

**UK – Thailand**

**Capacity Building in Software and Hardware Infrastructures and Data Handling through Astronomy 2017**

**Case for Support**

This template should be completed using: Arial (or an equivalent) and a minimum font size of 11. A minimum of single line spacing and standard character spacing must be used. Margins must not be less than 2cm and the document must stay within the page lengths specified for each section.

1. **General Information**

**Project Title** *[up to 150 characters]*

|  |
| --- |
| Using astronomy surveys to train Thai researchers in handling Big Data |

**Theme**

*Please identify which themes your proposal covers:*

* Mechatronics/telescope control
* VLBI engineering, technology and research
* Data handling
* Outreach to support STEM education programme in schools

All projects should include some element of outreach activity.

|  |
| --- |
| Data handling |

1. **Previous track record of applicants and links between proposed partners**

*Please summarise how the UK and Thai partners will work together and any previous interactions or experience of working internationally and collaboratively. Please take into account track records of the applicants and institutions and provide details of any facilities that are required to undertake the project outside of the host institutions. [Maximum 1 side of A4]*

|  |
| --- |
| The Thai and UK researchers will collaborate to broaden and build upon the success of our previous Newton-funded project: “Using astronomy surveys to train Thai researchers in Big Data analysis”. During that first phase of the project, funded for 12 months from February 2017 (hereafter, Phase 1), our multi-disciplinary team has collaborated closely to successfully train Thai graduate students in advanced data handling techniques, with a specific focus on machine learning (ML) and database management (DM). This has only been achievable by combining the Thai data scientists’ knowledge of data handling techniques with the UK and Thai astronomers’ access to large datasets, and, crucially, their understanding of the data held therein.Through the use of instant messaging, teleconferencing, and face-to-face meetings we have shared our knowledge and experience throughout the team to foster a highly effective training environment.  The data scientists involved in the project have a strong track record in developing ML algorithms, including automated image analysis, (Boongoen, Eungwanichayapant, Iam-On, Uttuma) and setting-up and maintaining DM systems (Boongoen, Iam-On). The UK and Thai astronomers have track records in using non-ML techniques to analyse large astronomical datasets and are all involved in the Gravitational-wave Optical Transient Observatory (GOTO) project (see section 3 for a description of GOTO), which is the source of the large astronomical datasets that is being used throughout the project. Importantly, through our collaborative work during Phase 1 project, all data scientists and astronomers involved in the project now also have a good working knowledge of *each other’s* areas of expertise (i.e., the astronomers have learned from the data scientists, and vice versa). This cross-disciplinary knowledge make our team especially well-suited to exploiting astronomical datasets to train others in advanced data handling techniques, particularly those which a background in astronomy or other physical sciences.  We now request funds to enable us to build upon and capitalise on the wealth of experience we have gained during Phase 1 of our project. Over 24 months beginning January 2018 (hereafter Phase 2) we will use the requested funds to: (a) disseminate the knowledge and skills we have acquired to date through two 5-day practical workshops aimed at research staff and students based at NARIT and other Thai research institutes, (b) continue our successful graduate training programme in which students gain experience in advanced data-handling through cutting-edge research projects, and (c) establish a GOTO data centre based in NARIT which, as well as being an important research asset in itself, will be used as a training aid for NARIT staff and students both during the project and beyond.  For (a), all staff and students involved in the project will contribute their respective expertise to disseminate practical skills in DM and the basics of ML to around 50 trainees via a series of lectures, practical sessions, and homeworks held over five days during July 2018 and 2019. For the practical sessions and homeworks, we will preconfigure cloud-based systems into which the trainees will be able to login in order to gain hands-on experience. To meet this requirement economically, a public cloud service based on Amazon AWS (especially its education version) can be exploited for this training, thereby avoiding having to invest in expensive hardware systems and removing any security concerns that may arise if they were to log into NARIT’s servers. For (b), following the success of Phase 1, the data scientists and astronomers will jointly supervise four graduate students (3 MSc, 1 PhD) in advancing the research undertaken in Phase 1 (see Section 3) by (i) applying what we have learned from simulated data to real GOTO data, (ii) researching pixel-based deep-learning analysis, and (iii) researching how well the DM systems explored in Phase 1 scale to the large data rates provided by GOTO. Finally, for (c), a Thai data-science student will visit UK for a 3-month visit to work alongside GOTO scientists in setting-up a UK data centre. He will then take the knowledge gained from that experience back to Thailand to help set-up a mirror GOTO database in NARIT. |

