CIS*6180/Data*6300 (Winter 2025) (0.5 Credits)
Analysis of Big Data
School of Computer Science
University of Guelph

Instructor Information:

Instructor: Dr. Gurjit Randhawa
Office Location: Reynolds Building Rm 3326
Office hours: By appointment
Email: randhawg@uoguelph.ca
**for course-related emails, please put "$CIS*6180$" or "$Data*6300$" in the subject line*
Teaching Assistants: Dylan Lewis (dlewis14@uoguelph.ca), Office hours: TBD
Sakib Shahriar (shahrias@uoguelph.ca), Office hours: TBD

Course Schedule:

Lectures:
- *Time:* Fridays - 11:30 am to 2:20 pm
- *Location:* REYN 1101
***The first two lectures (Jan 10 and Jan 17) will be held at SSC 1304.**

Course Website:
Course material, announcements, and grades will be regularly posted on the course website, which can be found at https://courselink.uoguelph.ca/d2l/home/928577.

Course Description:

This course introduces software tools and data science techniques for analyzing big data. It covers big data principles, state-of-the-art methodologies for large data management and analysis, and their applications to real-world problems. Modern and traditional machine learning techniques and data mining methods are discussed and ethical implications of big data analysis are examined.

Required preparation

At the start of the course, students should be able to

- Program in one of the following:
  - MATLAB, Mathematica, R, Java, Python, C, or C++
  - Use of any other programming language/tool requires permission from the instructor

- Design algorithms and analyze an algorithm's complexity
- Work with basic SQL queries and Linux commands
- Open to learning basic concepts from multiple different disciplines

Learning Outcomes:

After completing this course, students will be able to:

- Understand and apply foundational concepts in data analysis, including similarity measures, dimensionality reduction, clustering, and classification techniques.
- Use software tools and frameworks, such as Hadoop and SQL-based systems, for querying, managing, and analyzing large datasets.
- Analyze and preprocess large-scale datasets to extract meaningful patterns and insights using machine learning and data mining techniques.
- Address ethical, legal, and governance challenges in big data analysis, with a focus on privacy, fairness, and compliance with regulations.
- Create and interpret data visualizations to effectively communicate analytical results.
- Design and execute a project that demonstrates the integration of data analysis methods, tools, and ethical considerations in solving real-world problems.

Recommended textbooks:

A textbook is not required. Course notes will be provided on the course website.
Recommended readings:

- Data Science Mindset, Methodologies, and Misconceptions, Zacharias Voulgaris, 2017, Technics Publications.
- The Data Science Handbook, Field Cady, 2017, John Wiley & Sons, Inc.
- Artificial Intelligence: A Modern Approach, 4th Ed., Stuart J. Russell and Peter Norvig, 2021, Prentice Hall.
- Data Mining: Practical Machine Learning Tools and Techniques, Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal., Morgan Kaufmann, 2016.
- Handbook of Statistics # 24 - Data Mining and Data Visualization, Edited by C.R. Rao, E.J. Wegman, J.L. Solka; Elsevier B.V.; 2005, Volume 24, Pages 1-644; ISBN: 0-444-51141-5, ISSN: 0169-7161.
- Applied Multivariate Statistical Analysis (Classic Version), 6th Edition - Richard A. Johnson and Dean W. Wichern.; ISBN-13: 9780134995397.

Course Overview:

Week	Dates	Topics
1	Jan 10	Course Introduction; Brief review of the fundamental concepts; Similarity and dissimilarity (distance) measures: Euclidean distance, Minkowski distance, Manhattan distance, Pearson distance, etc.
2	Jan 17	Data visualization and Dimensionality reduction techniques: Histogram, Scatter plot, Boxplot, Principal component analysis, Linear Discriminant Analysis, Classical multidimensional scaling, t-SNE etc.
3	Jan 24	Classification in Machine Learning: Linear Discriminant Analysis, Support Vector Machines, Decision trees, etc.
4	Jan 31	Clustering analysis: k-means algorithm, Partitioning around medoids, DBSCAN, Hierarchical Clustering, etc.
5	Feb 7	Project presentations covering problem statement, background and literature review, dataset description, project plan, etc.
6	Feb 14	Deep Learning
7	Feb 21	**Winter break**
8	Feb 28	**Midterm exam**
9 - 10	Mar 7, 14	Big Data Tools and Frameworks: Introduction to distributed systems for big data, including Hadoop and MapReduce, with an overview of SQL-based tools like Hive and their role in querying large datasets.
11	Mar 21	Ethics and Governance in Big Data: Discussion on ethical, legal, and social implications of big data, focusing on privacy, security, and responsible data management practices.
12 - 13	Mar 28, Apr 4	Final Project Presentations covering design/methodology, results, discussion, conclusion and future work, etc.

