DATA

JEROEN VEEN

HUGO ARENDS

HAN_UNIVERSITY
OF APPLIED SCIENCES

WORKFLOW

HAN_UNIVERSITY
OF APPLIED SCIENCES
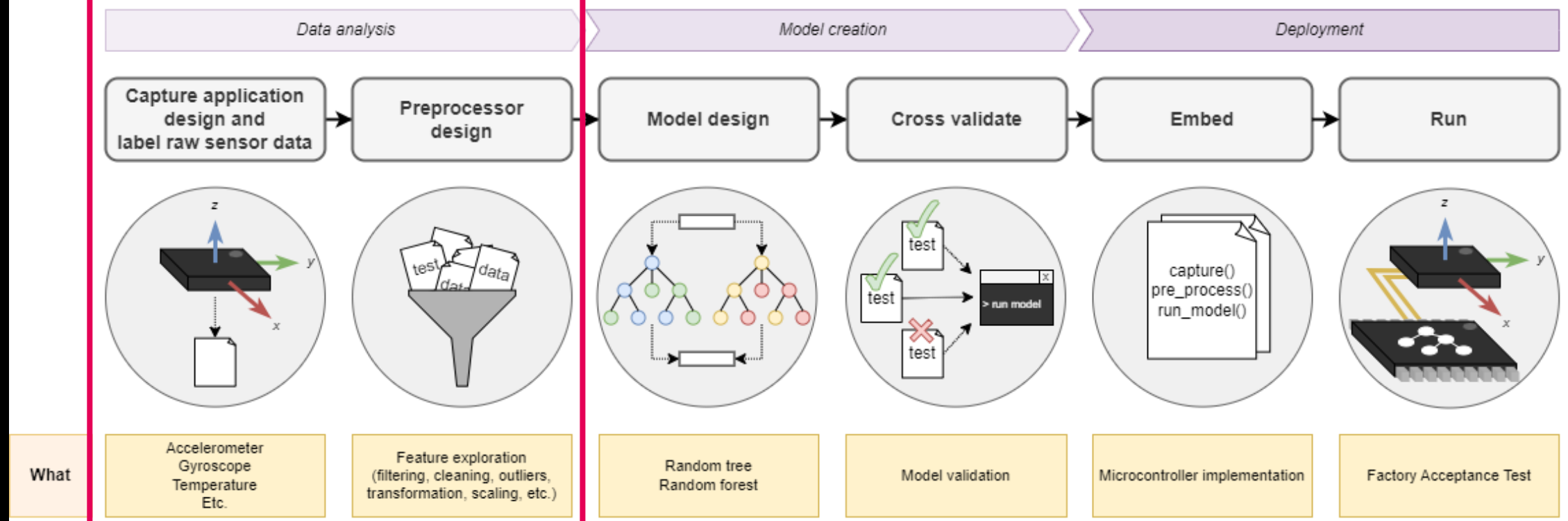
AGENDA

- Data collection
- SMART objective
- Data splitting
- Feature engineering
- Exploratory feature analysis

LABEL RAW SENSOR DATA

Input	Process	Output

Microcontroller development board

UART →

./capturing/data_recorder.py

→ ./data/captured/<label_1>.csv

./data/captured/<label_2>.csv

./data/captured/<label_3>.csv

TODO
LABEL_NAME: *the label name for the current recording.*
Tip: use short, yet descriptive names!

HAN_UNIVERSITY
OF APPLIED SCIENCES

LABEL RAW SENSOR DATA

Input	Process	Output

Microcontroller development board

UART →

 ./capturing/data_recorder.py

→ ./data/captured/<label_1>.csv

./data/captured/<label_2>.csv

./data/captured/<label_3>.csv

TODO
ATTRIBUTE_NAMES: *the attribute names according to your application. Tip: use short, yet descriptive names!*

HAN_ UNIVERSITY
OF APPLIED SCIENCES

LABEL RAW SENSOR DATA

Input	Process	Output

Microcontroller development board --- UART ---> ./capturing/data_recorder.py ---> 📄 *./data/captured/<label_1>.csv*

📄 *./data/captured/<label_2>.csv*

📄 *./data/captured/<label_3>.csv*

TODO
N_SAMPLES: *the number of samples the script records*
(per recording)

HAN_UNIVERSITY
OF APPLIED SCIENCES

LABEL RAW SENSOR DATA

📄 *./data/captured/stationary.csv*

```
label,timestamp1,timestamp2,x_out,y_out,z_out
stationary,216321,216321,0.244,0.244,1000.0
stationary,216330,216330,0.488,1.465,999.756
stationary,216340,216340,0.488,0.0,1000.732
stationary,216350,216350,0.244,1.953,1001.953
stationary,216360,216360,0.488,0.488,999.512
stationary,216369,216369,0.732,1.709,999.512
stationary,216379,216379,-0.488,-0.488,999.756
stationary,216389,216389,0.244,0.488,1001.221
stationary,216399,216399,-0.244,-0.732,1001.221
stationary,216408,216408,-0.488,-0.732,1000.488
stationary,216418,216418,0.732,0.977,1000.488
stationary,216428,216428,-1.221,-0.488,1000.977
```

LABEL RAW SENSOR DATA

📄 *./data/captured/stationary.csv*

```
label,timestamp1,timestamp2,x_out,y_out,z_out
stationary,216321,216321,0.244,0.244,1000.0
stationary,216330,216330,0.488,1.465,999.756
stationary,216340,216340,0.488,0.0,1000.732
stationary,216350,216350,0.244,1.953,1001.953
stationary,216360,216360,0.488,0.488,999.512
stationary,216369,216369,0.732,1.709,999.512
stationary,216379,216379,-0.488,-0.488,999.756
stationary,216389,216389,0.244,0.488,1001.221
stationary,216399,216399,-0.244,-0.732,1001.221
stationary,216408,216408,-0.488,-0.732,1000.488
stationary,216418,216418,0.732,0.977,1000.488
stationary,216428,216428,-1.221,-0.488,1000.977
```

Added by the Python script

label, timestamp1 and timestamp2 are fixed names

x_out, y_out and z_out are the configurable attribute names

HAN_UNIVERSITY
OF APPLIED SCIENCES

LABEL RAW SENSOR DATA

📄 *./data/captured/stationary.csv*

```
label,timestamp1,timestamp2,x_out,y_out,z_out
stationary,216321,216321,0.244,0.244,1000.0
stationary,216330,216330,0.488,1.465,999.756
stationary,216340,216340,0.488,0.0,1000.732
stationary,216350,216350,0.244,1.953,1001.953
stationary,216360,216360,0.488,0.488,999.512
stationary,216369,216369,0.732,1.709,999.512
stationary,216379,216379,-0.488,-0.488,999.756
stationary,216389,216389,0.244,0.488,1001.221
stationary,216399,216399,-0.244,-0.732,1001.221
stationary,216408,216408,-0.488,-0.732,1000.488
stationary,216418,216418,0.732,0.977,1000.488
stationary,216428,216428,-1.221,-0.488,1000.977
```

*Transmitted by microcontroller
at every sample time*

HAN_UNIVERSITY
OF APPLIED SCIENCES

LABEL RAW SENSOR DATA

📄 *./data/captured/stationary.csv*

```
label,timestamp1,timestamp2,x_out,y_out,z_out
stationary,216321,216321,0.244,0.244,1000.0
stationary,216330,216330,0.488,1.465,999.756
stationary,216340,216340,0.488,0.0,1000.732
stationary,216350,216350,0.244,1.953,1001.953
stationary,216360,216360,0.488,0.488,999.512
stationary,216369,216369,0.732,1.709,999.512
stationary,216379,216379,-0.488,-0.488,999.756
stationary,216389,216389,0.244,0.488,1001.221
stationary,216399,216399,-0.244,-0.732,1001.221
stationary,216408,216408,-0.488,-0.732,1000.488
stationary,216418,216418,0.732,0.977,1000.488
stationary,216428,216428,-1.221,-0.488,1000.977
```


Bunch
clear()
value()

Container object that extends dictionaries by enabling values to be accessed by key òr attribute.

```
d = {"labels":["a","b","c"], "vals":[0,1,2]}
print(d["labels"], d["vals"])
# print(d.label, d.vals) # ERROR

b = Bunch(labels=["a","b","c"], vals=[0,1,2])
print(b["labels"], b["vals"])
print(b.labels, b.vals)
```
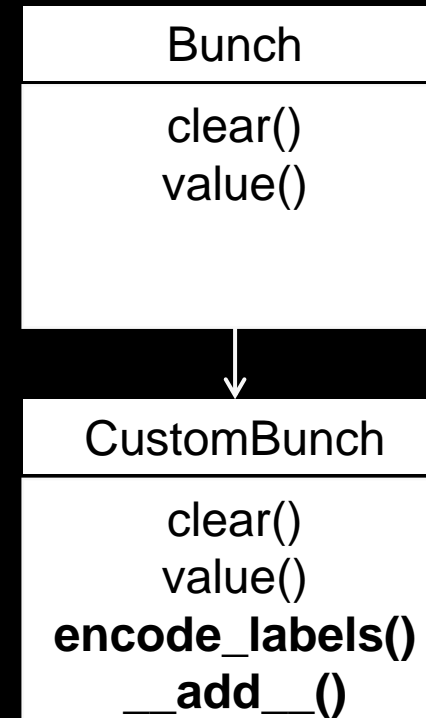
LABEL RAW SENSOR DATA

📄 *./data/captured/stationary.csv*

```
label,timestamp1,timestamp2,x_out,y_out,z_out
stationary,216321,216321,0.244,0.244,1000.0
stationary,216330,216330,0.488,1.465,999.756
stationary,216340,216340,0.488,0.0,1000.732
stationary,216350,216350,0.244,1.953,1001.953
stationary,216360,216360,0.488,0.488,999.512
stationary,216369,216369,0.732,1.709,999.512
stationary,216379,216379,-0.488,-0.488,999.756
stationary,216389,216389,0.244,0.488,1001.221
stationary,216399,216399,-0.244,-0.732,1001.221
stationary,216408,216408,-0.488,-0.732,1000.488
stationary,216418,216418,0.732,0.977,1000.488
stationary,216428,216428,-1.221,-0.488,1000.977
```

Bunch
clear()
value()

Container object that extends dictionaries by enabling values to be accessed by key òr attribute.

