

Draft syllabus. It will be finalized by January 8.

CSCI-4146/6409 --- Process of Data Science

Course Syllabus – Winter 2024

1. Instructor Information

Instructor:	Dr. Evangelos Milios	Office:	Online (Teams)
E-mail:	eem@cs.dal.ca	Office Hours:	Wed 0900-1000 Book a 10min slot here. Send me a personal msg on Teams at the time of the mtg.
Class Meeting Time:	TUE & THU 1600-1730	Class Room No:	Henry Hicks Acad 212 & Online (Teams) MS Team code:: sbtmnit
Tutorial Meeting Time:	FRI 0835-0925 (Foundations Review) FRI 1005-1125 (Regular)	Tutorial Room No:	Online (Teams) Goldberg 127 (for the Midterm exam)
Course Homepage:	https://dal.brightspace.com/d2l/home/312014		
Teaching Assistants	<ul style="list-style-type: none">TA1 Maksym Taranukhin, mk630647@dal.caTA2 Aman JaiswalTA3 Md Mahbub Alam	TA Office hours	On demand: Student requests a meeting from Maksym by email. Please do not send him private messages on Teams
Course Mail List:	---	MS Team Code	sbtmnit

2. Important Dates

- Summer term begins: **Jan. 8**
- Last day to register: **Jan. 22**
- Munro Day - University closed: **Feb. 2**
- Last day to drop without penalty: **Feb. 6**
- Nova Scotia Heritage Day - University closed: **Feb 19**
- Winter break: **Feb 19-23**
- Last day to drop with penalty: **Mar. 6**
- Good Friday - University closed: **Mar. 29**
- Friday classes will be held: **Apr. 8 & Apr. 9**
- Classes end: **Apr. 9**
- Final Exam period: **Apr. 11-23**
- [Important Dates - Academics - Dalhousie University](#)
- [Dalhousie Academic Calendars - View Calendar](#)

3. Course Description

A recent article in the Harvard Business Review regarded "Data Scientist" as the most appealing job of the 21st century. There are several reasons for this claim, but the main one comes from the diverse set of skills needed and the shortage of professionals with such background or experience. So, what's data science about after all? It's about asking the right questions to transform data into business value using statistics and algorithms. While other fields concentrate on finding previously unknown knowledge or searching for a specific pattern, data science focuses on answering deep questions and making conclusions understandable to the rest of the organization.

[Computer Science Academic Calendars](#)

4. Learning Outcomes

- Develop data analytic thinking: connecting business problems with data science solutions
- Learn to use the CRISP data mining process
- Learn to apply standard data exploration methods
- Apply predictive modelling and attribute selection with decision trees
- Apply model fitting to real data: classification, regression, support vector machines
- Understand and avoid overfitting using learning curves
- Understand and apply similarity computation
- Perform and analyze hierarchical clustering, learning cluster descriptions with supervised learning
- Evaluate machine learning models using confusion matrices, in the presence of unbalanced classes, cost-sensitive classifications
- Visualize model performance: ROC curves, AUC, Lift curves
- Understand probabilistic models: Naive Bayes, Evidence Lift
- Applying data science software in Python: SciKit Learn, NumPy, NLTK
- Solving data science problems for four different domains: spatio-temporal, stream, text mining, and risk/fraud analytics.

5. Course Rationale

This course offers undergraduate and graduate students an understanding of data science, including theoretical content and practical examples applied to large and complex datasets covering real-world scenarios. It explores content in machine learning, data mining and data analytics, focusing on industry context by answering questions and making conclusions accessible to others.

6. Class Format and Course Communication

- Class activities include:
 - o **"Before" the lecture:** Online videos and readings will be assigned before each class meeting to support students in the lecture understanding, which can be completed at their convenience before the class (asynchronously).
 - o **Lectures:** These lectures will be synchronous and recorded. Then the instructor goes deeper in the revised concepts and/or solves problems on the class content in a collaborative manner with the class (synchronously).
 - o **Tutorials** will be run synchronously and recorded. Tutorials will review background material required for the course and discuss practice problems through Python notebooks that will help students do the experimental questions in the assignments.
- **Assignments** will be submitted electronically to Crowdmark.
- All students must join the course MS Team using this code: **sbtmnit**
To join, open MS Teams, go to "Join or create a team" at the bottom of your teams listing, then "Join a team with a code", enter the above code.
- The class will make use of the following modes of interaction:
 - o Synchronous sessions will be available in-person, broadcast and recorded over **MS Teams**
 - o Students are encouraged to **self-organize** by creating study groups as private (no instructor or TA participation), or connect on other media of their choice (e.g. Facebook, Whatsapp or Discord).

- Course announcements will be distributed via email or MS Teams. It is the student's responsibility to check Brightspace, Teams and their Dal e-mail on a daily basis. To access your Dal e-mail see: <https://www.dal.ca/dept/its/o365/services/email.html>

7. Evaluation Criteria

The evaluation criteria depend on the specific course you are taking.

