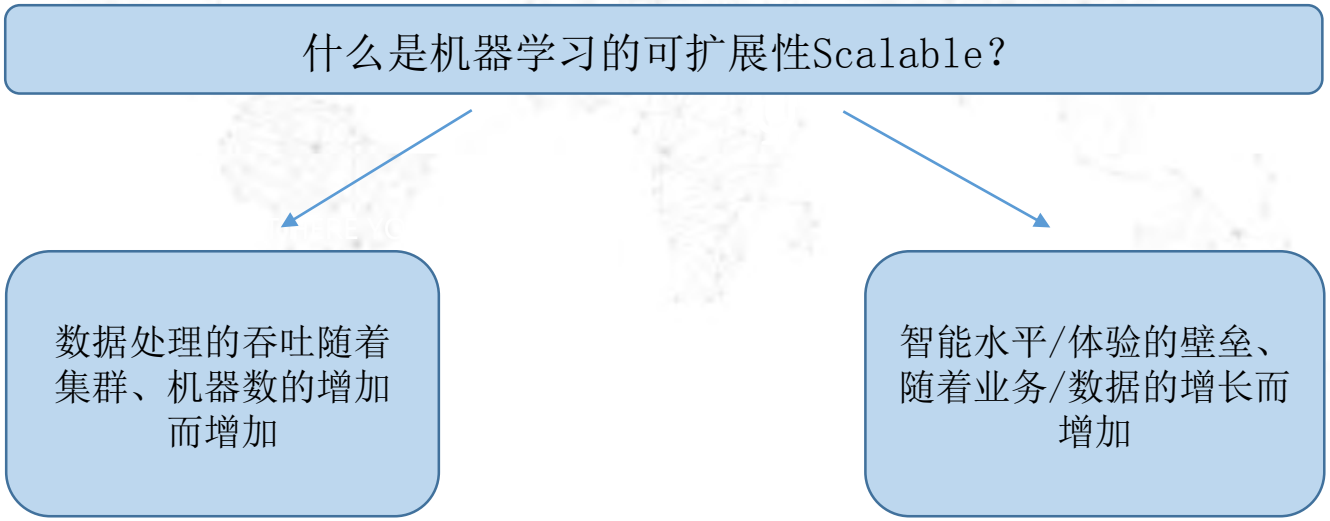
人工智能工业应用痛点及解决思路

第四范式 陈雨强

2017年 8月5日

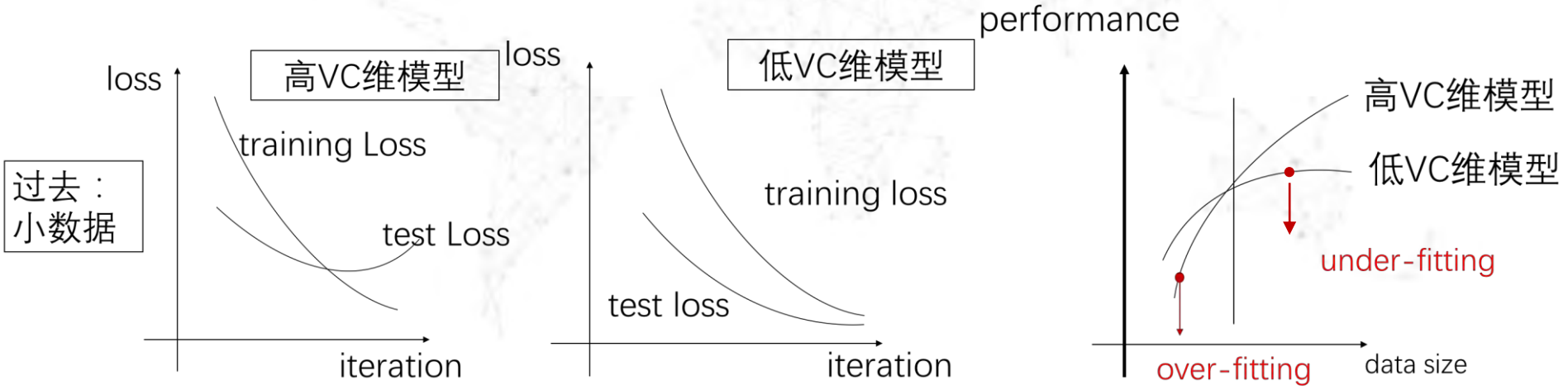
可扩展的机器学习系统

- ✓ 人工智能的兴盛是数据量变大、机器性能提升、并行计算发展的结果
- ✓ Scalable ML System ≠ Scalable System



工业大数据需要高VC维模型

- ✓ VC维（Vapnik-Chervonenkis Dimension）又称VC理论；
- ✓ VC维反映了函数集的学习能力，VC维越大则模型或函数越复杂，学习能力就越强；
- ✓ 模型一定要与待解决的问题相匹配，如果模型过于简单，而问题本身的复杂度很高，就无法得到预期的精度



可扩展的机器学习系统

机器学习＝数据＋特征＋模型

工业追求极高的VC维度

如何沿着模型走？

- ✓ 学术界主导（ICML，NIPS，ICLR）
  - 非线性的三把宝剑：Kernel，Boosting，Neural Network
  - 模型大部分单机可加载
  - 解决数据分布式问题，以及降低系统overhead

- ✓ 工业界针对应用定制模型
  - 基于思考或者观测得到的假设
  - 加入新的模型、结构，以加入更多参数
  - 典型案例：伽利略

如何沿着模型走？

- ✓ 工业界主导（KDD，WWW）
  - 模型相对简单粗暴
  - 分布式，工程挑战大
  - 高效并行并保证快速收敛
- ✓ 工业界中一般针对应用定制特征
  - 为什么有那么多特征，怎么产生这些特征
  - 如何理解这些特征
  - 人工智能爱因斯坦

没有免费的午餐定理：不存在万能模型

