Machine Learning and Ecosystem Informatics: Challenges and Opportunities

Tom Dietterich

School of Electrical Engineering and Computer Science
Oregon State University
http://web.engr.oregonstate.edu/~tgd

tgd@eecs.oregonstate.edu

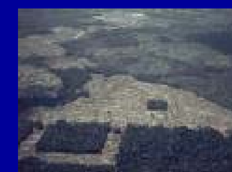
Threats to the Biosphere

Pollution including Greenhouse Gases





Habitat Loss and Fragmentation


Over-Harvesting






Lack of Scientific Knowledge

- Our understanding of ecosystem structure and function is poor
  - Extremely complex interactions
  - Operate at many temporal and spatial scales
  - Hard to do controlled experiments
  - Impossible to observe critical past events
- Long record of policy failures: “Ecological Surprises”
  - Doak et al. Ecology 39(4), 2008.
  - “Surprises are common and extreme”

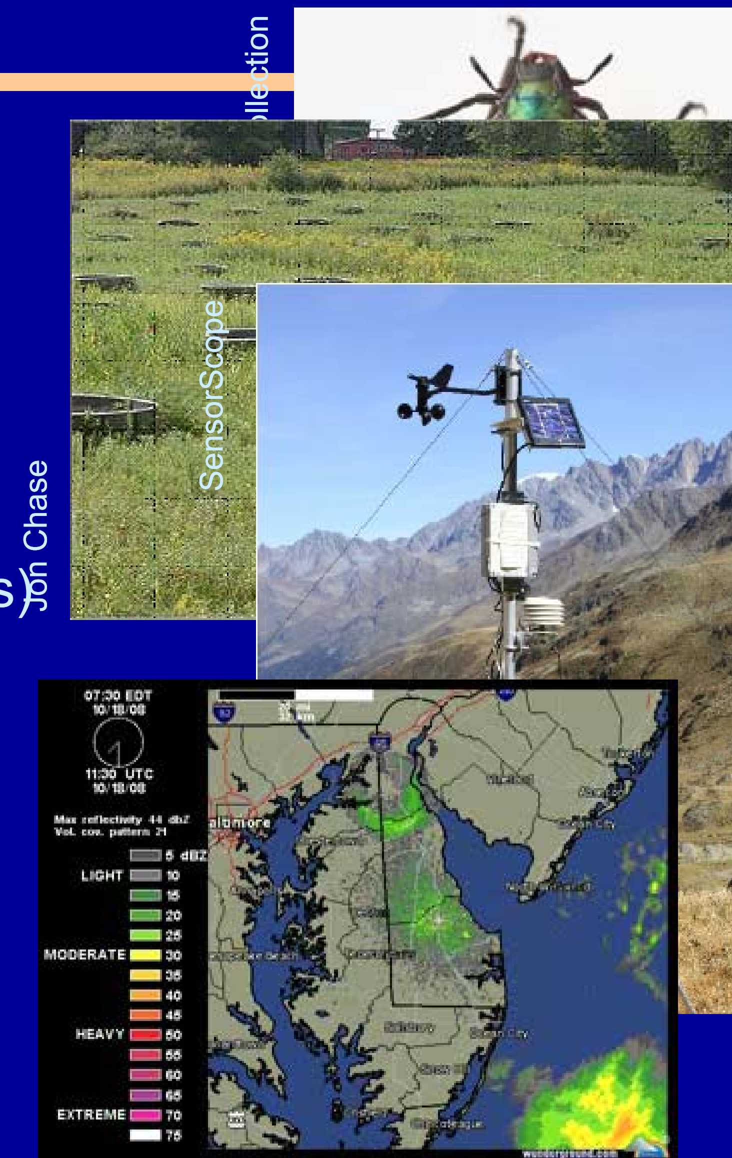
A Limiting Factor: Ecological Data

- Many ecological simulation models are based on little or no data
- Historical time series only extend back 100 years
- For almost all species, their location, population size, and interactions are unobserved

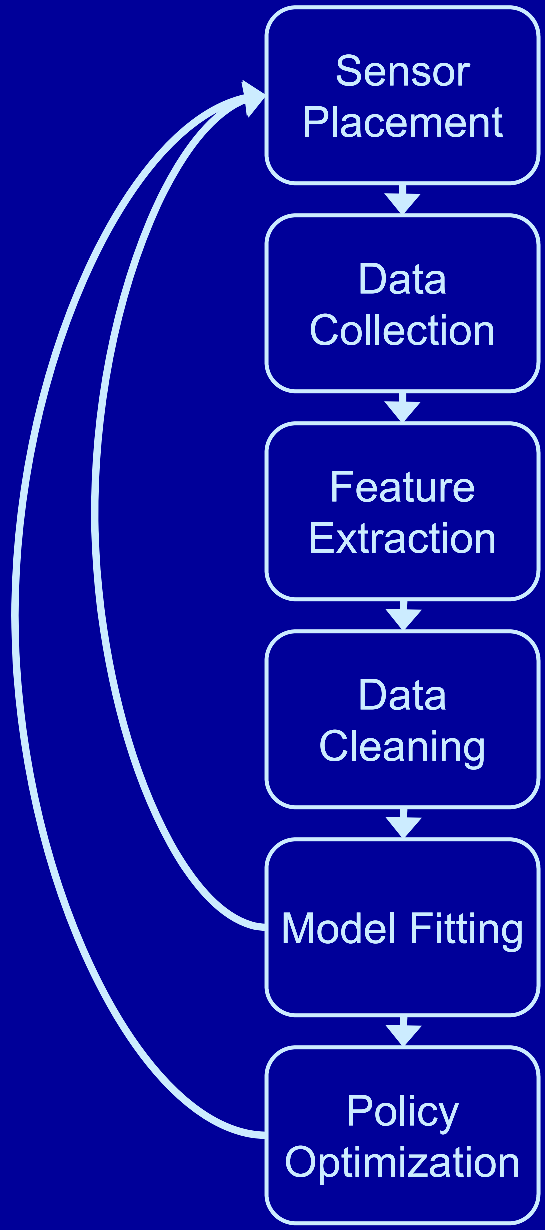
Transforming Ecosystem Science

Hypothesis Driven → Data Driven

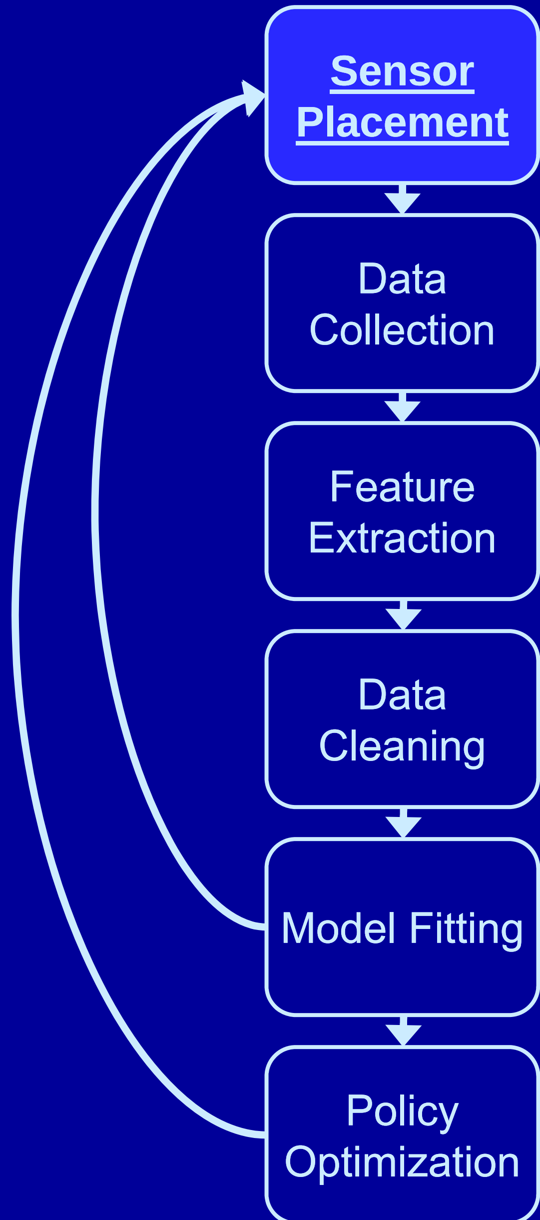
- Past approaches
  - Naturalists: museum collections
  - Artificial ecosystems (test tubes; barrels)
  - Isotope tagging of fluxes
- Emerging approaches
  - In-situ sensor networks
  - Radio/RFID tagging and tracking of organisms
  - Radar ornithology
  - Remote sensing



Data Pipeline

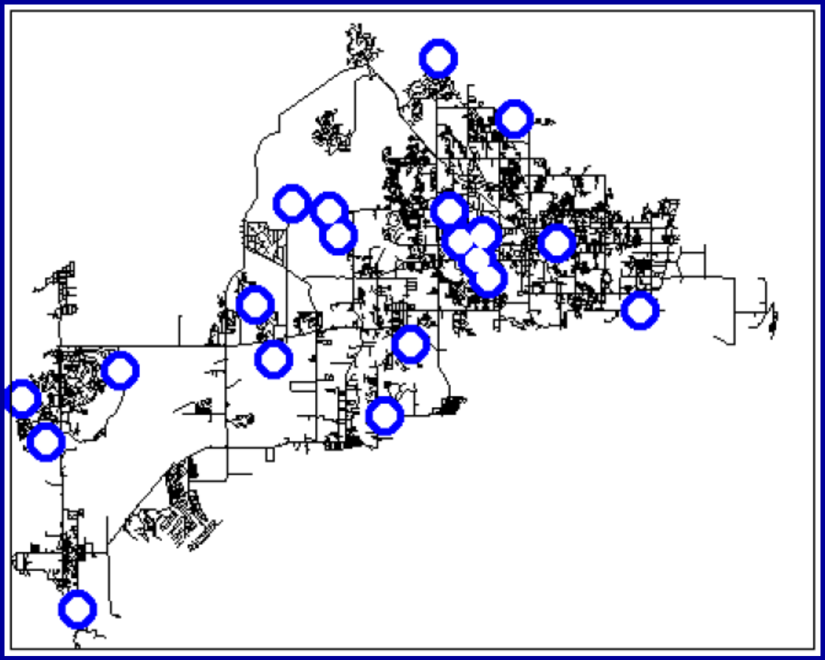


Data Pipeline



Optimal Sensor Placement

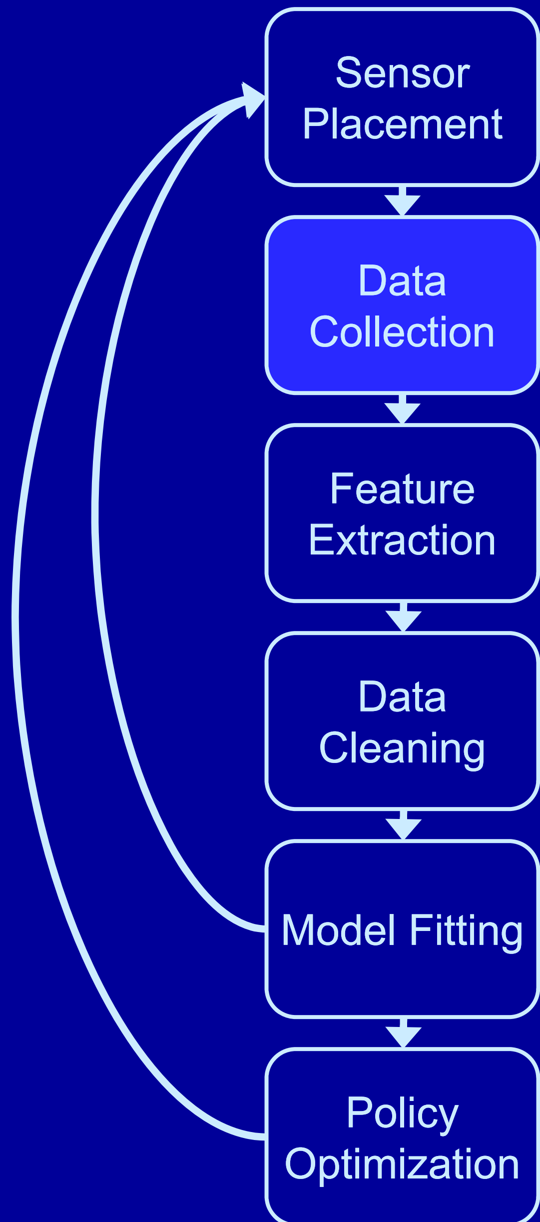
Optimal Sensor Placement for Environmental Data Collection



