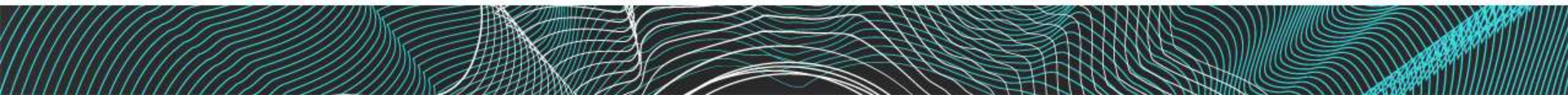


BML 百度大规模机器学习云平台实践

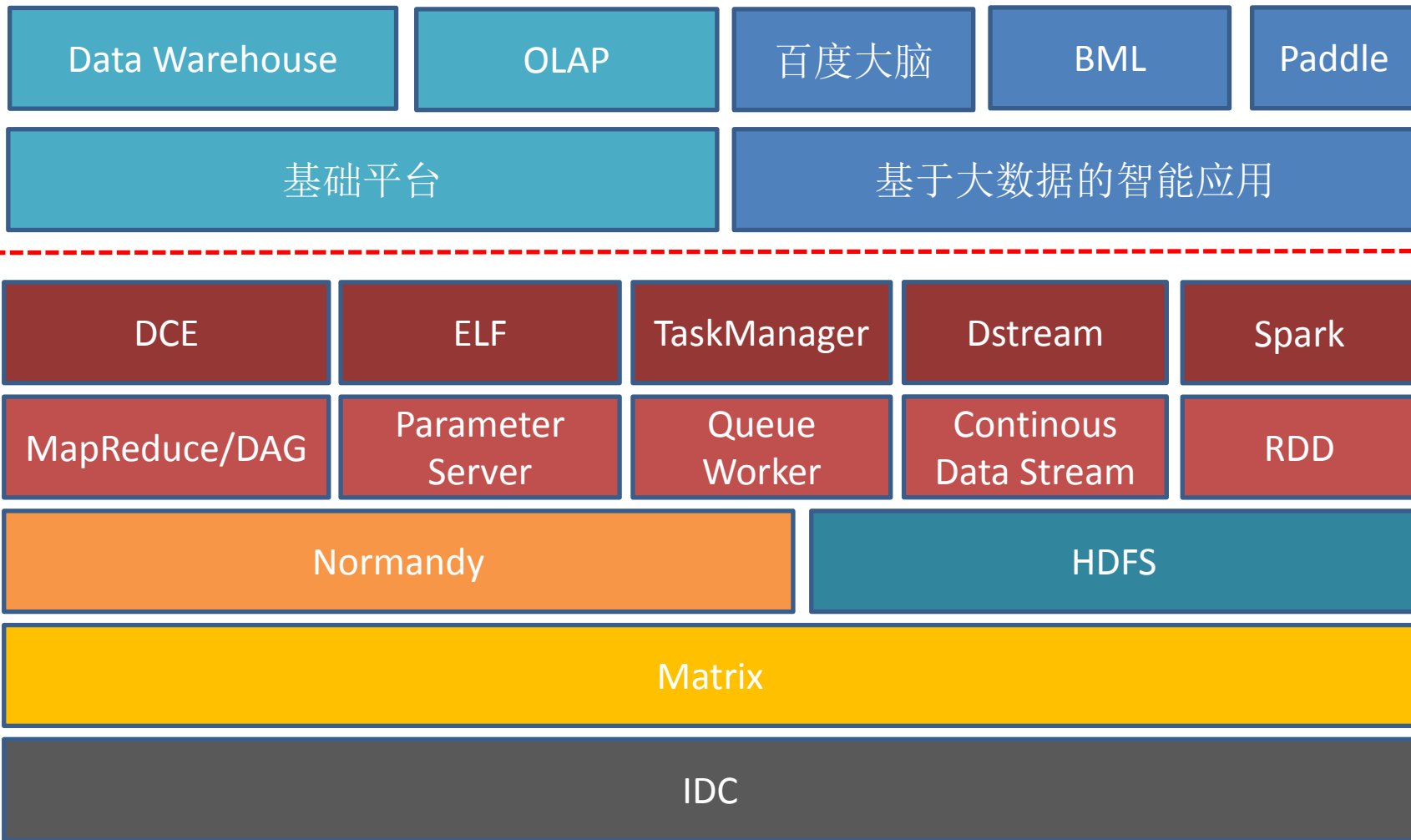
Practice of Baidu Large Scale Machine Learning Cloud

