Contents

Intention to submit document for the Work with US researchers BBSRC-NSF/BIO lead agency 2024 funding opportunity

Enabling Naturalistic, Long-Duration and Continual Neuroscience Experimentation with Advanced Machine Learning

October 23, 2024

1 Summary

Word limit: 2 A summary is not required for this section, please write 'N/A' in the textbox. Please still include a title for your project.

N/A

2 Core team

List the key members of your team and assign them roles from the following:

- project lead (PL)
- project co-lead (UK) (PcL)
- specialist
- professional enabling staff
- research and innovation associate
- technician
- researcher co-lead (RcL)

Only list one individual as project lead.

The core team section must only contain details of the UK applicants. The US applicant information should be listed in the 'US applicants' section.

Find out more about UKRI's core team roles in funding applications.

project lead (PL) Prof. Maneesh Sahani

project co-lead (UK) (PcL) Prof. Tiago Branco, Prof. Thomas Mrsic-Flogel

researcher co-lead (UK) (RcL) Dr. Joaquin Rapela, Dr. Dario Campagner

3 Application questions

3.1 Research theme

Word limit: 5 Please state the research theme you are applying under. Choose one of the following research themes:

1. biological informatics
2. understanding host-microbe interactions
3. synthetic cells and cellular systems
4. synthetic microbial communities

biological informatics

3.2 Vision

Word limit: 500

What are you hoping to achieve with your proposed work?

What the assessors are looking for in your response

Your vision should clearly address:

- one of the opportunity research themes (biological informatics, understanding host-microbe interactions, synthetic cells and cellular systems or synthetic microbial communities)
- the remit of the BBSRC and the NSF/BIO division associated with your chosen research theme

References may be included within this section, but this will count towards your word count.

Images are not required for this section.

3.2.1 Context

Conventional systems neuroscience experiments are typically short in duration and often place significant constraints on subject behavior to simplify data analysis. However, these restrictions may limit our ability to observe critical aspects of brain function and behavior that only manifest in more naturalistic and extended conditions.

At the Sainsbury Wellcome Centre (SWC) for Neural Circuits and Behaviour, we are pioneering Naturalistic, Long-Duration, and Continual (NaLoDuCo) foraging experiments in mice that span weeks to months. During these extended experiments, we collect high-resolution recordings of both behavioral and neural activity in naturalistic settings. In collaboration with the Gatsby Computational Neuroscience Unit (GCNU), we are developing novel analytical methods to interpret this new class of data.

This novel experimental approach will enable researchers to explore neural mechanisms underlying naturalistic behavior over extended periods for the first time, offering the possibility of uncovering insights across a wide range of phenomena, including long-term behavioral adaptation, neural plasticity, and learning. The data generated from NaLoDuCo experiments represent an entirely new resource in neuroscience, with the potential to drive breakthroughs and discoveries that are beyond the reach of traditional experiments.

Our vision is to empower research centers worldwide to adopt this groundbreaking approach. However, the scale and complexity of the data generated pose significant challenges in data acquisition, visualisation, and analysis. In this proposal, we will address these challenges, developing and sharing openly the necessary expertise, hardware, and software to enable this transformative type of experimentation on a global scale.

3.2.2 Focus areas

Below, we outline the key focus areas we aim to address (Figure 4). Challenges addressing these areas primarily revolve around the collection and analysis of continuously recorded, extremely large datasets–on the order of hundreds of terabytes–gathered from experiments spanning weeks to months.

While experiments in neuroscience that are naturalistic, long-duration, or continuous have been conducted in the past [e.g., 14, 19, 32], to the best of our knowledge, we are the first to integrate all three of these features in a single experimental paradigm. This combination introduces unprecedented complexities in data processing, as we aim to capture behavior and brain activity in their most ecologically valid, extended, and uninterrupted forms.

The focus areas of the proposed project are (Figure 4):

Data Collection & Management Efficiently gathering and organizing massive datasets over extended periods.

Data Sharing Providing easy access to large-scale datasets to researchers around the globe using cloud-based technologies.

Data Visualization Developing efficient web-based tools to visualize very large behavioral and neural datasets.

Spike Sorting Assigning spikes to neurons reliably, and tracking individual neurons across long-periods of time in real time.

