# Tracking Environmental Change through the Data Resources of the Bird-Monitoring Community

Mirek Riedewald, Rich Caruana, Daniel Fink, Wesley M. Hochachka, Steve Kelling, Art Munson, Ben Shaby, Daria Sorokina
Cornell University, Ithaca, NY
{mirek,caruana,mmunson,daria}@cs.cornell.edu; {df36,wmh6,stk2,bshaby}@cornell.edu
http://www.avianknowledge.net

## Our Vision

**Understand ecological processes at the continent scale.**

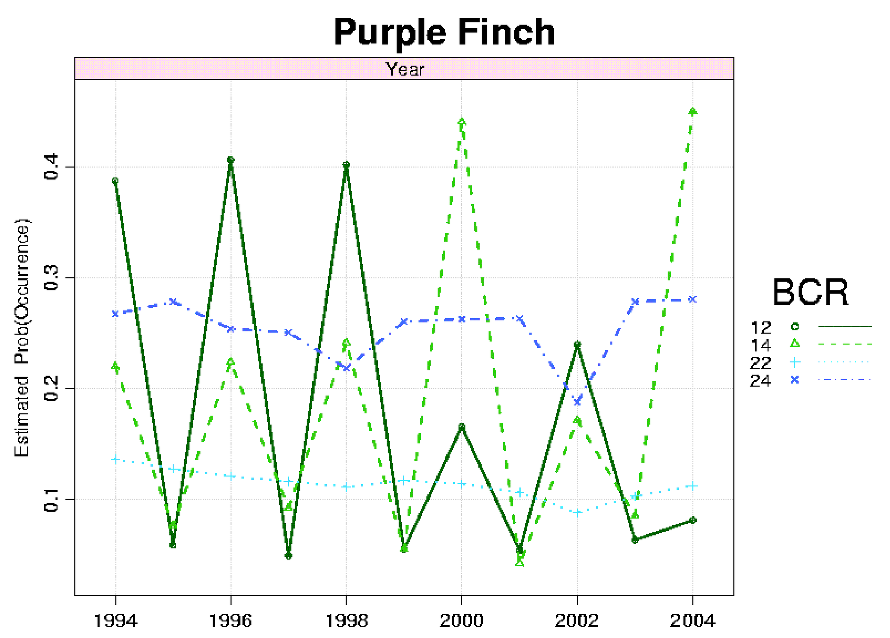
Examples: Where, how, and why are bird populations changing? Which features of the environment are the most important ones and how do they interact?

Traditional approach: Hypothesis, design study, collect data, test hypothesis
-Not feasible to collect data for large-scale analysis
-System complexity: Limited ability to come up with relevant hypothesis

**Our approach: Develop novel techniques for mining observational data**

Main idea: Find interesting patterns automatically, use them for hypothesis generation
Advantages: (1) Large amounts of observational data already available, (2) Might find surprising patterns that lead to new discovery

Challenges:
-Is observational data useful for deriving knowledge at all? We believe the answer is YES!
-Need novel methods to improve accuracy and confidence of results from observational data
  -Discovery of biological and ecological patterns, include domain knowledge
-Need to handle high-dimensional data from various protocols: Computational cost, protocol bias, observer bias, noise, missing values etc.



