GEORGE MASON UNIVERSITY | Department of Health Administration and Policy

College of Health and Human Services

Spring 2018

Syllabus	
Course information	HAP 880: Advanced Health Data Mining
Course placement	**( )** Core **(X)** Concentration **(X)** Elective in HSR PhD (Knowledge Discovery and Health Informatics Concentration), Health Informatics MS (Data analytics Concentration), Health Informatics and Data Analytics graduate certificate. **(X)** Pre-requisite(s): HAP 618, HAP 719, HAP 780 or permission of instructor This is a follow up course to HAP 780. It includes advanced topics that require both knowledge of basic data mining techniques and statistical methods. Before enrolling, students must be fluent in SQL, know basic data mining concepts, and have basic programming skills in Python.
Instructor	Janusz Wojtusiak PhD jwojtusi@gmu.edu Northeast Module, Room 108, Fairfax Campus
Course description	An advanced course in health data mining. Includes knowledge and skills needed to analyze health data using modern tools. Describes analytics of administrative and clinical data. Covers concepts and tools for big data analytics and NoSQL data analytics.
Course objectives	**Upon completion of the course, students will be able to:** 1. Understand application of data mining and analytics techniques applied to administrative and clinical data 2. Analyze data using modern open source analytics tools 3. Understand big data platforms and tools needed for analyzing health data 4. Apply data mining tools to administrative and clinical data 5. Use selected tools for NoSQL data analysis, including graph data

Required textbook(s) and/or materials	**Required Text:** Class notes and slides. **Additional Readings:** Sebastian Raschka, Python Machine Learning, Packt Publishing - ebooks Account (September 2015). Antonio Gulli, A collection of Data Science Interview Questions Solved in Python and Spark: Hands-on Big Data and Machine Learning (A Collection of Programming Interview Questions) (Volume 6), CreateSpace Independent Publishing Platform. Van Bruggen, Rik. Learning Neo4j. Packt Publishing Ltd, 2014. (can be found for free online)
Course requirements	**Computer requirements** This is a computationally intensive course and you are expected to access databases, software tools, and other contents. You will need: • Fast computer (multicore PC or Mac) with at least 50GB of free disk space and at least 4GB RAM (8GB+ recommended), Windows 7 or newer. Mac users may require more powerful computers to enable virtualization to run windows. • Fast internet connection • Microsoft office for viewing and preparing files • Other software will be provided in class (most likely: Weka, Python/Anaconda, PostgreSQL, Neo4j) If you do not have sufficient computer, you can request access to Health Informatics Learning Lab, located in Northeast Module, or use one of computer labs at GMU. It is responsibility of students to configure and maintain own computers, make sure that it is set up correctly and installed software (i.e., security) do not interfere with software used in class. **Expectations: Students are responsible for assigned readings, class content and material. Students are also responsible for finding right computer equipment that allows accessing the course materials, using data and software tools, and for checking email/blackboard on daily basis.** **Data mining is a very broad topic, which is condensed here into one semester course. This course requires students to participate in lectures and spend at least another 3-6 hours per week on assignments, reading, and project.** **Evaluation Methods:** If you are taking this course as part of a graduate level course, you will receive a grade. Your grade will depend on your participation, quality of your project work and your team work. Assignments and projects are graded based on multiple criteria that will be discussed in detail.

Always write all answers in own words. Do not copy-and-paste.

You can ask questions by sending email to the instructor. In most cases you will receive response within 48 hours.

Final Project

Data mining requires combining theoretical knowledge with practical skills. In order to develop skills in the context of health care applications, semester-long project is the most important component of the grade.

The project topics should be related to analyzing healthcare data in order to solve clinical or administrative problems. The project should include, but be not limited to: (1) problem description; (2) data selection; (3) data pre-processing; (4) selection DM methods; (5) application of methods; (6) analysis of results; (7) review of available literature and related work; (7) conclusions and description of impact on healthcare. Brief description of what you learned in the project. The project should be based on tools in which you need to 'write code' to analyze data, i.e., Python, R, not tools with GUI such as Weka or RapidMiner.

