ISSN 2186-7437

NII Shonan Meeting Report

No. 210

Advancing Mobility Data Science and Mobility AI

Flora Salim
Andreas Züfle
Mahmoud Sakr
Kyoung-Sook Kim
Peer Kröger

February 17–20, 2025

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Advancing Mobility Data Science and Mobility AI

Organizers:

Flora Salim
(University of New South Wales Sydney, Australia)

Andreas Züfle
(Emory University, USA)

Mahmoud Sakr
(Université libre de Bruxelles, Belgium)

Kyoung-Sook Kim
(National Institute of Advanced Industrial
Science and Technology (AIST), Japan)

Peer Kröger
(University of Kiel, Germany)

February 17–20, 2025

1 Background and introduction

The proliferation of handheld GPS enabled devices, spatial and spatio-temporal data is generated, stored, and published by billions of users in a plethora of applications. Multiple communities, in computer science, outside computer science, and in industry, have responded to the pertinent challenges and proposed solutions to individual problems. These communities include mobile data management, spatial data mining, geography, transportation engineering, spatial privacy, and spatial epidemiology. In addition, the AI and machine learning communities have also started exploring spatio-temporal and mobility data. Integrating these communities around the common interest of AI and data science around spatio-temporal and mobility-related problems is the best chance to achieve impactful end-to-end solutions to real world problems in our cities. This Shonan meeting followed the success of the Dagstuhl Seminar on Mobility Data Science held in January 2022, and expand it to the Mobility AI (or GeoAI) community.

Aims

The meeting was aimed at discussing three main topics. First, the foundation of mobility data science and mobility AI research, that is, the open large-scale datasets. Second, the reproducibility of algorithms and models. Finally, future directions, especially cross-discipline directions and cross-country collaborations. For each topic, we will discuss several research questions described as follows.

2 Overview of the meeting

The seminar was run fully in person with a handful of participants joining online. Participants come from different discipline areas and with different expertise, including databases and data mining, geography / geographical information science, AI and machine learning, agent-based simulation, and human-computer interaction.

We designed the seminat to be mostly interactive and hands on. We divide the day over two major broad sessions – morning and evening. The overall topics of discussions were set per session. Participants were encouraged to *post* ideas on the ideas board, which were then voted on by the rest of the group. Further, the groups that were assigned randomly then get to decide the topic that they would like to propose or discuss under the major umbrella topics set for that session.

The following are the major topics set by the organizers that guide the overall discussion over the sessions.

Open Large-scale Datasets.

The success of developing advanced models for various applications largely depends on high-quality large-scale datasets. In mobility data science and mobility AI research, most research works still rely on traditional datasets like Geolife, Gowalla, TaxiPorto, and FourSquare, which were introduced a decade ago. The meeting will discuss the challenges and questions towards developing large-scale open datasets such as:

- What are the challenges of collecting and publishing modern mobility data?
- How to assess the quality of mobility data?
- How to address the privacy concerns in releasing large-scale mobility data?
- Which data models facilitate the integration of heterogeneous data from multiple sources?

Spatio Temporal Foundation Models.

- What are the downstream tasks?
- How is this going to work across all domains, regions, time?

Reproducibility and Standardization.

In comparison to other research communities such as CV and NLP, the reproducibility in the mobility data science/AI research area is overlooked. For example, a proposed model or algorithm is rarely evaluated globally. We cannot guarantee the utility of a model in a new scenario even if its implementation is available. We will discuss questions related to the reproducibility, including:

- What are the main reproducibility issues in the community?
- Can we develop a centralized platform for cross-region reproduction?

- How can we design a reproducibility checklist for mobility data science/AI research?
- How to conduct benchmarking research for each mobility data science/AI research task?

New Directions.

Another important aspect of this meeting is to shape the future of mobility data science and mobility AI (e.g., in the next 10 years). We believe with proper designing, the research of mobility could bring many benefits to our daily life in multiple ways. Research questions discussed will include:

- What are the priority areas of focus for the advancement of mobility data science/AI?
- How can we conduct responsible mobility data science/AI research for other disciplines such as for Intelligent Transportation, business intelligence, emergency and disaster response?
- How can we strengthen the cross-country collaborations to leverage data with more diverse geographical characteristics?
- How to foster interdisciplinary collaborations on a regular basis?

