

大规模机器学习技术

夏粉

2015年3月30日

Outline

- 广告背景
- 大数据机器学习
- 深度学习与CTR
- 总结展望

搜索广告： Search Ads

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

北京美食

百度一下

推荐：[用手机随时随地上百度](#)

北京美食 首选国内领先的吃喝玩乐信息平台-易吃易乐 [bj.echiele.com](#)  推广链接

北京美食首选国内领先的吃喝玩乐信息平台-易吃易乐,每天有上百万网

● [易吃易乐](#) ● [餐饮美食](#) ● [休闲娱乐](#) ● [美容美发](#)

[找北京美食?来DaoDao.com](#) [www.daodao.com](#) 

找北京美食?DaoDao.com为您提供210000条北京市旅游点评/攻略.

北京美食-大众点评网

根据合理的商区、地标和美食商户分类系统,为你提供北京83892家美食商户,并通过海量亲身消费者的点评聚合,以各种评分、星级的标准让你选择。

[www.dianping.com/beijing/f...](#) 2013-7-5 - 百度快照

北京美食攻略 [北京美食推荐](#) [美食街,小吃,指南](#)-驴妈妈旅游网

驴妈妈旅游网关于北京美食攻略,包含更多北京特色美食小吃(美食,餐饮,娱乐),【旅游预订】打折门票,周边酒店,自由行及跟团游信息,就在([www.lv mama.com](#))

[www.lv mama.com/travel/zhongguo_beiji...](#) 2013-6-29 - 百度快照

北京有什么特色美食? [百度知道](#)

13个回答 - 提问时间: 2011年12月25日

最佳答案: 1.烤鸭:在北京您要是想吃到便宜实惠的烤鸭,您可以去便宜坊、大鸭梨、安贞烤鸭店。当然您要是想吃最地道的烤鸭那就去和平门的全聚德。 2.涮羊肉:地...

[zhidao.baidu.com/question/3585625...](#) 2013-1-27 - 百度快照

北京美食 [百度百科](#)

北京美食guide是一款让你随时随地掌握北京美食信息的手机软件。北京美食拥有详尽的地图,十多种美食分类。

[基本信息](#) - [软件介绍](#) - [安装指南](#) - [分辨率](#) - [软件截图](#)

[baike.baidu.com/](#) 2013-07-03

在北京市搜索北京美食 [百度地图](#)



A. [辣尚瘾\(人大店\)](#) - (010)82650566

★★★★★ 1229条评论

北京美食

外文名: Guide

版本: V1.8.0

软件大小: 3941KB

来自[百度百科](#)>>

相关食物



[北京小吃](#)



[北京烤鸭](#)



[门钉肉饼](#)



[褡裢火烧](#)




[护国寺小吃](#)

推广链接

北京美食 [餐厅预定有折扣](#)


上咕嘟妈咪,方便轻松享优惠北京美食;

咕嘟妈咪,不让亲朋排队等.

[www.gudumami.cn](#) 


北京美食 [金鼎鱼香渔村欢迎..](#)

金鼎鱼香生态渔村,特色全鱼宴,灶台柴锅水库鱼,柴锅柴鸡,特色烧鸽

[www.myjdyx.com](#) 

北京美食 [刷雅酷卡 乐享无限..](#)

找北京美食,精选北京美食折扣优惠!吃喝玩乐尽在雅酷卡网!

[www.vacool.com](#) 

展示广告：Display Ads

1

玻璃家具的挑选。

玻璃家居，一定要注意玻璃的厚度，可以的情况下，最好选择是钢化玻璃的，可以安全很多，坚硬很多。因为钢化后，耐热，即使是破碎后，碎片也不易伤人。

家中有孩子有老人的特别要注意。



2

塑料家具的选择。

塑料家居，一般为椅子，桌子，层架等小物品。塑料制品，需要对材料进行分析，最好有说明材料的等级。

注意好塑料的软硬度，承载力，还有塑料制品需要避免风化，暴晒，否则容易缩短寿命。

快快选择自己喜爱的颜色吧。

春季喝什么茶养颜美容

如何工作更具效率 50

春季菠萝滋润吃法 25

肠道排毒按摩手法示范 69

刚毕业大学生租房攻略 35

春日简单的盘发教程 73

愚人节
玩出新花样!

