37th Conference on Neural Information Processing Systems (NeurIPS 2023).

SimMTM: A Simple Pre-Training Framework for Masked Time-Series Modeling

Jiaxiang Dong*, Haixu Wu*, Haoran Zhang, Li Zhang, Jianmin Wang, Mingsheng Long✉
School of Software, BNRist, Tsinghua University, China
{djx20,whx20,z-hr20}@mails.tsinghua.edu.cn, {lizhang,jimwang,mingsheng}@tsinghua.edu.cn


Jiaxiang Dong


Haixu Wu


Haoran Zhang


Li Zhang


Jianmin Wang


Mingsheng Long

Time Series In Real World

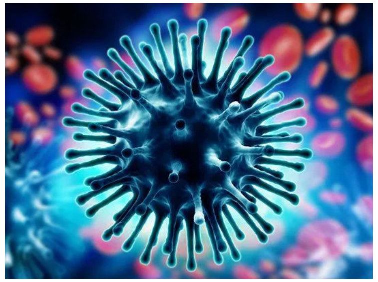
Data

Energy Consumption

Traffic Flow

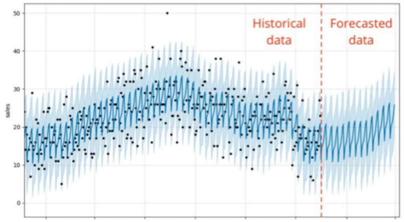
Economic Changes

Weather Variations

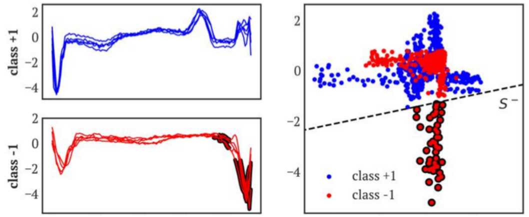


Disease Propagation

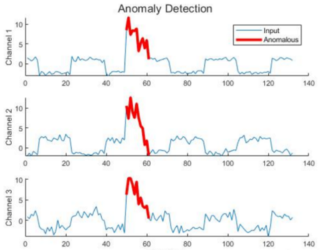
Task



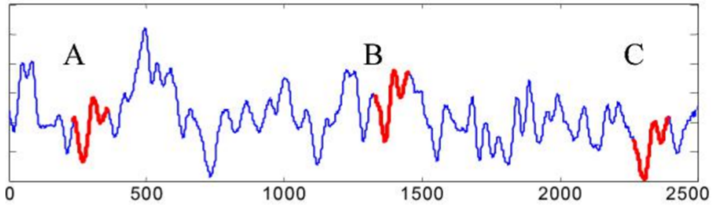
Forecasting



Classification

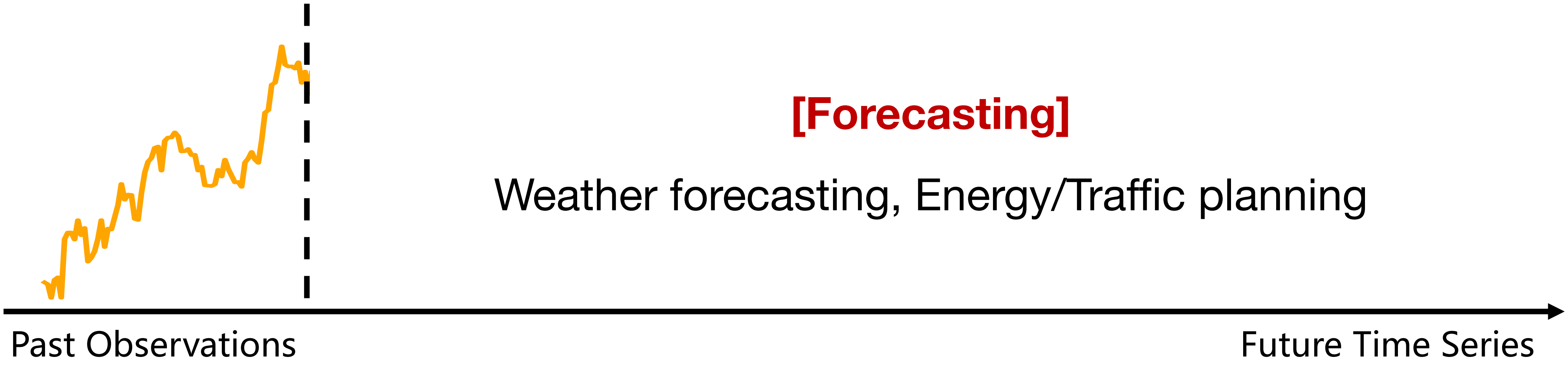


Anomaly Detection



Imputation

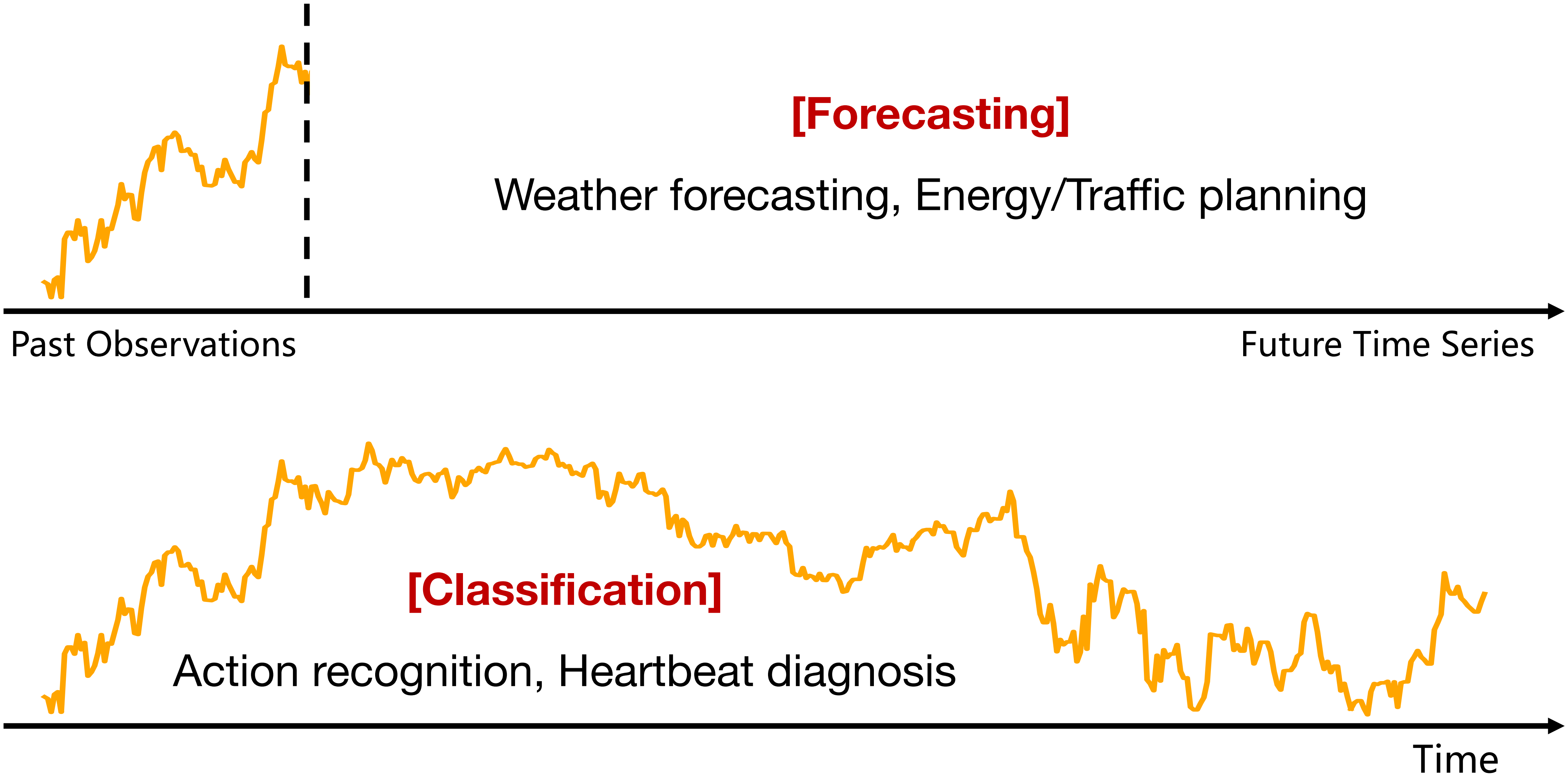
Time Series Analysis



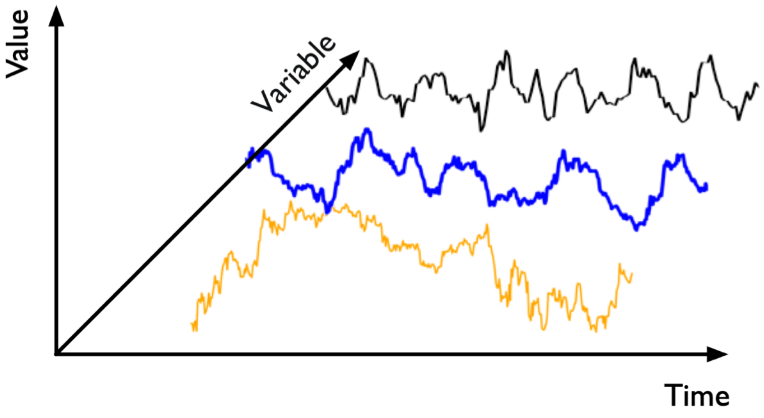
[Forecasting]

