Insight mining in time series data with applications for anomaly detection

Dieter De Paepe

GHENT UNIVERSITY

Insight mining in **time series data** with applications for anomaly detection

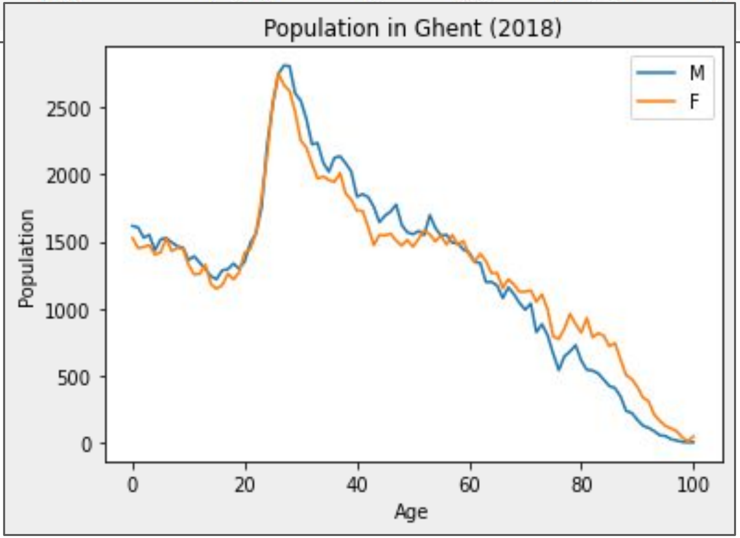
Traditional Data

Rows are records

Columns are features

Can be visualized

	CD_MUNTY_REFNIS	TX_MUNTY_DESCR_NL	CD_SEX	CD_NATLTY	TX_CIV_STS_NL	CD_AGE	MS_POPULATION
0	71024	Herk-de-Stad	F	BEL	Gehuwd	39	42
1	71037	Lummen	M	BEL	Gehuwd	82	24
2	71011	Diepenbeek	F	BEL	Gehuwd	42	51
3	71016	Genk	M	BEL	Gehuwd	63	277
4	71017	Gingelom	F	BEL	Gehuwd	30	14
...	...	...	...	...	...	...	...
465413	92141	La Bruyère	F	ETR	Gescheiden	64	1
465414	46024	Stekene	M	ETR	Gescheiden	67	2
465415	46024	Stekene	M	ETR	Gescheiden	74	1
465416	46024	Stekene	M	ETR	Gescheiden	81	1
465417							1



Source: Statbel

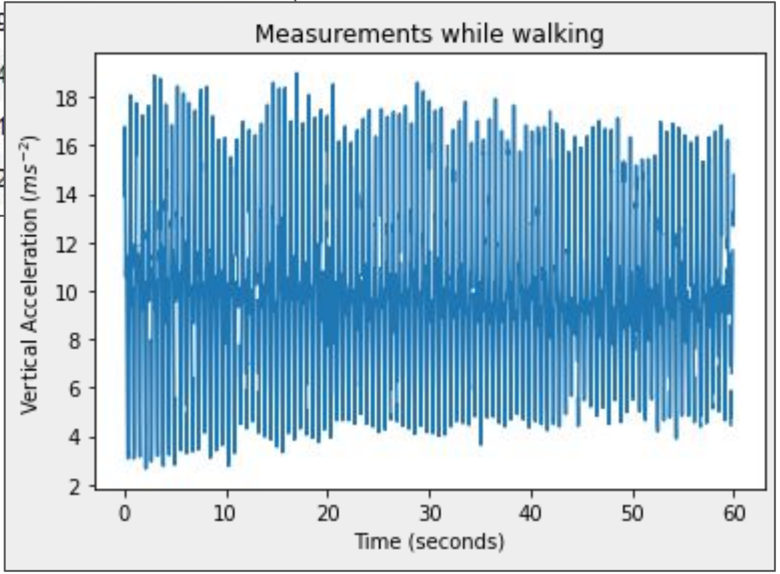
Time Series

Rows are records

Columns are features

Can be visualized

	time	x acc	y acc	z acc
227000	2278.38	0.829437	13.9132	-0.104062
227001	2278.39	0.468504	14.0607	-0.730616
227002	2278.40	-0.172953	15.0715	-1.907150
227003	2278.41	-0.396788	16.7580	-3.194480
227004	2278.42	-0.708866	16.6444	-3.432090
...	...	...	...	...
232996	2338.34	-0.106280	13.1876	-2.054700
232997	2338.35	-0.24809 [illegible]	[illegible]	[illegible]
232998	2338.36	0.08824 [illegible]	[illegible]	[illegible]
232999	2338.37	0.82391 [illegible]	[illegible]	[illegible]
233000	2338.38	1.50292 [illegible]	[illegible]	[illegible]



Time Series

Show change through time

Often periodic or repetitive

Capture behavior

	time	x acc	y acc	z acc
227000	2278.38	0.829437	13.9132	-0.104062
227001	2278.39	0.468504	14.0607	-0.730616
227002	2278.40	-0.172953	15.0715	-1.907150
227003	2278.41	-0.396788	16.7580	-3.194480
227004	2278.42	-0.708866	16.6444	-3.432090
...	...	...	...	...
232996	2338.34	-0.106280	13.1876	-2.054700
232997	2338.35	-0.24809		
232998	2338.36	0.08824		
232999	2338.37	0.82391		
233000	2338.38	1.50292		

Time Series are everywhere

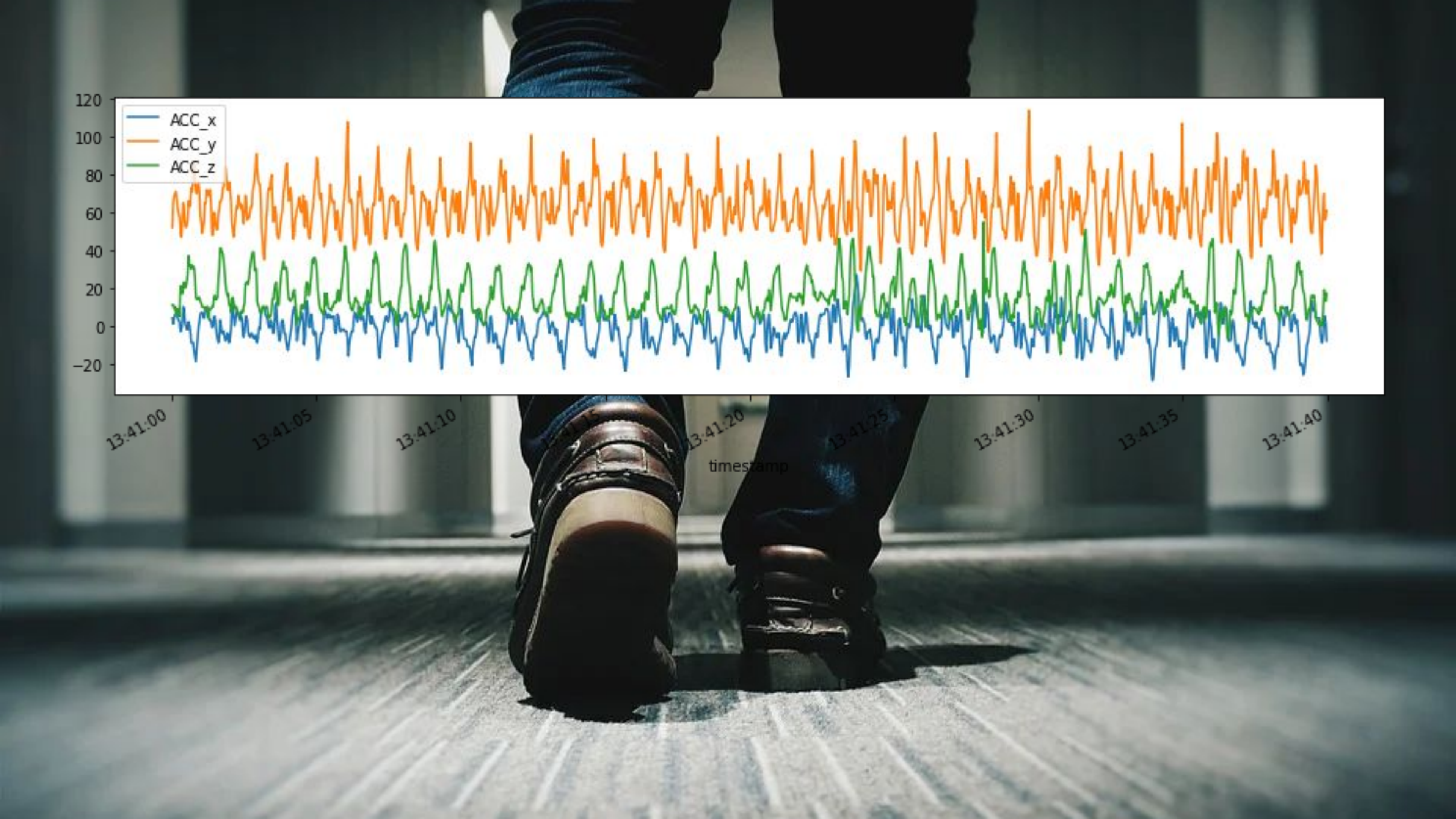
Time series are

omnipresent
new
valuable

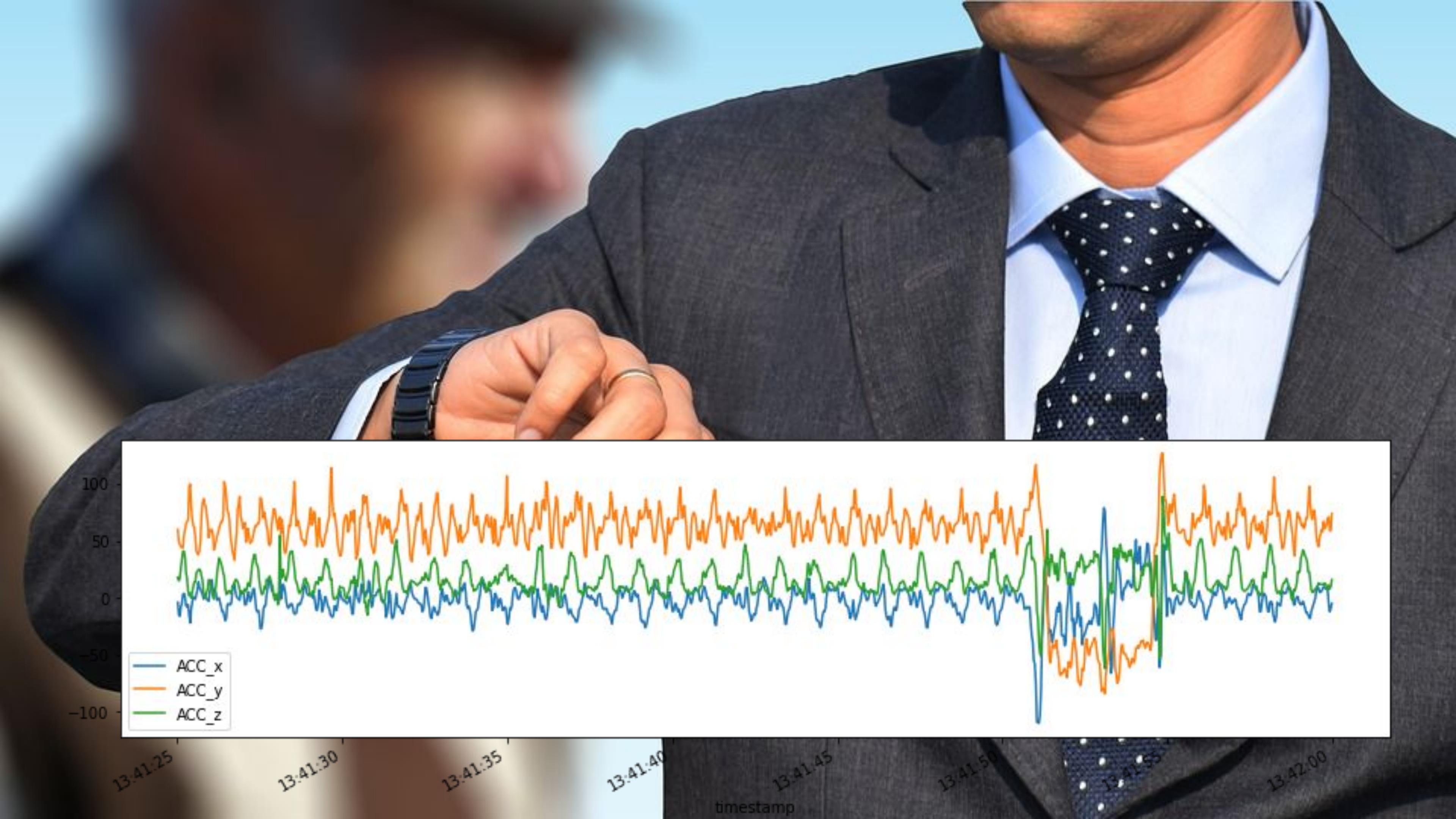
Time series are

omnipresent
new
valuable


ACC_x
ACC_y
ACC_z
100
80
60
40
20
0
−20
13:41:00
13:41:05
13:41:10
13:41:15
13:41:20
timestamp



ACC_x
ACC_y
ACC_z
120
100
80
60
40
20
0
−20
13:41:00
13:41:05
13:41:10
13:41:20
13:41:25
13:41:30
13:41:35
13:41:40
timestamp



100
50
0
−50
−100
ACC_x
ACC_y
ACC_z
13:41:25
13:41:30
13:41:35
13:41:40
13:41:45
13:41:50
13:42:00
timestamp


100
50
0
−50
−100
ACC_x
ACC_y
ACC_z

Verizon
2:17 PM
fitbit
Read the Alta HR 101 Guide
Today
3,380 steps
995
cals
1.37
miles
0
minutes
0
floors
11:26 PM
5:40 AM
76
resting bpm
5 hr 26 min
48 min awake
Dashboard

