# 科研指南：以软件工程研究为例

汇报人：陈翔

汇报时间：2023

# 个人简介

- 研究问题
  - AI4SE、软件仓库挖掘、软件测试与维护
- 发表论文
  - 在TSE、TOSEM、EMSE、IST、JSS、TRel、JCST、JSEP、ICSE、FSE、ASE、ICSME、ISSRE、SANER、软件学报、计算机学报、计算机研究与发展等国内外期刊会议发表高质量论文100多篇
  - 2篇ACM SIGSOFT 最佳论文（ICPC 2023和ICSE 2021）
- 更多信息可以访问
  - https://smartse.github.io/index.html

# 大纲

- 寻找研究问题
- 寻找解决方案
- 实验设计
- 论文写作

# 计算机主要研究方向（参考CCF列表）

- 计算机体系结构/并行与分布计算/存储系统
- 计算机网络
- 网络与信息安全
- 软件工程/系统软件/程序设计语言
- 数据库/数据挖掘/内容检索
- 计算机科学理论
- 计算机图形学与多媒体
- 人工智能
- 人机交互与普适计算
- 交叉/综合/新兴

**软件工程**

- 需求工程
- 软件测试
  - 测试用例生成
  - 测试用例Oracle
  - 回归测试
  - 变异测试
- 软件体系结构
- 软件维护
- 软件仓库挖掘
  - 缺陷预测
  - Stack Overflow挖掘
- 经验软件工程

# 确定研究点的三种方式（1）

- **请教该领域有丰富经验的研究人员**
  - 适用于研究起步阶段
  - 请教对象一般是该领域的资深学者（导师），对研究点有更好的把握能力
  - 研究点不存在好坏之分
    - 只要做的足够深入，都能有好的工作；但在特定的时期，某些领域可能比较活跃，杰出的成果相对来说比较多
  - 当前软件工程最热门的研究方向
    - AI4SE（用AI的技术解决软件工程问题）、LLM4SE
    - SE4AI（用软件工程的方法来提高AI系统的质量）

# 确定研究点的三种方式（2）

- **自己寻找研究点**
  - 确定研究领域内的顶级期刊和顶级会议；在计算机科学研究中，更需要**关注顶级会议发表的论文**
  - 阅读近三年顶级会议发表论文的摘要，总结近几年的研究点
  - 从中选择合适的研究点，选择依据包括
    - 个人兴趣（最重要）
    - 知识结构（如果从无到有，代价太高）
    - 能否获得必要的资源（例如实验对象、实验工具等）