百度开放云架构师 沈国龙

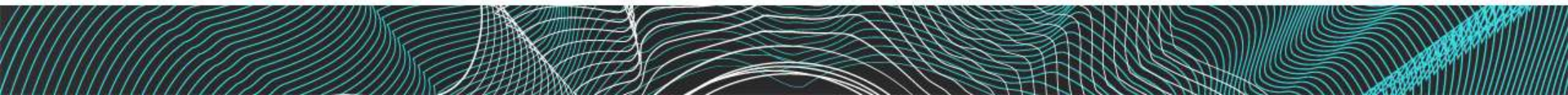
- 百度每天响应数十亿次的搜索
- 支持百万企业客户的推广需求
- 拥有20+用户过亿的移动产品
- 每天处理的数据量将近100个PB，相当于5000个国家图书馆的信息量总和
- 2013年，百度Hadoop单集群规模达到全球最大的1.3w台
- 2015年，全球Spark峰会唯一受邀主题演讲的中国企业
- 多位“人工智能”领域泰斗加盟，机器学习技术国内领先



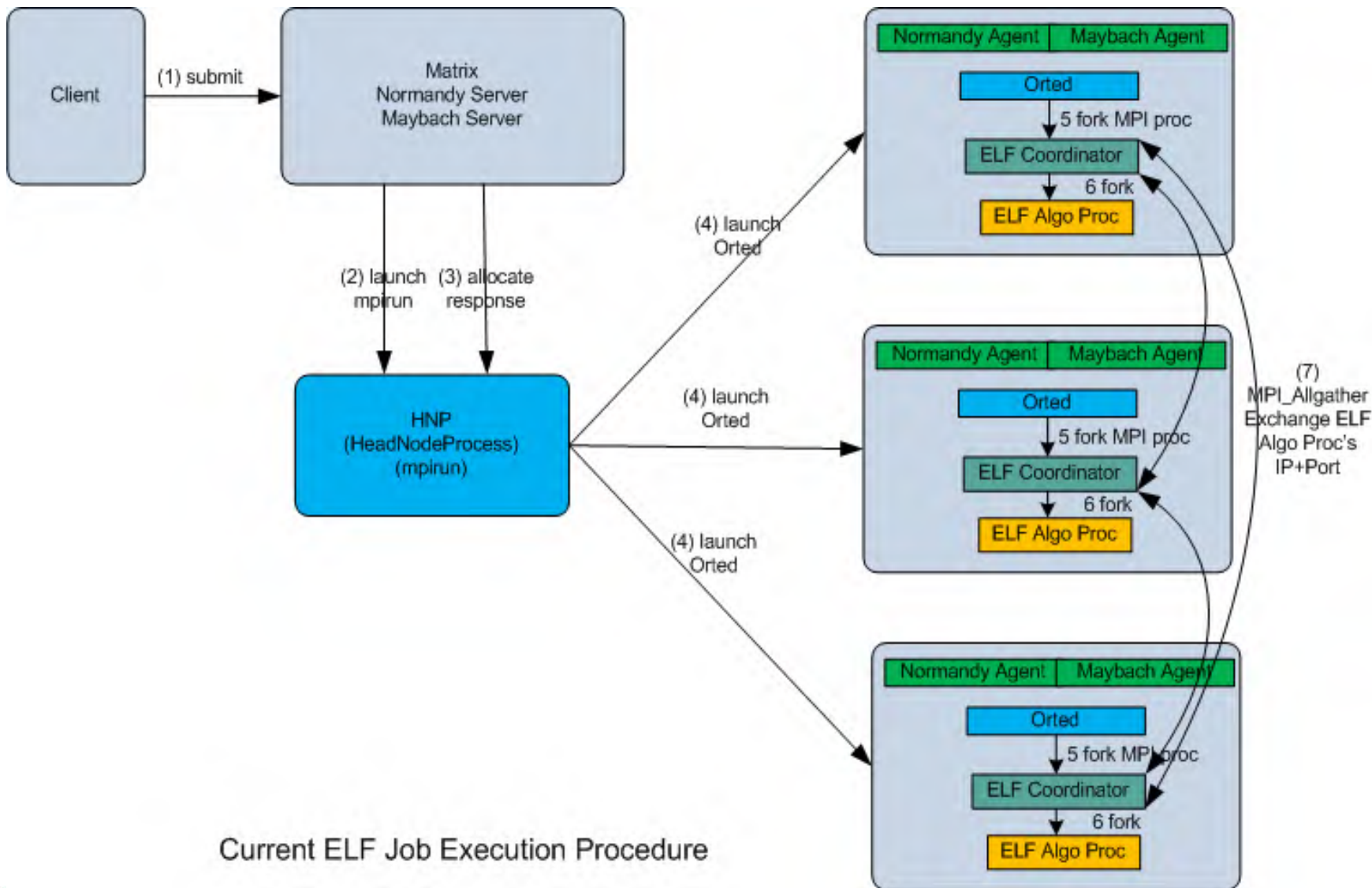


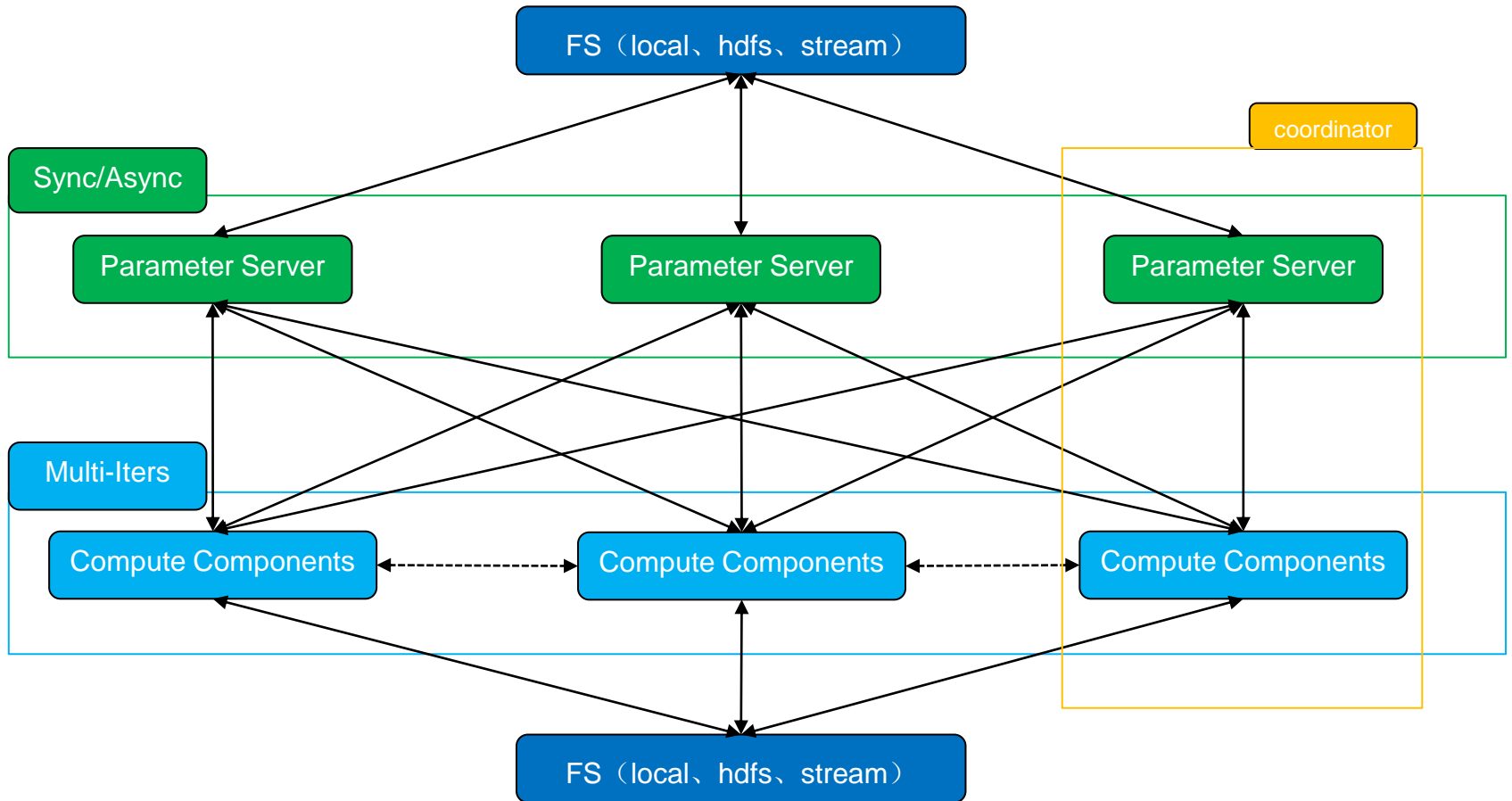


Essential Learning Framework



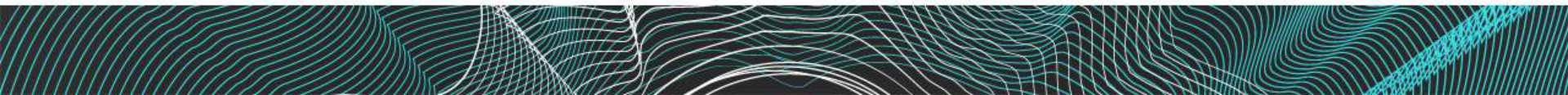
