Team S3J2 Proposal

### INTRODUCTION

Team S3J2 is pleased to present our proposal for the 2018 IDI Analytics competition. Team S3J2 consists of five team members: Jared Beekman, Dr. Jordan Thomas, Dr. Sean Vermillion, Sean Tatman, and Samantha Oleson. For the 2018 competition we will be participating in the TrackML Particle Tracking Challenge.

At CERN, physicists use the Large Hadron Collider to accelerate and then collide particles in the hopes to better understand matter and the origins of life. Analyzing the millions of experiments has created a challenge for scientists and has led to interdisciplinary collaboration between the physics and machine learning communities. This competition aims to leverage the machine learning community's ability to accurately and efficiently reconstruct particle trajectories given the resulting collision's 3D points left in silicon detectors.

This competition is hosted on Kaggle, and more detailed information can be found on their website: <https://www.kaggle.com/c/trackml-particle-identification>.

### DESCRIPTION OF APPROACH

Our approach will consist of three phases: (1) independent exploratory data analysis (EDA) of the data provided from Kaggle and any additional, pertinent data, (2) independent application of machine learning techniques, and (3) a combination of our best models for a merged model submission. We will use weekly meetings and discussions via email/GitHub as our primary communication to provide feedback and improve scores.

During phase 1 – independent exploratory data analysis (EDA) – each S3J2 team member will be given one of the 5 training datasets to work with. Team members will use data processing and engineering techniques they are most familiar with to explore and become familiar with the data. The first team discussion will consist of each team member sharing their EDA results, and any conclusions, inferences, or problems found through this analysis. The goal of this two-week phase is to gain a deep understanding of the variables and the relationships among them.

The remainder of the independent analysis, phase 2, will focus on generating particle trajectory reconstructions using machine learning and data visualization techniques. Each team member will use the same training dataset they analyzed during phase 1, to generate an algorithm that can accurately reconstruct particle paths. Our goal is to provide, at minimum, five submissions per week – one from each team member. As we prune our individual models, we will determine which models/methods have the highest accuracy and prepare our aggregation tool for the final phase.

In phase 3, we will provide final submissions using a merged model concept. Historical results during similar competitions have supported the merged model concept, where the best scores have blended diverse models to obtain better scores than individual models. The last week of the competition will focus on providing the aggregated submission, model documentation, and visualization.

Jared, Jordan, and Sean Vermillion will primarily work within the Python environment; Sean Tatman will work within the R coding environment; Samantha will work within Tableau. Other coding environments will also be employed if found useful.

### PROJECT STATUS AND SCHEDULE

Below is our estimated timetable for completion of the project:

|  |  |
| --- | --- |
| **Milestone** | **Deadline** |
| Download and distribute data via GitHub | 18 June 2018 |
| Complete Phase 1: EDA | 1 July 2018 |
| Phase 1: EDA Team Meeting | 2 July 2018 |
| Complete Phase 2: Independent Modeling | 22 July 2018 |
| Phase 2: Independent Modeling Meeting | 23 July 2018 |
| Complete Phase 3: Merged Model | 29 July 2018 |
| Phase 3: Merged Model Meeting | 30 July 2018 |
| Final Submission | 31 July 2018 |
| Documentation Submission to IDI | 6 August 2018 |

The team has licenses for all software currently in use except for Tableau. Samantha will use a free trial version of Tableau until we pursue purchasing IDI license(s). We have established a designated GitHub site as the primary file sharing location and version control software (VCS).

### MODEL VALIDATION

Model validation is accomplished through the use of test data sets. A standard 90% training, 10% test data set will be used for model training; however, custom splits will be explored during the competition.

Kaggle provides a test data set for model evaluation. Because this competition has a cash prize, the public leaderboard on Kaggle’s competition website only uses 29% of the test data set, reserving the remaining 71% for the final, private leaderboard for evaluation. This discourages manual overfitting that can artificially boost a submission’s score.

### BENEFIT TO IDI AND ITS CUSTOMERS

This competition will benefit IDI in multiple ways. Chiefly, each phase of this competition will foster insights for future modeling work. EDA is a common skill used across projects at IDI, and we will discover innovative techniques and visualizations that improve future work. By individually modeling during phase 2, we will improve individual skill sets in these areas as well as foster knowledge share. The merged model concept has proven effective in many IDI projects and this project will further our understanding of how to use it. Additionally, if IDI performs well it will be good publicity within the Kaggle community and may lead to new future customers. Lastly, improving our joint application of data visualization will benefit our customers seeking this kind of analytics. If these discoveries are worthy of publication, we may write a joint paper.