3 Overview of Sessions

Day 1 Morning: Self Introductions

This session started with a 3-minute introduction of all participants and their research. The list of the participants can be found in the next section.

Day 1 Afternoon: Discussions on Open Large-Scale Datasets

This session discussed the wide range of publicly available datasets, including trajectory, spatio-temporal, and synthetic datasets.

Day 1 Afternoon (Part 2): Group Discussions and Presentations on Open Large-Scale Datasets

In the second session we split up into five groups of five selected randomly to discuss the datasets needed to build a spatio-temporal foundation model. Different groups came with different views of large scale datasets, and the taxonomy thereof.

Day 2 - Morning: Group Discussions and Presentations on Open Large-Scale Datasets

We continued the discussions from the end of day 2, with groups reporting back their results.

Day 2 - Afternoon: Discussions on Spatio-Temporal Foundation Models

This session discussed the challenges of developing spatio-temporal foundation models. Three sub-groups were formed based on the development stages: Pre-training, Fine-tuning, and Systems and Interfaces.

Day 3 Morning: Writing Session

In this session we spent the whole morning to write up the discussions and tabulated the discussions in a working journal paper to be submitted post the workshop.

Day 4 Morning: Conclusion and Future Directions

We wrapped up the discussion from the previous days and discussed the plan for the journal paper and future directions. We closed the session with participants sharing the highlights from this Shonan seminar – the positives and the drawbacks, and their main takeaways.

4 List of Participants

- Taylor Anderson, George Mason University, USA
- Yang Cao, Institute of Science Tokyo, Japan
- Gilles Dejaegere, Université libre de Bruxelles, Belgium
- Zeinalipour Demetris, University of Cyprus, Cyprus
- Joon-Seok Kim, Emory University, USA
- Kyoung-Sook Kim, National Institute of Advanced Industrial Science and Technology (AIST), Japan
- Peer Kröger, Christian-Albrechts-Universität zu Kiel, Germany
- Yuxuan Liang, The Hong Kong University of Science and Technology (Guangzhou), Hong Kong
- Amr Magdy, University of California, Riverside, USA
- Gaspard Merten, Université libre de Bruxelles, Belgium
- Mohamed Mokbel, University of Minnesota, USA
- Jianzhong Qi, University of Melbourne, Australia
- Chiara Renso, ISTI - CNR, Italy
- Matthias Renz, Kiel University, Germany
- Hamada Rizk, Osaka University, Japan
- Mahmoud Sakr, Université libre de Bruxelles, Belgium
- Flora Salim, University of New South Wales Sydney, Australia
- Cyrus Shahabi, University of Southern California, USA
- Ranga Raju Vatsavai, North Carolina State University, USA
- Li Xiong, Emory University, USA
- Hao Xue, University of New South Wales, Australia
- Daisuke Yamamoto, Nagoya Institute of Technology Japan
- Du Yin, University of New South Wales, Australia
- Takuro Yonezawa, Nagoya University, Japan
- Claudius Zelenka, Christian-Albrechts-Universität zu Kiel, Germany
- Liang Zhao, Emory University, USA
- Andreas Züfle, Emory University, USA

Figure 1: Group photo of participants

5 Meeting Schedule

Check-in Day: February 16 (Sun)

- Welcome Banquet

Day1: February 17 (Mon)

- Talks and Discussions

Day2: February 18 (Tue)

- Talks and Discussions

Day3: February 19 (Wed)

- Talks and Discussions
- Group Photo
- Excursion and Main Banquet

Day4: February 20 (Thu)

- Talks and Discussions
- Wrap up

6 Summary of discussions

6.1 Discussion on Datasets

In the first afternoon session on Monday we discussed the topic of "Datasets for Spatio-Temporal Foundation Models". This discussion highlighted the complexity of selecting and utilizing datasets for training models that capture spatial and temporal dynamics. A key point is that spatio-temporal data is not limited to trajectories; it encompasses a broader spectrum, including origin-destination (OD) data, edge weights, and other mobility-related features. While OD data can sometimes be transformed into trajectory data, it often lacks granularity, particularly when zones are large, leading to a loss of critical movement information. The conversation also emphasizes the need for a taxonomy of datasets, categorizing available data, their utility, and associated challenges such as resolution, representation (raw, aggregated, or derived models), and encoding standards.

