# DATA 606 | Capstone in Data Science — Thursdays, Spring 2024

Instructor: Dr. Chaojie (Jay) Wang — E-mail: jaywang@umbc.edu
WebEx: https://umbc.webex.com/meet/jaywang — Office Hour: By appointment
Day/Time: Thursday, 7:10pm – 9:40pm
Classroom: Fine Arts 006 First Session, thereafter online using WebEx
Course Info: https://github.com/wcj365/UMBC-Data-Science-Capstone

**IMPORTANT DATES**

| *Phase* | *Classes/Dates* |
|---|---|
| PHASE I: Preliminary EDA & Proposal | Feb 1, Feb 8, Feb 15 |
| PHASE II: Data Preparation & EDA | Feb 22, Feb 29, Mar 7 |
| PHASE III: Model Training & Deployment | Mar 14, (Mar 21, No Class), Mar 28, Apr 4 |
| Final Presentation | Apr 11, Apr 18, Apr 25, Apr 29, May 2 |
| Final Deliverables Due: | Sunday, May 19 |

1. Students are required to use GitHub and push code to GitHub timely. Use your personal email for GitHub account. Don't use UMBC email for GitHub account. File naming conventions in GitHub:
    a. Draft proposal = **draft_proposal.md**
    b. Final proposal = **final_proposal.md**
    c. Project presentation = **capstone.pptx**
    d. Project report = **report.md**
2. Students are required to attend all classes.
3. Students are required to share their project proposal and work-in-progress in class.
4. Students are required to present their completed final projects in class:
    - Use PPT for presentation.
    - Walk through the project artifacts (source code, markdown, Notebooks, etc.)
    - Demonstrate the working software (Python, Jupyter notebooks, web apps, etc.)
5. Final Deliverable:
    a. Project Report (report.md). Report include link to PPT, and YouTube video
    b. Project presentation (capstone.pptx)
    c. Project video: Upload to YouTube
    d. **Due Sunday, May 19.**

**Student Projects from Previous Semesters**

- https://github.com/DATA-606-FALL-2022
- https://github.com/DATA-606-SPRING-2022
- https://github.com/DATA606-SUMMER-2022

**Course Description**

This is a semi-independent online course that provides the graduate Data Science students an opportunity to apply the knowledge, skills and tools they have learned to a real-world data science project. Even though real data sets are recommended, students can also use synthetic data sets to experience the entire lifecycle of a data science project. Typically, this cycle includes collecting, cleansing, and transforming the data, choosing best methods to solve the problem or prove the

hypothesis, implementation, and quantifying the robustness and accuracy of their model. The project can be conducted with industry, government, and academic partners, who can provide a data set.

**Prerequisites:** Completion of DATA 601, 602, 603, and 604.

**Course Learning Objectives**

Demonstrate

- Proficiency in managing a full life-cycle data science project
- Effective communication and presentation skills
- Competence in preparing insightful visualizations
- Competence in writing an article

**Course Materials**

Required Reading: None

Suggested Reading:

- Ryan Hodson, Ry's Git Tutorial, 2014, ASIN: B00QFIA5OC.
- Robert de Graaf, Managing Your Data Science Projects, Springer, 2019, ISBN 9781484249079.
- Cole Nussbaumer Knaflic, Storytelling with Data: A Data Visualization Guide for Business Professionals, Wiley, 2008, ISBN 9781119002253.
- Brian Godsey, Think Like a Data Scientist, Manning, 2017, ISBN 9781633430273.
- Kenneth S. Rubin, Essential Scrum: A Practical Guide to the Most Popular Agile Process, Addison-Wesley, 2012, ISBN-13: 978-0137043293.
- Ralph Hughes, Agile Data Warehousing Project Management: Business Intelligence Systems Using Scrum, Morgan Kaufmann, 2012, ISBN: 978-0123964632.

**Course Format**

- This is an online course, where all the material prepared by the instructor and students will be shared with each other electronically.
- There are three main phases (steps):
  - Project Proposal & Approval
  - Data Preparation & EDA
  - Model Training & Deployment
- Instructor and students will meet weekly to discuss project progress.
- Student are expected to update their profile/project page (P3) on the Data 606 course GitHub on a continuous basis.

**Implementation**

- Student will upload their dataset and codes to their GitHub repository. If the file size is larger than 25 MB but less than 1 GB, students are recommended to split the dataset into smaller sets (~23 MB) after data cleansing. If the file size is larger than 1 GB, then students are recommended to provide a short script in their repository to download the dataset to a local drive.

- Students will prepare and record presentations, upload their records to YouTube, and share their videos via their GitHub. These presentations can be viewed as a rehearsal for their main presentation which will be made to their classmates and instructor.

