

Dayanara Lebron-Aldea

Portfolio: <https://dlebron12.github.io> | Cel: 813-993-6577 | email: daya.lebron@gmail.com

Education:

M.S Statistics-University of California, Davis, California

09/16 – 12/18

B.S Bio-Mathematics- Universidad Metropolitana, San Juan, Puerto Rico

08/10 – 05/15

Published Papers:

Thissen, JB ..., **Lebrón-Aldea, Dayanara**, et.al (2018). *A novel variant of torque virus 7 identified in patients with Kawasaki disease*, Plos One 13(12): e0209683, doi: 10.1371/journal.pone.0209683

Sichtig, H., ..., **Lebrón-Aldea, Dayanara**, et.al (2018). *FDA-ARGOS: A Public Quality-Controlled Genome Database Resource for Infectious Disease Sequencing Diagnostics and Regulatory Science Research*, Nature Communications. doi: 10.1101/482059

Lebrón-Aldea, Dayanara, Dhurandhar, Emily Jane, Pérez-Rodríguez, Paulino, Klimentidis, Yann C., Tiwari, Hemant K, V. A. I. (2015). *Integrated genomic and BMI analysis for type 2 diabetes risk assessment*. Frontiers in Genetics, 6, 1–8. doi:10.3389/fgene.2015.00075

Skills/ Programming Skills/ Statistical Interests:

- Experience with parallel computing on HPC systems.
- 3+ years of experience in metagenomic sequence analysis and big data statistical analysis.
- Proficient coder in R and Python; Perl, Bash, Unix Tools and Shell; basic C++ knowledge.
- Machine Learning methods, High Performance Computing and Exploratory Analysis.
- Jupyter notebook, Pandas, Numpy, scikit-learn, Tensor Flow and basic SQL knowledge.
- Web-data scraping, data download using API's and visualizations with ggplot2 and pyplot.
- Collaborated with multidisciplinary teams to produce scientific papers and contributed to technical reports.

Languages: Proficiency in Spanish and English written and spoken.

Relevant Coursework: Machine Learning, Statistical Methods I & II, Computational Statistics, Data Technologies, Big Data and HPC, Data Science, Analysis of Variance and Experimental Design, Time Series Analysis:

Projects (available in portfolio):

- **Exploring Stock Correlation of the S&P 500 Index**
 - Web Scraped data from Wikipedia using Beautiful Soup and Downloaded 2008-2016 Stock Information for S&P using Yahoo Finance API.
 - Explored SP500 dataset by looking at correlation between stocks and observed behavior of market during time.



Dayanara Lebron-Aldea

Portfolio: <https://dlebron12.github.io> | Cel: 813-993-6577 | email: daya.lebron@gmail.com

- Built yearly networks of the S&P500 based on correlations of stock returns using Gephi an open sourced network analysis and visualization software.
 - Identified key market segments in which it would be profitable to invest; created portfolio using Minimum Spanning Trees (MST) in Network X and tested performance with data from 2016.
 - Scripts done in Python using Jupiter notebooks.
- **Detecting Credit Fraud**
 - Used Principal Components of dataset to protect data confidentiality.
 - Implemented Random Forest, Logistic Regression and SVM with a 10-fold cross validation, in sklearn, and a Neural Network with hyperbolic tangent kernel.
 - Used oversampling and under sampling techniques to reduce bias in our incomplete dataset.
 - Achieved a 98.5% accuracy LR classifier and 99.8% in SVM.
 - Scripts done in Python.
 - **Twitter Sentiment Analysis: Use of Classification Models for Inappropriate Content with Validation on Trump Tweets**
 - Analyzed 50k tweets and downloaded Trump Tweets using Twitter API.
 - Data Processing using python nltk to remove punctuations, stop-words, hashtags and other grammatical issues. Created TFIDF matrix from words in data.
 - Trained Several Classification Models to detect positive and negative sentiment using scikit-learn and compared via AUC and ROC curves. Models: Polarity based model, SVM using hinge loss with Gaussian/Sigmoid kernels, and Neural Networks using keras.
 - Achieved a 77.6% AUC classifier through 10-fold CV.
 - Validation with Trump Tweets showed majority of negative tweets started when discussing politics when being presidential candidate in 2016. Word Representation through word cloud.

