Linh D. Shinguyen, M.S.

Lnguyen1021@gmail.com | www.linkedin.com/in/shinguyenlinh

EXECUTIVE SUMMARY

Inquisitive data scientist with 8 years of experience working with various data types ranging from FHIR data, CT imaging data, and textual data, and has proven expertise collaborating with and leading cross functional teams; working with complex health data and finding data-driven solutions; and communicating data driven story telling.

CORE COMPETENCIES

Python | R | SQL| Git | Snowflake | FHIR
Data visualization & analysis | Data storytelling
Machine learning & Deep learning | Computer vision & NLP

PROFESSIONAL EXPERIENCE

Data Scientist

Dec 2023 - Present

Health Resources and Services Administration, Rockville, MD
Bureau of Primary Health Care, Office of Strategic Business Operations

UDS+ Data Modernization Technical Lead

- Served as the data lake house team technical lead in support of UDS+ Modernization efforts, enabling data-driven health outcome improvements for Health Center program awardees nationwide.
- Co-led project initiatives by facilitating technical discussions between development teams and senior leadership, ensuring alignment and timely achievement of key contracting milestones.
- Collaborated with the contracting team to optimize data architecture for ingesting UDS+ FHIR data into Snowflake, resulting in a scalable solution capable of processing millions of healthcare records efficiently.
- Co-managed the end-to-end ingestion process of preliminary UDS+ data from 80 volunteer health center awardees, successfully loading over 30 million records into Snowflake for initial analysis.

UDS+ Personal Health Information (PHI) Detection Project Lead

- Led a team of 2 junior data scientists to design and implement a solution for detecting potential Protected Health Information (PHI) within UDS+ FHIR data, ensuring adherence to Safe Harbor Guidelines.
- Architected and developed a scalable PHI detection framework using Python and Snowflake, integrating NER modeling with custom regular expressions to efficiently identify sensitive patient information.
- Successfully identified and confirmed over 2 million records containing PHI, as defined by Safe Harbor Guidelines and additional specifications from the UDS+ FHIR Implementation Guide, significantly enhancing data privacy and compliance.

UDS+ & Data Lake House Liaison for Data Division Business Intelligence (BI) Team

- Delivered expert technical support to 4 BI team members across multiple projects, ensuring seamless access to UDS+ data and data lake house resources while troubleshooting complex integration challenges.
- Engineered and implemented the Salesforce ticketing data pipeline into Snowflake, enabling the BI team to develop comprehensive service metrics dashboards in Tableau.
- Strategically curated and optimized UDS+ tables in Snowflake that powered 3 executive summary dashboards, providing critical program insights to division, office, and bureau leadership.
- Conducted targeted ad hoc data analysis for senior leadership, transforming raw UDS+ submissions into actionable intelligence that informed strategic decisions about Health Center program operations.

UDS+ Electronic Clinical Quality Measure (eCQM) Project Lead

- Directed a team comprising 2 junior data scientists and contracting personnel to develop scalable solutions in Snowflake for calculating 6 electronic Clinical Quality Measures (eCQMs 122, 124, 125, 130, 165, 349) according to guidelines from the Centers for Medicare & Medicaid Services (CMS).
- Partnered with the contracting team to architect and execute dynamic tables in Snowflake that automatically identified patient-level UDS+ data meeting eCQM inclusion/exclusion criteria and transformed individual data points into comprehensive health center-level performance measures.
- Generated and distributed technical reports detailing performance on 6 key eCQMs to all health centers that voluntarily submitted UDS+ data, providing valuable insights into quality measure adherence.

Data Science Fellow

Sep 2021 – Nov 2023

National Institute for Allergy and Infectious Diseases, Rockville, MD
Emerging Leaders in Data Science Fellowship

Module Development and Comparison of nnUNet and Deep learning Medical Image Segmentation (DMIST) models for Lung Lesion Segmentation

Integrated Research Facility (IRF) at Fort Detrick: AI & Imaging Team

- Engineered and implemented a comprehensive model comparison module in collaboration with IRF data scientists, enabling quantitative evaluation and systematically monitoring of deep learning model segmentation performance across multiple infection types and organ systems.
- Leveraged Git and GitHub for version control and streamlined collaboration throughout the development lifecycle of the model comparison module.
- Redesigned researcher and radiologist workflow by consolidating submission requirements from 4 files to a single file, reducing process complexity and accelerating model comparison initiation by 75%.
- Authored a script to automate the collation of CT prediction files organized into a consistent directory structure, enabling efficient visual comparison and analysis of segmentation results.
- Implemented and integrated a benchmark model (nnUNET) to establish baseline lung lesion prediction performance for models developed via the DMIST pipeline, enabling objective performance assessment.

