1. Application Details

- Education
  - Ph.D. Student in Mathematics, University of Colorado, Boulder, CO (2018–Present)
  - B.S. in Mathematics-Physics, The College of Idaho, Caldwell, ID (2012–2016)
    - Senior Study: Galois Theory for Differential Equations.
    - Advised by Dr. Jonny Comes
    - Full-tuition merit scholarship for undergraduate studies. [Awarded to 11 of 287 first-year students in 2012.]
- Experience
  - Student Visitor, Research Data Archive, National Center for Atmospheric Research, Boulder, CO (Present)
  - Software Engineering Intern, National Center for Atmospheric Research, Boulder, CO (Summer 2019)
    - Created an image repository for historical weather data (e.g., marine logbooks from 1870 to 1950) to support climate reanalyses. Funded by SIParCS. Mentored by Thomas Cram, Matt Mayernick, Steve Worley, Philip Brohan. [Python, SQL, XML]
  - Social Work Intern, United Way of Thurston County, Olympia, WA (2017–2018)
    - Supervised volunteers, interns, and work-studies at a 24-hour shelter for families experiencing homelessness. Funded by CNCS. Mentored by Lindsay Fujimoto, Abbigail Shirk. [JavaScript, community organization]
  - Social Work Intern, YMCA of Greater Houston, Houston, TX (2016–2017)
    - Managed data for the preferred communities refugee medical assistance program. Funded by TX-ESC. Mentored by Shaoli Bhadra, Danielle Bolks. [SQL, health records, immigration policy]
- Proposed Field(s) of Study
  - Mathematical Sciences: Computational and Data-enabled Science
- Proposed Graduate Study and Graduate School Information
  - Ph.D. in Mathematics
  - University of Colorado Boulder
- References
  - Agnés Beaudry (Math Dept., CU Boulder)
  - Thomas Cram (Research Data Archive, NCAR)
  - Matthew Mayernik (Library, UCAR)

2. Graduate Research Plan Statement

I propose to continue work on the classification and reduction of meteorological data from binary image files (e.g., ship log books) using methods in statistics/machine learning, applied topology, and the digital humanities.

As a student visitor at the National Center for Atmospheric Research (NCAR), I am extending my project as a summer intern to classify images of handwritten meteorological logbooks from ocean and land platforms (ships and weather stations) in order to harvest meteorological data from them for the International Comprehensive Ocean-Atmosphere Data Set (ICOADS). My immediate goals are

1. to gather images into (at least) one repository and associate each image to a persistent identifier,
2. to establish a common description framework for the image metadata, and
3. to provide bulk, programmatic access to image subsets.

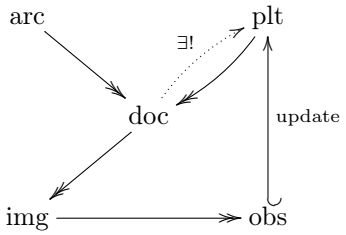
I am committed to accomplishing these three tasks in the next year, but would necessarily rely on a fellowship to extend this project for the duration of my graduate career. My goals for an extended project would be

1. to develop a framework for eliciting most useful classifications of individual images in a time-series,
2. to generalize the current geospatial metadata schema in the language of applied category theory, and
3. to study and reduce the statistical uncertainty associated to each image's geospatial metadata.

2.1. **Which images will be classified?** The archetypical image of a meteorological logbook contains 24 hours of weather observations (barometric pressure, sea-surface temperature, etc.) from an ocean-faring platform circa 1860 to 1960. The mathematical problem is to reduce each 6Mb image to a 2Kb time series of weather observations. To this end, I have develop a sophisticated metadata schema for images consisting of 5 categories

1. (`arc`) the archive of origin,
2. (`doc`) the source document,
3. (`img`) the binary image file itself,
4. (`obs`) the meteorological observations available from the image,
5. (`plt`) the platform responsible for producing the observations,

with functional dependencies given by the following diagram



Roughly, an image in the category img approximates a time-series $\sigma\colon [t_0, t_1] \to \mathscr{S}$, where $\mathscr{S}$ is the state space of meteorological variables. An observation is $\sigma(t)$ evaluated at a time t. A platform in the category plt approximates a time-series $\lambda\colon [t_0, t_1] \to M$, where M is the Earth's surface (as a manifold). An image is successfully classified if both σ and λ are determined up to some statistical uncertainty.

