Sarah Tan

CONTACT INFORMATION	Redwood City, CA https://shftan.github.io	512-716-9711 ht395@cornell.edu

OBJECTIVE

Seeking **full-time position** in applied machine learning research / data science. My dissertation research is on interpretability, causal inference methods, and social good applications.

EDUCATION

Cornell University

PhD Statistics; Minor in Computer Science   **2013 - Expected May 2019**
- Advisors: Giles Hooker, Martin Wells (Cornell Statistics)
- Committee member: Thorsten Joachims (Cornell Computer Science)
- External committee member: Rich Caruana (Microsoft Research)
- Affiliation: Cornell Algorithms, Big Data, and Inequality Program

University of California San Francisco (UCSF)

Visiting Graduate Student   **Jan 2018 - present**
- Host: Charles McCulloch (UCSF Epidemiology and Biostatistics)
- Ongoing Collaborations:
  - Zuckerberg San Francisco General Hospital (Probing the need and use cases for interpretability in clinical decision support systems)
  - Memory and Aging Center (Deep learning on MRI images for dementia diagnosis)

University of California Berkeley

BA (Honors) Statistics, Economics; Minor in Operations Research   **2006 - 2010**

PROFESSIONAL EXPERIENCE

Microsoft Research   Redmond, WA

Research Intern   Mentors: Rich Caruana, Kori Inkpen, Ece Kamar

Focus Areas: Interpretability, Algorithmic Fairness   **Summers 2017 & 2018**
- Developed global interpretability method for fully-connected neural nets to characterize the relationship between tabular data features and neural net predictions.
- Extended model distillation techniques to inspect criminal justice and credit risk scoring models for potential bias.

Data Science for Social Good   Chicago, IL

Summer Fellow   Mentor: Rayid Ghani (Obama 2012 Chief Data Scientist)   **Summer 2014**
- Developed predictive models to help a nonprofit identify clients at risk for attrition or noncompliance; wrote blog post to describe findings for non-technical audience.

Johnson Research Labs | Startup research lab   New York, NY

Research Scientist. Focus Area: Computational Social Science   **2012 − 2013**
- Topic modeling on tweets, news articles, and other media content to investigate the influence of social issue documentaries on public opinion and legislation
- Web scrapping to assemble database of movie information from various online media sources, including IMDB, Rotten Tomatoes, Box Office Mojo, etc.

New York City Health + Hospitals | Public hospitals system

Research Assistant (Part-Time), Statistics & Data Quality Group   **Oct 2011 − Aug 2013**
- Pulled data from electronic medical records and applied statistical models to develop predictive models of care quality, hospital readmissions, and adverse drug reactions.

For my complete work experience, please see my LinkedIn.

PUBLICATIONS

Tan, R Caruana, G Hooker, Y Lou. *Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation*, ACM/AAAI AI, Ethics, Society Conference (**Oral talk**), 2018.

- Media coverage: MIT Technology Review, Politico, Futurism, WorkFlow

Tan, S Makela, D Heller, K Konty, S Balter, T Zheng, JH Stark. *Using Bayesian Evidence Synthesis to Estimate Disease Prevalence Among Hard-To-Reach Populations*, Epidemics, 2018.
- Presented to NYC Health Commissioner

Tan, G Hooker, MT Wells. *Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable.* NIPS Interpretability Workshop, 2016.

Tan, G Hooker, MT Wells. *Probabilistic Matching: Incorporating Uncertainty to Correct for Selection Bias.* NIPS Causal Inference Workshop, 2016.

IB Vasi, E Walker, JS Johnson, **Tan**. *"No Fracking Way!" Media Activism, Discursive Opportunities and Local Opposition against Hydraulic Fracturing in the United States, 2010-2013*, American Sociological Review, 2015.
- **2 Best Paper Awards** from American Sociological Association
- Media coverage: The Guardian, The Atlantic, Pacific Standard

Tan, DI Miller, J Savage. *Proximity Score Matching: Locally Adaptive Matching for Causal Inference*, Atlantic Causal Inference Conference (Lightning talk), 2015.
- **1 of 3 Best Student Paper Awards** from American Statistical Association SSPA section

For all my publications, please see my Google Scholar.

WORK UNDER REVIEW	**Tan**, R Caruana, G Hooker, P Koch, A Gordo. *Learning Global Additive Explanations for Neural Nets Using Model Distillation*

Tan, J Adebayo, K Inkpen, E Kamar. *Investigating Human + Machine Complementarity for Recidivism Predictions*

X Zhang, **Tan**, P Koch, Y Lou, U Chajewska, R Caruana. *Interpretability is Harder in the Multiclass Setting: Axiomatic Interpretability for Multiclass Additive Models*

GRANTS AND FELLOWSHIPS		
• Microsoft Research Dissertation Grant ($25,000)		**2018**
• American Statistical Association Wray Jackson Smith Award ($1,000)		**2017**
• Engaged Cornell Grant for Community-Engaged Dissertation Research ($15,000)		**2017**
• Harmony Institute Research Fellowship ($15,000)		**2016**

INVITED TALKS		
• AT&T Labs Graduate Student Symposium		**Nov 2018**
• UC Santa Cruz Responsible Data Science Seminar. *Host: Lise Getoor*		**May 2018**
• Novartis Pharmaceuticals. *Host: Statistics Methodology Group*		**April 2018**
• UCSF Medical Interpretability Seminar. *Host: Gilmer Valdes*		**March 2018**

SERVICE		
• *Co-organizer*, ICLR Workshop "Debugging Machine Learning Models"		**2019**
• *Board of directors*, Women in Machine Learning organization		**2018**
• *Co-organizer*, Invited Session "New Advances in Causal Inference for Longitudinal and Survival Data" at International Conference on Health Policy Statistics (ICHPS)		**2018**
• *Scientific committee (student representative)*, ICHPS		**2018**
• *Co-organizer*, "Statistics for Social Good" Session at Joint Statistical Meetings		**2016**
• *Co-organizer*, Women in Machine Learning Workshop (600 attendees, 200 posters)		**2016**

PROGRAMMING SKILLS	R, Python

SOFTWARE	R package surfin: (Statistical Inference for Random Forests)