Feature Engineering - 2

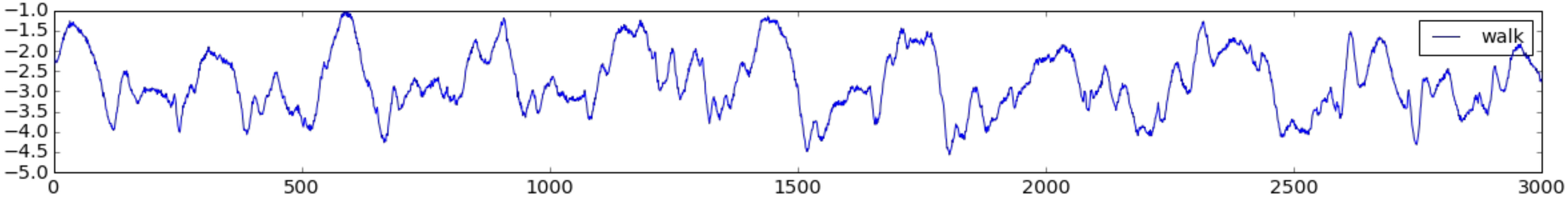
Feature Optimization

- Normalization
- Interaction between Features
- Binning
- Adaptive Binning
- Thresholding
- Scaling
- Log Transformation

Histograms

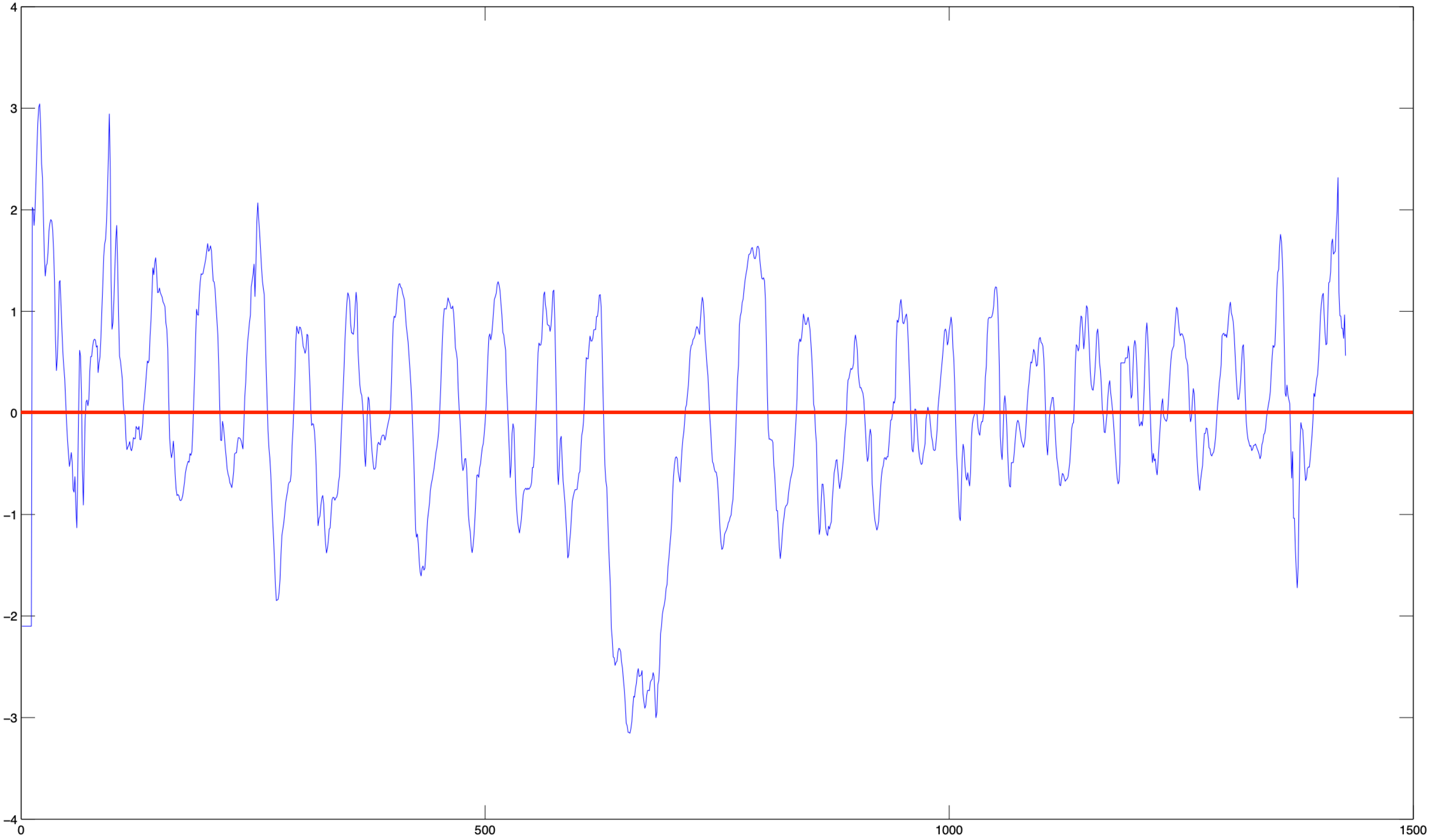
vs.

Time-Series Data



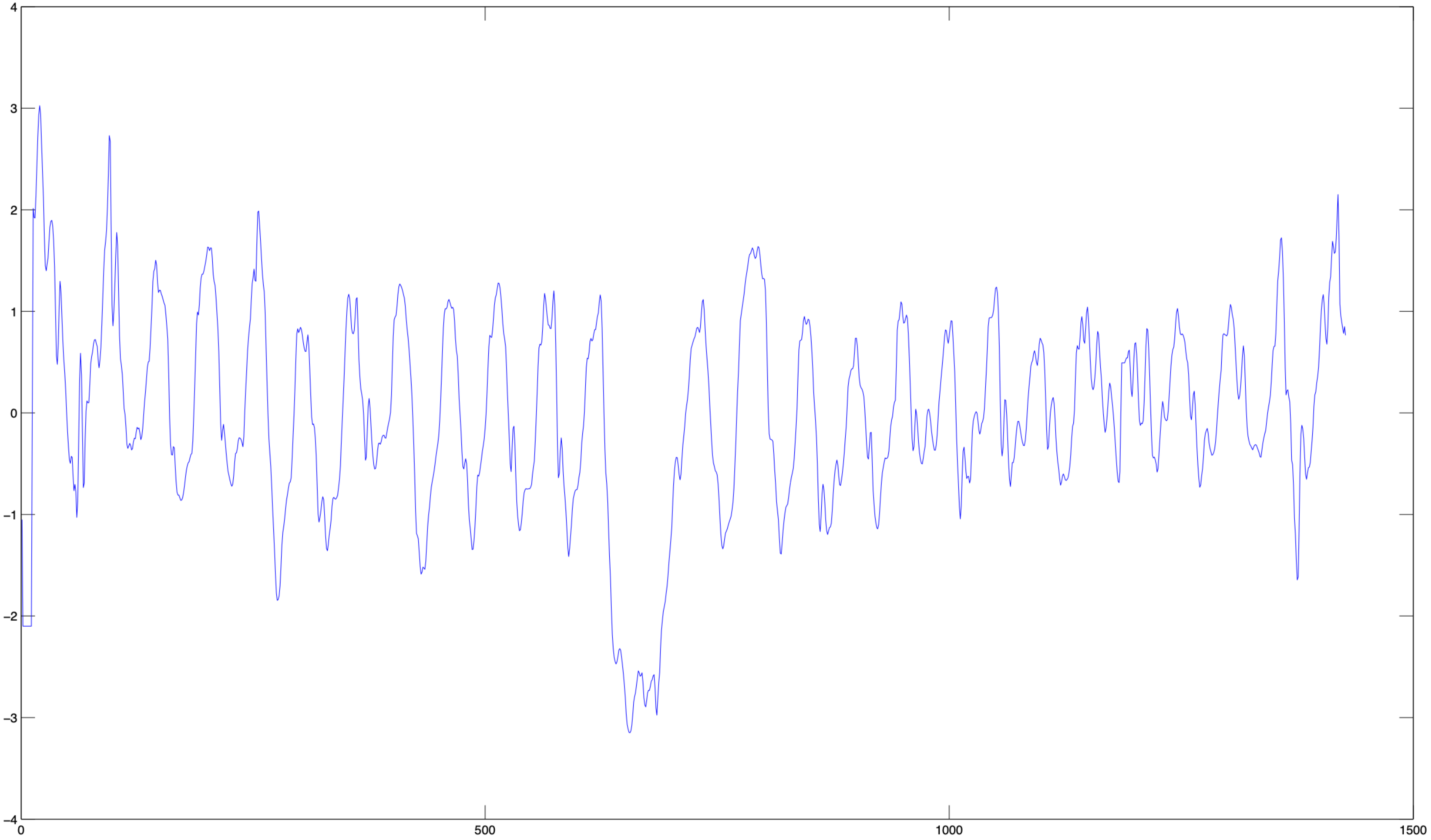
- Cleaning or preprocessing the data
- Correct representation
- Level/Magnitude
- Repetitions
- Shape of the curve