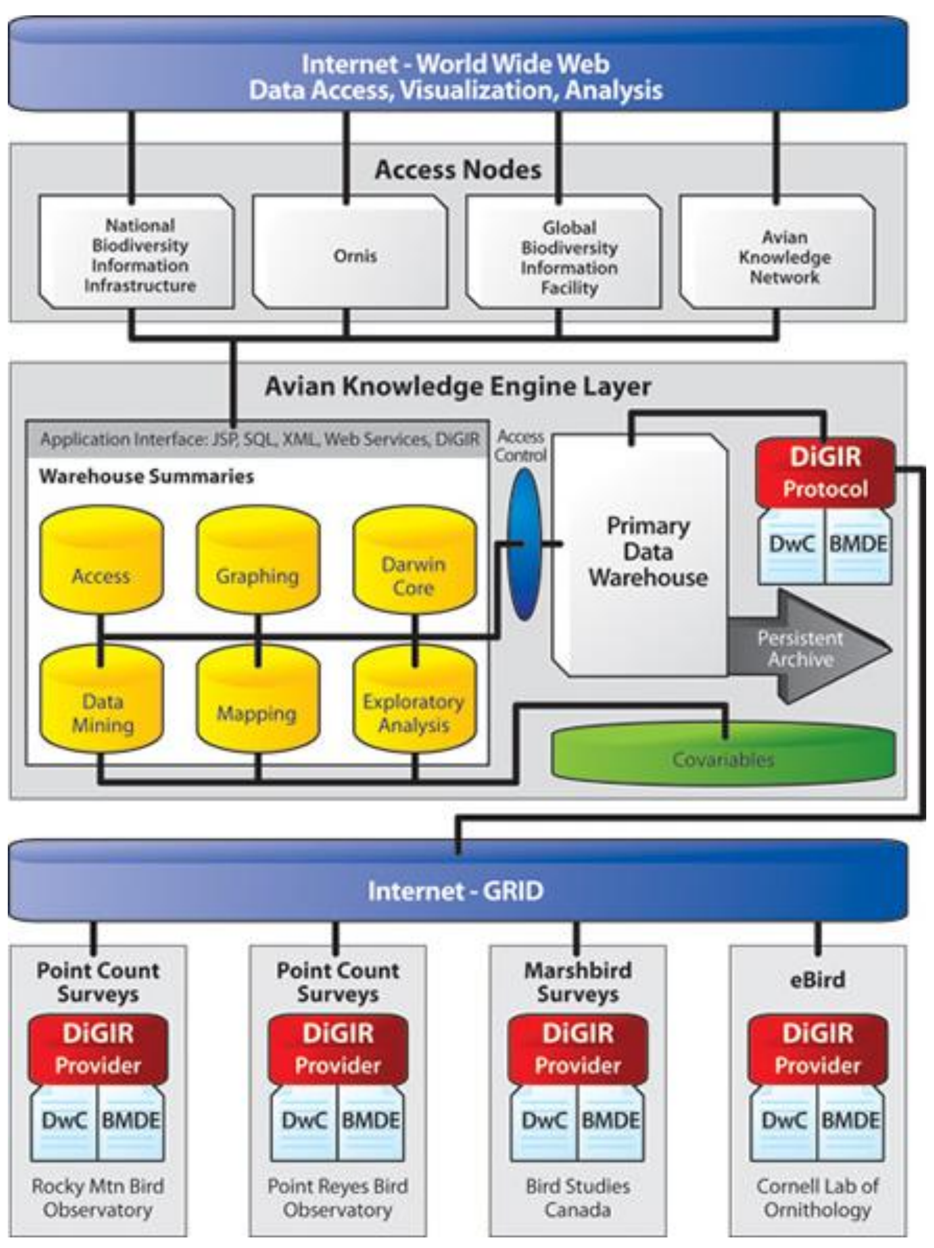

## Massive Data Collection and Sharing

Many partner organizations contribute data, for example:

Cornell Lab of Ornithology; Hawk Mountain; Avian Science Center; Redwood Sciences Lab US Forest Service; prbo; Conabio Comisión nacional para el conocimiento y uso de la biodiversidad; USGS science for a changing world, Patuxent Wildlife Research Center; Rocky Mountain Bird Observatory; Bird Studies Canada; HawkWatch International; Klamath Basin Bird Observatory

**Current data resources:**

-37 million observations
-200,000 locations
-Time period: 1900 to 2007
-2266 species and sub-species
-1000s of variables (habitat, observer effort, land cover, climate, weather, human demographics)
-Growing rapidly: Millions of new observations every year, more variables from GIS data