1. **Official Development Assistance (ODA) compliance**

*Please provide a statement and* ***evidence explaining******how*** *your proposed research is compliant with Official Development Assistance (ODA) guidelines. Proposals must contribute towards the economic development and welfare of Thailand. For more information on ODA please refer to the* [*http://www.newtonfund.ac.uk/about/what-is-oda/*](http://www.newtonfund.ac.uk/about/what-is-oda/) *and* [*RCUK Newton Fund Guidance*](http://www.rcuk.ac.uk/documents/international/odaguidancercukspecific-pdf)*. Your ODA compliance will be assessed from this statement so please ensure you consider this in detail.* ***If your proposal is not considered ODA compliant it will be rejected.***

*[Maximum 1 A4 page]*

|  |
| --- |
| As an upper middle-income country, Thailand has already successfully tackled many of the greatest problems of developing countries, such as basic infrastructure development. Instead, the primary economic challenge that Thailand now faces is to develop into a high-income economy. The Thai government has recognised that a key means of achieving this is by training a highly skilled workforce able to compete internationally in high-value-added sectors that rely heavily on innovation (<http://bit.ly/2dWCZ01>), such as those listed in section one. Today, many of these sectors involve the collection of large amounts of digital data, whether in the form of customer information, patient medical records, information of crop growth, logistical information about a production line or distribution network, stock prices, etc. Many of these sectors have been earmarked by the Thai Board of Investment (BOI) as eligible for investment promotion (http://bit.ly/2djPB06). To successfully develop within these economically important sectors, it is vital that the Thai economy has access to home-grown talent trained in handling large amounts of digital data. The main objective of our project is to increase the skill level and experience of Thai workers and students in this important area. It thus satisfies the ODA’s guidelines of “promoting the economic development and welfare of developing countries as its main objective” and providing “educational services” [points 3 & 6 of “Official Development Assistance – RCUK Newton Fund Guidance”].  As the first phase of our project has already demonstrated, the students that are the focus of our training package become experts in advanced data handling, capable of teaching others such skills in their chosen post-graduation sectors. Indeed, our current graduate students will all teach at the practical workshops planned for this second phase. In this way, the project represents sustainable development. The short-term success of the project will be measured by the destination of the students on completion of the project and whether the learned skills are being successfully applied to develop the Thai economy. This information will be acquired by contacting the students six months after completion of the project. It will then be compiled by the partners and fed back to NARIT and STFC. The project will be considered a success if the students are using their acquired data handling skills in their chosen areas of work, and especially if they are teaching others these skills. In this respect, the project will investigate whether the training is effective in addressing Thailand’s economic development problems, thereby satisfying the ODA’s guideline of “researching the problems of developing countries” [point 5 of “Official Development Assistance – RCUK Newton Fund Guidance”].  The training our project provides will be delivered by the UK and Thai partners via official Thai agencies: NARIT and Mae Fah Luang. It thereby satisfies the ODA’s guideline of being “provided by official agencies, including state and local governments, or by their executive agencies” [point 3 of “Official Development Assistance – RCUK Newton Fund Guidance”]. |

1. **Detailed Research Information**
2. **Current landscape**

*Describe how the current priorities and challenges in Thailand and the UK will be addressed through this project demonstrating knowledge and understanding of past and current work in the subject area*

