Challenges for Machine Learning in Computational Sustainability

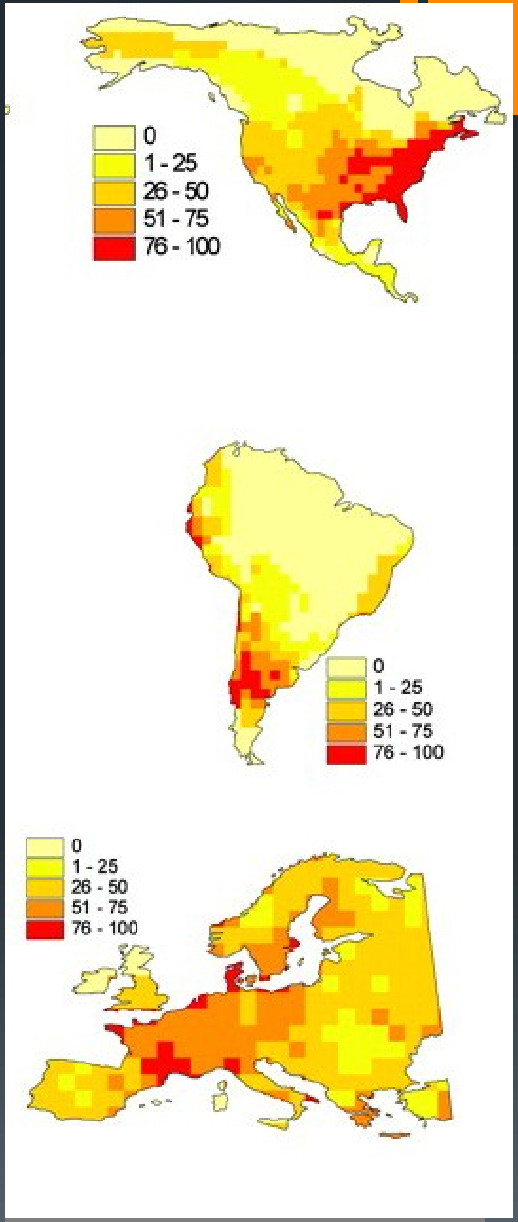
Tom Dietterich
Oregon State University
In collaboration with
Postdocs: Rebecca Hutchinson, Dan Sheldon, Mark Crowley
Graduate Students: Majid Taleghan, Kim Hall, Liping Liu
Economist: H. Jo Albers
and the Cornell Lab of Ornithology

NIPS 2012

Sustainable Management of the Earth’s Ecosystems

- The Earth’s Ecosystems are complex
- We have failed to manage them in a sustainable way
  - Example:
    - Species extinction rate of mammals ≈ 10-100 times historical rates
    - Mammalian populations are dropping rapidly worldwide

Ceballos & Erhlich, 2002



% mammal population lost

Why?

1. We did not think about ecosystems as a management or control problem
2. Our knowledge of function and structure is inadequate
3. Optimal management requires spatial planning over horizons of 100+ years

Computer Science can help!

1. We did not think about ecosystems as a management or control problem
2. Our knowledge of function and structure is inadequate
3. Optimal management requires spatial planning over horizons of 100+ years

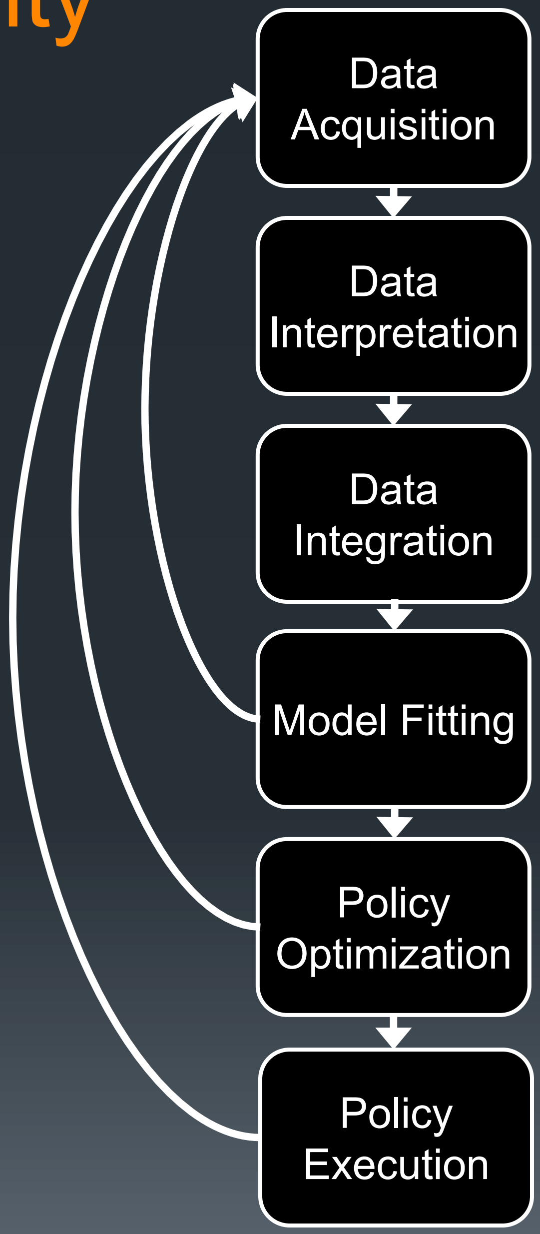
Sensors

Machine Learning

Optimization

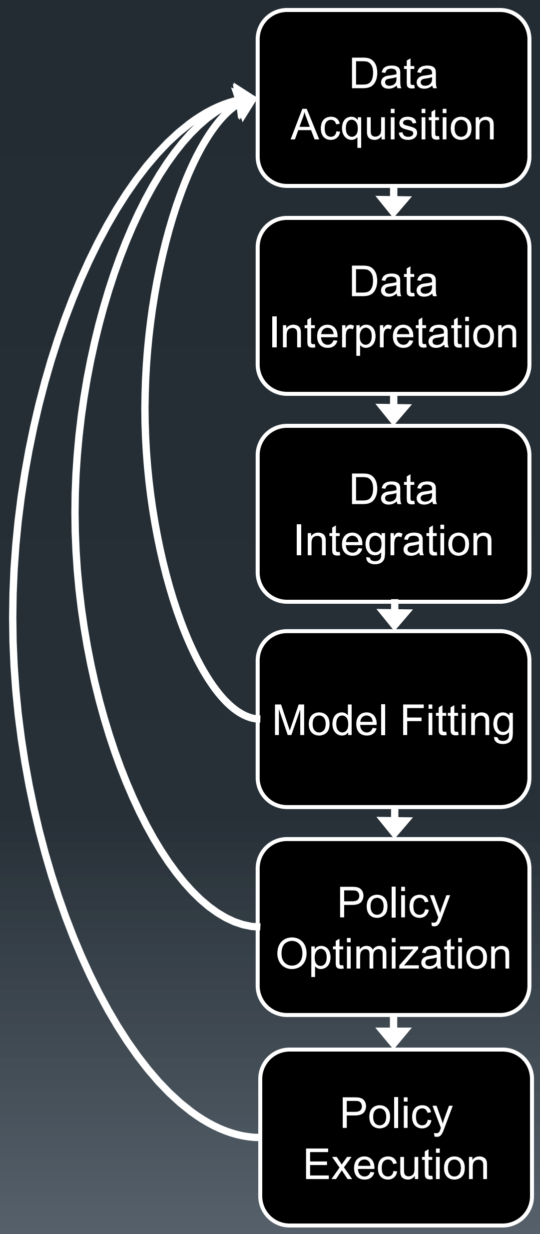
Computational Sustainability

- The study of computational methods that can contribute to the sustainable management of the earth’s ecosystems
- Data → Models → Policies



Outline

- Illustrative Research Challenges for each stage
- Drill down on three projects at Oregon State University
- Discussion: What are the distinctive aspects of computational sustainability problems?



Example Research Challenges
Data Acquisition

Data Acquisition

- 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 meteorology sensor station at a cost of EUR 200
  - Deploy 20,000 such stations across Africa
  - Where should sensors be placed?
    - Accuracy of reconstructed fields for precipitation, temperature, relative humidity, wind, etc.
    - Robustness to sensor failure, station loss

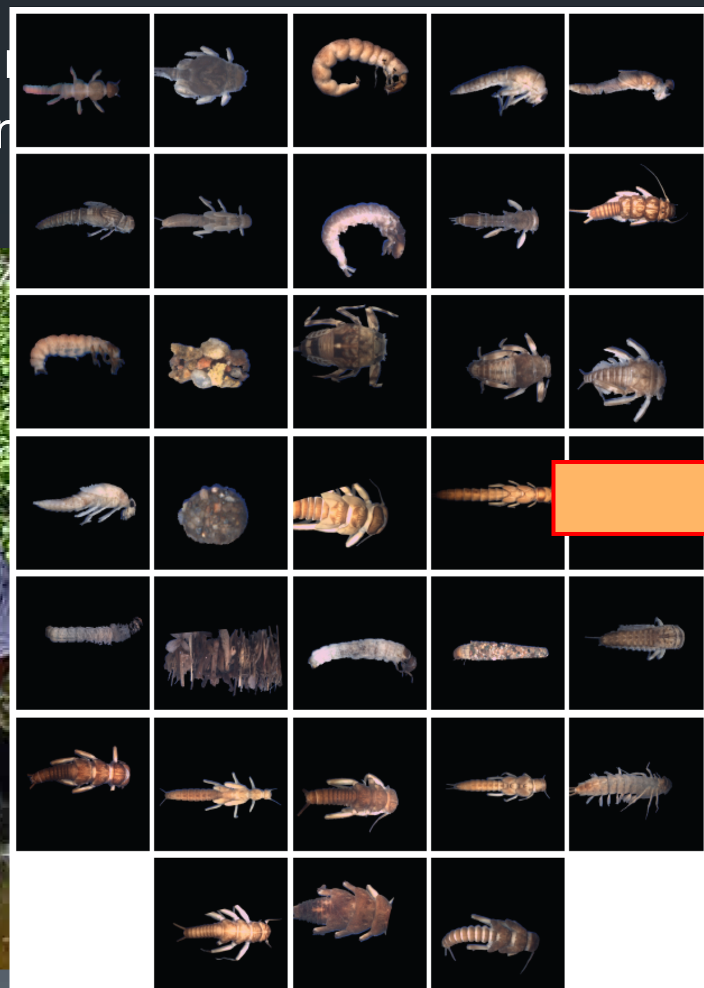
TU Delft Delft University of Technology

OSU Oregon State UNIVERSITY

Data Interpretation

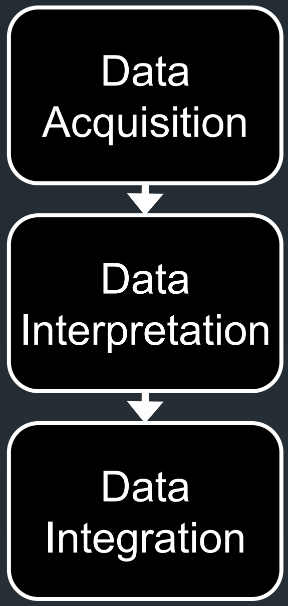
- Insect identification for population counting
- Raw data: image
- Interpreted data: Cou
- Challenge: Fine-Grai

www.epa.gov

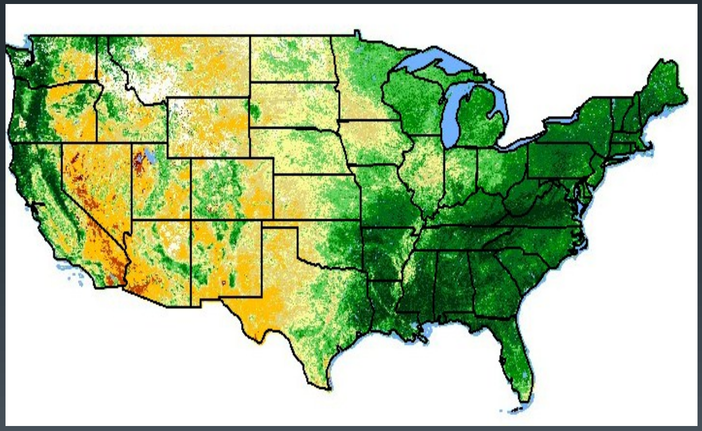


Species	Count
Limne	3
Taenm	15
Asiop	4
Epeor	25
Camel	19
Cla	12
Cerat	21

Data Integration



- Virtually all ecosystem prediction problems require integrating heterogeneous data sources
  - Landsat (30m; monthly)
    - land cover type
  - MODIS (500m; daily/weekly)
    - land cover type
  - Census (every 10 years)
    - human population density
  - Interpolated weather data (15 mins)
    - rain, snow, solar radiation, wind speed & direction, humidity
- Challenge:
  - Learn from heterogeneous data
    - without losing fine-grained information
    - without losing uncertainty in the data

