ICDE 2026
MONTRÉAL, CANADA

KAIST AI
Kim Jaechul Graduate School

Effective Dataset Distillation for Spatio-Temporal Forecasting with Bi-Dimensional Compression

Taehyung Kwon*

Yeonje Choi*

Yeongho Kim

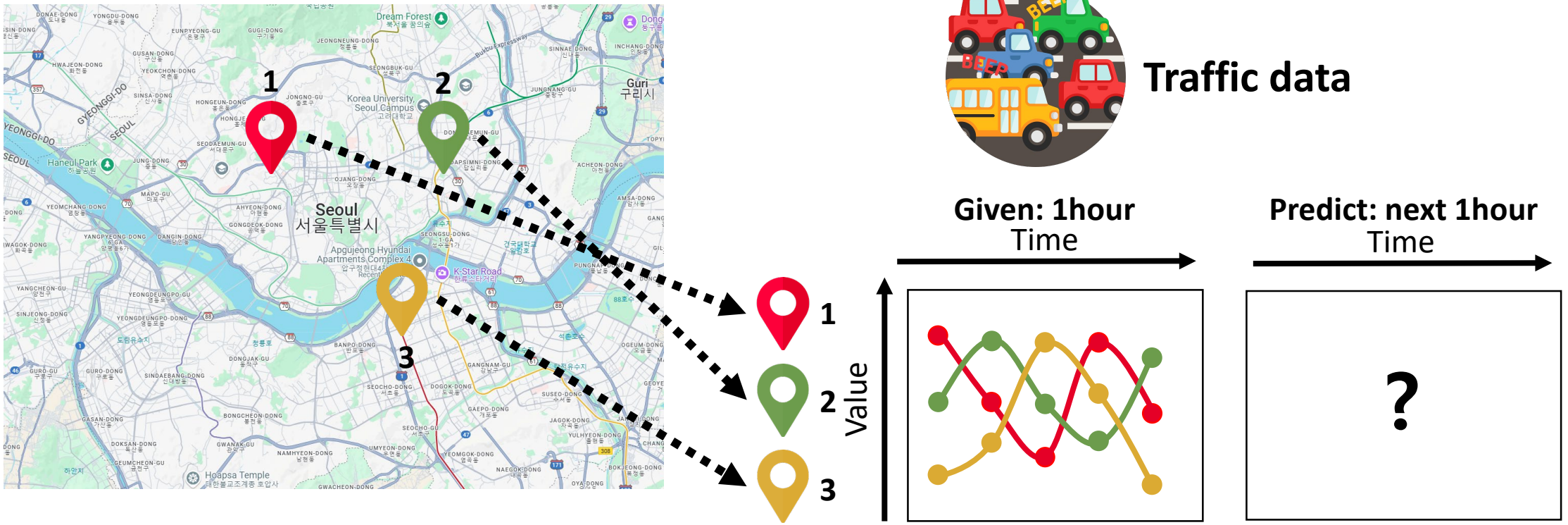
Kijung Shin

Outline

1. Introduction.
2. Proposed method.
3. Experiments.
4. Conclusion.

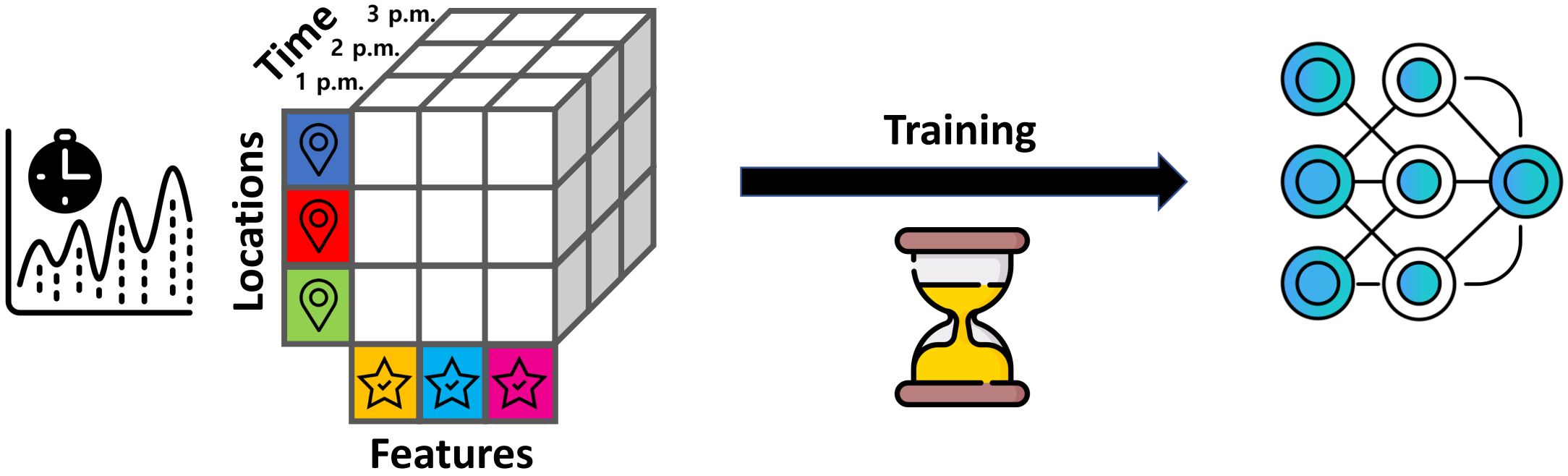
Spatio-Temporal Time Series Forecasting

- Data: multivariate time series collected at various locations at regular intervals.
- Given: all time series data over a certain time period,
- Task: to predict the time series for the subsequent (future) time period.



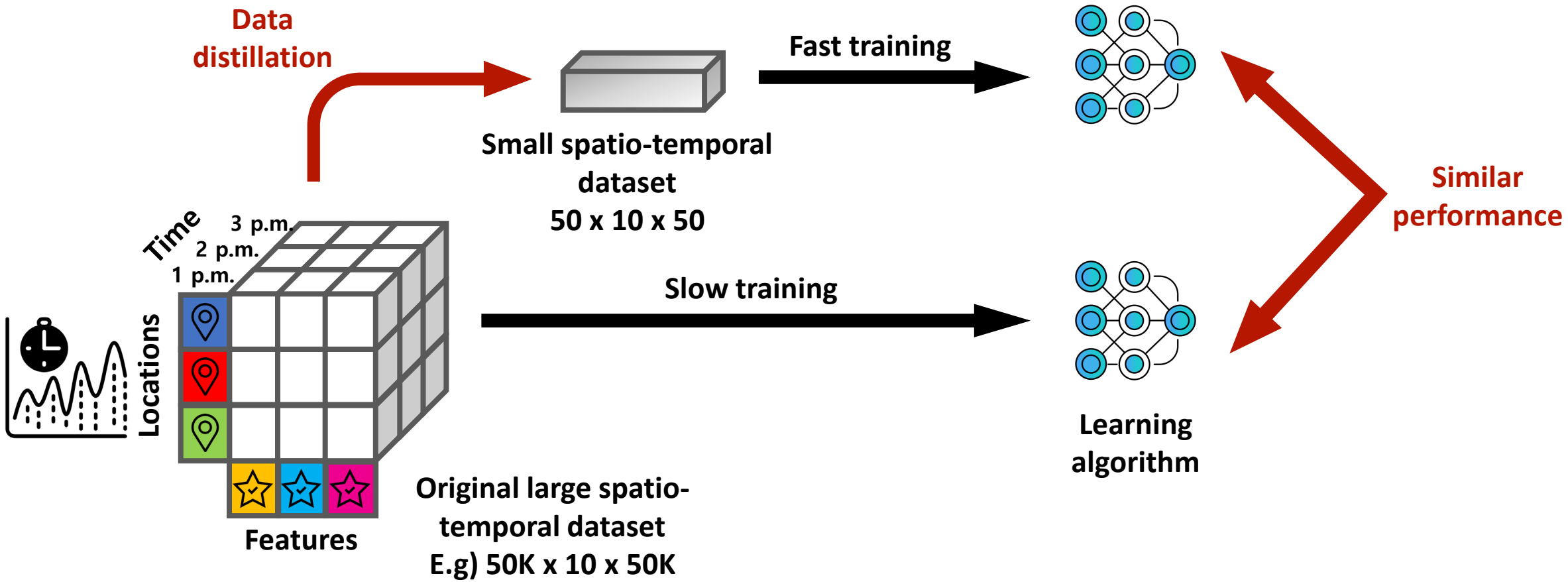
Challenges in Large-Scale Data

- Spatio-temporal datasets are often **massive in scale**, since they are collected from numerous **locations** over extended **periods**.
  - Machine-learning training on such a large-scale dataset can be time-consuming.