Weather forecasting, Energy/Traffic planning

Time Series Analysis

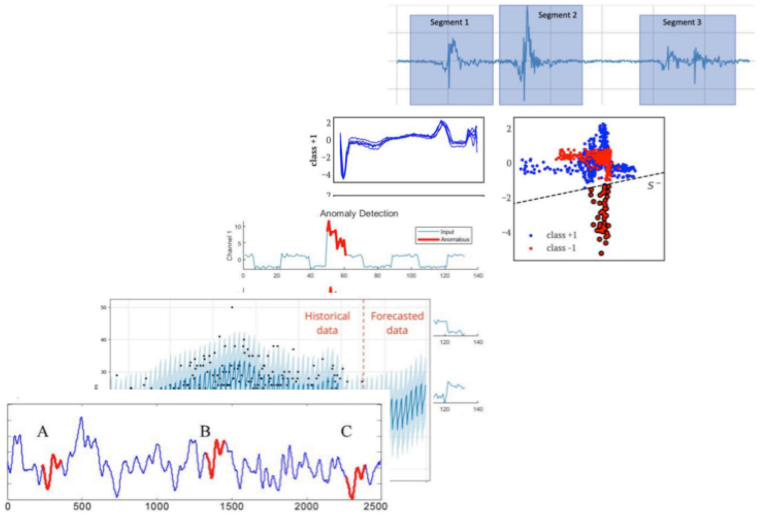


Pre-training and Fine-tuning in Time Series



Pre-training

Fine-tuning



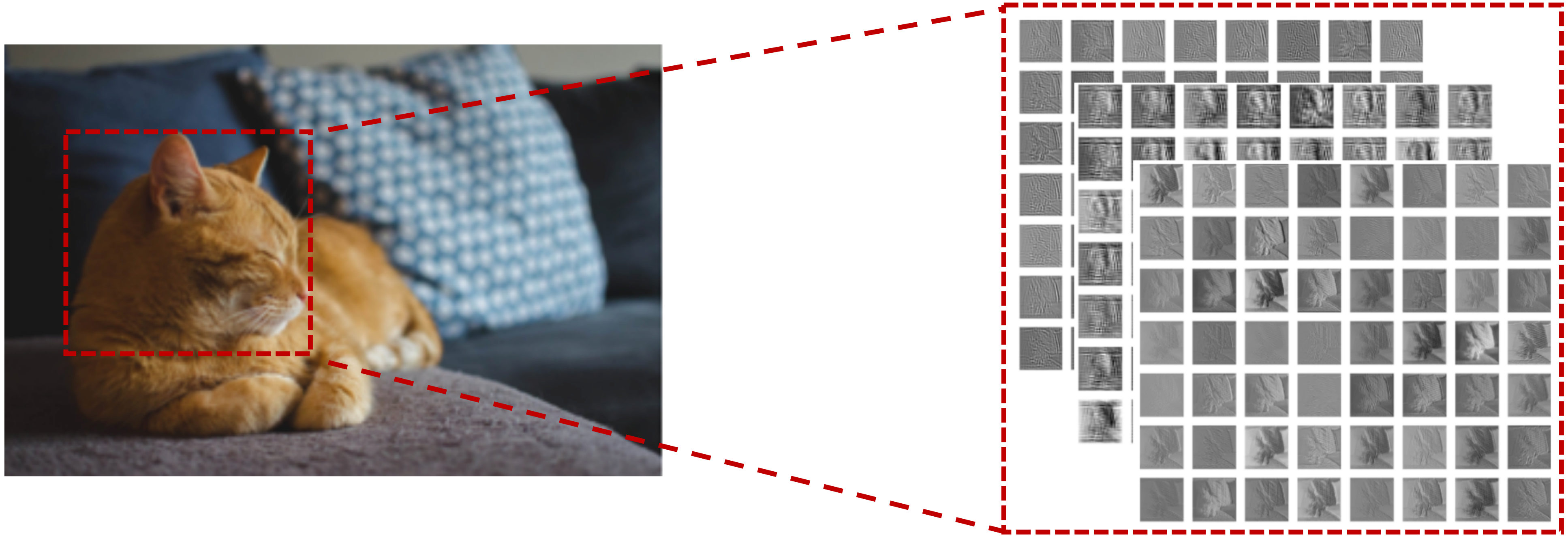
Large-scale time series data

Diversified time series analysis tasks

① Use the model as the carrier of knowledge.

② Learn transferable temporal representations.

Differences among Image, Language and Time Series in Pre-training



Basic visual element

Pre-training is important in *time series* domain.

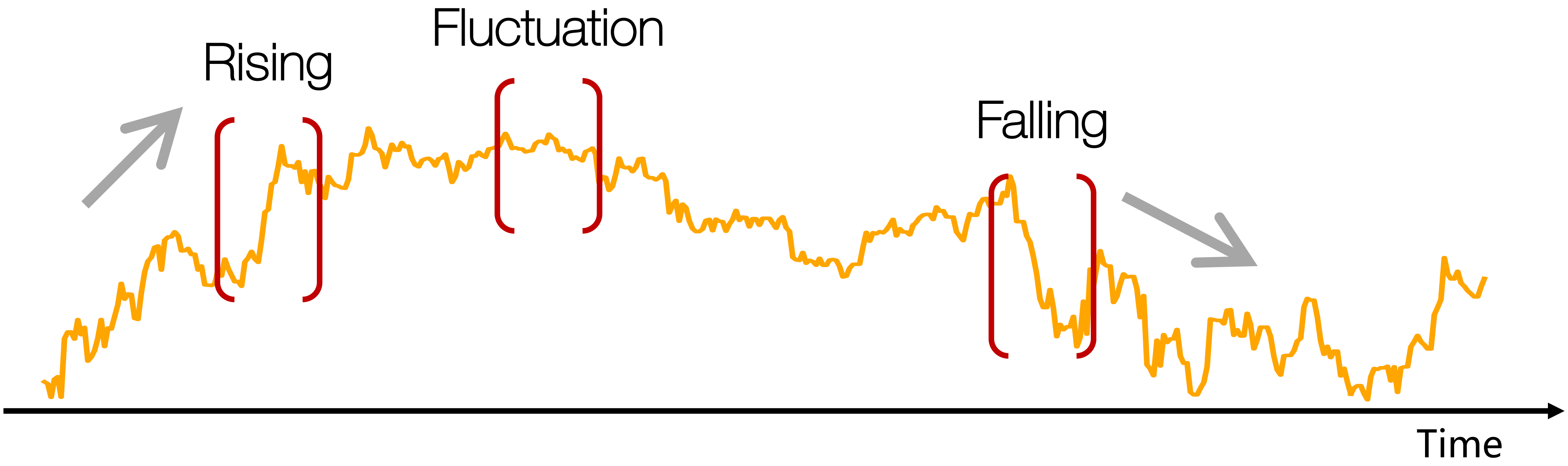
Semantic association & semantic information

Time series is a series of data points indexed (or listed or graphed) in time order.

