TiD 2017
质量竞争力大会

人工智能产品：质量保障方案探索

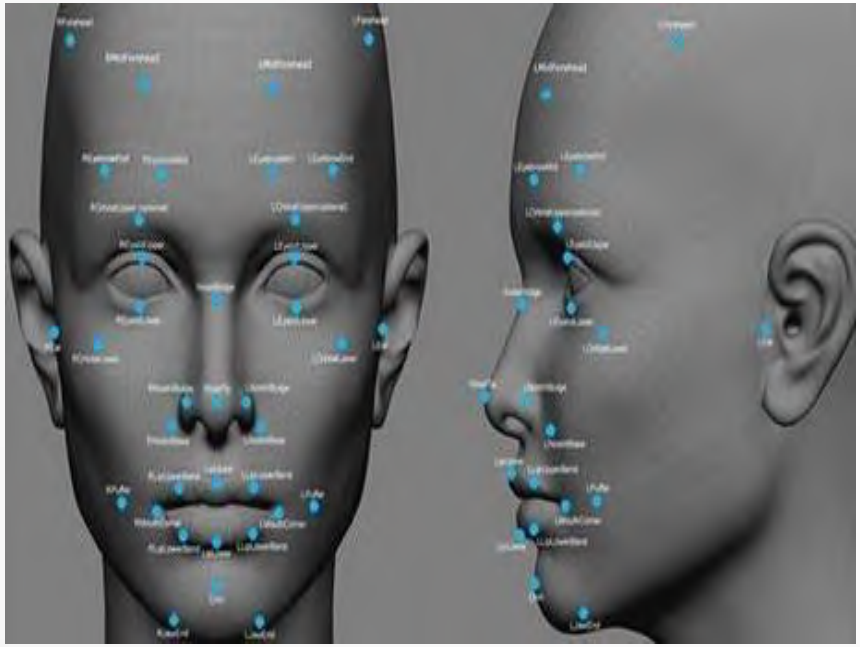
2017.7　　钱承君

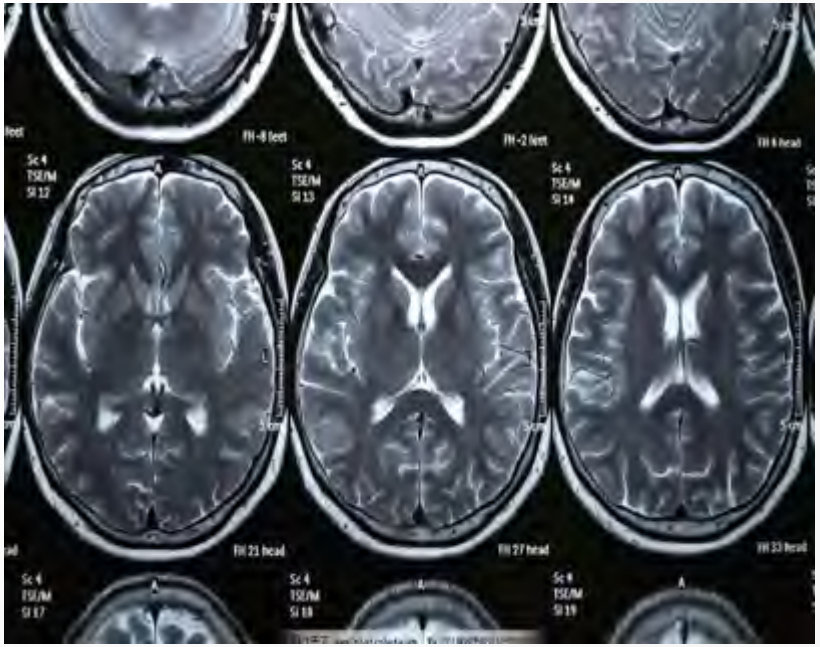
典型人工智能产品：无人车

典型人工智能产品：机器人

典型人工智能产品：推荐系统

典型人工智能产品：图像识别





通常被谈及的质量保障范畴

流程控制	• 甘特图 / 看板 • 敏捷 / 持续集成
测试设计	• 等价类 / 边界值 / 因果图 • 异常 / 容错处理 / 安全
测试执行	• 探索性测试 • 自动化
测试分析	• 缺陷根因 / 收敛分析 • 性能瓶颈分析