CustomBunch
clear()
value()
encode_labels()
__add__()

Hold data items and meta-data (e.g. attribute names, label names, etc)

HAN_UNIVERSITY
OF APPLIED SCIENCES

LABEL RAW SENSOR DATA

Input	Process	Output

./data/captured/<label_1>.csv

./data/captured/<label_2>.csv

./data/captured/<label_3>.csv

./capturing/data_plotter.py

./data/captured/
all_raw_time_series_data.png
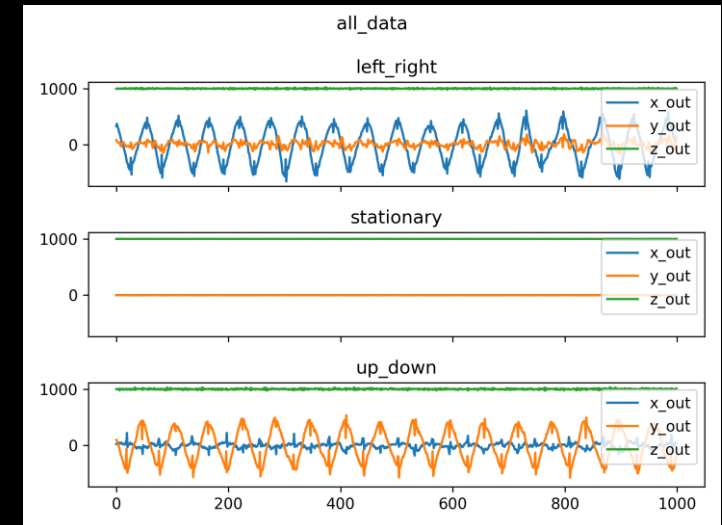
HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./../lib/filters.c

./../lib/filters.h

./preprocessing/filter_selection/
filter_functions_c2dll.py

*./data/preprocessed/features/
filters.dll*

PREPROCESSOR DESIGN

Input	Process	Output

./../lib/filters.c

./../lib/filters.h

./preprocessing/filter_selection/
filter_functions_c2dll.py

./data/preprocessed/features/
filters.dll

TBD
Shared Object (.so) file for Linux…

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./../lib/filters.c

./../lib/filters.h

*./preprocessing/filter_selection/
filter_functions_c2dll.py*

*./data/preprocessed/features/
filters.dll*

./../lib/normalizations.c

./../lib/normalizations.h

*./preprocessing/normalization_selection/
normalization_functions_c2dll.py*

*./data/preprocessed/features/
normalizations.dll*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./../lib/filters.c

./../lib/filters.h

./preprocessing/filter_selection/
filter_functions_c2dll.py

*./data/preprocessed/features/
filters.dll*

./../lib/normalizations.c

./../lib/normalizations.h

./preprocessing/normalization_selection/
normalization_functions_c2dll.py

*./data/preprocessed/features/
normalizations.dll*

./../lib/features.c

./../lib/features.h

./preprocessing/feature_selection/
feature_functions_c2dll.py

*./data/preprocessed/features/
features.dll*

PREPROCESSOR DESIGN

Input	Process	Output

./../lib/filters.c

./../lib/filters.h

./../lib/normalizations.c

./../lib/ normalizations.h

./../lib/features.c

./../lib/features.h

Source file in microcontroller IDE project

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN
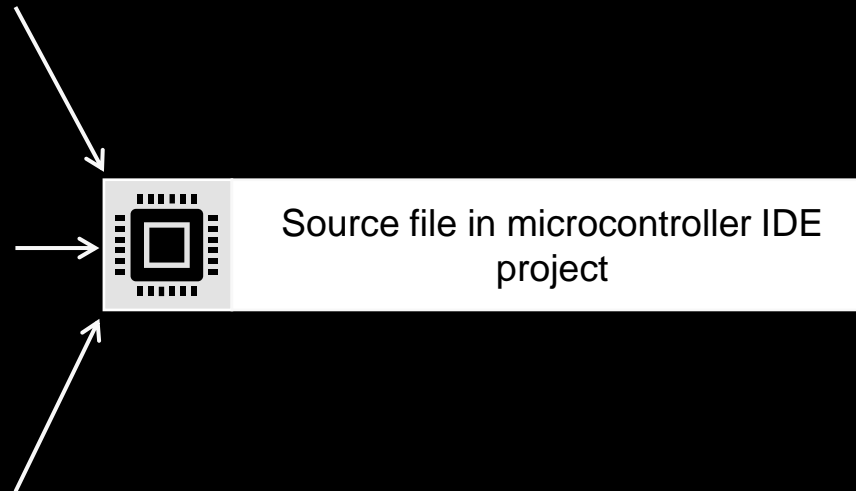
./../lib/filters.c

./../lib/filters.h

Supported filter functions

```
float fir(const float data, const float *coefs, float *x,
        const uint32_t n);
```

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/filters/
filters.dll   →   ./preprocessing/filter_selection/
filter_calculator.py   →

TODO

FILTER_FUNCTIONS: *create a list of filter functions to be calculated.*

Each filter function is calculated for each attribute!

x_out_fir
y_out_fir
z_out_fir

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

*./data/preprocessed/filters/
filters.dll*

 ./preprocessing/filter_selection/
filter_calculator.py

TODO
ARGS: *Set filter specific values, such as FIR coefficients.*

Check fir_coefs_calculator.py for helper functions!

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/filters/
filters.dll

./preprocessing/filter_selection/
filter_calculator.py

TODO
INPUT_DIR_PATH: *defaults to raw captured data*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/filters/
filters.dll →

./preprocessing/filter_selection/
filter_calculator.py

./data/captured/<label_1>.csv

./data/captured/<label_2>.csv

./data/captured/<label_3>.csv

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/filters/ filters.dll

./data/captured/<label_1>.csv

./data/captured/<label_2>.csv

./data/captured/<label_3>.csv

./preprocessing/filter_selection/ filter_calculator.py

./data/preprocessed/filters/ <label_1>.csv

./data/preprocessed/filters / <label_2>.csv

./data/preprocessed/filters / <label_3>.csv

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

📄 *./data/preprocessed/filters/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir,y_out_fir,z_out_fir
stationary,,,983.7379150390625,0.4375389218303833,20.670734405517578
stationary,,,919.2149047851562,0.4448928236961365,85.5930786328125
stationary,,,751.0918579101562,0.3765883147716522,251.9529418945312
stationary,,,500.9484558105469,0.4315893352031708,500.3926086425781
stationary,,,251.8128662109375,0.5671759843826294,748.8976440429688
stationary,,,85.34004211425781,0.789832353591919,915.3922729492188
stationary,,,20.548179626464844,0.9223169088363647,980.3436889648438
stationary,,,0.3947712481021881,0.969209194183496,1000.39705078125
stationary,,,0.3150411546230316,0.836254060268402,1000.188293457031
stationary,,,0.1824399381875918,0.530323147737427,1000.170654296875
stationary,,,0.040251106023788445,0.205855578184127,1000.414001464438
stationary,,,-0.08218221366405487,-0.0893305838108067,1000.657348632812
```

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

📄 *./data/preprocessed/filters/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir,y_out_fir,z_out_fir
stationary,,,983.7379150390625,0.43753892183303833,20.670734405517578
stationary,,,919.2149047851562,0.4448928236961365,85.5930786328125
stationary,,,751.0918579101562,0.3765883147716522,251.9529418945312
stationary,,,500.9484558105469,0.4315893352031708,500.3926086425781
stationary,,,251.8128662109375,0.5671759843826294,748.8976440429688
stationary,,,85.34004211425781,0.789832353591919,915.3922729492188
stationary,,,20.548179626464844,0.9223169088363647,980.3436889648438
stationary,,,0.3947712481021881,0.969209191833496,1000.397705078125
stationary,,,0.3150411546230316,0.836254060268421,1000.188293457012
stationary,,,0.1824399381875918,0.530323147737427,1000.170654296875
stationary,,,0.0402511060237845,0.205855578184127,1000.414001464438
stationary,,,-0.08218221366405487,-0.08933058381080627,1000.6573486328125
```

Label is preserved.

PREPROCESSOR DESIGN

📄 *./data/preprocessed/filters/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir,y_out_fir,z_out_fir
stationary,,,983.7379150390625,0.43753892183303833,20.670734405517578
stationary,,,919.2149047851562,0.4448928236961365,85.5930786132812
stationary,,,751.0918579101562,0.3765883147716522,251.95294189453125
stationary,,,500.9484558105469,0.4315893352031708,500.3926086425781
stationary,,,251.8128662109375,0.5671759843826294,748.8976440429688
stationary,,,85.34004211425781,0.789832353591919,915.3922729492188
stationary,,,20.548179626464844,0.922316908363647,980.3436889648438
stationary,,,0.3947712481021881,0.969209941833496,1000.397705078125
stationary,,,0.315041154623016,0.8362540602684021,1000.1882934570312
stationary,,,0.1824393818759918,0.530323147737427,1000.170654296875
stationary,,,0.04025110602378845,0.2058555781841278,1000.4140014648438
stationary,,,-0.08218221366405487,-0.0893305838108627,1000.6573486328125
```

Removed for further processing, because there is no way of knowing how long this takes on the microcontroller

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

📄 *./data/preprocessed/filters/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir,y_out_fir,z_out_fir
stationary,,,983.7379150390625,0.4375389218303833,20.670734405517578
stationary,,,919.2149047851562,0.4448928236961365,85.5930786138125
stationary,,,751.0918579101562,0.3765883147716522,251.9529418943125
stationary,,,500.9484558105469,0.4315893352031708,500.3926086425781
stationary,,,251.8128662109375,0.5671759843826294,748.8976440429688
stationary,,,85.34004211425781,0.789832353591919,915.3922729492188
stationary,,,20.548179626464844,0.9223169088363647,980.3436889648438
stationary,,,0.3947712481021881,0.969209194183496,1000.39705078125
stationary,,,0.3150411546230316,0.8362540602684021,1000.188293570312
stationary,,,0.1824399318759918,0.530323147737427,1000.17065429875
stationary,,,0.04025110602378845,0.2058555781841278,1000.414001464438
stationary,,,-0.08218221366405487,-0.08933058381080627,1000.65734863825
```