# CCF列表（第六版，2022）

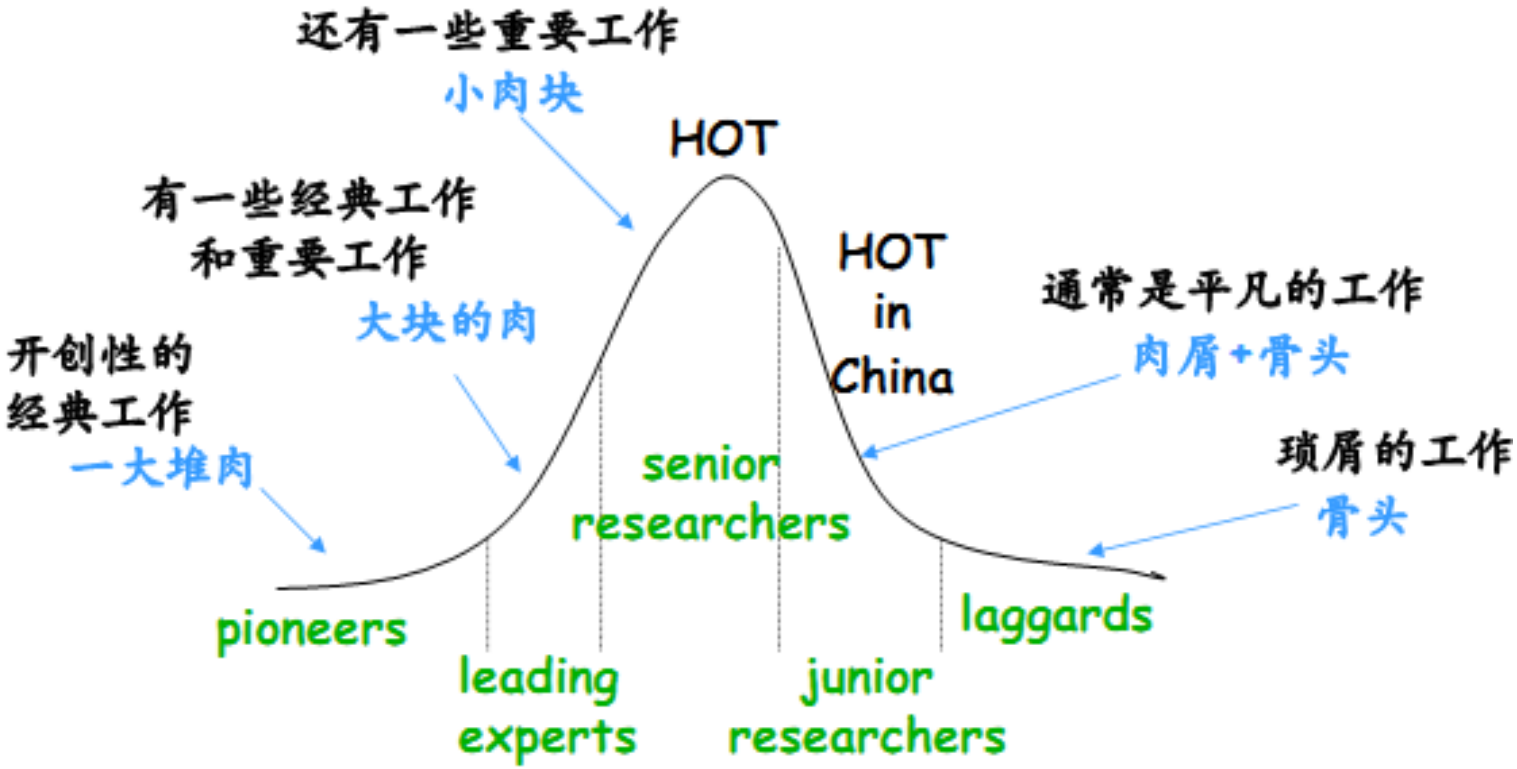
- 将期刊会议分成A类、B类和C类
- 以软件工程为例
  - A类期刊：TSE、TOSEM
  - A类会议：ICSE、FSE、ASE、ISSTA
  - B类期刊：**EMSE**、IST、JSS、 ASEJ、 JSEP、STVR、SPE
  - B类会议：ICSME、ICPC、SANER、ESEM、ISSRE
  - C类期刊：SQJ、IJSEKE
  - C类会议：**ICST**、**MSR**、APSEC、COMPSAC、QRS
  - A类中文期刊：中国科学：信息科学、软件学报、计算机学报、计算机研究与发展

# 软件工程的A类会议和期刊

- A类会议（通常录用率在20%左右）
  - ICSE、ESEC/FSE、ASE、ISSTA
- A类期刊（审稿人评价非常苛刻）
  - TSE、TOSEM
- 国内期刊（大概与CCF B类期刊水平差不多）
  - 计算机学报、软件学报、中国科学：信息科学
- 希望研究生至少从投B类期刊/会议开始

# 确定研究点的三种方式（3）

- 去工业界找问题
  - 来自实际开发过程
  - 容易落地


还有一些重要工作
小肉块
HOT
有一些经典工作
和重要工作
大块的肉
HOT
in
China
通常是平凡的工作
肉屑+骨头
开创性的
经典工作
一大堆肉
senior
researchers
琐屑的工作
骨头
pioneers
laggards
leading
experts
junior
researchers

# 常见论文检索系统

- 谷歌学术搜索引擎（翻墙）
  - 支持为个人建立学术档案（可以了解论文的引用次数、**定期推荐相关论文**）


不限语言 中文网页 简体中文网页

## 推荐的文章

- An Improved Method for Training Data Selection for Cross-Project Defect Prediction
  NA Bhat, SU Farooq
  Arabian Journal for Science and Engineering - 7 天前
- Genetic algorithm-based oversampling approach to prune the class imbalance issue in software defect prediction
  C Arun, C Lakshmi
  Soft Computing - 13 天前
- Feature Selection and Software Defect Prediction by Different Ensemble Classifiers
  N Shakhovska, V Yakovyna
  International Conference on Database and Expert System... - 11 天前

更多 1 周前发布的文章

- Unsupervised Learning to Heterogeneous Cross Software Projects Defect Prediction
  R Vashisht, SAM Rizvi
  International Conference on Innovative Computing and C... - 20 天前

