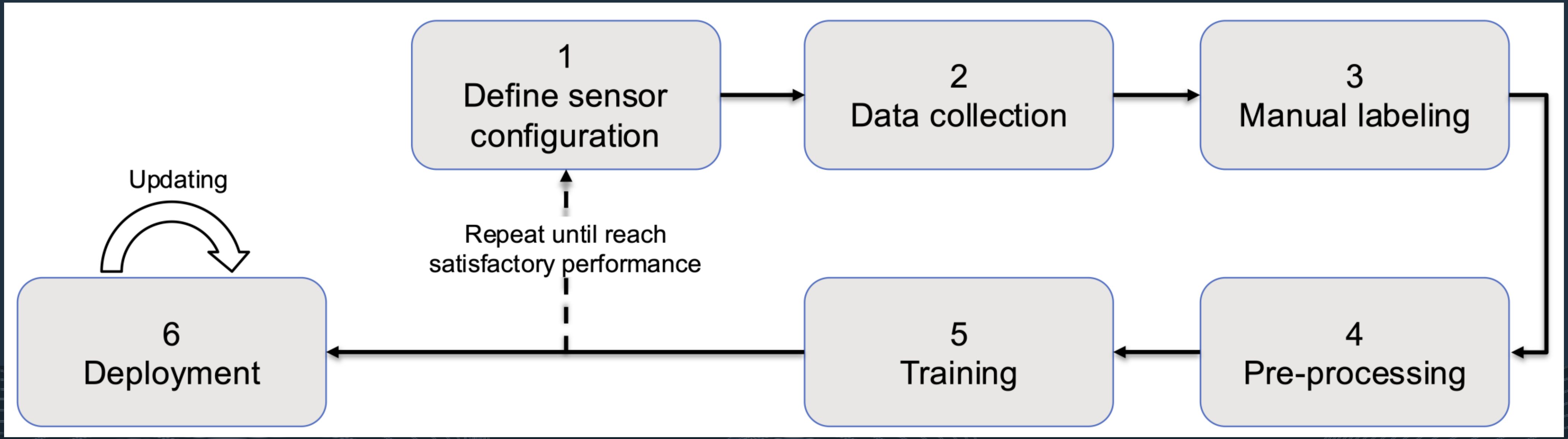
Processing data and Scikit-learn

Yao Zhang - Aalto University

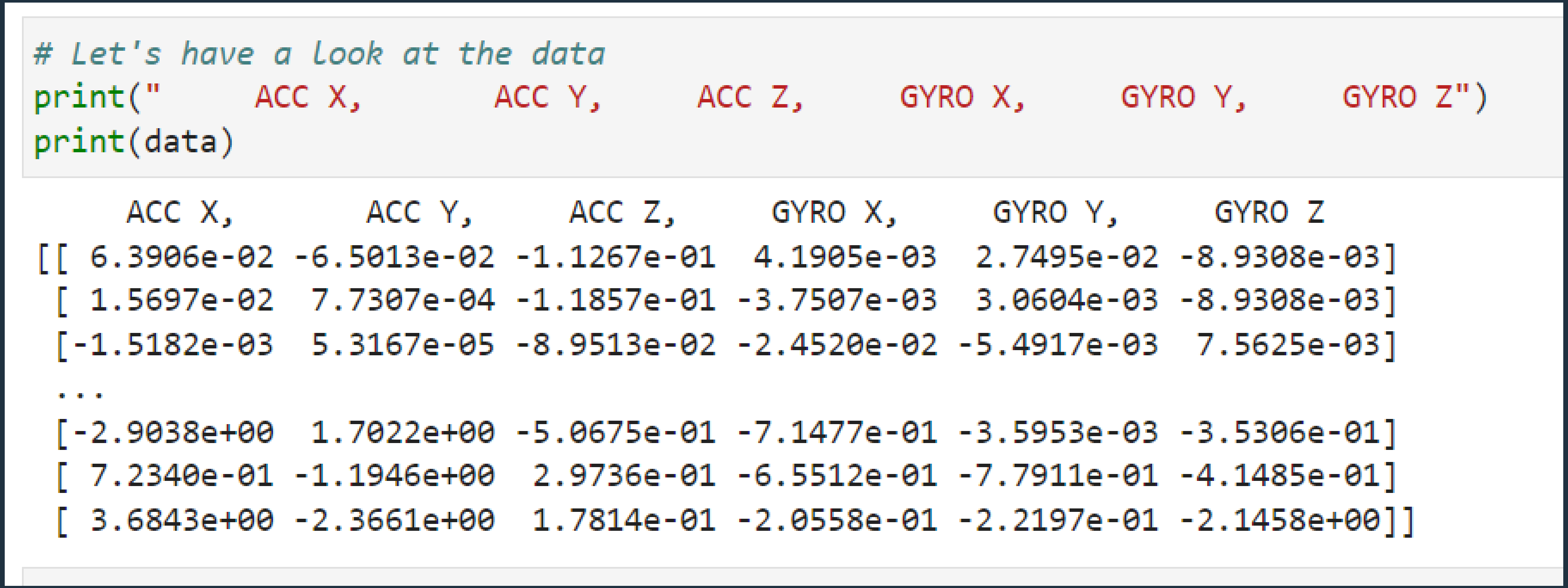
Expected Outcomes

- Learn how to load data from csv file and processing the data
- Learn the general processing data pipeline before used for training model