Time series are

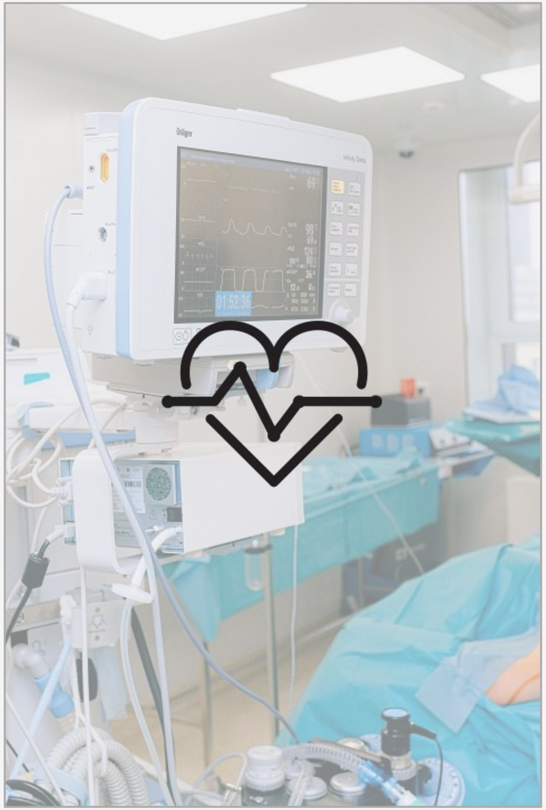
omnipresent

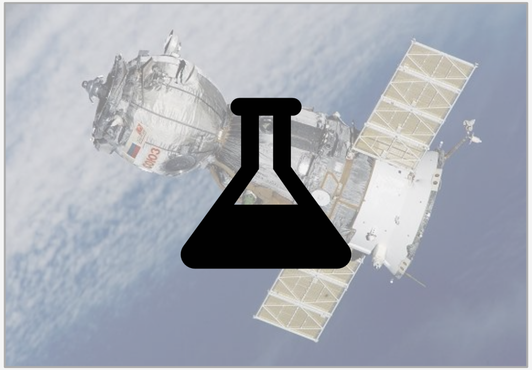
new

valuable

Insight mining | omnipresent
new
valuable

Value

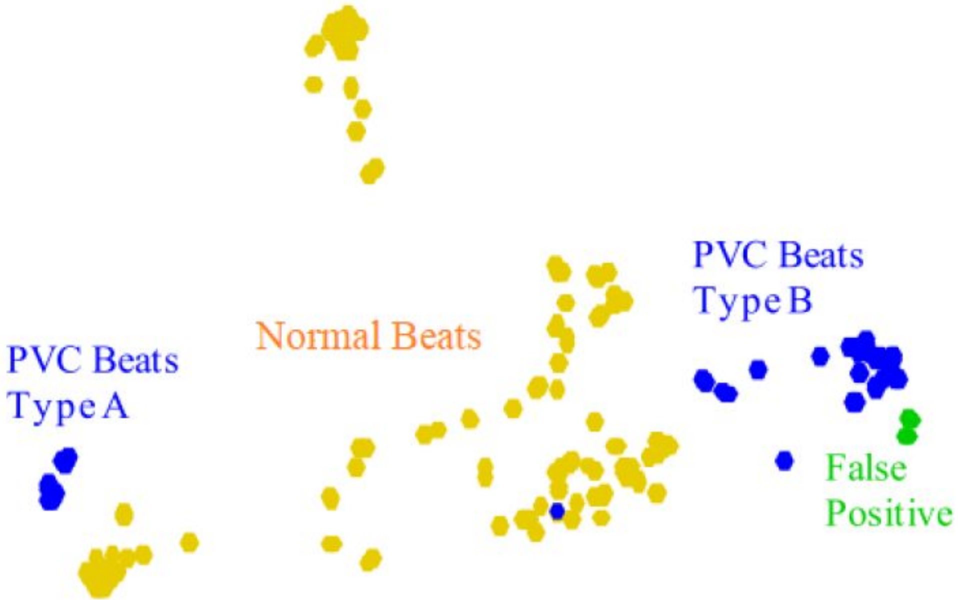


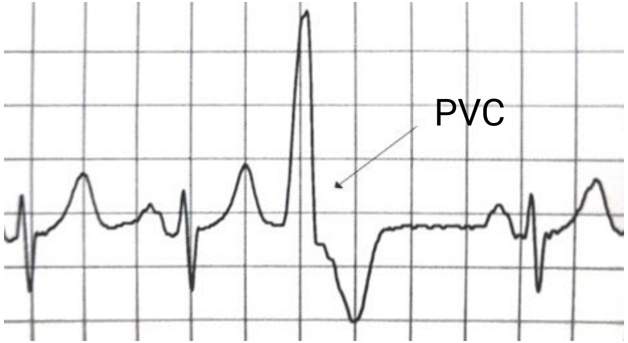


Insight mining in time series data with applications for anomaly detection

Value = anything useful

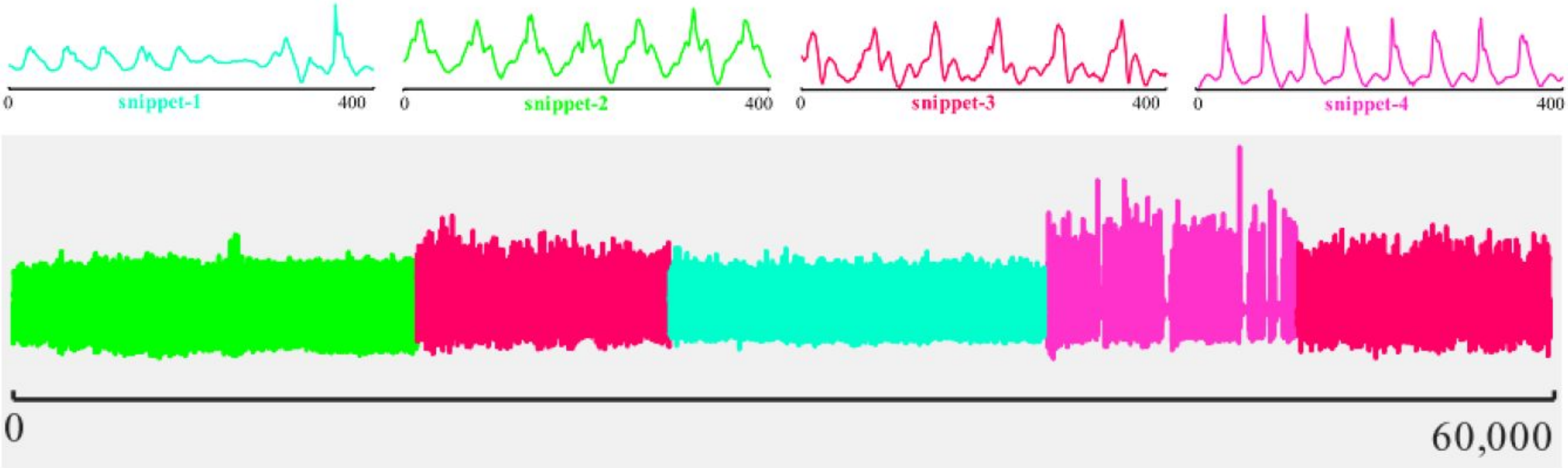
Visualizing content





Value = anything useful

Summarizing content



Value = anything useful

Detecting evolving patterns

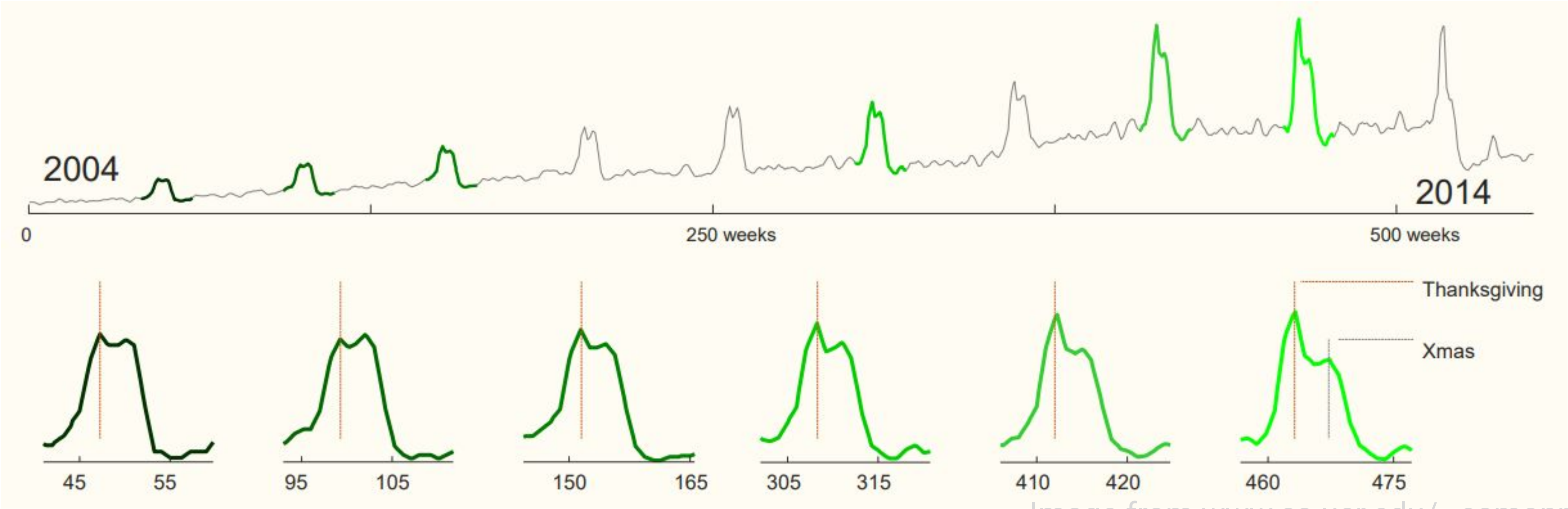
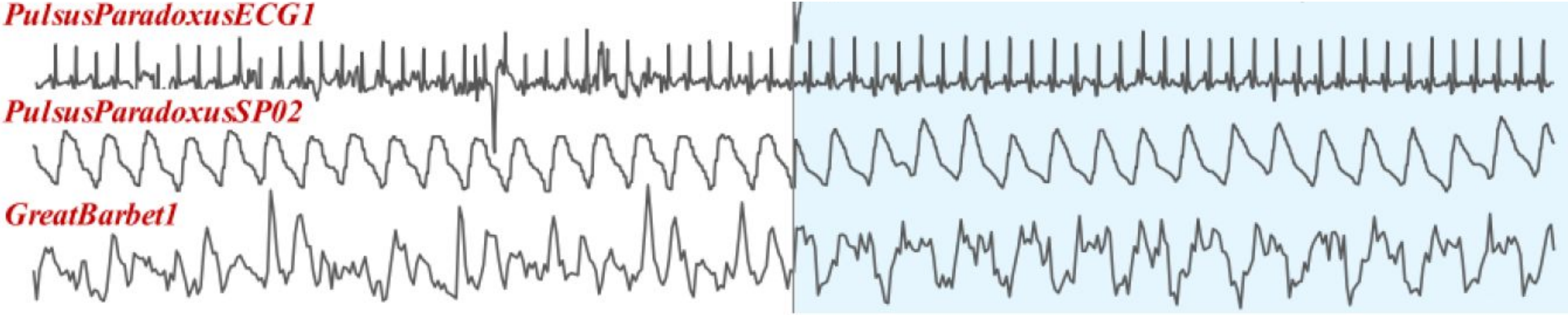


Image from www.cs.ucr.edu/~eamonn/

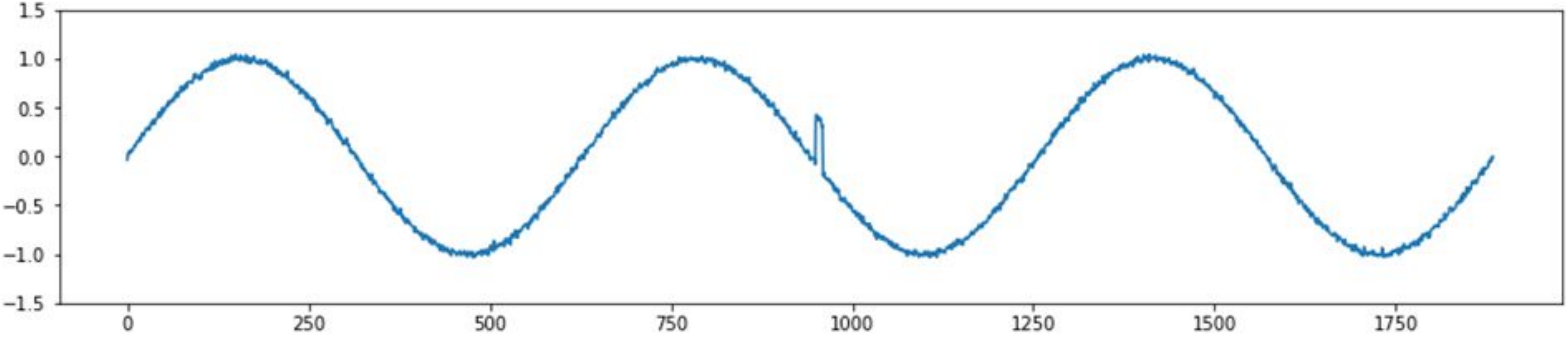
Value = anything useful

Detecting changepoints



Value = anything useful

Detecting anomalies



Insight mining in time series data with applications for **anomaly detection**

Anomaly detection

… for exploration

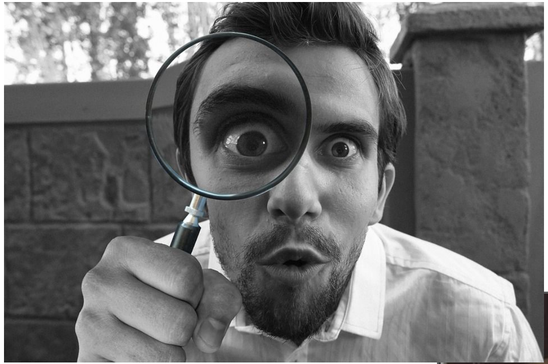
“We didn’t expect that!”

… for prevention

“Check your engine!”

… for reaction

“Call a doctor!”



Anomalies are vague

Highly subjective

E.g. yearly fire drill

Context dependent

E.g. weekdays versus holidays

Instantaneous or long-term

E.g. noise versus different behavior

Insight mining in time series data with applications for anomaly detection

Introduction

Matrix Profile

Contextual Matrix Profile

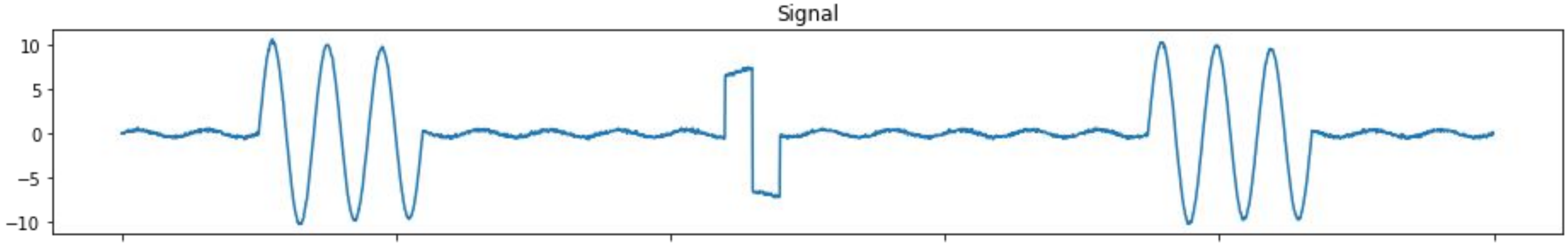
Noise Elimination

Radius Profile

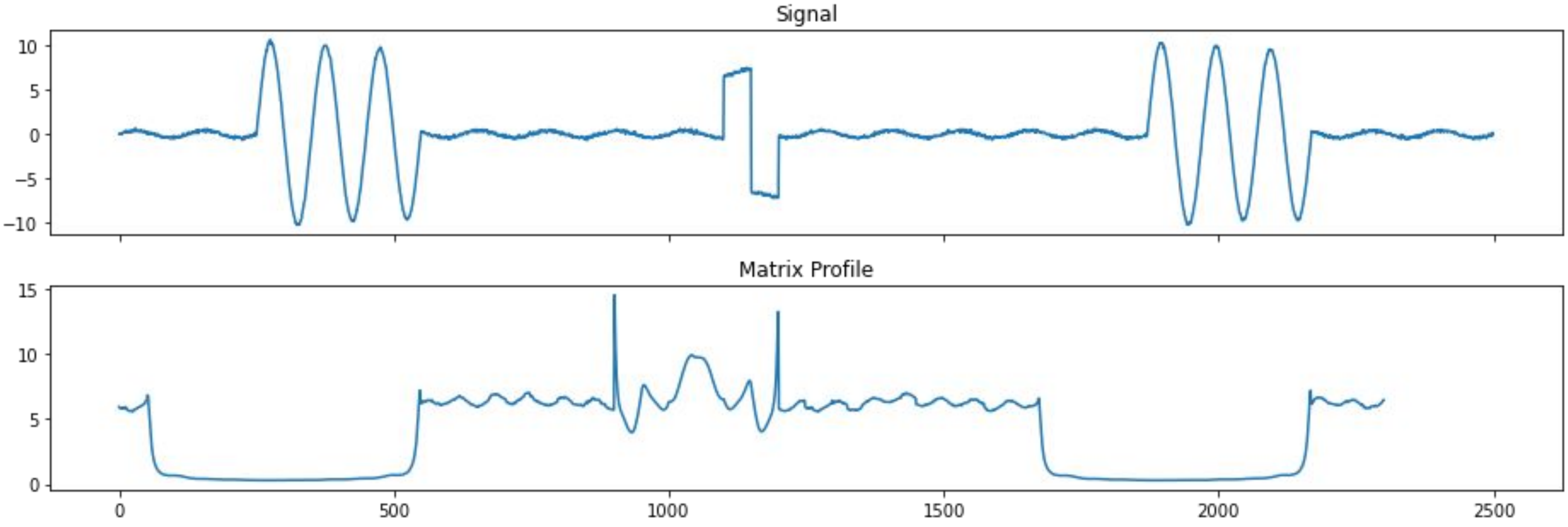
SDM-Framework

Conclusion

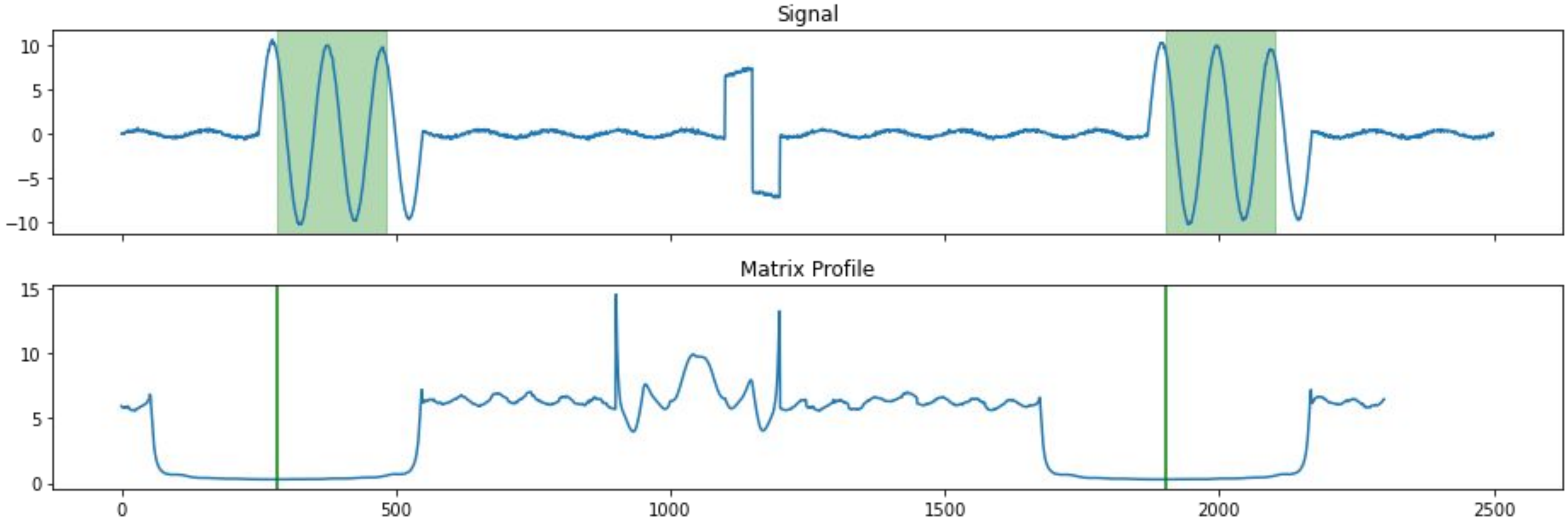
Matrix Profile | Example



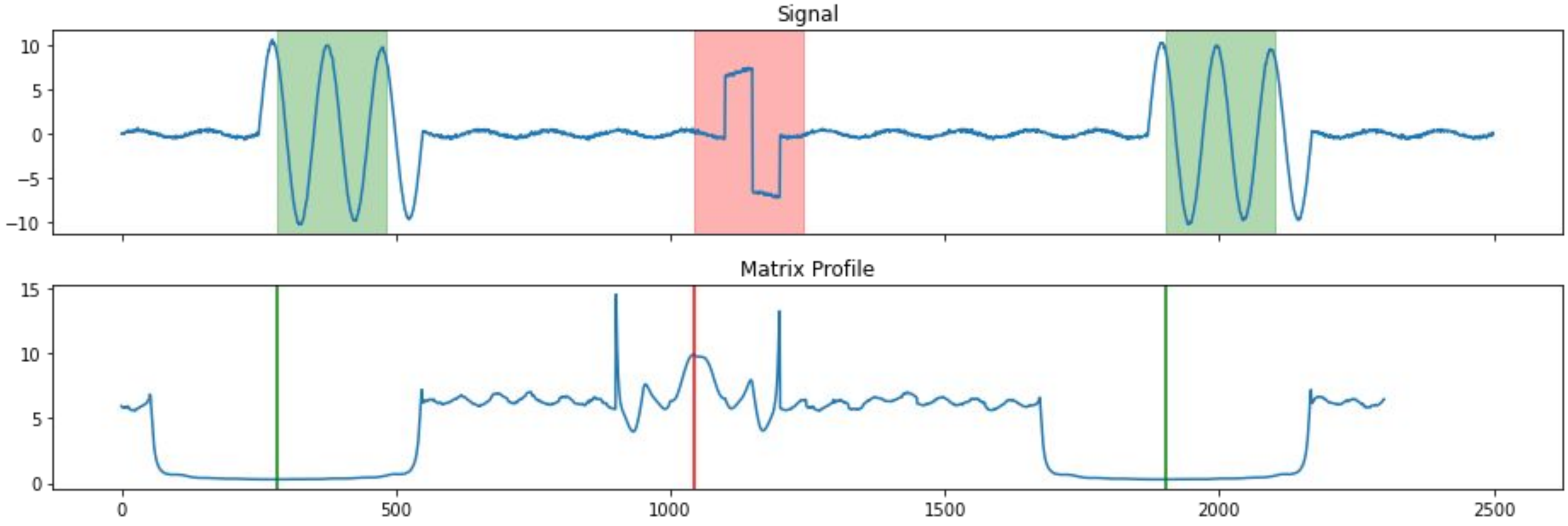
Matrix Profile | Example



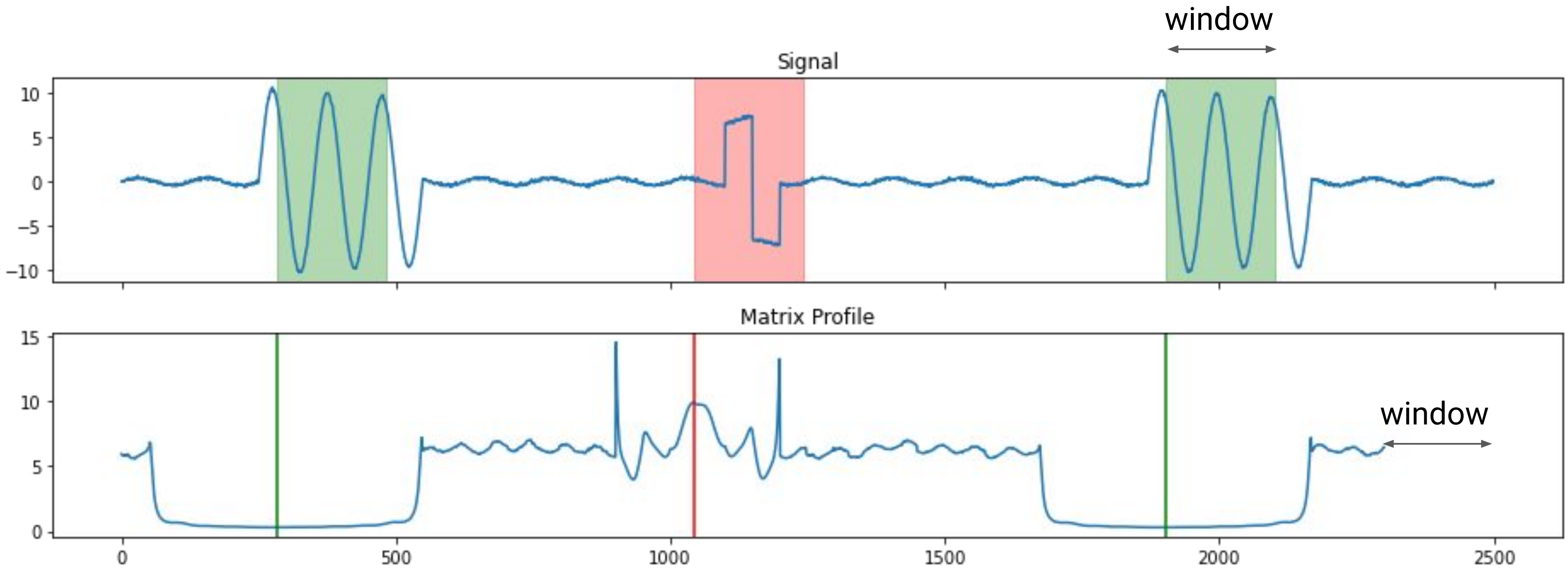
Matrix Profile | Example



Matrix Profile | Example



Matrix Profile | Example



Matrix Profile | Similarities

Given two sequences, define a distance measure

Manhattan distance

$$D_M(X,Y) = \sum_i |x_i - y_i|$$

Euclidean distance

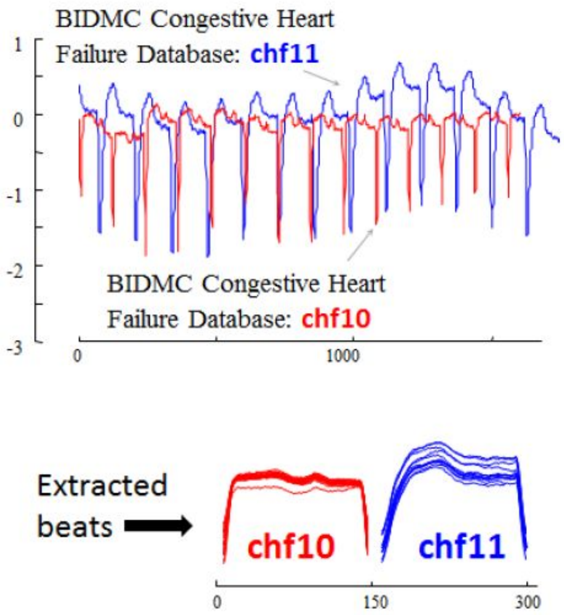
$$D_E(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$$