"We show that all algorithms that search for an extremum of a cost function perform exactly the same, when averaged over all possible cost functions. In particular, if algorithm A outperforms algorithm B on some cost functions, then loosely speaking there must exist exactly as many other functions where B outperforms A"

-- by Wolpert and Macready (1995) in No Free Lunch Theorem

所有的机器学习模型都是一个偏置

- 更多的模型假设 ⇒ 更少的数据
- 更简单的模型假设 ⇒ 更多的数据支持与特征刻画

工业界机器学习中并没有免费的午餐，要做出对业务问题合适的选择

然AI还远未普及

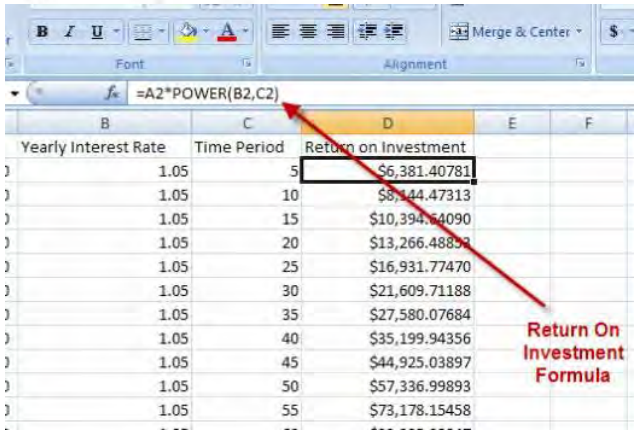
与Hadoop相比

工业界应用机器学习的难题

工业界应用机器学习的难题

- ✓ 需要AI应用平台
  - Tenserflow，Mxnet，Caffe等工具日趋丰富
  - 但是，足够了么？

- ✓ 为什么人工智能还没有真的大规模应用到每个企业
  - Hadoop为什么用的人多？
  - 先驱知识要求太多
  - 能做AI的还是研究/应用机器学习科学家
  - 核心机器学习算法平台只降低了一部分门槛
  - 更大的应用基础：降门槛 > 算法效果