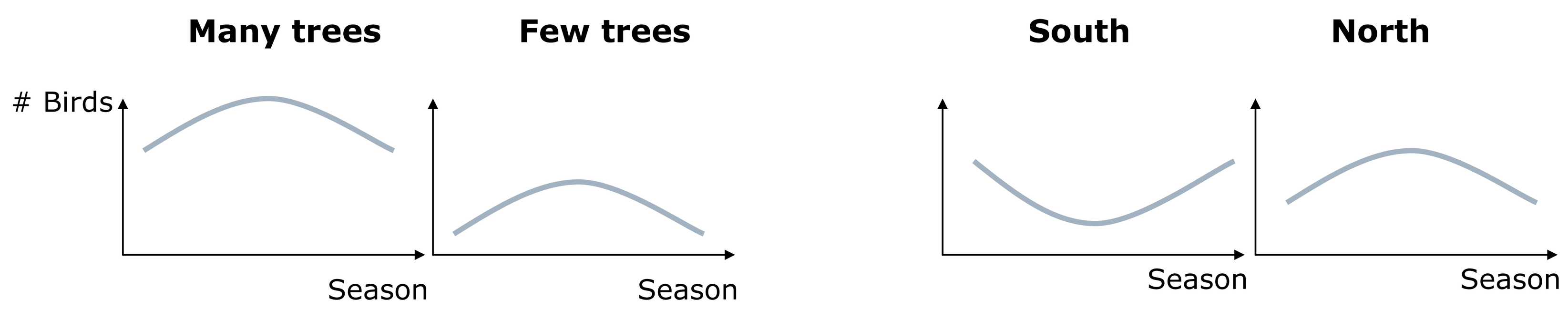



## Promising Data Mining Results

**Detecting interactions between variables**

Statistical interaction = non-additive effects among two or more variables in a function

-F $(x_1,\ldots,x_n)$ shows no interaction between $x_i$ and $x_j$ when
  $F(x_1,x_2,\ldots x_n) = G(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) + H(x_1,\ldots,x_{j-1},x_{j+1},\ldots, x_n)$
  (G does not depend on $x_i$, H does not depend on $x_j$)
-Example: $F(x_1,x_2,x_3) = \sin(x_1+x_2) + x_2x_3$
  -x1, x2 interact; x2, x3 interact; x1, x3 do not interact



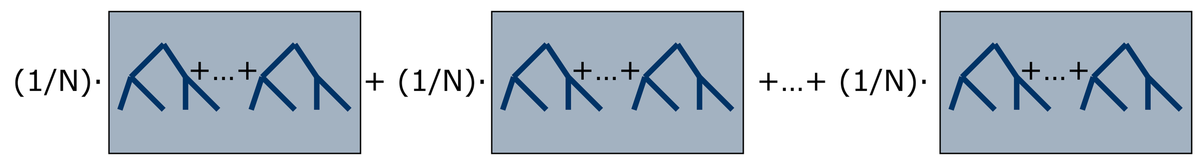

New interaction detection approach:
1. Build a model from the data (no restrictions).
2. Build a restricted model – this time do not allow interaction of interest.
3. Compare their predictive performance.
  -If restricted model as good as unrestricted – no interaction
  -If restricted model significantly worse – interaction

Discovered interaction: (year, latitude) for house finch
-Corresponds to an eye-disease that affected house finches during the decade covered by the dataset

