Title

High Fidelity Data Collection: Managing the Collection Process Throughout the Deployment Lifecycle (SYS 19)

Authors

Laura Balzano
Nithya Ramanathan
Tom Sc hoellhammer
et al.

Publication Date

2006

High Fidelity Data Collection Throughout the Deployment Lifecycle

Laura Balzano, Nithya Ramanathan, Tom Schoellhammer, Deborah Estrin, Eddie Kohler, Mani Srivastava

What's the Problem?

SensorNet Data Travels a Protracted and Perilous Path!

- **(See Diagram Below!) Sensing channel … through the sensor … collected by the board … transmitted across an ad-hoc wireless network … collected at a base-station … uploaded to a server … cleaned … uploaded to a data-base … end-user**

System Pitfalls and Shortcomings Impact Collection and Delivery of Data

- **E.g. bad sensor coupling, faulty communication channel**
- **Often, user intervention required to identify and fix such problems, resulting in drawn out deployment cycles, during which the impacts of faults accumulate as unusable data**
- **Often an end user cannot definitely determine if data are usable without better understanding the context in which they were collected**

Traditional Approach Focuses on Post-Deployment Operations

- **Manual/visual cleaning of data and time-stamps**
- **Apply physical/statistical models to identify outliers**
- **Guesswork as to environmental context… e.g. "Did wind-blown grass or a truck trigger this motion sensor?"**

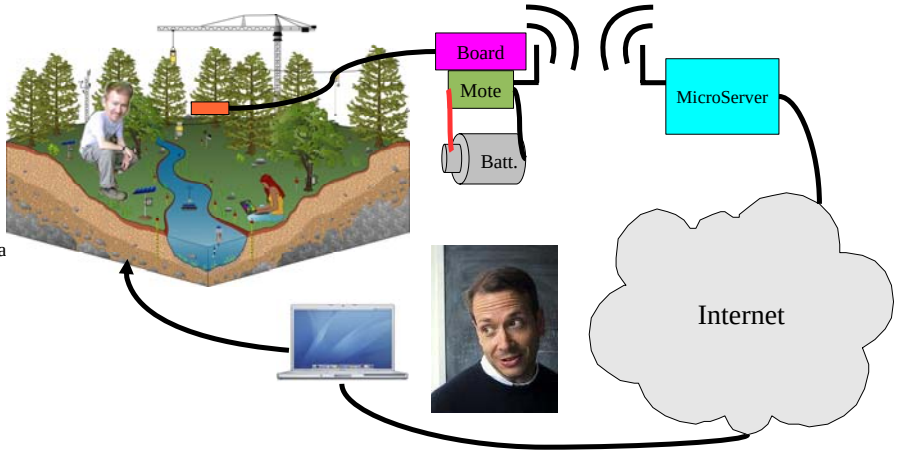
Carefully integrating human input into the collection path is more useful than completely autonomously collected data!!

The Puzzle Comes Together

Design Systems Incorporating Human Input into Collection Path

Design systems to work with humans to:

- **Diagnose and/or fix** problems impacting the *generation and delivery* of high quality data to the end-user
- **Document the context** by gathering extra sensory *observations, such as physical samples or human observations,* that provide contextual information useful for determining data integrity and easing data analysis



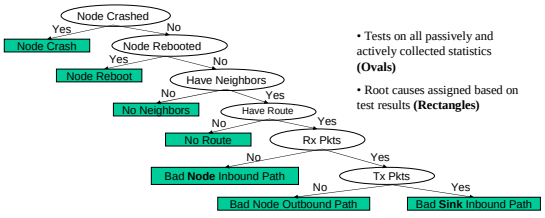
Network Integrity: Sympathy as a Sensor Network Debugger

Sympathy is designed to identify faults impacting the *delivery* of data (i.e. network faults) by:

- **Collecting Metrics** from every node such as *neighbor table and number of packets received*
- **Identifying Network Failures** using a *decision tree* (right) to find the root cause
- **Localize Failures** to identify *where data is lost*

Sympathy then suggests **actions** associated with the fault **that a user can take in the field** to fix these network bugs

- Actions include: replacing the battery, moving a node, or tracking a software bug



Confidence for Data Quality Management

Confidence is designed to identify **actions a user can take in the field** to remediate faults impacting the *generation* of data (i.e. data quality faults) by

- **Classifying data** based on 4 pre-specified attributes (e.g. gradient, standard deviation)
- Use *instance-based learning* techniques to identify data that most closely resembles current data point using euclidean distance in the attribute space
- Suggestion actions associated with closest resembling data points
  - -- Actions include: taking a physical sample, or checking sensor connections

Deployment Buddy (Working Title)

Span the networked sensing deployment and analysis life cycle

- Explore *experimental design* strategies prior to deployment
  - •Given a phenomenon model, what design seems best for parameter estimation?
- Capture deployment metadata and irregularities during deployment
  - •Decisions made in the field impact data analysis
- Facilitate data analysis, using phenomenon specific models, and recorded installation metadata