Graphical Models and Flexible Classifiers: Bridging the Gap with Boosted Regression Trees

Thomas G. Dietterich

with Adam Ashenfelter, Guohua Hao, Rebecca Hutchinson, Liping Liu, and Dan Sheldon

Oregon State University
Corvallis, Oregon, USA

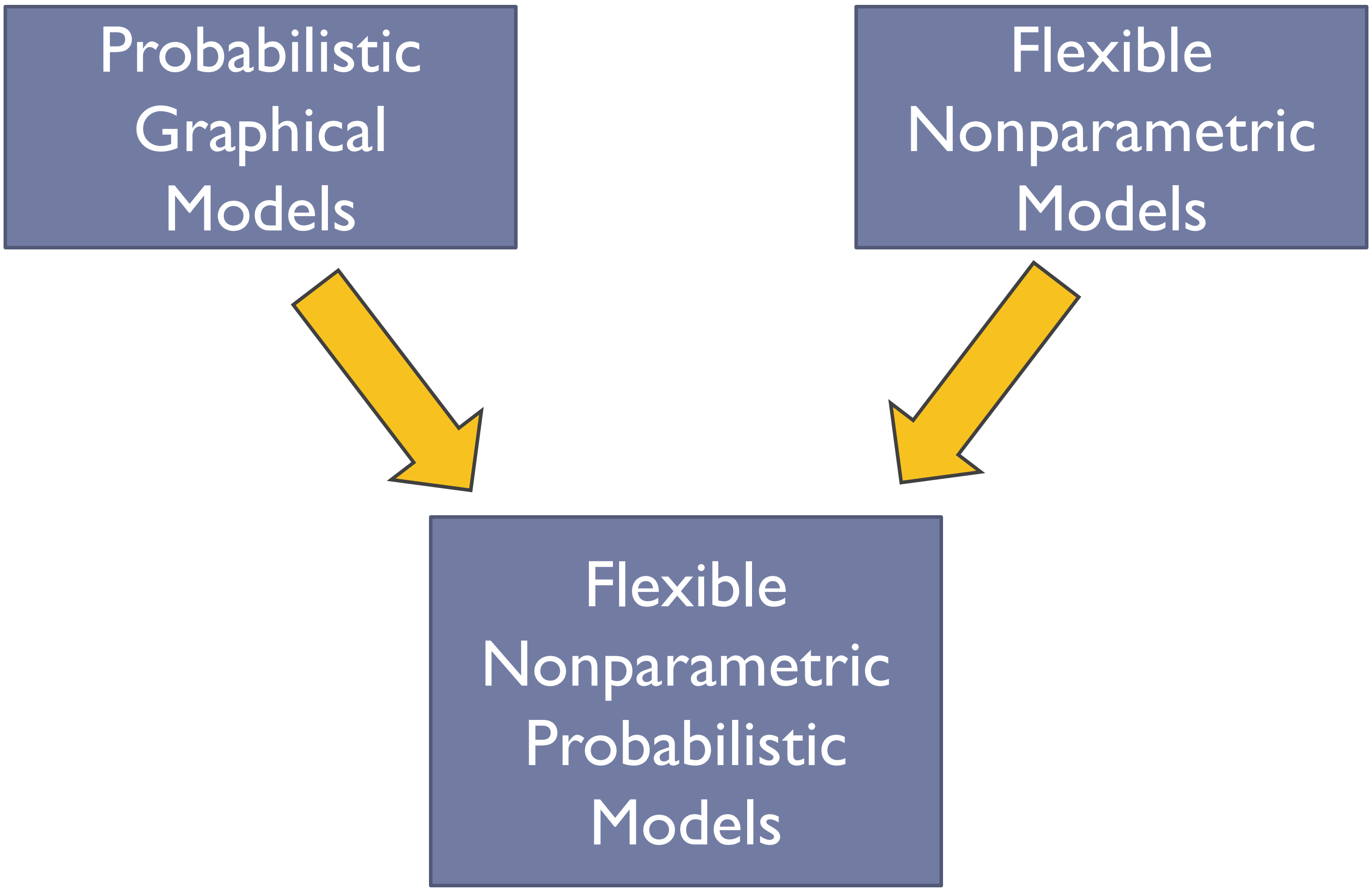
The Distinguished Speakers Program
is made possible by