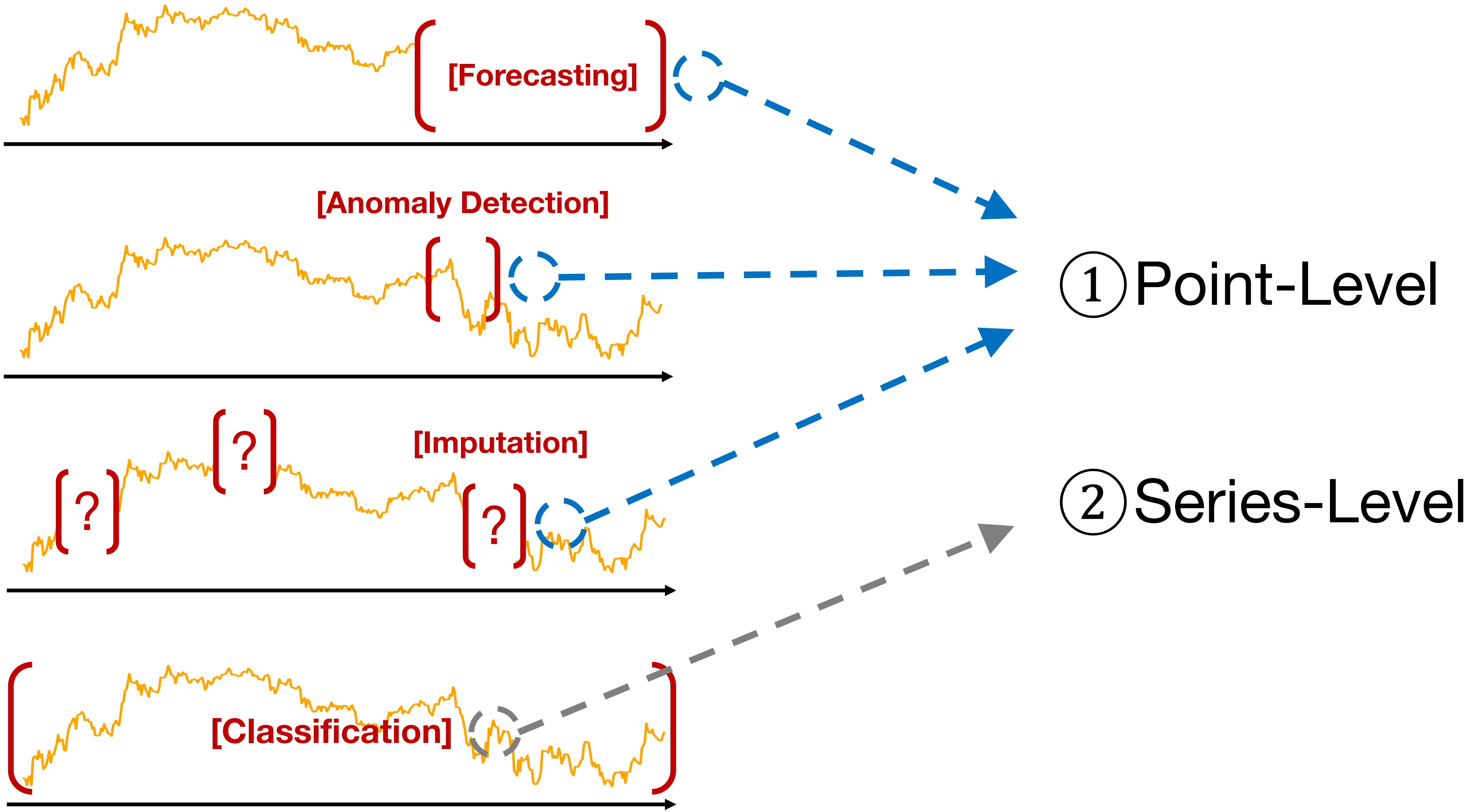
Temporal Variations Modeling in Time Series

More information of time series is in **temporal variations**,

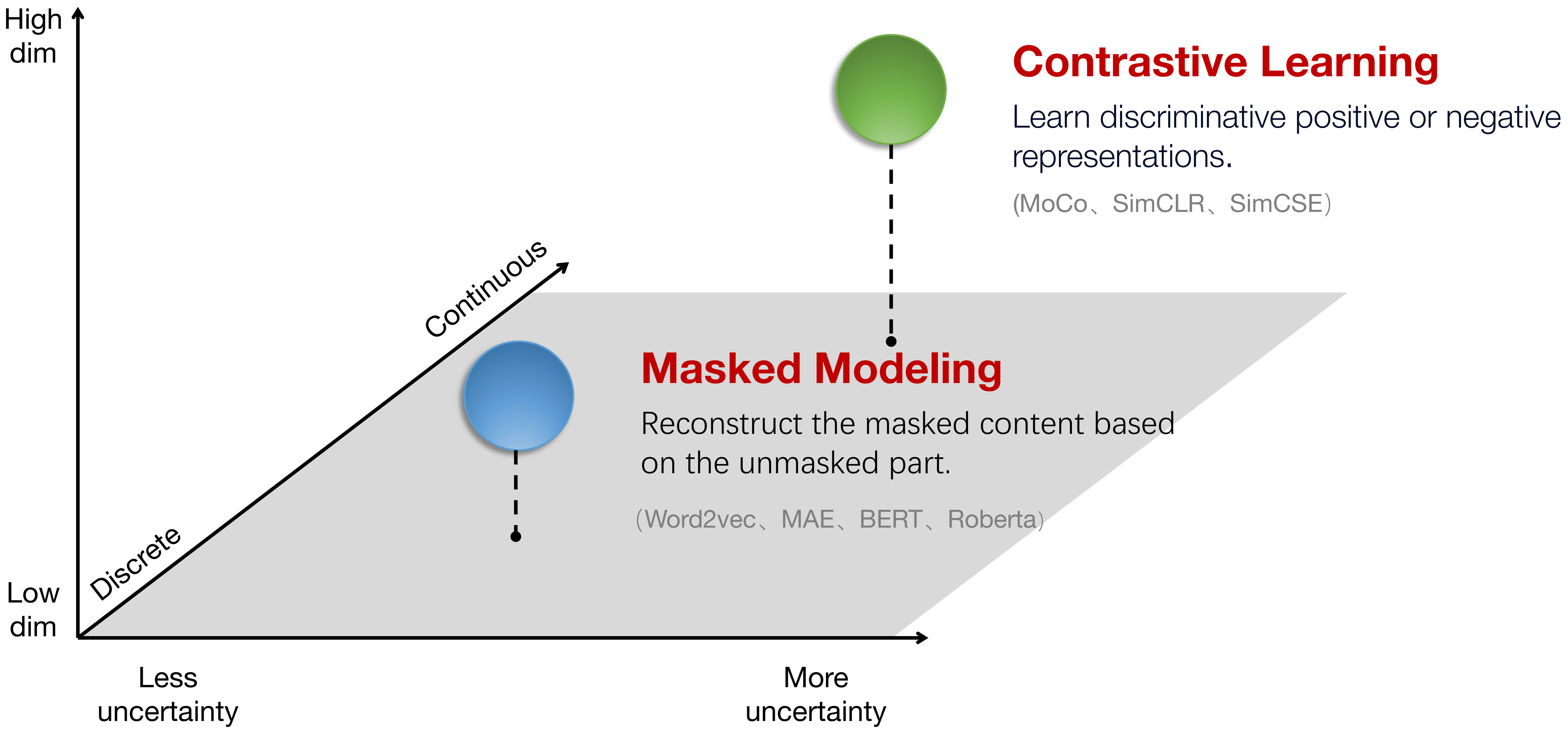
such as continuity, periodicity, trend and etc.



