Course Code
Big Data Analysis
Module 1, 2024-25

Course Information

Instructor: Ding Ma
Office: PHBS Building, Room

Phone: 86-755-2603-
Email:
Office Hour:

Teaching Assistant:
Phone:
Email:

Classes:
Lectures: Monday and Thursday, 10:30 am - 12:20 pm
Venue: PHBS Building, Room

Course Website:
TBA

1. Course Description

1.1 Context

Course overview:
This course introduces the fundamental theory and applications of big data analysis, and teaches some commonly and widely used data processing techniques in various fields. We will cover data sampling, visualization, pattern recognition, and prediction, various machine learning and deep learning methods.

In an era where data is the new oil, the ability to analyze and extract meaningful insights from massive datasets is crucial for gaining a competitive edge in any industry. The "Big Data Analysis" course is designed to equip students with the skills and knowledge necessary to harness the power of big data, transforming raw information into actionable insights that drive strategic decision-making and innovation.

This course begins by introducing the fundamental concepts of big data, including its characteristics, challenges, and the importance of data-driven decision-making in today's digital economy. Students will explore the entire data lifecycle, from data collection and storage to processing and analysis, gaining a comprehensive understanding of how big data pipelines are constructed and managed.

A significant portion of the course will focus on advanced data analysis techniques, including machine learning algorithms that are scalable to big data environments. Students will learn how to apply these techniques to perform predictive analytics, anomaly detection, and trend analysis on massive datasets. The course will also cover the use of deep learning models in big data contexts, providing students with the skills to develop AI-driven solutions that can process and learn from extensive data streams.

Furthermore, this course will explore the practical challenges of working with big data, such as data privacy, security, and ethical considerations. Students will examine case studies from various industries, including finance, healthcare, retail, and social media, to understand how big data is used to solve real-world problems and create value. They will also learn about the regulatory frameworks that govern the use of big data and the importance of maintaining ethical standards in data analysis.

In the final stages of the course, students will engage in hands-on projects that simulate real-world big data analysis scenarios. These projects will involve working with large datasets, using industry-standard tools and techniques to analyze data, and presenting findings in a way that is both insightful and actionable for business stakeholders. Through these projects, students will not only reinforce their technical skills but also develop critical thinking and problem-solving abilities that are essential for a successful career in the field of big data.

By the end of this course, students will be proficient in the use of big data tools and techniques, capable of designing and implementing scalable data analysis solutions, and prepared to tackle the complex challenges of analyzing and interpreting large datasets. Whether you are aspiring to become a data scientist, a data engineer, or a business analyst, this course will provide you with the foundational knowledge and practical experience needed to excel in the fast-evolving world of big data.

Prerequisites:

Fair Knowledge of programming and statistics

1.2 Textbooks and Reading Materials

- No official textbooks, we use the following two books as References.
- James Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor "An introduction to statistical learning with Applications in Python."
- Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman "The elements of statistical learning: data mining, inference, and prediction."

2. Learning Outcomes

2.1 Intended Learning Outcomes

Learning Goals	**Objectives**	**Assessment (YES with details or NO)**
1. Our graduates will be effective communicators.	1.1. Our students will produce quality business and research-oriented documents.	YES
	1.2. Students are able to professionally present their ideas and also logically explain and defend their argument.	YES
2. Our graduates will be skilled in team work and	2.1. Students will be able to lead and participate in group for projects, discussion, and presentation.	YES

leadership.	2.2. Students will be able to apply leadership theories and related skills.	
3. Our graduates will be trained in ethics.	3.1. In a case setting, students will use appropriate techniques to analyze business problems and identify the ethical aspects, provide a solution and defend it.	
	3.2. Our students will practice ethics in the duration of the program.	YES
4. Our graduates will have a global perspective.	4.1. Students will have an international exposure.	YES
5. Our graduates will be skilled in problem-solving and critical thinking.	5.1. Our students will have a good understanding of fundamental theories in their fields.	YES
	5.2. Our students will be prepared to face problems in various business settings and find solutions.	YES
	5.3. Our students will demonstrate competency in critical thinking.	YES

2.2 Course specific objectives

See section 1.1 Context.

2.3 Assessment/Grading Details

Attendance 15%, Assignments 20%, Exams 30%, Final Project 35%

The level of background knowledge may vary among students, but it will be ignored in grading.

2.4 Academic Honesty and Plagiarism

It is important for a student’s effort and credit to be recognized through class assessment. Credits earned for a student work due to efforts done by others are clearly unfair. Deliberate dishonesty is considered academic misconducts, which include plagiarism; cheating on assignments or examinations; engaging in unauthorized collaboration on academic work; taking, acquiring, or using test materials without faculty permission; submitting false or incomplete records of academic achievement; acting alone or in cooperation with another to falsify records or to obtain dishonestly grades, honors, awards, or professional endorsement; or altering, forging, or misusing a University academic record; or fabricating or falsifying of data, research procedures, or data analysis.

All assessments are subject to academic misconduct check. Misconduct check may include reproducing the assessment, providing a copy to another member of faculty, and/or communicate a copy of this assignment to the PHBS Discipline Committee. A suspected plagiarized document/assignment submitted to a plagiarism checking service may be kept in its database for future reference purpose.

Where violation is suspected, penalties will be implemented. The penalties for academic misconduct may include: deduction of honour points, a mark of zero on the assessment, a fail grade for the whole course, and reference of the matter to the Peking University Registrar.

For more information of plagiarism, please refer to *PHBS Student Handbook*.

3. Topics, Teaching and Assessment Schedule (Tentative)

Week	Dates	Topics
1	Sep 2, 5	Introduction to Big Data Analytics: Understanding Central Limit Theorem (CLT) and Python for Data Analysis
2	Sep 9, 12	Financial Data Analysis: Multiple Testing, FWER, FDR, and Linear Regression in Fund Manager Data
3	Sep 14, 19	Stock Market Prediction: Applying Logistic Regression, LDA, and Naïve Bayes to S&P 500 Index Data
4	Sep 23, 26	Portfolio Optimization: Cross-Validation and Bootstrap Methods with Portfolio Data
5	Oct 9, 10	Predictive Modeling in Sports: Ridge Regression, Lasso, and Model Selection Techniques on Hitters Salary Data
6	Oct 14, 17	Midterm Exam and Project Proposals: Designing Big Data Projects with Real-World Applications
7	Oct 21, 24	Housing Market Analysis: Tree-Based Methods (Bagging, Random Forests, Boosting) on Boston Housing Data
8	Oct 28, 31	Advanced Neural Networks: Exploring CNNs and RNNs with Image and Document Data
9	Nov 4, 7	Crime Data Analytics: Unsupervised Learning Techniques (PCA, K-Means, Hierarchical Clustering)

4. Miscellaneous