Institute of Cyber-Physical Systems			2nd semester according to the sample curriculum 2025–26/2			
Subject name:	Code:	Credit:	Number of lessons			
				ea	tgy	lab
Data preparation and transformation	NKXAF1EMNF	5	full-time, semester	2	0	2
Responsible person for the subject: Balázsné Dr. Kail Eszter			Classification: Senior lecturer			
Subject lecturer(s): Angyalné Dr. habil Alexy Márta, Dr. Alwahab Dhulfiqar Zoltán						
Prerequisites:						
Way of the assessment:	midterm grade					

The curriculum	
Goal:	The aim of the course is to provide students with a thorough understanding of the purpose, methods, and underlying rationale of data preparation, preprocessing, and transformation tasks, which are essential to the proper functioning of data models and data analysis workflows. An additional objective of the course is to examine key issues in data collection, storage, and loading. Following the theoretical presentation of the methods introduced in the lectures and an explanation of their practical importance, the practical sessions enable students to gain hands-on experience by implementing these methods in Python.
Course description:	The course syllabus is structured around the four main stages of data preprocessing—data cleaning, data transformation, data integration, and data reduction—and covers the following topics: CRISP-DM; data loading; data exploration; handling missing data; outlier detection and treatment; data manipulation; data transformation; normalization; aggregation; dimensionality reduction and expansion; data compression; discretization; data quality assurance; and an overview of the tools and techniques used to implement these tasks.

Semester schedule	
Semester week	Topic
1.	Theory: General introduction to the CRISP-DM methodology; key aspects of problem identification and the collection of relevant data. Practice: The Pandas module, Pandas DataFrame, and main data types. (python review1)
2.	Theory: Practical aspects of data preparation, modeling, model evaluation, and deployment within the CRISP-DM methodology. Practice: Data loading from various data sources (relational databases and other common source types). (python review 2)
3.	Theory: Descriptive statistics. Data types and measurement scales (nominal, ordinal, interval, and ratio scales); characteristics of time series. Practice: Basic descriptive statistical tasks in a Pandas environment; fundamentals of data visualization using Matplotlib. (Python in data preperation part 1)
4.	Theory: Data transformation (normalization, standardization, discretization, feature derivation, aggregation). Practice: Data transformations and transformation operations in Pandas; one-hot encoding and scaling techniques. (python in data preperation part 2)
5.	Theory: Data merging, data integration, and data reduction (sampling, dimensionality reduction). Practice: Data integration and data reduction tasks in a Pandas environment.(python modules)
6.	Theory: Data preprocessing and levels of data cleaning. Level 1 data cleaning: establishing primary data structures and formats, indexes, and column names.

	Practice: Basic data cleaning tasks in a Pandas environment.
7.	Theory: Level 2 data cleaning: information extraction; transformation of data structures and data formats. Practice: Data transformations and data cleaning using SQL statements.
8.	Theory: Level 3 data cleaning: handling missing, outlier, and erroneous values. Practice: Handling missing data in a Pandas environment.
9.	Theory: Data quality assurance (data quality dimensions and metrics). Practice: Introduction to NumPy; operations with NumPy arrays.
10.	Theory: Introduction to Big Data technologies. Practice: Data preparation and processing in Big Data environments; Big Data file formats.
11.	Theory: Batch and stream processing in Big Data environments; data protection issues, pseudonymization, and anonymization. Practice: Databricks platform; batch and stream processing examples; basic data preparation tasks on the Databricks platform; demonstration of a modeling example in a Big Data environment (Apache Spark MLlib).
12.	Written theoretical midterm exam
13.	Midterm retake opportunity

Mid-semester requirements	
Requirements for obtaining the midterm grade / course signature:	A minimum of 36 points must be achieved in the theoretical midterm exam, and at least 20 points must be obtained from the points awarded for the semester assignment.

Written midterm exams

Semester week	Topic
12.	Written theoretical midterm exam
13.	Midterm retake opportunity

Method for determining the midterm grade

The final score determining the course grade is calculated as the sum of the following two components, according to the specified point thresholds:
The score achieved in the theoretical midterm exam (maximum 70 points)
The score awarded for the semester assignment (maximum 30 points)

Method of retaking the exam

Method for retaking the midterm exam / midterm grade / course signature:	The theoretical midterm exam may be retaken in the final week of the semester and at the beginning of the examination period. Submission of the semester assignment may be made up at the beginning of the examination period.

Type of the exam (to be filled out only for subjects with exams)

not relevant

Calculation of the exam mark (to be filled only for subjects with exams)

not relevant

Point thresholds for each grade:

88–100 points: excellent (5)
76–87 points: good (4)
64–75 points: satisfactory (3)
52–63 points: pass (2)
0–51 points: fail (1)

Literature

Mandatory:	Presentation materials related to the lectures and practical work 1. Roy Jafari: Hands-On Data Preprocessing in Python, Packt Publishing, 2022, ISBN: 1801072132
Recommended:	1. Jake VanderPlas: Python Data Science Handbook, O'Reilly Media, 2022, ISBN: 1098121228 2. Ofer Mendelevitch, Casey Stella, Douglas Eadline: Practical Data Science with Hadoop and Spark, Addison-Wesley Professional, 2016, ISBN: 0134024141
Other:	Uploaded materials to Moodle