

Research Interests

Trustworthy machine learning, survival analysis, time series analysis, nonparametric prediction, healthcare

Academic Appointments

Carnegie Mellon University Pittsburgh, PA 1/2017 – present
Assistant Professor of Information Systems, Heinz College of Information Systems and Public Policy
Affiliated Faculty, Machine Learning Department

Education

Massachusetts Institute of Technology Cambridge, MA 9/2010 – 6/2015
Ph.D. in Electrical Engineering and Computer Science, 6/2015 GPA: 5.0/5.0
▪ Thesis: *Latent Source Models for Nonparametric Inference*
Received the George M. Sprowls award for best thesis in Computer Science
▪ Advisors: Polina Golland, Devavrat Shah
▪ Thesis reader: Sahand Negahban
Electrical Engineer (post-master's degree focused on coursework), 6/2014 GPA: 5.0/5.0
S.M. in Electrical Engineering and Computer Science, 6/2012 GPA: 5.0/5.0
▪ Thesis: *Deformation-Invariant Sparse Coding*
▪ Advisor: Polina Golland

University of California, Berkeley Berkeley, CA 8/2006 – 5/2010
B.S. with dual majors:
Electrical Engineering and Computer Sciences (*Highest Honors*)
Engineering Mathematics and Statistics (*Highest Honors*) GPA: 3.98/4.0

Selected Professional Experience

Co-Founder and Advisor, CoolCrop Bangalore, India 6/2016 – present
Providing refrigeration and marketing analytics to farmers. We provide refrigerators to farmers in rural India along with market forecasts to help them decide when and where to sell their harvest. I help develop machine learning approaches to produce price forecasting.

MITx Digital Learning Fellow & Postdoctoral Associate, MIT Cambridge, MA 9/2015 – 12/2016
New edX course on computational probability and inference. I was the primary instructor and course developer for a new free online intro-college/advanced-high-school-level course covering basic probability, probabilistic graphical models, and learning probability distributions (see below under the TEACHING section for details). I also managed an on-campus pilot of 15 students taking the residential version of the course to get their feedback on the online material being developed. During a preliminary stage of courseware development, I also supervised an undergraduate student and a recent PhD grad in polishing notes and making the material more accessible to a broader audience.

Co-Founder and Data Scientist, GridForm Cambridge, MA 1/2014 – 12/2016

Providing power to people. Over 1 billion people live in areas without electricity. To help renewable energy companies plan electrification projects, I worked on automatically analyzing satellite images of large tracts (> 100km²) of predominantly rural land. My team developed software that finds villages, simulates how to wire them up, and estimates the cost of thousands of wiring configurations per village. We regularly communicated with energy companies to make sure what we provide them is useful. We won the \$10,000 grand prize at the 2014 MIT IDEAS Global Challenge competition. My focus was on building machine learning algorithms for object recognition in satellite images.

Research & Development Intern, Data Science Team, Celect, Inc. Boston, MA 5/2015 – 8/2015

Predictive analytics. I developed machine learning solutions for recommending products to people for different retailers. My work involved deep learning/neural networks, collaborative filtering, and other prediction tools for big data.

Research Assistant, MIT Computer Science and Artificial Intelligence Lab Cambridge, MA 9/2010 – 6/2015

Theory for analyzing social data and medical images. To forecast whether a news topic will go viral on Twitter, we can compare it to past news topics with similar Tweet activity. More generally, we can make a prediction based on an observation by looking at similar past observations. My Ph.D. thesis develops theory to understand when, why, and how well such so-called “nearest-neighbor” methods work in applications such as forecasting trends on Twitter (time series classification), recommending products to people in systems like Netflix (online collaborative filtering), and finding where a human organ is in medical images by looking at image patches (patch-based image segmentation).

Modeling brain activation patterns. Brain activation patterns can vary significantly across people when they read the same sentences. My master’s thesis captures this variability in a new model for how different people’s brains respond to the same functional task such as language processing.

Real-time medical image analysis. Many algorithms for analyzing images use dimensionality reduction that is not fast enough for real-time applications. I developed a method that speeds up dimensionality reduction, enabling real-time medical imaging analysis such as tracking a patient’s breathing during radiation therapy and monitoring tissue heating when a patient is in an MR scanner.

