面向未来的泛内容AI平台建设实践

阿里巴巴 总监 蔡龙军

SACC2017

Agenda

- 文娱泛内容AI平台背景
  - 互联网发展的背景，对内容产业带来的冲击和生机
  - 建立娱大脑必要性
- 文娱泛内容AI平台整体规划
  - 三维立体分析平台规划
  - 基础设施深度学习平台DeepDriver介绍
- 文娱泛内容AI平台投资采买分析能力建设
- 文娱泛内容AI平台营销分析能力建设

SACC2017

内容的时代：这世界很酷

《军师》~70亿，男性高知群体的最爱

Day & night 追凶 25亿

虐恋总是要100~200亿虐！！

爱情春风引爆暑期56亿

战狼2 票房 57亿！

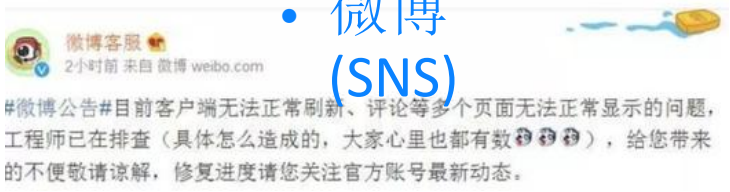
铁打的四大神兽，流水的明星

这些内容去哪里看呢？YouKu

这种变化影响是什么？

鹿晗恋爱了....，是件互联网大事

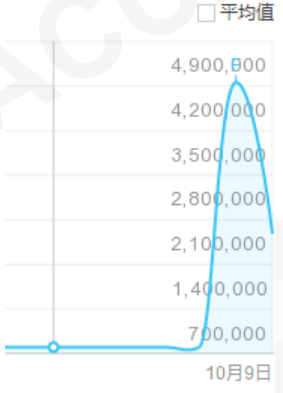
- 微博(SNS)



- 优酷(视频)

- 淘宝(电商)

- 百度(搜索)



- 王者荣耀(游戏)

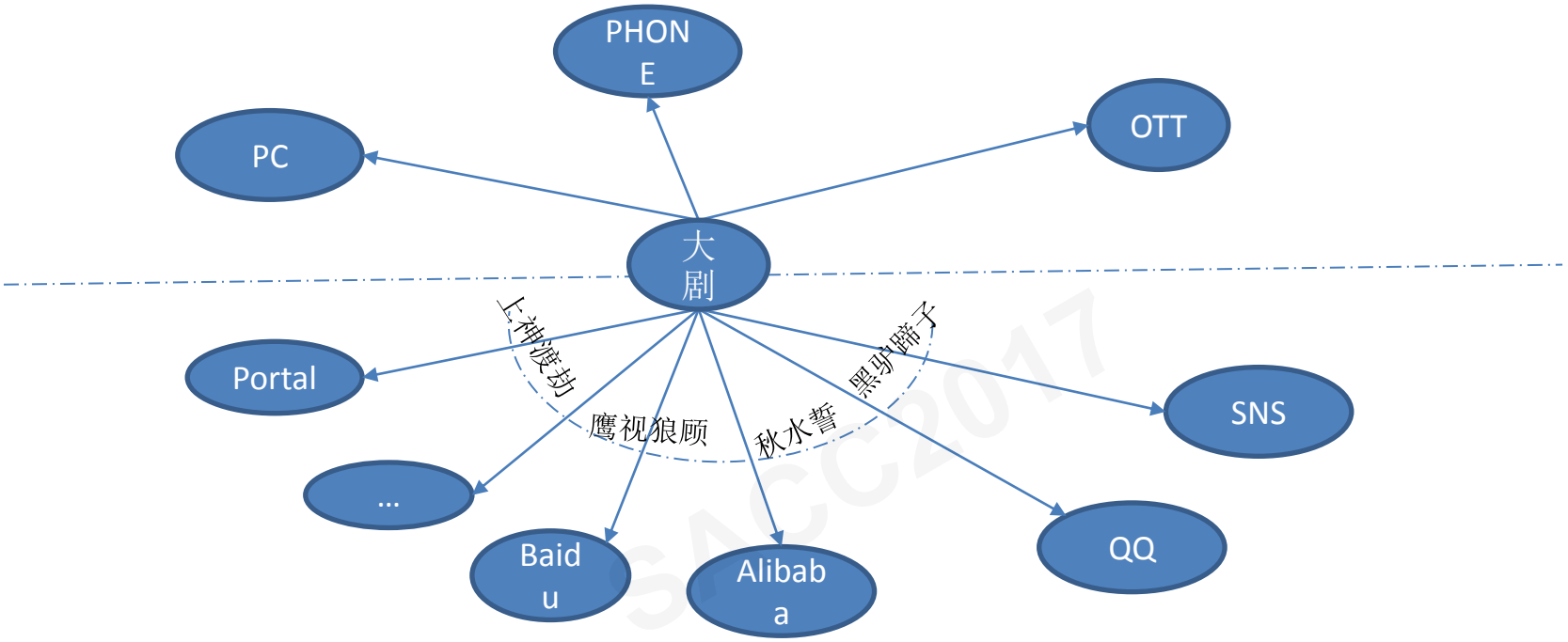
暴打“关羽”，“姓关的就没有好东西”

- 婚纱摄影(传统)

“鹿晗就不是好东西，没拍完，我一个女客户穿婚纱跑了...”

