

Education

- 2019 – **Stanford University**, *PhD Candidate in Computer Science*
- Advisor: Christopher Ré
- 2015-2019 **Princeton University**, *B.S.E in Operations Research and Financial Engineering (ORFE), certificate in Applications of Computing*, GPA: 3.962/4
- Graduated Summa Cum Laude
 - Senior Thesis: *A Quantum Version of the Multiplicative Weights Algorithm* (recipient of the Ahmet S. Çakmak Thesis Prize)
 - Thesis advisor: Elad Hazan

Research Interests

I am interested in using a theoretical lens to improve modern machine learning techniques from the perspective of data. Recently, I have been focusing on problems in data selection, data labeling, and data representations, especially in the setting where there are multiple input signals or objectives. Moving forward, I am particularly excited about developing a more principled understanding of how models learn from data.

Publications and Preprints

- **Skill-it! A data-driven skills framework for understanding and training language models.**
Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, Christopher Ré.
ICML Workshop on Data-centric Machine Learning Research, 2023.
- **Embroid: Unsupervised Prediction Smoothing Can Improve Few-Shot Classification .**
Neel Guha, Mayee F. Chen, Kush Bhatia, Azalia Mirhoseini, Frederic Sala, Christopher Ré.
In submission, 2023.
- **A case for reframing automated medical image classification as segmentation.**
Sarah Hooper, Mayee F. Chen, Khaled Kamal Saab, Kush Bhatia, Curtis Langlotz, Christopher Ré
In submission, 2023.
- **Ask Me Anything: A simple strategy for prompting language models.**
Simran Arora*, Avaniika Narayan*, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Christopher Ré.
International Conference on Learning Representations (ICLR), 2023. **Notable top-25%.**
- **Anomaly Detection with Multiple Reference Datasets in High Energy Physics.**
Mayee F. Chen, Benjamin Nachman, Frederic Sala.
Journal of High Energy Physics, 2023. *Machine Learning and the Physical Sciences (ML4PS) Workshop, NeurIPS*, 2022.
- **Reducing Reliance on Spurious Features in Medical Image Classification with Spatial Specificity.**
Khaled Saab, Sarah M. Hooper, Mayee F. Chen, Michael Zhang, Daniel Rubin, Christopher Ré.
Machine Learning for Healthcare (MLHC), 2022.
- **Shoring Up the Foundations: Fusing Model Embeddings and Weak Supervision**
Mayee F. Chen*, Daniel Y. Fu*, Dyah Adila, Michael Zhang, Frederic Sala, Christopher Ré.
Uncertainty in Artificial Intelligence (UAI), 2022. **Best Student Paper Runner-Up Award, Oral Presentation.**
- **Perfectly Balanced: Improving Transfer and Robustness of Supervised Contrastive Learning**
Mayee F. Chen*, Daniel Y. Fu*, Avaniika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, Christopher Ré.
International Conference on Machine Learning (ICML), 2022.
- **TABi: Type-Aware Bi-Encoders for Open-Domain Entity Retrieval**
Megan Leszczynski, Daniel Y. Fu, Mayee F. Chen, Christopher Ré.
Findings of ACL, 2022.
- **The Details Matter: Preventing Class Collapse in Supervised Contrastive Learning**
Mayee F. Chen*, Daniel Y. Fu*, Michael Zhang, Kayvon Fatahalian, Christopher Ré.
AAAI Workshop on Artificial Intelligence with Biased or Scarce Data, 2022. **Best Paper Award.**

- **An Adversarial Model of Network Disruption: Maximizing Disagreement and Polarization in Social Networks.**
Mayee F. Chen and Miklos Z. Racz.
IEEE Transactions on Network Science and Engineering (TNSE), 2021.
- **Mandoline: Model Evaluation under Distribution Shift**
Mayee F. Chen*, Karan Goel*, Nimit Sohoni*, Fait Poms, Kayvon Fatahalian, and Christopher Ré.
International Conference on Machine Learning (ICML), 2021.
- **Comparing the Value of Labeled and Unlabeled Data in Method-of-Moments Latent Variable Estimation**
Mayee F. Chen*, Benjamin Cohen-Wang*, Steve Mussmann, Frederic Sala, and Christopher Ré.
AISTATS, 2021.
- **Fast and Three-rious: Speeding Up Weak Supervision with Triplet Methods**
Daniel Y. Fu*, Mayee F. Chen*, Frederic Sala, Sarah M. Hooper, Kayvon Fatahalian, and Christopher Ré.
International Conference on Machine Learning (ICML), 2020.

