Making Sense of Suppressions and Failures in Sensor Data: A Bayesian Approach

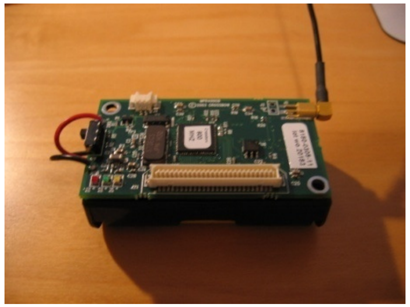
Adam Silberstein
Yahoo! Research

Jun Yang, Kamesh Munagala
Duke CS

Gavino Puggiono, Alan Gelfand
Duke ISDS

Introduction

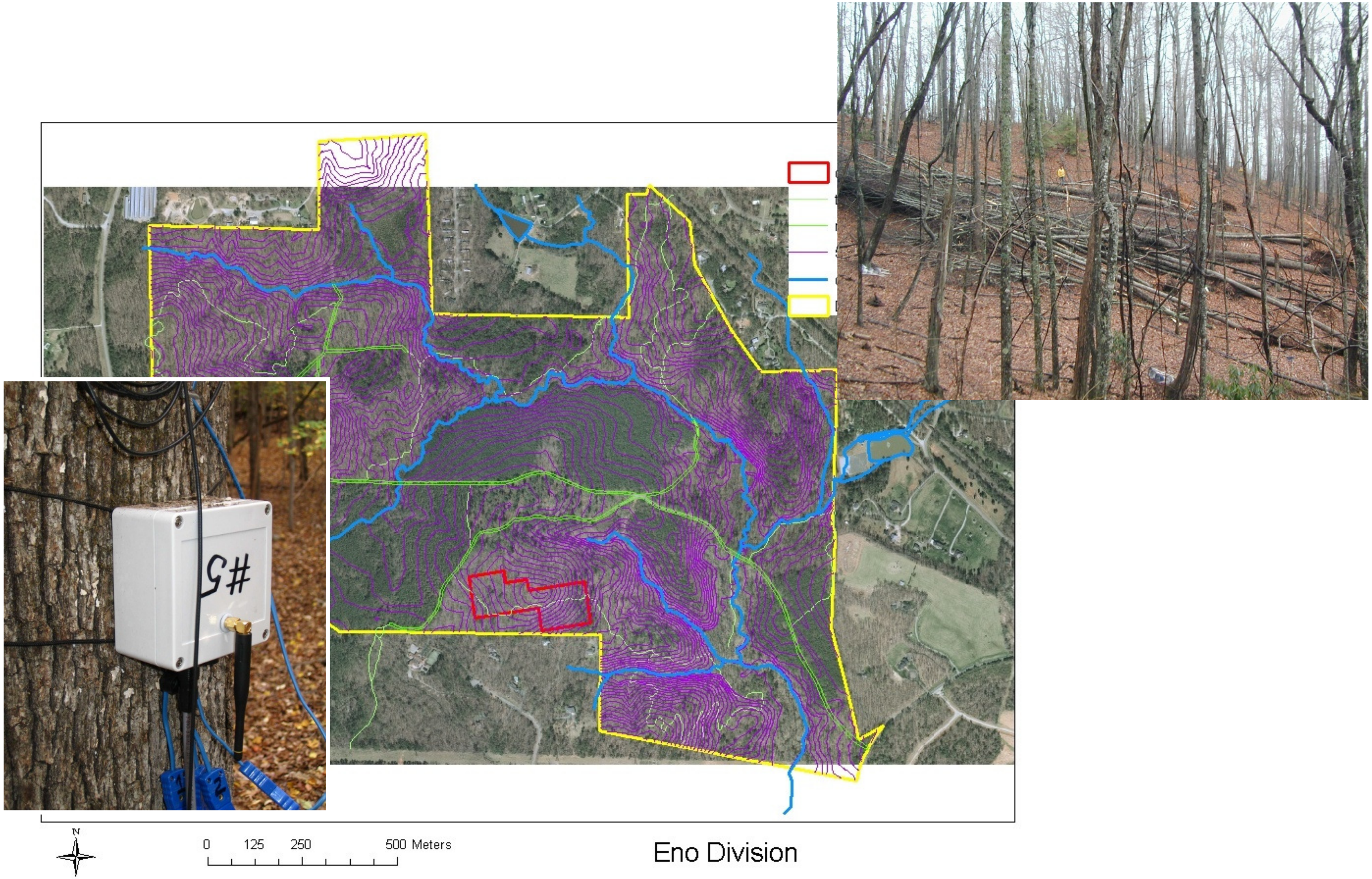
- What is a sensor network?
  - A collection of nodes
  - Node components
    - Sensors (e.g. temperature)
    - Radio (wireless) communication
    - Battery power



