Evaluating Reliability of Static Analysis Results Using Machine Learning

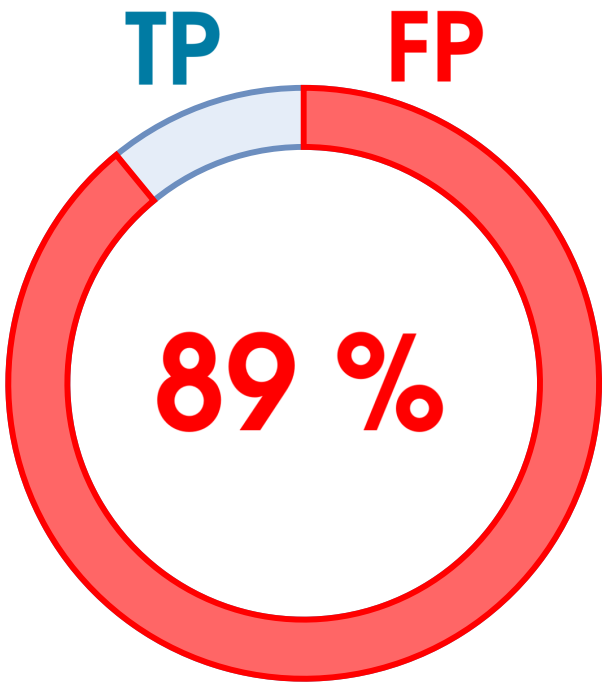
Tomáš Beránek
Supervisor: prof. Ing. Tomáš Vojnar, Ph.D.
Consultants: Ing. Viktor Malík, Mgr. Marek Grác, Ph.D.

Brno University of Technology, Faculty of Information Technology

February 1, 2023

Motivation – Meta Infer

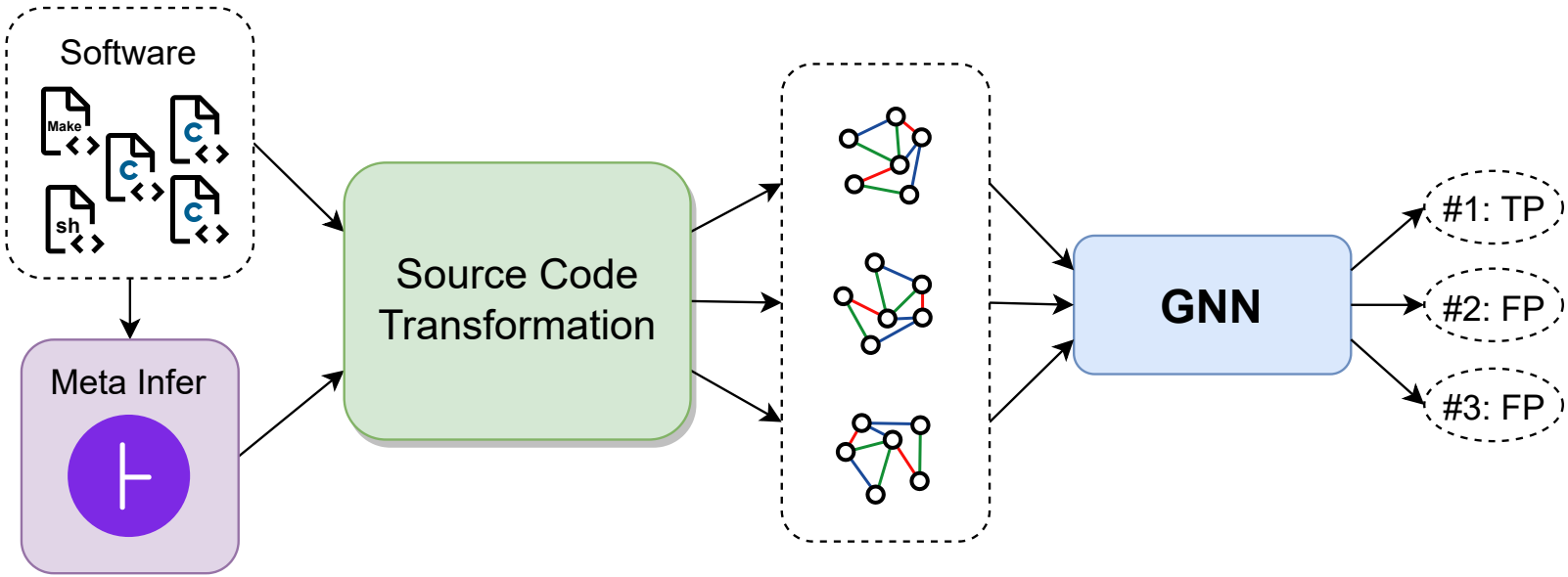
- Advantages:
  - highly scalable,
  - easy to use,
  - can analyze variety of software (with a wrapper).
- Disadvantages:
  - too many False Positives (FP) (almost 90 %!).



