correlation one

DS4A Colombia / 2020:
Practicum Guidelines

This document will provide you important information about the final project, or Practicum. The Practicum is an essential part of the DS4A course and is where you will be able to apply all you learn throughout the course.

Basic Information

The Practicum is an opportunity for you to demonstrate the knowledge you gain from the course, as well as to provide immediately useful outputs. Teams will consist of 5 - 7 participants each. There will be two types of teams throughout the program:

- **Institutional Projects** -- Teams will work on specific problems provided by public or private entities
- **Autonomous Projects** -- Projects will consistent entirely of self-directed projects

The Practicum is meant to incorporate skills from the entire curriculum and is broken up into three main components:

- A production-level application centered around the problem the team has selected
- A final report which details what the application does, how it does it, and the design decisions that were made in creating it
- A final presentation where participants pitch their projects and applications

When applicable, the entities for Institutional Projects will provide the teams with what they deem to be necessary data for the Practicum, but teams may supplement this with external data sources as they see fit.

We encourage teams to use technologies taught in class, but they are not required to do so.

correlation one

Project Scoping

The project scoping phase consists of two steps:

Clarifying the Problem: Institutional Project teams will not be given an explicit, well-specified task and told to do X, Y, and Z, so it will be up to them to translate the generic problem(s) presented to them into an actionable Practicum. Problem(s) are often communicated in a high-level, vague manner, so Institutional Project teams will likely need to gather additional information in order to ascertain the precise nature of the problem.

For Autonomous Project teams working on their own project idea, conforming to a specific need is not required, but this step is still extremely important because it will force them to be precise about exactly what problem they are solving. Too often teams that do not do this end up with a problem that is too vague, ambitious, or difficult given the time constraints.

Writing the Proposal: Once teams have a clear, precise understanding of the problem at hand, they can begin constructing their proposed solution. (The likely components of such a proposed solution will be discussed later in this document.) A few very important factors to consider when evaluating proposal quality are:

- **Feasibility**
  - Can your team accomplish this given your current skills and the skills you will be learning throughout the program?
  - Can your team complete this in the time frame of the program?
- **Impact**
  - Does the proposal directly address the need(s) identified in the previous step?
  - (For Institutional Project teams) Will the final product provide a material improvement over existing solutions/operations in terms of efficiency, accuracy, etc.?
- **Usability**
  - (For Institutional Project teams) Will the final product be integrable into existing systems and/or workflows?
  - Is the final product easy for one to familiarize oneself with and use?

Here is a Scoping Document Template that you can use as an example.

Once your team has finished your proposal and your TA has OK-ed it, you can move onto the actual deliverables.

Application

As mentioned, the application MUST work in real-time as it will be presented to government officials and private sector representatives. The specific components of the final application are going to depend heavily on what problem the team has chosen to address. It may need to include some components that are NOT discussed below. However, at the very minimum, each application MUST have the following three technical components: Front-End User Interface, Data Pipeline, and Analytical Tools & Models.

Front-End User Interface

If the team's application is going to be directly used by business end-users or even other data professionals, it will almost certainly require an interpretable and easy-to-use front-end interface. This means that:

- The front-end should make it clear to users how to get value out of the application and what that value is. This might involve a combination of meaningful outputs and visualizations.
- The front-end should be easily accessible via the Internet or some other universal access point
- There should be accompanying documentation for the entire application, which should be hosted along with the application.

We highly recommend using Dash (by Plotly) for this component, although participants are welcome to use any tool that they prefer.

Data Pipeline

The team's application needs to work in real-time during its final presentation (to be discussed later) and (for Institutional Project teams) should be able to be put into production without much additional effort. This means that the application ought to be hosted in the cloud (NOT on one's personal machine). The expectation is to use a cloud services provider (we recommend AWS) with AT LEAST the following minimum components:

- **Database (e.g. AWS RDS):** This should persist all relevant data that the team will be using. Teams must load the information in this database to the front-end for use.
- **Data Analysis & Computation (e.g. AWS EC2):** These should perform all computation on the data itself that is relevant to the application. They can be structured in a manner of the team's choice - microservices, chron jobs, scripts, etc. However, they must live in and run off of the AWS compute engine (NOT a local machine)

There may be more necessary components (it is up to your team and your TA to determine what those are), but we have outlined just the most common & essential ones.