Association for
Computing Machinery

Advancing Computing as a Science & Profession

For additional information, please visit http://dsp.acm.org/

Combining Two Approaches to Machine Learning

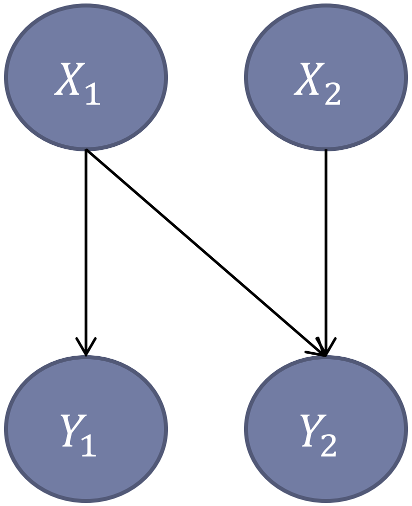


Outline

- Two Cultures of Machine Learning
  - Probabilistic Graphical Models
  - Non-Parametric Discriminative Models
  - Advantages and Disadvantages of Each
- Representing conditional probability distributions using non-parametric machine learning methods
  - Logistic regression (Friedman)
  - Conditional random fields (Dietterich, et al.)
  - Latent variable models (Hutchinson, et al.)
- Ongoing Work
- Conclusions

Probabilistic Graphical Models

- Nodes: Random variables
  - X_1, X_2, Y_1, Y_2
- Edges: Direct probabilistic dependencies
  - $P(Y_1|X_1), P(Y_2|X_1, X_2)$



- Joint probability distribution is the product of the individual node distributions
  - $P(X_1, X_2, Y_1, Y_2) = P(X_1)P(X_2)P(Y_1|X_1)P(Y_2|X_1, X_2)$

Probabilistic Graphical Models (2)

- Can be learned from training data, even when some of the random variables are unobserved (latent or missing)
  - Mixture models (e.g., Gaussian mixture models)
  - Train with EM, gradient descent, or MCMC
- Can represent dynamical processes (Markov models, Dynamic Bayesian Networks)
- Provide probabilistic predictions
  - Useful for integrating into larger systems
- Provide a powerful language for designing and expressing models of complex systems
  - Useful for capturing background knowledge

Probabilistic Graphical Models (3)

- How should the conditional probability distributions be represented?
  - Conditional Probability Tables (CPTs) with one parameter for each combination of values
  - $\frac{2^N}{2}$ parameters

X_1	X_2	Y_1	$P(Y_1 \mid X_1, X_2)$
0	0	0	$1-\alpha$
0	0	1	α
0	1	0	$1-\beta$
0	1	1	β
1	0	0	$1-\gamma$
1	0	1	γ
1	1	0	$1-\delta$
1	1	1	δ

Probabilistic Graphical Models (3)

- How should the conditional probability distributions be represented?
  - Log-linear models (logistic regression)

$$\log \frac{P(Y_1 = 1|X_1, X_2)}{P(Y_1 = 0|X_1, X_2)} =$$

$$\alpha' + I[X_1 = 1]\beta' + I[X_2 = 1]\gamma'$$

  - $\operatorname{expit} u = 1/(1 + \exp(-u))$
  - N parameters

X_1	X_2	Y_1	$P(Y_1\|X_1, X_2)$
0	0	0	$1 - \operatorname{expit} \alpha'$
0	0	1	$\operatorname{expit} \alpha'$
0	1	0	$1 - \operatorname{expit}(\alpha' + \gamma')$
0	1	1	$\operatorname{expit}(\alpha' + \gamma')$
1	0	0	$1 - \operatorname{expit}(\alpha' + \beta')$
1	0	1	$\operatorname{expit}(\alpha' + \beta')$
1	1	0	$1 - \operatorname{expit}(\alpha' + \beta' + \gamma')$
1	1	1	$\operatorname{expit} \alpha' + \beta' + \gamma'$

Advantages and Disadvantages of Parametric Representations

Advantages

- Each parameter has a meaning
- Supports statistical hypothesis testing: "Does X_1 influence Y_1?"
  - $H_0: \beta' = 0$
  - $H_a: \beta' \neq 0$

