Step 1: Pattern mining

Semi-trust (humans must decide this)

Learn idiosyncratic patterns in code

Self-supervision; no labels

Step 1.0: Source Code Repositories

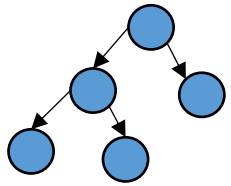
Codebase

Step 1.1: Mine patterns in control structures

Source code parser

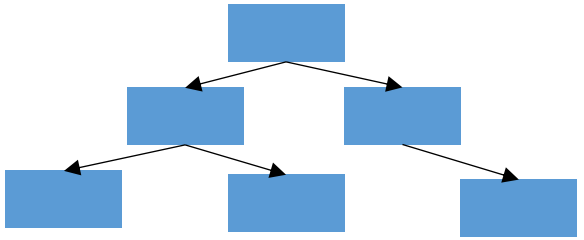
Patterns

Step 1.2: Build representation for patterns



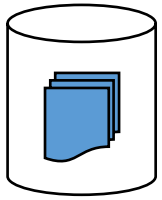
Syntax trees for patterns

Step 1.3: Self-supervised clustering using decision tree



Step 2: Scanning for idiosyncratic patterns

Step 2.0: Target Code Repository



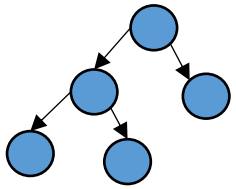
Codebase

Step 2.1: Mine patterns in control structures

Source code parser

Patterns

Step 2.2: Build representation for patterns



Syntax tree for pattern

Step 2.3: Find pattern and its "nearest" patterns in the decision tree.

Nearest patterns from training dataset

Step 2.4: Is pattern an anomaly?

$$\frac{n_0 * 100}{\sum_{i=0}^{\max\_cost} \max(n_i)} < \alpha \; \forall (p, n) \in C$$