RGC Ref. No.:
UGC/FDS16/E01/16
(please insert ref. above)

**RESEARCH GRANTS COUNCIL**
**COMPETITIVE RESEARCH FUNDING SCHEMES FOR**
**THE LOCAL SELF-FINANCING DEGREE SECTOR**

**FACULTY DEVELOPMENT SCHEME (FDS)**

**Completion Report**
*(for completed projects only)*

***Submission Deadlines:***
1. *Auditor's report with unspent balance, if any: within **six** months of the approved project completion date.*
2. *Completion report: within **12** months of the approved project completion date.*

## Part A: The Project and Investigator(s)

### 1. Project Title

Design and development of a big data system for predicting harmful algal blooms

### 2. Investigator(s) And Academic Department(s) / Unit(s) Involved

| Research Team | Name / Post | Unit / Department / Institution |
|---|---|---|
| Principal Investigator | Prof. HO Kin-chung / Dean | School of Science and Technology / The Open University of Hong Kong |
| Co-Investigator(s) | Mr. TONG Bruce Kwong-bun / Senior Lecturer | School of Science and Technology / The Open University of Hong Kong |

### 3. Project Duration

| | Original | Revised | Date of RGC / Institution Approval *(must be quoted)* |
|---|---|---|---|
| Project Start Date | 1 January 2017 | N/A | N/A |
| Project Completion Date | 31 December 2018 | 12 May 2018 | 6 June 2018 |
| Duration *(in month)* | 24 | 16 months 12 days | 6 June 2018 |
| Deadline for Submission of Completion Report | 31 December 2019 | 12 May 2019 | 6 June 2018 |

## Part B: The Final Report

## 5. Project Objectives

5.1 Objectives as per original application

1. *To investigate different methods and models on detecting harmful algal blooms (HABs)*

2. *To investigate machine learning algorithms for detecting and identifying HABs.*

3. *To set up an experimental Big Data test-bed prototype, and install Apache Hadoop-based computing platform across all computing units.*

4. *To collect water sampling data from Environment Protection Department (EPD) and Agriculture, Fisheries and Conservation Department of Hong Kong Government.*

5. *Upon collecting Big Data from different means, e.g., satellite images, apply ML algorithm on the Big Data computing testbed on identifying HABs using ML with accuracy measurements*

5.2 Revised objectives

Date of approval from the RGC: N/A

Reasons for the change: Nil

*1.*

*2. ...*

5.3 Realisation of the objectives
*(Maximum 1 page; please state how and to what extent the project objectives have been achieved; give reasons for under-achievements and outline attempts to overcome problems, if any)*

The goal of the project is to investigate, design, create and establish a responsive method which shall automatically discover and report events of harmful algal blooms (HABs) in the coastal areas surrounding Hong Kong.

For the objective 1, at the early phase of the project, we spent the initial few months on understanding the biological nature of different algal types, and investigating different methods reported in published articles, online websites, refereed journals and conference papers on harmful algal blooms.

For the objective 2, some reported investigation methods were based on sampled and measured chemical and biological data, e.g., dissolved oxygen, nitrogen, phosphorus, etc. to test different machine learning (ML) algorithms. Examples of ML algorithms reported

were random forest, support vector machine (SVM) etc. Most of these results would be applicable to certain types of phytoplankton. Some of them may not be reproducible if certain chemicals were not measured in our sampled water.

We included a large amount of data that assisted in our derivation of different algorithmic methods in discovering harmful algal blooms. Hence, it was important to know the amount of data in Hong Kong that would be available to us and enable our research. For the objective 4, we got in touch with both the Agriculture, Fisheries, and Conservation Department (AFCD) and the Environmental Protection Department (EPD) of the Hong Kong Government. We received all biological and chemical data from the sampled water in different water sampling stations surrounding Hong Kong. However, the aggregate data from both department was only about 15 Mbytes in size.

With the collected water sampling data, there arose issues, which were that the sampling locations were mostly fixed, and sampling period between successive samplings was too wide in time. As an example, the EPD collected data once per month, i.e., 12 sets of data per year. This made the two successive measurement data less correlated. Fortunately, with our investigation, remote sensing was another approach that could be feasible for our investigation. There are different Earth observing Satellites (EoSs) going around the globe. They take the snapshots of the surface of the Earth which may make our research feasible. Each image scanned by a satellite is about 1 GByte in size.

For the Big Data testbed in the objective 3, we spent time on understanding different operating models and the open-source software systems. Our testbed consisted of three computing systems (each of them is installed with many RAMs, at least 32 GBytes), one 8 TBytes NAS, and one connecting switch. The software systems installed are the latest versions of Apache Hadoop and Apache Spark.

For the objective 5, we used deep learning algorithms to test on extracted satellite images. The EoS systems being investigated are the Sentinel-2/3 from ESA (European Space Agency) and Landsat-7/8 from NASA. Since the Sentinel-2 satellite system takes a shorter period in going around the Earth than the other three EoS systems, it was therefore our main research focus on analyzing the data provided by Sentinel-2 at the current stage.

