十年架构 成长之路

SACC 第十届中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2018

2018年10月17-10月21日 北京海淀永泰福朋喜来登酒店

构建企业级机器学习平台

猎聘大数据研究院 单艺

自我简介

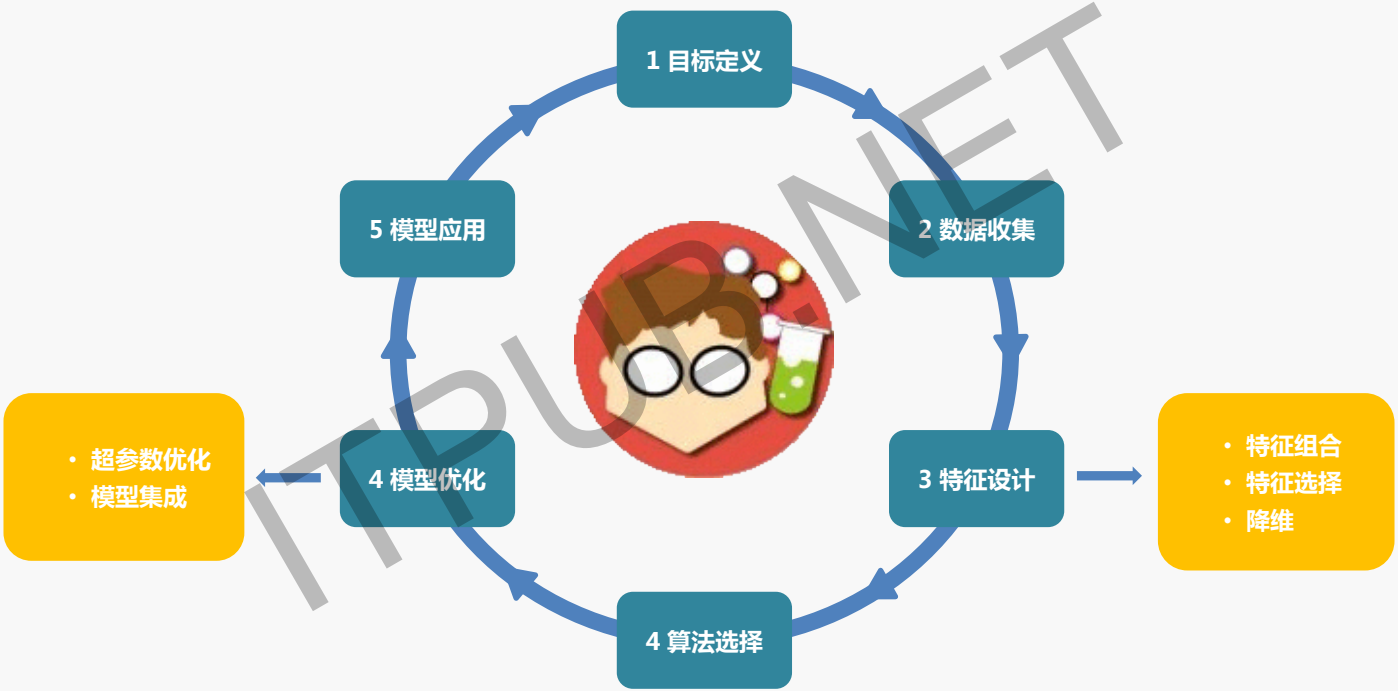
- 现任猎聘首席数据科学家，负责人工智能和大数据研发
- 曾任职于美国硅谷的Altera、Yahoo！和奥美广告
- 专注于机器学习、推荐系统、自然语言处理和大数据
- 毕业于清华大学和美国University of Arizona

机器学习应用场景

更多幕后的应用

- 用户和职位画像：职能分类、用户分级、能力模型
- 精准营销：定向产品推广、用户求职行为预测
- 平台运营：自动化订单分配、HR行为预测
- 销售自动化：客户分类分级、商机预测、商机分配

机器学习应用开发流程

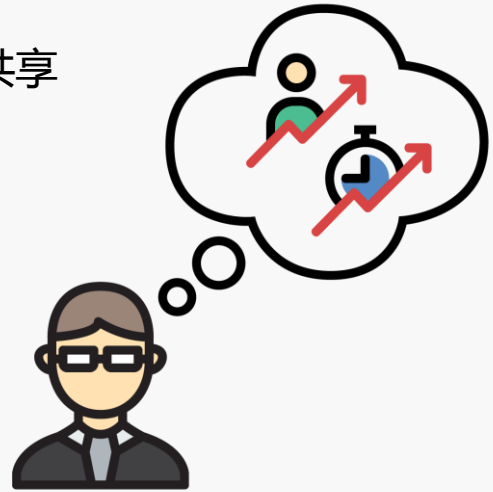


问题和挑战

- 全流程对于工程和算法要求高
- 数据处理繁琐、易错
- 项目特征“孤岛”，开发成本高
- 模型效果优化严重依赖经验
- 部署和运维手工作业，稳定性差
- 重复发明“轮子”，质量参差不齐