Data Analysis Evaluating existing methods, and developing new ones, when necessary, to study key behavioral and neural-coding problems with NaLoDuCo experimental data (Figure 2).

Inference-Driven Experimentation Creating a new type of experimentation driven by real-time behavioral and neural inferences.

We are a unique team to implement this project. The SWC is a world leader in experimental neuroscience, working closely with the GCNU, a renowned authority in computational neuroscience and machine learning. Both institutions share the same building and collaborate extensively. NeuroGEARS Ltd. is a key business partern for the implementation of the NaLoDuCo experimental framework, while Catalyst Neuro has played a pivotal role in developing and operating the DANDI archive, in collaboration with Dr. Jeremy Magland, an expert in spike sorting, data visualization, and cloud computing.

3.3 Approach

Word limit: 500

How are you going to deliver your proposed work?

What the assessors are looking for in your response

Your approach should give an overview highlighting:

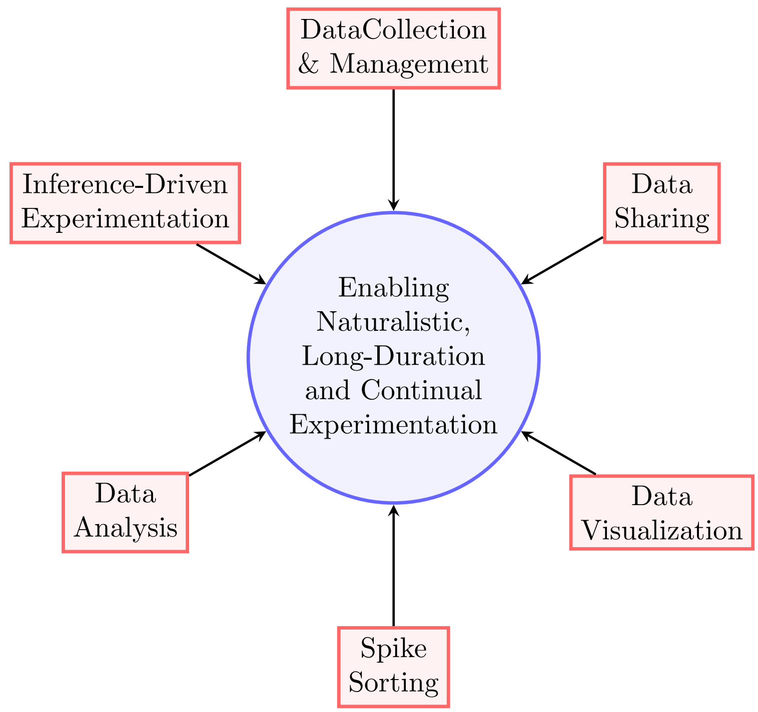
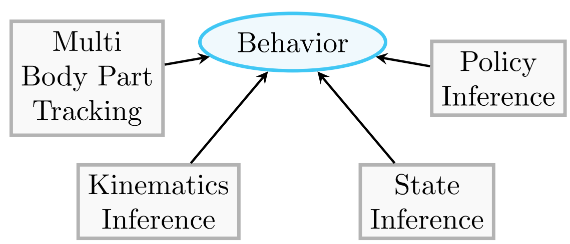
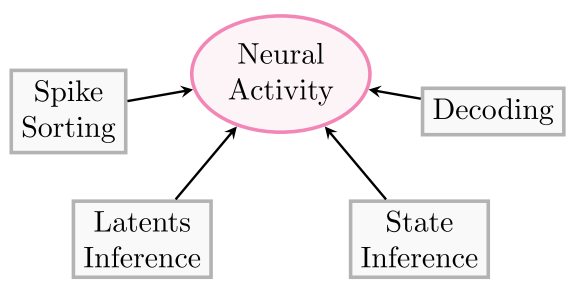


Figure 1: Project theme (blue) and focus areas (red).



(a)



(b)

Figure 2: Behavioral (a) and neural (b) data analysis problems to address.

- a clear description of the objectives and methodology for the proposed work, including the contributions of the UK and US teams
- the potential outputs and outcomes of the proposed work

References may be included within this section, but this will count towards your word count.

Images are not required for this section.

3.3.1 Data collection & management

We have developed a new platform that allows housing of mice in large arenas (>2m diameter), while manipulating and monitoring their behaviour at high spatiotemporal resolution [Figure 3, 2]. We have openly shared software for supporting data acquistion [11] and management [12] in this arena. Using this platform we have collected several week long datasets both with single mouse and multiple mice.