- 百度新三代机器学习计算框架，依赖于百度多年机器学习算法开发以及分布式计算经验
- 设计上汲取了常见计算框架Hadoop、Spark、MPI的精华。拥有和Hadoop一样简单的编程模式，比Spark更快的性能，以及比MPI更易用的接口
- 基于数据流的编程模式，让用户通过简单的map-reduce就能轻松写出高效的并行程序
- 计算过程进行托管，提供了包括多轮数据迭代处理、异步更新、并行通信等功能，让用户不在考虑底层的实现细节，专注算法自身逻辑
- ELF还拥有性能一流的参数服务器（Parameter Server），可用于存储万亿规模参数





- 易用
 - 对用户屏蔽多机分布式细节，用户就像写单机程序一样
 - 编写Online异步SGD仅需要200行代码；
 - 100+行就能实现大规模分布式LR算法；
 - 分布式LDA算法只需要500行代码
- 高效
 - 组件分布式多线程实现，同时支持细粒度的线程控制
 - 节点间通信依赖高效的baidu-rpc
 - 深度优化hashtable，专用于Parameter Server
 - 支持多种不同的参数读取和更新方式

Baidu Machine Learning



算法丰富

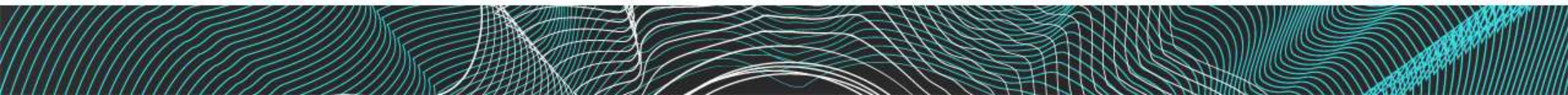
- 自 2009 年开始研发大规模逻辑回归（广告CTR）,已包含20多种并行机器学习算法
- 性能极致
- 所有算法均为分布式实现,经历数年持续优化,应用大量计算/通信优化技术,速度业界一流
- 久经考验
- 百度重要业务线上使用,包括百度推广（凤巢、网盟）广告 CTR 预估,搜索排序等
- 全流程支持
- 预处理、特征分析、模型训练、评估、预测

