

# Michiharu Yamashita

michiharu@psu.edu | mickeymst.github.io

## EDUCATION

---

- Pennsylvania State University** Aug 2020 – Present  
*Ph.D. in Information Science and Technology (Advisor: Prof. Dongwon Lee)* University Park, PA
- Thesis (expected): Advancing the Future of Work: Machine Learning in Job and Human Resource Domain
- Tokyo Institute of Technology** Apr 2015 – Mar 2017  
*M.Eng. in Computational Intelligence and Systems Science (Advisor: Prof. Kazuo Yano)* Tokyo, Japan
- Salutatorian (2nd place in the department)**
  - Thesis: Machine Learning for Work Environment Optimization to Improve Well-Being through Wearable Sensor
- University of Tsukuba** Apr 2011 – Mar 2015  
*B.S. in Management Science and Engineering (Advisor: Prof. Ushio Sumita)* Tsukuba, Japan
- Thesis: Network Analysis and Visualization for Mobile Applications' Competitiveness

## RESEARCH EXPERIENCE

---

- PIKE Research Group at Penn State** Aug 2020 – Present  
*Research Assistant (Advisor: Prof. Dongwon Lee)* University Park, PA
- Research Topics: NLP, RecSys, Graph Neural Networks, Adversarial Attacks, Computational Job Marketplace
  - Developing user-oriented job recommender systems, LLMs, and robust/secure models for online job marketplaces.
- Megagon Labs** Apr 2017 – Mar 2018  
*Research Engineer* Tokyo, Japan
- Research Topics: NLP, Entity Matching, Entity Extraction
  - Developed various entity matching models and embedded the modules into the company system.
- Hitachi Central Research Laboratory (Tokyo Institute of Technology)** Apr 2015 – Mar 2017  
*Research Assistant (Advisor: Prof. Kazuo Yano)* Tokyo, Japan
- Research Topics: Network Science, Wearable Sensors, People Analytics
  - Developed machine learning models for wearable sensors' data and visualized the sensor data into graph.
- Sumita Research Group at University of Tsukuba** Jan 2014 – Mar 2015  
*Research Assistant (Advisor: Prof. Ushio Sumita)* Tokyo, Japan
- Research Topics: Network Analysis, Visualization, Operations Research, Ranking Algorithm
  - Developed a competitive score algorithm for app installing with graph embeddings.

## EMPLOYMENT EXPERIENCE

---

- Visa** May 2024 – Aug 2024  
*Machine Learning Research Intern* Austin, TX
- Developed transformer-based foundational models from scratch for payment-domain downstream tasks on sequential tabular transaction datasets (Submitting a patent and a paper).
- Indeed** May 2022 – Aug 2022  
*Data Science and Research Intern* Austin, TX (Remote)
- Developed robust job recommender models to deliver high-quality matches to job seekers and employers.
  - Developed a large language model for multiple job-domain downstream tasks.
- Freelance** Jun 2019 – Jul 2020  
*Machine Learning Engineer* Tokyo, Japan
- Conducted machine learning projects with a big tech company, a tech startup, and an education startup.
  - Developed job mobility prediction models, machine learning models, query optimization tools, etc.
- Recruit Holdings** Apr 2017 – Jan 2019  
*Machine Learning Engineer* Tokyo, Japan
- Developed recommendation systems, multi-view click prediction models, and GIS-based applications.
  - Developed a pedestrian congestion visualization algorithm using GIS data and OpenStreetMap.
  - Developed the geo-topic model to obtain the user interest from POI.
- Ohma** Mar 2016 – Mar 2017  
*Software Engineer (Advisor: Prof. Yutaka Matsuo)* Tokyo, Japan
- Developed multiple ML-related systems: network visualization, entity extraction, face recognition, search engine.

- 16 Yeon-Chang Lee, JaeHyun Lee, **Michiharu Yamashita**, Dongwon Lee, and Sang-Wook Kim. CAPER: Enhancing Career Trajectory Prediction using Temporal Knowledge Graph and Ternary Relationship. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025
- 15 **Michiharu Yamashita**, Thanh Tran, and Dongwon Lee. OpenResume: Advancing Career Trajectory Modeling with Anonymized and Synthetic Resume Datasets. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024
- 14 Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason Samuel Lucas, **Michiharu Yamashita**, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. Authorship Obfuscation in Multilingual Machine-Generated Text Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024 (EMNLP)*, 2024
- 13 **Michiharu Yamashita**, Thanh Tran, and Dongwon Lee. Fake Resume Attacks: Data Poisoning on Online Job Platforms. In *Proceedings of the ACM Web Conference 2024 (WWW)*, 2024
- 12 Jason Lucas, Adaku Uchendu, **Michiharu Yamashita**, Jooyoung Lee, Shaurya Rohtagi, and Dongwon Lee. Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023
- 11 Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, **Michiharu Yamashita**, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023
- 10 **Michiharu Yamashita**, Jia Tracy Shen, Thanh Tran, Hamoon Ekhtiari, and Dongwon Lee. JAMES: Normalizing Job Titles with Multi-Aspect Graph Embeddings and Reasoning. In *2023 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2023
- 9 Yunqi Li, **Michiharu Yamashita**, Hanxiong Chen, Dongwon Lee, and Yongfeng Zhang. Fairness in Job Recommendation under Quantity Constraints. In *AAAI 2023 Workshop on AI for Web Advertising*, 2023
- 8 Jingyi Xie\*, **Michiharu Yamashita\***, Zekun Cai\*, and Aiping Xiong. A User Study on the Feasibility of Topic-aware Misinformation Warning on Social Media. In *Proceedings of the Human Factors and Ergonomics Society (HFES)*, 2022 (\* denotes co-first)
- 7 **Michiharu Yamashita**, Yunqi Li, Thanh Tran, Yongfeng Zhang, and Dongwon Lee. Looking Further into the Future: Career Pathway Prediction. In *ACM WSDM 2022 Workshop on Computational Jobs Marketplace*, 2022
- 6 Jia Tracy Shen, **Michiharu Yamashita**, Ethan Prihar, Neil Heffernan, Xintao Wu, Ben Graff, and Dongwon Lee. MathBERT: A Pre-trained Language Model for General NLP Tasks in Mathematics Education. In *NeurIPS 2021 Workshop on Math AI for Education*, 2021 (**Best Paper Award**)
- 5 Jia Tracy Shen, **Michiharu Yamashita**, Ethan Prihar, Neil Heffernan, Xintao Wu, Sean McGrew, and Dongwon Lee. Classifying Math Knowledge Components via Task-Adaptive Pre-Trained BERT. In *International Conference on Artificial Intelligence in Education (AIED)*. Springer, 2021
- 4 **Michiharu Yamashita**, Shota Katsumata, and Yusuke Fukasawa. Discovery of User Preferences from Big Geospatial Data Using Topic Models. In *2018 IEEE International Conference on Big Data (BigData)*. IEEE, 2018
- 3 **Michiharu Yamashita**, Hideki Awashima, and Hidekazu Oiwa. A Comparison of Entity Matching Methods between English and Japanese Katakana. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology at EMNLP*, 2018

