Machine Learning in Ecosystem Informatics and Sustainability

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






Needed: Robust Optimal Policy Based on Sound Science

- 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”

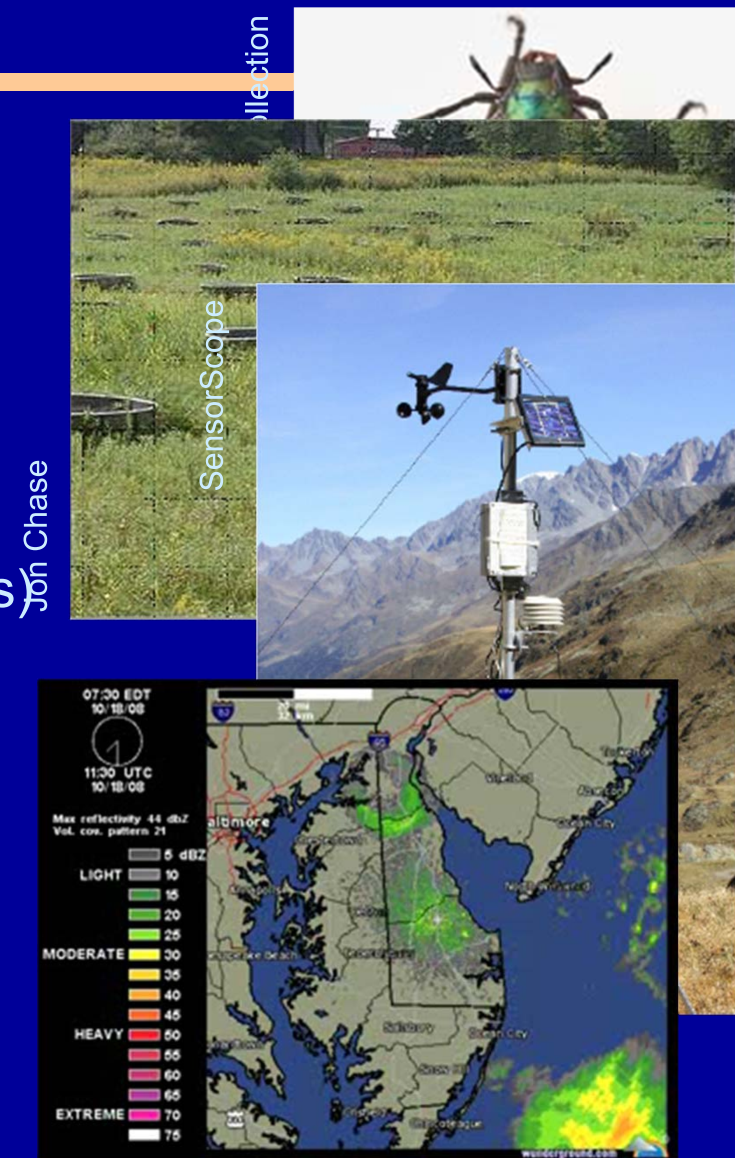
A Limiting Factor: Ecological Data

- Many ecological simulation models are based on little or no data
- Historical time series only extend back 100 years
  - Oldest continuous data set at HJ Andrews Experimental Forest is 1909-present
  - Most begin in 1990s
- Location, population size, interactions for virtually all species are unobserved

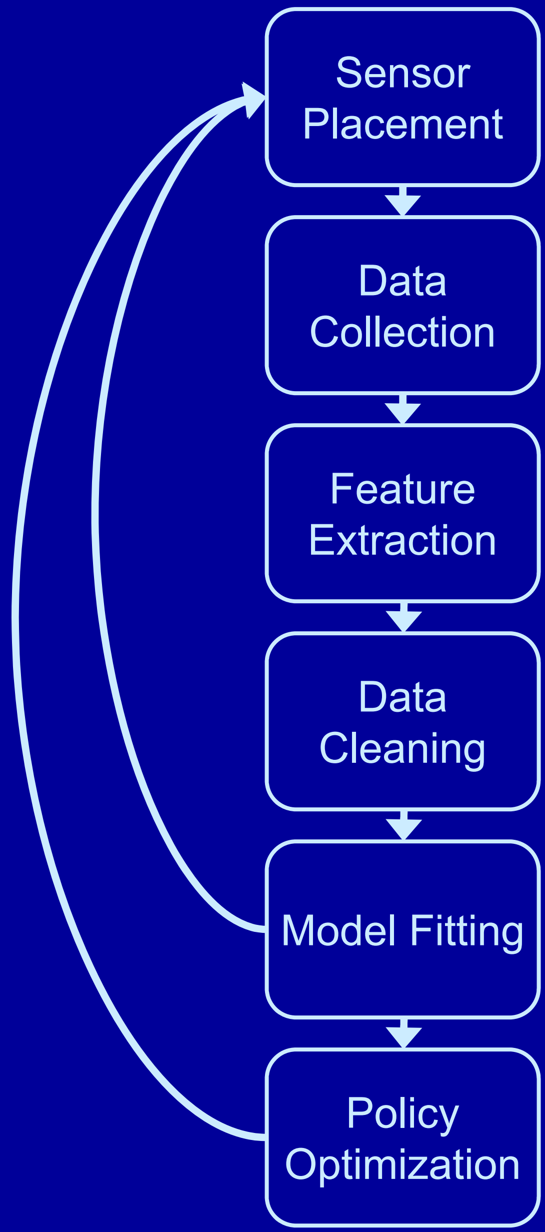
Ecosystem Sciences

Hypothesis Driven → Data Driven

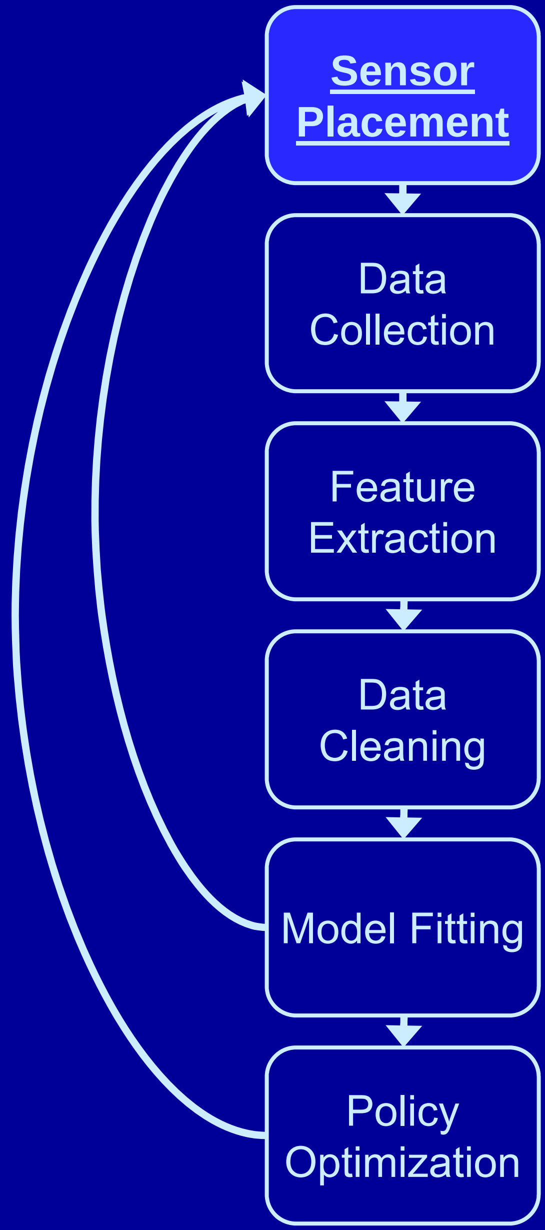
- Past approaches
  - Naturalists: museum collections
  - Artificial ecosystems (test tubes; barrels)
  - Isotope tagging of fluxes
- Emerging approaches
  - In-situ sensor networks
  - Radio/RFID tagging and tracking of organisms
  - Radar ornithology
  - Remote sensing



Data Pipeline

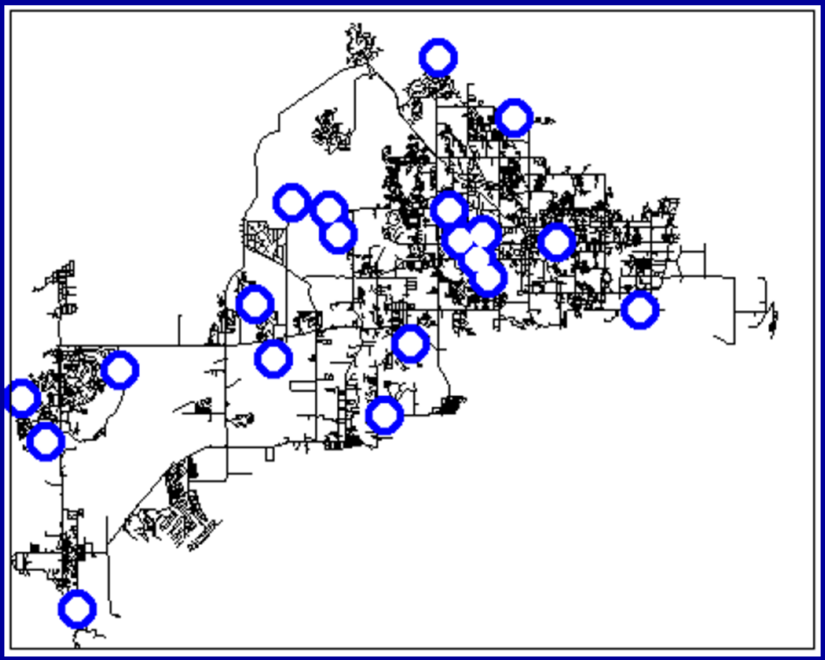


Data Pipeline



Optimal Sensor Placement

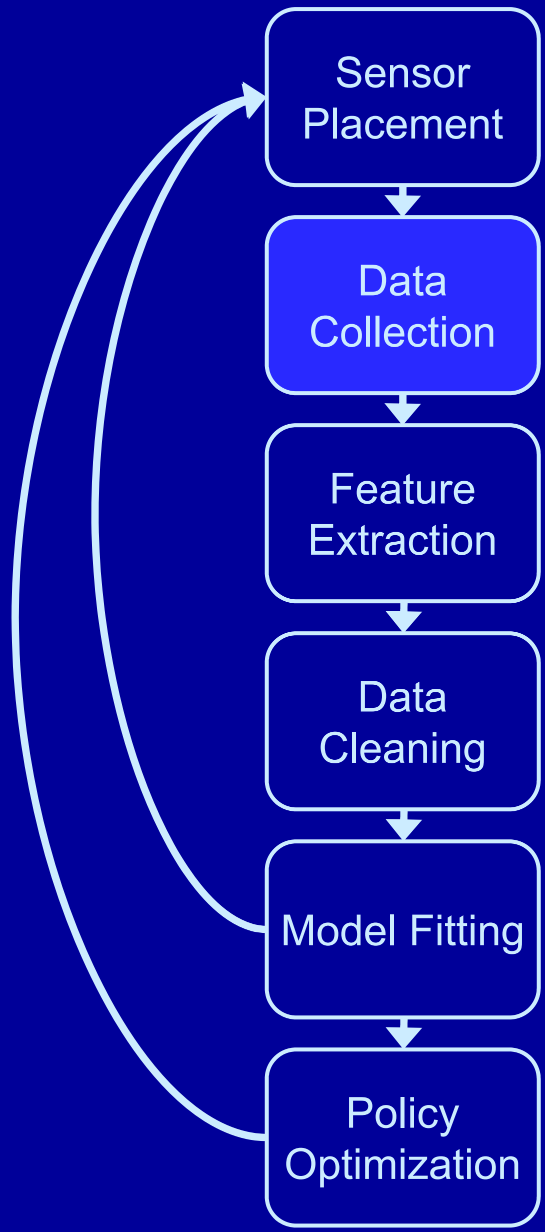
Optimal Sensor Placement for Environmental Data Collection



Leskovec et al, KDD2007

- Objectives
  - detection probability
  - improving model accuracy
  - improving causal understanding
  - improving policy effectiveness

