



University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science / Data Science

Department: **Electrical & Computer Engineering and Computer Science**

COURSE:	DSCI 6007 – Distributed & Scalable Data Engineering	
MIDTERM PROJECT	Due date: October 17, 2022	
PROJECT TITLE:	<p>1. Form your Project Title based on your problem or challenge that you will work on.</p> <p>2. Midterm Project will be the start of your Final Project</p> <p>//example: Analyze real-time Twitter Sentiment</p>	
PROJECT EXAMPLE:	<p>3. Analyze real-time Twitter Sentiment</p> <p>4. Bank Model-CRIPSP-DM</p> <p>//Links provided</p>	
SEMESTER:		
MIDTERM INFORMATION	Submit Project Title:	10/10/2022
	Submit project:	10/19/2022 (9:00 am)
	Present project:	10/19/2022
Group Names:	Name 1:	
	Name 2:	
	Name 3:	
	Name 4:	

Midterm Project- Distributed & Scalable Data Engineering

Midterm -project Guidelines

Getting ready for your midterm-project!

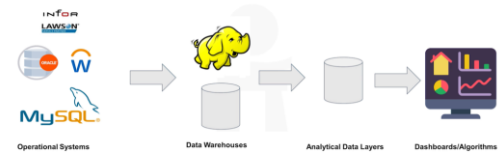
1. Team Leaders: Get the contact information (school email) for each team member and meet to devise a strategic plan on completing your Midterm Project. There is a large variety of people in your groups, with different fields of expertise available: from Mathematicians to Computer Scientists to industry experts [and others, I'm sure!], make sure to discuss each team members strongest skills and decide what part of the project each one will complete!
2. Coming up with the topic: Brainstorm and research on the topic of your midterm-project. Use the **Technical Report Template** given to you in Midterm Project Folder (posted on canvas-under Project Module)

The outline of your project description should be as follows:

Midterm Project (due **October 19 at 9:00am**)

1. **Problem Forming**: Title of the project.
2. **Problem Challenge| Industry**: Business Understanding.
 - o A good project should include an interesting challenge that you would like to work on. So, for example, a project that begins with a single clean data set and a clear task (e.g., predict a given response), isn't a good choice for your project.
3. **Data Set**: Data understanding- sources, cleaning, wrangling, management
4. **Pitch deck** (template attached): summary of your project, challenge, data, model, tools that you will use to answer the problem (**submit as part of Midterm Project- due date **October 19 at 9:00am****)
Pitch deck video (3 minutes): summary of your project, challenge, data, model, tools that you will use to answer the problem (**submit as part of Midterm Project- due date **October 19 at 9:00am****)

Pitch deck should include	
Your project title	Give your project a title
Your team	<p>List all your team members and shortly the role that each team member is going to take in this project, ex:</p> <ul style="list-style-type: none"> • Team leader • Data Scientist • Data Modeling • App development-environment for model deployment • Storytelling and data visualization • Etc.
Your challenge	It can be one challenge, or your project is tackling 3 small challenges of a big challenge
Your solution (Midterm Project)	<p>Here you just describe what you are going to build to find a solution to the problem.</p> <p>At this point do not worry for the technological solution, just list:</p> <ul style="list-style-type: none"> • Data that you will need • Where you will get • An overview of the data science pipeline for your solution that includes CRISP methodology standard <p>///CRISP methodology</p> <p>Document your CRISP-DM methodology</p> <ul style="list-style-type: none"> • Business Understanding • Data Understanding • Data Preparation • Modeling • Evaluation • Deployment <p>Note 1: Give a graphical view of your ETL (all tools, technologies, platforms that you will use)</p>

	 <p>Note 2: Give a graphical view of your CRISP-DM solution (Power point posted on MidtermProject folder)</p> <p>//Midterm Project ends here</p>
Your solution (Final Project)	///Will discuss later on October 20th

///Final Project starts below

5. **Data analysis:** Data Preparation-statistics, database schemas
6. **Data Modeling:** create the model, machine learning
7. **Model evaluation:** choose the best model for your problem. Adopt agile deployment.
8. **Model Deployment:** create an environment to deploy your model
9. **Data Visualization:** Communication of results: summarization & visualization
10. **Operationalization:** creating added value, end-user point of view

Your Technical Report includes:

- Title of the project.
- Document your CRISP-DM methodology
 - **Business Understanding**
 - **Data Understanding**
 - **Data Preparation**
 - **Modeling**
 - **Evaluation**
 - **Deployment**

Step1: Problem Forming

This is the most challenging part: coming up with the problems yourselves. But worry not, we got your back!