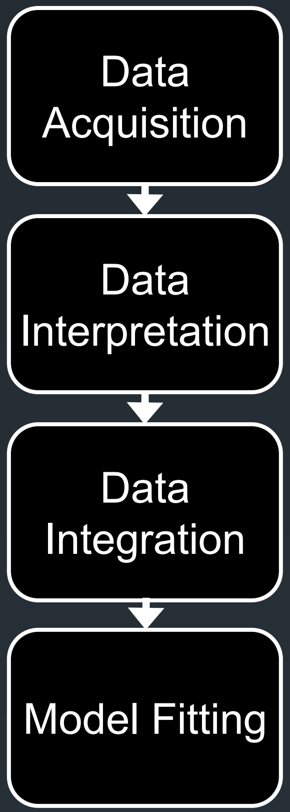


Landsat NDVI: http://ivm.cr.usgs.gov/viewer/

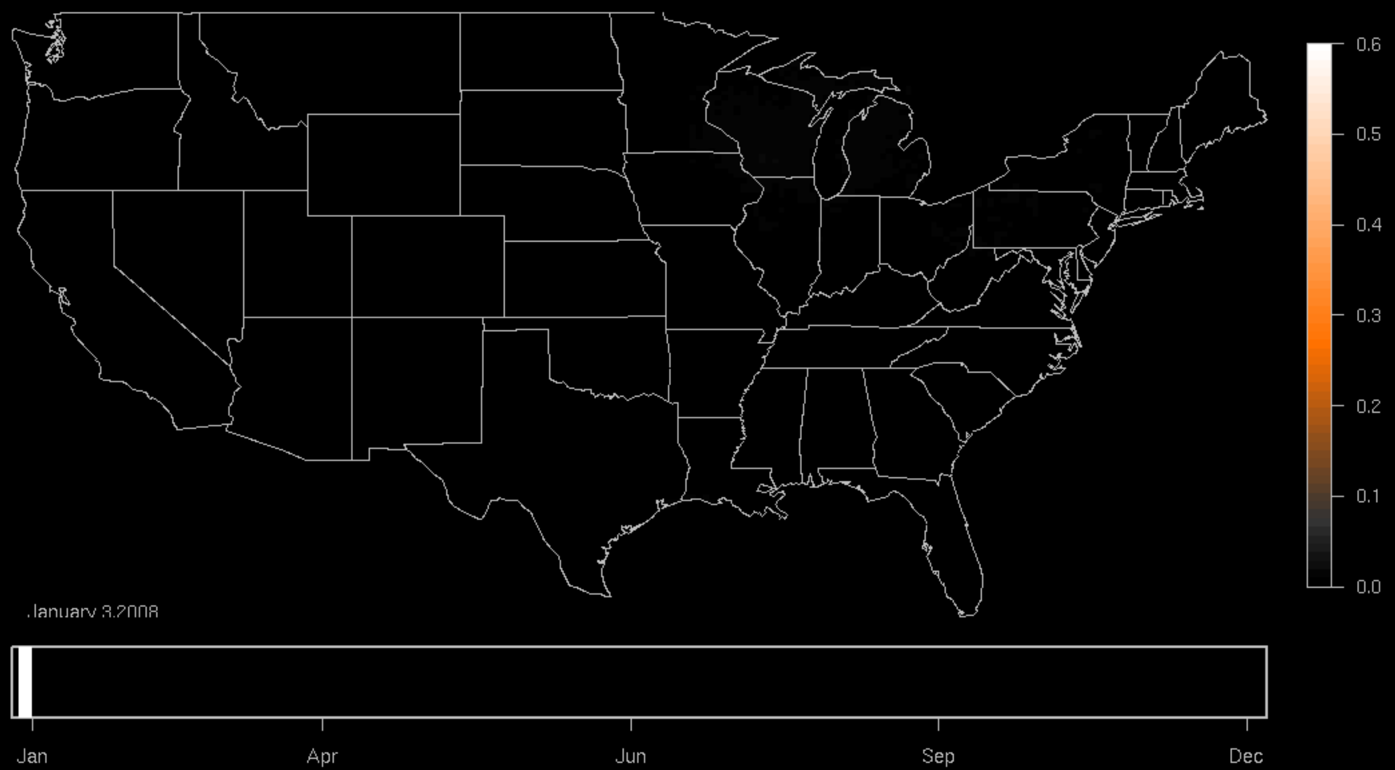
Model Fitting

- 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

- Challenges
  - The variables of interest are all latent
    - Latent distribution of species
    - Latent dynamics
  - The data are very messy



State of the Art: STEM Model of Bird Species Distribution

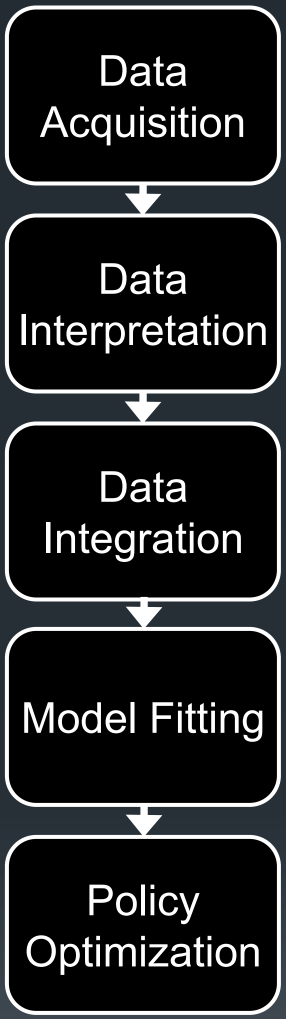


Indigo Bunting

slide courtesy of Daniel Fink

Policy Optimization

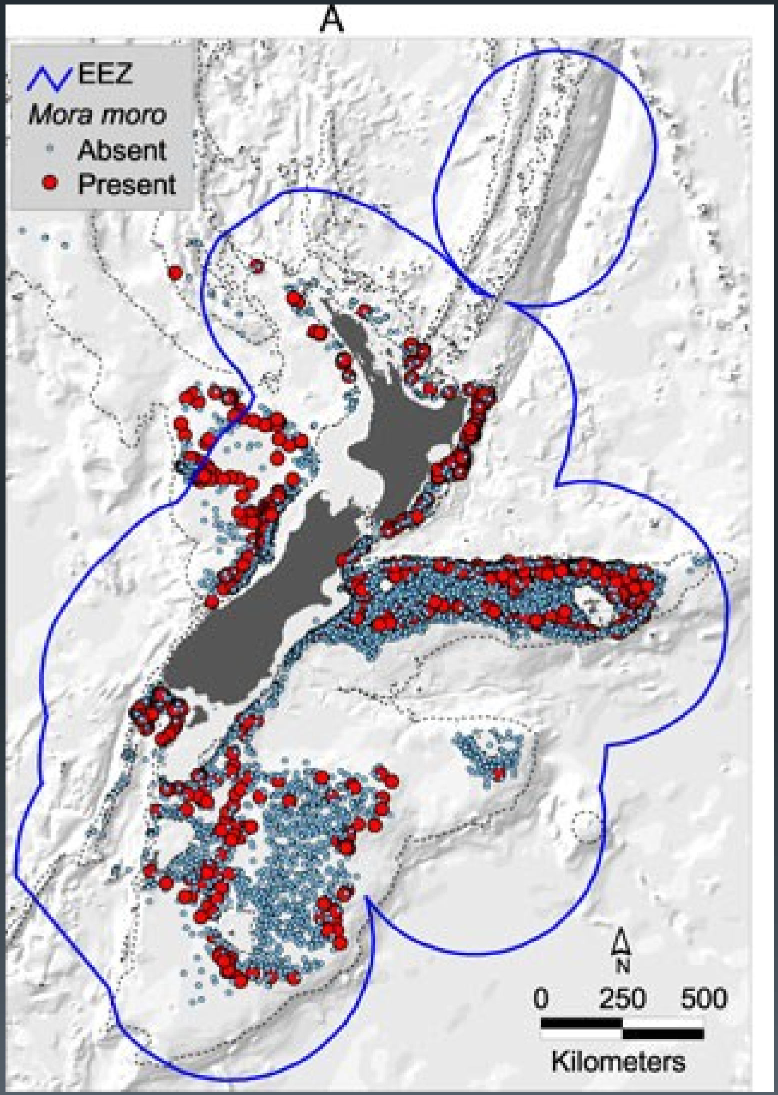
- Challenges
  - Long time horizons (100+ years)
  - The system model is uncertain, so the optimization needs to be robust to this uncertainty
  - The state of the system covers large spatial regions (scales exponentially in region size)
  - System dynamics only available via simulation or sampling



Leathwick et al, 2008

State of the Art: Reserve Design from a Species Distribution Model

Observations

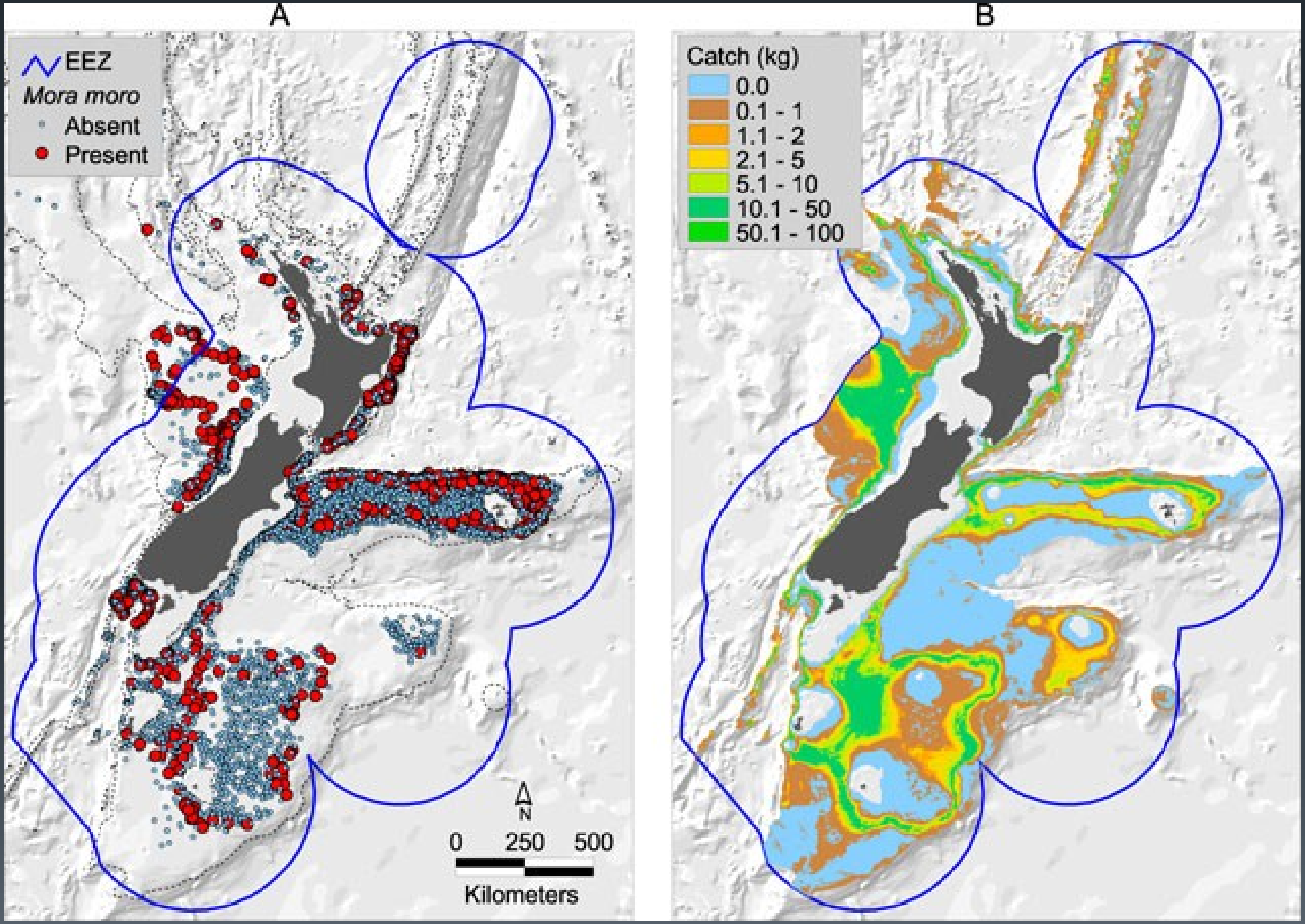


Data Acquisition → Data Interpretation → Data Integration → Model Fitting → Policy Optimization

