Professor Pavan Ramdya
EPFL

Background

Until this part of the semester, we've focused on how to think about other people's science through presentations and critique of primary literature. We've also practiced programming the behaviours of artificial agents in Webots using Python.

For this mini-project, we would like you to use these technical and conceptual skill by giving you hands-on experience using Python to analyse new data with the intention of discovering something new about how animals control their movements. Specifically, you will focus on understanding how signals between the brain and motor control centres control limb movements. We will be looking at two important classes of neurons:
Descending neurons (DNs) are defined by having their cell bodies and dendrites in the Brain and their axons primarily projecting to the insect motor control centre, the ventral nerve cord (VNC, roughly equivalent to the insect 'spinal cord').
Ascending neurons (ANs) are the opposite of DNs. They have their cell bodies and dendrites in the VNC and their axons project to the brain.
Previous theoretical studies in models and robots (Ijspeert et al. *Science* 2007) suggested that simply driving a small subset of DNs could flexibly elicit multiple behaviours (e.g., swimming and walking) by recruiting downstream motor circuits in new ways. Such a possibility could not be tested in real animals until recently when the building of transgenic strains of *Drosophila* (one strain in Bidaye et al., *Science* 2014 and >100 strains in Namiki et al. *Elife* 2018) allowed scientists to target these neurons for studies of their anatomy and function and found that activation of individual DNs can trigger distinct behaviours, such as backward walking.
While exciting, strongly activating single neurons is a very artificial setting and these initial studies provided very limited details about how multiple DNs together controlled leg kinematics and behavioural dynamics and what the role of ANs is.
Thanks to recent developments, it is now possible to record from many neurons connecting the brain and VNC at once using functional two-photon imaging (Chen et al. *Nature Communications* 2018).
Using this very recent technique, we recorded the activity of 123 neurons that connect the brain and the VNC at once, while a fly is freely moving on an air-suspended spherical treadmill. You will be working with a novel dataset that no-one has ever analysed before.
With this data in hand, you have the opportunity to answer some fundamental questions regarding how groups/populations of neurons control behaviour in animals, and also to try out some new data analysis approaches to study the relationship between neuronal signals and behaviours in an unsupervised manner – taking "pesky" humans out of the loop in neuroscience.

The dataset

The dataset consists of recordings from one fly (R57C10-Gal4 > UAS-GCaMP6f, UAS-tdtom) across 12 trials of around 250s each.
For each trial, we provide the following two pandas dataframes:
Neural data: COBAR_neural.pkl

- Fluorescence traces of 123 neurons in the cervical connective expressing the genetically encoded Calcium indicator GCaMP6f (Chen et al. Nature 2013), imaged with a 2-photon microscope at a sampling rate of around 16 Hz. The exact time stamp of each sample is in the signal "t"

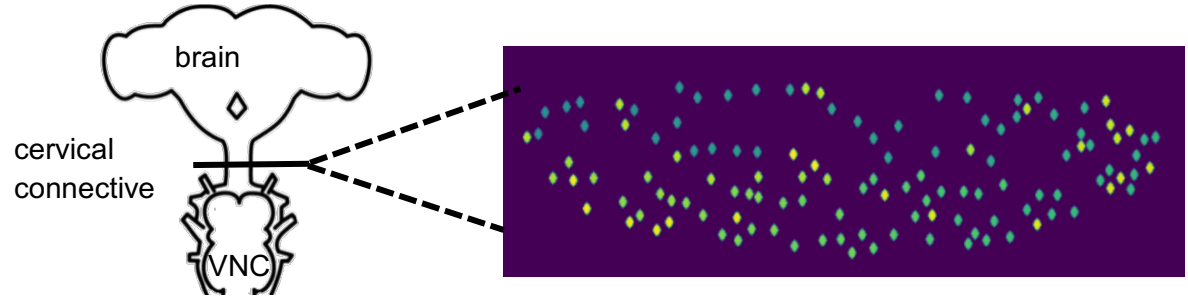


Figure 1: Left: The fly's central nervous system consisting of the brain and the ventral nerve cord (VNC). Right: Location of 123 neurons in a cross-section of the cervical connective.

Behavioural data: COBAR_behav.pkl

Behavioural variables are recorded/computed at a higher frame rate (100 Hz) to capture very fast movements of the fly. The behaviour of the fly is recorded with 7 cameras surrounding the fly. Using the marker-less pose-estimation algorithm DeepFly3D (Günel et al. Elife 2019), we track the joint positions over time. After alignment of the pose to a template fly (See Lobato et al. BioRxiv 2021), the joint angles are computed. Based on the joint angles, we trained a simple classifier, to predict the behaviour of the fly for every single frame.