[3月30日北京家居团购博览会!](#)

中国超大规模一站式家居建材装修采购展家居团购博览会, 鼎级家装设计, 家居团购一线品..

[jctg.com](#)

[北京 家居网.装房子,买家具...](#)

居然之家明码实价直降10%,"以旧换新"补贴5%!购物满10000还赠油卡灯具券,更多精彩.

[www.juran.cn](#)

[4月12日中国网友大型家具家..](#)

涵盖建材团购,家具团购,家具团购,地板团购,瓷砖团购,木门团购,橱柜团购,卫浴团购,洁..

[bj.wzcbd.com](#)

[板式家具设备选山东明美==山..](#)

选板式家具设备就到山东明美.板式家具设备一站式销售服务,全程专人负责调试,安排,培.

广告与点击率预估

广告核心问题

- 给定环境下，用户与广告的最佳匹配

流量变现

$$profit = PV * CTR * ACP$$

方法

- 依赖机器学习和大数据，做精准CTR预估



广告系统介绍

搜索
广告
系统

Google
AdWords

百度凤巢

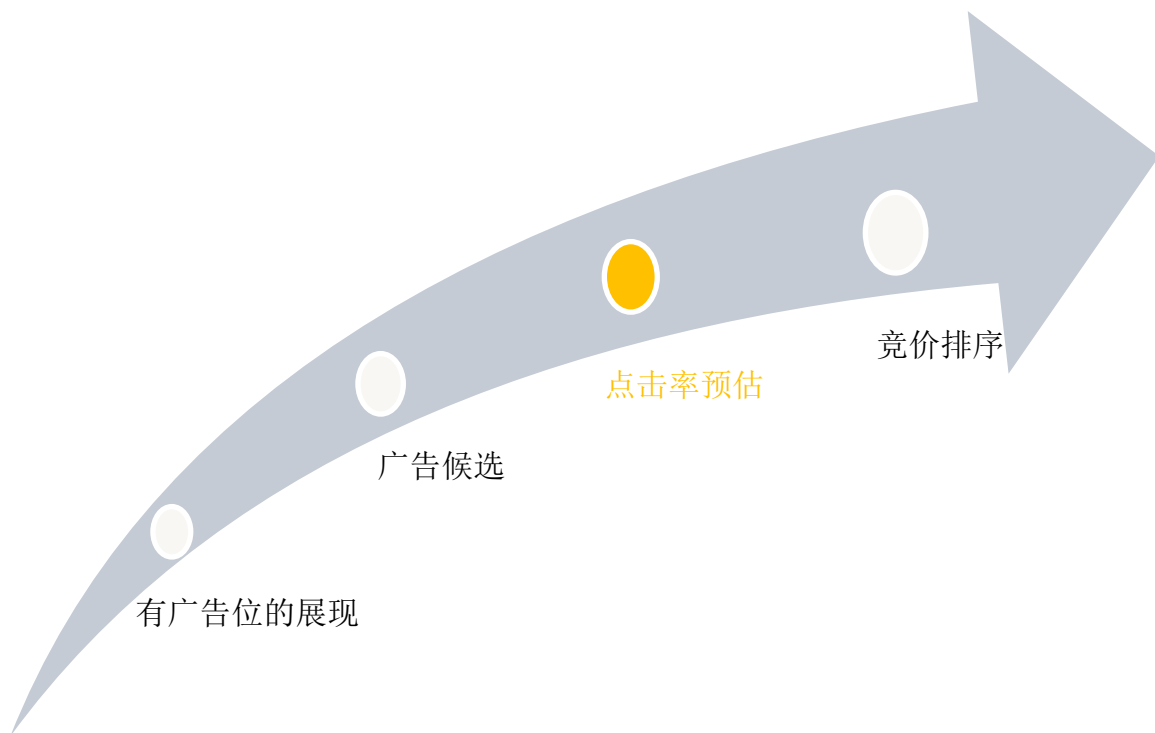
展示
广告
系统

Google
AdSense

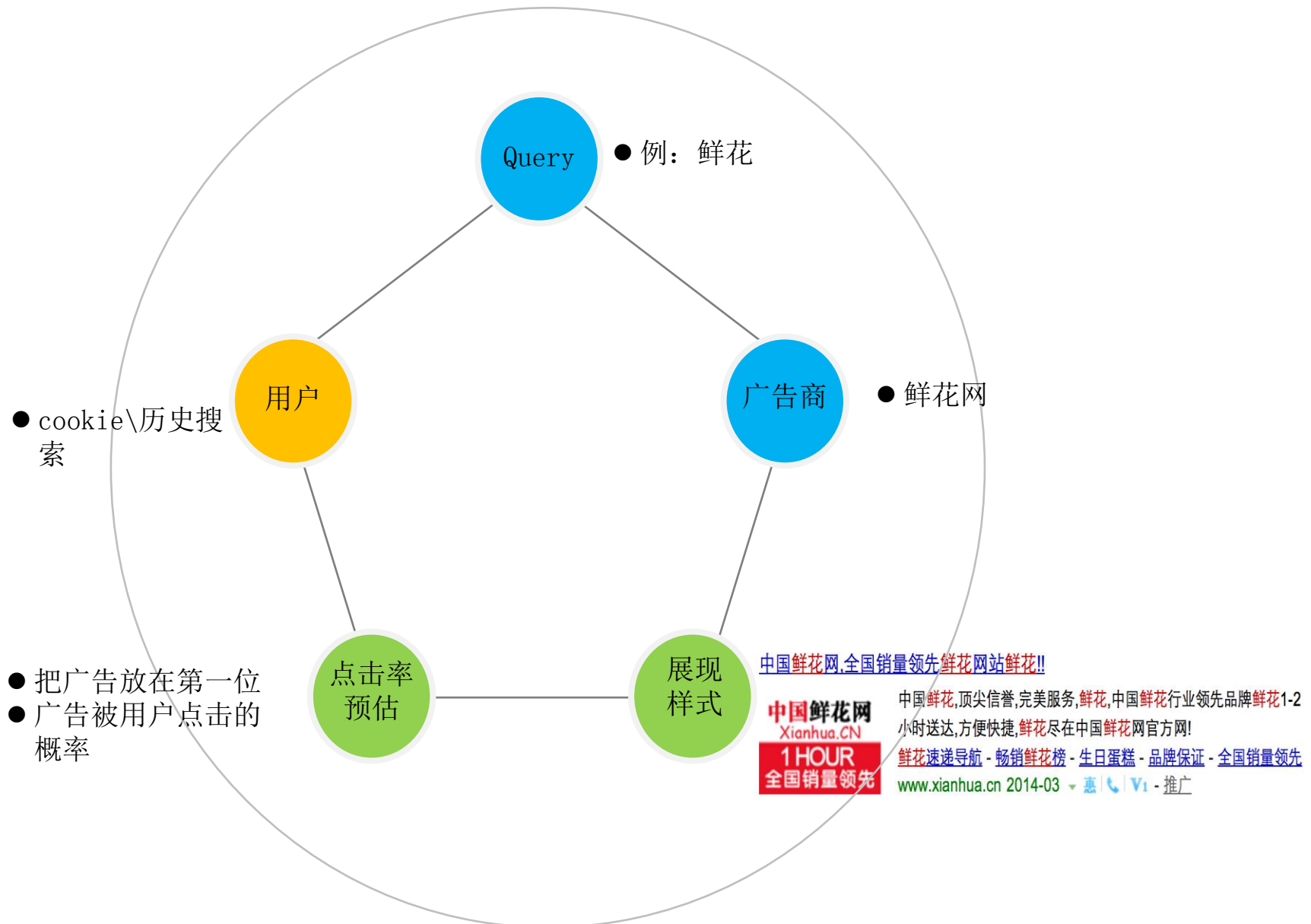
百度网盟

RTB
(Real Time
Bidding)

