Machine Learning and Computational Sustainability

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. Optimal management requires spatial planning over horizons of 100+ years

Computer Science can help!

1. Lack of knowledge of function and structure
2. Spatial planning

Sensors

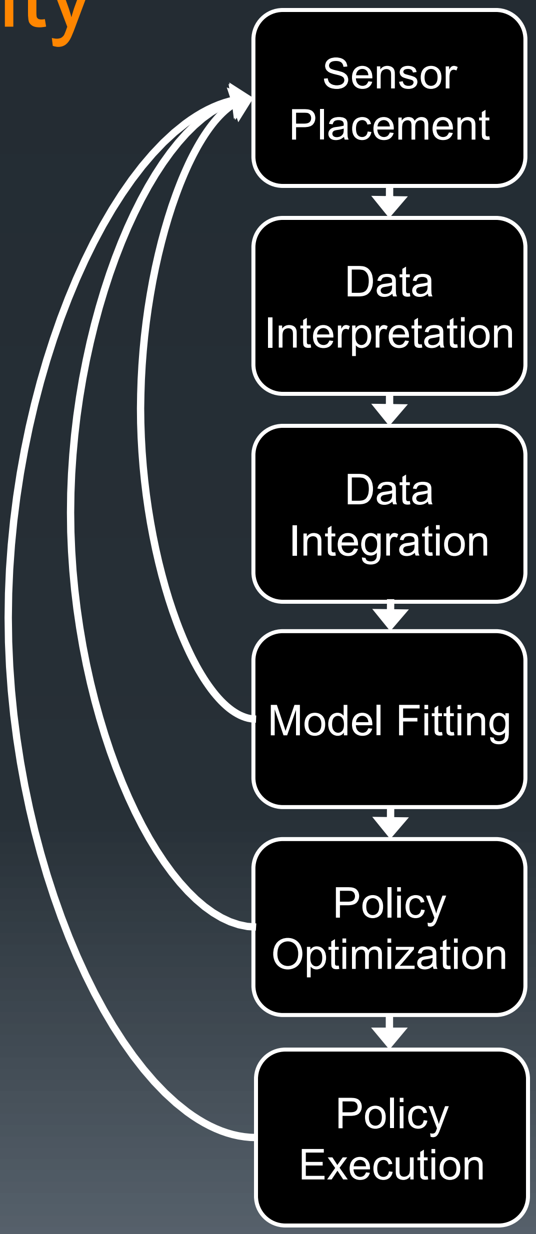
Machine Learning

Optimization

Computational Sustainability

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

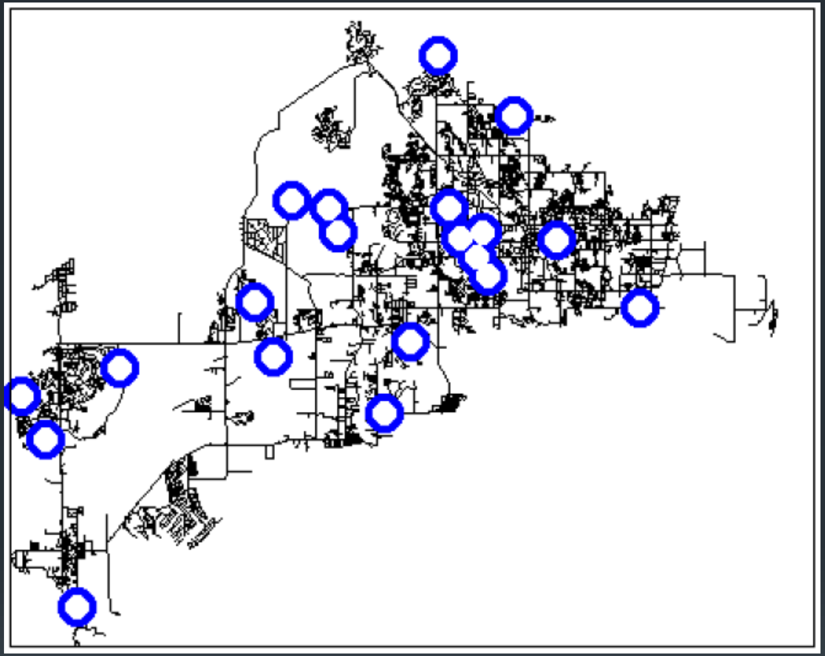
- Data → Models → Policies



Example Research Efforts

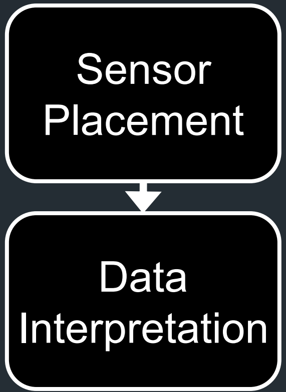
Sensor Placement

- 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

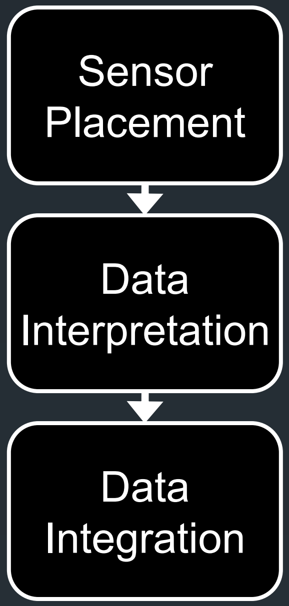


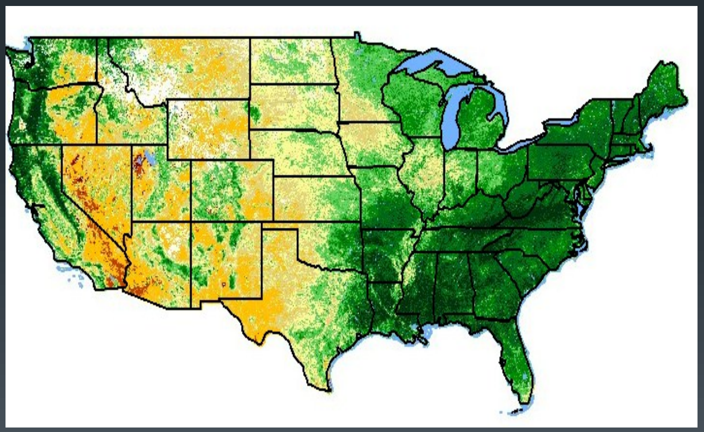
- Insect identification for population counting
- Raw data: image
- Interpreted data: Count by species

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

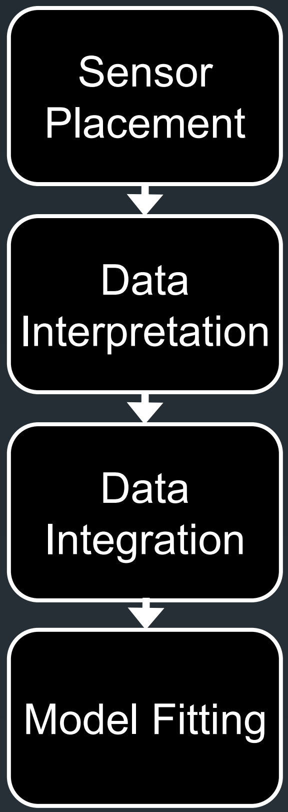




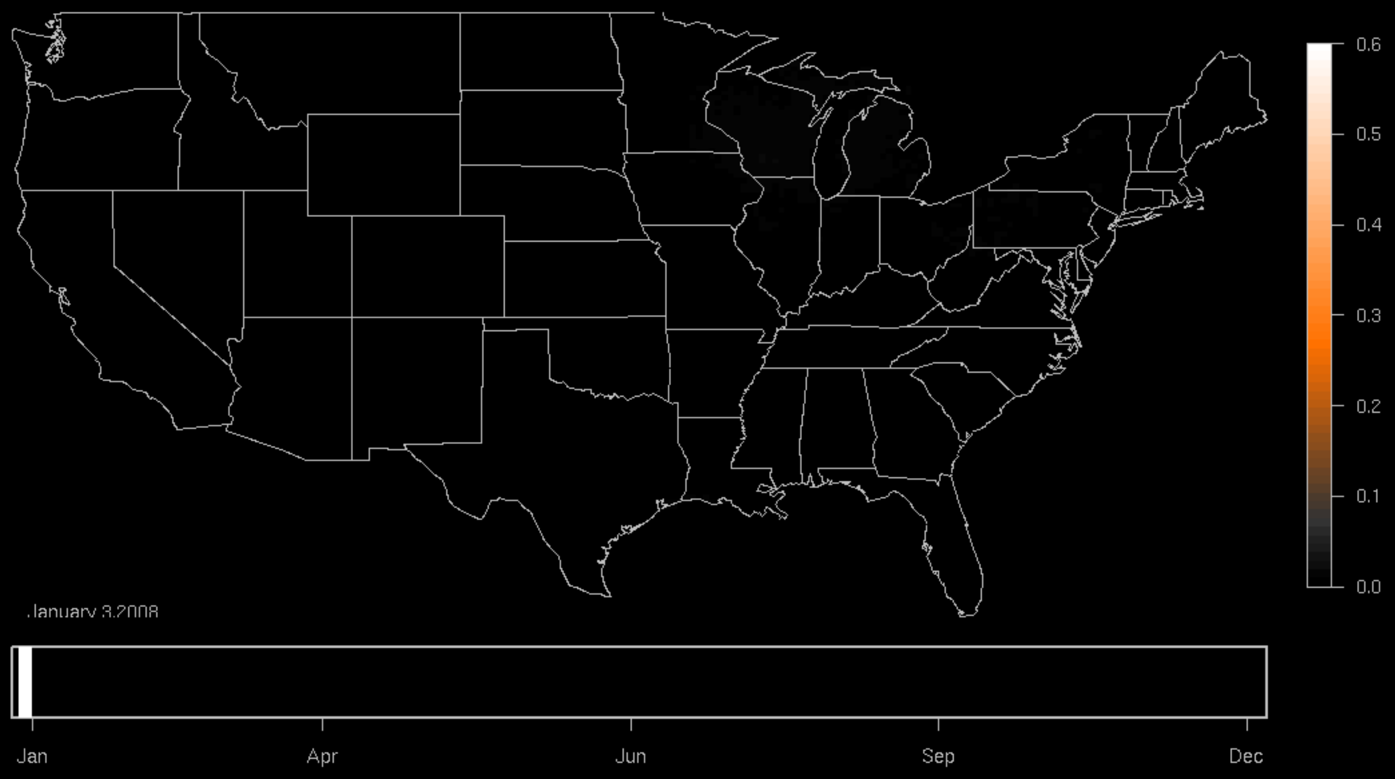
Landsat NDVI:
http://ivm.cr.usgs.gov/viewer/