- Joint positions of 5 joints per leg, each in xyz, including Coxa, Femur, Tibia, Tarsus, Claw → in total 5*6*3 = 90 variables
- Joint angles. 7 per leg including Coxa_yaw, Coxa, Coxa_roll, Femur, Femur_roll, Tibia, Tarsus. → in total 7*6 = 42 variables
- Output of behavioural classifier. The classifier was trained on images of multiple flies including the following behavioural labels: abdominal grooming, antennal grooming, eye grooming, foreleg grooming, hindleg grooming, resting, walking. Resting and walking are the behaviours, which are observed the most. Sometimes, the fly pushes the ball with its abdomen.
    - Prediction → the behavioural prediction of the classifier
    - Entropy → A measure of uncertainty of the classifier. If high, then the classifier is not sure.
    - Probabillity_Walking/… → the probability of each individual behaviour
- "twop_frame": This signal allows you to match behaviour frames (0:25199 per trial), to neural data frames (0:4039 per trial). The signal has a negative value if no neural frame matches the behavioural frame.

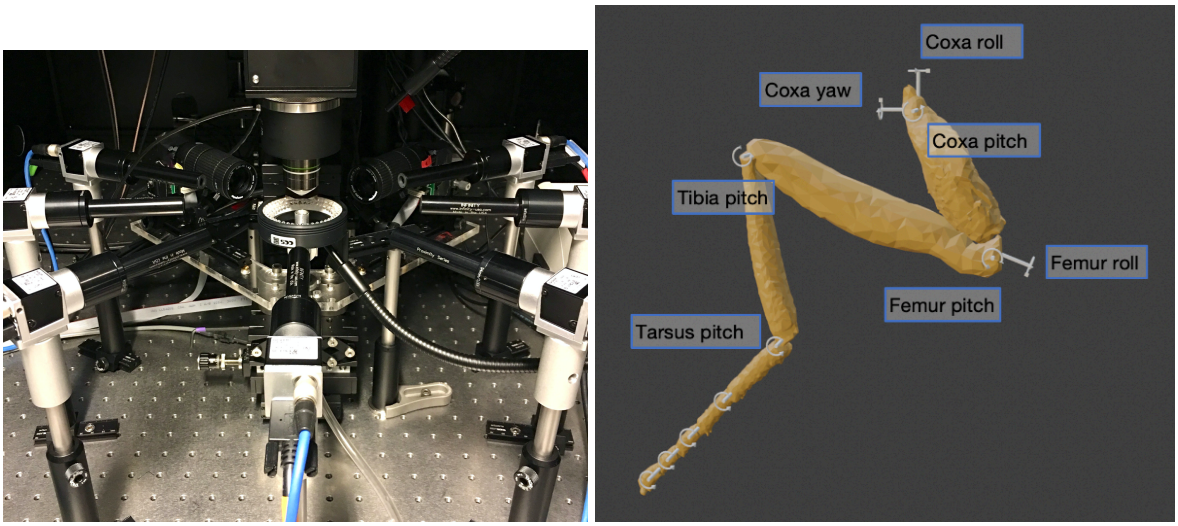


Figure 2: Left: recording set-up with 7 cameras surrounding the fly and the objective of the two-photon microscope Right: A leg of a drosophila with the joint names and joint angles

Mini-project structure

Week 8 - 20.04.: Familiarisation with data, pre-processing, plotting

1. Familiarise yourselves with the data by plotting time traces
2. Data pre-processing
3. Understanding the types of behaviours, that a fly performs, while manually correcting behavioural labels of an automated classifier

Week 9 – 27.04.: Analysing behavioural and neuronal data separately

1. Build a behavioural classifier
2. Perform dimensionality reduction & clustering on neuronal/behavioural data

Week 10+11 – 04.05.+11.05.: Correlating Neuronal activity with behavioural activity
Week 12 – 18.05.: Finalise data analysis & prepare mini-project report
Week 13 – 25.05.: Submit mini-project report & prepare presentations
Week 14 – 01.06.: Present mini-projects

We encourage you to be creative in analysing these data but at least address the following questions:
In general, we do not expect you to answer each individual question in the report/presentation. Make sure to think about each of the given questions and use them as a guide to design your data-processing / analysis pipeline. It should be clear from the report that you have spent some thoughts on each section. Rather than writing a question-answer style report, try to make the motivation of your analysis clear and put together a coherent story.

General questions for the report/presentation introduction

1. What is the experimental paradigm the data was generated with?
2. Why do you think the experiment were performed this way?
3. How might the experimental protocol, or data acquisition scheme have been better designed?
4. What are the advantages/disadvantages of studying spontaneous behaviour as opposed to a strict experimental design with many repetitions of the same experimental paradigm?