3.3.2 Data sharing

The large dataset sizes generated by NaLoDuCo experiments, on the order of hundreads of terabytes, make it impractical to distribute data to users, and require to bring users to data. Fortunately, cloud technologies are now mature to allows this. We will store data in the Distributed Archives for Neuroscience Data Integration (DANDI), which uses Amazon S3 buckets, and we will provide software to visualize and analyze data in Amazon EC2 instances, to avoid costly data transfers.

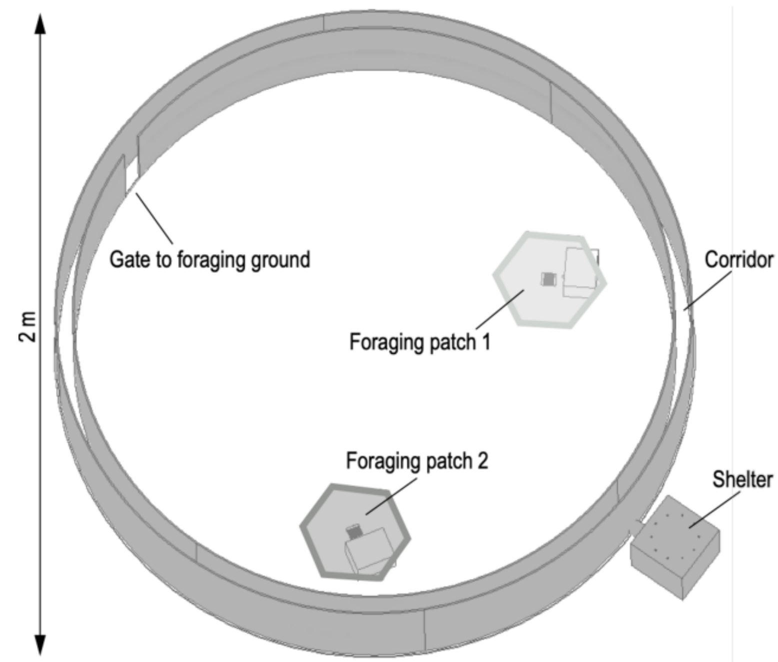
3.3.3 Data visualisation

Our visualisation tools need to display very large datasets at different temporal scales, from milliseconds to weeks and months, and they need to be web based. We will use multi-resolution visualization techniques, which store data at various resolutions, and use the approriate resolution for each zoom level. Web-based visualisation will be optimized using web workers [8].

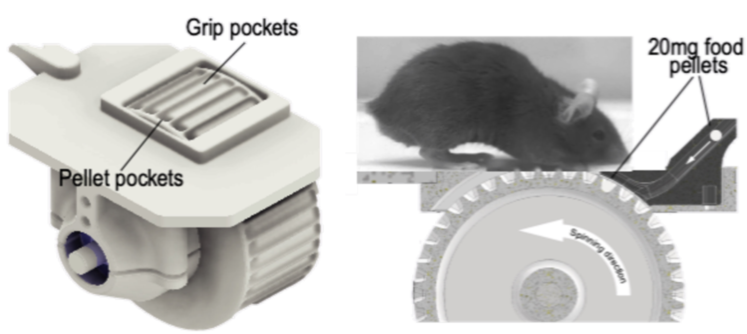
3.3.4 Spike sorting

Spike sorting is specially challenging in NaLoDuCo experimentation since we want to track individual neurons of freely moving mice for weeks to months. In addition, we need online spike sorting, to allow experiments driven by real-time machine learning inference, as described below.

We will evaluate methods for tracking neurons over long periods of time [e.g., 34, 31] and for online sorting [e.g., 26, 29].



(a)



(b)

Figure 3: Foraging arena (a) and feeder (b). The floors of the arenas are tessellated to form honeycombs of modular hexagonal tiles (a), each of which can be equipped with a newly designed underground feeder (b). Pellets are dispensed onto a foraging wheel once the mouse has spun it for a pre-defined programmable distance threshold using its forepaws (fictive digging). The arena contains up to six scale-equipped nesting modules that allows housing of mice in the arena and weight monitoring. Behavioural monitoring is achieved by an array of high-speed cameras (up to 15), by which mouse location, mouse identity and body parts can be track in real time. Long term monitoring of neural activity is performed using Neuropixels probes.

