Yiqun T. Chen

Website: https://yiqunchen.github.io/
Google scholar: https://tinyurl.com/yiqunc-scholar
Email: yiqunc@stanford.edu

PROFESSIONAL EXPERIENCE

Postdoctoral Scholar
Fall 2022 – Now — Stanford University
Hosted by Dr. James Zou and supported by a **Stanford Data Science Fellowship**.

EDUCATION

Ph.D. in Biostatistics — GPA: 3.9/4.0
Fall 2017 – Summer 2022 — University of Washington, Seattle
Dissertation Committee: Dr. Daniela Witten (Chair), Dr. Alex Luedtke, Dr. Amy Willis, Dr. Jamie Morgenstern, Dr. Kelley Harris.
Relevant Coursework: Theoretical Statistics, Advanced Regression Methods for Independent and Dependent Data, Causal Inference, Convex Optimization.
Thomas R. Fleming Excellence in Biostatistics Award

B.A. in Statistics, Computer Science, and B.S. in Chemical Biology — GPA: 3.9/4.0
Fall 2013 – Summer 2017 — University of California, Berkeley
High Distinction in General Scholarship

PUBLICATIONS AND PREPRINTS

Statistical Machine Learning Methodology:

1. **Chen YT**, Zou J (2023+). A Simple But Hard-to-Beat Foundation Model for Genes and Cells Built From ChatGPT.
   - Preprint: `https://www.biorxiv.org/content/10.1101/2023.10.16.562533v1`
   - A preliminary of this work has been accepted as an oral presentation at MLCB 2023 ($< 15\%$ acceptance rate).
2. **Chen YT**, Gao LL (2023+). Testing for a difference in means of a single feature after clustering. The manuscript has been submitted to AISTATS for review.
3. **Chen YT**, Witten DM (2023). Selective inference for k-means clustering. In press at *Journal of Machine Learning Research*.
   - Link: `https://www.jmlr.org/papers/v24/22-0371.html`
   - Awarded student research award for this work at 35th New England Statistics Symposium in 2022.
4. **Chen YT**, Jewell SW, and Witten DM (2023). More powerful selective inference for the graph fused lasso. To appear in *Journal of Computational and Graphical Statistics* arXiv link: `https://arxiv.org/abs/2109.10451`.
5. **Chen YT**, Jewell SW, and Witten DM (2022). Quantifying uncertainty in spikes estimated from calcium imaging data. To appear in *Biostatistics*. arXiv link: `https://arxiv.org/abs/2103.07818`.
   - Awarded the Best Student Oral Presentation at WNAR 2021 for this work.

Applications in Data Science:

6. **Chen YT**, Zou J (2023+). TWIGMA: A dataset of AI-Generated Images with Metadata From Twitter. Accepted by *NeurIPS 2023*; arXiv link: `https://arxiv.org/abs/2306.08310`.
7. Liang W, Rajani N, Yang X, Ozoani E, Wu E, **Chen YT**, SS Daniel, M Daniel, Zou J (2023+) What's documented in AI? Systematic Analysis of 32K AI Model Cards. Under submission at *Nature Machine Intelligence*.
8. **Chen YT**, Smith AD, Reinecke K, and To A (2023). Why, when, and from whom: considerations for collecting and reporting race and ethnicity data in HCI. In *CHI'23*.
    - Awarded Best Paper Honorable Mention (Top 5% of all submissions to CHI 2023); also covered in `https://www.khoury.northeastern.edu/awards-ethics-cross-college-collaboration-northeastern-at-chi-2023`.
    - This paper reviewed six years of CHI proceedings and interviewed 15 authors to understand the current practice of collecting race and ethnicity data in the field of human-computer interaction. It highlights important considerations for researchers when they decide to collect such data.
9. **Chen YT**, Smith AD, Reinecke K, and To A (2022). Collecting and Reporting Race and Ethnicity Data in HCI. In *CHI'22 Extended Abstracts*.
10. **Chen YT**, Gopinath R, Tadakamalla A, Ernst MD, Holmes R, Fraser G, Ammann P, Just R. Revisiting the relationship between fault detection, test adequacy criteria, and test set size. In: *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE).* 2020:237-249.
    - This paper addressed a long-standing paradoxical observation in the field of software engineering from a statistical perspective, and has been cited by 40+ papers in the field since publication.

