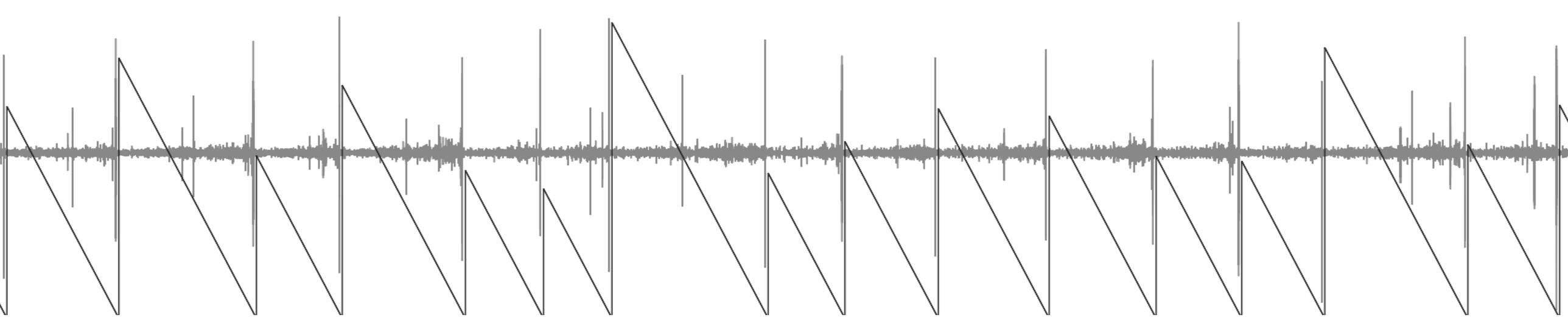


Shaking Earth: LANL Earthquake Prediction Challenge

Nelli Fedorova

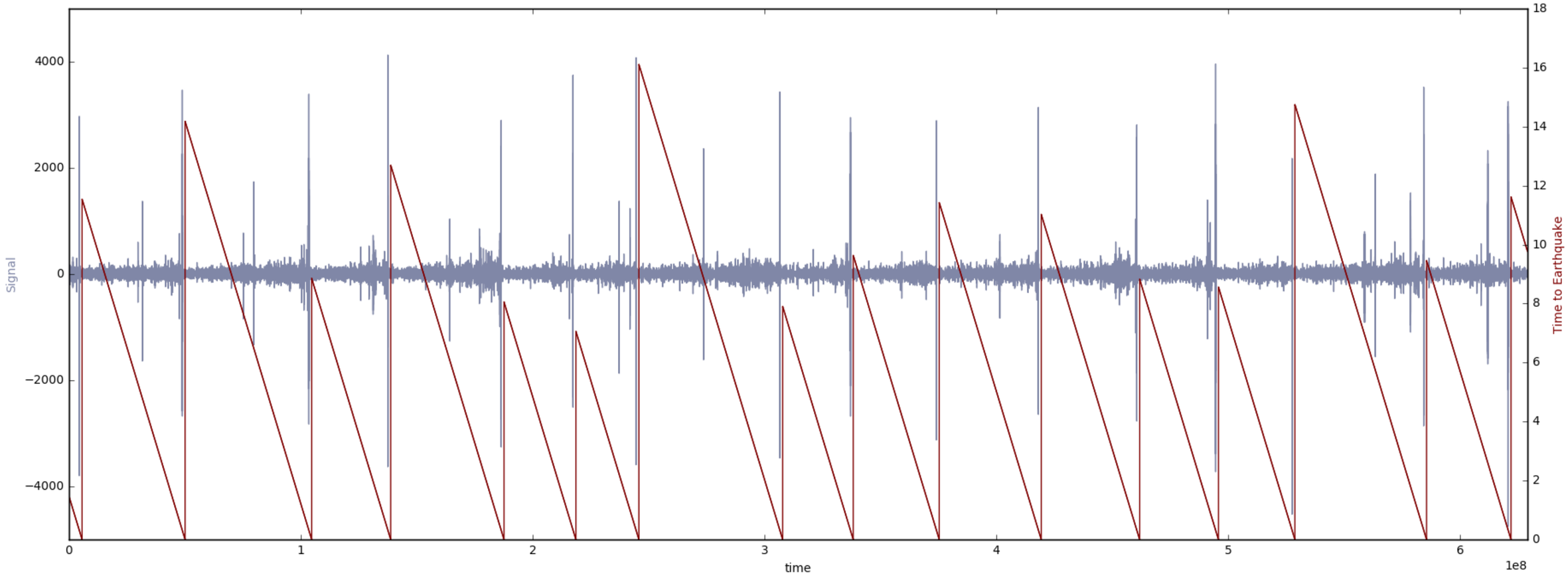
Data

https://www.kaggle.com/c/LANL-Earthquake-Prediction

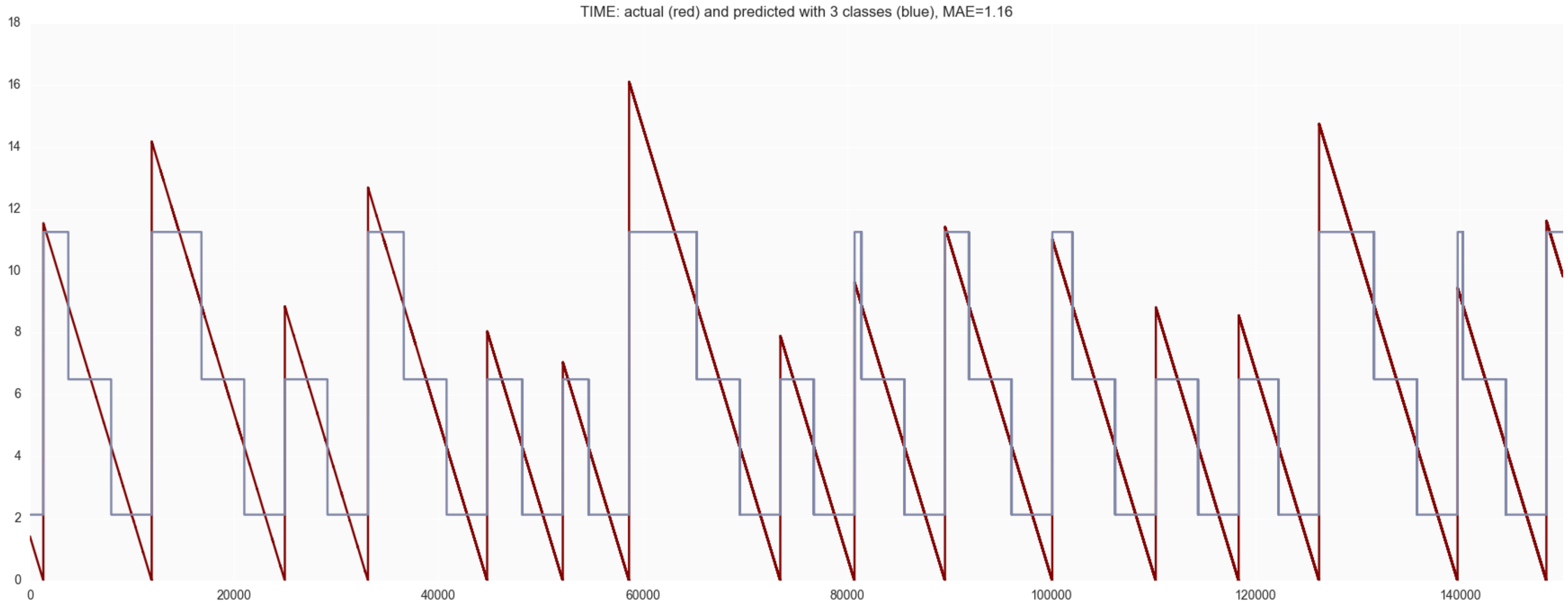
- **Goal**
  - predict time left to the next laboratory earthquake based on seismic signal of length 150,000
- **Data**
  - **Training:** ~629,100,000 subsequent observations with 17 earthquakes
  - **Test:** 2624 separate fragments of length 150,000
- **Performance measure**
  - MAE (mean absolute error)
- **Best Kaggle score (August 2019)**
  - MAE = 1.08 (public, 13%), 2.26 (private, 87%)

Training data: 17 earthquakes, ~629M points

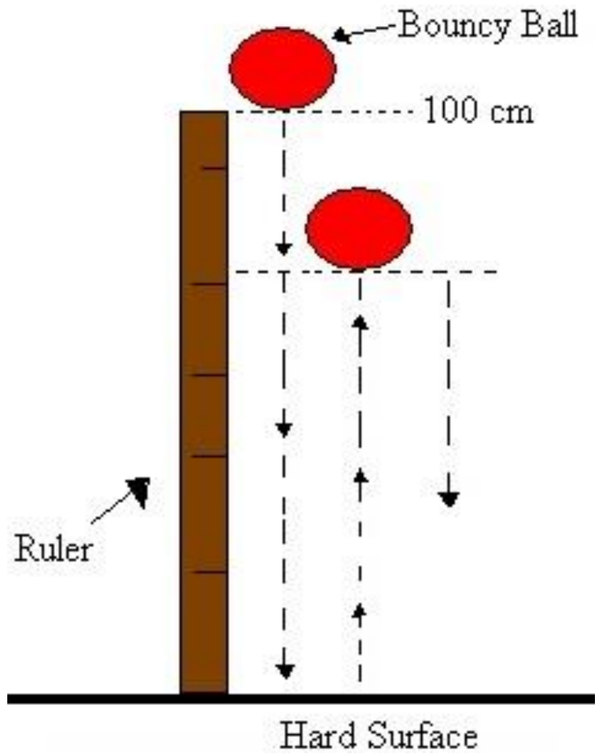
https://www.kaggle.com/c/LANL-Earthquake-Prediction

