tbanalytics99@gmail.com **Thomas J. Balmat** 919-820-1057

Recent Projects

High Performance Computing Cluster R Consultation

Research Computing, Duke University

The Duke Computing Cluster (DCC) is a high performance computing environment available to researchers at Duke for intensive computation. It consists of 1,360 compute nodes, more than 45,000 virtual CPUs, 980 GPUs, and 270TB of RAM.

Contributions

- Assist faculty, post-doc, and PhD student researchers in improving performance of R scripts in high data volume, high demand computational settings
- Adapt sequential algorithms to parallel by defining disjoint mathematical or data subsets
- Implement parallel algorithms using functions from the parallel package (mcmapply, parApply)
- Improve computational efficiency using parallel methods, Rcpp, and openMP
- Implement very high efficiency matrix operations on GPUs, by integrating custom Cuda algorithms with R
- Develop Slurm scripts for batch submission of R scripts
- Assess computational performance to specify optimal core and memory resource requests
- Compile and install R packages, by identifying necessary operating system resources
- Document common computing cluster R configuration procedures to be used by researchers
- Prepare and document solutions to computational problems as examples of computational efficiency and effective use of DCC resources (developed using R, Rcpp, Slurm, GPUs, emphasis on matrix operations)

Federal Office of Personnel Management Human Capital Project

Social Science Research Institute, Duke University

The Social Science Research Institute is home to fourteen centers for research in the social sciences. I am involved in the U.S. Federal Employee Human Capital Project, which is composed of faculty and researchers from diverse disciplines including Political Science, Economics, Sociology, and Statistics. The project conducts research into the policy value and effectiveness of the skill, experience, and education assets of non-military employees of the U.S. federal government. This involves intensive data exploration and management, development of novel statistical modeling methods, and development of high performance solutions to computationally complex models. Data sets include large, publicly available and private databases of federal employee career profiles and historical economic, demographic, and socio-political data.

Contributions

- Design and implement a SQL database to house and secure over 600,000,000 observations spanning more than forty years of historical data related to U.S. federal workforce human capital, the economy, and governmental policy sourced from Freedom of Information Act (FOIA) requests and sources such as FedScope, The Current Population Survey, The Centers for Disease Control, and customized data sets provided by researchers at affiliate institutions such as Emory University, Vanderbilt University, and UCLA
- Assist researchers with SQL query development and general data access, transformation, and efficiency improvement

- Research and develop high performance computational strategies and functions for fitting high dimension ($10^7 \times 10^5$ and beyond) fixed effects regression, quantile regression, logistic regression, and multinomial models using R, SQL integration, parallel processing, and C
- Assess the performance of research computing servers using high dimension regression and probability models, measure results across wide input parameter intervals, report findings, and recommend alternative configurations

Resources for the Human Capital project

- Research paper: Efficient Large Fixed Effects Regression Algorithm, including sections on efficient computation of robust, clustered standard errors for large problems (author)
  - Paper
  - Supplementary material
  - R package
  - Rcpp code
- Cholesky decomposition algorithms for solving large FE regression problems
  - Cholesky decomposition
  - $(\boldsymbol{X}'\boldsymbol{X})^{-1}$
  - R package
  - Rcpp code
- Research paper: Elections, Ideology, and Turnover in the U.S. Federal Government (contributor)
- Presentation: Efficient Solution of Large Fixed Effects Regression Problems (presented at useR!2017)
- Presentation: Approaches to Solving Large Computational Problems (presented at Joint Statistical Meetings 2017)
- Study: Research computing server performance evaluation

Human Capital Synthetic Data Project
Department of Statistical Science, Duke University

The Synthetic Data Project, led by faculty of the Department of Statistical Science, with participating faculty from the Fuqua Business School and Department of Computer Science, along with members from the Office of Information Technology at Duke and faculty from the Political Science Department at Emory University, develops and applies theory and methods for generating privacy protecting versions of the authentic federal human capital database that maintains career longitudinal and covariate relationships, for 3,500,000 employees over twenty four years, while limiting risk of revealing sensitive information on individual employees.

