Deep Learning Embeddings for Data Series Similarity Search

Qitong Wang

Université de Paris, LIPADE

Themis Palpanas

Université de Paris, LIPADE
French University Institute (IUF)

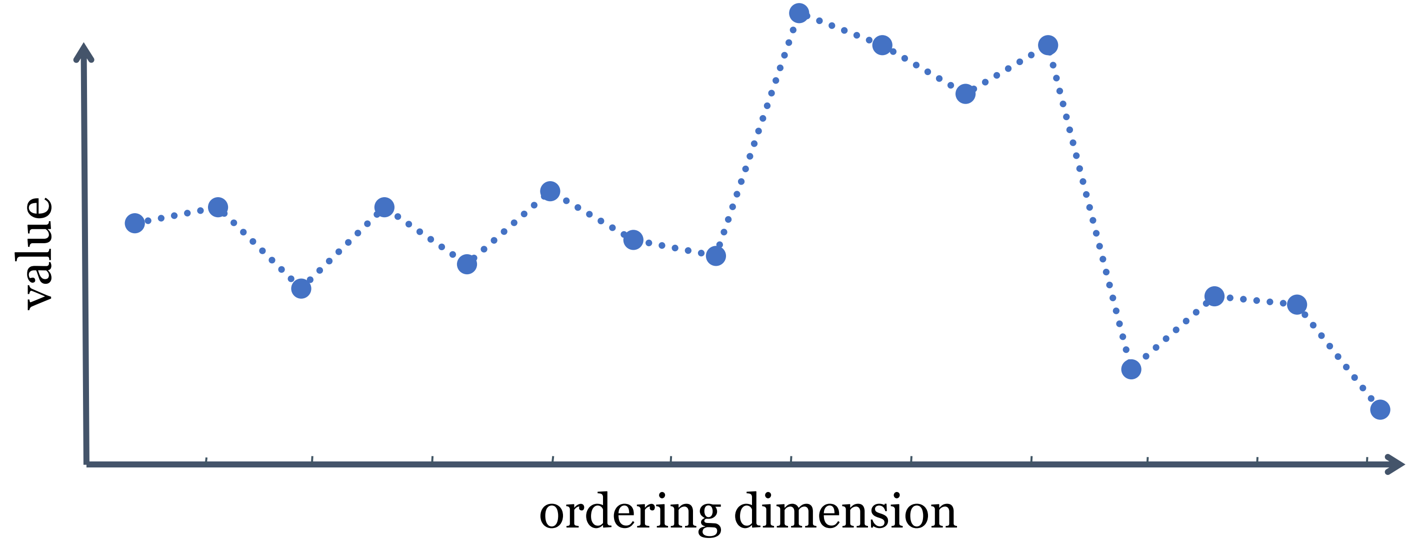
ACM SIGKDD 2021, Singapore

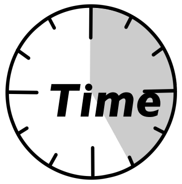
Université de Paris

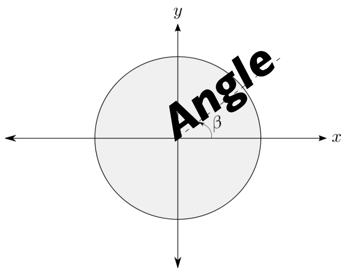
institut universitaire de France

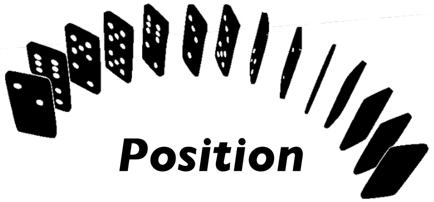
Data Series

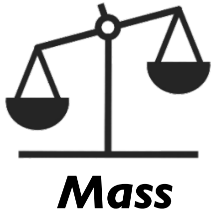
- Sequence of points ordered along some dimension

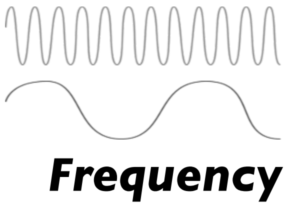






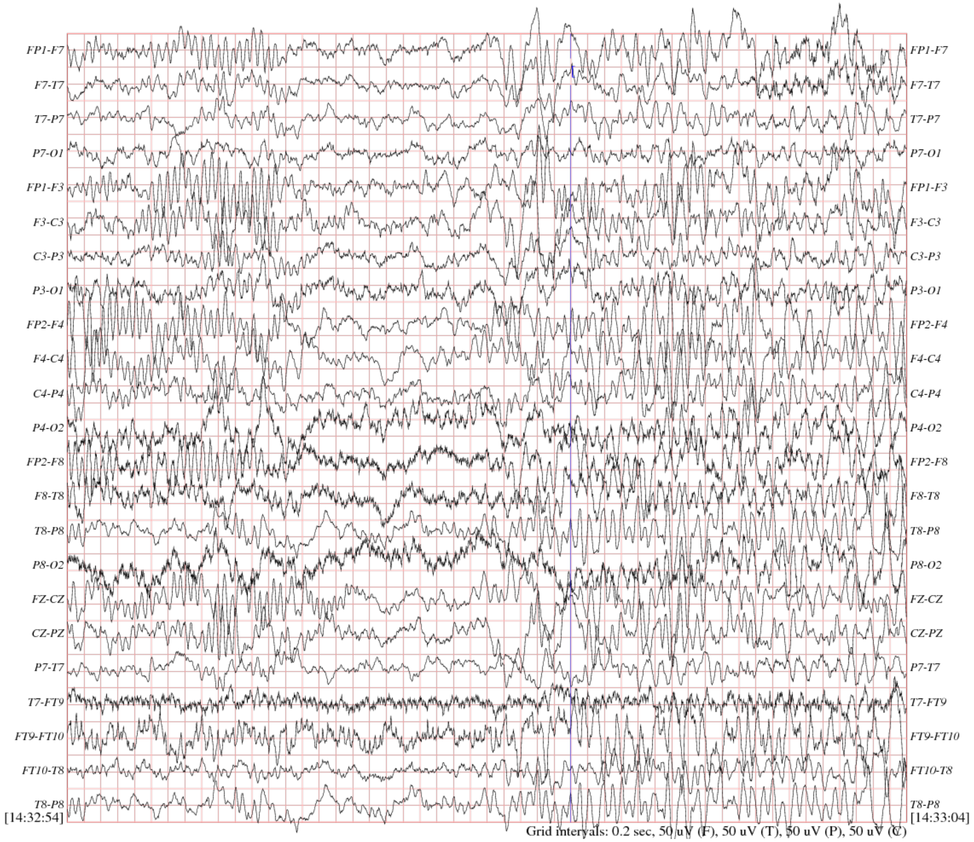




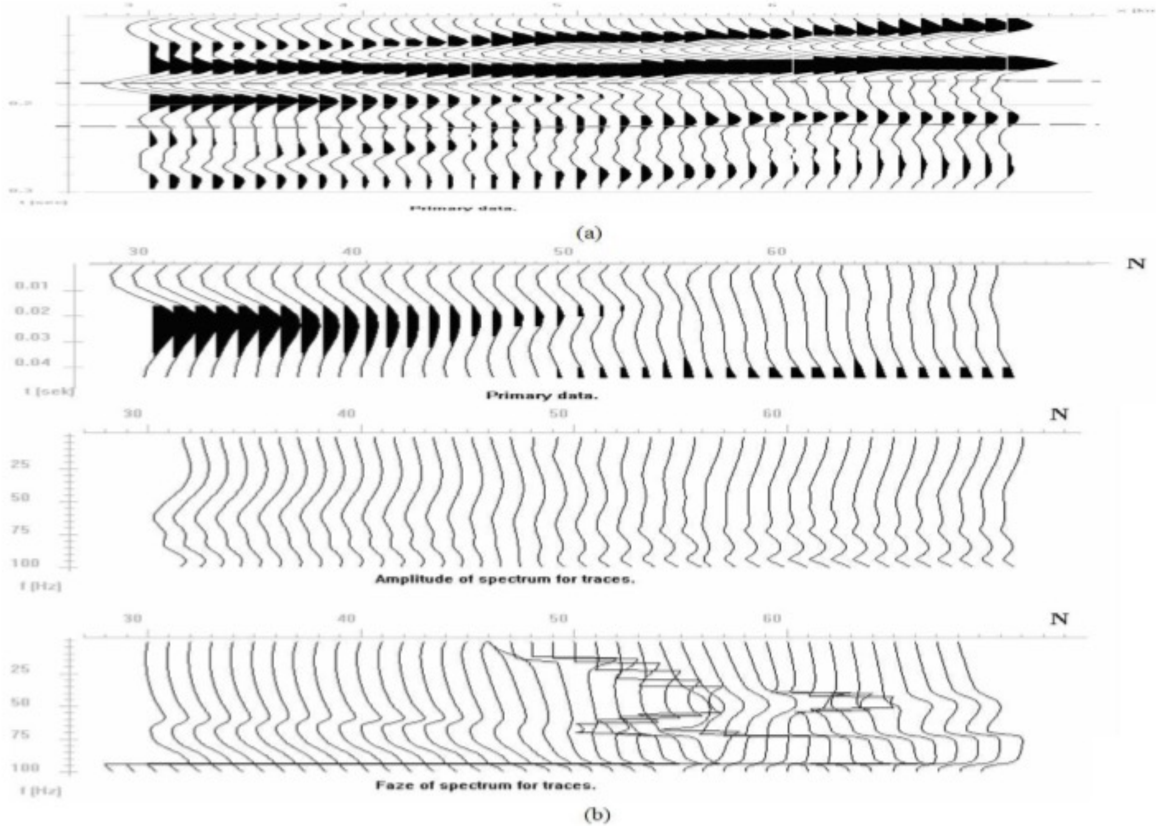


Data Series from Various Domains

Electroencephalography (EEG) Data[1]

