深度之眼
deepshare.net

NLP文本分类挑战赛 模型集成&可视化

导师：阿水

深度之眼
deepshare.net

目录

深度之眼
deepshare.net

1 模型集成

Model Ensemble

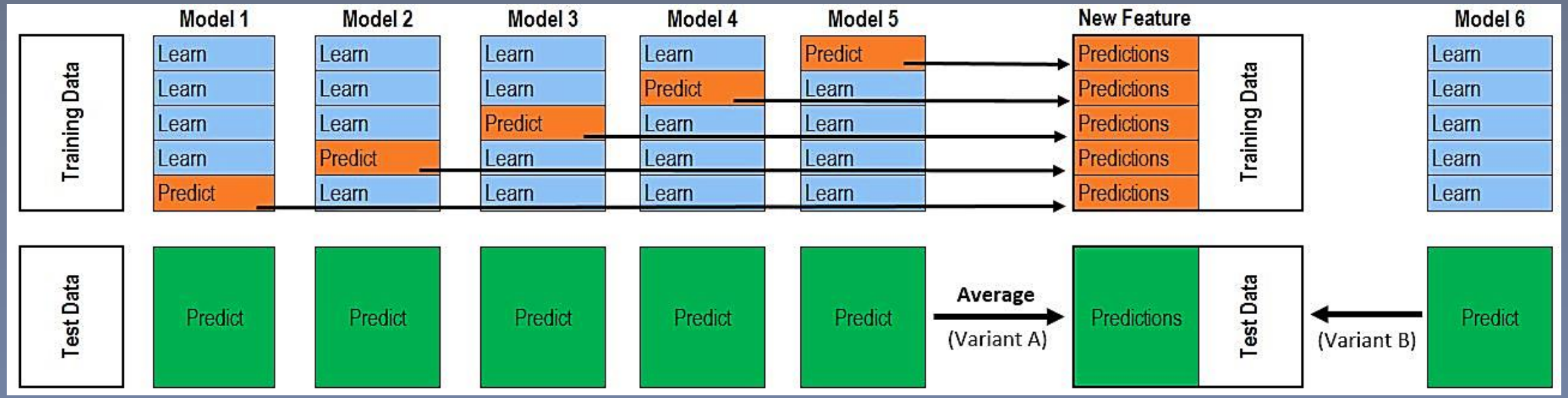
1 模型集成

Model Ensemble

Stacking

在交叉验证的过程中对模型进行多折训练，对训练集和测试集统计进行预测；

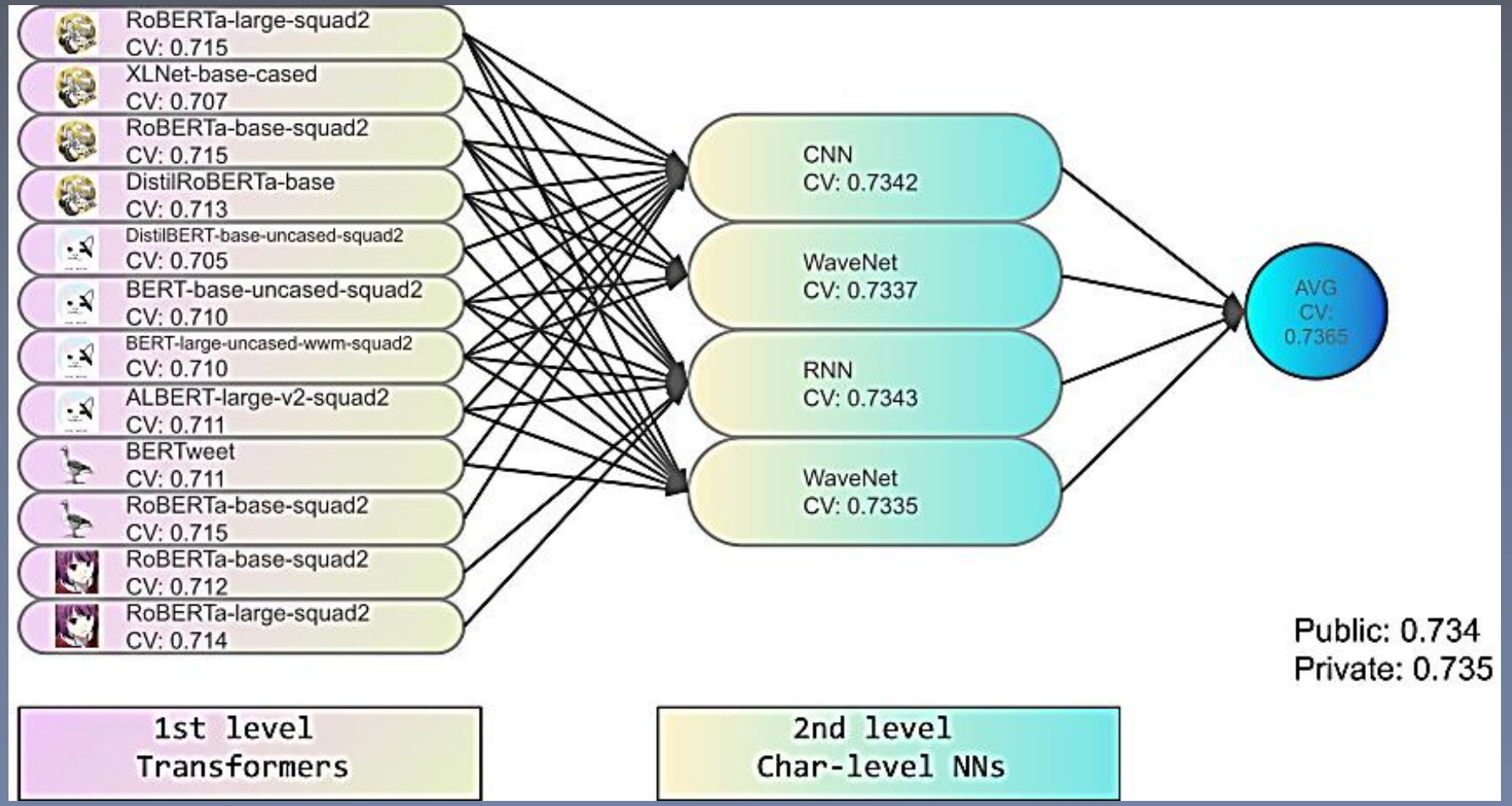
out of fold可以用来进行增加特征，也可以用来进行stacking；



https://www.kaggle.com/artgor/using-meta-features-to-improve-model#Feature-generation

