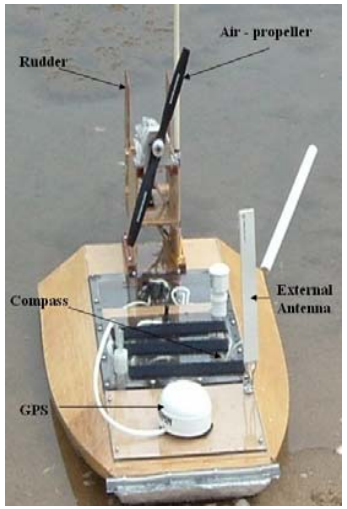
The AMBROSia Project

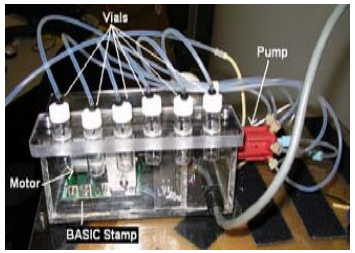
Data-Driven Sensing and Fault Modeling

Ramesh Govindan
ramesh@usc.edu

David Caron, Abhimanyu Das, Amit Dhariwal, **Leana Golubchik**, David Kempe, Carl Oberg, Abhishek Sharma, Beth Stauffer, Gaurav Sukhatme, Bin Zhang

Motivating Application





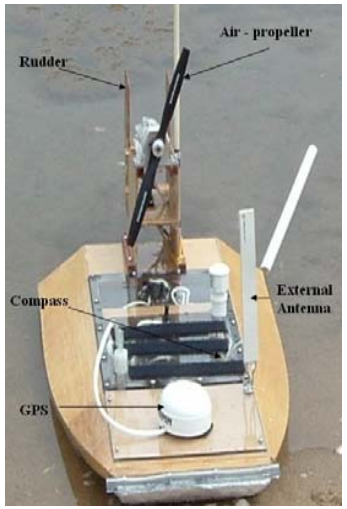
Map and Sample Hydrographic Features

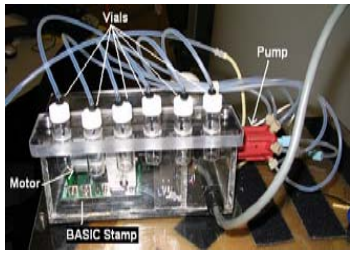
Public Policy Implications: Safety of Water Sources

