



Baidu Machine Learning

百度机器学习云平台

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

大数据技术探索和价值发现



百度基础架构部高级架构师

刘伟

- BML 平台概况
  - 主要特点
  - 平台承载的业务
- BML 平台功能
  - 预处理
  - 通用机器学习算法
  - 深度学习算法
  - 可视化算法
  - API 和 ELF
- 平台应用案例
  - Logistic Regression 应用
  - 深度学习应用



- 主要特点:

- 支持算法丰富，自 2009 年开始研发第一个逻辑回归算法，目前 BML 平台已包含 20 多种大规模并行机器学习算法
- 性能极致，所有算法均为分布式实现，经历数年持续优化，用上几乎所有可能的计算/通信优化技术，速度业界一流
- 经历线上大规模使用部署考验，承载公司内各种重要业务线上使用，包括凤巢广告 CTR 预估，搜索排序
- 易用，通用，支持多种方式调用，文档详尽，配置灵活



# BML 平台概况 – 承载的主要业务



百度

手机搜索，随时随地



地图

搜索功能完备的网络地图



百度推广

获得新客户和合作伙伴



百度联盟

与百度合作，变得更强



百度杀毒

更快，更安全



百度翻译App

您的掌上翻译专家



百度外卖

在线外卖订餐产品



百度糯米

省钱更省心，全场随便退！



新闻

搜索浏览最热新闻资讯



百度私有云

bpc.baidu.com



百度开放云



BML机器学习云平台



## ● 预处理:

- 分词
- TF-IDF 计算
- 专名识别
- 采样
- 特征转换/抽取
- 特征统计
- 大规模 ID 化
- 大规模相似度计算
- 大规模 Join

## ● 分类算法:

- 逻辑回归
- SVM
- 最大熵分类
- 深度神经网络
- 随机森林
- GBDT

## ● 聚类算法:

- K-Means
- 凝聚层次聚类
- 谱聚类



- 主题模型:
  - PLSA
  - 半监督 PLSA
  - LDA
- 推荐算法:
  - 矩阵分解
  - 协同过滤
  - KNN
- 深度学习:
  - RBM
  - RNN
  - Word2Vec
  - LSTM
  - CNN
- 可视化:
  - PCA
  - MDS
  - t-SNE



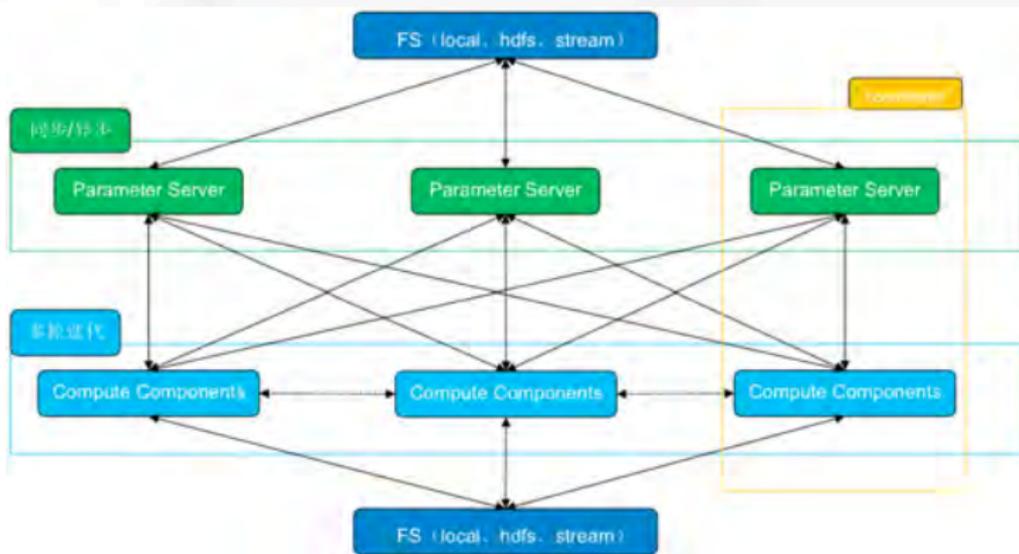
- 序列模型
  - CRF
  - 半监督 CRF
  - HMM
- 机器学习 Solver:
  - SGD, ASGD
  - L-BFGS
  - ADMM
- 在线学习:
  - online DNN
  - FTRL
  - online HMM
  - online SVM



- API:
  - 统一样本，模型，特征，各种词典的格式
  - Restful API 调用各种机器学习方法
  - 模型上线部署
  - 在线预测、模型分析和可视化 API
  - 访问基于百度大数据预训练的各种模型
- ELF(Essential Learning Framework):
  - 适合在百度公有云上自研算法使用
  - 类 Spark 的全内存 DAG 计算引擎
  - 高性能，高可用的分布式 Parameter Server

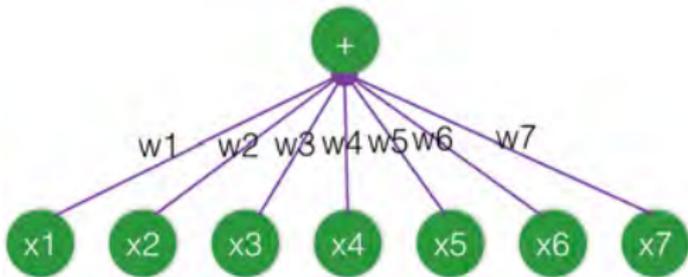


- ELF 架构图：



- 逻辑回归应用：

- 典型场景：点击率建模
- 数据：各种用户点击日志
- BML-LR 算法特点：
  - 支持数百 T 样本数据训练，千亿特征，千亿样本
  - 支持连续值，离散值
  - 支持 L-BFGS 和 SGD 两种算法求解