如何解决特征工程

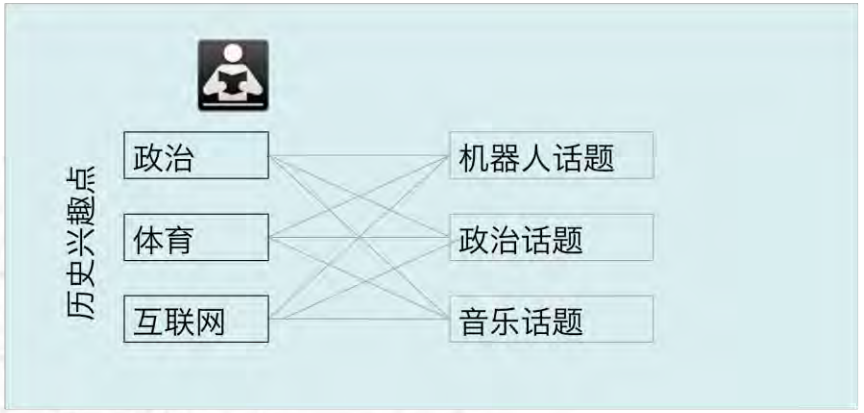
- ✓ 特征工程在工业界是巨大的难关
  - 什么是特征工程？现在的平台已经足够了吗？
  - 需要对机器学习与业务都非常理解
  - 不同的算法，要使用不同的特征工程达到同一个目标

- ✓ 以新闻推荐为例
  - 一阶特征：每个用户直接喜欢什么
  - 二阶特征：用户的扩展兴趣（喜欢大数据的人，可能对机器学习也感兴趣）
  - 不同模型如何添加？

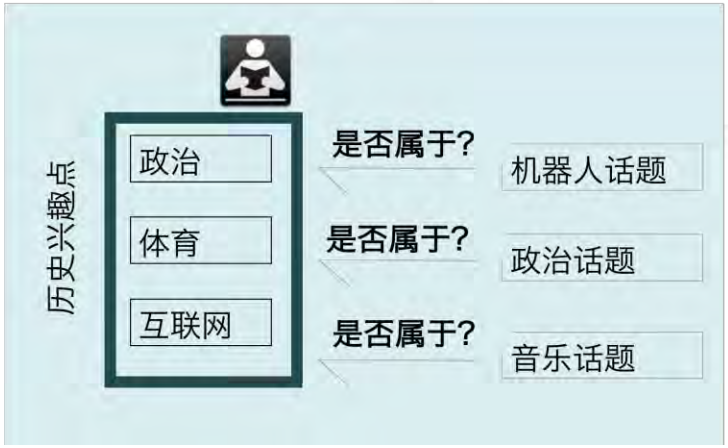
- Feature Transformers
  - Tokenizer
  - StopWordsRemover
  - *n*-gram
  - Binarizer
  - PCA
  - PolynomialExpansion
  - Discrete Cosine Transform (DCT)
  - StringIndexer
  - IndexToString
  - OneHotEncoder
  - VectorIndexer
  - Interaction
  - Normalizer
  - StandardScaler
  - MinMaxScaler
  - MaxAbsScaler
  - Bucketizer
  - ElementwiseProduct
  - SQLTransformer
  - VectorAssembler
  - QuantileDiscretizer
  - Imputer