Direct application of existing software to publically available datasets is not sufficient. The projects must demonstrate significant efforts in data manipulation, processing, and mining. Projects must also illustrate understanding of applied techniques as well as healthcare problem.

Data Access

Access to most data requires completion of CITI training (www.citiprogram.org). After creating an account and selecting GMU as organization, complete the Group 2 Biomedical Research Investigators course. Several datasets are available in the DSHI center at GMU and access requires additional background check.

Teaching methods	(X) Lecture ( ) Group work ( ) Independent research ( ) Field work ( )Papers ( ) Guest speakers ( ) Student presentations ( ) Case Studies (X) Lab ( ) Class discussion ( ) Other _____
Evaluation	Weekly Assignments 40% Semester-long project 50% Participation 10%
Grading Scale	96+ A 90-95 A - 86-89 B + 80-85 B 75-79 B - 70-74 C 0-70 F

Mason Honor Code	The complete Honor Code is as follows: *To promote a stronger sense of mutual responsibility, respect, trust, and fairness among all members of the George Mason University community and with the desire for greater academic and personal achievement, we, the student members of the university community, have set forth this honor code:* **Student members of the George Mason University community pledge not to cheat, plagiarize, steal, or lie in matters related to academic work.** *(From the 2017-18 Catalog – catalog.gmu.edu)*
Individuals with Disabilities	The university is committed to providing equal access to employment and educational opportunities for people with disabilities. Mason recognizes that individuals with disabilities may need reasonable accommodations to have equally effective opportunities to participate in or benefit from the university educational programs, services, and activities, and have equal employment opportunities. The university will adhere to all applicable federal and state laws, regulations, and guidelines with respect to providing reasonable accommodations as necessary to afford equal employment opportunity and equal access to programs for qualified people with disabilities. Applicants for admission and students requesting reasonable accommodations for a disability should call the Office of Disability Services at 703-993-2474. Employees and applicants for employment should call the Office of Equity and Diversity Services at 703-993-8730. Questions regarding reasonable accommodations and discrimination on the basis of disability should be directed to the Americans with Disabilities Act (ADA) coordinator in the Office of Equity and Diversity Services. *(From the 2017-18 Catalog – catalog.gmu.edu)*
E-mail Policy	Web: masonlive.gmu.edu Mason uses electronic mail to provide official information to students. Examples include notices from the library, notices about academic standing, financial aid information, class materials, assignments, questions, and instructor feedback. Students are responsible for the content of university communication sent to their Mason e-mail account and are required to activate that account and check it regularly. Students are also expected to maintain an active and accurate mailing address in order to receive communications sent through the United States Postal Service. *(From the 2017-18 Catalog – catalog.gmu.edu)*
	I plan to record selected lectures for the future use of online students. The recorded videos will be posted online for students to view. The camera will be facing the screen and instructor. Because live interaction with class is recorded, some of your questions and voice may be also recorded. If you do not wish to be on the final recording, please do not say anything in class and let me know. I will ask for your help to review the final versions of recordings to ensure that you are completely edited out.

Tentative Weekly Schedule

The schedule below is approximate and may be changed to adapt to students' needs and requests, new material, and for other reasons. Due dates and assignments are subject to change and will be provided weekly.

Wk	Topics
1/24	Review of Data Mining Methods for Healthcare Applications Introduction to Data Mining in Python. Pandas, numpy, and sklearn, libraries.
1/31	Reading the data: text files, SQL, XML, JSON. Manipulating & exploring the data.
2/7	Classification & regression of claims data in sklearn
2/14	Classification & regression of claims data in sklearn
2/21	Sensitivity Analysis
2/28	Clustering & outlier analysis of claims data in sklearn
3/7	Building and applying models from clinical data
3/21	More clinical data, MIMIC
3/28	Introduction to Hadoop, HDFS, and Map-Reduce. Big Data in Healthcare Applications.
4/4	Data Mining Tools in Hadoop
4/11	Apache Spark, data mining in Hadoop
4/18	Graph Databases, Neo4j
4/25	Graph Databases, Neo4j
5/2	Student Project Presentations