- 广告 CTR 预估:(ref:CIKM 2013 Kai Yu)

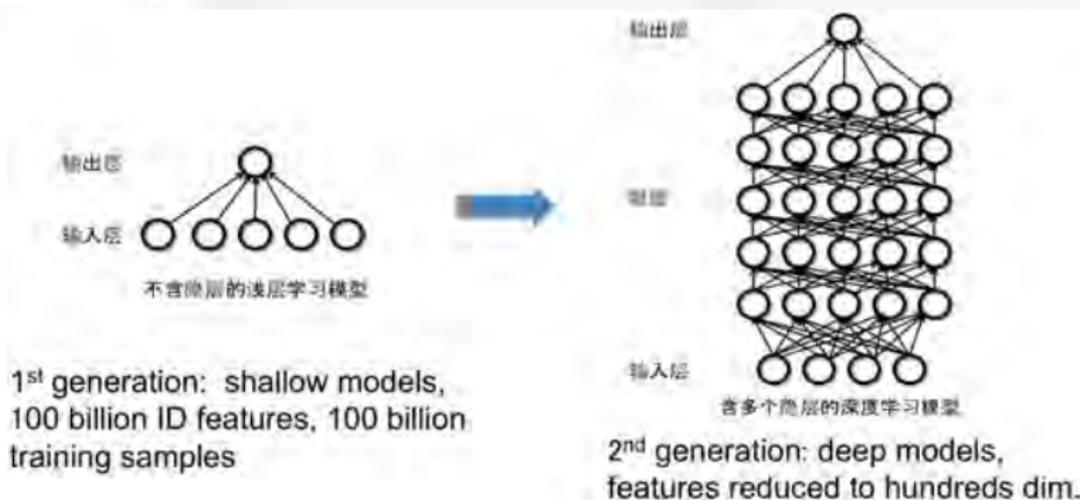


Figure: CTR 预估神经网络模型图



- 搜索 LTR 排序:(ref:CIKM 2013 Kai Yu)

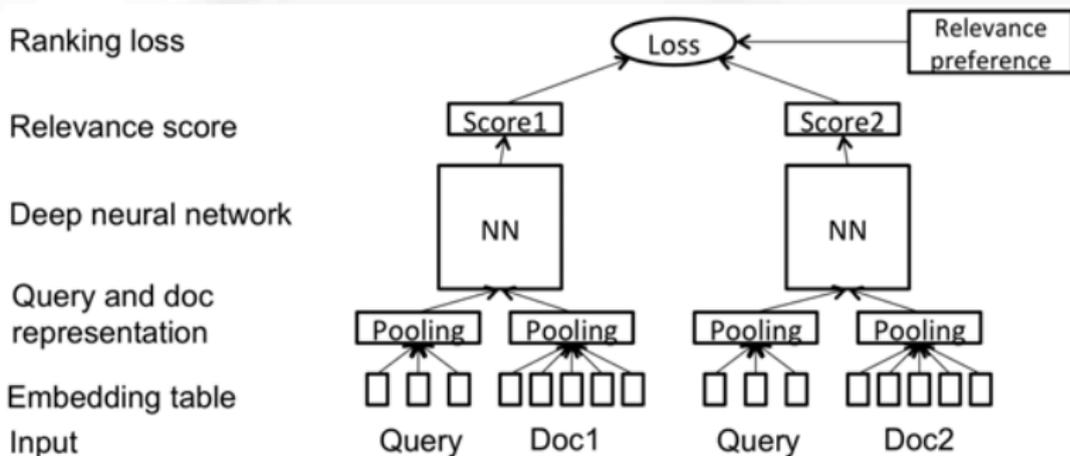


Figure: LTR 排序神经网络模型图



- BML Word2Vec 效果:

```
word=苹果, id=359289
printx: 0.0125495 -0.0188294 -0.0318404
        -0.0425401 0.0194138 0.0407652 -0.09011
0951118 -0.00359665 -0.00810341 -0.03609
594268 0.0191381 0.0391772 -0.0876153 0.
[0]: 苹果 1
[1]: 推出 0.965608
[2]: 购买 0.949781
[3]: 品类 0.945782
[4]: 面市 0.945037
[5]: 富士康 0.94402
[6]: 直销 0.943164
[7]: 实惠 0.942807
[8]: 水货 0.940582
[9]: 中档 0.938433
```

```
word=北京, id=388718
printx: -0.0198855 -0.0838369 -0.109926 0.
        .0095244 0.0720015 0.105703 -0.0727412 0.1
0.135655 0.045565 0.0196295 -0.0660551 -0.
0852344 -0.0243348 -0.0520941 0.216335 0.6
[0]: 北京 1
[1]: 海口 0.985943
[2]: 昆明 0.983359
[3]: 贵阳 0.982533
[4]: 乌鲁木齐 0.982477
[5]: 多伦多 0.982107
[6]: 拉萨 0.978163
[7]: 宜昌 0.977319
[8]: 首尔 0.975819
[9]: 芝加哥 0.974543
```

Figure: Word2Vec 训练效果图



- BML Word2Vec 效果:



Figure: 训练结果向量可视化图



ITPUB

ChinaUnif

ITPUB

THANKS