# Data Science Fellowship

## Capstone Project Requirements

Completion of a capstone project is an integral part of the Data Incubator's fellowship program.

## Rationale

The capstone project is a chance for students to drive a project from conception through delivery. This demonstrates their abilities to apply the skills taught in the lectures and miniprojects. It also requires them to manage all stages of a project. The clear business use case and provided value of a good capstone project are generally more impressive than any technical or implementation details.

The best capstone projects will be shown off at Pitch Night, but all students will find that the projects help in their interview process. Some interviewers will ask about the capstone project directly, while others may ask the interviewee to talk about any project they've worked on. Many interviews will have questions asking for how a situation had been or would be handled. The experience from the capstone project offers concrete examples to use in these cases.

## Requirements

The details of an excellent capstone project will depend on the project itself, so we will not give detailed requirements. Nonetheless, all capstone projects must meet a few minimum requirements:

1. A clear business objective
2. Data ingestion
3. Visualizations
4. A demonstration of at least one of the following:
    a. Machine learning
    b. Distributed computing

    c. An interactive website
5. A deliverable

**1. Business Objective:** The project should have a clear and demonstrable business objective. This may be direct, building a product for consumers, or indirect, conducting analysis to improve the business of a company or government. It should be clear who the user of the product is, and how they will gain a benefit from the product.

**2. Data Ingestion:** The project should involve data processing beyond simply loading an existing data set. This may involve gathering data through web scraping or API calls, combining and harmonizing data from multiple sources, or non-trivial processing of existing data sets. Those projects using existing data sets should endeavour to find a unique perspective on that data. The amount of data being processed is highly dependent on the project itself. Many exemplary projects work with a gigabyte or more of data.

**3. Visualizations:** The project should contain at least **two** distinct types of visualizations. Types of visualizations include line plots, scatter plots, maps, connection graphs, low-dimensional projections of data, and word clouds. Distinct types of interactivity are also acceptable. These may be produced with tools such as Matplotlib, Seaborn, Pandas, Bokeh, ggplot, D3.js, or others.

**4a. Machine Learning:** The project should include the creation and use of one or more machine learning models. The project must involve **two** or more of the following topics: regression, classification, unsupervised learning, cross validation, analysis of feature importance, anomaly or outlier detection, deep learning, ensemble models, feature engineering, time series analysis, and natural language processing.

**4b. Distributed Computing:** The project should involve some distributed analysis of the data. This may involve processing in MapReduce, Spark, or another framework on top of Hadoop. Other frameworks may be approved at the instructors' discretion.

**4c. Interactive Website:** The project should result in a website that allows users to interact with the data. This may be in the form of an interactive exploration of the data or customized recommendations or predictions for users. This interactivity may be client-side, via Javascript, or server-side, via Flask or another web framework.

**5. Deliverable:** A deliverable should describe the work performed on the capstone as well as its primary results. This deliverable may take the form of (a portion of) a website, a Jupyter notebook, or some other type of document. It should describe the tools used, the process of data ingestion and analysis, and the major results. It should also include or link to the visualizations of point 3.