机器学习在蘑菇街风控业务中的应用

王帅（花名莲华）

蘑菇街技术专家

2017.07.22

个人简介

- 1、2013-2014，航天五院，图像数据挖掘算法策略；
- 2、2014-2015，百度，搜索相关性算法策略；
- 3、2015年至今，蘑菇街，风控算法策略。

提纲

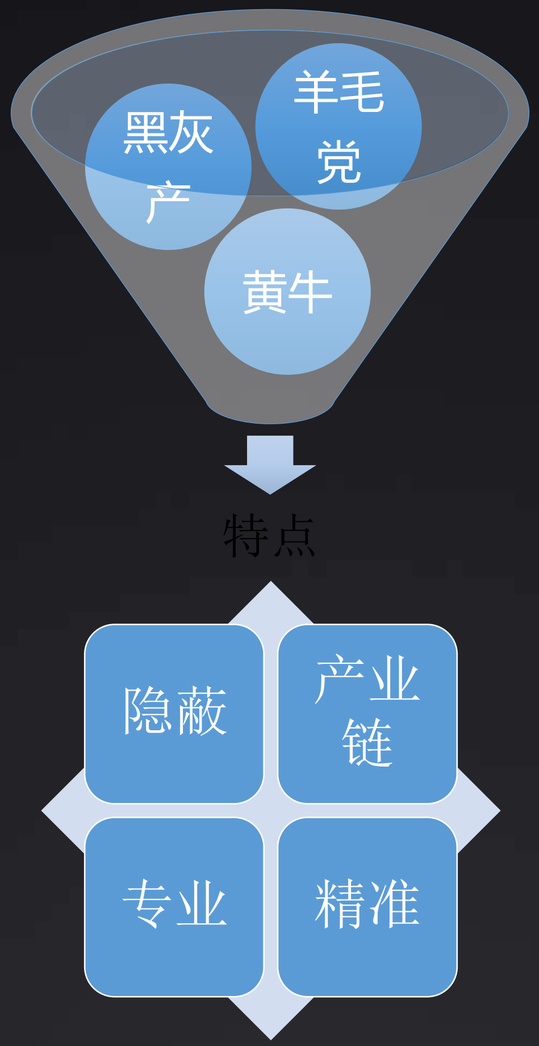
01 业务介绍

02 机器学习初探

03 新技术探索

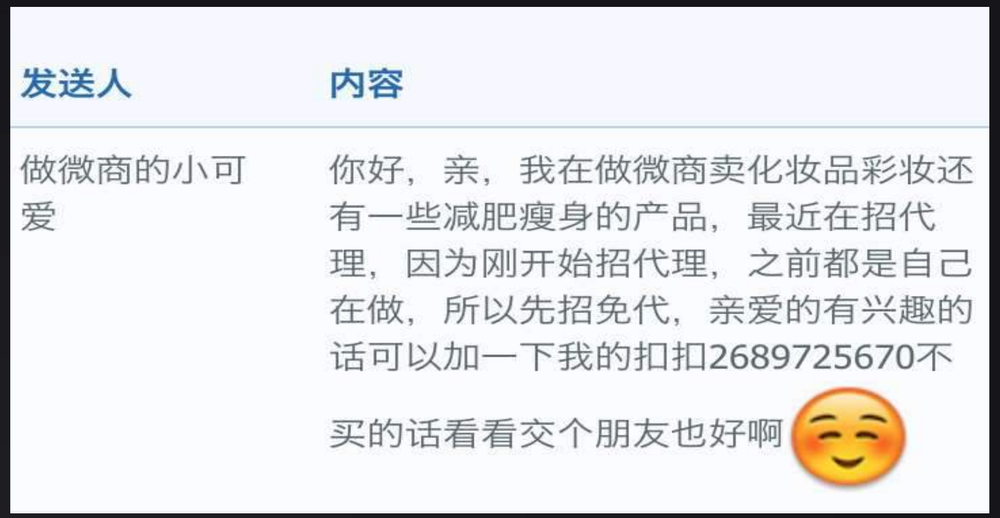
04 未来展望

一、蘑菇街风控业务



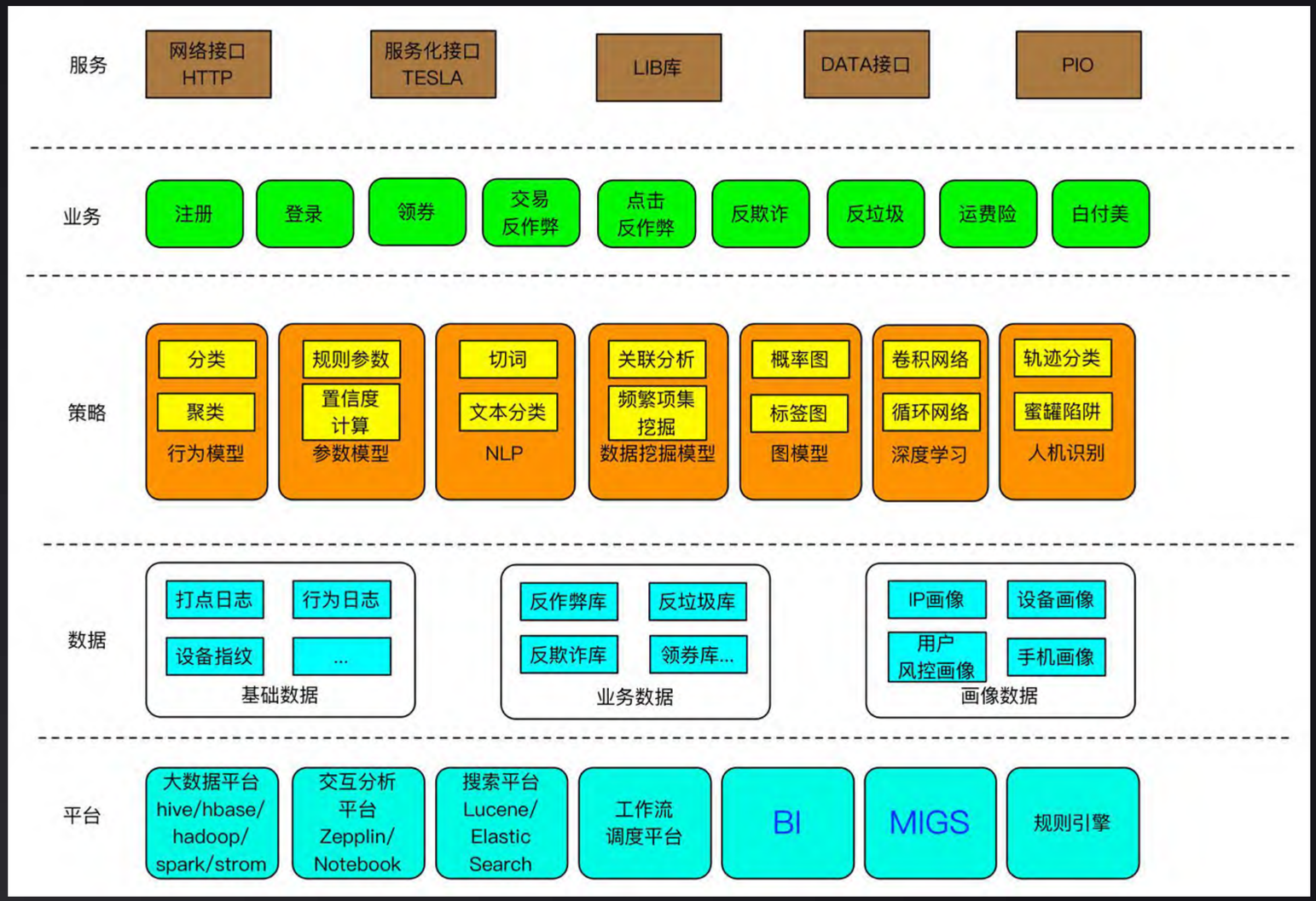
一、蘑菇街风控业务

- 1、账户风控
- 2、交易风控
- 3、内容风控
- 4、信贷风控



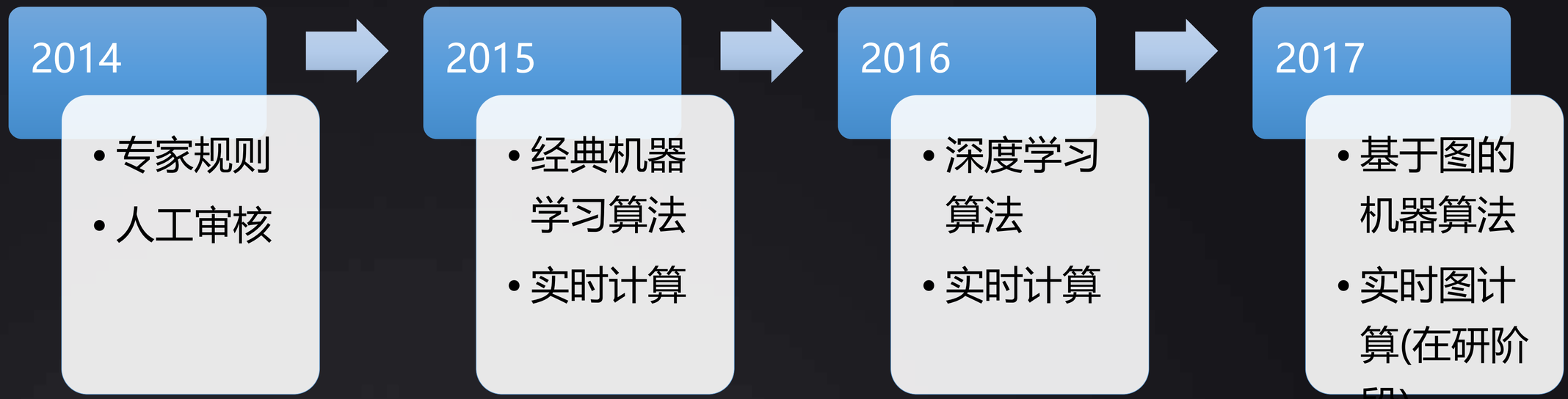
一、蘑菇街风控业务

- 蘑菇街风控策略架构
  - 平台
  - 数据
  - 策略
  - 业务
  - 服务



二、机器学习初探

- 风控策略演进时间线：



二、机器学习初探

- 传统方法：
  - 专家规则系统+人工审核

- **基于专家规则系统的优点**
  - **优点：**
  - 简单，直观，可控，解释性强，规则随时可以添加
  - 在特定领域的识别能有一定效果
  - 能够利用专家的经验常识
  - 快速响应业务变化，处理紧急情况

规则列表

ruleid	规则名称	部门	所属应用	关联事件	执行环境
