


大连理工大学
信息检索研究室
Information Retrieval Laboratory of DUT

信息检索与文本挖掘

11/6/2019

汶东震

目录


大连理工大学
信息检索研究室
Information Retrieval Laboratory of DUT

入门：评测与数据集

11/6/2019

汶东震

前言与背景

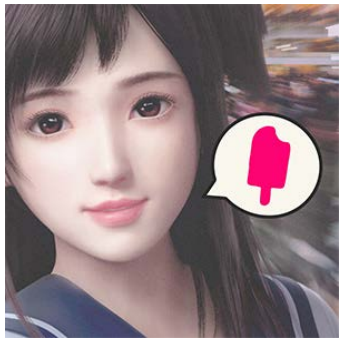
- **人工智能大势所趋**

习近平：推动我国新一代人工智能健康发展

2018-10-31 19:47:56　来源：新华网

- **产学研结合**
  - AI研究如火如荼
  - AI产品百花齐放
  - AI市场前景广阔

科大讯飞 iFLYTEK



NiuTrans

HIKVISION 海康威视

评测简介

数据竞赛与数据科学家

评测

- **学术评测**
  - 目的：学术交流、技术进步、构建标准数据集
    - ImageNet、SemEval、LETOR
- **公司评测**
  - 目的：解决实际问题、提高公司影响力
    - 国外：Quora垃圾评论分类、Google识图
    - 国内：看山杯、平安医疗问句匹配

举办单位

BAAI 北京智源人工智能研究院

中国平安 PINGAN
金融 · 科技

国家电网
STATE GRID

评测类型

重磅比赛

全部 >

U-RISC 神经元识别大赛

用人工神经网络发现生物神经…

主办方: 智源研究院

奖励: ￥100,000

时间轴:2019-10-31 ~ 2020-01-15

OAG-WholsWho 赛道二

智源 - AMiner 姓名排歧大赛 赛…

主办方: 智源、AMiner

奖励: ￥50,000

时间轴:2019-09-30 ~ 2019-12-02

OAG-WholsWho 赛道一

智源 - AMiner 姓名排歧大赛 赛…

主办方: 智源、AMiner

奖励: ￥50,000

时间轴:2019-09-30 ~ 2019-12-02

DigSci 科学数据挖掘大赛…

DigSci 科学数据挖掘大赛 2019

主办方: AMiner · Microsoft · biendata

奖励: ¥40,000

时间轴:2019-10-02 ~ 2019-10-12

www.biendata.com

评测类型

学术评测

全部 >

已结束

平安医疗科技疾病问答迁...

主办方: 平安医疗科技

奖励: ¥9000

时间轴:2019-09-18 ~ 2019-11-03

已结束

成语阅读理解大赛

主办方: 中国计算机学会、清华大学...

奖励: ￥24,000

时间轴:2019-06-25 ~ 2019-09-18

已结束

CCIR 2019 基于电子病历...

主办方: 中国信息检索学术会议

时间轴:2019-07-30 ~ 2019-09-10

已结束

CCKS 2019 公众公司公告...

主办方: CCKS & 东南大学

奖励: ￥15,000

时间轴:2019-07-20 ~ 2019-07-25

www.biendata.com

评测类型

其他比赛

全部 >

已结束

CS 492 Tsinghua University

清华大学计算机系2019年...

清华大学 CS 492 数据挖掘的...

时间轴:2019-05-10 ~ 2019-06-16

已结束

AMiner

2017 高级机器学习第三次...

仅限清华机器学习学生参加

时间轴:2017-04-08 ~ 2017-06-10

已结束

mobike CUP

(练习赛) 摩拜

主办方: 摩拜

时间轴:2017-9-27 ~ 2017-12-16

已结束

知乎

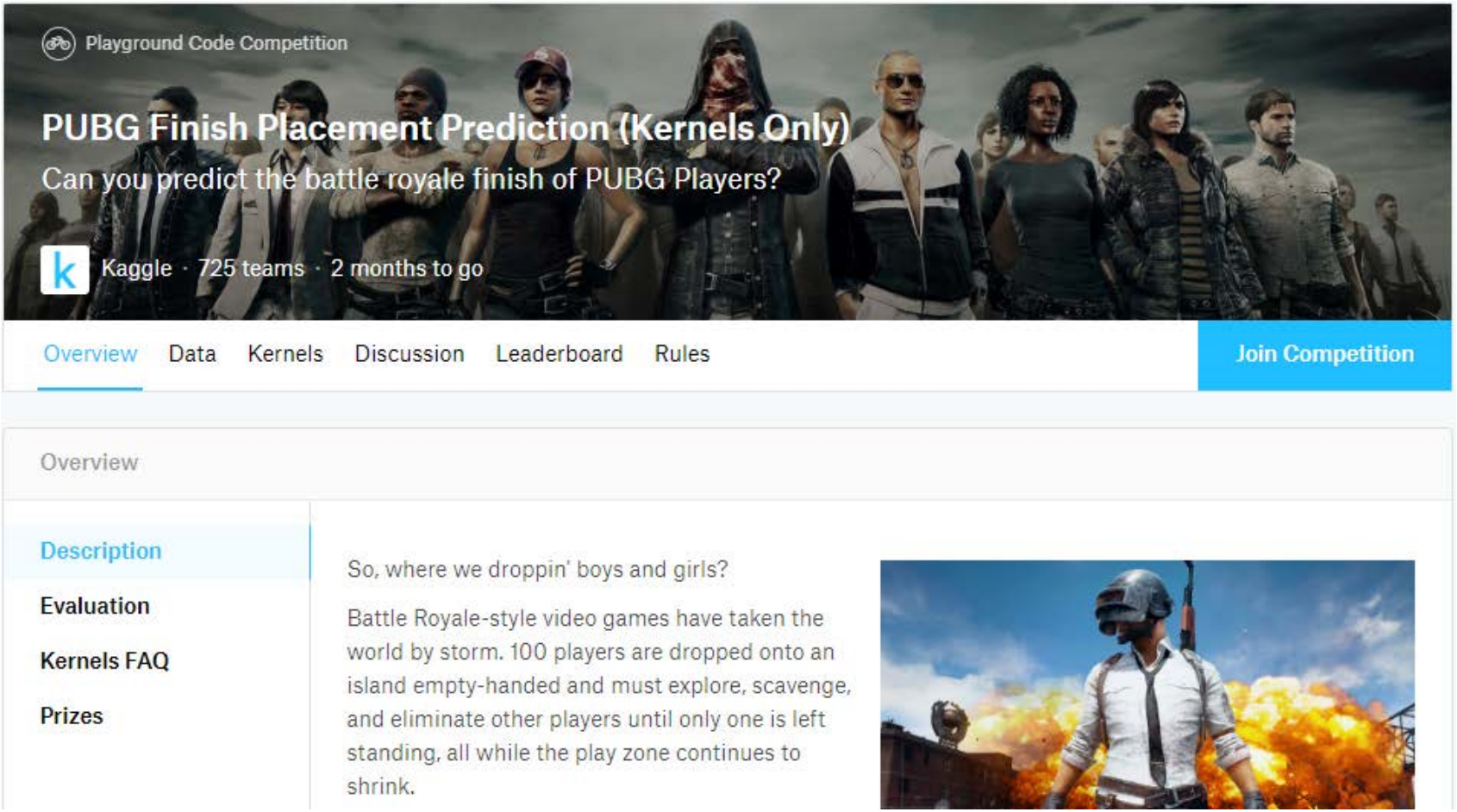
(练习赛) 2017 知乎看山杯...