The general pipeline for develop a smart wearable



CSV file

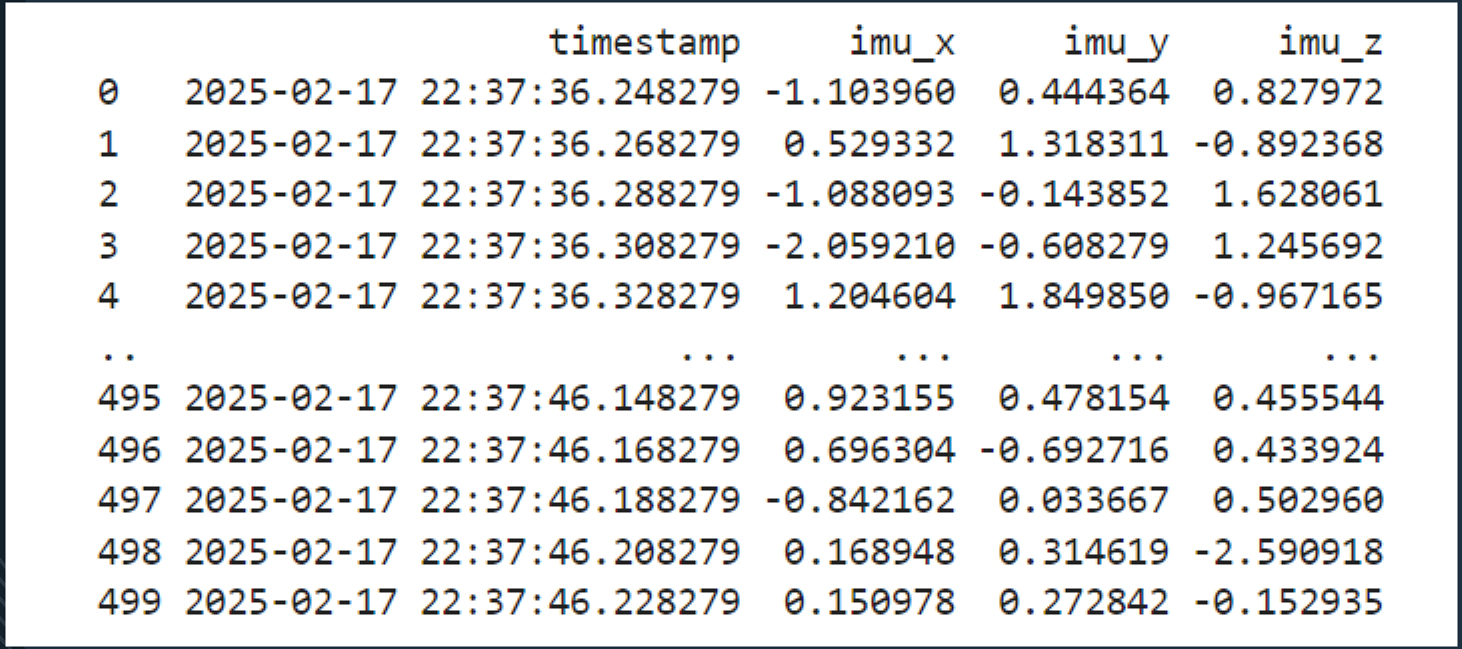


```
# Let's have a look at the data
print("     ACC X,       ACC Y,     ACC Z,     GYRO X,     GYRO Y,     GYRO Z")
print(data)
```

