Utrecht University

Course Syllabus: Data Wrangling and Data Analysis (INFOMDWR)

Department	Department of Information and Computing Sciences
Course title	Data Wrangling and Data Analysis (INFOMDWR)
Academic Quarter	Block 1
Quarter Start Date	07/09/2020
Quarter End Date	6/11/2020
Class Schedule	Lectures: Mon. and Tue. 9:00 – 10:45 Tutorials : Mon. and Tue. 13:15 – 15:00 Seminars : Thu. 13:15 – 15:00 Exam 1: Thu. 15:30 – 15:50 Exams 2 & 3: Fri. 15:30 – 15:50

Instructor(s)				
Name	Email (@uu.nl)	Phone (Office)	Office Location	Office Hours
Hakim Qahtan	a.a.a.qahtan	030 253 5407	BBG-461	--

Teaching Assistant(s)	
Name	Email (@uu.nl)

Course Information	
Course Objectives	In this this course, you will learn to: 1. Know, explain, and apply data retrieval from existing relational and nonrelational databases, including text, using queries build from primitives such as select, subset, and join both directly in, e.g., SQL and through a rjson interface. 2. Know, explain, and apply common data clean-up procedures, including missing data and the appropriate imputation methods and feature selection. 3. Know, explain, and apply methodology to properly set-up data analysis experiments, such as train, validate, and test and the bias/variance trade-off. 4. Know, explain, and apply supervised machine learning algorithms, both for classification and regression purposes as well as their related quality measures, such as AUC and Brier scores. 5. Know, explain, and apply non-supervised learning algorithms, such as clustering and (other) matrix factorization techniques that may or may not result in lower-dimensional data representations. 6. Be able to choose between the different techniques learned in the course and be able to explain why the chosen technique fits both the data and the research question best.
Course Description from Program Guide	Data do not fall from heaven, but are created, manipulated, transformed, and cleaned - in any data analysis, therefore, the

	treatment of the data itself is just as important as the modeling techniques applied to them. In this course, you will learn to perform predictive data analysis to gain insights for science and business applications, while simultaneously keeping track of where these data originated and handling them yourself. The course consists of two parts, data wrangling and data analysis, which are intertwined. Each week, you will do a series of increasingly complex computer exercises with online short exams each Thursday and Friday.
Required Knowledge	Demonstratable knowledge of Statistics up to regression and analysis of variance, as well as some experience in programming in languages such as R and Python are the pre-requisites.
References	• Introduction to Statistical Learning (James et al.) http://www-bcf.usc.edu/~gareth/ISL/ • R for Data Science (Grolund & Wickham) https://r4ds.had.co.nz/ • Data Science at the Command Line (Janssen) https://www.datascienceatthecommandline.com/ • Abraham Silberschatz, Henry F. Korth, S. Sudarshan "Database System Concepts" • Wes McKinney "Python for Data Analysis" • Raghu Ramakrishnan, Johannes Gehrke "Database Management Systems" • Bleifuß, Tobias, Sebastian Kruse, and Felix Naumann. Efficient Denial Constraint Discovery with Hydra. Proceedings of the VLDB Endowment (PVLDB). 11(3):311-323, 2017 • Loukides, M. "What is data science? The future belongs to the companies and people that turn data into products" • Jiawei Han, Micheline Kamber, Jian Pei "Data Mining: Concepts and Techniques" • Ian H. Witten, Eibe Frank "Data Mining: Practical Machine Learning Tools and Techniques" • Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze "An Introduction to information retrieval" https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf • …
Office Hours	--

Tentative Course Schedule

Wk.	Week ID	Topic	Date	Staff
36	W1	Orientation	31-08-2020 – 04-09-2020	MC & AS
37	W2_1	Course Introduction + Boolean Queries + Data collection and extraction (SQL Queries)	07-09-2020	HQ
	W2_2	Data collection and extraction (SQL Queries + Data extraction using R or Python)	08-09-2020	HQ
		Lab + Exams	Exam 1: 10-09-2020 Exams 2 & 3: 11-09-2020	
38	W3_1	Advanced SQL	14-09-2020	HQ
	W3_2	Data consistency (Integrity Constraints)	15-09-2020	HQ
		Lab + Exams	Exam 1: 17-09-2020 Exams 2 & 3: 18-09-2020	
39	W4_1	Heterogeneous Data Integration	21-09-2020	HQ
	W4_2	Entity Linkage	22-09-2020	HQ
		Lab + Exams	Exam 1: 24-09-2020 Exams 2 & 3: 25-09-2020	
40	W5_1	Data Visualization	28-09-2020	DO
	W5_2	Exploratory Data Analysis	29-09-2020	DO
		Lab + Exams	Exam 1: 01-10-2020 Exams 2 & 3: 02-10-2020	
41	W6_1	Data Preparation 1 (Cleaning + Transformation)	05-10-2020	HQ
	W6_2	Data Preparation 2 (Reduction + Normalization)	06-10-2020	HQ
		Lab + Exams	Exam 1: 08-10-2020 Exams 2 & 3: 09-10-2020	
42	W7_1	Missing Data and Imputation (1)	12-10-2020	DO
	W7_2	Missing Data and Imputation (2)	13-10-2020	DO
		Lab + Exams	Exam 1: 15-10-2020 Exams 2 & 3: 16-10-2020	
43	W8_1	Regression, Classification and Evaluation (1)	19-10-2020	HQ
	W8_2	Regression, Classification and Evaluation (2)	20-10-2020	HQ
		Lab + Exams	Exam 1: 22-10-2020 Exams 2 & 3: 23-10-2020	
44	W9_1	Clustering (1)	26-10-2020	DO
	W9_2	Clustering (2)	27-10-2020	DO
		Lab + Exams	Exam 1: 29-10-2020 Exams 2 & 3: 30-10-2020	
45	W10_1	Text Mining	02-11-2020	HQ
	W10_2	Dashboard Design	03-11-2020	HQ
		Lab + Exams	Exam 1: 05-11-2020 Exams 2 & 3: 06-11-2020	

* The reading material will be decided by the instructor and will be specified during the lecture.