HEX UCR Anomaly Benchmark Datasets 2021

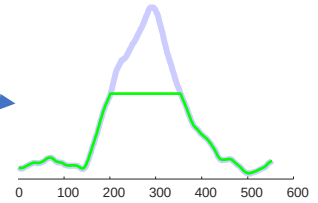
Dataset Design Principles I

- **Remove the threshold question**. Anomaly detection typically consists of two parts. A) Find the candidate region(s) that *might be* anomalies. B) Test if, under your model, you should flag these as anomalies. Here we remove 'B', by telling the world that there is *exactly one anomaly* in the test data. We do this because 'B' can depend on external factors (misclassification costs etc.), and we think that in most domains, if you can do 'A' robustly, 'B' will be easy.

- **Try to have diverse datasets**.

- **Use real data only in the case** you can be sure the anomaly is the *only* (or by a large margin, *most significant*) anomaly.

- **For synthetic data, model something in the real world** (when possible). For example:
  - This anomaly models the actor falling down.
  - This anomaly models a terrorist attack in Melbourne, and the police blocking the road for an hour.
  - This anomaly models a nurse placing her hand under the respiration strap.
- **Avoid Goldilocks**. For at least some problems, make multiple versions….
  - One that is obvious, probably *any* algorithm can find them.
  - One that is more subtle
  - One that is very very subtle, probably *no* algorithm can find them.



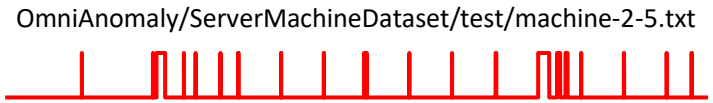
This anomaly models a nurse placing her hand under the respiration strap

`UCR_Anomaly_resperation1_100000_110260_110412.txt`

- **Implication of the above**: It is virtually certain that 100% accuracy is not possible.

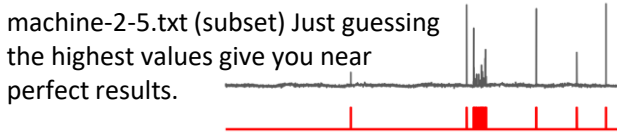
Dataset Design Principles II

- **Only one anomaly per dataset**. Some datasets (see machine-2-5.txt) have so many anomalies that it confuses scoring metrics.



- **Anomalies can appear anywhere**. Some datasets (esp. Yahoo) have most of the anomalies at the end (run-to-failure). But this is an unfair clue to give algorithms.

- **Anomalies should not (in general) be the highest/lowest values**. Many datasets (Yahoo/ServerMachineDataset/NASA) have some problems that are trivial to solve with the "pick the highest value" heuristic.



- **(More general version of previous rule)** Anomalies should not (in general) have very different means/standard deviations/skewness or other simple summary statistics. Such things *can* be anomalies, but we can typically find them with 60-year old change detection algorithms.

- **Avoid always inserting anomalies at round numbers,** say at location 10,000 or 11,500. As that gives a clue to the algorithms.

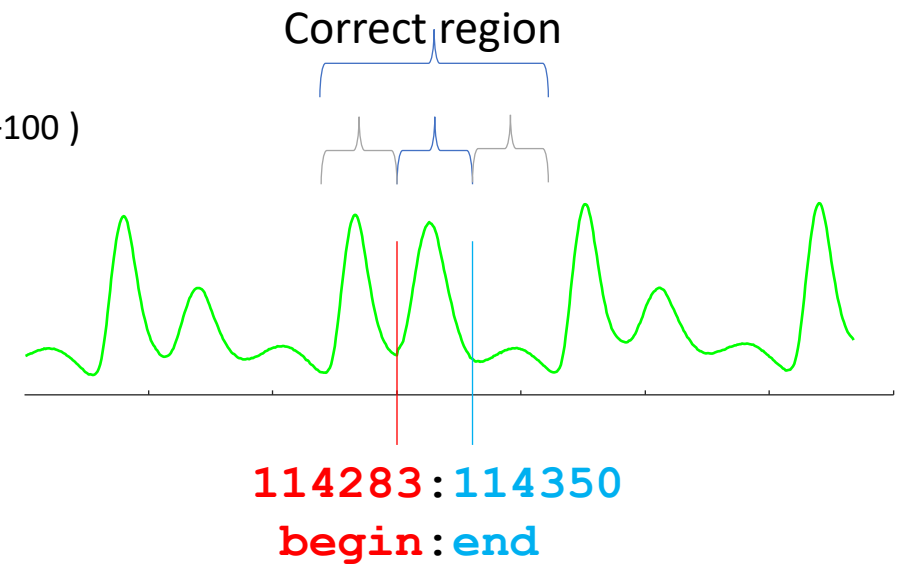
Scoring Function Design Principles I

Avoid complex and opaque scoring functions We want a scoring function that..

- Is a single number, for easy comparisons.
- Does not have spurious location precision. If the ground truth says the anomaly is at say 1250, and an algorithm reports 1247 or 1254, it should be counted as correct. This problem is compound by the fact that different algorithms report the *leading edge*, the *center* or the *trailing edge* of a sliding window.
- Has a binary score for each example, that can be combined to a real number for the full collection.
- Reports a number close to zero for a "random dart" algorithm (i.e. the default rate) and close to one for a perfect algorithm.

My suggestion

- Let length of anomaly be L, $L = \texttt{end} - \texttt{begin} + 1$
- Let the prediction of an algorithm be an integer P
- P is labeled as correct if: $\min(\texttt{begin}-L, \texttt{begin}-100) < P < \max(\texttt{end}+L, \texttt{end}+100)$
- Why the '100' case? Some anomalies can be as short as a single point.



012_UCR_Anomaly_**tiltAPB1**_100000_114283_114350.txt

- Dataset number
- Mnemonic name
- From 1 to **X** is training data
- **Begin** anomaly
- **End** anomaly

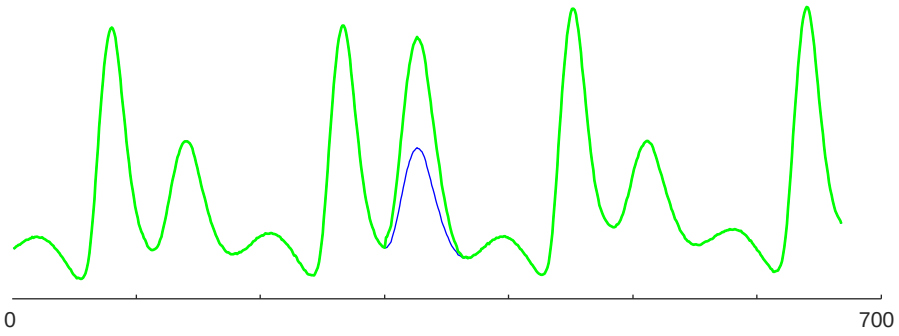
Sample format

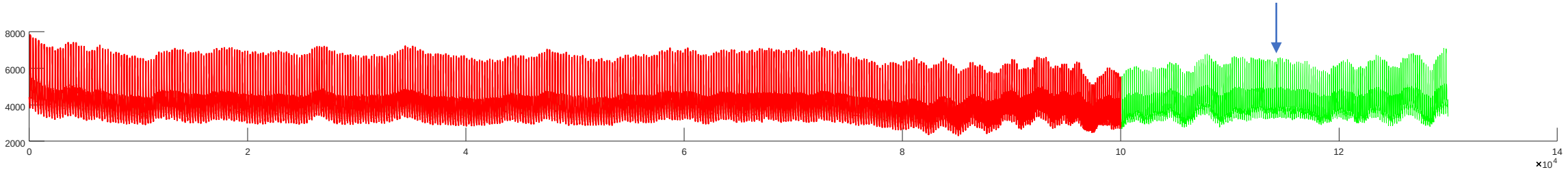
Note the structure of the file names
Files are '-ascii' format.
Note the layout of the plots

The data comes from a healthy male on a tilt table.
At first, he is supine, at around 80000, the table is tilted forward. The trace is his APB.

The anomaly is synthetic. There is a secondary peak after the dicrotic notch. It is normally about half the size of the peak systolic pressure. For one randomly chosen beat, we made it much greater, almost as big as the main peak.

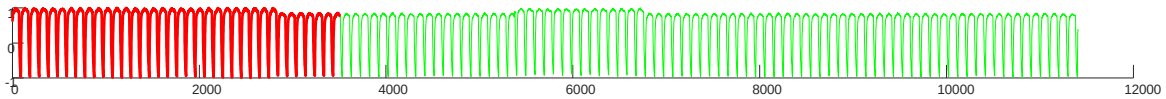
Blue is original data; green is data after anomaly was introduced



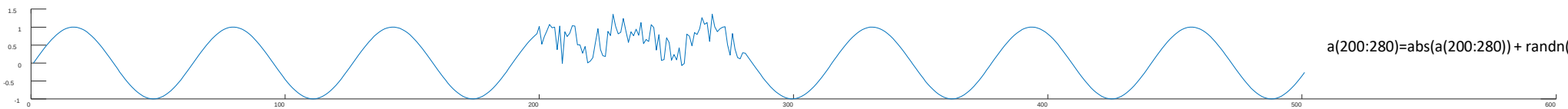


Notes

- Every dataset should be considered completely independent of the other datasets. Any algorithm that considers data from two or more datasets at the same time should be considered cheating.
- However, an algorithm *could* use external data. For example
  - An algorithm could “realize” the dataset is an ECG (without human intervention, obviously), it could then search the web for similar ECGs, and use that information to help build the model.
  - Note, we do not think this is likely to help, but it is within the spirt of fair evaluation
- You can assume:
  - The training data is free of anomalies
  - The test data has only one anomaly. Or equivalently, if it has more than one, exactly one is much much more obvious/significant than the others.



- You should **predict the center** of the anomaly:
  - Some algorithms, by default, output the beginning, the middle, or the end of the anomaly. A good scoring function will have some “wiggle room” to allow for this, so this is probably a non-issue.
  - However, it is probably best to output the *center* of where your algorithm thinks the anomaly is. For example, in the below, the anomaly starts at 200, and is of length 80. It would be best to output 240.




```
a(200:280)=abs(a(200:280)) + randn(size(abs(a(200:280))))/4;
```

Known issues

- This will be updated if new issues are brought to our attention

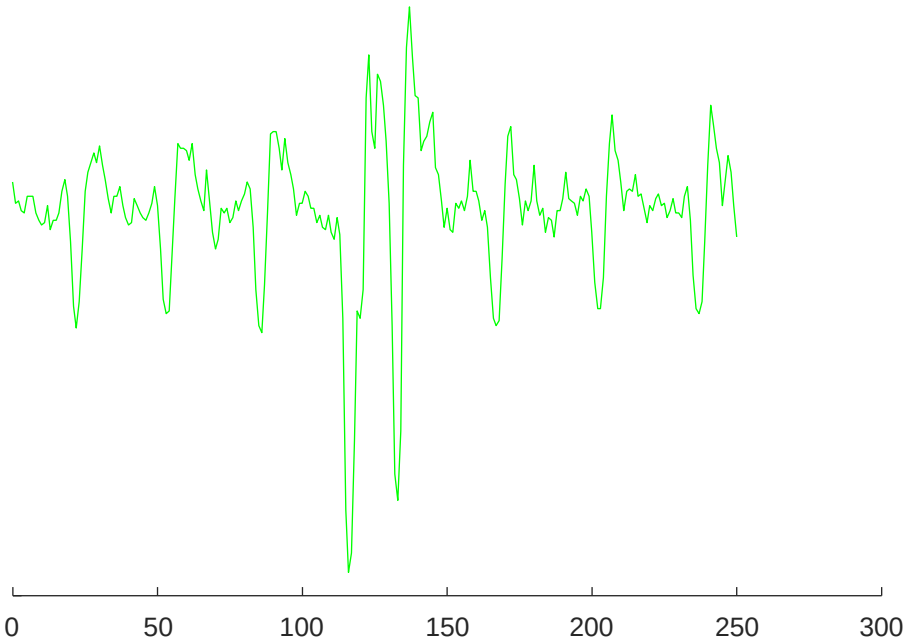
- Some files are formatted differently: tabulator between values instead of line break. This can lead to a problem when reading the files, e.g. when using pandas.read_csv(...). This applies to file number 204, 205, 206, 207, 208, 225, 226, 242 and 243.

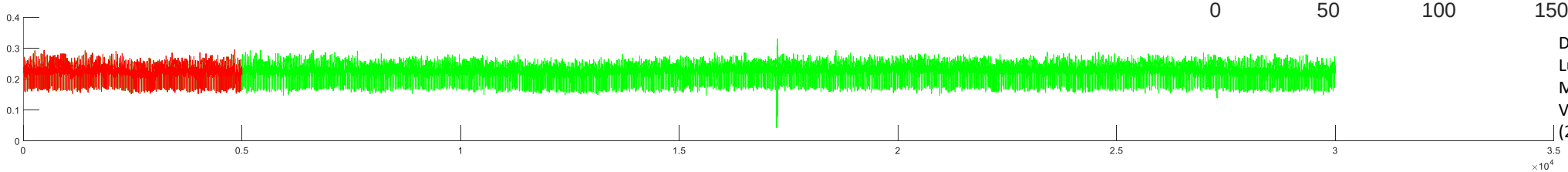
Key to Datasets

UCR_Anomaly_Lab2Cmac011215EPG1_5000_17210_17260.txt

This is a real dataset from an insect, and Asian Citrus Psyllid, recorded using an EPG apparatus.

Here the anomaly is completely natural, probably the insect moved its stylet to a new vein.



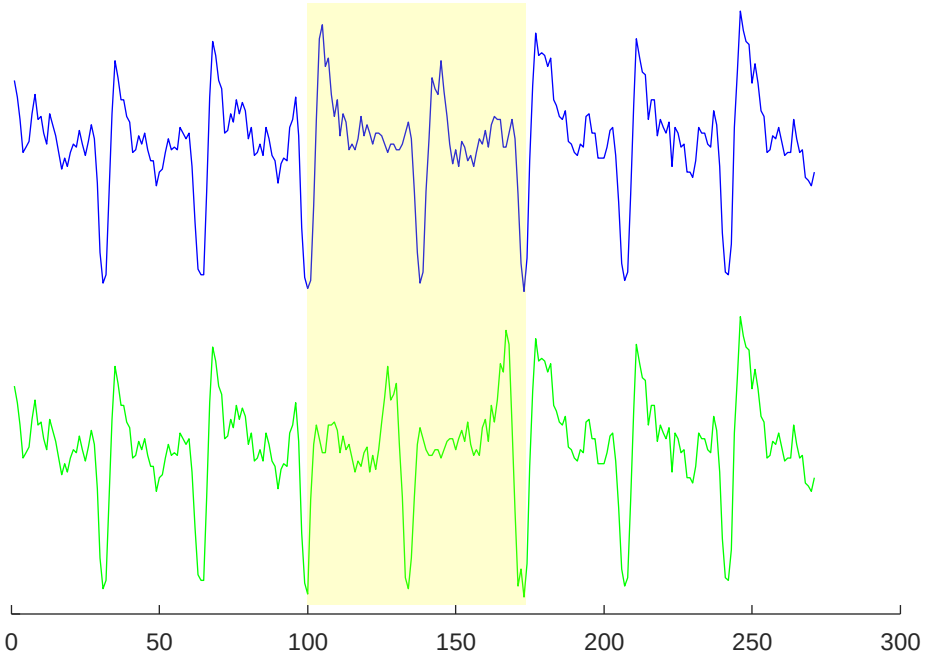


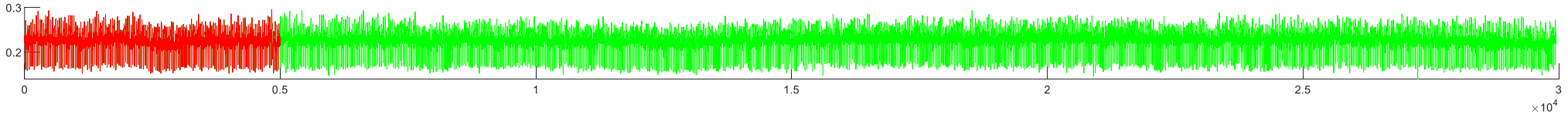
Denis S. Willett, Justin George, Nora S. Willett, Lukasz L. Stelinski, and Stephen L. Lapointe. 2016. Machine Learning for Characterization of Insect Vector Feeding. PLOS Computational Biology (2016).

UCR_Anomaly_Lab2Cmac011215EPG2_5000_27862_27932.txt

This is a real dataset from an insect, and Asian Citrus Psyllid, recorded using an EPG apparatus.

Here we removed the natural anomaly and then reversed the direction of two beats.



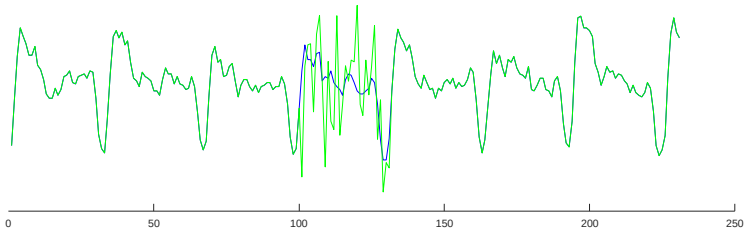


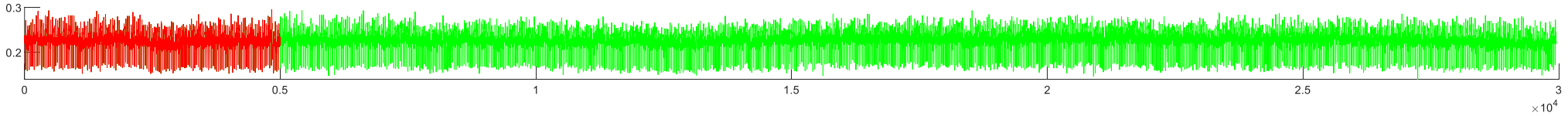
Denis S. Willett, Justin George, Nora S. Willett, Lukasz L. Stelinski, and Stephen L. Lapointe. 2016. Machine Learning for Characterization of Insect Vector Feeding. PLOS Computational Biology (2016).

UCR_Anomaly_Lab2Cmac011215EPG3_5000_16390_16420.txt

This is a real dataset from an insect, and Asian Citrus Psyllid, recorded using an EPG apparatus.

Here we removed the natural anomaly and added a lot of noise to a single beat.





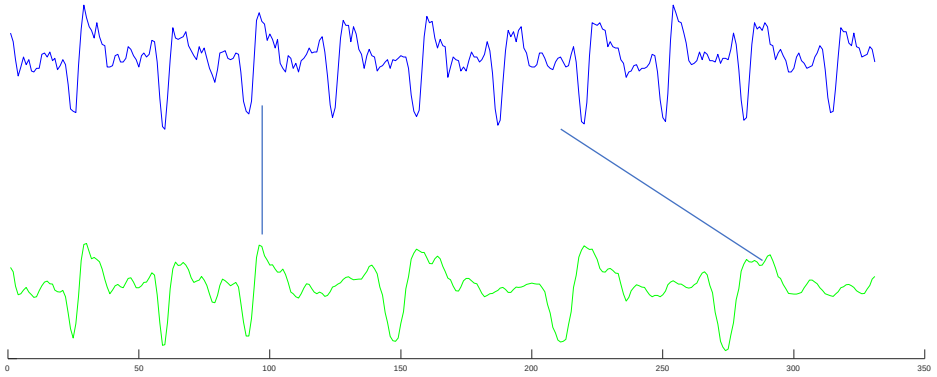
Denis S. Willett, Justin George, Nora S. Willett, Lukasz L. Stelinski, and Stephen L. Lapointe. 2016. Machine Learning for Characterization of Insect Vector Feeding. PLOS Computational Biology (2016).

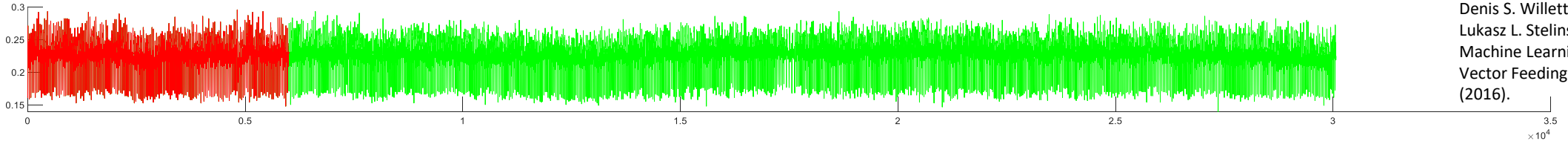
UCR_Anomaly_Lab2Cmac011215EPG4_6000_17390_17520.txt

This is a real dataset from an insect, and Asian Citrus Psyllid, recorded using an EPG apparatus.

Here we removed the natural anomaly and then “slowed down” about three beats with interpolation.

We also truncate a few random points from the beginning, just so this would not align with the other datasets in this series.

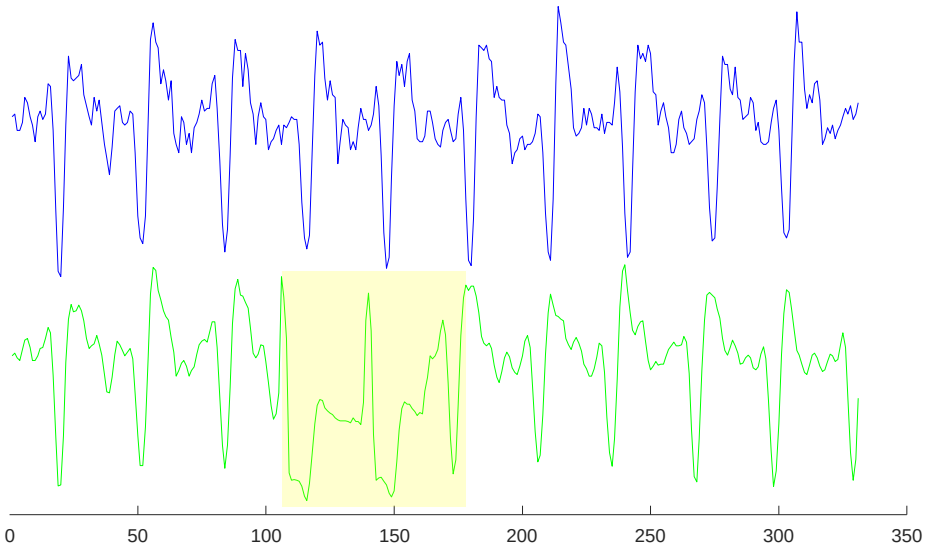


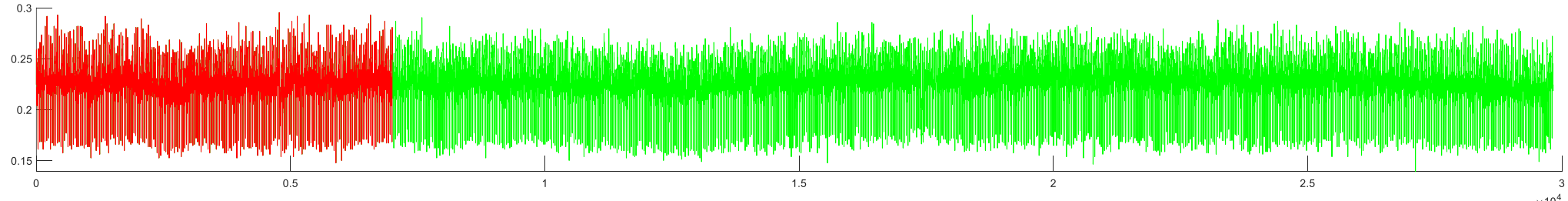


Denis S. Willett, Justin George, Nora S. Willett, Lukasz L. Stelinski, and Stephen L. Lapointe. 2016. Machine Learning for Characterization of Insect Vector Feeding. PLOS Computational Biology (2016).

UCR_Anomaly_Lab2Cmac011215EPG5_7000_17390_17520.txt

This is a real dataset from an insect, and Asian Citrus Psyllid, recorded using an EPG apparatus.
Here we removed the natural anomaly.
We inserted about 1.5 human heartbeats (from UCR_Anomaly_ECG3_8000_17000_17100.txt)
We made a quick effort to edit the human beats to have approximately the same mean, variance and periodicity as the original data, and to smooth the joins.
We also truncate a few random points from the beginning, just so this would not align with the other datasets in this series.

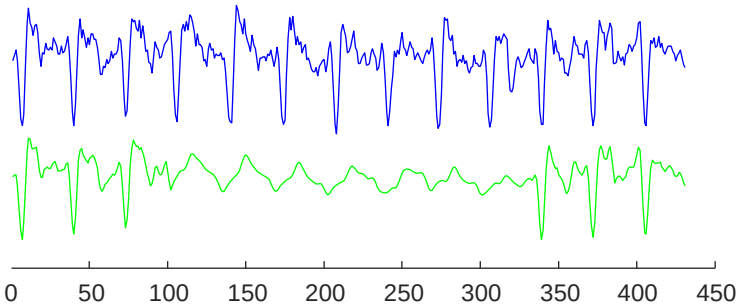


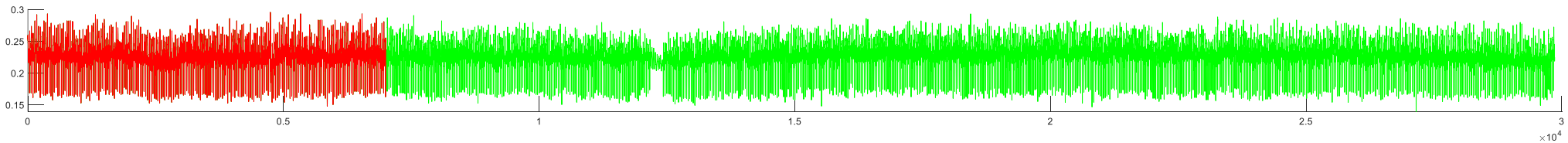


Denis S. Willett, Justin George, Nora S. Willett, Lukasz L. Stelinski, and Stephen L. Lapointe. 2016. Machine Learning for Characterization of Insect Vector Feeding. PLOS Computational Biology (2016).

UCR_Anomaly_Lab2Cmac011215EPG6_7000_12190_12420.txt

This is a real dataset from an insect, and Asian Citrus Psyllid, recorded using an EPG apparatus.
Here we removed the natural anomaly.
We greatly smoothed a section. This both smooths the data, and reduces the variance.
We also truncate a few random points from the beginning, just so this would not align with the other datasets in this series.



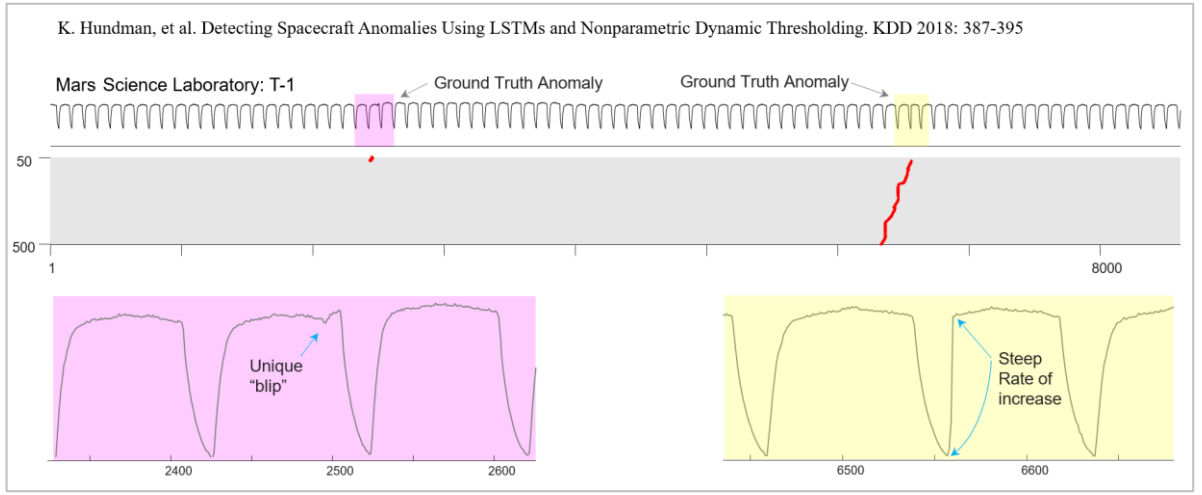


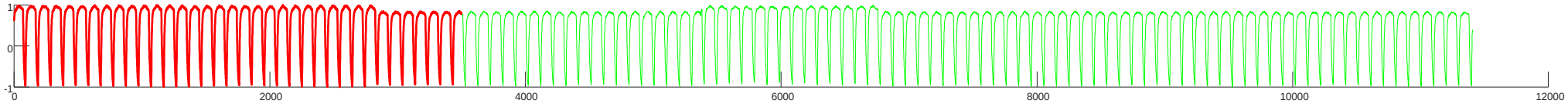
Denis S. Willett, Justin George, Nora S. Willett, Lukasz L. Stelinski, and Stephen L. Lapointe. 2016. Machine Learning for Characterization of Insect Vector Feeding. PLOS Computational Biology (2016).

UCR_Anomaly_TkeepFirstMARS_3500_5365_5380.txt

This is a real dataset from NASA spacecraft, that appeared in a KDD 2018 paper.
We joined the train and test sets.
The original dataset had two anomalies.
We carefully edited the data, into two datasets, each with *one* anomaly.
This one contains “unique blip”.
There is also a small amount of level change in this datasets.
However, note that it occurs in both the train and test sets, so it is not the anomaly.

This figure is from the original test split

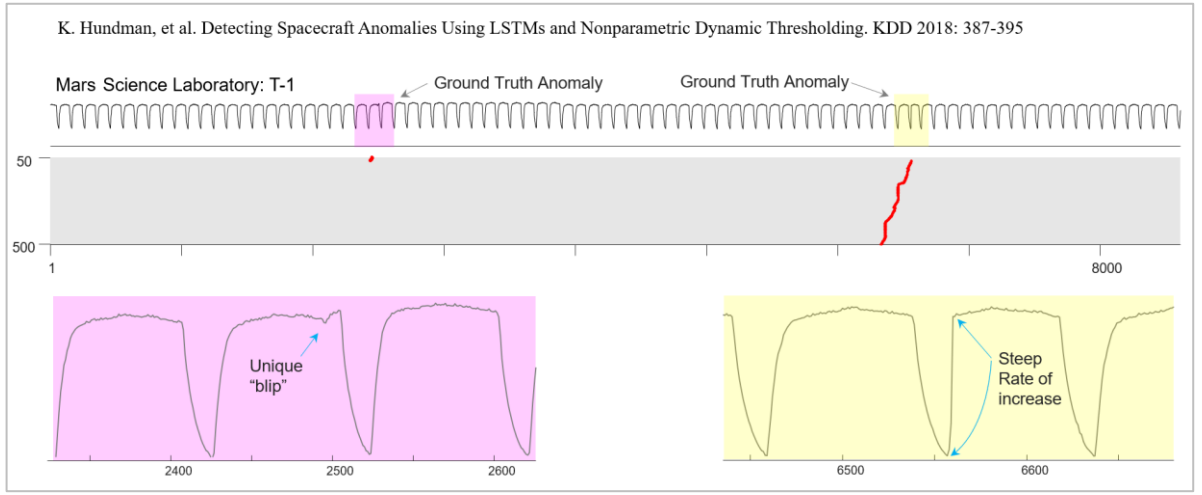


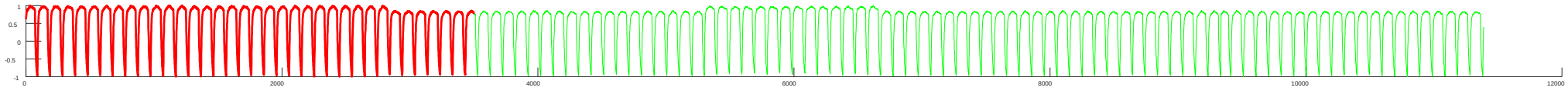


UCR_Anomaly_TkeepSecondMARS_3500_9330_9340.txt

This is a real dataset from NASA spacecraft, that appeared in a KDD 2018 paper.
We joined the train and test sets.
The original dataset had two anomalies.
We carefully edited the data, into two datasets, each with *one* anomaly.
This one contains "steep rate of increase".
There is also a small amount of level change in this datasets.
However, note that it occurs in both the train and test sets, so it is not the anomaly.

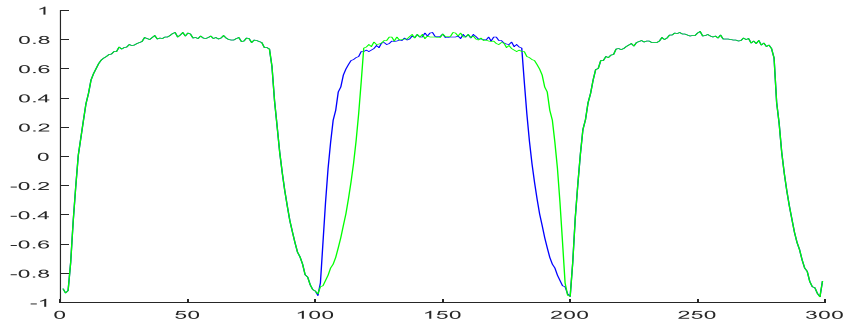
This figure is from the original test split

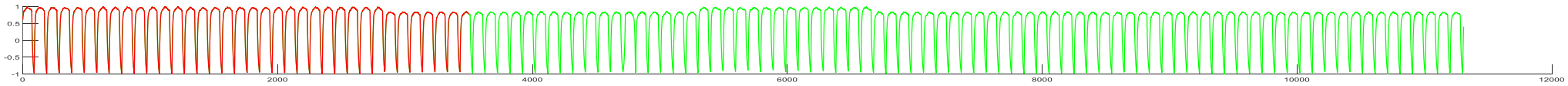




UCR_Anomaly_TkeepThirdMARS_3500_4711_4809.txt

This is a real dataset from NASA spacecraft, that appeared in a KDD 2018 paper.
We joined the train and test sets.
The original dataset had two anomalies.
We carefully removed both the anomalies, and added one of our own.
We flipped one cycle backwards
There is also a small amount of level change in this datasets.
However, note that it occurs in both the train and test sets, so it is not the anomaly.

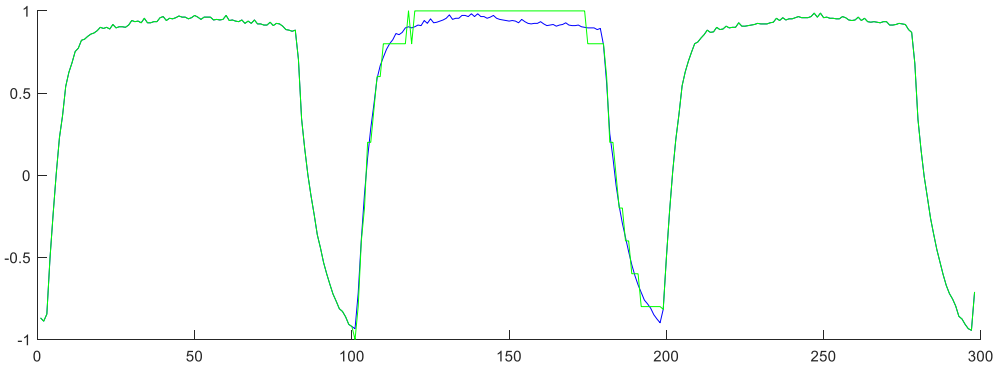


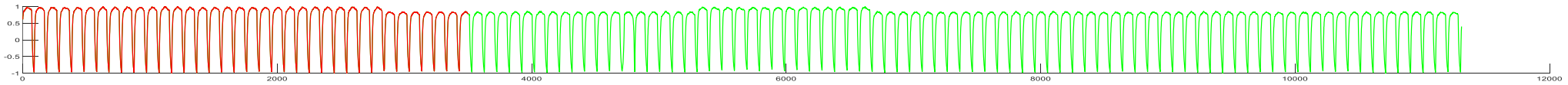


UCR_Anomaly_TkeepForthMARS_3500_5988_6085.txt

This is a real dataset from NASA spacecraft, that appeared in a KDD 2018 paper.
We joined the train and test sets.
The original dataset had two anomalies.
We carefully removed the anomalies, and added one of our own.
We reduced the bit dept of one cycle.
There is also a small amount of level change in this datasets.
However, note that it occurs in both the train and test sets, so it is not the anomaly.

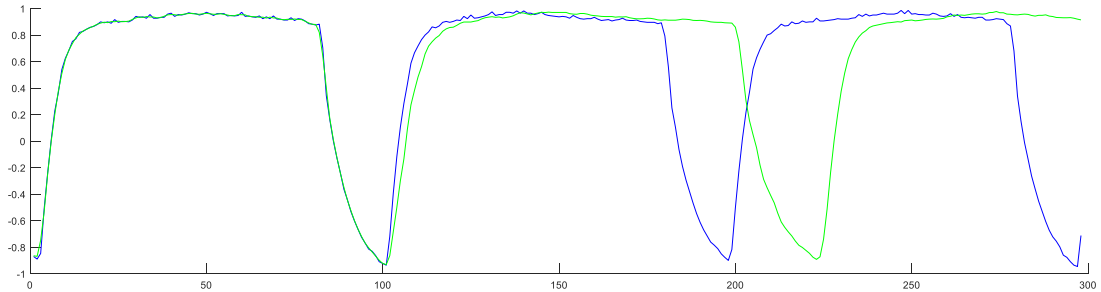
```
sub = T(start_anomaly:end_anomaly  )  ;
sub = sub*5;
sub = round(sub);
sub = sub /5;
T(start_anomaly:end_anomaly  ) = sub;
```

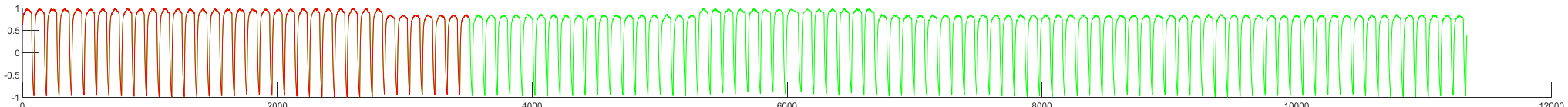




UCR_Anomaly_TkeepFifthMARS_3500_5988_6085.txt

This is a real dataset from NASA spacecraft, that appeared in a KDD 2018 paper.
We joined the train and test sets.
The original dataset had two anomalies.
We carefully removed the anomalies, and added one of our own.
We made one beat about 25% slower than the rest
There is also a small amount of level change in this datasets.
However, note that it occurs in both the train and test sets, so it is not the anomaly.





UCR_Anomaly_gaitHunt1_18500_33070_33180.txt

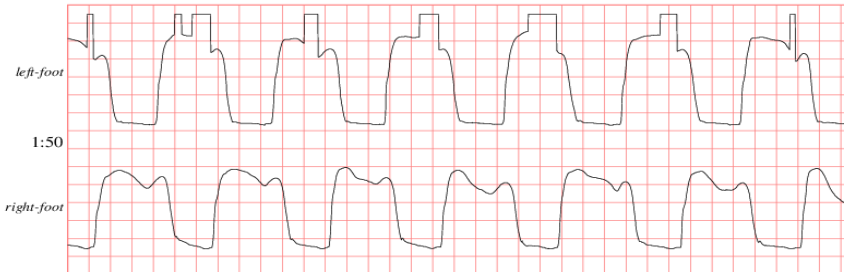
Selected input: record gaitndd/hunt17 , from 0:00.000 to 5:00.000

Gait in Neurodegenerative Disease Database (gaitndd)

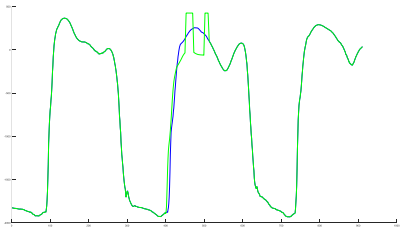
This dataset comes from someone walking on a force plate in a biomechanics lab.
The individual had a mostly symmetric gait, however, after some time, the left foot sensor developed a fault.

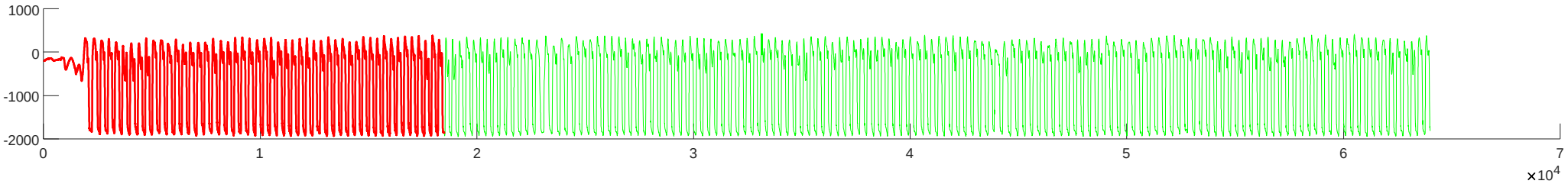
This allows us to create an almost 100% natural dataset. We simply took faulty data from the left foot, and used it to replace some right foot data. We shifted it by a half cycle, slightly smoothing it to remove cut and paste artifacts, and reduced it amplitude so it was not necessarily the tallest peak.

Fault in left sensor



Blue is original data, green is data after anomaly was introduced





UCR_Anomaly_gaitHunt2_18500_31200_31850.txt

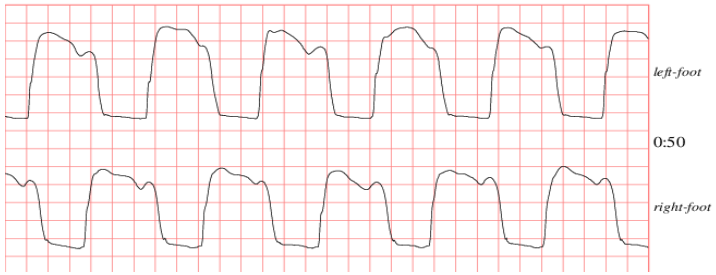
Selected input: record gaitndd/hunt17 , from 0:00.000 to 5:00.000

Gait in Neurodegenerative Disease Database (gaitndd)

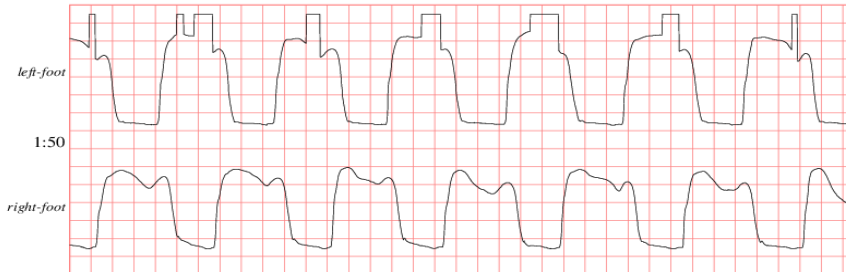
This dataset comes from someone walking on a force plate in a biomechanics lab.
The individual had a mostly symmetric gait, however, after some time, the left foot sensor developed a fault.

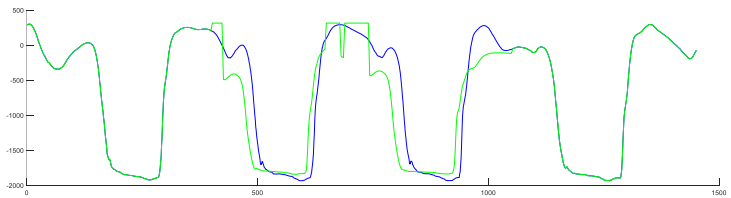
This allows us to create an almost 100% natural dataset. We simply took faulty data from the left foot, and used it to replace some right foot data. We shifted it by a half cycle, slightly smoothing it to remove cut and paste artifacts, and reduced it amplitude so it was not necessarily the tallest peak .