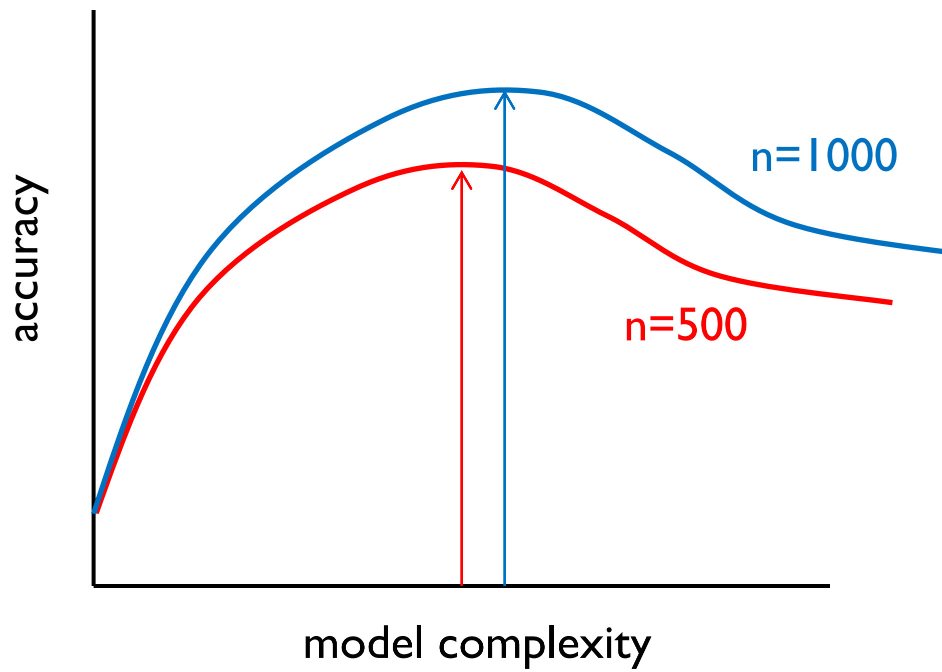
Disadvantages

- Model has fixed complexity
  - Will typically either under-fit or over-fit the data
- Designer must decide about interactions, non-linearities, etc. etc.
  - Wrong decisions lead to highly biased models and invalidate hypothesis tests
  - Correlated variables cause trouble
  - Difficult for problems with many features
- Data must be transformed to match the parametric form
  - Discretized
  - Square root or log transforms

Fundamental Theorem of Statistical Learning

- Three-way tradeoff
  - amount of data
  - complexity of the model
  - prediction accuracy

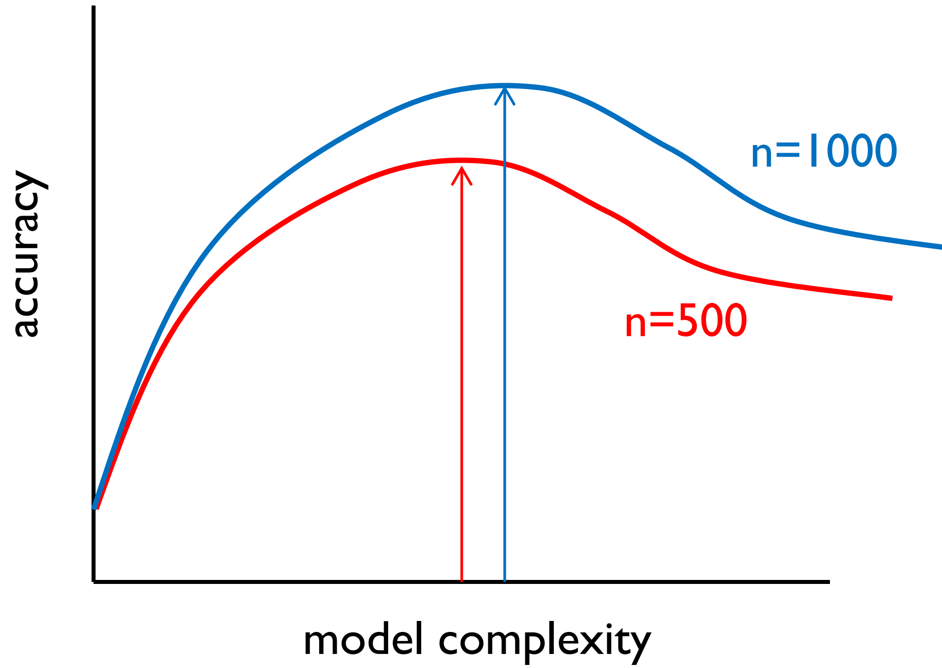
- To achieve optimum accuracy, model complexity should be tuned to the amount of data
  - "Structural Risk Minimization" (Vapnik)