SACC2017

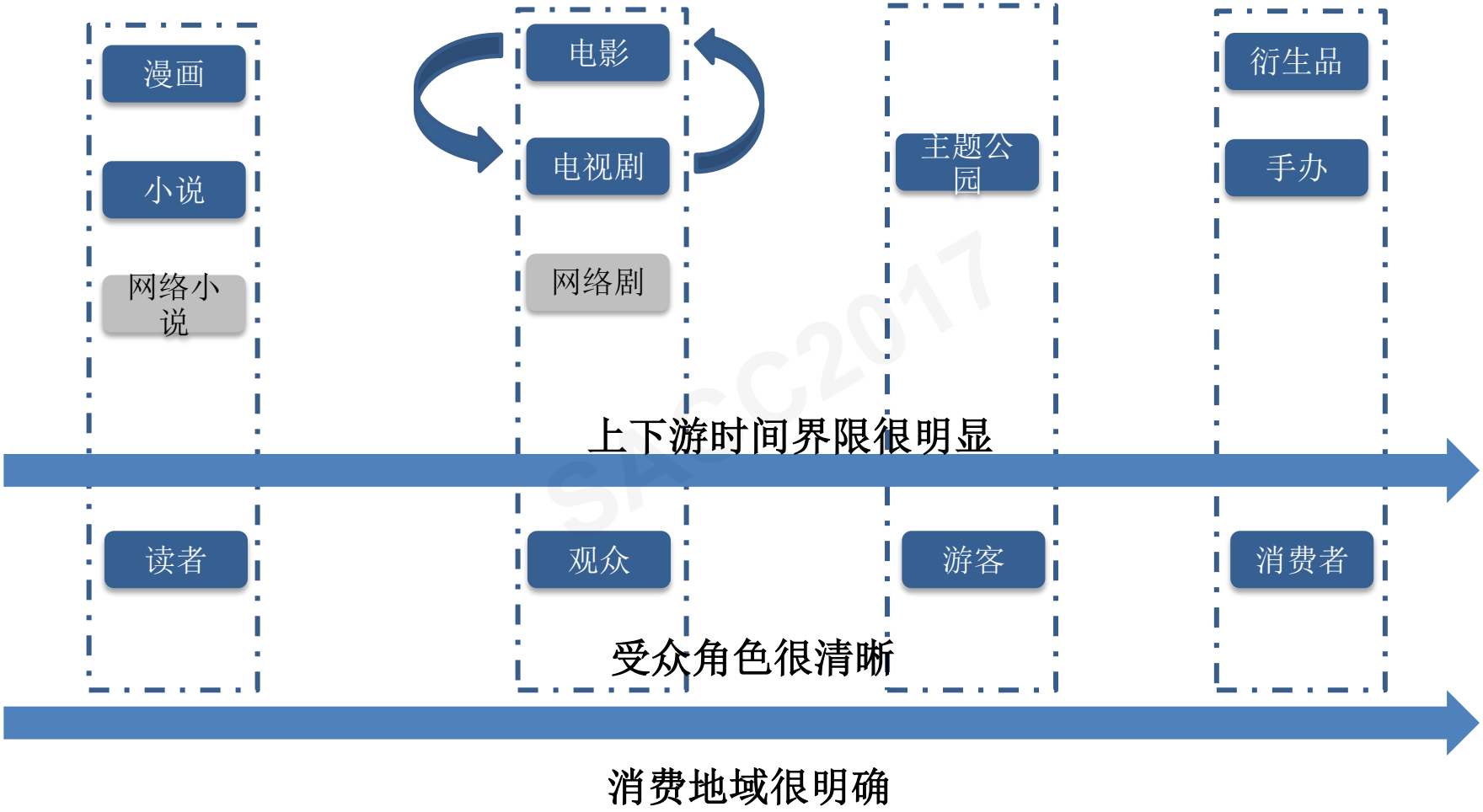
互联网娱乐化



- 那么把控文娱就要理解整个互联网和文娱相关的数据！

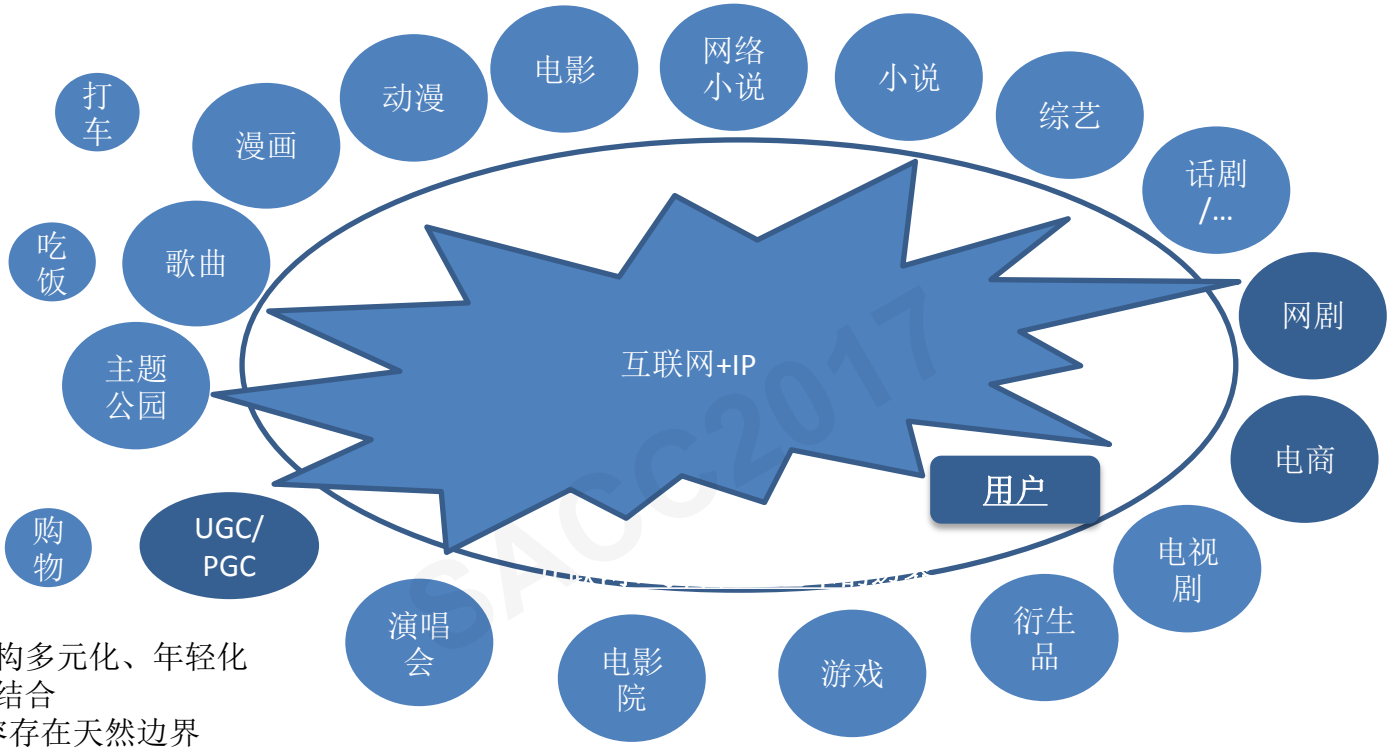
SACC2017

曾经的文化娱乐产业是条河



SACC2017

破坏原有土壤结构，也带来了新的形式



1.内容形式、结构多元化、年轻化
2.人才各种跨界结合
3.互联网+和内容存在天然边界
4.实时与预知

制作思路和内核都有了新的变化

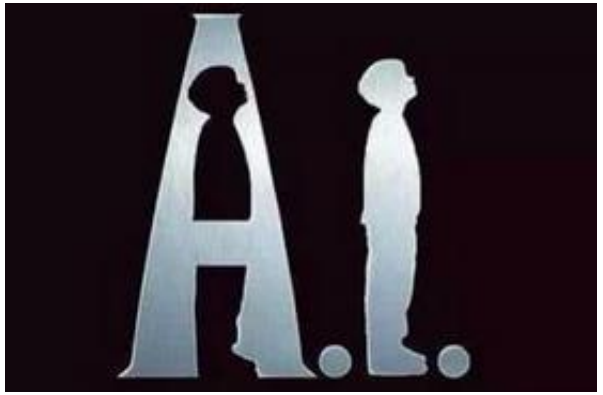
SACC2017

帮助人们理解内容：文娱泛内容AI平台

有生之年看不完的数据

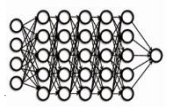
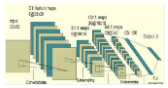
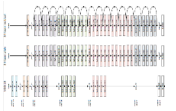
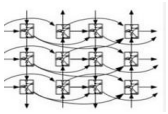
所以，只能...