Crossbow Mica2

WiSARD

Duke Forest Deployment



Getting All the Data

- Scientists often want ALL the data!
  - No aggregates (e.g. mean)
- Continuous reporting
  - Repeatedly transmit readings to root
    - Explicitly construct central DB and use traditional processing techniques
    - Radio costs too high!
      - ➢ Cost to transmit a bit over radio ~1000 times more than to execute machine instruction

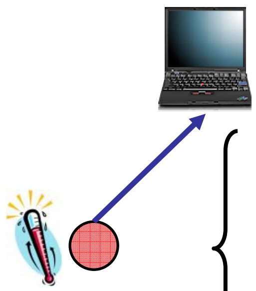
➢ Push processing into network with suppression

Outline

1. Suppression
2. Failure!
3. Coping using redundancy
4. ***BaySail***
   - Inference of missing readings, parameters

Suppression

- Push-based communication
  - Only report deviations from a model
- Value-based Temporal Suppression
  - *model: $temp_t = temp_{(t-1)}$*

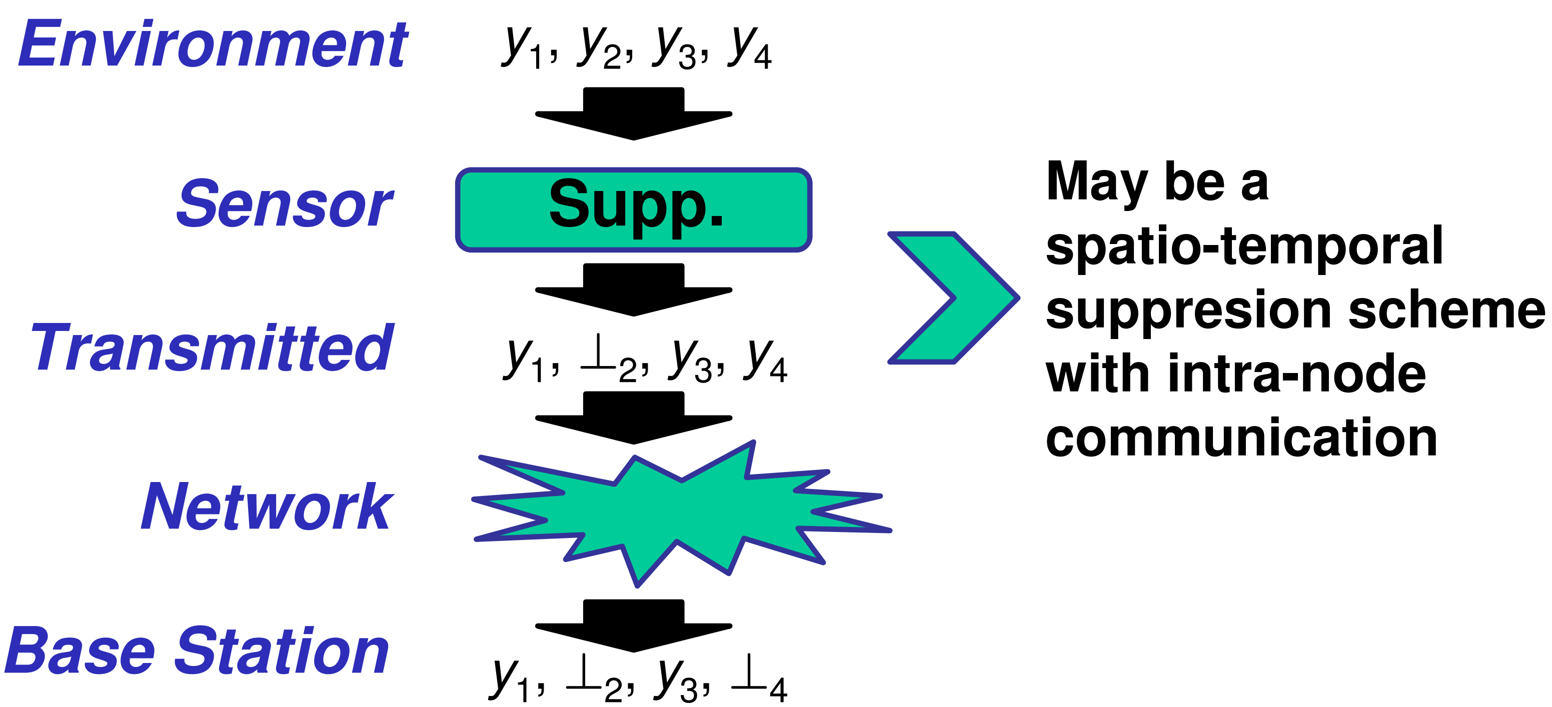


```
if (curr_temp != last_sent_temp) {
  transmit(temp);
  last_sent_temp=curr_temp;
}
```

- In practice, include error tolerance

The Catch for Suppression

- What about reports generated, but lost to failure?



