COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Big data analysis	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: Julius Venskus **Other(s):**	Department of Econometric Analysis

Study cycle	Level of course	Type of the course unit (module)
First	Advanced	Compulsory

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
Face-to-face	Second (spring) semester	English

Requirements for students	
Prerequisites: descriptive statistics, basics of parametric hypothesis testing, estimation theory and causal modelling; basics of Python; ability to understand English at the level of independent user (B1 according to CEFR classification). Familiarity with machine learning, Python procedural and objective oriented programming, and arbitrary relational data base management system would be an advantage.	**Additional requirements (if any):**

Course (module) volume in credits	Total student's workload	Contact hours	Self-study hours
10	250	66	184

Purpose of the course unit (module): programme competences to be developed (the number in the brackets coincides with that given in the official description of the programme)

- creatively solve nonstandard theoretical and empirical problems (1);
- critically analyse and correctly apply the results presented in the scientific literature (2);
- apply the interdisciplinary knowledge (4);
- prepare raw empirical data for the econometric analysis and professionally operate the econometric software (10);
- analyse big data (9);
- evaluate the adequacy of statistical models and modify the models appropriately (8);
- know and understand at advanced level the problems and principles of data science (5).

Learning outcomes of the course unit (module); after completing the course students should:	Teaching and learning methods	Assessment methods
• be familiar with typical statistical models applied to big data analysis; • be able to use software tools designed for big data analysis; • be able to formalize practical problems and select appropriate statistical models suitable for the data at hand; • read and critically analyse results presented in literature as well as implement newly suggested methods.	Lectures, problem solving and reading, assignments, individual tasks accomplishment	Tests, evaluation of individual assignments

Content: breakdown of the topics	Contact hours							Self-study work: time and assignments	
	Lectures	Tutorials	Seminars	Exercises	Laboratory work	Internship/work placement	**Contact hours**	**Self-study hours**	**Assignments**
1. Introduction. Main concepts (big data, big data analytics, etc.). Examples. Brief review of currently available analytical methods along with appropriate software.	1						**1**	**10**	Read ch. 1 of [2]. Find an introductory article on big data analysis and read it on your own.
2. Distributed and parallel computing for Big Data. Typical tasks and theoretical models. Main concepts of parallel computing, distributed computation systems and architectures. Computational task splitting into parallel tasks. Advanced Python for parallel computing. Examples. Review of existing methodologies and practical approaches. Big data handling: cluster computing, batch and real time processing, NoSql database management systems. Short survey of typical machine learning models frequently encountered in big data analysis.	6				7		**13**	**65**	Solve a set of individual tasks assigned by the lecturer (assignment depends on the level of the students at hand and is not, therefore, detailed here). Regularly accomplish exercises designed for gaining of appropriate skills.
3. Software. Hadoop, Python, IPython, Jupyter notebook, Apache Spark, NoSql database management systems. A brief review of other relevant software.	18				16		**34**	**65**	Split into groups consisting of several students. Choose some software tool designed for big data analysis and undiscussed by the lecturer. Get familiar with it on your own. Present the tool to your classmates during the seminar.
4. Real data analysis. Worked examples of various real data analysis encompassing formulation of a problem and a full solution.	4						**4**	**24**	Carefully work on your own through several examples pointed out by the lecturer.
5. Tests and exam.					14		**14**	**20**	Prepare for tests.
Total							**66**	**184**	

Assessment strategy	Weight,%	Deadline	Assessment criteria
Test 1	15	3rd study week	The test consists of at most 5 practical tasks intended to check the level of knowledge obtained. The total weight of these tasks equals to 1 point. The weight of each task ranges from 0.1 to 1 point. Tasks are designed to be solved in a written form or by making use of computer and appropriate software. Each task is evaluated as follows: a) the task is divided into
Test 2	15	6th study week	
Test 3	15	9th study week	

Test 4	15	12th study week	parts and each part is assigned an appropriate amount of points; b) if student accomplishes the part without mistakes, the whole amount of that part is attained; otherwise, the amount is reduced considering the mistakes made; c) the parts are evaluated independently.
Project type assignment	20	13-15 study weeks	The first assignment corresponds to the topic number 3 of the table „Contents: breakdown of the topics“ and is described in a column „Assignments“. The lecturer evaluates the following: 1) the amount of work done; 2) the quality of delivered presentation. Each group of students is allowed to show their work to the lecturer and to get critical comments before delivering the presentation. The mistakes pointed out are not taken into consideration when the evaluation takes place.
Individual assignment	20	The final examination session; however, the student is allowed to report at any time during the regular semester	At the beginning of the semester, every student gets a set of the tasks to be accomplished individually. He is allowed to report whenever he is finished. That is, evaluation can take place both during the semester and during the exam. The lecturer takes into account not only the solutions presented, but also the results of a short interactive interview conducted during the time of reporting.
External examination is not allowed.			

Author	**Year of publication**	**Title**	**Issue of a periodical or volume of a publication**	**Publishing place and house or web link**
Compulsary reading				
1. V. Skorniakov	2018	Introduction to Big Data Analysis. Lecture Notes		https://klevas.mif.vu.lt/~visk/BigData/
2. Balamurugan Balusamy, Nandhini Abirami. R, Seifedine Kadry, and Amir H. Gandomi	2021	Big Data. Concepts, Technology, and Architecture		John Wiley & Sons, Inc.
3. Giancarlo Zaccone	2020	Python Parallel Programming Cookbook - Second Edition		Packt, https://github.com/PacktPublishing/Python-Parallel-Programming-Cookbook-Second-Edition
4. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani	2013	An Introduction to Statistical Learning: with Applications in R		Springer
Optional reading				
5. Vignesh Prajapati	2013	Big Data Analytics with R and Hadoop		Packt Publishing
6. Alpaydin, Ethem	2016	Introduction to Machine Learning, Third Edition		The MIT Press
7. Sebastian Raschka	2015	Python Machine Learning		Packt publishing
8. Max Kuhn, Kjell Johnson	2013	Applied Predictive Modeling		Springer
9. Brett Lantz	2015	Machine Learning with R - Second Edition		Packt Publishing
10. Trevor Hastie, Jerome H. Friedman, Robert Tibshirani	2016	The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition		Springer
11. Bart Baesens	2014	Analytics in a Big Data World		John Wiley & Sons, Inc.