Machine Learning for Computational Sustainability

Tom Dietterich

Oregon State University

In collaboration with Dan Sheldon, Sean McGregor, Majid Taleghan, Rachel Houtman, Claire Montgomery, Kim Hall, H. Jo Albers
and the Cornell Lab of Ornithology

ICGG 2012

Data → Models → Policies

- Managing Earth's ecosystems is a control problem
- We need
  - A model of the "plant"
  - An optimal controller

- Our research:
  - Machine learning methods for modeling ecosystem dynamics from data
  - Approximate Stochastic Dynamic Programming algorithms for creating optimal controllers

Outline

- Model of the plant
  - Towards a Dynamical Model of Bird Migration
- Optimal control
  - Managing wildfire in Eastern Oregon
  - Managing invasive plant species in the US Intermountain West

Bird Distribution and Migration

- Management:
  - Many bird populations are declining
  - Predicting aircraft-bird interactions
  - Siting wind farms
  - Night-time lighting of buildings (esp. skyscrapers)
  - How will climate change affect bird migration and survival?

- Science:
  - What is the migration decision making policy for each species
    - When to start migrating?
    - How far to fly each night?
    - When to stop over and for how long?
    - When to resume flying?
    - What route to take?

Why bird migration is poorly understood

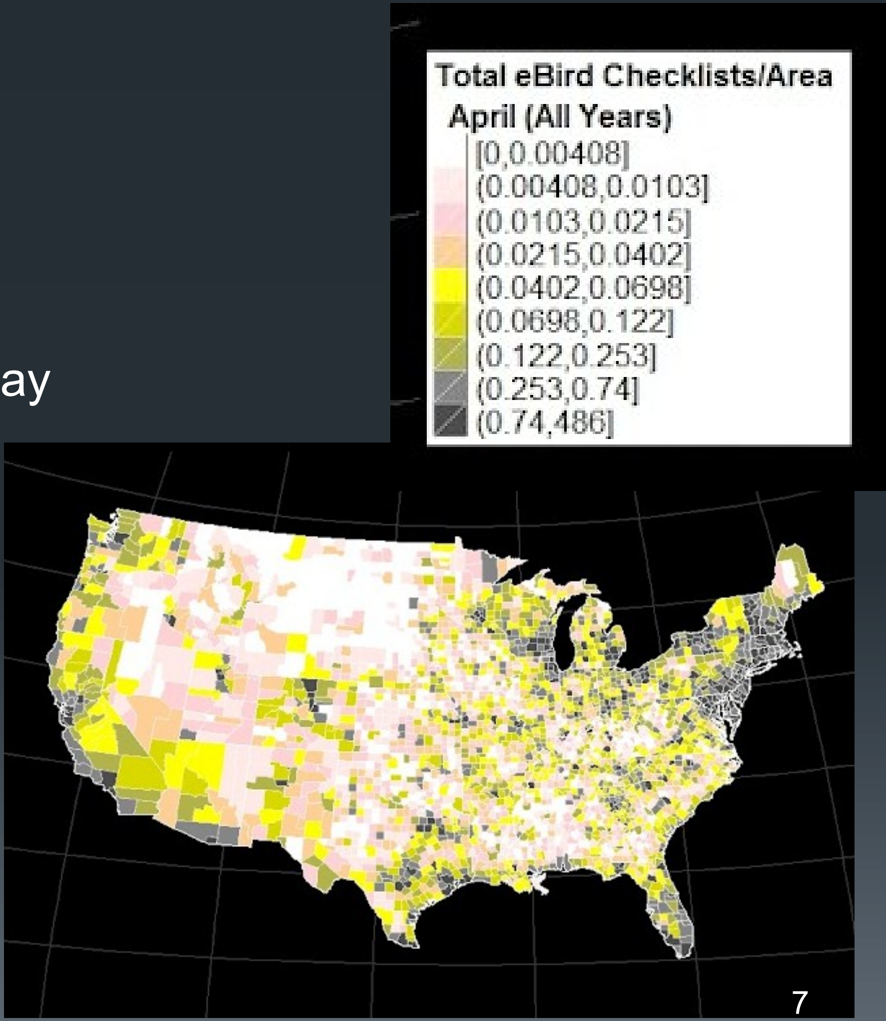
- It is difficult to observe
  - Takes place at continental scale (and beyond)
  - Impossible for the small number of professional ornithologists to collect enough observations
  - Very few birds have been individually tracked

What Data Are Available?