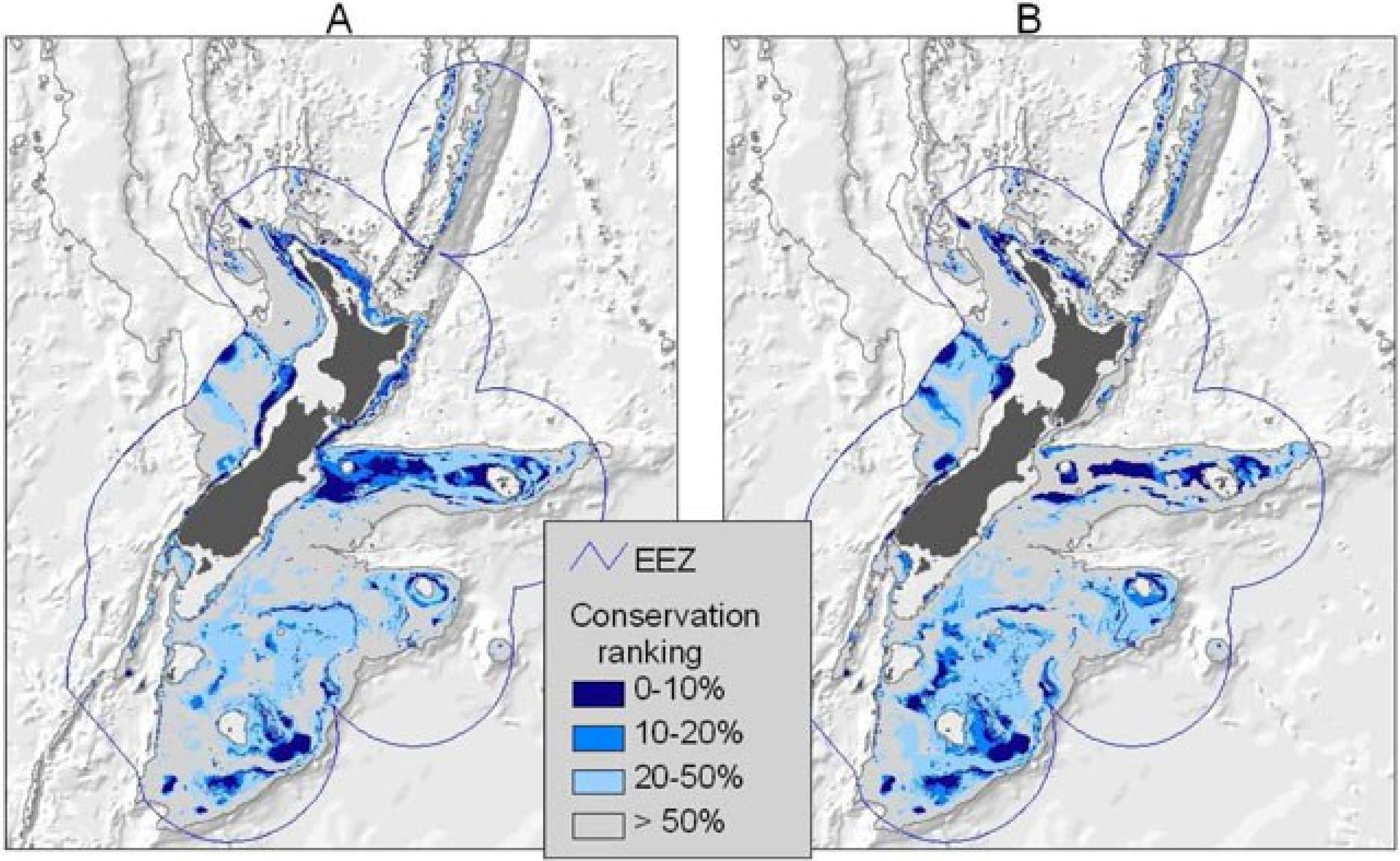
Leathwick et al, 2008

State of the Art: Reserve Design from a Species Distribution Model

Observations | Fitted Model



Leathwick et al, 2008



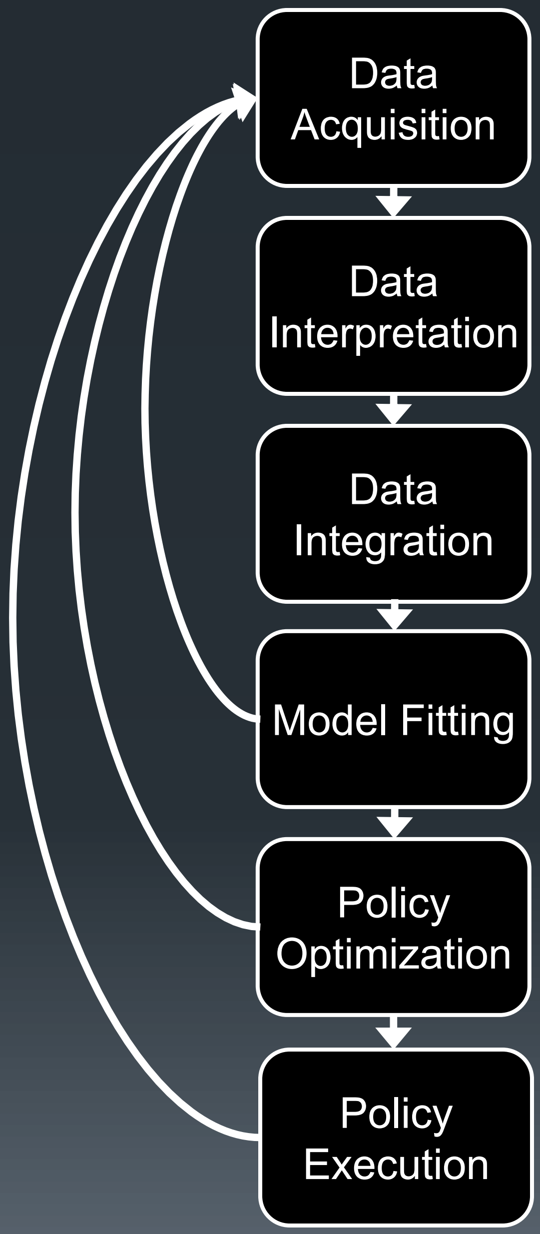
Disregarding costs to fishing industry

Full consideration of costs to fishing industry

Leathwick et al, 2008

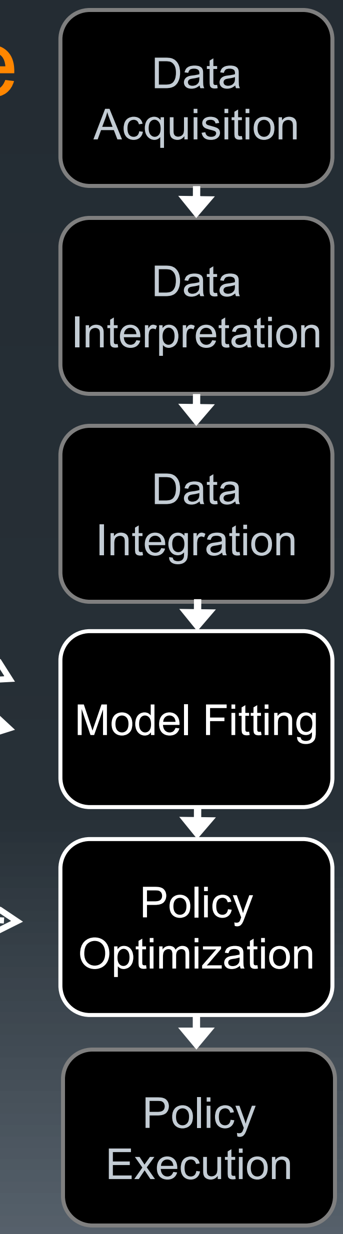
Policy Execution

- 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)



Drill Down: Three Projects at Oregon State

- Species Distribution Modeling with Imperfect Observations
  - Explicit Observation Models
  - Flexible Latent Variable Models
- Models of Bird Migration
  - Collective Graphical Models
- Policy Optimization
  - Controlling Invasive Species
  - Algorithms for Large Spatial MDPs



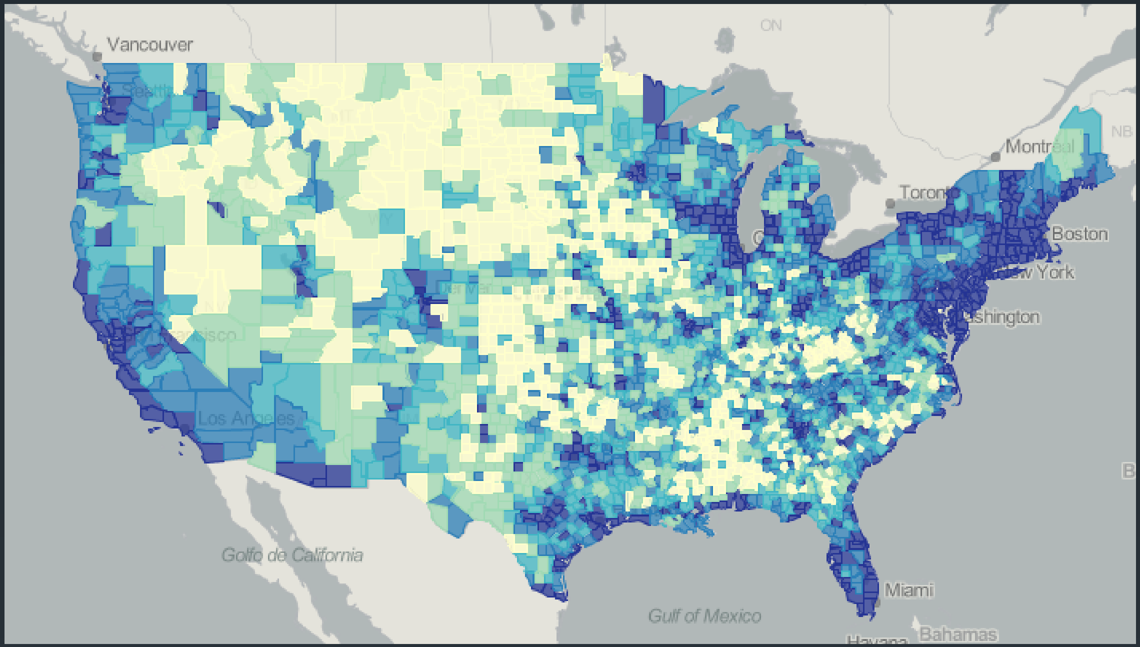
Project eBird
www.ebird.org

- 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

Species Distribution Modeling from Citizen Science Data:

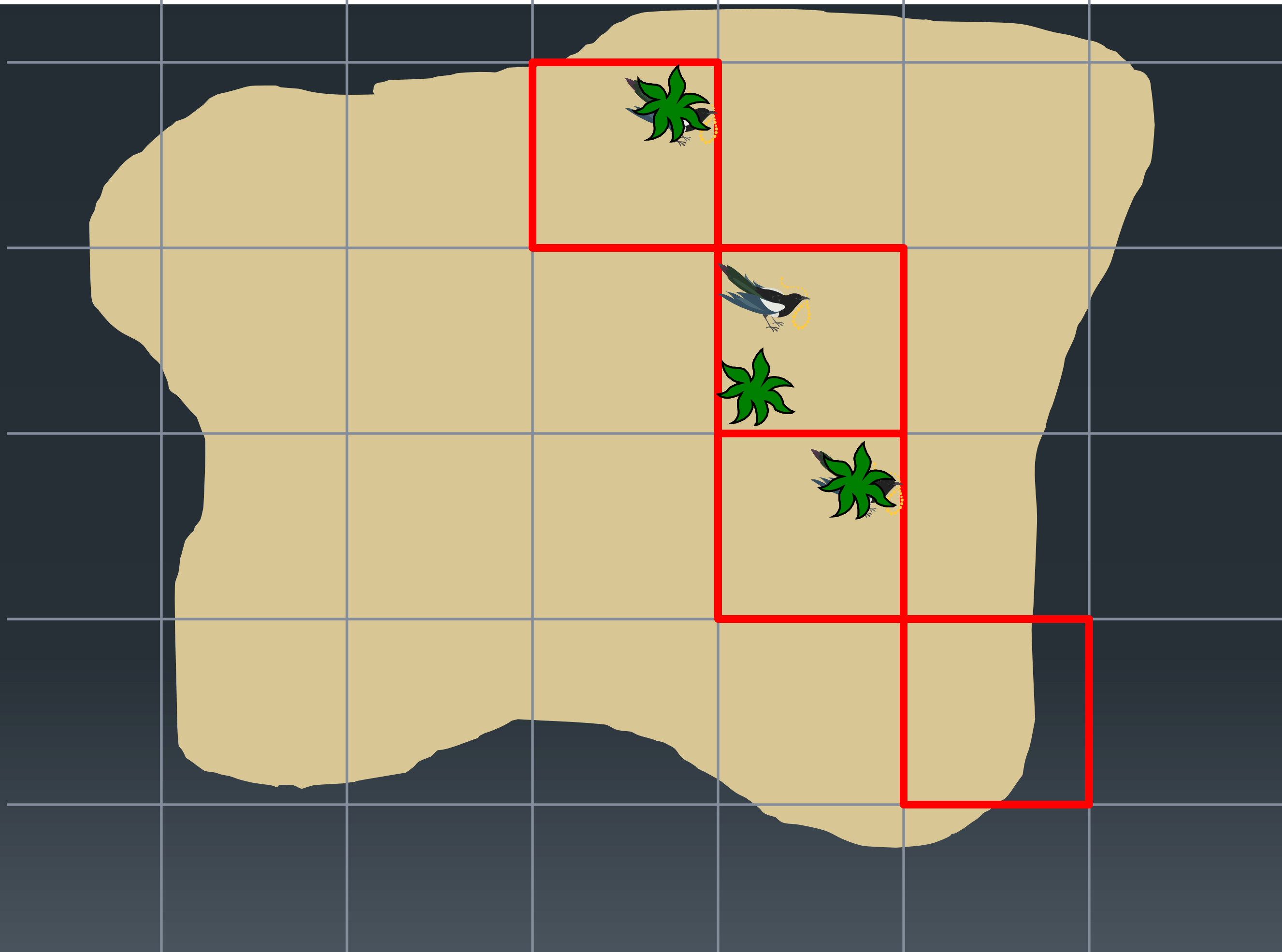
- eBird data issues
  - imperfect detection
  - variable expertise
  - sampling bias
  - ...