传统测试方法的外延

众包、监控、用户体验评测

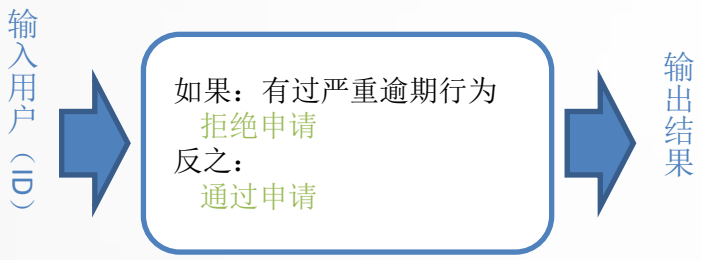
更**完整**、更**高效**、更**低成本**地

戳一戳，怼一怼，看看对不对

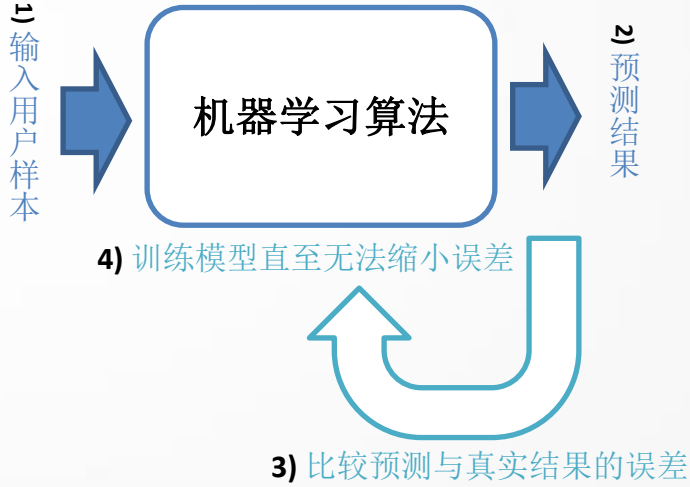
编外：简易版机器学习介绍

案例：判断是否准许特定用户的信用卡申请

传统思路（规则化思路）



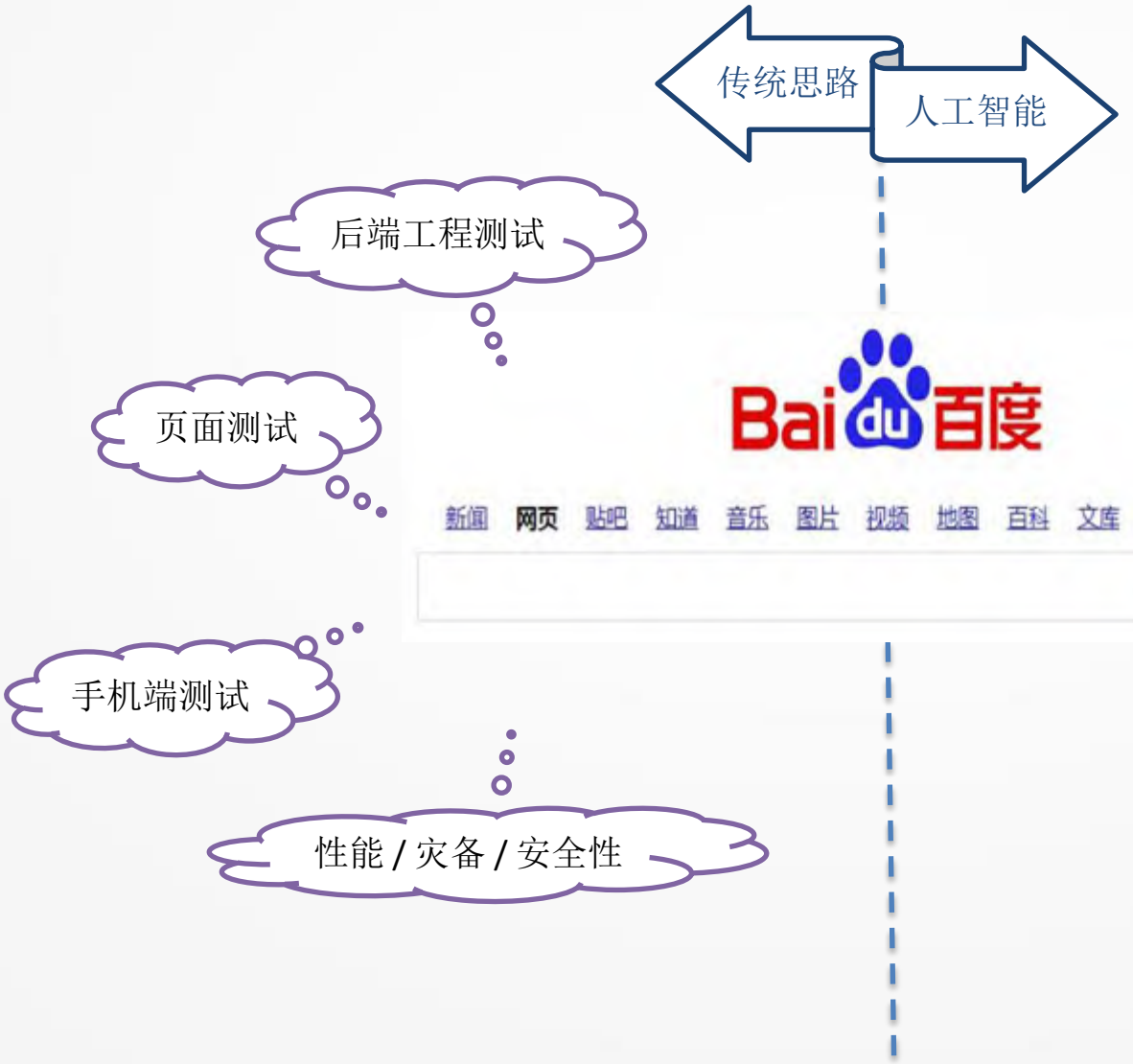
机器学习（持续自优化的思路）



“A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.”

-- Tom Mitchell, Carnegie Mellon University

传统软件 vs人工智能



找到所需信息

更多样的内容

更新鲜的内容

更少干扰和无效信息

TiD2017

人工智能产品：质量保障思路初探

	验收标准	评测数据	评测手段	结果分析
语音识别	近场识别 **字准率 > 95%** 说明：识别结果与标注答案对齐后，去除掉插入、删除、替换三种错误后，字级别正确的比例	1）手百多模线上采样数据 2）输入法线上采样数据	1）设备按实际使用场景摆放，正交人工做测试输入获得结果数据 2）通过软件手段合成与回放音频，获得更大批量结果数据	1）多测试方法间结果趋于一致、或有合理解读 2）与历史版本比，结果趋于一致、或有合理解读 3）测试数据分布吻合用户实际使用分布 4）修正和解读指标计算、样本选择带来的结果偏差