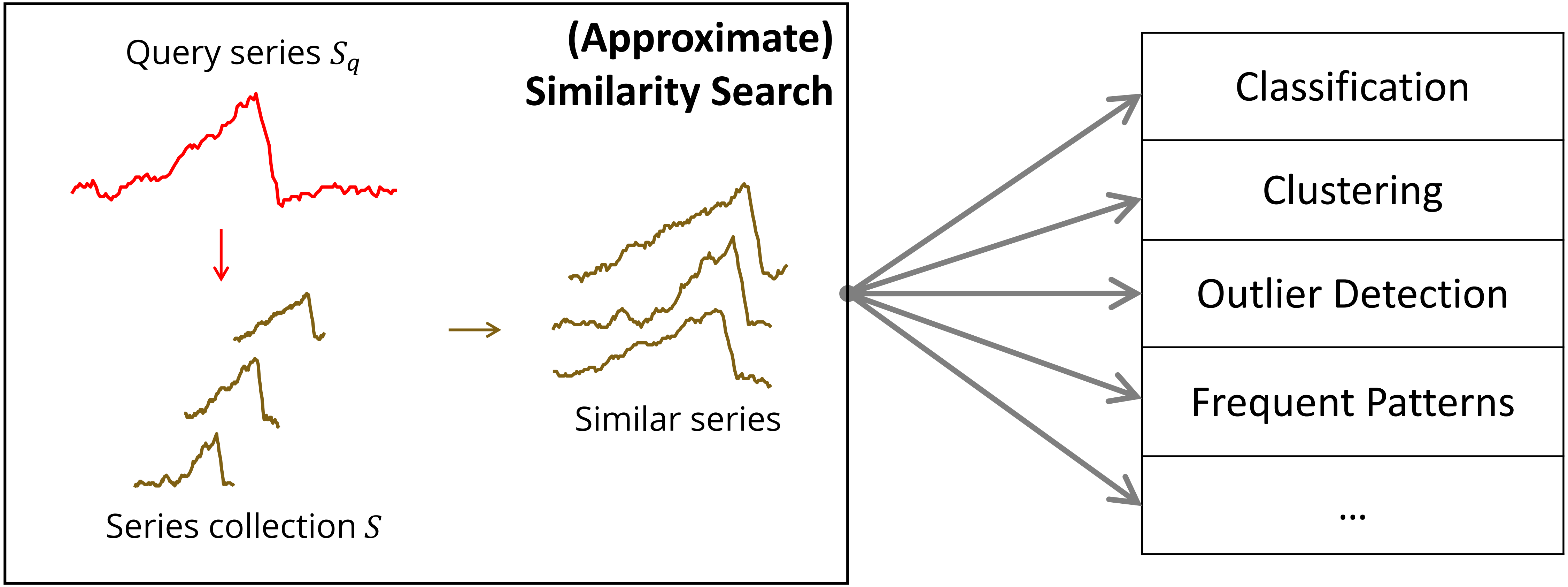


Seismic Data[2]



1. Ali Shoeb. Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment. PhD Thesis, MIT, 2009.
2. Phase and polarity assessment of seismic data. https://wiki.seg.org/wiki/Phase_and_polarity_assessment_of_seismic_data, fetched June 22, 2021.

Data Series Similarity Search

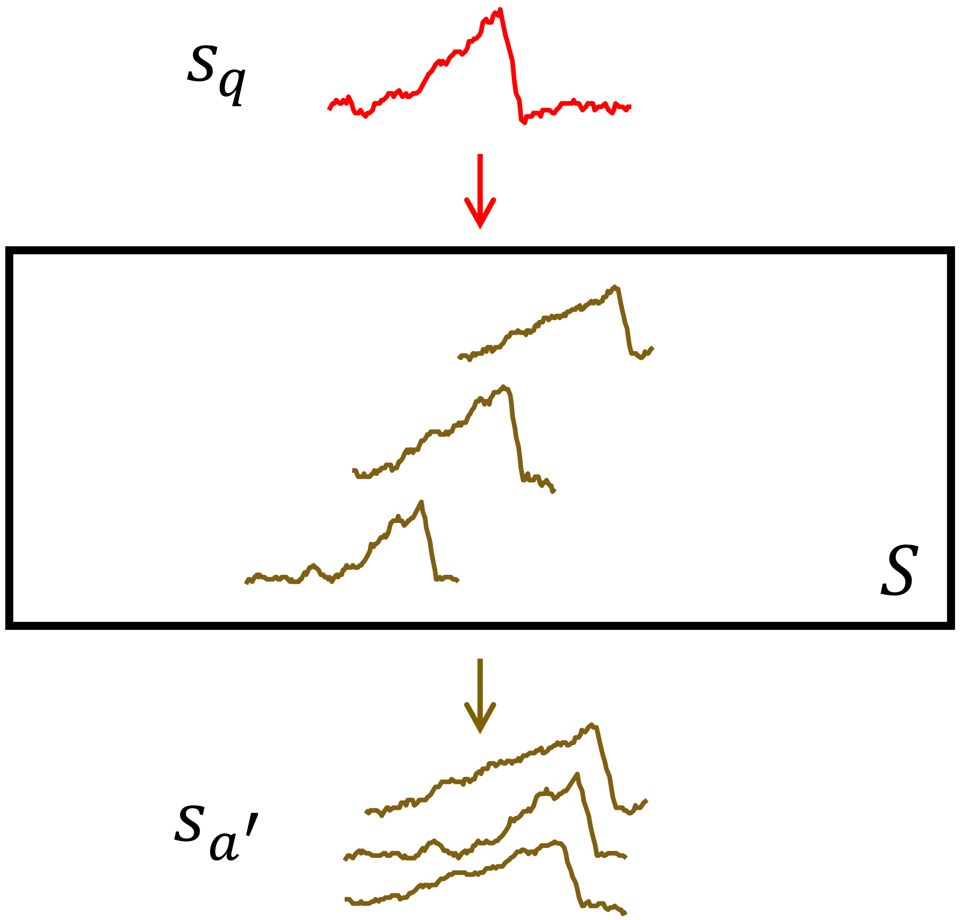


Data Series Approximate Similarity Search

- Similarity search
  - given a series set S, a query series s_q and a similarity measure $d(\cdot,\cdot)$
    - d is commonly the Euclidean distance
  - find the closest series in S to s_q, i.e.,

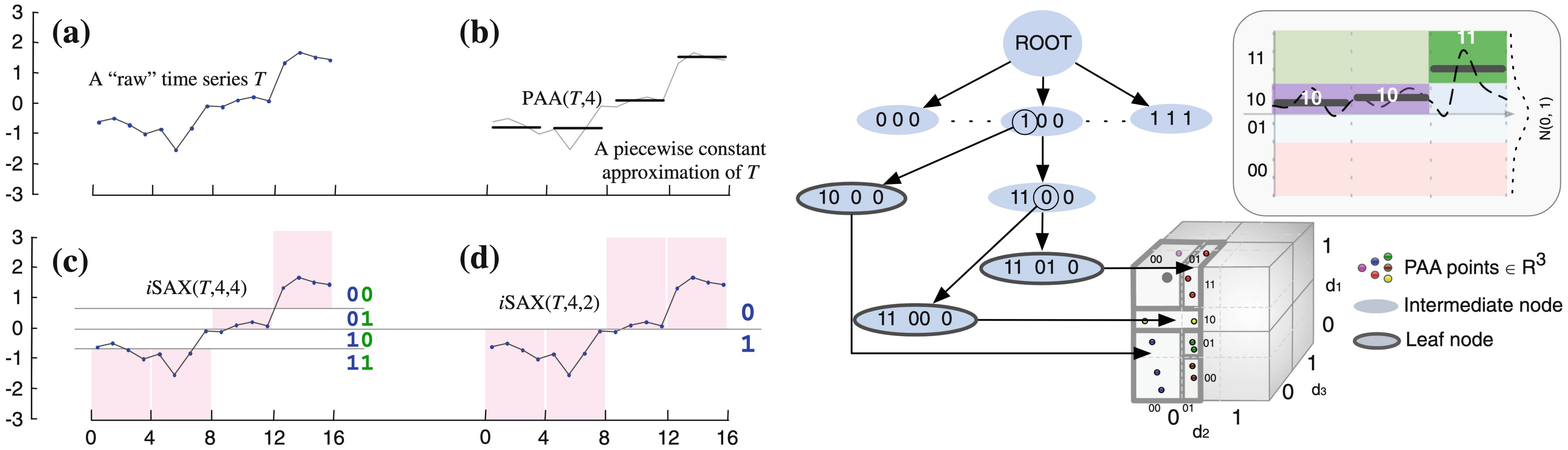
$$s_a = \underset{s_i \in S}{\arg\min}\, d(s_q, s_i)$$

