



2017云栖大会·成都峰会  
THE COMPUTING CONFERENCE



# 阿里云机器学习技术与应用

主讲人：阿里云高级专家 刘吉哲



2017云栖大会·成都峰会  
THE COMPUTING CONFERENCE

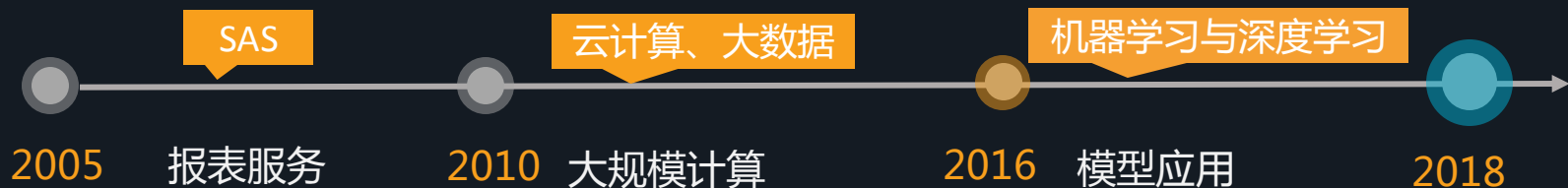
阿里云

云栖社区  
yq.aliyun.com





## 由BI转向机器学习



- 数据化运营需求公司已达**200**万家（来自国家统计局）
- 约**20**万家企业意识到机器的价值
- 2016-2018约 **60%** 大中型公司将机器学习用于商业分析