- Birdwatcher count data: eBird.org
- Doppler weather radar
- Night flight calls

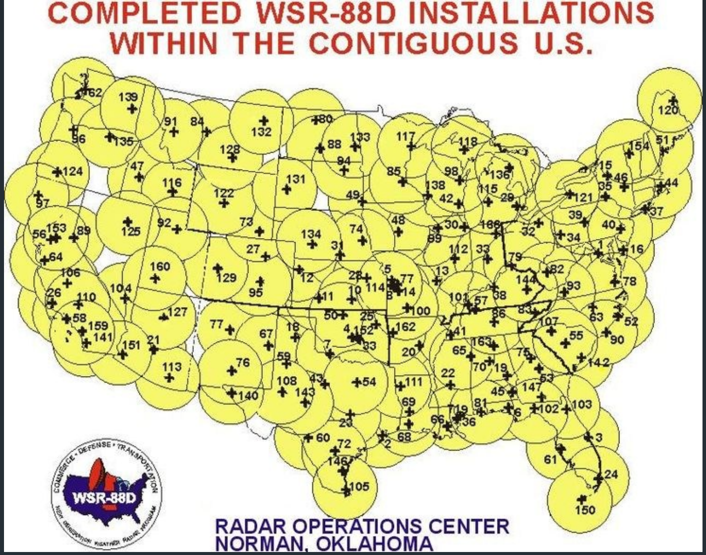
eBird Data

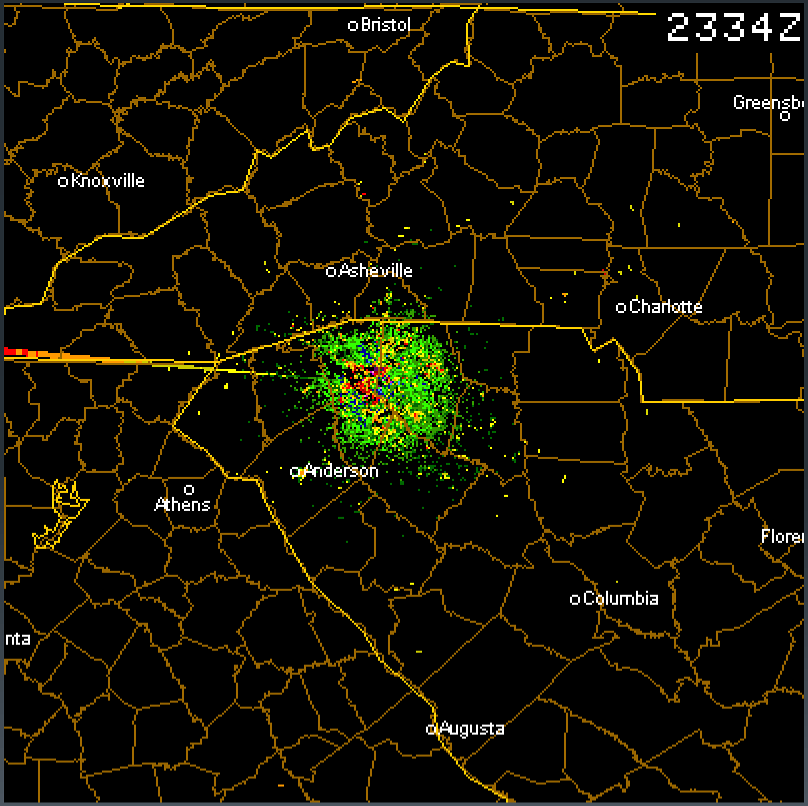
- Bird watchers record their observations in a database through eBird.org.
  - “Citizen Science”
- Dataset available for analysis
- Features
  - LOTS of data!
    - ~3 million observations reported last May
  - All bird species (~3,000)
  - Year-round
  - Continent-scale
- Challenges
  - Variable quality observations
  - No systematic sampling design