A crucial debate arised around high-resolution data. While valuable, it is not always necessary, as models can achieve good results with limited high-quality data. The use of synthetic or simulated data is another point of contention; while simulations must be carefully calibrated to real-world data to avoid learning artificial biases, fields like chemistry and astrophysics successfully leverage simulated data to infer knowledge. The distinction between algorithmically generated and model-driven simulations is also becoming increasingly blurred.

Data collection remains a significant challenge, particularly for academic research, which lacks access to large-scale datasets collected by industry. The discussion suggests exploring agent-based models and LLM-driven simulations for generating trajectories, as seen in computer vision, where pretraining on randomly generated images has proven effective. Future steps involve a deeper exploration of different mobility data types (e.g., indoor vs. outdoor) and the development of a multi-view taxonomy to systematically address open problems in spatio-temporal modeling.

We discussed the increasing volume of collected data which creates challenges in managing and storing this data. Current practices like using compression techniques are insufficient in the long term. Innovative approaches involving AI and ML could enable more effective ways to retain only the necessary data and 'decay' or delete the rest efficiently. Such strategies would allow for the restoration of important data from compact forms when needed.

Finally, we agreed that defining and standardizing metadata for both real and synthetic datasets is essential. This would facilitate the integration and detailed understanding of various datasets, enabling more effective data utilization and management through formats like Geoparquet, which provide a structured approach to handling extensive datasets efficiently.

A diverse array of datasets is critical for advancing mobility research. Here is a list of mobility-related datasets that was listed by seminar participants :

- **YJMob100K**: A city-scale and longitudinal dataset of anonymized human mobility trajectories. Available at: `https://zenodo.org/records/10142719`
- **NOAA OISST V2**: Optimum Interpolation Sea Surface Temperature. More information: `https://psl.noaa.gov/data/gridded/data.noaa.`

oisst.v2.html

- **Pseudo-PFLOW Dataset**: https://pflow.csis.u-tokyo.ac.jp/data-service/pseudo-pflow/
- **New York Taxi**: TLC Trip Record Data. Access the data at: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- **COVID-19 Mobility Data Network**: https://www.covid19mobility.org/
- **U.S. Dynamic Human Mobility Flow During COVID-19**: https://github.com/GeoDS/COVID19USFlows
- **Semantic Trails**: https://figshare.com/articles/dataset/Semantic_Trails_Datasets/7429076
- **UCR Star**: An interactive tool for accessing public geospatial datasets. https://star.cs.ucr.edu
- **UCR Spider**: A tool for generating synthetic spatial data. https://spider.cs.ucr.edu
- **MobilityTwin.Brussels**: Provides real-time and static datasets for public transport and bike positions.
- **OpenUAS**: A dataset representing urban usage patterns with Area2Vec embeddings. https://github.com/UCLabNU/OpenUAS
- **Yelp Open Dataset**: Includes business reviews, photos, check-ins, and attributes. https://business.yelp.com/data/resources/open-dataset/
- **National Household Travel Survey NextGen OD data**: https://nhts.ornl.gov/od/
- **LODES**: Origin-Destination commuting flows data. https://lehd.ces.census.gov/data/
- **Semantic trajectories (Geolife and OSM)**: https://github.com/chiarap2/MAT_Builder/tree/master/datasets
- **PISA**: a dataset that includes the foot traffic flow data of POIs, used for prompt-based or LLM-based forecasting. https://github.com/HaoUNSW/PISA
- A comprehensive survey on trajectory datasets can be found in the ACM publication: https://dl.acm.org/doi/10.1145/3440207
- **XXLTraffic**: An Extremely Long Traffic Dataset that contrains California in USA and NSW in Australia. https://github.com/cruiseresearchgroup/XXLTraffic