Leskovec et al, KDD2007

- Objectives
  - maximize detection probability
  - improve model accuracy
  - improve causal understanding
  - improve policy effectiveness

Data Pipeline

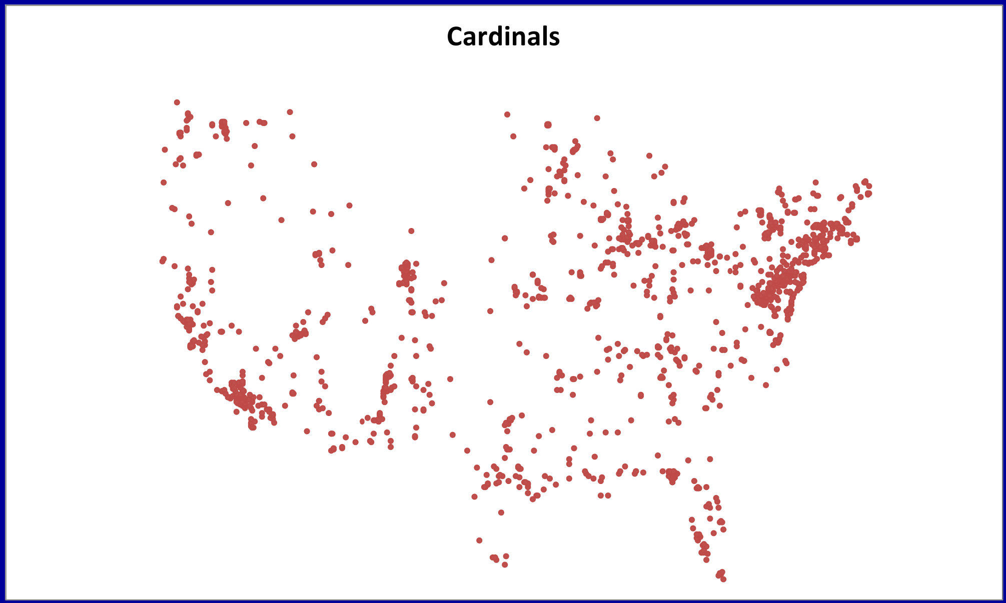


Optimal Sensor Placement

Detectability
Errors / Noise
Sampling Bias

Sampling Bias: ebird.org

- Citizen science collected by amateur bird watchers
- Strong bias toward where people live
- Explicit models of sampling bias

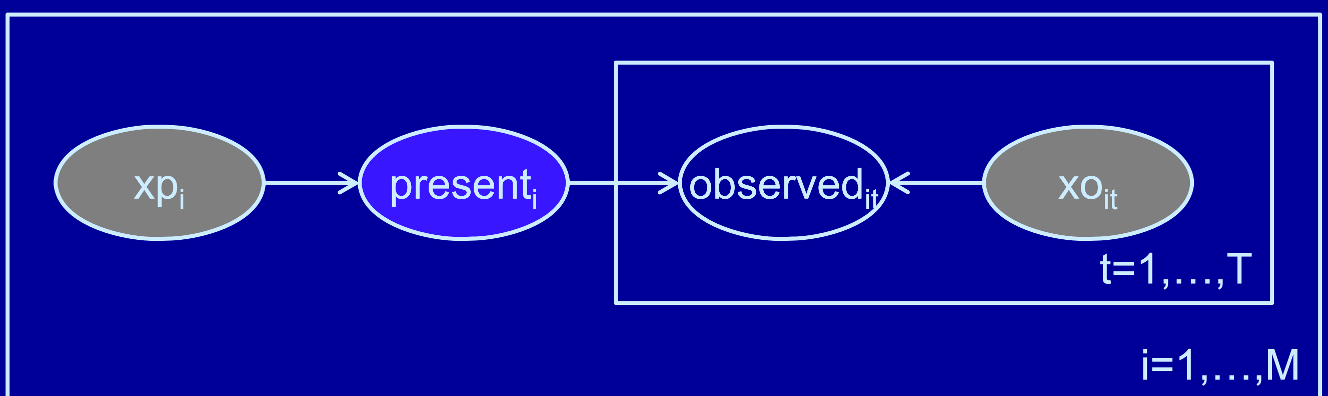


Phillips, Dudik, Elith, Graham, Lehmann, Leathwick, Ferrier: Sample Selection Bias and Presence-only Distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1), 181-197. 2009.

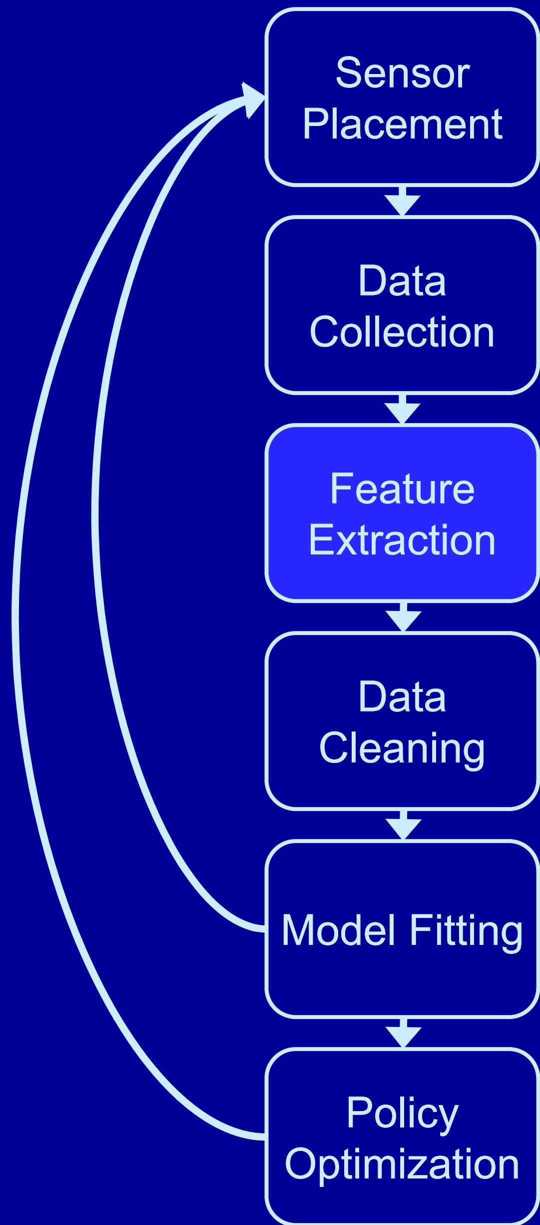
Detectability

- Birds may be present but not detected by observer
- Coupled models of detectability and presence can be fit simultaneously

  Royle, Dorazio (2008). *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities.*



Data Pipeline



Optimal Sensor Placement

Detectability
Errors / Noise
Sampling Bias

Species classification
Recognizing individuals
Tracking individuals

The BugID Project:
Rapid Throughput Arthropod Counting

- Arthropods are a powerful data source
  - Found in virtually all environments
    - streams, lakes, oceans, soils, birds, mammals
  - Easy to collect
  - Provide valuable information on ecosystem function
    - Consume the primary producers: bacteria, fungi, plants
    - Are consumed by more charismatic organisms: birds, mammals, fish
- Problem: Identification is time-consuming and requires scarce expertise
- Solution: Combine robotics, computer vision, and machine learning to automate classification and population counting




Data Pipeline

Optimal Sensor Placement

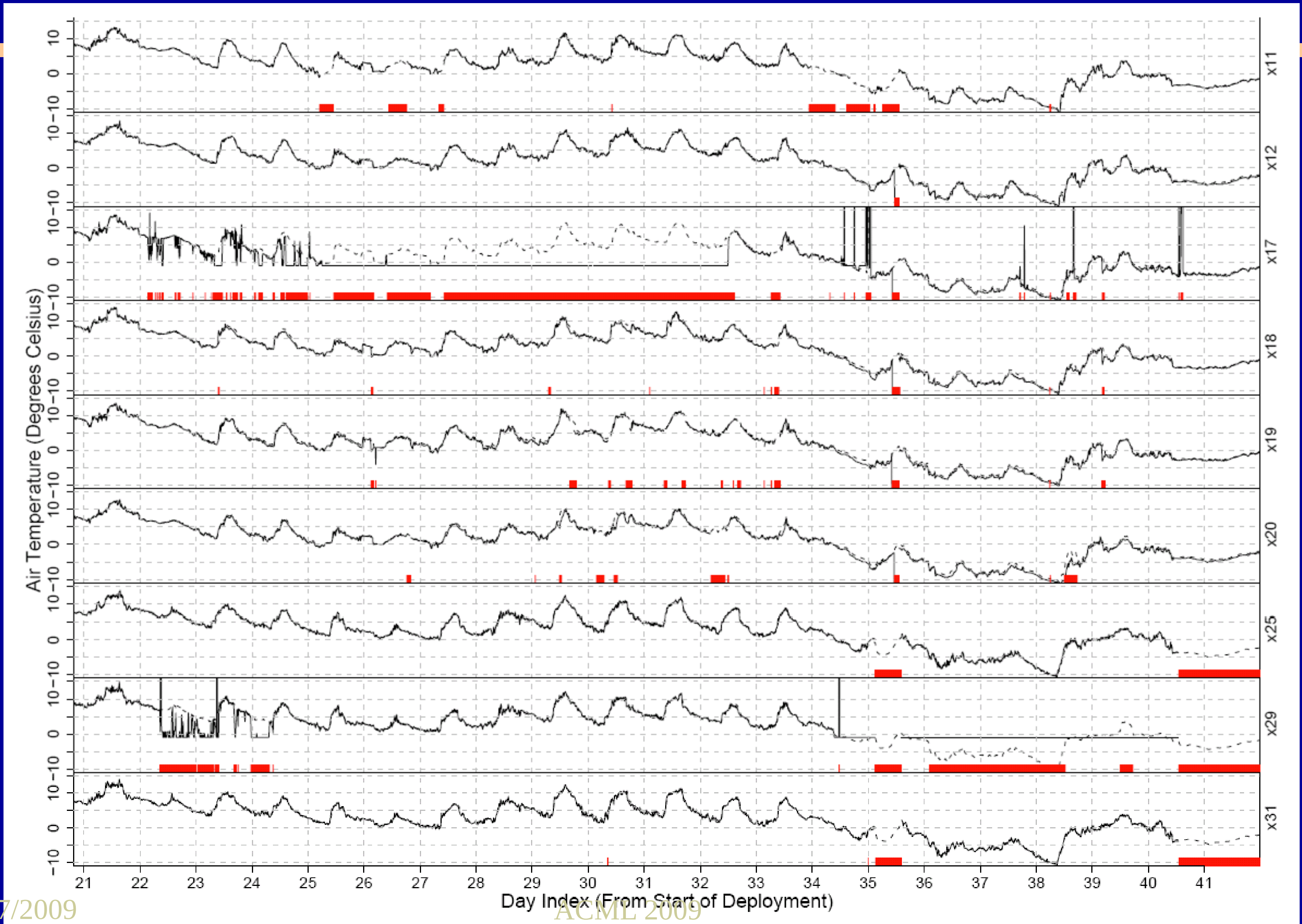
Detectability
Errors / Noise
Sampling Bias