	远场识别 **五米，字准 > 92%** 说明：距离麦克风一定距离场景下，模拟家居场景摆放，综合得出评估结论	1）现场人工录制数据 2）软件合成数据 3）加噪数据，噪音源包含聊天、电视节目、音箱播放音乐、环境噪音		

基础验收指标

Cat detector： return TRUE when there is a cat.

	实际是猫（P）	实际不是猫（N）
判断是猫（T）	TP 	FP 
判断不是猫（F）	 FN	 TN

precesion = TP/(TP+FP)

判定为猫的图片中，有多少比例是真的猫

recall = TP/(TP+FN)

实际为猫的图片中，有多少比例被判定为猫

F = 2RP/(R+P)

accuracy = (TP+TN)/(P+N)

是猫的判定为猫，不是猫的判定不是猫

Specificity = TN/(TN + FP)
Missing Alarm = FN/(TP + FN)
False Alarm = FP / (TP + FP)

TiD2017

案例：色情图片检测

场景	指标	数值
色情	召回率	96%
	准确率	98%
	F-Score	0.97
正常	召回率	98%
	准确率	84%
	F-Score	0.90

1000张色情图，混杂正常样本若干

总共有1000张色情图，判断对了960张，占比96%，漏判了40张
所有被判为色情的图有980张，判对960张，判错20张，判对占比98%

1000张正常图，混杂色情样本若干

总共有1000张正常图，判断对了980张，占比98%，漏判了20张
所有被判为正常的图有1167张，判对980张，判错187张，判对占比84%

	实际色情（P）	实际正常（N）
判断色情（T）	960	20
判断正常（F）	40	?