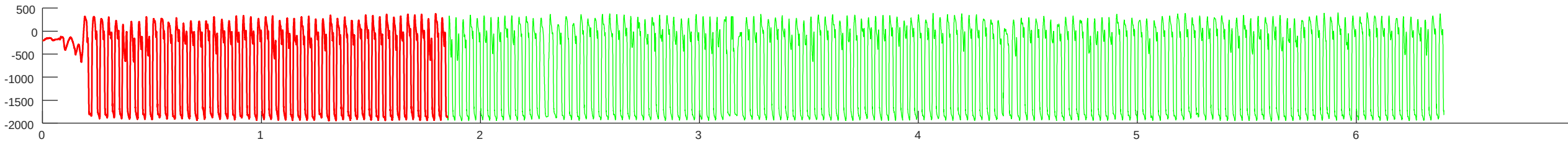
First minute or so…



Fault in left sensor







UCR_Anomaly_gaitHunt3_23400_38400_39200.txt

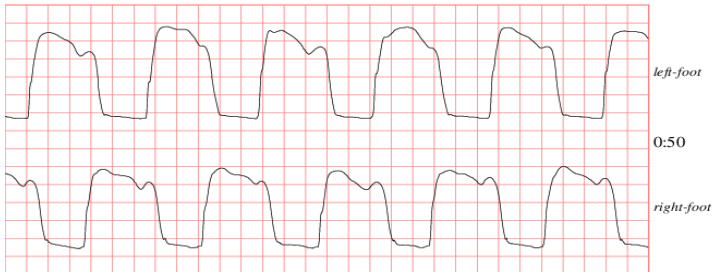
Selected input: record gaitndd/hunt17 , from 0:00.000 to 5:00.000

Gait in Neurodegenerative Disease Database (gaitndd)

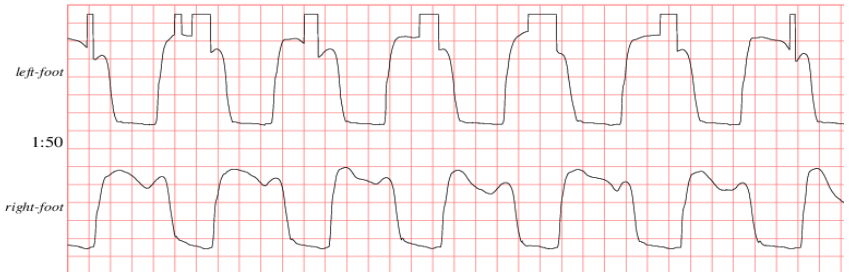
This dataset comes from someone walking on a force plate in a biomechanics lab.
The individual had a mostly symmetric gait, however, after some time, the left foot sensor developed a fault.

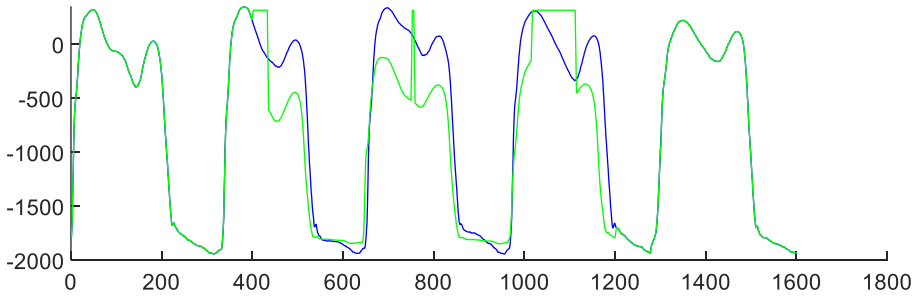
This allows us to create an almost 100% natural dataset. We simply took faulty data from the left foot, and used it to replace some right foot data. We shifted it by a half cycle, slightly smoothing it to remove cut and paste artifacts, and reduced it amplitude so it was not necessarily the tallest peak .

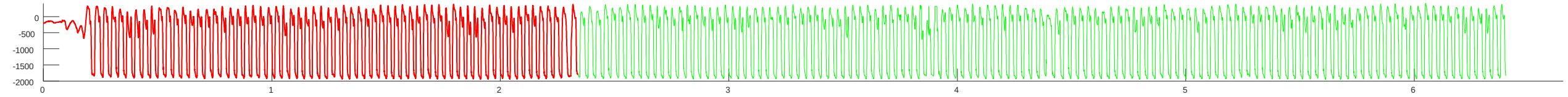
First minute or so…



Fault in left sensor







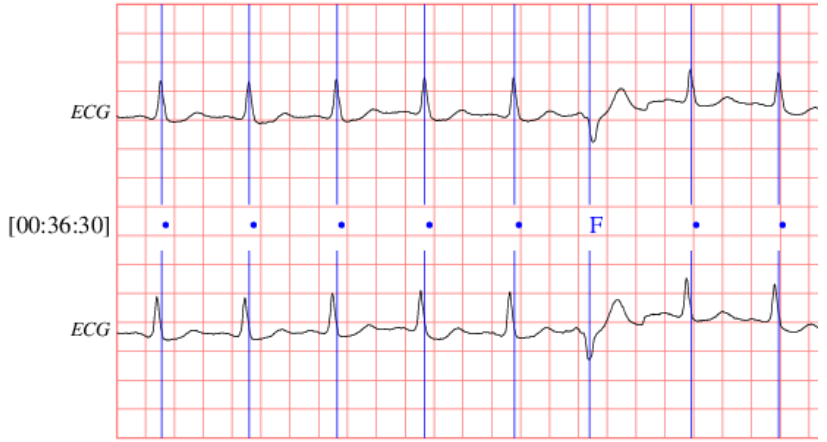
UCR_Anomaly_sddbECG_20000_33270_33400.txt

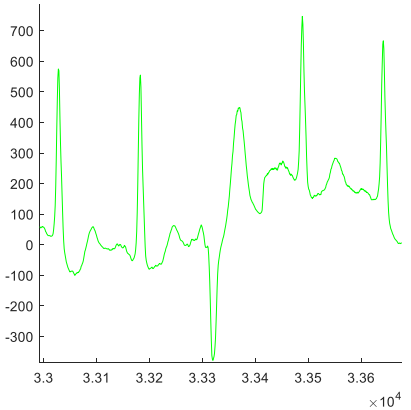
Selected input: record sddb/30 , annotator atr , from [00:34:20.000 1] to [01:34:20.000 1]

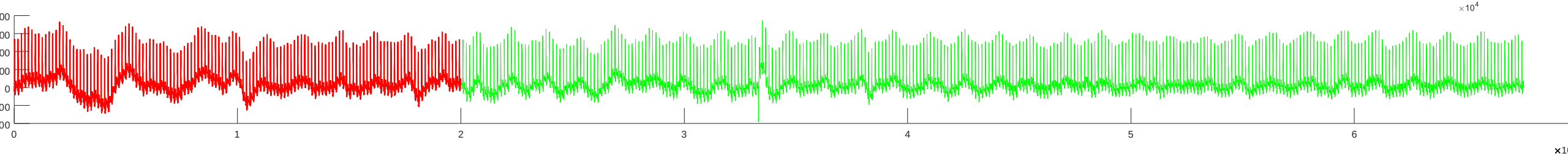
Sudden Cardiac Death Holter Database (sddb)

This is a completely natural dataset. We found a long stretch of ECG that had a single anomaly (F = Fusion of ventricular and normal beat).

The data was *wandering baseline*, which I suspect will confuse many algorithms.







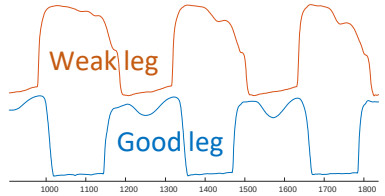
UCR_Anomaly_gait1_20000_38500_38800.txt

Selected input: record gaitndd/hunt14 , from 0:00.000 to 5:00.000

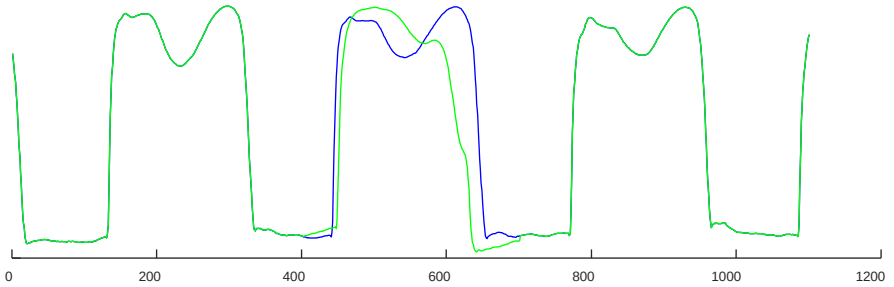
Gait in Neurodegenerative Disease Database (gaitndd)

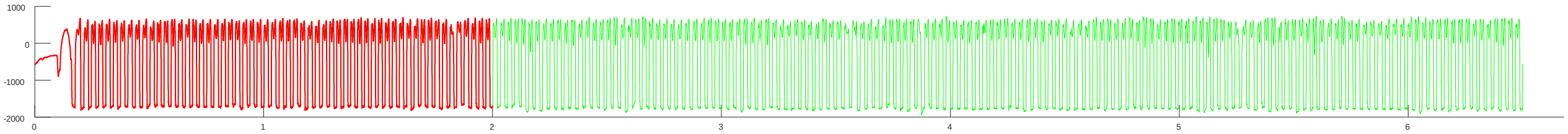
This dataset comes from someone walking on a force plate in a biomechanics lab.
The individual had Huntington's disease, and a highly asymmetric gait (a limp).
The dataset is from the good leg (his right)
The anomaly is that at a random location we swapped in the weak leg (shifting it by a half cycle, so it lines up correctly) for one cycle.

This models an individual having a sudden, but thankfully short lived, weakness or pain in their leg.



Blue is original data, green is data after anomaly was introduced



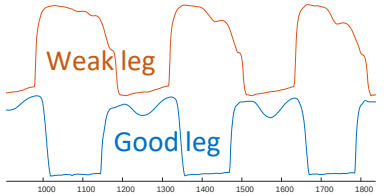


UCR_Anomaly_gait2_22000_46500_46800.txt

Selected input: record gaitndd/hunt14 , from 0:00.000 to 5:00.000

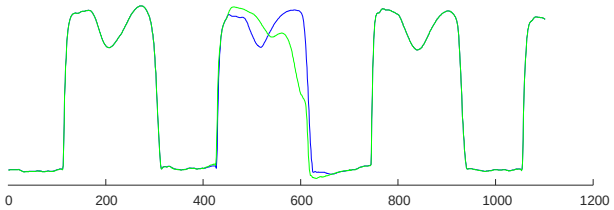
Gait in Neurodegenerative Disease Database (gaitndd)

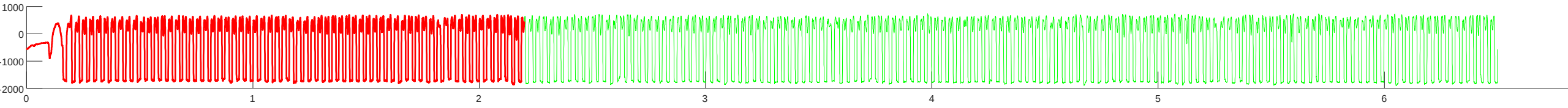
This dataset comes from someone walking on a force plate in a biomechanics lab.
The individual had Huntington's disease, and a highly asymmetric gait (a limp).
The dataset is from the good leg (his right)
The anomaly is that at a random location we swapped in the weak leg (shifting it by a half cycle, so it lines up correctly) for one cycle.



This models an individual having a sudden, but thankfully short lived, weakness or pain in their leg.

(This is very similar to UCR_Anomaly_gait1_20000_38500_38800.txt)





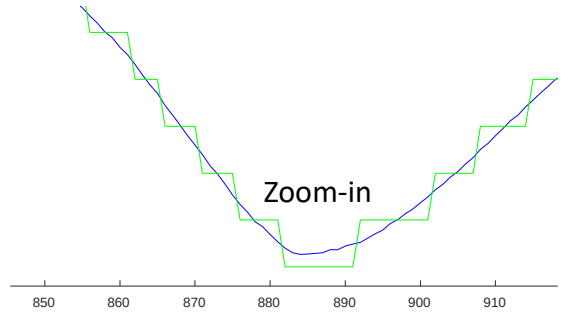
UCR_Anomaly_gait3_24500_59900_60500.txt

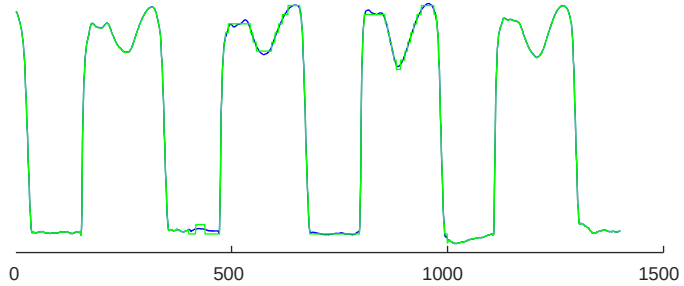
Selected input: record gaitndd/hunt14 , from 0:00.000 to 5:00.000

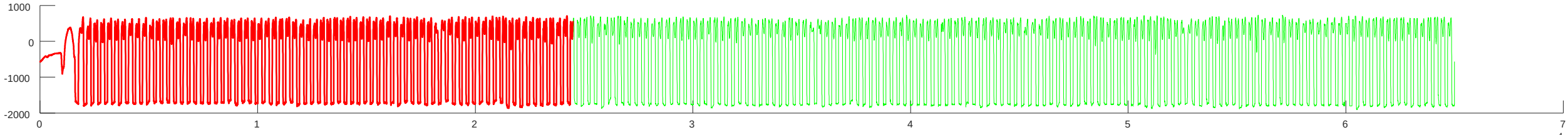
Gait in Neurodegenerative Disease Database (gaitndd)

This dataset comes from someone walking on a force plate in a biomechanics lab.
The individual had Huntington's disease, and a highly asymmetric gait (a limp).
The dataset is from the good leg (his right)
The anomaly is that for two steps, we reduce the bit dept significantly

- Divide by 100
- Round
- Multiple by 100







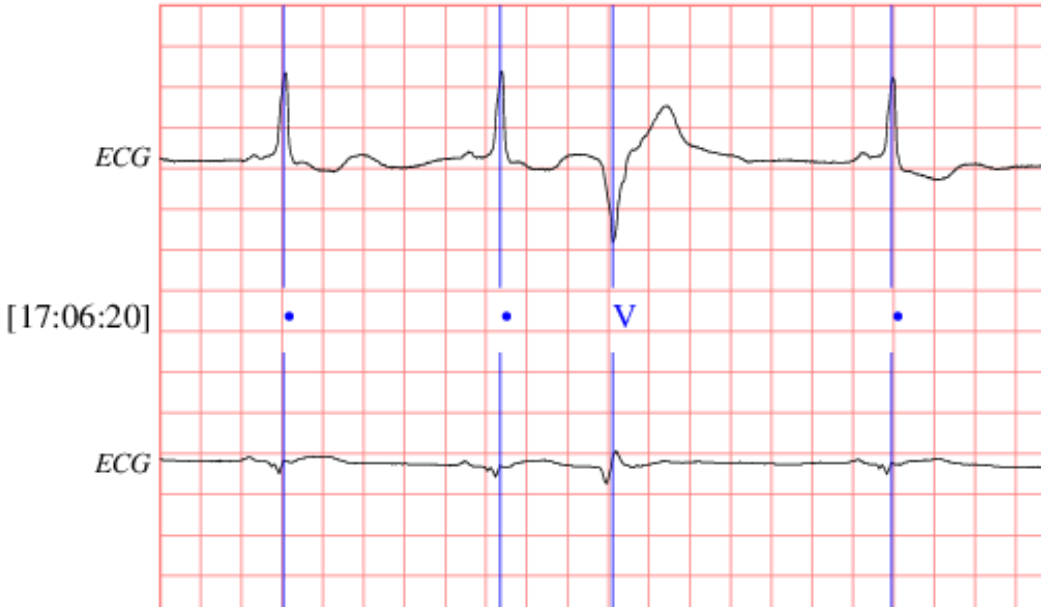
UCR_Anomaly_sddb49_20000_67950_68200.txt

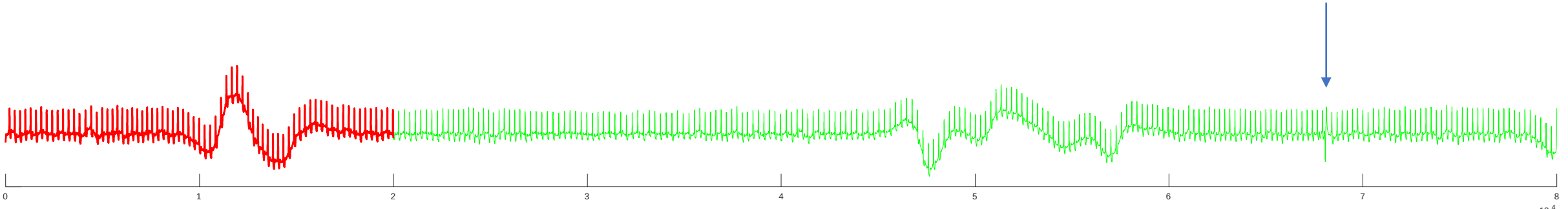
Selected input: record sddb/49 , annotator atr , from [17:01:50.000] to [18:01:50.000]

Sudden Cardiac Death Holter Database (sddb)

This is a completely natural dataset. We found a long stretch of ECG that had a single anomaly (V beat).

The data was *wandering baseline*, which I suspect will confuse many algorithms.





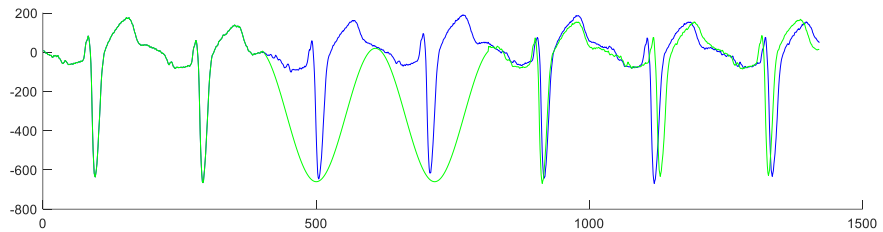
UCR_Anomaly_1sddb40_35000_52000_52620.txt

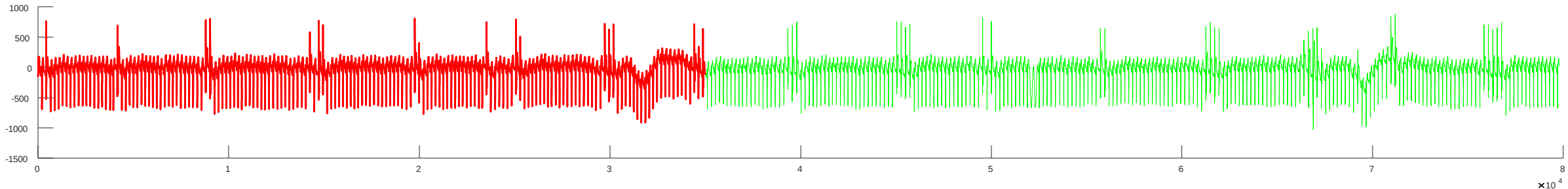
Selected input: record sddb/40 , annotator atr , from [03:27:00.000 1] to [04:27:00.000 1]

Sudden Cardiac Death Holter Database (sddb)

In this dataset there are a mixture of normal beats, and PVC beats (often in mini clusters) and wandering baselines.

We inserted 1.5 periods of a pure sine wave, with about the same mean, standard deviation and period as the real beats.





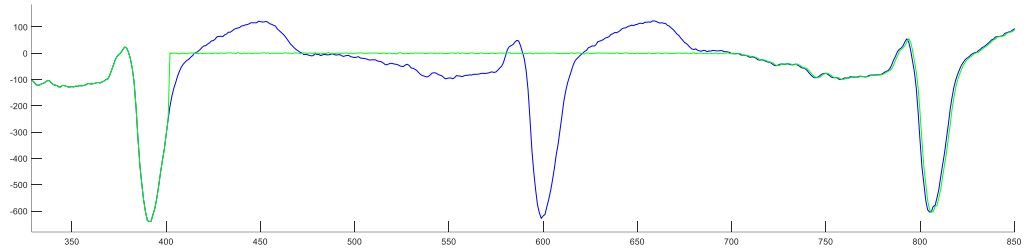
UCR_Anomaly_2sddb40_35000_56600_56900.txt

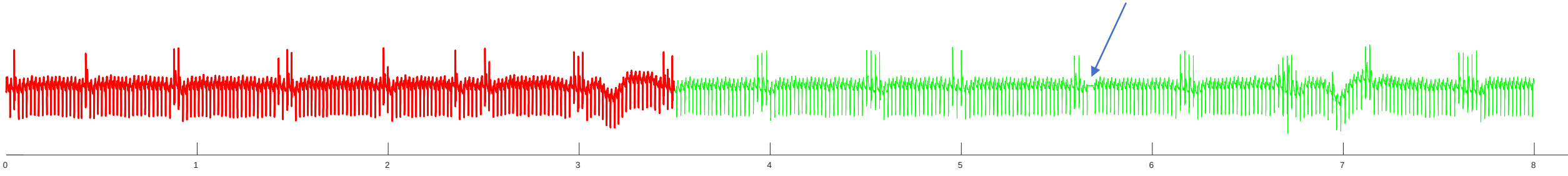
Selected input: record sddb/40 , annotator atr , from [03:27:00.000 1] to [04:27:00.000 1]

Sudden Cardiac Death Holter Database (sddb)

In this dataset there are a mixture of normal beats, and PVC beats (often in mini clusters) and wandering baselines.

We replaced 300 datapoints (about one beat) with a “constant” section of very low amplitude noise.





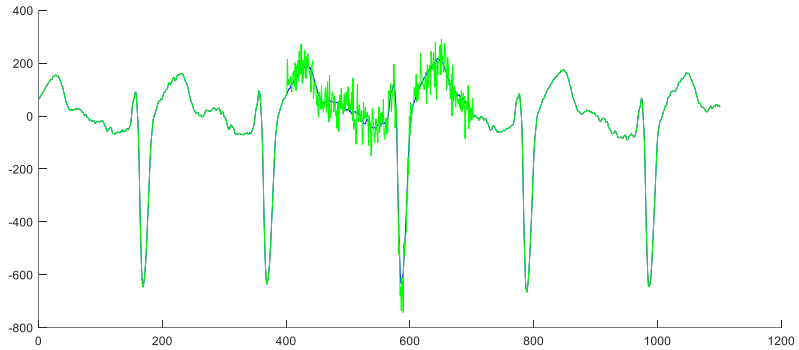
UCR_Anomaly_3sddb40_35000_46600_46900.txt.txt

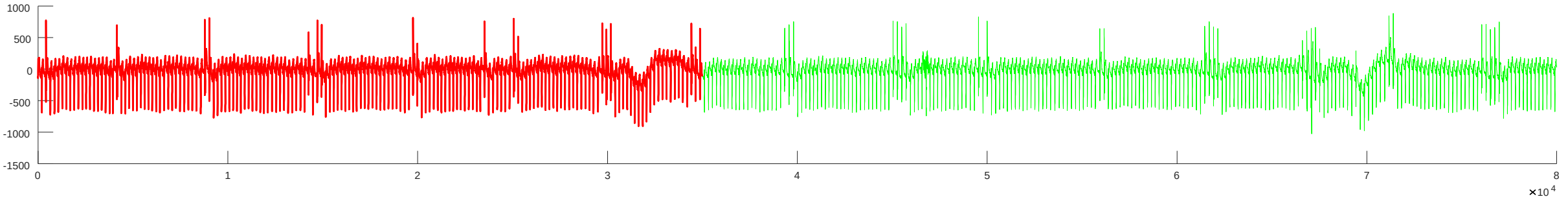
Selected input: record sddb/40 , annotator atr , from [03:27:00.000 1] to [04:27:00.000 1]

Sudden Cardiac Death Holter Database (sddb)

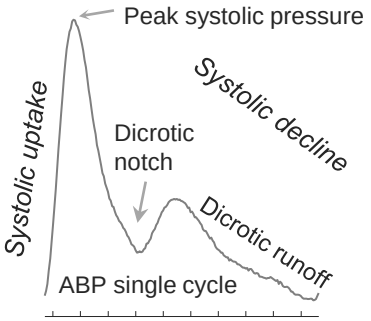
In this dataset there are a mixture of normal beats, and PVC beats (often in mini clusters) and wandering baselines.

We added noise to a region of length 300



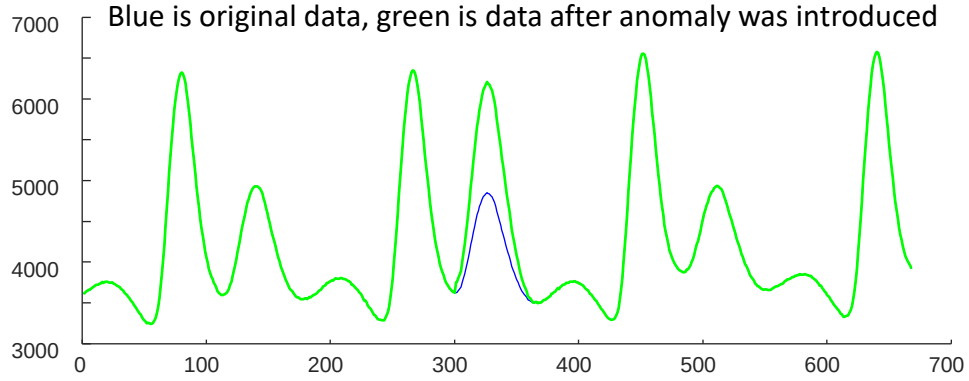


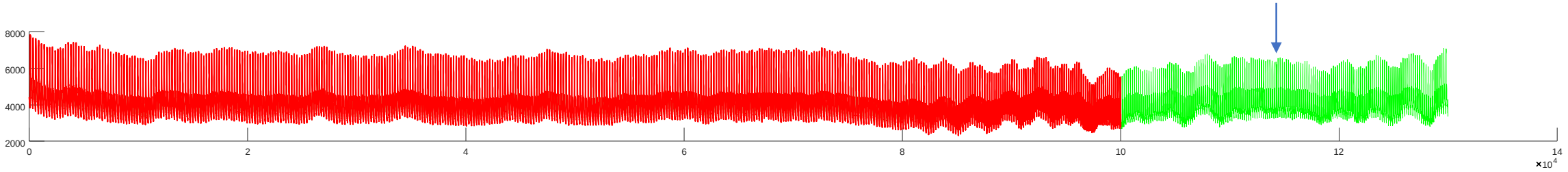
UCR_Anomaly_tiltAPB1_100000_114283_114350.txt



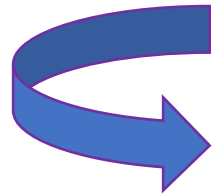
The data comes from a healthy male on a tilt table.
At first, he is supine, at around 80000, the table is tilted forward. The trace is his APB.

The anomaly is synthetic. There is a secondary peak after the dicrotic notch. It is normally about half the size of the peak systolic pressure. For one randomly chosen beat, we made it much greater, almost as big as the main peak.



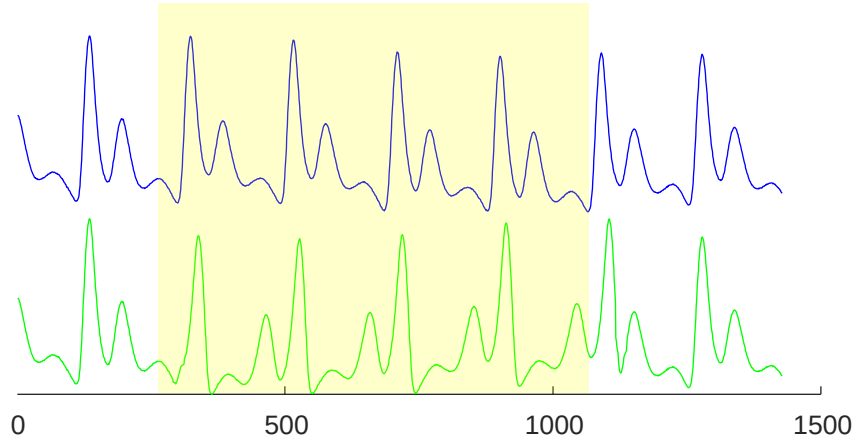


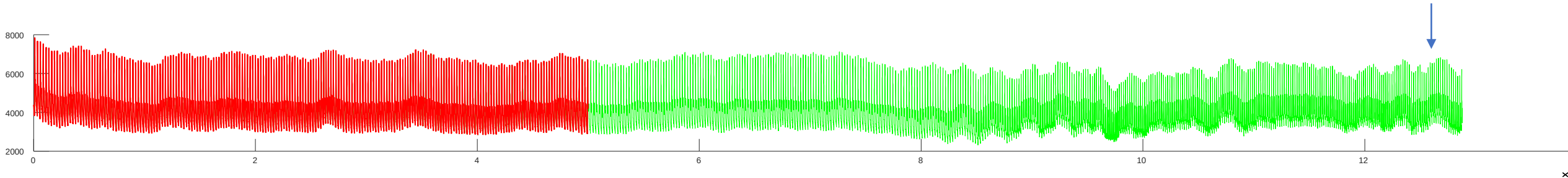
UCR_Anomaly_tiltAPB2_50000_124159_124985.txt



The data comes from a healthy male on a tilt table.
At first, he is supine, at around 80000, the table is tilted forward. The trace is his APB.

The anomaly is synthetic. We simply reverse the direction of the time series for about four beats.

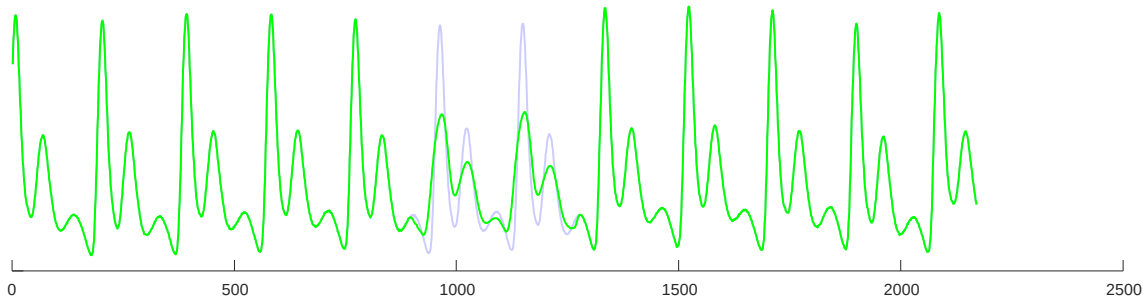


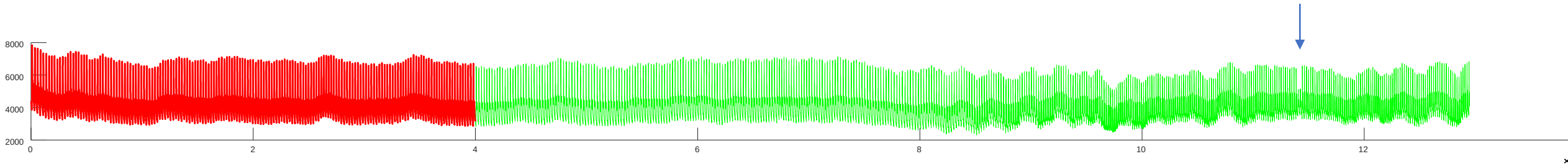


UCR_Anomaly_tiltAPB3_40000_114000_114370.txt

The data comes from a healthy male on a tilt table.
At first, he is supine, at around 80000, the table is tilted forward. The trace is his APB.

The anomaly is synthetic. We ran MATLAB's default smoothing on two beats. This has the effect of “dulling” the peaks and valleys.

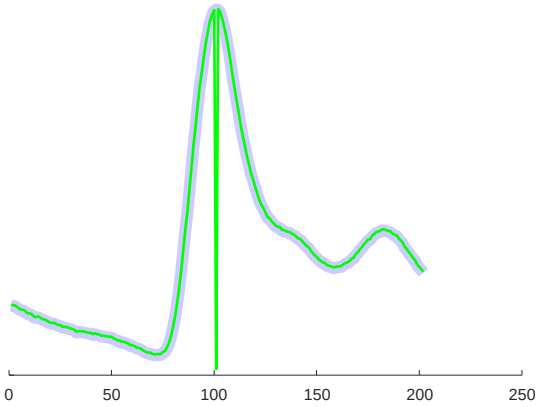


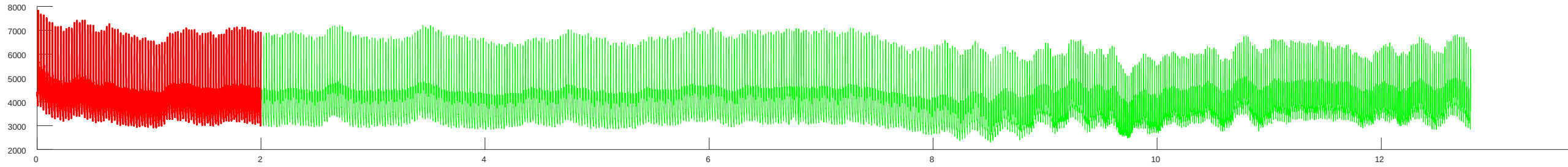


UCR_Anomaly_tiltAPB4_20000_67995_67996.txt

The data comes from a healthy male on a tilt table.
At first, he is supine, at around 80000, the table is tilted forward. The trace is his APB.

The anomaly is synthetic. We added a dropout at the top of a peak.

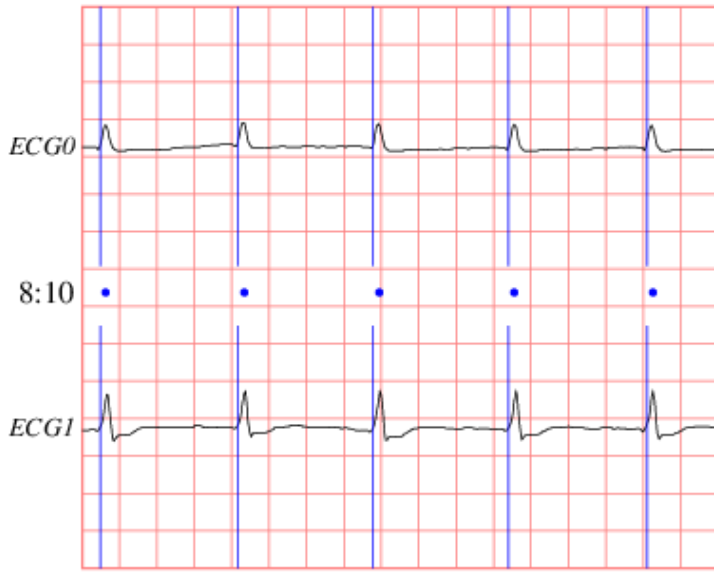


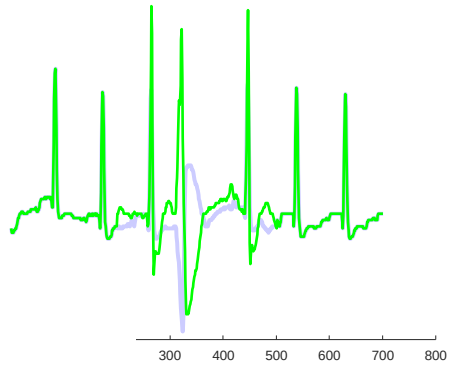


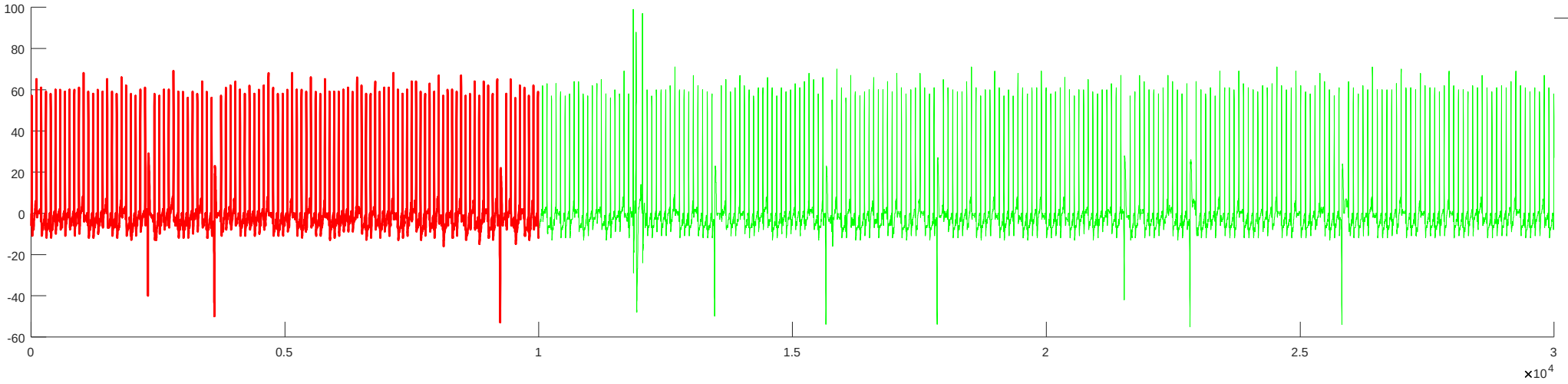
UCR_Anomaly_ECG1_10000_11800_12100.txt

This dataset was prepared from a two-lead ECG trace that contains some PVCs. Note that both train and test have multiple PVCs, that is not the anomaly.

The anomaly is synthetic. At a random location, we swapped in ECG1 for ECG2. That random location happened to include a PVC.



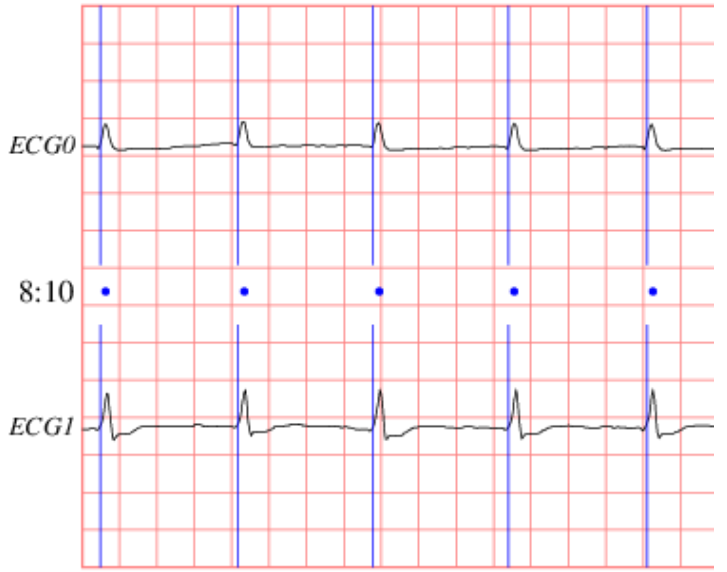




Selected input: record afpdb/n04 , annotator qrs , from 8:10.000 to 8:20.000

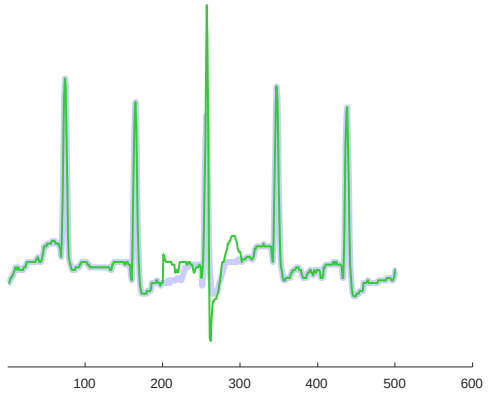
PAF Prediction Challenge Database (afpdb)

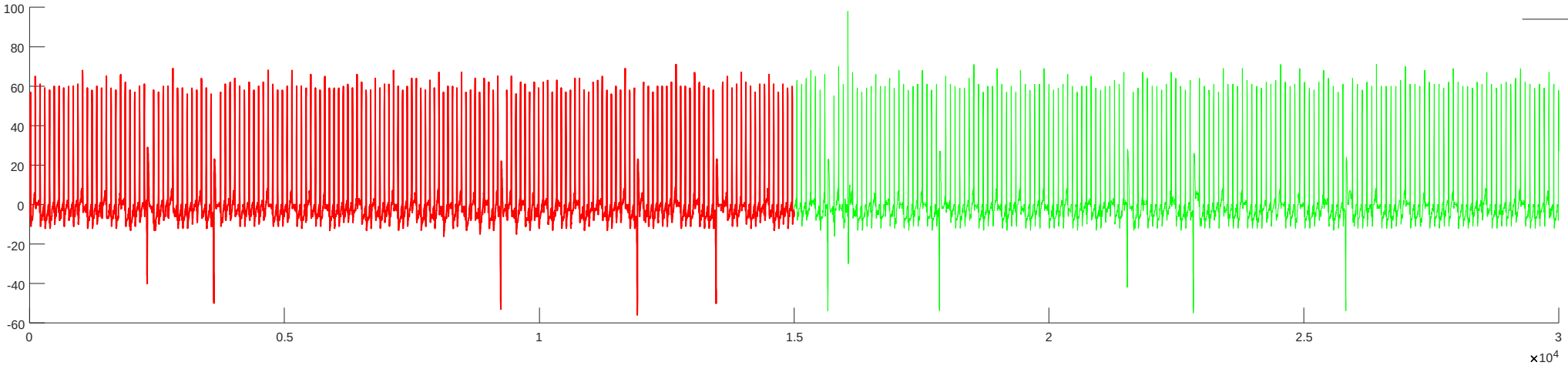
UCR_Anomaly_ECG2_15000_16000_16100.txt



This dataset was prepared from a two-lead ECG trace that contains some PVCs. Note that both train and test have multiple PVCs, that is not the anomaly.

The anomaly is synthetic. At a random location, we swapped in ECG1 for ECG2. This is a more subtle version of `UCR_Anomaly_ECG1_10000_11800_12100.txt`

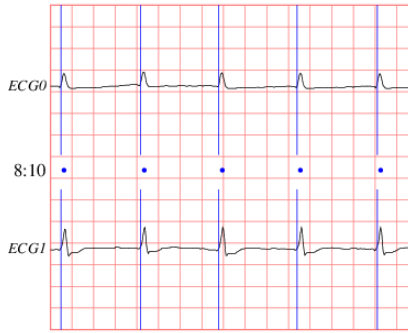




Selected input: record afpdb/n04 , annotator qrs , from 8:10.000 to 8:20.000

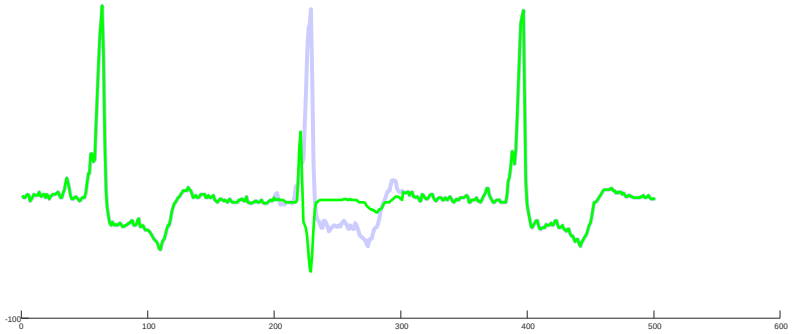
PAF Prediction Challenge Database (afpdb)

UCR_Anomaly_ECG3_8000_17000_17100.txt



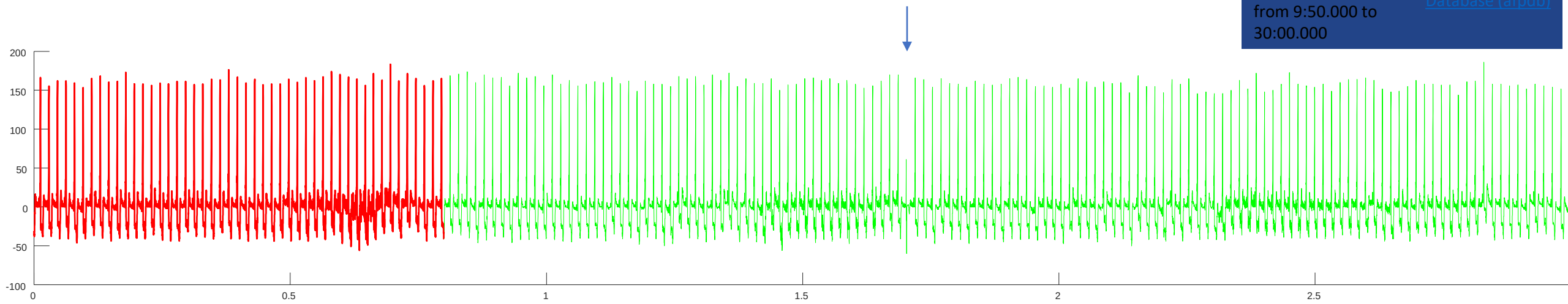
This dataset was prepared from an ECG trace.
Note that both train and test have noise levels that wax and wane, that is not the anomaly

The anomaly is synthetic. At a random location, we swapped in ECG1 for ECG2.