Research Experience:

Data Scientist

04/16/18-12/18 (8 months)

Supervisors: Jonathan Allen and Marisa Torres;
AKIMA LLC, Lawrence Livermore National Lab

- Supported research and software development in the area of data mining and metagenomics.
- Researched, designed and implemented analysis techniques to compare large and complex metagenomic datasets.
- Programmed parsers in Python for data mining purposes as well as conversion of genome

- information to json, fasta and other formats as inputs for our software's using Bio package.
- Create custom genome dense databases for metagenomic analysis.
- Use of genomic tools to detect divergent/ conflicting reads in reference database, as well as genome contamination removal.

Information Systems Post College Appointee

02/28/16-02/27/18 (2 years)

Supervisor: Jonathan Allen;

Global Security, Lawrence Livermore National Lab

- Collaborated in project with the FDA creating a Reference Genome Database for Infectious Disease Genomic Analysis.
- Collaborated in project with the Department of Health in Oasaka, Japan, carrying metagenomic analyses of their clinical samples and summarizing results.
- In charge of processing and launching LMAT requests from outside collaborations and extracting/producing useful information from the taxonomical outputs of LMAT.
- Mentored summer bioinformatics students on project and poster presentation skills.
- Designed and developed quality control pipeline for candidate genomes prior to be added to LMAT database.
- Scripts made in R, Python, Bash and Perl; HPC systems used: **Catalyst** and **Herd**.
- Finding and extracting contaminant fragments in candidate sequences. BLAST on Human Genomes.
- Performed quality control on 100,000+ genomes from GENBANK and built new database.

Computation Intern (CO-OP)

06/15/15-02/27/16 (6 months)

Supervisor: Jonathan Allen;

Computation Directorate, Lawrence Livermore National Lab

- Analyzed TCGA RNA-Seq datasets with the purpose of finding relationship between bacterial/viral presences in liver and pancreatic cancer.
- Scripts and Data Analysis done in R.
- Worked on external xenosurveillance project on malaria presence on mosquitoes in Africa.
- Human Genome Project Analysis and Visualization with ggplot2.
- Identification and compilation of contaminant sequences across sequencing centers.
- Developed pipeline that produces "BIOM" tables that eases interpretation and analysis of taxonomical info of genomes.
- Prepared technical reports and papers to communicate to internal and external superiors.

Research Assistant – Statistical Genetics

05/27/14-08/08/14 (2.5 months)

Mentors: Ana I. Vazquez, Hermant Tiwari, Yogasudha Veturi;

Department of Public Health, University of Alabama at Birmingham

- Performed analysis to estimate genetic variances of obesity related genes with the use of genetic and physical data from 5,000+ patients.
- Worked under the TIGER STUDY proposal (University of Texas Austin).
- In charge on communicating advances in the research through oral presentations and written communications to the Principal Investigator.
- G-BLUP; Mixed Models; Quality Control Data.

- HPC Cluster used for analysis: **Cheaha**
- Data Analysis conducted in R

Research Assistant – Statistical Genetics

06/03/13-08/06/13 (2 months)

Research Mentors: Ana I. Vazquez, Emily Dhurandar, & Paulino Perez;
University of Alabama at Birmingham

- Performed Single Variant Analysis to determine genes associated with increasing Type 2 diabetes risk.
- Built predictive models for Type 2 Diabetes using clinical and genetical information of subjects, with model selection through a 10-fold cross validation and AUC parametric.
- Compared predictive ability of models by using multiple classification models: Logistic Regression, Support Vector Machine and Neural Networks.
- Scripts made in R and Bash Scripts.
- Predictive model build including genetic information achieved almost a 80% accuracy among all classifiers.
- Published first authored peer review paper in Frontiers Journal.

Intern – Biology

05/11-08/11 (3 months)

Research Mentor: Declan McCabe
EPSCOR, University of Vermont, Burlington VT

- Planned and Conducted Research in the area of Macroinvertebrates in Biology; Sampling strategy, research methodology and literature review.
- Research focused on Collecting Samples from both urban and forest streams in order to see the effect of sedimentation in community productivity through the collection of oxygen measurements along different sites of the river as well as community abundance.
- Research strategy included adding nuisance factors such as canopy cover, day-night light and sedimentation effect to create case- control groups and also add these variables that can affect the level of oxygen available through photosynthesis in flora within the river.
- Findings showed that sedimentation greatly affects the amount of oxygen available for the community of macroinvertebrate to thrive in such rivers; the most affected are urban rivers since are the most prevalent to human deposits of sediments.

