



人工智能多源异构大数据处理实践经验分享

刘岩

泰康保险集团股份有限公司



报告主题



云计算



大数据



人工智能



融合应用

1. 科技浪潮是普惠的商机

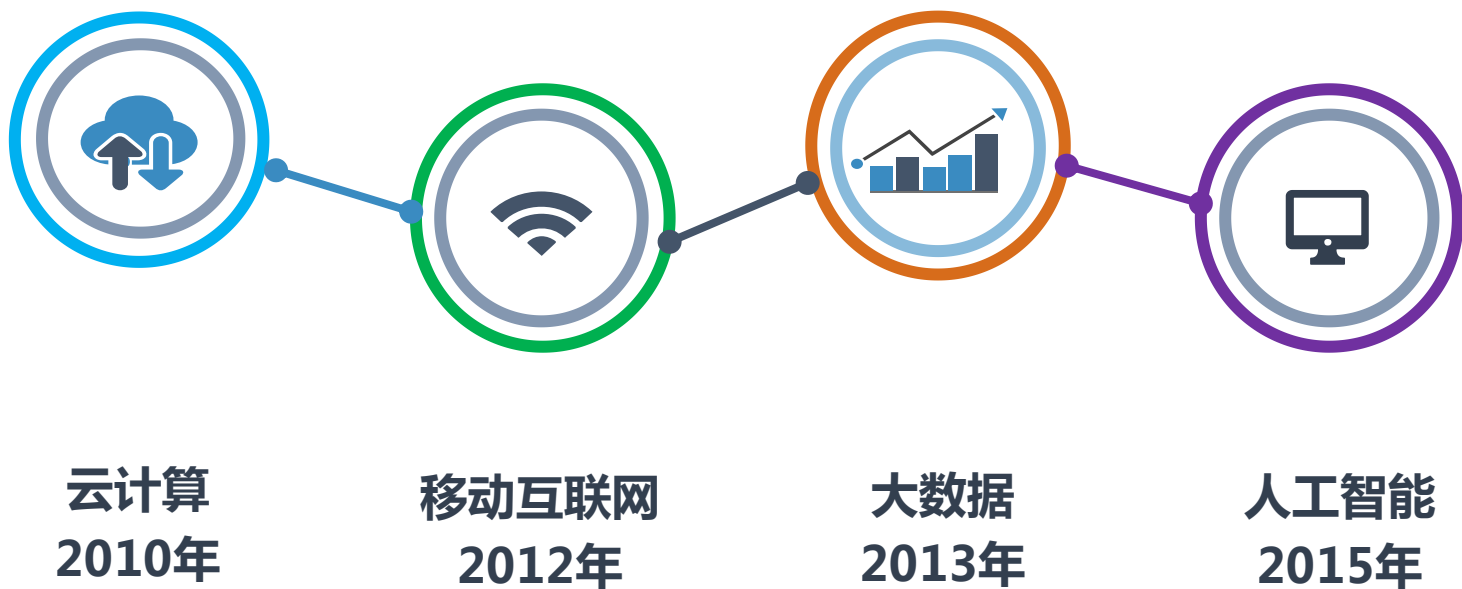
2. 智慧城市大数据人工智能处理

3. 金融保险大数据人工智能处理

科技浪潮是普惠的商机

Open your eyes on science

近年科技浪潮主题的演变



曾经备受争议的“炒作”，已经变成了普惠的科技与商机！

科技生产力普及！= 炒作



激活投资

资金向社会主驱动力流动



创新政策

政策开放
