

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

Case for support

1.1 Background

As outlined in our ODA statement, there is strong evidence that improving access to advanced data handling and analysis techniques is one of the most effective ways of increasing productivity in businesses and organisations. This is most sustainable when using home-grown talent to provide these services. As such, providing relevant, high quality training in these digital skills to Thai students and workers is a major element in helping Thailand graduate from a middle to an upper income economy.

Facilitated by STFC funding via the Newton and GCRF programs our team of educators and researchers are developing a highly effective data science training programme for Thai students. Within this programme, students work with external research organisations and businesses to gain first-hand experience of using data science to solve real-world problems. To date, this programme has been highly productive, as evidenced by:

- two students presenting peer-reviewed proceedings at three international conferences (one of whom was awarded an “Excellent Oral Presentation” award at the 10th International Conference on Machine Learning and Computing);
- the publication of five peer-reviewed journal articles or conference proceedings, one of which was recently awarded the “Best Conference Paper” award at the 2018 IEEE International Conference on Knowledge, Innovation and Invention;
- our successful ongoing collaboration with five external businesses and organisations;
- keen interest from four additional national businesses and organisations to work with us in solving their data handling and analysis needs.

By building on these successes and expanding our external partnerships to a national level, the next phase of our project – for which we now request funding – will establish Mae Fah Luang University (MFU, the Thai PI institute) as a hub for data science research and training within northern Thailand. This will enable future students, businesses and organisations within Thailand to benefit indefinitely from STFC’s current investment.

Our collaboration started in February, 2017, with a 12-month STFC Newton project followed by a further 24-month Newton grant. For the Newton project, our team works with Thai graduate students to research and develop Machine Learning (ML)-based algorithms to analyse the large amounts of astronomical data generated by the Gravitational-Wave Optical Transient Observatory (GOTO). GOTO surveys the full observable night sky every two weeks, delivering data on roughly seven million astronomical sources *every night*. The primary *research* goals of our Newton project are to: (a) develop a data warehouse that is capable of storing large amounts of data that is updated on a daily basis, yet is readily analysed using complex algorithms and (b) develop fully-automated ML algorithms to quickly and robustly categorise sources detected by GOTO. Our research in this area has been recognised as cutting edge within the field of data science, with our students winning awards for their presentations at two major international data science conferences (see attachments). With our Newton projects focussing on storing and analysing large amounts of astronomical data, the science, technology and expertise developed within our collaboration has originated from work associated with STFC’s core Science Programme.

1.2 Broadening the impact of our research to national businesses and organisations

During our Newton-funded project, it became evident to us that the technologies we are developing and the training we are providing would be of great use to a broad range of organisations and businesses within Thailand. With this in mind, we successfully applied for a STFC GCRF Foundation Award to research how we could adapt the technology and training we developed during our Newton project to increase the productivity of Thai businesses and organisations. This funding has been used to (i) hold two networking events between our students and five external partners (one in March, 2018, the next in September, 2018), (ii) purchase a data server onto which partners’ data has been uploaded, and (iii)

provide training and supervision to five teams of 5-7 students while they collaborate with each of our external partners. Despite being only half way through the 12-month project, our students are already being praised by our external partners for their achievements to date (see letter from Thanapiriya).

With our GCRF Foundation Award, we have provided training to approximately thirty MFU students, giving them the confidence and skills needed to use our in-house computer infrastructure to address the data-related needs of Thai businesses and organisations. During that “capacity-building” phase, we focussed on small-to-medium-sized partners with a regional presence to ensure we didn’t overstretch our teams during that early stage. The progress made and lessons learned over the past six months make us confident that we are capable of expanding our network of external partners to a national level, including some large enterprises (see the descriptions of the partners below). Doing so will give MFU the exposure needed to become self-sustaining in terms of attracting a variety of future external partners and establishing it as a hub for data science research and training within northern Thailand.

To support the expansion of our project to a national level, we request funds to host another series of networking events for a new cohort of students to establish collaborations with a set of new external partners, plus Thanapiriya – the largest of our first partners. We also request funding to support the supervision of these students throughout the collaboration with the external partners. All our external partners have already agreed to work with us throughout the 24-month project (see Letters of Support). Below, we provide a brief description of the external partners and their data needs:

- **Thai AirAsia** is part of the AirAsia group, the largest budget airline operating in South East Asia. They are renowned among the air industry for their high levels of aircraft utility. A barrier to achieving even higher levels of aircraft utility is the grounding of aircraft while waiting for the delivery of replacement parts. This waiting could be reduced if AirAsia were able to predict when parts were likely to require replacement. AirAsia engineers will share their flight data with the students working on this project, and work with them to adapt our technologies to better predict when parts will require replacement.
- **Thanapiriya** is a chain of supermarkets spread across various locations in Northern Thailand. The business has grown rapidly over the past decade, and are now looking to expand into more locations. A problem they face when opening new stores is how to predict customer demographics, which has knock-on effects on how to stock the new store. Their aim is to work with a small group of students and researchers to research effective means to characterize their customers on a branch-by-branch basis and predict appropriate levels of stock.
- **Deevana Hotel Group** is a group of eight hotels based in the south of Thailand. Deevana use promotions to attract guests to their hotels. At present they use a scatter-gun approach to advertise these promotions, whereas they wish to investigate whether more targeted campaigns based on customer need can deliver better results. Deevana will work with our students to research how machine learning – in particular clustering analysis – could identify different types of customers and test which promotions are most effective at attracting different customer types.
- **Biophics** is a Center of Excellence for biomedical and health informatics operating out of the Faculty of Tropical Medicine at Mahidol University, Bangkok. They specialise in the collection and analysis of data relating to the incidence and spread of infectious diseases within Thailand. Biophics wishes to work with our students to develop a ML-based algorithm to identify outliers in Tuberculosis (TB) patient records that may help to identify instances of drug resistance. This is a “rare event” problem similar to the one we have developed solutions to for analysing GOTO data.
- **OpenLandscape** is a provider of cloud computing infrastructure and services in Thailand. In order to grow, OpenLandscape needs to attract customers, but also be able to predict when those customers will place greatest load on their systems. OpenLandscape would like to work with our students to (a) analyse their customer database and usage logs to classify customers according to their usage type, and (b) analyse the usage logs of their hardware to predict those times when there is heaviest load on their systems.

Following the first twelve months of the project, we will hold another networking event to help attract more external partners to work with a new cohort of students. We have also budgetted for two Thai researchers to visit the UK twice during the project to help expand the project to include UK partners.

By working closely with external partners, our team of researchers are gaining deep insights into the skills shortages experienced by Thai businesses. This makes our team extremely well-suited to providing highly relevant training to broader sections of Thai society beyond those that are able to attend university, including those already in full-time employment. To capitalise on this opportunity, we request funding to develop a distance-learning course to teach the data-handling skills that are in high demand by Thai businesses. This will be achieved, in part, through a Massive Open Online Course (MOOC). We will consult our partners during the design process to ensure that the MOOC addresses the skills gaps they face and to ensure that the courses are accessible to current and potential employees. On successful completion of the course, participants will have the opportunity to attend one of four workshops that we will hold at MFU during the final seven months of the project. The aim of this workshop will be to enable participants to collaborate with members of our team to apply the technologies they have learned during the MOOC to solve either their own data-related problems, or one of a suite of pre-defined problems based on using our team's own data.

1.3 Description of work to be undertaken

Over the 24-month period of the grant, we will (in broadly chronological order; see Gantt Chart):

1. Host the first networking event in Chiang-Rai, Thailand (where most of the Thai co-Is are based), attended by all UK and Thai team members, students, and representatives from all our pre-identified external partners. At the event, our team will deliver a series of short (5-10 minute) presentations to highlight our data-led solutions developed during the first GCRF-funded project. Representatives from our external partners will be asked to describe their business/organisation and what they seek to achieve from the collaboration
2. Each partner will be assigned a staff researcher as a main contact, who will also lead the supervision of a group of final-year project students. The role of the primary contact is to coordinate fortnightly discussions with the external partner to provide progress updates and receive feedback. Within the first three months of the project, the students will visit and work with each external user for at least two weeks to ensure they fully understand the data and problem.
3. Around month four each team will transition to the implementation phase by focusing on adapting our technology and techniques to the needs of the external partners. During the fortnightly meetings the teams will demonstrate their solutions and encourage the client to suggest improvements.
4. During month seven, we will host another meeting between UK and Thai team members and representatives from all external partners. This will allow us to revisit the tasks accomplished in the first half, specify the limitations/difficulties identified during the past six months, and revise the working plan for the rest of the projects. Plans will be made for further on-site visits and internships. Around this time, we will also start work on developing the MOOC, following consultation with our external partners.
5. Our work will focus on delivering the agreed systems by adapting our own and further researching new technologies to meet our partners' needs. As this happens, the purpose of the fortnightly meetings will progressively shift toward feedback sessions, during which the primary contact will demonstrate our systems and allow the end user to suggest improvements.
6. By month ten, the team will deliver beta versions of our solutions. At this time, with the use of new virtual classroom at MFU, the team will train the end user on how to use the systems and collect any immediate feedback they may have. Following feedback, students will work on delivering the final products by month twelve.
7. We will hold a networking event in month 12 to discuss the outcomes, successes/drawbacks, impact-to-date, and future directions of the research and partnerships. This event will be attended by other potential partners to work with a new cohort of final-year students during months 13-24.

