## ISIS-ML Project Application Form

*Please read the notes at the end before completing the form. Completed forms should be sent to* [anders.markvardsen@stfc.ac.uk](mailto:anders.markvardsen@stfc.ac.uk) *and be a maximum of five pages.*

|  |
| --- |
| **Project title**  “Enough is enough!” – a framework for stopping experiments with sufficient statistics |
| **Project lead** (name, division, group)  **Jos Cooper, IDM, NR** |
| **Other project contributors** (available from Project Lead’s end. This includes helping with understanding the data, being available for recurrent catch-up and other. Include names, availability and optionally areas of expertise)  Andrew McCluskey (Bayes and information theory)  James Durant (Python) |
| **Non-expert description of the problem, highlighting the role of ML in solving it, and why you think ML will be useful** (without technical terms or formulas (up to 300 words))  Neutron data is usually acquired for an arbitrary length of time, often decided by either historical (“one hour per angle always works”) or subjective (“the data looks good now”) factors. We aim to develop a statistics based tool, which can both predict model parameters, and estimate how much longer the data needs to be counted in order to reduce the uncertainty in those parameters to a desired level.  We wish to have this be able to run in near real time, allowing experiments to be stopped when they reach their uncertainty threshold, saving beamtime; or to tell the experimenters to carry on counting, saving experimental datasets.  This will be developed initially for reflectometry, but much of the methodology will be completely transferrable to most other techniques used at ISIS, simply requiring a different “front end” for parameter estimation / model selection.  Our approach will require a combination of traditional methods, Bayesian model selection, nested sampling, and parameter estimation, as well as neural networks, for calculations which would take too long to fully compute otherwise. |
| **Project summary** (expand on above and include technical terms as needed (up to 500 words))  The goal of this project requires several steps to be performed in a short space of time. Most of the steps constitute the normal workflow which the data analysis would take, though these would be automated and sped up. The last step, which is not currently performed, is the projection of the uncertainties to a dataset with arbitrarily increased counting statistics. This then leaves the experimenter in an informed position, both about their current data, as well as the state of the data after more neutrons have been counted.  To briefly summarise the work flow:   * A dataset is generated by the instrument, in the case of reflectivity this is usually in the form of a set of points of reflectivity as a function of momentum transfer, with every point having an error bar in reflectivity based on counting statistics. * The first step in our process is to assign a model to the reflectivity curve. This can be done either naively using something like a neural network which has been trained to recognise reflectivity curves, or through a user defined input. * The model parameters then need to be refined. There are many existing methods for parameter space searching and optimisation for reflectivity, such as genetic algorithms or simulated annealing. * Once the data and the model agree to within some tolerance, further analysis is possible. Using methods such as Markov chain Monte Carlo (MCMC), nested sampling, or bootstrapping it is possible to determine the uncertainty in (and if desired the correlations between) parameters.   All of these steps are possible and commonly performed, though the uncertainty evaluation is currently not always performed. Our project aims to chain together and automate each of these steps, while trying to make speed gains whenever possible, then go a step further to try to predict parameter uncertainties with increased counting statistics. All of this should occur in a short enough timescale that it will be possible to inform experiments which are running, in near real time.  Each step which is automated represents gains to the users in terms of utility, but the greatest gains are in the parameter uncertainty prediction. This will enable measurement times to be shortened, and is not specific to any technique. |
| **Description of the data that the ML methods are applied to** (Please provide as much information as possible, such as whether it is simulated and/or real, size of individual samples (such as in GB), overall size (number of samples), data format, how they are accessed and whether it has any missing or incomplete parts)  The data to be used here will be a combination of simulated data (an unlimited supply of this is possible), and real data which is able to be time sliced into chunks with varying counting statistics. This data is historic and it would be possible to obtain ~hundreds of such datasets, each of which could be split into counting time dependant subsets. All data is the end result of a Mantid processing chain and can be stored in any format, most likely HDF5 when concatenated, but commonly individually as ascii (X,Y,E). An individual dataset would be ~kb when stored as an asci. |
