1. ambient researchers

- Jeff Erickson, University of Illinois, Urbana, IL, USA
- Ulrich Bauer, Technische Universität München, Germany
- Kathryn Hess, EPFL, Lausanne, Switzerland
- Henry Adams, Colorado State University, Fort Collins, CO, USA
- Nicole Sanderson, Lawrence Berkeley National Laboratory, CA, USA

2. application materials

- Personal Information
- Education, Work and Other Experience
- electronic transcripts
  - College of Idaho
  - University of Idaho
  - University of Colorado Boulder
- Proposed Field(s) of Study
  - MATHEMATICAL SCIENCES: Topology
  - MATHEMATICAL SCIENCES: Computational and Data-enabled Science
  - COMPUTER AND INFORMATION SCIENCES & ENGINEERING: Communication and Information Theory
- Proposed Graduate Study and Graduate School Information
  - Ph.D. in Mathematics
  - University of Colorado Boulder
- the names and email addresses of at least three reference letter writers
  - Agnés Beaudry (Math Dept., CU Boulder)
  - Thomas Cram (Research Data Archive, NCAR)
  - Matthew Mayernik (Research Data Archive, NCAR)

3. personal, relevant background and future goals statement

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.

% what this self study? how long for? what are you credentials? what about the location or the style or domain of these? % might not be the right skills

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 [GS18]. 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.

% what were the two talks? center my contributions or process around "broadening my knowledge"?

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 the NSF GRFP would complement my theoretical strengths, which are afforded to me by my enrollment in a pure mathematics department.

% courses that I should take % what access? also to teach? access to resources? TIME! % passion killing me for teaching

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, 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.

% glad to identify emergent themes

4. graduate research plan statement

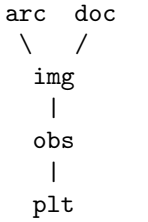
% say explicitly that a proto-type has been built, but

The robustness of a data management plan determines its utility. Just imagine designing a metadata scheme that fails to account for partial information, or setting up a database under the assumption that future users will only want a single set of hard-coded queries. My goal this summer, then, is to lead the project "Building a Historical Data Image Archive to Support Climate Research..." by designing and prototyping a robust repository for historical climate documents.

% including well defined research questions % needs some mathematical rigor % pick methods and justify expertise (computational topology) % already skilled? no % professor to name drop? yes % plan to take courses? most definitely % am a seasoned beginner to pursue

From exposure to data management in healthcare, in personal records, and in open source development, 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?

- mathematical problem: reduction of 6Mb image to a 2Kb time series
  - e.g., 24 hours of meteorological observations from the doomed USS Jeannette, circa 1879
- method I: develop sophisticated metadata schema
  - categories: archive, document, image, observation, platform
  - 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 $\rho(t)$ 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 a smooth manifold.



- method II: topological data analysis
  - 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.

% what classes?

- project: classifying pre-1960 climate records and harvesting meteorological data for data assimilation
  - 1. Gather images into (at least) one repository.
  - 2. Establish a common description framework for image metadata.
  - 3. Provide bulk, programmatic access to image subsets.

% only two references % less focus on cumulative efforts

- ambient scientific context
  - builds on: metadata schema for historical climate records
    - Quantitative (e.g. analysis of documents)
      - "Assessing the uptake of persistent identifiers by research infrastructure users" http://n2t.net/ark:/85065/d7q24214
    - Technology/system design & implementation
      - "Building geoscience semantic web applications using established ontologies" http://n2t.net/ark:/85065/d728098t
    - Reporting on/informing professional practice
      - "Modernizing library metadata for historical weather and climate data collections" http://n2t.net/ark:/85065/d71v5hm7
  - builds on: research datasets for computational climate modelling
    - International Comprehensive Ocean-Atmosphere Data Set (Gil Compo, NOAA)
  - builds on: image classification techniques
    - On Characterizing the Capacity of Neural Networks using Algebraic Topology (William Guss, CMU)
  - builds on: computational techniques for time-series analysis
    - Lagrangian Data Assimilation and its Applications to Geophysical Fluid Flows (Laura Silvinki, NOAA)
    - Computational Topology Techniques for Characterizing Time-Series Data (Nikki Sanderson, CU Boulder)
- 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

% include user testing

```
- 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)
```

% how am I contributing to math? % theoretical or methological contributions % differential geometry and statistical? % how to break up the intellectual merit of the pursuit? % 1. intellectual merit in computational topological % 1. statistical analysis % 1. informatics