2.2. **What is the broader impact of reducing meteorological data from these images?** To uniquely identify records used to construct historical climate models.

- open issues on github
  - metadata schema for digital images of ship logbooks
  - semi-sequential UUIDs (partially ordered set structure, persistent identifiers)
  - initialize (local and remote versions) of test database "images"
  - realize spatio-temporal metadata for images (differential geometric interpretation)

  - tools for pre-ingest metadata validation (requires information theory)
  - how to query the database locally? how to bulk download images matching a query?
  - implement boundary polygons at logbook level (requires statistical modelling)
  - standardize filetype schema with Zaihua Ji
  - standardize metadata schema with Bob Dattore
  - programmatic metadata ingest
  - streamlined process for uploading images and metadata to the RDA, post-metadata verification
  - partnering with CU Boulder and Rocky Mountain Advanced Computing Consortium for image hosting
  - accomodating two use cases: Philip Brohan (UK Met Office) and Kevin Wood (UW/NOAA)
- deliverable: research data archive image module (https://github.com/NCAR/rda-image-archive)
  - python package to support (meta)data munging for images at the RDA in NCAR
    * tools for the metadata provider
      · `metadata`
      · `uuid`
      · `bundle`
      · `database`
    * RESTful API for queries

2.3. **What is the intellectual merit of extending methods from statistics, applied topology, and the digital humanities to do so?** The use of advanced statistics and machine learning in the context of historical climate data is a novel intellectual contribution to the informatics and climate field.

I am motivated by two heuristics: to create *long content* with *low noise-to-signal* ratio. The utility of pre-1960 climate records is apparent—my project specific interest is rather in formatting and deploying a repository of such records. By which gateways will the data be accessed? Should we create additional interfaces for non-specialists? How can historical climate records be visualized? How can crowd-sourced data entry be verified?

This project builds on:

- the digital humanities by providing persistent identifiers for each image
  - Mayernik, M. S., & Maull, K. E. (2017). Assessing the uptake of persistent identifiers by research infrastructure users. Plos One, 12, e0175418. doi:10.1371/journal.pone.0175418
- the field fo applied topology by developing categorical techniques for describing time-series metadata and formal reductions of this metadata
  - Computational Topology Techniques for Characterizing Time-Series Data (Nikki Sanderson, CU Boulder, https://arxiv.org/abs/1708.09359)
- image classification techniques
  - On Characterizing the Capacity of Neural Networks using Algebraic Topology (William Guss, CMU, https://arxiv.org/abs/1802.04443)
- computational techniques for time-series analysis
  - Lagrangian Data Assimilation and its Applications to Geophysical Fluid Flows (Laura Silvinki, NOAA)

3. Personal, Relevant Background and Future Goals Statement

I aim to become a stronger researcher, a more deliberate teacher, and a competent mathematician during my time as a Ph.D. student at the University of Colorado in order to contribute to the broader climate science community in Boulder.

3.1. How do I plan to become a stronger researcher?

- 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 the NSF GRFP would complement my theoretical strengths, which are afforded to me by my enrollment in a pure mathematics department.
- By contributing to open source software development.
- By understanding the formal semantics of metadata.

3.2. What would be the broader impact of my work as a teacher?

- I plan to develop open source curriculum for statistics and applied topology.
- I am presently working with Suchit Sharma through the Boulder Valley School District mentoring program to design linear algebra curriculum for image classification.
- 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, 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.

3.3. What would be the intellectual merit of my career in mathematics?