Contributions

- Design and implement SQL instances of six versions of synthetic federal employee human capital data
- Develop SQL queries to integrate authentic and synthetic data for data transformation and research model fitting
- Collaborate with human capital, data privacy, and statistics researchers to develop and fit models for assessing agreement of covariate relationships with those observed in authentic data, risk of private authentic information disclosure, and performance of synthesis probability models
- Analyze numerical and statistical privacy protecting algorithms for improvement of efficiency and compactness; test, certify, and report mathematical properties of resulting probability distributions
- Develop efficient statistical and graphical methods for fitting hundreds to thousands of (typically regression, probability density, covariance, Bayesian posterior, or Markov) models to various subsets of data for wide scale model evaluation
- Develop statistical classifier algorithms, logistic models, and Receiver Operating Characteristic methods to measure probabilities of revealing authentic individual information from generated synthetic data using training patterns derived from publicly available data sets

Resources for the Synthetic Data project

- Research paper: Providing Access to Confidential Research Data Through Synthesis and Verification
  - Paper (coauthor)
  - Synthesis models supplement (author)
  - Authentic/synthetic model and covariate distribution validation supplement (author)
  - Verification measure sensitivity supplement (author)
  - Verification measure regression variance sensitivity analysis (author)
  - Verification measure sensitivity simulation (author)
- Research paper: Empirical Evaluation of Privacy in Synthetic Data (coauthor, forthcoming)
- Research article: Improved Efficiency Bayesian Multinomial Differential Privacy Algorithm (coauthor, in-progress)
- Studies
  - Authentic Gender Disclosure Risk Assessment Using Synthetic Observations (author)
  - Generation of Differentially Private Synthetic Data, Utility of Random Decision Trees (author)
  - Synthetic Data Generation Using Variable-Depth Random Decision Trees, OPM Pay Disparity Models (author)
  - Differentially Private Synthetic Attribute Generation, Random Decision Tree Branch Splitting (author)
  - Variable Depth Random Decision Trees, Branch Growth (author)
  - Global Sensitivity (GS) Laplace Mechanism Applied to Proportions vs. Frequencies (author)

Hi-Host Phenome Project
Department of Molecular Genetics and Microbiology, Duke University

The Hi-Host Phenome Project (H2P2) seeks to interpret human genetic variation through a study of cell biology. H2P2 uses a genome-wide association study (GWAS) of cellular traits to provide a means to decipher how human genetic variation regulates cellular pathways, while also providing a system for experimental validation and mechanistic studies. The project employs a data set of more than 10,000,000,000 observations that record the association of more than 15,000,000 genetic variants in more than 500 subjects to phenotypic responses for approximately 150 pathogens (diagram). The goal of the project is to provide an efficient and intuitive on-line platform for local and public geneticists and cell biologists to conduct real-time research and genotype-phenotype association hypothesis testing by using underlying fast data query and model fitting processes.

Contributions

- Design and implement a SQL database for efficient query of over 10 billion observations across multiple tables
  - Records for more than 8 billion genetic variants (15 million single nucleotide polymorphisms for each of more than 500 subjects) were partitioned into 493 disjoint tables by SNP (the choice of 493 yielded maximum balance of SNP observations per table; choice of partitioning by SNP resulted from a study of expected primary query filtering parameters)
  - More than 2.2 billion GWAS records (one for each SNP and pathogen) were partitioned into 149 disjoint tables by pathogen (partitioning by pathogen provided balanced tables with optimum alignment to expected query parameters)
  - Partitioning of high dimension genotype and GWAS tables reduced typical query execution time from more than ten minutes to less than ten seconds
  - An optimized SNP table was designed that implements Plink-like binary encoding of individual genetic variant, which reduced overall database size by 40% diagram. Storage and query of variants requires efficient in-process parsing of bit segments of SNP records and was implemented entirely within SQL.
- Design and implement an R/Shiny web application for public GWAS research query and association exploration