灵活度提升



优化产业

行业界限
逐渐消退
服务导向
性行业增加



普惠大众

享受科技
改变生活的
成果

大数据不同行业需求的差异

1

政府
大数据

痛点在多源，数据缺标准，大数据交换与融合平台建设优先

痛点在异构，种类繁多，非结构化大数据处理需求优先

保险
大数据

2

智慧城市大数据AI处理

Open your eyes on city

智慧城市大数据运营中心建设的外在形式

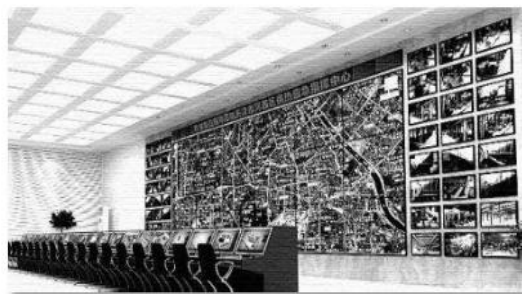
三个中心：云计算&数据中心、城市运营&管理中心、体验&展示中心

三个平台：数据共享交换平台、大数据挖掘与分析平台、数据网格可视化平台

四个数据库：公共基础数据库、公共服务数据库、公共业务数据库、特需情报数据库



云计算&数据中心



城市运营&管理中心



体验&展示中心

智慧城市大数据运营中心建设的内在需求

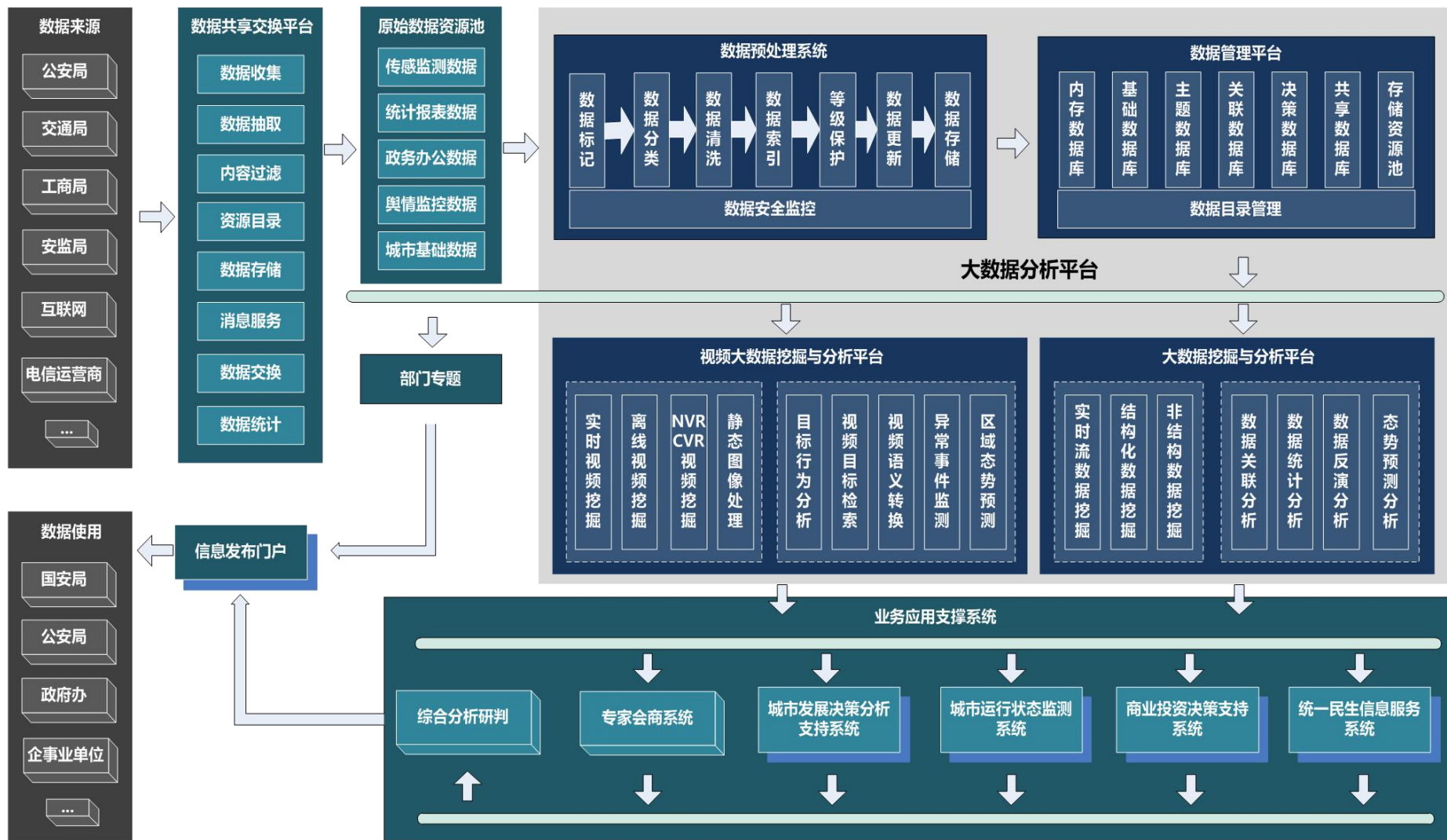
政府行政部门

面向城市运营的数据集市、
数据报表、数据共享查询

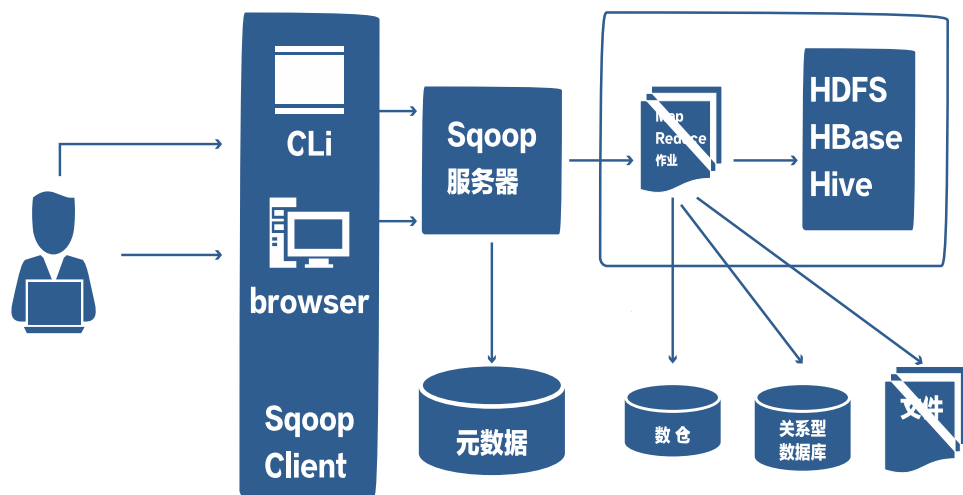
面向刑侦、技侦的知识推
荐、数据关联检索

政府安全部门

智慧城市大数据运营中心建设一般框架

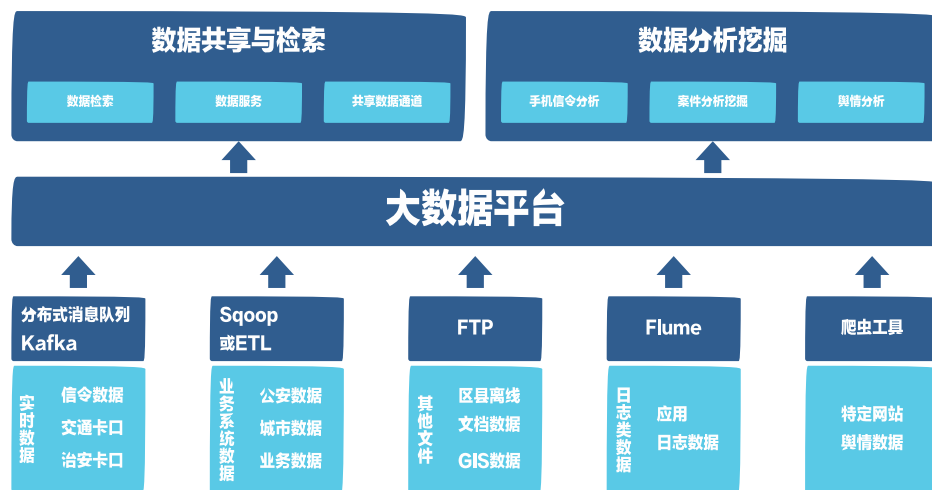