主办方: 知乎

时间轴:2017-08-16 ~ 2017-11-16

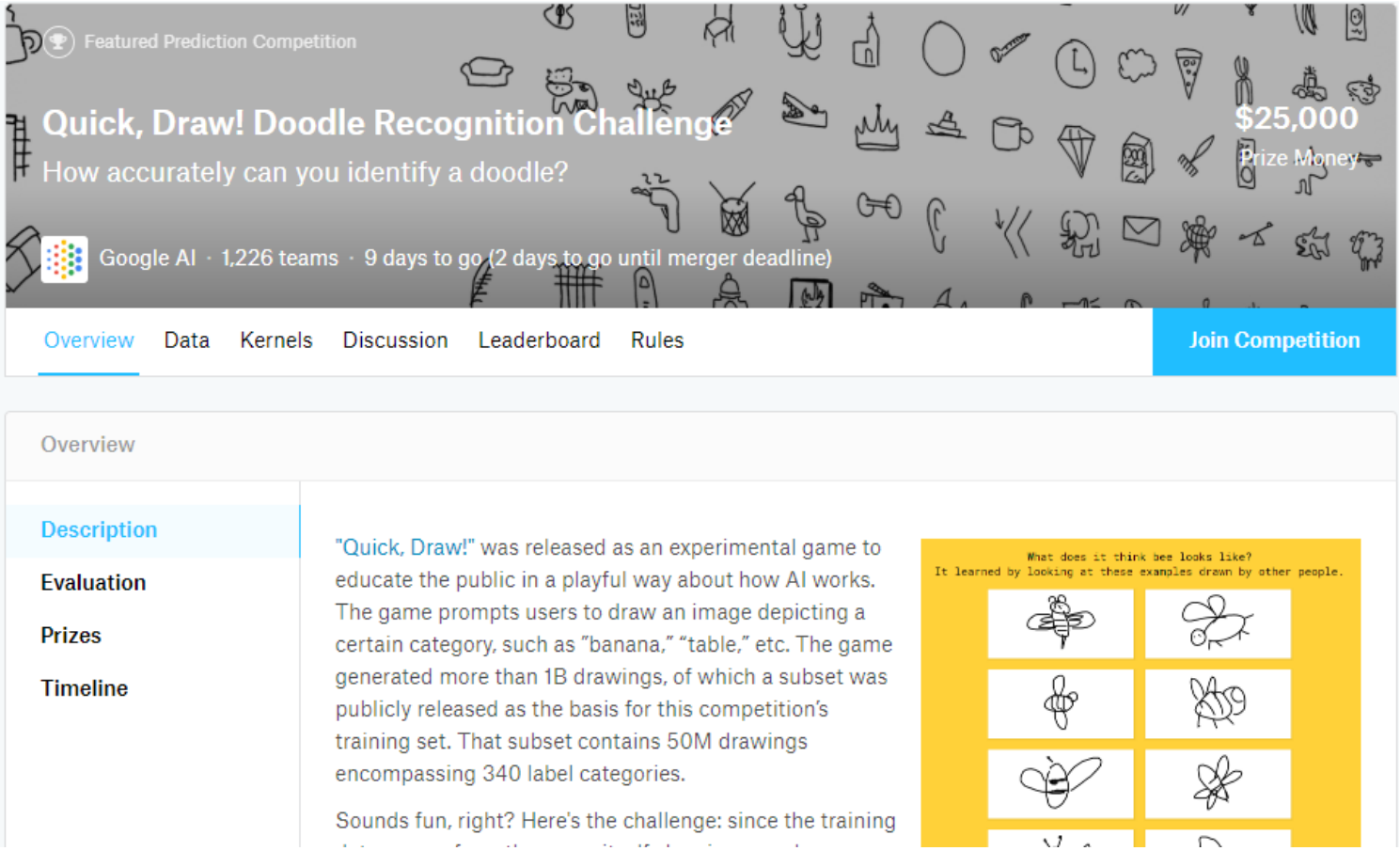
www.biendata.com

评测类型



https://www.kaggle.com/c/pubg-finish-placement-prediction

评测类型



https://www.kaggle.com/c/quickdraw-doodle-recognition

参赛方式

- **按照评测方式区分**
  - 实时打榜：线上提交，实时结果
  - 一锤定音：线下自测，一次提交
- **按照参与方式区分**
  - 线下训练：自备设备，线下训练
  - 在线运行：提交代码，在线运行

数据集和评价指标

人在江湖

数据集

- 数据科学研究的基础
- 评价算法、模型、方法理论
- Benchmark与Baseline
- 公开公正公平的比较
- 华山论剑中的“华山”

图像领域

- **MNIST**
  - 手写数字识别
  - 训练集6w样本
  - 测试集1w样本
  - 28*28矩阵
  - 784像素

http://yann.lecun.com/exdb/mnist/

图像领域

- **ImageNet**
  - 计算机视觉（CV）领域
  - 高分辨率、高质量（标注）、大规模（1300w）
  - 涵盖任务：
    - 图像分类
    - 目标定位、目标检测
    - 视频序列的目标检测

http://www.image-net.org/

文本领域

- **SQuAD**
  - Stanford Question Answering Dataset
  - 斯坦福问答数据集（机器阅读理解）
  - 15w问题数据

SQuAD2.0
The Stanford Question Answering Dataset

Article: Endangered Species Act
Paragraph: "*. . . Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

https://rajpurkar.github.io/SQuAD-explorer/

文本领域

- **GLUE**
  - 通用语言理解评测
  - 多任务数据集
  - 涵盖自然语言推断、自然语言理解、情感分析、语义匹配、蕴含识别
  - 高手云集

General Language Understanding Evaluation

Rank	Name	Model
1	T5 Team - Google	T5
2	ALBERT-Team Google Language	ALBERT (Ensemble)
3	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)
4	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)
5	Facebook AI	RoBERTa

https://gluebenchmark.com/leaderboard/

信息安全领域

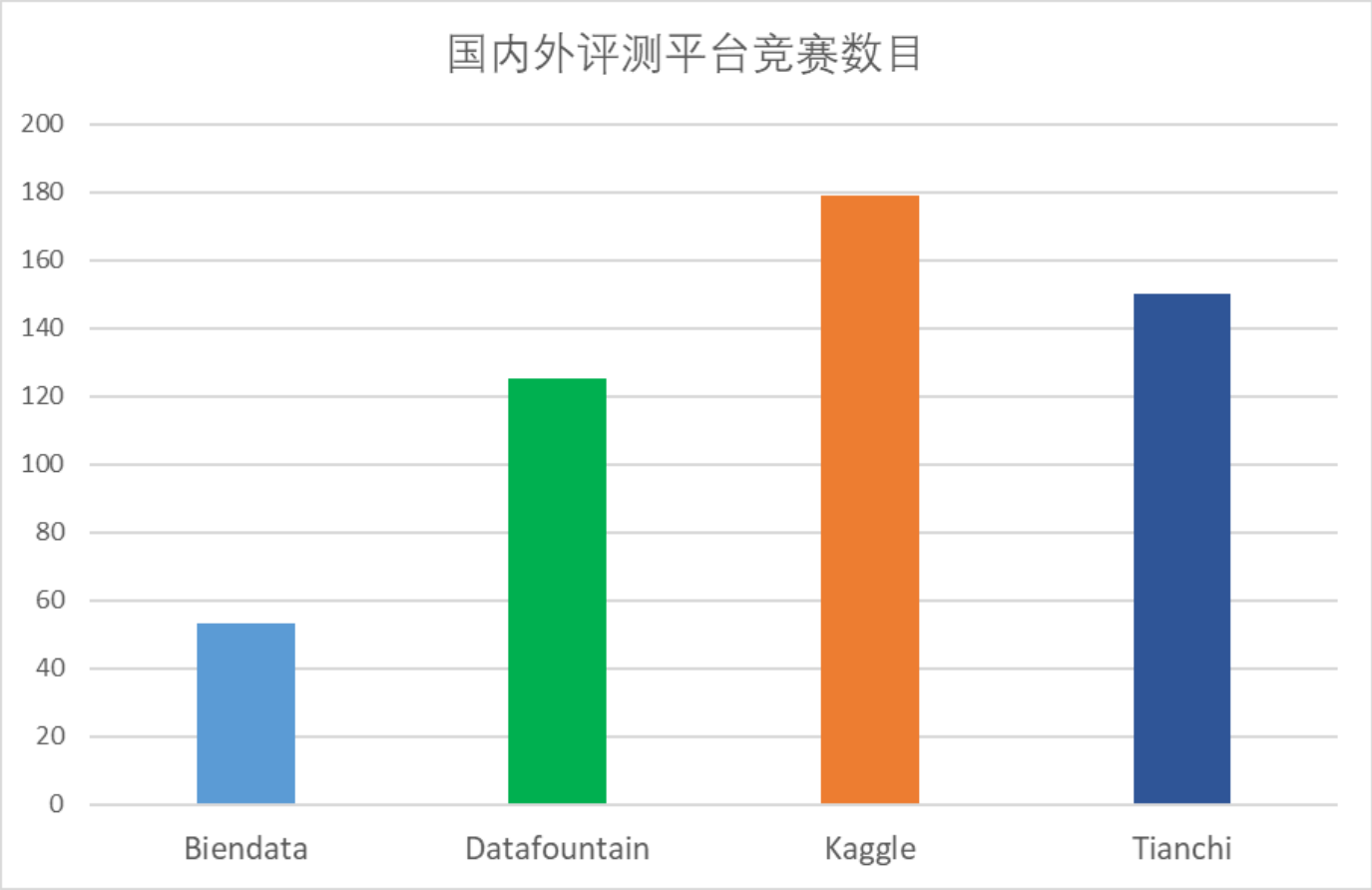
- **KDD99**
  - 1998年DARPA入侵检测评估数据
  - 模拟不同用户、不同流量、不同攻击手段
  - 网络入侵智能检测的Benchmark