Research Assistant, UC Berkeley Video and Image Processing Lab Berkeley, CA 5/2008 – 5/2010

Backpack with sensors for automated indoor modeling. I developed algorithms for tracking where a backpack is inside a building using data from laser scanners attached to the backpack. Thus, a person wearing the backpack could obtain accurate information on the backpack’s location over time even within buildings without reliable GPS access. The location information also fed into our team’s algorithms for constructing 3D models of building interiors.

Analyzing aerial images of cities. Urban planners often use labeled maps of cities to make decisions. I worked on algorithms for automatically locating where buildings, trees, road, and water are in aerial LIDAR images of cities.

Intern, Real-time Vision and Modeling Group, Siemens Corporate Research Princeton, NJ 7/2006 – 8/2006

Creating animated simulations of human crowds. I coded a PYTHON crowd simulator interfacing with POSER to create 3D animations of people in public indoor locations such as a subway. These synthetic animations could be used as test data for computer vision algorithms that recognize different human actions.

Teaching Experience

Instructor, *Unstructured Data Analytics for Policy* (CMU 94-775) Every spring term since Spring 2018

New half-semester course for public policy master's students. This course covers Python programming for data analysis, exploratory data analysis with a focus on processing text, and predictive data analysis with a focus on text and images. Students form teams to work on final projects that apply unstructured data analysis to address specific policy questions.

Instructor, *Unstructured Data Analytics* (CMU 95-865) Fall/spring terms since Fall 2017

New half-semester course for master's in information systems management students on analyzing data such as text and images that lack a pre-defined model and are hence often referred to as "unstructured". The course is heavily practical with all assignments done in Python Jupyter notebooks. Students learn to work with large datasets with the help of cloud computing on Amazon Web Services. I developed a suite of brand new lectures covering basic natural language processing, exploratory data analysis including manifold learning and topic modeling, and predictive data analysis including deep neural networks for analyzing images and time series. I led a team of teaching assistants to put together brand new homework assignments and a final exam.

Instructor, *Computational Probability and Inference* (MIT 6.008.1x, offered on edX) Fall 2016

New intro-college/advanced-high-school-level course covering basic probability, probabilistic graphical models, and learning probability distributions. All three of these main topics were covered with heavy emphasis on coding. The course prerequisites were comfort in Python programming and calculus. I developed nearly all the course notes, 75% of the videos, numerous new exercises, all the autograders, and a new two-part final project on learning structure in agriculture data and forecasting market prices. This course is based largely on the first half of an existing residential MIT course 6.008 that I helped develop (see below for the course *Introduction to Inference*) but differs in that its presentation has been made to be accessible to a much broader audience.

Computer Science Instructor, Middle East Entrepreneurs of Tomorrow (MEET) Summers 2015, 2016

MEET brings promising high school Palestinian and Israeli students together for a three-year program that teaches them computer science, entrepreneurship, and a deeper understanding of the conflict in the region. MEET aims to educate and empower these young leaders to create positive political and social change in the Middle East. I co-taught a three-week introduction to web development in Jerusalem for the students in their second year of MEET. For their final project, the students worked in binational teams to develop socially responsible web apps that promote cultural diversity, which they presented to Google Tel Aviv. Our material coverage included HTML, CSS, Bootstrap, Python, SQLAlchemy, Flask, and JavaScript. I created new lectures and labs on building web apps with Flask.

Teaching Assistant, *Introduction to Inference* (MIT 6.S080, now 6.008) Fall 2012, Spring 2014

New undergraduate core Electrical Engineering and Computer Science course introducing statistical inference and probabilistic graphical models. I taught for the class during the first two semesters that it had ever been offered. I developed substantial portions of the courseware, including Khan Academy style videos for students, a series of PYTHON robot tracking coding projects, new recitation notes, new problem sets, and more. For my teaching and courseware development, I won a departmental teaching award as well as the top campus-wide teaching award given to grad students.

Lecturer, *Statistics for Research Projects* (MIT 6.S085) January 2014

Two-week introduction to statistics course with an emphasis on recognizing when and how to apply different statistical tools to research problems. I updated the lectures to fold in ideas from predictive analytics and mathematical statistics.

