Samuel J Thomas

Data Scientist
Address: 70 W Clear Lake Ln, Westfield IN 46074
samuel.thomas522@gmail.com +1 317 696 9214 linkedin.com/in/samueljthomas1/ | *Updated:* February 17, 2025

Education

Indiana University, Ph.D. Biostatistics May 2021

Purdue University, M.S. Mathematics

University of Notre Dame, B.S. Electrical Engineering

Society of Actuaries, Associate of the Society of Actuaries (ASA)

Employment

Data Scientist

General Services Administration Aug 2024–present

Lead Data Scientist on project assigning 13k GSA employees to their nearest possible duty station from President Trump's Return to Office Executive Order. This project combines data from Human Resources, Public Building Services, and geographical data from Arcgis. At the request of Human Resources, I am using R software to allocate personnel based on capacity and other employment characteristics such as job rank.

Developed a data model in Python and SQL to combine data from disparate data sources including Accounting Transactions, Federal Procurement data, and agency Budget data. The purpose of this data model is to extract insights into Information Technology costs and spend in the federal government. This model has been presented to the A-suite and approved for further development.

Developed and successfully pitched a proposed Datathon (similar to a hackathon for data) topic on estimating the financial risk of natural catastrophe hazards to the Public Building Services organization within GSA. This dataset combines publicly available data from the federal real properties data with FEMA's natural hazard risk model and financial measures per building, including revenue and expenses.

Led a project to automatically identify types of Google scripts being developed across the agency based on structured data and unstructured text data. This effort identified 12 unique types of Google scripts based on a customized NLP algorithm based on spectral clustering. Contributed this algorithm to the publicly available GovCXAnalyzer package.

Principal Data Scientist

Guidewire Software 2022–Aug 2024

Worked with internal actuaries to develop predictive modeling capabilities to help insurers properly reserve for bodily injury claims. This project involved developing an injury severity model as well as litigation likelihood models. I have developed several models using scikit-learn, xgboost, and catboost with customized data preprocessing.

Completed redesign and implementation of a risk model for cyber catastrophe events for the Cyence product. This model aligns risk of attack to a combination of historical experience, cyber expert judgment, and customer feedback. Improved the computational efficiency of this model from a runtime of 12 hours for 10k simulations to 40 minutes using coarse-grained parallelism and vectorized numpy calculations in AWS EMR.

Lead catastrophe modeler responsible for Mass Ransomware and Mass Data Breach model development. Use PyMC to develop a Bayesian Generalized Linear Mixed Effects model with random intercepts for industry categorization. PyMC uses Hamiltonian Monte Carlo, an advanced Bayesian MCMC technique, to fit regression models.

Using PyTorch to develop an autoencoder to impute firmographic and technographic features for companies we lack such data.

Lead modeler and prototype developer for Merger and Acquisition cyber-risk model intended for a leading financial services client. This model is used to quantify the cyber risk of merging two organizations. Adjustments in tail risk and average annual loss are based on the combined risk profile of the two companies. Provided direction to contractors on the report and chart design.

Designed and developed an original algorithm to quantify supply risk in a Bayesian network. Algorithm is implemented in python with a polynomial time complexity. This algorithm replaced the initial proposal with an exponential time complexity, which was too slow for a production application. Developed the original set of inputs or priors for this model based on cyber risk data, industry studies, original research, and associations of sources of revenue between companies (correlation matrix).

Led second-place team in Guidewire hackathon. Developed NLP model to mine cyber attack data for new ransomware gangs and categorizing types of attacks.

Senior Data Scientist

Capital Group Companies 2008, 2010–2022

Developed the forecasting methodology and implementation for the American Funds Service Company budget process. This process has been adapted and reproduced for multiple departments in the organization. I developed regression models with ARMA errors using Rob Hyndman's forecast package in R. These models related macroeconomic indicators to fund account volumes and were used in various scenario analysis. This forecast model was consistently accurate within 3% of account volumes.

Developed a custom forecast package in R called mlts (multi-level time series) to facilitate the annual forecast process. This package automatically combines hierarchies of low level forecast. This approach facilitated business adjustments on any of hundreds of forecasts for use in the budget process.

Developed machine learning and statistical models to redesign the North American Distribution sales territories. These models optimized the geographic size and location of territories for over 200 sales professionals across the United States. One direct impact of this analysis was the creation of a new sales territory region in the San Francisco area. Modeling techniques used included K-means clustering, 2-dimensional optimization, and a customized adaptation of genetic algorithms.

Developed an attribution model to inform the business regarding which types of events are most profitable to the organization. The model automatically estimates lift in sales and visits for many sponsored events in Azure Databricks. This model incorporates a deep learning model called an autoencoder for advisor matching, and a Bayesian generalized linear model for attribution. Long-term plans including evolving this model for recommendations.

Estimated impact of mutual fund placements on recommended lists using a generalized linear mixed effects model. This model is used to quantify $50B in sales opportunity, delineated at the mutual fund level.

Developed a sequence analysis using the TraMineR package to quantify the sales preferences of 40k financial advisors. This analysis demonstrated that the firm's sales in one product did not

cannibalize sales in another product. An initiative is planned to identify leads for the newer product, based on the results of this analysis.

Led a team of 3 consultants to build analytics for three new departments of the North American Client Group. These departments work with the home offices of financial services firms for product placement and sales initiatives.

Developed a simulation model in R to estimate the sales lift for a major coverage redesign initiative. This model is based on historical and projected market share, based largely on activity data reported through salesforce.com.

Extensive experience with using internal and external data to drive business decisions related to sales analytics. Worked with senior business leadership to build the foundations for reporting and analytics.

Fit a Generalized Linear Mixed Effects model (GLMM) to sales and activity data to identify the most significant drivers of work for Internal Wholesalers (IW). This model was used to assign IW coverage.