Current methods provide sparse sampling, labor-intensive

Microbial Observing and Sampling in Freshwater and Marine Ecosystems

Features





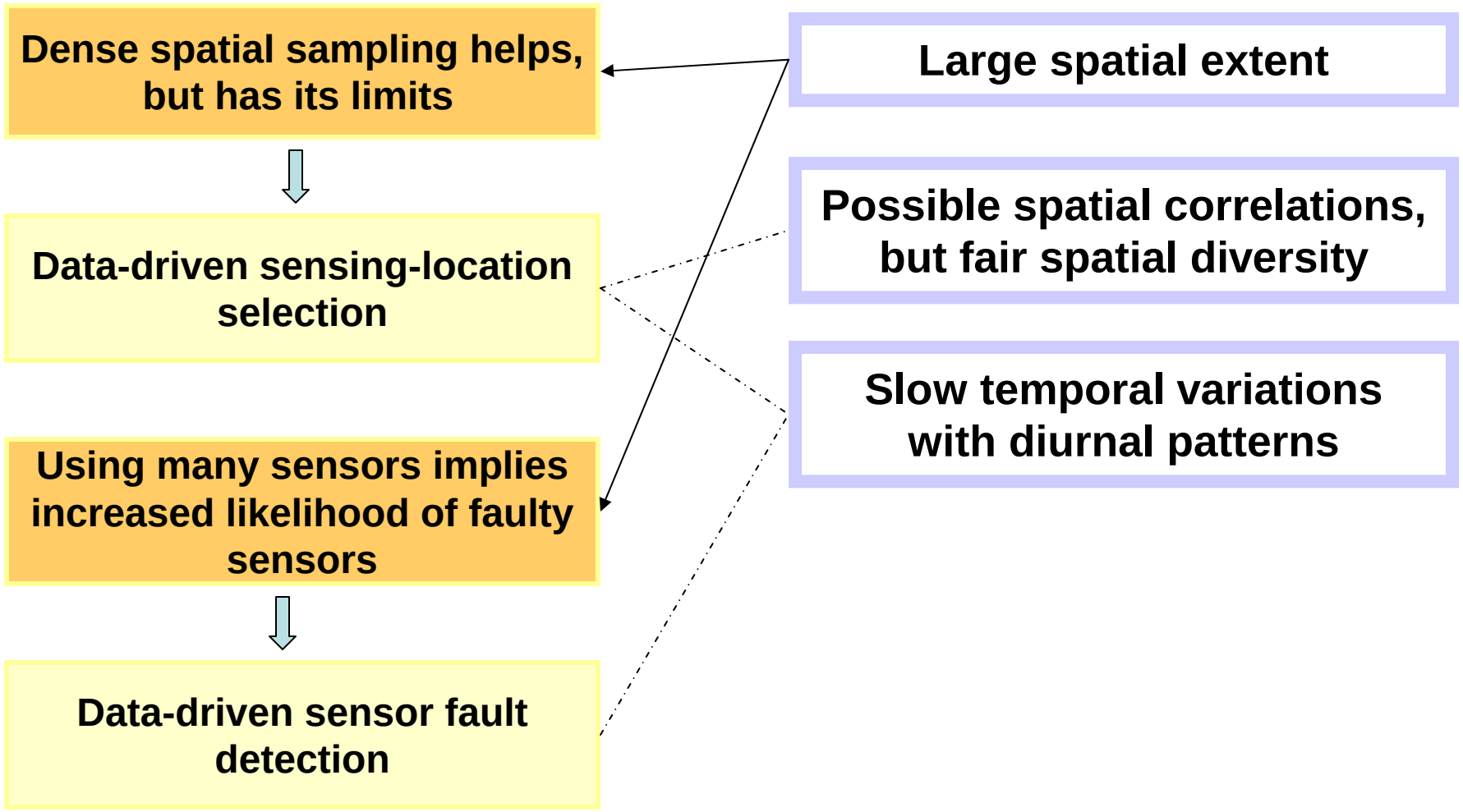
Large spatial extent

Possible spatial correlations, but fair spatial diversity

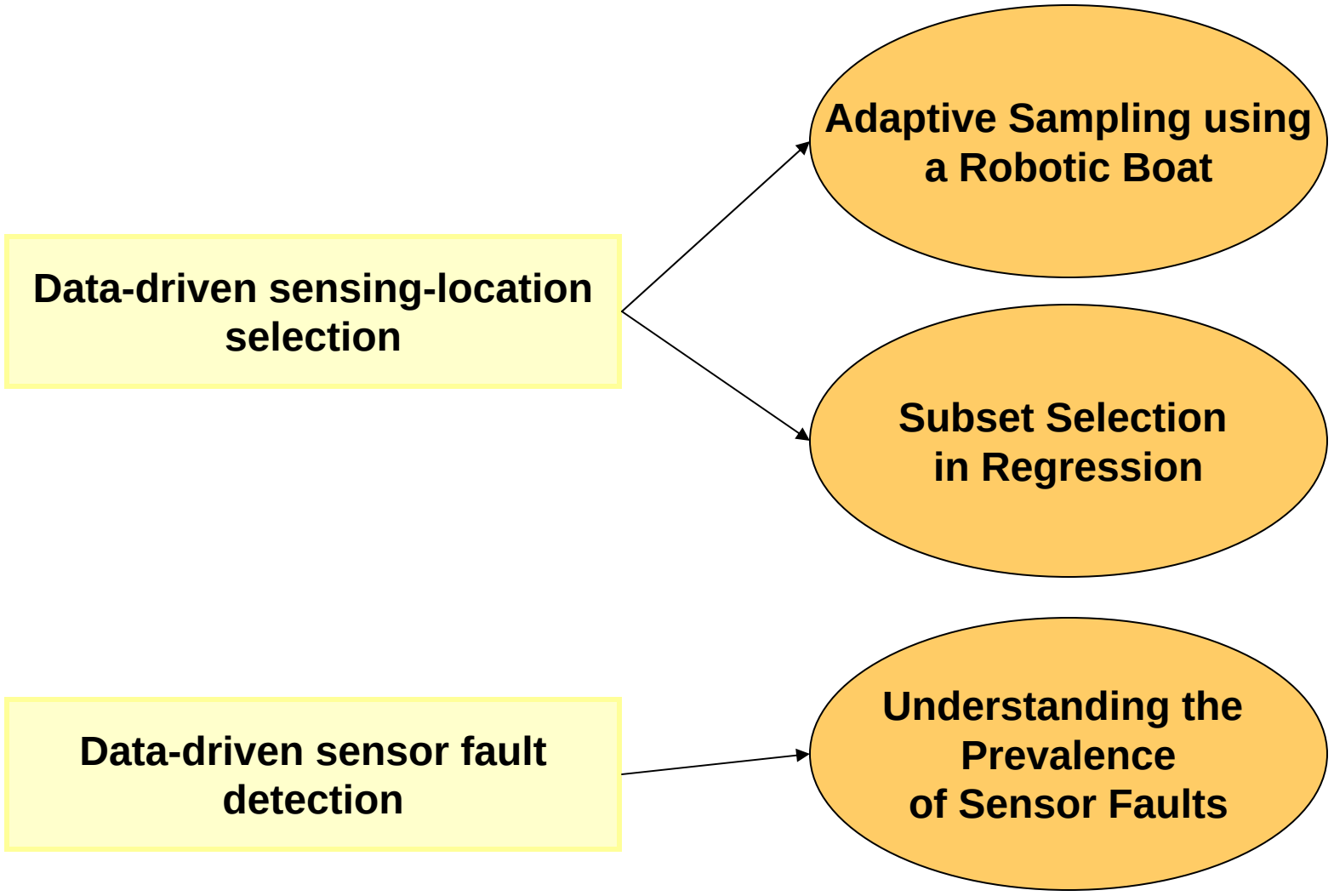
Slow temporal variations with diurnal patterns