- For non-reported values, the base station cannot distinguish failures from suppressions

Coping With Failure

- Focus on simple temporal suppression
- Learn *ALL* missing values

Two Coping Strategies

System-level acks + re-transmissions	**Application-level redundancy**
• Sender re-sends until receiver returns acknowledgement ➢ **Minimize chance report not received**	• Augment existing reports ➢ **Minimize impact of missing report**

Redundancy

- Temporal Suppression with error tolerance
  - Report only if reading changes beyond ε since last reported
- 5 report types

Name	**Payload Addition**
Standard	Node reading
Counter	Incrementing report number
Timestamp	Last *n* report times
Timestamp D	Last *n* report times + direction bits
History	last *n* times + readings

- Increasing payload, increasing info

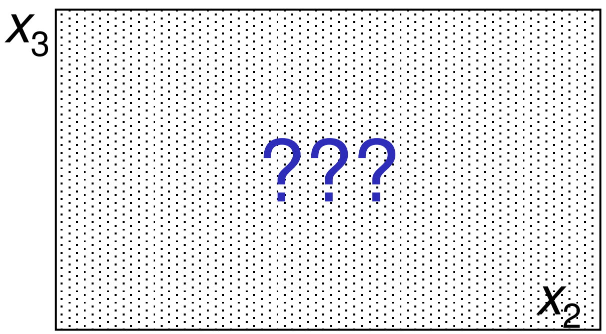
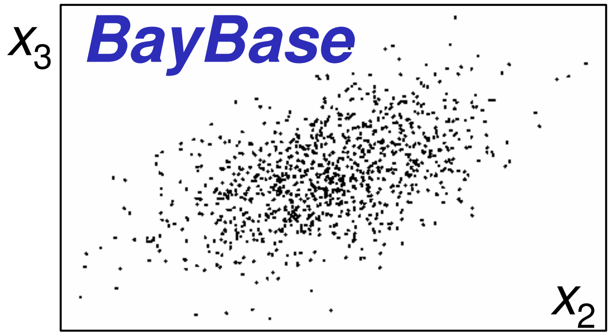
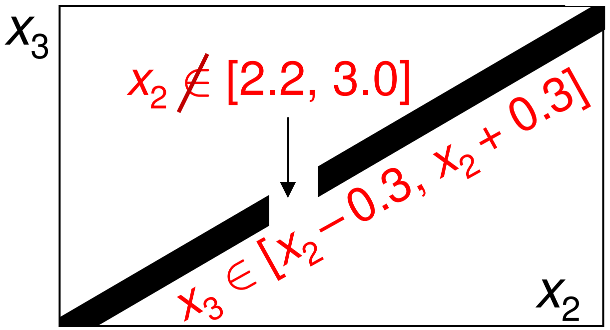
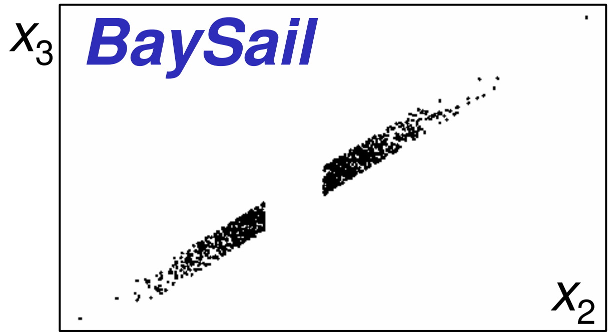
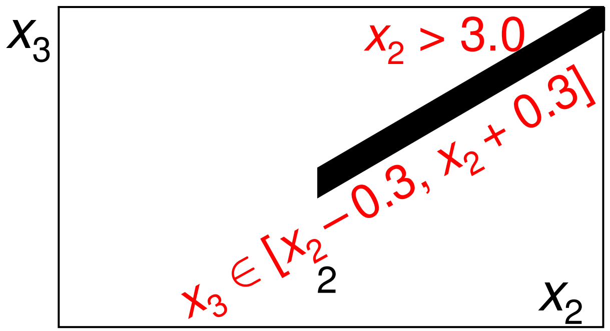
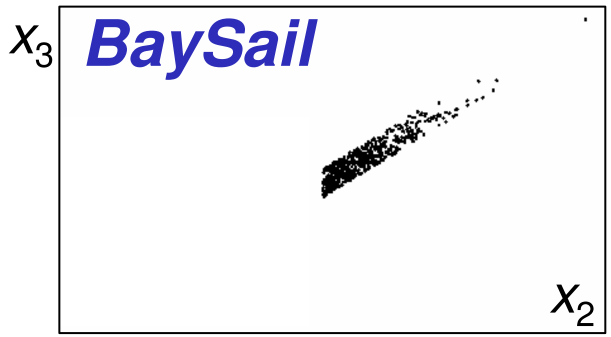
TinyOS Implementation