http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

数据竞赛平台

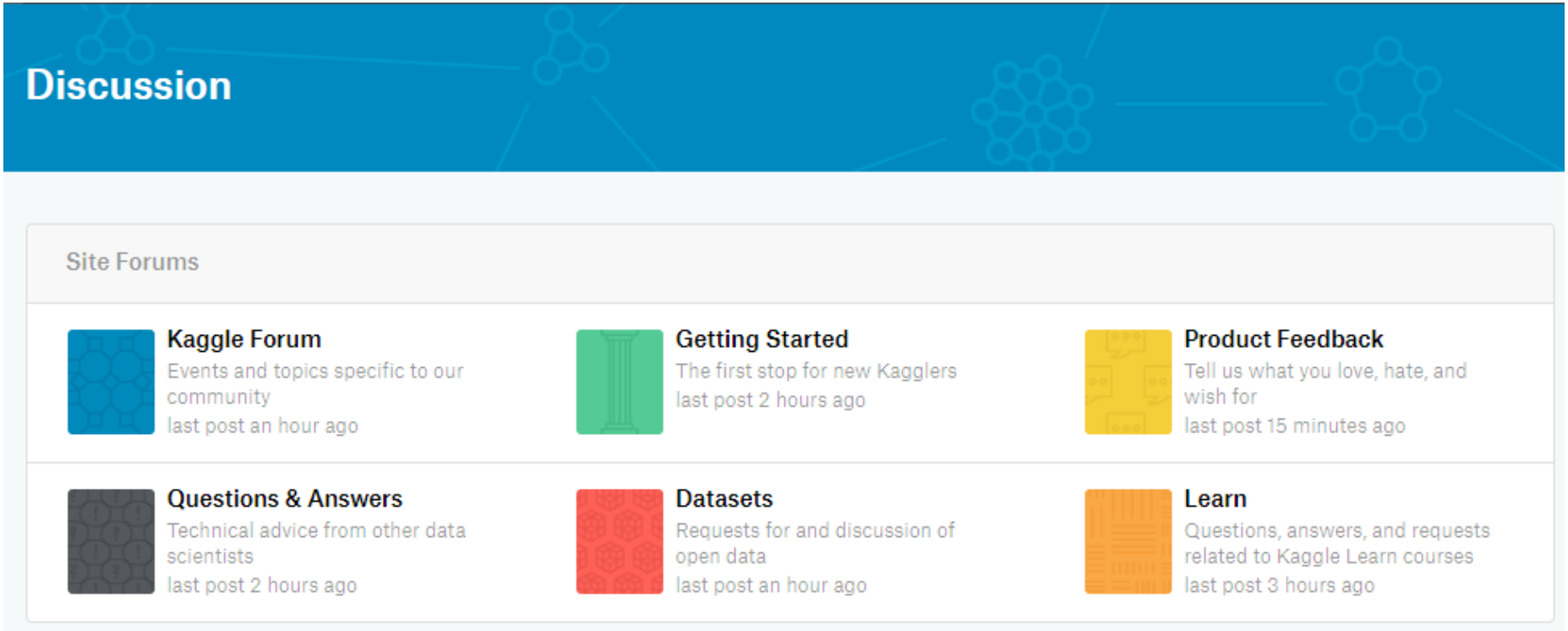
Here we go~

主流平台



https://www.biendata.com/; https://www.kaggle.com;
https://www.datafountain.cn/; https://tianchi.aliyun.com/home/

社区--平台



https://www.kaggle.com/discussion

Kaggle平台

- **性质**
  - 众包模式的数据科学平台
  - 数据科学家社交平台
  - 多学科交互论坛
  - 产学研耦合的枢纽

kaggle

Making Data Science a Sport

Biendata

https://www.biendata.com/

Datafountain

https://www.datafountain.cn/

天池大数据平台

https://tianchi.aliyun.com/home/

数据科学三步走

- **第一步：有的放矢，找准目标**
-
-


大连理工大学
信息检索研究室
Information Retrieval Laboratory of DUT

军械库：数据科学工具链

11/6/2019

汶东震

前言

- **从数据中来 到数据中去**
  - 数据科学涉及大量数据处理、分析、可视化内容
  - 针对任务目标，使用机器学习模型拟合数据分布
  - 使用一套工具解决一系列问题
- **构建自己的军械库**
  - 分解问题，选择工具，寻找最佳实践
  - 抽象流程，重构代码，增强代码复用

编程语言 编程 职业发展 数据分析 修改

哪种编程语言吸金度最高? 修改

在众多行业中，程序员属于高薪职业。无论是在国外还是国内，的工作岗位。 搜集了一天全国各公司发布在拉钩网的招聘数据，

关注问题 写回答 邀请回答 添加评论

程序员 PHP Java C（编程语言） C++ 修改

哪种编程语言有相对更好的发展和薪水? 修

计算机系本科生，现在大四。准备读研。 下定决心想钻研一门编程间花在哪一门语言上比较合适。 C的指针很好玩，学校主要教的C+

关注问题 写回答 邀请回答 添加评论

选择一门语言

修改

编程 修改

学习哪门编程语言最有前途，最好赚钱，需求量高?

本人软件专业学生一枚，想知道要学习哪门语言最有前途

修改

关注问题 写回答 邀请回答 添加评论 分享 举报 ···

学习哪种编程语言薪酬高? 修改

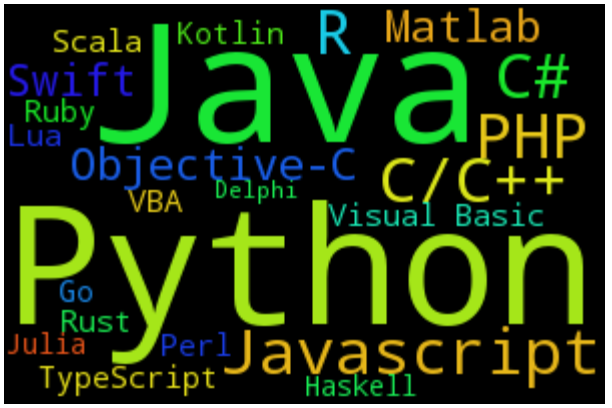
想问下现在哪种语言从事的 开发工资相对较高

修改

关注问题 写回答 邀请回答

编程语言

- 科研？开发？
- 效率？效果？
- 术业有专攻



Nov 2018	Nov 2017	Change	Programming Language	Ratings	Change
1	1		Java	16.746%	+3.51%
2	2		C	14.396%	+5.10%
3	3		C++	8.282%	+2.94%
4	4		Python	7.683%	+3.20%
5	7	^	Visual Basic .NET	6.490%	+3.58%

Python

- **特性**
  - 命令式语言、面向对象
  - 胶水语言、脚本语言

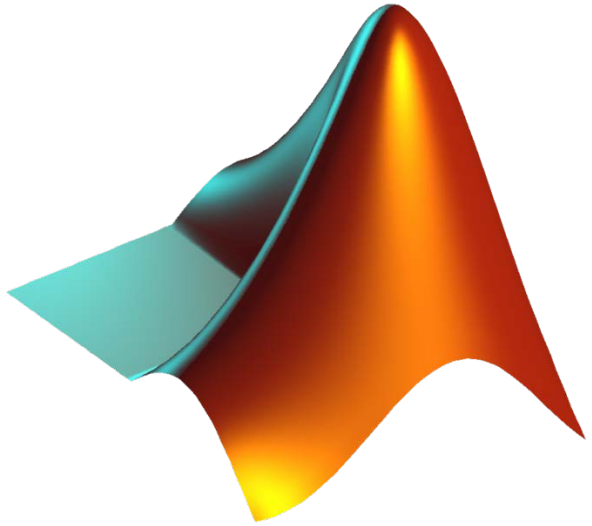
- **领域**
  - 科学计算、系统脚本、网站开发、深度学习
- **There should be one-- and preferably only one --obvious way to do it.**

R

- **特性**
  - 生于统计分析，专于统计分析
  - 语法简单，适合非CS专业学习
  - 完整的统计—绘图工具链
- **领域**
  - 金融、科学计算
  - 统计分析

Matlab

- **特性**
  - 各类行业专业化工具
  - 图形化界面、专属服务
  - Simulink
- **领域**
  - 航空航天、电力电子
  - 控制优化、信号滤波



Java

- **特性**
  - 良好的跨平台特性
  - 众多开源项目
  - DeepLearning4J

其他

- Julia
- Fortran
- Common Lisp
- Prolog
- Haskell
- C/C++

SWI Prolog

总结

- **合适的？适合的？**
  - 解决特定问题的最佳实践
  - 适合个人使用习惯
- **天下武功唯快不破**
  - 上手快、速度快、开发快、落地快
- **选库多的**

挑选合适的工具

工欲善其事

利其器

- **合适的工具**
  - 提高开发效率，减少错误发生
  - 专事专干，最佳实践
  - 需求指导一切，根据需求选择工具链

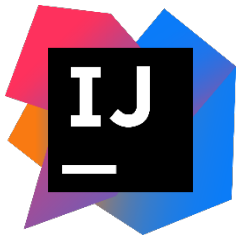
- **不合适的工具**
  - 工具引发bug
  - 好的工具!=适合的工具