2017云栖大会·成都峰会  
THE COMPUTING CONFERENCE

阿里云

云栖社区  
yq.aliyun.com

技术门槛高

数据量大

硬件成本高

机器学习开发面临的挑战





2017云栖大会·成都峰会  
THE COMPUTING CONFERENCE



高性能云端计算  
降低存储和计算成本

阿里云机器学习  
PAI

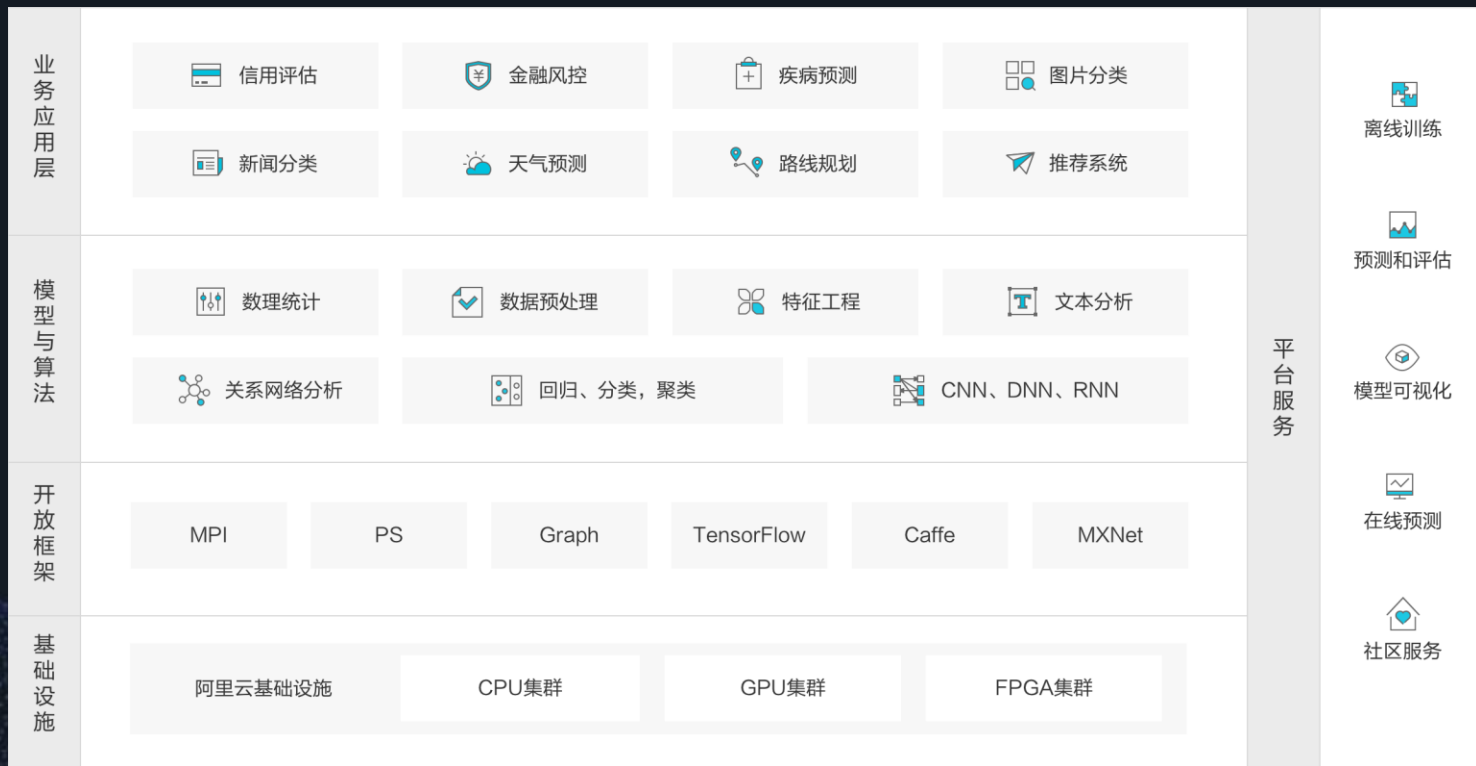
工具、算法库  
降低技术门槛

# 目录

## Contents

- 1 阿里云机器学习
- 2 关键技术
- 3 应用案例

# 阿里云机器学习 PAI 2.0





最新功能

- 深度学习DNN算法
- 云栖社区功能



新手引导

- PAI产品概述
- 新手入门



常见问题

- F&Q
- 客服在线



云栖社区

- 【玩转数据系列二】机器学习应用没那么难，这次教你玩心脏病预测

开发流程

算法组件介绍

常用算法推荐

数据

更多

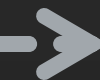
基础



新建空白实验

基础

农业贷款预测的回归算法...



更多

数据预处理

特征工程

机器学习模型训练

模型评估

学习

离线/在线服务

推荐

通过协同过滤算法实现商品推荐。

1255位用户



加载更多



## 阿里云机器学习 (PAI)

### 预处理工具组件

- 采样与过滤
- 数据合并
- 拆分
- 数据合并
- 格式转换
- 类型转换
- ...

### 特征工程

- 特征重要性评估
- 特征变换
- 特征选择
- 特征生成

### 统计分析

- 假设检验
- 协方差
- 直方图
- 散点图
- 概率密度图
- ...