Microbial Observing and Sampling in Freshwater and Marine Ecosystems

Dense Sensing

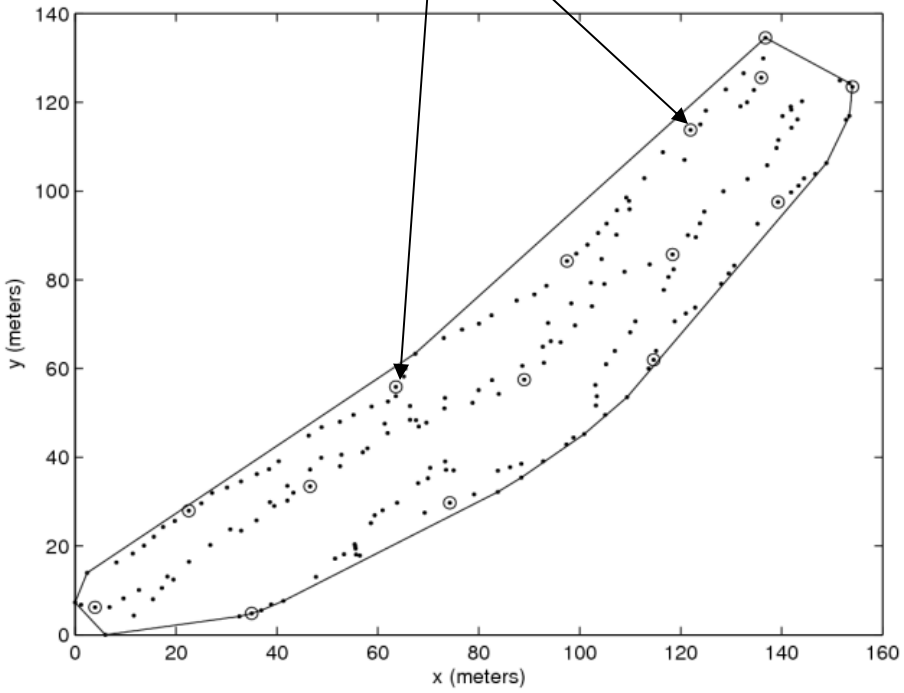


Talk Outline



Adaptive Sampling

Static Sensor Locations



Lake Fulmor

If each static sensor makes a measurement in its vicinity, and the total energy available to a mobile robot is known, what path should the mobile robot take to minimize the Iterative mean-squared error associated with the reconstruction of the entire field?

Assume robot knows readings from static sensors

Using a Regression Model

$$Y_i = m(X_i) + v^{1/2}(X_i)\varepsilon_i$$

Non-parametric model for scalar field $m(X_i)$

Minimize:

$$\sum_{i=1}^{n}\{Y_i - \alpha - \beta^T(X_i - x)K_H(X_i - x)\}^2$$

$$K_H = |H|^{1/2}\, K(H^{1/2}u)$$

Weighted Local Linear Regression

IMSE(x_1,x_2,…x_n)

Closed form for mean-squared error from using these points to reconstruct field

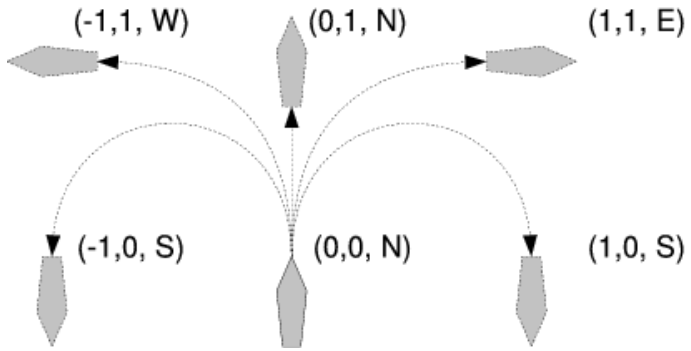
Computing the Boat's Path

Path defined as a sequence of locations $P = (x_1, x_2, ..., x_m)$

The *gain* of location x is defined as
$G(x) = IMSE(x_1, ..., x_n) - IMSE(x_1, ..., x_n, x)$

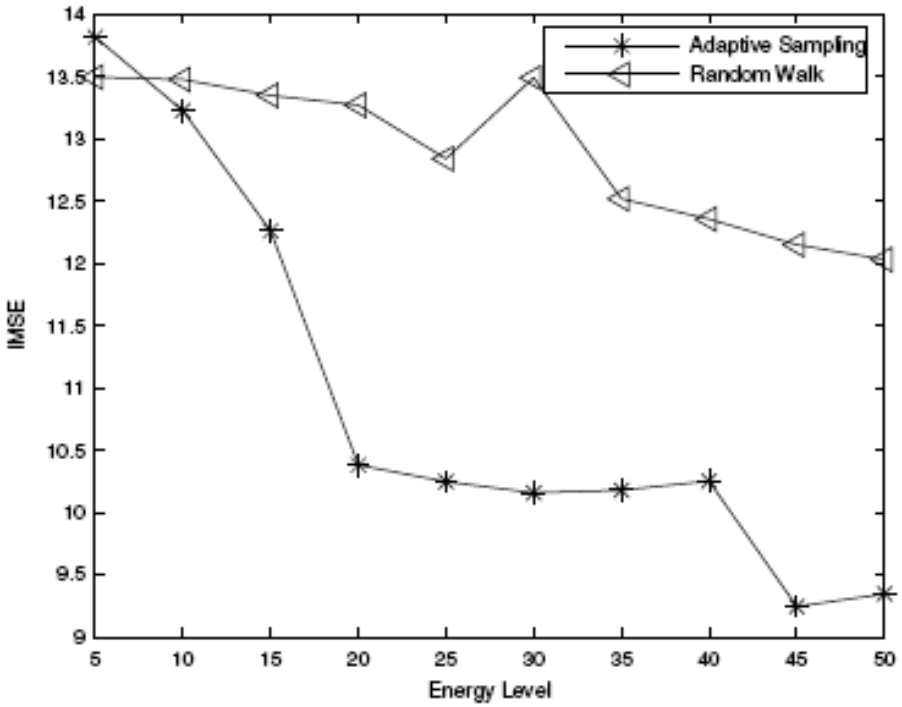
The gain of a path P is defined as
$G(P) = IMSE(x_1, ..., x_n, P) - IMSE(x_1, ..., x_n)$

