Zachary Webel

Data Scientist & Applied Mathematician

Washington, D.C.

561-225-4251

zow2@georgetown.edu

@zac-webel

zacharywebel

Profile

Experienced Data Scientist and Applied Mathematician with a **passion for leveraging analytical skills to extract valuable insights from complex datasets**. Possessing a strong background in mathematics and statistics, coupled with expertise in machine learning and data mining techniques, I am adept at transforming raw data into actionable solutions. A **collaborative and innovative problem-solver**, I thrive in multidisciplinary teams, effectively communicating technical concepts to non-technical stakeholders.

Skilled engineer with over 8+ years working with Python, 5+ years working with Machine and Deep Learning and experience with TensorFlow, Keras, Scikit Learn, PyTorch, PostgreSQL, R, C++, Docker, CUDA, MLOps, Generative AI, Large Language Models (LLM), BigData, and AWS. Continuously staying updated with the latest industry trends and advancements, I am **committed to applying cutting-edge methodologies to deliver impactful solutions that address complex business challenges**.

Education

08-2022 – 08-2023

Master of Science: Health Informatics & Data Science
Georgetown University – Washington, D.C.
GPA: 4.0

Courses Taken:

- Massive Health Data Fundamentals
- Utilizing Data in Electronic Medical Records
- Precision Health Informatics
- Image Informatics
- Advanced Health Informatics
- AI for Health Applications
- Digital Health Applications
- Capstone
- Evidence-Based Data Analysis in Population Health

05-2019 – 05-2022

Bachelor of Science: Applied & Computational Mathematics
Florida State University Honors College – Tallahassee, FL.
GPA: 3.9 – Magna Cum Laude

Courses Taken:

- Ordinary Differential Equations
- Linear Algebra
- Partial Differential Equations
- Numerical Analysis
- Math Modeling
- Complex Variables
- Discrete Mathematics
- Vector Calculus
- Calculus I, II, and III
- Economics
- Statistics
- Physics A & B

04-2019

Shanghai Jiao Tong University: Machine Learning
Shanghai, China

.

I was a member of the Machine Learning immersion program at the Shanghai Jiao Tong University in Shanghai, China. While in the program, I had the opportunity to attend global panels and engage in insightful discussions with AI experts from around the world. I also took a course in Machine Learning which provided me with a deeper understanding of the theoretical foundations and practical applications of Artificial Intelligence.

Experience

08-2023 - Current

Artificial Intelligence Researcher – NLP
Food and Drug Administration (FDA), Washington, D.C.

As an Artificial Intelligence Researcher at the FDA, I built a nucleotide transformer language model to predict mutations in virus RNA coding for glycoproteins. My model was able to simulate the evolution of the virus genome to provide insight to guide viral countermeasures towards future versions of the virus before the virus has evolved to that stage.

- Leveraged artificial intelligence techniques to enhance understanding of virus evolution and guide proactive measures.
- Collaborated with a team of researchers and scientists to develop and refine the language model.
- Employed cutting-edge AI technologies and methodologies to analyze and predict virus mutations.

07-2023

Data Scientist – Private Intelligence NLP
Stealth Startup, Washington, D.C.

As a Data Scientist, I created and deployed dynamic topic models to model the evolution of scientific research in a specific technology field for a target country.

- Deployed 3b parameter language model summarizing scientific paper abstracts.
- Created a knowledge graph of foreign institutions/scientists collaborating on technology advancements.

08-2022 – 08-2023

Data Scientist
Georgetown University, Washington, D.C.

While completing my Master's in Data Science at Georgetown University, I completed many projects, a few of which I would like to highlight below:

Generative Sequential Language Model - Virus Protein Generation
For this project I created a two-branched neural network which has shown the ability to generate a novel protein sequence with a third party (Proteinter) predicted function of virus tail fiber and cell binding when given an entirely random seed sequence and user desired molecular properties.

- Created, trained, and evaluated model architecture with ~ 4 million parameters.
- User inputs 30 amino acid seed and desired molecular properties for the sequence, the model at each time steps generates a 21-class probability distribution representing 20 amino acids and a stop flag.
- Created a script to pull and clean protein sequences of interest from the NCBI Entrez protein database.
- Tokenized all sequences and performed word2vec to generate custom amino acid embeddings.
- Developed a sub-sequence data collection script to translate each amino acid into a corresponding embedding vector with a window size of 30 and a step size of 1. The ending position +1 amino acid was then one-hot encoded for output, and at each time step, I concatenated three molecular properties for processing.

Self-Attention LSTM Separate Branch Model – Polypeptide Sequence Generation Given User Defined Secondary Structure Properties
For this project I created a neural network which takes a seed sequence and 3 desired secondary structure percentages as input and is shown to construct novel polypeptide sequences with an average absolute percent difference of under 3% for alpha helix turn and beta sheet desired properties calculated by bio seqUtils protein analysis.

- Created, trained, and evaluated model architecture with ~ 6 million parameters.
- Developed a model that constructs a transformed representation of the input sequence through three branches. A self-attention layer is applied to the sequence of embeddings sequentially and the third branch expands the desired secondary structure properties into nonlinear higher dimensions. The branches are then concatenated and used to generate a 21-class probability distribution.
- Created a generation script for inference to generate sequences.

Genetic Biomarker Discovery – Stage 2 Invasive Bladder Cancer
For this project I ran statistical analysis on gene expression data from Stage 2 invasive bladder cancer vs clinically verified normal samples to be able to predict the occurrence of Stage 2 invasive bladder cancer in patients.

- Isolated 4 significantly differentially expressed genes: CDC20, TPO2A, SRPX, IQGAP3.
- Performed Principal Component Analysis on the gene expression data, reducing the dimensions to 2 which then was passed into a K-means algorithm that aided in defining the separations between malignant and nominal samples.