ARUN SHARMA

+1 612-946-5123 | arunshar@umn.edu | LinkedIn | Google Scholar | Website | Github | Hugging Face | OpenReview

SUMMARY

I build physics-informed AI systems that make spatial reasoning reliable for real-world deployment, from detecting GPS spoofing to enable safer autonomous navigation, to modeling energy-efficient vehicle routing, to downscaling climate data for coastal flood risk assessment. My research fuses deep generative models (diffusion models, transformers), physics-informed learning, and geostatistical priors to address data incompleteness, distribution shift, and violations of physical laws in spatial ML. CS PhD from UMN (2025). At Esri, I shipped a maritime anomaly detection pipeline on AWS that improved classification accuracy from 55% to 73% and reduced API latency by 30%. Seeking MLE and Research Scientist roles.

EDUCATION

University of Minnesota, Twin Cities – Ph.D., Computer Science **2018 – 2025**
Advisors: Prof. Shashi Shekhar | **Committee:** Prof. Vipin Kumar, Prof. Ravi Janardan, and Prof. Ying Song
Dissertation: Distortion-Aware Spatial Data Science | **Doctoral Dissertation Fellowship (2022–23)**
State University of New York at Buffalo – M.S., Computer Science **2016 – 2018**

TECHNICAL SKILLS

Languages: Python, Java, R, SQL, Scala, C/C++, Julia, Rust, Go, Shell (Bash), MATLAB, JavaScript (Node.js), TypeScript.
Machine Learning: MLflow, XGBoost, MLlib, LLMs (fine-tuning & inference), VLMs, RAG, LangChain, VectorDB (FAISS, Chroma, Pinecone), Hugging Face Transformers/TRL, LoRA, PEFT, RLHF, OpenAI API, Triton Inference Server, ONNX.
Agentic AI: Chain-of-Thought (CoT), ReAct, Tool/Function Calling, Agent Traceability, Tool-Call & Token Optimization, Multi-Agent Orchestration (LangGraph, AutoGen), MCP, Prompt Caching.
Big Data & Distributed Systems: PySpark, Flink, Presto, Airflow, Kafka, RabbitMQ, Delta Lake, Docker, Terraform, Kubernetes.
ML Frameworks & Deployments: PyTorch, TensorFlow, JAX, ONNX, SageMaker, Lambda, ECS, Bedrock, vLLM/SGLang, Temporal-Nexus.
RL & Alignment: RLHF/RLAIF/RLVR, PPO, DPO, GRPO, Constrained RL, Reward Hacking Mitigation, Agent Traceability, Tool-use Safety.

EXPERIENCE

ESRI – Environmental Systems Research Institute **Research Scientist Intern, May 2023 – Dec 2023**

- Improved detection of illegal vessel behavior (route deviations, dark shipping) from 55% to 73% accuracy by independently designing and deploying an end-to-end anomaly detection pipeline integrating Transformer-based models, Evidential Deep Learning (EDL), and AWS SageMaker on a large-scale maritime dataset (~500M AIS records), with automated retraining via Lambda, ECS, and Step Functions.
- Reduced maritime route query latency by 40% to support real-time vessel tracking by designing and implementing a scalable Graph-based Traffic Representation and Association (GTRA) framework leveraging PySpark and GeoAnalytics APIs over a large-scale AIS maritime dataset.
- Cut model retraining time by 35% and API latency by 30% to enable faster anomaly alerting by implementing model quantization and optimizing inference with SageMaker Multi-Model Endpoints (MME), Step Functions, SQS, CloudWatch, and CI/CD pipelines.

University of Minnesota, Twin Cities **Graduate Research Assistant Aug 2018 – Aug 2025**

