文本的结构化信息提取

yibwang@thoughtworks.com 王祎

ThoughtWorks®

业务部门的需求……

XX科联招标中心关于平面动漫、智能实训设备采购(GXZT2018-G1-00035-KLXY)招标公告 2018年01月29日 12:09　公告概要：公告信息：采购项目名称XX科联招标中心关于平面动漫、智能实训设备采购(GXZT2018-G1-00035-KLXY)招标公告品目货物/其他货物/其他不另分类的物品采购单位XX市职业教育中心行政区域XX市公告时间2018年01月29日 12:09获取招标文件时间 2018年01月29日 12:09至 2018年01月29日 12:09 获取招标文件的地点XX市公共资源交易中心一楼大厅（XX市XX区XX商业步行街与XX路交叉口东南150米XX大厦）开标时间2018年01月29日 12:09开标地点XX市公共资源交易中心一楼交易厅（XX市XX区金城商业步行街与XX路交叉口东南150米XX大厦）预算金额￥600万元（人民币）联系人及联系方式：项目联系人李晓明项目联系电话0000-8888888采购单位XX市职业教育中心采购单位地址XX市XX区XX路101号采购单位联系方式联系人:李晓明、王建国 联系电话： 0000-8888888代理机构名称XX科联招标中心代理机构地址XX分部地址：XX市XX大道888号XX中心9楼0901室代理机构联系方式联系 人：李建国 联系电话： 0000-8888888 文XX科联招标中心受XX市职业教育中心委托，根据《中华人民共和国政府采购法》等...

期望内容

内容是否与计算机设备相关

招标编号

采购目标

采购预算

采购单位

联系方式

联系人

开标时间

……

目录 CONTENTS

» 结构化信息提取简述

» 结构化信息提取步骤

» 从模型到服务

ThoughtWorks®

结构化信息提取简述

什么是结构化信息抽取

将文本中的非结构化信息自动提转成结构化数据的过程。

千岛湖，即新安江水库，位于浙江省杭州市淳安县境内，小部分连接杭州市建德市西北，是为建新安江水电站拦蓄新安江上游而成的人工湖，1955年始建，1960年建成。水库坝高105米，长462米；水库长约150千米，最宽处达10余千米；最深处达100余米，平均水深30.44米，在正常水位情况下，面积约580平方千米，蓄水量可达178亿立方米，在最高水位时拥有1078座大于0.25平方千米的陆桥岛屿，并以2平方千米以下的小岛为主，岛屿面积共409平方千米。杭州千岛湖与加拿大渥太华西南200多千米的金斯顿千岛湖、湖北黄石阳新仙岛湖并称为"世界三大千岛湖"。

