GLIS 692 - *Winter 2021*

Data Science for Information Professionals

*McGill School of Information Studies*

*Instructor: Eleonore Fournier-Tombs*

*TA: Cyntia Tarlao*

# Course Description

This course introduces the field of data science with a focus on the application of its various tools and methodologies in a professional context. It will cover the theoretical background and context of data science as a new field and allow students to become familiar with the basics of the end-to-end data science workflow, including data modelling, descriptive and predictive analytics, technical implementation and results reporting.

The course is acceptable for technical and non-technical audiences, but for non-technical students will require the willingness to learn some basics of programming in R.

# Learning Outcomes

At the end of this course the student should be able to

1. Identify and understand the main elements of a data scientists' toolkit
2. Construct analytical questions from abstract requirements
3. Select appropriate data sources and methodologies to answer analytical questions
4. Discuss the basic mathematical constructs of data science
5. Understand the concepts of descriptive and predictive analytics and their application
6. Demonstrate comfort with the various analytical tools of the R environment
7. Develop simple visualizations to communicate analytical outcomes
8. Discuss and critique the societal impacts and uses of data science.

# Course content

* Basic tool kit: Introduction to data science and R
* Basic tool kit: R for data science
* Basic tool kit: Visualizations and plotting
* Basic tool kit: Data analytics and model building
* Computational analysis: Regression
* Computational analysis: Classification
* Computational analysis: Clustering
* Computational analysis: Text analysis – Part 1
* Computational analysis: Network analysis
* Computational analysis: Methodology review
* Working in data science: Algorithmic bias
* Working in data science: The future of data science

# Guest lectures:

* Laurence Mathieu-Léger, Senior Producer and Video Journalist, Quebecor – Data, Inclusion and US Politics
* Manu Singh – Data Science and Political Science, Columbia University
* Roberto Rocca – Data Journalism

# Readings

1. Dietrich, D., Heller, B., & Yang, B. (2015). Data Science and Big Data Analytics: Discovering. Analyzing, Visualizing and Presenting Data. Indianapolis, IN: Wiley.

eBook available through McGill Libraries: <http://mcgill.worldcat.org/oclc/898154112>

1. Grolemund, Garrett & Wickham, Hadley. (2017). R for Data Science: <https://r4ds.had.co.nz/>
2. Siegel, Eric (2016). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie or Die. Hoboken, NJ: Wiley.

eBook available through McGill libraries: <http://mcgill.worldcat.org/oclc/924684043>

# Course outline

|  |  |  |  |
| --- | --- | --- | --- |
| Week | Class content | Readings and Videos (to complete before class) | Workshop |
| 1. Jan 11-17 2021 | Basic tool kit: Introduction to data science and R | * Week 1 videos * Week 1 quiz * Instructions to install R in Grolemund et al (2017) Introduction | Installation of R-studio and basic R exercises |
| 2. Jan 18-24 2021 | Basic tool kit: R for data science | * Week 2 videos * Week 2 quiz * Grolemund et al (2017). Ch. 4 * Dietrich et al (2015). Ch. 1 | Discussion workshop on data science question framing  R and R studio exercises |
| 3. Jan 25-31 2021 | Basic tool kit: Visualizations and plotting | * Week 3 videos * Week 3 quiz * Grolemund et al (2017). Ch. 3 & Ch. 27 * Dietrich et al (2015). Ch. 3 | R and R studio exercises - analysis and plotting using melanoma dataset  Assignment 1 work |
| 4. Feb 1-7 2021 | Basic tool kit: Data analytics and model building | * Week 4 videos * Week 4 quiz * Grolemund et al Ch. 7 & Ch. 23 * Dietrich et al (2015). Ch. 2 * Siegel (2016). Introduction | R and R studio exercises – Human Freedom Index |
| 5. Feb 8-14 | Computational analysis: Regression | * Week 5 videos * Week 5 quiz * Dietrich et al (2015). Ch. 6 * Siegel (2016). Ch. 1 | R and R studio exercises – cost of US Airlines |
| 6. Feb 15-21 | Computational analysis: Classification | * Week 6 videos * Week 6 quiz * Dietrich et al (2015). Ch. 7 * Siegel (2016). Ch. 2 | Discussion workshop on modeling in data science  R and R studio exercises – predicting presence of frogs in micro-eco systems |
| 7. Feb 22-28 2021 | Computational analysis: Clustering | * Week 7 videos * Week 7 quiz * Dietrich et al (2015). Ch. 4 * Siegel (2016). Ch. 3 | Discussion workshop on Canadian customer segments  R and R studio exercises – United Nations voter blocks |
| MARCH BREAK (March 1-7 2021) | | | |
| 8. Mar 8-14 2021 | Computational analysis: Text analysis – Part 1 | * Week 8 videos * Week 8 quiz * Dietrich et al (2015). Ch. 9 * Siegel (2016). Ch 4 | R and R studio exercises – LDA analysis of research papers  Create a text message spam classifier |
| 9. Mar 15-21 2021 | Computational analysis: Network Analysis | * Week 9 videos * Week 9 quiz * <https://www.jessesadler.com/post/network-analysis-with-r/> * Siegel (2016). Ch. 5 | R and R studio exercises – modeling the Marvel Universe network |
| 10. Mar 22-29 | Computational analysis: Methodology review | * Week 10 videos * Week 10 quiz * Siegel (2016). Ch. 6 | Review exercises and class discussion |
| 11 Mar 30-Apr 4 | Working in data science: Algorithmic bias | * Week 11 videos * Week 11 quiz * Siegel (2016). Ch. 7 | Work on assignment 3 |
| EASTER BREAK (April 5-April 11, 2021) | | | |
| 12 Apr 12-Apr 18 | Working in data science: The future of data science | * Week 12 videos * Week 12 quiz * Reading TBD | Discussion workshop on jobs in data science |

# Assignments and learning outcome mapping

The main assignment for this course will be the development of a data science project, which will be separated into three assignments, building upon each other. In the first assignment, the students will conduct a peer review of a data science paper, in order to understand typical components of research papers and practice critical analysis are. In the second assignment, they will describe their chosen problem statement, proposed methodology and data sources for their project, as well as exploratory analysis. In the third assignment, they will bring all elements together to submit a complete data science report, including results and recommendations.

The quizzes are meant to improve retention and reflection after having watched each videos. They will not be graded, however students are highly encouraged to complete them on a weekly basis in order to be able to interact critically with the material that is provided.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Description | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Due | % |
| A0 | Retention and critical thinking quizzes | X | X | X | X | X | X | X | X | Week 12 | 0 |
| A1 | 3-page article peer review | X | X | X |  | X |  |  |  | February 5, 2021 | 25 |
| A2 | 5-page description of problem statement, proposed methodology and data sources, descriptive analytics and visualizations. |  |  |  |  | X | X | X |  | March 19, 2021 | 40 |
| A3 | Final paper including revised A2, analytical results and recommendations |  | X | X |  | X | X | X |  | April 18, 2021 | 45 |

# Distance learning - lecture and video format

Each week, there will be a 15-20 minutes video to watch, which will introduce the key learning elements for the week. The live lecture time will last between 60 and 90 minutes each week, and will focus on questions, live coding exercises, groups discussion and review of assignments and deliverables. The live lecture will be recorded, and attendance will not be mandatory. Key elements discussed, such as assignments and deadlines, will be also communicated with the students separately.