Week 8: Familiarisation with data, plotting

Part 1: Plotting the data

1. Neural data

Plot time traces of individual neurons across multiple trials.

a. Are they stable across all trials or do their activity patterns change?
b. Are all neurons active during all trials or are some of them only active for some trials? What could be a reason for that?
c. Are there some neurons which are much more active than others?
d. Can you already find neurons which have very similar signals?
e. What could be a good summary feature of a neuron? Its mean? Its maximum? Its standard deviation?
f. Do all neurons have the same baseline fluorescence? If no, do you think this affects the analysis? Does the baseline change over time?

2. Behavioural data

Plot time traces of individual angles/joints across multiple trials

a. Are they stable across all trials?
b. Do some of them look very similar?
c. Why do you think we want to look at joint angles in addition to the animal's pose?

Plot the 3d pose over time.

For your final report make sure that plots always have x- and y- axis labels and a title that is understandable. If you show multiple lines/data points in one plot, provide a legend and/or a colour bar. If you plot time series, always have time on the x-axis, not samples. Choose the scale of the x- and y-axis such that it is easy to see the message you want to convey with the plot.

Part 2: Data pre-processing

1. $\Delta F/F$

   In the first part, you have observed different baseline fluorescence values. In order to account for that, one frequently calculates the so called $\Delta F/F$ ("Delta F over F"): $\Delta F/F = (\mathrm{F} - \mathrm{F}_0)/F_0$, where F_0 is the baseline fluorescence.

   a. What would be a good way to compute the baseline fluorescence? Do you think taking the minimum provides a stable estimate? How could you stabilise this? (One idea might be to compute a moving average and then find the minimum or the 1% quantile).
   b. Should we use the same baseline for all trials or a different baseline for different trials?

2. Noise reduction

   Measurements are inherently noisy. Filtering is a common strategy to remove noise. A couple of adequate methods might be low-pass filtering (e.g. Gaussian or Butterworth filter) or outlier removal (e.g. median filtering), but feel free to explore different methods. Use plotting as a means to see whether the signals are contaminated by noise and how your denoising affects them

   a. Do the signals look noisy to you? What could be sources of noise?
   b. How would you remove the noise? What do you consider noise, what signal?
   c. Does denoising limit your further analysis?
   d. How do parameters of the filters affect the signal appearance?
   e. In case you low-pass/high-pass filter the data, which cut-off frequencies did you use?

   For your final report, provide examples that show your denoising approach and justify the choice of the parameters.

Part 2: Understanding different types of behaviour

To analyse behaviours, it is important to understand the types of behaviours the fly is capable of. We have previously trained an automatic classifier for the data, but it is not perfect. While familiarising yourself with the data, one of your tasks will be to annotate the behaviours of one trial (around 4 minutes) of behavioural recordings. **Each group will be assigned one trial** and can find the corresponding files **in the following google drive.** We ask you to manually correct the automatic labels to:

- familiarise yourself with the behaviours that a fly performs when on a spherical treadmill.
- provide you with an even more exciting dataset for next weeks, where you can try to build an automatic classifier to classify behaviour and can get a better understanding of

the relationship of individual neuron's activity with behaviour. Your colleagues will thank you for your labelling and you will be grateful for their work!

We have uploaded videos for each trial that include:
- the frame number,
- the time of the frame,
- a prediction from a not so optimal automatic classifier,
- the skeleton of the fly obtained by pose estimation superimposed on the video.

We suggest the following procedure:
1. Split up the trial amongst the two/three of you (i.e. person 1 does the first third and so on…)
2. Open the Google sheets file we provide. Go to the tab with the trial corresponding to your group. You should see something like this:

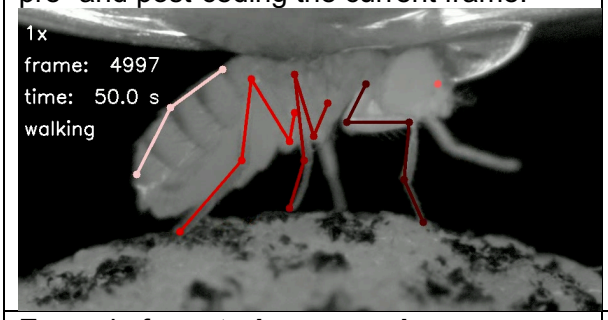
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Date	Genotype	Fly	Trial	Frame	Manual	Prediction	Entropy	Probability a	Probability a	Probability e	Probability fe	Probability h	Probability re	Probability walking	
2	210301	J1xCI9	1	0	0		resting	0.04588146	0.00018622	6.77E-05	6.31E-05	5.00E-05	0.00076335	0.99301667	0.00585303	
3	210301	J1xCI9	1	0	1		resting	0.03788282	0.00010744	4.22E-05	4.50E-05	4.33E-05	0.00064927	0.99440149	0.00471129	
4	210301	J1xCI9	1	0	2		resting	0.04215738	9.77E-05	5.89E-05	4.14E-05	4.23E-05	0.00069808	0.99361343	0.0054482	
5	210301	J1xCI9	1	0	3		resting	0.08693023	0.00015473	8.11E-05	7.60E-05	6.32E-05	0.0014947	0.98454088	0.0135894	
6	210301	J1xCI9	1	0	4		resting	0.14975949	0.00086742	0.00018146	0.00015372	8.15E-05	0.00512976	0.97133694	0.02224923	
7	210301	J1xCI9	1	0	5		resting	0.02738627	0.00025493	6.12E-05	6.59E-05	2.97E-05	0.00138638	0.99645751	0.00174443	
8	210301	J1xCI9	1	0	6		resting	0.15533683	0.00084799	0.00015796	0.0001748	7.27E-05	0.02399906	0.96975515	0.00499238	
9	210301	J1xCI9	1	0	7		resting	0.53454091	0.00104967	0.00052602	0.00058225	0.00016907	0.05300987	0.84974626	0.09491687	
10	210301	J1xCI9	1	0	8		resting	0.72716019	0.00116573	0.00093445	0.00087925	0.00030854	0.08645749	0.7605777	0.14967685	
11	210301	J1xCI9	1	0	9		resting	0.98798299	0.0028817	0.00152032	0.00113702	0.00049234	0.16928835	0.59253678	0.23214349	
12	210301	J1xCI9	1	0	10		resting	0.70457788	0.00250639	0.0015661	0.00122597	0.00041019	0.02885294	0.73395579	0.23148262	
13	210301	J1xCI9	1	0	11		resting	0.4687878	0.00367804	0.00237686	0.00193504	0.0007531	0.01342941	0.87138121	0.10644635	
14	210301	J1xCI9	1	0	12		resting	0.13407128	0.00120521	0.00078812	0.00034034	0.00018006	0.00455807	0.97630982	0.01661837	
15	210301	J1xCI9	1	0	13		resting	0.082866	0.00021749	0.00018557	0.00010726	4.85E-05	0.00149387	0.98566809	0.01227919	
16	210301	J1xCI9	1	0	14		resting	0.06465497	0.00016527	0.00021894	8.73E-05	4.02E-05	0.00097812	0.98939496	0.00911528	
17	210301	J1xCI9	1	0	15		resting	0.02908164	0.00010485	7.83E-05	4.29E-05	1.95E-05	0.00010803	0.99576909	0.00387738	
18	210301	J1xCI9	1	0	16		resting	0.0683422	0.00020942	0.00011544	6.32E-05	2.98E-05	0.00017492	0.98797119	0.01143601	

The **columns E, F, and G are the most important.**
E contains the frame number, which you can also see printed in the video
F is currently empty. This is where you should fill in the behaviours you find
G contains the automatically generated predictions (the following columns contain some more info about the classifier)

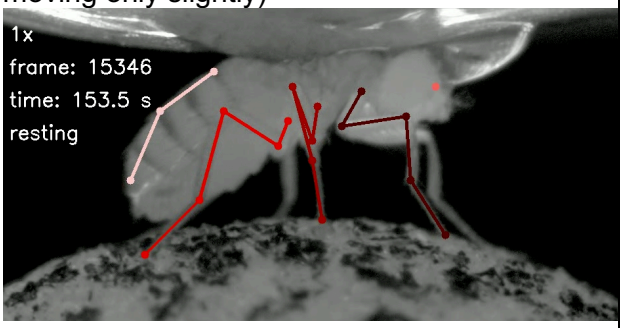
3. Open the video corresponding to your trial. We suggest to open it with a video player that allows you to move forward a single frame at a time. In Quicktime, this is possible by clicking the → and ← button on your keyboard. With VLC media player, look at this website for help. Usually the 'E' key is setup as a so called hotkey to forward by one frame.

