11 Jan 2019

SIParCS Program at NCAR

To whom it may concern,

What summary statistics are available from point cloud data sets, when one should make a minimal interpretation of the point cloud? In my doctoral study of topological data analysis (TDA), I am interested in this genre of agnostic model selection problems. Closely related is the SIParCS project question: what statistics are available from an aggregate of imperfect models, provided that one is allowed to mung, discard, and summarize the space of such models? My goal for SIParCS, then, is to lead "Applying Machine Learning to Maximize Information Extraction from Imperfect Climate Models" by performing model agnostic data analysis.

1. In terms of technical sophistication, my basic computing toolkit includes version control (git, duplicity), high level (R, Python, Haskell) and low level (C++) programming languages, and markup languages (`ipynb`, `Rmd`, XML, HTML, pandoc markdown, LaTeX). I am accustomed to both Unix-like (MacOS, Ubuntu/Debian) and Windows operating systems. I am comfortable performing computations on the Google Cloud Platform, and I have introductory knowledge of XSEDE science gateways from participating in SGCI webinars.

2. Regarding machine learning: this fall I contributed to CU Boulder's StatOptML (Statistics, Optimization, and Machine Learning) seminar. I applied TDA to consider "is depth needed for deep learning?" following Guss and Salakhutdinov's empirical study [1]. By using persistent homology to catalog feed forward neural architectures, I introduced myself to tensorflow and Ripser, software that I would be enthusiastic to deploy in collaboration with peers and faculty at NCAR.

   Prior to joining the PhD program at CU, I strongly considered a master's study of *sediment transport* and *ground water contamination*. Between this precursory interest and my undergraduate work in physics, I have an introductory understanding of climate modelling.

3. Statistical and computational experience would support my career goal to do topological data analysis. There are limited opportunities in my department to train with powerful computational tools. I believe that SIParCS would complement my theoretical strengths, which are afforded to me by my enrollment in a *pure* mathematics department.

4. In terms of unique contributions to the program: After my undergrad, I took two years to perform stipended service work. For a year in Houston, I developed scalable resources for refugee case management, including a crowd-sourced map of clinics and languages spoken. I wrote bug reports for the implementation three SQL databases, and, when Texas cut funding for Refugee Medical Assistance, I contributed to a data management plan for refugees transitioning from state to federal medical care. For a year in Olympia, I served as a community organizer at a 24/7 homeless shelter. I relied on distributed version control, and became a staunch advocate for deploying "early, often, and with redundant backups". In all, this background experience allows me to contribute to a inclusive research environment. I strive (i) to collaborate, e.g., to focus my effort on tasks where I have a comparative advantage, (ii) to make incremental contributions on the work of others, and (iii) to be transparent, so that others may work off of my contributions.

[1] W. H. Guss and R. Salakhutdinov, "On Characterizing the Capacity of Neural Networks using Algebraic Topology," *arXiv:1802.04443 [cs, math, stat]*, Feb. 2018 [Online]. Available: http://arxiv.org/abs/1802.04443. [Accessed: 21-Nov-2018]

Thank you for your consideration,

Colton Grainger