- **CSCI 4146**
 - o **Assignments (4%)**
 - Four assignments of equal weight
 - Late assignments will not be accepted.
 - All assignments will be checked for plagiarism.
 - Assignments will be done in teams of 3.
 - o **Midterm exam (48%)**
 - It will be taken individually, in-person, in-class, closed book
 - o **Project (46%)**
 - Late projects will not be accepted.
 - A project proposal and a project presentation will be required.
 - The project will be checked for plagiarism.
 - The project will be done in teams of 3.
 - The project mark will be based on oral presentations (proposal, progress report, final presentation) followed by oral examinations
 - o **Seminar presentation quizzes (2%)**
 - Short quizzes on understanding the seminar presentations
 - One quiz per presentation, est ~20 quizzes
 - Individual
 - pass/fail of a random subset of seminar presentations
- **CSCI 6409**
 - o **Assignments (4%)**
 - Four assignments of equal weight
 - Late assignments will not be accepted.
 - All assignments will be checked for plagiarism.
 - Assignments will be done in teams of 3.
 - o **Midterm exam (28%)**
 - It will be taken individually, in-person, in-class, closed book
 - o **Seminar presentations (20%)**
 - Group seminar in-class presentations of research papers in Data Science
 - o **Project (46%)**
 - Late projects will not be accepted.
 - A project proposal and a project presentation will be required.
 - The project will be checked for plagiarism.
 - The project will be done in teams of 3.
 - The project mark will be based on oral presentations (proposal, progress report, final presentation) each followed by a Q&A session.
 - o **Seminar presentation quizzes (2%)**
 - Short quizzes on understanding the seminar presentations
 - One quiz per presentation, est ~20 quizzes
 - Individual
 - pass/fail of a random subset of seminar presentations
- **Math quiz (P/F):** An optional short diagnostic math quiz must be completed before **January 20**. If you get an F, you should make an appointment to discuss your math preparedness for the course with the instructor.
- **Academic Integrity Module** <https://dal.brightspace.com/d2l/home/178166> (Required):
 - o Must be completed by February 1 to pass the course
 - o Must receive 75% or better on each of the four quizzes in this module

Notes

- **Grading:** All class components will be graded in the range 0-10, based on the rubrics published in section 16 of this syllabus.
- In CSCI 4146, a minimum grade of C (or 6/10) is required in this course if it is core to your FCS degree, or if it will be used as a prerequisite for a subsequent CSCI course.
- As of 2019, undergraduate students who receive a grade lower than C (or 6/10) in the same required CS course twice, will be dismissed.
- In CSCI 6409, a minimum grade of B- (7/10) is required to pass the course.
- Overall average will be a number 0-10 (no rounding on intermediate calculations of averages). The final letter grade will be based on the natural breaks in the sorted list of overall averages for all students.
- A student of CSCI4146 must pass (6/10) all three components of the course individually (assignments, midterm and project) to pass the course.
- A student of CSCI6409 must pass (7/10) all four components individually (assignments, midterm, seminar presentations and project) to pass the course.
- The meaning of the numeric and letter grades is described in [Section 17.1 of the Academic Regulations, Undergraduate Calendar](#)

8. Tutorials

Two tutorials will be offered every week, the **regular tutorial** and the **foundations' review tutorial**. The regular tutorial will elaborate on the lecture content. The foundations' review tutorials aim to support students who have not taken a machine learning course or never used Python and sklearn. Both tutorials will be offered online and will be recorded on Teams. All students are expected to have taken courses on (and be willing to review) calculus, probability and statistics, and linear algebra.

9. Student Declaration of Absence

The Student Declaration of Absence policy shall apply to the assignments only. https://www.dal.ca/campus_life/safety-respect/student-rights-and-responsibilities/academic-policies/student-absence.html
The student has a maximum of two (2) SDAs per course per semester. The student **must** notify the instructor of their inability to meet a deadline **before** the deadline by contacting the instructor or submitting the completed SDA. Upon notification, the student has 3 days after the deadline to submit the SDA.

10. Academic Standards

Failure to properly attribute sources in your work will be treated as an academic standards issue and points may be deducted for not following citation requirements. For example, forgetting to quote text taken from other sources, failure to include in-text citations, or failure to include required information in the citations or references. Please see the resources on proper citation provided by the Dalhousie Writing Center (<https://dal.ca/libguides.com/c.php?g=257176&p=5001261>).

When a source is referenced, the reference should be specific, for example, the specific web page instead of a large website, and a chapter instead of just the book title. The reader should be able to locate the relevant passage being referenced without search.

Please note that if it appears that the error was made with intent to claim other people's work as your own such as a lack of both citations and references, an allegation of plagiarism will be submitted to the Faculty Academic Integrity Officer, which could result in consequences such as a course failure.

If in doubt, always ask me, and consult the Academic Integrity website, https://www.dal.ca/dept/university_secretariat/academic-integrity.html

Use of Generative AI (e.g. chatGPT)

Use of Generative AI (Gen AI) is allowed, and it may be a helpful tool in understanding the course material. Two things to keep in mind are: (a) Gen AI is not a search engine, and you cannot rely on the existence or correctness of sources that it returns in response to a search-like prompt. Gen AI often hallucinates, i.e. it may produce incorrect facts or even fictitious references. You should diligently proofread and fact-check using good judgment all answers returned by Gen AI. Gen AI-supported search engines like Bing Chat are not free of hallucinations; (b) Gen AI tends to produce fluffy and vague text in response to general prompts. You should provide prompts that are as specific and complete as possible (to constrain Gen AI) and carefully edit the returned text to ensure that it is tight. Fluffy and vague text and hallucinatory responses in

https://www.dal.ca/dept/clt/e-learning/AI_Resource.html
https://www.dal.ca/dept/clt/e-learning/AI_Resource/communicating-with-your-students.html

For the up-to-date schedule [visit here](#)

