SYLLABUS

1. Data about the program of study

1.1	Institution	Technical University of Cluj-Napoca
1.2	Faculty	Automation and Computer Science
1.3	Department	Computer Science
1.4	Field of study	Computer Science and Information Technology
1.5	Cycle of study	Master of Science
1.6	Program of study / Qualification	Cybersecurity Engineering / Master
1.7	Form of education	Full time
1.8	Subject code	9.1

2. Data about the subject

2.1	Subject name	***Big Data and Machine Learning for Cybersecurity***
2.2	Course responsible/lecturer	Conf.dr.ing. Camelia LEMNARU - camelia.lemnaru@cs.utcluj.ro
2.3	Teachers in charge of seminars	Conf.Dr.ing. Ciprian OPRIȘA - ciprian.oprisa@cs.utcluj.ro

2.4 Year of study	I	2.5 Semester	2	2.6 Type of assessment (E - exam, C - colloquium, V - verification)	E

2.7 Subject category	Formative category: DA – advanced, DS – speciality, DC – complementary	DS
	Optionality: DI – imposed, DO – optional (alternative), DF – optional (free choice)	DO

3. Estimated total time

3.1 Number of hours per week	4	of which	3.2 Course	2	3.3 Seminar	0	3.3 Laboratory	2	3.3 Project	0
3.4 Total hours in the curriculum	56	of which	3.5 Course	28	3.6 Seminar	0	3.6 Laboratory	28	3.6 Project	0

3.7 Individual study:	
(a) Manual, lecture material and notes, bibliography	32
(b) Supplementary study in the library, online and in the field	18
(c) Preparation for seminars/laboratory works, homework, reports, portfolios, essays	43
(d) Tutoring	0
(e) Exams and tests	2
(f) Other activities	0

3.8 Total hours of individual study (sum (3.7(a)...3.7(f)))	94
3.9 Total hours per semester (3.4+3.8)	150
3.10 Number of credit points	6

4. Pre-requisites (where appropriate)

4.1	Curriculum	Data bases
4.2	Competence	Statistics and probabilistic calculus

5. Requirements (where appropriate)

5.1	For the course	blackboard, beamer, computers
5.2	For the applications	blackboard, beamer, computers

6. Specific competences

Professional competences	**C2. Investigate and analyze cyber-criminality actions and malware using advanced methods such as reverse engineering and behavior monitoring.** • **C2.1.** Advanced knowledge of classifications and characteristics of different cybersecurity attacks and malware. • **C2.2.** Be able to analyze and understand new kinds of malware, the new techniques they use to attack, gain persistence, escalate privileges etc., and be able to compare them with known attack techniques. • **C2.4.** Capability to identify and assess theoretical and practical limitations of existing automatic malware analysis tools and propose improvements, where and if possible. • **C2.5.** Capability to derive new classes of attacks and exploitation techniques, supposed to be used by new malware, and propose the appropriate methods to identify and classify them correctly. **C4. Design and develop highly secure software, security solutions and tools.** • **C4.2.** Be able to identify new situations and scenarios when it is needed to develop a new cybersecurity solution or use an existing one. Be able to analyze proposed cybersecurity solutions and compare them with existing ones. • **C4.3.** Capability to develop complex secure software, complying with recommended good practices of built-in security and secure coding. Capability to develop software tools used for cybersecurity pentesting and assessment. • **C4.5.** Capability to develop software modules and tools that could provide a high degree of cybersecurity. Capability to propose new methods to assess the cybersecurity of computing systems and devices and ways to improve it. **C5. Develop rigorous and efficient security solutions to complex real-life problems and situations. Be able to use security mathematical tools and models, engineering approaches and technologies specific and appropriate for the information and computing system security field.** • **C5.1.** Knowledge of complex relationship between cybersecurity and real-life aspects. Knowledge of mathematical theory some cybersecurity mechanisms and solutions are based on. • **C5.2.** Be able to analyze and understand new complex real-life scenarios from the cybersecurity perspective. Be able to identify needed cybersecurity solutions and derive new ones for new particular cases. • **C5.3.** Capability to apply mathematical and computer engineering theoretical models to analyze, assess and address real-life cybersecurity and privacy issues and challenges. • **C5.5.** Capability to run research activities and projects aimed to derive applicable cybersecurity solutions, implement their hardware and/or software prototype.
Cross competences	N/A

7. Discipline objectives (as results from the *key competences gained*)

7.1	General objective	Acquiring the ability to analyse large datasets. Considering the increasing number of malicious programs in the wild, the goal is to learn how to handle large collections of data, design, implement and evaluate malware detection and classification models.
7.2	Specific objectives	1. Acquire the ability to use scripting languages and databases to handle large datasets. 2. Design and implement distributed systems, understand and use the Map-Reduce paradigm. 3. Understand and learn algorithms and techniques for searching in large collections of data. 4. Understand and learn Machine Learning algorithms suitable for malware classification and detection