Data Pipeline

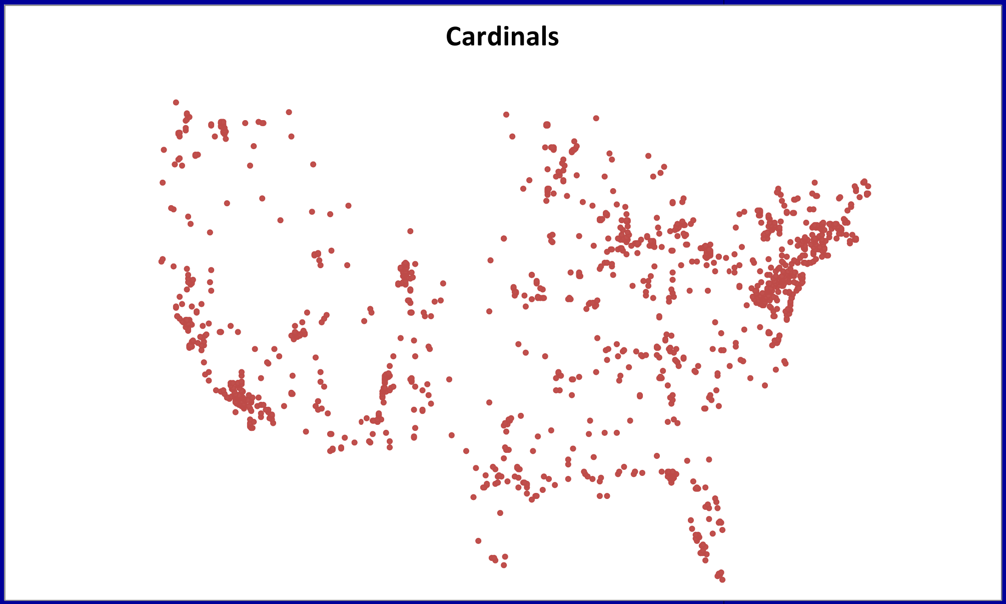


Optimal Sensor Placement

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



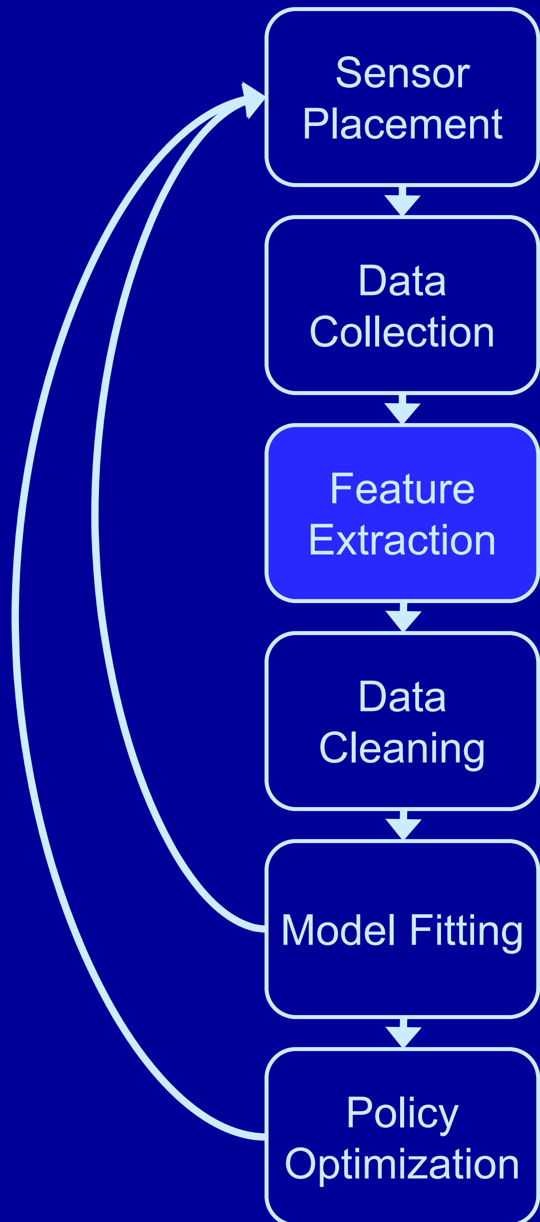
Phillips, Dudik, Elith, Graham, Lehmann, Leathwick, Ferrier: Sample Selection Bias and Presence-only Distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1), 181-197. 2009.

Detectability

- Birds in Forested Landscapes protocol
  - Step 1: 2 minutes silent listening and observing
  - Step 2: Play "con-specific" mating calls and listen/observe
  - Step 3: Play "predator mobbing" tape and listen/observe
- Coupled models of detectability and occurrence 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

<u>Species classification</u>
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

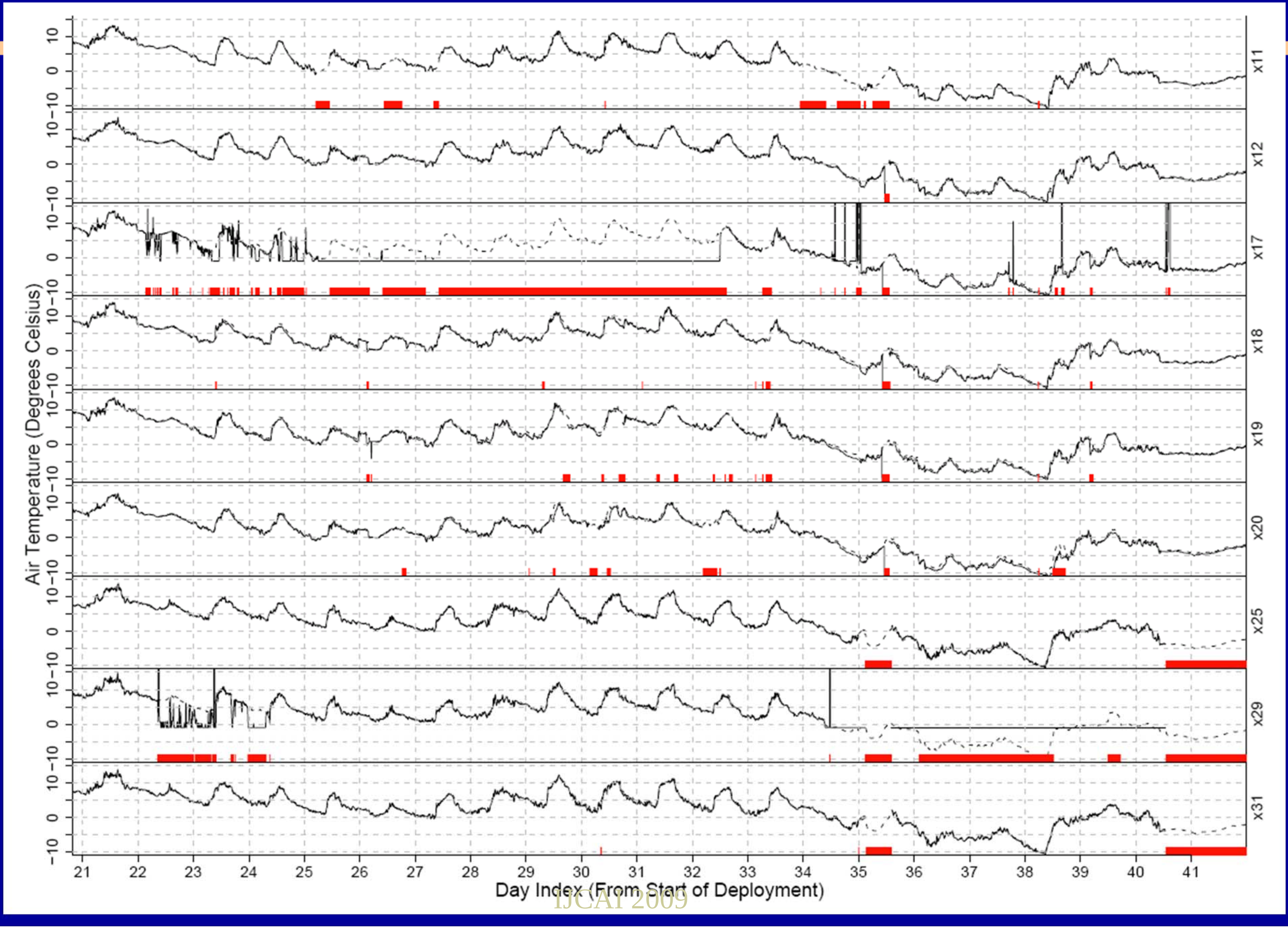
Optimal Sensor Placement

Detectability
Errors / Noise
Sampling Bias

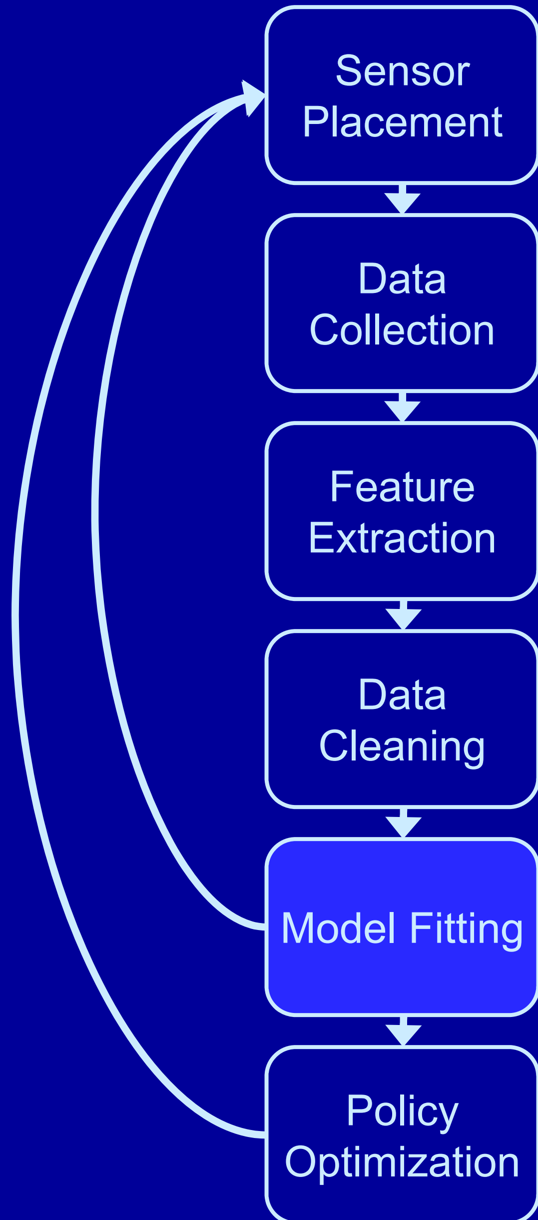
Species classification
Recognizing individuals
Tracking individuals

