Mining Software Repositories
What is MSR?

- Mining Software Repositories (MSR) uses data available in repositories to support development activities
  - For example, defect assignment, software validation, evolution and planning
- Increased amount of data available in software archives through large open source projects
- Process applies data extraction and analysis to make decisions and predictions
Repositories

• Three repository types are the focus of research activity:
  • Bug Repositories (e.g. Bugzilla)
  • Source Repositories (e.g. CVS, SVN)
  • Communication Repositories (e.g. mailing lists)
Research focus areas

• Bug repositories
  • Defect triage
  • Estimate hours for a fix to plan schedules
  • Report quality

• Source Repositories
  • Code clones
  • Code evolution and project growth
  • Bug detection

• Communication Repositories
  • Mailing list interaction
Process

• Extract data from the repository (e.g. Eclipse Mylyn demo)
• Generally apply a machine learning algorithm or data mining algorithm to the data (e.g. WEKA demo)
  • Typical strategies:
    – Supervised learning
    – Unsupervised learning (e.g. kNN)
Data Analysis

- In analyzing the effectiveness of the application, typically use precision and recall

- Precision
  - How well provide concise recommendations \cite{Yin04}

- Recall
  - How well recommend items in the correct solution \cite{Yin04}
Future direction

- Increased support through tooling
- Ease of use for the end user
Resources


