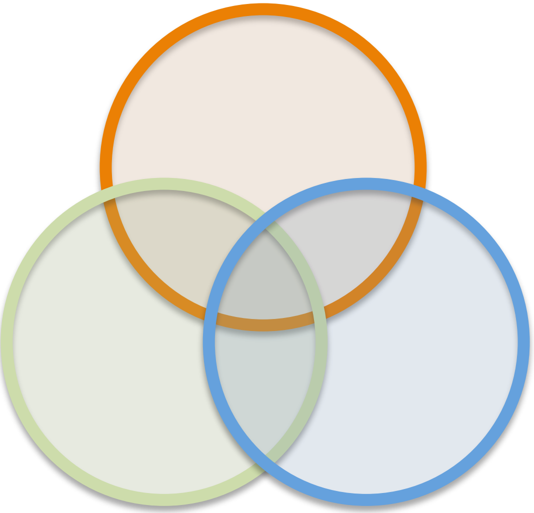
Machine Learning for the Quantified Self



Lecture 2

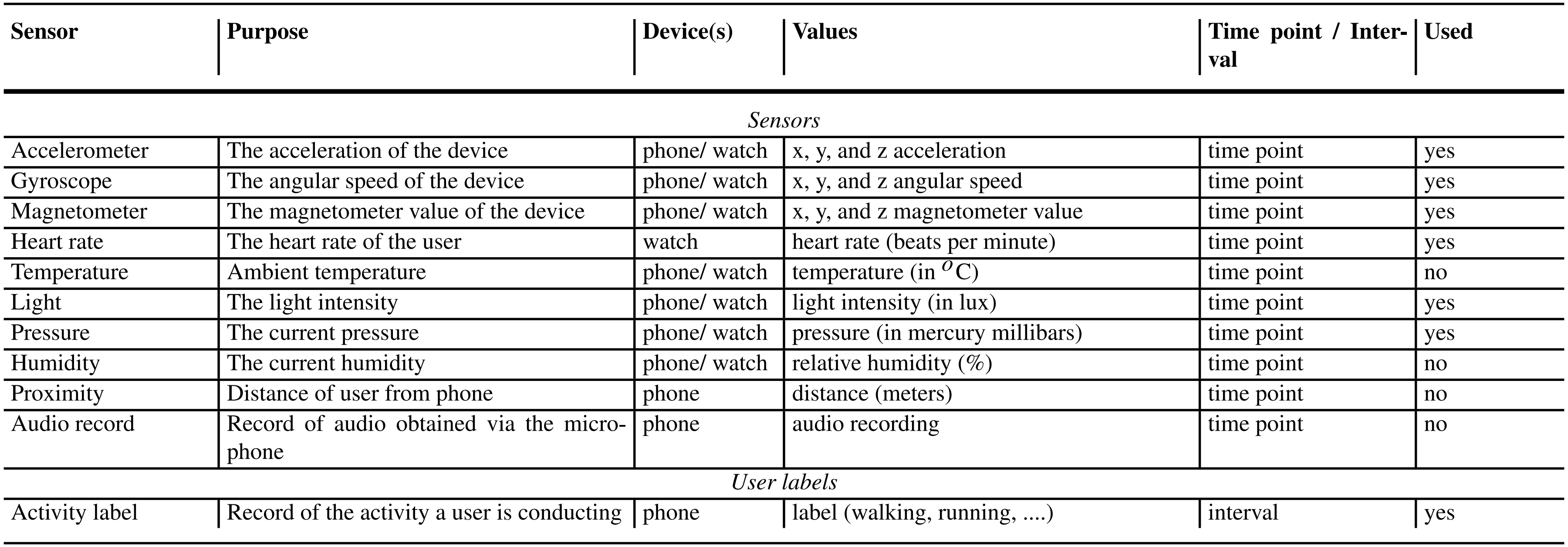
Basic of Sensory Data

Dataset (1)

- During the course we will use a running example provided by CrowdSignals.io
- People share their mobile sensors data (smart phone and smart watch) and get paid for annotating their data with activities