Preprocessing



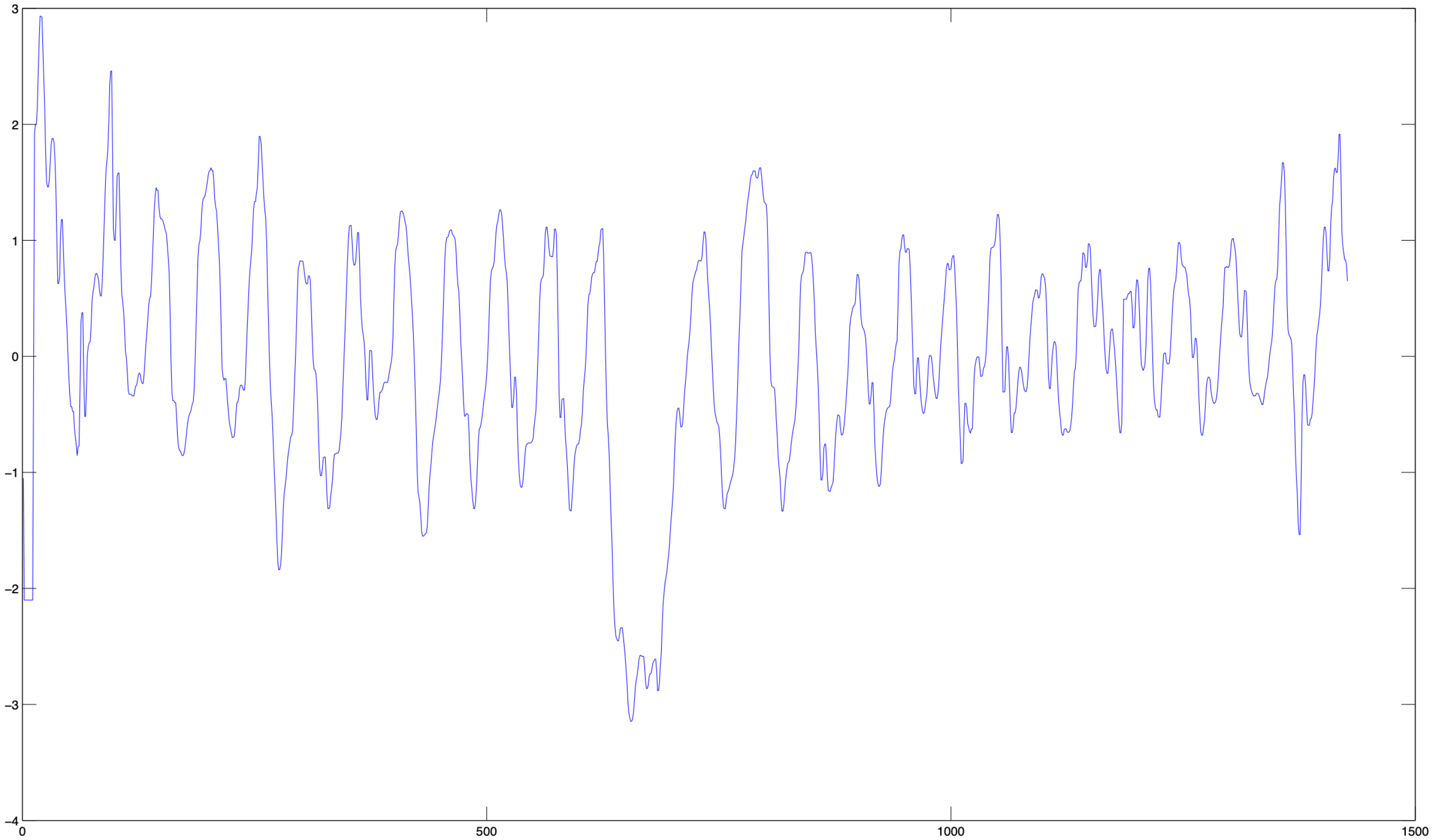
Zero Crossing on clean data

Preprocessing



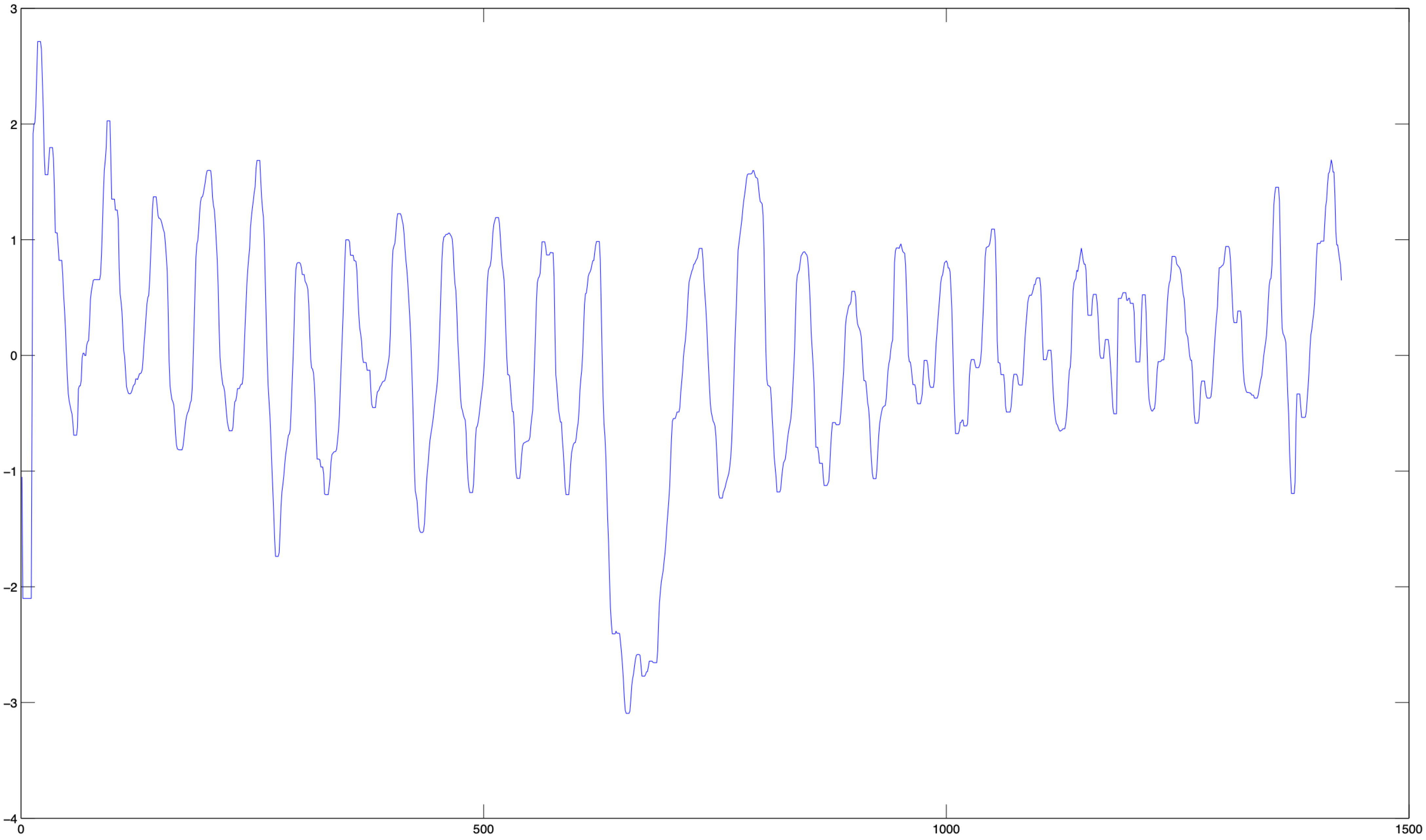
Median Filtering, size = 2

Preprocessing



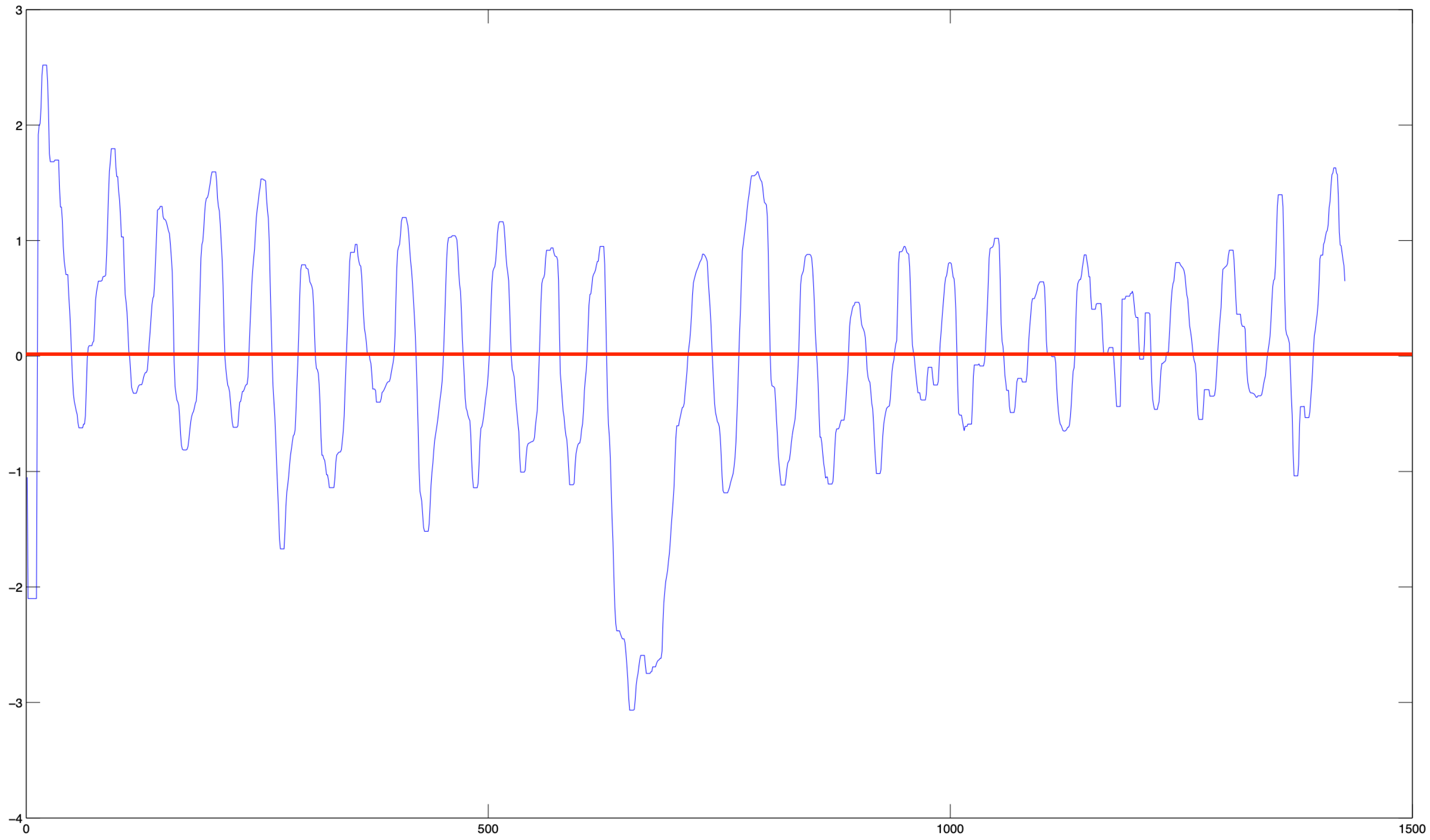
Median Filtering, size = 4

Preprocessing



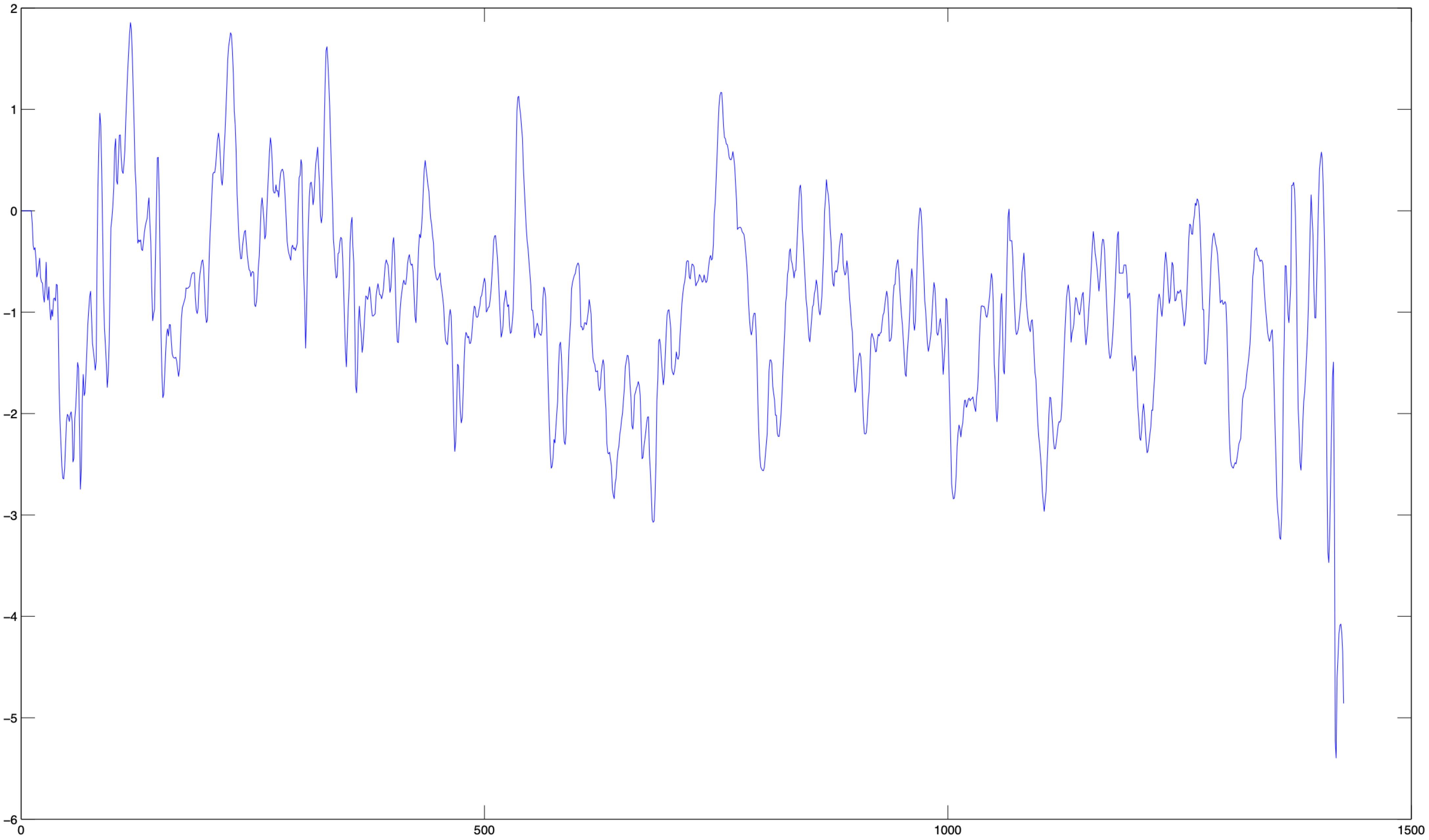
Median Filtering, size = 6

Preprocessing



