# Jake Sauter | Curriculum Vitae

78 Old Sylvan Lake Rd, Hopewell NY, 12533

845-891-9910 • jake.sauter3@gmail.com • jakesauter.github.io

Bioinformatics Data Scientist with background in statistical modelling and deep learning. I aim to collaborate with researchers that wish to gain a deeper understanding of various biomedical data modalities (sequencing, imaging, flow cytometry, etc.) through data science and statistical techniques such as dimensionality reduction, clustering analysis and predictive/inferential modelling. I am to share data visualization insights and deployed models through generative Rmarkdown reports, R/Python libraries, and web-based R-Shiny apps to attain real-world impact of cutting edge statistical and and deep learning models for patient outcome prediction and disease understanding.

## Relevant Experience

- **Northeast Information Discovery** — **Canastota, NY**
  *Data Scientist* — *March 2019 – August 2020*

  Played an influential role in a small research and development team formed with the sole purpose of designing, building and deploying mission-critical signal processing capabilities. Made use of Statistical and Machine Learning techniques as well as Linux scripting automation on a daily basis while working within modern agile processes. Responsible for training Deep Learning models and conducting report-driven analysis of developments and improvements.

- **Center for the Neural Basis of Cognition** — **Carnegie Mellon University/University of Pittsburgh**
  *Undergraduate Researcher* — *June 2018 – August 2018*

  Completed a 2 week Computational Neuroscience bootcamp filled with lectures from many leading researchers in the field. Following the lecture period, 8 weeks of research were completed under advisement from professor Marc Coutanche (Univ. of Pitt.) in the Learning in Neural Systems Lab. Wrangled data for statistical analysis of information density in various brain regions, presented in thoughtfully designed poster for presentation of experimental findings.

- **NSF MedIX REU** — **Depaul University/University of Chicago**
  *Undergraduate Researcher* — *June 2017–August 2017*

  Completed a 10 week program as a member of a small research team under advisement of Dr. Jacob Furst and Dr. Daniela Stan Raicu. Acclimation into the field was facilitated by reading scientific publications; this allowed for the development of skills involved with scientific reading. After a successful submission to the SPIE Houston 2018 conference, the paper was accepted for poster presentation, with the manuscript to be published in the conference proceedings later in 2018. Work was done in Python and is available for viewing on Github.

- **iD Tech Camps** — **Vassar College/American University**
  *Technical Coordinator/Instructor* — *June 2016–August 2016*

  Executed key role responsible for the technological well being of the camp. Accomplished a successful summer with 15 other instructors under the advisement of a director and assistant director for 7 weeks, gaining strong team working and communication skills. Provided an opportunity to teach at American University located tech-camp in Washington DC to fill open teaching position due to my proven effectiveness as an instructor in weeks prior.

## Education

### Academic Qualifications

- **Weill Cornell Graduate School of Medical Sciences** — **NYC**
  *GPA: 4.0, Master's of Computational Biology* — *8/20-2/22*

- **SUNY Oswego** — **Oswego**
  *GPA: 3.54, Major: Applied Mathematics, Minor: Computer Science; Phi Kappa Phi Honors Society* — *8/16–12/18*

- **SUNY Fredonia** — **Fredonia**
  *GPA: 3.91, Major: Computer Science, Courses in: Mathematics, Web Development; Golden Key Award* — *8/15–5/16*

### Research Publications

- **An Evaluation of Consensus Techniques for Diagnostic Interpretation** : Published in SPIE Houston 2018 Conference Proceedings on work done at Depaul University during an NSF REU that took place the summer of 2017.

### Coursework Website

- **http://cs.oswego.edu/~jsauter**: Coursework for Systems Programming, Introduction to AI, AI and Heuristics, and Natural Language Processing available for viewing as well as chronological list of academic accomplishments.