1 模型集成

Model Ensemble



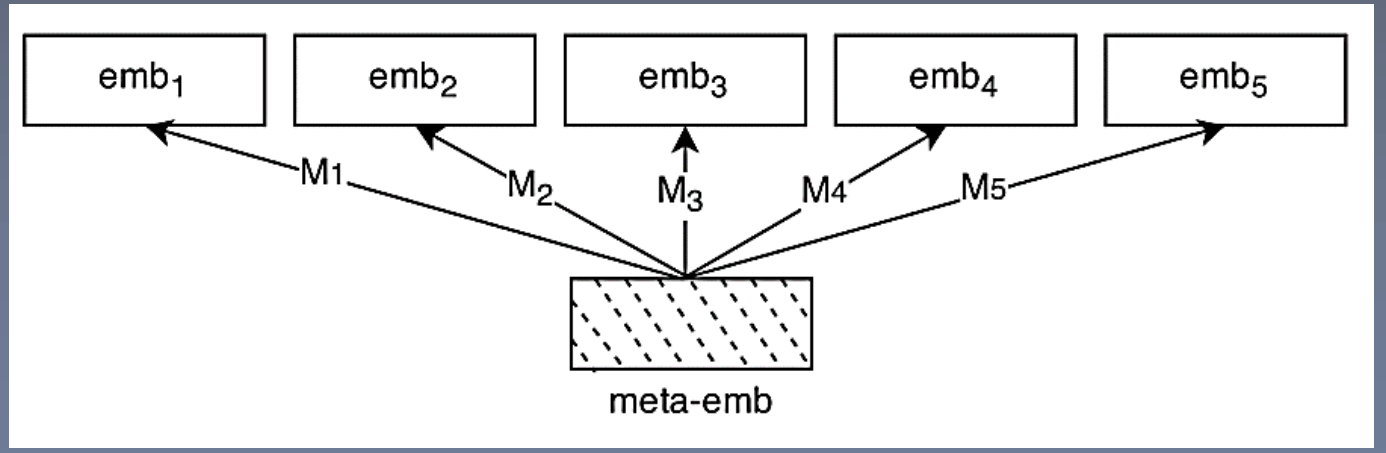
https://www.kaggle.com/c/tweet-sentiment-extraction/discussion/159477

1 模型集成

Model Ensemble

□ 词向量集成

- ✓ 对词向量进行平均（Blend）
- ✓ 对词向量进行拼接（Concat）
- ✓ 对词向量进行学习（Meta）



https://www.kaggle.com/c/quora-insincere-questions-classification/discussion/71778/
https://www.aclweb.org/anthology/K18-1028/

