



**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 January 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 productive training environment.

The Thai data science partners have a strong track record in developing ML algorithms, including those for automated image analysis, (Boongoen, Anant, Iam-On, Uttuma) and setting-up and maintaining DM systems (Boongoen, Iam-On). The UK partners have extensive understanding of the methods and outputs of the pipelines used to process data from the Gravitational-wave Optical Transient Observatory (GOTO; see section 3 for a description of GOTO), which is the source of the large astronomical datasets that are used throughout the project. The Thai astronomy partners (Sawangwit, Awiphan) are familiar with and have ready access to the computing infrastructure based at NARIT that will be used throughout the project. Importantly, through our collaborative work during Phase 1, all data scientists and astronomers involved in the project now also have a good working knowledge of *each other’s* areas of expertise. This cross-disciplinary knowledge makes our team especially well-suited to exploiting astronomical datasets to train others – especially astronomers – in advanced data handling techniques.

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) establish a GOTO data centre based in NARIT that will double as a “Big Data” training resource for the broader Thai scientific community, (b) continue our successful graduate training programme in which students gain experience in advanced data-handling through cutting-edge research projects, and (c) disseminate our team’s knowledge of data handling through two 5-day practical workshops aimed at research staff and students based at NARIT and other Thai research institutes.

For (a), a Thai data-science student will visit UK for a 3-month visit to obtaining training in database design and implementation by working alongside GOTO scientists in establishing a UK data centre. He will then take the knowledge gained from that experience back to Thailand to assist in establishing a GOTO data centre within NARIT. However, rather than simply being a mirror of the UK data centre, it is important that the NARIT data centre also meets the data handling training needs of Thai researchers. To ensure this, we will test various different software and hardware systems by leasing facilities provided by Amazon Web Services (AWS). This will provide an extremely cost-effective way of ensuring the effectiveness of the data centre as a research asset and training resource for Thai scientists prior to investing in major infrastructure. 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 deep-learning analyses for source classification, including pixel-based, and (iii) researching how well the DM systems explored in Phase 1 scale to the large data rates provided by GOTO. Finally, for (c), staff and students involved in the project will contribute their expertise to disseminate practical skills in DM and ML to up to 30 trainees per 5-day workshop held each July of the grant (i.e., up to 60 trainees in total).

## 2. 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/> and [RCUK Newton Fund Guidance](#). 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. There is clear evidence 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 (Bakhshi & Mateos-Garcia, 2012), 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 Phase 2. 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"].

**References:** Bakhshi & Mateos-Garcia, 2012, The Rise of the Datavores, [www.nesta.org.uk/sites/default/files/rise\\_of\\_the\\_datavores.pdf](http://www.nesta.org.uk/sites/default/files/rise_of_the_datavores.pdf);

### 3. Detailed Research Information

#### a) 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 advanced data-handling skills. A secondary objective is the establishment of a Thai data centre to store, 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 achieved our Phase 1 goals, 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 “imbalanced” 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). Imbalanced 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 imbalanced data in GOTO difference imaging two of our graduate students have been researching techniques to artificially boost the information available from the true positives and, in doing so, are 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 one of the students has also been selected to give an oral presentation (attached to the JE-S application as it has yet to be published). 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, our graduate student (Pruangpharch) has been researching the suitability of different data management systems (DMS) to store GOTO data. In addition to setting up a relational database (as traditionally used by astronomy surveys e.g., SDSS), the student has also researched 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, experience with HDFS provides our trainees with skills that are directly transferrable to many economically important sectors. There are clear *potential* benefits for developing an 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 Pruangpharch has set up a small PostgreSQL relational database, but has yet to test how it compares to HDFS nor how it scales to the size and rate of data provided by GOTO.

**b) Description of proposed work [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 in 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 our 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, 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.

**2. Achieving the project's training goals**

To achieve our training goals, our project will consist of three main components:

- an extended visit to the UK by Thai researchers and students to collaborate on establishing a UK GOTO data centre;
- a series of 9 to 18-month research projects (see Gantt chart) during which four Thai students (3 MSc, 1 PhD) will establish a Thai GOTO data centre and research techniques to automatically analyse the data held therein;
- two 5-day practical workshops to be held each July during which we will use the data centre as training assets to instruct Thai staff and students based in NARIT and elsewhere on data handling techniques;

In the following, we detail the features and training benefits of each component:

**2.1 Extended visits to the UK**

To date, the GOTO collaboration's priorities have, necessarily, focussed on hardware development (telescope, dome construction etc.), control software and data processing (as opposed to analysis). With GOTO now fully operational, a major priority is the establishment of GOTO data centres to provide ready data access. This represents an excellent training opportunity for Thai students and staff to work alongside GOTO scientists to establish the first fully-functional GOTO UK data centre in preparation for setting-up a data centre in Thailand. To fully exploit this opportunity, two Thai staff and two Thai students will visit the UK to collaborate on designing and setting-up GOTO's UK database.