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

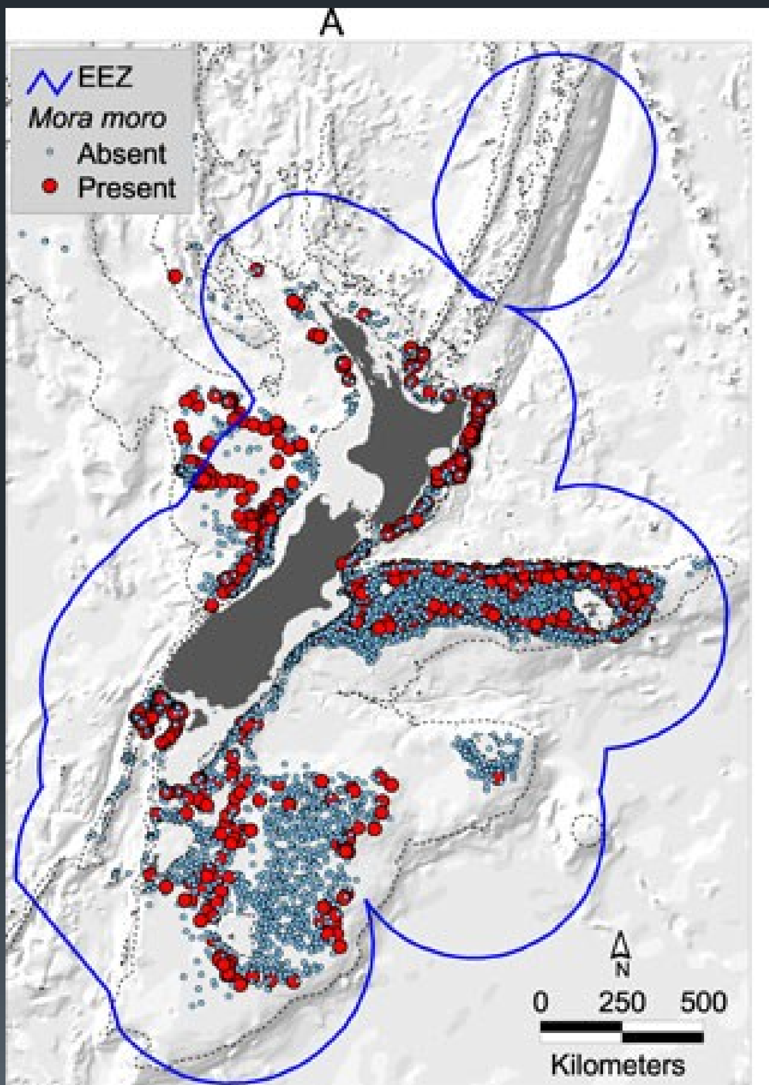


Indigo Bunting

slide courtesy of Daniel Fink

Policy Optimization

Observations



Leathwick et al, 2008

Sensor Placement

Data Interpretation

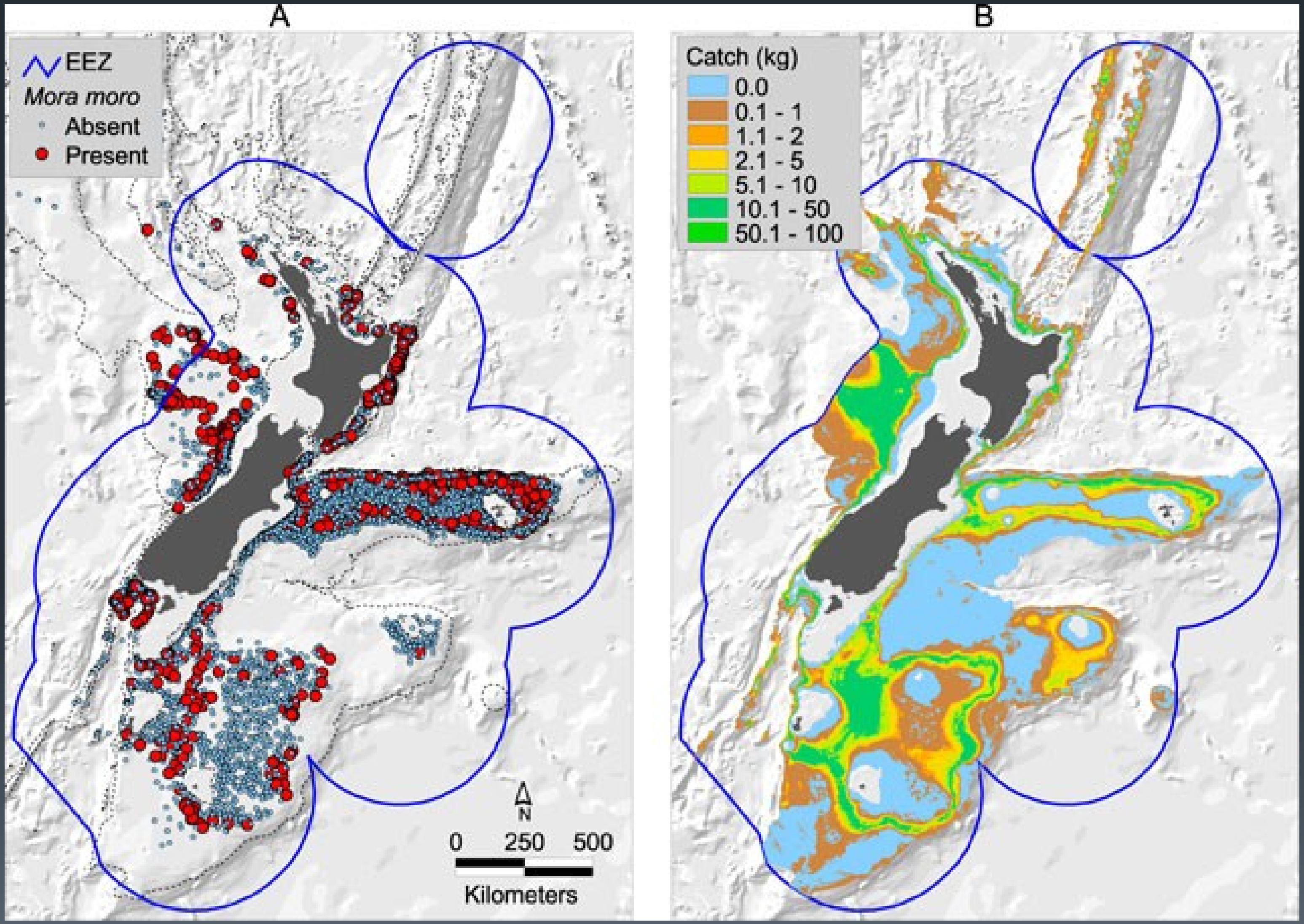
Data Integration

Model Fitting

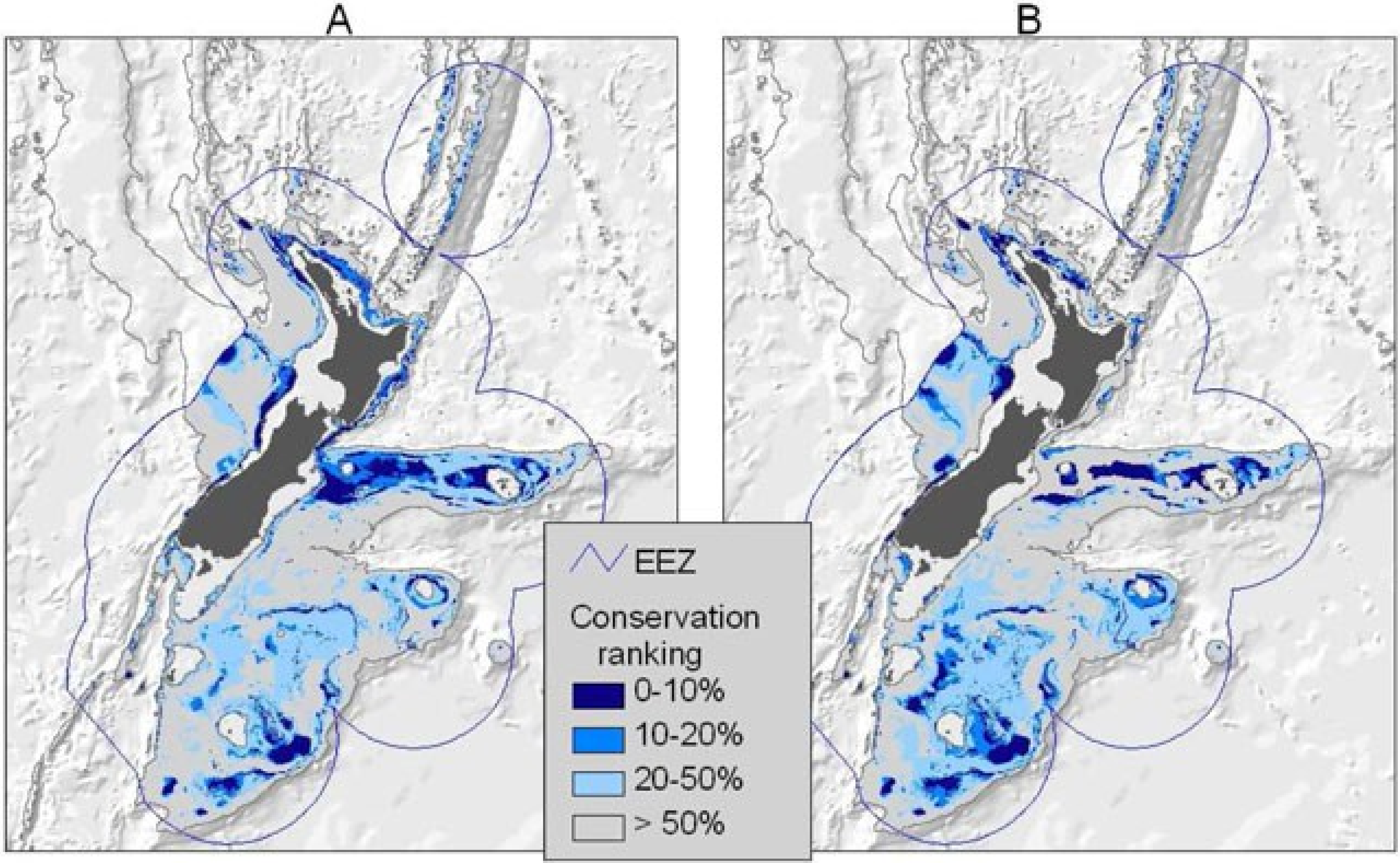
Policy Optimization

Policy Optimization

Observations | Fitted Model



Leathwick et al, 2008



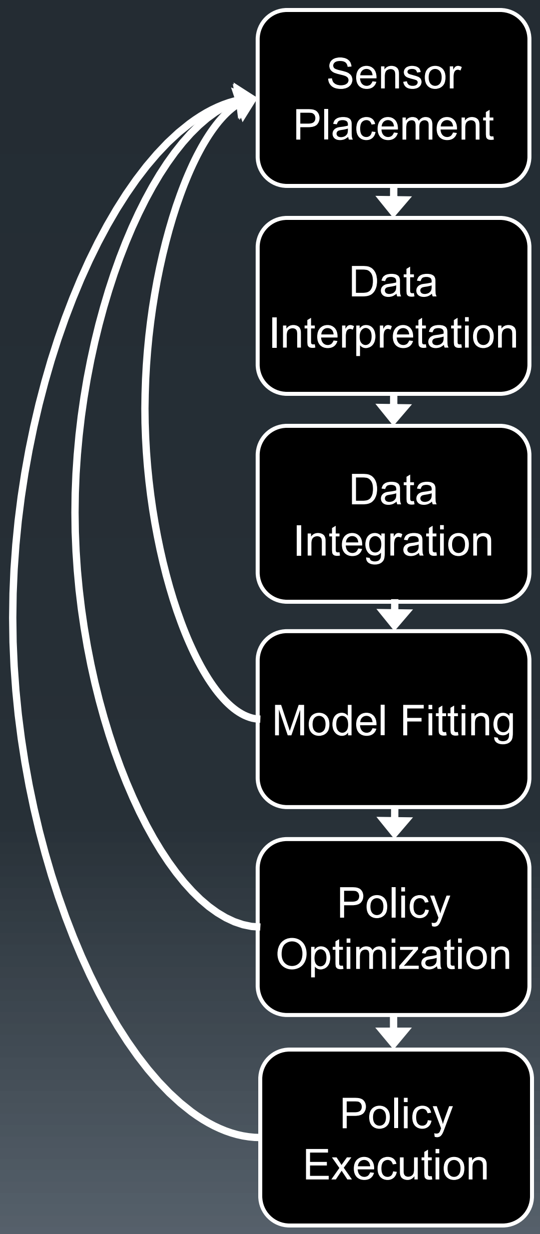
Disregarding costs to fishing industry

Full consideration of costs to fishing industry

Leathwick et al, 2008

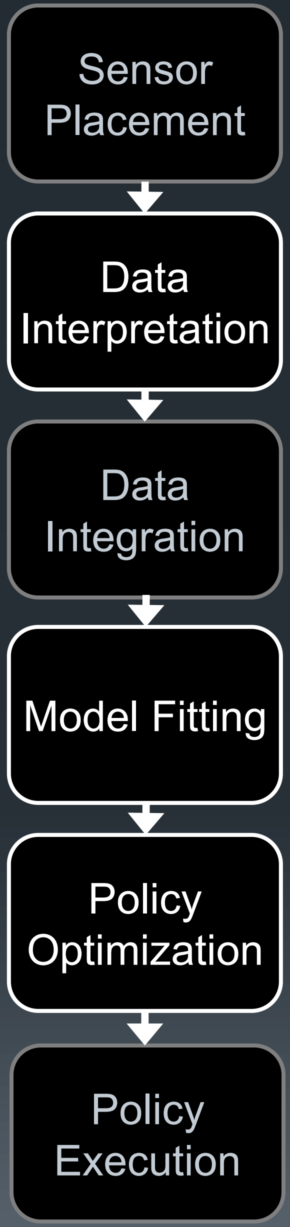
Policy Execution

- 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
  - Project TAHMO
- Model Fitting
  - Explicit Observation Models
  - Flexible Latent Variable Models
- Policy Optimization
  - Managing Fire in Eastern Oregon
  - Algorithms for Large Spatial MDPs



Project TAHMO
20,000 hydro-met stations for Africa

- Africa is very poorly sensed
  - Only a few dozen weather stations reliably report data to WMO (blue points in map)

- Project TAHMO (tahmo.org)
  - TU-DELFT & Oregon State University
  - Design a complete hydrology/meteorology sensor station at a cost of EUR 200
  - Deploy 20,000 such stations across Africa

Project TAHMO
20,000 hydro-met stations for ~~Africa~~ South America ??

- South America is also very poorly sensed

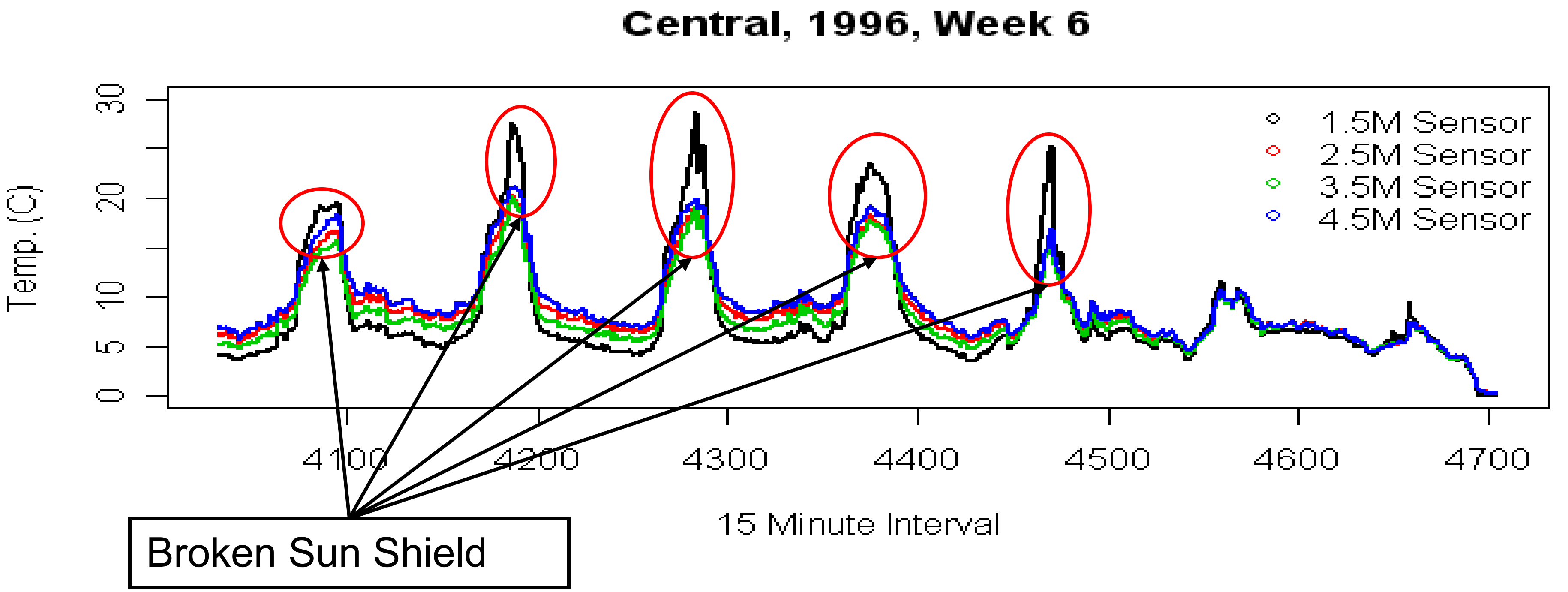
Challenges

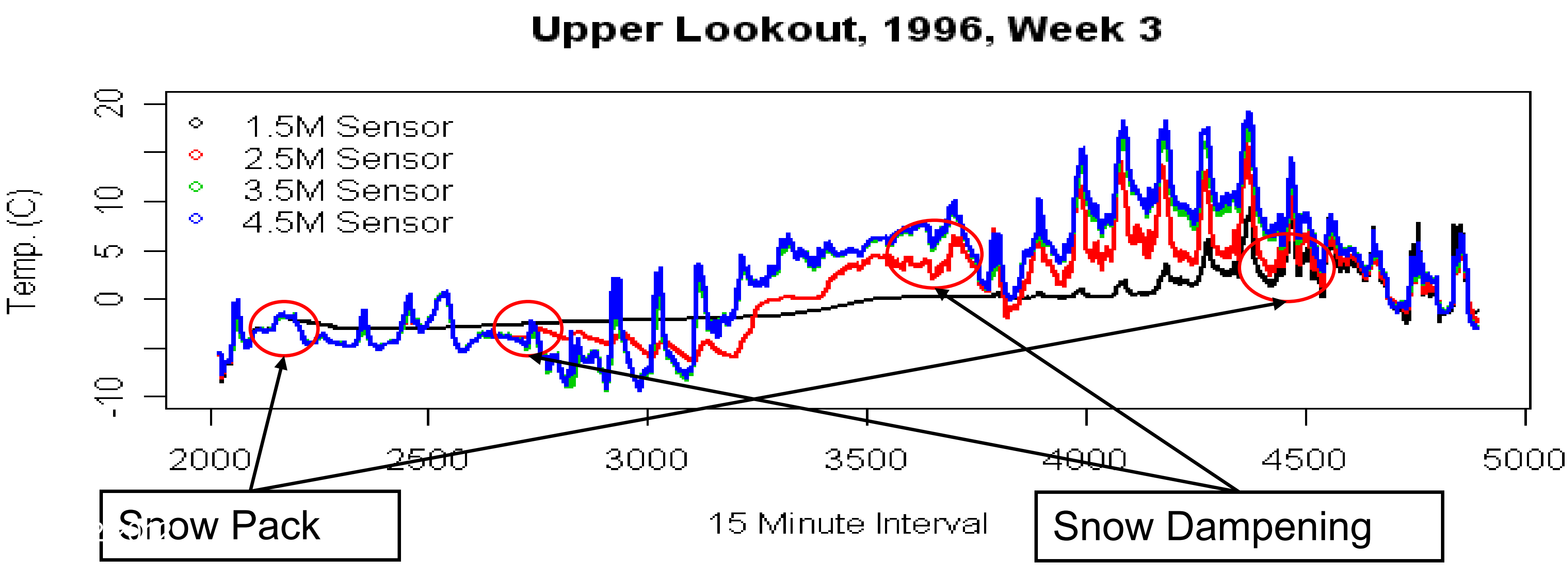
- Sensor Placement
  - Multiple criteria:
    - accuracy of reconstructing maps of
      - temperature, precipitation, solar radiation, wind speed and direction, relative humidity
    - accuracy of estimates of composite variables
      - Evapo-transpiration
    - robustness to sensor failure
    - accessibility and safety

- Continent-scale Data Quality Control
  - Sensors fail for infinitely many reasons
  - Detect failures and impute missing data

A Problem Closer to Oregon...