5.4 Summary of objectives addressed to date

| **Objectives** *(as per 5.1/5.2 above)* | **Addressed** *(please tick)* | **Percentage Achieved** *(please estimate)* |
|---|---|---|
| 1. Background study on HABs | ✓ | 100% |
| 2. Reported Machine Learning (ML) methods for algal blooms | ✓ | 100% |
| 3. Testbed construction | ✓ | 100% |
| 4. Sampling data from AFCD and EPD | ✓ | 100% |
| 5. ML on Predicting HABs | ✓ | 100% |

## 6. Research Outcome

6.1 Major findings and research outcome
*(Maximum 1 page; please make reference to Part C where necessary)*

For the project, we started off by investigating the water sampling data from both the AFCD and EPD departments of the Hong Kong Government. The data gave us glimpses about the potential impact on causing algal blooms due to the concentrations of different nutrients and chemicals data. However, there were two main drawbacks regarding the collected sampled raw data: 1) the sampling frequencies were low for both departments; 2) the sampling locations were mostly fixed. There would be occurrences that some algal bloom events would be missed and unreported.

We therefore have resorted to another approach, the remote, sensing, in identifying any potential HABs. For oceanic and land information, there are different Earth observing Satellite (EoS) programmes, and some of them offer free downloads of scanned images of Earth surfaces. The programmes that we are interested in were: 1) Sentinel-2/3 from ESA (European Space Agency); 2) Landsat 7/8 from NASA (National Aeronautics and Space Administration). Each scanned image granule may take more than 1 GByte in size, and, unfortunately, the image types are different among them, For our project, due to the excellent scanned image spatial and good temporal quality, we focus on using the retrieved data from the Sentinel-2 programme.

For our testbed, it runs Linux operating system. The Big Data software platform uses the distributed, free and open-source software packages: Apache Hadoop and Apache Spark. For the radiometric resolution, the Sentinel-2 uses twelve sensing bands for spectral significance responses on each scanned pixel. Each retrieved image is about 1 GByte in size, and the format is in JPEG2000.

Deep learning algorithms are among the best and commonly used techniques for image and audio processing today. In this project, a three-layer neural network construct was used to run on Spark for scanned image processing. This basic model offered a simple output, that is, either yes or no on algal bloom. We trained our neural network using satellite images with the reported algal bloom data from AFCD. Although AFCD might not catch all algal bloom events, we attempt to retrieve those images on the dates that AFCD positively reported.

All in all, for this project, we developed software programmes for the remotely sensed data:

1. Regular daily image retrieval from Copernicus website, which is the access point to the Sentinel images
2. Data conversion into text format for data storage and processing (in future, data wrangling can be made based our devised storage files)
3. Deep learning neural network design that would be trained with known algal bloom events, and for testing with newly retrieved data.

As of today, all processes run on the testbed with acceptable quality.

*6.2* Potential for further development of the research and the proposed course of action *(Maximum half a page)*

We constructed a simple functioning Big Data testbed. We have not yet investigated the different operating components that could be added to the Big Data computing platform architecture. For example, data streaming mechanisms for data feeding is not in the current architecture. Another important hot topic is the deep learning neural network designs. We only adopt a simple neural network in the current design, more sophisticated with deeper layers should be investigated. In the current design, we have not yet evaluated different operating non-linear functional features, for example, on the uses of different activation functions in networks. In future, we are aware of the new research programmes being announced for the Greater Bay Area. Our research could be expanded and extended in broader research and geographical area.

## 7. Layman's Summary

*(Describe in layman's language the nature, significance and value of the research project, in no more than 200 words)*

Our research aim is to create a novel algal bloom prediction system. This software system can be used to alarm 1) the fish farms if any imminent harmful algal bloom might arise and kill the fish stocks in Hong Kong; 2) the general public not to consume certain fishes which be toxic for human beings; and 3) the swimmers who might have gone to swim in beaches with high concentration of harmful algae. For our system to operate, we adopted the scanned images of the Earth observing satellites for algal bloom research. Currently, a simple model has been created to run on a large-scale distributed computing platform, and this platform is capable of handling a large amount of data and expanding to run on thousands of computing system.