- Application-level Redundancy
  - Simple to implement
    - 40-50 lines of additional code to a tutorial example
- Lower-level redundancy
  - Activate “acks” in MAC-layer code
  - Re-transmissions in application code
- Failure Rates
  - Tied to distance, clearance, battery, etc.
  - Independent over time
  - 30% failure rate with maximum 2 re-transmissions gives <3% effective failure rate

Suppression-Aware Inference

- Redundancy + knowledge of suppression scheme $\Rightarrow$ hard constraints on missing data
  - Temporal suppression with $\varepsilon = 0.3$, prediction = last reported
  - Actual: $(x_1, x_2, x_3, x_4) = (2.5, 3.5, 3.7, 2.7)$
  - Base station receives: (2.5, nothing, nothing, 2.7)
  - With ***Timestamp*** (r=1)
    - (2.5, failed, suppressed, 2.7)
    - $|x_2 - 2.5| > 0.3$; $|x_3 - x_2| \leq 0.3$; $|2.7 - x_2| > 0.3$
  - With ***Timestamp+Direction Bit*** (r=1)
    - (2.5, failed & increased, suppressed, 2.7 & decreased)
    - $x_2 - 2.5 > 0.3$; $-0.3 \leq x_3 - x_2 \leq 0.3$; $x_2 - 2.7 > 0.3$
  - With ***Count***
    - One suppression and one failure in x_2 and x_3; not sure which
    - A very hairy constraint!
- Posterior: $p(\mathbf{X}_{mis}, \Theta | \mathbf{X}_{obs})$, with $\mathbf{X}_{mis}$ subject to constraints

Using Redundancy

	Just data	Bayesian, model-based AR(1) with uncertain parameter
No knowledge of suppression		
Knowledge of suppression & ***Timestamps***		
Knowledge of suppression & ***Timestamps+ Direction Bits***		

BaySail Key Features

1. Estimates missing readings/parameters
2. Bayesian provides posterior distributions, not just single point estimates
3. Missing data not generically missing
   - Constrain possible settings using suppression scheme and redundancy
4. Computing posteriors is hard
   - Gibbs' sampling iteratively generates samples of reading time series and of each parameter
5. Combine simple, low-cost in-network reporting with efficient out-of-network inference