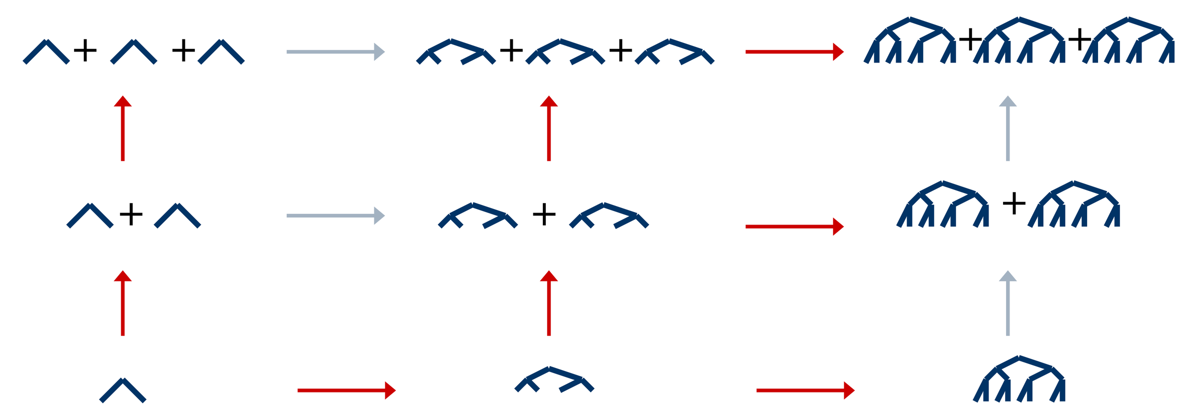
**Novel regression technique—Grove of Trees** [Best student paper at ECML 2007]

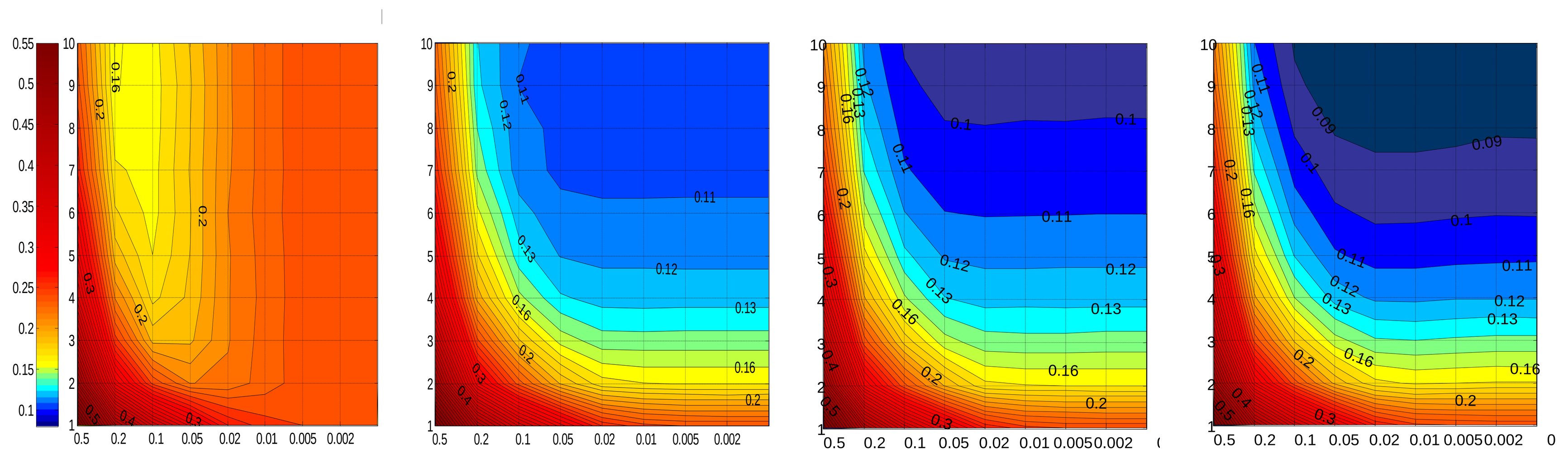
Originally developed as tool for interaction detection procedure
-Ensemble of trees: combines additive models and bagging
  -Main features: Large trees and additive structure
-Outperforms state-of-the-art ensembles like stochastic gradient boosting

Bagged Grove:




Training a Grove:





Bagged Groves trained as classical additive models | Layered training | Dynamic programming | Randomized dynamic programming

## Next Steps

-Combining statistical techniques with data mining to model domain knowledge
-Efficient search for interesting patterns among trend plots
-Detecting multi-way interactions
-Standard observation protocol mapping to combine data from different sources

This work has been supported by the National Science Foundation through grants EF-0427914, IIS-0612031, and IIS-0748626 and by a grant from the Leon Levy Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.