Problem Statement and Goals
Software Engineering

Team #18, Gouda Engineers
Aidan Goodyer
Jeremy Orr
Leo Vugert
Nathan Perry
Tim Pokanai

Table 1: Revision History

Date	Developer(s)	Change
Date1	Name(s)	Description of changes
Date2	Name(s)	Description of changes
...	...	...

1 Problem Statement

Current behavioral neuroscience research on Obsessive-Compulsive Disorder (OCD) is limited by the lack of accessible tools for managing and analyzing large-scale animal model data sets. With roughly 20,000 trials worth of rat behavioural data constituting one of the most extensive data sets in the field, the data set contains video recordings of rat behaviour with corresponding spatial-temporal tracking data, and additional research files. Although the data set is publicly accessible and contains rich and diverse forms of data, it is not presented in a user-friendly way for researchers to analyze. The current state of the data set impacts researchers by making it difficult to synchronize trajectories with video recordings and extract meaningful insights from correlated data types. This gap in infrastructure reduces the scientific utility of an extensive data set, ultimately slowing the pace of discovery through research.

[You should check your problem statement with the problem statement checklist. —SS]

[You can change the section headings, as long as you include the required information. —SS]

1.1 Problem

1.2 Inputs and Outputs

[Characterize the problem in terms of "high level" inputs and outputs. Use abstraction so that you can avoid details. —SS]

Inputs from users will consist of various query prompts or data visualization requests. These should be easily done by users without technical experience. This will include:

1. Submitting queries by applying filter criteria related to the experimental sessions or the data itself. These could be but are not limited to: type of file, selected study/experiment, drug injection used in the trial, rat bodyweight, date and time of trial, etc...

2. Submitting queries using natural language. "show me trials with strong checking behavior after 5 injections" or "find sessions where rats showed compulsive patterns" are examples of this

3. Making requests for data visualizations such as behavioral metrics (seperated by injection type for example) or visually plotting trajectories of the rats based on their (x,y) coordinates.

Outputs to users will be constrained to two main categories.

Category 1: raw data records returned from a query for the inputting user to view or extract

Category 2: Data visualizations which include plots of rat trajectories or statistical or graphical displays of behavioural metrics of the rat subjects.

1.3 Stakeholders

The application produced from this project will affect multiple stakeholders, which will either be affected directly or indirectly affected:

1. **Direct Stakeholders:**
    - Behavioural Neuroscience Researchers: Researchers like Dr. Szechtman and Dr. Dvorkin-Gheva will be end users of the platform and will be the core benefitors from the user-friendly search, visualization, and analysis of data.
    - Data Scientists: This group of users will benefit from the structured database, API query system, and processing tools included in the data processing pipeline.

- Graduate Students and Lab Members: These will be very frequent users of the application, who will benefit from the simplified preprocessing, visualization, and analysis tools.

2. **Indirect Stakeholders:**

   - OCD Clinicians: Clinicians don't use the platform directly, but they may incorporate researched insights into a clinical understanding of OCD by refining trials and treatments.
   - Patients with OCD: This stakeholder group may indirectly benefit with an improved quality of life from research findings, such as new treatments, accelerated by this platform.
   - Global Research Community: These collaborating institutions, potentially involved in different areas of research, may indirectly benefit from the streamlined data and its reusability, even if the platform isnt't built directly for them.

1.4 Environment

[Hardware and Software Environment —SS]

The system must be designed to provide efficient search and querying capabilities over a large, multi-terabyte dataset. The raw data files will remain externally hosted, while our system will manage the metadata and indexing required for efficient access. We anticipate the need for cloud-based database and backend infrastructure in order to load and serve experiment data for users worldwide.

A publicly accessible web-based platform is required to make the behavioural dataset accessible to the global research community. As such, we must anticipate users accessing the system from a range of web-capable devices, including desktops, laptops, and mobile devices. Users may also be accessing the system from a broad range of network conditions, from high-latencty connections to low bandwidth environments. As such, the system must be designed to be efficient and responsive under varying network conditions.

The system will also expose natural language querying capabilities to enable researchers with limited technical expertise to access the dataset. This necessitates the need for natural language processing tools or large language models to be integrated into the backend infrastructure. As such, the system design must be open to the integration of third-party APIs and services to support these capabilities.

2 Goals

2.1 Goal 1: Sound and Complete DBMS

A complete database schema must be developed for the purposes of this project. This means that all pieces of data in the FRDR repository have been accounted

for and are uniquely identified within tables in the DBMS. This also means that all metadata related to each piece of data are accessible for querying. Finally, the relations between data must be properly represented. Video, track files and trajectory diagrams related to the same session must have relationships between them.

Furthermore, a proper query framework must be set up and able to access the DBMS. The MVP form of this would just be writing SQL into a database request and recieve the correct result. This MVP would essentially just need Rest APIs set up to make a request to the database.