易用的实力派

- 上手容易,隐藏算法细节,指定简单参数即可完成全过程

团队强悍

- “百度大脑”的支持团队; 获得三个“百度最高奖”的团队; 多位常年从事机器学习的专家

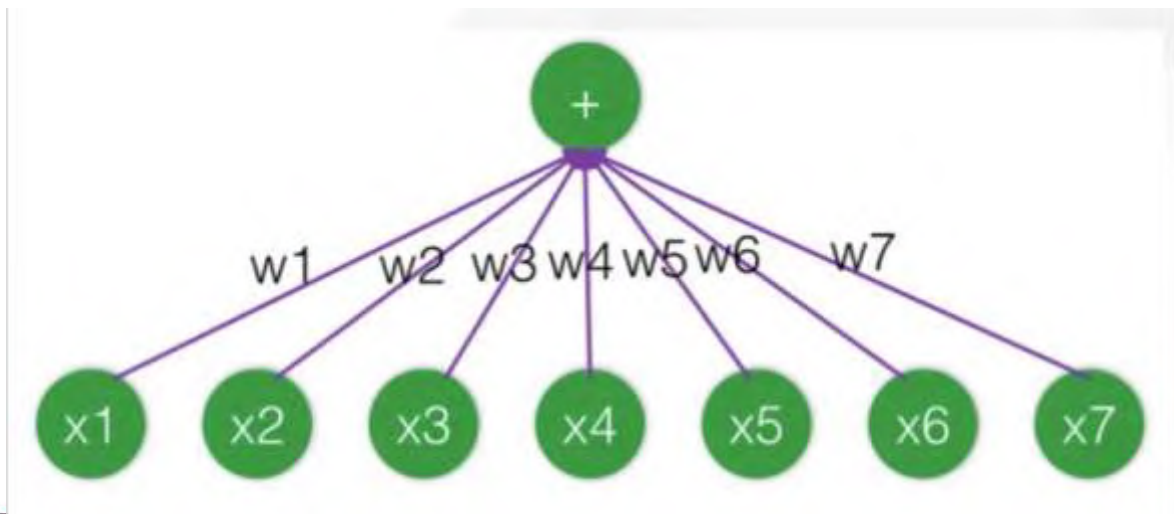


点击率建模

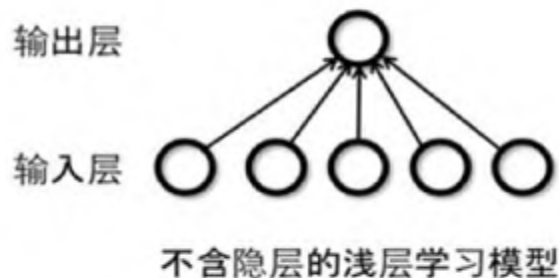
- 应用算法：逻辑回归、GBDT + FFM
- 数据：各种用户点击日志

BML 逻辑回归算法特点：

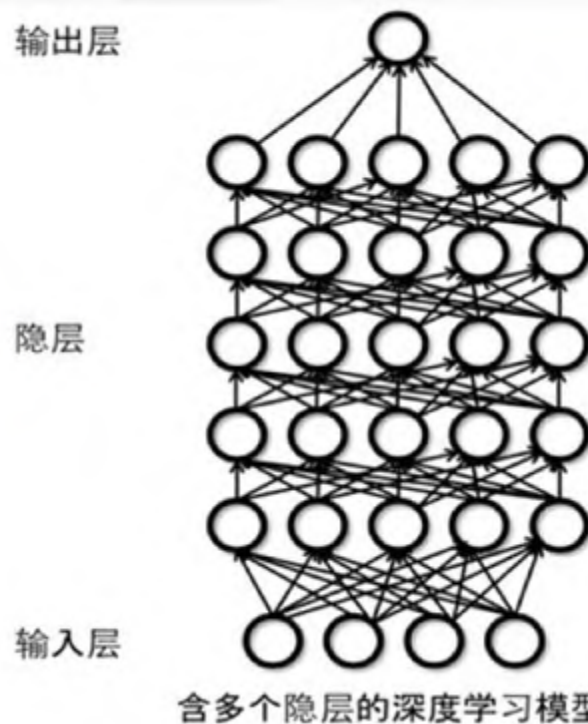
- 支持数百 T 样本数据训练，千亿特征，千亿样本，支持连续值/离散值
- 支持 L-BFGS 和 SGD 两种算法求解



广告CTR预估



1st generation: shallow models,
100 billion ID features, 100 billion
training samples



2nd generation: deep models,
features reduced to hundreds dim.