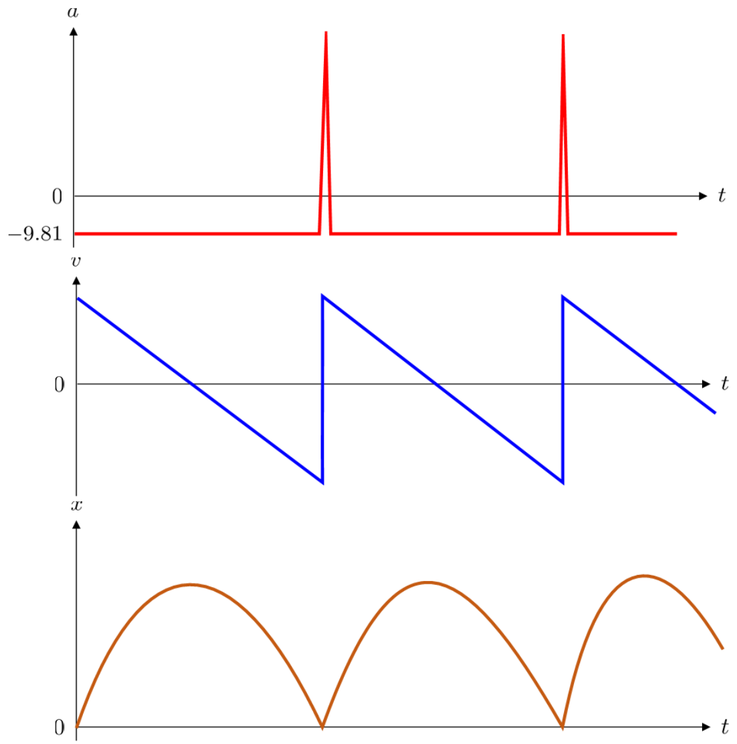


Training data: 3 classes, MAE~1.16

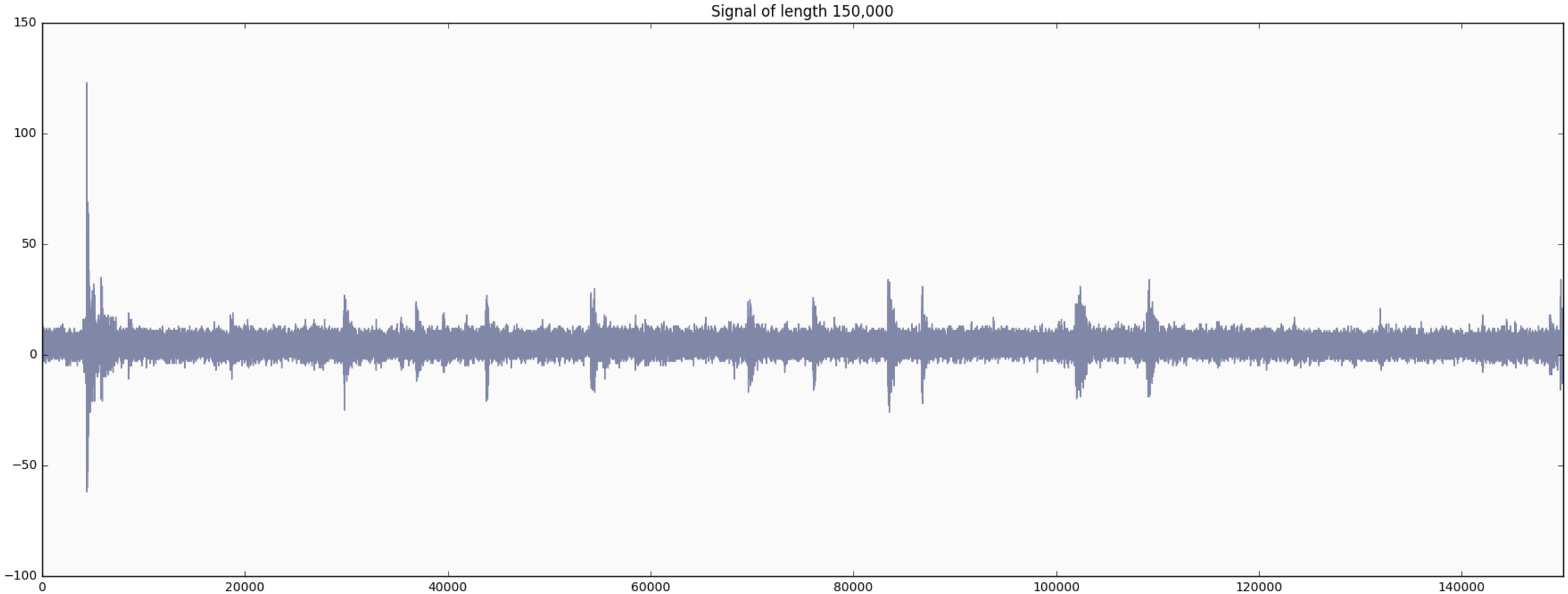


Bouncy ball analogy

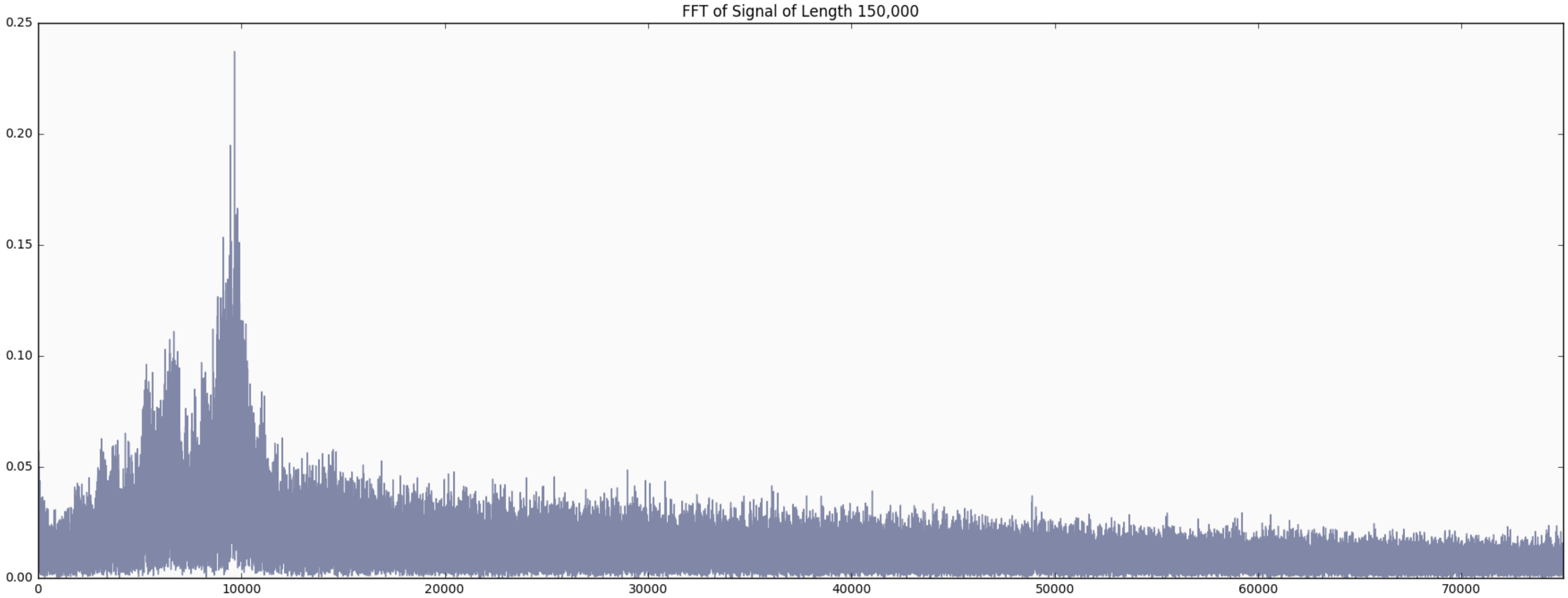




Seismic signal of length 150,000



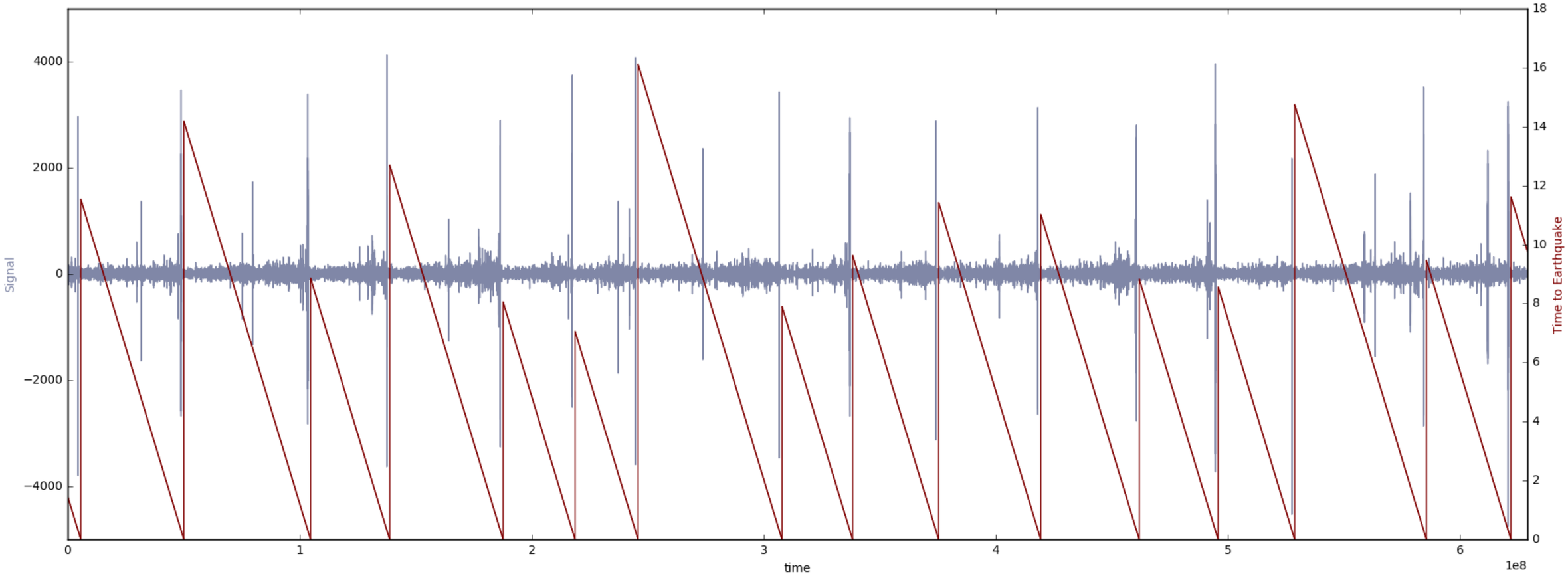
FFT coefficients



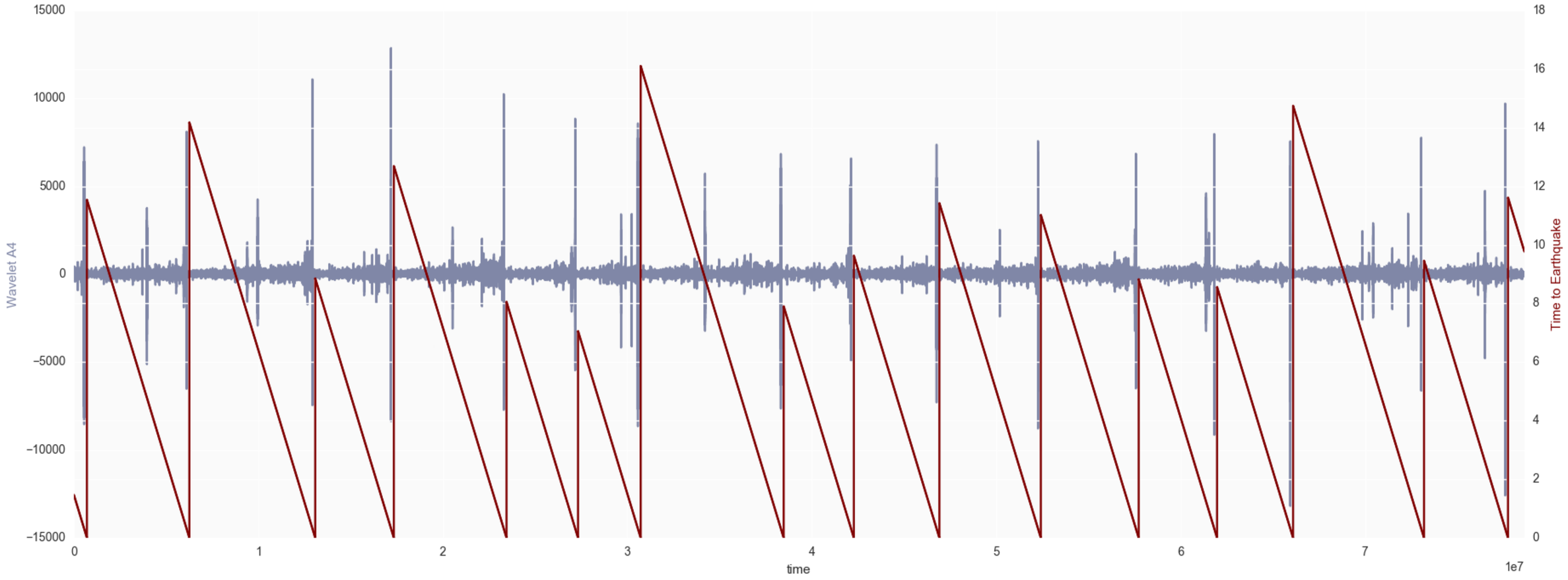
Plan

1. Compute descriptors for each signal, its FFT and wavelet (A,D) at different levels
