Yan Shuo Tan

yanshuo@nus.edu.sg LinkedIn Google Scholar

Academic Appointments

Assistant Professor *2022—*
Department of Statistics and Data Science, National University of Singapore
- Faculty Affiliate, Institute of Data Science
- Faculty Affiliate, NUS AI Institute

Neyman Visiting Assistant Professor *2020—2021*
Department of Statistics, University of California, Berkeley
- *Mentor: Prof. Bin Yu*

Post-doctoral Fellow *2021—2022*
Department of Statistics, University of California, Berkeley
- *Mentor: Prof. Bin Yu*

Post-doctoral Fellow *2018—2019*
Department of Statistics, University of California, Berkeley
- *Mentor: Prof. Bin Yu*

Patrick J. McGovern Research Fellow *2018—2018*
Simons Institute, University of California, Berkeley

Education

Ph.D., Mathematics *2013—2018*
University of Michigan, Ann Arbor, USA
- Advisors: *Prof. Roman Vershynin & Prof. Anna Gilbert*

B.S., Mathematics (Honors) *2009—2013*
University of Chicago, Chicago, USA

Research Interests

Theoretical and applied aspects of statistical machine learning and data science.
- Theory and methodology for tree-based methods and ensembles
- Bayesian tree ensemble modeling
- In-context learning for tabular data

Grants

MOE AcRF Tier 1 Grant (NUS Cross Faculty Grant) `A-8004458-00-00`, $250,000 *2026*

MOE AcRF Tier 1 Grant `A-8002498-00-00`, $250,000 *2024*

NUS Start-up Grant `A-8000448-00-00`, $180,000 *2022*

Achievements & Awards

NUS Inauguration Grant *2022*

US Junior Oberwolfach Fellowship *2019*

Patrick J. McGovern Research Fellowship *2018*

Paul R. Cohen Prize *2013*
Awarded to graduating seniors with the highest record in mathematics (5 awarded per year)

Phi Beta Kappa *2012*

Journal Publications

(∗ equal contribution or alphabetical ordering, † student/postdoc supervised, ‡ corresponding author)

1. **Yan Shuo Tan**, Jason M Klusowski, and Krishnakumar Balasubramanian, "Statistical-Computational Trade-offs for Recursive Adaptive Partitioning Estimators", *Annals of Statistics*, 2026.
2. Qiong Zhang*, **Yan Shuo Tan***, and Jiahua Chen, "Byzantine-Tolerant Distributed Learning of Finite Mixture Models", *Journal of the Royal Statistical Society, Series B*, 2026.
3. **Yan Shuo Tan***, Chandan Singh*, Keyan Nasseri*, Abhineet Agarwal*, James Duncan, Omer Ronen, Matthew Epland, Aaron Kornblith, and Bin Yu, "Fast Interpretable Greedy-Tree Sums", *Proceedings of the National Academy of Sciences*, vol. 122, no. 7, pp. e2310151122, 2025.
4. **Yan Shuo Tan** and Roman Vershynin, "Online Stochastic Gradient Descent with Arbitrary Initialization Solves Non-Smooth, Non-Convex Phase Retrieval", *Journal of Machine Learning Research*, vol. 24, no. 58, pp. 1–47, 2023.
5. Chandan Singh, Keyan Nasseri, **Yan Shuo Tan**, Tiffany Tang, and Bin Yu, "imodels: A Python Package for Fitting Interpretable Models", *Journal of Open Source Software*, vol. 6, no. 61, pp. 3192, 2021.
6. John Lipor, David Hong, **Yan Shuo Tan**, and Laura Balzano, "Subspace Clustering Using Ensembles of K-Subspaces", *Information and Inference: A Journal of the IMA*, vol. 10, no. 1, pp. 73–107, 2021.
7. Nick Altieri*, Rebecca L Barter*, James Duncan*, Raaz Dwivedi*, Karl Kumbier*, Xiao Li*, Robert Netzorg*, Briton Park*, Chandan Singh*, **Yan Shuo Tan***, Tiffany Tang*, Yu Wang*, Chao Zhang*, and Bin Yu*, "Curating a COVID-19 Data Repository and Forecasting County-Level Death Counts in the United States", *Harvard Data Science Review*, 2020.
8. Raaz Dwivedi*, **Yan Shuo Tan***, Briton Park, Mian Wei, Kevin Horgan, David Madigan, and Bin Yu, "Stable Discovery of Interpretable Subgroups via Calibration in Causal Studies", *International Statistical Review*, vol. 88, pp. S135–S178, 2020.
9. **Yan Shuo Tan** and Roman Vershynin, "Phase Retrieval via Randomized Kaczmarz: Theoretical Guarantees", *Information and Inference: A Journal of the IMA*, vol. 8, no. 1, pp. 97–123, 2019.
10. **Yan Shuo Tan**, "Energy Optimization for Distributions on the Sphere and Improvement to the Welch Bounds", *Electronic Communications in Probability*, 2017.

