Liyun (Roy) Zeng

Mobile: (323) 286-5197 | **Email:** liyunzeng@gmail.com | **LinkedIn :** https://www.linkedin.com/in/liyunzeng/

OBJECTIVE

- To obtain the Data Scientist or Machine Learning Engineer position on machine learning and Stat modeling applications.

EDUCATION

- **Ph.D.** candidate in Statistics, University of Arizona, Tucson, AZ. 2017- 2022 (Expected Graduation: *May 2022*). GPA: **4.0**
- **M.S.** in Computer Science, University of Arizona, Tucson, AZ. May 2021. GPA: **4.0**
- Ph.D. in Molecular and Cellular Biology (*Bioinformatics*), University of Washington, Seattle, WA. August 2010. GPA: 3.73
- B.S. in Biological Science, Sichuan University, Chengdu, Sichuan, China. July 2005. GPA: 3.79

RELEVANT COURSEWORK

Calculus, Linear Algebra, Real Analysis, Discrete Structures, Data Structures and Algorithms (**Java**), Theory of Computation, Operating Systems (**C**), Computer Networking (**C++**), Design and Analysis of Algorithms, Database Systems (**C++**), Statistical Machine Learning and Data Mining (**R/Python**), Machine Learning Theory, Online Learning and Multi-armed Bandit (**Python**), Deep Learning (**Python**), Computer Vision (**MATLAB**), Information Retrieval (**Java**), Statistical Natural Language Processing (**Python**), Fundamentals of Optimization, Analysis of High Dimensional Data (**R**), Statistical Computing (**R**), Theory of Probability and Statistics, Design of Experiment (**SAS**), Adv Regression Analysis (**R**), Bayesian Statistical Theory (**R**), Stochastic Modeling (**R**)

SKILLS SET

• **MACHINE LEARNING:** Regression, Classification, Predictive Modeling, Feature Engineering and Selection, GLM, LASSO, SVM, Kmeans, Decision Tree, Random Forest, XGBoost, Recommendation Systems, Contextual Bandit, NLP, Information Retrieval, Text Mining, Computer Vision, Anomaly Detection, Network Association Analysis, Deep Learning, GAN, VAE

• **STATISTICS:** Probability and Statistical Inference Theory, Bayesian Statistics, Probabilistic Programming, Probabilistic Graphical Model, Stochastic Modeling, Machine Learning Theory, High Dimensional Data Analysis, Linear and Non-linear Programming, Convex Optimization, Design of Experiment, Survival Analysis, Statistical Computing, Biostatistics, Bioinformatics

• **PROGRAMMING:** Python, R, SQL, MATLAB, SAS, Julia, Go, Java, C++, Linux, Git, PyTorch, Pyro, Tensorflow, Keras

• **BIG DATA/CLOUD:** AWS, Amazon SageMaker, Domino Lab, GitLab, PySpark, Hadoop

PROFESSIONAL EXPERIENCE

Sanofi, Bridgewater, NJ May 2021 to Jul 2021

Data Scientist Intern, Applied machine learning for pre-clinical vaccine development RNA-seq data analysis and predictive modeling

- Established the RNA-seq analytic pipeline from pre-clinical data for downstream statistical analysis and ML modeling.
- Innovated RNA-seq data deconvolution to estimate cell type abundance and cell type-specific differential gene expression.
- Applied genetic association analysis with GWAS for RNA-seq deconvoluted DE genes at a 15% increase of targets accuracy.

Amazon, Seattle, WA May 2020 to Aug 2020

Applied Scientist Intern, Machine Learning Algorithm for efficient Non-Prime Amazon customers purchase behavior prediction

- Achieved feature engineering data aggregation using **SQL**, **Spark** and **Python** in **AWS** for 52M*257 super high-dim data.
- Applied regression analysis for repurchase frequency and monetary prediction in fixed time for 81% AUC in **SageMaker**.
- Innovated survival analysis of repurchase event and interval prediction in flexible time with comparable performance.
- Established business driver analysis with 23 new features as significant drivers and improve ~5% model performance.
- Designed business metric analysis for efficient customer targeting by varying threshold with confusion matrix evaluation.

