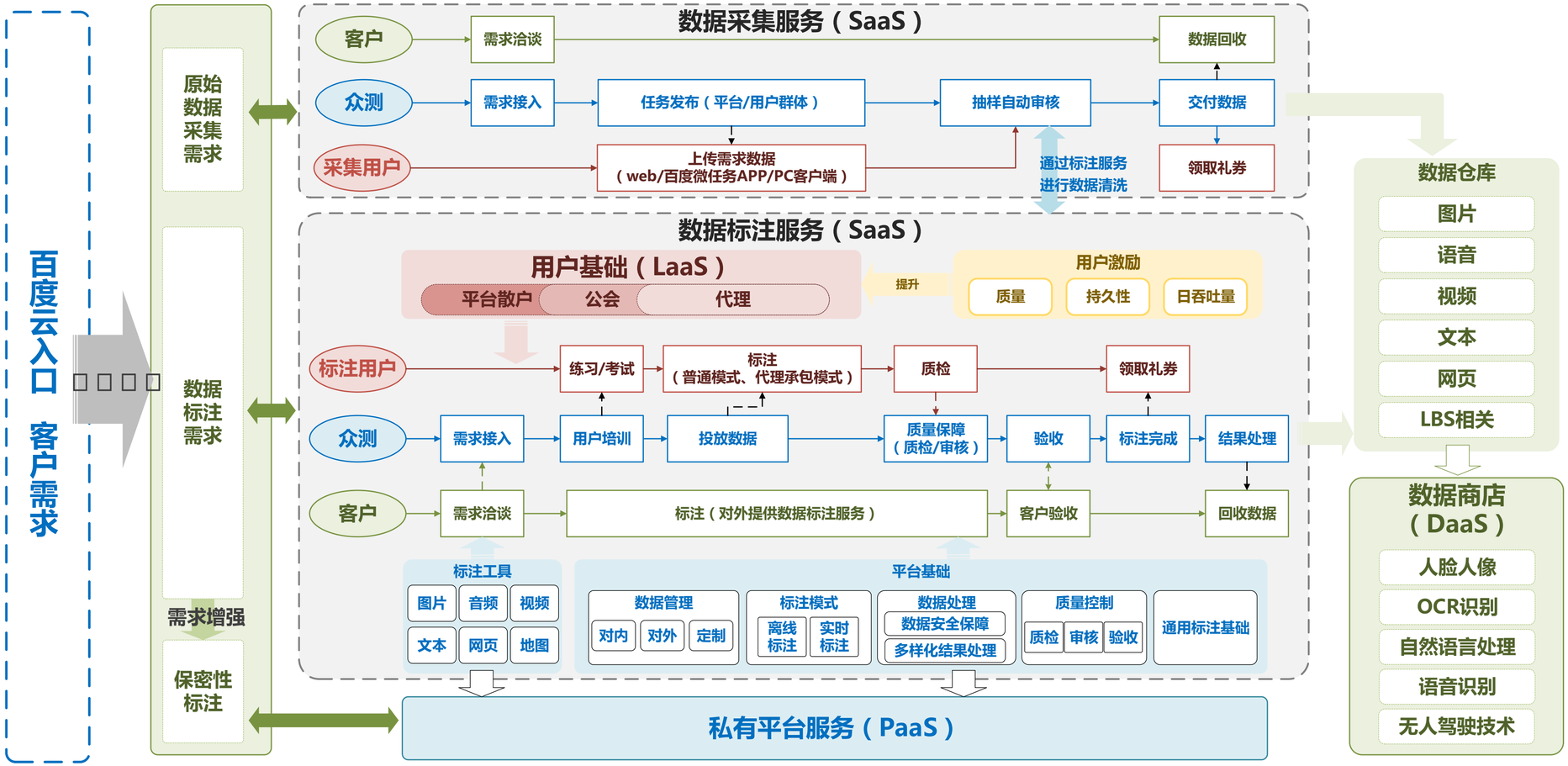
- 采集指定地区的汉语方言数据
- 通过手机自带麦克录制
- 四川话 / 上海话 / 湖南话等8种方言
- 安静 / 吵闹环境录制

语音数据转写标注

- 中文方言、普通话
- 转写准确率98%

TABLE OF CONTENTES

百度数据产品矩阵


百度云入口 客户需求
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!