#### UNIVERSITY OF WISCONSIN-MILWAUKEE School of Information Studies

#### INFOST (691) – Data Analysis for Data Science Section 210, 211 - Online Spring 2019

# **SYLLABUS**

Instructor: Margaret Kipp E-mail: kipp@uwm.edu (best contact method) Fax: 414-229-6699 Office: NWQB 2574 Office Hours: TBA

# CATALOG DESCRIPTION

Introduces basic concepts, background, theoretical, practical and technological aspects of data analysis for data science. 3 credits

# **GENERAL DESCRIPTION**

This course provides an introduction to data analysis for data science. Data science has developed as a set of methods for analysing massive data sets to extract useful knowledge. A data scientist is a person who has the skills and knowledge to perform these analyses. This course will cover topics necessary to develop data science analysis skills such as statistics, probability, forecasting, prediction, optimization, clustering, and visualization.

# PREREQUISITES

- 582 Introduction to Data Science, or instructor consent. (Note: 799 Research Methods and 370 Data Analysis and Visualization for the Information Professional are also acceptable as prerequisites).
- Junior Standing. For 500 and 600 level courses it is recommended that an undergraduate student first consult with the appropriate instructor and/or advisor concerning the applicability of this specific course.
- Basic computer facility and technology literacy as listed in the SOIS policy are required, especially basic facility with Excel (or an Excel-like spreadsheet programme): <u>http://uwm.edu/informationstudies/academics/graduate/mlis/?</u> target=curriculum/#computer-literacy
- This course assumes basic statistical knowledge: measures of central tendency, basic probability, and some simple hypothesis testing such as the t-test, chi square test and correlation.

# **OBJECTIVES/OUTCOMES**

Upon completion of the course, students will be able to:

- 1. effectively develop researchable questions; (Paper and project)
- 2. identify and apply appropriate statistical methods for analysing data; (Basic Statistical Analysis, Choosing a Statistical Test, Advanced Statistical Analysis, Project)
- 3. identify and apply appropriate data science analysis techniques to analyse data; (Optimization, Forecasting and Prediction, Project)

- 4. critically evaluate tools for working with data; (Basic Statistical Analysis, Choosing a Statistical Test, Advanced Statistical Analysis, Optimization, Forecasting and Prediction, Project)
- 5. address multilingual and multicultural issues in data creation and analysis; (Short Paper, Project, Readings and Discussions)
- 6. identify emerging trends and stay current with issues in data science. (Readings and Discussions)

# ALA COMPETENCIES (for MLIS students)

- 1. Information, communication, assistive, and related technologies as they affect the resources, service delivery, and uses of libraries and other information agencies.
- 2. The application of information, communication, assistive, and related technology and tools consistent with professional ethics and prevailing service norms and applications.
- 3. The principles and techniques necessary to identify and analyse emerging technologies and innovations in order to recognize and implement relevant technological improvements.

# METHOD

Lecture/Discussion/Readings/Examples/Exercises – to achieve a satisfactory understanding of the course material and to fulfil requirements of the assignments, students are expected to attend the lectures, read and comment on the readings, participate in discussions and inclass exercises, and explore examples and tutorials.

#### TIME COMMITMENT

This course requires a weekly time commitment. General university guidelines indicate that a 3 credit course requires a minimum 144 hour time commitment over the course of a term. This time commitment represents a minimum of 9-10 hours of work per week per course. For an onsite class 3 of these hours represent onsite instruction in a classroom; in an online class this time would be spent on independent reading, discussions and in-class exercises.

Each week you may be required to read notes and readings from the reading list associated with that class, participate in discussions, write summaries of readings, complete in-class exercises, explore examples, or complete assignments and projects. It is your responsibility to plan your time in order to complete all activities based on the schedule outlined in this syllabus.