- Approximate similarity search
  - (efficiently) find $s_{a'}$,$d(s_q, s_{a'}) \approx d(s_q, s_a)$



State-of-the-art: iSAX Family of Indexes[3,4]

Raw series → PAA approximation → SAX symbolization → iSAX index

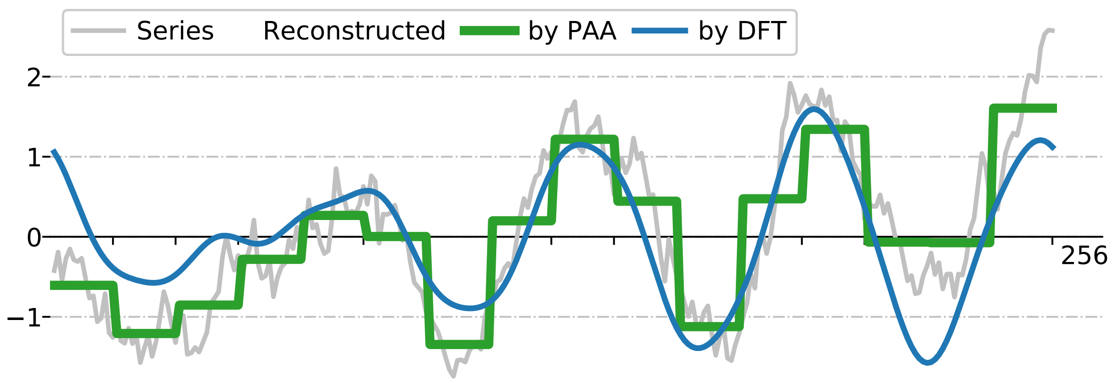


3. Alessandro Camerra, et al. Beyond One Billion Time Series: Indexing and Mining Very Large Time Series Collections with iSAX2+. KAIS 39(1):123-151, 2014.
4. Themis Palpanas. Evolution of a Data Series Index - The iSAX Family of Data Series Indexes. CCIS 1197, 2020.

