Objectives and Specific Aims

Joaquín Rapela and Maneesh Sahani

March 12, 2025

1 Objectives

1. **Co-develop a new type of neuroscience experimentation** that enables animals to express ethologically relevant behaviours over extended periods in naturalistic environments, while capturing high-precision behavioural and neural activity measurements.

2. **Create a global resource to support research groups worldwide in adopting this new experimental framework**.

3. **Enable global access to the data generated by these experiments** by developing web-based platforms for data access, visualisation, and analysis, while also allowing users to conduct custom analysis.

2 Specific Aims

1. **Develop software for neural and behavioural time-series visual exploration**, operating on continuous and epoched data.

2. **Design efficient statistical data analysis methods** to process long-duration and non-stationary time-series at scale.

3. **Develop batch and online spike sorting methods** to process weeks-to month-long continual electrophysiology recordings and support real-time machine learning inferences.

4. **Build real-time machine learning algorithms** to enable intelligent neural manipulations.

2.1 Challenges

long-duration most conventional machine learning method implementations for neural data analysis will not scale to datasets of the size generated in NaLoDuCo experiments. For example, (with classical methods) it will not be possible to invert a data matrix from a month-long experiment to perform batch linear regression.

Possible solutions:

1. **distributed computing** some algorithms, like batch linear regression, can be efficiently distributed across several computers and be used in very large datasets. However, not all algorithms parallelise well, due to sequential dependencies or high-communication costs, and can be distributed efficiently.

2. **online methods** some batch methods, like batch linear regression, have equivalent online methods, like online Bayesian linear regression, which can process infinite data streams.

non-stationarity alternatives

1. detect performance degradation → retrain
2. use non-stationary methods

2.2 Methods to implement

For behavioural data, we will investigate methods to:

track multiple body parts of animals (e.g., [7] and a switching-linear-dynamical method using RFIDs that we will develop),

infer kinematics of foraging mice (e.g., [8,9]),

segment behaviour into discrete states (e.g., [10] and a hierarchical HMM that we will develop),

infer rules that govern mice behaviour from behavioural observations only (i.e., policy inference) (e.g., [11]).

For neural data, we will investigate methods to:

estimate low-dimensional continual representations of neural activity (i.e., latents inference) (e.g., [12]),

segment neural activity into discrete states (e.g., [13]),

decode environment variables from neural activity (e.g., [14]).

3 Use Case

A user begins by **visualising continuous behavioural measurements**, such as the kinematics (speed and acceleration) of a mouse during a months-long experiment measured by the IMU of an ONIX probe.

Next, she **examines the results of a machine learning analysis**, such as behavioural states inferred by a Switching Hidden Markov Model (SHMM) using kinematic data.

The SHMM was initially trained on the first two hours of the experiment and was periodically retrained to **adapt to non-stationarities**, such as sensor fluctuations, changes in motivation, fatigue, or learning.

She then **visualises epoched data**, such as SHMM states aligned to key events—e.g., the onset of a foraging bout in the richer patch at a specific time of day.

Curious about the neural basis of these SHMM states, she checks the neural recordings but realises they have not yet been spike-sorted. She runs our **offline spike sorting method**, developed for very long continuous recordings, and performs quality control on its results.

Based on quality metrics, she detects drift in the recorded signal, adjusts the drift correction parameters, and reruns the sorting algorithm.

After examining the sorted spikes from a large neural population, she finds it challenging to interpret activity across so many neurons. So she decides to summarise the population spiking activity by **estimating continuous latent variables**.

She then returns to the behavioural visualisation software, integrates machine learning indices corresponding to these latent variables, and **visualises behavioural data aligned to the newly estimated latents**.

Through these explorations, she hypothesises that a peak in a neural latent variable from the prefrontal cortex signals the moment when mice decide to begin a foraging bout.

To test this, she **runs an online machine learning model** to estimate latent variables from prefrontal cortex activity, predicting when this peak will occur. She then **optogenetically inactivates the neural population** at the forecasted time.

Because inactivation prevented the mouse from initiating a foraging bout, her hypothesis was supported.

Notes:

1. for a week-long experiment, the size of the behavioural and neural recordings exceeds 200 terabytes. Due to the large datasets sizes, we will **bring users to data**, instead of data to users. Data will be stored in the cloud, computation will run on the cloud and user computers will only display small data and analysis results.

2. we will **share data, hardware specifications and open-source software openly**.