293	UUidU[redacted]	detailcheat	WallDetailCheat	1	依赖系统自带的表达式定义来执行校验
294	detail_[redacted]y	detailcheat	WallDetailCheat	1	依赖系统自带的表达式定义来执行校验
295	user_t[redacted]	detailcheat	WallDetailCheat	1	依赖script脚本来执行
298	detail_[redacted]y	detailcheat	WallDetailCheat	1	依赖系统自带的表达式定义来执行校验
300	detail_[redacted]	detailcheat	WallDetailCheat	1	依赖script脚本来执行
476	Userid[redacted]	detailcheat	WallDetailCheat	1	依赖系统自带的表达式定义来执行校验
477	UserCli[redacted]	detailcheat	WallDetailCheat	1	依赖系统自带的表达式定义来执行校验

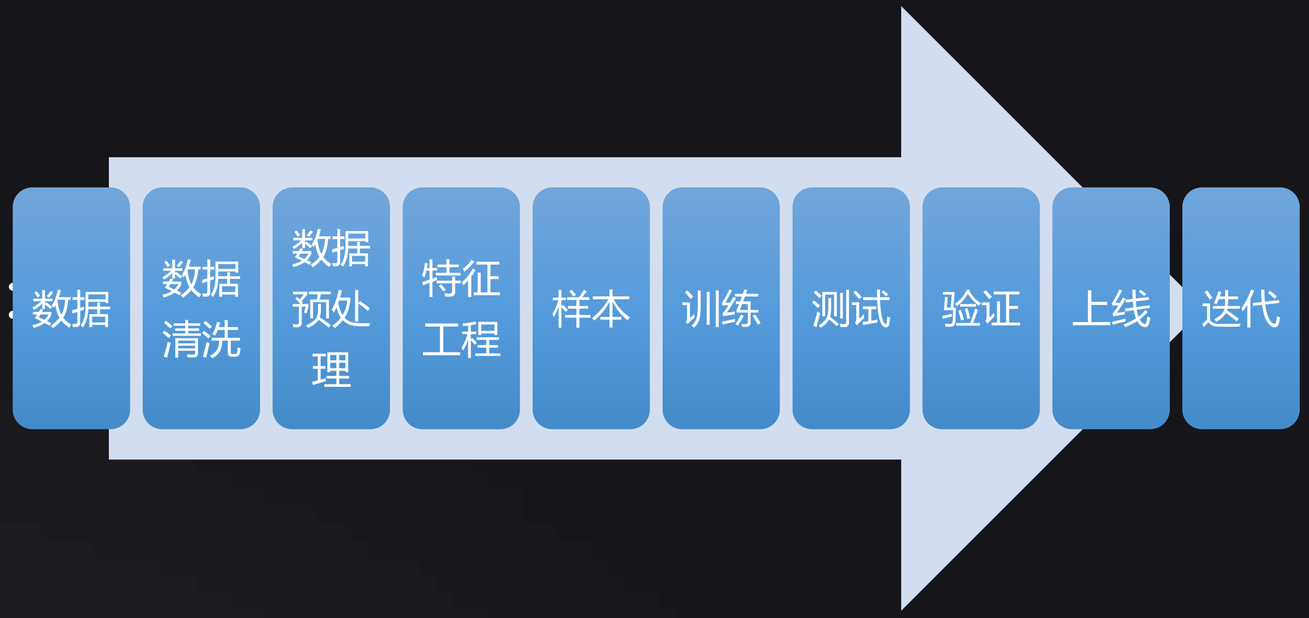
二、机器学习初探

- **基于专家规则系统的缺点**
  - 容易被黑灰产攻破，规则效果衰减较快，所以需要人工持续优化规则，耗费大量人力物力
  - 业务变得复杂时，规则因子增多，规则提炼变得困难
  - 泛化能力差：
    - 高度的领域相关性
    - 可移植性差

