Arun Sharma
Research Statement

My research investigates novel spatial data science and GeoAI techniques to address rising societal challenges that are critical to the well-being of citizens, the resilience of communities, environmental sustainability, and social equity. My primary focus is on developing innovative approaches to identify adversarial behaviors, particularly those involving data distortion, such as missing training data. For example, maritime trajectory data is increasingly being used for monitoring purposes to protect marine biodiversity. Large-scale commercial fishing has led to the over-exploitation of marine resources, threatening the diversity of life in oceans and seas. However, current methods often underperform due to insufficient training data or missing values while also struggling to enable out-of-sample predictions. Hence, my thesis explores novel spatial data science methods for imputing missing data, leveraging known physics-based knowledge.

1. Background:

Spatial data has tremendous value and is a necessary component in many important societal applications. In recent years, spatial technologies such as Google Maps, Waze, Uber, Lyft, Grubhub, Lime, and autonomous driving have revolutionized our everyday lives. Location data on mobile devices generates hundreds of billions of dollars in revenue annually [1] with applications in energy, health, retail, etc. The world's economy also heavily relies on location and time data from over 2 billion GPS receivers [2], as this data is essential to applications in banking, air travel, law enforcement, emergency services, and telecommunications. Meanwhile, new types of spatial data are emerging at unprecedented scale and volume. In transportation, a single vehicle can generate onboard diagnostic data at 25 GB per hour [3]. Earth observation data is being collected at increasing spatial resolution and frequency because of its value in transportation, infrastructure security, resource mapping, agriculture monitoring, etc. Spatial data research is funded by many National AI Institutes (AI-CLIMATE), and multiple Harnessing the Data Revolution (HDR) Institutes have been established. However, spatial data science and GeoAI face several unique challenges.

On the one hand, the knowledge derived from spatial data provides essential context for understanding and interpreting spatial patterns of events and objects across different areas. On the other, many scientific domains ignore or remain unaware of spatial data's special nature. They largely adopt a spatial one-size-fits-all (OSFA) approach, where traditional data mining and machine learning methods are trained without accounting for the geographic properties that give rise to diverse geophysical and cultural phenomena. It is similar to choosing a general practitioner over a specialist when treating a complex medical condition. Extracting meaningful and useful patterns from spatial or spatiotemporal datasets is complex due to the diverse data types and intricate relationships involved. Moreover, spatiotemporal data samples do not follow a uniform distribution across all regions and periods; instead, different geographic areas and temporal intervals exhibit distinct distributions. Finally, while there exist many spatial data mining and statistical methods to address these issues, they tend to underperform when there is insufficient training data or missing values. This is my current area of interest, namely, how to handle missing or distorted data, especially as these relate to adversarial behavior.

2. Research Accomplishments

My research explores the broad area of anomaly detection in spatial data science to understand adversarial patterns generated via data distortion that point to likely data manipulation, including missing-value imputations. Such adversarial pattern detection methods require physical interpretation, and current machine learning methods often fail due to the lack of large training samples. Therefore, my thesis explores a physics-based approach to investigate adversarial behavior where no prior training data is present. The problem of understanding data distortion for adversarial behavior is societally important to meet grand challenges of environmental sustainability, such as maritime safety and regulation [4, 5, 6]. One example of data distortion in the maritime domain occurs through *denial of service* when vessels engaged in illicit activities turn off their devices to mask their location [7].

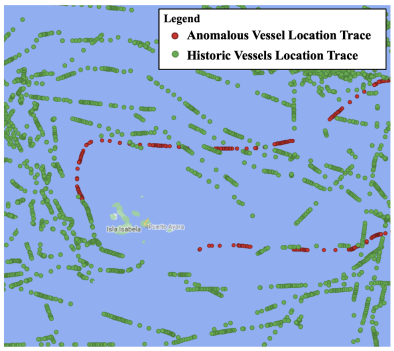


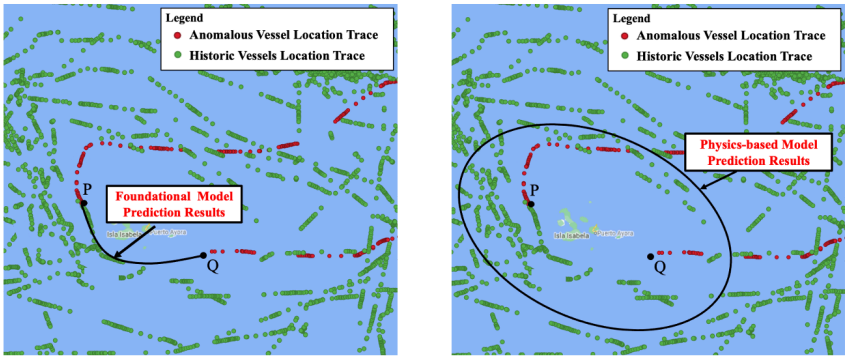
Fig. 1: Historical AIS Locations (in green) and possible illegal fishing vessel (in red)