3.3.5 Data analysis

The very large size of NaLoDuCo experimental data, the fact that the statistics of these data change across time, and the requirement for real-time and close-loop inference create new challenges to conventional machine learning methods. We will evaluate existing methods targeting the experimental problems in Figure 2 and, if necessary, modify them, or create new ones, to address the previous challenges.

For behavioral data, we will evaluate methods to:

- track multiple body parts of animals [e.g., 20, 23, 1, and a switching-linear-dyanamical method using RFIDs that we will develop],
- infer kinematics of foraging mice [e.g., 24, 3],
- segment behavior into discrete states [e.g., 33, 13, and a hierarchical HMM that we will develop],
- infer the rules that govern mice behavior from behavioral observations only (i.e., policy inference) [e.g., 36, 35].

For neural data, we will evaluate methods to:

- estimate low-dimensional continual representations of high-dimensional spiking activity (i.e., latents inference) [e.g., 18, 9, 22, 28],
- segment neural activity into discrete states [e.g., 4, 10],
- decode environment variables from neural activity [e.g., 7, 15, 30].

3.3.6 Inference-driven experimentation

We call inference-driven experimentation to a type of experimentation driven by machine learning inferences on neural or behavioral data, where the result of these inferences can change the experiment in real time.

We will apply inference-driven experimentation to test if patterns of neural activity are causally related to foraging behaviors. We would first check that a pattern of neural activity always precedes a given foraging behavior. We would then detect the occurence of the pattern and in real time optogenetically inactivate the neurons responsible for the pattern. If the behavior dissapears the causality argument would be supported.

For this we will use the Bonsai ecosystem for experimental control [17] and online machine learning functionality that we are adding to Bonsai [25], funded by a BBSRC award [21].

3.4 US applicants

Word limit: 200

Please provide the following details of the US applicants on this application:

1. name
2. institute
3. job title
4. role in project (for example, project lead or project co-lead)
5. email address

Please also indicate who the lead US applicant will be.

NSF will use this information to confirm applicant eligibility.

Please do not include details of US applicants in the 'Core team' section.

3.5 Resources

Word limit: 200

Please provide the following:

- overall estimates for costings and staffing full time equivalent (FTE) for both the UK and US components
- clear separation of UK and US costings, in pounds sterling and US dollars (USD) respectively

The overall budget should be below the maximum £2 million combined funder contribution

If there is more than one UK or US team associated with the application, please combine their estimates together.

A detailed calculation and breakdown of resources is not required at this stage, nor is a justification of costs.

The following is an example of how this might look.

UK Resources:

Total cost estimate: £600,000

Research council contribution: £480,000

0.2 FTE time, 1.0 FTE PDRA, 0.5 FTE technician

US Resources:

Total cost estimate: $300,000

1.0 FTE PDRA or 1.0 FTE doctoral researcher

Total funder contribution estimate:

£716,475 (£480,000 + £236,475 ($300,000 at exchange rate 0.79))

References

References

[1] Dan Biderman, Matthew R Whiteway, Cole Hurwitz, Nicholas Greenspan, Robert S Lee, Ankit Vishnubhotla, Richard Warren, Federico Pedraja, Dillon Noone, Michael M Schartner, et al. Lightning pose: improved animal pose estimation via semi-supervised learning, bayesian ensembling and cloud-native open-source tools. *Nature Methods*, pages 1–13, 2024.

[2] D. Campagner, J. Bhagat, G. Lopes, L. Calcaterra, J. Ahn, A. Almeida, F. J. Carvalho, B. Cruz, A. Erskine, C. Lo, T. T. Nguyen, A. Pouget, J. Rapela, T. Ryan, J. Reggiani, and S. SWC Foraging Behaviour Working Group. Aeon: an open-source platform to study the neural basis of ethological behaviours over naturalistic timescales. In *Society for Neuroscience Abstracts*, page PSTR033.03 / I26, 2024. Presented at the Society for Neuroscience Annual Meeting.

[3] Subhash Challa, Mark R. Morelande, Darko Mušicki, and Robin J. Evans. *Fundamentals of Object Tracking*. Cambridge University Press, 2011.

