JONGHYUN YUN

Data Scientist, PhD in Statistics

@ jonghyun.yun@gmail.com jyun.netlify.app github.com/jonghyun-yun in jongyun-yun

- **Summary** PhD in statistics with 10+ years working experience in industry and academia. Proficiency in advanced statistical modeling, (un)supervised learning, visualization, Bayesian inference, and big-data analytic tools including Python, R, Spark, SQL. Capable of developing innovative approaches, building and deploying a scalable ML system, overcoming granularity and scalability issues, and managing trainees and assistants.

FEATURED ONGOING PROJECTS

ML system to integrate fraud detection modules

- Design and deploy a large scale system (Spiderweb) that integrates ML prediction and SME's feedback to capture fraud rings. The system has increased the loyalty points recovery from compromised accounts by 94.7%.
- Design feedback loops to incorporate SME's knowledge into the system using R Shiny UI.
- Create and deploy Wiki pages using Hugo to provide better understanding of Spiderweb to stakeholders.

Loyalty fraud detection using anomaly detection technique

- Create 10+ hand-made features to follow SME's fraud investigating processes. Engineer train and test data sets and create the feature selection procedure based on the isolation forest. Perform the error analysis.
- Develop and deploy ML algorithms to detect and track fraudster's resources to facilitate the detection of account takeover before the account monetization.

Sequence of actions data (log data) analysis (code)

- Develop a novel ML approach to analyze timestamped log data leveraging NLP and survival models.
- Identify behavioral differences between groups of action sequences, and develop software packages.

Network dependence analysis using time to events (code)

- Develop Cox models equipped with latent space to discover complex patterns between connection time and outcome in bipartite network models.
- Apply the model to identify test-taker's proficiency using accuracy and response times in Duolingo.

EMPLOYEMENT HISTORY

Cybersecurity Data Scientist

American Airlines

02/2022 – Present Fort Worth, TX, USA

- Develop loyalty fraud detection models to capture compromised accounts at early stages of account takeover.
- Create and maintain databases of malicious resources used by fraudsters.
- Design and deploy a large scale ML system to create fraud incident reports and to leverage feedback from SME to reinforce the detection performance. The system has increased the fraud detection efficiency by 94.7%.

Data Scientist

Institute of Statistical Data Intelligence

09/2019 – Present Mansfield, TX, USA

- Develop ML methods for prediction, time series, causal inference, segmentation for big data. Apply NLP and survival models to analyze timestamped log data. Processing, cleansing and validating the integrity of data.
- Develop novel graph learning to discover dynamic interaction b/w customers and items. Parallel programming for complex Bayesian inference. Present analysis and visualization, and developing software packages.

Assistant Professor of Statistics

Department of Mathematics, University of Texas at Arlington

09/2016 – 08/2019 | Arlington, TX, USA

- Responsible for bringing innovative machine learning approaches to studies broadly related to statistics, engineering, business, and biomedical fields, and continuously growing and sustaining research lab infrastructure.
- Designed data science courses including data mining and regression analysis. Created hands-on examples for R and Python programming. Mentored and trained junior scholars. Managed staff of teaching assistants.

Assistant Professor of Statistics

Department of Mathematical Sciences, University of Texas at El Paso

08/2015 – 06/2016 | El Paso, TX, USA

- Responsible for developing statistical methods in biomedical research, translating meaningful findings back to the community, supporting researchers in Border Biomedical Research Center.

Postdoctoral Researcher

Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center

09/2012 – 07/2015 | Dallas, TX, USA

- Developed innovative statistical methods to detect genomic markers using multiple sequencing data. Collaborated to apply the designed method to cancer research. Presented outcomes to all levels of audience.

EDUCATIONAL HISTORY

PhD in Statistics

Department of Statistics, University of Illinois at Urbana-Champaign

09/2006 – 08/2012 | Champaign, IL, USA

- Research in Monte Carlo methods for high-dimensional models with focus on solar weather prediction, target tracking and data assimilation. Dissertation on *Ensemble Filtering of State Space Models*. Advised by Yuguo Chen.

MA in Applied Statistics

Department of Applied Statistics, Yonsei University

03/2004 – 02/2006 | Seoul, South Korea

- Research in high-dimensional prediction models with applications in smart wearable and word frequency. Thesis on *Bandwidth Selection in Dimension Reduction Regression*. Advised by Hakbae Lee.