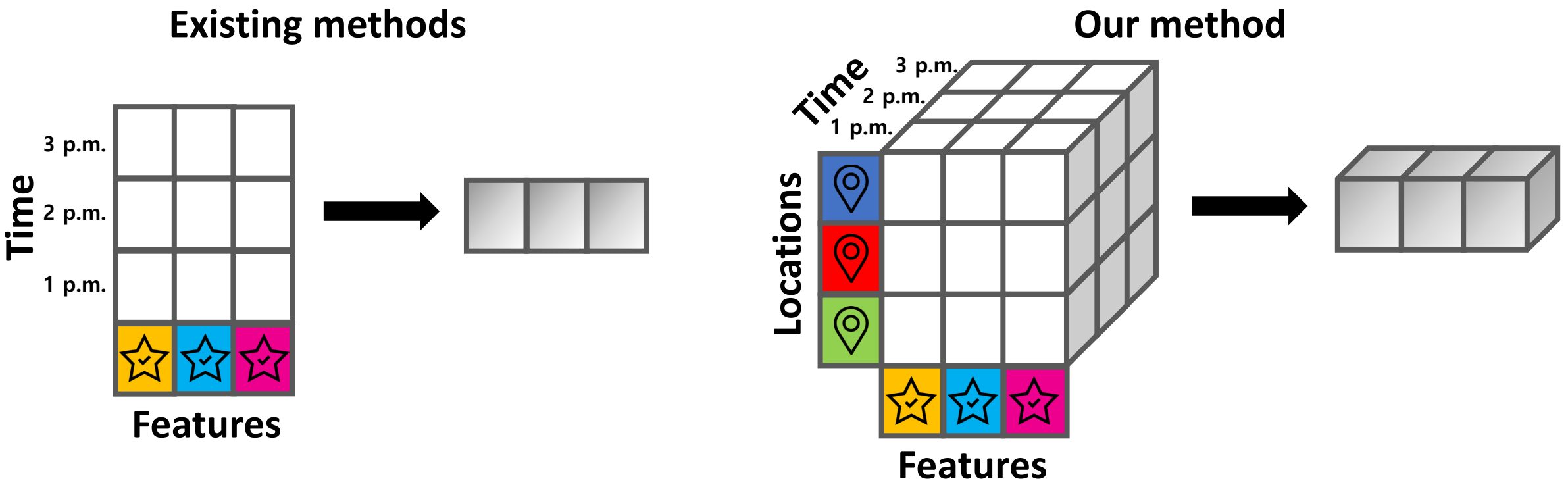
Goal of Proposed Research

- We address **spatio-temporal time series dataset distillation,** which aims to synthesize a compact dataset that enables effective training in place of the original dataset.



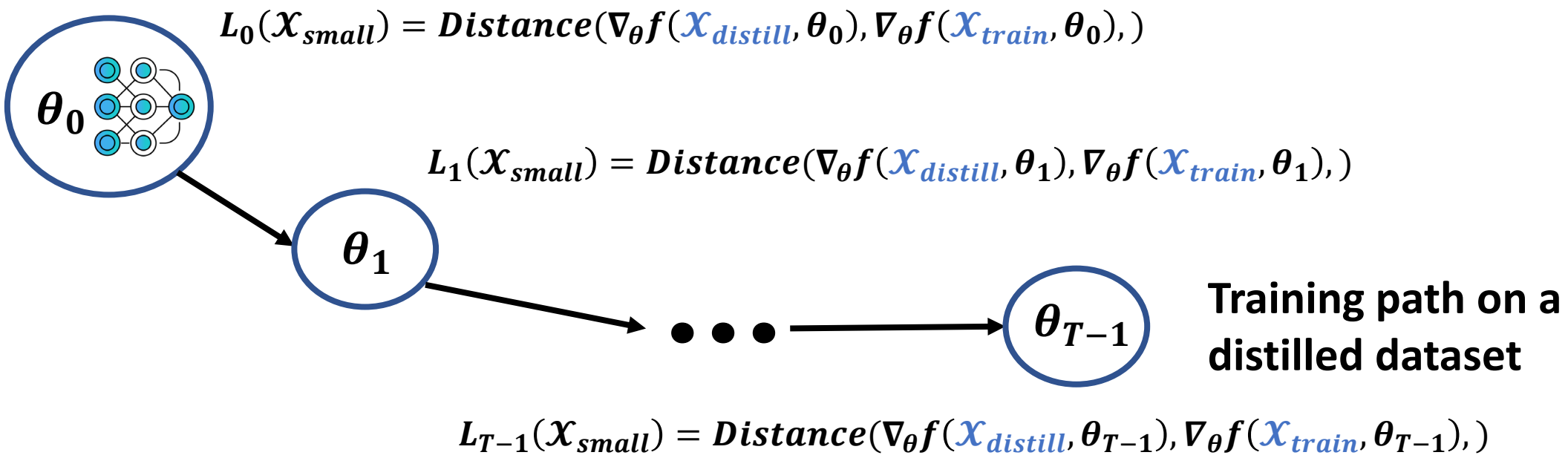
Preliminaries: Existing Methods for Time Series Distillation

- CondTSF and TimeDC are time-series data distillation methods for forecasting problems.
- However, they are limited to single-location time series data (without spatial information).



Preliminaries: Gradient Matching

- Our method is based on gradient matching, which is a widely adapted technique for dataset distillation.
- Gradient matching minimizes the distance between the gradients from distilled and original data.



Outline

1. Introduction.
2. Proposed method.
3. Experiments.
4. Conclusion.

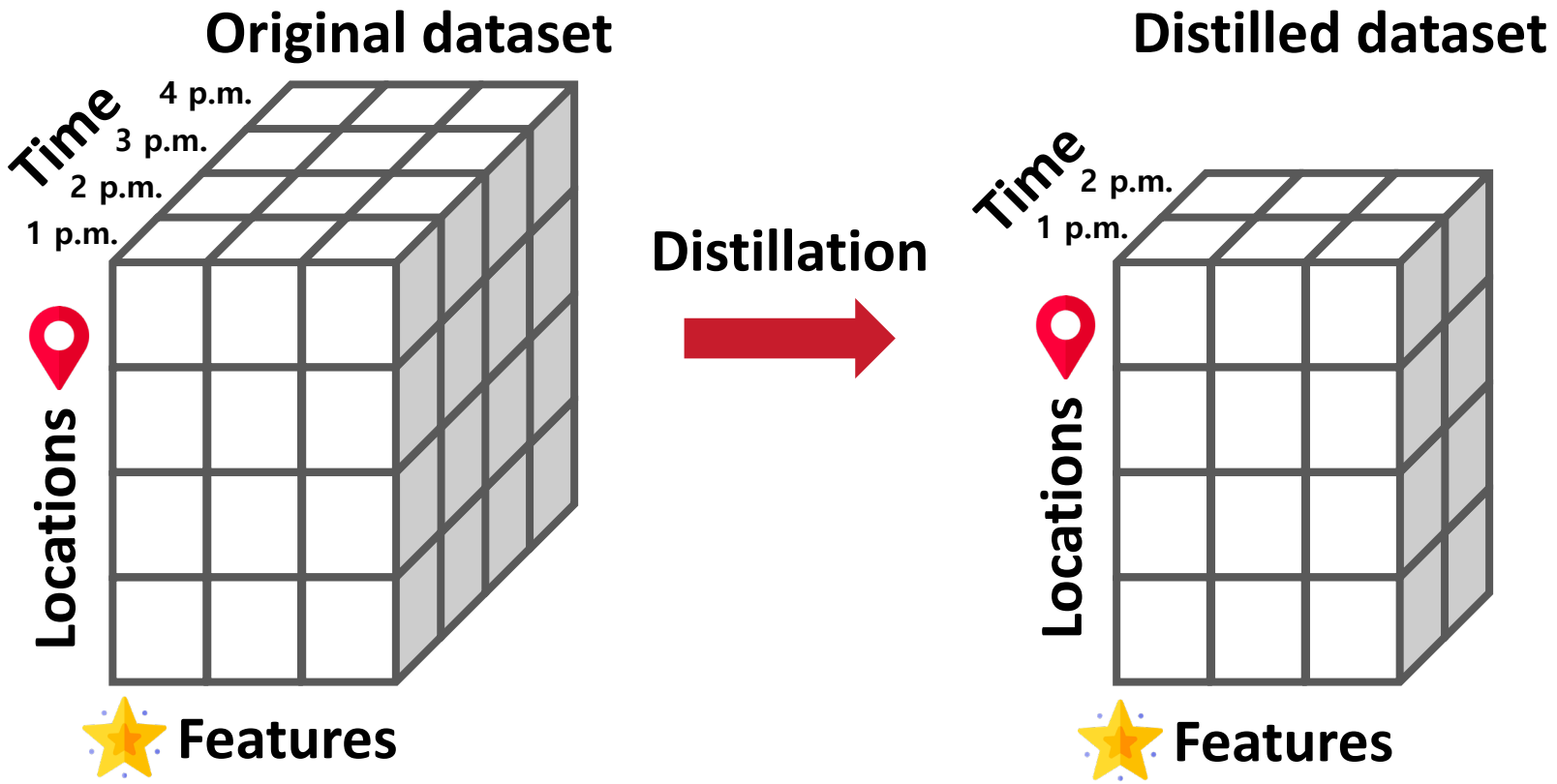
Overview of STemDist

- We propose STemDist, a novel dataset distillation method specialized for spatio-temporal time series datasets.