Find highest-gain path using a simple BFS strategy taking robot's energy into account

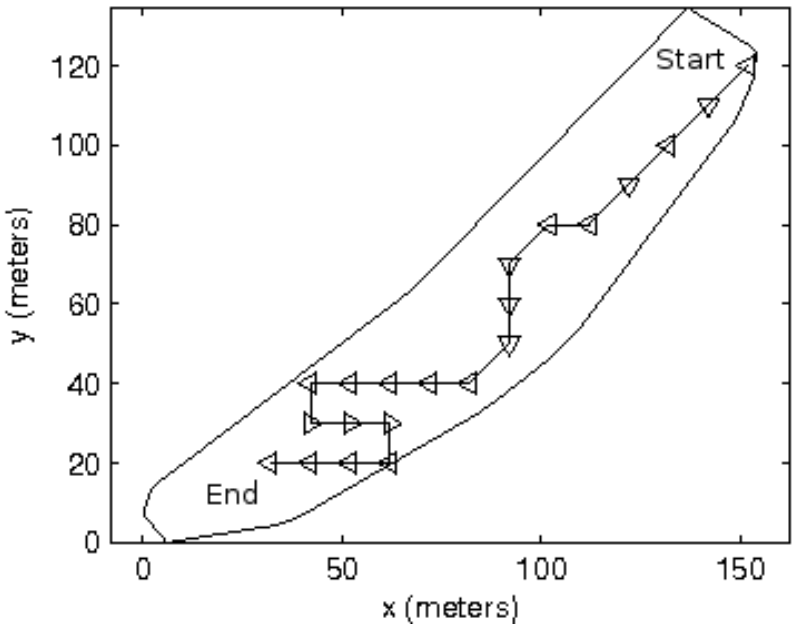


Energy consumption is a function of maneuver

How Well Does it Work?



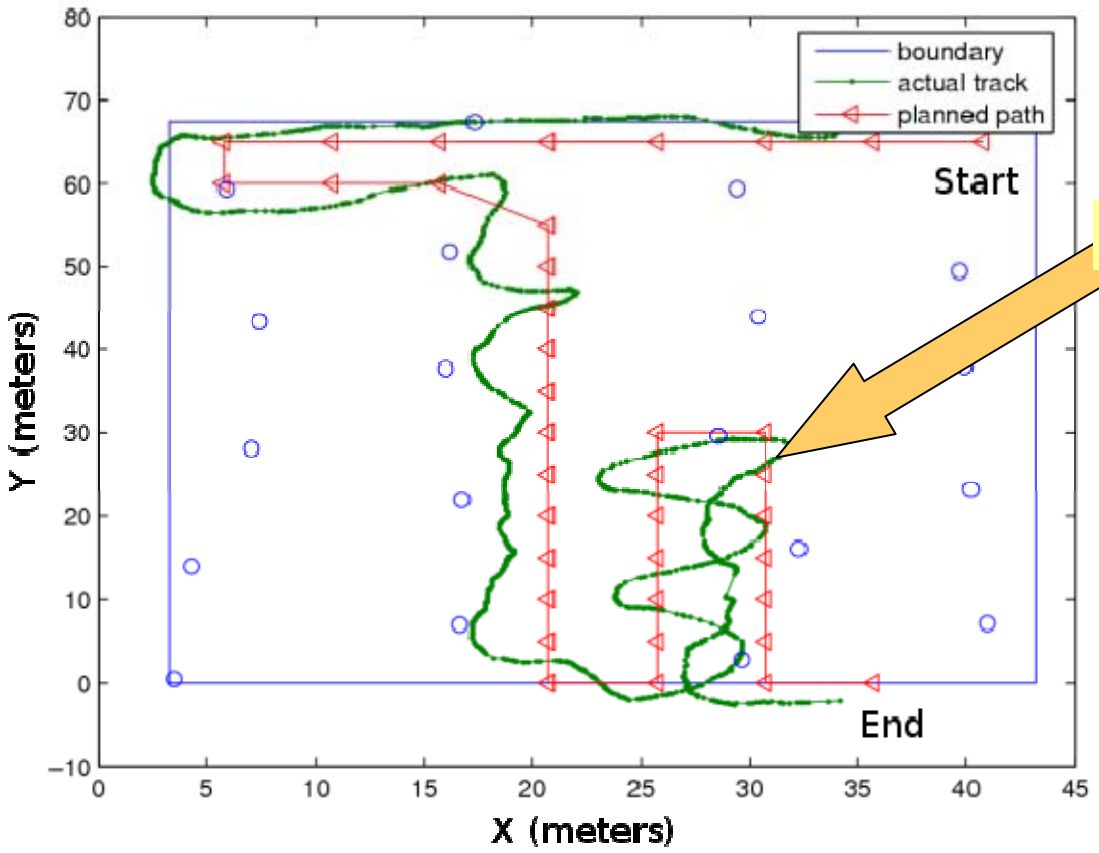
Lower reconstruction error than random walk