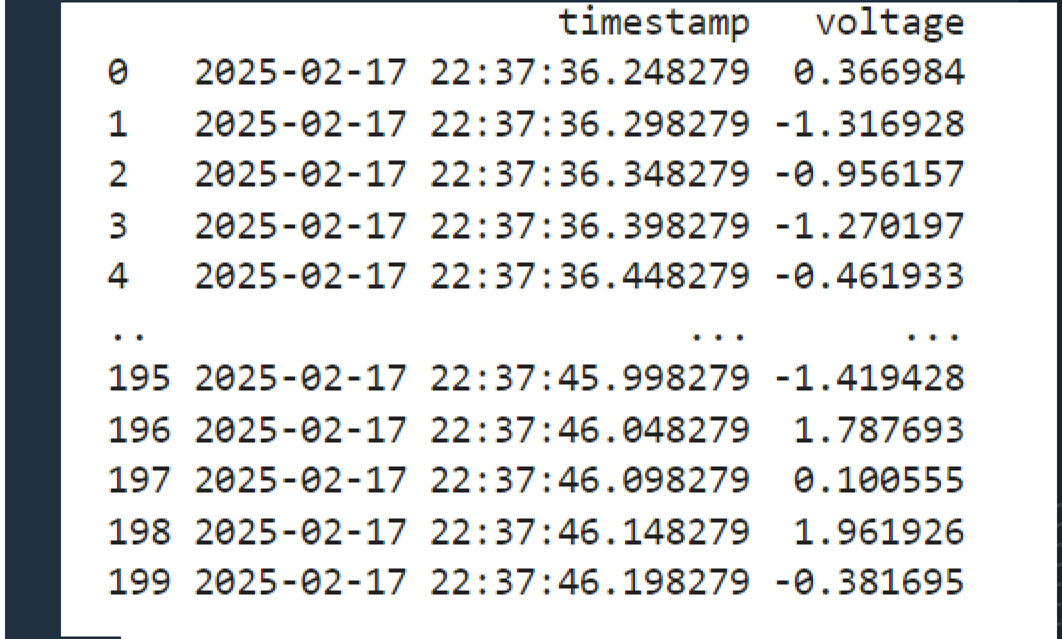
```
     ACC X,       ACC Y,     ACC Z,     GYRO X,     GYRO Y,     GYRO Z
[[ 6.3906e-02 -6.5013e-02 -1.1267e-01  4.1905e-03  2.7495e-02 -8.9308e-03]
 [ 1.5697e-02  7.7307e-04 -1.1857e-01 -3.7507e-03  3.0604e-03 -8.9308e-03]
 [-1.5182e-03  5.3167e-05 -8.9513e-02 -2.4520e-02 -5.4917e-03  7.5625e-03]
 ...
 [-2.9038e+00  1.7022e+00 -5.0675e-01 -7.1477e-01 -3.5953e-03 -3.5306e-01]
 [ 7.2340e-01 -1.1946e+00  2.9736e-01 -6.5512e-01 -7.7911e-01 -4.1485e-01]
 [ 3.6843e+00 -2.3661e+00  1.7814e-01 -2.0558e-01 -2.2197e-01 -2.1458e+00]]
```

- A csv file store the data we collected, n_dimensional(column) means the sensor dimension.

Data combination

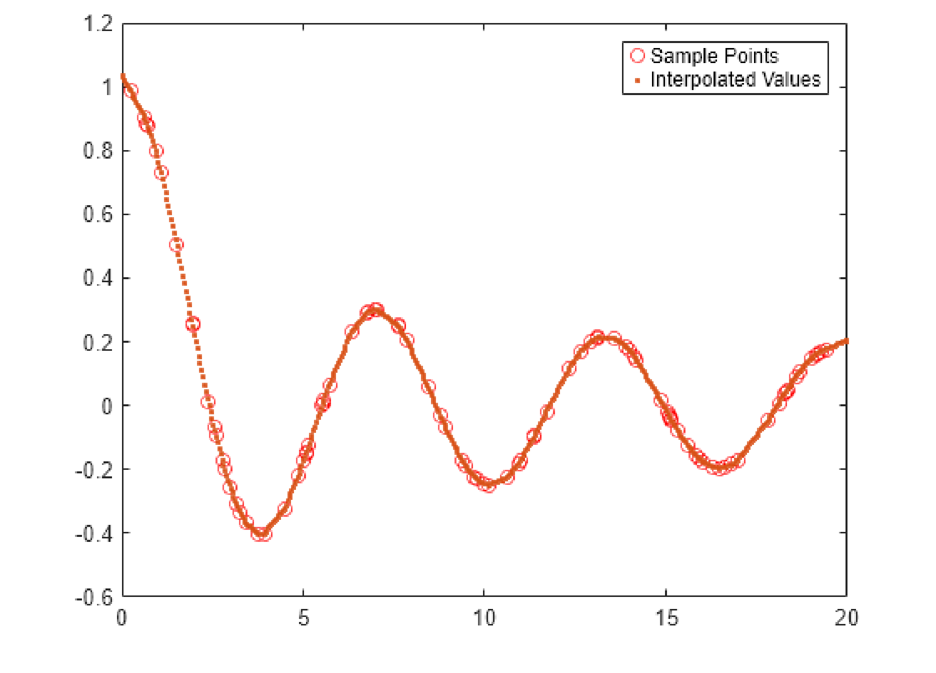


```
                      timestamp     imu_x     imu_y     imu_z
0   2025-02-17 22:37:36.248279 -1.103960  0.444364  0.827972
1   2025-02-17 22:37:36.268279  0.529332  1.318311 -0.892368
2   2025-02-17 22:37:36.288279 -1.088093 -0.143852  1.628061
3   2025-02-17 22:37:36.308279 -2.059210 -0.608279  1.245692
4   2025-02-17 22:37:36.328279  1.204604  1.849850 -0.967165
..                         ...       ...       ...       ...
495 2025-02-17 22:37:46.148279  0.923155  0.478154  0.455544
496 2025-02-17 22:37:46.168279  0.696304 -0.692716  0.433924
497 2025-02-17 22:37:46.188279 -0.842162  0.033667  0.502960
498 2025-02-17 22:37:46.208279  0.168948  0.314619 -2.590918
499 2025-02-17 22:37:46.228279  0.150978  0.272842 -0.152935
```

