Hao He (she/her)

haohe@andrew.cmu.edu https://hehao98.github.io/ hehao98 +1 949 351 1893

Data Science & Machine Learning Skills

- **Statistical & ML Modeling**: Causal Inference, Quasi-Experiments, Difference-in-Differences, Panel Regression, Counterfactual Analysis, XGBoost, Ensemble Methods, Anomaly Detection, Time-Series Analysis, A/B Testing
- **Programming Languages**: Python, R, JavaScript, Java, C, C++, C#
- **Libraries & Frameworks**: PyTorch, Scikit-learn, Pandas, NumPy, Apache Spark, Google BigQuery, MongoDB, MySQL, Docker, Jupyter, CodeQL, Linux, Shell, Networkx, HTML, CSS, Vue.js
- **Domain Expertise**: Fraud Detection, Network Analysis, Recommendation Systems, Large-Scale Data Processing, Empirical Research Methodology, Software Supply Chain Security, Program Analysis, Mining Software Repository

Education

- **Carnegie Mellon University** Pittsburgh, PA, USA
  Ph.D. in Software Engineering; Adv: Bogdan Vasilescu & Christian Kästner *Aug 2023 - Dec 2026 (Exp.)*
- **Carnegie Mellon University** Pittsburgh, PA, USA
  Master of Science in Software Engineering *Aug 2023 - Dec 2024*
- **Peking University** Beijing, China
  Ph.D. Student in Computer Software and Theory (transferred to CMU) *Sep 2020 - Jul 2023*
- **Peking University** Beijing, China
  Bachelor of Science in Computer Science and Technology *Sep 2016 - Jul 2020*

Experience

- **Carnegie Mellon University** Pittsburgh, PA
  Graduate Research Assistant *Aug 2023 - Current*
  - Applied quasi-experimental simulation and panel regression to measure the effect of dependency management strategies on 30K+ open-source projects over a 12-month period (FSE'25 Distinguished Paper Award)
  - Designed and implemented *Abandabot*, a context-aware LLM-based prototype to provide dependency abandonment recommendations using API call-site analysis and retrieval-augmented generation (ICSE'26)
  - (Work in Progress) Applying state-of-the-art difference-in-difference framework and program analysis techniques to unveil the impact of LLM agent assistant on development velocity and software quality
- **Socket Inc (https://socket.dev/)** Remote
  Software Engineering Research Intern *Jun 2024 - Aug 2024*
  - Designed and implemented *StarScout*, a fraud detection system processing the entire GitHub (20 TiB+ of data) using Google BigQuery to identify fraudulent starring activities and fake promotion campaigns
  - The anomaly detection method identified 6M+ fake stars across 331M repositories with 81% precision
  - The research is deployed in production as a Socket Alert, published in ICSE'26, and reported by media
- **Peking University** Beijing, China
  Graduate Research Assistant *Sept 2020 - July 2023*
  - Designed and implemented *GFI-Bot*, an end-to-end machine learning application to recommend "good first issues" to GitHub newcomers based on historical issue resolution data among popular GitHub repositories
  - The data pipeline collects and incrementally updates GitHub repository & user data for model training
  - The underlying XGBoost classifier was trained on 53K+ GitHub issues and achieved 0.853 AUC
  - The research was recognized as two top-tier conference publications (ICSE'22, FSE'23)
  - Led a variety of other data-science-driven projects under the topic of open-source software sustainability, all of which were published in major software engineering venues (e.g., ICPC'22, ASE'23, TSE'23)

- **Huawei Technologies, Co., Ltd.** Beijing, China
  Research Software Engineer Intern *Sep 2020 - Apr 2022*
  - Designed and implemented a data processing pipeline to mine library migration patterns from 20K+ open-source projects and generate recommendations in an internal IDE plugin (SANER'21, ICSE'21)
  - Designed and implemented a static and dynamic analysis tool for Python package API extraction and breaking change analysis, as part of a larger project that assists package updates in Python

Selected Publications

- **[ICSE'26]** Designing Abandabot: When Does Open Source Dependency Abandonment Matter?
  Courtney Miller, ***Hao He**, Weigen Chen, Elizabeth Lin, Chenyang Yang, Bogdan Vasilescu, Christian Kästner (*Joint First Author) *The 2026 IEEE/ACM International Conference on Software Engineering,* [PDF]

- **[ICSE'26]** Six Million (Suspected) Fake Stars in GitHub: A Growing Spiral of Popularity Contests, Spams, and Malware
  Hao He, Haoqin Yang, Philipp Burckhardt, Alexandros Kapravelos, Bogdan Vasilescu, Christian Kästner
  The 2026 IEEE/ACM International Conference on Software Engineering, [PDF]

- **[FSE'25]** Pinning Is Futile (🏆 **Distinguished Paper Award**)
  Hao He, Bogdan Vasilescu, Christian Kästner
  The 2025 ACM International Conference on the Foundations of Software Engineering [PDF]

- **[ASE'23]** Understanding and Remediating Open-Source License Incompatibilities in the PyPI Ecosystem
  Weiwei Xu, ***Hao He**, Kai Gao, and Minghui Zhou (*Joint First Author)
  The 2023 38th IEEE/ACM International Conference on Automated Software Engineering. [PDF]

- **[TSE'23]** Automating Dependency Updates in Practice: An Exploratory Study on GitHub Dependabot
  Runzhi He, ***Hao He**, Yuxia Zhang, and Minghui Zhou (*Joint First Author)
  IEEE Transactions on Software Engineering, Aug 2023. [PDF]

- **[ICSE'22]** Recommending Good First Issues in GitHub OSS Projects
  Wenxin Xiao, ***Hao He**, Weiwei Xu, Xin Tan, Jinhao Dong, and Minghui Zhou (*Joint First Author)
  The 2022 IEEE/ACM 44th International Conference on Software Engineering. [PDF]

- **[ICPC'22]** Demystifying Software Release Note Issues on GitHub (🏆 **Distinguished Paper Award**)
  Jianyu Wu, **Hao He**, Wenxin Xiao, Kai Gao, and Minghui Zhou
  The 2022 IEEE/ACM 30th International Conference on Program Comprehension. [PDF]

- **[ESEC/FSE'21]** A Large-Scale Empirical Study on Java Library Migrations: Prevalence, Trends, and Rationales
  Hao He, Runzhi He, Haiqiao Gu, and Minghui Zhou
  The 2021 ACM 29th Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. [PDF]

- **[SANER'21]** A Multi-Metric Ranking Approach for Library Migration Recommendations
  Hao He, Yulin Xu, Yixiao Ma, Yifei Xu, Guangtai Liang and Minghui Zhou
  The 2021 IEEE 28th International Conference on Software Analysis, Evolution and Reengineering. [PDF]

  Full List: https://scholar.google.com/citations?user=eL6RHssAAAAJ&hl=en

Languages & Services

- **Languages**: English (TOEFL 114/120), Japanese (JLPT N1, 145/180), Chinese (Native)
- **Program Committee**: ICSE'23 Artifact Evaluation, MSR'23 & MSR'24 Junior PC
- **Peer Review**: ACM Transactions on Software Engineering and Methodology, Journal of Systems and Software, Empirical Software Engineering, Journal of Software: Evolution and Process, Information and Software Technology