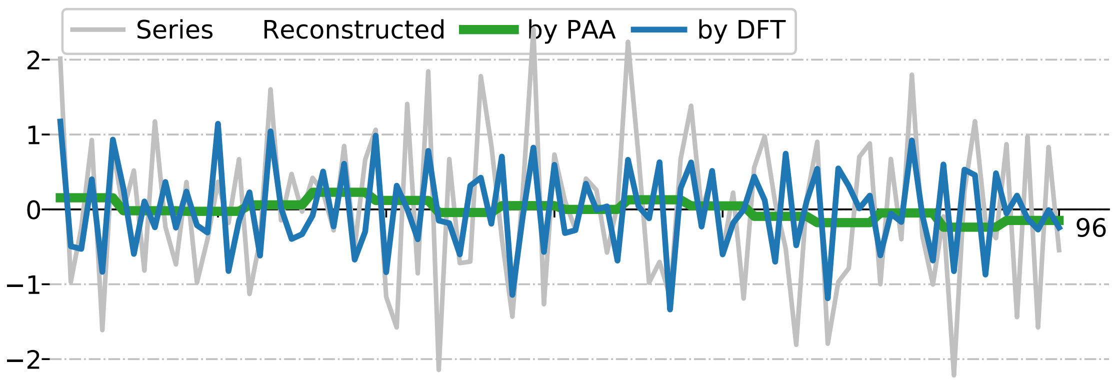
Limitations of (PAA-based) iSAX

Depends on whether PAA successfully profiles the dataset

→ **Need for better summarizations**

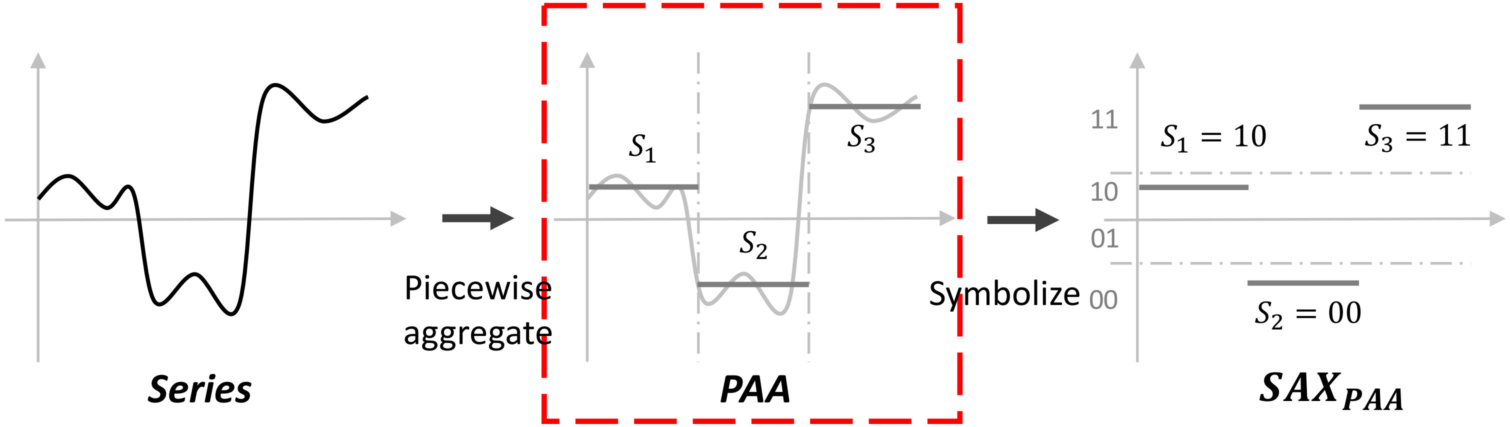


PAA (and DFT) works
to approximate and reconstruct
a RandomWalk series



PAA (and DFT) fails
to approximate and reconstruct
a Deep1B series

DEA: *D*eep *E*mbedding *A*pproximation



PAA-based
SAX symbolization

Replace PAA by DEA for SAX symbolization and iSAX index

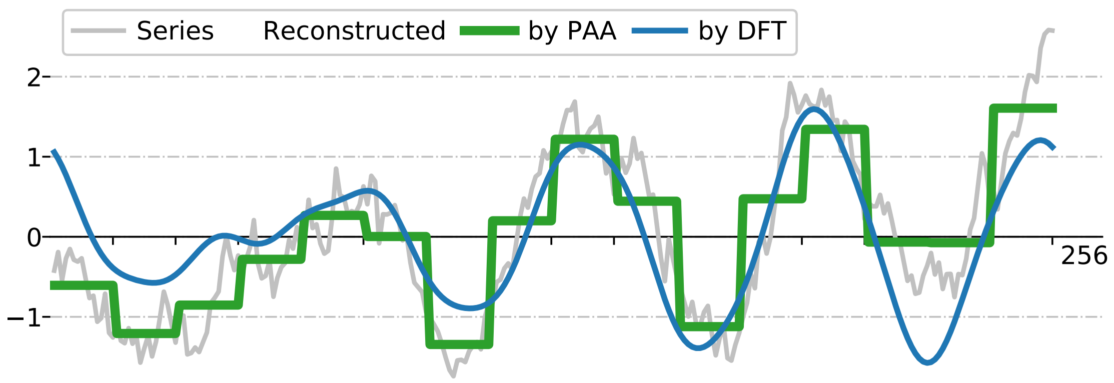
$[e_1, e_2, e_3]$

DEA

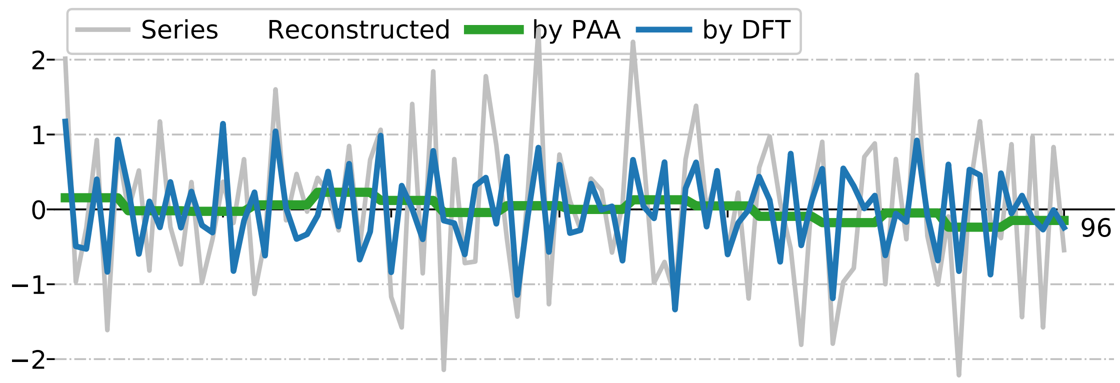
(PAA-based) iSAX → DEA-based iSAX

Depends on whether PAA successfully profiles the dataset

→ **Need for better summarizations**



PAA (and DFT) works
to approximate and reconstruct
a RandomWalk series

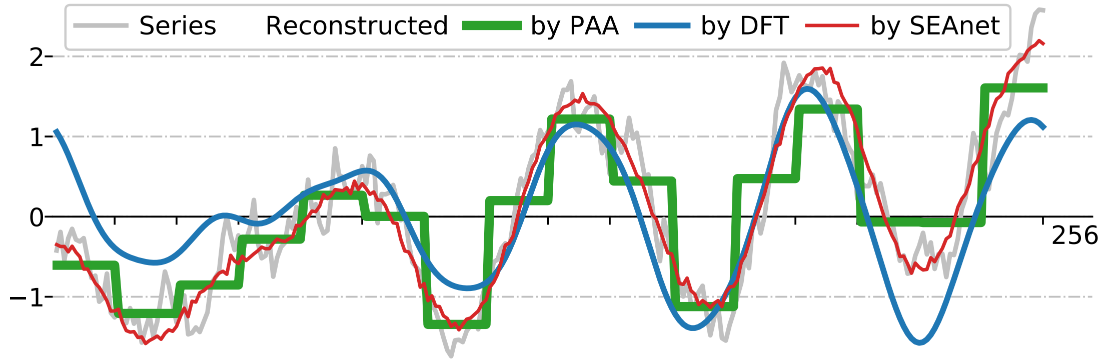


PAA (and DFT) fails
to approximate and reconstruct
a Deep1B series

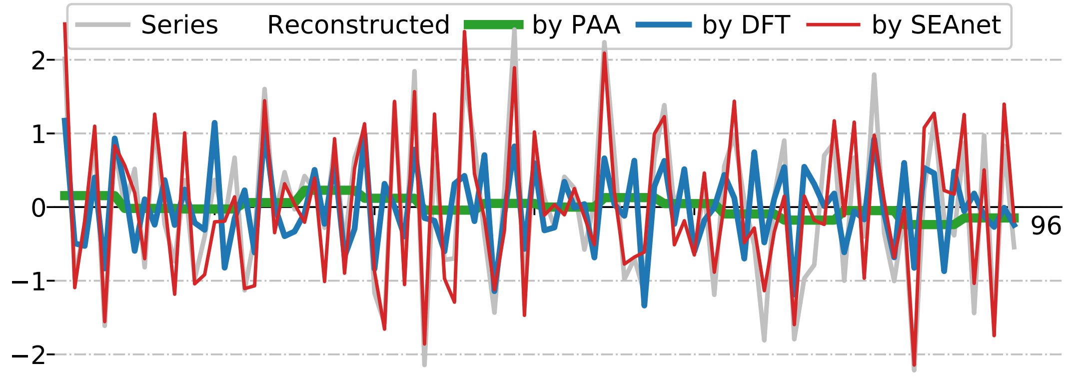
(PAA-based) iSAX → DEA-based iSAX

- DEA better profiles diversified dataset than PAA

✓ Fulfill the need for better summarizations

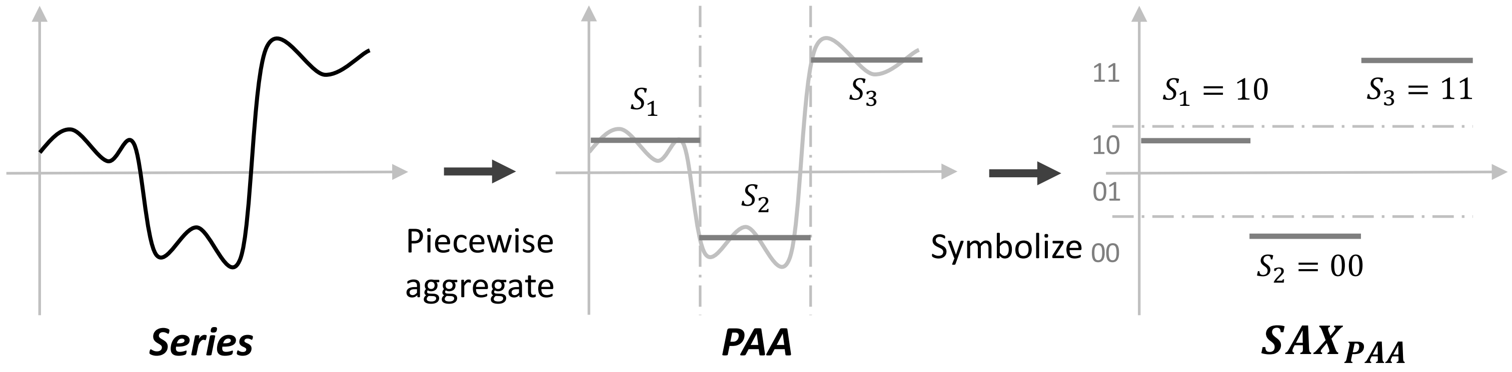


DEA works
to approximate and reconstruct
a RandomWalk series



DEA works
to approximate and reconstruct
a Deep1B series

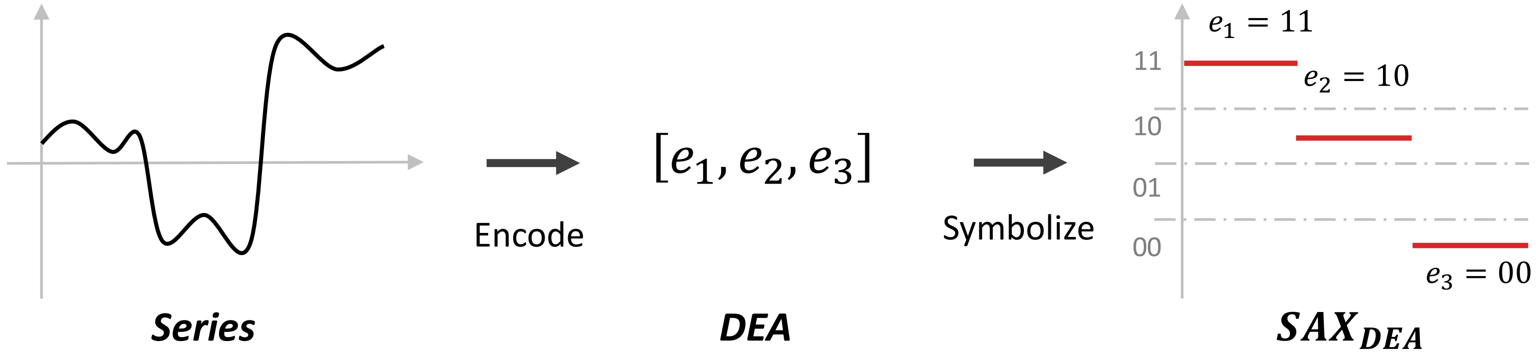
DEA-based iSAX



PAA-based SAX symbolization

Replace PAA by DEA for SAX symbolization and iSAX index

DEA-based SAX symbolization



How to generate high-quality DEA on massive data series collections for approximate similarity search?

Challenges

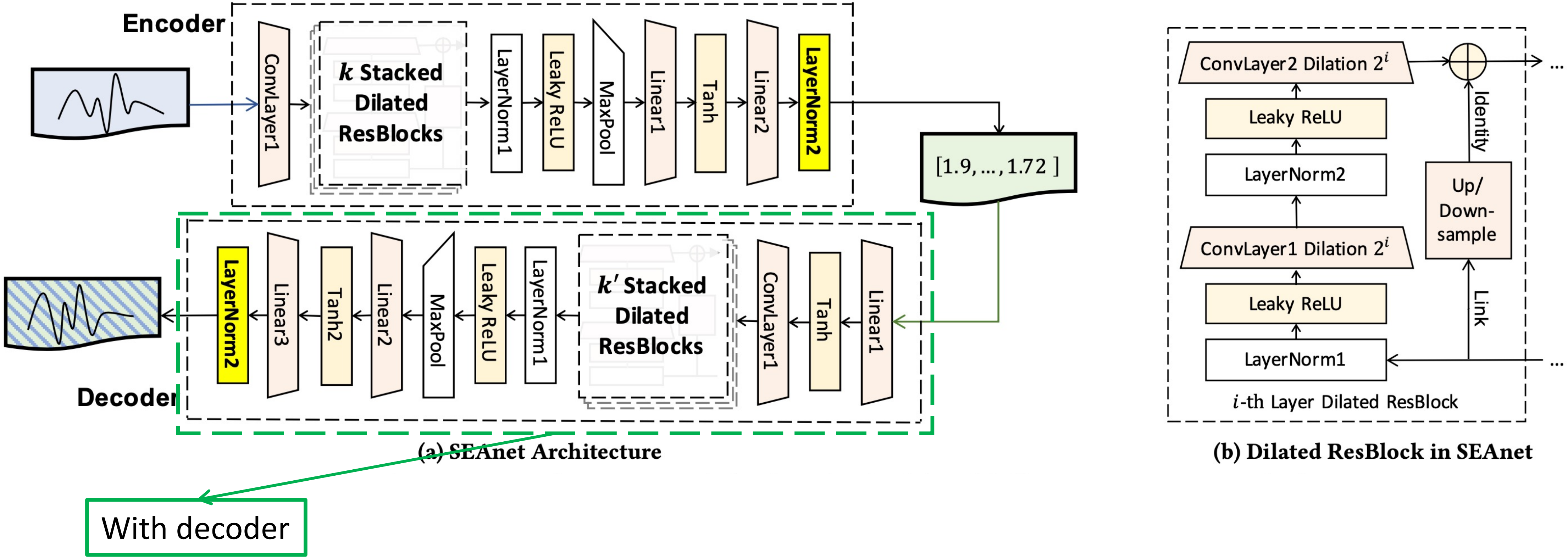
1. Effective architecture for similarity search?

