Final Project: VAST Challenge (EDA Condition)


Deadlines (all submissions are due before class time)**:**

- Week 1: ~~3/9 by 11:59am~~ 3/10 by 5:59pm (progress report): **10%** of the final grade
- Week 2: 3/16 by 11:59am (progress report): **10%** of the final grade
- Week 3: 3/23 by 11:59am (progress report): **10%** of the final grade
- Week 4: 3/30:
  - Check-In #1: 3/30 by 11:59pm (in-class check-in): **10%** of the final grade
- Weeks 5 and 6:
  - Check-In #2: 4/13 by 11:59am (in-class check-in): **20%** of the final grade
- Final Project Demo: 5/8 12pm-2pm: **40%** of the final grade
  - All source code and final project book submitted by 11:59am on 5/8

Description: The goal of your final project will be to apply your knowledge about Visual Analytics to solving the VAST challenge. VAST challenge datasets [1] are modeling real-world phenomena in interesting and mostly multivariate datasets.

The Challenge: You will be working with the 2017 VAST challenge dataset! The description of the challenge is available here: https://www.cs.tufts.edu/comp/178/eda. Your team will tackle MC1 for your final project. Note that this mini-challenge has a "ground-truth" answer. Your job is to develop a complete visual analytics system to analyze the data.

Requirements: Your job is to develop a complete visual analytics system to analyze the data. The system must have **at least two coordinated visualizations**. There needs to be a strong coding/programming component to your system (e.g. submitting a Tableau Dashboard does not meet this requirement).

You are allowed to use existing libraries and web frameworks, but they cannot be the primary contribution of your final system. For example, using a charting library like Plotly is discouraged – if you are using Plotly to create visualizations for your final system, you will need to demonstrate significant additional contributions elsewhere (e.g., in advanced data analysis techniques). However, if you want to use a library to create a custom slider, that is perfectly fine.

If you have questions about whether a particular library or framework is allowed, please contact us.

You must create a process book and document all of your work (explained later).

Week 1 (3/2 – 3/9): Begin process book and familiarize team with all phase requirements

Tasks:

1) Familiarize yourselves with the given VAST challenge (storyline, dataset description).
2) Download the dataset of MC1.
3) Start a ***process book*** to structure and document your project progress throughout the project.
    a. The goal of this process book is a fully archivable and transparent documentation of the project. You will describe all facets of a successful VA project: the project structure and goals, the analysis approach, the intended workflow design and the visualizations that you envision and later implement.
    b. A complete and thorough process book will allow us to grade your project without having seen the VA tool (your final project system)!
    c. Note that this process book is a "living document." This means that your final system does not have to adhere perfectly to these initial ideas. Nonetheless, you should document any changes to your approach and solution as they occur.
4) Complete the Qualtrics survey
    - We're using Qualtrics to document your process
    - https://tufts.qualtrics.com/jfe/form/SV_0dK0YtKVQI17kgK

Deliverables (due 3/9):

1) Completed the Qualtrics Survey (each team member needs to complete this)
2) Start your process book (see below for the required information)
3) Submit the process book

Required format for the process book:

Your process book needs to include:

a. **Team Name** and the names of your team members
b. A **general description of your understanding of the challenge**
c. A **description of the data (at least 2 paragraphs)**
d. What are the three most important analysis questions that you think will address the challenge?

Week 2 (3/9 – 3/16): Explore and Describe Data and Formulate Analysis Questions

Tasks:

1) You should start to explore with the data by using existing tools (e.g. Tableau, Python, etc.) to explore the data to understand its characteristics.
2) As you do so, you should document any interesting findings and characteristics about the data.
3) You should document these findings as possible analysis questions that you will pursue in your final project.

Deliverables (due 3/16):

1) Update process book with the result of your data exploration
2) Submit the updated process book

Required addition to the process book:

You need to add the following to your process book:

a. Record of your initial exploration of the data (e.g., screenshots)
b. Document any interesting findings you encountered when exploring the data
c. Describe why these findings are interesting, formulate them as possible analysis questions, and report why they might be relevant to solving the challenge.

