RIDER UNIVERSITY

College of Liberal Arts and Sciences

Department of Computer Science & Physics

CSC 410 Data Engineering

CRN:21912 , Section: I1, Spring 2020, Credits: 3.0

Building: SCI 220, Day: MW, Time: 02:50 pm – 04:20 pm

Instructor Information

- **Instructor:** Dr. Md Liakat Ali
- **Office:** Science Hall, Room 204 D
- **Office Hours:**

MON, TUE, THU 12-1:00 pm, or by appointment

E-mail: mdali@rider.edu

Course Description

- **Course Description:** This course serves as an introduction to the interdisciplinary and emerging fields of data engineering and data science. Students learn to combine tools and techniques from computer science, statistics, data visualization and the social sciences to solve problems using data. Central themes include: the data engineering and data science processes; tools for working with both big and small datasets, statistical modeling, and machine learning. Specific topics and tools include: data wrangling and munging, machine learning algorithms, statistical models, data visualization, data pipelines, ethics, Hadoop, Spark, R, Python, and MapReduce.

Course Description

- **Prerequisites and Restrictions:** CS 410 is a required course for CS majors. The prerequisites for this course are CS 230, Probability for Computer Science.

Text Book

- There is no required specific text book for the class, but it is highly recommended that you should have one of the following books:
  - Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, by EMC Education Services, Wiley and Sons. (2015).
  - Introducing Data Science: Big Data, Machine Learning, and More, using Python Tools, by Davy Cielen, Arno D. B. Meysman, and Mohamed Ali, Manning Publications. (2016). Available in Online: https://www.manning.com/books/introducing-data-science?query=Introducing%20Data%20Science

Course Management System

- Canvas will be used to provide and submit all homework assignments, classwork, project and grades.
- Make sure that you check your total score in Canvas regularly.
- Any questions regarding Canvas points, you have to ask/email me **before Final Exam**. No email will be answered after final exam.

Course objectives

- Upon completion of this course, the student should be able to:
  - Explain the difference between data engineering, data science and data analytics.
  - Describe the various types of data: structured, unstructured, natural language, machine-generated, graph-based or networked, audio, image, video, and streaming.
  - Recognize problems solvable with machine learning and data science algorithms.
  - Explain and use data visualization techniques.
  - Discuss about classification, clustering, text mining, and information extraction
  - Analyze the distributing data storage and processing with frameworks, Hadoop, MapReduce, and NoSQL ("non SQL" or "non-relational") databases and data modeling.
  - Discuss and use of Python and Python data engineering libraries, Scikit-learn and StatsModels

Requirements and expectations

- Canvas: Supplemental course site. Software: Latest version of R and RStudio Desktop: available at- https://cran.rstudio.com/, https://rstudio.com/products/rstudio/download/. Latest version of Python available at https://www.python.org/downloads/ Latest version of PyCharm Python IDE available at https://www.jetbrains.com/pycharm/
- scikit-learn Machine Learning in Python: https://scikit-learn.org/stable/
- Commitment to attend class and log-on periodically to Canvas to check for announcements, check calendar, post.
- During all classes you are expected to take substantive notes and to develop a list of terms/concepts as instructed by the instructor.
- Participate in all quizzes, tests, and in-class work. For all assignments, you are expected to submit within due date, no assignments will be accepted after the due date.

Classroom Decorum

- Be on time! Class will begin promptly at the scheduled time. Quizzes will usually be given at the start of the class so if you are late you may miss a quiz.
- Do your best to remain in the room during the period. Exiting and entering during the period breaks the concentration of your fellow students, and makes it hard for you to get the full value of the class.
- Turn off all cell phones, no earbuds, no texting, and anything else that would cause a distraction to yourself or others around you. If there's an emergency situation, you have to inform your instructor.
- Participate in class discussions; ask questions, no eating or drinking in class. Be aware of body language; it can speak volumes! No sleeping in class
- Students are permitted to use computers/laptops during class for note-taking and other class-related work only. Those using computers/laptops during class for purposes not related to the class (like e-mailing, instant messaging, game playing or internet surfing) will be asked to leave the classroom for the remainder of the class period.

Image Reference: https://www.shutterstock.com

Graded Course Activities

- Visit Canvas for details about assignment, classwork, project, and exams.
- All assignments and classwork for this course will be submitted electronically through Canvas unless otherwise instructed.
- Points you receive for graded activities will be posted to the Canvas.