My thesis explores solutions to three broad computational challenges. First, spatial regions affected by data distortion require further scanning of large satellite imagery at high computational cost, potentially delaying decision-making by human analysts and domain experts. Fig. 1 shows a real-world example of data distortion [8], where an actual fishing vessel switched off its location trace device (i.e., denial of service) before entering a Galapagos marine-protected habitat and then switched it back on 15 days later, even though vessels must broadcast their location signals via satellite to prevent potential accidents or collisions [9]. Second, current methods for identifying non-trivial adversarial behavior (e.g., denial of service) fail due to the non-availability of training samples. Third, monitoring the world's ships for abnormal activity requires handling billions of location-trace signals broadcasted every minute, each annotated with geographic location and multiple ship attributes such as speed, direction, draft, etc. The result is terabytes of data to analyze, and such a high update rate and processing results in high computation costs.

I have addressed these challenges in two main lines of research: physics-guided anomalous trajectory-gap patterns and spatiotemporal optimization for non-trivial anomalous signatures. I've divided my key contributions toward solution quality and then described computationally efficient algorithms for both areas. My work has resulted in 5 research publications, including ACM Transactions in Intelligent Systems and Technology (TIST), and prestigious conferences in spatial data science such as the ACM Conference of Geographic Information Systems (SIGSPATIAL), International Conference on Geographic Information Science (GIScience), and International Conference on Spatial Information Theory (COSIT).

2.1 Physics-guided Anomalous Trajectory-Gap Patterns

My work is based explicitly on explainable methods to model adversarial patterns where no prior training data is present or is completely unsupervised. Current data-driven methods fail to capture such adversarial behavior, making it reasonable to use a physics-based approach to model uncertainty. First, we spatially quantify a trajectory gap based on the idea that an individual object's possible movements are constrained by certain limits of space and time. I represent these constrained movement possibilities as a space-time prism [10] by considering physics-based attributes (e.g., ship's speed). This results in a more precise trajectory reconstruction technique than the classical assumption of shortest or most frequented path interpolation derived by current methods. The method captures certain out-of-sample predictions which are often missed by the shortest or most likely path assumption. Figure 2(a) shows a foundation model (FM) that trains on historical location data for the entire study area and outputs the most likely path an object would have taken while traveling from P to Q. By contrast, Figure 2(b) shows that we can leverage a simple physics-based model (e.g., a space-time prism) to constrain the object's movement in the form of a geo-ellipse, based on maximum speed (but unconstrained acceleration for simplicity) and without any prior knowledge of existing data [11].



(a) Trajectory Gap Prediction via Foundational Model

(b) Trajectory Gap Prediction via Physics-based Model

Fig 2: Comparison of proposed physics-based model with related work

(a) Explainable Anomaly Measure: In my approach, I compute an abnormal gap measure (AGM) for gaps falling within a GPS coverage area (i.e., the probability of an object's location not being reported

during a gap despite GPS coverage). The given anomaly detection method is explainable because, since 2005, ships have been legally required to report their location signals. Therefore, signal gaps that occur despite GPS coverage may indicate possible abnormal behavior. The results were also verified by a real-world case study of the Galapagos marine-protected habitat [7], where the proposed AGM scores based on a geo-ellipse captured such denial-based abnormal behavior more accurately than linear interpolation estimation. The explainability of these results increases the confidence in algorithmic reasoning, which is critical to the responsible use of these machine learning systems. This work was published at the *International Conference on Spatial Information Theory 2022*, a flagship conference in spatial data science [10].

(b) Computational Efficiency: In the following work, I devised faster algorithms using caching, hierarchical indexing, and on-the-fly AGM computations. The baseline approach first performs spatiotemporal intersections and then coalesces qualified intersected regions to reduce overall redundant computations. It then uses additional space to cache the polygonal shape. Such caching avoids unnecessary gap enumeration without compromising correctness and completeness. To further optimize, we introduced novel preprocessing and hierarchical indexing, which effectively leverage space-time partitioning to index and filter out non-intersecting gaps. We also devised a novel dynamic region merge to compute abnormal gap measures more efficiently while preserving correctness and completeness. This extended contribution was accepted in *ACM Transactions on Intelligent Systems and Technology* [11] (Impact Factor: 9.061).

