DAMA600 Module Outline

1. GENERAL

SCHOOL	School of Science and Technology		
PROGRAMME	Data Science and Machine Learning		
LEVEL OF STUDIES	Level 7 of the Hellenic and European Qualifications Framework		
MODULE CODE	DAMA600	**SEMESTER**	2/3
MODULE TITLE	Mining of Massive Datasets		
INDEPENDENT TEACHING ACTIVITIES *if credits are awarded for separate components of the course, e.g. lectures, laboratory exercises, etc. If the credits are awarded for the whole of the course, give the weekly teaching hours and the total credits*		**HOURS**	**CREDITS**
Weekly workload: 32-33 hours x 13 weeks		420	15 ECTS
MODULE TYPE *Compulsory/Elective/ Mandatory Optional*	Compulsory		
PREREQUISITE MODULES	None		
LANGUAGE OF INSTRUCTION and EXAMINATIONS	English		
IS THE MODULE OFFERED TO ERASMUS STUDENTS	Yes		
MODULE WEBSITE (URL)	The Module has a dedicated space in HOU's digital learning platform (http://courses.eap.gr, http://study.eap.gr), which students and tutors can access using their credentials.		

2. LEARNING OUTCOMES

Learning outcomes

The learning outcomes, specific knowledge, skills and competences of an appropriate level, which the students will acquire with the successful completion of the Module are described.

Consult Appendix A

- *Description of the level of learning outcomes for each qualifications cycle, according to the Qualifications Framework of the European Higher Education Area*
- *Descriptors for Levels 6, 7 & 8 of the European Qualifications Framework for Lifelong Learning and Appendix B*
- *Guidelines for writing Learning Outcomes*

Knowledge:

Upon successful completion of the Module, students will be able to:

- Describe the challenges of mining large-scale datasets and discuss the relevant computing architectures.
- Define and apply similarity measures, and use techniques such as shingling and minhashing for data comparison.
- Design locality-sensitive hashing (LSH) schemes to perform efficient similarity search.
- Use and evaluate scalable algorithms for frequent itemset mining.

Skills:

Upon successful completion of the Module, students will be able to:

- Analyze and model large-scale graph data using algorithms like PageRank and community detection.
- Implement data stream processing algorithms using sampling and sketching methods (e.g., Count-Min Sketch).
- Apply clustering techniques adapted for large datasets, including K-means and hierarchical clustering methods.
- Develop scalable recommender systems using collaborative filtering and matrix factorization.

Competences:

Upon successful completion of the Module, students will be able to:

- Understand dimensionality reduction methods including SVD and CUR decompositions.
- Describe and implement scalable classification methods for large data (e.g., decision trees, naïve Bayes).
- Apply machine learning algorithms in a distributed computing framework like MapReduce.
- Evaluate the efficiency, scalability, and applicability of massive data mining techniques in real-world scenarios.

General Competences

Taking into consideration the general competences that the degree-holder must acquire (as these appear in the Diploma Supplement and appear below), at which of the following does the Module aim?

Search for, analysis and synthesis of data and information, with the use of the necessary technology
Adapting to new situations
Decision-making
Working independently
Team work
Working in an international environment
Working in an interdisciplinary environment
Production of new research ideas
Project planning and management
Respect for difference and multiculturalism
Respect for the natural environment
Showing social, professional and ethical responsibility and sensitivity to gender issues
Criticism and self-criticism
Production of free, creative and inductive thinking

The general skills that the students will acquire are:

- Search for, analysis and synthesis of data and information, with the use of the necessary technology
- Adapting to new situations
- Decision-making
- Working independently
- Team work
- Project planning and management
- Showing social, professional and ethical responsibility and sensitivity to gender issues
- Production of free, creative and inductive thinking

3. SYLLABUS

Purpose of Module

This module equips students with specialized knowledge in mining and analyzing massive datasets, focusing on scalable algorithms and big data frameworks. Unlike traditional data science courses, it emphasizes techniques designed to handle data that exceeds the capacity of main memory and must be processed using distributed systems. Students will explore the architecture and principles of systems like MapReduce and Spark, which support large-scale data processing. They will learn methods for efficient similarity search, including minhashing and locality-sensitive hashing, tailored to high-dimensional data. The course covers algorithms for mining frequent patterns and association rules at scale, going beyond conventional in-memory approaches.

