深度之眼
deepshare.net

NLP文本分类挑战赛竞赛流程&Pipline

导师：阿水

深度之眼
deepshare.net

目录

深度之眼
deepshare.net

1 竞赛流程

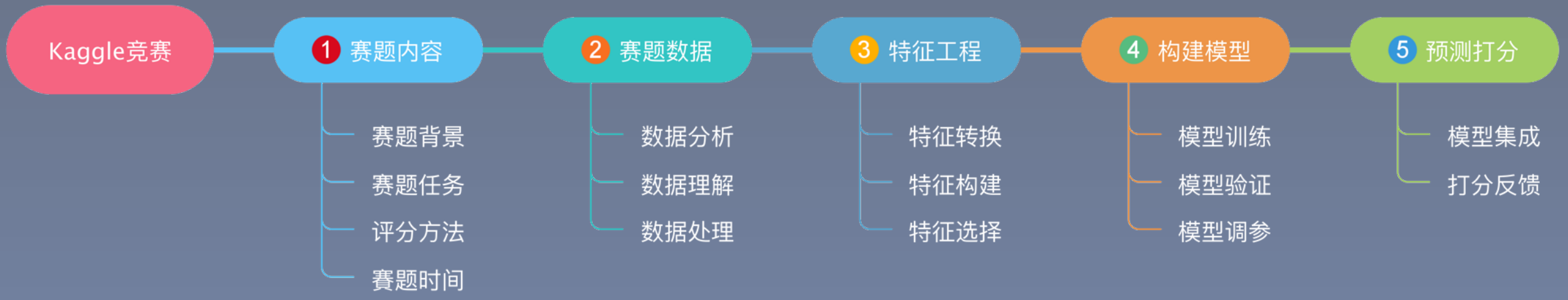
Competition Pipeline

1 竞赛流程

Competition Pipeline

解题流程：流程化、规范化

思考：你做比赛有什么习惯呢？

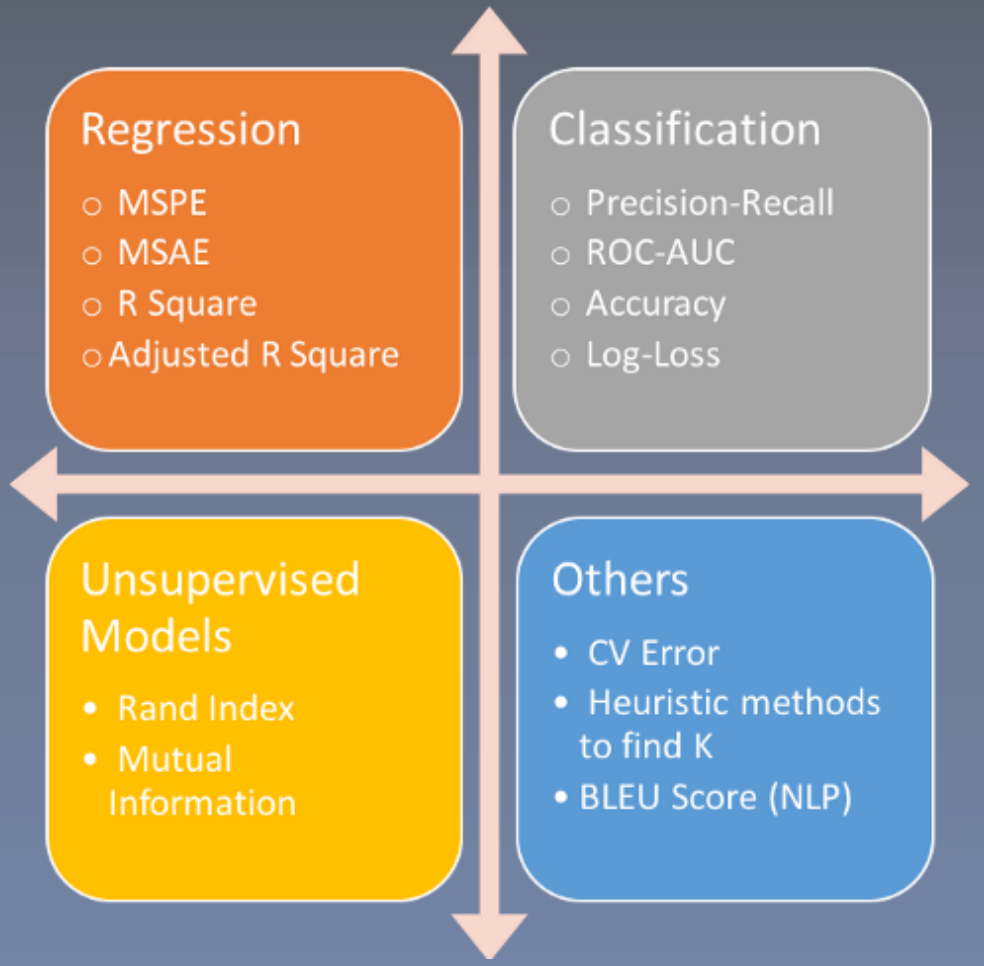


1 竞赛流程

Competition Pipeline

赛题类型可以根据背景、**赛题任务**和**赛题数据**进行分类：

- ✓ 结构化赛题
- ✓ 图像赛题
- ✓ 文本赛题
- ✓ 语音赛题

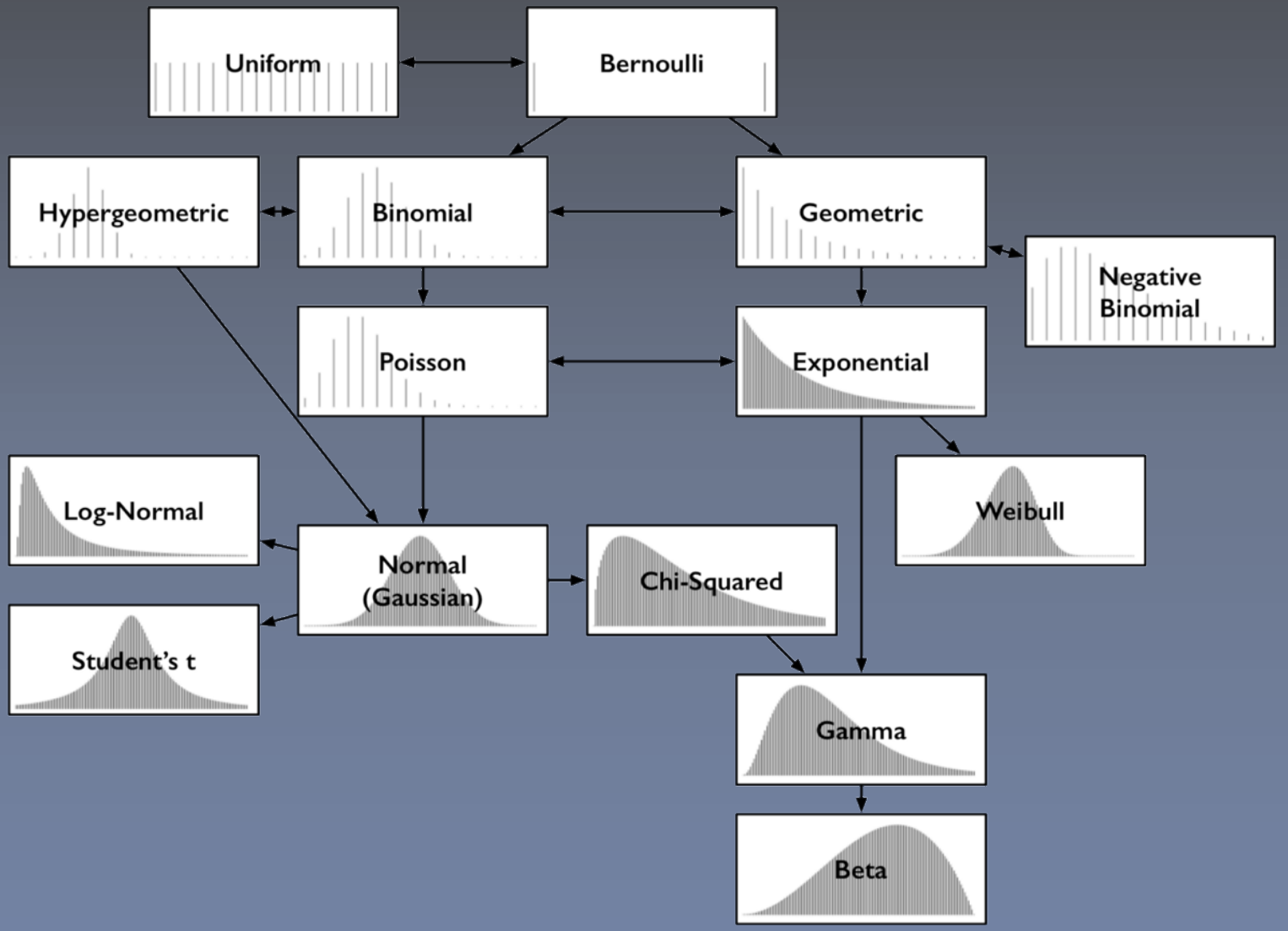


1 竞赛流程

Competition Pipeline