2.2 Spatiotemporal Optimization for Non-Trivial Known Anomalous Signatures

I have also applied my method to a more sophisticated abnormal pattern where two or more objects switch off their location broadcasting devices to rendezvous and perform illicit activities. Here, I first delineate a possible rendezvous area that is narrowed down via a time-slicing technique [12] and then further minimize data distortion using novel spatial filters before it is sent to human analysts for domain interpretation.

(a) Time-slicing approach: In this work, I further refined the space-time prism by providing a tighter filter within geometric constraints defined at each temporal slice within each gap lifetime. During each time instant *t,* we perform a circular intersection from the start and end points of the ellipse foci using speed and time elapsed as the radius. The geometry derived from the two circular intersections is defined as a lens. For instance, $Ellipse_1$ and $Ellipse_2$ denote two trajectory gaps, and the intersection of the two ellipses shows a possible rendezvous area. However, at a given time instant *t*, the rendezvous area is the intersection of two lenses, $Lens_1$ and $Lens_2$. The experiments evaluated the effectiveness of the proposed time-slicing filtering by quantifying the compactness of the space-time approximation [12, 13, 14].

(b) Computational efficiency: In the next work [13], I addressed computational efficiency via time prioritization and a combination of static and dynamic filtering strategies. The time prioritizer narrows down the time-slicing operation proposed in [12] by estimating the actual duration when two space-time prisms are intersected. The results have been accepted in the flagship journal ACM Transactions on Intelligent Systems and Technology [11] (Impact Factor: 9.061). For spatial networks, we also devised a Dual Convergence Trajectory Gap-Aware Rendezvous Detection (DC-TGARD) algorithm to scalably leverage the symmetry property of the geo-ellipse in the space-time prism by simultaneously filtering nodes in opposite directions (bi-directional pruning) until the lenses meet at the midpoint of the ellipse. More details and preliminary findings [14] were published at the *30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL)* (acceptance rate 23.8%).

3. Research Plan

3.1 Short Term: Generative Models for Detecting Unknown Deception-based Adversarial Signatures

My work quantifying data-distortion uncertainty within a trajectory gap addressed denial-based anomaly detection, that is, instances where objects deliberately withhold their location signals. This work has garnered considerable interest from domain experts, encouraging me to explore additional behavioral and geometric patterns, notably those generated via deception-based activity. When an entity behaves deceptively, it attempts to conceal its movements by intentionally emitting false signals (i.e., spoofing) [15]. Other known trajectory generation methods and microsimulators (e.g., SUMO, MATSim, etc.) incorporate realistic human behaviors encompassing leverage available data, mobility models, and social theory to create trajectory datasets that, while simulated, successfully replicate the real-world dynamics, relationships, and emergent properties visible in human movement. However, such microsimulations are typically limited to transportation

science, such as generating vehicle or human mobility trajectory data, and may not work effectively in other domains (e.g., maritime contexts). Finally, machine learning-based trajectory generators typically do not incorporate physical models, reducing their interpretability and explainability. This lack of transparency is a significant barrier for human analysts who must verify ground truth via satellite imagery. I plan to leverage current generative AI models, such as denoising diffusion probabilistic models (DDPM), with known physics knowledge via the kinematic bicycle model (KBM). The preliminary work has been submitted and is currently under review for the SIAM Conference on Data Mining (SDM25).

3.2.1 Long-Term 1: Knowledge-Guided Foundational Models for Trajectories

Foundation models (FMs) can be adapted to a wide range of downstream tasks through fine-tuning, few-shot, or even zero-shot learning. However, these tasks often require large power generation or transmission capacity from data centers, resulting in the consumption of non-renewable resources [12]. Despite their broad application, we have not yet seen significant efforts to develop foundation models for geospatial artificial intelligence (GeoAI), particularly in anomaly detection in trajectories. Training on large amounts of denial and deception-based behavior offers infinite possibilities for tailoring models to specific anomaly types. Additionally, many data-driven foundational models underperform on certain downstream tasks due to insufficient training data. Such models often incur high computational costs, even when training on small datasets. Thus, there is a need for foundational models trained in a task-agnostic manner capable of supporting a wide variety of anomaly detection tasks. Pre-training embedding vectors using unsupervised or self-supervised objectives is a common practice, and mobility trajectories share many characteristics with natural language sentences. Modeling frequent contextual information, such as road networks and vehicle characteristics, enriches semantic understanding and further reduces computational costs.

