COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Data Analysis Methods	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: dr. Donatas Narbutis	Faculty of Physics

Study cycle	Type of the course unit (module)
Second	Optional

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
Face-to-face	1st semester	Lithuanian, English

Requirements for students	
Prerequisites: Basics of calculus, Basics of probability theory	**Additional requirements (if any):** Basics of programming

Course (module) volume in credits	Total student's workload	Contact hours	Self-study hours
5	140	64	76

Purpose of the course unit (module): programme competences to be developed

To provide theoretical knowledge and practical skills necessary to perform analysis of experimental and observational data, to simulate numerical experiments, to process and homogenize data, to visualize multidimensional data-sets and search for patterns, to select methods for statistical analysis and to perform it, to evaluate the reliability of the results, to formulate and to base the conclusions of the analysis.

Learning outcomes of the course unit (module)	Teaching and learning methods	Assessment methods
Will be able to prepare data in various formats for analysis.	Lectures, laboratory works, independent work.	Laboratory works, exam.
Will be able to select the method of data visualization and explain visualization errors.	Lectures, laboratory works, independent work.	Laboratory works, exam.
Will be able to choose methods for statistical analysis.	Lectures, laboratory works, independent work.	Laboratory works, exam.
Will be able to assess the reliability of the analysis results, to formulate and to base the conclusions of the analysis.	Lectures, laboratory works, independent work.	Laboratory works, exam.
Will be able to interpret the data of real experiment within limits of constructed simplified numerical experiment and explain the physical meaning of model parameters.	Lectures, laboratory works, independent work.	Laboratory works, exam.

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. Overview of data analysis methods. Measurements and data acquisition, error sources. Python in data analysis. Numpy and matplotlib libraries.	2				3			7	Literature, tasks.
2. Digital data formats. Data structures and formats. Reading and saving various file formats. Pandas library (homogenization, reductions, selection, filtering, grouping).	2				2			6	Literature, tasks.
3. Data preparation. Errors in data, filtering, sorting. Basic statistical calculations. Interpolation and extrapolation of data. Large-scale processing of map-reduce algorithm.	2				3			7	Literature, tasks.
4. Data visualization. Importance of visualization and approaches. Plotly library. Histograms, two-dimensional diagrams. Bad and good visualization practices. Multidimensional data. Interactive analysis with TopCat.	2				3			7	Literature, tasks.
5. Numerical experiment. Linear relationships. Random numbers and their generation. Uniform and Gaussian distributions, their properties and application. Modeling of measurement errors.	3				3			7	Literature, tasks.
6. Model parameter estimation. Estimating parameters of linear relationship. The least squares method. The influence of data-point number and measurement errors on the reliability of model parameter estimates.	3				3			7	Literature, tasks.
7. Digital images. Image registration, pixels. Poisson distribution, its properties and application. Image distortions, camera sensitivity. Image matching.	3				3			7	Literature, tasks.
8. Signal detection. Time series, filtering, convolution, correlation. Spatial filtering, object detection in images.	4				3			7	Literature, tasks.
9. Non-linear relationships. Optimization packages and methods (Levenberg-Marquard, Markov Chain Monte Carlo, genetic). Formulation of the model. Criterion of model goodness. Local and global minimums. Model complexity and selection principles.	5				3			7	Literature, tasks.
10. Data classification. Scikit-learn library. Training and testing datasets. Decisions tree classification. Dimension reduction, linear and logistic regression. Clustering. Model selection.	3				3			7	Literature, tasks.
11. Image classification. Keras library. Structure of neural network and its adaptation for image analysis. Data preparation. Training process and control. Reliability of results.	3				3			7	Literature, tasks.
Total	**32**				**32**			**76**	

Assessment strategy	Weight,%	Deadline	Assessment criteria
Laboratory work	50	1-15 weeks of semester	Students must complete and present reports for 8 laboratory work projects per semester. During the laboratory work hours the aim and tasks of each project are formulated, the possibilities of code implementation are clarified. Work is completed during self-study hours. One laboratory work project is allocated 4 hours of laboratory work time and 9.5 hours self-study time. The reports must be presented until the end of semester. Each report must be provided as the Jupyter notebook with a program code, comments and illustrations. Reports are evaluated by the following criteria: - Achieved goal, objectives, proper description, software code and illustrations, sound results and based conclusions – 10. - Achieved goal, but there are at least one essential or several minor comments on the fulfillment of objectives, description, program code, illustrations or conclusions – 9. - Partially achieved goal, there are two essential comments to be corrected – 8. - Partially achieved goal, but more than half of the tasks should be corrected – 7. - At least two tasks have been performed – 6. - At least one task has been performed – 5. - The Jupyter notebook is submitted, containing description of the formulated goal and tasks, and thoughts on the methods that could be used to solve tasks presented with discussion of the reliability of the methods – 4-1. - Nothing is provided – 0. An average of 8 laboratory project evaluations is computed (L).
Exam	50	16th week of semester	During the semester, a creative research project is arranged, which can be based on combination of all or at least some laboratory project works. The theme of the creative project has to be arranged until the middle of the semester. The report of the creative project has to be sent prior to the exam in PDF format. The evaluation criteria for an exam's creative project report are the same as those for laboratory project reports provided above. During the exam, the results of the creative project are presented and discussed, 5 questions concerning the project have to be answered. The exam's maximum rating (E) is equal to evaluation of the report of creative project when detailed answers to at least 4 out of 5 questions are provided. A correspondingly lower rating is given if fewer questions are answered. If there are no answers given to all 5 the questions, the assessment of the exam equals the evaluation of the report of creative project minus 4 points. Final rating P = 0.5 L + 0.5 E.

Author	Year of publication	Title	Issue of a periodical or volume of a publication	Publishing place and house or web link
Compulsory reading				
Jake VanderPlas	2016	Python Data Science Handbook		Oreilly https://github.com/jakevdp/PythonDataScienceHandbook
Wes McKinney	2017	Python for Data Analysis	2nd Edition	Oreilly

Optional reading				
Scikit-learn developers	2018	Scikit-learn user guide		http://scikit-learn.org/dev/_downloads/scikit-learn-docs.pdf
John V. Guttag	2016	Introduction to Computation and Programming Using Python	2nd Edition	https://mitpress.mit.edu/books/introduction-computation-and-programming-using-python-second-edition