- **Q1. How can we form distilled datasets to reduce the training cost?**
- Q2. How can we accelerate the distillation algorithm?
- Q3. How can we make distillation better capture the original data?

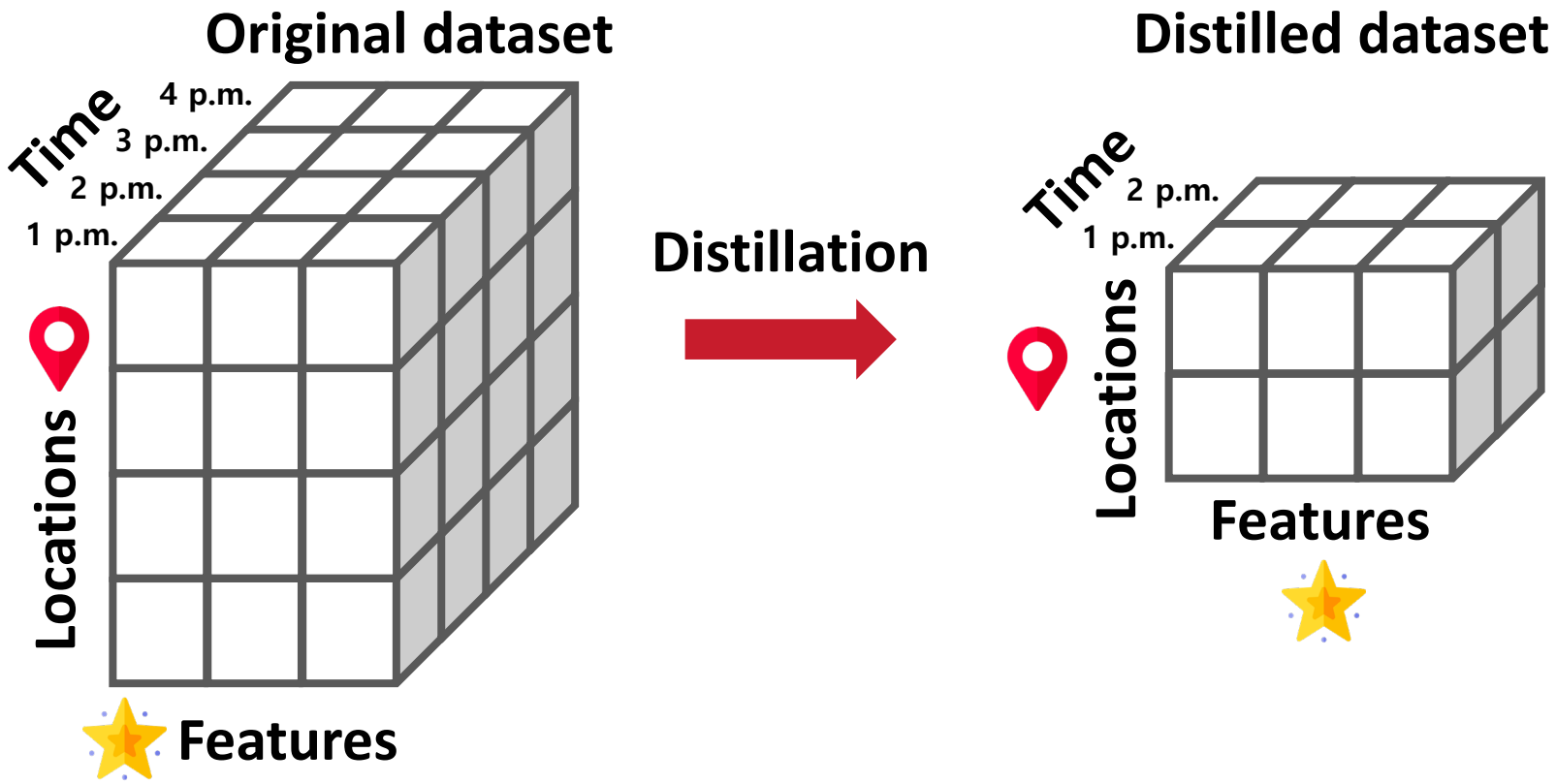
Motivation: Limitation of Existing Methods

- Simply applying existing methods only reduce the number of time steps.
- Distilled datasets still have many locations, leading to high training costs.



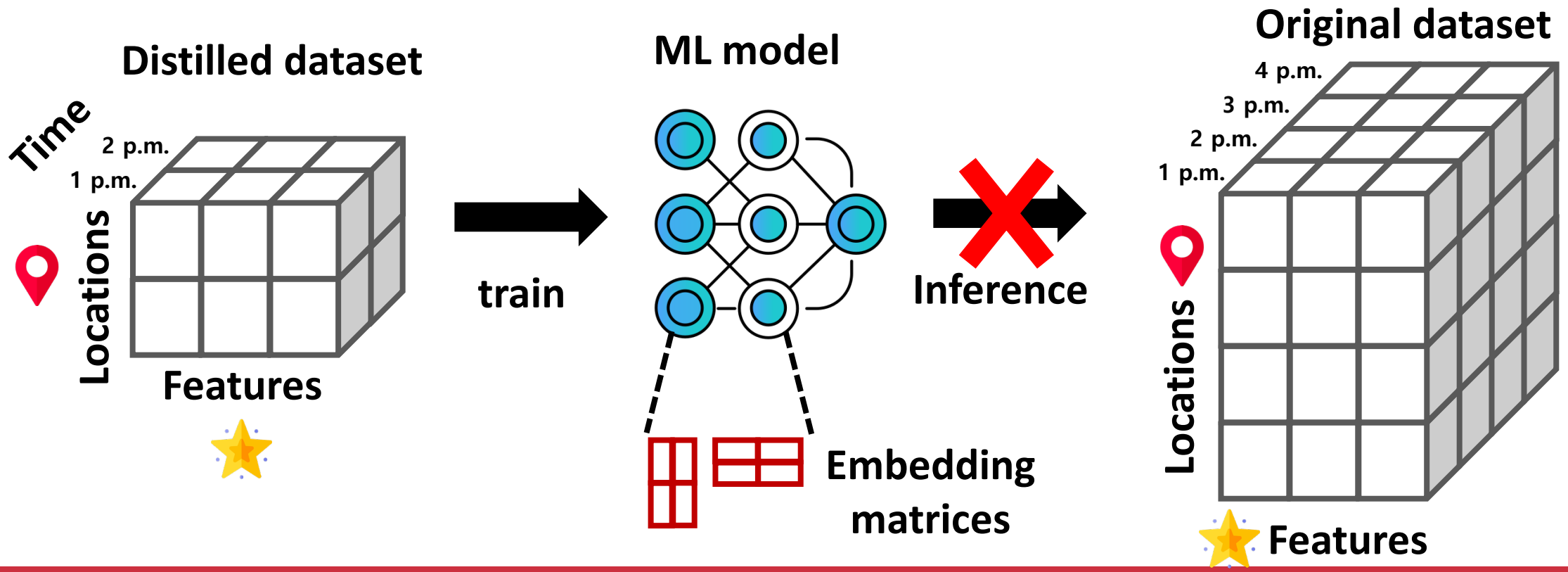
A1. Our Idea: Simultaneous Compression of Two Dimensions

- Our method reduces both the number of time steps and locations.
- Distilled datasets have a small number of locations, leading to lower training costs.



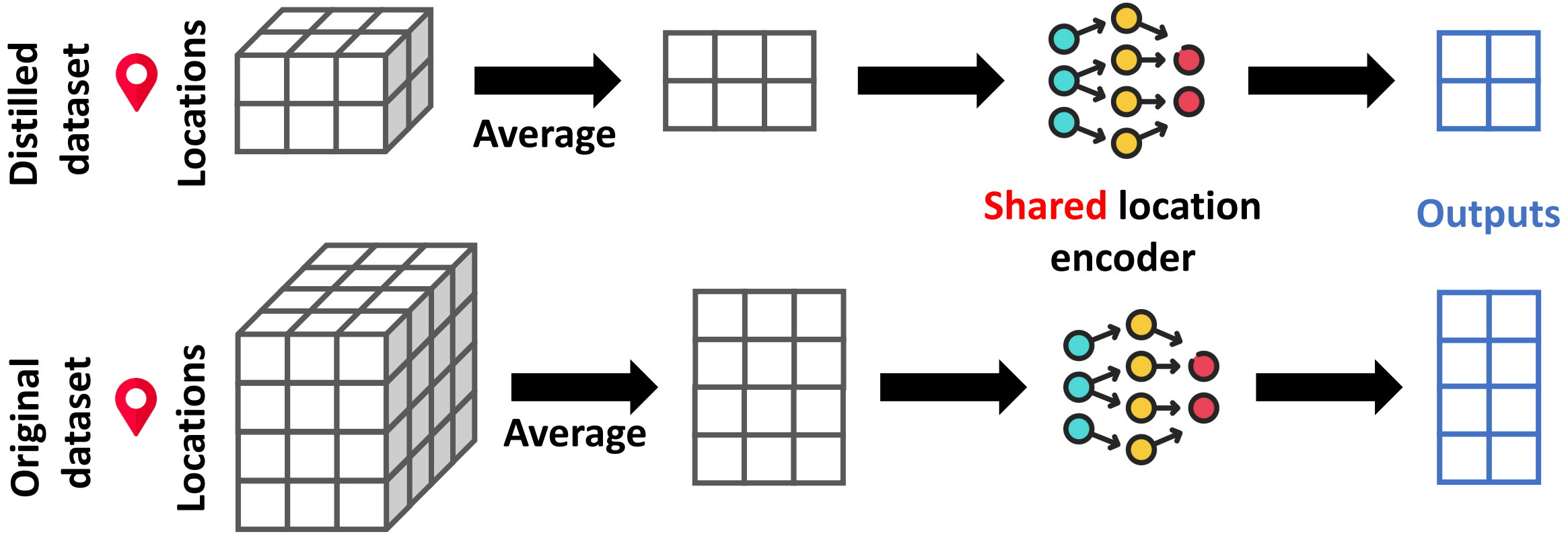
Challenges in Implementing Our Idea

- When the numbers of locations vary, ML models trained on the distilled dataset cannot be used on the original test dataset.
- This is because ML models learn location embedding matrices with the number of rows equal to the number of locations in the dataset.