Sensor failures
Networking failures
Recognition errors

Multi-Sensor Anomaly Detection



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

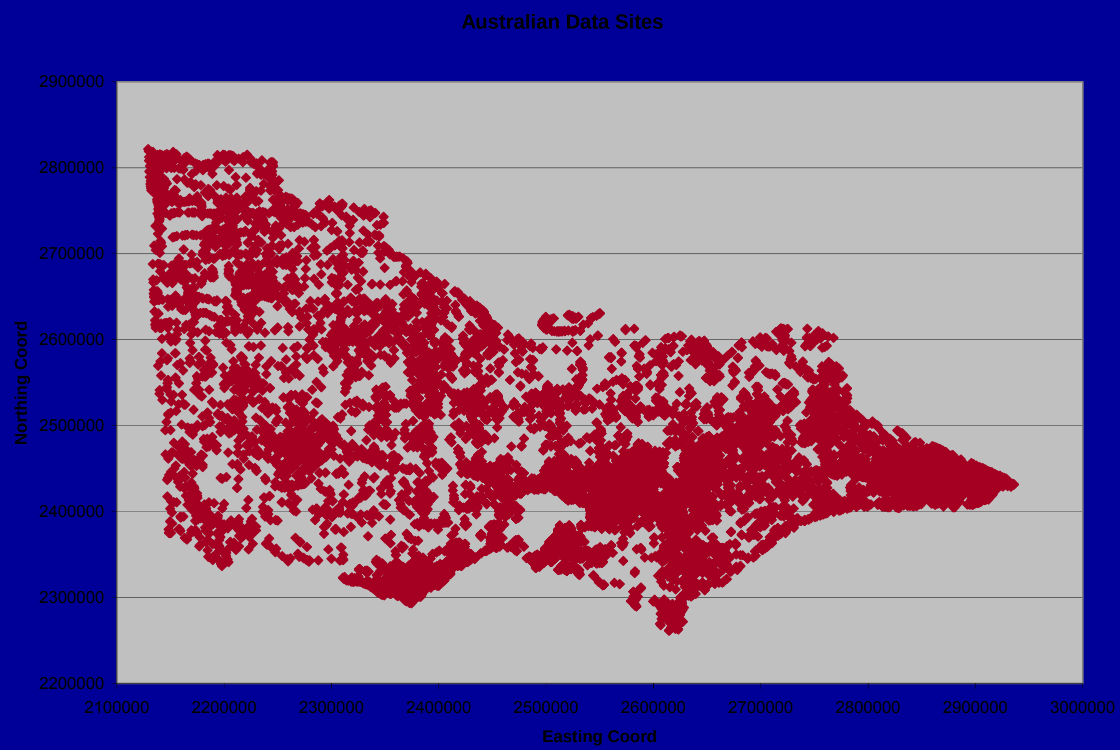
Coupling Multiple Problems

Species Distribution Models

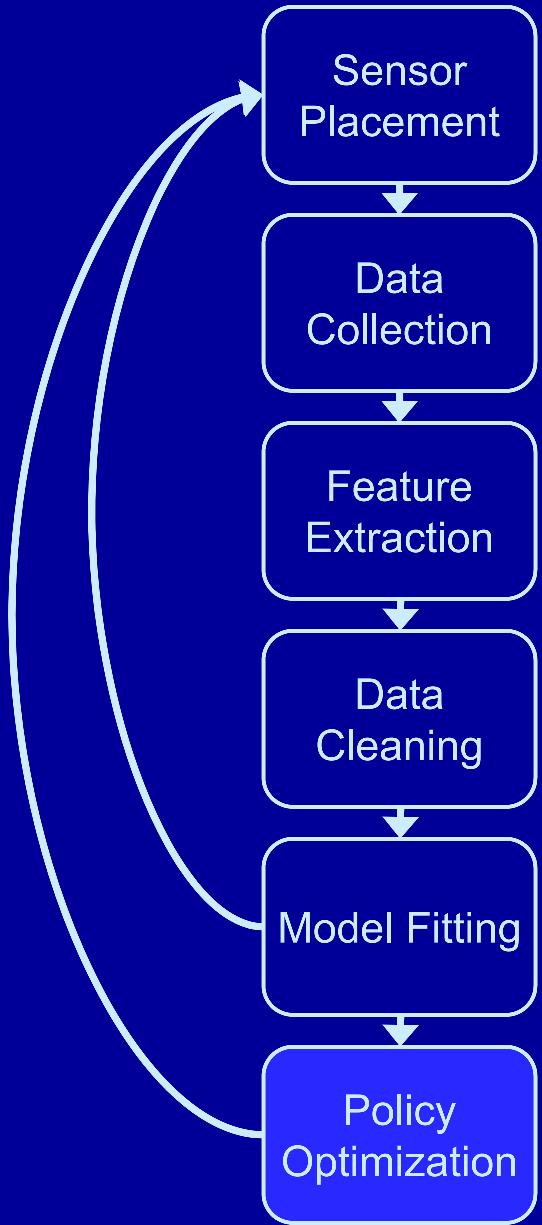
- What are the environmental/biological requirements for a species?
- Given:
  - Environmental features (elevation, soil properties, weather) of a site
  - Presence, presence/absence, or abundance of K species
- Find:
  - Probability that each of the K species will be found at new sites
  - Extrapolation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

Plants in Victoria

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



Data Pipeline



Optimal Sensor Placement

Detectability
Errors / Noise
Sampling Bias

Species classification
Recognizing individuals
Tracking individuals

Sensor failures
Networking failures
Recognition errors

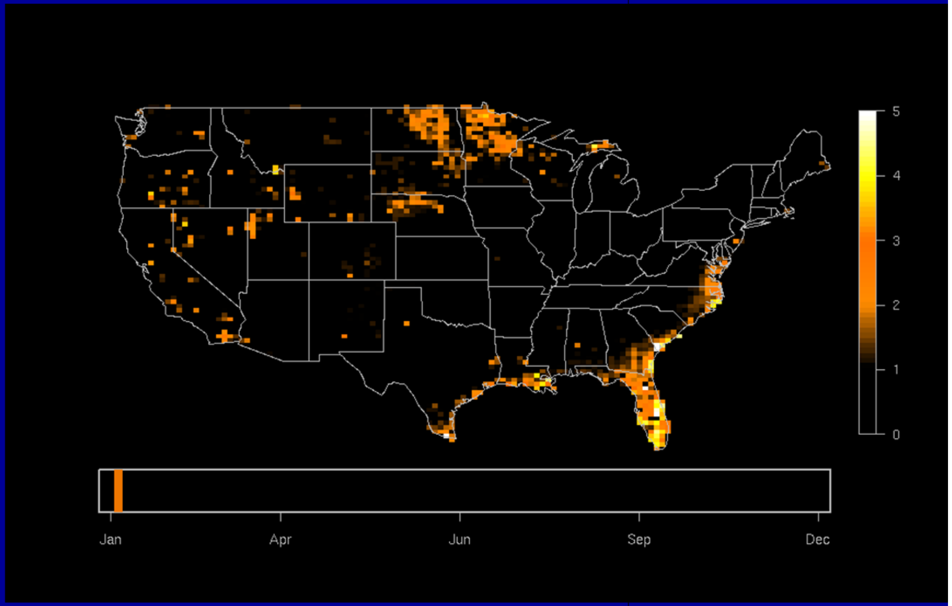
Species distribution models
Behavioral models
Dynamical systems models

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

- BugID Project: Arthropod Counting
- Automated Data Cleaning for Wireless Sensor Network Data

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
- **Three applications:**
  - stoneflies in freshwater streams
  - soil mesofauna
  - freshwater zooplankton

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”

Application 3: Freshwater Zooplankton

Daphnia

Bosmina

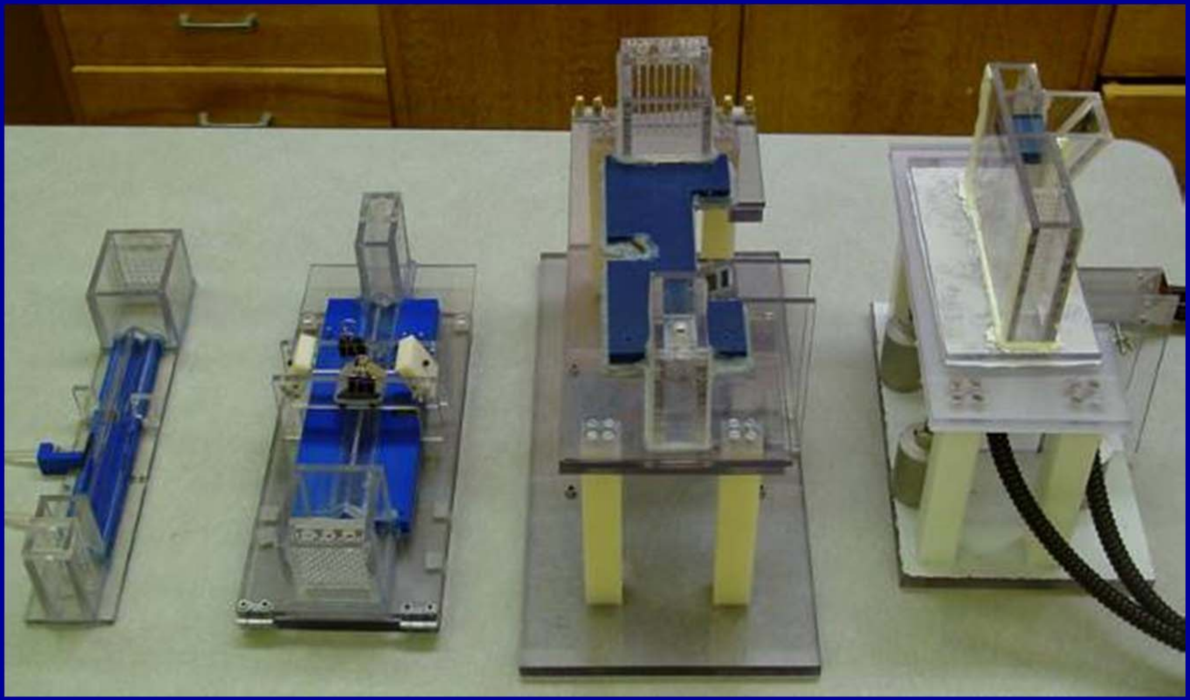
Polyphemus
(cladocerans)

Cyclops
(copepod)

- Measure biodiversity in freshwater lakes
- 70 species

Images from Microscopy-UK.

Image Capture Apparatus



Stonefly Imaging

Soil Mesofauna Imaging

Robotic Extraction of Specimens

Computer Vision Challenges(1)

- Highly-articulated objects with deformation

Computer Vision Challenges(2)

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

tergites → become → wings

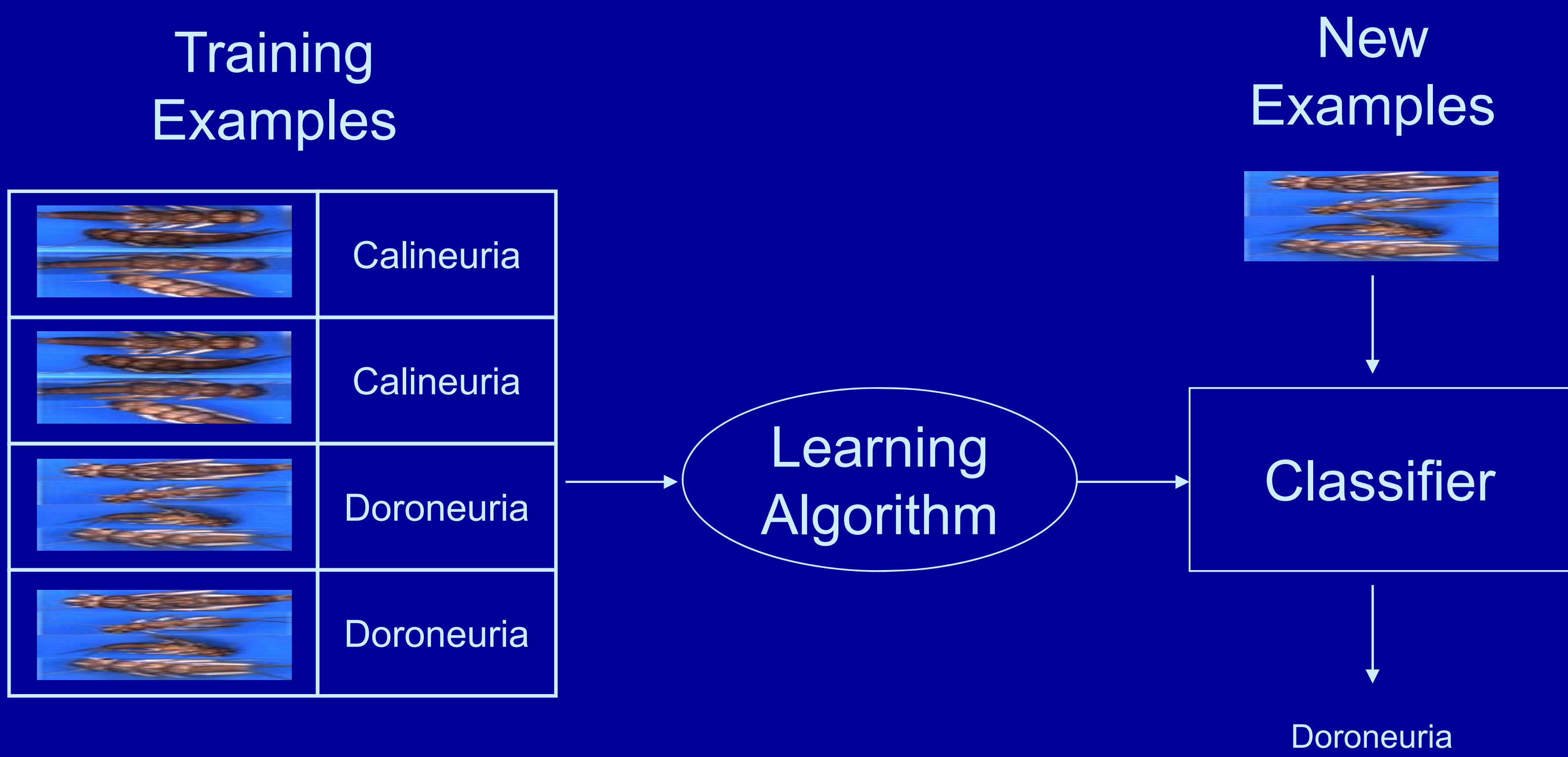
Computer Vision Challenges(3)