2. Efficient learning on massive datasets?
   - 100 million 256-length series ≈ 100GB

Solutions

✓ SEAnet: *SE*ries *A*pproximation *net*work
   - exponentially dilated ResNet + Sum of Squares (SoS) regularization

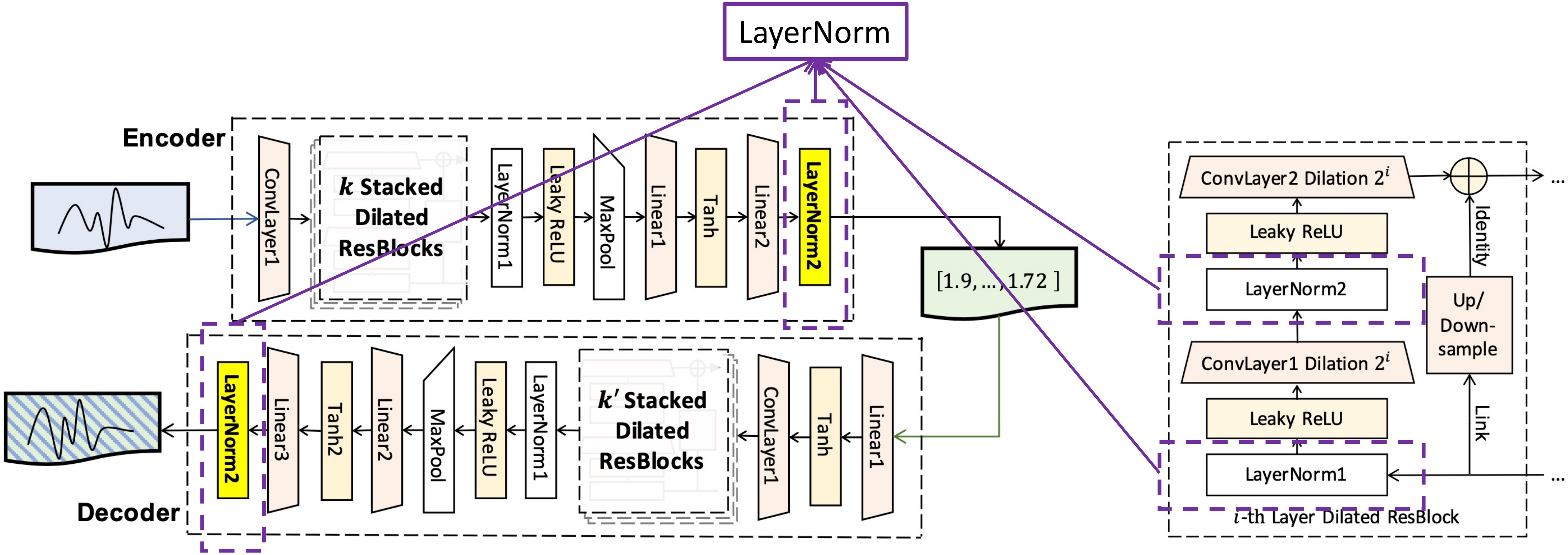
SEAnet Architecture



(a) SEAnet Architecture

(b) Dilated ResBlock in SEAnet

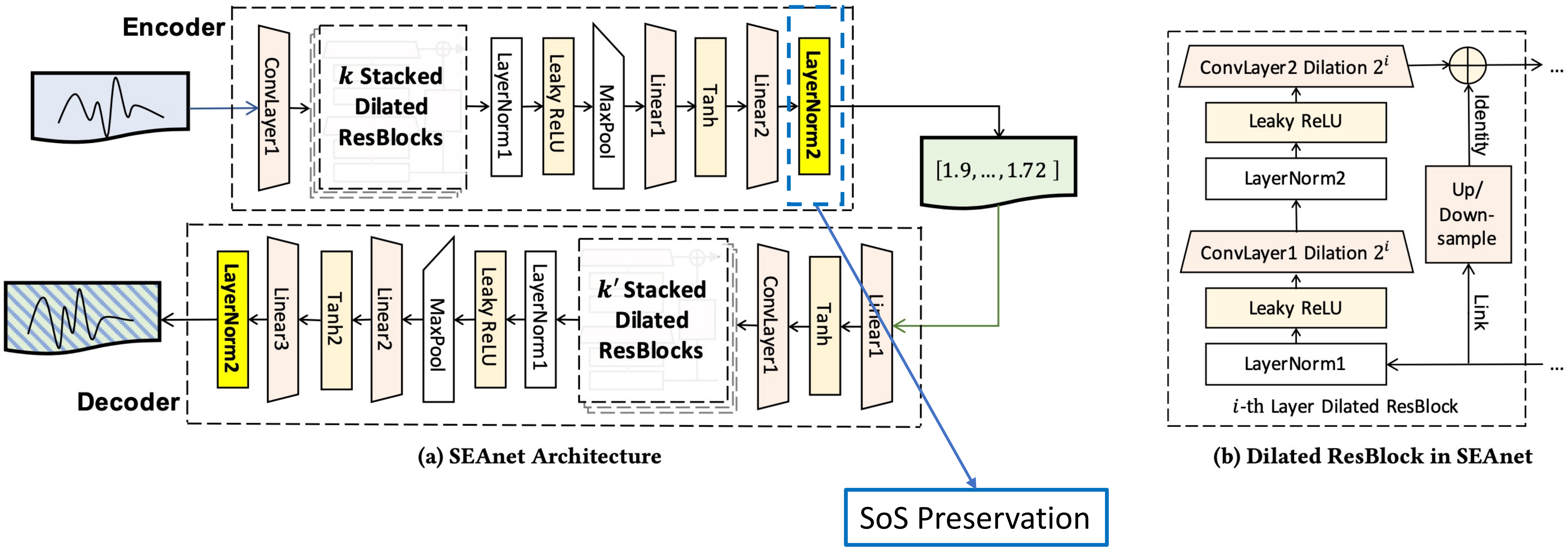
SEAnet Architecture



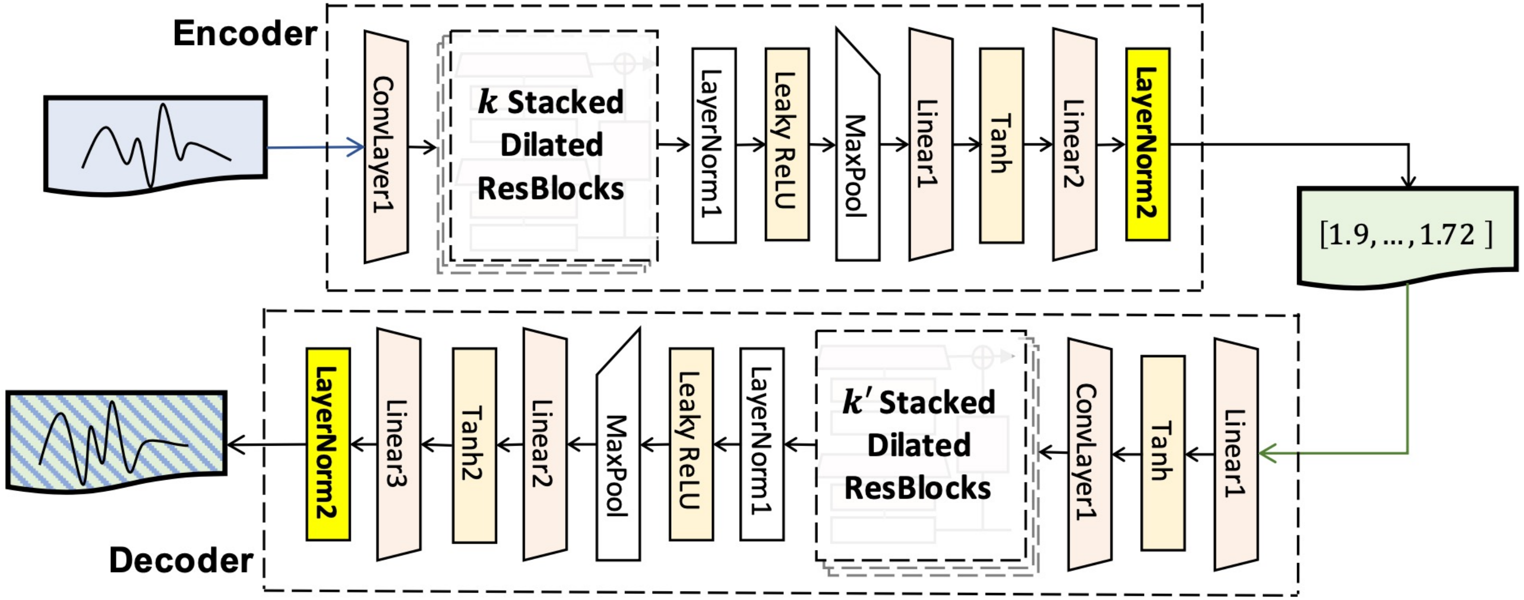
(a) SEAnet Architecture

(b) Dilated ResBlock in SEAnet

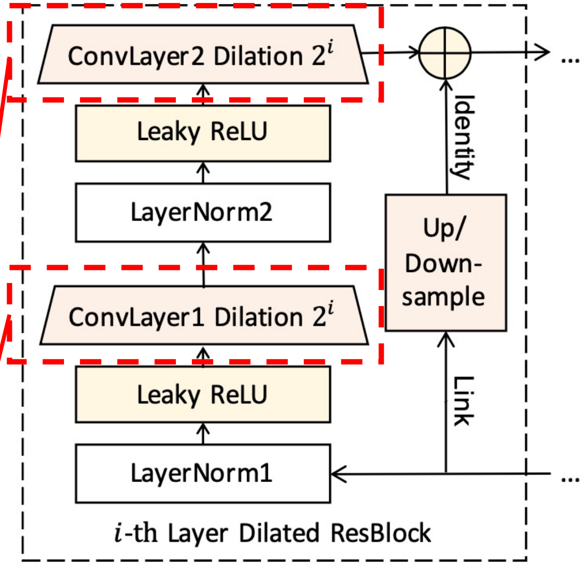
SEAnet Architecture



SEAnet Architecture



(a) SEAnet Architecture



(b) Dilated ResBlock in SEAnet

Exponentially-increasing dilations

SEAnet Training

- Loss $L = L_C + \alpha L_R$
  - Compression error (pairwise) L_C

$$L_C = \frac{1}{N_p} \sum_{(S_i,S_j)\in S\times S} \left| \frac{1}{\sqrt{m}} d(S_i, S_j) - \frac{1}{\sqrt{l}} d\left(\phi(S_i), \phi(S_j)\right) \right|$$

  - Reconstruction error L_R

$$L_R = \frac{1}{N_S} \sum_{S_i\in S} \frac{1}{\sqrt{m}} d(S_i, \psi \cdot \phi(S_i))$$

  - $\frac{1}{\sqrt{m}}$ and $\frac{1}{\sqrt{l}}$: scaling coefficients under SoS regularization
    - m: series length, l: DEA length, ϕ/ψ: en-/decoder mapping

Sum of Squares Preservation

- Sum of Squares (SoS)
  - $\sum_{i,j} M_{i,j}^2$
    - $M_{i,*}$ denotes series, $M_{*,j}$ denotes position

⇒measures preserved information
- in linear dimensionality reductions on z-normalized datasets
  - where SoS ⟺ total variances
- by selecting the largest eigenvalues

⇌fix SoS, to learn the transformation
- i.e., nonlinear encoder mapping
- SoS works as a regularizer

⟲

⇒fix transformation (linear), to preserve SoS

SoS-Preservation Regularization

- Regularize SEAnet by SoS preservations:
  1. z-normalize embeddings (LayerNorm2)
