Challenges for Machine Learning in Ecological Science and Environmental Management

Tom Dietterich

Oregon State University

In collaboration with Ethan Dereszynski, Rebecca Hutchinson, Dan Sheldon, Weng-Keen Wong, Claire Montgomery and the Cornell Lab of Ornithology

Sustainable Management of the Earth's Ecosystems

- The Earth's Ecosystems are complex
- We have failed to manage them in a sustainable way
- Why?

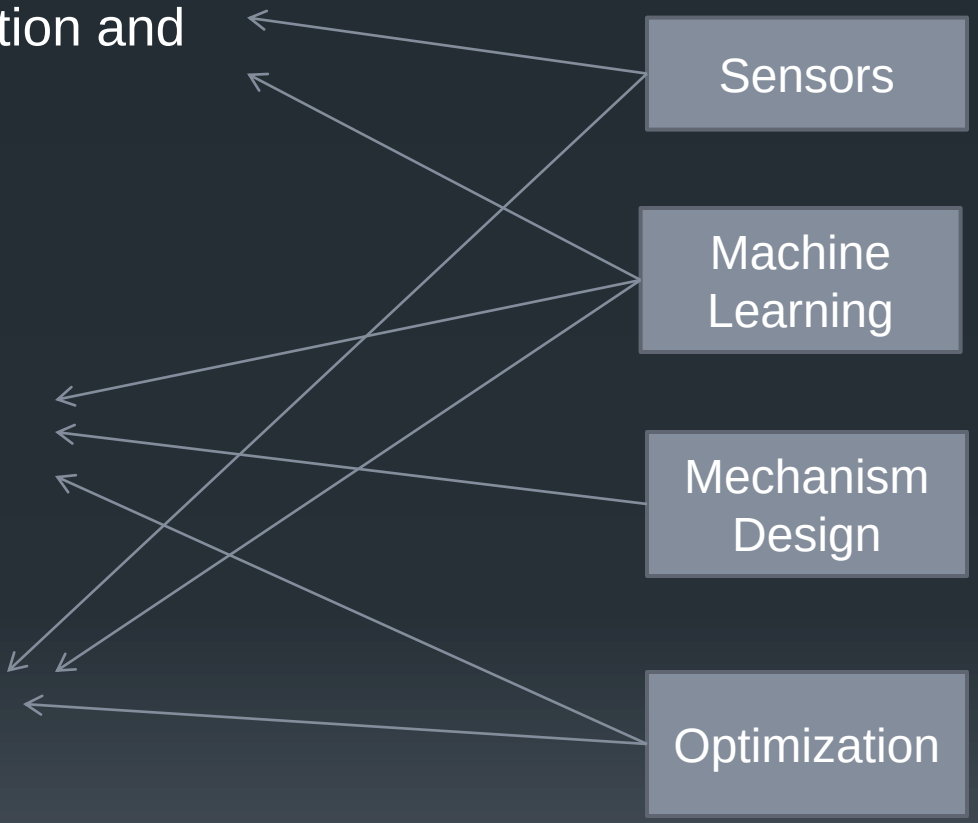
1. Our knowledge of function and structure is inadequate
   - Doak et al (2008): Ecological Surprise

2. We focused only on part of the larger system
   - We have ignored
     - human / social components
     - spatial aspects
     - interactions among multiple species

3. We simplified the systems to make them manageable
   - High-efficiency agriculture relies on expensive, non-sustainable exogenous inputs: energy, fertilizer, pesticides, herbicides

Computer Science can help!

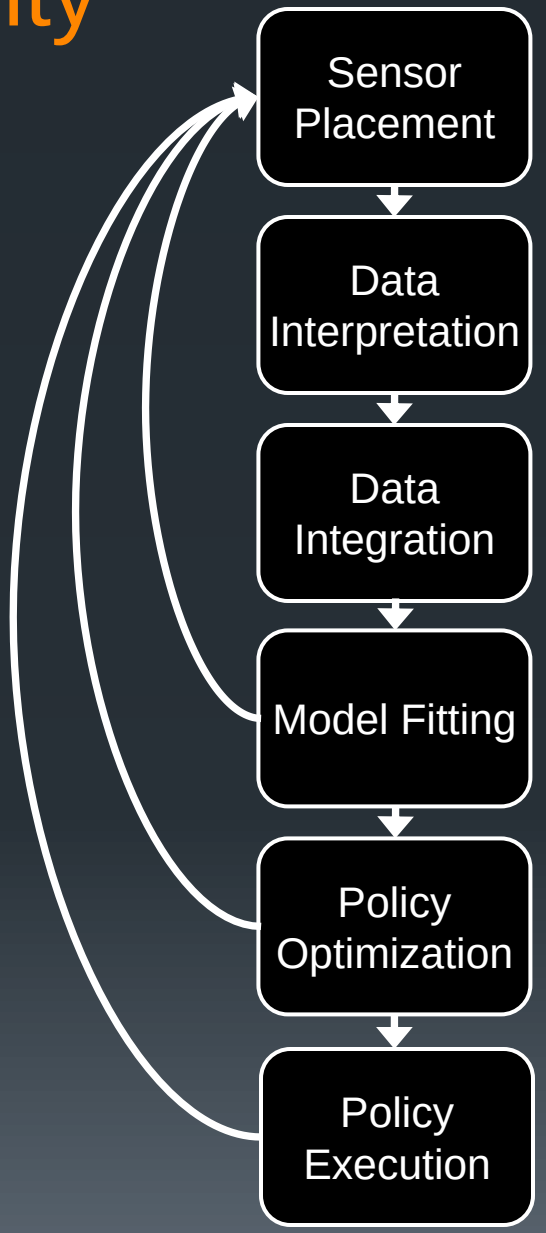
1. Lack of knowledge of function and structure
2. Focus on subsystems
3. Simplified systems using exogenous inputs



Computational Sustainability

- The study of computational methods that can contribute to the sustainable management of the earth's ecosystems
  - biological
  - social
  - economic

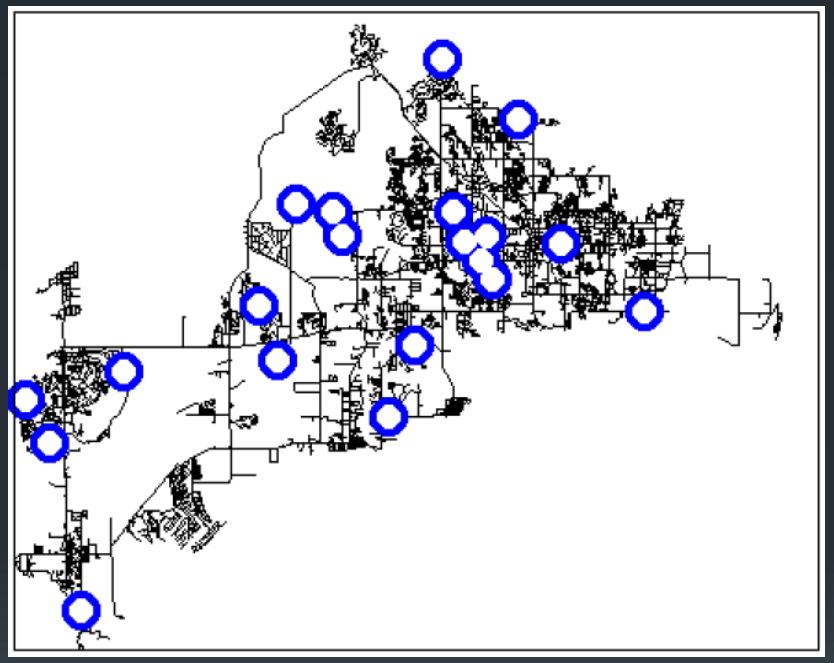
- Data → Models → Policies



Example Research Efforts

Sensor Placement

- Objectives
  - detection probability
  - improving model accuracy
  - improving causal understanding
  - improving policy effectiveness
- Key Tool: Submodular Functions
  - Formulate the problem in terms of a submodular objective
  - Greedy algorithm then works well and has provable performance



Leskovec et al, KDD2007

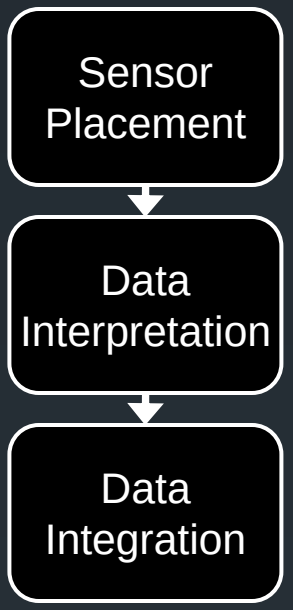
Data Interpretation

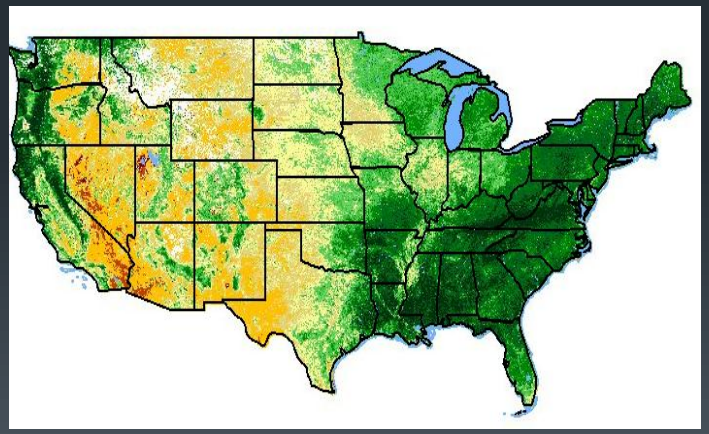
- Insect identification for population counting
- Raw data: image
- Interpreted data: Count by species

Species	Count
Nilaparvata lugens	12
Sogatella furcifera	8
Laodelphax striatellus	0
Cnaphalocrocis medinalis	0
Chilo suppressalis	45
Sesamia inferens	18

Data Integration

- Integrating heterogeneous data sources to predict when migrating birds will arrive:
  - Landsat (30m; monthly)
    - land cover type
  - MODIS (500m; daily/weekly)
    - land cover type
    - “greening” index
  - Census (every 10 years)
    - human population density
    - housing density and occupation
  - Interpolated weather data (15 mins)
    - rain, snow, solar radiation, wind speed & direction, humidity
  - Integrated weather data (daily)
    - warming degree days
  - Digital elevation model (rarely changes)
    - elevation, slope, aspect

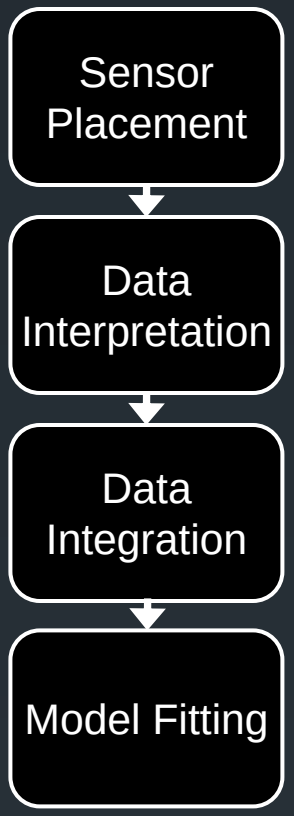




Landsat NDVI:
http://ivm.cr.usgs.gov/viewer/

Model Fitting

- Species Distribution Models
  - create a map of the distribution of a species
- Meta-Population Models
  - model a set of patches with local extinction and colonization
- Migration and Dispersal Models
  - model the trajectory and timing of movement