抽取出的结构化信息

景点名称：千岛湖

位置：浙江省杭州市淳安县

高：105米

长：462米

面积：约580平方千米

蓄水量：178亿立方米

为什么要做结构化信息抽取

构建领域知识库

构建面向特定任务的知识库，可以再次基础上实现智能知识服务等

辅助商业决策

进行特定目标信息的发现和识别，减少人工内容提取，辅助进行商业决策

智能服务的基础

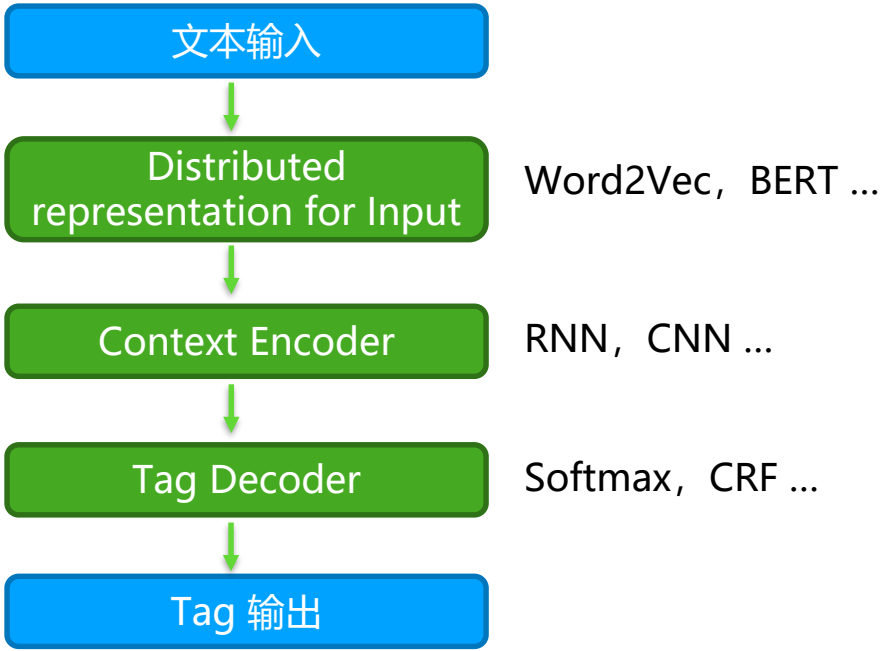
从文本中抽取出的信息框架可以用于信息检索、问答系统、情感分析等应用

常用方案

传统方法

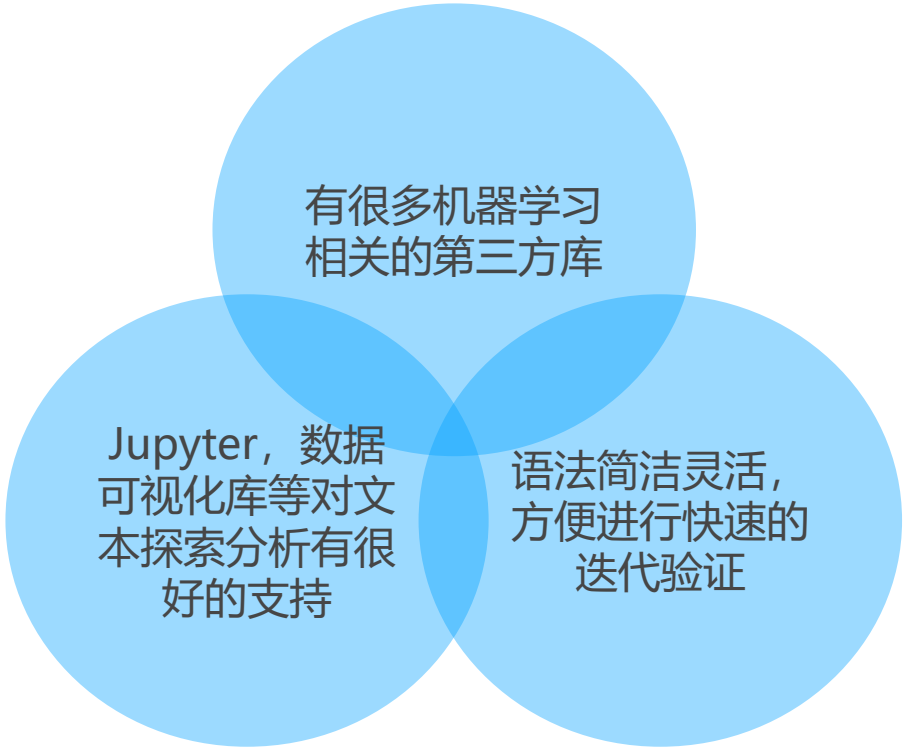
- 基于规则的方法
- 无监督学习的方法
  - 聚类
- 基于特征的监督学习方法
  - HMM
  - 决策树
  - 最大熵模型
  - 支持向量机
  - CRF

深度学习方法



为什么使用Python

- 对文本进行探索分析
- 需要使用机器学习算法
- 快速进行方法验证



如何提取结构化信息

结构化信息提取步骤

文本数据获取

文本清洗

- 文本中不合理的换行
- 句子中有干扰的空格

文本整理： 句子的拆分与合并

公告在《
XX省
政府采购网》上发布。公告期限 201 8 年 1 月 5 日至 201 8 年 1 月 18 日。 7.联系方式 采 购 人:
云南大学
地 址:
xx市
xxxxx号 联 系 人:
xxx
联系电话:
0001-00000000
采购代理机构:
云南中咨海外咨询有限公司
地 址：xx路xxxx 17幢1单元4楼（xx路与xx路交叉口） 联 系 人：李xx 联系电话： 0001-0000000转 6000 传 真： 0000-0000000
开户银行：中国XX银行XXX市区支行 帐号： 0000000000000000000 版权所有:

文本清洗

Before Clean

Python is a populer programming language. It was created by Guido van Rossum, and released in 1991.