7*24并行计算+AI



SACC2017

文娱泛内容AI平台：覆盖全生命周期

依托DeepDriver不断深入AI

传播生命周期、4P/4C、STEPPS经典理论在业务上应用

智能化需求挖掘

智能化运营

智能化制作

智能化营销

智能化投资

生命周期整合

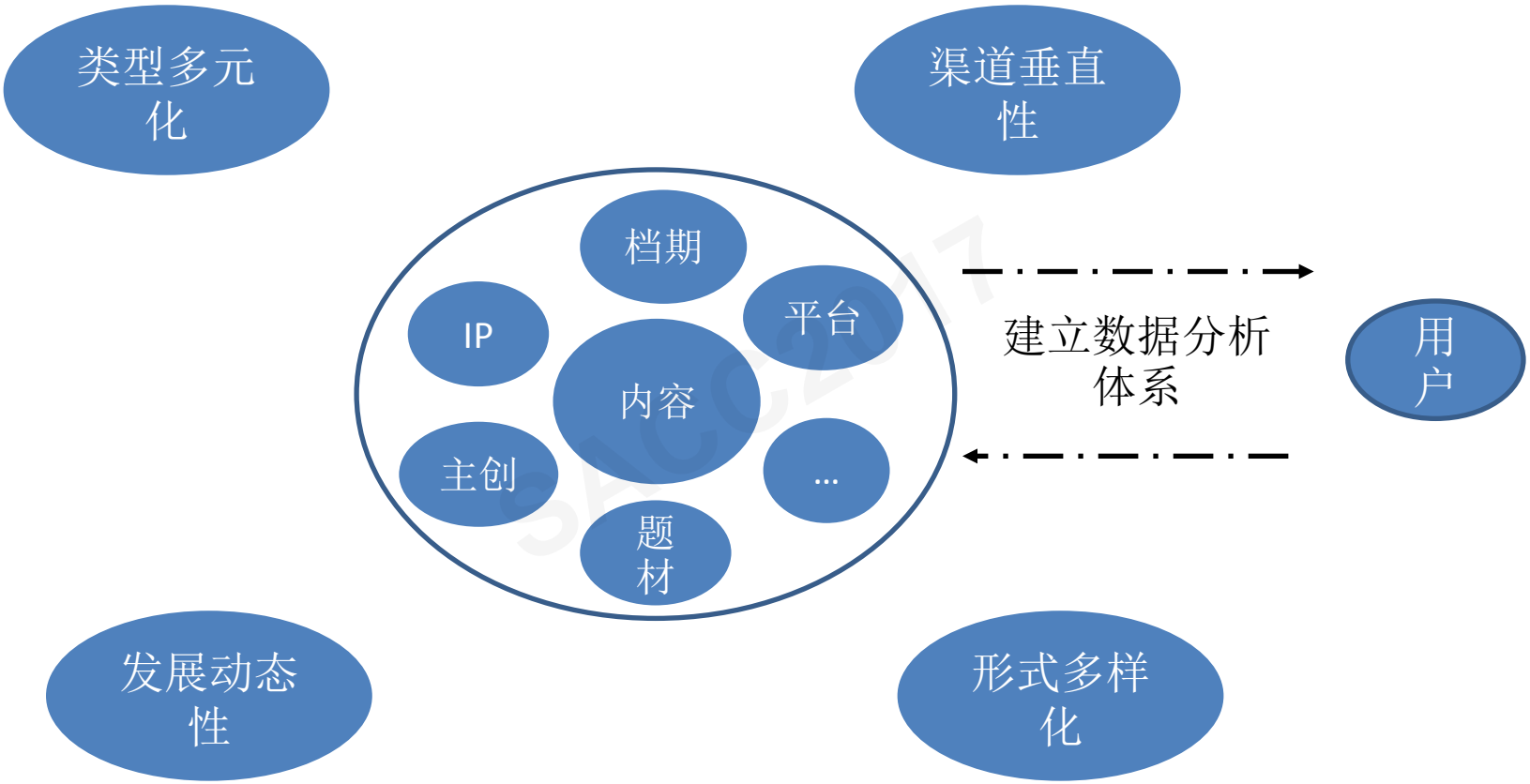
可传播的抽象或者具象对象实体

歌曲 ..... 出版物 话剧 小说 漫画 事件 游戏 资讯 地点 新闻 音乐 电视剧 综艺 体育 电影 动漫

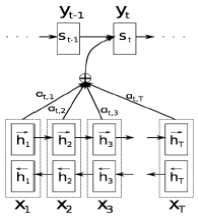
贯通多文本、全生命周期上下游数据网络体系

SACC2017

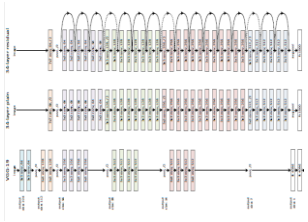
数据体系建设：从上帝角度进行思考



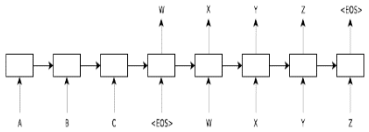
自研的深度学习平台DeepDriver



AttentionModel

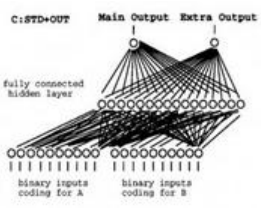


VGG&ResNet



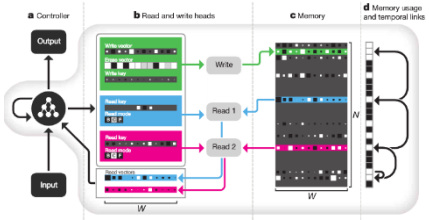
Encoder-Decoder

复杂DL层

