CSD 115: Elements of Data Science and Machine Learning

Cohort IV, Year I, Semester I 2024/2025

Instructor:	Alon B. Muhame	**Time:**	Tue, 4:30-6:30pm (in person)
Guest Instructor:	Caleb Bii	**Time:**	Wed, 4:30-6:30pm (Virtual)
Email:	alon.muhame@witu.org	**Location:**	Computer Lab Roof top

Course Pages:

1. `https://github.com/alon-muhame/witidatascienceml`

Office Hours: After class, or anytime when you see me around.

Course Description: This course is primarily designed for a beginner student interested in data science, machine learning and general data ethics in computing. The goal of this class is to provide a solid introduction to data science and some introductory algorithms used in data science, finding relationships in data, and distributions across data pattern with an emphasis on practical implementation in python coding environment. I will provide a set of resources including slides and notes, worked examples for code in python-jupyter notebook.

More generally, this is a course where I focus on providing my understanding and intuition of data science, as used by many practitioners and data scientists. This means that this is not a course where we will spend a lot of time on the formal mathematical derivations and other details (beyond what is necessary to get the work done), but instead focus on the important algorithms and frameworks that guides most of data science processes. I'll also do my best to communciate how any of these topics fit together. Key to note will be how these concepts and frameworks are applied in industry practice. This is a course very much focused on communication and artisanship of data and science. By the end of the course, my hope is for three things:

Course learning outcomes:

(a) Knowledge:

- Gain an understanding of data science and machine learning principles and their uses
- Learn to interpret and judge statistical/data science related information
- Learn how to effectively present data information to general audiences

(b) Skills:

- Setup a Python environment in the jupyter notebook ikernel/google colab
- Import and export of data from and to csv, excel and online databases
- Perform descriptive analysis of numerical data using jupyter notebook ipython kernel
- How to collect data, summarize data using graphical and numerical techniques, and examine data for patterns and relationships
- Analyze data so you can draw valid and appropriate conclusions using jupyter notebook ipython kernel

(c) Competences:

- Work collaboratively on code projects with other class members/group members via version control
- Use online communities to find existing code and get help

CSD 115: Elements of Data Science and Machine Learning

Cohort IV, Year I, Semester I 2024/2025

Instructor:	Alon B. Muhame	**Time:**	Tue, 4:30-6:30pm (in person)
Guest Instructor:	Caleb Bii	**Time:**	Wed, 4:30-6:30pm (Virtual)
Email:	alon.muhame@witu.org	**Location:**	Computer Lab Roof top

Course Pages:

1. `https://github.com/alon-muhame/witidatascienceml`

Office Hours: After class, or anytime when you see me around.

Course Description: This course is primarily designed for a beginner student interested in data science, machine learning and general data ethics in computing. The goal of this class is to provide a solid introduction to data science and some introductory algorithms used in data science, finding relationships in data, and distributions across data pattern with an emphasis on practical implementation in python coding environment. I will provide a set of resources including slides and notes, worked examples for code in python-jupyter notebook.

More generally, this is a course where I focus on providing my understanding and intuition of data science, as used by many practitioners and data scientists. This means that this is not a course where we will spend a lot of time on the formal mathematical derivations and other details (beyond what is necessary to get the work done), but instead focus on the important algorithms and frameworks that guides most of data science processes. I'll also do my best to communciate how any of these topics fit together. Key to note will be how these concepts and frameworks are applied in industry practice. This is a course very much focused on communication and artisanship of data and science. By the end of the course, my hope is for three things:

Course learning outcomes:

(a) **Knowledge:**

- Gain an understanding of data science and machine learning principles and their uses
- Learn to interpret and judge statistical/data science related information
- Learn how to effectively present data information to general audiences

(b) **Skills:**

- Setup a Python environment in the jupyter notebook ikernel/google colab
- Import and export of data from and to csv, excel and online databases
- Perform descriptive analysis of numerical data using jupyter notebook ipython kernel
- How to collect data, summarize data using graphical and numerical techniques, and examine data for patterns and relationships
- Analyze data so you can draw valid and appropriate conclusions using jupyter notebook ipython kernel

(c) **Competences:**

- Work collaboratively on code projects with other class members/group members via version control
- Use online communities to find existing code and get help

- Communicate precisely technical findings using visualization, tables, plots
- Ability to tell a story using data using jupyter-notebook ipython kernel

Assignments: There will be problem sets every week. These will involve mainly computer code exercises in which you will be asked to analyze data sets using the concepts and skills learnt. This class will continue using python and especially on the previous knowledge and concepts learnt in mathematical computing class. Solutions will be handed out written in Python (relying on the ipython kernel - Jupyter Notebooks). Since there will be a fair number of problem sets, and in order to allow me to post the solutions quickly on the webpage for the course, I will not accept late problem exercise. If you anticipate difficulty meeting the deadline, you can ask me for the problem exercises earlier to give you additional time to work on it. The assignment grades will be based on the problem exercises, divided evenly over the problem sets. The problem exercises will contribute 30% of the course marks and then the test will contribute 20%.

