Overview of Statistics & Data Practices

Dave Fearon, Matt Mayernik, Keith Mayoral, Sheela Nair, Andrew Parker, Alberto Pepe, Erick Romero, Abhishek Sharma, Jillian Wallis, Yuan Yao, Richard Guy, Mike Taggart, Christine Borgman, Deborah Estrin, Leana Golubchik, Ramesh Govindan, & Mark Hansen

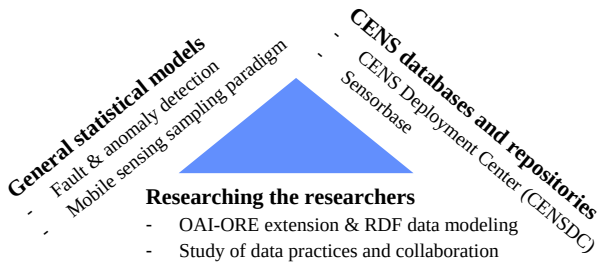
Introduction: Now that we have data where do we go from here?

New Focus on CENS Data

- **Maturing of sensing instrumentation and systems at CENS**

  For so long the question was always "are we getting any data" from our systems and instruments, but now we are able to ask "are we getting good data?" With this shift from quantity to quality the research emphasis has shifted from data collection technologies and systems to ensuring data cleanliness and dissemination.
- **New problems means new approaches**
  - Clean data on the way in through the use of fault detection algorithms or new adaptive sampling algorithms to catch the interesting data
  - Access points for data discovery and interpretation
  - Enhanced description as surrogates for other trust establishing processes

Three Research Approach Areas



Problem Description: The fine line between data and noise

- **Novel instruments with novel needs**

  Mobile sensing research encounters different hurdles than static and remote sensing network research. These needs have emerged from the repeated deployment and testing of equipment and systems.
- **Developing standards of best practice for mobile sensing networks**

  As a Center we are the first of our kind to explore these issues, thus we have the opportunity and responsibility to set the standards of best practice for use of these technologies. The diversity present within the Center allows for opportunities to push the limits of mobile sensing on a regular basis. Developing and extending standards from tinyOS to ORE, as well as participation in these communties is extremely vital.
- **Bridging between multi-discipline research cultures**

  Making the research interesting and valuable to all of the CENS collaborators is not a trivial matter. Each researcher needs to balance their own research and professional interests with the needs of the larger group, and includes a certain measure of sacrifice to enjoy the benefits of collaboration.
- **Coordinating multi-institution research activities**

  Coordinating research programs across a large distances requires awareness, organization, and communication both of research findings and institutional memory.
- **Supporting data reuse and sharing within and beyond CENS**

  CENS is an investment of funding and researcher time, it would be a waste to not be able to reuse CENS collected data.

Proposed Solution: A three-pronged approach

Statistical Models (poster 2)

- **Learning from our own frustrations**

  All of our field deployments and testing have lead to a focus on identifying faulty equipment and data in the field, so that these faults can be mitigated. The result of this is the development of a compilation of fault types and increasingly more sophisticated algorithms for the identification of faults both in the field and in datasets which can be applied beyond CENS.
- **Scarcity of resources requires agile thinking**

  There are never enough sensors or people to deploy them to really capture everything, rather than focusing on this as an intractable problem the focus has shifted to utilizing the "mobile" part of mobile sensing, for reconfiguring networks in real-time. Smarter algorithms for deployment lead to higher quality data collection.

Bespoke CENS Systems (poster 3)

- **Looking for off-the-shelf solutions or existing standards**

  This is the best way to ensure interoperability with other projects and not waste effort reinventing the wheel. Unfortunately this search did not always yield an existing solution to our problems.
- **Iterative improvement from user feedback**

  Sensorbase and CENSDC have been up and running for at least 3 and 2 years respectively, providing ample opportunity for the systems to be used, feedback provided, and improvements integrated.

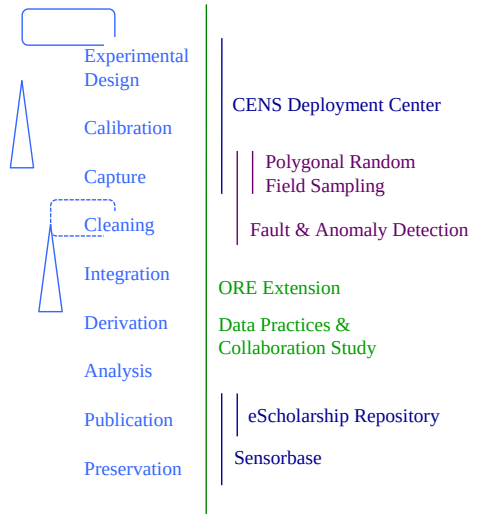
Social Studies of Scientists

- **Understanding the process to improve the process**

  Using interviews of CENS researchers and participant observation, we have developed a lifecycle of CENS data. Understanding of this lifecycle has allowed us to identify cross-disciplinary cyberinfrastructure needs to support data reuse and sharing and propose systems, such as the CENSDC to fill these holes.
- **Aggregation of research products throughout lifecycle for more robust data annotation**

  Throughout the data lifecycle products are thrown off: deployment plan, versions of datasets, publications, etc. By linking these items together it makes data access and reuse more feasible in the long-term. Extension of metadata standards such as OAI Object Reuse and Exchange, would allow for these products to be aggregated.

CENS Data Lifecycle - S&DP Research Projects



Ex. ORE Aggregated Resource

How ORE could aggregate products during the three phases of production in the research lifecycle.