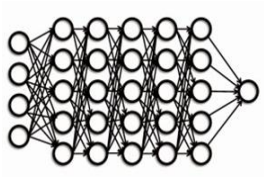


Seq2Seq

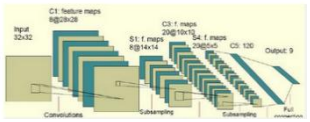
MTL



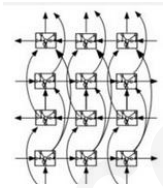
DNC



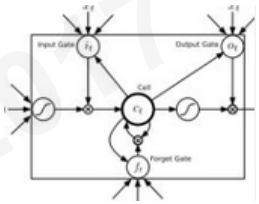
DNN



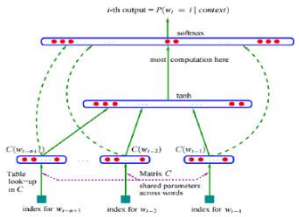
CNN



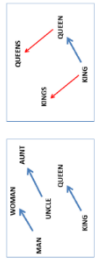
基础DL层



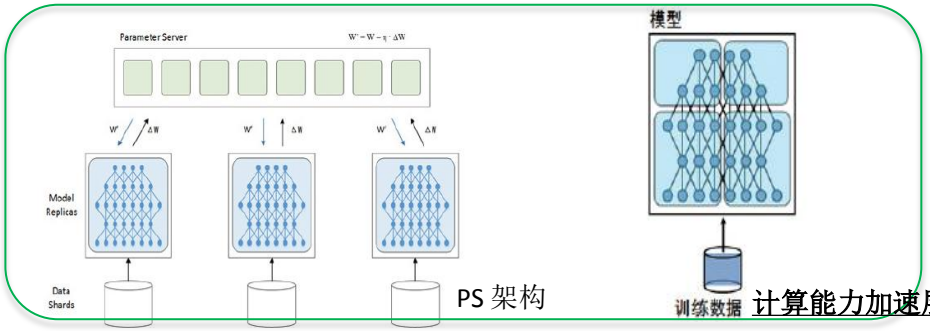
RNN/LSTM



NN-LM



W2V



PS 架构

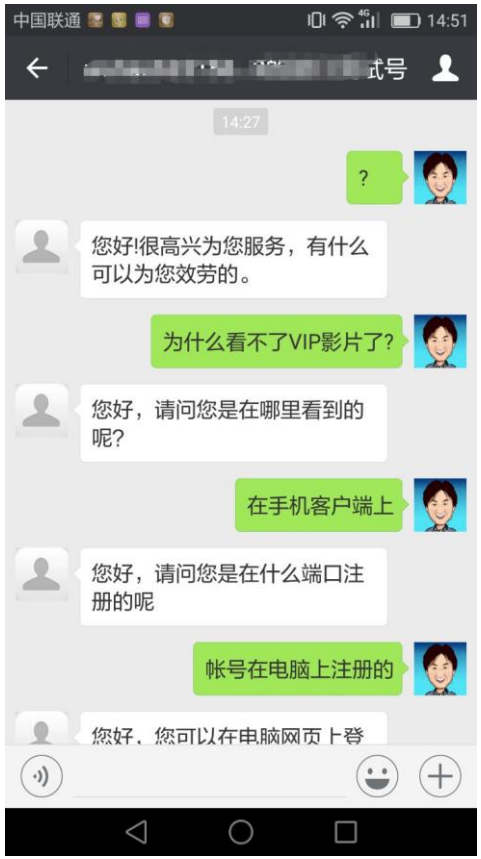
计算能力加速层

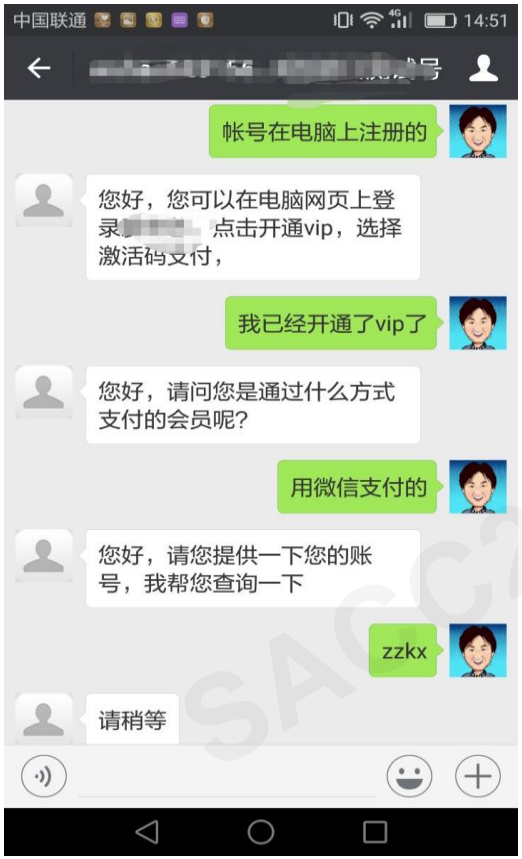
FPGA

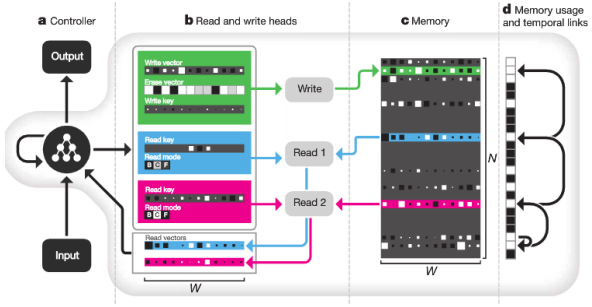
芯片级

SACC2017

DeepDriver应用的例子



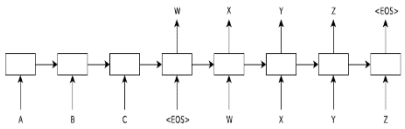


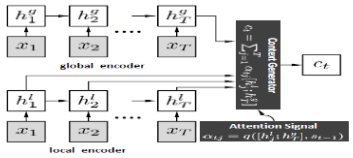