Carvana, Phoenix, AZ May 2019 to Aug 2019

Predictive Modeling Intern, Machine Learning Algorithm for Carvana sales and marketing strategy by efficient predictive modeling

- Achieved experimental design, data aggregation, data munging and predictive modeling using **T-SQL, R** and **Python**.
- Applied *logistic regression* to predict the best calling strategy for the maximum customer contact rate with ~60% AUC.
- Optimized high-dimension (84*1750) market data for efficient feature selection and sales prediction by *adaptive LASSO.*

Bank of China, Chengdu, China Jul 2016 to Aug 2017

Data Analyst Intern, Data analysis and statistical modeling in risk assessment and business development for key account support

- Achieved risk management and prediction modeling by data mining and management using **Excel**, **R**, **Python** and **SQL**.
- Applied machine learning using *logistic regressions*, *regularization*, *random forests*, *boosting* to reduce 15% churn rate.
- Optimized data analysis and visualizations for monthly sales and revenue reports for better regional key account support.

MaxCyte, Los Angeles, CA Nov 2015 to Jun 2016

West Coast Field Application Scientist, commercial support for west coast customers using MaxCyte flow transfection system

- Supported MaxCyte flow transfection system with 25% + contribution to a sales revenue increase of west coast business.
- Acquired sales data analysis and visualization using **Excel** and **Salesforce** to present dashboards to senior management.

Thermo Fisher Scientific, Shanghai, China/Rochester, NY Jan 2014 to Jul 2015

APAC Field Application Scientist, commercial support for Asia-Pacific customers using Thermo Fisher Bioproduction products

- Achieved APAC bioproduction VaB products with 15% more contribution to a total revenue increase of VaB business.
- Optimized sales, revenue and customer feedback data for marketing campaigns to drive APAC commercial performance.
- Advanced data visualization and monthly APAC revenue report using **Excel** and **Salesforce** for BD senior management.

GenScript, Nanjing, China/Piscataway, NJ May 2012 to Dec 2013

Senior Scientist, Bioproduction of antibodies and recombinant proteins, bioinformatics and statistical modeling of DoE optimization

- Optimized large sequence database by bioinformatic machine learning pipeline to increase protein expression >100 folds.
- Directed bioprocess team of 9 subordinates applying mixed effect model for DoE increase > 50% purification efficiency.

University of Southern California, Zilkha Neurogenetic Institute, Los Angeles, CA Oct 2010 to Apr 2012

Postdoctoral Research Associate, Bioinformatics and statistical analysis of Autism genetics using stem cell model and NGS technology

- Advanced bioinformatics analysis of RNA-seq data by **Python** via **Spark/Hadoop** identity *NRXN1* as key gene in autism.
- Optimized large-scale DNA-seq and ChIP-seq data for gene network identified > 100 key genes of patient-specific iPSCs.

ACADEMIC DESIGN PROJECTS

C/C++ Developer

Log-Structured File System (LFS) Implementation Spring 2020

- Designed the LFS with log, file and directory layers and integrated with *fuse* library with full function for vim applications.
- Implemented data consistency with checkpoint and crash recovery with log cleaner and support basic Linux commands.

Minibase Database Management System Implementation Fall 2018

- Designed the Minibase relational database management system (DBMS) using C/C++ programming language.
- Implemented the heapfile, buffer management, B+ tree, concurrency control and recovery for efficient DBMS function.

Java Developer

hack4equality Hackathon in Los Angeles (Work with Oliver Li, SDE@Meredith) Sep 2016

- Supported "My safe zone" project to help the minority resettled in a new country to find safe places or direct services.
- Data mining on open data resources such as data.lacity.org, geohub.lacity.org, data.weho.org, and ArcGis for modeling.