Week	Month	Date	Day	Lecture topic	Suggested Readings and Materials						Regular tutorial topics (F 1000-1130)	Fundamentals Review tutorial topics (F 0830-0930)	Assignments, quizzes and Project	quiz
					DSB	FMLPDA, 2e								
1	Jan	9	Tu	Data Analytic Thinking	Ch 1: Data analytic thinking	Ch 1 - ML for Predictive data analytics	1					FRT0: Tools for data science (recording)	Math diagnosis quiz	
				Data Science Solutions	Ch 2: Business problems & Data science solutions	Ch 2 - Data to Insights to decisions								
				Data exploration		Ch 3, Data exploration, App. A	1	2	3	4				
		11	Thu	Predictive Modelling (entropy graphs, decision tree induction)	Ch 3: Intro to predictive modelling - decision trees	Ch 4 - Information based learning					T1: Data science project structure / business understanding	FRT1: Building models of data (decision trees)	Assignment 1 out (Data Exploration)	
				-->										
2		16	Tue	Fitting a model to data (optimizing loss function, SVM, Logistic Regression)	Ch 4: Fitting a model to data	Ch 7 - Error based learning								
		18	Thu	Fitting a model to data (optimizing loss function, SVM, Logistic Regression)										
				-->										
		19	Fri								T2: Data exploration	FRT2: Building models of data (error-based models)		
3		23	Tue	Classification and Regression Trees	Ch 12 - Ensemble models		1	2						
		25	Thu	Ensembles & Random Forests, XGBoost (Regression and Classification)			1	2	4	5	6			
				-->									Assignment 1 due - Assignment 2 out (Predictive modeling)	
		26	Fri								T3: Families of supervised models	FRT3: Ensemble learn		
4		30	Tue	Project proposal presentations ~10min x 8										
	Feb	1	Thu	Project proposal presentations ~10min x 8										
		2	Fri	Munro Day - University closed									Days W 3-5	
5		6	Tue	Neural Networks-Deep learning		Ch 8 - Deep learning							Assignment 2 due - Assignment 3 out (Ensemble methods)	
		8	Thu	Neural Networks-Deep learning		Ch 8 - Deep learning								
		9	Fri	-->										
6		13	Tue	Neural Networks-Deep learning							T4: Ensemble learning	FRT4: Deep learning intro		
		15	Thu	T5: Case study										quiz
		16	Fri	-->										
		20	Tue	WINTER BREAK							T6: Deep learning			
		22	Thu	WINTER BREAK										
		23	Fri	WINTER BREAK										
7		27	Tue	Project Progress Report - 5min each										
		29	Thu	Overfitting and its avoidance, What is a good model	Ch 5: Overfitting and its avoidance, Ch 8: Visualizing model performance	Ch 9 - Evaluation								
					Ch 7: Decision Analytic thinking - what is a good model,	Ch 9 - Evaluation							Assignment 3 due - Assignment 4 out (Overfitting, Cost sensitive learning)	
Mar		1	Fri	Cost-sensitive learning							T7: Performance metrics - comparing models			
8		5	Tue	Similarity & Clustering	Ch 6 - Similarity, neighbours and clusters	Ch 5 - Similarity based learning, Ch 10 - Unsupervised learning	DBScan							
		7	Thu	Similarity & Clustering			1							
		8	Fri											
9		12	Tue	Presentation & reporting skills seminar									Assignment 4 due	
		14	Thu	Presentation & reporting skills seminar										
		15	Fri	MIDTERM EXAM (IN-CLASS)										
10		19	Tue	Intro-Text mining			1							
		21	Thu	Intro - Spatiotemporal Data			1	2	3	4				
		22	Fri	Intro - Streaming data			1	2	3	4				
11		26	Tue	Presentations-Text mining (10min x 8)										
		28	Thu	Presentations - Spatiotemporal Data (10min x 8)										
		29	Fri	Presentations - Streaming data (10min x 8)										
12	Apr	2	Tue	Final project presentations (20min/pres x 4)										
		4	Thu	Final project presentations (20min/pres x 4)										
		5	Fri	Final project presentations (20min/pres x 4)										
13		8	Mon	Final project presentations (20min/pres x 4)										
		9	Tue	Final project presentations (20min/pres x 4)										

12. Suggested Texts and Resources

- **Main textbooks**

- [DSB] Foster Provost and Tom Fawcett: [Data Science for Business](#), O'Reilly, 2013 - [Novanet e-book](#) - [Dal e-book](#)
- [FMLPDA] J. Kelleher, B. MacNamee and Aoife D'Arcy: [Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples and Case Studies](#) - MIT Press, 2nd ed. 2020)
- [KT] J. Kelleher and Brendan Tierney: [Data Science](#), MIT Press, 2018. - [Dal e-book](#)

- **Programming oriented textbooks for Machine Learning**

- [AG] Aurelien Geron: [Hands-on Machine Learning with Scikit-Learn and Tensorflow](#), O'Reilly, 2017. [Dal e-book](#)
- [FC] Francois Chollet: [Deep Learning with Python](#), Manning, 2018
- [BKL] Steven Bird; Ewan Klein; Edward Loper: [Natural Language Processing with Python](#), O'Reilly Media, Inc. 2009 (available in html form [here](#))
- [RARS] R. Arumugam and R. Shanmugamani: [Hands-on Natural Language Processing with Python](#), Packt, 2018
- [BSD] B. Srinivasa-Desikan: [Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras](#), Packt, 2018.
- [WM] Wes McKinney: [Python for Data Analysis](#), O'Reilly, 2013 - [Dal e-book](#)
- [HRL] T. Hope, Y. Resheff, I. Lieder: [Learning TensorFlow: a Guide to Building Deep Learning Systems](#), O'Reilly. 2017. [Dal e-book](#)
- [NS] Nishant Shukla with Kenneth Fricklas: [Machine Learning with TensorFlow](#), Manning, 2017
- [TS] T. Segaran: [Programming Collective Intelligence: Building smart Web 2.0 applications](#) O'Reilly, Aug. 2007, 1st ed., ISBN 0-596-52932-5 (available on [Safari](#))