1 模型集成

Model Ensemble

Test Time Augmentation（TTA）

- ✓ 预测句子的开始maxlen 字符；
- ✓ 预测句子的开始maxlen/2字符 + 末尾maxlen/2字符；
- ✓ 将句子按照长度进行拆分；

深度之眼
deepshare.net

2 伪标签

Pseudo Label

2 伪标签

Pseudo Label

伪标签： 将测试集进行标注并进行有效训练；

- ✓ 步骤1：根据训练集训练模型；
- ✓ 步骤2：对测试集样本进行预测；
- ✓ 步骤3：将测试集的训练集一起训练；

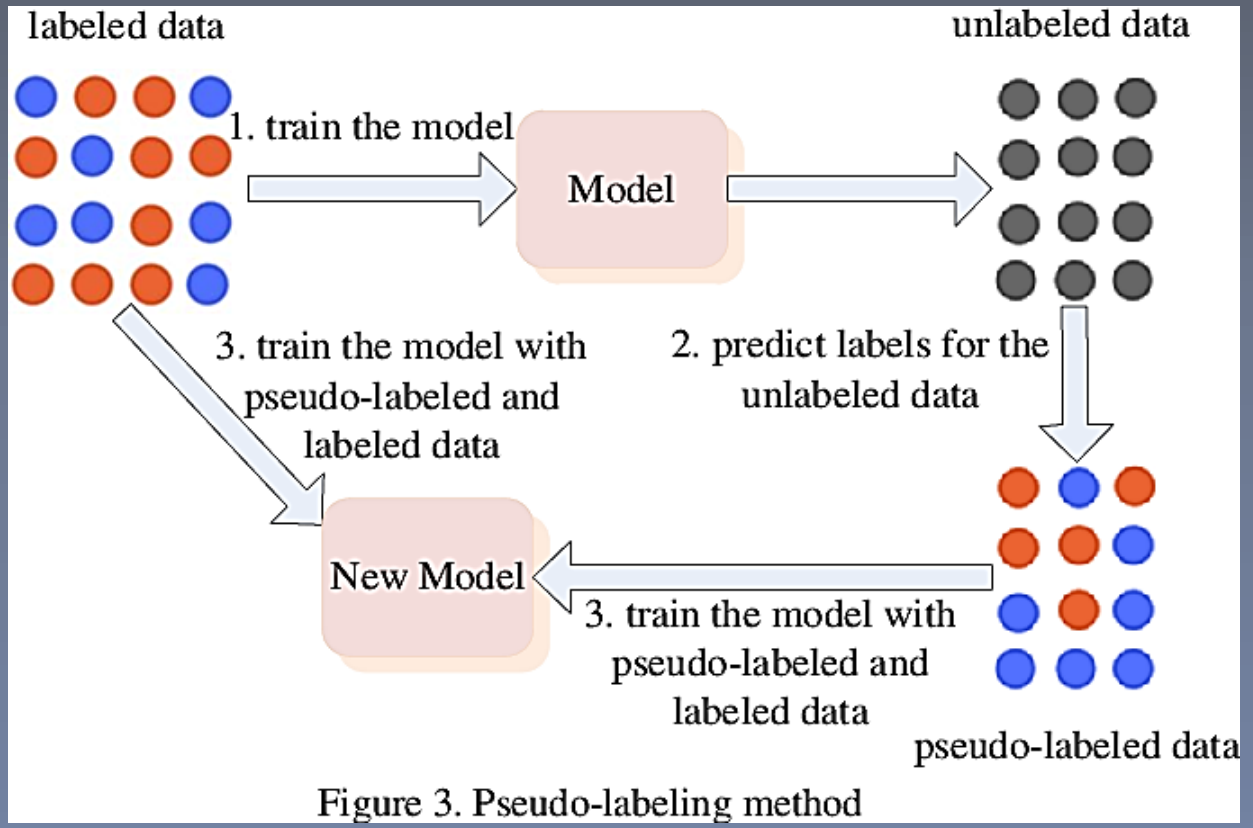


Figure 3. Pseudo-labeling method

2 伪标签

Pseudo Label

伪标签：

- ✓ 并不是所有的场景都适用，一般适用于分类；
- ✓ 分类空间更小，且能够配合soft label同时使用；
- ✓ 伪标签优先将置信度高的样本加入训练；

案例：

https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/discussion/160862

https://www.kaggle.com/c/tweet-sentiment-extraction/discussion/159477

