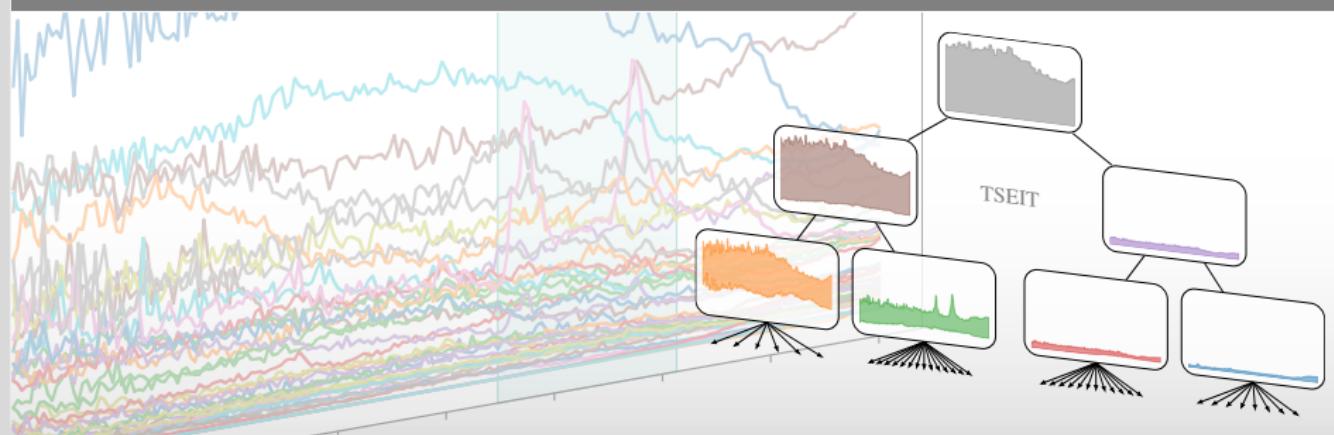


Efficient k -NN Search of Time Series in Arbitrary Time Intervals

Final Presentation for Master's Thesis

Janek Bettinger | March 23, 2018

SYSTEMS FOR INFORMATION MANAGEMENT GROUP — INSTITUTE FOR PROGRAM STRUCTURES AND DATA ORGANIZATION

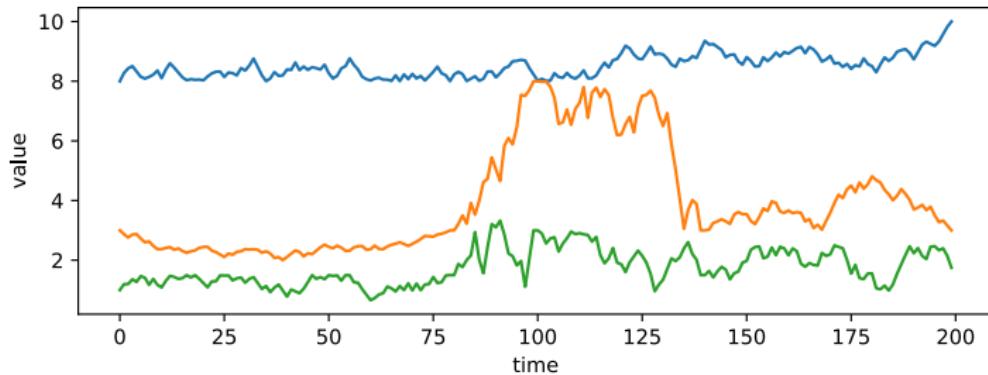


— 1 —

Motivation

Time Series

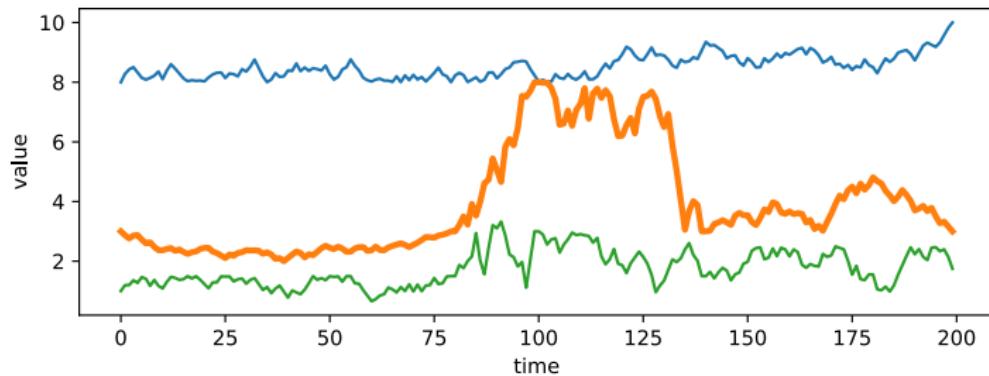
- time series might describe over time:
 - sensor values, e.g.
temperature, energy usage
 - stock prices
 - election polls
 - heartbeats
 - brainwaves



k -Nearest Neighbors (k -NN) of a Time Series

- the k closest time series regarding a distance measure
- for classification, clustering, regression, exploratory data analysis

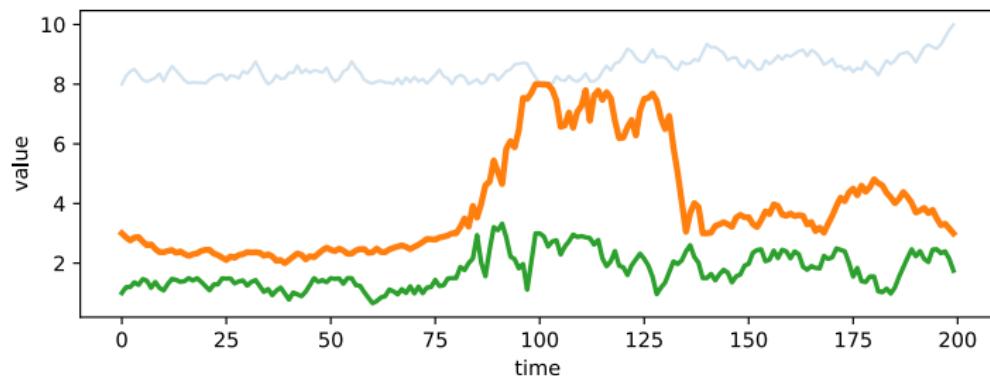
with $k = 1$:



k -Nearest Neighbors (k -NN) of a Time Series

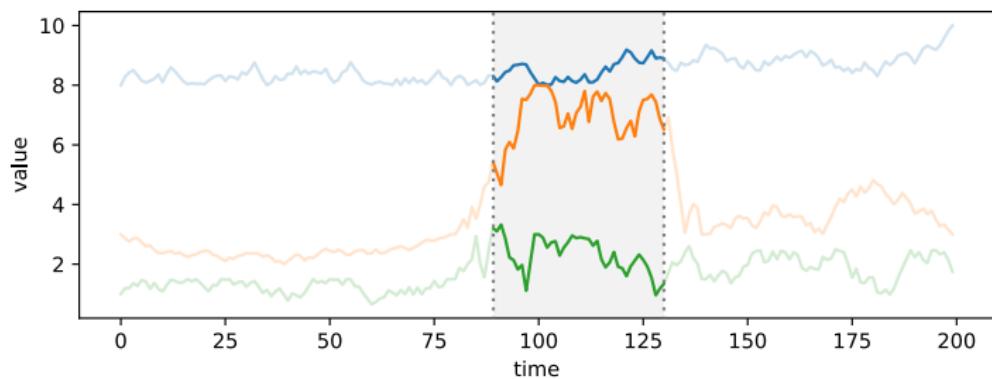
- the k closest time series regarding a distance measure
- for classification, clustering, regression, exploratory data analysis

with $k = 1$:

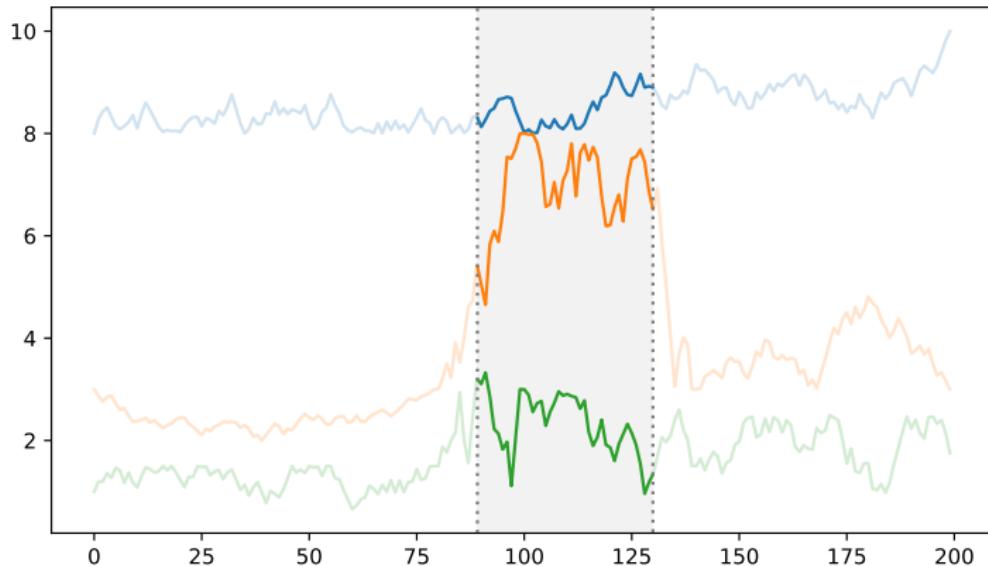


k -Nearest Neighbors — Time Intervals

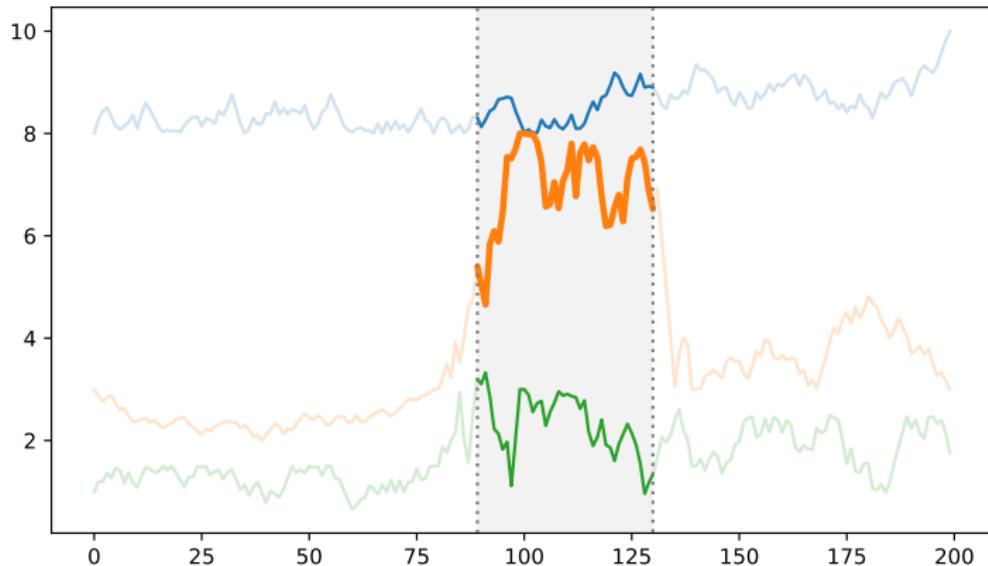
- often only specific time intervals are of interest, e.g.
 - latest period
 - period around a special event
 - e.g., temperature or energy usage during a particular month
- state-of-the-art techniques not suitable



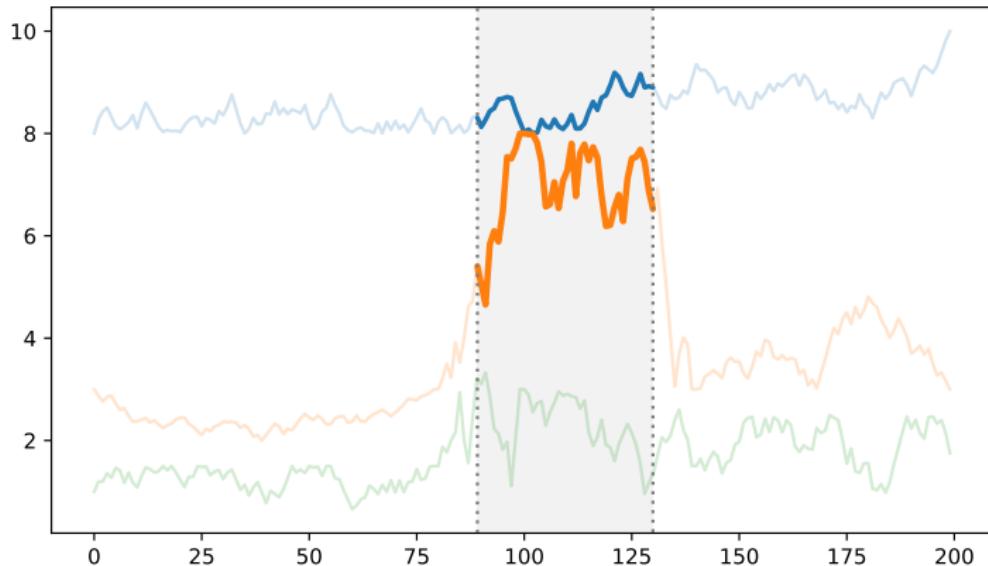
k -Nearest Neighbors — Time Intervals I



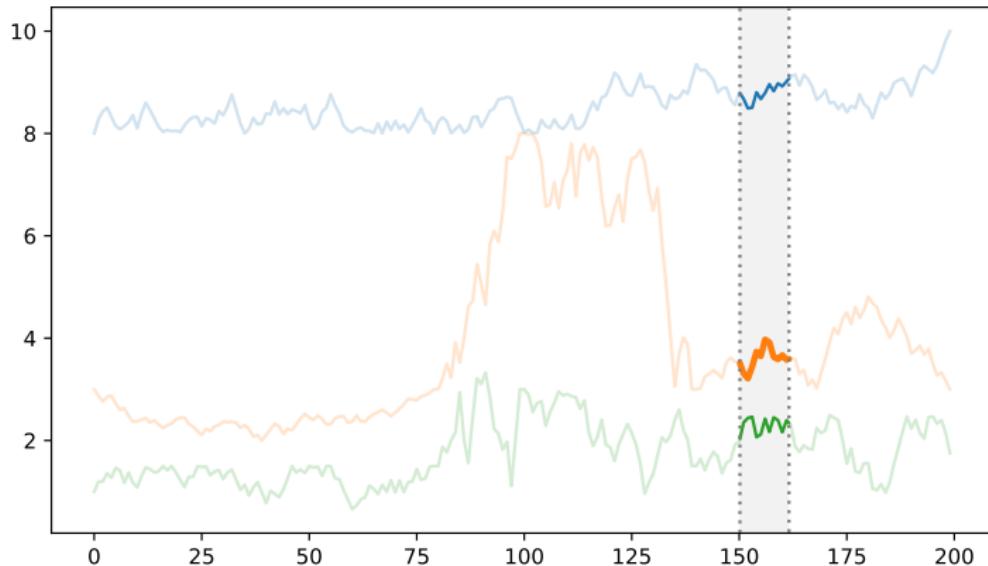
k -Nearest Neighbors — Time Intervals I



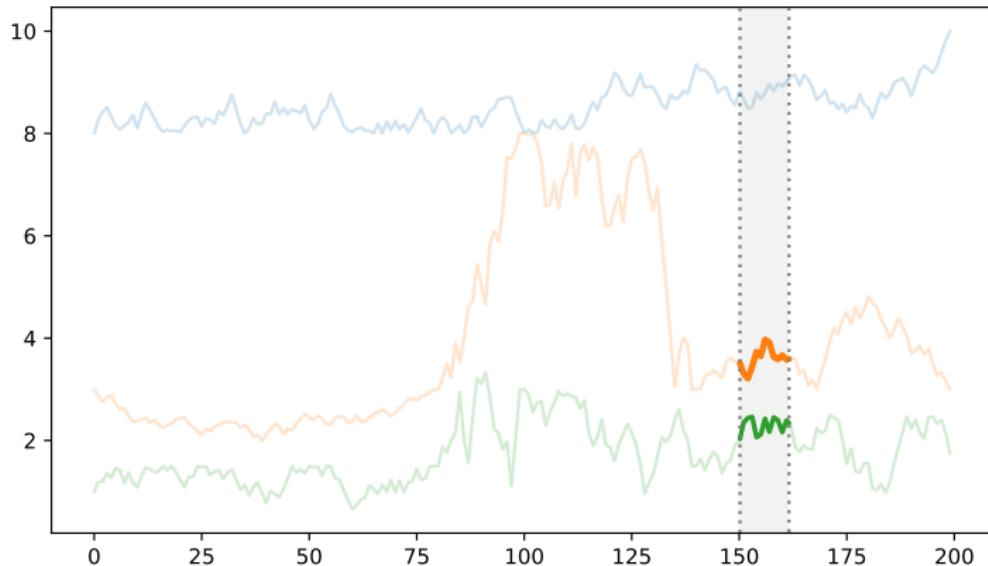
k -Nearest Neighbors — Time Intervals I



k -Nearest Neighbors — Time Intervals II



k -Nearest Neighbors — Time Intervals II



Content

- The Data Structure TSEIT

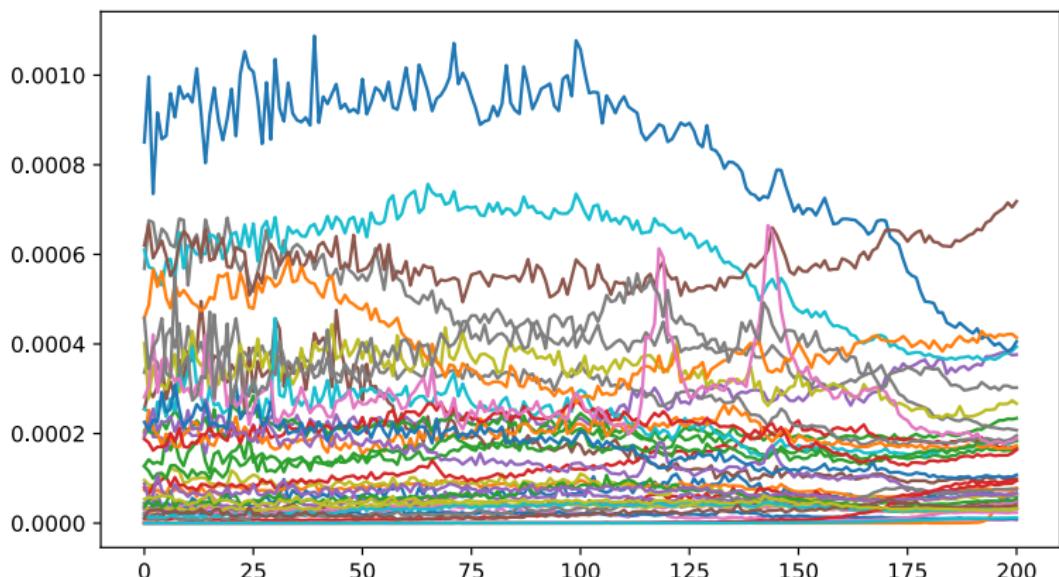
- k -NN Querying
- Index Creation
- Implementation