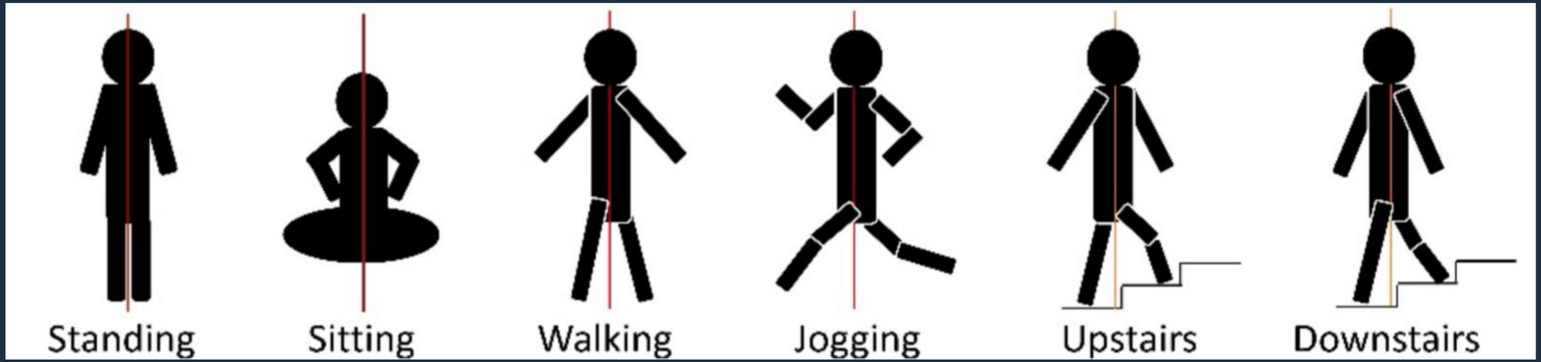


```
                      timestamp   voltage
0   2025-02-17 22:37:36.248279  0.366984
1   2025-02-17 22:37:36.298279 -1.316928
2   2025-02-17 22:37:36.348279 -0.956157
3   2025-02-17 22:37:36.398279 -1.270197
4   2025-02-17 22:37:36.448279 -0.461933
..                         ...       ...
195 2025-02-17 22:37:45.998279 -1.419428
196 2025-02-17 22:37:46.048279  1.787693
197 2025-02-17 22:37:46.098279  0.100555
198 2025-02-17 22:37:46.148279  1.961926
199 2025-02-17 22:37:46.198279 -0.381695
```

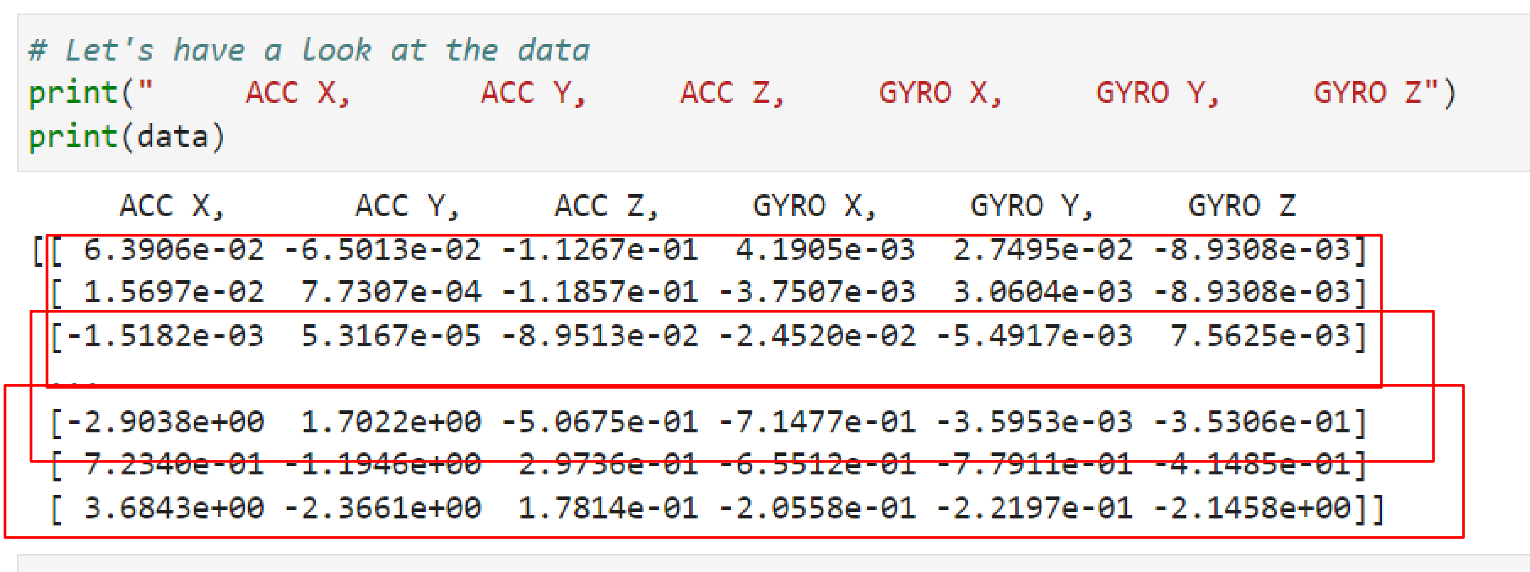
Remember add timestamps, keep an almost same sampling frequency, detect Nan values, interpolation for missing data