	实际色情（P）	实际正常（N）
判断色情（T）	?	20
判断正常（F）	187	980

场景一：图搜下架色情图片

场景二：贴吧禁发色情图片

TiD2017

线上曾经存在的问题（图搜色情）

用户向的验收指标（User Side Matrix）

预判的

实际发生的

色情图片召回率

80% → 95%

漏检数量降低四倍

即使只有少量色情图片存在，会因以下搜索策略被推高，实际展现量远高于推算预期。

图搜色情图片流量

2% → 0.5%

同比降低四倍

图片点击率高

判定为受欢迎

前置展现排序

获得更高点击

备注：非真实数据

TiD2017

指标制定过程常见错误

指标与用户需求偏离

- 色情图片检出率 vs 色情图片误检率

指标不对终端用户负责

- 色情图片检出率 vs 线上色情图展示率

忽视指标间联动与完整

- 准确与召回不可兼得
- 复杂模型带来更好的效果，但有更大的功耗

评测数据集举例

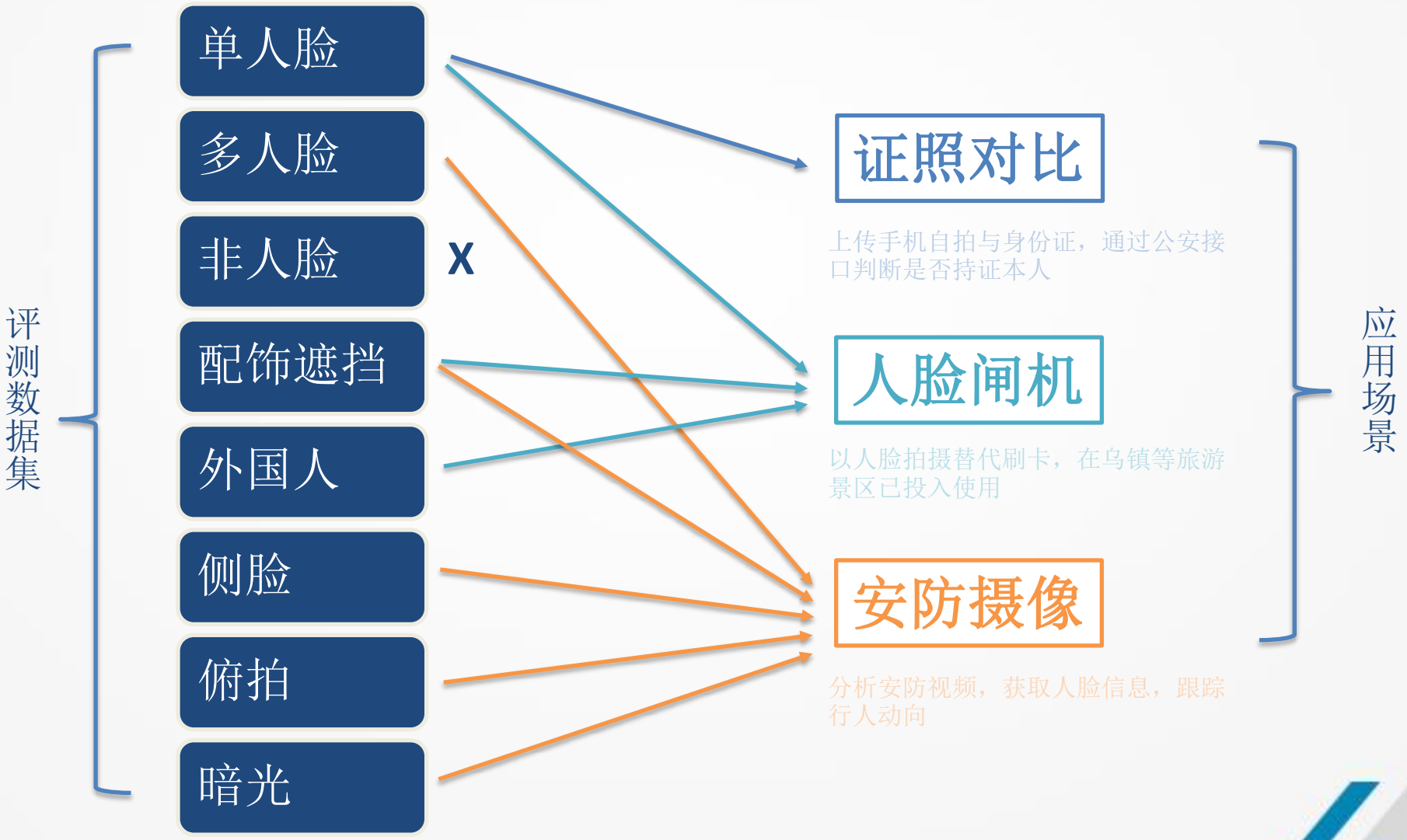
语音识别

- 发音特质：男女老少，口音，声调，语速
- 语言内容：语种，长句，中英混合，专名
- 噪音环境：人声，交通，空调，电视，音乐，通讯设备
- 应用场景：车载离线，远场识别，麦克风阵列

人脸检索

- 基础：多人脸，人脸尺寸
- 干扰：遮挡，侧脸，帽子，眼镜
- 光源：暗光，反光，曝光过度，隔玻璃反光
- 图像：模糊，失焦，黑白
- 细分：老人，外国人，卡通人脸

针对场景构建评测集


评测数据集
单人脸
多人脸
非人脸
X
配饰遮挡
外国人
侧脸
俯拍
暗光
证照对比