为了克服上述缺点，需要机器学习技术！

二、机器学习初探

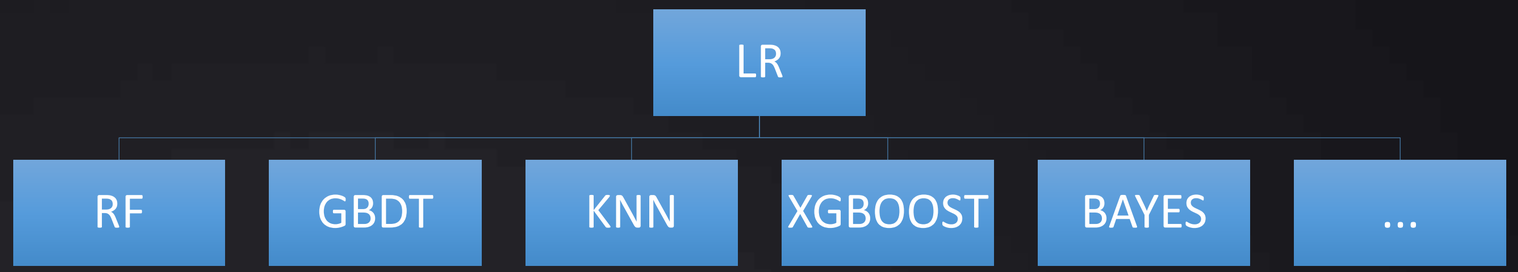
- 经典机器学习算法建模一般流程：



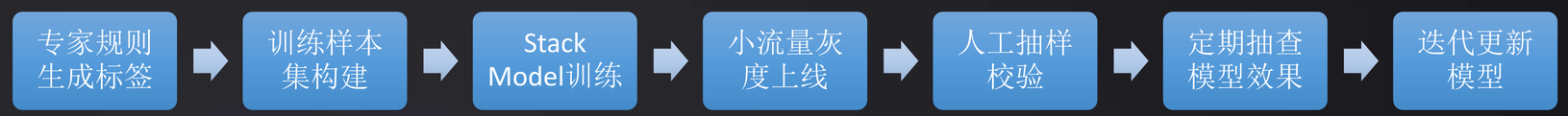
- A)特性工程（深入业务）：
- 特征主体：
  - 用户：以买家，卖家的维度来聚合特征
  - 设备（设备指纹）：以设备为维度聚合特征
  - IP：以IP为维度聚合特征
  - 手机：以手机号为维度聚合特征
- 特征维度：
  - 属性特征：如是否实名等等；
  - 基础特征：基本时间单位内的用户行为，如喜欢，搜索等等；
  - 组合特征：基础特征的组合，如1天，7天，15天等等；
  - 统计特征：如喜欢的均值，方差，标准差等等；
  - 稀疏特征：one-hot编码的特征，如手机类型等等。

二、机器学习初探

- B)机器学习建模：
  - 基于经典机器学习方法的stack model
    - 第一层：RF，GBDT，KNN， XGBOOST等等
    - 第二层：LR逻辑回归



机器学习模型应用于线上业务的流程：

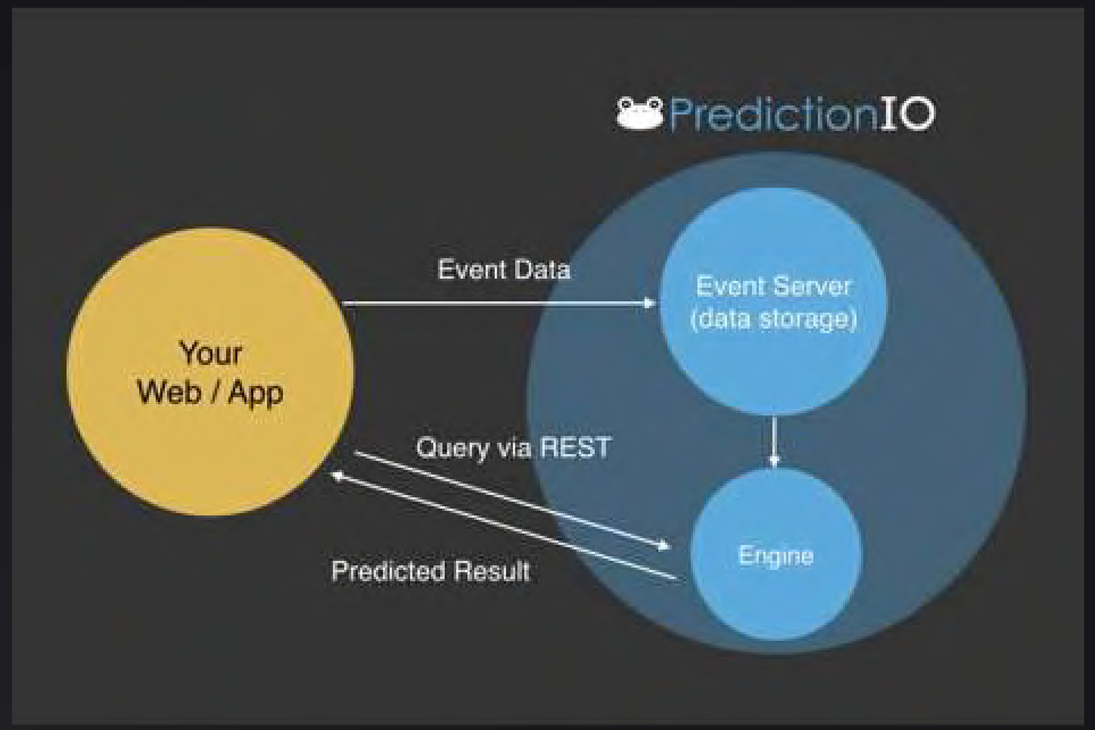


二、机器学习初探

- 为什么需要实时化改造？
  - T+1清算留给灰产一定的活动时间
  - 难以实施监控业务的风险情况

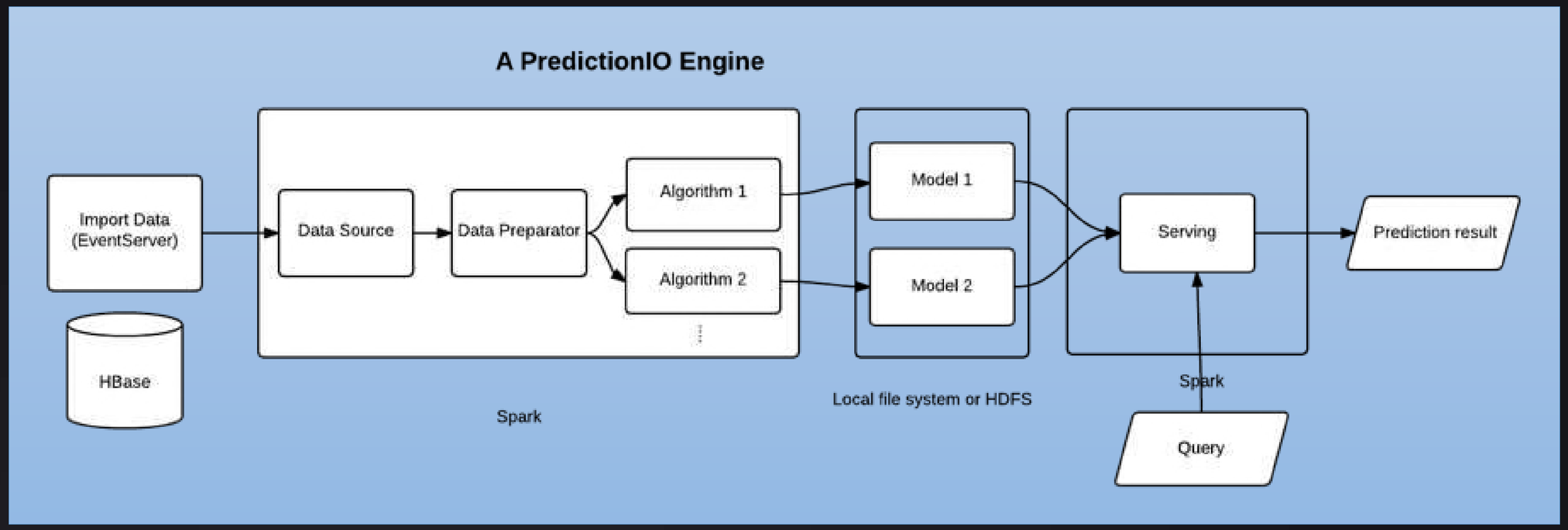
- Prediction IO ：
  - https://github.com/apache/incubator-predictionio
  - http://predictionio.incubator.apache.org/

- 原因：
  - GITHUB上开源10大机器学习工具之一，star 10000+
  - 通用的开源机器学习算法平台，快速独立地构建机器学习算法引擎
  - 通用的机器学习应用模板，可以快速针对业务进行部署