### 常用机器学习算法

- 二分类
- 多分类
- 聚类
- 回归
- 评估
- 预测

### 垂直应用领域

- 文本分析
- 搜索推荐
- 图像处理
- 网络分析
- 金融板块

### 深度学习

- TensorFlow
- Caffe
- MXNet
- 自定义算法



# 算法脱胎于内部业务

- 01 MPI 算法和SDK  
定向广告，微贷风控
- 02 PS 算法和SDK  
淘宝主搜推荐系统
- 03 深度学习算法  
芝麻信用、蚂蚁证件审核



# 深度学习框架

1 TensorFlow  
支持版本 1.0

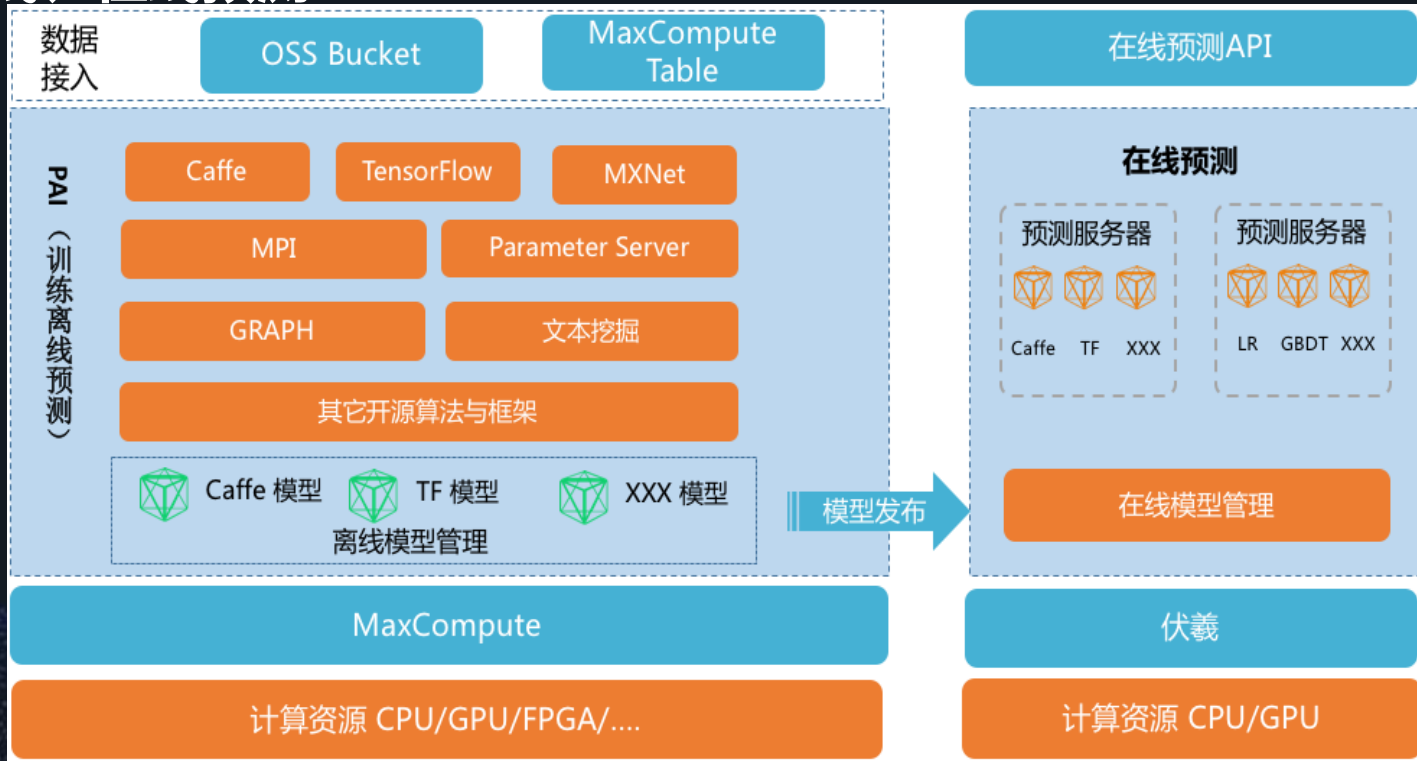
2 Caffe  
支持版本 RC3

3 MXNet  
支持版本 0.9.5





# 模型离线、在线预测



# 目录

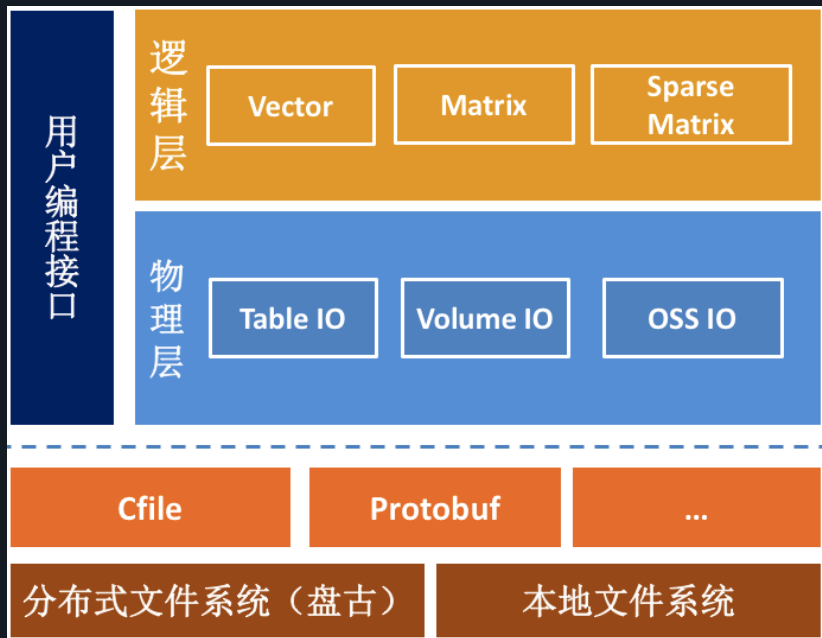