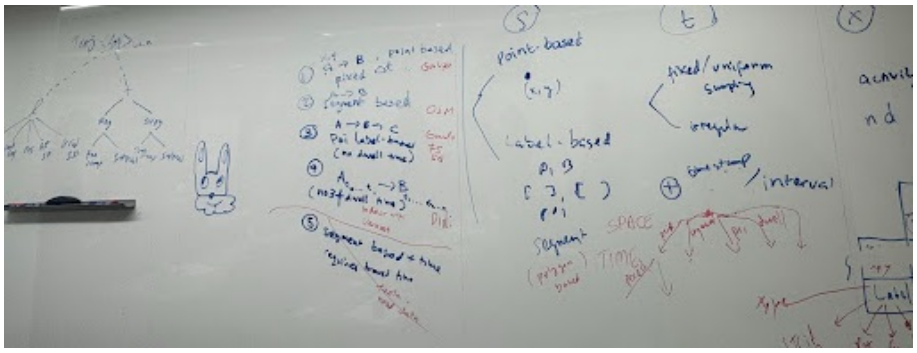

Figure 2: A taxonomy of trajectory data

6.2 Discussion on Taxonomy of Spatio-Temporal Data

In the second afternoon session on Monday, we asked the large team to split into five separate groups of 5-6 people each to continue the discussion on datasets for spatio-temporal foundation models. The five groups met in separate rooms for an hour and then reported their results back to the larger group in the last hour of the day.

Taxonomy of Trajectory data

The discussion emphasizes that tasks should drive data selection, as different trajectory types (e.g., drive routes, trips) vary in structure and timestamps. Trajectory data can be classified as discrete or continuous, often requiring extensive preprocessing to address missing values. Two key definitions are proposed: Definition A, which categorizes trajectories based on movement patterns (e.g., point-to-point, segmented, check-ins, and dwell-time-based data), and Definition B, which classifies trajectories by spatial (point-based vs. label-based) and temporal (regular vs. irregular) perspectives, forming a 3D framework of time, space, and labels. Figure 2 summarizes the discussions of Group 1.

Taxonomy of Spatio-temporal data

We highlighted that a clear taxonomy of spatio-temporal data is essential to understand what types of information can be derived from different data sources and how they interrelate. Transferability of models is a critical challenge, as models trained in one city may not generalize well to another due to differences in infrastructure, mobility patterns, and socio-economic factors; identifying the necessary conditions for successful transfer remains an open problem. Data format and accessibility also play a significant role, with datasets ranging from public and semi-public sources to proprietary or private business-controlled data, impacting both research transparency and model applicability.

We also discussed that a comprehensive taxonomy of spatio-temporal data must account for multiple dimensions that influence data utility and applicability. Key factors include source technology (e.g., GPS, sensors, simulations, social media) and data format (e.g., point, trajectory, raster, graph, OD, or time series). The taxonomy must also consider spatial and temporal granularity, cov-