Sliding window

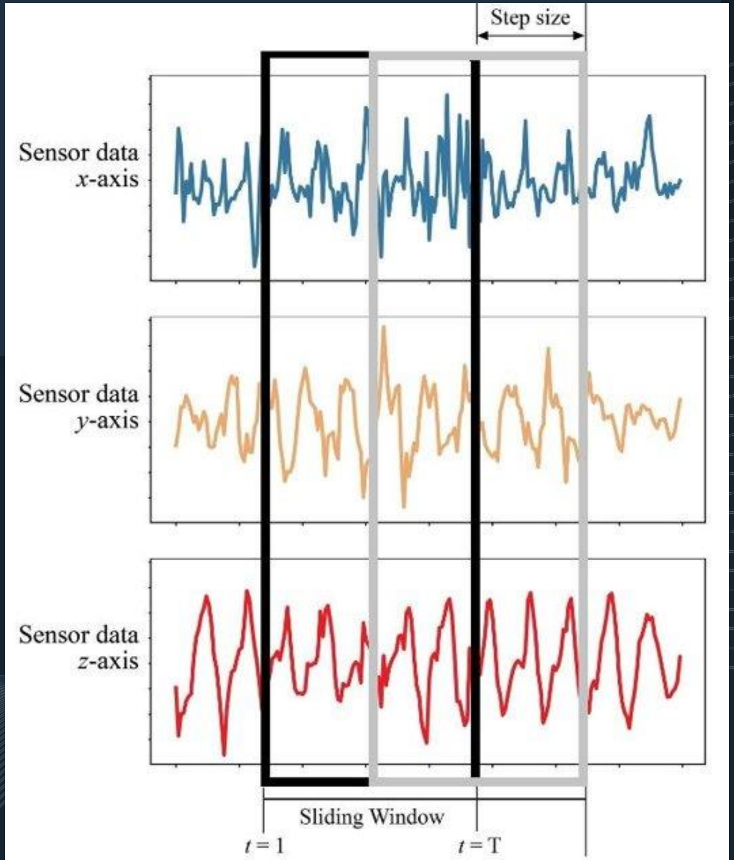


- Using a sliding window go through the data, for example: window length = 50, step size = 25



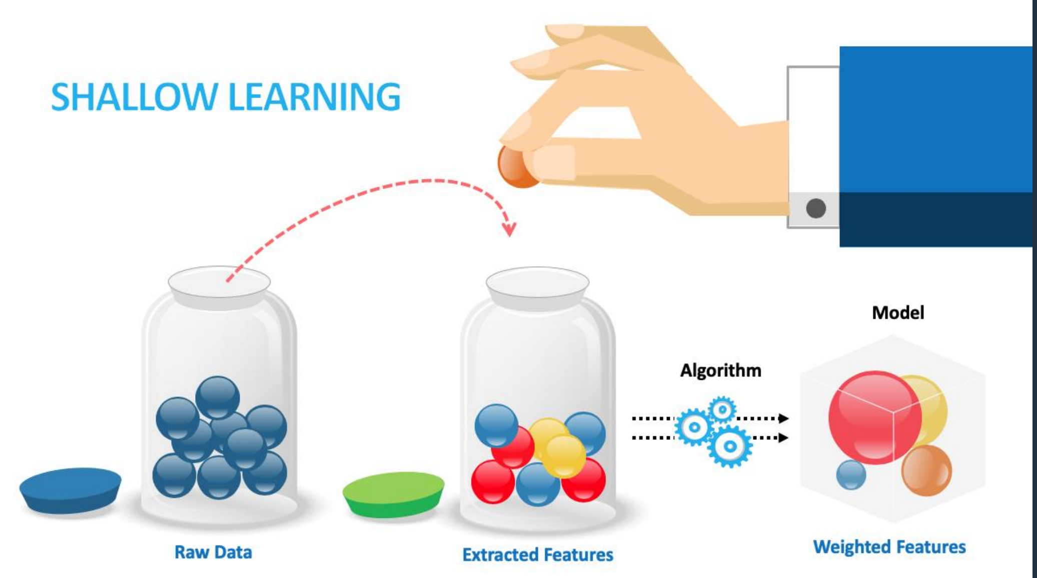
```
# Let's have a look at the data
print("     ACC X,         ACC Y,      ACC Z,      GYRO X,      GYRO Y,      GYRO Z")
print(data)
```