Selected input: record afpdb/n20 , annotator qrs , from 9:50.000 to 30:00.000

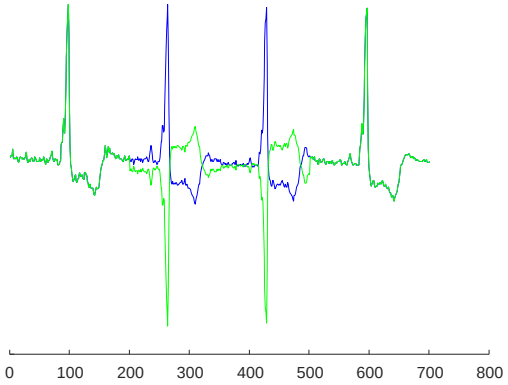
PAF Prediction Challenge Database (afpdb)



UCR_Anomaly_ECG4_5000_16800_17100.txt

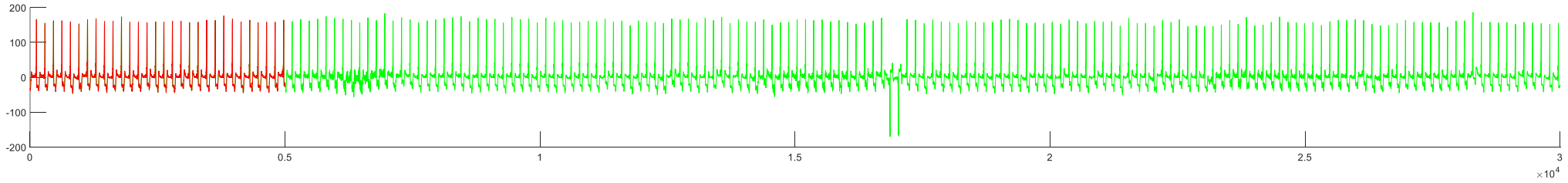
This dataset was prepared from an ECG trace.
Note that both train and test have noise levels that wax and wane, that is not the anomaly

The anomaly is synthetic. At a random location, we turned the data, about two beats worth, upside down (by multiplying the subsequence by -1)



Selected input: record afpdb/n20 , annotator qrs , from 9:50.000 to 30:00.000

PAF Prediction Challenge Database (afpdb)



UCR_Anomaly_resperation1_100000_110260_110412.txt

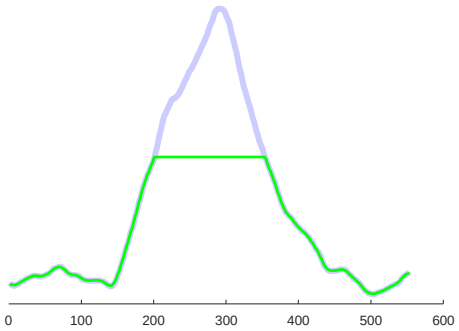
This is a respiration dataset.

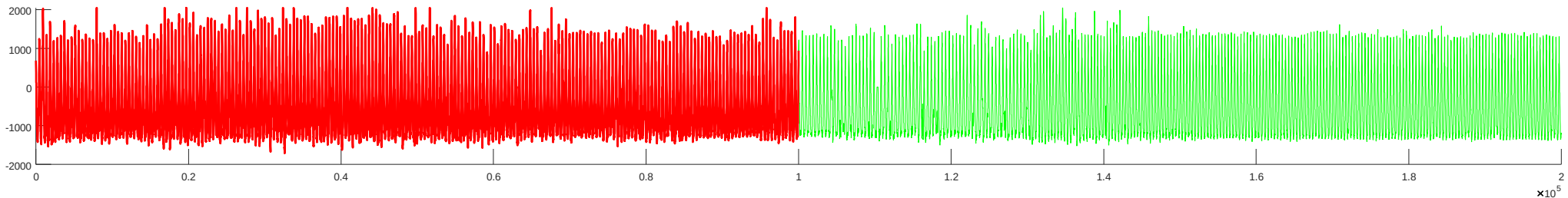
Many of these datasets have hard upper and lower limits, due to

1) Bit depth: Here we see a few peaks maximized at 2047 (which is $2^{12}/2 - 1$)
2) Mechanical limits of the elasticity of the chest band (hard*ish* limits, produces *almost* constant values)

The mechanical limits can change with movement, or with a nurse placing her hand under the band for some reason etc.

Here we model such an event, for a random breath cycle, the maximum value was hard limited to near zero.





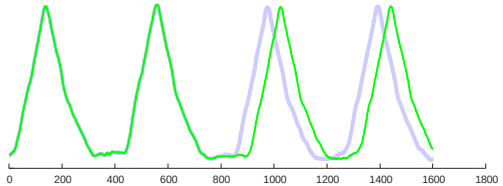
Selected input: record mimicdb/037/037 , annotator abp , from [13:30:35.000 14/08/1994] to [14:30:35.000 14/08/1994]

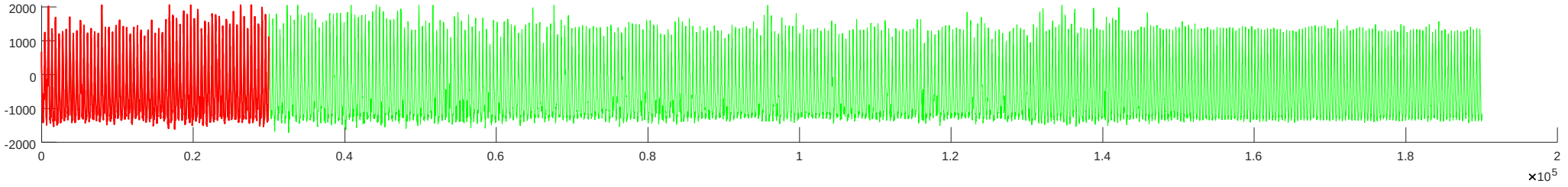
MIMIC Database (mimicdb)

UCR_Anomaly_resperation2_30000_168250_168251.txt

This is a respiration dataset.
The breath cycles here are very regular (implying a deep sleep).
Here we created an anomaly by increasing the time the patient held their breath (by a subtle amount).

This models sleep apnea





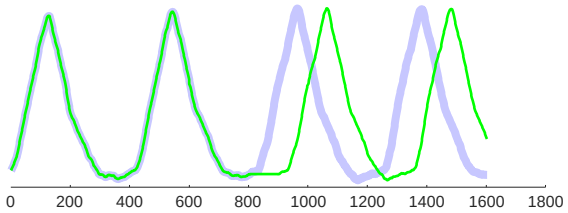
Selected input: record mimicdb/037/037 , annotator abp , from [13:30:35.000 14/08/1994] to [14:30:35.000 14/08/1994]

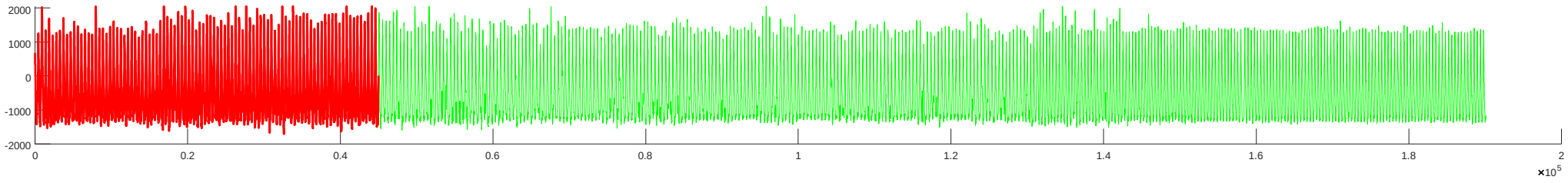
MIMIC Database (mimicdb)

UCR_Anomaly_resperation3_45000_158250_158251.txt

This is a respiration dataset.
The breath cycles here are very regular (implying a deep sleep).
Here we created an anomaly by increasing the time the patient held their breath (by a significant amount).

This models sleep apnea





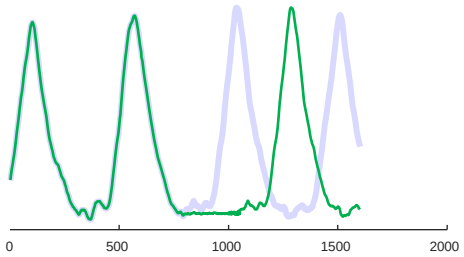
Selected input: record mimicdb/037/037 , annotator abp , from [13:30:35.000 14/08/1994] to [14:30:35.000 14/08/1994]

MIMIC Database (mimicdb)

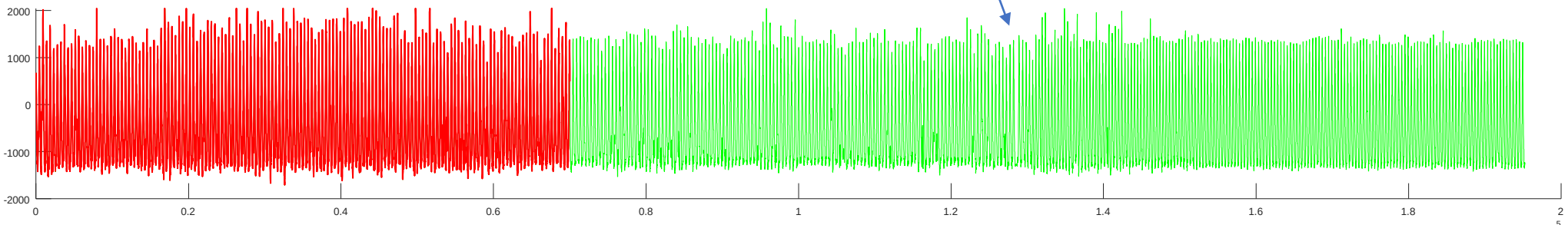
UCR_Anomaly_resperation4_70000_128430_128431.txt

This is a respiration dataset.
The breath cycles here are very regular (implying a deep sleep).
Here we created an anomaly by increasing the time the patient held their breath (by a very large amount).

This models sleep apnea



visible at this scale

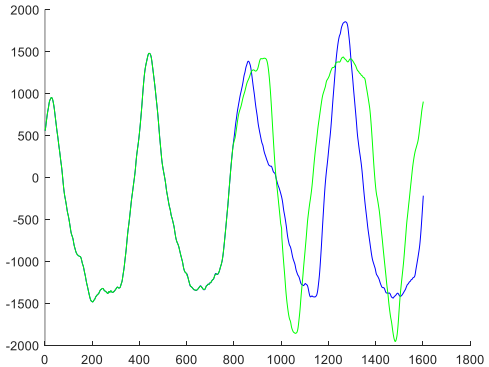


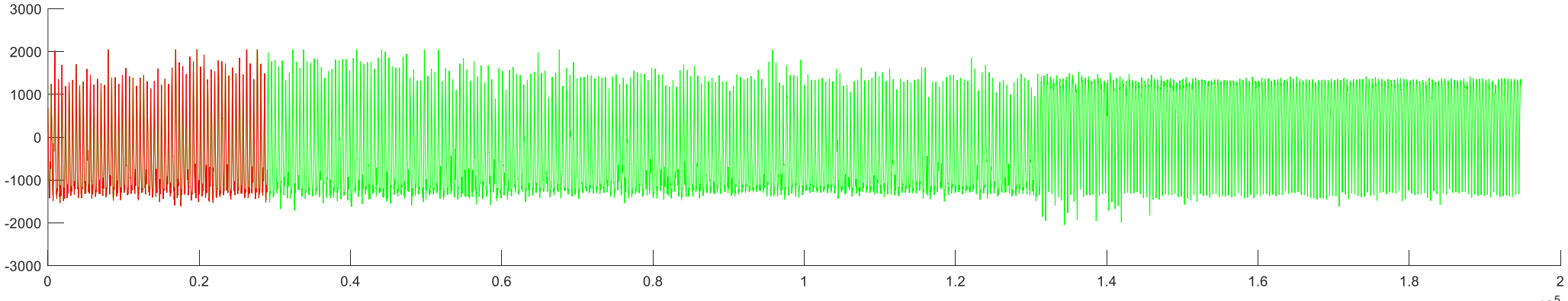
Selected input: record mimicdb/037/037 , annotator abp , from [13:30:35.000 14/08/1994] to [14:30:35.000 14/08/1994]

MIMIC Database (mimicdb)

UCR_Anomaly_resperation8_29000_131230_131231.txt

This is a respiration dataset.
The breath cycles here are very regular (implying a deep sleep).
Here we created an anomaly by simply turning the data upside down at a random location.
We cleaned up and smoothed the “join” so as to make it less visible.



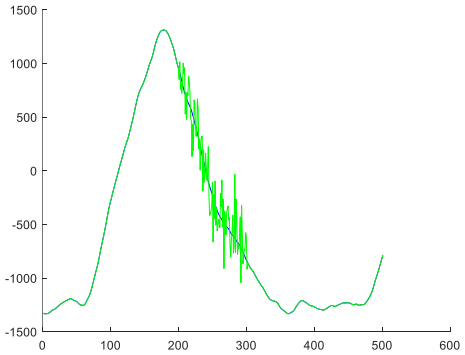


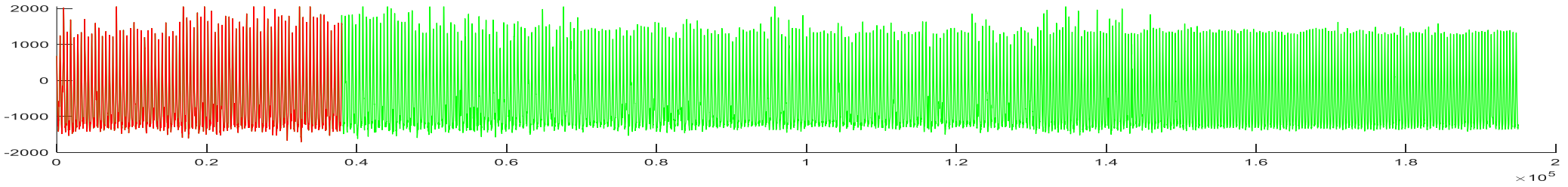
Selected input: record mimicdb/037/037 , annotator abp , from [13:30:35.000 14/08/1994] to [14:30:35.000 14/08/1994]

MIMIC Database (mimicdb)

UCR_Anomaly_resperation9_38000_143411_143511.txt

This is a respiration dataset.
The breath cycles here are very regular (implying a deep sleep).
Here we created an anomaly by adding noise.





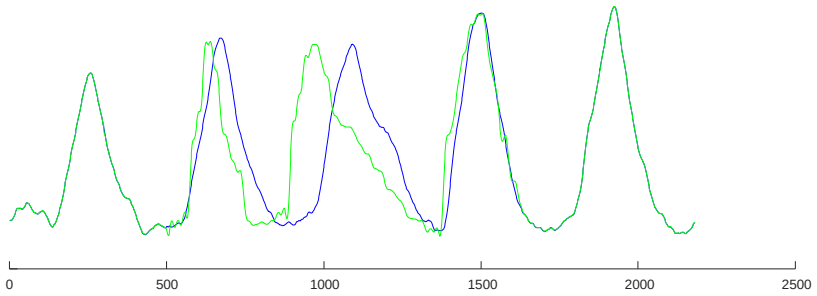
Selected input: record mimicdb/037/037 , annotator abp , from [13:30:35.000 14/08/1994] to [14:30:35.000 14/08/1994]

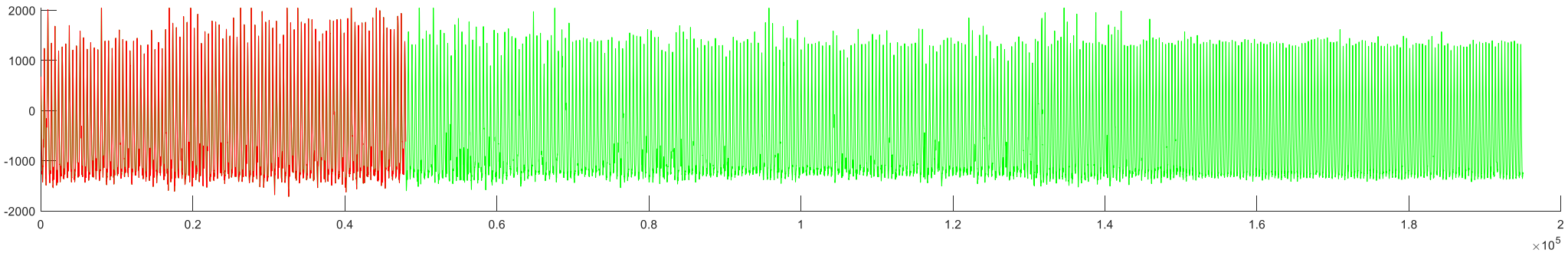
MIMIC Database (mimicdb)

UCR_Anomaly_resperation10_48000_130700_131880.txt

This is a respiration dataset.
The breath cycles here are very regular (implying a deep sleep).
Here we created an anomaly by adding time warping.

```
function [warped_T] = add_warping_one_time_series(T,p)
    i = randperm(length(T));
    i = sort(i(1:end-floor(length(T) * p)));
    warped_T = smooth(resample(T(i),length(T),length(i)),1);
```





Selected input: record mimicdb/037/037 , annotator abp , from [13:30:35.000 14/08/1994] to [14:30:35.000 14/08/1994]

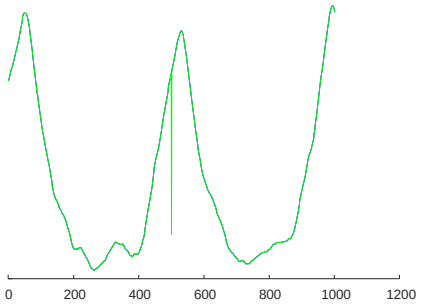
MIMIC Database (mimicdb)

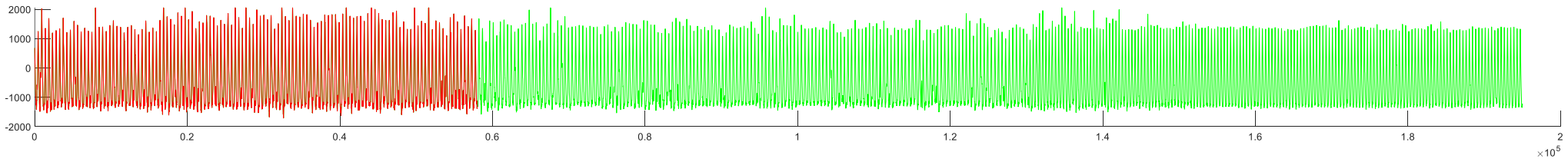
UCR_Anomaly_resperation11_58000_110800_110801.txt

This is a respiration dataset.
The breath cycles here are very regular (implying a deep sleep).

In some legacy systems, missing values are encoded at -999.

Here we randomly inserted such a value.





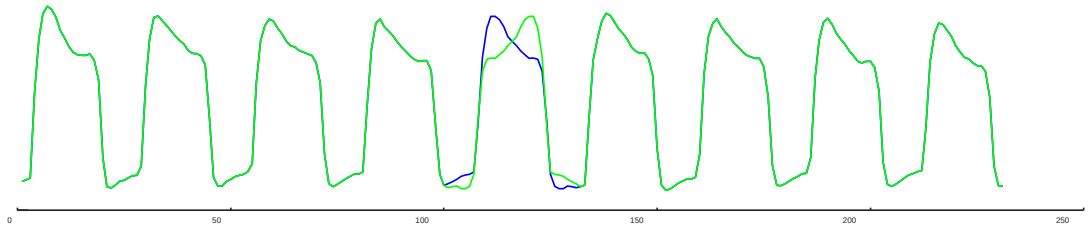
Selected input: record mimicdb/037/037 , annotator abp , from [13:30:35.000 14/08/1994] to [14:30:35.000 14/08/1994]

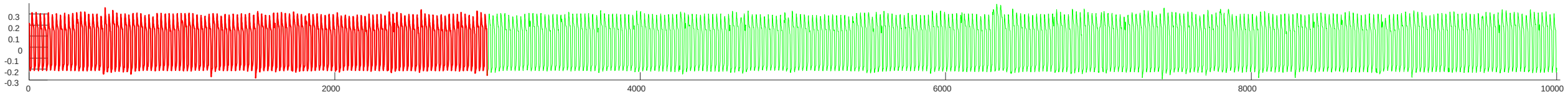
MIMIC Database (mimicdb)

UCR_Anomaly_insectEPG1_3000_7000_7030.txt

This is an insect EPG dataset, an Asian Citrus Psyllid.
This is a trace of insect feeding behavior.

We reversed a single cycle

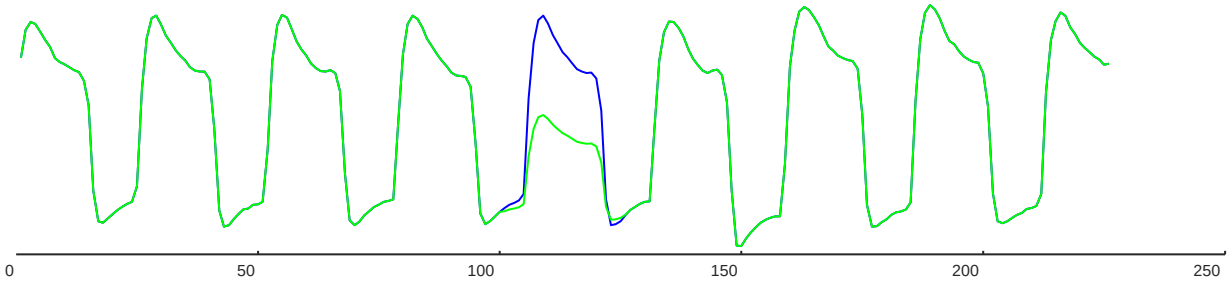


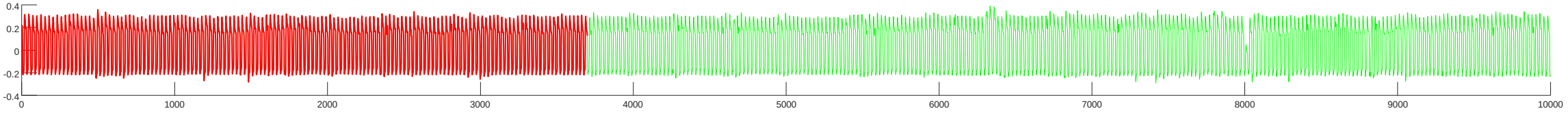


UCR_Anomaly_insectEPG2_3700_8000_8025.txt

This is an insect EPG dataset, an Asian Citrus Psyllid.
This is a trace of insect feeding behavior.

We reduced the amplitude of a single cycle by half

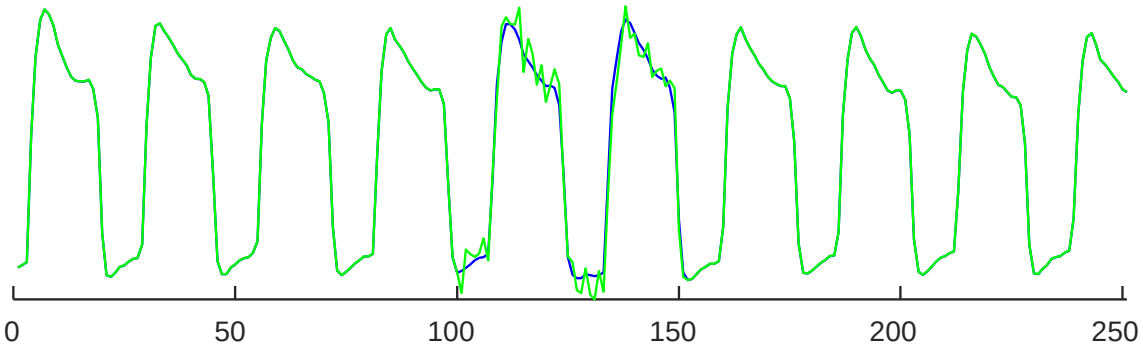


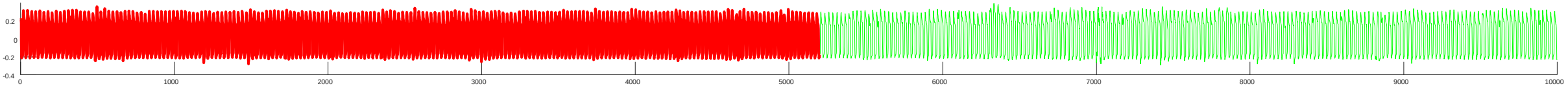


UCR_Anomaly_insectEPG3_5200_7000_7050.txt

This is an insect EPG dataset, an Asian Citrus Psyllid.
This is a trace of insect feeding behavior.

We added noise to two cycles

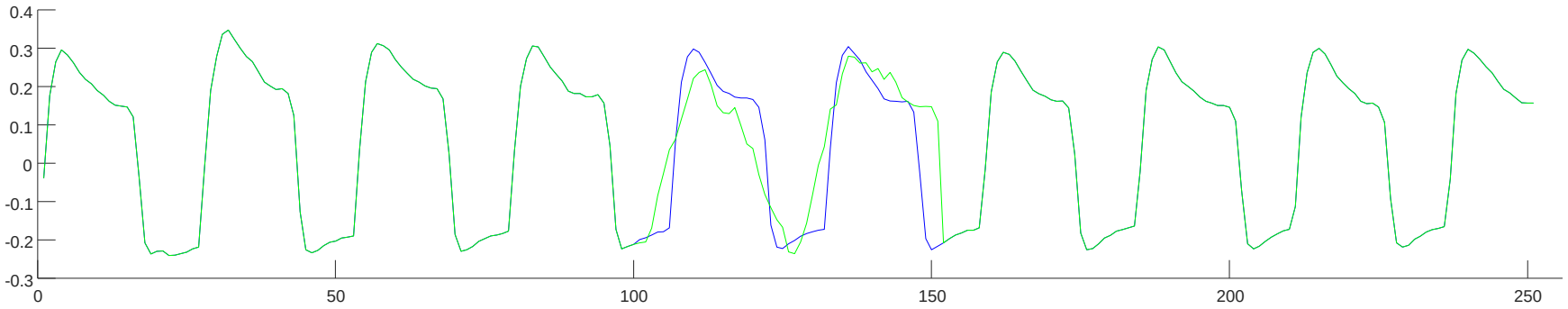


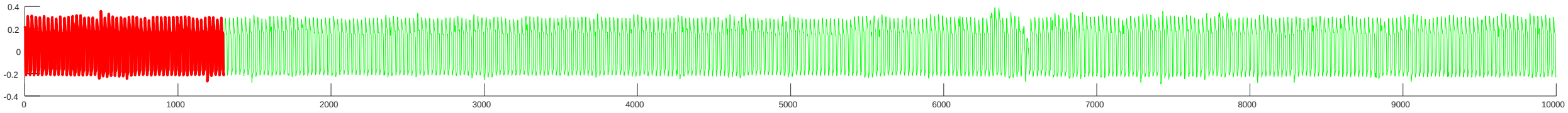


UCR_Anomaly_insectEPG4_1300_6508_6558.txt

This is an insect EPG dataset, an Asian Citrus Psyllid.
This is a trace of insect feeding behavior.

We added distorted two cycles by adding noise, then smoothing.

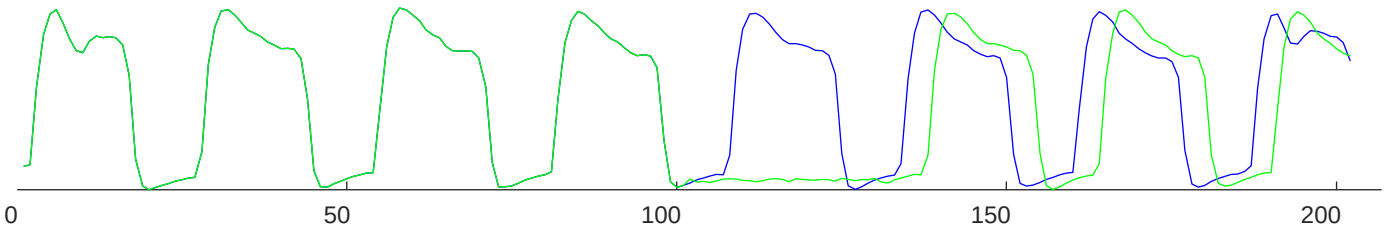


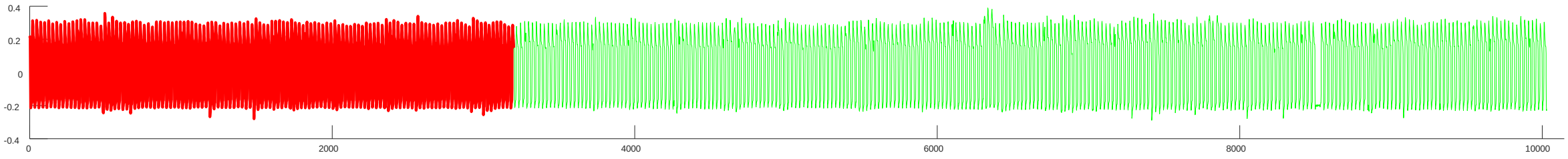


UCR_Anomaly_insectEPG5_3200_8500_8501.txt

This is an insect EPG dataset, an Asian Citrus Psyllid.
This is a trace of insect feeding behavior.

We added some space between two cycles



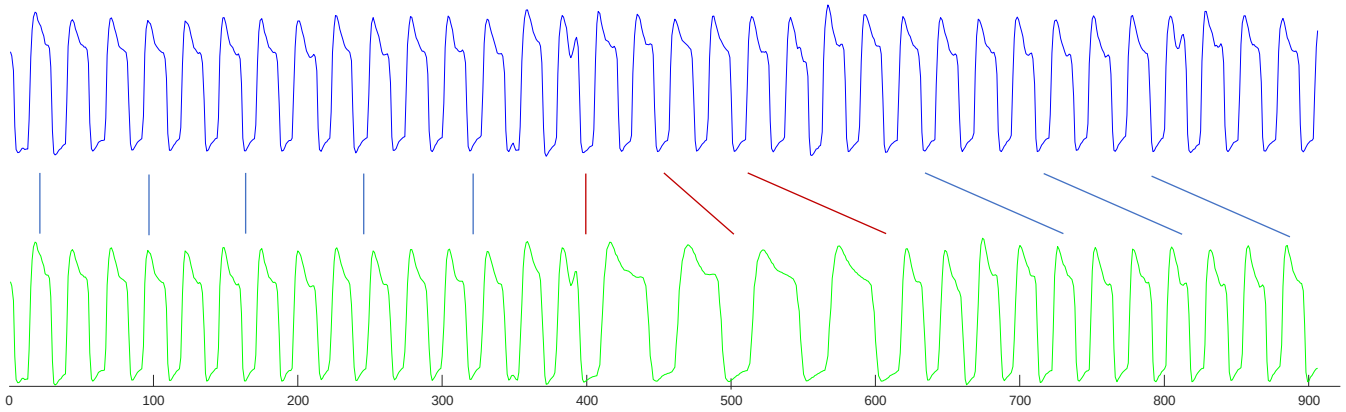


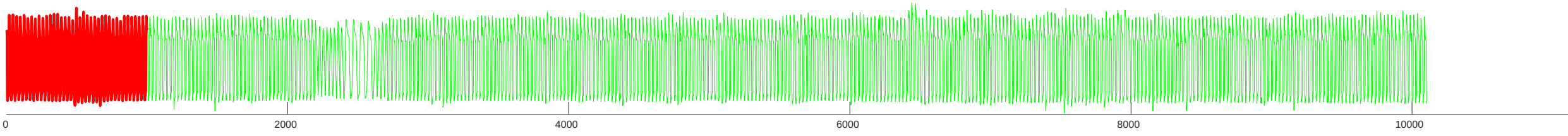
UCR_Anomaly_insectEPG6_1000_2400_2505.txt

This is an insect EPG dataset, an Asian Citrus Psyllid.
This is a trace of insect feeding behavior.

For four cycles, we slowed down the behavior by a factor of two, slightly smoothing the result to remove interpolation artifacts.

```
T = [ T(1:start_anomaly);   T(start_anomaly:0.5:end_anomaly) ;   T(end_anomaly:end)   ];
T(start_anomaly-200:end_anomaly+200) = smooth(T(start_anomaly-200:end_anomaly+200),1);
```



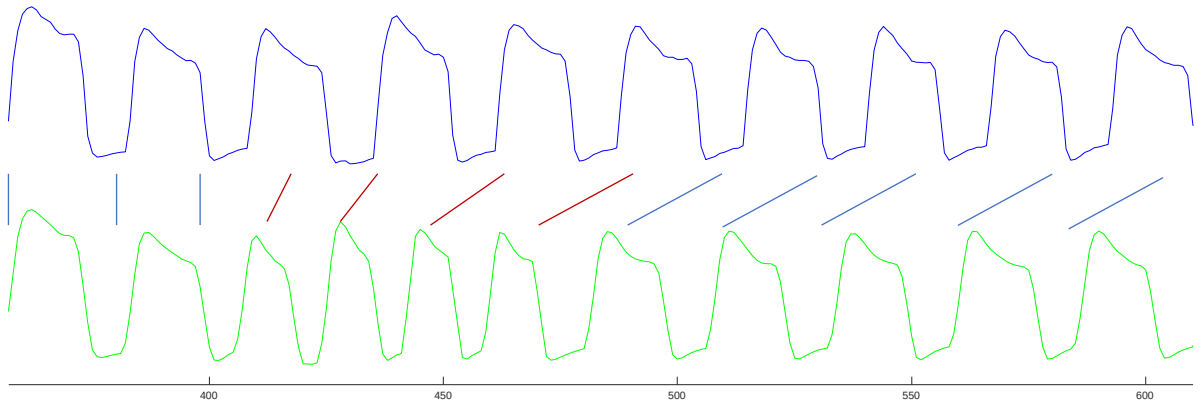


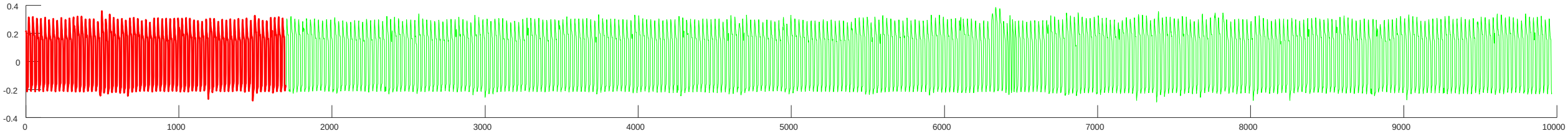
UCR_Anomaly_insectEPG7_1700_6400_6505.txt

This is an insect EPG dataset, an Asian Citrus Psyllid.
This is a trace of insect feeding behavior.

For four cycles, we speed up the behavior by a factor of 50%, slightly smoothing the result to remove interpolation artifacts.

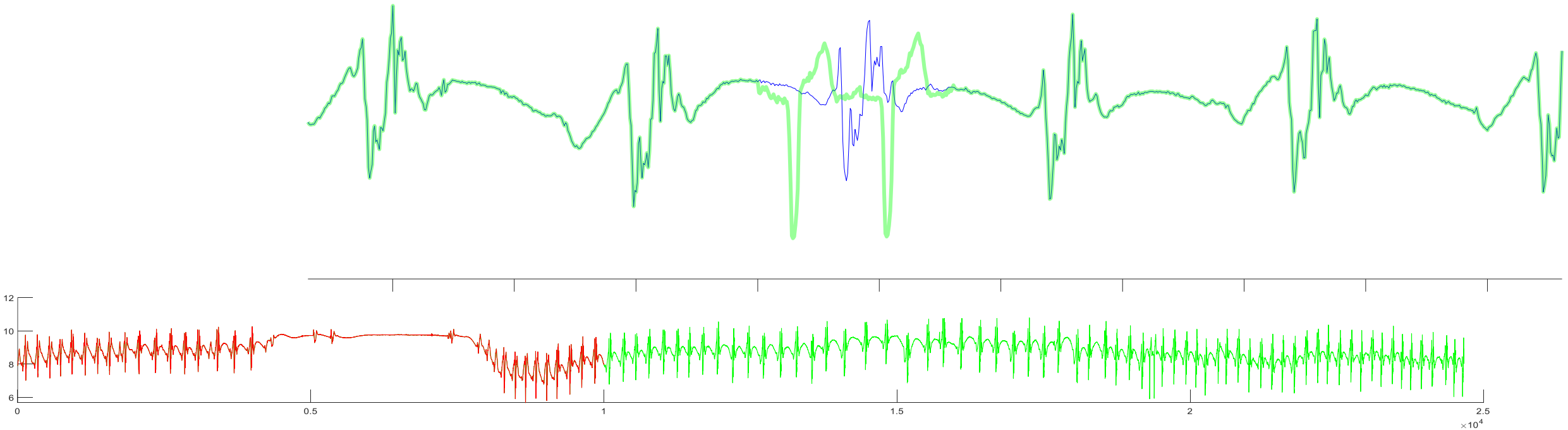
```
T = [ T(1:start_anomaly);   T(start_anomaly:1.5:end_anomaly) ;   T(end_anomaly:end)  ];
T(start_anomaly-200:end_anomaly+200) = smooth(T(start_anomaly-200:end_anomaly+200),1);
```





UCR_Anomaly_MesoplodonDensirostris_10000_19280_19440.txt

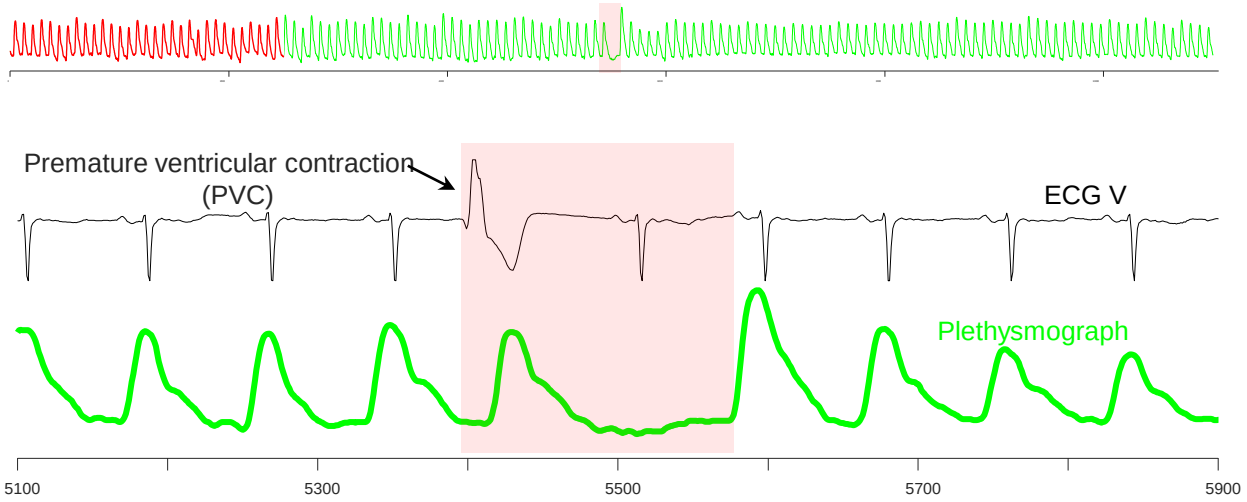
This data comes from an accelerometer attached to a Blainville's beaked whale.
We replace one swim cycle with two human heartbeats, scaled (*very* approximately) to have about the same mean and variance as the data they replaced

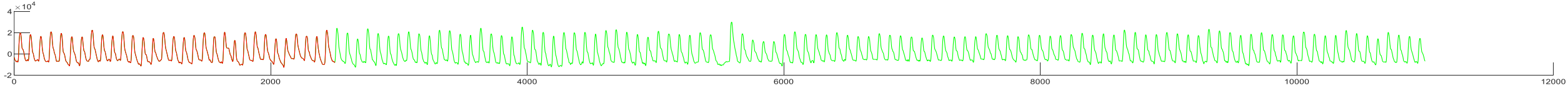


UCR_Anomaly_BIDMC1_2500_5400_5600.txt

Selected input: record bidmc/bidmc01 , annotator ann1 , from 0:00.000 to 1:00.000

Here the anomaly is a little subtle, how can we be so confident that is it semantically an anomaly? We can make this assertion because we examined the electrocardiogram that was recorded in parallel. This was the only region that had an abnormal heartbeat, a PVC. Note that there is a slight lag in the timing, as an ECG is an electrical signal, and the pleth signal is mechanical. However, the scoring functions typically have a little “play” to avoid the brittleness of requiring spurious precision.



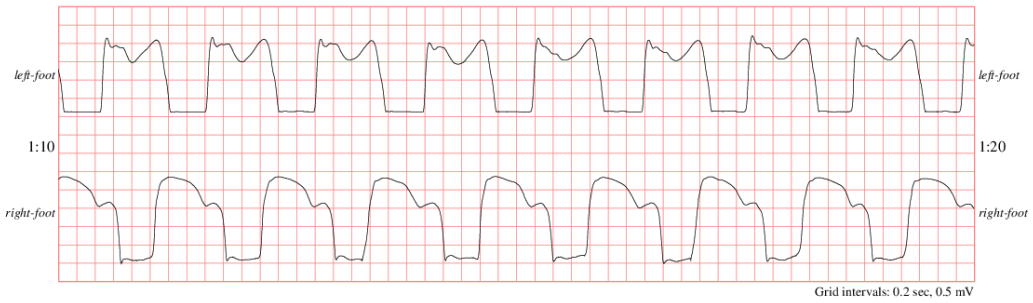


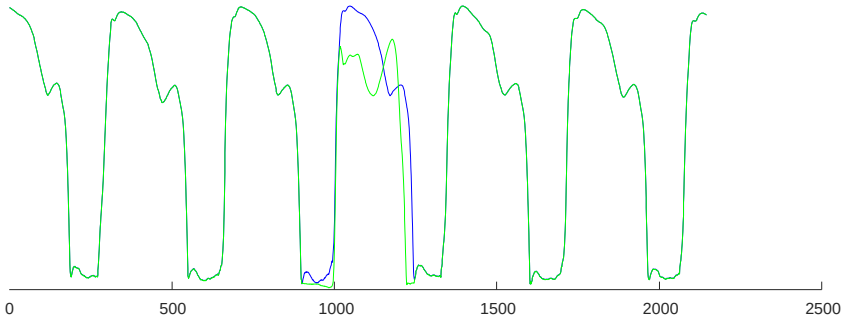
UCR_Anomaly_park3m_60000_72150_72495.txt

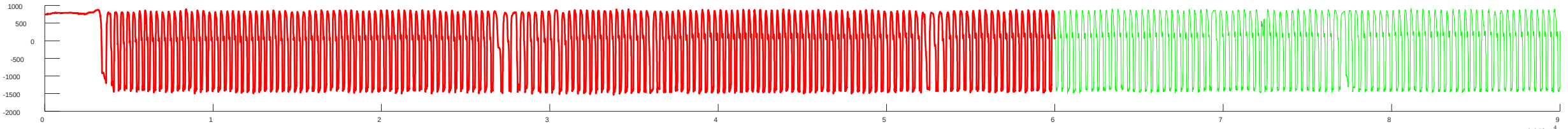
Selected input: record gaitndd/park3 , from 0:00.000 to 5:00.000

Gait in Neurodegenerative Disease Database (gaitndd)

We took the right foot of a person with an asymmetric gait because of Parkinson's disease, and swapped in one cycle from his left foot,







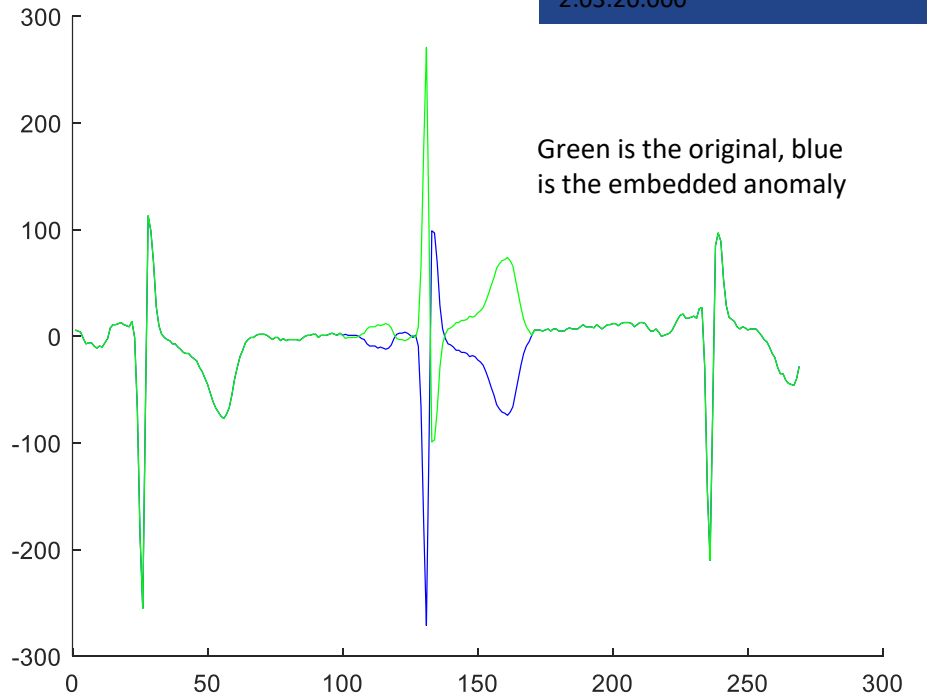
UCR_Anomaly_apneaecg_10000_12240_12308.txt

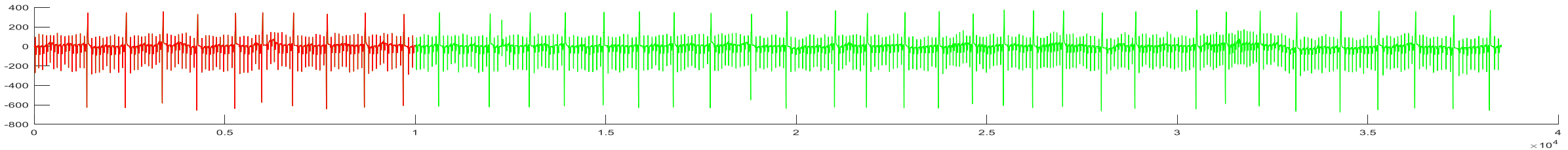
Selected input: record apnea-ecg/b05 , annotator apn , from 1:03:20.000 to 2:03:20.000

Apnea-ECG Database (apnea-ecg)

This dataset contains a mixture of normal beats, with *approximately* every tenth beat being a Premature ventricular contraction.