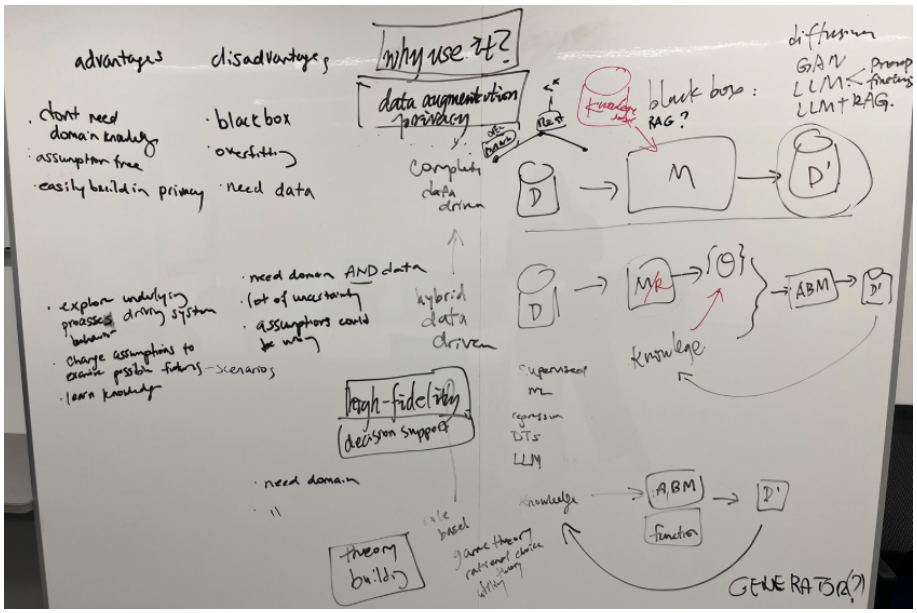

Figure 3: Data-driven versus Theory-driven synthetically generated spatio-temporal datasets

erage, and access level (public, paid, or private), as well as privacy constraints (e.g., de-identified, aggregated, or regulated).

Additional critical aspects include update frequency (historical vs. real-time), uncertainty levels (certain vs. noisy), and the classification of moving objects (humans, vehicles, animals, etc.), with potential constraints. Furthermore, understanding the source organization (e.g., owner, publisher, or simulator) and the intended task (e.g., mobility analysis, map matching) is crucial for evaluating dataset suitability.

Taxonomy of Synthetic and Simulated Datasets

We discussed various methods for synthesizing spatio-temporal data, focusing on both agent-based simulation and generative AI approaches. Within simulation-based methods, the group distinguished between top-down (data-driven) approaches, which resimulate existing real-world datasets, and bottom-up (theory-driven) approaches, which generate simulations without direct data input and later calibrate parameters to align with observed real-world patterns. These two categories of synthetic data generation are illustrated in Figure 3, providing a conceptual overview of their distinctions.

Additionally, a classification of synthetic datasets along the data-driven vs. theory-driven spectrum is presented in Figure 4, offering a structured framework for comparing different synthetic data methodologies. Understanding these approaches is essential for evaluating the reliability and applicability of synthetic data in spatio-temporal modeling.

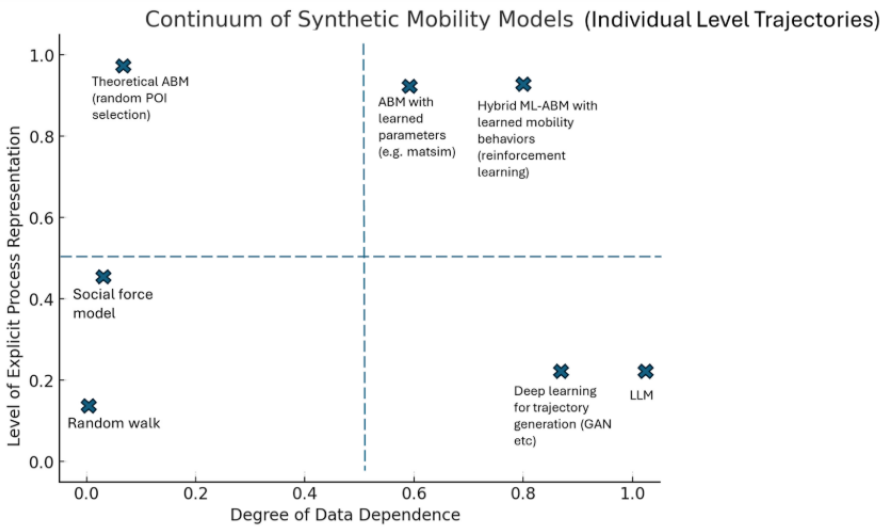

Figure 4: Classification of existing synthetic spatio-temporal datasets

7 Spatio-Temporal Foundation Models

The research focuses on identifying the necessary components for building human mobility foundation models, emphasizing transferability across different cities. Key components include spatial data (e.g., trajectories), modality (simplified mobility characteristics), and context information (e.g., land use, POI). The process involves extracting general features and biases separately, applying spatial normalization, and ensuring broad spatial and temporal coverage, with simulated data used to impute missing information. Evaluation dimensions span multiple tasks, including classification (e.g., anomaly detection, urban/rural classification), clustering, regression (e.g., travel time, edge weights), and generative tasks (e.g., next-location prediction, trajectory reconstruction). Additionally, optimization tasks (e.g., departure time prediction), fusion with large language models (LLMs), and transferability methods (zero-shot and few-shot learning) are essential for assessing model performance across diverse datasets and scenarios.