Business Question Understanding Industries.

How to take the available data and pair that with the right mathematical model to formulate a solution

Recommended Courses:

- <https://developers.google.com/machine-learning/problem-framing/problem>
- <https://www.coursera.org/learn/advanced-models-for-decision-making?specialization=analytics-for-decision-making>
- <https://www.linkedin.com/learning/paths/develop-critical-thinking-decision-making-and-problem-solving-skills?u=2359714>

Step 2: Problem Understanding

Example projects

Example twitter-project topics, to get an idea of how to form the topic yourselves.

1. [Sentiment analysis - Wikipedia](#)
2. <https://online.datasciencedojo.com/course/Sentiment-Pipeline-for-Live-Tweets#ccnTab-2>
3. <https://code.datasciencedojo.com/datasciencedojo/tutorials/tree/master/Building%20Real-Time%20Sentiment%20Pipeline%20for%20Live%20Tweets>
4. [Building Real-Time Sentiment Pipeline for Live Tweets · master · Data Science Dojo / tutorials · Code](#)
5. [Twitter Sentiment Visualization \(ncsu.edu\)](#)
6. [Tweet Sentiment Visualization App \(ncsu.edu\)](#)

Step 3: Data set

Example data sets:

<https://careerfoundry.com/en/blog/data-analytics/where-to-find-free-datasets/>

Step 4: Pitch Deck Session

Use the template posted on canvas

Pitch sessions: Selling your idea

On **October 19th**, we will host pitching sessions. All groups should prepare a 3-minute (no more - no less than that) selling pitch of their application. The pitch is directed to their predefined target group, and thus should not be technical. Stress the selling points of your application: why is this useful? why is it interesting to this audience? etc. You can accompany the pitch with slides or a demo of your application. Focus on selling your idea, it should be pitch-like.

[Here you can find a few useful tips on coming up with a great pitch](#)

[Kevin Hale - How to Pitch Your Startup - YouTube](#)

Midterm Project Due Date

- Each team member submits individually (the same team files)
- Submit your project by **October 19, 2022 (morning 9:00am)**

MidTerm Project Final deliverables

The project should be submitted by everyone (all members submit the same project) on canvas on the corresponding submission section. Final deliverables should include:

1. Your **GitHub account** Create a team github account and start using that for Midterm Project and Final Project. All the Midterm Project files should be stored on your team's github.
2. **Your Pitch Deck** (video link, website, etc) (use the template posted on canvas)
3. **Your Pitch Deck** (document, slides, etc.)

Submission deadlines

Deadline for group projects is **October 19, 2022**

Assessment Criteria

Some key-points to be taken into consideration for your work.

- **Presentation and pitch:** This is the overall impression you get from the presentation of the problem and its solution. How clear is the problem description? Are the results communicated well? How convincing was the pitch? [depending on the topic's purpose]
- **Value:** What is the value of the final deliverable?
- **Effort:** Do you think there was a sufficient amount of work put into the project?
- **Idea:** How original is the topic idea in your opinion? How useful/innovative is it? Could it have more applications than the one described in this case?
- **Topic scope:** How suitable was this project for this course? Are all stages of the data life cycle considered?
- **Actionable:** How actionable is the outcome? Could it be used in real life applications?
- **Deliverables:** Your Project files.

