# Jerome Siljan

JeromeSiljan@gmail.com • (469) 733-8585 • GitHub • LinkedIn
Data Scientist | Software Engineer

## Summary:

Data scientist adept in the entire data pipeline, with strong mathematical and programming skills. Proficient in Python, Docker, SQL and skilled in managing virtual environments as well as other industry practices, including version control, evident in a robust GitHub portfolio.

## Technical Skills:

**Languages:** C, C++, Python, R, SQL, Java, ARM Assembly, Verilog (HDL)
**Certifications:** Alteryx Core Designer, Certified Entry Level Python Programmer (PCEP)
**Tools/Libraries:** BigQuery, pyspark, tensorflow, scikit-learn, Apache Hive, Docker, opencv, pandas, matplotlib, numpy, flask, SQLite3, AutoSys git, BitBucket, Confluence, Alteryx

## Education:

Bachelor’s of Science in Computer Engineering at The University of Texas at Arlington
Spring 2024

## Work Experience:

**Buxtonco – Data Analyst**

- Develop custom analytic solutions using consumer intelligence to aid in real estate and marketing decisions for retail, restaurant, and healthcare clients to optimize their potential and performance.
- Implemented a feature enabling data analysts to craft custom SQL queries within a Python framework. This innovation bridged the gap between analysts and the data pipeline, greatly boosting productivity.
- Analyzed consumer data and site location data for patterns of behavior and other qualitative findings.

**Citigroup – PBWMT Summer Analyst**

- Developed a high-impact program that integrated Impala querying functionality into an HDFS database, processing approximately 20 thousand records daily using AutoSys.
- Collaborated with a team to realize a document scanning and OCR application that streamlines the credit card application process for users at Citi.
- Led an overhaul of legacy code, replacing hardcoded implementations with a configuration file based solution to significantly improve readability and maintainability.

## Projects:

**Forest for Flood**

- Trained and tuned random forest to assign flood insurance policies based on features such as latitude, longitude, elevation, etc.
- Processed over 50 million data points from FEMA’s flood insurance database with pyspark’s RDDs to ease fast, multithreaded processing.
- Optimized random forest by adjusting hyperparameters with a grid search.
- Analyzed and culled features in the preprocessing stage to reduce training time and improve accuracy.

**Monocle**

- This project was built to solve CBRE's challenge from the 2022 Texas A&M Datathon.
- Used machine learning to identify, read, and cluster text in order to sort them into sequences using **CRAFT** and **tesseract**.
- Designed and implemented a custom clustering algorithm that groups text based on their coordinates.
- Increased OCR accuracy from 2% to 80% by including text detection (using **CRAFT**) in the preprocessing stage.
- Developed **flask** web UI that interfaces with the OCR component to display results.
- Successfully organized and managed the work load between a team of engineers in order to complete the project in a tight timeline.