SI CHEN

(540)-988-2118 ◇ chensi@vt.edu
Google Scholar ◇ Github ◇ LinkedIn ◇ Twitter ◇ Webpage

RESEARCH EXPERTISE

AI Security, **Generative AI**, **Deep Learning**, **Machine Learning**, **Data Valuation**, **Privacy**

EDUCATION

Ph.D., Computer Engineering (CPE), Virginia Tech *Aug. 2019 - April. 2025*

- Master of Science in CPE
- Graduate Research Assistant @ Responsible Data Science Lab
- Student spotlight @ Sanghani Center for Artificial Intelligence & Data Analytics
- Advisor: Ruoxi Jia

B.S., School of Information and Electronics, Beijing Institute of Technology *Sep. 2015 - Jun. 2019*

INTERNSHIP HIGHLIGHTS

Innopeak Technology, Inc *June. 2023 - Aug. 2023*
AI Research Intern @ Seattle Research Center
Conducted research aimed at tracing factual knowledge within Language Models back to their training corpus.

Samsung, Inc *May. 2022 - Aug. 2022*
AI Research Intern @ Samsung Research America
Developed practical defenses against backdoor attacks on image classifiers, effectively eliminating the need for clean in-distribution data.

Innopeak Technology, Inc *June. 2021 - Aug. 2021*
AI Research Intern @ Seattle Research Center
Led a project focused on model inversion attacks, leveraging generative models (i.e., GANs) to enhance the quality of recovered samples from a target face recognition model.

SELECTED PUBLICATIONS & MANUSCRIPTS

(i) **Data-Centric Defense: Shaping Loss Landscape with Augmentations to Counter Model Inversion**
Si Chen, Feiyang Kang, Nikhil Abhyankar, Ming Jin and Ruoxi Jia
In Submission.

(ii) **Turning a Curse into a Blessing: Enabling In-Distribution-Data-Free Backdoor Removal via Stabilized Model Inversion** TMLR 2023
Si Chen, Yi Zeng, Tianhao Wang, Won Park, Xun Chen, Lingjuan Lyu, Zhuoqing Mao and Ruoxi Jia
Transactions on Machine Learning Research.

(iii) **Just Fine-tune Twice: Selective Differential Privacy for Large Language Models** EMNLP 2022
Weiyan Shi, **Si Chen**, Chiyuan Zhang, Ruoxi Jia and Zhou Yu
The 2022 Conference on Empirical Methods in Natural Language Processing.

(iv) **Adversarial Unlearning of Backdoors via Implicit Hypergradient**[video] ICLR 2022
Yi Zeng, **Si Chen**, Won Park, Z. Morley Mao, Jin Ming and Ruoxi Jia
The International Conference on Learning Representations.

(v) **Label-Only Model Inversion Attacks via Boundary Repulsion** CVPR 2022
Mostafa Kahla, **Si Chen** and Ruoxi Jia
Proceedings fo the IEEE / CVF Computer Vision and Pattern Recognition Conference.

(vi) **Knowledge-Enriched Distributional Model Inversion Attacks** [video] ICCV 2021
Si Chen, Mostafa Kahla, Ruoxi Jia and Guo-Jun Qi
Proceedings of the IEEE/CVF International Conference on Computer Vision.

(vii) **Zero-Round Active Learning** ArXiv Preprint
Si Chen, Tianhao Wang and Ruoxi Jia
ArXiv Preprint, 2021.

(viii) **One-Round Active Learning** TMLR 2023
Tianhao Wang, **Si Chen** and Ruoxi Jia
Transactions on Machine Learning Research.

(ix) **Data2Model: Predicting Models from Training Data** SaTML 2023
Yingyan Zeng, Tianhao Wang, **Si Chen**, Hoang Anh Just, Ran Jin and Ruoxi Jia
1st IEEE Conference on Secure and Trustworthy Machine Learning.

SELECTED PROJECTS

Project ①: Backdoor Mitigation ii,iv *Advisor:* ***Prof. Z. Morley Mao & Prof. Ruoxi Jia***

- Propose a universal backdoor removal framework with and without access to clean in-distribution data.
- Effectively reduce Attack Success Rate (ASR) to $\leq 10\%$ while maintaining high Accuracy in defending against various types of backdoor attacks.
- Formulate backdoor removal as a bilevel minimax optimization and solve with implicit hypergradient.

Project ②: Generative Model Inversion (MI) Attacks i,vi,v *Advisor:* ***Prof. Ruoxi Jia & Dr. Guo-Jun Qi***

- Propose frameworks of MI attacks under both white-box and black-box (hard labels only) settings.
- Boost the attack accuracy of the SOTA MI attacks by 150% and generalize better to a variety of datasets and models. The attack accuracy is $> 90\%$ on CelebA dataset.
- Present a novel inversion-specific GAN that can better distill knowledge from the target model; recover the private training distribution instead of single data points compared with prior works.

Project ③: Selective Differential Privacy for Large Language Models iii *Advisor:* ***Prof. Ruoxi Jia*** & **Prof. Zhou Yu**

- Propose a specifically designed two-step fine-tune strategy to prevent transformer-based models from privacy leakage.
- On both the task of natural language understanding and language generation, achieves better privacy guarantee with higher accuracy/ lower perplexity than DPSGD.
- Present selective differential privacy notion and corresponding policy function.

Project ④: Active Learning Under Limited Interaction with Data Labeler vii,viii *Advisor:* ***Prof. Ruoxi Jia***

- Propose a one-round active learning framework which selects data to be labeled all at once. Further extend the framework to the zero-round setting, which avoids the necessity for labeled data in the domain of interest.
- Achieve SOTA performance on various active learning benchmarks in the one-round setting.
- Learn a model that predicts data utility for a set of data and use it to guide the selection of unlabeled data.

Project ⑤: Sematic Image to Image Translation *Advisor:* ***Prof. Jia-Bin Huang***

- Proposed a novel Semantic Generative Adversarial Network to generate images with attributes specified explicitly.
- Our designed method demonstrated superior performance with lower Mean Average Error than Pix2pix, enabling attribute manipulation in generated images.

PROFESSIONAL SERVICES

PC Member:	36th & 37th AAAI Conference on Artificial Intelligence (AAAI-22, AAAI-23, AAAI-24)
Reviewer:	Conference on Neural Information Processing Systems (Neurips' 23)
Reviewer:	International Conference on Computer Vision (ICCV'23)
Reviewer:	2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)
Reviewer:	IEEE Transactions on Dependable and Secure Computing
Reviewer:	IEEE Transactions on Multimedia
Reviewer:	IEEE Transactions on Circuits and Systems for Video Technology
Reviewer:	The 28th &29th ACM International Conference on Multimedia

TECHNICAL STRENGTHS

Programming:	Python, Matlab, C, R, EasyX, Arduino, Verilog HDL, Assembly language
Frameworks:	Pytorch, Tensorflow, Sklearn, Numpy

SELECTED COURSEWORK

Deep Learning, Advanced Machine Learning, Computer Vision, Optimization Techniques, Statistical Inference, Bayesian Statistics, Theoretical Statistics, Linear Algebra, Data Structure