平台目标

- 服务人群：机器学习工程师、数据分析师、业务研发工程师
- 简化机器学习模型的开发、测试和部署，提升效率，降低成本
- 提供便利的数据处理和特征管理工具，提升数据和特征质量，促进共享
- 提供全面的监测功能，保证线上服务的稳定、可靠和性能
- 提供高性能的特征计算服务，实现毫秒级的响应
- 提供实时的训练数据生成服务，保证数据质量，避免“穿越”问题
- 运用AutoML技术自动优化模型构建，优化模型效果



主要功能

模型服务

模型管理 | 预测服务 | 实验管理 | 日志落地 | 指标监测

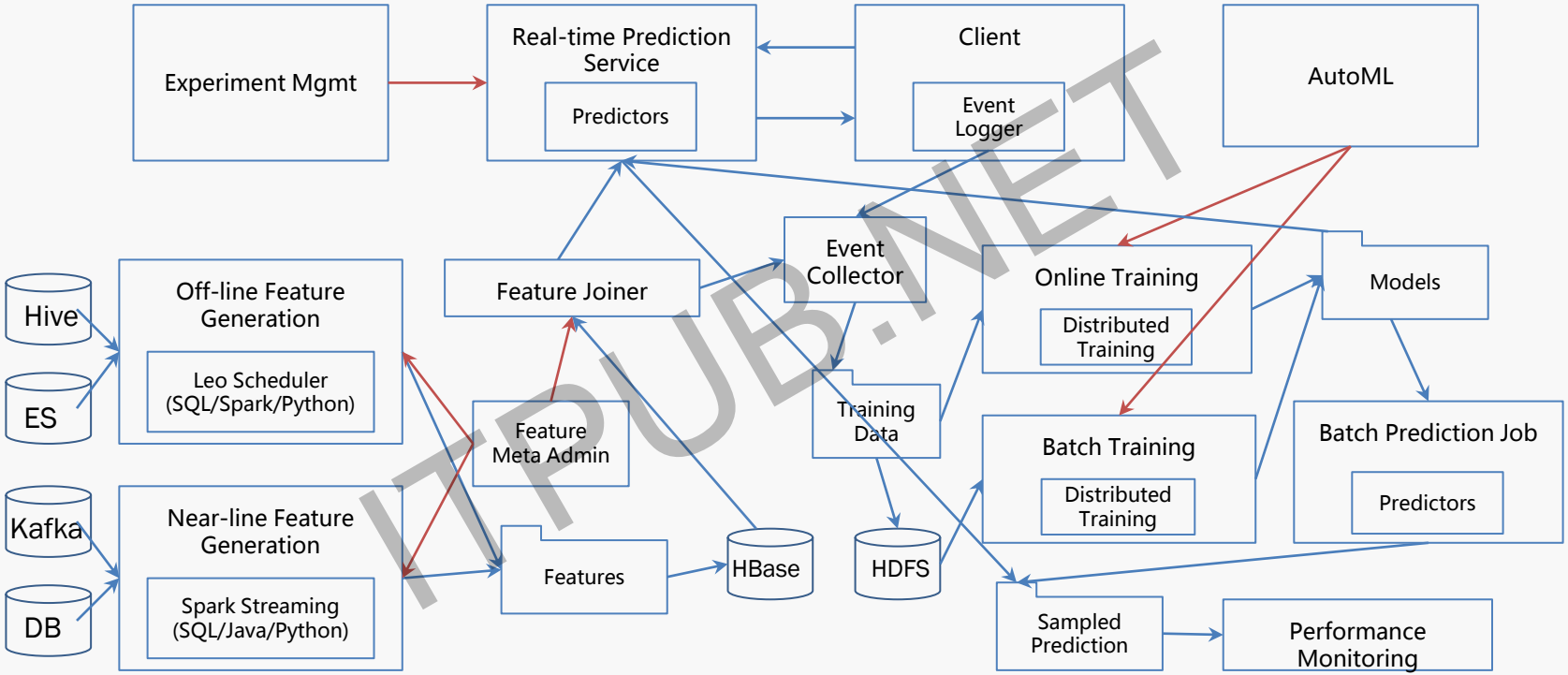
模型构建

特征组合 | 特征筛选 | 降维/聚类 | 模型训练 | AutoML