Solutions for Implementing Our Idea

- We allow ML models to handle datasets of any number of locations.
- This is accomplished by replacing embedding matrices with the outputs of shared **location encoders**, which can handle arbitrary number of locations (e.g., Self-attention-based encoders).



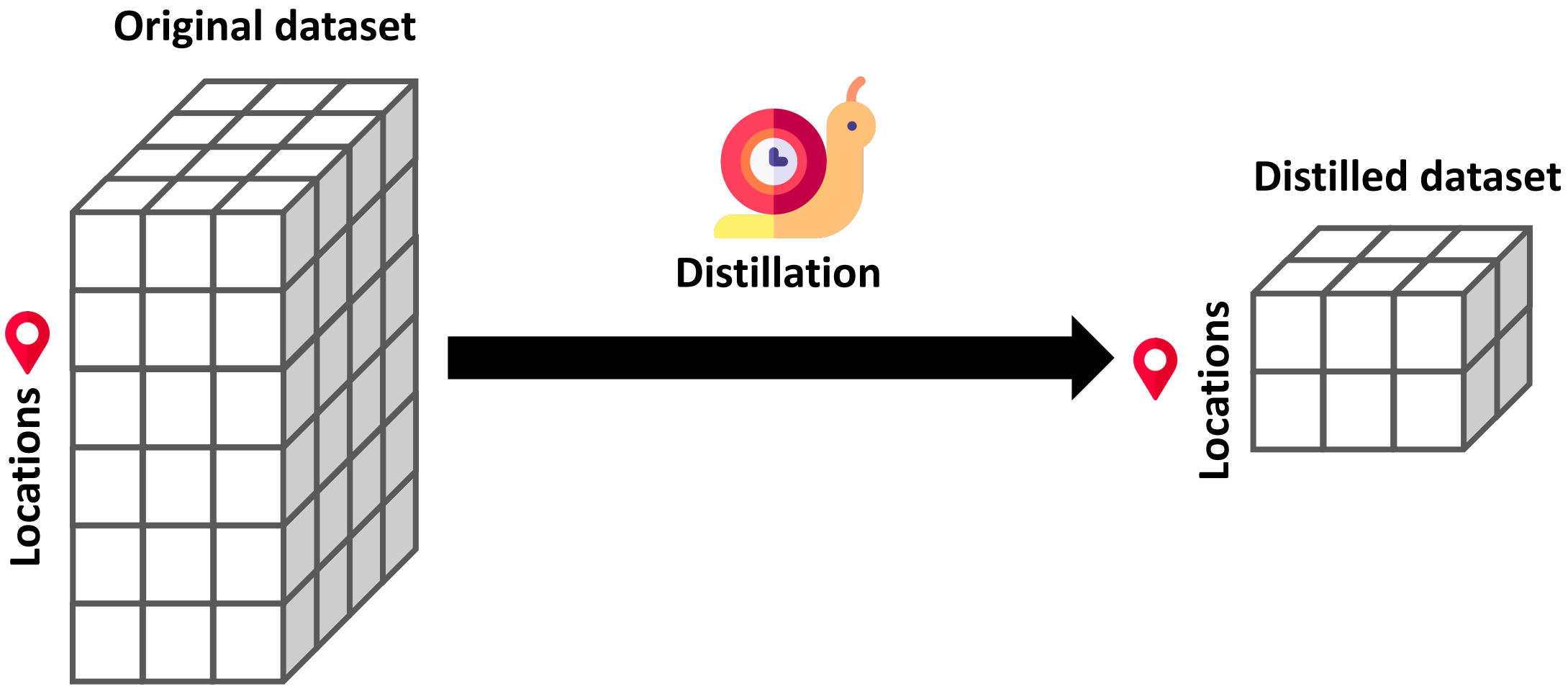
Overview of STemDist

- We propose STemDist, a novel dataset distillation method specialized for spatio-temporal time series datasets.

- Q1. How can we form distilled datasets to reduce the training cost?
- **Q2. How can we accelerate the distillation algorithm?**
- Q3. How can we make distillation better capture the original data?

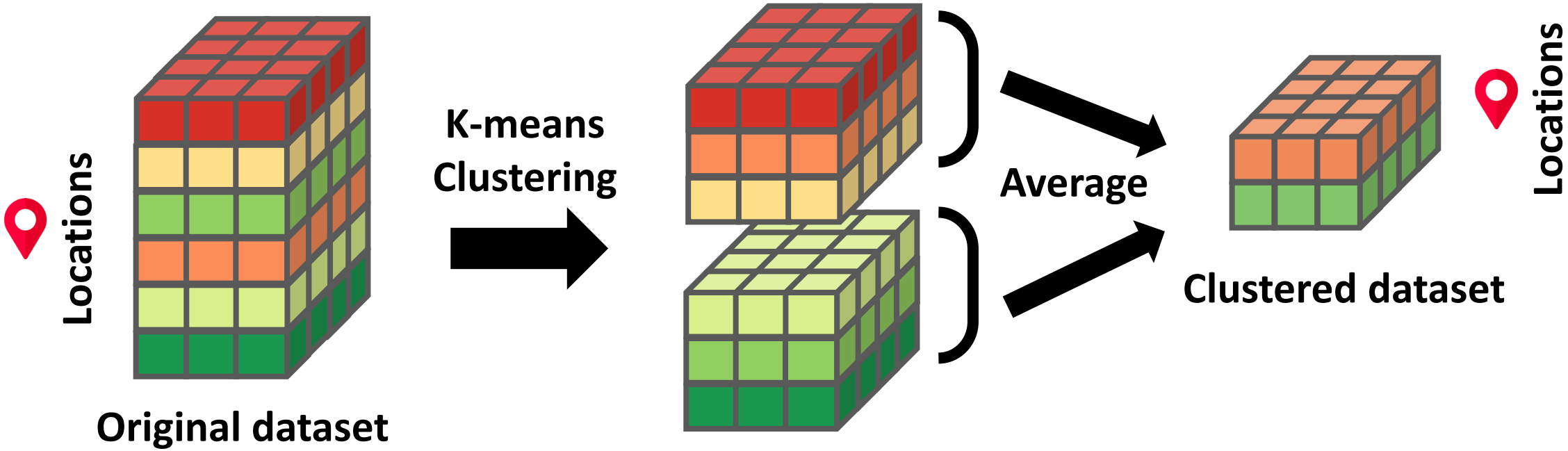
Motivation: Slow Distillation

- A large number of locations in the original dataset slow down distillation.



A2. Our Idea: Location Reduction in the Original Dataset

- We accelerate the distillation process.
- This is accomplished by reducing the number of locations in the original dataset by clustering to match that of the distilled dataset.



