* 1. **国内外研究现状**I wanna have a try

本小节简要回顾现存的缺陷预测方法。这些方法大致上可以分为二类：1、通过分类技术预测软件模块是否缺陷，2、通过回归技术预测软件模块缺陷的数目。这里将以上二类以以下三种方式讨论。

* + 1. **本项目软件缺陷预测方法**

本项目软件缺陷预测是软件工程数据挖掘领域中一个研究的热点。[静态软件缺陷预测方法研究-陈翔]该方法通过分析软件代码或者开发过程设计出与软件缺陷相关的度量元。通过挖掘软件历史仓库(software historical repositories)来创建缺陷预测数据集。目前，可以挖掘与分析的软件历史仓库包括项目所处的版本控制系统(例如SVN，CVS，Git等)、缺陷跟踪系统(例如Bugzilla，Mantis，Jira或Trac)或相关开发人员的电子邮件等。最后，对搜索到的数据进行清洗得到可用缺陷预测数据集，并构建缺陷预测模型。应用模型对目标项目进行缺陷预测。

Akiyama等人[付-13]最早提出了软件缺陷预测方面的量化关系式，研究计算并给出了软件缺陷与代码行数之间的关系式：D = 4.86+0.018L，即代码行数与软件缺陷个数呈线性关系。这种量化关系式子最早是基于人们的经验估计得到。当时普遍认为平均每完成60行代码就会出现一个软件缺陷。Akiyama通过统计分析早期软件缺陷与代码行数的内在联系，提出了该量化关系式。在软件工程领域最早应用于估计软件缺陷数量。

支持向量机[yx-25-27]、神经网络[yx]、决策树以及贝叶斯等方法都是缺陷预测领域的分类方法。这些方法使用软件度量指标来预测软件模块是否存在缺陷。但是，数据集中缺陷倾向和无缺陷类别的高度不平衡性质降低了预测性能。这些问题也是本项目软件缺陷预测领域最常涉及到的问题。因此，现已提出特征选择和类别不平衡学习方法来解决以上问题。比如，Wang 等[improving cross-17]利用深度信赖网络从源代码中自动学习语义特征。