2 伪标签

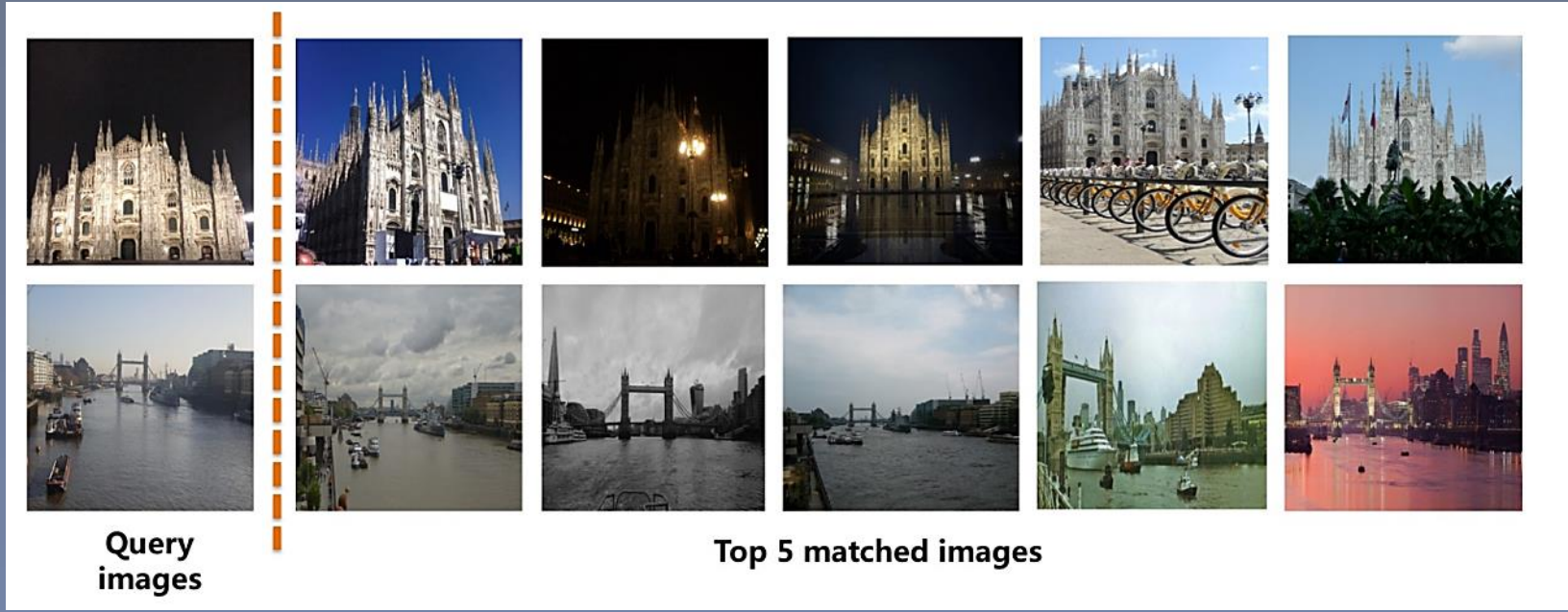
Pseudo Label

□ Google Landmark Retrieval Challenge

https://www.kaggle.com/c/landmark-retrieval-challenge/

赛题介绍：构建一个地标图像检索系统；

赛题难点：赛题数据量大，需要有效进行训练；



2 伪标签

Pseudo Label

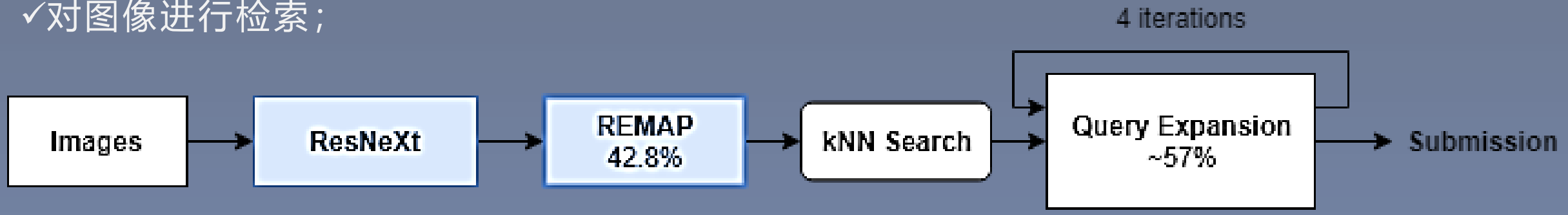
□ Google Landmark Retrieval Challenge

赛题思路：

✓提取图像特征；

✓对图像进行索引；

✓对图像进行检索；



深度之眼
deepshare.net

3 模型可视化

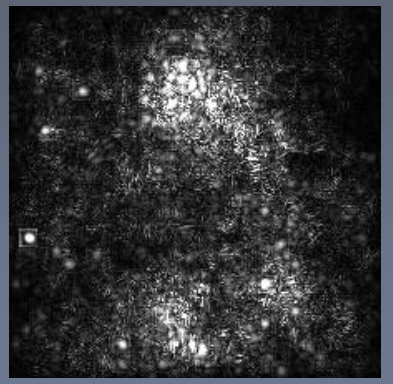
Model Visualization

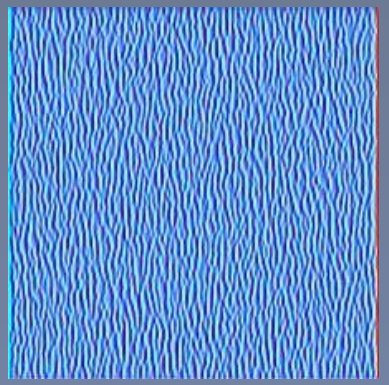
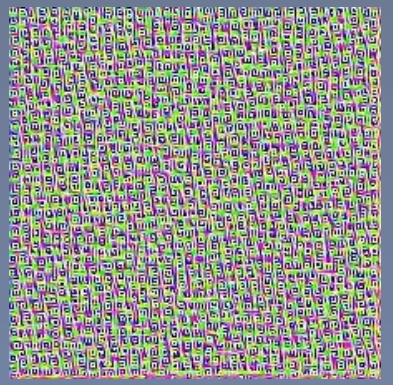
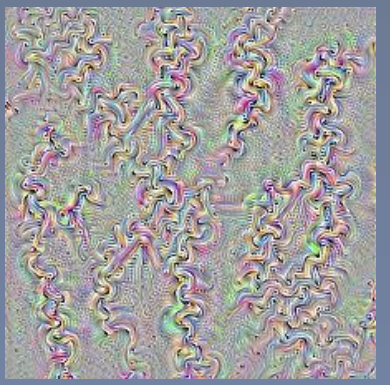
3 模型可视化