Here we simply flipped one normal beat upside down.





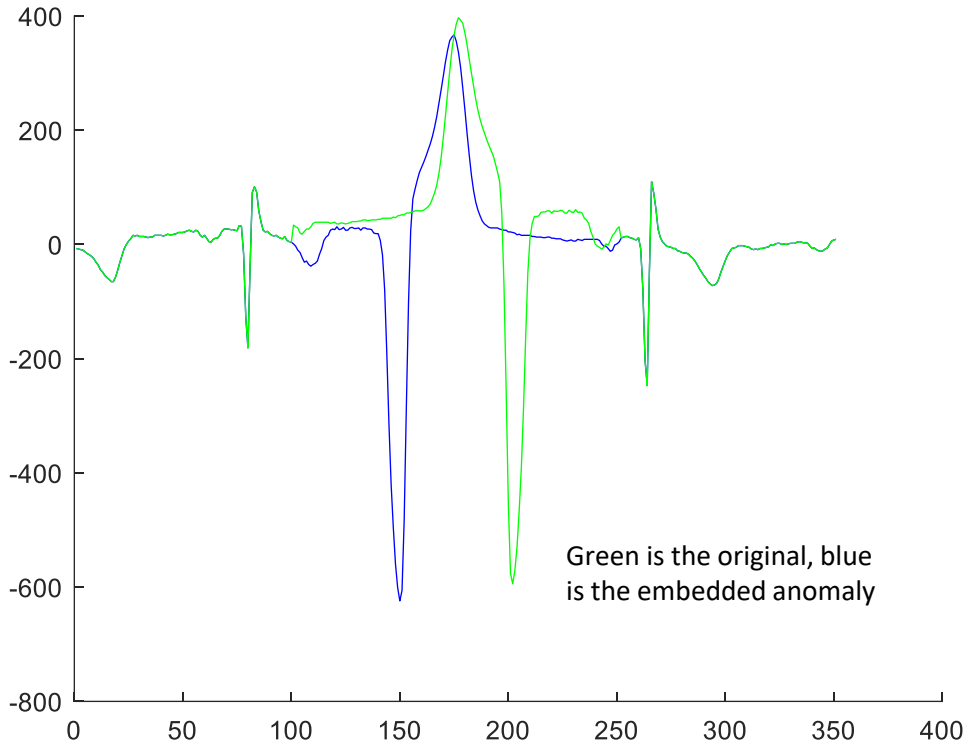
UCR_Anomaly_apneaecg2_10000_20950_21100.txt

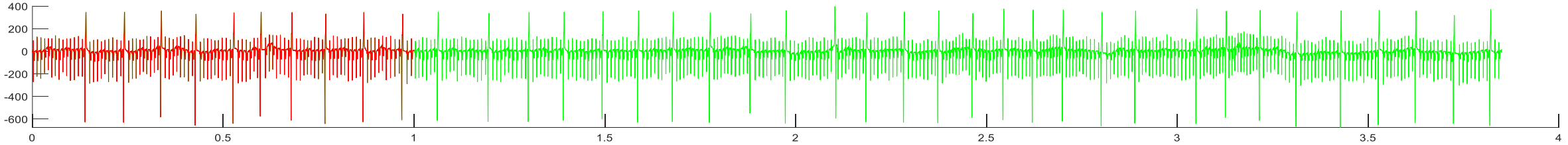
Selected input: record apnea-ecg/b05 , annotator apn , from 1:03:20.000 to 2:03:20.000

Apnea-ECG Database (apnea-ecg)

This dataset contains a mixture of normal beats, with *approximately* every tenth beat being a Premature ventricular contraction.

Here we simply flipped one PVC beat left to right down.





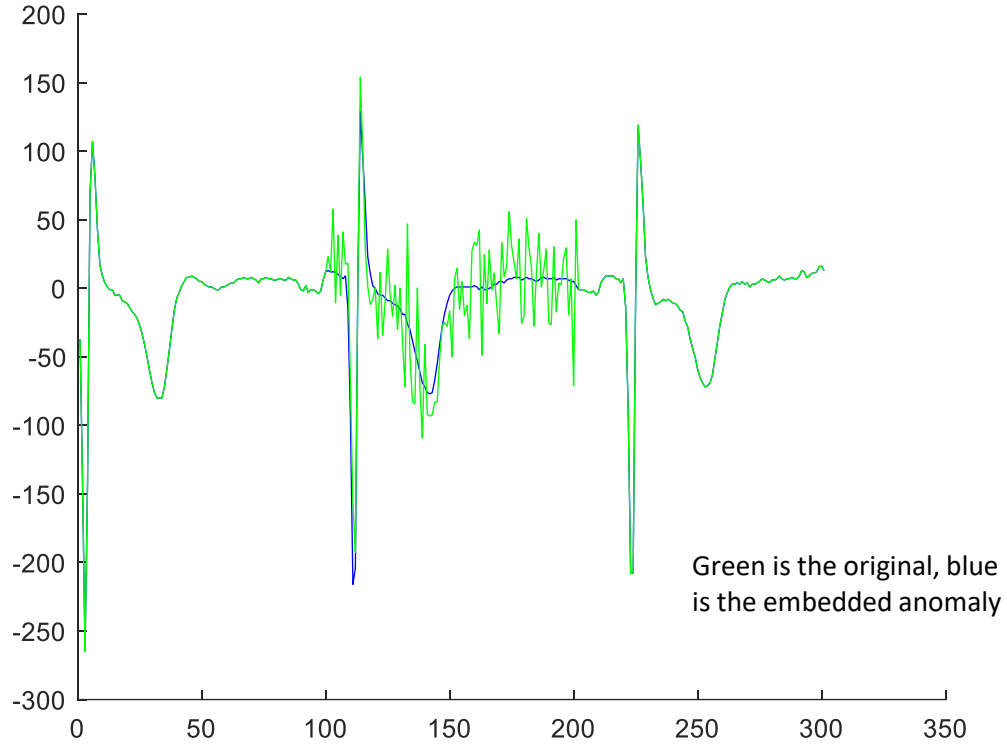
UCR_Anomaly_apneaecg3_5000_11111_11211.txt

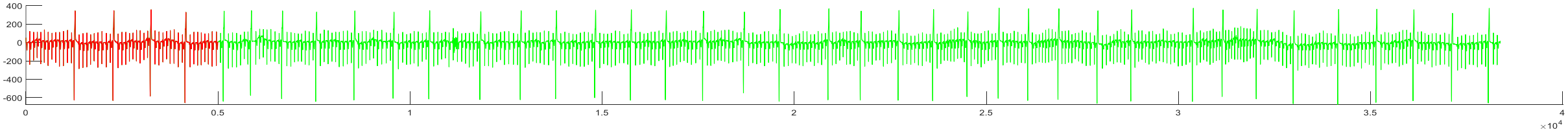
Selected input: record apnea-ecg/b05 , annotator apn , from 1:03:20.000 to 2:03:20.000

Apnea-ECG Database (apnea-ecg)

This dataset contains a mixture of normal beats, with *approximately* every tenth beat being a Premature ventricular contraction.

Here we simply added noise to a short region.





UCR_Anomaly_apneaecg4_6000_16000_16100.txt

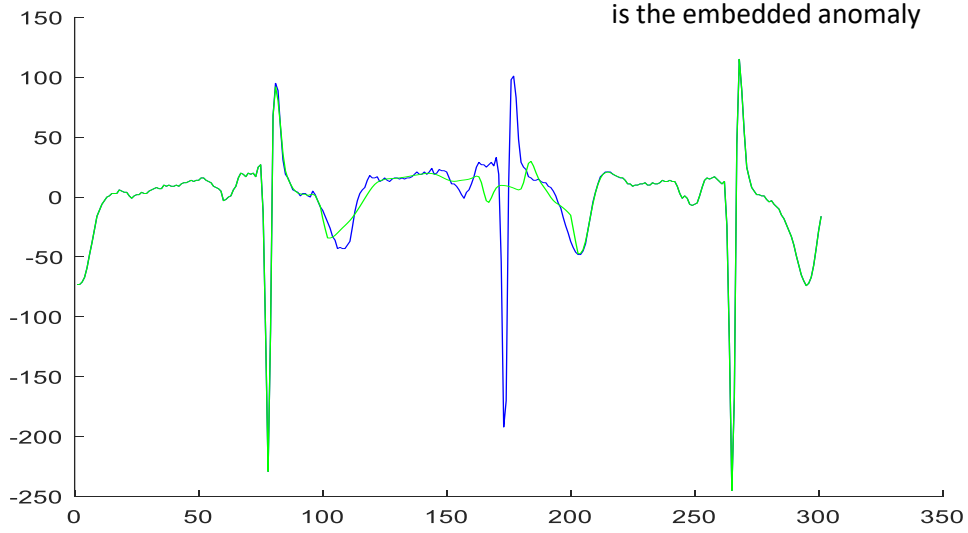
Selected input: record apnea-ecg/b05 , annotator apn , from 1:03:20.000 to 2:03:20.000

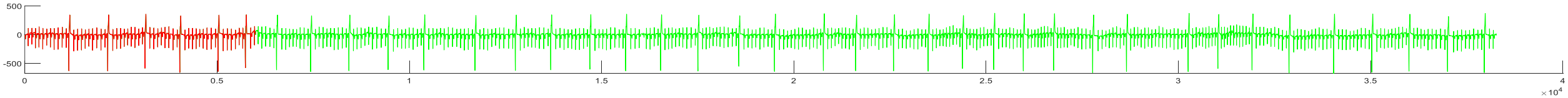
Apnea-ECG Database (apnea-ecg)

This dataset contains a mixture of normal beats, with *approximately* every tenth beat being a Premature ventricular contraction.

Here we *oversmoothed* a single beat. This tends to depress the peak and valley

Green is the original, blue is the embedded anomaly

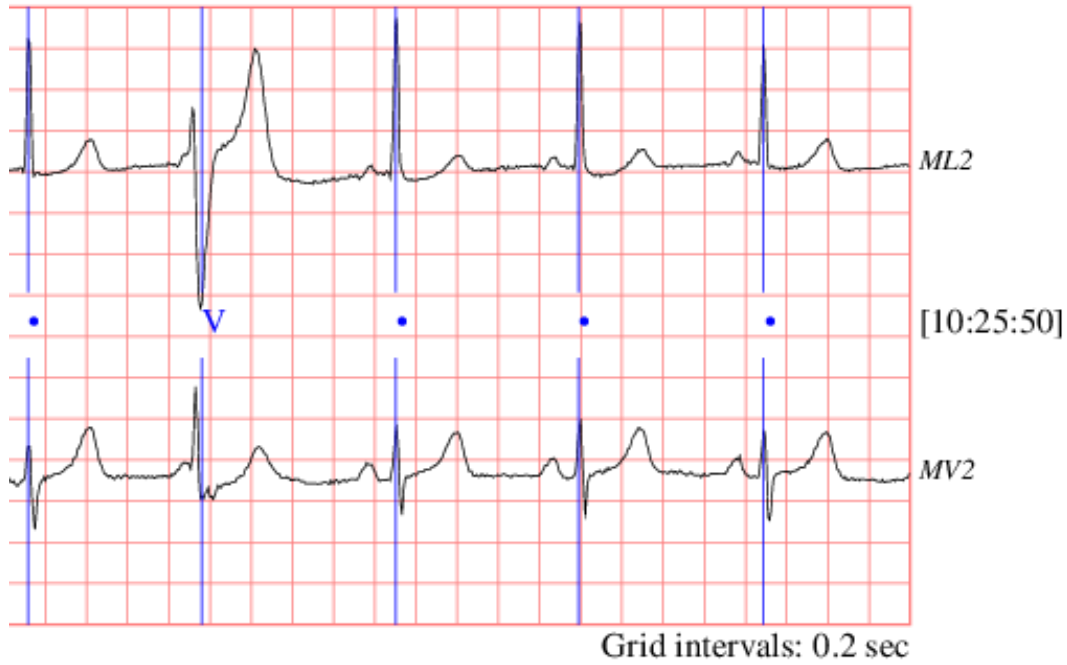




UCR_Anomaly_s20101m_10000_35774_35874.txt

This dataset contains normal beats, with a single ventricular beat between 35774 and 35874.

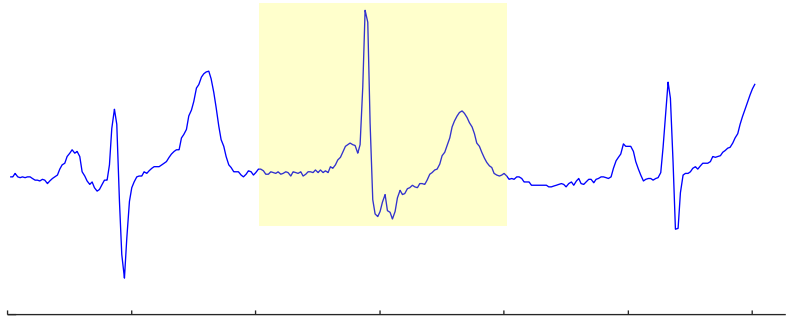
Here we use MV2, which is the more subtle of the two traces.

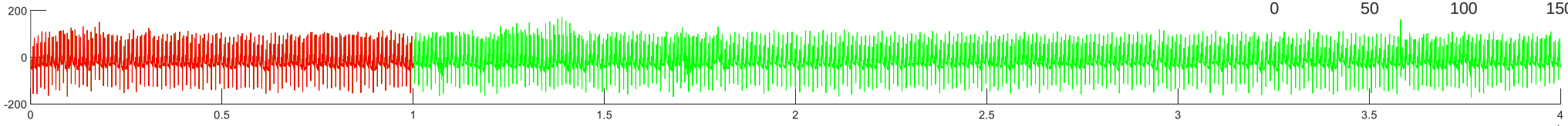


Selected input: record ltstdb/s20101 , annotator atr , from [10:25:40.000 01/11/1994] to [10:25:50.000 01/11/1994]

Long Term ST Database (ltstdb)

pschart -a atr -c "" -C -E -G -CG 1 .5 .5 -Cs 0 0 0 -H -l -P 300x250 -m 20 20 5 5 -M -n 0 \ -S 4 2 -t 25 -T "" -v 10 -w 0.5 script >chart.ps
convert -density 100x100 chart.ps chart.png

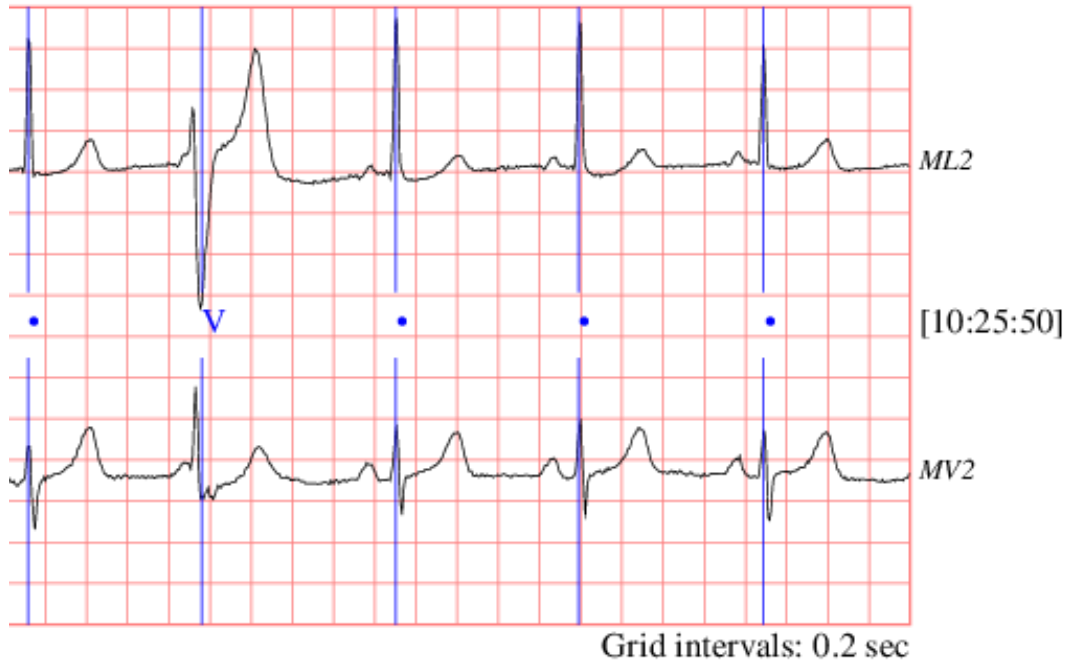


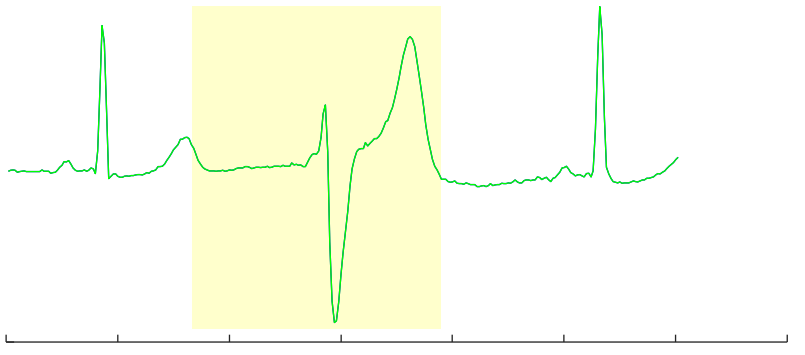


UCR_Anomaly_s20101mML2_12000_35774_35874.txt

This dataset contains normal beats, with a single ventricular beat between 35774 and 35874.

Here we use ML2, which is the more obvious of the two traces.

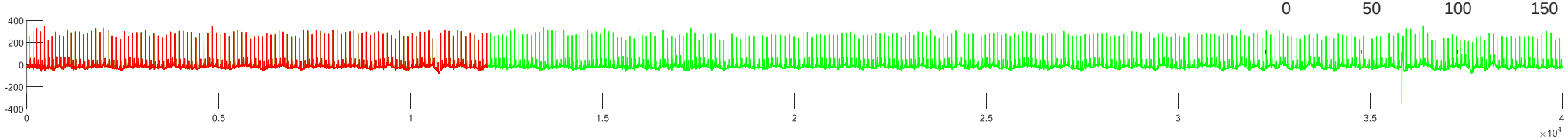




Selected input: record ltstdb/s20101 , annotator atr , from [10:25:40.000 01/11/1994] to [10:25:50.000 01/11/1994]

Long Term ST Database (ltstdb)

pschart -a atr -c "" -C -E -G -CG 1 .5 .5 -Cs 0 0 0 -H -l -P 300x250 -m 20 20 5 5 -M -n 0 \ -S 4 2 -t 25 -T "" -v 10 -w 0.5 script >chart.ps
convert -density 100x100 chart.ps chart.png

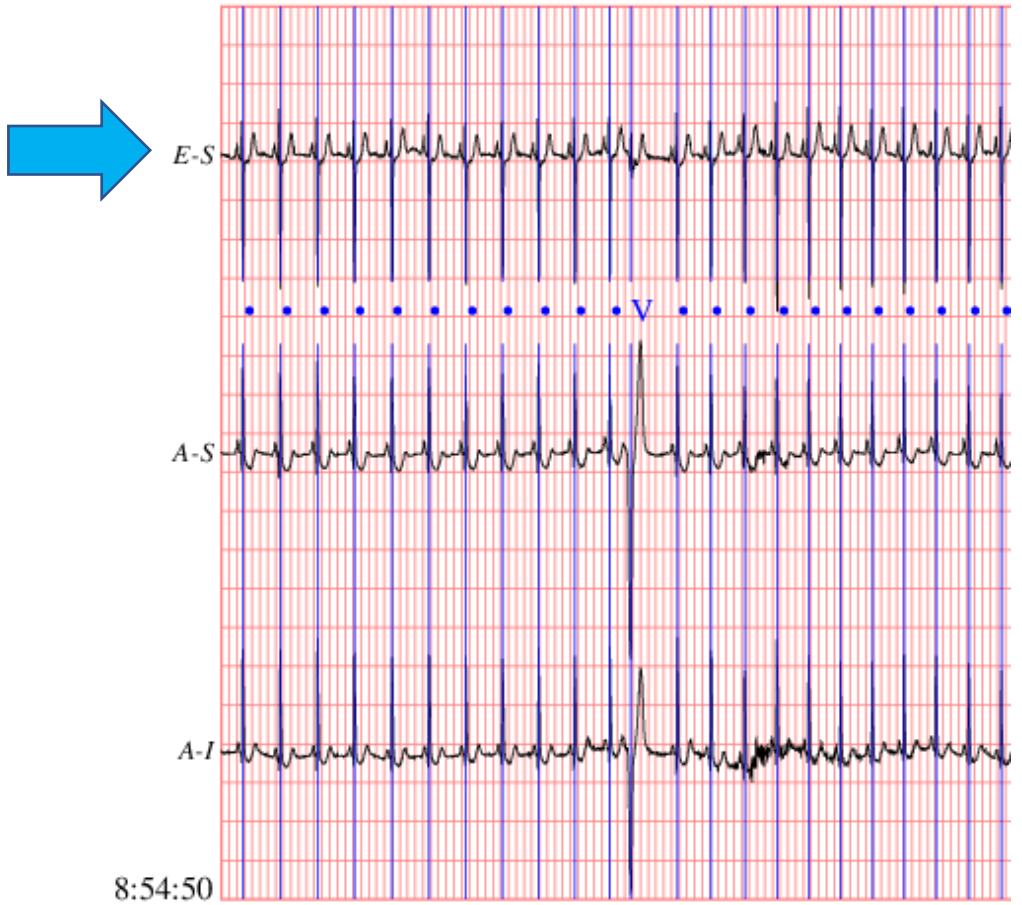


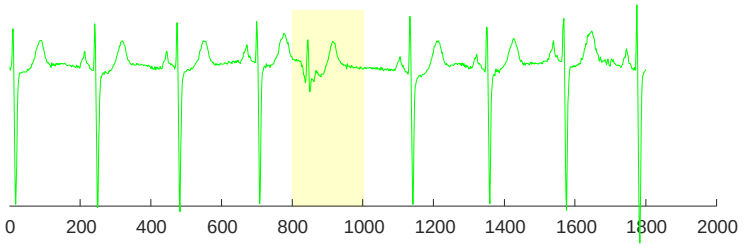
UCR_Anomaly_ltstdbs30791ES_20000_52600_52800.txt

This dataset contains normal beats, with a single ventricular beat between 52600 and 52800.

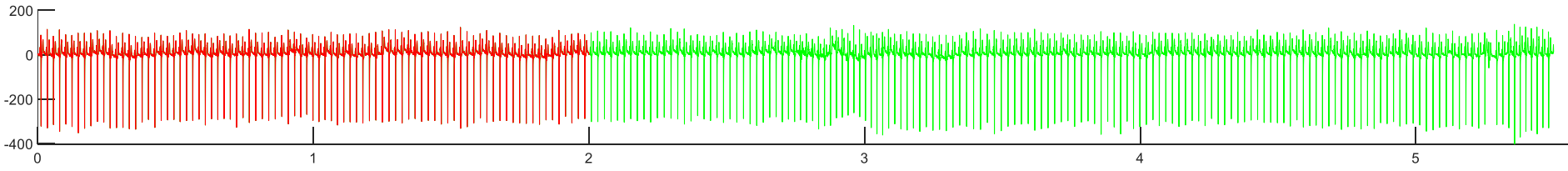
Here we use E-S, which is the more subtle of the three traces.

There are some noisy regions in the test set, which some algorithms might consider anomalies. However, there are just as noisy regions in the training set.
Likewise, there is slight wandering baseline in the test data, but we also see this in the train data.





pschart -a atr -c "" -C -E -G -CG 1 .5 .5 -Cs 0 0 0 -H -I -P 360x250 -m 20 20 5 5 -M -n 0 \ -S 4 2 -t 5 -T "" -v 10 -w 0.5 script >chart.ps
convert -density 100x100 chart.ps chart.png



Selected input: record ltstdb/s30791 , annotator atr , from 8:54:50.000 to 8:55:50.000

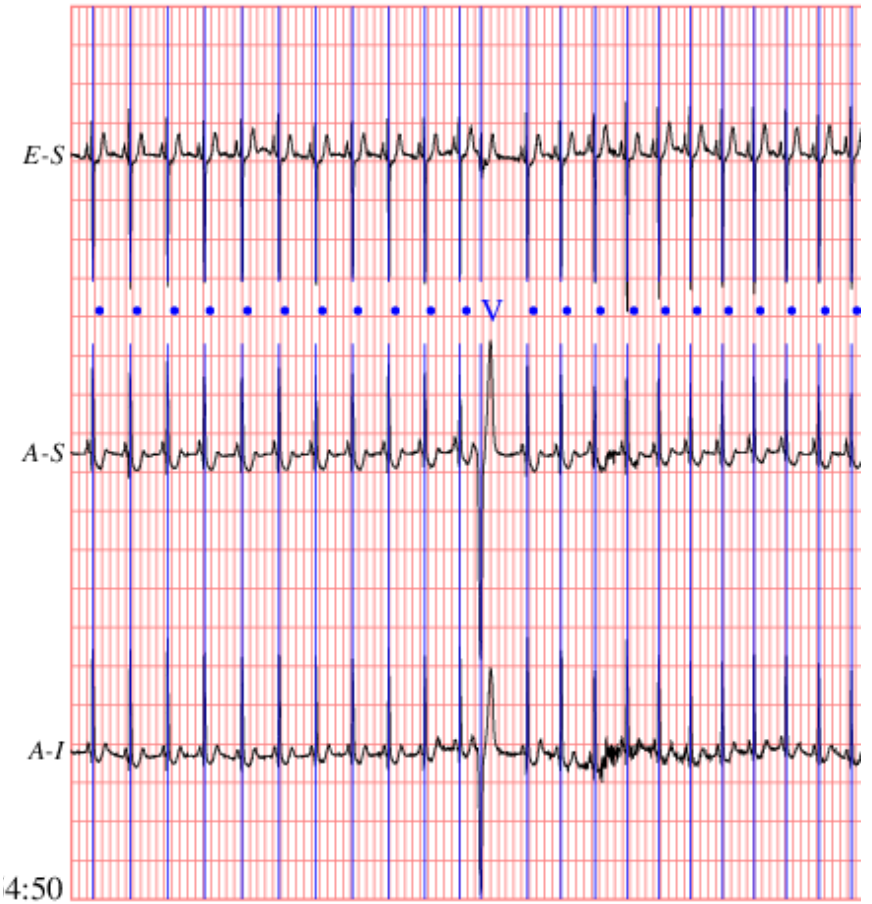
Long Term ST Database (ltstdb)

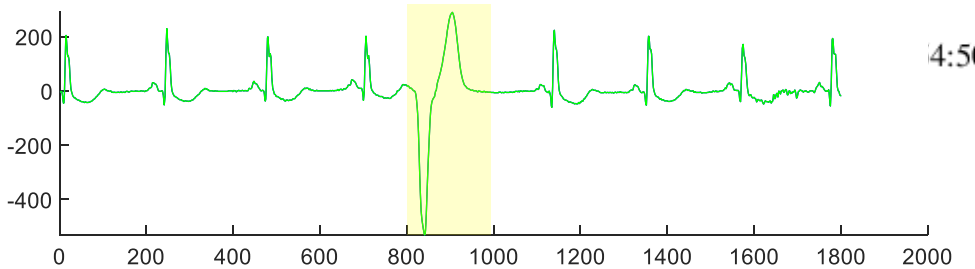
UCR_Anomaly_ltstdbs30791AS_23000_52600_52800.txt

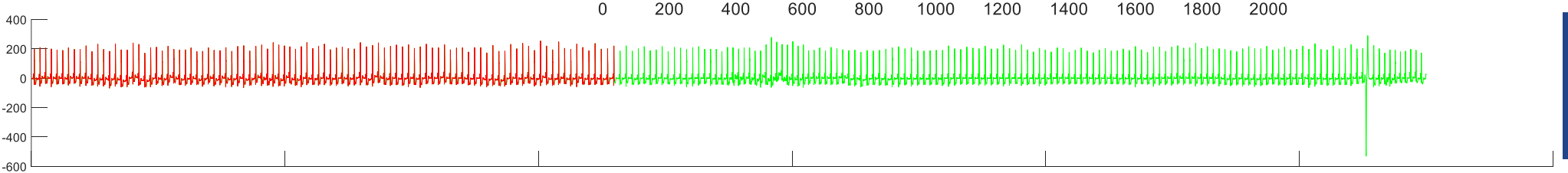
This dataset contains normal beats, with a single ventricular beat between 52600 and 52800.

Here we use A-S, which is the least subtle of the three traces.

There are some noisy regions in the test set, which some algorithms might consider anomalies. However, there are just as noisy regions in the training set.
Likewise, there is slight wandering baseline in the test data, but we also see this in the train data.







pschart -a atr -c "" -C -E -G -CG 1 .5 .5 -Cs 0 0 0 -H -I -P 360x250 -m 20 20 5 5 -M -n 0 \ -S 4 2 -t 5 -T "" -v 10 -w 0.5 script >chart.ps
convert -density 100x100 chart.ps chart.png

Selected input: record ltstdb/s30791 , annotator atr , from 8:54:50.000 to 8:55:50.000

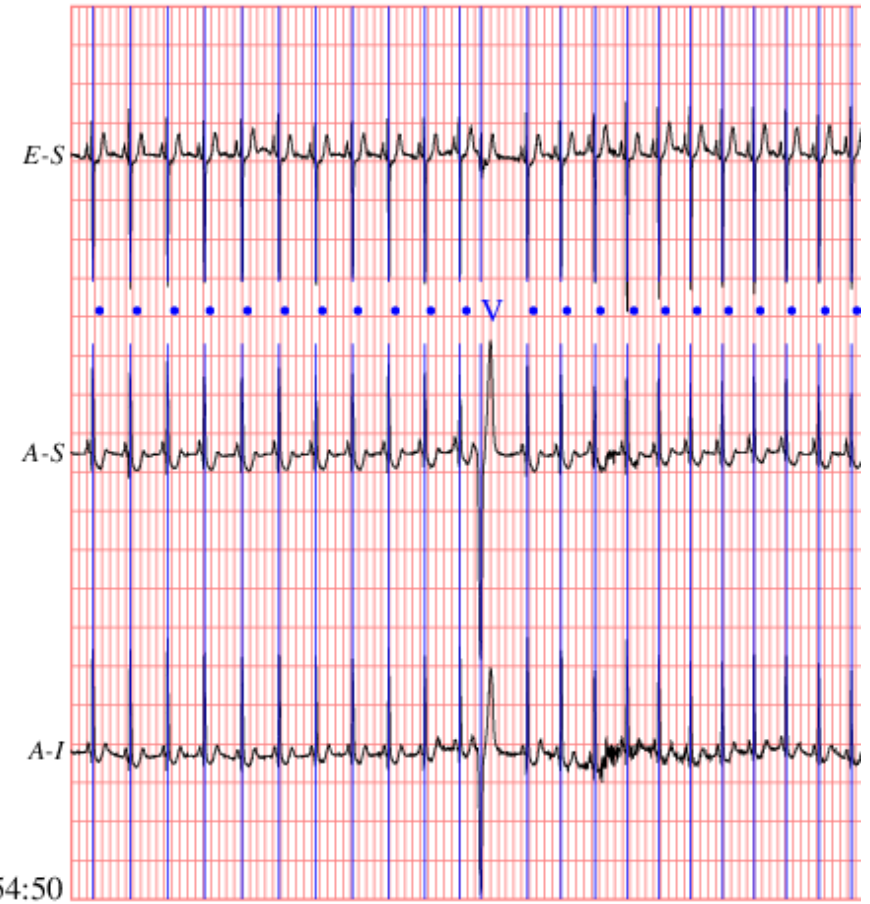
Long Term ST Database (ltstdb)

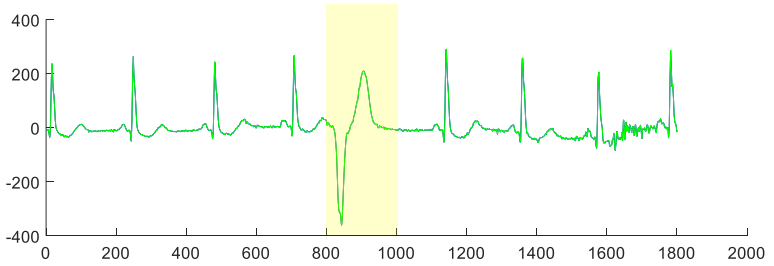
UCR_Anomaly_ltstdbs30791AI_17555_52600_52800.txt

This dataset contains normal beats, with a single ventricular beat between 52600 and 52800.

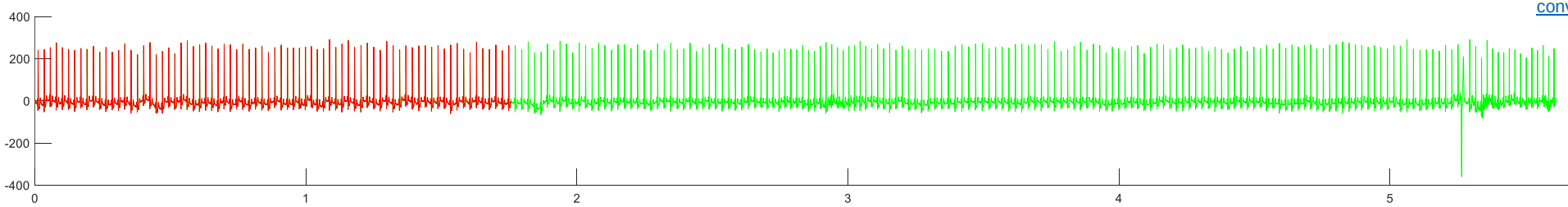
Here we use A-I.

There are some noisy regions in the test set, which some algorithms might consider anomalies. However, there are just as noisy regions in the training set.
Likewise, there is slight wandering baseline in the test data, but we also see this in the train data.





pschart -a atr -c "" -C -E -G -CG 1 .5 .5 -Cs 0 0 0 -H -I -P 360x250 -m 20 20 5 5 -M -n 0 \ -S 4 2 -t 5 -T "" -v 10 -w 0.5 script >chart.ps
convert -density 100x100 chart.ps chart.png



Selected input: record ltstdb/s30791 , annotator atr , from 8:54:50.000 to 8:55:50.000

Long Term ST Database (ltstdb)

UCR_Anomaly_qtdbSel100MLII_4000_13400_13800.txt

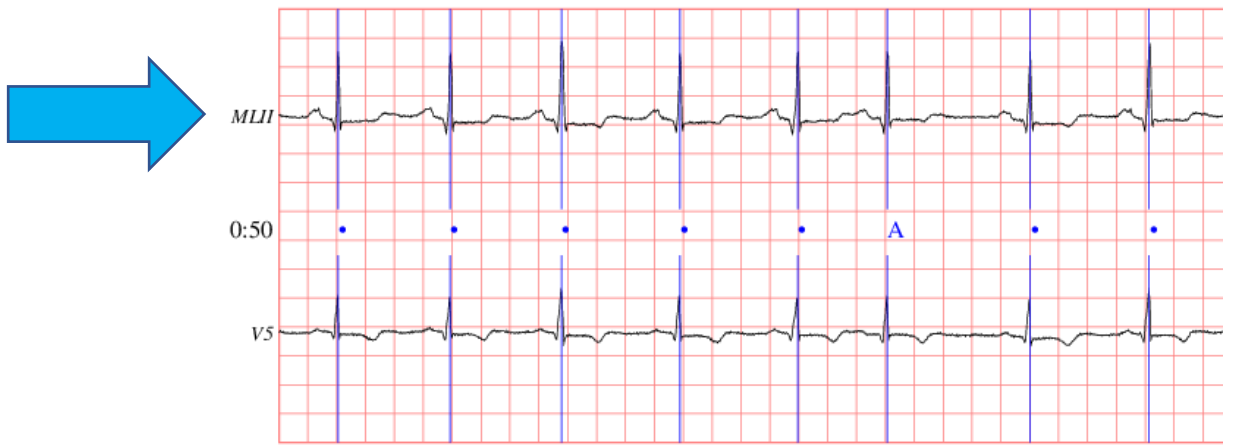
pschart -a atr -c "" -C -E -G -CG 1 .5 .5 -Cs 0 0 0 -H -l -P 300x250 -m 20 20 5 5 -M -n 0 \ -S 4 2 -t 25 -T "" -v 10 -w 0.5 script >chart.ps
convert -density 100x100 chart.ps chart.png

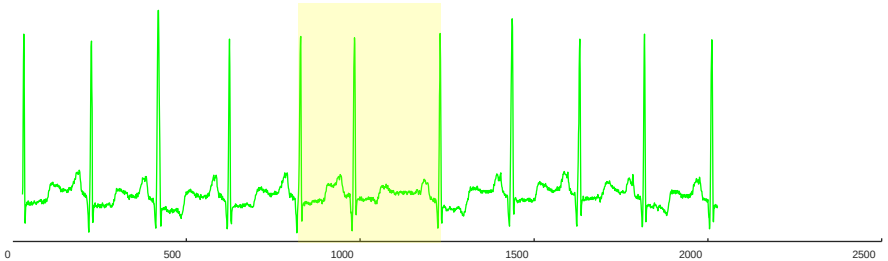
Selected input: record qtdb/sel100 , annotator atr , from 0:50.000 to 1:00.000

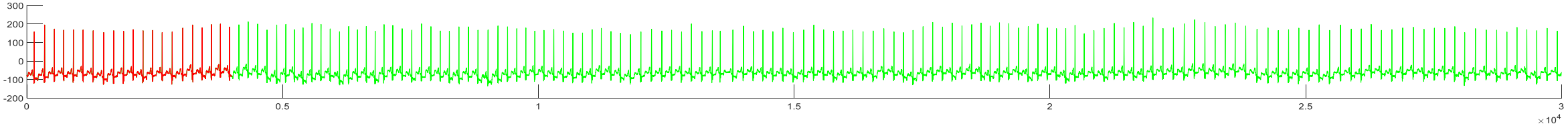
QT Database (qtdb)

This dataset has a single bad heartbeat, an Atrial premature beat between 13400 and 13800.
By most standards, this is a subtle anomaly.

This is the MLII trace







UCR_Anomaly_qtdbSel11005V_4000_12400_12800.txt

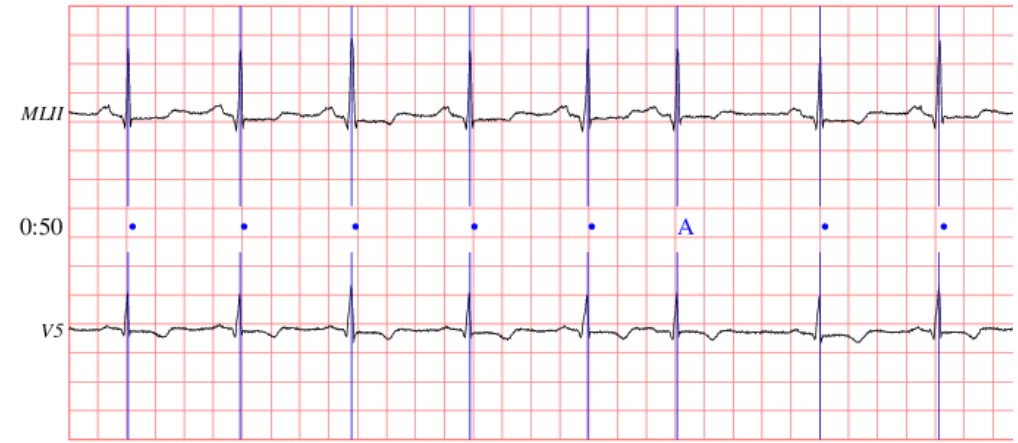
pschart -a atr -c "" -C -E -G -CG 1 .5 .5 -Cs 0 0 0 -H -l -P 300x250 -m 20 20 5 5 -M -n 0 \ -S 4 2 -t 25 -T "" -v 10 -w 0.5 script >chart.ps
convert -density 100x100 chart.ps chart.png

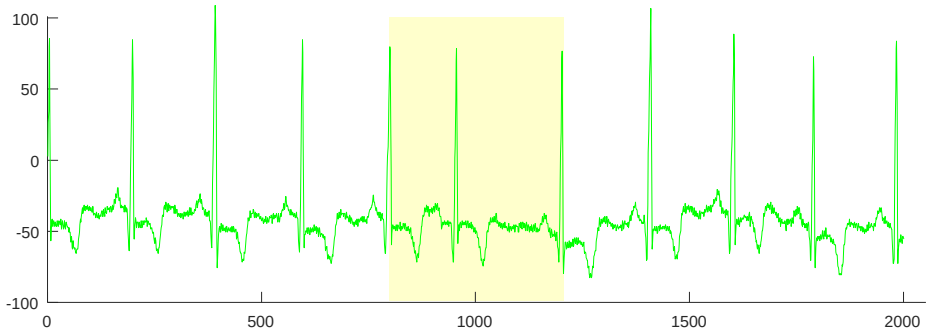
Selected input: record qtdb/sel100 , annotator atr , from 0:50.000 to 1:00.000

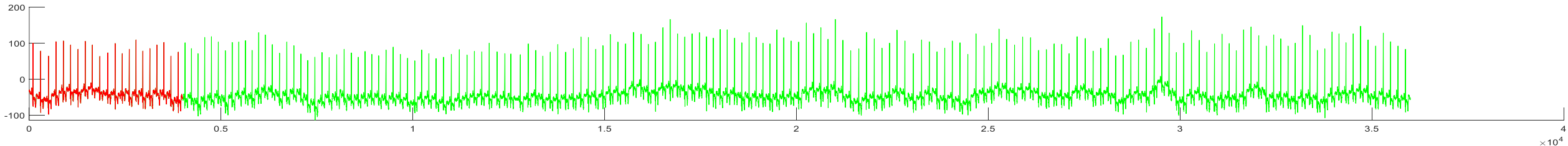
QT Database (qtdb)

This dataset has a single bad heartbeat, an Atrial premature beat between 12400 and 12800.
By most standards, this is a subtle anomaly.