Tom Auer http://geocommons.com/maps/137230

Imperfect Detection

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

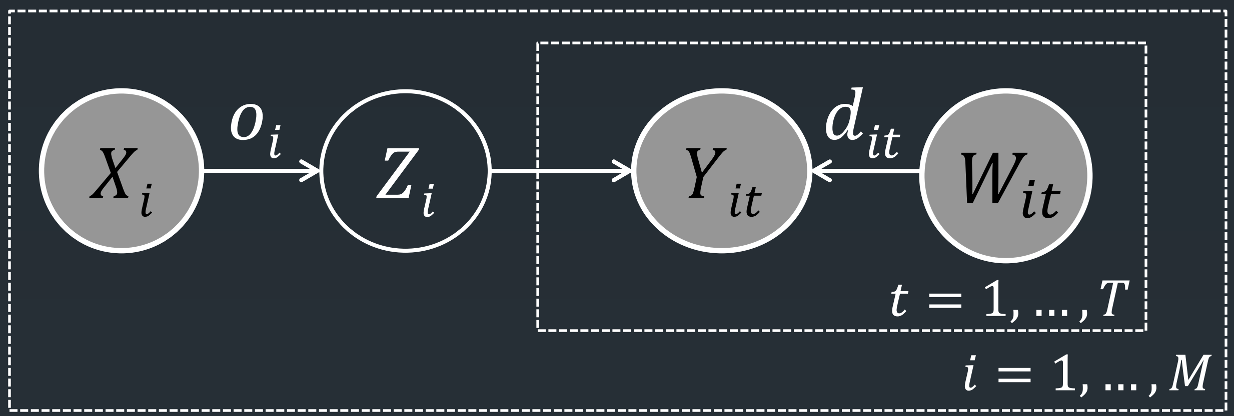


Multiple Visits to the Same Sites

		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, 2002



$Z_i \sim P(Z_i | X_i)$: 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

$P(Y_{it} = 1 | Z_i, W_{it}) = Z_i d_{it}$

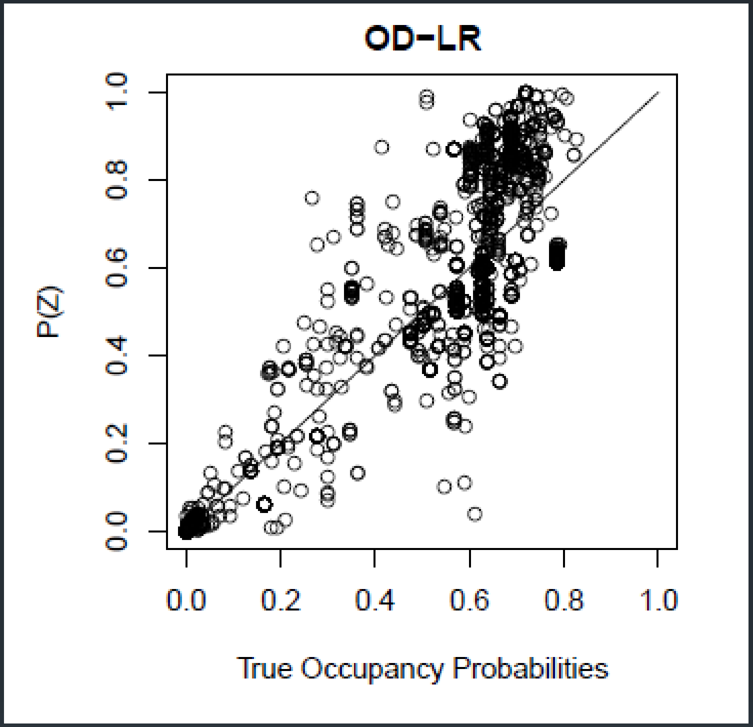
$d_{it} = G(W_{it})$ "detection probability"

Standard Approach: Log Linear (logistic regression) models

- $\log \frac{F(X_i)}{1-F(X_i)} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_J X_{iJ}$
- $\log \frac{G(W_{it})}{1-G(W_{it})} = \alpha_0 + \alpha_1 W_{it1} + \cdots + \alpha_K W_{itK}$
- Fit via maximum likelihood

Results on Synthetic Species with Nonlinear Dependencies

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

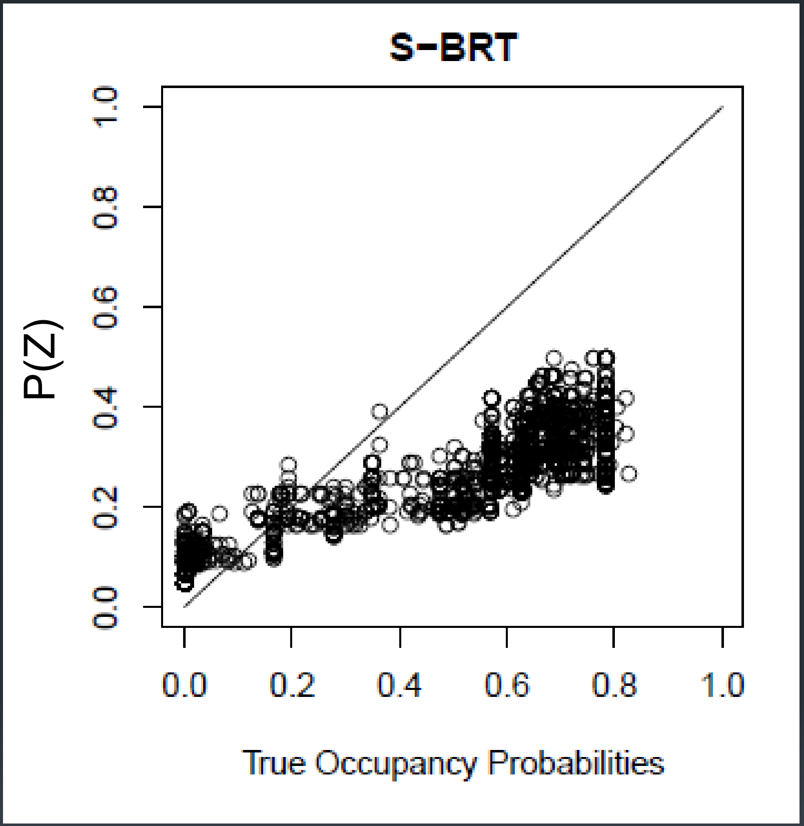


A Flexible Predictive (non-Latent) Model

- 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

Predictive Model 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

Probabilistic Graphical Models

- Advantages
  - Supports latent variables
- Disadvantages
  - Hard to use
    - Model must be carefully designed
    - Data must be transformed to match model assumptions
  - Model has fixed complexity so either under-fits or over-fits

Flexible Nonparametric Models

- Advantages
  - Model complexity adapts to data complexity
  - Easy to use "off-the-shelf"
- Disadvantages
  - Do not support latent variables

The Dream

Probabilistic Graphical Models

Flexible Nonparametric Models

Flexible Nonparametric Probabilistic Models

A Simple Idea: Parameterize F and G as boosted trees

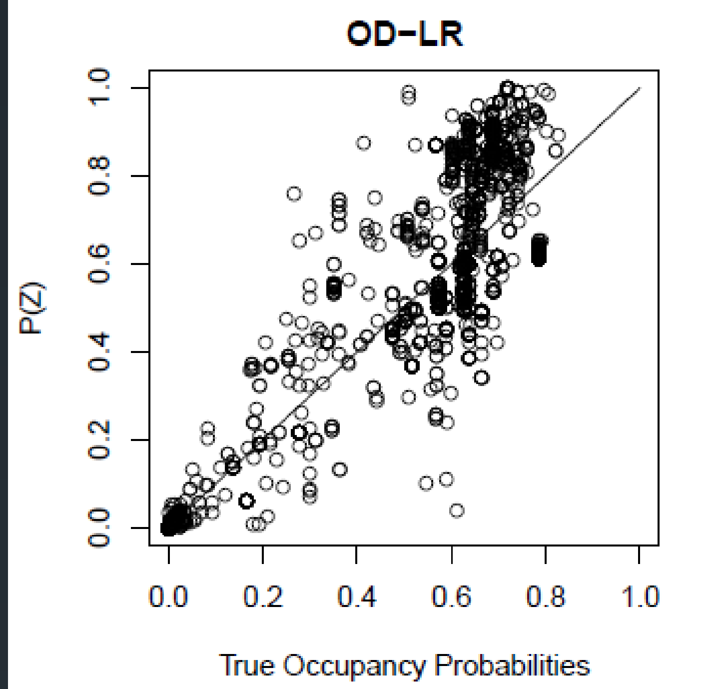
- $\log \frac{F(X)}{1-F(X)} = f^0(X) + \rho_1 f^1(X) + \cdots + \rho_L f^L(X)$
- $\log \frac{G(W)}{1-G(W)} = g^0(W) + \eta_1 g^1(W) + \cdots + \eta_L g^L(W)$
- Perform functional gradient descent in F and G

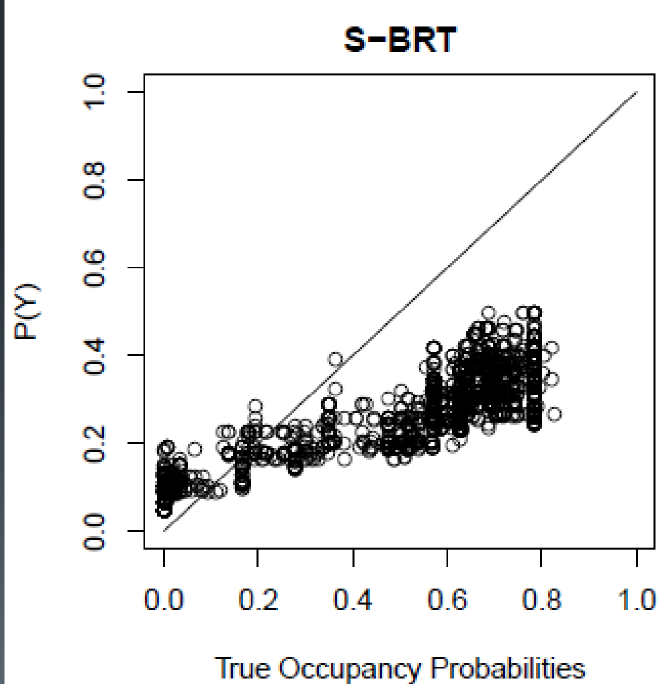
- See also...
  - Kernel logistic regression
  - Non-parametric Bayes
  - RKHS embeddings of probability distributions

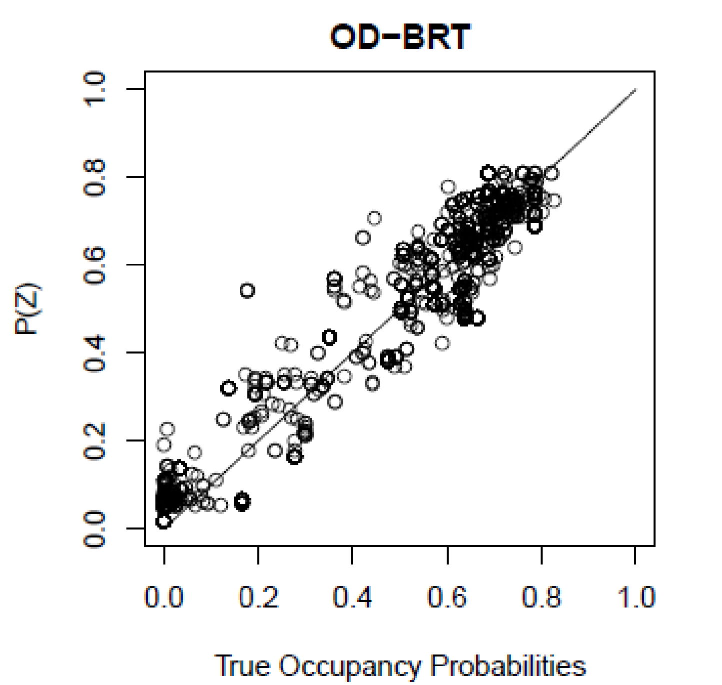
Results: OD-BRT

(Hutchinson, Liu & Dietterich, AAAI 2010)