CrowdSignals.io

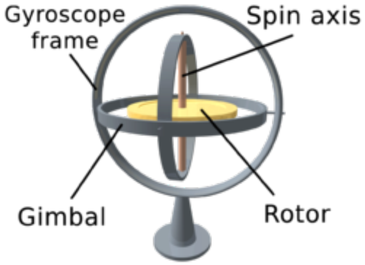
Dataset (2)



Sensor	**Purpose**	**Device(s)**	**Values**	**Time point / Interval**	**Used**
			Sensors		
Accelerometer	The acceleration of the device	phone/ watch	x, y, and z acceleration	time point	yes
Gyroscope	The angular speed of the device	phone/ watch	x, y, and z angular speed	time point	yes
Magnetometer	The magnetometer value of the device	phone/ watch	x, y, and z magnetometer value	time point	yes
Heart rate	The heart rate of the user	watch	heart rate (beats per minute)	time point	yes
Temperature	Ambient temperature	phone/ watch	temperature (in ^{o}C)	time point	no
Light	The light intensity	phone/ watch	light intensity (in lux)	time point	yes
Pressure	The current pressure	phone/ watch	pressure (in mercury millibars)	time point	yes
Humidity	The current humidity	phone/ watch	relative humidity (%)	time point	no
Proximity	Distance of user from phone	phone	distance (meters)	time point	no
Audio record	Record of audio obtained via the microphone	phone	audio recording	time point	no
			User labels		
Activity label	Record of the activity a user is conducting	phone	label (walking, running, ...)	interval	yes

Mobile phone measurements (examples)

- Accelerometer
  - Measures the changes is forces upon the phone in the x-y-z plane
- Gyroscope
  - Orientation of the phone compared to the earth's surface



- Magnetometer
  - Measures x-y-z orientation compared to the earth's magnetic field

The raw data

- What does the raw CrowdSignals data look like?
  - Separate tables per measurement
  - Specific time point measurements:

sensor_type	device_type	timestamps	rate
heartrate	smartwatch	1454956086325639687	175.000
heartrate	smartwatch	1454956086684549167	176.000
heartrate	smartwatch	1454956087523516770	175.000

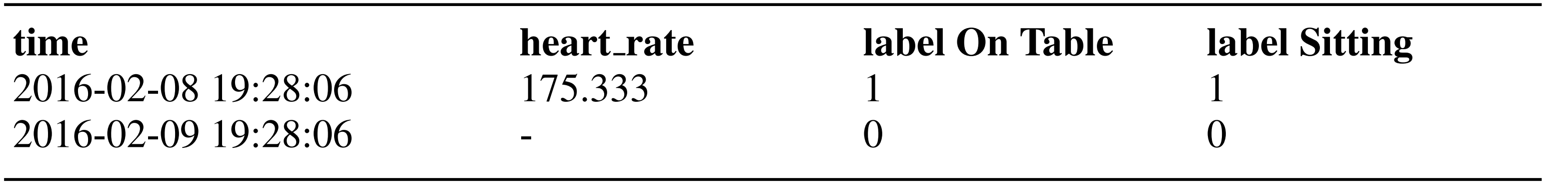
  - Interval measurements

sensor_type	device_type	label	label_start	label_end
interval_label	smartphone	On Table	1454956132985999872	1454956366574000128
interval_label	smartphone	On Table	1454956393088000000	1454956578385999872
interval_label	smartphone	On Table	1454956608515000064	1454956813323000064
interval_label	smartphone	Sitting	1454956894057999872	1454957092968000000

Transforming the raw data (1)

- Need to combine these table, but how?
- Select a *step size* Δt you want to consider in the data
  - this will represent one discrete time step
  - start at the earliest time point in the data
  - find all measurements for each single attribute associated with each interval $[t, t+\Delta t)$
  - we consider categorical features (e.g. label) as a number of binary features
  - combine their values (e.g. average for heart rate or accelerometer or sum for the label)

Transforming the raw data (2)