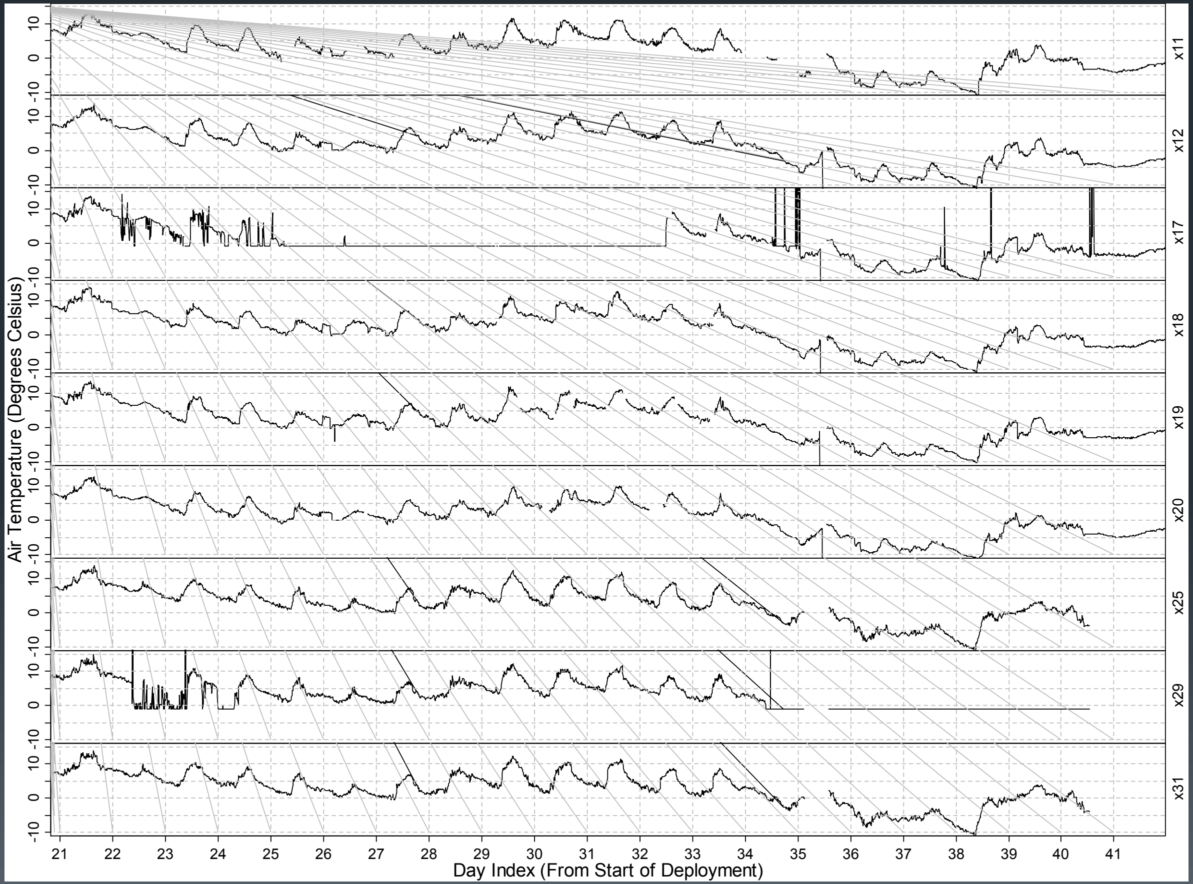
thermometers at 1.5, 2.5, 3.5, and 4.5m


Central, 1996, Week 6
Temp. (C)
0
5
10
20
30
4100
4200
4300
4400
4500
4600
4700
15 Minute Interval
1.5M Sensor
2.5M Sensor
3.5M Sensor
4.5M Sensor
Broken Sun Shield


Upper Lookout, 1996, Week 3
Temp. (C)
-10
0
5
10
20
2000
2500
3000
3500
4000
4500
5000
15 Minute Interval
1.5M Sensor
2.5M Sensor
3.5M Sensor
4.5M Sensor
Snow Pack
Snow Dampening

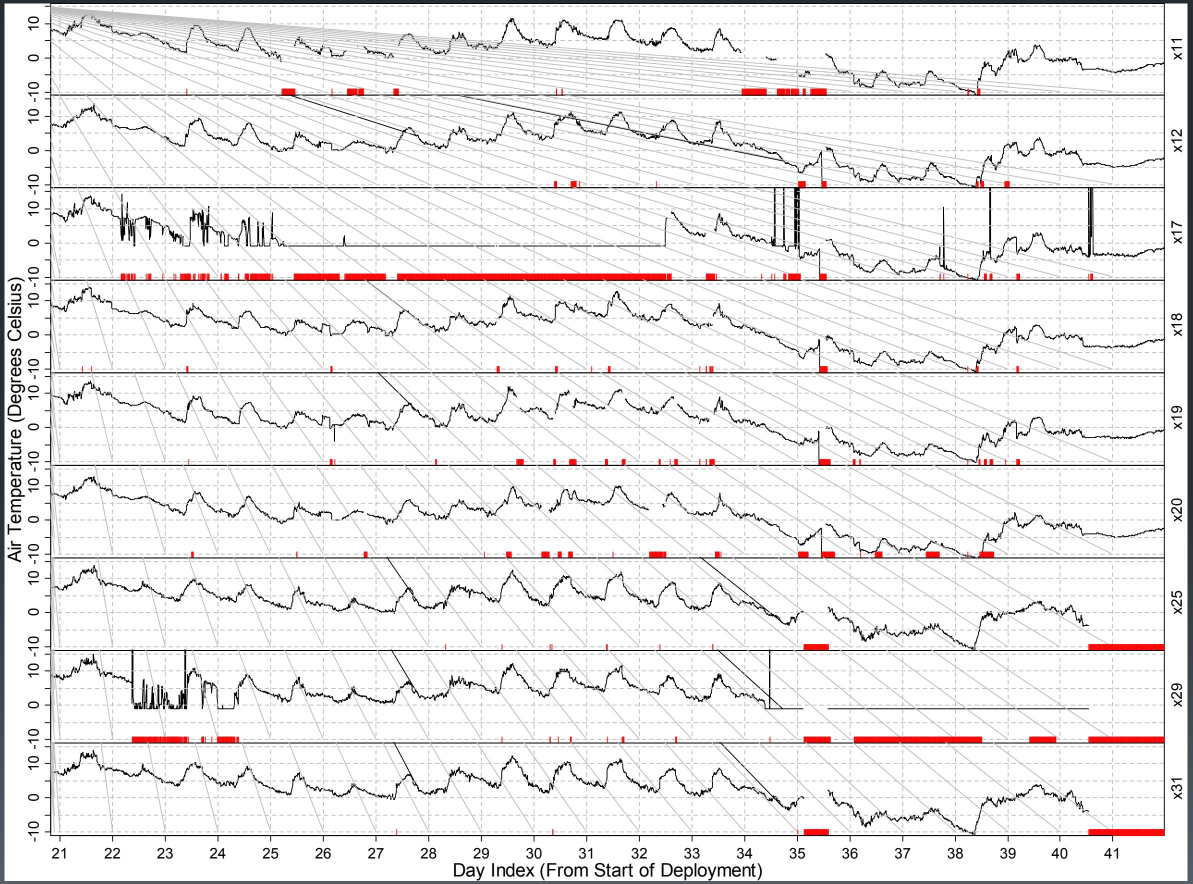
Functions of a Data Cleaning Method

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



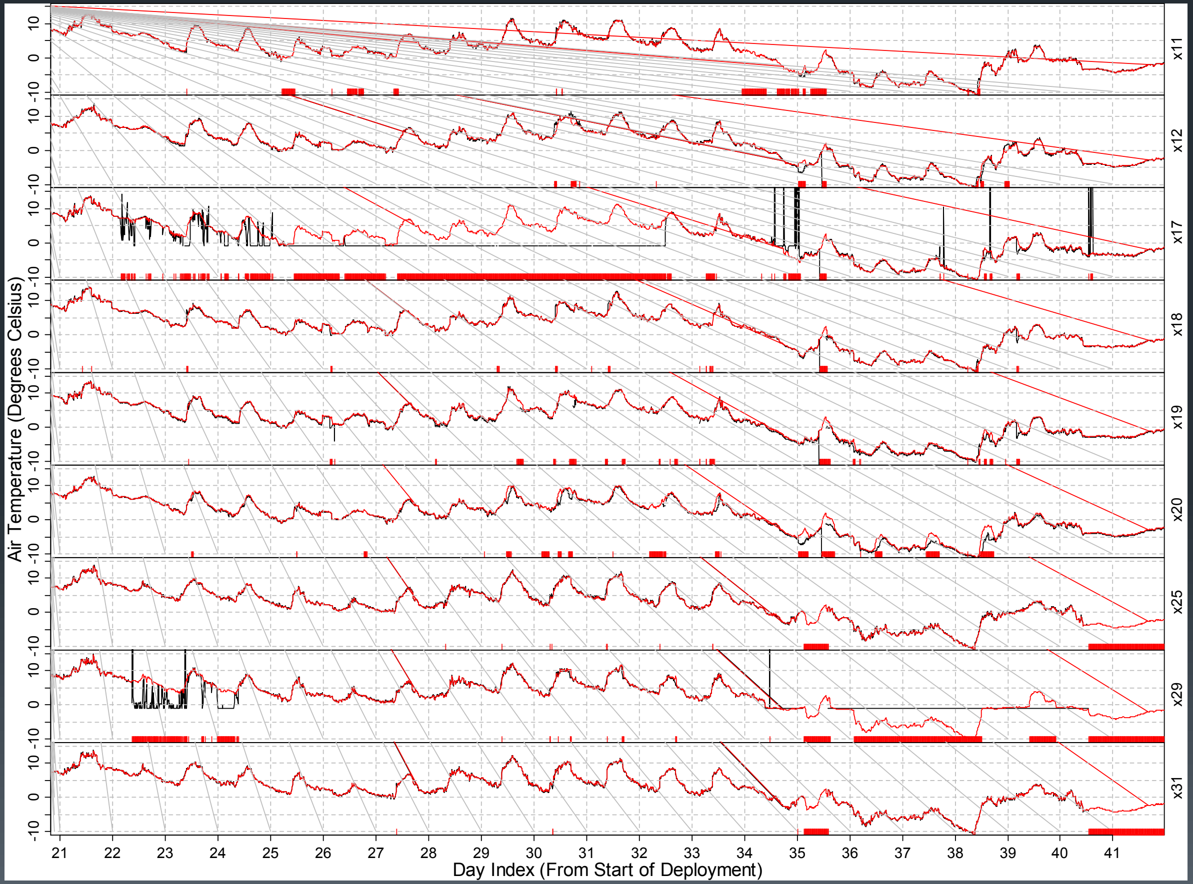
Functions of a Data Cleaning Method

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



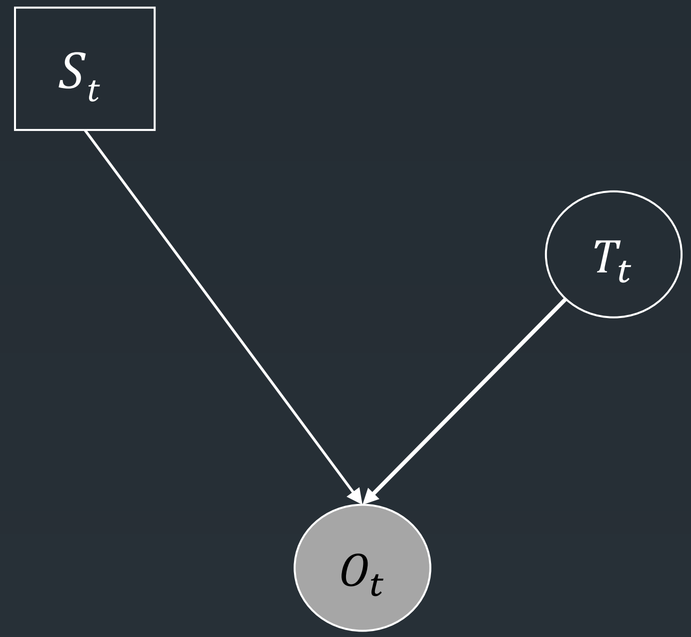
Functions of a Data Quality Control Method

