Zhihao (Johnson) **Du**

zhihao617@berkeley.edu | +1 (510) 833-4417

Links

Github
github.com/JohnsonJDDJ

Personal Website
zhihao.myxd.place
(for info on all projects)

Skills

Languages
Python (since 2017)
SQL (since 2019)
Java (since 2020)
R (since 2020)
C (since 2021)
HTML, CSS (since 2017)

Coursework
Neural Networks
Statistical Learning
Service Operations Design
Database Systems
Reproducible Data Science
General Linear Models
Data Structures
Machine Structures
Linear Algebra
Probability Theory

Tools
Git (since 2020)
Jupyter (since 2020)
MS Azure (since 2022)
DBeaver (Summer 2021)
Command Line
Microsoft Office

Education

Note: Applying for master's program with expected graduation date: 05/2024.

University of California, Berkeley
B.A. Statistics, B.A. Computer Science
GPA: 3.75/4 | Expected graduation date: 05/2023

Experience

ETL Engineer Intern
DataCVG Shanghai
Shanghai, China | 05/2021 - 08/2021
Worked under the database services team at the Business Intelligence software company. Engineered directly on the client "FosunPharma"'s pharmaceutical database system by designing and implementing extract-transform-load (ETL) pipelines on relational data:

- Collectively designed target relational database architecture with ER diagrams.
- Individually wrote queries using DBeaver for the ETL pipeline that combined two source databases.
- Debugged architecture failures through long diagnostic process and extensive communication with PM and client representatives.

Projects

Project AEI (Artificial Emotional Intelligence)
Real-time audio emotion classification using neural networks | 01/2022 - Now
Advised by Prof. Dacher Keltner, assembled a police aggression discernment and early warning system powered by a parallel CNN Transformer neural network using pytorch. The system is capable to classify emotions from streaming real-life audio speech data:

- Trained the parallel network using large emotional databases (RAVDESS, SAVEE) through MS Azure ML cloud compute platform.
- Preprocessed streaming audio data into melspectrograms and augmented with Gaussian white noise using pyaudio, librosa and pytorch.

zilean
Python package for data mining and engineering pipelines | 05/2022 - Now
Advised by Prof. Fernando Pérez, developed python package "zilean" that bridges the Riot Games API with traditional python ML APIs through implementing customizable data crawler objects and domain specific data manipulator objects.

- Designed highly customizable and fully automatic data crawler using rate limiting API request algorithms through requests (python package).
- Constructed data mining algorithms incorporating domain specific knowledge for large semi-structured data from the API using native python libraries.
- Capable of creating data pipelines producing multidimensional structured data ready for ML or DL tasks.
- Utilizing this tool, predicted League of Legends match outcomes with game statistics before the 16 minute mark by tuning a random forest classifier and a XGBoost classifier through sklearn and hyperopt.

HOYO Lab
Data Modelling in R | 03/2021 - 08/2021
Predicted Genshin Impact's damage mechanism by building a basic linear model in R with feature mapping using original data collected through in-game simulations. Displayed the result and hosted an interactive model using R ShinyApp online. Used the model to wrote game tutorials receiving 400,000+ views.