Flexible Machine Learning Models

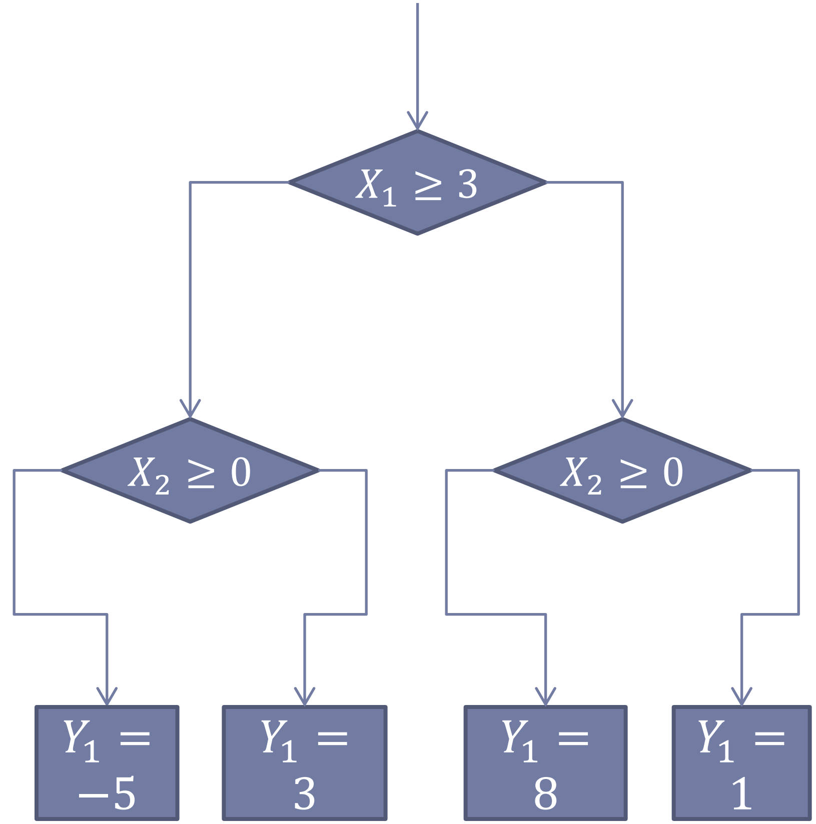
- Support Vector Machines
- Classification and Regression Trees

- Key advantage: Can tune the complexity of the model to the complexity of the data



Another Advantage: Interactions and Nonlinearities

- SVMs:
  - Polynomial kernels capture interactions and polynomial nonlinearities
  - Gaussian kernels capture nonlinearities, however, interactions are embedded in the distance function (typically Euclidean)
- Classification and regression trees
  - Interactions are captured by the if-then-else structure of the tree
  - Nonlinearities are approximated by piecewise constant functions



$$Y_1 = -5 \cdot I(X_1 \geq 3, X_2 \geq 0) + 3 \cdot I(X_1 \geq 3, X_2 < 0) + \\ 8 \cdot I(x_1 < 3, X_2 \geq 0) + 1 \cdot I(X_1 < 3, X_2 < 0)$$

Tree-Based Methods

Advantages

- Flexible Model Complexity
  - Controlled by depth of tree
- Can handle discrete, ordered, and continuous variables
  - No normalization or rescaling needed
- Can handle missing values
  - Proportional distribution
  - Surrogate splits
- Best "off the shelf" method (Breiman)

Disadvantages

- Poor probability estimates
- Do not support hypothesis testing
- Cannot handle latent variables
- High variance, which can be addressed by
  - Boosting
  - Bagging
  - Randomization

Can we combine the best of both?

Probabilistic Graphical Models

- Probabilistic semantics
- Structured by background knowledge
- Latent variables and dynamic processes

Non-Parametric Tree Methods

- Tunable model complexity
- No need for data scaling and preprocessing
  - Discrete, ordered, or continuous values

Existing Efforts

- Dependency Networks
  - Heckerman et al. (JMLR 2000):
    - Bayesian network where each $P(X|Y)$ is a decision tree (with multinomial output probabilities)
  - Trained to maximize pseudo-likelihood
  - Requires all variables to be observed
- RKHS embeddings of probabilities distributions
  - Song, Gretton & Guestrin (AISTATS 2011)
    - Tree-structured graphical model (undirected)
    - No explicit latent variables
- Bayesian semi-parametric methods
  - Dirichlet processes (Blei, Jordan, et al.)

Outline

- Two Cultures of Machine Learning
  - Probabilistic Graphical Models
  - Non-Parametric Discriminative Models
  - Advantages and Disadvantages of Each
- Representing conditional probability distributions using non-parametric machine learning methods
  - Logistic regression (Friedman)
  - Conditional random fields (Dietterich, et al.)
  - Latent variable models (Hutchinson, et al.)
- Ongoing Work
- Conclusions

Representing $P(Y|X)$ using boosted regression trees

- Friedman: Gradient Tree Boosting (2000; Annals of Statistics, 2011)
- Consider logistic regression:
  - $\log \frac{P(Y=1)}{P(Y=0)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_J X_J$
  - $D = \{(X^i, Y^i)\}_{i=1}^N$ are the training examples
  - Log likelihood:
    - $LL(\beta) = \sum_i Y^i \log P(Y = 1|X^i;\beta) + (1 - Y^i) \log P(Y = 0|X^i;\beta)$

Fitting logistic regression via gradient descent

- Let $\beta^0 = g^0 = \mathbf{0}$
- For $\ell = 1, \dots, L$ do
  - Compute $g^\ell = \nabla_\beta LL(\beta)\big|_{\beta=\beta^{\ell-1}}$
    - g^ℓ = gradient w.r.t. β
  - $\beta^\ell := \beta^{\ell-1} + \eta_\ell g^\ell$ take a step of size η_ℓ in direction of gradient
- Final estimate of β is
  - $\beta^L = g^0 + \eta_1 g^1 + \cdots + \eta_L g^L$