线性模型，学习一阶特征

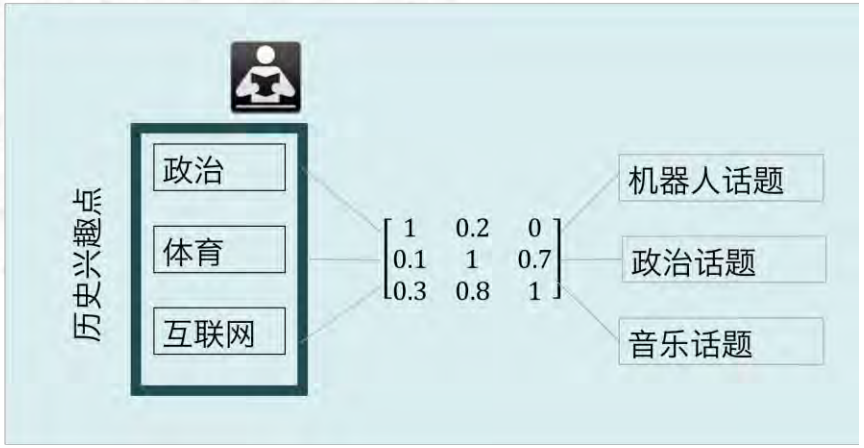
线性模型，学习二阶特征



非线性宏观特征模型，学习一阶特征



非线性宏观特征模型，学习二阶特征



如何解决特征工程

- ✓ 特征工程在工业界是巨大的难关
  - 需要对机器学习与业务都非常理解
  - 不同的算法，要使用不同的特征工程达到同一个目标

	一阶特征（用户直接喜欢什么）	二阶特征（用户的扩展兴趣）
线性模型	用户ID -- 待推荐新闻话题 （不用统计，形成高维特征）	用户历史兴趣点 -- 待推荐新闻话题 （需要统计，形成高维特征）
非线性模型（宏观特征）	用户历史兴趣点是否包含待推荐新闻话题（需要统计，形成单个特征）	用户历史兴趣点，兴趣相关性矩阵，待推荐新闻话题（需要统计，单特征）
非线性模型（精细特征）	用户ID，待推荐新闻话题 （不用统计，形成高维特征）	用户历史兴趣点，待推荐新闻话题 （需要统计，形成高维特征）