Doppler Weather Radar

- Weather radar detects migrating birds

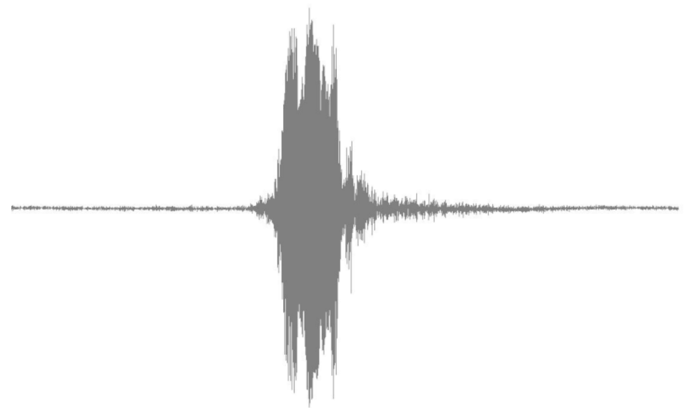


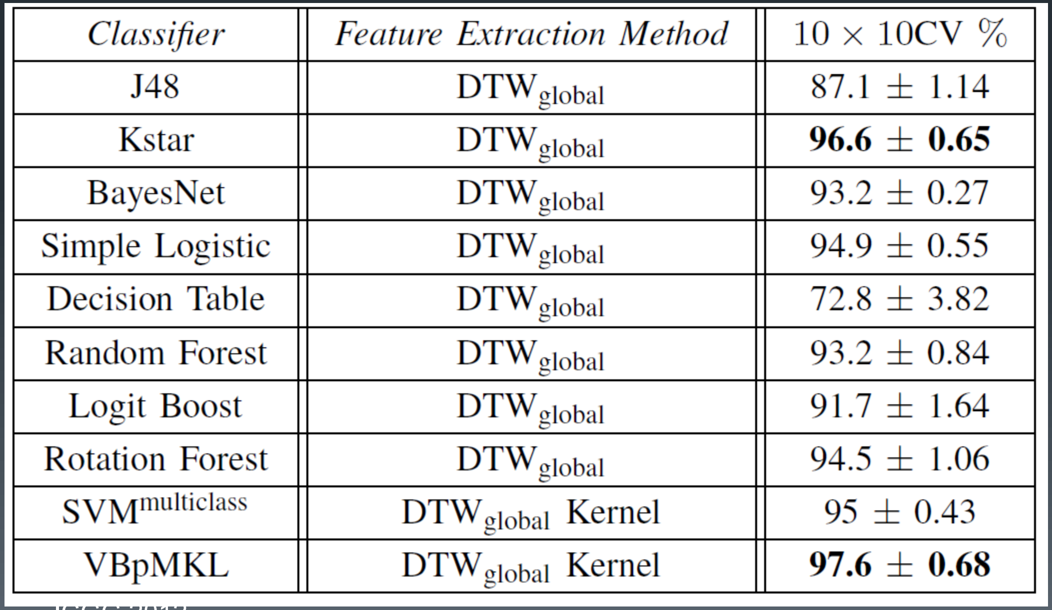


- Can estimate total biomass
- No species information
- Archived data available back to 1995

Night Flight Calls

- Many species of migrating birds emit flight calls that can be identified to species or species group
- New project at Cornell to roll out a large network of recording stations
- Automated detection and classification
- DTW kernel
  - Damoulas, et al, 2010
  - Results on 5 species
  - Clean recordings





Classifier	*Feature Extraction Method*	$10 \times 10\text{CV}$ %
J48	$\text{DTW}_{\text{global}}$	87.1 ± 1.14
Kstar	$\text{DTW}_{\text{global}}$	$\mathbf{96.6 \pm 0.65}$
BayesNet	$\text{DTW}_{\text{global}}$	93.2 ± 0.27
Simple Logistic	$\text{DTW}_{\text{global}}$	94.9 ± 0.55
Decision Table	$\text{DTW}_{\text{global}}$	72.8 ± 3.82
Random Forest	$\text{DTW}_{\text{global}}$	93.2 ± 0.84
Logit Boost	$\text{DTW}_{\text{global}}$	91.7 ± 1.64
Rotation Forest	$\text{DTW}_{\text{global}}$	94.5 ± 1.06
$\text{SVM}^{\text{multiclass}}$	$\text{DTW}_{\text{global}}$ Kernel	95 ± 0.43
VBpMKL	$\text{DTW}_{\text{global}}$ Kernel	$\mathbf{97.6 \pm 0.68}$

Bird Migration Modeling

- Given observations from ebird, radar, flight calls
  - Reconstruct migration behavior

- Given observations + weather forecast
  - Predict migration behavior for next 24 hours, next 5 days

Machine Learning Challenge

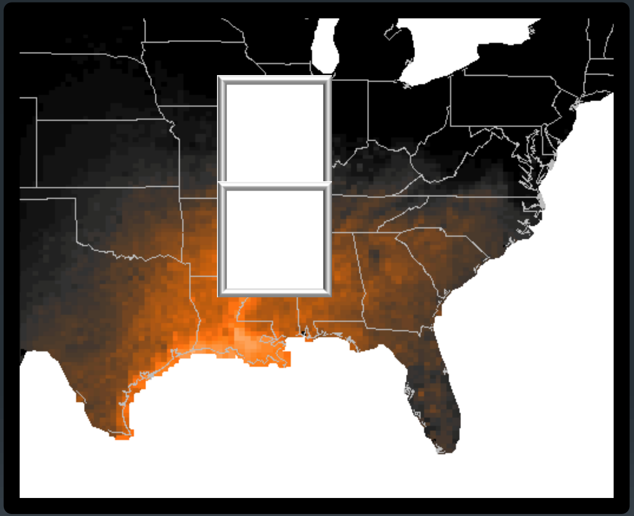
- Migration most naturally described at level of individual behavior, but we can only observe population-level statistics
  - We need a modeling technique to link the two
- Our Approach: Collective Graphical Models

Modeling Approach

- Place a grid of cells over North America
- State of a bird at time t = cell it occupies at time t