Model Visualization

模型可视化：

- ✓ 可视化模型决策；
- ✓ 可视化模型梯度；
- ✓ 可视化模型参数；

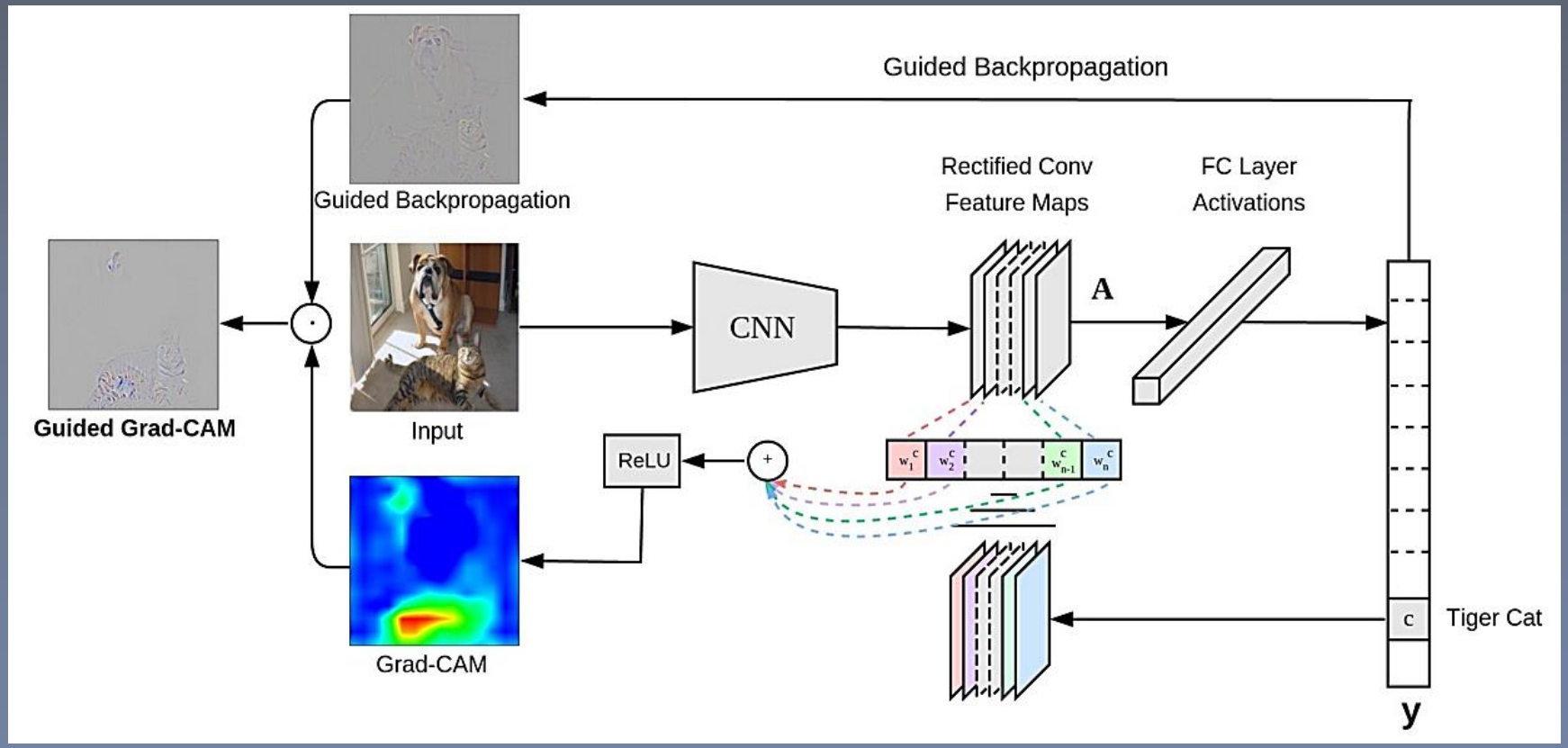




3 模型可视化

Model Visualization

思路1：

通过梯度可视化



https://arxiv.org/pdf/1610.02391.pdf

3 模型可视化

Model Visualization

思路2：

对输入进行掩码；

```
import eli5
from eli5.lime import TextExplainer

te = TextExplainer(random_state=42)
te.fit(doc, pipe.predict_proba)
te.show_prediction(target_names=twenty_train.target_names)
```

y=alt.atheism (probability **0.000**, score **-9.663**) top features

Contribution?	Feature
-0.360	<BIAS>
-9.303	Highlighted in text (sum)

as i recall from my bout with kidney stones, there isn't any medication that can do anything about them except relieve the pain. either they pass, or they have to be broken up with sound, or they have to be extracted surgically. when i was in, the x-ray tech happened to mention that she'd had kidney stones and children, and the childbirth hurt less.