- 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



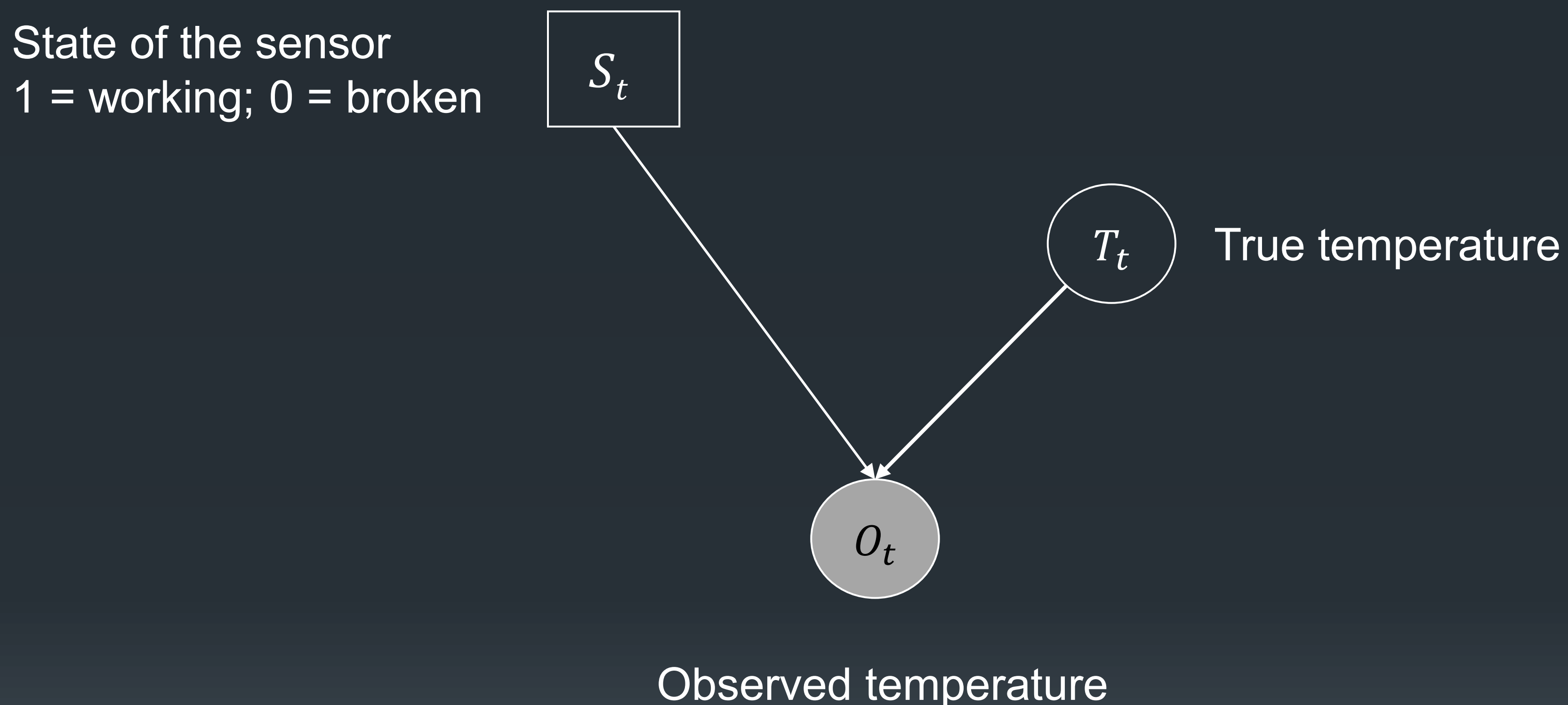
True temperature

Observed temperature

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

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

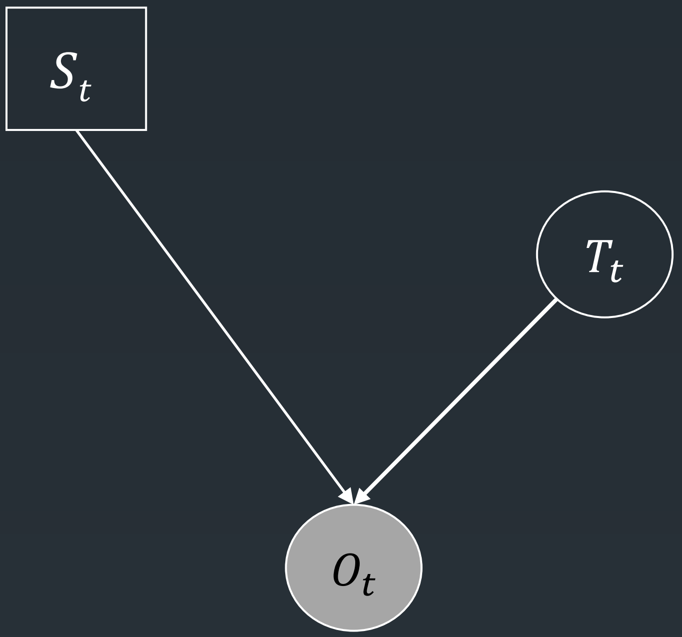
Anomaly Detection Via Probabilistic Inference



Query: What is the most likely value of S_t?

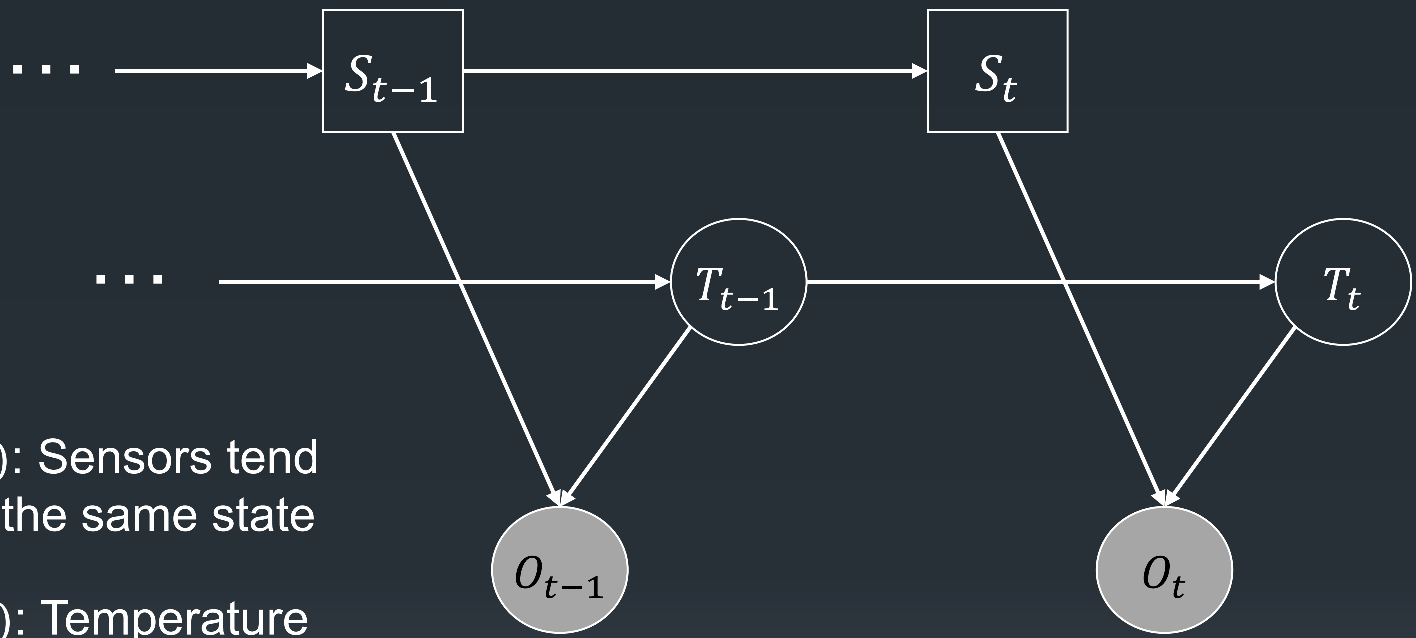
$$\underset{s}{\operatorname{argmax}} P(S_t = s | O_t)$$

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

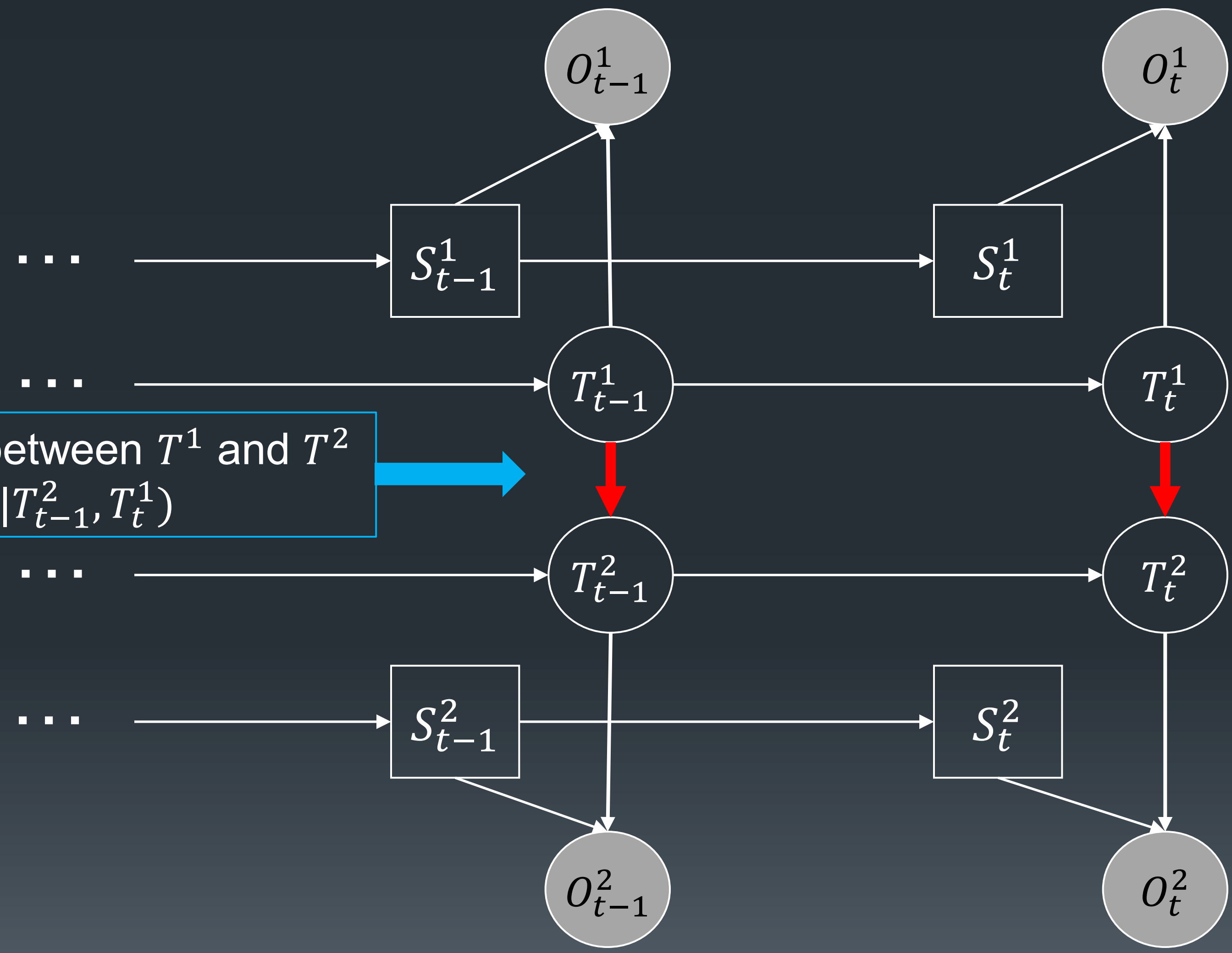


$P(S_t|S_{t-1})$: Sensors tend to stay in the same state

$P(T_t|T_{t-1})$: 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



Dependency between T^1 and T^2

$P(T^2_t | T^2_{t-1}, T^1_t)$

Probabilistic Inference is Infeasible in the Single Sensor Model

- Single sensor Markov model query: $\underset{S_t, S_{t-1}, \ldots}{\operatorname{argmax}} \; P(S_t, S_{t-1}, \ldots \mid O_t, O_{t-1}, \ldots)$
  - Requires time exponential in the length of the time series

- Solution:
  - Commit to each S_t in time order
  - $\hat{S}_1 \coloneqq \underset{s}{\operatorname{argmax}} \, P(S_1 = s \mid O_1)$
  - $\hat{S}_2 \coloneqq \underset{s}{\operatorname{argmax}} \, P(S_2 = s \mid \hat{S}_1, O_2)$
  - …
  - Also bound the variance of T_t

- Each of these inferences is easy

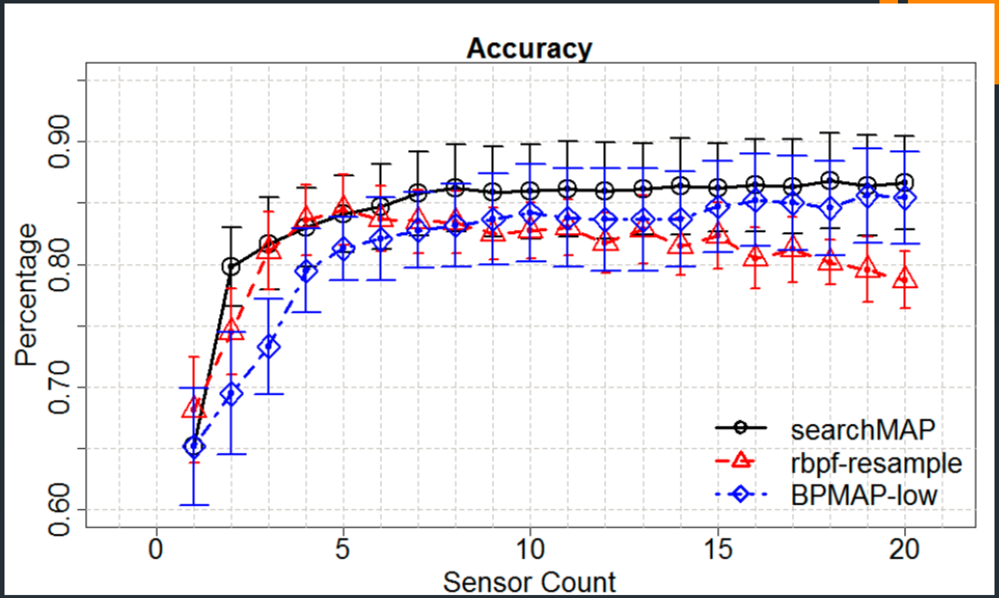
Probabilistic Inference is Infeasible in the Multiple Sensor Model

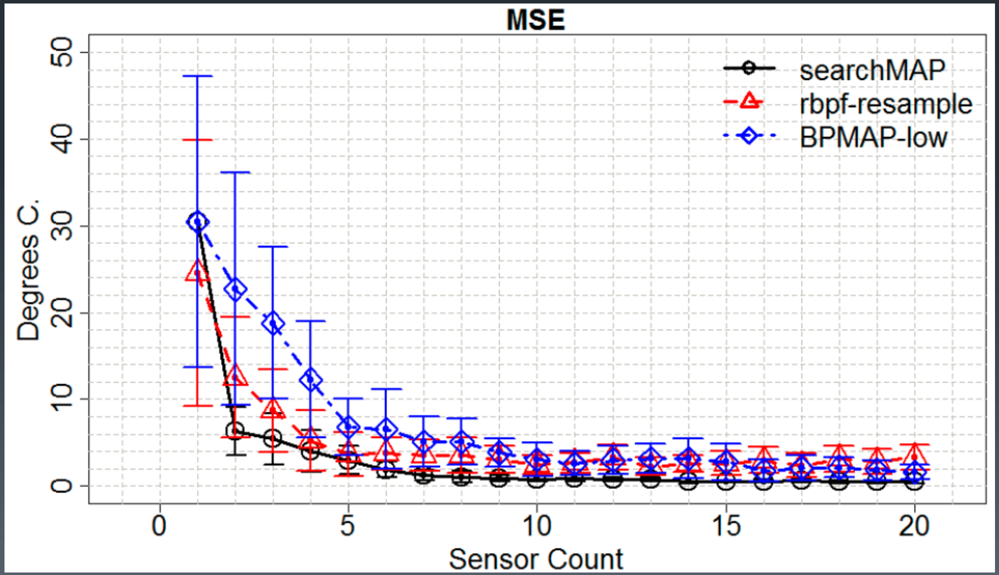
- Even if we commit to values for $S_t^1, S_t^2 \dots, S_t^K$ for K sensors, we must compute an intermediate data structure of size 2^K

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

Comparison of Approximate Inference Methods

- Faults injected into clean data
  - randomized spike, bias (offset), and flatline faults generated from a first-order Markov model
- Algorithms
  - Loopy BP MaxProduct (best of EP and BP-related methods)
  - Rao-Blackwellized particle filters
  - SearchMAP

