

# Van Minh Nguyen

✉ vmnguyen251@gmail.com | ☎ (321)-368-9310 | 📍 Melbourne, FL USA  
🔗 Portfolio | 🔗 LinkedIn Profile

## Education

### Florida Tech

PH.D. OPERATIONS RESEARCH

- Researched on Double Stochastic Processes (Branching Process with random offspring).
- Researching Electronics Health Record embedding for Machine Learning/Deep Learning.

Melbourne, FL

Aug 2020 - Present

### Florida Tech

M.S. OPERATIONS RESEARCH, GPA: 4.0

Melbourne, FL

Aug 2018 - May 2020

### Florida Tech

B.S. BIOCHEMISTRY (BIOLOGY EMPHASIS), GPA: 3.43

Melbourne, FL

Aug 2014 - May 2018

## Work Experience

### Florida Tech

RESEARCH LAB - DEVOPS ENGINEER

- Deploy and maintain (bare-metal) servers for deep learning lab lead by Dr. Ryan T. White
- Deploy multi-user environments with GPU support for deep learning researches.
- Deploy Zero Trust Architecture for granting role-based access to servers (RBAC). Authentication backed with Keycloak and SAML, authorization backed with Cloudflare Access and Google OAuth.
- Secure deployments with HTTPS for web server, short-lived SSH certs for remote access, and VNC over HTTPS for remote control.

Melbourne, FL

Jan 2022 - Present

### Truveta

GRADUATE INTERN

- Cost-effective and scalable SOTA Named Entity Recognition (NER) pipelines for clinical notes information extraction and de-identification. Saved \$2 million in annual cost compared to using existing third party solutions, and optimized to reduce 75% operating cost overhead compared with original baseline.
- Prototyped a clinical notes annotation tool for internal annotator team.
- Threat modeling and mitigation for deployment of mentioned tools, including common attack vectors on platform and PHI leakage threats.

Seattle, WA

Jan 2022 - May 2022

### Truveta

GRADUATE INTERN

- Developed an ETL pipeline to measure data quality of Truveta Health Data Model (THDM), garnering trusts from health providers. Contributed to new partnerships with multiple after Private Review in Aug 2021, and with Microsoft in Sept 2021.
- Designed a Monte Carlo sampling model to generate synthetic patient data for stress-testing and identify bottleneck in ETL process.
- Developed an annotation recommender system for medical concept normalization. Reduced internal annotators' workload by 80%

Seattle, WA

May 2021 - Aug 2021

## Skills

### Software & Tools

Python, Spark/PySpark, Hadoop Streaming, MySQL, Redis, TensorFlow, PyTorch, ONNX

### Operating Systems

Windows, Ubuntu Server, Arch Linux, macOS

### Cloud Infrastructure

Microsoft Azure, Amazon AWS

### Other Skills

Data Processing & Analysis, Stochastic Modeling, Reverse Engineering, Penetration Testing

## Projects

### Sentiment Analysis on MyAnimeList User Ratings

DATA MINE, ETL TO EXTRACT USER REVIEW FROM MAL. PREDICT USER RATING BASED ON DATA-MINED REVIEW TEXT WITH SENTIMENT ANALYSIS. 94% TESTING ACCURACY WITH ENSEMBLE MODEL

Redis, Docker, TensorFlow

[Project link]

### GPU-supported PySpark Notebook with DeltaLake

DATA ANALYSIS CONTAINER WITH JUPYTER NOTEBOOK SERVER WITH GPU-RAPIDS AI FOR GPU-ACCELERATED DATA TRANSFORMATION, SPARK/PYSPARK FOR DISTRIBUTED AND SCALABLE ETL, WITH ADDITION OF ALL COMMON DATA SCIENCE LIBRARIES. AIMING FOR FEATURE PARITY WITH DATABRICKS WHILE STAYING MINIMAL.

Docker, Compose, GitHub CD/CI

[Repo link]

### Mobile Game Assets Decryption and Datamining tool

REVERSE ENGINEERED GAME FOR ASSETS DECRYPTION FOR GAME "DATE A LIVE: SPIRIT PLEDGE". SCHEDULED PIPELINE FOR CONTINUOUS UPDATE, DECRYPT, AND DATAMINE DATA.

Frida, Ghidra, GitHub CD/CI

[Repo link],[Data link]