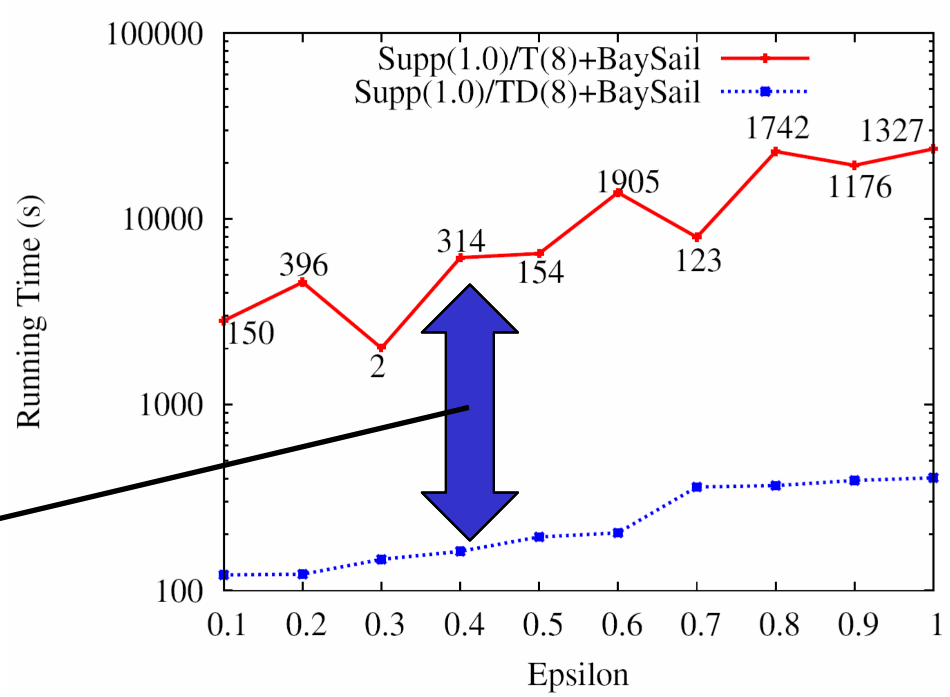
BaySail Experimental Example

- Simple model of soil moisture
  - $y_{s,t} = c_t + \phi\, y_{s,t-1} + \varepsilon_{s,t}$
    - c_t is a series of known precipitations
    - $\phi \in (0,1)$ controls how fast moisture escapes soil
    - $\mathrm{Cov}(Y_{s,t}, Y_{s',t'}) = \sigma^2 (\phi^{|t-t'|}/(1-\phi^2)) \exp(-\tau \|s - s'\|)$
    - τ controls strength of spatial correlation over distance
- *Prior*: $1/\sigma^2 \sim$ Gamma, $\phi \sim U(0,1)$, $\tau \sim$ Gamma
- *Joint Posterior*: $p(Y_{mis}, \phi, \sigma^2, \tau \mid Y_{obs})$ subject to constraints

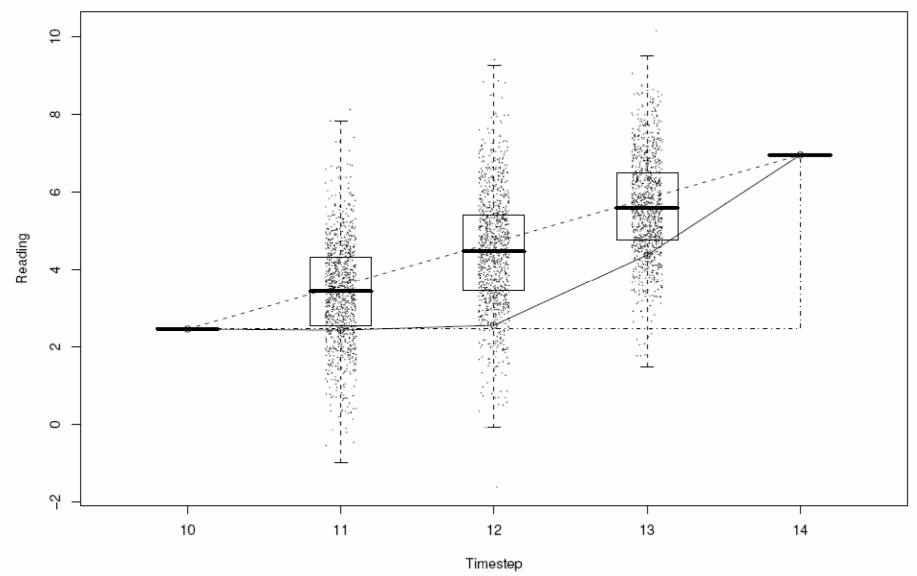
Why the Direction Bit?

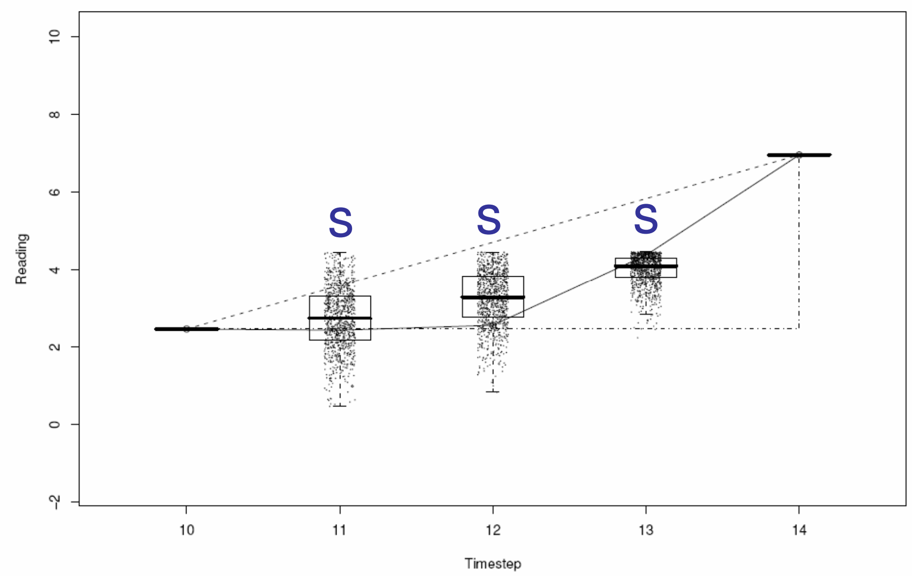
- TS gives OR constraints: $|x_2 - x_1| > \varepsilon$
  - Inefficient *rejection* sampling
- TS+D gives linear constraint: $x_1 - x_2 > \varepsilon$
  - Allows for more efficient sampling [Rodriguez-Yam et al. 04]

>100x improvement… the major reason for the direction bit!