Visual studio

- 多种语言支持
- Win平台开发最佳IDE
- 个人用户免费
  - Visual Studio Community
  - VS Code

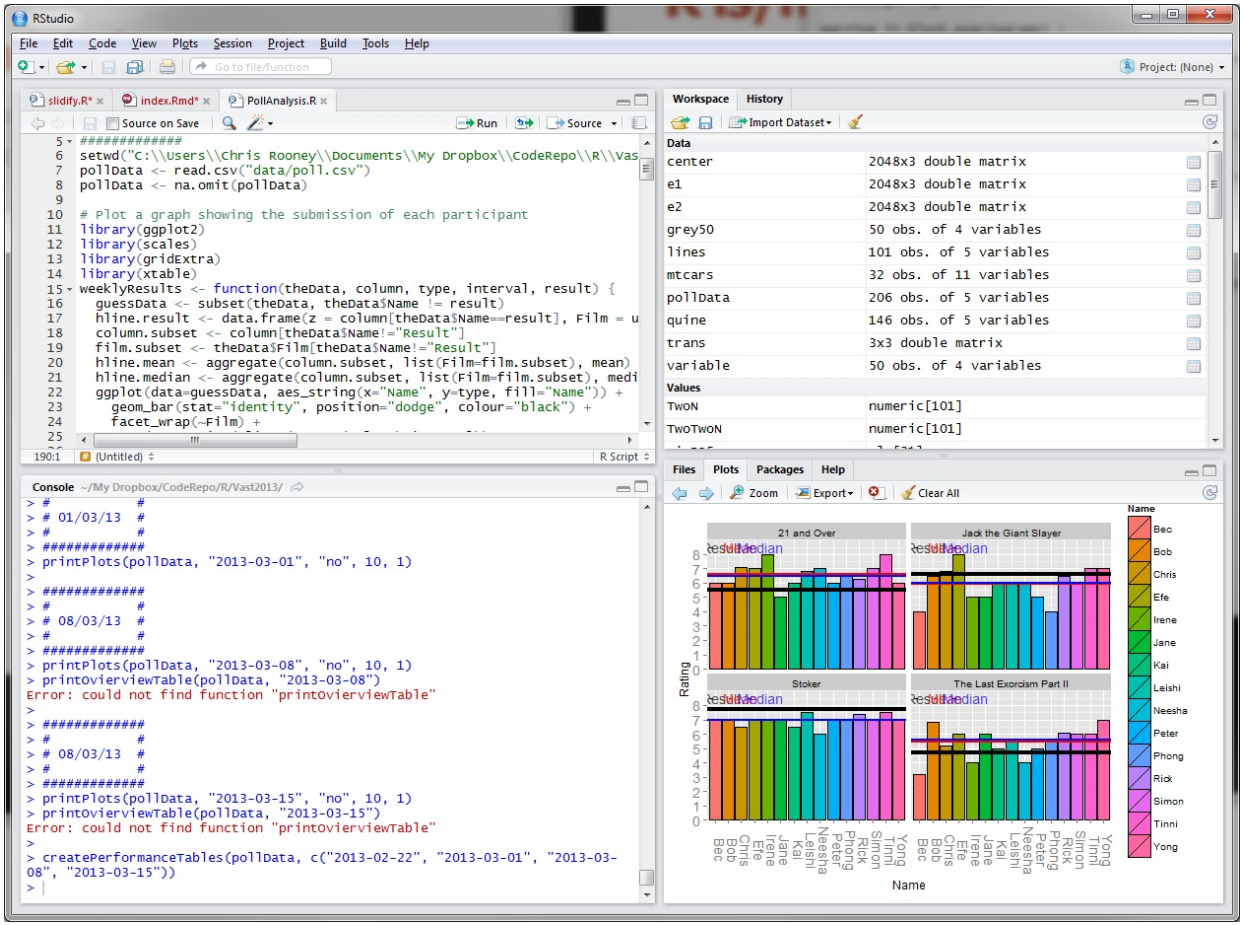
JetBrains

- **全套开发套件**
  - 多种语言支持：R、Go、Php
  - IntelliJ替代eclipse
  - Pycharm
- **申请免费教育版**
  - 校园邮箱验证
  - https://www.jetbrains.com/student/



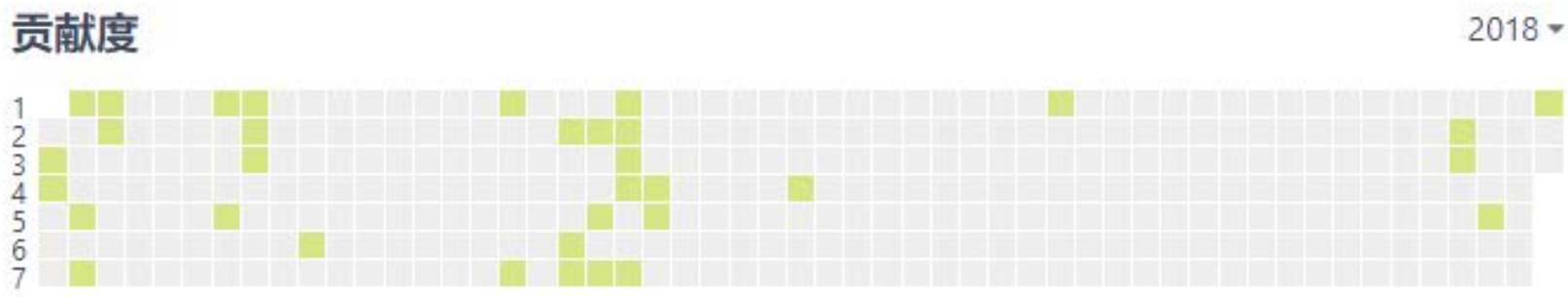
R-studio

- **R语言最佳IDE**



版本管理

- **Git**
  - 分布式版本管理
  - 追踪代码的各个版本
- **Git!=Github**
- **使用Git进行团队合作**

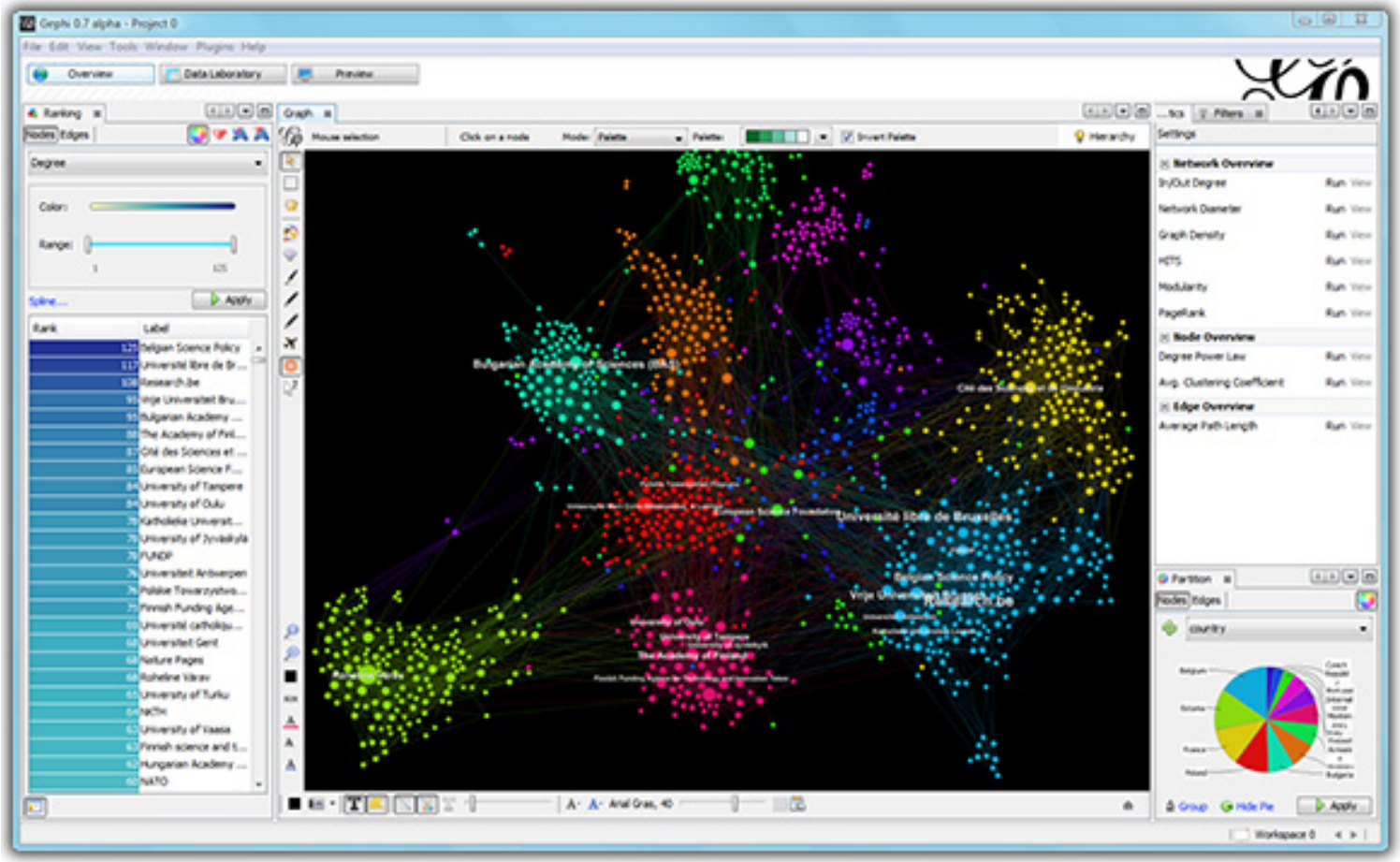


Linux使用

- **Xshell**
  - 字符终端、控制台模式登录
- **Xftp**
  - 文件传输、无需自己搭建ftp
- **Xmanage**
  - 图形化linux远程桌面

网络分析

- Gephi


Gephi 0.7 alpha - Project 0
Overview
Data Laboratory
Preview
Ranking
Graph
Filters
Degree
Color:
Range:
Apply
Rank
Label
Settings
Network Overview
Node Overview
Edge Overview
Partition
country