BA in Business Administration and Applied Statistics

Yonsei University

03/1997 – 02/2004 | Seoul, South Korea

- Related studies in economics, finance, operations research, marketing, and accounting. Minor in mathematics

STRENGTHS

General skills Project leadership, Interdisciplinary collaboration, Mentorship

Data science skills Advanced statistical modeling, ML, RL, NLP, Anomaly detection, Predictive modeling, Dimension reduction, Data visualization, Time Series, Hidden Markov model, Bayesian inference, Monte Carlo method, ML System design, Causal inference, Multiple hypothesis testing

Areas of experience Cybersecurity, Fraud detection, Biostatistics, Bioinformatics, Genomic data analysis, Smart infrastructure, Item response model, Network model, Sequence of actions analysis

Technical skills R, Python, C/C++, Spark, PyTorch, TensorFlow, SQL, MATLAB, Scala, Git, Parallel computing, Linux, Bash, Lisp, Hugo, Markdown, LaTeX

PUBLISHED INTELLECTUAL CONTRIBUTIONS

Refereed Journal Articles

1. Jin, I. H., Jeon, M., Schweinberger, M., Yun, J. & Lin, L. Multilevel Network Item Response Modelling for Discovering Differences between Innovation and Regular School Systems in Korea. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* (2022).
2. Yun, J., Ryu, K. R. & Ham, S. Spatial Analysis Leveraging Machine Learning and GIS of Socio-Geographic Factors Affecting Cost Overrun Occurrence in Roadway Projects. *Automation in Construction* **133,** 104007 (2022).
3. Yun, J., Kang, S., Tehrani, A. D. & Ham, S. Image Analysis and Functional Data Clustering for Random Shape Aggregate Models. *Mathematics* **8,** 1903 (2020).
4. Yun, J., Shin, M., Jin, I. H. & Liang, F. Stochastic Approximation Hamiltonian Monte Carlo. *Journal of Statistical Computation and Simulation* **90,** 3135–3156 (2020).
5. Nam, J. H., Yun, J., Jin, I. H. & Chung, D. hubViz: A Novel Tool for Hub-Centric Visualization. *Chemometrics and Intelligent Laboratory Systems* **203,** 104071 (2020).
6. Cai, L., Li, Q., Du, Y., Yun, J., Xie, Y., DeBerardinis, R. J. & Xiao, G. Genomic Regression Analysis of Coordinated Expression. *Nat Commun* **8,** 2187 (2017).
7. Yun, J., Yang, F. & Chen, Y. Augmented Particle Filters. *Journal of the American Statistical Association* **112,** 300–313 (2017).
8. Chen, B., Yun, J., Kim, M. S., Mendell, J. T. & Xie, Y. PIPE-CLIP: A Comprehensive Online Tool for CLIP-seq Data Analysis. *Genome Biol* **15,** R18 (2014).
9. Kwon, I., Xiang, S., Kato, M., Wu, L., Theodoropoulos, P., Wang, T., Kim, J., Yun, J., Xie, Y. & McKnight, S. L. Poly-Dipeptides Encoded by the C9orf72 Repeats Bind Nucleoli, Impede RNA Biogenesis, and Kill Cells. *Science* **345,** 1139–45 (2014).
10. Yun, J., Wang, T. & Xiao, G. Bayesian Hidden Markov Models to Identify RNA-Protein Interaction Sites in PAR-CLIP. *Biometrics* **70,** 430–440 (2014).

Non-Refereed Articles

1. Yun, J. & Chen, Y. Comments on "Particle Markov Chain Monte Carlo Methods" by C. Andrieu, A. Doucet, and R. Hollenstein. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **72,** 332–333 (2010).

Book Sections

1. Wang, T., Yun, J., Xie, Y. & Xiao, G. in *Hidden Markov Models* 177–184 (Humana Press, New York, NY, 2017).

Software

1. Yun, J. *Statistical Data Intelligence Tools for Cost-Overrun Analysis of Roadway Construction Projects* 2021. github.com/jonghyun-yun/dico.
2. Yun, J. *TEMPEST: Latent Space Competing Risk Model for Accuarcy and Reponse Time Data* https://github.com/Jonghyun-Yun/TEMPEST.
3. Yun, J. *Process Data Modeling for PIACC Data* 2021+. https://github.com/Jonghyun-Yun/proda.
4. Alvarez, H. & Yun, J. *Baseball Statistics Collecting Functions from HTML Tables* 2017. https://github.com/jonghyun-yun/brscrap.git.
5. Yun, J. *A MATLAB Toolbox to Identify RNA-protein Binding Sites in HITS-CLIP* 2013. https://qbrc.swmed.edu/labs/xiaoxie/download/README1.pdf.
6. Yun, J. *R Package for PAR-CLIP Analysis* 2013. https://qbrc.swmed.edu/labs/xiaoxie/download/README2.pdf.

Working Papers

1. Yun, J., Jin, I. H. & Jeon, M. Analysis of Connection Times in Bipartite Network Data: Development of the Latent Space Accumulator Model with Applications to Assessment Data. *Journal of the American Statistical Association* (2022+). To be submitted.
2. Yun, J., Ick Hoon, J. & Minjeong, J. Analysis of Time-Stamped Action Sequences (2022+).

PRESENTATIONS

Invited Talks

11/2021 "Latent Space Accumulator Model for Analyzing Bipartite Networks with Connection Times and Its Applications to Item Response Data", *Autumn annual conference of the Korean statistical society*, virtual.

02/2017 "Integrative modeling approaches for next-generation sequencing data", *Colloquim Series*, Texas A&M University-Commerce.