# ACCOMMODATIONS

If you need accommodations due to illness, disabilities, scheduling conflicts with religious observances, or other life events (e.g. military service) contact the instructor as soon as possible, preferably by the third week of class as per university policy. Official documentation may be required depending on the nature of the considerations requested per university policy (<u>http://www4.uwm.edu/secu/docs/faculty/1895R5\_Uniform\_abus\_Policy.pdf</u>).

# **TEXTBOOK AND READINGS**

Foreman, John W. 2015. Data Smart. Wiley. ISBN: 9781118661468 [Required]

Fischetti, Tony. 2015. Data Analysis with R. Packt. ISBN: 9781785288142 [Required] <u>https://www.packtpub.com/big-data-and-business-intelligence/data-analysis-r</u>

Toomey, Dan. 2014. R for Data Science. Packt. ISBN: 9781784390860 [Required] <u>https://www.packtpub.com/big-data-and-business-intelligence/r-data-science</u>

Knell, Robert. 2013. Introductory R: A Beginner's Guide to Data Visualisation, Statistical Analysis and Programming in R. Amazon Digital Services LLC. [Required] <u>https://www.amazon.com/dp/B00BU34QTM/</u> [Also available from iTunes, Google Books, etc.]

Lane. n.d. Online Statistics Education: An Interactive Multimedia Course of Study. <u>http://onlinestatbook.com/2/index.html</u> [Open Textbook]

Readings are listed in the course outline for each class. Readings should be completed before the class. Other course materials, including this syllabus, are available through D2L (<u>http://d2l.uwm.edu/</u>).

Changes may be made to the readings as the term progresses. Changes will be announced in D2L ahead of the classes for which changes will occur.

Class	Date	Topics	Readings (complete before class)	In-Class		
1	Jan 23	Introduction to Data Analysis for Data Science	<ul> <li>Loukides. 2010. What is Data Science? (12p) http://www.cloudera.com/content/da m/cloudera/Resources/PDF/What_is Data_Science_OReilly.pdf;</li> <li>Zhu &amp; Xiong. 2015. Defining Data Science. (8p) http://arxiv.org/abs/1501.0509 [cs.DB];</li> <li>Chandrasekaran. 8 July 2013. Becoming a Data Scientist. http://nirvacana.com/thoughts/becom ing-a-data-scientist/;</li> <li>Foreman. 2015. Data Smart. Wiley. Chapter 1;</li> <li>Fischetti. 2015. Data Analysis with R. Packt_Chapter 1;</li> </ul>	<ul> <li>Installing Software</li> <li>Using Excel/Calc, PSPP, or R to import data</li> </ul>		
2	Jan 30	Dealing with Messy Data; Ethics in Data	<ul> <li>Rahm. (n.d.) Data Cleaning: Problems and Current Approaches, University of Leipzig, Germany. (11p) <u>http://lips.informatik.uni-</u> <u>leipzig.de/files/2000-45.pdf;</u></li> <li>O'Leary. 2015. Big Data and Privacy: Emerging Issues. <i>Intelligent</i> <i>Systems, IEEE</i> 30(6): 92-96. (D2L);</li> <li>Kugler. 2016. What Happens When Big Data Blunders? <i>Communications</i></li> </ul>	<ul> <li>Ethics in Data Discussion</li> <li>Cleaning Data in a Spreadsheet</li> <li>Using OpenRefine</li> </ul>		

