Project Summary:

Name of project: SensorDat: Real-time sensor testbed for improved provenance and data quality

Project lead and contact details: Mike Daniels, daniels@ucar.edu, (303) 497-8793
Project partners and contact details:
Renée F. Brown, rfbrown@unm.edu
Scotty Strachan, strachan@unr.edu
Matthew Bartos, mdbartos@umich.edu

Proposed start and end date: July 31, 2018 - January 31, 2019

Budget Requested: $7,840-$9,840

Budget Summary:

Cloud-hosting fees: AWS CHORDS instances	Up to 4 AWS EC2 t2.large instances @ $67/mo per instance based on current pricing.	$1340
Initial instrument integration with CHORDS AWS instances	Attach one or more continuous environmental data streams, including from sites associated with the U.S. Long Term Ecological Research (LTER) Network, to ESIP CHORDS instances. This would take 4-6 hours of student development time per instance under the guidance of the project partners. The process is clearly documented via chordsrt.com	$500
QC detection script development	Develop scripts to tap into data streams, develop basic QC check capability (e.g. out of range, spikes, etc.). We estimate this would take ½-1 month of a student's time under the guidance of project partners.	$2000-4000
QC annotation/repair assessment	Develop requirements, and begin to test the performance	$4000

	of more sophisticated anomaly detection techniques (e.g. statistical and machine-learning based approaches). To provide a baseline of requirements to build from, we estimate one month of student time, with guidance from the project partners.	

Project Outline:

Project description: Real-time sensors are increasingly being used for scientific analysis and discovery in earth science research. The Internet of Things (IoT) concept describes an environment in which small, inexpensive sensors become ubiquitous and stream their data to the Internet in real-time. However, sensors used for scientific research purposes require additional sophistication due to issues surrounding standards and metadata requirements, spatial and temporal coverage, data quality considerations, measurement specifications, and geolocation information. To address the IoT as it could be applied to the geosciences, an NSF-funded project called "Cloud-Hosted Real-time Data Services for the Geosciences" (CHORDS, see chordsrt.com) was proposed and funded in 2016 to explore the use of real-time data in a scientific context. Through the work on this project thus far, it has become clear that many of the needs of NSF-funded measurement teams associated with this project are similar, and therefore could be extended to other scientific domains. In particular, through systems like CHORDS, it is now possible to address data quality issues in real-time so that problems are caught quickly, ultimately improving measurement quality. In addition, metadata standards are evolving to help researchers discover streaming data in their area of interest, and to describe features needed for proper interpretation of the data (e.g., units of measurement, spatial and/or temporal coverage). Through the ESIP lab, we would like to 1) extend the use of CHORDS to real-time data streams that are outside of the traditional NSF Geosciences domain, including new varieties of sensors that take advantage of IoT miniaturization, and 2) develop advanced workflows focused on automated data quality and data quality annotation and/or correction.

Project objectives, significance and impact:

- Demonstrate cloud-hosted streaming of new real-time data sources outside of typical NSF Geosciences funded teams. This will demonstrate that there are common aspects of real-time data handling that span broad scientific areas. Through this testbed, we would work to build further the community of envirosensing researchers.
- Develop a testbed of scripts to assess and identify data quality issues in real-time. This will assist in handling data issues as soon as they appear so that the damage to the resulting dataset is mitigated. This testbed will set the stage for community development of robust production workflows for sensor-based science and cloud-hosted data services.
- Develop plans for how we might standardize ways to annotate these data for data quality issues and/or make corrections. Pilot a standard for manual annotations and corrections to

sensor data that will allow applications to return a "view" onto the data without modifying the raw sensor data in place. Lack of annotation schemes is currently a huge gap in practice, so this will bolster the provenance of data streams so that limitations and quality issues in a dataset are clearly documented.

Description of key project steps and timeline:
Summer 2018

- Announce the SensorDat lab testbed at the 2018 ESIP Summer Meeting via the EnviroSensing sessions.

Early Fall 2018

- Establish project GitHub repository.
- Create ESIP AWS instance of CHORDS in the cloud.
- Announce availability via the ESIP newsletter and the EnviroSensing Cluster.
- Attach one or more continuous environmental data streams, including from sites associated with the U.S. LTER Network, to ESIP CHORDS instances.

Mid-Fall 2018

- Develop scripts to tap into data streams, develop basic QC check capability.
- Develop requirements, and/or test the performance of more sophisticated anomaly detection techniques.
- Develop a plan describing requirements for annotation of data quality issues and/or ways to make corrections through a system like CHORDS.

Early Winter 2018

- Work with new data streams as they become available.
- Create a poster for the ESIP Winter Meeting describing lessons learned.

Late Winter 2018

- Code wrap up and documentation.
- Complete final report.

Outreach:

What groups/audiences will be engaged in the project?
The team will engage members of the ESIP EnviroSensing Cluster and the ESIP sensor community at large. If this project is approved, we will announce this at the 2018 ESIP Summer Meeting and through ESIP Newsletters.

How will you judge that project has had impact?
Impact will be assessed in terms of the number of data streams we are able to test in the lab's testbed, and through the quality and extent of the data quality scripts and annotation tools. In addition, we will measure the success of the project by determining the degree to which these tools and techniques are used and developed beyond the period of the incubation project.

How will you share the knowledge generated by the project?
Information about the project will be shared on the ESIP EnviroSensing site and through resulting publications, posters, and presentations given relating to the project. An ESIP GitHub repository will be created to store all project documentation and artifacts.

Project Partners:

Description of project partners (individuals and/or organizations) and their involvement:
Mike Daniels, NCAR, will lead the project and coordinate the activities. He will also work with NCAR software engineers to help deploy and support CHORDS instances.
Renée F. Brown, UNM and co-Chair of the ESIP EnviroSensing Cluster, will lead the effort to integrate streaming data from the McMurdo Dry Valleys and/or Sevilleta LTER sites. Her team will work to apply QC algorithms in the cloud through CHORDS.
Scotty Strachan, UNR and co-Chair of the ESIP EnviroSensing Cluster, will lead the effort to integrate streaming data from NevCAN (Nevada Climate-ecohydrology Assessment Network) sites. His team will work to apply QC algorithms in the cloud through CHORDS.
Matthew Bartos, UMich, will help to integrate streaming data and work to develop data quality annotations. Matt will also research the applicability of modern signal processing tools for real-time anomaly detection (including probabilistic methods and machine learning-based techniques like decision tree forests and support vector machines).

All of the partners will help to promote the lab testbed within the context of the ESIP and the ESIP EnviroSensing cluster.

How will this project engage members of the ESIP community?
The team will promote and engage the ESIP community through the ESIP EnviroSensing Cluster, and will promote participation at the ESIP summer and winter meetings.