赛题数据分析：单个字段的分析；

- ✓ 类别变量；
- ✓ 数值变量；
- ✓ 时序变量；

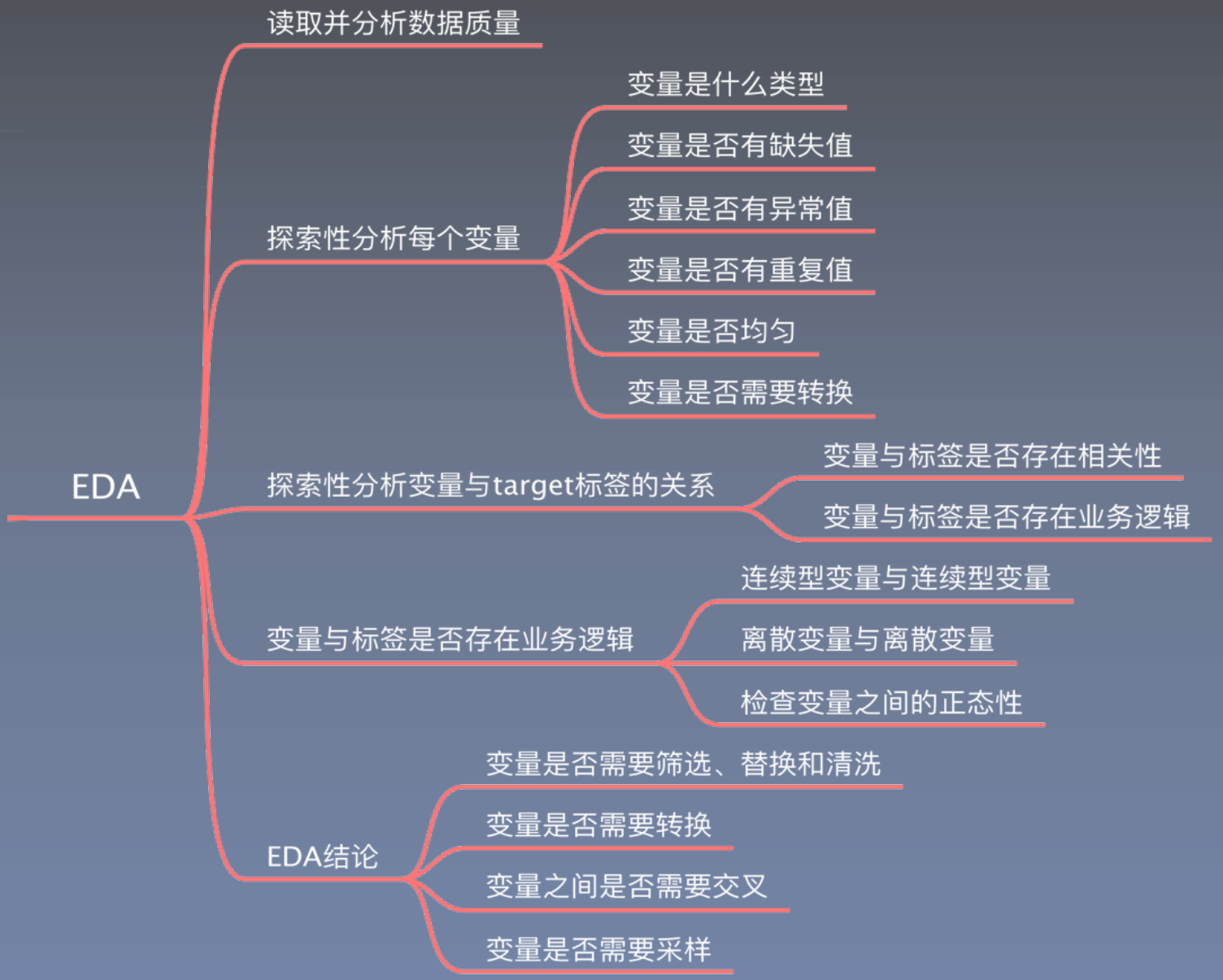


1 竞赛流程

Competition Pipeline

数据分析思路：

- ✓ 分析单个变量；
- ✓ 分析多个变量；

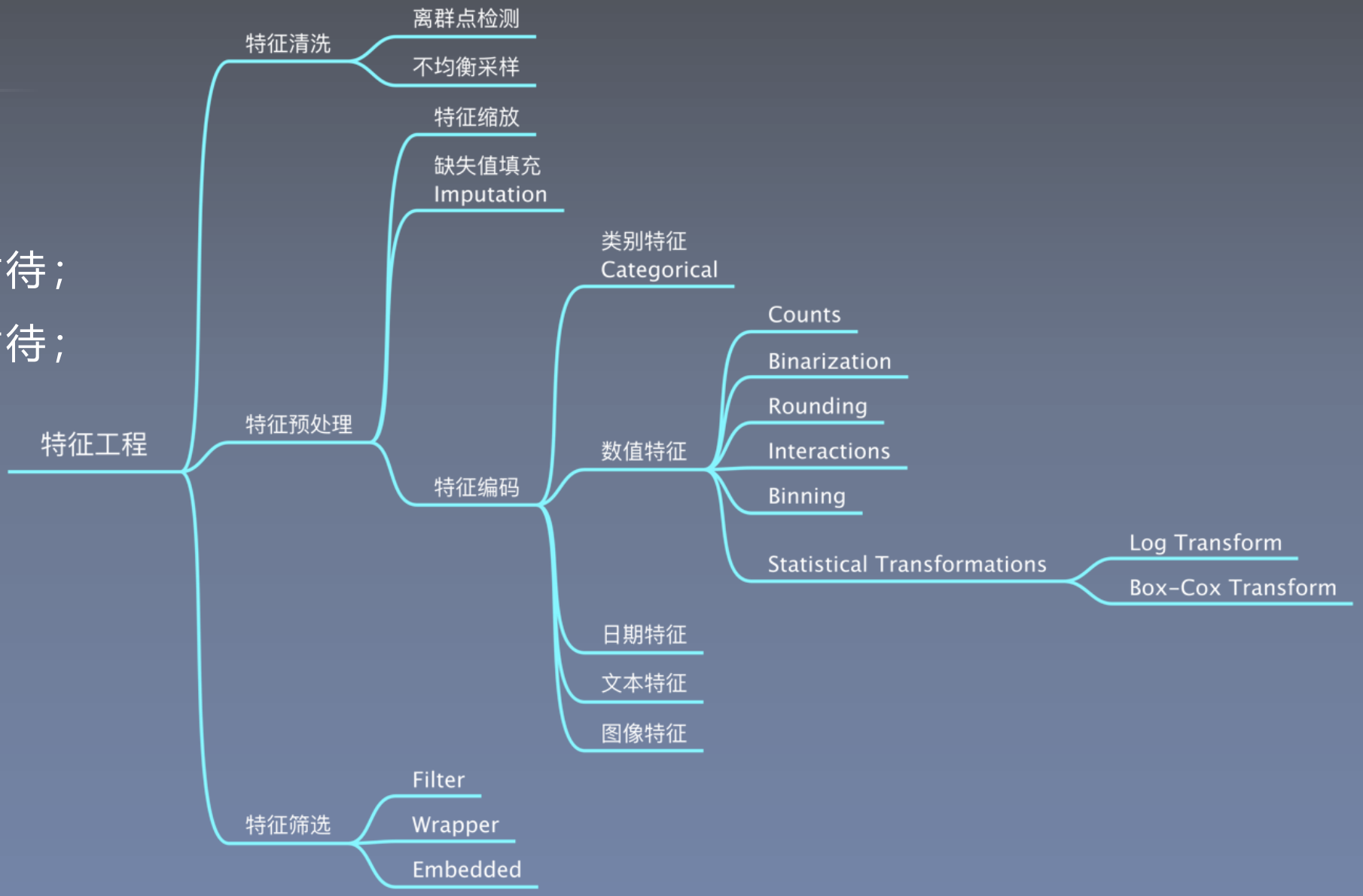


1 竞赛流程

Competition Pipeline

特征工程：

- ✓ 不同类型数据特殊对待；
- ✓ 不同类型赛题特殊对待；

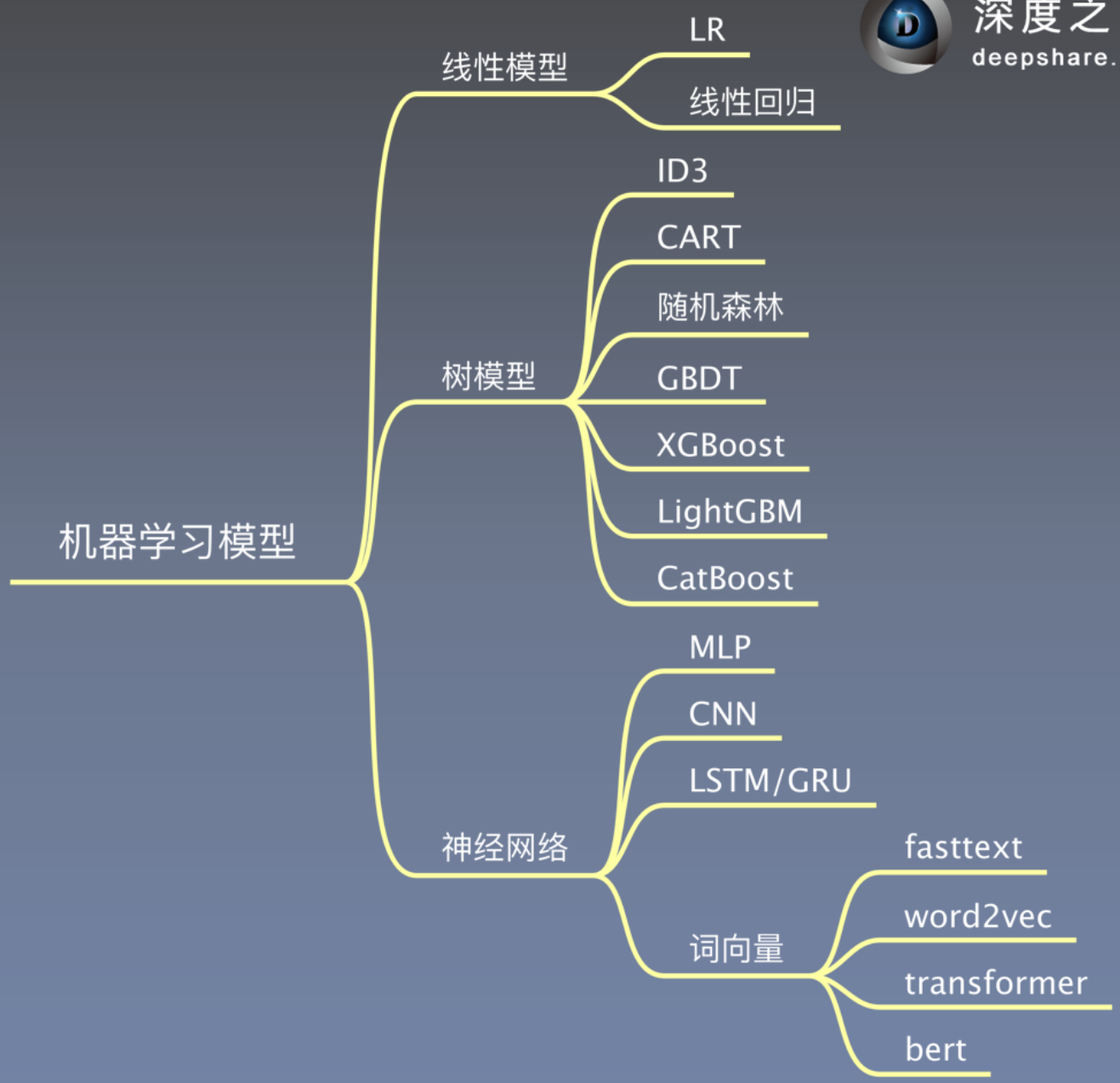


1 竞赛流程

Competition Pipeline

机器学习模型：

- ✓ 不同的模型有不同的偏好；
- ✓ 结构化数据优先考虑树模型；
- ✓ 非结构化数据优先考虑深度学习；

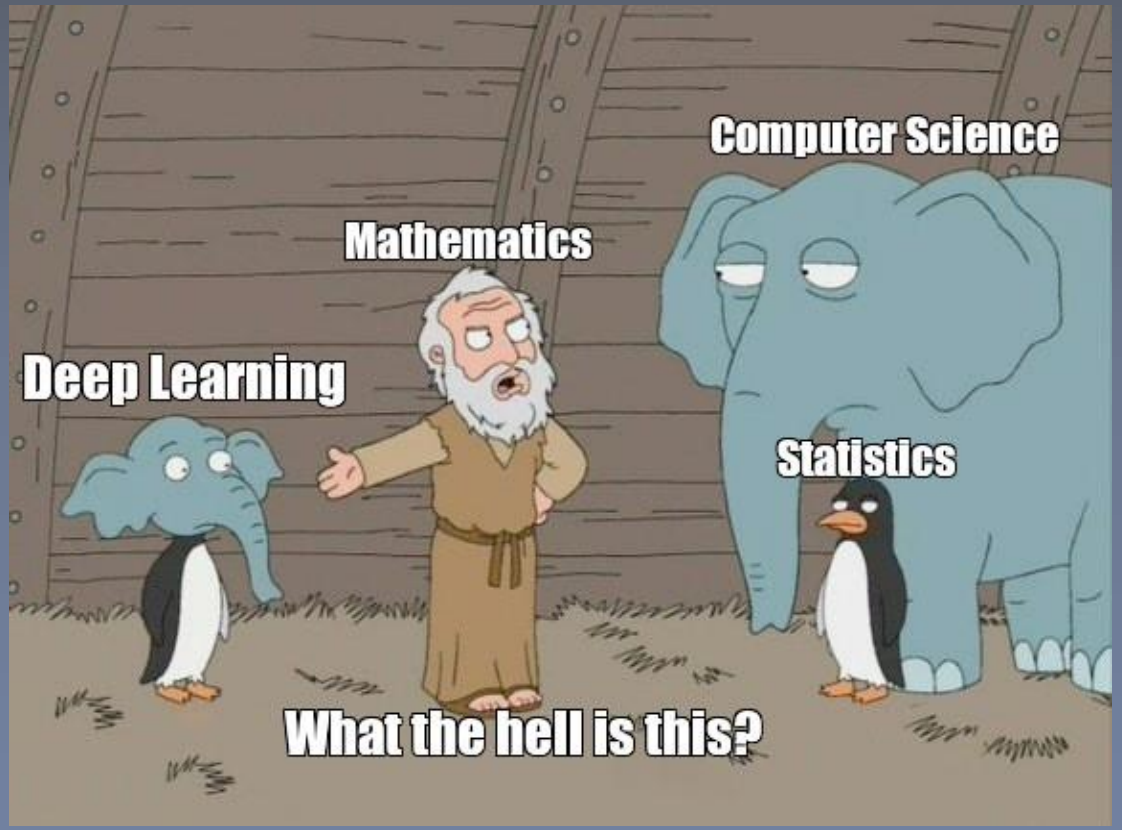