上传手机自拍与身份证，通过公安接口判断是否持证本人
人脸闸机
以人脸拍摄替代刷卡，在乌镇等旅游景区已投入使用
安防摄像
分析安防视频，获取人脸信息，跟踪行人动向
应用场景

TiD2017

评测数据获取

人工标注

- 自标注
- 自建标注团队
- 众测模式

数据复用

- 行标数据，采购数据
- 线上数据 / 用户数据

数据合成

- 生成不同噪音背景下的语料
- 生成同一人脸不同曝光下的图像

案例一

优化思路：用“验证”代替“标注”

案例二

优化思路：用“模拟”代替“采集”

TiD2017

评测集的拟合

模拟数据与实测数据拟合

- 通过模拟手段扩容样本
- 该场景下的实测数据
- 对比关键指标分布一致性
- 修正数据合成方案

评测样本集间拟合

- 多样本集间获取交集
- 比对交集部分的偏差，标识相对置信度
- 对部分评测集进行修正取舍

案例一

实测方法：距离设备0-5米距离，固定摆放，每个关键点不同录制人多次朗读语料。

模拟方法：通过近场（0米）录制语料，计算得出不同距离的语料情况。

当上述结论趋于一致，认为模拟得出的数据有较高的仿真度，可用于验收。

案例二

样本一：公司 ERP 系统导出信息。（性别/年龄）

样本二：合作友商导入CRM数据。

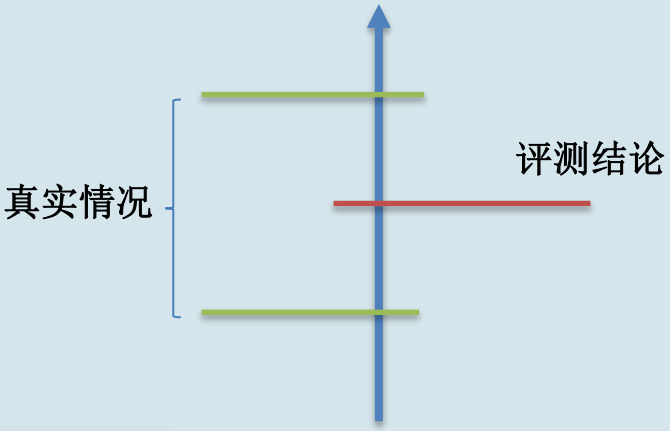
样本三：用户注册时自填信息。

样本四：依据用户行为标注推导出信息。

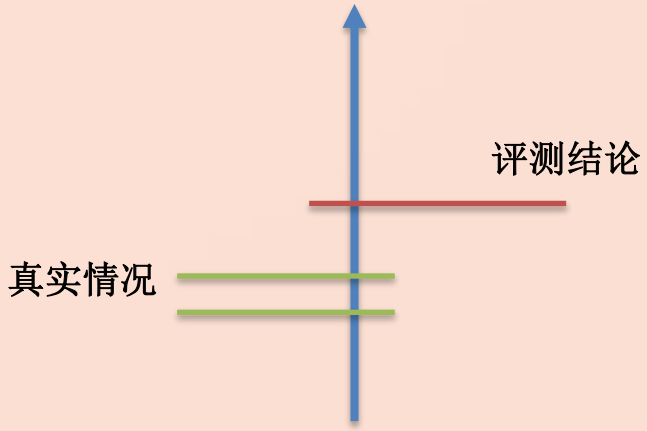
各用于评测的样本来源不一致，置信度不一致。依据样本间相对关系，做为样本置信分析的依据。