06/2016 "Model based identification of RNA-protein binding sites", Bioinformatics Session, *International Workshop on Applied Probability*, Toronto, ON, Canada.

10/2015 "Comparative analysis of CLIP-seq under multiple experimental conditions", *Border Biomedical Research Center Seminar*, *UT El Paso*, El Paso, TX, USA.

08/2014 "Statistical strategies for identification of the RNA-protein binding site in CLIP-seq", Biometrics Section, *2014 Joint Statistical Meetings*, Boston, NY, USA.

10/2014 "Statistical models to identify RNA-protein binding sites from CLIP experiments", *Computational and Systems Biology Seminar*, UT Southwestern, Dallas, TX, USA.

10/2011 "Augmented particle filters", *Robert Bohrer Student Workshop in Statistics*, University of Illinois at Urbana-Champaign, Champaign, IL, USA.

Poster Presentation

02/2014 "Identification for RNA-protein binding sites in CLIP-seq", *7th Annual Bayesian Biostatistics and Bioinformatics Conference*, Houston, TX, USA.

PROFESSIONAL AND UNIVERSITY SERVICE

Professional Service

06/2016 Co-chair, Bioinformatics session at *2016 International Workshop on Applied Probability* at Toronto, ON, Canada.

University Service (UTA)

09/2017 – 08/2019 Department advisory committee.
09/2016 – 08/2019 Math preliminary exam B subcommittees.
01/2017 – 05/2017 Undergraduate affairs committee.
01/2019 – 08/2019 College of Science Data science working group.
04/2018 Judge, College of Science Aces Research Symposium.

University Service (UTEP)

Spring 2016 Math Club Zero committee

Referee/Reviewer Work (Journals)

- *Journal of the American Statistical Association*, *Journal of Computational and Graphical Statistics*, *Computational and Mathematical Methods in Medicine*, *Journal of Statistical Software*, *Journal of Probability and Statistics*, *Bayesian Analysis*, *International Journal of Data Science*, *Genes*, *Mathematics*, *International Journal of Environment Research and Public Health*, *Antibiotics*, *Axioms*, *Healthcare*

TEACHING ACTIVITIES

University of Texas at Arlington

Spring 2019 MATH6312 - Data Mining (10 students)
Fall 2018 MATH3316 - Statistical Inference (57 students)
Spring 2018 MATH5358 - Regression Analysis (13 students)
Fall 2017 MATH5312 - Mathematical Statistics I (12 students)
Spring 2017 MATH5392 - Selected Topics in Mathematics (Data Mining) (12 students)
MATH5313 - Mathematical Statistics II (6 students)
Fall 2016 MATH5312 - Mathematical Statistics I (14 students)

University of Texas at El Paso

Spring 2016 STAT5474 - Introduction to Data Mining (14 students)
Fall 2015 STAT5354 - Post-genomic Analysis (5 students)
BINF5113 - Math Seminar for Bioinformatics (4 students)

University of Illinois at Urbana–Champaign

Spring 2012 STAT200 - Statistical Analysis (51 students)
Summer 2011 STAT100 - Statistics (30 students)
01/2010 – 05/2011 STAT400-Statistics and Probability I (Discussion Section Leader)

Spring 2010 (59 students), Fall 2010 (60 students), and Spring 2011 (93 students)

08/2006 – 12/2009 Teaching Assistant: STAT100-Statistics, STAT400-Statistics and Probability I, STAT410-Statistics and Probability II, STAT424-Analysis of Variance, STAT429-Time Series Analysis, STAT510- Mathematical Statistics I, and STAT511-Mathematical Statistics II.

Yonsei University

12/2005 Preliminary Calculus

03/2005 – 12/2005 Discussion Section Leader: STA2101-Calculus (65 students) and STA2102-Linear Algebra (67 students).

03/2004 – 12/2004 Teaching Assistant: STA1001-Introductory Statistics, STA1001-Introductory Statistics, STA3102-Multivariate Statistical Analysis, and BC682-Statistical Methods for Behavioral Sciences.

DIRECTED STUDENT LEARNING

Graduate Supervised Research

09/2017 – 09/2019 Anthony Thomas (*Statistics*, UT Arlington)

Project: *Bayesian hierarchical dynamic factor models*

09/2017 – 12/2017 Mario Garza (M.S. *Statistics*, UT Arlington)

Project: *Forecasting sales using a finite-state HMM: an inventory control exercise*

5 M.S. Student Committees

09/2016 – 08/2019 Daniel Sang Le, Nidhi Kiran Dawda, Zachary Loucks, Hongbo Yu

Statistics, UT Arlington

09/2015 – 08/2016 Tun-Lee Ng

Statistics, UT El Paso

6 Ph.D. Student Committees

09/2016 – 08/2019 Souad Sosa, Izzet Sozucok, Geoffrey Schuette, Yi Liu, Mahmoud Jawad, Piyachart Wiang-nak

Statistics, UT Arlington

Undergraduate Supervised Research

Spring 2018 Henry Alvarez (*Mathematics*, UT Arlington)

Project: *Developing a software package to collect baseball statistics*