CS2545 - Data Science for Big Data Analytics

Fall 2021 - Course Outline

Schedule:

CS2545	**M, W and F**	**12:30-1:20**	**Lecture: In-person* (HC10)** **Handson: ADM** **Lab for hands-on sessions –** Remote access only; *see course schedule on D2L*	**S. Ray**

* Note that a few lectures during the beginning of the course will be taught with Alternative delivery (AD) method as *synchronous* activities

Office hours: TBD

Course description

Data science enables one to bring structure to large quantities of data and make analysis possible. The purpose of the course is to introduce students to the fundamentals of data science and prepare them in dealing with the challenges of Big Data analytics. It covers advanced Python programming and Python libraries for data analysis. It presents data visualization techniques and statistical methods, as well as data exploration techniques such as data cleaning and munging, manipulating data, rescaling and dimensionality reduction. It includes an introduction to machine learning with linear regression, classification and clustering and presents special data analysis topics of time-series analysis. Also, it introduces data analysis approaches with relational databases and big data frameworks such as Dask.

Evaluation (tentative)

1.	Assignments:	20%
2.	Hands-on activity:	20%
3.	Final Exam:	40%
4.	Midterm Exam:	15%
5.	Class activities & participation:	5%

Textbooks

[T1] Python Data Science Handbook: Essential Tools for Working with Data. By: Jake VanderPlas. O'Reilly Media.

[T2] Think Statistics - Exploratory Data Analysis in Python. By: Allen B. Downey.

[T3] Python Data Analytics: Data Analysis and Science Using Pandas, matplotlib, and the Python Programming Language. By: Fabio Nelli. Publisher: Apress.

References (not textbook, but recommended reading)

[R1] Think Python 2nd Edition. By Allen B. Downey

Note: Some of these are available as course reserves at the Engineering & CS Library

Topics (Syllabus)

1. Introduction to Python

i) Basics

ii) Advanced concepts

iii) Python tools and libraries for data analysis, such as, IPython notebook, NumPy and SciPy

2. Data Wrangling and Exploration

i) Introduction to Pandas

ii) Working with data: Cleaning and Munging, Manipulating Data

iii) Transformation, Rescaling, Dimensionality Reduction

3. Data Visualization

i) Visualization fundamentals, Infographics, Interactive Visualization, Mapping

ii) matplotlib, Bar Charts, Line Charts, Scatterplots

4. Statistics

i) Statistics basics: Describing a set of data, Central tendencies, Outlier

ii) Probability basics, Pmf, Cdf, Pdf, Modeling distributions, Estimation

iii) Relationships between variables, Correlation and causation

iv) Hypothesis and Inference: Statistical Hypothesis Testing, Confidence Intervals

5. Machine Learning Introduction

i) Basics of machine learning

ii) Basic machine learning techniques:

Linear Regression

Classification

Clustering

6. Data Engineering: Data Manipulation at Scale

i) Accessing data from relational databases

ii) Scaling data analysis with Dask

Notes

- The ***hands-on***s will be conducted with Alternative delivery (AD) method as *synchronous* activities, where students are expected to be online and interact with the professor at scheduled times. This means that you should be prepared to be available and online during the hands-on times. Any conflict between this schedule and your own (e.g. due to timezones) will be up to you to resolve.

- Note that ***a few lectures*** during the beginning of the course will be taught with Alternative delivery (AD) method as *synchronous* activities. This is due to fact that some students are encountering travel delays due to the ongoing pandemic. Subsequent lectures will be taught in-person at HC10. Please check the course schedule for details, along with Teams meeting link.