*[Maximum 1 side of A4].*

|  |
| --- |
| The primary objective of our project is to help address the challenge of developing Thailand from an upper-middle to a high-income economy through training in high-level data-handling skills. A secondary objective is the development of data handling systems to organise and analyse large amounts of astronomical survey data obtained by the Gravitational-wave Optical Transient Observatory (GOTO) on La Palma, Spain. GOTO is a major new observatory, funded by the UK Universities of Warwick, Sheffield, and Leicester, Monash University (Australia), NARIT (Thailand) and the Observatory of Armagh (UK). GOTO surveys the entire northern sky roughly every week, producing vast amounts of data (~250 gigabytes per night) in the process. It was commissioned in 2017, and so is already generating data for the project. In this subsection, we highlight how our team has worked our goals during Phase 1 of the project, while in the following subsection we describe how we will build upon this work during Phase 2.  During Phase 1 our team has provided training-through-research in both ML and DM. For the ML component, our focus has been to automatically discriminate between false and true positive in “difference imaging”, whereby two images of the same patch of sky taken at different times are subtracted from each other to reveal those sources that have changed in brightness. False positives arise due to effects such as differences in the point spread function (PSF) between the two images and imperfect image alignment. They typically outnumber true positives by many hundreds to one. Such “unbalanced” data is problematic for basic ML algorithms as they struggle to find the defining characteristics of the rare true sources, leading to low success rates (i.e., identifying <10% of true positives). Unbalanced ML problems arise in many economically important sectors such as, for example, in detecting rare instances of fraud amongst the many millions of legal daily transactions. To tackle the problem of unbalanced data in GOTO difference imaging two of our graduate students has been researching techniques to artificially boost the information available from the true positives and, in doing so, is obtaining significantly higher success rates (i.e., identifying >90% of true positives). A description of this work has recently been accepted in the peer-reviewed proceedings of the 10th International Conference on Machine Learning and Computing, at which the student has also been selected to give an oral presentation. To date, however, this research has been based on simulated data, so a key goal of Phase 2 is to research the applicability of our techniques to real GOTO data.  For the DM component, another of our graduate students has been researching the suitability of different data management systems (DMS) to store GOTO data. In addition to setting up a relational database system (as traditionally used by previous astronomy surveys such as SDSS), the student has also set up our own non-relational database based on the Hadoop Distributed File System (HDFS). HDFS is used extensively within the tech industries for its flexibility, scalability and speed advantages over relational databases. As such, this experience provides our trainees with skills that are directly transferrable to many economically important sectors. There are clear *potential* benefits for developing a full HDFS system for GOTO, not least its scalability and more rapid data access for advanced ML-based analytics (thereby marrying both research components of the project). However, we are unaware of any other implementation of HDFS by astronomy surveys, so it is vital that this be tested alongside more traditional relational DMSs. In Phase 1 our student has learned how to set up a PostgreSQL relational DMS and a small HDFS system, but has yet to test how either scale to the size and rate of data provided by GOTO. |

1. **Description of proposed work**

*Please describe the research and partnership activities proposed. Please also identify any risks and mitigation strategies.*