| **Describe any previous/current ML or Bayesian techniques that have been applied to this problem/data**  Over the last year and a half we (with the assistance of the SciML group) have developed a neural network based data pipeline for naive model selection allowing rapid fitting of a simple model to user data. This has been written up and we are in the process of responding to reviewer comments. My student this year is tidying up the code, and will begin working on this project when he is up to speed with it. We are collaborating with Andrew McCluskey on this, as he has a similar project student looking at model selection using Bayesian statistics. |
| **More broadly, what other related work has previously been done in this area and by whom**  A relatively large amount of work has been done on model selection and applications of Bayesian methods for many years. D. Sivia discusses the application of Bayes theory to neutron data in “Data Analysis: A Bayesian Tutorial” (2005), he also has many papers applying these techniques to reflectometry data. Almost all of the work has been done on data which has already been taken, with an aim to correctly parameterise the problem. Some work has recently been done on both experimental design and the information content of reflectivity curves by F. Heinrich (e.g. “Optimization of reflectometry experiments using information theory” (2019)). However very little seems to exist for *on stream* working. |
| **Potential wider impact of the project to ISIS and elsewhere**  The development of the end goal, e.g. a tool which is able to tell you how long to count, for a given level of uncertainty in some parameters would save large amounts of beamtime, both at ISIS and other facilities (x-ray as well as neutron). However, in addition to this, many of the steps towards this goal are valuable in their own right. For example, one of the outcomes might be a massively lower barrier to applying Bayesian methods to reflectivity data, making it routine, thus maximising the information output of each experiment. Additionally these tools would make optimising experimental design for, e.g. interfacial magnetism experiments, simpler, giving a good idea of what effects would be measureable in the finite counting time of an experiment. |
| **Potential start date and duration** (all projects will be reviewed after ~six months period and every six months thereafter. However, specify a start date for when you see you are ready to start the project, and what you expect the length of the project to be from that date)  This project has already started, so any additional effort would be able to start right away. It is the aim to have at least some conclusion in one year’s time, see milestones. |
| **Project deliverables** (what will be the deliverable: software, algorithms, ML models, reports etc)  The project will mainly deliver software, written in Python. We will make all work open source, including trained model weights for any neural networks.  The work will be published in a peer reviewed journal, and presented at a relevant conference (e.g. ICANS or similar). |
| **Project milestones** (Examples include (i) Data cleaning and data transformations to facilitate ML (~2 months); (ii) Creating xyz simulated data (1-3 months); (iii) Applying ML method abc to data (~1 month) (iv) Creating internal stage report summarising results (~1 months))   1. An implementation of on stream model selection and optimisation (i.e. close to real time). This can be in the form of naive neural networks, or with user input for data priors. (2 months for NN, 1 month for user inputs) 2. Uncertainty estimation for parameters derived for model using nested sampling or MCMC methods, ideally in close to real time as well. (1 month) 3. Implementation of parameter uncertainty prediction as a function of counting statistics, in close to real time. (3 months) 4. Processing stream to parse results between all preceding parts and optimisation for constant running on a beamline. (3 months) 5. Preparation of manuscript and code for publication. (2 months) 6. Future work is in applying some of these methods to other neutron/x-ray techniques. |
| **Any other information** |

Notes

1. The data must be from ISIS, or from other data source where it is clear that analysing these in some way has the potential to provide benefit to ISIS, or both. Further the data must be open access or will be turned in open access data (either after an embargo period or other mechanism)
2. The Project Lead will be requested to provide brief periodic updates on progress.
3. A successful project will primarily be resourced by the SciML group in SCD, and depending on ongoing recruitment etc. start dates of successful projects may vary.
4. The project needs to be well defined and should have clear outcomes. It can be longer-term project, but all projects will be reviewed after the initial six months, to assess their suitability for continuation, and with no guarantee of extension. Further, all projects continue to be reviewed and assessed on ~six months basis
5. The Steering Group will assess the forms. A presentation to the Steering Group may be required.
6. Any created data, software and results from the project must be made open, under an appropriate license (such as CC-by for data and BSD for software). Help with Data Management will be provided.