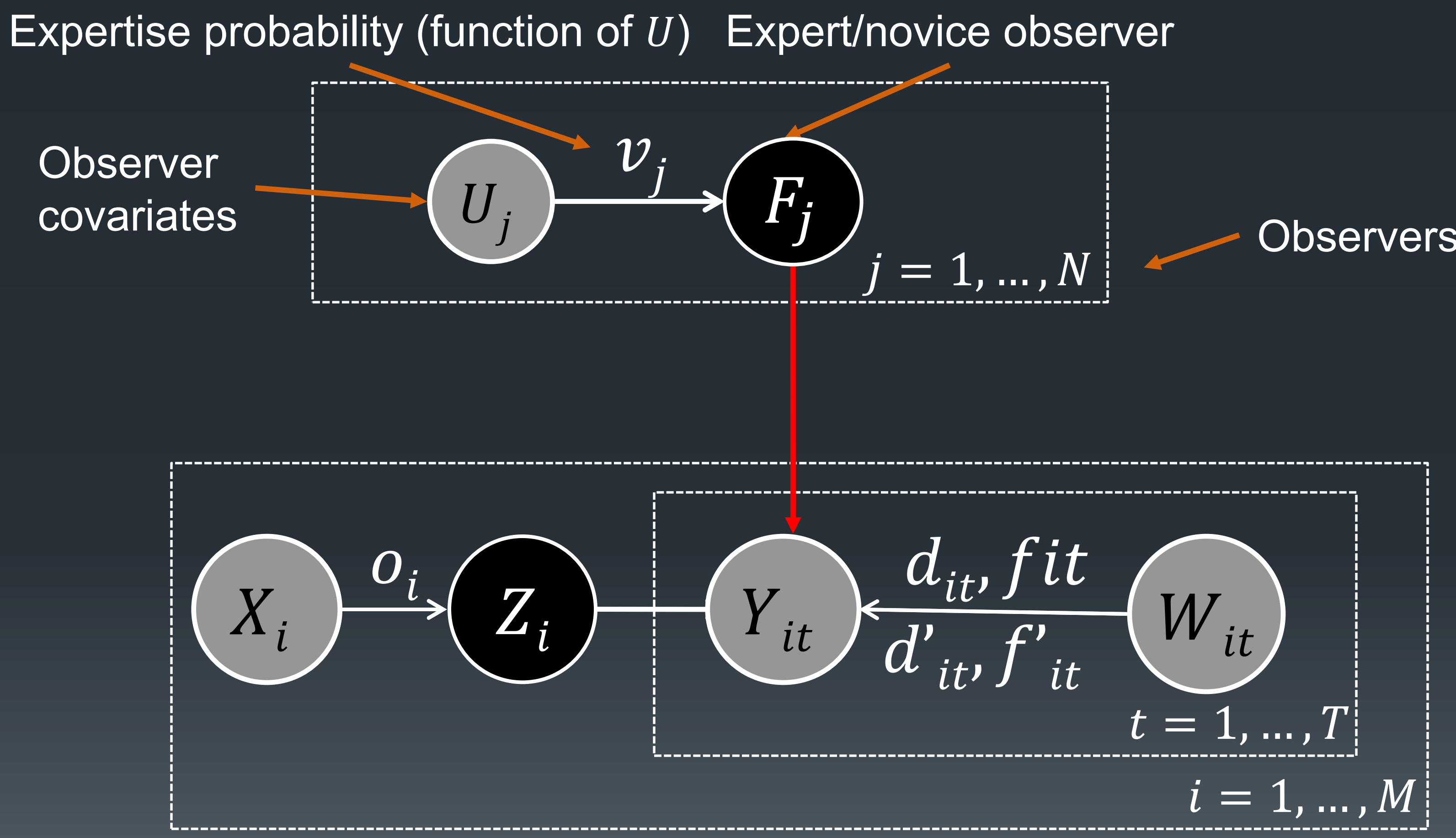
- Occupancy probabilities are predicted very well



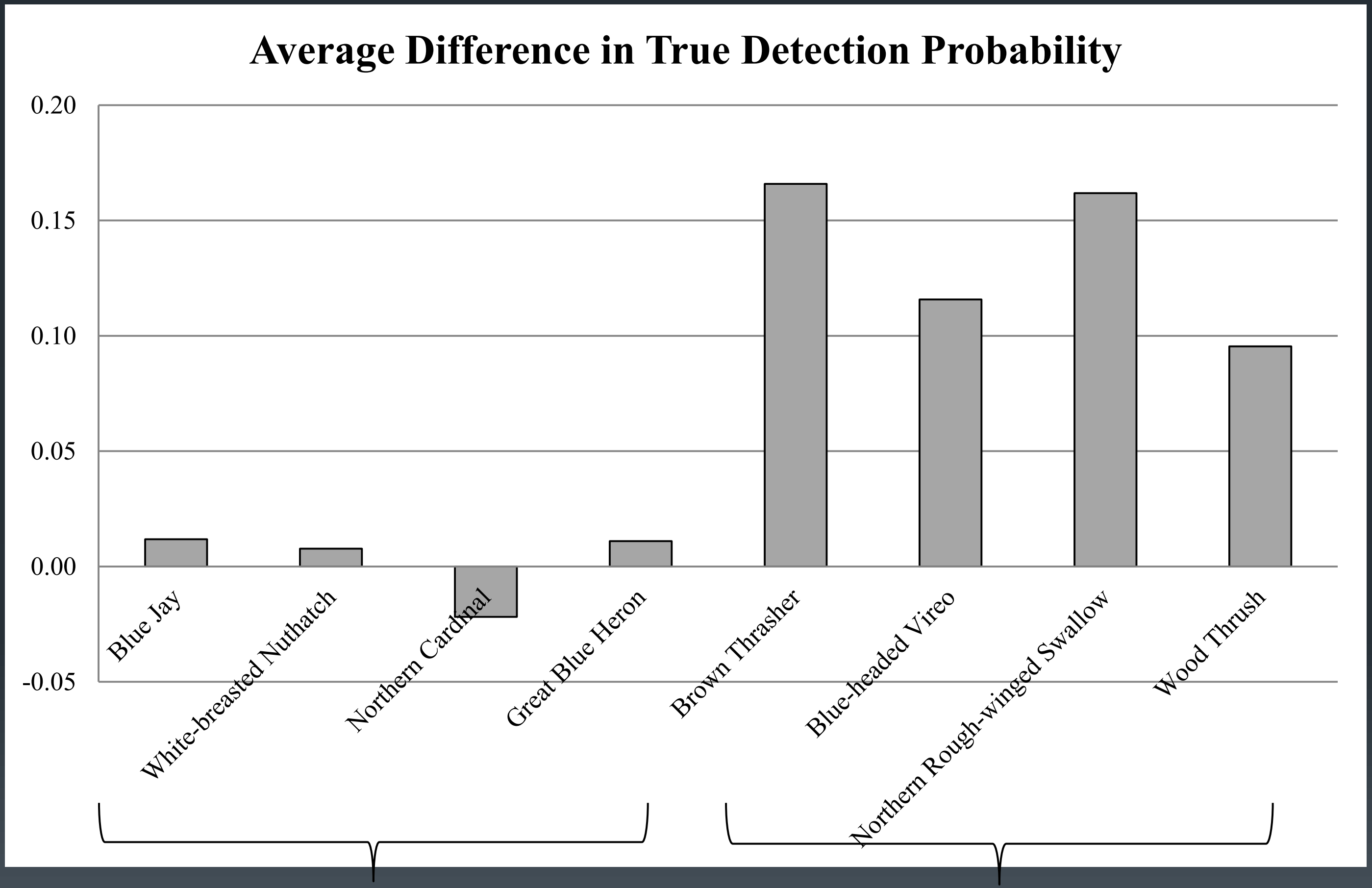




Handling Variable Expertise



Expert vs. Novice Differences



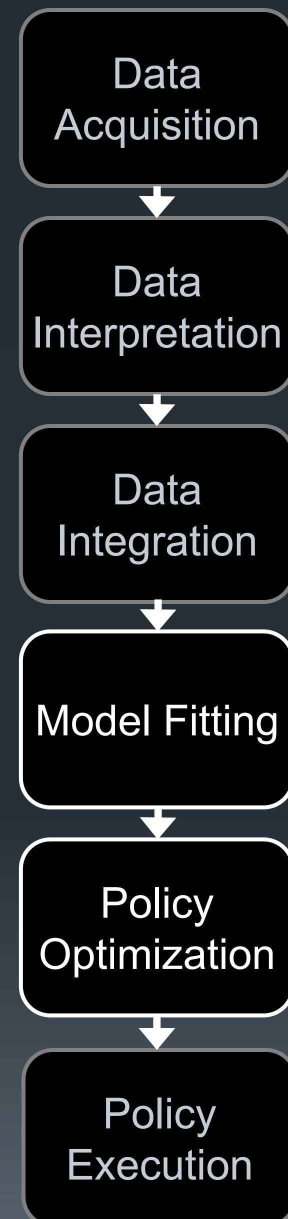
Common birds

Hard-to-detect birds

Yu, et al, 2010

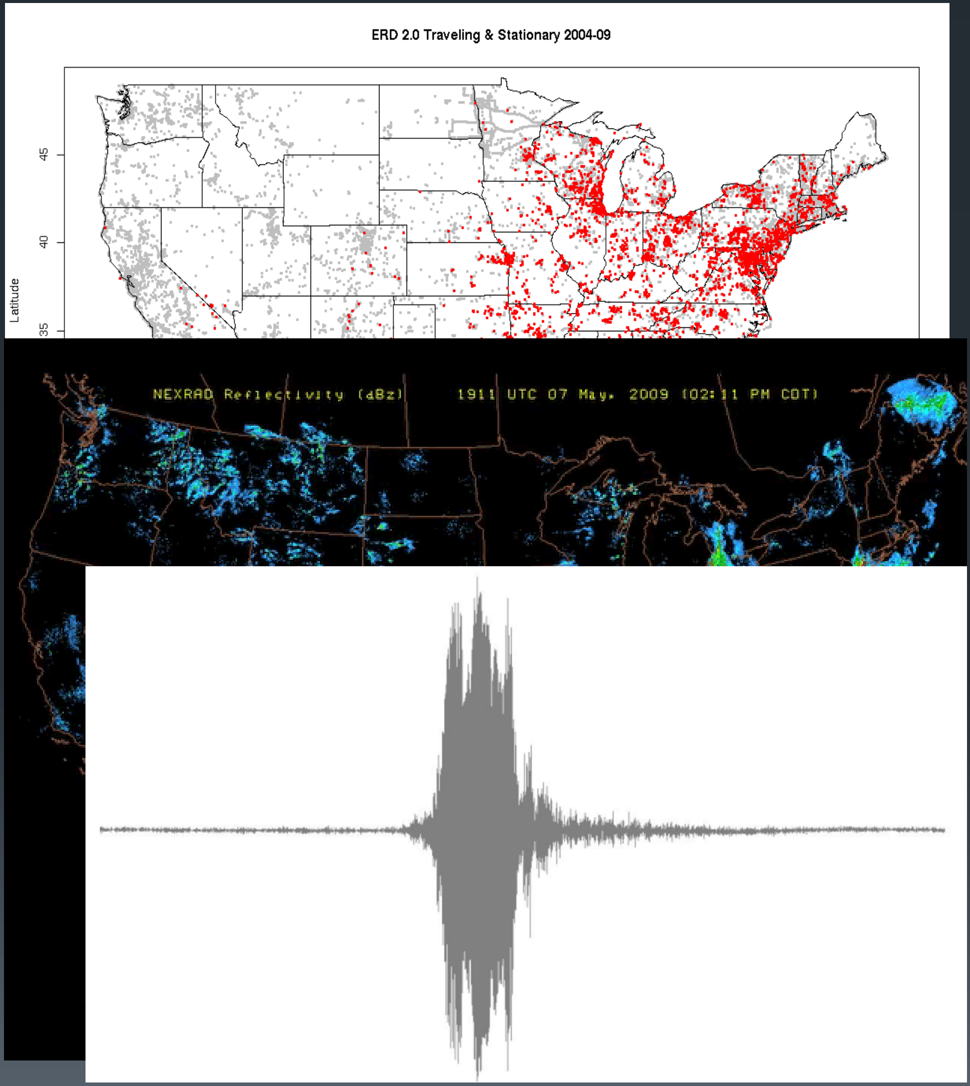
Drill Down:
Three Projects at Oregon State

- Species Distribution Modeling with Imperfect Observations
  - Explicit Observation Models
  - Flexible Latent Variable Models
- Models of Bird Migration
  - Collective Graphical Models
- Policy Optimization
  - Controlling Invasive Species
  - Algorithms for Large Spatial MDPs



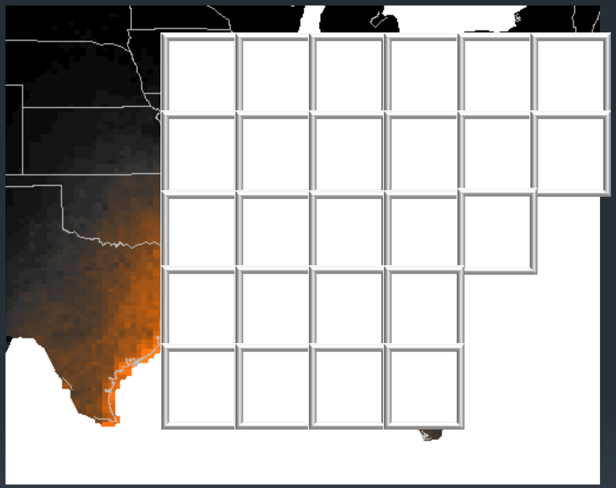
BirdCast: Understanding and Forecasting Bird Migration

- Available data:
  - eBird observations
  - NEXRAD weather radar
  - acoustic monitoring stations
  - weather data
  - weather forecast
- Goals:
  - predict spatial distribution of each species 24- and 48-hours in advance
  - understand what factors drive bird migration
    - wind speed and direction?
    - temperature?
    - relative humidity?
    - absolute or relative timing?
    - food availability?



