JEREMY J SAMUELSON

Data Scientist & Machine Learning Engineer

(770) 206-0424
contact@rhodata.solutions

I am an accomplished Data Scientist, Machine Learning Engineer, and Consultant with extensive experience creating and operationalizing solutions with demonstrated impact for client organizations in multiple sectors, including financial services and healthcare. I have an extremely consistent track record of leveraging data, creativity, and advanced mathematical techniques to help clients achieve their goals and grow their profits. I'm highly skilled in using analytics, statistics, probability, deterministic and stochastic optimization, data mining and exploration, statistical modeling, data modeling, machine learning, artificial intelligence, and deep learning to find lean, timely, actionable insights and practical solutions to real-world business problems.

Technical Skills

Programming

- R (11 years)
  - Statistical and predictive models
  - Data sanitation, validation, and transformation using dplyr, reshape2, and testthat
  - Data mining and exploration with ggplot2
  - Computer intensive statistical hypothesis testing methods, such as Monte Carlo simulation and bootstrapping
  - Designing and implementing custom R packages to meet specific data science team needs (using roxygen2 package for clear, professional-quality built-in package documentation)
  - Design and implement Web apps/dashboards, such as sales reports, OLAP cubes, project management dashboards, and model diagnostics using Shiny, shinydashboard, and plotly
  - Professional quality reporting and report automation with ggplot2 and knitr
  - Automated data collection using rvest, RCurl, and RSelenium
  - Expertise in all core "tidyverse" packages
- Python (11 years)
  - Statistical models, machine learning, deep learning, NLP/text mining
  - Data mining and exploration with Pandas and matplotlib
  - Building custom Python modules
  - Web apps/dashboards using plotly and Dash
  - Automated data collection using ScraPy, BeautifulSoup, and RoboBrowser
  - Expertise in NumPy stack (NumPy, SciPy, Pandas, matplotlib)
  - Deep learning with Theano, TensorFlow, and PyTorch
  - High performance computing for data analysis and machine learning tasks with Dask, Numba, Cython, and the NVIDIA RAPIDS stack (CuPy, CuDF, CuML)
  - Deployment of models to APIs using Flask
- C / C++ (11 years)
  - High performance computing programs for deep learning inference
  - Modified TensorFlow source code to expand functionality of TensorFlow's standard quantization implementation
  - Expanded the functionality of the TensorFlow Profiler tool to create an in-depth tool for profiling deep learning inference performance and identifying performance bottlenecks
  - Deploying optimized machine learning and deep learning models, saved as protocol buffer operation graph files, using TensorFlow for C++

- SQL (11 years)
    - PostgreSQL and MySQL Databases
    - Submit SQL queries to pull data directly into Python and R sessions using SQLite and SQLAlchemy
    - Query cloud-based relational data sources, such as AWS RDS and RedShift
    - Integrate relational databases with enterprise SAS systems using Proc SQL
- Scala (4+ years)
    - Big data analytics involving distributed cloud processing with Spark and SparkML
    - Statistical analysis, modeling, and machine learning with Breeze, Scalalab, Smile, and Spark MLlib
    - Data visualizations (usually for model diagnostics) with Breeze-vis
- MATLAB (5+ years)
    - Machine learning and signal processing applications to build proofs of concept for complex modeling applications
    - Creation of professional, publication-quality visualizations and animations to effectively display data narratives
    - Creation of simulations of business processes to identify optimal and attractor states of operational systems
- F# (4+ years)
    - Business process modeling and statistical analysis
    - Software development for embedding data science and analytical algorithms and models into software solutions developed in other .NET languages (usually C#)
- SAS (8+ years)
    - Big data analytics in the cloud utilizing the SAS Cloud Suite for Enterprise analytics
    - Statistical modeling and model interpretation to provide descriptive, predictive, and prescriptive modeling solutions