- Small between-class differences

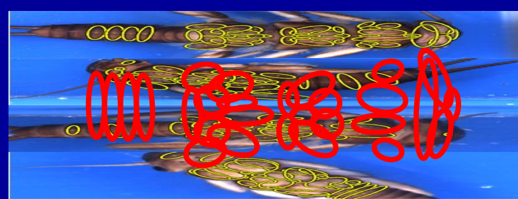
Calinueria

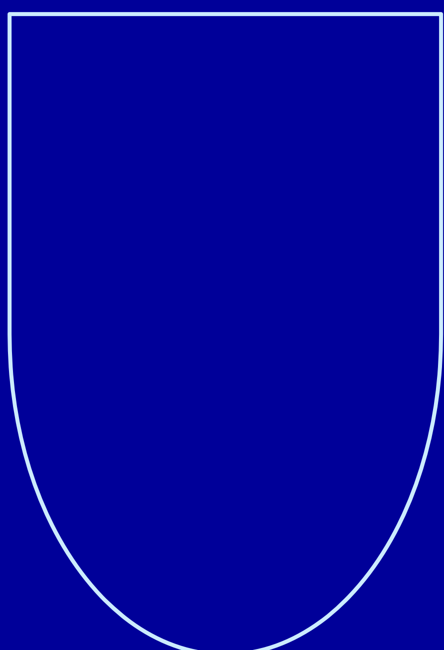
Doronueria

Machine Learning



Region-Based Approaches: Convert Image to Bag of Patches

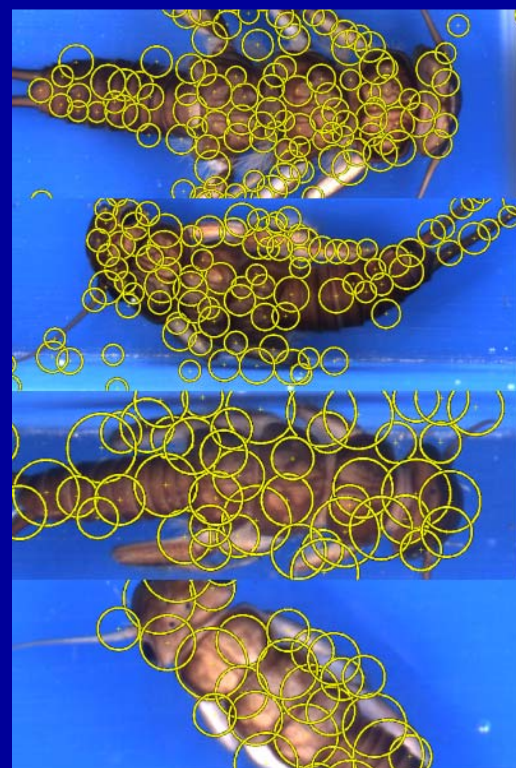




- Handles
  - Occlusion
  - Rotation, translation
  - Scale (with scale-independent patch representation)
  - Partial out-of-plane orientation
  - Articulation / Pose
- Problem:
  - How to define the patches?
  - How to represent each patch?
  - How to classify a BAG of patches?

Defining the Patches: Interest Region Detectors

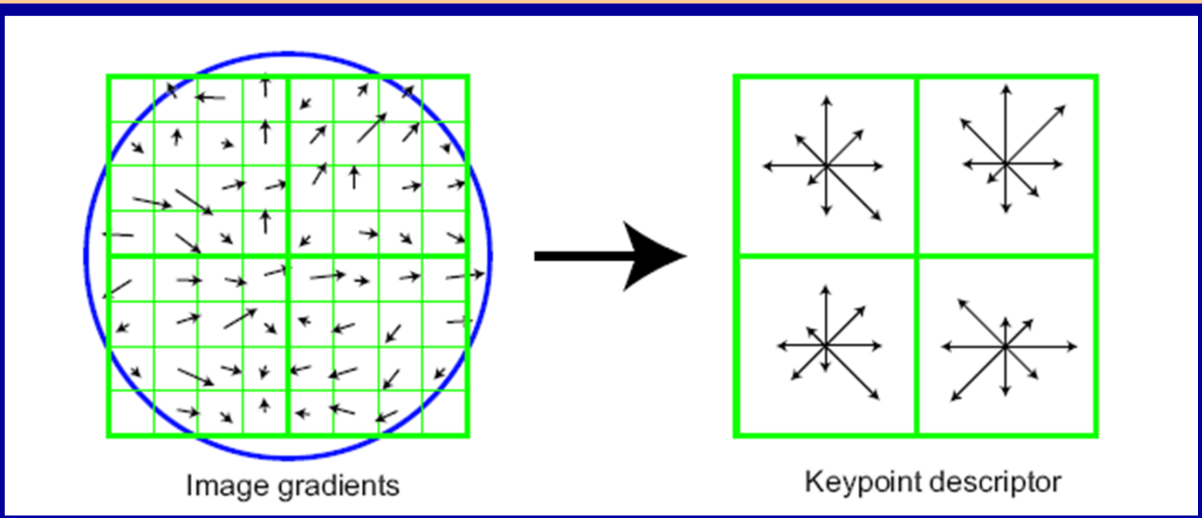
Hessian-Affine Detector



Kadir Entropy Detector

PCBR Detector

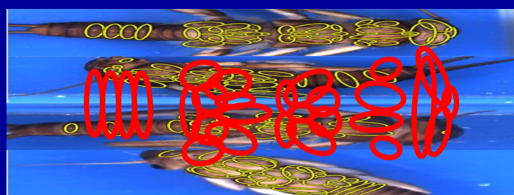
Representing the Patches: SIFT (Lowe, 1999)



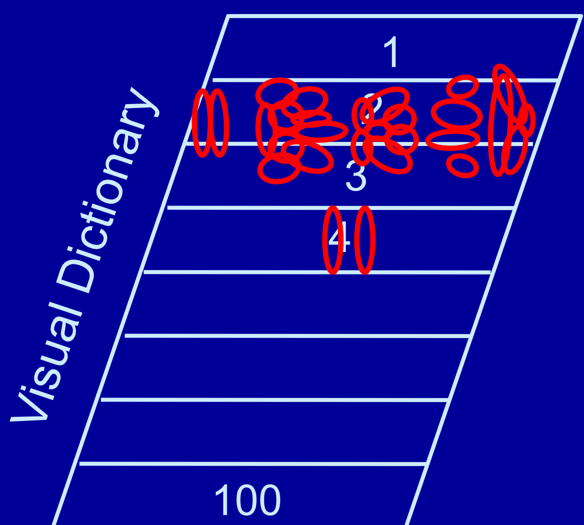
(Lowe, 1999)

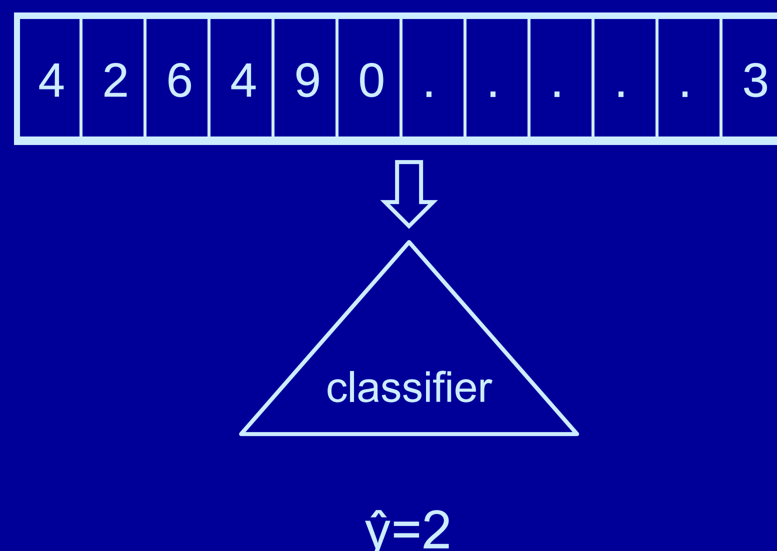
- Morph ellipse into a circle
- Compute intensity gradient at each pixel in 16x16 region
- Rotate whole circle according to dominant intensity gradient
- Weight gradients by a gaussian distribution (indicated by circle)
- Collect into histograms within each 4x4 region (gives 16 histograms)
- Result: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

Classify Bag of Patches
Method 1: Visual Dictionaries



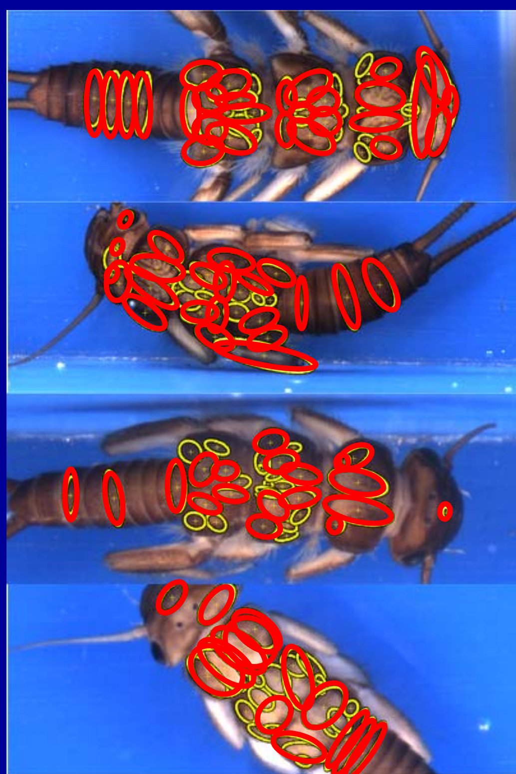
- “look up” each patch in dictionary and count into a feature vector
- feature vector is then given to the classifier