Deep Learning CT Segmentation Model Comparisons: UNET, UNETR, Swin-UNETR
IRF at Fort Detrick: AI & Imaging Team

- Built and evaluated 3 deep learning architectures to perform image segmentation analysis on 94 non-human primate lung CT scans infected with SARS-COV2 to support critical infectious disease research.
- Applied the Swin-UNETR deep learning algorithm for SARS-COV-2 lung lesion segmentation in CT scans, achieving 61% accuracy despite the challenging nature of pathological tissue identification.
- Trained a UNETR deep learning model for human lung segmentation in CT images, attaining an accuracy of 98%.
- Improved Swin-UNETR lung lesion segmentation model performance on human CT scans by 10% through the implementation and evaluation of transfer learning techniques, comparing pretrained and non-pretrained models.
- Collaborated directly with junior radiologists to gather precise ground truth annotations, leading to enhanced CT segmentation model accuracy and reliability through high-quality training data.

Multimodal Analysis of Multiple Sclerosis
Division of Intramural Research (DIR): Laboratory of Clinical Immunology and Microbiology (LCIM)

- Conducted exploratory data analysis on multi-modal omics data (patient demographics, immunophenotyping, bulk RNA sequencing, Somamer scan) from cerebrospinal fluid to characterize individuals with multiple sclerosis (MS).
- Applied dimension reduction and clustering techniques (UMAP, t-SNE, PCA) to try and identify key features distinguishing 41 MS patients from 40 controls, informing the understanding of disease differentiation.

Prognostic modeling of post-acute sequelae of SARS-Cov-2 (PASC)
Division of Clinical Research (DCR): Bioinformatics Research Branch (BRB)

- Assisted in univariate analysis to characterize differences among individuals who developed post-acute sequalae of SARS-COV-2 (PASC) and those who did not among persons infected with COVID-19.
- Developed and interpreted random forest and elastic net models utilizing Shapley values to explore predictive risk factors and diagnostic indicators associated with PASC.

Estimation of data generation from grants utilizing research performance progress reports (RPPRs): A Pilot Natural Language Processing (NLP) Project
Division of Allergy, Immunology, and Transplantation (DAIT)

- Collected, wrangled, and meticulously cleaned 900 unstructured NIAID RPPR documents to establish a gold standard corpus, enabling the development of robust Natural Language Processing (NLP) models.
- Extracted "data generated" entities from 250 grant documents to curate a high-quality training dataset for NLP applications.
- Developed a Named Entity Recognition (NER) model to automatically classify RPPR documents based on "data generated" criteria, providing insights into the NIAID data landscape.

Database and Data Science Intern

Feb 2020 – Aug 2020

Tractor Supply Company, Nashville, TN

- Redesigned and optimized 5 complex Tableau Desktop dashboards into 7 mobile-responsive analytics solutions, enabling C-suite executives to access critical business metrics seamlessly across mobile devices.
- Improved marketing department analysts' workflows by optimizing 5 key performance metric calculations, resulting in more efficient data processing and analysis.
- Pioneered an innovative data refresh protocol using Tableau Prep that delivered comprehensive cost-benefit analysis for new marketing software acquisition, directly informing strategic technology investment decisions.
- Analyzed 3 million Tennessee store transactions using R to conduct market basket analysis, uncovering critical association rules and customer purchasing behavior patterns during COVID-19.

Laboratory Safety Fellow

Jan 2018 – Sep 2018

Centers for Disease Control and Prevention, Atlanta, GA
Centers for Surveillance, Epidemiology, and Laboratory Services

- Executed comprehensive data cleaning for laboratory proficiency testing data spanning 94 distinct analytes, ensuring data integrity.
- Collaborated with the branch statistician to analyze historical proficiency testing data, calculating failure rates to identify

and monitor performance trends over a 22-year period (1994-2016).
- Applied knowledge of CLIA regulations to systematically identify laboratories exhibiting multiple high-failure events, subsequently generating concise summary reports for branch leadership review and action.

Health Promotion Intern
Aug 2017 – Jan 2018

Centers for Disease Control and Prevention, Atlanta, GA
Centers for Surveillance, Epidemiology, and Laboratory Services

- Partnered with a health scientist to successfully collect comprehensive proficiency testing data from laboratories across the United States.
- Conducted a needs assessment within the office to identify opportunities and strategies for promoting a healthy work-life balance among employees.
- Collaborated with building management and gym partners to drive engagement in wellness initiatives, designing and distributing flyers to improve communication and promote health-related events, which aimed to increase gym and fitness class usage.