This is the V5 trace. It is shifted by 1000 datapoints from
UCR_Anomaly_qtdbSel100MLII_4000_13400_13800.txt







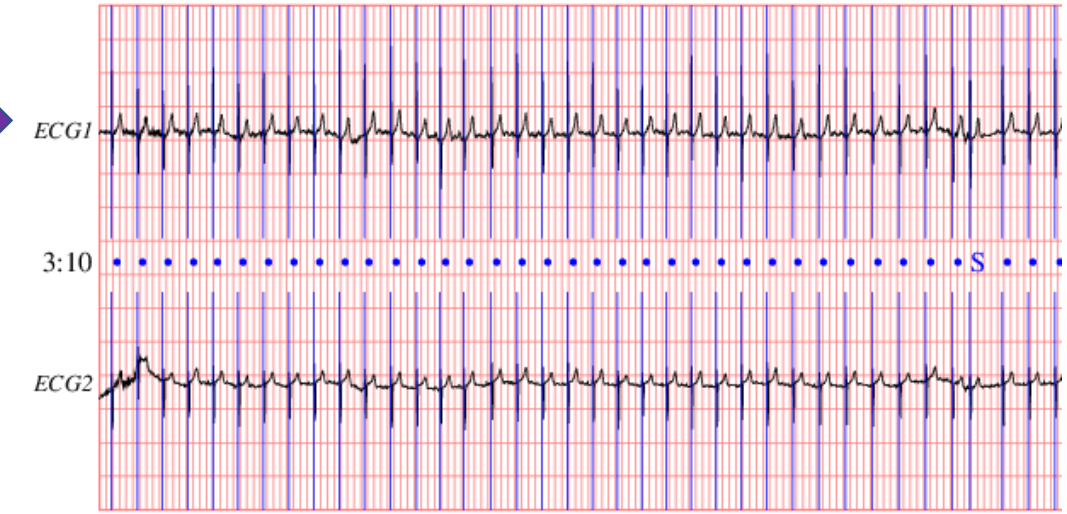
UCR_Anomaly_sel840mECG1_17000_51370_51740.txt

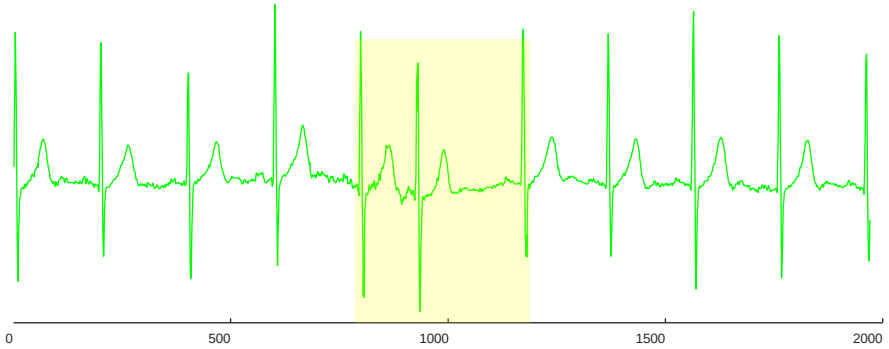
wfdb2mat -r qtdb/sel840 -f 10 -t 899.996 -l s1000000 >sel840m.info

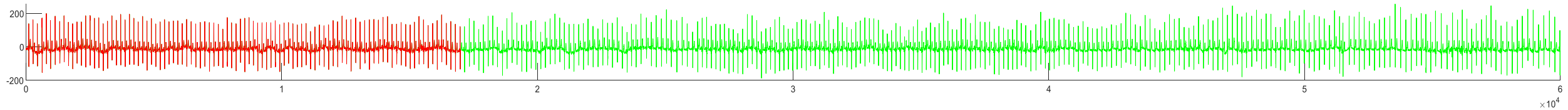
Selected input: record qtdb/sel840 , annotator atr , from 0:10.000 to 14:59.996

QT Database (qtdb)

This dataset has a single bad heartbeat, a supraventricular beat between 51370 and 51740.
This is ECG1
By most standards, this is a subtle anomaly.







UCR_Anomaly_sel840mECG2_20000_49370_49740.txt

wfdb2mat -r qtdb/sel840 -f 10 -t 899.996 -l s1000000 >sel840m.info

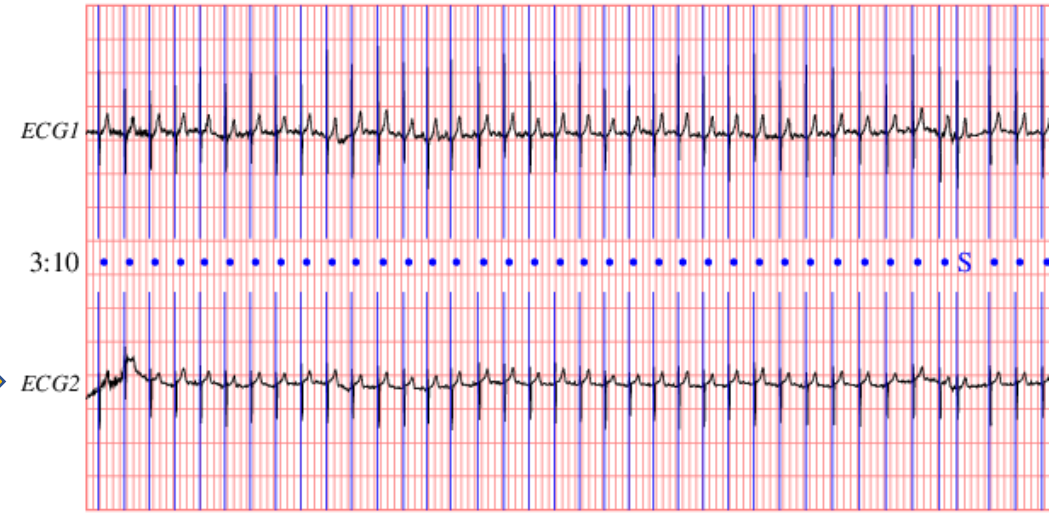
Selected input: record qtdb/sel840 , annotator atr , from 0:10.000 to 14:59.996

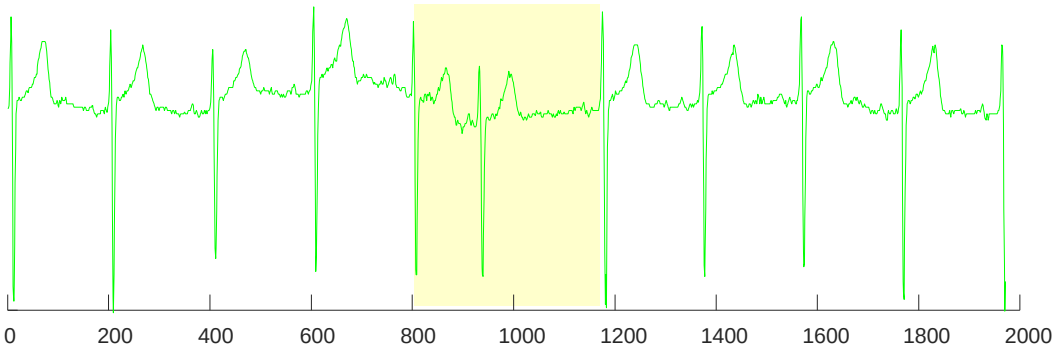
QT Database (qtdb)

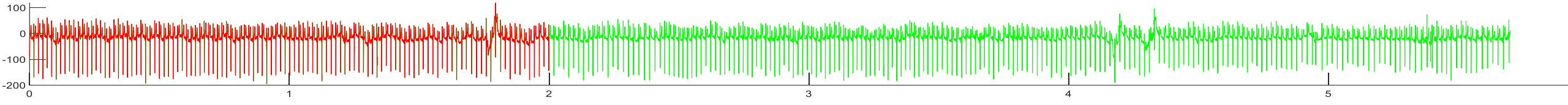
This dataset has a single bad heartbeat, a supraventricular beat between 49370 and 49740.

This is ECG2

This is shifted by 2000 datapoints from `UCR_Anomaly_sel840mECG1_17000_51370_51740.txt`
By most standards, this is a subtle anomaly.





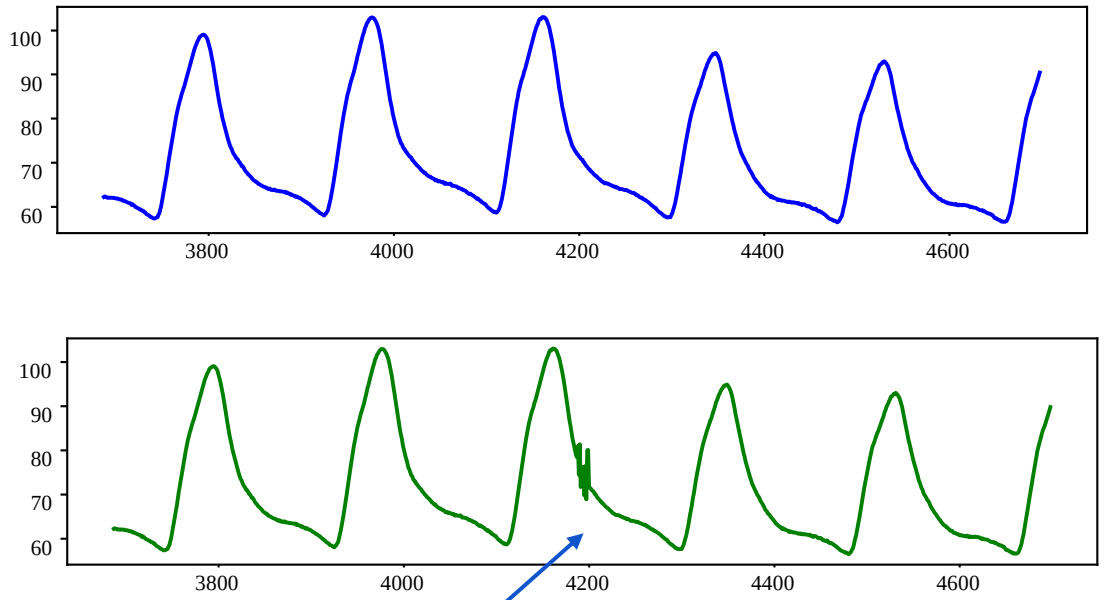


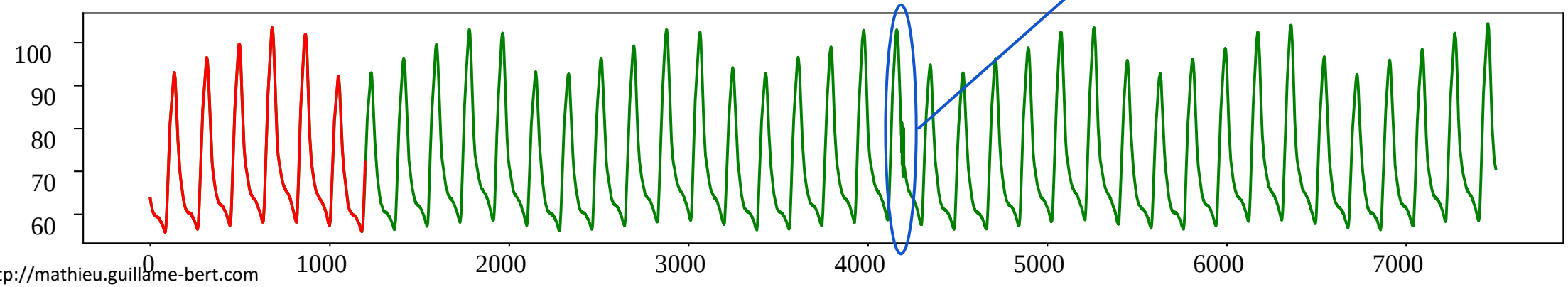
UCR_Anomaly_InternalBleeding16_1200_4187_4199.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We generated random numbers to add noise from 4187 to 4199.

Blue is original data, green is data after anomaly was introduced



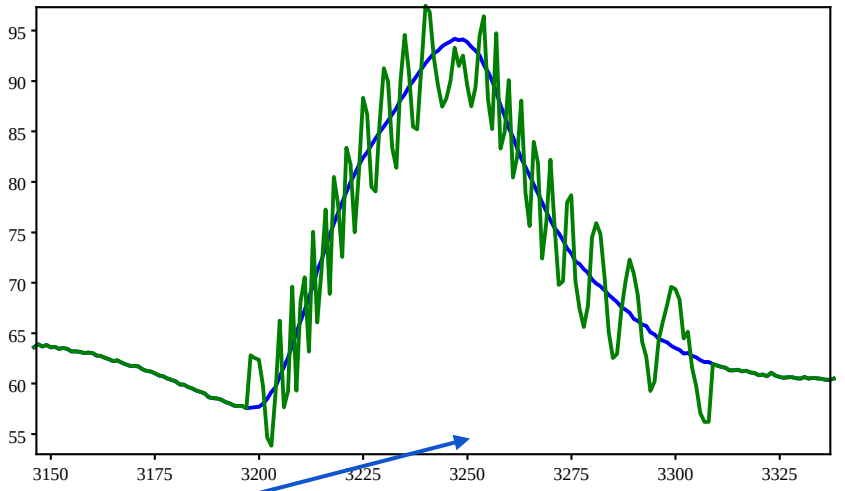


[1]: http://mathieu.guillame-bert.com

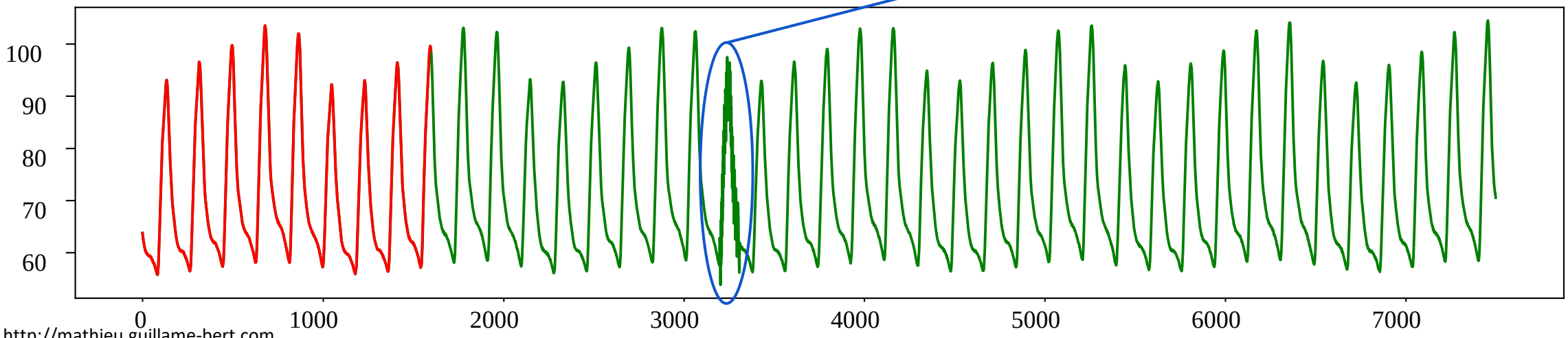
UCR_Anomaly_InternalBleeding17_1600_3198_3309.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

Blue is original data, green is data after anomaly was introduced



The anomaly is synthetic. We added a series of sine waves in one cycle to make a piece of smooth data become fluctuate and rough.

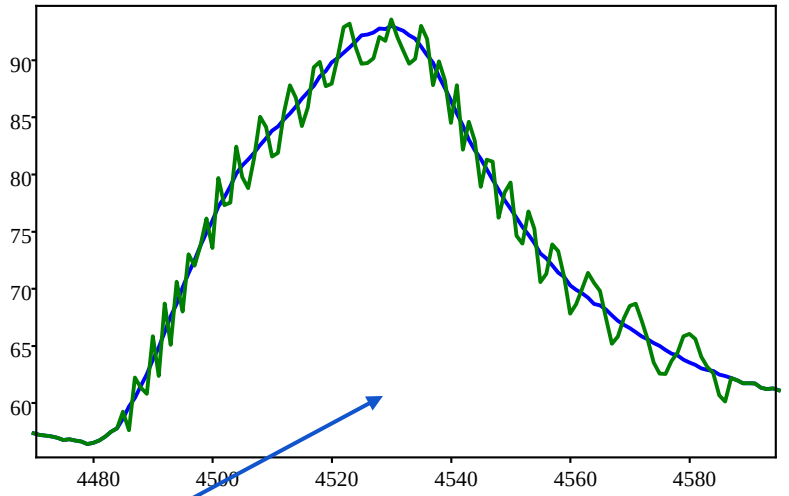


[1]: http://mathieu.guillame-bert.com

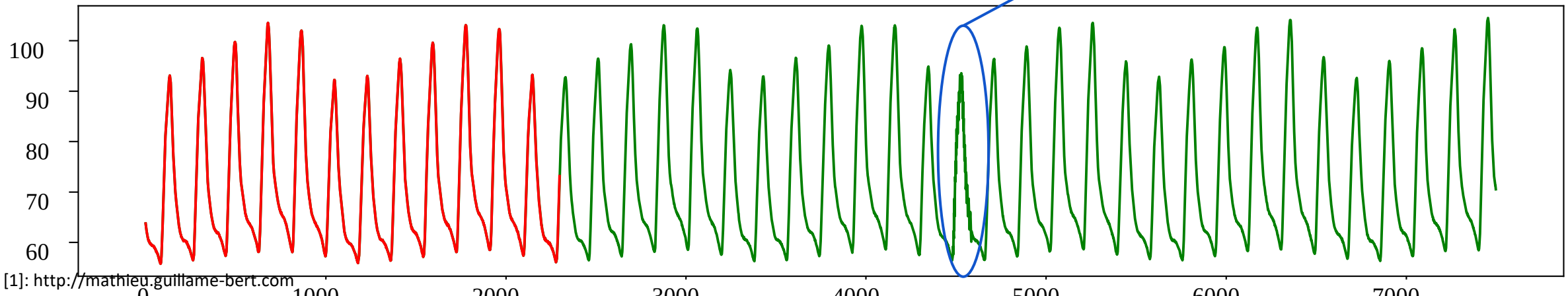
UCR_Anomaly_InternalBleeding18_2300_4485_4587.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

Blue is original data, green is data after anomaly was introduced



The anomaly is synthetic. We added a series of sine waves in one cycle to make a piece of smooth data become fluctuate and rough. This is a more subtle version of UCR_Anomaly_InternalBleeding16_1200_4187_4199.txt.



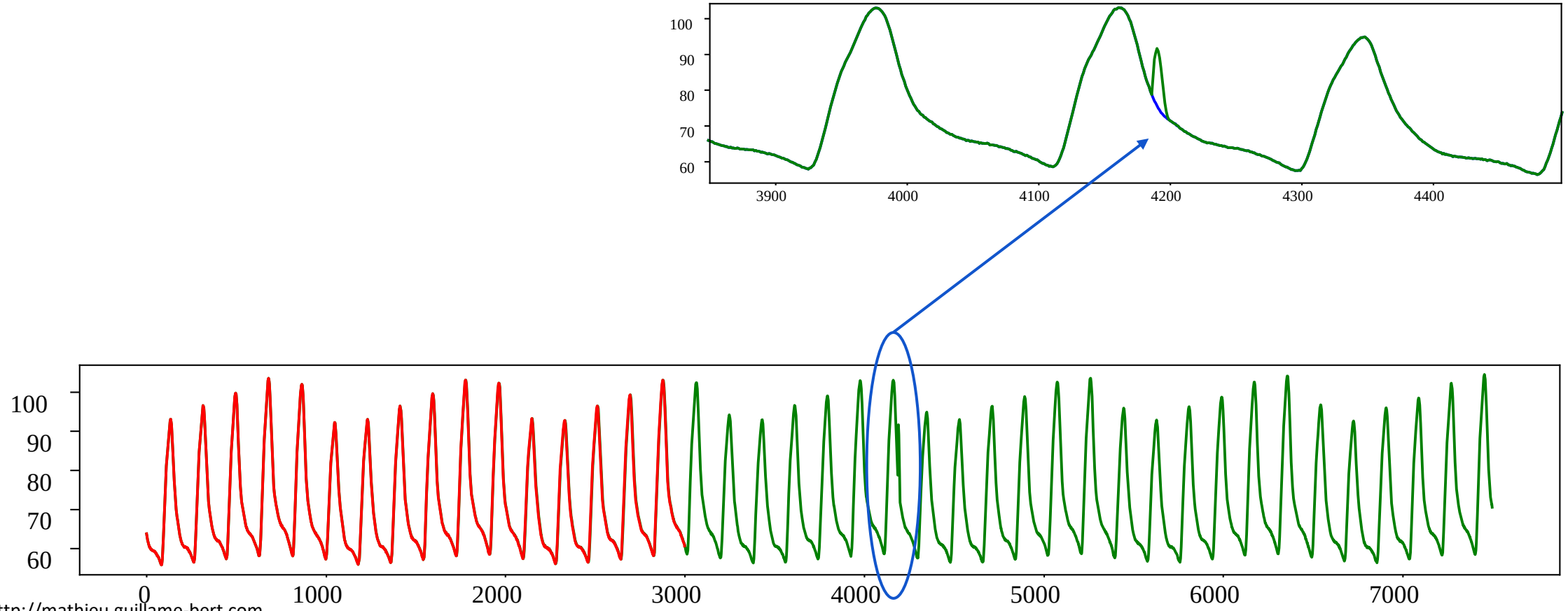
[1]: http://mathieu.guillame-bert.com

UCR_Anomaly_InternalBleeding19_3000_4187_4197.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We added a secondary peak to a cycle of data that has only one peak.

Blue is original data, green is data after anomaly was introduced



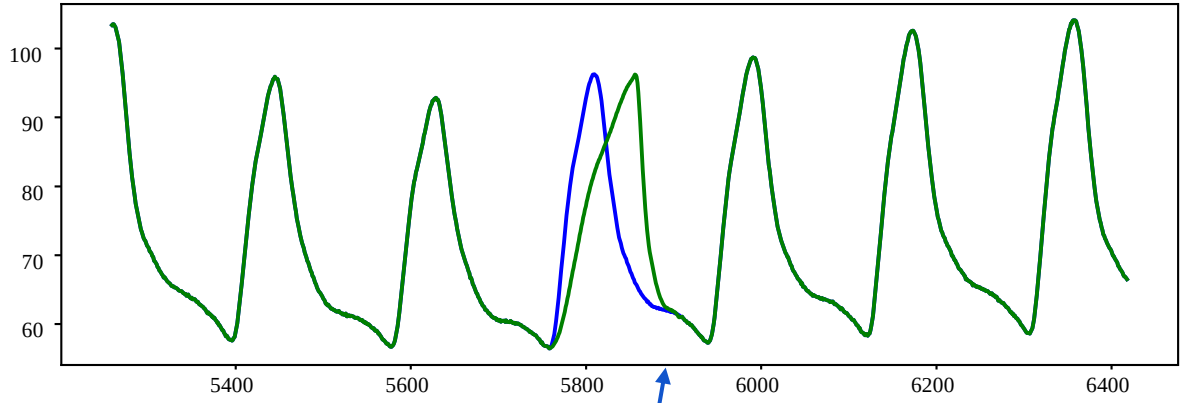
[1]: http://mathieu.guillame-bert.com

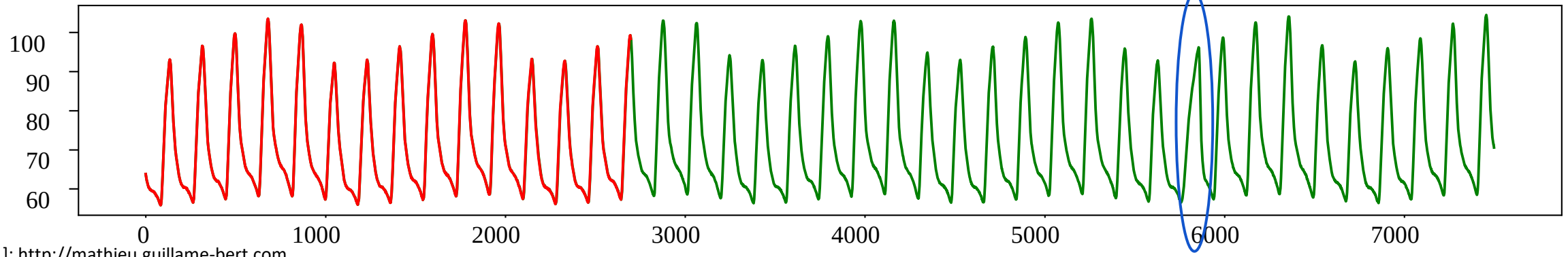
UCR_Anomaly_InternalBleeding20_2700_5759_5919.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We performed resampling (upsampling + downsampling) operations to skew the data in one cycle.

Blue is original data, green is data after anomaly was introduced



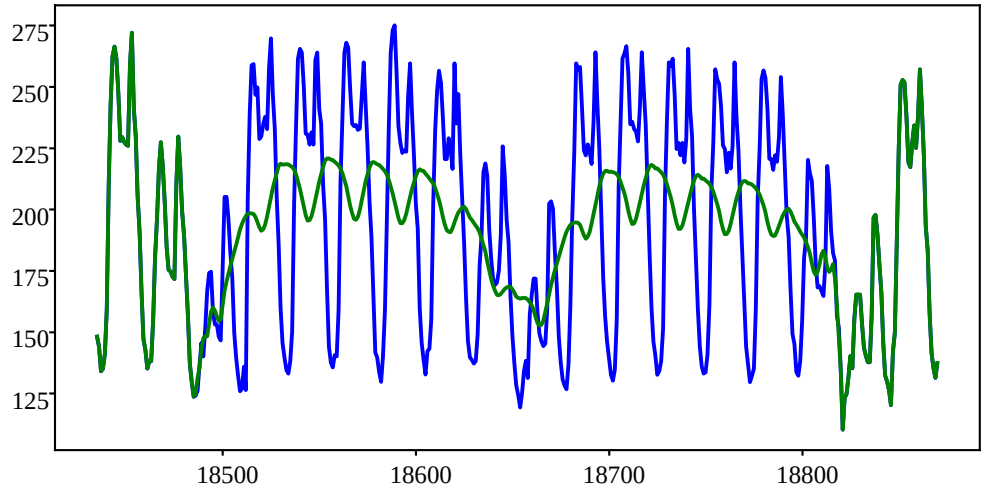


[1]: http://mathieu.guillame-bert.com

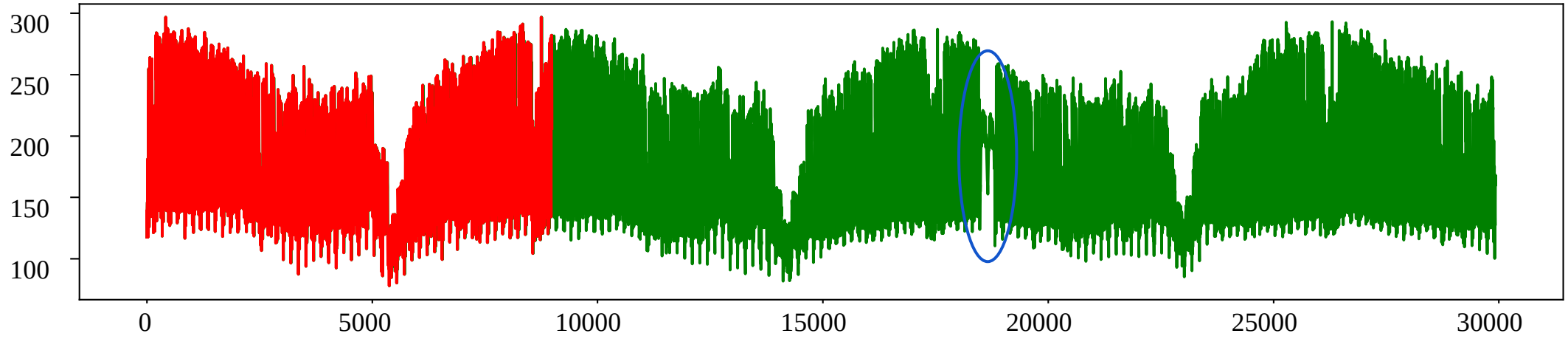
UCR_Anomaly_PowerDemand1_9000_18485_18821.txt

The data comes from an Italian power demand dataset. (from 1/1/1995 to 5/31/1998)

The anomaly is synthetic. We ran the moving average algorithm on random two-week power supply data (2/9/1997 3:00 - 2/23/1997 3:00).



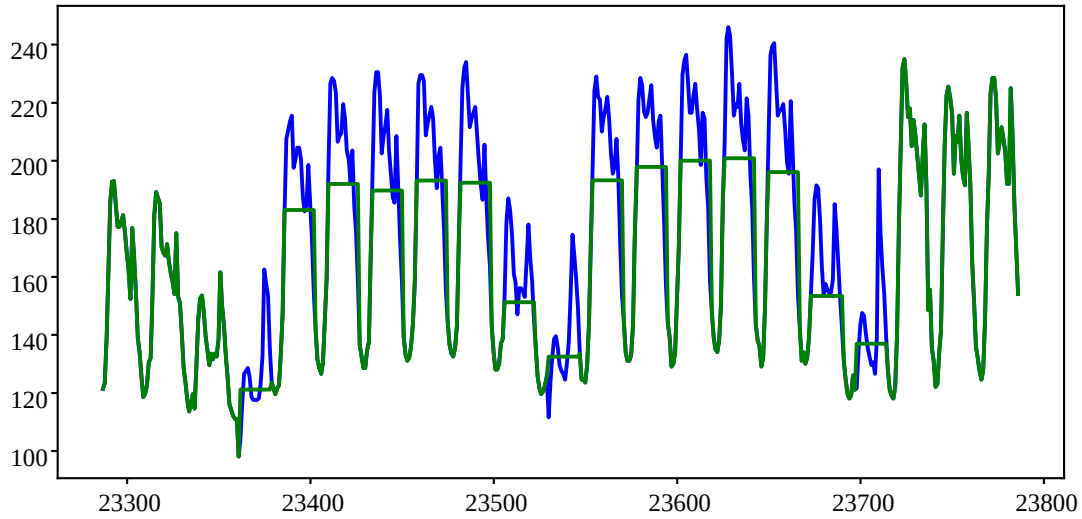
Blue is original data, green is data after anomaly was introduced



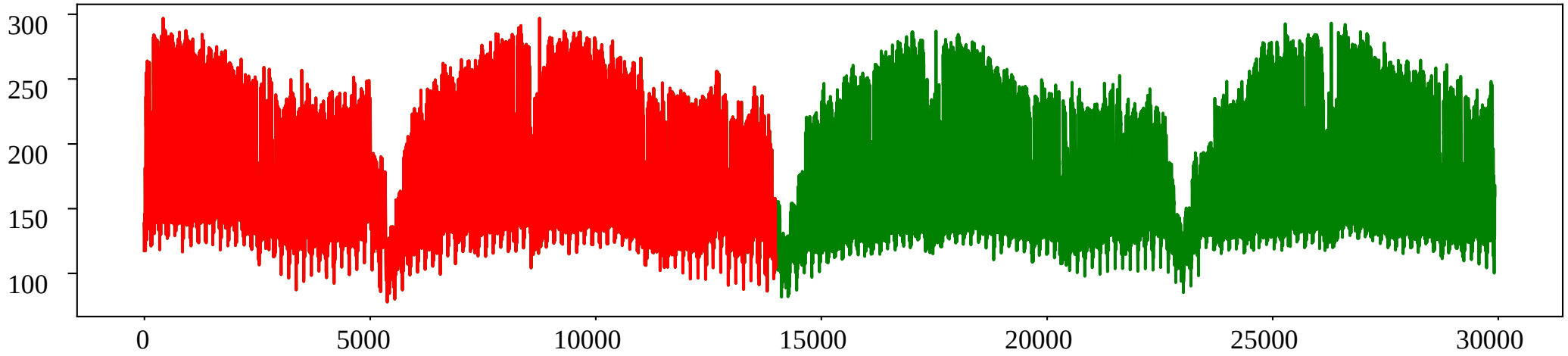
UCR_Anomaly_PowerDemand2_14000_23357_23717.txt

The data comes from an Italian power demand dataset. (from 1/1/1995 to 5/31/1998)

The anomaly is synthetic. We created an anomaly by removing peaks and adding constant regions for half a month's power supply data. (8/31/1997 - 9/15/1997)

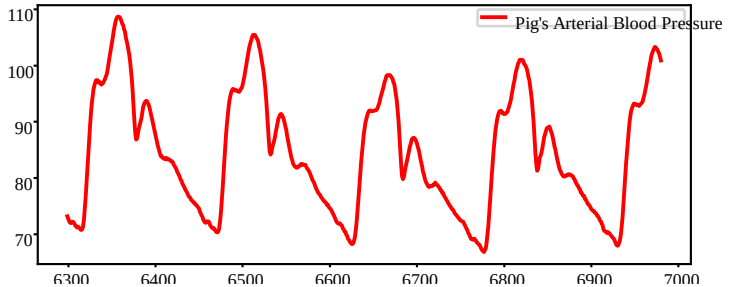


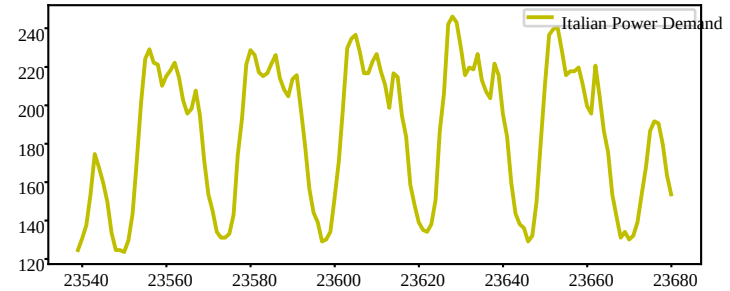
Blue is original data, green is data after anomaly was introduced



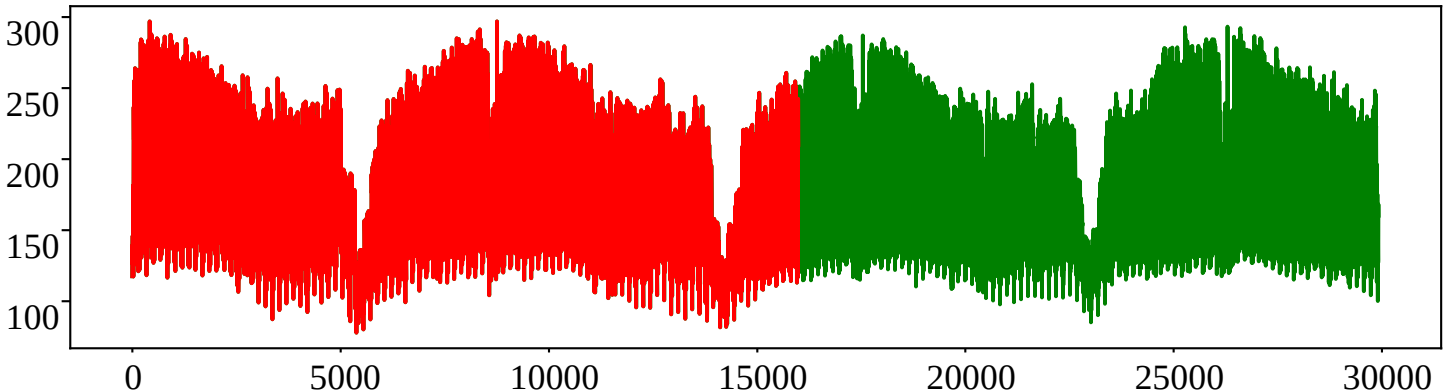
UCR_Anomaly_PowerDemand3_16000_23405_23477.txt

The data comes from an Italian power demand dataset. (from 1/1/1995 to 5/31/1998) The red curve indicates the pig arterial blood pressure, and the yellow curve indicates the Italian power demand. These two curves look similar.

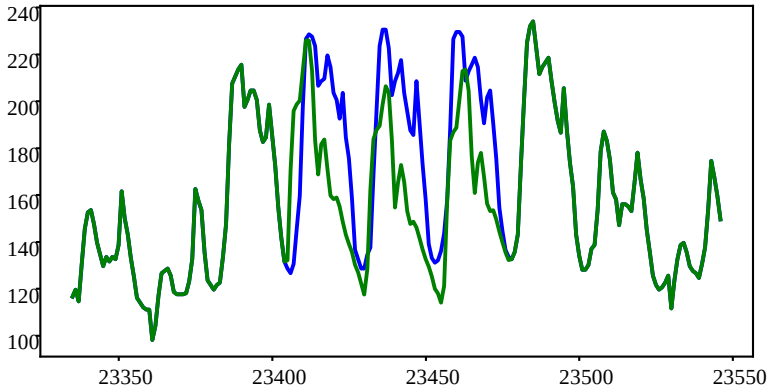




The anomaly is synthetic. We replaced random three-day power supply data with a segment of arterial blood pressure data. (The X of arterial blood pressure data is scaled by resampling and the Y is adjusted by a level shift.)

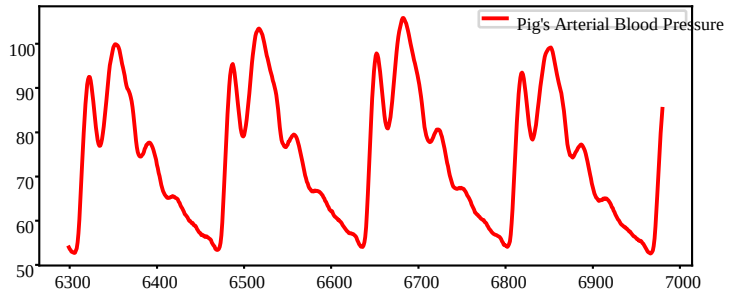


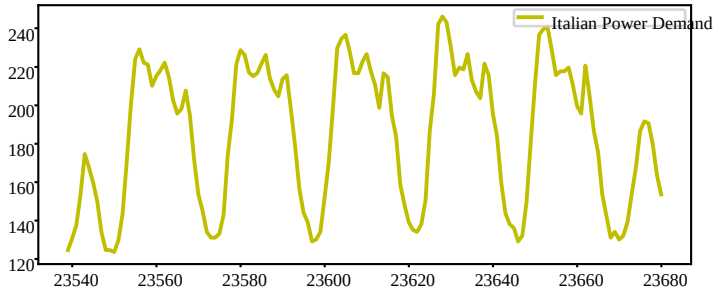
Blue is original data, green is data after anomaly was introduced



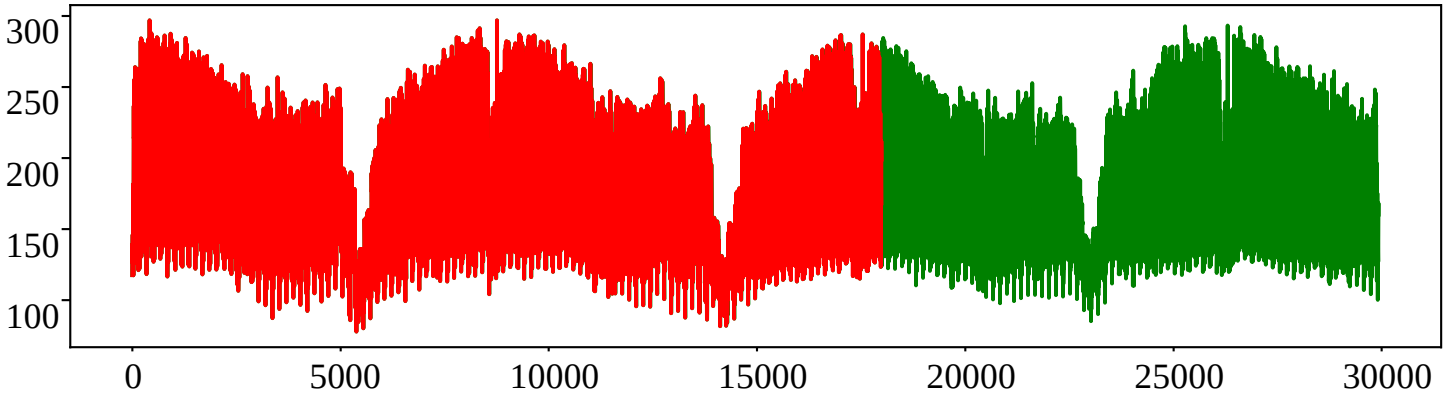
UCR_Anomaly_PowerDemand4_18000_24005_24077.txt

The data comes from an Italian power demand dataset. (from 1/1/1995 to 5/31/1998) The red curve indicates the pig arterial blood pressure, and the yellow curve indicates the Italian power demand. These two curves look similar.

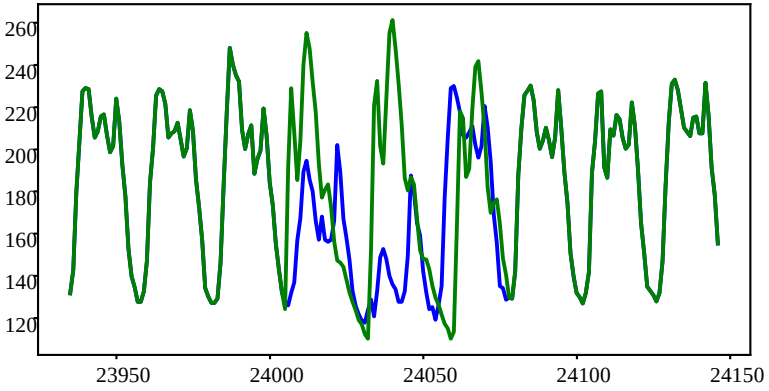




The anomaly is synthetic. We replaced random three-day power supply data with a segment of arterial blood pressure data. (The X of arterial blood pressure data is scaled by resampling and the Y is adjusted by a level shift.)



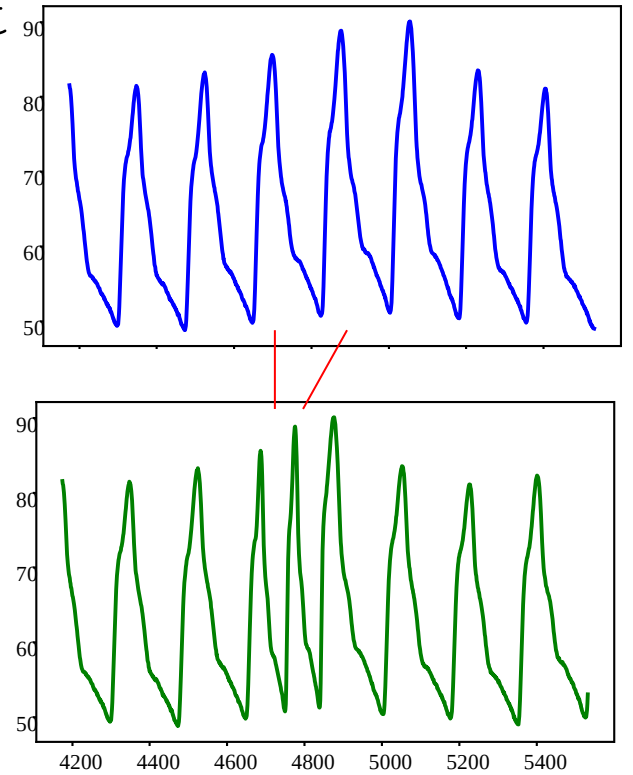
Blue is original data, green is data after anomaly was introduced



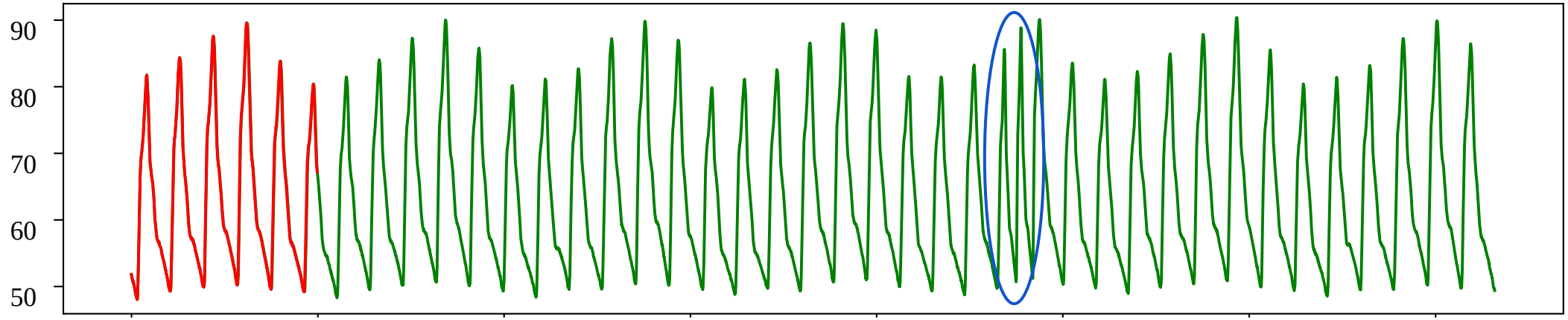
UCR_Anomaly_InternalBleeding4_1000_4675_5033.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We performed a downsampling operation on a short fraction of normal arterial blood pressure data. From 4675 to 5033, the values at every two time points are combined by averaging.