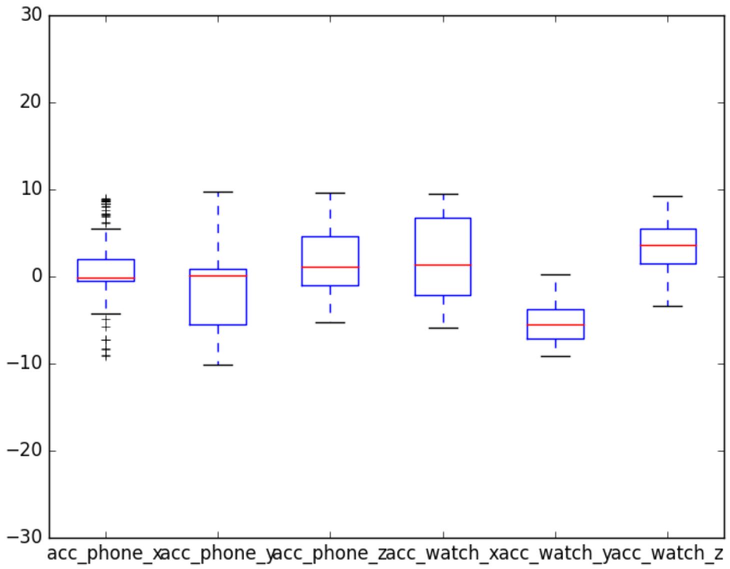


time	heart_rate	label On Table	label Sitting
2016-02-08 19:28:06	175.333	1	1
2016-02-09 19:28:06	-	0	0

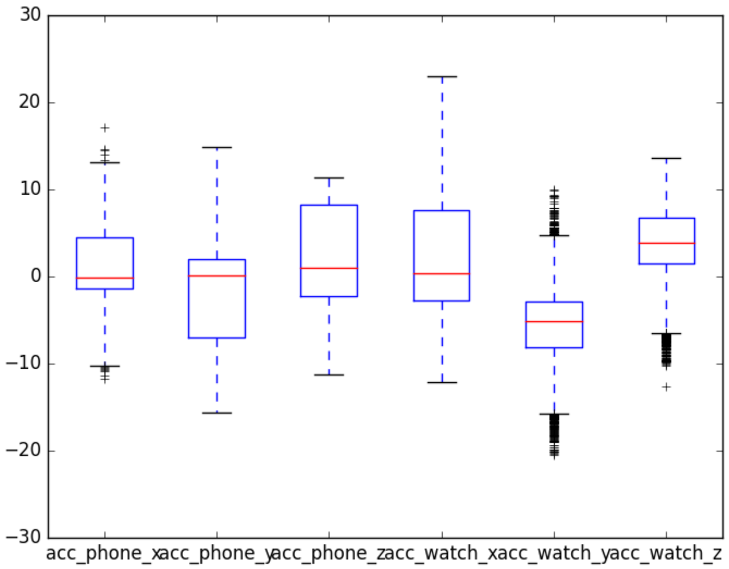
Exploring the data (1)

- Let us consider a dataset from CrowdSignals which covers around 2 hours of data
- Imagine we take a step size of Δt = 1 minute and Δt = 250 milliseconds
- What difference would you expect in the spread of the data?
- What are the pros and cons of a higher value for Δt (less fine grained)?