Species classification
Recognizing individuals
Tracking individuals

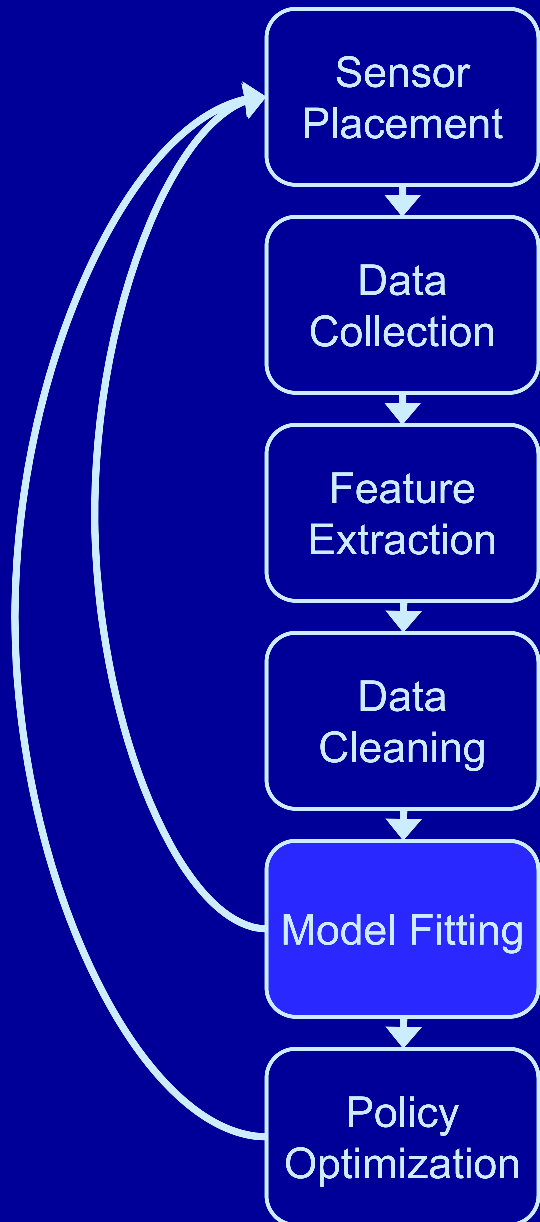
Sensor failures
Networking failures
Recognition errors

Multi-Sensor Anomaly Detection

[Dereszynski & Dietterich, submitted]



Data Pipeline



Optimal Sensor Placement

Detectability
Errors / Noise
Sampling Bias

Species classification
Recognizing individuals
Tracking individuals

Sensor failures
Networking failures
Recognition errors

Species distribution models
Behavioral models
Dynamical systems models

Coupling Multiple Problems

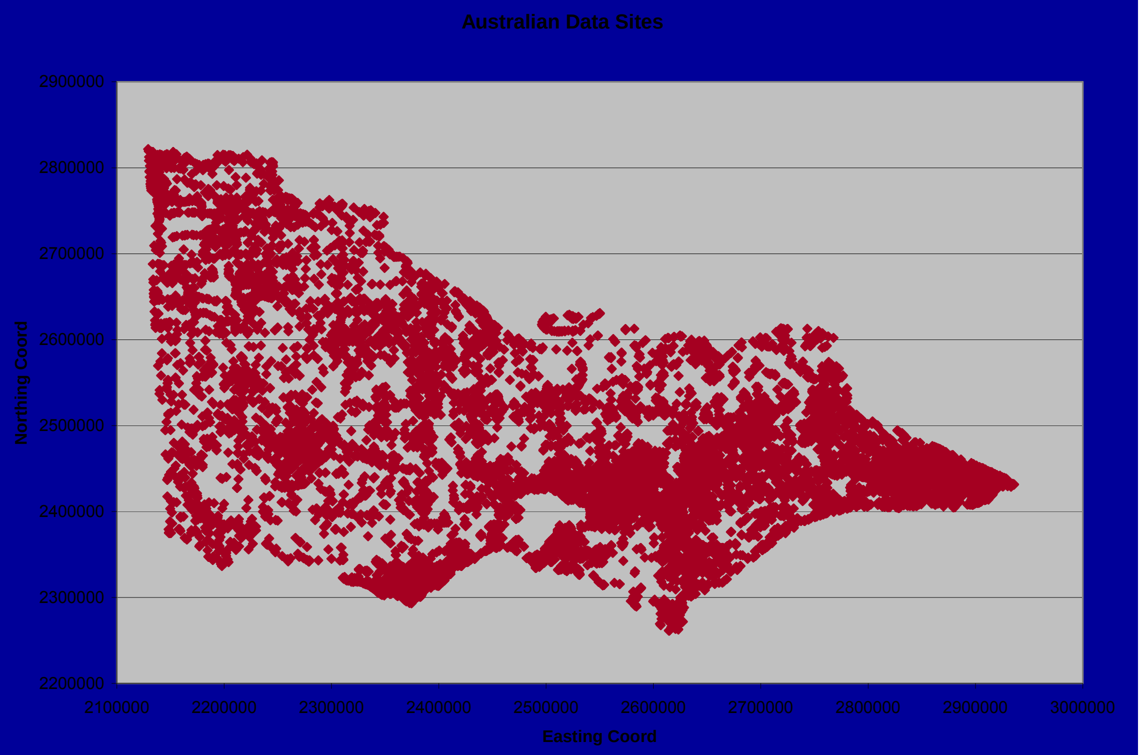
Species Distribution Models

- What are the environmental/biological requirements for a species?
- Given:
  - Environmental features (elevation, soil properties, weather) of a site
  - Presence, presence/absence, or abundance of K species
- Find:
  - Probability that each of the K species will be found at new sites
  - Extrapolation to global climate change scenarios

Plants in Victoria

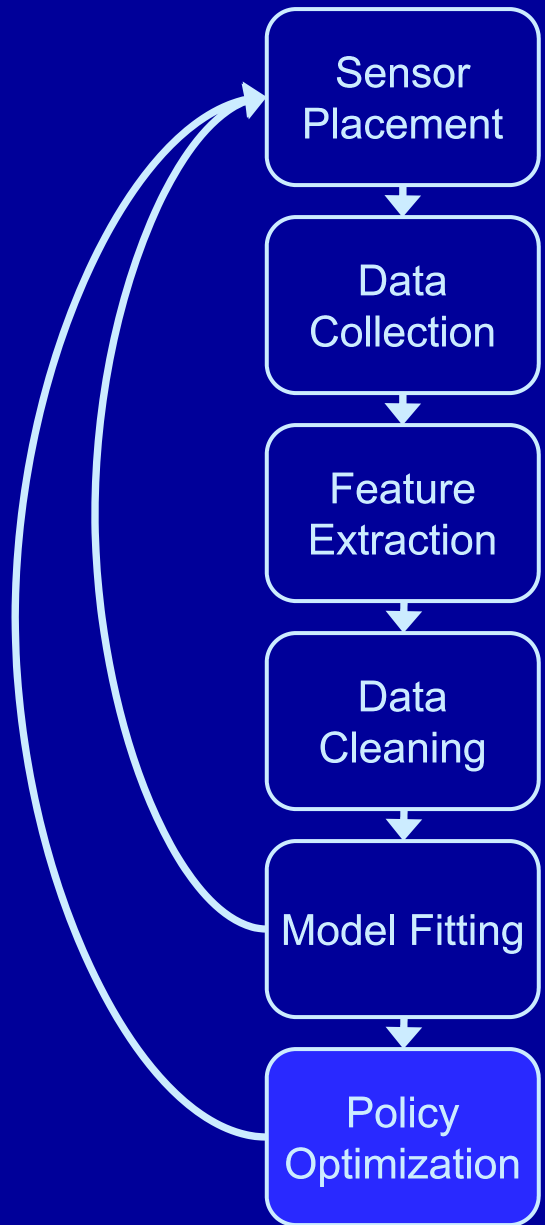
[Arwen Lettkeman]

- 5,605 plant species measured at >113,000 sites
- 83 environmental features



Source: Matt White, Arthur Rylah Institute

Data Pipeline



Optimal Sensor Placement

Detectability
Errors / Noise
Sampling Bias

Species classification
Recognizing individuals
Tracking individuals

Sensor failures
Networking failures
Recognition errors

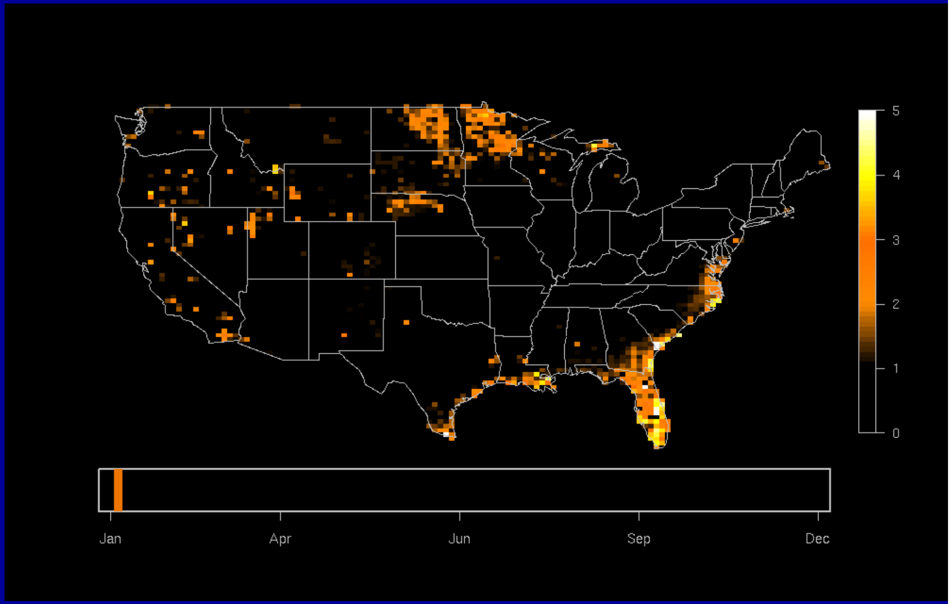
Species distribution models
Behavioral models
Dynamical systems models

Large spatio-temporal MDPs
Optima that are robust to model uncertainty

Coupling Multiple Problems

Robust Reserve Design

- Given:
  - Species distribution model
  - Budget
- Find:
  - Set of reserves to purchase that are good habitat for the species and fit within the budget
- Robust to uncertainties in the model (and climate, etc.)
  - Optimize the machine learning to be more accurate where land is cheaper to acquire?
  - Joint optimization of model fitting and optimization?