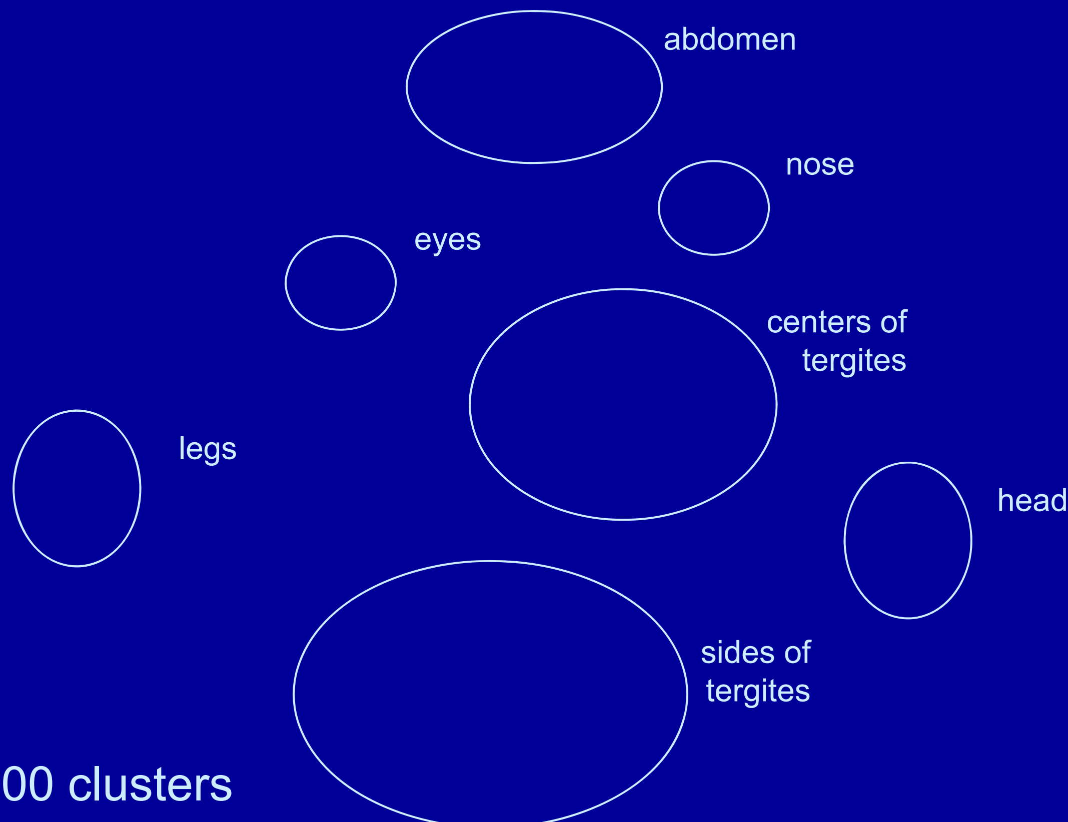




Learn visual dictionary via clustering

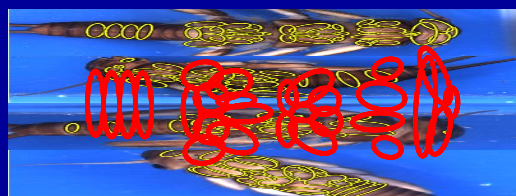
- Gaussian Mixture Model (k=100) with diagonal covariance matrices (EM, initialized with K-means)

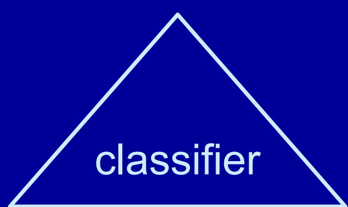




100 clusters

Classify Bag of Patches
Method 2: Multiple-Instance Classifier





$\hat{y}$=2

2	8	1	3	0	0	6	4	2

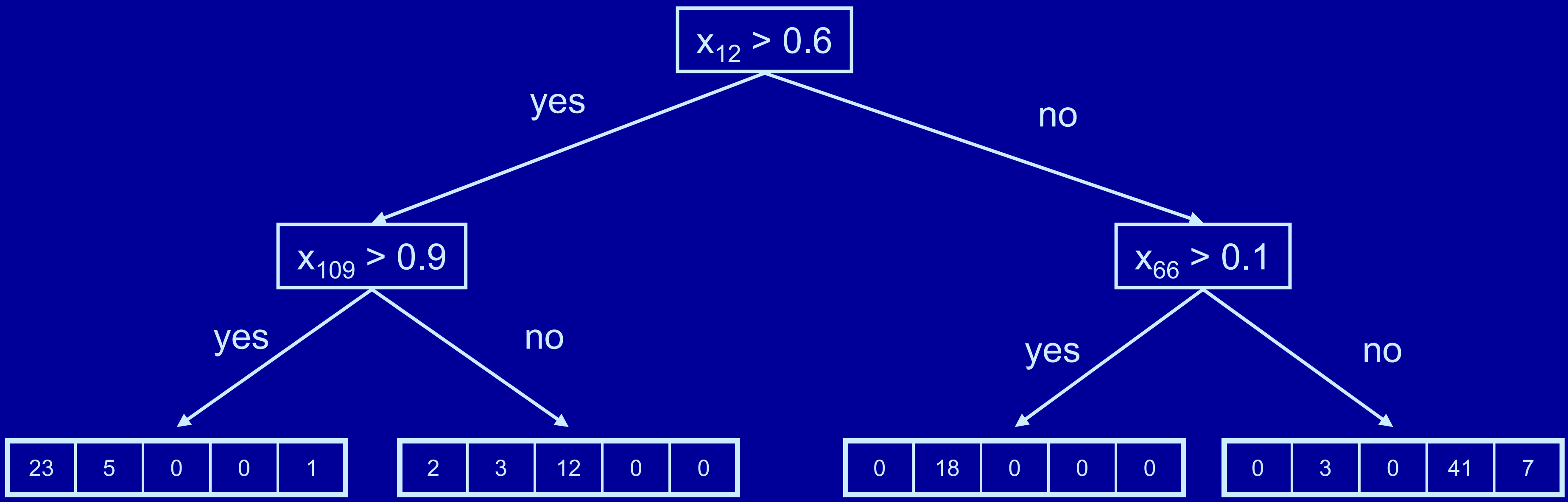
votes

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

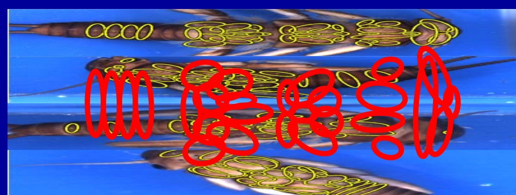
- The classifier predicts the class of the image separately from each patch
- These vote to make the final decision

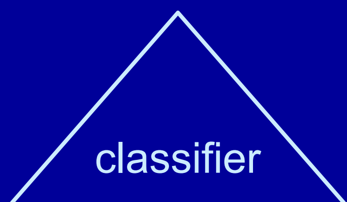
Improved Multiple-Instance Classification

- 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





2	3	12	0	0

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

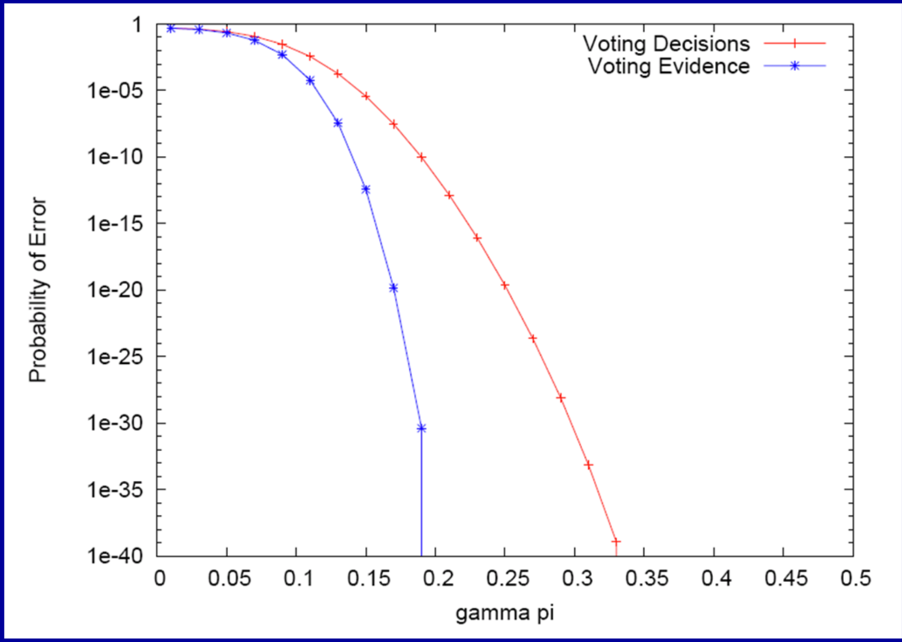
87	14	34	6	61

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



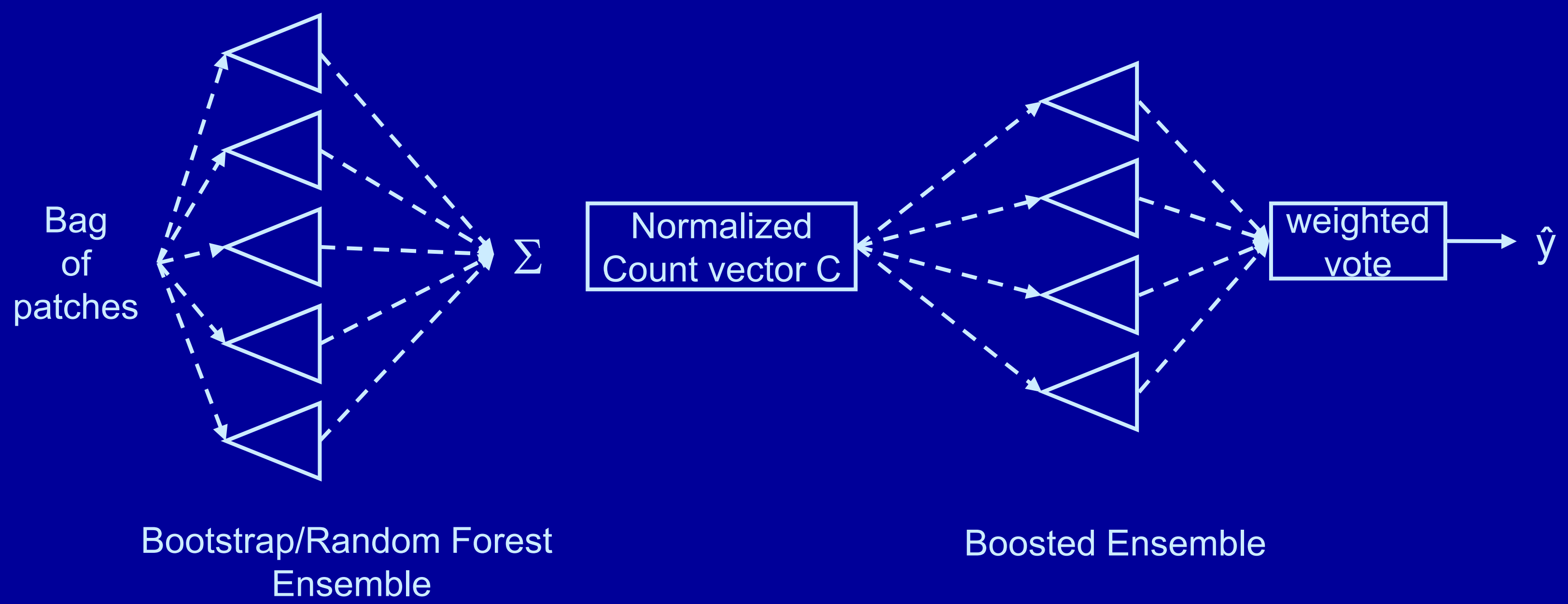
γ = margin of decision tree nodes
π = fraction of non-noise patches

Ensemble Learning

- Idea: Learn multiple evidence trees and have them vote
- Question: How to construct multiple diverse trees?
  - **Bootstrapping**: train each tree on a different bootstrap sample
    - Majority vote
  - **Boosting**: train each tree based on a sample containing 50% points misclassified by the previous trees and 50% points correctly classified by previous trees
    - Focuses subsequent trees on the misclassified points
    - Weighted vote
  - **Random Forests**: at each node, randomly sample a subset of features and choose the best split from among them
    - Majority vote