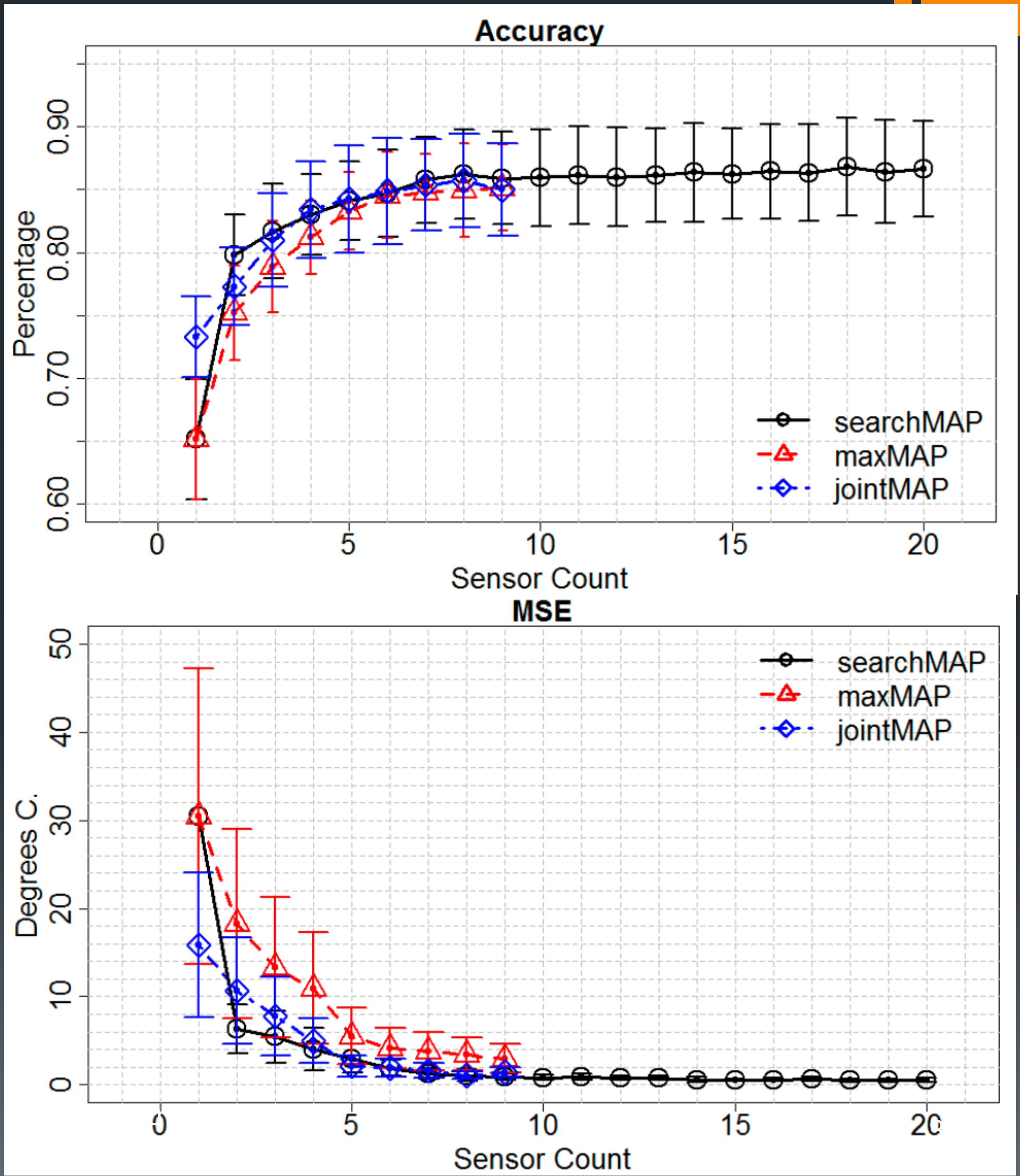




Approximate Inference with Many Sensors vs. Exact Inference with Fewer Sensors?

- For a given target sensor, order the remaining sensors by their mutual information to the target
- Exact (within time-step) inference is feasible for ≤ 9 sensors

- Conclusions:
  - searchMAP is better, even for < 9 sensors!
  - Its bias toward all sensors working seems to be slightly advantageous
  - only slight benefit of >9 sensors

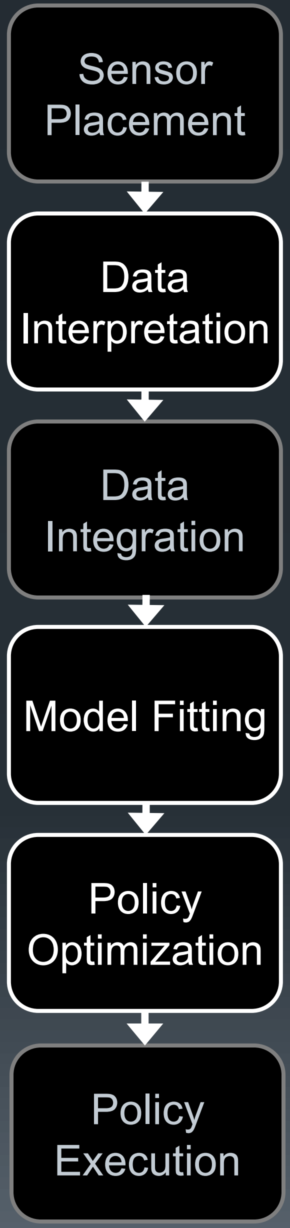


Next Steps

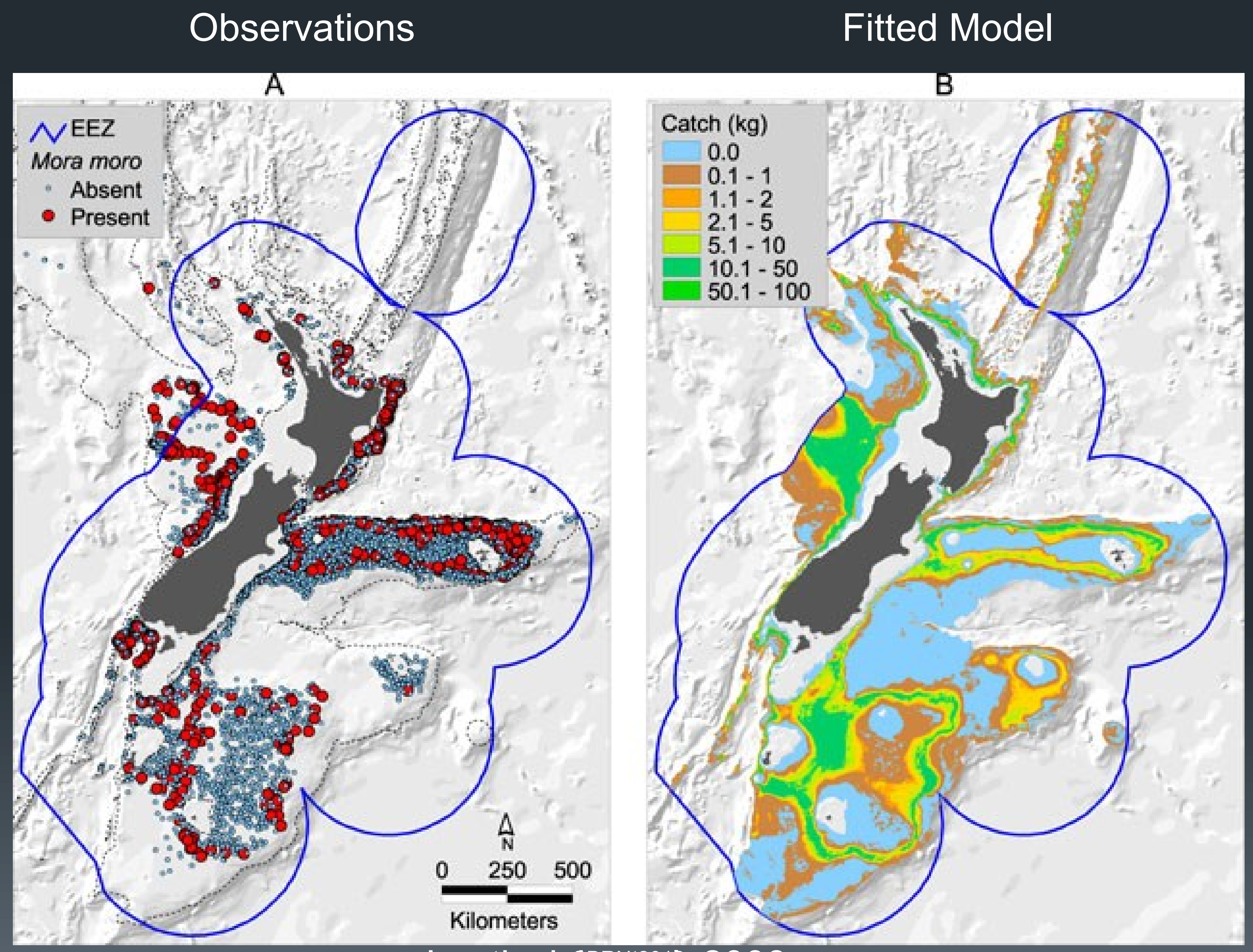
- Improved models for multiple types of sensors
  - temperature
  - precipitation
  - wind speed
  - wind direction
  - relative humidity
  - soil radiation
  - soil moisture
  - These are not jointly Gaussian!
- Methods that work at multiple spatial scales
  - continent scale

Outline:
Three Projects at Oregon State

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



Species Distribution Modeling



Leathwick et al, 2008

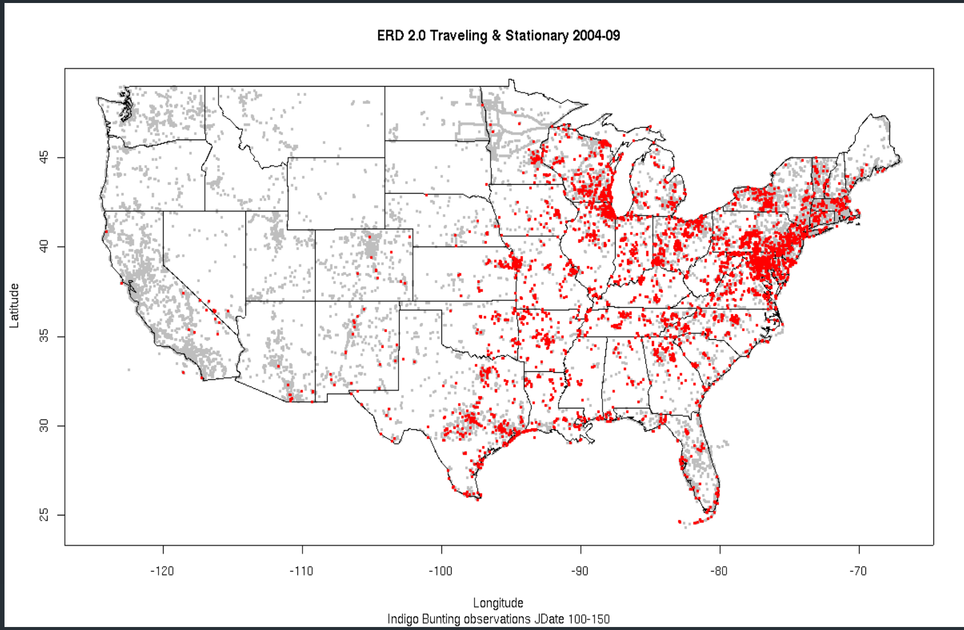
Project eBird
www.ebird.org

- Volunteer Bird Watchers
  - Stationary Count
  - Travelling Count
- Time, place, duration, distance travelled
- Species seen
  - Number of birds for each species or ‘X’ which means ≥ 1
- Checkbox: This is everything that I saw

- 8,000-12,000 checklists per day uploaded
- We need more observers in South America!!

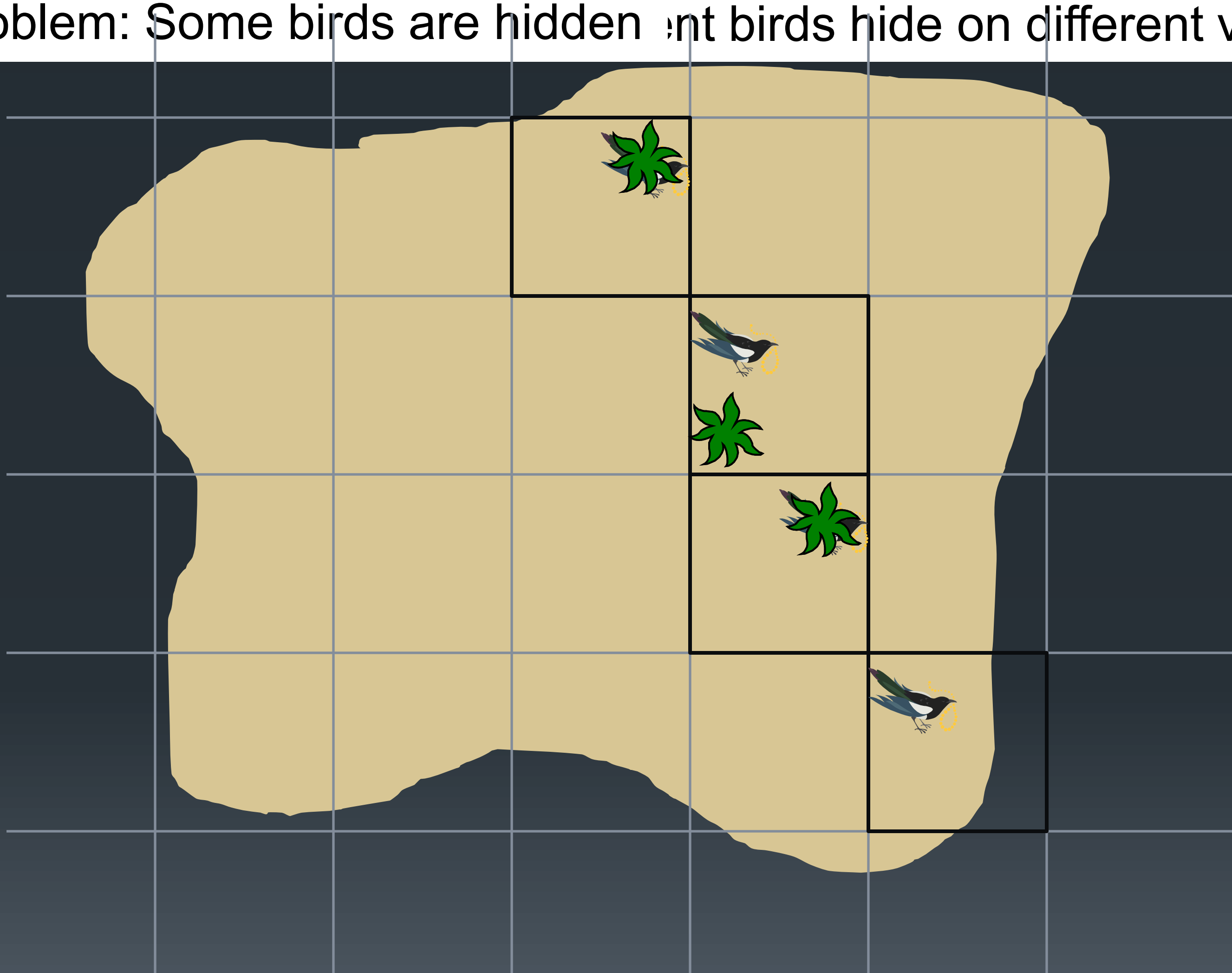
A Species Distribution Modeling Problem:

- eBird data
  - 12 bird species
  - 3 synthetic species
  - 3124 observations from New York State, May-July 2006-2008
  - 23 features



Imperfect Detection

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

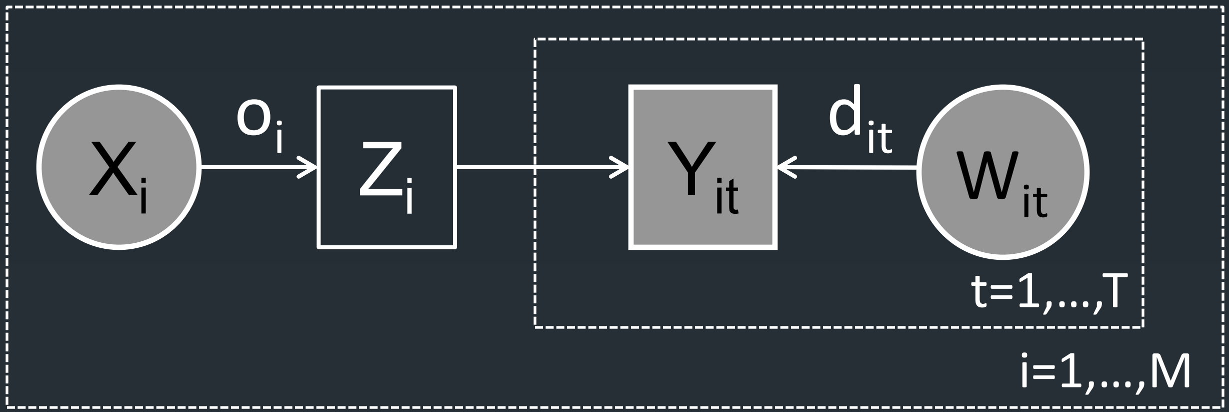


Multiple Visits to the Same Sites

		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

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

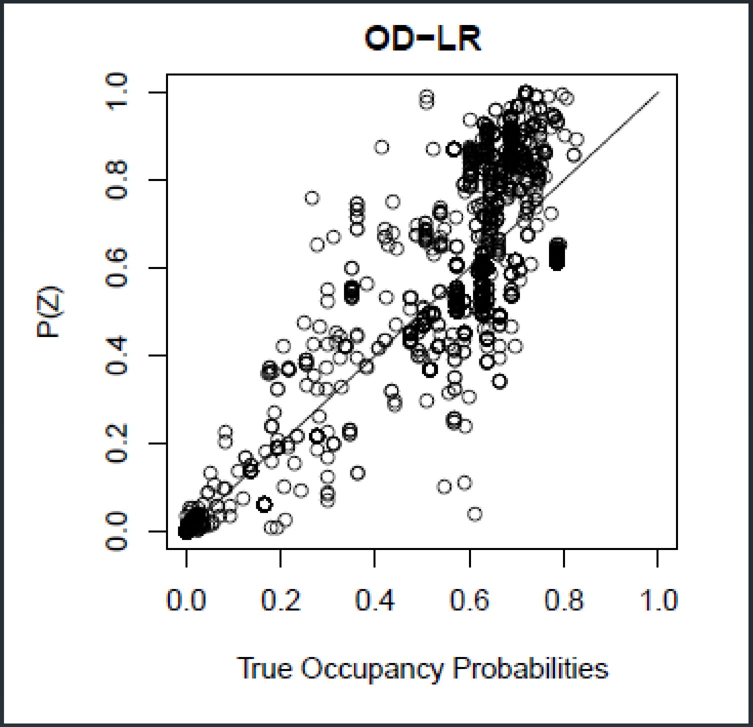
$$d_{it} = G(w_{it})$$ "detection probability"

Standard Approach: Log Linear (logistic regression) models

- $\log \frac{F(X_i)}{1-F(X_i)} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_J X_{iJ}$
  - Same as $F(X_i) = \frac{1}{1+\exp(-\boldsymbol{\beta}^T \boldsymbol{X})}$
- $\log \frac{G(W_{it})}{1-G(W_{it})} = \alpha_0 + \alpha_1 W_{it1} + \cdots + \alpha_K W_{itK}$
  - Same as $G(W_{it}) = \frac{1}{1+\exp(-\boldsymbol{\alpha}^T \boldsymbol{W}_{it})}$
- Fit via maximum likelihood
- Can apply hypothesis tests to assess importance of covariates
  - $H_0: \beta_1 = 0$
  - $H_a: \beta_1 > 0$

Results on Synthetic Species with Nonlinear Interactions

- Predictions exhibit high variance because model cannot fit the nonlinearities well

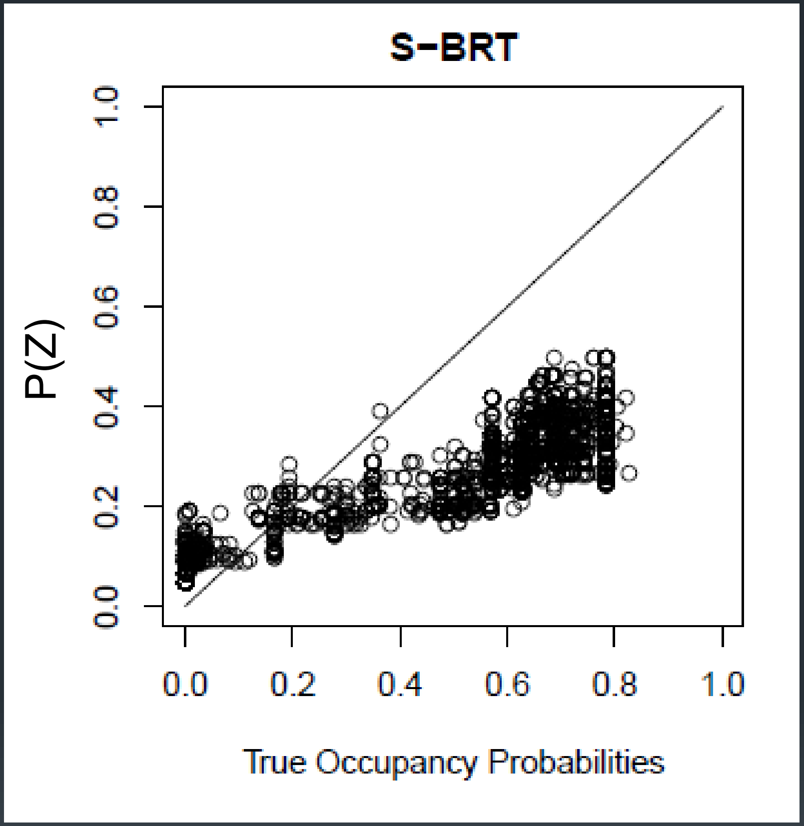


A Flexible Predictive Model

- Predict the observation y_{it} from the combination of occupancy covariates x_i and detection covariates w_{it}
- Boosted Regression trees
  - $\log \frac{P(Y_{it}=1|X_i,W_{it})}{P(Y_{it}=0|X_i,W_{it})} = \beta_1 tree_1(X_i, W_{it}) + \cdots + \beta_L tree_L(X_i, W_{it})$
  - Fitted via functional gradient descent (Friedman, 2001, 2010)
- Model complexity is tuned to the complexity of the data
  - Number of trees
  - Depth of each tree

Results

- Systematically biased because it does not capture the latent occupancy
  - Underestimates occupancy at occupied sites to fit detection failures
- Much lower variance than the Occupancy-Detection model, because it can handle the non-linearities



Two Approaches: Summary

Probabilistic Graphical Models	Flexible Nonparametric Models

- Advantages
  - Supports latent variables
  - Supports hypothesis tests on meaningful parameters
- Disadvantages
  - Model must be carefully designed (interactions? non-linearities?)
  - Data must be transformed to match modeling assumptions (linearity, Gaussianity)
  - Model has fixed complexity so either under-fits or over-fits

- Advantages
  - Model complexity adapts to data complexity
  - Easy to use “off-the-shelf”
- Disadvantages
  - Cannot support latent variables
  - Cannot provide parametric hypothesis tests

The Dream

Probabilistic Graphical Models

Flexible Nonparametric Models

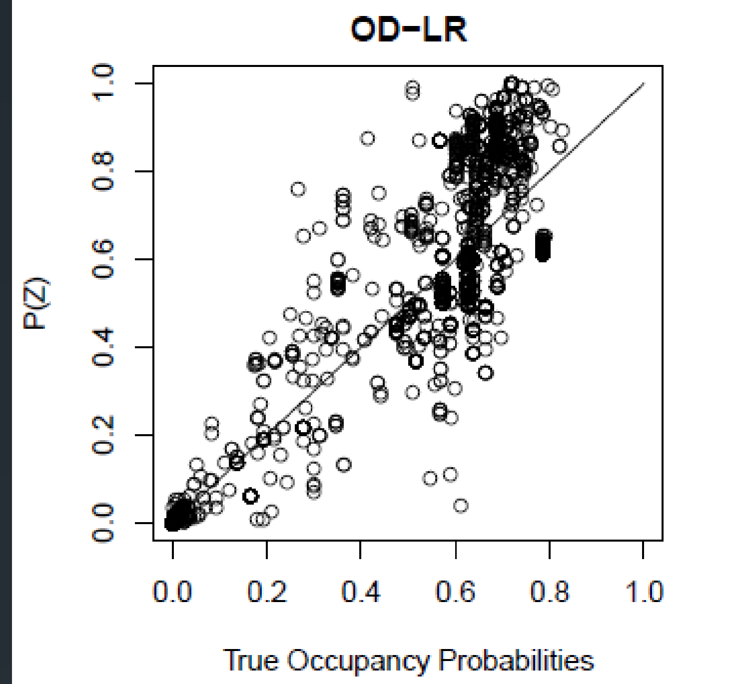
Flexible Nonparametric Probabilistic Models

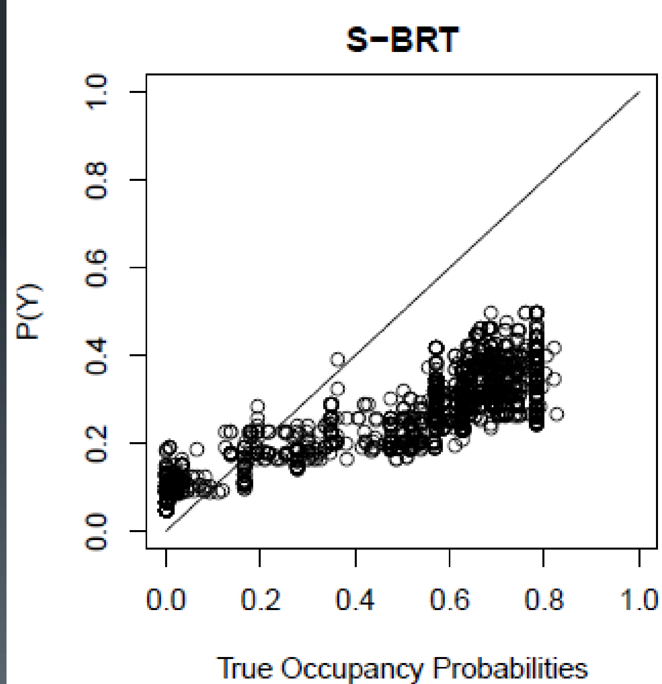
A Simple Idea: Parameterize F and G as boosted trees

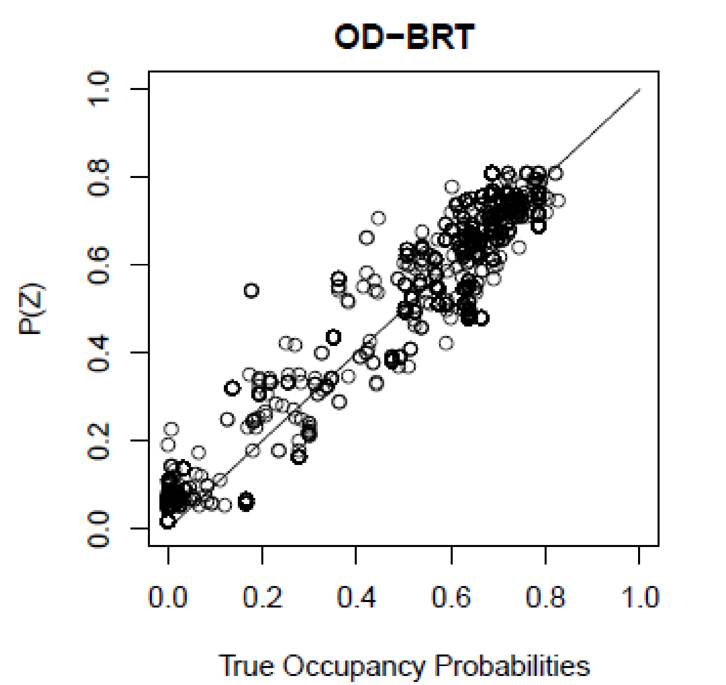
- $\log \frac{F(X)}{1-F(X)} = f^0(X) + \rho_1 f^1(X) \ + \cdots + \rho_L f^L(X)$
- $\log \frac{G(W)}{1-G(W)} = g^0(W) + \eta_1 g^1(W) + \cdots + \eta_L g^L(W)$
- Perform functional gradient descent in F and G

Results: OD-BRT

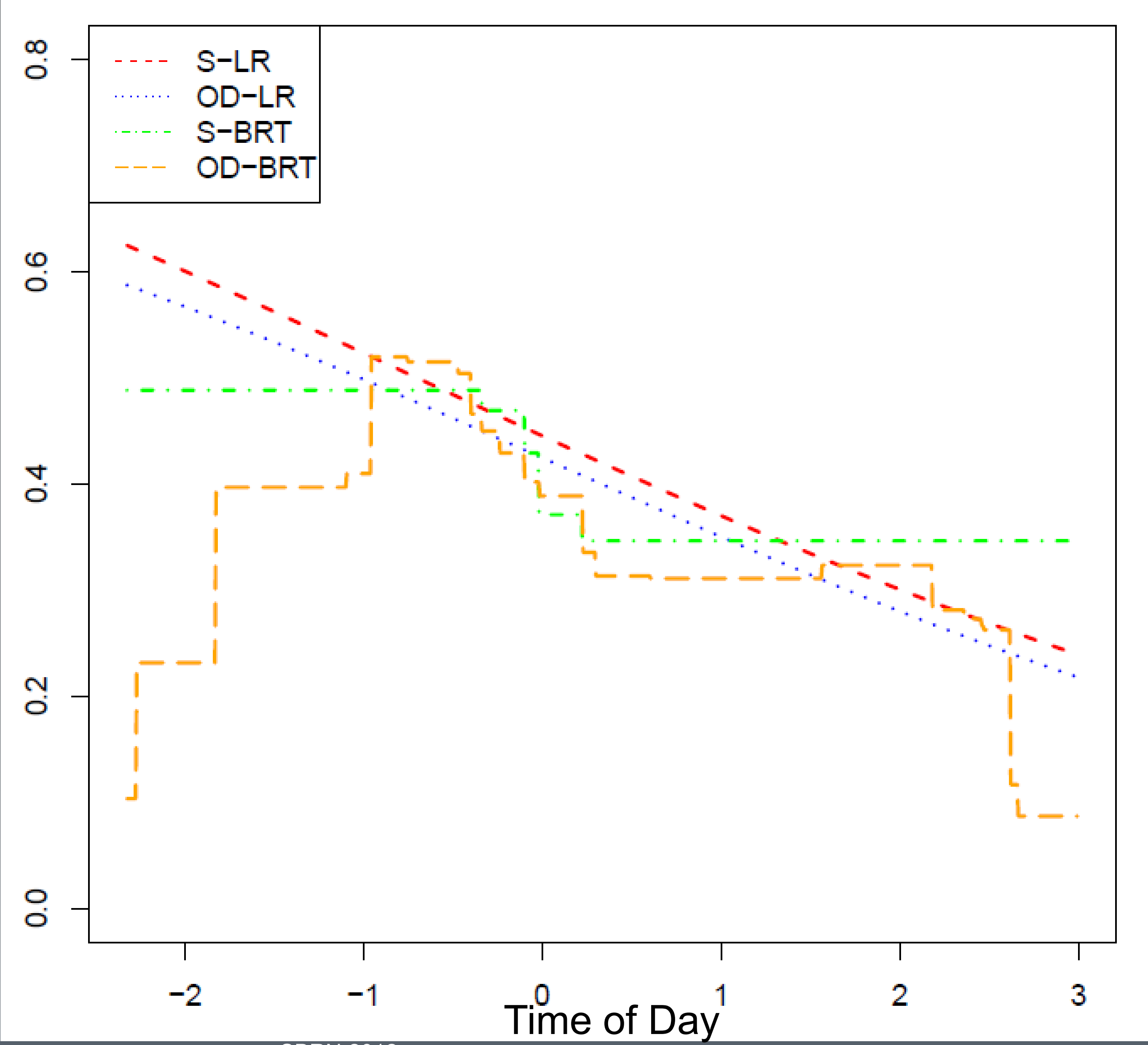
- Occupancy probabilities are predicted very well