- Evaluation

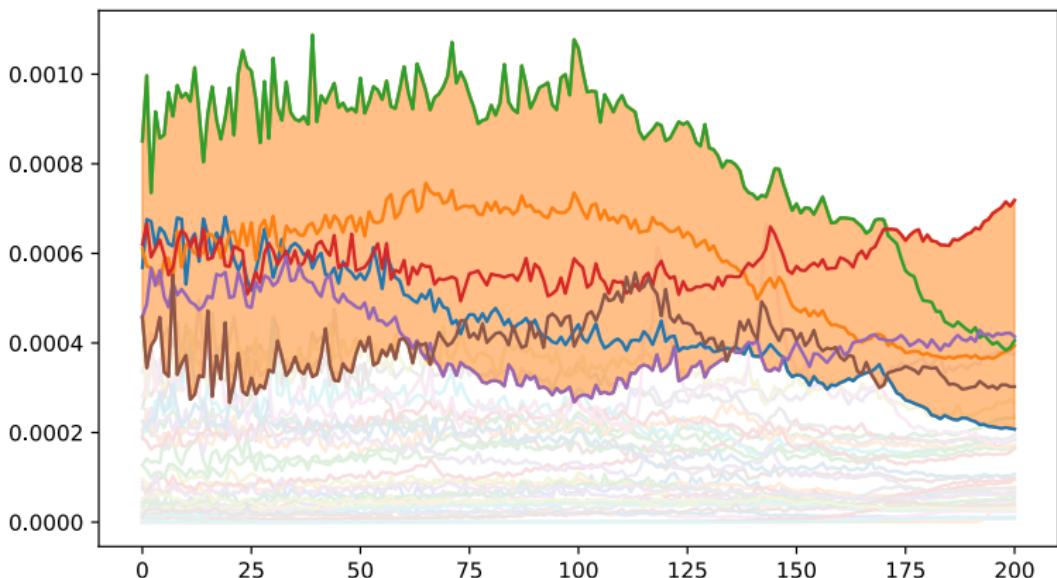
— 2 —

The Data Structure TSEIT

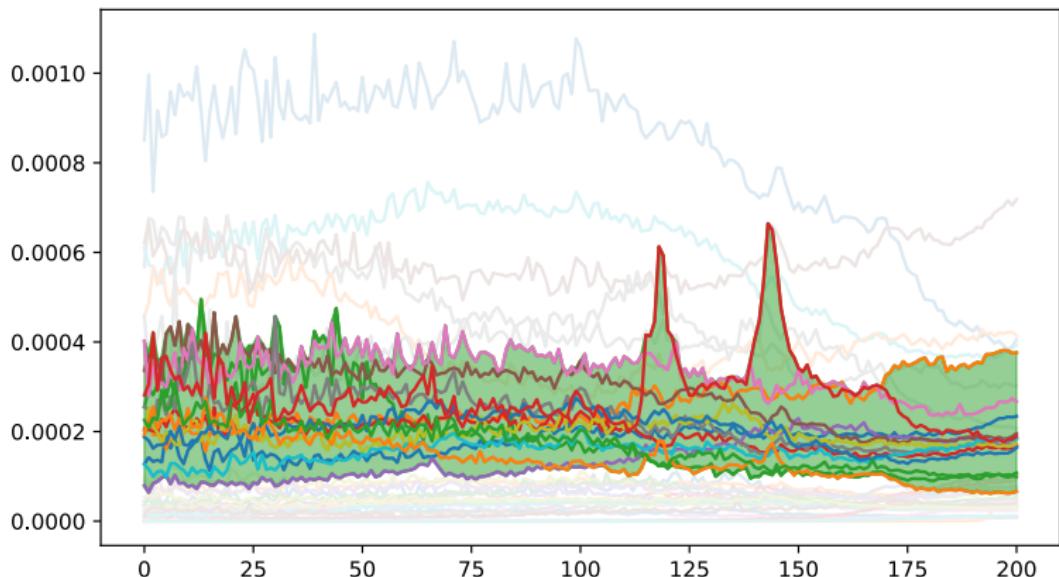
Example — 40 Time Series



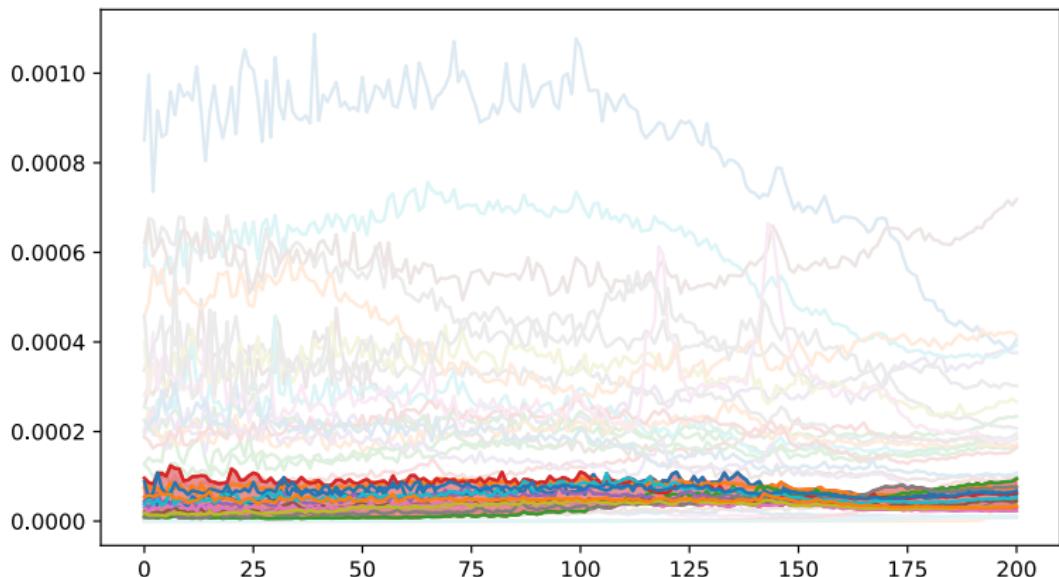
Example — 1st Group of Time Series



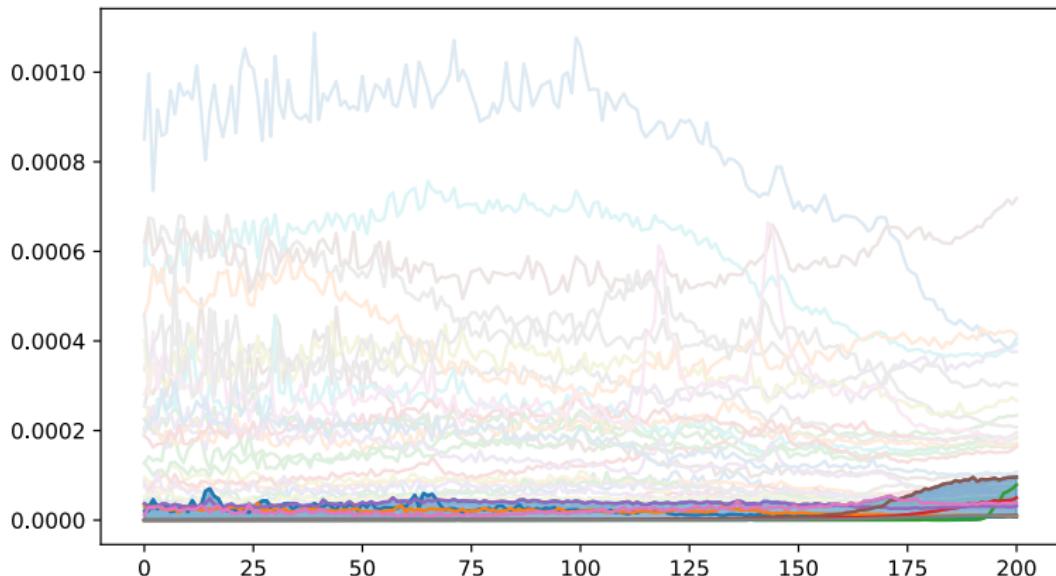
Example — 2nd Group of Time Series



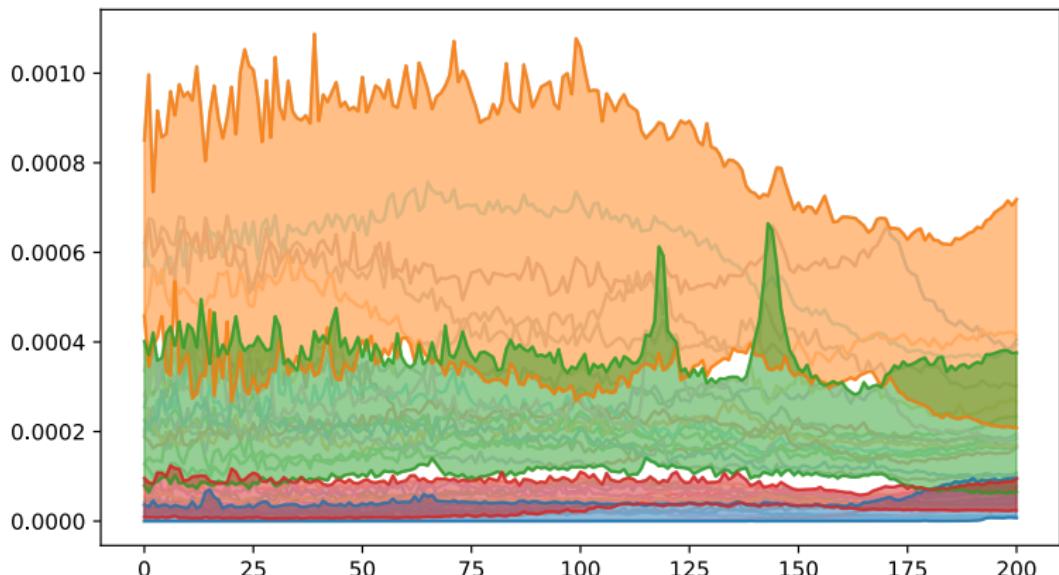
Example — 3rd Group of Time Series



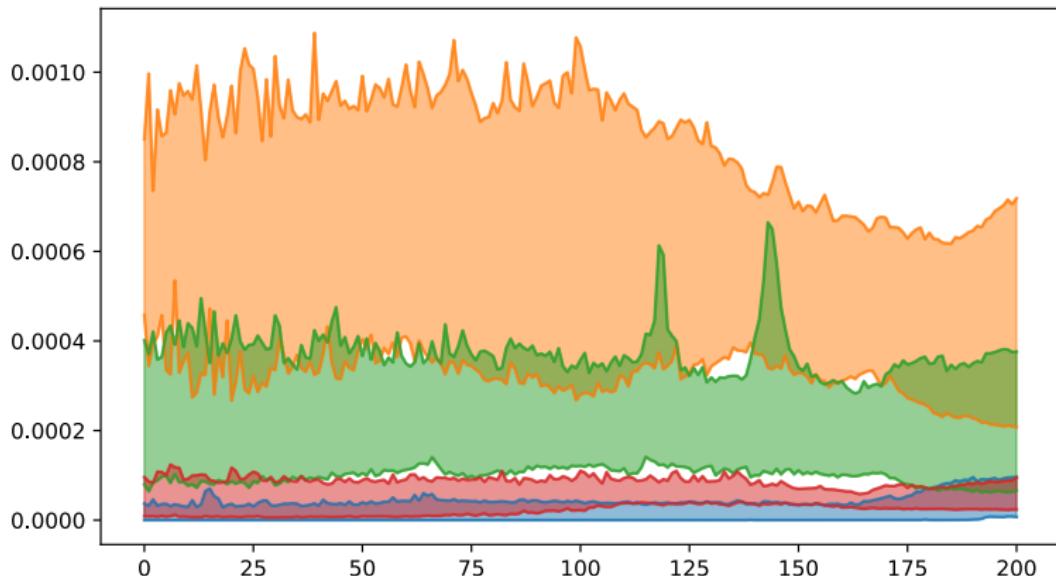
Example — 4th Group of Time Series