1 竞赛流程

Competition Pipeline

机器学习模型：

- ✓ 不同的模型有不同的偏好；
- ✓ 结构化数据优先考虑树模型；
- ✓ 非结构化数据优先考虑深度学习；

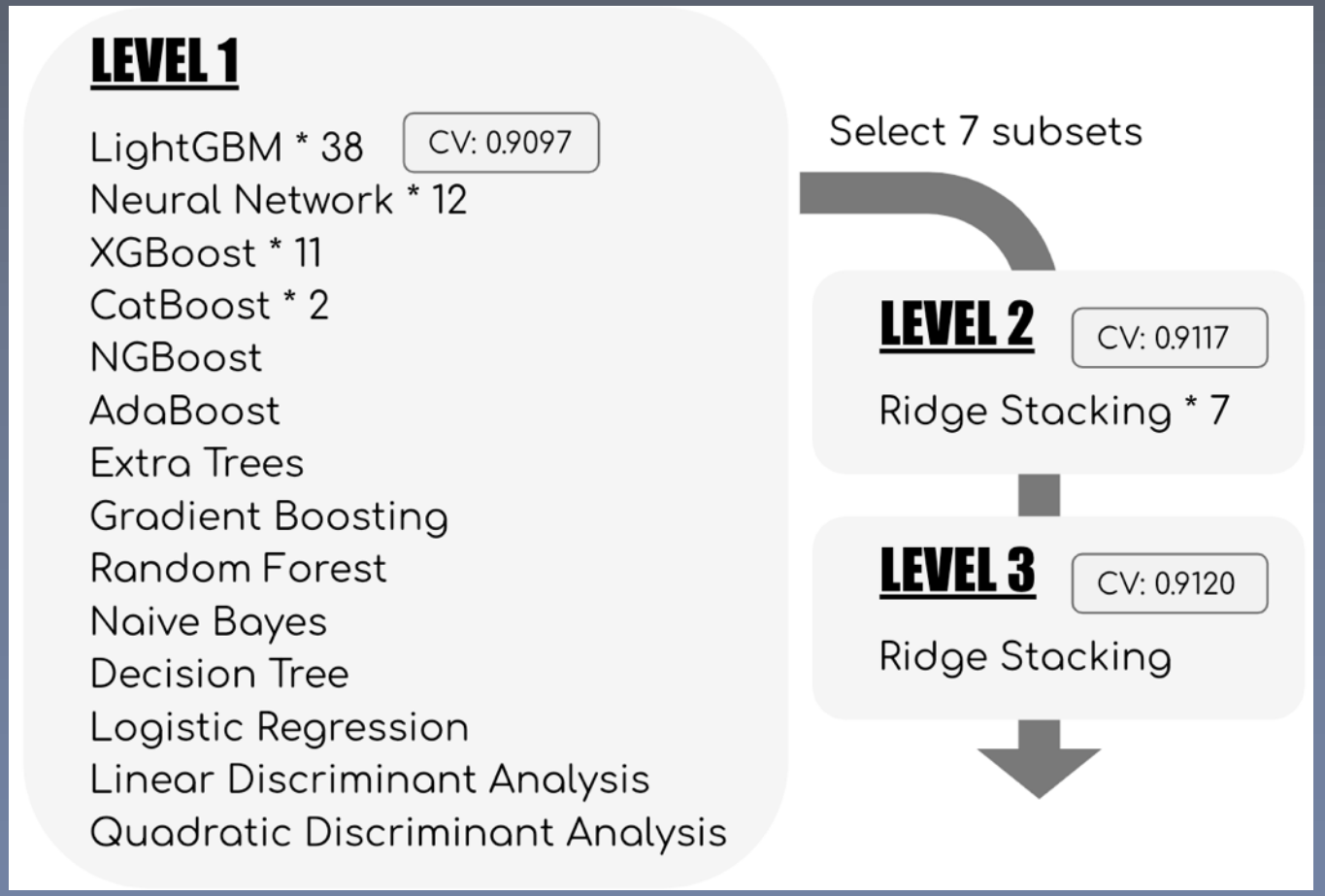


1 竞赛流程

Competition Pipeline

模型集成：

- ✓ Vote
- ✓ Blend
- ✓ Stacking



1 竞赛流程

Competition Pipeline

吾日三省吾身：会了？懂了？上分了？

1 竞赛流程

Competition Pipeline

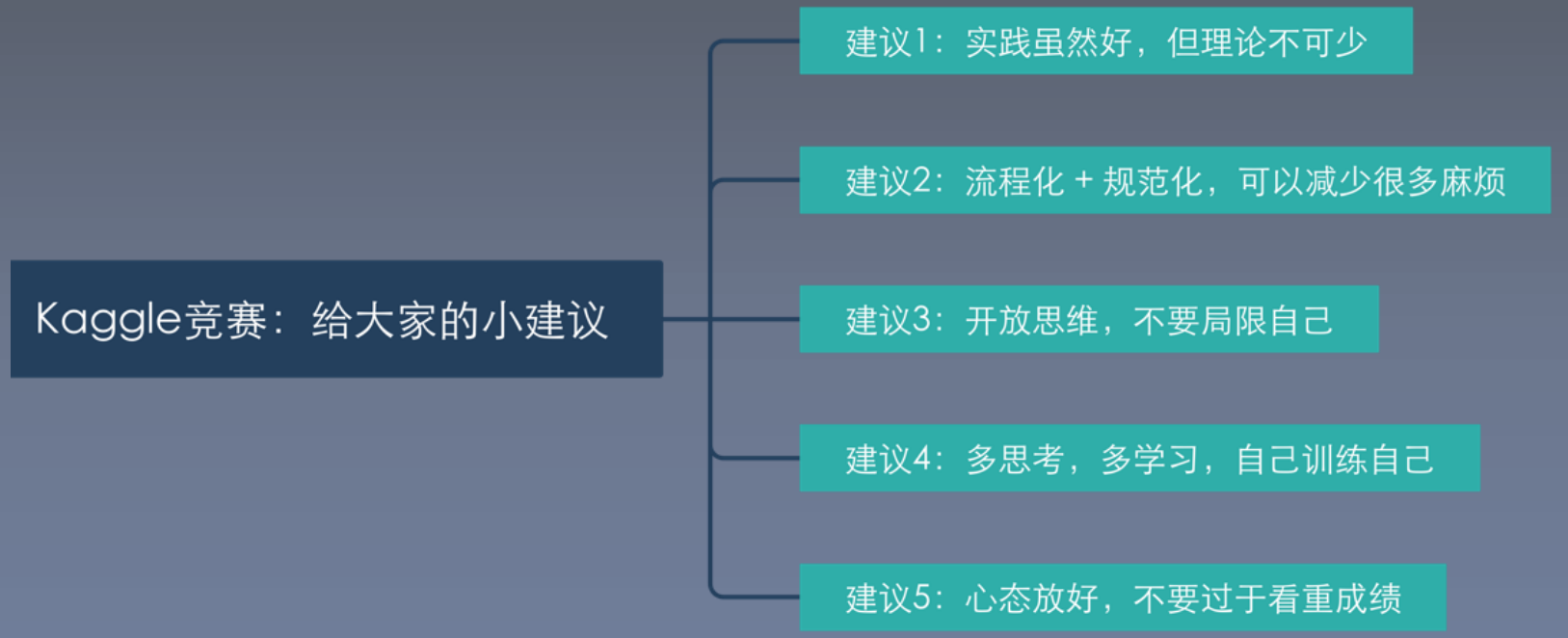
阿水给大家的**小建议**：

✓特征工程（不同特征）；

✓机器学习模型（不同模型）；

✓模型超参数（不同参数）；

✓模型集成；



深度之眼
deepshare.net