Figure: CTR 预估神经网络模型图

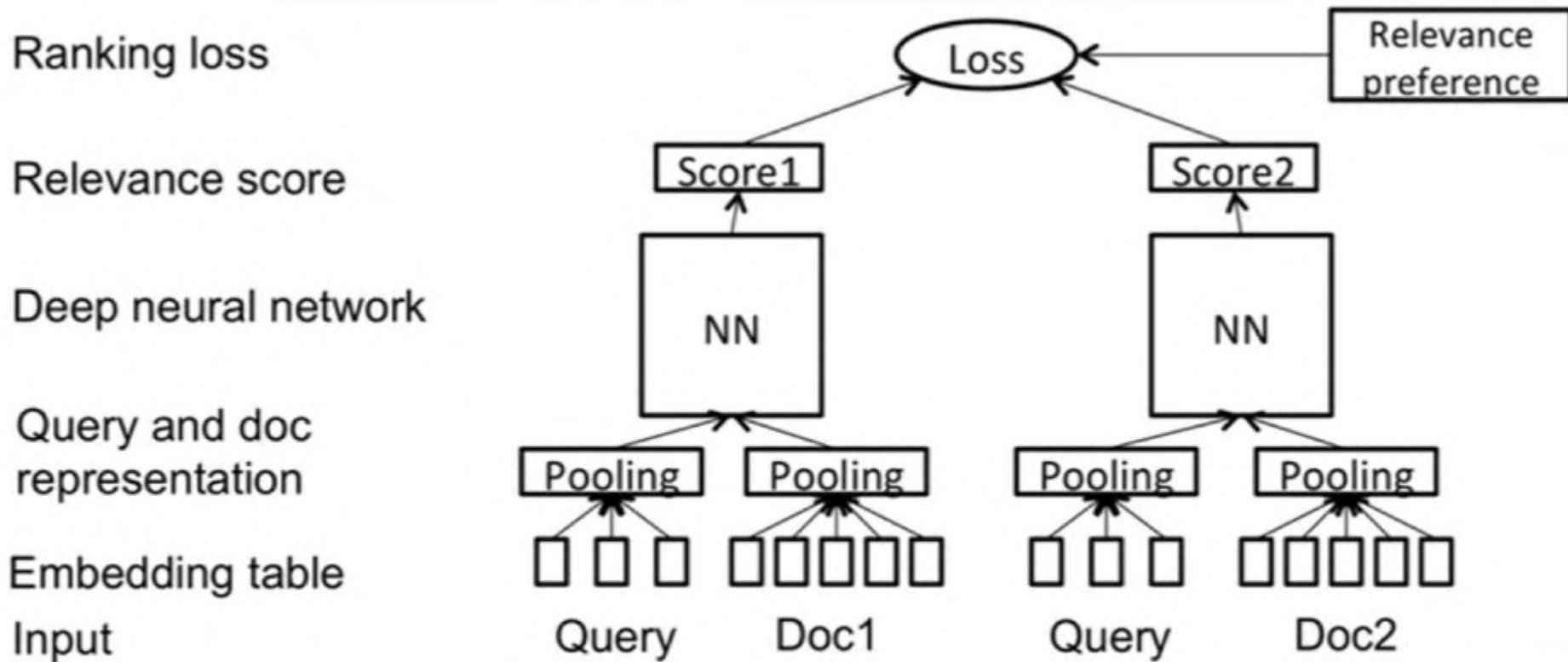
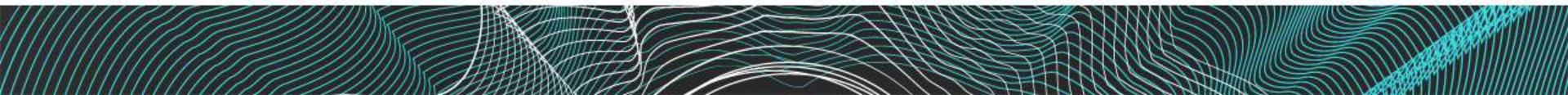


Figure: LTR 排序神经网络模型图

- 百度外卖提升商户推荐排行榜转化率
- 降低百度云端杀毒的误差率
- 深度学习算法，预测硬盘故障
- 语音识别应用**BML**，降低错误率
- 针对搜索用户做直达号的个性化推荐，提升转化率
- 地图推荐商家，应用机器学习提升**CTR**
- 客服问答系统，实现问题自动归类
- 糯米应用**BML**，实现精准营销和店铺推荐