#### COURSE OUTLINE

			Of The ACM 59(6): 15-16. (D2L);		
			<ul> <li>Fischetti. 2015. Data Analysis with R. Packt. Chapter 11:</li> </ul>		
3	Feb 6	Review of Basic Statistical Analysis	<ul> <li>Online Statistics Education: An Interactive Multimedia Course of Study, Chapter 1-3, 5 <u>http://onlinestatbook.com/2/index.ht</u> <u>ml;</u></li> <li>Fischetti. 2015. Data Analysis with R. Packt. Chapter 2, 4;</li> </ul>	•	Descriptive Statistics Normal distributions
4	Feb 13	Review of Univariate and Bivariate Analysis; Comparison of Means; Power	<ul> <li>Online Statistics Education: An Interactive Multimedia Course of Study, Chapter 6-7, 11-12 <u>http://onlinestatbook.com/2/index.ht</u> <u>ml;</u></li> <li>Fischetti. 2015. Data Analysis with R. Packt. Chapter 5, 6;</li> </ul>	• • •	Comparison of means Z-test T-test Correlation
5	Feb 20	Choosing a Statistical Test; Non- Parametric Tests	<ul> <li>Online Statistics Education: An Interactive Multimedia Course of Study, Chapter 13, 17-18 <u>http://onlinestatbook.com/2/index.ht</u> <u>ml;</u></li> </ul>	• •	Choosing a test Chi-square Mann- Whitney/Wilc oxon Rank Sum Spearman's r
6	Feb 27	Advanced Bivariate Analysis	<ul> <li>Online Statistics Education: An Interactive Multimedia Course of Study, Chapter 14-15 <u>http://onlinestatbook.com/2/index.ht</u> <u>ml;</u></li> <li>Fischetti. 2015. Data Analysis with R. Packt. Chapter 8;</li> <li>Toomey. 2014. R for Data Science. Packt. Chapters 4 (not multivariate or robust), 5 (not polychoric, tetrachoric, heterogeneous or partial correlation);</li> </ul>	•	ANOVA Linear Regression Multiple Regression
7	Mar 6	K-means and Hierarchical clustering	<ul> <li>Foreman. 2015. Data Smart. Wiley. Chapter 2, 10 (matching sections);</li> <li>Toomey. 2014. R for Data Science. Packt. Chapters 1, 6;</li> </ul>	•	K-means clustering
8	Mar 13	Naïve-Bayes and Optimization	<ul> <li>Foreman. 2015. Data Smart. Wiley. Chapters 3,4;</li> <li>Zhang. 2016. Naïve Bayes classification in R. Annals of Transactional Medicine 4(12): 241. <u>https://www.ncbi.nlm.nih.gov/pmc/arti cles/PMC4930525/;</u></li> </ul>	•	Naïve-Bayes Optimization (Linear Programmin g)

9	Mar 20	Spring Break -	No Readings	•	
10	Mar 27	Social Network Analysis	<ul> <li>Foreman. 2015. Data Smart. Wiley. Chapter 5;</li> <li>Hanneman, Robert A. and Mark Riddle. 2005. Introduction to social network methods. Chapters 1, 3, 5 <u>http://faculty.ucr.edu/~hanneman/;</u></li> </ul>	•	SNA Gephi
11	Apr 3	Supervised Prediction and Ensemble Modelling	<ul> <li>Foreman. 2015. Data Smart. Wiley. Chapters 6, 7, 10 (matching sections);</li> <li>Toomey. 2014. R for Data Science. Packt. Chapter 11;</li> </ul>	•	Logistic regression Ensemble Modelling
12	Apr 10	Forecasting and Outlier Detection	<ul> <li>Foreman. 2015. Data Smart. Wiley. Chapters 8, 9, 10 (matching sections);</li> </ul>	•	Forecasting Outlier Detection
13	Apr 17	Visualizing Data in R	<ul> <li>Toomey. 2014. R for Data Science.</li> <li>Packt. Chapters 7-9 (skip latticist in chapter 7 and mapping);</li> </ul>	•	Visualization
14	Apr 24	General R Programming	<ul> <li>Fischetti. 2015. Data Analysis with R. Packt. Chapter 13;</li> <li>Ognyanova. 2016. Network Analysis and Visualization with R and igraph. Section 1 <u>http://kateto.net/networks- r-igraph;</u></li> <li>Martin. n.d. A beginner's introduction to the R programming language. Skim as required. <u>http://www.biostat.jhsph.edu/~ajaffe/ docs/undergradguidetoR.pdf;</u></li> </ul>	•	General R Programmin g
15	May 1	Special Topics (MDS and Mapping in R)	<ul> <li>Daróczi. 2016. Multidimensional Scaling with R. <u>https://www.r-statistics.com/2016/01/multidimensional-scaling-with-r-from-mastering-data-analysis-with-r/;</u></li> <li>Everitt and Hothorn. 2013. A Handbook of Statistical Analyses Using R. Chapter 14: MDS. <u>https://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_multidimensional_scalingpdf;</u></li> <li>Toomey. 2014. R for Data Science. Packt. Chapters 7 (mapping);</li> <li>Gilbert. 2012. Making Maps with R. <u>http://www.molecularecologist.com/2</u></li> </ul>	•	MDS Mapping