Calculated by the Python script

Naming convention:

<attribute name>_<filter name>

27

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/filters/ <label_1>.csv

./data/preprocessed/filters/ <label_2>.csv

./data/preprocessed/filters/ <label_3>.csv

./preprocessing/filter_selection/filter_plotter.py

./data/preprocessed/filters/ all_filters_time_series_data.png

HAN_UNIVERSITY
OF APPLIED SCIENCES
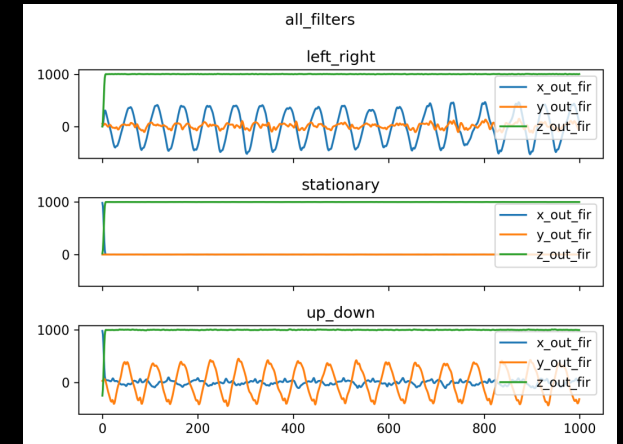
PREPROCESSOR DESIGN
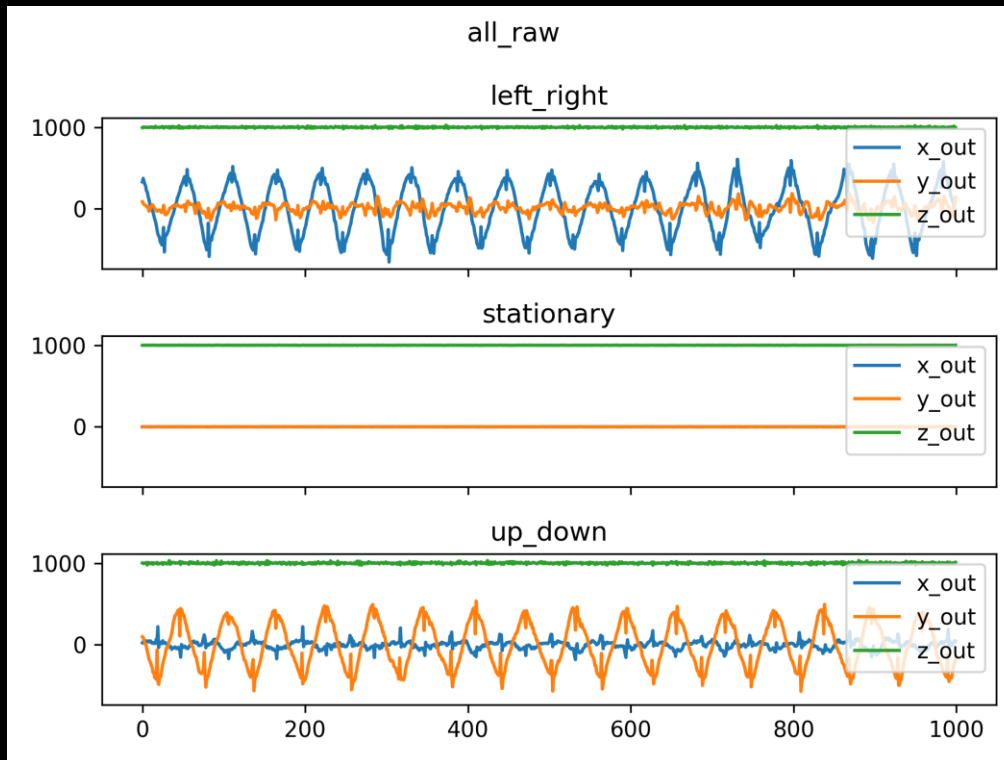
FIR low pass filter example

f_s=100 Hz and f_{cutoff}=2 Hz

29

HAN_UNIVERSITY
OF APPLIED SCIENCES
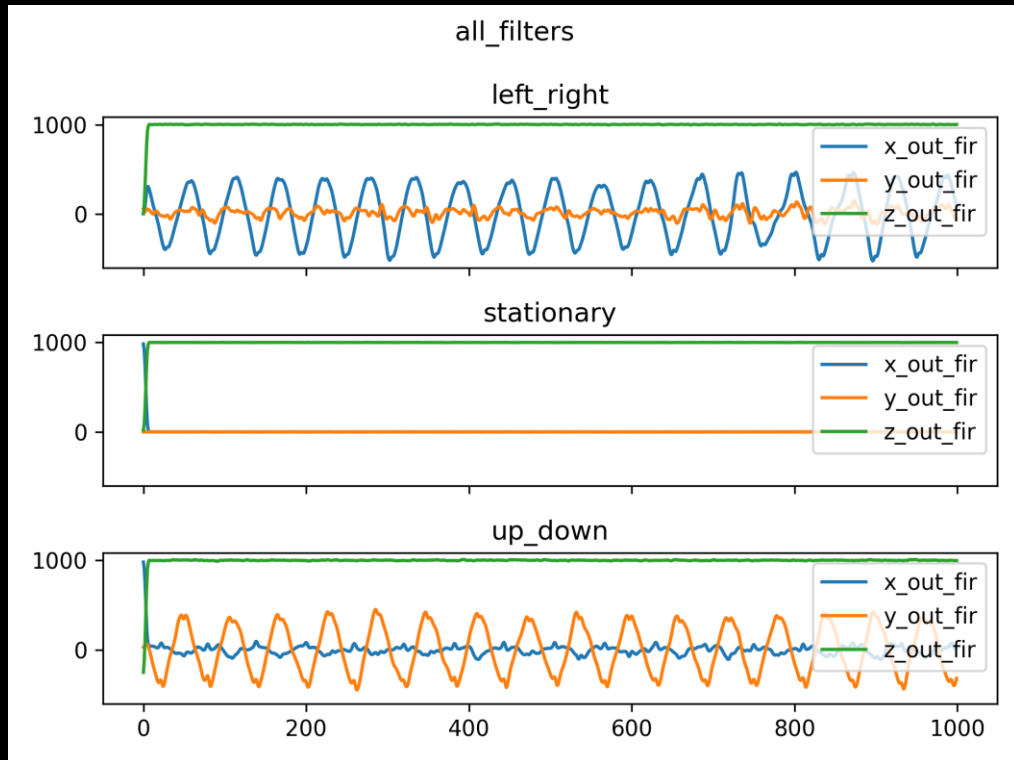
PREPROCESSOR DESIGN

Supported normalization functions

./../lib/normalizations.c

```
float rescale(const float data, const float from[2], float to[2]);
float clip(const float data, const float min[1], float max[1])
```

./../lib/normalizations.h

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/normalizations/normalizations.dll → ./preprocessing/normalization_selection/normalization_calculator.py →

TODO

NORMALIZATION_FUNCTIONS*: create a list of normalization functions to be calculated.*

Each normalization function is calculated for each attribute!

x_out_..._rescale
y_out_..._rescale
z_out_..._rescale

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/normalizat ions/normalizations.dll → ./preprocessing/normalization_selec tion/normalization_calculator.py →

TODO
ARGS: Set normalization specific values, such as

Rescale:
FROM: natural range of the sensor data
TO: new range of the sensor data

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/normalizat ions/normalizations.dll → *./preprocessing/normalization_selec tion/normalization_calculator.py* →

TODO
ARGS: Set normalization specific values, such as

Clip:
MIN: minimum new value
MAX: maximum new value

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/normalizat ions/normalizations.dll → ./preprocessing/normalization_selec tion/normalization_calculator.py →

TODO
INPUT_DIR_PATH: *set input files.*

Defaults to filtered files, but raw files can also be selected if filtering is not needed

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/normalizations/normalizations.dll → ./preprocessing/normalization_selection/normalization_calculator.py

./data/…/<label_1>.csv

./data/…/<label_2>.csv

./data/…/<label_3>.csv

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/normalizations/normalizations.dll

./data/…/<label_1>.csv

./data/…/<label_2>.csv

./data/…/<label_3>.csv

./preprocessing/normalization_selection/normalization_calculator.py

./data/preprocessed/ normalizations/<label_1>.csv

./data/preprocessed/ normalizations/<label_2>.csv

./data/preprocessed/ normalizations/<label_3>.csv
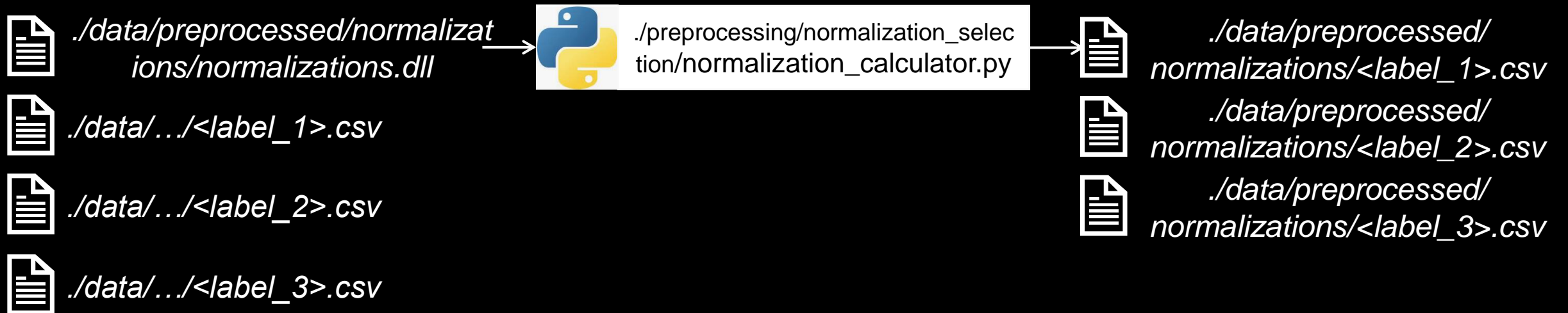
PREPROCESSOR DESIGN