2 模型训练与验证

Train and Validation

2 模型训练与验证

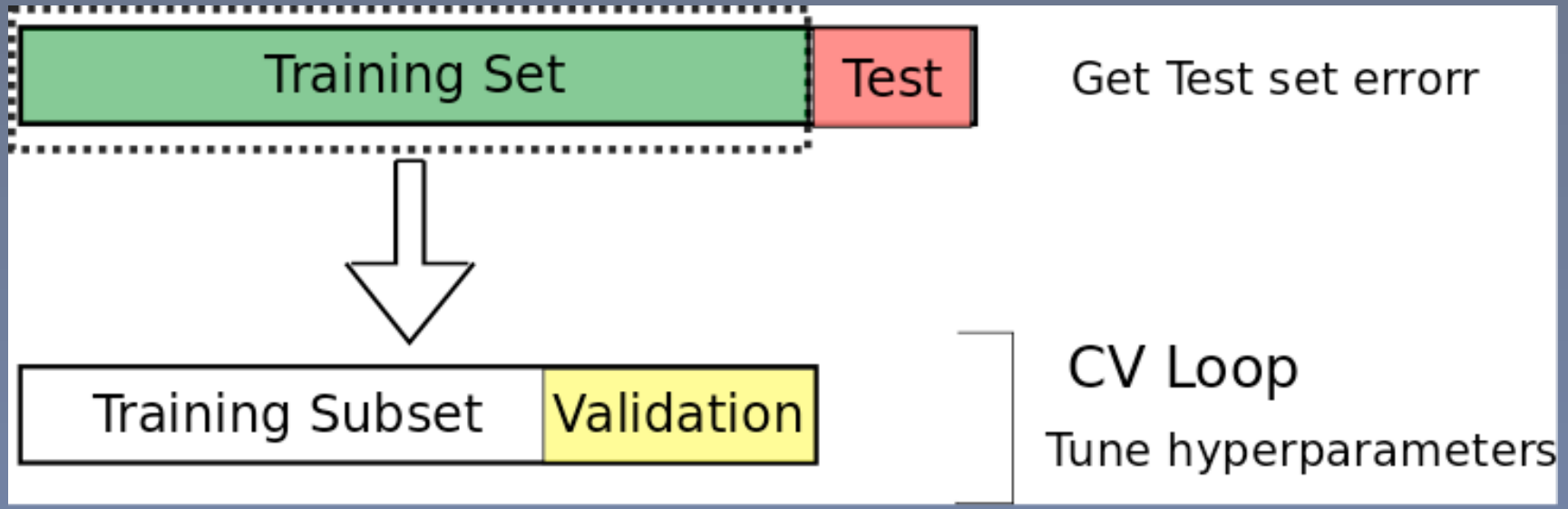
Train and Validation

数据集按照使用用途可以划分为：

✓训练集（Training Set）：进行模型训练和参数更新；

✓验证集（Validation Set）：进行模型验证集和参数选择；

✓测试集（Test Set）：进行验证模型精度；



2 模型训练与验证

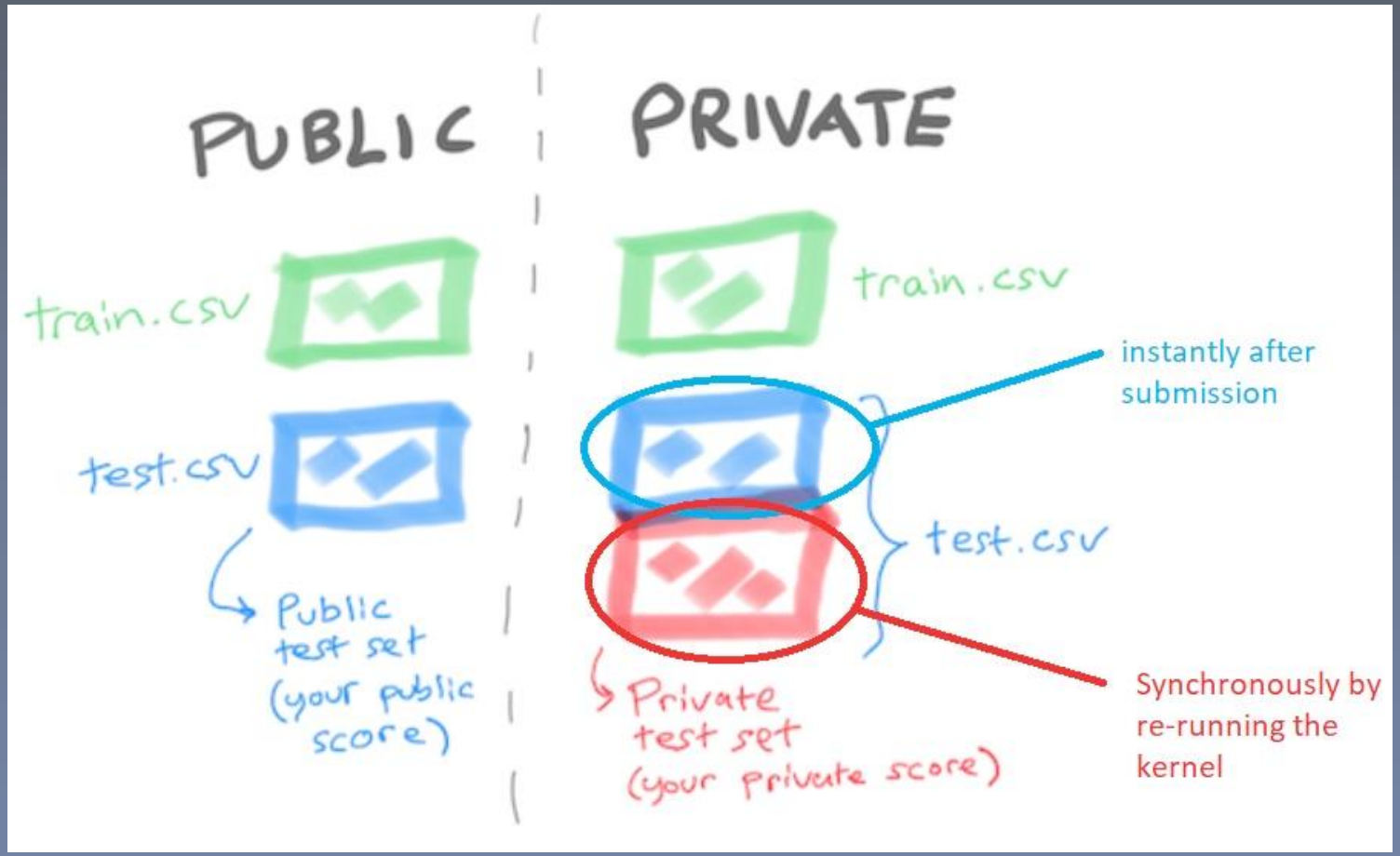
Train and Validation

需要注意：

✓测试集一般不能用来训练；

✓测试集可能分为AB榜单；

✓只要有反馈，就有过拟合；



2 模型训练与验证

Train and Validation

在竞赛中测试集在某些情况也可以加入训练，伪标签：

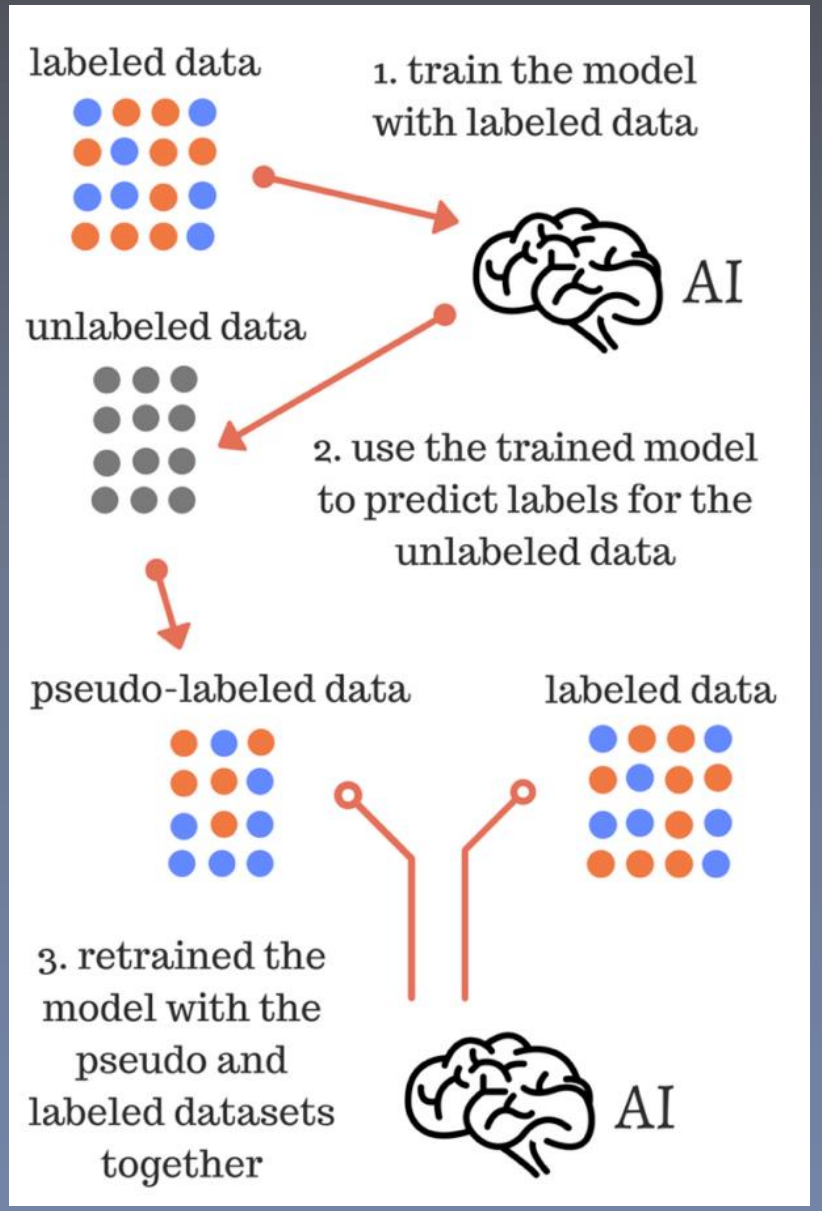