Applications in Biology and Epidemiology:

11. **Chen YT***, Gui H*, Yao C, Adu-Brimpong J, Javitz S, Golovko V, Ko J, Daneshjou R, Chiou A. (2023+) A Novel Triage and Referral Pathway for Single Lesions Improves Skin Cancer Risk Stratification and Efficiency in Dermatology Clinics. Submitted to *JAMA Dermatology*; * indicates joint first authorship
    - We evaluated the efficacy in a large-scale cohort study; this work has led to a grant application to further improve the risk stratification of patients referred to Dermatology using multi-modal data and deep learning methods.
12. Schurr MS, **Chen YT**, Raue PJ, Arean PA, Alexopoulos GS, Renn BN (2023+). Changes in Executive Dysfunction During the Course of Brief Psychotherapies for Late-Life Depression and Relation to Treatment Response. Submitted to *American Journal of Geriatric Psychiatry*.
    - I mentored the student author on the longitudinal data analysis component of the paper.
13. Veseli I, **Chen YT**, Schechter MS, Vanni C, Fogarty EC, ..., Murat Eren, A. (2023+). Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. To appear in *eLife*.
    - Applied selective inference and stability selection methodology to real data applications and assisted with overall statistical analyses.

14. **Chen YT**, Williamson BD, Okonek T, Wolock C, Spieker AJ, Hee Wai YT, Hughes JP, Emerson SS, and Willis AD (2023). rigr: Regression, Inference, and General Data Analysis Tools in R. *Journal of Open Source Software*, 7(80), 4847, https://doi.org/10.21105/joss.04847

    - This paper introduced a package that facilitates common data analysis routines in biomedicine and epidemiology, with an emphasis on straightforward, modern regression modeling in `R`.

15. **Chen YT***, Marquez C*, Atukunda M, Chamie G, Balzer LB, ..., Charlebois ED, Havlir DV, and Petersen ML (2023). The Association Between Social Network Characteristics and TB Infection Among Adults in Nine Rural Ugandan Communities. To appear in *Clinical Infectious Diseases*; * indicates joint first authorship.

    - Preliminary results from this paper were accepted as an oral presentation at the 23rd International AIDS Conference in 2019.
    - This work led to lasting collaboration and a paid consulting position as part of NIHR01AI151209 (Nov 16, 2020 - Oct 31, 2025).

16. **Chen YT**, Brown LB, Chamie G, Kwarisiima D, Ayieko J, Kabami J, Charlebois E, Clark T, Kamya M, Havlir DV, Petersen ML, and Balzer LB (2021). Social networks and HIV care outcomes in rural Kenya and Uganda. *Epidemiology*, 32(4):551-559.

    - Awarded a New Investigator Scholarship for CROI 2020 for a preliminary of this work.

17. Brown L, Balzer L, Kabami J, Kwarisiima D, Sang N, Ayieko J, **Chen Y**, Chamie G, Charlebois E, Camlin C, Cohen C, Bukusi E, Kamya MR, Moody J, Havlir D, Petersen M (2020). The influence of social networks on antiretroviral therapy initiation among HIV-infected antiretroviral therapy-naive youth in rural Kenya and Uganda. *J Acquir Immune Defic Syndr.* 83(1):9-15.

18. **Chen Y**, Zheng W, Brown LB, Chamie G, Kwarisiima D, Kabami J, Clark TD, Sang N, Ayieko J, Charlebois ED, Jain V, Balzer L, Kamya MR, Havlir D, Petersen M, the SEARCH Collaboration (2019). Semi-supervised record linkage for construction of large-scale sociocentric Networks in resource-limited settings: an application to the SEARCH study in rural Uganda and Kenya. arXiv preprint. arXiv link: `http://arxiv.org/abs/1908.09059`.

19. Jakobson C, **Chen Y**, Slininger M, Valdivia E, Kim E, Tullman-Ercek D (2016). Tuning the catalytic activity of subcellular nanoreactors. J Mol Biol. 428(15):2989-2996.