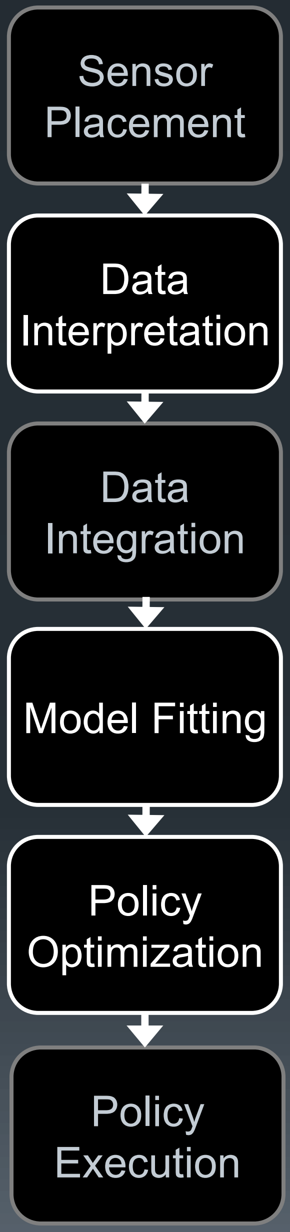


Partial Dependence Plot: Detection probability of Blue Jay vs. Time of Day



Outline:
Three Projects at Oregon State

- Data Interpretation
  - Project TAHMO
  - Automated Data Cleaning
- Model Fitting
  - Explicit Observation Models
  - Flexible Latent Variable Models
- Policy Optimization
  - Managing Fire in Eastern Oregon

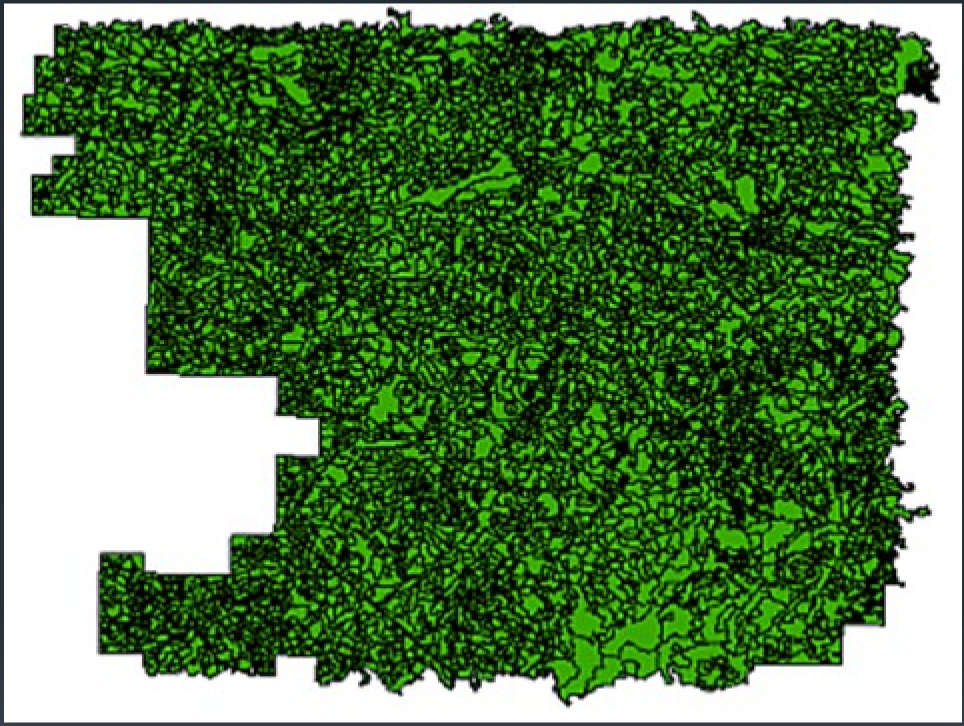


Managing Wildfire in Eastern Oregon

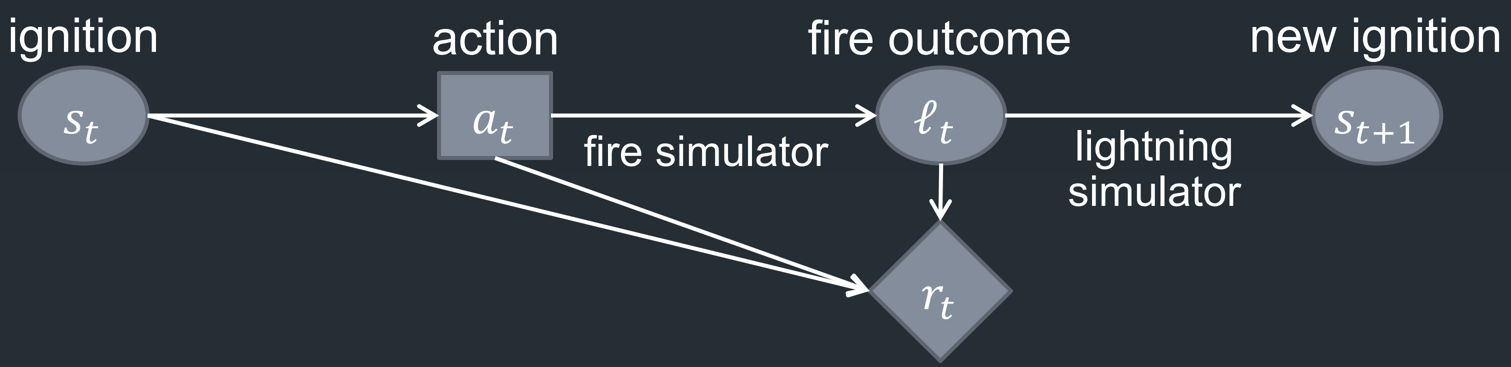
- Natural state (hypothesized):
  - Large Ponderosa Pine trees with open understory
  - Frequent “ground fires” that remove understory plants (grasses, shrubs) but do not damage trees
- Fires have been suppressed since 1920s
  - Large stands of Lodgepole Pine
  - Heavy accumulation of fuels in understory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

Study Area: Deschutes National Forest

- Goal: Return the landscape to its "natural" fire regime
- Management Questions:
  - LET-BURN: When lightning ignites a fire, should we let it burn?
  - FUEL TREATMENT: Which units should have mechanical fuel removal?
- ~4000 management units

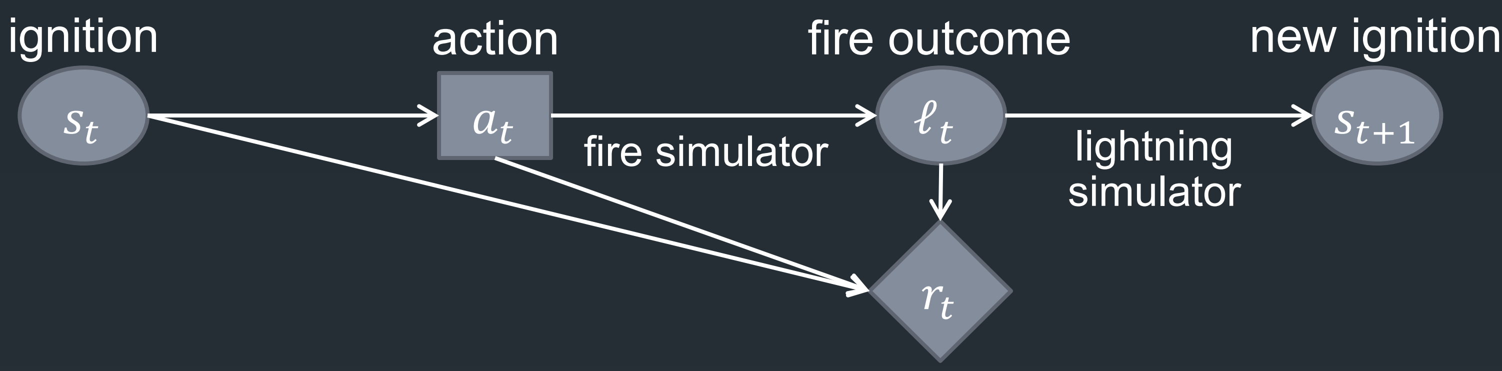


Formulating LETBURN as a Markov Decision Process $\langle S, A, R, T, \gamma \rangle$



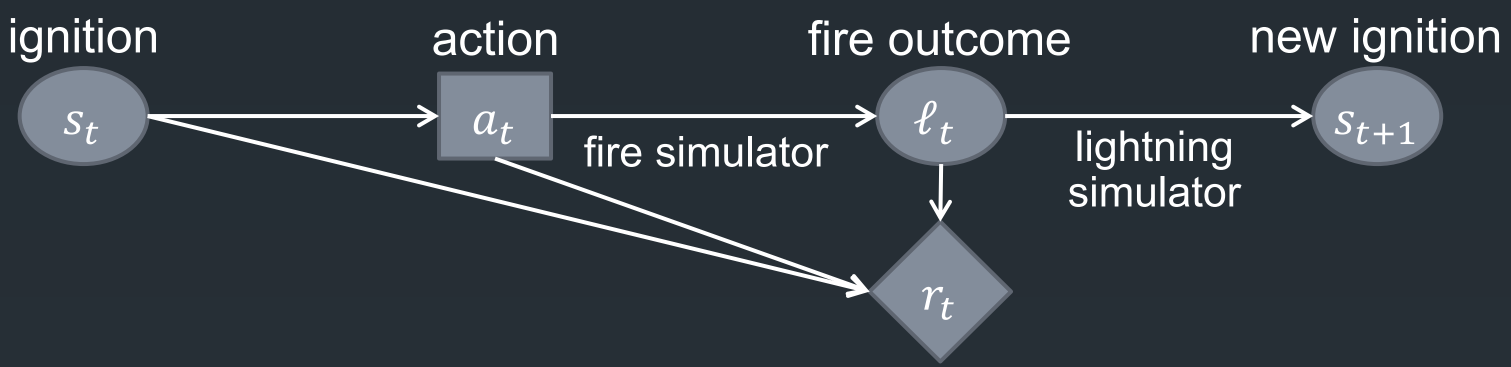
- State space: S
  - 4000 management units; each unit is in one of 25 local states
  - Global state space is 25^{4000}
- Action space: A
  - At fire ignition time t, $a_t \in \{LETBURN, SUPPRESS\}$
- Reward function: $R(s, \ell, a)$
  - Cost of lost timber value
  - Cost of lost species habitat
  - If SUPPRESS, then cost of fire suppression

Formulating LETBURN as a Markov Decision Process



- Transition function: $T(s_{t+1}|s_t, a_t)$
  - $T(s_{t+1}|s_t, a_t) = P(\ell_t\,|s_t, a_t) \cdot P(s_{t+1}|s_t)$
  - Includes forest growth at the end of each fire season
- Discount factor γ
- Optimization goal
  - Maximize sum of discounted rewards:
  - $\mathbb{E}_T[r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots]$

Solving the MDP
No existing methods...



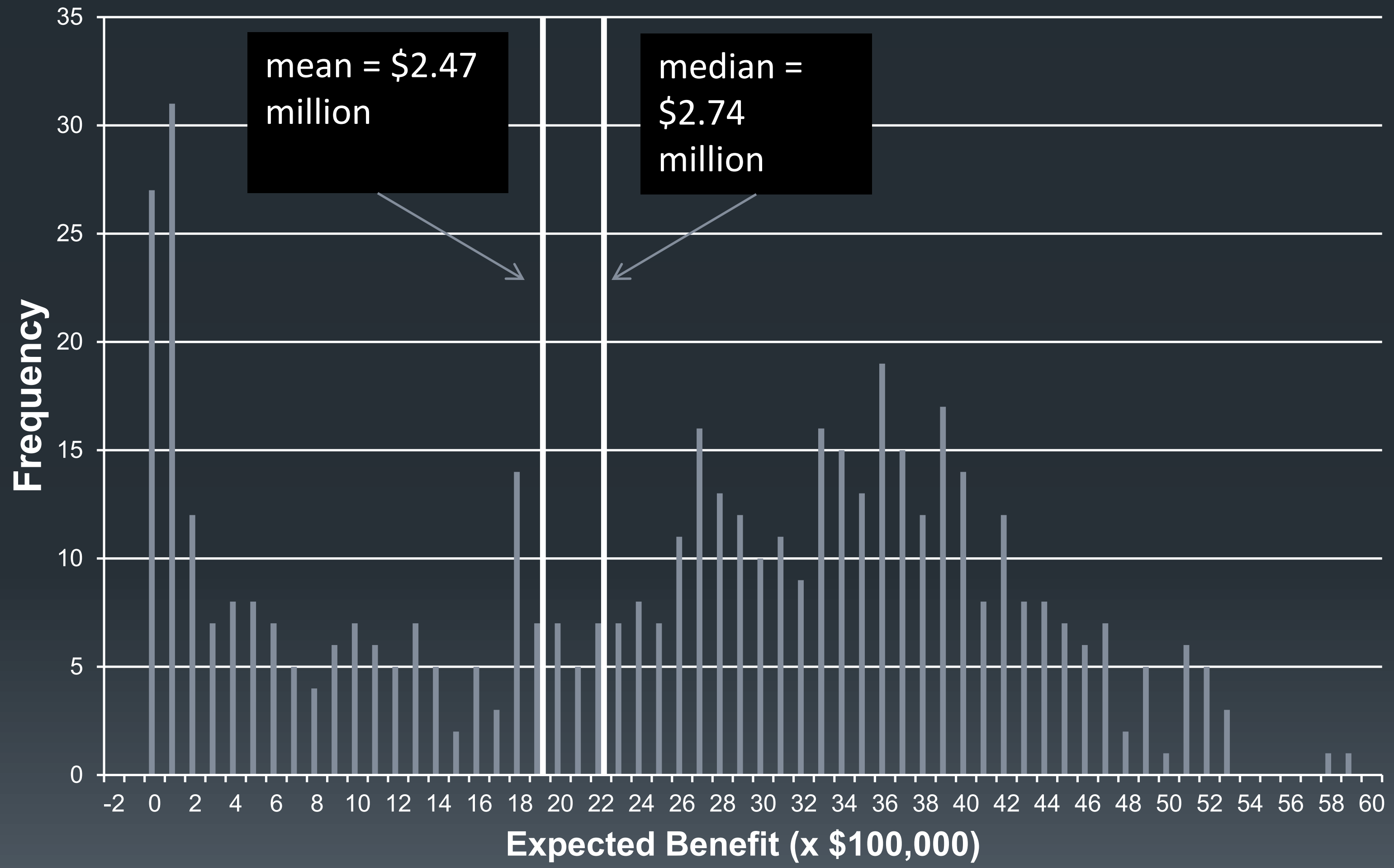
- Promising Approaches:
  - Approximate Policy Iteration (Fern & Givan, 2005)
    - represent the policy as a classifier
    - train using Monte Carlo trials
  - Policy Gradient (Williams, 1992)
    - represent the policy as a function
    - train via Monte Carlo gradient estimates

A Simpler Problem

- Is there any benefit to allowing fires to burn for just one year?
  - Year 1: LETBURN
  - Years 2-100: SUPPRESS

- Evaluate via Monte Carlo trials

Expected Benefit of LETBURN
(Suppress all fires after year 1)



Next Steps

- Single Year LETBURN Study:
  - Several model improvements
  - Include standard forest harvest policy
  - Include more accurate timber value

- 100-year Dynamic LETBURN Study
  - Needed: MDP algorithms that can scale to the immense state space
    - Approximate Policy Iteration? (Fern et al.)