Time

Cell		1	2	3
	A	87	61	22
	B	13	39	78



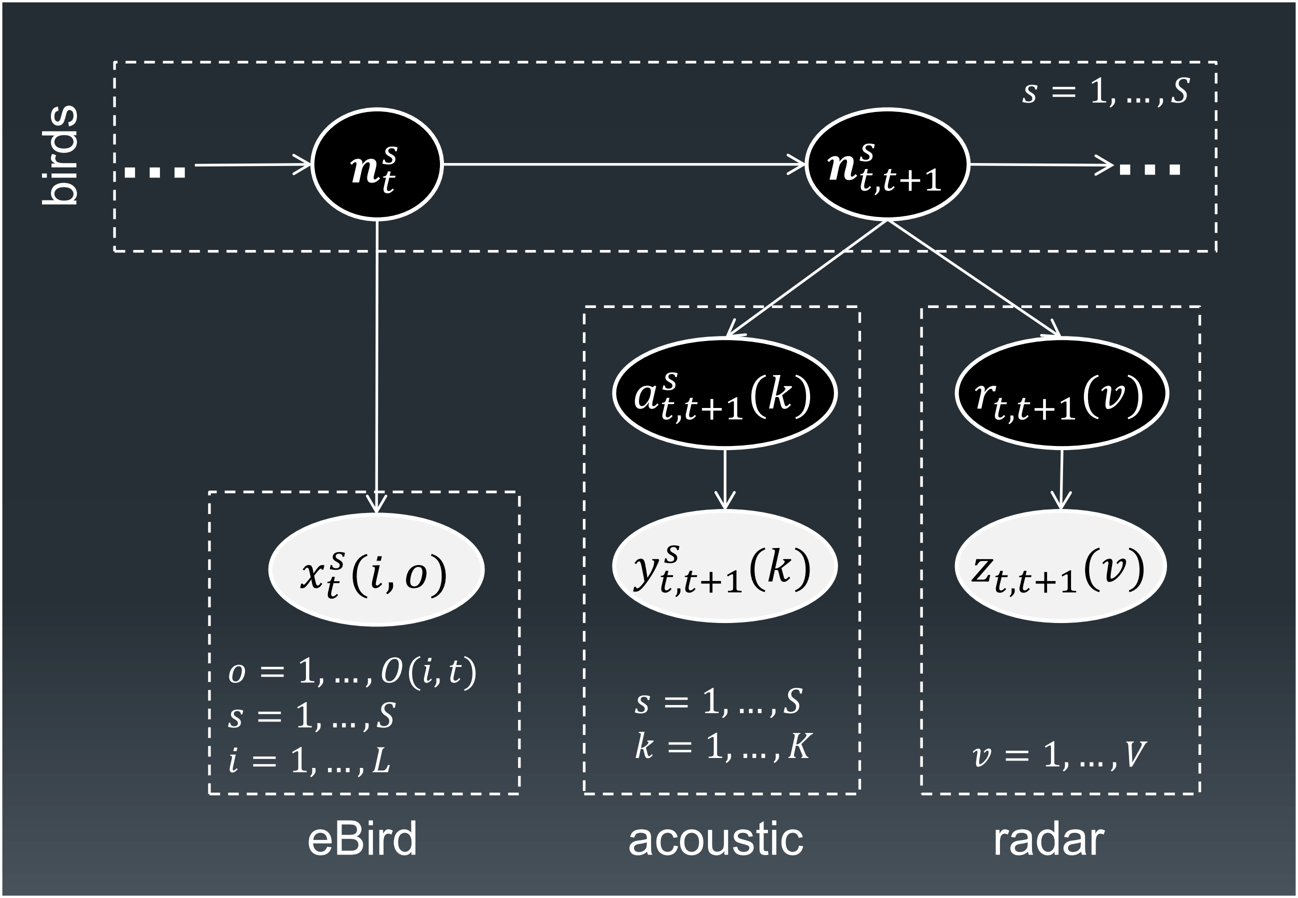
- Aggregate data: does not track individual birds