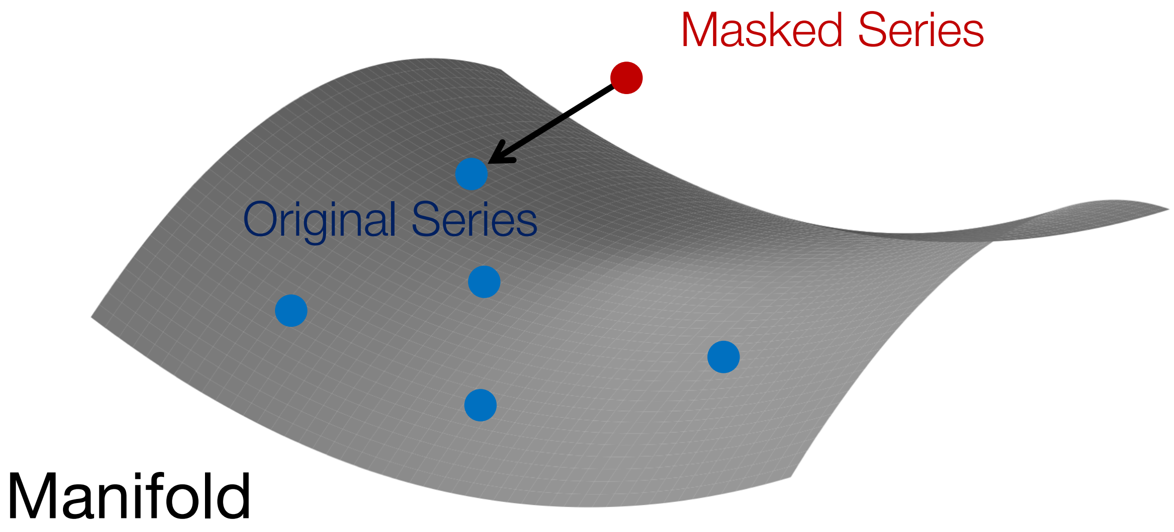
Different Tasks Need Different Level Representation



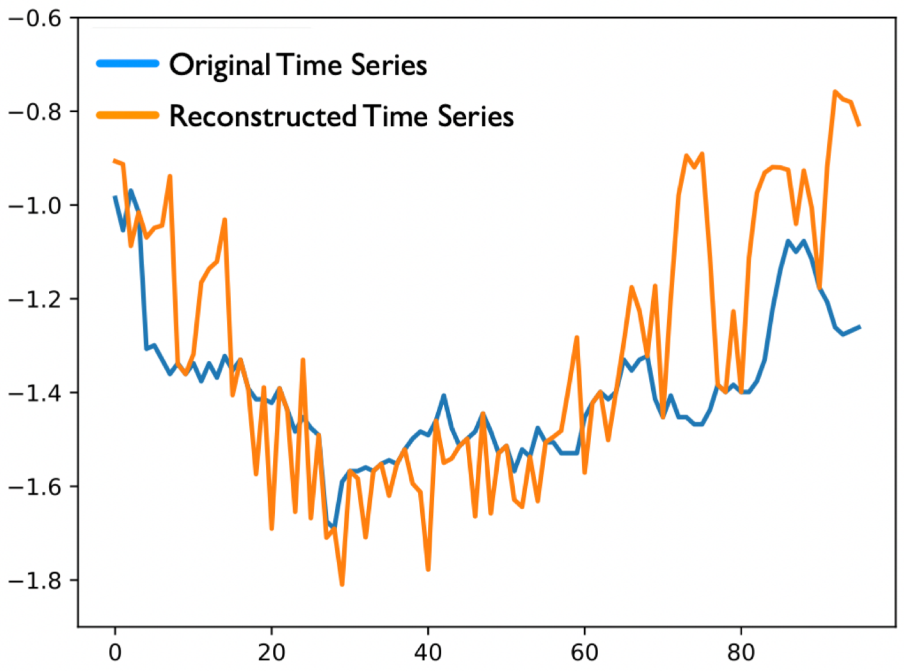
Pre-training Methods in CV and NLP



Canonical Masked Modeling



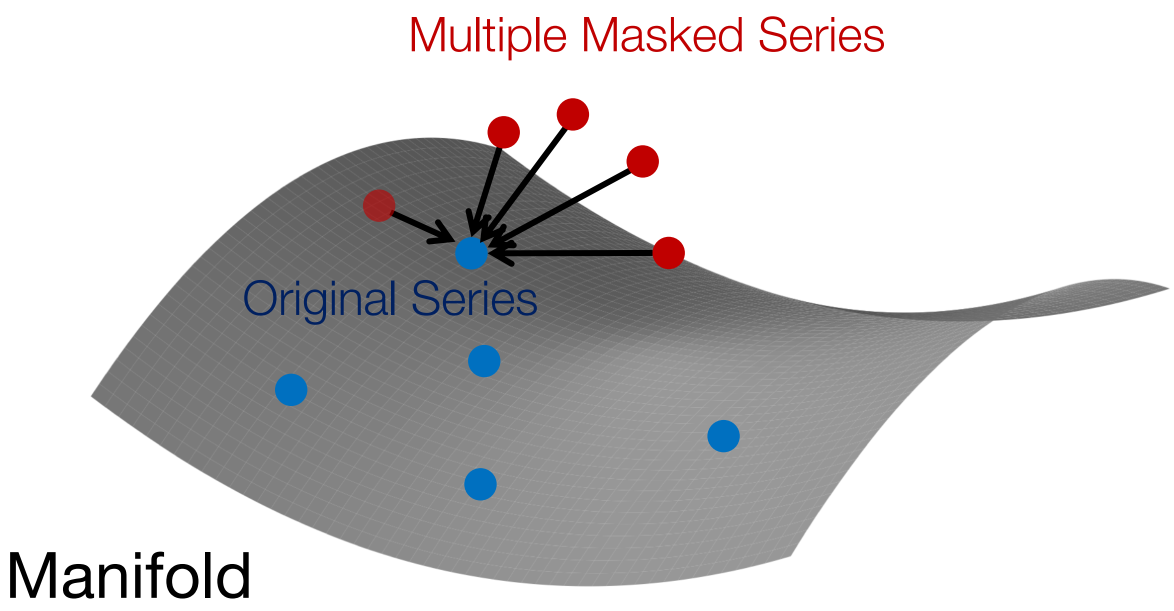
Difficult to Reconstruct



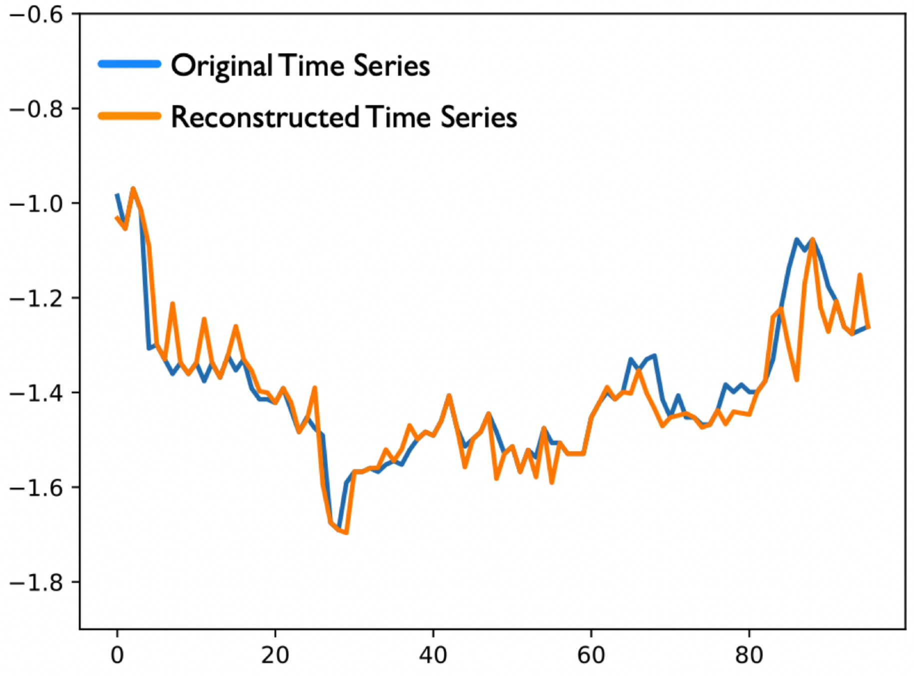
✓ Direct Reconstruction

Directly masking a portion of time points will seriously ruin the temporal variations of the original time series.