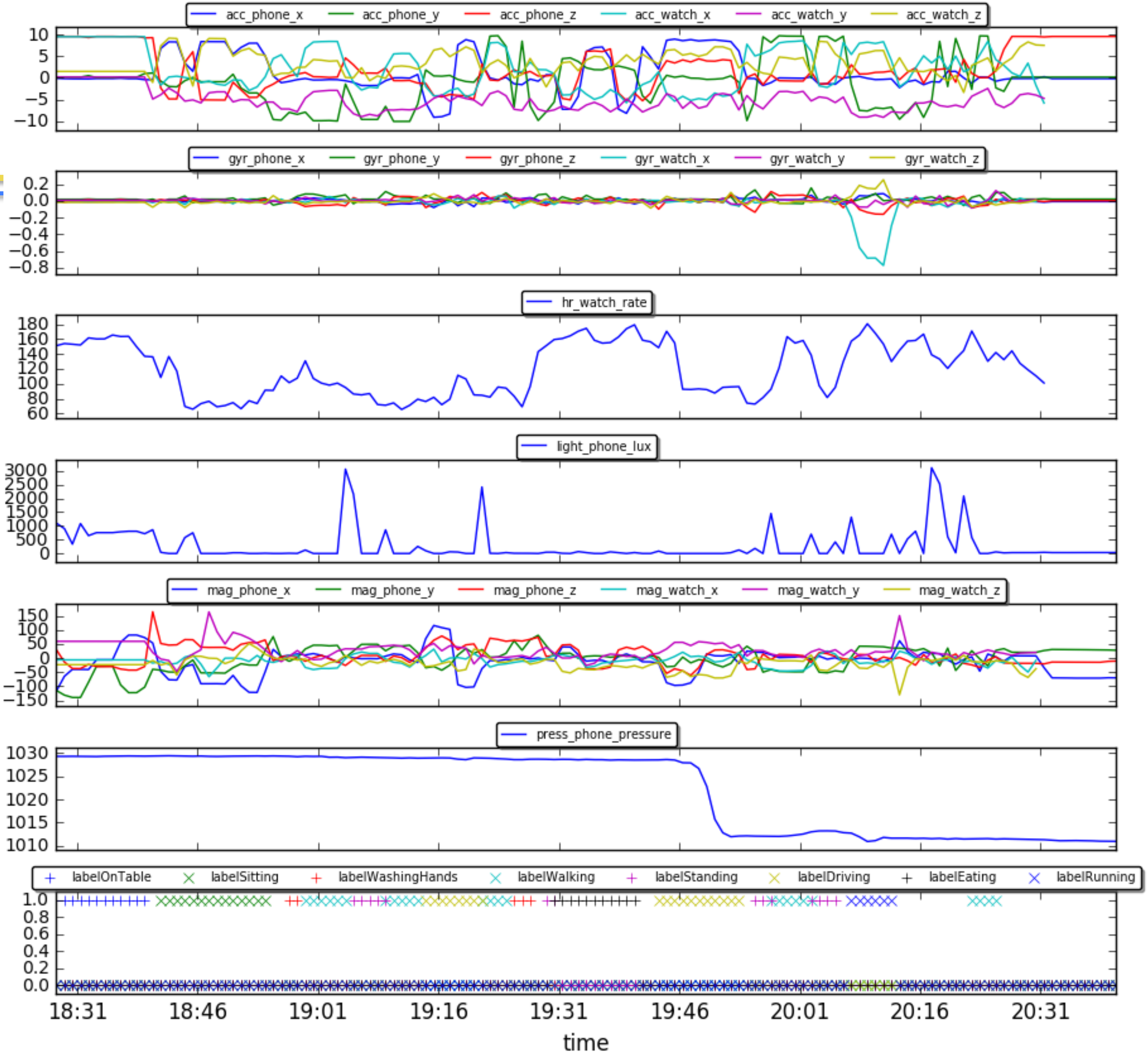
Exploring the data (2)

