SuSi: A Tool for the Fully-Automated Classification of Android Sources and Sinks

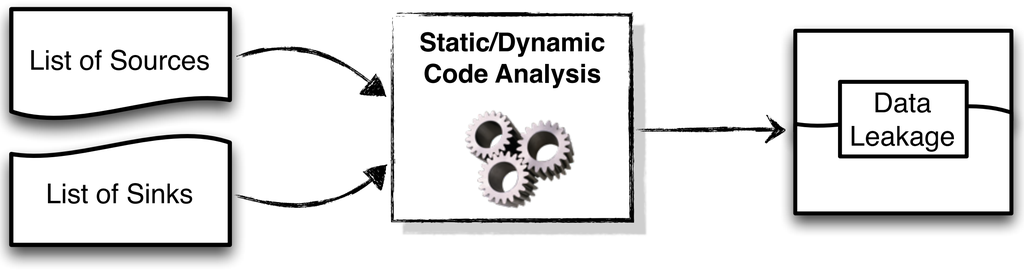
Steven Arzt, Siegfried Rasthofer and Eric Bodden (TU Darmstadt / EC SPRIDE)

Motivation and Goal

- **Information-flow tools** require specifications of sources and sinks
- Current code analysis approaches usually only consider a small **hand-selected set** of sources and sinks known from literature
- But those lists are incomplete, causing many **data leaks** to go undetected

Main Goal:

Fully automated generation of a categorized list of sources and sinks for Android applications.



Sources and Sinks

Android Source:

Sources are calls into resource methods returning non-constant values into the application code.

Android Sink:

Sinks are calls into resource methods accepting at least one non-constant data value from the application code as a parameter, if and only if a new value is written or an existing one is overwritten on the resource.

Categories

- Sources are categorized into domain-specific categories:

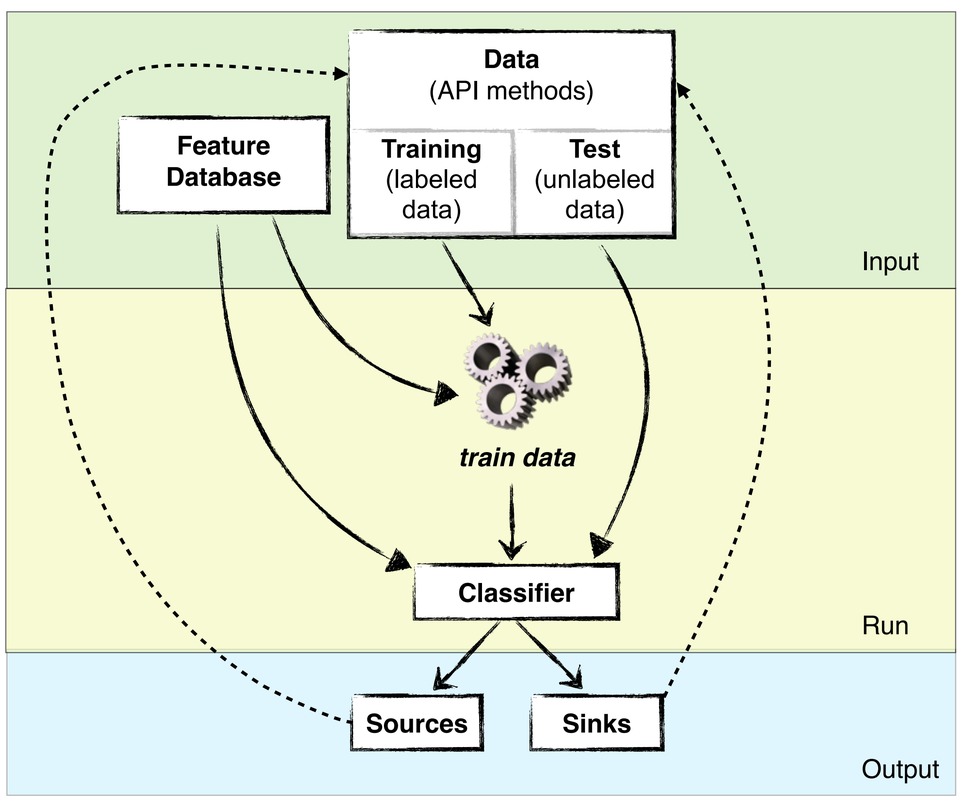
SMS/MMS | Location Information | Calendar Information | Contact Information | Other

- Sinks are categorized into domain-specific categories:

SMS/MMS | Bluetooth | Near-Field-Communication | Email | Other

- New categories can easily be added:
  - Label API methods for the new category
  - Add category-specific features into the feature database
- Categories can be used to semantically define flows between sources and sinks (e.g., only interested in flows: location information via SMS)

Methodology



- Training data is created from randomly picked and hand-annotated examples (*labeled data*)
  - Sources/sink training set (1st run)
  - Categories training set (2nd run)
- Meaningful features extracted from data samples (*feature database*)
- *Input:* Android API methods (*unlabeled data*), trained Classifier and *Feature Database*
- *1st Run:* Train the classifier for sources/sinks and evaluate all Android methods
- *2nd Run:* Train the classifier for categories and evaluate it on the sources/sinks from 1st run
- *Output:* Categorized list of sources and sinks

Features

- Fully-automated approach
- Android version independent
- Very fast classification
- Provides the most comprehensive publicly available list of sources and sinks
- General approach could be adopted to other platforms like J2EE, PHP, C++, etc.

Category	True Positives	False Positives
Sources	0.907	0.008
Sinks	0.852	0.034
Neither/nor	0.954	0.122
Weighted Average	0.926	0.090

Table 1: Source/Sink Cross Validation

Emmy Noether-Programm, Deutsche Forschungsgemeinschaft DFG

http://sse.ec-spride.de/
http://sse-blog.ec-spride.de

SECURE SOFTWARE ENGINEERING GROUP