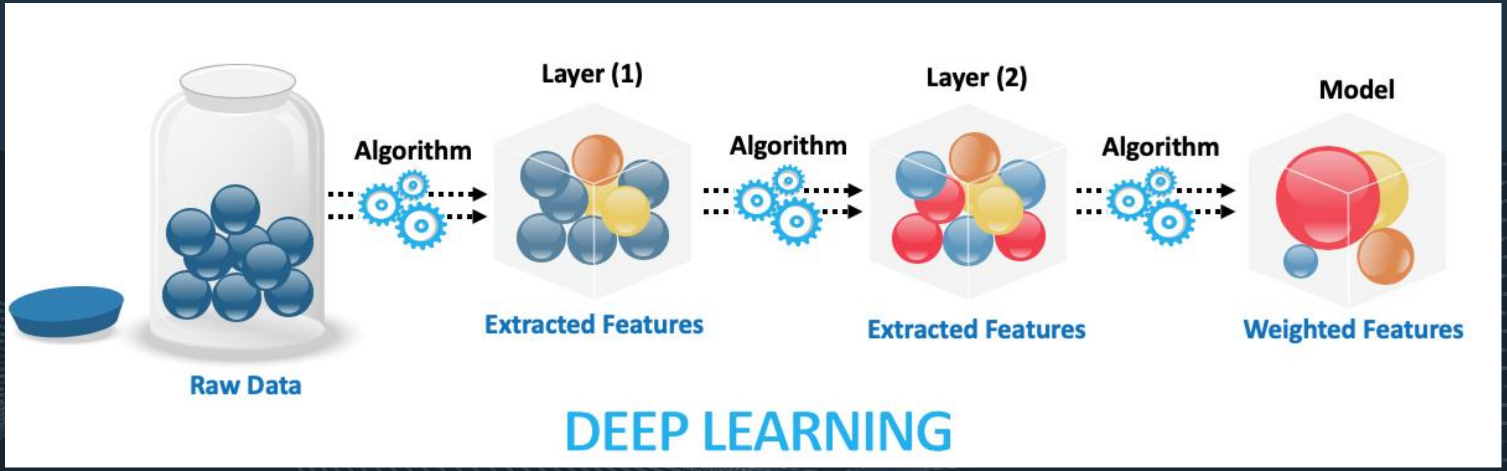
```
     ACC X,         ACC Y,      ACC Z,      GYRO X,      GYRO Y,      GYRO Z
[[ 6.3906e-02 -6.5013e-02 -1.1267e-01  4.1905e-03  2.7495e-02 -8.9308e-03]
 [ 1.5697e-02  7.7307e-04 -1.1857e-01 -3.7507e-03  3.0604e-03 -8.9308e-03]
 [-1.5182e-03  5.3167e-05 -8.9513e-02 -2.4520e-02 -5.4917e-03  7.5625e-03]
 ...
 [-2.9038e+00  1.7022e+00 -5.0675e-01 -7.1477e-01 -3.5953e-03 -3.5306e-01]
 [ 7.2340e-01 -1.1946e+00  2.9736e-01 -6.5512e-01 -7.7911e-01 -4.1485e-01]
 [ 3.6843e+00 -2.3661e+00  1.7814e-01 -2.0558e-01 -2.2197e-01 -2.1458e+00]]
```



Manually feature extraction is important

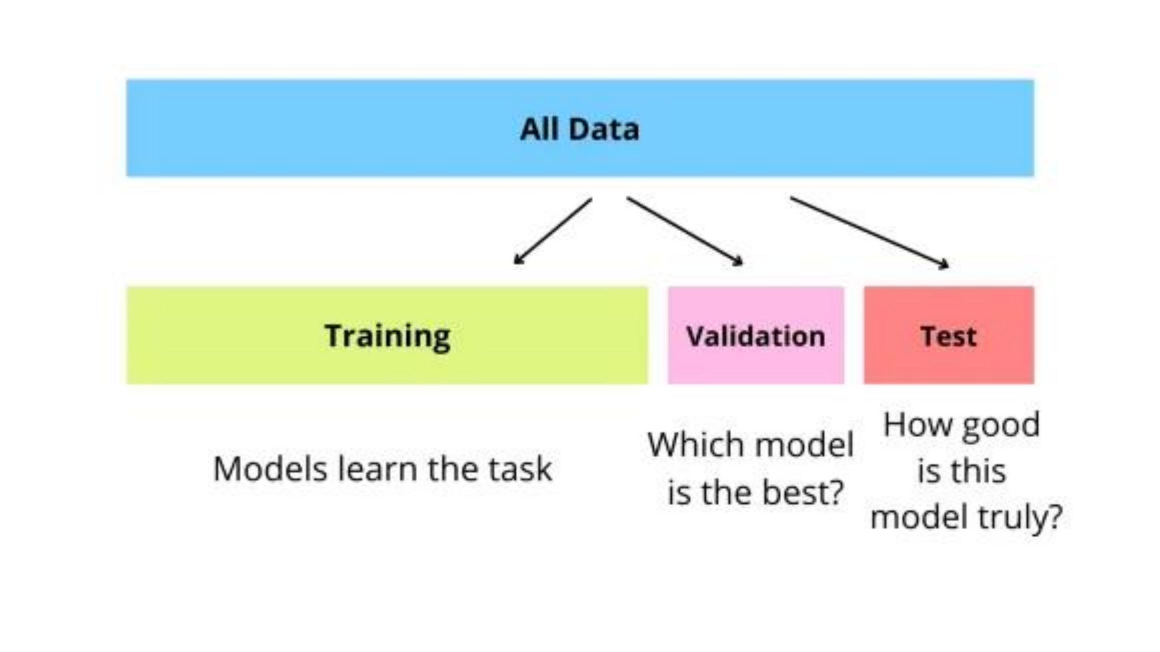
- Statistical features (mean, variance, skewness, kurtosis, entropy)
- Autoregressive coefficients, FFT-based features
- Peak detection, frequency-domain transforms