Pretraining

Building a unified foundation model for spatio-temporal data presents challenges due to irregular timestamps, spatial granularity differences, and semantic incompatibilities across data types, such as ship trajectories versus check-in locations. Different datasets—e.g., GPS trajectories, check-in data, and POI locations—vary in their structure, frequency, and resolution, requiring a framework that can generalize across them. A potential solution is to encode heterogeneous data types into a shared latent space using specialized encoders that preserve their inherent characteristics. Additionally, domain-specific tokenizers are necessary to process different modalities separately, ensuring that each data type maintains its unique vocabulary (e.g., ships cannot logically check in

at McDonald's). By leveraging spatio-temporal points (coordinates + timestamps) as base tokens, the model can flexibly represent both sparse and dense trajectories. This approach enables pretraining strategies where tokens are generated independently but later aligned through embedding techniques, allowing the model to generalize across diverse downstream tasks. Furthermore, fine-tuning on specific tasks can activate relevant domain experts, optimizing task performance while preserving cross-domain adaptability. The key challenge remains in defining the optimal encoding and tokenization strategy to effectively connect disparate data representations within a robust foundation model.

Downstream Tasks

In the exploration of trajectory data applications, various functionalities are essential, organized into categories like similarity-based tasks, prediction-based tasks, and trajectory understanding, with additional focus on data augmentation tasks. Critical within these applications is trajectory anomaly detection, where atypical movement patterns are identified that diverge from normative behaviors. This function is vital across sectors such as fraud detection, safety monitoring, and event detection.

Another pivotal function is trajectory classification, which organizes movement patterns into predefined categories based on spatial and temporal characteristics. This is crucial for urban planning and transportation management, allowing for the differentiation between various types of transportation modes and routine activities through the analysis of location sequences.

Trajectory forecasting and recovery are also significant, focusing on predicting future movements and reconstructing incomplete trajectory data, respectively. Forecasting leverages historical data to predict future locations with considerations of human intentions and external conditions, whereas recovery deals with filling in missing trajectory points to ensure data completeness.

Extended capabilities include trajectory generation, which synthesizes realistic human movement data for privacy enhancement and machine learning purposes; travel time estimation, which calculates times for different routes by considering various spatio-temporal factors; and trajectory-based recommendations, which propose potential future locations or social connections based on past movement patterns. These features showcase the system's ability to deliver a broad spectrum of analytical tools that can cater to needs ranging from traffic simulation to enhanced social networking applications.

Additionally, model collapse -— the degradation of model performance when applied to out-of-distribution data — poses risks to transferability, highlighting the need for robust adaptation strategies.

Systems and Interfaces

To advance research and practical applications in mobility foundation models, we envision the development of a centralized platform for browsing, testing, and training models. Despite numerous papers claiming to provide foundation models for trajectory analysis, few offer readily available implementations. This platform would serve as a "Hugging Face for Mobility", enabling users to evaluate existing models on diverse downstream tasks, upload datasets, and transfer trained models across different cities. A core requirement is an exten-

sible model database for managing and updating model weights, along with a pluggable interface that standardizes input/output formats and supports new tasks. Additionally, a dataset repository would centralize mobility datasets while incorporating metadata for spatial and temporal attributes. Given the unreliability of OpenStreetMap (OSM) data, we propose a structured repository to store essential attributes like maximum speed, lane count, and flow speed, indexed by OSM node and segment IDs. This would enable users to access the latest dataset versions, contribute new data, and visualize completion levels on an interactive map. Ultimately, the platform aims to provide a standardized, collaborative space for improving spatio-temporal mobility models.

8 Summary of new findings

The collection of real data poses its set of challenges, particularly for academic researchers who lack the extensive resources that industries possess to gather large-scale data via methods like crowdsourcing. Despite this, the value of real data in training robust AI models cannot be understated. Furthermore, the discussion extends to the need for a canonized method to encode spatio-temporal (ST) data, enhancing the models' ability to learn and generalize from such data effectively.