From Prediction IO official website

三、机器学习初探



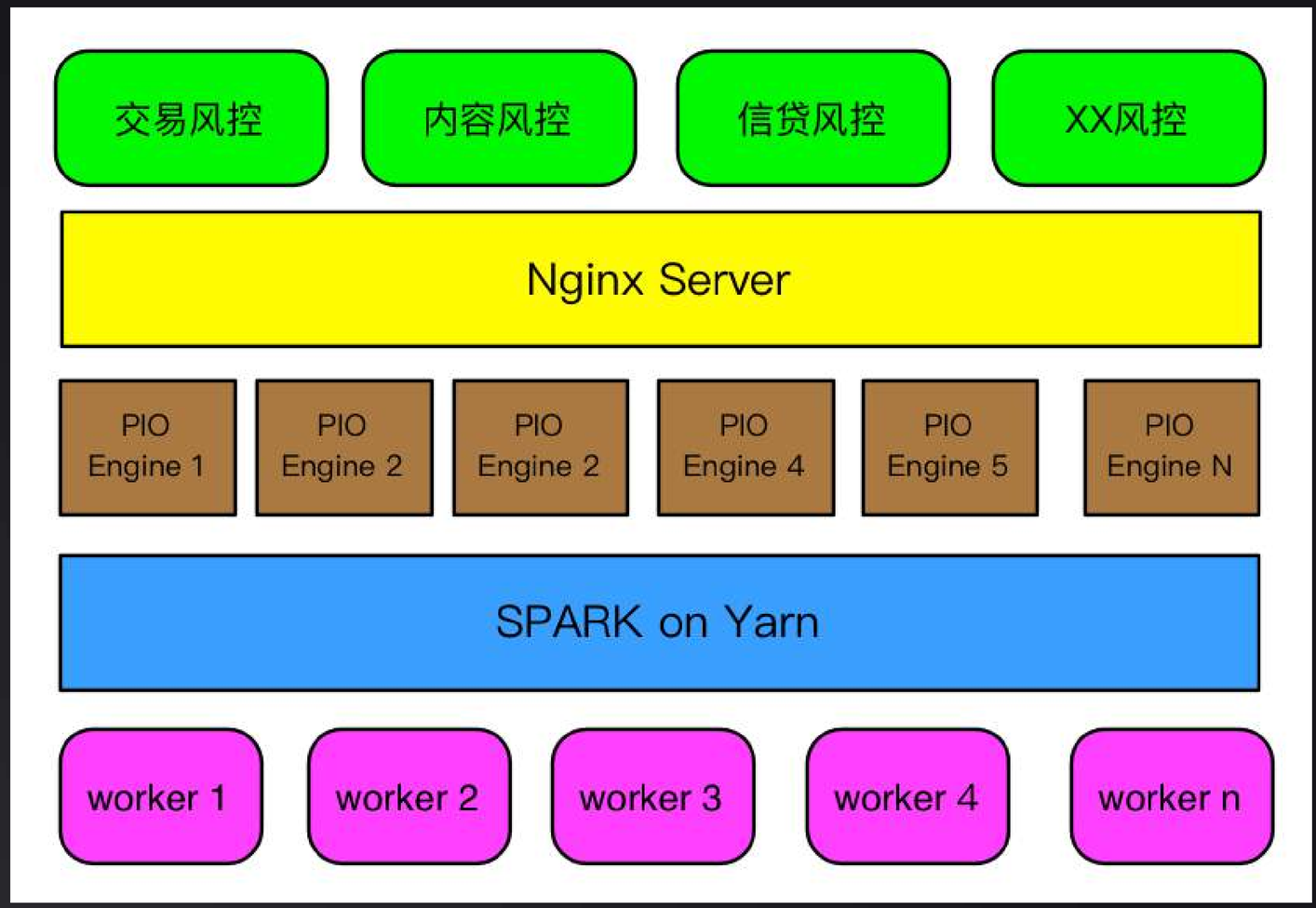
From Prediction IO official website

- 数据：[D] Data Source and Data Preparator: 实时特征+离线特性
- 算法：[A] Algorithm: RF，GBDT
- 服务：[S] Serving: Http
- 评估：[E] Evaluation Metrics: Cross-Validation

二、机器学习初探

- 计算能力问题
  - SPARK on YARN，集群提高运算能力

- 单点问题:
  - NGINX + 一个业务部署多个PIO



PIO架构图

二、机器学习初探

- 问题和改进：
- 数据存储：Mysql/Hbase/MongoDB
  - 利用hbase作为event collector的存储时，全表查询可能超时，而且预测时RT过高。
  - 将事件服务存储替换成Kv-Store（公司自研的基于Redis的存储中间件）

Engine Information

Training Start Time	2016年1月22日 星期五 下午05时36分18秒 CST
Training End Time	2016年1月22日 星期五 下午05时38分07秒 CST
Variant ID	2
Instance ID	AVJosQ-azMvvHRLqTcVZ

Server Information

Start Time	2016年1月22日 星期五 下午05时49分22秒 CST
Request Count	11
Average Serving Time	0.4541 seconds
Last Serving Time	0.1620 seconds

Engine Information

Training Start Time	Wednesday, November 2, 2016 4:38:52 PM CST
Training End Time	Wednesday, November 2, 2016 4:40:50 PM CST
Variant ID	5
Instance ID	AVgkMOLVLlqOujk0Z5sv

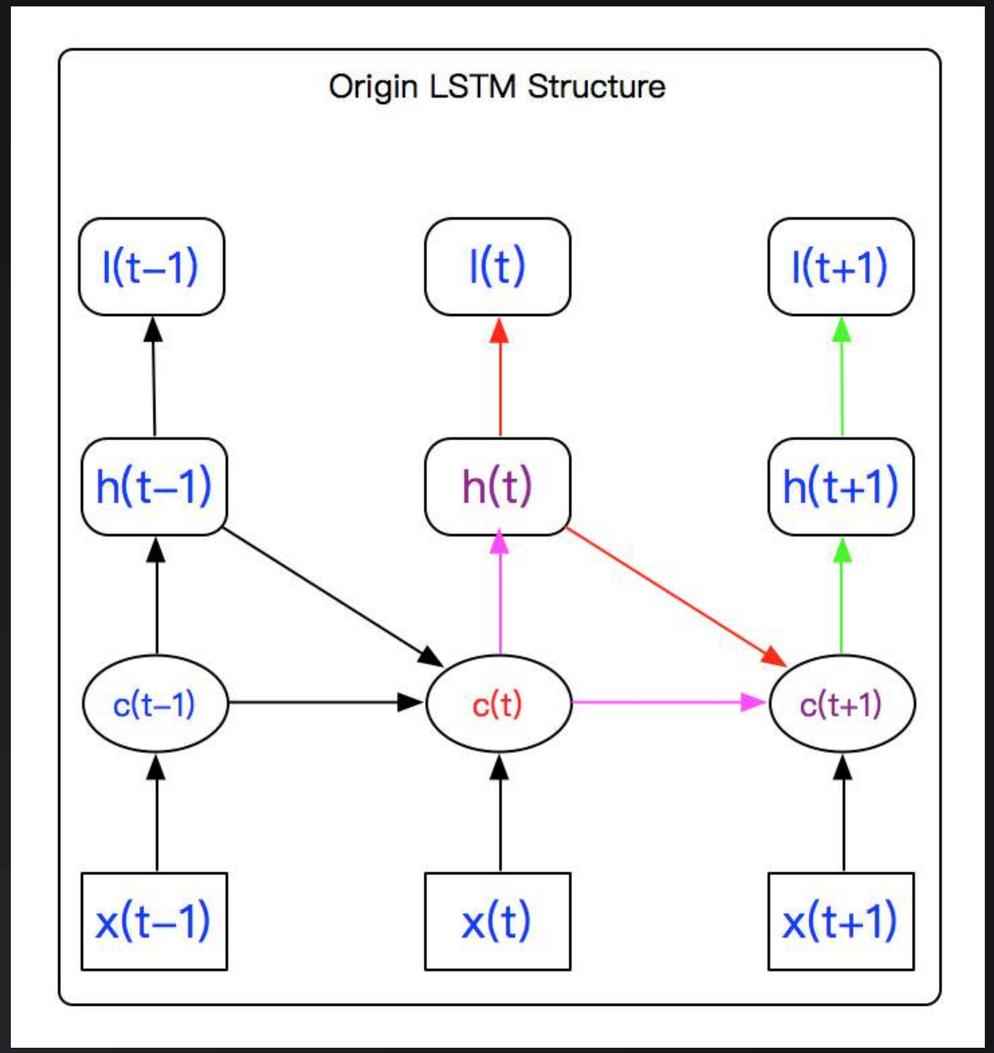
Server Information

Start Time	Friday, December 2, 2016 11:17:24 AM CST
Request Count	47743979
Average Serving Time	0.0090 seconds
Last Serving Time	0.0090 seconds