2. Select descriptors with strongest Pearson's and Spearman's correlation with TIME
3. Build KNN
   - Predict TIME
   - Use to select training points near TEST data
4. Build and cross-validate FF, XGB, SVR, RCV, LGB
   - Predict TIME
   - Use to select best descriptors

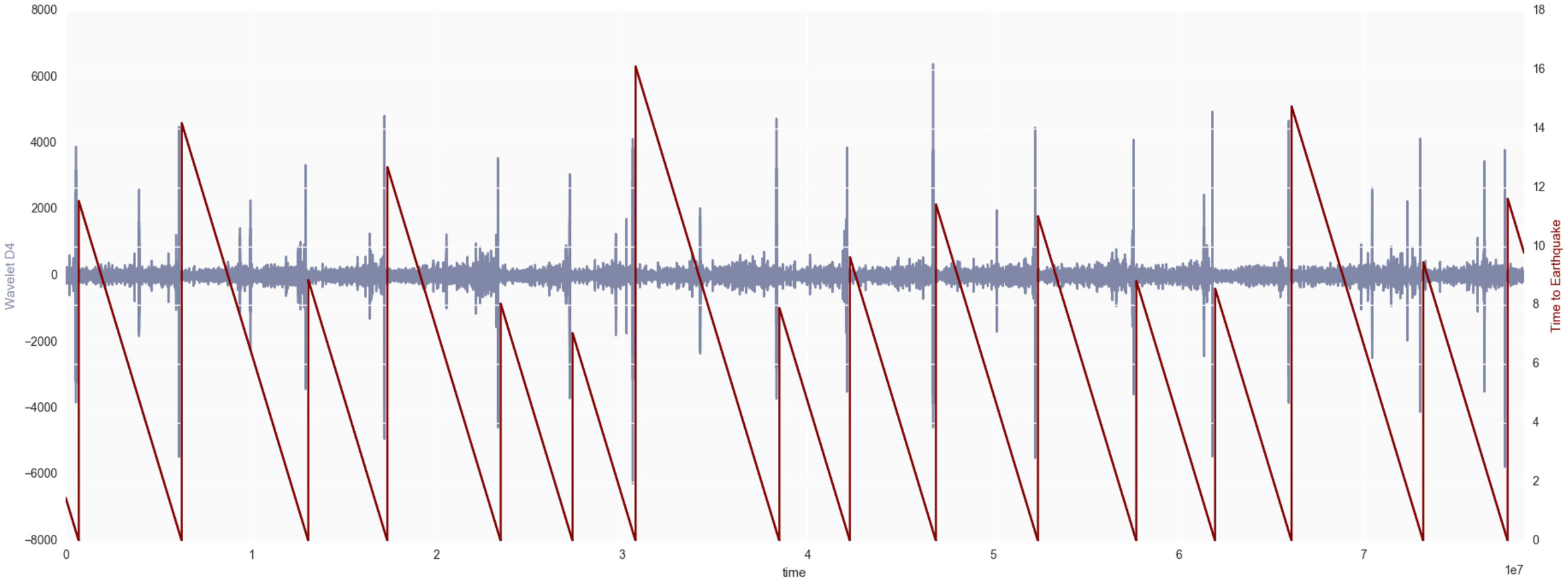
Original Data (dim=150,000)



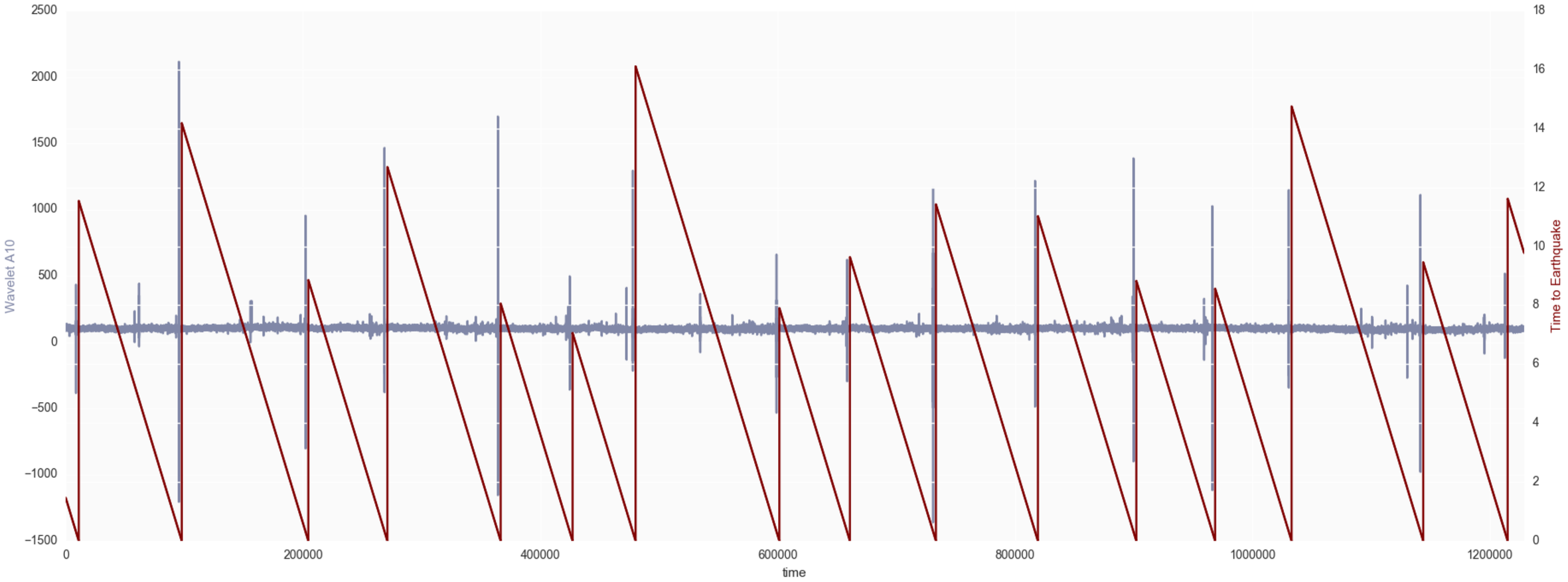
Haar Wavelet A4 (dim=18,750)



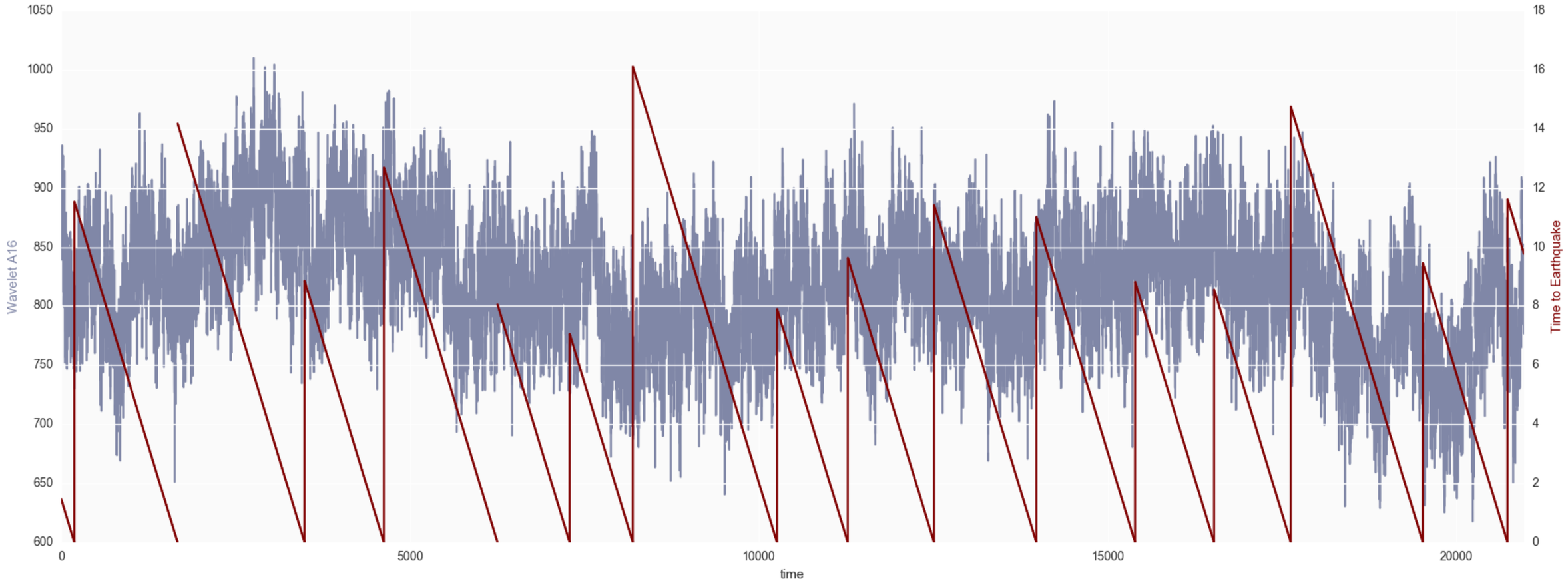
Haar Wavelet D4 (dim=18,750)