## Contents

- 1 阿里云机器学习
- 2 关键技术
- 3 应用案例



## 支持结构化、非结构化数据

- 结构化
  - MaxCompute 表
    - 数据是无序的
    - 特征表达有限
  - KV表
  - Map表
    - SQL select支持
- 非结构化
  - 图像、语音、视频
  - 支持向结构化数据转换



# MPI 算法开发

- 开发

- 平台与算法解耦
- MPI 计算框架与算法分离

- 调试

- 本地开发、调试
- 线下、线上无缝集成

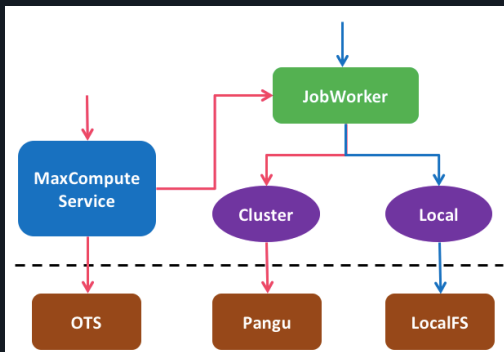
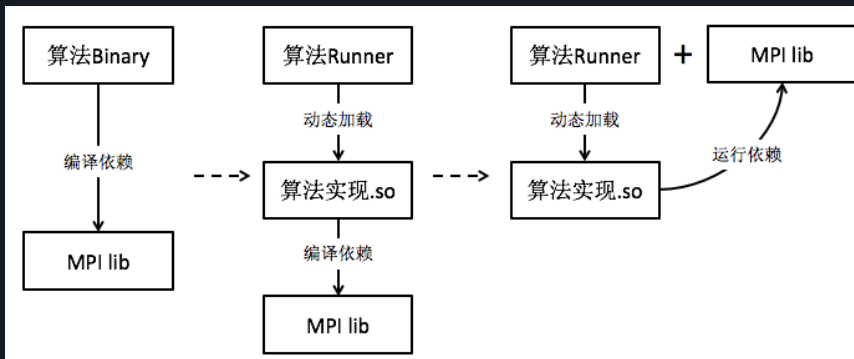
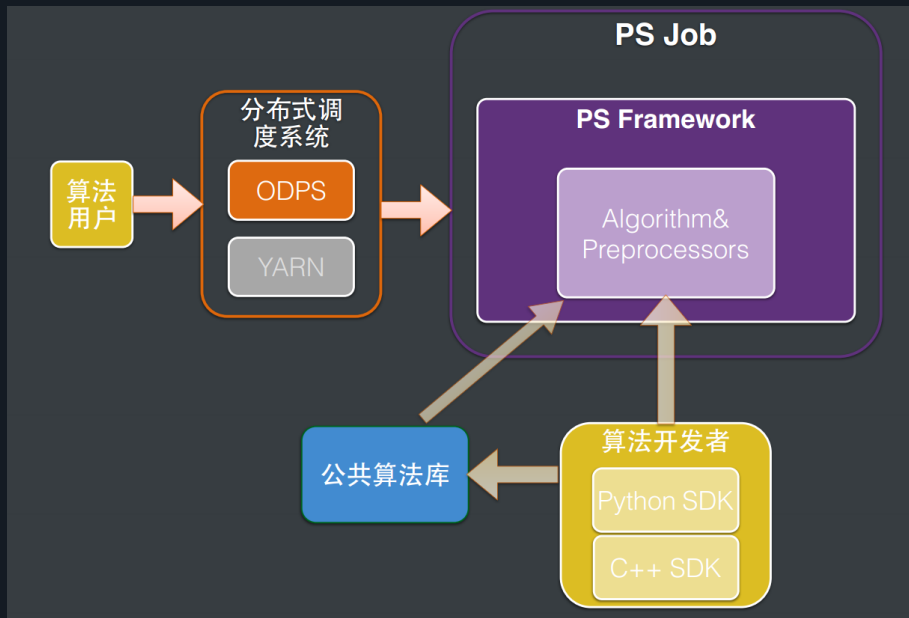