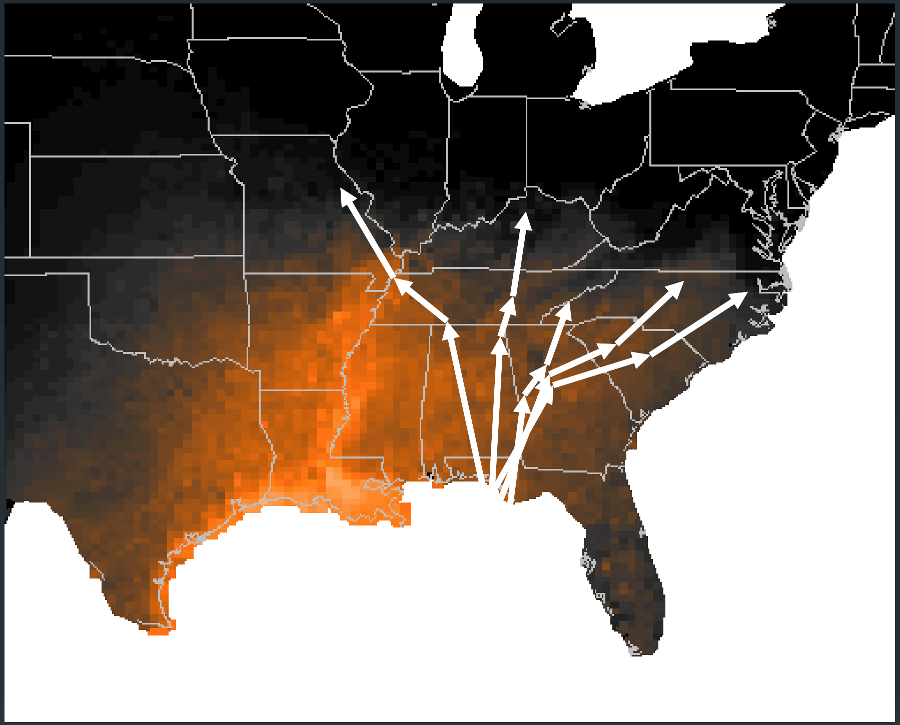
Modeling Goal: Spatial Hidden Markov Model

- Define a grid over the US
- Let n_i^t be the number of birds in cell i at time t
- Learn a probability transition matrix that depends on the features
  - wind, temperature, time, etc.



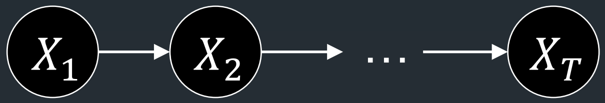
Problem:
We have only aggregate data

- The data we wish we had:
  - tracks of individual birds
- The data we have:
  - ebird: aggregate counts of anonymous birds
  - radar: birds per km^3 summed over all species
  - ...

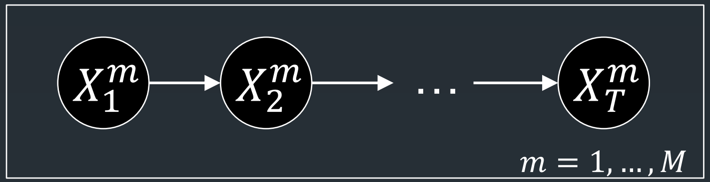


Solution:
Collective Graphical Models

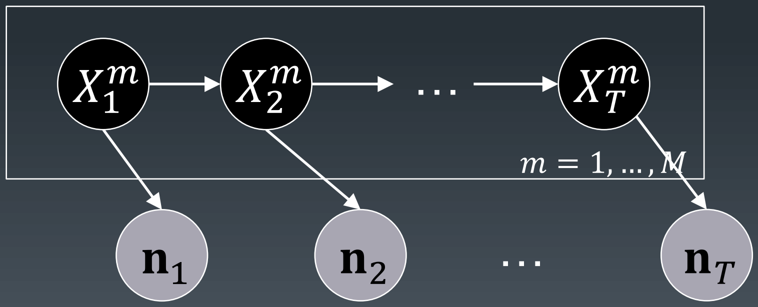
Individual model:
Markov chain on grid cells



Population model:
iid copies of individual model

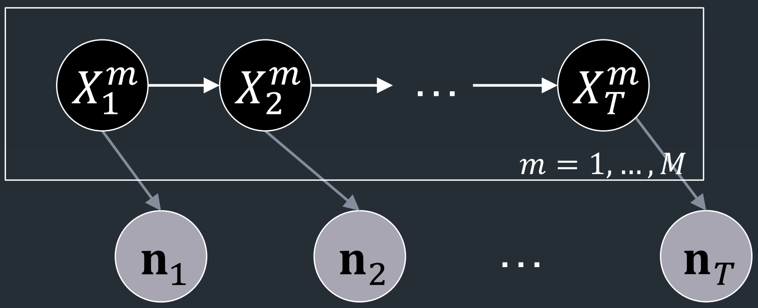


Derive aggregate observations

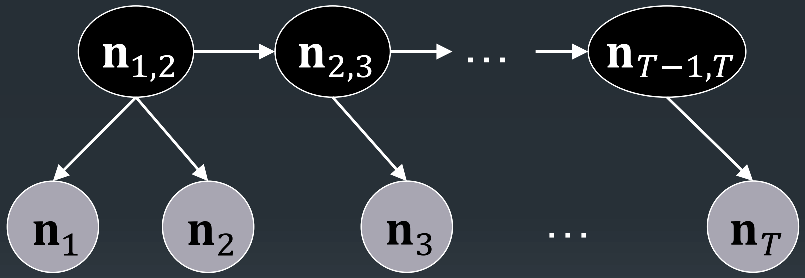


Solution: Collective Graphical Models (2)

Derive aggregate observations



Marginalize out individuals: chain-structured model on sufficient statistics

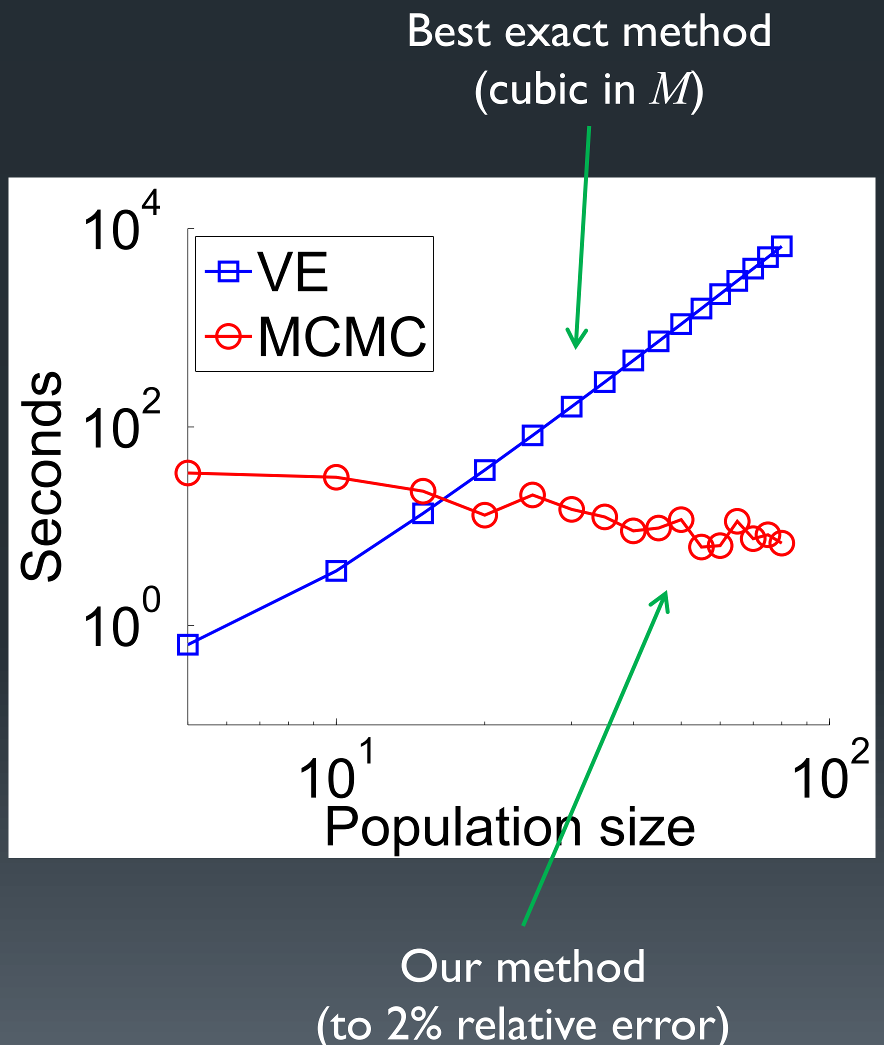


Transition counts

Note: MAP estimates of $\mathbf{n}_{ij}$ are sufficient statistics of the individual model
We don't need to reconstruct individual tracks to fit the individual model

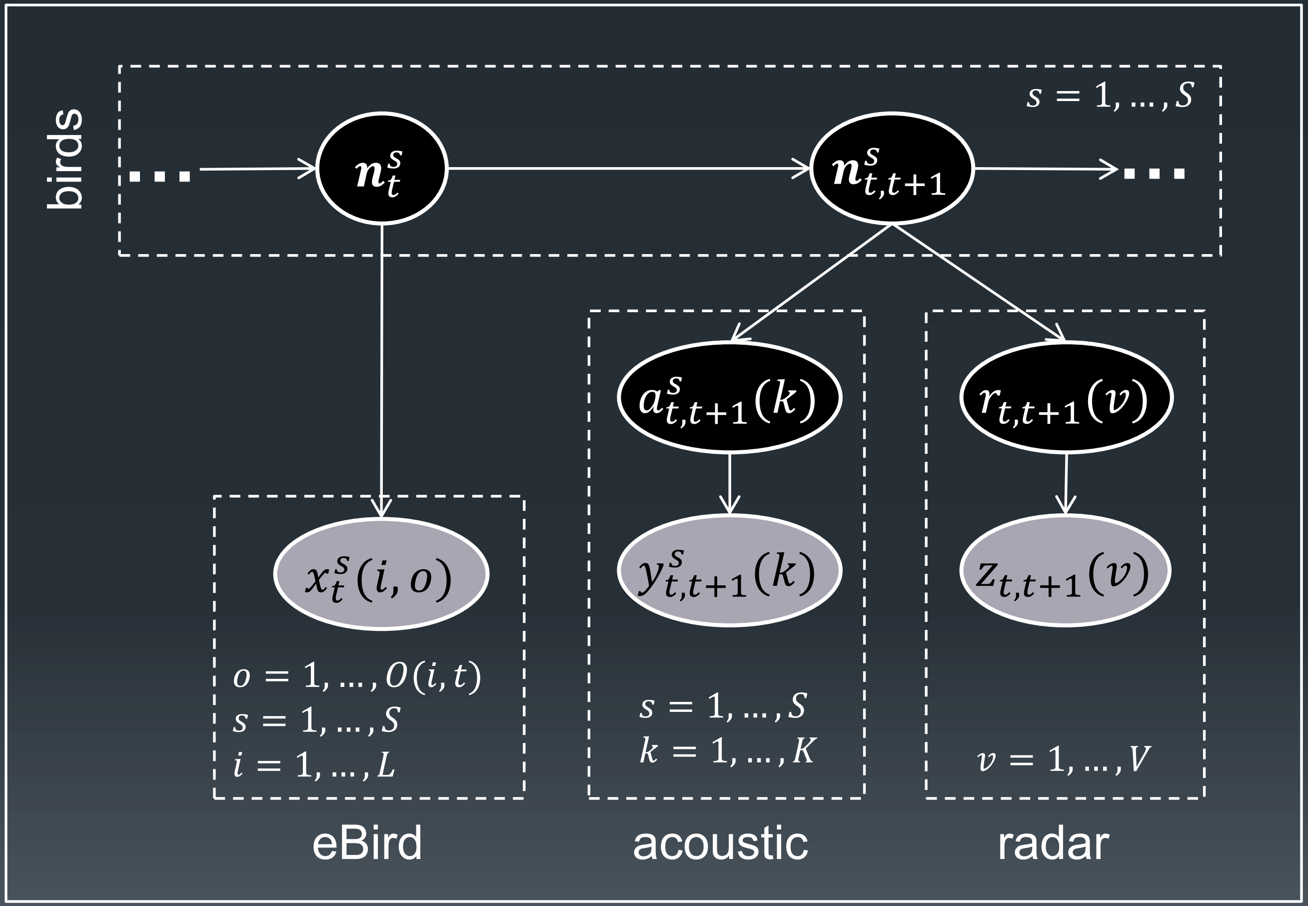
Inference in Collective Graphical Models (Sheldon & Dietterich, NIPS 2011)

- Model Fitting via EM
  - Requires sampling from $P(\boldsymbol{n}_{t,t+1} | \boldsymbol{n}_1, \ldots, \boldsymbol{n}_T)$
    - posterior distribution of "flows" through the HMM trellis
  - Fast Gibbs Sampler that respects Kirchoff's laws
    - running time is independent of population size

