# 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

# 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。







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 V2

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 V2

4月12日中国网众大型家具家..
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 V2

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

流量变现

方法

- 给定环境下，用户与广告的最佳匹配

$$profit = PV * CTR * ACP$$


- 依赖机器学习和大数据，做精准CTR预估

# 广告系统介绍

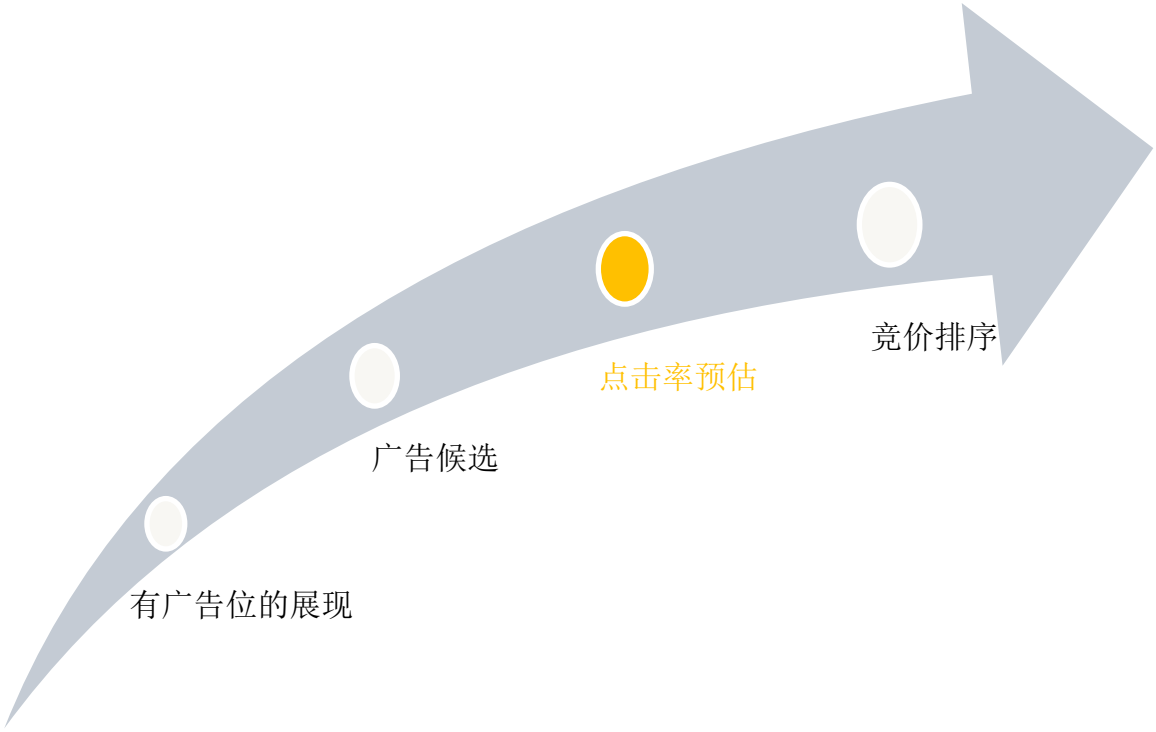
搜索广告系统

- Google AdWords
- 百度凤巢

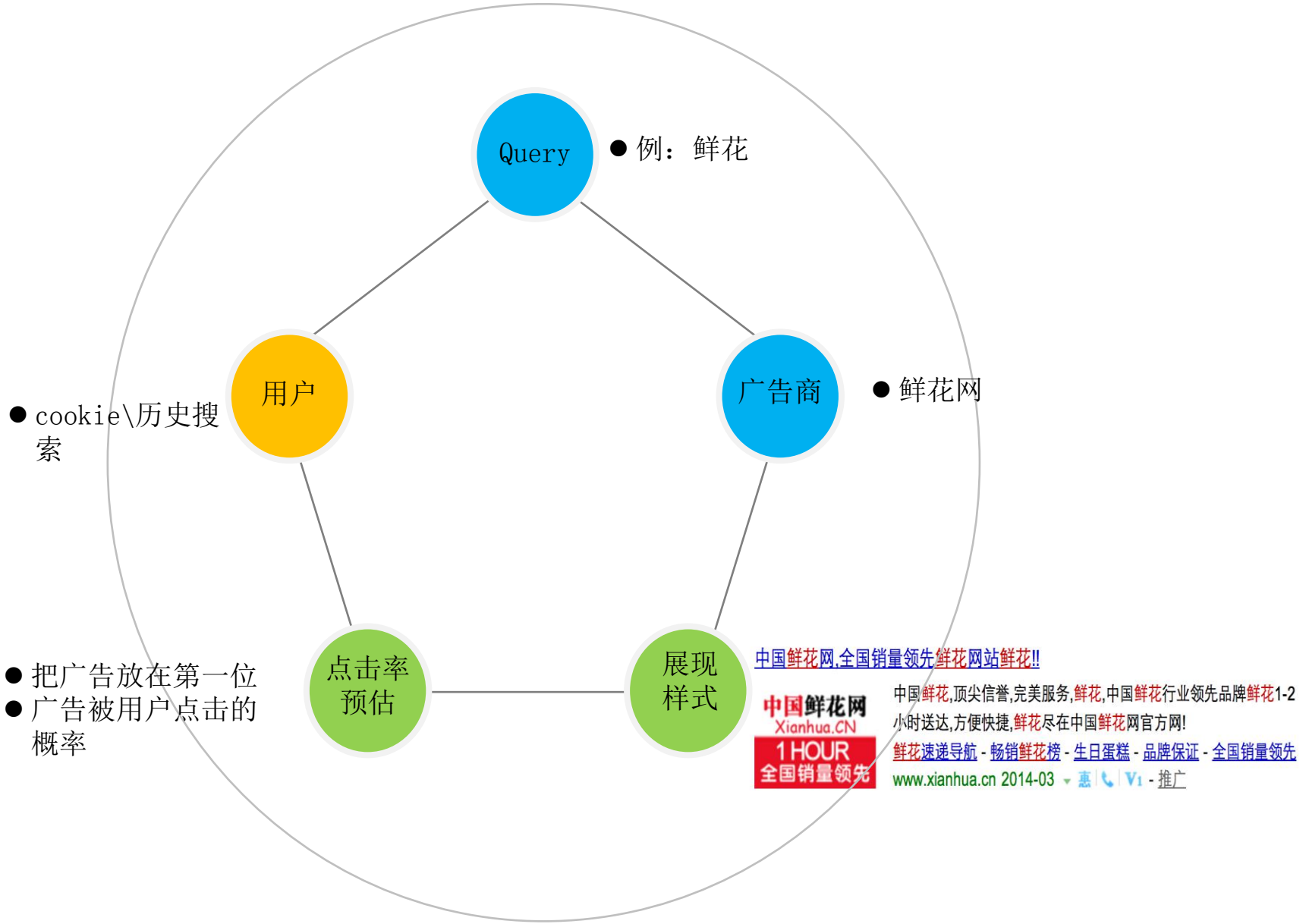
展示广告系统

- Google AdSense
- 百度网盟
- RTB (Real Time Bidding)

大致流程



# 点击率（CTR）预估问题


Query
例：鲜花
用户
cookie\历史搜索
广告商
鲜花网
点击率预估
把广告放在第一位
广告被用户点击的概率
展现样式
中国鲜花网,全国销量领先鲜花网站鲜花!!
中国鲜花网
Xianhua.CN
1 HOUR
全国销量领先
中国鲜花,顶尖信誉,完美服务,鲜花,中国鲜花行业领先品牌鲜花1-2小时送达,方便快捷,鲜花尽在中国鲜花网官方网!
鲜花速递导航 - 畅销鲜花榜 - 生日蛋糕 - 品牌保证 - 全国销量领先
www.xianhua.cn 2014-03 - 推广

