1.11会议记录：

1. 提出了3个模型挖掘方向：

①构建分类标准，对代码进行分类。已有的分类标准用来预处理。可以考虑结合知识图谱。

②背景：github上克隆的代码超过90%。对克隆代码检测：本质是对代码相似性的度量

③一些编程语言的语言规范、语言特性、流行编程方式：通过在代码仓库中挖掘数据进行实证。

2.数据来源

爬取github仓库：手段：Java+JGit或者python

目标：获取github仓库数据，构建成文件（或构建数据库）

3.大致流程：收集数据 🡪 预处理、初步分析、统计数据 🡪 构造模式、数据挖掘

要善用已有手段处理数据，但这不是创新点。

要善用各种已有数据库和论文库， https://lib.nju.edu.cn

综合已有手段，也是创新点。创新在综合上。

4.挖掘方向：需要了解确定的方向相应的领域的知识，如获取的操作系统代码（Linux等）就需要了解操作系统方面的知识。

5.项目重点知识需求：软工基础开发与机器学习