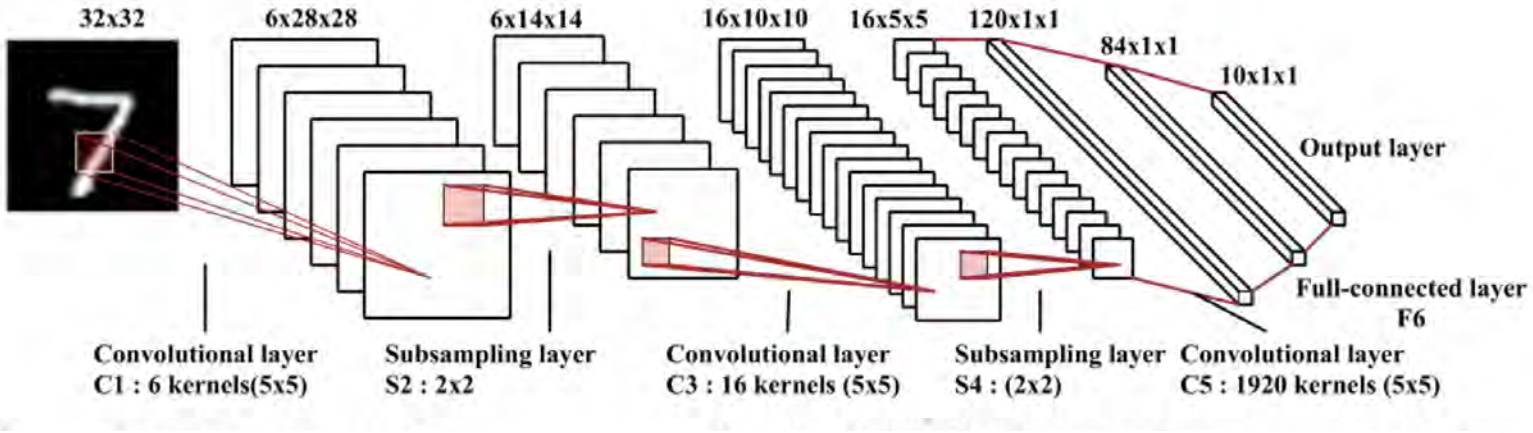
如何解决特征工程

- ✓ 特征工程是非常大的难题
  - 需要对机器学习与业务都非常理解
  - 不同的算法，同样的特征，获得效果不同

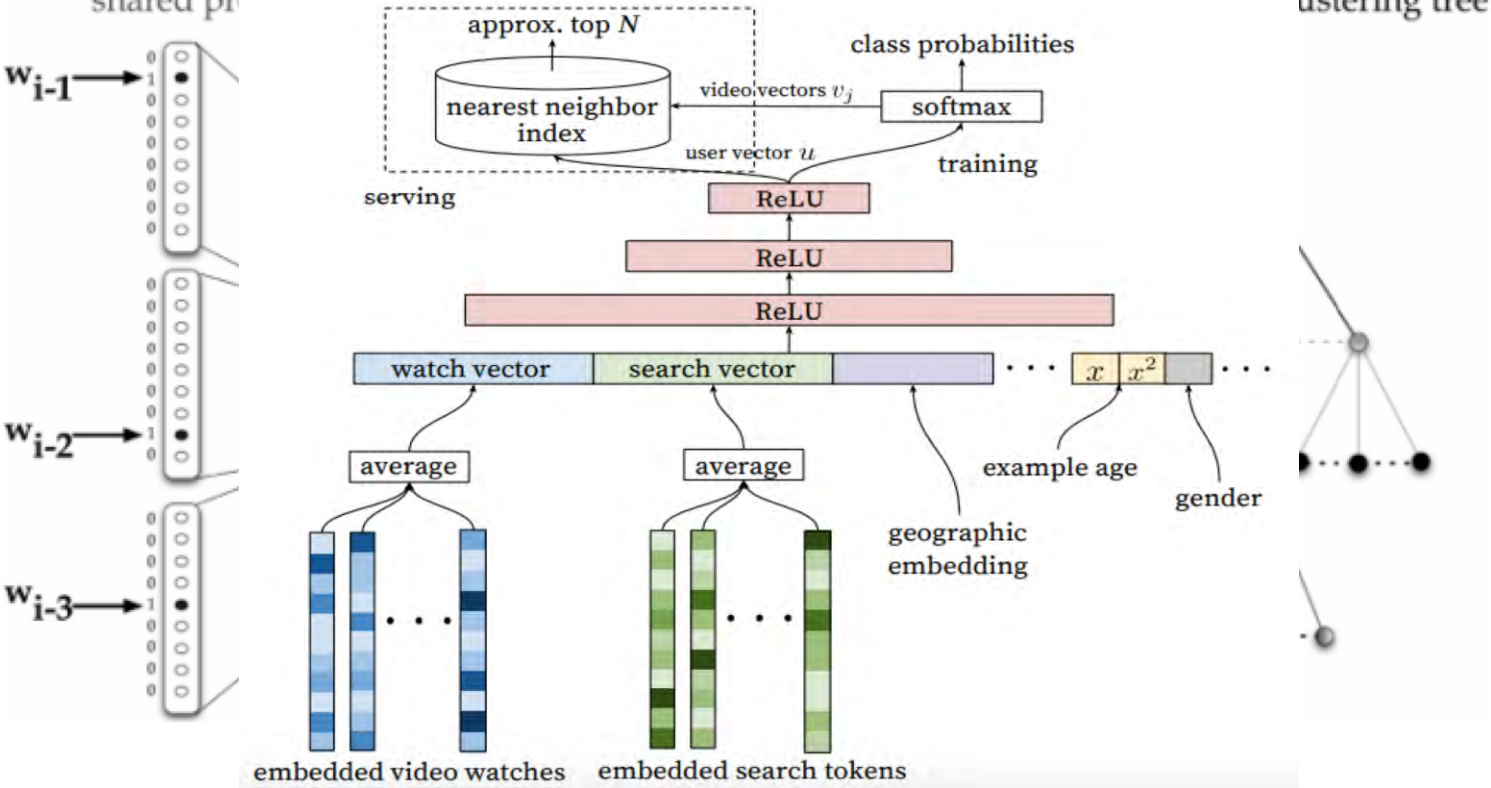
- ✓ 如何进行自动的特征工程
  - 隐式特征组合（NN，FM）
  - 半显式显示特征组合（GBDT）
  - 显式特征组合（特征叉乘）

