

# Yanna Shen

📞 206-290-3069 ✉ [yanna.cshen@gmail.com](mailto:yanna.cshen@gmail.com) 💼 [www.linkedin.com/in/chengyang-shen](http://www.linkedin.com/in/chengyang-shen) 🌐 <https://yannacs.github.io/>

## EDUCATION

**Master of Science in Data Analytics Engineering (GPA: 3.96 / 4.00)**

Sept. 2023 - May 2025

*Northeastern University, Seattle, WA*

- **Relevant Coursework:** Data Analytics Engineering, Deterministic Operations Research, Data Mining, Data Management and Database Design, Visualization for Analytics, Cloud Computing, Data Structure, Algorithms, Statistics

**Bachelor of Engineering in Computer Science and Technology (Honors)**

Sept. 2017 - Aug. 2021

*Xiamen University, Malaysia*

## SKILLS

**Certificate:** Tableau Certified Data Analyst, AI Literacy, Graduate Leadership, DA on AWS, Applied AI  
**Programming Languages:** Python, SQL(MySQL, NoSQL, MongoDB), C, C++, JAVA, JSP, MATLAB, Shell, MIPS  
**Technologies:** Data Cleaning, Data Analytics, Data Mining, Data Processing, Data Pipeline, ETL, Data Ingestion, Data Warehouse, Data Storage, Database Design, Database Normalization, Pandas, Numpy, Scikit-Learn, Matplotlib, Time Series Analysis, Text Analysis, Sentiment Analysis, Clustering(K-Means, KNN), Feature Selection, Feature Engineering, Principal Component Analysis(PCA), t-SNE, TensorFlow, Anaconda, pip, Machine Learning(ML), CV, NLP, Tableau, Streamlit, UML, MS Office, Excel, AWS, GenAI, BigQuery, Spark, Kafka, html, Linear Optimization

## EXPERIENCE

**Data Engineer Intern | Uplift Northwest, Seattle (ETL, ML, Tableau)**

June - Sept. 2024

- Conducted comprehensive database assessment, identifying normalization issues and implementing schema optimizations
- Designed and deployed ETL pipelines to integrate disparate data sources, creating a unified data warehouse
- Developed interactive Tableau dashboards with drill-down capabilities, enabling stakeholders to track key performance metrics
- Created machine learning predictive models for client outcomes using scikit-learn, implementing feature engineering to improve model performance
- Automated data quality monitoring and reporting processes, reducing manual workload by 25% while improving data reliability

**Data Analyst | Zhejiang Earthview Image Inc., China (Geospatial Data, CV, Analytics)**

Sept. 2020 - May 2022

- Developed data processing pipeline for satellite imagery, implementing ETL workflows for terabytes of remote sensing data
- Created computer vision models to automate detection of environmental conditions in rural areas, achieving 87% classification accuracy
- Designed and executed A/B testing framework to evaluate platform features, analyzing user engagement metrics across different interfaces
- Implemented statistical analysis of platform usage patterns, identifying key drivers of user retention and feature adoption
- Contributed to award-winning research project leveraging machine learning for environmental monitoring, recognized in national competition

**Data Science Researcher | Prof. Stephen Coggeshall's Research Group, Online (Python, Big Data, ML)**

Feb. - Aug. 2020

- Performed comprehensive data cleaning and transformation on **one million** financial transactions, addressing missing values and outliers while preserving data integrity
- Conducted rigorous feature engineering and selection using statistical methods (SelectKBest) and performance-driven approaches (ROC-AUC wrapper)
- Designed experimental methodology to evaluate model performance using out-of-time validation to simulate real-world deployment scenarios
- Optimized Gradient Boosting Decision Tree algorithm through Bayesian hyperparameter optimization, achieving industry-leading **99.07%** classification accuracy
- Published analysis methodology and findings in IEEE conference proceedings (DOI: 10.1109/MLBDBI51377.2020.00025)