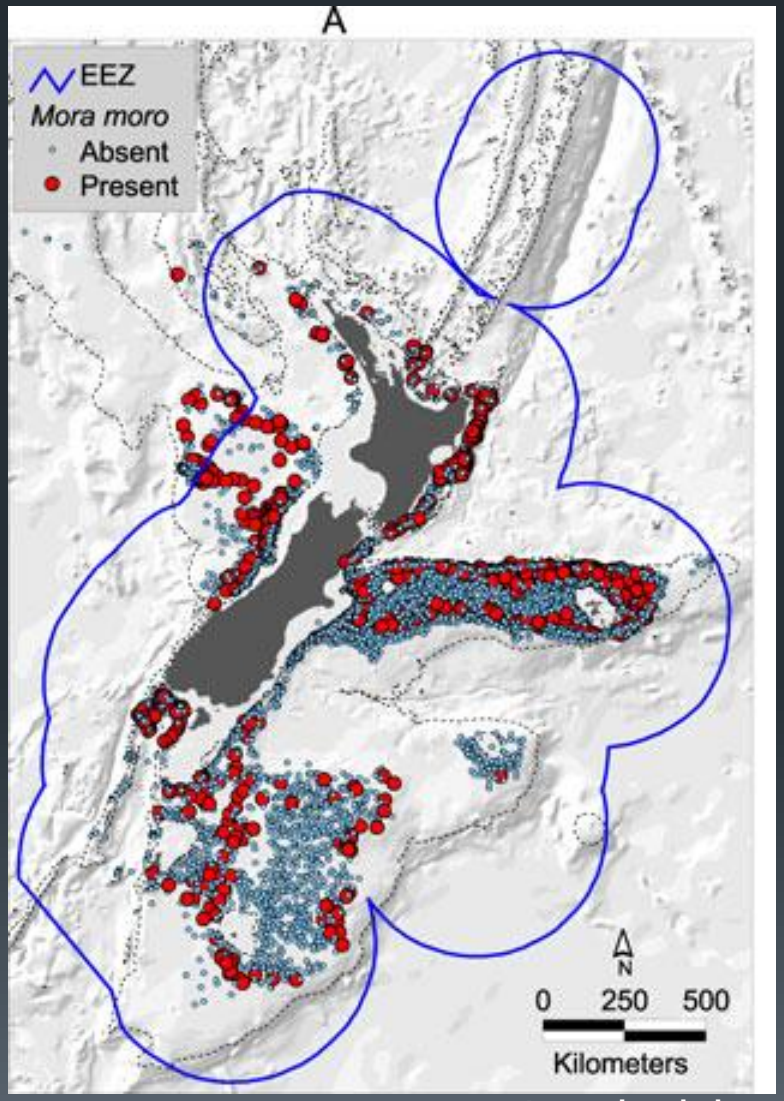
Example Fitted Model: STEM Model of Bird Species Distribution

Indigo Bunting

slide courtesy of Daniel Fink

Policy Optimization

Observations



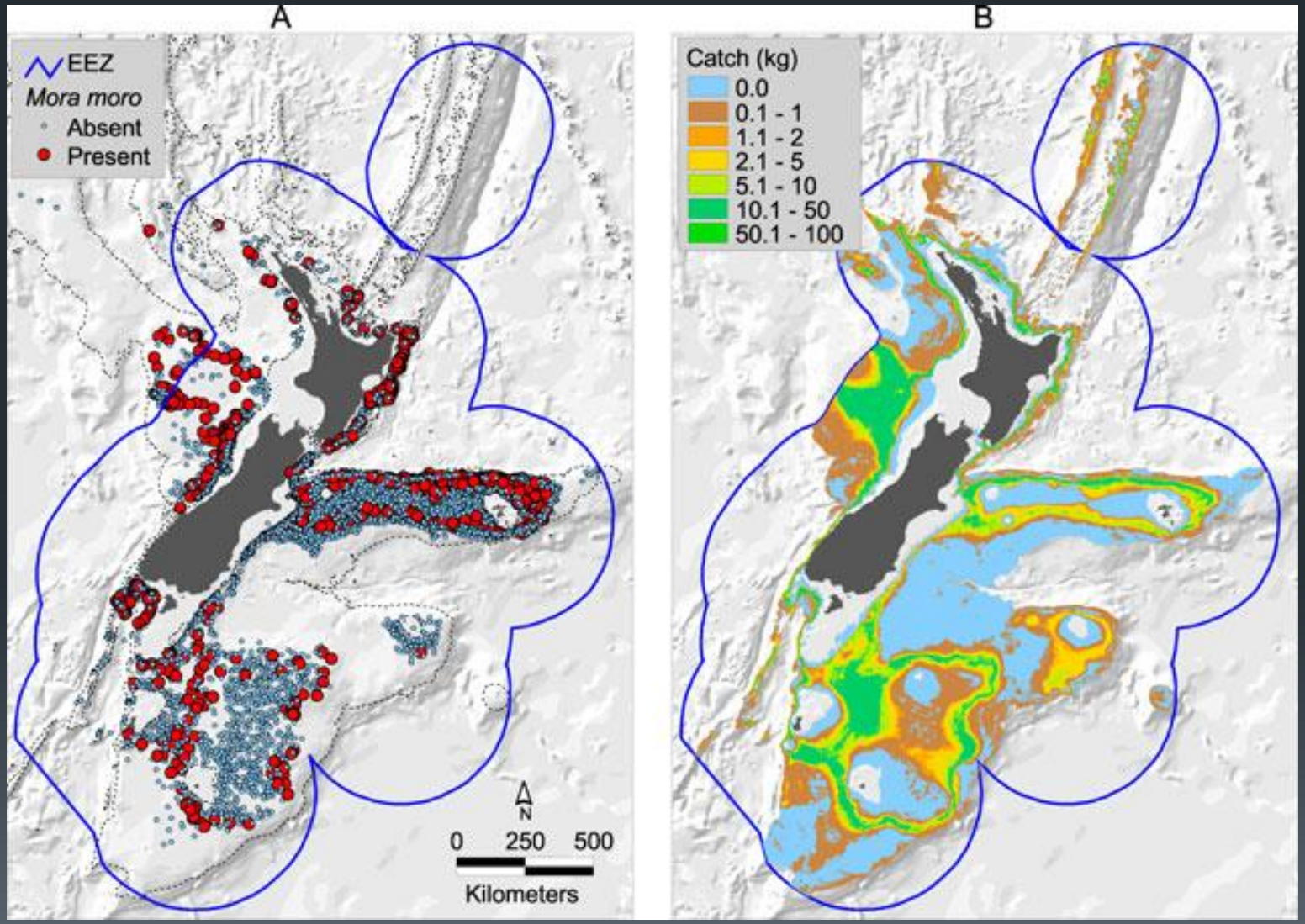
Leathwick et al, 2008

- Sensor Placement
- Data Interpretation
- Data Integration
- Model Fitting
- Policy Optimization

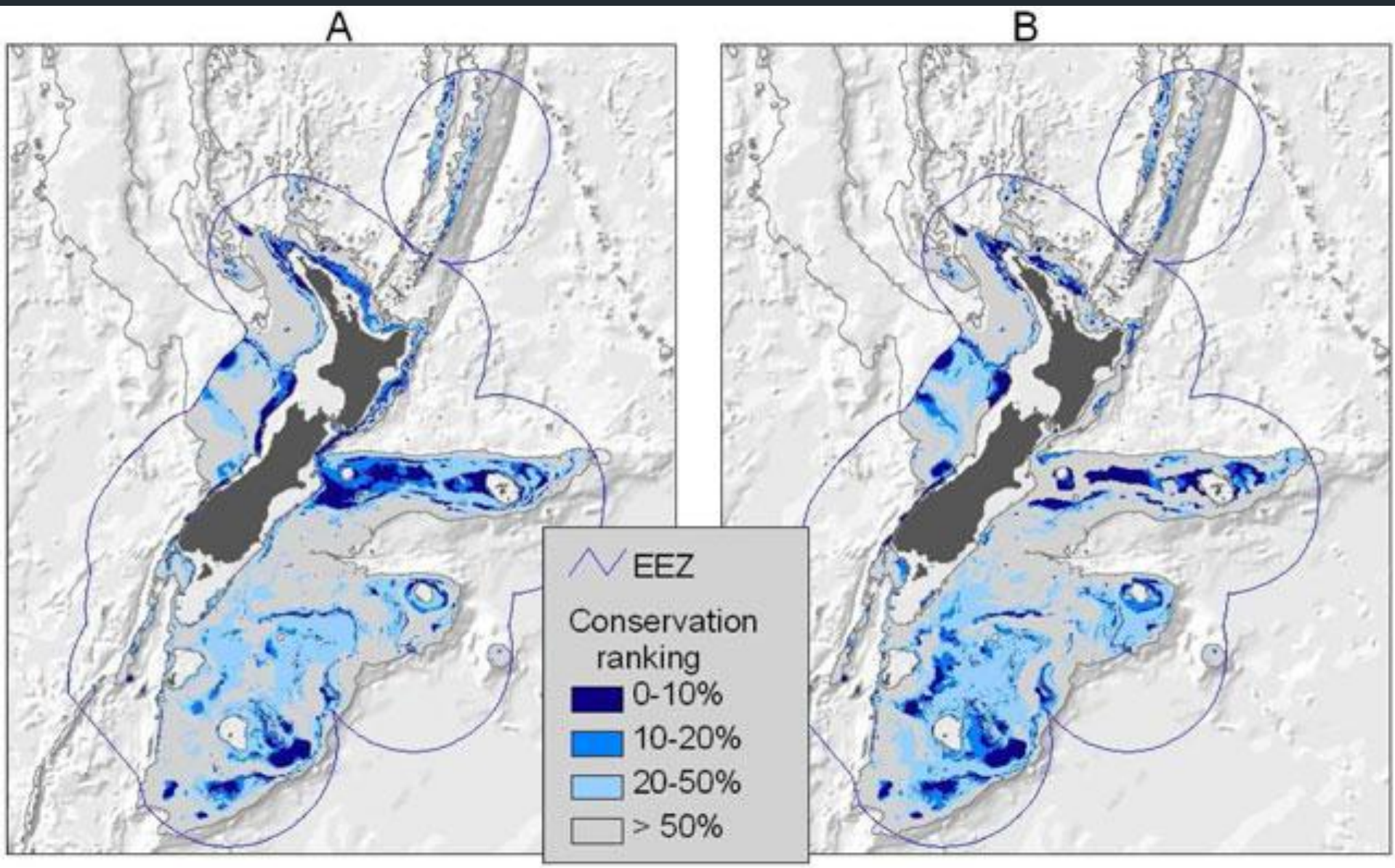
Policy Optimization

Observations

Fitted Model



Leathwick et al, 2008



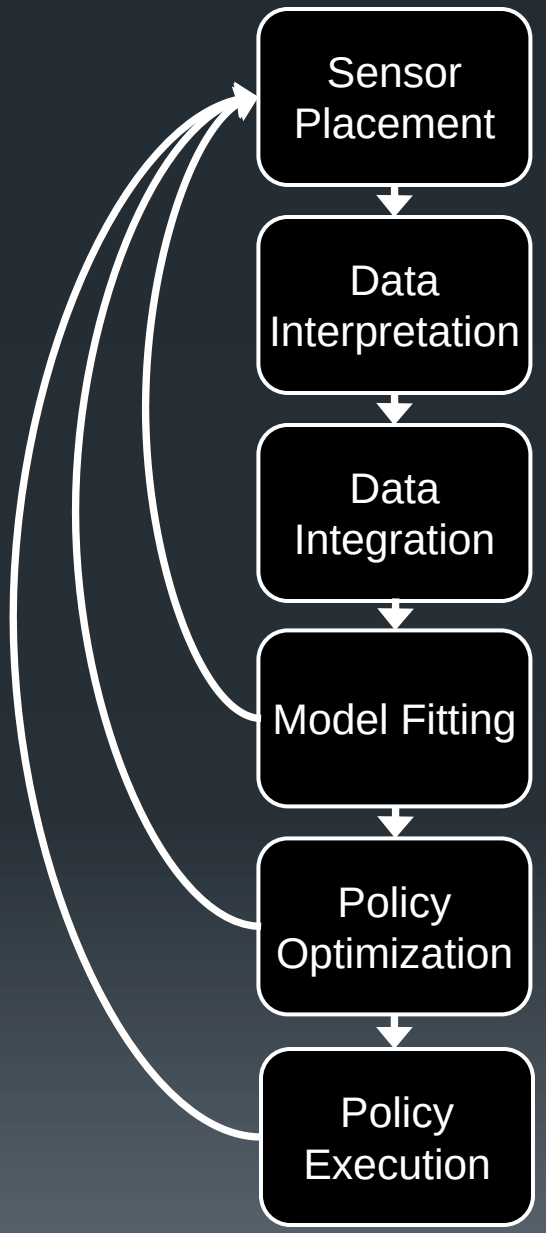
Disregarding costs to fishing industry

Full consideration of costs to fishing industry

Leathwick et al, 2008

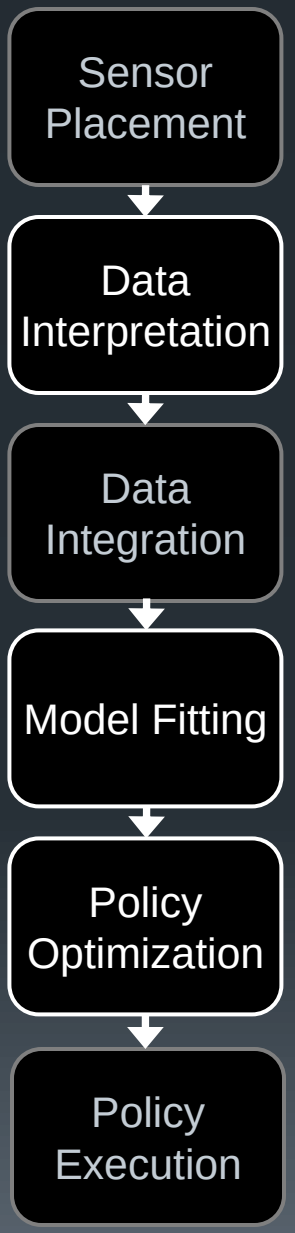
Policy Execution

- Repeat
  - Observe Current State
  - Choose and Execute Action
- Need to continually improve our models and update our policies
- Challenge: We must start taking actions while our models are still very poor.
  - How can we make our models robust to both the “known unknowns” (our known uncertainty) and the “unknown unknowns” (things we will discover in the future)



