

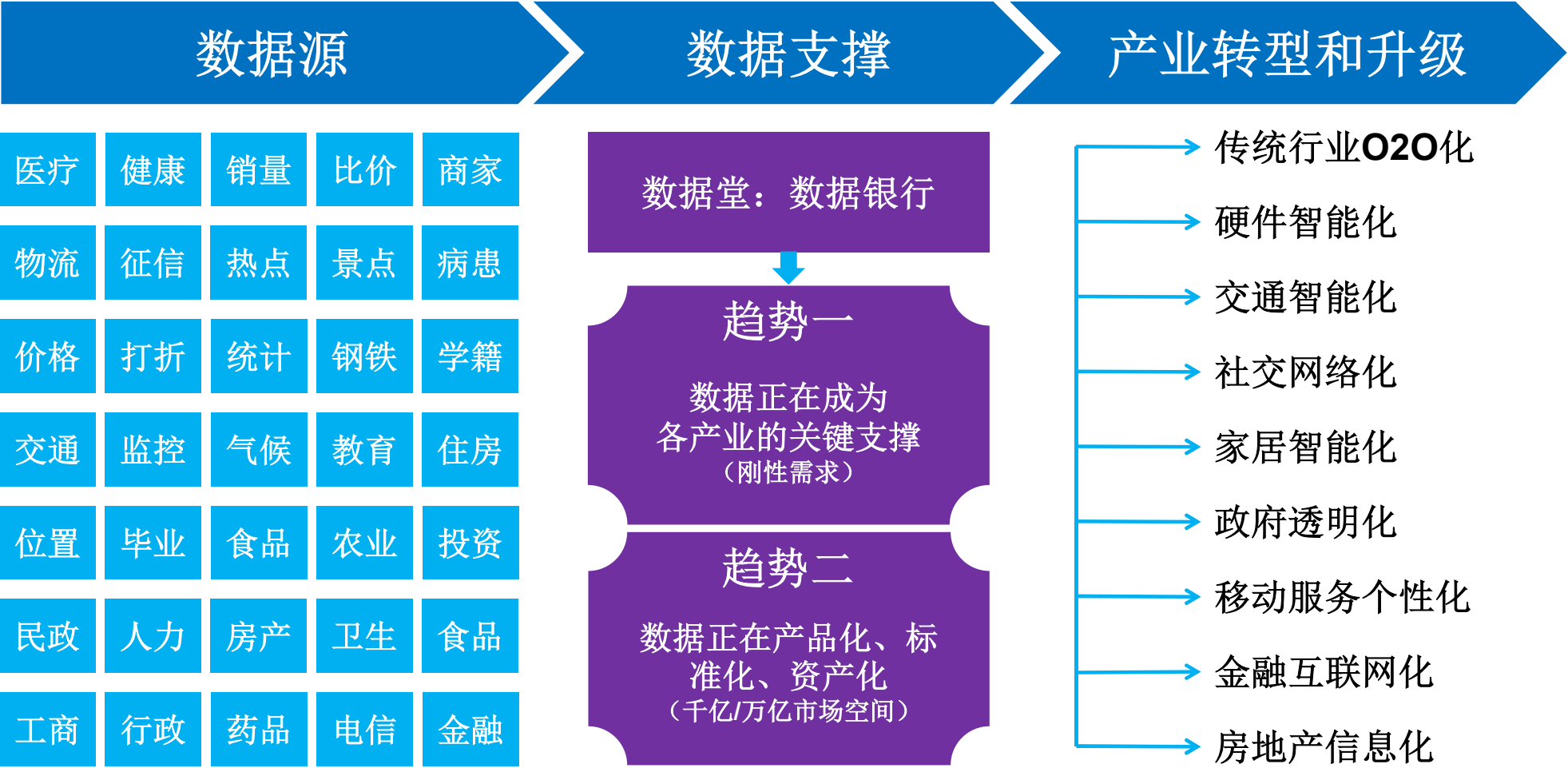


NEEQ: 831428

数据堂：数据银行

数据堂（北京）科技股份有限公司

定位：卡位数据资产，服务产业升级



数据银行：数据资源与获取方式

基础数据

有关位置、人、企业等基本数据，以便在空间、时间以及身份等维度上关联各个领域数据：

- 全国公民身份证信息
- 全国超精细格网化人口数据
- 企业基本信息数据（工商）
- 全国银联刷卡记录数据库
- 地图位置数据
- 全国天气、空气质量数据

IT智能化

移动终端、车载终端、家电、物联网的智能化需要数据+智能化技术：

- 图像识别：图片导购、安全监控、地图增值、图像检索；
- 语音识别：语音助手、语音导航、家电控制；
- 语义理解：人机交互、舆情监控、客服中心；

商家+商价

移动互联网、传统互联网业务O2O化，需要线下数据在物理世界落地：

- 家：家政、家装、亲子、教育、运动、健身、娱乐。
- 衣：商场、店铺
- 食：餐饮、美食
- 住：酒店、宾馆
- 行：车/船/飞机订票

医疗健康

医疗健康个性化、移动化、智能化需要整合医院、医生、医疗等数据：

- 医院、医生、药企、医疗器械企业基本信息采集与整合；
- 基于病历、体检等基本医疗信息的个人健康追踪；
- 基于个人健康的个性化医疗保险。

智能交通

交通涉及整个国计民生，其智能化、数据化、互联化需要数据支持：

- 实时交通路况
- 车载设备及车联网
- 停车、违章、公交、出租车
- 基于视频的交通安全
- 陆运、海运、空运

数据堂数据：45000套，500T

众包采集

独家



线下大数据(40万众客)

