

Milestone 1

Mataeo Anderson, Jeffrey Gregory

AAI-3303 - September 25, 2025

Problem Definition

Jeffrey and I are going to undertake entering a Kaggle machine-learning competition for our course project. The competition is the MABe challenge: Social Action Recognition in Mice. An overview of the competition can be found [here](#).

Context and Overview:

The problem is important because manually reviewing footage to describe animal behaviours is an extremely time intensive, and boring task. It requires researchers to watch footage for hours and painstakingly annotate every action and behaviour they observe and then take careful, precise, notes detailing the video, the animal subjects, the action or behaviour, etc. It would be very beneficial if this task could be automated using machine learning to train a model to accurately describe these actions and behaviours, and even predict new ones. A well-trained model could observe and predict patterns, behaviours, and interactions that a human observer may miss. It is a very interesting problem because animal behaviour has often been somewhat foreign to humans, always striving to understand their languages, their behaviours, their interactions. But it has always eluded us why animals behave how they do in most common-place situations. Machine learning could fundamentally change our understanding of the animal kingdom.

Specific Objective:

We will be trying to classify animal actions. The competition provides a data set of coordinates of mouse body-parts during pre-annotated interactions and actions, from which the model (ideally) can distinguish patterns of movement to ascribe to the different actions. This will be supervised learning as the data set already has labelled actions with data describing the mouse movements during those actions and trying to classify un-labelled movements in the test data.

Scope and Feasibility:

The data set we will use has been provided by the kaggle competition. It is suitable to use because it is suitably large without being unnessecarilly so for this task. It contains information that should be suitable to use to train a model to classify interactions between mice based on the labelled sets of training data provided in the kaggle data set.

Expected Outcome: We expect the ML solution will provide insight into new animal behaviours and interactions that human observers may have missed. It should also allow for some degree of automation when labelling and annotating future interactions and hopefully provide a foundation for expanding into other members of the animals kingdom to make meaningful observations about the interactions of other species as well.

Dataset Source

The dataset we will use is freely available on Kaggle through the competition. It can be found through [this link](#).

It can be accessed directly using a Kaggle notebook and importing the data into the notebook. It can also be downloaded locally for uses such as doing EDA on your local machine.

The data set is essentially 2 directories containing parquet files; one directory's parquet files contains the "features" (tracking) and one directory's parquet files contain the labels (annotations). Reading all the parquet files into dataframes and loading them all into memory uses about 14GB of memory from Kaggle's cloud computing, using a Kaggle notebook.

The annotation parquet files all have many many rows but the same number of columns. The columns are agent_id (mouse performing an action), target_id (mouse they perform it on/with), action (the action performed), start_frame (when the action was started/initiated), and end_frame (when the action ended).

The tracking parquet file columns are video_frame (when the data was recorded), mouse_id (the mouse the data is associated with), bodypart (the bodypart being described by the x and y coordinates), x, and y (the x and y coordinates of the body part being described).

There are also 2 CSV files with different metadata about the mice, labs, and videos, etc. One for the training sets, and an example one for how the test sets will be structured.

The key features of the data are the video, the labelled actions, the mice IDs, and the mouse body parts with x and y coordinates per frame (can be restructured/visualized to show a projection in 2D of the mouse movements.)

Team Contract

Roles & responsibilities: Because this is only a small two-person team, we would like to split the roles and responsibilities somewhat equally as this seems fair. Our idea was that both of us would work through the project in parallel and then make weekly/bi-weekly merges to a unified github repository after discussing our work over the period together. We thought parallelizing the workflow for this challenge might offer better efficiency for the kind of problem we are trying to solve.

Communication Expectations: As mentioned, we are aiming for weekly or bi-weekly meetings. I'd imagine we will also discuss this before and after our classes together on Tuesdays and Thursdays. Discord and text messaging are the preferred platforms of communication.

Deadlines & accountability: This isn't perfectly set in stone due to our small team size. If any work is missed, we will need to momentarily assess with whom it would be most efficient to offload the extra work, contextually. We'd like to aim to make 3 submissions to the Kaggle competition over the run of the project's lifetime if possible, working out to roughly one attempt a month.

Conflict resolution: Again, being a two-person team our conflict resolution structure is likely to remain relatively relaxed and interpersonal. If we cannot come to an agreement for any reason whatsoever then some sort of future third-party arbiter would necessarily need to be agreed upon.

Signed:

Jeffrey Gregory

Mataeo Anderson

09/25/25

Progress Reports Plan

How will we track progress?

We will track progress using our communication platforms including SMS and Discord while using a sole GitHub repository as the primary source of truth over the life of the project. Each team member will have a local branch which we will regularly merge back into the main after discussing work done, as a group, with the goal of incremental and parallelized development over the period of the competition.

Short progress update outline:

For each progress update on each milestone, we will likely submit a comprehensive report containing:

- All commits made to the repository since the last report.
- A detailed explanation of each commit, why it was made, and its functionality that goes beyond the standard commit comment.
- Any relevant outside materials, sources, or references that have been used either directly or as inspiration to the project.
- Note any changes in our meta-level plan to complete our project and enter into the Kaggle competition.
- An abstract, high-level explanation of the project as it exists in its current state that explains our architectural, engineering, and design decisions.
- Any other relevant comments, information, or material we might contextually decide is necessary to include in our explanation.