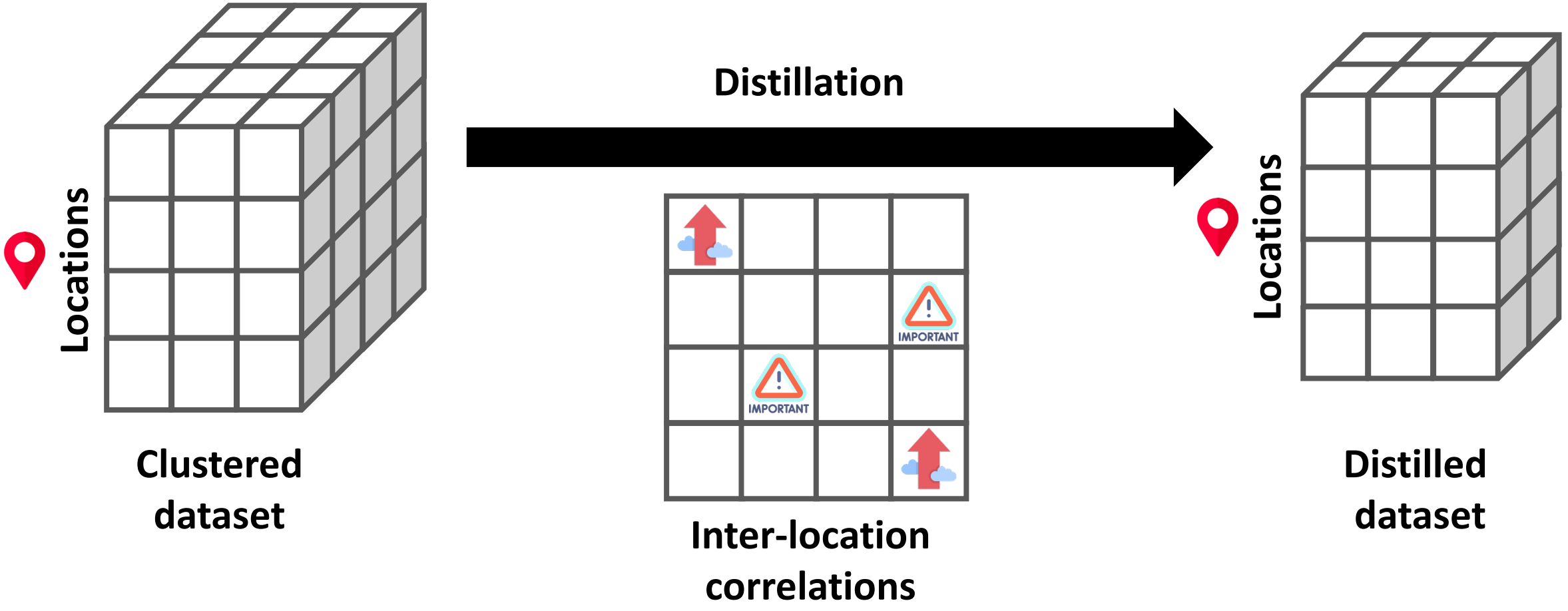
Overview of STemDist

- We propose STemDist, a novel dataset distillation method specialized for spatio-temporal time series datasets.

- Q1. How can we form distilled datasets to reduce the training cost?
- Q2. How can we accelerate the distillation algorithm?
- **Q3. How can we make distillation better capture the original data?**

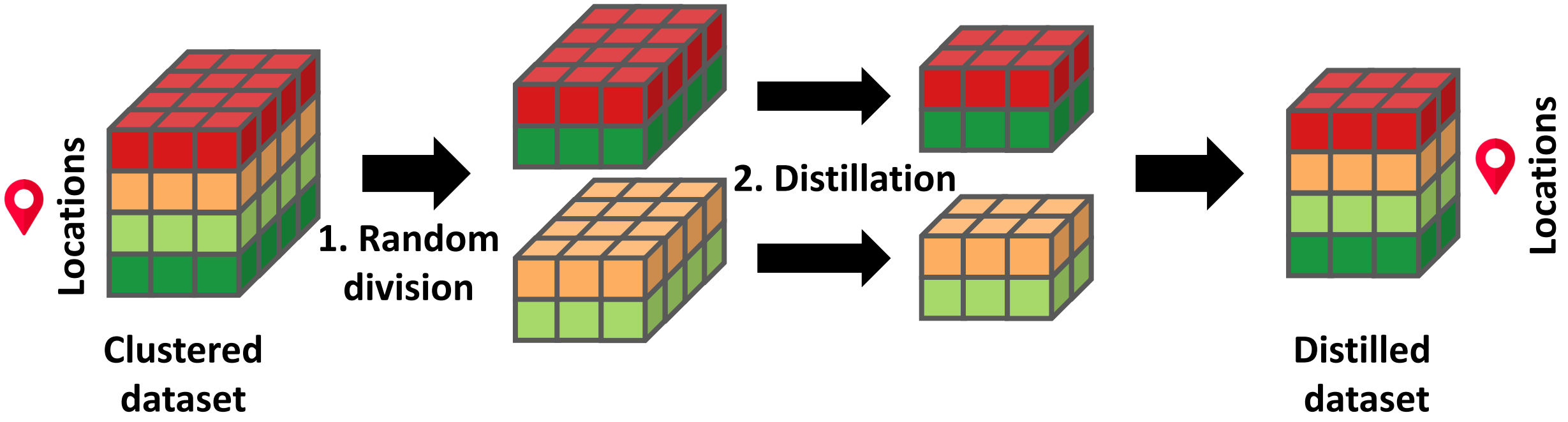
Motivation: Using All Locations is Ineffective

- Using all locations during distillation may overlook weak but important inter-location correlations.



Solution: Use Locations in Smaller Units

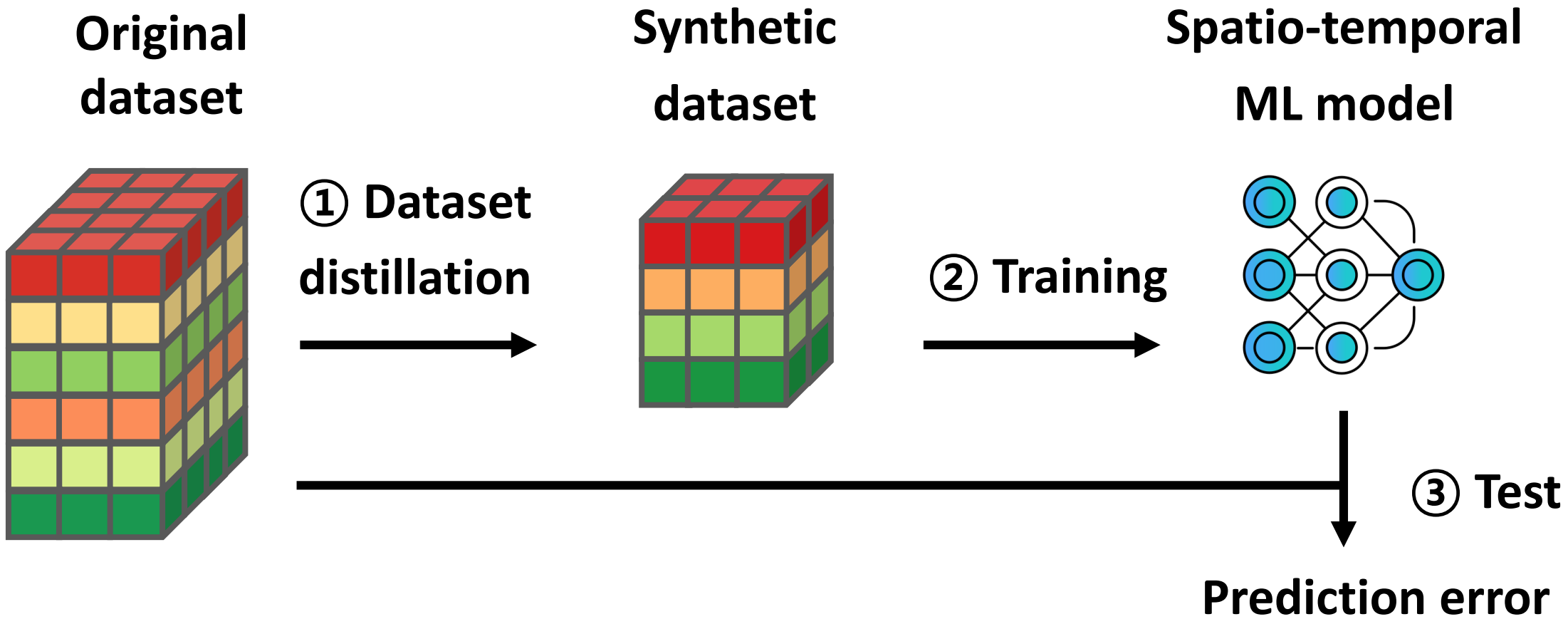
- We distill the dataset in units of location subsets.
  - Weak inter-location correlations can be captured easily during distillation since a smaller number of correlations within subsets are considered at a time.
  - Location subsets are randomly chosen for every distillation iteration.