Blue is original data, green is data after anomaly was introduced

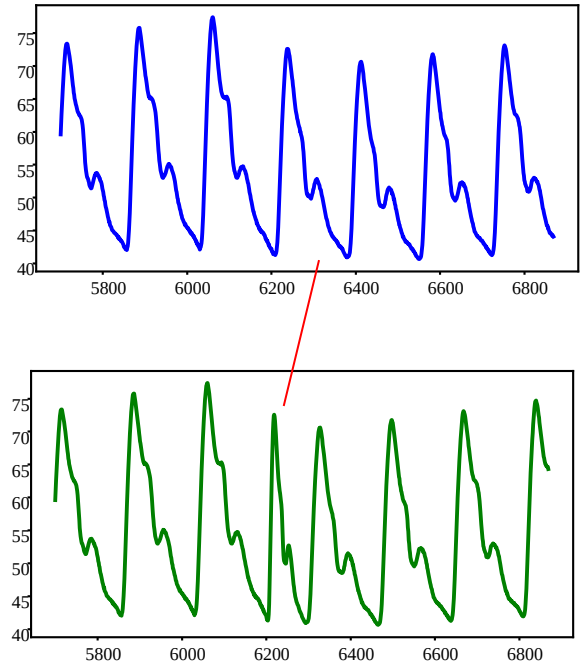


[1]: http://mathieu.guillame-bert.com

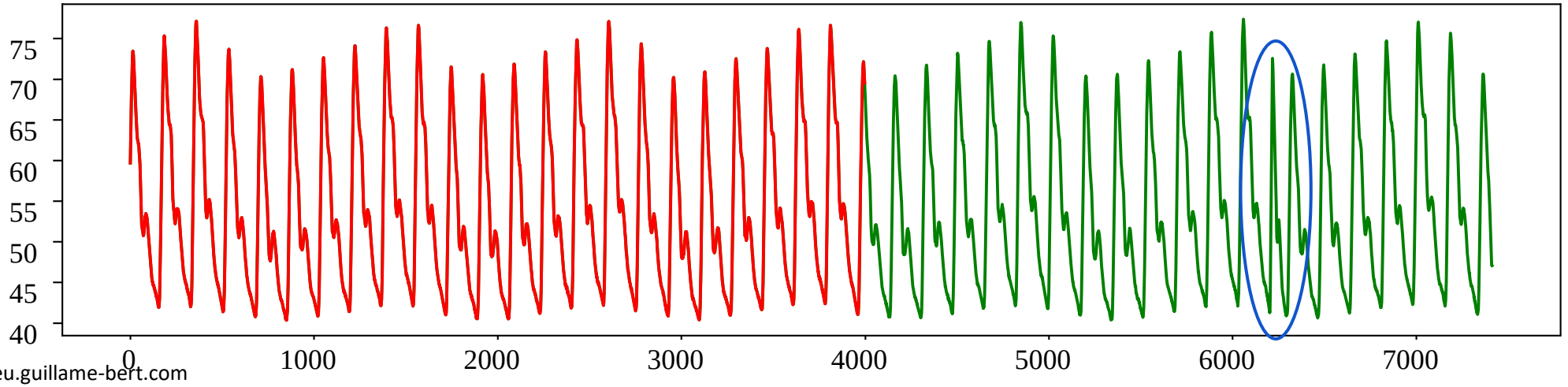
UCR_Anomaly_InternalBleeding5_4000_6200_6370.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We performed a downsampling operation on a short fraction of normal arterial blood pressure data. From 6200 to 6370, the values at every two time points are combined by averaging.
This is a more subtle version of UCR_Anomaly_InternalBleeding4_1000_4675_5033.txt



Blue is original data, green is data after anomaly was introduced

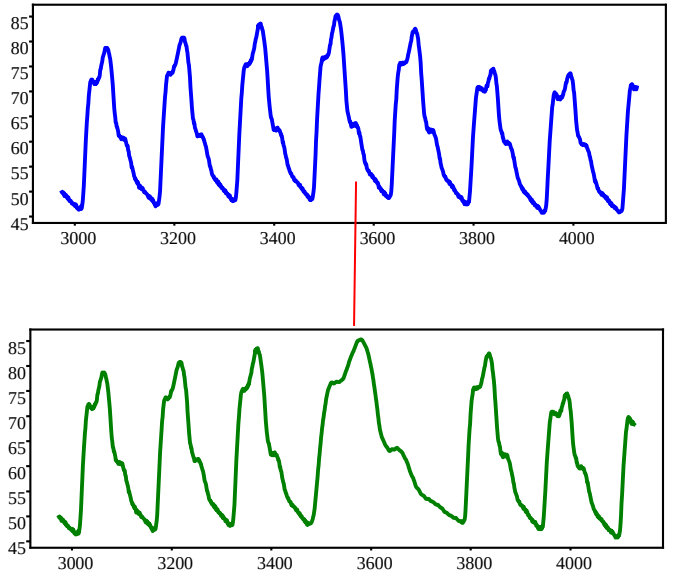


[1]: http://mathieu.guillame-bert.com

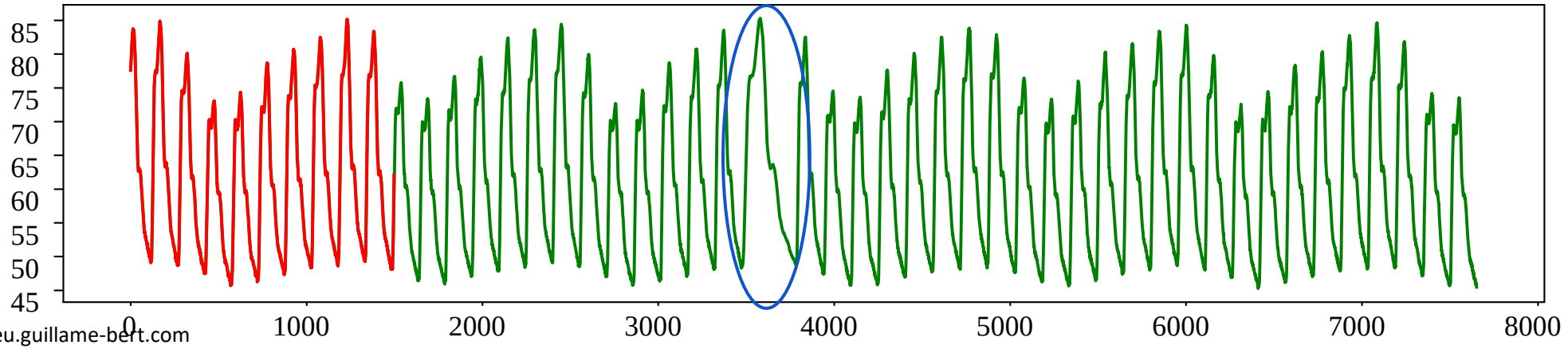
UCR_Anomaly_InternalBleeding6_1500_3474_3629.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We performed an upsampling operation on a short fraction of normal arterial blood pressure data. From 3393 to 3570, the frequency of the time series is doubled, and the missing data is filled with the value of the previous time point.



Blue is original data, green is data after anomaly was introduced

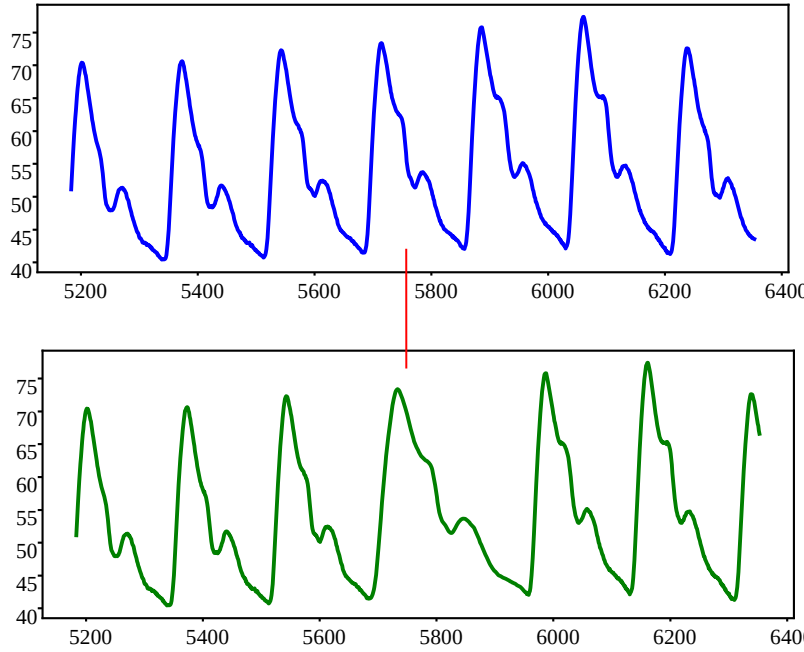


[1]: http://mathieu.guillame-bert.com

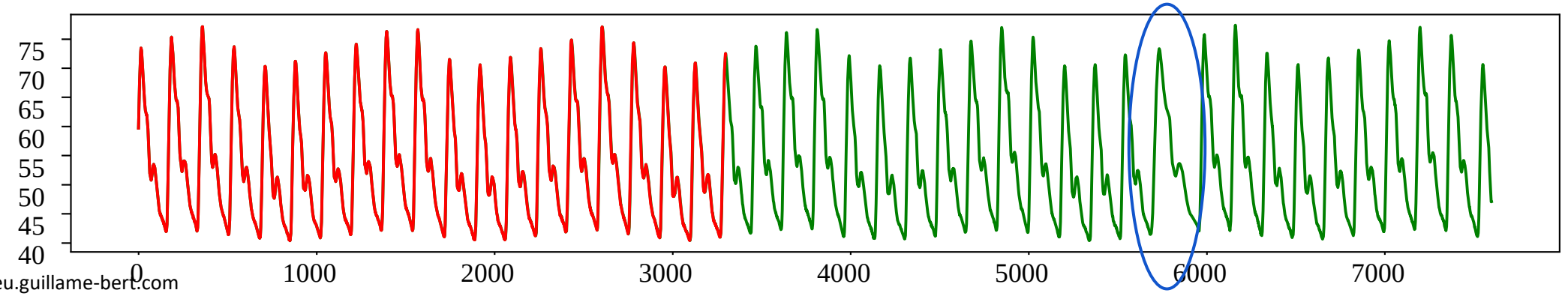
UCR_Anomaly_InternalBleeding15_1700_5684_5854.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We performed an upsampling operation on a short fraction of normal arterial blood pressure data. From 5684 to 5854, the frequency of time series is 1.6 times higher than before, and the missing data is filled with the value of the previous time point. This is a more subtle version of UCR_Anomaly_InternalBleeding6_1500_3474_3629.txt.



Blue is original data, green is data after anomaly was introduced

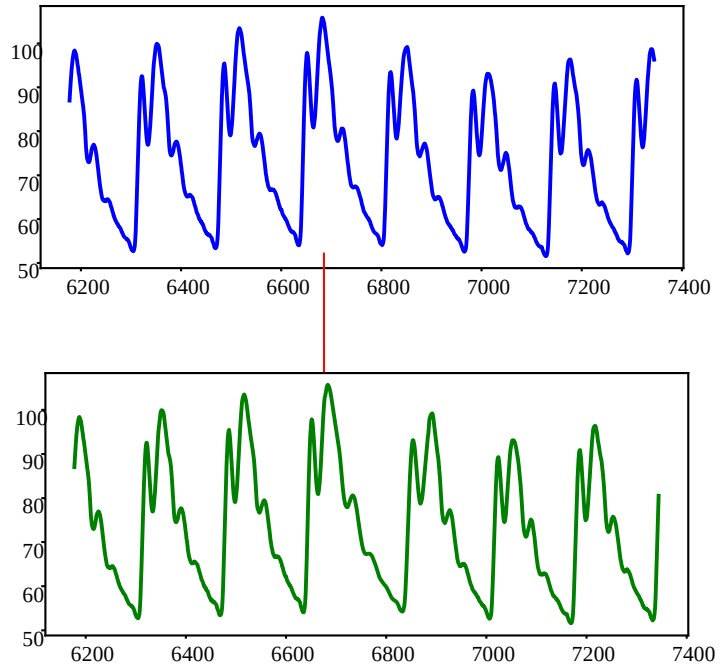


[1]: http://mathieu.guillame-bert.com

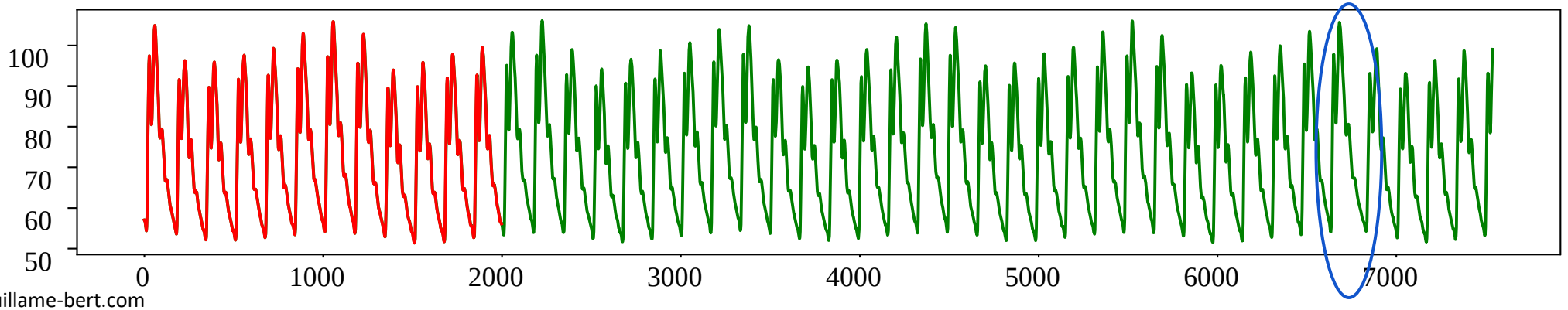
UCR_Anomaly_InternalBleeding7_2000_6678_6846.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We performed an upsampling operation on a short fraction of normal arterial blood pressure data. From 6678 to 6846, the frequency of time series is 1.25 times higher than before, and the missing data is filled with the value of the previous time point. This is a more subtle version of UCR_Anomaly_InternalBleeding6_1500_3474_3629.txt.



Blue is original data, green is data after anomaly was introduced

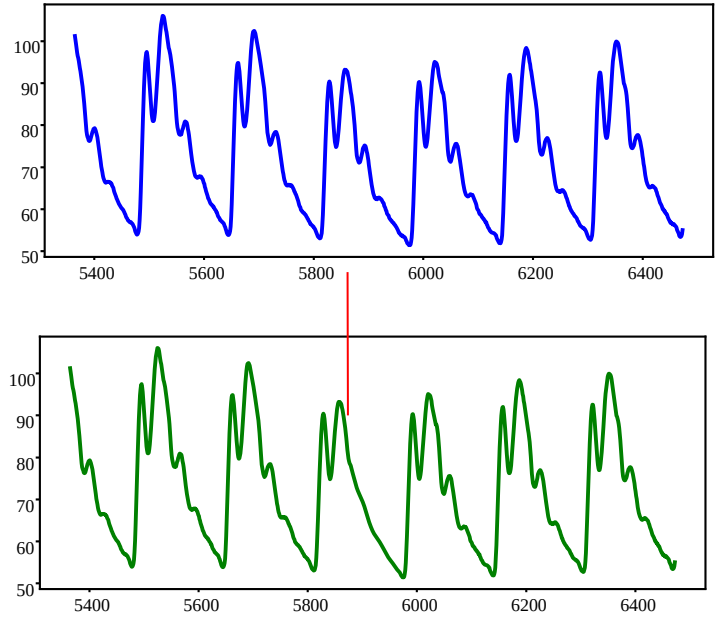


[1]: http://mathieu.guillame-bert.com

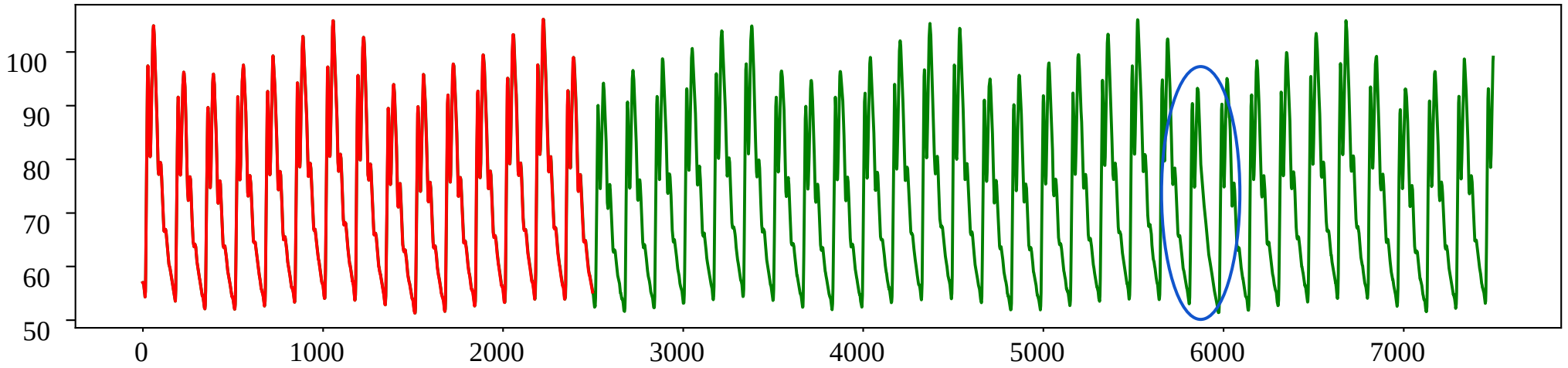
UCR_Anomaly_InternalBleeding8_2500_5865_5974.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We ran the moving average algorithm on a short fraction of normal arterial blood pressure data to eliminate about 4 extreme points within one cycle.



Blue is original data, green is data after anomaly was introduced

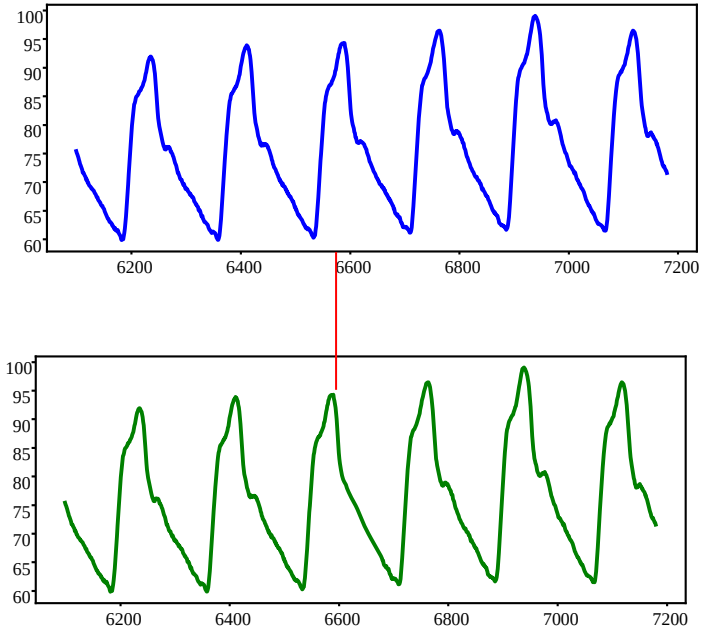


[1]: http://mathieu.guillame-bert.com

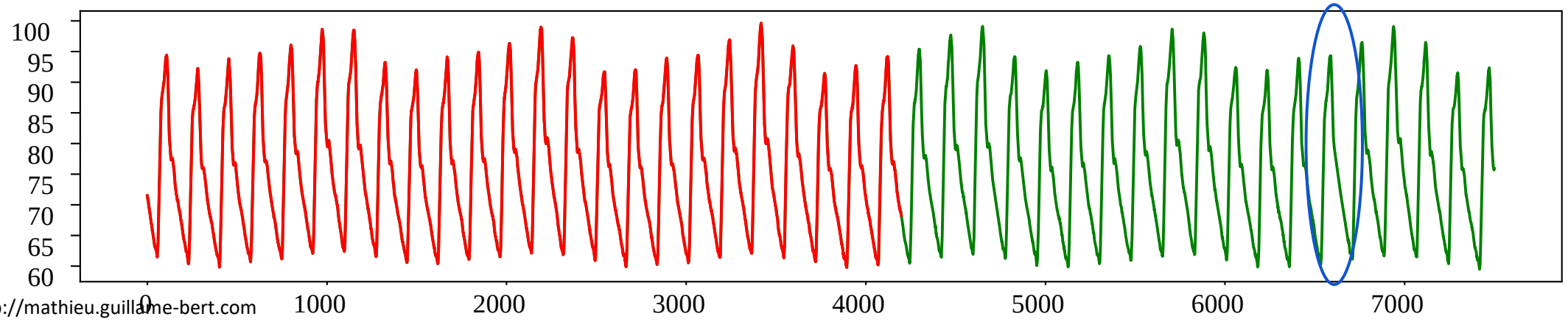
UCR_Anomaly_InternalBleeding9_4200_6599_6681.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We ran the moving average algorithm on a short fraction of normal arterial blood pressure data to eliminate about 2 extreme points within one cycle.



Blue is original data, green is data after anomaly was introduced



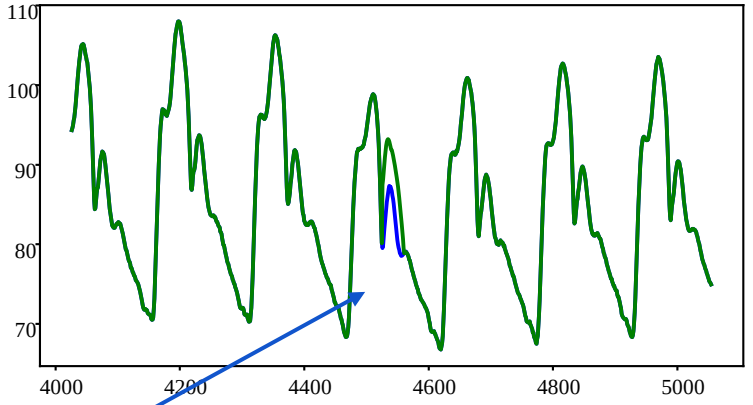
[1]: http://mathieu.guillame-bert.com

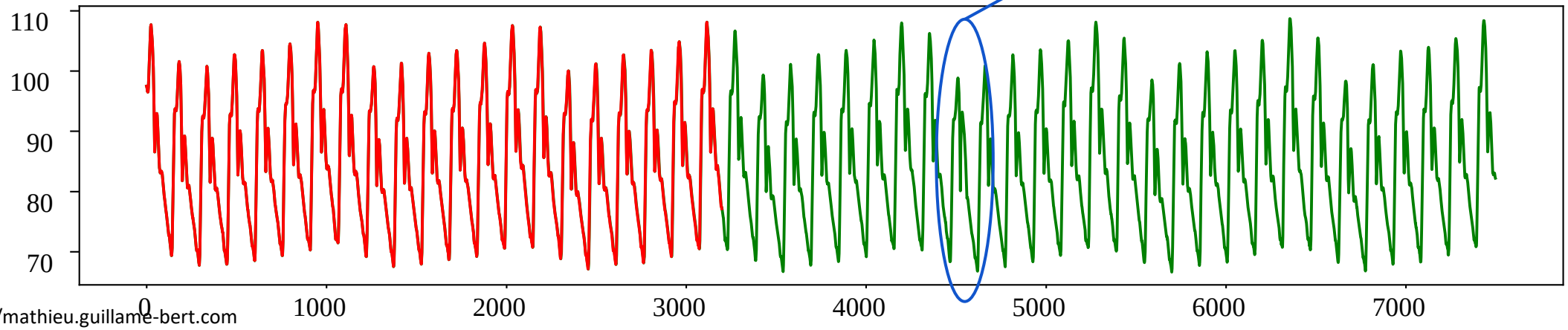
UCR_Anomaly_InternalBleeding10_3200_4526_4556.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We increased the amplitude of on a short fraction of normal arterial blood pressure data to 110% of the original to make the secondary peak in one cycle greater.

Blue is original data, green is data after anomaly was introduced





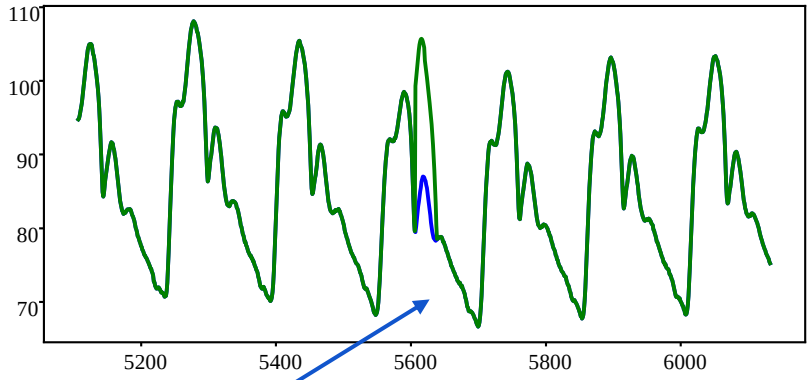
[1]: http://mathieu.guillame-bert.com

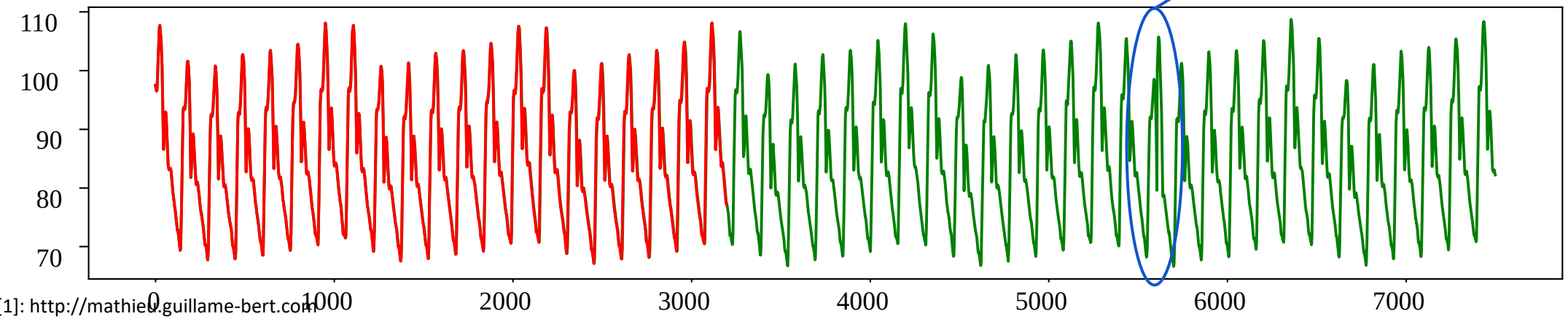
UCR_Anomaly_InternalBleeding14_2800_5607_5634.txt

The data comes from an internal bleeding dataset[1]. From the dataset, we extracted the arterial blood pressure measurements of pigs.

The anomaly is synthetic. We increased the amplitude of on a short fraction of normal arterial blood pressure data to 125% of the original to make the secondary peak in one cycle greater than the main peak.

Blue is original data, green is data after anomaly was introduced



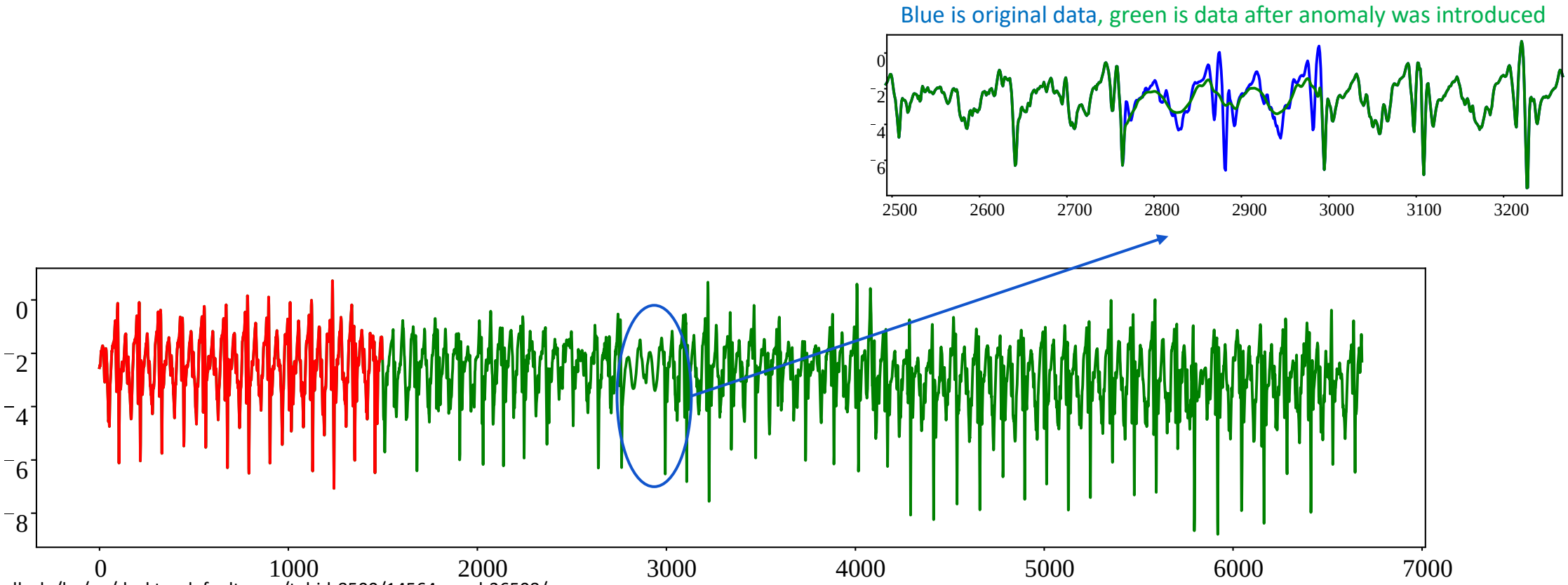


[1]: http://mathieu.guillame-bert.com

UCR_Anomaly_WalkingAceleration1_1500_2764_2995.txt

The dataset is built based on an activity recognition system benchmark data[2]. We extracted an acceleration data measured by the sensor when an experimental subject is walking. Since the acceleration is 3-dimensional, we obtained our data by averaging the values in x, y and z axes.

The anomaly is synthetic. We first merged all data for the walking activity, and ran the moving average algorithm on a small fraction of the merged walking data (about two cycles).

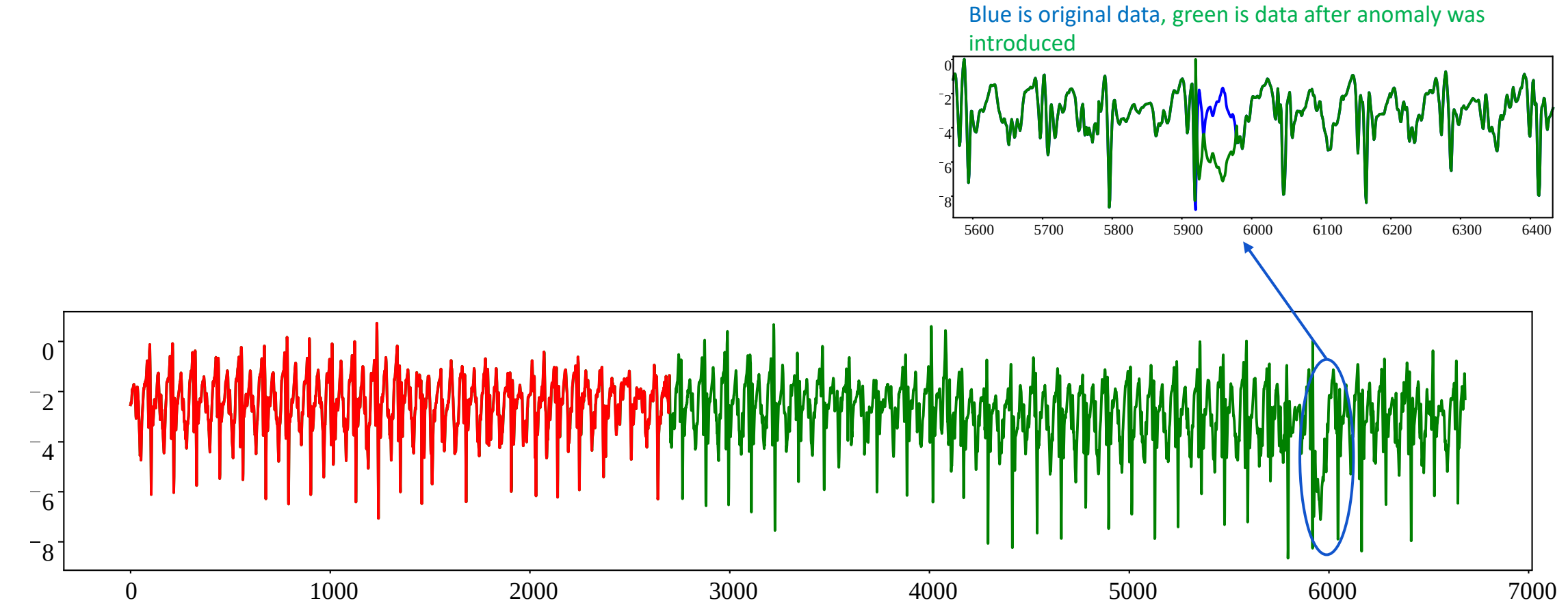


[2]: https://www.dlr.de/kn/en/desktopdefault.aspx/tabid-8500/14564_read-36508/

UCR_Anomaly_WalkingAceleration5_2700_5920_5979.txt

The dataset is built based on an activity recognition system benchmark data[2]. We extracted an acceleration data measured by the sensor when an experimental subject is walking. Since the acceleration is 3-dimensional, we obtained our data by averaging the values in x, y and z axes.

We first merged all data for the walking activity. An anomaly is created by turning a half cycle's data upside down.



[2]: https://www.dlr.de/kn/en/desktopdefault.aspx/tabid-8500/14564_read-36508/

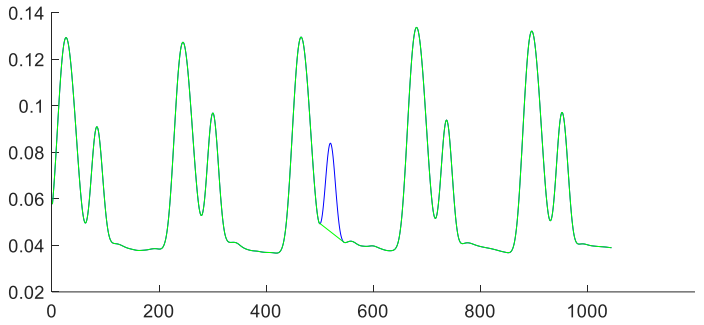
UCR_Anomaly_GP711MarkerLFM5z1_5000_6168_6212.txt

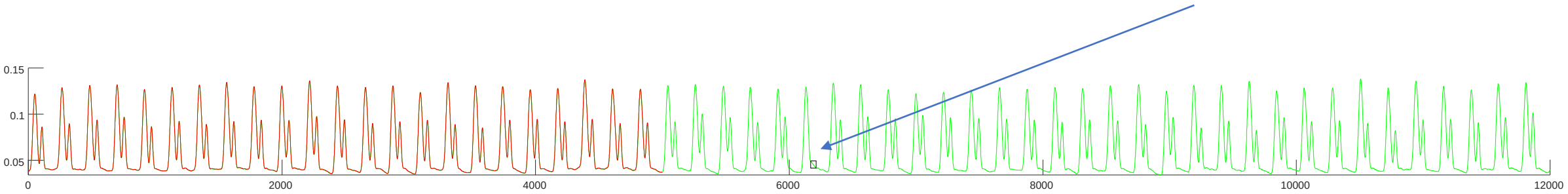
The data comes from subject 7 in the GaitPhase Database [1]. The subject was required to walk on a split-belt treadmill at walking speed of 1.1 m/s.
This dataset represents the vertical direction of a 3D marker's position. The marker was placed on subject's left shoe above the second metatarsal head.
There are totally 55 gait cycles in the dataset.

The anomaly is synthetic. For a random gait cycle, we created the anomaly by removing the second peak.

```
T(start_anomaly:end_anomaly) = interp1([start_anomaly end_anomaly],
    [T(start_anomaly) T(end_anomaly)], start_anomaly:end_anomaly);
```

Blue is original data, green is data after anomaly was introduced





[1] https://www.mad.tf.fau.de/research/activitynet/gaitphase-database/

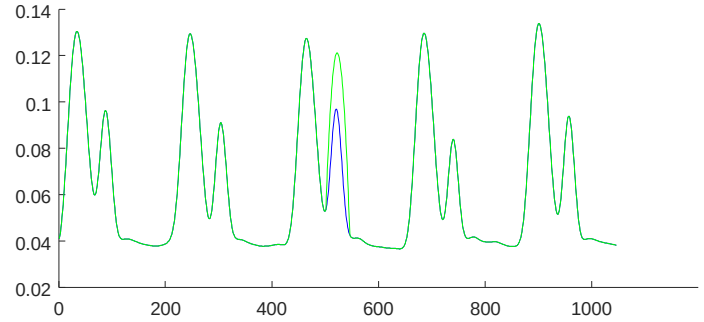
UCR_Anomaly_GP711MarkerLFM5z2_5000_5948_5993.txt

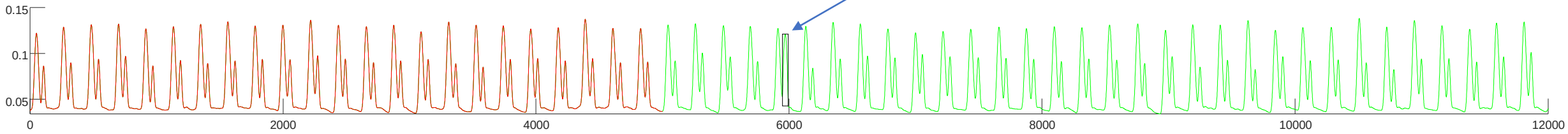
The data comes from subject 7 in the GaitPhase Database [1]. The subject was required to walk on a split-belt treadmill at walking speed of 1.1 m/s.
This dataset represents the vertical direction of a 3D marker's position. The marker was placed on subject's left shoe above the second metatarsal head.
There are totally 55 gait cycles in the dataset.

The anomaly is synthetic. For a random gait cycle, we made the second peak much higher, almost as high as the main peak.

```
[max_val,max_locidx] = max(T(start_anomaly:end_anomaly));
T(start_anomaly:end_anomaly) = interp1(
    [start_anomaly start_anomaly+max_locidx end_anomaly],
    [T(start_anomaly) max_val*1.25 T(end_anomaly)],
    start_anomaly:end_anomaly, 'pchip');
```

Blue is original data, green is data after anomaly was introduced





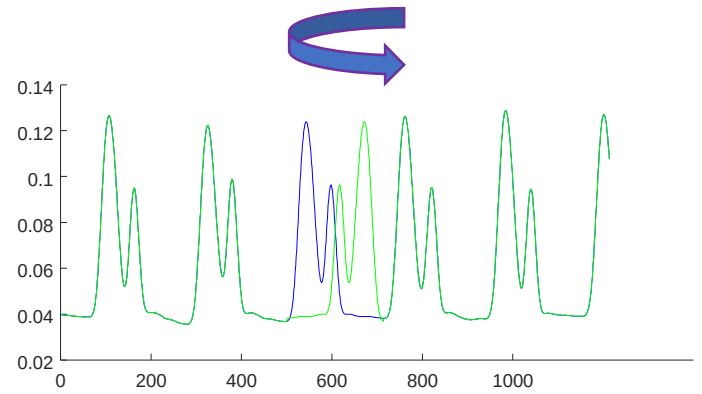
[1] https://www.mad.tf.fau.de/research/activitynet/gaitphase-database/

UCR_Anomaly_GP711MarkerLFM5z3_5000_7175_7388.txt

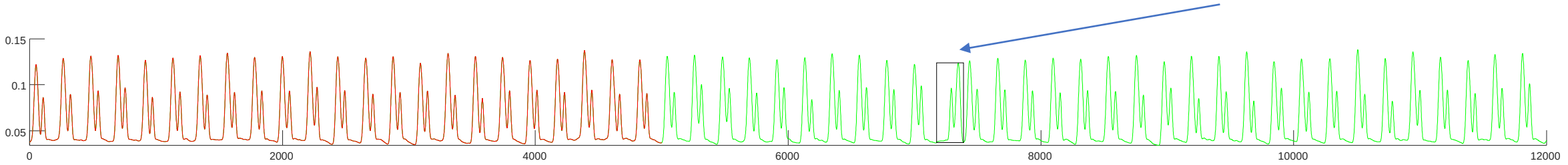
The data comes from subject 7 in the GaitPhase Database [1]. The subject was required to walk on a split-belt treadmill at walking speed of 1.1 m/s.
This dataset represents the vertical direction of a 3D marker's position. The marker was placed on subject's left shoe above the second metatarsal head.
There are totally 55 gait cycles in the dataset.

The anomaly is synthetic. We created the anomaly by reversing the direction of a random gait cycle.

```
T(start_anomaly:end_anomaly) = flip(T(start_anomaly:end_anomaly));
```



Blue is original data, green is data after anomaly was introduced



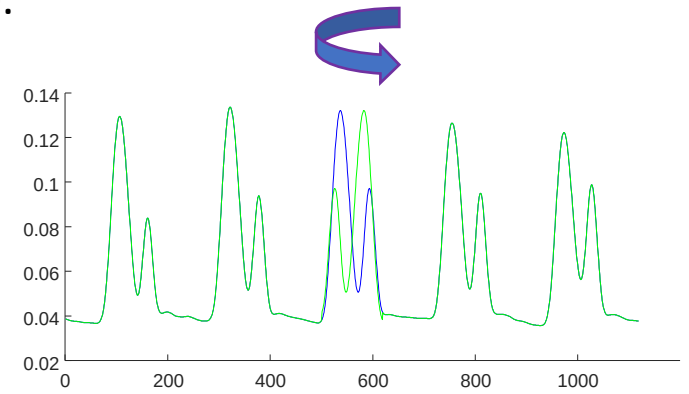
[1] https://www.mad.tf.fau.de/research/activitynet/gaitphase-database/

UCR_Anomaly_GP711MarkerLFM5z4_5000_6527_6645.txt

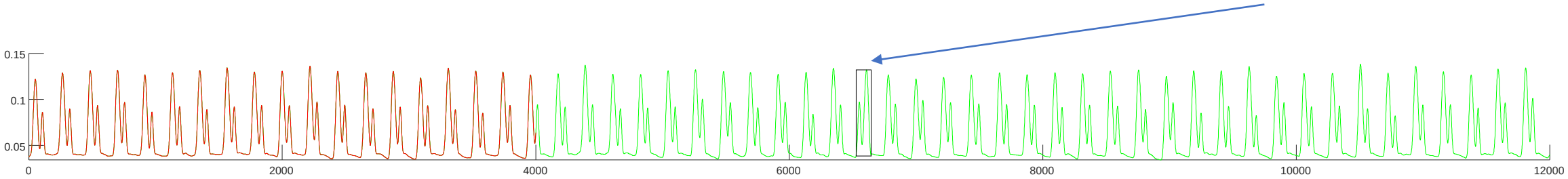
The data comes from subject 7 in the GaitPhase Database [1]. The subject was required to walk on a split-belt treadmill at walking speed of 1.1 m/s.
This dataset represents the vertical direction of a 3D marker's position. The marker was placed on subject's left shoe above the second metatarsal head.
There are totally 55 gait cycles in the dataset.