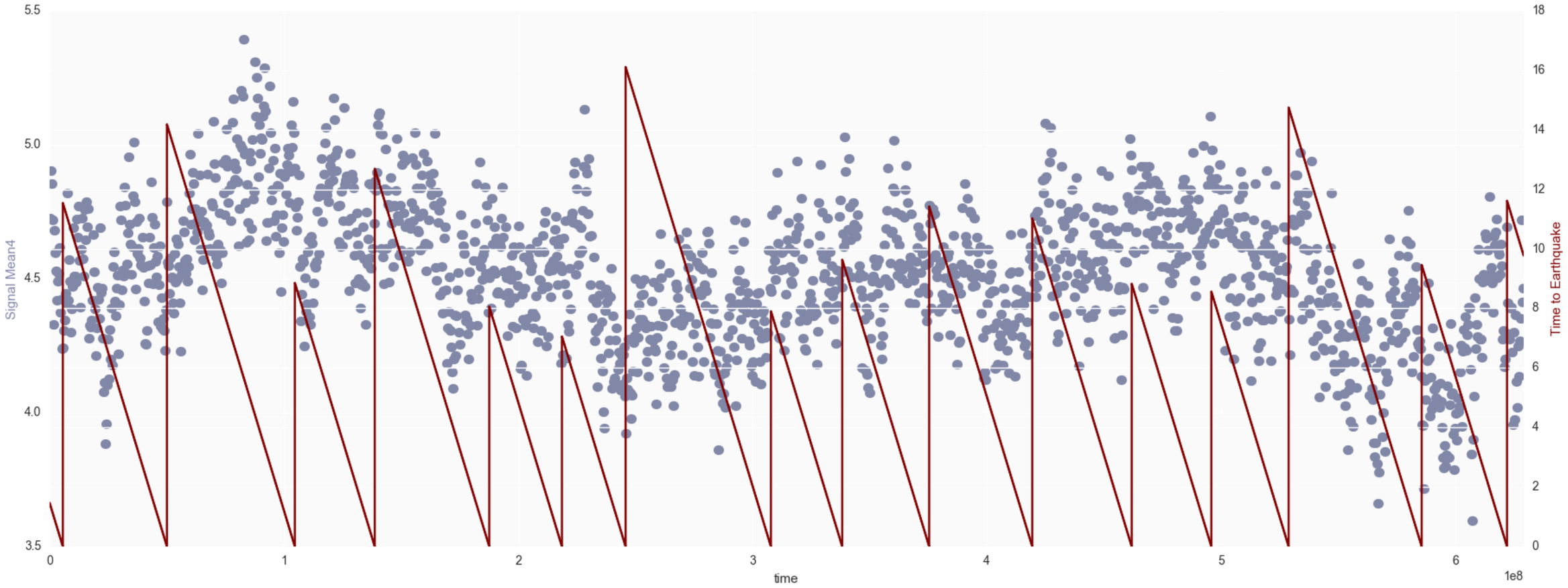
Haar Wavelet A10 (dim=293)



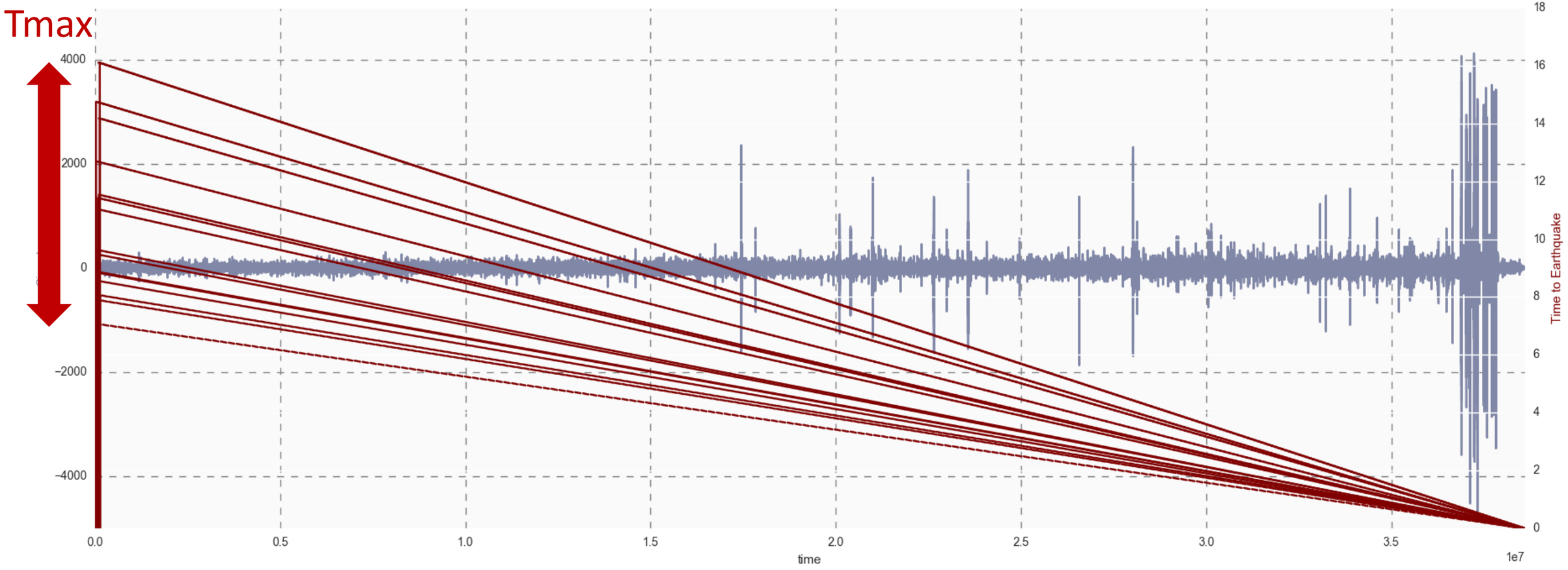
Haar Wavelet A16 (dim=4)



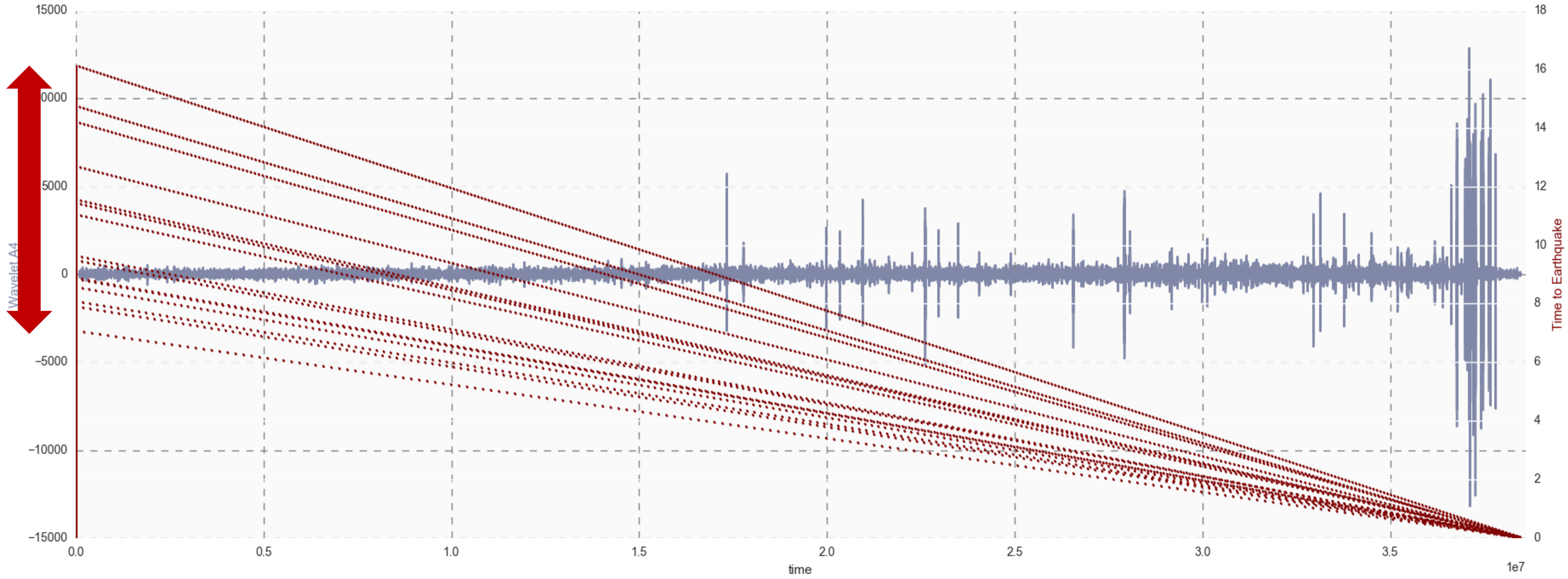
Signal mean


Signal Mean4
5.5
5.0
4.5
4.0
3.5
0
1
2
3
4
5
6
time
1e8
Time to Earthquake
18
16
14
12
10
8
6
4
2
0

