UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

COURSE DESCRIPTION– Spring 2014

COURSE NAME	**Data Analytics: Introduction, Methods and Practical Approaches**
COURSE CODE	INF2190H
DESCRIPTION	The influx of data that is created, gathered, stored and accessed has given birth to some new areas of data analysis. The terms "predictive analytics", "big data" and "data science" are prevalent in scientific as well as broad audience publications and often make part of new business opportunities. Understanding the significance of techniques that perform analytics and knowing how to interpret their results offers a unique advantage in the performance of information professionals within an organization. This course provides an introduction to the field of analytics, and therefore the extensive use of data, statistical and quantitative analysis, exploratory and predictive models to mine and discover unexpected but useful glimpses of previously unknown information. We discuss standard data mining algorithms that can be applied on both structured and unstructured data and experience their impact on decision making situations. The students will actively participate in the delivery of this course through case and project presentations.
INSTRUCTOR	Periklis Andritsos
TIME SPAN	January to April 2014
LECTURE TIME	9am – 12pm on Tuesdays
OFFICE HOURS	12pm – 1pm on Tuesdays or by appointment
PREREQUISITES	- Recommended INF1343. - Recommended that students have some basic statistics background.
COURSE OBJECTIVES	The objective of the course is to introduce the notions of "Data Analytics" and provide an overview and hands-on experience of tools that perform analytical tasks. Specifically, we will focus on data mining techniques such as clustering, classification, and association rules. The course is designed so that the students acquire the theoretical foundations from statistics that will help them identify and solve any problems that arise when mining large repositories of data.
OVERVIEW	• Introduction to Data Mining • Data types • Data preparation • Basic and advanced probability and Statistics • Introduction to Data Warehouses • Association Rules • Cluster Analysis • Classification

- Text Processing and Analysis
- Overview of Big Data Technologies
- Class Presentations

Student Learning Objectives

By the end of the this course the students should be able to:

- Understand the theory behind three main techniques in Data Analytics.
- Apply data analytic techniques across multiple large data sets.
- Apply Methods for pre-processing data for large Data Analytics (e.g. discretizing numerical values or replacing empty values).
- Use open source tools, e.g. WEKA, for performing data analysis.
- Explain results using appropriate graphs and visualizations.

Relationship to Masters of Information (MI) Program-Level Student Learning Outcomes

Data Science has become a significant practice in several disciplines that deal with large amounts of data and information extracted. This course will help students understand and adapt in the practice of Data Analyst and Scientist (Outcome 1). With the knowledge acquired in INF2190, the students will be able to lead in taking decisions and provision of information services for many disciplines (Outcome 2). The course will allow students to develop their own goals and continue in life-long intellectual growth beyond graduation (Outcome 6).

READINGS

Every week, readings will be posted on the blackboard. We will also and primarily look at an Open Source Data Mining Tool, namely WEKA[1]. The readings will mostly be some research papers as well chapters from the following books:

- "Data Mining, Concepts and Techniques" by Jiawei Han, Micheline Kamber and Jian Pei. 2nd Edition. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2005. 800 pages. **Referred as [HK]**
- "Data Mining, Practical Learning Tools and Techniques with Java Implementations" by Ian H. Witten and Eibe Frank, Morgan Kaufmann. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 1999. 371 pages. **Referred as [WF].**
- "Bad data Handbook: Mapping the World of Data Problems" by Q.Ethan McCallum. O'Reilly Media, 2012. 264 pages. **Referred as [EC].**
- "Statistics Hacks: Tips & Tools for Measuring the World and Beating the Odds" by Bruce Frey. O'Reilly Media, 2006. 358 pages. **Referred as [BF].**
- "Hadoop: The Definitive Guide" by Tom White. O'Reilly Media, 2012.

[1] http://www.cs.waikato.ac.nz/ml/weka/

688 pages. **Referred as [TW].**

- "Mining of Massive Datasets" by Anand Rajaraman, Jure Leskovec and Jeffrey Ullman. Cambridge University Press, 2011. 326 pages. **Referred as [RLU].**

ASSESSMENT

Students will form groups of up to three members. The coursework will include three parts; the first will be submitted in the first 4 weeks and it will be a proposal on the group's project. On week 6, there will be a midterm test, in which every student will be assessed individually. The test will be with close books and notes. A group presentation will be held on Week 12. Students will be asked to find a problem in their path/concentration or specialty (e.g. Information Systems and Design or Library Information Systems) that requires further analysis. Details of each piece of coursework and distribution of marks will be done as follows:

Assignment (week 5) Feb 10th, 2014 : After students have been split into groups they will need to choose a data analytic project to work on. Each group will submit a unique proposal that will:

1. state the problem clearly and the motivation;
2. choose at least two data sets that will contribute to a solution (e.g. from the Open Data Ontario initiative);
3. state the possible techniques that can be used towards a solution;

Worth: 20%

In-class midterm (week 6) Feb 24th, 2014: The midterm will last for 1 hour and will concern main concepts that have been presented during the class. It will include multiple choice questions. The exam will be a "closed-books" and "close-notes" exam.
Worth: 40%

In-class presentation (week 12) April 8th, 2014 : The groups will submit a paper describing their work. The paper will be up to 6 pages long and further instructions for its contents will be given during the semester. During lecture time, a final in-class presentation will be give by each group, during which they will have to:

1. present the problem;
2. showcase example scenarios where the problem can have an impact;
3. present the data sets and techniques in detail
4. showecase the solutions found

Emphasis will be give in demonstrating that the results are interesting and can be used to change how an existing process or a new process can be implemented as efficiently and effectively as possible.
Worth: 40% (20% for the paper + 20% for the presentation)

Self and peer assessment: Assignment submission will include self and peer assessment forms that must be completed by each group member separately. When students upload their assignments (on Blackboard) they will be asked to upload a special form discussing the teamwork. These forms are strictly

confidential and will be provided during the course (on Blackboard).