- **Online Tutorials**

- [Practical Machine Learning Tutorial with Python](#) (series of short tutorials)

- **General Programming Resources**

- [ML] Mark Lutz: [Learning Python](#), 5th Edition, O'Reilly Media, Inc. 2013
- [KC] Kristina Chodorow: [MongoDB: The Definitive Guide](#), O'Reilly, 2nd Ed., 2013. [Safari](#).
- [KKWZ] H. Karau, A. Konwinski, P. Wendell, M. Zaharia: [Learning Spark: Lightning Fast Data Analysis](#), O'Reilly, 2015. [Dal e-book](#).
- Petar Zecevic and Marko Bonaci: [Spark in Action](#), Manning, 2016

- **Conceptual and algorithms textbooks**

- [LRU] Jure Leskovec, Anand Rajaraman, Jeff Ullman: [Mining of Massive Datasets](#), Cambridge Univ. Press, 2nd ed, 2014 (pdf available from [mmds.org](#))
- [MR] Matthew Russell: [Mining the Social Web](#), O'Reilly, 2011.
- [BL] Bing Liu: [Web Data Mining](#), Springer, 2007, ISBN: 978-3642072376 (view online or download from [Springerlink](#))
- [MRS] Manning, Raghavan and Schuetze: [Introduction to Information Retrieval](#), Cambridge University Press, 2008
- [GBC] Ian Goodfellow and Yoshua Bengio and Aaron Courville, [Deep Learning](#), MIT Press, [pdf](#) [html](#)
- [CA] Charu Aggarwal, [Data Mining: The Textbook](#) (Springer), May 2015. (pdf accessible through Dalhousie library)

- **Data visualization**

- [SM] Scott Morgan, [Interactive Data Visualization for the Web](#), O'Reilly, 2013

- **Statistics basics**

- [LSH] David M. Lane, David Scott, Mikki Hebl, Rudy Guerra, Dan Osherson, and Heidi Zimmer, [Introduction to Statistics](#)
- [ID] Barbara Illowsky, Susan Dean, [Collaborative Statistics](#), Connexions 2013

- **Applications textbooks**

- [SK] Scott Spangler and Jeffrey Kreulen: Mining the Talk: [Unlocking the business value in unstructured information](#) ISBN-10: 0-13-233953-6; ISBN-13: 978-0-13-233953-7; Published: Jul 19, 2007; Edition: 1st.

13. Machine Learning Resources

- **Datasets**
 - [UCI Machine Learning Repository](#) (data sets)
 - [Kaggle](#)
 - [Canadian Open Government Data](#)
 - [OECD Open Government Data](#)
 - [USA Open Government Data](#)
 - [Nova Scotia Open Government Data](#)
 - [India Open Government Data](#)
- **Python data analysis and machine learning resources**
 - **Anaconda.** A Python Data Science Platform that includes Python all packages below. Install from <https://www.anaconda.com/download/>. Pick the option that installs Python 3. Before installing, uninstall a previous version if you have one (on Windows uninstall Python from add/remove programs of the Control Panel). Launch Anaconda Navigator from the desktop, and launch Jupyter Notebook. Load a notebook from the Jupyter Notebook browser tab.
Read [markdown documentation](#) for formatting text in Jupyter notebooks. To expose the formatting, execute the text cell. To introduce a new line, enter two spaces and return.
See the [markdown cheatsheet](#).
 - **Math:** mathematical functions <https://docs.python.org/3/library/math.html>
 - **NumPy:** Numerical computation in Python, <https://numpy.org/>
 - **SciPy:** optimization, integration, interpolation, eigenvalue computations, algebraic equations, differential equations, statistics etc. <https://scipy.org/>
 - **Pandas:** basic data analysis in Python, <https://pandas.pydata.org/>
 - **Matplotlib:** Python Plotting, <https://matplotlib.org/>
 - **Statsmodels:** Statistical modelling in Python, <https://www.statsmodels.org/stable/index.html>
 - **PyStan:** Bayesian inference in Python, <https://pystan.readthedocs.io/en/latest/>
 - **SciPy.org** Python-based ecosystem of open-source software for mathematics, science and engineering.
 - **Scikit-learn** Python-based machine learning library
 - **Keras:** the Python Deep Learning library
 - **Pytorch:** another library/environment for Deep Learning
 - **Keras vs Pytorch**
 - **Google Colaboratory.** A programming and execution environment for Python as a web service that allows collaborative editing of python notebooks, with GPU acceleration on top!
 - **Gensim:** topic modelling for humans, <https://radimrehurek.com/gensim/>
- **High performance computing & data management**
 - **Mongodb,** <https://www.mongodb.com/>
 - **Apache Spark:** <https://spark.apache.org/>
 - **Apache Mahout** Scalable machine learning library (recommendation mining, text clustering, text classification) on Hadoop
- **Visualization**
 - **D3,** <https://d3js.org/>
 - **Gephi:** the open graph viz platform, <https://gephi.org/>
 - **Mayavi:** 3D Data visualization in Python. [link](#)
- **Other tools and resources**
 - [Reinforcement Learning Repository at UMass](#)
 - [Scilab: a free scientific software package](#) (a free equivalent of Matlab)
 - [Octave](#) (a free GNU project alternative to Matlab)
 - [WEKA:](#) Data Mining Software in Java
 - [RapidMiner](#) (formerly YALE: Yet Another Learning Environment (incorporates WEKA))
 - [Dragon toolkit for Language Modeling, Text Retrieval, and Text Data Mining](#)
 - [LingPipe](#) (NLP applications of ML)
 - [Topic Modelling Toolbox](#) (Matlab) using Latent Dirichlet Allocation (LDA)
- **Large scale project resources**

