LUYANG SI

Champaign, IL | (+1) 447-902-1417 | ls36@illinois.edu | LinkedIn | GitHub

EDUCATION

University of Illinois Urbana-Champaign, Champaign, IL 08/2022-08/2025
BS in Information Sciences + Data Science Major GPA: 3.65/4.0
Relevant Courses: Data Management; Model & Learning in Data Science; Data Visualization

TECHNICAL SKILLS

- *Programming Languages:* Python, Java, C++, Command Line, R.
- *Data Analysis:* SQL, R, Power BI, Tableau, LDA, pandas, spaCy, BERT.
- *Full stack:* GitHub, Hugging Face, Jupyter Notebook, VS Code, Android Studio, HTML, CSS.

WORK EXPERIENCES

Narciszmade, *Data Analyst/Information Consultant,* Savannah, GA 11/2025-Present

- Develop the sales data analysis pipelines using Python and R.
- Build and manage databases using SQL.

Business Intelligence Group, *Information Consultant*, Champaign, IL 08/2024-12/2024

- Optimized stock management systems of a national food distributor.
- Developed an inventory forecasting model by using Random Forest and trained it with historical sales data and external factors such as stock market performance, GDP, and import/export data.

EasyTransfer, O*verseas Market Operator, Beijing, CN* 05/2023-08/2023

- Analyzed data on users' behaviors and attitudes on financial products to design advertising strategies.
- Coordinated with organizations such as student council, bank, and financial service agencies.

CERTIFICATION

- IBM Data Engineering Professional Certification
- Google Project Management Professional Certification
- Google Business Intelligence Professional Certification

PROJECTS

Research Dataset Recommendation System (Link)

- Built intelligent tool matching research questions to relevant datasets from 25+ sources (CFPB, FRED, Census, WRDS), reducing manual search time from 2 hours to <5 minutes.
- Conducted end-to-end analysis of CFPB consumer complaints using statistical modeling (OLS regression, t-tests) with reproducible code and publication-ready visualizations.

Canvas Platform Data Ingestion (Link)

- Built a production-style, medallion-layered pipeline (raw/cur/meta) that ingests Canvas-style JSONL into Azure SQL for traceability, then transforms into analytics-ready curated tables.
- Implemented rerun-safe + incremental ingestion using watermarking, enabling repeatable runs and reduced reprocessing for operational workflows.
- Added operational metadata for run auditing, data quality checks, and schema-drift detection to support monitoring and debugging.

Wikipedia API Extraction (Link)

- Built a data extraction + candidate-generation pipeline that pairs EN ↔ ZH Wikipedia pages across sensitive identity categories (e.g., race, gender/sex, nationality, age, religion).
- Computed document- and sentence-level semantic similarity using multilingual embeddings, surfacing low-similarity "mismatch" sentence pairs for downstream review.

Crossref Retraction Metadata Analysis (Link)

- Analyzed how consistently retraction flags (e.g., "retracted", "removal notice") appear in publication titles and how well they reflect true retraction status.

- Tracked indexing drift by investigating 208 DOIs marked retracted in April 2023 but not in July 2024, highlighting discrepancies that can affect research integrity.