University of Ottawa

School of Electrical Engineering and Computer Science

CSI4142 Introduction to Data Science

Winter 2019

This course provides an introduction to data science, following a data driven discovery perspective. We will focus on how to create a repository for analytics and mining (a so-called data mart), and we will also cover a number of techniques and algorithms that were developed to explore large-scale data.

Formal Calendar description

Data preparation: organization, basic statistics, cleaning, and integration; Data warehousing and multi-dimensional analysis; Data mining techniques: pattern mining, classification, clustering, outlier and anomaly detection; model evaluation; Big data, analytics, and cloud computing; Data visualization and visual data analytics.
Prerequisites: CSI2132, (CSI3120 or SEG2106), MAT2377 or (MAT2371 and MAT2375).

Professor's details

Herna L Viktor, PhD
Email: hviktor@uottawa.ca
Office: SITE Building Room 5-100
Office Hours: Friday 11h00-12h00 (or by email appointment)

Recommended Texts

The notes are based on parts of the following books:

1. **Data Mining, Concepts and Techniques, 3rd Edition, Jiawei Han, Micheline Kamber and Jian Pei, Morgan Kauffman Publishers, 2012, ISBN 978-0-12-381479-1.**
2. **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, (Selected Chapters), 3rd Edition, Ralph Kimball and Margy Ross, Wiley, 2013, ISBN 978-1-11-853080-1.**
   This book contains a number of useful case studies that illustrates the fundamental concepts.

Final grade

Your final grade will be calculated as follows.

Team project (3 students)	**40**
Midterm	**25**
Final Exam	**35**

Some important information is listed below.

1. **The team project will involve the design and implementation of a data mart, as well as the exploration of this data mart using online analytic processing (OLAP) and data mining techniques. Complete this project in a team of 3 students. The project will be done in three phases, completed during the term:**
    a. **Conceptual design: Due on 5 February 2019.**
    b. **Physical design, data staging and OLAP queries: Due on 12 March 2019.**
    c. **BI dashboard, data mining and information visualization: Due on 2 April 2019.**

2. **The completed final team project is due on 2 April 2019. Teams are required to demonstrate their projects in a 15-20 minute timeslot. Note that all team members are required to attend the project demonstration.**

3. **You are allowed to use any full-fledged DBMS of your choice, such as PostgreSQL (with Jason), or MySQL. You are also welcome to use Hadoop or Spark.**
    a. **You are encouraged to use Scikit-Learn or R for the data mining portion of this course. Both are widely used in the data science community and will strengthen your CV.**
    b. **Other options are the WEKA data mining tool, Matlab and Mathematica.**

Overview of Lectures

The following topics will be covered. Please refer to the slides and the recommended texts.

Week of	Topic	Reference
07/01/2019	Introduction and course outline	Notes
14/01/2019	Store: Conceptual Modeling	Kimball 1,2, 17, 18 + CS*; Han 4, 5
21/01/2019	Store: Physical Design and Aggregation	Kimball 1, 2, 17, 18 + CS*; Han 4, 5
28/01/2019	Store: Data staging (ETL)	Kimball 19,20 + CS*
04/02/2019	Explore: Analytics via OLAP queries	Notes; Han 4, 5
11/02/2019	Explore: Data mining fundamentals	Han 1
18/02/2019	*Reading week*	
25/02/2019	Midterm on Friday 01/03/2019	All up to now
04/03/2019	Explore: Getting to know your data	Han 2, 3
11/03/2019	Explore: Finding frequent patterns	Han 6
18/03/2019	Explore: Finding groupings	Han 8
25/03/2019	Explore: Classification and prediction	Han 10
01/04/2019	Explore: Finding anomalies and outliers	Han 12

CS* refers to the Case Studies that are discussed in Chapters 3 to 16 of the textbook by Kimball and Ross.