3 Missing Values Cluster



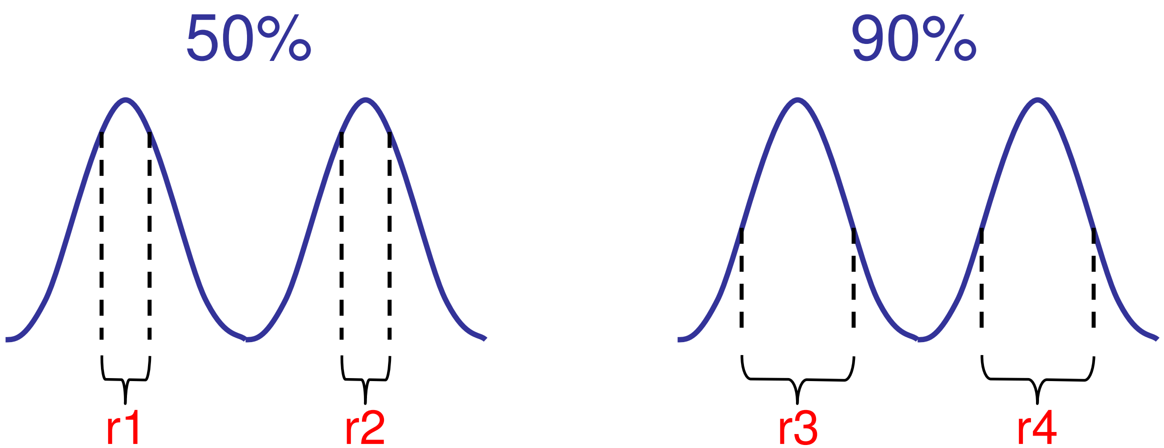


BayBase: Conditioning on model and endpoints

BaySail: Conditioning on model, endpoints, and that missing values are suppressions

Metrics

- Compare posterior mean to actual?
  - Mean misleading for bimodal distributions
- High density regions (hdr)
  - Given percentage *x*, return minimal length range(s) of values such that *x*% of sample's probability density contained in range(s)
  - Ensure hdr covers actual reading x% of time