Hidden Markov Model

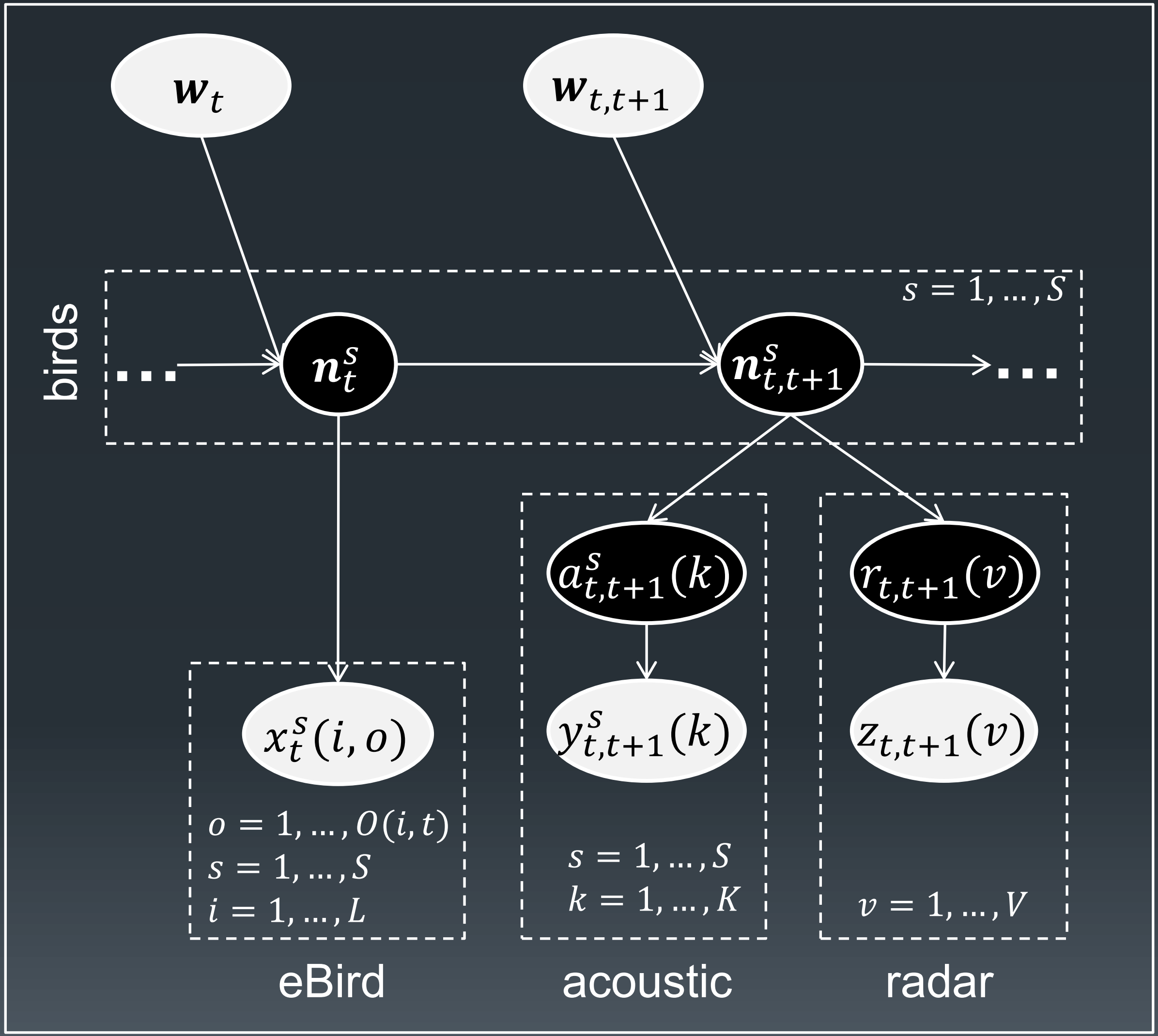
- Let $n_t^s(i)$ be the number of species s in cell i at time t
- Let $\boldsymbol{n}_t^s$ be a vector of $n_s^t(i)$ across all cells $i = 1, \dots, N$
- HMM dynamics: $P(\boldsymbol{n}_t^s | \boldsymbol{n}_{t-1}^s, \mathbf{x}_\mathrm{t})$
  - $P(n_t^s(i) | \boldsymbol{n}_t^s, \boldsymbol{x}_t)$ is the probability distribution of the number of birds in cell i at time t given the locations of the birds in all previous cells in the previous time step $t - 1$
  - $\boldsymbol{x}_t$ is a vector of attributes describing system inputs (weather, temperature, wind, vegetation) that influence bird behavior
- HMM observations: eBird counts, radar, acoustic detections