  - Develop a query tool to explore associations between researcher specified SNPs, phenotypes, and genes
    - Results presented in tabular format with selectable rows by SNP and phenotype
    - Row selection generates boxplot of phenotypic response by genotype, with filtering features
    - Table download feature permits further local analysis of selected GWAS results
  - Develop a interactive heatmap to explore Spearman (rank) correlations of specified phenotypic responses
  - Develop a Manhattan plot of specified phenotype-genotype regression estimates for identification of association of phenotype to chromosome-gene position
- Develop global-map plots indicating volume of H2P2 site visits and resources accessed by location, using SQL activity records and source IP addresses taken from Shiny activity logs

Resources for the H2P2 project

- Research paper: An Atlas of Genetic Variation Linking Pathogen-Induced Cellular Traits to Human Disease (contributor)
- SQL encoding of Plink BED-style genetic variants

Interactive Cross-Phenotype Analysis of GWAS Database (iCPAGdb)
Department of Molecular Genetics and Microbiology, Duke University

iCPAGdb employs an R/Shiny web app to compare associations in two genome wide association study (GWAS) data sets: one uploaded by a researcher and one stored within iCPAGdb. Traits with significant association within each GWAS set and common chromosomal loci are identified and graphically presented. The idea is to define potential compound traits X:Y, with common genetic association, where X and Y are taken from independent studies). My role on the project is to develop and maintain the web app, graphical functions, numerical functions, and data/SQL interfaces.

Contributions

- When compared with the previous generation of computational solutions, iCPAGdb improves memory 153 efficiency with built-in functions connecting to SQL GWAS and LD proxy databases and 154 improves computational efficiency and speed by utilizing multiple CPUs. For the 155 NHGRI-EBI GWAS Catalog, the growth of GWAS findings and improvements of iCPAGdb over the previous version of CPAG led to a 27.7-fold increase in direct cross-157 phenotype associations and a 47.7-fold increase in indirect cross-phenotype 158 associations.
- Implementation of an algorithm using Fishers exact test for identifying pleiotropic effects (possible association of a genetic variant with traits from two GWAS studies being compared)
- Development of a web app using R, Shiny Server, SQL, and various R packages (Shiny, ggplot2, DT)

Resources for the iCPAGdb project

- Research paper: An atlas connecting shared genetic architecture of human diseases and molecular phenotypes provides insight into COVID-19 susceptibility (coauthor)
- iCPAGdb web app design considerations (author)
- Pleiotropy exploration app in R/Shiny
  - Explanation
  - Shiny App

Brain Computer Interface Project
Pratt School of Engineering, Duke University, with collaborators from East Tennessee State University

The Brain Computer Interface (BCI) project seeks to improve the effectiveness of human-machine communication systems used by amyotrophic lateral sclerosis (ALS) patients in a clinical setting. Data from electro encephalogram (EEG), P300 speller (the P300 speller is a visual panel device used to generate characters and symbols based on input excitation signals), and Tobii eye tracker (the eye tracker provides eye-gaze coordinates used in determining P300 panel target location) signals are encoded in concurrent streams. Accurate interpretation of signal state transitions and associated speller targets is essential to accurate reporting of a subject's intended message. A major goal of the project is to make signal data available in an open source format by converting proprietary BCI2000 files (as generated by the EEG, P300, and eye tracker equipment) into European Data Format (EDF) standard files.

Contributions

- Develop R scripts to interpret files containing BCI2000 source files, containing over fifty signal streams of 10,000 or more time stamps, and generate concurrent graphs of signals, equipment states, and selected speller symbols
- Develop parallel R scripts to convert approximately 10,000 propriety BCI2000 signal files to the open European Data Format (EDF)
- Assist researchers interpret EDF file contents for use in machine learning algorithms designed to improve the accuracy of subject message interpretation
- Develop R scripts for graphical presentation of BCI2000 and EDF time sequence signal data
- Develop algorithms in R for highlighting and reporting patterns and temporal inconsistencies of EEG, P300, and eye tracker signals

Resources for the BCI project

- Research paper: An Open, Diverse and Machine Learning Ready P300-based Brain-Computer Interface Dataset (coauthor)
- Example EEG, P300, Tobii eye tracker signal graphs

Urea Cycle Disorder Project
School of Nursing, Duke University, with collaborators from Childrens National Hospital, Baylor College of Medicine, and University of Nebraska College of Medicine

The Urea Cycle Disorder (UCD) project employs a clinical data set to be explored in a visual, ad-hoc manner for identification of network-style associations between participant demographics, physiology, metabolic signals, cognitive performance, treatments, and UCD diagnosis signals.