- Enabled detection of GPS spoofing for maritime safety and autonomous vehicle security by leading development of Pi-DPM, a physics-informed diffusion probabilistic model pairing a transformer Context-Informed Encoder (spatial+temporal attention over neighbor trajectories) with a physics-informed decoder embedding a single-axle Kinematic Bicycle Model (S-KBM) prior and physics regularizers on velocity, acceleration, heading, curvature, and turning rate; detected GPS-spoofed and AI deep-fake trajectories; outperforming 6 SOTA baselines (DiffWave, DiffTraj, ControlTraj) at 5/10/20% anomaly rates with strong within- and cross-domain transfer across maritime and urban domains.
- Improved coastal flood risk assessment by downscaling coarse climate projections (25km → ~1km) to fine-grained sea-level maps by co-leading a cross-disciplinary collaboration to design a Kriging-informed Conditional Diffusion Probabilistic Model, a U-Net conditional diffusion conditioned on Universal Kriging interpolations under a Matérn variogram with a variogram-based regularizer penalizing empirical-vs-observed variogram discrepancy in reverse diffusion; downscaled Copernicus CDS and CMIP6 HighResMIP (EC-Earth3P) coarse projections to fine-resolution sea-level anomaly and eddy kinetic energy maps, outperforming bicubic, CNN, GAN, and baseline diffusion on RMSE/MAE/PCC/CRPS across Eastern/Western North America.
- Built GeoTrace-Agent, a production-grade multi-agent framework for auditable spatiotemporal reasoning over heterogeneous trajectory and Earth-observation sources, by designing a typed PlanGraph chain-of-thought planner that decomposes natural-language spatial queries into deterministic sub-tasks over AIS feeds, OSM road networks, Copernicus weather, Sentinel imagery, and space-time-prism tools; integrated specialized agents for Hägerstrand prism reasoning, STAGD-DRM abnormal trajectory-gap detection, TGARD/DC-TGARD rendezvous discovery, and S-KBM kinematic validation, exposed through MCP and JSON-RPC A2A protocols with OpenTelemetry tracing, Postgres HITL review queues, semantic caching, tool-call deduplication, and adaptive prompt compression; reduced per-query token spend by ~40%, lowered cost from ~$0.054 to ~$0.034/query versus no-cache ablations, and enforced physical feasibility by rejecting infeasible regions before user-facing output.
- Designed Pi-GRPO, a physics-informed reinforcement-learning stack for trajectory generation and trajectory-reasoning policies, by unifying PPO, DPO, and GRPO under a shared hybrid reward composed of an unbounded hard S-KBM violation penalty, a 95th-percentile jerk/curvature soft envelope, a Pi-DPM reconstruction-likelihood term, and an optional cross-encoder preference model; implemented GRPO group-baseline advantages without a value head, DPO with a physics-aware γ_phys augmentation, PPO with bounded adaptive KL control, vLLM-backed online rollouts with prefix caching, content-addressed checkpoints, W&B/OpenTelemetry reward decomposition, and HITL-to-DPO preference-data curation from GeoTrace-Agent; trained on ~11K preference triples and demonstrated reward-hacking resistance by reducing hard-violation rate from 18% under vanilla DPO to 0% with physics-augmented DPO, while maintaining bounded KL behavior across 3,000 PPO steps.

SELECTED PROJECTS

- Generated realistic synthetic vehicle trajectories for autonomous driving simulation and traffic prediction by co-designing GCDM (Geo-lucid Conditional Diffusion Model), a two-stage cascaded generative diffusion injecting attributed road-network priors via map-informed latent variables, spatial cross-attention, and a physics-based trajectory decomposition (map-derived base + learned residual); produced trajectories with stronger geo-distribution similarity (density, trip error) and higher velocity/acceleration/jerk fidelity than 4 baselines (DiffTraj, Sashimi, DiffWave, EETG) on Porto (~1.7M trips) and Harbin taxi data, while reducing downstream traffic in/out-flow prediction error.
- Enabled real-time energy-vs-travel-time route planning for electric vehicles by co-designing MBOR (Multi-Level Bi-Objective Routing), introducing three novel data structures, a boundary multigraph, Multi-level Encoded Pareto Frontier View (MEPFV), and 2D cost-interval pruning with multi-edge pruning, to precompute and retrieve complete Pareto-optimal path sets over fragmented road networks; achieved >10× online-runtime improvement on real-world road network data for real-time bi-objective EV route planning.
- Reduced manual investigation area for illegal maritime activity (fishing, oil transfers, trans-shipments) by up to 80% by leading a multi-year, 4-publication research program developing STAGD (Space-Time Aware Gap Detection) with a Dynamic Region Merge (DRM) criterion combining comparison-less temporal indexing, R*-tree hierarchical spatial indexing, and maximal-union merge over space-time prism geo-ellipses; proposed TGARD and Dual Convergence DC-TGARD exploiting ellipse symmetry with bi-directional pruning and early-stopping; experimentally faster and more accurate than baselines AIS data (500K ships) for homeland-security and public-safety applications.
- Accelerating computationally expensive ecosystem models for sustainable agriculture: developed SM-Hybrid, a novel surrogate model for the Daycent agroecosystem model capable of explicitly modeling spatial autocorrelation (within-site dependencies) and tele-connections (cross-site long-range dependencies) using hybrid spatial neural network architectures. Predicted land emissions (GPP, Ra, Rh, NEE, crop yield) across Iowa/Illinois/Indiana 2× faster than Daycent while achieving higher accuracy than SM-ANN baselines, enabling rapid scenario evaluation for climate-smart agricultural policy.
- Enabling precision immunotherapy through spatial analysis of tumor microenvironment: developed domain-adapted spatially-lucid neural networks for multi-category point set classification in non-Euclidean space, using a spatial ensemble framework with place-calibration parameters, weighted-distance learning rate, and spatial domain adaptation across heterogeneous tissue place-types. Leveraged a multi-task architecture with spatial mix-up masking and spatial contrastive predictive coding for self-supervised learning on MxIF oncology data to classify immune-tumor relationships for immunotherapy.
- Community-engaged spatial epidemiology for pandemic response: built an entity-relationship model, system architecture, and implementation to analyze aggregated mobile device data for understanding COVID-19 policy interventions on mobility. Delivered county-level analytics on long-duration visits to high-risk business categories and fine-resolution device count maps to epidemiologists, analysts, and policymakers. Collaborated directly with the Minnesota Department of Management and Budget; work cited in testimony to the MN House Transportation Finance Committee.