Diagram of the Migration Model

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



HMM with added features

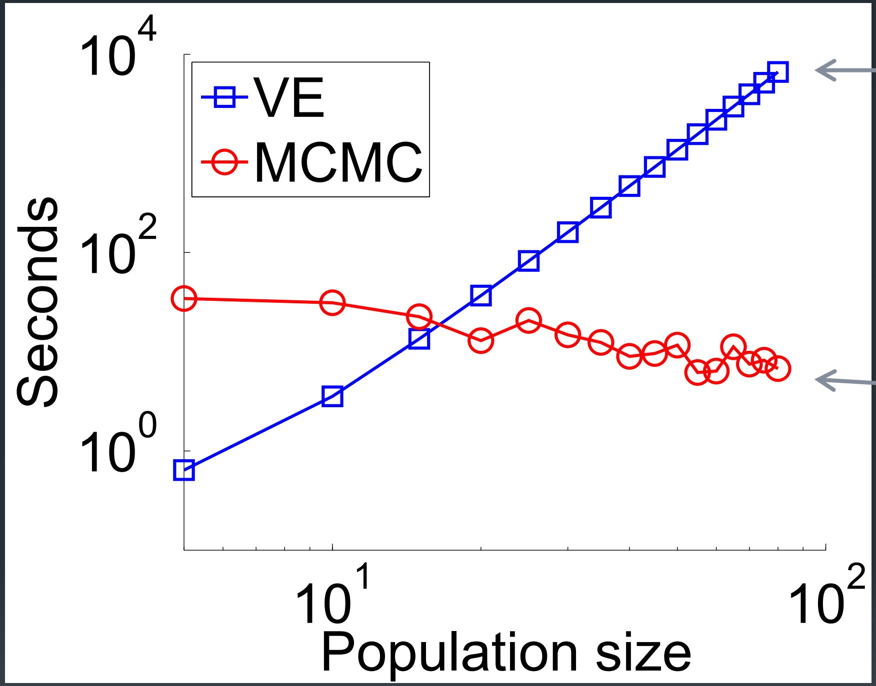


Computational Challenge

- Inference in this HMM is intractable
  - Population size > 5 Billion
- Solution: Gibbs Sampling over the “bird flows” in the HMM
- Sheldon & Dietterich (NIPS 2011) presents a fast Gibbs sampler that takes time *independent of the population size*

Results

- Running time



Best exact method
(cubic in M)

Our method
(to 2% relative error)

- Running time *independent of population size*
  - Previous best: exponential

[Sheldon & Dietterich, NIPS 2011]

Bird Migration Summary

- Fitting Dynamical Models to Multiple Data Sources
  - eBird + radar + night flight calls
  - Collective Graphical Models: General Methodology
  - Fast Gibbs sampler for CGMs (independent of population size)

- Stay tuned for BirdCast
- Follow our progress at www.ebird.org

Outline

- Model of the plant
  - Towards a Dynamical Model of Bird Migration
- Optimal control
  - Managing wildfire in Eastern Oregon
  - Managing invasive plant species in the US Intermountain West

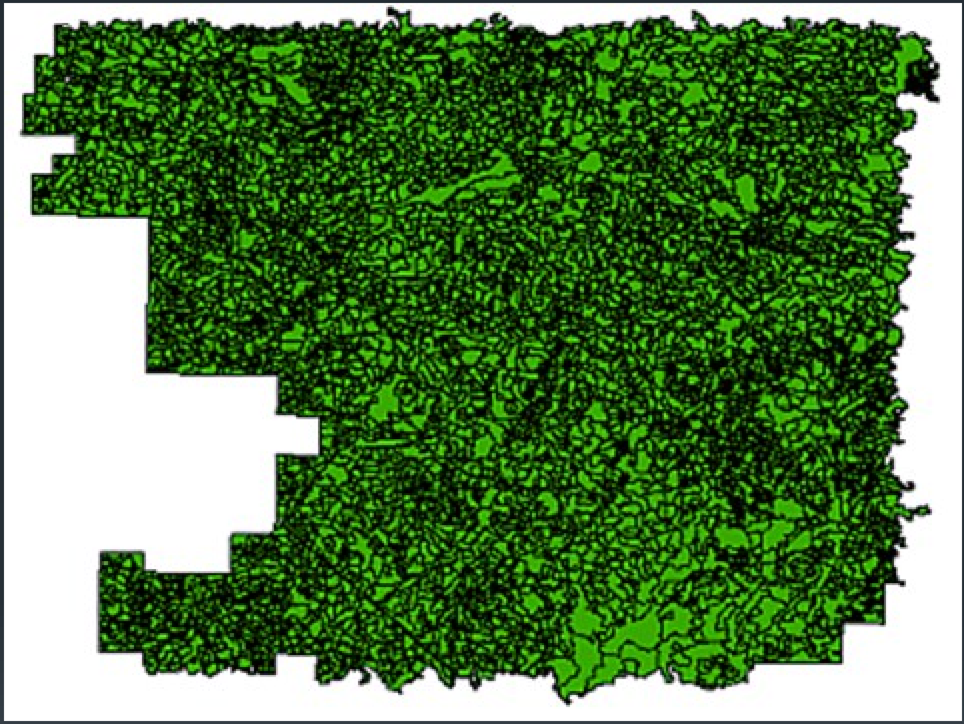
Managing Wildfire in Eastern Oregon

- 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 Question:
  - LET-BURN: When lightning ignites a fire, should we let it burn?
- ~4000 management units