Teaching Assistant, *Algorithms for Inference* (MIT 6.438) Fall 2011

Introductory graduate-level course on probabilistic graphical models. I made Khan Academy style videos and helped typeset the first complete set of lecture notes for the class. I also delivered three lectures for the Fall 2013 class.

Teaching Assistant, *Structure and Interpretation of Systems and Signals* (UCB EE20) Fall 2007, Spring 2008, Fall 2008

Undergraduate core Electrical Engineering and Computer Sciences course introducing the math behind processing images, audio, and video, and behind systems for communication (e.g., radio) and control (e.g., robotic manipulators). My third time teaching for this course (Fall 2008), I was the head teaching assistant, managing seven other teaching assistants. My teaching effectiveness was recognized by a campus teaching award.

Selected Awards and Honors

NSF CAREER award 2021 for “Real-Time Nonparametric Machine Learning for Healthcare with Guarantees” (award number 2047981)

Best paper award at AAI Workshop on AI for Behavior Change 2021

Best paper (theoretical track) at INFORMS Data Mining and Decision Analytics Workshop 2019

MIT George M. Sprowls Award 2015

Best Ph.D. thesis in Computer Science at MIT; selectivity: a few graduating Ph.D. students chosen each year from the EECS department

MIT Goodwin Medal 2015

Top campus-wide teaching award given to graduate students; selectivity: typically 1 student across all of MIT chosen per year (occasionally 2 students)

MIT IDEAS Global Challenge 2014: \$10,000 Grand Prize Winning Team

Annual MIT innovation and entrepreneurship competition focused on public service; selectivity: 4 teams out of 37

MIT EECS Frederick C. Hennie III Teaching Award 2013

Departmental teaching award; selectivity: ~1-3 students in the EECS department chosen per year

National Defense Science and Engineering Graduate Fellowship 2012-2015 (award year: 2012)

Three-year funding for graduate school; selectivity: ~200 US graduate students per year

National Science Foundation Graduate Research Fellowship 2011-2012 (award year: 2010)

Three-year funding for graduate school (I forfeited most of this due to other fellowships); selectivity: ~2,000 US graduate students per year

Siebel Scholarship 2011-2012

Funding for final year of my master's degree program; selectivity: within MIT EECS, 5 grad students chosen per year

MIT Irwin Mark Jacobs and Joan Klein Jacobs Presidential Fellowship 2010-2011

Funding for first year of graduate school; selectivity: ~15 out of the ~150 admitted EECS students per year

Eta Kappa Nu Outstanding Electrical and Computer Engineering Student Award 2010

For academic achievements, moral character, and service: 1 student chosen per year across universities with accredited electrical and computer engineering programs (nominations are done by Eta Kappa Nu chapters but the recipient need not be a member of Eta Kappa Nu)

UC Berkeley Engineering Science Departmental Citation 2010

Top graduating senior in UC Berkeley's Engineering Science program; selectivity: 1 student per year

UC Berkeley Outstanding Graduate Student Instructor Award 2009

Campus-wide teaching award (received as an undergraduate teaching assistant); selectivity: ~10% of teaching assistants

Intel International Science and Engineering Fair 2006: 1st place team

For high school research in computer science; selectivity: 7 teams out of 236

Papers

Note: "♣" denotes an author list that is alphabetical by last name, as is customary in fields like mathematics and theoretical computer science