多源大数据的抽取问题



拥抱HADOOP生态
及各类开放式工具

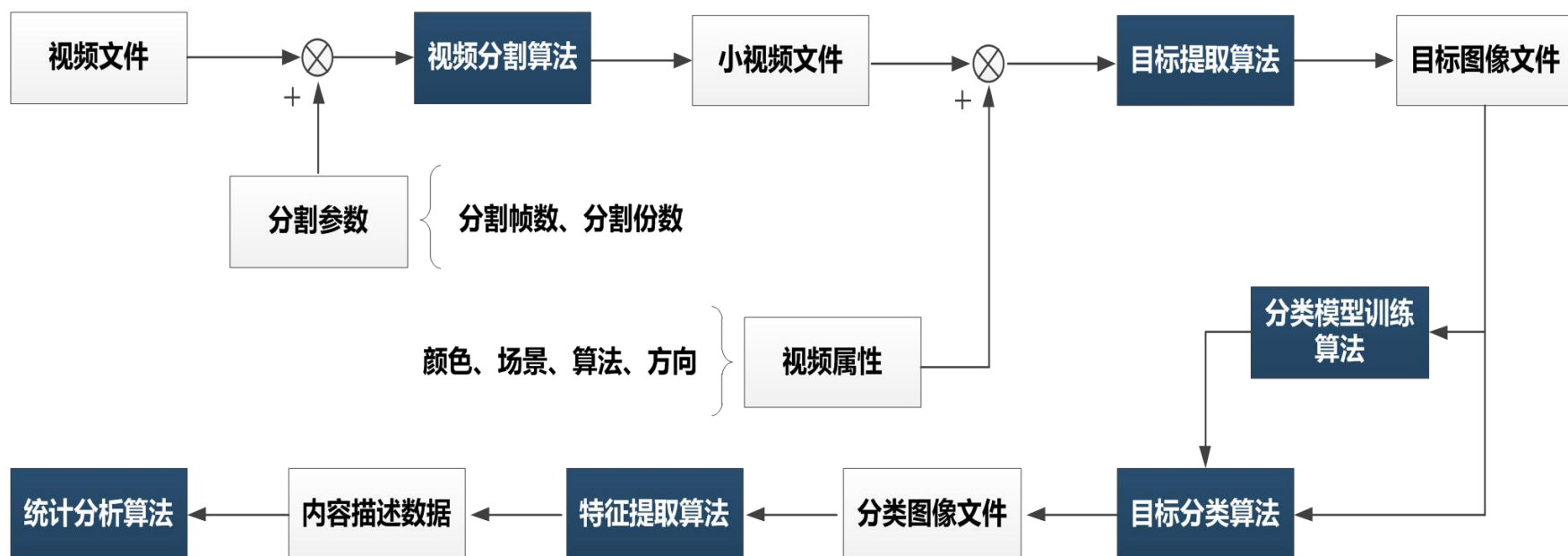
优先解决数据收集与共享
问题



非结构化视频数据在HADOOP中的处理问题

1、实时视频流截取：帧累加 & 分镜

2、离线视频文件切分：帧累加 & 分镜 & 关键目标



大规模小图像文件的存储问题

1、常规方案：Fast-DFS

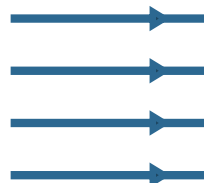
2、改进方案：基于HBASE结构的文件合并存储



一级索引



二级索引



三级索引