- Source: https://www.linkedin.com/pulse/introduction-shallow-machine-learning-ayman-mahmoud/

Data split



All Data

Training

Validation

Test

Models learn the task

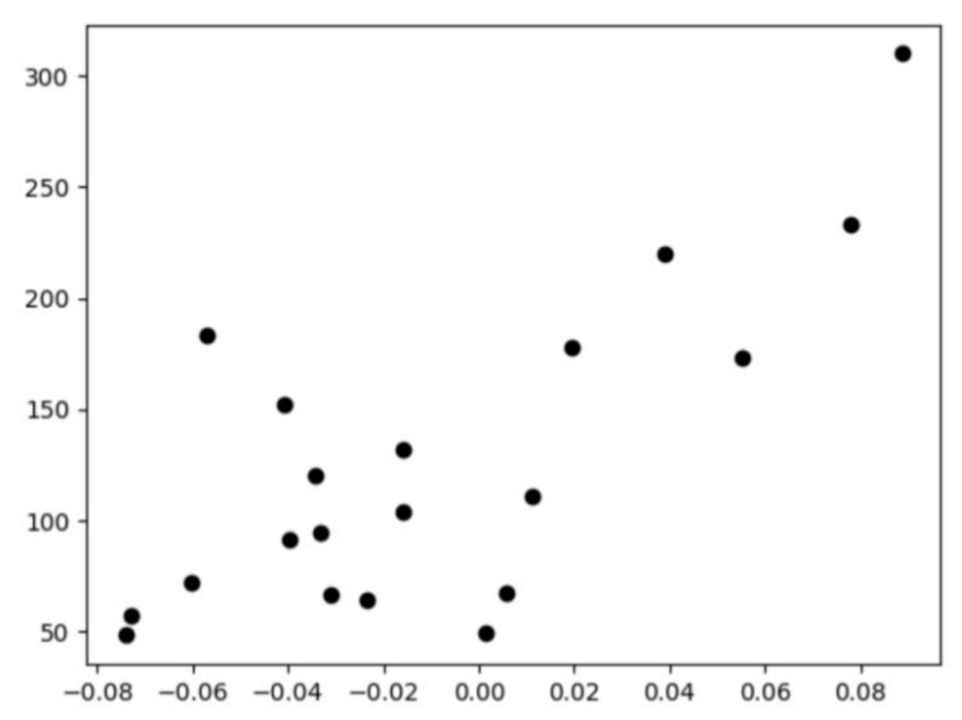
Which model is the best?

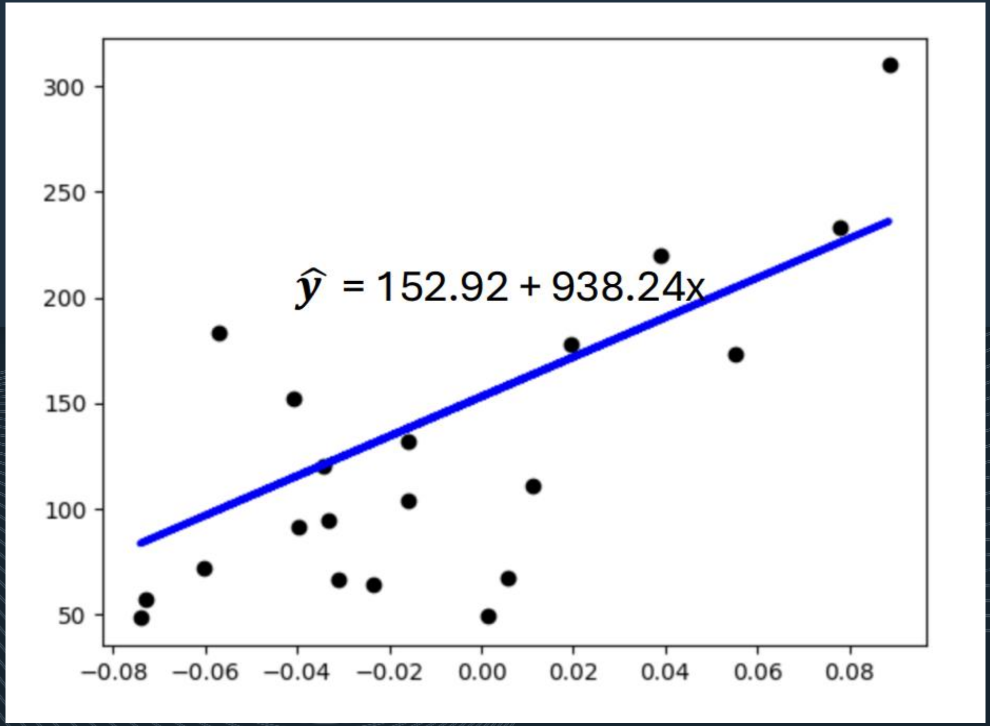
How good is this model truly?