# Xiang Chen （陈翔）

已关注

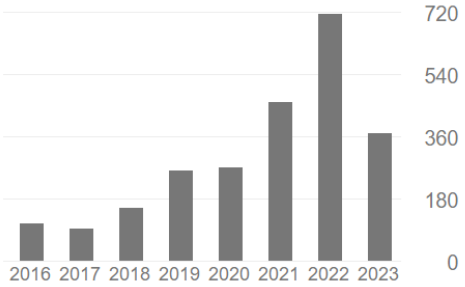
Associate Professor, Nantong University, China
在 ntu.edu.cn 的电子邮件经过验证 - 首页

AI4SE　Software Repository Mining　Software Defect Prediction　Empirical Software Enginee…

| 标题 | 引用次数 | 年份 |
|---|---|---|
| Variable strength interaction testing with an ant colony system approach<br>X Chen, Q Gu, A Li, D Chen<br>2009 16th Asia-Pacific Software Engineering Conference, 160-167 | 117 | 2009 |
| MULTI: Multi-objective effort-aware just-in-time software defect prediction<br>X Chen, Y Zhao, Q Wang, Z Yuan<br>Information and Software Technology 93, 1-13 | 107 | 2018 |
| Empirical studies of a two-stage data preprocessing approach for software fault prediction<br>W Liu, S Liu, Q Gu, J Chen, X Chen, D Chen<br>IEEE Transactions on Reliability 65 (1), 38-53 | 98 | 2015 |
| Applying particle swarm optimization to pairwise testing<br>X Chen, Q Gu, J Qi, D Chen<br>2010 IEEE 34th Annual Computer Software and Applications Conference, 107-116 | 92 | 2010 |
| Software defect number prediction: Unsupervised vs supervised methods<br>X Chen, D Zhang, Y Zhao, Z Cui, C Ni<br>Information and Software Technology 106, 161-181 | 88 | 2019 |

## 引用次数　查看全部

| | 总计 | 2018 年至今 |
|---|---|---|
| 引用 | 2751 | 2240 |
| h 指数 | 30 | 27 |
| i10 指数 | 71 | 60 |




## 开放获取的出版物数量　查看全部

25 篇文章　69 篇文章

无法查看的文章　可查看的文章

根据资助方的强制性开放获取政策

# 常见论文检索系统

- DBLP系统
  - https://dblp.uni-trier.de/
- ACM、IEEE、Springer等
- Arxiv
  - https://arxiv.org/list/cs.SE/recent（与软工相关的论文）

# 论文的阅读

- 论文不需要按照顺序去阅读
  - 通过阅读标题、摘要和引言**确定：**
    - **论文关注的问题**
    - **解决方法**
    - **具体的实验设计和主要的实验结果**
- 需要批判性思维（Critical Thinking）
  - 问题的假设是否合理？
  - 方法是否合理？
  - 实验设计是否合理？
- 重现论文共享的代码是找到idea的一个重要方式

# 寻找研究问题

- 问题创新最为重要
  - Stack Overflow标题自动生成
  - 已有工作：基于帖子中的代码片段来生成标题
  - 我们的问题创新
    - 额外考虑帖子中的问题描述，是否可以提高生成标题的质量？（SANER 2022）
    - 开发人员如果能提供一些标题片段，是否可以自动补全剩余片段？（ICPC 2023）
    - 如果能有用户标题的历史修改日志，是否可以自动润色标题？（TSE 2023）

# 寻找解决方案

- 具备研究领域基础知识
  - 适当的学习经典教材和经典慕课
- 能有自己熟悉的工具与方法
  - 多目标优化方法、集成学习、迁移学习、特征选择、类不平衡方法
- 借鉴其他领域的最新进展
  - 机器学习领域的最新进展是否可以用于软件测试领域？
  - 这些最新的方法如何结合问题特征进行扩展？

# 实验设计

- 实验对象的选择
  - 利用已有的（有超过baseline的压力） Vs 自己去构造（相对容易取得更好的结果）
- 评测指标/Human Study
- 与基准方法的比较（是否超过state-of-the-art (SOTA)）
- 方法框架内的影响因素的识别 （消融实验）
- 结果分析
  - 图、表、显著性分析方法
- 实验脚本和数据的共享（建议上传到github网站）
- 相关工具的开发

# 论文的写作

- 了解不同类型论文的框架
  - 综述性论文
  - 方法类论文（推荐这一类研究工作）
  - 实证研究类论文
- 使用Latex排版
  - 使用Share Latex网站联机修改
  - 去Arxiv上找一篇论文的Latex代码，直接研究