### Game Data Management System (MySQL, JAVA, JSP, Git)

Jan. - Apr. 2025

- Designed a normalized relational database schema **ER-Diagram** for storing and analyzing character data in MMORPG game, with clear description of relationship's modalities and cardinalities by crow's foot notation;
- Created physical data models with appropriate relationships, enforced data integrity constraints in **MySQL**, and implemented complex SQL queries to generate meaningful reports by joining;
- Developed a **JAVA-based data access layer** that enforced data typing and validation before database persistence, preventing **SQL injection vulnerabilities** through proper implementation of JDBC abstractions for RDBMS interaction;
- Built a **JSP-based Web** with filtering and interactive capabilities to analyze game statistics across multiple dimensions, displaying cross-relational data insights for stakeholders.

### Clustering and 3D Visualization Dashboard (Python, Streamlit, Scikit-learn, Plotly)

Sept. - Dec 2024

- Developed interactive data clustering tool that automatically determines optimal k using an ensemble approach;
- Implemented **3D visualizations** comparing K-means and Hierarchical clustering algorithms;
- Built **ETL** pipeline for custom datasets with automated feature mapping and performance metric calculation;
- Achieved **90%** reduction in time required for cluster analysis compared to manual parameter tuning methods.

### Co-Creator: AI-Powered Content Analytics Platform (Data Pipeline, NLP, OpenCV)

Apr. - May 2024

- Architected and implemented data extraction pipelines that process **YouTube API** data to identify high-performing content patterns and engagement metrics;
- Developed a real-time video processing system using **OpenCV** that extracts and analyzes visual features from uploaded content to drive optimization recommendations;
- Designed **NLP**-based data transformation workflows to analyze script content, identifying key engagement factors and semantic patterns correlated with viewer retention;
- Engineered a metadata analytics engine that processes historical performance data to generate **statistically optimized** titles, descriptions, and keyword sets;
- Built ETL processes integrating multiple data sources to create predictive models for content performance across various creator categories and audience segments;
- Implemented a data **visualization dashboard** that transforms complex content metrics into actionable insights, using statistical analysis to quantify recommendation impact.

### Type 2 Diabetes Risk Prediction Model (Python, Feature Engineering, Machine Learning)

Jan. - Apr. 2024

- Analyzed healthcare survey dataset with **445k** records among **328** variables to identify key diabetes risk indicators;
- Implemented feature selection methodology combining **correlation analysis(heatmap)**, **PCA**, and **chi-square testing** to identify key health predictors;
- Engineered data transformation pipeline addressing class imbalance in healthcare data through **weighted classification** techniques;
- Developed and compared multiple classification models (**Decision Tree**, **Naive Bayes**) with cross-validation to optimize prediction reliability;
- Achieved **72%** sensitivity in diabetes detection (from baseline **51.6%**), prioritizing recall over precision for early intervention;
- Identified and quantified eight critical risk factors including age, exercise habits, and socioeconomic indicators with highest predictive value.

### Word Frequency Analysis (Python, NLP)

Nov. - Dec. 2023

- Conducted comprehensive word frequency analysis on longitudinal Twitter dataset, extracting communication pattern insights;
- Performed statistical modeling to validate **Zipf's law** application in social media communication contexts;
- Implemented **term frequency** normalization methodology to account for varying corpus sizes across time periods;
- Applied **TF-IDF** (Term Frequency-Inverse Document Frequency) weighting to identify contextually significant terms;
- Created temporal analysis framework to track evolution of key terms and topics over multiple years;
- Developed visualization suite demonstrating statistical linguistic distributions through logarithmic transformations.

### Mobility Prediction System (Python, Time Series Algorithms)

Oct. - Nov. 2023

- Built **time series** engine processing mobility data with automated cleaning and aggregation;
- Implemented algorithms separating trend, seasonal, and residual components in temporal data;
- Used statistical forecasting algorithms including Exponential Smoothing (**ES**) and **ARIMA** models for prediction tasks;
- Created visualization library with interactive components to display temporal patterns and prediction intervals;
- Built model evaluation framework with **cross-validation** to quantify forecast accuracy and reliability;
- Optimized algorithm performance for handling large movement datasets with millions of data points.