✓当模型精度教高时；

✓但比赛规则允许时；

模型对测试进行预测，并将预测结果与训练集一起再训练：

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

https://www.kaggle.com/c/kuzushiji-recognition/discussion/112712



2 模型训练与验证

Train and Validation

模型根据训练阶段分为：**过拟合** 与 **欠拟合**

✓过拟合无法避免，只能缓解；

✓缓解过拟合的方法：

- 增加数据量（数据扩增）；
- 做正则化（L1或L2）；
- 做交叉验证（Easy Stopping）；
- 增加随机性（Dropout、样本采样）；

2 模型训练与验证

Train and Validation

模型根据训练阶段分为：**过拟合** 与 **欠拟合**

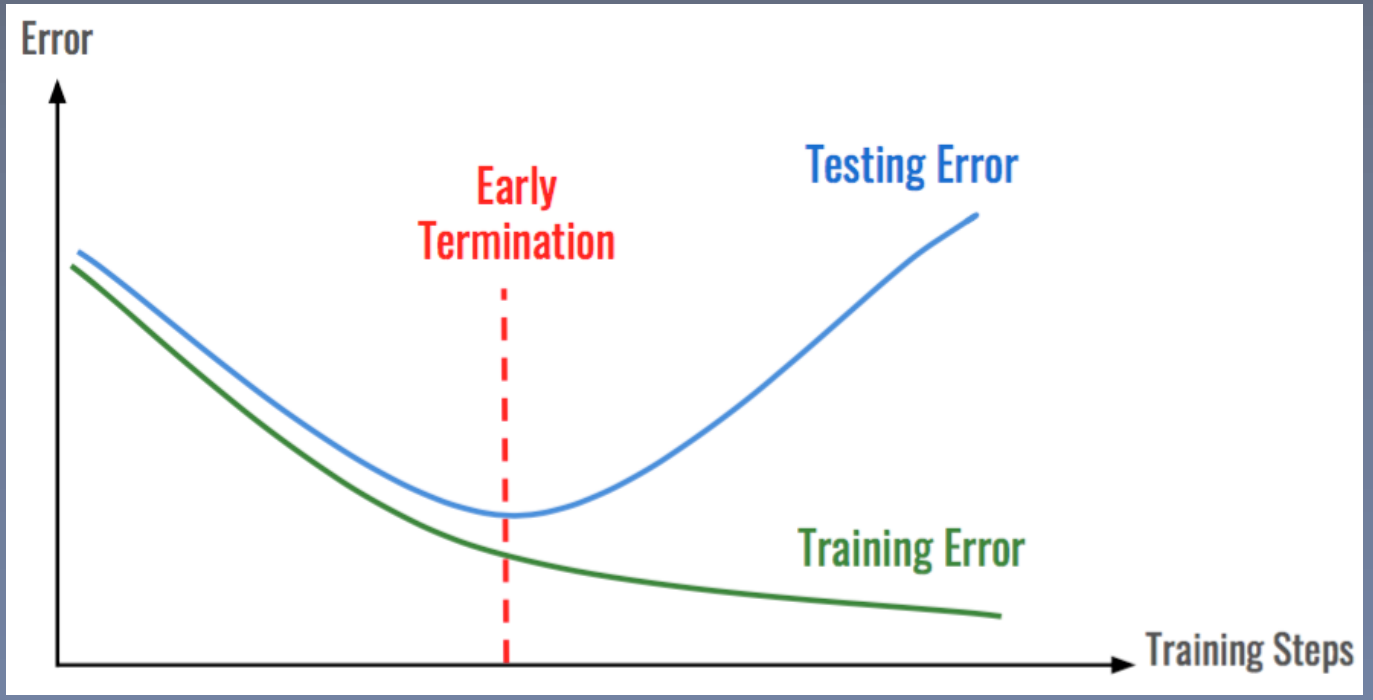
✓缓解过拟合的方法：

- 做交叉验证（Early Stopping）；

在划分有验证集后，可以进行控制停止训练：

✓ 对传统机器学习方法和深度学习都适用；

✓ 数据、超参数不同会影响训练轮数；

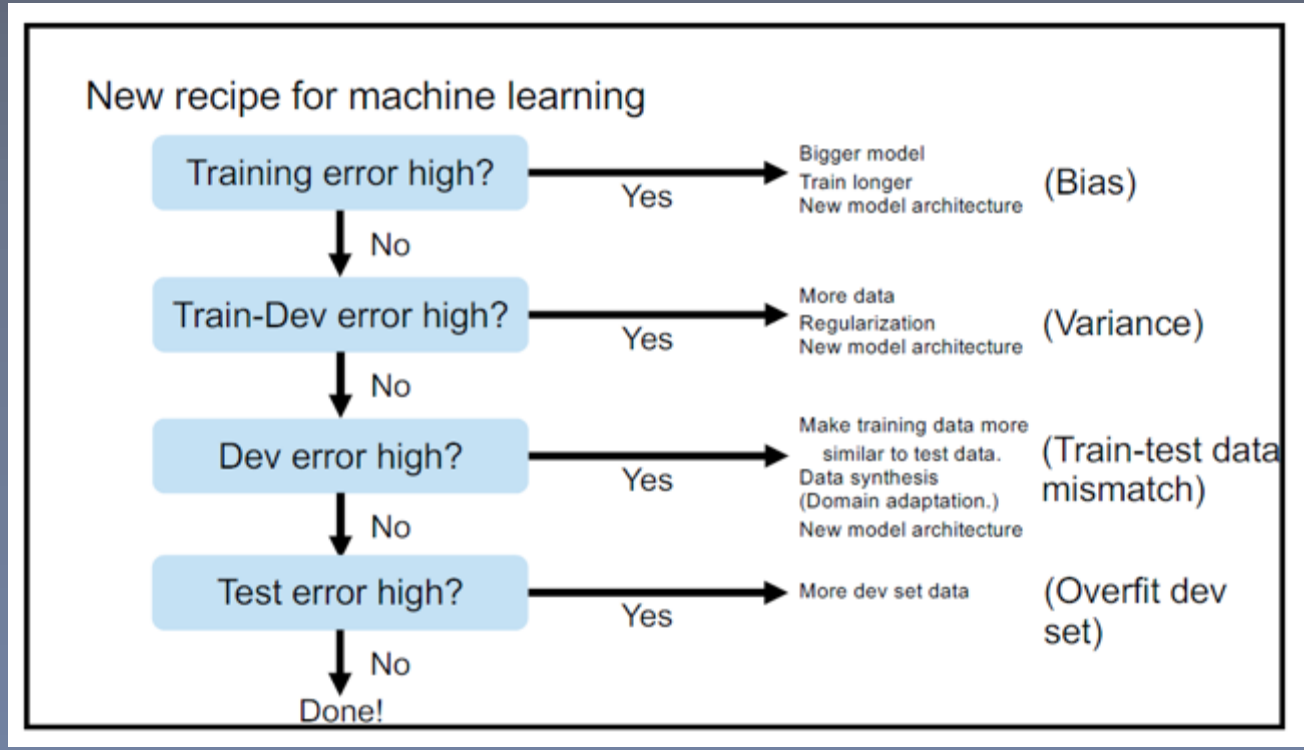


2 模型训练与验证

Train and Validation

找到模型的上限：验证集精度 vs 测试集精度

Error Type	Formula
Bias	(Training Error) - (Human Error)
Variance	(Train-Dev Error) - (Training Error)
Train/Test Mismatch	(Dev Error) - (Train-Dev Error)
Overfitting of Dev	(Test Error) - (Dev Error)