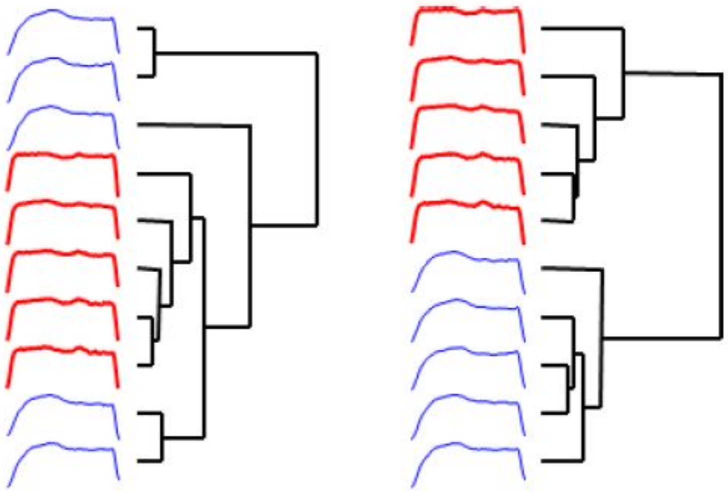
Z-normalized Euclidean distance

$$D_{ZE}(X,Y) = D_E\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right)$$

Matrix Profile | Z-normalized Euclidean Distance

Most used because it compares shape

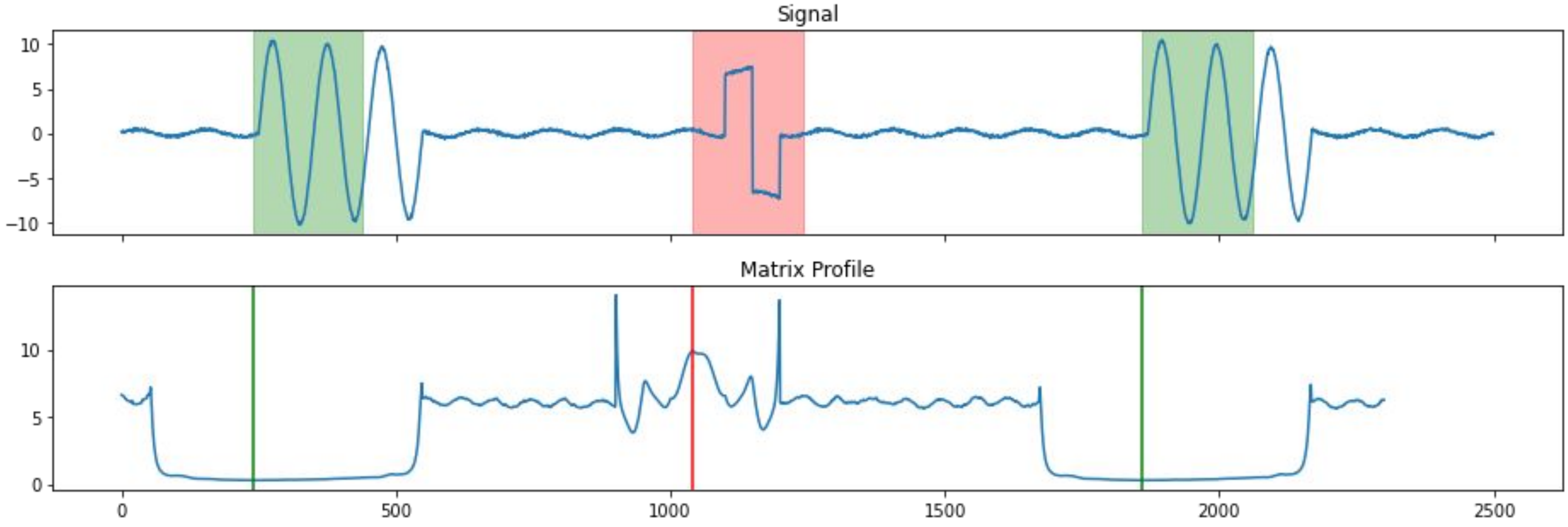




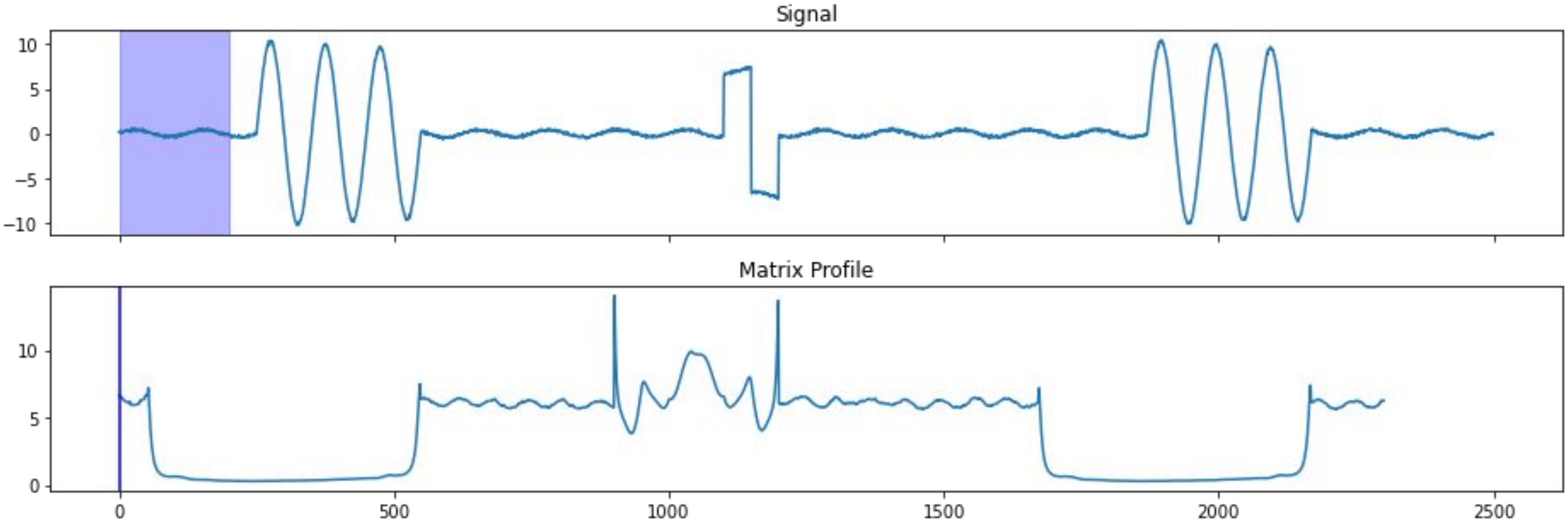
Euclidean Distance

Z-Normalized
Euclidean Distance

Matrix Profile | Calculation

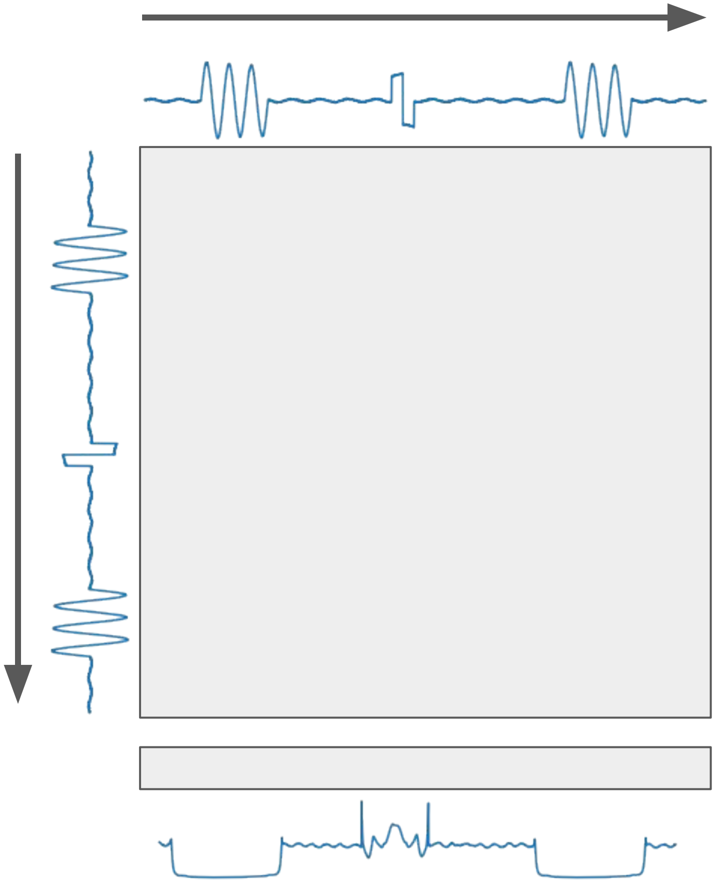


Matrix Profile | Calculation



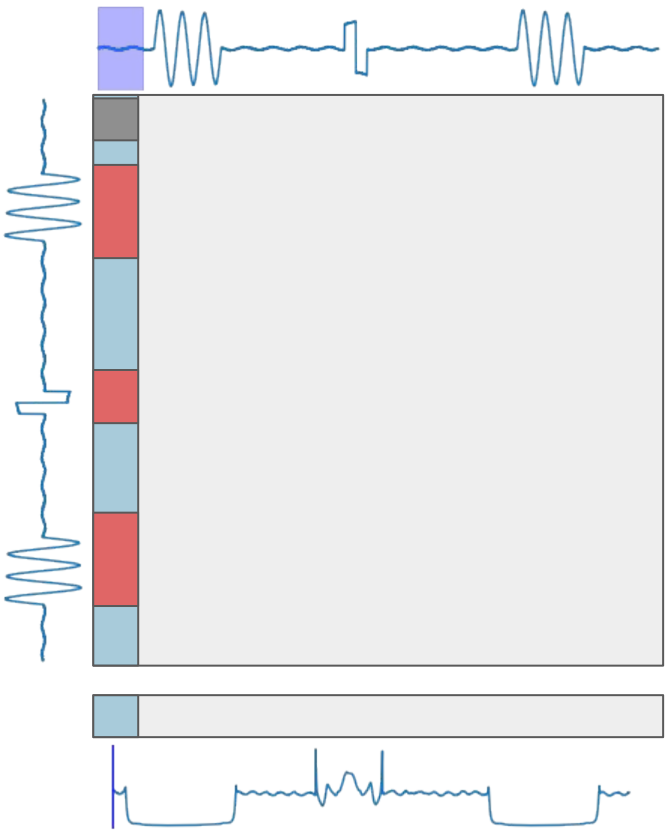
Matrix Profile | Calculation

Distance matrix visualizes all distances



Matrix Profile | Calculation

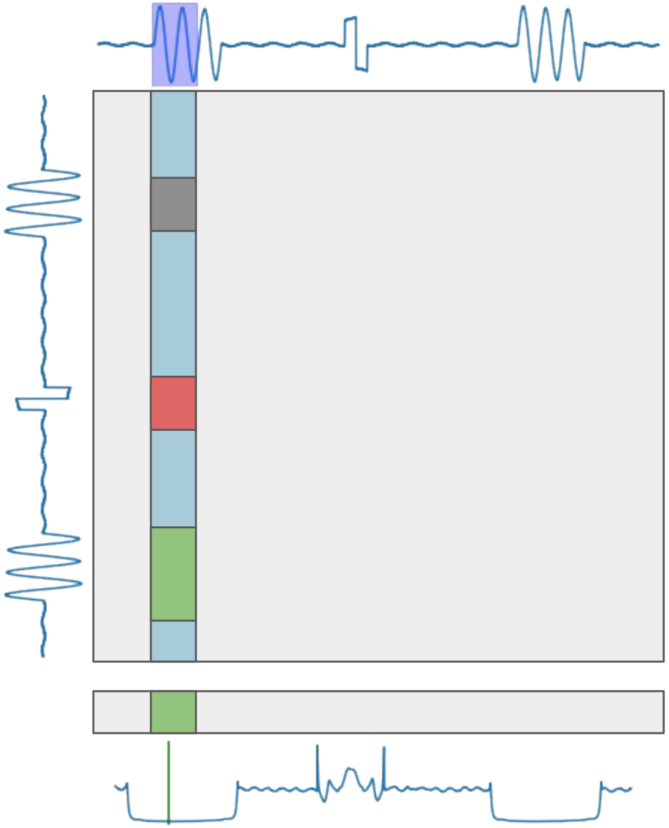
No clear pattern results in neutral distance