Computed path

Simulation using real data

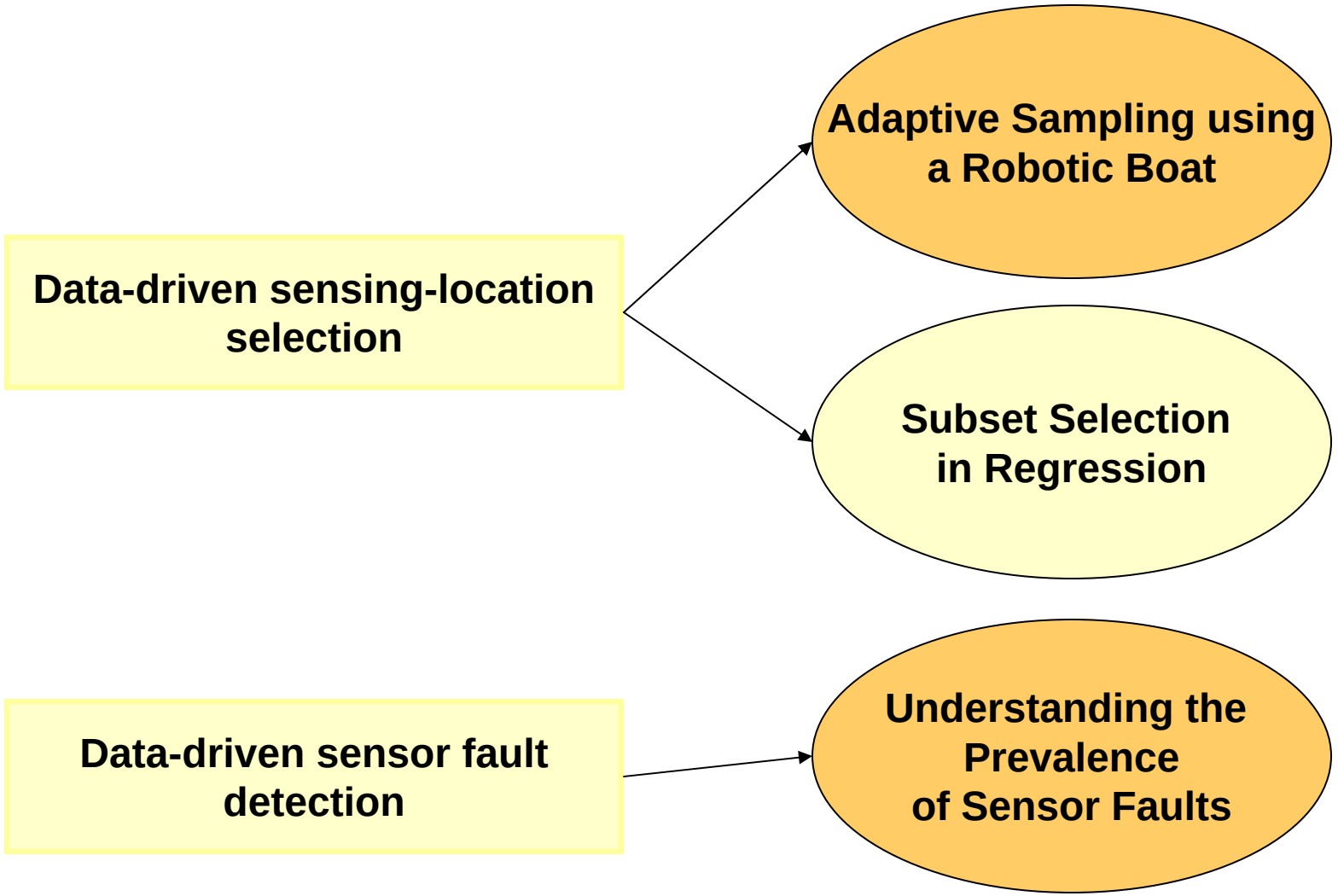
How Well Does it Work?



Drift due to wind!

Actual boat track in real experiment!

Talk Outline



Subset Selection

Assume the correlations between sensors can be learned

Identify smallest set of sensor readings to retrieve that yields maximum information about variable of interest

Examples of variables: average/max temperature, light, chlorophyll concentrations

Formalization

- Random variables to sample: X_i
- Goal: estimate random variable Z
  - Example: average or max temperature
- Given: (n+1) * (n+1) covariance matrix
  - Between X_is and Z, from past data

Select size-*k* subset of X_i that minimizes mean square regression error

Problem Characteristics

Equivalent to sparse signal approximation over dictionaries

Known to be NP-hard, hard to approximate

We focus on tractable, relevant special cases

Our Results

If the X_i are "nearly independent", standard greedy heuristics give good approximation bounds

Improves recent results by Gilbert *et al.*

If the covariance matrix is "structured", dynamic programming with rounding gives good bounds