WEEKLY SCHEDULE

Week 1 (7/1)
::::: Introduction to Big Data and Data Mining

Readings
- Big Data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May 2011.
- Challenges and Opportunities with Big Data: A community white paper developed by leading researchers across the United States. February 2012.
- [HK], Chapter 1.
- [WF], Chapter 1.

Week 2 (14/1)
::::: Probability and Statistics
- [BF], Chapter 1.

Week 3 (21/1)
::::: Data pre-processing
- [HK], Chapter 2.
- [EC], Chapter 1.

Week 4 (28/1)
::::: Introduction to Data Warehouses
- [HK], Chapter 3.
- An Overview of Data Warehousing and OLAP technology. Surajit Chaudhury and Umeshwar Dayal. ACM Sigmod Record Volume 26, Issue 1, March 1997.

Week 5 (11/2)
::::: Association Rules
- [HK], Sections 5.1, 5.2, 5.3: Mining Frequent Patters and Associations.
- [WF], Chapter 4.5: Mining Association Rules.
- Fast Algorithms for Mining Association Rules. Rakesh Agrawal and Ramakrishnan Srikant. VLDB Conference, 1994.
- Hands-on example.

Week 6 (25/2)
::::: Cluster Analysis
- Data Clustering Techniques. Tech. Report CSRG-443, U. of Toronto, Dep. Of Computer Science, March 2002.

- [HK], Chapter 7.
- [WF], Chapter 6.
- Hands-on example.

Week 7 (4/3)
::::: Classification
- [HK], Chapter 6.
- [WF], Sections 4.3 & 4.6.
- Hands-on example.

Week 8 (11/3)
::::: Text Processing and Analysis
- Overview and Semantic Issues of Text Mining. Anna Stavrianou, Periklis Andritsos and Nicolas Nocoloyannis. In SIGMOD Record, 36(3), pp. 23-34, September 2007.
- Text-Mining Tutorial. Marko Grobelnik, Dunja Mladenic. Learning Methods for Text Understanding and Mining, Grenoble, France, 2004.
- [HK] Section 10.4.
- National Centre for Text Mining: Source of several articles and tools.
- Hands-on example.

Week 9 (18/3)
::::: Big Data Technologies
- Map/Reduce Technologies, [TW] Chapter 2
- Hadoop, [TW] Chapter 3

Week 10 (25/3)
::::: Big Data Visualization
- This will be a showcase of tools that can be used to visualize big data. We will look at: Tableau Software, D3 Visualization, dbTouch and other technologies that have appeared in this arena.

Week 11 (1/4)
::::: What is Predictive Analytics
- Wojciech Gryc has from Canopy Labs, a Toronto-based Big Data company involved in predictive analytics has tentatively agreed to give a invited talk to students of the class.
- I also want to give the student the opportunity for some time asking me questions regarding their projects.

Week 12 (8/4)
::::: In-class presentation

General Expectation

1. Communication Policy: Please do not email questions to the instructors or TAs. If you have a question, there is a pretty good chance that other people in the course have the same question or, at least, will benefitt from the answer. Please post all the questions to Blackboard (forum threads to be announced) so everyone in the course can benefitt from your questions and our answers. Questions posted to Blackboard will be answered within two (2) business days. Students are encouraged to post answers to the questions of other students where appropriate.

IMPORTANT: Please prefix the subject of your emails to the instructor and TA with "INF2190H" and include some more details, e.g., "INF2190: book appointment October 1st".

2. Readings: It is important to complete the required readings before the lecture in order to fully benefit from the class activities.

3. Late policy: Late submission of an assignment carries a penalty of one grade (e.g. from B+ to B) for each week, to a maximum of two weeks; submissions will not be accepted after two weeks. Exceptions will be made only when supported by appropriate documentation.

4. Academic Integrity: The essence of academic life revolves around respect not only for the ideas of others, but also their rights to those ideas and their promulgation. It is therefore essential that all of us engaged in the life of the mind take the utmost care that the ideas and expressions of ideas of other people always be appropriately handled, and, where necessary, cited. For writing assignments, when ideas or materials of others are used, they must be cited. You may use any formal citation format, as long as it is used consistently in your paper, the source material can be located and the citation verified. What is most important is that the material be cited. In any situation, if you have a question, please post it to Blackboard. Such attention to ideas and acknowledgment of their sources is central not only to academic life, but life in general. Please acquaint yourself with UofT's Code of Behaviour on Academic Matters.

5. Participation and Attendance: Discussion and interaction in

the classes are important ways to learn. Sharing your experiences and ideas with your classmates is central to your learning experience in this course. As such, you should attend and participate in every class. There will also be exercises and discussions that you will participate in within your groups in your class. Some of the activities will be very helpful in completing your assignments.

6. Students with Special Needs or Health Considerations: All students are welcome in this course and we will make every e ort to ensure a meaningful, respectful, and positive learning experience for everyone. If there are special considerations that you require to help you successfully fulfil the requirements of the course, please feel free to see one of the instructors, the Faculty of Information Student Services, and/or contact the Accessibility Student Office as soon as possible so we can ensure you are able to successfully meet the learning objectives for this course.

7. Writing Resources: Please review the material you covered in Cite it Right, familiarize yourself with the How Not to Plagiarize site and UofT's policy, and consult the Office of English Language and Writing Support as necessary.