[4] Zhe Chen, Sujith Vijayan, Riccardo Barbieri, Matthew A Wilson, and Emery N Brown. Discrete-and continuous-time probabilistic models and algorithms for inferring neuronal up and down states. *Neural computation*, 21(7):1797–1862, 2009.

[5] Kelly B Clancy and Thomas D Mrsic-Flogel. The sensory representation of causally controlled objects. *Neuron*, 109(4):677–689, 2021.

[6] Karl Deisseroth. Optogenetics: 10 years of microbial opsins in neuroscience. *Nature neuroscience*, 18(9):1213–1225, 2015.

[7] Xinyi Deng, Daniel F Liu, Kenneth Kay, Loren M Frank, and Uri T Eden. Clusterless decoding of position from multiunit activity using a marked point process filter. *Neural computation*, 27(7):1438–1460, 2015.

[8] MDN Web Docs. Web workers. `https://developer.mozilla.org/en-US/docs/Web/API/Web_Workers_API`, 2024. Web Workers makes it possible to run a script operation in a background thread separate from the main execution thread of a web application. The advantage of this is that laborious processing can be performed in a separate thread, allowing the main (usually the UI) thread to run without being blocked/slowed down.

[9] Lea Duncker and Maneesh Sahani. Temporal alignment and latent gaussian process factor inference in population spike trains. In *Advances in Neural Information Processing Systems*, pages 10445–10455, 2018.

[10] Sean Escola, Alfredo Fontanini, Don Katz, and Liam Paninski. Hidden markov models for the stimulus-response relationships of multistate neural systems. *Neural computation*, 23(5):1071–1132, 2011.

[11] SWC Foraging Behavior Working Group. aeon_acquisition repository. https://github.com/SainsburyWellcomeCentre/aeon_acquisition, 2024. Task control and acquisition systems for Project Aeon.

[12] SWC Foraging Behavior Working Group. aeon_mecha repository. https://github.com/SainsburyWellcomeCentre/aeon_mecha, 2024. Project Aeon's main library for interfacing with acquired data. Contains modules for raw data file input/output, data querying, data processing, data quality control, database ingestion, and building computational data pipelines.

[13] Alexander I Hsu and Eric A Yttri. B-soid, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature communications*, 12(1):5188, 2021.

[14] Hueihan Jhuang, Estibaliz Garrote, Xinlin Yu, Vinita Khilnani, Tomaso Poggio, Andrew D Steele, and Thomas Serre. Automated home-cage behavioural phenotyping of mice. *Nature communications*, 1(1):68, 2010.

[15] Fabian Kloosterman, Stuart P Layton, Zhe Chen, and Matthew A Wilson. Bayesian decoding using unsorted spikes in the rat hippocampus. *Journal of neurophysiology*, 111(1):217–227, 2014.

[16] Mikhail A Lebedev and Miguel AL Nicolelis. Brain–machine interfaces: past, present and future. *TRENDS in Neurosciences*, 29(9):536–546, 2006.

[17] NeuroGEARS Ltd. Bonsai. https://bonsai-rx.org/, 2024. A visual language for reactive programming.

[18] Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural information processing systems*, 24, 2011.

[19] Dun Mao, Eric Avila, Baptiste Caziot, Jean Laurens, J David Dickman, and Dora E Angelaki. Spatial modulation of hippocampal activity in freely moving macaques. *Neuron*, 109(21):3521–3534, 2021.

[20] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

[21] Thomas Mrsic-Flogel. Machine intelligence for neuroscience experimental control. https://gow.bbsrc.ukri.org/grants/AwardDetails.aspx?FundingReference=BB%2FW019132%2F1, 2023. BBSRC award BB/W019132/1.

[22] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 15(10):805–815, 2018.

[23] Talmo D Pereira, Nathaniel Tabris, Arie Matsliah, David M Turner, Junyu Li, Shruthi Ravindranath, Eleni S Papadoyannis, Edna Normand, David S Deutsch, Z Yan Wang, et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495, 2022.

[24] Joaquin Rapela. Linear dynamical systems in python. https://github.com/joacorapela/lds_python, 2024. Python code to estimate Gaussian linear dynamical systems.

[25] Joaquin Rapela, Nicholas Guilbeault, and Goncalo Lopes. Bonsai machine learning. https://bonsai-rx.org/machinelearning/, 2024. Machine learning functionality for experimental control in Bonsai.

