PERSONAL, RELEVANT BACKGROUND, AND FUTURE GOALS STATEMENT

COLTON GRAINGER

My career goal is to be a research scientist at US national laboratory, focusing on the design of robust mathematical models for complex systems. My immediate goals are

- to complete a prototype[1] of the research data archive image module (`rdaim`) as a student visitor at the National Center for Atmospheric Research (NCAR);
- to build skill as a mathematician by deepening my understanding of statistics, scientific computing, and applied algebraic topology; and
- to develop rigorous, engaging, and inclusive open source curriculum[2] for lower division undergraduate mathematics courses at the University of Colorado Boulder.

Intellectual Merit

As a graduate student, I feel the use of advanced statistics in the context of historical climate data would be a novel intellectual contribution to the informatics and climate field. I am passionate to design a dynamic, iteratively refinable, metadata schema with thought to the future use of the 2KB observational data (from each 6MB unreduced image) for *data assimilation* and climate modelling.

Boulder's Ambient Research Community. Because Boulder hosts the University of Colorado and two atmospheric research laboratories,—NCAR and the National Oceanic and Atmospheric Administration—constituting a diverse population of climate scientists and academics, I have proposed a graduate research plan in the Research Data Archive at NCAR. I feel I have a comparative advantage to produce generalizable, interdisciplinary knowledge *for climate science* by collaborating with expert statisticians, computational scientists, and topologists in Boulder.

Moreover, NCAR is in the process of acquiring and hosting the ~60 TB collection of unreduced meteorological images described in my graduate research plan. Without a robust mathematical model for metadata reduction, the images may be archived in a distressingly[3] unpartitioned state.

In the status quo, images have only been effectively classified through human effort,[4] in a page-by-page fashion through recovered documents. However, page-by-page human classification of these images is not timely for the data assimilation needs of climate reanalyses. At the present rate, the estimated ~100 TB collection of historical meteorological documents (~60 TB available to NCAR and ~40 TB in diaspora among international archives) would require ~300 years to classify.

[1] Source code: `https://github.com/NCAR/rda-image-archive`.

[2] E.g., Fall 2019, I am teaching sophomore-level statistics (`https://math2510.coltongrainger.com`) from *OpenIntro Statistics v4.0* (`https://openintro.org`) published in the creative commons (CC BY-SA 3.0).

[3] It would be fair to point out that in the status quo, the National Archives and Records Administration does provide excellent document-level metadata for meteorological records, but I would argue the desired granularity of data required from these documents for climate modelling necessitates at least two further categorical refinements of document-level metadata: first down to the image-level and further down to observation-level, both of which are statistically and mathematically information rich.

[4] According to Philip Brohan: trials to classify even *typed pages* of Naval logbooks circa 1950 with Amazon's `textract` (an optical character recognition system) yielded only 85% accuracy as compared to a faithful human transcription of the documents.

Considering that policy-makers need now to be informed by predictive climate models that are well validated and interpolated through historical meteorological data, this timeline is unacceptable.

Invariants for categorical metadata. Having worked this summer to understand the rich spatio-temporal aspects of image metadata for meteorological records, I am excited to formally describe and locate the images and the associated reduced observations in ambient categories that are both physically meaningful and mathematical well behaved.

For example, treating the location of an ocean platform as a path in a smooth manifold and describing the observational data faithfully recorded by such a platform as a time-series of sections into the state space of meteorological variables over the manifold suggests that a mathematical method partially grounded in algebraic topology and differential geometry yet also informed by inferential statistics could produce a set of formal rules for reducing the uncertainty associated both to the location of the platform over time and the observational data itself.

From an information theoretic standpoint, the reduction of a 6MB file to a 2KB file, without significant human effort and without any relevant context, ought to be difficult. An opposing view, from the perspective of algebraic topology and differential geometry, might aim to exploit as much structure as is available (e.g., by finding "topological" invariants for objects in the category of platforms that are preserved under metadata reductions) to generate statistical priors for unreduced image metadata in series with fully classified images. In the later school of thought, my proposed graduate research plan aims to study the formal mathematical properties of intermediate time-scale (~50 years) physical models of natural processes.

Broader Impacts

Service.

- Having taught 9 weeks of sophomore level statistics I am beginning to recognize that the mathematics department at CU Boulder would benefit from curriculum that is *open source* and *published freely*. In particular, I have observed that the students who withdraw from mathematics courses because they unable to pay for currently expensive online course materials have been primarily students of color, or students with limited English proficiency.
- Drawing on positive experiences from my undergraduate institution, the College of Idaho, in being introduced to open source software for mathematics (`SAGE`, `PreTeXt`, `WeBWorK`) as a sophomore, I have begun to develop open source curriculum[5] for students in my statistics course. If awarded a fellowship, I would pursue the further creation and curation of open source content.
- Moreover, I aim to reach underrepresented and underresourced students at CU Boulder by facilitating a summer bridge program through the Laboratory for Interdisciplinary Statistical Analysis in advanced statistical methods for image classification. To this end, I am presently working with Suchit Sharma through the Boulder Valley School District mentoring program to design linear algebra curriculum for image classification.

Assets.

- After my undergrad, I took two years to perform stipended service work. For a year in Houston, TX, 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,

[5] `https://math2510.coltongrainger.com/guide/`

I contributed to a data management plan for refugees transitioning from state to federal medical care. For a year in Olympia, WA, 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.