Outline:
Three Projects at Oregon State

- Data Interpretation
  - Automated Data Cleaning
- Model Fitting
  - Explicit Observation Models
- Policy Optimization
  - Managing Fire in Eastern Oregon

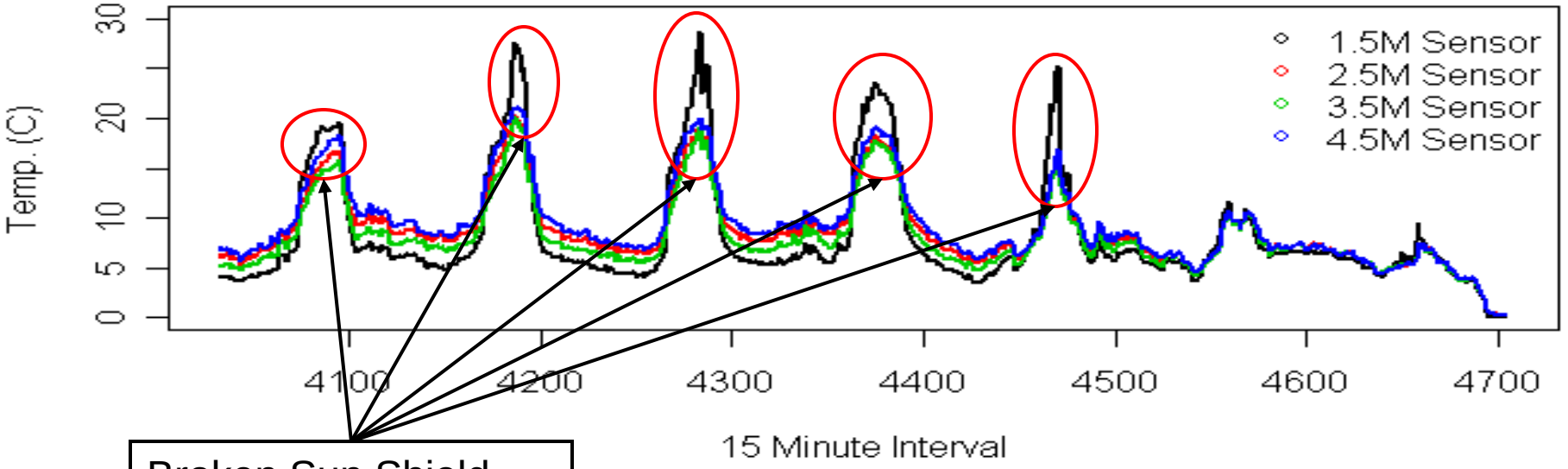


Automated Data Cleaning for Sensor Networks

- Ethan Dereszynski's PhD Work
- He will graduate in Spring 2012

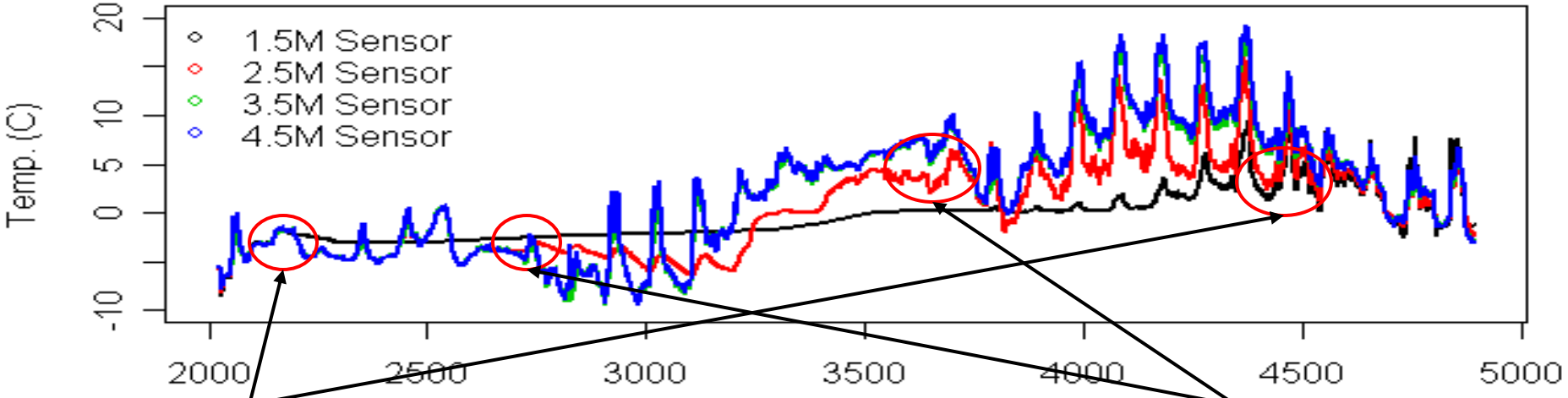
Upper Lookout Met. Station

Central, 1996, Week 6



Broken Sun Shield

Upper Lookout, 1996, Week 3

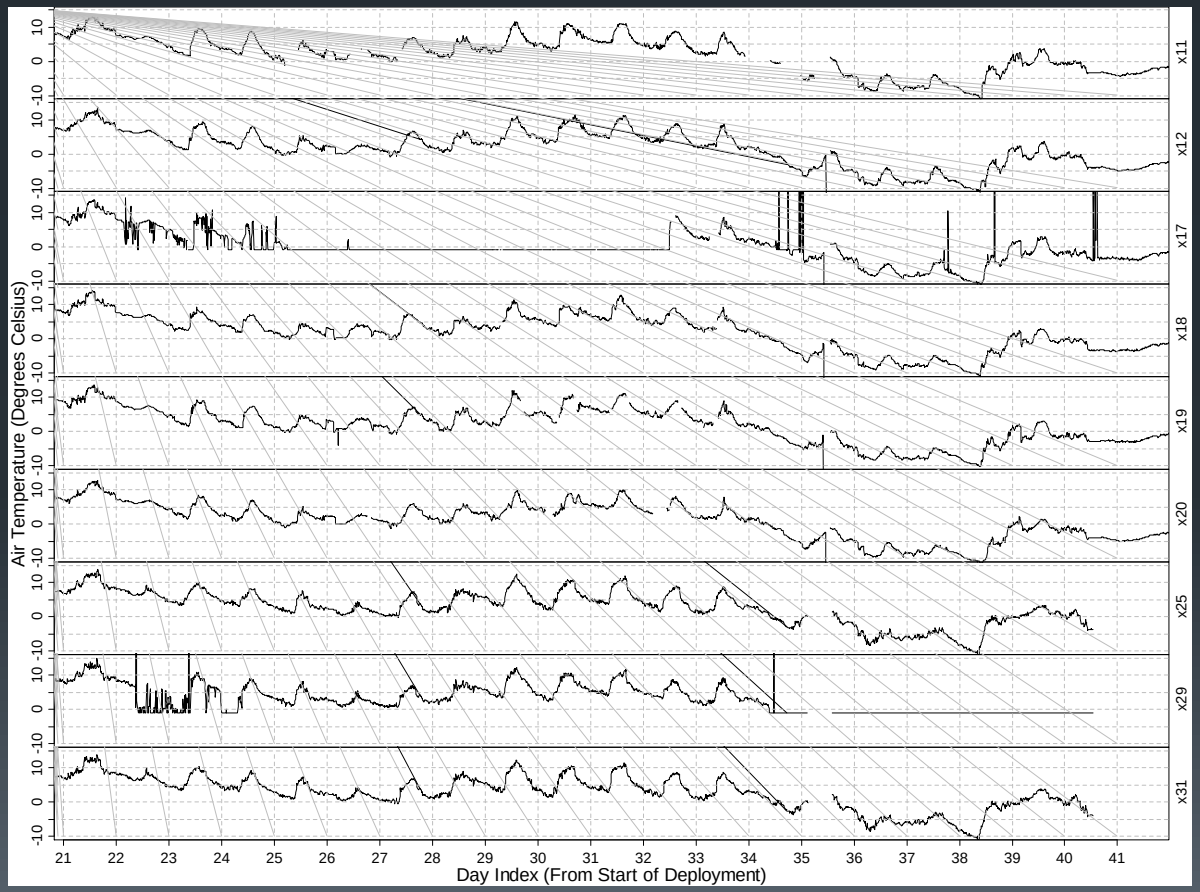


Snow Pack

Snow Dampening

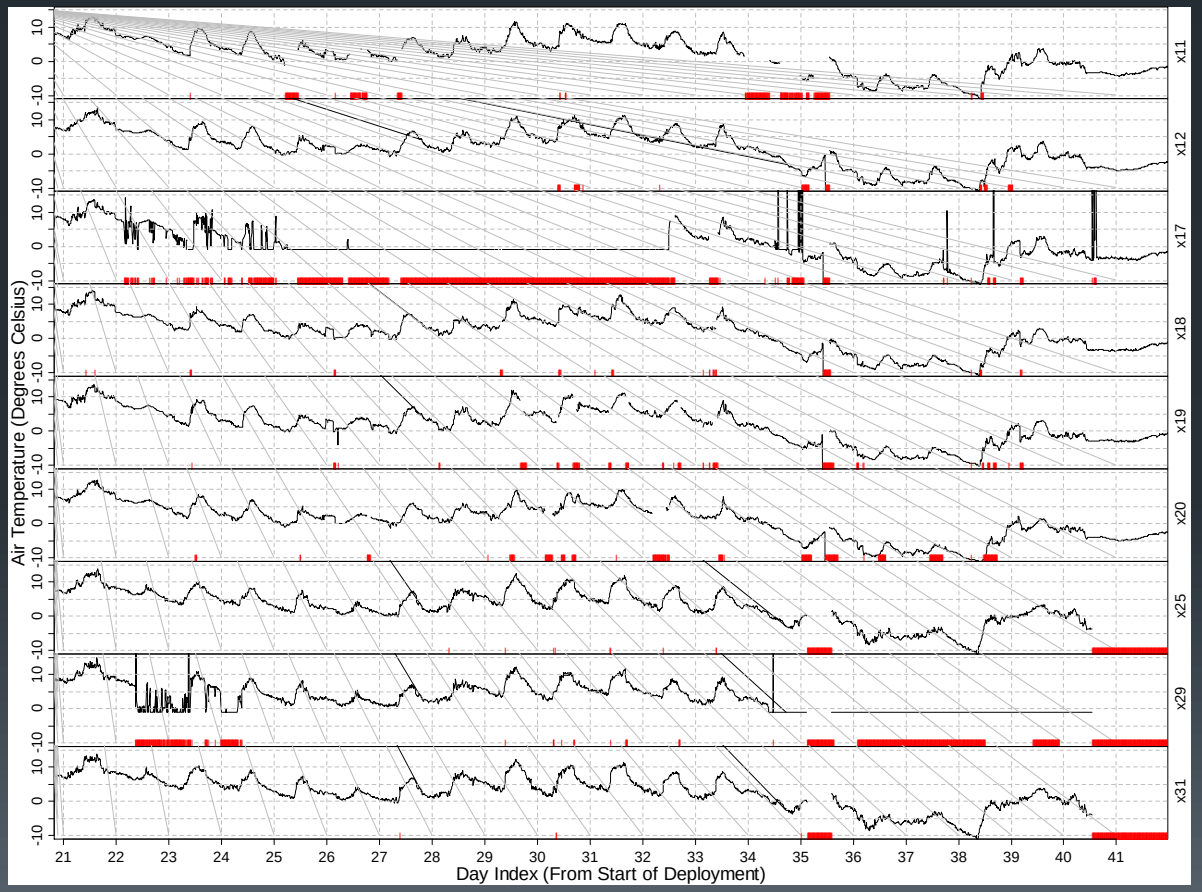
Functions of a Data Cleaning Method

- An ideal method should produce two things given raw data:



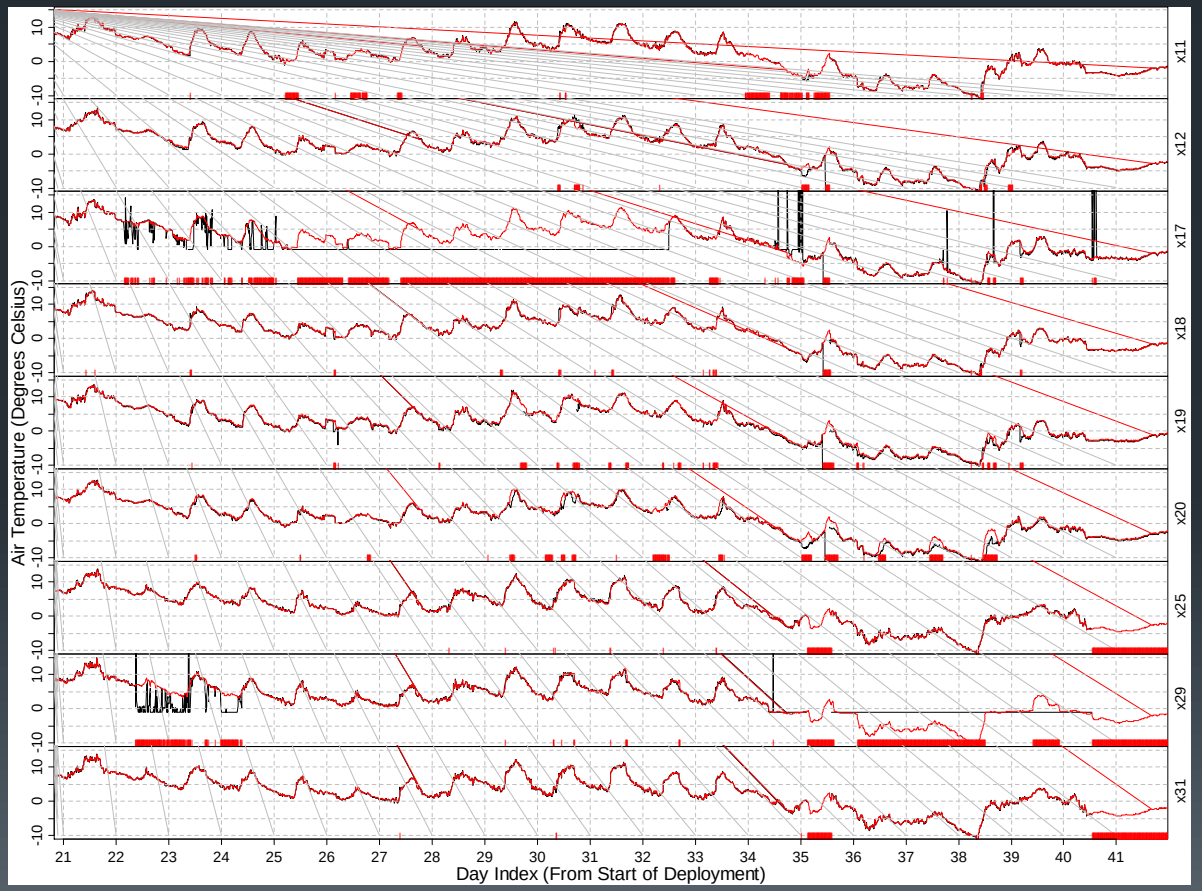
Functions of a Data Cleaning Method

- An ideal method should produce two things given raw data:
  - A label that marks anomalies



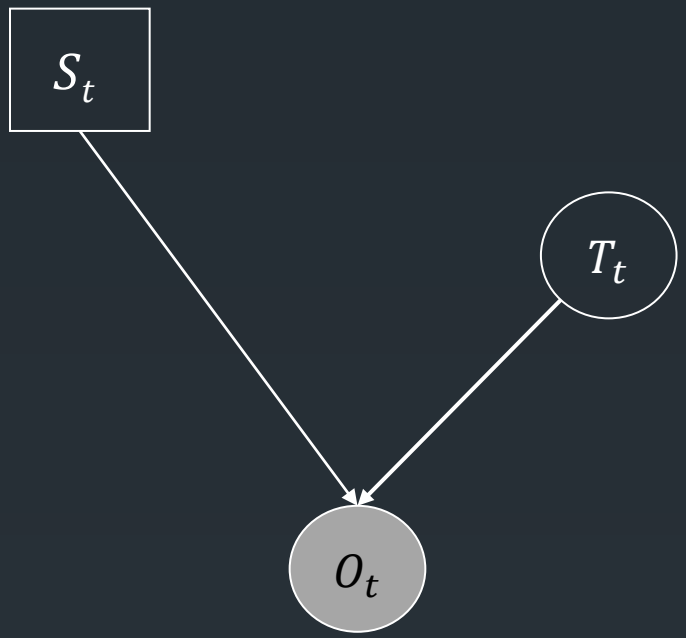
Functions of a Data Cleaning Method

- An ideal method should produce two things given raw data:
  - A label that marks anomalies
  - An imputation of the true value (with some confidence measure)



Method: Probabilistic Modeling Using a Bayesian Network with Hidden Variables

State of the sensor
1 = working; 0 = broken

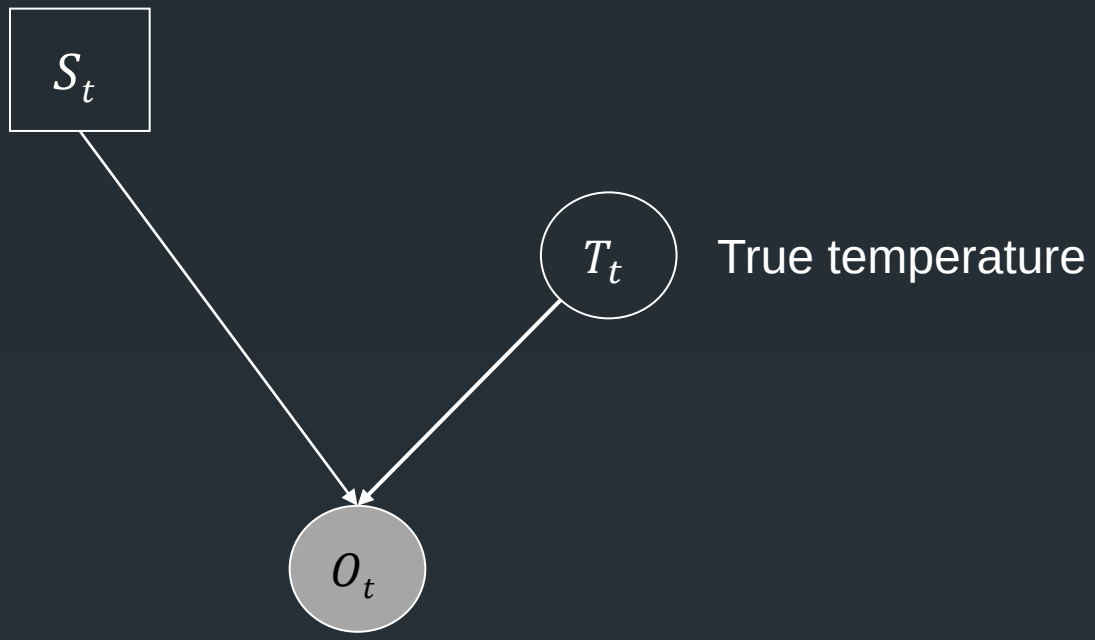


$$P(O_t = o | S_t = 1, T_t = x) = \text{Normal}(o\ |x, \epsilon^2)$$

$$P(O_t = o | S_t = 0, T_t = x) = \text{Normal}(o\ |0{,}1000)$$

Anomaly Detection Via Probabilistic Inference

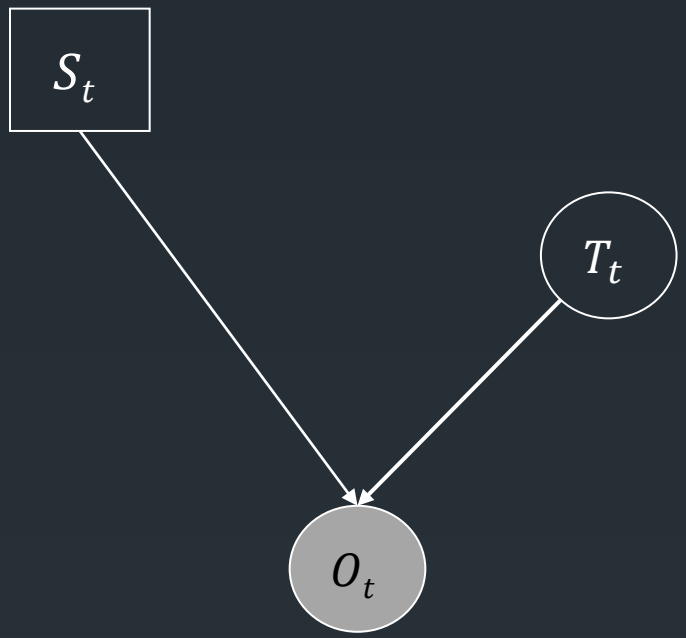
State of the sensor
1 = working; 0 = broken



Query: What is the most likely value of S_t?

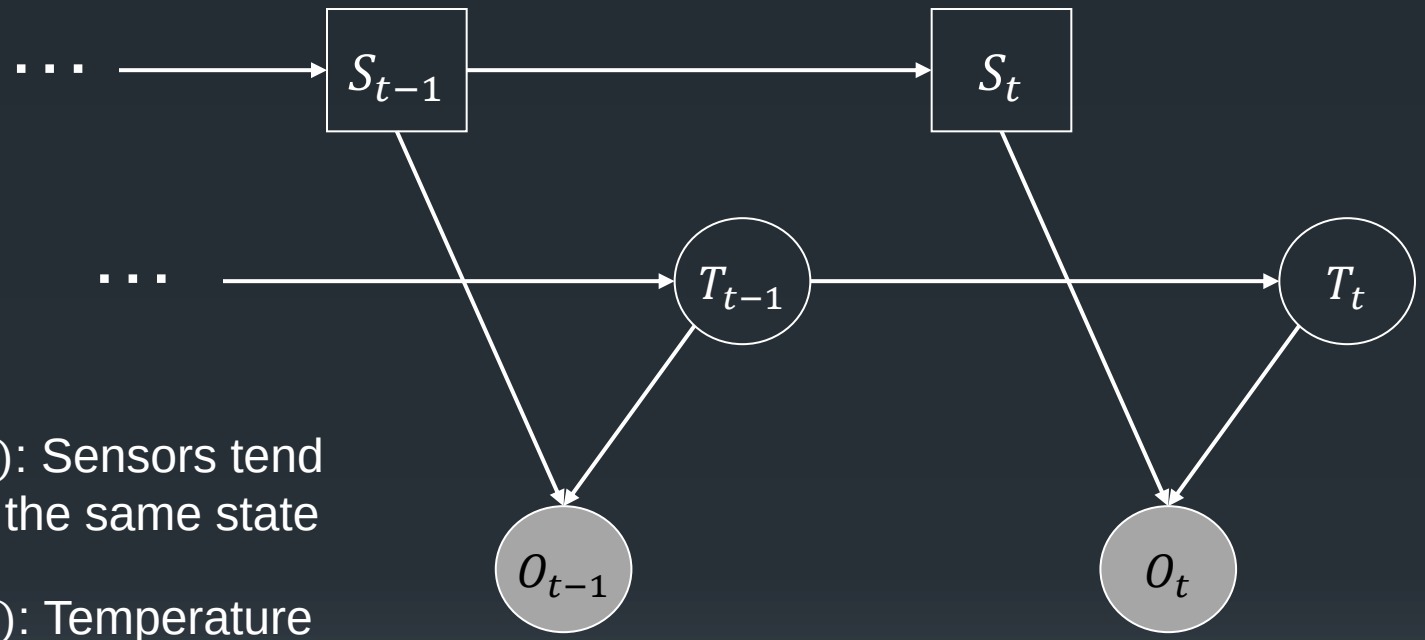
$$\underset{s}{\operatorname{argmax}} P(S_t = s | O_t)$$

Imputation Via Probabilistic Inference



Query: What is the most likely value of T_t? $\underset{x}{\operatorname{argmax}} P(T_t = x | O_t)$

