Chau Nguyen

Washington, DC Metro Area | Email: cn490@georgetown.edu
Github: chav-ngvyen | LinkedIn: chavngvyen | Website: https://www.chav-ngvyen.com

Professional Experience

Massive Data Institute — September 2021 - Present

MDI Scholar (with Dr. Toni-Lee Sangastiano - Department of Art and Art History, Georgetown Univeristy)

- Utilizes transfer learning to finetune pre-trained image classification neural nets to find banners in paintings from museum archives

MDI Scholar (with Office of Evaluation Sciences - General Services Administration)

- Prototyped tool to qualify low-income renters for self-attestation for ERA in partnership with US Treasury and Digital Services.
- Analyzed spending patterns of recipients of $350 billions in fundings from Coronavirus State and Local Fiscal Recovery Funds.

Fraym — June 2021 - September 2021

Data Science Intern

- Streamlined crash detection by grouping past error logs using K-means clustering algorithm to find similarities in documents.
- Improved core product by implementing tryCatch clauses around gaps in codebase to prevent similar crashes in the future.
- Developed process to match job IDs with AWS logStreams using boto3, SQLALchemy and pandas; presented in DS study group.
- Presented results from analysis of job failures and highlighted actionable insights to improve codebase in company-wide meeting.
- Created API endpoints to query geometries and raster data from internal PostgreSQL database; used fastAPI and GeoAlchemy2.

International Monetary Fund — December 2016 - August 2020

Research Analyst

- Traveled to Samoa and Tuvalu, engaged in policy discussions with countries' leaders, contributed to drafting of Staff Reports.
- Used SQL to query data of 275,000 firms in 75 countries and compiled results for Corporate Vulnerability Utility index.
- Identified and addressed bugs causing CVU scripts to crash, created detailed documentation for use as reference in the future.
- Created eight interactive Tableau dashboards to visualize and analyze data of over 70 countries in China's Belt and Road Initiative.
- Managed four countries' macroeconomic databases, performed statistical checks for data consistency across all economic sectors.
- Contributed more than 200 descriptive charts and 60 tables published in over ten official IMF publications and five journal articles.

Education

Georgetown University | M.S. Data Science for Public Policy — Expected Graduation: May 2022

Relevant courses: Supervised ML (naive Bayes, LDA, decision tree, random forest, ensembles, SVM); Unsupervised ML (PCA, clustering, topic modeling, sentiment analysis); Advanced Statistics (A/B testing, causal inference, IVs, diff-in-diff, regression discontinuity); Neural Nets & Deep Learning (computer vision, time series, transformers, auto-encoders, reinforcement learning); Massive Data with SQL Specialization (relational databases, cloud computing, Hadoop, Spark, AWS); Text as Data (computational linguistics, NLP pipelines).

University of California, Berkeley | B.A. Economics — Graduation: May 2016

Publication: Raymond J. Hawkins and Chau N. Nguyen (2017). Macroeconomic dynamics and the IS puzzle. Economics Discussion Papers, No 2017-20, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2017-20

Projects

- Scrollable Interactive Tutorial on Density-based Clustering Algorithms (D3.js, JavaScript, Python, HTML, SVG, CSS)
  Online "scrollyboard" with self-updating graphics to explain technical differences between DBSCAN and HDBSCAN - unsupervised algorithms used to cluster unlabeled data. Used Python for clustering simulations and D3.js for embedded interactive visualizations.
- Doyle, Christie, or LeBlanc? - A Deep Learning Approach to NLP and Authorship Detection (TensorFlow, NLTK, shell script)
  Created NLP pipeline to preprocess novels. Trained RNNs with embedding layer to learn penmanship, used authors' writings in other domains to test for accuracy. Wrote Python script to tune hyperparameters, bash script to launch process from command line.
- Hyper-parameter Tuning Convolutional Neural Net Layers to Predict Forest Fires (TensorFlow, keras, sklearn)
  Performed feature engineering and tuned layers of a convolutional neural network to accurately predict the spread of forest fires.
- Is it the Crowd or the Stadium? - An ML Approach to Home-field Advantage in the NFL (scikit-learn, pandas, beautifulsoup)
  Wrote Python script to web-scrape, clean, pre-process game-level information for 20 NFL seasons and train classification model. Developed new approach to treat games as dyads, quantify home-field advantage using in-dyad differences and predict outcomes.
- Colorblind tutorial (OpenCV, Jupyter Notebook, matplotlib, data visualization, colorblind, accessibility)
  Tutorial on how to set up Python modules to check data visualizations for colorblindness accessibility within any Jupyter Notebooks.

Technical Skills

- **Programming languages:** Python (preferred), R, Stata.
- **Python modules:** General: Jupyter notebooks, scikit-learn, pandas, NumPy, fastAPI. | AWS: boto3 | NLP: SpaCy, NLTK, gensim. Deep Learning: keras, TensorFlow, PyTorch | SQL: MySQL, SQLAlchemy | Webscrapping: requests, beautifulsoup, selenium.
- **R packages:** Data wrangling & analysis (baseR, tidyverse), geospatial data (raster, sp, sf), statistical training/ ML (caret).
- **Data visualization & storytelling:** Python: matplotlib, seaborn, Altair, bokeh, plotly | R: ggplot2, gnuplot | Tableau | D3.js.
- **Project management & collaboration:** Comfortable with Agile development practices, Jira, Git workflow and version control.
- **Other:** Experience with LaTeX, Linux (Ubuntu) environment, Unix command line, shell scripting, ArcGIS, shape files, geometries.
- **Languages:** English (fluent), Vietnamese (native).