

2.11 Statistics and data practices (SDP)

Data, statistical models and inferential procedures permeate CENS deployments, from the four founding scientific application areas to the more recent urban sensing campaigns and “n-of-one” personal data collection exercises. Rather than attempt to synthesize that work here, we will instead focus on three classes of activities not well represented elsewhere: 1) General statistical models for embedded sensing, with specific applications to data quality and continuous sampling, 2) Significant CENS-designed and supported databases and repositories, and 3) Studies into the data lifecycle for embedded sensing systems. On their own, these three tell a story of data sharing and reuse, of application-driven system design, and of the multiple levels of analysis that take place simultaneously in an interdisciplinary center like CENS.

General statistical models: Applications to data quality and continuous sampling

Fault and anomaly detection

This year, our fault detection work proceeded on two fronts. In the first we considered various model-based approaches that flag as anomalous measurements that differ significantly from expectations. We report on two in our project summaries: A multiplicative seasonal ARIMA time series model, and a (Gaussian) Bayesian model designed to quantify “surprise” in one or more sensor readings. The second broad direction for our fault detection work is non-parametric in nature, depending on a set of well-chosen data features that are combined in a flexible “signature” that describes normal operation of a sensor or a cluster of sensors. Signatures are also constructed for various fault types, and, as new data are collected, these are compared to identify likely anomalous measurements. Both the model- and signature-based methods share common themes and technical challenges: Both depend on the underlying network topology for setting realistic constraints on data sharing between system components; both rely on local updates that allow the system to learn new, non-faulty behavior; both require well-documented deployment data from several application areas for training and experimentation; and both begin with some characterization of the fault process. In connection with this last point, we are pleased to report that our taxonomy of fault types, “Sensor Network Data Fault Types” will appear in the 2009 Proceedings of the ACM Transactions on Sensor Networks.

A new sampling paradigm

In applications involving mobile sensors, we have often carried over many notions that are rooted in static sensing. As a result, we often employ mobile nodes in a way that mimics static data collection: We move a mobile sensor to a grid point where a static sensor might have been, let it dwell in that location for some period of time until a sample is taken, and then move the node to another point. By contrast, over the last year, we have developed sampling schemes for situations in which we can sample (near) continuous paths using a single mobile node. Our test problem involves modeling dynamic light patterns in a forest understory. In addition to adapting many field estimation schemes to work with highly structured input data, we are developing structural models based on polygonal random fields.

Significant CENS-developed Databases and Repositories

The previous studies would not have been possible without a stable base of well-documented data, from both short-term experiments and as well as long-term stable deployments. The development of effective fault detection methods, for example, requires detailed information from a variety of deployments. For each we ask: What equipment was involved? Were there known issues with certain observations or ranges of observations? Are other measurements available from nearby devices or from other instrumentation that could provide useful context? In this section we document three CENS-developed databases and repositories that provide uniform access to CENS data and research products. These projects have begun to converge, enabling CENS data, deployment records, and publications to be linked together.

CENS Deployment Center (CENSDC)

The CENS Deployment Center (CENSDC) is a database for researchers to document their deployment activities through the creation of pre-deployment plans and post-deployment feedback. CENSDC encourages researchers to

capture their experiences both in and out of the field to help inform future deployments. In so doing, we cull tacit information about equipment setups, deployment locations and field preparations, all of which play a critical role in data collection techniques. CENSDC designers have worked closely with CENS researchers as they plan for and ultimately execute detailed sensing experiments. Research is underway on a handheld mobile device for data collection in the field, from which data will flow into CENSDC and Sensorbase.

Sensorbase

For several years now, Sensorbase has provided a database backbone infrastructure for CENS. It provides a means to store all sensor data gathered in the field in a user-friendly environment (with a web interface for interactive use and a collection of APIs for programmatic access via SOAP and RESTful services). Sensorbase is built on a set of data permissions and privacy constraints that allow users to easily manage how their data are shared with others. This year, we completed a major code integration effort that ended with the creation of a central repository for distributing Sensorbase source code. In addition, we migrated our operation to a new, vastly more powerful database server and made public a host of new services for data upload and access.

Studying the data lifecycle: Watching the watchers

Finally, we present systematic studies of how data are collected, stored, shared and reused within CENS. This work takes place in a larger context of research into the data lifecycle and the design of effective cyberinfrastructure to support interdisciplinary collaborations.

Towards a Virtual Organization for Data Cyberinfrastructure

As part of our overarching research agenda to understand CENS as a collaboratory and the cyberinfrastructure (CI) necessary to support such a collaboratory, we established a CI Virtual Observatory for the study of data, data analysis, and visualization. In this one-year Small Grant for Exploratory Research, we developed a research instrument and sampling method to probe, in a consistent fashion, several CI projects, including CENS. The instrument is being used by collaborators at several institutions, each of which obtained IRB approval. Ultimately, we see this work providing guidance as to what constitutes a viable cyberinfrastructure.

Object Reuse and Exchange; RDF data modeling

This project concerns the design and development of tools that allow efficient reuse and exchange of information objects resulting from embedded sensing research applications. We are experimenting with the Open Archive Initiative's Object Reuse and Exchange protocol (OAI-ORE) to describe, publish and share aggregations of information objects produced at different stages of the scientific lifecycle in environmental sensing research. We are developing a sensor-specific vocabulary to describe the relationships among these information objects, using the Resource Description Framework data model to capture a) information about deployments (stored in the CENS Deployment Center), b) sensor data (stored in Sensorbase.org) and c) scientific publications (stored in the California Digital Library's eScholarship repository). Although these data are organized and managed as separate entities in disjointed data archives, we view them as building blocks of the same scholarly production chain. Our present work is aimed at weaving together these resources using the OAI-ORE data model.