采购&代理

重点



行业大数据

数据爬虫



线上（互联网）大数据

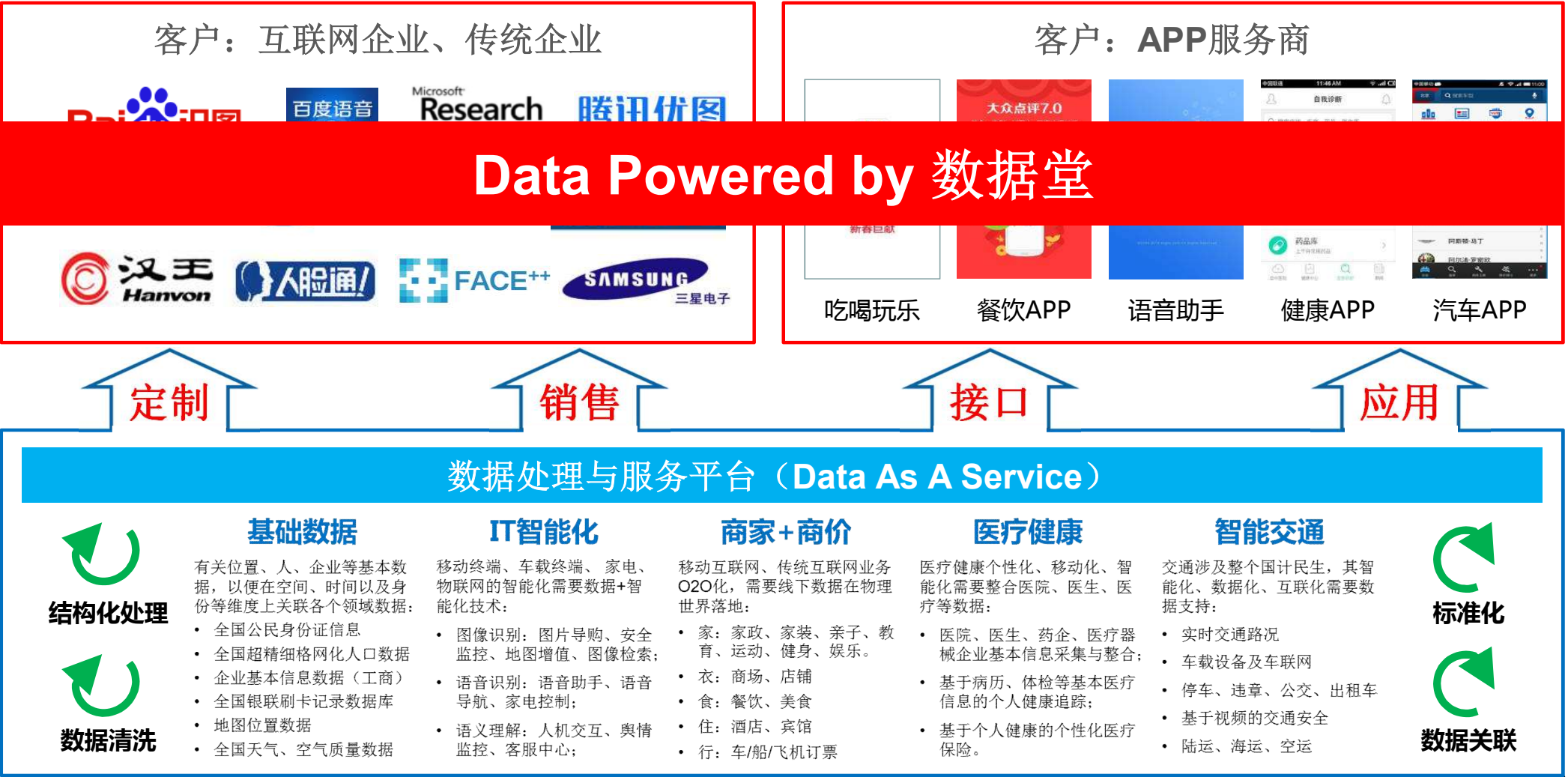
合作

布局



政府大数据

数据银行：数据处理与数据服务



数据获取优势 - 众包平台

数据堂众包平台 - 线下大数据



案例

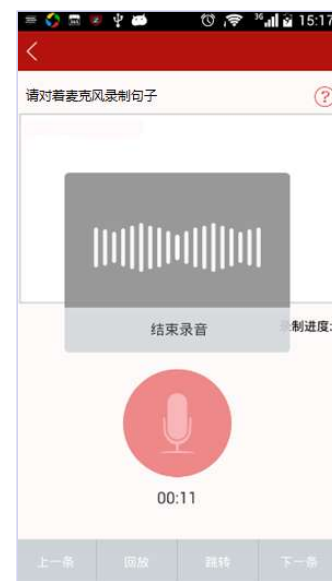
人脸图像



购物小票



语音录音



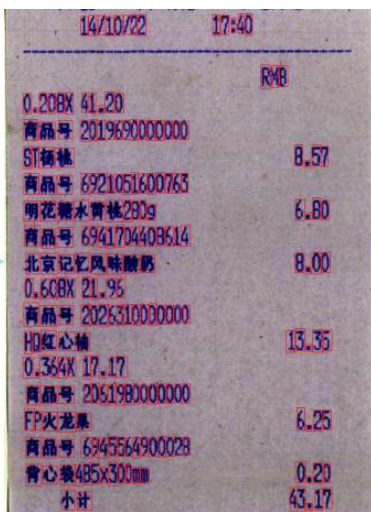
- 40万全球（以中国为主）实名兼职人员（众客）；
- 采集其它企业无法获取的独家线下数据；
- 采集案例：购物小票，商家店铺，脚尺寸，语音等；
- 为传统产业O2O化提供重要数据支撑。

核心模式：

- 保证数据可重复使用，可重复销售
- 保证数据堂拥有版权，有权利销售

技术优势：非结构化数据处理 – 小票识别

- 数据中**80%**是非结构化形态存在，包括语音、图像、视频、文本等，绝大多数客户无处理能力。
- 购物小票是反映线上商业活动的媒介，数据堂拥有小票（图像）**OCR**的核心技术。
- 已支撑**10万张**小票数据的处理，**100万个**数据标签。



购买日期	商品编号	商品名	购买数量	商品单价
2014/10/22 17:40	2019690000000	ST杨桃	0.208	41.20
2014/10/22 17:40	6921051600763	明花糖水黄桃280g	1	6.80
2014/10/22 17:40	6941704408614	北京记忆风味酸奶	1	8.00
2014/10/22 17:40	2026310000000	HQ红心柚	0.608	21.96
2014/10/22 17:40	2061980000000	FP火龙果	0.364	17.17
2014/10/22 17:40	6945564900028	背心袋485x300mm	1	0.20