	1	1		1		
			012/09/making-maps-with-r/;			
			<ul> <li>Anderson. n.d. Making Maps with R.</li> </ul>			
			<u>http://eriqande.github.io/rep-res-</u>			
			web/lectures/making-maps-with-			
			<u>R.html;</u>			
16	May	Wrapup and	<ul> <li>Invasion of the data scientists: Hot</li> </ul>	•	Data	
	8	Data Science	job of 2016 expands beyond tech		Science Job	
		as a Career	http://www.denverpost.com/business/		Ads	
			ci 29451303/invasion-data-			
			scientists-hot-job-2016-expands-			
			beyond;			
			<ul> <li>Data Scientist: The Sexiest Job No</li> </ul>			
			One Has			
			http://www.informationweek.com/big-			
			data/big-data-analytics/data-			
			scientist-the-sexiest-iob-no-one-			
			has/d/d-id/1112832:			
			<ul> <li>Interviews with Data Scientists</li> </ul>			
			http://www.datascienceweekly.org/pd			
			f/DataScienceWeekly-			
			DataScientistInterviews-Vol1-			
			April2014.pdf (one interview):			
	1	1		1		

# ASSIGNMENTS

Assignment	Graduate	Undergrad	Associated	Due Date
You may use whichever statistical			Classes	
package you choose for the				
assignments, but you must use one				
from each class of software				
(Spreadsheet, Stats, R) over the				
course of the term.				
<b>Basic Statistical Analysis</b>	5	10	1-3	Class 4
Use one of the suggested analysis				
tools from class 1 to perform a basic				
statistical analysis (descriptive				
statistics, basic graph) on a given				
data set. Analyse your findings and				
write a short report about the data				
(600 words).				
Choosing a Statistical Test	10	10	4-5	Class 6
Choose a relevant statistical test to				
analyse the given data set and apply				
it. Analyse your findings and write a				
short report. Include a justification of				
your choice of statistical test (600				
words).				
Advanced Statistical Analysis	10	10	6	Class 7

Analyse the given data using ANOVA or multiple regression. Analyse your findings and write a short report (800 words).				
<b>Optimization</b> Use K-means clustering or another method to try to optimize the given data set. Write a report on your findings (800 words).	10	10	7-8	Class 10
<b>Forecasting and Prediction</b> Use forecasting and prediction techniques to extrapolate tendencies in the given data set. Write a report on your findings (800 words).	10	10	10-12	Class 13
<b>Project</b> Analyse a given data set using techniques covered in this course and produce a short report, including graphs. Justify your chosen technique and interpret the results. (800U/1200G words or equivalent, charts, tables, etc. do not count towards the word limit) Note: Graduate students are expected to provide additional critical analysis and reflection of the data including potentially locating and citing appropriate supporting materials from published sources.	20	35	13-14	Proposal: Class Class 8 Project: Last class
<b>Short Paper</b> Write a short paper on a data science related topic (1000 words).	20	n/a	All	Proposal: Class 3 Paper: Class 11
Participation (see below)	15	15	All	Last class

# Working with Classmates

All assignments except the short paper and participation may be completed in pairs or trios.