Functional Gradient Descent Boosted Regression Trees

- Breiman (1996), Friedman (2000), Mason et al. (NIPS 1999): Fit a logistic regression model as a weighted sum of regression trees:

$$\log \frac{P(Y=1)}{P(Y=0)} = tree^0(X) + \eta_1 tree^1(X) + \cdots + \eta_L tree^L(X)$$

- When "flattened" this gives a log linear model with complex interaction terms

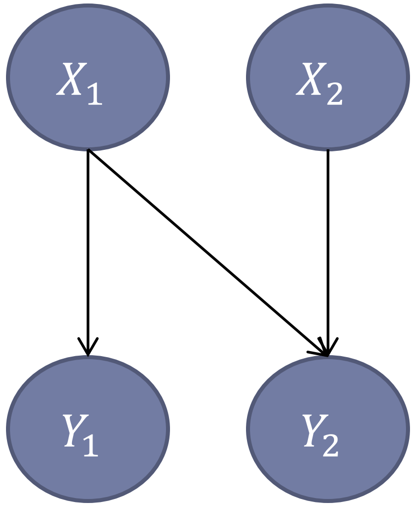
L2-Tree Boosting Algorithm

- Let $F^0(X) = f^0(X) = \mathbf{0}$ be the zero function
- For $\ell = 1, \ldots, L$ do
  - Construct a training set $S^\ell = \{(X^i, \tilde{Y}^i)\}_{i=1}^N$
    - where $\tilde{Y}$ is computed as
    - $\tilde{Y}^i = \left.\frac{\partial LL(F)}{\partial F}\right|_{F=F^{\ell-1}(X^i)}$ // how we wish F would change at X^i
  - Let f^ℓ = regression tree fit to S^ℓ
  - $F^\ell := F^{\ell-1} + \eta_\ell f^\ell$
- The step sizes η_ℓ are the weights computed in boosting
- This provides a general recipe for learning a conditional probability distribution for a Bernoulli or multinomial random variable

L2-TreeBoosting can be applied to any fully-observed directed graphical model

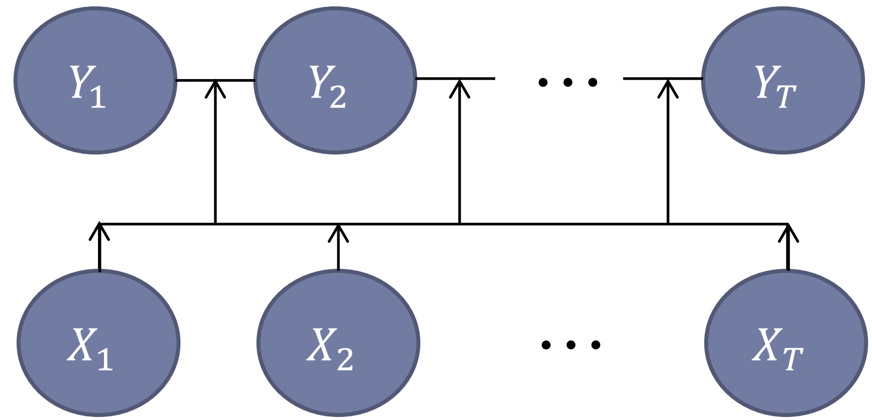
- $P(Y_1|X_1)$ as sum of trees
- $P(Y_2|X_1, X_2)$ as sum of trees

- What about undirected graphical models?