Multiple Masked Modeling



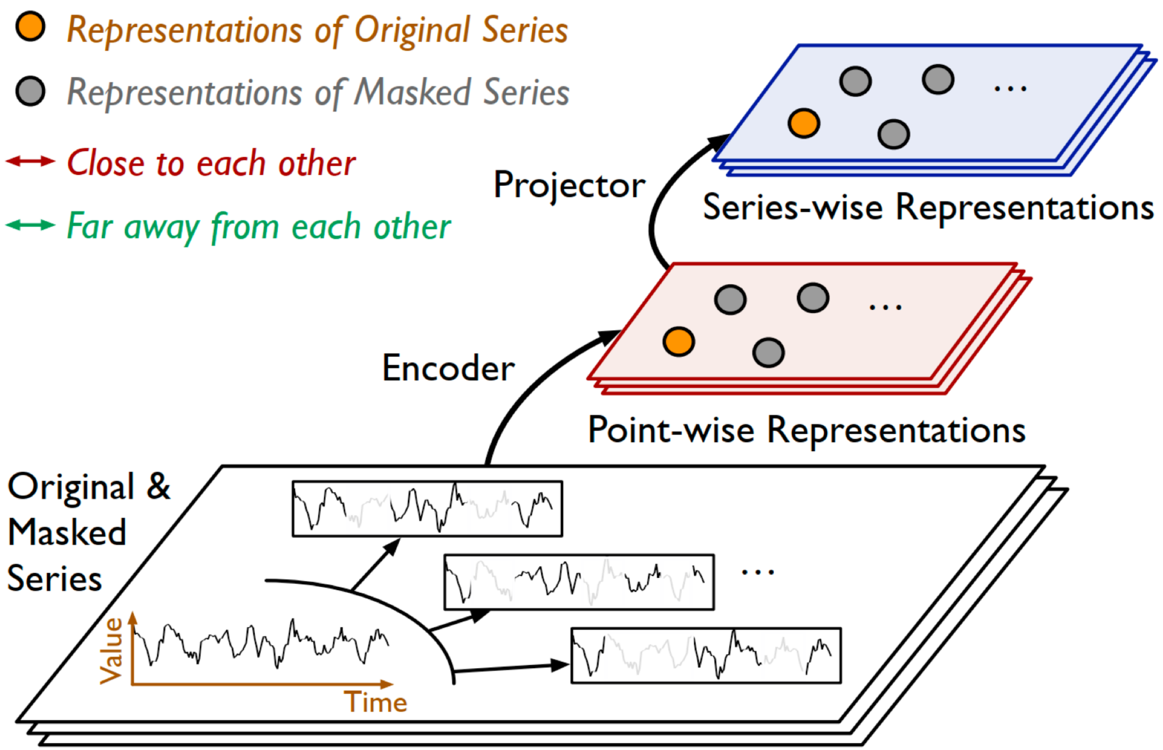
Benefit Masked Modeling



✓ Neighborhood Aggregation

Multiple randomly masked series will complement each other.

Overall design of SimMTM

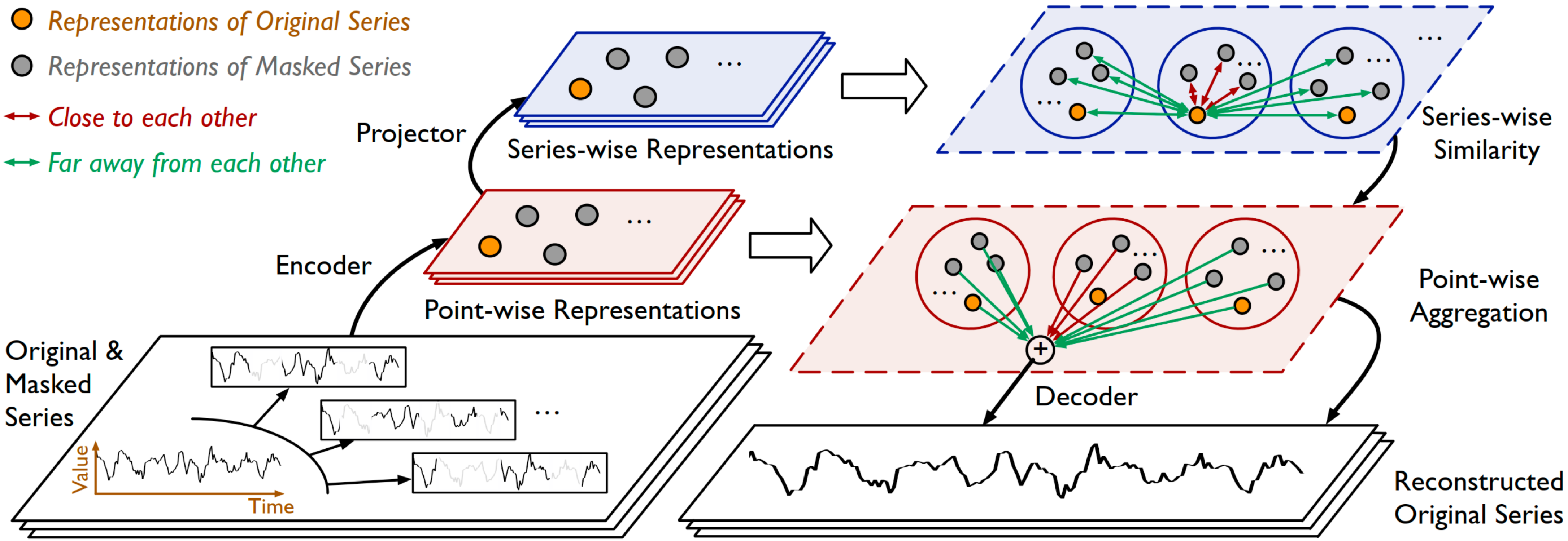


Generate original & masked series representations.

① Point-wise Representations

② Series-wise Representations

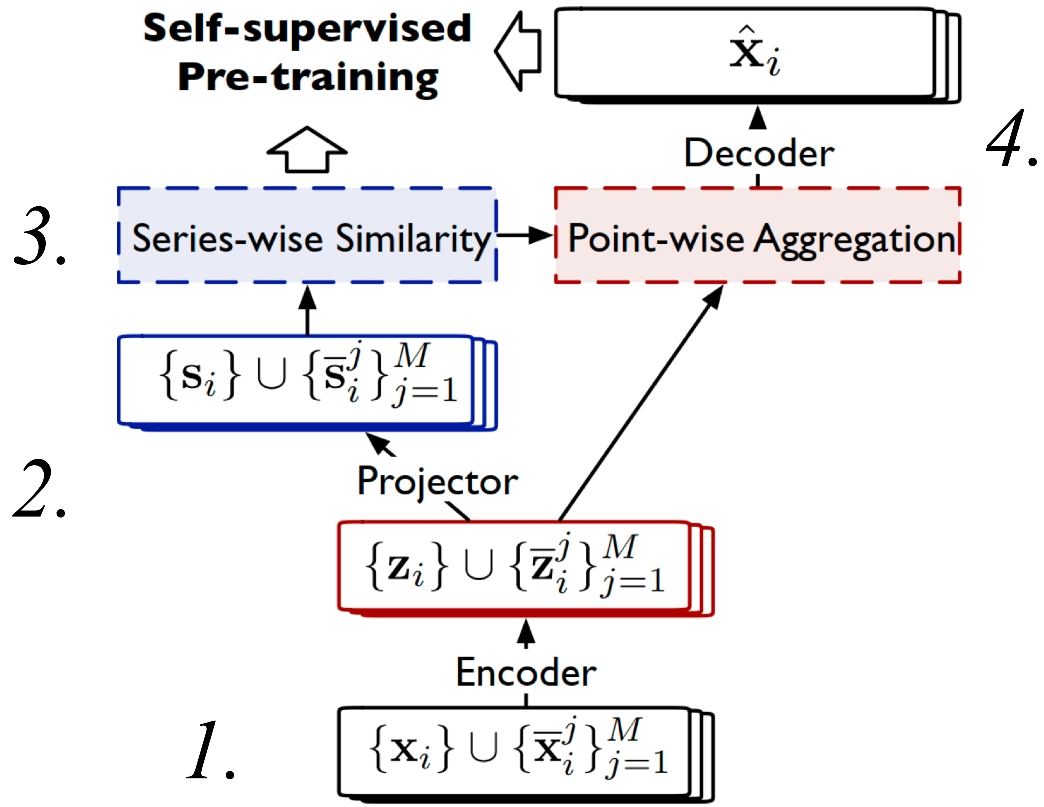
Overall design of SimMTM



① Series-wise Similarity ② Point-wise Aggregation

Multiple masked series complete each other and adaptive aggregate weight.