Exam: There will be two in class exams: test and final semester exam. The final examination will only contribute 50%. It will test understanding and application of theory and practical code. The exam will basically seek to examine transfer of knowledge from theory to practicals.

Mark-up policy: Make-up oral exams will only be given to student with written medical excuses at least 24 hours prior to the exam.

Academic integrity: You can work together on the problem exercises and discuss them with classmates, but you need to write up the code individually and hand them in separately.

Ownership of course materials: The course materials provided to you online and in class are yours for your personal use for the semester and to keep as a reference for your career. I expressly do not give permission for you to sell or freely distribute my course materials, including exams, quizzes, homework assignments, or solutions.

Is Maths helpful for this course? There are a lot of debates online about how much math a data scientist/ analyst really needs to know. This depends on the background and view of one has in a transition to data science or machine learning. There is however, a danger of applying methods when you don't know the mathematical basis for how they work is that you can make dangerous misinterpretations of their results.

This is particularly true for a good foundation in statistics, which is the bedrock of most modeling techniques in data science & machine learning . The idea that a data scientist or machine learning enthusiast may just run data science models without a bare foundational knowledge of some of the statistical theory and methodology underlying those models is a really a danger and may lead to disinformation.

My advice: Try to get friendly with technology early, it really can be your friend in this course. Not having access to technology (e.g., tools and a laptop), will never be a suitable justification for not completing an assignment.

Main References:

- Allen B. Downey, Elements of Data Science, `https://allendowney.github.io/ElementsOfDataScien`
- Thomas Nield, Essential Math for Data Science: `https://www.essentialmathfordatascience.com/`
- Kieran Healy, Data Visualization: A Practical Introduction: `https://socviz.co/`
- Yuli Vasiliev, Python for Data Science: `https://www.oreilly.com/library/view/python-for-data/9781098130275/`

Course Outline

Day	Date	Topic	Assignment/ Test/Exam
Tue	27/08	Elements of data science: variables/values, lists, arrays, loops and files, dictionaries, scheme for data science for beginners	
Tue	3/09	DataFrame and Series: importing and transforming data in csv, excel, txt	Ex 1
Tue	10/09	Plotting: One-, two- dimensional plots, line charts, scatter plots, Histograms, box plots,	
Tue	17/09	Distributions: probability (PDFs and CDFs) & simulations (using Jupyter notebooks), comparing distributions, modelling distributions	Ex 2
Tue	24/09	Describing date and relationships: Exploring data, relationships, correlation,	
Tue	1/10	Inference under linear regression: types of variables, assumptions, cost functions and intuition behind mathematics, fitting a linear regression, interpreting coefficients, visualizations of results,	Ex 3
Tue	8/10	Test I	Test 1
Tue	15/10	Hypothesis testing under linear regression	
Tue	22/10	Linear classifiers in Python: intro and application to a worked example	
Tue	29/10	Group based project	Group project
Tue	5/11	Sampling and resampling in Python: variance testing, resampling framework, means estimation,	Ex 4
Tue	12/11	Linear algebra and matrices behind recommendation systems	
Tue	19/11	Project for group: Building a recommendation matrix	Group project

Figure 1: My tentative plan for the semester

Day	Date	Topic	Assignment/ Test/Exam
Tue	27/08	Elements of data science: variables/values, lists, arrays, loops and files, dictionaries, scheme for data science for beginners	
Tue	3/09	DataFrame and Series: importing and transforming data in csv, excel, txt	Ex 1
Tue	10/09	Plotting: One-, two- dimensional plots, line charts, scatter plots, Histograms, box plots,	
Tue	17/09	Distributions: probability (PDFs and CDFs) & simulations (using Jupyter notebooks), comparing distributions, modelling distributions	Ex 2
Tue	24/09	Describing date and relationships: Exploring data, relationships, correlation,	
Tue	1/10	Inference under linear regression: types of variables, assumptions, cost functions and intuition behind mathematics, fitting a linear regression, interpreting coefficients, visualizations of results,	Ex 3
Tue	8/10	Test I	Test 1
Tue	15/10	Hypothesis testing under linear regression	
Tue	22/10	Linear classifiers in Python: intro and application to a worked example	
Tue	29/10	Group based project	Group project
Tue	5/11	Sampling and resampling in Python: variance testing, resampling framework, means estimation,	Ex 4
Tue	12/11	Linear algebra and matrices behind recommendation systems	
Tue	19/11	Project for group: Building a recommendation matrix	Group project

Figure 1: My tentative plan for the semester