After Clean

python popular program language. it create by Guido van Rossum, and release in 1991.

文本清洗

统一文本：英文中的大小写等

去除无意义的字符

文本纠错

- 拼写校正
- 错字校正

构建停用词表，去除停用词

词干提取（有词形变化的语言）

停用词

停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为Stop Words（停用词）

文本预处理

构建分词词典

文本分词

当样本数据不足时，进行样本增强

文本标注

常用：B-，I-，E-，S-，O……

Before

千岛湖，即新安江水库，位于浙江省杭州市淳安县境内，小部分连接杭州市建德市西北，是为建新安江水电站拦蓄新安江上游而成的人工湖，

After

千岛湖	NAME
，	O
即	O
新	B-NAME
安江	I-NAME
水库	E-NAME
，	O
位于	O
浙江省	B-LOC
杭州市	I-LOC
…	

文本预处理

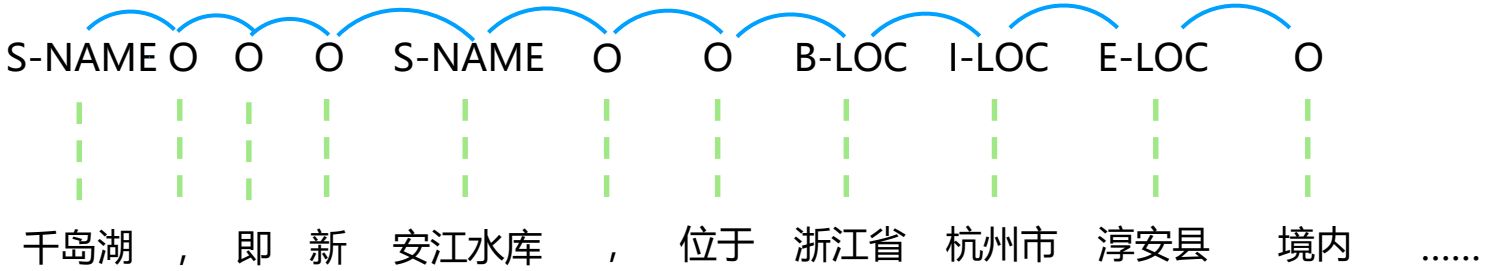
- 选择合适的文本表示

传统	基于字	基于词	混合
One-hot Tf-idf	ELMo …	GloVe Word2Vec …	Bert XLNet …

- 选择算法

CRF, CNN, RNN….