23. (Working paper) "Consumer Behavior in the Online Classroom: Using Video Analytics and Machine Learning to Understand the Consumption of Video Courseware"
Mi Zhou, George H. Chen, Pedro Ferreira, Michael D. Smith
To appear in the *Journal of Marketing Research* (accepted July 2021)
22. (Working paper) "Influence via Ethos: On the Persuasive Power of Reputation in Deliberation Online"
Emaad Ahmed Manzoor, George H. Chen, Dokyun Lee, Michael D. Smith
(Currently under review)
Best paper at AAAI Workshop on AI for Behavior Change 2021
21. "Neural Topic Models with Survival Supervision: Jointly Predicting Time-to-Event Outcomes and Learning How Clinical Features Relate"
Linhong Li, Ren Zuo, Amanda Coston, Jeremy C. Weiss, George H. Chen
Artificial Intelligence in Medicine, August 2020
20. "Predicting Mortality Risk in Viral and Unspecified Pneumonia to Assist Clinicians with COVID-19 ECMO Planning"
Helen Zhou*, Cheng Cheng*, Zachary C. Lipton, George H. Chen, Jeremy C. Weiss
Artificial Intelligence in Medicine, August 2020
(Also presented at the *International Conference on Machine Learning (ICML) Workshop on Global Health*, July 2020)
19. "Deep Kernel Survival Analysis and Subject-Specific Survival Time Prediction Intervals"
George H. Chen
Machine Learning for Healthcare, July 2020
18. "Missing Not at Random in Matrix Completion: The Effectiveness of Estimating Missingness Probabilities Under a Low Nuclear Norm Assumption"
Wei Ma*, George H. Chen* (* = equal contribution)
Neural Information Processing Systems (NeurIPS), December 2019
Best paper (theoretical track) at INFORMS Data Mining and Decision Analytics Workshop 2019
17. "Truck Traffic Monitoring with Satellite Images"
Lynn H. Kaack, George H. Chen, M. Granger Morgan
Computing and Sustainable Societies (COMPASS), July 2019
(Also presented at the *International Conference on Machine Learning (ICML) Workshop on Climate Change*, June 2019)
16. "Nearest Neighbor and Kernel Survival Analysis: Nonasymptotic Error Bounds and Strong Consistency Rates"
George H. Chen
International Conference on Machine Learning (ICML), June 2019
15. "An Interpretable Produce Price Forecasting System for Small Farmers in India using Collaborative Filtering and Adaptive Nearest Neighbors"
Wei Ma, Kendall Nowocin, Niraj Marathe, George H. Chen
Information and Communication Technologies and Development (ICTD), January 2019
14. "Explaining the Success of Nearest Neighbor Methods in Prediction"
George H. Chen, Devavrat Shah
Foundations and Trends in Machine Learning, May 2018
13. "Survival-Supervised Topic Modeling with Anchor Words: Characterizing Pancreatitis Outcomes"
George H. Chen, Jeremy C. Weiss
Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning for Health, December 2017
(Also presented at *Society for Medical Decision Making North American Meeting*, October 2017)

12. "Toward Reducing Crop Spoilage and Increasing Small Farmer Profits in India: a Simultaneous Hardware and Software Solution"
George H. Chen, Kendall Nowocin, Niraj Marathe
Information and Communication Technologies and Development (ICTD), November 2017
11. "A Latent Source Model for Patch-Based Image Segmentation"
George H. Chen, Devavrat Shah, Polina Golland
Medical Image Computing and Computer Assisted Interventions (MICCAI), October 2015
10. "Targeting Villages for Rural Development Using Satellite Image Analysis"
Kush R. Varshney, George H. Chen, Brian Abelson, Kendall Nowocin, Vivek Sakhrani, Ling Xu, Brian L. Spatocco
Big Data, March 2015
9. "A Latent Source Model for Online Collaborative Filtering"
♣ Guy Bresler, George H. Chen, Devavrat Shah
Neural Information Processing Systems (NeurIPS), December 2014
Selected for spotlight (one of 62/1678 submissions)
8. "A Latent Source Model for Nonparametric Time Series Classification"
♣ George H. Chen, Stanislav Nikolov, Devavrat Shah
Neural Information Processing Systems (NeurIPS), December 2013
7. "Sparse Projections of Medical Images onto Manifolds"
George H. Chen, Christian Wachinger, Polina Golland
Information Processing in Medical Imaging (IPMI), June-July 2013
6. "Deformation-Invariant Sparse Coding for Modeling Spatial Variability of Functional Patterns in the Brain"
George H. Chen, Evelina G. Fedorenko, Nancy G. Kanwisher, Polina Golland
Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning and Interpretation in Neuroimaging,
December 2011
5. "Indoor Localization and Visualization Using a Human-Operated Backpack System"
Timothy Liu, Matthew Carlberg, George Chen, Jacky Chen, John Kua, Avidesh Zakhor
International Conference on Indoor Positioning and Indoor Navigation (IPIN), September 2010
4. "Indoor Localization Algorithms for a Human-Operated Backpack System"
George Chen, John Kua, Stephen Shum, Nikhil Naikal, Matthew Carlberg, Avidesh Zakhor
International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), May 2010
3. "Classifying Urban Landscape in Aerial LIDAR Using 3D Shape Analysis"
Matthew Carlberg, Peiran Gao, George Chen, Avidesh Zakhor
International Conference on Image Processing (ICIP), November 2009
2. "2D Tree Detection in Large Urban Landscapes Using Aerial LIDAR Data"
George Chen, Avidesh Zakhor
International Conference on Image Processing (ICIP), November 2009
1. "Image Augmented Laser Scan Matching for Indoor Dead Reckoning"
Nikhil Naikal, John Kua, George Chen, Avidesh Zakhor
International Conference on Intelligent Robots and Systems (IROS), October 2009