Predicted winter distribution of tree swallows (Fink, et al., unpublished)

Outline: Three Challenges for Machine Learning

- Object Recognition for Arthropod Counting
- Multiple Species Prediction
- Spatio-Temporal Optimization for Forest Management

Automated Rapid-Throughput Arthropod Population Counting

- **Goal:**
  - technician collects specimens in the field by various means
  - robotic device automatically manipulates, photographs, classifies, and sorts the specimens

- **Two applications:**
  - stoneflies in freshwater streams
  - soil mesofauna

Application 1: Stonefly populations in freshwater streams

- differentially sensitive to many pollutants
- live in rivers; reliable indicator of stream health
- difficult and expensive for people to classify (particularly to genus or species levels)
- hundreds of species

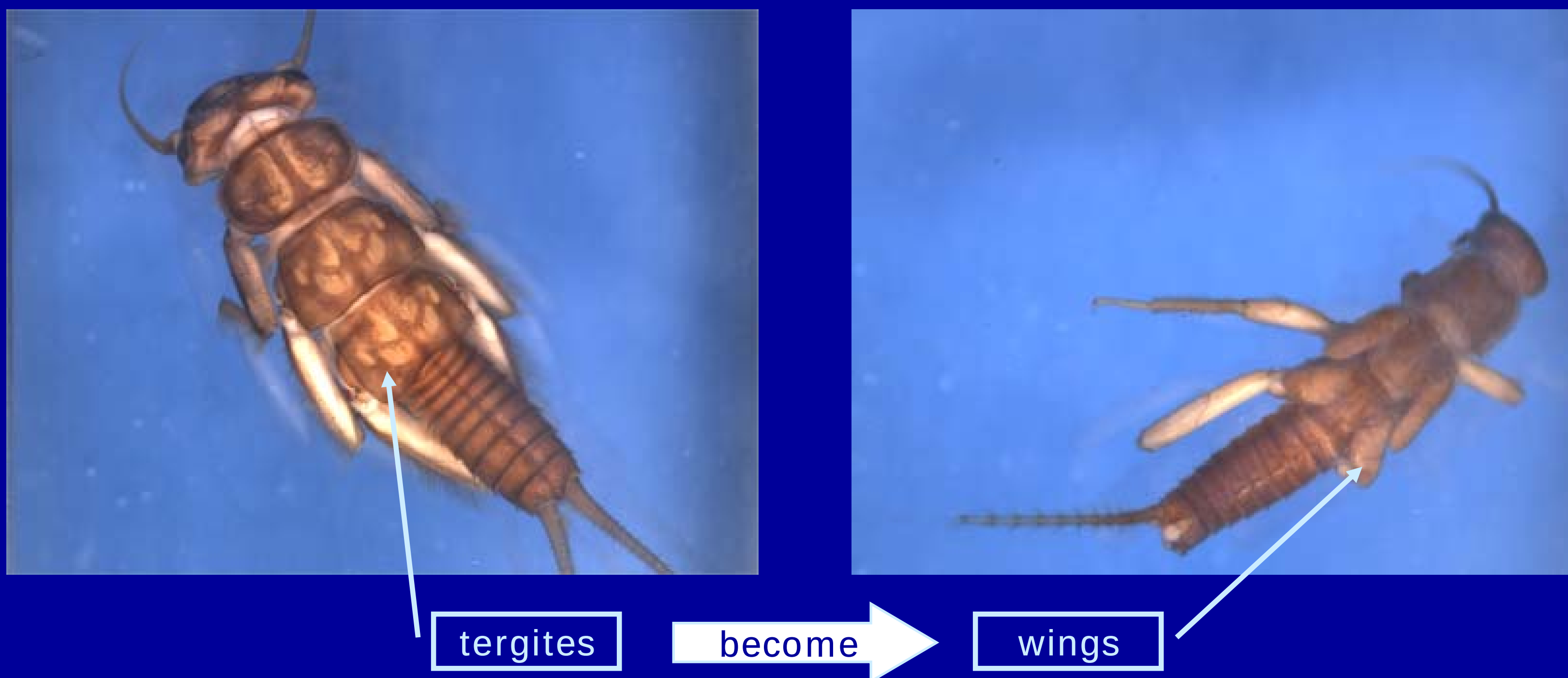
Application 2: Small arthropods in soil: “soil mesofauna”

Computer Vision Challenges(1)

- Highly-articulated objects with deformation

Computer Vision Challenges(2)

- Huge intra-class changes of appearances due to development and maturation



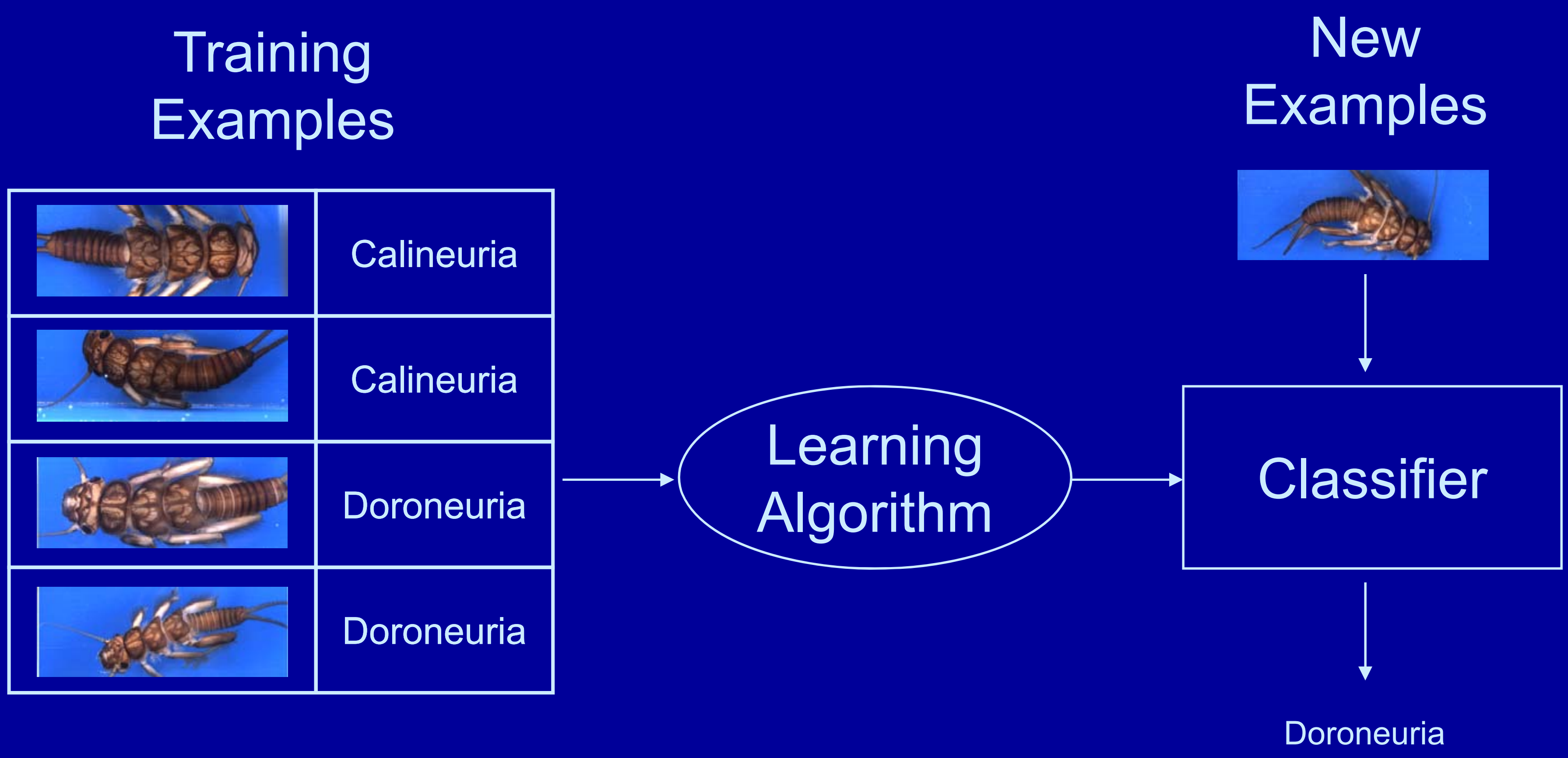
Computer Vision Challenges(3)

- Small between-class differences

Calinueria

Doronueria

Machine Learning

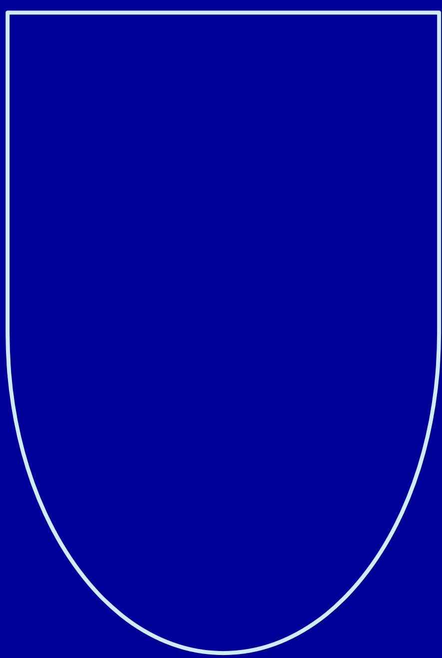


State of the Art in Object Recognition

- “Bag of Keypoints” based on visual dictionaries
  - 85% correct

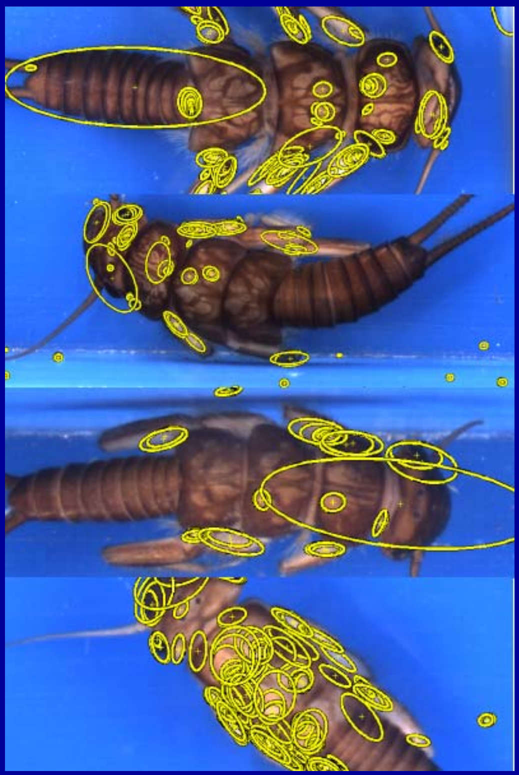
- New method: multiple-instance classification of bags of regions
  - 95% correct

Region-Based Approaches: Convert Image to Bag of Patches

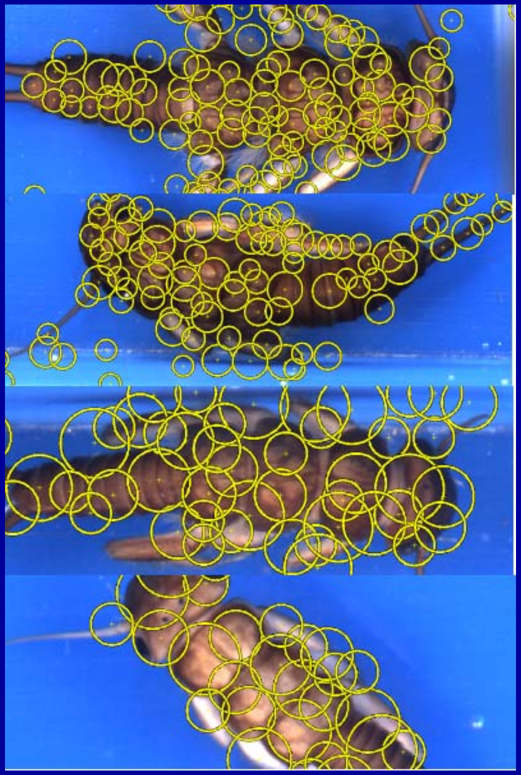


- Handles
  - Occlusion
  - Rotation, translation
  - Scale (with scale-independent patch representation)
  - Partial out-of-plane orientation
  - Articulation / Pose
- Problem:
  - How to define the patches?
  - How to represent each patch?
  - How to classify a BAG of patches?