- 例如给个短文：
- “小明带个球去操场，小明抱着球回到教室，小明抱着作业去老师办公室”
- 测试：

“小明在哪里”，答案是“办公室”

“球在哪里”，答案是“教室”





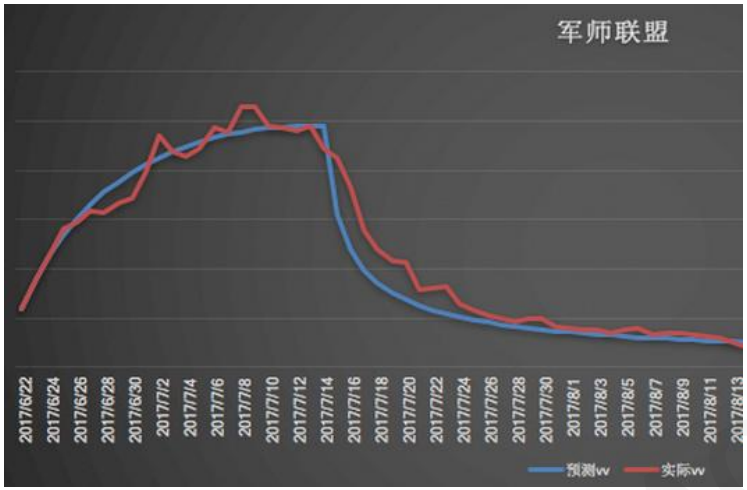
- 纯JAVA开发，轻量级(几万行代码，极少第三方依赖包)
- 着眼于先进的人工智能算法实践
- 开源平台https://github.com/LongJunCai/DeepDriver

SACC2017 云智未来 9th
IT168.com ChinaUnix ITPUB

文娱泛内容AI平台投资采买分析能力建设

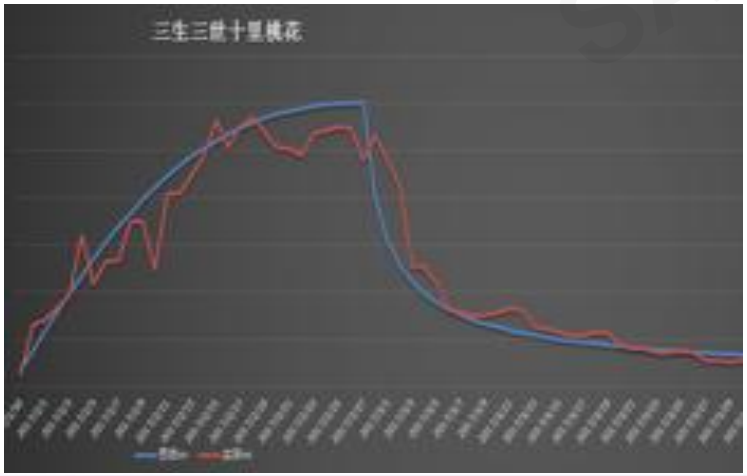
- 预测准确率：提前**1年预测：80%+；**

准确率92%

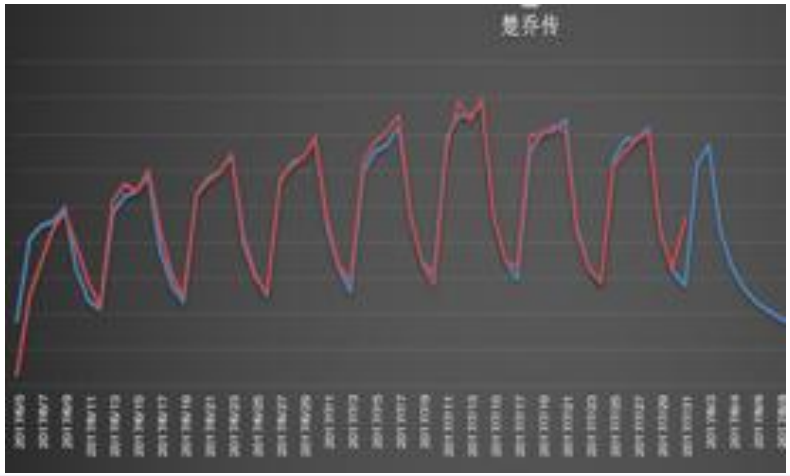


准确率97%+

准确率90%



准确率92%

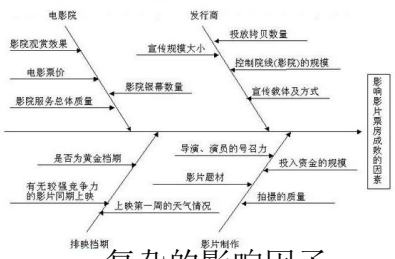


黑盒学习存在的困难与挑战

- 常规的预测思路：数据+简单模型、数据+逼近能力强黑盒模型
- 常规思路本质：黑盒模型逼近真相
- 存在问题：
  - 复杂机制很难通过样本进行覆盖
  - 很难深入理解问题本质
  - 很难跨领域进行举一反三学习