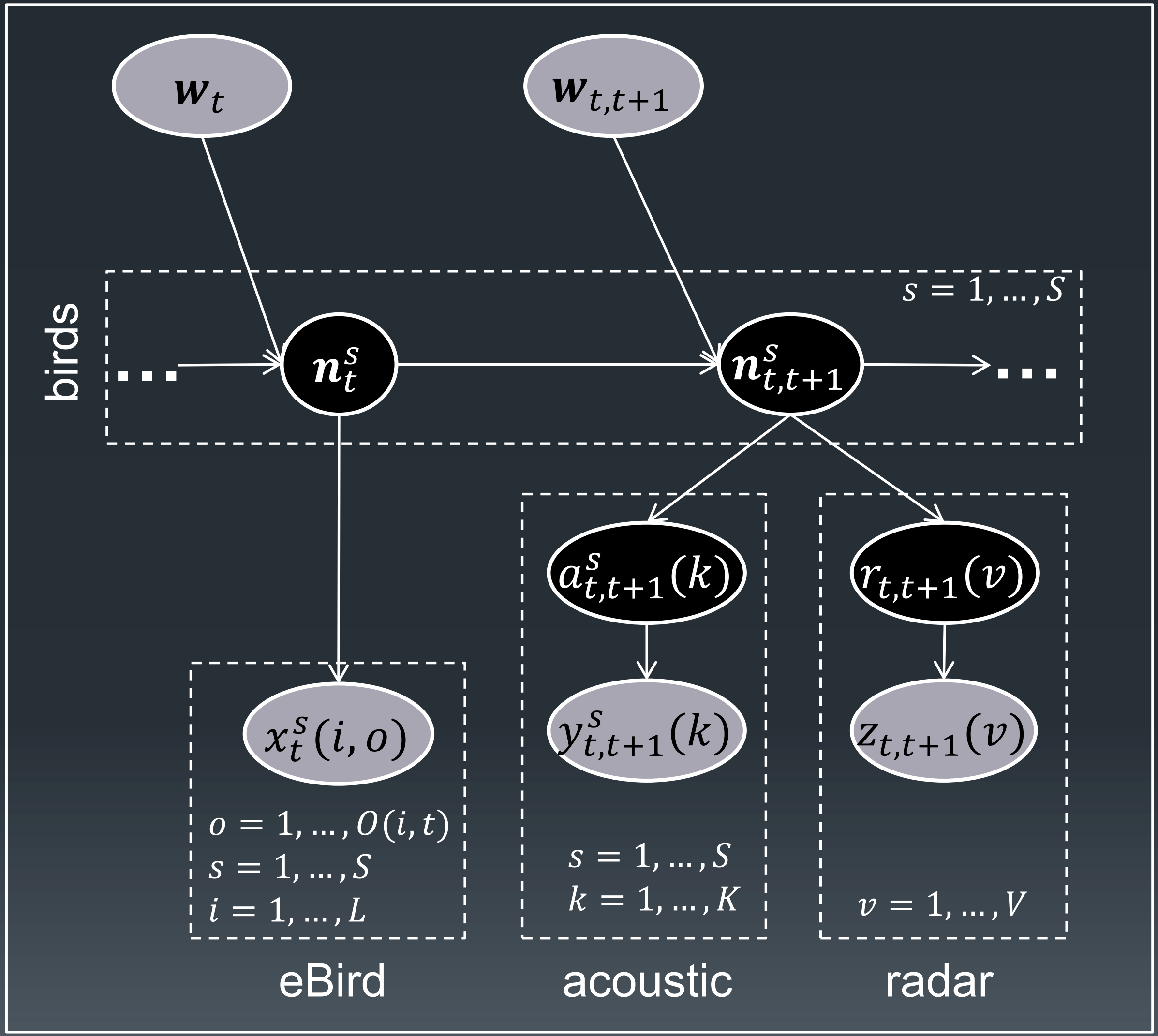


The Migration Model

- Species s
- Observers o
- Sites i
- Acoustic stations k
- Radar sites v

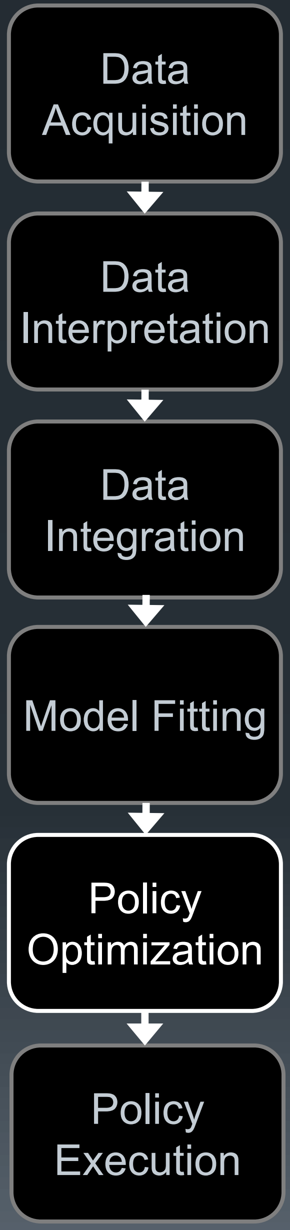


With Added Covariates



Drill Down:
Three Projects at Oregon State

- Species Distribution Modeling with Imperfect Observations
  - Explicit Observation Models
  - Flexible Latent Variable Models
- Models of Bird Migration
  - Collective Graphical Models
- Policy Optimization
  - Controlling Invasive Species
  - Algorithms for simulator-defined MDPs

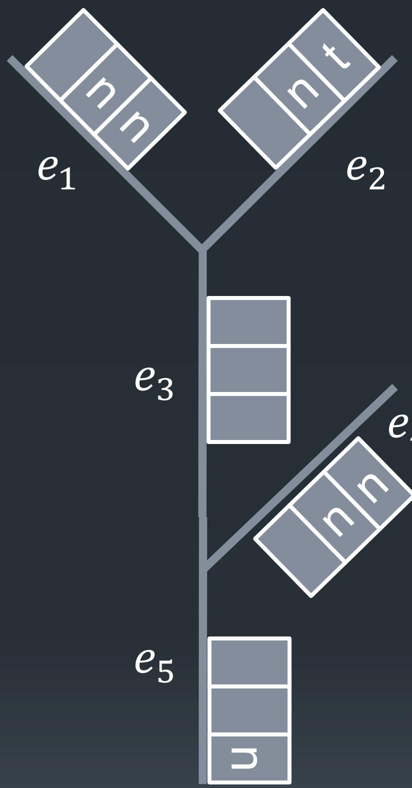


Invasive Species Management in River Networks

- Tamarisk: invasive tree from the Middle East
  - Out-competes native vegetation for water
  - Reduces biodiversity
- What is the best way to manage a spatially-spreading organism?

Markov Decision Process

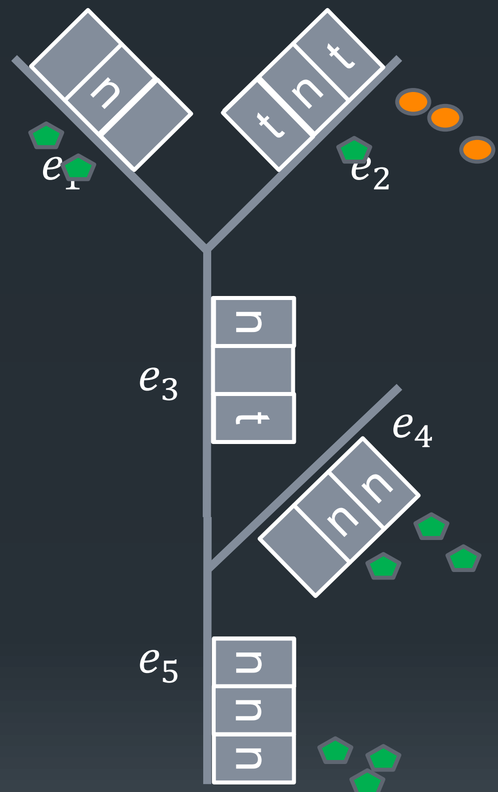
- Tree-structured river network
  - Each edge $e \in E$ has H “sites” where a tree can grow.
  - Each site can be
    - {empty, occupied by native, occupied by invasive}
  - # of states is 3^{EH}
- Management actions
  - Each edge: {do nothing, eradicate, restore, eradicate+restore}
  - # of actions is 4^{E}



Dynamics and Objective

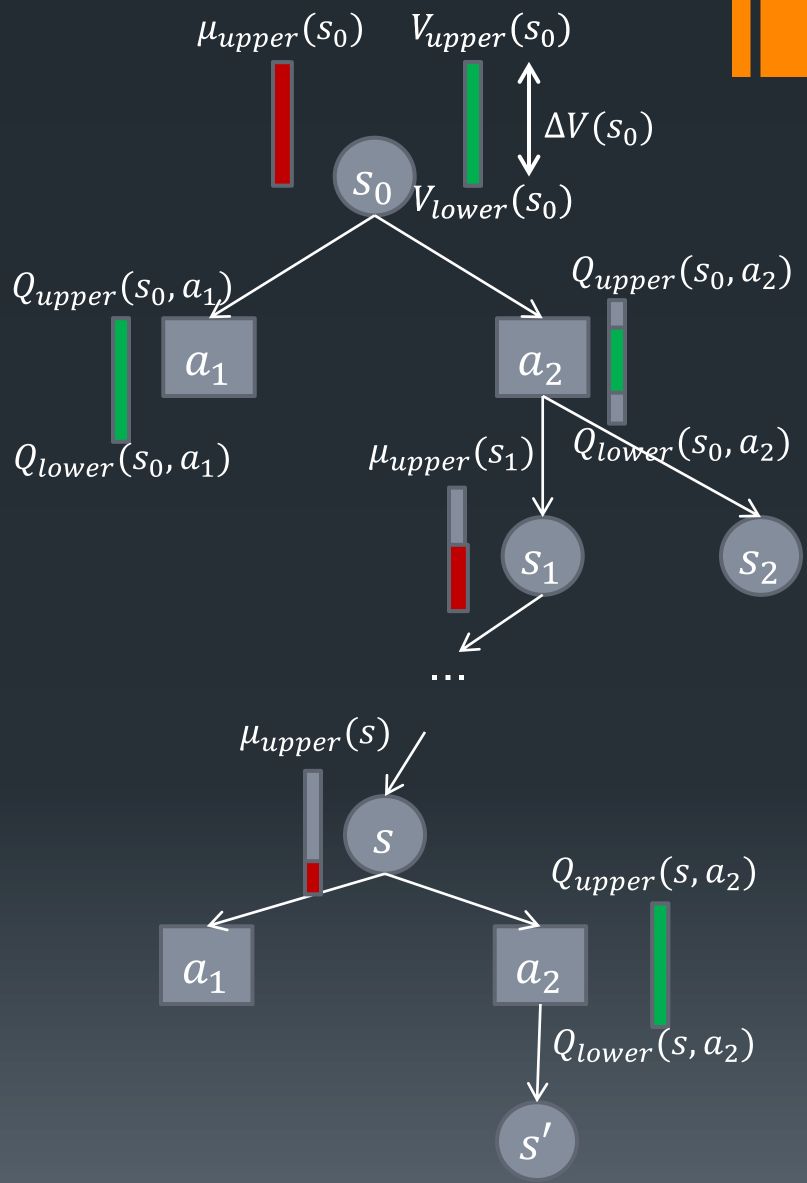
- Dynamics:
  - In each time period
    - Natural death
    - Seed production
    - Seed dispersal (preferentially downstream)
    - Seed competition to become established
  - Couples all edges because of spatial spread
  - Inference is intractable

- Objective:
  - Minimize expected discounted costs (sum of cost of invasion plus cost of management)
  - Subject to annual budget constraint



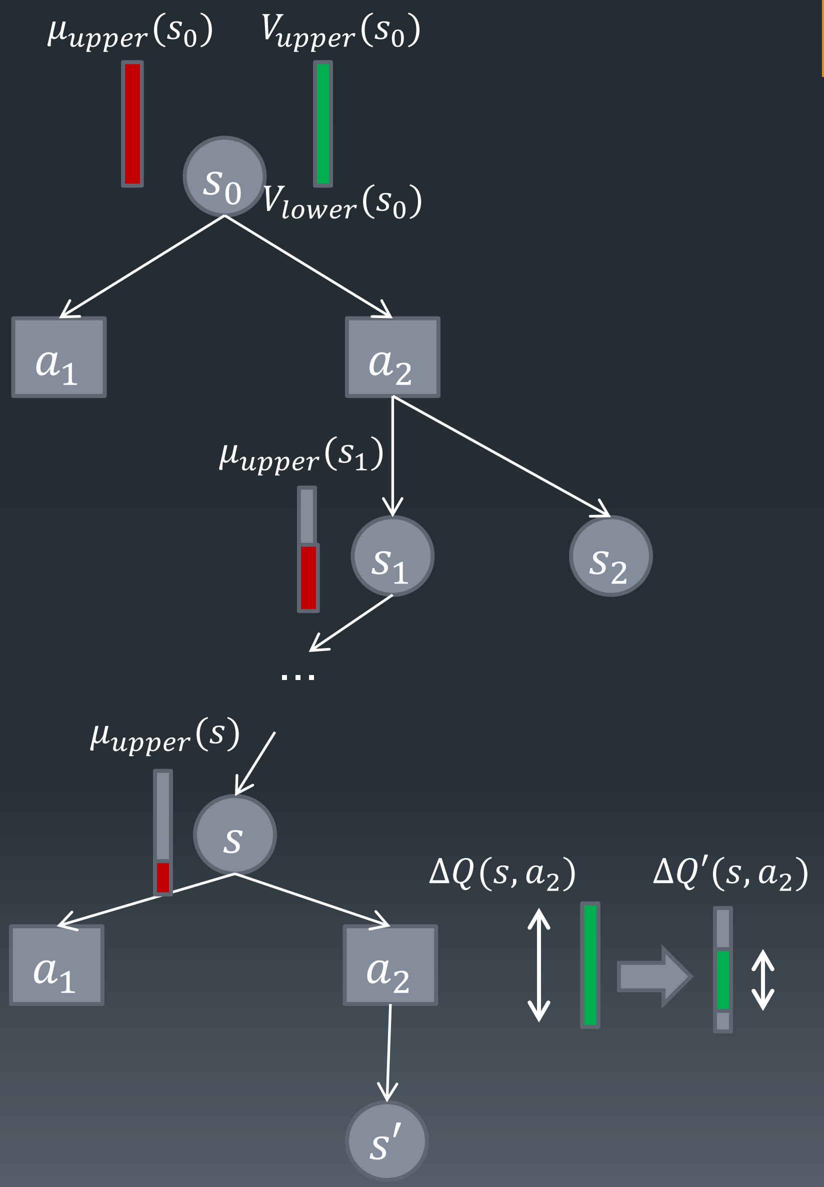
Algorithm DDV

- Goal: Compute PAC-optimal policy while minimizing simulator calls
- Explicit representation of the MDP (Transition matrix and Q table)
- Confidence intervals $Q_{lower}(s,a)$ and $Q_{upper}(s,a)$
- Confidence interval on $V(s_0)$
- Upper bound on discounted state occupancy probability $\mu_{upper}(s)$
  - $\mu^{\pi}(s) = \sum_t \gamma^t P(s^t = s | s^0 = s_0, \pi)$
- Measure of uncertainty:
  - $\Delta V(s_0) = V_{upper}(s_0) - V_{lower}(s_0)$