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

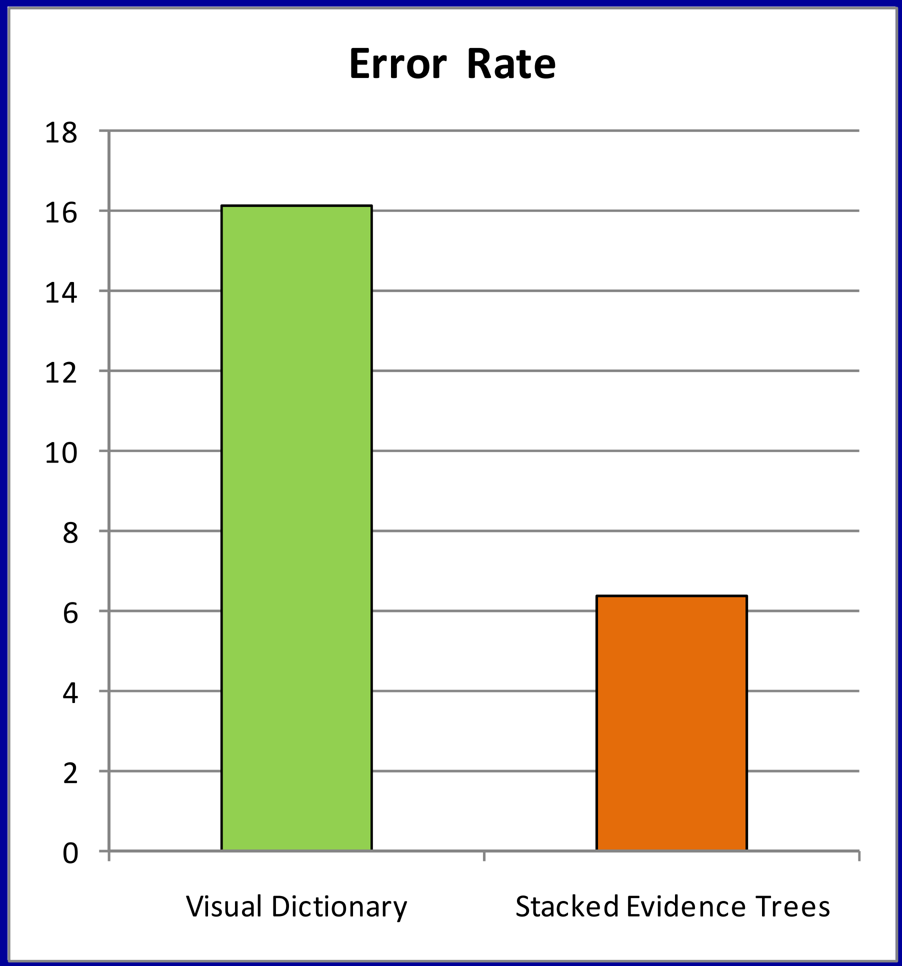
Yor

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



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

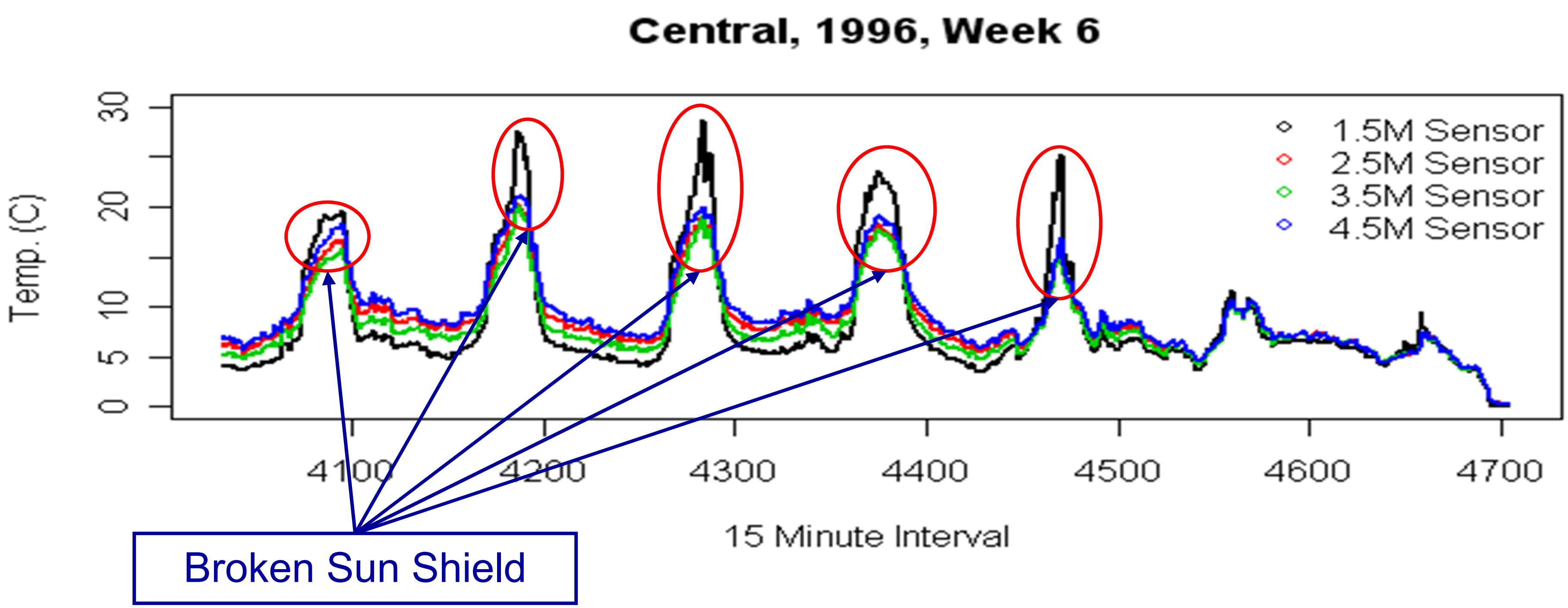
Next Steps

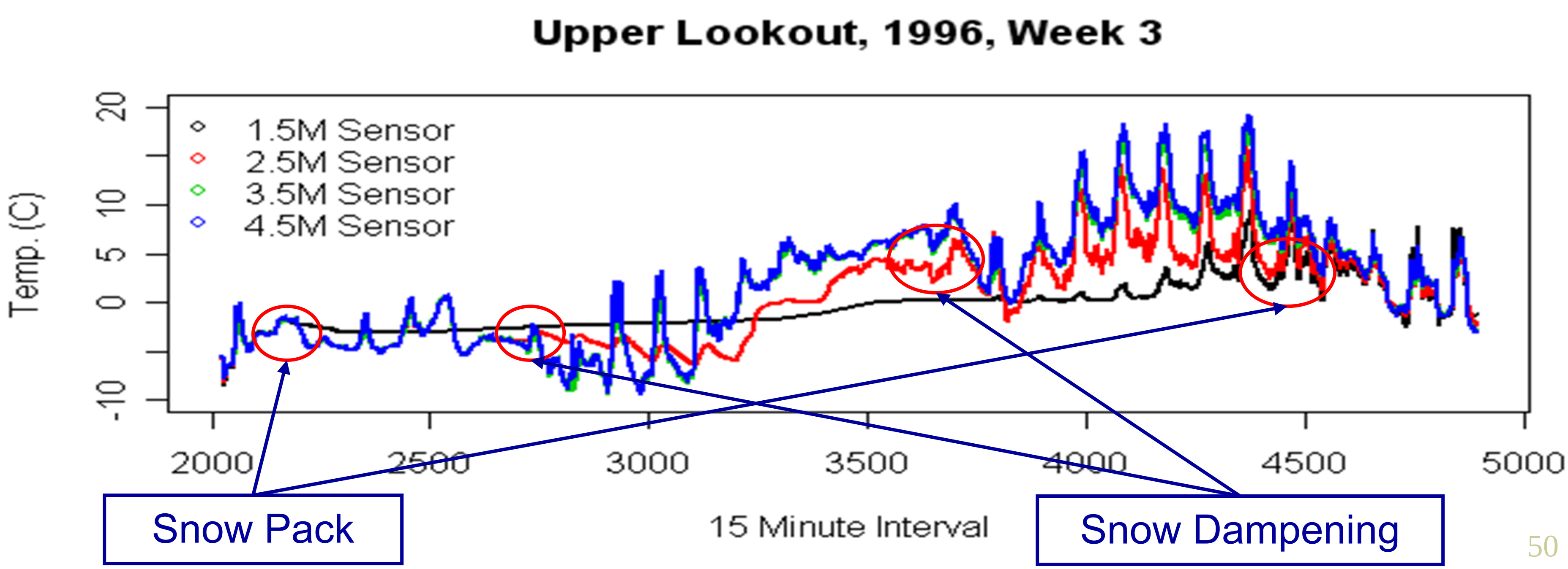
- Stoneflies
  - Detecting and Rejecting “Distractors”
  - Extending coverage to Ephemeroptera (mayflies) and Trichoptera (caddis flies)
  - EMAP study
- Soil Mesofauna
- Freshwater Zooplankton
- Moths
- Shellfish Larvae

Outline

- BugID Project: Arthropod Counting
- Automated Data Cleaning for Wireless Sensor Network Data

Upper Lookout Met. Station


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

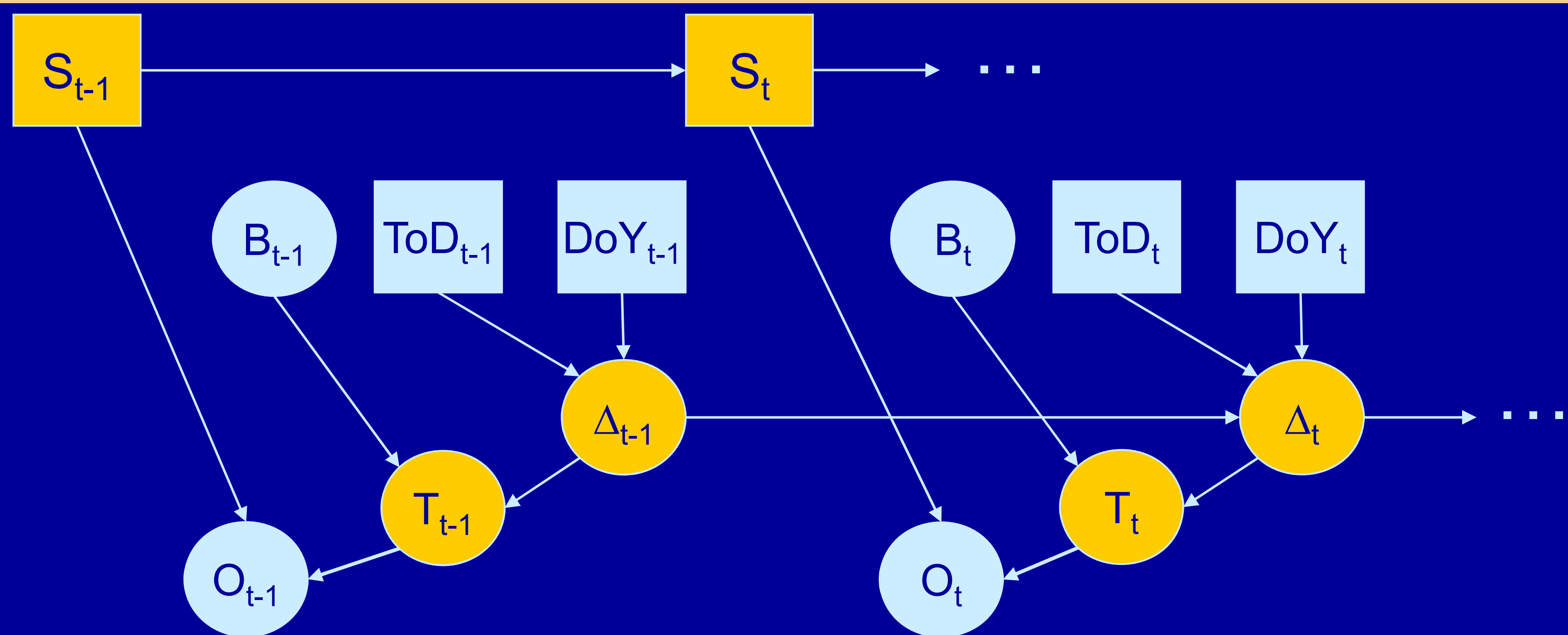

Upper Lookout, 1996, Week 3
Temp. (C)
-10
0
5
10
20
1.5M Sensor
2.5M Sensor
3.5M Sensor
4.5M Sensor
2000
2500
3000
3500
4000
4500
5000
15 Minute Interval
Snow Pack
Snow Dampening