Please annotate the following behaviours:
- **resting**: when the fly is standing still on the ball without walking/grooming/pushing/...
- **walking**: when the fly is walking (irrespective of the direction) or turning
- **abdominal_pushing**: when the fly is pushing its abdomen against the ball (this behaviour is not included in the automatically generated labels, but is nonetheless frequent)
- **anterior_grooming**: when the fly is grooming (i.e. rubbing its front legs against) its eyes, antennae or front legs (this is supposed to summarise the automatically generated labels antennal_grooming, eye_grooming, and foreleg_grooming, because they are often hard to discriminate for the untrained eye)
- **posterior_grooming**: when the fly is grooming (i.e. rubbing its hind legs against) the abdomen or hind legs (this is supposed to summarise the automatically generated labels abdominal_grooming, and hindleg_grooming, because they are often hard to discriminate for the untrained eye). This behaviour is very rare, so don't worry if you don't find any instances.

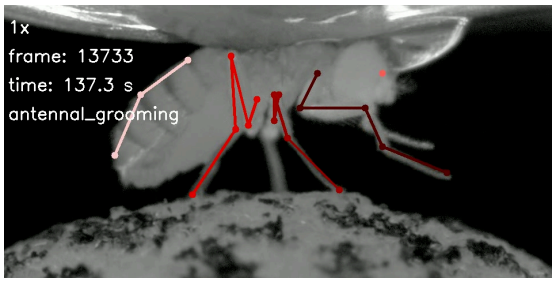
Example for **walking**:
the fly is frequently, and in a coordinated manner, regularily lifting multiple of its legs. Here: left front leg and right mid leg are lifted. Best judged by looking at the frames pre- and post-ceding the current frame.



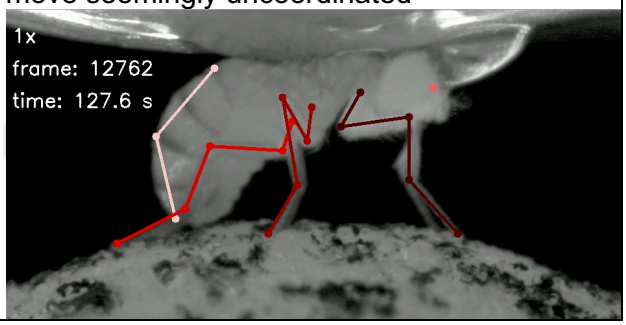
Example for **resting**:
all legs are on the ball; no leg is moving (ball moving only slightly)



Example for **anterior_grooming**:
both front legs are lifted up. They touch each other or the head.



Example for **abdominal_pushing**:
abdomen is bent downwards and pushed against the ball. Often, the other legs are move seemingly uncoordinated



Tricks to speed up annotation:

- use copy and paste a lot, e.g., copy from the “Prediction” column, if the classifier was correct or drag down from the previous frame if the behaviour persists.
- Scroll through the video to identify changes in behaviour. Use the same annotation between two consecutive changes

While annotating, think about the following questions:

- What is “a behaviour”? Is it hard to differentiate between different behaviours?
- Are the behaviours you observe very stereotyped or do they vary along a spectrum?
- Would you have chosen different behavioural categories?

Week 9: Analyse neural & behavioural data separately

Starting this week, we will analyse the data more deeply, for example by using methods from modern machine learning. There are many different types of analyses that are possible. The reviews we posted on Moodle provide some examples of methods that have been previously used (Cunningham and Yu 2014, Valletta et al. 2017). During today's class we introduced a few methods that you can try out to explore the neuronal and behavioural data you have in your hands. This week, we will focus on analysing behavioural data and neuronal data separately, but next week, we will try to link the two to gain some insides into how neurons can control behaviour.

At this point it might make sense that each member of your group tries out a different analysis method or even a different part of the steps below and you discuss regularly to understand how different methods can lead to different results or whether one or the other might be more appropriate.

Part 1: Neuronal data analysis

The following steps should help you to get started with the analysis of neuronal data. You do not have to perform all of them. You can also decide to follow up one an interesting part in more detail.

1. Use PCA with your denoised $\Delta F/F$, with each neuron as a feature and each time point as a sample. You can try using data from one trial only, or from all trials at once.
    a. Plot the first few principal components (PCs) over time. Which temporal features can you observe? Can you observe them in particular neurons? Do they depend on a particular set of neurons? (Hint: look at the so called "loadings" that map neurons to principal components)
    b. How many principal components do you need to explain 90 % of the variance in the neuronal data? Do you think this is a lot? Do you think this hints at an efficient/robust way to convey signals?
    c. Plot a scatter plot of the first two PCs (i.e., PC1 on the x-axis and PC2 on the y-axis) How much variance do these two explain? Can you see clusters?
        i. Hint: You are plotting a lot of points. Try to make the individual dots transparent or use a technique similar to kernel density estimation.
2. Use another dimensionality reduction technique, such as t-SNE or u-MAP to reduce the number of features to 2.
    a. Plot the scatter plot again. Does the plot look different than before? What has changed? Can you see clusters? What could these clusters correspond to?
    b. Are time points that are close to each other in similar locations in your 2D plot or in different locations?
3. Use PCA again, but this time use the time points as features and each neuron as a sample (i.e., use the transform of your data array for (1.))
    a. Would it make sense to reduce the number of features by averaging across some of them?
    b. Reduce the number of features to 2 using your preferred method.
    c. Plot a scatter plot. What do individual dots correspond to?
    d. Look at the signals over time for neurons that are close to each other in the scatter plot or that are far apart. What is their correlation coefficient? Does it make sense what you see? Why would multiple neurons have similar signals?

