



Subject card

Subject name and code	BIG DATA ANALYSIS - DATA PROCESSING METHODS, PG_00070144						
Field of study	InfoBioChem						
Date of commencement of studies	February 2027		Academic year of realisation of subject		2027/2028		
Education level	second-cycle studies		Subject group		Obligatory subject group in the field of study Subject group related to scientific research in the field of study		
Mode of study	Full-time studies		Mode of delivery		at the university		
Year of study	1		Language of instruction		Polish		
Semester of study	2		ECTS credits		3.0		
Learning profile	general academic profile		Assessment form		exam		
Conducting unit	Department of Biotechnology and Microbiology -> Faculty of Chemistry -> Faculties of Gdańsk University of Technology						
Name and surname of lecturer (lecturers)	Subject supervisor		dr inż. Paweł Wityk				
	Teachers						
Lesson types	Lesson type	Lecture	Tutorial	Laboratory	Project	Seminar	SUM
	Number of study hours	15.0	0.0	30.0	0.0	15.0	60
	E-learning hours included: 0.0						
Learning activity and number of study hours	Learning activity	Participation in didactic classes included in study plan		Participation in consultation hours		Self-study	SUM
	Number of study hours	60		2.0		13.0	75
Subject objectives	The aim of the course is to introduce students to the concept of Big Data and the principles of properly acquiring, organizing, describing, storing, preparing, and analyzing large biological datasets. Using biological/biochemical data as examples, students will learn about the structure of a valid dataset, the importance of file formats and metadata, quality control and filtering methods, as well as the fundamentals of descriptive statistics and statistical inference. Students will acquire practical skills in data organization and preprocessing, performing statistical analyses using JASP, and processing and interpreting data with open-source software. The course also aims to develop skills in clearly presenting and critically discussing data obtained during the execution of a master's thesis.						

Learning outcomes	Course outcome	Subject outcome	Method of verification
	[K7_K05] is ready for continuous learning, developing interdisciplinary knowledge, and updating their skills in a dynamically changing scientific and professional environment.	is prepared to independently update their knowledge regarding the management and analysis of biological data, critically evaluate data quality and the methods employed, and reliably present the results of their own research.	[SK1] Assessment of group work skills [SK2] Assessment of progress of work [SK4] Assessment of communication skills, including language correctness
	[K7_W07] has a thorough knowledge and understanding of the structure and operating principles of components and systems related to the field of study, including theories, methods, and complex relationships between them, as well as selected specific issues relevant to the curriculum InfoBioChem	is familiar with the structure of a typical biological dataset and understands the significance of metadata, identifiers, control samples, and replication, as well as the impact of missing data, outliers, filtering, and data transformation on analysis results.	[SW2] Assessment of knowledge contained in presentation [SW3] Assessment of knowledge contained in written work and projects
	[K7_W05] has knowledge of theoretical computational and computer methods used to solve problems appropriate to the curriculum InfoBioChem.	knows and understands basic concepts related to Big Data, principles of data description and storage, commonly used file formats, and the fundamentals of descriptive statistics and statistical inference used in the analysis of biological data.	[SW3] Assessment of knowledge contained in written work and projects [SW2] Assessment of knowledge contained in presentation
	[K7_U07] can apply machine learning and artificial intelligence methods to analyze large sets of biological data.	is able to prepare and filter data, calculate and interpret basic statistics, select and perform the appropriate analysis, and clearly present the results using JASP or similar software. They are able to discuss the data obtained during the preparation of their master's thesis and formulate conclusions based on the analysis.	[SU1] Assessment of task fulfilment [SU2] Assessment of ability to analyse information [SU3] Assessment of ability to use knowledge gained from the subject [SU4] Assessment of ability to use methods and tools