Matrix Profile | Calculation

No clear pattern results in neutral distance

Matching pattern gives low distance

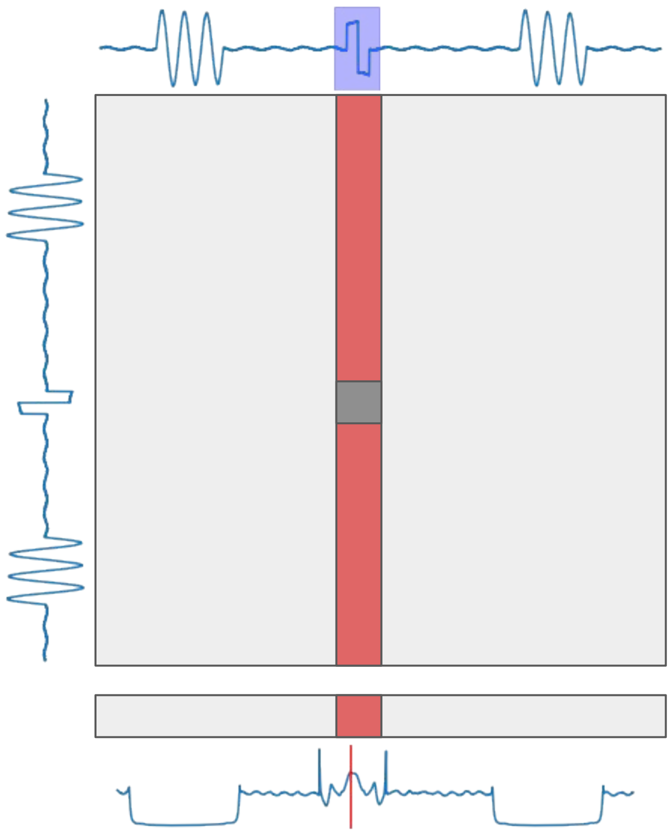


Matrix Profile | Calculation

No clear pattern results in neutral distance

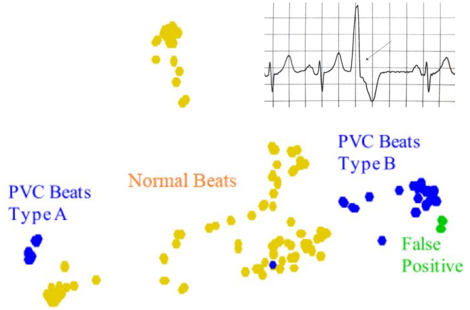
Matching pattern gives low distance

Lone pattern results in high distance

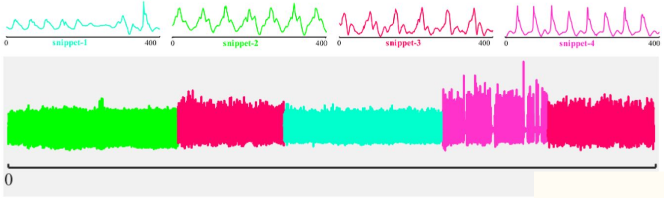


Matrix Profile | Insights

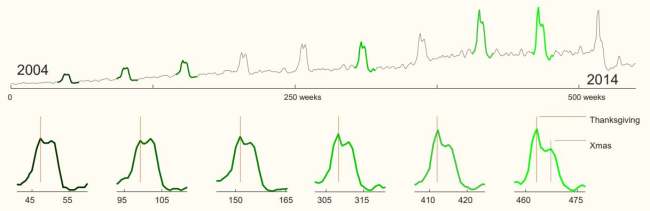
Visualization



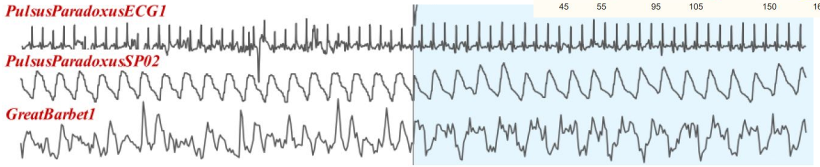
Summarization



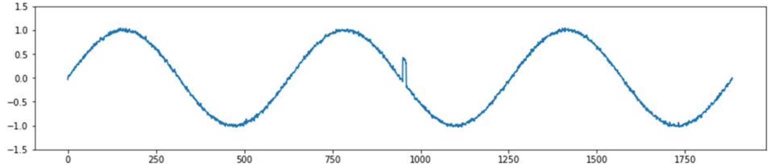
Finding evolving patterns



Segmentation



Anomaly detection



Matrix Profile | Limitations

Periodicity

Discover & visualize

Anomalies

Matrix Profile | Limitations

Periodicity

Noise in signals

- Affects perceived shape
- Impedes insights

Matrix Profile | Limitations

Periodicity

Noise in signals

Repetition

- Across time series
- Within single time series

Matrix Profile | Limitations

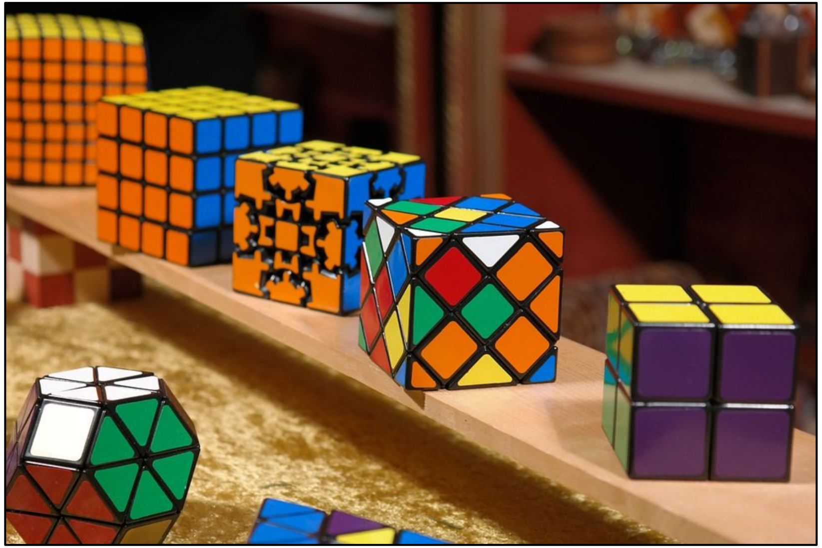
Periodicity

Noise in signals

Repetition

Integration

- Shared functionality
- Single workflow



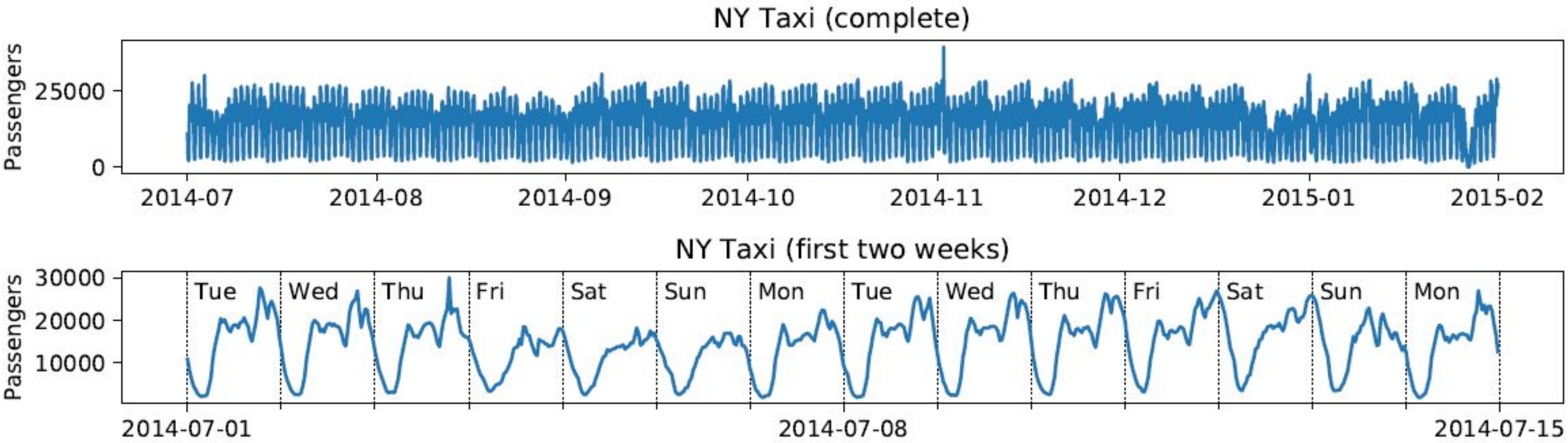
Introduction
Matrix Profile
Contextual Matrix Profile
Noise Elimination
Radius Profile
SDM-Framework
Conclusion

Periodicity
Noise
Repetition
Integration

Introduction
Matrix Profile
Contextual Matrix Profile
Noise Elimination
Radius Profile
SDM-Framework
Conclusion

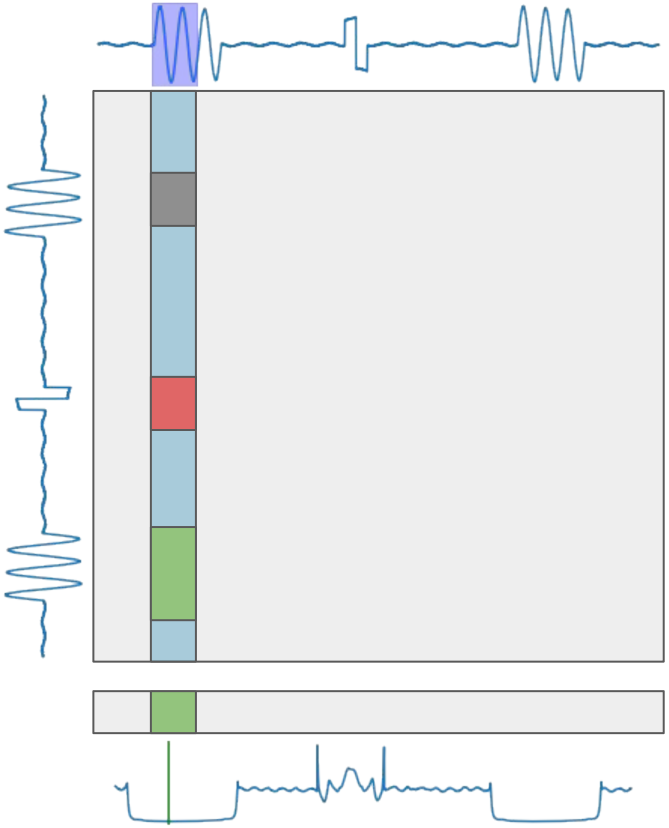
Contextual Matrix Profile | Example

Dataset of taxi passengers in New York



Contextual MP | Calculation

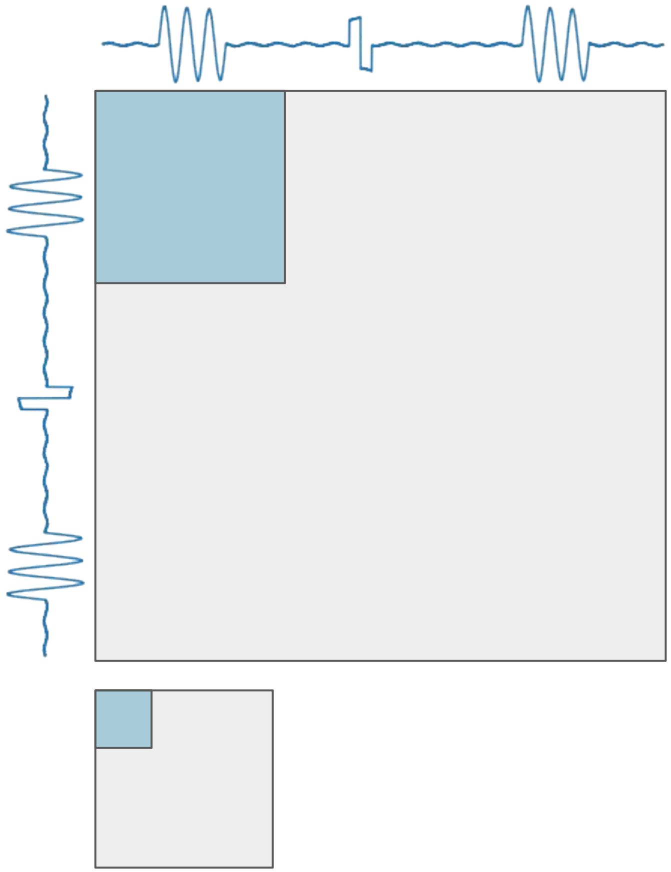
Distance matrix visualizes all distances



Contextual MP | Calculation

Distance matrix visualizes all distances

Find best match in region

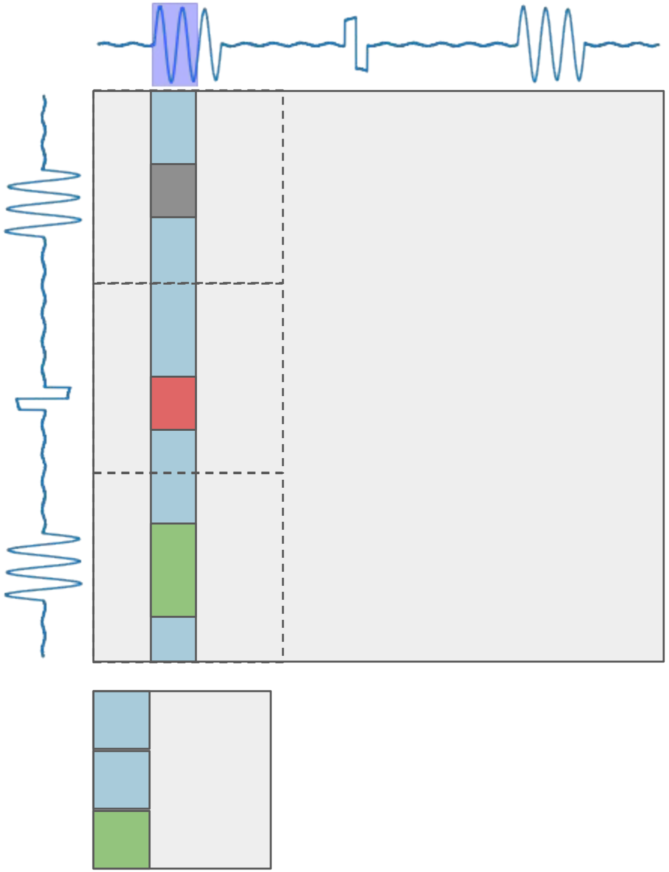


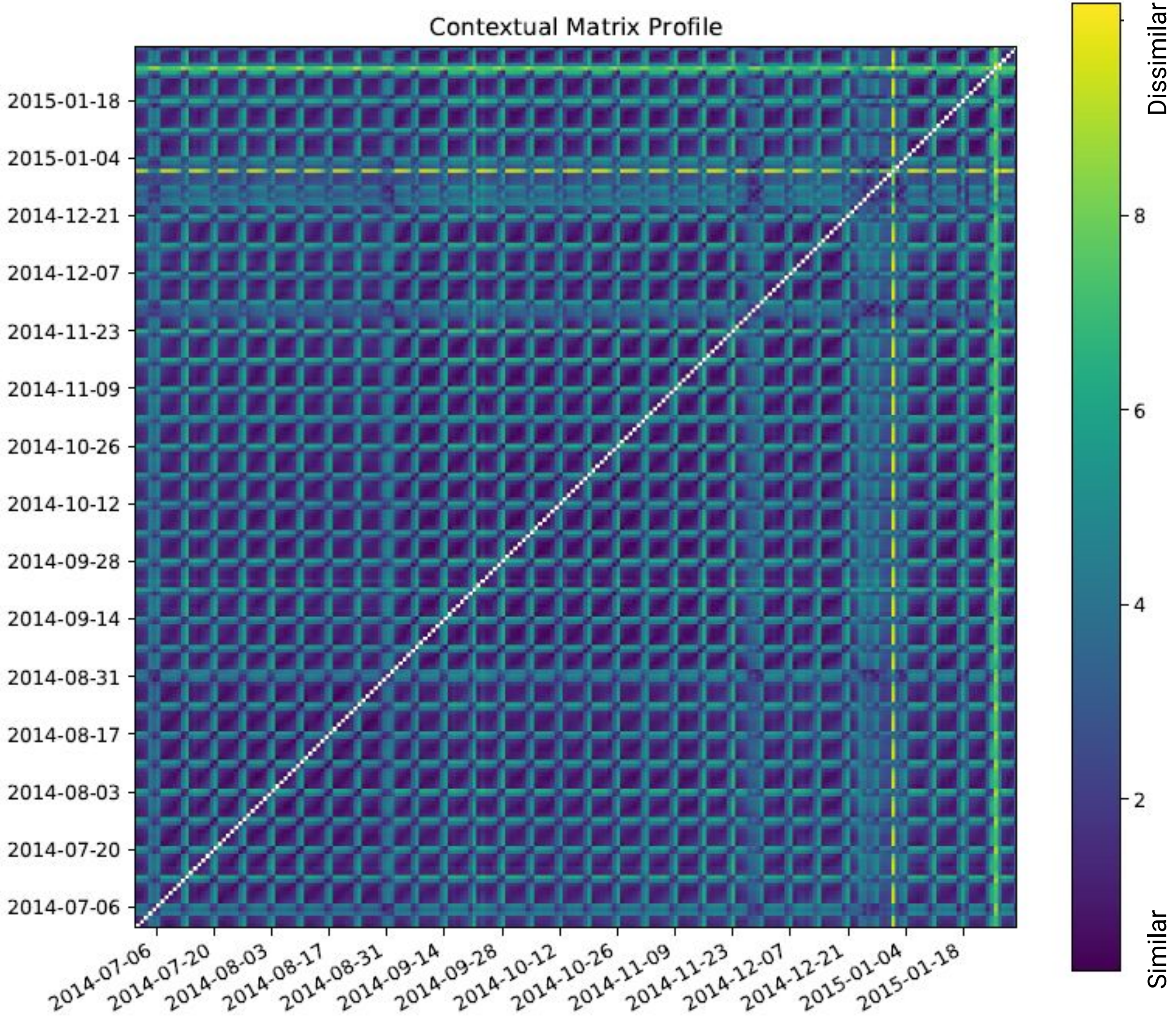
Contextual MP | Calculation

Distance matrix visualizes all distances

Find best match in region

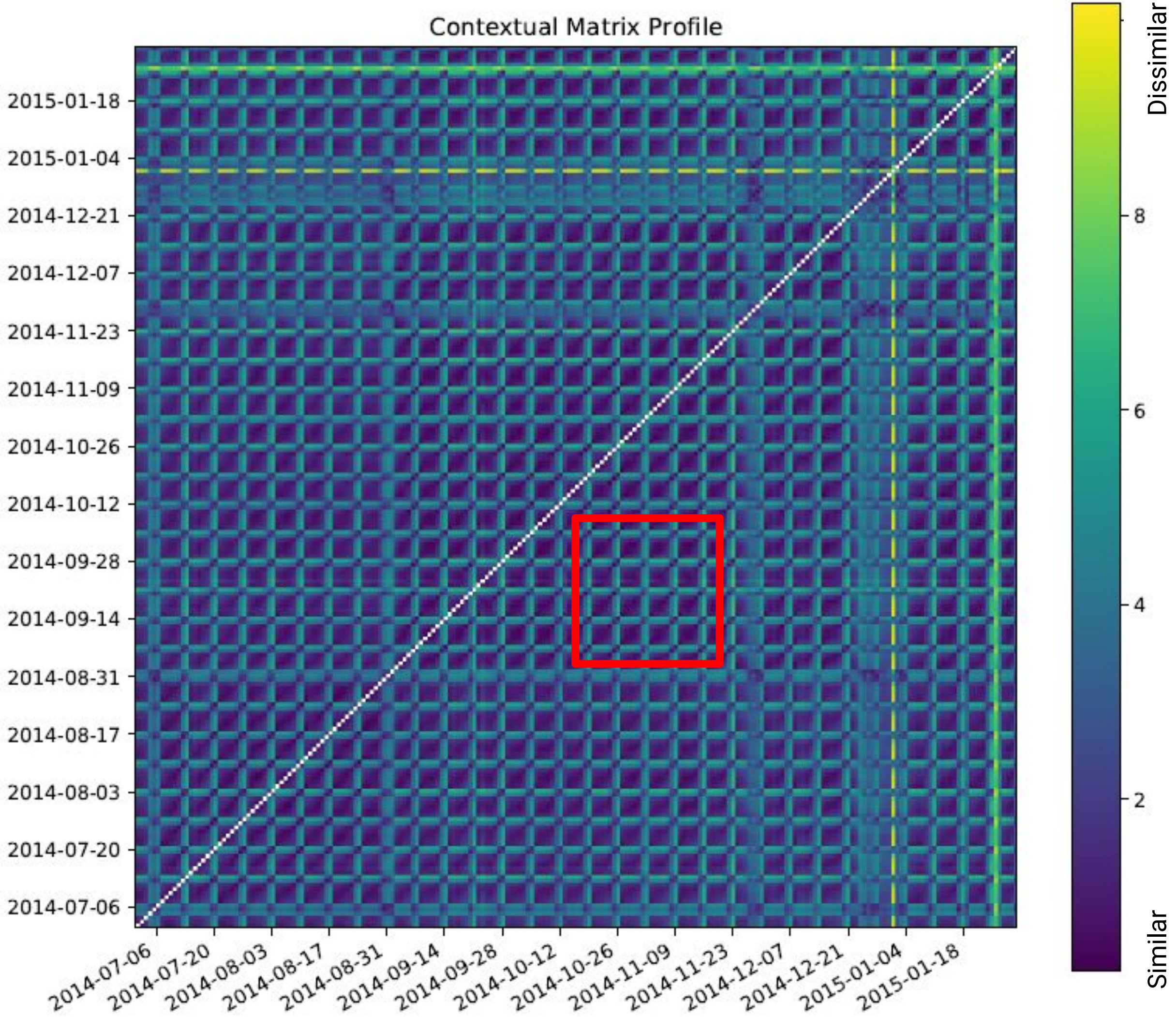
Calculate distances as usual



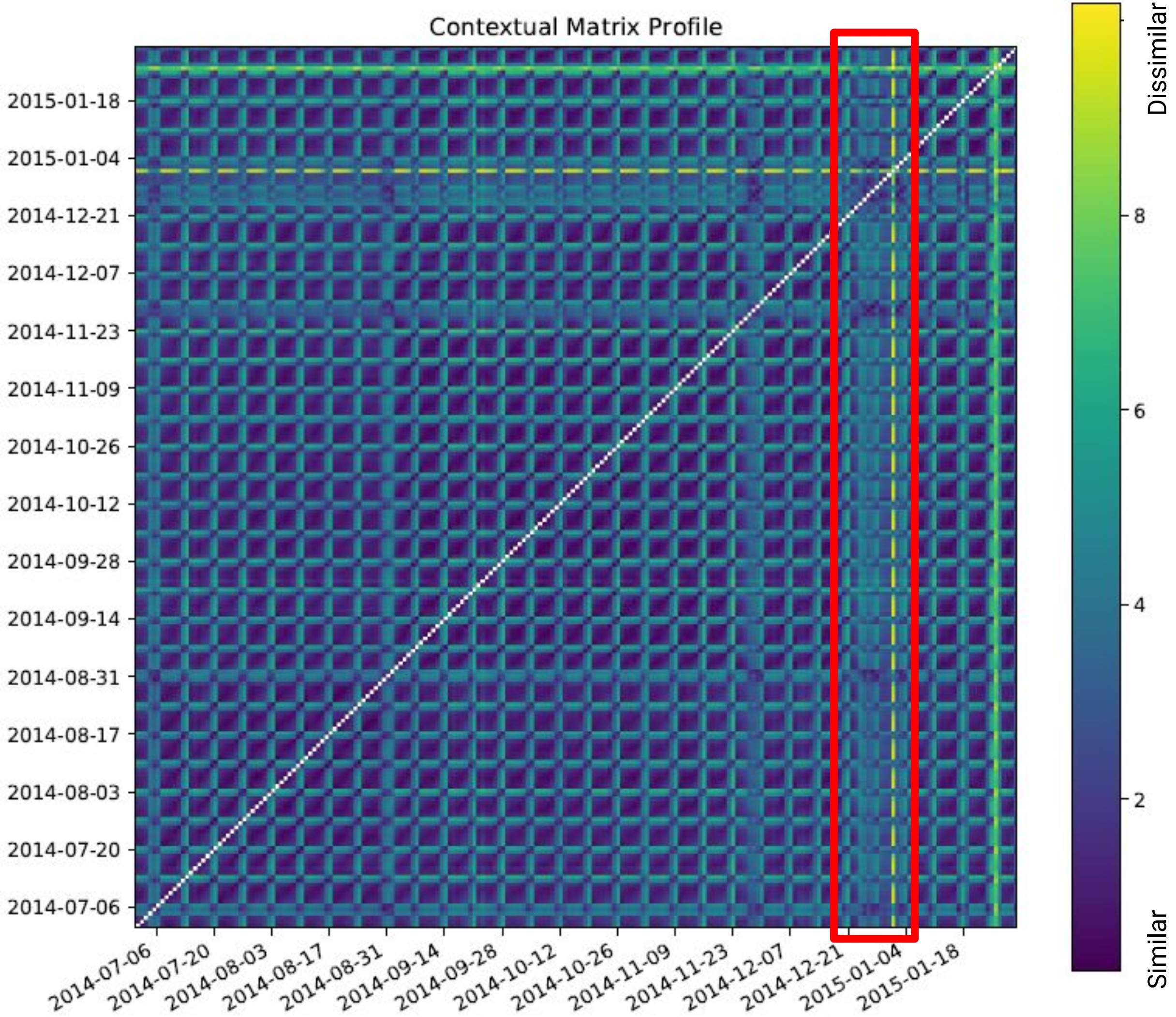


Contextual Matrix Profile
2015-01-18
2015-01-04
2014-12-21
2014-12-07
2014-11-23
2014-11-09
2014-10-26
2014-10-12
2014-09-28
2014-09-14
2014-08-31
2014-08-17
2014-08-03
2014-07-20
2014-07-06
Dissimilar
Similar
8
6
4
2