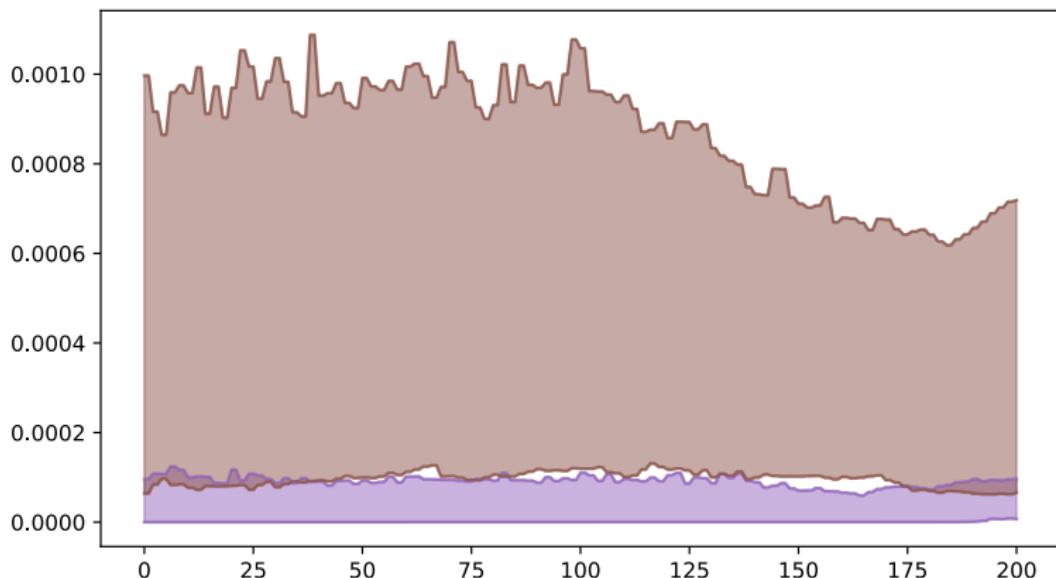
Example — 4 Envelopes



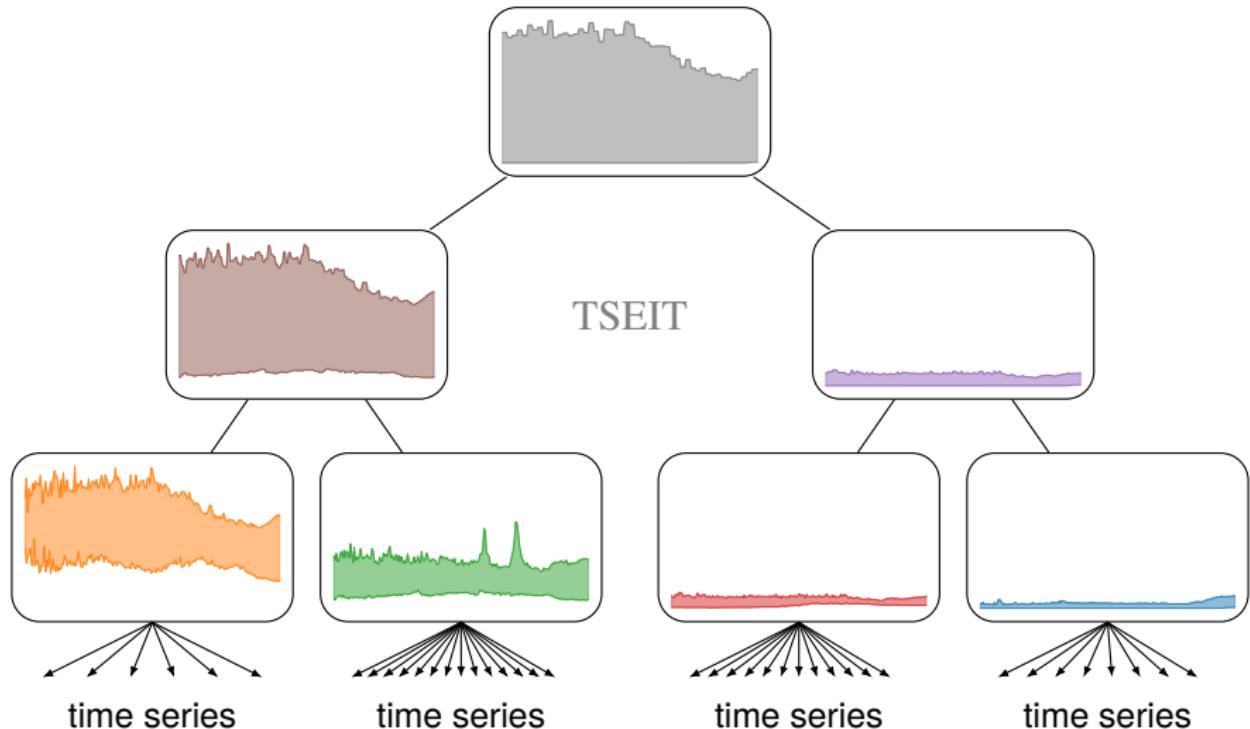
Example — 4 Envelopes



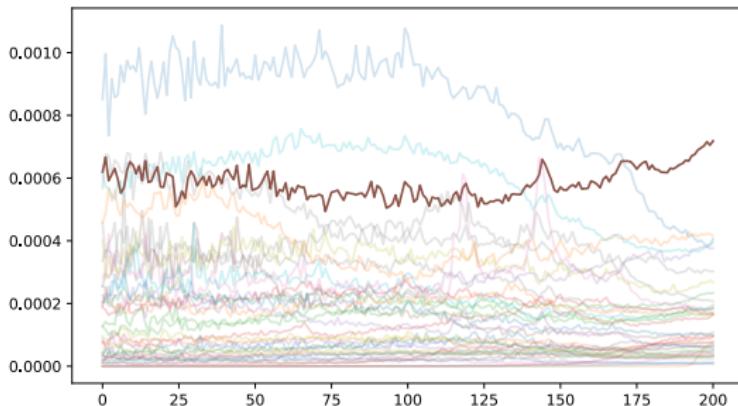
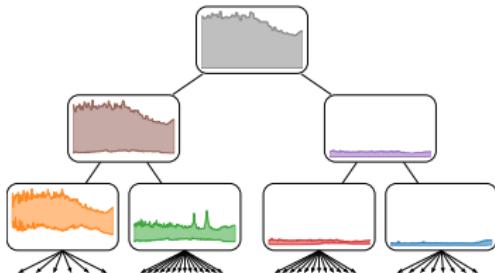
Example — 2 Combined Envelopes



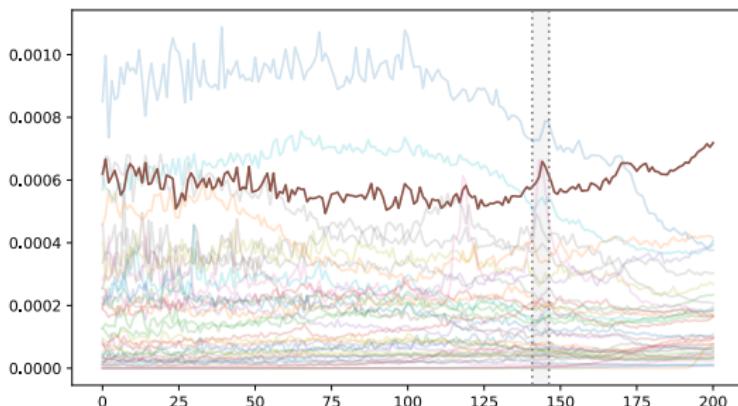
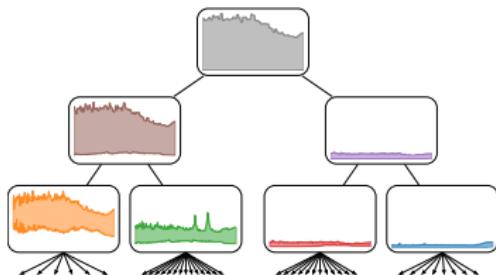
Example — Time Series Envelopes Index Tree



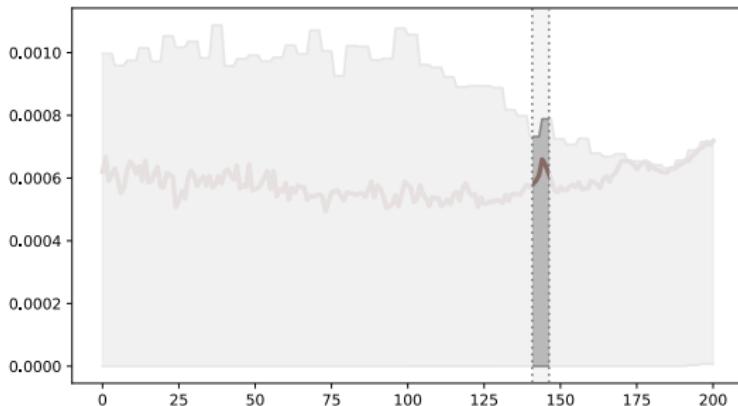
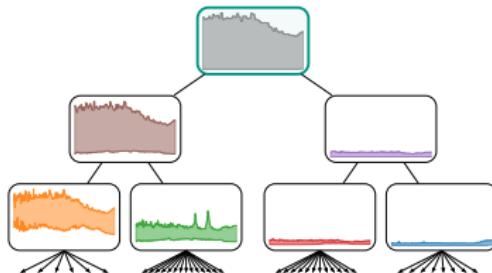
k -NN Querying