图 — Cluster  
示 — Local



## 参数服务器/PS

- 参考实现
  - Google DistBelief
  - Limu Parameter Server 架构
- 广告推荐
  - 100T, 260亿特征
  - 8个小时收敛
- PS-SDK接口
  - Python/C++混合编程
  - 提供本地调试



# 文本分析

## 文本分析算法 ( API )

- 词频统计
- TF-IDF
- PLDA
- Word2Vector
- Doc2Vector
- 关键字抽取
- 文章相似度
- 文本摘要
- 条件随机场预测

正文

我来说两句(1人参与)

扫描到手机

来源: 央视网 | 2016-10-12 09:36:28 |

应柬埔寨王国国王西哈莫尼邀请, 国家主席习近平将于10月13日至14日对柬埔寨进行国事访问。

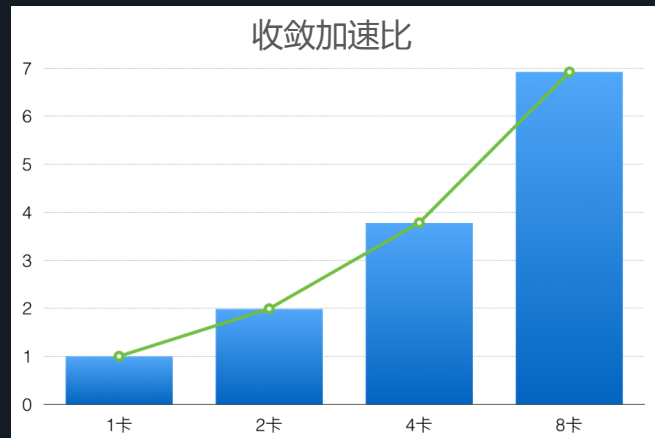
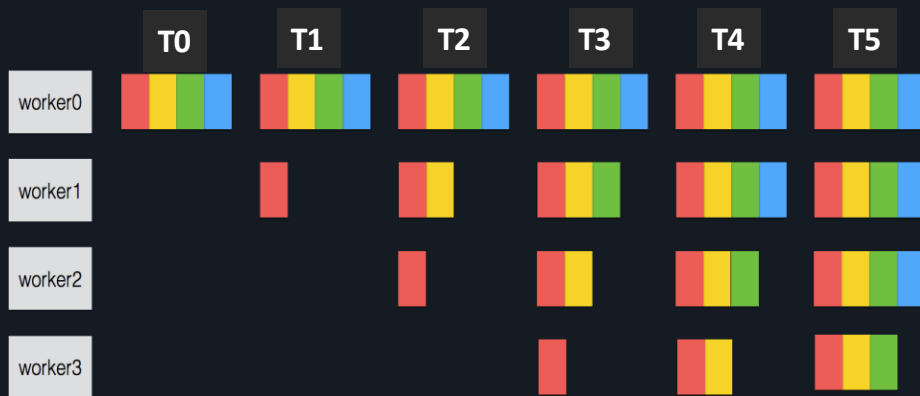
柬方对于此次访问非常重视, 为迎接即将到访的中国国家主席习近平, 柬埔寨的主要道路都挂上了中国国旗, 中柬标志性的建筑物天安门和吴哥窟也出现在市中心的广场上。