8. Contents

8.1. Lecture (syllabus)	Number of hours	Teaching methods	Notes

Introduction to scripting languages: Python	2	Blackboard illustrations and explanations, beamer presentations, discussions, short challenges	
Data processing in Python	2		
Relational and non-relational databases: ACID properties, relational algebra, CAP theorem	2		
Map-Reduce – paradigm and execution framework	2		
Map-Reduce algorithms complexity	2		
Simple search techniques for big data: indexing, hashing	2		
Advanced search techniques for big data: similarity search, determine program similarity	2		
Advanced search techniques for big data: reversed index, locality-sensitive hashing	2		
Link analysis: Page Rank, SEO techniques	2		
Clustering techniques: K-means, hierarchical clustering	2		
Advanced clustering techniques for big data	2		
Building prediction models: linear/logistic regression, decision trees	2		
Advanced classifiers: Support Vector Machines, perceptron, Neural Networks	2		
Dimensionality reduction	2		

Bibliography

1. Mining of Massive Datasets (Rajarman, Anand – 2011 – Cambridge)
2. Pattern Recognition and Machine Learning (Bishop, Christopher – 2007 – Springer)
3. MongoDB: The Definitive Guide (Chodorow, Kristina – 2013 – O'Reilly) (2nd ed)
4. Data Science for Business: What you need to know about data mining and data-analytic thinking (Provost, Foster – 2013 – O'Reilly)
5. Learning Python (Lutz, Mark – 2013 – O'Reilly) (5th ed)
6. Research papers, list provided at the beginning of each semester

8.2. Seminar / Laboratory / Project	Number of hours	Teaching methods	Notes
Introduction to Python	2	Brief reviews, blackboard illustrations and explanations, tutorials, roadmaps, short live demos and guidance of code development, discussions, homework	
Data structures in Python	2		
Specific libraries and functions for working with collections of data	2		
Feature extraction for potentially malicious programs	2		
Data storage and access: databases, indexing	2		
Map-Reduce algorithms for malware processing	2		
Computing program similarity	2		
Building a reversed index, using Map-Reduce	2		
Search for similar programs in large collections of data: locality sensitive hashing	2		
Clustering techniques for malware detection: part 1	2		
Clustering techniques for malware detection: part 2	2		
Spam and malware classifiers: part 1	2		
Spam and malware classifiers: part 2	2		
Evaluation and verification	2		

Bibliography

1. Mining of Massive Datasets (Rajarman, Anand – 2011 – Cambridge)
2. Pattern Recognition and Machine Learning (Bishop, Christopher – 2007 – Springer)

3. MongoDB: The Definitive Guide (Chodorow, Kristina – 2013 – O'Reilly) (2nd ed)
4. Data Science for Business: What you need to know about data mining and data-analytic thinking (Provost, Foster – 2013 – O'Reilly)
5. Learning Python (Lutz, Mark – 2013 – O'Reilly) (5th ed)
6. Research papers, list provided at the beginning of each semester

9. Bridging course contents with the expectations of the representatives of the community, professional associations, and employers in the field

This aspect will be achieved by recurrent discussions with the relevant industry employers (cybersecurity domain). Big Data courses are delivered within other master programs, but very few focus on computer and information security. Both malware and spam detection and classification require, from a practical standpoint, working with large collections of data, which requires big data analysis and machine learning. For example, there are several master programs which teach big data and business analytics, teaching methods which can be successfully applied to the data/computer security domain:

- *Big Data*, Masters in Computer and Information Security, University of Liverpool, UK http://www.liv.ac.uk/study/online/programmes/information-technology/msc-computer-and-information-security/module-details/
- *Big Data Management and Security*, Graduate Certificate Program, Missouri University of Science and Technology, USA, http://dce.mst.edu/credit/certificates/bigdatamanagementandsecurity/
- CS246, *Mining Massive Data Sets*, Stanford, http://web.stanford.edu/class/cs246/
  CSE 599, *Machine Learning for Big Data*, Computer Science & Engineering, University of Washington

10. Evaluation

Activity type	10.1 Assessment criteria	10.2 Assessment methods	10.3 Weight in the final grade
10.4 Course	Ability to define and explain concepts and methods specific to course's field. Attendance frequency, interest, and interactivity during lecture classes.	Written exam, including online quiz tests (e.g. on Moodle platform) and presentation(s) of different subjects / paper in the course's field during semester time.	60%
10.5 Laboratory	Capability and ability to give correct and functional solutions to problems specific to course's field. Attendance frequency, interest, and interactivity during lecture classes.	Evaluate lab activity. Evaluate lab assignments (homework). Evaluate solutions of problems given in a final lab exam.	40%

10.6 Minimum standard of performance

***Lecture*.** Attending **minimum 50%** of lecture classes, to be allowed to take the final examination. Minimum final grade must be 5 for the exam to be considered passed.

***Lab*.** Attending **all lab classes** (one lab could be recovered during the semester, and one more during re-examination sessions). Minimum lab grade must be 5 to be allowed at final exam.

By the end of the course, the students should be able to work with big datasets, both structured and unstructured, using sequential and distributed algorithms (e.g. Map-Reduce). The main operations students should have assimilated are: search in large collections of data, classification and clustering, building and evaluating prediction models.

Date of filling in:		Title Surname Name	Signature
	Lecturer	Conf.dr.ing. Camelia LEMNARU	
	Teachers in charge of application	Conf.dr.ing. Ciprian OPRIȘA	

Date of approval in the Computer Science Department 20.02.2024	Head of department Prof.dr.ing. Rodica Potolea
Date of approval in the faculty of Automation and Computer Science 22.02.2024	Dean Prof.dr.ing. Mihaela Dinsoreanu