- 2 Kent Kawai, **Michiharu Yamashita**, and Yutaka Matsuo. Face Recognition System Based on Convolutional Neural Network Robust to Occlusion and Low Quality Images. In *The 31st Annual Conference of the Japanese Society for Artificial Intelligence*. JSAI, 2017
- 1 **Michiharu Yamashita**, Nobuo Sato, and Kazuo Yano. Enhancing Collective Happiness by Controlling Room Temperature Using Wearable Sensor Data. In *The 2016 IEICE General Conference*. IEICE, 2016

---

## OPEN-SOURCE AND PUBLIC-FACING PRODUCT DEVELOPMENT

- OpenResume** Dec 2024  
*PIKE Research Group at Penn State*
- Developed and released anonymized and synthetic resume datasets for career modeling and job prediction tasks.
  - Validated the performance across key job prediction tasks.
- BigGorilla (Open-source Components for Data Integration)** Apr 2017 – Mar 2018  
*Megagon Labs*
- Developed an entity matching and entity extraction module and NLP related frameworks.
  - Applied NLP modules into the companies and promoted open-source components BigGorilla.
- Spysee2 (People Search Engine)** Mar 2016 – Apr 2017  
*Ohma, Inc.*
- Developed a network visualization search engine SPYSEE2 which had 1M+ visits per month.
  - Developed entity extraction, entity linking, and face recognition systems from unstructured and noisy web data.
  - Crawled millions of web pages efficiently using programs on AWS.
- Mobile App Competitiveness Visualization** Apr 2014 – Mar 2015  
*Fuller, Inc.*
- Developed a ranking algorithm for mobile app competitive analysis.
  - Developed a network visualization using app downloading flow.

---

## TEACHING EXPERIENCE

- IST 597: Explainable AI** Fall 2024  
*Teaching Assistant at Pennsylvania State University* *University Park, PA*
- REU class: Foundation of Representation Learning** Summer 2023  
*Lecturer at NSF Research Experiences for Undergraduates Program* *University Park, PA*
- REU class: Research Management** Summer 2023  
*Lecturer at NSF Research Experiences for Undergraduates Program* *University Park, PA*
- REU class: Foundation of Word Embeddings** Summer 2022  
*Lecturer at NSF Research Experiences for Undergraduates Program* *University Park, PA*
- REU class: Research Management** Summer 2022  
*Lecturer at NSF Research Experiences for Undergraduates Program* *University Park, PA*

---

## SERVICE

- Reviewer**
- 2023-present ACL Rolling Review (ARR)
  - 2022-present ACM SIGKDD (KDD)
  - 2024 EACL 2024 NLP4HR Workshop
  - 2024 ECML-PKDD 2024 AI4HR & PES Workshop
  - 2024 Nature Scientific Reports
  - 2023 IEEE Transactions on Computational Social Systems
  - 2023 IEEE Transactions on Big Data

## HONORS AND AWARDS

---

2023 Student Volunteer Grant for EMNLP 2023  
2023 IEEE CIS Travel Grant for DSAA 2023  
2021 Best Paper Award at NeurIPS 2021 Workshop on Math AI for Education  
2020 Funds for Excellence in Graduate Recruitment Scholarships from Penn State  
2018 Recruit Holdings The Best Freshman Award 2018  
2018 Recruit Holdings R&D MVP Award 2017  
2017 Full Repayment Exemption of Graduate Student Loan \$20,000 for Excellent Achievement  
2017 Salutatorian at Tokyo Institute of Technology  
2017 The Second Best Master Thesis Award from Tokyo Institute of Technology  
2015-2017 Full Tuition Exemption from Tokyo Institute of Technology  
2011-2015 Full Tuition Exemption from University of Tsukuba

## TECHNICAL SKILLS

---

**Languages:** Python, Java, C/C++, SQL, Swift, R, JavaScript, HTML/CSS

**Frameworks:** Tensorflow, Keras, PyTorch, Flask, Elasticsearch

**Developer Tools:** Docker, Google Cloud Platform, AWS, SageMaker