Towards Robust and Scalable Search of Binary Code and Data

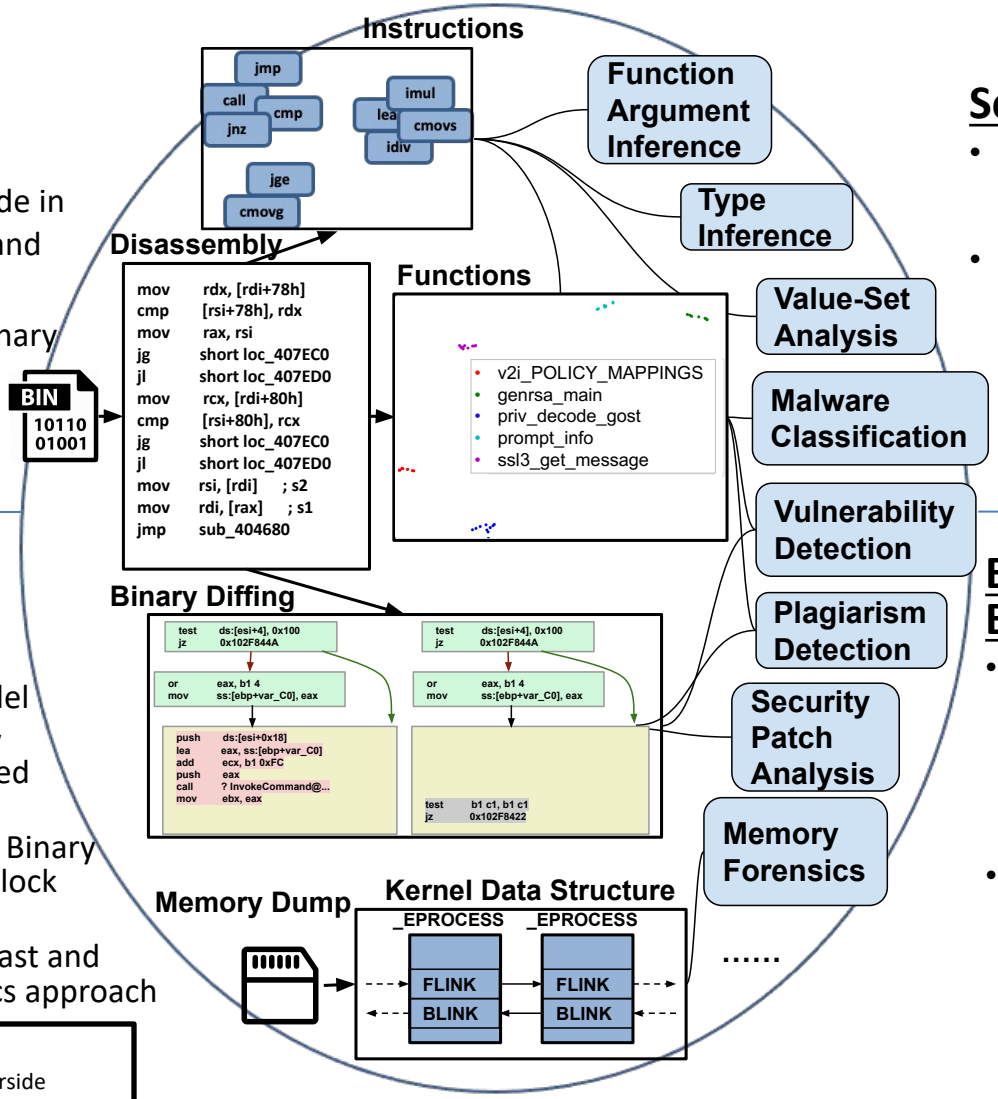
UC RIVERSIDE

Challenge:

- Identify semantically equivalent or similar code in different architectures and compilation settings
- Identify objects from binary data such as memory dumps and documents.
- Scale to a large dataset.

Solution:

- **DeepDi** [Usenix'22]: A fast disassembler
- **PalmTree** [CCS'21]: An assembly language model
- **Gemini** [CCS'17]: Binary similarity detection based on function embedding
- **DeepBinDiff** [NDSS'20]: Binary diffing based on basic-block representations
- **DeepMem** [CCS'18]: A fast and robust memory forensics approach



Scientific Impact:

- Security applications built on top will benefit from the project
- Stimulate more research in the direction of deep learning-based binary analysis
  - Gemini got 388 citations by 05/08/2022

Broader Impact and Broader Participation:

- The project can be applied to security task such as vulnerability detection, which will guard the security for PC or mobile users
- Reverse engineer, security analyst and security audit company will benefit from the project

Award Number: 1719175

Institution: University of California, Riverside
Principal Investigator: Heng Yin (heng@cs.ucr.edu)