张豪哲 HAOZHE ZHANG

www.linkedin.com/in/haozhe-zhang-data
haozhestat.github.io
haozhe.zhang@outlook.com
+1-(425) 503-6250
Seattle, Washington

EDUCATION

Iowa State University, Ph.D., Statistics, 2014 - 2019
Dissertation: Topics in functional data analysis and machine learning predictive inference
2nd Place and 5th Place at Data Mining Cup 2016
11 publications on reputable scholarly journals including the top-ranking J. Am. Stat. Asso.
Presidential Scholars Fellowship, George W. Snedecor Award, Holly and Beth Fryer Award

University of Science and Technology of China, B.S., Statistics, 2010-2014
Hua Loo-Keng Talent Program in Mathematics, School of the Gifted Young
China National Scholarship, Outstanding Student Scholarship, National Encouragement Scholarship

CAREER SUMMARY

Data & Applied Scientist II, Microsoft AI Platform 05/2019 – PRESENT

- Lead developer and project owner of the open-sourced Python library "*shrike*" that supports running compliant ML pipelines on commercial and consumer data in Azure.
- Contributed to the propotype, benchmark, and productionization of the Python library "*flaml*" - a cost-effective hyperparameter optimization and learner selection method invented by *Microsoft Research*.
- Built and productionized end-to-end ML pipelines for *smart compose* and *email triage* projects.
- Improved the core algorithm of Azure automl by driving and accomplishing three sub-projects (*meta-learning, smart ensemble*, and *data privacy*).
- Conducted applied research on lightweight & fast automl, confidential ML, GPT-3, and privacy-preserving cross-silo federated learning.
- Researched and implemented a computationally efficient method to construct confidential intervals for evaluation metrics of machine learning pipelines.

Research Intern, eBay 05/2018 – 08/2018

- Developed a semi-supervised recommendation algorithm, utilizing label propagation, for the look-alike system of eBay first-party ads business.
- Constructed a unified large-scale user graph, scalable to *50+* million active eBay sellers, with interactive graph visualization based on *igraph* and *d3js*.
- Released internally data-driven user segmentation results by performing community detection on the constructed user graph.
- Designed and conducted back-testing and online A/B testing experimentations on the proposed graph-based recommendation algorithm on the eBay marketplace platform, for comparison with other state-of-the-art benchmarks, e.g., SVD-based collaborative filtering and hybrid methods.

Data Scientist Intern, Liberty Mutual 06/2017 – 08/2017

- Trained and validated deep neural network architectures (e.g., *CNN, LSTM, Res-Net, U-net*) on *100+* GB vehicle telematics data and driver accident records using *Tensorflow* framework in AWS.
- Compared the predictive performance of the deep learning models with classic machine learning methods (e.g., *xgboost, lightGBM*), and explored ensemble opportunities.
- Tested and improved the reliability of the internal AWS-based machine learning platform.

Research Assistant, Iowa State University 05/2015 – 05/2019

- Proposed a new method to construct *random forest* prediction intervals with theoretical guarantees and developed a *regression-enhanced random forest* method to address out-of-distribution challenge.
- Studied and modeled noisy phenotypic data derived from *crowdsourced* images annotated by *Amazon Mechanical Turk* workers.
- Analyzed the Zillow real estate data and London housing price data by sparse functional modeling.
- Built and maintained a web-crawling platform to automatically download real-time weather and air-pollution data for research projects.
- Participated in 5+ interdisciplinary consulting analytic projects in economics, finance, and sociology.

Research Intern, Okinawa Institute of Science and Technology Winter & Summer 2013

- Performed automatic feature recognition on Google satellite images.
- Implemented cluster analysis on image data using the nonparametric mean-shift algorithm.
- Analyzed the polygon class distribution results using Pearson's Chi-squared test.
- Worked on a mathematical model to explain geometric patterns of neuron bifurcations.

SELECTED PUBLICATIONS

- **Zhang, H.**, & Li, Y. (2021). Unified principal component analysis for sparse and dense functional data under spatial dependency. *Journal of Business & Economic Statistics.*
- Liang, D., **Zhang, H.,** Chang, X., & Huang, H. (2021). Modeling and regionalization of China's $PM_{2.5}$ using spatial-functional mixture models. *Journal of the American Statistical Association.*
- **Zhang, H.,** et al. (2020). Random forest prediction intervals. *The American Statistician.*
- **Zhang, H.,** Nettleton, D., & Zhu, Z. (2017). Regression-enhanced random forests. In *JSM Proceedings, Section on Statistical Learning and Data Science.*
- Liang, X., Zou, T., Guo, B., Li, S., Zhang, **H., Zhang**, S., Huang, H., & Chen, S. X. (2015). Assessing Beijing's $PM_{2.5}$ pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences.*

SKILLS AND QUALIFICATIONS

- **Programming:** Python, C, R
- **Database:** SQL, spark, MongoDB, Kusto
- **Tools:** docker, databricks, git, Jupyter, PowerBI, bash/pwsh, vim
- **Framework:** pytorch(-lightning), hugging face, lightgbm, catboost, pyspark, spacy, scikit-learn
- **Classic Machine Learning:** GBDT, bagging methods, lasso & ridge regression, elastic net, SVM, PCA, t-SNE, collaborative filtering, ensemble, embedding
- **Deep learning:** CNN, RNN, transformer
- **Optimization:** convex optimization, Bayesian optimization, linear programming, gradient descent
- **Math, Stat & Prob:** statistical inference (A/B testing), experimental design, time series analysis, Bayesian modeling, econometrics, measure theory, functional analysis, perturbation theory, concentration inequality
- **Special Topics:** automated machine learning, federated learning, differential privacy