数据

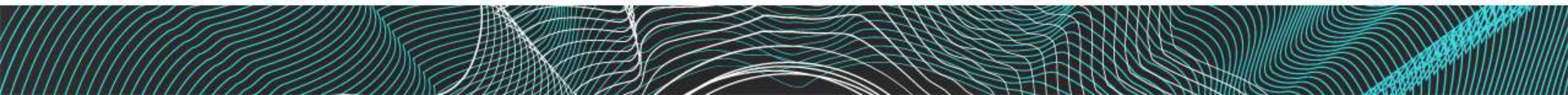
- 数据收集和多套数据的打通
- 清晰、明确、“洁净”的数据源
- Online & Offline数据的结合

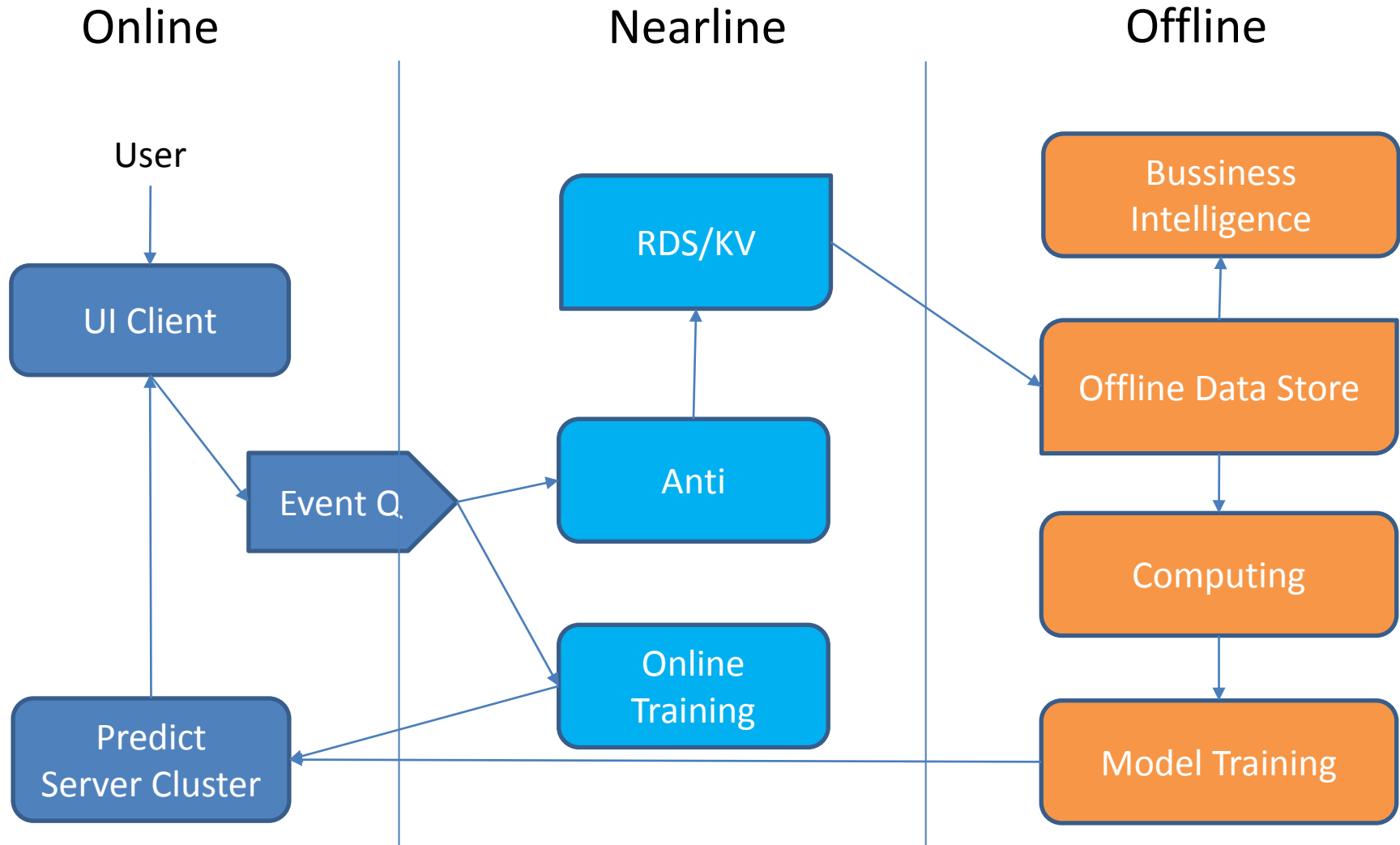
系统

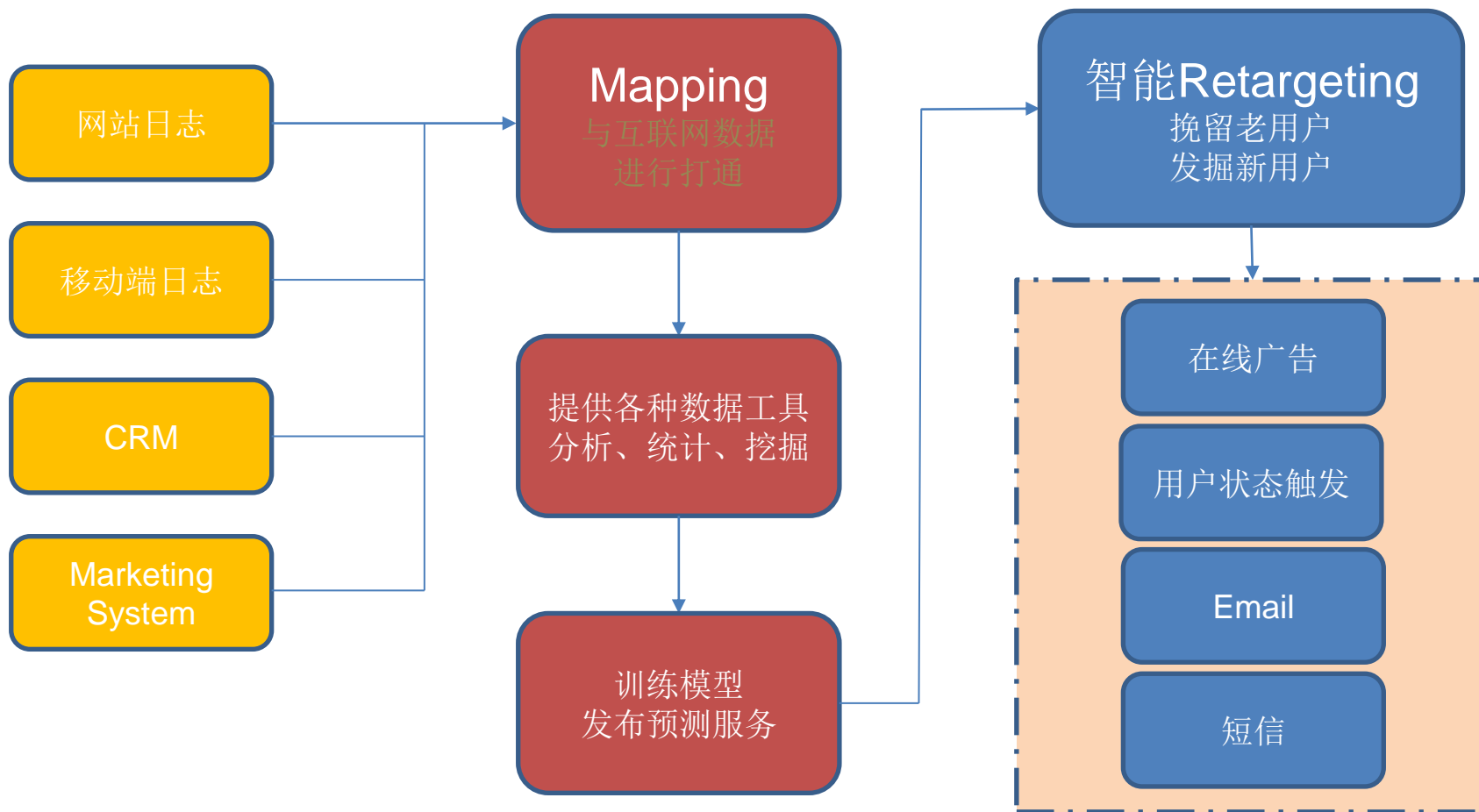
- 快速、低成本的实现
- 支持规模快速扩张的高效算法库
- AB Test和模型迭代机制

评价标准

- 推荐系统为例：覆盖率、置信度、差异性、采纳率、新颖性、隐私性、预测Auc、NDCG、收入波动、人工使用体验等指标
- 对整体系统的影响







用户的发现 -> 了解 -> 锁定 -> 状态触发 -> 推广行为