  2. scale series/DEA by its length in loss functions

- Benefits
  - ✓regularize by preserving SoS → higher-quality embeddings
  - ✓stabilize gradients and latent values (by decreasing Var) → better model convergence

	Before Scaling		Scaled by $\sqrt{256/m}$		Scaled by $\sqrt{1/m}$	
Length m	Mean	Var	Mean	Var	Mean	Var
256	22.605	0.999	22.605	0.999	1.4128	0.0039
128	15.969	0.998	22.583	1.9961	1.4115	0.0078
96	13.820	0.997	22.569	2.6597	1.4105	0.0104
16	5.5692	0.984	22.277	15.743	1.3923	0.0615
8	3.8772	0.967	21.933	30.944	1.3708	0.1209

How to generate high-quality DEA on massive data series collections for approximate similarity search?

Challenges

1. Effective architecture for similarity search?

2. Efficient learning on massive datasets?
   - 100 million 256-length series ≈ 100GB

Solutions

✓SEAnet: *SE*ries *A*pproximation *net*work
- exponentially dilated ResNet + Sum of Squares (SoS) regularization

✓SEAsam: *SEA-sam*pling
- sampling based on a sortable series summarization

SEAsam

- Intuition
  - Sampling by dataset's intrinsic distribution

  ✓Draw samples by equal-intervals from the **ordered** set

→Observation
  - every subsequent bit in one SAX symbol contains a decreasing amount of information about the location
    - ≈ space-filling curves

- How to order series in a dataset?

✓Order by **InvSAX**[5]
  - interleaving SAX's bits

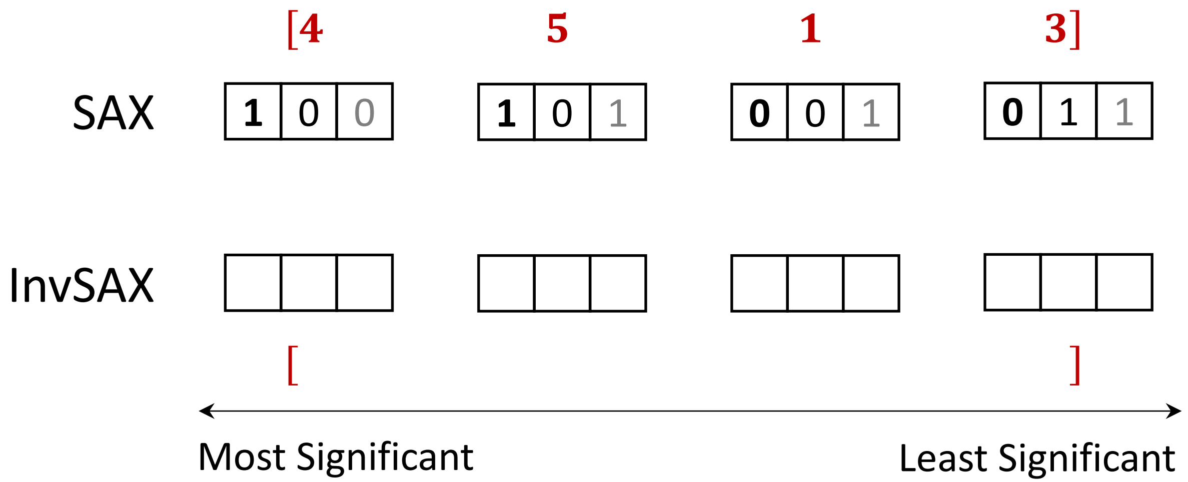
  ⇒ all significant bits across each SAX symbol precede all less significant bits

5. Haridimos Kondylakis, Niv Dayan, Kostas Zoumpatianos, Themis Palpanas. Coconut: A Bottom-Up Approach for Building Data Series Indexes. VLDB, 2018

InvSAX transformation

- interleaving SAX's bits

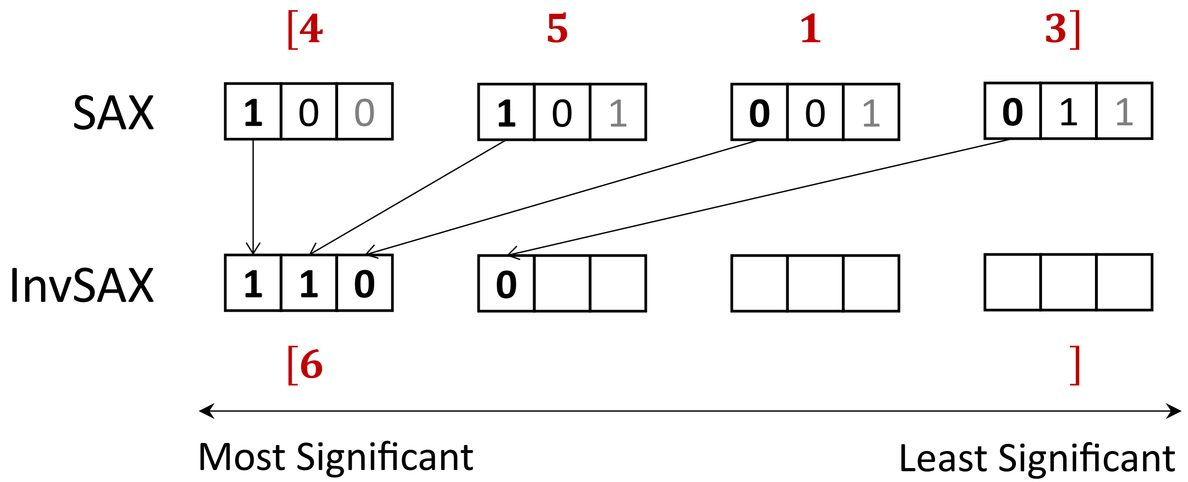
⇒ all **significant bits** across each SAX symbol precede all less significant bits