https://eli5.readthedocs.io/en/latest/tutorials/black-box-text-classifiers.html

3 模型可视化

Model Visualization

思路3：

通过shap值计算；

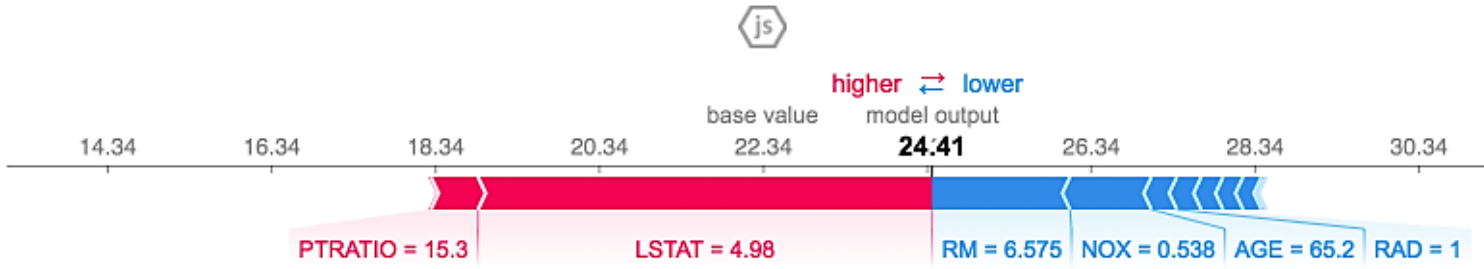
```python
import xgboost
import shap

# load JS visualization code to notebook
shap.initjs()

# train XGBoost model
X,y = shap.datasets.boston()
model = xgboost.train({"learning_rate": 0.01}, xgboost.DMatrix(X, label=y), 100)

# explain the model's predictions using SHAP
# (same syntax works for LightGBM, CatBoost, scikit-learn and spark models)
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X)

# visualize the first prediction's explanation (use matplotlib=True to avoid Javascript)
shap.force_plot(explainer.expected_value, shap_values[0,:], X.iloc[0,:])
```



