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	--
Daniel Oberski	d.l.oberski	030 253 9039	Padualaan 14 Room C1.109	--
Erik-Jan van Kesteren	e.vankesteren1		Padualaan 14 Room C1.22	--
Ayoub Bagheri	a.bagheri		Padualaan 14 Room	--

Teaching Assistant(s)	
Name	Email (@uu.nl)
Ali	a.katsheh

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.

Utrecht University

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	RF1. Introduction to Statistical Learning (James et al.) http://www-bcf.usc.edu/~gareth/ISL/ RF2. R for Data Science (Grolund & Wickham) https://r4ds.had.co.nz/ RF3. Data Science at the Command Line (Janssen) https://www.datascienceatthecommandline.com/ RF4. Abraham Silberschatz, Henry F. Korth, S. Sudarshan "Database System Concepts" RF5. Wes McKinney "Python for Data Analysis" RF6. Raghu Ramakrishnan, Johannes Gehrke "Database Management Systems" RF7. Bleifuß, Tobias, Sebastian Kruse, and Felix Naumann. Efficient Denial Constraint Discovery with Hydra. Proceedings of the VLDB Endowment (PVLDB). 11(3):311-323, 2017 RF8. Loukides, M. "What is data science? The future belongs to the companies and people that turn data into products" RF9. Jiawei Han, Micheline Kamber, Jian Pei "Data Mining: Concepts and Techniques" RF10. Ian H. Witten, Eibe Frank "Data Mining: Practical Machine Learning Tools and Techniques" RF11. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze "An Introduction to information retrieval" https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf RF12. Jure Leskovec, Anand Rajaraman, Jeff Ullman, "Mining Massive Datasets" http://www.mmds.org RF13. Stef van Buuren, "Flexible Imputation of Missing Data" https://stefvanbuuren.name/fimd RF14. DL Oberski, "Mixture models: latent profile and latent class analysis" https://daob.nl/wp-content/uploads/2015/06/oberski-LCA.pdf RF15. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, "Introduction to information retrieval" https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf
Office Hours	--

Tentative Course Schedule				
Wk.	**Week ID**	**Topic**	**Reading**	**Staff**
36	W1	Orientation		MC & AS
37	W2_1	Course Introduction + Boolean Queries + Data collection and extraction (SQL Queries)	RF4 (CH 1 – 3.3)	HQ
	W2_2	Data collection and extraction (SQL Queries + Data extraction using R or Python)	RF4 (CH 3.4 – 4.1) RF5 (CH 5, 6, 8)	HQ
		Lab + Exams	Exam 1: 10-09-2020 Exams 2 & 3: 11-09-2020	Remindo
38	W3_1	Advanced SQL	RF6 (CH 5.1—5.7 and CH 8)	HQ
	W3_2	Data consistency (Integrity Constraints)	RF4 (CH 4.4 and 8.3) RF7	HQ
		Lab + Exams	Exam 1: 17-09-2020 Exams 2 & 3: 18-09-2020	Remindo
39	W4_1	Heterogeneous Data Integration	RF12 (CH 3)	HQ
	W4_2	Entity Linkage	RF12 (CH 3)	HQ
		Lab + Exams	Exam 1: 24-09-2020 Exams 2 & 3: 25-09-2020	Remindo
40	W5_1	Data Visualization	RF2 (Selected sections)	FSW
	W5_2	Exploratory Data Analysis	RF2 (Selected sections)	FSW
		Lab + Exams	Exam 1: 01-10-2020 Exams 2 & 3: 02-10-2020	Remindo
41	W6_1	Data Preparation 1 (Cleaning + Transformation)	RF9 (CH 3, 12)	HQ
	W6_2	Data Preparation 2 (Reduction + Normalization)	RF9 (CH 3)	HQ
		Lab + Exams	Exam 1: 08-10-2020 Exams 2 & 3: 09-10-2020	Remindo
42	W7_1	Missing Data and Imputation (1)	RF13 (Selected sections)	FSW
	W7_2	Missing Data and Imputation (2)	RF13 (Selected sections)	FSW
		Lab + Exams	Exam 1: 15-10-2020 Exams 2 & 3: 16-10-2020	Remindo
43	W8_1	Regression, Classification and Evaluation (1)	RF9 (CH 8, 9) RF10 (CH 5)	HQ
	W8_2	Regression, Classification and Evaluation (2)		HQ
		Lab + Exams	Exam 1: 22-10-2020 Exams 2 & 3: 23-10-2020	Remindo
44	W9_1	Clustering (1)	RF1 (CH 10.3) RF14 (1.1, 1.2)	FSW
	W9_2	Clustering (2)		FSW
		Lab + Exams	Exam 1: 29-10-2020 Exams 2 & 3: 30-10-2020	Remindo
45	W10_1	Text Mining	RF3 (CH 3, 5) RF15 (CH 2.2, 6.2, 6.3)	HQ
	W10_2	Dashboard Design	http://shiny.rstudio.com/tutorial http://shiny.rstudio.com/images/shiny-cheatsheet.pdf http://www.shinyapps.io http://www.showmeshiny.com/ http://deanattali.com/blog/building-shiny-apps-tutorial/	HQ
		Lab + Exams	Exam 1: 05-11-2020 Exams 2 & 3: 06-11-2020	Remindo

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