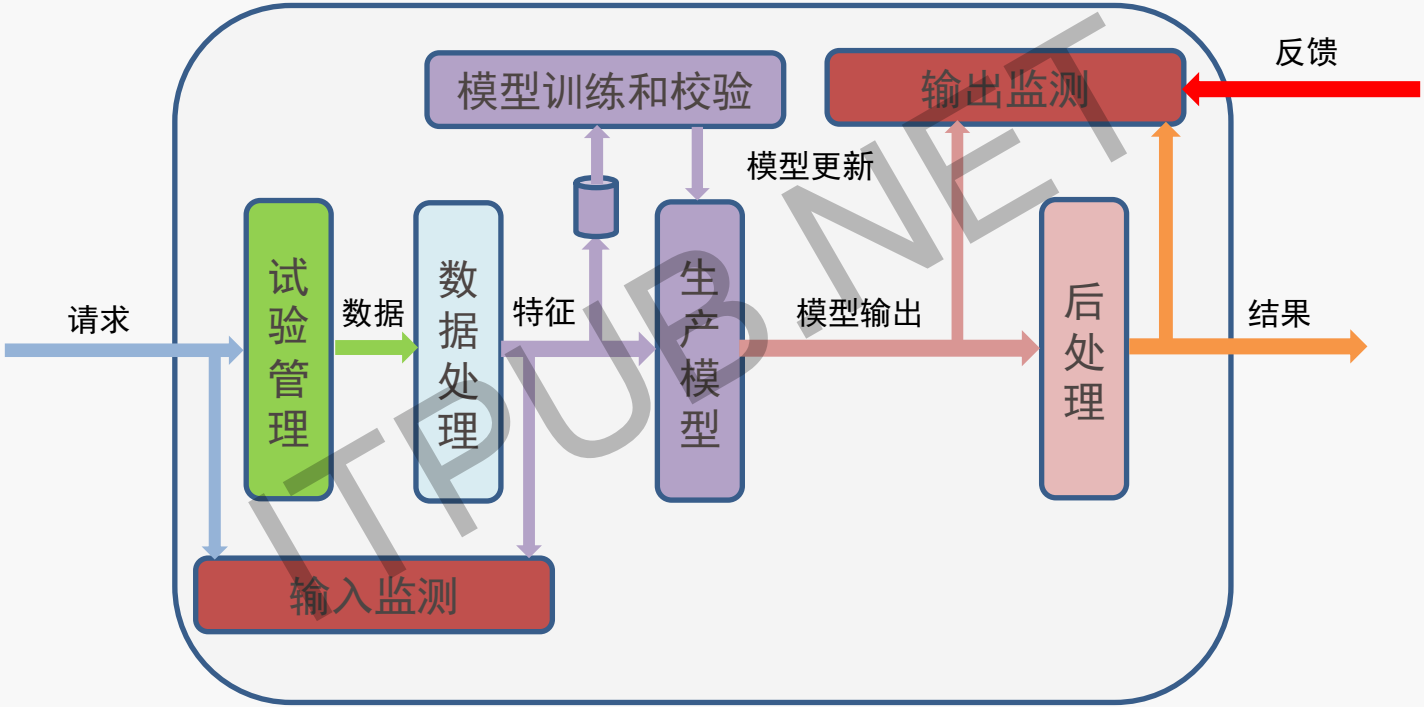
特征计算

元数据管理 | 特征管理 | 特征生成 | 特征获取 | 特征监测

平台架构

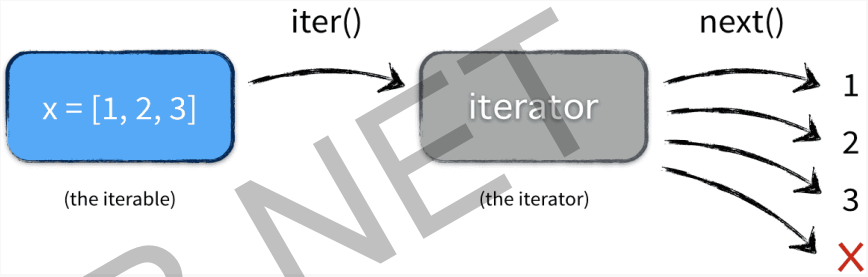


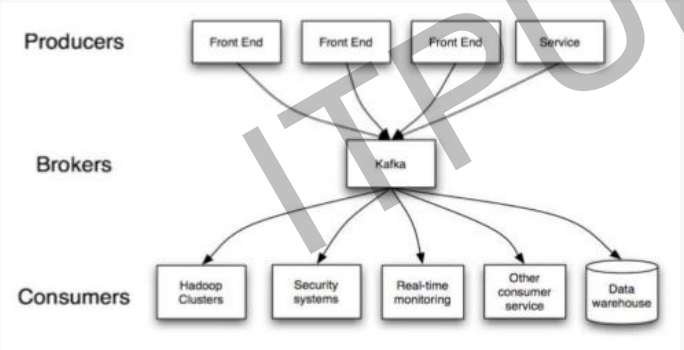
数据流概览

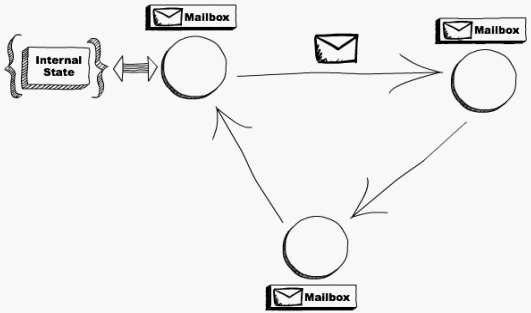


数据处理常用设计模式

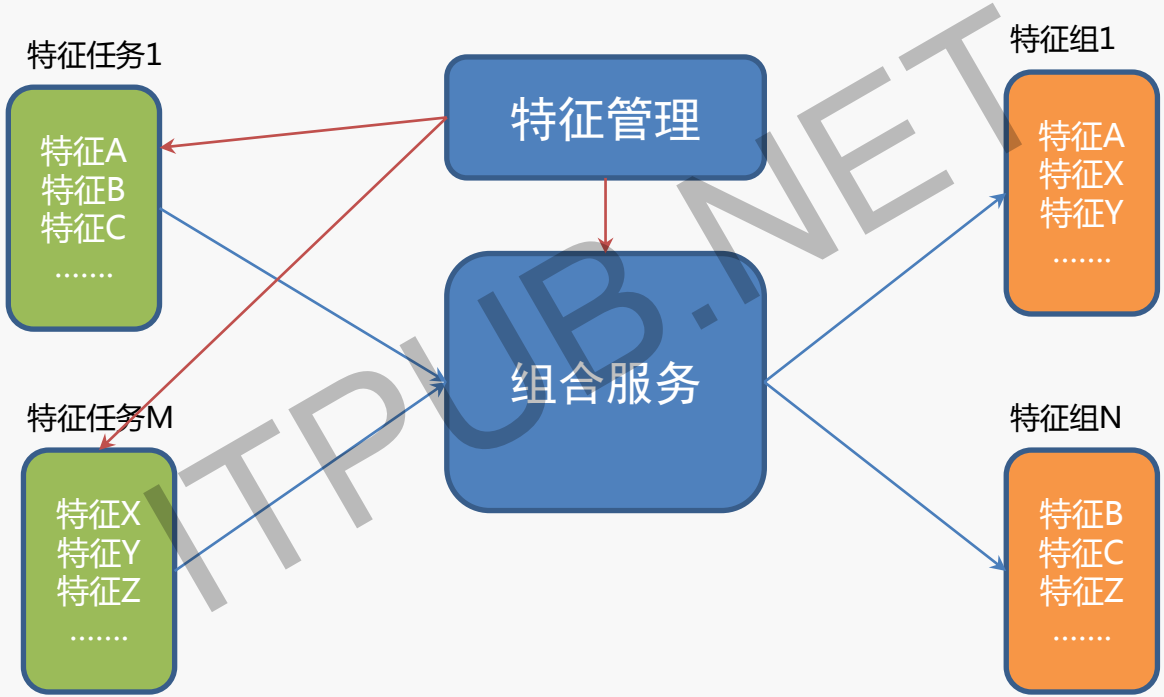
- Iterators
- Pub/Sub
- Actor model
- Caching joins