Assignments completed in pairs/trios must identify all work partners by full name at the top of the assignment. You must each submit the same assignment to the dropbox. If you simply assisted each other but did not do the whole assignment together, you must also note this at the top of the assignment. Unacknowledged borrowing is seen as plagiarism, so be sure to document your teamwork to avoid this.

# Formatting Guidelines for Assignments

Assignments should be written using Arial or another Sans-Serif style font. Do not use red for emphasis or to highlight your answers to questions. Remove all extraneous information before submission (e.g. assignment instructions or tips).

Use whatever citation format you prefer, but do not use footnotes. If you are not using a common format such as MLA or APA you should include information about which style guide you are using in the assignment.

Paper submissions will not be accepted. All assignments must be typed on a computer and submitted electronically. Handwritten submissions will not be accepted, even if scanned and submitted electronically.

Assignments may not be submitted in Pages, Microsoft Works, or Microsoft Project as I cannot open these formats. You should save these as a PDF instead. Other common file formats should be acceptable including Open Office formats. If you are using an unusual format you can always check with me first before submission to ensure I can open it.

#### Due Dates and Assignment Submission

All assignments and projects should be submitted through D2L to the appropriate dropbox before midnight (Central Time) on the due date. Points for late assignments will be reduced 10% per day late after the due date. The dropbox will remain open for the submission of late assignments until the late penalty reaches 100%.

Participation items should be submitted to the appropriate discussion group (see the participation section below) before the discussion group closes. Discussion groups will be open for 1 week before and 1 week after the date of the associated class.

Emailed submissions will only be accepted as a backup to a D2L submission (or in case of D2L errors).

Everything must be submitted by the Last Class (this includes all assignments, papers, projects, and participation). All project and assignment deadlines are in the syllabus. For discussion deadlines check the discussion groups or the D2L calendar. The D2L calendar also contains all project and assignments deadlines. It is your responsibility to keep track of deadlines using the tools provided or by creating your own calendar of deadlines.

Items submitted early will not be evaluated until their due date. Students are encouraged to complete all Associated Classes listed under Assignments before submitting the assignments since the material in these classes constitutes preparation for the assignments. Submission well before the due date is not encouraged.

#### Extensions

Students must contact the instructor before each due date listed under Assignments for any extensions. Extension requests made prior to the due date do not require any documentation as long as they are not longer than a week. Simply provide a date/time by which you will submit the assignment. After the deadline the penalties listed under Due Dates and Assignment Submission will be enforced. Material submitted late after an extension will also be subject to these penalties. Plan your time accordingly.

# **Extra Credit or Other Special Considerations**

Per university policies (see <u>http://www4.uwm.edu/secu/docs/other/S29%2Ehtm</u>) extra credit assignments and other special consideration are not possible. Students should make use of the extensions policy outlined above or provide appropriate documentation of special circumstances as outlined elsewhere in the syllabus.

#### Participation

Students are expected to participate in discussion and in-class exercises as a demonstration of their ability to articulate key concepts. Discussion will include individual and group components. Participation is mandatory and constitutes one quarter of the points available for this class. Participation will consist of all of the following: individual summaries of readings, participation in group discussions, contributed articles, and responses to others.

Participation will consist of all of the following:

- Completion of the Syllabus Quiz
  - The syllabus quiz must be completed in the first 2 weeks of class. Bonus points will automatically be entered in D2L.
- Individual Summaries of Readings
  - Post 3 summaries of the weekly readings to the appropriate weekly discussion group based on the class associated with each reading.
  - You must post 3 summaries in total, but you may choose the classes for which you wish to contribute the summaries.
  - Sign up for 3 sets of readings on the signup sheet posted in the news section of D2L.
  - Responses need not exceed 300 words.
  - Summaries posted before the date of the class earn a half bonus point each. Be sure to mark this on your course completion checklist to ensure you receive the bonus.