The Reconstruction Process of SimMTM



① Masking

$$\{\overline{\mathbf{x}}_i^j\}_{j=1}^M = \mathrm{Mask}_r(\mathbf{x}_i), \quad \mathcal{X} = \bigcup_{i=1}^N \left(\{\mathbf{x}_i\} \cup \{\overline{\mathbf{x}}_i^j\}_{j=1}^M\right).$$

② Representation Learning

$$\mathcal{Z} = \bigcup_{i=1}^N \left(\{\mathbf{z}_i\} \cup \{\overline{\mathbf{z}}_i^j\}_{j=1}^M\right) = \mathrm{Enocder}(\mathcal{X}),$$

$$\mathcal{S} = \bigcup_{i=1}^N \left(\{\mathbf{s}_i\} \cup \{\overline{\mathbf{s}}_i^j\}_{j=1}^M\right) = \mathrm{Projector}(\mathcal{Z}),$$

③ Series-wise similarity learning

$$\mathbf{R} = \mathrm{Sim}(\mathcal{S}) \in \mathbb{R}^{D\times D}, D = N \times (M+1),$$

$$\mathbf{R}_{\mathbf{u},\mathbf{v}} = \frac{\mathbf{u}\mathbf{v}^{\mathsf{T}}}{\|\mathbf{u}\|\|\mathbf{v}\|}, \mathbf{u}, \mathbf{v} \in \mathcal{S},$$

④ Point-wise aggregation

$$\widehat{\mathbf{z}}_i = \sum_{\mathbf{s}' \in \mathcal{S}\backslash\{\mathbf{s}_i\}} \frac{\exp(\mathbf{R}_{\mathbf{s}_i,\mathbf{s}'}/\tau)}{\sum_{\mathbf{s}''\in\mathcal{S}\backslash\{\mathbf{s}_i\}} \exp(\mathbf{R}_{\mathbf{s}_i,\mathbf{s}''}/\tau)} \mathbf{z}',$$

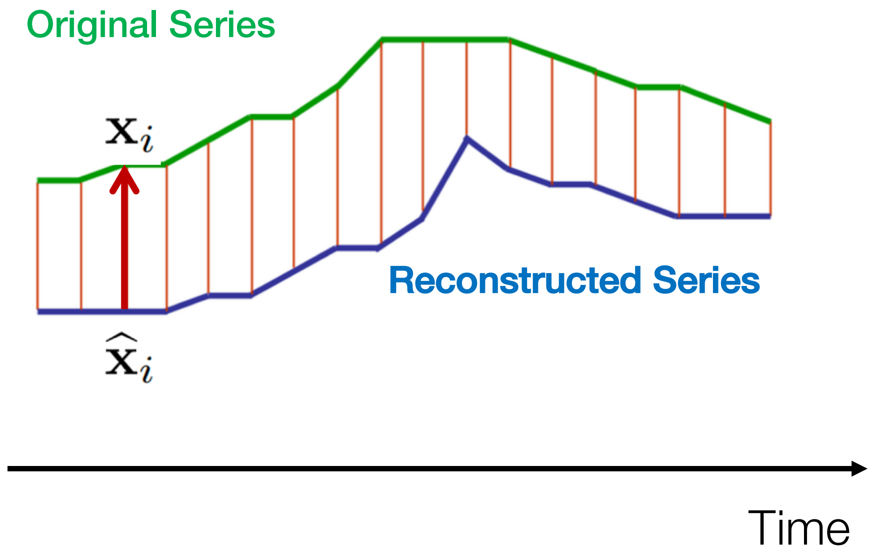
$$\{\widehat{\mathbf{x}}_i\}_{i=1}^N = \mathrm{Decoder}(\{\widehat{\mathbf{z}}_i\}_{i=1}^N),$$

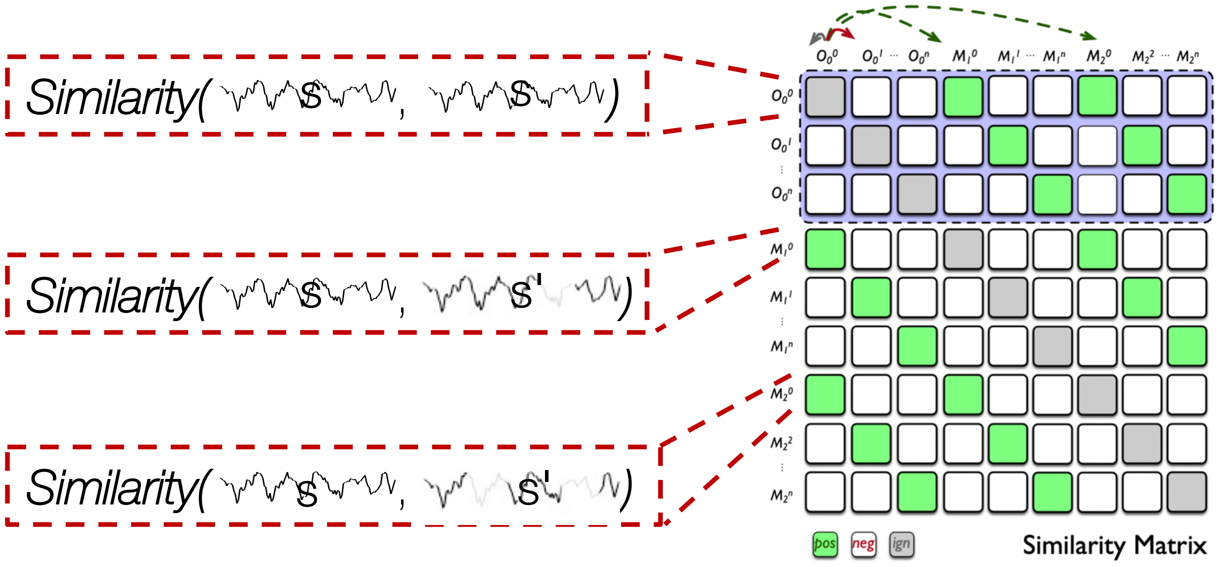
SimMTM：A Simple Time Series Self-supervised Pre-training

$$\boxed{\min_{\Theta} \mathcal{L}_{\text{reconstruction}} + \lambda \mathcal{L}_{\text{constraint}},}$$

① $$\mathcal{L}_{\text{reconstruction}} = \sum_{i=1}^{N} \|\mathbf{x}_i - \widehat{\mathbf{x}}_i\|_2^2.$$

② $$\mathcal{L}_{\text{constraint}} = -\sum_{\mathbf{s}\in\mathcal{S}} \left(\sum_{\mathbf{s}'\in\mathcal{S}^+} \log \frac{\exp(\mathbf{R}_{\mathbf{s},\mathbf{s}'}/\tau)}{\sum_{\mathbf{s}''\in\mathcal{S}\setminus\{\mathbf{s}\}} \exp(\mathbf{R}_{\mathbf{s},\mathbf{s}''}/\tau)} \right),$$





pos Positive pairs

$$\left(\{\mathbf{s}_i\} \cup \{\overline{\mathbf{s}}_i^j\}_{j=1}^M\right) \sim \left(\{\mathbf{s}_i\} \cup \{\overline{\mathbf{s}}_i^j\}_{j=1}^M\right),$$

neg Negative pairs

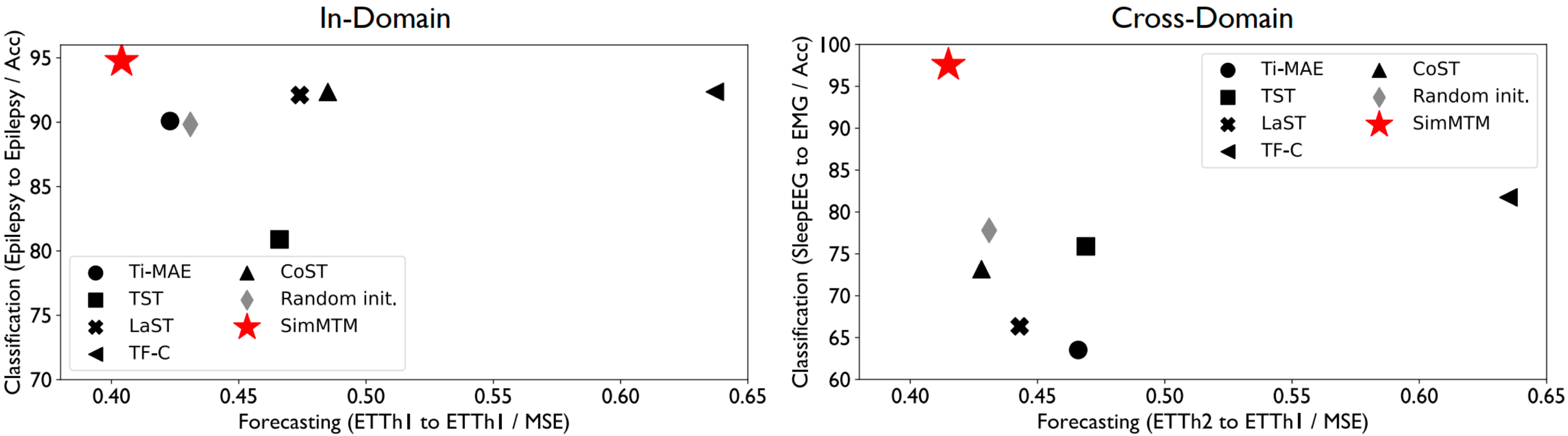
$$\left(\{\mathbf{s}_i\} \cup \{\overline{\mathbf{s}}_i^j\}_{j=1}^M\right) \nsim \left(\{\mathbf{s}_k\} \cup \{\overline{\mathbf{s}}_k^j\}_{j=1}^M\right), i \neq k$$