4K72
7V53

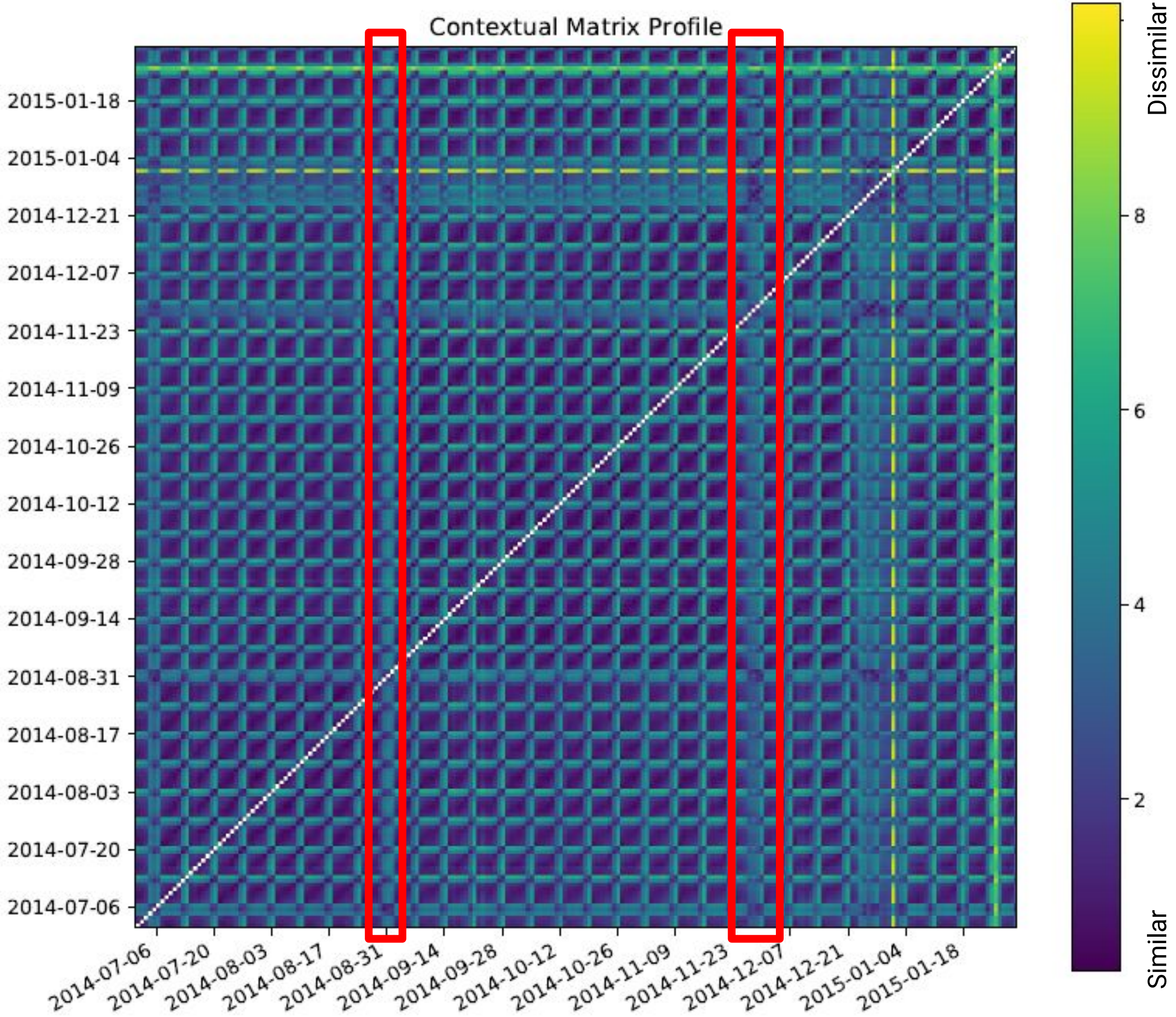


Weekday vs Weekend



Weekday vs Weekend

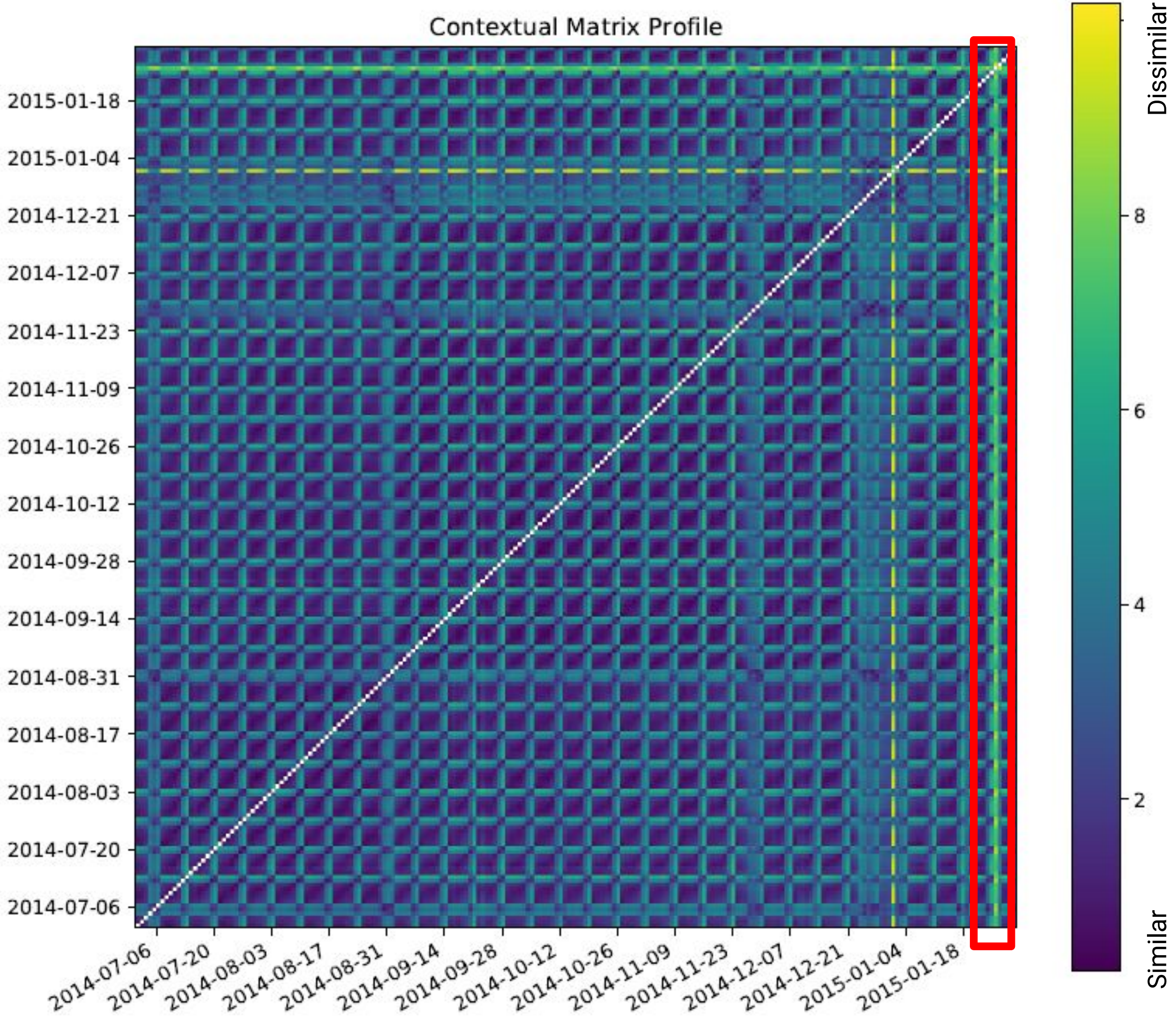
Christmas - New Year



Weekday vs Weekend

Christmas - New Year

Labor Day & Thanksgiving

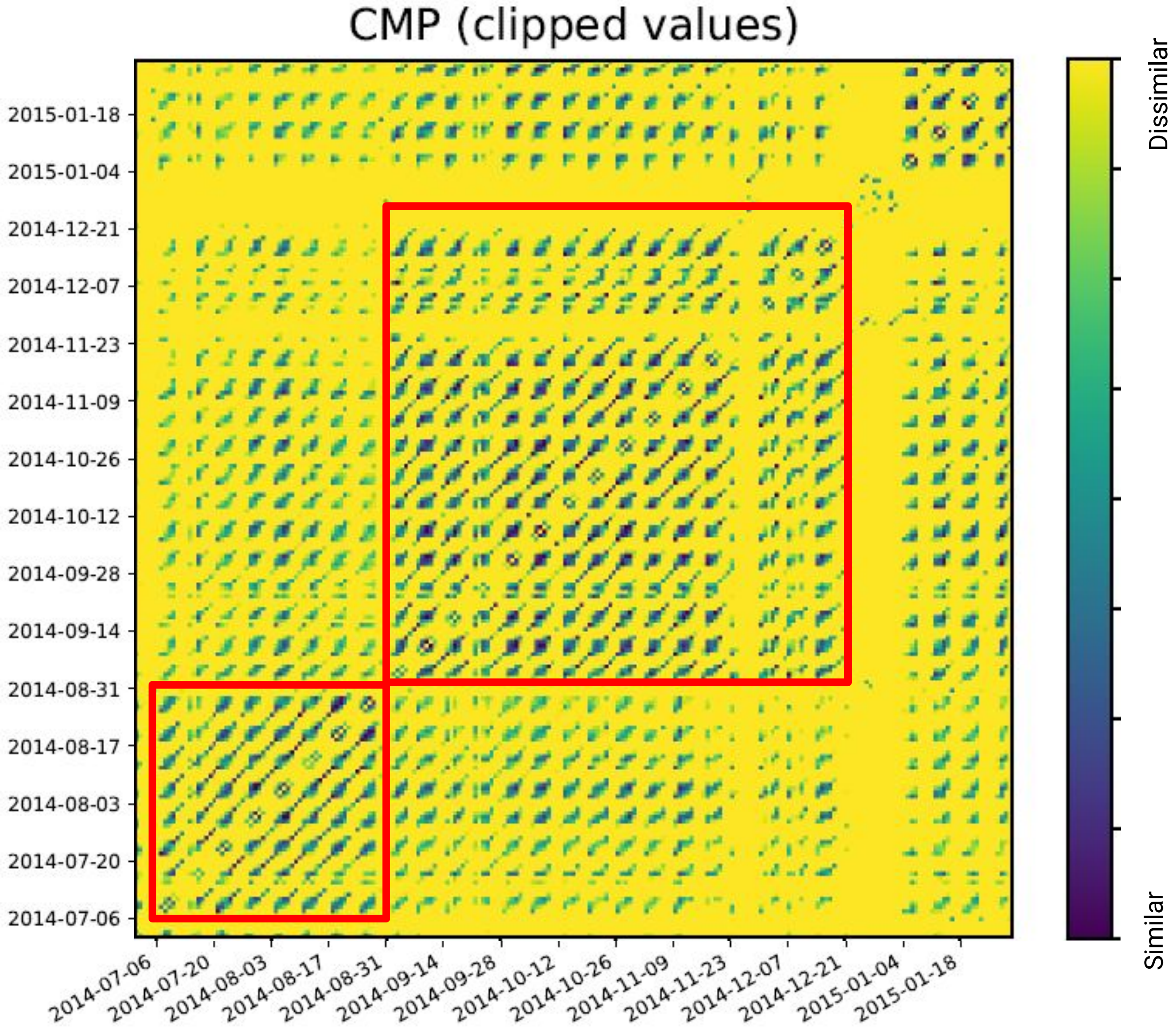


Weekday vs Weekend

Christmas - New Year

Labor Day & Thanksgiving

Blizzard



Weekday vs Weekend

Christmas - New Year

Labor Day & Thanksgiving

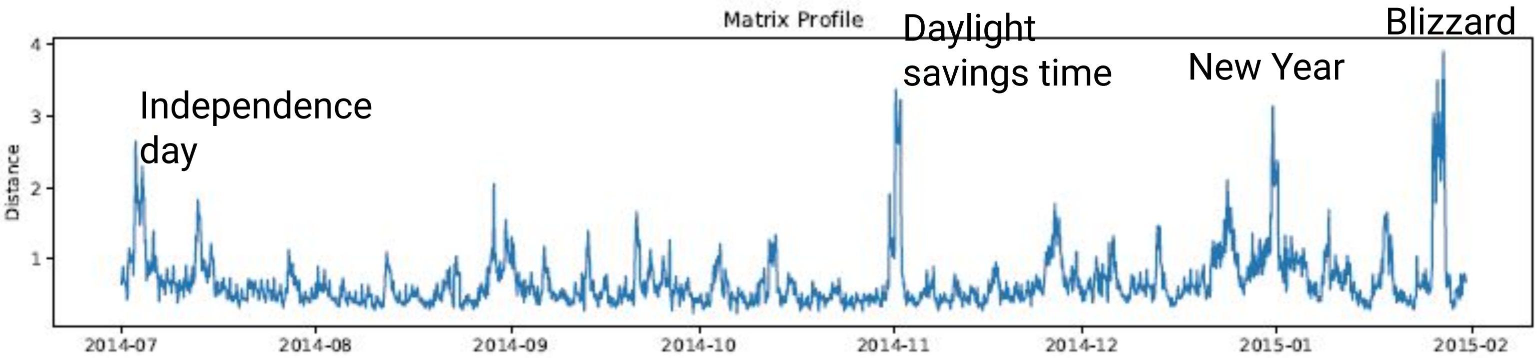
Blizzard

Start of school

Matrix Profile versus Contextual MP

Matrix Profile finds distinct patterns

E.g. blizzard, holidays, transition wintertime



Matrix Profile versus Contextual MP

Matrix Profile finds distinct patterns

E.g. blizzard, holidays, transition wintertime

Contextual MP additionally finds:

periodicity: weekday/weekend, school period

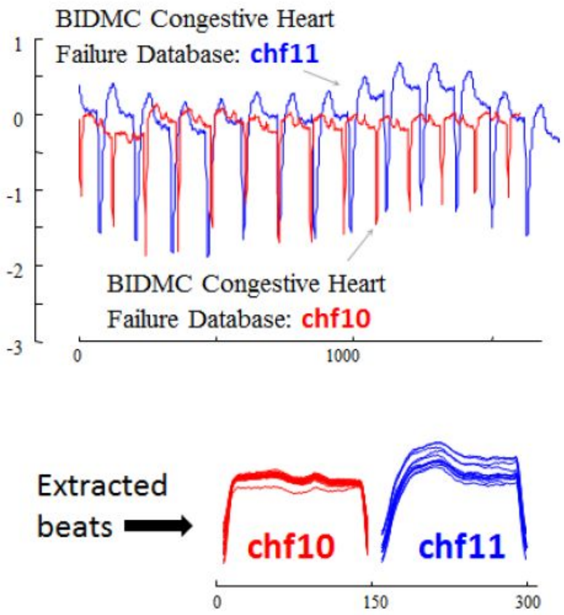
deviating patterns: Christmas period, additional holidays

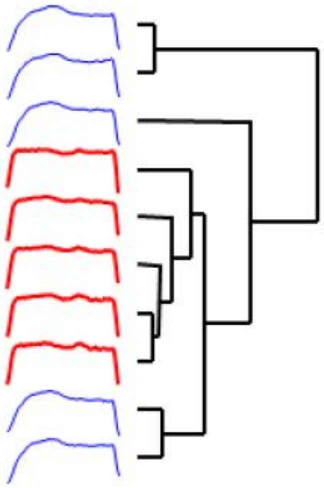
Introduction
Matrix Profile
Contextual Matrix Profile
Noise Elimination
Radius Profile
SDM-Framework
Conclusion

