DS4A Latin America:
Participant Practicum Guidelines

This document will provide you important information about the final project, or Practicum, of Correlation One's DS4A Softbank LatAm Training program. It will also detail your expected deliverables, and give tips on how to achieve the best results.

Basic Information

The Practicum is an opportunity for teams to demonstrate their newfound skills. Teams will consist of 4 - 6 participants each. There will be two types of teams throughout the program:

- Type 1 teams will consist of ~50% Softbank portfolio company employees and ~50% non-Softbank participants, or 100% Softbank portfolio company employees
- Type 2 will consistent entirely of non-Softbank participants

For Type 1 teams, the core of the Practicum will be a real-world data science challenge provided by the assigned Softbank portfolio company. For Type 2 teams, the Practicum will be self-directed. Type 2 teams can adopt projects of interest to Softbank portfolio companies, or they can work on their own project idea.

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

- A cloud hosted 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 Softbank portfolio companies 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.

Teams will have to complete specific milestones throughout the program. Your assigned TA will ensure that these are completed on time.

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

Project Scoping

The project scoping phase consists of two steps:

Clarifying the Problem: Type 1 teams will not be given an explicit, well-specified challenge by their assigned Softbank portfolio company, and it will be up to them to use creativity to translate the business problem presented to them into an actionable Practicum. Business need(s) are often communicated in a high-level manner, so Type 1 teams will likely need to gather additional information in order to ascertain the precise nature of the business problem.

For Type 2 teams working on their own project idea, conforming to a specific business need is not required, but this step is still extremely important because it will force you to be precise about exactly what problem you 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.

Type 1 teams may not receive all necessary information for their projects in the first week. In such a scenario, the members of those teams who are Softbank portfolio company employees should make best efforts to obtain that information from their companies as soon as possible. Team members should use the extra time to work on their Extended Case assignments while they wait for that information.

Writing the Proposal: Once your team has a clear, precise understanding of the problem at hand, you can begin constructing your proposed solution. 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 Type 1 teams) Will the final product provide a material improvement over the company's existing solutions/operations in terms of efficiency, accuracy, etc.?
- Usability
    - (For Type 1 teams) Will the final product be integrable into the company's existing systems and/or workflows?
    - Is the final product easy for one to familiarize oneself with and use?

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

Application

The specific components of the final application are going to depend heavily on what business need your team has chosen to address, so it may or may not include everything discussion below. In fact, it may also need to include some components that are NOT discussed below. Nevertheless, below is an explanation of the most common/likely components a typical application will contain:

Front-End User Interface

If your 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

Your team's application needs to work in real-time during its final presentation (to be discussed later) and (for Type 1 teams) should be able to be put into production at the company 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 your 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 your 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 assigned 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 your own analytical tools & models that you've embedded into the back-end. Often times, 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, **all of a team's code MUST be commented and submitted.**

Final Report

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

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 you set out to solve. You 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: You should talk about the technologies used as well as implementation details. You should discuss how the front-end passes and receives information to and from the other components. You should discuss which features you chose to include (e.g. visualizations) and why they are important.
  - Database: You should discuss what type of databases you used, and the main data tables you set up. You should talk about how you designed them and why, as well as the technology tools you used throughout.
  - Additional topics that ought to be covered if relevant for your project:
    - Code/program design paradigms used
    - Flow charts/diagrams indicating how the different parts interact with each
- Data Analysis & Computation: In this section you should provided a clear exposition of the mathematical tools used and it usually consist of the following:
  - Datasets + Data Wrangling & Cleaning: This should point to your data sources and explain the data cleaning process performed. It should provide a proper *justification* for the procedures used.
  - Exploratory Data Analysis: You should selectively showcase important and/or relevant portions of your investigative process. You should include data visualizations alongside the observations and insights you 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: You should walk through your analysis steps and why you made the choices you did at each step. You should explain why your model is valid.

- Conclusions and Future Work: In this section, your team should provide concrete, actionable conclusions based on your work. You are also encouraged to mention how your application can be expanded and improved.

Your TA will meet with you for at least 15 minutes in the afternoons to discuss your progress and any issues you are running into. They will help guide you in the right direction based on this.

Presentation

There are going to be two versions of the final presentation. For Type 1 teams, the presentation will be showcased to the assigned Softbank portfolio company on the last two days of the program (May 29 - 30). It should showcase the importance of the problem as well as the analysis and conclusions. It should also have a live demo/exposition of your application, if at all possible. The presentation should also include a basic tutorial about how to use the application you created.

For Type 2 teams, the presentation will be showcased at either the career fair and to your fellow participants (if doing your own project idea) or to a Softbank portfolio company (if you are adopting one of the project challenges that said Softbank portfolio company put out).

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:

- Happen on-site, using Powerpoint and other interactive digital tools
- Be scheduled around the calendars of the company audience
- Be a group effort, with each participant talking through a section

Teams will need to be polished when showing their work to their assigned portfolio companies. As such, 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 TA, who will give feedback and ensure that the presentation is at a level suitable to the intended audience.

Short Version Format: There will also be a short version of this presentation which will be given at the career fair and to fellow participants. This version will:

- Be redacted of all sensitive information (i.e. any information that the portfolio company would not want to make public)

- Focus on the overall process (scoping, methods and tools used, end results & deliverables) over specific content

Timeline of Deliverables

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

Application

The timeline for the application is as follows:

- Weeks 1 and 2
  - Team formation and (if applicable) Softbank portfolio company assignment
  - Scoping work
    - Team members who are employees of the assigned portfolio company should provide any useful background information about the company and their needs/problems
    - All team members responsible for condensing these needs/problems into a well-defined project
    - All team members responsible for determining what additional information is needed to
- Week 3
  - Idea should be finalized
- Week 4
  - 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
  - Design of the front end should be completed and shown to the TA
- Week 8
  - Design of the back end should be completed and shown to the TA
- Week 9
  - Front end infrastructure completed
- Week 10
  - Back end infrastructure completed
  - 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
- Week 11
  - The application should be live with every component working as expected

Final Report

Teams should be continuously adding to the final report every single week - it should NOT be reserved for the end.

- Week 3
  - 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 10
  - 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 9
  - Preliminary presentation outline to be shown to your TA
- Week 10
  - Presentation draft completed
  - Presentation run-through with your TA during office hours
- Week 11
  - Final presentation completed based on TA feedback
  - (If applicable) Final presentation given to Softbank portfolio company
  - Redacted final presentation given at career fair

Accolades

1 - 2 teams per site will receive accolades and superior distinction. So do a good job - there will be recognition and prizes at the end!