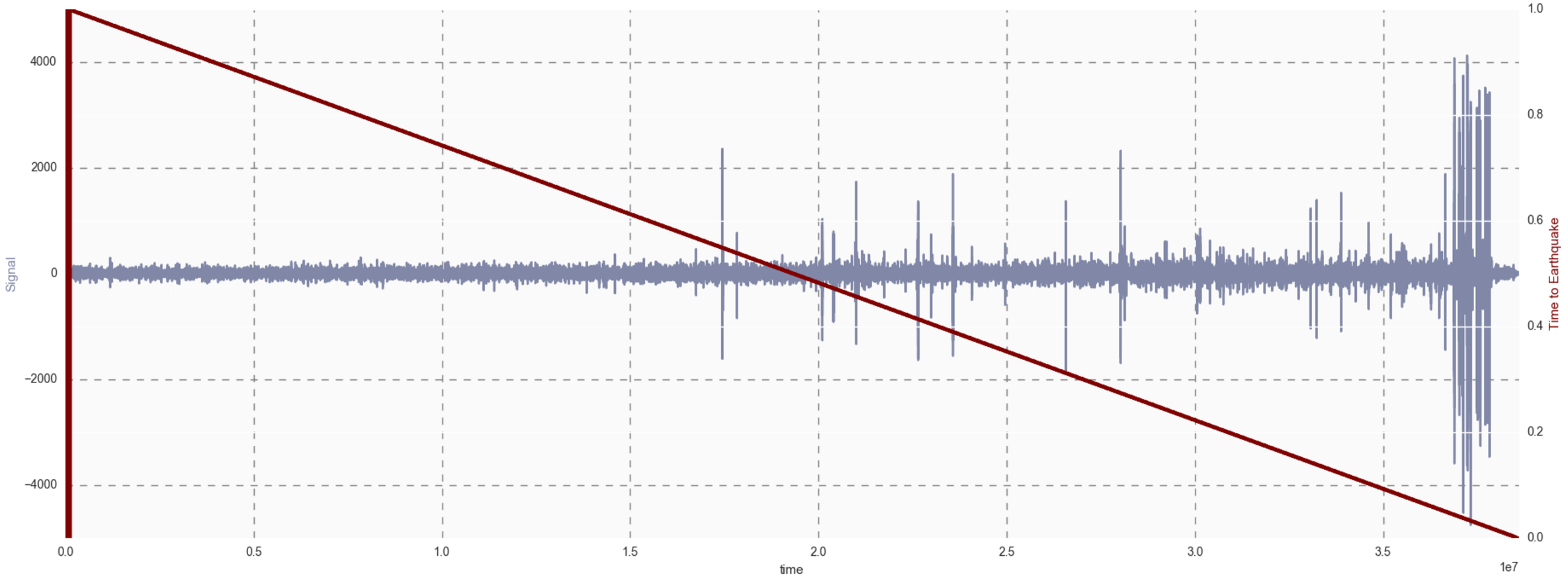
Normalized Earthquakes
Signal (dim=150,000) & TIME



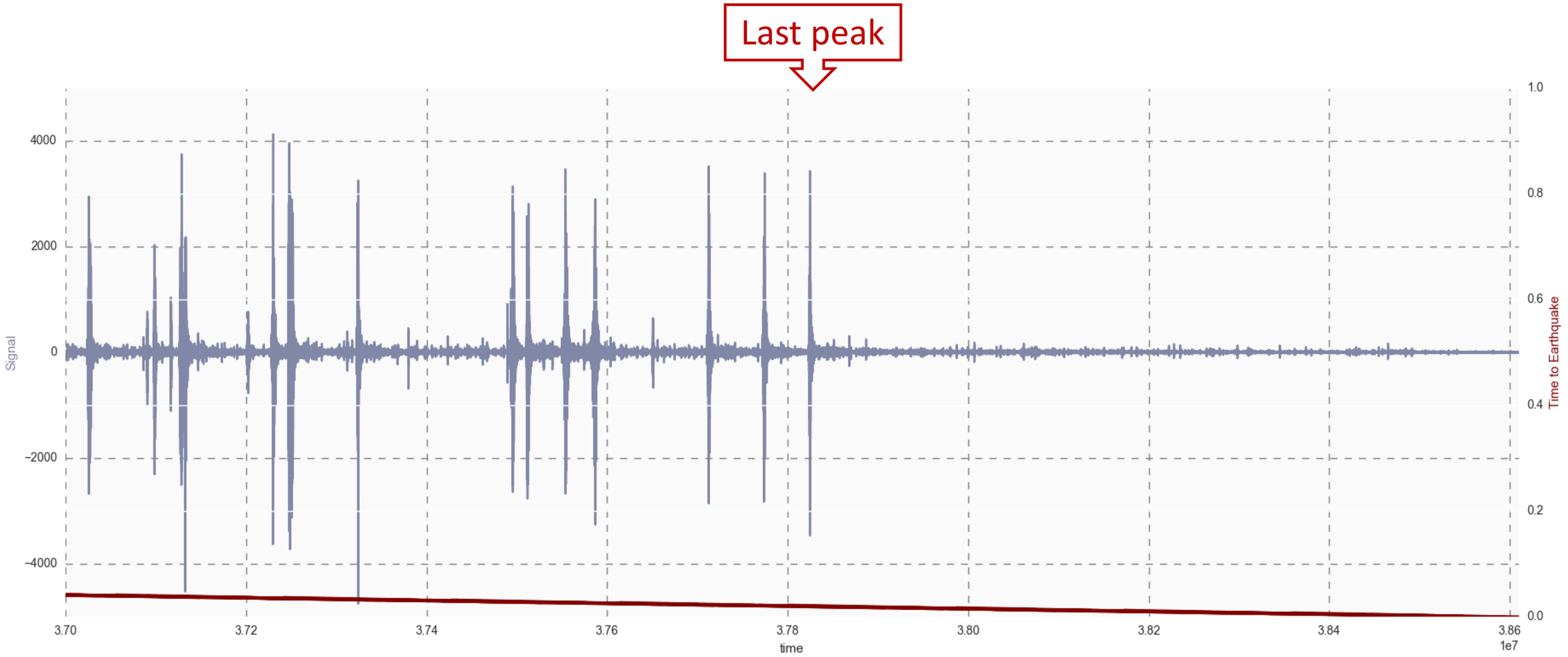
Normalized Earthquakes
Wavelet A4 (dim=18,750) & TIME



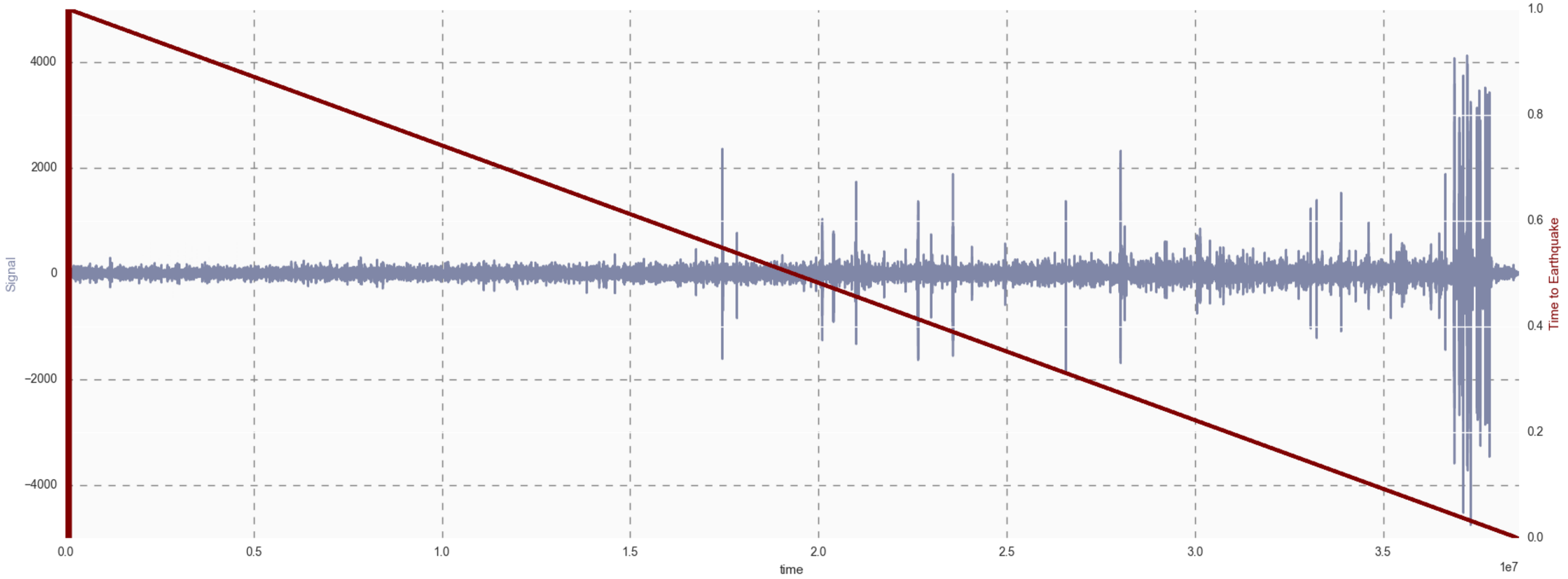
Normalized Earthquakes
Signal & normalized TIME