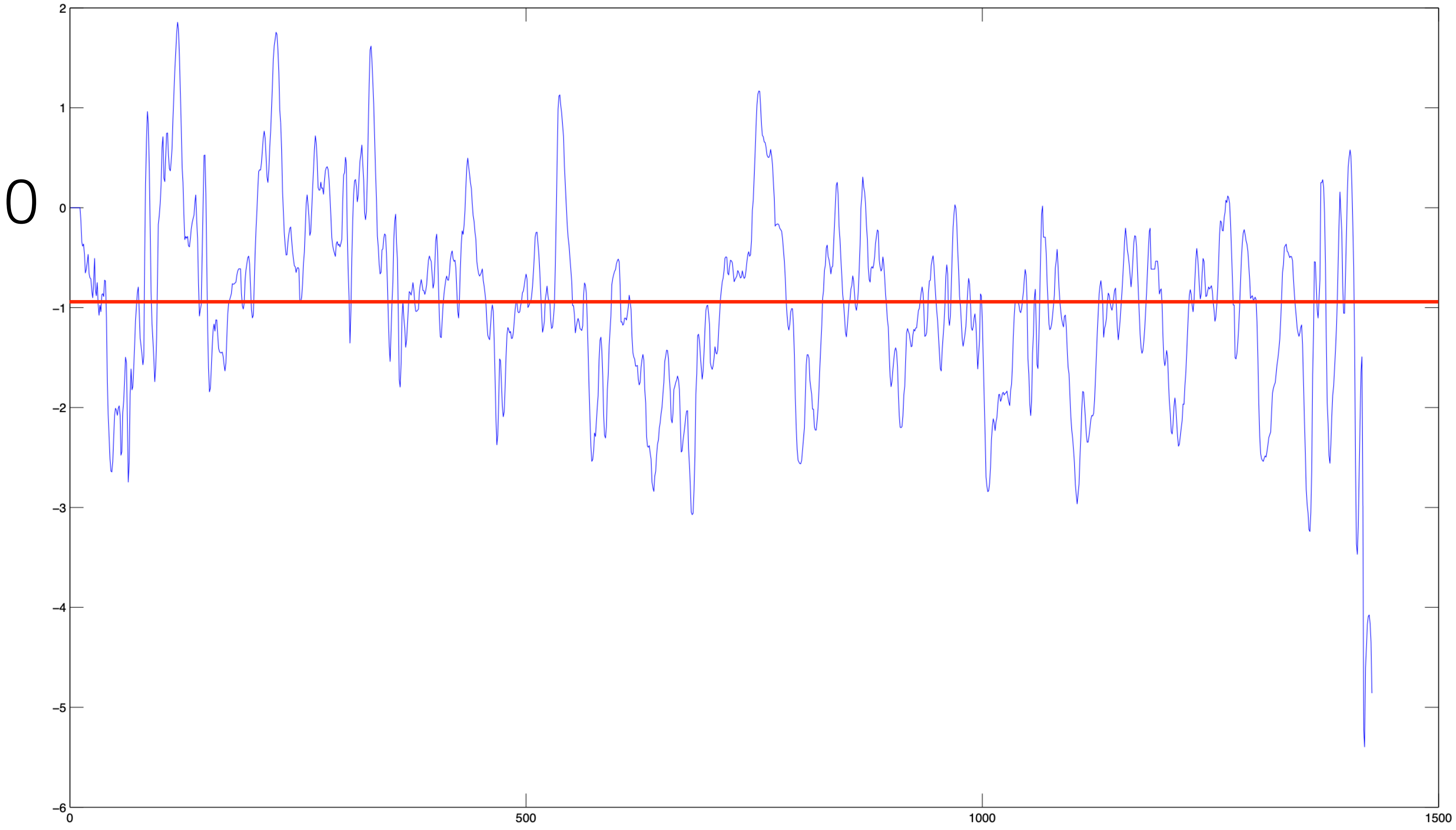
Median Filtering, size = 8

Preprocessing



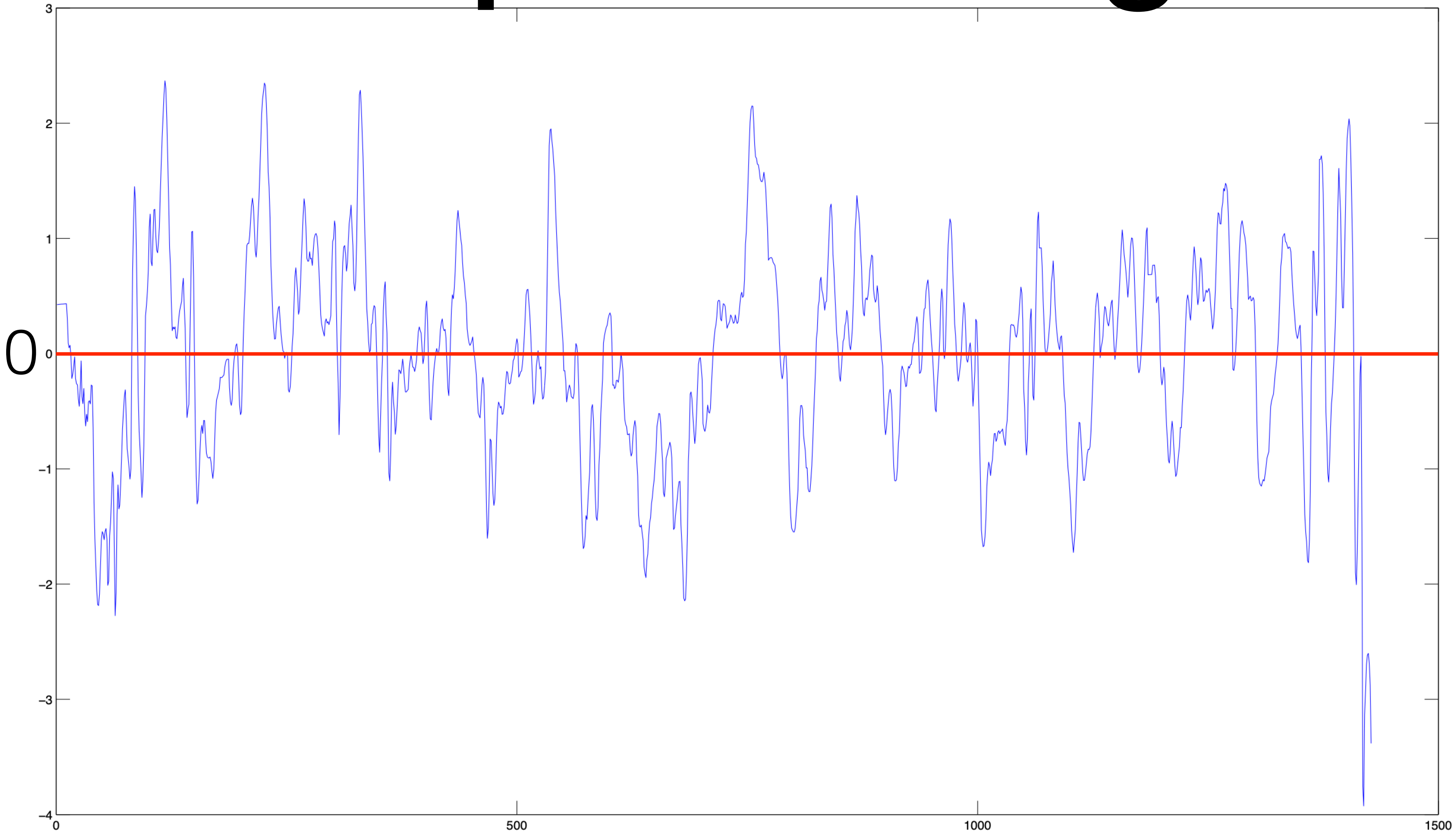
It might not always be clean

Preprocessing



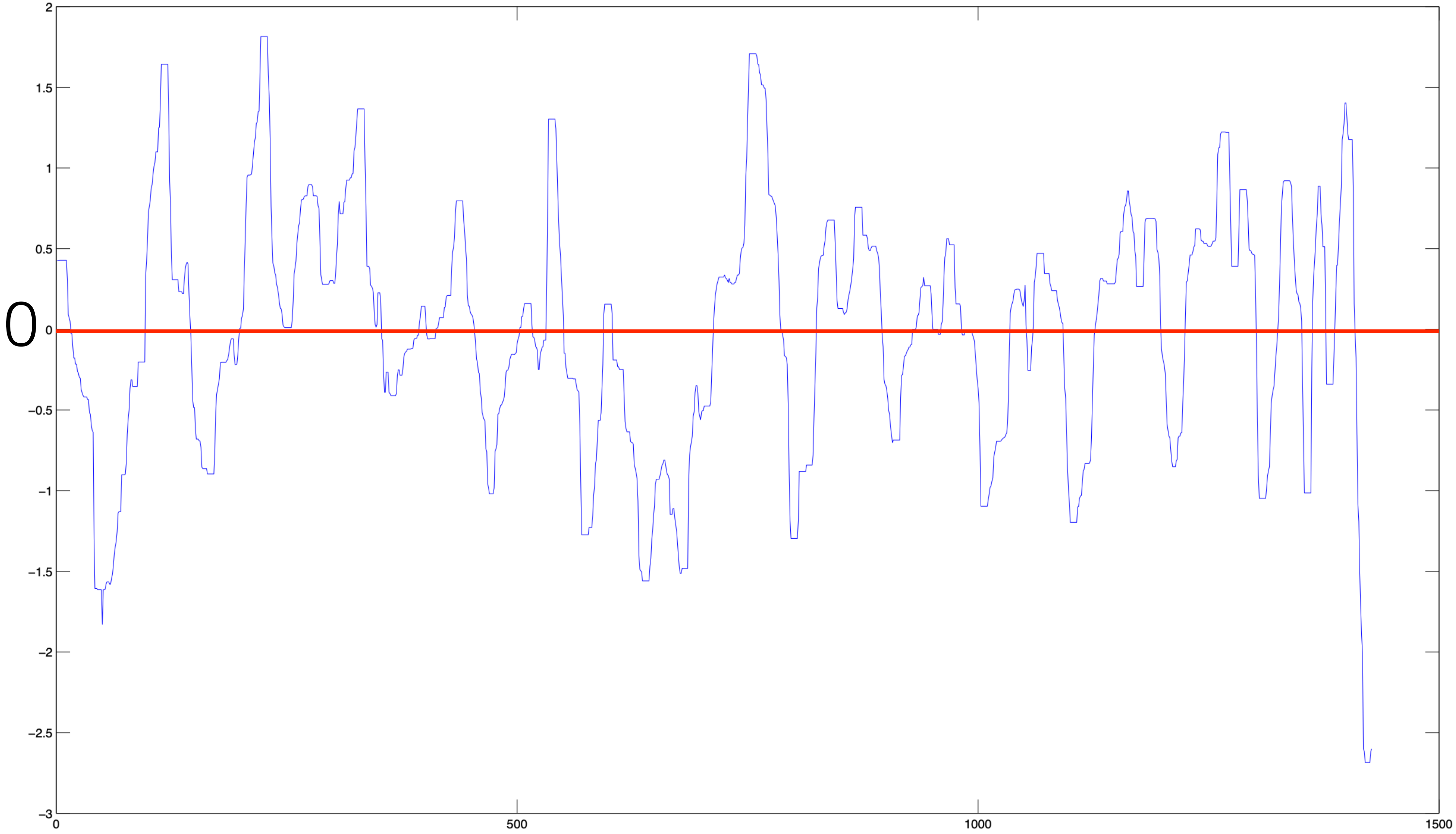
Mean is not zero

Preprocessing



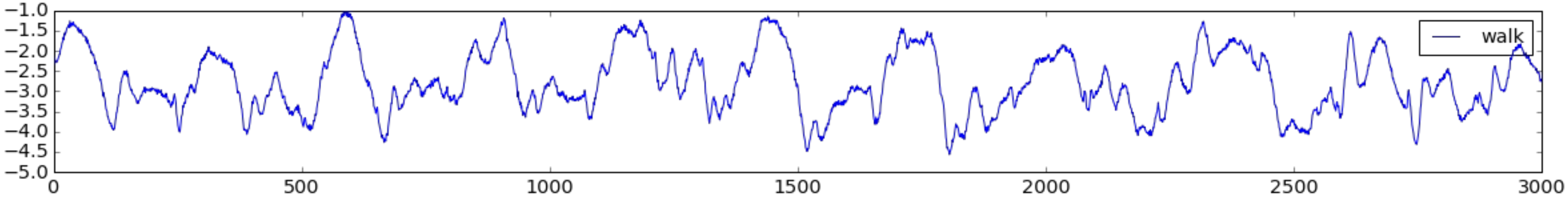
After de-meaning

Preprocessing



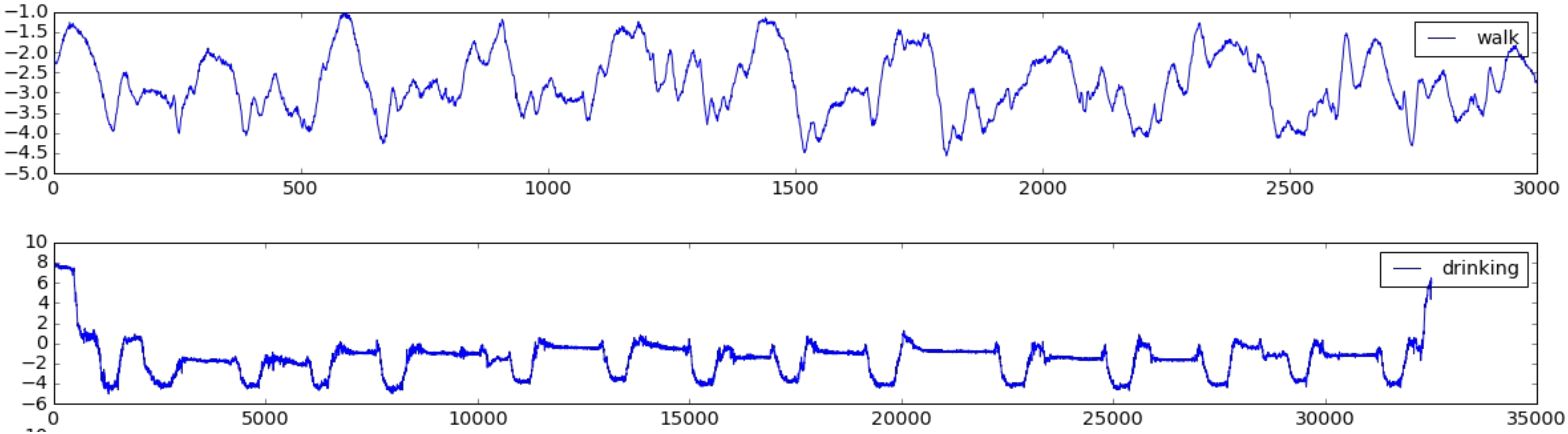