📄 *./data/preprocessed/normalizations/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir_rescale,y_out_fir_rescale,z_out_fir_rescale
stationary,,,0.9837380647659302,0.00043761730194091797,0.020670771598815918
stationary,,,0.9192149639129639,0.00044488906860635156,0.08559310436248779
stationary,,,0.7510918378829956,0.0003765821456909179,0.2519528865814209
stationary,,,0.5009485483169556,0.0004316568374633789,0.5003926753997803
stationary,,,0.2518129348754883,0.0005671977996826172,0.7488977909088135
stationary,,,0.0853014225006104,0.0007898807525634766,0.9153923988342285
stationary,,,0.02054822444915771,0.0009223222732543945,0.9803438186645508
stationary,,,0.0003948211669921875,0.0009692907333374023,1.0003976821899414
stationary,,,0.00031507015228271484,0.0008362531661987305,1.0001883506774902
stationary,,,0.000182509422302461,0.0005303621292114258,1.0001707077026367
stationary,,,4.029273986816406e-05,0.0002058744430541992,1.0004141330718994
stationary,,,-8.213520050048828e-05,-8.928775787353516e-05,1.00065755844162
```

PREPROCESSOR DESIGN

📄 *./data/preprocessed/normalizations/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir_rescale,y_out_fir_rescale,z_out_fir_rescale
stationary,,,0.9837380647659302,0.00043761730194091797,0.020670771598815918
stationary,,,0.9192149639129639,0.0004448890686035156,0.0855931043624879
stationary,,,0.7510918378829956,0.0003765821456909179,0.251952886581429
stationary,,,0.500948548316956,0.0004316568374633789,0.5003926753997803
stationary,,,0.2518129348754883,0.0005671977996826172,0.7488977909088135
stationary,,,0.0853401422506104,0.0007898807525634766,0.9153923988342285
stationary,,,0.02054822449157715,0.0009232322732543945,0.9803438186645508
stationary,,,0.0003948211669921875,0.000969290733374023,1.000397682189414
stationary,,,0.0003150701522271484,0.000836253166198305,1.000188506774902
stationary,,,0.000182509423022461,0.0005303621292114258,1.000170707026367
stationary,,,4.02927398681406e-05,0.0002058744430541992,1.0004141330718994
stationary,,,-8.21352005004828e-05,-8.928775787353516e-05,1.00065755844116
```

Label is preserved.

38

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

📄 *./data/preprocessed/normalizations/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir_rescale,y_out_fir_rescale,z_out_fir_rescale
stationary,,,0.9837380647659302,0.00043761730194091797,0.02067077159881591
stationary,,,0.9192149639129639,0.0004448890686035156,0.0855931043624877
stationary,,,0.7510918378829956,0.0003765821456909179,0.25195288658142
stationary,,,0.5009485483169556,0.0004316568374633789,0.5003926753997803
stationary,,,0.2518129348754883,0.0005671977996826172,0.7488977909088135
stationary,,,0.08534014225006104,0.0007898807525634766,0.9153923988342285
stationary,,,0.02054822449157715,0.0009232322732543945,0.9803438186645508
stationary,,,0.000394211669921875,0.000969907333374023,1.0003976821899414
stationary,,,0.0003150701522827148,0.000836253166198730,1.000188350677490
stationary,,,0.000182509423022461,0.0005303621292114258,1.0001707077026367
stationary,,,4.029273986816406e-05,0.0002058744430541992,1.0004141330718994
stationary,,,-8.213520050048828e-05,-8.928775787353516e-05,1.000657558441162
```

Removed for further processing, because there is no way of knowing how long this takes on the microcontroller

PREPROCESSOR DESIGN

📄 *./data/preprocessed/normalizations/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir_rescale,y_out_fir_rescale,z_out_fir_rescale
stationary,,,0.9837380647659302,0.00043761730194091797,0.020670771598815918
stationary,,,0.9192149639129639,0.00044488906860035156,0.0855931043624779
stationary,,,0.7510918378829956,0.0003765821456909179,0.251952865814209
stationary,,,0.500945483169556,0.0004316568374633789,0.5003926753997803
stationary,,,0.2518129348754883,0.0005671977996826172,0.7488977909088135
stationary,,,0.0853401422506104,0.0007898807525634766,0.915392988342285
stationary,,,0.02054822449157715,0.0009223222732543945,0.9803438186645508
stationary,,,0.000394821166992187 5,0.0009692907333374023,1.000397682189414
stationary,,,0.0003150701522827148 4,0.000836253166198730 5,1.000188350677490 2
stationary,,,0.0001825094223022461,0.0005306212921142 58,1.0001707077026367
stationary,,,4.029273986816406e-05,0.0002058744430541992 2,1.000414133071899 4
stationary,,,-8.213520050048828e-05,-8.928775787353516e-05,1.000657558441162
```

Calculated by the Python script

Naming convention:

*<attribute name>_
<filter name>*

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/ normalizations/<label_1>.csv

./data/preprocessed/ normalizations/<label_2>.csv

./data/preprocessed/ normalizations/<label_3>.csv

./preprocessing/normalization_selection/normalization_plotter.py

./data/preprocessed/normalizations/ all_filters_time_series_data.png
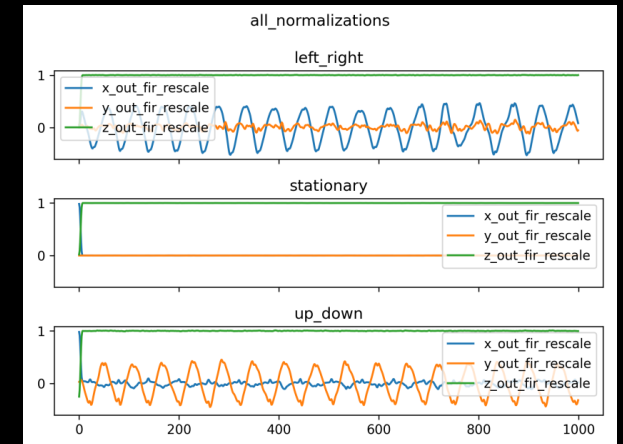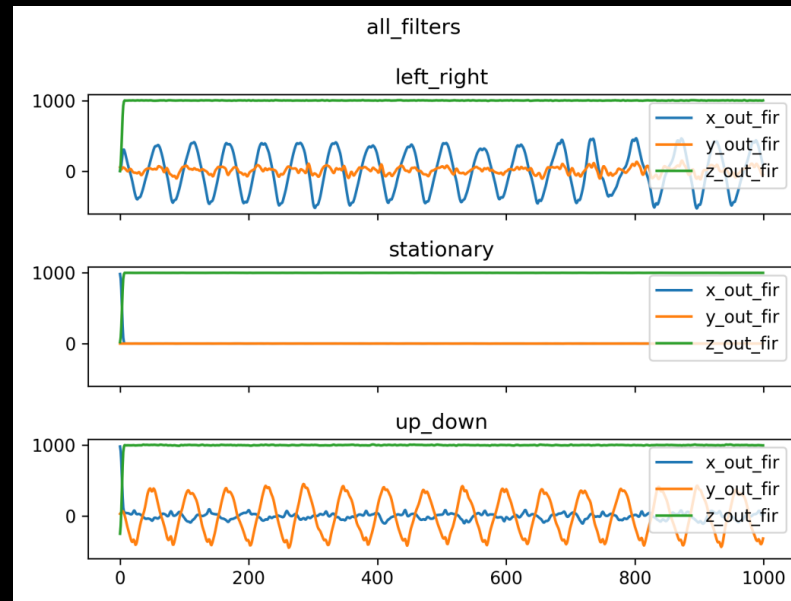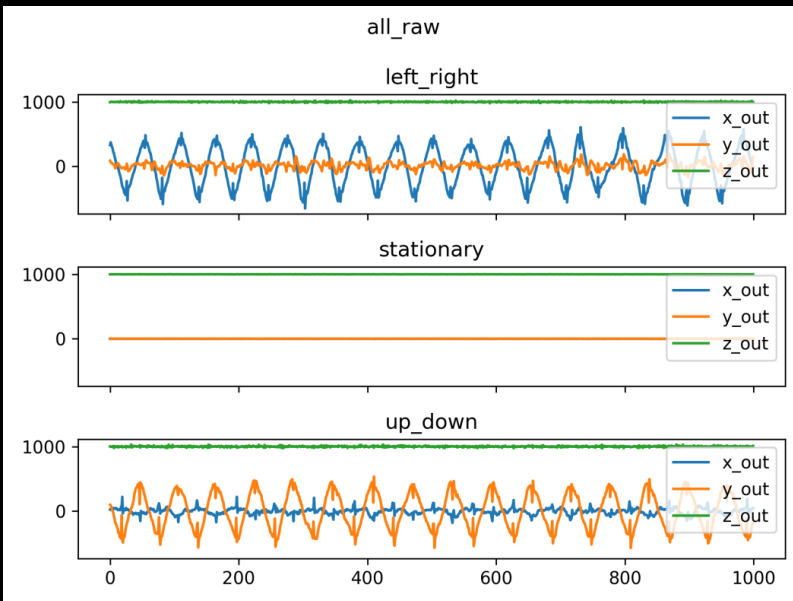
HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