Normalized Earthquakes (zoomed)
Signal & normalized TIME

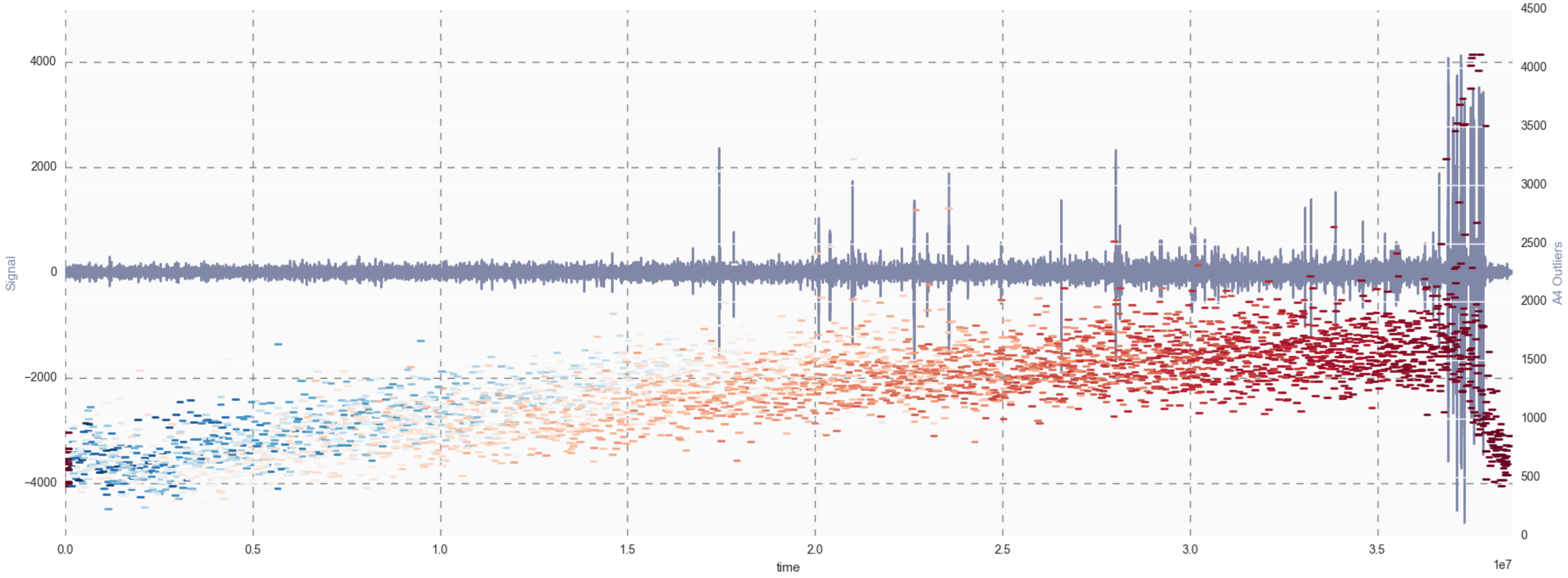


Normalized Earthquakes
Signal & normalized TIME



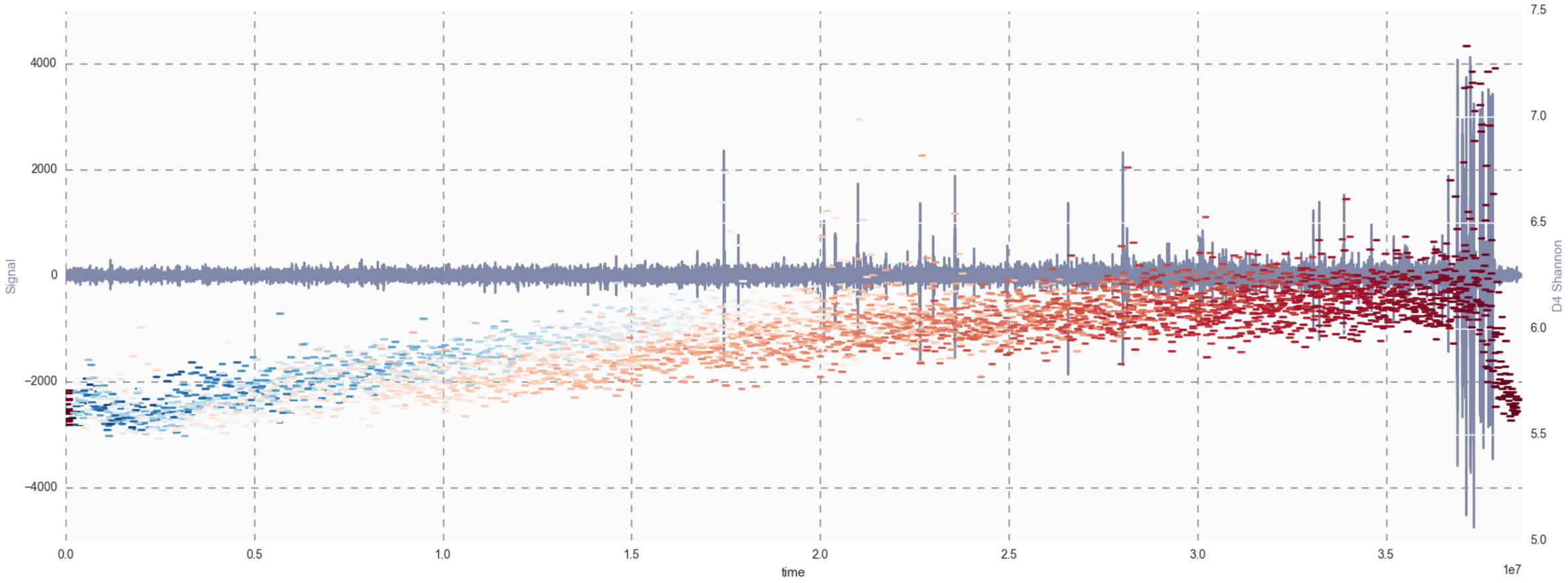
Normalized Earthquakes Signal & Number of Peaks

Time to Earthquake: small (red), medium (orange), large (blue)



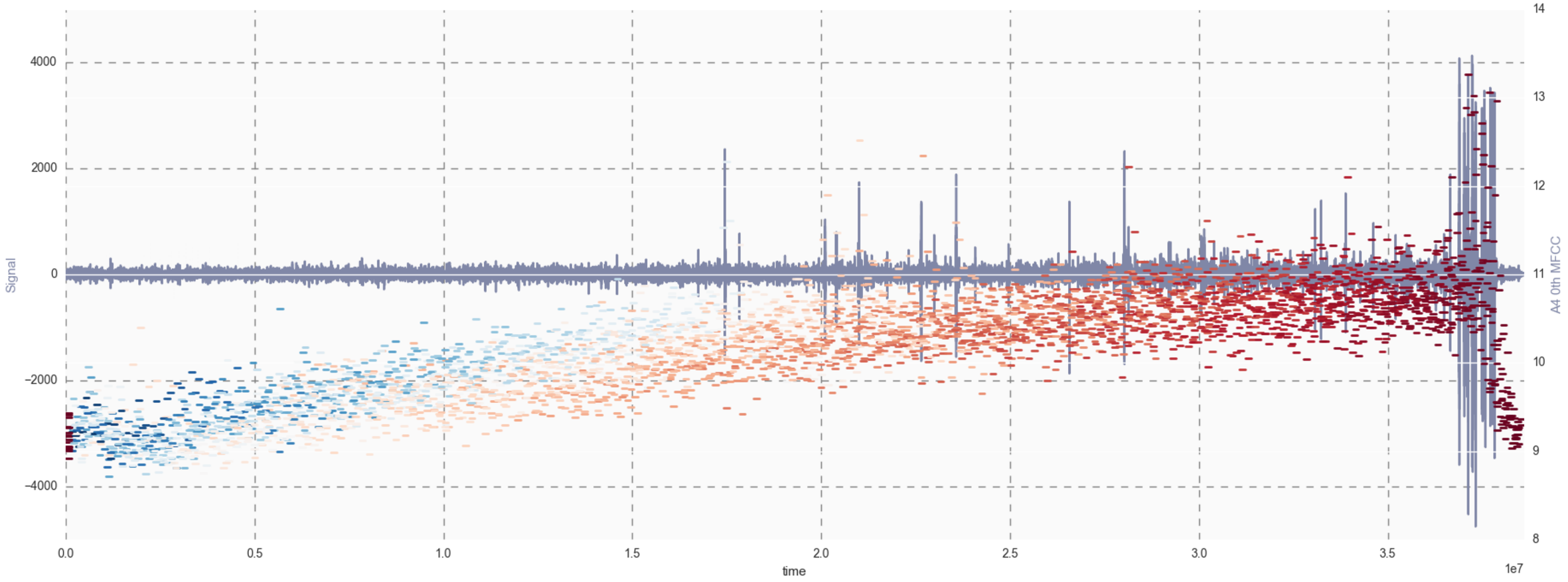
Normalized Earthquakes
Signal & Shannon's Entropy

Time to Earthquake: small (red), medium (orange), large (blue)

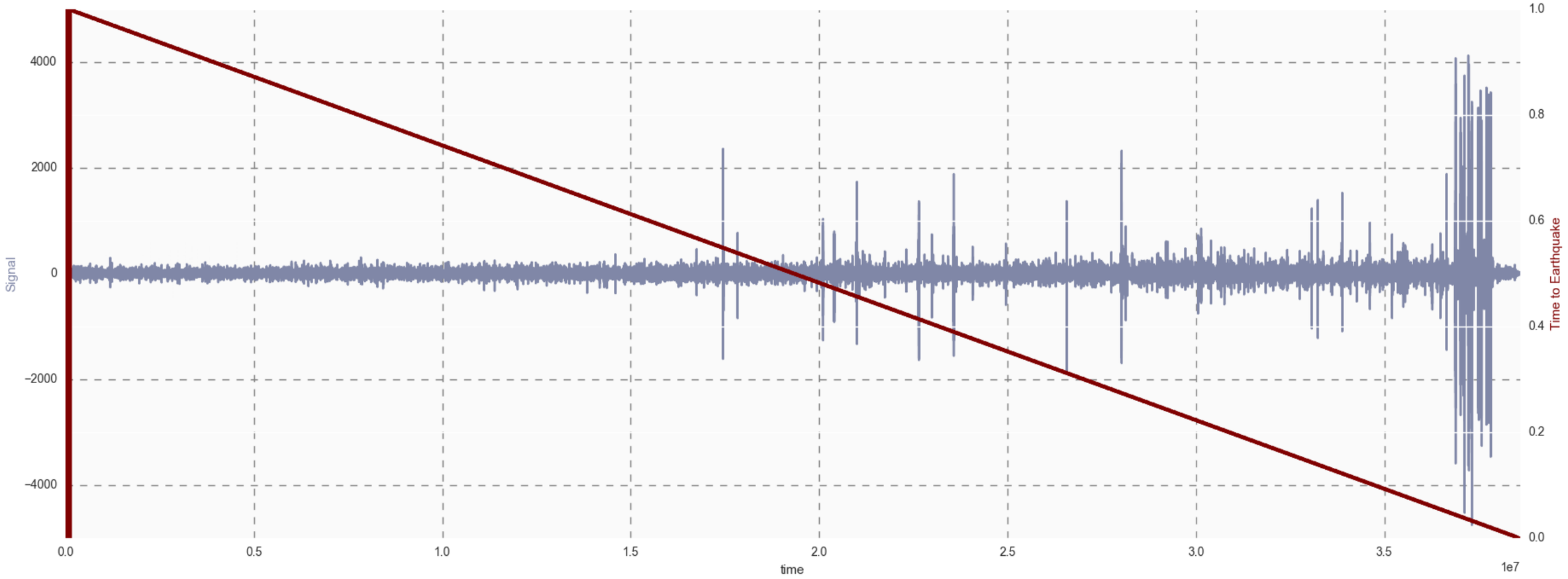


Normalized Earthquakes Signal & MFCC

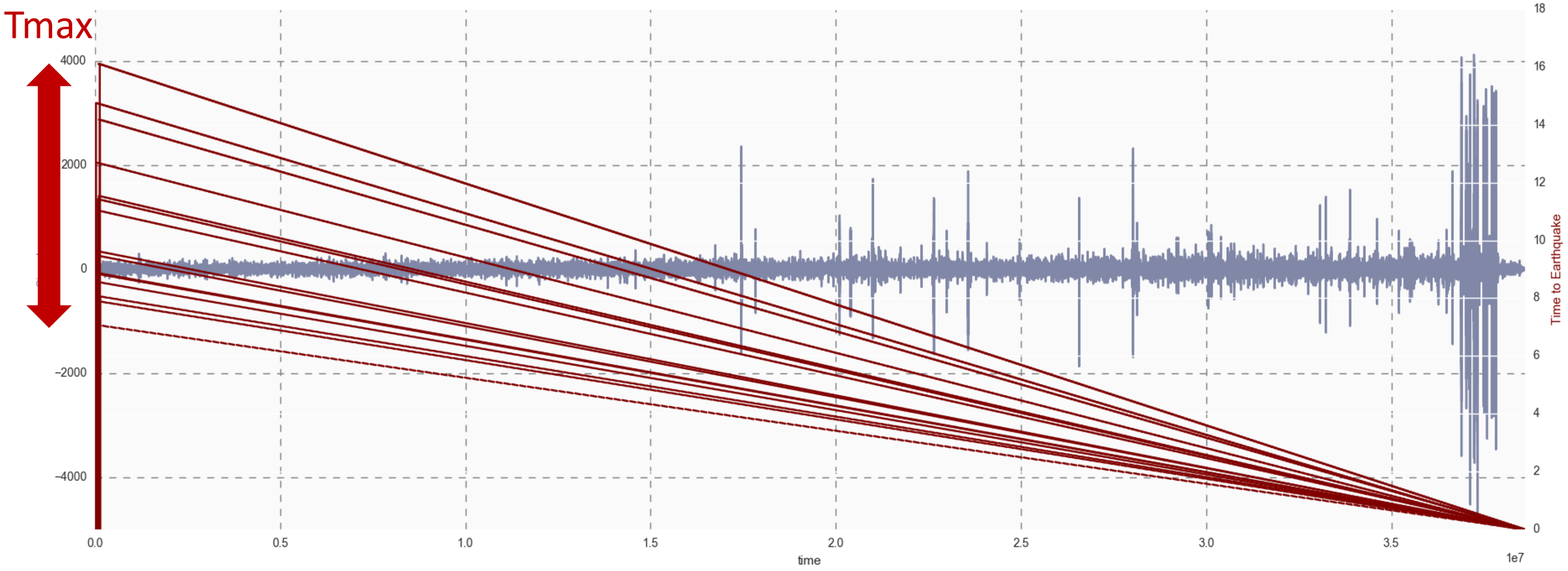
Time to Earthquake: small (red), medium (orange), large (blue)



