# Leon Yuan

TEL: 469-251-3061 | Email: yuanli@smu.edu | GitHub: LiYuan199701

Bilingual: Mandarin and English

## EDUCATION

**Southern Methodist University** — **Dallas, Texas**

- ✧ ***Ph.D. in Statistical Science*** *(third-year, GPA: 3.82/4.0)* — *Aug.2022 – May 2026*
- ✧ **Awarded the Scheuren for outstanding performance on the PhD qualification theory exam** — *May 2023*

**Vanderbilt University** — **Nashville, Tennessee**

- ✧ ***M.S. in Data Science*** *(GPA: 3.89/4.0)* — *Aug.2020 – May 2022*

**Sichuan University (Rank #11 in China)** — **Chengdu, China**

- ✧ ***B.S. in Statistics*** *(Major GPA: 3.5/4.0)* — *Sep.2015 – June 2019*

**University of California, Berkeley.** — **Berkeley, California**

- ✧ ***Berkeley Global Access Program (Start - Discover)****(GPA: 3.6/4.0)* — *Aug.2019 – May 2020*
- ✧ ***Berkeley Summer Session (2020)(Numerical Analysis & Linguistic Data)*** — *Jun.2020 – Aug.2020*

**Yale University** *(GPA: 4.0/4.0)* — **Berkeley(online)**

- ✧ ***Yale Summer Session (2020): Multivariable Calculus & Intro Computing & Programming*** — *May 2020 – Jun.2020*

**Duke University** — **Nashville(online)**

- ✧ ***Duke Summer Session (2021): Bayesian and Modern Statistics (graduate-level)*** — *May 2021 – Jun.2021*

**University of Chicago** — **Nashville(online)**

*Data & Policy Summer Scholar Program Lectured and Advised by a Professor, Austin Wright* — *Aug.2021 – Sep.2021*

- ✧ ***Data Analytics in Public Policy & Introduction to Programming in R***
- ✧ ***Capstone Research Project: The Economic Impact of Retail Mask Policies***

**Stanford University** — **Nashville(online)**

- ✧ ***CS234: Reinforcement Learning*** — *Jan.2022 – Mar.2022*
- ✧ ***XCS224N, Natural Language Processing with Deep Learning*** — *Jun.2021 – Aug.2021*

## Professional Experience

**Data Scientist Internship at Amazon** — **Boston, MA**

*Supervised by Matt Tucker in the Amazon Alexa team* — *June 2023 – August 2023*

- ✧ Build a varied of time series models including, Holt-Winters model, dynamic regression with SARIMA error, Seasonal ARIMA, Neural Network AR model, Prophet, Neural Prophet, Ensemble modeling, Hierarchical Nested model, and Grouped Cross model for time series data
- ✧ Use R and Python to implement all above models under the same framework to compare and select the optimal model with time series cross validation criterion
- ✧ Use Bootstrap technique in time series to build my own ensemble modeling for forecasting and prediction interval

## Academic (Research) Experience

**Build and Merge two existing databases, protein structure databank and kinetics databank** — **Nashville**

***Published Paper:*** ***https://pubs.acs.org/doi/pdf/10.1021/acs.jpcb.1c05901***

*Advised by Prof. Zhongyue(John) Yang from Chemistry and RA. Bailu(Lucy) Yan from Vanderbilt* — *Sep.2020 – Sep.2021*

- ✧ Used PDB file from raw laboratory records to extract desired data fields, I used python with varied packages, such as pypdb, pandas, NumPy forming a user-defined class to realize extracting and converting into a data frame for further statistics analysis

- ✧ Used GraphQL-based Data API from websites connected to local python to form data stream which is not available in the PDB raw files
- ✧ Found the relations between PDB API and UniProt API and built the relationship database which serves as an enzyme computation platform

**Build Private Acquisition Agreements Database, Predict Probability of Female lawyers on deals** **Nashville**

*Part-time research assistant*

*advised by Prof. Tracey George from School of Law at Vandy, Prof. Albert Yoon, and Prof. Mitu Gulati* *Mar.2021 – Sep.2021*

- ✧ Used *kutools* on Window to batch convert 2470 .doc deals to .docx and *Adobe Acrobat Action Wizard* to batch convert 895 pdf to .docx with the auxiliary of *bash command line* to move, remove, create a directory, count file
- ✧ Used R and packages *readtext* and *tidyverse* to read in all 3365 .docx files and used R regular expression and non-greedy algorithms to extract text variables and values to populate the data table then built a relationship of database by encoding
- ✧ Imputed missing Gender and Regressed Gender on Law firm rank, Law School rank, Industry Sector, Deal Values by ordinal logistic regression and some other correlation analysis to decide influential factors helping Gender/Female on Deals

**Use transformers models from Huggingface to predict sentiment labels and cluster audio embeddings** **Nashville**

*Part-time research assistant advised by Senior Data Scientist, Charreau Bell, Ph.D.* *Aug.2021 – Dec.2021*

- ✧ Used different models like valhalla/distilbart-mnli-12-3 for zero-shot classifies four labels: reprimand, praise, neural, opportunity to respond, fine-tune models by using different labels to improve accuracy by observing confusion matrix
- ✧ Used Wav2Vec2Processor, Wav2Vec2Model to transcribe teacher audio files and used mean of last hidden state embedding pipelined with PCA, t-SNE and umap to reduce dimensions, then did K-means clusters and supervised cluster
- ✧ Added human ground true labels and human start timestamp and end timestamp to segment audio files for clusters models, Used Stratified Split technique to divide the whole data CSV into training, validation, test set with the same distribution

**Theory and Applications of Artificial Neural Networks** **Online in Nashville**

*Advised by Prof. Roman Kuc from Electrical Engineering at Yale* *May 2021 – Jul.2021*

- ✧ Initialized one new method for image data augmentation and initialized one new committee vote method to improve generalized accuracy and used Boxplot to illustrate reproducible generalization error
- ✧ Learned the basic concepts of Deep Neural Network and understand how it works, then write codes to tune and experiment with simulated digit audio data by ourselves using soundpy package
- ✧ Use our own recorded spoken digit data to train our speech recognization deep learning model by teams using Python and Keras library

**Probability, Bayesian Statistics Inference and Model Selection** **Shanghai, CHN**

**Team Leader**

*Advised by Prof. Joseph Chang(Chair) from Statistics and Data Science at Yale University* *July 2019 – Aug.2019*

- ✧ Explored different Bayesian approaches to model selection, including AIC, BIC, and Laplace Approximation, derivated the Laplace Approximation method, and simulated its performance in R
- ✧ Used R programming to realize the comparison of different estimators, proposed new estimators for computing the marginal likelihood, and used MCMC/JAGS and importance sampling, with 10% precision improvement
- ✧ Analyzed hot hand phenomenon using JAGS/Gibbs packages

## SKILLS

***Computer Skills*: R**, **Python**, **Git/GitHub**, **SQL**, **LaTeX**, C/C++, Matlab, SAS, Markdown, JavaScript

R packages often used by me: **h2o, tidyverse, tidymodels, ggplot2, stringi, kableExtra, caret, randomForest, lubridate**

Python packages often used by me: **PyTorch, Tensorflow, transformers**, **NumPy, pandas, matplotlib, sklearn, keras**

Most projects are collaborated and shared in teams on **GitHub** with **Git** on local.