FIR low pass filter example

Rescale example

f_s=100 Hz and f_{cutoff}=2 Hz

to_min=-1 and to_max=1

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

FIR low pass filter example

Clip example

f_s=100 Hz and f_{cutoff}=2 Hz

min=0 and max=1000

HAN_UNIVERSITY
OF APPLIED SCIENCES
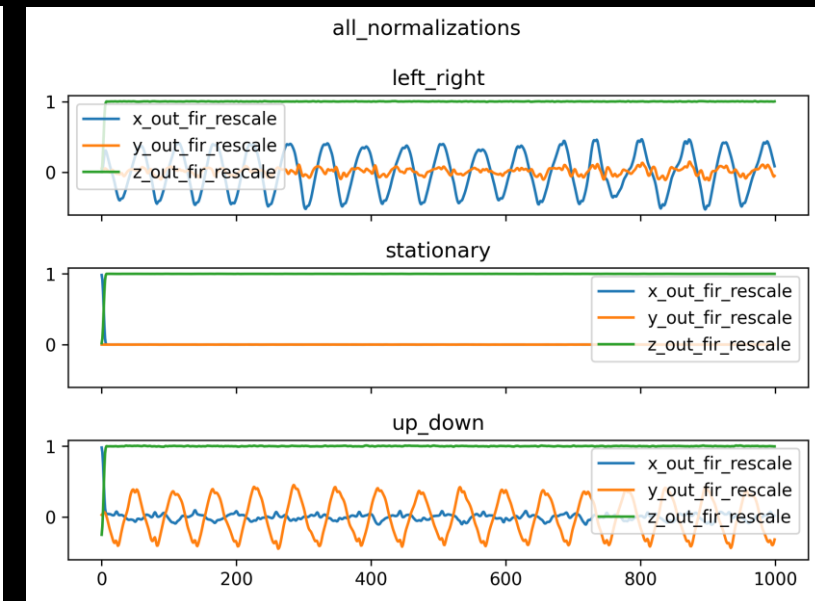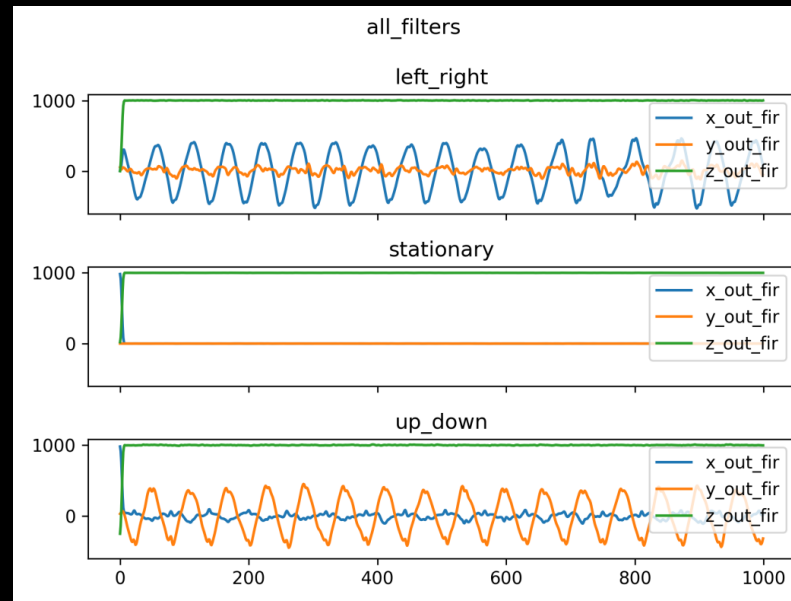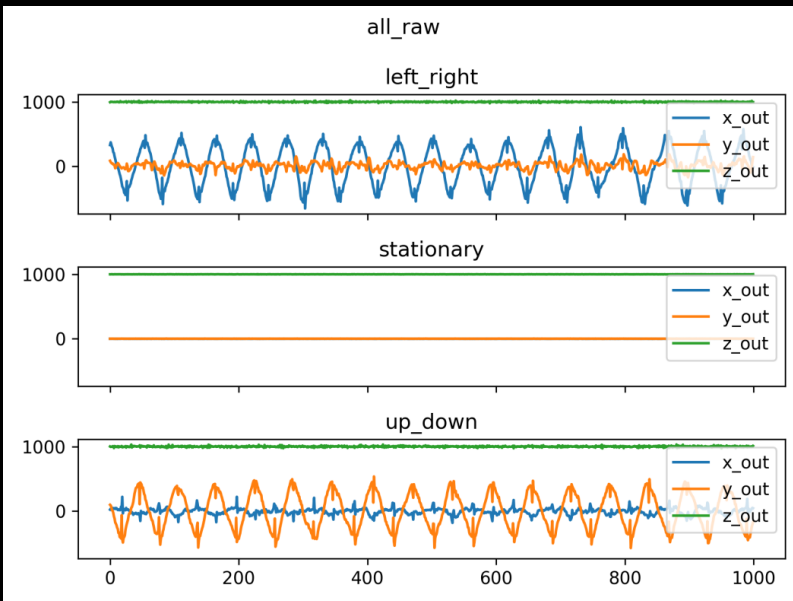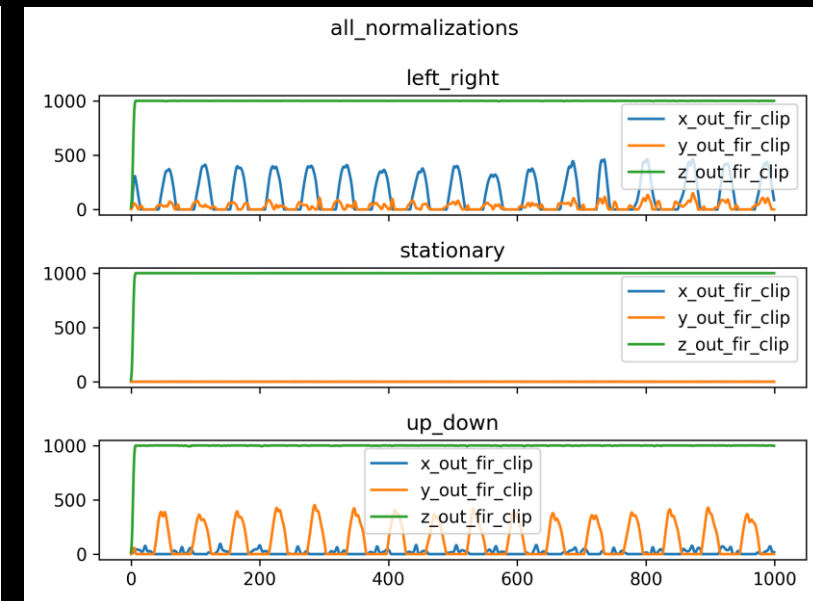
PREPROCESSOR DESIGN

Supported feature functions

📄 *./../lib/features.c*

📄 *./../lib/features.h*

```c
float min(float *data, const uint32_t n);
float mean(float *data, const uint32_t n);
float variance(float *data, const uint32_t n);
float energy(float *data, const uint32_t n);
float peak_to_peak(float *data, const uint32_t n);
```

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/features/ features.dll → ./preprocessing/feature_selecti on/feature_calculator.py →

TODO

FEATURE_FUNCTIONS: *create a list of feature functions to be calculated.*

Each feature function is calculated for each attribute!

x_out_..._mean
y_out_..._mean
z_out_..._mean

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/features/ features.dll → ./preprocessing/feature_selection/feature_calculator.py →

./config.py
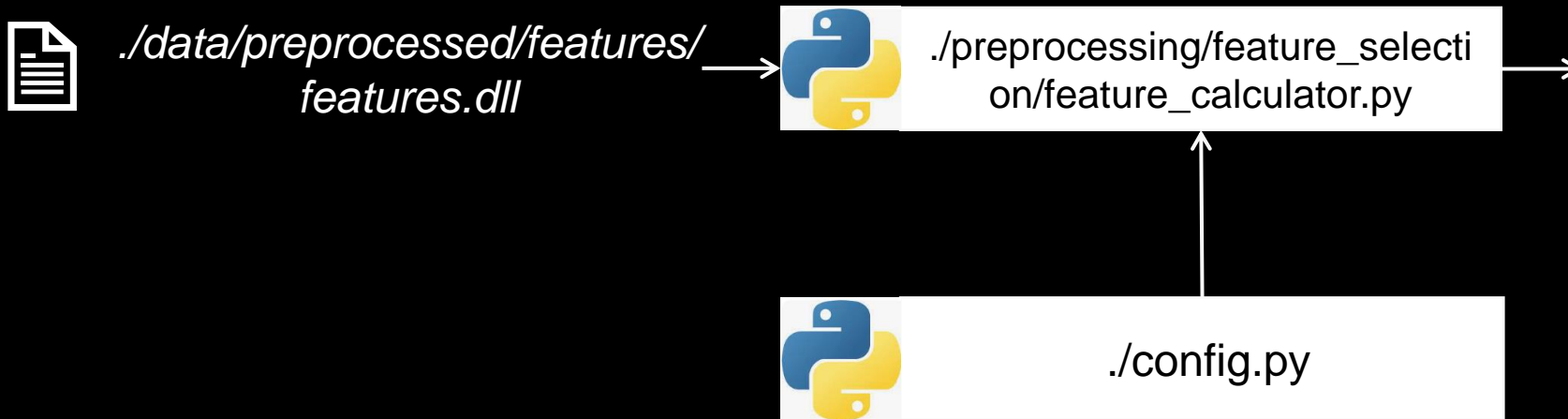
TODO
BLOCK_SIZE: *number of items used in feature calculation*
BLOCK_TYPE: *BLOCK or SLIDING*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

BLOCK_SIZE: *4*
BLOCK_TYPE: *BLOCK*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

3	5	7	6	10	30	43	45

t

BLOCK_SIZE: *4*
BLOCK_TYPE: *BLOCK*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

| 3 | 5 | 7 | 6 | 10 | 30 | 43 | 45 | 65 | 76 | 4 | 34 |

t

***BLOCK_SIZE**: 4*
***BLOCK_TYPE**: BLOCK*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

| 3 | 5 | 7 | 6 | 10 | 30 | 43 | 45 | 65 | 76 | 4 | 34 | 45 | 67 | 54 | 30 |

t

BLOCK_SIZE: *4*
BLOCK_TYPE: *BLOCK*

PREPROCESSOR DESIGN

3	5	7	6

t

BLOCK_SIZE: *4*
BLOCK_TYPE: *SLIDING*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

BLOCK_SIZE: *4*
BLOCK_TYPE: *SLIDING*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

BLOCK_SIZE: *4*
BLOCK_TYPE: *SLIDING*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

BLOCK_SIZE: *4*
BLOCK_TYPE: *SLIDING*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

BLOCK_SIZE: 4
BLOCK_TYPE: SLIDING

PREPROCESSOR DESIGN

BLOCK_SIZE: *4*
BLOCK_TYPE: *SLIDING*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

BLOCK_SIZE: *4*
BLOCK_TYPE: *SLIDING*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

