ANDY WANG

awang124@jh.edu • (443) 465-1358 • linkedin.com/in/awang124 • awang124.github.io

EDUCATION

Johns Hopkins University — Baltimore, MD
Bachelor of Science in Applied Mathematics & Statistics — Expected May 2027
Additional Major in Computer Science — Cumulative GPA: 3.89 / 4.0

Coursework: Machine Learning, Deep Learning, Bayesian Statistics, Mathematical Image Analysis, Time Series Analysis, Statistical Learning Theory, Stochastic Processes, Probability, Optimization

SKILLS

Programming: Python (PyTorch, TensorFlow / Keras, Scikit-Learn, Pandas, NumPy), R, SQL, MATLAB
ML: Deep Learning (CNNs, RNNs, Transformers), Unsupervised Learning (Clustering, Anomaly Detection)
Data & Statistics: Time Series Forecasting (ARIMA, LSTM), Spectral Analysis, Statistical Hypothesis Testing

RESEARCH EXPERIENCE

Mario Micheli's Research Group
Johns Hopkins Department of Applied Mathematics & Statistics | Baltimore, MD — May 2025 – Present

- Developed a novel probabilistic approach to optical flow estimation, to exploit spatial correlation, capture inherent motion uncertainty, and provide confidence measures for safety-critical applications.
- Formulated Bayesian estimation of true optical flow fields through spatial Gaussian process regression (GPR) conditioned on noisy observations obtained via CNN/RNN-based methods (FlowNet, RAFT).
- Derived full posterior distributions for latent flow fields, and optimized GP kernel hyperparameters through gradient-based likelihood maximization, implemented with PyTorch.
- Demonstrated GPR effectively denoises deterministic deep learning methods and corrects erroneous estimates along textureless edges on real and synthetic image sequences (e.g. Yosemite).
- Discovered posterior covariance eigenvalues align with orthogonal image gradients and quantify optical flow difficulty, correlating with regions of occlusion, limited texture, and rapid motion.
- Improved performance by modifying GPR to model noise using data-dependent variance.
- Extending existing work, using semantic segmentation to capture different motions in localized regions, and applying spatiotemporal (3-dimensional) Gaussian processes on flow field sequences.

Laboratory of Computational Intensive Care Medicine
Johns Hopkins School of Medicine | Baltimore, MD — Sep 2024 – May 2025

- Investigated correlation between Heart Rate Variability (HRV) and clinical outcomes in Traumatic Brain Injury (TBI) patients for use in phenotype discovery, outcome prediction, and personalized treatment.
- Queried 398 million MIMIC-III records using PostgreSQL, identifying 347 TBI patients and extracting clinical data (vitals, demographics, treatment outcomes) and ECG waveforms.
- Engineered Python pipeline to clean ECGs (outlier removal, missing value interpolation), find optimal signal windows via Fourier analysis, and extract 31 HRV features, including statistical time-domain metrics, frequency-domain power bands, and nonlinear measurements.
- Implemented novel Transformer-based model using TensorFlow to generate temporal data embeddings.
- Designed an iterative unsupervised learning algorithm, combining Transformer with K-Means clustering, to assign patients to one of three distinct HRV phenotypes.
- Discovered clusters exhibit significant differences in GCS score / ICU Length of Stay, via Chi-Square / Kruskal-Wallis testing ($p < 0.05$), enabling future research into personalized TBI treatment strategy.

John Edison's Research Group
Johns Hopkins Whiting School of Engineering | Baltimore, MD May 2024 – August 2024

- Utilized C++ and Arduino libraries to transmit/receive timestamps between DWM1000 UWB transceivers in order to measure distances between these devices with accuracy ±10cm.
- Designed and implemented a triangulation algorithm to determine the coordinates of a moving device relative to four fixed devices using distance measurements.
- Used Python to receive positional data and timestamps from the fixed devices via serial communication, log data in CSV format, and visualize of positions over time.
- Collaborated with JHU Makerspace to create cases for the devices wearable by athletes to generate insights on various sports strategies using players' positional data.

PROJECTS

Cardiac Arrhythmia Diagnosis Aug 2024

- Applied Python and WFDB to read 24 hours of ECG signals from the MIT-BIH Arrhythmia database.
- Utilized BioSPPy to split signals into 116942 individual heartbeats and NumPy to engineer 360 features from normalized signal values.
- Performed PCA on dataset, reducing training time by 77% while preserving 95% variance.
- Ensembled and fine-tuned Gradient Boost, Random Forest, and Support Vector classifiers to create a final model yielding 0.91 F1-score in predicting heartbeat abnormality.
- Employed K-Means clustering to determine cardiac arrhythmia based on model probability output.

Apple Stock Price Prediction Jul 2024

- Performed time series analysis on 2022-23 Apple stock prices using Pandas and Matplotlib, examining stationarity and autocorrelation and applying trend decomposition to identify patterns and noise.
- Trained LSTM, ARIMA, and Prophet models to forecast 2024 prices, yielding a lowest RMSE of 2.592 dollars per share compared to ground truth.
- Deployed a Shiny web application in R to forecast and visualize any company's stock using the LSTM.

Apache Spam Email Detection Jul 2024

- Utilized Python, HTML, and regular expressions to parse and clean 6047 emails from the Apache SpamAssassin database, lemmatizing words and removing hyperlinks, escape characters, etc.
- Implemented word-count vectorization algorithm from scratch and built custom Scikit-Learn data transformation pipeline to process emails into a sparse feature matrix.
- Trained and fine-tuned Random Forest, Gradient Boost, and Logistic Regression classifiers, yielding 96% recall and 94% accuracy in flagging spam.

California House Price Prediction Jun 2024

- Conducted exploratory data analysis on 20640 California census districts with Pandas and Matplotlib, considering feature distributions and correlation to identify best predictors of median house price.
- Engineered geospatial features via K-Means clustering and the RBF kernel to reduce RMSE by 30%.
- Fitted and fine-tuned Random Forest, Ridge, and Lasso regressors, yielding a lowest RMSE of $43K.
- Compared each feature's GINI impurity decrease to determine with statistical significance that median income is the best predictor of median house price.

Northwind Sales Analysis Jun 2024

- Analyzed 31700 entries from 6 tables of the Northwind Sales PostgreSQL database via R library DBI.
- Wrote complex SQL queries involving window functions, CTEs, and joins to calculate cumulative revenue, compare employee performance, identify high-value customers and best-selling products, etc.
- Created stored procedures and views to calculate total costs of orders, return products whose sales exceed a given percentile, query information schemata, etc., in order to answer difficult questions.
- Elegantly visualized engineered data using ggplot2 to communicate insights to a business audience.

TEACHING EXPERIENCE

Gateway Python Teaching Assistant
Johns Hopkins Department of Computer Science | Baltimore, MD Sep 2024 – May 2025

- Held weekly office hours and project workshops for 430 students, fielding questions, reinforcing programming techniques, and ensuring students are equipped to succeed.
- Graded assignments and provided personalized feedback for 68 students.

SAT Mathematics Tutor
Capital Educators | Rockville, MD Sep 2023 – May 2024

- Tutored 81 high school students weekly in advanced SAT math, and reinforced studying / test-taking strategies, yielding an average increase of 100 points in students' SAT math scores.
- Collaborated with tutors and supervisors in planning and composing lessons and diagnostic tests.