Experiment: Overall

Tasks	Datasets	Semantic
Forecasting	ETTh1,ETTh2	Electricity
	ETTm1,ETTm2	Electricity
	Weather	Weather
	Electricity	Electricity
	Traffic	Transportation
Classification	SleepEEG	EEG
	Epilepsy	EEG
	FD-B	Faulty Detection
	Gesture	Hand Movement
	EMG	Muscle Responses

- ✓ Two typical time series analysis tasks: **Forecasting and Classification.**
- ✓ Under multiple experiment settings: **In- and Cross domain， Unified and Official implementation Encoder.**
- ✓ Compared to **6 advanced baselines in 12 databases.**

Experiment: Overall



SimMTM outperforms other baselines significantly in all settings!

Experiment: Forecasting

Models	SimMTM		Random init.		Ti-MAE [21]		TST [56]		LaST [42]		TF-C [57]		CoST [46]		TS2Vec [55]	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.409**	**0.428**	0.431	0.448	0.423	0.446	0.466	0.462	0.474	0.461	0.637	0.638	0.485	0.472	0.446	0.456
ETTh2	**0.353**	**0.390**	0.395	0.427	0.380	0.386	0.404	0.421	0.499	0.497	0.398	0.398	0.399	0.427	0.417	0.468
ETTm1	**0.348**	**0.385**	0.356	0.387	0.366	0.391										
ETTm2	**0.263**	**0.320**	0.279	0.336	0.267	0.325										
Weather	**0.230**	0.271	0.239	0.275	0.234	**0.265**										
Electricity	**0.162**	**0.256**	0.212	0.300	0.205	0.296										
Traffic	**0.392**	**0.264**	0.490	0.316	0.475	0.310										
Avg	**0.308**	**0.331**	0.343	0.356	0.336	0.346										

Models	SimMTM		Ti-MAE [21]		TST [56]		LaST [42]		TF-C [57]		CoST [46]		TS2Vec [55]	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2 → ETTh1	**0.415**	**0.430**	0.466	0.456	0.469	0.459	0.443	0.471	0.635	0.634	0.428	0.433	0.517	0.486
ETTm1 → ETTh1	**0.422**	**0.430**	0.495	0.469	0.475	0.463	0.426	0.441	0.700	0.702	0.620	0.541	0.484	0.482
ETTm2 → ETTh1	**0.428**	**0.441**	0.464	0.456	0.453	0.450	0.503	0.507	1.091	0.814	0.598	0.548	0.616	0.550
Weather → ETTh1	**0.456**	0.467	0.462	0.464	0.465	**0.456**	-	-	-	-	0.518	0.487	0.463	0.460
ETTh1 → ETTm1	**0.346**	**0.384**	0.360	0.390	0.373	0.393	0.353	0.390	0.746	0.652	0.370	0.393	0.699	0.557
ETTh2 → ETTm1	0.365	**0.384**	0.383	0.402	0.391	0.409	0.475	0.489	0.750	0.654	**0.363**	0.387	0.694	0.557
ETTm2 → ETTm1	**0.351**	**0.383**	0.390	0.410	0.382	0.402	0.414	0.464	0.758	0.699	0.385	0.412	0.423	0.420
Weather → ETTm1	**0.350**	**0.383**	0.411	0.423	0.368	0.392	-	-	-	-	0.382	0.403	0.382	0.395
Avg	**0.392**	**0.413**	0.429	0.434	0.422	0.428	0.436	0.460	0.780	0.693	0.458	0.451	0.535	0.488

SimMTM consistently outperforms other pre-training methods for in- and cross-domain settings.

Experiment: Classification

Models	**SimMTM**	Random init.	Ti-MAE [21]	TST[56]	LaST[42]	TF-C[57]	CoST[46]	TS2Vec[55]
Epilepsy → Epilepsy	**94.75**	89.83	90.09	80.89	92.11	93.96	92.35	92.33
SleepEEG → Epilepsy	**95.49**	89.83	73.45	82.89	86.46	94.95	93.66	94.46
SleepEEG → FD-B	**69.40**	47.36	70.88	65.57	46.67	69.38	54.82	60.74
SleepEEG → Gesture	**80.00**	42.19	65.54	75.12	64.17	76.42	73.33	73.33
SleepEEG → EMG	**97.56**	77.80	63.52	75.89	66.34	81.74	73.17	80.92
Avg	**87.44**	69.40	72.70	76.07	71.15	83.29	77.47	80.36

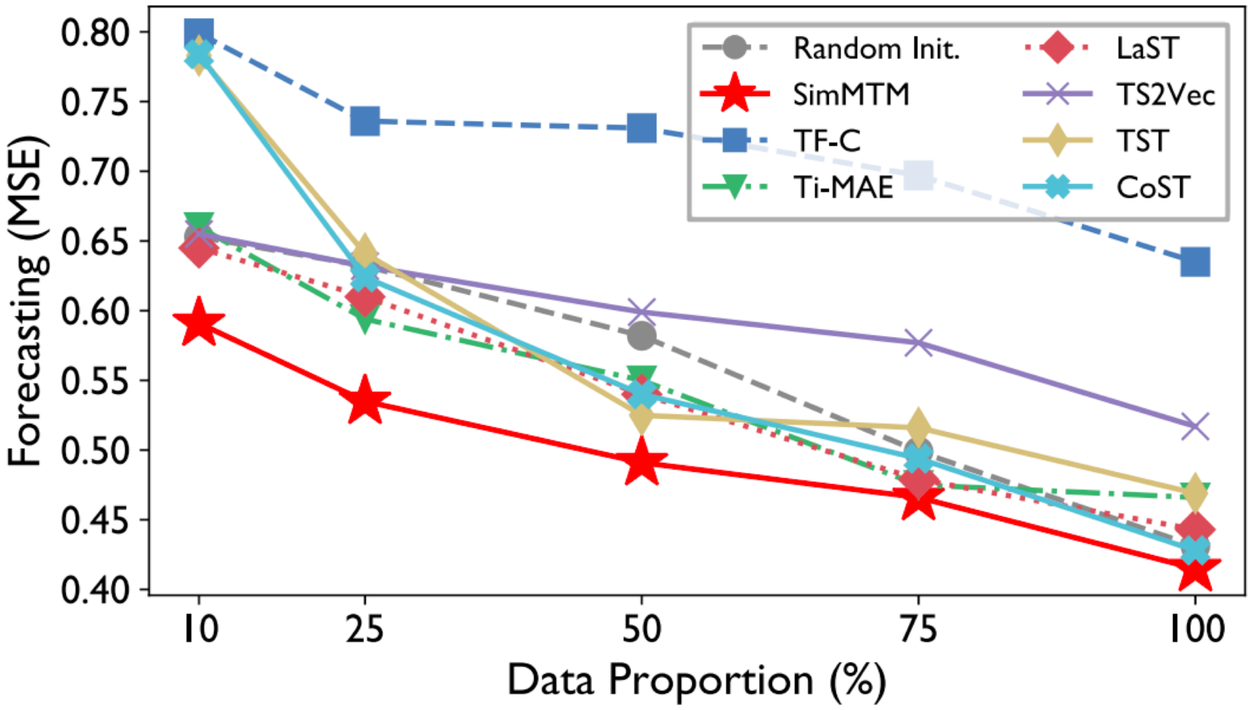
SimMTM surpasses all advanced time series pre-training baselines.