HONORS AND ACHIEVEMENTS

DOCTORAL DISSERTATION FELLOWSHIP **2022 - 2023**
University of Minnesota, Twin Cities

TEACHING EXPERIENCE

SPATIAL DATA SCIENCE RESEARCH **Spring 2024**
- Guest lectures on scientific methodology, fostering research skills, and effectively communicating research ideas.
- Supervised diverse projects and mentored multiple graduate and undergraduate students from interdisciplinary fields.

SPATIAL DATA SCIENCE **Fall 2019**
- Responsible for handling class lectures, queries, homework assignments, labs, exams, lecture slides, etc.
- Guest lecture topics: Spatial Indexing, Networks, and Data Mining.

ADVANCED DATABASE SYSTEMS **Spring 2019**
- Responsible for handling class lectures, queries, homework assignments, labs, exams, lecture slides, etc.
- Guest lecture topics: Concurrency Control, Database Security, and Data Mining.

DATA STRUCTURES AND ALGORITHMS **Fall 2018**
- Instructed weekly recitation sessions with over 40+ students and graded 400+ students.

SERVICES AND LEADERSHIP

SUST 4096 SUSTAINABILITY INTERNSHIP, UNIVERSITY OF MINNESOTA, MN (Mentor) **Spring 2024**
- Supervised an undergraduate student on a course project based on KGML for Sustainable and Precision Agriculture.
- Presented a lightning talk and a poster at the Annual AI-LEAF PI Meeting in UMN Saint Paul Campus, MN[1].

HONORS MENTORS CONNECTION, WAYZATA HIGH SCHOOL, MN (Mentor) **Fall 2023 - Spring 2024**
- Advised a high school student to analyze a real-world case study of signal spoofing behavior in the maritime domain.
- The student helped analyze circular patterns, enabling us to create a taxonomy of spoofing behavior in open waters.

MINNESOTA DEPARTMENT OF MANAGEMENT AND BUDGET, MN (Service) **May 2020 - June 2021**
- Reported county-level mobility traffic to epidemiologists, analysts, and policymakers for informed decision-making.[2]
- Advised multiple high school and undergraduate students who were considering a research career.

[1] https://cse.umn.edu/aiclimate/news/2024-ai-climate-annual-review-meeting
[2] Future of transportation in a post-pandemic world, Prof. Shashi Shekhar, Transportation Finance and Policy Committee, Minnesota House of Representatives, Jan. 14th, 2021.

INVITED PRESENTATIONS AND POSTERS

Poster: Towards Physics-guided Generative Foundation Models **2025**
The 1st ACM SIGSPATIAL International Workshop on Generative and Agentic AI for Multi-Modality Space-Time.
Authors: Arun Sharma, Majid Farhadloo, Mingzhou Yang, Bharat Jayaprakash, William Northrop, and Shashi Shekhar

Poster: Spatial Distribution-Shift Aware Knowledge-Guided Machine Learning **2024**
AAAI Bridge on Knowledge-Guided ML: Bridging Scientific Knowledge and AI, 2024.
Authors: Arun Sharma, Majid Farhadloo, Mingzhou Yang, Ruolei Zeng, Subhankar Ghosh, and Shashi Shekhar

Presentation: Towards Distortion-Aware Spatial Data Science **2023**
Michigan Institute of Data Science and Society (MIDAS), University of Michigan, Ann Arbor
Authors: Arun Sharma and Shashi Shekhar

