UNIVERSITY OF TWENTE.

BIG GEODATA ANALYTICAL METHODS AND DISTRIBUTED COMPUTING

INTRODUCTION

Mahdi KHODADADZADEH
Mahdi FARNAGHI
Serkan GIRGIN
February 2022

ITC FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

LEARNING OBJECTIVES

- **LO1** Explain to peers the fundamentals of big geodata processing.
- **LO2** Compare various big geodata solutions.
- **LO3** Design workflows that run on the cloud and consider options for efficient computing.
- **LO4** Prepare and maintain a code repository.
- **LO5** Interpret the analytical results and demonstrate their reproducibility.

TOPICS

- Introduction to Jupyter Notebook and JupyterLab
- Introduction to big geodata (including the seven Vs: Volume, Velocity, Variety, Variability, Veracity, Value and Visualization)
- Principles of big geodata modelling and analysis (clustering, classification and regression tasks).
- Big geodata solutions (e.g. Google Earth Engine vs. HADOOP/SPARK or DASK-based solutions).
- Code versioning

COURSE SCHEDULE

- Ten days of lectures, tutorials and practicals (February 17 - March 3)
- About 4 days for project work and evaluation/exam
- Materials (including recordings) and announcements will be published in Moodle

Slot	Time (BD)	Time (NL)	Content
1	8:30-12:30	3:30-07:30	Work on exercises / reading materials / tutorials
2	12.30-13.30	07.30-08.30	Lunch break
3	13.30-14.00	08.30-09.00	Online feedback sessions
4	14.00-16.30	09.00-11.30	Online lectures

TIMETABLE

- Mix of online lectures and practicals (exercises and guided discussions)
- Plan and use well the self-study slots

Day	Content
#1 (17 Feb)	Introduction to the course, project and CRIB Introduction to Jupyter Notebook and JupyterLab
#2 (20 Feb)	Review Phase I and II on CRIB
#3 (22 Feb)	Big geodata and Machine Learning
#4 (23 Feb)	Git and GitHub Distributed computing
#5 (24 Feb)	Google Earth Engine
#6 (27 Feb)	Project group work
#7 (28 Feb)	Cloud computing
#8 (1 Mar)	How to setup cloud VRE
#9 (2 Mar)	Project group work
#10 (3 Mar)	Projects presentation and assessment

GROUP PROJECT

Big Spatio-temporal Data Analytics: Advanced Machine Learning Modeling with Python

- Each group will have 3 participants (i.e., 3 groups).
- Four days of work
- We will schedule “question hours” during the project work
- We will monitor the discussion forum

From: R. Zurita-Milla, R. Goncalves, E. Izquierdo-Verdiguier and F. O. Ostermann, "Exploring Spring Onset at Continental Scales: Mapping Phenoregions and Correlating Temperature and Satellite-Based Phenometrics," in *IEEE Transactions on Big Data*, vol. 6, no. 3, pp. 583-593, 1 Sept. 2020

PROJECT QUESTION HOURS

- Bring in your questions to the meeting
- Use the discussion forum
- Be active, help others
- Teamwork is very appreciated

Slot	Date	Time (BD)	Time (NL)
1	Monday, 28 February 2022	13.30-14.00	08.30-09.00
2	Tuesday, 29 February 2022	13.30-14.00	08.30-09.00

