Caterina Urban
ANTIQUE Research Team
Inria & École Normale Supérieure
caterina.urban@inria.fr

Static Analysis of Jupyter Notebooks

The advent of big data — the manipulation and analysis of massive quantities of data – has revolutionized the world of software development in the past decade. Every day, data scientists develop short computer programs to clean, analyze, process, and visualize data. As we rely more and more on these data-manipulating programs for making decisions, we become increasingly vulnerable to programming or technical mistakes. Mistakes that do not cause software failures can have serious consequences, since they give no indication that something went wrong. Just to name a recent case, a simple technical mistake made during data processing last year, caused nearly 16,000 cases of Covid-19 between September 25th to October 2nd to go unreported from official figures in the UK. As a consequence, Public Health England was unable to send out the relevant contact-tracing alerts[1]. Mistakes in medical applications can be deadly.

Goals. This goal of this project is to contribute to an on-going effort to develop an abstract interpretation-based static analyzer to aid data scientists develop Jupyter notebooks. The project combines both theoretical and practical implementation work with the objective of designing and implementing novel abstract domains for analyzing data frame-manipulating programs. An extensive experimental evaluation will target real-world Jupyter notekbooks which make use of the most popular Python data science libraries, i.e., NumPy, Pandas, etc.

Useful Prerequisites. The following skills would be helpful, but can also be learned during the project:

- Background in static analysis and abstract interpretation
- Familiarity with Python and its data science ecosystem

Contacts

- Caterina Urban
  caterina.urban@inria.fr

[1] https://www.bmj.com/content/bmj/371/bmj.m3891.full.pdf