COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Applications of distributed computing to telecommunication network modeling	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: assoc. prof. Rimvydas Aleksiejūnas **Other(s):**	Faculty of Physics, Vilnius University

Study cycle	Type of the course unit (module)
Master (graduate)	Mandatory

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
Classroom training	3 (autumn) semester	Lithuanian

Requirements for students	
Prerequisites: Basic knowledge of computer programming	**Additional requirements (if any):**

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		
Students will get familiar with Apache Spark, PySpark, IPython / Jupyter Notebook programming environments, gain experience with big data analytics and distributed computing, will use practical exercises of statistical model applications for signal propagation and telecommunications network performance (coverage, capacity and reliability) modeling.		
Learning outcomes of the course unit (module)	**Teaching and learning methods**	**Assessment methods**
- Will be able to apply statistical models for big data processing. - Will acquire theoretical knowledge and gain analytical data processing abilities, will be able to apply knowledge in solving practical problems of telecommunication network modeling.	Self-study laboratory exercises, studying literature, discussions.	Laboratory works: performing practical tasks, automatic tests.
- Gain knowledge in big data analytics and distributed computing for telecommunication network modeling.	Teaching problem-solving skills, demonstrations, discussion.	Written exam, closed-ended questions.

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. Big data processing principles and techniques: MapReduce, Apache Hadoop and Spark. Virtual work environment, Vagrant, VirtualBox.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
2. Introduction to Python programming. Basics of functional programming and lambda functions. Working with Python libraries: numpy, scipy, matplotlib, pandas, scikits-learn and statsmodels. Forecasting of telecommunication network services using scikits-learn tools.	2				2		**4**	**6**	Literature reading, completion of lab assignments and self-study tasks, improvement of Python programming skills.
3. Apache Spark framework, PySpark interface, IPython / Jupyter Notebook environment. Command line interface, remote server configuration. Installation and configuration.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
4. Apache Spark architecture and programming model. Data storage in HDFS and HBase. Distributed computing using resilient distributed dataset (RDD) model. Data transformations and actions. Key-value pairs and broadcast variables.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
5. Data import into Spark system. Data filtering and aggregation. Analysis of HTTP web server logfile and graphical display of results.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
6. Machine learning library Apache Spark MLlib. Analysis of algorithms for classification, regression and clustering. K-nearest neighbors classifier (k-NN). Fastest descent method. Dimensionality reduction using principal component analysis (PCA) and singular value decomposition (SVD) methods. Prediction of telecommunication service usage and network outage forecasting using MLlib library.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
7. LTE radio access layer, power budget calculations and interference analysis. LTE heterogeneous networks. Base station coverage prediction and statistical analysis using Apache Spark tools.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
8. Review of radio wave propagation models. Application of linear regression for analysis of signal propagation models. Radio wave propagation model calibration based on measurement data using regression models.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
9. Land cover type classification using clustering algorithms. Random forest algorithm. Classification of categorical and numerical data.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
10. Clustering and classification algorithms for mobile data traffic analytics. Application of MapReduce algorithm for statistical analysis of mobile data traffic.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
11. Statistical analysis of mobile network quality	2				2		**4**	**4**	Literature reading,

indicators. Correlation analysis of user location, signal level, network loading and quality of service. Open access databases: NetRadar, OpenCellID and OpenSignal. Statistical analysis of mobile network quality indicators for the selected area.									completion of lab assignments and self-study tasks.
12. Machine learning methods for wireless sensor networks. Clustering of mobile sensor data. Analysis of WISDM Actitracker sensor data. Use of triaxial accelerometer measurements for determining human activity. Effects of mobile receiver's position on the quality of communication link.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
13. Real-time analytics with Apache Spark streaming tools. Statistical analysis of call detail records (CDR). HTTP web server log file analysis using Spark streaming package.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
14. Internet of things and data mining. Statistical analysis of wireless sensor data. Apache Storm package. Temporal series analysis using spark-ts library.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
15. Anomaly detection in network traffic. Anomalies in sensor networks. Anomaly detection in web data traffic using principal component analysis.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
16. Geoprocessing using Apache Spark. Telecommunication network planning tasks using geographical data analysis. Localization and RF fingerprinting in wireless networks.	2				2		**4**	**4**	Literature reading, completion of lab assignments and self-study tasks.
Preparation for the exam								**10**	
Total	**32**				**32**		**64**	**76**	

Assessment strategy	**Weight,%**	**Deadline**	**Assessment criteria**
Laboratory work: performing practical tasks	70	During the semester	Evaluation of the acquired knowledge and analytical skills, the ability to apply knowledge to solve practical telecommunication network modeling tasks, prepared in the Jupyter Notebook environment. Each laboratory work consists of practical assignment according to single lecture's topic with step-by-step instructions and about 20-40 automatic tests, and 1-2 self-study short tasks. **5-9 points:** Estimated proportionally to the number of tests completed correctly: 50% - 5 points, 60% - 6 points, etc. **10 points:** More than 90% of tests completed and at least one self-study task performed.
Written exam	30	During the session	Evaluation of knowledge from big data analysis and applications of distributed computing methods for telecommunication network modeling. The exam consists of 20 closed-ended questions with possible several correct answers. Each answer is scored on a scale of 0 to 0.5, and the result is summed up: **0 points:** At least one wrong answer is selected or not all the correct answers are selected; **0.5 point:** One or all available correct answer options are selected.

Author	**Year of**	**Title**	**Issue of a periodical**	**Publishing place and house or web link**

	publication		or volume of a publication	
Compulsary reading				
E. Alpaydin	2014	Introduction to machine learning		Cambridge, Massachusetts: The MIT Press
J. Grus	2015	Data Science from Scratch		O'Reilly Media, Inc.
C. O'Neil and R. Schutt	2013	Doing Data Science		O'Reilly Media, Inc.
Optional reading				
V. Čekanavičius and G. Murauskas	2000, 2002, 2009	Statistika ir jos taikymai	I-III dalys	Vilnius: TEV
A. Smola and S. V. N. Vishwanathan	2008	Introduction to machine learning		Cambridge University Press
W. McKinney	2013	Python for data analysis: Data wrangling with Pandas, NumPy, and IPython		O'Reilly Media, Inc.