# University of Ottawa
# School of Electrical Engineering and Computer Science
# CSI4142 Introduction to Data Science

## Winter 2018

## 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.
Prerequisite: CSI2132, (CSI3120 or SEG2106), MAT2377 or (MAT2371 and MAT2375).

## Texts used for reference

The notes will be based on information as contained in a number of texts. The following two texts are the most relevant:

1. Data Mining, Concepts and Techniques, 3rd Edition, Jiawei Han, Micheline Kamber and Jian Pei, Morgan Kauffman Publishers, ISBN-10: 9380931913, 2011.
2. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Edition, Ralph Kimball and Margy Ross, Wiley, ISBN-10: 1118530802, 2013. (Note that chapters 3 to 16 of this book discuss multiple case studies: a great practical approach to grasping the concepts.)

## Professor's details

Herna L Viktor, PhD
Professor of Computer Science, School of EECS, uOttawa

| | |
|---|---|
| Office: | SITE Building, Room 5-100 |
| Office Hours: | Tuesdays 11h00-12h00 or by email appointment |
| Email: | hviktor@uottawa.ca |
| Phone: | 613 562 5800 2341 |


## Your final grade

| | |
|---|---|
| Two Individual Assignments | 10 |
| Team project (3 students) | 30 |
| Midterm on Tuesday 27th February, during lecture time | 25 |
| Final Exam | 35 |
| **TOTAL** | **100** |

## About the Team Project and Assignments

Submit all your work through the Virtual Campus. Some important information items are listed below.

1. The two assignments are individual.
    a. The first assignment will be posted on January 24th and it is due two weeks later.
    b. The second assignment will be posted on March 7th and it is due two weeks later.
2. 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.
3. All teams will use the same data, as agreed in class, by majority voting.
4. Students are encouraged to suggest a project no later than January 19th. (Note that this is optional and that no-one will be penalized if not participating.)
5. Some potential sources to explore include:
    a. Canada's Federal and Provincial Open Data Repositories, including:
        i. http://open.canada.ca/en/open-data
        ii. https://www.ontario.ca/search/data-catalogue
    b. World Bank Data: https://data.worldbank.org/
    c. "Awesome Public Datasets": https://github.com/caesar0301/awesome-public-datasets
    d. The Paradise Papers: https://offshoreleaks.icij.org/pages/database
6. The team project is due on April 5th. Teams are required to demonstrate their projects in a 15-20 minute timeslot. Note that all team members are required to attend the project demonstrations.
7. You are allowed to use any full-fledged DBMS of your choice, such as PostgreSQL. You are also welcome to use Hadoop or Spark.
8. You are encouraged to use the R programming language for the data mining portions of the team project. (R is widely used in the data science community and will strengthen your CV. Some machine learners are also now switching to Python.) Other options are the WEKA data mining tool, Python, Matlab and Mathematica. You may also use any statistical package to initially explore the data.

## Course Outline 2018

| Week of | Topic | References |
|---|---|---|
| 08-Jan | - Introduction<br>- Getting to know your data | <br>Han 1-2 |
| 15-Jan | - Conceptual data mart design | Kimball 1-2 |
| 22-Jan | - Physical data mart design 1 | Han 3<br>Kimball 18 |
| 29-Jan | - Physical data mart design 2 | Han 3<br>Kimball 18 |
| 05-Feb | - Data Staging | Kimball 19-20 |
| 12-Feb | - Online Analytical Processing (OLAP) | Notes<br>Han 3 |
| 19-Feb | *Reading Week* | |
| 26-Feb | *Midterm on Tuesday February 27th during lecture time*<br>- Project Joint Application development on Friday March 2nd | All up to now<br>Notes |
| 05-Mar | - Implementation Guidelines and Project Guidance<br>*Midterm Review* | Notes |
| 12-Mar | - Finding grouping: Cluster analysis | Han 10 |
| 19-Mar | - Finding Patterns and Associations | Han 6 |
| 26-Mar | - Classification, Outliers and Anomalies 1<br>*Easter Break from Friday March 30th to Monday April 2nd* | Han 9, 11 |
| 02-Apr | - Classification, Outliers and Anomalies 2 | Han 9, 11 |
| 09-Apr | - Data at Scale: Perspective on Big Data Analytics<br>*Last class on Tuesday April 10th* | Kimball 21 |

Some important rules to remember:

- As per academic regulations, class attendance is mandatory. Also, as per academic regulations, students should attend 80% of the class to be allowed to write the final examinations.
- All components of the course must be fulfilled; otherwise students will receive an INC as a final mark (equivalent to an F).
- For more information about academic fraud regulations, please visit the following uOttawa website: www.uottawa.ca/academic/info/regist/crs/0305/home_5_ENG.htm)