*[Maximum 2 sides of A4]*

|  |
| --- |
| **1. Project synopsis**  The goal of our project is to train Thai students and staff based at NARIT and other Thai research institutes advanced data handling techniques. Our Phase 1 work has demonstrated the effectiveness of using large astronomical datasets from GOTO as a training resource. In Phase 2 we will build upon and capitalise on this experience by establishing a GOTO data centre accessible by staff and students based at NARIT and other Thai research institutes. In addition to data storage, this facility will run ML algorithms to automate the analysis of data held within the data centre, potentially feeding through to NARIT’s TST, TNO, and Regional Observatories for (semi-) automated follow-up observations. The establishment alone of this data centre will provide a wealth of training opportunities for Thai researchers and students. Once established, this facility will be an important research asset for NARIT scientists and a valuable “Big Data” training resource for the broader Thai research community beyond the lifetime of the funding.   1. **Achieving the project goals**   To maximise the impact of our training, our project will consist of two main components:   * two week-long practical workshops to be held in June/July of each year of funding (i.e., 2018 and 2019); * a series of 18-month research projects during which four Thai students (3 MSc, 1 PhD) will continue to research optimal techniques to organise and analyse large astronomical datasets.   The practical workshops will be attended by up to 50 staff and students based at various institutes, including NARIT. Part of their goal will be to teach the new cohort of graduate students the science background to the project and technical skills that will help them get started with their research projects. After their first year of research, these same students will contribute to the second practical workshop by sharing the knowledge and skills they have acquired over the past 12 months of research. Below, we describe in detail what will be covered in the practical workshops and the focus of the research projects. First, however, we provide some context with a brief description of GOTO and the challenges the research component will attempt to overcome.  *2.1 The GOTO project*  The primary science goal of GOTO is to identify transient optical sources that are the counterparts of gravitational-wave (GW) sources detected with the Laser Interferometer Gravitational-wave Observatory (LIGO). It can achieve this by rapidly scanning the large (~100 sq. deg) error circle of the GW detection as soon as it receives a trigger from LIGO. When not following-up GW triggers, GOTO conducts an all-sky survey. This survey mode is a critical aspect of the GOTO project, as it provides an up-to-date reference map against which the post-trigger observations can be compared to identify any transient sources that may be associated with the GW event. GOTO was commissioned earlier this year and, having recently undergone a hardware upgrade, is currently delivering data.  *2.2 GOTO’s Data Handling Challenge*  On a typical night of observations, GOTO will observe between five and ten million sources (predominantly stars and galaxies, but also asteroids, satellites, etc). All these sources must be detected, measured, categorised and the resulting information recorded. The detection and measurement stages are solved problems in astronomy, even for such large data rates as GOTO. By contrast, categorizing such large numbers of sources on short timescales presents a significant challenge to astronomers. With manual classification an impossibility on these scales, we must instead turn to ML-based classification algorithms. Further, while storing large amounts of data is now feasible in terms of cost, organising this data in such a way that it is easily accessible on sufficiently short timescales for automatic analysis represents another significant unsolved challenge for GOTO and other high data rate facilities, including those used in industry.  *2.3 The practical workshops*  We will hold two week-long workshops during the 24-month period of the funding; the first in July 2018, the second 12 months later. They will be held in Chiang Mai and will be open to up to 50 attendees from various institutes, but with a particular emphasis on training NARIT staff and students in data handling. The focus of these workshops will be on disseminating practical skills that our team have acquired over the previous ~18 months of the project. There will be a particular emphasis on database management, covering topics including an introduction to relational and non-relational database, how to set up a cloud-based relational (i.e., SQL-based) database, and how to ingest data into and query these types of database. We will also cover the basic principles of ML and introduce attendees to “off-the-shelf” machine learning packages such as Python’s SciLearn, Matlab’s ML toolbox and Google’s TensorFlow. Teaching will be via lectures, hands-on experience (i.e., logging into cloud-based servers and working through pre-assigned tasks) and homeworks and will be conducted by staff team members and, crucially, graduate students from the first phase of funding. Where possible, teaching will be conducted in Thai to ensure maximal inclusivity. By the end of each workshop, attendees will be able to set up a relational database of arbitrary size and have a foundation in ML basics with which to build upon to suit their own analysis needs. All material, including lectures, will be made openly available online and staff and students will offer post-workshop assistance to attendees.  *2.4 The research projects*  To continue to train students in advanced data handling techniques (that they can then disseminate to others), it is vital that we continue our highly successful postgraduate research projects. One of our current MSc students has been researching the suitability of different DMS for the GOTO project. In the next round of funding, this student will spend three months (April-June, 2018) based in the UK applying his knowledge to working with GOTO scientists in setting-up the GOTO database. He will then take the additional skills acquired during this experience back to Thailand to work with NARIT staff in setting up a mirror GOTO database in Thailand. He will also continue to research the benefits and pitfalls of different DMS as the GOTO project’s needs evolve to include greater amounts of more diverse data.  For the ML aspect of our project, our current MSc student will build upon the research she has carried out during the first phase by applying what we have learned regarding automated false-positive rejection from simulated data to real GOTO data. To be truly successful, GOTO must also be able classify different types of source (e.g., star, galaxy, supernova, neutron-neutron star merger) in order to prioritise follow-up observations. For this, we will work with our current ML MSc student and an additional MSc student in using both simulated and real GOTO data to research ML-based techniques to perform source classification. Following the success of our true/false positive work, we will start by exploring feature-based analysis, in which the input to the ML algorithms are measured features of the sources (e.g., size, brightness, etc). However, we will also explore the possibility of pixel-based analysis, in which the algorithm considers the imaging data directly, more akin to human classification. |