大致流程



点击率（CTR）预估问题



Outline

- 广告背景
- 大数据机器学习
- 深度学习与CTR
- 总结展望

大数据机器学习

大型分布式模型训练

核心技术

大规模线性Logistic
Regression模型

现状

点击率
(CTR)
预估

未来

问题规模：

- 数据存储和管理：**上万台机器**
- 数据量：**百亿到千亿级**
- 特征数：**百亿到千亿级**（稀疏离散值特征）

大规模深度学习模型



CTR预估的机器学习流程

特征生成

- 把广告展现成一个向量



概率模型

- 把向量变成点击率



线上预测

- 把模型用到新的广告展现上



模型训练

- 从历史数据学习模型参数

离散特征生成

- 假设 : 10000 查询; 1000 用户; 100 广告
- 查询(q): 1, 2, ..., 10000
- 用户(u): 1, 2, ..., 1000
- 广告(a): 1, 2, ..., 100
- 原始特征向量: q=1,u=2,ad=3

$$\left[\begin{array}{c} \left[\underbrace{1, 0, \dots, 0}_{10000\text{-dim q-vector}} \right] \left[\underbrace{0, 1, 0, \dots, 0}_{1000\text{-dim u-vector}} \right] \left[\underbrace{0, 0, 1, 0, \dots, 0}_{100\text{-dim ad-vector}} \right] \end{array} \right]$$

高阶特征生成

- **1st 阶:** 3 种单维度特征 q,u,ad

$$\left[\begin{array}{ccc} \underbrace{[1, 0, \dots, 0]}_{10000\text{-dim } q\text{-vector}} & \underbrace{[0, 1, 0, \dots, 0]}_{1000\text{-dim } u\text{-vector}} & \underbrace{[0, 0, 1, 0, \dots, 0]}_{100\text{-dim } ad\text{-vector}} \end{array} \right]$$

- **2nd 阶:** q*u 查询和用户特征组合

$$\left[\underbrace{0, 0, \dots, 0, 1, 0, \dots, 0}_{10000 \times 1000 \text{ dim } q \times u \text{ vecotr}} \right]$$

离散特征影响



维数约简

- 离散到离散: **Hashing**

$$\left[\underbrace{0, 0, \dots, 0, 1, 0, \dots, 0}_{10000 \times 1000 \text{ dim } q \times u \text{ vecotr}} \right] \rightarrow \left[\begin{array}{c} \underbrace{0, \dots, 1, \dots}_{1000 \text{dim hash table}} \end{array} \right] \left[\begin{array}{c} \underbrace{0, \dots, 1, \dots}_{1000 \text{dim hash table}} \end{array} \right]$$

- 离散到统计: **statistics**

$$\left[\underbrace{0, 0, \dots, 0, 1, 0, \dots, 0}_{10000 \times 1000 \text{ dim } q \times u \text{ vecotr}} \right] \rightarrow \left[\begin{array}{cc} \underbrace{0.4}_{\text{historic CTR}} & \underbrace{103}_{\text{historic show}} \dots \end{array} \right]$$

- 更多先进技术?