DevOps Tools

- LaTeX (11 years)
    - Incorporated LaTeX with R and Python to generate highly professional, visually appealing reports, presentations, and whitepapers using the knitr and Pweave packages
    - Data, source code, and model documentation
- Git (10 years)
    - Version controlling source code on all projects
    - Supervising updates, feature additions, and R&D tasks on the level of individual development team members through branching, testing, and merging
- Linux (6 years)
    - Used Linux Ubuntu command line to interface with cloud-based computing and storage resources
    - Bash shell scripting to efficiently automate processes
- Amazon Web Services (4 years)
    - Used AWS to automate, deploy, and operationalize machine learning solutions at scale
    - Utilized AWS Kinesis to stream data from sources for real-time processing and analysis
- Docker (3+ years)
    - Containerization of projects for portability of solutions and dependencies
    - Utilized containers to house models within APIs to expose solutions to other applications
- Kubernetes (2+ years)
    - Utilized Kubernetes to implement microservices architecture when necessary for operationalization of solutions

Integrated Development Environments

- PyCharm (Python)
- IntelliJ IDEA (Scala)
- CLion (C/C++)
- RStudio (R)
- Vim
- Sublime
- Jupyter Notebooks
- Visual Studio

Communication Skills

- Excellent listener – able to identify business needs, goals, and challenges of customers/clients/stakeholders, and team members
- Strong ability to clearly convey complex analyses and their interpretations to both technical and non-technical audiences
- Perpetually optimistic – never criticizing, condemning, or complaining, but instead seeking to understand and show genuine appreciation whenever possible
- Strong ability to put people at ease, but always remain professional
- Strong ability to find the core issue(s) of any matter and dispel unimportant details
- Strong ability to communicate solution proposals and rationale to teammates

Management Skills and Experience

- Experience with management of both local and international development teams distributed across the United States, Europe, and India
    - Currently managing a team of Software Engineers and Data Scientists distributed across multiple cities in the United States and India
    - As Principal Data Scientist at Enhance IT, managed 17 client projects in the United States, Europe, and India
    - Managed software development process (SDLC) from feature analysis and design to deployment
    - Authored specification documents showing the intended design and outlining intended feature implementation roadmap
    - Assigned tasks and supported team members on an individual basis as needed to ensure deadlines were kept and features met specifications
    - Coordinated meeting schedule with team members on both continents
- Expertise in fostering a culture of continuous improvement that enables and encourages the growth and development of each team member
- Team purpose, values, practices, and expectations are always clearly communicated to team members
- Ability to prioritize tasks to meet production deadlines
- Experience with Agile methodologies
- Experience with workflow management and productivity software (Trello, Jira)

Work Experience

SimpleMachines, Inc.
Principal Data Scientist
San Jose, CA
February 2020 – Present

SimpleMachines, Inc. is an exciting technology start up based in Silicon Valley. SimpleMachines, Inc. is developing new, proprietary hardware acceleration technology to achieve faster performance in deep learning inference than is possible with existing GPU architectures. As the Principal Data Scientist and Machine Learning Engineer, I am responsible for the correct and efficient implementation of deep learning models in the SimpleMachines, Inc. machine learning and deep learning software stack.

- Leading the Data Science and Software Engineering teams in the correct implementation of important benchmark deep learning models
- Implementing benchmark computer vision models, including ResNet50, Inception, VGG-16, and SSD using a modified build of the TensorFlow deep learning framework
- Implementing benchmark natural language processing (NLP) models, including LSTM models, Seq2Seq models (word2vec, doc2vec), attention models (BERT, BERT-SQuAD), and translation models (GNMT) using a custom modified build of the TensorFlow deep learning framework
- Meeting with clients to analyze custom client deep learning models and optimize model performance on the custom hardware
- Analyzing all models using TensorFlow Tensorboard to visualize the op graph of each model
- Use analysis of each model op graph to fully document and characterize all operations for the correct implementation of each operation in the software stack and optimal integration with the proprietary hardware
- Perform experiments and analysis to find optimal parameters for model quantization to maximize compute performance while maintaining prediction "accuracy"
- Modifying existing TensorFlow quantization graph transforms to extend and improve implementation for higher-quality and more granular quantization of model operations
- Modifying the TensorFlow Profiler tool to include custom performance analyses and to create a stand-alone performance analysis tool for customers
- Performing in-depth performance analysis of each model using Python and TensorFlow to identify performance bottelnecks
- Visualizing and presenting insights using common Python tools for Data Analysis, including Pandas, matplotlib, seaborn, Dash, and plotly
- Comparing achieved model performance to that of competitors by analyzing performance data on the ML-Perf platform