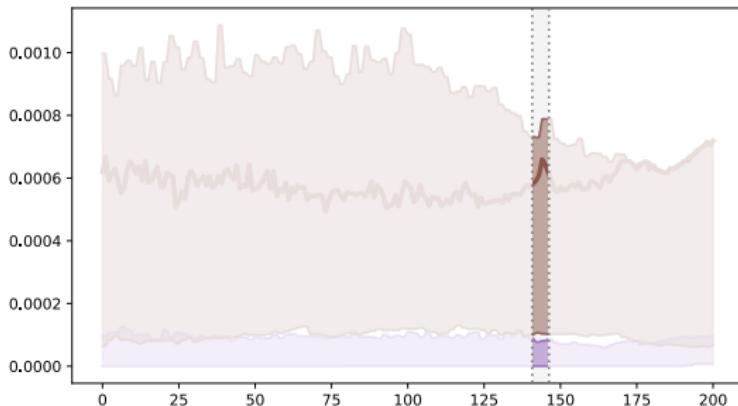
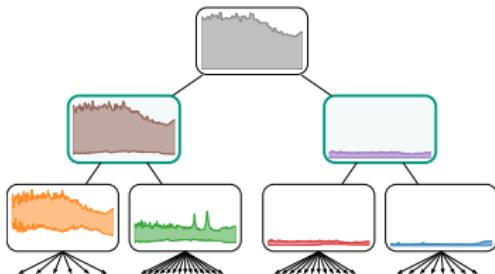
k -NN Querying



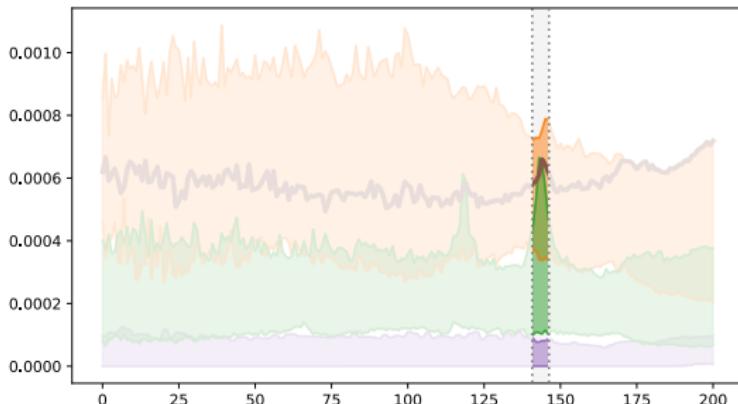
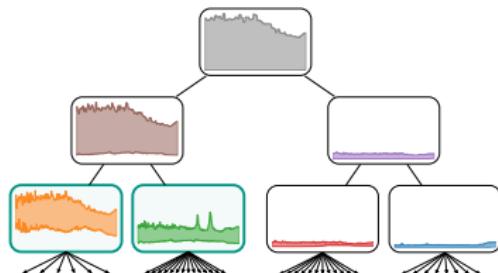
k -NN Querying



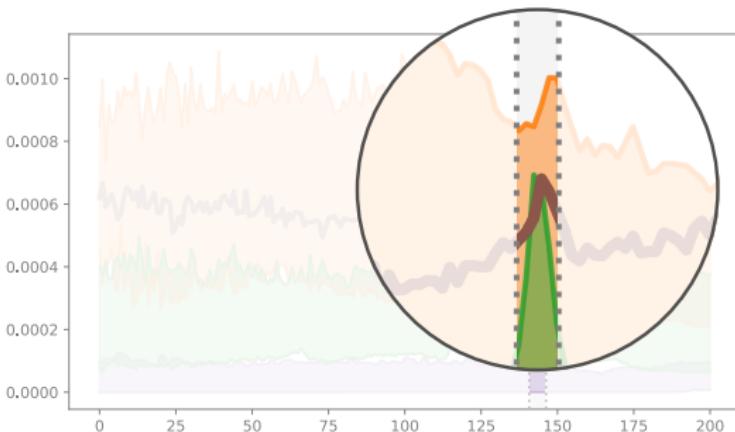
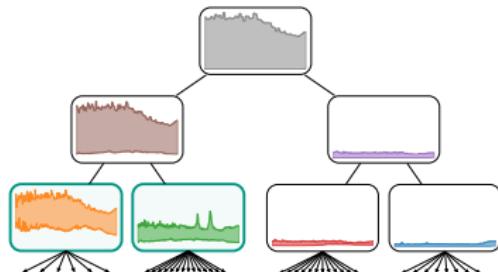
k -NN Querying



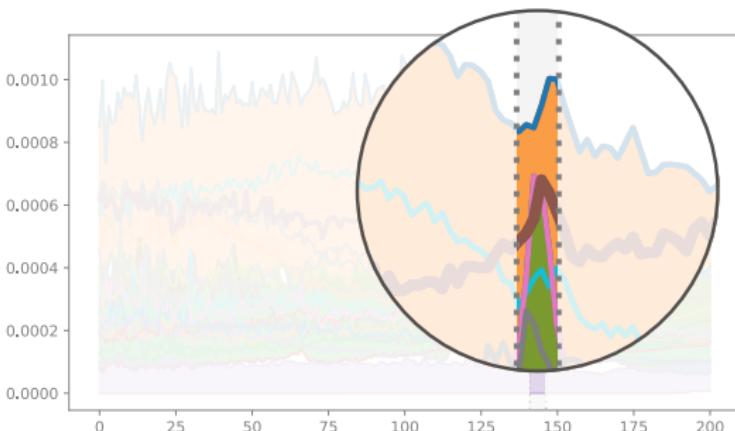
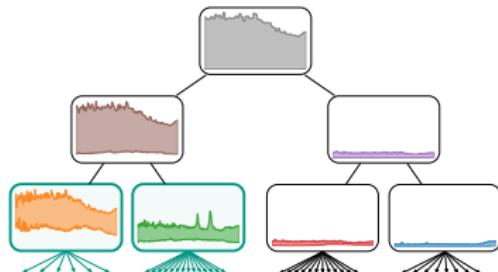
k -NN Querying



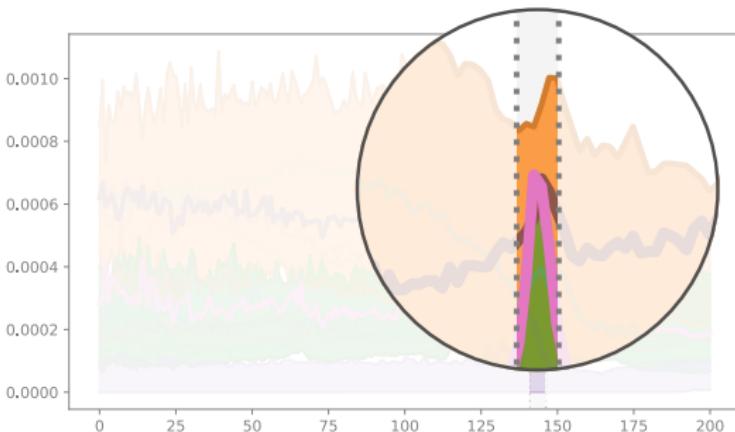
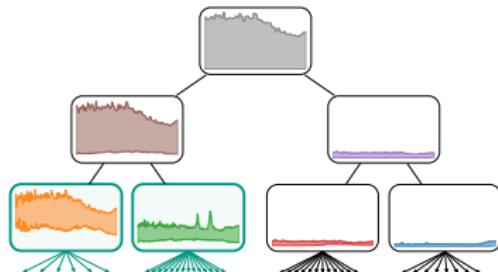
k -NN Querying



k -NN Querying



k -NN Querying

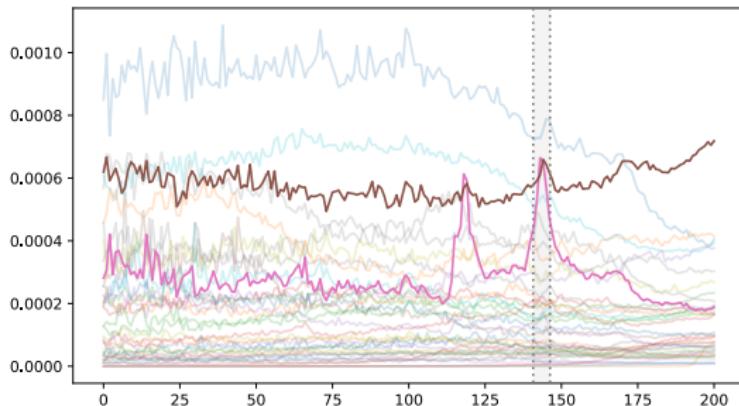
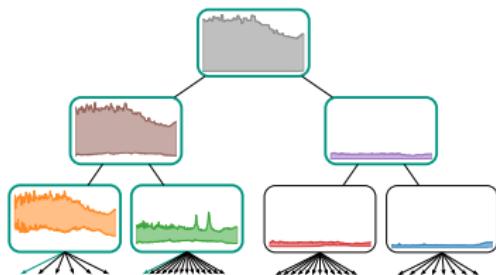