Outline

1. Introduction.
2. Proposed method.
3. Experiments.
4. Conclusion.

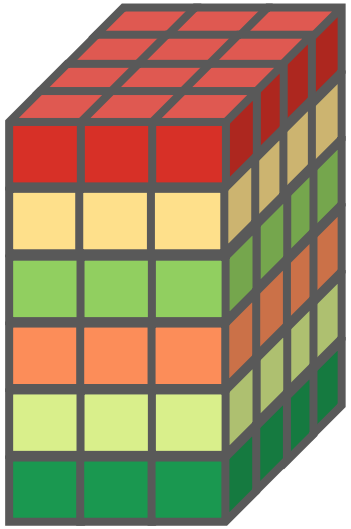
Experimental Settings



Experimental Settings

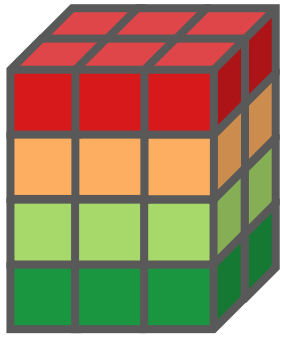
- 3 traffic datasets
- 2 weather datasets

Original dataset



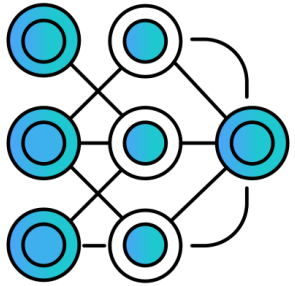
① Dataset distillation

Synthetic dataset



② Training

Spatio-temporal ML model



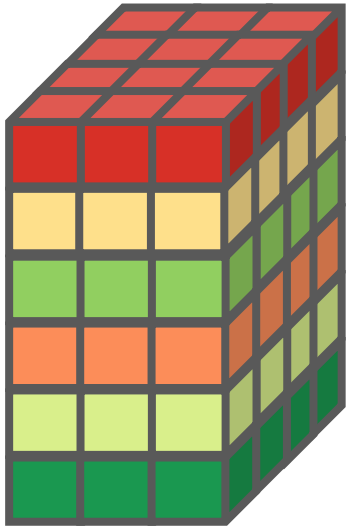
③ Test

Prediction error

Experimental Settings

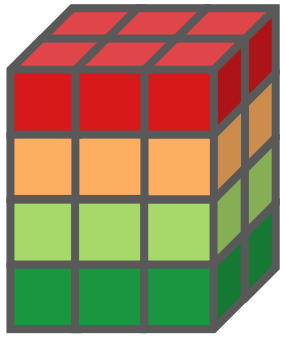
- StemDist (proposed)
- 12 competitors

Original dataset

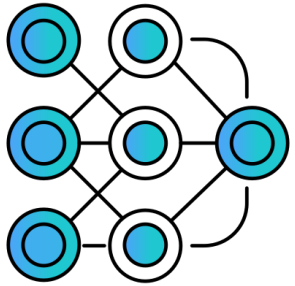


① Dataset distillation

Synthetic dataset



② Training



③ Test

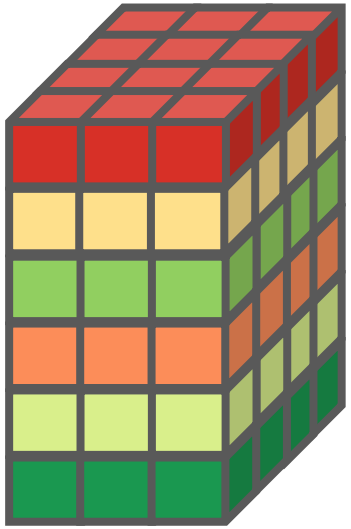
Prediction error

Experimental Settings

- StemDist (proposed)
- 12 competitors

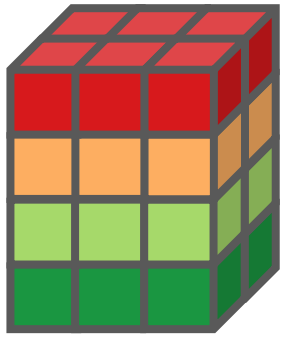
- MTGNN (default)
- 5 other models

Original dataset



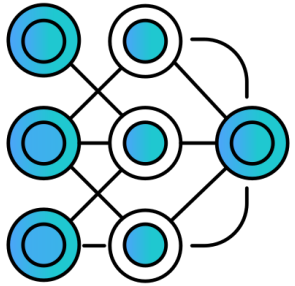
① Dataset distillation

Synthetic dataset



② Training

Spatio-temporal ML model



③ Test

Prediction error

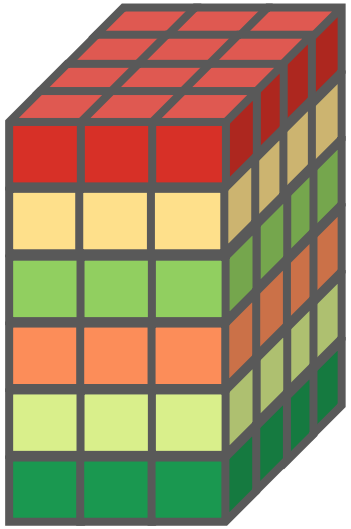
Experimental Settings: Main Evaluation

- StemDist (proposed)
- 12 competitors

Compare!

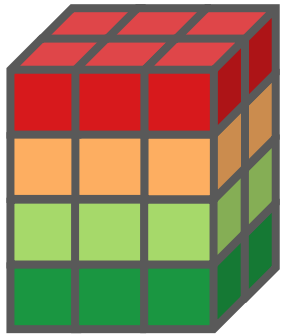
Varying

Original dataset



① Dataset distillation

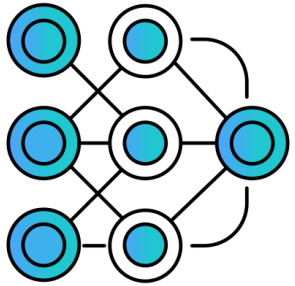
Synthetic dataset



② Training

Fixed

Spatio-temporal ML model



③ Test

Measure: Prediction error

STemDist Enables Effective Training

: **Best working method**

- STemDist consistently outperforms all competitors.

Datasets	GBA		GLA		ERA5		CA		CAMS	
Methods	Relative MAE	Relative RMSE	Relative MAE	Relative RMSE	Relative MAE	Relative RMSE	Relative MAE	Relative RMSE	Relative MAE	Relative RMSE
Random	0.380	0.453	0.313	0.410	0.482	0.520	0.323	0.431	0.660	0.796
K-Center	0.319	0.400	0.269	0.356	0.461	0.524	0.289	0.355	0.883	0.883
Herding	0.467	0.506	0.331	0.428	0.471	0.522	0.472	0.521	0.584	0.730
CRAIG	0.291	0.354	0.258	0.323	0.430	0.470	0.287	0/354	0.563	0.685
DC	0.281	0.348	0.243	0.311	0.392	0.442	0.254	0.331	0.630	0.735
DM	0.338	0.409	0.298	0.380	0.420	0.467	0.321	0.405	0.528	0.674
MTT	0.339	0.414	0.343	0.439	O.O.M	O.O.M	O.O.M	O.O.M	O.O.M	O.O.M
DATM	0.430	0.517	0.332	0.402	O.O.M	O.O.M	O.O.M	O.O.M	O.O.M	O.O.M
IDM	0.358	0.420	0.300	0.381	0.438	0.481	0.315	0.394	0.567	0.709
Frepo	0.397	0.465	0.350	0.439	0.482	0.510	0.361	0.436	0.688	0.802
CondTSF	0.337	0.407	0.311	0.394	0.449	0.489	0.322	0.403	0.555	0.692
TimeDC	0.350	0.422	0.344	0.445	0.437	0.480	0.360	0.448	0.585	0.729
STemDist	**0.251**	**0.317**	**0.211**	**0.277**	**0.385**	**0.426**	**0.226**	**0.292**	**0.522**	**0.653**
Original data	0.207	0.264	0.182	0.250	0.368	0.424	0.120	0.253	0.500	0.661

Compression ratio: 0.5 %

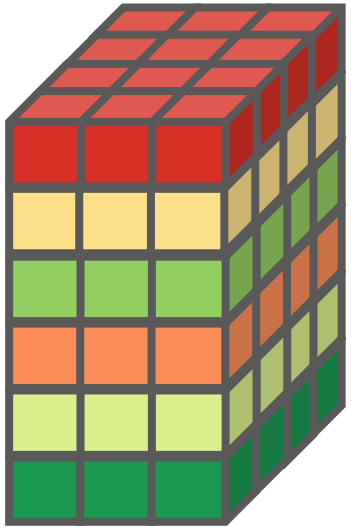
Experimental Settings: Cross-Model Evaluation

- StemDist (proposed)
- Top-performing competitors

Compare!

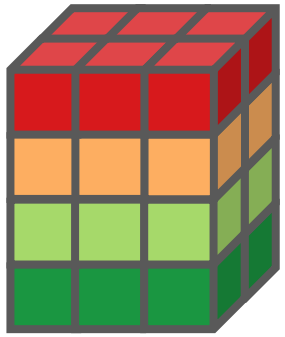
Varying

Original dataset



① Dataset distillation

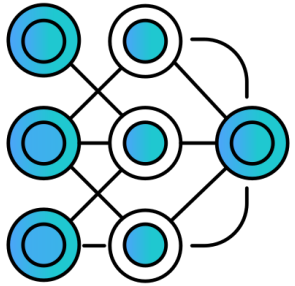
Synthetic dataset



② Training

Fixed

Spatio-temporal ML model



③ Test

Measure: Prediction error

STemDist Enables Effective Training

- Synthetic datasets from STemDist are effective for training various models.