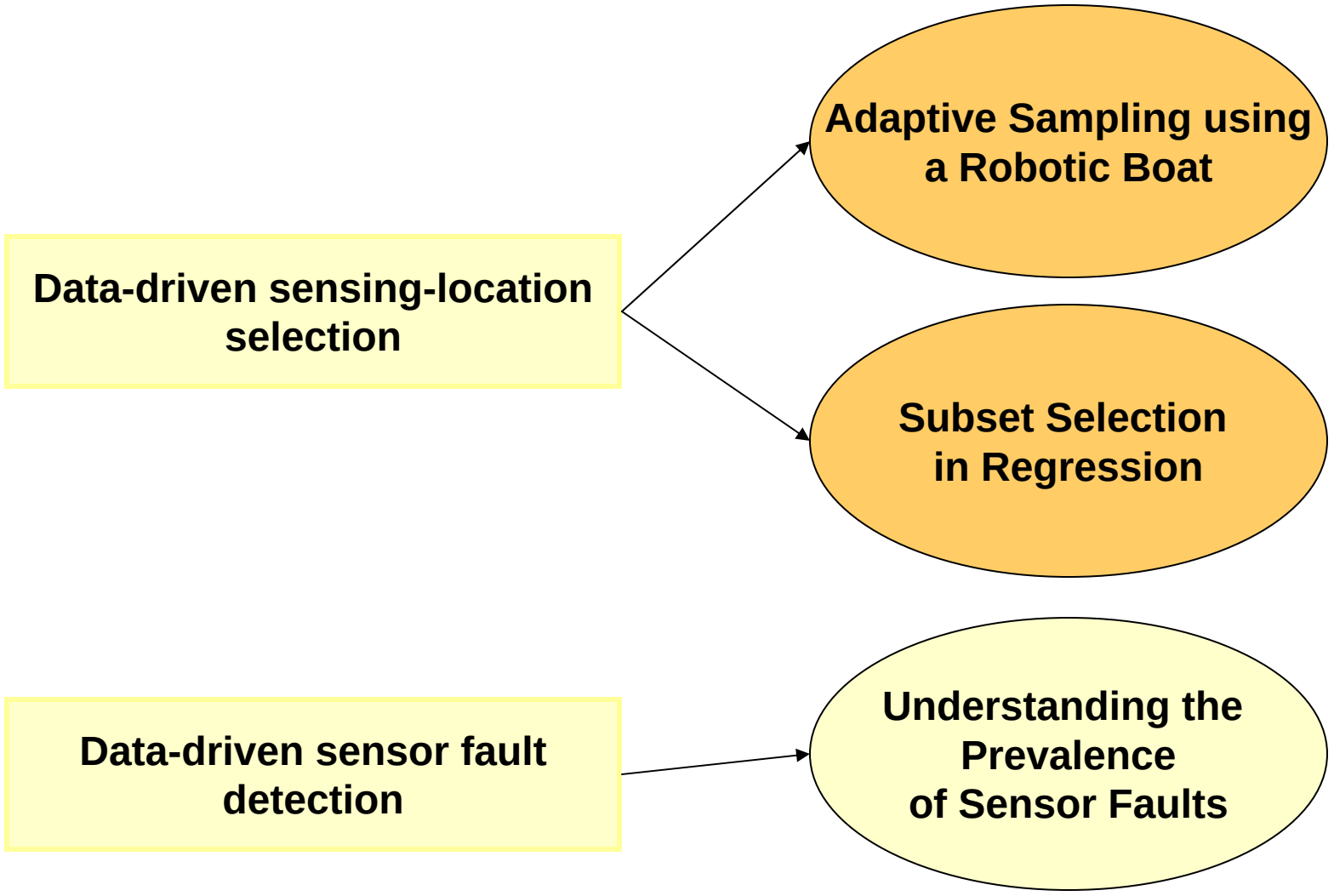
Analysis uses sophisticated matrix perturbation techniques

Ongoing Work: Other Cases

X_i are embedded in a metric, and their spatial correlations decrease with distance

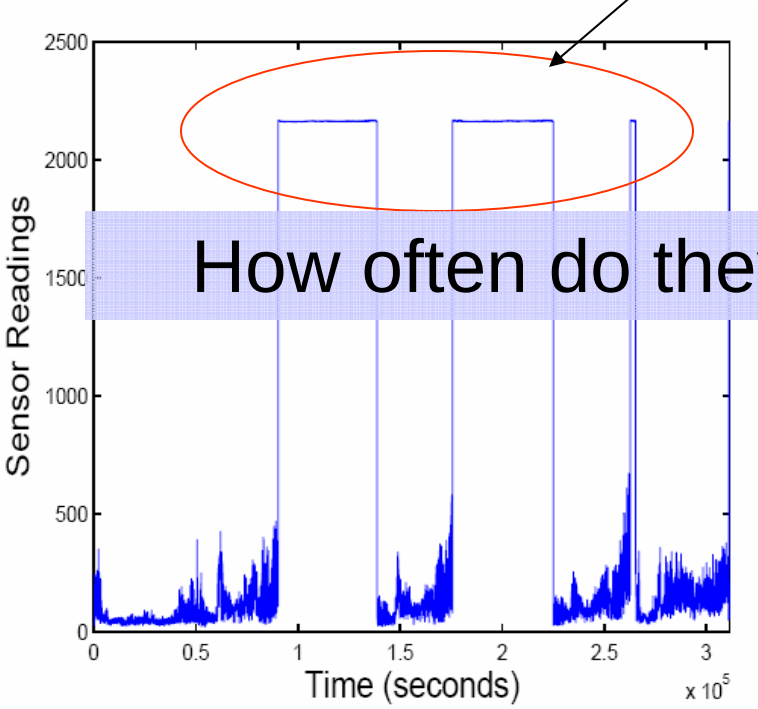
X_i are chosen by an adversary, but are constrained by a distance function