Motivation

TSEIT

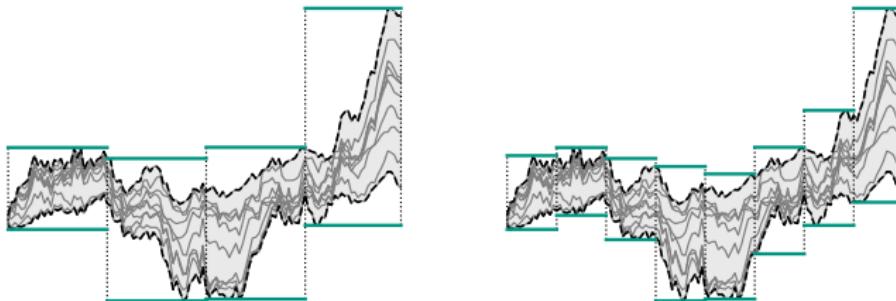
Evaluation

k -NN Querying



Tree Characteristics

- height-**balanced**
- envelopes are stored **segmented**
 - saves runtime and storage
 - higher resolution in deeper tree levels



Index Creation

- **aims:** well-filled leaf envelopes
with minimum area and minimum overlap
- **novel clustering algorithm** based on k -means
 - for splitting nodes
 - supports a minimum cluster size
- **reinsertion** of time series to reorganize the tree

Implementation



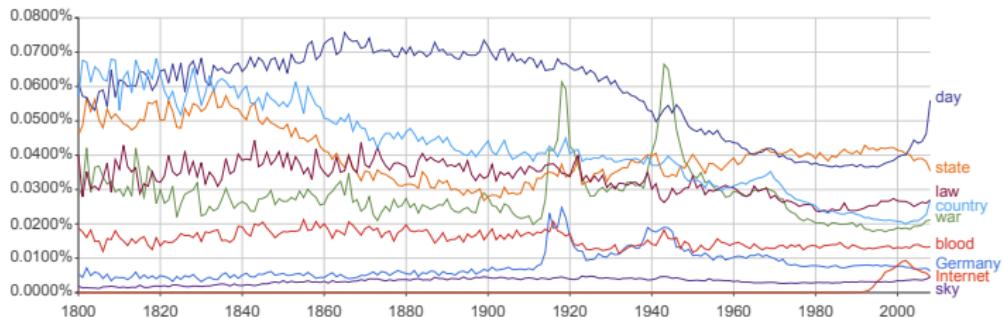
- time series and index are stored in a **PostgreSQL table**
- a PL/**Python** trigger function maintains the index
- a PL/Python function performs the k -NN queries

— 3 —

Evaluation

Datasets

- based on the *Google Books American English n-gram* datasets
- time series describe rel. number of occurrences per *n*-gram and year



n-gram

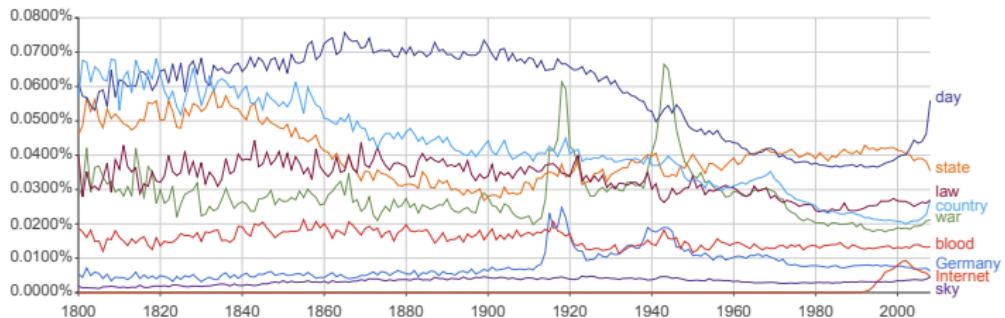
Contiguous sequence of *n* words from a given sentence.

Example: “*This is great*” contains

- 1-grams: “*This*”, “*is*”, “*great*”
- 2-grams: “*This is*”, “*is great*”

Datasets

- based on the *Google Books American English n-gram* datasets
- time series describe rel. number of occurrences per *n*-gram and year



■ ~ 7 million time series

- t.s. length: 209*
- 1-grams
- 11 GB

■ ~ 103 million time series

- t.s. length: 209*
- 2-grams
- 155 GB

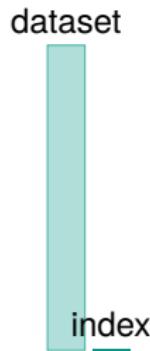
* corresponds to the years 1800–2008

Index Size

- ~ 8 MB per 1 million time series or
 $\sim 0.5\%$ of the dataset size*

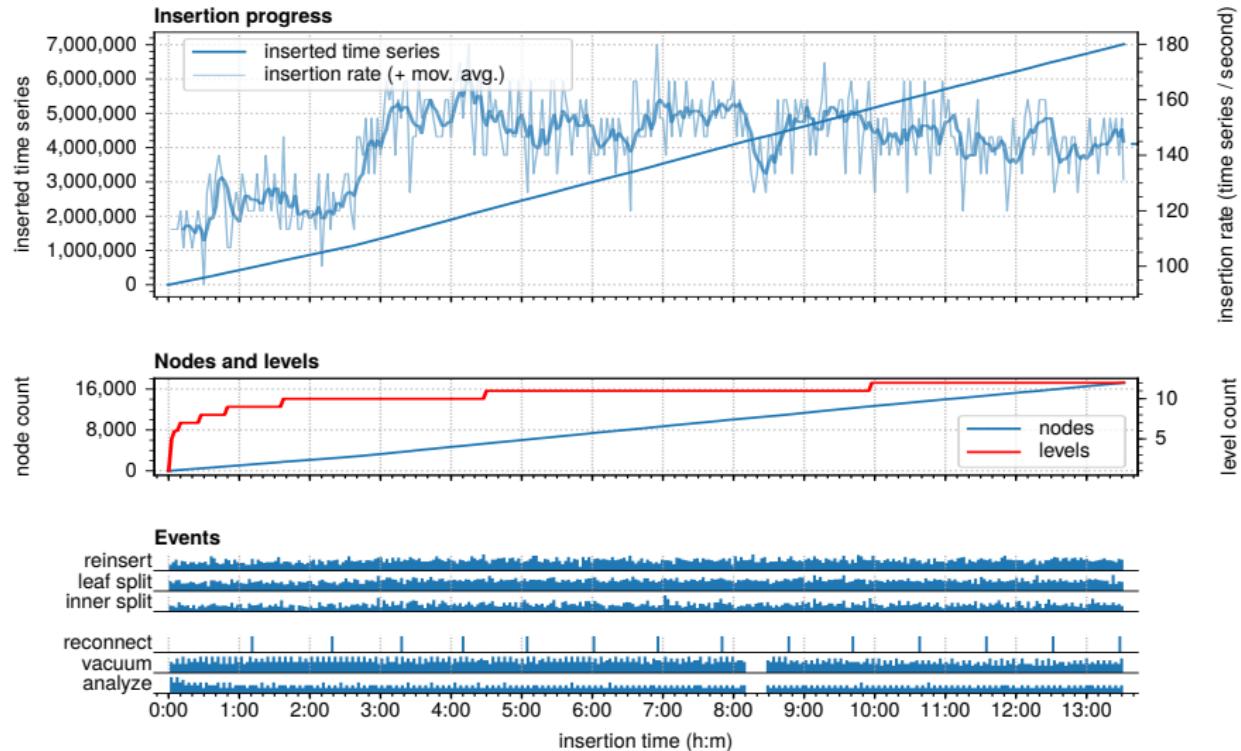
- 7-million dataset: 56 MB
- 103-million dataset: 821 MB

→ index **fits into memory**

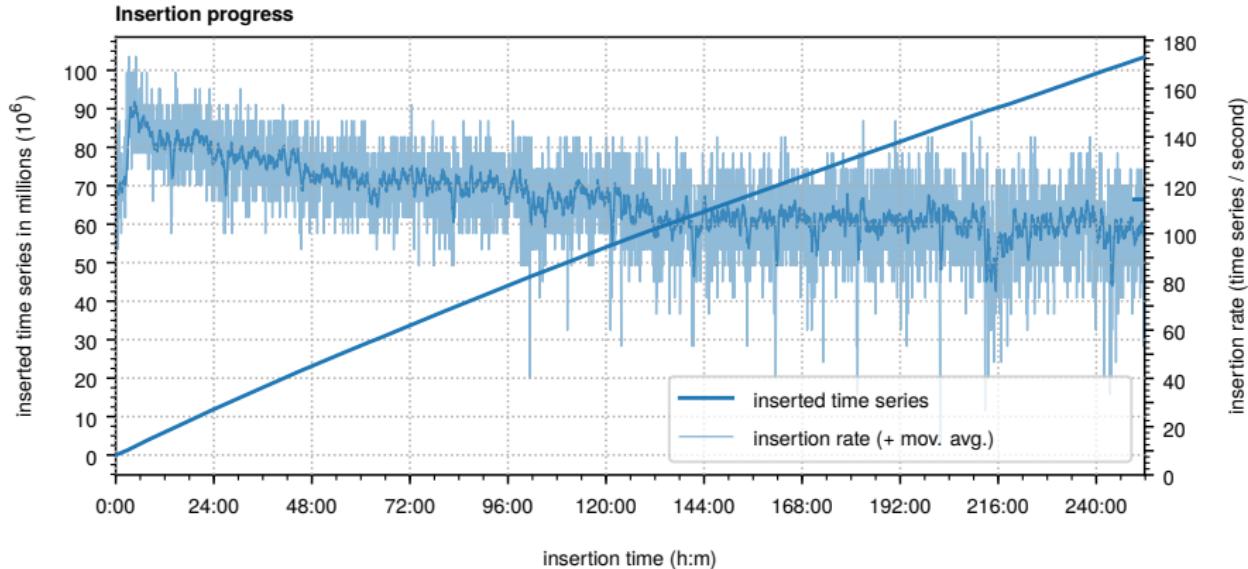


* default configuration
minimum segment length: 1,
max. 1000 time series per leaf