Periodicity
Noise
Repetition
Integration

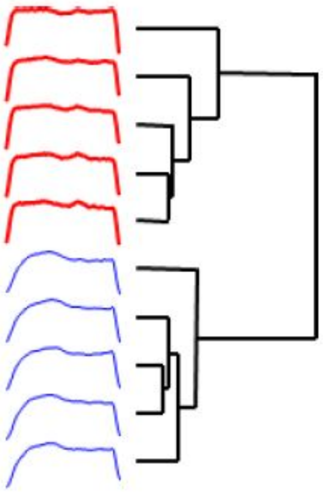
Noise | Z-normalized Euclidean Distance

Most used because it compares shape





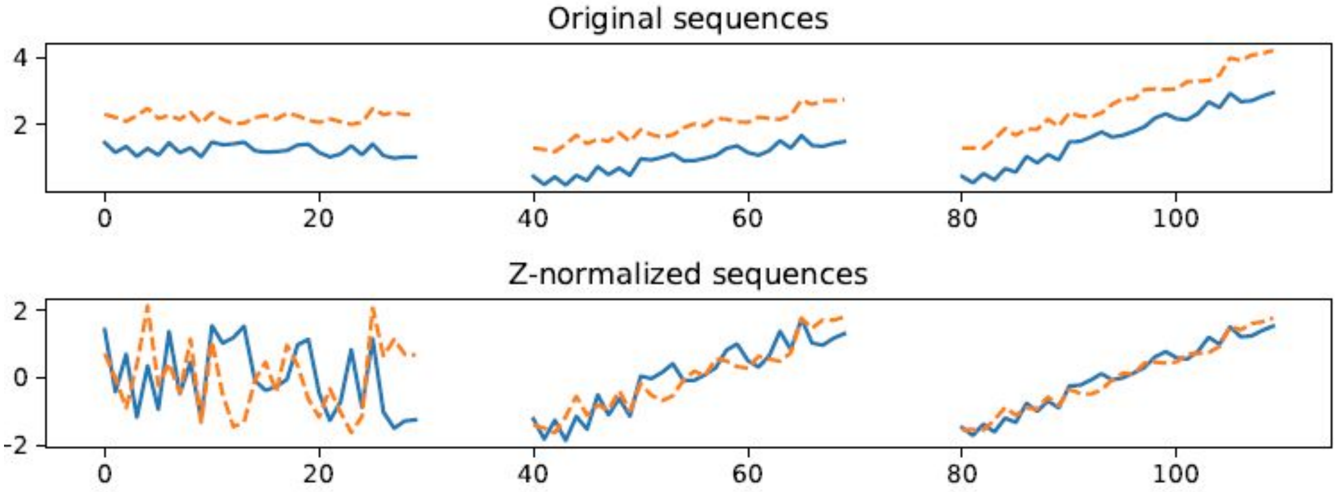
Euclidean Distance



Z-Normalized
Euclidean Distance

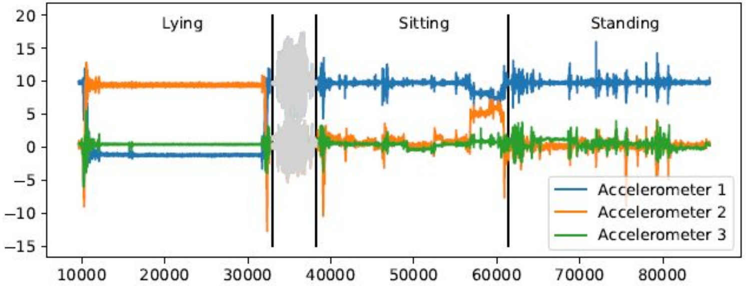
Noise | Z-normalized Euclidean Distance

In rare cases, noise defines shape of the data

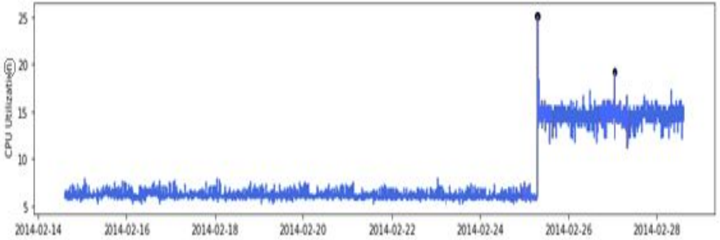


Noise is pretty common

Most sensors experience noise

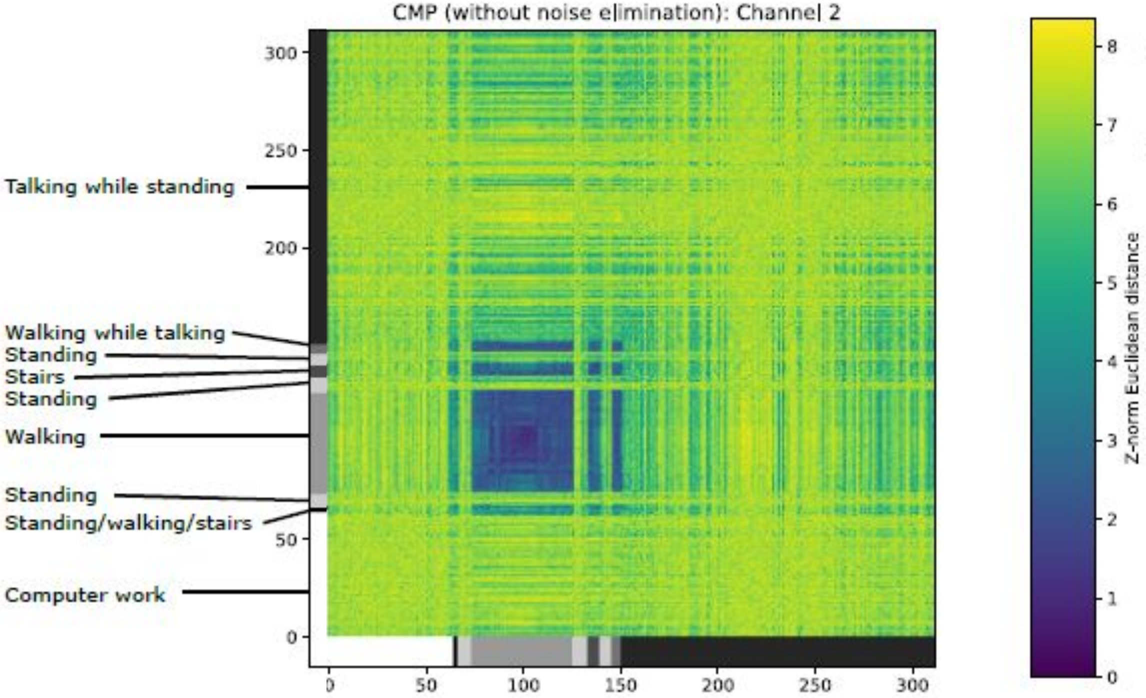


Systems behavior is similar to noise



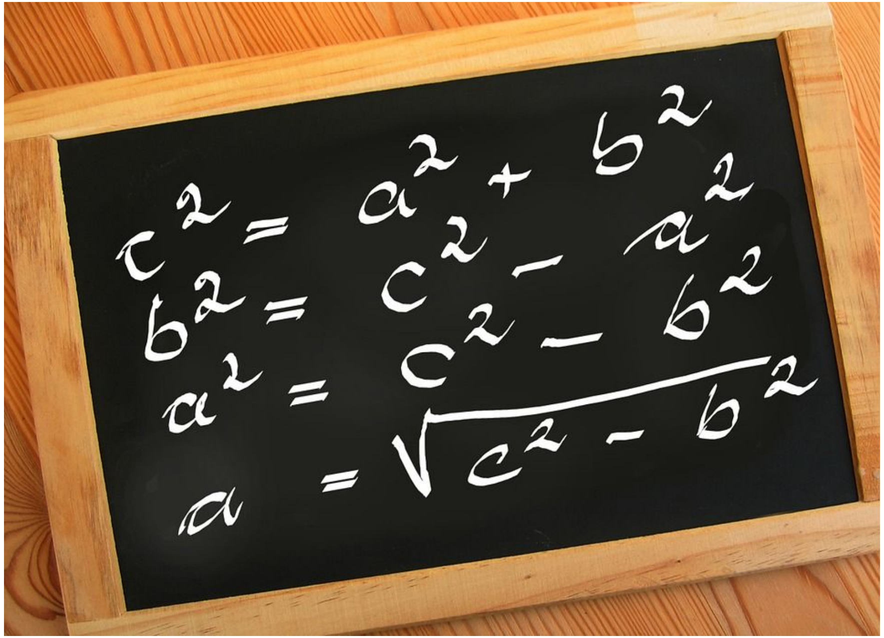
Noise | Examples with noise elimination

Visualization on activity dataset



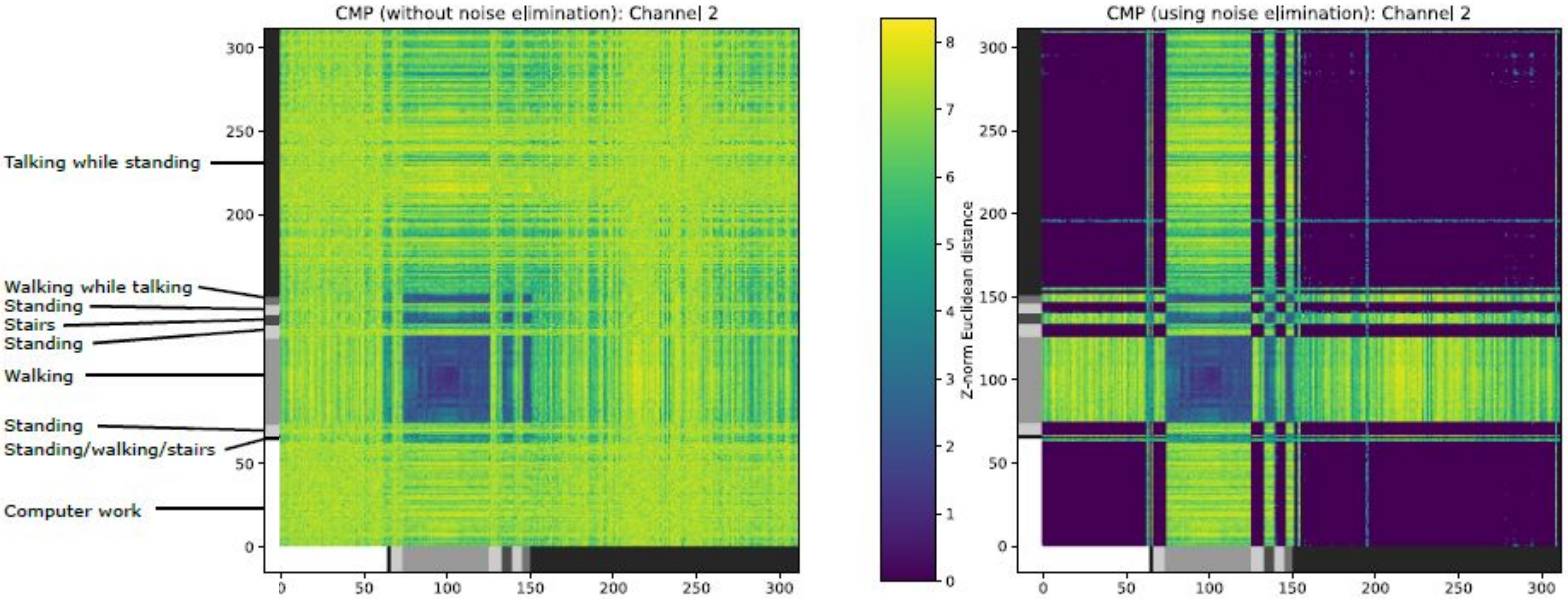
Noise | Approach

Analytically estimate the effect of noise & deduct this estimate



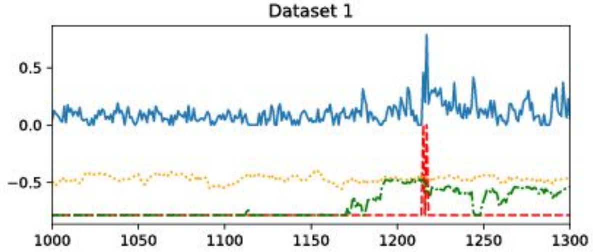
Noise | Examples with noise elimination

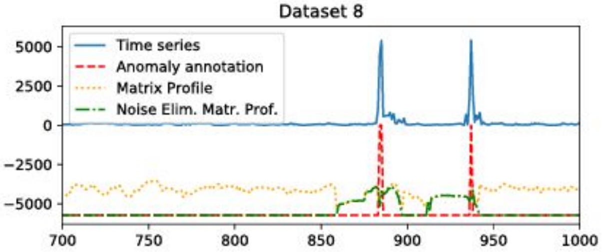
Visualization on activity dataset



Noise | Examples with noise elimination

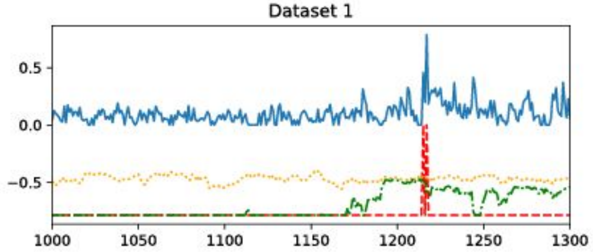
Anomaly detection on system monitoring

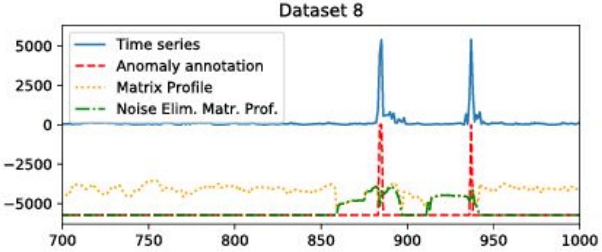


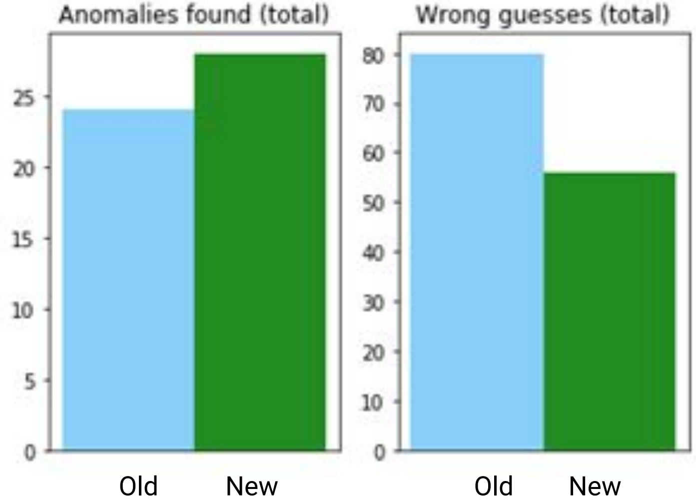


Noise | Examples with noise elimination

Anomaly detection on system monitoring



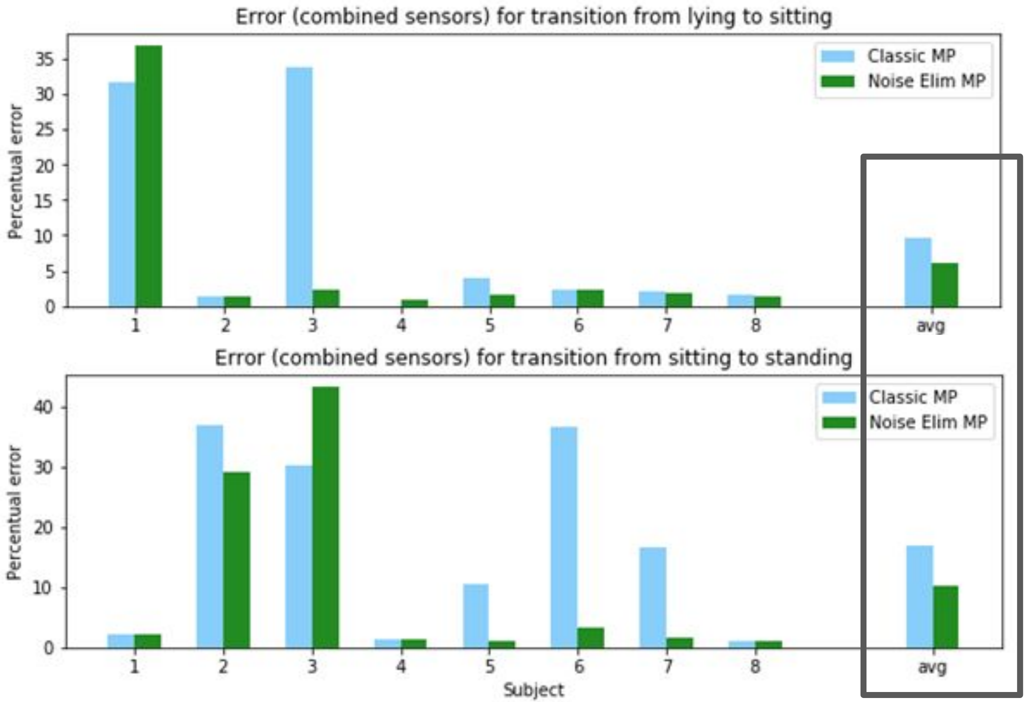




More anomalies found in **less time**

Noise | Examples with noise elimination

Segmentation on activity dataset

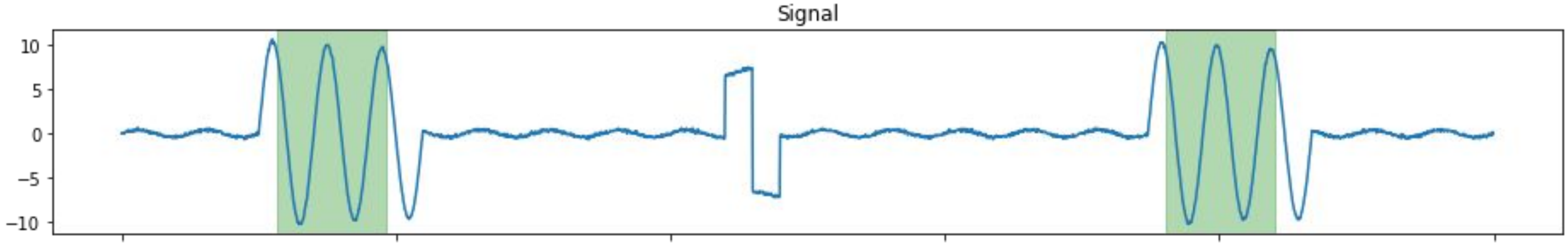


Lower error

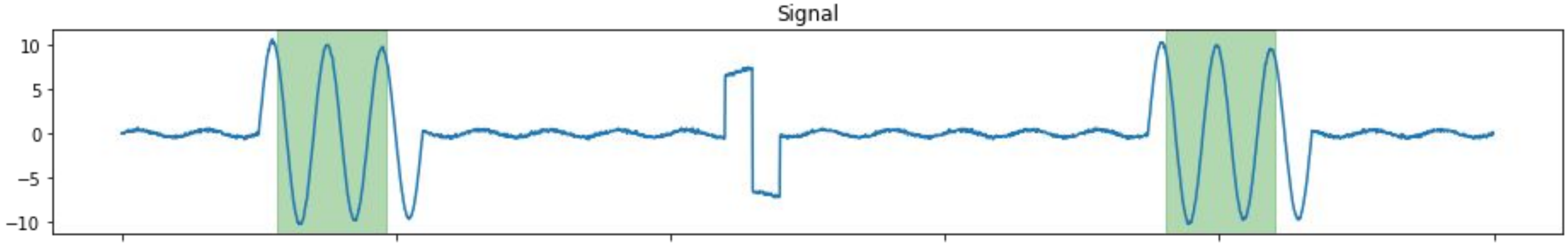
Periodicity
Noise
Repetition
Integration

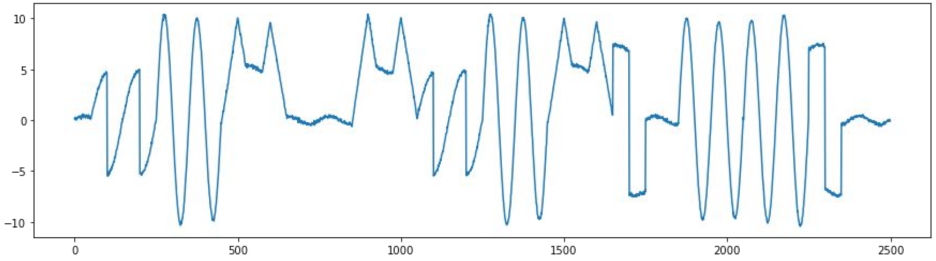
Introduction
Matrix Profile
Contextual Matrix Profile
Noise Elimination
Radius Profile
SDM-Framework
Conclusion