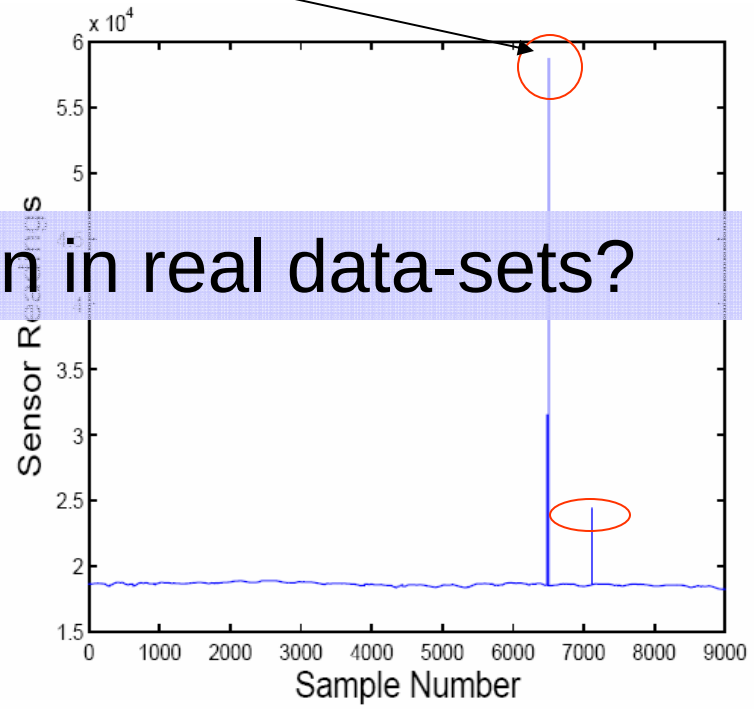
Talk Outline



Data from Real Deployments

Transient Faulty Sensor Readings!





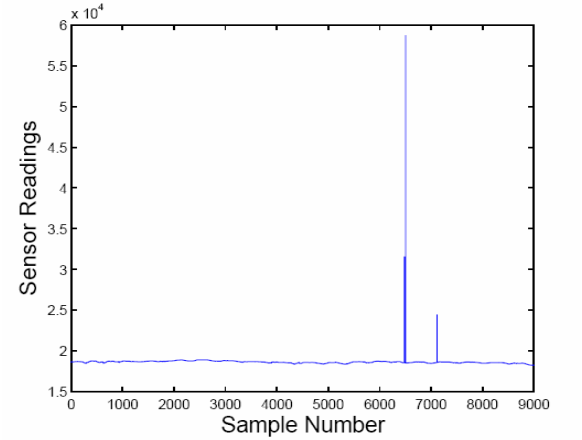
How often do they happen in real data-sets?

Lake Water Chlorophyll concentration, NAMOS deployment

Pressure sensor readings, Great Duck Island

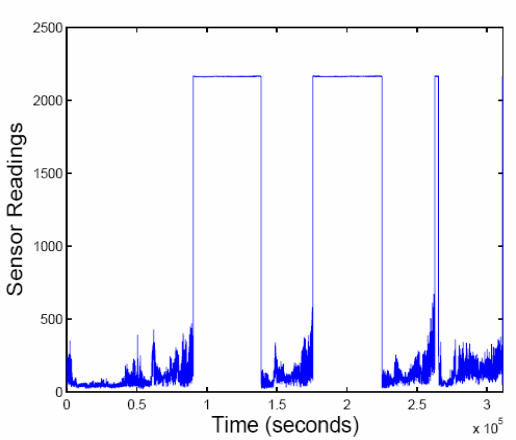
Taxonomy of Faults

Single sample spikes



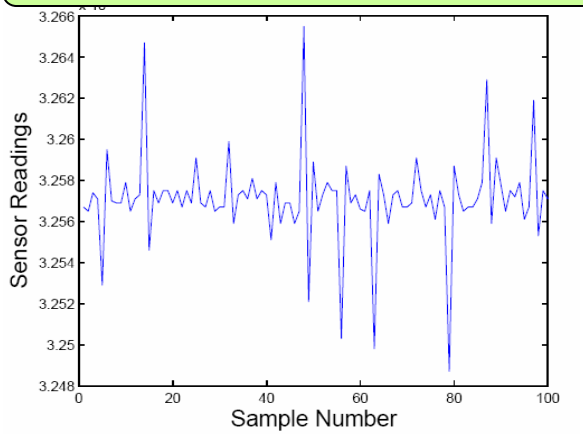
SHORT

Same value for N (contiguous in time) samples



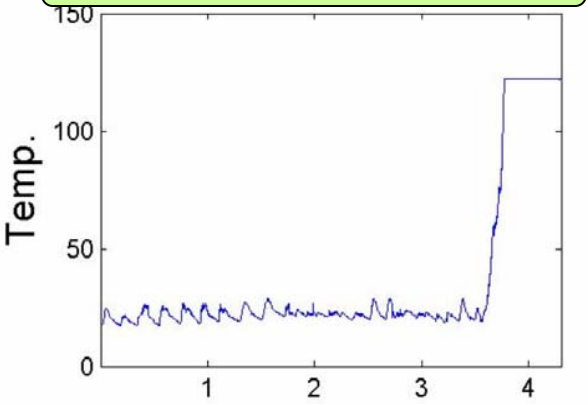
CONSTANT (Stuck at)

Sample variance increases



NOISE

Combination of faults



NOISE + CONSTANT

Fault Detection Methods

Heuristic Rule-Based Methods

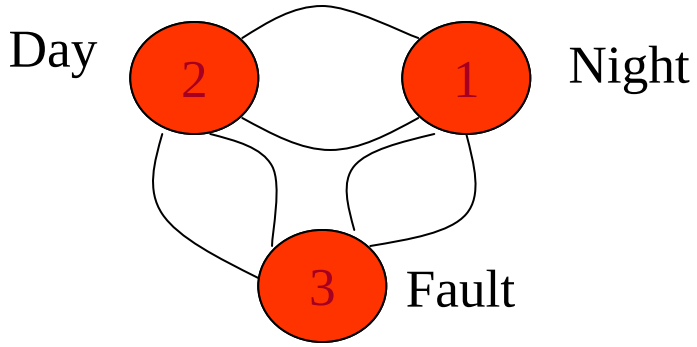
$$\frac{value(t) - value(t_0)}{t - t_0} > threshold$$