TECHNICAL PROJECTS

Improving Customer Email Conversion Rate Mar 2021 – May 2021

Vanderbilt Data Science Institute industry partner

- Directed and mentored a team of 4 first-year graduate students in developing machine learning solutions that successfully increased client email conversion rates, establishing clear objectives and providing technical guidance throughout the project lifecycle.
- Prepared 43 GB of client data for modeling initiatives through comprehensive data wrangling, cleaning, and feature engineering, leveraging parallelization techniques on Vanderbilt's high-performance computing cluster to handle large-scale data.
- Segmented customers into distinct groups by applying principal component analysis (PCA) for dimensionality reduction and K-means clustering, identifying cohorts with similar purchasing behaviors.

Customer Base Exploratory Data Analysis Jan 2021 – Mar 2021

Vanderbilt Data Science Institute industry partner

- Led a team of 4 first-year graduate students in conducting an exploratory data analysis project utilizing customer data to deliver actionable insights on digital product performance to an industry client.
- Designed and implemented an interactive Tableau dashboard that enabled the client to visualize national market trends and track product performance metrics over time, supporting data-driven strategic decision-making.
- Identified that 10% of businesses were delinquent on payments through data analysis, enabling the industry partner to initiate targeted payment collection efforts and recover potential revenue

Sub-Volcanic Discolored Water Pixel Classification Jan 2021 – May 2021

Vanderbilt Department of Earth and Environmental Sciences

- Engineered a sophisticated random forest model in Google Earth Engine leveraging SENTINEL-II satellite imagery to detect discolored water pixels resulting from subvolcanic eruptions, enabling remote monitoring of underwater geological activity.
- Enhanced model performance by creating custom satellite spectral band features and implementing hyperparameter optimization, significantly improving classification accuracy from 73.2% to 84.1%.
- Discovered previously unidentified regions of subvolcanic eruption activity through application of the optimized machine learning model, contributing valuable data to geological research and hazard monitoring efforts.

Optimization of Onsite and Remote Nurses for Physician Offices in Texas Jan 2021 – Mar 2021

Vanderbilt Data Science Institute industry partner

- Collaborated with 3 data science graduate students to develop an interactive Streamlit dashboard that utilized hierarchical modeling and regression analysis to optimize the staffing mix of onsite and remote nurses required for physician office operations across various Texas locations.
- Partnered with a health consulting agency's data analytics team to analyze the time allocation of essential operational tasks for both onsite and remote nurses, subsequently developing a regression model to accurately estimate task time distributions.
- Leveraged the time distribution regression model to construct a financial model that estimated the overall operational costs for both onsite and remote nursing staff, providing stakeholders with clear insights into the costs and benefits of different staffing options.

Validation of Cervical Pre-Cancer Billing Claims Data June 2019 – Feb 2021

Vanderbilt University Medical Center

- Partnered with a PhD researcher to process and standardize 8,500 ICD-9 and ICD-10 billing records spanning 2008-2017 for Davidson County, TN, implementing robust Python data cleaning protocols that ensured analytical integrity.
- Developed and validated a random forest classification model (scikit-learn) that accurately identified 89% of cervical precancer events within billing records, with model performance confirmed against biopsy data.
- Co-authored research findings stemming from this analysis, resulting in a publication in the peer-reviewed journal JNCI Cancer Spectrum. https://doi.org/10.1093/jncics/pkaa112.

Catastrophic Loss Prediction Jan 2020 – May 2020

Vanderbilt Data Science Institute industry partner

- Collaborated with industry partner and classmates to develop predictive models for catastrophic losses in high-risk regions, focusing on rare severe weather events across a critical 10-year period (2008-2018).
- Delivered comprehensive financial impact analysis of severe weather events through advanced exploratory data techniques, providing the industry partner with actionable insights for risk management and resource allocation.
- Identified and recommended valuable secondary data sources that significantly informed and improved subsequent iterations of the machine learning prediction models.

Latin American Survey Interview Adherence Aug 2019 – Dec 2019

Vanderbilt Data Science Institute industry partner

- Analyzed survey data for 60 interviewers in collaboration with Latin American stakeholders, successfully identifying abnormal survey behaviors in 10% of interviewers, which led to targeted retraining to ensure adherence to protocol.
- Developed a KNN clustering model in Python, integrating one year of survey, geospatial, and time series data to generate comprehensive surveyor profiles, enabling the client to gain deeper insights into surveyor behavior patterns.

EDUCATION

M.S. in Data Science
Vanderbilt University, Nashville, TN
Capstone: Classification of Sub-Volcanic Discolored Water Pixels Using Satellite Data; Advisor: Tushar Mittal, Ph.D., Kristen Fauria, Ph.D.

B.S. in Health Promotion and Behavior
University of Georgia, Athens, GA

CERTIFICATES

Certificate in Data Analytics and Visualization
Georgia Institute of Technology, Atlanta, GA