图书馆有WMSR、IEEE、ACM库，DBLP（根据作者信息、会议信息），Google学术，在上面找相关文章，看摘要。复现文章内容。论文页码超过10页优先（已有实现）

先收集数据[爬取github数据]->预处理、初步分析【统计数据】->构造模式，数据挖掘

用已有手段处理github数据（统计、模式、数据挖掘）

【江贺-大连理工】【周明辉-北大】

启发：获取数据、确立问题、建立模型、数据统计，挖掘

代码分类（分类标准），已有分类标准，自行设计分类标准 🡪 结合知识图谱

热点：知识图谱（知识图谱结合分类）-具备相应领域的知识，如获取的操作系统代码就要了解其构成

软工+机器学习 基础方法的应用 以上的结合应用是创新点

方向：克隆代码的检测-代码相似性的度量（clone代码在github上占比例90%）

先构建数据集：代码信息、commit信息、数据

语言规范、语言特性：通过在代码仓库中实证

构造数据集（并自动更新数据集工具）—— 可视化工具 —— 可视化工具可以用于数据挖掘等

数据获取（复用）；数据库模式；自动更新功能