支持复杂的拍摄场景

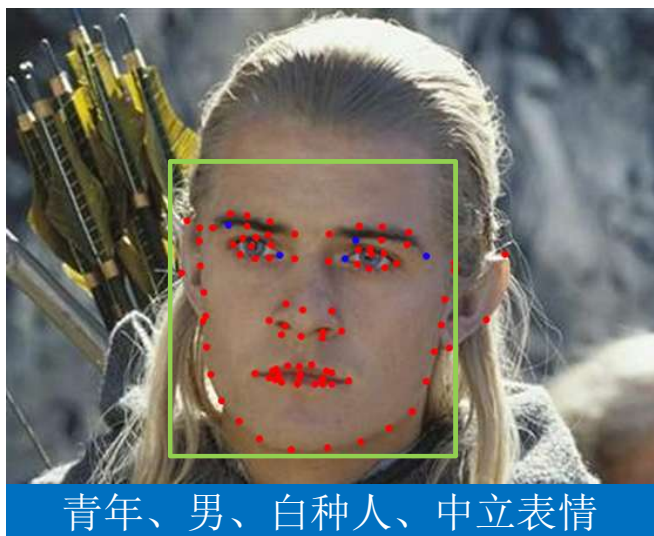
自动处理**倾斜**、**光线暗**等拍摄不佳的图片，识别率**90%**以上。