BLOCK_SIZE: *4*
BLOCK_TYPE: *SLIDING*

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/features/features.dll

./data/.../<label_1>.csv

./data/.../<label_2>.csv

./data/.../<label_3>.csv

./preprocessing/feature_selection/feature_calculator.py

./config.py

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/features/ features.dll

→

 ./preprocessing/feature_selection/feature_calculator.py

→

 ./data/preprocessed/features/ <label_1>.csv

./data/…/<label_1>.csv

./data/preprocessed/features/ <label_2>.csv

./data/…/<label_2>.csv

 ./config.py

./data/preprocessed/features/ <label_3>.csv

./data/…/<label_3>.csv

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

📄 *./data/preprocessed/features/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir_rescale_variance,y_out_fir_rescale_variance,z_out_fir_rescale_variance
stationary,,,0.025755632668733597,1.0166539254896634e-07,0.025553688406944275
stationary,,,4.993831126398618e-08,7.614894315111087e-08,1.7571814225902926e-07
stationary,,,7.28732914012653e-08,9.023828795307054e-08,1.0345027590119571e-07
stationary,,,5.80202552669168e-08,5.29892894007844e-08,1.366340285358092e-07
stationary,,,7.567471982383722e-08,9.929044608725235e-08,1.751279370409975e-07
stationary,,,6.224084359018889e-08,1.2927046100230655e-07,1.268510061436245e-07
stationary,,,9.401501444017413e-08,1.4411538984404615e-07,1.432363490039279e-07
```

PREPROCESSOR DESIGN

📄 *./data/preprocessed/features/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir_rescale_variance,y_out_fir_rescale_variance,z_out_fir_rescale_variance
stationary,,,0.025755632668733597,1.0166539254896634e-07,0.025553688406944275
stationary,,,4.993831126398618e-08,7.614894315111087e-08,1.7571814225902926e-07
stationary,,,7.28732914012653e-08,9.023828795307054e-08,1.0345027590119571e-07
stationary,,,5.80202552691068e-08,5.29892894007844e-08,1.36634028535809e-07
stationary,,,7.567471982383722e-08,9.929044608725235e-08,1.751279370409975e-07
stationary,,,6.224084359018889e-08,1.2927046100230655e-07,1.268510061436245e-07
stationary,,,9.401501444017413e-08,1.4411538984404615e-07,1.43236349039279e-07
```

Number of samples depends on BLOCK_SIZE and BLOCK_MODE

PREPROCESSOR DESIGN

📄 *./data/preprocessed/features/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir_rescale_variance,y_out_fir_rescale_variance,z_out_fir_rescale_variance
stationary,,,0.025755632668733597,1.0166539254896634e-07,0.025553688406944275
stationary,,,4.993831126398618e-08,7.6148943151111087e-08,1.7571814225902926e-07
stationary,,,7.28732914012653e-08,9.023828795307054e-08,1.0345027590119571e-07
stationary,,,5.80025526691068e-08,5.29892894007844e-08,1.366340285358092e-07
stationary,,,7.567471982383722e-08,9.929044608725235e-08,1.751279370409975e-07
stationary,,,6.224084359018889e-08,1.2927046100230655e-07,1.268510061436245e-07
stationary,,,9.401501444017413e-08,1.4411538984404615e-07,1.43236349003927 9e-07
```

Label is preserved.

PREPROCESSOR DESIGN

📄 *./data/preprocessed/features/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir_rescale_variance,y_out_fir_rescale_variance,z_out_fir_rescale_variance
stationary,,,0.025755632668733597,1.0166539254896634e-07,0.025553688406944275
stationary,,,4.993831126398618e-08,7.6148943151111087e-08,1.7571814225902926e-07
stationary,,,7.28732914012653e-08,9.023828795307054e-08,1.0345027590119571e-07
stationary,,,5.80202552691068e-08,5.29892894007844e-08,1.36634028535092e-07
stationary,,,7.56747198238372e-08,9.929044608725235e-08,1.75127937040975e-07
stationary,,,6.224084359018889e-08,1.2927046100230655e-07,1.268510061432455e-07
stationary,,,9.401501444017413e-08,1.4411538984404615e-07,1.432363490039279e-07
```

Removed for further processing, because there is no way of knowing how long this takes on the microcontroller

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

📄 *./data/preprocessed/features/stationary.csv*

```
label,timestamp1,timestamp2,x_out_fir_rescale_variance,y_out_fir_rescale_variance,z_out_fir_rescale_variance
stationary,,,0.025755632668733597,1.0166539254896634e-07,0.025553688406944275
stationary,,,4.993831126398618e-08,7.614894315111087e-08,1.7571814225902926e-07
stationary,,,7.28732914012653e-08,9.023828795307054e-08,1.0345027590119571e-07
stationary,,,5.80202552691068e-08,5.29892894007844e-08,1.36634285358092e-07
stationary,,,7.567471982383722e-08,9.929044608725235e-08,1.751279370409975e-07
stationary,,,6.224084359018889e-08,1.2927046100230655e-07,1.268510614362455e-07
stationary,,,9.401501444017413e-08,1.4411538984404615e-07,1.432363490039279e-07
```

Added by the Python script

Naming convention:

<attribute name>_<feature name>

HAN_UNIVERSITY
OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/features/<label_1>.csv → *./preprocessing/feature_selection/feature_plotter.py* → *./data/preprocessed/features/all_features_time_series_data.png*
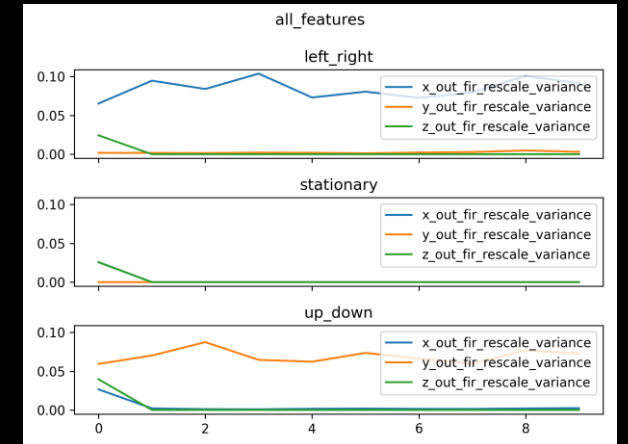
./data/preprocessed/features/<label_2>.csv

./data/preprocessed/features/<label_3>.csv

HAN_UNIVERSITY OF APPLIED SCIENCES

PREPROCESSOR DESIGN

Input	Process	Output

./data/preprocessed/features/<label_1>.csv

./data/preprocessed/features/<label_2>.csv

./data/preprocessed/features/<label_3>.csv

./preprocessing/feature_selection/feature_plotter.py

./data/preprocessed/features/.png*

HAN_UNIVERSITY
OF APPLIED SCIENCES
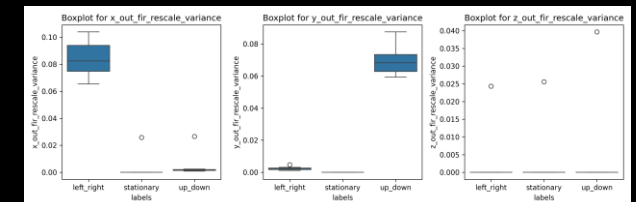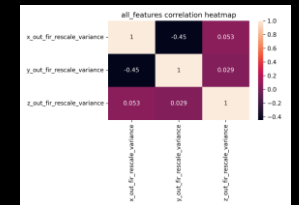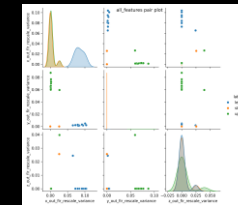
DEFINE YOUR OBJECTIVE

• What do you want to achieve?
  - > Define a SMART objective

• What classes apply?
• What data is available?
• What attributes are present?
• What data should be collected?

• What features matter?
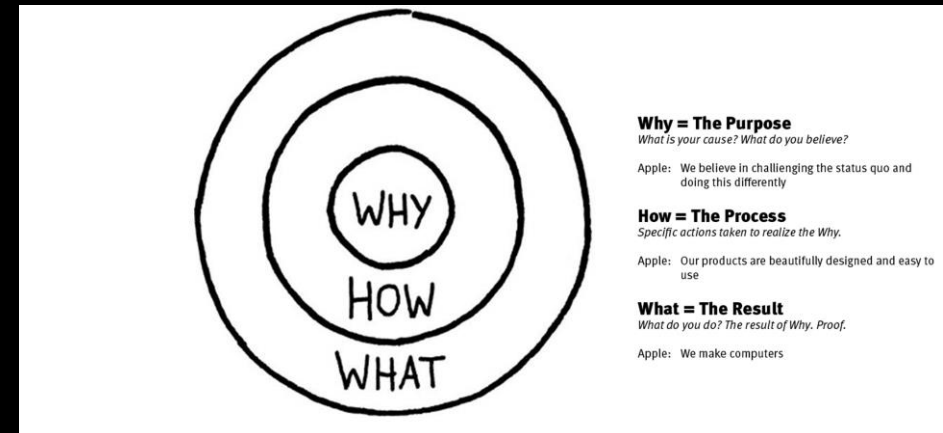
DATA MATTERS MORE THAN ALGORITHMS!