This is a subtle version of `UCR_Anomaly_GP711MarkerLFM5z3_5000_7175_7388.txt`.
Instead of reversing the whole gait cycle, we only flipped the two peaks in the cycle.

```
T(start_anomaly:end_anomaly) = flip(T(start_anomaly:end_anomaly));
```



Blue is original data, green is data after anomaly was introduced



[1] https://www.mad.tf.fau.de/research/activitynet/gaitphase-database/

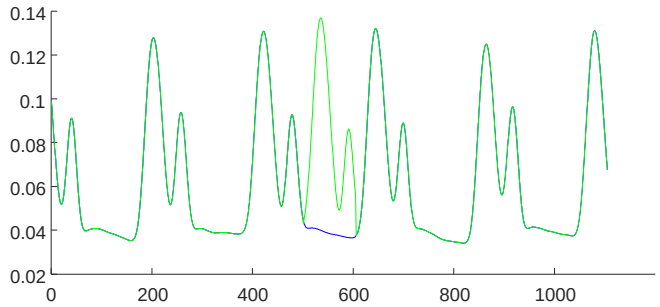
UCR_Anomaly_GP711MarkerLFM5z5_5000_8612_8716.txt

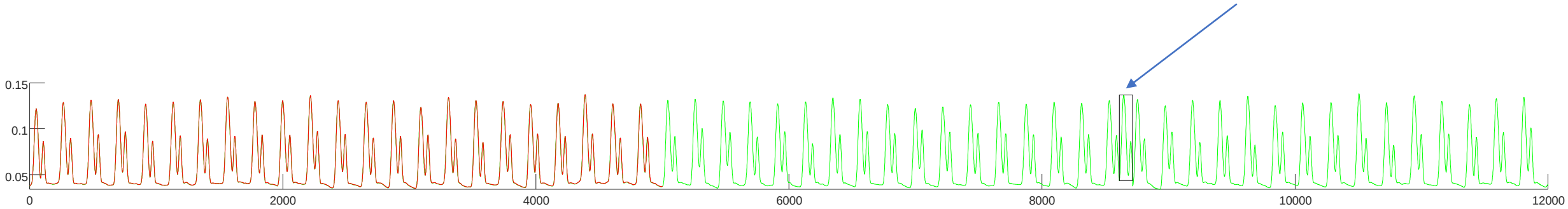
The data comes from subject 7 in the GaitPhase Database [1]. The subject was required to walk on a split-belt treadmill at walking speed of 1.1 m/s.
This dataset represents the vertical direction of a 3D marker's position. The marker was placed on subject's left shoe above the second metatarsal head.
There are totally 55 gait cycles in the dataset.

The anomaly is synthetic. For a random gait cycle, we replaced its swing phase with corresponding stance phase of subject's right foot.

Blue is original data, green is data after anomaly was introduced

```
T(start_anomaly:end_anomaly) = UCR_Original_GP711MarkerRFM5z(start_anomaly:end_anomaly);
```





[1] https://www.mad.tf.fau.de/research/activitynet/gaitphase-database/

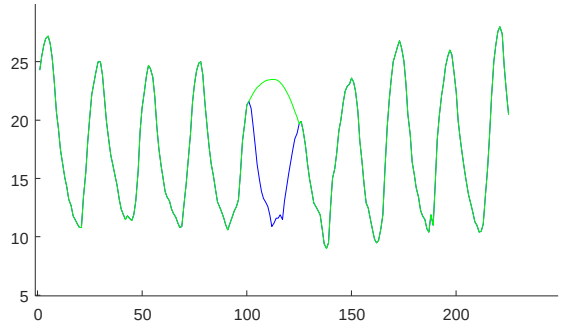
UCR_Anomaly_CIMIS44AirTemperature2_4000_5703_5727.txt

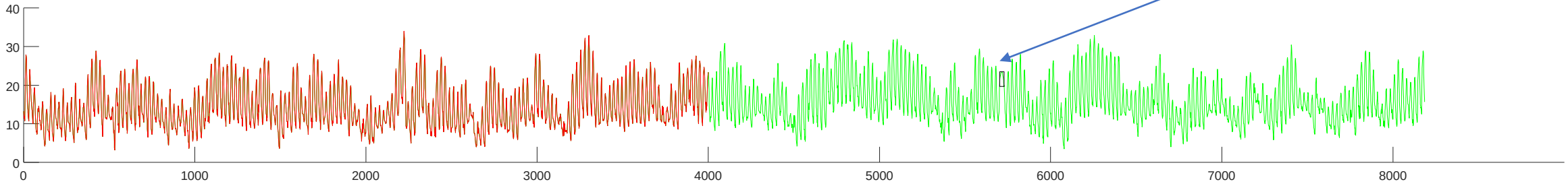
The data comes from the public weather data report from CIMIS station 44 in Riverside [1].
This dataset consists of hourly air temperature between 03/01 and 03/31 from 2009 to 2019.

The anomaly is synthetic. We removed the valley between randomly chosen 2 days and replaced the original two peaks with a single peak via cubic interpolation. The length of the original data is not changed but a wider peak was placed.

```
peak1 = T(start_anomaly);
peak2 = T(end_anomaly);
[~,valley_locidx] = min(T(start_anomaly:end_anomaly));
T(start_anomaly:end_anomaly) = interp1(
    [start_anomaly start_anomaly+valley_locidx end_anomaly],
    [peak1 3*max([peak1 peak2])-2*mean([peak1 peak2]) peak2],
    start_anomaly:end_anomaly, 'pchip');
```

Blue is original data, green is data after anomaly was introduced





[1] ftp://ftpcimis.water.ca.gov/pub2/annualMetric/

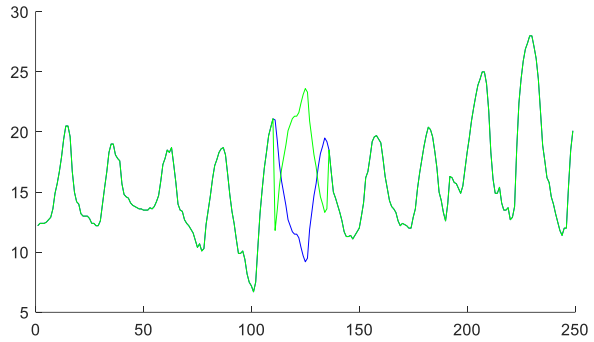
UCR_Anomaly_CIMIS44AirTemperature3_4000_6520_6544.txt

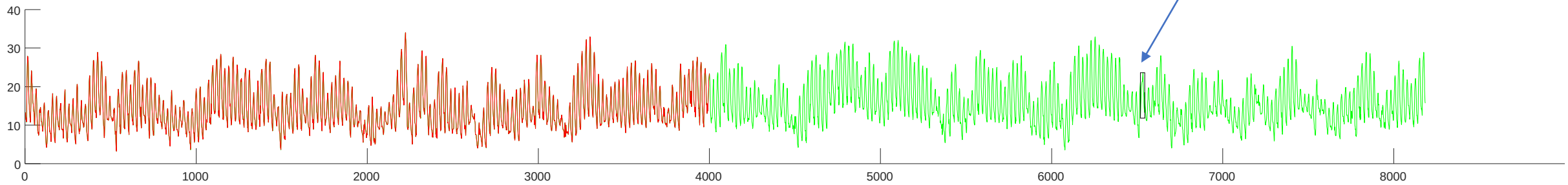
The data comes from the public weather data report from CIMIS station 44 in Riverside [1].
This dataset consists of hourly air temperature between 03/01 and 03/31 from 2009 to 2019.

The anomaly is synthetic. At a random location, we flipped a 12-hour data over across the mean value of the original data. The original valley now becomes an anomaly peak.

```
T(start_anomaly:end_anomaly) = 2*min(T(start_anomaly:end_anomaly))-
    T(start_anomaly:end_anomaly)+mean(T(start_anomaly:end_anomaly));
```

Blue is original data, green is data after anomaly was introduced





[1] ftp://ftpcimis.water.ca.gov/pub2/annualMetric/

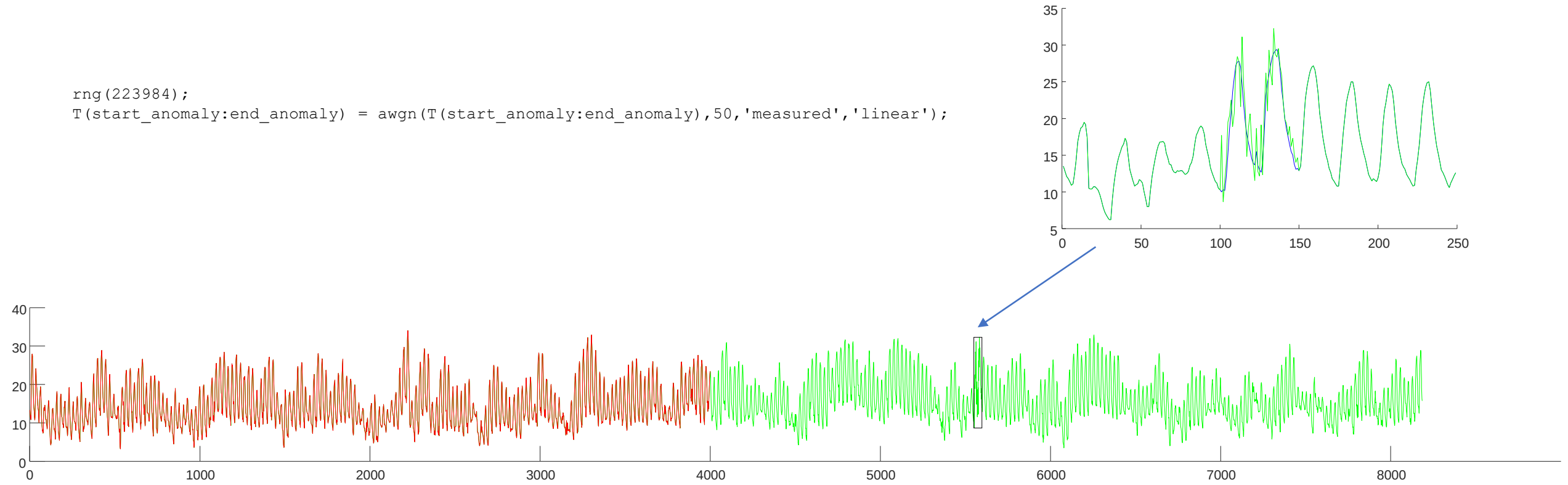
UCR_Anomaly_CIMIS44AirTemperature4_4000_5549_5669.txt

The data comes from the public weather data report from CIMIS station 44 in Riverside [1].
This dataset consists of hourly air temperature between 03/01 and 03/31 from 2009 to 2019.

The anomaly is synthetic. For a random consecutive 48 hours, we created the anomaly by adding noise, to simulate some electrical interference to the temperature sensor.

Blue is original data, green is data after anomaly was introduced

```
rng(223984);
T(start_anomaly:end_anomaly) = awgn(T(start_anomaly:end_anomaly),50,'measured','linear');
```



[1] ftp://ftpcimis.water.ca.gov/pub2/annualMetric/

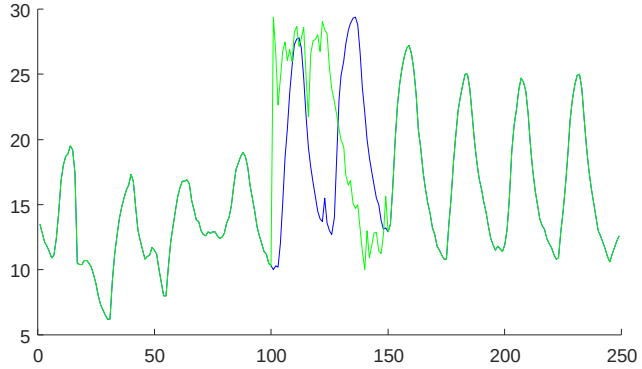
UCR_Anomaly_CIMIS44AirTemperature5_4000_4852_4900.txt

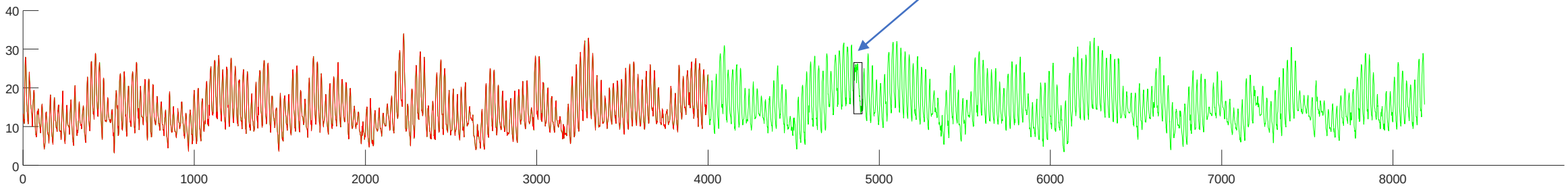
The data comes from the public weather data report from CIMIS station 44 in Riverside [1].
This dataset consists of hourly air temperature between 03/01 and 03/31 from 2009 to 2019.

The anomaly is synthetic. For a randomly chosen 48 hours, we replaced the original data with random walk to simulate a sudden failure of the temperature sensor. The random walk was carefully generated so that it fits within the value range of the original 48-hours data.

Blue is original data, green is data after anomaly was introduced

```
rng(485248);
min_temp = min(T(start_anomaly:end_anomaly));
max_temp = max(T(start_anomaly:end_anomaly));
anomaly = cumsum(randn(end_anomaly-start_anomaly+1, 1));
anomaly = (anomaly-min(anomaly))/(max(anomaly)-min(anomaly));
anomaly = anomaly*(max_temp-min_temp)+min_temp;
T(start_anomaly:end_anomaly) = anomaly;
```





[1] ftp://ftpcimis.water.ca.gov/pub2/annualMetric/

UCR_Anomaly_CIMIS44AirTemperature6_4000_6006_6054.txt

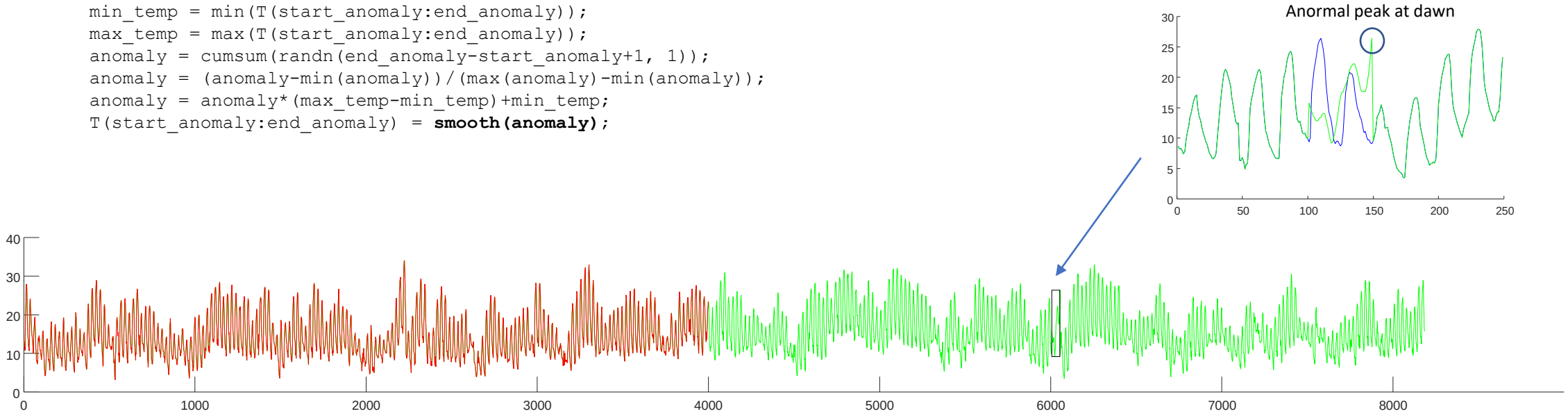
The data comes from the public weather data report from CIMIS station 44 in Riverside [1].
This dataset consists of hourly air temperature between 03/01 and 03/31 from 2009 to 2019.

The anomaly is synthetic. Same way as `UCR_Anomaly_CIMIS44AirTemperature4_4000_4852_4900.txt` to introduce the anomaly, but a much subtle version. Instead of directly replacing the original data with the random walk, we smoothed it before substitution to make it less obvious.

We also carefully chose the random seed, so that there is an abnormal peak temperature at dawn, which is quite impossible in real life. Because at dawn, it is usually the coolest time in a day.

```
rng(459728);
min_temp = min(T(start_anomaly:end_anomaly));
max_temp = max(T(start_anomaly:end_anomaly));
anomaly = cumsum(randn(end_anomaly-start_anomaly+1, 1));
anomaly = (anomaly-min(anomaly))/(max(anomaly)-min(anomaly));
anomaly = anomaly*(max_temp-min_temp)+min_temp;
T(start_anomaly:end_anomaly) = smooth(anomaly);
```

Blue is original data, green is data after anomaly was introduced



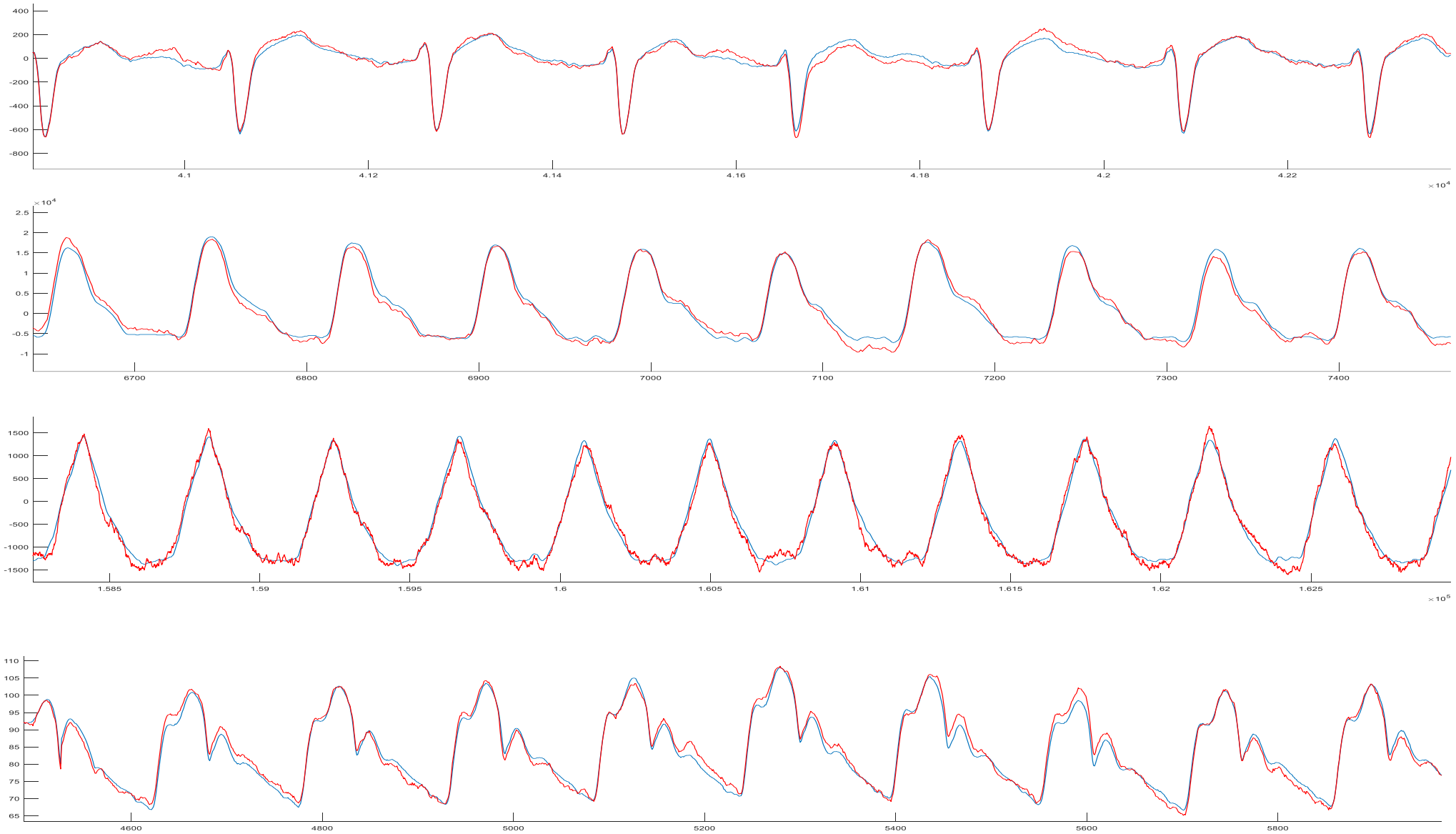
[1] ftp://ftpcimis.water.ca.gov/pub2/annualMetric/

- The remaining datasets, upto 200, are duplicates of the previous datasets, but..
  - Some with some **distort**ion added. If `temp` is the dataset, then…
    - `distort = temp + (smooth(randn(size(temp)),25) * std(temp) );`
  - The name of the file reflects this. For example
    - `UCR_Anomaly_DISTORTEDECG4_5000_17000_17100.txt`
  - The distortion adds a little noise and wandering baseline.

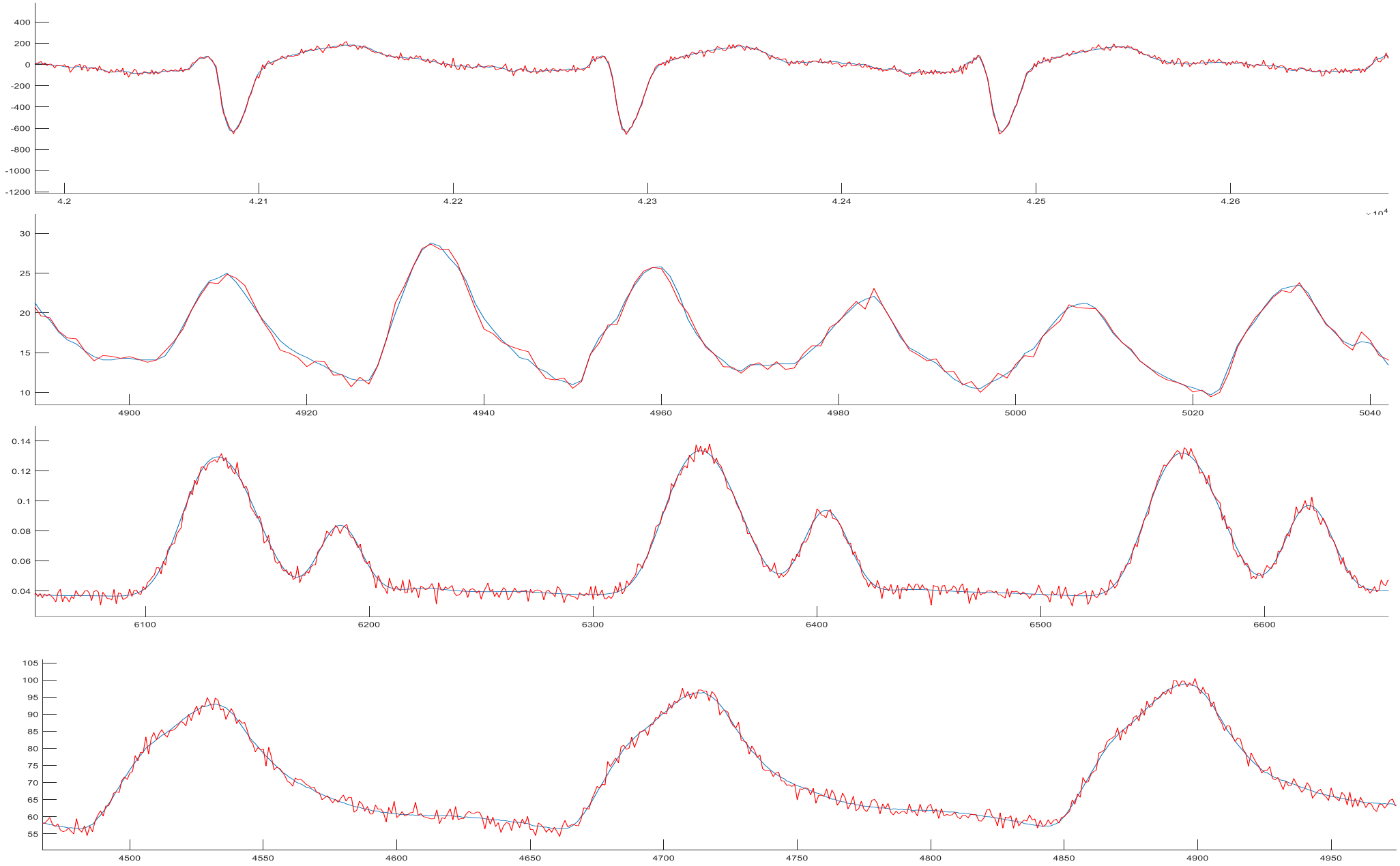
- and
  - Some with some **noise** added. If `temp` is the dataset, then…
    - `NOISE = temp + (randn(size(temp)) * (std(temp)/10));`
  - The name of the file reflects this. For example
    - `102_UCR_Anomaly_NOISEMesoplodonDensirostris_10000_19280_19440.txt`
  - The noise is simple Gaussian noise

- In both cases, this makes little visual difference to the time series, see the next two slides.

- Samples of `distort = temp + (smooth(randn(size(temp)),25) * std(temp) );` The red time series are distorted



- Samples of `NOISE = temp + (randn(size(temp)) * (std(temp)/10));` The red time series are noisy



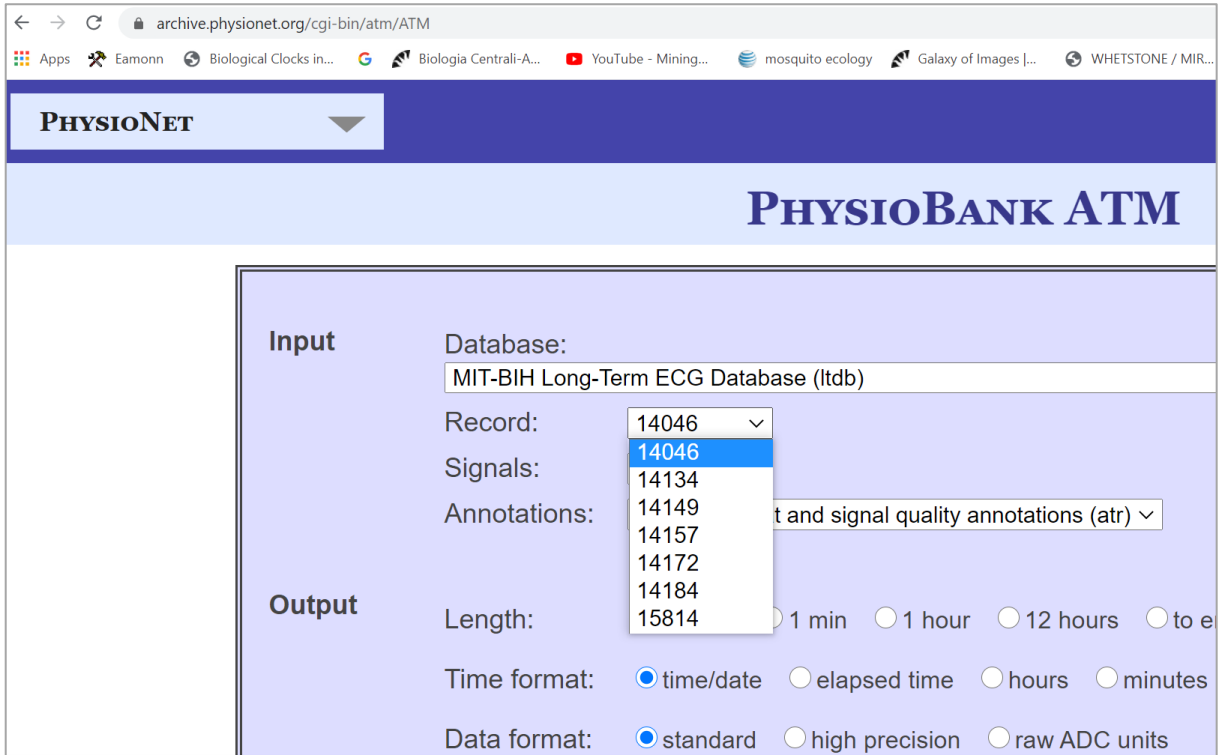
For these files we considered data from MIT-BIH Long-term ECG database.

These dataset have many arrythmias, sometimes multiple types of arrythmias. However, in every case we made sure that the arrythmias *also* exist in the training data, so they are not novel in the test data.

We will not bother creating a slide for each dataset, but in the following few slides we show some examples of the anomalies we created.

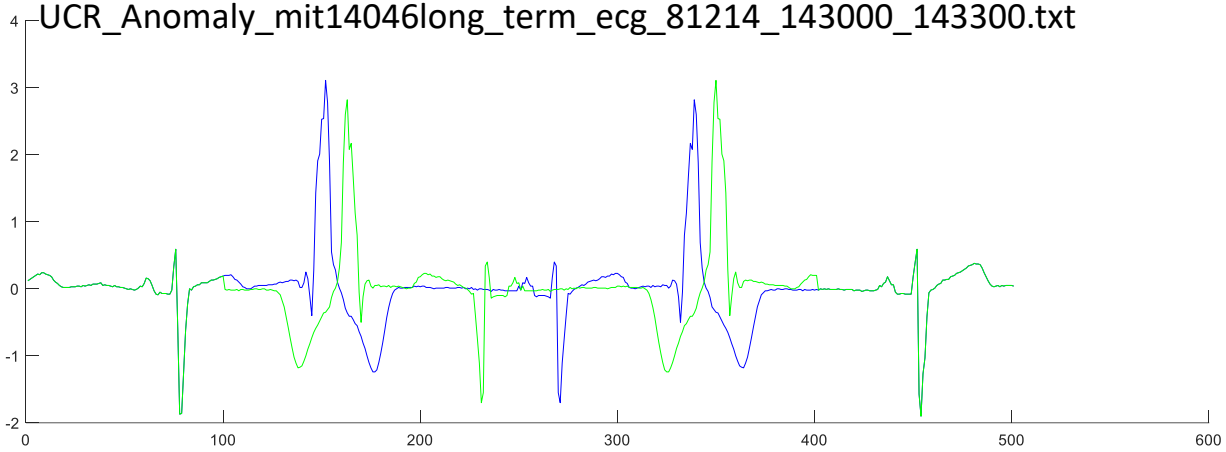
We create these with cardiologist Dr. Greg Mason

Our philosophy was: If a trained cardiologist had a lot of time to view the data, would *they* find the anomaly we added?
Some of the anomalies we added would easily pass this test, for some others it is closer to the edge...



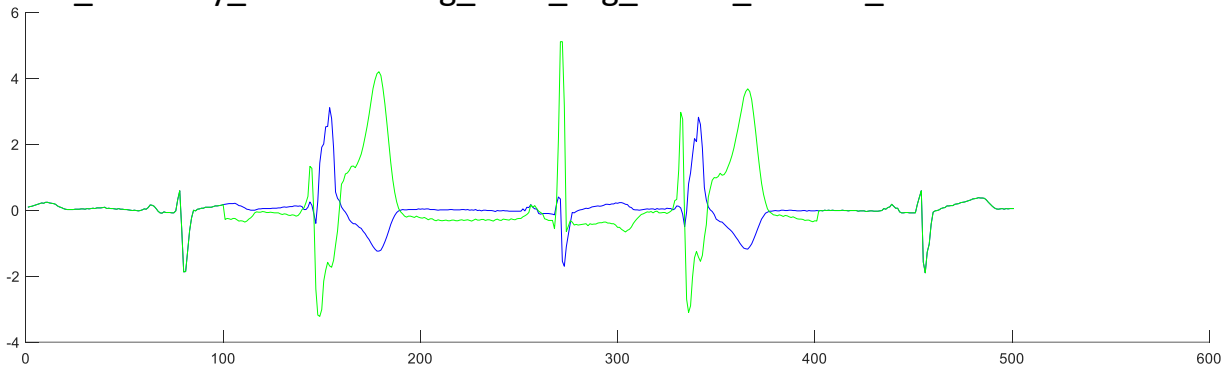
Reverse two beats (they happened to be arrythmias)

UCR_Anomaly_mit14046long_term_ecg_81214_143000_143300.txt



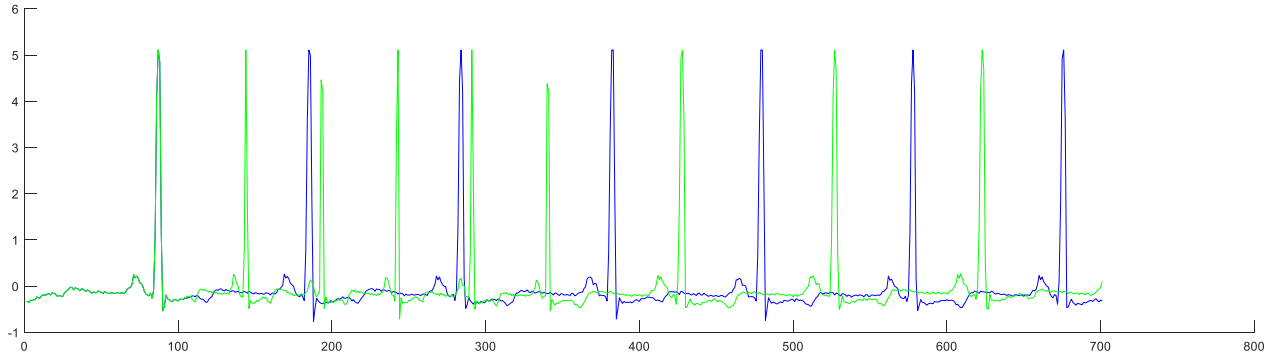
Swap in lead 2 for lead 1 a random region

UCR_Anomaly_mit14046long_term_ecg_96123_123000_123300.txt

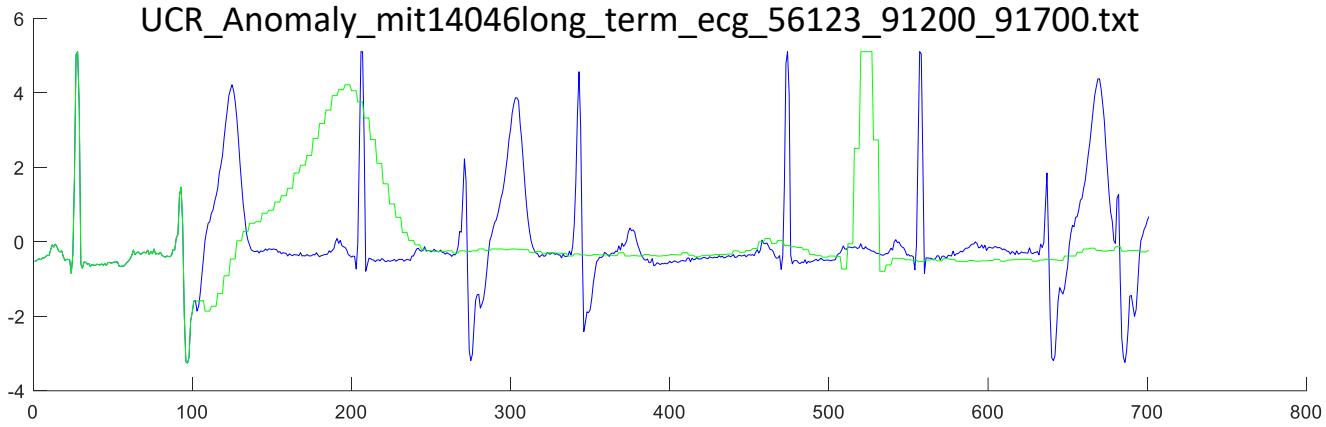


For a region containing about 8 normal beats, increase the speed by a factor of 2

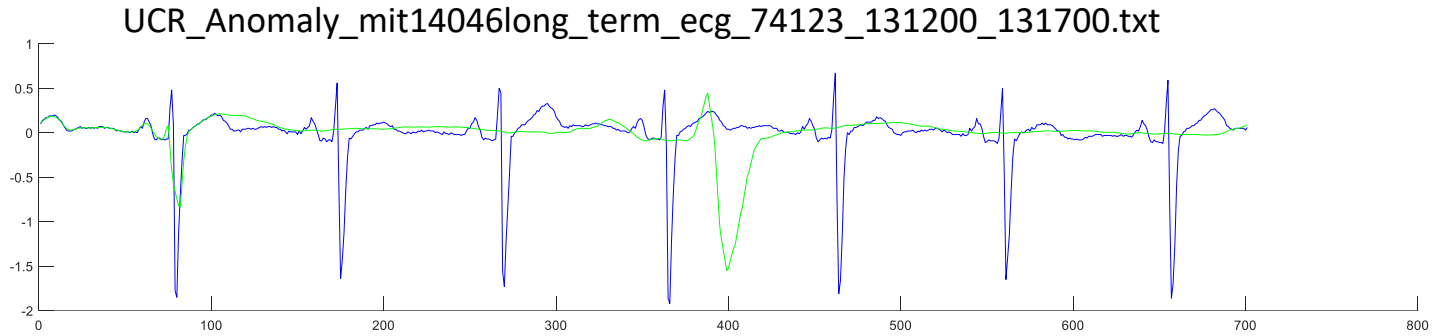
UCR_Anomaly_mit14046long_term_ecg_76123_191200_191700.txt



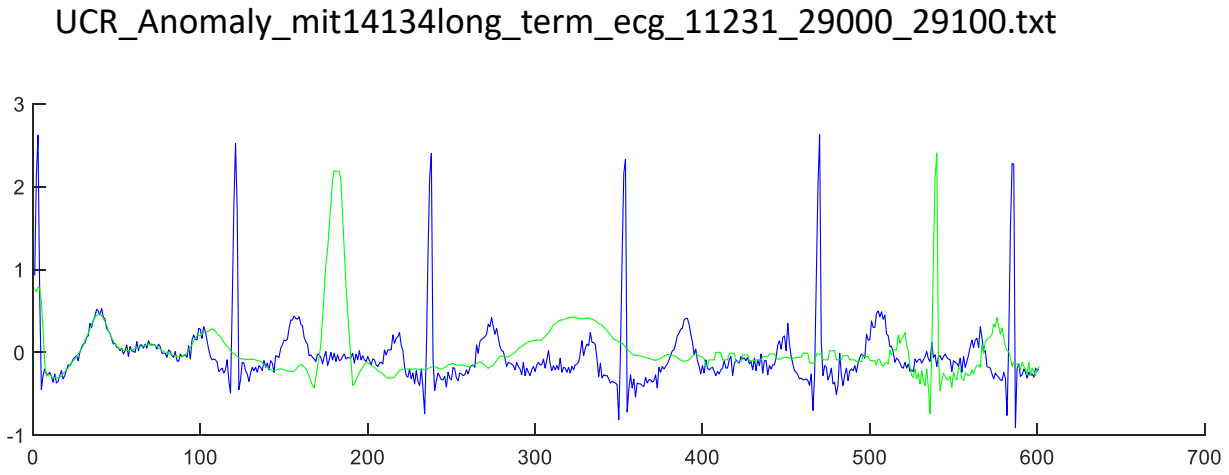
For a random region decrease the speed by a factor of 4, and leave in the “staircase” artifacts.



For a random region, decrease the speed by a factor of 4, and smooth out artifacts.

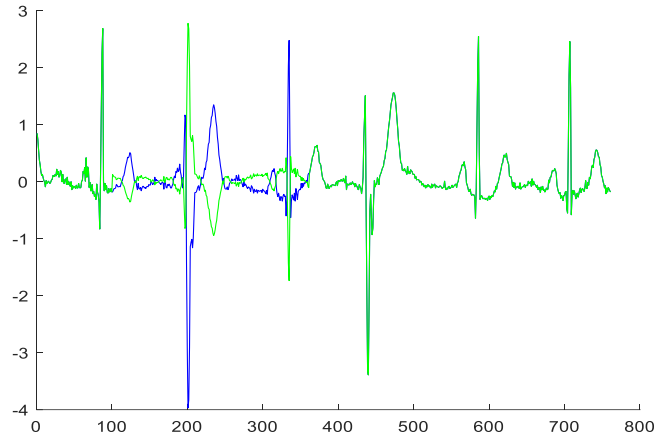


For a random region, decrease the speed by a factor of 4, and smooth out artifacts.