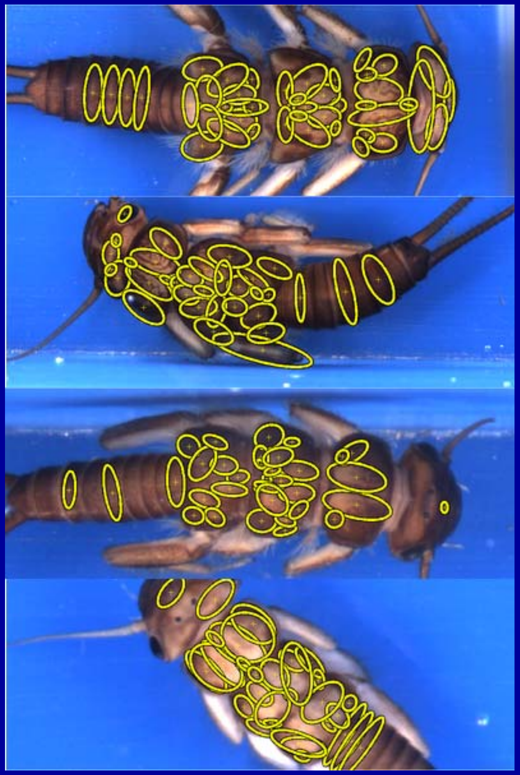
Defining the Patches: Interest Region Detectors



Hessian-Affine Detector

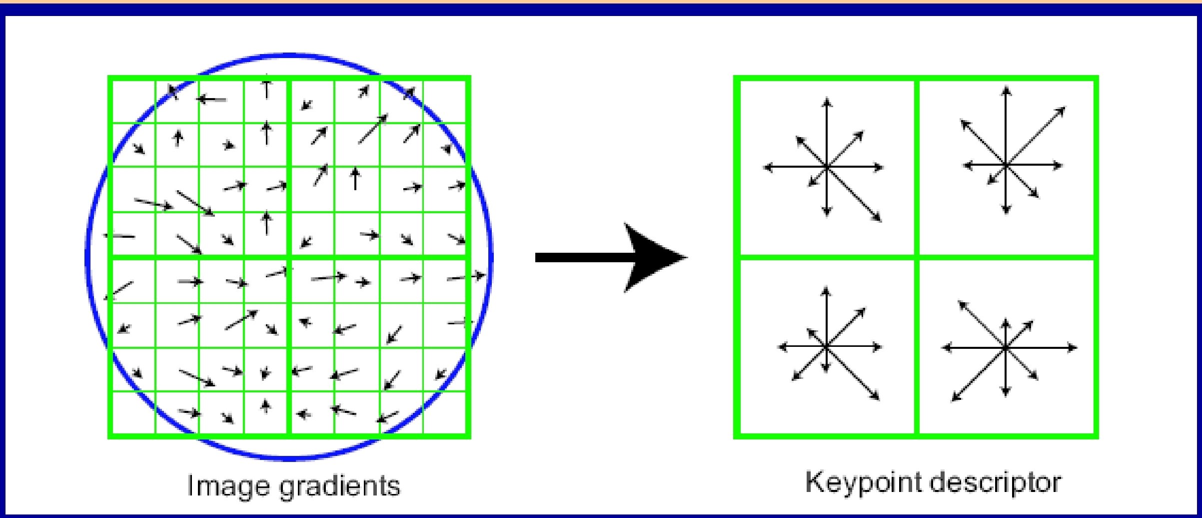


Kadir Entropy Detector



PCBR Detector

Representing the Patches: SIFT (Lowe, 1999)

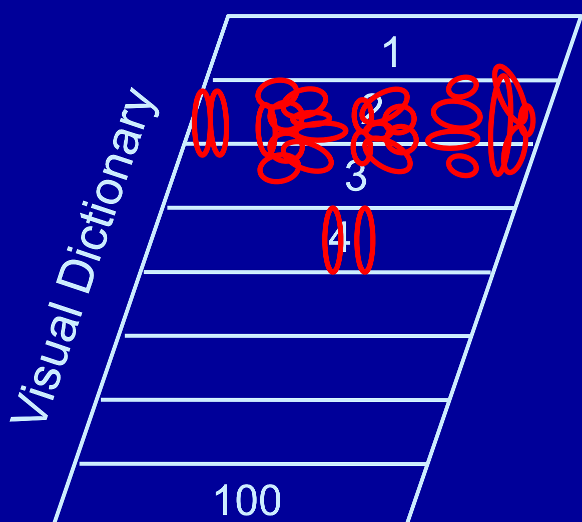


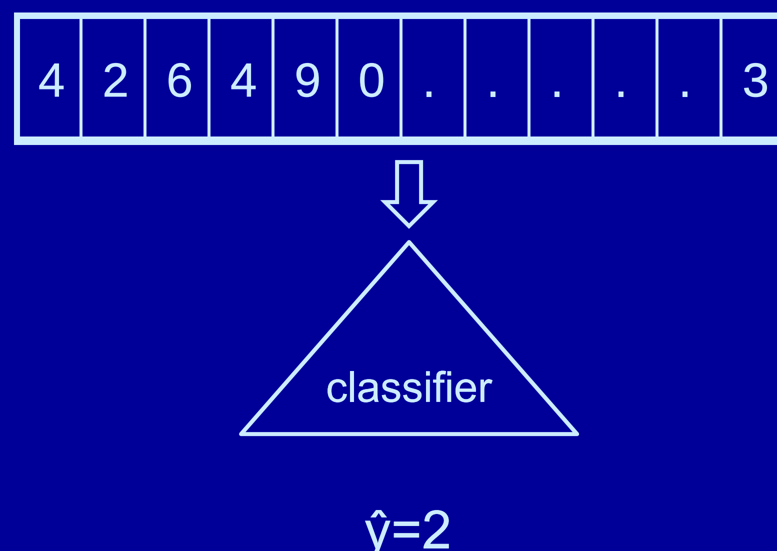
(Lowe, 1999)

- Morph ellipse into a circle
- Compute intensity gradient at each pixel in 16x16 region
- Rotate whole circle according to dominant intensity gradient
- Weight gradients by a Gaussian distribution (indicated by circle)
- Collect into histograms within each 4x4 region (gives 16 histograms)
- Result: 128-element vector normalized to have Euclidean norm 1

Classify Bag of Patches
Method 1: Visual Dictionaries

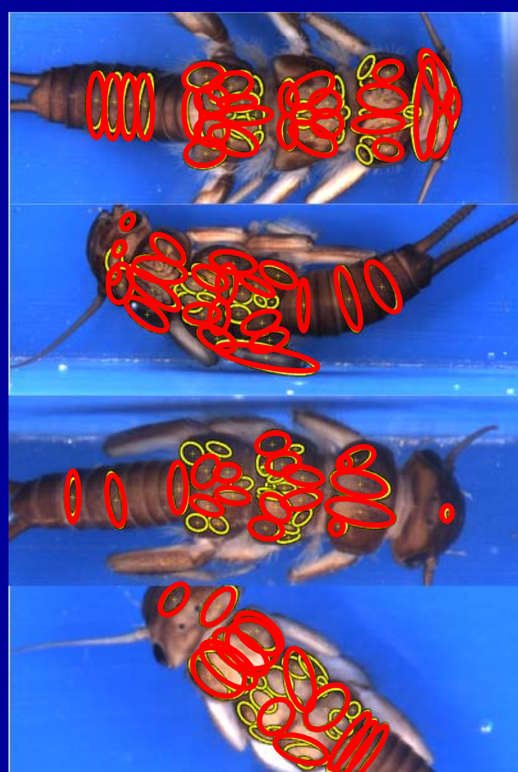
- "look up" each patch in dictionary and count into a feature vector
- feature vector is then given to the classifier

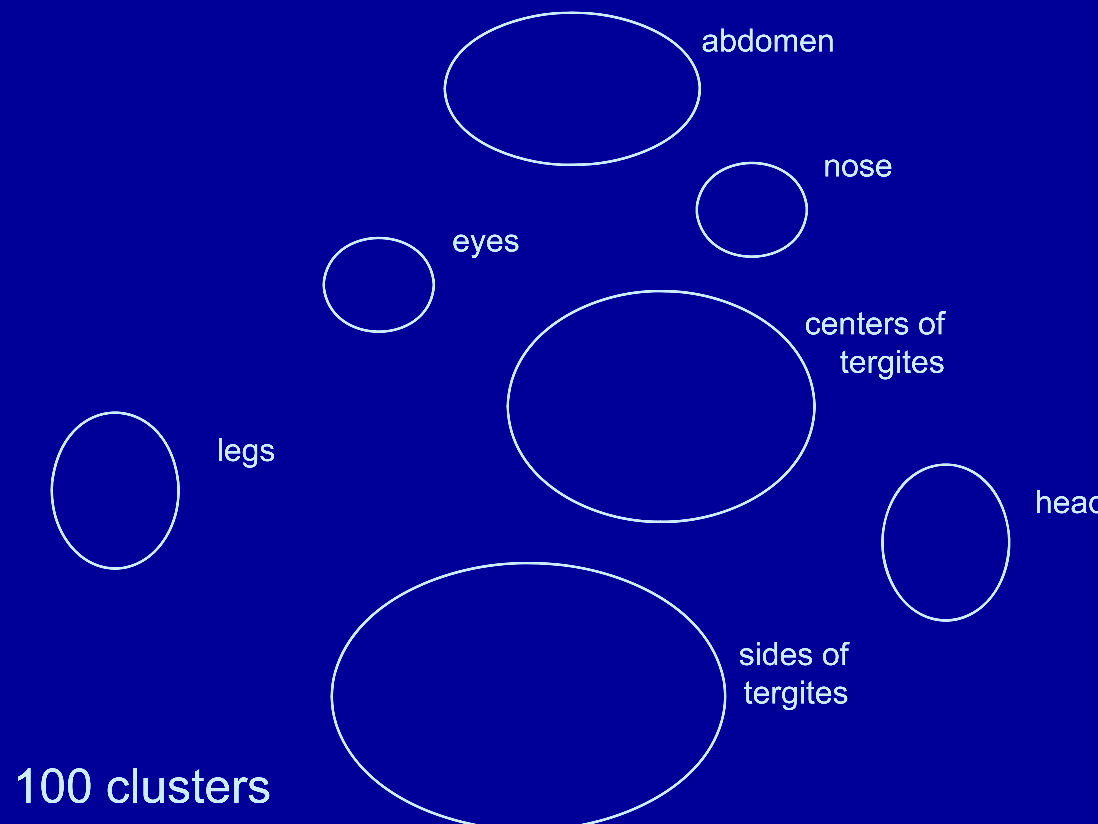




Learn visual dictionary via clustering

- Gaussian Mixture Model (k=100) with diagonal covariance matrices (EM, initialized with K-means)

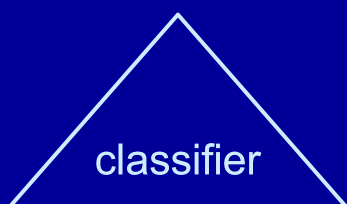




Issues with Visual Dictionaries

- Unsupervised
  - Several efforts to construct discriminative dictionaries (Moosman et al., 2006)
- Lose information
  - 128-element SIFT contains 1024 bits, a bag of 256 SIFTs contains 256K bits
  - Keyword histogram from 2700-element dictionary contains ~2700bits