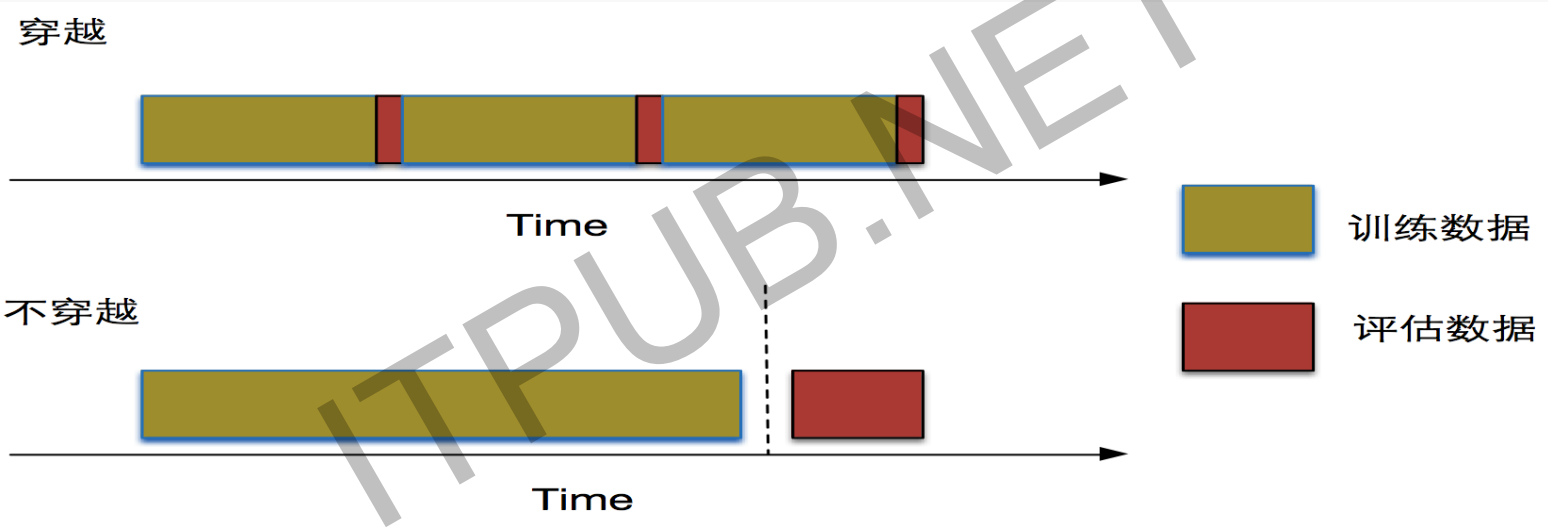




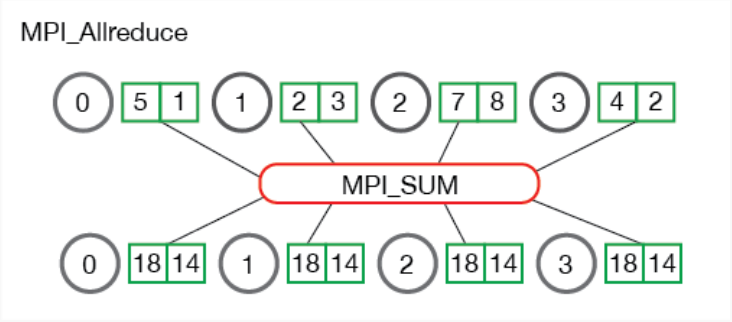
特征任务与特征组

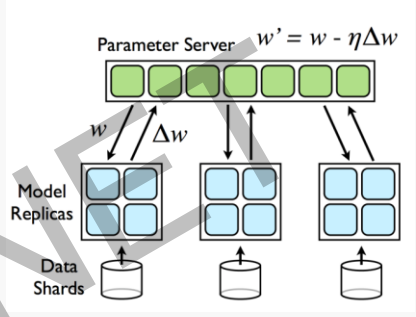


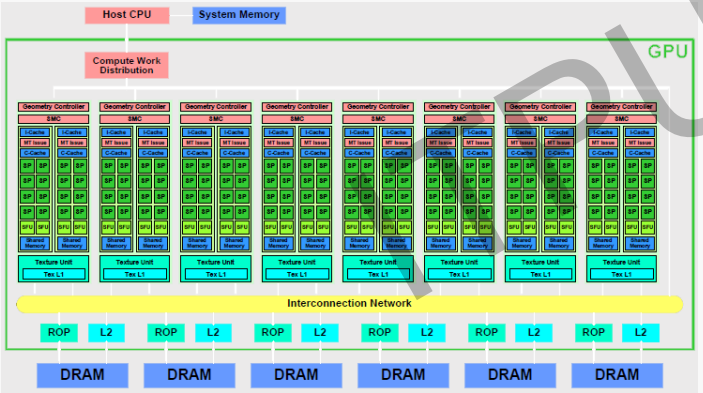
避免“穿越”