文本摘要

习近平 将 对 柬埔寨 进行 国事 访问



# Caffe 多机多卡



AlexNet On ImageNet Data

- 数据分片, 1-N 变为 P2P
- AlexNet 8卡上得到将近7倍性能提升





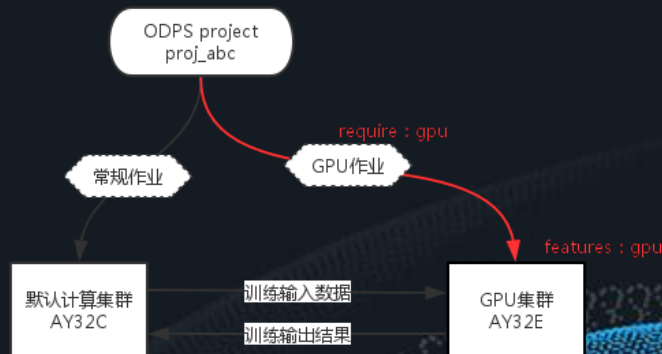
# 异构(CPU/GPU)集群调度

Job DAG 中允许不同Action类型节点



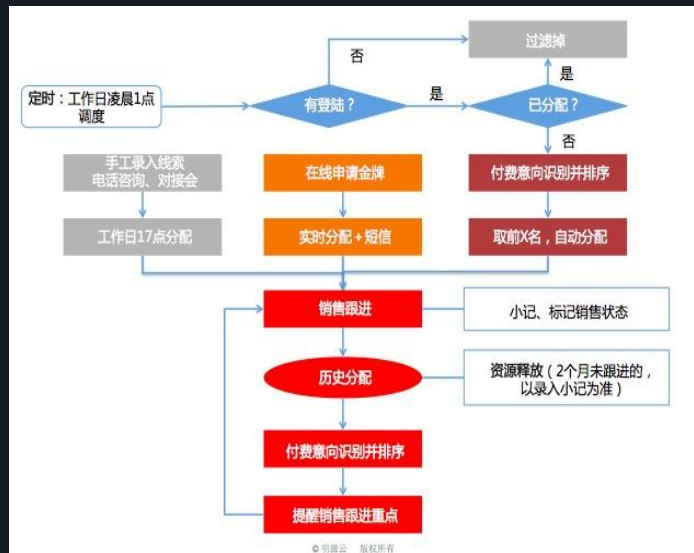
扩展Job XML, 基于 tag 确定运行集群

```
<resource>
  <instance_count>3</instance_count>
  <cpu>100</cpu>
  <mem>4000</mem>
  <virtual_resources>
    <r name="GPU" value="100" />
  </virtual_resources>
</resource>
```