模型: Logistic Regression

- 模型假设

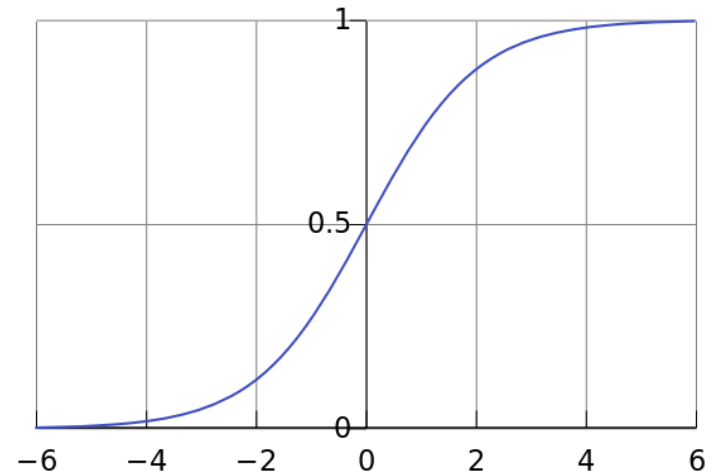
- 输入向量 x 、输出点击率 ctr、模型参数 w

$$\text{ctr} = \frac{1}{1 + \exp(-w^\top x)}$$

- 模型训练

- 训练数据
 - x : 特征向量
 - y : $\{-1, +1\}$, -1: 未点检, +1: 点检
- 求解优化问题: $(x_1, y_1), \dots, (x_n, y_n)$

$$\min_w \sum_{i=1}^n \ln(1 + \exp(-w^\top x_i y_i))$$



模型: Logistic Regression

- 正则化

- 减少模型大小

$$\min_w \sum_{i=1}^n \ln(1 + \exp(-w^\top x_i y_i)) + C \|w\|_1, \quad \|w\|_1 = \sum_{j=1}^d |w_j| \quad w = [w_1, \dots, w_d]$$