**PHASE I:** PROJECT PROPOSAL & PLANNING

In the first few weeks of the semester, students are required to determine a data science project idea, which will allow them to

- Carry out a complete data science project from end to end.
- Demonstrate ability to carry out a literature search and summarize the state of the art.
- Demonstrate ability to translate the project objects into a realistic work plan.
- Demonstrate ability to design and implement required software using programming languages and/or platforms.
- Demonstrate proficiency in preparation and walk through of a presentation.

Students are free to choose the subject of their capstone project, but **they need to get instructor's approval** before they start working on their project. Deadline is provided in the beginning of this syllabus.

Some crucial points:

1. Students have to have a research question which might be answered with data science. Exploratory data analyses are not accepted. It must include model training, evaluation, tuning, and deployment.
2. Students have to determine a dataset which is relevant to the subject and large enough.
3. Datasets used in previous data science competitions (such as Kaggle) are not accepted.
4. Students have to get the instructor's approval before they start working on the data.

In previous semesters, our students worked on datasets from various fields such as healthcare, finance, medical imaging, meteorology, geophysics, payment, sports and many others. Students are encouraged to visit https://sites.google.com/umbc.edu/data606 to learn some of the projects carried out last year.

Other details:

- Even though the individual capstone projects are encouraged, students might work in small teams depending on the project's complexity of the problem. As the project and problem statements warrant, students may be permitted to organize into teams of two to three participants.
- Projects focusing on global challenges in areas such as climate, energy, natural resources, transportation, and healthcare are highly encouraged.
- Students are encouraged to read as much as possible during this stage to gain a substantial knowledge on the subject they choose.
- Once the student chooses a project pitch and dataset, s/he should contact the instructor (**jaywang@umbc.edu**) as soon as possible for approval. The email should include a brief description of the idea (what is it about, which dataset will be used, how large dataset is, which methods are likely to be used, etc.)
- Students, who cannot find a project idea, should contact the instructor as soon as possible. Data Science faculty might help with a list of possible projects and students can align themselves with problem statements corresponding to their individual interests. Students are encouraged to mention their strengths in data science and specific area of their interest (i.e. health data, HR data, real estate data, e-shopping data, etc.).

**PHASE II:** Data Preparation and EDA
As soon as the project is chosen, students will

- make a brief literature/industry research to determine similar projects to their project,
- learn from those studies' outcomes and differentiate their project than others,
- plan and document the details of the planned implementation
- get familiar with their datasets and carry out transformations and cleansing, if necessary, and
- carry out some basic exploratory data analysis on their data sets.
- Communicate/document the EDA results with visualizations

**PHASE III:** Model Training and Deployment
In the second part, students will

- complete their data exploration stage (the dataset should be completely ready after appropriate cleansing and transformations; student is expected to be familiar with all the major patterns and trends in dataset)
- construct their model (i.e. if it is a regression problem, then students should have their codes ready that are compatible with the dataset; if it is a neural network implementation, then students should complete at least one successful training, etc.)
- come up with concrete outcomes (either their initial hypothesis work and why or "these are the outcomes with this much prediction according to method-x, etc.),
- discuss the performance of their model (in terms of accuracy and/or confidence levels)
- select the best model among alternatives and further tuning the selected model
- deployment of the model for predictive analysis.
- students are encouraged to develop prediction web apps or dashboards using Streamlit or Dash
- elaborate what could be done differently or what can be done next

At the end of this final phase, each student will present their completed project in class.

**Course Grading Criteria**

| | |
|---|---|
| Success of the project | %40 |
| Clarity, quality, and effectiveness of the presentations | %30 |
| Content, format, and richness of the GitHub | %20 |
| Punctuality | %10 |

Final grade will be computed as follows: 85% - 100% A, 70% - 84% B, 55% - 69% C, 40% - 54% D, <40% F.

Students who cheat will get an F (no excuse is acceptable).

**About Presentations**

- In the first slide, students should have the title of project, their name, and time (i.e. Summer 2022).
- In the first presentation, students are encouraged to introduce themselves. However, students with privacy concerns do not have to share their names or faces.
- This web-page has some (slightly outdated) information about how to record presentations https://www.csee.umbc.edu/~simsek/DATA606/videos.html

**Resources**

UMBC provides a range of writing assistance, which can be found in the following:

- The Writing Center https://lrc.umbc.edu/tutor/writing-center/
- Research Guides & Tutorials https://lib.guides.umbc.edu/tutorial

Please visit https://covid19.umbc.edu/ for UMBC Policies and Resources during COVID-19.

**Course Policies**

- Failure to follow guidelines for each phase, including the required format, style, length, submission, etc., may result in at least one-letter-grade reduction.
- Incomplete assignments will not be accepted unless an extension has been agreed to in advance. Emergency situations will be handled on a case by case basis with appropriate justification and/or documentation.
- Incomplete grades will not be entertained unless extenuating circumstances warrant and your request is made before the last week of class.