Improving the Model: Markov Model of Temperature

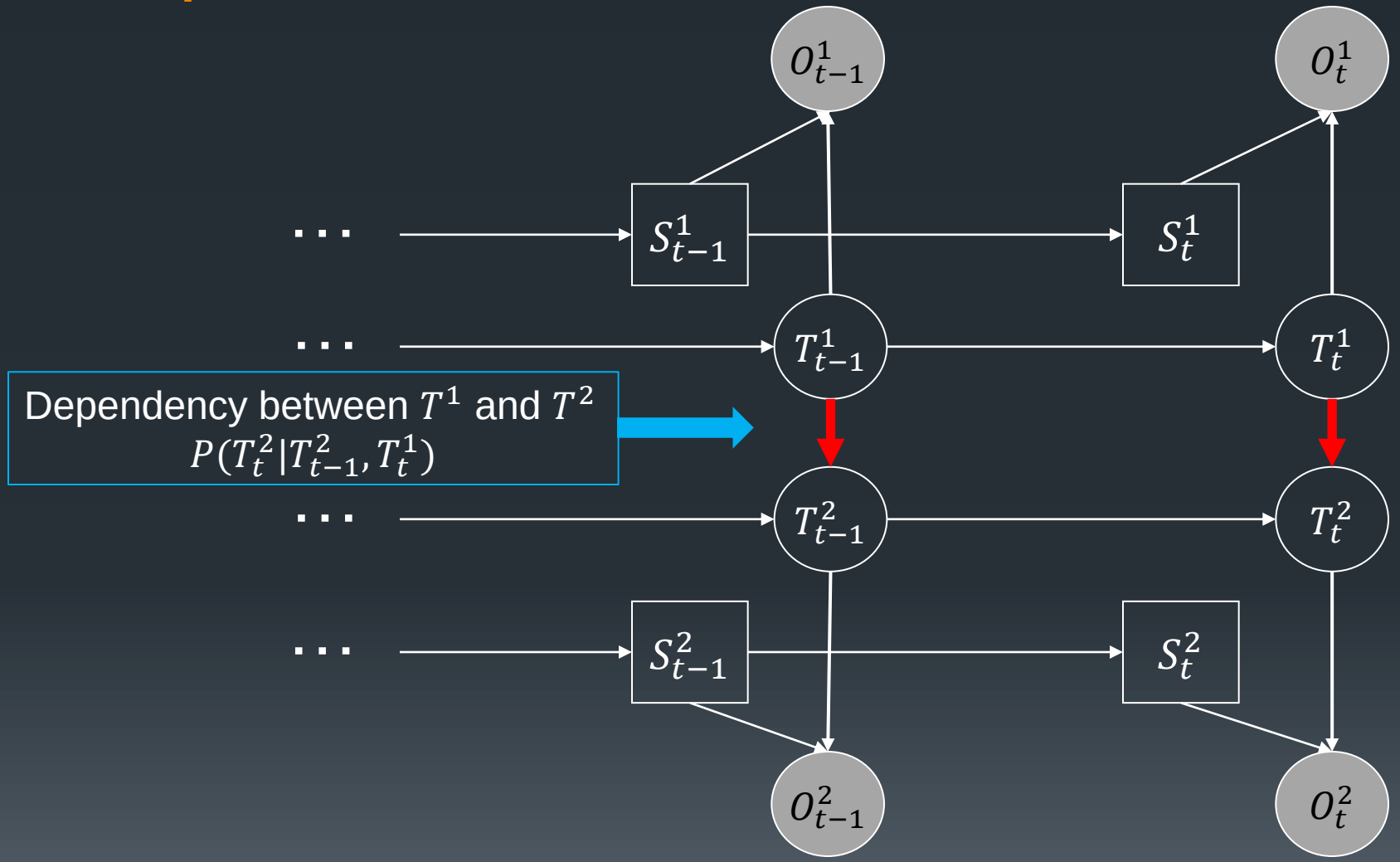


$P(S_t|S_{t-1})$: Sensors tend to stay in the same state

$P(T_t|T_{t-1})$: Temperature changes slowly (15 minute time step)

Query: $\underset{S_t}{\operatorname{argmax}} P(S_t|O_t, O_{t-1}, \dots)$

Improving the Model: Multiple Sensors



Probabilistic Inference is Infeasible in the Single Sensor Model

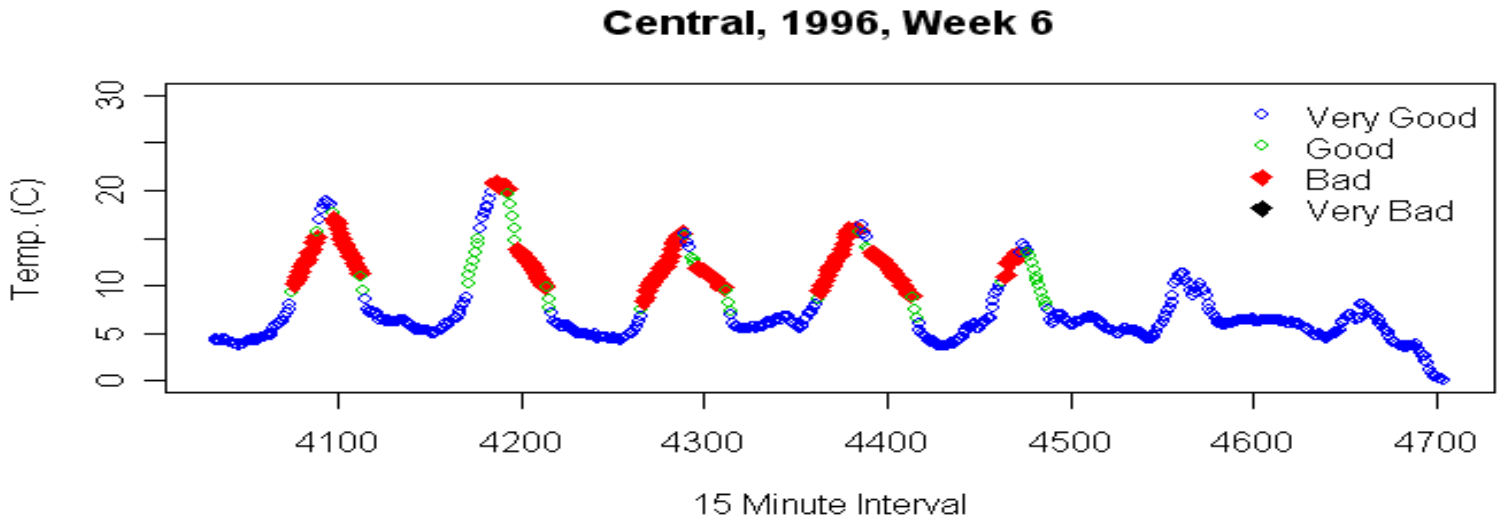
- Single sensor Markov model query: $\underset{S_t, S_{t-1}, \ldots}{\operatorname{argmax}} P(S_t, S_{t-1}, \ldots | O_t, O_{t-1}, \ldots)$
  - Requires time exponential in the length of the time series
- Solution:
  - Commit to each S_t in order
  - $S_1 \coloneqq \underset{s}{\operatorname{argmax}} P(S_1 = s | O_1)$
  - $S_2 \coloneqq \underset{s}{\operatorname{argmax}} P(S_2 = s | S_1, O_2)$
  - …
  - Also bound the variance of T_t
- Each of these inferences is easy

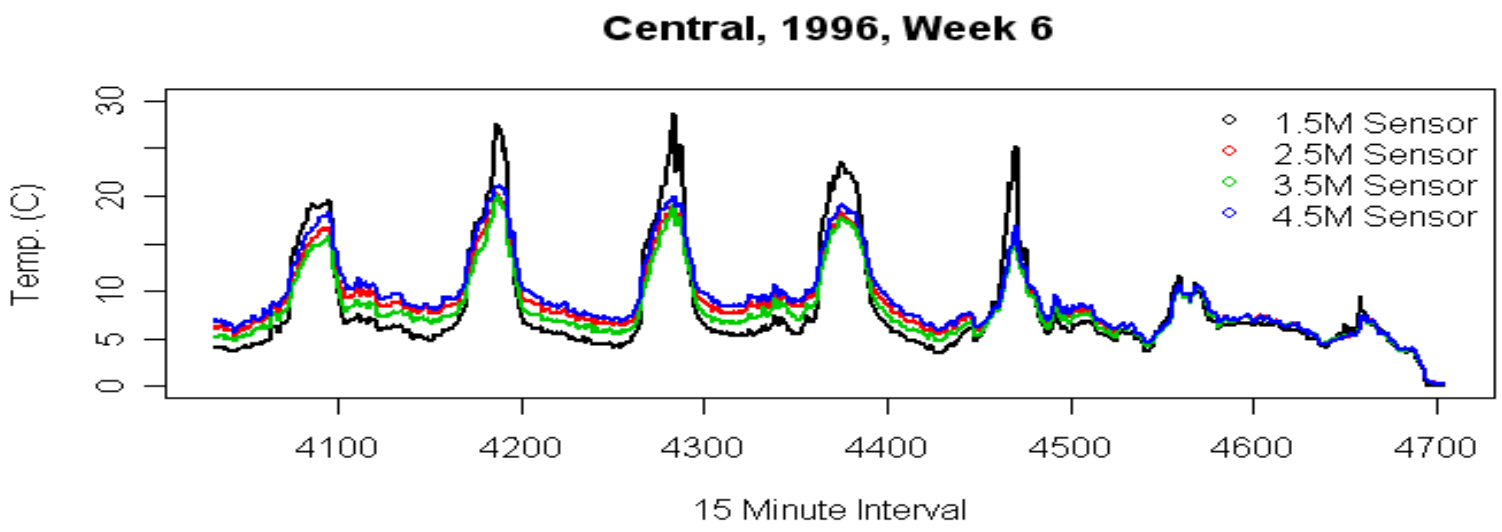
Probabilistic Inference is Infeasible in the Multiple Sensor Model

- Even if we commit to values for $S_t^1, S_t^2 \ldots, S_t^K$ for K sensors, we must compute an intermediate data structure of size 2^K

- Possible Solution: SearchMAP. At each time t,
  - Start with $(S_t^1, \ldots, S_t^K) = \mathcal{S}_t = (1, 1, \ldots, 1)$ // all sensors working
  - Perform a greedy search to maximize $P(\mathcal{S}_t | O_t^1, \ldots, O_t^K)$ by "breaking" one sensor at a time
  - Polynomial in K

