Caterina Urban
INRIA — Team ANTIQUE
École Normale Supérieure
45 rue d'Ulm, 75005 Paris
caterina.urban@inria.fr

A General Framework for Input Data Usage Static Analysis

M2 Research Internship Proposal, 2019-2020

Context. Nowadays, thanks to advances in machine learning and the availability of vast amounts of data, computer software plays an increasingly important role in assisting or even fully automating tasks in our daily lives. As a consequence, however, we become increasingly vulnerable to *programming errors*. In particular, programming errors that do not cause failures can have serious consequences since code that produces an erroneous but plausible result gives no indication that something went wrong. In 2012, flawed code caused losses of billions of dollars for JP Morgan Chase & Co. [4]. In medical applications, programming errors are deadly [2]. It is thus paramount to develop methods and tools that can keep up with these developments and ensure that data science software behaves correctly and reliably.

Goals and Objectives. In recent work [5], we have proposed a static analysis for automatically detecting *unused input data*. Its key ingredient is tracking *syntactic dependencies* between the input data and the outcome of the program.

The goal of this internship is to extend this static analysis. We envision multiple directions that can be explored:

1. Using *semantic* notions of dependency;
2. Designing a complementary analysis for *guaranteeing* input data *usage*;
3. Investigating other input data usage errors such as *accidental duplication*.

In addition to addressing (some or all of) the theoretical questions mentioned above, the internship will also include the implementation of the newly designed static analyses and the experimental evaluation of their practical usefulness on realistic Python programs. The implementation will be integrated into the Lyra static analyzer, developed by the host team. If time permits and the intern is interested, the implementation could also be extended to support programs using common data science libraries such as NumPy.

Prerequisites. The internship requires a background in program analysis and abstract interpretation. Knowledge of the Python programming language is also required. Familiarity with NumPy is a plus, but not a requirement.

Practical Information. The internship will take place in the INRIA research team ANTIQUE, hosted at École Normale Supérieure, Paris. A successful internship may provide opportunities for a funded PhD on a follow-up subject.

References

[1] Roberto Giacobazzi and Isabella Mastroeni. Abstract Non-Interference: Parameterizing Non-Interference by Abstract Interpretation. In *POPL*, pages 186–197, 2004.

[2] Nancy G. Leveson and Clark Savage Turner. Investigation of the Therac-25 Accidents. *IEEE Computer*, 26(7):18–41, 1993.

[3] Isabella Mastroeni and Damiano Zanardini. Abstract Program Slicing: An Abstract Interpretation-Based Approach to Program Slicing. *ACM Transactions on Computational Logic*, 18(1):7:1–7:58, 2017.

[4] Michael Cavanagh and JPMorgan Chase & Co. Report of JPMorgan Chase & Co. Management Task Force Regarding 2012 CIO Losses, 2013.

[5] Caterina Urban and Peter Müller. An Abstract Interpretation Framework for Input Data Usage. In *ESOP*, pages 683–710, 2018.