Normalized Earthquakes
Signal & normalized TIME



Normalized Earthquakes Signal & TIME



Selecting best descriptors

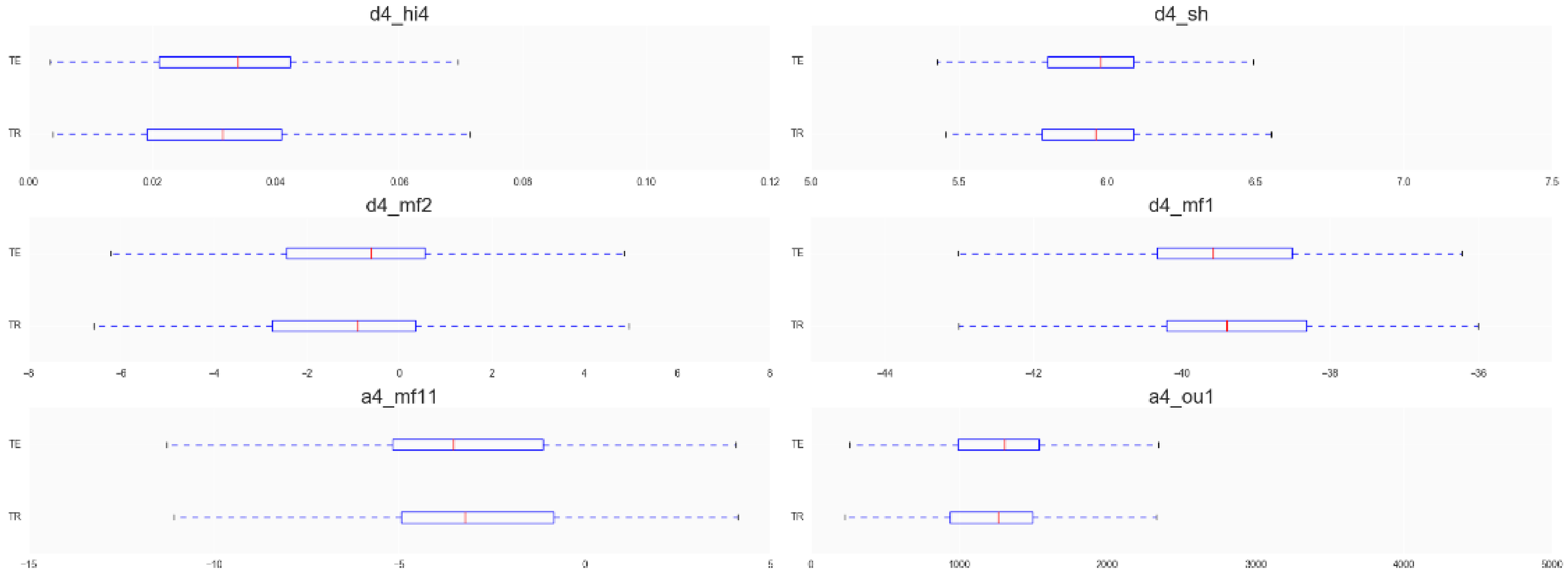
name	*Pearson*	*Spearman*	*descriptor*
d4_hi4	**0.66**	**0.68**	**Number of values in [10,20]**
d4_le	**0.65**	**0.68**	**Log energy**
d4_sh	**0.64**	**0.68**	**Shannon's entropy**
d4_hi1	**0.62**	**0.68**	**Number of values in [-100,-10]**
a4_sh	**0.64**	**0.67**	**Shannon's entropy**
d4_hi3	**0.61**	**0.67**	**Number of values in [0,10]**
d4_mf0	**0.65**	**0.67**	**0th MFCC**
a4_mf0	**0.65**	**0.67**	**0th MFCC**
d4_hi2	**0.60**	**0.67**	**Number of values in [-10,0]**
d4_hi5	0.55	0.65	
d4_mf2	0.64	0.65	
a4_hi2	0.62	0.65	
a4_mf1	0.63	0.64	
d4_mf1	0.62	0.64	
a4_hi5	0.61	0.63	
a4_hi1	0.54	0.63	
a4_mf11	0.62	0.63	
d4_ou1	**0.58**	**0.62**	**Number of outliers**
a4_ou1	**0.59**	**0.62**	**Number of outliers**
a4_hi6	0.49	0.61	

Inter-correlated descriptors!

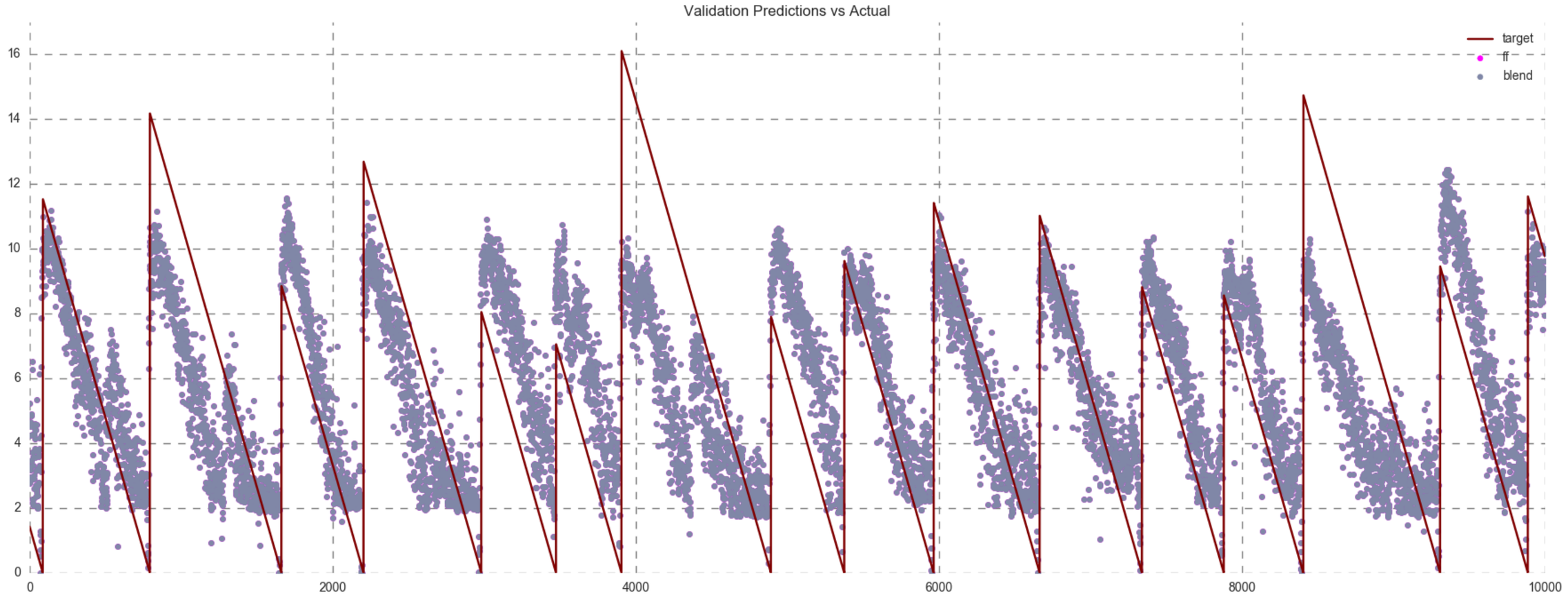
Selecting best descriptors: groups

group of inter-correlated >0.98	*pearson*	*spearman*
d4_hi4, d4_le	**0.66**	**0.68**
d4_sh, d4_hi1, a4_sh, d4_mf0, a4_mf0	**0.64**	**0.68**
d4_hi3	0.61	0.67
d4_hi2	0.60	0.67
d4_hi5, a4_hi1	0.55	0.65
d4_mf2	**0.64**	**0.65**
a4_hi2	0.62	0.65
a4_mf1	0.63	0.64
d4_mf1	**0.62**	**0.64**
a4_hi5	0.61	0.63
a4_mf11	**0.62**	**0.63**
d4_ou1	0.58	0.62
a4_ou1	**0.59**	**0.62**
a4_hi6	0.49	0.61
a4_hi4	0.56	0.59