4. Try to cluster the 2D plot that you have found in (3.) using an automatic clustering algorithm (e.g. k-means, GMM, watershed). How many clusters do you find/do you choose?
    a. Compare different clustering algorithms. Do they give the same results?
    b. Would you choose the clusters similarly if you were to manually cluster the data?
    c. Do you observe the same clustering when you analyse single trials compared to when you analyse data from all trials?

For your final report we ask you to describe your analysis and think about the following questions:

1. Can you find neurons that have the same/similar signals?
2. What could be a reason for an organism to have multiple neurons conveying the same/similar signal?
3. Can you find "clusters" of neurons that have the same signals?
4. How many signal components do you need to describe most (e.g. 90%) of the variance in neuronal activity? Is this a lot?
5. Do these findings depend on the analysis method you choose?

Part 2: Behavioural analysis

The following steps should help you to get started with the analysis of behavioural data. Part 2.1 (feature extraction) forms the foundation for both parts 2.2 (clustering) and 2.3 (classification), but 2.2 and 2.3 can be performed independently. You do not have to perform all steps. You can also decide to follow up on an interesting part in more detail.

Part 2.1: Identifying features of behaviour

1. Think about what might be important features of behaviours.
    a. Do you think the 3D pose at one time point uniquely specifies a behaviour?
    b. Are joint angles a better measure?
    c. Do you think we need information about temporal dynamics to analyse behaviours? Which methods can we use to extract them?
2. Perform PCA on the 3D-pose and/or joint angle data. Use time as samples and pose/joint angles as features.
    a. Plot some of the top PCs over time. Which temporal feature dominate them? Oscillations? Slow time-scale transients? Something else?
    b. How many PC do you need to explain 90% of the variance? Is this similar to the analysis in part 1 about neural variance?
3. Combine different feature extractions methods to yield a 2D embedding of behavioural features (e.g. PCA + Wavelet + t-SNE as in Berman et al. 2014)
    a. Which sequence of algorithms makes the most sense?
    d. Plot a scatter plot (i.e., feature 1 on the x-axis and feature 2 on the y-axis). Can you see clusters in your behavioural embedding?
        i. Hint: You are plotting a lot of points. Try to make the individual dots transparent or use a technique similar to kernel density estimation.
    b. Are time points that are close to each other in time in a similar location of your behavioural embedding?
    c. Does the density of points at different locations in the 2D behavioural map change across trials?

Part 2.2: Clustering behaviours

1. Do different behaviours correspond to different locations in your behavioural embedding from part 2.1?
    a. Hint: colour each dot depending on the manual behavioural labels that you have generated last week.
    b. Do you think you can easily perform clustering in this 2D embedding? How would you cluster the data?
    c. If not, how do you think you can improve the features used for clustering? Do you think that more than two features might help?
2. Use a clustering algorithm (e.g., k-means or GMM) to cluster the behavioural data.
    a. How many clusters would you define? Why?
    b. How does the clustering change when you pick different numbers of clusters?
    c. Do clusters found by your algorithm closely correspond to the manual behavioural labels that you identified? Is this good or bad?
    d. Does the assignment to clusters change across time (i.e., trials)?

Part 2.3: Classifying behaviours

1. Do different behaviours correspond to different locations in your behavioural embedding from part 2.1?
    a. Hint: colour each dot depending on the manual behavioural labels that you have generated last week.
    b. Do you think you can train a classifier based on these two features to classify different behaviours?
    c. If not, how do you think you can improve the features used for classification? Do you think that more than two features might help?
2. Use different sets of features and different classifiers to train a behavioural classifier.
    a. Use the following as a baseline: all joint angles and 25 wavelet features with frequencies between 1 and 50 Hz, without PCA, random forest classifier.
    b. Can you do better than this selection of features and the classifier?
    c. Which features work best? Why do you think this is the case?
    d. How do your results compare to the automatic classification that is provided in the data-frame?
    e. Do you think you could easily use this classifier on data from another fly?