Week 3 (3/16 – 3/23): Define Goals, Tasks and Project Timeline

Tasks:

1) Based on the findings last week, you will begin to formulate the project plan. This means that you will develop a project timeline and milestones
2) The milestones should be specific and should include tasks (what will be done) and goals (what will be the outcome).

Deliverables (due 3/23):

1) Update process book with the timeline and milestones.
2) Submit the updated process book

Required addition to the process book:

You need to add the following to your process book:

a. Provide a project timeline, complete with milestones that include tasks and goals
b. Describe how each team member will contribute towards these milestones.

Week 4 (3/23 – 3/30): Iterate and Refine
Progress Update and In-Class Report (Due: March 30)

Tasks:

1) Your task is to finalize your data exploration and complete your project timeline and milestones
    - You should review your previous findings (from week 2)
    - Discuss as a team whether the analysis questions derived from these findings are reflective of what you want your final project to be.
    - Iterate over the exercises from week 2 and week 3 to finalize your project plan
2) Prepare for Check-In #1
    - We will be using class time on March 30th to do an in-person check-in to assess your progress and give you feedback.
    - Prepare a short slide deck and give us a presentation
        i. Your presentation should be about 5-6 minutes
        ii. Your slides should have at least 6 slides that include
            - (Slide 1): Team name and members' names
            - (Slide 2): Your understanding of the challenge and the data
            - (Slide 3): Your data exploration findings
            - (Slide 4): How your findings became analysis questions
            - (Slide 5): Your favorite analysis questions (the ones you think are the most promising)
            - (Slide 6): Your project timeline and milestones.

Deliverables (due 3/30):

1) Update and submit your process book
2) Presentation slides

Required addition to the process book:

You need to add the following to your process book:

a. A finalized list of analysis questions
b. Describe your rationale for these analysis questions given your data exploration findings.
c. Completed project timeline and milestones.

Weeks 5 and 6 (3/30 – 4/13): Data Preparation and In-Class Report (Due: April 13th)

Tasks:

1) Your task in this stage is to develop and design the programs and tools needed to solve the challenge. Tools like Tableau will not be sufficient to fully meet your analysis needs as this dataset and the analysis goals are unique and specific. You

will start to develop some of these tools in order to prepare your data (e.g., data cleaning, data transform, etc.), perform data analysis (e.g., with statistics), and visual analysis to solve the challenge. Specifically:

- You should report your progress and/or plan towards developing the data analysis and visual analytics system to answer your analysis questions. We will not provide specific deliverables because your plans will likely depend on your approach. However, your outcomes should include:
    i. The necessary programs/scripts to perform the data preparation and data transformation.
    ii. (Optional) Report the data analysis techniques that you will apply. For example, if you are using machine learning to do data analysis, you should have completed the program/script that can run the machine learning algorithm on the sample data. This program/script will likely need to be updated in the future when the whole system is put together (e.g., to tune the parameters). However, if your project involves data analysis, the initial functional version of this program/script should be completed.
    iii. Report a design of your visual analytics system. Describe the features of this system and how they can be used to solve the challenge. Note that you do not need to start the development of the system at this point. You just need to: (a) document the features or capabilities of your system (in terms of how they will be useful in solving the challenge), and (b) provide a sketch/drawing of what the system might look like.

2) Prepare for Check-In #2
    - We will be using class time on April 13th to do the second in-person check-in to assess your progress and give you feedback.
    - Similar to your previous check-in, you will prepare a short slide deck and give us a presentation
        i. Your presentation should be about 5-6 minutes
        ii. Your slides should have at least 6 slides that include
            1. (Slide 1): Team name and members' names
            2. (Slide 2): Your **finalized list of analysis questions**
            3. (Slide 3): Describe how you come to these analysis questions
            4. (Slides 4-5): What data transformations you have implemented and, if appropriate, what data analysis (e.g., statistics and machine learning) you have implemented. **Provide some explanation** to these decisions.
            5. (Slide 6): Your sketch and design for your final VA system. Explain how this design will help you solve the challenge.