Awards:

- 2016 Travel Fellowship for BPDM Workshops and KDD Conference (awarded to only 10% of applicants)
- 2016 -2017 GEM Fellowship - Full Fellow (LLNL-UCDAVIS)
- 2016-2017 NSF Research Training Group Grant
- 2015 LLNL Poster Symposium - Most Outstanding Poster Presentation in Computation.
- 2013 UAB Research Expo (Third Place) Best Poster Presentation in Public Health
- 2013 MGE@MSA/WAESO Research Conference - Honorable Mention for Poster Presentation
- 2012 AGMUS Research Symposium Best Poster Presentation in Biostatistics
- 2012-2014 Mathematics Alliance Scholar
- 2011-2014 NSF Bio-Mathematics Scholarship

- Spring 2011: AGMUS Bio-Mathematics Excellence List
- 2011 SACNAS National Conference Best Poster Presentation in Applied Mathematics.
- 2011 Model Institute of Excellence (MIE) Award

Teaching Experience & Leadership:

Mentor Assistant for Bio-informatics Interns S Program, Global Security, Lawrence Livermore National Laboratory	Summer 2016
Mentor Teacher in Biostatistics Saturday Research Academy, Universidad Metropolitana, San Juan, PR	Fall 2012-14
Calculus 1 Tutor for Pre-College Students. Free Lance Work	Fall 2013
Mathematics Tutor, Learning Zone Center, Universidad Metropolitana, San Juan, Puerto Rico	Fall 2013

Presentations:

Unexpected finding of contaminants during careful screening of sequences for LMAT database update.

- Poster Presentation at the 2016 Poster Symposium at Lawrence Livermore National Laboratory, Livermore CA, August 2016.

Microbial Community Profiling of an Unexpected Source: The Human Genome

- Poster Presentation at the 2015 Poster Symposium at Lawrence Livermore National Laboratory, Livermore CA, August 2015.
- Undergraduate Presentations

Population Dynamics of the Green Iguana in Puerto Rico: a Pest Control Method

- Poster Presentation at the 2015 Joint Mathematics Meeting, San Antonio, TX, January 2015.
- Gene Variance Associated to Anthropometric Measurements in Obesity Associated Genes
- Oral Presentation at the 2015 UMARCH, University of Puerto Rico Mayaguez, March 2015.
 - Oral Presentation at the 2014 Undergraduate Research Symposium, San Juan, P.R, September, 2014
 - Poster Presentation at the 2014 UAB Research Expo, University of Alabama at Birmingham, July, 2014

Genome Enabling Models for Type 2 Diabetes Risk Assessment

- Poster Presentation at the 71st Joint Annual NIS and BKC Meeting, Houston TX, March, 2014
- Poster Presentation at “Posters y Postres”, Universidad Metropolitana, February, 2014
- Poster Presentation at the 2013 Undergraduate Research Symposium, San Juan, P.R, September, 2013
- Poster Presentation at the 2013 UAB Research Expo, University of Alabama at Birmingham, July, 2013

Using Actigraphy Watches to Measure Sleep Activity in Subjects with Obstructive Sleep Apnea

- Poster Presentation at the 2013 Eleventh Annual MGE@MSA/WAESO Student Research Conference, Arizona State University, March, 2013

- Poster Presentation at the 2013 Richard Tapia Computer Science and Diversity National Conference, Washington, D.C., February, 2013
 - Poster Presentation at SACNAS National Conference, Seattle Convention Center, Washington, October, 2012
 - Presentation to Bio-mathematics Undergraduates at the Universidad Metropolitana, Cupey, Puerto Rico, October, 2012
 - Poster Presentation at the 2012 AGMUS Research Symposium, San Juan, Puerto Rico, September, 2012
 - Presentation at ISIB Symposium: Iowa Summer Institute, College of Public Health, University of Iowa, July, 2012.
- Community Productivity with Sand Exposure in Urban and Forested Streams and the Effect of Fine Sediments in Grazers and Scrapers
- Poster Presentation at the 69th Joint Annual NIS and BKC Meeting, Nashville TN, March, 2012.
 - Presentation to Bio-Mathematics Undergraduates at the Universidad Metropolitana, Cupey, Puerto Rico, November, 2011
 - Poster Presentation at SACNAS National Conference, San Jose Convention Center, San Jose, California. October, 2011
 - Poster Presentation at 2011 AGMUS Research Symposium, Hotel Conrad, San Juan, Puerto Rico, September, 2011
 - Poster Presentation at Math Weekend, Universidad Metropolitana, Cupey, Puerto Rico, September, 2011