SELECTED TALKS

1. (August 2023; invited) Joint Statistical Meetings 2023, Toronto, ON, Canada.
2. (June 2023; invited) ICSA 2023 Applied Statistics Symposium, Ann Arbor, MI, USA.
3. (November 2022; invited) "Selective inference for k-means clustering" at Rising Stars in Data Science conference, Chicago, IL, USA.
4. (August 2022; invited) "Selective inference for k-means clustering" at COMPSTAT 2022, Bologna, Italy.
5. (August 2022; contributed) "Selective inference for k-means clustering" at Joint Statistical Meetings 2022, Washington DC, USA.
6. (May 2022; contributed) "Selective inference for k-means clustering" at 35th New England Statistics Symposium, Storrs, CT, USA.

- Student Paper Award

7. (May 2022; contributed) "Collecting and Reporting Race and Ethnicity Data in HCI" at CHI 2022, New Orleans, LA, USA.
8. (April 2022; invited) "Selective inference for k-means clustering" at Young Data Science Researcher Seminar at ETH Zürich, virtual.
9. (April 2022; invited) "Selective inference for k-means clustering" at International Seminar on Selective Inference, virtual.
10. (January 2021; invited) "Selective inference for k-means clustering", James Zou group at Stanford, virtual.
11. (January 2021; invited) "Selective inference for k-means clustering", Kathryn Roeder group at CMU, virtual.
12. (December 2021; contributed) "Selective inference for k-means clustering" at UW Biostatistics student seminar, and at SLAB LAB group meeting, Seattle, WA, USA.
13. (June 2021; contributed) "Quantifying uncertainty in spikes estimated from calcium imaging data" at the 2021 WNAR Annual Meeting of International Biometric Conferences, Anchorage, AK, USA (virtual due to the COVID-19 pandemic).
    - Best Student Oral Presentation
14. (September 2020; contributed) "Revisiting the relationship between fault detection, test adequacy criteria, and test set size" at the 2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE), Melbourne, Australia (virtual due to the COVID-19 pandemic).
15. (May 2020; invited) "Social networks and HIV care outcomes in rural Kenya and Uganda" at the UCSF social network and HIV workshop, San Fransisco, CA, USA, (canceled due to the COVID-19 pandemic).
16. (March 2020; contributed) "HIV+ persons in rural Uganda with fewer social connections have lower HIV suppression" at the 2020 Conference on Retroviruses and Opportunistic Infections (CROI), Boston, MA, USA, (virtual due to the COVID-19 pandemic).
    - New Investigator Scholarship for CROI 2020

HONORS & AWARDS

- Selected to participate in the Cornell Young Researchers Workshop Fall 2023
- Best Paper Honorable Mention CHI (Top 5% of all submissions) Winter 2023
- Rising Stars in Data Science, University of Chicago DSI Autumn 2022
- Thomas R. Fleming Excellence in Biostatistics Award Spring 2022
  - UW Biostatistics' most prestigious student honor
- NESS Student Research Award Spring 2022
- Outstanding Teaching Assistant Award, School of Public Health, University of Washington Spring 2022
- Data Science Postdoctoral Fellowship at Stanford Spring 2022
- Biostatistics Retention and Success Scholarship Winter 2022
- Best Student Oral Presentation at WNAR 2021, for "Quantifying uncertainty in spikes estimated from calcium imaging data" Spring 2021
- Scholarship for 6th Seattle Symposium in Biostatistics Fall 2020

- New Investigator Scholarship for CROI 2020 Spring 2020

  UC Berkeley
- Dean's List (Awarded to top 4% Students) Fall 2013 – Spring 2017
- Percy Lionel Davis Award for Excellence in Scholarship in Mathematics Spring 2017
- Senior Research Award, College of Chemistry Spring 2017
- Best Poster Presentation, Institute of International Studies Spring 2017
- Scholarship for Research Merit, Institute of International Studies Fall 2016
- Scholarship for Academic Excellence, International Office Fall 2014
- Elected to join Phi Beta Kappa Fall 2014

SELECTED TEACHING EXPERIENCE

Teaching and Mentoring at Stanford University

- Selected to participate in the Inclusive Mentorship Workshop, organized by the Office of Postdoctoral Affairs.
- Finalist for the university-wide "Peer Mentors for Postdoctoral Scholars" program.