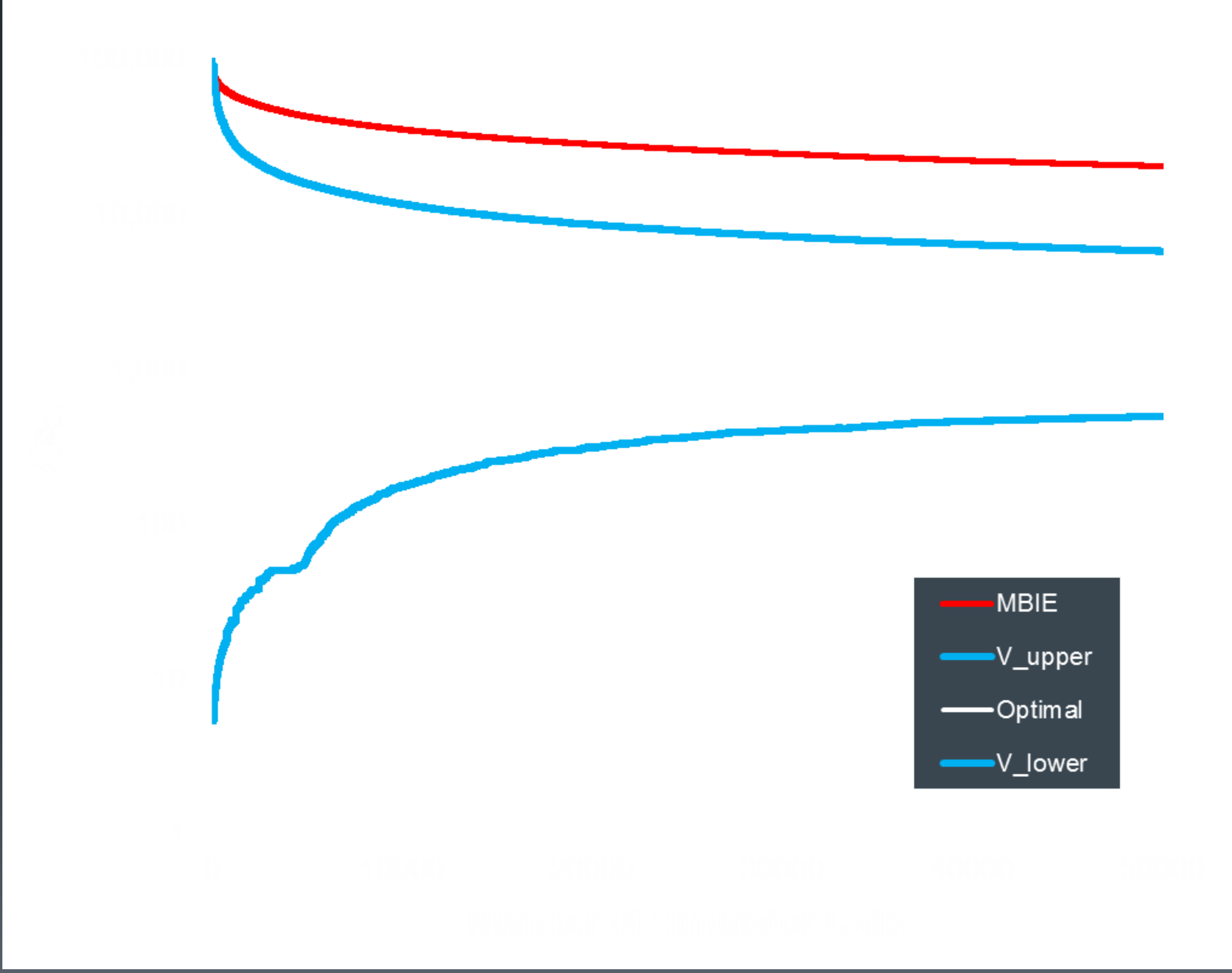
Algorithm DDV

- Exploration heuristic:
  - Exploring (s, a_2) will cause a local reduction in $\Delta Q(s, a_2) = Q_{upper}(s, a_2) - Q_{lower}(s, a_2)$
  - The impact of this on $\Delta V(s_0)$ can be approximated by $\mu_{upper}(s)[\Delta Q(s, a_1) - \Delta Q'(s, a_1)]$
  - Explore the (s, a) that maximizes $\mu_{upper}(s)[\Delta Q(s, a) - \Delta Q'(s, a)]$



Results on "RiverSwim" benchmark

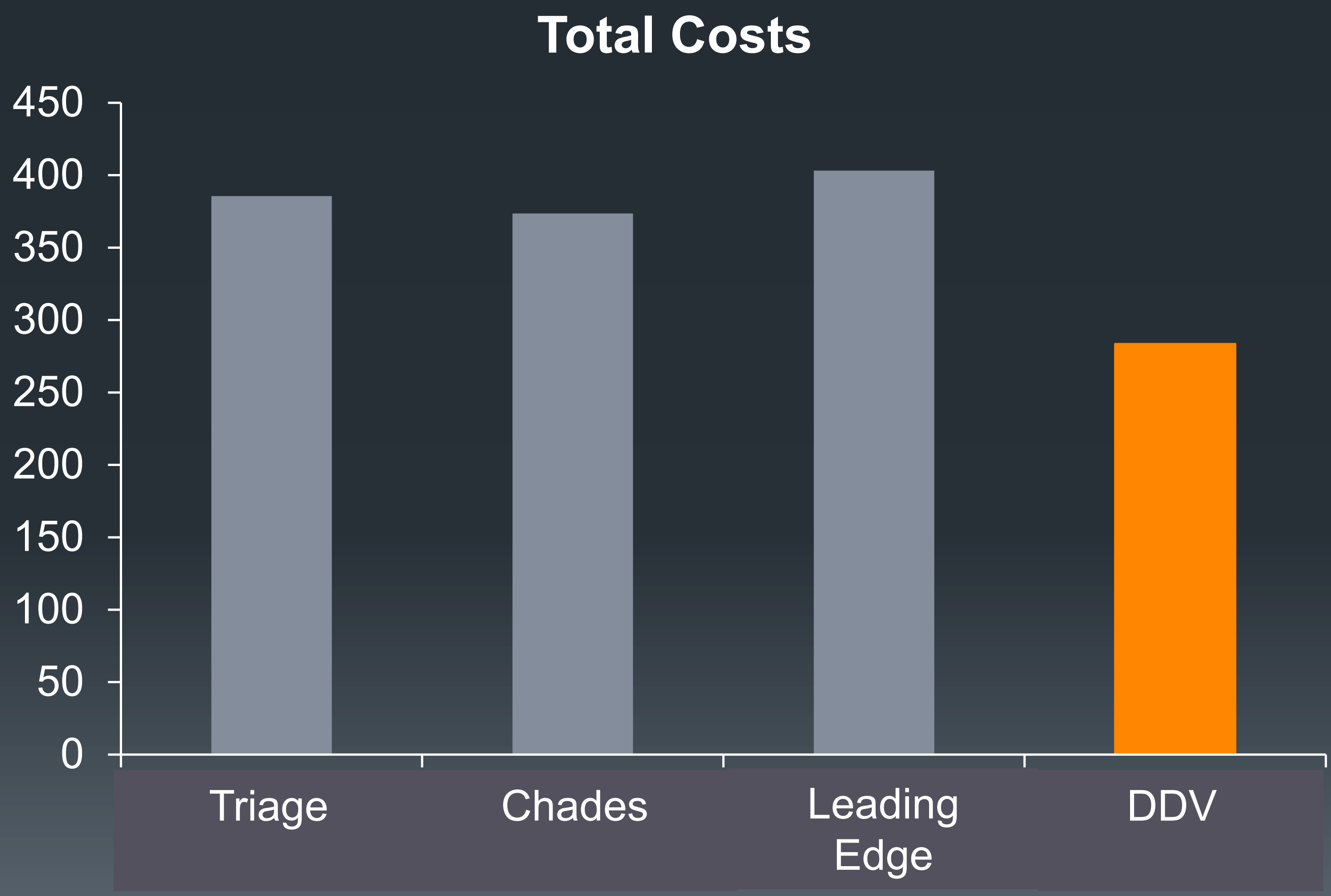
- Comparison with Strehl & Littman (2008) Model-Based Interval Estimation (MBIE)
- DDV reduces the uncertainty in $V(s_0)$ much faster than MBIE
  - note log scale
- Both algorithms have PAC guarantees



Published Rule of Thumb Policies for Invasive Species Management

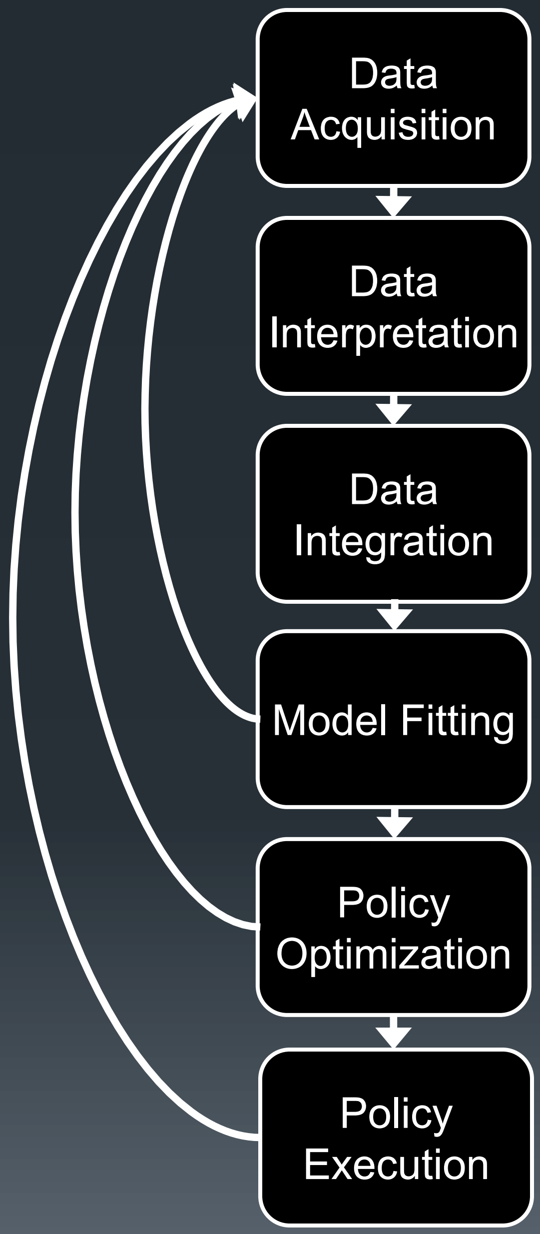
- Triage Policy
  - Treat most-invaded edge first
  - Break ties by treating upstream first
- Leading edge
  - Eradicate along the leading edge of invasion
- Chades, et al.
  - Treat most-upstream invaded edge first
  - Break ties by amount of invasion
- DDV
  - Our PAC solution

Cost Comparisons: Rule of Thumb Policies vs. DDV



Summary

- Data → Models → Policies
- Three projects at Oregon State:
  - Species Distribution Modeling with Imperfect Observations
    - Flexible Latent Variable Models
  - Models of Bird Migration
    - Collective Graphical Models
  - Policy Optimization
    - Algorithms for simulator-defined MDPs



Distinctive Characteristics of Sustainability Problems

- Goal is typically to encourage or prevent spatial spread
  - Encourage spread of endangered species
  - Manage spread of fire
  - Prevent spread of diseases and invasive species
  - Over long time horizons
  - Resulting MDPs are immense
  - Dynamics are typically available only via a simulator

- Data are extremely noisy, heterogeneous, and incomplete
  - Need to learn latent process dynamical models from this data

- Optimization is based on learned models
  - Need to be robust to incorrect models
  - Need to be robust to the unknown unknowns
  - Risk sensitive:
    - avoid species extinctions
    - avoid catastrophic fires

Computational Sustainability

- There are many opportunities for computing to contribute to sustainable ecosystem management
- There are many challenging machine learning research problems to be solved
- Institute for Computational Sustainability: http://www.computational-sustainability.org/

Thank-you

- Rebecca Hutchinson, Liping Liu: Boosted Regression Trees in OD models
- Dan Sheldon: Collective Graphical Models
- Steve Kelling, Andrew Farnsworth, Wes Hochachka, Daniel Fink: BirdCast
- H. Jo Albers, Kim Hall, Majid Taleghan, Mark Crowley: Tamarisk
- Carla Gomes for spearheading the Institute for Computational Sustainability

- National Science Foundation Grants 0705765, 0832804, and 0905885

Questions?