After median filtering

Level/Magnitude



- Max
- Min
- Mean
- Median
- Quantiles

Repetitions



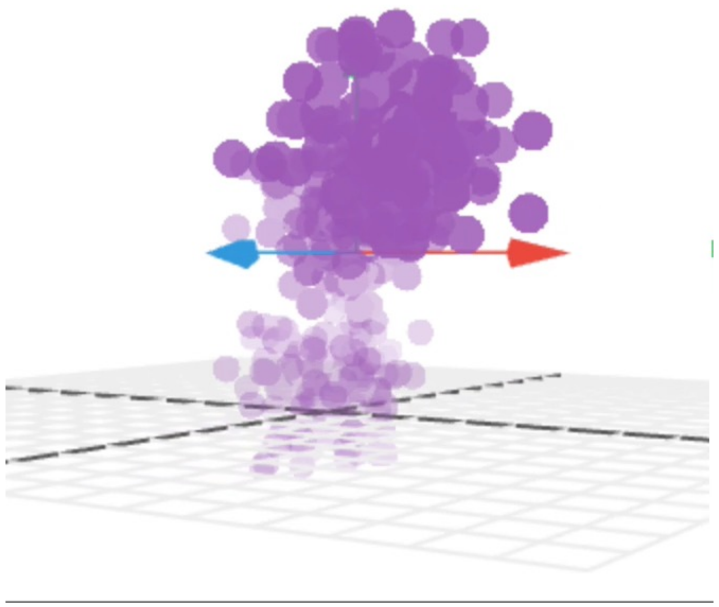
- Zero Crossing
- Frequency Analysis
- Auto-Correlation

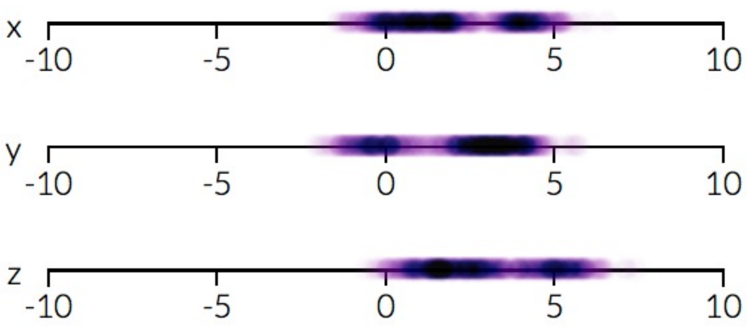
Correct Representation

(e.g., Axes)

- Let us look at some sensor examples
- Domain-dependent
- Magnitude
- Principal Component Analysis