Radius Profile | Use Case

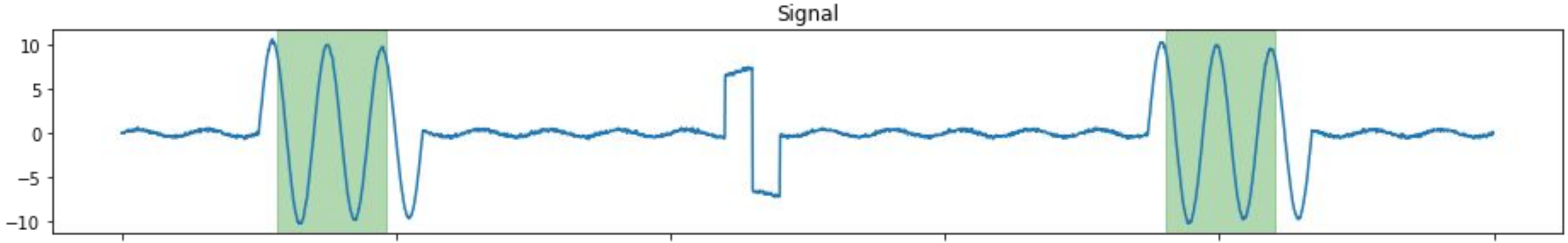


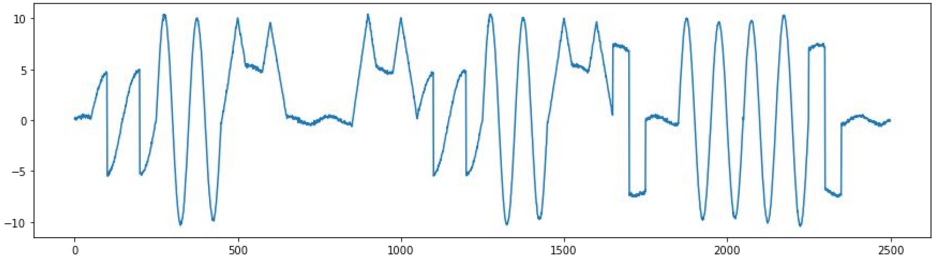
Radius Profile | Use Case

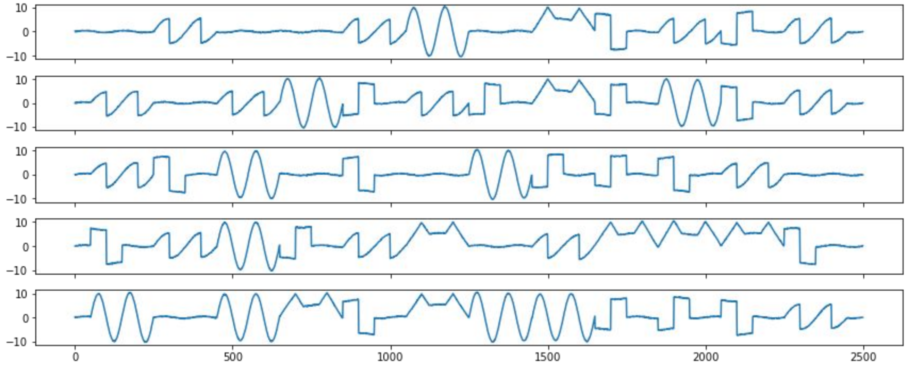




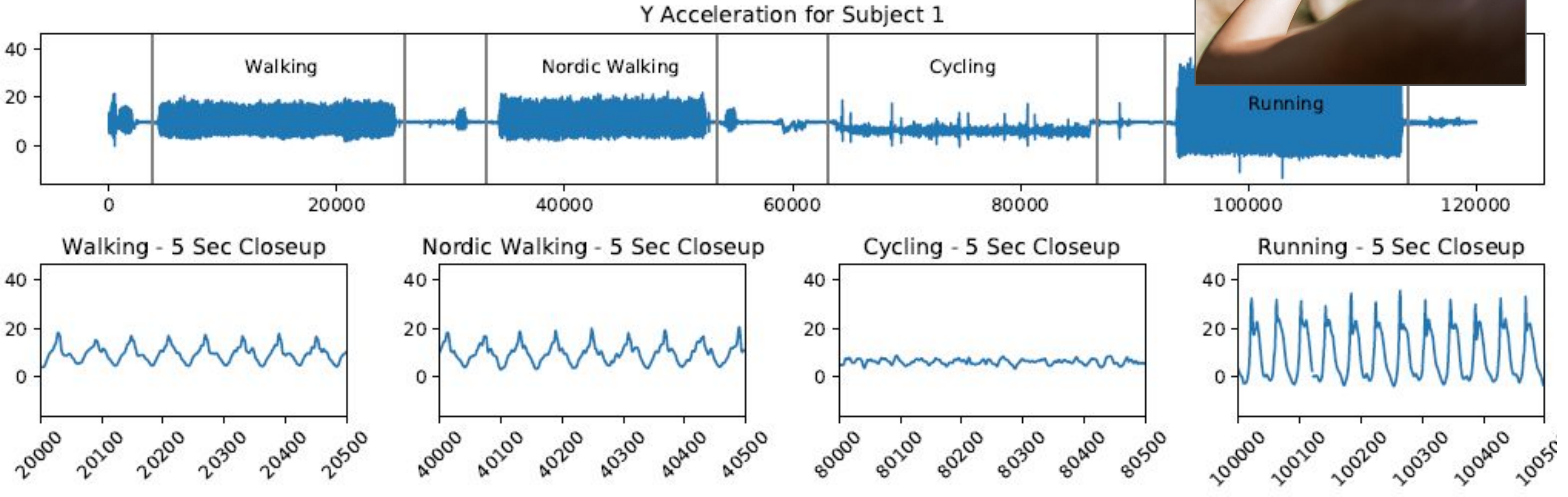
Radius Profile | Use Case





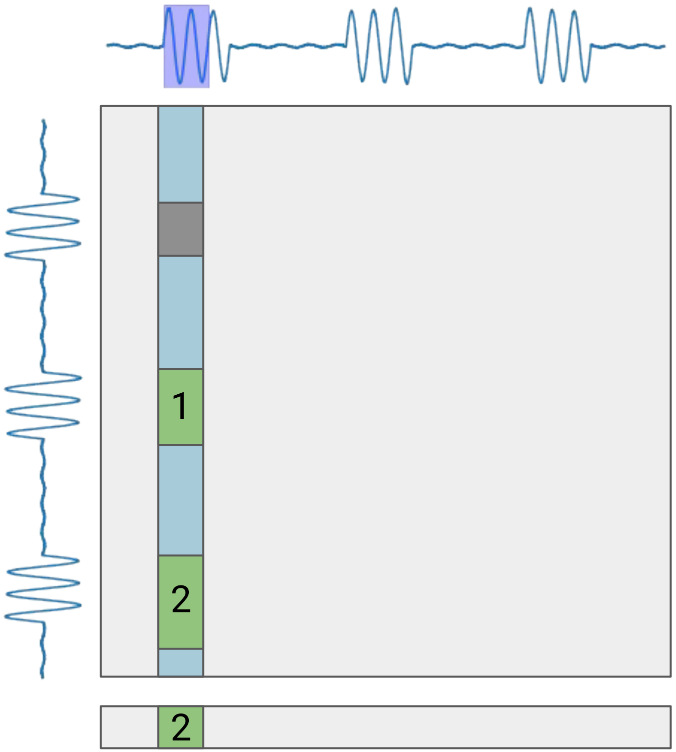


Radius Profile | Example

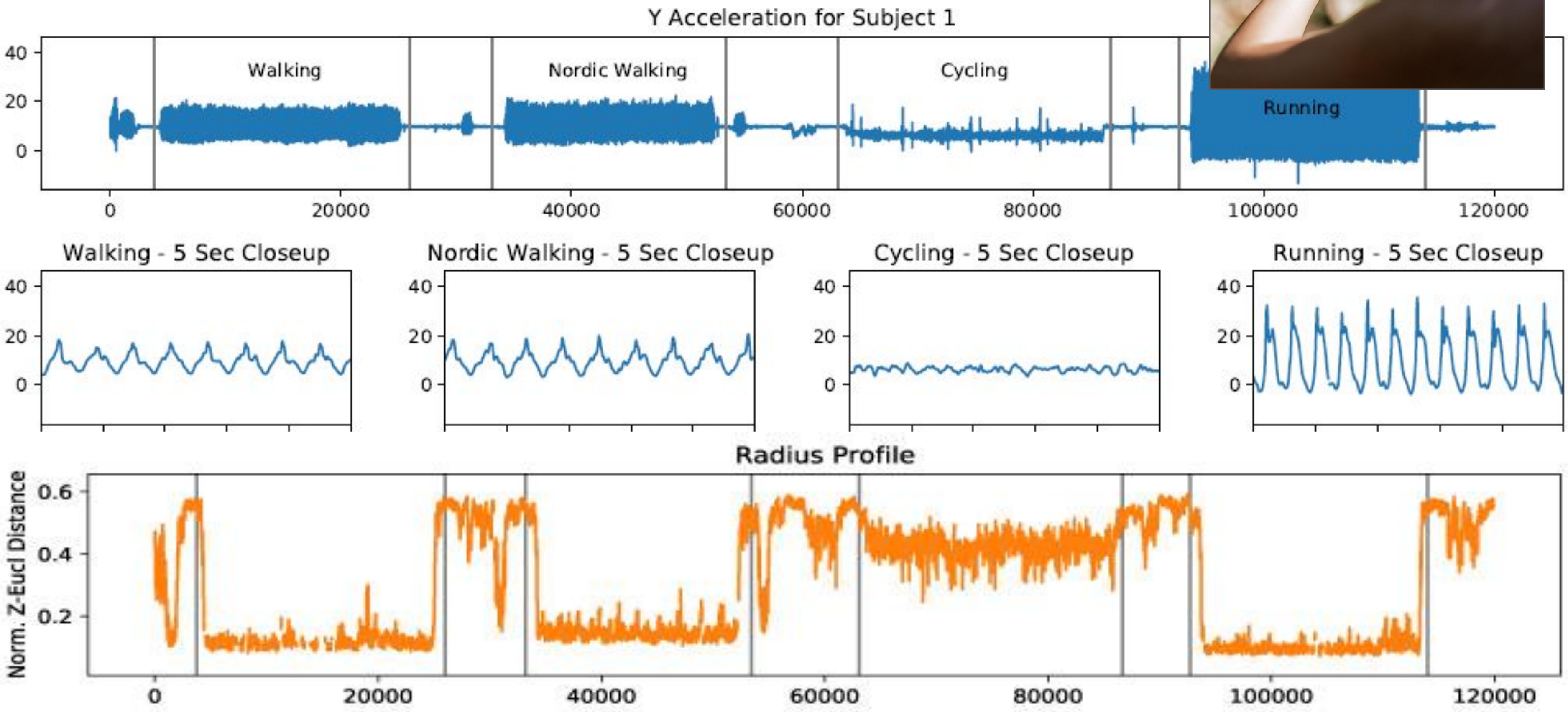


Radius Profile | Calculation

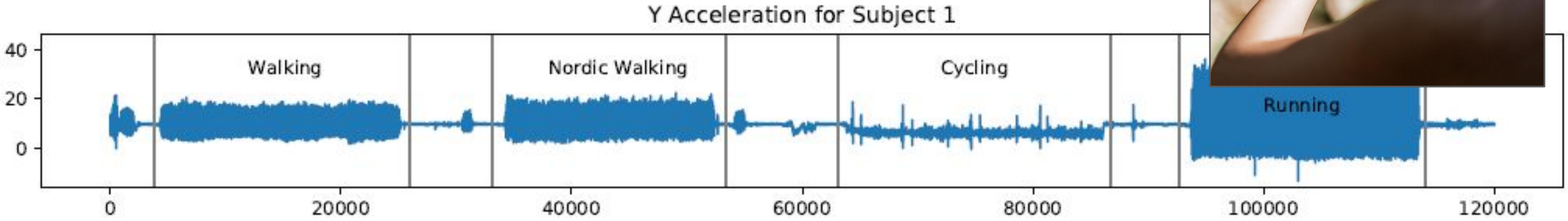
Distance matrix visualizes all distances

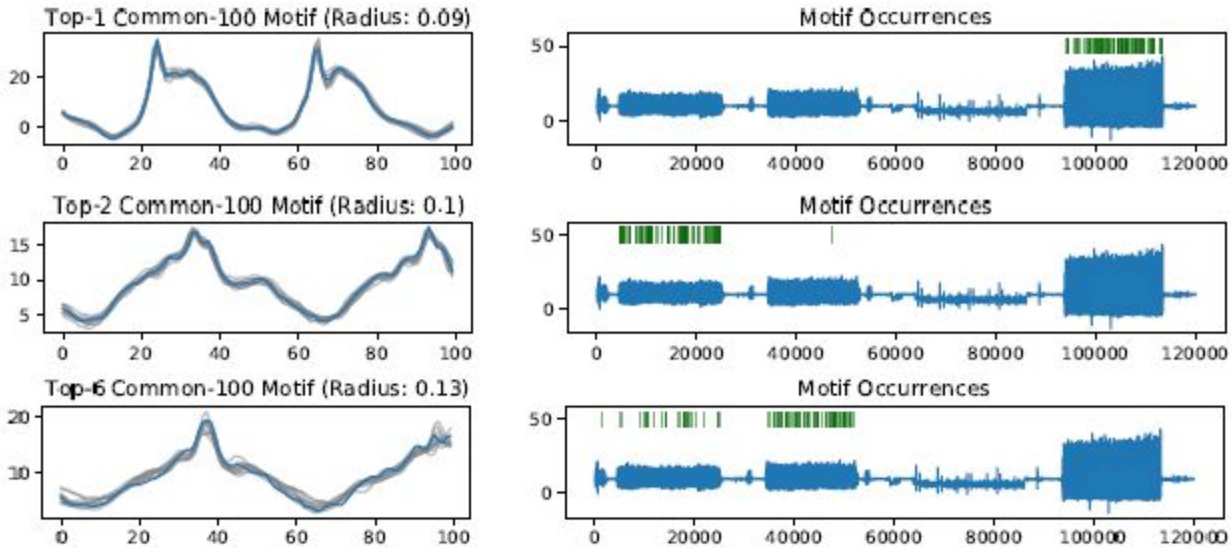


Radius Profile | Example

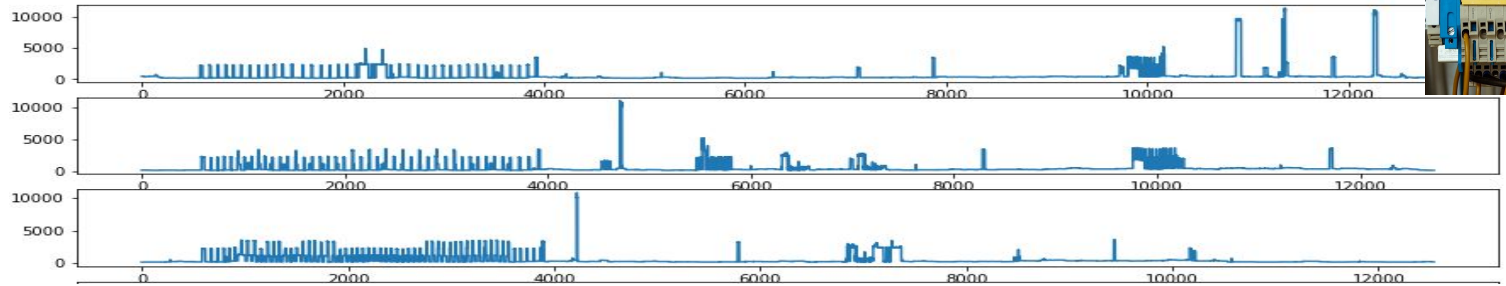


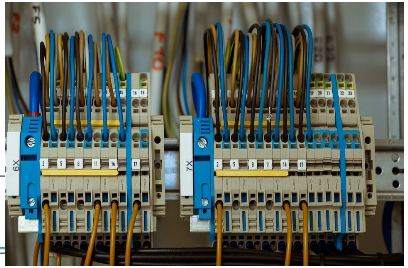
Radius Profile | Example





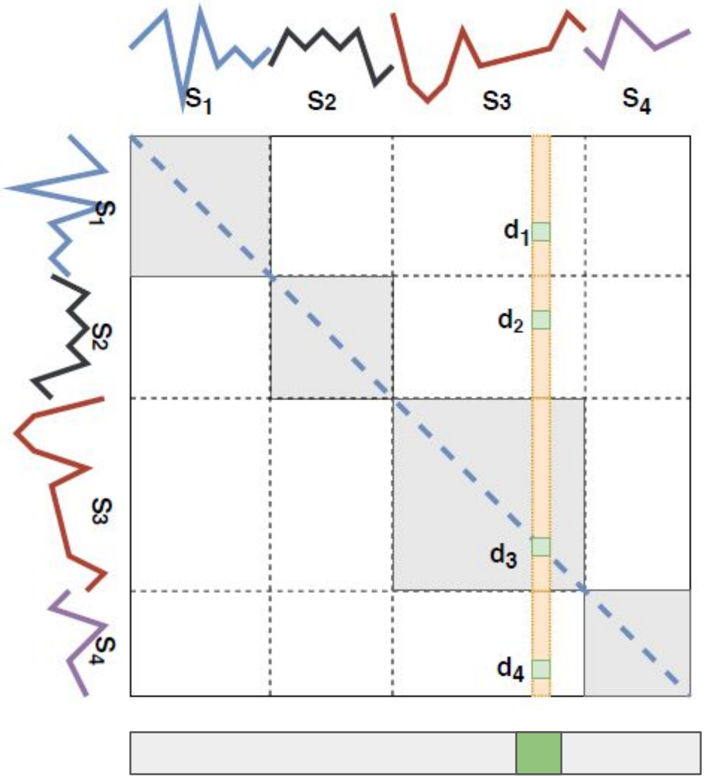
Radius Profile | Example



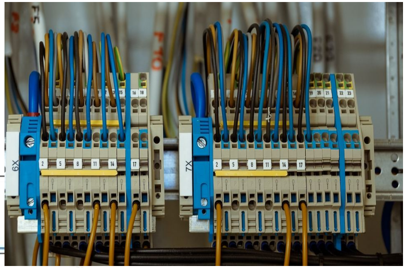


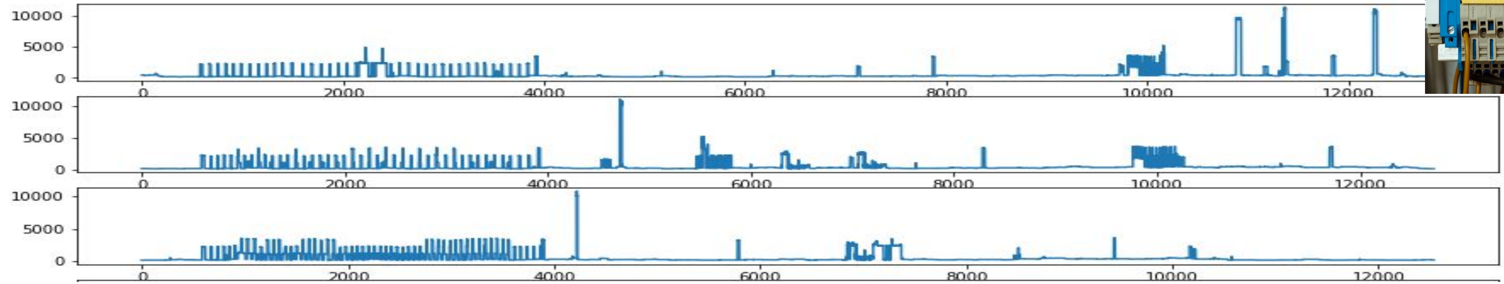
Radius Profile | Calculation

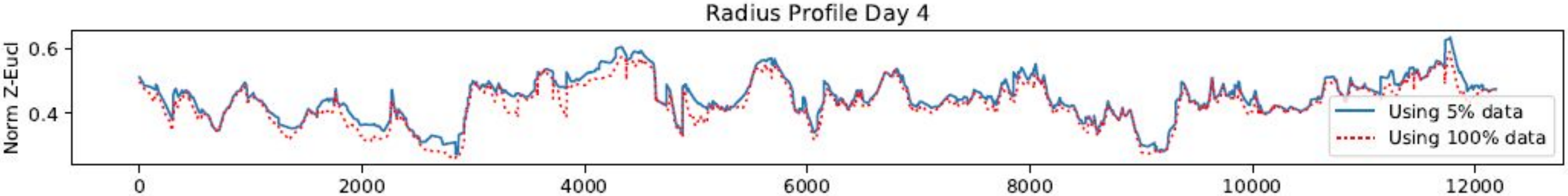
Distance matrix visualizes all distances

