# Yuxin Xiao

yuxin102@mit.edu | xiaoyuxin1002.github.io | xiaoyuxin1002 | xiaoyuxin1002 | @YuxinXiao6

## Education

**Massachusetts Institute of Technology (MIT)** — Cambridge, MA

Ph.D. in Social and Engineering Systems — 09/2022 - Present

- Advised by **Prof. Marzyeh Ghassemi**; GPA: 5.00/5.00
- Affiliated with Institute for Data, Systems, and Society (IDSS) and Computer Science and Artificial Intelligence Laboratory (CSAIL)
- Advanced courses (received A/A+): Mathematical Statistics, Econometrics

**Carnegie Mellon University (CMU)** — Pittsburgh, PA

M.S. in Machine Learning — 08/2020 - 12/2021

- Advised by **Prof. Eric P. Xing** and **Prof. Louis-Philippe Morency**; GPA: 4.12/4.33
- Advanced courses (received A/A+): Advanced Machine Learning: Theory and Methods, Advanced Deep Learning, Convex Optimization, Probabilistic Graphical Models, Machine Learning with Large Datasets, Probability and Mathematical Statistics

**University of Illinois at Urbana-Champaign (UIUC)** — Urbana, IL

B.S. in Computer Science; B.S. in Statistics, Mathematics — 08/2016 - 05/2020

- Advised by **Prof. Jiawei Han** and **Prof. Hari Sundaram**; GPA: 3.93/4.00
- Advanced courses (received A/A+): Data Mining Principles, Advanced Information Retrieval, Advanced Social and Information Networks

## Awards & Honors

2020 **C. W. Gear Outstanding Undergraduate Award**, UIUC (2 at UIUC)

2020 **CRA Outstanding Undergraduate Researcher Award (Honorable Mention)**, CRA (4 at UIUC)

2019 **IEEE BigData 2019 Student Travel Award**, IEEE BigData

2016-2020 **Dean's List**, UIUC

2012-2015 **Senior-Middle 1 (SM1) Scholarship**, Ministry of Education, Singapore

## Publications

(* indicates equal contribution)

**In the Name of Fairness: Assessing the Bias in Clinical Record De-identification** — FAccT 2023

Yuxin Xiao*, Shulammite Lim*, Tom Joseph Pollard, Marzyeh Ghassemi *(oral presentation)* — [Paper], [Code]

**Uncertainty Quantification with Pre-trained Language Models: A Large-Scale Empirical Analysis** — EMNLP 2022

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, Louis-Philippe Morency *(findings)* — [Paper], [Code]

**SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction** — NAACL 2022

Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, Jiawei Han *(oral presentation)* — [Paper], [Code]

**Amortized Auto-Tuning: Cost-Efficient Bayesian Transfer Optimization for Hyperparameter Recommendation** — Preprint 2021

Yuxin Xiao, Eric P. Xing, Willie Neiswanger — [Paper], [Code]

**Heterogeneous Network Representation Learning: A Unified Framework with Survey and Benchmark** — TKDE 2020

Carl Yang*, Yuxin Xiao*, Yu Zhang*, Yizhou Sun, Jiawei Han *(230+ citations, 290+ GitHub stars and forks)* — [Paper], [Code]

**Discovering Strategic Behaviors for Collaborative Content-Production in Social Networks** — WWW 2020

Yuxin Xiao, Adit Krishnan, Hari Sundaram *(oral presentation)* — [Paper], [Code]

**Non-local Attention Learning on Large Heterogeneous Information Networks** — IEEE BigData 2019

Yuxin Xiao*, Zecheng Zhang*, Carl Yang, Chengxiang Zhai *(oral presentation)* — [Paper], [Code]

## Research Experience

**Healthy ML Group** — MIT

Supervisor: **Prof. Marzyeh Ghassemi** — 09/2022 - Present

**In the Name of Fairness: Assessing the Bias in Clinical Record De-identification**

- Examined the bias in de-identifying clinical discharge notes for patients of various demographic groups by conducting a large-scale empirical evaluation of nine popular de-identification baseline methods of different categories
- Prepared and annotated 100 note templates based on real-world clinical discharge notes and 16 name lists that were representative of different gender, race, name popularity, and the decade of popularity
- Investigated three factors that affected the de-identification quality and proposed an effective and method-agnostic solution by fine-tuning de-identification methods with clinical context and diverse names

### Bosch Center for Artificial Intelligence — Renningen, Germany

Supervisor: **Prof. Annemarie Friedrich**, **Prof. Heike Adel** — 06/2023 - 08/2023

**InstructTransfer: Transfer Tuning of Human-Language Instructions for Black-Box Large Language Models**

- Developed an automatic and cost-effective instruction-tuning strategy for black-box language models (e.g., ChatGPT) by transferring knowledge from prior tuning tasks via an optimal-transport-based task similarity kernel
- Outperformed existing tuning baselines on downstream natural language generation applications with only one-eighth of the tuning cost

### MultiComp Lab — CMU

Supervisor: **Prof. Louis-Philippe Morency** — 04/2021 - 04/2022

**Uncertainty Quantification of Pre-trained Language Models: A Large-Scale Empirical Analysis**

- Presented a holistic empirical analysis on how to compose a well-calibrated pre-trained language model-based prediction pipeline
- Examined how different pre-trained language models worked with various uncertainty quantifiers under distribution shifts and on diverse tasks
- Investigated the influence of pre-training datasets and strategies, model sizes, fine-tuning strategies, and uncertainty quantifiers on the calibration quality of pre-trained language models based on thousands of empirical observations
- Inspected the relationship between calibration quality and other aspects of model performance such as accuracy, sharpness, and robustness

