Xuan (Lily) Yang

+1 984-309-5936 | xuan.yang@duke.edu
Durham, NC, USA

Education

Duke University — Durham, NC, USA
PhD in Computer Science; ***Advisor: Dr. Jian Pei*** — *Sep. 2023 – Now*
(Open to extending the internship part-time after the summer term if needed)

Zhejiang University — Hangzhou, China
M.S. in Computer Science; ***GPA: 3.98/4.00*** — *Sep. 2020 – Mar. 2023*

Zhejiang University — Hangzhou, China
B.S. in Digital Media Technology; ***GPA: 3.96/4.00*** — *Sep. 2016 – Jun. 2020*

Research Interest

Data-centric AI; Multi-agent LLM System; Data Valuation

Selected Publications

1. **Xuan Yang**, Jian Pei. Local Shapley: Efficient Data Valuation for Model Training. Under Review.
2. **Xuan Yang**, Furong Jia, Roy Xie, Xi Xiong, Jian Li, Monica Agrawal. Batch-of-Thought: Cross-Instance Learning for Enhanced LLM Reasoning. Under Review.
3. **Xuan Yang**, Yang Yang, Chenhao Tan, Yinghe Lin, Zhengzhe Fu, Fei Wu, Yueting Zhuang. Unfolding and Modeling the Recovery Process after COVID Lockdowns. *Scientific Reports* 2023.
4. **Xuan Yang**, Yang Yang, Jintao Su, Yifei Sun, Shen Fan, Zhongyao Wang, Jun Zhan, and Jingmin Chen. Who's Next: Rising Star Prediction via Diffusion of User Interest in Social Networks. *IEEE Transactions on Knowledge and Data Engineering*, 2022
5. **Xuan Yang**, Quanjin Tao, Xiao Feng, Donghong Cai, Xiang Ren, and Yang Yang. Multimodal Learning with Graph Alignment on Social Media. Preprint.
6. Taoran Fang, Zhiqing Xiao, Chunping Wang, Jiarong Xu, **Xuan Yang**, Yang Yang. DropMessage: Unifying Random Dropping for Graph Neural Networks. **Distinguished Paper Award**, *AAAI* 2023.

Selected Research Projects

Golden Data Curation for Large-scale Model Training — Durham
Duke University — *Sept. 2025 – Now*

- Formulated a Data Shapley framework over *probabilistic data distributions*, measuring expected contribution of different datasets to prioritize high-value data sources for large-scale LLMs training.
- Identified the chain effect in which 'golden' datasets influence the generation and composition of massive synthetic corpora for LLM post-training, and analyzed how golden-data and synthetic-data curation policies interact to preserve quality and diversity.

Batch-of-Thought: Cross-Instance Learning for Enhanced LLM Reasoning — Bellevue, WA
Tiktok — *Jun. 2025 – Sep. 2025*

- **Introduction:** Proposed a training-free Batch-of-Thought (BoT) method and a multi-agent reflection framework (BoT-R) that jointly reason to exploit mutual information beyond isolated inference.
- **Industrial impact:** Developed scalable multi-agent reasoning system that has been officially launched on Tiktok *E-commerce*, increasing pre-GMV fraud recall by **26%** and reducing case decision cost by **84%**.
- **Academic output:** *Batch of Thought: Cross-Instance Learning for improved LLM Reasoning,* Under Review.

Local Shapley: Efficient Data Valuation for Machine Learning Models Durham, NC

Duke University *Apr. 2024 – Oct. 2025*

- **Introduction:** Investigated Shapley-based data valuation through the lens of *locality*, designing algorithms that provably preserve fairness and uniqueness while delivering orders-of-magnitude speedups.
- **Contribution:** Developed an exact locality-aware method, *LSMR* (model reuse), and an approximation delivering $10^6\times$ lower time-to-ε at parity or better on data curation utility. Extend methods to GNN trainings.
- **Academic output:** *Local Shapley: Efficient Data Valuation for Model Training*, Under Review.

Unfolding and Modeling the Economic Recovery after COVID Lockdowns Hangzhou, China

Institute for Artificial Intelligence, Zhejiang University *Jun. 2021 – Feb. 2022*

- **Introduction:** Propose a GNN-based method over spatiotemporal electricity data to characterize post-lockdown urban recovery dynamics, informing economic resilience and policy planning.
- **Social impact:** Conducted a case study in Hangzhou city revealing heterogeneous recovery trajectories, cross-sector chain effects and policy impacts. The study was covered on the **front page of *Zhejiang Daily***.
- **Academic output:** *Unfolding and Modeling the Recovery Process after COVID Lockdowns, Scientific Reports 2023*

Alleviate Recommendation System Disequilibrium Hangzhou, China

Data and Technology Department, Alibaba *Jan. 2021 – May. 2021*

- **Introduction:** Framed the "rising-star" problem and proposed *RiseNet*, a GNN-incorporated time-series framework, to mitigate unfair exposure in online marketing recommendations.
- **Industrial impact:** Developed the RiseNet framework and a family-marketing recommendation model for *Taobao* (the largest e-commerce platform in China), boosting CTR by 2.3% over the production baseline.
- **Academic output:** *Who's Next: Rising Star Prediction via Diffusion of User Interest in Social Networks, IEEE Transaction on Knowledge and Data Engineering,TKDE*

Internship Experiences

Tiktok Bellevue, WA

***Research Intern**, Risk Control team* *May. 2025 – Nov. 2025*

Alibaba Group Hangzhou, China

***Research Intern**, Data Assets and Algorithm team* *Oct. 2020 – Dec. 2021*

Stanford University Palo Alto, CA

***Research Assistant**, Center for Magnetic Nanotechnology* *Jan. 2019 – Mar. 2019*

National University of Singapore Singapore

***Research Assistant**, Big Brain, BIGHEART* *Jun. 2018 – Aug. 2018*

Selected Honors

Excellent Postgraduate students' award	*2023*
Graduate of Triple A graduate, Zhejiang University	*2021 - 2022*
Tencent Technology Excellence Scholarship	*2021 - 2022*
First-class Academic Prize, Zhejiang University	*2021 - 2022*
Award of Honor for Graduate, Zhejiang University	*2021 - 2022*