In the context of streaming data, students will understand models and techniques for real-time processing, such as sketches and approximate counting. A strong emphasis is placed on mining structured data like graphs, where students will study PageRank, HITS, community detection, and triangle counting—especially relevant in web and social network analysis. The course also introduces scalable recommendation systems using collaborative filtering and matrix factorization methods. Techniques for dimensionality reduction, such as CUR decompositions and random projections, are discussed with an emphasis on their scalability and suitability for large datasets. Machine learning content focuses on the efficient implementation of classification and clustering algorithms for massive datasets.

Students will also examine how to design algorithms under resource constraints, and how trade-offs in approximation, speed, and accuracy are managed at scale. Throughout the course, theoretical foundations are paired with practical assignments involving large datasets and distributed environments. Unlike DAMA510, which focuses on statistical models and introductory machine learning, this course prioritizes the engineering and algorithmic challenges of working with truly massive data. By the end of the module, students will be capable of designing, implementing, and evaluating scalable data mining pipelines using contemporary frameworks.

4. TEACHING and LEARNING METHODS - EVALUATION

<table>
<tr><td>DELIVERY
Face-to-face, Distance learning, etc.</td><td>- Distance teaching and learning with three (3) Group Counseling Meetings (GCMs) of 4-hour duration during the academic semester on weekends.
- Personal communication and feedback (advisory role of Adjunct Faculty).</td></tr>
<tr><td>USE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY
Use of ICT in teaching, laboratory education, communication with students</td><td><u>During GCMs and teaching the following are used</u>:
- Remote meetings tools (webex, Teams),
- Presentation software (PowerPoint, educational video - animations etc.),
- Specialized software/databases for the subjects under study.
In addition, students use office automation tools, web browsers and e-reader for digital books.
<u>Communication with the students is supported by</u>:
- The digital platform of HOU ((<u>https://courses.eap.gr/login/index.php</u> / <u>https://study.eap.gr/login/index.php</u>) (course information, educational material posts, announcements, messages, examination results, user groups, discussion forums etc.).
- e-mail and messaging.</td></tr>
<tr><td>TEACHING METHODS
The manner and methods of teaching are described in detail.
Lectures, seminars, laboratory practice, fieldwork, study and analysis of bibliography, tutorials, placements, clinical practice, art workshop, interactive teaching, educational visits, project, essay writing, artistic creativity, etc.
The student's study hours for each learning activity are given as well as the hours of non-directed study according to the principles of the ECTS</td><td>
<table>
<tr><th>Activity</th><th>Semester Workload</th></tr>
<tr><td>3 GCMs (x 4 hours)</td><td>12</td></tr>
<tr><td>2 Educational Activities (x 10 hours)</td><td>20</td></tr>
<tr><td>2 Semester Assignments (x 30 hours)</td><td>60</td></tr>
<tr><td>Individual Study time (25 hours x 13 weeks)</td><td>325</td></tr>
<tr><td>Final examination</td><td>3</td></tr>
<tr><td>Total Workload</td><td>420</td></tr>
</table>
</td></tr>
<tr><td>STUDENT PERFORMANCE EVALUATION
Description of the evaluation procedure
Language of evaluation, methods of evaluation, summative or conclusive, multiple choice questionnaires, short-answer questions, open-ended questions, problem solving, written work, essay/report, oral examination, public presentation, laboratory work, clinical examination of patient, art interpretation, other
Specifically-defined evaluation criteria are given, and if and where they are accessible to students.</td><td><u>Students' evaluation – Grade assessment of a Module</u>:
a. Two (2) multiple-choice (quiz) Educational Activities (Q), which contribute equally to the final grade with a value of 5% each.
b. Two (2) Semester Assignments (A) which contribute equally to the final grade with a value of 10% each.

The scoring of educational activities and assignments is activated only if the student succeeds an overall score equal to or above the base (≥50%) in the final or repeat exams.

c. Final or repeat exams (E) contribute to the final grade of the module by 70%.

The Final Grade of the module is calculated as follows (with 10 being the maximum Grade):

Final Grade = (Q1 x 5%) + (Q2 x 5%) + (A1 x 10%) + (A2 x 10%) + (E x 70%)

<u>Language of evaluation</u>: English</td></tr>
</table>

5. INDICATIVE BIBLIOGRAPHY

- Recommended bibliography:

- J. Leskovec, A. Rajaraman & J.D. Ullman (2020). Mining of Massive Datasets (3rd edition). Cambridge University Press
- P.-N. Tan, M. Steinbach, A. Karpatne & V. Kumar (2021). Introduction to Data Mining (2nd Edition). Pearson.
- R. Zafarani, M.A. Abbasi & H. Liu (2014). Social media mining: an introduction. Cambridge University Press.

Additional digital (and multimedia) material will be made available online.