FUEL TREATMENT

- 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)

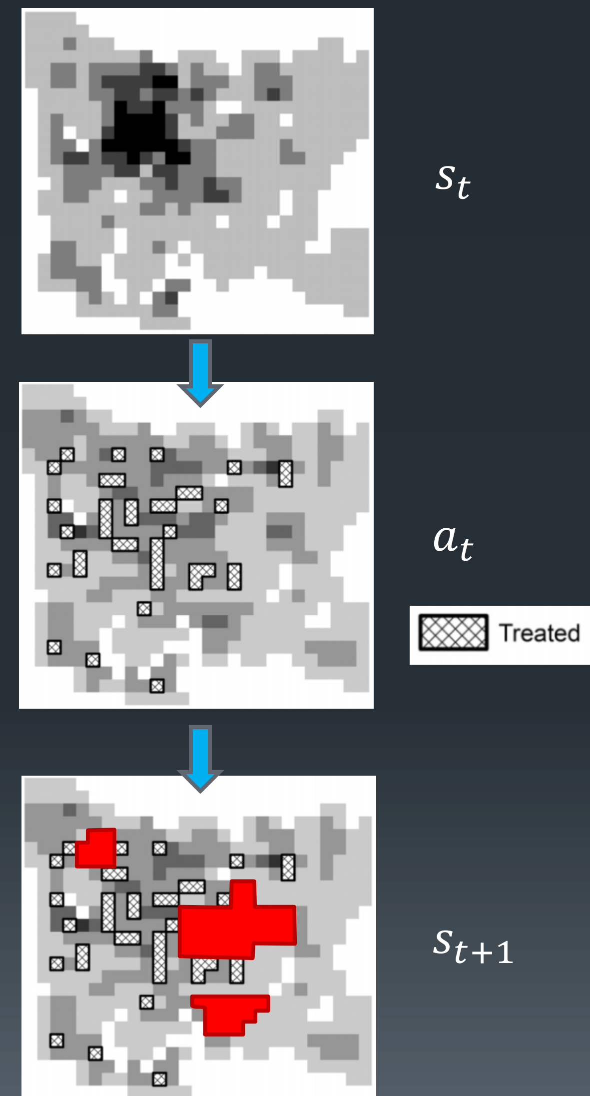
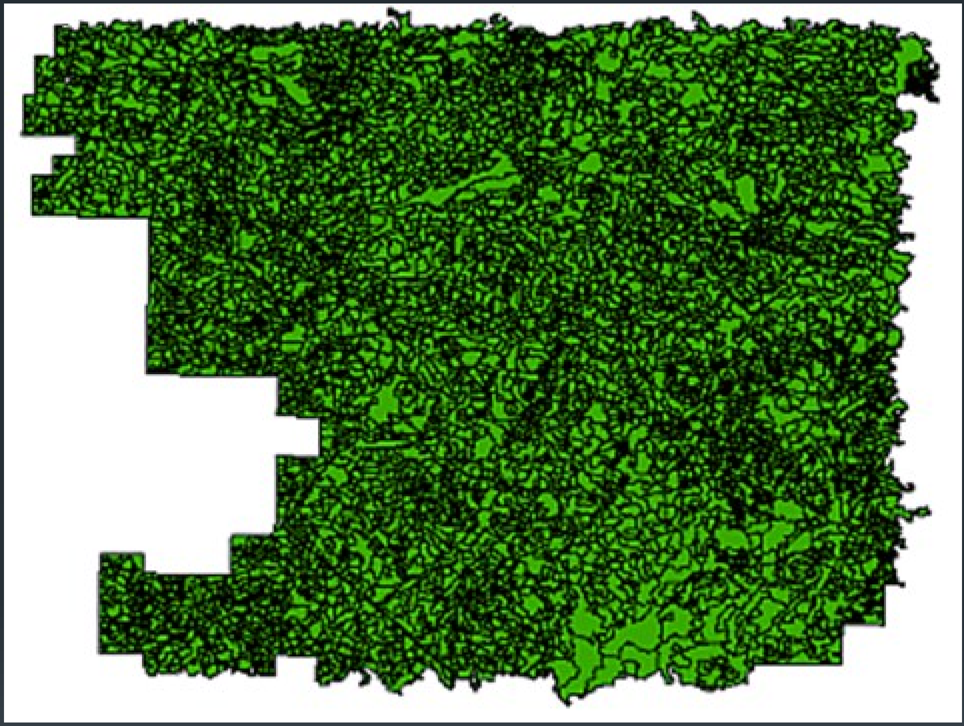


Image: Wei et al, 2008

Formulation as a Markov Decision Process

- State of each MU:
  - Age of trees (years)
    - {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

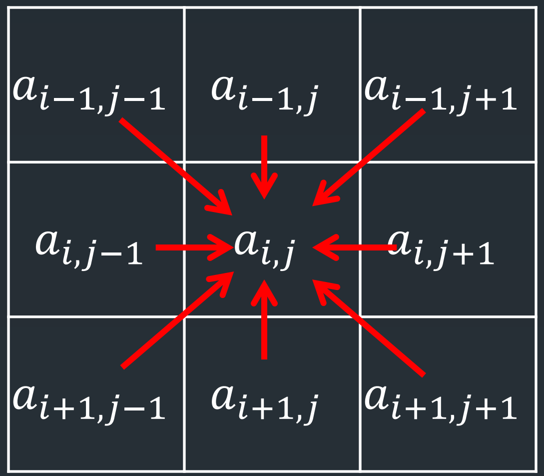
- Actions in each MU each decade
  - Do nothing
  - Fuel treatment (costs money)
  - Harvest trees (makes money, but increases fuel)
  - Harvest + Fuel
  - 4^{4000} possible actions over landscape



Study area in Deschutes National Forest

Solving Spatial MDP

- No existing methods
- Promising Approach: Equilibrium Policy Gradient
  - Define a pixel policy $\pi(\theta, \eta(ij))$ that chooses an action for pixel i, j based on a neighborhood $\eta(ij)$
  - Define a Markov Chain as in Gibbs sampling
  - Sample an landscape action vector from the stationary distribution of the chain
- It is possible to compute the policy gradient of this MC equilibrium policy
  - Crowley, Nelson, & Poole (AAAI 2011)

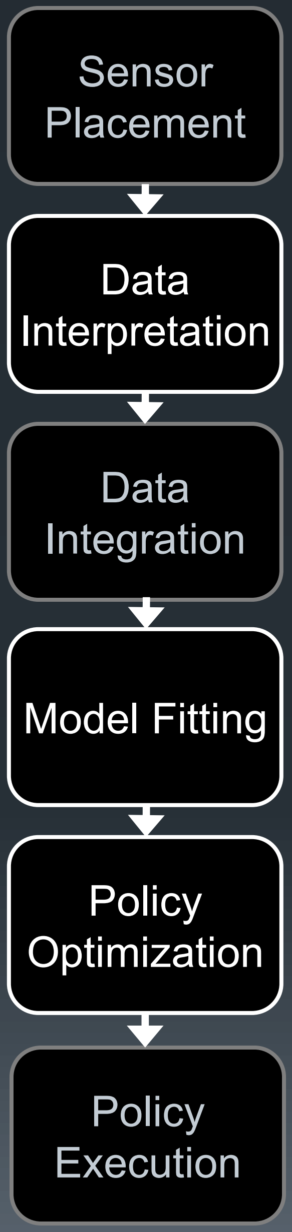


Open Problems

- Risk-sensitive solutions
  - Maximize expected value while keeping the probability of catastrophic fires below ϵ
- Visualize the resulting policy

Summary

- Data Interpretation
  - Automated Data Cleaning
  - Probabilistic modeling + approximate inference
- Model Fitting
  - Explicit Observation Models
  - Combine flexible machine learning with latent variable models
- Policy Optimization
  - Managing Fire in Eastern Oregon
  - Monte Carlo optimization



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

- Institute for Computational Sustainability: http://www.computational-sustainability.org/

Thank-you

- Ethan Dereszynski: Automated Data Cleaning
- John Selker: Project TAHMO
- Rebecca Hutchinson: Boosted Regression Trees in OD models
- Claire Montgomery, Rachel Houtman, Sean McGregor, Mark Crowley: Fire challenge

- National Science Foundation Grants 0705765, 0832804, and 0905885

The Distinguished Speakers Program is made possible by

Association for Computing Machinery

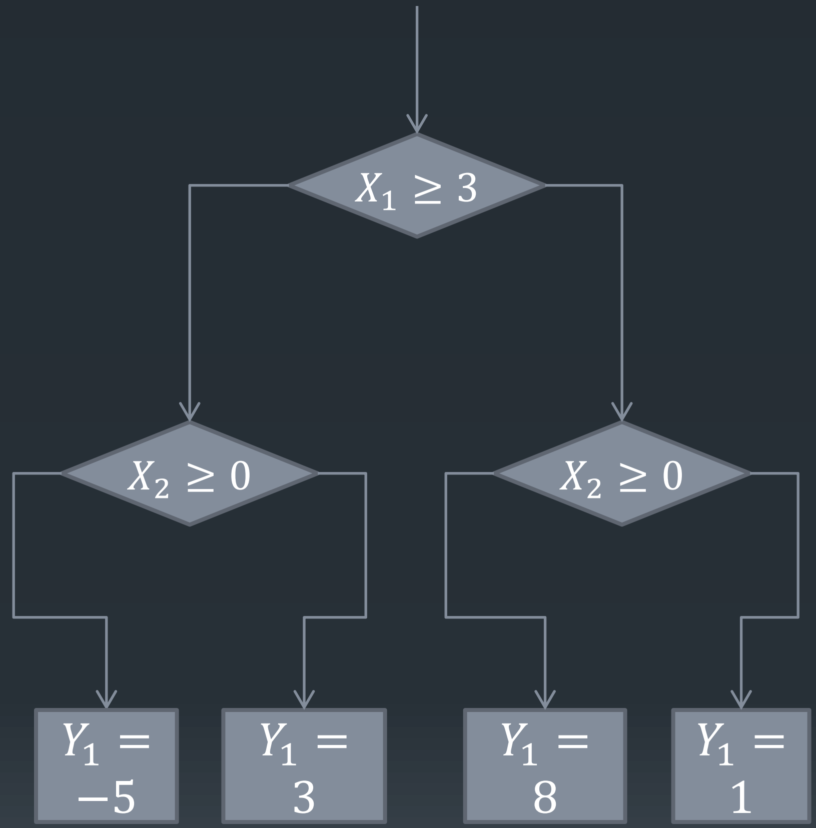
Advancing Computing as a Science & Profession

For additional information, please visit http://dsp.acm.org/

Questions?

Regression Trees

- Classification and regression trees
  - Interactions are captured by the if-then-else structure of the tree
  - Nonlinearities are approximated by piecewise constant functions



$$Y_1 = -5 \cdot I(X_1 \geq 3, X_2 \geq 0) + 3 \cdot I(X_1 \geq 3, X_2 < 0) + \\ 8 \cdot I(x_1 < 3, X_2 \geq 0) + 1 \cdot I(X_1 < 3, X_2 < 0)$$

Representing $P(Y|X)$ using boosted regression trees

- Friedman: Gradient Tree Boosting (2000; Annals of Statistics, 2011)
- Consider logistic regression:
  - $\log \frac{P(Y=1)}{P(Y=0)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_J X_J$
  - $D = \left\{\left(X^i, Y^i\right)\right\}_{i=1}^{N}$ are the training examples
  - Log likelihood:
    - $LL(\beta) = \sum_i Y^i \log P\left(Y = 1 \middle| X^i; \beta\right) + \left(1 - Y^i\right) \log P\left(Y = 0 \middle| X^i; \beta\right)$

Fitting logistic regression via gradient descent

- Let $\beta^0 = g^0 = \mathbf{0}$
- For $\ell = 1, \dots, L$ do
  - Compute $g^\ell = \nabla_\beta LL(\beta)\big|_{\beta=\beta^{\ell-1}}$
    - g^ℓ = gradient w.r.t. β
  - $\beta^\ell \coloneqq \beta^{\ell-1} + \eta_\ell g^\ell$ take a step of size η_ℓ in direction of gradient

- Final estimate of β is
  - $\beta^L = g^0 + \eta_1 g^1 + \cdots + \eta_L g^L$

Functional Gradient Descent Boosted Regression Trees

- Friedman (2000), Mason et al. (NIPS 1999), Breiman (1996)
- Fit a logistic regression model as a weighted sum of regression trees:

$$\log \frac{P(Y=1)}{P(Y=0)} = tree^0(X) + \eta_1 tree^1(X) + \cdots + \eta_L tree^L(X)$$

- When "flattened" this gives a log linear model with complex interaction terms

L2-Tree Boosting Algorithm

- Let $F^0(X) = f^0(X) = 0$ be the zero function
- For $\ell = 1, \dots, L$ do
  - Construct a training set $S^\ell = \{(X^i, \tilde{Y}^i)\}_{i=1}^N$
    - where $\tilde{Y}$ is computed as
    - $\tilde{Y}^i = \left.\frac{\partial LL(F)}{\partial F}\right|_{F=F^{\ell-1}(X^i)}$ how we wish F would change at X^i
  - Let f^ℓ = regression tree fit to S^ℓ
  - $F^\ell := F^{\ell-1} + \eta_\ell f^\ell$
- The step sizes η_ℓ are the weights computed in boosting
- This provides a general recipe for learning a conditional probability distribution for a Bernoulli or multinomial random variable