: Best working method

Graph Wavenet

Datasets	GBA	GLA
Methods	Relative RMSE	
K-Center	0.427	0.367
DC	0.395	0.312
CondTSF	0.416	0.387
STemDist	**0.310**	**0.273**
Original data	0.227	0.212

STGCN

Datasets	GBA	GLA
Methods	Relative RMSE	
K-Center	0.714	0.746
DC	0.812	0.866
CondTSF	0.697	0.734
STemDist	**0.501**	**0.459**
Original data	0.294	0.295

FourierGNN

Datasets	GBA	GLA
Methods	Relative RMSE	
K-Center	0.419	0.395
DC	0.403	0.376
CondTSF	0.419	0.421
STemDist	**0.384**	**0.339**
Original data	0.307	0.284

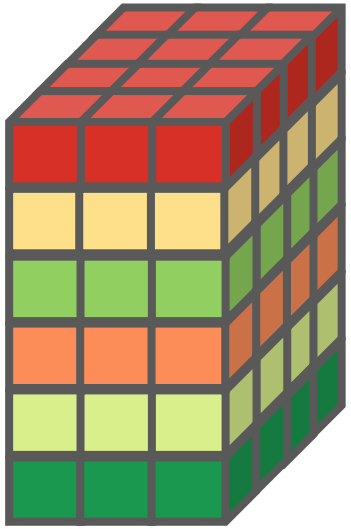
Experimental Settings: Training Efficiency Analysis

- StemDist (proposed)
- 12 competitors

Compare!

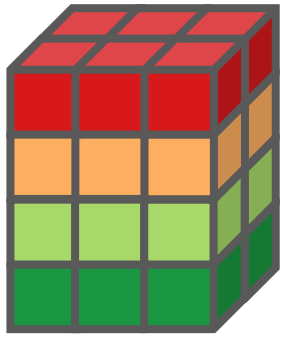
Varying

Original dataset



① Dataset distillation

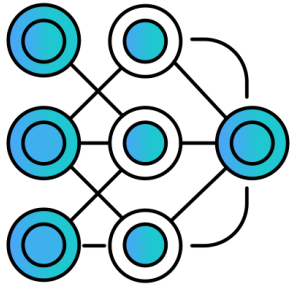
Synthetic dataset



② Training

Fixed

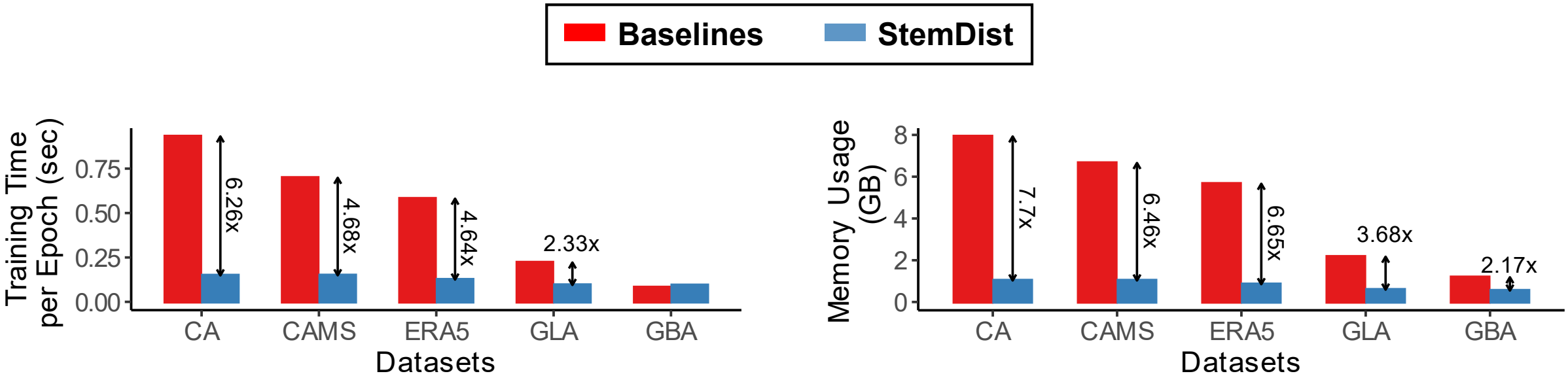
Spatio-temporal ML model



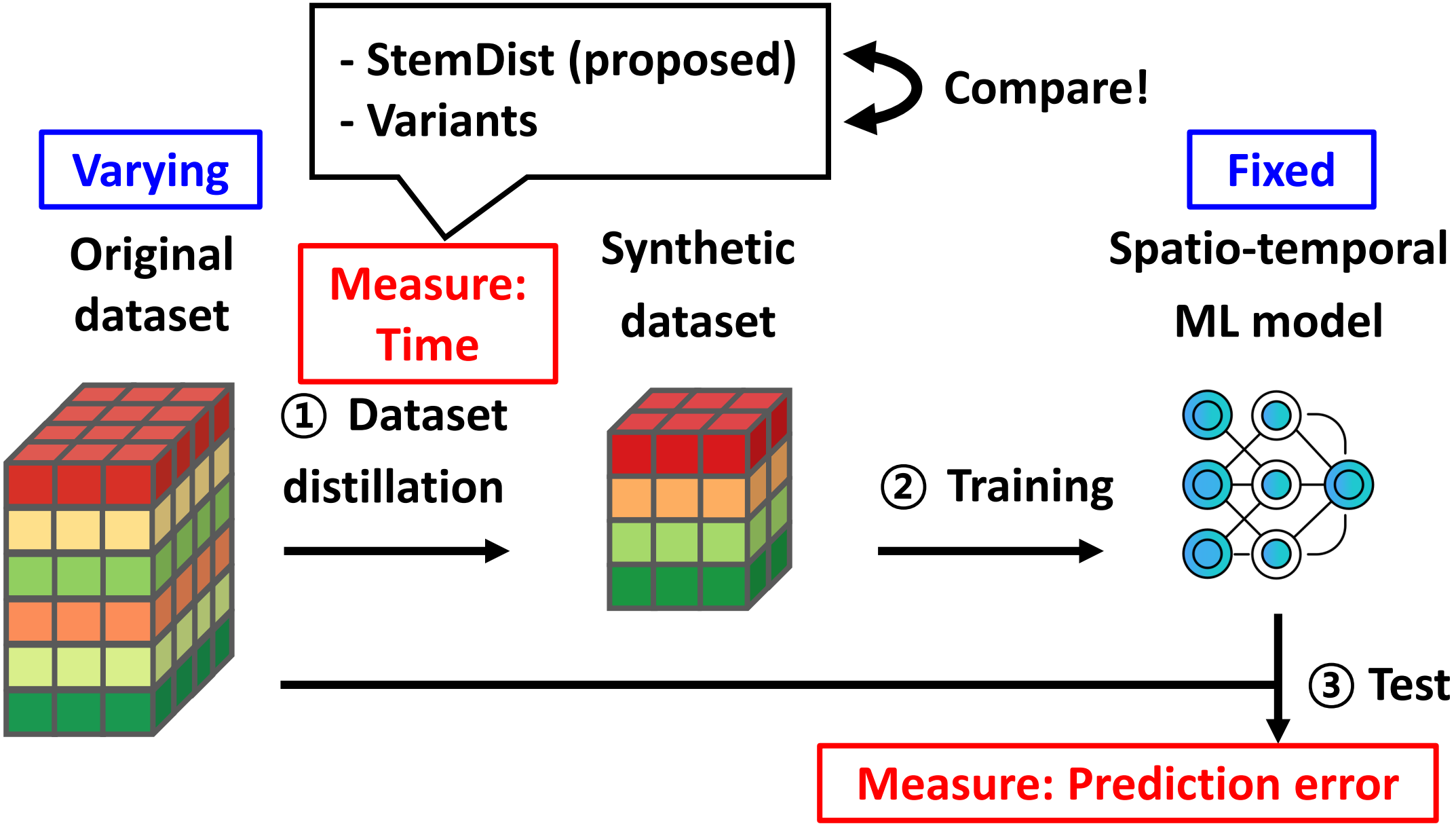
Measure: Time & memory

STemDist Enables Efficient Training

- Training on datasets distilled by STemDist is up to 6.3x faster with up to 7.7x less memory.
  - Recall that baselines reduce only temporal dimension.
  - Under the same budget, STemDist reduces both temporal and spatial dimensions.