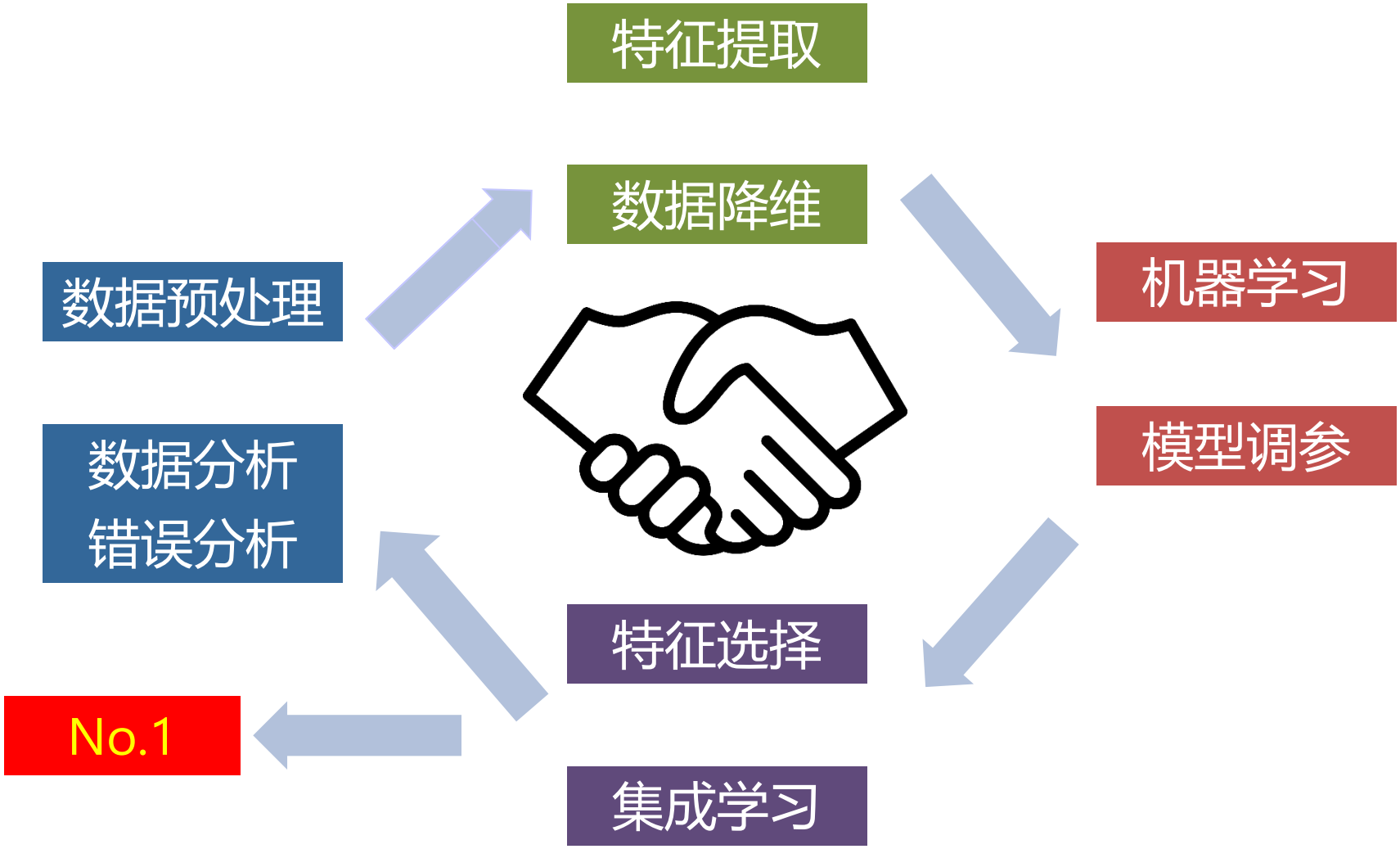
文本编辑器

- ~~Notepad++~~
- VsCode
- Sublime
- Vim

数据科学工作流

从文本任务谈起

非典型工作流



数据预处理

- **中文数据处理**
  - 分词、分短语
  - 词性标注
  - 中文数字转换
  - 中文命名实体标注
  - 语法树分析

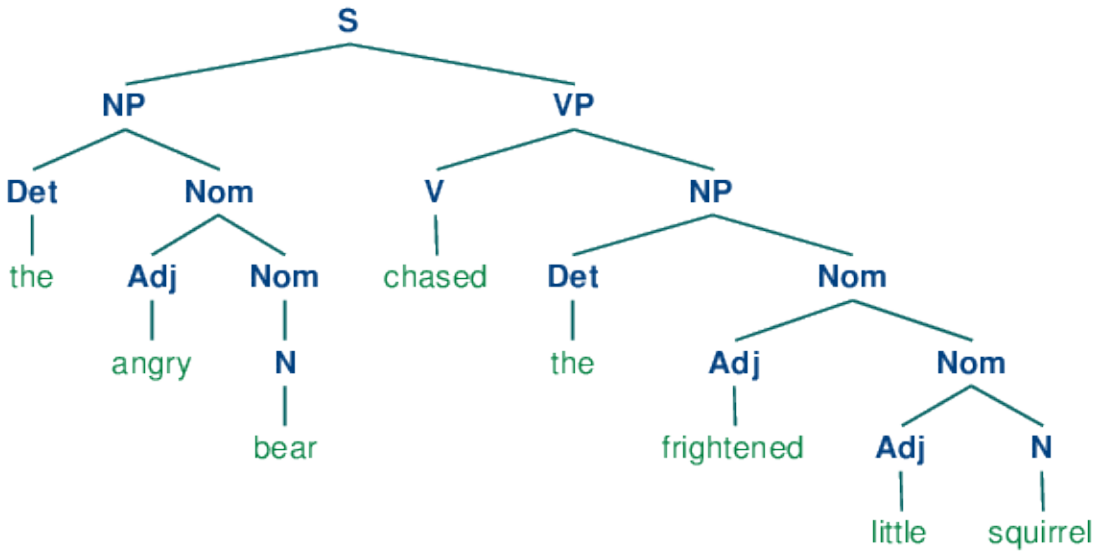
我爱北京天安门
(我 r),(爱 v),(北京 ns),(天安门 ns),

数据预处理

- **英文数据处理**
  - 词干化
  - 词性标注
  - 命名实体标注
  - 语法树分析

```
stemer.stem('exciting')
```

```
'excit'
```



数据预处理

- **根据任务进行数据处理**
  - 词组替换、量词识别替换归一化
  - POS标注、实体标注、关系识别
  - 数值特征归一化、离散化、对数映射
  - Category类型数据重新编码

https://www.kaggle.com/matleonard/categorical-encodings

数据预处理

- **工具**
  - Jieba、NLTK
  - 哈工大LTP
  - 清华大学NLP Toolkit
  - Stanford NLP Toolkit
  - Hanlp

数据分析

- **分布分析**
  - 单个篇章的字、词、句数目，词句比
  - 文档集上篇章长度，字词数目分布
  - 分类任务中，标签在文档上的分布
  - 带有时间戳数据，计算数据在时间上的分布
  - 具有地理位置特性的看地理位置分布

数据分析

- **字词分析**
  - 文本和标注之间关联性
  - 任务相关的关键字发现
- **数值分布**
  - 中位数、四分之一位数、平均数
  - 箱线图、直方图、散点图、饼图

错误分析/异常值分析

构建领域词典

异常分析（数值）

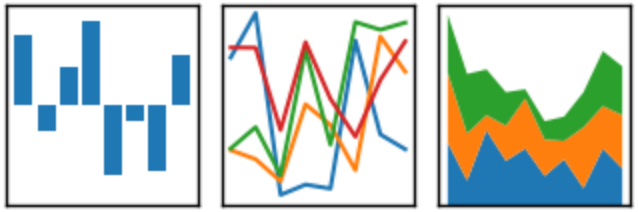
分布分析

数据分析

- **工具**
  - Pandas
  - Matplotlib
  - Bokeh
  - networkX

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Bokeh

matplotlib

NetworkX

特征提取

- **文本特征**
  - Ngram、Tfidf
  - BM25(信息检索)
  - 深度表示：word2vec、glove、elmo、bert
- **数值特征**
  - 字词句长度，关键词数目等等
  - 数值特征：加减乘除、取对数（log）

特征提取

- **分布特征**
  - 部分数值在时间上的分布
  - 部分数值在地域上的分布
  - 其他各种
- **网络信息（关系信息）**
  - 出度、入度、稠密程度
  - 图嵌入（network embedding）

特征提取

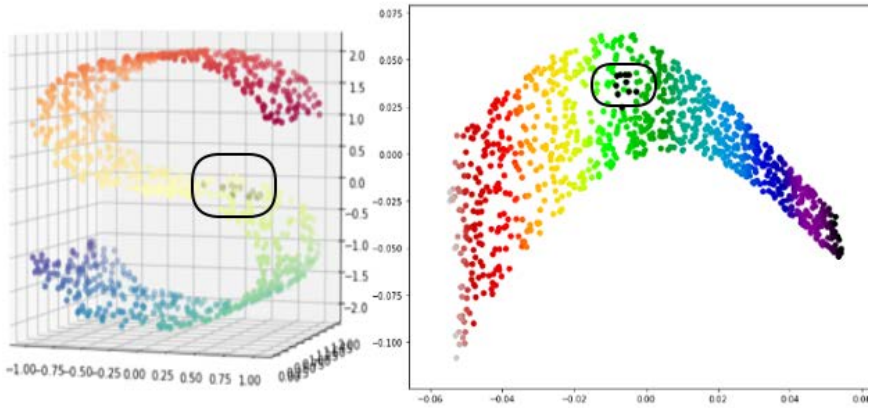
- **工具**
  - Sklearn.feature_extraction.text
  - Sklearn.feature_extraction
  - networkX
  - LINE
  - Graph neural network

数据降维

- **为什么要降维**
  - 特征维度过大，影响算法效率
  - 特征稀疏，影响空间利用率和特征有效性
  - 潜在语义分析，通过矩阵分解的方法
  - 进行可视化、文本聚类

数据降维

- **无监督**
  - PCA（主成分分析）
  - NMF（非负矩阵分解）
  - TSNE
    （t分布随机临近嵌入）
  - LLE（局部线性嵌入）
  - LE（拉普拉斯映射）
  - SE（光谱嵌入）
- **有监督**
  - LDA（线性判别分析）
  - Lasso