[26] Ueli Rutishauser, Erin M Schuman, and Adam N Mamelak. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *Journal of neuroscience methods*, 154(1-2):204–224, 2006.

[27] Maneesh Sahani. *Latent variable models for neural data analysis*. Phd thesis, California Institute of Technology, Pasadena, CA, May 1999. Available at https://www.proquest.com/docview/304498531.

[28] Omid G Sani, Hamidreza Abbaspourazad, Yan T Wong, Bijan Pesaran, and Maryam M Shanechi. Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nature Neuroscience*, 24(1):140–149, 2021.

[29] Gopal Santhanam, Maneesh Sahani, Stephen I Ryu, and Krishna V Shenoy. An extensible infrastructure for fully automated spike sorting during online experiments. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2, pages 4380–4384. IEEE, 2004.

[30] Ardi Tampuu, Tambet Matiisen, H Freyja Ólafsdóttir, Caswell Barry, and Raul Vicente. Efficient neural decoding of self-location with a deep recurrent network. *PLoS computational biology*, 15(2):e1006822, 2019.

[31] Enny H van Beest, Célian Bimbard, Julie MJ Fabre, Sam W Dodgson, Flóra Takács, Philip Coen, Anna Lebedeva, Kenneth D Harris, and Matteo Carandini. Tracking neurons across days with high-density probes. *Nature Methods*, pages 1–10, 2024.

[32] Benjamin Voloh, David J-N Maisson, Roberto Lopez Cervera, Indirah Conover, Mrunal Zambre, Benjamin Hayden, and Jan Zimmermann. Hierarchical action encoding in prefrontal cortex of freely moving macaques. *Cell reports*, 42(9), 2023.

[33] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

[34] Augustine Xiaoran Yuan, Jennifer Colonell, Anna Lebedeva, Michael Okun, Adam S Charles, and Timothy D Harris. Multi-day neuron tracking in high-density electrophysiology recordings using earth mover's distance. *Elife*, 12:RP92495, 2024.

[35] Hao Zhu, Brice De La Crompe, Gabriel Kalweit, Artur Schneider, Maria Kalweit, Ilka Diester, and Joschka Boedecker. L(m)v-iql: Multiple intention inverse reinforcement learning for animal behavior characterization. *arXiv preprint arXiv:2311.13870*, 2023.

[36] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

A More details about the vision

A.1 Context

Conventional systems neuroscience experiments are typically short in duration and often place significant constraints on subject behavior to simplify data analysis. However, these restrictions may limit our ability to observe critical aspects of brain function and behavior that only manifest in more naturalistic and extended conditions.

At the Sainsbury Wellcome Centre (SWC) for Neural Circuits and Behaviour, we are pioneering Naturalistic, Long-Duration, and Continual (NaLoDuCo) foraging experiments in mice that span weeks to months. During these extended experiments, we collect high-resolution recordings of both behavioral and neural activity in naturalistic settings. In collaboration with the Gatsby Computational Neuroscience Unit (GCNU), we are developing novel analytical methods to interpret this new class of data.

This novel experimental approach will enable researchers to explore neural mechanisms underlying behavior over extended periods for the first time, offering the possibility of uncovering insights across a wide range of phenomena, including long-term behavioral adaptation, neural plasticity, and learning. The data generated from NaLoDuCo experiments represent an entirely new resource in neuroscience, with the potential to drive breakthroughs and discoveries that are beyond the reach of traditional experiments.

Our vision is to empower research centers worldwide to adopt this groundbreaking approach. However, the scale and complexity of the data generated pose significant challenges in data acquisition, visualisation, and analysis. In this proposal, we will address these challenges, developing and sharing openly the necessary expertise, hardware, and software to enable this transformative type of experimentation on a global scale.

A.2 Focus areas and their challenges

Below, we outline the key focus areas we aim to address (Figure 4), along with their associated challenges. These challenges primarily revolve around the collection and analysis of continuously recorded, extremely large datasets–on the order of hundreds of terabytes–gathered from experiments spanning weeks to months.

While experiments in neuroscience that are naturalistic, long-duration, or continuous have been conducted in the past [e.g., 14, 19, 32], to the best of our knowledge, we are the first to integrate all three of these features in a single experimental paradigm. This combination introduces unprecedented complexities in data processing, as we aim to capture behavior and brain activity in their most ecologically valid, extended, and uninterrupted forms.