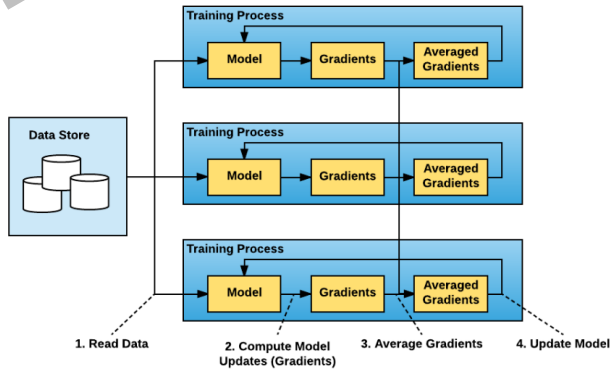


模型训练

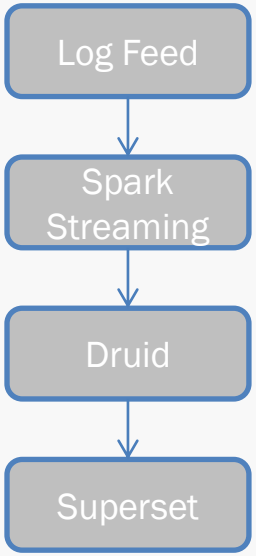


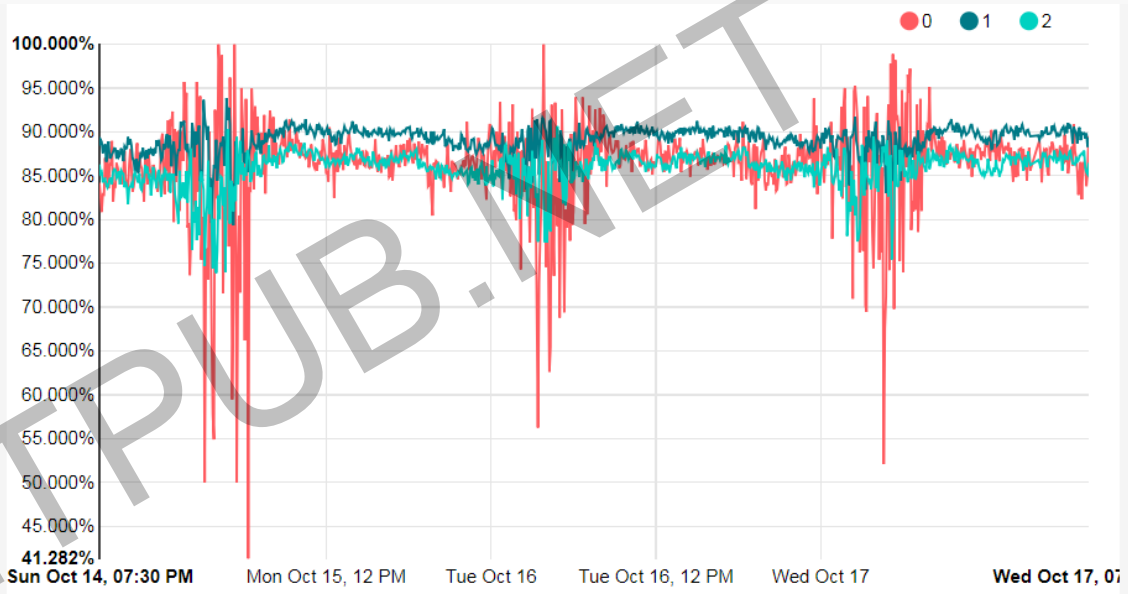






监测


Log Feed
Spark Streaming
Druid
Superset

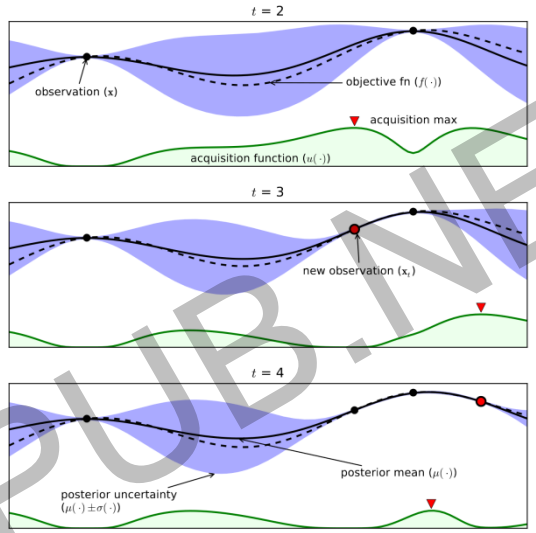

0
1
2
100.000%
95.000%
90.000%
85.000%
80.000%
75.000%
70.000%
65.000%
60.000%
55.000%
50.000%
45.000%
41.282%
Sun Oct 14, 07:30 PM
Mon Oct 15, 12 PM
Tue Oct 16
Tue Oct 16, 12 PM
Wed Oct 17
Wed Oct 17, 07

建模中的超参数

Item CF	相似度算法、相似度因子权重...
Matrix Factorization	隐因子数、正则化权重...
Neural Networks	结构、层数、每层神经元数、dropout比例
GBDT	提升次数、树的最大深度、学习率、样本采样率、特征采样率...
Random Forest	树的数量、树的最大深度、样本采样率、特征采样率...
Logistic Regression	正则化权重、正则化方法
Gradient Descent	学习率、批次大小、迭代次数...

AutoML方法

- 贝叶斯优化：
  - 高斯过程回归
  - SMAC
  - TPE
- 谱模型
- Bandit算法
  - Hyperband算法
- Network Architecture Search
  - Network Controller + Reinforcement Learning
  - DART



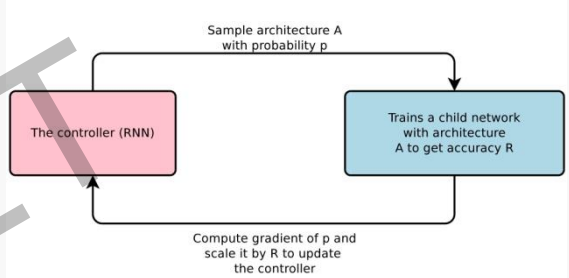


Figure 1: An overview of Neural Architecture Search.