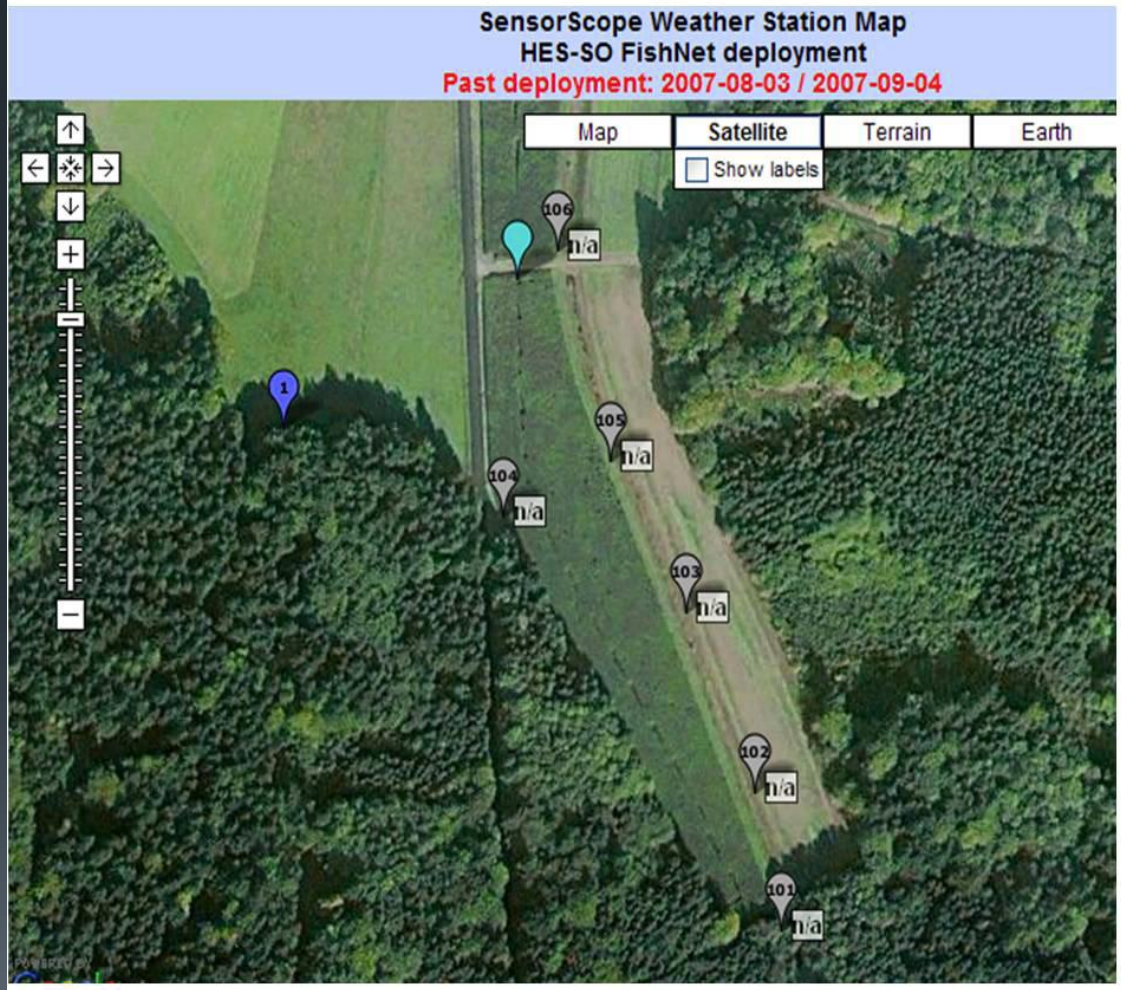
Single Sensor Results

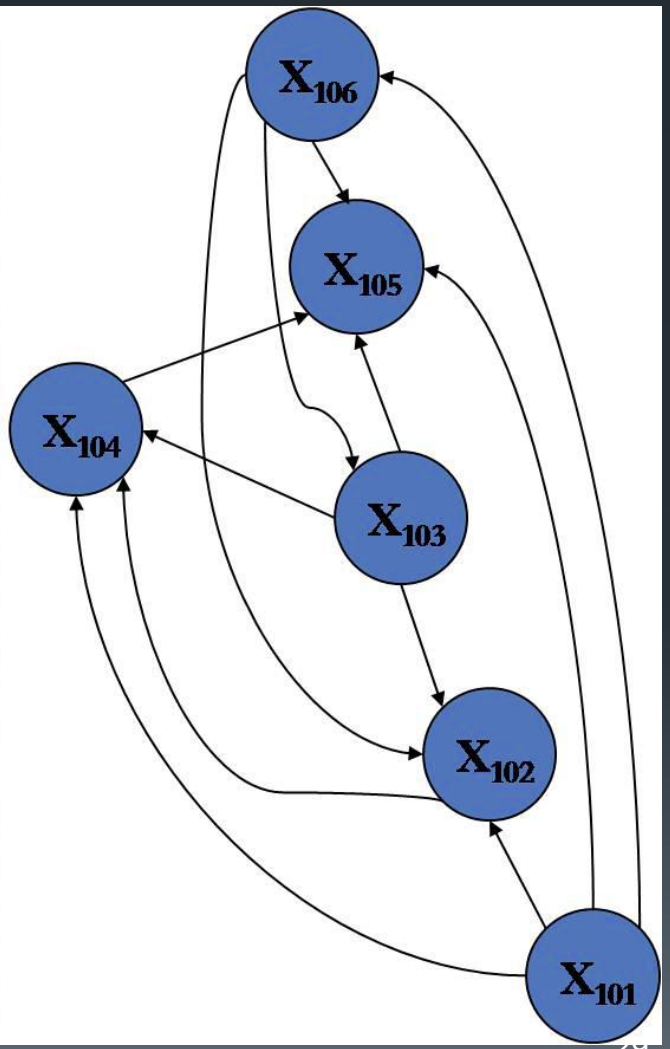




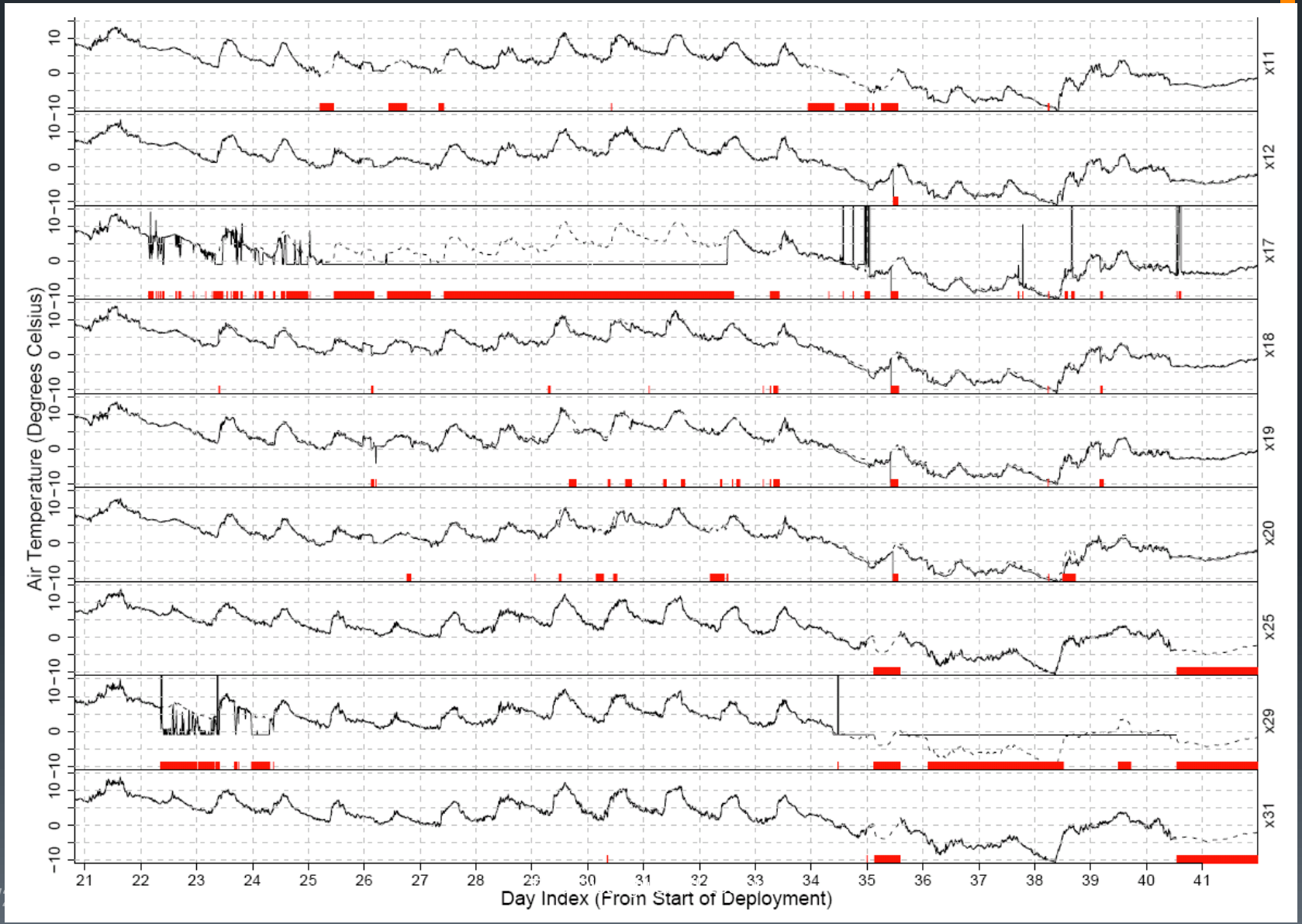
Multiple Sensor Application (EPFL, Switzerland)

Learned Network Structure





Multi-Sensor Anomaly Results



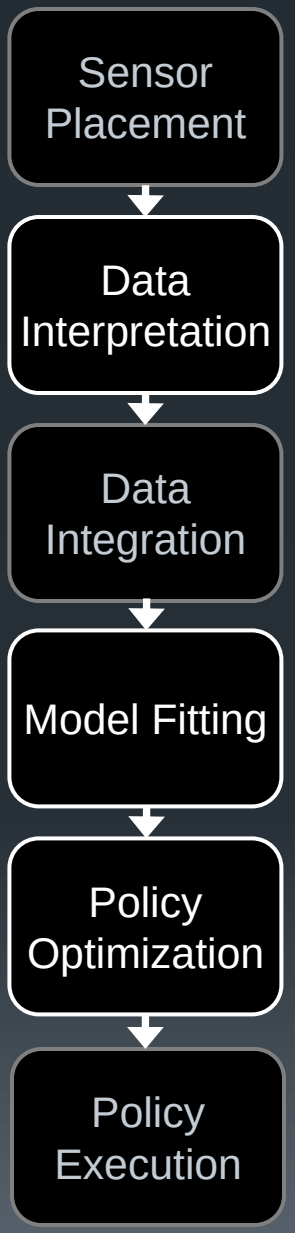
Additional Challenges

- Anomaly detection should operate at multiple time scales. How?

- Integrating heterogeneous sensors
  - Solar radiation
  - Wind speed and direction
  - Precipitation
  - Snow depth

Outline:
Three Projects at Oregon State

- Data Interpretation
  - Automated Data Cleaning
- Model Fitting
  - Explicit Observation Models
- Policy Optimization
  - Managing Fire in Eastern Oregon



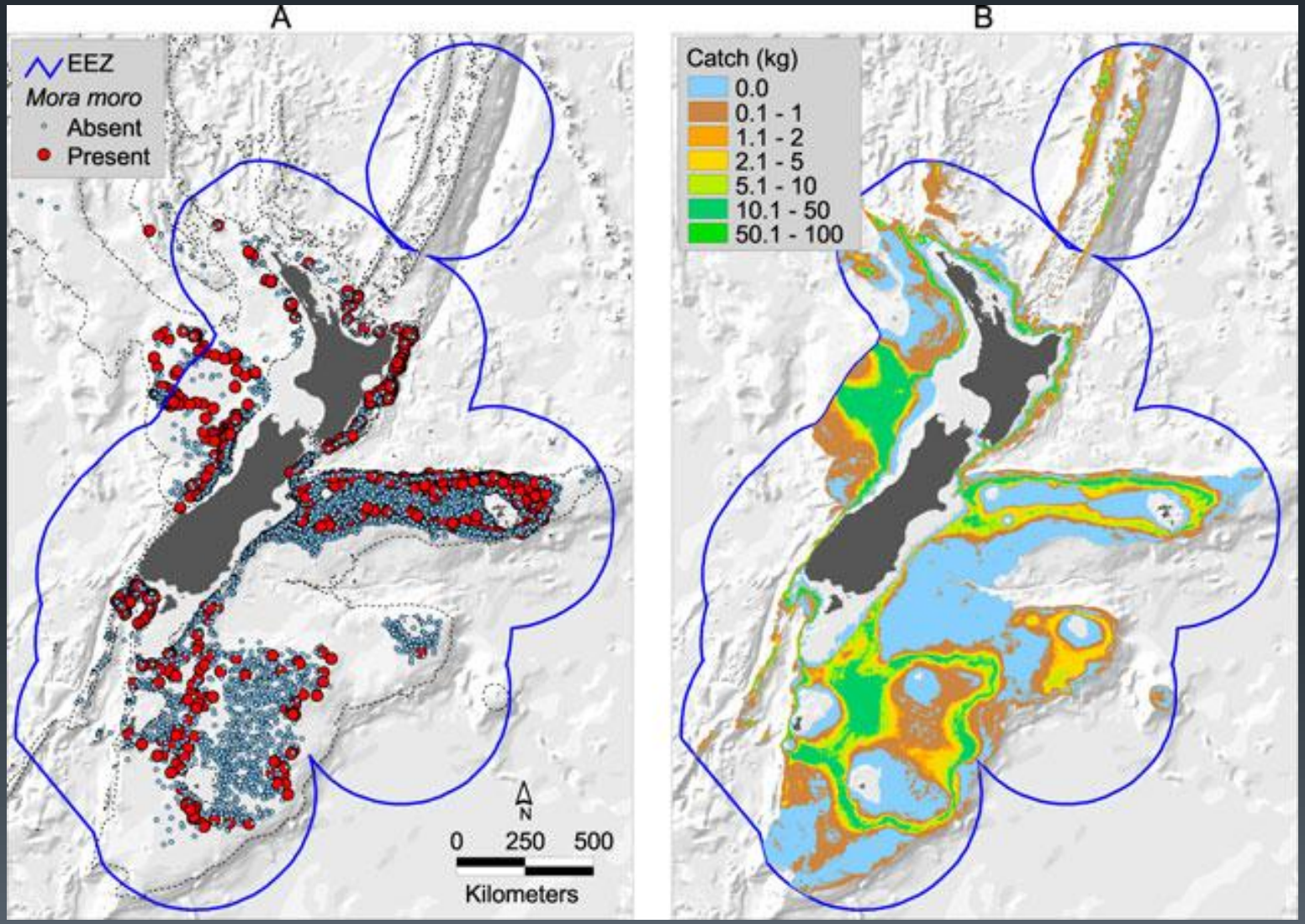
Flexible Species Distribution Modeling For Imperfect Detection