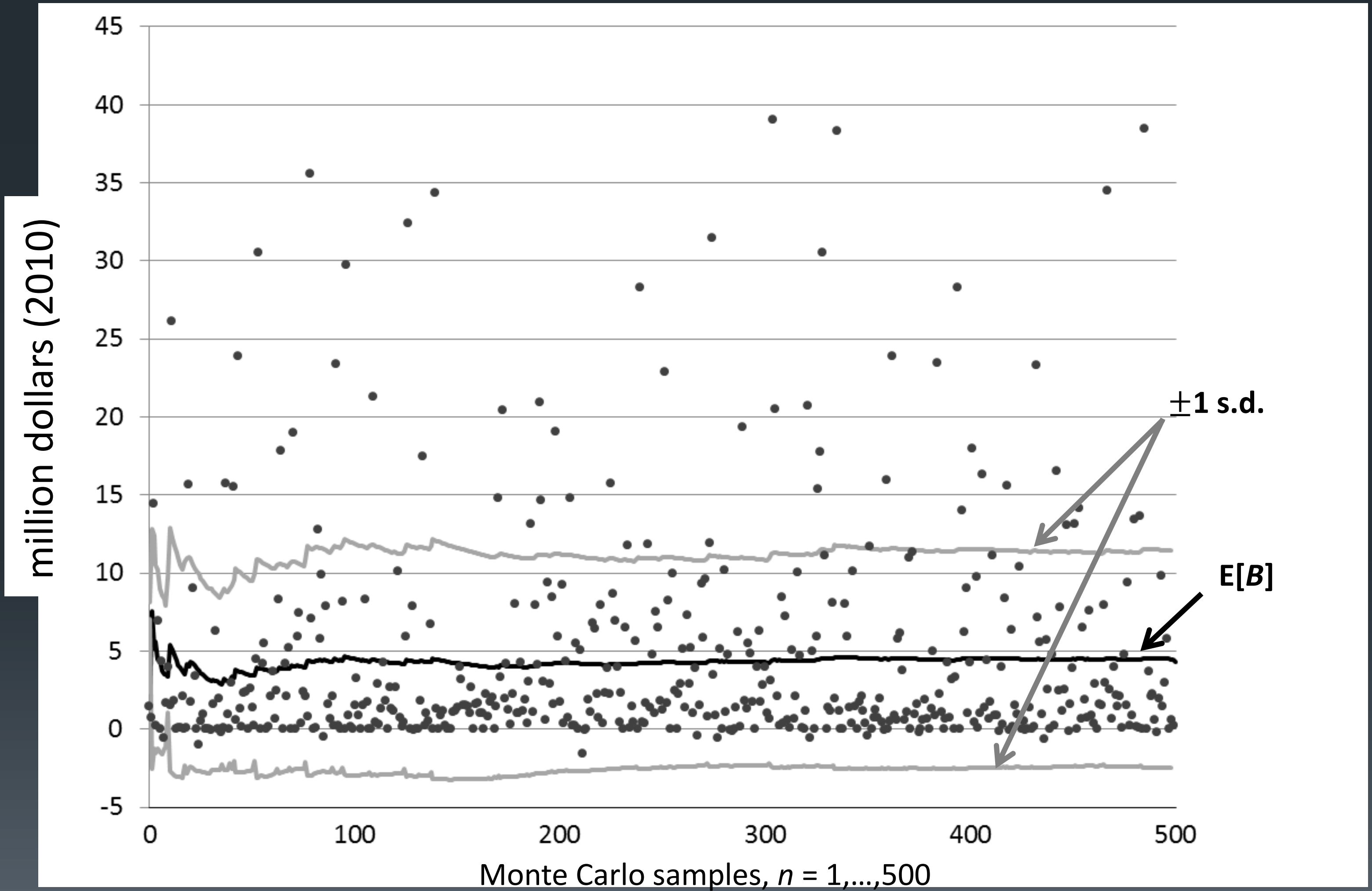
LET BURN

- Model of System Dynamics
  - Model of lightning strikes and fire ignitions
  - Model of fire spread (suppressed or not suppressed)
  - Model of fire duration (suppressed or not suppressed)
  - Model of suppression costs
  - Model of forest growth (including fuel accumulation)

- For year = 1, ..., 100
  - For fire = 1, ..., NFires
    - ignite fire
    - decide whether to SUPPRESS or LET BURN (control input)
    - burn fire
  - Grow vegetation
- Objective: Expected discounted benefit

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: Approximate Dynamic Programming algorithms that can scale to immense state space
    - 4000 management units
    - Each unit in one of 25 states
    - 25^{4000} possible states

Outline

- Model of the plant
  - Towards a Dynamical Model of Bird Migration
- Optimal control
  - Managing wildfire in Eastern Oregon
  - Managing invasive plant species in the US Intermountain West

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?