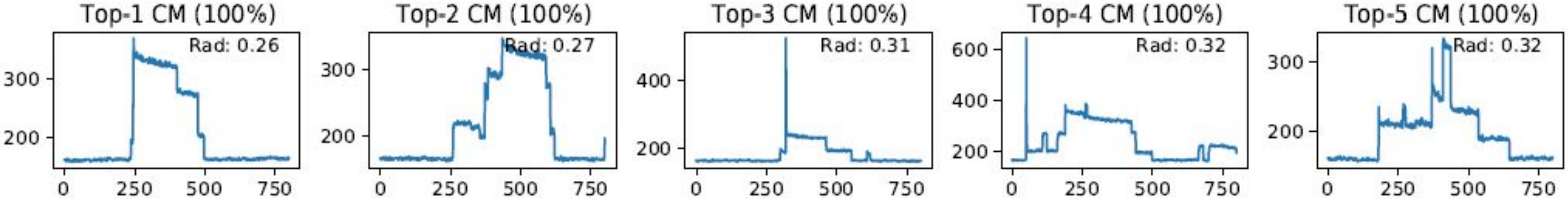


Radius Profile | Example





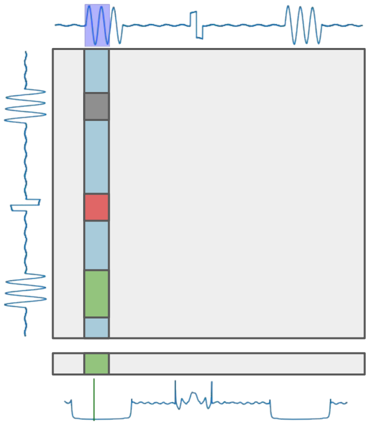




Periodicity
Noise
Repetition
Integration

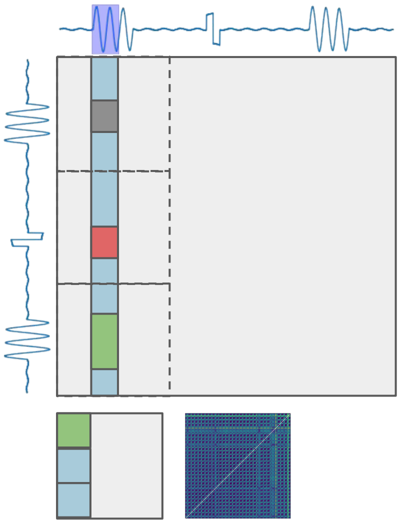
Introduction
Matrix Profile
Contextual Matrix Profile
Noise Elimination
Radius Profile
SDM-Framework
Conclusion

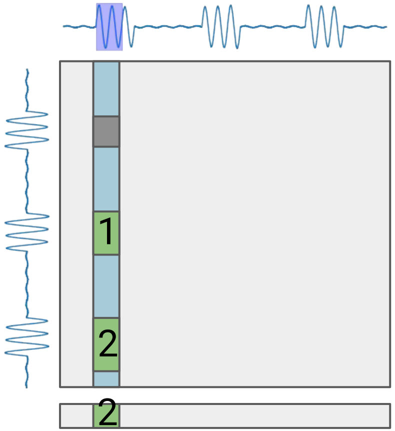
Distance matrix as a foundation



Matrix Profile

Contextual Matrix Profile





Radius Profile

Matrix Profile | Similarities

Given two sequences, define a distance measure

Manhattan distance

$$D_M(X,Y) = \sum_i |x_i - y_i|$$

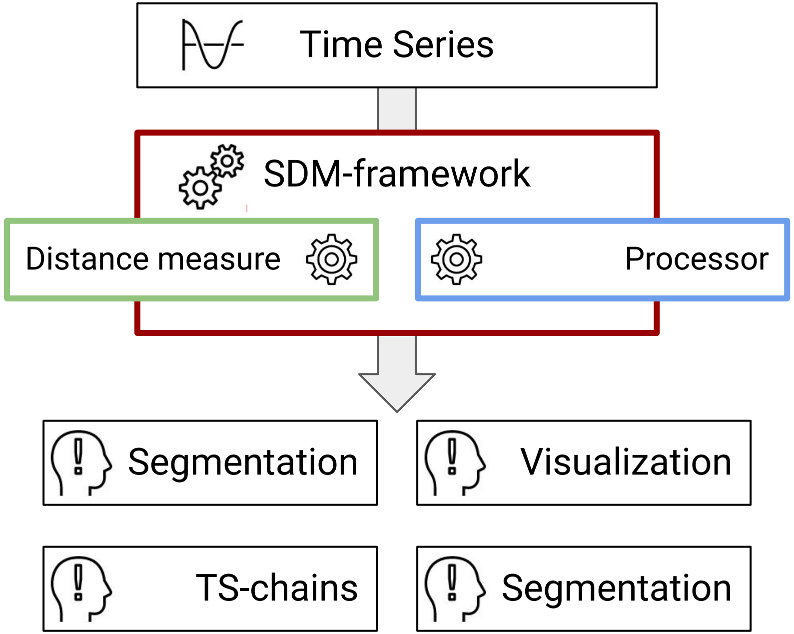
Euclidean distance

$$D_E(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$$

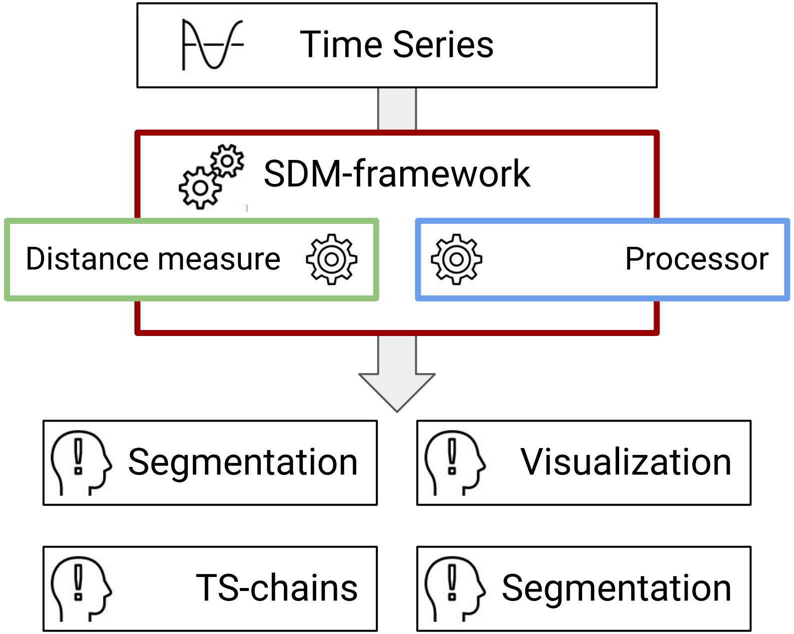
Z-normalized Euclidean distance

$$D_{ZE}(X,Y) = D_E\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right)$$

Series Distance Matrix framework



Series Distance Matrix framework

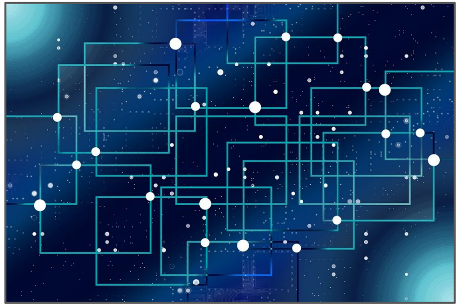


$D_M(X, Y)$

$D_E(X, Y)$

$D_{ZE}(X, Y)$

Series Distance Matrix framework



Time Series

SDM-framework

Distance measure

Processor

$D_M(X,Y)$

$D_E(X,Y)$

$D_{ZE}(X,Y)$

Matrix Profile

Contextual MP

Radius Profile

Segmentation

Visualization

TS-chains

Segmentation

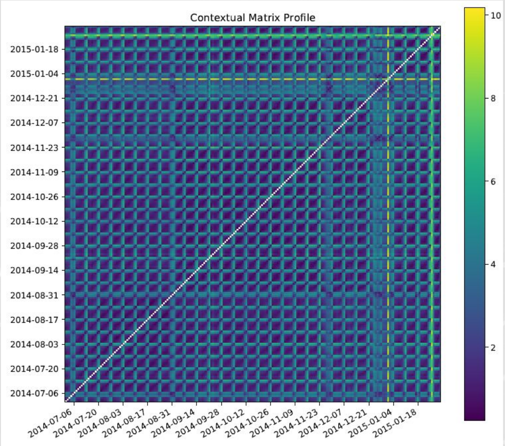
Available online

https://github.com/predict-idlab/seriesdistancematrix

Periodicity
Noise
Repetition
Integration

Introduction
Matrix Profile
Contextual Matrix Profile
Noise Elimination
Radius Profile
SDM-Framework
Conclusion

Periodicity
Noise
Repetition
Integration

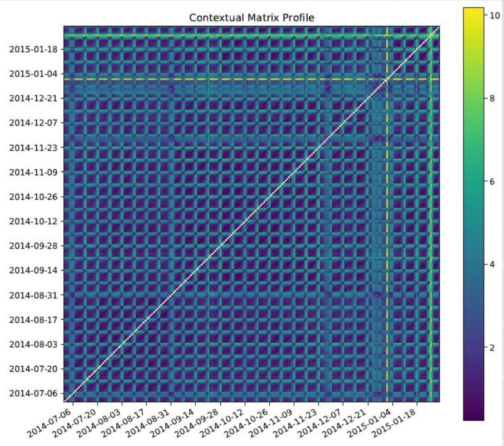


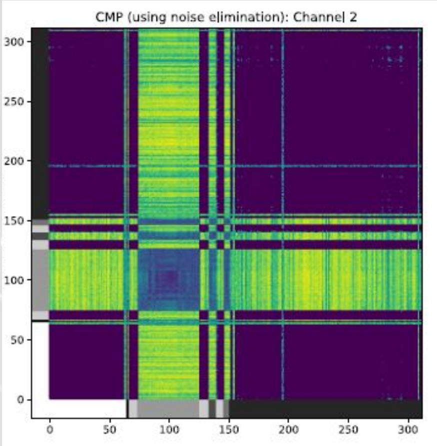
Periodicity

Noise

Repetition

Integration





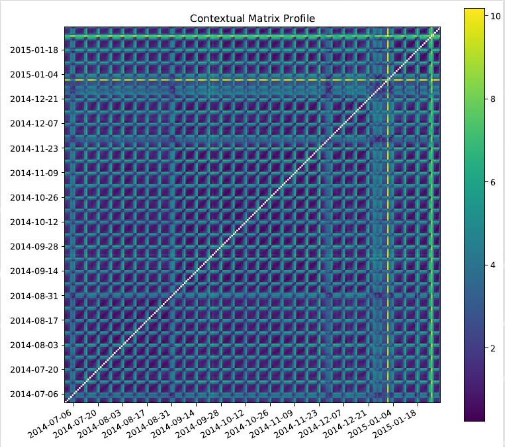
Periodicity

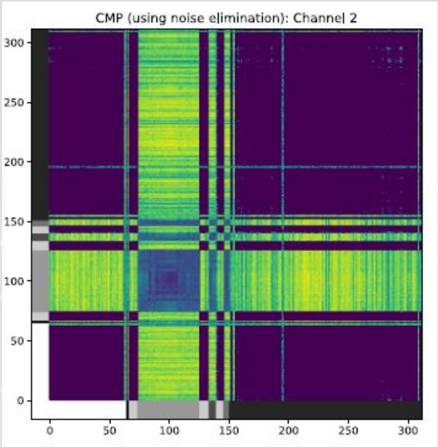
Noise

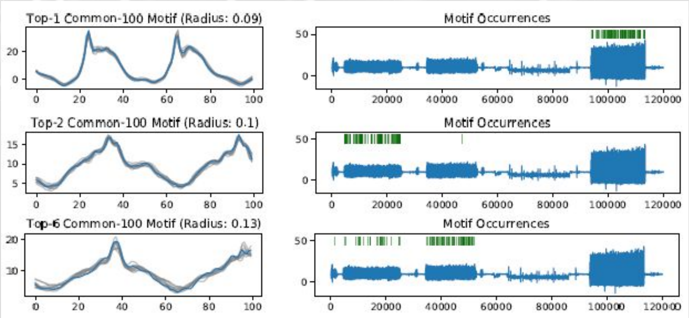
Repetition

Integration

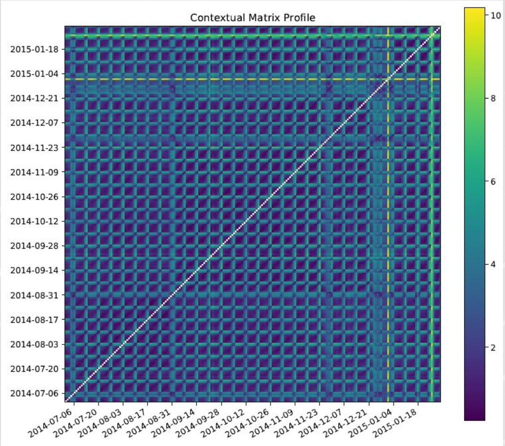
Periodicity
Noise
Repetition
Integration

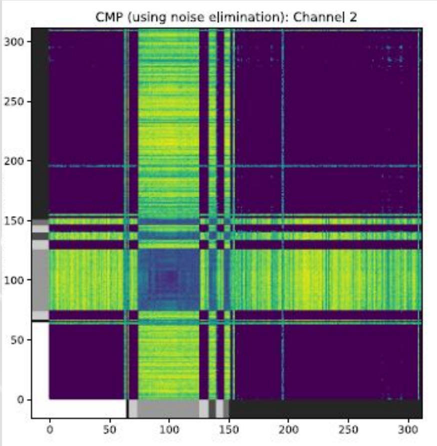


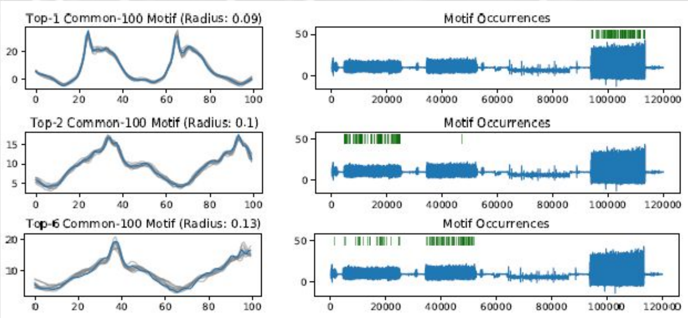


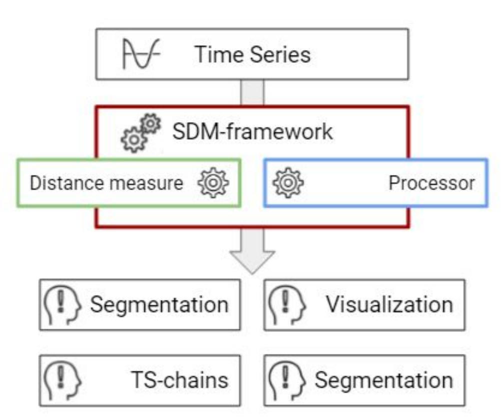


Periodicity
Noise
Repetition
Integration

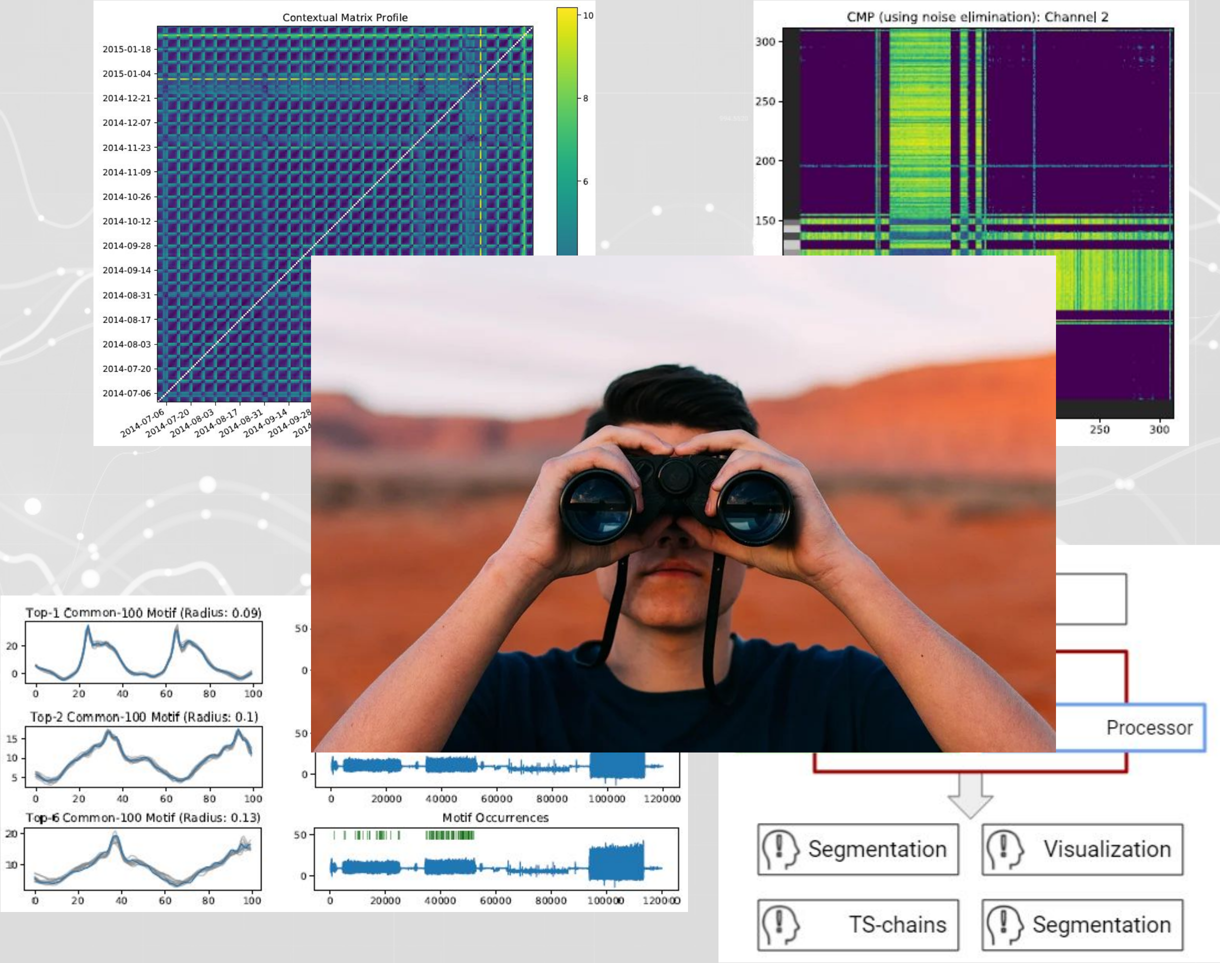








Periodicity
Noise
Repetition
Integration



Insight mining in time series data with applications for anomaly detection

Dieter De Paepe

GHENT UNIVERSITY

Questions