Classify Bag of Patches
Method 2: Multiple-Instance Classifier



$\hat{y}=2$

2	8	1	3	0	0	6	4	2

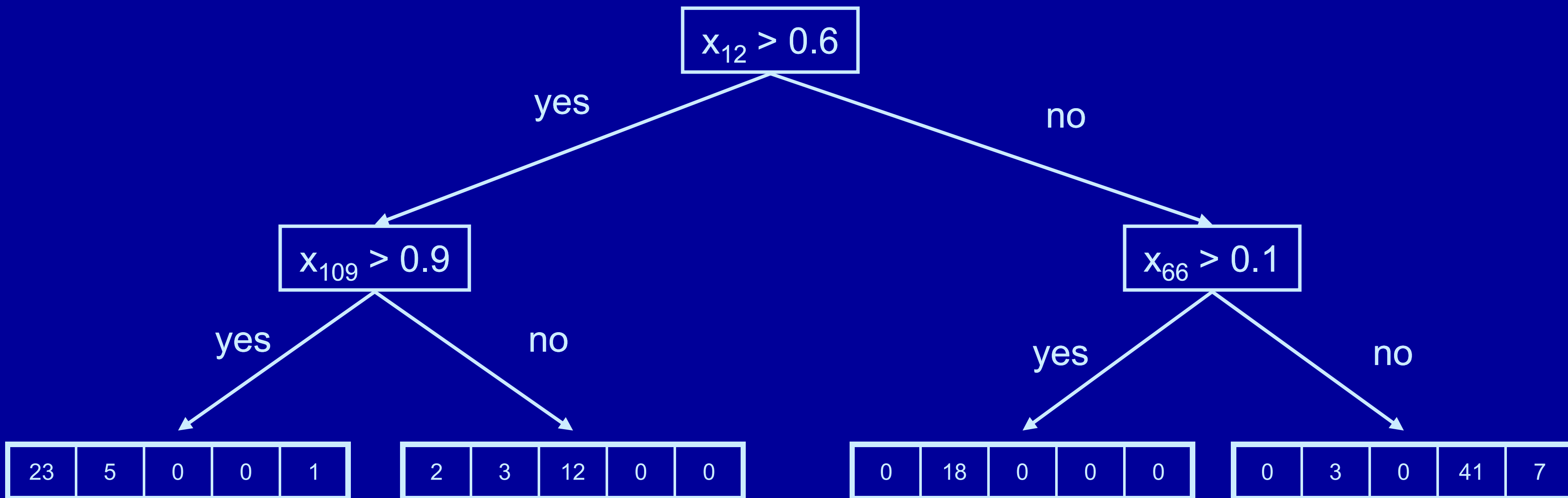
votes

Final prediction: $\hat{y}=2$

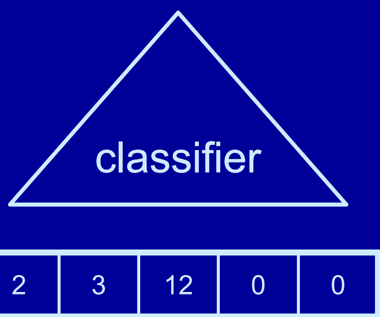
- The classifier predicts the class of the image separately from each patch
- These vote to make the final decision

Improved Multiple-Instance Classification

- Evidence Trees: Like decision trees, but store the "evidence" in each leaf
- Given an input, output the evidence



Classify Bag of Patches
Voted Evidence Trees



- The classifier predicts the class of the image separately from each patch
- These vote to make the final decision

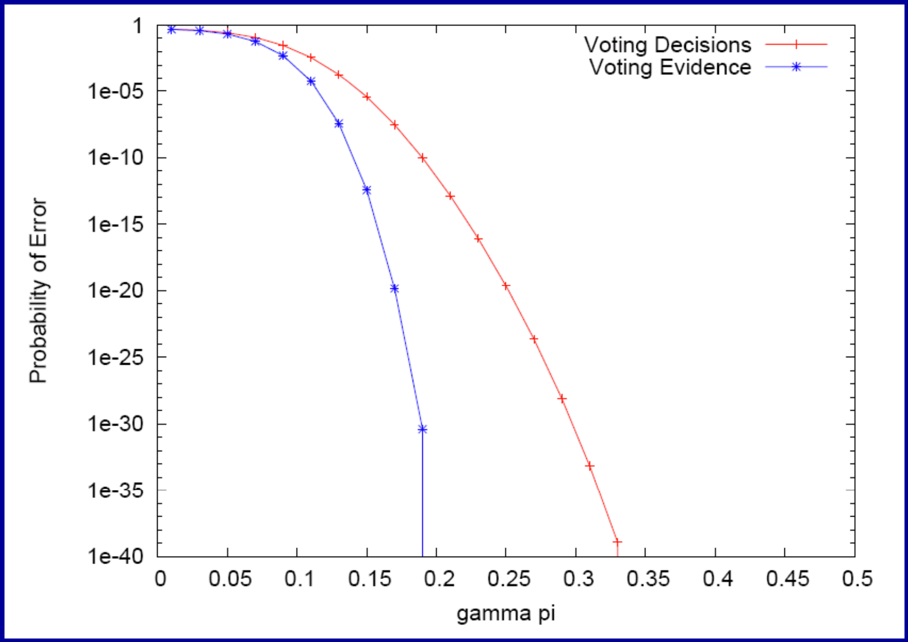
87	14	34	6	61

votes

Final prediction: $\hat{y}=1$

Theorem: Voting Evidence is Better than Voting Decisions

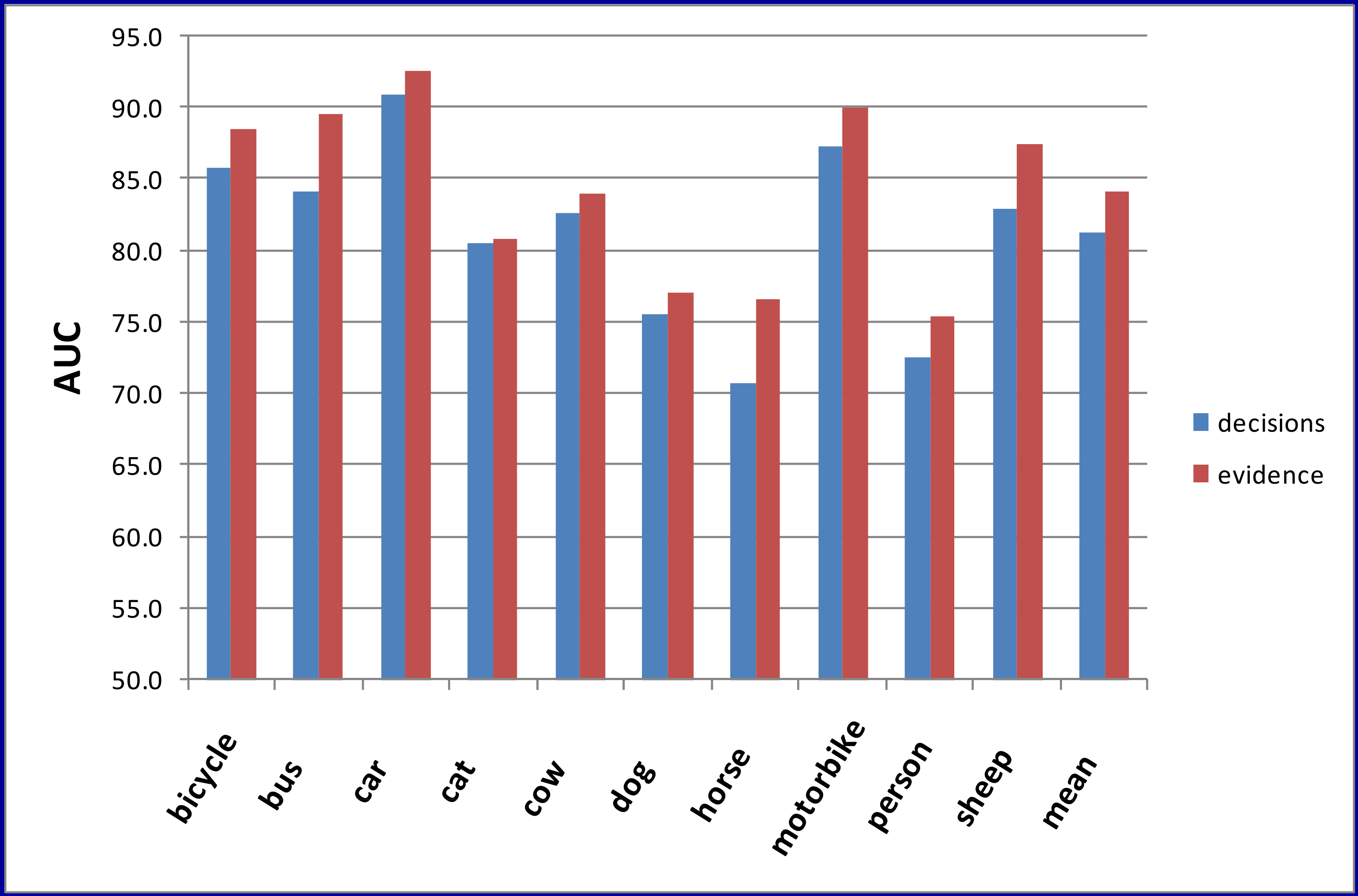
- Intuition: When voting decisions, there are two opportunities to make a mistake:
  1. Making the wrong decision at each leaf
  2. Making the wrong decision when combining the votes
- With evidence trees, the first opportunity is avoided



γ = margin of decision tree nodes
π = fraction of non-noise patches

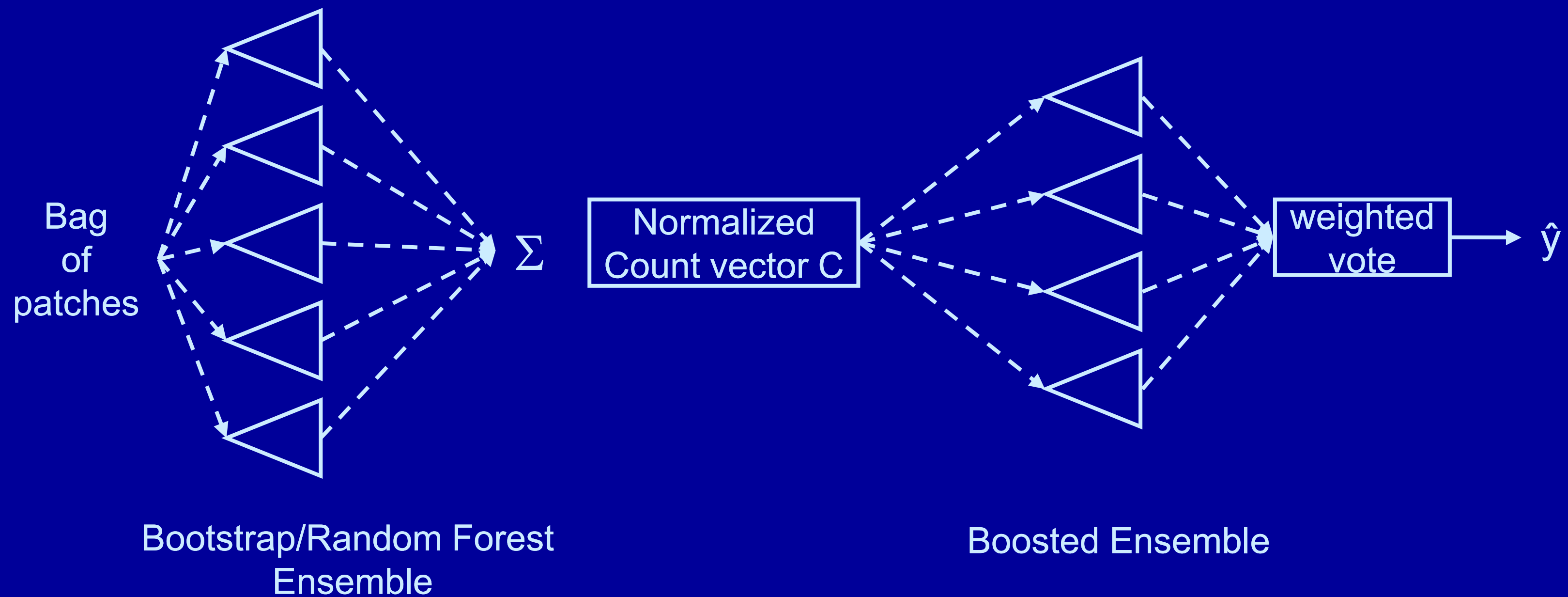
Voting Decisions vs. Voting Evidence

PASCAL 2006 VOC



Final Classifier:
Stacked Random Forests

1. Each patch is processed by a ***random forest*** of evidence trees
2. Evidence is summed and normalized to produce C
3. C is classified by a second-level ***boosted decision tree ensemble***