Subject contents	Course content – lecture		
	<div>1. The concept of Big Data: characteristics of large datasets, data sources, and examples of applications in biology, biotechnology, and chemistry.</div> <div>2. Qualitative and quantitative data; observations, variables, types, and measurement scales.</div> <div>3. Organizing a proper data table: one row per observation, one column per variable; variable names, identifiers, and data dictionaries.</div> <div>4. Metadata and dataset documentation: data provenance, sample descriptions, units, measurement conditions, dates, and versions.</div> <div>5. Data formats and their applications: CSV, TSV, XLSX, TXT, JSON, and specialized formats; principles of import, export, and conversion.</div> <div>6. Data storage, organization, and sharing: directory structure, file naming, versioning, backups, security, and data integrity.</div> <div>7. FAIR principles, analysis reproducibility, and sound research data management.</div> <div>8. Data quality control: completeness, consistency, duplicates, recording errors, missing values, and outliers.</div> <div>9. Filtering and preparing data for analysis; the impact of cleaning, transformation, normalization, and scaling on the results.</div> <div>10. Descriptive statistics: count, mean, median, quartiles, minimum, maximum, variance, standard deviation, and coefficient of variation.</div> <div>11. Describing data distribution and selecting visualizations: tables, histograms, box plots, scatter plots, and bar charts.</div> <div>12. Fundamentals of statistical inference: population and sample, hypotheses, p-value, confidence interval, effect size, and test assumptions.</div> <div>13. Comparing groups and analyzing relationships: basic parametric and non-parametric tests, correlation, and regression.</div> <div>14. Specifics of metabolomic data as an example of multidimensional data; Multiple comparisons, PCA, and the fundamentals of biological interpretation.</div> <div>15. Correct reporting, presentation, and interpretation of data; common errors and misuses.</div>		
	Course content – laboratory		
	<div>1. Preparing a proper data table: variable names, identifiers, units, metadata, and group coding.</div> <div>2. Saving and converting XLSX, CSV, and TXT files; decimal separators, character encoding, and import validation.</div> <div>3. Data organization: sorting, filtering, identifying duplicates and errors, marking missing values, and using basic formulas.</div> <div>4. Calculating mean, median, standard deviation, coefficient of variation, and other descriptive statistics in JASP.</div> <div>5. Creating summary tables and basic charts; selecting the presentation method based on data type.</div> <div>6. Importing data into JASP, defining variable types, and applying filters.</div> <div>7. Descriptive statistics and data distribution assessment in JASP; identifying outliers.</div> <div>8. Comparing two or more groups in JASP; checking assumptions and selecting parametric or non-parametric tests.</div> <div>9. Correlation analysis and regression basics in JASP; interpreting tables, charts, p-values, confidence intervals, and effect sizes.</div> <div>10. Preparing biological data for import: layout of samples, groups, and metabolites, and handling missing values.</div> <div>11. Data filtering, imputation, normalization, transformation, and scaling in MetaboAnalyst.</div> <div>12. Univariate comparisons, multiple testing correction, and result visualization in MetaboAnalyst.</div> <div>13. PCA, cluster analysis, and heatmaps in MetaboAnalyst; interpreting similarities, differences, and outliers.</div> <div>14. Enrichment and metabolic pathway analysis; formulating biological conclusions.</div> <div>15. Developing a complete example: from the data table through cleaning and analysis to exporting results, charts, and a brief report.</div>		
	Course content – seminar		
	<div>1. Preparation of a presentation on data obtained during the master's thesis project.</div> <div>2. Presentation of the research problem, data acquisition methods, dataset type and size, and storage formats and methods used.</div> <div>3. Description of the data structure, variables, metadata, research groups, replication, and any control samples.</div> <div>4. Discussion of quality control procedures, filtering, handling of missing values, and data preparation for analysis.</div> <div>5. Presentation of descriptive statistics, tests, tables, and charts used, along with a justification for the chosen methods.</div> <div>6. Interpretation of results in the context of the research question and current knowledge in the field.</div> <div>7. Discussion of dataset and analysis limitations, potential sources of error, and possible next steps.</div> <div>8. Discussion of the presented data, feedback from the group and the supervisor, and refinement of the communication of scientific results.</div>		
Prerequisites and co-requisites	Basic knowledge of biology, biochemistry, and mathematics. Basic ability to work with computers, files, and spreadsheet software. Ability to read scientific literature in English. No prior knowledge of JASP, or programming is required. Students should have access to data obtained during their master's thesis; if the dataset is not yet complete, partial or pilot data may be used. Only appropriately anonymised data approved for educational use may be used during classes.		
Assessment methods and criteria	Subject passing criteria	Passing threshold	Percentage of the final grade
	Exam (lecture)	60.0%	40.0%
	Laboratory classes (report)	60.0%	40.0%
	Seminar (presentation)	60.0%	20.0%

Recommended reading	Basic literature	James G., Witten D., Hastie T., Tibshirani R., Taylor J., An Introduction to Statistical Learning: with Applications in Python, Springer, 2023 Tosi, Davide, Redon Kokaj, and Marco Roccetti. "15 years of Big Data: a systematic literature review." <i>Journal of Big Data</i> 11.1 (2024): 73. https://jasp-stats.org/ https://www.metaboanalyst.ca/
	Supplementary literature	Luo, Senlin, and Limin Pan. <i>Big Data Analysis: Theory and Technology</i> . Springer Nature, 2026.
	eResources addresses	
Example issues/ example questions/ tasks being completed	<ol style="list-style-type: none"> 1. Detection of errors, duplicates, missing values, and outliers in a dataset. 2. Calculation and interpretation of the mean, median, standard deviation, coefficient of variation, and basic graphs. 3. Filtering, normalisation, transformation, and scaling of metabolomics data. 4. Selection and performance of basic statistical analyses in JASP: group comparisons, correlation, and regression. 5. Interpretation of PCA results, heat maps, and statistical comparison results. 6. Presentation and discussion of data obtained during the master's thesis: data structure, methods used, results, limitations, and conclusions. 	
Practical activities within the subject	Not applicable	

Document generated electronically. Does not require a seal or signature.