https://www.youtube.com/watch?v=F1ka6a13S9I

2 模型训练与验证

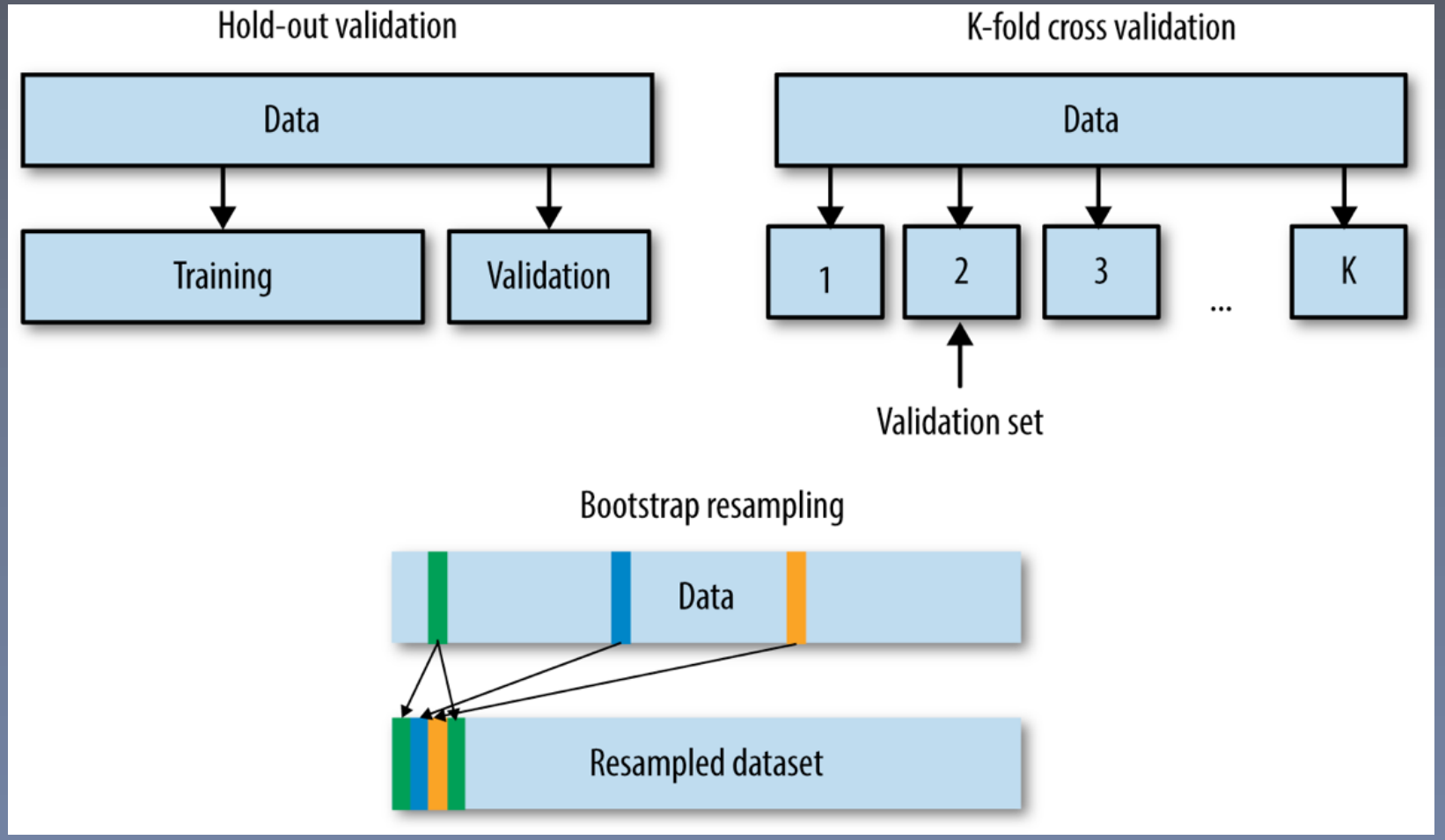
Train and Validation

数据划分（模型评估）方法：

✓留出法（Hold-out）

✓K折交叉验证（K-fold CV）

✓自助采样（Bootstrap）



深度之眼
deepshare.net

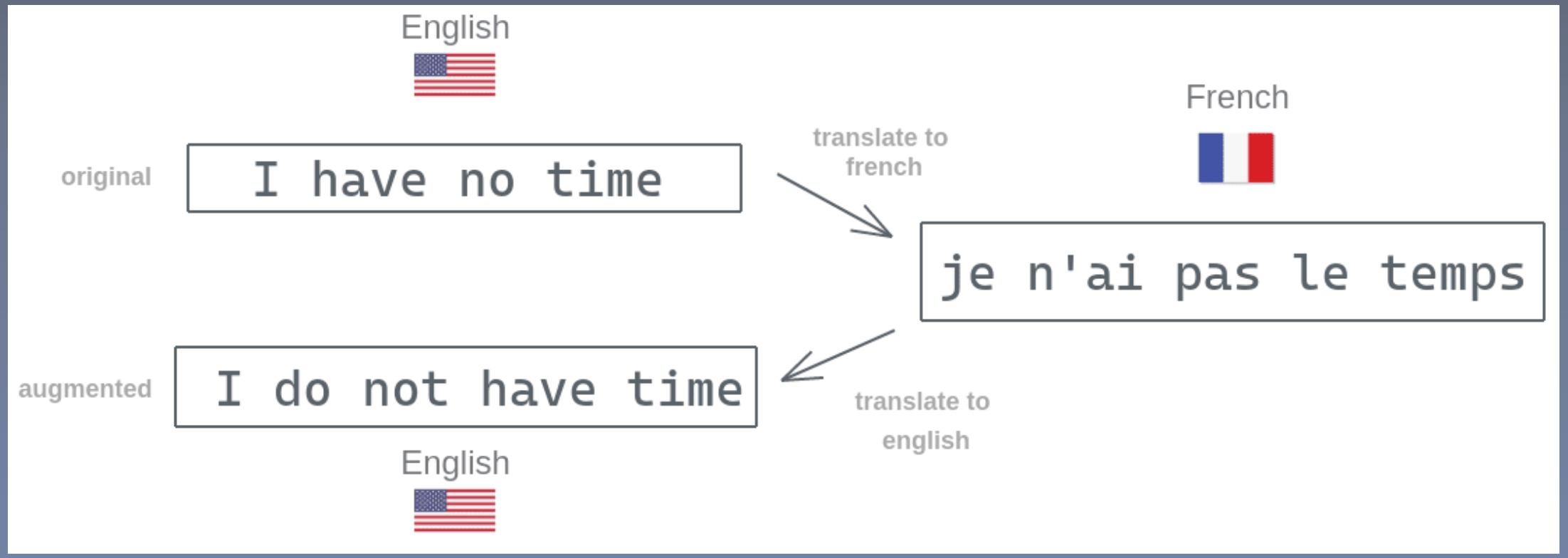
3 数据扩增方法

Data Augmentation

3 数据扩增方法

Data Augmentation

回译： 翻译系统增加语言多样性；



3 数据扩增方法

Data Augmentation

深度之眼
deepshare.net

Easy Data Augmentation

- ✓ 随机插入；
- ✓ 相似词替换；
- ✓ 随机删除；
- ✓ 交换句子位置；

	Sentence
Original	The quick brown fox jumps over the lazy dog
Synonym (PPDB)	The quick brown fox climbs over the lazy dog
Word Embeddings (word2vec)	The easy brown fox jumps over the lazy dog
Contextual Word Embeddings (BERT)	Little quick brown fox jumps over the lazy dog
PPDB + word2vec + BERT	Little easy brown fox climbs over the lazy dog

https://neptune.ai/blog/data-augmentation-nlp