Approach: Learn a Very Accurate Model of Normal Behavior

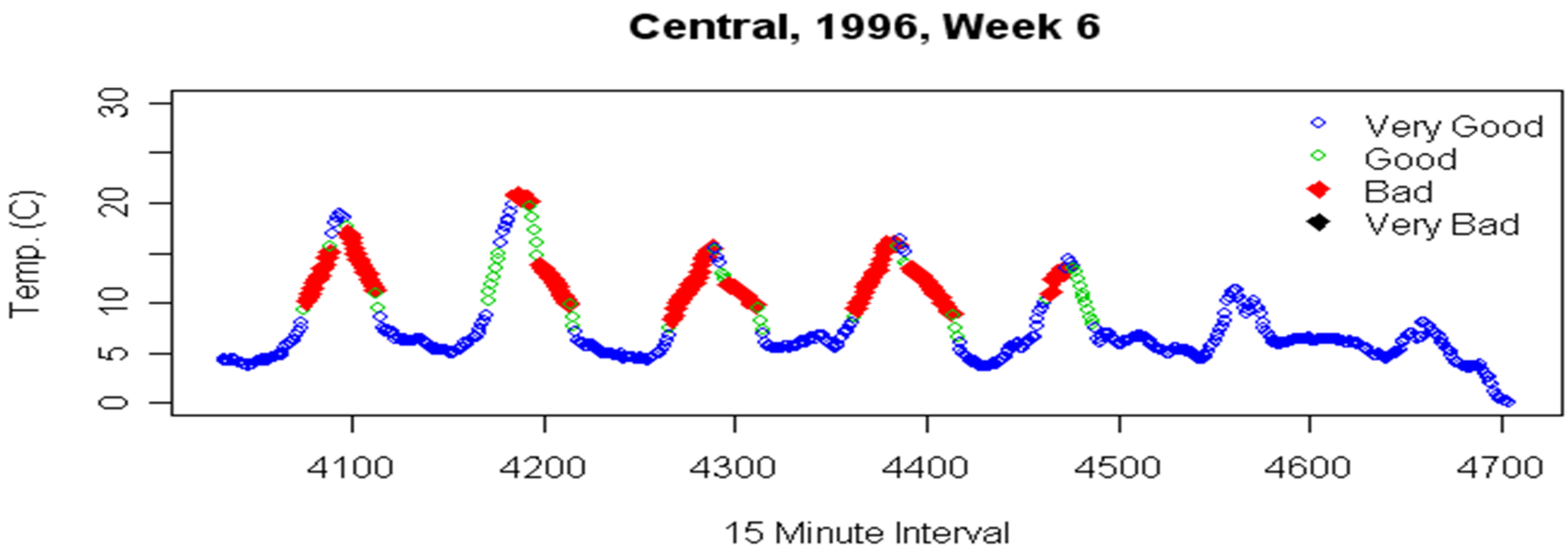
P(current observation | previous observations)

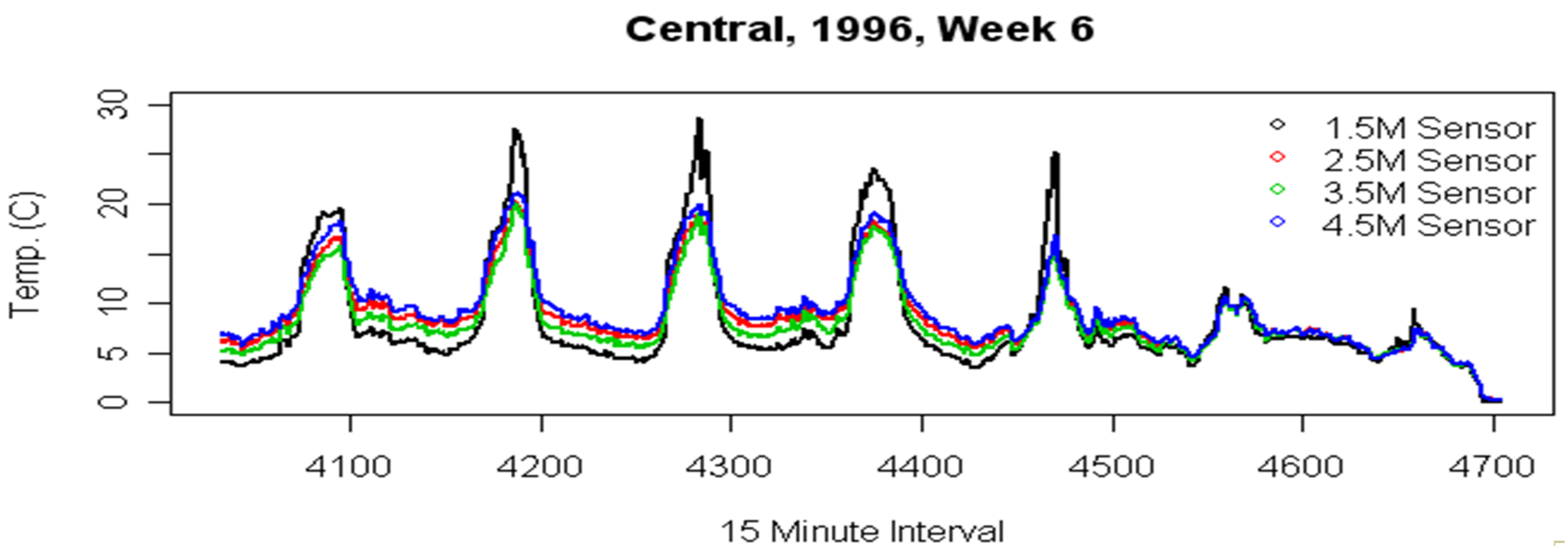
- If predicted probability is too low, then declare an anomaly

Single Sensor
Bayesian Network Model



S: Sensor State (Very Good, Good, Bad, Very Bad)
ToD: Time of Day (the quarter-hour)
DoY: Day of Year (365 day year)
B: Baseline Temperature
Δ: Deviation from Baseline
T: Predicted Temperature
O: Observed Temperature


Central, 1996, Week 6
Temp. (C)
0
5
10
20
30
4100
4200
4300
4400
4500
4600
4700
15 Minute Interval
Very Good
Good
Bad
Very Bad


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

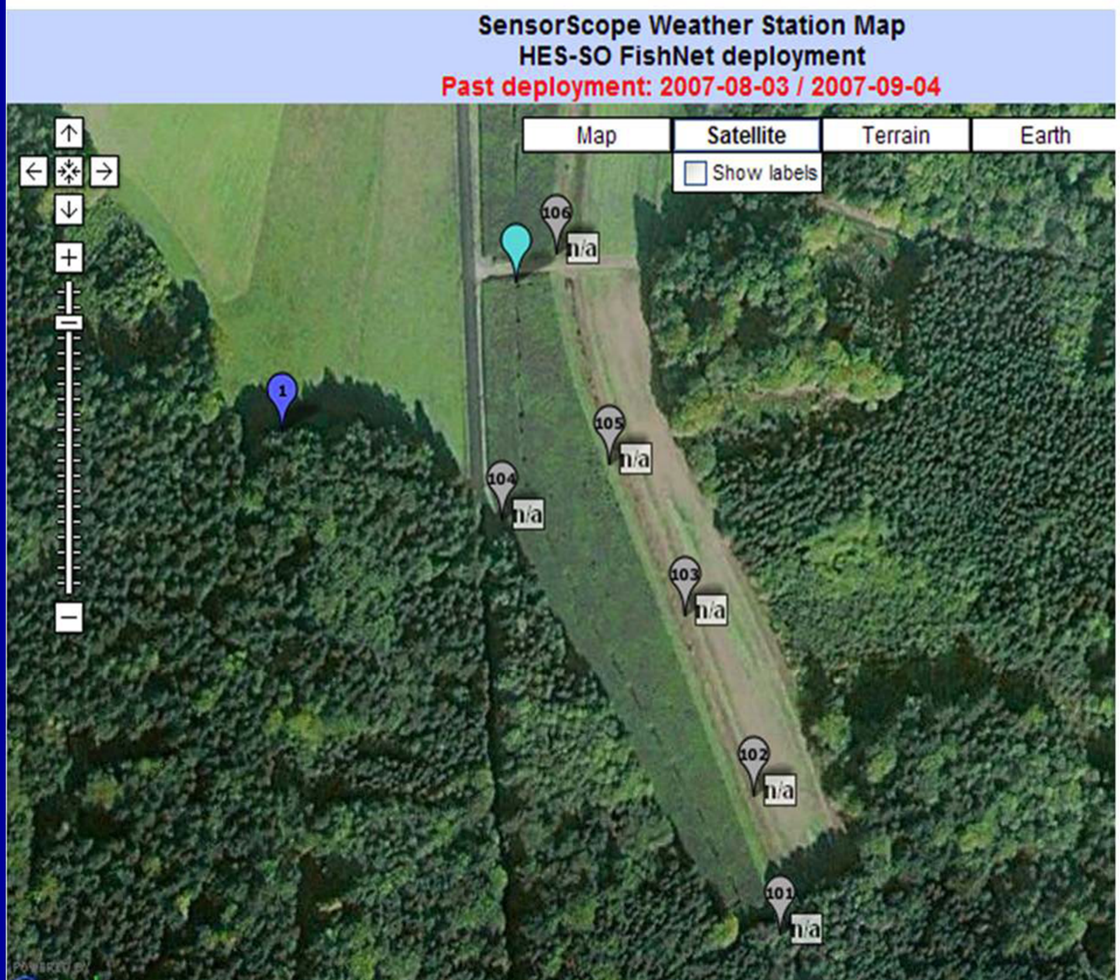
Assessment

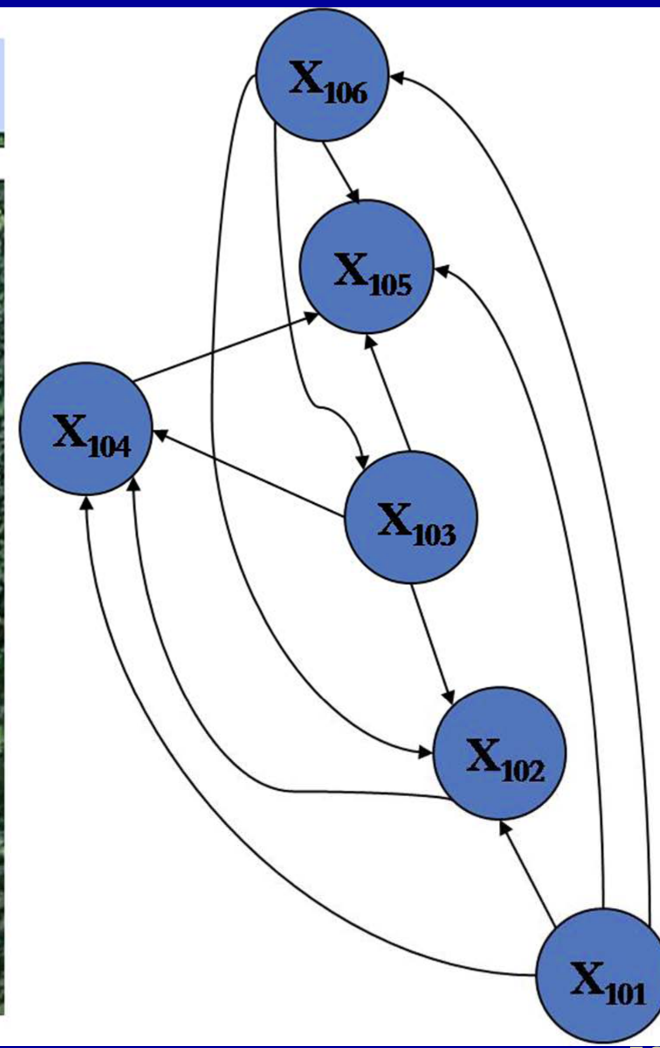
- Assessment:
  - near 100% recall for anomalies
  - 5.3% false positive rate
  - would allow expert to ignore 94% of data = 15x speedup in manual cleaning time