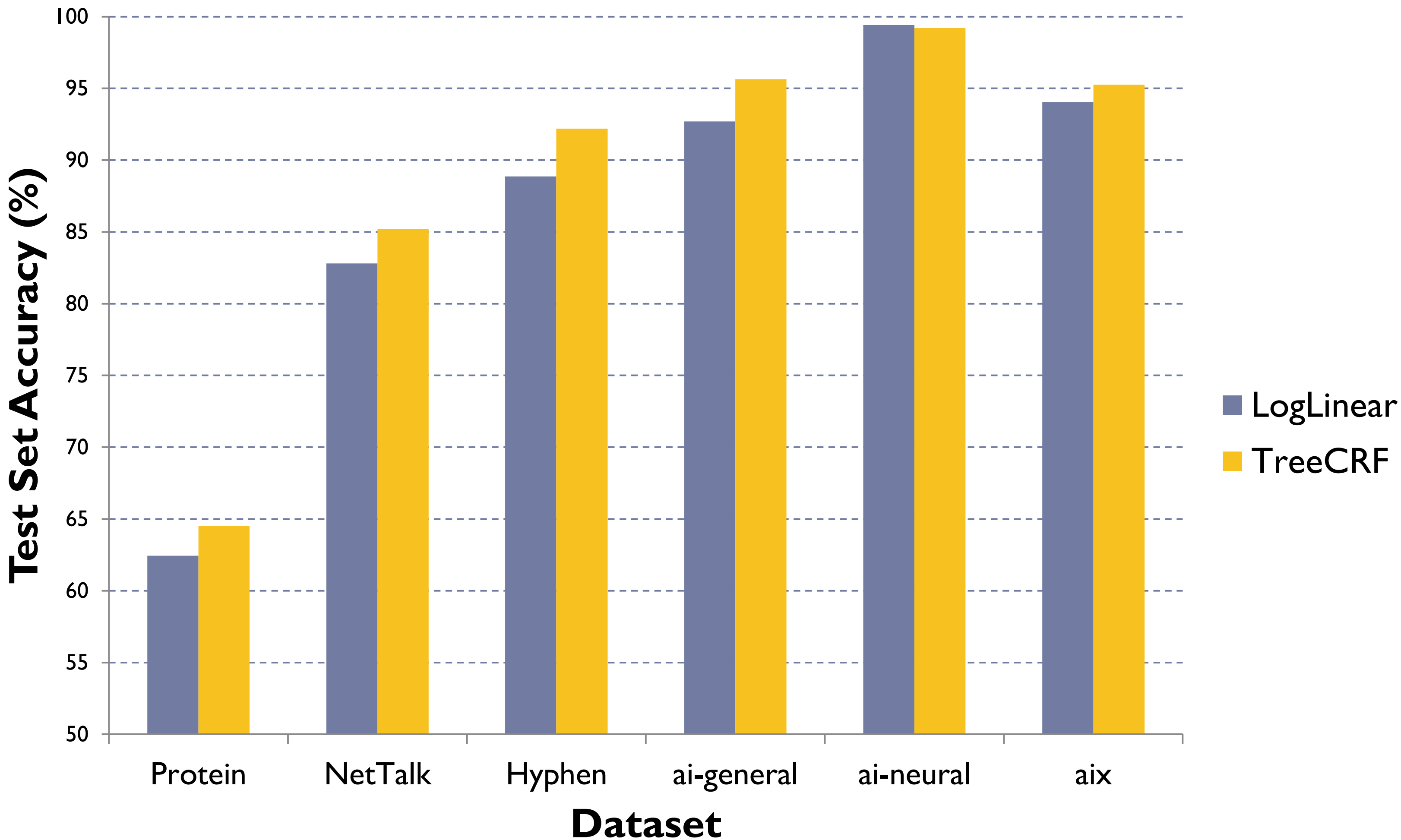
Tree Boosting for Conditional Random Fields

- Conditional Random Field (Lafferty et al., 2001)
  - $P(Y_1, \dots, Y_T | X_1, \dots, X_T)$
  - Undirected graph over the Y's conditioned on the X's.
  - $\Phi(Y_{t-1}, Y_t, X)$ = log linear model



- Dietterich, Hao, Ashenfelter (JMLR 2008; ICML 2004)
  - Fit $\Phi(Y_{t-1}, Y_t, X)$ using tree boosting
  - A form of automatic feature discovery for CRFs

Experimental Results



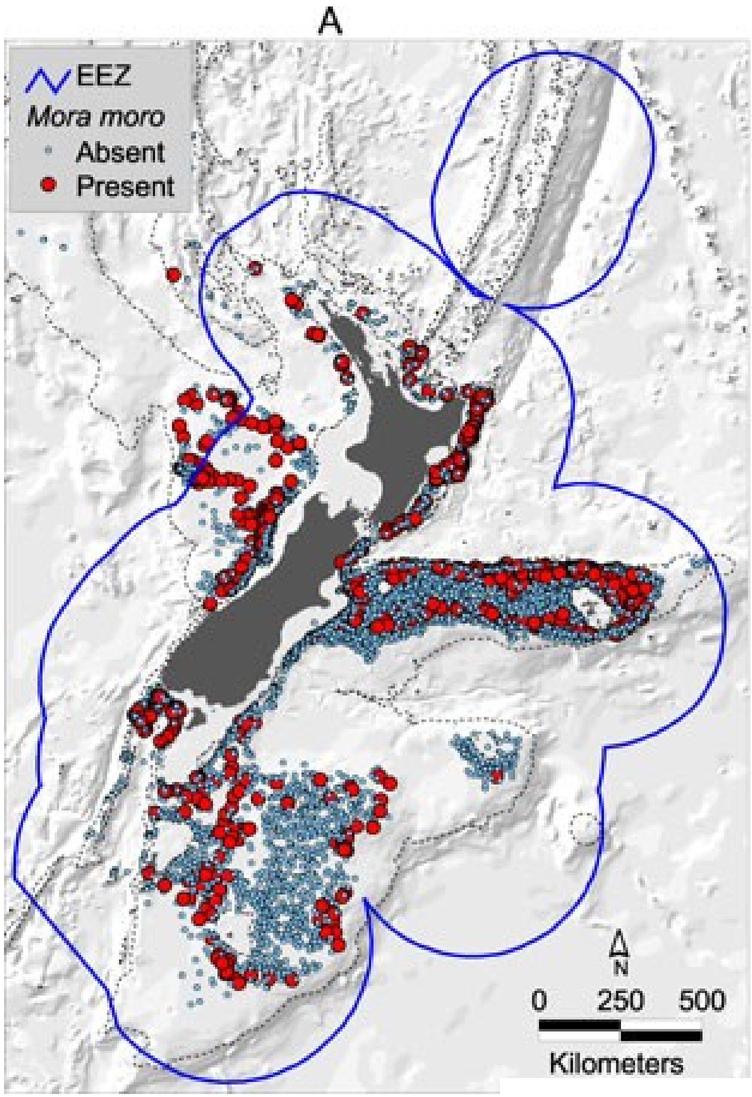
All differences statistically significant $p<0.005$ or better