InvSAX transformation

- interleaving SAX's bits

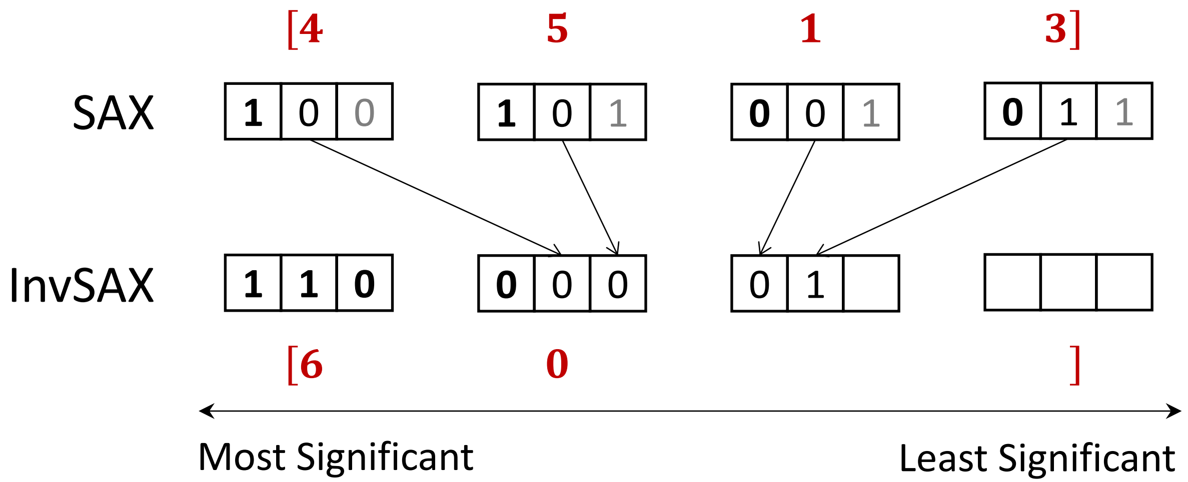
⇒ all **significant bits** across each SAX symbol precede all less significant bits



InvSAX transformation

- interleaving SAX's bits

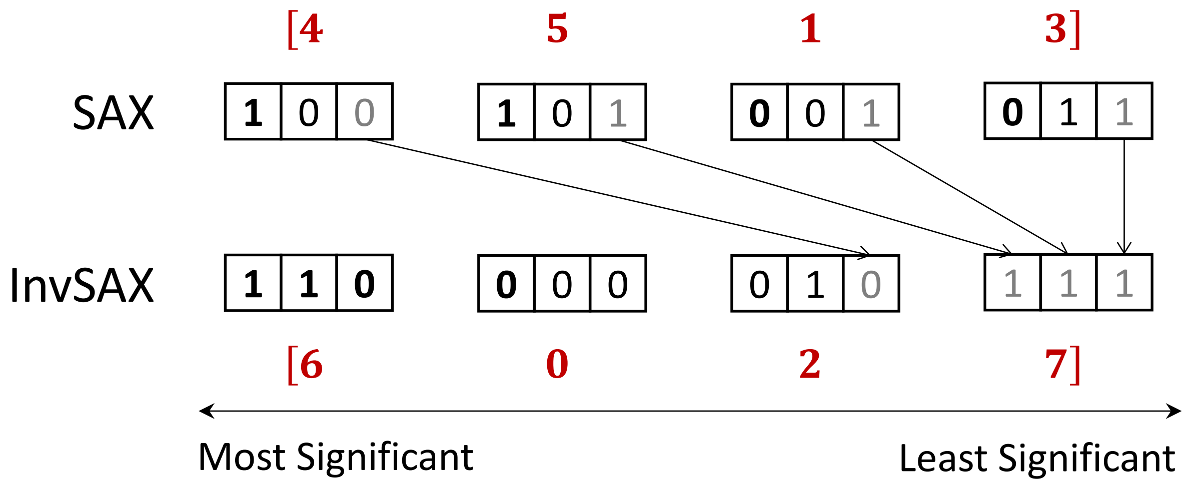
⇒ all **significant bits** across each SAX symbol precede all less significant bits



InvSAX transformation

- interleaving SAX's bits

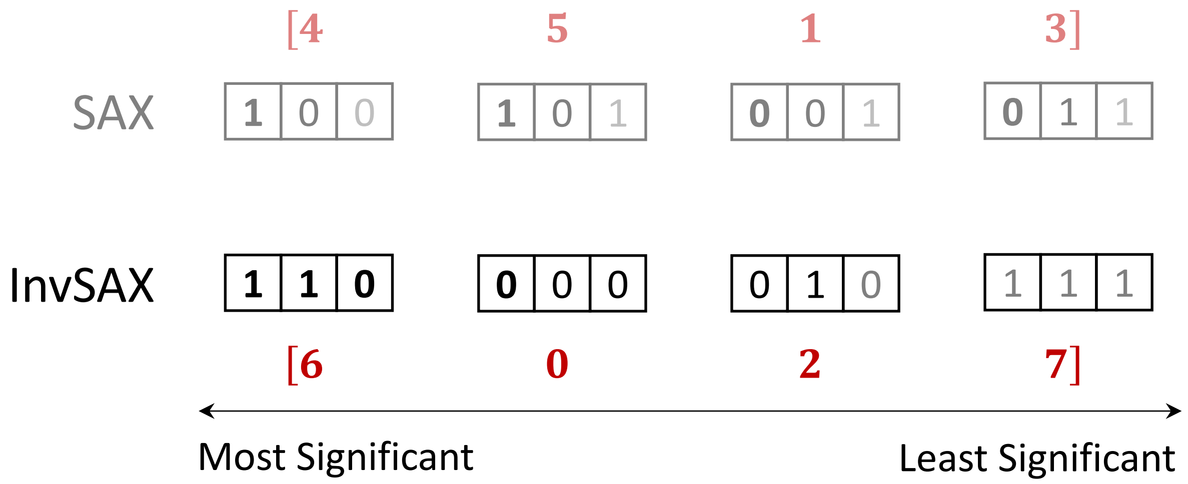
⇒ all **significant bits** across each SAX symbol precede all less significant bits



InvSAX transformation

- interleaving SAX's bits

⇒ all **significant bits** across each SAX symbol precede all less significant bits



Experimental Setup

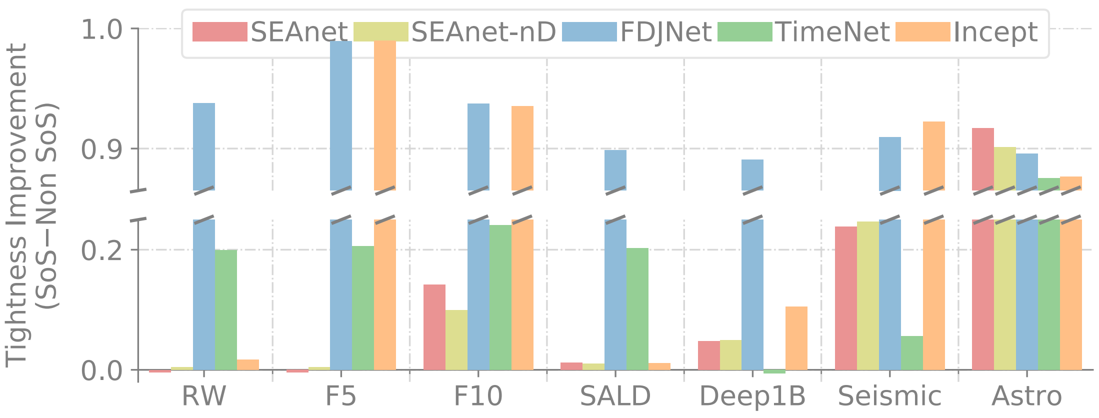
- Datasets
  - 3 synthetic (RandWalk, F5, F10)
  - 4 real (Seismic, SALD, Deep1B, Astro)
- Comparison methods
  - PAA
  - SEAnet-nD (SEAnet without decoder)
  - TimeNet[5], FDJNet[6], InceptionTime[7]
    - Adapted for similarity search
- **Hyper-parameter tunning**
  - **~5,000 models trained**

