Van Minh **Nguyen**

Melbourne, FL USA

vmnguyen251@gmail.com | https://n0k0m3.github.io/ | n0k0m3 | minhnguyen251

Education

Florida Tech *Melbourne, FL*

PH.D. OPERATIONS RESEARCH, 4.00 GPA *Aug 2020 - Dec 2023*

- Determine bacteria mutation rate with double stochastic branching process with random offspring.
  - Publication: Determination of Mutation Rates with Two Symmetric and Asymmetric Mutation Types. Symmetry. 2022; 14(8):1701.
- Investigating privacy-focused, longitudinal (temporal) generation of synthetic Electronic Health Records (EHR) with Differential Privacy.

M.S. OPERATIONS RESEARCH, 4.00 GPA *Aug 2018 - May 2020*

B.S. BIOCHEMISTRY (BIOLOGY EMPHASIS), 3.43 GPA *Aug 2014 - May 2018*

Work Experience

Engage-AI.org *Remote*

DATA ENGINEER - CONTRACTOR *May 2023 - Present*

- Developed and refined the Engage AI Data Platform, a cost-efficient cloud-based data management solution, leveraging Cloudflare R2, DuckDB, and Apache Superset for object storage, data processing, and interactive visualization.
- Collaborated with data analysts to understand their data requirements, refining and optimizing the platform based on feedback.

Truveta *Seattle, WA*

RESEARCH INTERN *Jan 2022 - May 2022*

- Developed and deployed scalable NER pipelines for clinical notes information extraction and de-identification using SparkNLP and PyTorch, saving $2M annually and reducing operating costs by 75%.
- Conducted threat modeling using OWASP Threat Dragon and recommended mitigation strategies for pipeline deployment.
- Created a clinical notes annotation tool prototype based on Label Studio and INCEpTION for internal use.

GRADUATE INTERN *May 2021 - Aug 2021*

- Built an ETL pipeline for measuring Truveta Health Data Model quality using Spark and Azure Pipelines, leading to a Microsoft partnership and integration into Truveta Studio.
- Designed a synthetic patient data model for stress-testing and bottleneck identification in the ETL process, generating millions of records in 1 hour.
- Developed an annotation recommender system for medical concept normalization, reducing annotators' workload by 80%.

Florida Tech *Melbourne, FL*

SUICIDE PREVENTION RESEARCH - DEPARTMENT OF COMPUTER ENGINEERING AND SCIENCES *Jan 2023 - Present*

- Enhanced data scraping pipelines for Twitter and Reddit, reducing ingestion time by 60x.
- Developed a prototype model for predicting suicidal tendencies from social media posts using NLP features
- Analyzed monthly word statistics and word clouds of suicidal posts over a 5-year period

MLOPS TECHNICIAN - NEURAL TRANSMISSIONS LAB *Jan 2022 - Present*

- Managed on-premise server deployment and maintenance for research lab, utilizing Kubernetes on Ubuntu Server.
- Set up multi-user research environments with GPU support and role-based access control using JupyterHub, Kubernetes, and Keycloak.
- Secured deployments using HTTPS, DNS configuration, short-lived SSH, and VNC over HTTPS.

TEACHING ASSISTANT - DEPARTMENT OF MATHEMATICAL SCIENCES *Aug 2018 - Present*

- Provided instruction, grading, and tutoring for courses including Calculus, Stochastic Modeling, Neural Networks, and Machine Learning

Projects

3D Reconstruction of satellite using Dynamic Neural Radiance Fields *[Publication]*

Applied *instant-ngp* and *D-NeRF* for efficient 3D model reconstruction of satellite from a single view camera video of the real satellite object, enabling improved space debris removal and on-orbit servicing.

Sentiment Analysis on MyAnimeList User Ratings *[Project link]*

Implemented a transformer module and action-memory within Soft Actor-Critic and TD3 architectures, enabling the agent to "remember" its previous actions and effectively predict the next ones. Tested on *highway-env*, an OpenAI Gym environment for autonomous driving decision-making tasks

GPU-supported PySpark Notebook with DeltaLake *[Repo link]*

Docker container for data analysis with Jupyter notebook server, RAPIDS AI, PySpark for GPU-accelerated, distributed and scalable ETL, aiming for feature parity with Databricks - a popular cloud-based data analytics platform.

Skills

Programming & Deep Learning	Python, R, C#, TensorFlow, PyTorch, ONNX
Big Data & Cloud Platforms	Spark/PySpark, Hadoop Streaming, Microsoft Azure, Databricks, Kubeflow, MLFlow
Deployment & Databases	Docker, Kubernetes, Azure Pipelines, Cloudflare Zero Trust, SQL (MariaDB), NoSQL (MongoDB, Redis)
Analytics & Modeling	Data Mining, Data Processing & Analysis, Statistical Modeling, Stochastic Modeling, Mathematical Analysis