Objectives of the Thesis

- Identify false positives using deep learning.
  - Choose a suitable kind on NN:
    - Graph Neural Network (GNN).
  - Choose a suitable code representation:
    - Code Property Graph (CPG).
- Input: source files and build scripts.
- Output: probability of being FP for each reported issue.

- The false positive detection system consists of:
  - source code transformation,
  - trained model.

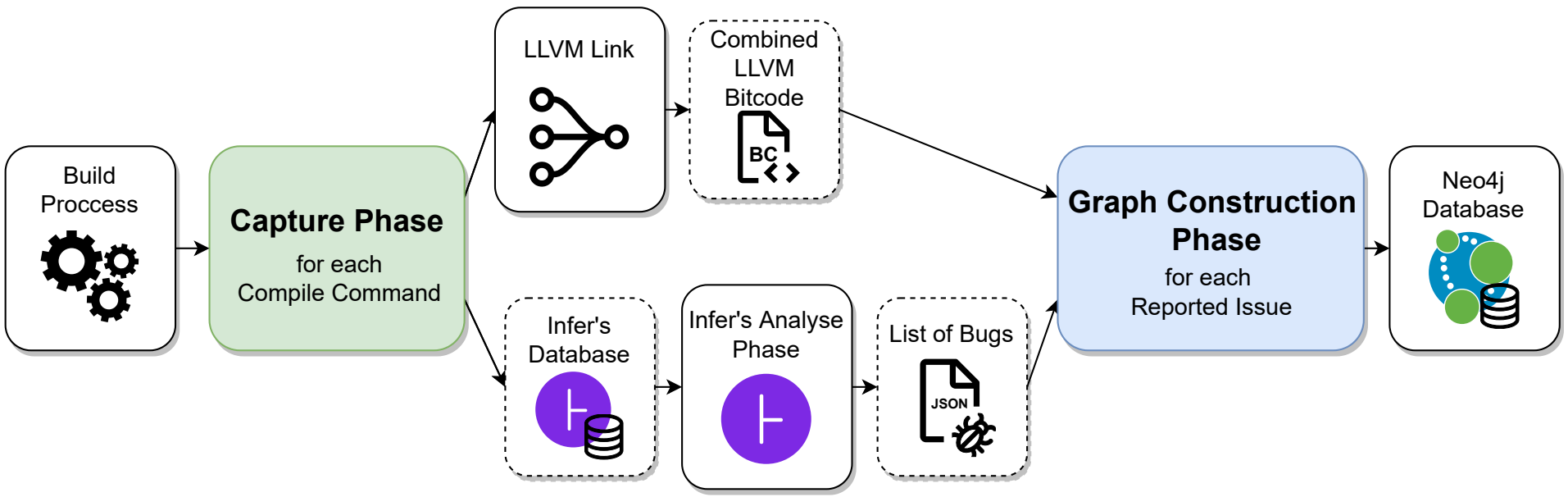


Existing Approaches for Graph Construction

- Existing approaches often use Joern.
- Disadvantages of existing approaches:
  - not considering conditional compilation,
  - inability to automatically identify required source files.
- The proposed solution also has a slightly different use case.