Indexing 7 Million Time Series



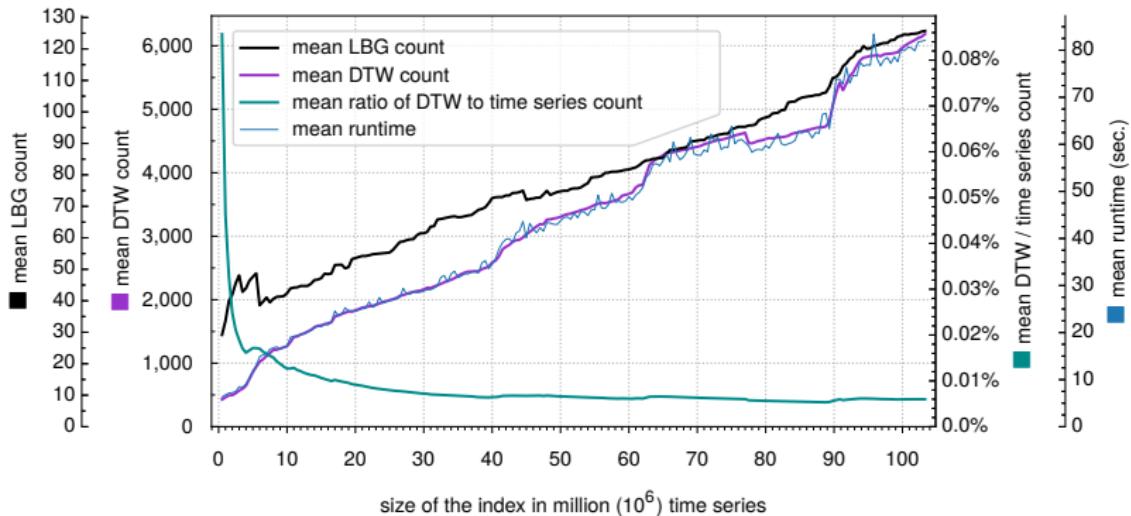
Indexing 103 Million Time Series



- insertion rate decreases only linearly by a small factor
- mean insertion rate of 113.3 ts/sec

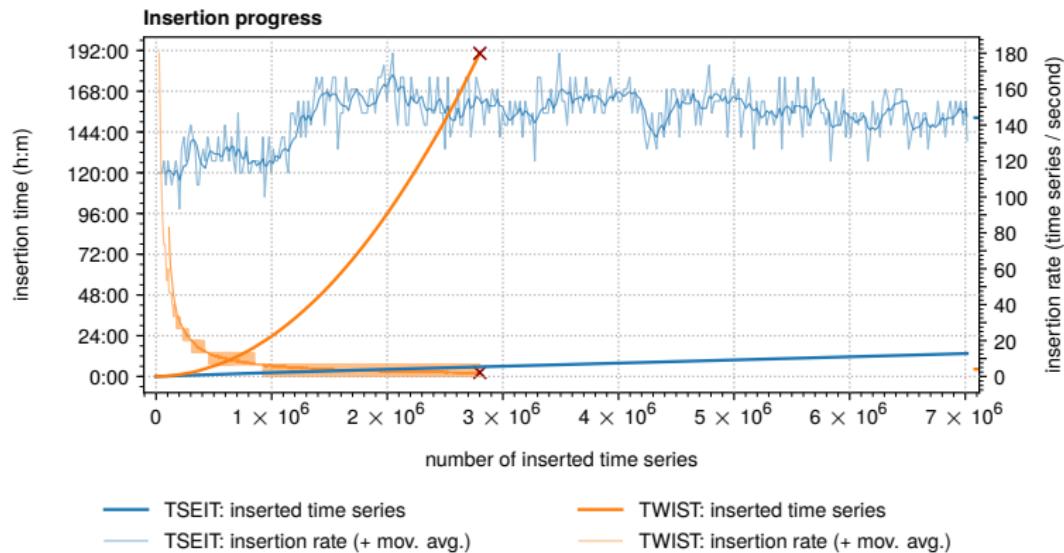
Querying up to 103 Million Time Series

for 100 sample 1-NN queries:



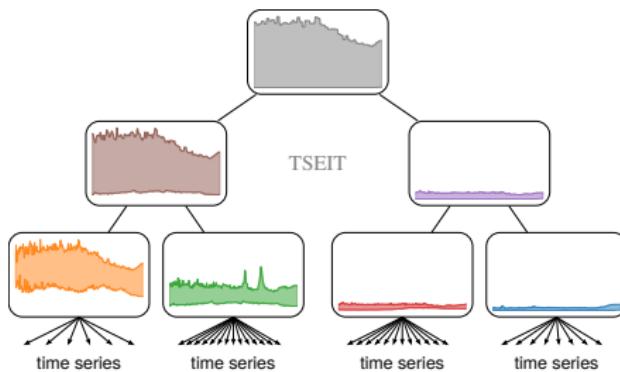
- query runtime is mainly affected by the number of DTW calculations
- number of DTW calculations settles at 0.006% of all time series

Comparison between TSEIT and TWIST



- TWIST scales poorly: insertion rate decreases logarithmically
- in 13.5 hours TWIST: 0.75 million time series
 TSEIT: 7 million time series

TSEIT enables the efficient k -NN search of time series
in arbitrary time intervals without any false dismissals
by storing groups of similar time series in a height-balanced tree.



Thank you!

Exemplary k-NN Queries — 7 Million Dataset

```
# SELECT * FROM knn('and', start_index => -1, end_index => -1,
#                   k => 4, radius => -1, verbose => True);
get time series...
[knn bsf: inf      ] level 10 - node 5318          LBG dist: 0.10528418421862709
[knn bsf: inf      ] level 10 - node 12631         LBG dist: 0.0
[knn bsf: inf      ] level 9 - node 5319          LBG dist: 0.0
[knn bsf: inf      ] level 9 - node 8744          LBG dist: 0.1052840820344287
[knn bsf: inf      ] level 8 - node 1777          LBG dist: 0.10528379661854458
[knn bsf: inf      ] level 8 - node 5317          LBG dist: 0.0
[knn bsf: inf      ] level 8 - node 12597         LBG dist: 0.10528373423066201
...
[knn bsf: inf      ] level 2 - node 134           LBG dist: 0.1276305340647704
[knn bsf: inf      ] level 2 - node 6622          LBG dist: 0.0
[knn bsf: inf      ] level 2 - node 14926         LBG dist: 0.12407300009380696
[knn bsf: inf      ] level 1 - node 6482          LBG dist: 0.0
[knn bsf: inf      ] level 1 - node 6621          LBG dist: 0.09944154604173609
[knn bsf: inf      ] level 0 - node 6448          LBG dist: 0.0
[knn bsf: inf      ] level 0 - node 6481          LBG dist: 0.051786927665999495
[knn bsf: 2.25377e-02] level 0 - node 6448         min. DTW: 0.001016920096783093
```

LBG calculations: 28
 examined inner nodes: 11
 examined leaf nodes: 1
 LB_Kim calculations: 5
 DTW calculations: 5

k	ts_id	name	distance	start_index	end_index	stats
1	6076076	to	0.00101692009678309	0	208	...
2	2896104	in	0.0155376628564426	0	208	...
3	459271	a	0.0220354034133428	0	208	...
4	4412877	of	0.0225376950785585	0	208	...

Time: 367.575 ms

Exemplary k-NN Queries — 7 Million Dataset

```
# SELECT * FROM knn('Microsoft', start_index => 190, end_index => -1,
#                   k => 5, radius => -1, verbose => True);
get time series...
[knn bsf: inf      ] level 10 - node 5318          LBG dist: 2.0242240655347458e-10
[knn bsf: inf      ] level 10 - node 12631         LBG dist: 0.0
[knn bsf: inf      ] level  9 - node 5319          LBG dist: 0.0
[knn bsf: inf      ] level  9 - node 8744          LBG dist: 2.010737231312652e-10
...
[knn bsf: 1.79091e-10] level  1 - node 170          LBG dist: 1.1233211648369213e-09
[knn bsf: 1.79091e-10] level  1 - node 9762         LBG dist: 3.1510635881383756e-10
[knn bsf: 1.79091e-10] level  0 - node 7368          min. DTW: 8.842193648181429e-10
[knn bsf: 1.79091e-10] level  2 - node 301           LBG dist: 4.172609075245131e-10
[knn bsf: 1.79091e-10] level  2 - node 13576         LBG dist: 4.3011495719803036e-10
[knn bsf: 1.79091e-10] level  3 - node 1068          LBG dist: 4.808619308227447e-10
[knn bsf: 1.79091e-10] level  3 - node 5487          LBG dist: 3.928554057541039e-10
[knn bsf: 1.79091e-10] level  3 - node 2317          LBG dist: 4.489510651596557e-10
[knn bsf: 1.79091e-10] level  3 - node 16387         LBG dist: 4.204858013159738e-10
[knn bsf: 1.79091e-10] level  3 - node 16976         LBG dist: 4.4345386367506687e-10
```

LBG calculations: 75
 examined inner nodes: 31
 examined leaf nodes: 12
 LB_Kim calculations: 7989
 DTW calculations: 2954

k	ts_id	name	distance	start_index	end_index	stats
1	2968781	IP	1.1112064920618e-10	190	208	...
2	1146578	Click	1.1620426968856e-10	190	208	...
3	3146497	Java	1.34770384524882e-10	190	208	...
4	1531883	dialog	1.40257448743664e-10	190	208	...
5	5465057	Server	1.79090613709668e-10	190	208	...

Time: 900.352 ms

Exemplary k-NN Queries — 7 Million Dataset

```
# SELECT * FROM knn('Germany', start_index => 130, end_index => 150,
#                   k => 2, radius => -1, verbose => True);
get time series...
[knn bsf: inf      ] level 10 - node 5318          LBG dist: 2.452010665370238e-07
[knn bsf: inf      ] level 10 - node 12631         LBG dist: 0.0
[knn bsf: inf      ] level 9 - node 5319          LBG dist: 0.0
[knn bsf: inf      ] level 9 - node 8744          LBG dist: 2.451516379639861e-07
...
[knn bsf: inf      ] level 0 - node 9291          LBG dist: 0.0
[knn bsf: inf      ] level 0 - node 14924         LBG dist: 0.0
[knn bsf: 1.04658e-08] level 0 - node 9291        min. DTW: 8.793723645729129e-09
[knn bsf: 5.56403e-09] level 0 - node 14924        min. DTW: 4.720827730487925e-09
[knn bsf: 5.56403e-09] level 1 - node 7720          LBG dist: 0.0
[knn bsf: 5.56403e-09] level 1 - node 14925         LBG dist: 0.0
[knn bsf: 5.56403e-09] level 0 - node 127           LBG dist: 9.584258272321103e-07
[knn bsf: 5.56403e-09] level 0 - node 6966           LBG dist: 0.0
[knn bsf: 3.43002e-09] level 0 - node 6966           min. DTW: 3.044039632242908e-09
[knn bsf: 3.43002e-09] level 0 - node 128            LBG dist: 0.0
[knn bsf: 3.43002e-09] level 0 - node 7719            LBG dist: 0.0
[knn bsf: 3.04404e-09] level 0 - node 128            min. DTW: 1.5346986393409813e-09
[knn bsf: 3.04404e-09] level 0 - node 7719            min. DTW: 4.5235610212726406e-09

LBG calculations:      34
examined inner nodes:  14
examined leaf nodes:   5
LB_Kim calculations: 2358
DTW calculations:    504
```

k	ts_id	name	distance	start_index	end_index	stats
1	4756547	peace	1.53469863934098e-09	130	150	...
2	583527	blood	3.04403963224291e-09	130	150	...