• Massive amounts of training data is needed
• Labelling is tedious and error prone
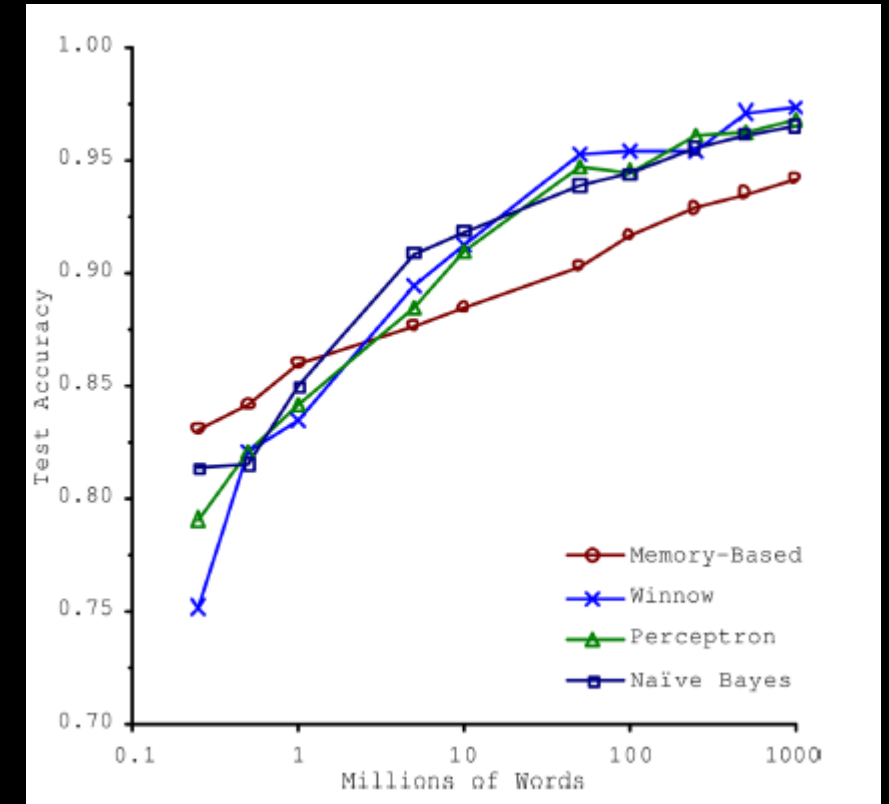
Source: Peter Norvig et al 2009

SPLITTING DATA

- Slice data into three subsets: Training, validation and test data

~60%	~20%	~20%
Training Set	Validation Set	Test Set

- Make sure that your subsets meet the following conditions:
  - Large enough to yield statistically meaningful results.
  - Representative of the data set as a whole.
    E.g. don't pick a test set with different characteristics than the training set.

STRATIFIED SAMPLING

- Make sure the subsets properly reflect the population!

population

sample

TRAINING, VALIDATION, TESTING

• Never train on test data!

HAN_UNIVERSITY
OF APPLIED SCIENCES
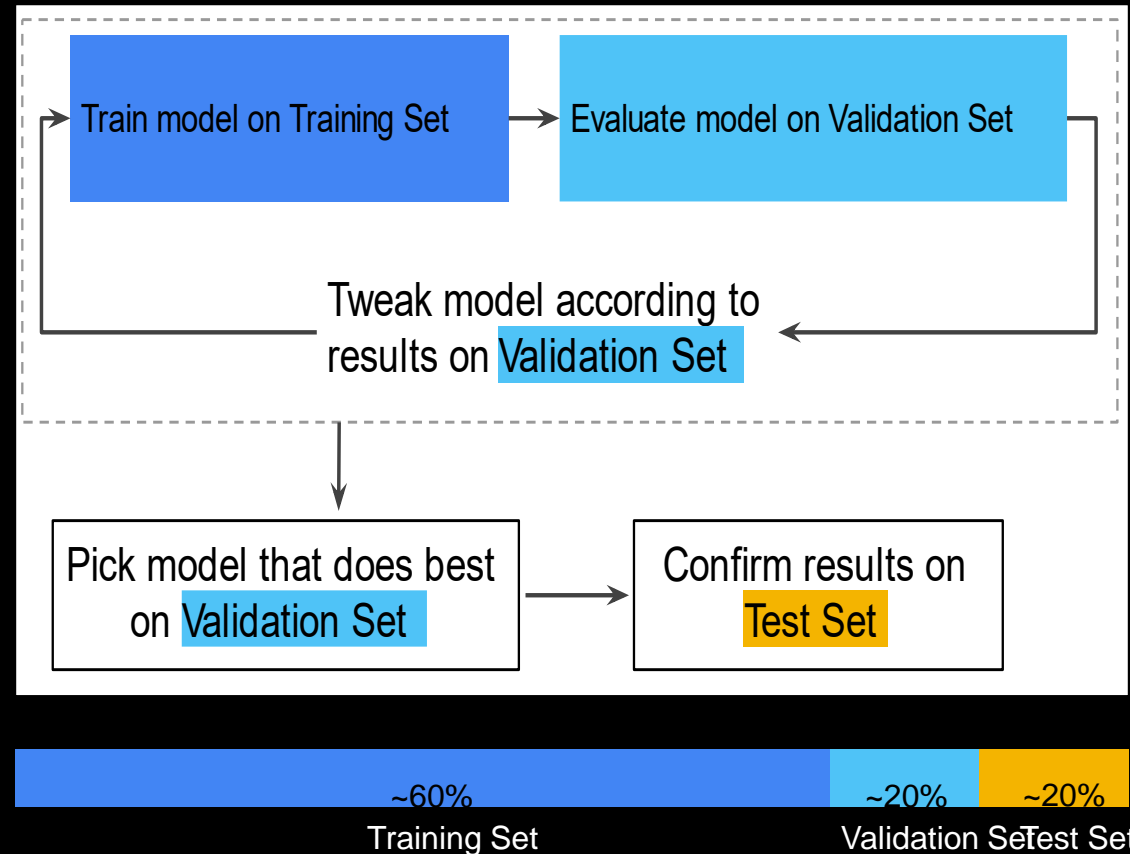
STRATIFIED SAMPLING

- Make sure the subsets set properly reflect the population

sample

population
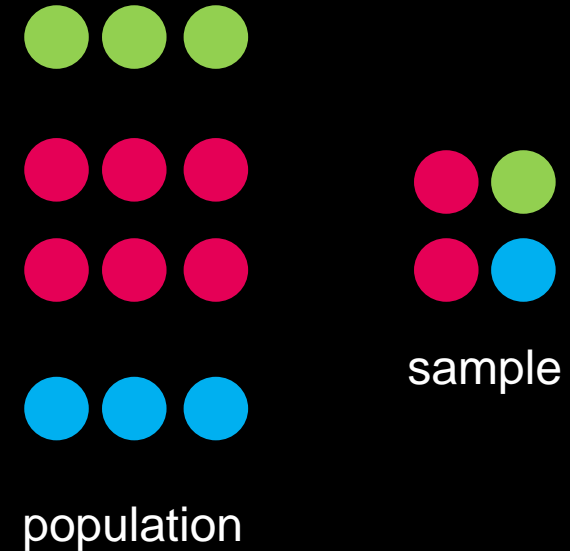
Never train on test data.
If you are seeing surprisingly good results on your evaluation metrics, it might be a sign that you are accidentally training on the test set. For example, high accuracy might indicate that test data has leaked into the training set.

FEATURE ENGINEERING

- Select features
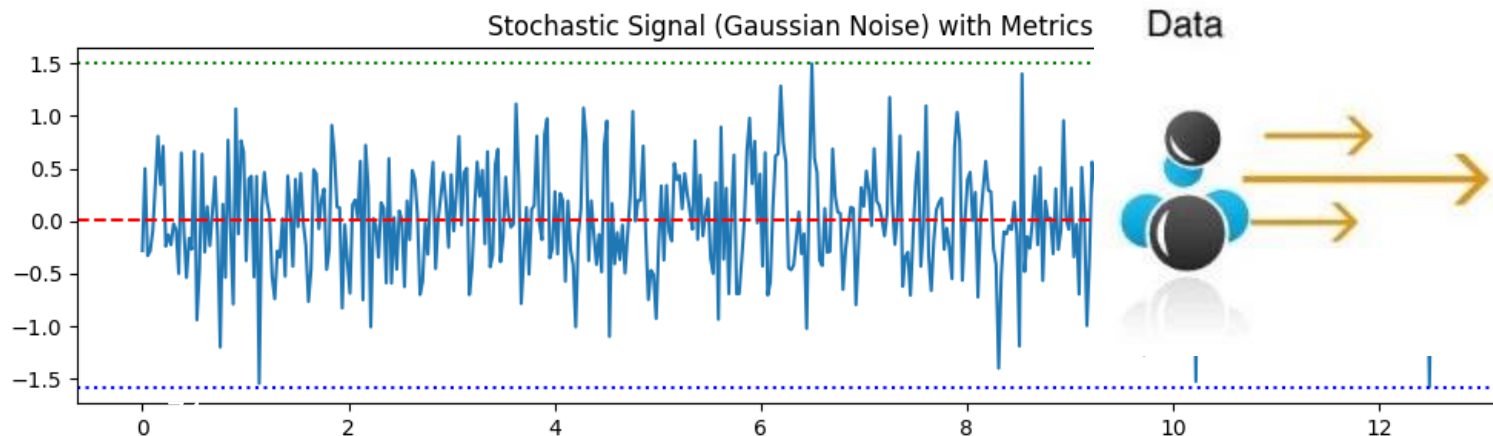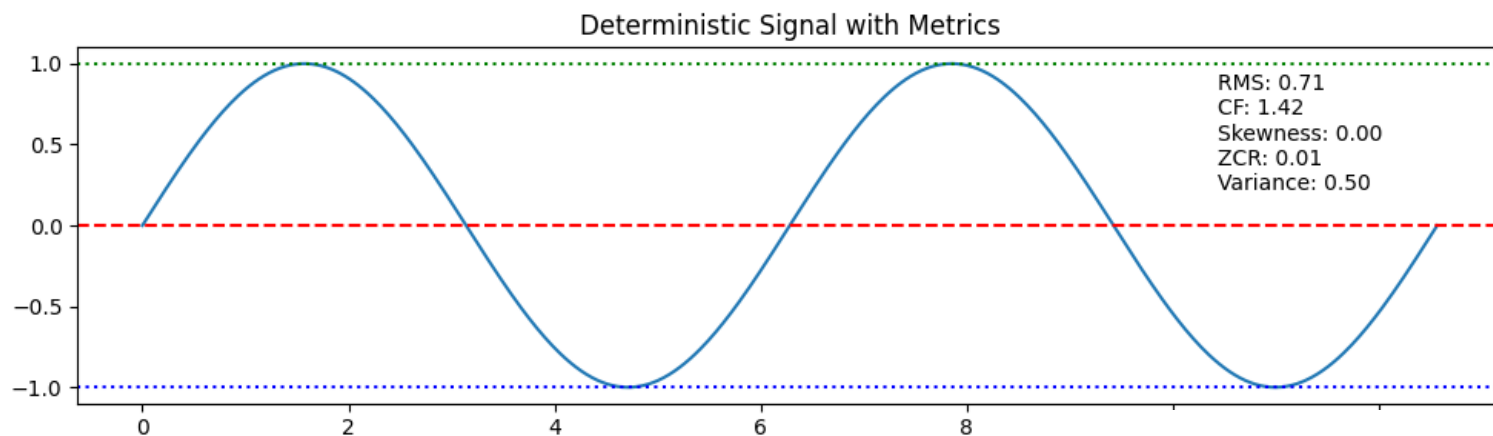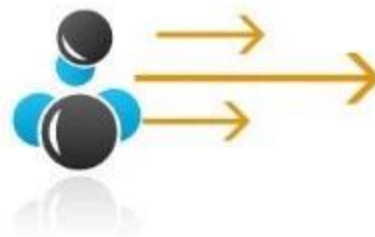
Examples
- Root Mean Square (RMS)
- Crest Factor
- Skewness and Kurtosis
- Zero Crossing Rate