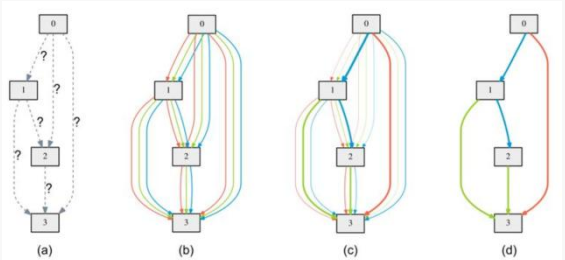
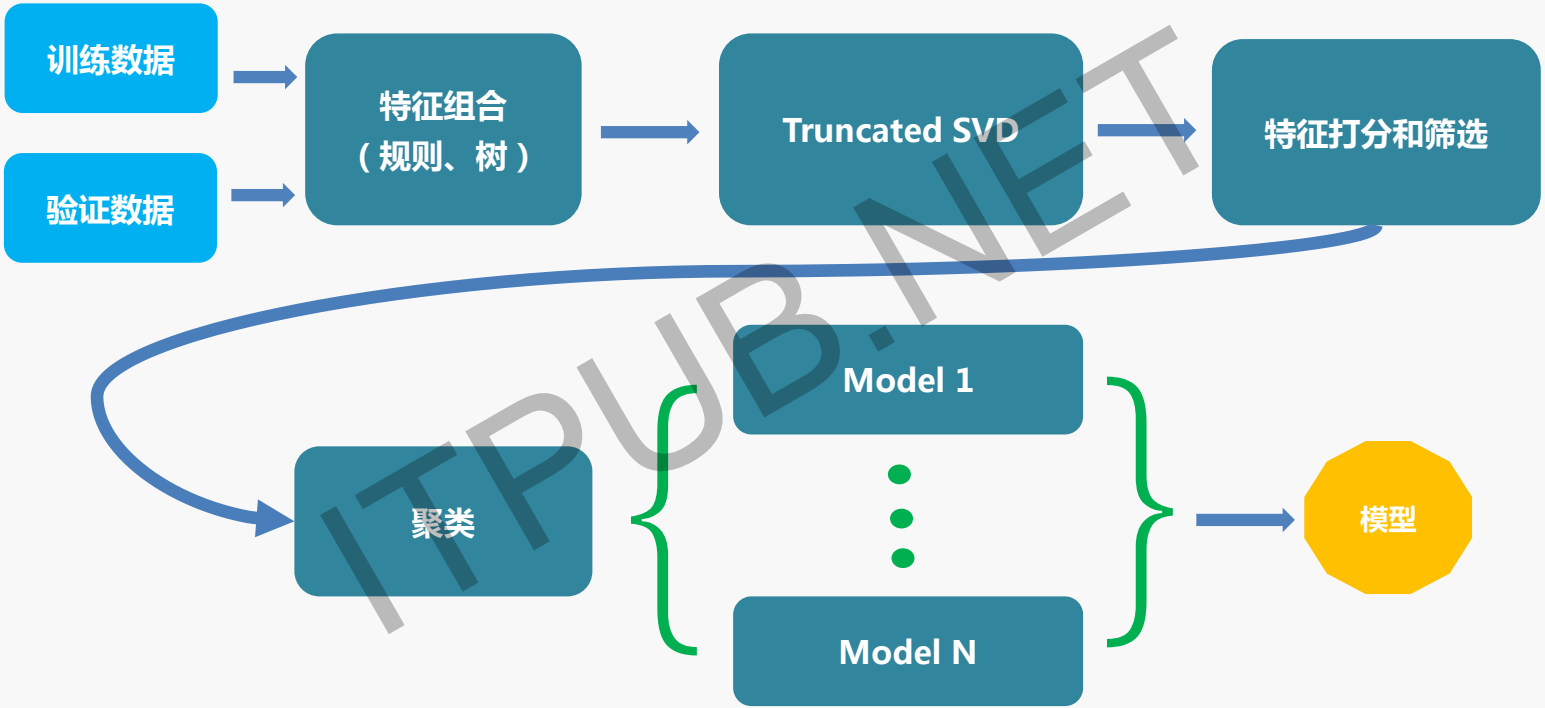