# 论文投稿

- 会议（计算机领域更加重视会议）
  - 3~4个审稿人
  - 论文篇幅固定（10页正文+2页参考文献）
  - 固定时间返回审稿意见
  - 竞争激烈（优中选优，大概20%的录用率）
  - 开销大（注册费+差旅费）

# 论文投稿

- 期刊
  - 论文篇幅不限制
  - 审稿周期长（半年到1年，取决于审稿人的速度）
  - 审稿结论：Reject、Major revision、Minor revision、Accept
  - 审稿意见需要逐条回复
  - 大部分期刊不收版面费
    - 有的期刊会对超页进行收费，例如TSE，超过一页收取220美元
    - 有的期刊会选择Open access

# 投稿注意事项

- 一稿多投
- 禁止抄袭
  - 抄袭他人
  - 抄袭自己

# 会议扩充到期刊

- 需要扩展50%到100%（需要方法、实验等有新的内容）
- 已有的内容建议重写，否则查重会有问题
- 在Introduction中需要明确指出新增的内容
- 基本上
  - C类会议可以投到C类期刊
  - B类会议可以投到B类期刊
  - A类会议可以投到A类期刊

# 论文的审稿

- 单盲 Vs 双盲
  - 提交论文
  - 评审论文
- 会议论文（双盲） 期刊论文（单盲）也有会议开始公开审稿人的审稿意见
- 首先看论文的整体是否专业
  - 排版是否规范、结构是否规范、图表是否有特色
- 问题是否新颖
- 解决方案是否新型
- 实验设计是否合理

# AI4SE的论文check list（1）

- 研究的问题是否解决了实际问题？研究的问题是否具有一定的难度？
- 深度学习方法是否针对研究问题有定制？可以考虑使用经典的深度学习方法作为baseline，随后通过性能比较，来说明直接使用深度学习方法是不可行的。因此研究问题本身具有一定的挑战性
- 论文所提方法针对已有基准方法是否有提升？基准方法是否考虑周全。数据集的划分是否多次（如果1次，需要说明这是相关研究经常使用的实验设置）？评测指标是否考虑齐全？
- 论文是否针对定制部分，进行了消融实验的设置。同时针对定制部分，最好能够有为什么这么定制的解释。

# AI4SE的论文check list （2）

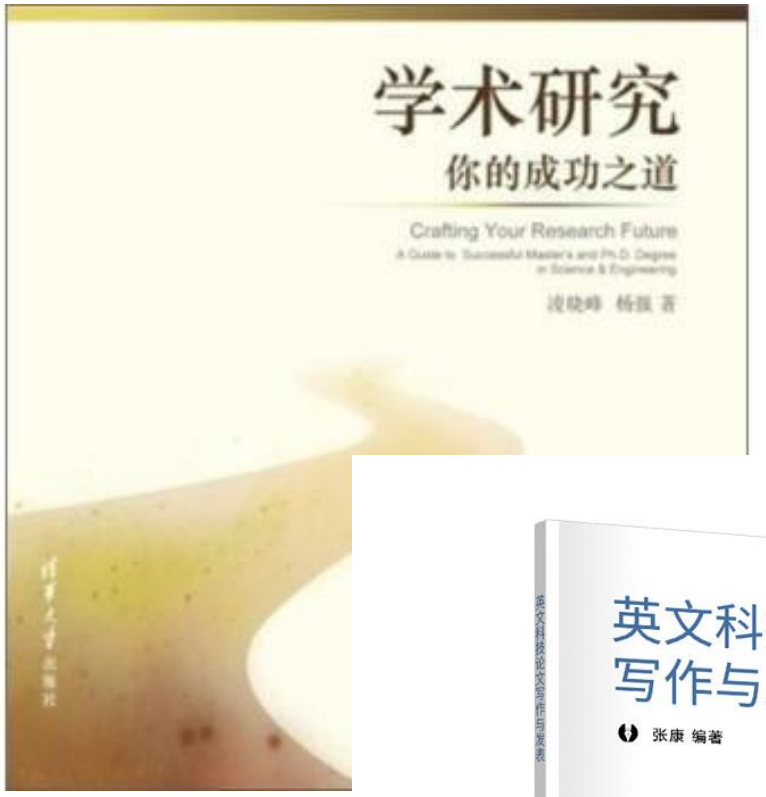
- 深度学习中的超参取值对结果的影响，实验是否有分析？
- 论文的有效性，需要进行human study。human study的方法是否设计合理？
- 论文的limitation，简单通过几个例子，说明为什么我们方法的效果不好
- 论文方法，最好能配套开发对应的原型工具
- 论文的相关工作，尽量分类，最后强调一下novelty
- 论文共享模型、数据集并提供详细的说明