Course Format

- Two lectures / week.
- Class format: approximately 45-60 minutes lecture and 30-45 minutes class exercise/class work
- Participation in the class exercise and class work session is mandatory.
- All tests will be **closed book**. You **can not** use textbook/notes. You are not allowed to copy other's work.
- There will be
  - Assignments
  - Mid term
  - Final exam
  - Class Work
  - Group Project

Grading Criteria

Grading Criteria	
Points	Description
20	Group Project
18	Assignments
18	Mid Term
30	Final Exam
09	Classwork
05	Class Participation
100	Total Points

Letter Grade	Percentage	Quality Points
A	93.0-100%	4.0
A-	90.0-92.9%	3.7
B+	87.0-89.0%	3.3
B	83.0-86.9%	3
B-	80.0-82.9%	2.7
C+	77.0-79.9%	2.3
C	73.0-76.9%	2
C-	70.0-72.9%	1.7
D	60.0-69.9%	1
F	0-59.9%	0

Participation

- ATTENDANCE IS MANDATORY. Class participation is worth 05% of your final grade.
- Class participation is critical in this course. Much of the work is additive/cumulative. Consequently, frequent absences will adversely affect your grade.
- You will also lose credits for classwork points if you miss classes. If you have perfect attendance, your Professor will consider boosting your grade if it is borderline.
- If you miss a class, it will be your responsibility to determine what you missed and to make up the work.

Assignments

- There are six assignments in this course.
- Each assignment should be submitted on the Canvas website assignment page.
- Completed assignments must be submitted electronically via canvas no later than 11:59 PM on the due date listed in the syllabus.
- No assignment will be accepted after due date.
- **Do not email me your assignment, as it will not be accepted**.

Classwork and Project

- There will be six classwork in this course.
- If you miss a class, you will miss class work points. There will be no makeup for classwork.
- Proofs are needed for exceptions or true emergencies.
- There is a team project with maximum two persons in each team. The project should involve analytics on real world data of significant size and must be approved by the instructor.
- More about project coming soon!

Exam

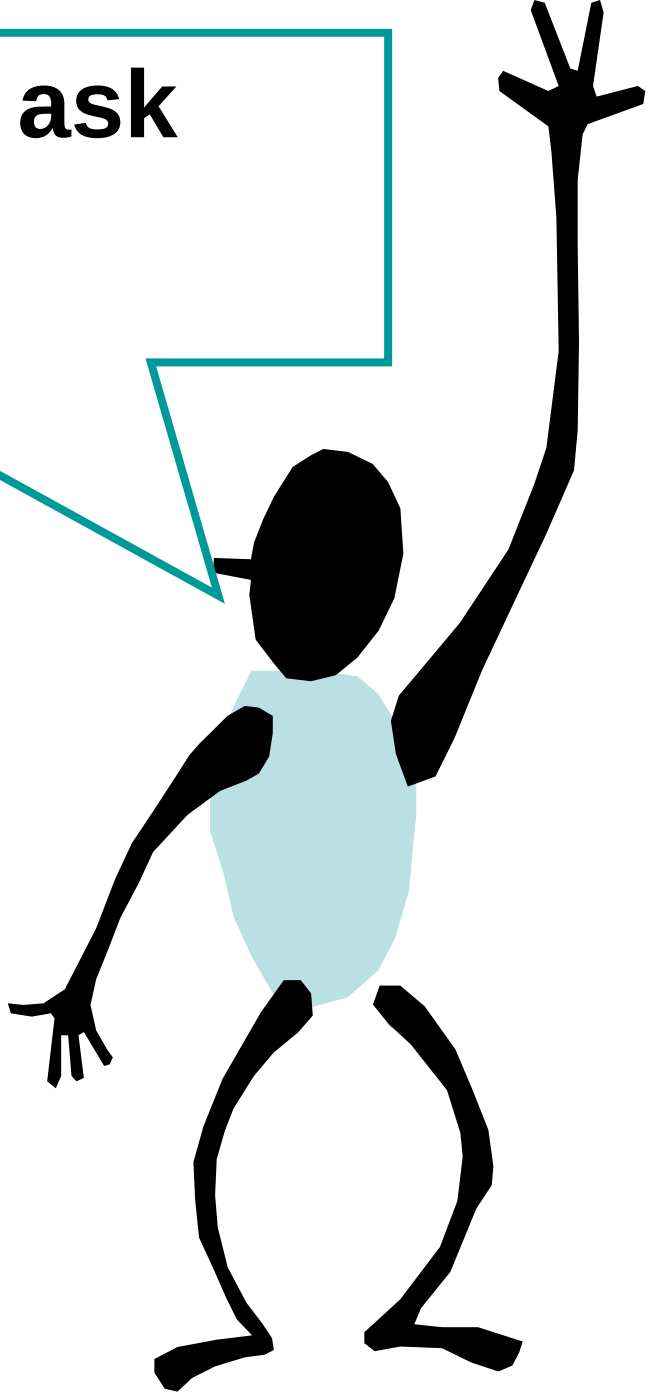
- One midterm and one final exam will be given in the semester session. All tests are in class.
- **All exams will be closed-book. You can not use textbook/ notes**. You are not allowed to copy others work.
- The exams will test assigned readings and material discussed in class. The final exam will **not** be comprehensive in nature and will cover the material after the second test.
- Exams cannot be made up, cannot be taken early, and must be taken in class at the scheduled time. Proofs are needed for exceptions or true emergencies