Selected Service

Taught a survival analysis tutorial "A Tour of Survival Analysis, from Classical to Modern" at the following venues:

- *SIGMETRICS* 2021, virtual
- *CHIL* 2020, virtual (co-taught with Jeremy C. Weiss with more of a healthcare emphasis)

Conference organizing committee:

- *CHIL* 2022 proceedings chair (joint with Gerardo Flores and Tom Pollard)
- *CHIL* 2021 proceedings chair (joint with Emily Alsentzer and Matthew McDermott) and track chair (joint with Rumi Chunara and Alistair Johnson for the track "Impact and Society")

Workshop organizing committees:

- *KDD Machine Learning for Consumers and Markets Workshop* 2021, virtual (joint with Wen Wang, Han Zhao, and Dokyun Lee)
- *APPAM 2019 Pre-Conference Workshop on ML and Public Policy* 2019, Denver, CO (co-led a session with Jens Ludwig "What New Challenges Arise When Applying ML Tools to Policy Problems?")
- *NeurIPS Nearest Neighbors Workshop* 2017, Long Beach, CA (joint with Devavrat Shah and Christina Lee Yu)

Reviewer for:

- *Neural Information Processing Systems* (2016, 2017, 2018, 2019, 2020, 2021)
- *International Conference on Machine Learning* (2018, 2019, 2020, 2021)
- *International Conference on Learning Representations* (2021, 2022)
- *Machine Learning in Healthcare* (2017, 2018, 2019, 2020, 2021)
- *Machine Learning for Health* (2017, 2018, 2019, metareviewer in 2020, metareviewer in 2021)
- *Conference on Health, Inference, and Learning* (metareviewer in 2021)
- *AAAI* (2017, 2018, 2020)
- *The Annals of Statistics* (2019, 2020)
- *Management Information Systems Quarterly* (2020)
- *Transactions on Pattern Analysis and Machine Intelligence* (2020)
- *Machine Learning* (2015, 2019)
- *IJCAI* (2020)
- *Digital Signal Processing* (2015)
- *Medical Image Computing and Computer Assisted Interventions* (2013)

Head of Curriculum, Middle East Entrepreneurs of Tomorrow (MEET) 10/2015 – 10/2016

Worked with the other Head of Curriculum and the team stationed in Jerusalem to establish the entrepreneurship and computer science curriculum for the 2016-2017 MEET school year that brings together Palestinian and Israeli high school students to learn and work together

Member, Task Force on the Future of MIT Education 4/2013 – 7/2014

Discussed and made recommendations for the future of MIT education both on and off campus, accounting for the growing prevalence of digital learning tools; I also led student outreach events and helped assemble student and faculty surveys on MIT education

Vice President of Resources, MIT Sidney-Pacific Graduate Community 4/2012 – 4/2013

Helped run the largest MIT graduate dorm Sidney-Pacific (700 residents) as one of the five executive officers, working with 50 other graduate student officers, the housemasters, and the building manager to put on copious amounts of events, start a community garden, launch a new website, overhaul building resources, handle severe conflicts, and more; I managed 10 officers and the \$140,000 budget