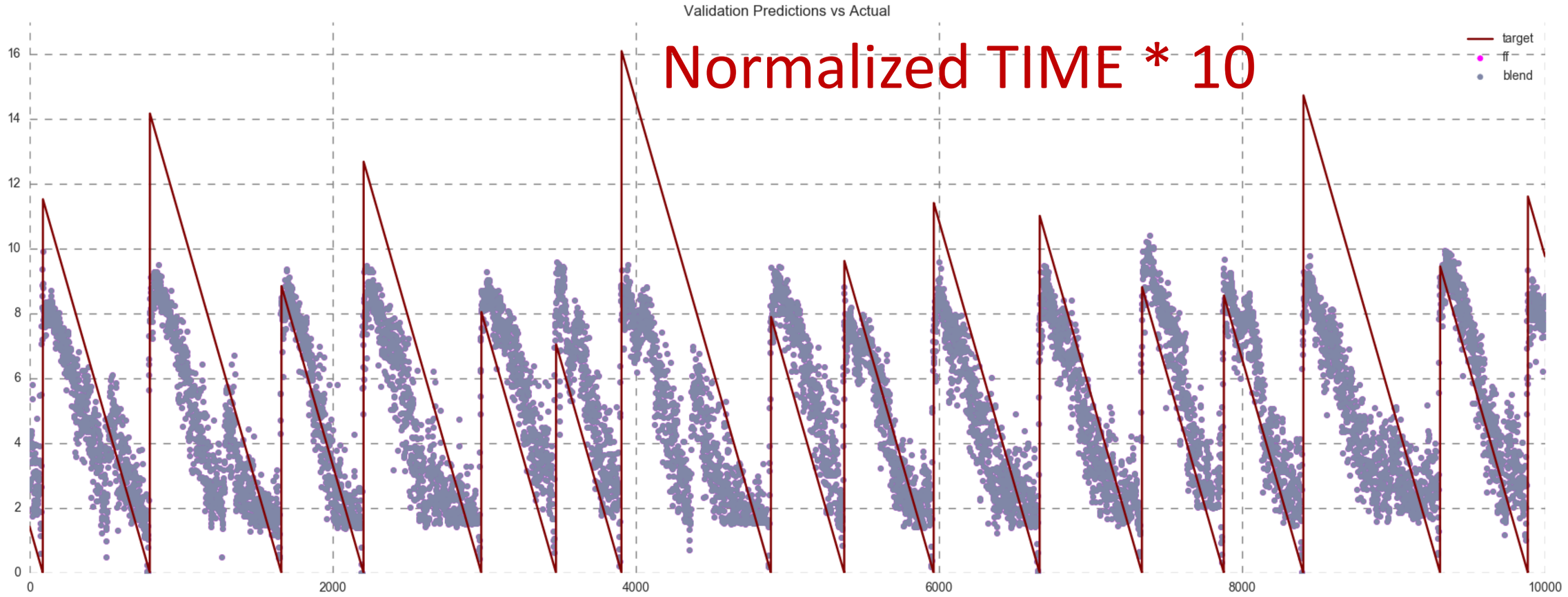
Training and TEST data distribution



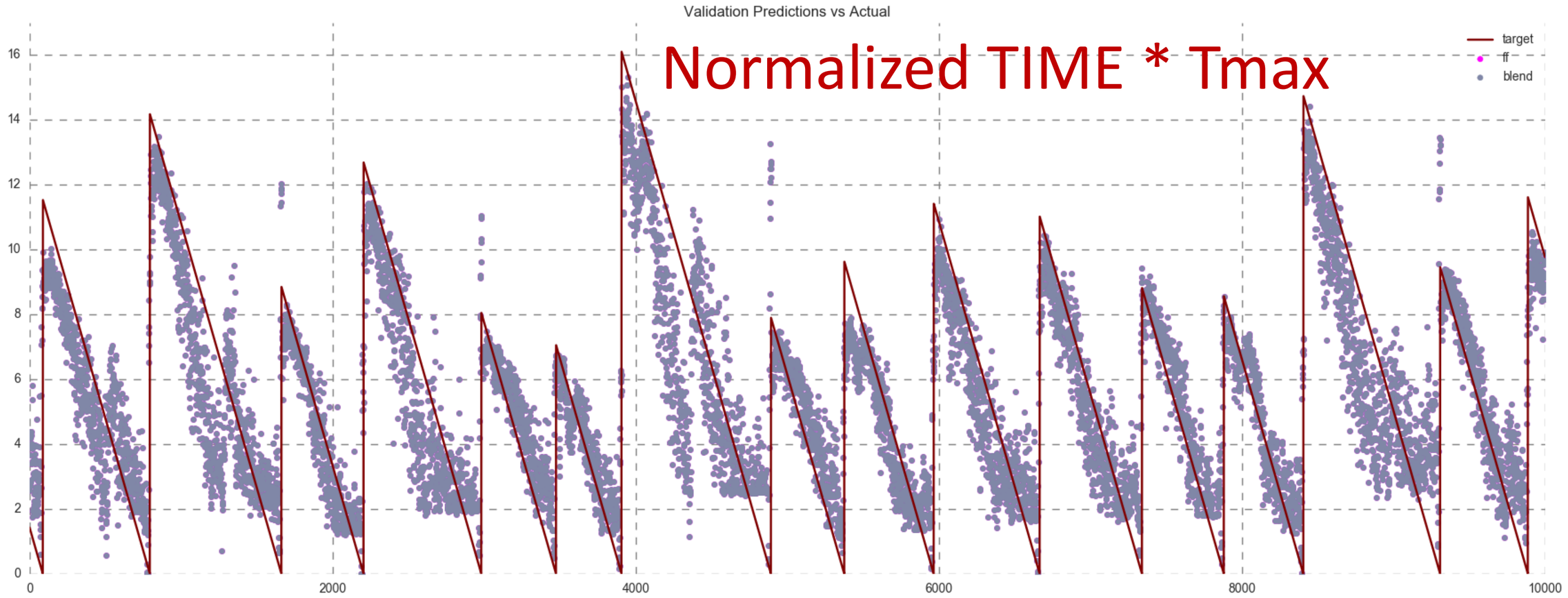
10,000 random training points
FF CV MAE=2.17, Kaggle Private MAE=2.43



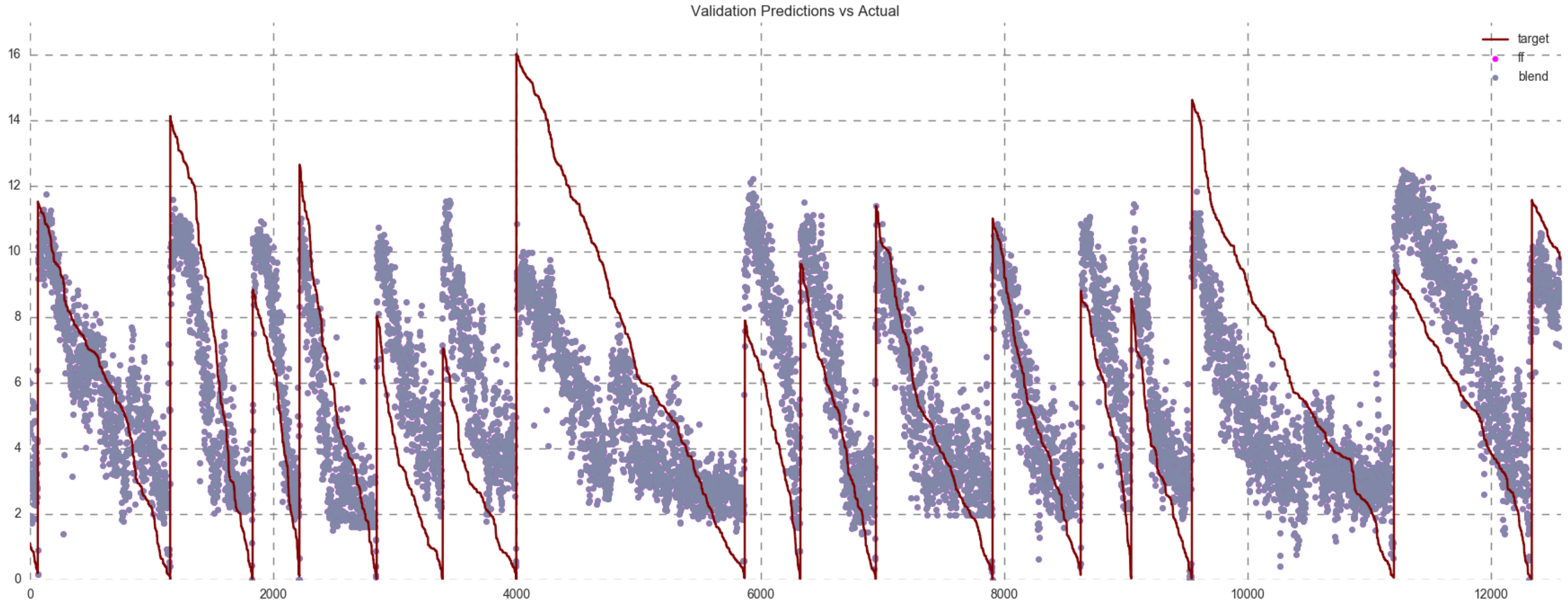
10,000 random training points
FF CV MAE=2.15, Kaggle Private MAE=2.44



10,000 random training points
FF CV MAE=1.35

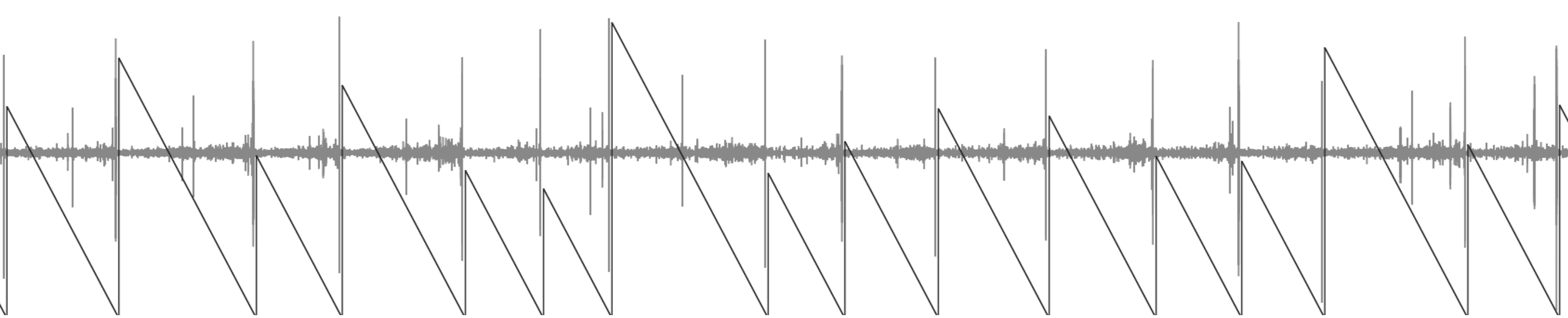


~12,000 nearest neighbors (K=5) for TEST points
FF CV MAE=2.24, Kaggle Private MAE=2.42



Study Summary & Future Research

- Computed ~100 descriptors for each signal, its FFT and wavelet (A,D) and selected ~10 groups of descriptors with correlation above 0.65…0.7 with TIME
  - Shannon's entropy, log energy, 0^{th}, 1^{st}, 11^{th} MFCCs, number of observations in [10,20] (all computed for A_4 and D_4), etc.
  - Other descriptors?
- Built KNN
  - Kaggle Private MAE=2.43 (in Kaggle Top 50)
  - Used to select training data for FF
- Built and cross-validated LR, RCV, FF, SVR, XGB, LGB
  - Kaggle Private MAE=2.42 (in Kaggle Top 30)
  - Normalized TIME is easier to predict
  - Need Tmax to predict TIME?
- Used ~60,000 training samples (0.01%)
  - More samples?



Shaking Earth: LANL Earthquake Prediction Challenge

Nelli Fedorova

TIME < 7.06

Long earthquakes are still visible!