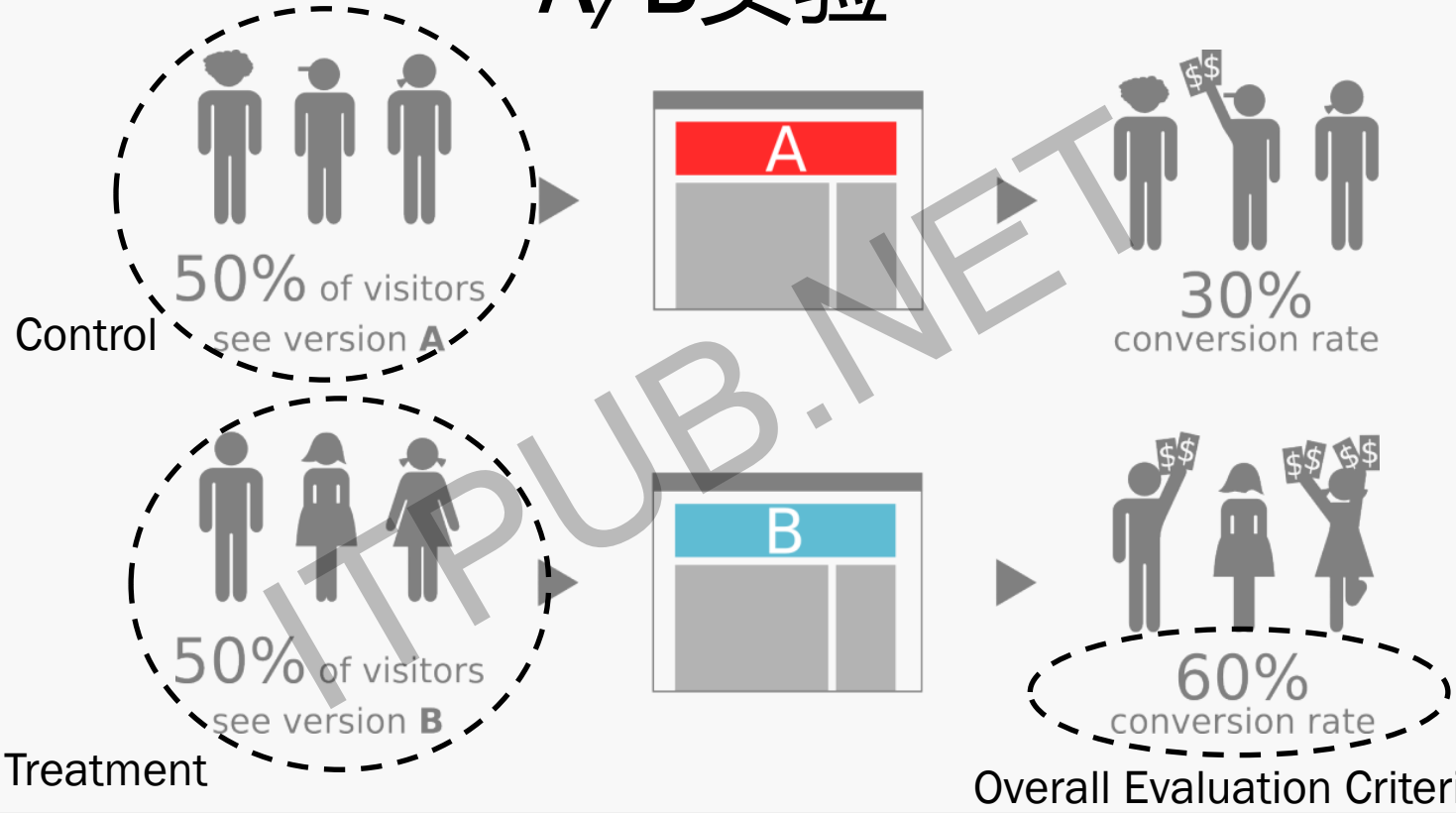