DIGITAL SIGNAL PROCESSING

• Filtering

Deterministic Signal and Highpass Filtered Signal

RMS: 0.87
CF: 1.73
Skewness: -0.00
ZCR: 0.01
Variance: 0.04
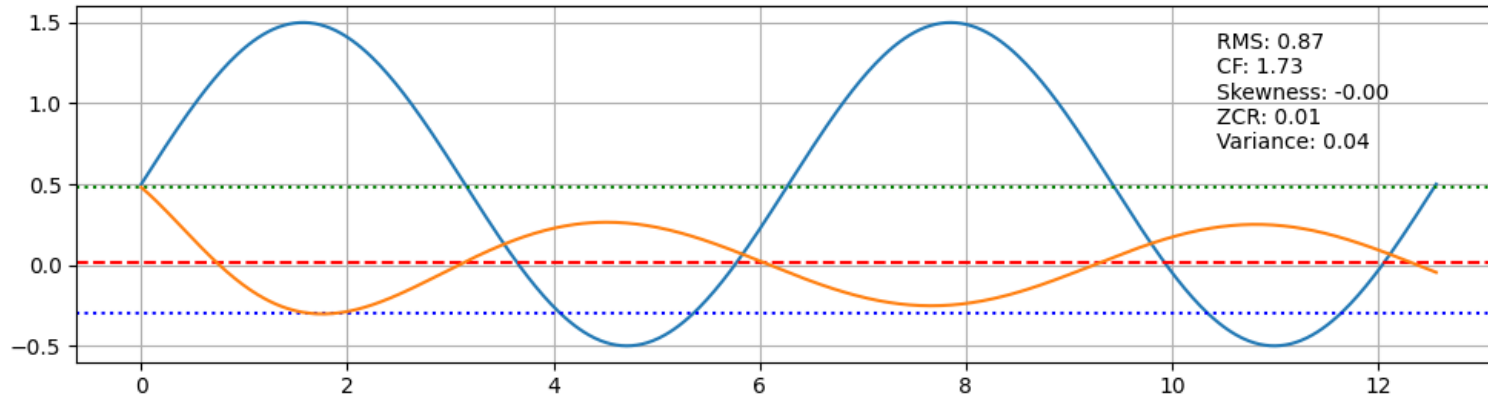
Stochastic Signal and Lowpass Filtered Signal

RMS: 1.01
CF: 2.72
Skewness: -0.06
ZCR: 0.20
Variance: 0.45
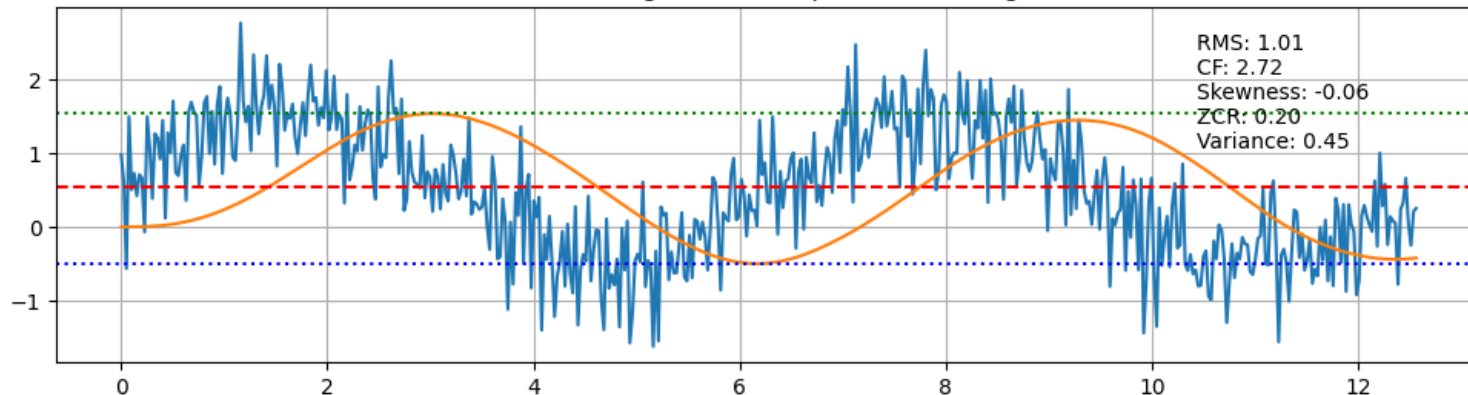
stochastic process

share price series

electrocardiogram

dynamical system model

electroencephalogram

space recording

heart beat interval sequence

air temperature

rainfall

random number sequence

audio

iterative map model

correlated noise process

QUALITIES OF GOOD FEATURES

• Informative
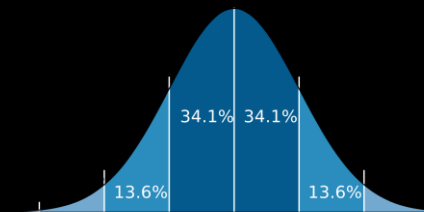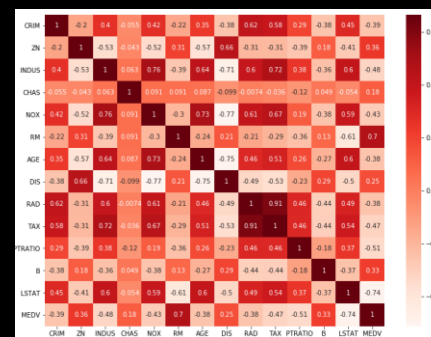• Discriminating
• Independent
• Nearly unique
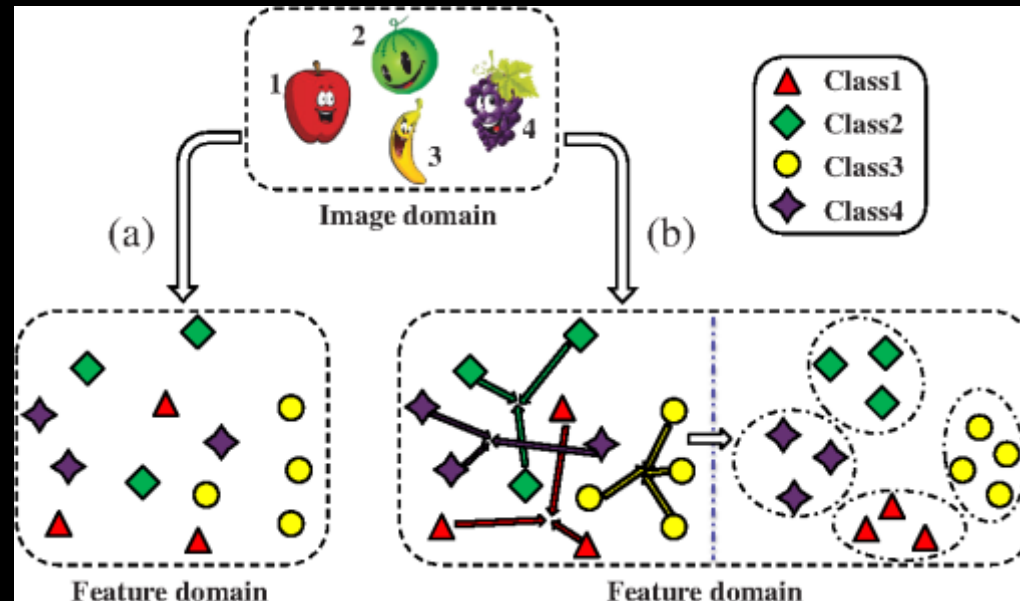
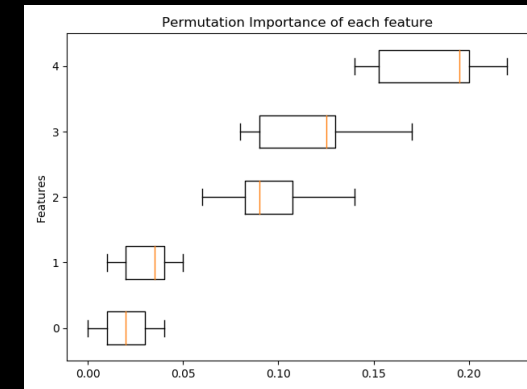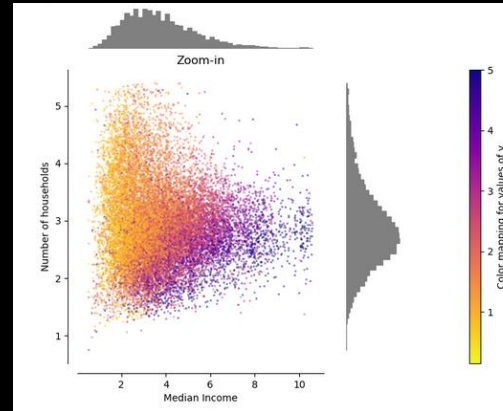

• NB later on feature scaling may be required

WHAT MAKES A GOOD FEATURE?

https://www.youtube.com/watch?v=N9fDIAflCMY&feature=youtu.be

TOOLS FOR EXPLORATORY DATA ANALYSIS

- Univariate analysis
- Histogram
- Scatterplot
- Boxplot
- Correlation heatmap

DATA PREPARATION ALGORITHMS

- Feature computation
- Cleaning, e.g. Fix or remove outliers
- Scaling, e.g. standardize or normalize features.