三、新技术探索

3.1深度学习

- 为什么需要深度学习?
- 1) Feature engineering + 统计机器学习模型走到尽头，90%，80%以上
- 2) Deep learning，本质上是，Representation learning
- 3) 用户行为，时序关系，LSTM

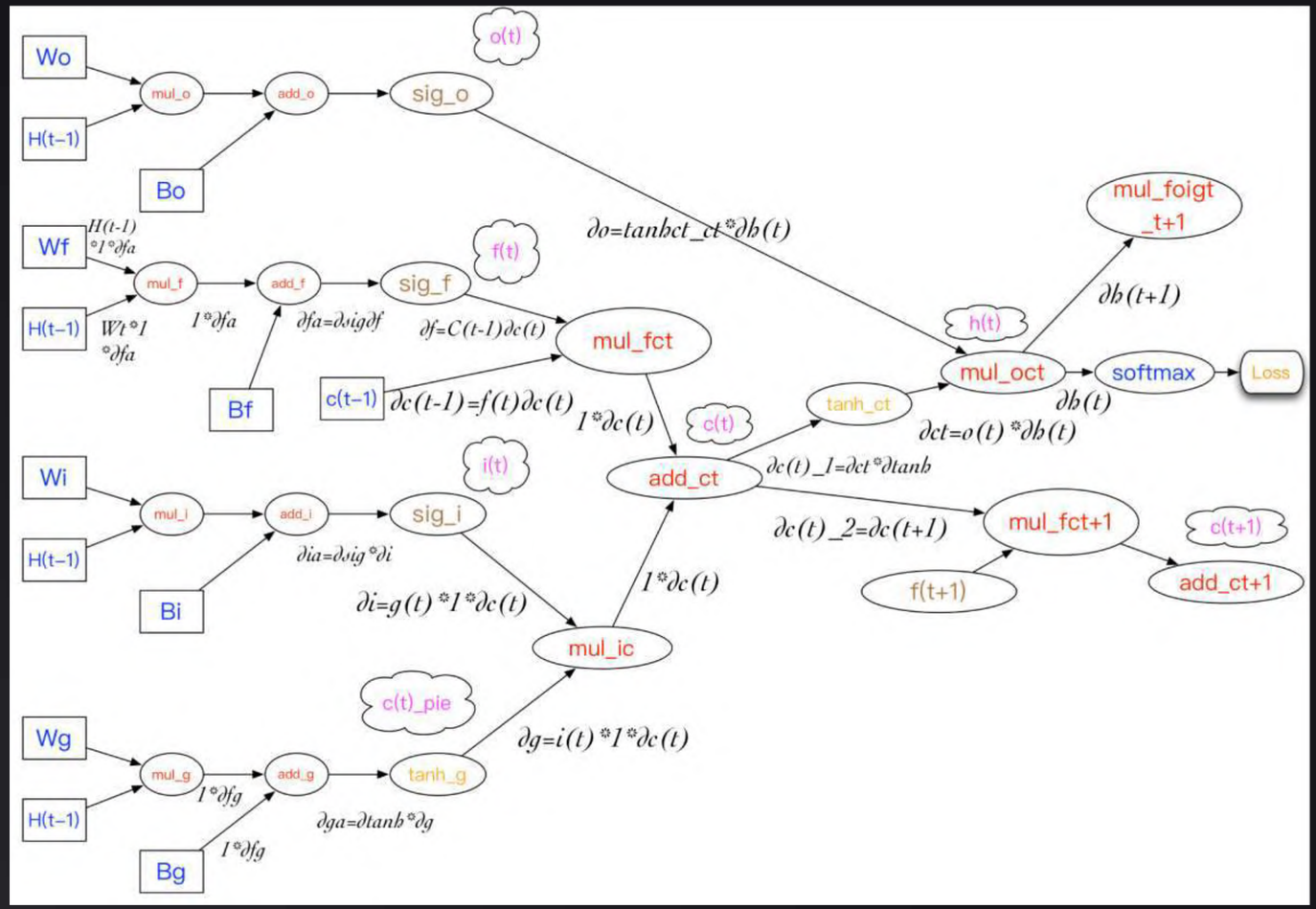


三、新技术探索

3.1深度学习

- 弯路：
- 将人工处理过的特征放入CNN，未有明显提升

- LSTM
- 1、在自然语言处理中有大量成功的使用
- 2、适用于存在时序关系的数据
- 3、用户浏览互联网的行为有自然的时序关系

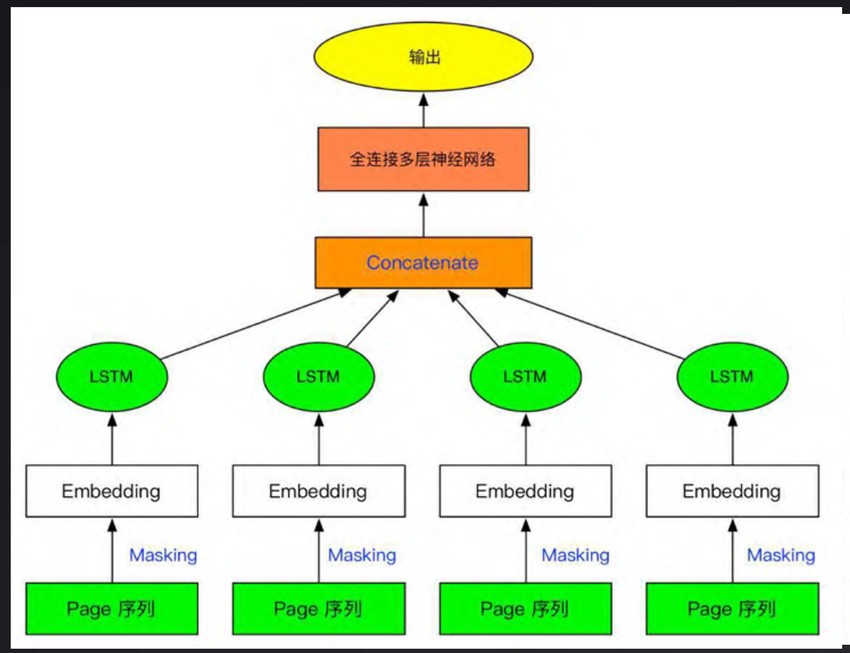


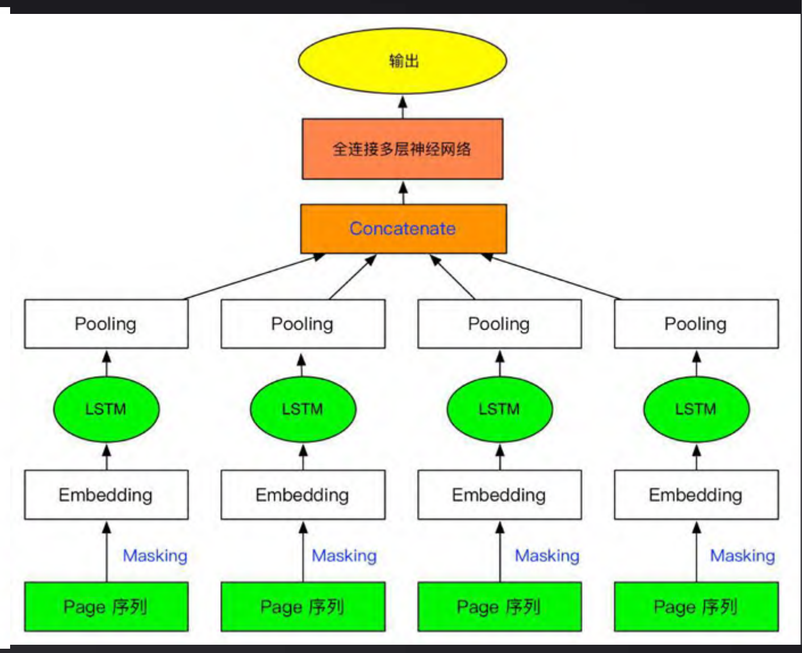