具有生命周期

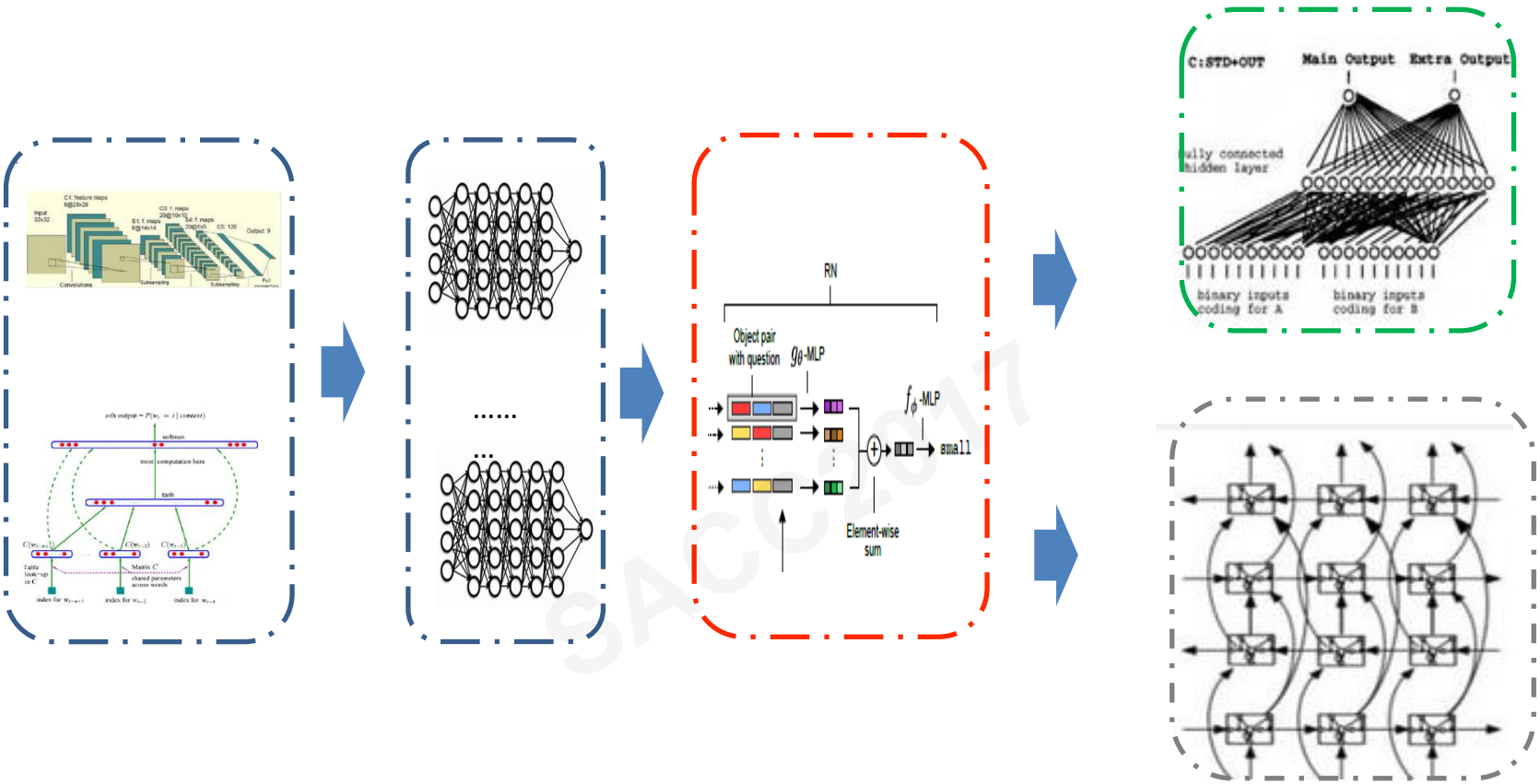
竞争博弈



复杂的影响因子

SACC2017

基于机制建模构造复变预测模型

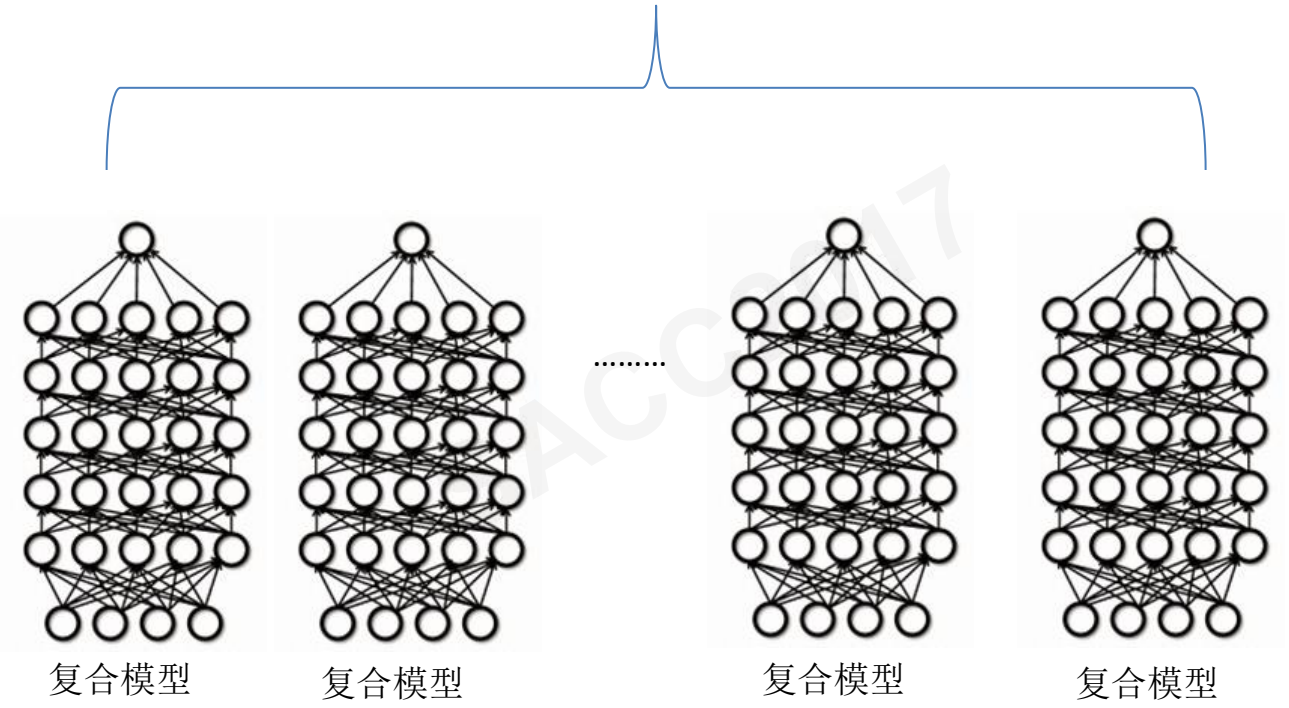


- 建立了一个端到端的混合预测模型：Embedding CNN+DNN+RN+MTL+LSTM

SACC2017

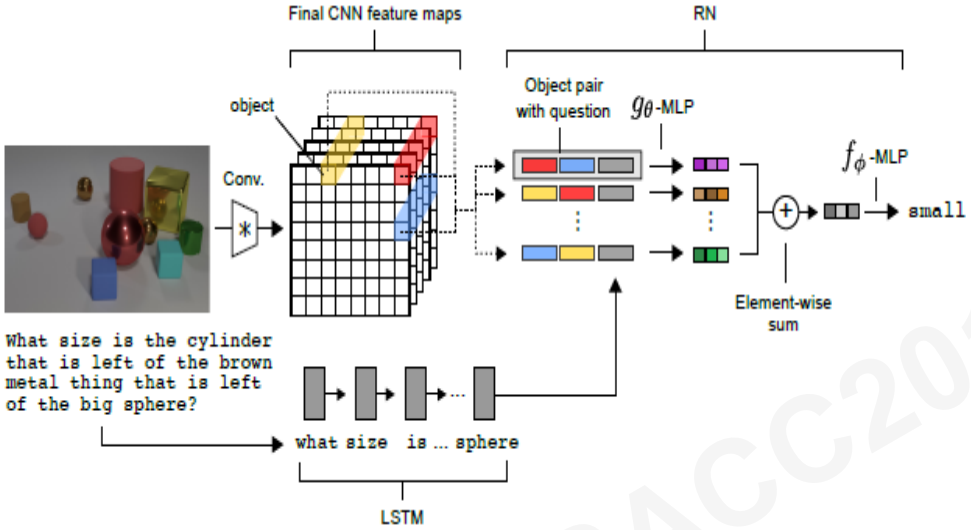
最终模型：组合多个模型结果

Thousands of DL make dynamic decisions



SACC 2017 云智未来 9th
IT168.com ChinaUnix ITPUB

Relation Net



$$RN(x_1, x_2, \ldots, x_n) = f_\varphi\left[\sum_{i=1}^{n-1} g_\theta(x_i, x_n)\right]$$

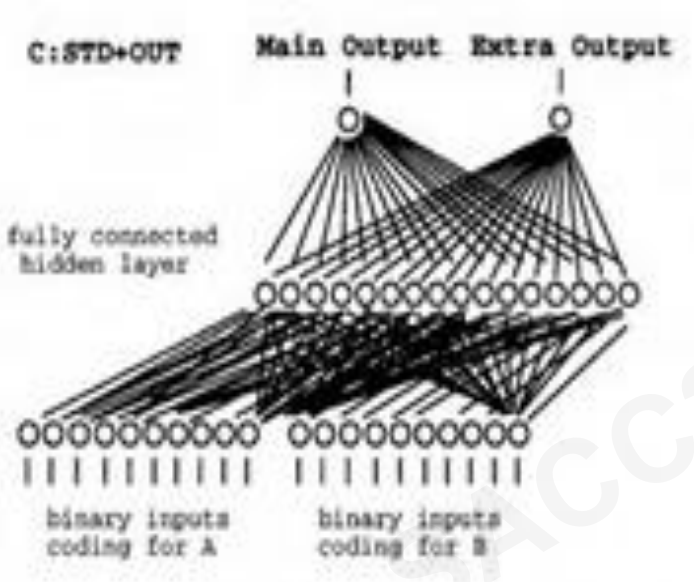
f_φ 和 g_θ是多层感知器 MLP

- 准确率有超过5%的提升

Adam Santoro, *A simple neural network module for relational reasoning*

SACC2017

MTL—Multiple Task Learning