https://github.com/slundberg/shap

深度之眼
deepshare.net

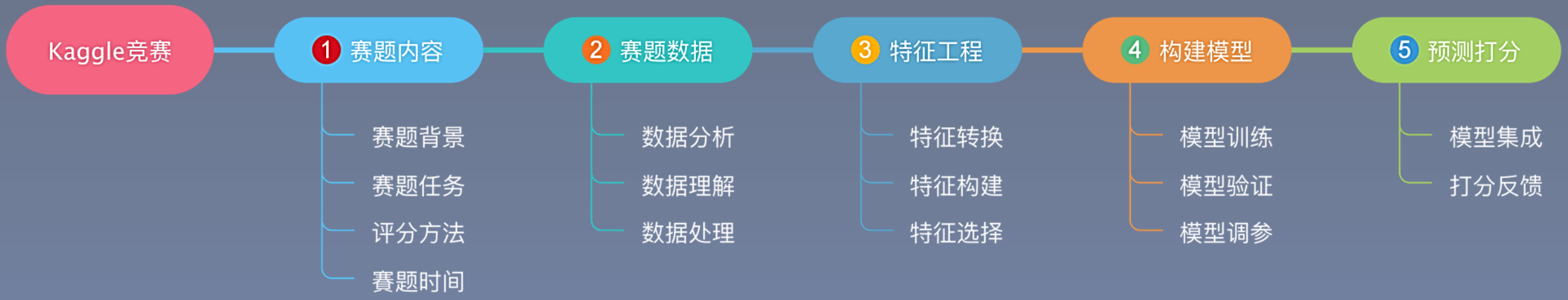
4 知识点总结

Pseudo Label

4 知识点总结

All about Class

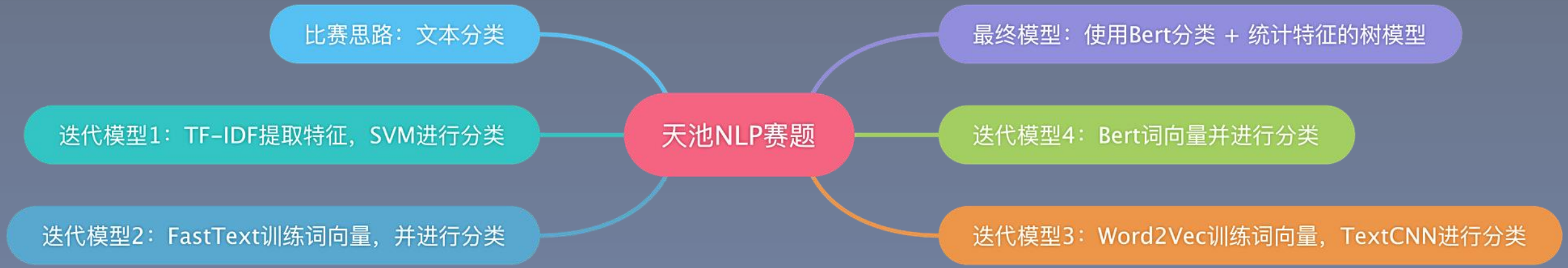
1、比赛流程



4 知识点总结

All about Class

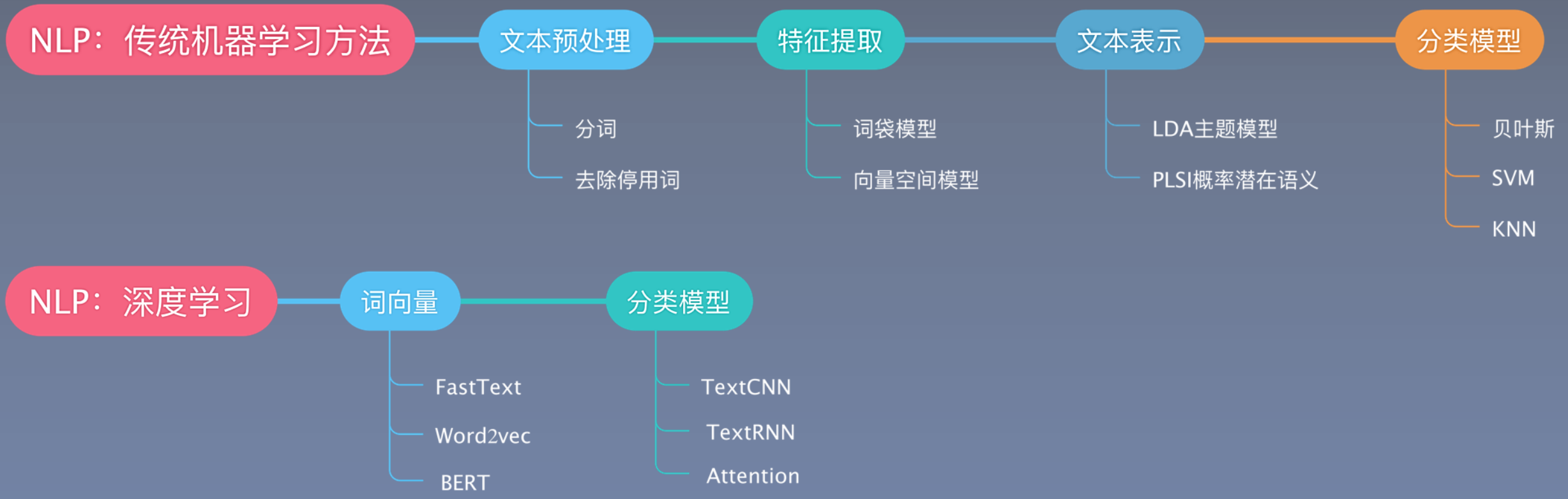
2、文本分类流程



4 知识点总结

All about Class

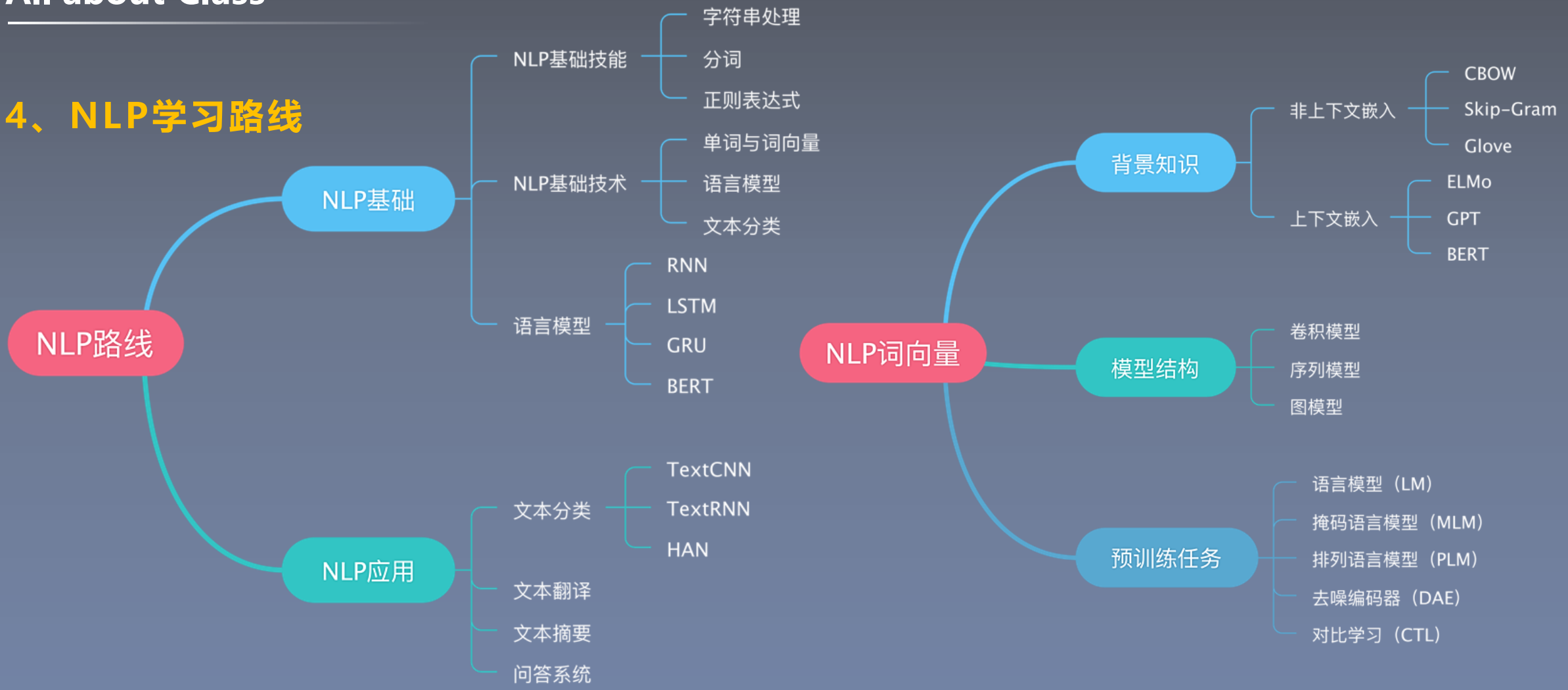
3、NLP发展路线



4 知识点总结

All about Class

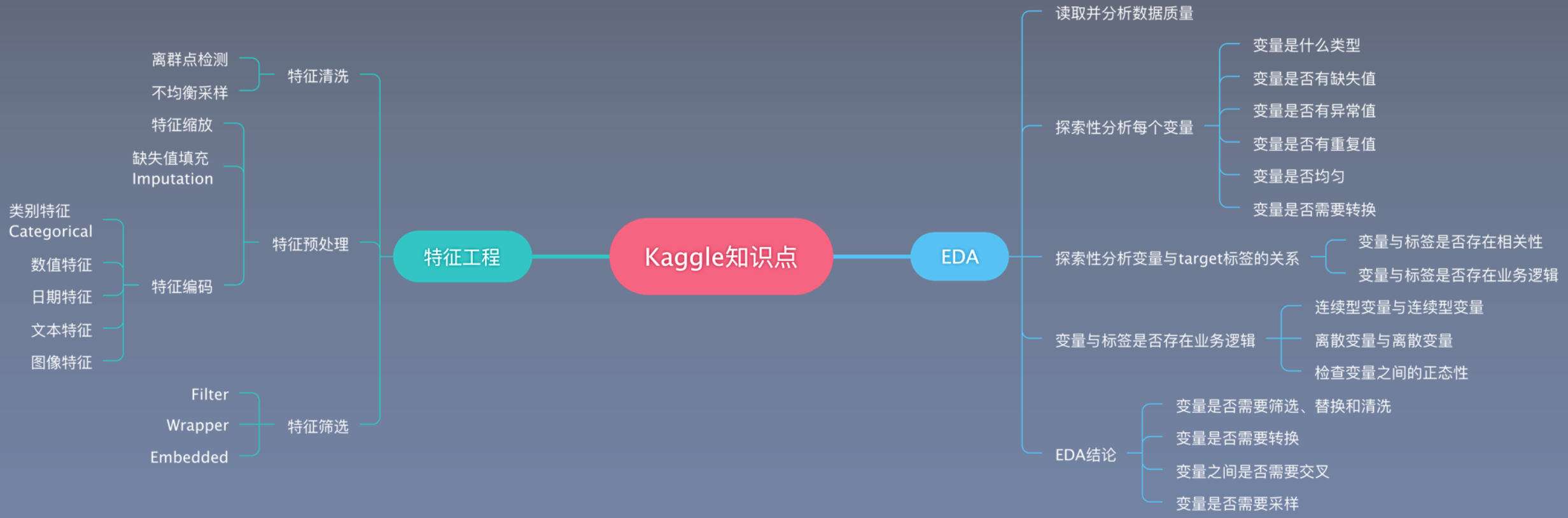
4、NLP学习路线



4 知识点总结

All about Class

本次比赛只是一小步，还有很多知识点~



深度之眼
deepshare.net

5 Q&A

Ask me anything

5 Q&Q

Ask me anything

1、Null importance

https://www.kaggle.com/ogrellier/feature-selection-with-null-importances

2、Stacking

https://www.kaggle.com/amiiiney/price-prediction-regularization-stacking

3、Pseudo Label

https://www.kaggle.com/nvnnghia/yolov5-pseudo-labeling

请让我们一起立一个flag！

我承诺：

4周努力上TOP100！

结语

再小的细节，也值得被认真对待

深度之眼
deepshare.net

联系我们：

电话：18001992849

邮箱：service@deepshare.net

Q Q： 2677693114

公众号

客服微信