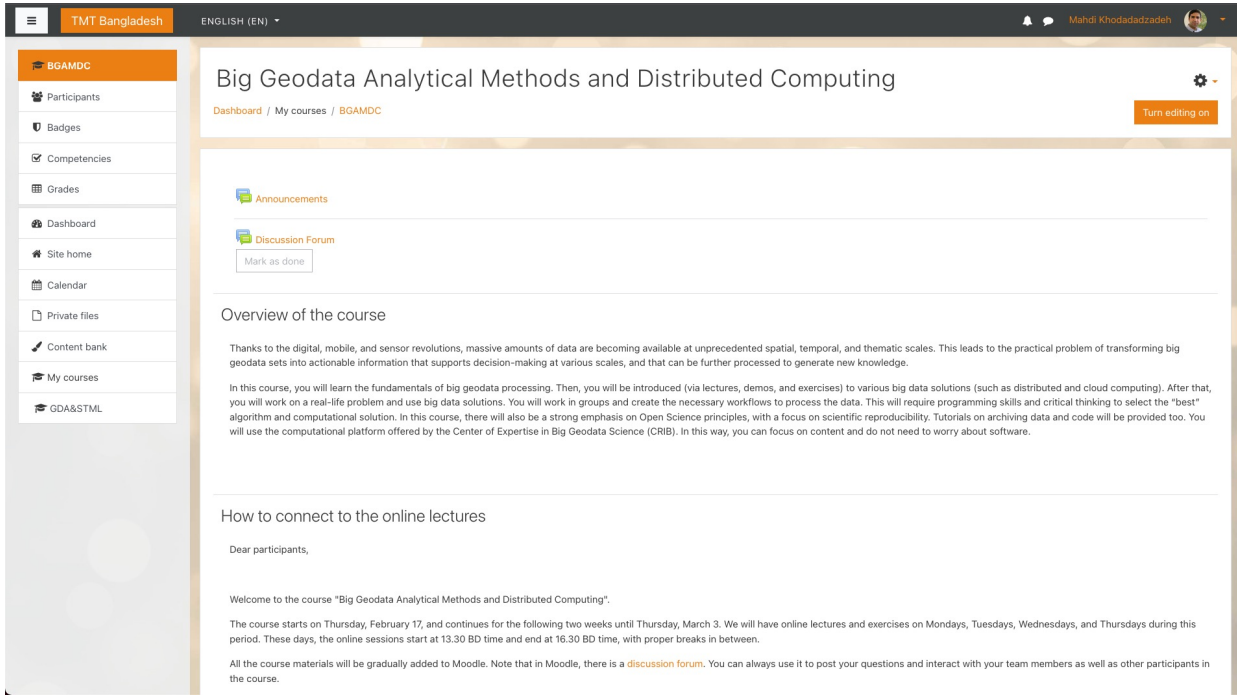
ASSESSMENT

The assessment will be based on group projects (50%) and individual oral examinations (50%)

- The project assessment will be done based on 15 mins presentation and preparation of a well-documented Jupyter notebook
- The oral examination will be done after the project presentation (±10 mins per participant)

MOODLE

Materials will be added on the go … check regularly



QUESTIONS?

UNIVERSITY OF TWENTE.

GROUP PROJECTS

Mahdi KHODADADZADEH
Mahdi FARNAGHI
Serkan GIRGIN
February 2022

TOPICS

- Big Spatio-temporal Data Analytics: Advanced Machine Learning Modeling with Python
  - Data management
  - Data exploration
  - Data analysis / modelling
    - Machine Learning
    - Distributed Computing
  - Big data solutions

INPUT DATA, METHOD AND RESULTS

- The input data should be explored sufficiently
- Proper tools need to be applied to identify outliers, missing data or other relevant features
- The workflow should rely on advanced and/or distributed processing solutions that operates on a big dataset
- A reasonable and justified choice of methods should be made
- The methods should be used correctly

- Results need to be discussed and sufficiently interpreted

PROJECT ASSESSMENT

- The project assessment will be done based on 15 mins presentation and preparation of a well-documented Jupyter notebook → A code repository with all the code (one or more scripts) developed during the project.
- The main notebook must explain the overall workflow of the project and any operation or task that was done manually.
- It should have the following sections: introduction, methods, results, and conclusions.
- It should provide a compact but informative background to the research problem.
- It should clearly state the objective of the project work and the contribution of each group member.

PROJECT PROPOSALS

1) A comparative study of decision tree, random forest, supervised and unsupervised classification for monitoring crop and fallow of Indian Sundarban region (Debolina Sarkar)
2) Crop type mapping in Godagari Upazila using sentinel-2 time series data and machine learning algorithms (Shakhawat Hossain)
3) Rice production area estimation (Hasan Md. Hamidur Rahman)
4) Prediction of soil fertility and crop productivity through machine learning algorithms (Mustafa Kamal Shahadat)
5) Hyperparameter optimization and performance assessment of supervised algorithms for land cover classification (Suman Biswas)
6) Crop land classification (Istiak Ahmed)
7) Prediction of the crop condition (healthy or damaged) throughout the harvesting season (Afroza Begum)

GROUPS

Group 1	Group 2	Group 3
Afroza Begum	Ummy Kulsum	Debolina Sarkar
Md Golam Mahboob	Istiak Ahmed	Shakhawat Hossain
Hasan Md. Hamidur Rahman	Mustafa Kamal Shahadat	Suman Biswas

Discuss these points:

- Which project do you select?
- What is the dataset? Is it Big?
- Does your project need data preparation?
- What are the possible machine learning algorithms?
- Which distributed processing solution will you use?
- Will you use GEE?
- What is the expected outcome?

ASSIGNMENT

- Send in a one-page description of your project, by February 27.
  - One paragraph introduction
  - One paragraph dataset
  - One paragraph method
  - One paragraph expected outcomes

- Start working on your projects!

QUESTIONS?