评测样本选择

样本数量



样本分布



TiD2017

评测数据选取常见错误

样本数量

- 数量量过小，例：拿个位数样本判断准确率
- 数据偏差，例：一千张图片中只有五张色情图

样本分布

- 样本泛化，例：导航和智能音箱场景下的中英混合识别
- 场景偏差，例：部署人脸识别时隔着玻璃
- 正负例数量不对等，例：人脸闸机需要彩照等负例

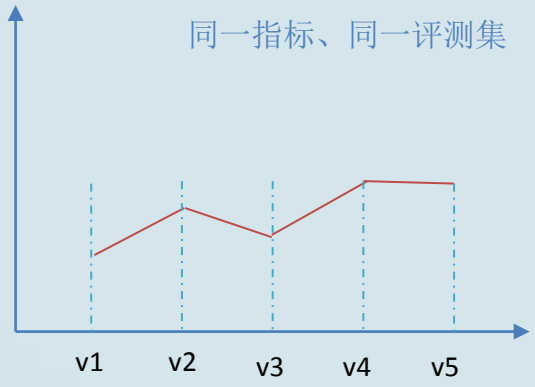
评测集选择

- 过拟合，测试数据直接用于模型训练
- 把自身优势场景用于竞品对标，不做交叉验证

评测结果：可连续对标

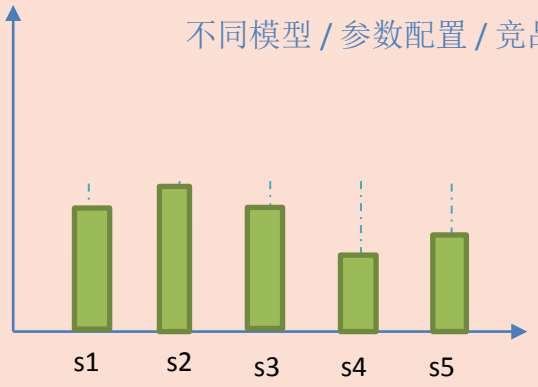
纵向对标

同一指标、同一评测集



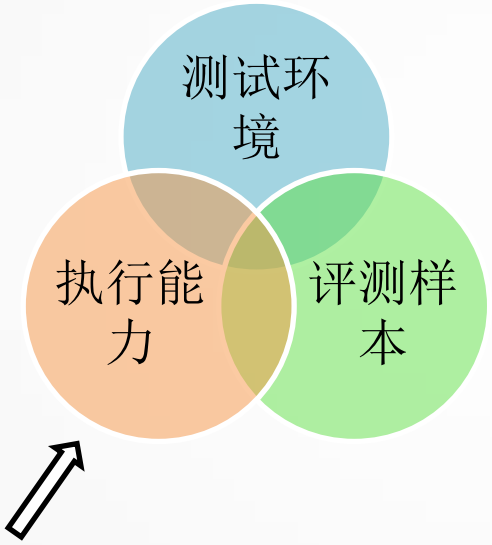
横向对标

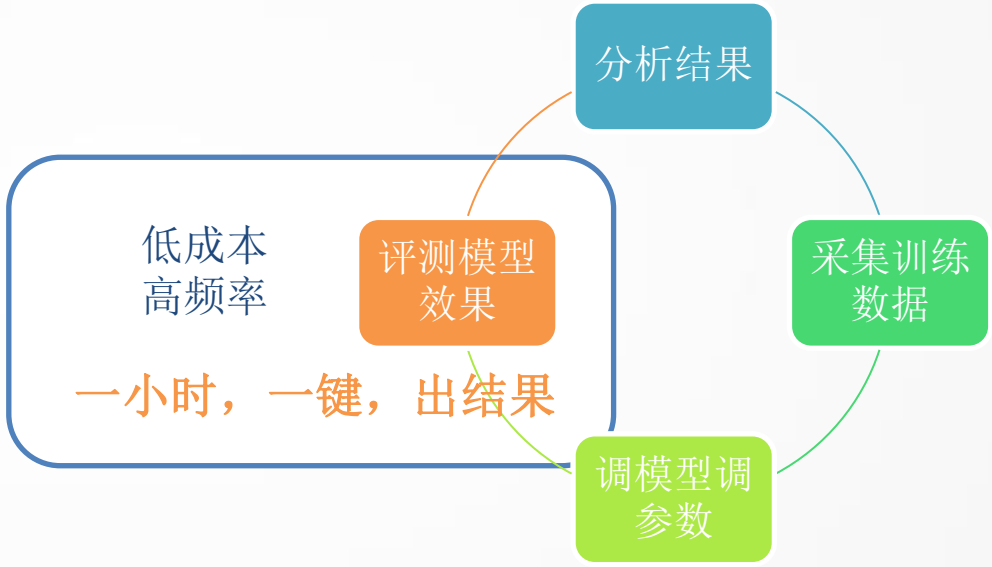