3.2.2 Long-Term 2: Other Risks and Challenges for GeoAI Foundational Models

Another long-term task is to investigate challenges related to GeoAI interpretability, robustness, bias, data quality, and adaptability. To enhance interpretability, future KGFMs could incorporate techniques to trace and explain how specific domain knowledge influences predictions. To improve robustness, future KGFMs must handle various uncertainties and environmental factors, such as noise, lighting variations, distribution shifts, and out-of-sample cases, while providing uncertainty quantification (UQ). Addressing biases will require ensuring that training data covers diverse scenarios, including clouds, polar regions, rare events, subpixel objects, and slow geological processes. Improving data quality will involve managing gaps, anomalies, and hotspots, as well as deception-based behavior where objects deliberately broadcast false location signals to deceive the end-users in data sources.

Finally, an open problem for GeoAI is how to achieve model generalizability across space while still allowing the model to capture spatial heterogeneity. Given geospatial data with different spatial scales, we desire an FM that can learn general spatial trends while still memorizing location-specific details. Will this generalizability introduce unavoidable intrinsic model bias in downstream GeoAI tasks? Will memorized localized information lead to an overly complicated prediction surface for a global prediction problem? Large-scale training data will likely amplify this problem, so careful consideration will be required, and I am excited to make contributions in these areas over the next 5 to 10 years.

References:

[1] Big data: The next frontier for innovation, competition and productivity. McKinsey Global Institute, May 2011.
[2] "The World Economy Runs on GPS. It Needs a Backup Plan." Bloomberg, Jul. 2018.
[3] "What's driving the connected car?" McKinsey & Company, Sep. 2014
[4] Goal 14: Conserve and sustainably use the oceans, seas, and marine resources, Sustainable Development Goals, United Nations.
[5] Launching the Grand Challenges for Ocean Conservation, World Wide Fund for Nature.
[6] Presidential Memorandum - Comprehensive Framework to Combat Illegal, Unreported, and Unregulated Fishing and Seafood Fraud, Office of Press Secretary, The White House, June 17, 2014.
[7] How Illegal Fishing Is Being Tracked From Space, Sarah Gibbens, National Geographic, December 3, 2018.
[8] Stoyanovich, Julia, Serge Abiteboul, Bill Howe, H. V. Jagadish, and Sebastian Schelter. "Responsible data management." Communications of the ACM 65, no. 6 (2022): 64-74.
[9] Maritime Safety, International Maritime Organization, https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx

[8] Miller, Harvey J. "Time geography and space-time prism." International encyclopedia of geography: People, the earth, environment and technology 1 (2017).
[10] Sharma, Arun, Jayant Gupta, and Shashi Shekhar. "Abnormal Trajectory-Gap Detection: A Summary (Short Paper)." In 15th International Conference on Spatial Information Theory (COSIT 2022). Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.
[11] Sharma, Arun, Subhankar Ghosh, and Shashi Shekhar. "Physics-based Abnormal Trajectory Gap Detection." ACM Transactions on Intelligent Systems and Technology.
[12] Sharma, Arun, Xun Tang, Jayant Gupta, Majid Farhadloo, and Shashi Shekhar. "Analyzing trajectory gaps for possible rendezvous: A summary of results." In 11th International Conference on Geographic Information Science (GIScience 2021)-Part I. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.
[13] Sharma, Arun, and Shashi Shekhar "Analyzing Trajectory Gaps to Find Possible Rendezvous Region." ACM Transactions on Intelligent Systems and Technology (TIST) 13, no. 3 (2022): 1-23.
[14] Sharma, Arun, Jayant Gupta, and Subhankar Ghosh. "Towards a tighter bound on possible-rendezvous areas: preliminary results." In Proceedings of the 30th International Conference on Advances in Geographic Information Systems, pp. 1-11. 2022.
[15] Fake Signals and American Insurance: How a Dark Fleet Moves Russian Oil, Christiaan Triebert, Blacki Migliozzi, Alexander Cardia, Muyi Xiao and David Botti, NYTimes, May 30, 2023.
[16] How AI Is Fueling a Boom in Data Centers and Energy Demand, Andrew Chow, Time, June 12, 2024.