SELECTED PUBLICATIONS

[1] Towards Physics-informed Diffusion for Anomaly Detection in Trajectories: A Summary of Results
Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Geospatial Anomaly Detection (GeoAnomalies '25)
Arun Sharma, Mingzhou Yang, Majid Farhadloo, Subhankar Ghosh, Bharat Jayaprakash, and Shashi Shekhar
[2] Towards Kriging-informed Conditional Diffusion for Regional Sea-Level Data Downscaling: A Summary of Results
32nd International Conference on Advances in Geographic Information Systems (SIGSPATIAL '24)
Subhankar Ghosh*, **Arun Sharma***, Jayant Gupta, Aneesh Subramanian and Shashi Shekhar (*Both authors contributed equally to this paper)
[3] Physics-based Abnormal Trajectory-Gap Detection
ACM Transactions in Intelligent Systems and Technology, 2024.
Arun Sharma, Subhankar Ghosh, and Shashi Shekhar
[4] Analyzing Trajectory Gaps for Possible Rendezvous Regions
ACM Transactions in Intelligent Systems and Technology, 2022
Arun Sharma and Shashi Shekhar
[5] Towards a Tighter Bound on Possible-Rendezvous Areas: Preliminary Results
30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22)
Arun Sharma, Jayant Gupta, Subhankar Ghosh, and Shashi Shekhar
[6] Geo-lucid Conditional Diffusion Models for High Physical Fidelity Trajectory Generation
Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25)
Mingzhou Yang, **Arun Sharma**, Majid Farhadloo, Bharat Jayaprakash, and Shashi Shekhar
[7] Towards Surrogate Models with Hybrid Spatial Neural Networks: A Summary of Results
Proceedings of the 8th ACM SIGSPATIAL International Workshop on Geospatial Simulation (GeoSIM '25), pp. 57–69, 2025
Shengya Zhang, **Arun Sharma,** Majid Farhadloo, Mingzhou Yang, Yao Zhang, Mu Hong, Licheng Liu, David Mulla, and Shashi Shekhar
[8] Towards Fine-Tuning-Based Site Calibration for Knowledge-Guided Machine Learning: A Summary of Results
5th Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE)
Ruolei Zeng, **Arun Sharma**, Shuai An, Mingzhou Yang, Shengya Zhang, Licheng Liu, David Mulla, and Shashi Shekhar
[9] Spatially-Delineated Domain-Adapted AI Classification: An Application for Oncology Data
SIAM International Conference on Data Mining (SIAM SDM) 2025
Majid Farhadloo, **Arun Sharma**, Alexey Leontovich, Svetomir N. Markovic, and Shashi Shekhar
[10] Towards Spatially-Lucid AI Classification in Non-Euclidean Space: An Application for MxIF Oncology Data
SIAM International Conference on Data Mining (SIAM SDM) 2024
Majid Farhadloo, **Arun Sharma**, Jayant Gupta, Alexey Leontovich, Svetomir N. Markovic, and Shashi Shekhar
[11] Spatiotemporal Data Mining: A Survey
Handbook of Spatial Analysis for the Social Sciences, Edward Elgar, 2022
Arun Sharma, Zhe Jiang, and Shashi Shekhar
[12] Understanding COVID-19 effects on mobility: A community-engaged approach
25th AGILE Conference on Geographic Information Science, 2022
Arun Sharma, Majid Farhadloo, Yan Li, Jayant Gupta, Aditya Kulkarni, and Shashi Shekhar
[13] WebGlobe: A cloud-based framework for interacting with climate data
International Workshop on Analytics for Big Geospatial Data (SIGSPATIAL) 2018
Arun Sharma, SM Arshad Zaidi, Varun Chandola, Melissa R Dumas, and Budhendra L Bhaduri

REVIEWER: NeurIPS, ICML, ICLR, CVPR, ECCV, AAAI, IJCAI, SIGKDD, CIKM, ICDM, SDM, MLSys, SIGSPATIAL, TKDE, JMLR

REFERENCES (CONTACT INFORMATION)

Prof. Shashi Shekhar	Prof. Vipin Kumar	Prof. Ying Song	Prof. Varun Chandola	Prof. David Mulla
5-203 Kenneth H. Keller Hall	5-225C Kenneth H. Keller Hall	548 Social Sciences Building	212 Capen Hall (CARA Suite)	564 Borlaug Hall
200 Union Street SE	200 Union Street SE	267 19th Ave S	University at Buffalo,	1991 Upper Buford Cir,
Minneapolis, MN 55455	Minneapolis, MN 55455	Minneapolis, MN 55455	Buffalo, NY 14260	St Paul, MN 55108
Email: shekhar@umn.edu	Email: kumar001@umn.edu	Email: yingsong@umn.edu	Email: chandola@buffalo.edu	Email: mulla003@umn.edu.
Phone: 612-624-8307	Phone: 612-624-8023	Phone: 612-625-1112	Phone: (716) 645-4747	Phone: (612) 625-6721
Personal Website	Personal Website	Personal Website	Personal Website	Personal Website