- 求解算法

$$w_t \leftarrow w_{t-1} - \eta_t d_t, \quad d_t: \text{梯度方向(梯度或者牛顿方向)}$$

- LBFGS: 使用 1st 阶梯度近似Hessian矩阵
- 坐标梯度下降: 使用单维特征梯度
- 随机梯度下降(SGD): 使用单个样本梯度

分布式计算架构

数据并行

- 每台机器存储所有参数
- 每台机器存储部分数据

模型并行

- 每台机器存储所有数据
- 每台机器存储部分参数

数据&模型并行

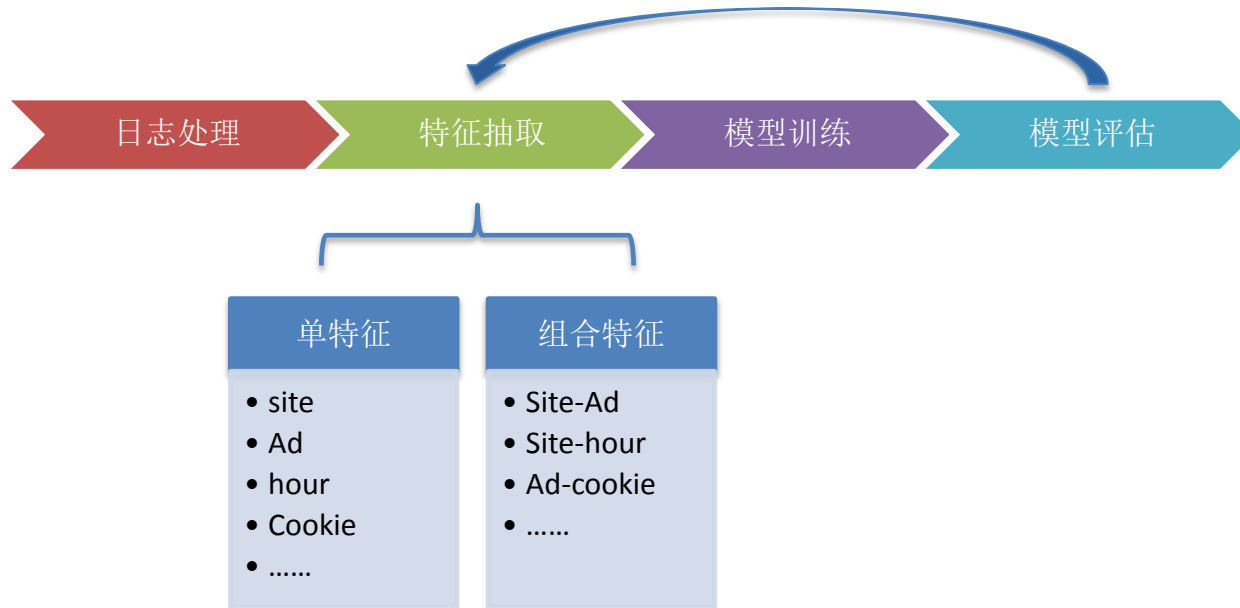
- 每台机器存储部分数据
- 每台机器存储部分参数

Outline

- 广告背景
- 大数据机器学习
- 深度学习与CTR
- 总结展望

人工特征工程

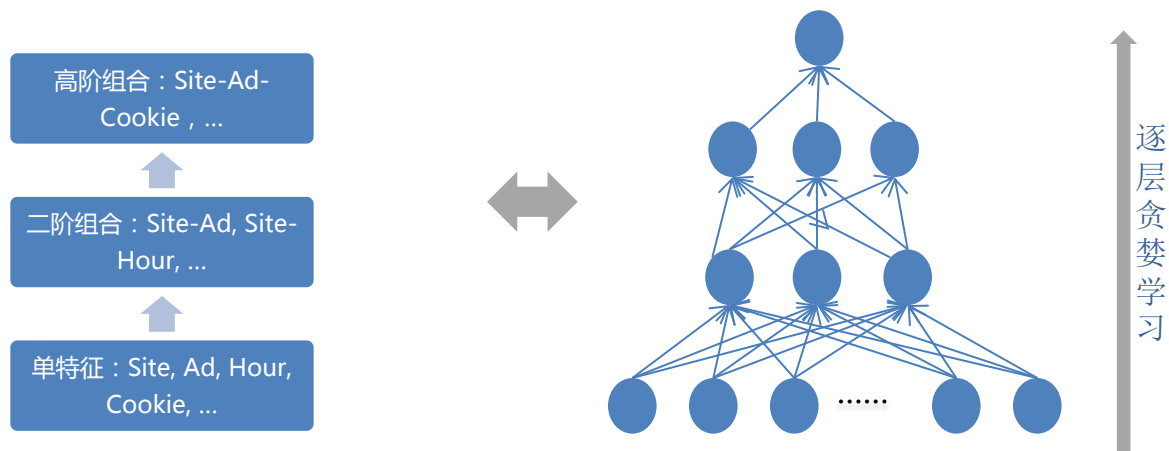
- CTR 预估模型



- 需要加入组合特征来提升LR的表达力
- 假设有N个单特征类，组合特征类： $2^{N-1} \approx C_N^1 + C_N^2 + C_N^3 + \dots + C_N^{N-1}$
- 人工挖掘，先验知识给出候选特征集合，依次加入模型训练
 - 耗时！耗力！

深度特征学习技术

- 特征学习
 - 深度学习在语音、图像上取得突破性进展
 - 广告数据特征维数非常高（单特征百亿），尚无大规模稀疏特征学习算法
- **DANOVA**: 首个直接应用于大规模稀疏特征的深度特征学习算法



- 上线效果
 - 特征挖掘效率提升上千倍
 - CTR, CPM显著增长

Outline

- 广告背景
- 大数据机器学习
- 深度学习与CTR
- 总结展望

大数据点击率预测技术发展

一代：人工规则

二代：40%+ 点击率提升

简单特征
小规模非线性模型

Yahoo, Facebook,
Microsoft, etc.

三代：10%+ 点击率提升

高维特征
大规模线性模型
模型实时更新

Google, Baidu, etc.

流式计算
模型在线更新



百度新一代模型
20% + 点击率提升



大规模



复杂模型



实时更新

技术全球领先

Thanks !