Analytical Tools & Models

Your team's application should be driven by its own analytical tools & models that you have embedded into the back-end. Oftentimes, this involves a predictive model, but the project need not be predictive in nature. It could be a descriptive project; regardless, there should be a high level of data science involved to generate the insights.

Implementation-wise, these will likely live in the Data Analysis & Computation section of the Data Pipeline; however, your team should provide a full exposition of this in your documentation. See the Data Analysis section in the final report (to be discussed later) for additional details.

Additionally, although not directly part of the report (to be discussed next), **all of your team's code MUST be commented and submitted.**

correlation one

Final Report

The final report is a document that catalogues your team's execution of their Practicum. It will also serve as a supplementary source of documentation.

The final report is extraordinarily important and is arguably just as important as the application itself. This is because it will be the primary means by which your team will stay on track. The final report ensures that your team can catch mistakes before it is too late to fix them.

Content & Format: The report should be added to each week as your team encounters new issues and solves them. It should **NOT** be written in the last week. A typical layout will consist of:

- **Introduction:** This should introduce the problem/need, and summarize your team's process of scoping this into an actionable solution. It should state the context of your team's application and the exact problem they set out to solve. Your team should explain how your solution is distinct from existing approaches to the problem and what value it adds over those.
- **Application Overview:** This should cover what the application does, what the primary use cases are, and how a user would interact with it.
- **Data Engineering:** At minimum, this should contain at least two sub-sections:
    - **Interactive Front-end:** Teams should talk about the technologies they used as well as implementation details. They should discuss how the front-end passes and receives information to and from the other components. They should discuss which features they chose to include (e.g. visualizations) why they are important.
    - **Database:** Teams should discuss what type of databases they used, and the main data tables they set up. They should talk about how they designed them and why, as well as the technology tools they used throughout.
    - Additional topics that ought to be covered if relevant for their project:
        - Code/program design paradigms used
        - Flow charts/diagrams indicating how the different parts interact with each
- **Data Analysis & Computation:** In this section participants should provided a clear exposition of the mathematical tools used and it usually consist of the following:
    - **Datasets + Data Wrangling & Cleaning:** Teams should point to their data sources and explain the data cleaning process performed. It should provide a proper *justification* for the procedures used.
    - **Exploratory Data Analysis:** Teams should selectively showcase important and/or relevant portions of their investigative process. They should include data

visualizations alongside the observations and insights they gathered from them. It is imperative to add context and/or comments along with the data visualizations. Not doing so is not good practice.
    - **Statistical Analysis & Machine Learning:** Teams should walk through their analysis steps and why they made the choices they did at each step. They should explain why their model is valid.
- **Conclusions and Future Work:** In this section, the team should provide concrete, actionable conclusions based on their work. They are also encouraged to mention how their application can be expanded and improved.

For each section AND subsection, your team must clearly indicate what the percentage split in effort was among the team members.

Examples: Here are some examples of good reports from prior DS4A programs:

- Example 1
- Example 2
- Example 3
- Example 4

Presentation

The presentation will be showcased on the last day of the program. It should discuss the importance of their problem as well as their analysis and conclusions. It MUST have a live demo/exposition of their application. The presentation should also include a basic tutorial about how to use the application you created.

The most successful presentations usually have a great amount of interactive and meaningful data visualization and very subtle technical exposition.

Format: The presentations will generally:

- Use Powerpoint and other interactive digital tools
- Be approximately 10 minutes
- Focus on the overall process (scoping, methods and tools used, end results & deliverables) over specific technical details

More details on the presentation will be shared later in the course.

Teams will need to be polished when showing their work as it will be presented to both government officials and private sector representatives. You are expected to dedicate some time to putting together and rehearsing a presentation script which covers the following:

- What problem does your application try to solve? Why is this an important problem?
- On a high level, how does your application work? How do you use data science & engineering technologies and methods to do what it does?
- How does a user interact with and get value out of your application? Where do they put in their inputs and how do they receive the desired outputs? How should that user interpret and use those outputs?