LSTM BPTT

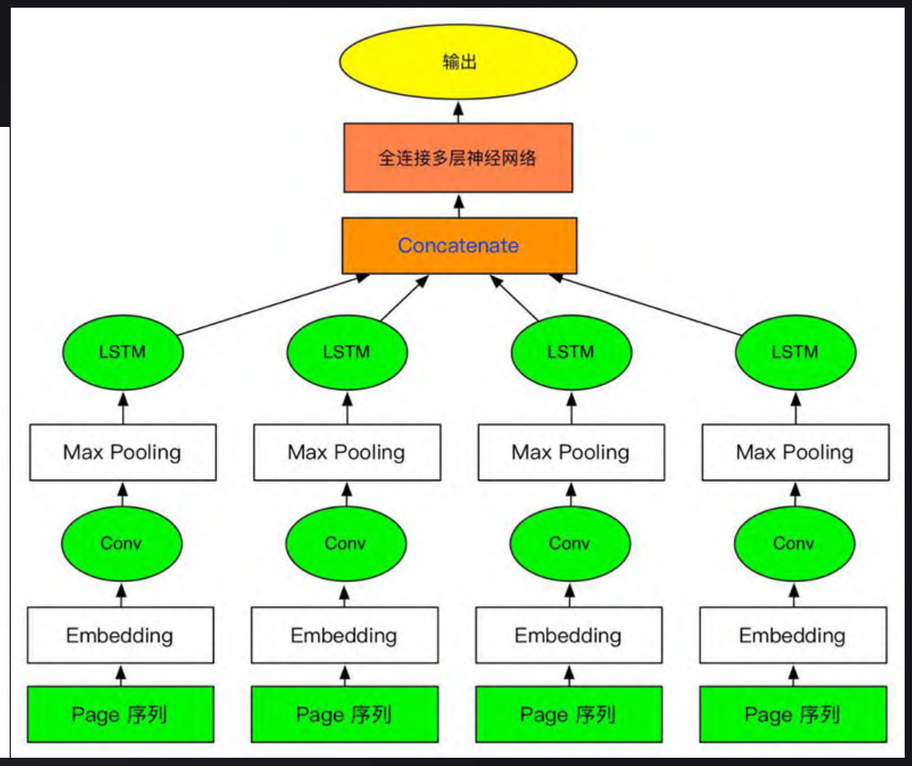
三、新技术探索

3.1深度学习

- 1、masking，卷积层都能够提高网络的分类效果；
- 2、卷积层对效果的提升更明显；
- 3、加入人工特征能够进一步提高效果。





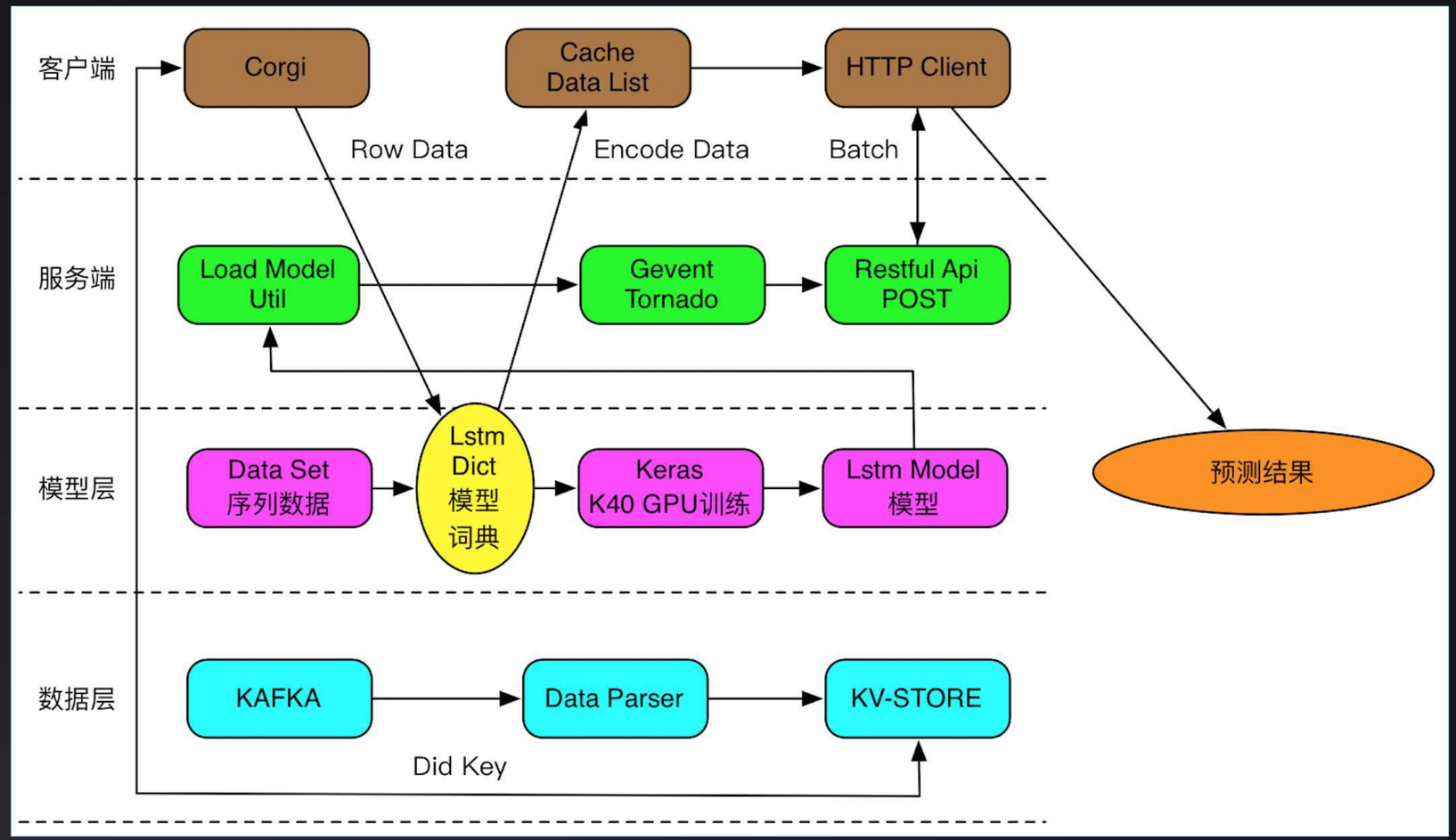


附： http://mp.weixin.qq.com/s/aOVMuhD--iJpMZtDXKgn4g LSTM算法原理及其在美联风控中的应用

三、新技术探索

3.1深度学习

- 实时化的几个关键问题：
- 1、速度较慢：异步生效，通过批处理提高吞吐能力
- 2、模型效果随时间衰减：效果监控及时迭代更新
- 3、特征获取：数据的平滑处理，取当前时间前一天内的特征