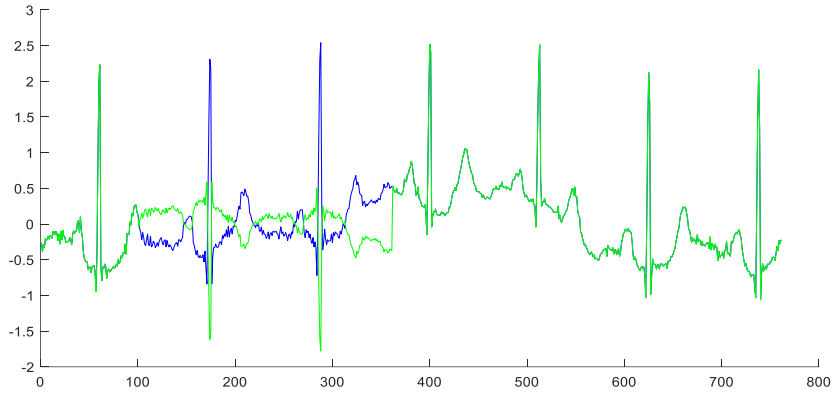
Flip a random region upside-down, by multiplying it by -0.7

UCR_Anomaly_mit14134long_term_ecg_8763_57530_57790.txt



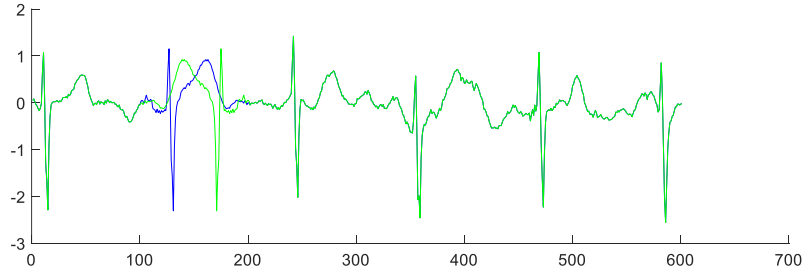
(again) Flip a random region upside-down, by multiplying it by -0.7

UCR_Anomaly_mit14134long_term_ecg_8763_47530_47790.txt



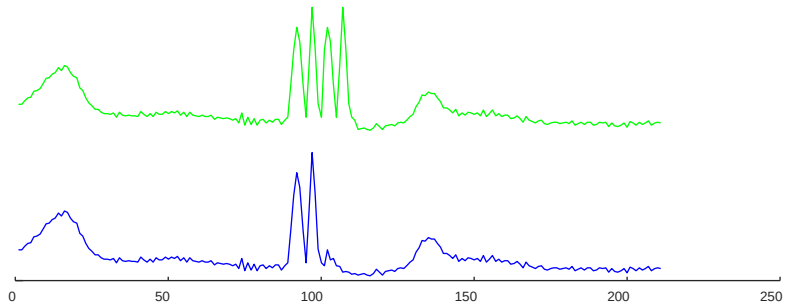
Reverse a single beat

UCR_Anomaly_mit14134long_term_ecg_19363_19510_19610.txt



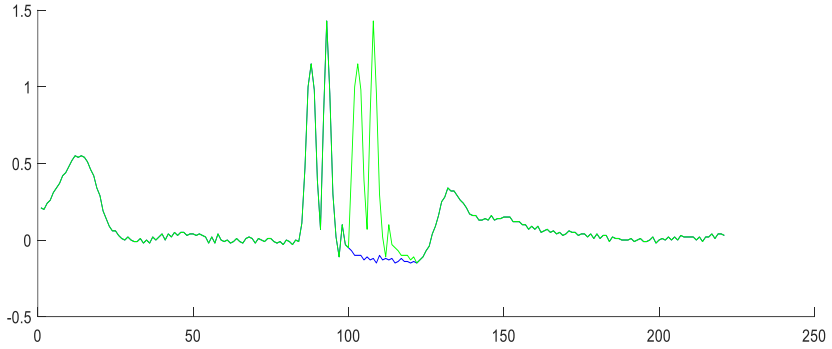
Normally there are two peaks, we “stuttered” the data so that there are 4 peaks

UCR_Anomaly_mit14134long_term_ecg_16363_57960_57970.txt



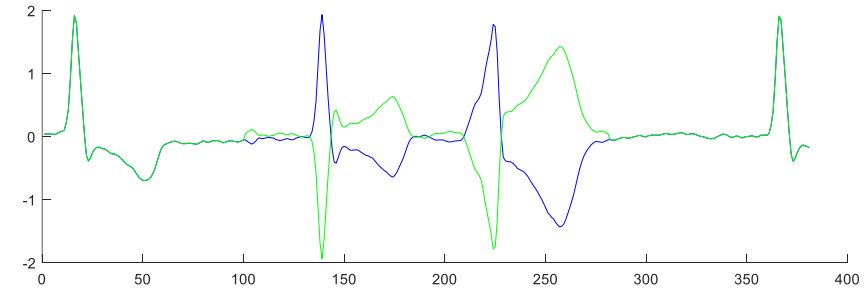
(again) Normally there are two peaks, we “stuttered” the data so that there are 4 peaks

UCR_Anomaly_mit14134long_term_ecg_11361_47830_47850.txt



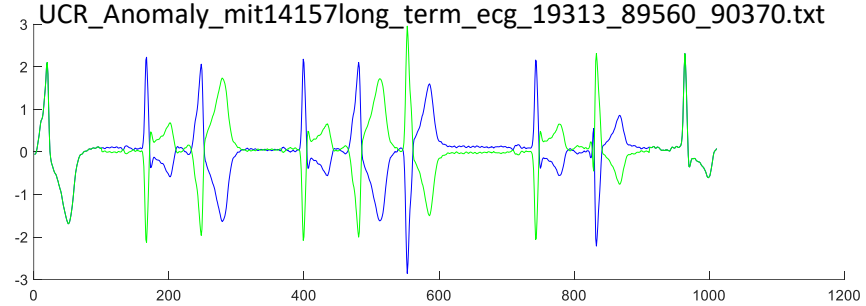
Flip a section upside down by multiplying it by -1

UCR_Anomaly_mit14157long_term_ecg_21311_72600_72780.txt



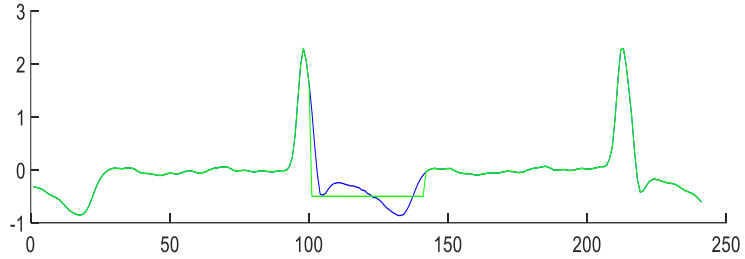
Flip a (longish) section that happened to contain some arrythmias upside down by multiplying it by -1

UCR_Anomaly_mit14157long_term_ecg_19313_89560_90370.txt



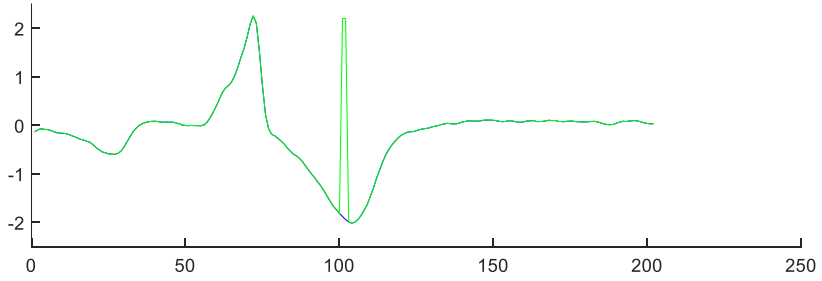
UCR_Anomaly_mit14157long_term_ecg_19313_46350_46390.txt

Hard limited a section of a beat to -0.5 (plus a tiny amount of noise)

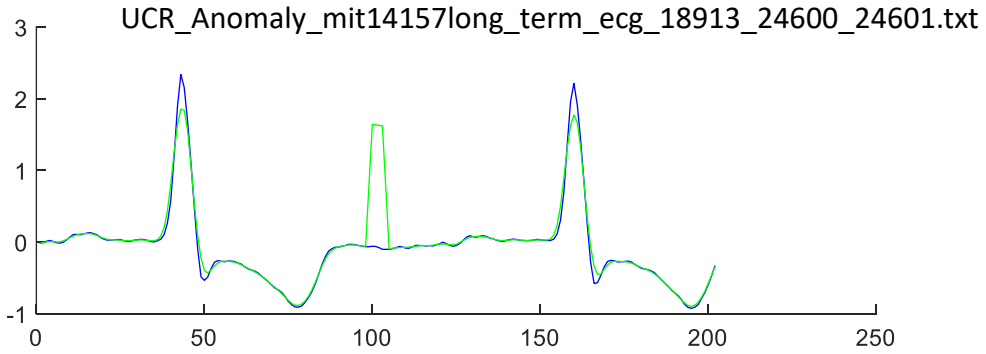


UCR_Anomaly_mit14157long_term_ecg_18913_75450_75451.txt

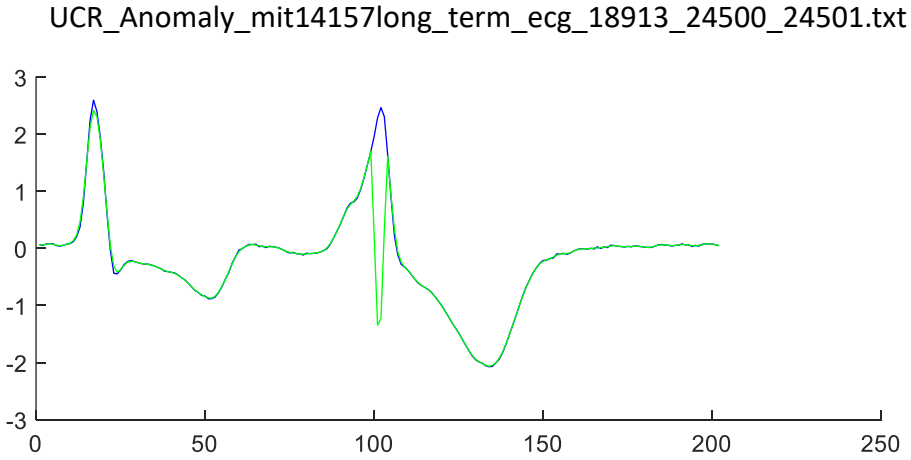
Added a spike (two points long)

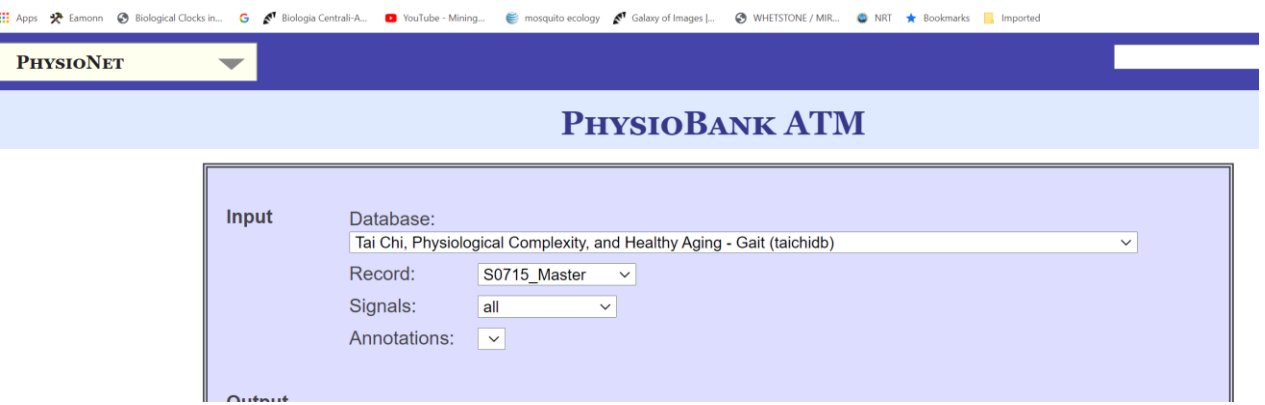


Added a spike (two points long), and slightly smoothed the area around it to blend it in more.



Added a dropout (two points long), and slightly smoothed the area around it to blend it in more.





The original data only has 4 unique values

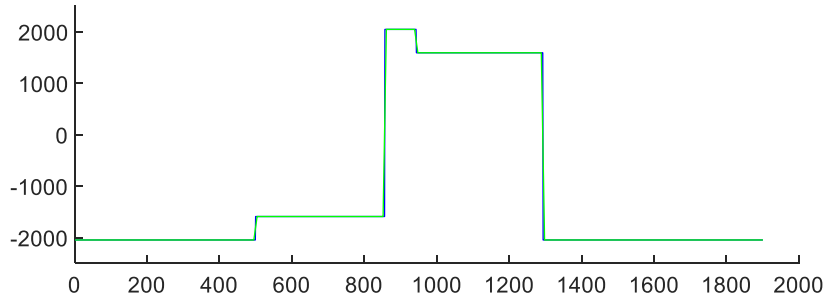
-2047 -1591 1592 2047

By *slightly* smoothing a section, we created a region with more unique values

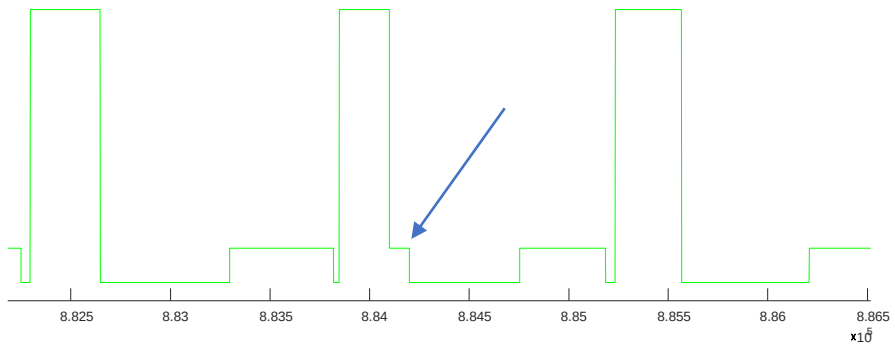
The original data has a small “step” before the main step. In one location we made a step after the main step.

The large step sometimes has an “overshoot” bump at the beginning. In one location we made an “overshoot” bump at the end of the step.

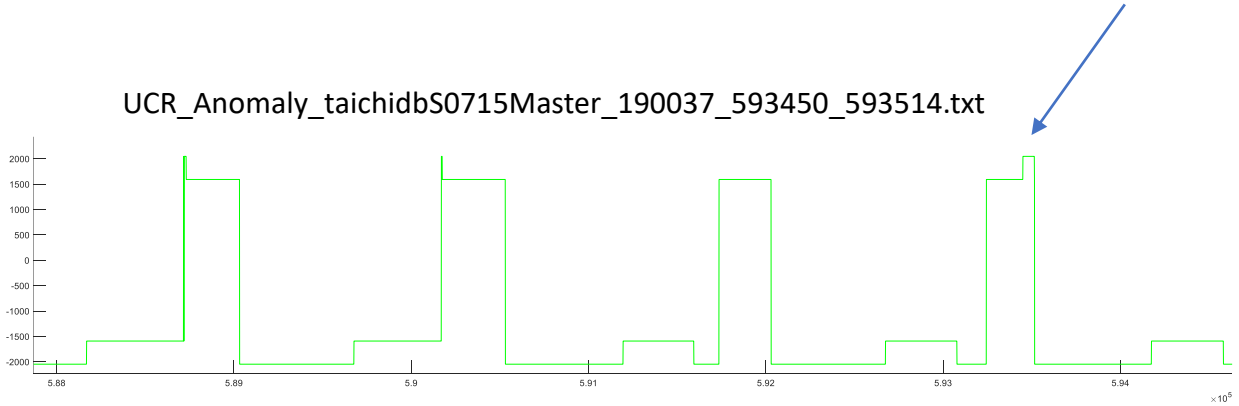
UCR_Anomaly_taichidbS0715Master_250000_837400_839100.txt



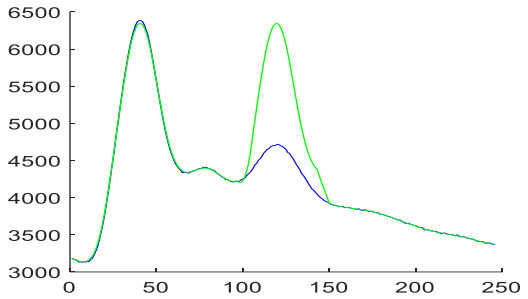
UCR_Anomaly_taichidbS0715Master_240030_884100_884200.txt



UCR_Anomaly_taichidbS0715Master_190037_593450_593514.txt

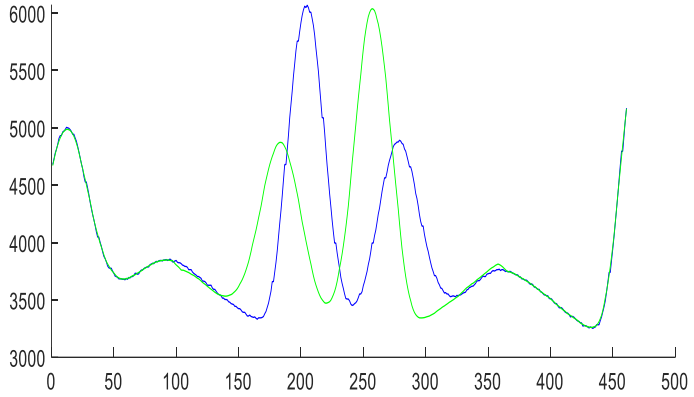


We “stuttered” the first bump of the APB cycle to enlarge the second bump



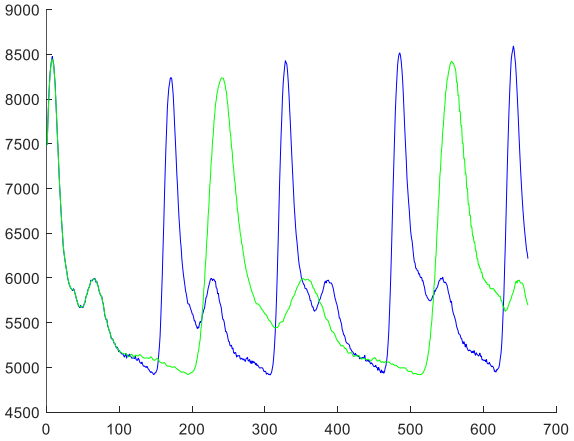
UCR_Anomaly_tilt12744mtable_100000_203355_203400.txt

We reversed a single APB cycle



UCR_Anomaly_tilt12744mtable_100000_104630_104890.txt

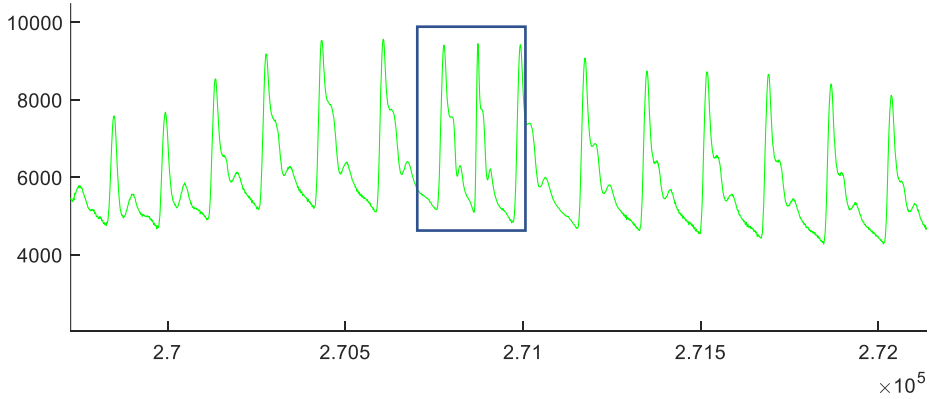
We slowed down an ABP beat by a factor of two.



UCR_Anomaly_tilt12754table_100013_104630_104890.txt

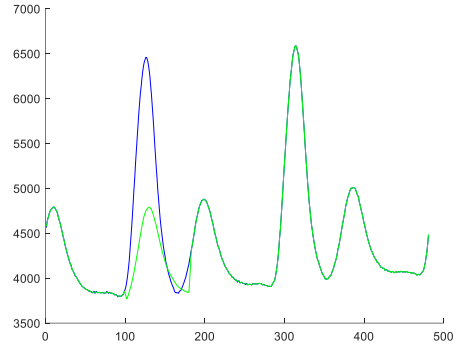
We sped up two ABP beats by a factor of two.

UCR_Anomaly_tilt12754table_100211_270800_271070.txt



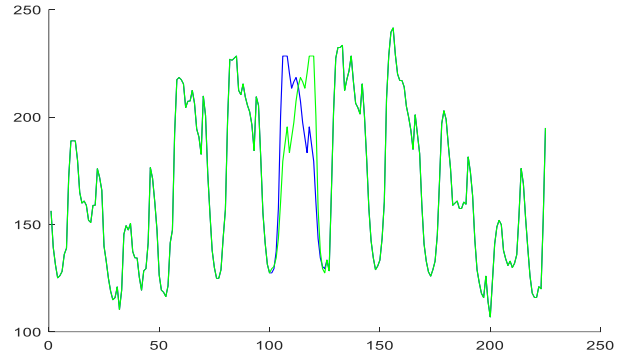
Stutter a small peak into the location of a large peak.

UCR_Anomaly_tilt12755mtable_50211_121900_121980.txt



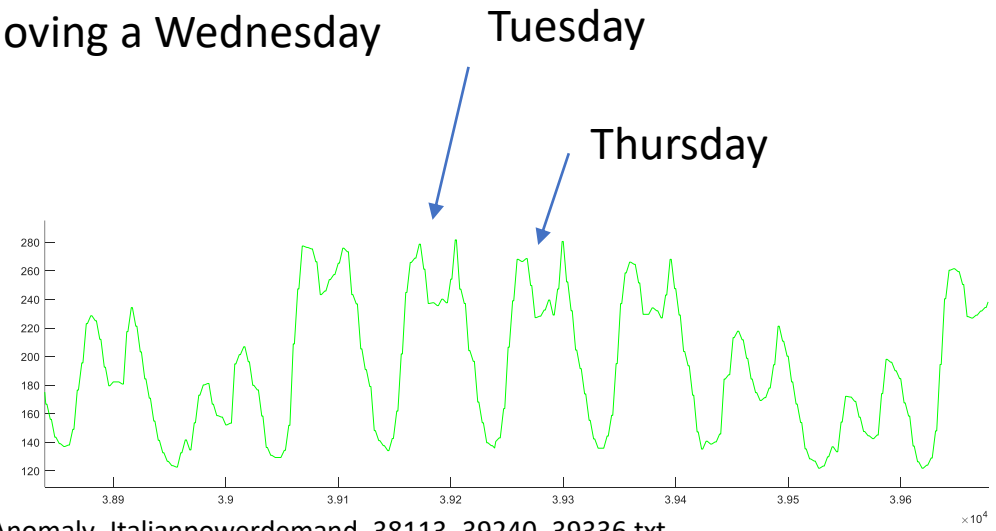
For Italian power demand
There are already natural “anomalies” (holidays, weather events etc). However, one year of training data should be enough to seen all these events at least once.
So, the anomalies we embedded are obvious, and hopefully more significantly that any natural anomalies.

Reversed the direction of a Wednesday



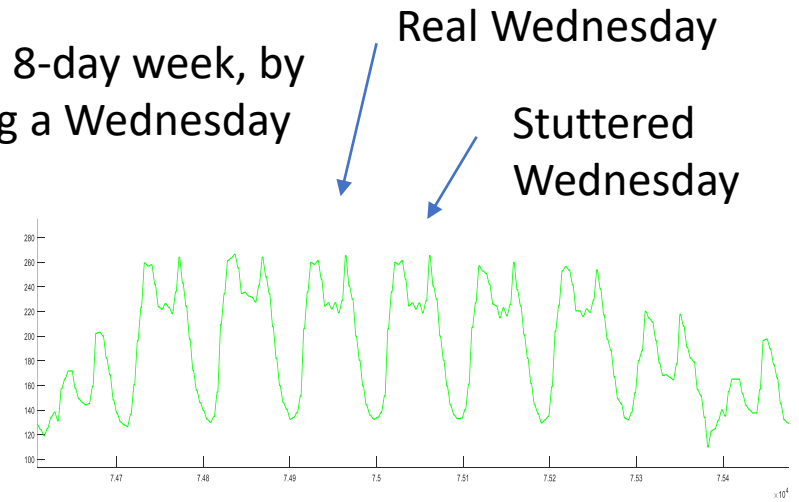
UCR_Anomaly_Italianpowerdemand_8913_29480_29504.txt

Made a 6-day week, by removing a Wednesday



UCR_Anomaly_Italianpowerdemand_38113_39240_39336.txt

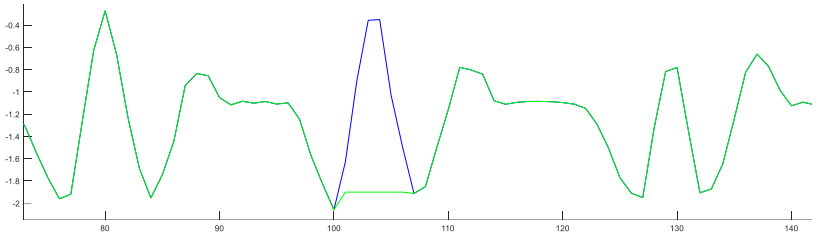
Made an 8-day week, by stuttering a Wednesday



UCR_Anomaly_Italianpowerdemand_36123_74900_74996.txt

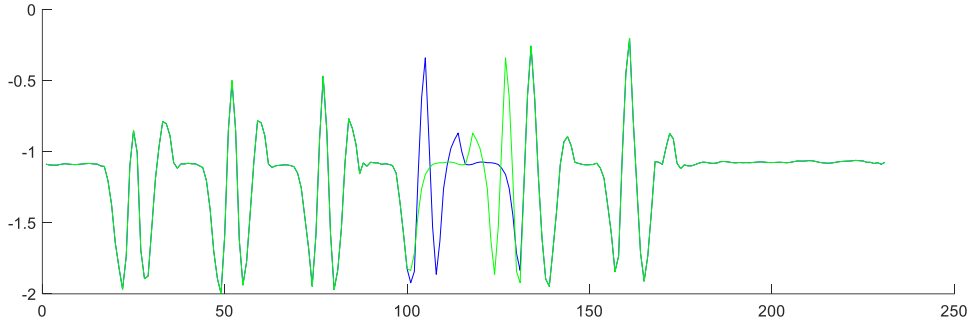
WeAllWalk

Flatted part of the gait cycle.



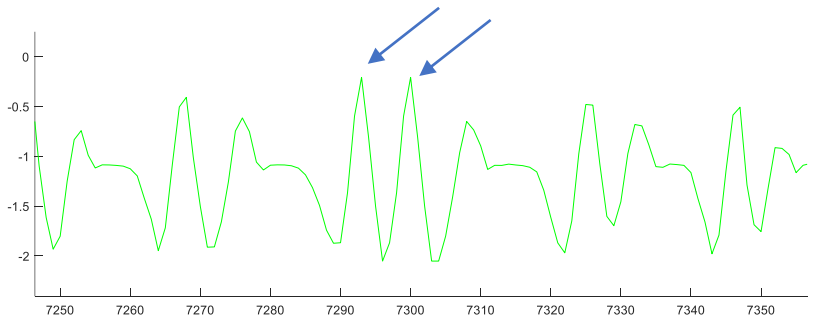
UCR_Anomaly_weallwalk_2000_4702_4707.txt

Reverse 1 gait cycle.



UCR_Anomaly_weallwalk_2753_8285_8315.txt

“stuttered” part of gait cycle

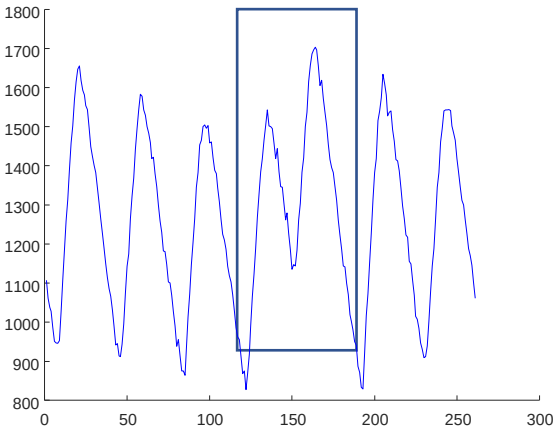


UCR_Anomaly_weallwalk_2951_7290_7296.txt

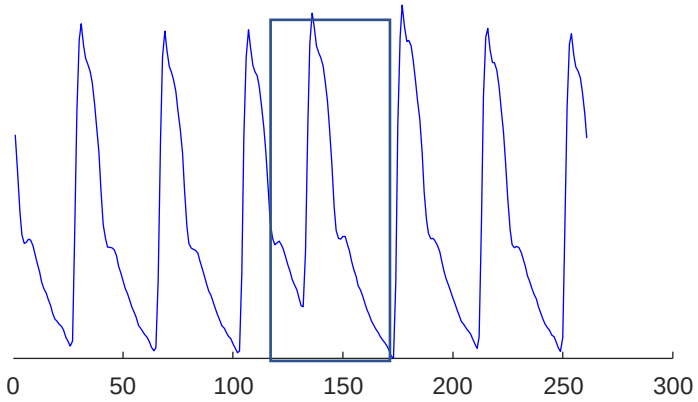
CHARISten

ECG, arterial blood pressure (ABP), and intracranial pressure (ICP)

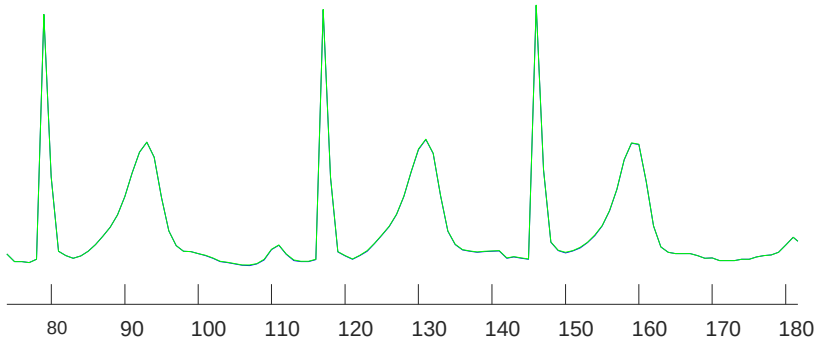
Natural anomaly of a single arrythmia.
We shifted the starting point, so that the anomaly would not always be in the same spot for the 3 datasets we made.



UCR_Anomaly_CHARISten_3165_26929_26989.txt



UCR_Anomaly_CHARISten_25130_29080_29140.txt



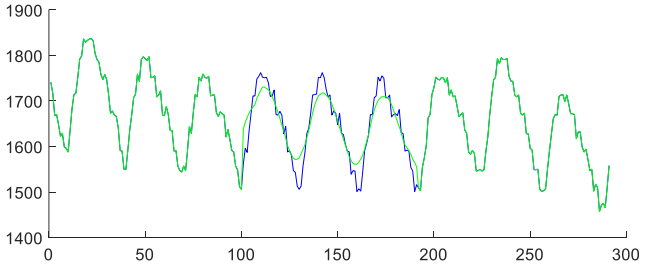
UCR_Anomaly_CHARISten_5130_27929_27989.txt

CHARISfive

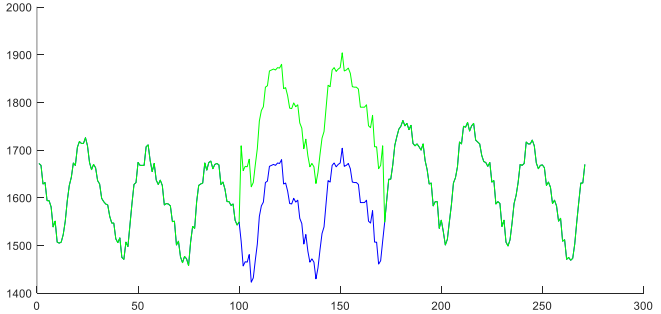
ECG, arterial blood pressure (ABP), and intracranial pressure (ICP)

We smoothed three beats, which also has the effect of reducing the size of the peaks and valleys

UCR_Anomaly_CHARISfive_9812_28995_29085.txt



We added 200 to two beats, shifting them up.

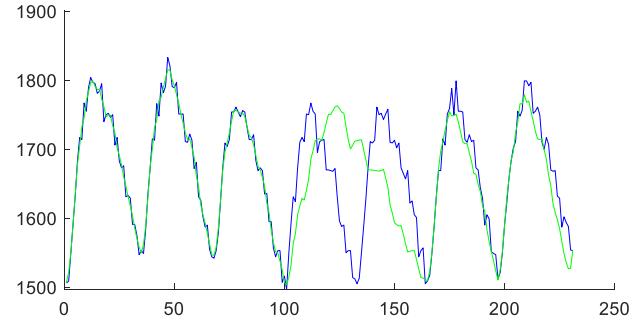


UCR_Anomaly_CHARISfive_12412_15000_15070.txt

This dataset
202_UCR_Anomaly_CHARISfive_10411_10998_11028.txt
Was originally mislabeled as
202_UCR_Anomaly_CHARISfive_11411_10998_11028.txt

Meaning that it was impossible to solve.
It is now fixed.

We made one beat twice as long as normal



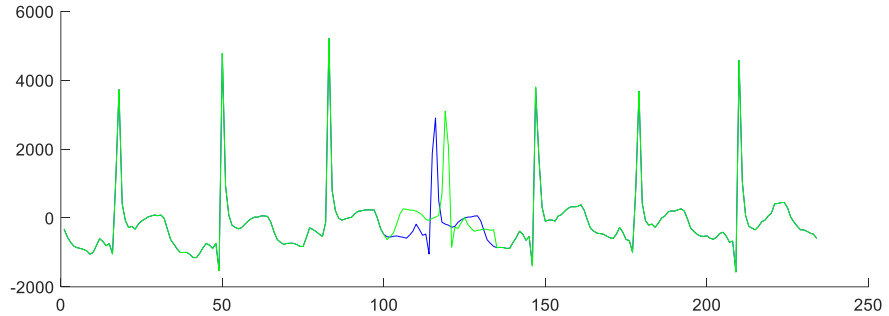
UCR_Anomaly_CHARISfive_11411_10998_11028.txt

Same for
203_UCR_Anomaly_CHARISfive_11812_10995_11028.txt
Now fixed, and renamed to
203_UCR_Anomaly_CHARISfive_10500_10995_11028.txt

CHARISfive

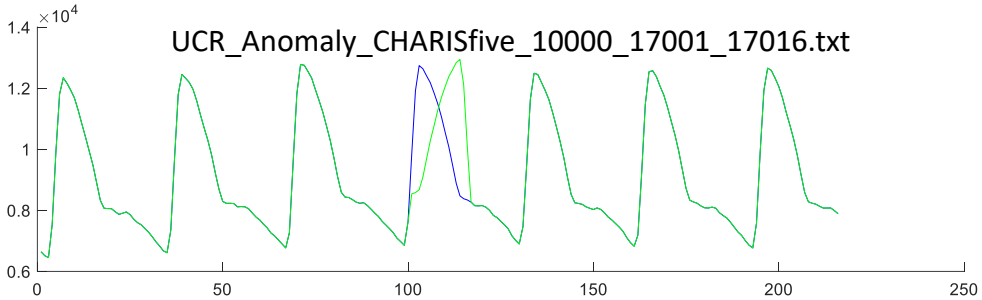
ECG, arterial blood pressure (ABP), and intracranial pressure (ICP)

We flipped a beat



UCR_Anomaly_CHARISfive_11812_10995_11028.txt

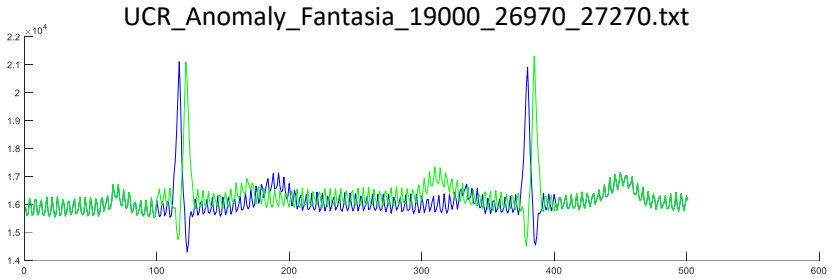
We flipped a beat



UCR_Anomaly_CHARISfive_10000_17001_17016.txt

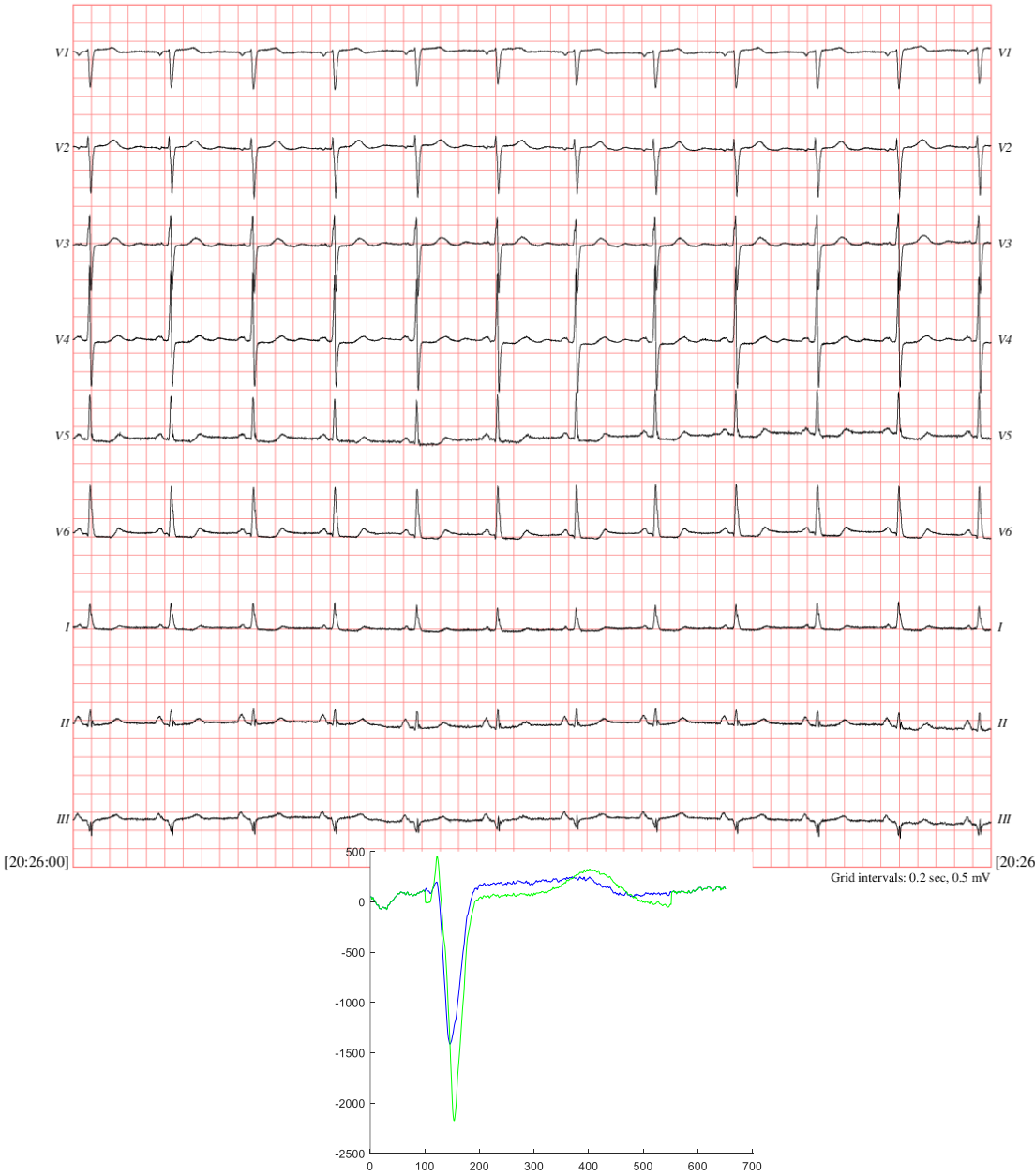
Fantasia Database

We flipped a beat



STAFFIIIDatabase

Here we swapped in short regions from a different trace



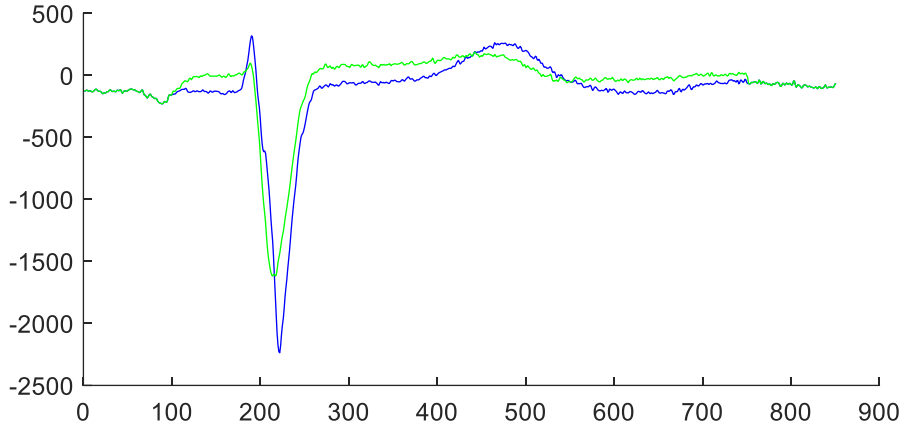
V2 for V1

UCR_Anomaly_STAFFIIIDatabase_33211_126920_127370.txt

STAFFIIIDatabase

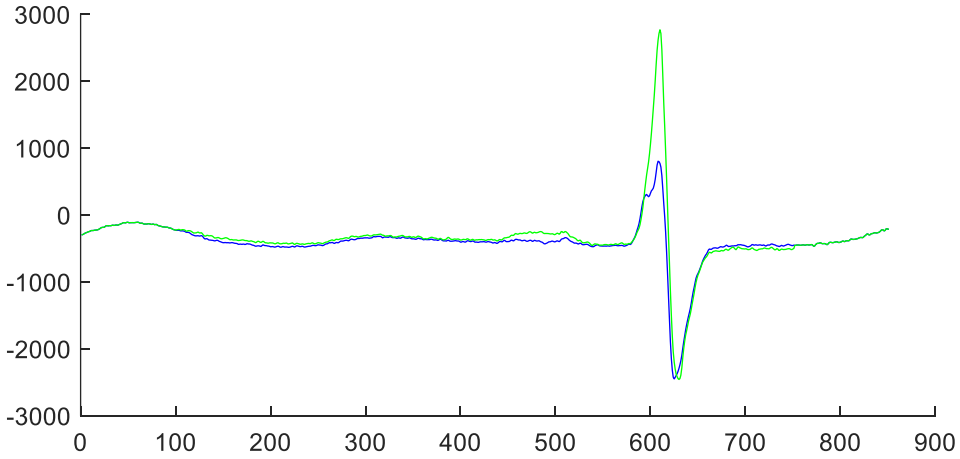
Here we swapped in short regions from a different trace

V2 for V1



UCR_Anomaly_STAFFIIIDatabase_36276_106720_107370.txt

V3 for 4

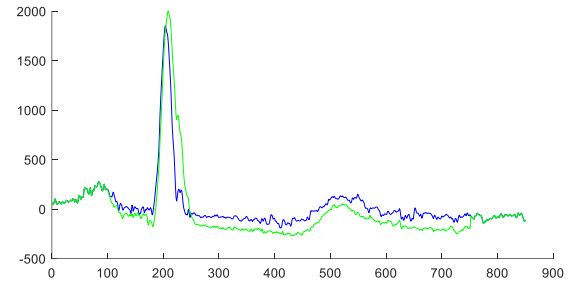


UCR_Anomaly_STAFFIIIDatabase_37216_160720_161370.txt

STAFFIIIDatabase

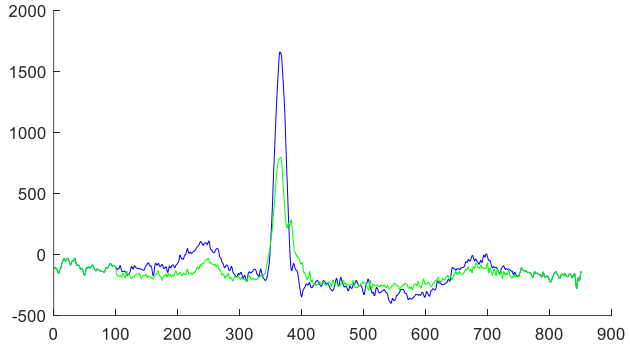
Here we swapped in short regions from a different trace

V6 for V4



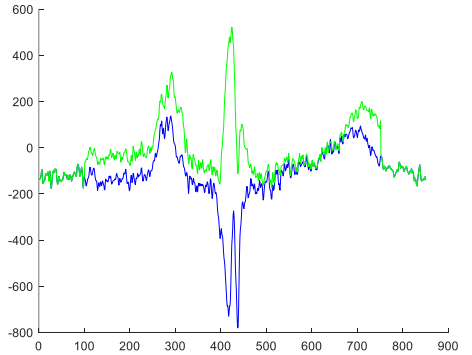
UCR_Anomaly_STAFFIIIDatabase_34211_125720_126370.txt

V7 for V4



UCR_Anomaly_STAFFIIIDatabase_38211_150720_151370.txt

8 for 9

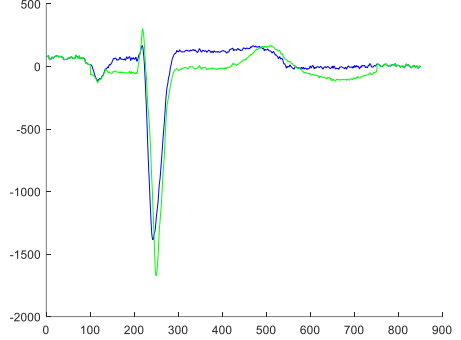


UCR_Anomaly_STAFFIIIDatabase_43217_250720_251370.txt

STAFFIIIDatabase

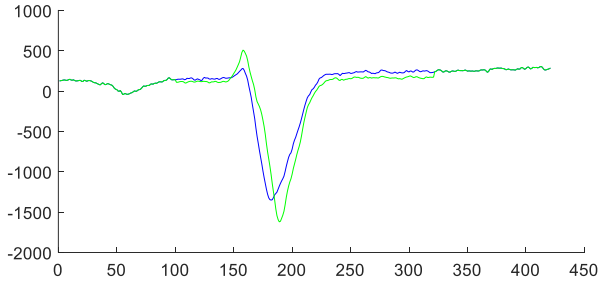
Here we swapped in short regions from a different trace

V2 for V1



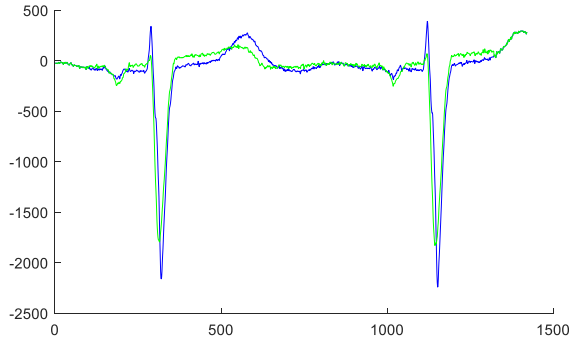
UCR_Anomaly_STAFFIIIDatabase_41117_210720_211370.txt

V2 for V1



UCR_Anomaly_STAFFIIIDatabase_41612_64632_64852.txt

V1 for V2



UCR_Anomaly_STAFFIIIDatabase_45616_163632_164852.txt