The specified list of topics is tentative and will be adjusted as needed to keep the course flexible and cover the most material.

Student Evaluation:

Item	**Release**	**Due**	**Weight(%)**
Assignment 1	February 3	February 14	15
Midterm test	February 28 - in class		25
Assignment 2	March 10	March 21	15
Group project	***Deliverable***	***Due***	
	First presentation	February 7	5
	Report-part#1	February 7	5
	Report-part#2	March 7	5
	Report-complete	April 4	10
	Implementation	April 4	10
	Final presentation	March 28, April 4	10
**Report-part#1: Title, 1. Introduction (including problem statement, background, and detailed literature review), 2. Materials and methods (Dataset details)*			
**Report-part#2: build on part#1; complete 2. Materials and methods*			
**Report-complete: Title, 1. Introduction, 2. Materials and methods, 3. Results, 4. Discussion, 5. Conclusion, References*			

- All assignments are individual assignments.
- All submissions will be due at 11.59 pm on the date indicated.
- *Midterm test:* The test covers all the material presented in lectures up to the winter study break. It will be an in-class test and will contribute 25% towards your final grade.
- *Group project:* More details about the various project components will be shared later.

Lateness and Incomplete work:

Late submissions will not be accepted. Any assessment that you miss will automatically be given a grade of 0. No make-up of any components will be given. This includes failing to complete Programming Assignments or Project deliverables on time, and it includes failing to participate in the midterm examination.

However, if you have a valid reason for missing one or more components of your grade under extenuating circumstances, please contact the instructor within one week of the missed component deadline. Appeals after this period may not be considered. Please note that a vacation is not a suitable reason for missing work, just as having last-minute technological troubles is not a valid excuse. Remember that technology can be fickle; do not wait until the last minute to complete your work!

Accessibility

The University of Guelph is committed to creating a barrier-free environment. Providing services for students is a shared responsibility among students, faculty and administrators. This relationship is based on respect for individual rights, the dignity of the individual and the University community's shared commitment to an open and supportive learning environment. Students requiring service or accommodation, whether due to an identified, ongoing disability, or for a short-term disability should contact Student Accessibility Services (SAS) as soon as possible. For more information, contact SAS at 1.519.824.4120 ext 56208 or accessibility@uoguelph.ca or wellness.uoguelph.ca/accessibility.

Academic Integrity

The University of Guelph is committed to upholding the highest standards of academic integrity and it is the responsibility of all members of the University community faculty, staff, and students to be aware of what constitutes academic misconduct and to do as much as possible to prevent academic offences from occurring. University of Guelph students have the responsibility of abiding by the University's policy on academic misconduct regardless of their location of study; faculty, staff and students have the responsibility of supporting an environment that discourages misconduct. All students who take a SoCS course must pass the Academic Integrity Self Test. For educational purposes, instructors impose conditions on assignments that may limit students' permission to collaborate with others or to utilize external sources (including, but not limited to, software, data, images, text, etc.). Any permitted utilization must be done with proper references. Aiding and abetting is a punishable offence; students must be careful not to help others commit offences by giving out solutions or providing access to computer accounts. Instructors may use automated tools to detect possible cases of academic misconduct. Please note: Whether or not a student intended to commit academic misconduct is not relevant for a finding of guilt. Hurried or careless submission of assignments does not excuse students from responsibility for verifying the academic integrity of their work before submitting it. Students who are in any doubt as to whether an action on their part could be construed as an academic offence should consult with a faculty member. The Academic Misconduct Policy is detailed in the Graduate Calendar here.