Awards and Honors

- 2021 NSF GRFP Honorable Mention
- 2019 Ahmet S. Çakmak Prize, *Princeton University*, for innovative senior thesis research.
- 2018 Phi Beta Kappa, *Princeton University*, one of 28 early inductees.
- 2017 Tau Beta Pi Engineering Honor Society, *Princeton University*
- 2017 Shapiro Prize for Academic Excellence, *Princeton University*, awarded to top 2-3% of the class.

Work Experience

- 6/23-9/23 **Research Intern**, *Microsoft*, Redmond, WA, Office of Applied Research
- 2023 **Course Assistant**, *Stanford University*
 - CS228: Probabilistic Graphical Models (Winter)
- 2016 – 2019 **Grader for Computer Science Department**, *Princeton University*
 - COS226: Algorithms and Data Structures (lead grader), COS326: Functional Programming, COS340: Reasoning about Computation, COS324: Introduction to Machine Learning, and COS445: Economics and Computing
- 6/18-8/18 **Quantitative Trading Intern**, *IMC Trading*, Chicago, IL, Fixed Income, Currencies, and Commodities Desk
- 5/17-8/17 **Software Engineering Intern**, *Google*, Mountain View, CA, Advertiser Platform Team
Worked on AdWords Next Overviews, frontpage data analytics for ads campaigns
- 5/16-8/16 **Engineering Practicum Intern**, *Google*, Mountain View, CA, Cloud/Cluster/Kernel team
Worked on an infrastructure tool for pushing configuration and data updates to services within Google

Presentations

- 8/23 Talk at Allen AI: Skill-it! A data-driven skills framework for understanding and training language models
- 2023 Stanford Generative AI and Foundation Models Workshop: Embroid: Correcting Large Language Models With Auxiliary Embeddings
- 2022 Panelist at NeurIPS 2022 Tutorial on Theory and Practice of Efficient and Accurate Dataset Construction
- 2022 Machine Learning for Fundamental Physics at Lawrence Berkeley National Laboratory: Anomaly Detection with Multiple Reference Datasets
- 2022 Snorkel AI Machine Learning Whiteboard Talk: Liger: Fusing weak supervision with foundation model embeddings
- 2021 MedAI Talk Series: Correcting distribution shift in the ML model evaluation process
- 2021 DAWN Research Workshop: Mandoline: Model Evaluation under Distribution Shift
- 2020 Google x Stanford Summit: Labeled vs Unlabeled Data in Latent Variable Graphical Models

Coursework

Relevant graduate courses taken:

- Information Theoretic Lower Bounds in Data Science, Principles of Data-Intensive Systems, Convex Optimization II, Randomized Algorithms

Relevant undergraduate courses taken:

- *ORFE Courses*: Probability Theory (graduate-level course), Optimization, Decision Modeling for Business Analytics, Monte Carlo Simulation, Financial Mathematics, Analysis of Big Data, Probability and Stochastics, Statistics
- *Computer Science Courses*: Optimization for Machine Learning (graduate-level seminar), Computer Networks, Operating Systems, Economics and Computing, Introduction to Machine Learning, Neural Networks, Functional Programming, Programming Systems, Algorithms and Data Structures

Leadership and Activities

Reviewer, NeurIPS ('21-'23), ICML ('21-'23), ICLR ME-FoMo ('23), AISTATS ('23), UAI ('20, '23), KDD ('20)

At Stanford University:

- CS PhD Admissions Committee (2020-2022)
- WiML PhD Application Mentorship Program (2022)
- CS Student Applicant Support Program Volunteer (2020-2022)
- Graduate WiCS Mentor (2021-2022)
- CS Undergrad Mentorship Program (2021-2022)
- XTRM Kpop Cover Group: dance captain (2019–), Alliance Dance Team (2019–2020)

Skills

Advanced: Python, C, Java Intermediate: Go, OCaml, R, Dart, PyTorch Basic: Matlab, Julia