Nils Matteson

linkedin.com/in/nilsmatteson nilsmatteson.com nilsmatteson@icloud.com Madison, WI

Data Science & CS Senior with expertise in **Distributed Systems** and **MLOps**. Proven track record architecting high-performance engines in Go/Rust and deploying RAG pipelines on AWS. Seeking ML Infrastructure, Backend, or Data Science roles.

Education

University of Wisconsin–Madison Madison, WI

B.S. Data Science, Minor in Computer Science *Expected May 2026*

- **Systems & AI**: Big Data Systems (CS 544), Machine Learning (STAT 479), Artificial Intelligence (CS 540), Machine Organization (CS 354), Programming III (CS 400), Intro to Computer Engineering (CS 252).
- **Data Science & Math**: Data Science Modeling I & II (STAT 240/340), DS Programming II (CS 320), Linear Algebra (MATH 340), Discrete Math (MATH 240).

Technical Skills

Languages: Python, Rust, Go, C++, SQL, TypeScript/JavaScript
ML & AI: PyTorch, Transformers, LLMs, Stable Diffusion, Scikit-learn, Hugging Face, OpenCV
Systems & Cloud: AWS, GCP, K8s, Docker, gRPC, Kafka, Redis, Distributed Systems, Postgres
DevOps & Web: CI/CD, Git, Linux, GitHub Actions, React, Next.js, WebSockets, WebGL

Experience

Research Cyberinfrastructure, UW–Madison DoIT Madison, WI

AI Workflows Research Collaborator *Jan 2026 – Present*

- Deploying **WattBot RAG Playground** (KohakuRAG) as a production-grade Streamlit service on AWS to enable AI energy estimation queries for research teams.
- Building interactive chat interface wrapping hierarchical RAG pipeline with **multi-query retrieval**, cross-query reranking, and ensemble inference with abstention-aware voting.
- Architecting AWS deployment using Bedrock for LLM inference and S3 for pre-built SQLite vector indices, implementing cost optimization and on-prem GPU benchmarking (GB10).

Selected Projects

Sentinel: Distributed Log Streaming Engine *Go, gRPC, Protobuf, LSM Trees, Raft*

Sole Architect of a distributed message queue (5,600+ lines) handling high-throughput log streaming.

- Engineered custom LSM-tree storage engine with skip list memtable achieving **1.7M writes/sec** and **3.9M reads/sec**; implemented CRC32 checksums, bloom filters, and crash-safe WAL.
- Built **Raft consensus layer** for fault-tolerant leader election and log replication; designed gRPC streaming API with Kafka-style topic/partition semantics and offset tracking.

Madison Metro: Autonomous ML Bus Prediction System *Python, XGBoost, Flask, Postgres, React*

End-to-end MLOps pipeline with automated retraining, model registry, and live inference serving.

- Architected data pipeline ingesting **50K+ daily GTFS-RT observations** via polling daemon with validation and deduplication before PostgreSQL feature store ingestion.
- Built autonomous ML pipeline: nightly XGBoost on 14-day window with **automated A/B deployment** (F1 >1% lift) to improve arrival accuracy over static schedules.
- Deployed Flask inference API (**sub-50ms P95 latency**) and React dashboard with real-time tracking, delay heatmaps, and drift monitoring; CI/CD via GitHub Actions.

Synapse: Real-time Collaborative Whiteboard *Rust, WASM, WebSockets, CRDTs (Yjs)*

Sole Developer of a lock-free distributed canvas supporting 50+ concurrent editors.

- Built Rust WebSocket server (Actix) with **CRDT state sync** (Yjs) for conflict-free concurrent editing without operational transforms or central locking.
- Compiled rendering to **WebAssembly** achieving **60 FPS** with 10K+ vector objects; architected horizontal scaling via Redis Pub/Sub with session affinity for zero-downtime.