## Part C: Research Output

### 8. Peer-Reviewed Journal Publication(s) Arising Directly From This Research Project

*(Please attach a copy of the publication and/or the letter of acceptance if not yet submitted in the previous progress report(s). All listed publications must acknowledge RGC's funding support by quoting the specific grant reference.)*

| **The Latest Status of Publications** | | | | **Author(s)** *(denote the corresponding author with an asterisk*)* | **Title and Journal / Book** *(with the volume, pages and other necessary publishing details specified)* | **Submitted to RGC** *(indicate the year ending of the relevant progress report)* | **Attached to this Report** *(Yes or No)* | **Acknowledged the Support of RGC** *(Yes or No)* | **Accessible from the institutional repository** *(Yes or No)* |
|---|---|---|---|---|---|---|---|---|---|
| **Year of Publication** | **Year of Acceptance** *(For paper accepted but not yet published)* | **Under Review** | **Under Preparation** *(optional)* | | | | | | |
| 2017 | | | | K. L. E. Law, Kin Chung Ho | Remote Sensing and Big Data for Environmental Affairs," in *Smart Cities and Green Environment*, Environmental Publication House, Hong Kong, pp.39-58, 2017, ISBN 9-789881-777829. (book chapter) | Yes | No | No* | Yes |
| | | | | | | | | | |
| | | | | | | | | | |

*Due to technical reasons during editorial stage, the authors were unable to acknowledge the RGC in funding relevant research project. The PI and Co-I would like to express their apologies and affirm to avoid such a mistake in future.

**9. Recognized International Conference(s) In Which Paper(s) Related To This Research Project Was / Were Delivered**
*(Please attach a copy of each conference abstract)*

| **Month / Year / Place** | **Title** | **Conference Name** | **Submitted to RGC** *(indicate the year ending of the relevant progress report)* | **Attached to this Report** *(Yes or No)* | **Acknowl-edged the Support of RGC** *(Yes or No)* | **Accessible from the institutional repository** *(Yes or No)* |
|---|---|---|---|---|---|---|
| July/ 2018/ Napier, New Zealand | Neural Networks Trained with Known Algal Bloom Events and Satellite Images for Bloom Detection (Accepted) | New Zealand Marine Sciences Society Conference 2018 (http://www.nzmss2018.co.nz/) | No | Yes [Attachment 1] | Yes | Yes |
| August/ 2017/ Hong Kong | Remote Sensing and Algal Blooms: A Study of the Coastal Areas in Hong Kong | The 2nd Annual Conference of OUHK Institute of Research in Innovative Technology & Sustainability (IRITS) (http://plbpc001.ouhk.edu.hk/irits2017/) | Yes (2017) | No | Yes | Yes |
| | | | | | | |

**10. Whether Research Experience And New Knowledge Has Been Transferred / Has Contributed To Teaching And Learning**
*(Please elaborate)*

We have produced three publications. One is in form of a book chapter, the other two are in the form of oral presentation in conferences. One of the presentation is supposed to be made in July 2018. This is our goal to describe our design mechanisms in detail. We hope all these publication manuscripts or slides can assist tech and knowledge transfers.

List of publications:

1. K. L. E. Law, K. C. Ho, "Neural Networks Trained with Known Algal Blooms Events and Satellite Images for Bloom Detection," accepted to appear in New Zealand Marine Sciences Society (NZMSS) Conference 2018, July 3-5, 2018, Napier, New Zealand.

2. K. L. E. Law, " Remote Sensing and Algal Blooms: A Study of the Coastal Areas in Hong Kong," 2nd IRITS (Institute of Research for Innovation Technology and Sustainability) Symposium, The Open University of Hong Kong, Aug. 22-25, 2017, Hong Kong.

3. K. L. E. Law, Kin Chung Ho, "Remote Sensing and Big Data for Environmental Affairs," in *Smart Cities and Green Environment*, Environmental Publication House, Hong Kong, pp.39-58, 2017, ISBN 9-789881-777829.

**11. Student(s) Trained**
*(Please attach a copy of the title page of the thesis)*

| Name | Degree Registered for | Date of Registration | Date of Thesis Submission / Graduation |
|---|---|---|---|
| N/A | | | |
| | | | |
| | | | |

**12. Other Impact**
*(e.g. award of patents or prizes, collaboration with other research institutions, technology transfer, teaching enhancement, etc.)*

The PI of this project is actively involved in another research team which focuses on the riverine research in the Pearl River region. The team will likely take the research results from this project and expand it to cover the whole Greater Bay Area in future.

## 13. Public Access Of Completion Report

*(Please specify the information, if any, that cannot be provided for public access and give the reasons.)*

| **Information that Cannot Be Provided for Public Access** | **Reasons** |
|---|---|
| N/A | |

**RESEARCH GRANTS COUNCIL**
**COMPETITIVE RESEARCH FUNDING SCHEMES FOR**
**THE LOCAL SELF-FINANCING DEGREE SECTOR**

**FACULTY DEVELOPMENT SCHEME (FDS)**

**Completion Report - Attachment**
*(for completed projects only)*

| | |
|---|---|
| **RGC Ref. No.:** | UGC/FDS16/E01/16 |
| **Principal Investigator:** | Professor HO Kin-chung |
| **Project Title:** | Design and development of a big data system for predicting harmful algal blooms |

## Statistics on Research Outputs

| | Peer-reviewed Journal Publications | Conference Papers | Scholarly Books, Monographs and Chapters | Patents Awarded | Other Research Outputs (Please specify) |
|---|---|---|---|---|---|
| No. of outputs arising directly from this research project [or conference] | | 1 presented/ published, 1 accepted for presentation | 1 book chapter published | | |