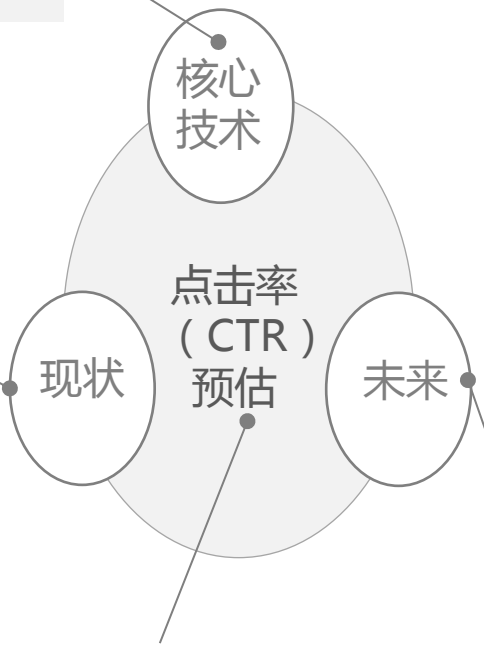
# Outline

- 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

大规模线性Logistic Regression模型




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

# 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

$$\Big[\ [\ \underbrace{1,0,\ldots,0}_{\text{10000-dim q-vector}}\ ][\ \underbrace{0,1,0,\ldots,0}_{\text{1000-dim u-vector}}\ ][\underbrace{0,0,1,0,\ldots,0}_{\text{100-dim ad-vector}}]\ \Big]$$

# 高阶特征生成

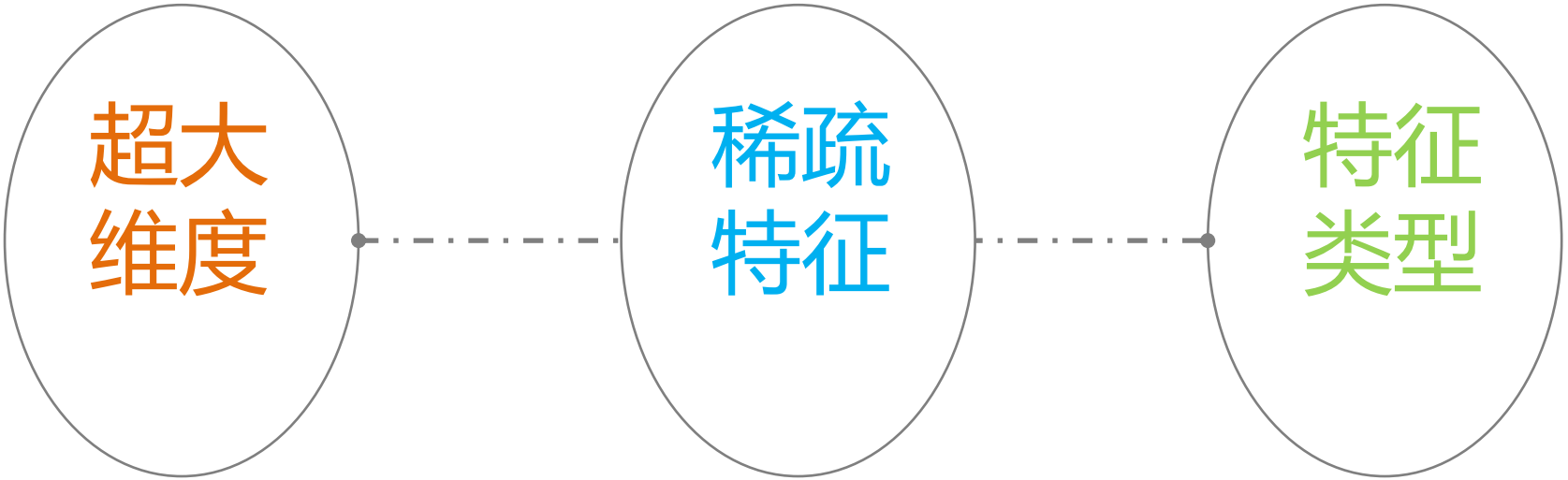
- **1st 阶：** 3 种单维度特征 q,u,ad

$$\Big[\ \underbrace{[1,0,\ldots,0]}_{\text{10000-dim q-vector}}\ \underbrace{[0,1,0,\ldots,0]}_{\text{1000-dim u-vector}}\ \underbrace{[0,0,1,0,\ldots,0]}_{\text{100-dim ad-vector}}\ \Big]$$