A pivotal area of discussion is the distinction between synthetic and simulated data. While synthetic data is generated through models that may or may not adhere closely to real-world dynamics, simulated data is typically algorithmically crafted to mimic specific real-world processes. The nuances between these types and the blurred lines in their definitions within the simulation community need clearer articulation. Additionally, the utility of simulated data is contingent on its generation rules; if these rules are too simplistic or detached from reality, the model's learning potential is constrained. Therefore, simulation models must be calibrated against real-world data to ensure their efficacy and relevance.

In many fields, such as chemistry and astrophysics, reliance on simulated data is common due to the prohibitive costs of experimental data collection. These fields have demonstrated that valuable insights can still be achieved from well-designed simulations. This precedent suggests that mobility data science could also benefit from sophisticated simulations, especially when real data is scarce or unavailable.

9 Feedback from Participants

In the final session, we asked participants to share their positive aspects and areas for improvement of this Shonan workshop. We summarize the feedback received as follows:

Positive Aspects

- Knowledge Sharing & Learning – Many participants highlighted the wealth of knowledge gained, including new research papers, emerging methodologies, and technical insights. The ability to learn directly from experts, rather than just through published papers, was particularly appreciated.

- Networking & Collaboration – The event facilitated new connections among researchers and experts in mobility, foundation models, and AI, leading to potential future collaborations.
- Diverse Perspectives – The seminar brought together researchers from different domains, providing a variety of viewpoints on mobility foundation models, from trajectory data to urban planning.
- Idea Generation & Conceptualization – Discussions sparked new ideas, especially around the potential of foundation models for geospatial applications, trajectory-based learning, and their relevance across different fields (epidemiology, ecology, urban planning).
- Community Building – There was enthusiasm for creating a stronger community around mobility foundation models, potentially leading to joint projects and shared datasets.

Challenges & Areas for Improvement

- Lack of Industry Representation – Many felt the absence of industry stakeholders limited discussions on practical applications and real-world challenges of mobility foundation models.
- Fragmented Discussions – While discussions were rich, they often diverged in multiple directions, making it difficult to converge on clear research priorities or solutions.
- Unclear Applicability of Foundation Models – Some participants questioned whether mobility foundation models are truly feasible or beneficial beyond existing techniques, particularly in the context of geocoding and spatial-temporal relationships.
- Insufficient Time & Structure – The seminar lacked a hands-on component where researchers and students could work together on prototypes, due to limited time and participation constraints.
- Noise & Technical Issues – Attendees joining remotely on Zoom found it challenging to follow discussions due to background noise and technical limitations in the meeting environment.

10 Identified issues and future directions

The group identified the following key takeaways and future directions:

- Potential of a Unified Geocoding Foundation Model – Discussions pointed toward the need for a robust geocoding foundation model that can understand location semantics worldwide.
- Data Standardization is Crucial – Many agreed that before foundation models can be effective, mobility datasets need to be better structured and standardized.
- Cross-Domain Collaboration Needed – Bringing in perspectives from other fields (e.g., epidemiology, urban planning) could help shape meaningful use cases for mobility foundation models.
- Building a Stronger Research Community – Participants expressed interest in forming a more structured community around mobility foundation models, with shared datasets, projects, and potential summer schools.
- Future Workshops Should Include Industry & Applied Use Cases – To enhance practical impact, future events should invite industry representatives to share real-world challenges and applications.
- Overall, the seminar was highly valued for fostering discussion and networking, but there is a need for more structured collaboration, hands-on work, and industry engagement to advance the field effectively.
- A comprehensive taxonomy of datasets is essential. This taxonomy should delineate which datasets are available, their potential applications, and the inherent challenges they pose, such as issues with resolution. It should also categorize the types of data representations available, ranging from raw data to aggregated or compressed forms and derived models. There's a need to critically evaluate whether high-resolution data is necessary for all tasks, as even limited amounts of high-quality data can yield robust results.
- Looking forward, examining different types of mobility data (e.g., indoor vs. outdoor, moving objects) and identifying existing and open problems in a multi-view taxonomy could enrich the field.

The discussions during this Shonan seminar will be further expanded into a journal paper.