For your final report we ask you to describe your analysis and think about the following questions:

1. How can we define different “behaviours”? What features did you find were best for defining a behaviour? Which methods could you use to extract features about the dynamics of the movement?
2. How can you classify behaviour from 3D pose and joint angle time-series? Which method worked best?
3. Can you find clusters in behaviours? How do they align with your manual behavioural labels or the automatic classification?
4. What are advantages/disadvantages of analysing behavioural data with automated algorithms as opposed to manual analysis?
5. What are the advantages/disadvantages of performing clustering of behavioural data as opposed to classification?

Week 10: Combining neural & behavioural data

Starting this week, we will try to find links between neuronal activity of Ascending, and Descending Neurons and behaviour, for example by using methods from modern machine learning. There are many different types of analyses that are possible. In part 1, we will try correlative methods and in part 2, we will attempt to "read the mind" of the fly: we will try to classify the behaviour of the fly based on neuronal activity alone. During last week's and today's classes we introduced a few methods that you can try out to explore the neuronal and behavioural data you have in your hands. If you already know another analysis method that has previously been applied to neuronal and behavioural data or think another method that you know might possibly be applied, you are more than welcome to try it out. If you want some feedback on whether this analysis might make sense, feel free to discuss your ideas with the TAs. As we mentioned before, no one has studied this kind of data before. As a result, this part of the mini-project is may be most exciting because it is exploratory.

At this point it might make sense that each member of your group tries out a different analysis method or even a different part of the steps below and you discuss regularly to understand how different methods can lead to different results or whether one or the other might be more appropriate.

In general, **also think about and answer the following questions in your final report**:

1. What are some improvements one might make on the data quality or technical approach to increase the resolution of what we can conclude from these data?
2. If you were given one year to analyse this and similar/more extensive datasets, what would you try to study and why?

Down-sampling the behavioural data

You might remember that the behavioural data was recorded at 100 Hz, while the neuronal activity was recorded at ~16 Hz. This difference is due to technical limitations of our microscopy system. In order to correlate neuronal activity and behaviour, it therefore might make sense to bring them to the same sampling rate (i.e., down-sample the behavioural data). We uploaded a new Jupyter Notebook showcasing a function that you can use to perform down-sampling. In the notebook, the mean() function is used for down-sampling, but there might be cases where another function, e.g. std(), max(), or min(), etc., might be more appropriate.

1. As a first step, down-sample your behavioural data using the provided function. What would be the best way to down-sample categorical data, such as the behavioural labels?

Part 1: Correlating neuronal activity and behaviour

In this part will try to understand which information the neurons of our dataset "carry". For now, we will assume that a correlation between the activity of one neuron and an external variable means that this neuron "encodes information" about this external variable. What are the limitations of this approach? Can you think of ways to overcome these limitations?

Part 1.1: Identifying correlations of individual neurons

1. Compute the mean activity of a neuron during different behaviours. I.e., average the neuron's activity separately during walking, resting, and so on. Repeat the analysis for every neuron. How does this analysis compare to computing a correlation coefficient?
   Hint 1: You could visualise the results as colours of a matrix with individual neurons on one axis and behaviours on the other axis.
   Hint 2: Standardising the neuronal time-series to have mean zero and unit variance (i.e., z-scoring) may help with the visualisation.

   a. Can you see neurons that are much more active during walking than during all the other behaviours? Or neurons that are active mostly during resting?
   b. How many neurons have different activity levels for different annotated behaviours? Are there equally many for each of the different behaviours? Are neurons active during multiple behaviours or only one?
   c. How might you determine the statistical significance of your results?
2. Identify a few interesting neurons which have different activity levels for different behaviours and plot their activities over time together with their behavioural signals (e.g., a binary signal indicating whether the fly is walking or not).
   a. Do these plots confirm the results from step 1?
   b. Are these neurons "commanding" the behaviour or "reporting" the behaviour?
3. Alternatively, the activity of the neurons might be related to the joint angles. Compute the correlation coefficient between individual neurons and individual joint angles.
   a. Is a Pearson correlation coefficient (scipy.stats.pearsonr()) appropriate? (*Hint*: it assumes Gaussian distributed variables). If not, you could use the Spearman rank-order correlation coefficient (scipy.stats.spearmanr()).
   b. Compute the correlation coefficient for each pair of joint angles and neurons. What is the maximum/minimum correlation coefficient you can find?
4. Identify a few interesting pairs of neurons and joint angles and plot them as scatter plots.
   a. Can you see the correlation by eye? Does it confirm what you saw before?
5. In general, are the neurons' activity more correlated to behavioural categories or to joint angles? What might this tell you about the role of Ascending and Descending Neurons?