- Developing, testing, and implementing proprietary approximations of activation functions commonly used in artifitial neural networks to further optimize compute performance of all deep learning models implemented within the software stack
- Maintaining the stability of the source codebase through designing and implementing unit test and integration tests prior to merging new code into the master Git branch
- Engaging with potential customers, clients, or investors to answer technical questions and assist with further development of the business
- Design and create demonstrations of models currently implemented in the software stack for client presentations as needed
- Regularly meet with company leadership to provide updates regarding the development of the software stack
- Lead and organize development efforts through the use of Agile SCRUM

Enhance IT
Principal Data Scientist
Atlanta, GA
December 2018 – January 2020

After the successful launch of the data science practice in the UK, I was brought to Atlanta to establish the data science practice for Botster.ai's sister company in the US. Established and documented the same best practices utilized in the UK, with some minor adaptations to better fit the culture of business in the US. Currently expanding the practice to realize previously unexplored revenue streams for the data science practice within this organization. Also, evaluating and transforming internal analytics and key performance indicators to yield more accurate and useful insights within the company.

- Established improved filters and criteria for hiring data science talent
- Expanding training curriculum to give more thorough and complete examples in areas where junior data scientists seem to struggle with comprehension and application
- Digitizing curriculum to make the material more accessible for review and self-study
- Engage with implementation partners to explore avenues for partnership and mutual benefit
- Engage with clients to assess client needs and opportunities, and to propose possible data science and machine learning solutions
- Leading all ongoing client projects to ensure a high standard of quality for deliverables, and most importantly, to ensure projected results and profit impacts are realized
- Implemented a time series model for a large telecommunications client. The model is able to project expected customer service center call volumes with very high

accuracy and detect when volumes are anomalously high or low to alert the client of potential network failures.
- The solution was built, tested, and deployed using Python
- The solution is deployed to a stateful API in an AWS serverless environment
- Implemented a propensity model and lead scoring model to prioritize leads, and to evaluate the effectiveness of sales agents for a large payment systems client using Python
- Implemented a time series model to forecast revenue performance of various investment instruments for an investment banking client using R
- Implemented and deployed a Customer Lifetime Value model to help guide sales representatives for an investment banking client using R
- Built a custom Independent Gaussian Mixture Model to detect fraudulent claims for an insurance client. Custom implementation outperformed all other proposed solutions built with canned modeling APIs, such as SciKit-Learn and statsmodels.api
- Implemented a Hidden Markov Model for survival analysis to predict time to hospitalization for recently discharged patients for a large insurance client
- Identified redundant and contradictory KPIs within Enhance IT's existing internal analytics reports
- Identified statistical modeling and machine learning use cases to infer correlation of KPIs with desired results, and to provide descriptive, prescriptive, and predictive insights from the available data

Botster.ai
Principal Data Scientist
London, UK
April 2018 – December 2018

Established the data science practice within an existing IT consultancy, based in Whitechapel, London, looking to add machine learning, deep learning, statistical modeling, and advanced analytics to their offerings. Defined and lead all data science initiatives within the organization including, but not limited to, defining hiring criteria for data science talent, creating training curriculum for junior data science consultants and leading trainings, engaging with clients to determine needs and advise regarding the feasibility of machine learning projects, creating and leading data science teams for each client project, and driving results and projected profit impacts for the client.

- Created training curriculum founded upon the mathematical theory and principles of machine learning, with a view toward practical application in industry
- Lead daily training sessions to ensure junior data scientists were given a clear and thorough exposition of the material

- Successfully implemented a custom and proprietary deep learning solution for an energy sector client. The solution automates much of the process for identifying potential natural resource deposits utilizing satellite images and seismic reading data.
- Used MATLAB to build a proof of concept for the model and model training in a local environment with a relatively small training set
- The final solution is deployed to an on-premises environment on the client site.
- The production model was recoded from MATLAB to Scala and Spark.
- Automated data ingestion using Kafka and Spark Streaming
- Advised on data monetization strategy for a large investment banking client
- Implemented a fraud detection model for said investment banking model to detect internal financial crime
- The fraud detection model was a custom ensemble method implemented in Python.
- The fraud detection model was able to detect 94% of the type of fraud targeted while maintaining a precision of over 90%, which was the desired precision threshold. There was no existing machine learning solution to which performance could be compared.