隐式特征组合

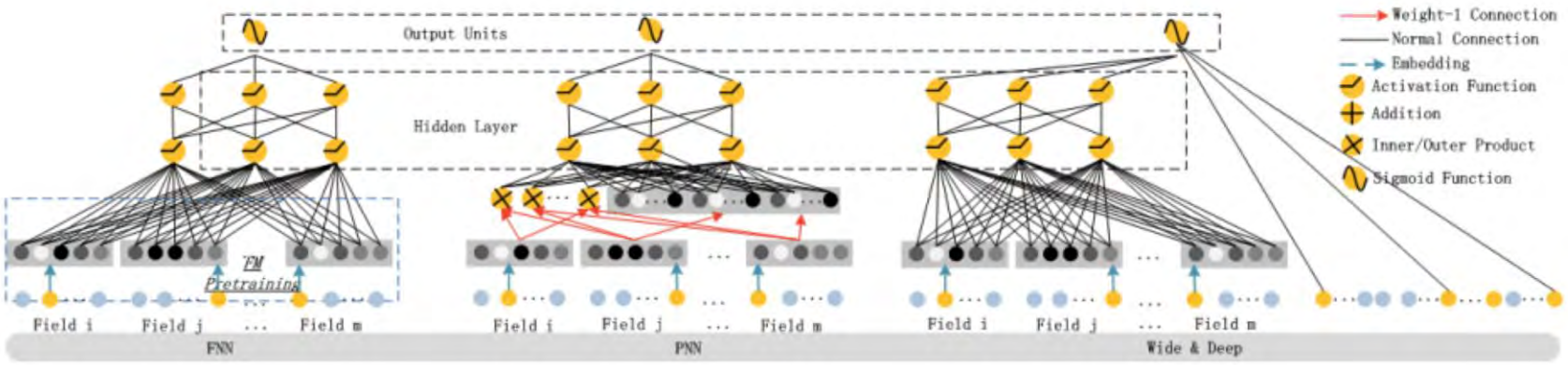
- ✓ 主要特点
  - 对连续值特征天然友好
  - 最大的成功：语音图像
  - 高级离散变量处理相对更复杂
  - 隐式组合，基本无可解释性
- ✓ 对离散特征需要Large Scale Embedding
  - Embedding NN
  - FM, FNN, PNN
  - DeepFM



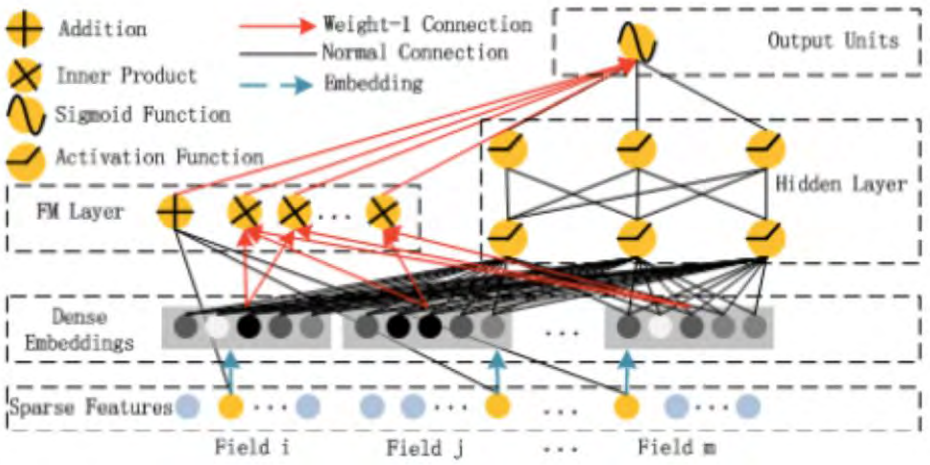
shared pr… …ustering tree



隐式特征组合



- ✓ 对离散特征需要Large Scale Embedding
  - Embedding NN
  - FM, FNN, PNN
  - DeepFM
  - 第四范式DSN



半隐式特征组合

- ✓ 主要是森林类算法
  - 为什么是“半隐式”
  - 看起来可以解释，实际上并不可解释
  - 看起来在做特征组合，实际上是层次贪心的副产物
- ✓ 主要特点
  - 理解容易，相对鲁棒，效果优秀
  - Off-the-shelf
  - 离散特征非常难解，无现有方案
- ✓ 第四范式HE-TreeNet，GBM
  - 解决大规模离散特征的树模型
  - 研发基于Embedding, Ensembling, Stacking的系列树算法