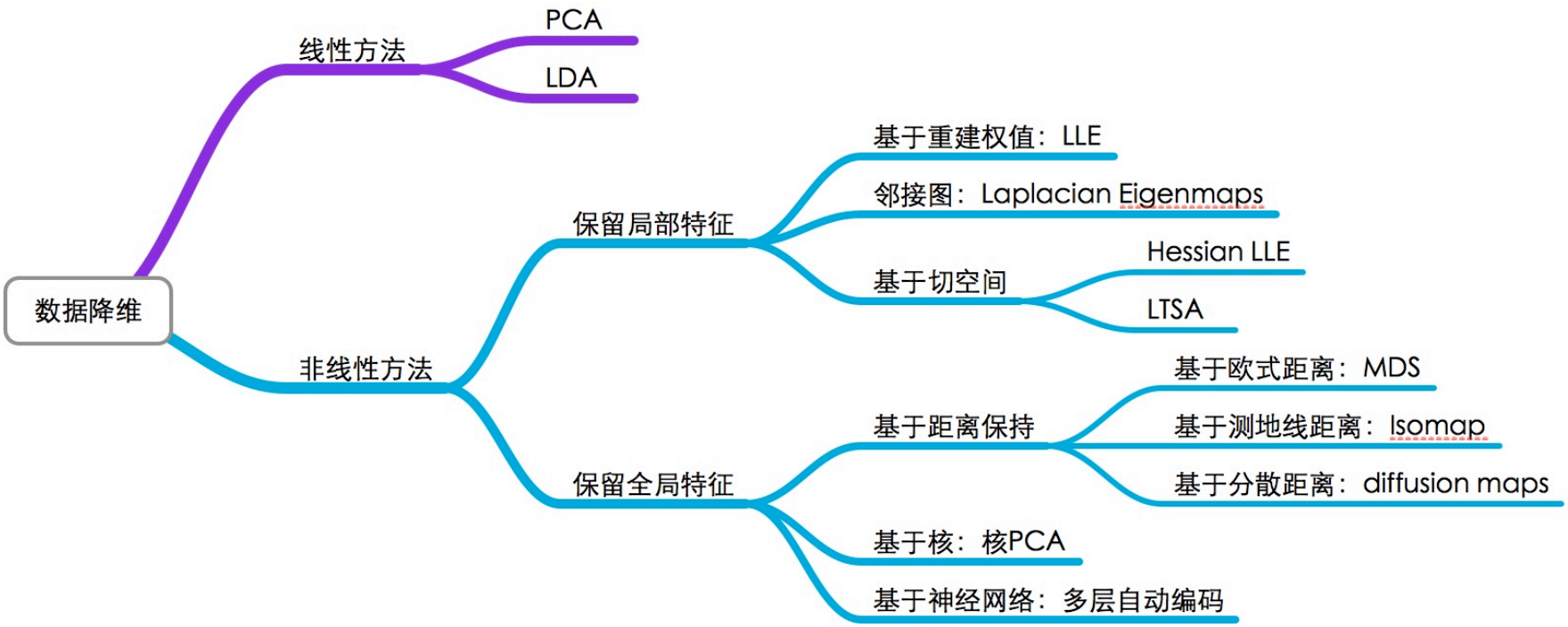


https://blog.paperspace.com/dimension-reduction-with-lle/

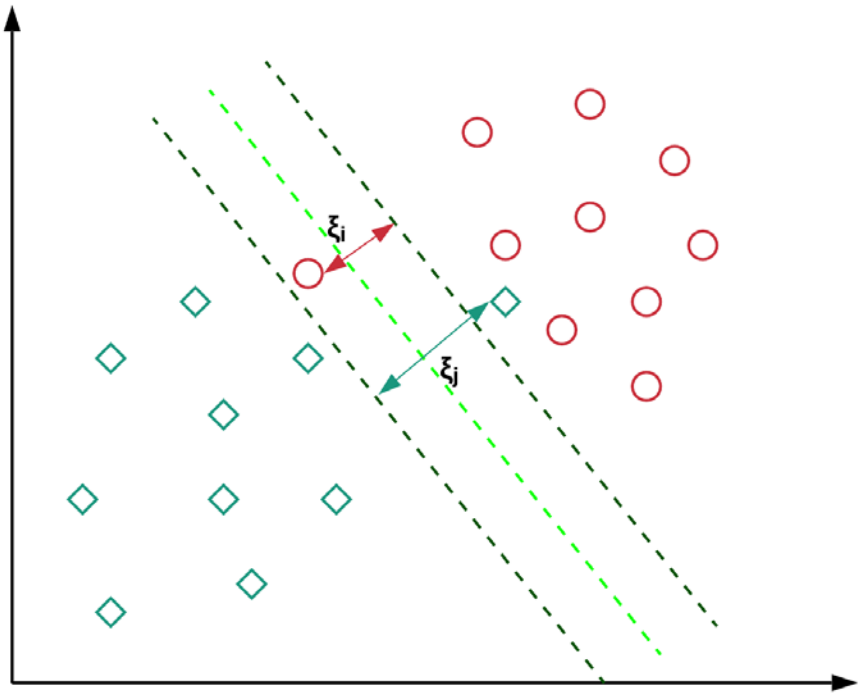
数据降维

- 工具
  - sklearn.decomposition
  - sklearn.discriminant_analysis
  - sklearn.manifold.LocallyLinearEmbedding
  - sklearn.manifold.SpectralEmbedding
  - sklearn.manifold.TSNE

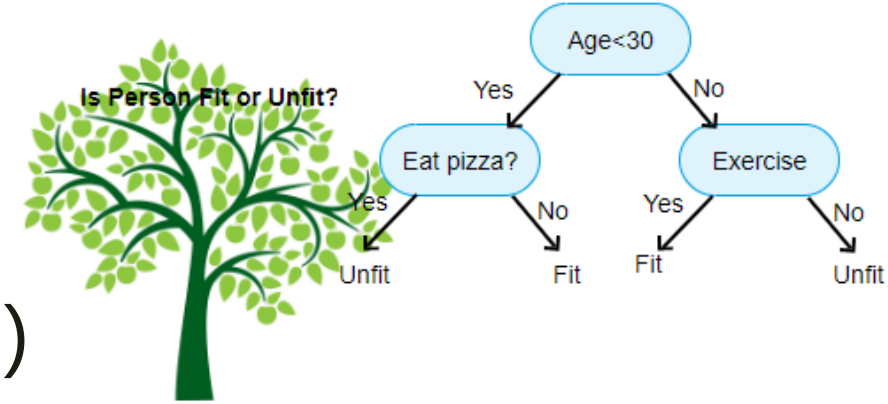
数据降维


数据降维
线性方法
PCA
LDA
非线性方法
保留局部特征
基于重建权值：LLE
邻接图：Laplacian Eigenmaps
基于切空间
Hessian LLE
LTSA
保留全局特征
基于距离保持
基于欧式距离：MDS
基于测地线距离：Isomap
基于分散距离：diffusion maps
基于核：核PCA
基于神经网络：多层自动编码

- **线性模型**
  - 线性回归
  - 逻辑回归
- **支持向量模型**
  - 支持向量回归（SVR）
  - 支持向量分类（SVC）



- **贝叶斯模型**
  - 朴素贝叶斯、高斯模型
  - 多项式模型、伯努利模型
- **决策树模型**
  - 决策树
  - 随机森林（ensemble）
  - GBDT（xgboost、lightgbm）



- **深度学习**
  - CNN结构（Capsule）
  - RNN结构（LSTM、GRU）
  - Seq2seq+Attention
  - Bert、ELMO、GPT-2

http://nlpprogress.com/
https://github.com/RedditSota/state-of-the-art-result-for-machine-learning-problems

机器学习

- **机器学习工具**
  - sklearn
  - Gensim
  - xgboost、lightgbm
- **深度学习框架**
  - Tensoflow
  - Pytroch
  - mxnet

特征选择

- 为什么
  - 特征维度过高影响模型训练效率
  - 筛选较好特征便于下一级模型融合
  - 筛除有害特征避免错误扩散
  - 部分机器学习模型要求特征之间不能存在线性相关性

特征选择

- **工具**
  - Sklearn.feature_selection
  - 通过logisticRegression输出特征重要性
  - 手工测试
    - 由简入繁
    - 由少向多
    - 设计框架自动筛选、避免重复工作

集成学习

- **概念**
  - 多个分类器针对同一任务进行训练，形成多个模型并将其结合使用
- **方法**
  - Bagging
  - Stacking
  - Boosting

- **Bagging**
  - 同一数据集中多次采样训练

Input: Data set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$;
Base learning algorithm $\mathcal{L}$;
Number of learning rounds T.

Process:
for $t = 1, \cdots, T$:
$\mathcal{D}_t = Bootstrap(\mathcal{D})$; % Generate a bootstrap sample from $\mathcal{D}$
$h_t = \mathcal{L}(\mathcal{D}_t)$ % Train a base learner h_t from the bootstrap sample
end.

