**Background**

In this project, we analyze a subset of data collected by Steinmetz et al. (2019). While this document provides the basic understanding of the experiments, it is highly recommended that one consults the original publication for a more comprehensive understanding in order to improve the quality of the analysis report.

In the study conducted by Steinmetz et al. (2019), experiments were performed on a total of 10 mice over 39 sessions. Each session comprised several hundred trials, during which visual stimuli were randomly presented to the mouse on two screens positioned on both sides of it. The stimuli varied in terms of contrast levels, which took values in {0, 0.25, 0.5, 1}, with 0 indicating the absence of a stimulus. The mice were required to make decisions based on the visual stimuli, using a wheel controlled by their forepaws. A reward or penalty (i.e., feedback) was subsequently administered based on the outcome of their decisions. In particular,

* When left contrast > right contrast, success (1) if turning the wheel to the right and failure (-1) otherwise.
* When right contrast > left contrast, success (1) if turning the wheel to the left and failure (-1) otherwise.
* When both left and right contrasts are zero, success (1) if holding the wheel still and failure (-1) otherwise.
* When left and right contrasts are equal but non-zero, left or right will be randomly chosen (50%) as the correct choice.

The activity of the neurons in the mice’s visual cortex was recorded during the trials and made available in the form of spike trains, which are collections of timestamps corresponding to neuron firing. In this project, we focus specifically on the spike trains of neurons from the onset of the stimuli to 0.4 seconds post-onset. In addition, we only use 18 sessions (Sessions 1 to 18) from four mice: Cori, Frossman, Hence, and Lederberg.

**Data structure**

A total of 18 RDS files are provided that contain the records from 18 sessions. In each RDS file, you can find the name of mouse from mouse\_name and date of the experiment from date\_exp.

session=list()

**for**(i **in** 1:18){

session[[i]]=readRDS(paste('./Data/session',i,'.rds',sep=''))

*# print(session[[i]]$mouse\_name)*

*# print(session[[i]]$date\_exp)*

}

Five variables are available for each trial, namely

* feedback\_type: type of the feedback, 1 for success and -1 for failure
* contrast\_left: contrast of the left stimulus
* contrast\_right: contrast of the right stimulus
* time: centers of the time bins for spks
* spks: numbers of spikes of neurons in the visual cortex in time bins defined in time
* brain\_area: area of the brain where each neuron lives

Take the 11th trial in Session 5 for example, we can see that the left contrast for this trial is 1 the right contrast is 0, and the feedback (i.e., outcome) of the trial is 1. There are a total of 1077 meurons in this trial from 10 areas of the brain. The spike trains of these neurons are stored in session[[5]]$spks[[11]] which is a 1077 by 40 matrix with each entry being the number of spikes of one neuron (i.e., row) in each time bin (i.e., column).

**Question of interest**

The primary objective of this project is to build a predictive model to predict the outcome (i.e., feedback type) of each trial using the neural activity data (i.e., spike trains in spks), along with the stimuli (the left and right contrasts). Given the complexity of the data (and that this is a course project), we break the predictive modeling into three parts as follows.

Part 1 . Exploratory data analysis. In this part, we will explore the features of the data sets in order to build our prediction model. In particular, we would like to (i) describe the data structures across sessions (e.g., number of neurons, number of trials, stimuli conditions, feedback types), (ii) explore the neural activities during each trial, (iii) explore the changes across trials, and (iv) explore homogeneity and heterogeneity across sessions and mice.

Part 2 . Data integration. Using the findings in Part 1, we will propose an approach to combine data across trials by (i) extracting the shared patters across sessions and/or (ii) addressing the differences between sessions. The goal of this part is to enable the borrowing of information across sessions to enhance the prediction performance in Part 3.

Part 3 . Model training and prediction. Finally, we will build a prediction model to predict the outcome (i.e., feedback types). The performance will be evaluated on two test sets of 100 trials randomly selected from Session 1 and Session 18, respectively. The test sets will be released on the day of submission when you need to evaluate the performance of your model.

**Project report outline**

The final submission of the course project is a report in HTML format, along with a link to the Github repository that can be used to reproduce your report. The project report must be legible and the exposition of the report is part of the grading rubrics. For consistency in grading, please follow the outline listed below.

* Title.
* Abstract.
* Section 1 Introduction. Introduce the objective and briefly review the background of this data set.
* Section 2 Exploratory analysis.
* Section 3 Data integration.
* Section 4 Predictive modeling.
* Section 5 Prediction performance on the test sets.
* Section 5 Discussion.