Experimental Study
9 Taxa of Stoneflies

Cal

Dor

Hes

Iso

Mos

Pte

Swe

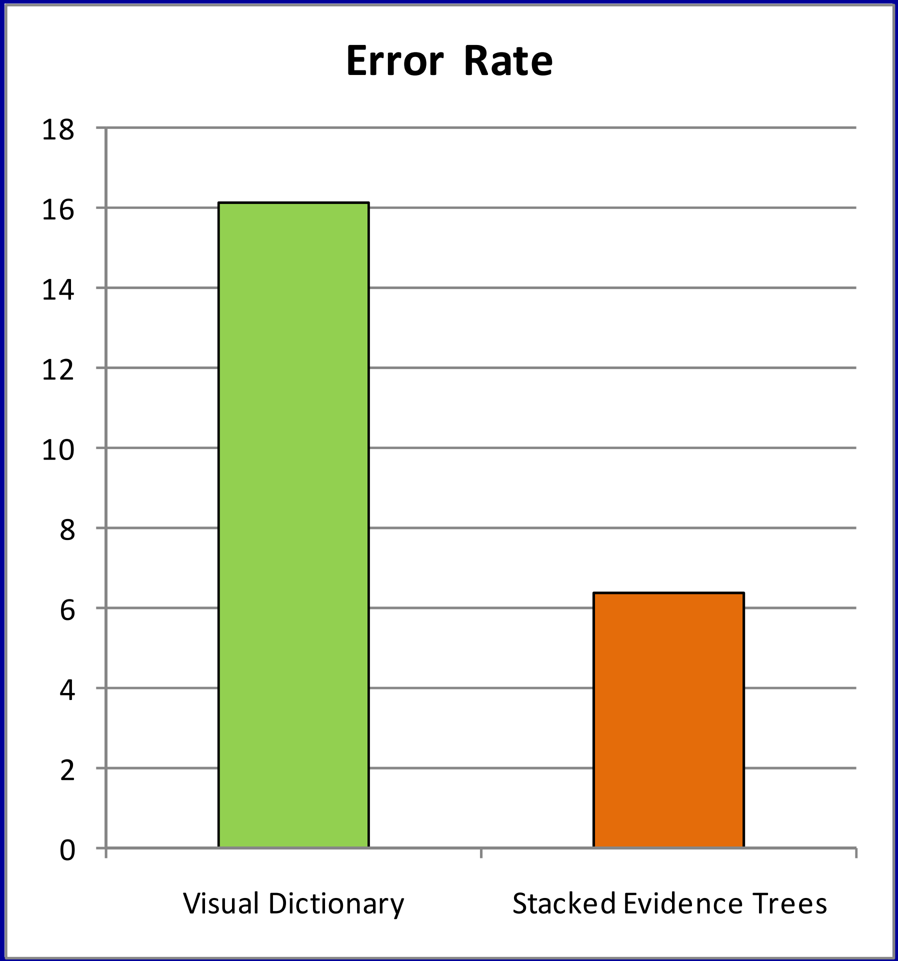
Yor

Zap

STONEFLY9 Dataset

- 3826 images
- 773 specimens
- 9 classes
- Error estimation by 3-fold cross-validation
  - all images of a specimen belong to the same fold

Comparison of Methods



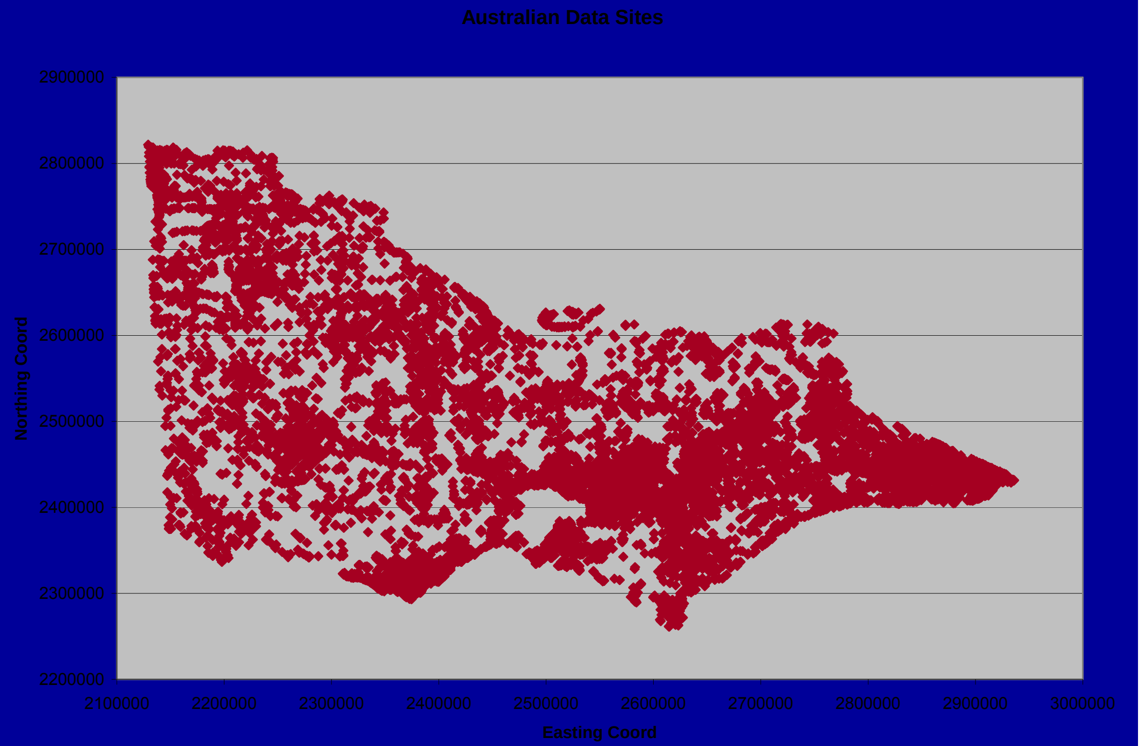
Outline: Three Challenges

- Object Recognition for Arthropod Counting
- Multiple Species Prediction
- Spatio-Temporal Optimization for Forest Management

Plants in Victoria

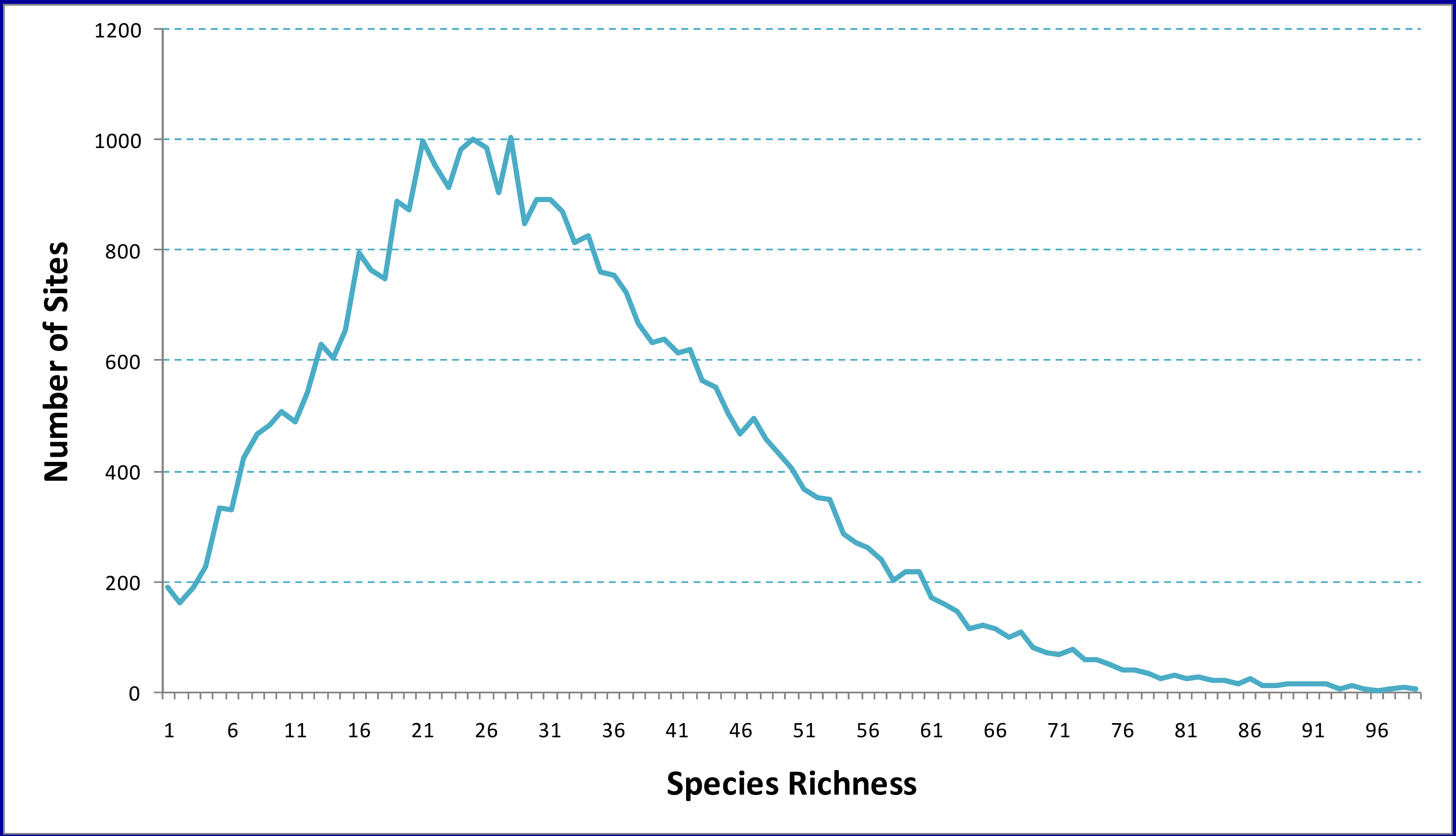
[Arwen Lettkeman]