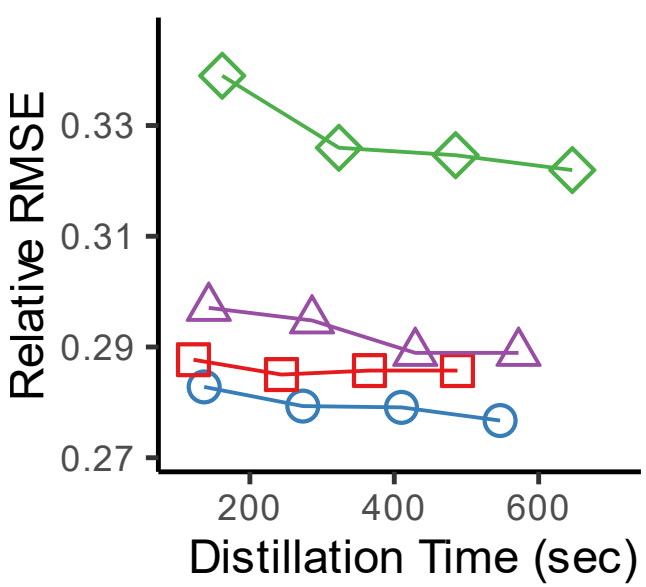


Experimental Settings: Ablation Study

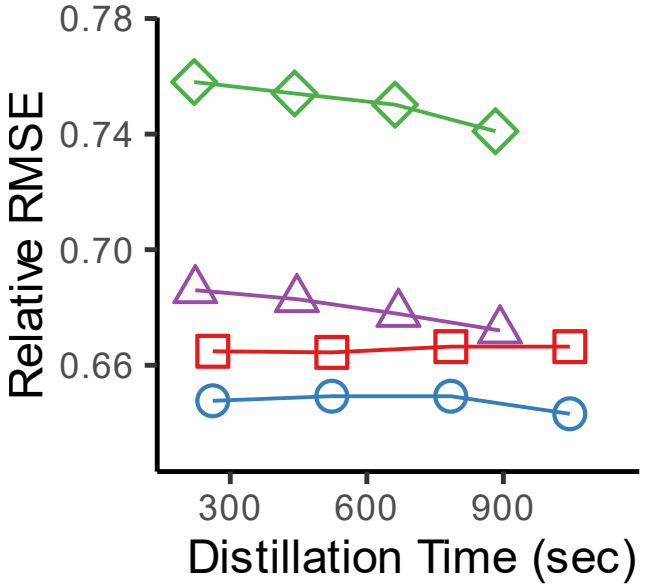


All Components of STemDist are Useful

- Each component of STemDist contributes to effective distillation.



GLA



CAMS

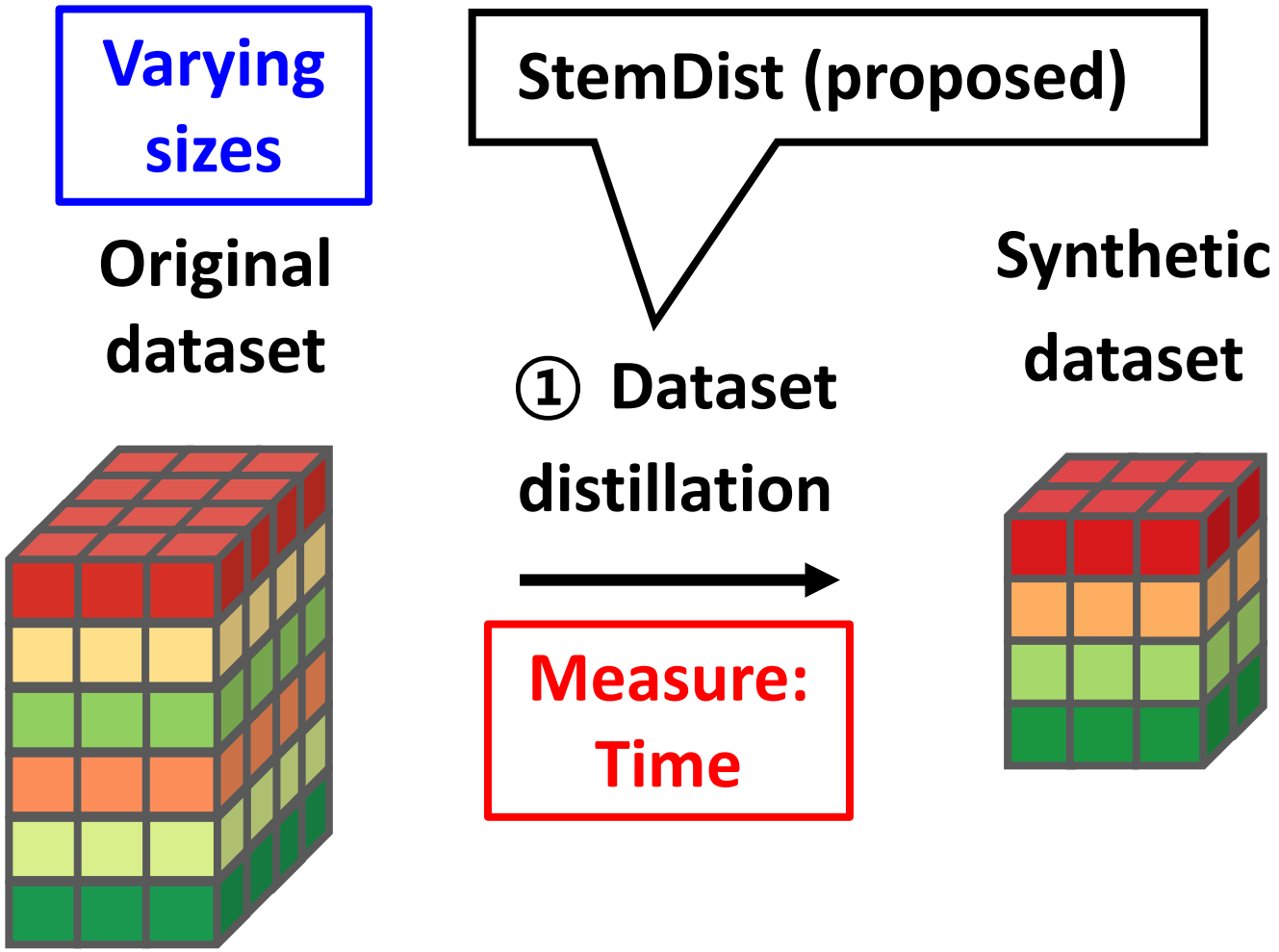
STemDist

STemDist **without subset-based granular distillation.**

STemDist **without subset-based granular distillation and clustering of locations.**

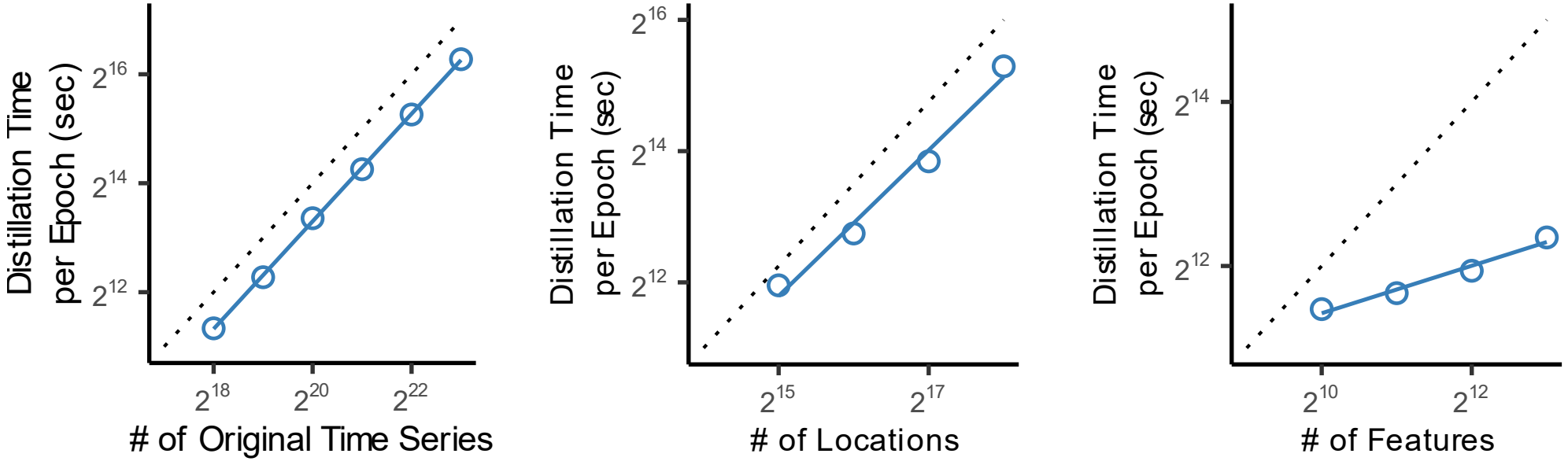
Naïve gradient matching.

Experimental Settings: Scalability Analysis



STemDist is Scalable

- The distillation time of STemDist increases (sub)linearly with dataset size.
  - (a) the number of time series.
  - (b) the number of locations.
  - (c) the number of features.



Outline

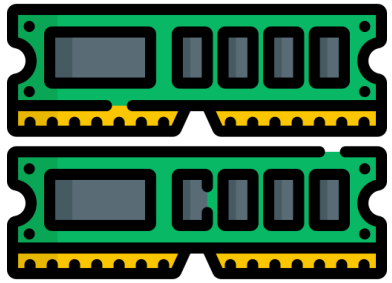
1. Introduction.
2. Proposed method.
3. Experiments.
4. Conclusion.

Conclusion

- We propose **STemDist**, a novel dataset distillation method for spatio-temporal time series.
- The main idea of STemDist is bi-dimensional compression, with three core components: (1) location encoder, (2) location clustering, and (3) subset-based granular distillation.

Fast model training

Memory-efficient model training



Effective

ICDE 2026
MONTRÉAL, CANADA

KAIST AI
Kim Jaechul Graduate School

Effective Dataset Distillation for Spatio-Temporal Forecasting with Bi-Dimensional Compression

Paper:

Github:

Taehyung Kwon* Yeonje Choi* Yeongho Kim Kijung Shin