From Stars to Baht: Broadening the economic impact of astronomical data handling techniques in Thailand

Data management plan

We will ensure from the outset that protocols are in place to systematically manage all data associated with the project. By the nature of the research, the project will collect and produce diverse datasets. The raw data we will receive from our external partners will be in the form of digital values and strings in either ascii (.csv) or excel spreadsheets. One of the aims of this project is to make this data more searchable and accessible for our external partners by setting up relational and/or non-relational databases to contain this data. The data derived from our ML analysis can be stored more simply, in ascii files, for example.

The main risk associated with handling data from most of our external partners is their potentially commercially sensitive nature. To mitigate the risk of data loss during transfer from the partners to our servers, all data will be transferred in an encrypted state (whether by physical media such as USB stick or over the internet). Once on site at MFU, the raw data will immediately be backed-up on MFU's multiple backup servers. One of our external partners, Biophics, stands out in terms of data risk due to the sensitive medical data that they hold. Biophics is based within the Faculty of Tropical Medicine at Mahidol University, the top ranked University within Thailand to study medicine. As such, its members of staff are extremely well-versed in Thailand's legislation surrounding the sharing of medical records. They have assured us, including within their letter of support, that all data will be fully anonymised to prevent any possible identification of individuals.

Once the original raw data are backed-up, work will start on researching the best means to incorporate the data into a distributed non-relational database. The data management system we will use for this is *Hadoop*, which is widely used throughout the IT sector, thus ensuring the long-term readability of the data. Most of our data analysis will take place within the *Orange* data mining software environment, which is an open-source and under constant development. These steps ensure that the data collected and produced by the project will be readable for at least 10 years.

The main data that will be of use to other researchers are descriptions of the technologies we will develop to address the needs of the external partners. As with any research, our project will involve trial and error - testing different databasing and analysis techniques. The main publishable outcome of the work will be the results from these tests in terms of accuracy, reliability and speed of the database system and the ML-based analysis. By signing the Letter of Support, our external partners already have an understanding of this. Where possible (and after full consultation with the external partners), raw data will be made publicly available, together with any non-commercially sensitive derived data.

The outcome of our research – the completed database design and working model – will be made available by enacting the Release to Public policies of Mae Fah Luang University. We will package our data analysis algorithms in an *Orange* module as well as providing a self-contained software package complete with user interface to the external partner. Any non-commercially sensitive code will be shared publicly using the GitHub software sharing website.