1. **Human Participation**

*Please provide the following information* ***if your project involves any kind of human participation*** *(please note that this is a requirement of the call). Any human participation must abide by UK and Thailand standards whether it is taking place in the UK or Thailand. Failure to complete this section may result in your proposal being rejected. If the project does not involve humans, please write ‘Non applicable’:*

* *Please indicate* ***where*** *the recruitment of the human participants/ samples/ tissue will take place and the appropriate agreements.*
* ***Please identify any ethical or health and safety issues arising from any involvement of people, human samples or personal data in the research proposal. Please give details of how these will be addressed and any specific risks mitigated.***
* *Please explain how the proposed research will be carried out to a high ethical standard* ***and how the research will abide by relevant legal requirements in the UK and Thailand.***
* *Please indicate the ethical approvals and* ***research governance arrangements that*** *will be sought/ have been sought and will be in place ahead of starting the grant in both the* ***UK and Thailand****.**(This may include arrangements for supporting and providing expert ethics advice to researchers, should unanticipated ethics issues arise, throughout the lifecycle of the grant.)*
* *If you’re using human samples/tissue please also provide information on the following:*
* *That what is being supplied is suitable for the research being undertaken.*
* *That the quantity of tissue being supplied is suitable, but not excessive for achieving meaningful results.*

*[Maximum 1 side of A4]*

|  |
| --- |
| Not applicable |

1. **Justification of animal use (if applicable)**

*Sufficient information and justification regarding* ***any animal research proposed****, regardless of country, must be provided. Any use of animals must abide by UK and Thailand standards whether it is taking place in the UK or Thailand. If the project does not involve animal use, write ‘Non applicable’*

*Applications including the use of animals should fully justify the animal use, including the following.*

*A statement that:*

* *They will adhere to all relevant national and local regulatory systems in the UK and Thailand.*
* *They will follow the guidelines laid out in the* [*Responsibility in the use of animals in bioscience research document*](https://www.nc3rs.org.uk/sites/default/files/Responsibility%20in%20the%20use%20of%20animals%20in%20bioscience%20research%20-%20July%202015.pdf) *and ensure that work is carried out to UK standards.*
* *Before initiation of the proposed research work, appropriate approvals from Institutional and/or central animal ethics committees will be obtained for experimental protocols to be adopted in their projects from both the UK and* ***Thailand (this is a requirement regardless of where the research is taking place).*** *Successful proposals may be expected to provide copies of these permissions before funding is released.*

*Please also detail the following:*

* *Please indicate* ***where*** *the animal research will take place (UK or overseas) and through which funder the resources are being sought. Applicants should include confirmation that animal welfare standards at the UK and Thailand institutions meet the requirements outlined above.*
* *Justification of the choice of design and numbers of animals and interventions.*
* *Adequate information concerning methodological issues.*
* *Information on the planned procedures to minimise experimental bias (for example, randomisation protocols, blinding) should be outlined or an explanation included as to why such procedures are not appropriate.*
* *Power calculations.*
* *The rationale for the experimental design.*
* *Any additional information which was not included in the proposal document but which is pertinent to the animal research proposed and which the funders should be aware of.*

*[Maximum 1 side of A4]*

|  |
| --- |
| Not applicable |

1. **Work plan**

*Please provide a Gantt chart, or diagrammatic work plan, for the project including milestones [Maximum 1 side of A4]*

**Signed by the UK and Thailand Partner**

*Date UK PI*

*\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Date Thai PI*

*\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Input more as needed*