- Rebecca Hutchinson (PhD CMU 2009)
- Finishing Postdoc 6/30/2012

Species Distribution Modeling

Observations

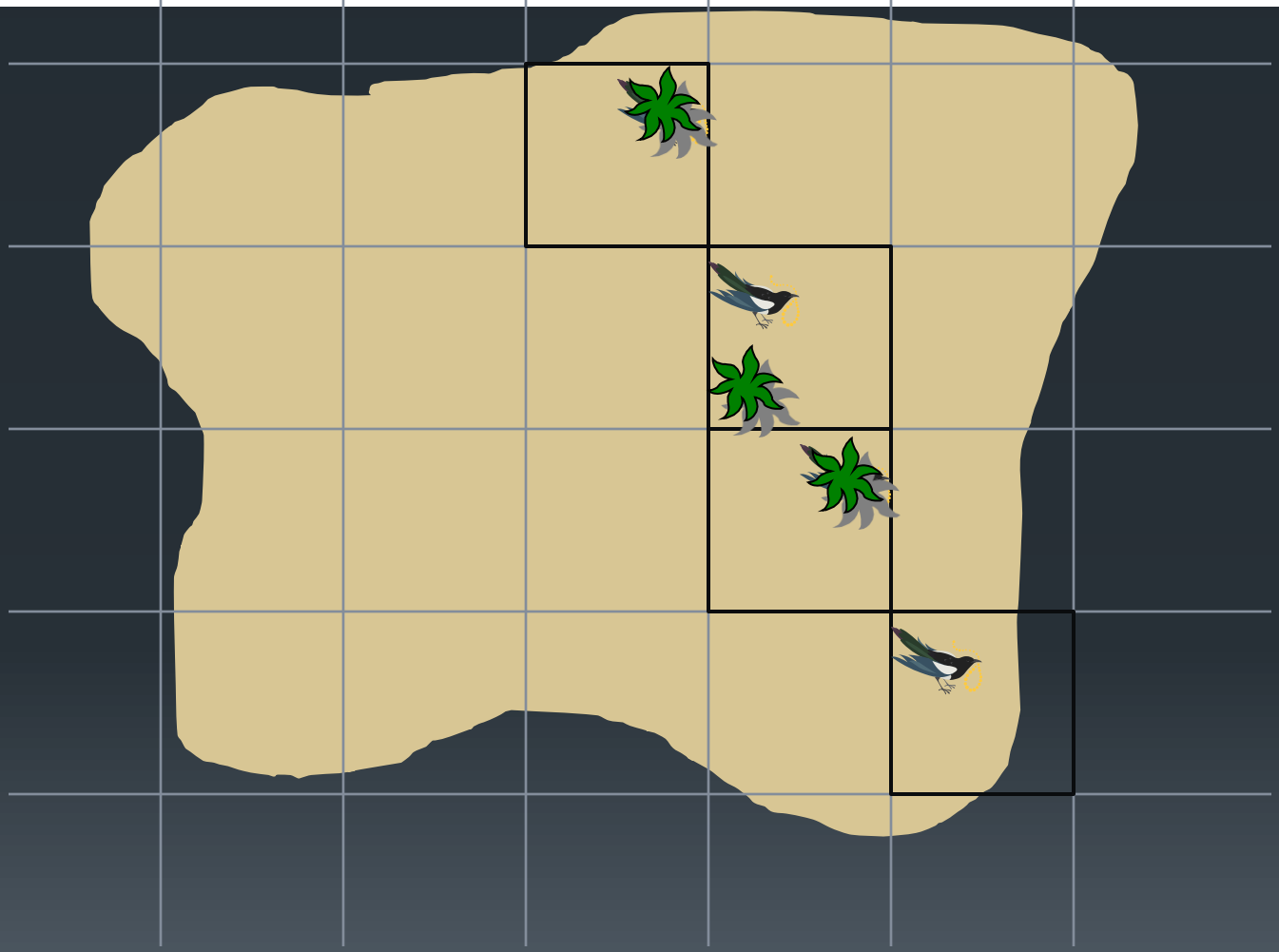
Fitted Model



Leathwick et al, 2008

Imperfect Detection

Pai Problem: Some birds are hidden ent birds hide on different visits

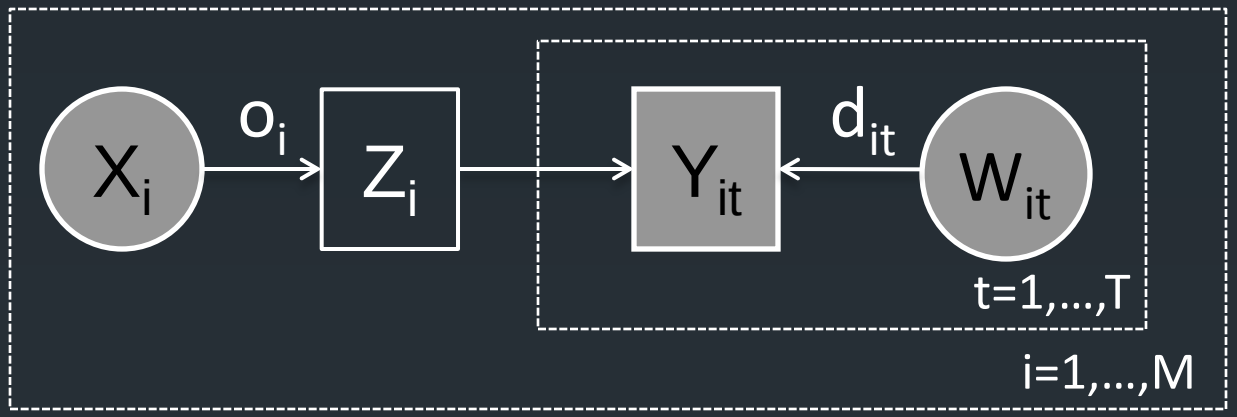


Multiple Visit Data

		Detection History		
Site	*True occupancy (latent)*	Visit 1 (rainy day, 12pm)	Visit 2 (clear day, 6am)	Visit 3 (clear day, 9am)
A (forest, elev=400m)	*1*	0	1	1
B (forest, elev=500m)	*1*	0	1	0
C (forest, elev=300m)	*1*	0	0	0
D (grassland, elev=200m)	*0*	0	0	0

Occupancy-Detection Model

MacKenzie, et al, 2006



$z_i \sim P(z_i|x_i)$: Species Distribution Model

$P(z_i = 1|x_i) = o_i = F(x_i)$ "occupancy probability"

$y_{it} \sim P(y_{it}|z_i, w_{it})$: Observation model

$P(y_{it} = 1|z_i, w_{it}) = z_i d_{it}$

$d_{it} = G(w_{it})$ "detection probability"

The Power of Probabilistic Graphical Models

- Probabilistic graphical models have many advantages
  - Excellent language for representing models
  - Learning and reasoning via probabilistic inference
  - Support hidden (latent) variables

- However, they have disadvantages
  - Designer must choose the parametric form of each probability distribution
  - Must decide on the number and form of interactions
  - Data must be scaled and transformed to match model assumptions
  - Somewhat difficult to adapt the complexity of the model to the amount and complexity of the data

Important Contribution of Machine Learning: Flexible Models

- Classification and Regression Trees
  - Require no model design
  - Require no data preprocessing or transformation
  - Automatically discover interactions as needed
  - Achieve high accuracy via boosting
- Support Vector Machines
  - Still require data preprocessing and transformation
  - Powerful methods for tuning model complexity automatically

Combining Probabilistic Graphical Models with Flexible Models

- Major open problem in machine learning

- Current efforts:
  - Kernel (SVM) methods for computing with probability distributions
  - Bayesian Non-Parametric Models: Dirichlet process mixture models
  - Our approach: Boosted regression trees

Flexible Occupancy-Detection Models

- Recall:
  - $F(X_i) = o_i$ is the occupancy probability
  - $G(W_{it}) = d_{it}$ is the detection probability
- Standard approach
  - Represent F and G as logistic regression models
- Our Idea:
  - Represent F and G using boosted regression trees
  - Learn them via boosting
  - This can be done using functional gradient descent (Mason & Bartlett, 1999; Friedman, 2000; Dietterich, et al, 2008; Hutchinson & Dietterich, 2011)

Experiment

- Algorithms:
  - Supervised methods:
    - S-LR: logistic regression from $(x_i, w_{it}) \rightarrow y_{it}$
    - S-BRT: boosted regression trees $(x_i, w_{it}) \rightarrow y_{it}$
  - Occupancy-Detection methods:
    - OD-LR: F and G logistic regressions
    - OD-BRT: F and G boosted regression trees
- Data:
  - 12 bird species
  - 3 synthetic species
  - 3124 observations from New York State, May-July 2006-2008
  - All predictors rescaled to zero mean, unit variance

Synthetic Species

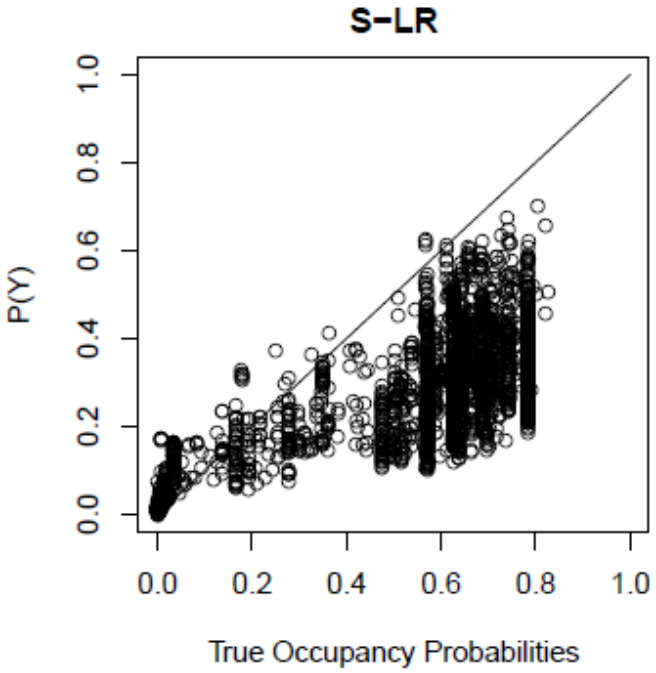
- Synthetic Species 2: F and G nonlinear

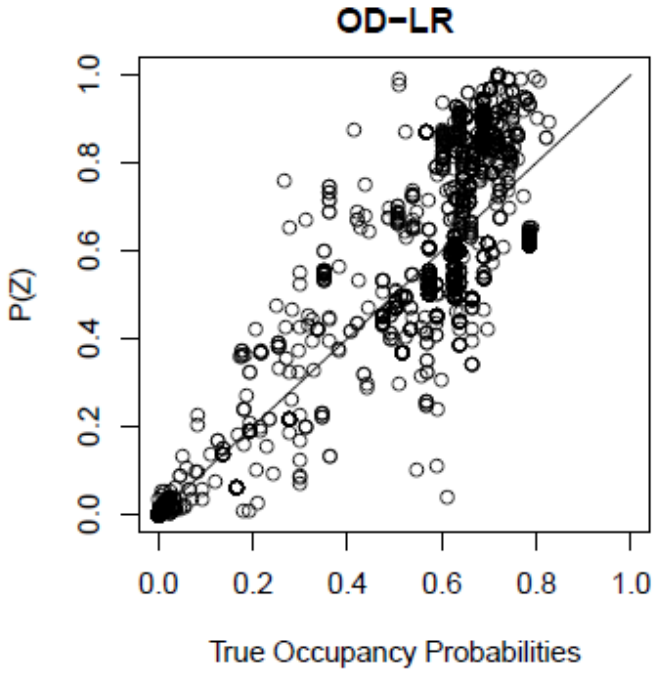
$$o_i \propto \exp\left(-2\left[x_i^{(1)}\right]^2 + 3\left[x_i^{(2)}\right]^2 - 2x_i^{(3)}\right)$$

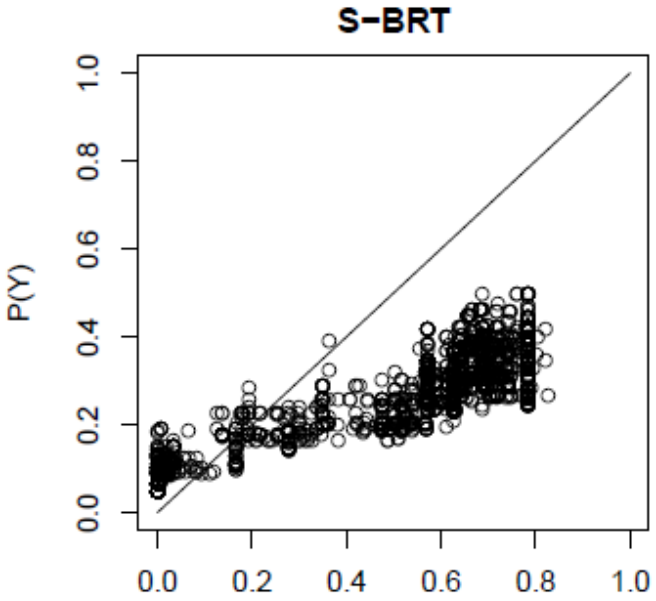
$$d_{it} \propto \exp\left(\exp\left(-0.5w_{it}^{(4)}\right) + \sin\left(1.25w_{it}^{(1)} + 5\right)\right)$$