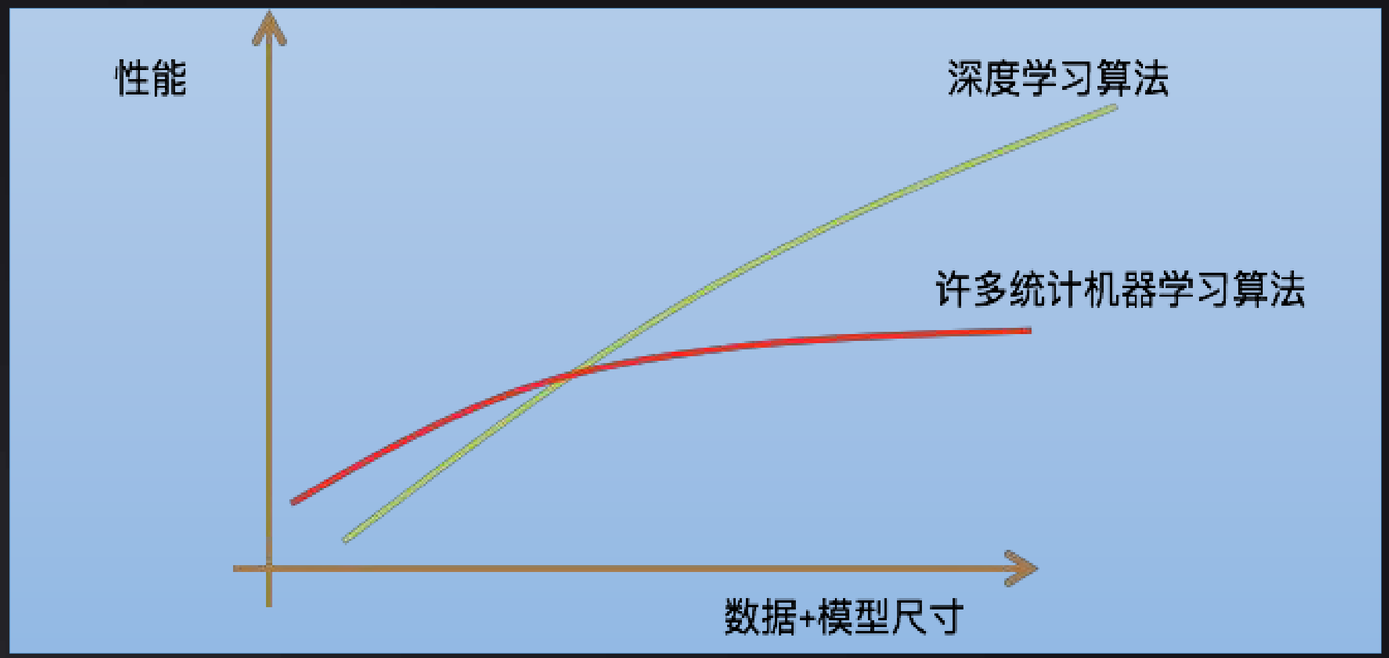


架构图

三、新技术探索

3.1深度学习

- 1、模型的尺寸对效果有正向影响
- 2、尺寸更大的模型需要更多数据
- 3、标签数据更丰富时，效果更好
- 4、数据不足时，深度学习的效果可能差
- （from NG）



三、新技术探索

3.1深度学习

- a）深度学习本身具有丰富悠久的历史，但是从不同的角度出发有很多不同得名，所以历史上其流行有过衰减的趋势。
- b）随着可用训练数据量的逐渐增加，深度学习的应用空间必将越来越大。
- c）随着计算机硬件和深度学习软件基础架构的改善，深度学习模型的规模（尺寸）必将越来越大。
- d）随着时间的推移，深度学习解决复杂应用的精度必将越来越高。
- （From Yoshua Bengio《Deep Learning》一书）

三、新技术探索

3.2基于图的机器学习

- 为什么需要基于图的机器学习？
- “粗刷”的方式打击殆尽，目前的挑战主要来源于团伙形式的“精刷”

- 图模型为什么有效？
- 图挖掘则是通过“全局性”来深挖作弊行为，而不仅仅利用“局部性”信息。

- 怎么做？
- 1）全站事件的前置团伙挖掘；
- 2）风险事件中的后置团伙挖掘。

有组织的灰产群

三、新技术探索

3.2基于图的机器学习

- **前置团伙挖掘**
  - FRAUDAR（KDD2016），Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, Christos Faloutsos. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016, San Francisco, USA

- 原理：该算法要解决的问题是找出站内最善于伪装的虚假账户簇。其原理是：
  - 虚假账户会通过增加和正常用户(目标)的联系来进行伪装，而这些伪装（边）会形成一个很紧密的子网络；
  - 这样就可以通过定义一个全局的嫌疑程度的度量，再移除二部图结构中的边，使得剩余网络结构对应的度量的值最大，就找到了最紧密的子网络，而这个网络就是最可疑的。

- 如何在风控中使用？
  - 用户关注其他用户或者用户浏览购买商品的二部图；
  - 虚假用户与目标之间就会形成一个“dense”的子网络。

三、新技术探索

- 算法的核心计算过程可以简要描述如下，具体可以参考原论文中Algorithm1的伪代码：
  - a、建立优先树（一种用于快速移除图结构边的树结构）；
  - b、对于二部图中的任意节点，贪心地移除优先级最高（由优先树得到）的节点，直至整个网络结构为空；
  - c、比较上述每一步得到的子网络结构对应的全局的度量，取该值最大的子网络结构，那么该子网络结构就是最紧密的子网络，也就是最可疑的团伙。

  - 其中最关键的地方是定义了一个全局度量，该Metric的定义是（目标度量），可以理解成子网络结构中每个点的平均可疑程度
    - $g(s)=f(s)|s|$
  - 其中，
    - $f(s)=fv(s)+fe(s)$

- 通过这样的定义，很容易可以得出4条性质：
  - 固定其他条件，包含更高可疑度节点的子网络比包含较低可疑度节点的子网络更可疑；
  - 固定其他条件，增加可疑的边使得子网络更可疑；
  - 如果节点和边的可疑程度固定，那么大的子网络比小的子网络更可疑；
  - 总的可疑程度相同，那么包含节点数少的子网络更可疑。

三、新技术探索

3.2基于图的机器学习

- 改进：
- 原始论文中只找出最dense的子网络，改进点是找出最dense的子网络后，将这个子网络对应的邻接矩阵的列设置为0，通过循环地执行FRAUDAR算法，删除当前子网络，可以得到嫌疑程度从高到低的多个子网络。

三、新技术探索

3.2基于图的机器学习

- 效果展示
- 典型例子：
  - 17万+用户的大团伙

垃圾账户的图片

三、新技术探索

3.2基于图的机器学习

- **后置团伙挖掘（社区发现算法，Louvain，Fast unfolding of communities in large networks）**
- 原理：
- Louvain算法是基于模块度（Modularity）的社区发现算法，该算法在效率和效果上都表现比较好，并且能够发现层次性的社区结构，其优化的目标是最大化整个图属性结构（社区网络）的模块度。

- 其中需要理解的核心点有：
- a、模块度Modularity的定义，这个定义是描述社区内紧密程度的值Q；

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j),$$

- b、模块度增量delta Q，即把一个孤立的点放入一个社区C后，计算Modularity的变化，其中计算过程的要点是，首先计算1个点的Modularity，和社区C的Modularity，再计算合并后新社区的Modularity，新社区的Modularity减去前两个Modularity就是delta Q。

$$\Delta Q = \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

- 对上述公式的理解是，将delta Q展开其等价于1/2 *(k_i,in/m - Sum_tot/m *ki/m)，其中k_i,in/m表示的是将孤立的节点和社区C放在一起对整个网络Modularity的影响，而Sum_tot/m和ki/m分别表示孤立的节点和社区C分开式分别对整个网络Modularity的影响，所以他们的差值就反应了孤立的节点放入社区C前后对整个网络Modularity的影响。

三、新技术探索

3.2基于图的机器学习

- 算法的计算过程如下：
- a、每个点作为一个community，然后考虑每个community的邻居节点，合并到community，然后看delta Q；找到最大的正delta Q，合并点到community；多进行几轮，至不再变动，那么结束；其中存在的问题是，不同的节点访问顺序将导致不同的结果，试验中发现这个顺序对结果影响不大，但是会在一定程度上影响计算时间。
- b、将新的community作为点，重复上述过程。那么如何确定新的点之前的权重呢？答案是将两个community之间相邻的点之间的权重和作为两个community退化成一个点后的新的权重。

