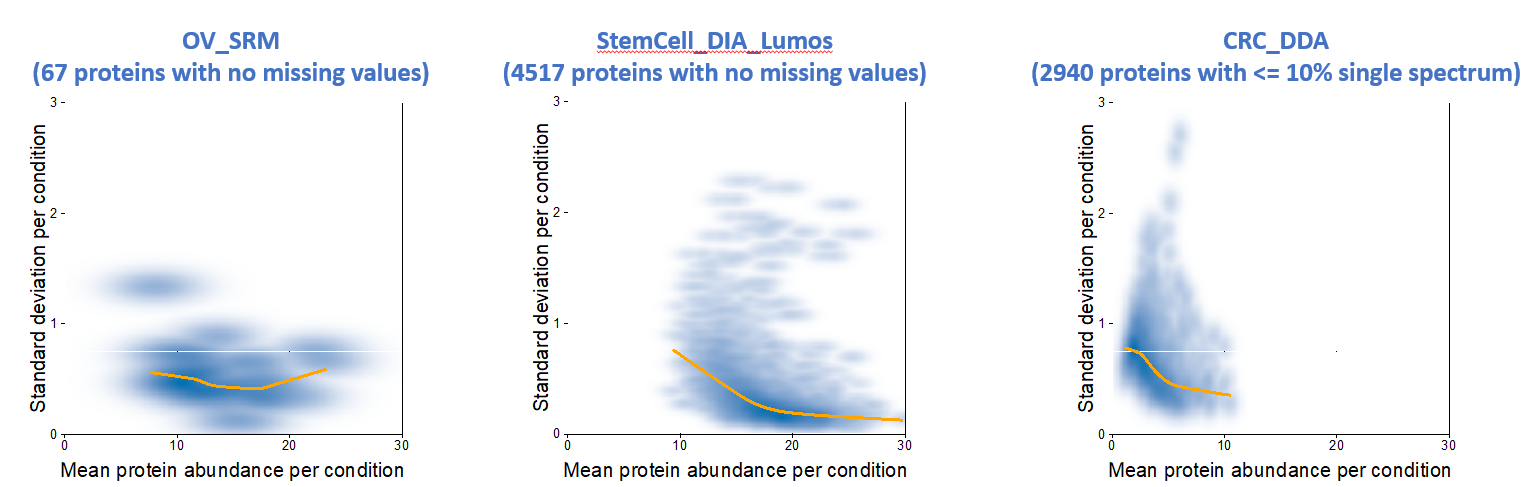
Spring 2020 Project Proposal

**Overview:** MSstats is an R-package which enables statistical analysis for Quantitative Mass Spectrometry-Based Proteomics. This package along with its ecosystem of packages namely MSstatSampleSize & MSstatsBioData allow users to experiment by simulating multiple datasets. The need to simulate dataset arises from the fact that data acquisition for proteomic experiments are time consuming and the cost to curate this data is astronomical. The current implementation of this simulation is purely command line based with human input required at every step of running the simulation, visualization that are generated as a byproduct of these simulations are static and do not integrate well with a dashboard environment that can be tested out in depth.

The following figures are examples of mean & standard deviation plots which are generated at the data exploration stage.



**Prior Work:** There exists an implementation of the web-application which is in its infant state and does not provide for all the requirements of the researcher, the shiny application can be found [here](https://github.com/Vitek-Lab/MSstatsSampleSize-Shiny)

**Proposed Work:** This project will focus on automating the simulation process and application of various knowledge discovery algorithms to enhance decision making by a research scientist. This will be accomplished by developing a Shiny Web Application in form of an R-package that can either be deployed to the cloud or that can be run locally on a researcher’s machine. The idea as it stands is to enable a researcher to provide a dataset that has been pre-processed as required by the web-app and augment the exploration process. The application will provide methods to run variety of experiments on the fly using various machine learning algorithms and the ability to tune parameters as and when required. As an addendum to prevent possible faults in the development process a design documentation will be deployed with set goals for each stage. Robust tests will be developed to minimize application crashes and disconnects.

Simulating Dataset, as stated in the Overview acquiring Proteomic data is both time and cost intensive as an alternative to these expensive methods data will be simulated. The simulation experiment will depend on the statistical information of some sample data, this will be further used by classification algorithms to discover and validate the simulations.

**Relation to Data Science:** Simulating experiments that are reproducible have been the crux of all Data Science projects. Being able to operationalize these experiments as a proof of concept software is expected out of a Data Scientist, which can communicate the results with the prospective customer effectively in a user-friendly manner. With this project a web-application will help operationalize the Statistical experiments as defined within the MSstats ecosystem and visualize the finding ergonomically while maintaining robustness of the application and maintaining agile development practices.

**Timeline:** The MVP of this project should be completed in the second half of the Spring 2020 semester, with the latter part dedicated to developing tests and feature updates as requested. The initial steps of the project will be to understand the workings of the MSstats packages and develop a design documentation determining the workflow and the various requirements of the application.

The second step would be to develop an MVP which is able to operationalize all basic requirements which will be achieved by a Shiny Web Application.

The final step will be to develop testing infrastructure around the built app to help test out newer functionalities that might be added out later which will enable the determine the compatibility of the application.

**Resources:**

* [MSstats: A Statistical Tool for Quantitative Mass Spectrometry-Bases Proteomics](http://msstats.org/)
* [MSstatsSampleSize: Simulation tool for optimal design of high-dimensional MS-based proteomics experiment](https://bioconductor.org/packages/release/bioc/html/MSstatsSampleSize.html)
* [R Shiny Tutorials](https://shiny.rstudio.com/tutorial/)