Qualitatively different methods give confidence in fault detection results

Linear Least-Squares Based Method

$$f_1(t_2) = m_{t_1} + \frac{\lambda_{t_1 t_2}}{\lambda_{t_2}}(t_2 - m_{t_2})$$

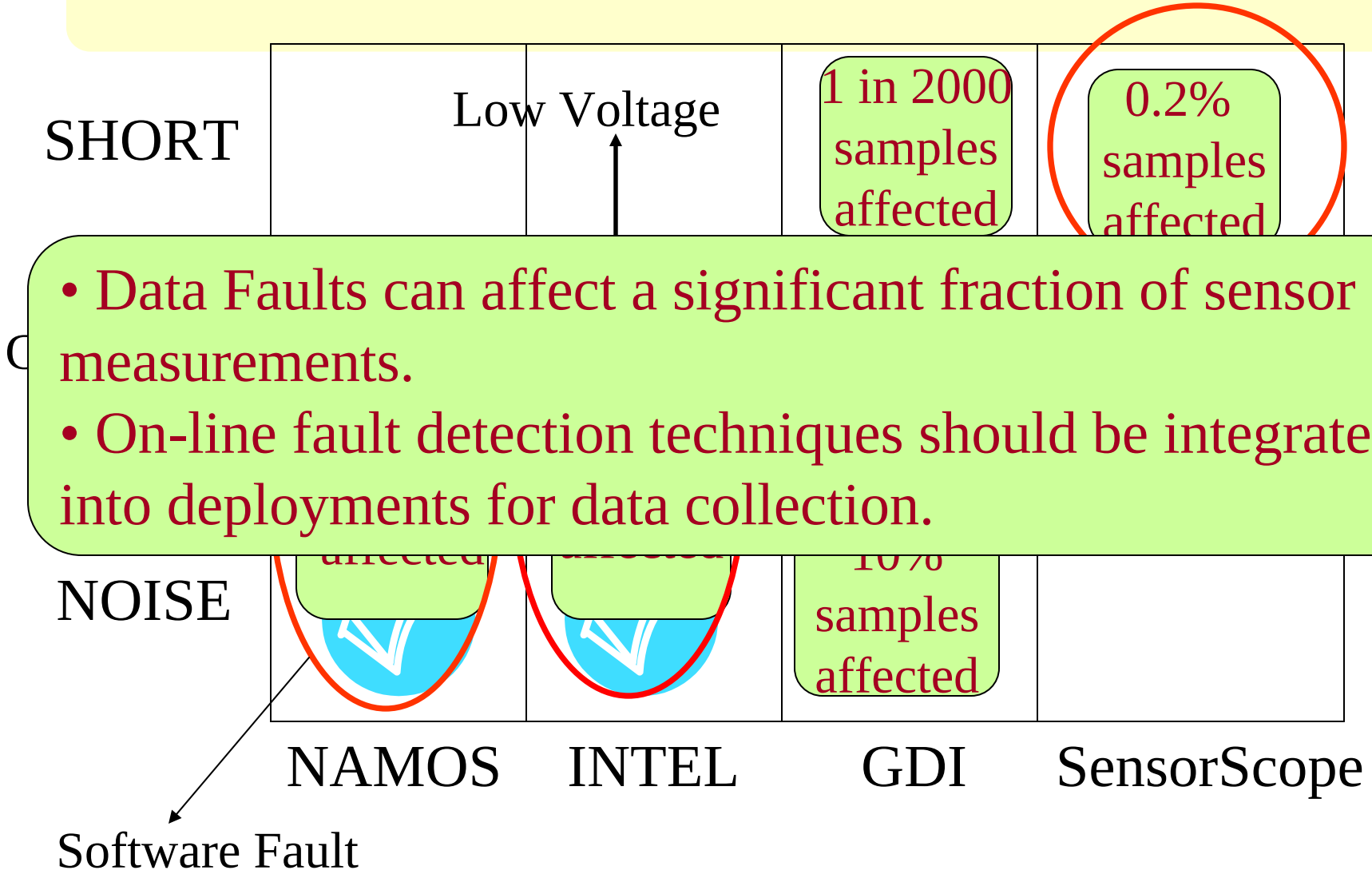
HMM-Based Method



Real world data sets

Data set	# Sensor Nodes	Phenomena sensed	Duration/ Sampling	Year
NAMOS (Lake water)	9 buoys	Chlorophyll, Temp.	24 hours/ 8-10 sec.	2006
Great Duck Island	15	Light, Temp. Pressure, Humidity	3 months/ 5 min.	2002
Intel Lab, Berkeley	54	Temp., Humidity, Light	1 month/ 30 sec.	2004
SensorScope	31	Temp., Humidity, Light, Wind speed…	Ongoing/ 30 sec.	2006-07

Prevalence of Data Faults

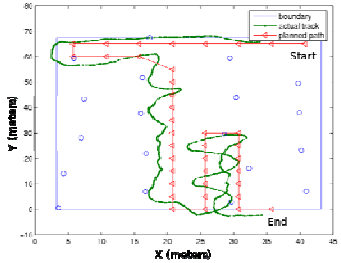


- Data Faults can affect a significant fraction of sensor measurements.
- On-line fault detection techniques should be integrated into deployments for data collection.

Summary

Data-driven sensing-location selection

Adaptive Sampling using a Robotic Boat



Subset Selection in Regression

Greedy Heuristic

Dynamic Programming

Data-driven sensor fault detection

Understanding the Prevalence of Sensor Faults