特征提取：CRF模型

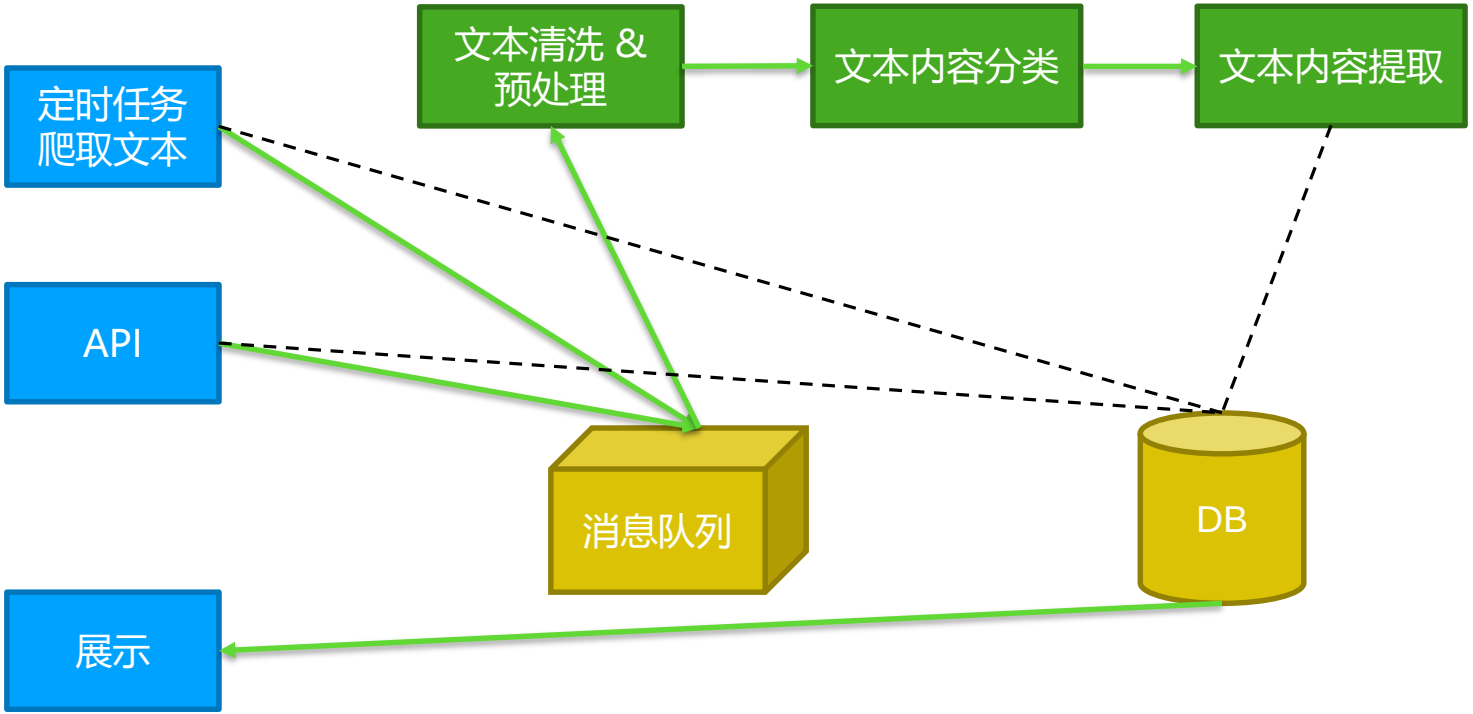


- 可以表达词与词之间的依赖
- 可以表示与前后多个标记之间的依赖关系
- 统计全局概率，可以得到全局最优解

从模型到服务

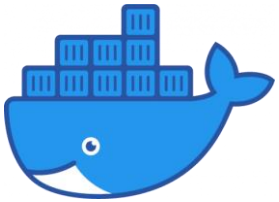
构建服务

- 服务构建



构建服务

服务部署



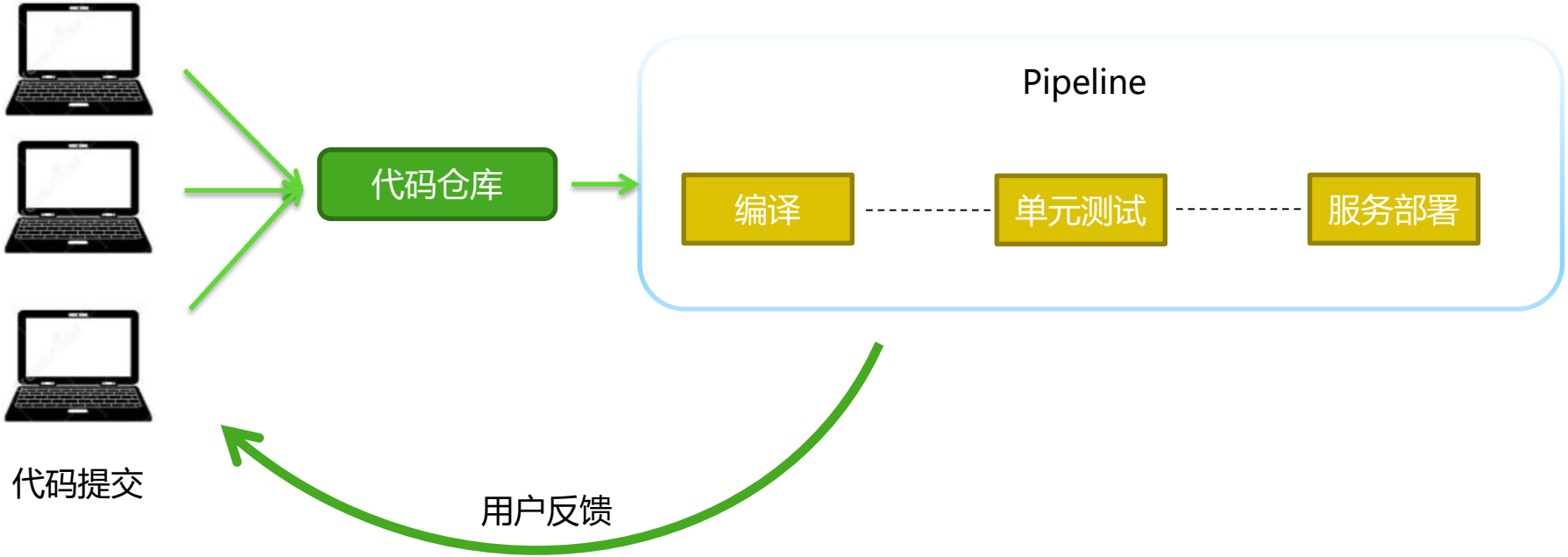