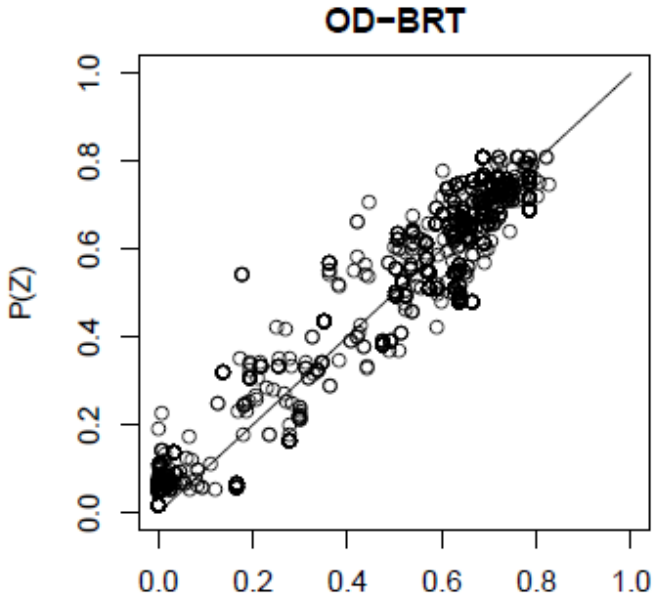
Predicting Occupancy

Synthetic Species 2



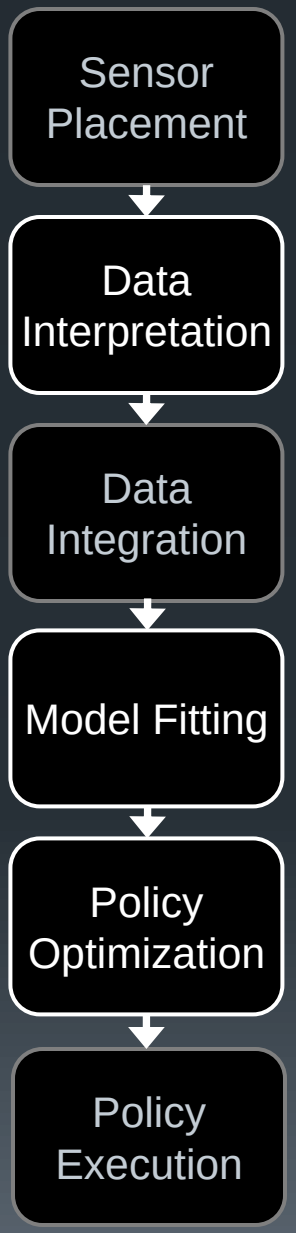






Outline:
Three Projects at Oregon State

- Data Interpretation
  - Automated Data Cleaning
- Model Fitting
  - Explicit Observation Models
- Policy Optimization
  - Managing Fire in Eastern Oregon

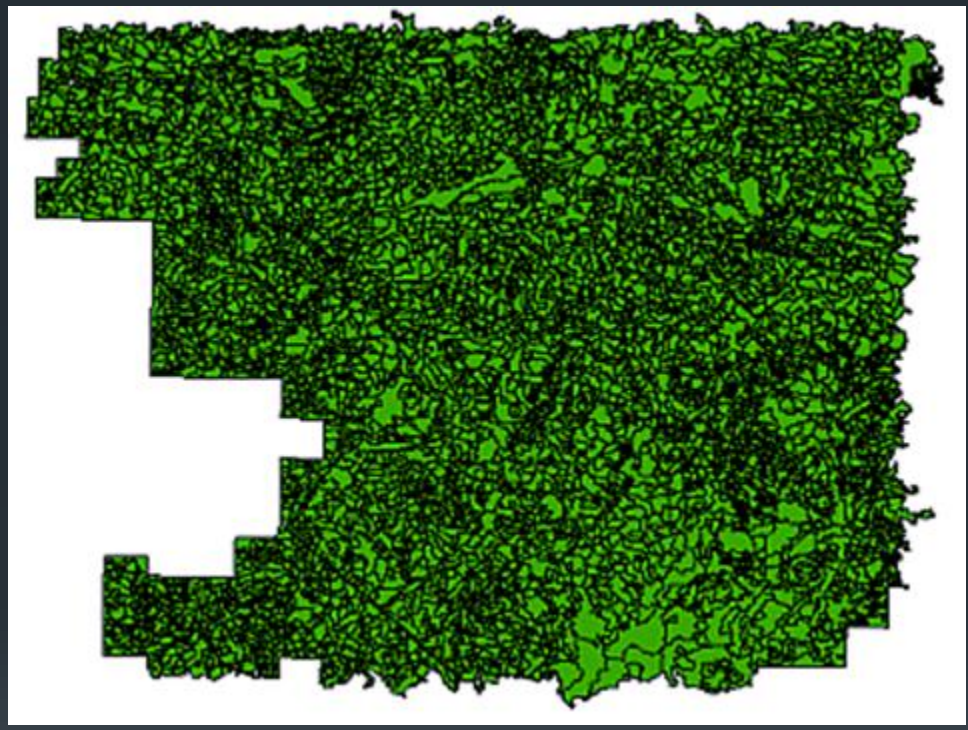


Managing Wildfire in Eastern Oregon

- Natural state (hypothesized):
  - Large Ponderosa Pine trees with open understory
  - Frequent “ground fires” that remove understory plants (grasses, shrubs) but do not damage trees
- Fires have been suppressed since 1920s
  - Large stands of Lodgepole Pine
  - Heavy accumulation of fuels in understory
  - Large catastrophic fires that kill all trees and damage soils
  - Huge firefighting costs and lives lost

Study Area: Deschutes National Forest

- ~4000 Management Units
- Goal: Return the landscape to its “natural” fire regime
- Which management units should be treated each year?



Game Against Nature

- For each time step t
  - Our turn:
    - Observe current state s_t (i.e., state of all MUs)
    - Choose action vector a_t
    - Execute the actions in the MUs
  - Nature's turn:
    - Stochastically ignite and burn fires on the landscape (Implemented by ignition model + fire spread model)
    - Grow trees and fuel (Implemented by forest growth model)

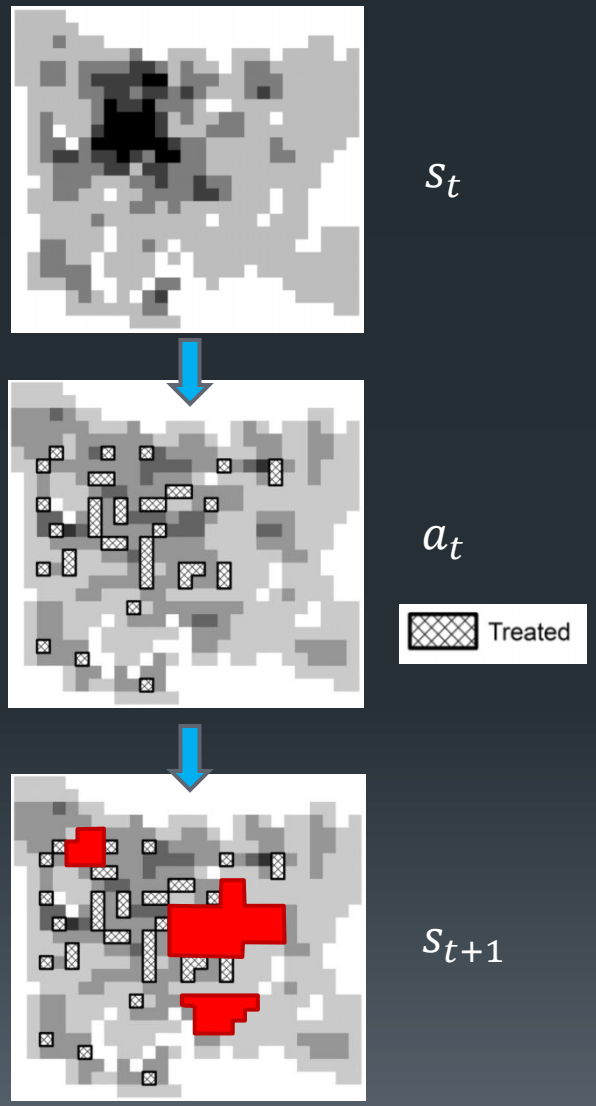
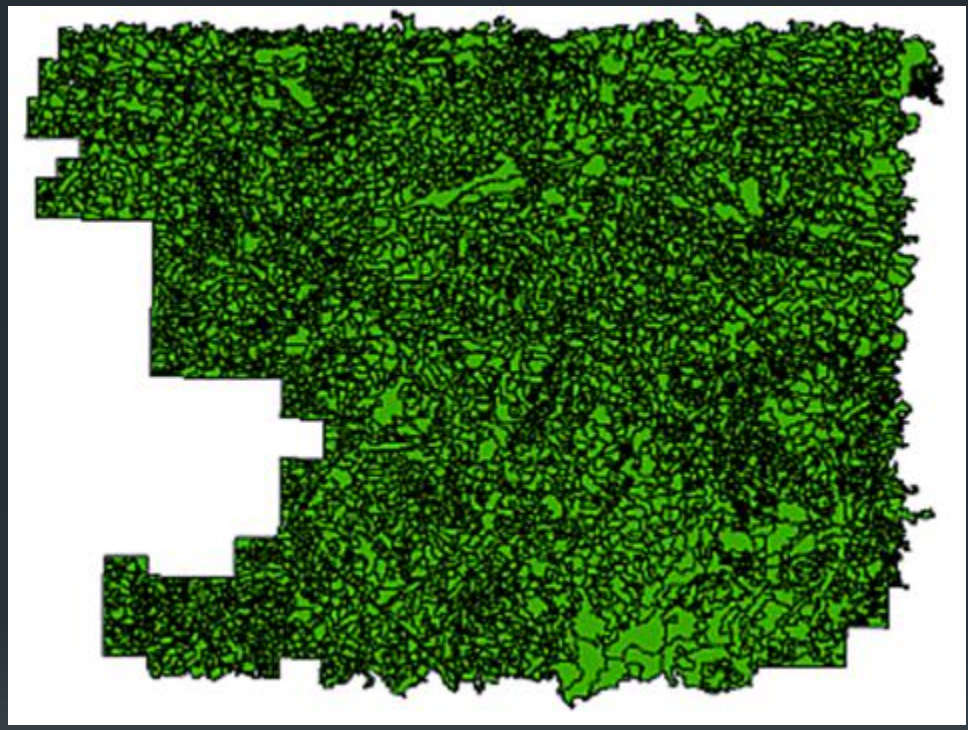


Image: Wei et al, 2008

Formulation as a Markov Decision Process

- State of each MU:
  - Age of trees
    - {0-9, 10-19, 20-29, 30-39, 40-49}
  - Amount of fuel
    - {none, low, medium, high, very high}
  - 25 possible combinations
  - 25^{4000} possible states for the landscape
- Actions in each MU each decade
  - Do nothing
  - Fuel treatment (costs money)
  - Harvest trees (makes money, but increases fuel)
  - Harvest + Fuel
  - 4^{4000} possible actions over landscape



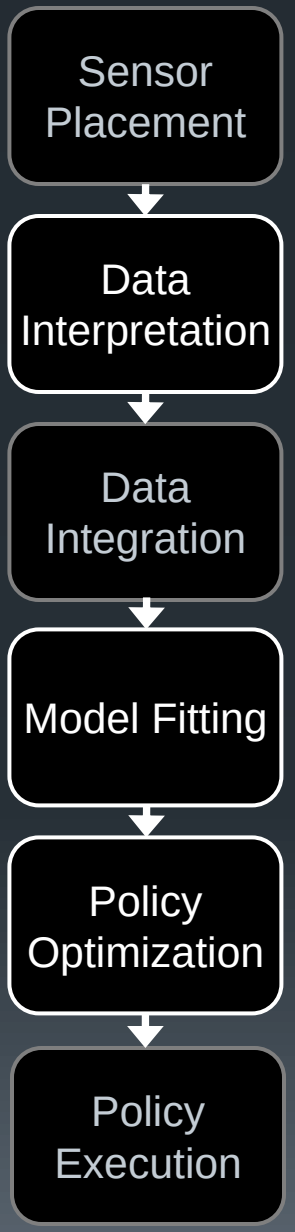
Study area in Deschutes National Forest

Open Problem: Solving This MDP

- One-shot Method [Wei, et al., 2008]
  - Run 1000s of simulated fires to generate fire risk map and fire propagation graph
  - Formulate and solve Mixed Integer Program to compute optimal one-shot solution
- Challenge:
  - Develop methods that can solve the MDP over long time horizons

Summary

- Data Interpretation
  - Automated Data Cleaning
- Model Fitting
  - Explicit Observation Models
- Policy Optimization
  - Managing Fire in Eastern Oregon



Computational Sustainability

- There are many opportunities for computing to contribute to a sustainable planet
- There are many challenging computer science research problems to be solved
- CCC is sponsoring Computational Sustainability tracks at leading conferences this coming year including ICML and AAAI
- Institute for Computational Sustainability: http://www.computational-sustainability.org/

Thank-you

- Ethan Dereszynski: Automated Data Cleaning
- Rebecca Hutchinson: Boosted Regression Trees in OD models
- Claire Montgomery, Rachel Houtman, and Sean McGregor: Fire challenge

- National Science Foundation Grants 0705765, 0832804, and 0905885