DAVID LIM

limddavid@gmail.com | 818-939-4235

EDUCATION

University of North Carolina, Chapel Hill *Class of 2022*
PhD. in Biostatistics

- My dissertation research consisted of: (1) unsupervised clustering methods for cancer genomics data, and (2) unsupervised and supervised deep learning architectures to properly handle missingness in high-dimensional data.
- Research conducted under the supervision of Dr. Naim Rashid and Dr. Joseph Ibrahim.

University of California, Los Angeles *Class of 2013*
B.S. in Applied Mathematics; B.S. in Physics

PROFESSIONAL EXPERIENCE

GlaxoSmithKline | Collegeville, PA *June 2022 - Present*
Principal Statistician, Research Statistics

- Fit random forests, XGBoost, and similar models to predict dependencies of targets' fitness on multi-omic features.
- Spatial analysis of data from imaging mass spectrometry to identify patterns of spread of administered drugs.
- Utilized PCA, tSNE, and extremely high-dim data to assess probabilities of success in the clinic.
- Employed rigorous statistical modeling to identify significant effects of compounds of interest.
- Power analyses using simulations to assess required sample size for sufficiently powered experiments.

Syngenta AG | Durham, NC *November 2016 - February 2018*
Data Analyst

- Used random forests, boosting, and penalized regression to build prediction models using high-dim data.
- Performed cleaning of large databases, keeping information regarding thousands of crops up-to-date.

UNC Chapel Hill | Chapel Hill, NC *July 2014 - June 2018*
Research Assistant

- Regression analysis on MRI pixel intensities over time to determine the rate of release of sputum from the lungs.
- Microbiome analysis of the sputum of COPD patients to determine differential abundance of bacterial strains.

ORAL PRESENTATIONS

- "Handling Missing Electronic Health Records Data Using Importance-Weighted Autoencoders," ENAR 2021 (Baltimore, MD)
- "Missing Data in Deep Learning," ENAR 2020 (Nashville, TN)
- "FSCseq: Simultaneous Feature Selection and Clustering of RNA-Seq Data," JSM 2019 (Denver, CO)
- "Unsupervised Clustering and Variable Selection for RNA-seq Data," ENAR 2018 (Atlanta, GA)

PUBLICATIONS

- D. K. Lim, N. U. Rashid, J. B. Oliva, and J. G. Ibrahim. "Deeply-Learned Generalized Linear Models with Missing Data." Journal of Computational and Graphical Statistics, *Accepted*, 2023. DOI: 10.1080/10618600.2023.2276122
- D. K. Lim, N. U. Rashid, J. B. Oliva, and J. G. Ibrahim. "Unsupervised Imputation of Non-ignorably Missing Data Using Importance-Weighted Autoencoders." Statistics in Biopharmaceutical Research, *Under Review*, 2023. Pre-print: https://arxiv.org/abs/2101.07357
- David K. Lim. Naim U. Rashid. Joseph G. Ibrahim. "Model-based feature selection and clustering of RNA-seq data for unsupervised subtype discovery." Ann. Appl. Stat. 15 (1) 481 - 508, March 2021. https://doi.org/10.1214/20-AOAS1407

ADDITIONAL INFORMATION

- **Computing**: R, SAS, Python, Pytorch, Tensorflow, C++, SQL, bash shell scripting, HPC clusters, AWS, GCE, Github
- **Languages**: English (Native), Korean (Fluent)
- **Memberships**: American Statistics Association (2016 - Present)
- **Awards**: Genomics and Cancer Training Grant (2018 - 2021)