The two visiting Thai staff (Boongoen, Iam-on) are data scientists that have extensive experience of database development. They will visit the UK for two weeks in March, 2018, to discuss with GOTO scientists the database requirements. Joining them will be the two students, one of whom will remain in the UK for one month, the other for three months. The student staying for one month will receive training in GOTO's two data processing pipelines – the in-house GOTOflow and the other based on the Large Synoptic Survey Telescope's (LSST) pipeline. As well as enabling the student to use the pipelines and train others in their use, through this experience he will gain a thorough understanding of the data products held within GOTO's data centres which he can then disseminate to others. The student staying for three months (Pruangpharch) will work alongside the UK partners with further guidance from Boongoen and Iam-On to set-up GOTO's UK database system based on PostgreSQL. This experience will provide the student with extensive hands-on training in DMS and designing database infrastructure capable of organising vast amounts of digital data. Toward the end of his visit, the student will set-up a smaller cloud-based database built on the same principles as the UK data centre to be used as a training resource by attendees of the first 5-day practical workshop in July, 2018.

## 2.2 The research projects

Four postgraduate students will receive training-through-research by working on three research projects. The focus of these projects is to maximise the effectiveness of GOTO's Thai data centre as an asset for research and as a training resource for handling Big Data.

**(i) Adapt our ML algorithms to real GOTO data (1 MSc student):** Our Phase 1 research was based on simulated GOTO data. This was for two reasons: first, simulated data is a powerful tool during early development of ML algorithms as it allows us to test the outputs against known inputs; second, GOTO was not operational at the start of Phase 1. Having used simulated data in Phase 1 to develop a ML-based true/false-positive classification algorithm, in Phase 2 we will adapt this algorithm to real GOTO data. Even with our Phase 1 experience, overcoming the imperfections of real data while ensuring our algorithm can work quickly enough to handle GOTO's data rate is a significant research challenge.

**(ii) Automated source classification (1 MSc, 1 PhD student):** In addition to true/false positive classification, it is also vital that GOTO can classify different types of true sources (e.g., stars, galaxies, supernovae, AGN) in order to prioritise follow-up observations. During Phase 2, we will research ML-algorithms to classify sources in GOTO data. One student will research feature-based ML analyses, whereby features measured by the GOTO processing pipeline (e.g., colour, shape, etc) are used as inputs. Another student will research pixel-based ML algorithms, whereby the computer "looks" at the images directly and classifies based on the raw data (i.e., more akin to human classification).

**(iii) Data centre development (1 MSc student; Pruangpharch):** For GOTO to succeed, it is vital that it outputs automatically-prioritised targets for follow-up with other facilities, including those operated by NARIT. For this to happen as quickly as needed, all analysis must be performed within the data centres. To ensure a seamless interface between algorithm and data, we will research the benefits of both PostgreSQL and HDFS DMS in terms of speed and usability. First, a PostgreSQL database to mirror the UK GOTO database will be established, requiring roughly 1TB of storage capacity on NARIT's High Powered Computing system "Chalawan". The HDFS system, by contrast, is less straightforward to set up. As such, it will be more cost and time effective to first use a pre-configured system hosted on AWS Elastic MapReduce for testing before configuring our own system should these tests prove HDFS a better system overall. These tests will make establishing the Thai data centre itself an important research project within GOTO.

These research projects will make NARIT's data centre the primary testbed for cutting edge DM and ML analyses within GOTO, ensuring that it remains at the forefront of data handling research in astronomy, not only within the GOTO collaboration, but globally.

## 2.3 The practical workshops [outreach component]

We will hold two 5-day workshops held each July during Phase 2. They will be held in Chiang Mai and will be open to up to 30 attendees from NARIT and other Thai research institutes. The focus of these workshops will be on disseminating practical skills that our team have acquired during the project. We will cover topics including an introduction to SQL and HDFS, how to set up a cloud-based SQL 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 and homeworks and will be conducted by staff and graduate students involved in the project. For the practical sessions and homeworks, students will log into to a smaller, cloud-based reproduction of the Thai GOTO data centre hosted on AWS's Educate service and containing simulated data. This is to mitigate against (a) NARIT's data centre not being set-up prior to the first workshop and (b) security and data access issues that may arise should non-NARIT trainees log into the data centre proper. Where possible, teaching will be conducted in Thai to ensure maximal inclusivity.

#### 4. 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
----------------

#### 5. 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](#) 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

**Please note:** if the pages exceed the specified lengths your application may be rejected.

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
----------------

## 6. Work plan

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

See attached

### Signed by the UK and Thailand Partner

Date

UK PI

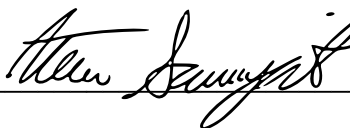
14 Nov 2017



Date

Thai PI

14 Nov 2017



Input more as needed

**Please note:** if the pages exceed the specified lengths your application may be rejected.