**SANGEETA**

**703-483-9183**

# Professional Overview

Versatile Data Scientist and Engineer (**Public Trust Clearance)** with 10+ years of IT domain experience in team leadership and hands- on involvement to combine data mining techniques through software development (SDLC). Creative thinker, problem solver, with technical know-how of predictive simulations and data analytics possessing sound knowledge of machine learning methods and algorithms.

* Proven track record in translating scientific and business needs into technical components and required computing and storage specifications
* Strong experience with large sets of structured and unstructured data and design of ETL processes to generate statistical and analytical report enabling strategic decision making
* Excellent at data cleansing, feature selection and algorithm design to run data analytics over suitable metrics
* Proficient in developing software to use statistical methods for regression tests and machine learning/artificial intelligence to extrapolate variables for predictive analytics and creating dynamic visualization to share actionable insights
* Hands-on at creating normalized data models for relational database management system to be implemented within software application
* Excellent at multi-tasking and communications with teams possessing varied technical/non-technical skills and background

# Education

## Ph.D.: Bioinformatics and Computational Biology 2015

George Mason University Fairfax, VA, USA

# Technical Qualifications

* + Development/Database: Python, C/C++, R (Statistical, Bioconductor, and Shiny), Unix/Linux shell scripting, Perl, Fortran, Oracle, SQL Server (SSMS, SSIS, SSAS), HTML/CSS, Jupyter, Airflow, Docker, Version Control: Git, SVN
  + Visualization/Data processing: Tableau Desktop, Tableau Server, Matlab
* Cloud services: AWS

## Added Skills:

* + Scientific Computing and Machine Learning (NumPy, SciPy, SciKit-Learn, TensorFlow, Keras, Theano, NLTK, Word2Vec, SpaCy, Matplotlib, Seaborn, pandas, Flask, Plotly), Computational / Systems Biology, Bioinformatics, CUDA, OpenMP

# Experience

## Data Scientist at ASCO’s CancerLinQ Alexandria VA (January 2020 – Present)

CancerLinQ (CLQ) connects cancer care team members across the country with real-world cancer data. Through CLQ participation, oncologists gain personalized guidance on treatment decisions by matching each patient’s care against quality standards and data from patients with similar characteristics. At CLQ, I lead the efforts for outbound delivery of data for curation and oversee internal systems to ensure integrity of incoming EHR data from participating medical practices.

* + Lead internal team project management, plan/track delivery schedule, monitor tasks completion and issues, and inspire communications dynamics
  + Supervise cross-team efforts to create a new application implementing principles of machine learning (ML) and natural language processing (NLP) to extract patterns in diagnosis type and treatment plan from real-world unstructured cancer data from EHR systems
  + Perform text analysis using Natural Language Processing techniques and Python NLTK library on unstructured content inherited from external data sources
  + Maintain automated pipelines to conduct data investigation and appropriate mapping of each incoming record from EHR into in-house system
  + Participate in building CI/CD pipeline to automate deployments using Docker
  + Update existing and devise new data models to support product features in a scalable way
  + Provide subject matter expertise to allocate and design optimal AWS infrastructure to handle millions of records of PHI including strategies to normalize structured data and locate corresponding unstructured data
  + Create and maintain documentation on internal and client facing business processes
  + Drive independent tasks and keep team and management in loop on next steps and action items
  + Coach peers to define, implement, and maintain statistical quality control metrics and methods to monitor and assess the quality and integrity of data within the production repository

## Skills used: AWS(EC2, S3, RedShift, IAM), RedShift, PostGreSQL, Python, Natural Language Processing(NLP), PySpark, Tableau, Airflow, Docker, Git, Windows, Mac

* **Database Developer/ Data Analyst/ Tableau Developer and Administrator at REI Systems, Inc. Sterling VA (November 2015 – January 2020)**

REI Systems, Inc. develops and manages, among others, a grants management application used by HRSA to audit/supervise their use of appropriated funds by beneficiary institutes.

* Research and build POCs to demonstrate for business development RFPs
* Perform data cleansing and feature engineering using Python libraries such as Pandas, NumPy and use harmonized data to build modern ML/NLP algorithms for applications including recommendation systems, risk monitoring, topic modeling
* Liaise with client stake holders on-site at HRSA/HHS with detailed data requests, report overviews and dashboards for data dissemination at national and state level
* Write highly complex yet clean, robust, and maintainable T-SQL scripts with script versioning, including stored procedures, triggers to implement business requirements
* Utilized SSIS for ETL data modeling, data migration to build brand new SQL Server data warehouse while ensuring data integrity
* Enhance and upkeep OLAP cube using SSAS
* Conduct impact analysis for change requests in the enterprise solution
* Fix errors and maintain database sanity and SQL Server Replication on Highly Available servers
* Adhere to and update SOP documentation, and Create new and maintain existing data models using ErWin
* Exhibit deep understanding of data flow to build interactive Tableau dashboards on live data
* Build dynamic visualizations with features such as filtering, top-down drilling, graphical representations, geographical maps et cetera with R and Python for backend data processing in Tableau Desktop
* Build, own and manage security across pre-production and production multi-node Tableau clusters to maintain server health, dashboard deployment activities, and user privileges for 2500 federal staff members
* Participate in daily stand-ups for task updates in Agile environment

## Skills used: T-SQL, SQL Server (SSMS, SSIS, SSAS), OLAP, OLTP, ETL, Python, R, Tableau Server, Tableau Desktop, SVN, Multi-node Tableau cluster set up and maintenance, ErWin, Windows, Agile, Advanced MS Excel (formulas, pivots, V-lookup, H-lookup)

* **Scientist at George Mason University Fairfax VA (Aug 2009 – October 2015)**

As a Doctoral researcher, worked on systems biology-based projects with varied scope and application, ranging from mathematical modeling to pattern recognition and prediction analytics. Outlines of the projects and their key features are highlighted here:

* Identify and integrate from disparate data repositories including publicly available resources, key components to create computational model for heart cell under various biological states
* Develop and deploy innovative algorithms to deterministically simulate using GPUs employing dynamic programming and statistical modeling to simulate time-series progression of biological states in healthy and non-healthy heart cell
* Create and implement statistical and probabilistic principles to perform image processing and compare digitally reconstructed images of neurons
* Design, execute and compare performance of Machine Learning models to predict protein structure using training set from public repositories
* Use SQL to extract known protein structures from public database and create visualization to compare predictions against known results on internal web page
* Publish work at conferences and journals
* **Skills used: Statistical modeling, ODE solving, Probabilistic modeling, Machine Learning, Image Processing, Web Development, MySQL, HTML, XML/CSS, Fortran, Python, CUDA, IDL, OpenMP, Linux, Windows, GPUs**

## Scientist at Zydus Research Center, Ahmedabad India (March 2009 – June 2009)

* + Conduct exhaustive literature research to capture activity pathway, flux rates, and metabolite concentrations for specialized biological process
  + Research appropriate technologies to create a Systems Biology based model to simulate biological function
  + Design stochastic 2-D spatial model and run simulations employing statistical methods to analyze system steady state while introducing computational perturbation to mimic normal and disease molecular pathways
  + Use results to identify potential drug targets to inform pre-clinical trials
  + **Skills used: CellDesigner, CoPASi, Statistical modeling, ODE solving, Stochastic simulation**