# 组内管理

- 希望1个研究生与若干本科生组队来攻克研究问题
- 每周组会，两位同学来讲，讲的内容：论文或者自己当前的研究工作或者需要报告的论文
- 计划
  - 研一：打基础，尽早确定研究点（硕士生课程通常压力不大），开始投论文/专利/软著（CCF B起步）
  - 研二：完成两个研究工作，校外实习（专硕毕业要求，如果自己联系实习，务必需要保证满足毕业要求）
  - 研三：有论文录用/找工作/完成毕业论文（3月份论文盲审）

# 参考文献


- 周志华，做研究与写论文，2007
- 彭明辉，研究生手册
- S.C. Cheung，On the IDOL route to be a researcher in students' days，2006
- Tao Xie，Research Skills，2006
- Michael Ernst，Choosing a venue: conference or journal? 2006


学术研究
你的成功之道
Crafting Your Research Future
A Guide to Successful Master's and Ph.D. Degree in Science & Engineering
凌晓峰 杨振 著

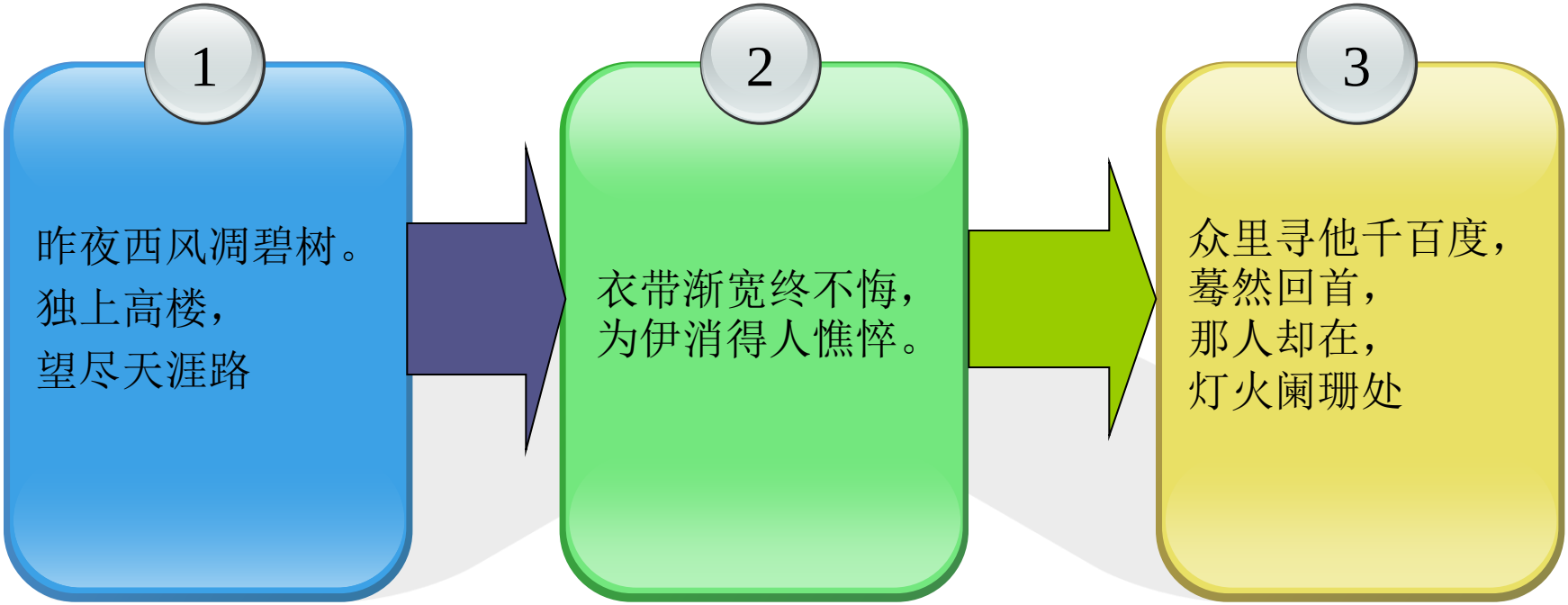

高质量读研：教你如何写论文、做科研
高质量读研：
教你如何写论文、做科研
张军平 著
时商与基础误区
想象力
寻导师的五维度
脱口秀与讨论班
科研流水线
论文合作技巧
创新力
评审与回复
合作共赢
心理调适
人民邮电出版社
凝练复旦大学教授的研究生指导经验、科研一线“干货”
郑南宁院士 · 周志华教授 · 马臻教授 · 五道口的颜学姐 | 联合推荐
中国科学院院士、“吴文俊人工智能最高成就奖”首位获得者
陆汝钤
作序
中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS