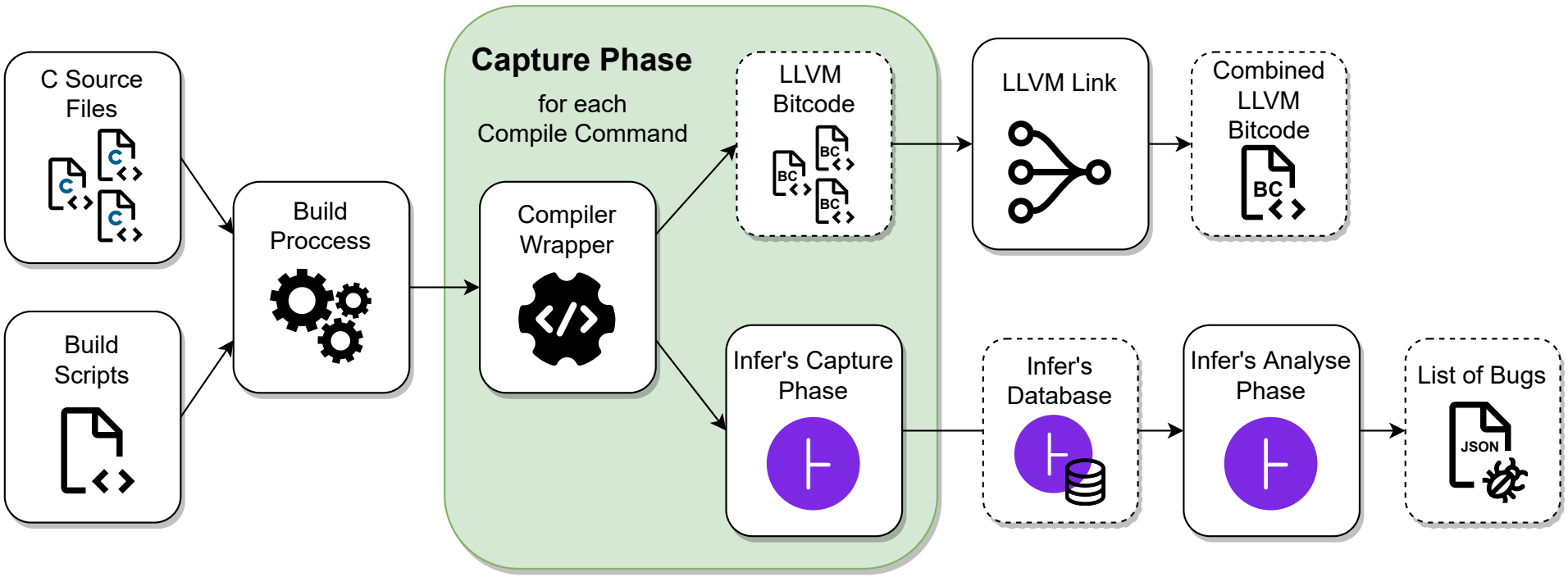
Graph Construction Pipeline

- Input is limited to C and a subset of C++.
- The output CPG is neither language- nor analyzer-dependent.
- Input: compilation commands (source files).
- Output: code property graph for each reported issue.



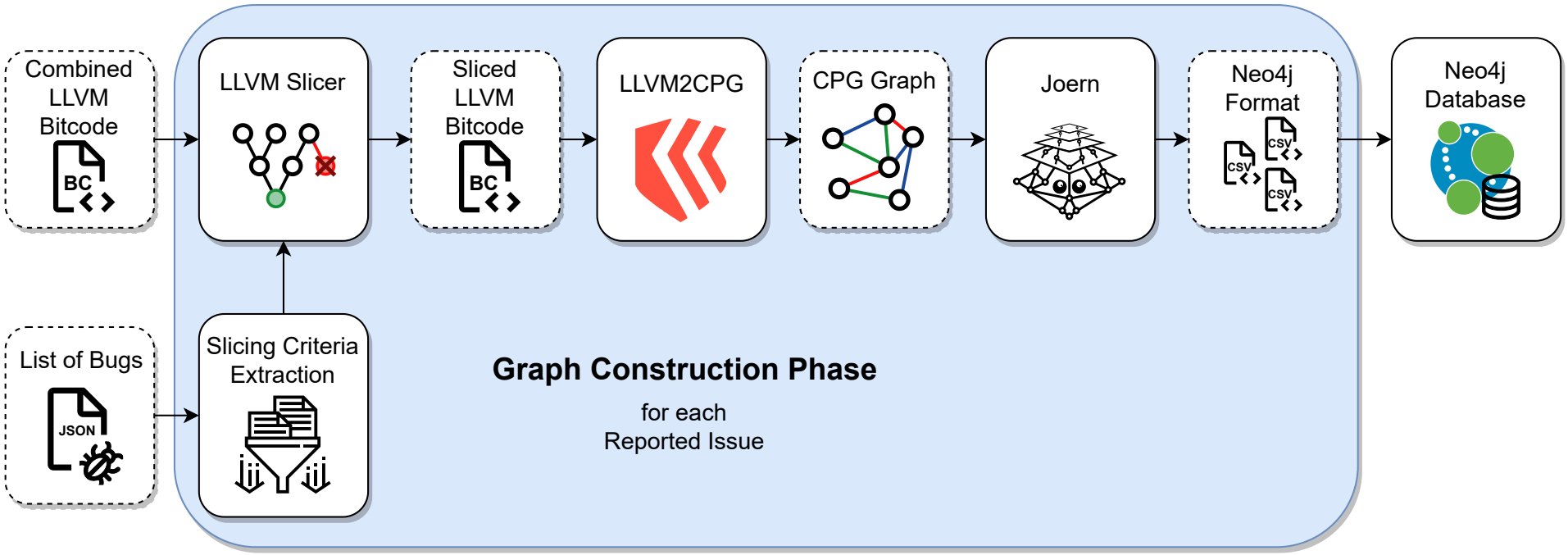
Capture Phase

- Input: compilation commands (source files).
- Output: combined LLVM bitcode and a list of issues.



Graph Construction Phase

- Input: combined LLVM bitcode and a list of issues.
- Output: code property graph for each reported issue.



Future Work

1. Automation of the graph construction pipeline.
2. Graph extraction from the D2A dataset.
3. GNN model architecture selection and training.
4. Implementation of self-training.
5. Integration with csmock.
6. Experiments on SRPM packages.