独特的语义**OCR**识别技术

在传统**OCR**文字识别算法的基础上，结合商品知识库的语义关联，识别率比传统算法**提高10%**。

技术优势：非结构化数据处理 — 人脸识别

- 数据堂拥有人脸识别的关键技术，其是图像处理的基本技术。
- 已支撑3000万张人脸图片数据的处理，4亿个数据标签。



青年、男、白种人、中立表情

支持不同粒度的人脸特征识别

识别人脸图片的人脸区域、表情、性别、年龄、人种、脸部关键点等**6种特征**，识别率**90%以上**

支持不同复杂场景

支持**不同姿态**（平、仰、俯、左、右、前）、**不同光线**（亮、中、暗）下的人脸识别

基于大量人脸图片数据的建模算法

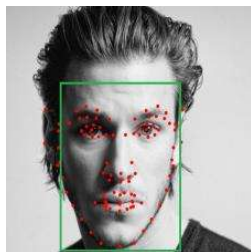
基于**千万级**人脸图片数据的深度学习建模



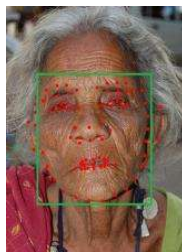
老人 黄种 女 微笑



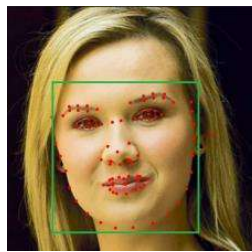
中年 黑种 女 中立



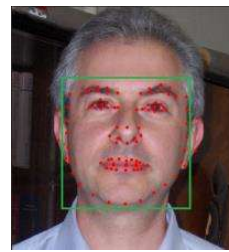
青年 白种 男 中立



老人 黑种 女 中立



青年 白种 女 微笑



老人 白种 男 中立



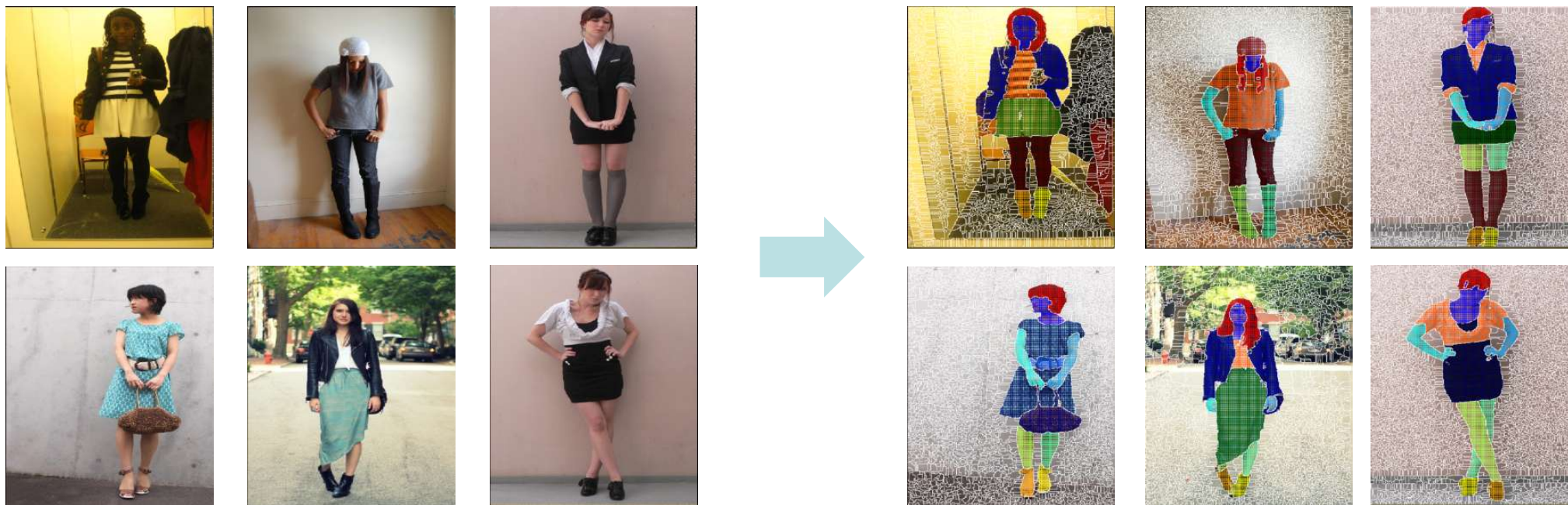
青年 黑种 男 微笑



青年 白种 女 中立

技术优势：非结构化数据处理 – 物体识别

- 数据堂拥有从图像中识别特定物体的核心技术，其是图片搜索和图片导购服务的基础。
- 已支撑**2000万**张物体图片数据的处理，**1亿**个数据标签。



精准的轮廓识别

精准识别帽子、头发、上衣、裙子、鞋子、包、眼镜等部位，识别误差小于10%

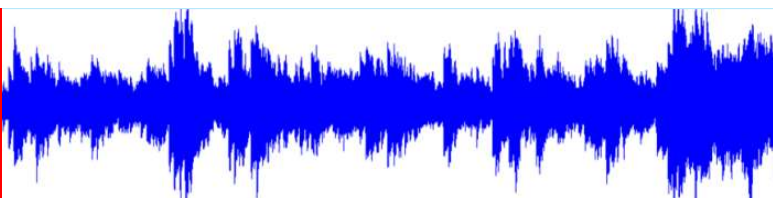
支持复杂的图片场景

可处理侧面、遮挡等复杂情况

技术优势：非结构化数据处理 – 语音/语义识别

- 数据堂拥有从语音中识别语音内容的技术，和从自然语言文字中理解语义的技术。
- 已支撑**40套**语音和**15套**文本数据的处理，**25亿**个数据标签