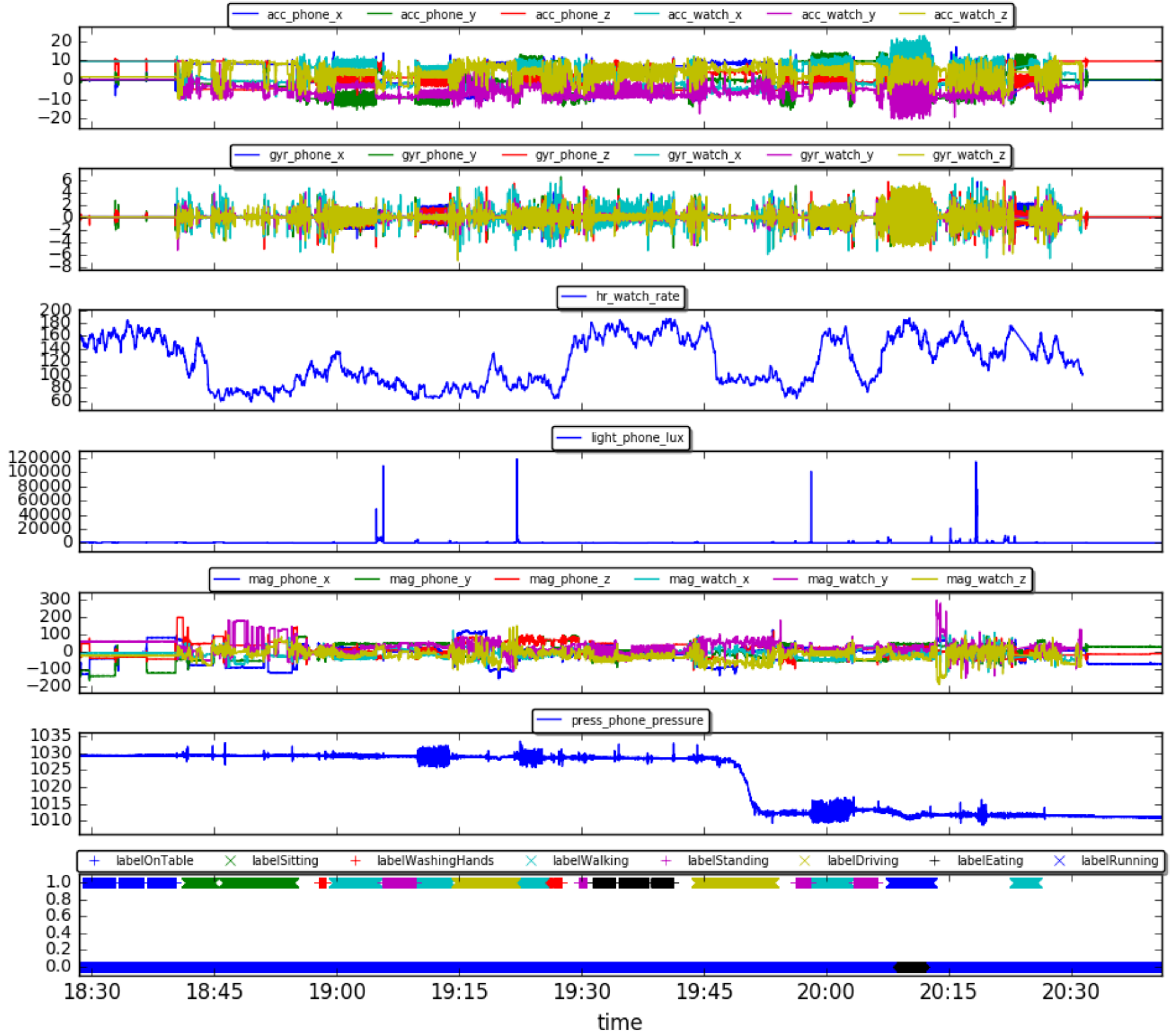


(a) $\Delta t = 60$ seconds



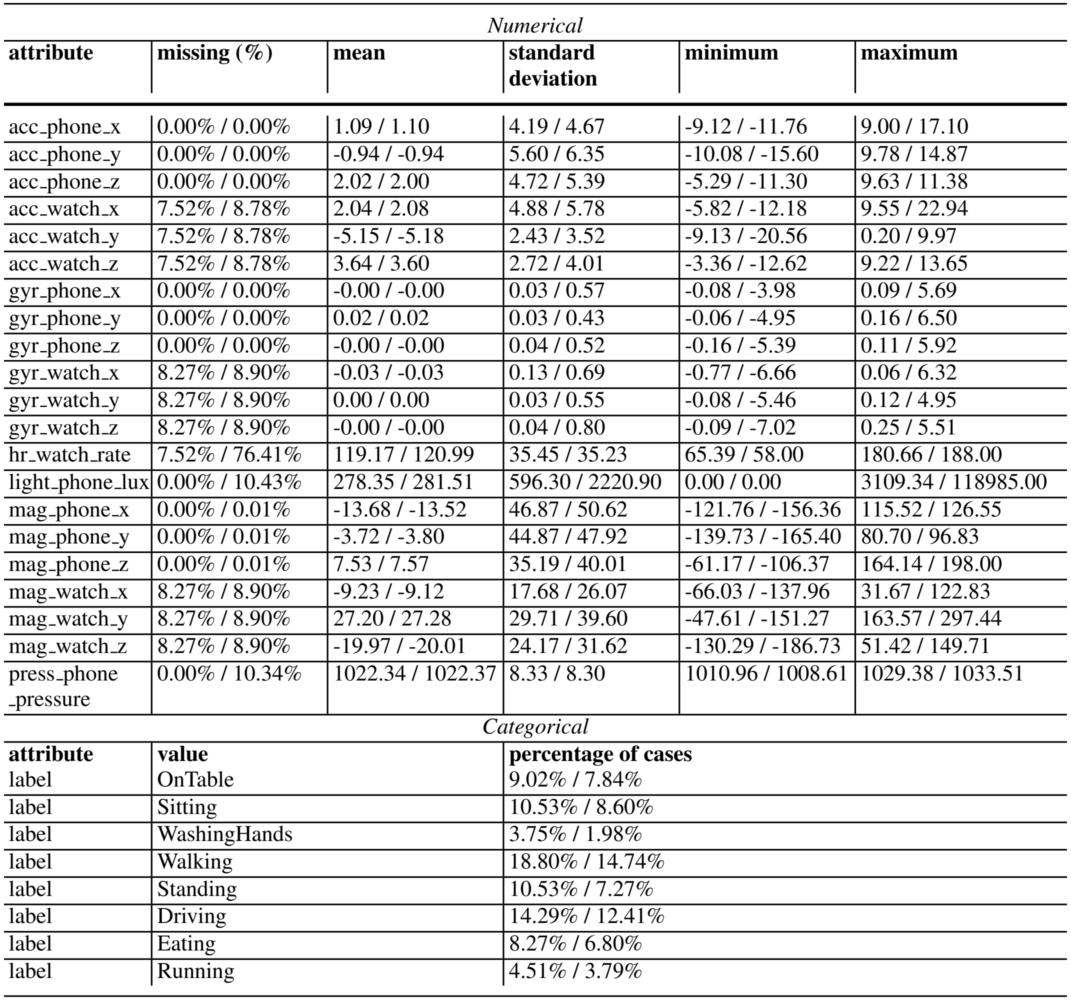
(b) $\Delta t = 0.25$ seconds


acc_phone_x
acc_phone_y
acc_phone_z
acc_watch_x
acc_watch_y
acc_watch_z
10
5
0
−5
−10
gyr_phone_x
gyr_phone_y
gyr_phone_z
gyr_watch_x
gyr_watch_y
gyr_watch_z
0.2
0.0
−0.2
−0.4
−0.6
−0.8
hr_watch_rate
180
160
140
120
100
80
60
light_phone_lux
3000
2500
2000
1500
1000
500
0
mag_phone_x
mag_phone_y
mag_phone_z
mag_watch_x
mag_watch_y
mag_watch_z
150
100
50
0
−50
−100
−150
press_phone_pressure
1030
1025
1020
1015
1010
labelOnTable
labelSitting
labelWashingHands
labelWalking
labelStanding
labelDriving
labelEating
labelRunning
1.0
0.8
0.6
0.4
0.2
0.0
18:31
18:46
19:01
19:16
19:31
19:46
20:01
20:16
20:31
time



acc_phone_x
acc_phone_y
acc_phone_z
acc_watch_x
acc_watch_y
acc_watch_z
20
10
0
−10
−20
gyr_phone_x
gyr_phone_y
gyr_phone_z
gyr_watch_x
gyr_watch_y
gyr_watch_z
6
4
2
0
−2
−4
−6
−8
hr_watch_rate
200
180
160
140
120
100
80
60
light_phone_lux
120000
100000
80000
60000
40000
20000
0
mag_phone_x
mag_phone_y
mag_phone_z
mag_watch_x
mag_watch_y
mag_watch_z
300
200
100
0
−100
−200
press_phone_pressure
1035
1030
1025
1020
1015
1010
labelOnTable
labelSitting
labelWashingHands
labelWalking
labelStanding
labelDriving
labelEating
labelRunning
1.0
0.8
0.6
0.4
0.2
0.0
18:30
18:45
19:00
19:15
19:30
19:45
20:00
20:15
20:30
time