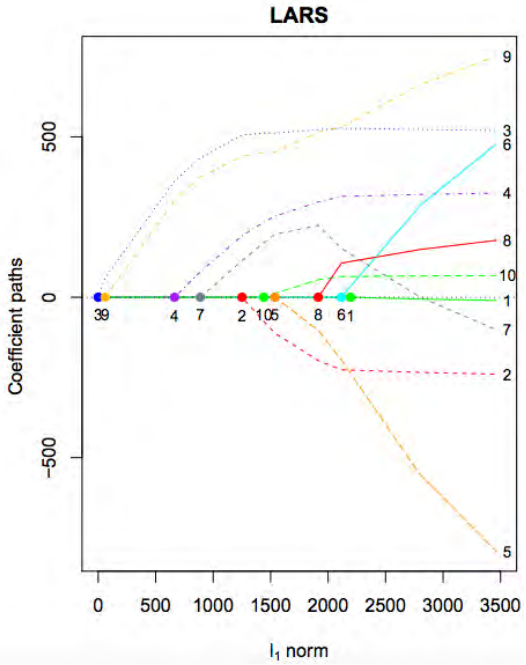
显式特征组合：问题

- ✓ 主要基于贪心与搜索
  - 正则化
  - Beam Search，MCTS
  - 遗传算法，模拟退火
- ✓ 问题特别的难
  - 围棋的状态空间 $< 3^{19\times19}$；而n个特征，选m个特征，限制最大k阶组合，状态空间为$C^{m}_{\Sigma_{i=2}^{k} C_n^i}$
  - 难以组合连续值特征？
- ✓ 显式特征组合优势
  - 可解释性：提供深度业务洞察
  - 可叠加性：增强所有机器学习算法

显式特征组合：现状

- ✓ State-of-art
  - Online Boosting Feature Selection: 单特征Weak Learner基于Adaboost的选择
  - Online Regularization：基于Lasso对梯度、权重截断

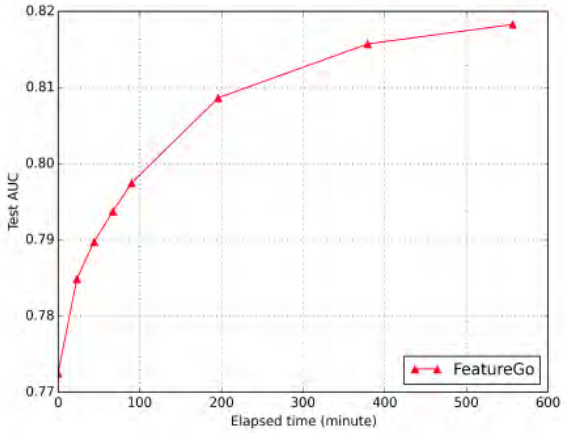
- ✓ 现有算法的问题
  - 并非为n选m个k阶以下特征设计
  - 多为副产物，对信息损失的比较大
  - 二阶组合为主，基本无法高阶特征组合

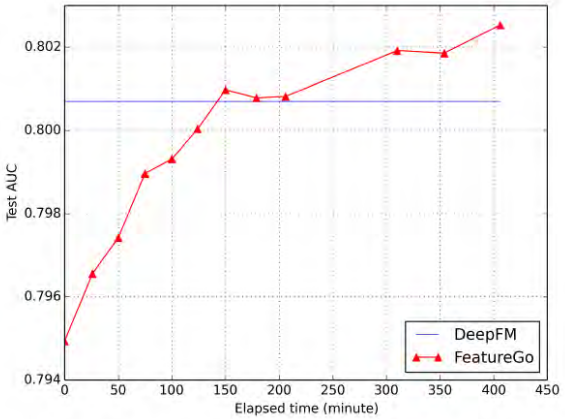


显式特征组合：FeatureGO

扫描二维码
第四范式·先知
FeatureGo, GBM, DSN触手可得！

- ✓ 第四范式FeatureGO算法
  - 基于MCTS，对特征组合状态进行估计
  - 调优的搜索剪枝技术
  - 利用LFC算法解决连续值特征组合问题
  - 组合特征可高达6阶