英文科技论文写作与发表
张康 编著
清华大学出版社
IC PAPER WRITING

# 总结

- 科研没有捷径，需要投入足够的时间（每周需要至少高效率的投入30个小时）
- 需要精通与课题最为相关的论文
- 需要尽可能的去多复现研究工作，熟悉数据和代码
- 需要与导师多交流
- 需要寻求合作的机会（单枪匹马是很难发表高质量论文的），与论文作者发邮件进行沟通等
- 需要多锻炼写论文的能力（建议背几篇模板的论文组织方法）
- 需要尽量投稿到A类期刊和会议，可以得到高质量的审稿意见，有益于积累自己的学术声誉和影响力

# 古今之成大事业大学问者必经过三种境界 by 王国维



# 智能软件工程——问题

- 分类问题
  - 软件缺陷预测（software defect prediction）
  - 即时软件缺陷预测（just-in-time defect prediction）
  - 漏洞预测/检测 （Software vulnerability prediction/detection）
  - 特定类型的缺陷报告识别（例如：安全缺陷报告、危险缺陷报告）
  - 恶意软件检测（Android Malware detection）
  - 代码克隆检测（code clone detection）
  - 重复缺陷报告检测
  - 代码分类（code classification）

# 智能软件工程——问题

- 生成问题
  - **代码到文本：**代码摘要生成（source code summarization）、提交消息生成（commit message generation）、代码评审（code review）、Bash代码注释生成（Bash comment generation）、log日志生成、方法名生成/推荐
  - 文本到代码：代码生成、漏洞代码生成、SQL代码生成、正则表达式生成、测试用例代码生成、Bash代码生成
  - **混合（代码+文本）到文本：**Stack Overflow Title Generation、Issue Title Generation
  - **代码到代码：**缺陷修复、漏洞修复
  - 文本到文本：查询重构

# 智能软件工程——问题

- **推荐/检索问题**
  - 代码推荐
  - API推荐
  - 缺陷定位

# 智能软件工程——问题

- 问答系统
  - 答案生成
  - APP review response
  - 答案总结

# 智能软件工程-研究角度

- 从数据集质量入手
  - 噪音的识别与移除
  - 特征选择 类不平衡方法
  - 数据集的扩充
  - 模块标记问题

# 智能软件工程-研究角度

- **分析实验设置对结果的影响**
  - 数据预处理方式
  - 数据集的划分方式
  - 评测指标的使用
  - 超参优化
  - 源代码解析工具的影响
  - 标识符切分

# 智能软件工程-研究角度

- 从代码表征角度入手—与程序分析方法结合
- 低资源编程语言的建模
  - **Few-shot learning**
  - **Zero-shot learning**
  - 迁移学习等
- 跨项目预测场景

# 智能软件工程-研究角度

- 简单方法
- 专家特征与手工特征的融合
- 信息检索方法与深度学习方法的融合
- 预训练+微调
- Prompt tuning
- 对比学习 （Contrastive Learning）
- 对偶学习
- 集成学习
- 强化学习
- 图神经网络
- 个性化模型
- 主动学习
- 基于输入过滤的方法

# 智能软件工程-研究角度

- 深度模型与知识图谱的融合
- 从模型的可解释性入手
- 从模型的鲁棒性入手
- 从模型的安全性/攻击入手
- 从模型的部署入手
- 从模型的公平性入手
- 从模型的泛化能力入手

# 智能软件工程-研究角度

- 智能系统的质量保障
  - 覆盖准则
  - 测试用例生成
    - Fuzzing
    - 蜕变测试
  - 测试用例选择
  - 变异测试
  - 缺陷检测
  - 缺陷定位
  - 缺陷修复
  - 对话系统
  - 翻译系统
  - 图像识别系统
  - 。 。 。

Thank you

Any Question?