数据索引



航班数据



通讯数据



住宿数据

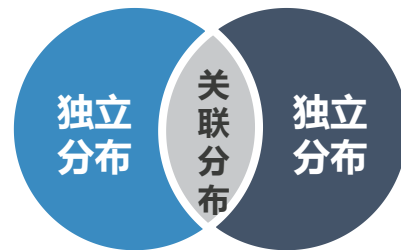
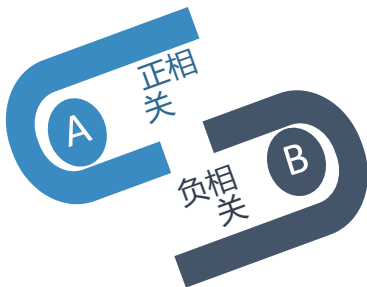


快递数据

不同来源数据的可信计算问题

数据的可信性判定、不同置信数据的融合处理

数据可信度标定



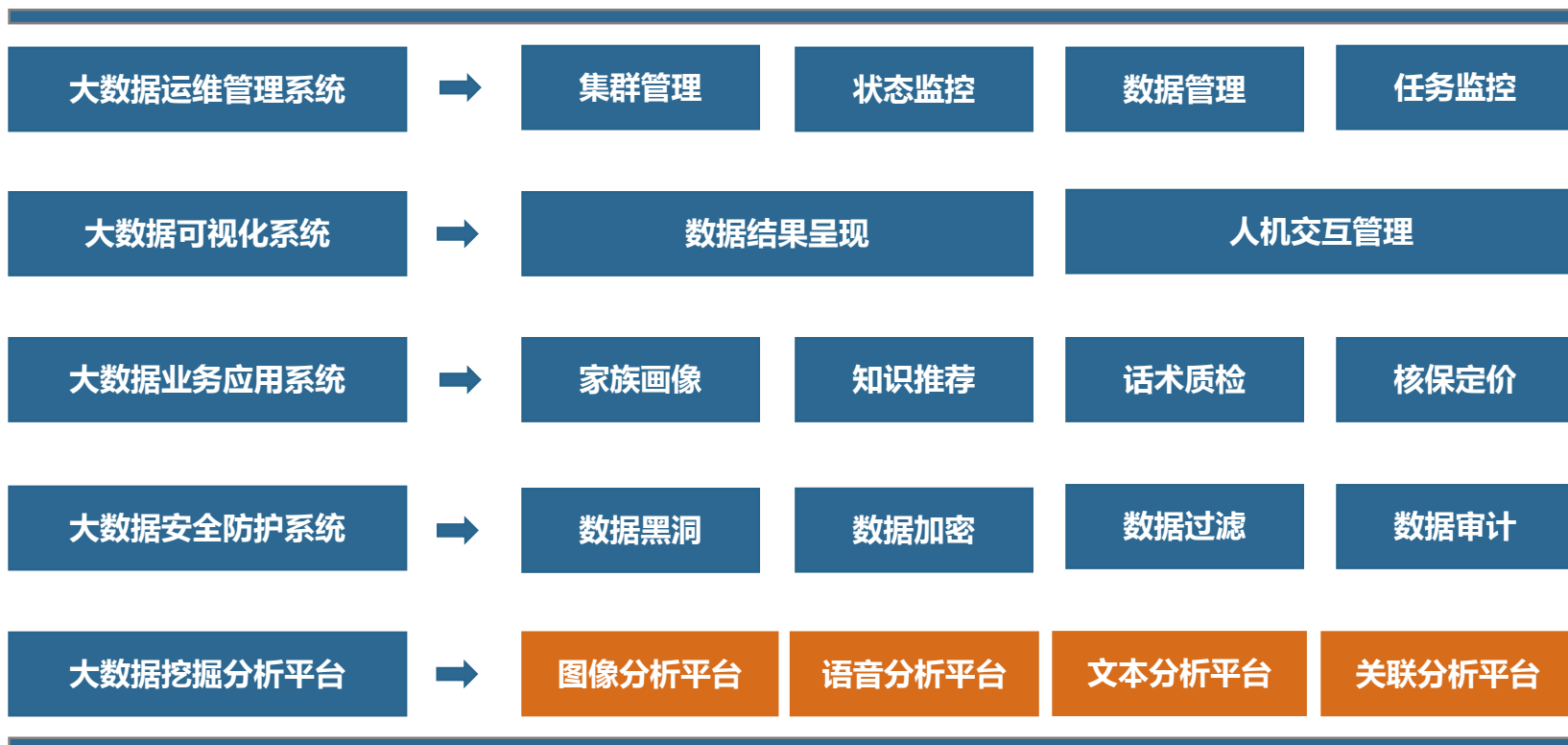
- 根据相关假设设计可信度计算模型：高斯分布、离散高斯积分分布等；
- 设定数据标尺将所有数据归一化到同一量纲和标准点；
- 可信模型设计与训练；
- 可信度计算与结果归一化。