Time: 191.804 ms

Exemplary k -NN Queries — 7 Million Dataset

```

# SELECT * FROM knn('Germany', start_index => 155, end_index => -1,
#                   k => 3, radius => -1, verbose => True);
get time series...
[knn bsf: inf      ] level 10 - node 5318          LBG dist: 2.1535228272938325e-07
[knn bsf: inf      ] level 10 - node 12631         LBG dist: 0.0
[knn bsf: inf      ] level 9  - node 5319          LBG dist: 0.0
[knn bsf: inf      ] level 9  - node 8744          LBG dist: 2.152780041990043e-07
...
[knn bsf: inf      ] level 1  - node 129           LBG dist: 0.0
[knn bsf: inf      ] level 1  - node 133           LBG dist: 8.348610342357409e-10
[knn bsf: inf      ] level 0  - node 9291          LBG dist: 0.0
[knn bsf: inf      ] level 0  - node 14924         LBG dist: 0.0
[knn bsf: 6.57233e-10] level 0  - node 9291          min. DTW: 4.740883488792752e-10
[knn bsf: 4.79416e-10] level 0  - node 14924         min. DTW: 4.681301632545856e-10
[knn bsf: 4.79416e-10] level 1  - node 7720          LBG dist: 7.530247245002756e-10
[knn bsf: 4.79416e-10] level 1  - node 14925          LBG dist: 0.0
[knn bsf: 4.79416e-10] level 0  - node 128           LBG dist: 0.0
[knn bsf: 4.79416e-10] level 0  - node 7719          LBG dist: 0.0
[knn bsf: 4.79416e-10] level 0  - node 128           min. DTW: 6.012067502745996e-10
[knn bsf: 4.79416e-10] level 0  - node 7719          min. DTW: 7.180280417408994e-10

LBG calculations:      32
examined inner nodes:  13
examined leaf nodes:   4
LB_Kim calculations: 2142
DTW calculations:    244

k | ts_id | name | distance | start_index | end_index | stats
---+---+---+---+---+---+---+---+---+
1 | 4831023 | Paris | 4.68130163254586e-10 | 155 | 208 | ...
2 | 1430013 | demand | 4.74088348879275e-10 | 155 | 208 | ...
3 | 2269496 | freedom | 4.79415545679705e-10 | 155 | 208 | ...

Time: 474.690 ms

```

Exemplary k-NN Queries — 103 Million Dataset

```
# SELECT * FROM knn('the United', start_index => -1, end_index => -1,
#                   k => 1, radius => -1, verbose => True);
get time series...
[knn bsf: inf      ] level 14 - node 119218          LBG dist: 1.826339364954095e-07
[knn bsf: inf      ] level 14 - node 235501          LBG dist: 0.0
[knn bsf: inf      ] level 13 - node 119219          LBG dist: 1.8252825135969383e-07
[knn bsf: inf      ] level 13 - node 235499          LBG dist: 0.0
[knn bsf: inf      ] level 12 - node 119217          LBG dist: 1.82325843055561e-07
[knn bsf: inf      ] level 12 - node 235498          LBG dist: 0.0
...
[knn bsf: inf      ] level  2 - node 14              LBG dist: 0.0
[knn bsf: inf      ] level  2 - node 456             LBG dist: 1.1185841457901548e-06
[knn bsf: inf      ] level  2 - node 222965          LBG dist: 6.997132874688362e-07
[knn bsf: inf      ] level  1 - node 6              LBG dist: 0.0
[knn bsf: inf      ] level  1 - node 9              LBG dist: 2.2953788473650518e-07
[knn bsf: inf      ] level  1 - node 227024          LBG dist: 3.0835804262542576e-09
[knn bsf: inf      ] level  0 - node 2              LBG dist: 0.02189835768101354
[knn bsf: inf      ] level  0 - node 3              LBG dist: 0.0
[knn bsf: 2.96448e-08] level  0 - node 3          min. DTW: 2.964480660174612e-08
[knn bsf: 2.96448e-08] level  0 - node 4          LBG dist: 4.347839287126069e-09
[knn bsf: 2.96448e-08] level  0 - node 224637          LBG dist: 9.464199988897262e-08
[knn bsf: 2.96448e-08] level  0 - node 4          min. DTW: 1.9423857190330249e-07
```

LBG calculations: 39
 examined inner nodes: 16
 examined leaf nodes: 2
 LB_Kim calculations: 1191
 DTW calculations: 295

k	ts_id	name	distance	start_index	end_index	stats
1	96222167	United States	2.96448066017461e-08	0	208	...

Time: 7600.031 ms (00:07.600)

Exemplary k-NN Queries — 103 Million Dataset

```
# SELECT * FROM knn('the United', start_index => 100, end_index => 130,
#                   k => 2, radius => -1, verbose => True);
get time series...
[knn bsf: inf      ] level 14 - node 119218          LBG dist: 8.062464804207799e-07
[knn bsf: inf      ] level 14 - node 235501          LBG dist: 0.0
[knn bsf: inf      ] level 13 - node 119219          LBG dist: 8.061609508853667e-07
[knn bsf: inf      ] level 13 - node 235499          LBG dist: 0.0
[knn bsf: inf      ] level 12 - node 119217          LBG dist: 8.059970881347681e-07
[knn bsf: inf      ] level 12 - node 235498          LBG dist: 0.0
...
[knn bsf: inf      ] level  3 - node 86027           LBG dist: 8.866930945679905e-07
[knn bsf: inf      ] level  3 - node 221902          LBG dist: 8.236557026781433e-07
[knn bsf: inf      ] level  2 - node 14              LBG dist: 0.0
[knn bsf: inf      ] level  2 - node 456             LBG dist: 1.0285315664521846e-06
[knn bsf: inf      ] level  2 - node 222965          LBG dist: 9.223030029750501e-07
[knn bsf: inf      ] level  1 - node 6               LBG dist: 0.0
[knn bsf: inf      ] level  1 - node 9               LBG dist: 5.878268680118461e-07
[knn bsf: inf      ] level  1 - node 227024          LBG dist: 3.0255240544908576e-08
[knn bsf: inf      ] level  0 - node 2               LBG dist: 0.0035411292104014776
[knn bsf: inf      ] level  0 - node 3               LBG dist: 0.0
[knn bsf: 1.81180e-08] level  0 - node 3           min. DTW: 1.0009375719301362e-08
```

LBG calculations: 37
 examined inner nodes: 15
 examined leaf nodes: 1
 LB_Kim calculations: 325
 DTW calculations: 185

k	ts_id	name	distance	start_index	end_index	stats
1	60130037	New York	1.00093757193014e-08	100	130	...
2	96222167	United States	1.81180182953422e-08	100	130	...

Time: 108.803 ms

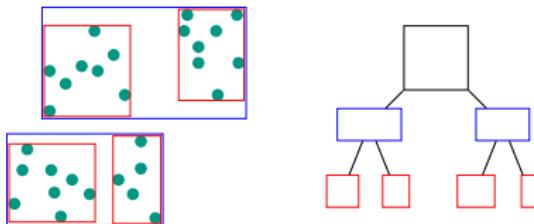
Fundamentals — Data Structures

TSEIT borrows concepts of:

■ R*-Tree

index for multidimensional data

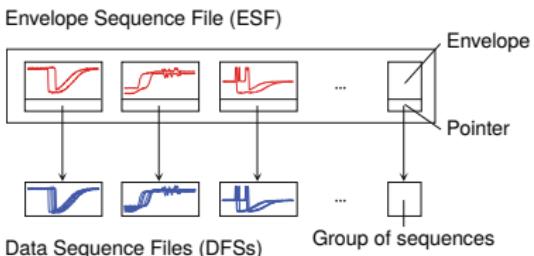
Beckmann et al: "The R*-tree: An Efficient and Robust Access Method for Points and Rectangles" (1990)



■ TWIST

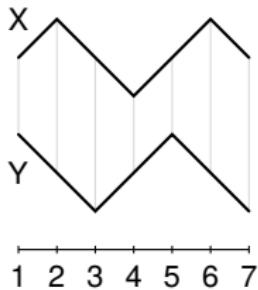
index for time series

Niennattrakul, Pongsakorn and Ratanamahatana: "Exact Indexing for Massive Time Series Databases under Time Warping Distance" (Nov. 2010)

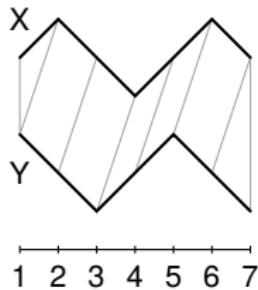


Fundamentals — Dynamic Time Warping

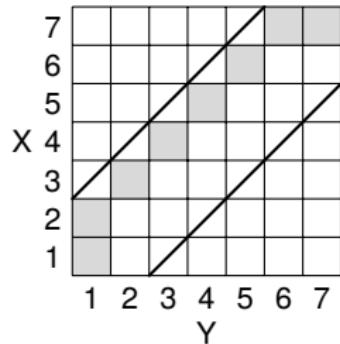
- probably the best distance measure for time series
- computational intensive: $\mathcal{O}(n^2)$ [n : length of time series]



(a) Euclidean distance



(b) DTW distance



(c) DTW distance matrix