Electronic clinical and research data coded with Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) can be used to accelerate the discovery of detailed clinical phenotypes, which are vital to understanding the pathology and management of rare and emerging disorders. The formal semantic relationships in SNOMED CT can support the exploration and analysis of data to recognize new clinical phenotypes, but methods and tools for using these relationships in clinical analytics or research are not readily accessible. This project takes a collaborative approach to semantic-based data exploration in a large research dataset on children with rare Urea Cycle Disorders (UCD). Our approach includes a multi-disciplinary team, a graph database representation of SNOMED CT relationships, and a prototype interactive data visualization tool.

Contributions

- Develop a Neo4j graph database to contain a clinical data set representing over one hundred Urea Cycle Disorder patients along with demographic, diagnostic, anatomical, and treatment codes defined by SNOMED CT
- Develop parameterized graph database queries to relate study participants by features specified in real time by clinical researchers (up to one hundred features at up to five levels of relational depth)
- Develop interactive network graph application to compose and execute, in real time, graph queries that encourage ad-hoc exploration of associations between groups of subjects by demographic features, medical

conditions, treatments, and subjective classification (app developed in R, using the Shiny, visNetwork, Neo4j, and http packages)

Resources for the UCD project

- Abstract for presentation to the American Medical Informatics Association (AMIA) 2021 Virtual Informatics Summit: Using SNOMED CT Relationships for Data Exploration and Discovery in Rare Diseases - An Interactive Data Visualization Tool (coauthor)
- Content for 2021 AMIA Virtual Informatics Summit: Slides demonstrating use of interactive network query and graph application to explore the relationship of 'Tremor' SNOMED CT Concepts to UCD and hyperammonemia diagnoses (author)
- Video of SNOMED CT Research Webinar: Using SNOMED CT relationships for data exploration and discovery in rare diseases An example in urea cycle disorders (assistant presenter)
- Research paper: ASL expression in ALDH1A1+ neurons in the substantia nigra metabolically contributes to neurodegenerative phenotype (contributor)

U.S. Circuit Court of Appeals Decision Study

Duke University School of Law

The Appeals Project studies the effect of variables such as homogeneity of panel (ideal point correlation), influential judge phenomenon, gender effect, inter-district differences, and historical trends.

Contributions

- Develop MySQL database, using Circuit Court appeals data sourced from LexisNexis, consisting of circuit identifier, judge panels, consenting/dissenting authors, opinion text, defendants, plaintiffs, legal topics, Shepard's treatment codes, and various other identifiers. Case records cover:
  - 1,083,600 decisions
  - 17 circuit courts
  - years 1973 through 2018
  - 3,230 judges
  - 80 million characters of opinion text
- Develop statistical and graphical methods to identify patterns in the data for basic data integrity analysis, panel composition, historical patterns, and for legal theory hypothesis testing (implement in SQL and R)
- Develop R/Shiny app for interactive query and review of cases and opinions by circuit, date, panel characteristics, treatments, etc.
- Develop text analysis algorithms and R/Shiny apps to compute and graph correlation of words appearing in opinion text
- Fit logistic regression models to model proportion decisions by Shepards treatment codes and internally developed case category identifiers

Resources for the Appeals project

- Research paper: Twenty-First Century Split: Partisan, Racial, and Gender Differences in Circuit Judges Following Earlier Opinions (contributor)
- Study: Analysis of text appearing in case opinions

Colab Shiny Courses

Miscellaneous Projects

- Problem Solver - mostly recreational probability problems and computational pastimes
- Other projects - interesting work related problems and solutions
- An application of Linear Approximation of Variance - useful for modeling dispersion of complex functions (contains generators and theoretical derivations of several arcane, but useful, probability distributions)
- Jamaica storm prep - a hypothetical do/don't go fishing decision process for Jamaican fishermen using World Bank damage estimates, NOAA historical hurricane category risk models, and Markov processes