金融保险大数据AI处理

Open your eyes on insurance

金融保险行业大数据应用热点

(1) 盘活数据资产，挖掘潜在客户；(2) AI附能，革新业务运营模式。



保险行业影像数据分析重点关注的内容

以数据密集型计算为主，对CPU主频、GPU板卡数要求较高。



人脸识别

身份认证与识别

业务办理自助
线下向线上迁移



字符识别

承保核保理赔

自动抽取数据
减少数据人工录入



影像分类质检

提升效能

影像资料完备性核实
影像资料质量检查

清醒地把握建设内容的成熟度及其相对性，不成熟的算法在适当技术架构和合适场景下是成熟的。

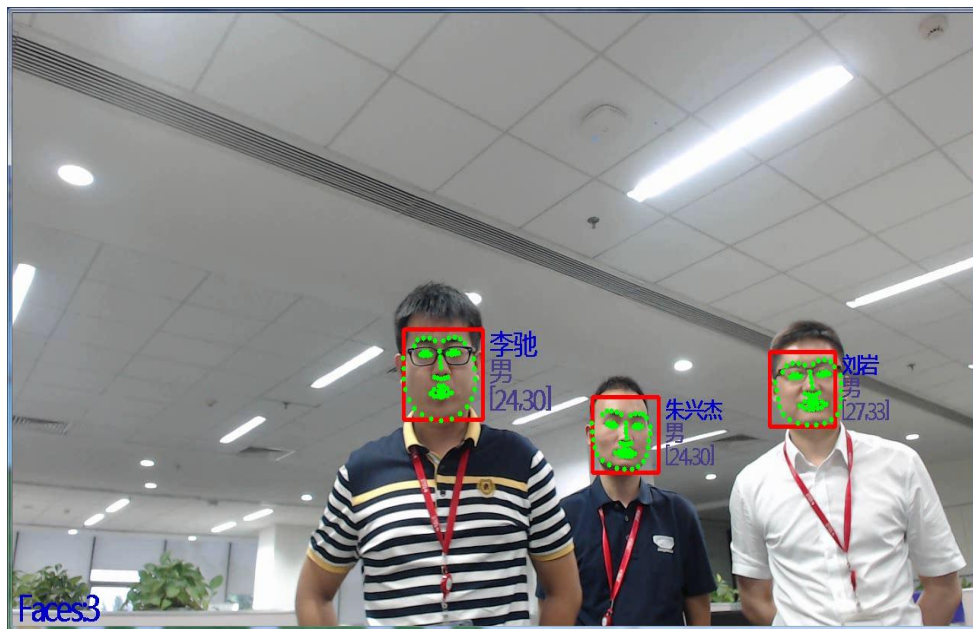
影像数据处理之人脸识别



人脸识别

身份认证与识别

业务办理自助
线下向线上迁移



人脸识别广泛应用于投保、核保、理赔、保全等保险业务中，主要用于打通线上业务自助办理流程。（含身份、年龄、性别、表情等）

影像数据处理之字符识别

人民邮电出版社
POSTS & TELECOM PRESS

IT大咖说
知识分享平台



字符识别

承保核保理赔

自动抽取数据
减少数据人工录入



医疗发票OCR中的药品名要基于全国医保目录库进行修正。