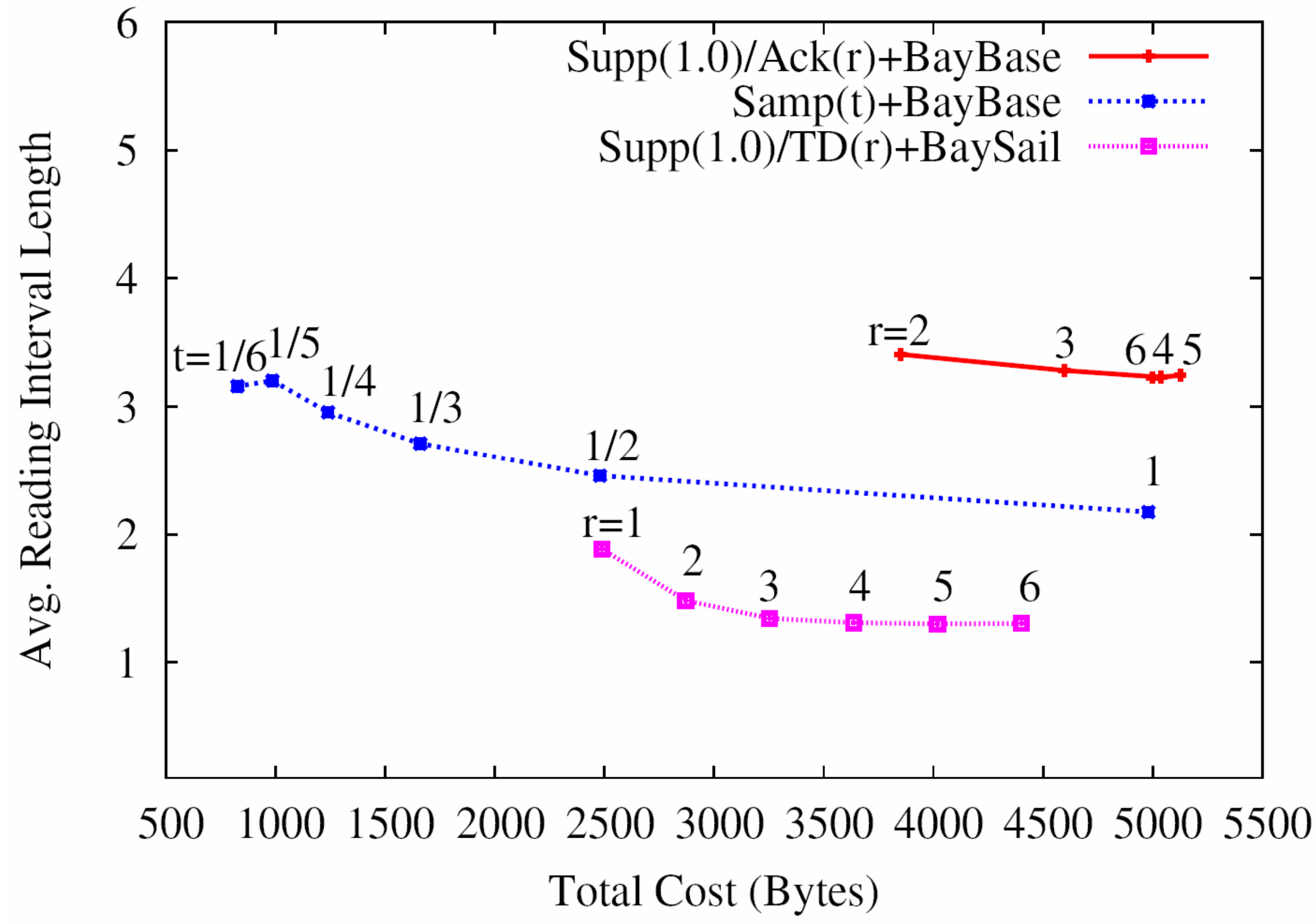


Cost vs. HDR Interval

- Parameters induce 60% suppression rate
  - $\sigma^2 = 1.0$, $\phi = 0.9$, $\varepsilon = 1.0$
- Failure rate 30%
- 3 Schemes
  - Samp(τ)
    - Fixed reporting every τ rounds
  - Supp/TD(r)
    - Timestamp + direction for last r reports
  - Supp/Ack(r)
    - Maximum r re-transmission attempts