2.1 Management plan

Our management plan for this next phase will build upon the successes of and lessons learned during the first, Foundation Award-funded, phase of the project. Following the success of the initial networking event at MFU, we will hold the same again at the start of this second phase. Prior to the networking event, teams consisting of at least one member of staff and 2-3 final year project students will be assigned to each external partner. During the networking event, the staff and students will speak with the external partners to gain a deeper understanding of the partners' data and desired outcome of the project (i.e., the "problem" they want solving). Our choice of using teams of 2-3 final-year students is based on our experiences from the first phase, in which larger teams of less experienced students were involved in the project. Unfortunately, we found that this first approach led to some students feeling less engaged with the project than others. By contrast, the model of using smaller groups of final-year students we have employed in our Newton-funded project has led to high levels of student engagement, so we will follow that same set-up here.

Following the initial networking meeting, Dr. Boongoen will coordinate the fortnightly partner meetings. At the end of each of their fortnightly meetings, the primary contact will write a brief meeting summary which will be discussed with, agreed, and signed-off by the external partner. These summaries will be collated and shared among the whole team prior to our fortnightly team meetings/telecons. During our minuted team meetings/telecons we will decide what short term actions should be taken, and by whom, to reach the intermediate milestones and the desired goals of the external partners, and whether any of these need to be reassessed with the partners. By pooling our resources and focussing on a limited number of external partners we mitigate the risk that the needs of any one partner will go unmet. Finalising, during the second networking event, a broad design of what will be delivered for Beta testing avoids the risk of not delivering a coherent system for Beta testing. The design can still evolve after this stage, but this approach ensures any evolution will be managed.

In a departure from the model adopted during the first phase, after the second networking event the students will have the opportunity to spend a limited amount of time (up to 2 weeks) as an intern based at the external partners (to which all partners have already agreed). The goal of the internships is two-fold: firstly, it will constitute a period of daily communication between students and partners, enabling rapid advances in product development to be made according to the partners' requirements and, secondly, give the students first-hand experience of working in industry.

At any time during the project there is, of course, the risk that an external partner may be unsatisfied with the proposed design of the system. If that happens, it will be the responsibility of Drs. Boongoen and Mullaney (as PIs) to negotiate a solution. Should a viable solution not be found, it is feasible that a partner may leave the collaboration. This risk to the project is mitigated by having multiple partners, while the team would still have learned valuable lessons from the experience to carry-over into future partnerships.

2.2 Track record of applicants

Our multi-disciplinary team of researchers is made up of astronomers, computer and data scientists, computer hardware specialists, a Lecturer in Business Management and Marketing, and a Sales and Marketing Consultant (pro bono). The UK PI (Mullaney) is an astronomer with extensive experience of analysing data from large astronomical surveys. He was the UK PI of our Newton-funded and GCRF Foundation Award-funded projects. The Thai PI (Boongoen) is a computer scientist with specialist expertise in developing ML algorithms. Prior to joining MFU, Dr. Boongoen worked as a member of the Thai Royal Air Force working on Big Data Analytics with a particular emphasis on security applications. He has extensive experience in managing research projects, having PI'd six successful grants in the last five years. Dr. Iam-on is a data scientist with expertise in database design, data mining, and developing ML algorithms for automated data analysis. Together with Arwatchananukul, Dr. Iam-On will oversee the database design elements of the project as well as contributing her expertise on the automated analysis aspects. Drs. Kirimasthong and Sigpant's expertise are also in the development of ML-based algorithms, which they will bring to bear as team leaders on the project. Mr. Nobnop and Mr. Prommool are involved in developing MFU's MOOCS, and so will assist in this part of the project. Dr. Wongwatkit's interests focus on the use of new technologies to enhance learning, and so will be particularly relevant in the latter stages of the project when developing the MOOCs and training the employees of the partners to use the systems we develop. Dr. Nupap will bring her experience as a consultant to several Thai businesses to help liaise between the researchers and external partners. Dr. Kurubanjerdjit has published research papers in the field of bioinformatics and has helped to develop secure databases to store medical information. As such, she will be closely involved in the Biophics project. Finally, Mr. Thongjua has over 20 years experience in Sales and Marketing, and has agreed to work with us on a pro bono basis to use his business network to identify and attract future partners.