Conference Publications

11. Ruizhe Deng†, Bibhas Chakraborty*, Ran Chen*, and **Yan Shuo Tan***‡, "BFTS: Thompson Sampling with Bayesian Additive Regression Trees", *International Conference on Machine Learning*, 2026 [**Spotlight**].
12. Jean Feng, Avni Kothari, Lucas Zier, Chandan Singh, and **Yan Shuo Tan**, "Bayesian Concept Bottleneck Models with LLM Priors", *Advances in Neural Information Processing Systems*, 2025.
13. Abhineet Agarwal*, **Yan Shuo Tan***, Omer Ronen, Chandan Singh, and Bin Yu, "Hierarchical Shrinkage: Improving the Accuracy and Interpretability of Tree-Based Models", *International Conference on Machine Learning*, pp. 111–135, 2022 [**Oral**].

14. **Yan Shuo Tan**, Abhineet Agarwal, and Bin Yu, "A Cautionary Tale on Fitting Decision Trees to Data from Additive Models: Generalization Lower Bounds", *International Conference on Artificial Intelligence and Statistics*, pp. 9663–9685, 2022.

15. **Yan Shuo Tan** and Roman Vershynin, "Polynomial Time and Sample Complexity for Non-Gaussian Component Analysis: Spectral Methods", *Conference on Learning Theory*, pp. 498–534, 2018.

Preprints

16. Chandan Singh, **Yan Shuo Tan**, Weijia Xu, Zelalem Gero, Weiwei Yang, Michel Galley, and Jianfeng Gao, "Agentic-imodels: Evolving Agentic Interpretability Tools via Autoresearch", arXiv preprint, 2026.

17. **Yan Shuo Tan**, Kenyon Ng, Ruizhe Deng†, Sumetha Loganathan†, Qiong Zhang, and Bibhas Chakraborty, "PFN-TS: Thompson Sampling for Contextual Bandits via Prior-Data Fitted Networks", arXiv preprint, 2026.

18. Zineng Xu†, Subhroshekhar Ghosh*, and **Yan Shuo Tan***, "On the Statistical Optimality of Optimal Decision Trees", arXiv preprint, 2026.

19. Tianqi Zhao, Guanyang Wang, **Yan Shuo Tan**, and Qiong Zhang, "TabClustPFN: A Prior-Fitted Network for Tabular Data Clustering", arXiv preprint, 2026.

20. Jean Feng, Avni Kothari, Patrick Vossler, Andrew Bishara, Lucas Zier, Newton Addo, Aaron Kornblith, **Yan Shuo Tan**, and Chandan Singh, "Human-AI Co-design for Clinical Prediction Models", major revision at *Nature Digital Medicine*.

21. Ruinan Jin, Gexin Huang, Xinwei Shen, Qiong Zhang, **Yan Shuo Tan**, and Xiaoxiao Li, "See-in-Pairs: Reference Image-Guided Comparative Vision-Language Models for Medical Diagnosis", arXiv preprint, 2025.

22. Qiong Zhang*, **Yan Shuo Tan***, Qinglong Tian*, and Pengfei Li, "TabPFN: One Model to Rule Them All?", minor revision at *Journal of the American Statistical Association*.

23. Xin Chen*, Jason M Klusowski*, **Yan Shuo Tan***, and Chang Yu*, "Revisiting Randomization in Greedy Model Search", arXiv preprint, 2025.

24. **Yan Shuo Tan**, Omer Ronen, Theo Saarinen, and Bin Yu, "On the Computational Efficiency of Bayesian Additive Regression Trees: An Asymptotic Analysis", in revision at *Annals of Statistics*.

25. Abhineet Agarwal*, Ana M Kenney*, **Yan Shuo Tan***, Tiffany Tang*, and Bin Yu, "Integrating Random Forests and Generalized Linear Models for Improved Accuracy and Interpretability", in revision at *Journal of Machine Learning Research*.

26. Xin Chen*, Jason M Klusowski*, and **Yan Shuo Tan***, "Error Reduction from Stacked Regressions", arXiv preprint, 2023.

27. Omer Ronen*, Theo Saarinen*, **Yan Shuo Tan***, James Duncan, and Bin Yu, "A Mixing Time Lower Bound for a Simplified Version of BART", arXiv preprint, 2022.