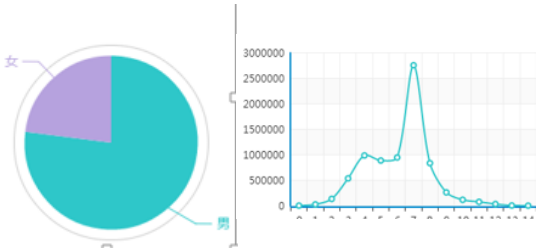
（1）隐式数据增加机制
（2）注意力集中机制
（3）窃听机制
（4）表示偏置机制
（5）正则化机制

- 泛化能力提升接近20%提升

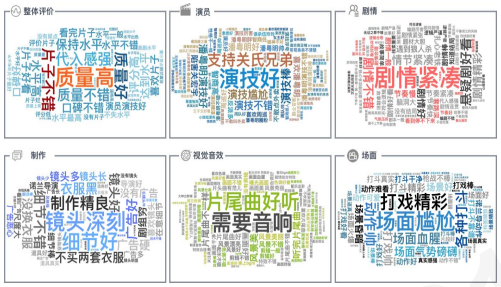
SACC2017

营销推广分析

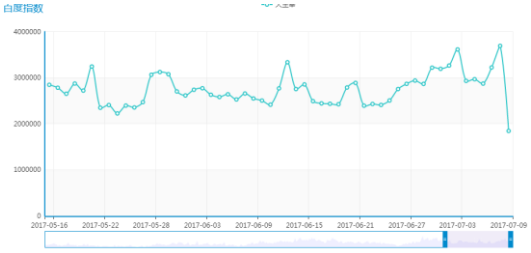
受众分析



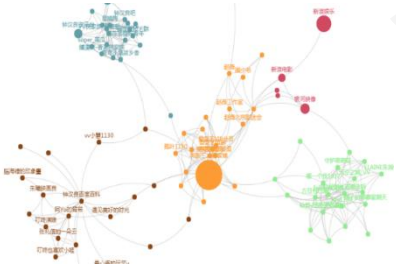
主题分析



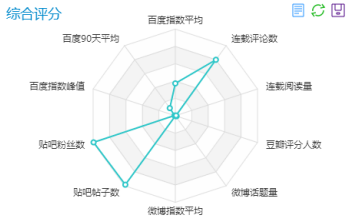
热度分析



社区分析



渠道分析



评分维度	该项最大数值	该IP该项数值	该项排名
百度指数平均	707420	264197	8
百度90天平均	1257186	121458	36
百度指数峰值	1266622154	3040703	16
贴吧粉丝数	5650486	5650486	1
贴吧帖子数	140744323	140744323	1
微博指数平均	5260694	2378	4532
微博话题量	77534000	83000	1025
豆瓣评分人数	289041	4551	560
连载阅读量	3653388141	14581100	365
连载评论数	1627814	1308432	4