Principal

Revelant Technologies 2008–2010

Leveraged data to identify opportunities for quality and efficiency improvement opportunities for the 1-800-MEDICARE service center.

Associate Actuary

Milliman 2007–2008

Assistant Actuary

WellPoint/Anthem 2005–2007

Independent Consulting

From 2006

Clients: Medscape/WebMD, Healthcare Performance Consulting, and Arks32

Provided statistical support for Medscape/WebMD for study design and sample size recommendations

Developed machine learning model (random forests) for a research project on developing "soft" sensors in an environmental application

Performed multiple statistical analyses and maintained access data base for Healthcare Performance Consulting.

Designed a statistical sampling plan for Medscape to estimate treatment plan adoption among physicians.

Quality assurance assistance for Medscape on Tableau dashboard development.

Teaching

Co-instructor, Biostatistics

Indiana University 2019, 2024

PBHL-B 646 Advanced Generalized Linear Models

Adjunct Professor, Mathematics

Ivy Tech University 2006–2014

Courses taught:

- Math 211 Calculus
- Math 136 College Algebra
- Math 135 Finite Math

PUBLICATIONS

Thomas, S. and Tu, W. (2022). *Riemannian Monte Carlo*. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). doi:10.1002/9781118445112.stat08392

Thomas, S. and Tu, W. (2020). *Hamiltonian Monte Carlo*. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). doi:10.1002/9781118445112.stat08243

Thomas, S. & Tu, W. (2021) *Learning Hamiltonian Monte Carlo in R*, The American Statistician, DOI: 10.1080/00031305.2020.1865198

Green, Brice and Thomas, Samuel, *Inference and Prediction of Stock Returns using Multilevel Models* (August 31, 2019). Available at SSRN: https://ssrn.com/abstract=3411358 or http://dx.doi.org/10.2139/ssrn.3411358

In Progress

Thomas, S. and Tu, W. 2020. *An R package for Bayesian Multivariate Generalized Additive Models*.

Software Packages

hmclearn: An R package to fit statistical models with Hamiltonian Monte Carlo. https://cran.r-project.org/web/packages/hmclearn/index.html

bayesGAM: An R package to fit semiparametric regression models using Hamiltonian Monte Carlo. https://cran.r-project.org/web/packages/bayesGAM/index.html

mlts: An R package to automatically develop forecasts and perform cross-validation for bottoms-up forecast models. Internal package for Capital Group Companies.

Skills

- Statistical Computation
- Bayesian Analysis: Markov Chain Monte Carlo
- R, Python, C++, SQL, Azure Databricks, Tensorflow

Talks

A Bayesian Analytical Software Based on Hamiltonian Monte Carlo. Regenstrief Institute, 12/4/2019. https://www.youtube.com/watch?v=sBA3lAoNhto

Using Fourier Series to Model Daily Seasonal Patterns of Redemptions. Capital Group Companies, Data Science Interest Group, 2018

Improving Capacity and Financial Planning, a Guide to Business Forecasting with Alteryx. Inspire 2016 Alteryx Conference, San Diego, CA.

Predicting At-Risk Plans Using the C5 Algorithm. Capital Group Companies, Data Science Interest Group, 2015

UseR 2012 at Vanderbilt University. UseR 2012 Vanderbilt University.

Academic Experience

Co-instructor for PBHL-B 646 Advanced Generalized Linear Models with Wanzhu Tu, Spring 2019, Indiana University.

Developed a convolutional neural network model in Tensorflow to classify minerals based on spectral imaging from the planet Mars.

Additional Professional Experience

Served as lead analyst in developing statistical models for the AFSG budget process. Models are used to develop hiring plans to support a >1,500-person service center and by Global Finance for revenue projections. Accurate forecasting is critical to the financial health of the organization.

Developed a machine learning algorithm (based on C5) to identify retirement plans at risk for attrition. At-risk plans, as identified by the model, are twice as likely to leave as non-risk plans.

Developing R Shiny app to streamline the AFSG long-range forecasting process. Reduction of 4 FTE in analyst time.

Using spacy NLP library in Python to explore open text messages from salesforce data

Serving as Math/Theory contributor in Capital Group Data Science guild

Capital Group Internal Consulting: statistical support/consulting for Client Analytics, Business Management, Work Management, Legal/Compliance, Human Resources, Quality, Retirement Plan Services, and American Funds Service Company.

Developed a mathematical model to estimate the steady-state account volumes based on queuing theory ($M/M/\infty$ queue).

Used PyMC3 (Python) to fit a generalized linear model using MCMC techniques (Hamiltonian Monte Carlo). Model identified economic factors associated with Plan Premier Participant counts.

Identified contact rate influencers using Generalized Least Squares (GLS) models. Daily seasonal factors were fit using Fourier series. Discussion with Data Science Interest group contributed to increased leveraging of AFS data for analysis for broader Capital Group applications.

Developed a custom optimization model in Python to estimate the number of Shareholder Services associates needed if processing work was outsourced. Model influenced decision to retain processing work in-house.

Developed a regression model to match the automated pricing results from a website. The model was developed using statistical software and translated to Excel for the client's use. Developed a predictive model based on machine learning algorithms for an environmental sensor application with over 400 variables.

Created an automated forecasting model to predict fuel demand in various locations in the UK. This model was used to anticipate geographical fueling needs for a trucking company. Used text mining to analyze open-responses to survey questions from a call center. This analysis was used to identify drivers of caller satisfaction.

Evaluated search paths commonly used by CSRs to find scripts using sequencing analysis in R. This analysis was used to identify opportunities to improve script searching capabilities. Identified opportunities for quality and cost savings for the 1-800-MEDICARE call center through data mining and statistical methods.