GABRIEL ALON
(650) 799-9295 | gabrielalon257@gmail.com | California
www.linkedin.com/in/gabrielalon | github.com/galonpy | Portfolio: galonpy.github.io

Overview: Senior Data Scientist with broad experience at A.I. companies. Strong at software, data, and product applications.

EXPERIENCE

BluelightAI February 2024 - Current
Senior Data Scientist

BluelightAI is a startup that discovers performance issues in ML and Generative AI models using embeddings, clustering and NLP.

- **Reduced** error pattern identification **time** for RAG and LLM pipelines **from hours to under 10 minutes** by building a state-of-the-art text summarization feature. Built using HuggingFace, Numpy, Pandas, Matplotlib and C-TFIDF research.
- Improved the **speed** of customer engagements and internal development cycles from **weeks to under 1 hour** by programming a Sentence Transformers Python integration to automate the data prep phrase of the product (embeddings-based vectorization).
- Brought in the company's first 1,400+ Python software downloads by inventing a **novel RAG evaluation** approach. Gave a **presentation** on RAG evaluation with BluelightAI's CTO, promoted by Zilliz (a vector database) to **10,000 enterprise users.**
- Developed a corporate partnership after discovering that **24% of ecommerce recommendations** returned worse search results after using BluelightAI's evaluation tool on the partner's neural network fine-tuning algorithm and sales dataset. Programmed an integration with the partner's vector database, pitched to their CEO, and then co-marketed with them to their 7000 users.
- Found thousands of code and medical related documents not yet used in standard LLM pre-training pipelines. Identified using BluelightAI's product on a sample of 140,000 from Red Pajama V2. Resulted in BluelightAI's first LLM developer user.
- Developed a Python integration to the **OpenAI API t**o enable exploratory text summarization of ML model errors with **custom prompt engineering** on top of clusters from topological analysis. Created and pitched a comparative analysis of this pipeline to a T-SNE based pipeline from OpenAI that led to a POC and license key at Together AI.
- Trained, fine-turned and evaluated Tensorflow and Hugging Face neural network embedding models for the ATIS benchmark to develop a POC LLM chatbot model repair solution and pitch deck for Salesforce. Detected 82% of simulated data drift.

Upwork May 2022 – February 2024
Freelance Data Scientist

- **Reduced** sales team **inefficiency by 12%** by removing unnecessary phone numbers in sales outreach excel spreadsheets of firms that were already called. Used SQL, Python Pandas, Regex, data normalization, fuzzy string matching and NLP.
- Discovered 27,000 high value government contract opportunities using SQL and Python Pandas for a startup founder. Contracts were used to help build a contract matching platform for small businesses to find eligible government contracts.
- Evaluated NASDAQ stock price Python algorithms for a client using published approaches on Optiver closing trades data.
- Conducted academic and market research on ML/NLP algorithms for identifying A.I. generated text as well state-of-the-art training approaches for sentence transformer neural network models for a stock trading client.

McD Tech Labs (Purchased by IBM) | Mountain View, California June 2019 - July 2021
Data Scientist

First data scientist in an A.I startup that was **acquired** and grew exponentially from 12 to 70 people.

- **Directly improved $14,000,000** worth of customer interactions (2% of sales a day across 100 stores) by fixing Python bugs in a logistic regression model that was causing the Siri-like voice AI to ignore customer speech. Implemented and deployed corrected evaluation metrics for precision and recall into production using scikit-learn.

- Discovered that the A.I automation rate was overestimated by 10%, by linking receipts and human interaction timestamps. Created performance tracking metrics and data visualizations with Python and SQL that summarized millions of A.I handled transactions of voice only food orders. Used pandas, scikit-learn, matplotlib, tableau and superset.

- Detected a 14% performance decline from machine learning model drift using time series models in Python, and data selection SQL via AWS RedShift. Identified which versions of the software were performing better over time. Used the causal inference technique difference-in-differences for the time series analysis.

- Wrote production level code with unit testing, CI/CD, static analysis, Jira tickets and code reviews.

- Presented a weekly performance report used for prioritizing engineering resources. This report highlighted product trends, root cause analyses of specific software failures, and proposed product requirements.

Wefi | Tel Aviv, Israel and Santa Cruz, California July 2015 - January 2016
Data Analyst (Part-Time)

- **Tableau** visual **published** on **Venturebeat.com** in an article titled "Attention Shoppers: Shopping with a store app…" that revealed the positive implications that retail apps have on in-store visits and total minutes in the retailer's physical store.
- Wrote SQL, Tableau, and Excel reports on broad mobile app usage patterns from a dataset of 1 million Android phones.

- Produced reports by custom request, i.e: informing investors prior to Square's IPO and relating ads to app behavior.

RESEARCH

University of Michigan May 2022 – February 2024

- Authored the **first effective solution** to an unsolved cybersecurity issue for LLMs like ChatGPT one month after the issue was covered by the **NYTimes**. Combined GPT-2's perplexity property with a gradient boosting model to achieve 90+% detection.

- Recipient of the Miller's scholarship for open-sourcing a YouTube video browser that provides personalized content filters. Fine-tuned a BERT model in Pytorch to filter out toxic content based on live user typed preferences. Deployed into a website for watching videos, using Streamlit and the OpenAI API. Led a team of four data scientists at the University of Michigan.

- Developed a healthcare prediction model to produce a text status of patient health based on NDC prescription records. Used Python and SQL to link CCS and NDC records from an insurance provider's claims. Built bayesian and random forest models.

SKILLS

Languages: Python (7 years) (Object Oriented, Data Structures/Algorithms)
SQL (5 years), Spark SQL, PostgreSQL, R, Unix, Bash, Shell, Git, Excel, HTML, CSS, Adobe Analytics, NoSQL, PySpark
Libraries: Pandas, Numpy, Scikit-learn, Matplotlib, Pytorch, Hugging Face, Langchain, NLTK, TensorFlow, Statsmodels, Seaborn,
Tools: Jupyter Notebook, GitHub, Linux Terminal, Tableau, Power BI, Excel, IDE, Azure, GCP, Scipy, ML Ops, VScode, Databricks, Hadoop
Datasets: Customer, B2B, Healthcare/Clinical, Web Scraping, DAG, Geospatial, JSON
Frameworks: Statistics, Causal Inference, Hypothesis and A/B Testing, Agile, Classification, Regression, Clustering, NLP

EDUCATION

University of Michigan, Ann Arbor May 2022 - August 2023
Master of Applied Data Science
3.7 GPA

University of California, San Diego Graduated 2019
Bachelor of Science in Management Science