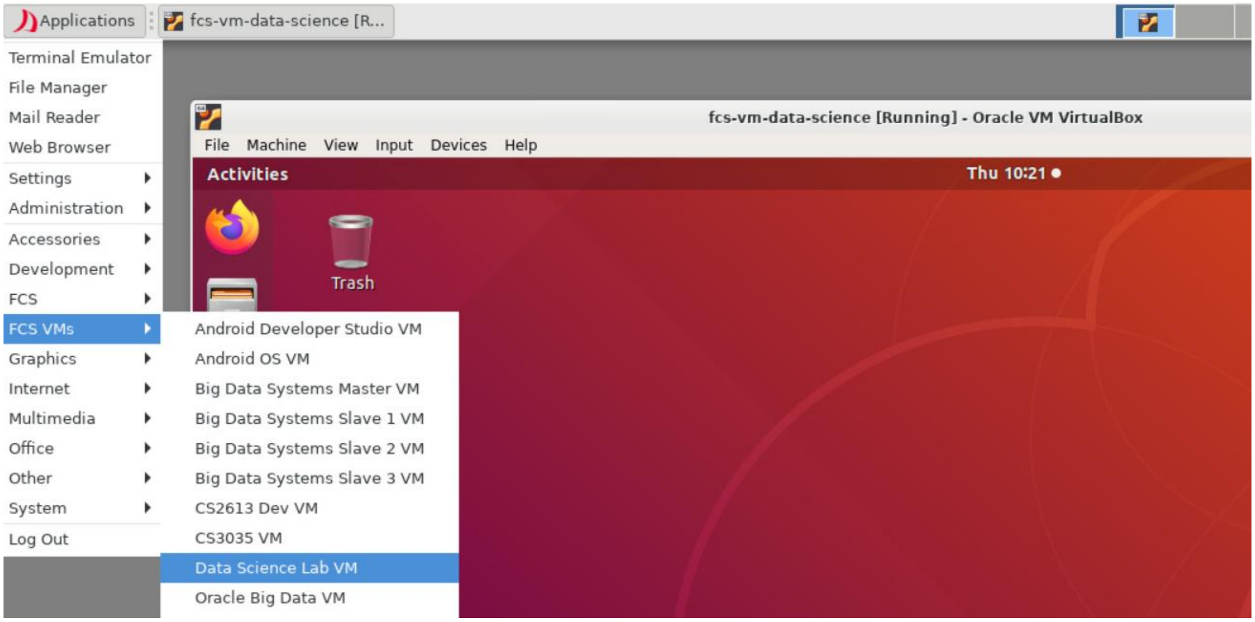
- Due to privacy concerns, I cannot guarantee availability of recordings or transcripts from synchronous activities.

- For the sessions delivered with Alternative delivery (AD) method, you will need an internet connection capable of watching streaming video, and participating in bidirectional audio calls. You are recommended to have a headset with built-in microphone to avoid noise. A webcam is optional, but can make group discussions more fun. We will use MS Teams for synchronous activities (labs and exams). The link to MS Teams for online meeting was provided above.

- For the hands-on sessions delivered with Alternative delivery (AD) method, you may need access to the *UNB VPN*, if you are not in campus. Note that this is a service provided by ITS, and you should contact them for help getting it working. Linux users may want to consider using openfortivpn to connect (although ITS will not support that client).

  After installing VPN, follow the instructions in the page below regarding how to connect to a lab machine remotely using SSH and VNC: https://www.cs.unb.ca/help/remote-lab-gui-access.shtml

  A VNC session is shown below:


Applications
fcs-vm-data-science [R...
Terminal Emulator
File Manager
Mail Reader
Web Browser
Settings
Administration
Accessories
Development
FCS
FCS VMs
Graphics
Internet
Multimedia
Office
Other
System
Log Out
fcs-vm-data-science [Running] - Oracle VM VirtualBox
File
Machine
View
Input
Devices
Help
Activities
Thu 10:21 ●
Trash
Android Developer Studio VM
Android OS VM
Big Data Systems Master VM
Big Data Systems Slave 1 VM
Big Data Systems Slave 2 VM
Big Data Systems Slave 3 VM
CS2613 Dev VM
CS3035 VM
Data Science Lab VM
Oracle Big Data VM