### Notable Projects

- **Master's Thesis: A Pipeline for Pan-Myeloid Flow Cytometry Data Processing and Clustering Analysis**

  Integrating expert-annotated flow cytometry data from a range of Acute Myeloid Leukemias, Myeloproliferative Neoplasms, and Myelodisplastic Syndromes into a various dimensionality reduction and clustering techniques for rare cell population identification and disease understanding. Apart of every step of the process from reverse-engineering and data-mining annotation programs, to proper data processing and normalization. Over the next months of my thesis, anticipations are to model the high-dimensional cell-marker distributional shifts between different Myeloid-based diseases.

- **Undergraduate Capstone: Applications of Statistics and Machine Learning to Genomic Clinical Group Classification**

  Project revolving around clinical group classification of cancer type from genomic data acquired through Microarrays, with the data specifically coming from the popular Golub (1999) data set. Throughout the course of the Capstone exposure to many data analysis techniques was facilitated, from dimensionality reduction (PCA) and cluster analysis to a wide array of classification techniques including Logistic Regression and SVM. After exploring many statistical methods for selecting differentially expressed genes, SAM was used to select the most differentially expressed genes at a controlled false discovery rate of 5%. The dimensionality of these selected genes was then reduced to be used as features for a classifier. This project had great success in classifying new samples and the entire documented project can be found on my Github in the **Molecular_Classification_Capstone repository.**

- **Lateralization of Brain Function**

  Project completed during an NIH-funded summer 2018 undergraduate research opportunity under the advisement of Dr. Marc Coutanche. The purpose of this project was to locate functionally correlating regions of interest (ROIs) (e.g. correlating a brain region in the left hemisphere that responds to faces to a similar region in the right hemisphere that responds similarly). Once these regions were located, representational analysis was performed through Multi-Voxel Pattern Analysis to test for the similarity of representations. Multi-Voxel representations (fMRI) of ROIs were used as features describing the stimuli that subjects are viewing. These features were then used to train a machine learning classifier on the reserved test set of fMRI voxel recodings, generating a testing accuracy score. The specific interest of this project was to determine if the combination of corresponding ROI data increased classification results over either single hemispheric ROI. Results suggested this was the case and indicates that the representations are both different and informative.

## Courses and Certifications

### Johns Hopkins Genomic Data Science Specialization

- Introduction to Genomic Technologies
- Genomic Data Science with Galaxy
- Python for Genomic Data Science
- Algorithms for DNA Sequencing
- Command Line Tools for Genomic Data Science
- Bioconductor for Genomic Data Science

### Tensorflow In Practice Specialization

- Introduction to TensorFlow for Deep Learning
- Convolutional Neural Networks in TensorFlow
- Natural Language Processing in TensorFlow
- Sequences, Time Series and Prediction

### UMichigan Applied Data Science with Python Specialization

- Introduction to Data Science in Python
- Applied Plotting, Charting and Data Representation
- Applied Machine Learning in Python

### UC Davis Learn SQL Basics for Data Science Specialization

- Data Wrangling, Analysis and AB Testing with SQL

## Technical and Personal Skills

- **Programming Languages:** Highly adaptable, mainly interested in the use of R or Python. More than 5 years of programming experience has lead to familiarity with: C, C++, Python, and Java. Have programmed for a summer in Matlab and have taken courses where I have used Assembly, Javascript, HTML, CSS, and PHP.
- **Proficient Technologies:** Comfortable in use of LaTeX. Very proficient in a collaborative git workflow, as well as use of SQLite in a Data Science environment. Spent several months developing dataset pipelines using **Tensorflow**, as well as training large deep learning models using **Keras** ported to **R**.
- **Linux:** Operated with Linux as primary operating system for close to 5 years. Excelled in a class introducing super users and programmers to the tools available through the Linux command line. Have made use of many command line tools that allow for faster and more automated solutions to interesting problems.
- **Research and Independence:** Accumulated research experience in which I was quite independent, and self-directed the research to successful results.