Teaching Assistant at University of Washington

- Summer Institute in Statistical Genetics (Instructors: Ken Rice & Ting Ye)
- Summer Institute in Statistics for Clinical & Epidemiological Research (Instructors: Katie Wilson & Anna Plantinga)
- Longitudinal Data Analysis (BIOSTAT 540; graduate-level; course rating: 4.7/5.0 ($n = 47$); Instructor: Katie Wilson)
- Categorical Data Analysis (BIOSTAT 536; graduate-level; Instructor: Katie Kerr)
  - Awarded the Outstanding Teaching Assistant Award at the University of Washington, School of Public Health.
- Introductory Laboratory Based Biostatistics (UCONJ 510; graduate-level; Instructor: Lloyd Mancl)
- Machine Learning for Biomedical and Public Health Data (BIOSTAT 546; graduate-level; Instructor: Daniela Witten)

Mentor for Directed Reading Program at University of Washington

- Mentored undergraduate students on the topic of identification in missing data and causal inference.
- Student presentation can be found at `https://spa-drp.github.io/writeups/win2021/suh_slides.pdf`.

Guest Lectures at University of Washington

- Machine Learning for Biomedical and Public Health Data (BIOSTAT 546). Guest lectures on decision trees, support vector machines (SVM), and principal component analysis (PCA).

Teaching Assistant at University of California, Berkeley

- Introduction to Machine Learning (CS 189/289A; advanced undergraduate-level; Fall 2016 & Spring 2017; Instructors: Ben Recht & Jitendra Malik)

- Discrete Mathematics and Probability (CS 70; undergraduate-level; Summer 2016 & Spring 2017; Instructor: Satish Rao)

SERVICE

Internal

Stanford University

- Event organizer at Data Science for Health Center Fall 2022-Now

University of Washington, Seattle

- Student Representative on the Admission Committee Winter 2022
- Student Organizer of the General Exam Information Session Winter 2022
- Panelist for Admitted Student Visit Day Winter 2021, Winter 2022
- Panelist for School of Public Health New Student Welcome Fall 2021
- Student Representative On the Curriculum Committee Fall 2020 – Spring 2021
- Member of the student organization Biostatistics Activities and Events Squad Fall 2018 – Fall 2021
- Peer mentor for Biostatistics Peer Mentoring Program Fall 2018 – Fall 2020
- Panelist for the department's internship workshop Fall 2018, Fall 2021

UC Berkeley

- Member of the Vice Chancellor's student committee Fall 2016

External

- Reviewer for *Biostatistics*, *Biometrika*, *Bioinformatics*, *Annals of Statistics*, *Journal of Machine Learning Research*, *Science Advances*, *Nature Machine Intelligence*, *Journal of Investigative Dermatology*, *American Journal of Epidemiology*, CHIL 2022, CHI 2022, CHI 2023, Neurips 2023, PSB 2022, PSB 2023

SELECTED INDUSTRY EXPERIENCE

Data Scientist Intern June 2019 – September 2019
Waymo LLC (formerly Google self-driving car project), Mountain View, CA

- Project: Modeling self-driving vehicles' planning and reaction time using tools from causal inference and machine learning.

Applied Scientist Intern June 2018 – September 2018
A9.com (now Amazon Search), Palo Alto, CA

- Project: Using deep learning-based language models for better summaries of the search queries on `amazon.com`.

SELECTED SOFTWARE

1. `rigr`:
   - Co-authors of `rigr`, an `R` package that streamlines data analysis for biomedical researchers and students in introductory to statistics classes.
2. `KmeansInference`:
   - An `R` package for testing for a difference in means between groups identified via k-means clustering.
   - Tutorials are available at `https://yiqunchen.github.io/KmeansInference/`.
3. `SpikeInference`:
   - An `R` (which serves as a wrapper for `c++` code) package that quantifies uncertainty for spikes estimated from calcium imaging data.
   - Tutorials are available at `https://yiqunchen.github.io/SpikeInference/`.