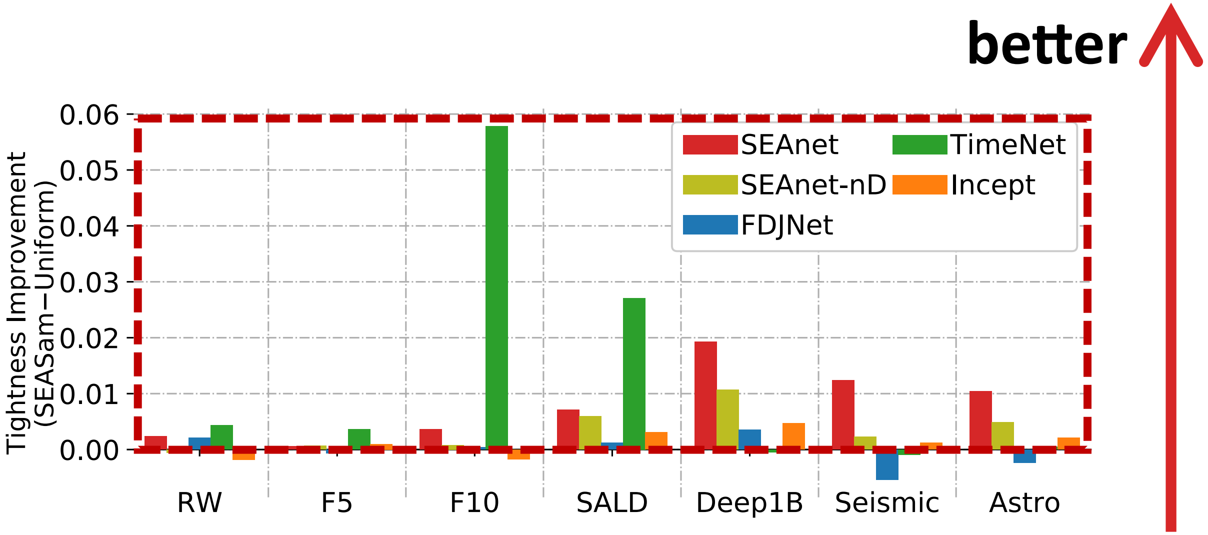
Dataset Statistics

Dataset	Length	Dataset Size	Training Size
RandWalk	256	100 million series	200k SEAsam samples
F5	256		
F10	256		
Seismic	256		
SALD	128		
Deep1B	96		
Astro	256		

6. Pankaj Malhotra, et al. TimeNet: Pre-trained deep recurrent neural network for time series classification. ESANN, 2017.
7. Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. NeurIPS, 2019.
8. Hassan Ismail Fawaz, et al. InceptionTime: Finding AlexNet for time series classification. DMKD, 2020.

SoS Preservation and SEAsam





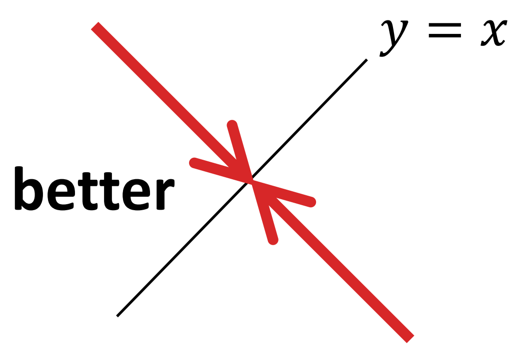
✓SoS Preservation improves tightness of approximate answers

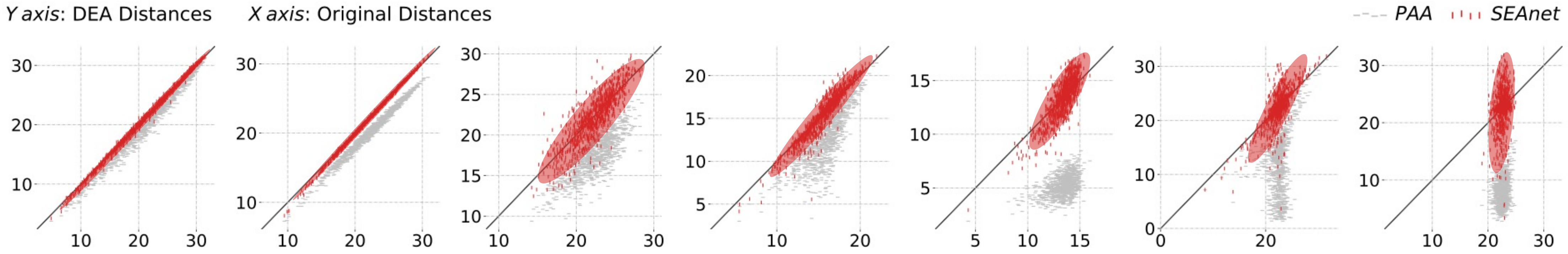
- vs. without SoS

✓SEAsam improves tightness of approximate answers

- vs. uniformly random sampling

DEA distances vs. Original distances

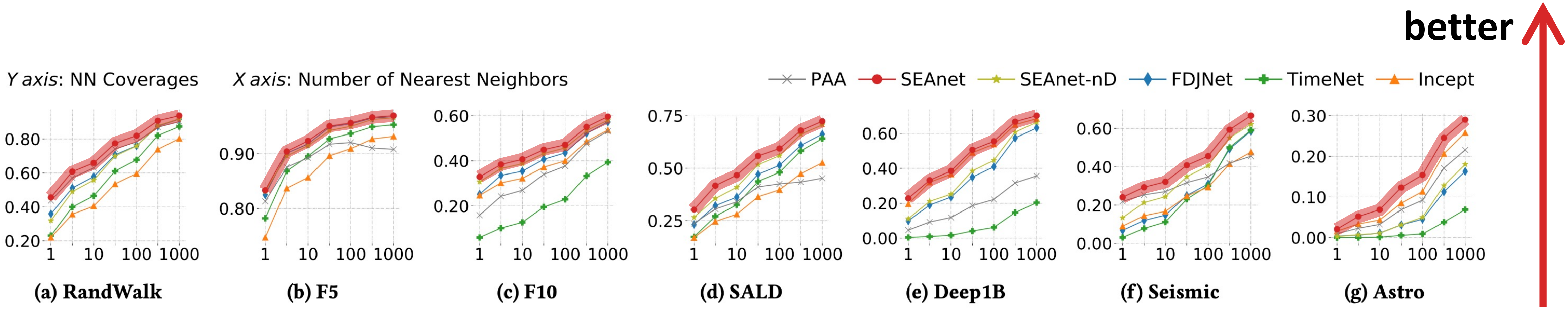




(a) RandWalk (b) F5 (c) F10 (d) SALD (e) Deep1B (f) Seismic (g) Astro

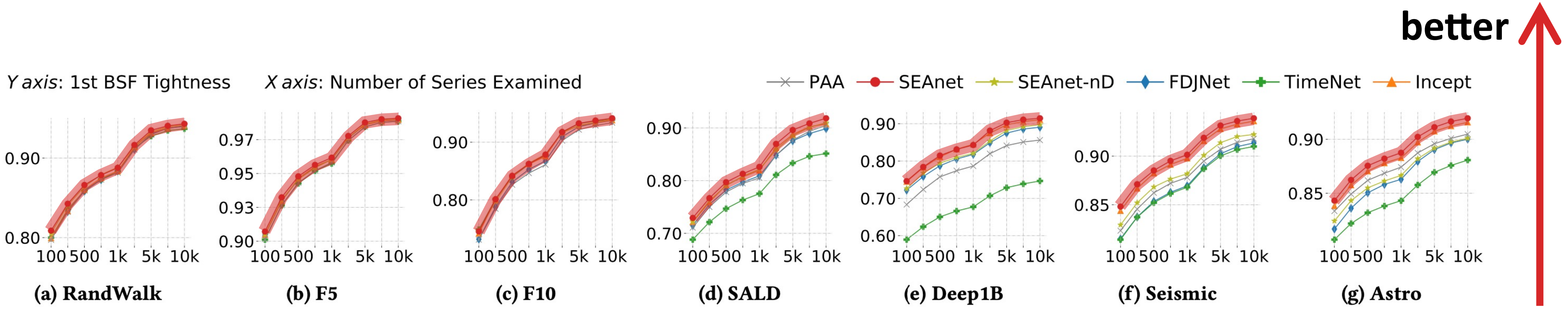
✓SEAnet better preserve original distances in the DEA spaces than PAA

Preserving neighborhood in DEA space



✓SEAnet well preserves the original neighbors in the embedding space

Approximate query answers' tightness



✓SEAnet provides tighter approximate answers

Conclusions

1. Proposed learned embeddings (DEA) as a replacement to traditional data series summarizations
2. Developed SEAnet to effectively learn DEA
   - designed using the novel Sum of Square (SoS) preservation regularization
3. Described SEAsam to efficiently train SEAnet on massive datasets

- DEA by SEAnet outperfoms PAA and other SOTA deep embeddings for data series approximate similarity search
- DEA and SEAnet lead to faster and more accurate data series processing/analytics
- several promising open research directions

Thanks!

Deep Learning Embeddings for Data Series Similarity Search

Qitong Wang and Themis Palpanas

ACM SIGKDD 2021, Singapore