# 目录 Contents

- 1 阿里云机器学习
- 2 关键技术
- 3 应用案例



- 在业务流程中融入机器学习
- 付费客户转化比由10.7% 提升到 **19.6%**



2017云栖大会·成都峰会  
THE COMPUTING CONFERENCE



阿里云

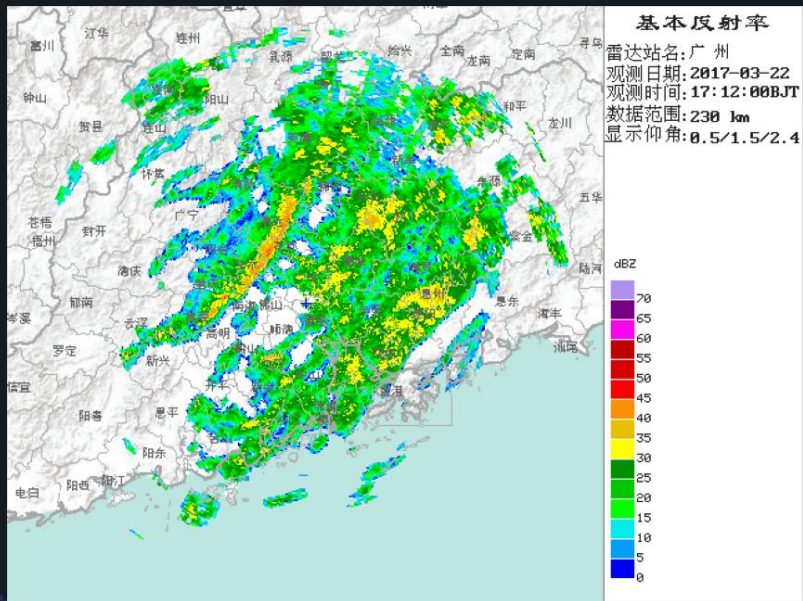
# 广告无打扰植入



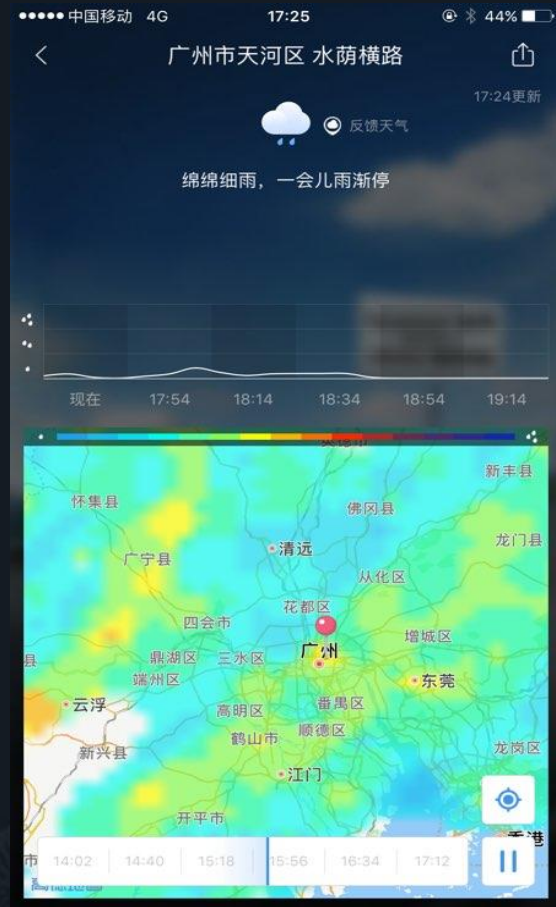
- 自动植入式广告，不影响观看体验，不停顿、无打扰
- 可根据视频场景和观众喜好植入定制化广告内容



# 墨迹天气



- 根据雷达回波图样本训练生成云图模型
- 通过在线预测实现区域短时天气预报



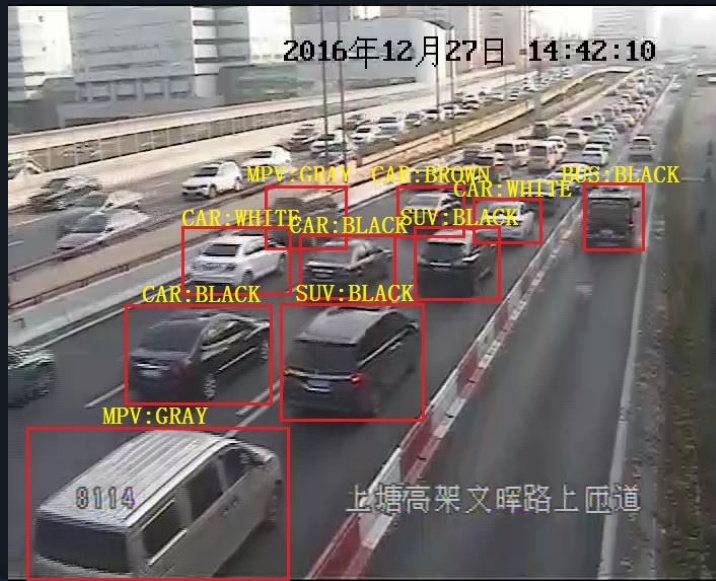




2017云栖大会·成都峰会  
THE COMPUTING CONFERENCE



# 杭州城市大脑



- 根据车型和品牌，计算每辆车所在位置和物理距离，计算通行速度
- 根据视频数据，进行异常检测，实时预警，提升事件处理效率



2017云栖大会·成都峰会  
THE COMPUTING CONFERENCE



# 天池算法大赛(Power By PAI)



2016

- 菜鸟分仓规划
- 阿里音乐流行趋势预测
- 最后一公里极速配送
- 阿里云安全算法挑战赛
- 机场客流量时空分布预测

- 阿里移动推荐算法
- 资金流入流出预测
- 新浪微博互动预测
- 淘宝穿衣搭配算法
- 公交线路客流预测

2015

- IJCAI-17 口碑商家客流量预测
- KDD CUP 2017
- 阿里聚安全算法挑战赛
- 更多赛事.....