Cheating

- Do not show your solutions to others!
- You are not allowed to read, copy, or rewrite the solutions written by others .
  - Copying materials from websites, books or any other sources is considered equivalent to copying from another student.
- If two people are caught sharing solutions, then both the copier and copies will be held equally responsible, which will result in zero point in homework.
- Cheating on an exam will result in failing the course.

Please feel free to ask questions

- Please talk with your instructor in the class, not with your friends!
- Help me know what people are not understanding. We do have a lot of material, It's your job to slow me down
- I usually add extra points in the final grade for those students who regularly participate in the class.



Feedbacks

- I appreciate your feedbacks. Your feedbacks help me know how I can better deliver my lectures, which will ultimately benefit you.
- If you have any concern regarding class, please email me/ discuss with me in the class or in my office/ drop your notes in my office
- I will also take a survey in the middle of the semester to get your feedback.

Content Calendar

Week	Date	Topic	Class Work	Assignment
1	01/27/2020	Introduction		
	01/29/2020	Introduction to Big Data Analytics		1 Due: 02/09/20
2	02/03/2020	Data Analytics Lifecycle		
	02/05/2020			
3	02/10/2020	Review of Basic Data Analytic Methods Using R		
	02/12/2020	**Project proposal presentation**		
4	02/17/2020	Advanced Analytical Theory and Methods: Clustering		2 Due: 02/25/20
	02/19/2020			
5	02/24/2020		1	
	02/26/2020	Advanced Analytical Theory and Methods: Association Rules		3 Due: 03/08/20
6	03/02/2020	Mid Term review		
	03/04/2020		2	
7	03/09/2020	**Mid Term Exam**		
	03/11/2020	**Mid Project Presentation**		
8	03/16/2020	No Class (Spring Recess)		
	03/18/2020			
9	03/23/2020	Advanced Analytical Theory and Methods: Regression		4 Due: 04/01/20
	03/25/2020			
10	03/30/2020		3	
	04/01/2020	Advanced Analytical Theory and Methods: Classification (Decision Trees)		5 Due: 04/12/20
11	04/06/2020		4	
	04/08/2020	Advanced Analytical Theory and Methods: Classification (Naïve Bayes)		6 Due: 04/19/20
12	04/13/2020		5	
	04/15/2020	Advanced Analytical Theory and Methods: Text Analysis		
13	04/20/2020		6	
	04/22/2020	Advanced Analytics-Technology and Tools: MapReduce and Hadoop/		
14	04/27/2020	**Students' Presentation**		
	04/29/2020	**Final Exam**		
15	**Final Project Presentation: May 7, 2020 Time: 3:30-5:30 PM**			

Group Project

- Purpose:
  - This is a team project with maximum two students in each team.
  - The project should involve analytics on real-world data of significant size and must be approved by the instructor.
  - The goal is to prepare you to apply machine learning algorithms/models to real-world tasks and to prepare you qualified for AI or machine learning research and higher study.
  - It is expected that each group member contributes equally in the project.
  - Only one group member needs to submit all assignments with tagging other students' name. Please do not submit separately.
  - Each group member should participate in the presentation to receive the presentation points.

Group Project

- There are few key deliverables which are described below:
  1. **Task 01- Project Proposal:**
     - Proposal submission Due- 02/16/2020
     - Proposal presentation: 02/12/2020
     - Proposal submission- 3 points, proposal presentation- 1 point, total 4 points.
  2. **Task 02- Mid Project Submission:**
     - Mid project submission Due- 03/16/2020
     - Mid project presentation: 03/11/2020
     - Mid project submission- 4 points, presentation- 1 point, total 5 points.