Deliverables (due 4/13):

1) Update your process book to include:

   a. Finalized list of analysis questions
   b. Description of your data preparation and data analysis programs/scripts
   c. Sketch of your VA system, along with justification for why this design will help you solve the challenge.
2) Source code of your programs/scripts for your data preparation and data analysis (see the Guideline for more detail)
3) A slide deck for your Check-In #2

Week 6 - end (4/13 – 5/8): Implement VA system

Tasks:

1) Implement your VA system: using the sketch that you have developed in the previous weeks, you will begin the development of your VA system.
2) Use everything you have developed (including your data analysis results and your VA system) to answer your own analysis questions and come up with the answer to the VAST Challenge.
3) If you have a new finding that you had not previously discovered but you believe is the correct final answer, you will need to document this finding and how you come to this realization separately.
4) Complete the "Answer Sheet" that came with the VAST Challenge.
5) Prepare for the final demo and presentation

Live Demo! (During the Final Exam Period on May 8th)

1) You will be giving a demo of your completed system during the final exam period for this class (on May 8th)
2) You will also need to submit the following before the start of the class (11:59am):
   i. Your final process book.
   ii. Video demo of your system (more below)
   iii. Slides to your presentation (more below)
   iv. Source code and material

Required submission items:

- Final Submission (40% of project grade; due on May 8th)
  - Of the 40%, 15% will come from your report and your presentation. The remaining 25% will be based on your visual analytics system
  - Updated and final process book with screenshots of the system, data analysis process, findings and conclusions
  - Report your approach
  - A set of findings from your analysis
  - Complete the "answer sheet" that comes with MC1
- **A video** demo of your system / solution

    - Note that one of the tasks in the "answer sheet" is to submit a video (more on this below).
- **In-Class presentation** of the built visual analytics system
    - At least one coordinated visualization system (can be built on top of lab submissions)
    - Prepared slides

Video Demo Requirements:
Given that the final exam period is 120 minutes and we have 13 teams in this class, each team will be limited to a 8-minute presentation. This is not a lot of time to showoff all of your hard work! So we'd like you to put together a video where you can demonstrate your system to your heart's content. To be clear, **the purpose of the video is for us to better grade your system/solution**. A few thoughts:

- We suggest that the video is about 5 minutes in length, but feel free to take more time if needed.
- We would suggest that you either:
    - Provide narration/voice-over – at least at the key moments so that we understand what we're seeing, or
    - Provide text captions in the video
- Your video should demonstrate how you can use your system to answer your analysis questions and the questions in the "answer sheet" (the two shouldn't be too different).
    - Be sure to make both the "process" (of using your system) and the "finding" clear
- [Optional] Feel free to put in all the info about your system. For example, the architecture (e.g., using a database backend), the pre-computation or pre-processing, the cool interactions you designed, etc.
    - We imagine that you'll be telling us all of this in your progress book. However, having the key points of your system in a video will make it a lot easier for us.
- Remember the videos you watched in Lab 2. Use those as inspirations for how to structure your video.
    - We don't imagine that your video will have that level of polish – video recording and editing take a lot of time (time that you should probably spend on building your system/solution). But try to make the video as coherent and easy to follow as possible.
- Note that I use the term "system/solution" to reinforce the idea that the primary contribution or novelty of your final project doesn't have to be the CMV itself but can be a holistic "solution" (with emphasis on data analysis, pre-computation, etc.)

Presentation Requirements:

Because each team only has 8 minutes to present, you need to use that time wisely! Your presentation should contain the following:

- Title page (1 page)
  - Project (or team) name
  - Names of members
- Analysis Questions page (1 page)
  - Motivation, goal, and the analysis questions of your project
  - Highlight the key analysis question that you'll use in your demo later
  - **Note: Do not use this time to talk about the data or the challenge itself. We know what the challenges are**
- Content page(s)
  - Present how you solved this challenge (what pre-computation you did, what is the architecture of your system, etc.)
  - Do not to be exhaustive. Focus on the key elements of your project that made your solution successful!
- Live Demo!!
  - We suggest a demo where you use your system to investigate one of the analysis questions (you probably won't have time to do more than that).
  - However, this is just a suggestion. You should demo your system in whatever way that best showcases your solution.
- Finding page
  - Conclude with the outcome of the investigation (of your analysis questions).
- Discussion and Conclusion
  - This is somewhat dependent on your finding / solution, but some possible discussion points include:
    - How confident are you in your finding. Why?
    - What aspects of your solution is critical to your success?
    - Are there things that you wish that you had done differently?
    - Do you have insights into this challenge that you can share?