- **2nd 阶：** q*u 查询和用户特征组合

$$\Big[\ \underbrace{0,0,\ldots,0,1,0,\ldots,0}_{10000\times1000\ \text{dim}\ q\times u\ \text{vecotr}}\ \Big]$$

# 离散特征影响




百亿到千亿量级

每个样本只有百个非零元

百个特征类型

# 维数约简

- 离散到离散：**Hashing**

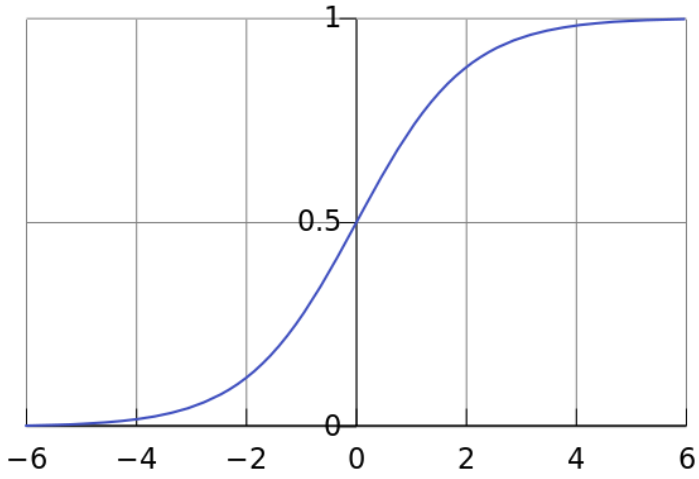
$$[\ \underbrace{0,0,\ldots,0,1,0,\ldots,0}_{10000\times 1000\text{ dim }q\times u\text{ vecotr}}\ ]\rightarrow\left[[\ \underbrace{0,\ldots,1,\ldots}_{1000\text{dim hash table}}\ ][\ \underbrace{0,\ldots,1,\ldots}_{1000\text{ dim hash table}}\ ]\right]$$

- 离散到统计：**statistics**

$$[\ \underbrace{0,0,\ldots,0,1,0,\ldots,0}_{10000\times 1000\text{ dim }q\times u\text{ vecotr}}\ ]\rightarrow\left[\underbrace{0.4}_{\text{historic CTR}}\quad\underbrace{103}_{\text{historic show}}\quad\cdots\right]$$

- 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

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

- **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：$(x_1, y_1), \ldots, (x_n, y_n)$