- [Splunk](#), searching, monitoring and analyzing machine generated data
- [Kafka](#), a distributed streaming platform
- [Scrum](#), agile team software development
- [Kanban](#), process management and improvement method
- **Online Kaggle micro Courses**
 - [Data Cleaning](#).
 - [Intermediate Machine Learning](#)
 - [Machine Learning explainability](#)
 - [Data Visualization](#)
 - [Feature Engineering](#)

14. Related courses

CSCI 3151: Foundations of Machine Learning

CSCI 3154: Introduction to AI with Gaming Applications

CSCI 3172: Web-centric computing

CSCI 4141: Information Retrieval (not offered)

CSCI 4152/6509: [Natural Language Processing](#)

CSCI 4155/6505: [Machine Learning](#)

CSCI 4144/6405: Data Mining & Data Warehousing

CSCI 6516: Deep Learning

CSCI 4163/6610: Human-Computer Interaction (support for human interaction with data)

CSCI 4166/6406: Visualization (data and information visualization is a prerequisite for human interaction with data)

CSCI 6515: Machine Learning for Big Data

CSCI 6612: Visual analytics

Prerequisites

CSCI-3130, and CSCI-3151

15. Standard syllabus content

Responsible Computing Policy

Usage of all computing resources in the Faculty of Computer Science must be within the Dalhousie Acceptable Use Policies (<http://its.dal.ca/policies/>) and the Faculty of Computer Science Responsible Computing Policy. For more information please see

https://www.dal.ca/content/dam/dalhousie/pdf/faculty/computerscience/policies-procedures/fcs_policy_local.pdf

Use of Plagiarism Detection Software

All submitted code may be passed through a plagiarism detection software, such as the plagiarism detector embedded in Codio, the Moss (<https://theory.stanford.edu/~aiken/moss/>) Software Similarity Detection System, or similar systems. If a student does not wish to have their assignments passed through plagiarism detection software, they should contact the instructor for an alternative. Please note, that code not passed through plagiarism detection software will necessarily receive closer scrutiny.

https://www.dal.ca/dept/university_secretariat/policies/academic/student-submission-of-assignments-and-use-of-originality-checking-software-policy-.html

Student Health and Wellness

Taking care of your health is important. As a Dalhousie student, you have access to a wide range of resources to support your health and wellbeing. Students looking to access physical or mental health & wellness services at Dalhousie can go to the Student Health & Wellness Centre in the LeMarchant Building. The team includes: registered nurses, doctors, counsellors and a social worker. Visit dal.ca/studenthealth to learn more and book an appointment today.

Students also have access to a variety of online mental health resources, including telephone/texting counselling and workshops/training programs. Learn more and access these resources at dal.ca/mentalhealth.

Culture of Respect¹

Every person has a right to respect and safety. We believe inclusiveness is fundamental to education and learning. Misogyny and other disrespectful behaviour in our classrooms, on our campus, on social media, and in our community is unacceptable. As a community, we must stand for equality and hold ourselves to a higher standard.

What we all need to do:

1. **Be Ready to Act:** This starts with promising yourself to speak up to help prevent it from happening again. Whatever it takes, summon your courage to address the issue. Try to approach the issue with open-ended questions like “Why did you say that?” or “How did you develop that belief?”
2. **Identify the Behaviour:** Use reflective listening and avoid labeling, name-calling, or assigning blame to the person. Focus the conversation on the behaviour, not on the person. For example, “The comment you just made sounded racist, is that what you intended?” is a better approach than “You’re a racist if you make comments like that.”
3. **Appeal to Principles:** This can work well if the person is known to you, like a friend, sibling, or co-worker. For example, “I have always thought of you as a fair-minded person, so it shocks me when I hear you say something like that.”
4. **Set Limits:** You cannot control another person’s actions, but you can control what happens in your space. Do not be afraid to ask someone “Please do not tell racist jokes in my presence anymore” or state “This classroom is not a place where I allow homophobia to occur.” After you have set that expectation, make sure you consistently maintain it.
5. **Find or be an Ally:** Seek out like-minded people that support your views, and help support others in their challenges. Leading by example can be a powerful way to inspire others to do the same.
6. **Be Vigilant:** Change can happen slowly, but do not let this deter you. Stay prepared, keep speaking up, and do not let yourself be silenced.

University Statements

This course is governed by the academic rules and regulations set forth in the University Calendar and the Senate. <https://academiccalendar.dal.ca/Catalog/ViewCatalog.aspx?pageid=viewcatalog&catalogid=111&loaduserredits=False>

Territorial Acknowledgement

Dalhousie University is located in Mi’kma’ki, the ancestral and unceded territory of the Mi’kmaq. We are all Treaty people. Dalhousie acknowledges the histories, contributions, and legacies of the African Nova Scotia people and communities who have been here for over 400 years.