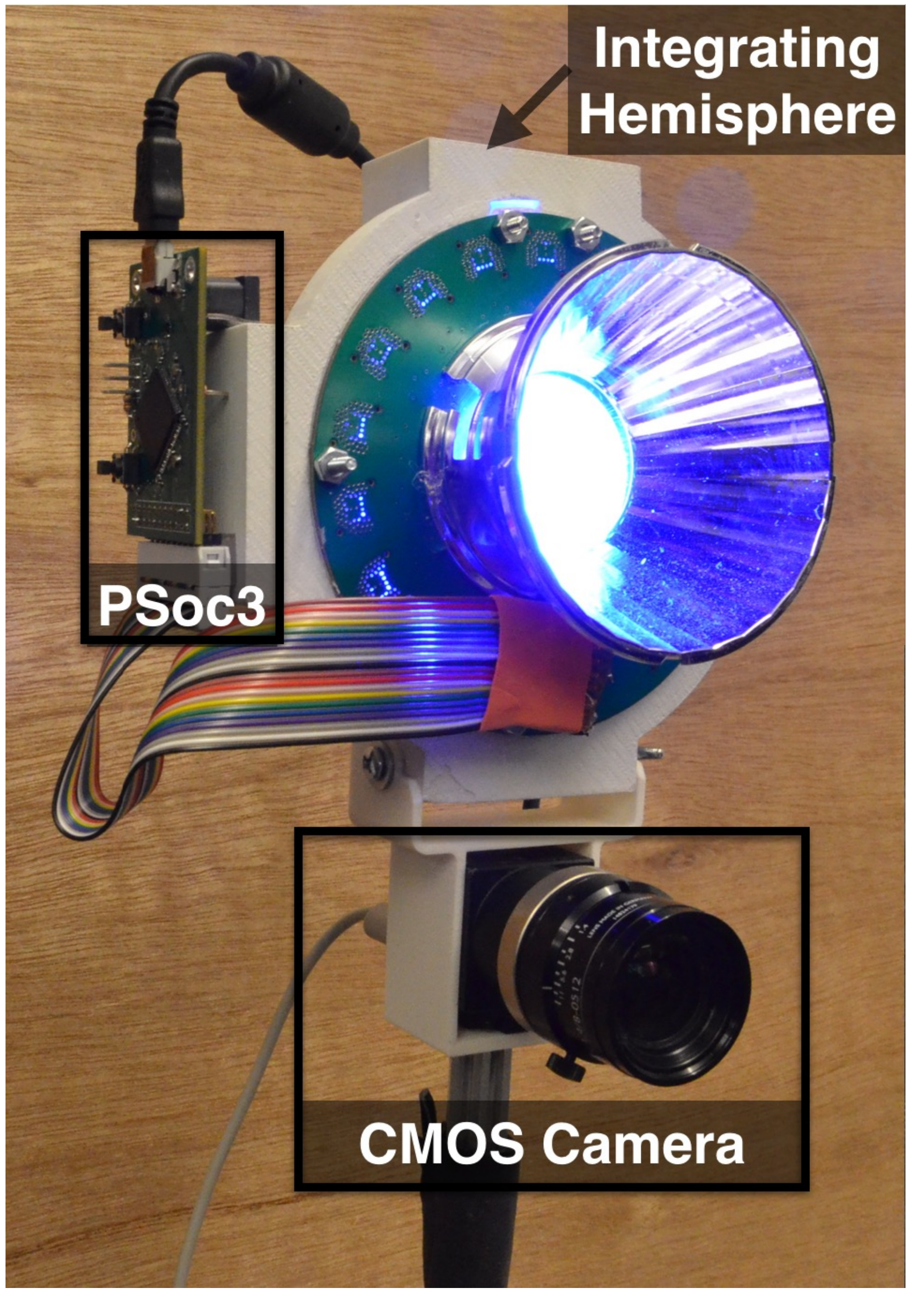
Principal Component Analysis



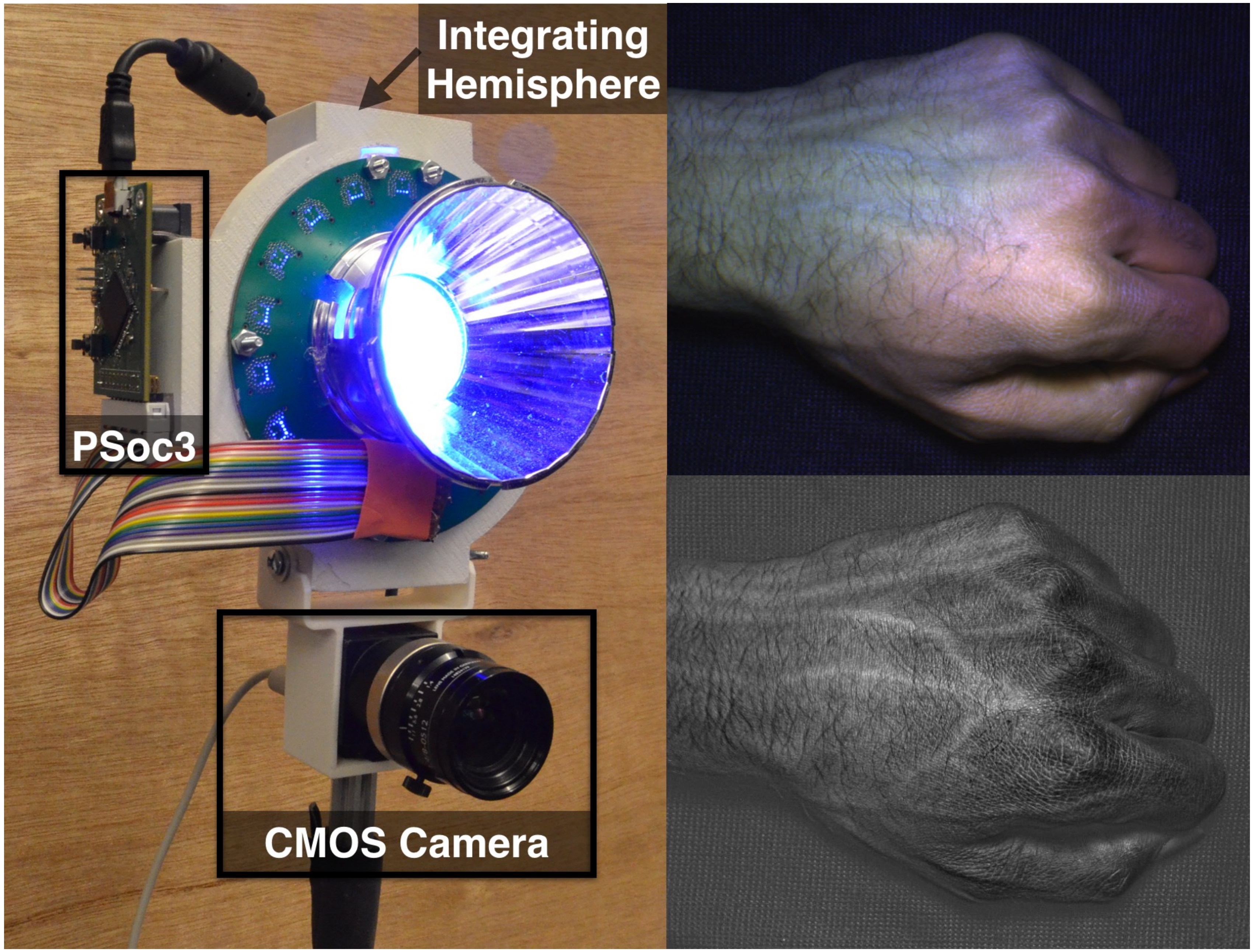


Ref: setosa.io/ev/principal-component-analysis/

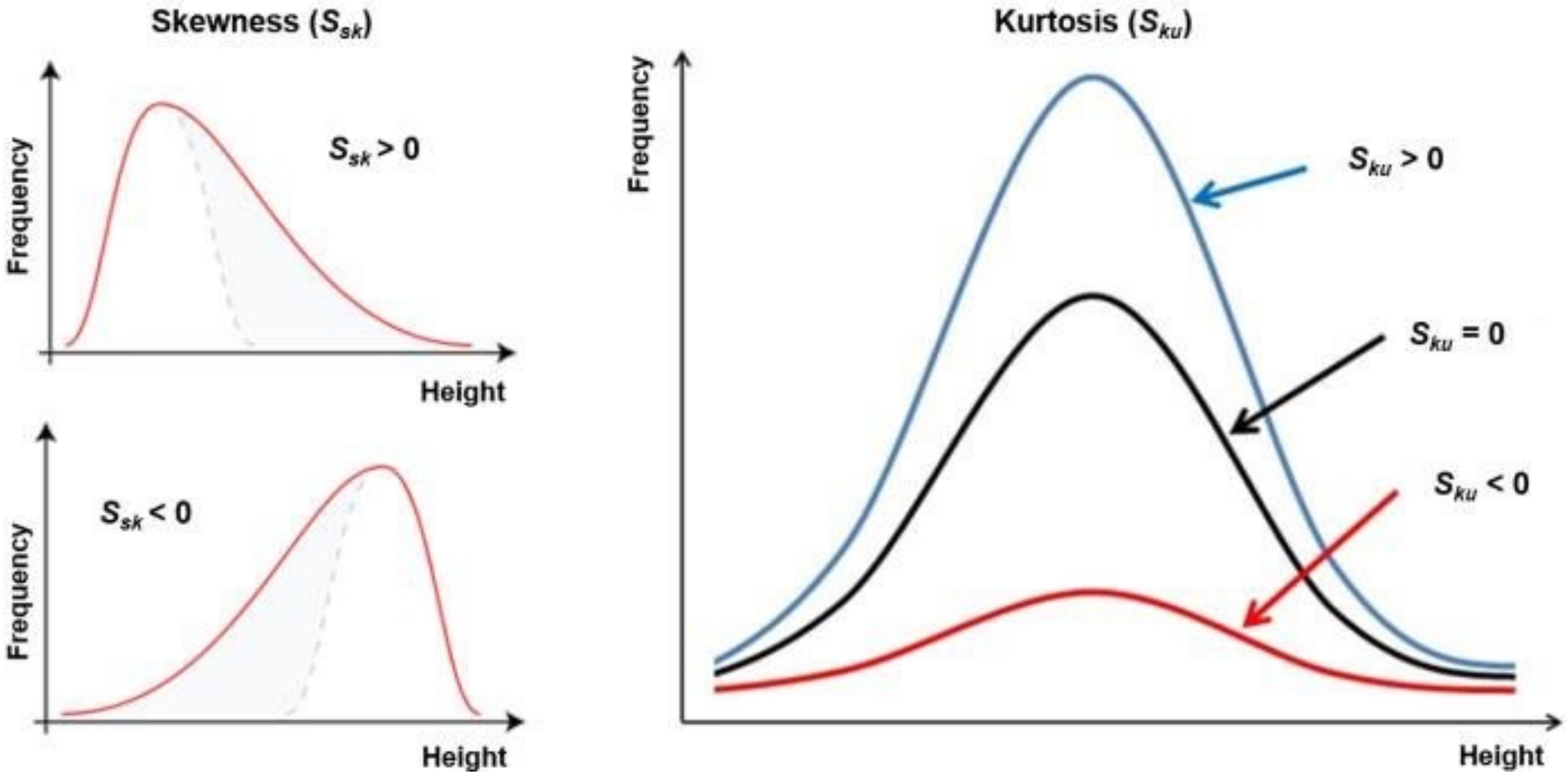
Principal Component Analysis



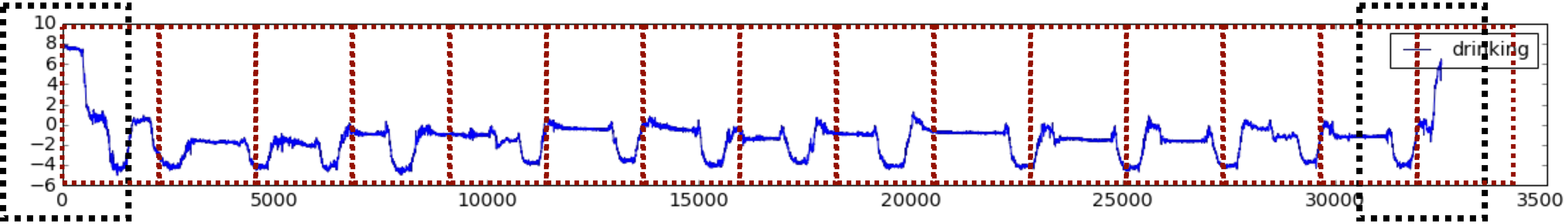
Principal Component Analysis



Shape of the Curve



Windowing



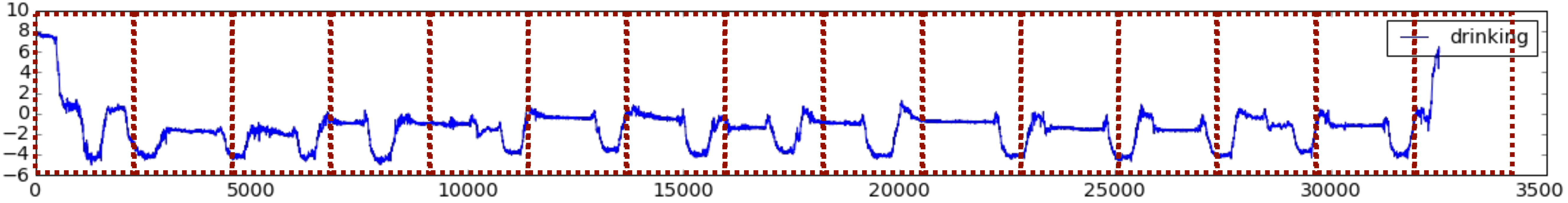
```
Duration, Number of Peaks, Max, nth Quantile, Skewness, Drinking/Walking
```

2 Strategies:

- Make a decision for each window
- Concatenate information from each window into a feature vector

```
Duration, Number of Peaks, Max, nth Quantile, Skewness, Drinking/Walking for w1
Duration, Number of Peaks, Max, nth Quantile, Skewness, Drinking/Walking for w2
Duration, Number of Peaks, Max, nth Quantile, Skewness, Drinking/Walking for w3
Duration, Number of Peaks, Max, nth Quantile, Skewness, Drinking/Walking for w4
- - - - - - - - - - - - - - - - - - COMBINE - - - - - - - - - - - - - - - - - -
```