Figure 1: An overview of DARTS: (a) Operations on the edges are initially unknown. (b) Continuous relaxation of the search space by placing a mixture of candidate operations on each edge. (c) Joint optimization of the mixing probabilities and the network weights by solving a bilevel optimization problem. (d) Inducing the final architecture from the learned mixing probabilities.

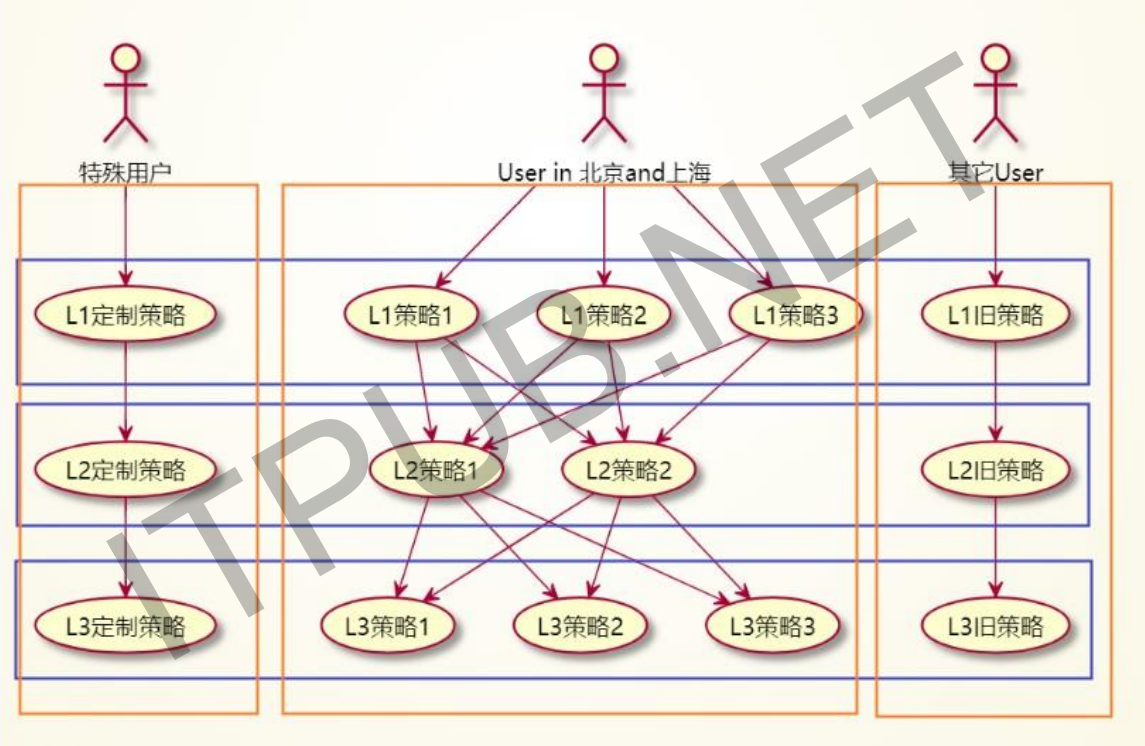
自动化建模



A/B实验



分层分域实验



灵活可配置的实验管理

```
"applicationId": "test_hash_domain",
"totalSize": 1000,
"domains": [
  {
    "domainId": "multi",
    "domainType": "MultiLayerHash",
    "ruleStrategy": "Hash",
    "description": "",
    "default": true
  },
  {
    "domainId": "fix",
    "domainType": "FixedGroup",
    "ruleStrategy": "ExpressionBase",
    "description": "Old Config No.0",
    "ruleExpression": "_ < 3000"
  }
],
```

```
"layers": [
  {
    "layerId": "L1",
    "description": "第一层",
    "policies": [
      {
        "domainIds": ["fix"],
        "name": "E1",
        "param": {
          "test2": 2
        }
      },
      {
        "domainIds":["multi"],
        "name": "E2",
        "size": 300,
        "param": {
          "test3": 3
        }
      },
      {
        "name": "E3",
        "default": true,
        "param": {
          "test3": 4
        }
      },
      {
        "name": "E4",
        "size": 200,
        "param": {
          "test3": 5
        }
      }
    ]
  },
```

```
{
  "layerId": "L2",
  "description": "第二层",
  "policies": [
    {
      "domainIds": ["fix"],
      "name": "F1",
      "param": {
        "test2": 2
      }
    },
    {

      "domainIds": ["multi"],
      "name": "F2",
      "size": 400,
      "param": {}
    },
    {
      "name": "F3",
      "size": -1,
      "default": true,
      "param": {
        "test3": 3
      }
    }
  ]
}
```

开源实验管理系统：Macaw

https://github.com/lpdig/macaw

THANKS