影像数据处理之分类质检

影像分类质检

提升效能

影像资料完备性核实
影像资料质量检查

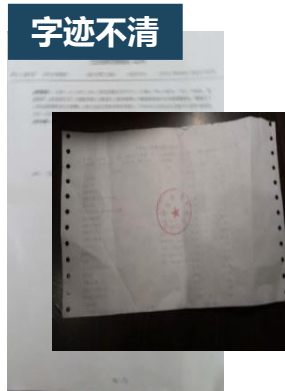
图像模糊



多张混拍



字迹不清



篡改+遮挡



← 影像分类

上传影像资料

选图

拍照

上传影像

及时发现不合格影像，业务处理时效由1~3天提升为几秒钟。

AI与影像大数据融合应用之健康预测

影像分类

内容OCR

文本理解

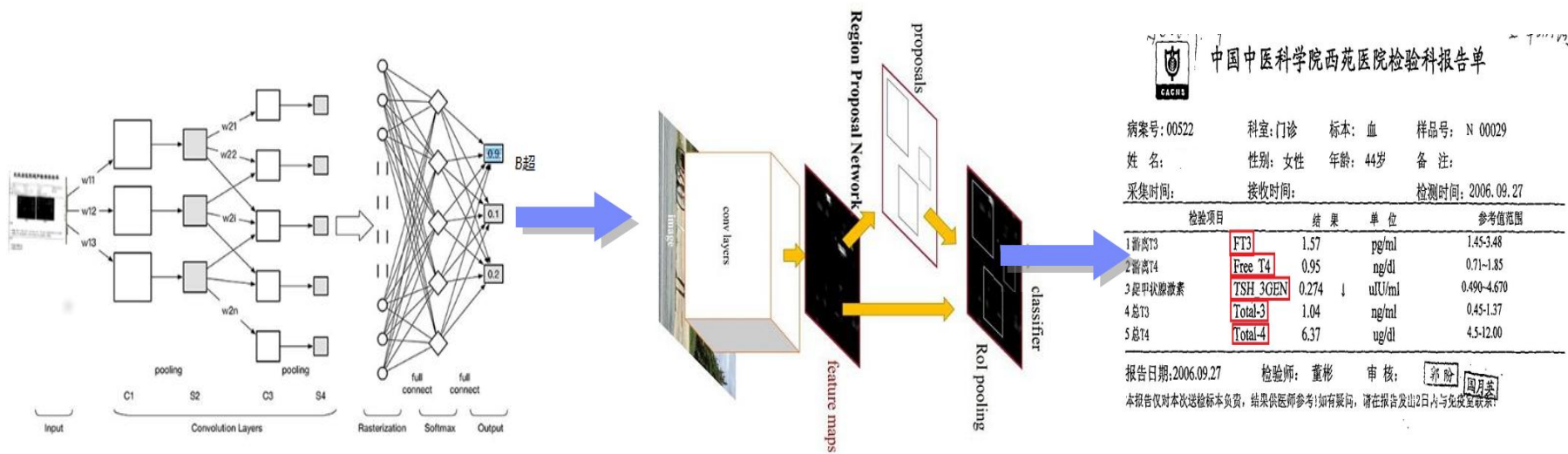
健康预测

卷积网络

切割识别

医学模型

神经网络



甲状腺自动核保诊断示例

保险行业语音数据分析重点关注的内容

需要缓存声学模型、语言模型、语义模型等，对内存要求较高。



语音识别

语音、语种、声纹

语音识别
身份认证



语音合成

个性交互

语音播报
个性合成



语义理解

语义映射

意图理解
精准回复

(1) 行业标准尚未统一，音频格式、设备接口等差异较大。

(2) 接口定义时需要注意不同格式和设备型号的支持。

AI与语音大数据融合应用之导航质检

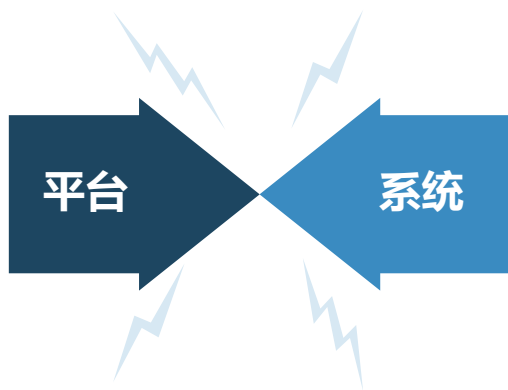