Internationalization

At Dalhousie, ‘thinking and acting globally’ enhances the quality and impact of education, supporting learning that is “interdisciplinary, cross-cultural, global in reach, and orientated toward solving problems that extend across national borders.” <https://www.dal.ca/about-dal/internationalization.html>

Academic Integrity

At Dalhousie University, we are guided in all of our work by the values of academic integrity: honesty, trust, fairness, responsibility and respect. As a student, you are required to demonstrate these values in all of the work you do. The University provides policies and procedures that every member of the university community is required to follow to ensure academic integrity. (read more: http://www.dal.ca/dept/university_secretariat/academic-integrity.html)

¹ Source: Speak Up! © 2005 Southern Poverty Law Center. First Printing. This publication was produced by Teaching Tolerance, a project of the Southern Poverty Law Center. Full “Speak Up” document found at: <http://www.dal.ca/dept/dalrespect.html>. Revised by Susan Holmes from a document provided April 2015 by Lyndsay Anderson, Manager, Student Dispute Resolution, Dalhousie University, 902.494.4140, lyndsay.anderson@dal.ca www.dal.ca/think.

Accessibility

The Student Accessibility Centre is Dalhousie's centre of expertise for matters related to student accessibility and accommodation. If there are aspects of the design, instruction, and/or experiences within this course (online or in-person) that result in barriers to your inclusion please contact: https://www.dal.ca/campus_life/academic-support/accessibility.html for all courses offered by Dalhousie with the exception of Truro.

Conduct in the Classroom — Culture of Respect

Substantial and constructive dialogue on challenging issues is an important part of academic inquiry and exchange. It requires willingness to listen and tolerance of opposing points of view. Consideration of individual differences and alternative viewpoints is required of all class members, towards each other, towards instructors, and towards guest speakers. While expressions of differing perspectives are welcome and encouraged, the words and language used should remain within acceptable bounds of civility and respect.

Diversity and Inclusion — Culture of Respect

Every person at Dalhousie has a right to be respected and safe. We believe inclusiveness is fundamental to education. We stand for equality. Dalhousie is strengthened in our diversity. We are a respectful and inclusive community. We are committed to being a place where everyone feels welcome and supported, which is why our Strategic Direction prioritizes fostering a culture of diversity and inclusiveness (Strategic Priority 5.2). (read more: <http://www.dal.ca/cultureofrespect.html>)

Student Code of Conduct

Everyone at Dalhousie is expected to treat others with dignity and respect. The Code of Student Conduct allows Dalhousie to take disciplinary action if students don't follow this community expectation. When appropriate, violations of the code can be resolved in a reasonable and informal manner—perhaps through a restorative justice process. If an informal resolution can't be reached, or would be inappropriate, procedures exist for formal dispute resolution. (read more: https://www.dal.ca/dept/university_secretariat/policies/student-life/code-of-student-con.html)

Fair Dealing Policy

The Dalhousie University Fair Dealing Policy provides guidance for the limited use of copyright protected material without the risk of infringement and without having to seek the permission of copyright owners. It is intended to provide a balance between the rights of creators and the rights of users at Dalhousie. (read more: https://www.dal.ca/dept/university_secretariat/policies/academic/fair-dealing-policy-.html)

Originality Checking Software

The course instructor may use Dalhousie's approved originality checking software and Google to check the originality of any work submitted for credit, in accordance with the Student Submission of Assignments and Use of Originality Checking Software Policy. Students are free, without penalty of grade, to choose an alternative method of attesting to the authenticity of their work, and must inform the instructor no later than the last day to add/drop classes of their intent to choose an alternate method. (read more: https://cdn.dal.ca/content/dam/dalhousie/pdf/dept/university_secretariat/policy-repository/OriginalitySoftwarePolicy.pdf)

Student Use of Course Materials

These course materials are designed for use as part of the CSCI courses at Dalhousie University and are the property of the instructor unless otherwise stated. Third party copyrighted materials (such as books, journal articles, music, videos, etc.) have either been licensed for use in this course or fall under an exception or limitation in Canadian Copyright law. Copying this course material for distribution (e.g. uploading material to a commercial third party website) may lead to a violation of Copyright law.

Learning and Support Resources

Please see https://www.dal.ca/campus_life/academic-support.html

16. MARK BREAKDOWN AND RUBRICS FOR COURSE COMPONENTS

All course components will be submitted to Crowdmark.

Marked coursework (except the midterm exam) will be returned to students via Crowdmark.

Mark calculations will be done on a spreadsheet, not on Brightspace.

16.1. Assignments

Assignments are due at 2359 on the following dates:

	Due Date	Description
Assignment 1	26/01	Data exploration
Assignment 2	06/02	Predictive modeling
Assignment 3	01/03	Ensemble methods
Assignment 4	12/03	Overfitting, cost sensitive learning

All assignments will be graded using the following grading scheme.