### SAILING Lab — CMU

Supervisor: **Prof. Eric Xing** — 09/2020 - 12/2021

**Amortized Auto-Tuning: Cost-Efficient Bayesian Transfer Optimization for Hyperparameter Recommendation**

- Proposed a multi-task multi-fidelity Bayesian optimization framework with a novel task kernel and acquisition function. Leveraged cheap-to-obtain low-fidelity observations to efficiently recognize promising hyperparameters for new tuning tasks via knowledge transfer
- Analyzed the cost-efficiency and flexibility of existing hyperparameter tuning baselines by surveying 36 methods based on seven specific criteria
- Computed a hyperparameter recommendation database to serve the research communities, which consisted of 27 unique image classification tuning tasks and 150 distinct configurations over a 16-dimensional nested hyperparameter space

### Data Mining Group — UIUC

Supervisor: **Prof. Jiawei Han** — 03/2019 - 05/2021

**SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction**

- Developed a multi-task learning framework for document-level relation extraction, which consisted of a broad spectrum of carefully designed tasks to explicitly teach pre-trained language models to capture the key sources of information—relevant contexts and entity types
- Boosted the model performance further via evidence-based data augmentation and ensemble inference, while preserving the computational cost by assessing the uncertainty of model predictions
- Achieved state-of-the-art relation extraction results on three benchmarks and outperformed the runner-up by 5.04% relatively in retrieving interpretable evidence for each extracted relation

**Heterogeneous Network Representation Learning: A Unified Framework with Survey and Benchmark**

- Analyzed and evaluated the performance of various types of heterogeneous network embedding models (proximity-preserving based, message-passing based, relation-learning based) on different applications (node classification, link prediction)
- Designed and implemented a unified and user-friendly experiment interface for fair and efficient comparison of 13 popular heterogeneous network embedding algorithms on four large-scale benchmark datasets; received over 260 stars and forks on the corresponding GitHub repository

### Crowd Dynamics Lab — UIUC

Supervisor: **Prof. Hari Sundaram** — 10/2018 - 04/2020

**Discovering Strategic Behaviors for Collaborative Content-Production in Social Networks**

- Investigated the research question of whether resource-limited individuals were able to discover strategic behaviors associated with high payoffs when producing collaborative/interactive content in social networks
- Proposed a novel dynamic dual graph attention network which modeled individuals' content production strategies under social influence through a generative process
- Conducted a thorough qualitative analysis on a real-world social network with over seven million nodes and 400 million edges, which revealed three strong empirical findings about the emergence of individuals' strategies during the content production stage

## Industry Experience

### Cars.com — Chicago, IL

**Software Engineer Intern**, Mobile Development Team — 01/2018 - 08/2018

- Collaborated with Mobile App Team and UI/UX Team under the Agile framework (e.g. Stand-up, JIRA, Code Review)
- Refactored the Consumer Review Page of the company's Shop App by using Dependency Injection (Dagger 2) and MVC architecture, which doubled the page loading speed and greatly improved users' scrolling experience
- Designed and implemented the updating mechanism of the local JSON database by using Room Persistence Library

### Lenovo — Shanghai, China

**Software Engineer Intern**, Department of Telco Carrier Enablement and Customization — 06/2017 - 08/2017

- Examined the results of Compatibility Test Suite (CTS) for Android tablets
- Designed and maintained the UI Automator Testing to automatically change the language of Android tablets

## Teaching Experience

**Massachusetts Institute of Technology (MIT)** Cambridge, MA

Teaching Assistant & Session Instructor

**14.310x Data Science for Social Scientists**, Instructors: *Prof. Esther Duflo*, *Prof. Sara Ellison*; Coordinator: *Dr. Karene Chu* 06/2022 - 08/2022

- Designed and lectured 13 weekly online recitations and office hours for 30 students from Aporta in Peru as a part of the MicroMasters Program in Statistics and Data Science (SDS)
- Launched a set of guidelines and marking rubrics to guide students' year-long data science project in the program for the local NGOs in Peru
- Developed a series of final exam questions on statistics, machine learning, and econometrics

**University of Illinois at Urbana-Champaign (UIUC)** Urbana, IL

Course Assistant

**CS446 Machine Learning**, Instructor: *Prof. Sanmi Koyejo* 08/2019 - 12/2019

- Assisted in developing homework and exams, graded assignments for over 120 students
- Prepared and organized the final course project on the topic of automated machine learning

**CS410 Text Information Systems**, Instructor: *Prof. Chengxiang Zhai* 08/2019 - 12/2019

- Developed a software tool to intelligently analyze students' interactions and content contributions. Promoted students' online discussion on Piazza with the tool and evaluated their class participation more accurately and fairly

**CS125 Intro to Computer Science**, Instructor: *Prof. William Chapman* 01/2017 - 05/2017

- Guided students through lab activities during discussion sessions, solved students' problems during office hours

## Activities

**Chinese Students and Scholars Association (CSSA), UIUC** Urbana, IL

Member, Department of Undergraduates 09/2016 - 09/2017

- Led and organized the Chinese Food Festival (A Bite of China) in UIUC
- Organized and participated in the Chinese Lunar New Year Celebration

## Skills

**Programming** Python, Java, Kotlin, C++, C, R, JavaScript, Go

**Frameworks** PyTorch, TensorFlow, Android Platform, MySQL, Apache Spark, Hadoop DFS

**Languages** English (Fluent), Chinese (Native)

**Interests** Piano, Guitar, Badminton, Photography