- 5,605 plant species measured at >113,000 sites
- 83 environmental features

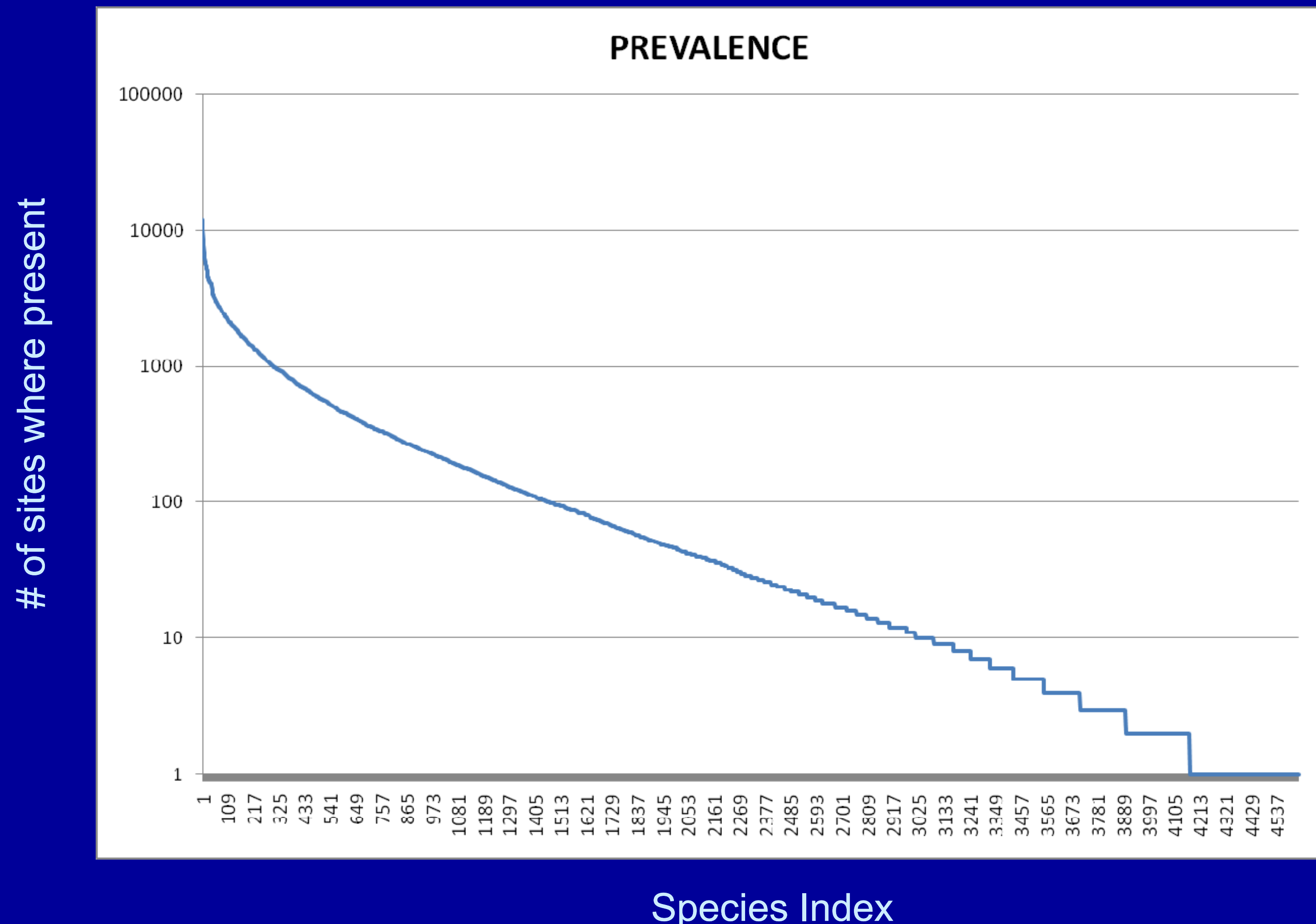


Source: Matt White, Arthur Rylah Institute

Labels are Sparse

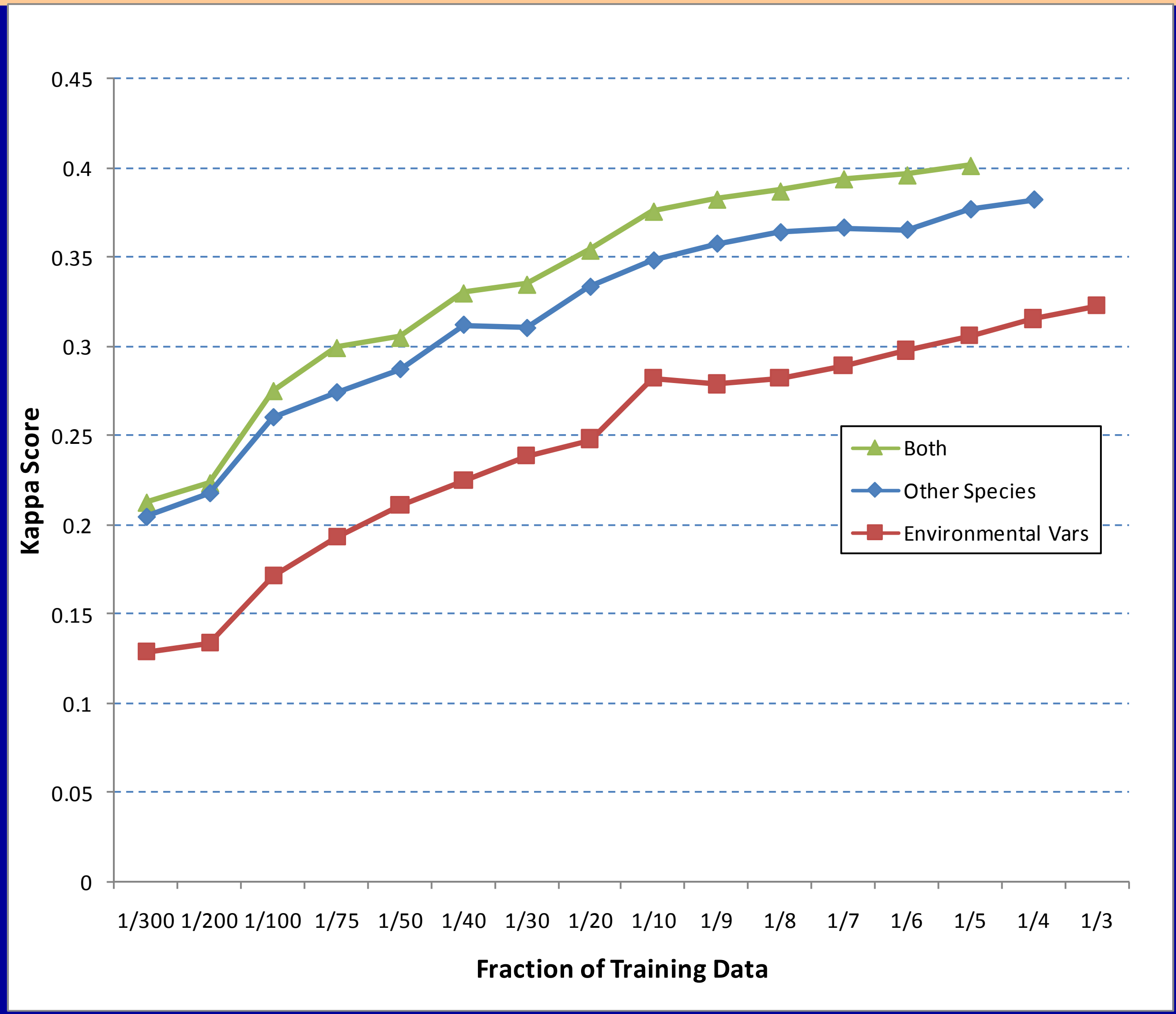


Many Species are Rare
Many Species are Common



Exploiting Multiple Species

- Experiment: Predict presence/absence of one species given
  - Environmental attributes
  - Presence/Absence of other species
  - Both



Multi-Label Classification

- Multiple-output neural networks
- Multiple-response decision trees
  - Zhang, H. 1998. Classification Trees for Multiple Binary Responses, *JASA*.
  - De'ath, G. 2002. Multivariate Regression Trees: A new technique for modeling species-environment relationships. *Ecology.*
- Conditional random fields
  - McCallum, A., Ghamrawi, N. 2204. Collective multi-label text classification. Tech Report.
- Conditional topic models
  - Mimno, D., McCallum, A. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. *UAI*.
- Stacking
  - Wolpert, D. 1992. Stacked generalization. *Neural Networks*
- Reduction to multi-class classification problems
  - Read, J., Pfaringer, B., Holmes, G. 2008. Multi-label classification using ensembles of pruned sets. *ICDM.*
  - Tsoumakis, G., Vlahavas, I. 2007. Random k-label sets: an ensemble method for multilabel classification. *ECML.*

Multi-Task Learning

- Train one model for each species, but use the other species as auxiliary tasks
- Train a joint model, but choose separate regularization constant and decision threshold for each species

Status

- No results yet!

Outline: Three Challenges

- Object Recognition for Arthropod Counting
- Multiple Species Prediction
- Spatio-Temporal Optimization for Ecosystem Management

Fires in the Western US

- Natural behavior – frequent low-intensity fires (every 15-20 years)
  - Favors Ponderosa Pine forests
    - thick bark to survive low-intensity fire
  - Takes out weaker trees – “natural thinning”
  - Result: Open stands of big, valuable trees

Fire Suppression Policy

www.mtmultipleuse.org/images/smokey.jpg

- William Greeley USFS chief 1920-9:

  "the conviction was burned into me that that fire prevention is the number 1 job of American foresters

  (Greeley, WB. 1951. *Forests and men*. NY: Doubleday.)
- 10:00 am policy: Contain every wildfire by 10:00 am the day after it is reported regardless of cost.

Result of this Policy

- Lodgepole pine becomes dominant tree
  - low economic value
  - vulnerable to pine bark beetle
  - dies and creates enormous fuel buildups
- Fires become catastrophic
  - most vegetation killed
  - most soil organic matter destroyed
  - very long recovery time
  - big CO_2 release

Adaptive Fire Treatment

- Choose what fires to allow to burn
- Perform “mechanical thinning” to reduce fuel loads

Formulation as a Markov Decision Process

- States:
  - Landscape divided into 100 management units (MUs)
  - Each MU has two state variables:
    - Age: age of trees {0-9, 10-19, 20-29, 30-39, 40-49}
    - Fuel: fuel load {very-low, low, medium, high, very-high}
    - 25^{100} states
- Actions:
  - Every 10 years for each MU: {grow, cut, fuel}
  - 3^{100} actions
- State transition function:
  - Actions are deterministic, but then fire burns stochastically depending on states and spatial arrangement of states
- Reward Function:
  - Value of timber cut and sold
  - Cost of fuel treatments

Status

- We have no algorithms that can handle such large spatio-temporal MDPs
- And there are typically 2000-3000 MUs

Summary

Sensor Placement — Optimal Sensor Placement

Data Collection — Detectability, Errors / Noise, Sampling Bias

Feature Extraction — <u>**Species classification**</u>, Recognizing individuals, Tracking individuals

Data Cleaning — Sensor failures, Networking failures, Recognition errors

Model Fitting — <u>**Species distribution models**</u>, Behavioral models, Dynamical systems models

Policy Optimization — <u>**Large Spatio-Temporal MDPs**</u>, Optima that are robust to model uncertainty

<u>**Coupling Multiple Problems**</u>

Acknowledgements

- Grant Support: US National Science Foundation, US Forest Service
- Species Distribution Models
  - Students: A. Lettkeman, P. Wilkins
  - Postdocs: R. Hutchinson
  - Faculty: W-K. Wong, T. Dietterich
- BugID:
  - Students: N. Larios, H. Deng, W. Zhang, N. Payet, M. Sarpola, C. Fagan, J. Yuen, S. Ruiz Correa
  - Postdocs: G. Martínez-Muñoz
  - Faculty: R. Paasch, A. Moldenke, D. A. Lytle, E. Mortensen, L. G. Shapiro, S. Todorovic, T. Dietterich
- Fire Management
  - Students: A. Gagnon, R. Houtman, J. Mark
  - Faculty: C. Montgomery, T. Dietterich, W-K. Wong

Questions?