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

# 模型: Logistic Regression

- **正则化**
  - 减少模型大小

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

- **求解算法**

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

  - 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
- 每台数据存储部分数据

## 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

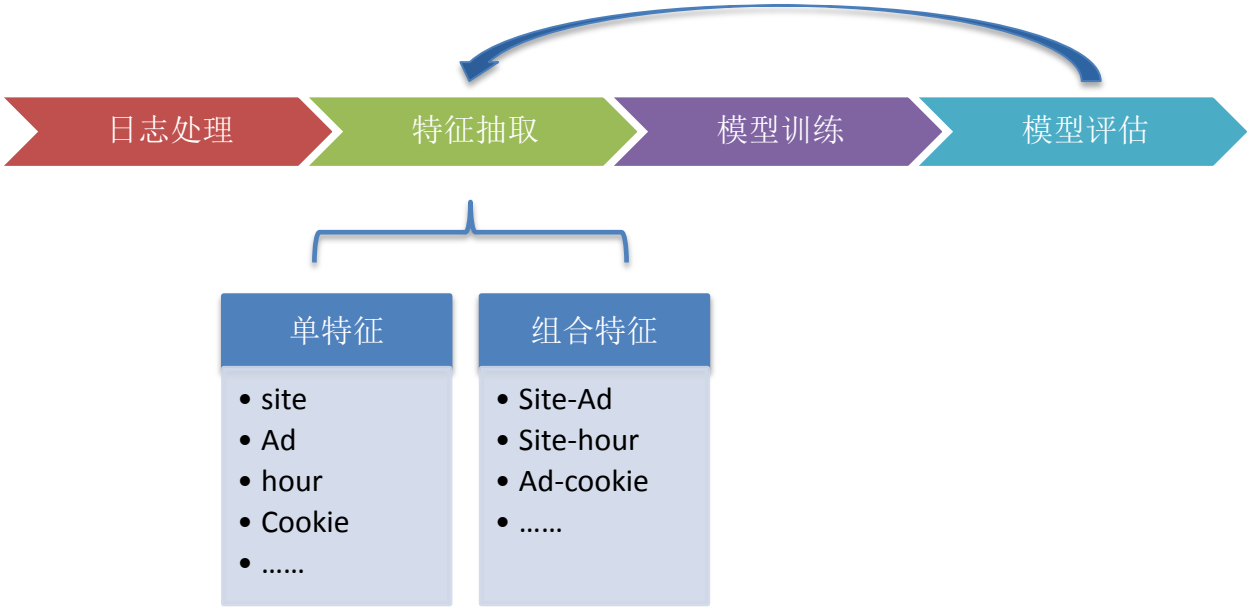
- 每台机器存储部分数据
- 每台机器存储部分参数

# 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

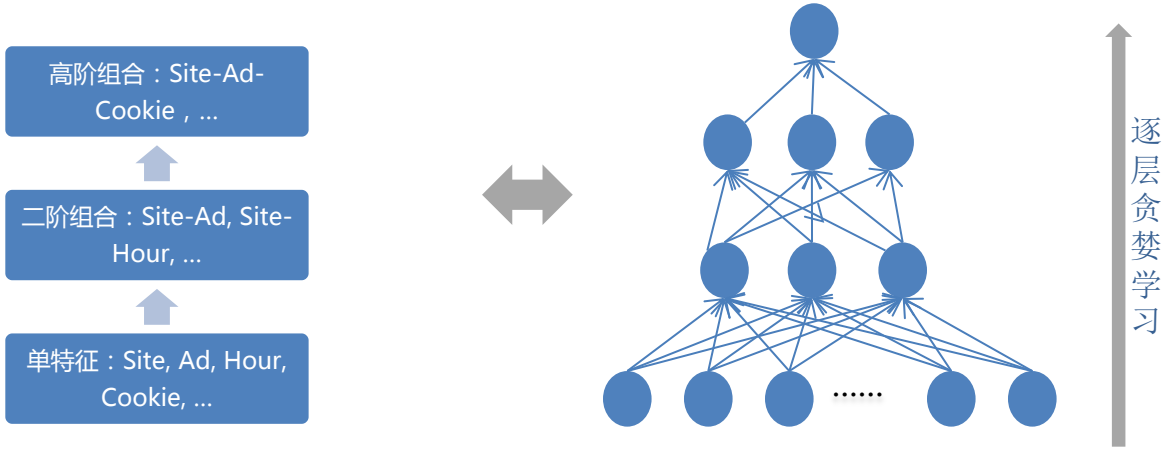
- **CTR 预估模型**




- 需要加入组合特征来提升LR的表达能力
- 假设有N个单特征类，组合特征类： $2^{N-1} \approx C_N^1 + C_N^2 + C_N^3 + \ldots + C_N^{N-1}$
- 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<br>小规模非线性模型 | Yahoo, Facebook, Microsoft, etc.

三代: 10%+ 点击率提升 | 高维特征<br>大规模线性模型<br>模型实时更新 | Google, Baidu, etc.

流式计算<br>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

Google,<br>Baidu,<br>etc.

# Thanks !