API构建

workflow

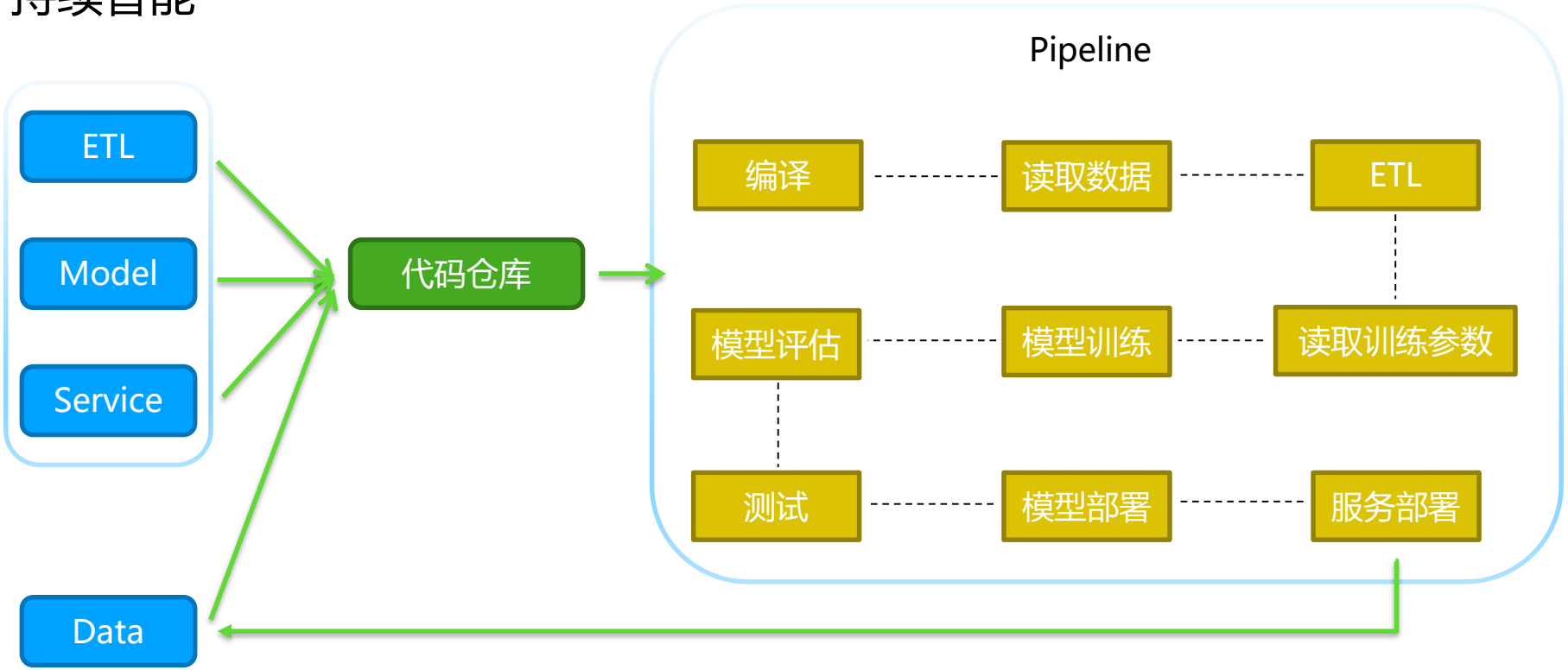
消息队列

DB

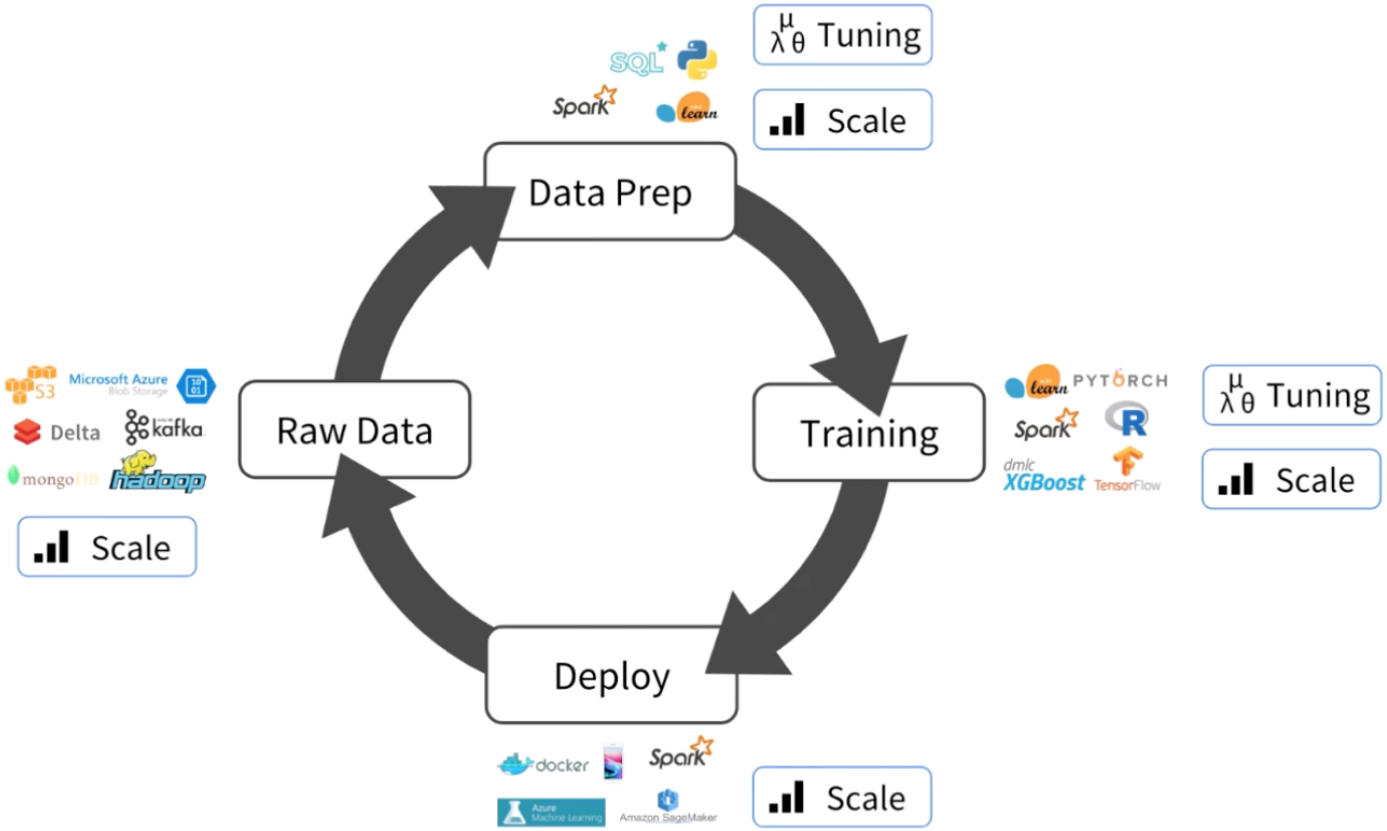
持续智能



持续智能


ETL
Model
Service
Data
代码仓库
Pipeline
编译
读取数据
ETL
模型评估
模型训练
读取训练参数
测试
模型部署
服务部署

持续智能


Tuning
Scale
Data Prep
Training
Deploy
Raw Data
Tuning
Scale
Scale
Scale

总结

- 是什么

  从非结构化的文本中抽取出结构化数据的过程。

- 为什么

  特定内容提取，辅助进行其他任务，辅助进行决策

- 怎么做

  文本内容爬取　　文本清洗　　文本预处理　　文本标注

  训练模型　　构建服务　　服务部署　　持续智能

ThoughtWorks®

Thanks!