Submission: (Due on May 8th by start of class time**)**

- Submit the following 4 items:
  - Your complete process book that includes your answers to the "answer sheet"
  - a video where you demo your system/solution.
  - slides of your presentation
  - a complete zip file of your source code, material, etc.
    - The purpose of this isn't so that we can re-run your system – we understand that this is not likely going to happen if you use a database or run your own webserver, etc. Instead, think of this as an archive – the same files you would have on, say, a Github repo.

References:

[1] Examples of 2014 VAST challenge submissions:

- https://www.youtube.com/watch?v=AQg4VZcaaG0
- https://www.youtube.com/watch?v=THLB7JKYThM
- https://www.youtube.com/watch?v=AiYnmnrOkLo

FAQ:

- I found some videos online of other VAST Challenge submissions! Can I draw inspiration from them?
    - Please do not seek out the answer to the VAST Challenge. They are available online, but we trust that you will refrain from doing so. This project really isn't any fun if you know the answer before you start.
    - The same applies to watching videos of other people's submissions. Not only is this a violation of the honor code, it's also just ***boring.***
- What should my visual analytics system look like?
    - Hopefully by now you've seen a number of examples of visual analytics tools and systems, ranging from academic prototypes to commercial products. You should use those as design inspirations. As you might have noticed, a very common design is a "coordinated multi-view" system (we will discuss these in a future lecture). We are requiring that your final project use this design concept.
    - Keep in mind that the purpose of this system is primarily for an analyst to investigate the MC challenge (and not for a wide, public audience). In this sense, first and foremost you should be developing a tool that can help you solve the problem! If your tool also has a communication component, that's a bonus.
- How should I design my visual analytics system?
    - We recommend that you consider the various frameworks and models we have discussed in the class, such as the Shneiderman's Mantra, the Keim Model, etc. One thing to keep in mind is that the nature of the 2017 VAST Challenge is that there is a culprit or a root cause that you are trying to discover. In this sense, ideally your visual analytics system shouldn't just be an ***exploratory tool.*** It would be really nice if your system has some capability to help the user perform ***confirmatory analysis*** (i.e. verify the outcome).
- How much data processing and precomputation should I do?
    - We highly encourage you to use any tools and techniques you have learned throughout this course. This means that you are welcome to throw your data into Python, Tableau, sklearn, etc. to do whatever exploration, analyses, pre-processing, data-cleaning, etc. you need to do.
- What if my solution is more "computational" than "visual"?
    - Visual analytics systems are a hybrid of automation and (human) interaction. Under this umbrella, your system can be more of one and less of the other. In other words, if your solution has a strong algorithmic

or computational contribution (e.g. using ML techniques to pre-process data and find clusters or outliers that you incorporate into the original data) but less novelty in the visualization, that's perfectly fine as a final project. Just make sure that there's still a (coordinated multi-view) visualization component that helps the user analyze the outcomes of the algorithm.

- How robust or generalizable does my system need to be?
    - For your final project, we do not require that your tool needs to be applicable to other data or problems beyond the intended use. In terms of robustness, we also don't care if your system crashes on unexpected inputs. Chances are that we will not ever use your tool (we'll just watch you use your tool during the demo), but you should still be prepared if we ask you to perform certain actions during your demo.
- How will my final project be graded?
    - Your final project will be graded based on: (1) your process book, (2) your visual analytics system, (3) your ability to use your system to solve the VAST Challenge, and (4) your demo and presentation.
    - To be more specific on point 2, your system needs to be of a coordinated multi-view design. Ideally it will have both a computational (automation) component and an interactive interface. Further, it should have the ability to help the user analyze the data and discover the culprit.
    - **Note that whether you successfully find the culprit is not our primary concern.** Obviously, we want you to be successful in finding the right answer (and your grade will reflect this). However, we are more interested in knowing your ***process of discovery*** for solving this challenge.