Model Generality

Dataset	ETTh1		ETTh2		ETTm1		ETTm2	
Model	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Transformer [39]	1.088	0.836	4.103	1.612	0.901	0.704	1.624	0.901
+ SimMTM	**0.927**	**0.761**	**3.498**	**1.487**	**0.809**	**0.663**	**1.322**	**0.808**
Autoformer [47]	0.573	0.573	0.550	0.559	0.615	0.528	0.324	0.368
+ SimMTM	**0.561**	**0.568**	**0.543**	**0.555**	**0.553**	**0.505**	**0.315**	**0.360**
NS Transformer [24]	0.570	0.537	0.526	0.516	0.481	0.456	0.306	0.347
+ SimMTM	**0.543**	**0.527**	**0.493**	**0.514**	**0.431**	**0.455**	**0.301**	**0.345**
PatchTST [26]	0.417	0.431	0.331	0.379	0.352	0.382	0.258	0.317
+ Sub-series Masking	0.430↓	0.445↓	0.355↓	0.394↓	**0.341**	0.379	0.258	0.318↓
+ SimMTM	**0.409**	**0.428**	**0.329**	0.379	0.348	**0.378**	**0.254**	**0.313**

SimMTM can consistently improve the forecasting performance of diverse base models.

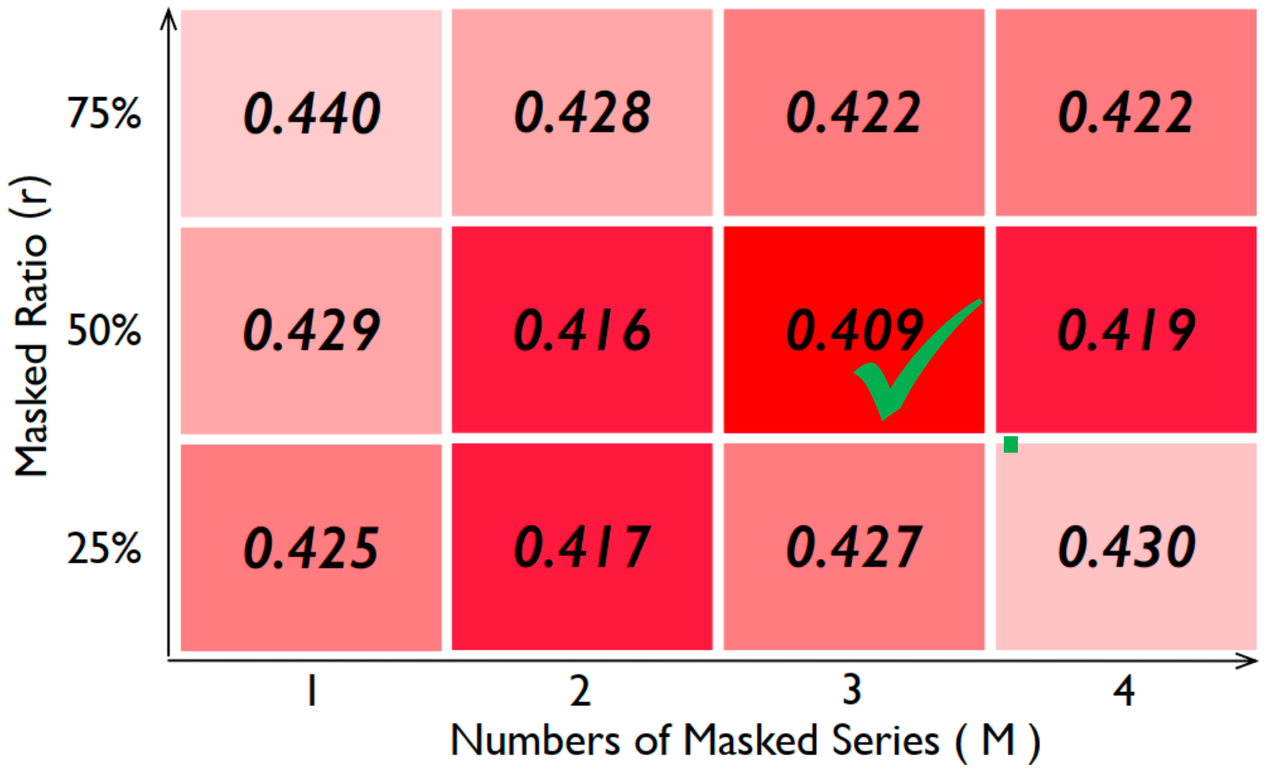
Limited Fine-tuning Data Scenarios



We pre-train a model and fine-tune it with different choices for the remaining proportions of training data.

✓ **SimMTM achieves significant performance gains in different data proportions.**

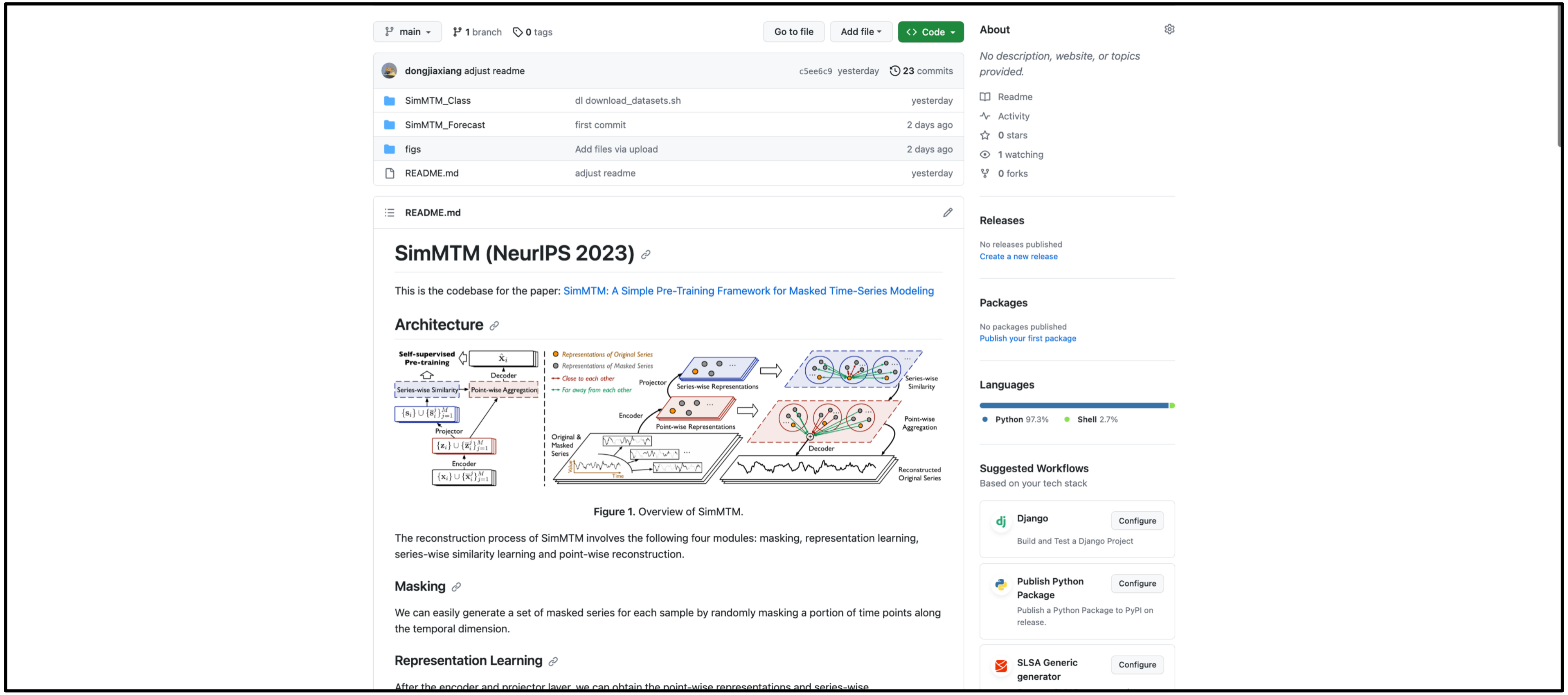
Masking Strategy



We explore the potential relationship between the masked ratio and the number of masked series used for reconstruction.

✓ **Choosing a reasonable balance between the masked ratio and the reconstructed numbers is critical when using SimMTM.**

Open Source



Code is available at https://github.com/thuml/SimMTM

Thank You!

djx20@mails.tsinghua.edu.cn

长按关注，获取最新资讯