Practicum Proposal: Fall 2019, 8-Week1 and 8-Week2

1. **Contact:** Lolita Mannik

[lmannik@regis.edu](mailto:lmannik@regis.edu)

720-609-9761

1. **Project Title:** Novel Database and Usage Metrics for NCAR’s Climate Model
2. **Project Description**

**Goal:** I am proposing a 16-week practicum project that uses and builds upon analysis from my summer internship at NCAR, which will help me develop my skills toward analyzing data in a scientific or medical field. I’d like to specifically showcase statistical and machine learning methods.

**Background:** The Community Earth System Model (CESM) is a highly configurable coupled climate model used by climate researchers across the world. Model configurability comes in the form of biologic, geologic, and atmospheric components mapped to surface and atmospheric grids, and is customized with physics options, all of which in turn affect the model’s computational performance – an important constraint in designing both model software and scientific experiments. In my internship, I created a SQL database by parsing text-based timing files of all experiments run on the Cheyenne supercomputer for the Coupled Model Intercomparison Project 6 (CMIP6) to learn how climate scientists use this sophisticated model.

**Internship:** My study identified which configurations are being used, their relative rates of usage, and their computational performance, which can vary significantly. For example, model cost and CPU hours were analyzed by configuration and case, and the effect of a system upgrade on model cost was assessed for those cases running both before and after the upgrade.

**Value:** This analysis allows NCAR to track configurations to inform model performance and development. Currently, no such collection of comprehensive metrics exists for CESM.

1. **Data Science Tasks**

**Data Engineering:** My current data collection includes timing files for all CMIP6 experiments on Cheyenne as of July 16, 2019. I propose to expand the collection to include more recent CMIP6 files, and to expand the collection to many other CESM experiments on Cheyenne. This involves mostly parsing text files to structured JSON (parser was developed by my mentor), converting that JSON to SQL, and storing them in a SQL database hosted on AWS (JSON to SQL converter was developed by me in Python). A significant outreach opportunity exists to contact CESM users at other institutions in order to collect and parse their timing files. I propose to lead the outreach and collection effort.

**Statistical Analysis:** While at NCAR, I got to look at the effect of a rare system upgrade on the Cheyenne supercomputer. The system was down for a week in early July 2019 and no one could run their simulations during that time. We identified 14 unique climate simulations that were running both before and after the upgrade and compared the mean model cost per case before and after the upgrade and found that in every case, the model cost was higher after the upgrade, meaning it was taking longer to run the same simulation. I also ran a T-test and found the difference to be statistically significant for the 4 cases with the highest % difference. A cursory look found performance degradation to correlate with an increase in the number of processors used in each case, but there is a lot more detective work that can be done here by looking deeper at the configurations involved, account for those cases that have adjusted their processor count, and plot post-upgrade performance week by week to track whether the system has reached post-upgrade performance.

**Machine Learning:** Aside from the upgrade analysis, and once more data is collected from Cheyenne and CESM users outside of NCAR, the data will need parsed, stored in a SQL database, and cleaned for outliers and nulls. Next, I’d like to perform feature engineering on the entire dataset and use machine learning to analyze and predict performance. I’d also like to use clustering techniques to see if the reveal any patterns.

1. **Data Description:** I am highly familiar with this data. My current dataset is 5,160 rows x 86 columns and could be substantially larger if we have wide cooperation with our data outreach effort. The variables include experiment IDs, components, grids, various run-time and simulation timers, settings related to parallel computing, and performance metrics.
2. **Methods:** I would like to use any applicable statistical methods that we learned in MSDS 660, such as regression techniques and parametric or non-parametric tests. Regarding machine learning, I’d like to use SVMs, k-means and Random Forest to correlate features and identify patterns in performance and to see if I can predict performance based on configuration.
3. **Potential Problems:** I feel that I already have strong mentorship support and a solid background with this project, but that I am somewhat limited by my knowledge of Python. I feel we have already spent 11 weeks this summer jumping over hurdles just to get to this point! But, as with most novel research, we need to remain flexible and tailor our analysis to where the data is leading us.
4. **Timeline:**

|  |  |  |
| --- | --- | --- |
| **Week** | **Activity** |  |
| Week 1 (Sept 5) | Tutorials on Python, MatPlotLib, regular expressions, etc. I really need them. |  |
| Week 2 (Sept 12) | Launch data collection outreach. Request data from CESM users outside of NCAR, giving them instructions and a deadline for timing file submission. | Data Collection Outreach |
| Week 3 (Sept 19) | Collect, parse, and clean new CMIP6 data from Cheyenne for upgrade analysis | Line arrow: Straight |
| Week 4 (Sept 26) | Identify all experiments spanning the upgrade and EDA, such as those with consistent settings before and after. | Line arrow: Straight |
| Week 5 (Oct 3) | Upgrade analysis and parametric testing | Line arrow: Straight |
| Week 6 (Oct 10) | Upgrade analysis, plotting and visualization | Line arrow: Straight |
| Week 7 (Oct 17) | Clean up upgrade analysis code and write up/prepare this segment for the final presentation. | Line arrow: Straight |
| Week 8 (Oct 24) | Assess data collected from outreach efforts. Work with NCAR mentors to parse the timing files and create JSON files. |  |
| Week 9 (Oct 29) | Convert new data JSON to SQL. Clean outliers, nulls, etc. |  |
| Week 10 (Nov 5) | Subsetting and statistical analysis. |  |
| Week 11 (Nov 12) | Continued statistical analysis. |  |
| Week 12 (Nov 19) | Subsetting and machine learning. |  |
| Week 13 (Nov 26) | Continued machine learning. |  |
| Week 14 (Dec 3) | Review and clean up code from stats/machine learning. |  |
| Week 15 (Dec 10) | Prepare presentation. |  |
| Week 16 (Dec 17) | Deliver final presentation. |  |