- 每天1万余名客服与电销人员录音的识别、分析质检。
- 每天几万次智能客服对话的实时识别、理解与应答。
- 每秒180余次导航语音的识别、理解。

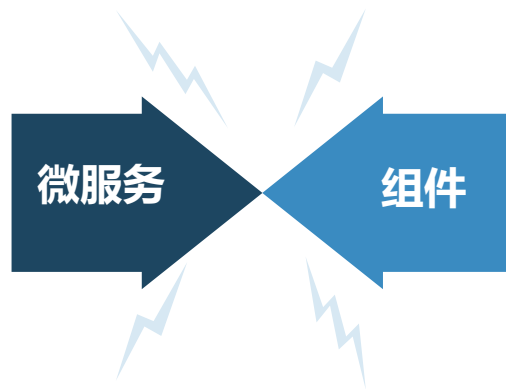


AI类架构设计需要注意的若干问题

将能力层和业务层分离；各层内功能单元微服务化。



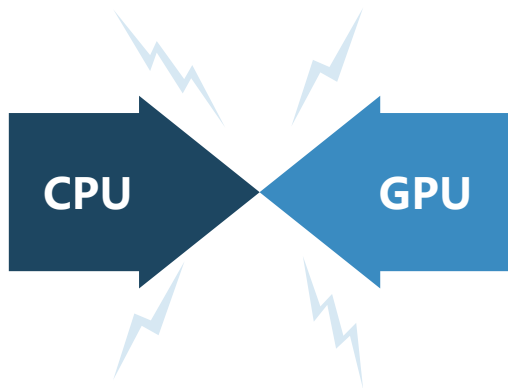
服务平台化，接口标准化



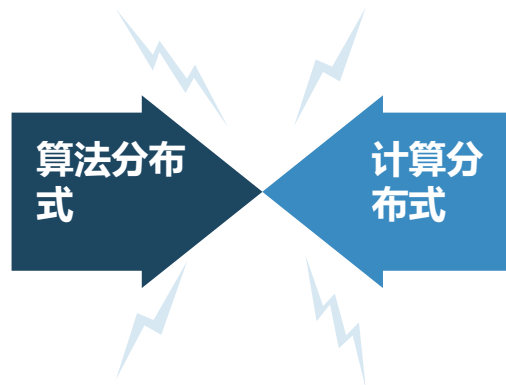
功能层次化，微服务化

以人脸识别技术为例，应用的系统有几十个，并且涵盖了公安部1：1、本地化1：1、本地化1：N、有感活体、无感活体、防照片攻击等，如果按垂直系统建，仅维护升级的难度就非常大。当面向整个行业提供服务的时候，如何办？

AI类**技术研发**需要注意的若干问题



主流 VS 热点



算法研发 VS 现成方案

- (1) 大多AI算法属于密集型计算，CPU弱虚拟化仍是主流需求。
- (2) 开源工具Tensorflow等让机器学习门槛降低，但要重视基础研究。
- (3) 算法分布式研发并未引起重视，传统数据分布式处理方式仍是主流。

报告完毕，欢迎批评指正！