Part 1.2: Identifying correlations of low-dimensional signals

1. Perform similar analyses as in Part 1.1 (steps 1-2), but for the principal components of neuronal activity that you identified in the previous week.
   a. Do individual principal components of neuronal activity strongly correlate with individual behaviours? Is the correlation higher than for individual neurons or lower? What might this tell you about how information is transmitted in the nervous system?
2. Look back at the results of your clustering last week.
   a. First, you were clustering time points (i.e., one dot in a scatter plot is one time point) for both behaviour and neuronal activity. Do clusters identified in neuronal activity correspond to clusters in behaviour data? How do you interpret this?
   b. Afterwards, you might have clustered different neurons (i.e., one dot is one neuron). Do neurons in one cluster have similar "behavioural encodings"? I.e., do they correlate with similar behaviours? How do you interpret this?

Part 1.3: "Explaining" neuronal activity

1. How much variance of neuronal activity can you explain from the current behaviour of the fly? (Musall et al 2019 shows an example of using behavioural variables to explain neuronal activity).
   a. Train a multi-variate regression model with binary behavioural variables (i.e., walking yes/no, grooming yes/no, …) as regressors (inputs) of your model and neuronal activity (possibly standardised) as the output of the model. You could use the regression object from sklean.linear_model.LinearRegression() or a custom implementation. Use the fraction of explained variance as a metric.

   b. How much variance can you explain for different neurons? Are there neurons, where you can explain a very high fraction of variance? Are there neurons where you cannot explain anything?
   c. Look at the regressors. How many different behaviours are important to explain the activity of individual neurons? Is it just one – e.g., you find a walking neuron – or is it a combination of different behaviours?
2. Now, instead, use the joint angles as regressors. Can you explain more variance in neuronal activity than with the behavioural categories as regressors? What does this tell you about the function of the neurons?

For your final report we ask you to describe your analysis and to answer the following questions:

1. What algorithmic strategies did you use to identify correlations between neural time-series and behavioural time-series?
2. What were the advantages and disadvantages of this approach?
3. What are alternative approaches you could take to overcome disadvantages of the approach you took?
4. Did individual neurons, or clusters of neurons relate to behavioural categories?
5. Which behaviours were most represented by neurons? Which were not well represented (i.e., few or no neurons)?
6. For (5), why do you think you observed these results? What is the meaning in the context of how the brain is organized and what it is designed to do?
7. Did you observe neurons that were active during resting? If so, what do you think they might be used for?
8. Do you think Descending and Ascending neurons encoding behaviours serve different functions? If so, what are those differences?
9. Did you observe neurons that tracked the movements of individual joints (not larger scale behaviours)? If so, which angles? If not, what might be some technical limitations that would make this difficult to observe from this dataset?

Part 2: Classifying behaviour from neuronal activity

In this part, we will try to “read the mind” of the fly: We will predict its behaviour from its neuronal activity.

Part 2.1: Predicting one behaviour

1. Use one neuron, that you identified in part 1.1 as a “walking neuron” to predict whether the fly is walking or not: Use logistic regression (e.g. sklearn.linear_model.LogisticRegression()) with one neuron’s activity as input and a binary variable that indicates whether the fly is walking or not as output of the model.
   a. How well can you classify walking?
2. Add more neurons as regressors.
   a. How many do you need to accurately predict walking? How does the classification accuracy improve with increasing numbers of neurons? *Hint*: add the neurons sorted by their correlation coefficient with the binary “walking” variable.
3. Include all neurons. Look at the weights of the classifier (the coef_ attribute of the sklearn LogisticRegression object). Are all neurons captured by the weights? How many neurons are actually taken into consideration by the classifier?

Hint: standardise the neuronal signals before training the classifier. This way, the weights become more interpretable.

4. Try the same thing with other behaviours (e.g., anterior grooming). Can you classify those with the same accuracy?

Part 2.2: Predicting multiple behaviours

1. Predict all behaviours at once with multi-class logistic regression. How much performance can your model achieve compared to predicting individual behaviours?
    a. Look at the weights. Do individual neurons contribute to the classification of multiple behaviours or just one?
2. Use a more advanced classifier (e.g., a neuronal network or a random forest) to classify multiple behaviours.
    a. Can you beat the logistic regression classifier? Why do you think this is the case?
    b. Are there any disadvantages to using a neuronal network over a regression technique?

For your final report we ask you to describe your analysis and think about the following questions:

1. Can you classify behaviour from neuronal activity?
3. How many neurons do you need? Is a single neuron enough to reliably classify one or multiple behaviours?
4. What are the advantages and disadvantages of the different methods you tried?