Computational Modeling of Par Protein Network Sep 2006

- Established mathematical models of the temporal and spatial dynamic of Par protein dynamics by C/Java and ODEs/PDEs.
- Visualized the spatial and temporal instability and traveling wave of par protein dynamics by implementing Java API.

PUBLICATIONS

- Zeng et al., Sparse learning and probability estimation with support vector machines. Journal of Computational and Graphical Statistics. 2022. In Preparation.
- Zeng et al., Scalable multiclass probability estimation with support vector machines. Journal of Data Science. 2021. Submitting.

- Zeng et al., Genomic and transcriptomic analysis reveals an oncogenic functional module in meningioma. Neurosurgical Focus. 2013;35(6): E3.
- Zeng et al., Functional impacts of NRXN1 knockdown on neurodevelopment in stem cell models. PLoS One. 2013: e59685.
- Zeng et al., DNA methylation in the malignant transformation of meningiomas. PLoS One. 2013: 8(1): e54114.
- Zeng L. Understanding Bcl-x_L and Mcl-1 GOF inhibition mechanism to mitochondrial metabolic changes. ProQuest. 2010.
- Zeng et al., Bcl-2 family proteins as regulators of oxidative stress. Semin. Cancer. Biol. 2009;19: 42-9.
- Zeng et al., Coloniality has evolved once in Stolidobranch Ascidians. Integrative and Comparative Biology. 2006; 46:255-268.
- Zeng et al., Molecular Phylogeny of Protochordate: Chordate Evolution. Canadian Journal of Zoology. 2005; 83:24-33
  Cited in Wikipedia [20]: https://en.wikipedia.org/wiki/Tunicate
- Zeng L. Apply Binary Notation Scale to Games. High-School Mathematics. 2001; 2: 12-16.

PRESENTATIONS

- Zeng L. The application of Quality by Design (QbD) with statistical modeling and optimization of DoE design in human vaccine and biopharmaceutical industry. China Biopharma Summit, Chengdu, China, 2014; China National Biotech Group Company Annual Meeting, Nanjing, China, 2014. Taiwan Biopharma Conference, National Taiwan University, Taipei, Taiwan, 2015.
- Zeng L, Zhang P, Lu W, Wang K. Statistical Modeling of NRXN1 functional significance in neurodevelopment using human embryonic stem cells (hESCs) and patient-specific induced pluripotent stem cells (hiPSCs). The American Society of Human Genetics 61th Annual Meeting and 12th International Congress of Human Genetics, Montreal, Canada, October 11-15, 2011.
- Zeng L, Mukhopadhyay B, Dawes AT. Mathematical and Statistical modeling of PAR protein dynamics. The First annual q-bio Conference on Cellular Information Processing, Los Alamos, NM, August 9-10, 2007.
- Zeng L, Swalla BJ. Statistical modeling and phylogenetic analysis for the evolution of coloniality in the Tunicates. Society for Integrative and Comparative Biology Annual Meeting, Orlando, FL, January 4-8, 2006.

HONORS AND AWARDS

- National Science Foundation (NSF) Research Training Group (RTG) Traineeship Grant in Data Driven Discovery, 2020-2022.
- APAC Customer Support Excellence Award, Thermo Fisher Scientific Lab Product Group, 2014-2015.
- Outstanding Contribution Award, GenScript Bioprocess Department, 2013.
- Travel Fellowship, Society for Integrative and Comparative Biology, 2006.
- Drs. Benjamin and Margaret Hall Foundation Fellowship in Biology, University of Washington, 2005-2006.
- UW-SCU International Scholarship, University of Washington, 2003-2004.
- Third-Class Bachelor Thesis for Academic Excellence, Sichuan University, 2005.
- First-Class Scholarship for Academic Excellence, Sichuan University, 2002, 2003, 2005.
- First prize in Mathematical Modeling Competition, Sichuan University, 2003.