Group Project

- Key deliverables (Cont'd.)

  3. Task 03-Advanced Analytics-Technology and Tools: Presentation

  – Presentation date: 04/22/2020 or 04/27/2020

  – Advanced analytics-technology and tools presentation- 3 points.

  4. Task 04-Final Project Submission:

  – Final project submission Due- 05/06/2020

  – Final project presentation: 05/07/2020

  – Final project submission- 6 points, presentation- 2 point, total 8 points.

See Project Guidelines for more details!

Data Analyst vs Data Engineer vs Data Scientist

Who is a Data Analyst, Data Engineer and Data Scientist?

Data Analyst	Data Engineer	Data Scientist
Data Analyst analyzes numeric data and uses it to help companies make better decisions.	Data Engineer involves in preparing data. They develop, constructs, tests & maintain complete architecture.	A data scientist analyzes and interpret complex data. They are data wranglers who organize (big) data.

Data Analyst vs Data Engineer vs Data Scientist: Salary

Data Analyst	Data Engineer	Data Scientist
$59000 /year	$90,8390 /year	$91,470 /year

https://www.edureka.co/

Data Analyst Vs Data Engineer Vs Data Scientist – Salary Differences

- On average, a **Data Analyst** earns an annual salary of **$67,377**
- A **Data Engineer** earns **$116,591 per annum**
- And a **Data Scientist**, on average, makes **$117,345 in a year**

https://data-flair.training/blogs/data-scientist-vs-data-engineer-vs-data-analyst/

Reference- https://www.edureka.co/ accessed on 1/20/2020

Data Analyst vs Data Engineer vs Data Scientist

Data Analyst	Data Engineer	Data Scientist
Data Warehousing	Data Warehousing & ETL	Statistical & Analytical skills
Adobe & Google Analytics	Advanced programming knowledge	Data Mining
Programming knowledge	Hadoop-based Analytics	Machine Learning & Deep learning principles
Scripting & Statistical skills	In-depth knowledge of SQL/ database	In-depth programming knowledge (SAS/R/ Python coding)
Reporting & data visualization	Data architecture & pipelining	Hadoop-based analytics
SQL/ database knowledge	Machine learning concept knowledge	Data optimization
Spread-Sheet knowledge	Scripting, reporting & data visualization	Decision making and soft skills

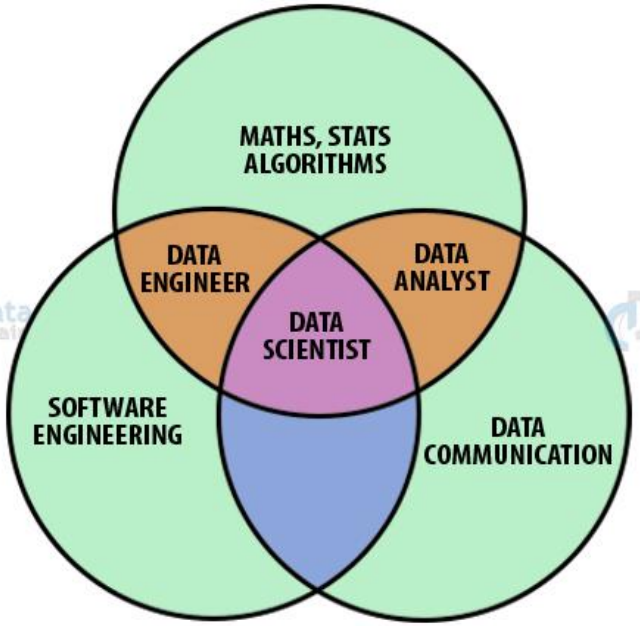
Data Analyst vs Data Engineer vs Data Scientist Skill Sets

Reference- https://www.edureka.co/ accessed on 1/20/2020

Data Engineer vs Data Scientist

- A **data analyst** is responsible for taking actionable that affect the current scope of the company. A **data engineer** is responsible for developing a platform that data analysts and data scientists work on. And, a **data scientist** is responsible for unearthing future insights from existing data and helping companies to make data-driven decisions.

- A data engineer is responsible for the development and maintenance of data pipelines. A data scientist uses dynamic techniques like Machine Learning to gain insights about the future.



https://data-flair.training/blogs/data-scientist-vs-data-engineer-vs-data-analyst/

Focus of Course

- Focus on quantitative disciplines – e.g., math, statistics, machine learning
- Provide overview of Big Data analytics
- In-depth study of a several key algorithms