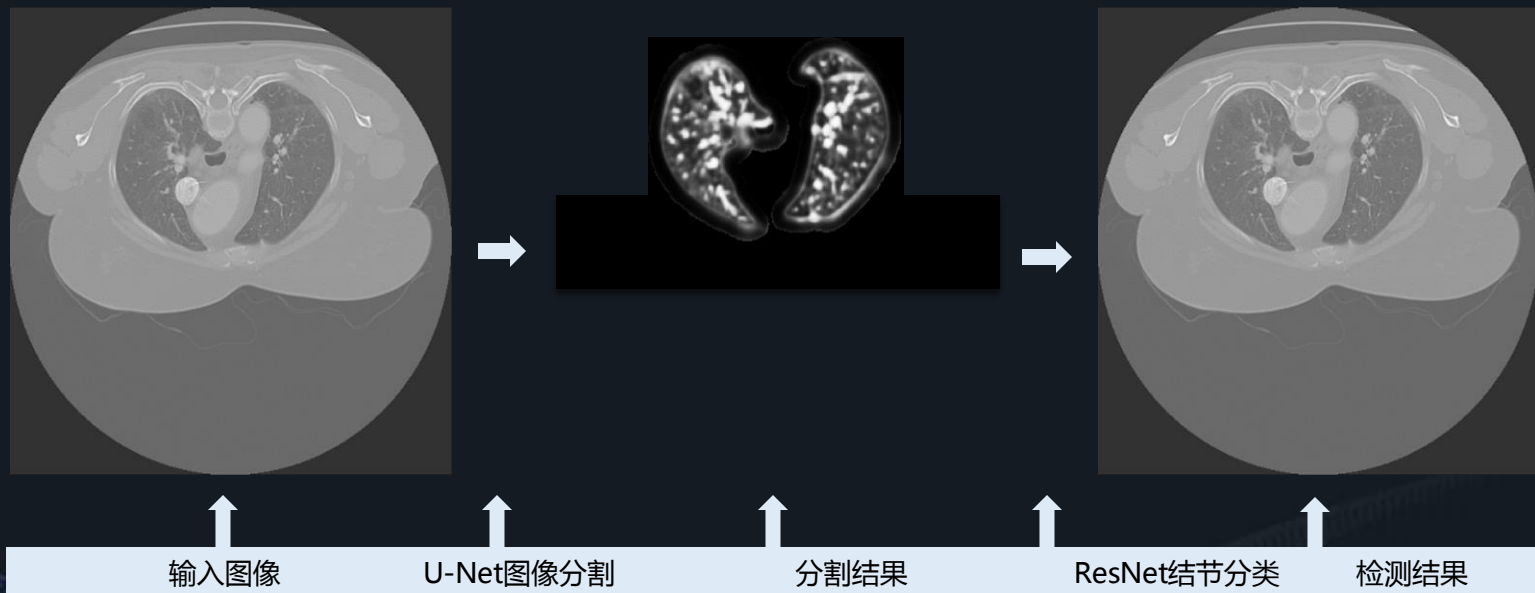
2017



2017云栖大会·成都峰会  
THE COMPUTING CONFERENCE



# 医学影像分析



- 阿里云与 Intel 合作，7月份在 PAI 产品上开展天池医学影像分析算法大赛

# 总结

- 机器学习与深度学习相结合的平台
- 强大的计算资源
- 一站式人工智能机器学习平台



2017云栖大会·成都峰会  
THE COMPUTING CONFERENCE



# 更多学习资料与支持请扫二维码进入



阿里云机器学习 PAI 官网



数加·阿里云机器学习钉钉群





2017云栖大会·成都峰会  
THE COMPUTING CONFERENCE

阿里云

云栖社区  
yq.aliyun.com

# 飞天·智能

## APSARA INTELLIGENCE

2017云栖大会·成都峰会

5月23日 成都世纪城天堂洲际大酒店