A.2.1 Data acquisition and management

At the SWC we have already performed foraging experiments in mice continuously collecting behavioral and experimental data 24 hours a day for seven days. We will share openly the specifications of the hardware used to build these experiments (e.g., instructions for building large foraging arenas, video cameras specifications, electrophysiological recording hardware), as well as the software we used for experimental control, data quality control, data access and management.

The data acquisition and management software used in our project is already publically available in GitHub[1]. This software is already being used by scientists at the Allen Institue for Neural Dynamics and at Northwester University. We will substantially improve its documentation to simplify its usage by external users.

Challenges related to data acquisition and management include data indexing to allow fast access to very large amount of saved data, online quality control and alert systems to guarantee that anomalities in data collection are detected and corrected with minimal delay, and syncrhonization between multiple data streams.

[1] https://github.com/SainsburyWellcomeCentre/aeon_mecha

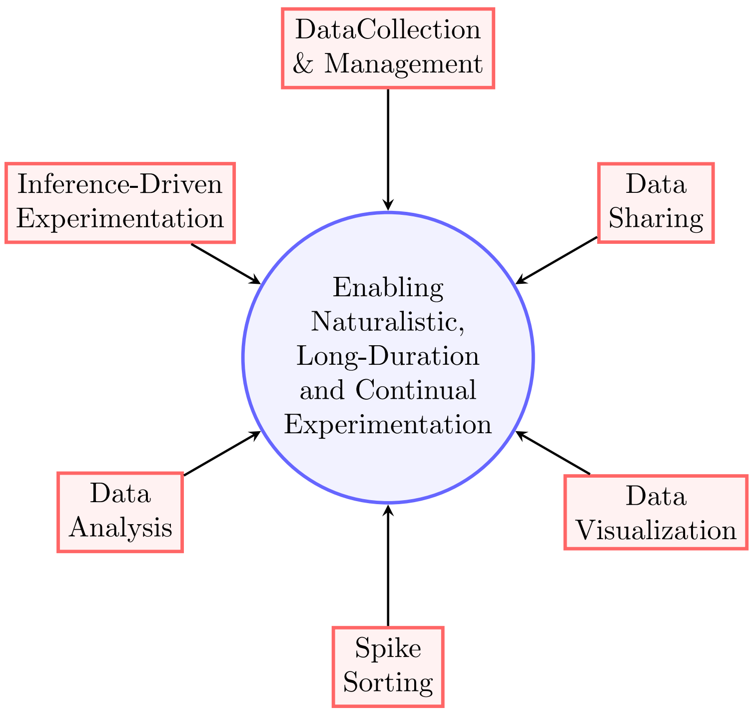


Figure 4: Project theme (blue) and focus areas (red).

A.2.2 Data dissemination

Datasets of the scale of hundreads of terabytes cannot be practically downloaded from data repositories. This is specially true for contiguous experiments where unique insights are extracted by characterizing full datasets, and not only parts of them. Therefore, we will store data in DANDI, which uses Amazon S3 buckets, and provide software in Amazon EC2 instances to visualize and analyze data on the cloud, avoiding costly data transfers. That is, the large dataset sizes of NaLoDuCo experiments make it impractical to distribute data to users and require to bring users to data. Fortunately, cloud technologies are now mature to allows this.

Importantly, if we distributed these very large datasets to users, only those in large research centers would have the computing power to process them. But, by deploying data and computing in the cloud, any person with Internet access around the world will be able to benefit from them. Storing large datasets in DANDI is free.

Dr. Ben Ditcher, founder of CatalystNeuro, has played a pivotal role in supporting the development and operations of the DANDI archive.

A.2.3 Data visualisation

Visualisations are essential for scientific discovery. For the proposed project visualisation present two major challenges. First, they need to display very large datasets at different temporal scales, from milliseconds to weeks and months. Second, as data and software will be deployed in the cloud, visualisation need to be web based. Standard visualization tools cannot display terabyte sized datasets. We will build custom web-based visualization tools to do this.

We have substantial experience building web-based visualization tools for neurophysiological data. Dr. Jeremy Magland is now developing Neurosift[2] a web-based visualizer for DANDI datasets.