MasterCard
Lead Data Scientist
Bellevue, WA
April 2017 – April 2018

Leading a team of data scientists embedded within a larger cross-disciplinary team of highly specialized consultants and subject matter experts in the fields of financial services and regulation. Utilizing an Agile SCRUM framework to lead the data science team in end-to-end implementation of statistical/machine learning models to improve early detection algorithms for credit and debit card fraud.

- Created a custom ensemble machine learning system, which improved the detection rate for fraudulent card transactions from 98.7% to over 99.9%.
- Consulted with regulatory and subject matter experts to gain a clear understanding of information and variables within data streams
- Researched current state-of-the-art methods and best practices for the specific problem domain
- Planned all phases of the development cycle before official project kick-off to determine project-specific needs and implement necessary solutions, including implementing custom libraries in Scala for tracking end-to-end dependency paths from data sources, to modeling algorithms, to diagnostic reporting

outputs within the project and checking for circular dependencies
- Supervised definition of features and raw feature space, as well as clear documentation of said features
- Guided development team data scientists in fitting several preliminary Bayesian and machine learning models in Scala and Python (with PySpark for data retrieval in Python) for improved understanding of data, and feature selection
- Utilized cloud/cluster computing resources for model optimization/tuning of hyperparameters, and cross-validation
- Implemented diagnostic reporting with Python for benchmarking performance across models and final model selection
- Following final model selection, supervised the translation of prototype code into optimized Scala code for improved performance, robustness, and maintainability of source code
- Lead team data scientists in the final round of model benchmarking and analysis to ensure no loss in model performance from prototyped to productionized implementation
- Supervised final rounds of testing and debugging, and finalizing model documentation
- Responsible for implementing and leading Agile KanBan management framework to guide the exploration and experimentation phases of the project
- Transitioned to, and lead, Agile SCRUM for deployment and operationalization of the solution
- Responsible for troubleshooting production-related issues and coordinating successful resolution
- Maintained integrity of source code throughout project via version controlling with Git

Vici Capital Partners
Senior Data Scientist
Salt Lake City, UT
March 2016 – March 2017

Utilized the Agile SCRUM framework to supervise a development team distributed across the United States and India in developing a suite of indirect procurement solutions to be used by senior-level procurement subject matter experts within Vici Capital Partners on future consulting projects. Following procurement consulting projects, the procurement solution suite is licensed by Vici Capital Partners to consulting clients for continued enterprise-wide spend data visibility and supplier management.

- Gathered features through interviews and online conferences with procurement subject matter experts
- Authored specification documents showing the intended design and outlining intended feature implementation roadmap

- Lead weekly scrum sessions, assigned tasks, and supported individual team members on an individual basis as needed to ensure deadlines were kept and features met specifications
- Performed research to locate publicly and privately available data sources for building a long-term knowledge base to inform various solutions within the suite
- Supervised junior analysts in using R to perform initial sanitation, validation, and transformation of data sources to prepare data for exploratory analysis
- Supervised team data scientists in exploratory analysis using R and Python to determine how data sources could be used to satisfy the implementation of software features
- Supervised and collaborated with data scientists and developers to implement Online Analytical Processing (OLAP) Cube in F# with powerful, unique, and nonstandard features, including agile, in-application categorization/recategorization of spend
- Guided programming team in India in the design and implementation of UI/UX functionality and layout (in C#)
- Met frequently with procurement subject matter experts, and other end-users, for iterations of feedback and redesign loops on UI/UX to ensure that functionality and workflow were intuitive and clear
- Researched techniques for optimizing runtimes to ensure UI/UX felt snappy and responsive
- Conveyed optimization techniques to programming team in India and guided programmers in optimizing runtime until all solutions met performance standards
- Implemented NLP techniques and logistic regression model in Python to suggest vendor name matches to assist users with normalization of vendor/supplier names in the data
- Used Python to implement unsupervised learning model (K-means clustering) to group suppliers based on categories of spend and propose optimal supplier consolidations
- Responsible for Troubleshooting production-related issues and coordinating successful resolution
- Managed feature updates and testing via repository branching with Git
- Supervised team members in building dashboards to diagnose model performance after deployment in R using shinydashboard.