SACC2017

营销推广分析-白夜追点啥？

- 整体集中在演员、整体评价和剧情上

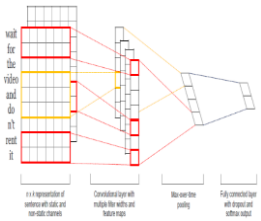
- 为潘粤明打Call，剧情精彩紧凑，场面镜头血腥恐怖，道具逼真，片尾曲好听（!!）

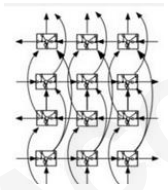
NLP模型

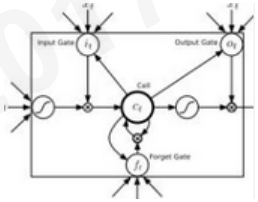
应用模型

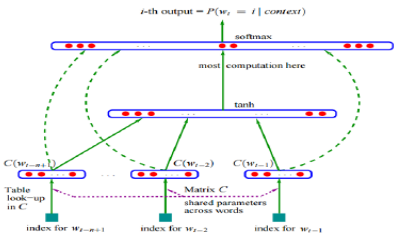
情感分类 | 意见提取与归类 | .... | 去噪音

通用模型









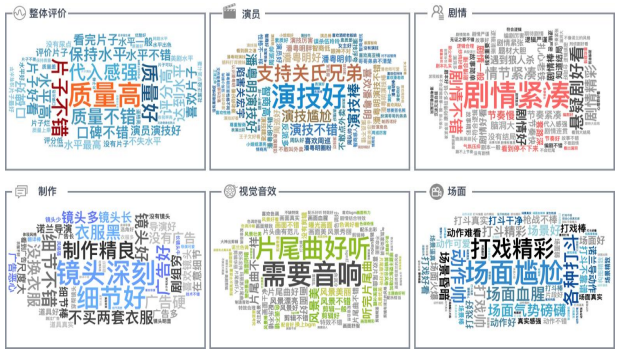
基础模型

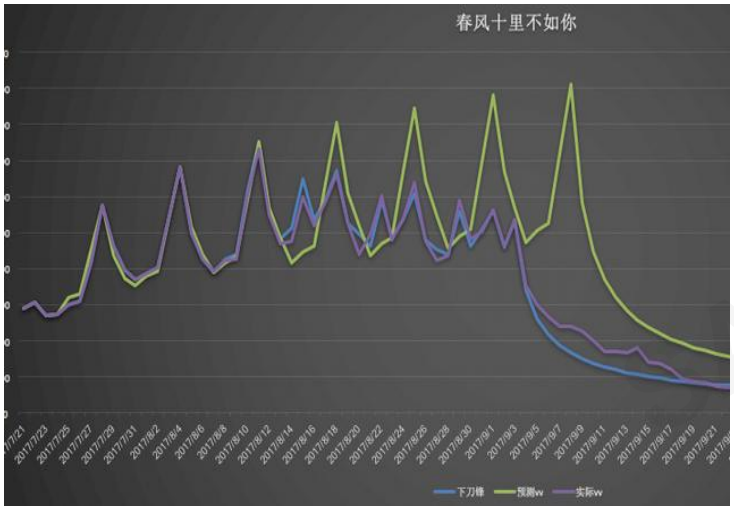
分词 | 命名实体识别 | .... | 语法分析

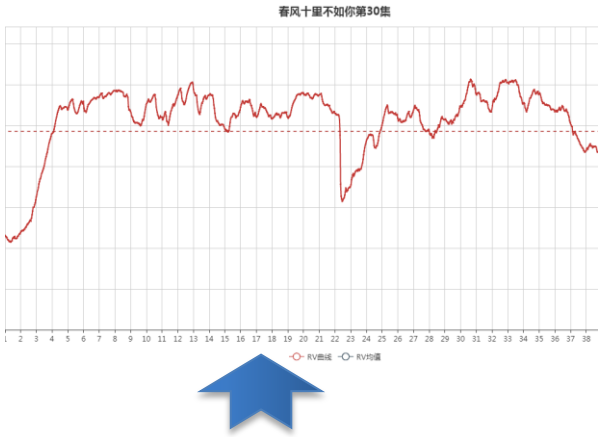
SACC2017

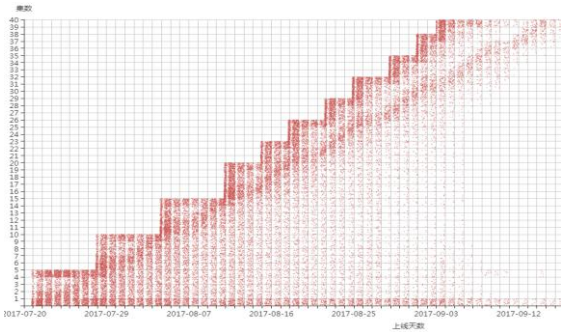
建立运营分析闭环

运营决策









SACC2017

总结

- 内容产业发展特点
  - 互联网连接人与高效服务，加速了文娱内容向互联网各种应用进行渗透
  - 移动互联网对内容产业的冲击改变了不同内容形式发展规律
- 建立内容三维立体AI分析平台可以有效帮助内容产业适应新发展趋势
- 针对发展呈现机理进行更深入的建模将成为一种趋势
- 随着人工智能技术的不断发展，模型的表达能力将得到进一步加强，也会推动行业奔向新的高度!

SACC2017

欢迎加微信！


THANKS