Windowing



Duration, Number of Peaks, Max, nth Quantile, Skewness, **<u>Drinking/Walking</u>**

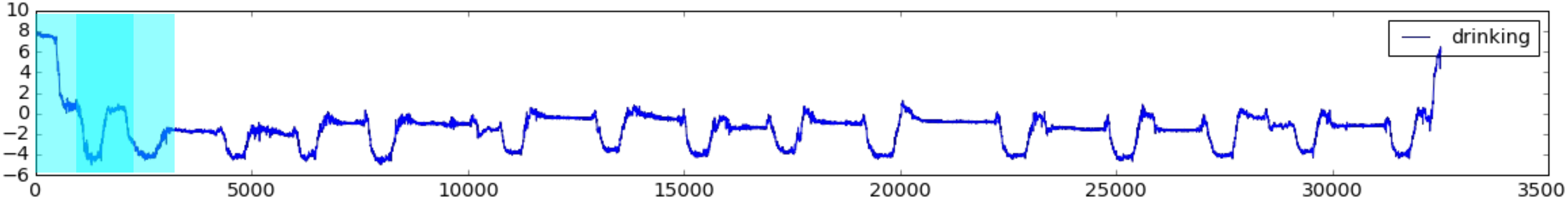
2 Strategies:

- Make a decision for each window
- Concatenate information from each window into a feature vector

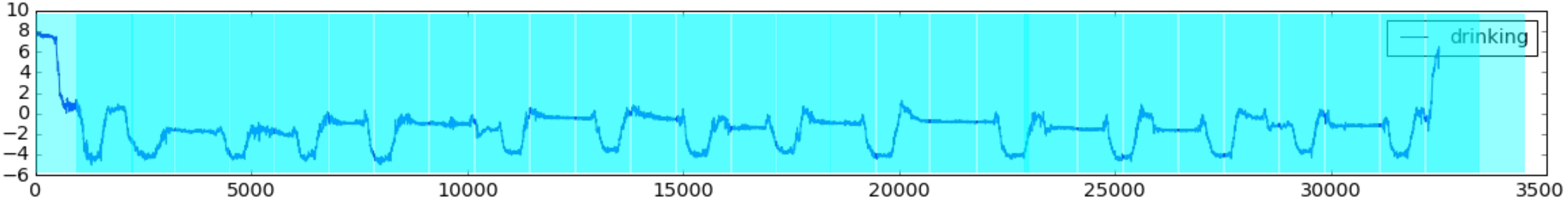
(Duration , Number of Peaks, Max, nth Quantile, Skewness)for w1,
(Duration , Number of Peaks, Max, nth Quantile, Skewness)for w2,
and so on for rest of the windows,
<u>Drinking/Walking</u>

} single feature vector

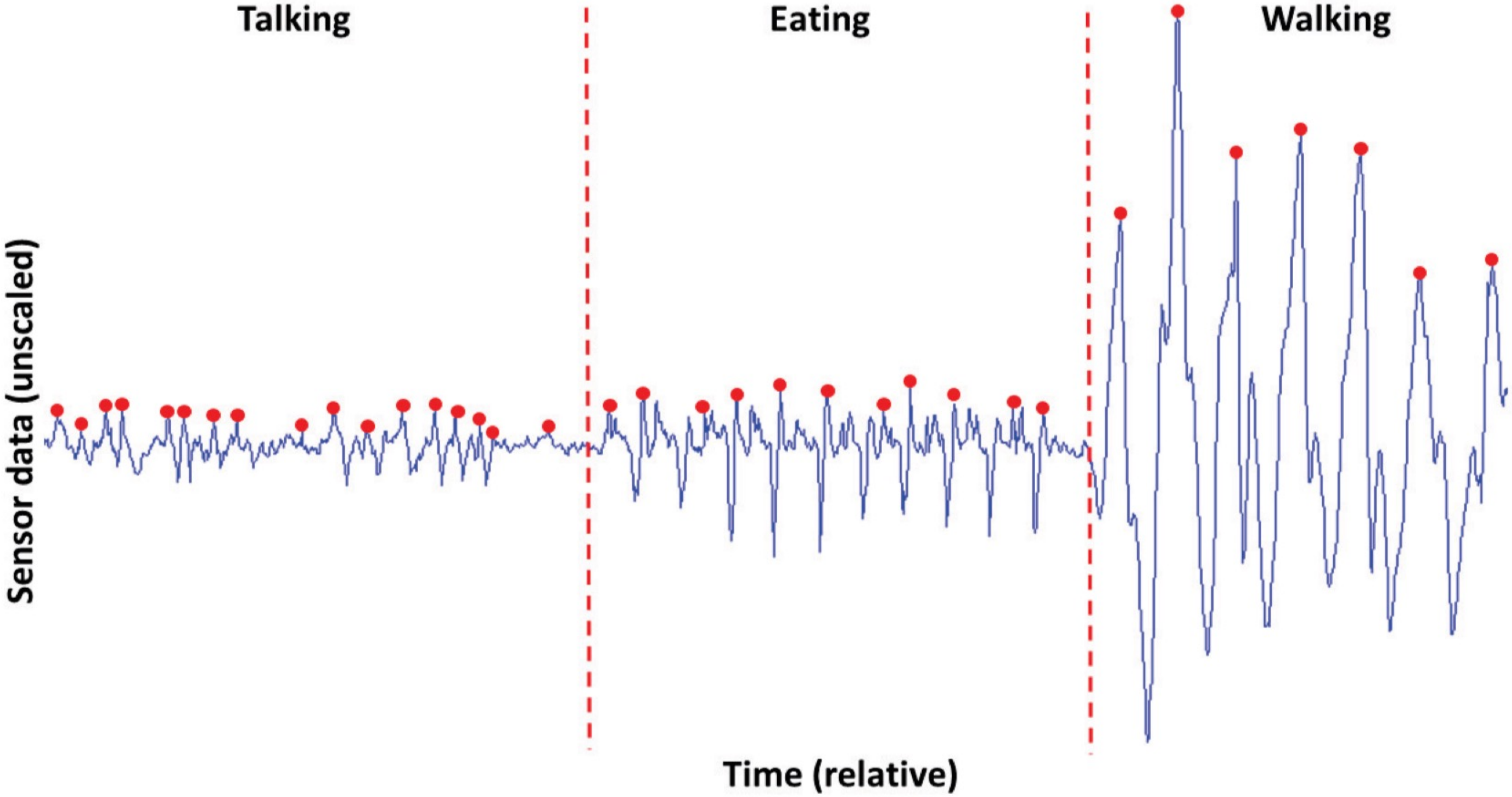
Windowing


10
8
6
4
2
0
−2
−4
−6
0
5000
10000
15000
20000
25000
30000
3500
drinking

Windowing

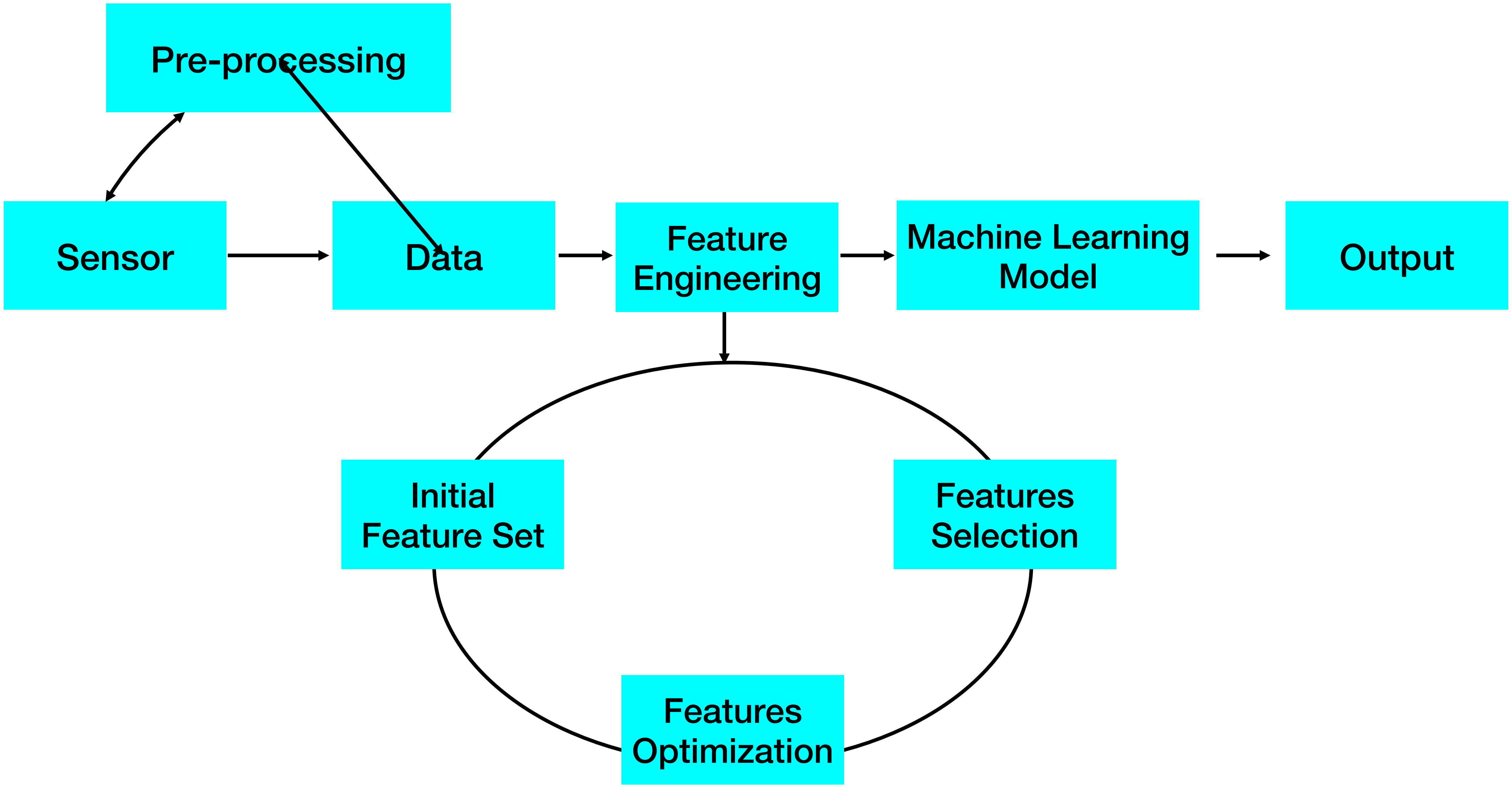

10
8
6
4
2
0
−2
−4
−6
0
5000
10000
15000
20000
25000
30000
3500
drinking

Do we keep running all the data through the whole ML pipeline?



