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

Case for support

We request funds to 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 handling and analysis needs.

1.1 Background: Using astronomy data to train Thai data scientists

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. However, Thai data scientists and their students typically lack access to very large amounts of digital data, restricting their training in "Big Data" science that is so crucial for the development of modern economies. To address this problem, we successfully applied for Newton funding (12-month project starting February, 2017) give the Thai data scientists in our team research experience of working with 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 has given our research team 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.

The primary *research* goals of our Newton project were 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 local businesses and organisations

At present, the impact of the our Newton project is limited to the people directly involved in the project (i.e., Thai and UK astronomers, Thai data scientists and their students), whereas our long-term ambition has always been to broaden the impact of our research to develop the wider Thai economy. In this regard the goal of this GCRF project is to use the technology and skills we have developed to increase the productivity of Thai businesses and organisations. To do this effectively, our team needs to first gain experience of working with such external parties to research how best to adapt our technologies and expertise to their data handling and analysis needs.

For this partnership-building project, we request funds to research how our data archiving and analysis techniques can be adapted to meet the needs of five pre-identified Thai businesses and organisations (hereafter, "external partners"; see Table 1). All our external partners are based in areas local to the Thai co-Is and have already agreed to work with us throughout the 12-month project (see Letters of Support). At this "capacity building" stage, we prefer to limit the number of external partners as it ensures that all parties have a thorough understanding of the goals and expectations of the project from the outset. In addition to our research goals, our aim is to learn lessons and define commonalities of approach that can be applied when we expand the project to work with larger numbers of external partners. Our ambition is to apply for follow-on GCRF funding to support this expansion.

Partner	*Data handling/analysis needs*
Thanapiriya plc	A food retail business that wishes to take multiple factors into account to predict optimum stock volumes. This is a categorisation task (the same type faced with GOTO data) in terms of increase/decrease/no-change of product demand.
TAPP Auto	A car sales business seeking to develop a system that can predict the depreciation curve of a given vehicle given multiple input factors. This is a regression and "missing data" problem, as commonly found in astronomy data analysis.
Pibulsongkram Raj. University	The Academic Resources Office wishes to identify the optimum online resources to meet the needs of different types of users. This can be achieved through ML-based Personalised Service Provision as often used by streaming services.
MFU	The Student Office seeks a system to identify the causes of the high student dropout rates in Thai Universities. This is, in part, a categorisation problem, since students can be grouped according to different drop-out factors.
M-Store	A retail complex based on MFU's campus seeks to increase footfall by targeting promotions to specific groups. A ML-based clustering analysis will help to identify different categories according to customer information.

Table 1: *Our external partners and a brief description of their data handling and analysis needs.*

1.3 Description of work to be undertaken

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

1. Have the Thai staff and students currently working on the project visit the UK. The purpose of this visit is provide the students with first-hand experience of presenting to and communicating with their first external partner – the GOTO collaboration.
2. Host a networking event in Chiang-Rai, Thailand, where most of the Thai co-Is and external partners are based. All team members and representatives from all external partners will attend. At the event, our team will deliver a series of short presentations to highlight our skills and technology, using GOTO as a case study. Representatives from our external partners will describe their businesses/organisations, the data they hold, and agree with the team their desired outcome from the project, together with an estimate of its impact.
3. Each external partner will be assigned at least one primary team member contact (this could be a postgraduate student, in which case a staff researcher will act as a secondary contact). The primary contact will attend on-site fortnightly meetings with the external partner to get to fully understand their data and analysis needs. Such close interaction is vital in order to discriminate between informative vs. non-informative data. After each meeting, the primary contact will report back to the rest of the team to ensure a collaborative effort is maintained.
4. We will host a second networking event in month 5. By this stage, the team will have a thorough understanding of the data and needs of each external partner, enabling the team to decide on broad design of what we will deliver to each partner. This design will be discussed and agreed-upon by the external partners during the networking event.
5. Our work will focus on delivering the agreed systems by adapting our own and researching new technniques 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 the 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 meet in order to reach each partner's desired outcome. After consultating 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 fornightly 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, in month 5, a broad design of what will be delivered for Beta testing avoids the risk of not delivering a coherent system. 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 Lecturer in Business Management and Marketing. The UK PI (Mullaney) is an astronomer with extensive experience of analysing data from large astronomical surveys. He is 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 offering 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.