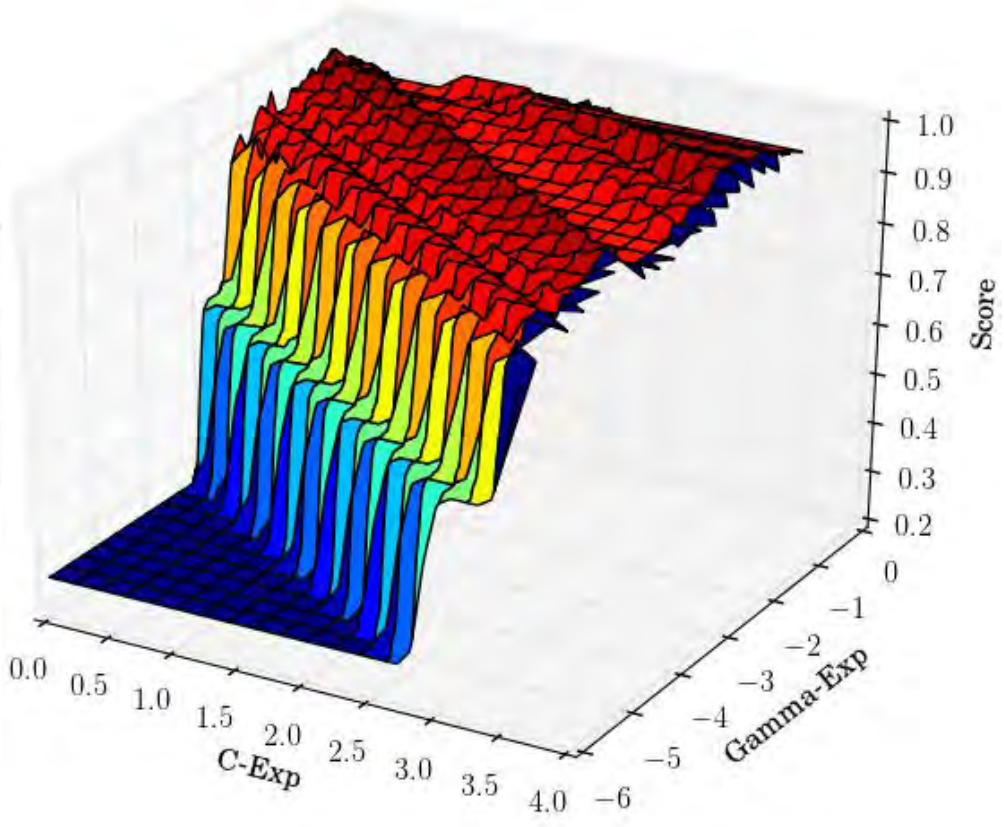
	Criteo	
	AUC	LogLoss
LR	0.7686	0.47762
FM	0.7892	0.46077
FNN	0.7963	0.45738
IPNN	0.7972	0.45323
OPNN	0.7982	0.45256
PNN*	0.7987	0.45214
LR & DNN	0.7981	0.46772
FM & DNN	0.785	0.45382
DeepFM	0.8007	0.45083
FeatureGO	0.8183	0.4501

算法背后的优化

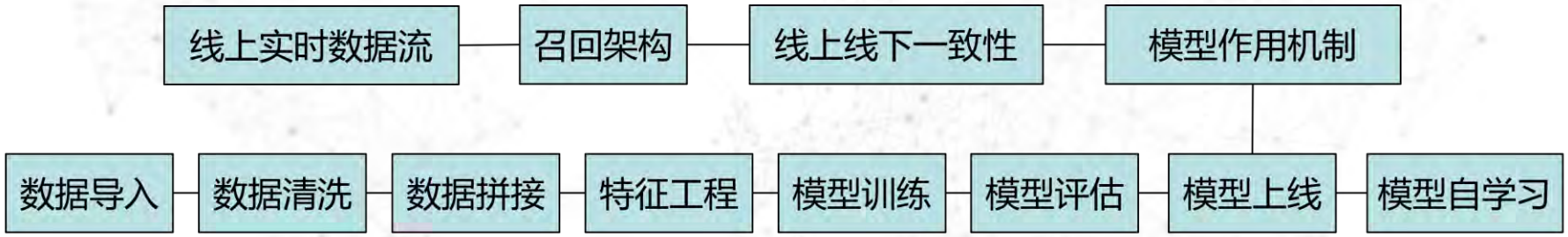
- ✓ Boosting
- ✓ Fully/Partially Corrective Learning
- ✓ Cross Parameter-server Sharing（CPS）
- ✓ 计算能力也是人工智能的一部分
  - 计算能力是新的性感，智能同样来自于计算
  - Google立出了榜样
  - 在第四范式，架构、工程优化与算法并重

如何解决调参

- ✓ 基于搜索的调参
  - Grid Search
  - Random Search
  - SMBO
- ✓ 如何做的更快更好
  - CPS
  - Dynamic Graph



AutoML：AI for Everyone



- 自动拼表：Domestic Knowledge Graph
- 模型可解释：Twice Learning
- 自动线上优化：强化学习
- …….

“第四范式 先知”试用版 免费对外开放


THANKS