Promontory Growth & Innovation
Data Scientist
Washington, DC
February 2014 – February 2016

As part of a cross-disciplinary team of consultants and subject matter experts, used creative problem solving, business acumen, application of mathematical theories and methods, data mining, statistical analysis, modeling, and machine learning methods to drive massive profit improvements for clients in the financial services and healthcare industries.

- Used Python on EMR cluster in AWS to perform data analysis on over 1 TB data set of medical procedures to find minimal cost equipment sets for each medical procedure performed by a large hospital client; analysis was used to propose minimal cost standardization of procedures
- Located relevant publicly available data sources to supplement data sources provided by hospital client; used R to sanitize, validate, and transform data to be fit for purpose
- Used R to implement regression analysis on data set incorporating hospital client data, AHCA data, and U.S. Census Bureau data on health insurance to model evolution of client's payor mix over time and benchmark client payor mix against local competing hospitals
- Created novel metrics for various hospital processes that were more informative than industry-standard metrics used by hospital client
- Used R to perform root cause analysis to identify process breakdowns within departments and provide information through the use of various data visualizations to find and communicate solutions to process breakdowns in both operating room and nursing staff scheduling
- Implemented linear programming using R to optimize Emergency Department nursing staff schedule to minimize overtime for a large hospital client
- Used R and PostgreSQL to sanitize, transform, and combine accounts payable, general ledger, cost center, and supplier data sets from many disparate sources to create a master database for enterprise spend analysis for a very large bank client
- Quickly implemented and deployed a basic Online Analytical Processing (OLAP) Cube with Python for multidimensional visualization of spending, since no suitable off-the-shelf solution could be found while on project
- Used Python to implement unsupervised learning algorithm (k-means clustering) for optimizing supplier consolidation in indirect procurement analysis for a large bank client
- As part of a development team, took part in the design, implementation, and analysis of a pricing model deployed for use by the sales quoting team of a biological supplies client; model was comprised of multiple layers, or sub-models, implemented in R and Stan (for probabilistic modeling)
- Built an online reporting dashboard using R and Shiny to automate the diagnosis of anomalous model outputs
- Used R and shinydashboard to create a sales reporting dashboard to be used by sales management to ascertain profit impact of the pricing model, as well as evaluate performance of individual sales representatives, identify customer purchasing

patterns, and identify strong and weak items in client's product catalog
- Validated cashflow models implemented in multiple technologies used to rate bonds and other securities in preparation for regulatory review with the SEC for a large ratings agency client
- Performed meta-analysis on portfolio data from past Promontory Growth and Innovation consulting projects; used R to implement a logistic regression model to predict the outcome of a given portfolio idea as a function of the idea's projected profit impact and risk classification
- Communicated results of meta-analysis to managing directors through reports generated with R and LaTeX using knitr package

SemanticBits
Data Scientist
Herndon, VA
January 2013 – January 2014

Utilized various common data science tools to enable clients in the healthcare industry to tame big data sources and extract lean, timely, and actionable insights into process improvement opportunities. As a team member on multiple software development projects, I provided expertise in implementing machine learning algorithms to be embedded in custom BI applications.

- Used Hadoop, MapReduce, and HQL to access big data sources in the cloud for cleaning, organizing, validating, and transforming raw data in preparation for data mining/ exploratory analysis
- Used SAS and SAS Cloud Suite in instances where client big data sources were implemented in SAS
- Used Python, R, or SAS to perform initial data exploration and analysis to determine whether data sources were fit for purpose, and to ascertain how data sources could be used to reach stated project goals
- Performed preliminary prototyping and design for data science-intensive features using Python
- Utilized matplotlib in Python to generate data visualizations to convey results, diagnostics, and useful insights to team members and team lead
- Translated prototype code from Python to Scala for more robust implementation of models/machine learning algorithms, and to enable UI/UX programmers to easily embed data science features within BI applications (UI/UX programmers used Java)
- Used Python or R (depending on the team) within LaTeX and the appropriate data visualization modules/packages to automate the periodic generation of reports on model performance and diagnostics, as well as other tests for adherence of algorithms to defined business logic
- Used LaTeX to create professional-quality, highly maintainable documentation for all algorithms, business logic, and

