Ruizhe(Jason) Dai

dai9@seas.upenn.edu | Philadelphia, PA | 831-225-4787 | www.linkedin.com/in/jason-dai9

Education

University of Pennsylvania, M.S.E in Data Science — Expected Dec 2026

- **GPA**: 4.0/4.0

University of California, San Diego, B.S. in Data Science & B.A. in Economics — Graduated June 2025

- **GPA**: 4.0/4.0
- **Honors**: Summa Cum Laude (Elite **top 2%** academic recognition)
- **Coursework (selected)**: Statistics, Machine Learning, Data Visualization, Databases, Distributed Systems.

Technical Skills

Languages & Analytics: Python (Proficient), SQL, Pandas, NumPy, Statistical Analysis, Tableau, Matplotlib
AI & Machine Learning: Generative AI (LLMs), LangChain, NLP, PyTorch, Scikit-Learn
Systems & Infrastructure: Spark, Hadoop, Distributed Systems, Docker, Kubernetes, AWS, Git, PostgreSQL

AI / ML Systems & Research Experience

Statistical ML Research Assistant, OPTIC Lab (Perelman School of Medicine) — Nov 2025 - Present

- Built an end-to-end ML forecasting and validation pipeline for disease progression using longitudinal HVF data.
- Applied **diffusion-based modeling** and uncertainty evaluation under label shift, for model reliability and safety.
- Designed cross-dataset evaluation workflows (UW $\rightarrow$ external cohort) to assess generalization under shift.
- Conducted large-scale validation and coverage analysis, achieving target coverage within $\pm$**2–3%**.

Research Assistant, Halıcıoğlu Data Science Institute – San Diego, CA — Jun 2024 – Sep 2025

- Prototyped **LangChain agents** with **prompt engineering** to automate Human-AI interaction data collection for **fine-tuning GPT-4o**, boosting visual grounding accuracy by **15%**.
- Designed experiments and automated pipelines to evaluate generative system performance.
- Presented findings to cross-disciplinary audiences; **co–first author** of work submitted to **CHI 2026**.

Industry Experience

Data Scientist Intern, China Asset Management – Beijing, China — Jul 2024 – Sep 2024

- Cleaned, explored, and analyzed large-scale financial datasets (100+ variables) in **Python (pandas)**, improving predictive performance to **F1 = 0.83 (+28%)**.
- Built **Tableau dashboards** to visualize model outputs and key trends, enabling investment teams to clearly interpret results and support data-driven decisions.

Machine Learning Intern, China Electronic Corporation – Beijing, China — Jul 2023 – Sep 2023

- Developed orchestration pipelines using Spark & Airflow for automated anomaly detection (**F1 = 0.80**).
- Developed Python scripts to process large-scale messaging data, translating raw events into actionable insights.
- Worked in **agile workflows** to integrate anomaly detection models into platforms for real-time troubleshooting.

Projects

Time Machine Earth: Spatiotemporal Data Visualization [GitHub link] | [Demo] — Nov 2024 – Dec 2024

- Built a full-stack data platform with **React, Node.js, and SQL** for interactive 3D globe visualization.
- Reduced SQL query latency by **97% (30s $\rightarrow$ 1s)** through indexing, caching, and query rewrites.

Leadership/Activities

Microeconomics (Econ 100A) Tutor, UC San Diego – San Diego, CA — Apr 2024 – Jun 2024

- Led weekly office hours and designed quizzes; clearly explained quantitative concepts to students with diverse backgrounds, demonstrating strong verbal communication skills **(perfect student rating)**.