Readings Interval

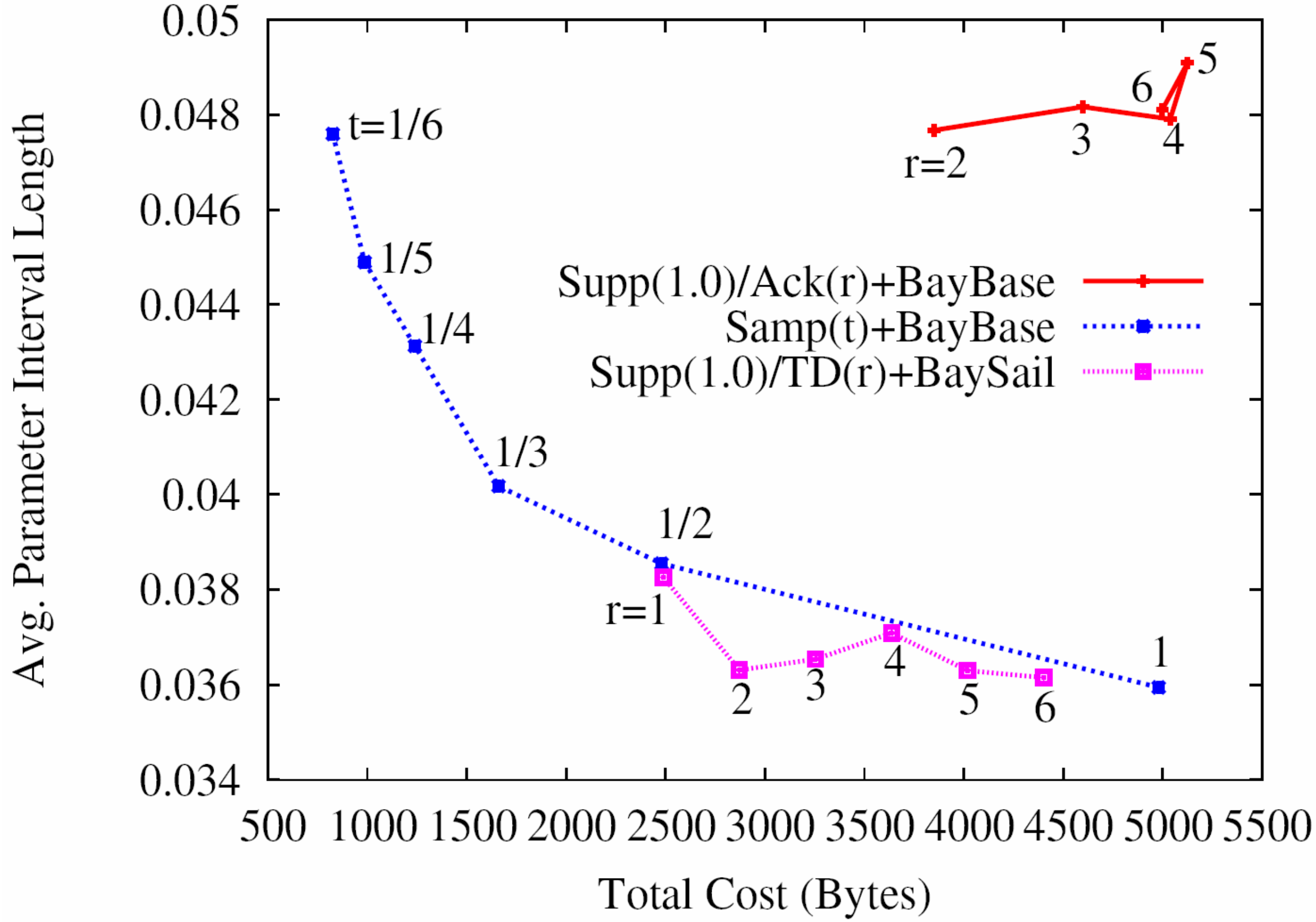
80% hdr



- BaySail demonstrates significant improvement

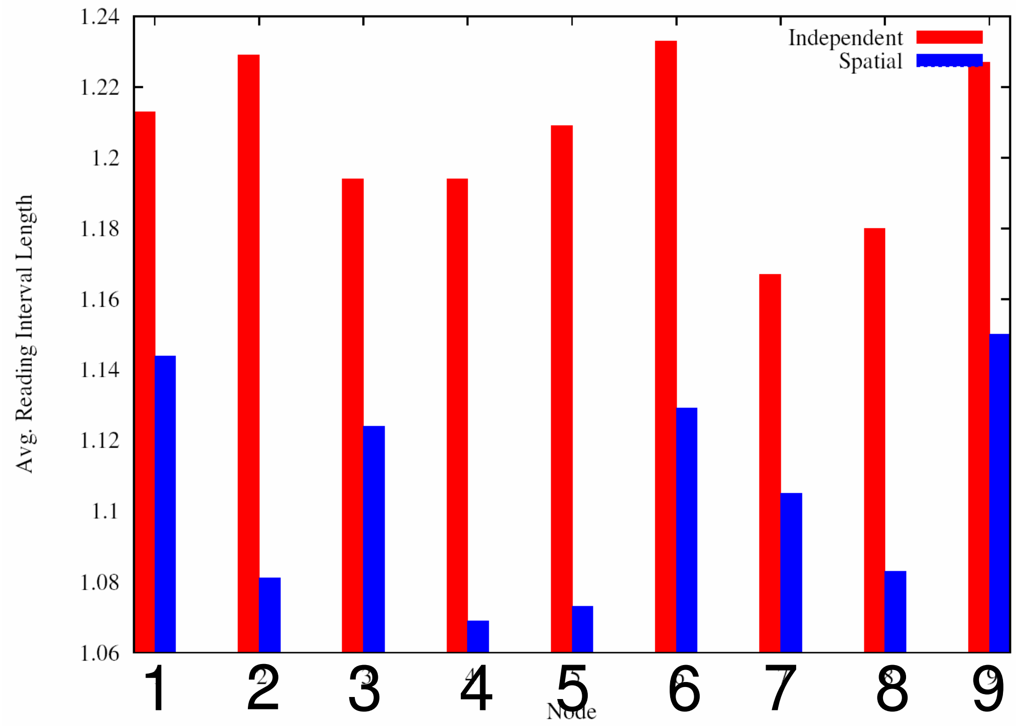
Phi Interval

80% hdr

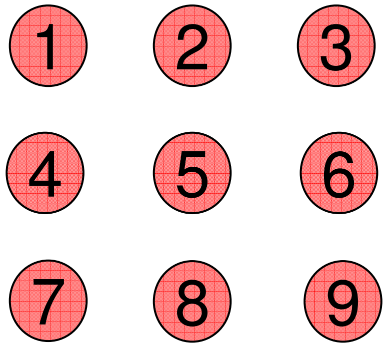


- ➢Choice has little effect for process parameter

Spatial Inference



3x3 Grid



Conclusion

- Suppression is a viable technique only when made robust to failure
- BaySail combines low-cost in-network redundancy with efficient out-of-network statistical inference
  - Generates posteriors distributions on raw missing values and process parameters
- Future Challenges
  - Sophisticated spatio-temporal schemes
    - Failure on in-network constraints
    - Failure of model parameter transmission
  - Storing query results