Assignment Component	Assessment Method	Weight
Functionality	Inspection by TA/instructor	0.5
Structure and design	Inspection by TA/instructor	0.3
Code clarity and best practices	Inspection by TA/instructor	0.2
Total		1

Rubric for assignment marking

- Functionality.
 - All the required features implemented?
 - Are the boundary cases handled?
 - Are the employed techniques the best choice?
- Structure and Design
 - Is the code properly decomposed and uses appropriate design patterns?
 - Is the code reusable, where applicable?
 - Is the code well documented?
- Code Clarity
 - A+/A/A- Code is very well organized, commented, readable, and maintainable. The code looks professional.
 - B+/B/B- Code is competently organized, commented and understandable. Requires minor tuning to bring it up to professional-looking code.
 - C+/C/C- Code is poorly organized, uncommented, or confusing. Requires significant effort to bring it up to professional-looking code.
 - F Code is unintelligible

16.2. Project

The course project is the student's opportunity to develop skills and expertise in Visual Analytics. The course project is designed to take students through all stages of development, from coming up with a new idea for a visual analytics application to designing, implementing, testing, and demonstrating the application. For manageability, the project is divided into different deliverables: proposal, implementation, report, and project demonstration.

Project Groups

The project should be done in groups of two to three students. Groups will be defined two weeks before the project proposal submission. Once the groups are set, changing them will only be possible before the project proposal deadline.

Project Goal

The end goal of the project is visual analytics software for a specific domain. The goal is to design and implement an application that uses machine learning or data mining techniques in the back end and information visualization methods in the front end in an interesting and useful manner and provides a significant amount of functionality.

Project Deliverables

There are different project deliverables and four deadlines:

1. Project Proposal and presentation
2. Project Implementation
3. Project Demonstration
4. Project Report and Presentation

All deliverables, except the Project Demonstration, must be submitted electronically via the course website (Brightspace). No late submissions are accepted. Deliverables 1,3, and 4 include a slide deck presented orally to the class. The mark will be based on the quality of the oral presentation, the slides, the written report and the competence in answering questions after the presentation. The slide deliverables are due before the presentation. The report is due after the presentation and should incorporate any feedback received on the presentation.

16.2.1 Project Proposal

The project proposal sets the application domain and problems to be solved for the student application. The project proposal should answer many of the questions asked before any development and implementation takes place. The proposal is intended to cement what the student is planning to do and how much she is expecting to accomplish.

The proposal should be between five and seven pages in length (11-point Times Roman font, single-spaced, 1" margins) and should be divided into the following sections:

- **Title Page:** This includes the project title and **acronym**, the group members, and email addresses.
- **Introduction, Motivation, and Benefits:** This should set the stage for the proposal, describing what the project will be and why it is interesting and important. Students may also want to include any background or relevant information.
- **Description of the dataset:** Size, dimensionality, list of attributes and their type. Open Government Data is a great source of data and project ideas. See the [associated links](#) under Machine Learning resources in this syllabus. Kaggle has a lot of datasets and code from which you can draw inspiration for your project.
- **What data science problem will your system solve?** This should go beyond classification or regression. You may explore the dataset, the feature correlations, hyperparameter tuning, and the explanation of the resulting predictions.
- **Functionality:** This describes all the functions that the resulting application will perform, divided into three parts:
 - o *Minimum Functionality:* the minimum amount of functionality that the group expects to implement;
 - o *Expected Functionality:* the functionality the group expects to implement, assuming most things go according to plan. Present the additional functionality on top of the minimum.
 - o *Bonus Functionality:* the functionality you would like to implement, assuming you managed to get everything done a lot earlier than expected. This is your opportunity to be creative. Present the additional functionality on top of the expected.

Under each of these subheadings, enumerate the functionality that you plan or wish to implement. Achieving Minimum Functionality will typically yield a B- grade; achieving Expected Functionality will typically yield a B+/A- grade; and achieving Bonus Functionality will typically yield an A/A+ grade.

- **High-Level Organization and Design:** Describe the high-level organization of the application, i.e., a functional decomposition.
- **Milestones and Timeline:** Lastly, include a timeline with milestones for completing the project. It is recommended to have a set of milestones for each week. Consider assigning (initially) roles and milestones to each group member.

In your proposal presentation, use 10 slides for presenting and additional slides with details for answering questions.

This assumes the standard 1 slide/min for a 10min presentation, and up to 10min for Q&A.

Slides should have minimum font size of 18pt, Arial font. Avoid narrative text, prefer bullet points instead.

Put your explanations in speaker notes. Submit both your pptx file and its pdf printout.

Consult the guidelines for good technical presentations in <https://web.cs.dal.ca/~eem/gradResources/gradResources.htm#oral>

16.2.2 Project Implementation and Report

The project deliverable should comprise the source code for the project and a final report. The report should be 10 to 15 pages in length (11-point font) and be divided into the following sections:

- **Title Page:** This includes the project title, the group members, and email addresses.
- **Abstract:** A brief 100-word summary of your project and its components.
- **Report:** Body of the report, containing sections and sub-sections to:
 - o Explain your project (what problem are you solving), discuss its importance, your solution to tackling the problem and your novelties in both visualization and machine learning part;
 - o Justify and explain the machine learning approach you used for solving the problem;
 - o Justify and explain each module in your implementation.
 - o Explain your evaluation metrics and how you evaluate the quality of your proposed approach.
 - o Cite all tools, libraries and external resources you used in your project. You must explain what you added or changed to any existing code and resources that you built on.
- **Conclusions and Future Work:** What do you learn from this project and a discussion of how it can be extended or improved if you had more time and inclination to do so?

16.2.3 Project Demonstration

Near the end of the term, each group will give a 15 to 20-minute presentation/demonstration of their final application, identifying which parts of the project were achieved and explaining how their application works. The Functionality portion of the Implementation will be evaluated at this point. You get points based on the quality of the presentation, the importance of the problem and the proposed solution. This is a required part of the project. All presentations and Q&A will be recorded on MS Teams.