Exploring the data (5)



Numerical

attribute	missing (%)	mean	standard deviation	minimum	maximum
acc_phone_x	0.00% / 0.00%	1.09 / 1.10	4.19 / 4.67	-9.12 / -11.76	9.00 / 17.10
acc_phone_y	0.00% / 0.00%	-0.94 / -0.94	5.60 / 6.35	-10.08 / -15.60	9.78 / 14.87
acc_phone_z	0.00% / 0.00%	2.02 / 2.00	4.72 / 5.39	-5.29 / -11.30	9.63 / 11.38
acc_watch_x	7.52% / 8.78%	2.04 / 2.08	4.88 / 5.78	-5.82 / -12.18	9.55 / 22.94
acc_watch_y	7.52% / 8.78%	-5.15 / -5.18	2.43 / 3.52	-9.13 / -20.56	0.20 / 9.97
acc_watch_z	7.52% / 8.78%	3.64 / 3.60	2.72 / 4.01	-3.36 / -12.62	9.22 / 13.65
gyr_phone_x	0.00% / 0.00%	-0.00 / -0.00	0.03 / 0.57	-0.08 / -3.98	0.09 / 5.69
gyr_phone_y	0.00% / 0.00%	0.02 / 0.02	0.03 / 0.43	-0.06 / -4.95	0.16 / 6.50
gyr_phone_z	0.00% / 0.00%	-0.00 / -0.00	0.04 / 0.52	-0.16 / -5.39	0.11 / 5.92
gyr_watch_x	8.27% / 8.90%	-0.03 / -0.03	0.13 / 0.69	-0.77 / -6.66	0.06 / 6.32
gyr_watch_y	8.27% / 8.90%	0.00 / 0.00	0.03 / 0.55	-0.08 / -5.46	0.12 / 4.95
gyr_watch_z	8.27% / 8.90%	-0.00 / -0.00	0.04 / 0.80	-0.09 / -7.02	0.25 / 5.51
hr_watch_rate	7.52% / 76.41%	119.17 / 120.99	35.45 / 35.23	65.39 / 58.00	180.66 / 188.00
light_phone_lux	0.00% / 10.43%	278.35 / 281.51	596.30 / 2220.90	0.00 / 0.00	3109.34 / 118985.00
mag_phone_x	0.00% / 0.01%	-13.68 / -13.52	46.87 / 50.62	-121.76 / -156.36	115.52 / 126.55
mag_phone_y	0.00% / 0.01%	-3.72 / -3.80	44.87 / 47.92	-139.73 / -165.40	80.70 / 96.83
mag_phone_z	0.00% / 0.01%	7.53 / 7.57	35.19 / 40.01	-61.17 / -106.37	164.14 / 198.00
mag_watch_x	8.27% / 8.90%	-9.23 / -9.12	17.68 / 26.07	-66.03 / -137.96	31.67 / 122.83
mag_watch_y	8.27% / 8.90%	27.20 / 27.28	29.71 / 39.60	-47.61 / -151.27	163.57 / 297.44
mag_watch_z	8.27% / 8.90%	-19.97 / -20.01	24.17 / 31.62	-130.29 / -186.73	51.42 / 149.71
press_phone_pressure	0.00% / 10.34%	1022.34 / 1022.37	8.33 / 8.30	1010.96 / 1008.61	1029.38 / 1033.51

Categorical

attribute	value	percentage of cases
label	OnTable	9.02% / 7.84%
label	Sitting	10.53% / 8.60%
label	WashingHands	3.75% / 1.98%
label	Walking	18.80% / 14.74%
label	Standing	10.53% / 7.27%
label	Driving	14.29% / 12.41%
label	Eating	8.27% / 6.80%
label	Running	4.51% / 3.79%

Machine Learning Tasks

- What kind of tasks could we identify in this dataset?
  1. a *classification* problem, namely predicting the label (i.e. activity) based on the sensors
  2. a *regression* problem, namely predicting the heart rate based on the other sensory values and the activity