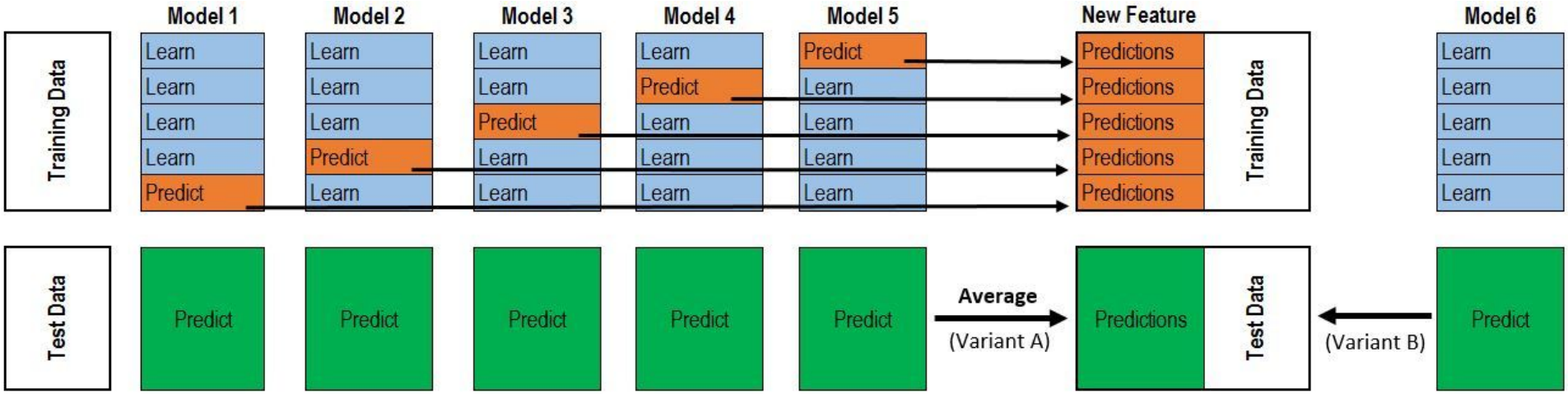
Output: $H(\boldsymbol{x}) = \text{argmax}_{y \in \mathcal{Y}} \sum_{t=1}^{T} 1(y = h_t(\boldsymbol{x}))$ % the value of $1(a)$ is 1 if a is $true$ and 0 otherwise

Fig. 2. The Bagging algorithm

https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/springerEBR09.pdf

集成学习

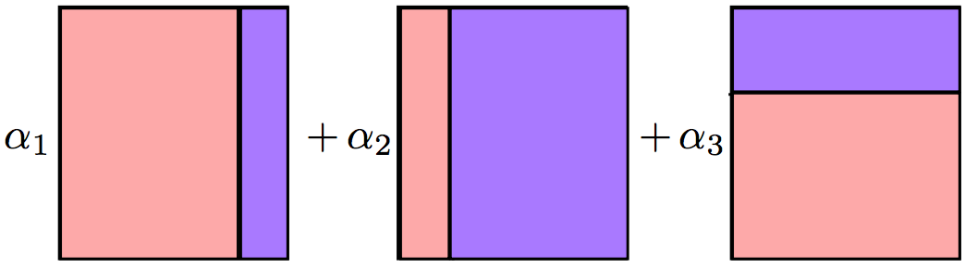
- **Stacking**

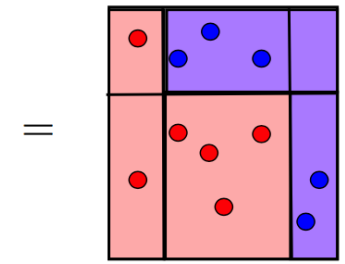


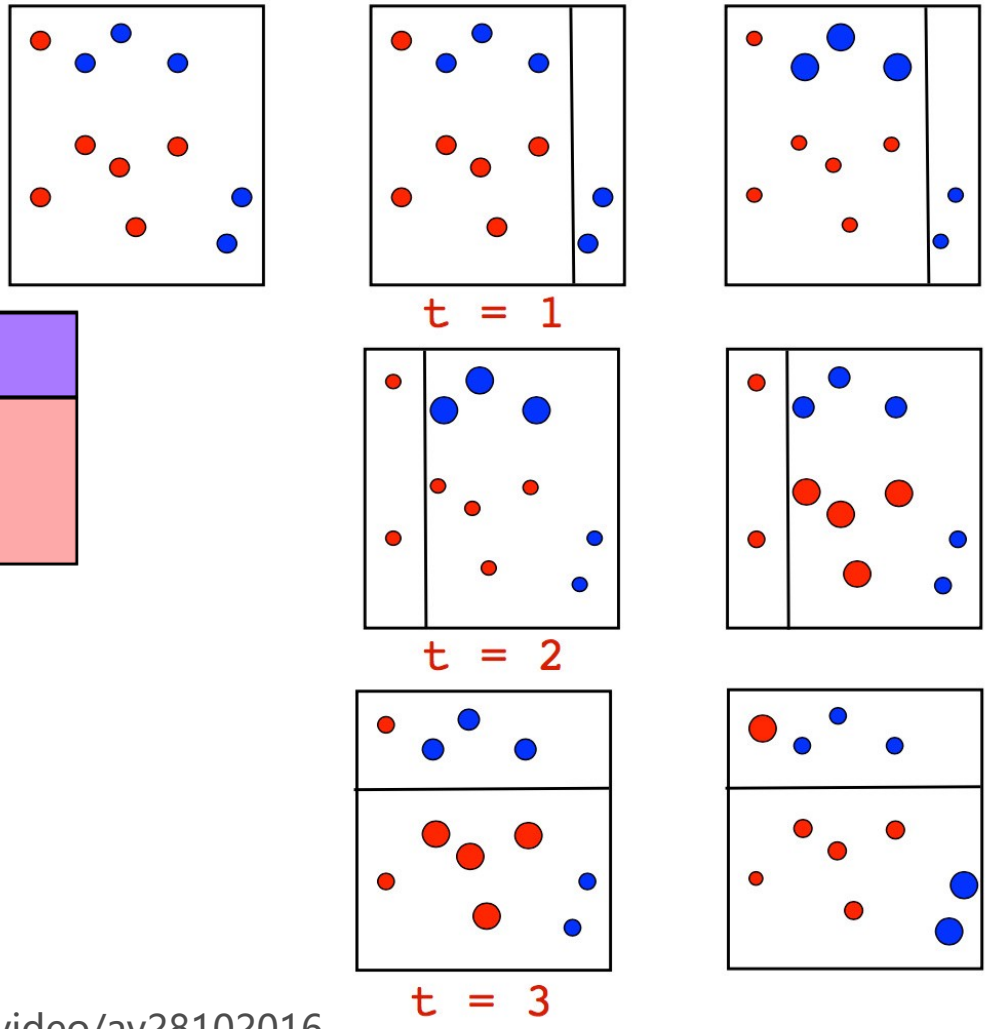
https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/springerEBR09.pdf

集成学习

- **Boosting**







Bosting25年, 周志华, https://www.bilibili.com/video/av28102016

集成学习

- **工具**
  - Sklearn.ensemble
  - 自己手动写
  - https://github.com/ikki407/stacking
  - https://github.com/MLWave/Kaggle-Ensemble-Guide
  - https://github.com/flennerhag/mlens

参数调优

- **调调调**
  - 祖传参数
  - 经验调参
  - 借助工具
- **方法**
  - 网格搜索
  - 随机搜索、贝叶斯优化

```
local_params = {
        'max_depth' : -1,
        'nthread': 12, # Updated from nthread
        'num_leaves': 64,
        'learning_rate': 0.05,
        'max_bin': 512,
        'subsample_for_bin': 200,
        'subsample': 1,
        'subsample_freq': 1,
        'colsample_bytree': 0.8,
        'reg_alpha': 5,
        'reg_lambda': 10,
        'min_split_gain': 0.5,
        'min_child_weight': 1,
        'min_child_samples': 5,
        'scale_pos_weight': 1,
        'num_class' : 1,
}
```

数据科学三步走

- 第二步：宝刀屠龙，百战百胜


实战：做一场评测

11/29/2018
汶东震

前言

- **个人能力的数据竞赛**
  - 洞察力、执行力、知识能力
- **团队协作的数据竞赛**
  - 一次数据竞赛是一个目标明确的短期项目
  - 一次成功的比赛取决于个人能力和团队协作
  - 做好项目管理有利于成员更大发挥各自优势
  - 做好一把锤子，到处都有钉子

组队

寻找队友、构建组织

寻找队友

- **线下**
  - 同学、同门、同届、同校
- **线上**
  - 数据竞赛社区：kaggle、datafountain
  - 比赛交流群：报名比赛，寻找合作
  - 邮件联系开源方案、博客作者