16.2.4 Project Evaluation and Grading

The project has the following breakdown:

- 20% Project Proposal and presentation (~Feb 1)
- 0% Progress report (~Mar 1)
- 20% Project Implementation (~Apr 1)
- 30% Project Demonstration (~Apr 1)
- 30% Project Report and presentation (~Apr 1)

The project proposal, final report, implementation, and demonstration will all be graded using the letter grade scale. All material is to be submitted electronically via the course Brightspace website.

Rubric for Project implementation

The implementation will be marked according to the following rubric:

	Exceptional: A+,A, A-	Acceptable: B+,B,B-	Substandard: C+,C,C-,D	Unacceptable: F
Functionality (70%)	All “expected functionality” and some “bonus functionality” is completed.	All “minimum functionality” and most “expected functionality” is completed.	All “minimum functionality” is completed.	Most of the “minimum functionality” is incomplete.
Code (30%) organization, commenting, readability, etc.,	Code is very well organized, commented, readable, and maintainable. The code looks professional.	Code is competently organized, commented and understandable. Requires minor tuning to bring it up to professional-looking code.	Code is poorly organized, uncommented, or confusing. Requires significant effort to bring it up to professional-looking code.	Code is unintelligible.

The Functionality part of the evaluation is assessed based on the Demonstration and report. Code quality will be evaluated by the instructor or TA at a later time.

Rubric for Written Submissions and Presentations

Each written submission will be graded according to the following rubric, based in part on Fleming, “Grading Rubric for Written Assignments,” CSCI 2100, 2011.

	Exceptional: A+,A, A-	Acceptable: B+,B, B-	Substandard: C+,C,C-,D	Unacceptable: F
Content (30%)	Ideas well organized and logically laid out always or almost always. Shows superior understanding of the subject.	Ideas well organized and logically laid out with competence. Shows commonplace understanding of the subject.	Minimal organization and logical progression of ideas. Shows partial familiarity with the subject.	Little or no organization or logical flow of ideas. Shows a great deal of misunderstanding about the subject.
Analysis (30%)	Addresses all points. Justifies all design decisions. Discusses criteria used in justifications. Considers the issues from multiple points of view.	Addresses most of the points. Justifies most design decisions. Identifies criteria used in justifications.	Addresses some of the points. Justifies some design decisions. Identifies some criteria in the analysis.	Addresses few of the points. Little or no justification and no identification of the criteria used.
Writing or Oral delivery (20%) Style and tone,	Shows exceptional use of tone and style. Speaks to the reader	Shows competent use of tone and style. Makes good word choices.	Shows minimal attention to tone and style. Shows poor usage or ineffective word	Shows little or no understanding of appropriate

standard conventions for grammar, word use, spelling, citations, headings, paragraphs, figures, and tables.	with precise, concise, appropriate language and choice of words. Always uses standard conventions. The document looks professional.	Mostly uses standard conventions. The document could use some editing.	variation. Does not consistently use standard conventions. The document requires significant editing.	tone. Uses inappropriate language and word choice. Standard conventions are flouted. Document is unreadable.
Question answering (20%)	Comprehensive answers to all the questions	Competent answers to most questions	Competent answers to some questions	Unable to answer questions

Note: To unlock the project grade, a minimum of B- (7/10) is required on all the Question Answering parts (proposal, demo and report). If this criterion is not met, the project receives a zero.

16.3. Midterm Exam Requirements

- A photo ID is required
- Closed book
- No dictionaries, notes, cell phones, PDAs, talking slide rulers, or other electronic aids are allowed.
- A non-programmable calculator is allowed.
- No questions to the proctor are allowed. If you don't understand something, make and clearly state a reasonable assumption and proceed.

16.4. Seminars

Groups of students (see "Project Group" section) will select research papers that are related to their project topic to present to their colleagues. As part of the presentation, students must answer these five standard questions according to the weights shown:

- | | |
|--|-----|
| 1. What is the problem the paper is trying to address? | 10% |
| 2. Why is this problem relevant? | 10% |
| 3. What are the solutions proposed in the papers (details)? | 20% |
| 4. How do the authors measure or assess the success of the work? | 20% |
| 5. What are the bottlenecks and limitations? | 20% |
| 6. What is the next step (future work)? | 20% |

Groups should select at least three papers for the presentation. The presentations will be approximately 10-15 minutes long, depending on the size of the class. The presentations will be recorded. The marks will be determined based on the oral presentation and slide quality based on the rubric for presentations.

16.5. Seminar quizzes

All students are expected to attend the seminar presentations will be required to submit their short answers on the key ideas of the seminar presentations **in their own words**. A random selection of seminar presentations (out of about 20) will be selected at the end of the presentations period for marking. A student who has attended the presentation and taken notes should be able to pass these quizzes. Presenters should not submit the quiz on their own presentation.

1. What is the key problem being addressed by the researcher?
2. What technique does the researcher use to solve the problem?
3. How does the author measure or assess the success of the work?

4. What is one specific concept or idea that you can take away from the seminar?
5. What question could you ask of the speaker?

The recommended process is to open a text or Word file with the above questions and fill in your answers while you watch the presentation. Aim to submit your answers on Crowdmark soon after.

You are expected to provide short answers (up to 60 words) to each question. Each of the five questions will be marked as pass/fail (1/0).

Your answers should reflect the fact that you paid attention and got some key messages from the seminar.