2.2 Goal 2: Sleek and Non-Technical UI

The main user base for this project will be academics with a psychiatric or animal related background and thus the user interface needs to allow for non-technical people to easily make rather complex queries. Our goal is for our front end system to have an intuitive interface for filtering and searching for data. This would include features such as adding easily adding filters related to metadata annotations, a certain study, or the date of the trial. These filters would then be converted into an equivalent query, sent to the DBMS and return the relevant data. In a similar way, this UI will include natural language searching in which a sentence for what the user wants will be similarily converted into an equivalent query and the data returned.

2.3 Goal 3: Data Visualization/Processing Capabilities

Due to the large amount of data in the repository, processing and visualization are important in helping users more easily interpret and draw conclusions from the data. The first part of this goal is to successfully implement an algorithm that can identify behaviours in the trials and group them accordingly. The MVP of this is simply an algorithm that can distinguish between compulsive and non-compulsive behaviour.

On the visualization side, graphical displays of important metrics should be an option for users to view through the user interface. The different visualizations will likely need to be developed on an ad-hoc basis depending on the user needs but an example of an MVP for this goal would be to provide graphical displays of compulsive behavioural metrics based on the type of injection the rat recieved.

3 Stretch Goals

3.1 Goal 1: Dataset-Agnostic Reusability

To give our project life outside of datasets for rats with OCD, the program should be made as an open platform extensible for any dataset. This would allow different users in other studies and different fields to sort and visualize their data to make it more accessible to others. By doing this, the program

would be future-proof allowing it to be used far beyond the scope of the project and for different uses.

3.2 Goal 2: Insights

Beyond basic data visualization, the program should have an extended toolkit of visualizations that provide insights on the data, more than the basic ones provide. This would be supported by data preprocessing techniques in the backend along with data pipelines and machine learning integration that would allow the program to provide more sophisticated insights that could go unnoticed by the human eye.

4 Extras

The first extra deliverable will be a performance report on areas of the software application where performance should be optimized. The second extra deliverable will be a user manual on the intended use of the software application.

[For CAS 741: State whether the project is a research project. This designation, with the approval (or request) of the instructor, can be modified over the course of the term. —SS]

[For SE Capstone: List your extras. Potential extras include usability testing, code walkthroughs, user documentation, formal proof, GenderMag personas, Design Thinking, etc. (The full list is on the course outline and in Lecture 02.) Normally the number of extras will be two. Approval of the extras will be part of the discussion with the instructor for approving the project. The extras, with the approval (or request) of the instructor, can be modified over the course of the term. —SS]

Appendix — Reflection

[Not required for CAS 741 —SS]

The purpose of reflection questions is to give you a chance to assess your own learning and that of your group as a whole, and to find ways to improve in the future. Reflection is an important part of the learning process. Reflection is also an essential component of a successful software development process.

Reflections are most interesting and useful when they're honest, even if the stories they tell are imperfect. You will be marked based on your depth of thought and analysis, and not based on the content of the reflections themselves. Thus, for full marks we encourage you to answer openly and honestly and to avoid simply writing "what you think the evaluator wants to hear."

Please answer the following questions. Some questions can be answered on the team level, but where appropriate, each team member should write their own response:

1. What went well while writing this deliverable?

   Our team was very aligned on the every part making the writing process very fluid. Through our open communication with each other it made the entire process much easier. We did not have any disagreements and could freely bounce ideas off of one another which helped speed up the brainstorming process and any critiques would be received in a professional manner where we could discuss with the team and find one answer. Furthermore, our discussions were noted in quick jot notes that helped us while writing our problem statement, goals, and stakeholders.

2. What pain points did you experience during this deliverable, and how did you resolve them?

   During the problem statement, we faced a couple of pain points. The first one being our understanding of different sections and ensuring the content of our sections matched what was required of us. To alleviate this pain point, we did through sweeps of what was provided and based our answers off of that. Furthermore, we did not know that there was a rubric until later in the problem statement deliverable, which would have been more help earlier on. Another pain point was that one of our group members laptop broke. We assigned them the later parts of the deliverable so they can contribute to the deliverable once they had a computer.

3. How did you and your team adjust the scope of your goals to ensure they are suitable for a Capstone project (not overly ambitious but also of appropriate complexity for a senior design project)?

   Our team adjusted the scope of our goals to make it more suitable for capstone by focusing on the problem at hand, rather than an ideal world solution, which we saved for the stretch goals. An example of this would be the use of machine learning to train on the dataset to provide us with better insights on the data. This would be an incredible feature, but due

to the complexity and time constraints of the course, could potentially be a stretch. For that reason, we focused on the problem our supervisor would like us to solve, then add extra features if possible.