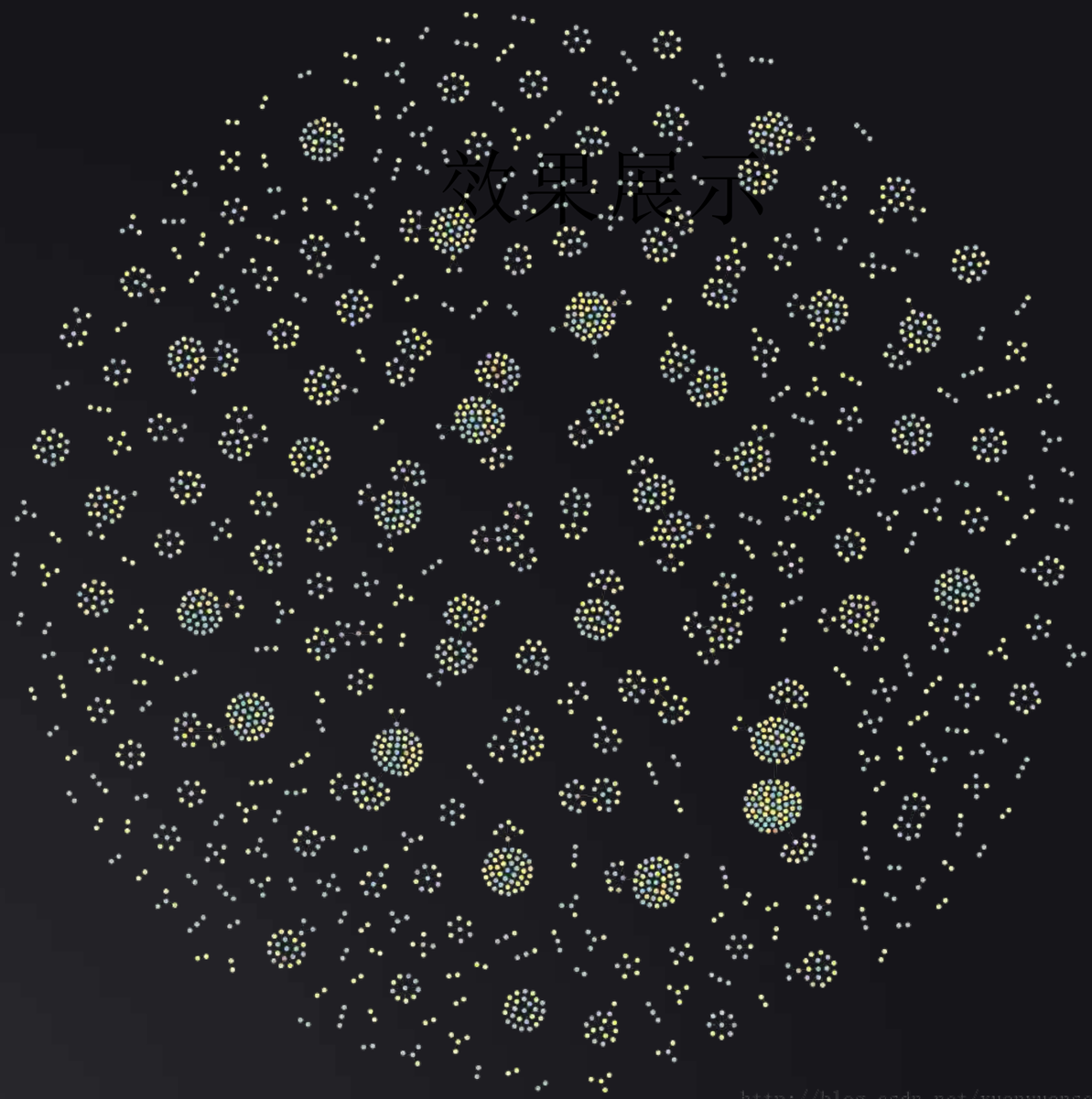
- 该算法的优点主要有3个：
- a、易于理解；b、非监督；c、计算快速，最后我们可以得到的结果是层次化的社区发现结果。

- 改进：A New Randomized Algorithm for Community Detection in Large Networks，其实现方式比较直接，就是考虑一个点周围的百分之多少点进行归并。可以在Spark下面通过类似于多路归并来实现。

三、新技术探索

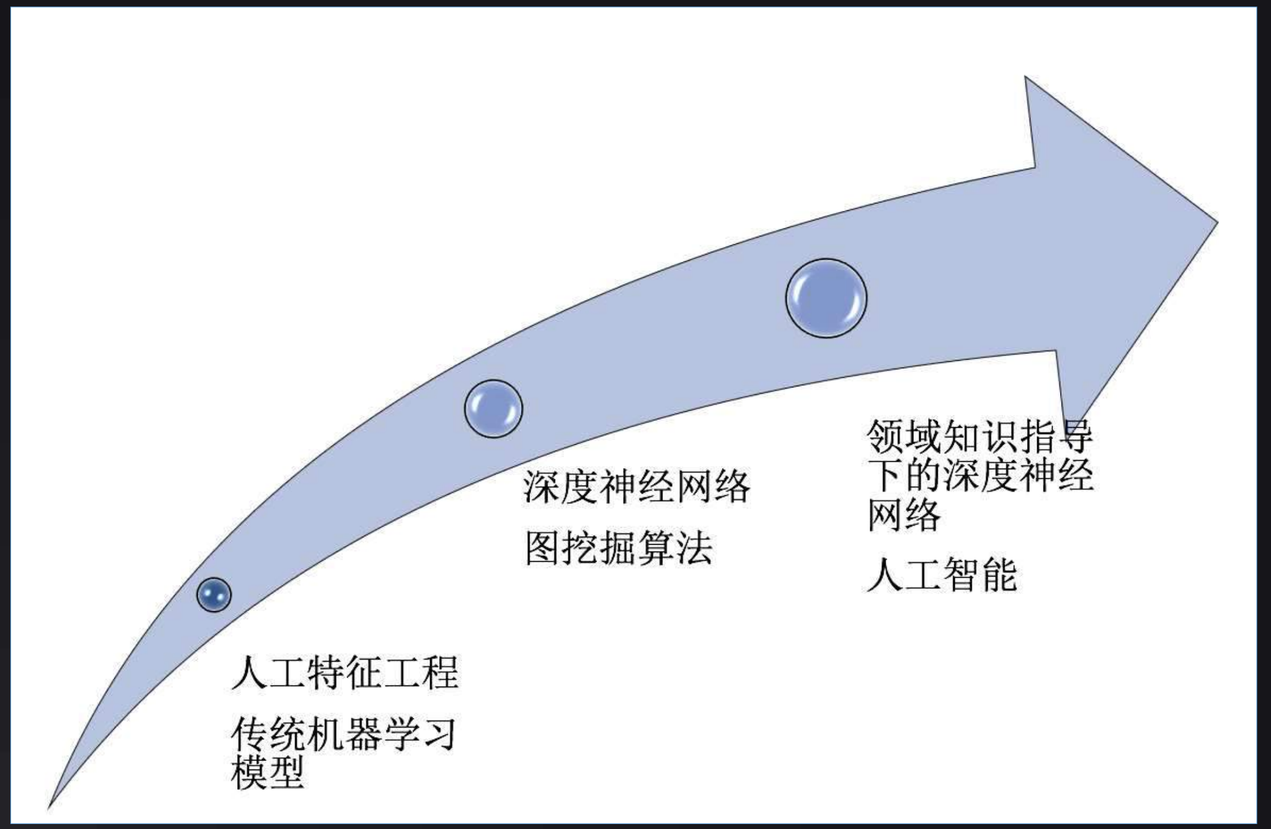
3.2基于图的机器学习

- 为什么是有效的？
  - 团伙内的成员总是有关联的
- 如何在风控场景中使用？
  - IP聚集？
  - 手法类似？
  - 主体相似性质？
  - 目标聚焦？



四、未来展望

- 风控算法策略演进之路
- a）特征工程，经典机器学习；
- b）深度学习，图模型；
- c）**领域知识+深度神经网络；人工智能**



一点点心得体会

- 在企业中算法工程师应该具备这四个能力。
- 1、业务理解。　　主动性。
- 2、算法积淀。　　必备功底。
- 3、实验与调优。　　核心能力。
- 4、应用与迭代。　　站在企业角度，业务效果的持续提升。

THANK YOU

蘑菇街风控策略团队招贤纳士！
lianhua@meili-inc.com