语音识别



明天天气非常好，我们去爬山如何？

支持多种语言识别

支持**普通话**、**英文**的识别，识别率**85%以上**。

支持不同场景的语音识别

支持安静、室内噪音、室外噪音和车载噪音等**4种典型场景**下的语音识别

基于大量数据的建模算法

基于**5万小时**语音数据及**T级**文本数据，利用深度建模算法的分布式建模

语义理解

凯凯**2.4AT**在冷车启动时，仪表盘上会有吱吱的声音，声音不大，但很烦人，开车的人会觉得郁闷。

品牌：丰田
车型：凯美瑞**2.4AT**
别称：凯凯，小凯

评价对象：
仪表盘

评价内容：有
吱吱的声音

倾向性：贬义
评价类别：噪音

支持多种语言的语义分析

支持**中文**、**英文**、**日文**的文本语义分析，识别率**85%以上**

三层语义理解模型

自动分析**句法结构**、**用户情感**、**查询意图**

可处理各类书面语和口语

可处理**新闻**、**短信**、**微博**、**微信**、**评论**、**论坛**等各种形式的文本

合伙人团队



齐红威, CEO

- NEC中国研究院, 研发部长
- 美国斯坦福大学, 访问学者
- 中科院自动化所, 工学博士



柴银辉, COO

- 2010-2013, 神州泰岳, COO
- 2003-2010, 腾讯产品、战略总监