System Dynamics

- Tree-structured river network
  - Each edge has H “sites” where a tree can grow.
  - Each site can be
    - {empty, occupied by native, occupied by invasive}
- In each time period:
  - Natural death
  - Seed production
  - Seed dispersal (preferentially downstream)
  - Seed competition to become established
- Management actions
  - {do nothing, eradicate, restore, eradicate+restore}

Economic Model

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

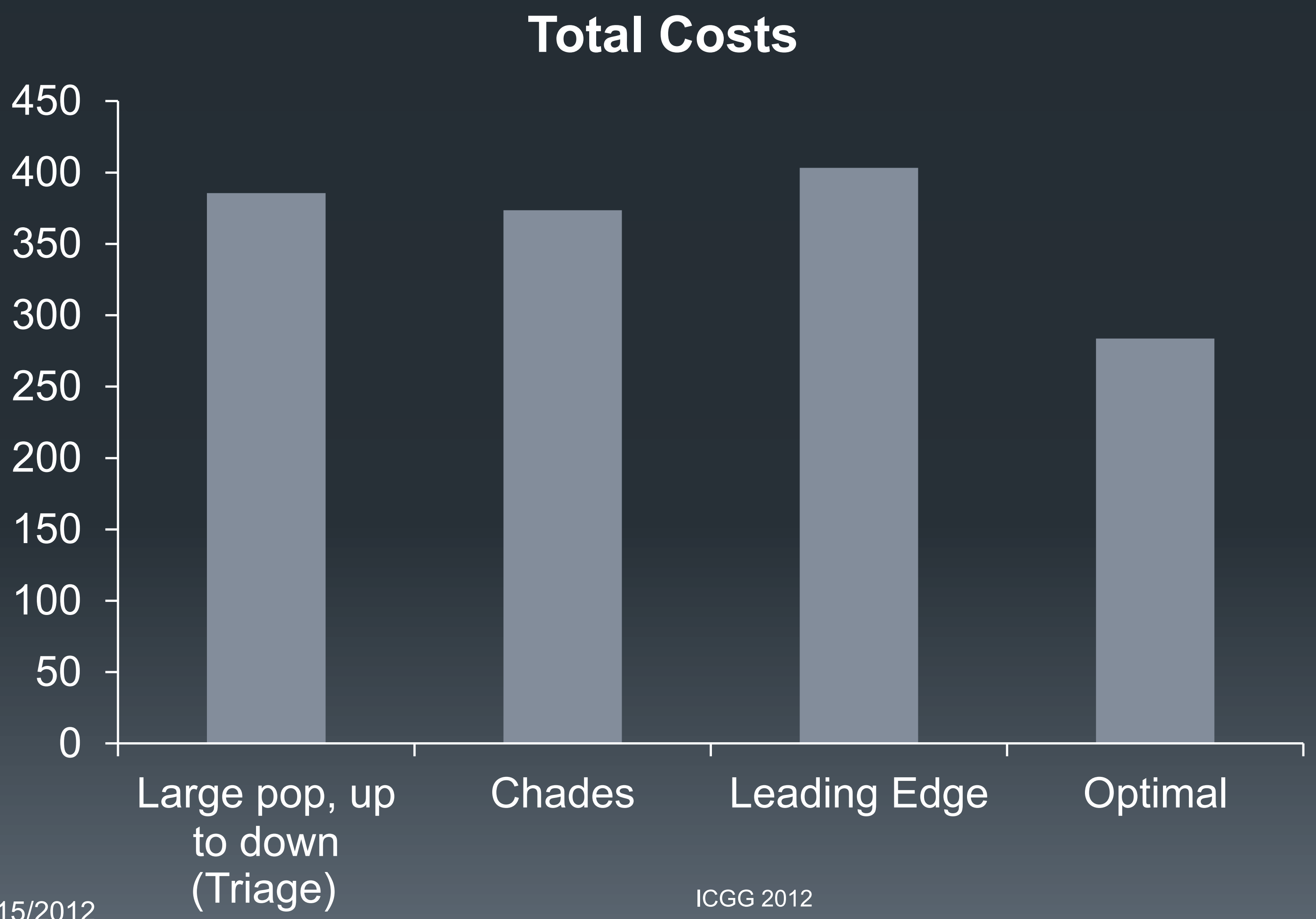
Algorithm

- Approximate computation of transition probabilities
  - Exact computation is intractable
  - Compute highly-accurate Monte Carlo estimates from simulator
  - [This is actually the rate-limiting step]
- Exact solution via value iteration
- Large memory: 14 Billion state-action combinations

Rule of Thumb Policies (from literature)

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

Cost Comparisons: Rule of Thumb Policies vs. Optimal



Summary

- Model of the plant
  - Towards a Dynamical Model of Bird Migration
- Optimal control
  - Managing wildfire in Eastern Oregon
  - Managing invasive plant species in the US Intermountain West

- MCMC algorithm for inference in Collective Hidden Markov Mode
  - eBird + radar + acoustics
- Approximate Dynamic Programming using Monte Carlo trials
- Exact Dynamic Program using Monte Carlo estimates of transition probabilities

Acknowledgements

- NSF Expeditions in Computing
  - Carla Gomes, PI (Cornell)
  - Wildfire management
  - Invasive species management

- NSF CDI BirdCast grant
  - Steve Kelling, PI (Cornell Lab of Ornithology)
  - Bird migration modeling

- NSF Bioinformatics Postdoc
  - Dan Sheldon