28. **Yan Shuo Tan**, "Sparse Phase Retrieval via Sparse PCA Despite Model Misspecification: A Simplified and Extended Analysis", arXiv preprint, 2017.

Invited Talks

(1) Understanding the Statistical Gaps in Tree-Based Learning *2025–2026*
- Gouxionghui Seminar, Online *2025*
- MBZUAI Seminar, Abu Dhabi, UAE *2026*
- UniMelb Seminar, Melbourne, Australia *2026*

(2) Sparse Modeling Revisited: From Linear Models to LLMs *2025*
- AIVP, RIKEN-AIP & A*STAR Joint Workshop, Tokyo, Japan *2025*

- ICSDS, Seville, Spain *2025*

(3) TabPFN: One Model to Rule Them All? *2025*
- EcoSta, Tokyo, Japan *2025*

(4) Lessons on Mixing for Bayesian Additive Regression Trees *2025*
- BayesComp, Singapore *2025*
- Workshop on Efficient Sampling Algorithms for Complex Models, IMS, National University of Singapore, Singapore *2025*

(5) Statistical-Computational Trade-offs for Recursive Adaptive Partitioning Estimators *2024–2025*
- One World MINDS Seminar, Online [video] *2025*
- Renmin University of China International Forum on Statistics, Beijing, China *2025*
- INFORMS International Meeting, Singapore *2024*
- RIKEN-AIP Deep Learning Theory Team Seminar, Tokyo, Japan *2024*

(6) A Mixing Time Lower Bound for a Simplified Version of BART *2022–2024*
- Workshop on Optimization in the Big Data Era, IMS, National University of Singapore, Singapore *2024*
- BIRS Workshop on Harnessing the Power of Latent Structure Models and Modern Big Data Learning, Hangzhou, China *2023*
- IMS Asia-Pacific Rim Meeting, Melbourne, Australia *2023*
- Workshop on the Mathematics of Data, IMS, National University of Singapore, Singapore *2022*

(7) MDI+: A Flexible Random Forest-Based Feature Importance Framework *2023*
- Young Mathematical Scientists Forum, IMS, National University of Singapore, Singapore *2023*

(8) Understanding and Overcoming the Statistical Limitations of Decision Trees *2022–2023*
- Neyman Seminar, Department of Statistics, University of California, Berkeley, USA *2023*
- Department Seminar, Department of Statistics and Data Science, National University of Singapore, Singapore *2022*
- Workshop on Machine Learning and its Applications, IMS, National University of Singapore, Singapore *2022*
- Workshop on Information Theory and Data Science, IMS, National University of Singapore, Singapore *2022*
- ESD Research Seminar, Singapore University of Technology and Design, Singapore *2022*

(9) Stable Discovery of Interpretable Subgroups via Calibration in Causal Studies *2020–2022*
- Bernoulli-IMS Young Researchers Pre-meeting, Seoul, South Korea *2022*
- ICSA Canada Chapter Symposium, Banff, Canada *2022*
- Causal Inference Research Group Seminar, University of California, Berkeley, USA *2021*
- Biostatistics Seminar, University of California, Berkeley, USA *2020*

(10) Implicit Regularization for Constant Step-Size SGD: Why It Works for Non-Smooth, Non-Convex, Low Rank Models *2019*
- Foundations of Data Science Program Reunion Workshop, Simons Institute, Berkeley, USA *2019*
- AMS Spring Central and Western Joint Sectional Meeting, Special Session on Sparsity, Randomness, and Optimization, Honolulu, USA *2019*
- ACO Seminar, University of California, Irvine, USA *2019*

(11) Efficient Algorithms for Phase Retrieval in High Dimensions *2017–2018*

- BLISS Seminar, University of California, Berkeley, USA *2018*
- Math-FLDS Seminar, University of Southern California, USA *2018*
- Probability Seminar, University of California, Irvine, USA *2017*
- ARC-TRIAD Seminar, Georgia Institute of Technology, USA *2017*
- SILO Seminar, University of Wisconsin, USA *2017*

Poster Presentations

(1) Revisiting Randomization in Greedy Model Search *2025*
- JSM, Nashville, USA *2025*

(2) Bayesian Concept Bottleneck Models with LLM Priors *2025*
- XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge, Singapore *2025*

Teaching Experience

National University of Singapore *2023–*

- **ST5209/X**: Analysis of Time Series Data (Semester II 2023, Semester II 2024, Semester II 2025, Semester II 2026)
- **ST5201X**: Statistical Foundations of Data Science (Semester I 2022)

University of California, Berkeley *2021–*

- **Data 102**: Data, Inference, and Decisions (Spring 2021)
  - Capstone course for the data science major. Develops the probabilistic foundations of inference in data science and builds a comprehensive view of the modeling and decision-making life cycle. Topics include: frequentist and Bayesian decision-making, false discovery rate, causal inference, general linear modeling, non-parametric modeling, bandits, and reinforcement learning.
- **Stat 210B**: Theoretical Statistics II (Spring 2020)
  - Second course on theoretical statistics for PhD students, focusing on high-dimensional statistics.