不同模型 / 参数配置 / 竞品



常见问题：前后测试报告给出差异很大的结论，数据不连续

评测过程：一键执行



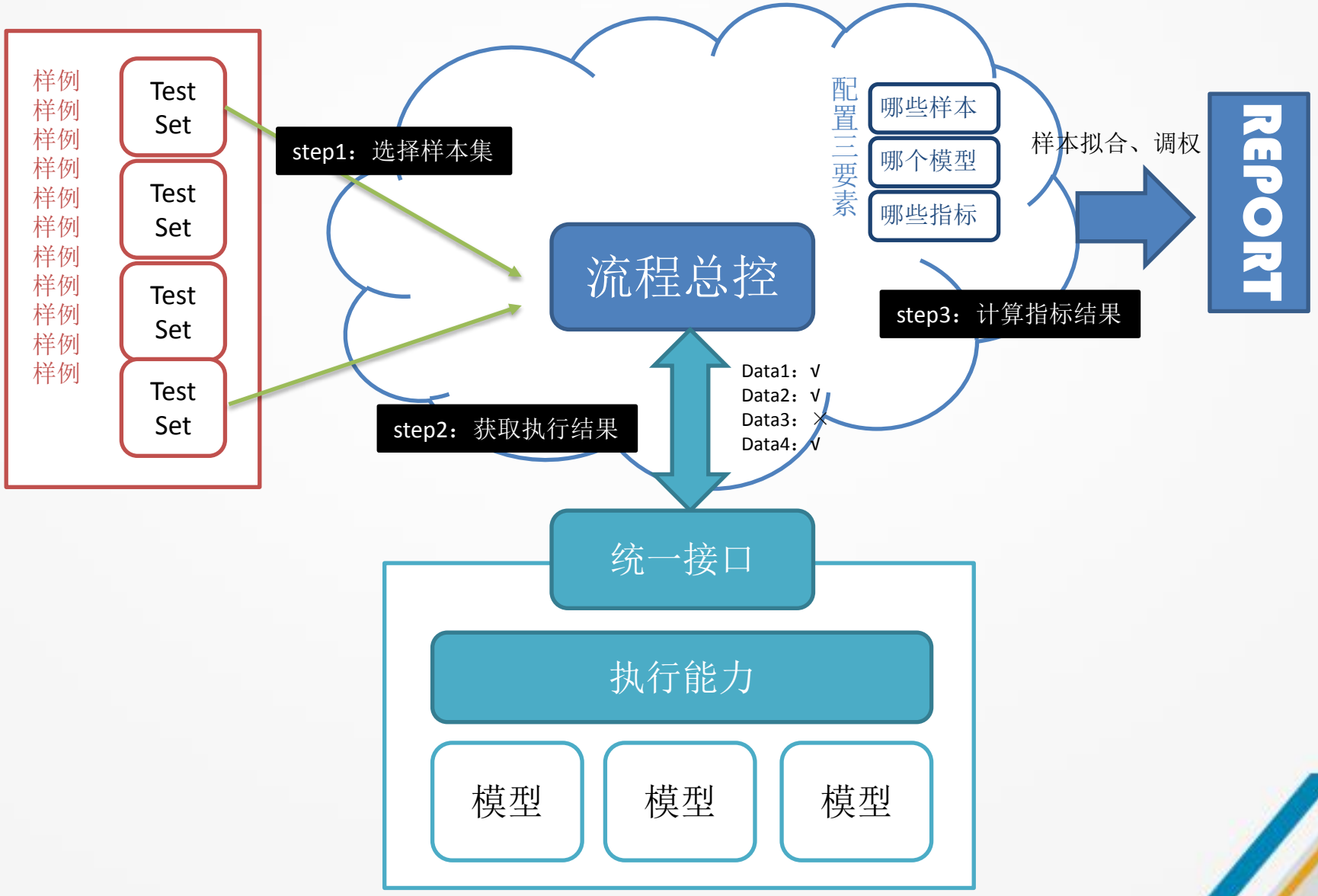


案例：语音语料复用

- 软件注入
- 耳机线导入
- 人工嘴

Learning Loop

评测系统的实现思路


样例
样例
样例
样例
样例
样例
样例
样例
样例
样例
样例
Test Set
Test Set
Test Set
Test Set
step1：选择样本集
step2：获取执行结果
流程总控
配置三要素
哪些样本
哪个模型
哪些指标
样本拟合、调权
step3：计算指标结果
REPORT
Data1：√
Data2：√
Data3：×
Data4：√
统一接口
执行能力
模型
模型
模型

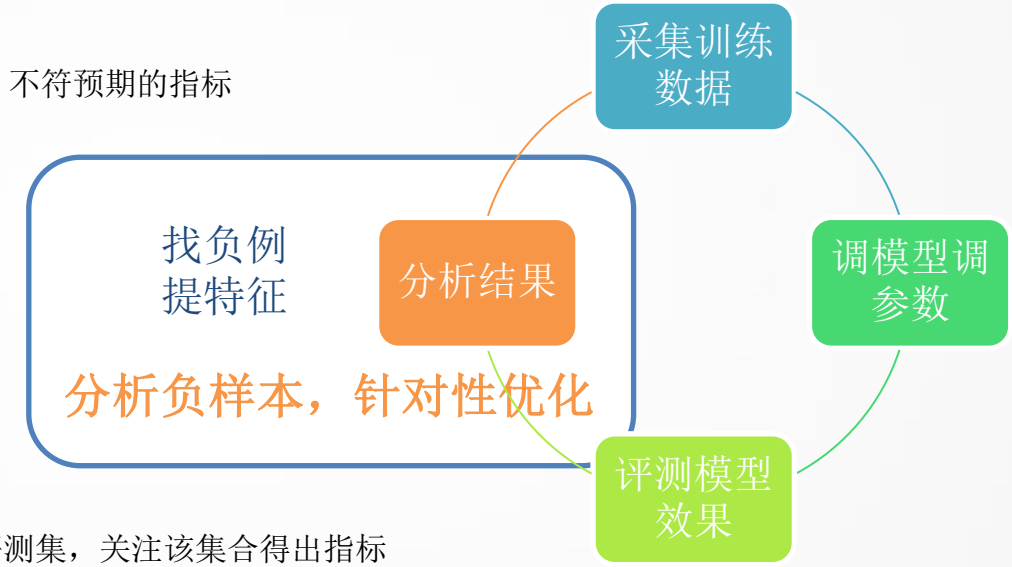
结果分析

找出指标　横向/纵向比对有较大波动，不符预期的指标

找出样本　正确变错误，错误变正确

分析负例特征　例：外国人识别率低

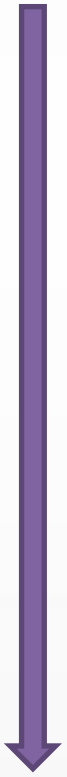
细分评测集　针对外国人建独立评测集，关注该集合得出指标



执行过程变体：针对邻接版本差集，抽样标注评估

质量保障过程类比

传统软件产品	人工智能产品
需求分析	验收标准
测试设计	评测数据
测试执行	评测手段
测试分析	结果分析



关键点：场景理解 + 测试设计

务虚谈：行业走向的个人观点

真正传统的领域
例如：航天所

软件工程化，外包盛行
例如：CMMI

互联网谈敏捷，自动化盛行
例如：Selenium

本质未变，基调收敛
承认手工测试，不同技能组合

测试设计 → 研发流程 → 自动化 → 移动应用

下一幕是什么？

感谢

THE END