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

Case for support

We request funds to extend our previous GCRF-funded project

research how the data archiving and analysis techniques we have developed as part of our previous STFC/Newton-funded project can be best adapted to address the needs of Thai businesses and organisations. To achieve this research aim, our multi-disciplinary team of Thai and UK scientists will spend 12 months working closely with a group of five pre-identified businesses and organisations based in Northern Thailand that have specific data archiving and analysis needs.

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 (details below) 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 MFU 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 present investment.

The collaboration between the PIs (based at Sheffield and MFU) started in February, 2017, with a 12-month Newton-funded project in which Thai data scientists in our team worked with Thai graduate students to 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 involved in that first phase 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 ongoing 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 we are developing a hybrid system that combines the structure of a relational database (such at that used by SDSS) with the flexibility of a non-relational database (which are popular within tech. industries). For the ML component, we are developing a new two-stage categorisation algorithm which uses unsupervised ML algorithms to quickly remove obvious artefacts from our data before passing

more ambiguous cases to a supervised ML algorithm. 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 GCRF funding call to fund research into how we could adapt the technology and training we have developed 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 work funded through the first round of GCRF funding has demonstrated our ability to successfully collaborate with external organisations. Our current external partners have a regional, as opposed to national or international, presence. While working with such small-to-medium size partners has proven beneficial in terms of our developing good working relationships, our ultimate 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 broad variety of external partners to work with our team and their students.

For this capacity-building project, we request funds to expand upon our first GCRF-funded project to work with four further external partners, each of which have a more national presence than the partners involved with the first phase. We also request funding to continue our collaboration with Thanapiriya. 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 (i.e., maximising the time a plane is in the air, and thus generating revenue). A barrier to achieving even higher levels of aircraft utility is the grounding of aircraft while they wait for the delivery of spare parts from various warehouses. This waiting could be reduced if AirAsia were able to predict when parts were likely to require replacement, and then act upon this when the aircraft was located at an airport with a parts warehouse. AirAsia engineers will share their flight data with our team so that we can adapt the techniques we have developed to identify rare sources in astronomy to better predict rare instances where parts require replacement. To assist with this development, Thai AirAsia have offered to host the students working on the project on a work-experience placement.

- **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. Thanapiriya was one of the partners involved in our first GCRF project and would like to continue to work with us. They would like to work with small group of students and researchers to investigate how we can adapt the work conducted during the first phase of our GCRF project 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. This region is a popular tourist destination, resulting in a particularly competitive hotel market and corresponding thin margins. Deevana has found that various promotions (e.g., reduced tariffs, free services/experiences, etc.) help boost traffic to their hotels, resulting in fewer empty rooms. At present, however, they use a scatter-gun approach to advertise these promotions, whereas more

targeted campaigns based on customer need may 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. In return, our staff and students will gain experience in a developing technology solutions for one of Thailand's most important economic sectors.

- **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 on developing 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. Being part of Mahidol University – Thailand's top-ranked medical school – Biophics is highly familiar with the anonymisation of patient data to the levels required by Thai law, and no patient data will be shared outside Thailand.

- **OpenLandscape** is a provider of cloud computing infrastructure in Thailand. As in western economies, it is predicted that many Thai companies will switch to rented cloud-based infrastructure from in-house computer hardware. In order to grow, OpenLandscape needs to attract more customers, but also be able to predict when those customers will place greatest load on their systems. To help achieve these aims, OpenLandscape would like to work with a group of 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 the usage logs of their hardware to predict when those times when there is heaviest load on their systems. Working with OpenLandscape will provide the students involved in the project with experience in one of Thailand's fastest growing industries.

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 and representatives from all our pre-identified external partners. During the event, our team will deliver a series of short (5-10 minute) presentations to highlight our data-led solutions developed during the previous batch of external users, and available expertise within the team. Representatives from our external partners will also be asked give short presentations to describe their business/organisation and what they seek to achieve from the collaboration

2. Each problem will be assigned a staff researcher as a main contact, who (and other staffs) co-supervise one student group (senior-project group). The role of the primary contact is to have fortnightly meetings with the external partner to present results and receive feedback. After each meeting, the primary contact will report back to the rest of the team to ensure a collaborative effort is maintained. Having identified possible data, the students will visit and work with each external user for 2 weeks to ensure the quality of collected/preprocessed data, with a chance to communicate directly with the user.

3. By month 4 the teams work will move to the implementation phase by increasingly focusing on adapting our technology and techniques to the needs of the users. During the fortnightly meetings the teams will demonstrate their solutions and allow 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 found during the past six months, and revise the working plan for the rest of the projects. This will include plans for further on-site visits to provide user training, and also the preparation for internships where students can help users to solve more problems.

5. Our work will focus on delivering the agreed systems by adapting our own and further researching new techniques 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. At the start of month ten, we will deliver "Beta versions" of the systems we have developed. We will train the external partners on how to use the systems and collect any immediate feedback they may have. The systems will then go through a two month-long Beta testing phase by the external partners, who will then report their experience back to the team. During months eleven and twelve we will address any feedback from the Beta testing phase to deliver the final product.

7. We will host a final networking event to discuss the outcomes, successes/drawbacks, impact-to-date, and future directions of the research and partnerships. This event will coincide with an outreach event to highlight to the public and invited representatives from other businesses/organisations the economic benefits of data science. Part of the goal of this event is to identify future partnerships.

1.4 Maximising the impact of the project

The steps we have taken to ensure that the project has maximal impact are described fully in our Pathways to Impact statement. Briefly, our choice of working very closely with a limited number of external partners is motivated by our desire to focus our efforts to ensure maximal impact in terms of our partners' productivity *and* building our team's capacity for working with external partners. This experience will establish protocols and ways of working to be taken forward to future partnerships.

In addition to the impact within Thailand, the project promises secondary benefits to UK research. With STFC's involvement in forthcoming data-intensive projects such as the LSST and SKA, it is vitally important that UK astronomers gain exposure to advanced data handling and analysis techniques. Further, the proposed work will give the team experience of working with data-intensive businesses – valuable preparation for research supported by the UK Government's Industrial Strategy.

2.1 Management plan

The success of each external partnership will be a team effort. During and following the first networking event we will set out a series of milestones to reach in order to achieve each partner's desired outcome. After consulting with the rest of the team, Drs. Boongoen and Mullaney will assign primary contacts by matching team members' skills and experience to the needs of the external partners. 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. There is the risk that an external partner may be unsatisfied with the proposed design of the Beta 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.