University of Michigan *2015–*

- Precalculus (Fall 2015)
  - Section of 20–25 students; responsible for lecturing, group work, materials, and grading.
- Calculus I (Fall 2013, Fall 2014, Winter 2016)
  - Section of 20–25 students; responsible for lecturing, group work, materials, and grading.
- Calculus II (Winter 2014, Winter 2015)
  - Section of 20–25 students; responsible for lecturing, group work, materials, and grading.
- Multivariate Calculus (Spring 2014)
  - Led lab tutorial sessions and graded.
- Differential Equations (Winter 2018)
  - Led lab tutorial sessions and graded.
- Precalculus (Course Co-coordinator) (Fall 2017)
  - Co-coordinated a course of ˜500 students. Trained and supervised new instructors; wrote homework and exams.

Academic Services

Students Advised

- Duan Shuo (BS) *2026–2028 (expected)*
- Zhao Yaya (PhD), *visiting from RUC* *2025–2026 (expected)*
- Le Thi Hong Ha (PhD) *2025–2029 (expected)*
- Martin Eppert (PhD), *co-advised with Subhroshekhar Ghosh* *2025–2029 (expected)*
- Sumetha Loganathan (PhD), *co-advised with Bibhas Chakraborty* *2025–2027 (expected)*
- Cai Yuchao (Postdoc) *2024–2026 (expected)*
- Deng Ruizhe (PhD), *co-advised with Bibhas Chakraborty* *2023–2027 (expected)*
- Zhuang Tianyi (PhD) *2023–2027 (expected)*
- Xu Zineng (PhD) *2023–2026 (expected)*
- Zeng Zitong (MSc) *2023–2024*
- Wang Yihe (BS) *2025–2025*
- Song Yida (BS) *2023–2024*

Thesis Advisory Committee

- Wenshan Qu (National University of Singapore) *2024–2027 (expected)*
- Nicolas Alexander Ihlo (University of Regensburg) *2025–2027 (expected)*

Other Mentoring

- BAIR Undergraduate Mentoring Program, University of California, Berkeley *2020–2020*
- Mentoring junior PhD students in Yu Group, University of California, Berkeley *2019–2022*

Reviewing Activities

- Annals of Applied Statistics *2020*
- Annals of Statistics *2025*
- Bernoulli *2025*
- Biometrika *2024*
- Conference on Learning Theory *2018, 2019*
- Conference on Neural Information Processing Systems *2023, 2025*
- Electronic Journal of Statistics *2024*
- Foundations of Computer Science *2018*
- IEEE Transactions on Information Theory *2023, 2026*
- IMA Journal of Numerical Analysis *2018*
- Information and Inference: A Journal of the IMA *2019*
- International Conference on Artificial Intelligence and Statistics *2021, 2024*
- Journal of Data Science *2024*
- Journal of Machine Learning Research *2025, 2026*
- Journal of the American Statistical Association *2024, 2025*
- Journal of the Royal Statistical Society: Series B *2022, 2024*
- Proceedings of the National Academy of Sciences *2020*
- Scandinavian Journal of Statistics *2026*
- Science China *2025*
- SMAI Journal of Computational Mathematics *2017*
- Statistical Methods in Medical Research *2025, 2026*
- Symposium on Discrete Algorithms *2019, 2020*

Other Service

- Graduate Research Committee Member, Department of Statistics and Data Science, National University of Singapore *2024–2026*
- Co-organizer, FIM-IMS Joint Workshop on Mathematics for Data Science, National University of Singapore *2026–2026*
- Co-organizer, IMS New Researchers Conference Asia, Institute of Mathematical Statistics *2026–2026*
- Young Researcher Committee Member, Bernoulli Society *2024–2026*
- Group Meeting Organizer, Yu Group, University of California, Berkeley *2020–2020*
- Student Probability Seminar Organizer, University of Michigan *2016–2017*
- Student Geometry Seminar Organizer, University of Michigan *2014–2015*