You are expected to do dry-runs of this with your TAs, who will give you feedback.

Examples: Here are some examples of good presentations from prior DS4A programs.

- Example 1
- Example 2

correlation one

Timeline of Deliverables

The three main components - application, final report, and presentation - each have their own timeline of deliverables, as outlined below:

Overview

	Milestones			Deliverable
Week	**Application**	**Report**	**Presentation**	
1	*Team formation + Start on idea formation*			
2	Work on idea formation			N/A
3	Idea should be finalized & Start on Scoping			Project Description Doc
4	Project Scoping Completed	Intro, PS, Scope, Plan (V1 top Vn)		Submit Report with Scoping
5	Datasets sourced			Submit Report with details about dataset
6	Basic EDA/Cleaning of datasets completed			Submit Report with basic EDA details
7	In-depth EDA, jupyter analysis, mockup of frontend	Report with EDA and Analysis		Submit Report with in-depth EDA details and frontend mockup
8	Frontend Design			Submit Report with Frontend design+databas e
9	Backend Design + Front End Infrastructure Complete	App design (visual and otherwise) & analysis	Presentation outline	Submit Report with app infrastructure design

10	Application infrastructure complete	Add most recent data analysis + Final back-end design + Start conclusion	Presentation draft + Presentation run through	Presentation draft + Updated report
11	Project complete	Finalize report with exec summary	Final presentation incorporating TA feedback	Final draft of report + Final Presentation
Finale				

Application

The timeline for the application is as follows:

- Weeks 1 - 3
    - Team formation and (if applicable) problem assignment
    - Scoping work
- Week 4
    - Idea should be finalized
    - Scoping work
- Week 5
    - Datasets sourced
- Week 6
    - Basic EDA of the datasets with the goal of prioritizing which ones will be used
    - Cleaning of these datasets
- Week 7
    - More in depth EDA of the datasets
    - Jupyter notebooks of the analysis to be shown to the TA
- Week 8
    - Design of the front end should be completed and shown to the TA
- Week 9
    - Design of the back end should be completed and shown to the TA
    - Front end infrastructure completed
- Week 10

    - Back end infrastructure completed
- Week 11
    - Databases hosted in the cloud. Proof of which must be shown to the TA
    - Analysis and notebooks should be completed and shown to the TA
    - The application should be live with every component working as expected

Final Report

Your team should be continuously adding to the final report every single week - it should NOT be reserved for the end. Timely progress on the final report allows your team to stay on top of what you are doing and help catch dead ends before they soak up too much time.

- Week 4
    - Report with introduction, problem statement and scope, and plan of execution.
    - Multiple problem versions chosen:
        - All versions should naturally build on top of each other
        - V1 should be pretty easy and reasonable
        - The last version can be moonshot - the idea is that they must implement for V1 first, then V2, etc. so as to guarantee some sort of finished build by the end of Week 10. If they can get to V2, V3, etc. by the end, great, if not, at the very least they have something done that they can present)
- Week 7
    - Report should be updated to reflect EDA done and include new sections on the analysis done
- Week 8
    - Document should be updated to include sections on:
        - front-end design & mockup
        - data analysis elements
- Week 9
    - Document should be updated to reflect the most recent data analysis and the final back-end design
    - Preliminary conclusion to be completed
- Week 11
    - Final report completed with conclusion and executive summary

Presentation

- Week 10
    - Preliminary presentation outline to be shown to the TA
    - Presentation run-through with the TA during office hours

- Week 11
    - Final presentation completed based on TA feedback
- Week 12
    - Final presentation delivered

Rubric

The following is the rubric for the project deliverables. Teams may use this as a guide for successful final projects:

Click here to see the grading rubric.

Prizes & Accolades

The top teams will receive prizes, accolades and superior distinction at the end of the program. These teams will be recognized by Correlation One and MinTIC, and will be publicly announced via social media and press releases. The practicum is not only a vehicle for you to learn and apply what you learn in the course, it is also a way for you to have a significant impact on an organization in Colombia.