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 the Thai economy to graduate from middle to upper income.

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 a "Best Presentation" prize);

- the publication of five peer-reviewed journal articles or conference proceedings;

- successfull collaborations with five external business at the regional level;

- keen interest from four national businesses to work with us in solving their data handling and analysis needs.

By building on these sucesses and expanding our external partnerships to a national level, the next phase of our project 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 to benefit indefinitely from STFC's investment.

Our collaboration started in February, 2017, with a 12-month Newton-funded project. For this project, our team worked with Thai graduate students 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*. This large, constantly-updated dataset gave the students the opportunity to develop the skills and techniques needed to handle and analyse the kinds of Big Data generated by many of today's industries. Following this first phase of the project, we were awarded a further two years' of Newton funding from February 2018 to continue our work with GOTO.

The primary *research* goals of our Newton project are to: (a) develop a database that is capable of storing large amounts of data that is updated on a daily basis and (b) develop fully-automated Machine-Learning (ML) algorithms capable of quickly and robustly categorising sources detected by GOTO. Our research has led to some novel solutions. For the database component are researching whether distributed non-relational databases are an effective means of storing and analysing large astronomical datasets. For the ML component, we are developing effective new methods to train algorithms to identify and classify very rare transient events among tens of thousands of astronomical sources. Since both goals are to address challenges associated with archiving and analysing large amounts of astronomical data, the science, technology and expertise involved in the proposed GCRF project 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 the first phase of our Newton-funded project, it became clear 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 to STFC's 2017 call for

GCRF Foundation funding 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. Through this project, our researchers and students have gained experience of working with five such external partners, and have been praised for their achievements so far (see support letter from Thanapiriya).

Our GCRF-funded project has demonstrated our ability to successfully collaborate with external organisations. Presently, our external partners have a regional, as opposed to national or international, presence. While working with these small-to-medium sized businesses and organisations has provided our team with invaluable experience of liasing with external partners, our long-term goal is to also work with partners with a national presence within Thailand. Doing so will provide MFU with the exposure needed to become self-sustaining in terms of attracting a variety of future external partners.

For this capacity-building project, we request funds to expand upon our GCRF Foundation Award-funded project to work with four further external partners, each of which have a more national presence than the partners involved with the our Foundation project. We also request funding to continue our successful collaboration with Thanapiriya – the largest of our first partners. All our external partners have already agreed to work with us throughout the 24-month project (see Letters of Support). At the end of the first 12 months, however, we will hold a networking event to attract other external partners. Below, we provide a brief description of the external partners and their data needs:

- **Thai AirAsia** is part of the AirAsia group, which is 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 they wait for the delivery of spare 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 so that we can adapt the technologies we have developed to better predict when parts will require replacement.
- **Thanapiriya** is a chain of supermarkets spread across various locations in Northern Thailand. They have expanded 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 the types of customer (sex, age, dispoable income, etc.) that are likely to visit the new store, 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 help attract guests to their hotels. At present they use a scatter-gun approach to advertise these promotions, whereas they would like to investigate whether more targeted campaigns based on customer need can deliver better results. Deevana wishes for our students to research how machine learning – in particular clustering analysis – could be used to better 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 would like 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 a group of our students to (a) analyse their customer database and usage logs to classify customers according to their usage type, which they will use to target potential customers with bespoke promotions and; (b) analyse the usage logs of their hardware to predict those times when there is heaviest load on their systems.

1.3 Description of work to be undertaken
1.3.1 Working with external partners
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 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 meetings 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 2 weeks to ensure they fully understand the data and problem.

3. Around month 4 each team's will transition to the implementation phase by increasingly 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 7, 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.

5. Our work will focus on delivering the agreed systems by adapting our own and further researching new technnologies 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.

1.3.2 Development of a Thai data science MOOC
With our partnerships with external organisations and businesses, we are able to provide world-class training in data science to Thai students enrolled at MFU, who can then take these skills to the workplace. However, this neglects the tens of thousands of other Thai citizens who could benefit from these skills, not least those already in full-time employment who are unable to take time off to attend University. The development in recent years of the Massive Open Online Course (MOOC) enables higher education institutes to greatly increase the impact of their teaching.

During the final 18 months of the funded project, our team will draw upon our broader teaching experience to develop a series of data science MOOCs. Crucially, we will use our unique collaborations with our external partners when developing these course. We will consult our partners during the design process to ensure that the course both meet the skills gaps that they have identified within the Thai workforce, while also ensuring 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 that final 7 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 using our team's own data.

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 experiences 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, a computer hardware specialist, and a Lecturer in Business Management and Marketing. 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 project, responsible for ensuring that the research meets the needs of the GOTO collaboration. The Thai PI (Boongoen) is a computer scientist with specialist expertise in developing ML algorithms. He has extensive experience in managing research projects, having PI'd four successful grants in the last four years. Dr. Iam-on is data scientist with expertise in database design, data mining, and developing ML algorithms for automated data analysis. She will oversee the database design elements of the project as well as contributing her expertise on the automated analysis aspect. Dr. Eungwanichayapant's background is in high energy astrophysics, with particular expertise in developing unsupervised ML algorithms to analyse data from Gamma Ray telescopes. Drs. Sawangwit and Awiphan are astronomers based in Thailand. Their research expertise lies in analysing large astronomical datasets and time varying data, and thus are very relevant to the project. Mr. Vattayasak will be the team's computer hardware expert: his specialism is in setting up

distributed networks of computers to host large, distributed databases. Finally, Ms. Noichankgkid is a Lecturer in Business Management and Marketing, with extensive management and finance experience prior to and during her academic career. Her experience will prove invaluable when liaising with and establishing the needs of our business partners.