Multiple Sensors

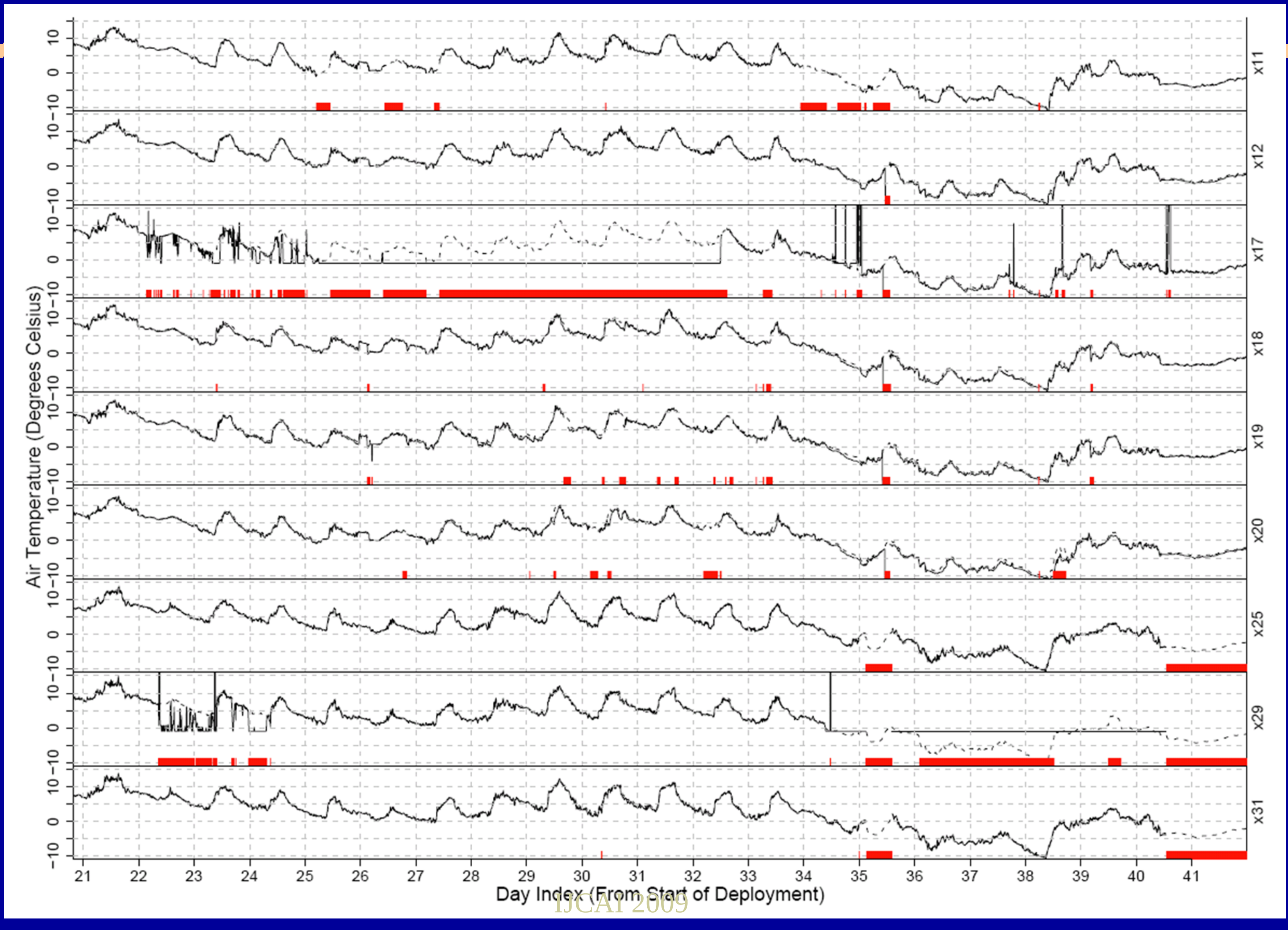
- Discover correlation structure among multiple sensors
- Exploit this to make more accurate inferences

Example: SensorScope
(EPFL, Switzerland)



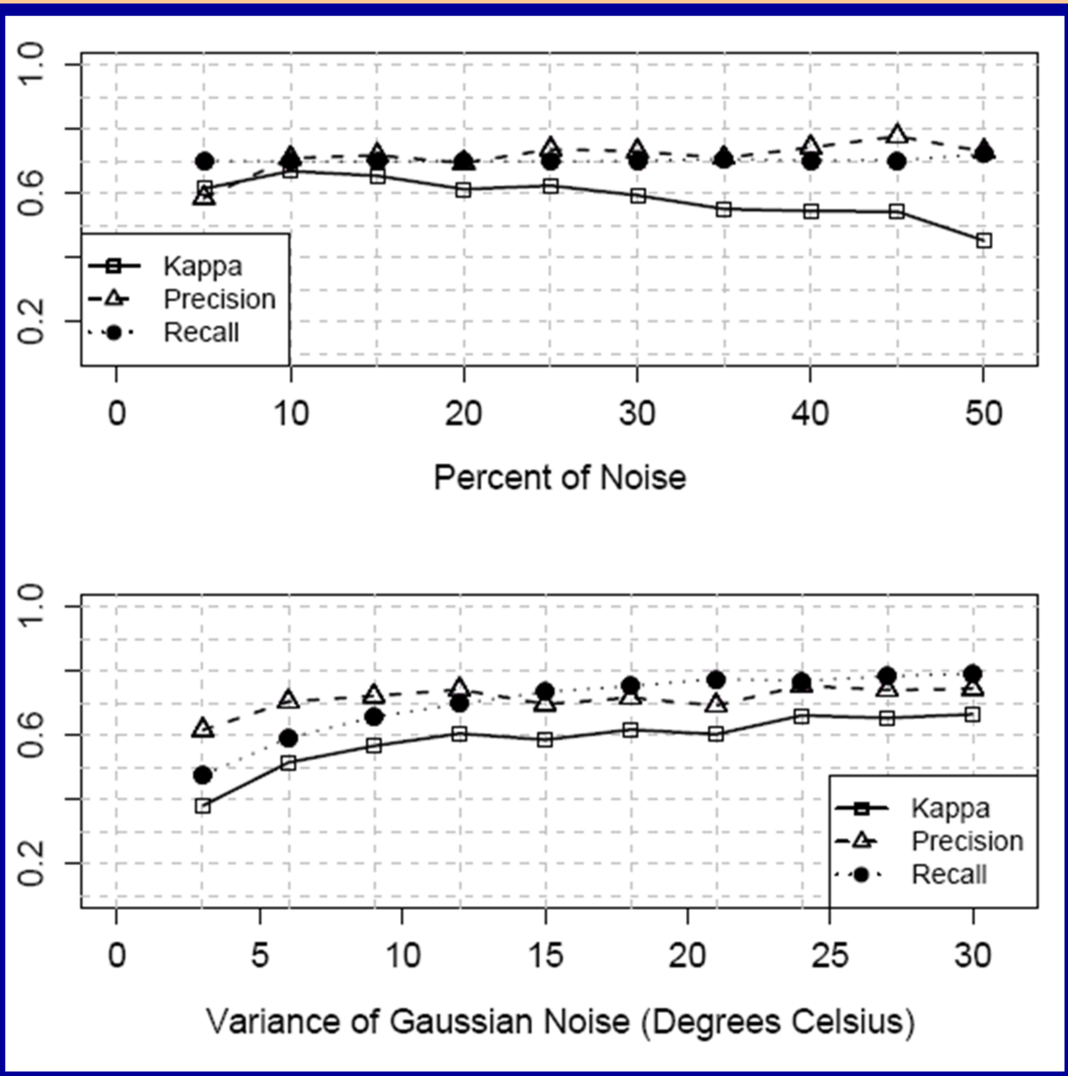


Multi-Sensor Anomaly Detection



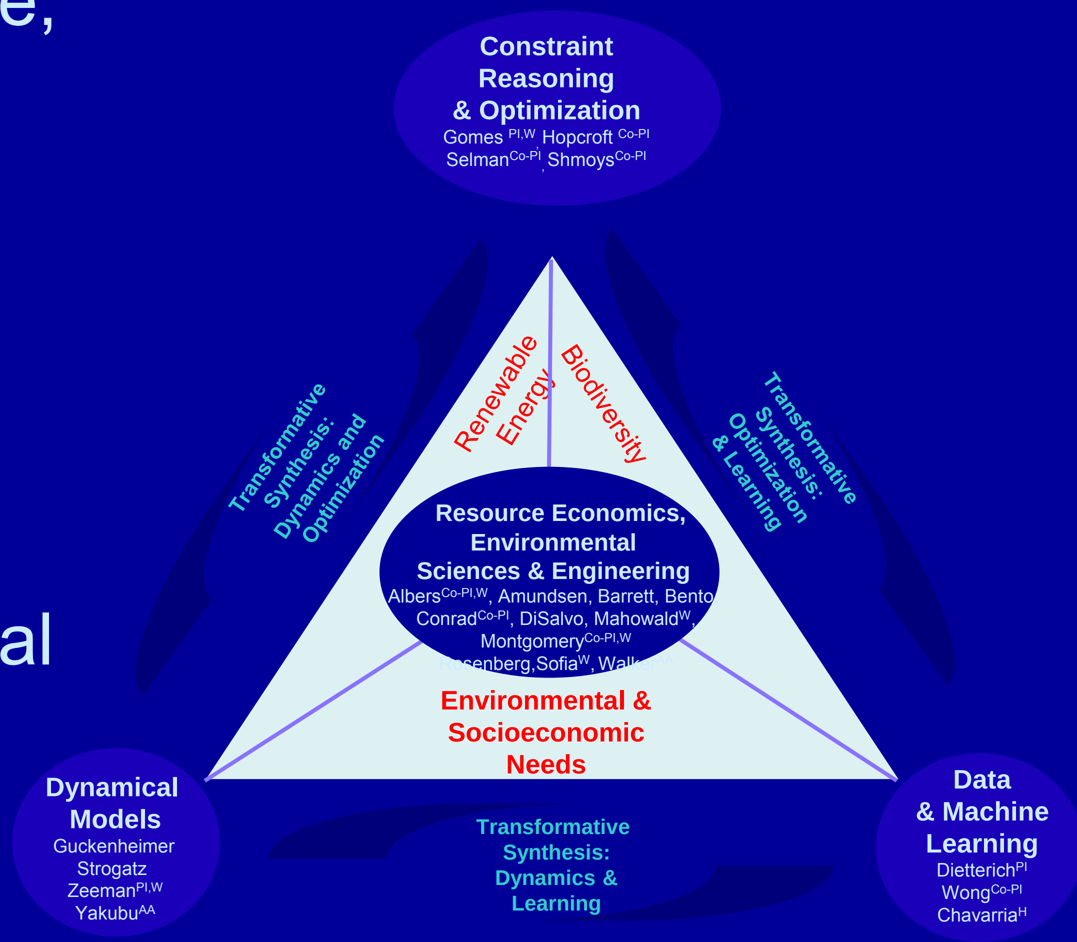
Multiple Sensor Evaluation

- Protocol:
  - Insert artificial anomalies
  - Measure how well we can detect them
- Results:
  - Robust to large amounts of noise
  - Insensitive to magnitude of noise except at very low levels



Institute for Computational Sustainability

- Cornell, Oregon State, Bowdoin, Howard U.
  - PI: Carla Gomes
  - co-PIs: Tom Dietterich, David Shmoys
- Goal: Identify and solve novel computational problems in ecological science, policy, and renewable energy



Summary

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

Optima that are robust
to model uncertainty

Coupling Multiple Problems

For More Information…

- Graduate program in Ecosystem Informatics: http://ecoinformatics.oregonstate.edu/
- Summer Institute in Ecosystem Informatics: http://eco-informatics.engr.oregonstate.edu/
- Institute for Computational Sustainability http://www.computational-sustainability.org/

Acknowledgements

- Grant Support: US National Science Foundation
- 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. G. Dietterich
- Data Cleaning: Ethan Dereszynski

Questions?