找到队友

- **合适的队友**
  - 能力相近（大腿很好、不要太粗）
  - 各有所长、对问题有实际经验
  - 交叉背景，复合型人才
  - 能够沟通交流

- **不合适的队友**
  - 摸鱼+心不在焉==团队毁灭者

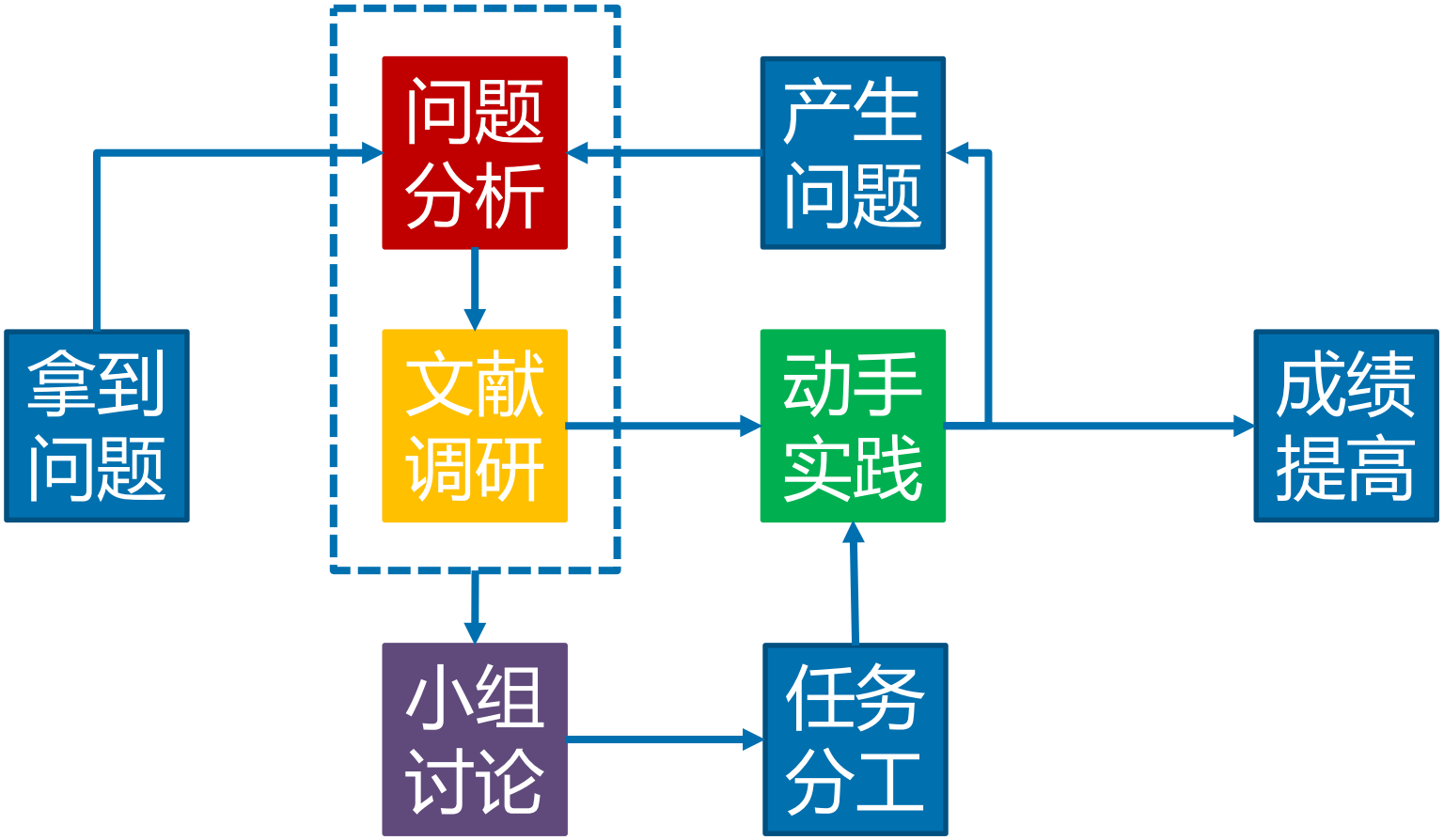
构建组织

- **一个QQ群**
  - 文件共享（数据、代码片段、学习资料）
  - 及时交流（线上开会）
- **一个私有的Git仓库**
  - 代码记录、追踪进度
  - 方便协作、规范协作
  - 随时记录、随时回滚
- **在线文档**
  - todo list、记录idea
  - 便于跟踪总体进度
  - 便于赛后复盘总结

项目管理

高水平、高要求

流程管理



工程管理

- **目录结构一致**
  - 多人共享、方便协作
  - 结构清晰、避免混乱
  - 方便后期整理代码

```
|--evaluation
    |--data
        |--raw_data
        |--processed_data
        |--feature_data
        |--other_data
        |--submission
    |--notebook_user1
        |--util_code
        |--1.process_notebook.ipynb
        |--2.analyze_notebook.ipynb
        ...
        ...
    |--notebook_user2
    ...
    ...
```

工程管理

- **命名方式一致**
  - 直观、明确
  - 包含必要信息

```
c_onehot_1gram_tsne_1d_interaction_df.pkl
c_onehot_1gram_tsne_2d_interaction_df.pkl
c_onehot_1gram_tsne_3d_interaction_df.pkl
c_onehot_2gram_tsne_1d_interaction_df.pkl
c_onehot_2gram_tsne_2d_interaction_df.pkl
c_onehot_2gram_tsne_3d_interaction_df.pkl
c_onehot_3gram_tsne_1d_interaction_df.pkl
c_onehot_3gram_tsne_2d_interaction_df.pkl
c_onehot_3gram_tsne_3d_interaction_df.pkl
```

  - 后缀表明文件格式（序列化方式）
- **数据文件、代码文件均需要清晰命名**

工程管理

- **数据格式一致**
  - 方便程序引用
  - 方便多人交互
  - 统一序列化格式
  - 推荐pandas.Dataframe

	qid1	qid2	label	q1_w_list
0	S107641	S103854	0.0	W105587 W101644 W102193 W106548 W104416
1	S103127	S110857	1.0	W106907 W107170 W100323 W101855 W104806
2	S120534	S135254	1.0	W108463 W108376 W101396 W107941 W104416 W106503
3	S102044	S106012	0.0	W100914 W102600 W104740 W106503 W102627 W10052...
4	S134333	S109235	0.0	W107170 W102193 W102254 W109339 W101995 W104416

人员管理

- **层次与梯队**
  - 队长：制定timeline、构建框架、监督项目
  - 技术帝：复杂模型攻坚、原型开发
  - 参与者： 调研、特征、筛选、调参、集成
- **各司其职**

进度管理

- **关键时间点**
  - 报名时间、数据发放时间、决赛提交时间
- **项目推进时间规划**
  - 一个训练测试周期所需时长预估
  - 流程分解，并行推进
- **单模型攻坚 vs 集成模型保底**

参赛策略

调研翔实、代码扎实

调研翔实

- **“文献”调研**
  - 相似比赛的开源方案
  - Kaggle社区相关分享
  - 解决问题相关论文
  - Github上现有代码
  - 师兄“祖传”解决方案

唯快不破

- **快速构建Baseline模型**
  - 搭建解决方案原型 有所对比
- **快速复现主流解决方案**
  - 优先寻找开源代码进行调试
- **快速迭代项目解决方案**
  - 优化代码各部分效率
  - 快速组合、运行、评估

代码扎实

- **最后一步、编写代码**
  - 分析理解问题更重要
- **开源代码、吃透再用**
  - 兼容你自己的代码体系
  - 避免精度误差向后扩散
- **构建用例、单元测试**
  - 科研代码也需要测试

其他建议

- 保持精力、少熬夜
- 策略性提交、多关注榜单
- 队友和睦、万事太平
- 记录每次提交结果，比对线上线下差异
- 记录Roadmap，随时跟进进度
- 有条件的话可以使用小号

数据科学三步走

-
-
- **第三步：披荆斩棘、共创佳绩**

考核方式

11/7/2019

汶东震

作业一

在线评测

- http://ir.dlut.edu.cn/OnlineJudge/

作业一

- **在线评测算法报告（建议内容）**
  - 对问题的分析
  - 数据分析、问题理解
  - 数据处理、特征提取
  - 机器学习、深度学习算法模型
  - 算法使用心得

*本页内容仅作为报告内容的指导性建议，并非绝对评判标准

作业二

- **自然语言处理领域综述**
  - 选择NLP、IR、TM领域感兴趣方向撰写综述
- **综述报告内容**
  - 研究领域简介（做什么？解决什么问题？）
  - 相关数据集（至少3个，对数据集进行介绍）
  - 当前主流研究方案（SOTA，介绍工作）
- **字数要求：5000~6000字（符号）**

*详细模版后续在群里给出

作业要求

- **本科生**
  - 垃圾邮件分类+领域综述报告
  - 在线评测选做（额外加分）
- **研究生**
  - 垃圾邮件分类+领域综述报告
  - 在线评测二选一（必须选择一个）
  - 在线评测算法报告

作业验收

- **本科生**
  - 邮件分类在线成绩
  - 领域综述报告

一份报告
一份在线成绩
如果选做在线评测，
需附上对应算法设计报告
方可加分

- **研究生**
  - 邮件分类在线成绩
  - 在线评测（二选一）成绩+ 算法设计报告
  - 领域综述报告

两份报告
两份在线成绩

***垃圾邮件分类任务无需撰写报告

相关数据

- **垃圾邮件分类**
  - 任务：http://ir.dlut.edu.cn/OnlineJudge/Detail/3
  - 链接: https://pan.baidu.com/s/1V-5RAFqcnPsZKW_iSP0WJw
  - 提取码: 8zaq
- **SMP 2016 用户画像**
  - 任务：http://ir.dlut.edu.cn/OnlineJudge/Detail/2
  - 数据&代码：https://github.com/liyumeng/SmpCup2016

报告要求

- **统一使用论文模版**
  - 大连理工大学硕士论文模版
  - http://gs.dlut.edu.cn/yjspy/xwgl/lwmb1.htm
- **填写清楚个人信息**
  - 姓名、班级、学号
  - 铅笔注明本科or硕士
  - 联系电话或邮箱

作业提交

- **最后时限**
  - 2019年12月29日（星期日）
- **提交方式**
  - 报告电子版：dutir_irtm@163.com
  - 邮件标题格式：11809000 东震 博1800
  - 评测成绩在线提交，榜上有名
  - 纸质报告：创新园大厦A923

其他

- 如对课程有意见、建议、投诉
- 请联系邮箱：

irlab@dlut.edu.cn

谢谢！