- 1999-2003, 华为智能网副总工



丰强泽, CTO

中科院计算所, 硕士&博士
NEC中国研究院



肖永红, 数采总监

中国科技信息研究所, 硕士
中科院情报中心



贺纯洁, 销售总监

南开大学, MBA
天宇朗通通信



吕菊梅, 财务总监

北京大学, 注册会计师
西普阳光

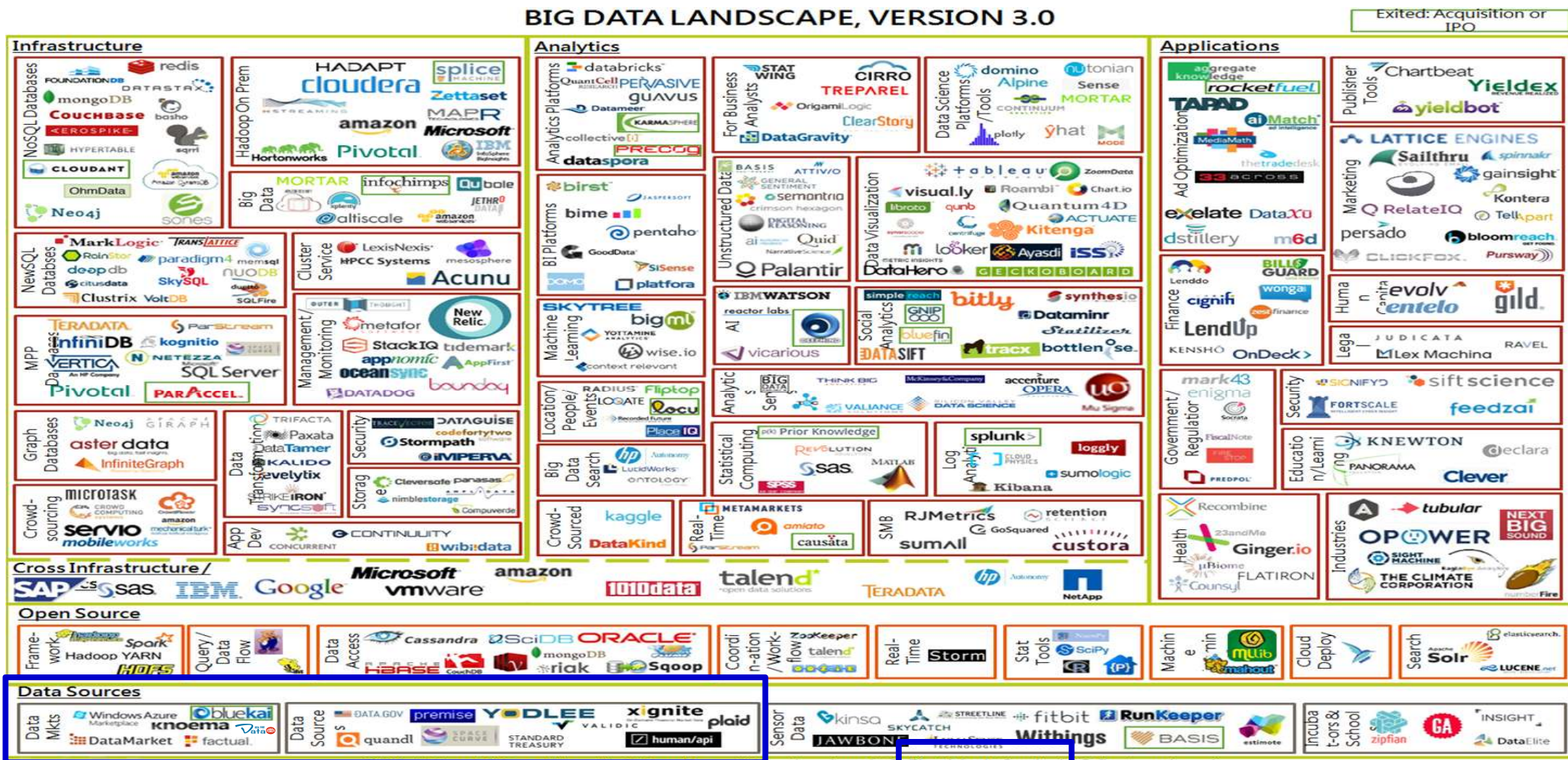


王建, 战略顾问

北方交通大学, 博士
10年公司创办经验

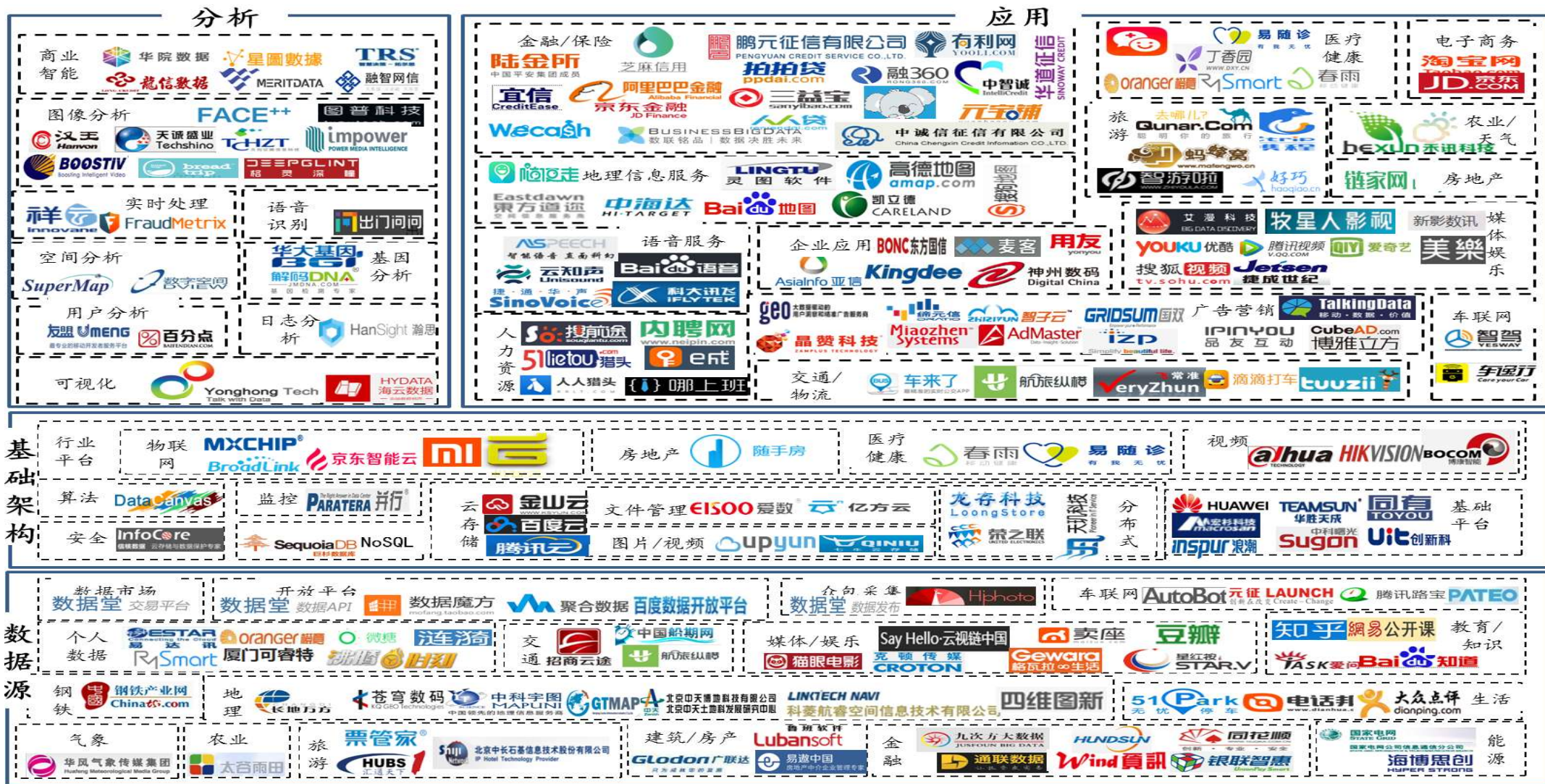
大数据生态链（美国）

BIG DATA LANDSCAPE, VERSION 3.0



© Matt Turck (@mattturck), Susan Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

大数据生态链（国内，数据堂独家版本）



数据堂业务定位（3）

线下数据

（众包采集、行业采购）

选择依据：

- 客户需要线下数据开展线上业务
- 和互联网企业能力差异化

线上业务

- 传统互联网企业
- 移动互联网企业

线下数据+线上数据

（众包采集、行业采购，互联网抓取）

不选择依据：

- 客户需求还处于尝试阶段
- 以传统IT解决方案为主

线下业务

- 传统IT企业
- 传统企业
- 咨询公司