Bedri et al.

Segmentation



Pre-processing

Sensor

Data

Feature Engineering

Machine Learning Model

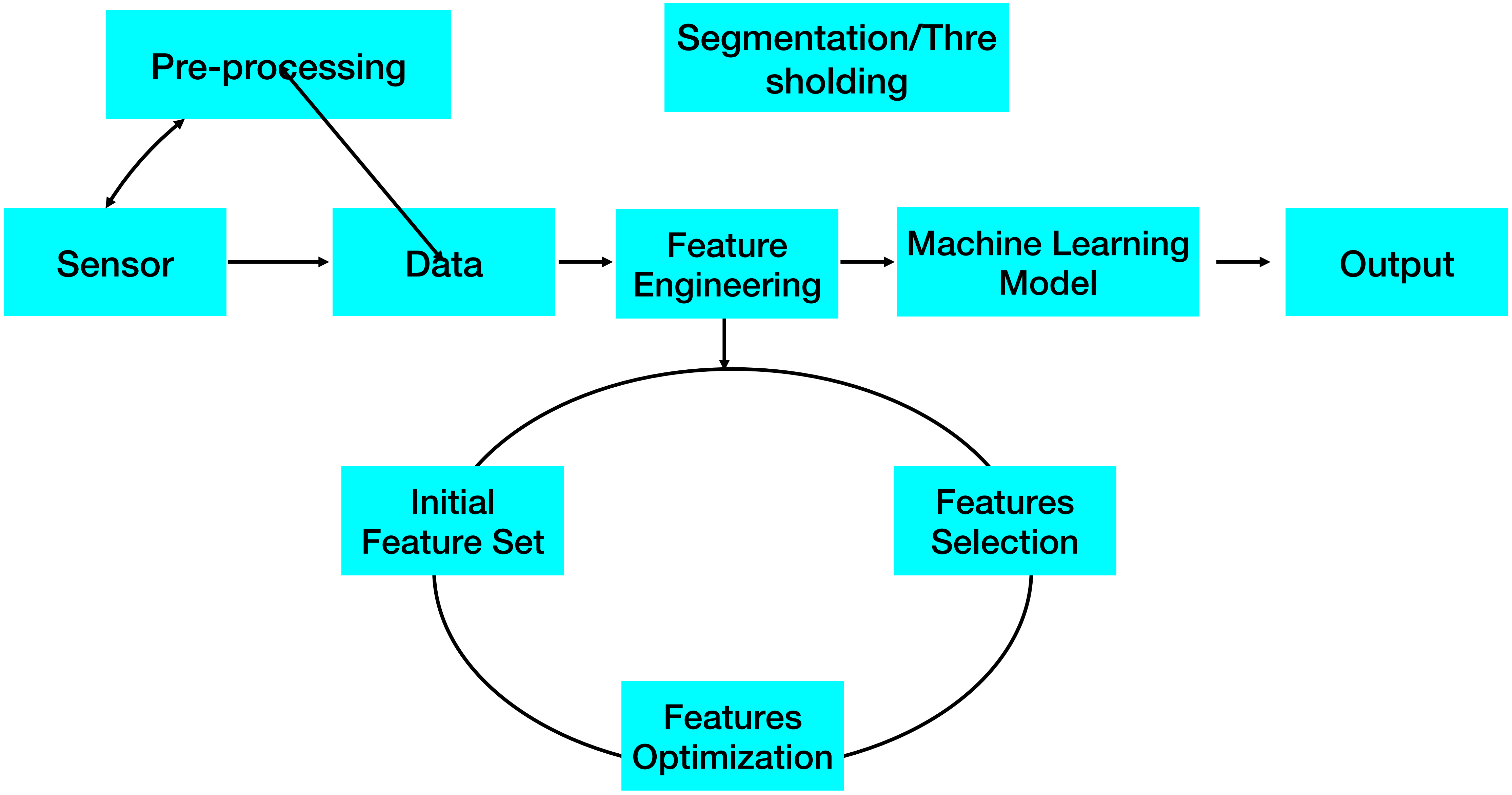
Output

Initial Feature Set

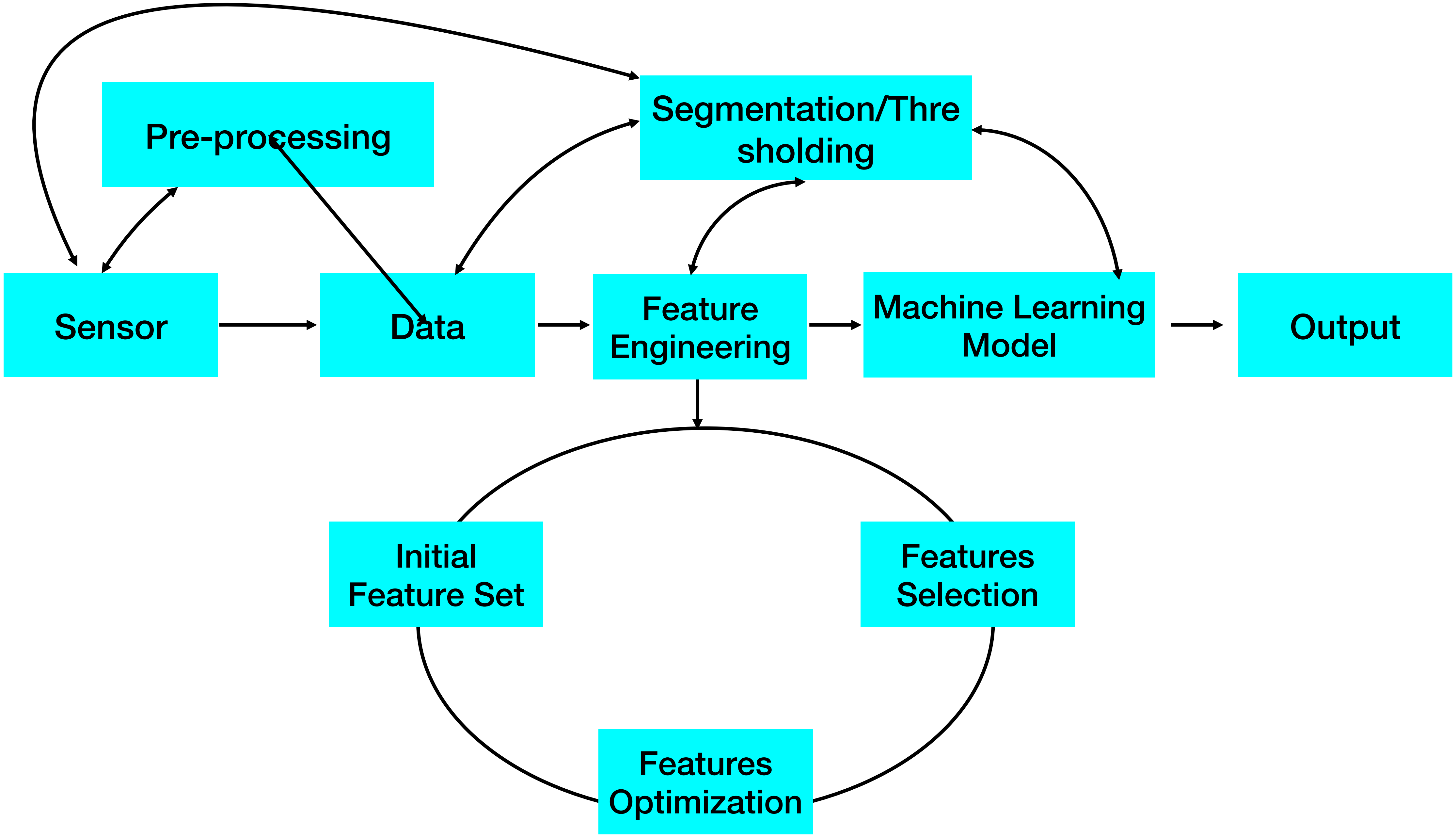
Features Selection

Features Optimization

Segmentation



Segmentation


Segmentation/Thre sholding
Pre-processing
Sensor
Data
Feature Engineering
Machine Learning Model
Output
Initial Feature Set
Features Selection
Features Optimization