A.2.4 Spike sorting

When electrodes are placed in the brain, they typically record spikes from multiple nearby neurons. Spike sorting attributes spikes to individual neurons.

Spike sorting is specially challenging for NaLoDuCo experiments. First, because these experiments require to track individual neurons of freely moving mice for weeks to months. Second, because spike sorting needs to be done online, to allow experiments driven by real-time machine learning inference, as described below.

Prof. Sahani pioneered the use of Bayesian inference methods for spike sorting [27]. Dr. Jeremy Magland has significantly advanced the field of spike sorting, particularly through his development of MountainSort[3] and his contributions to SpikeInterface[4].

[2] https://github.com/flatironinstitute/neurosift
[3] https://github.com/flatironinstitute/mountainsort5
[4] https://github.com/spikeinterface/spikeinterface

A.2.5 Data analysis

Advanced data analysis methods are indispensable to extract meaning from NaLoDuCo experimental data. However, analyzing this data is challenging for at least three reasons. First, important insights will most probably come from the characterization of complete datasets, and not form subsets extracted from them. Conventional batch methods cannot be used with datasets of the size produced by NaLoDuCo experiments. For instance, for learning, batch linear regression cannot load into memory and invert a data matrix with high-resolution observations from a one-month-long experiment. Thus, **online methods** that can process infinite data steams become mandatory.

Second, a pervasive assumption in most ML algorithms is stationarity; i.e., the assumption that the statistics of data do not change over time. But in long-duration and continuous experiments this assumption is most often violated as, for example, the arousal of subjects changes. Hence, the analysis of data generated by these experiments requires **adaptive methods**.

Third, statistical algorithms consist of two key stages: learning (or trainning) and inference (or prediction). The learning stage identifies model parameters, and the inference stage uses the learned model to make predictions, or infer latent variables, from new unseen data. Frequently training is performed on a small subset of a dataset, and inference is done on the remaining data. However, since in long-duration and continual experiments behavior and neural activity are generall not stationary, it is not optimal to train models on data subsets and use them to make inferences on the remaining data, since the state of the animal at training and inference times may be different. To overcome this difficulty we will use **continual learning methods**.

We will evaluate methods to analyze different aspects of behavior and neural activity (Figure **??**). We will test how these methods process very large datasets, how they handle non-stationary data, and how feasible is to retrain them to adapt to changing conditions. We will adapt these methods so that they better address these challenges and, when needed, develop new ones. We will carefully report the outcomes of these evaluations so that researchers performing NaLoDuCo experimentation can choose the best methods that suit their needs.

A.2.6 Experiments driven by real-time machine learning inference

Small animal experiments are usually controlled by simple static rules or direct behavioral observations. Funded by a BBSRC award[5] we are developing software to allow a new type of experimental control based on statistical inferences made on behavioral and/or neural measurements.

For example, after inferring latent variables from neural activity and observing that one of these latents have crossed a threshold, we can deliver a reward [as done in learning to control a BCI; 5], or perform an action [as done in motor imagery BCI; 16], or manipulate of neural activity [as done when studying the

[5] https://gow.bbsrc.ukri.org/grants/AwardDetails.aspx?FundingReference=BB%2FW019132%2F1

causal relation between a pattern of brain activity and behavior; 6]. We propose to further develop the previous software and use it to test causal effects of neural activity patterns on foraging decisions using our NaLoDuCo foraging experiments.

Buidling experiments driven by real-time machine learning inferences brings at least two challenges. The first one is a machine learning problem, how to build fast inferences that can operate in real time. The second one is a neuroscience problem, how to identify neuroscience experiments suitable to real-time control, and then perform the experiment with real-time control. Fortunately at the Gatsby Unit we are experienced on building advanced machine learning algorithms to address the first challenge. And at the SWC we perform many sophisticated animal experiments that could benefit from real-time experimental control.

In summary, we are pioneering a new paradigm in neuroscience experimentation, driven by advanced inferential methods applied to rich behavioral and neural recordings. This innovative technology has the potential to transform the field, enabling experiments that were previously unimaginable. By leveraging these sophisticated inferences, we may unlock new dimensions of knowledge that could not be achieved through simpler, conventional approaches. This breakthrough could open doors to insights that redefine our understanding of brain-behavior relationships.