statistical/machine learning models implemented for long term maintainability of source code by future programmers/data scientists assigned to the project
- Responsibly managed time to finish all tasks/sprints within assigned point allocations
- Regularly reported to team lead to ensure all features met production specifications

ManaHealth
Data Analyst
New York, NY
August 2010 – December 2012

Data analyst on advanced analytics team with many responsibilities, such as assisting clients with designing and implementing analytical data strategies for large claims and clinical data sets using predictive modeling, data mining, econometrics, and statistical methods. Drawing insights out of complex, often unstructured data. Assessing information from a range of data stored in disparate systems, integrating data, and mining data to answer specific business questions as well as identifying unknown trends and relationships in data.

- Worked collaboratively with teams of health services researchers, business analysts, and data stewards to draw insight and intelligence from large administrative claims datasets, electronic medical records and various healthcare registry datasets
- Enriched client data with third-party sources of information, such as US Census Bureau data on health insurance metrics, USPS data for defining zip codes (for stratification of location/temporospatial analysis), and many other examples, when needed
- Enhanced data collection procedures to include data and information relevant for building analytical systems
- Developed and tested hypotheses in support of research and product offerings, and communicated findings in a clear, precise, and actionable manner to clients using R
- Used R to program various computer-intensive hypothesis testing algorithms, such as Monte Carlo simulation and bootstrapping confidence intervals
- Processing, cleansing, and verifying the integrity of data used for analysis with R
- Ad-hoc analysis and presenting results using data visualizations generated in R and Python
- Implemented various machine learning techniques and algorithms (regression analysis, SVM, random forests, neural networks, and deep learning) in R and Python
- Applied statistical techniques, such as statistical sampling, hypothesis testing, regression, etc. to a variety of real-world business problems for clients using R and Python (depending on the client/project)

- Analyze and understand large amounts of data to determine suitability for use in models and then work on segmenting the data, creating variables, building models, and testing said models
- Performed Exploratory Data Analysis and Data Visualizations using R, and Tableau
- Used R to perform univariate and bivariate analyses to understand intrinsic and combined effects
- Established data architecture strategy, best practices, standards, and roadmaps
- Involved in analysis of business requirements, design, and development of high-level and low-level designs, unit and Integration testing
- Interfaced with the other departments to understand and identify desired insights and determine data needs and requirements
- Responded to operational data requests and created ad-hoc queries in SQL to support research and business analytics projects

Honeywell
Data Analyst
Atlanta, GA
July 2009 – August 2010

Worked with business analytics and infrastructure team to develop strategies to clearly outline how various enterprise platforms could enable business growth and productivity for clients by combining structured and unstructured data sources to identify opportunities for improved efficiency through data mining, exploratory analysis, statistical models, data visualization, and reporting

- Cleaned and organized big data sources using R, Python, and SAS
- Participated in all phases of data mining: data collection, data cleaning, developing models, validation, and visualization with R, Python, and SAS
- Collected data for analysis and modeling from relational database sources using SQL
- Interfaced with both relational and non-relational databases using SQL and NoSQL, respectively
- Utilized methods, such as Principal Component Analysis (PCA) in R, Python, and SAS to reduce the dimensionality of data sources with large number of features in preparation for modeling using R, Python, and SAS
- Imported data into dashboards on Tableau for accessibility and use by various levels of management
- Met with decision-makers, system owners, and end-users to define business requirements
- Translated business requirements into data requirements to inform data mining process

Education & Certifications

University of Arizona

Bachelor of Science in Mathematics – Statistics and Probability Emphasis

Minor with Department of Information, Science, Technology, and Arts – Data Science Emphasis

Amazon Web Services

AWS Certified Machine Learning – Specialty

NVIDIA

High Performance Computing with CUDA and Python