How understandable is it to your target group? How understandable is it to everyone? Could it stand alone without your pitch? Needs to describe what you did and how you did it in a technical **manner**/and non-technical manners (easy to understand)

- **Project analysis:** Data wrangling, data analysis, communication of results will be taken into account, based on each project's topic. For example, the visualizations of a project: How well do the visualizations communicate the point made? Could correlations between other variables have clearer results etc

Project Rubrics

Category I Requirements (Knowledge)	Poor 0.5 pts	Fair 3.5 pts	Good 11 pts	
Topic scope	Not submitted	Submitted but not suitable for the project	Submitted and suitable for the project	1 pts
Pitch Deck	Not included	Included but not complete	Completed	7pts
Project Pitch Video	Not included	Included but not complete	Completed	3 pts
Github-Communication of results	Not included	Included but not complete	Completed	0.5 pts

Category II (Presentation)	Poor 0 pts	Fair 1 pts	Good 2 pts	
Project Pitch	Not submitted	Submitted but isn't related to the Project	Submitted and presents the project	0.5
Project Presentation	Not submitted	Submitted but isn't completed	Submitted and presents the project	0.5
Category III (Originality)	Poor 0 pts	Fair 0.5 pts	Good 2 pts	
Idea	Not original/or innovative	Original/not innovative	Original/and innovative	0.5
Value/effort	Results are not useful	useful	Very useful.	1
Actionable	Not actionable	Actionable	Very actionable	1
Total Points				15 pts

Midterm Project Resources

Recommended Courses:

1. <https://developers.google.com/machine-learning/problem-framing/problem>
2. <https://www.coursera.org/learn/advanced-models-for-decision-making?specialization=analytics-for-decision-making>
3. <https://www.linkedin.com/learning/paths/develop-critical-thinking-decision-making-and-problem-solving-skills?u=2359714>
4. <https://online.datasciencedojo.com/course/Sentiment-Pipeline-for-Live-Tweets#ccnTab-2>
5. <https://code.datasciencedojo.com/datasciencedojo/tutorials/tree/master/Building%20Real-Time%20Sentiment%20Pipeline%20for%20Live%20Tweets>
6. [Building Real-Time Sentiment Pipeline for Live Tweets · master · Data Science Dojo / tutorials · Code](#)
7. [Twitter Sentiment Visualization \(ncsu.edu\)](#)
8. [Tweet Sentiment Visualization App \(ncsu.edu\)](#)
9. [2015/02-DataScrapingQuizzes.ipynb at master · cs109/2015 · GitHub](#)
10. [Justin Blinder](#)
11. by Healey and Ramaswamy

- a. http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/
- 12. [GitHub - bear/python-twitter: A Python wrapper around the Twitter API.](#)
- 13. [All Tutorials \(datasciencedojo.com\)](#)
- 14. [Course: What is a Data Engineer \(datasciencedojo.com\)](#)

Four useful references in scientific writing

- 15. Marie Davidian:
http://www4.stat.ncsu.edu/~davidian/st810a/written_handout.pdf
- Rod Little: <http://sitemaker.umich.edu/rlittle/files/styletips.pdf>
- Paul Halmos: <http://www.matem.unam.mx/ernesto/LIBROS/Halmos-How-To-Write%20Mathematics.pdf>
- George Gopen and Judith Swan:
<http://engineering.missouri.edu/civil/files/science-of-writing.pdf>

LINKs

[Build a True Data Lake with a Cloud Data Warehouse - Talend | Talend](#)

<https://www.upgrad.com/blog/data-science-vs-data-engineering-difference-between-data-science-data-engineering/>

<https://blog.panoply.io/what-is-the-difference-between-a-data-engineer-and-a-data-scientist>

<https://www.mastersindatascience.org/careers/data-engineer/>

<https://enterprise.2u.com/data-careerkickstarter-mini/>

[free student/educator licenses for Alteryx](#)

https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[What is Interactive Query in Azure HDInsight? | Microsoft Docs](#)

[Quickstart: Create Spark cluster in HDInsight using Azure portal | Microsoft Docs](#)

[Azure Quickstart Templates \(microsoft.com\)](#)