Source: https://medium.com/@rahulchavan4894/understanding-train-test-and-validation-dataset-split-in-simple-quick-terms-5a8630fe58c8

Linear Regression

Linear regression is a method for finding the straight line or hyperplane that best fits a set of points.



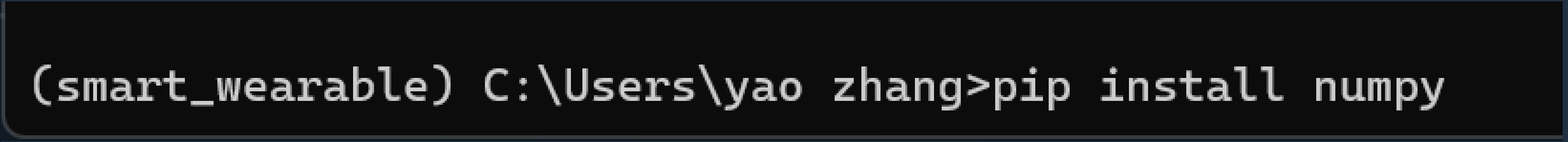


Package installation: numpy,sciket-learn,pandas

For Aalto Jupyter: no need for installation, already installed by Aalto

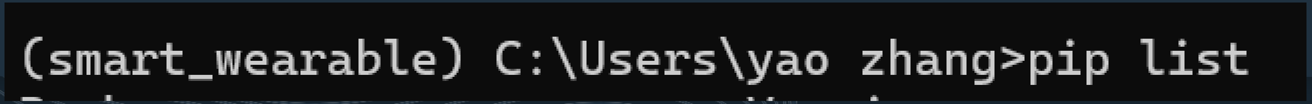
For your own local environment: install by pip

Example command:



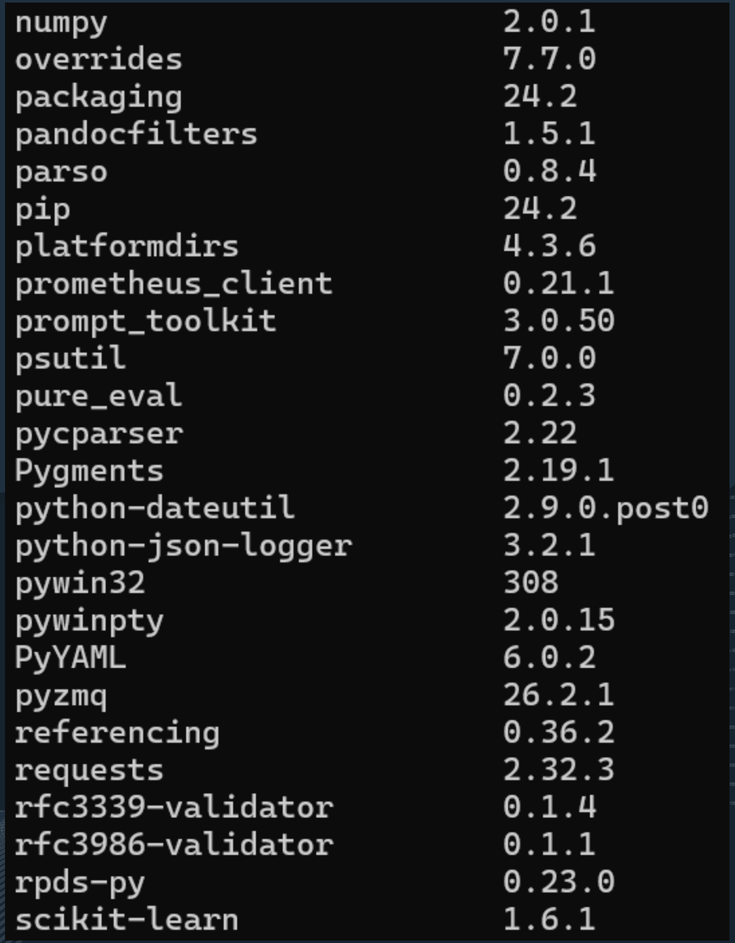
```
(smart_wearable) C:\Users\yao zhang>pip install numpy
```

How to check?



```
(smart_wearable) C:\Users\yao zhang>pip list
```

Link for instruction:
https://medium.com/@6unpnp/install-scikit-learn-d58f1415962d



```
numpy                        2.0.1
overrides                    7.7.0
packaging                    24.2
pandocfilters                1.5.1
parso                        0.8.4
pip                          24.2
platformdirs                 4.3.6
prometheus_client            0.21.1
prompt_toolkit               3.0.50
psutil                       7.0.0
pure_eval                    0.2.3
pycparser                    2.22
Pygments                     2.19.1
python-dateutil              2.9.0.post0
python-json-logger           3.2.1
pywin32                      308
pywinpty                     2.0.15
PyYAML                       6.0.2
pyzmq                        26.2.1
referencing                  0.36.2
requests                     2.32.3
rfc3339-validator            0.1.4
rfc3986-validator            0.1.1
rpds-py                      0.23.0
scikit-learn                 1.6.1
```