# Contributed Article

- Contribution of a new article, video, cartoon, etc. relevant to the class and a short summary (approximately 100 words) explaining its relevance to class. This should be posted to the appropriate weekly discussion group based on the topic. You may choose which week you wish to contribute this item.
- A signup sheet will be posted in the news section of D2L.

# • Individual Summaries of In-Class Exercises

- Participation in the in-class exercises included in most weeks. Post individual summaries to the appropriate weekly discussion group.
- You must post 6 summaries of in-class exercises in total, but you may choose the classes for which you wish to contribute the summaries.
- Responses need not exceed 300 words.

# • Participation in Weekly Discussions

- Participation in weekly discussions including reading and/or responding to weekly reading summaries and other information posted to the weekly discussion groups by classmates. Points will be allocated based on your reading level (i.e. many, few, nothing read) and/or your responses to others (i.e. many, few, no responses).
- Generally frequent participation requires that you participate at least once a week in most weeks.
- Submission of the Course Checklist to the participation dropbox

 The completed checklist with all required course elements listed submitted to the dropbox before the last class. You should complete as much as possible of the checklist. Use the checklist throughout the term to ensure you are on track to complete all course requirements.

#### Code of Conduct/Expectations for this Class

This is a professional programme and professional, courteous behaviour is expected of all participants. It is expected that class members will show consideration for all other members of the class and contribute in a constructive manner which is conducive to a good learning environment. Class members should consider the relevance and appropriateness of their contributions to the class before contributing to the class. Violations of these expectations will result in reduced participation points or other sanctions depending on severity.

#### Plagiarism and Referencing

It is expected that you will consult and cite the research and professional literature where merited and not rely solely on encyclopaedias, newspapers or unpublished, online sources. Papers where the majority of sources are blogs and Wikipedia (or similar sites) will not be accepted.

Use a common style manual for citations (e.g. APA, MLA, Chicago, etc.). Ideally you would choose a citation style guide you have used before, or one you are using in another class.

Plagiarism is the unacknowledged borrowing of ideas or material from someone else's work. It is considered an academic offence and can be considered grounds for failure in a course or expulsion from the programme. Cite all references and provide credit for all other materials. This applies to all material including images, sounds or videos. A citation (in the format of your choice) with a functioning URL (if relevant) is the minimum required for a reference. (http://guides.library.uwm.edu/content.php?pid=235714&sid=1949820#6509804)

You may not resubmit assignments already submitted in other courses or in a previous instance of this course, nor may you submit other people's work as your own. Plagiarism will be dealt with on a case by case basis but will result in a lowered mark on the assignment, failure on the assignment or failure in the course depending on severity and the number of plagiarized items submitted. Points lost through plagiarism may not be replaced by bonus points on other assignments.

96-100	Α	Superior work	74-76	С	Work is below standard
91-95	A-		70-73	C-	
87-90	B+		67-69	D+	
84-86	В	Satisfactory, but undistinguished work	64-66	D	Unsatisfactory work
80-83	B-		60-63	D-	
77-79	C+		Below 60	F	

#### GRADING SCALE

# UWM AND SOIS ACADEMIC POLICIES

The following link will take you to UWM pages/links which contain university policies affecting all UWM students. <u>http://uwm.edu/secu/wp-content/uploads/sites/122/2016/12/Syllabus-Links.pdf</u>

The following link will take you to pages/links which contain SOIS policies affecting all SOIS students. <u>http://www4.uwm.edu/sois/resources/formpol/policies.cfm</u>

Undergraduates may also find the *Panther Planner and Undergraduate Student Handbook* useful. <u>http://uwm.edu/studenthandbook/</u>

For graduate students, there are additional guidelines from the Graduate School. <u>http://uwm.edu/graduateschool/</u>

COSO This document is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States Licence except where other rights exist.