https://github.com/makcedward/nlpaug

深度之眼
deepshare.net

4 分布一致性

Distribution

4 分布一致性

Distribution

得分一致性：本地交叉验证得分、公开榜单、私有榜单

✓如何知道模型精度有提高？

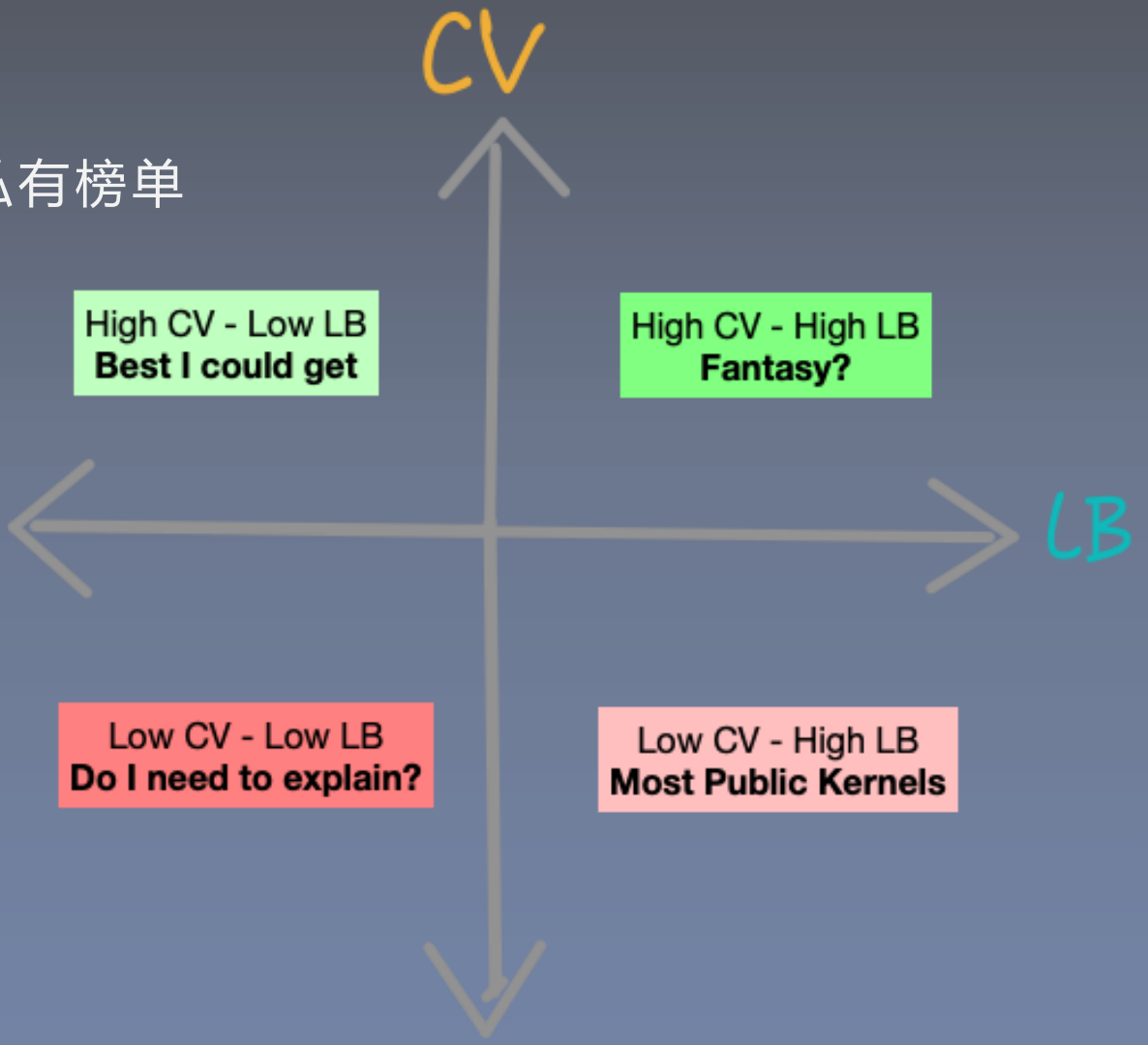
优先看本地验证得分，其次是公开榜单；

✓如何知道模型是否过拟合榜单？

本地验证得分提高，但榜单下降；

✓如何预测最终私有榜单得分？

使用本地验证得分和公开榜单进行预测；



4 分布一致性

Distribution

得分一致性：本地交叉验证得分、公开榜单、私有榜单

✓本地验证得分、公开榜单、私有榜单，保持一致；

✓本地验证得分、公开榜单，不一致；

4 分布一致性

Distribution

得分一致性：优先选择本地交叉验证的得分！

Trust your local CV, do not trust Lead board!

在每个Kaggle比赛中有很多分享：

✓Notebook（Kernel）

✓Discussion

但并不是的所有信息都是有用的，需要自己尝试！

4 分布一致性

Distribution

分布一致性：如何判断两个数据集分布是否一致？

✓方法1：通过数据分析的方式分析；

✓方法2：通过模型来验证（Adversarial Validation）

Adversarial Validation思路是构建一个分类模型，目的是分辨训练集和测试集的来源，这里假设使用AUC作为分类精度评价函数。

✓如果分类模型无法分辨样本（AUC接近0.5），则说明训练集和测试集数据分布比较一致；

✓如果分类模型可以分辨样本（AUC接近1），则说明训练集和测试集数据分布不太一致；

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