Tree Boosting for Latent Variable Models

- Both Friedman's L2-TreeBoosted logistic regression and our L2-TreeBoosted CRFs assumed that all variables were observed in the training data
- Can we extend Tree Boosting to <u>latent variable</u> graphical models?
- Motivating application: Species Distribution Modeling

Species Distribution Modeling

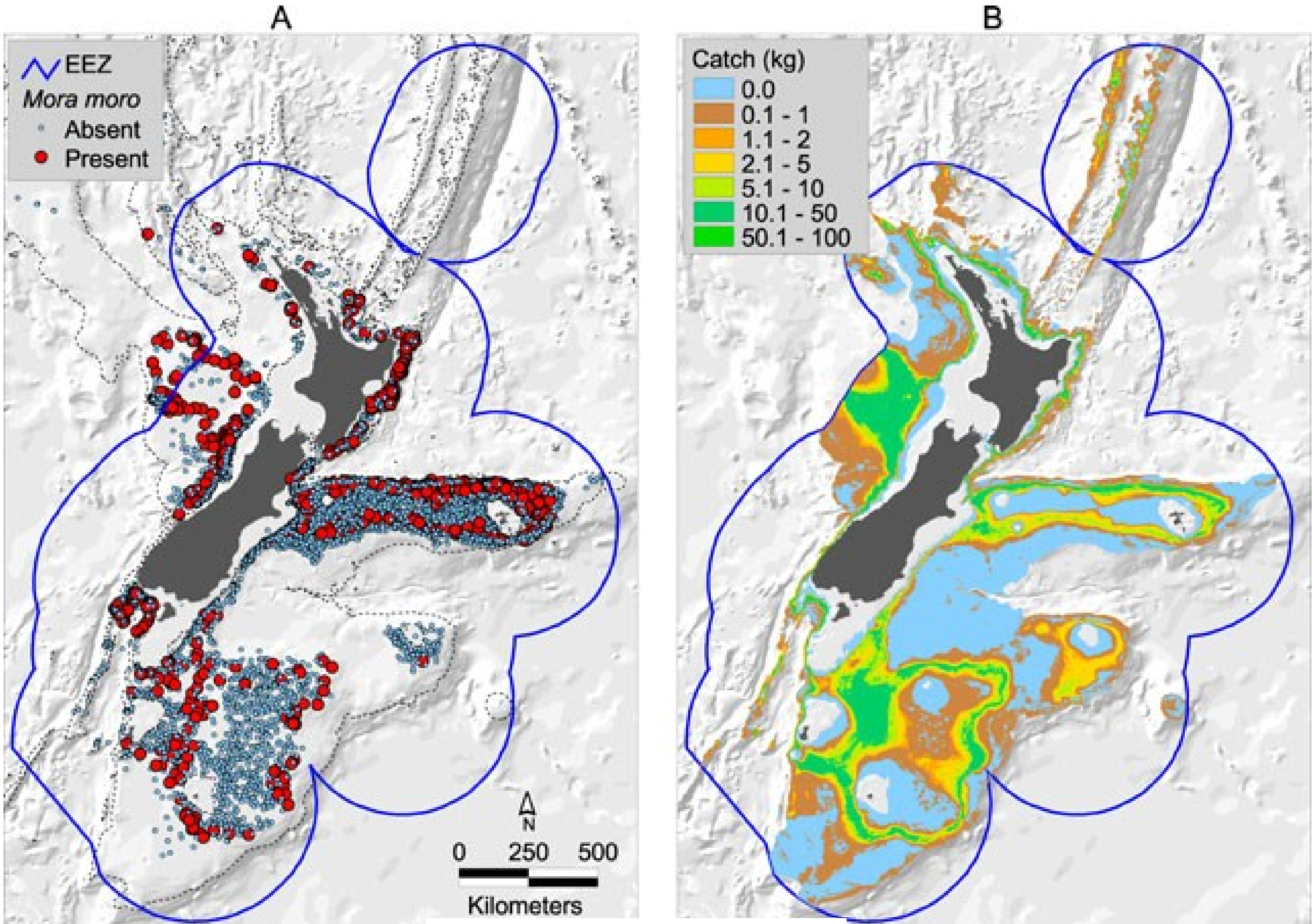
Observations



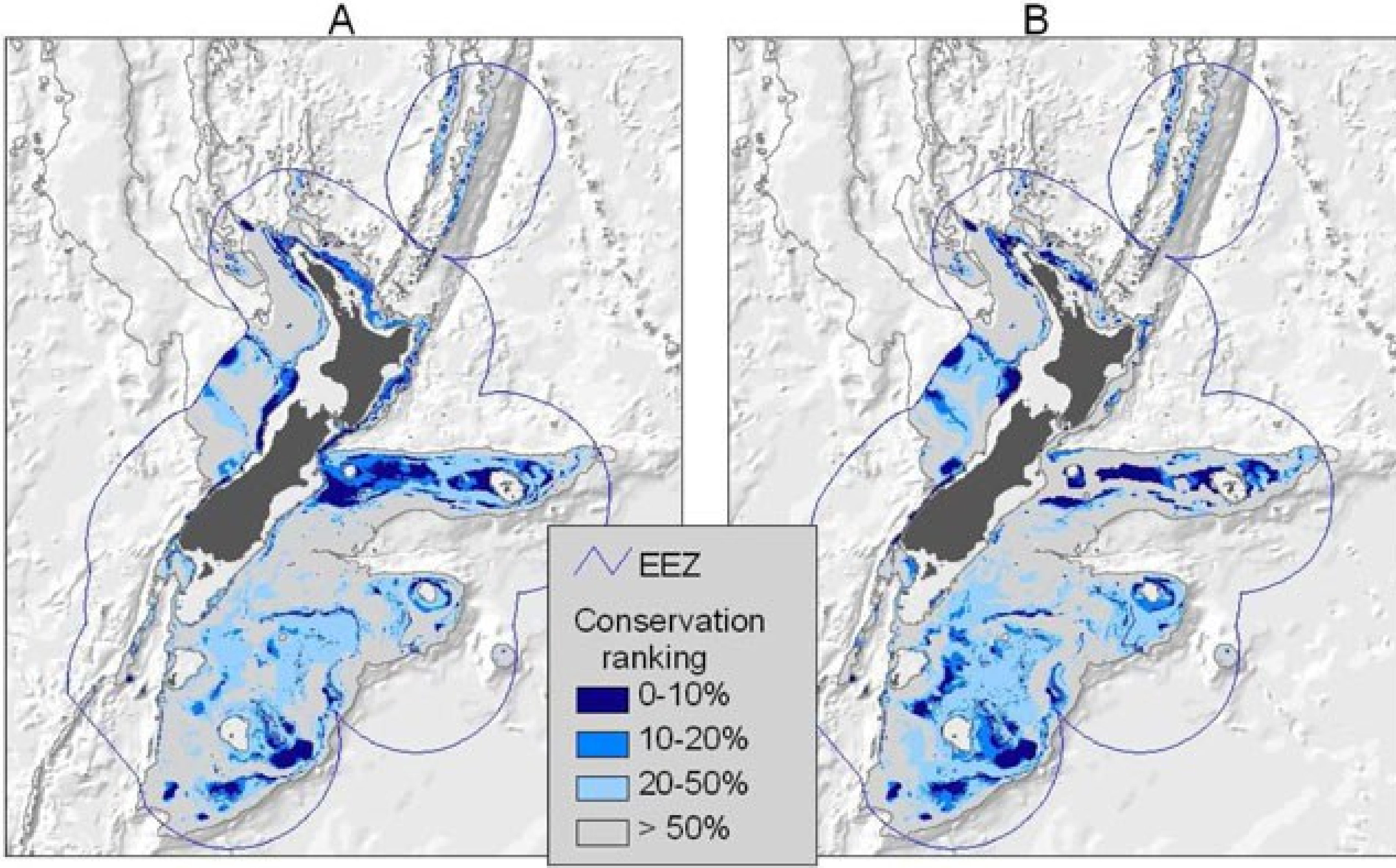
Leathwick et al, 2008

Species Distribution Modeling

Observations | Fitted Model



Leathwick et al, 2008



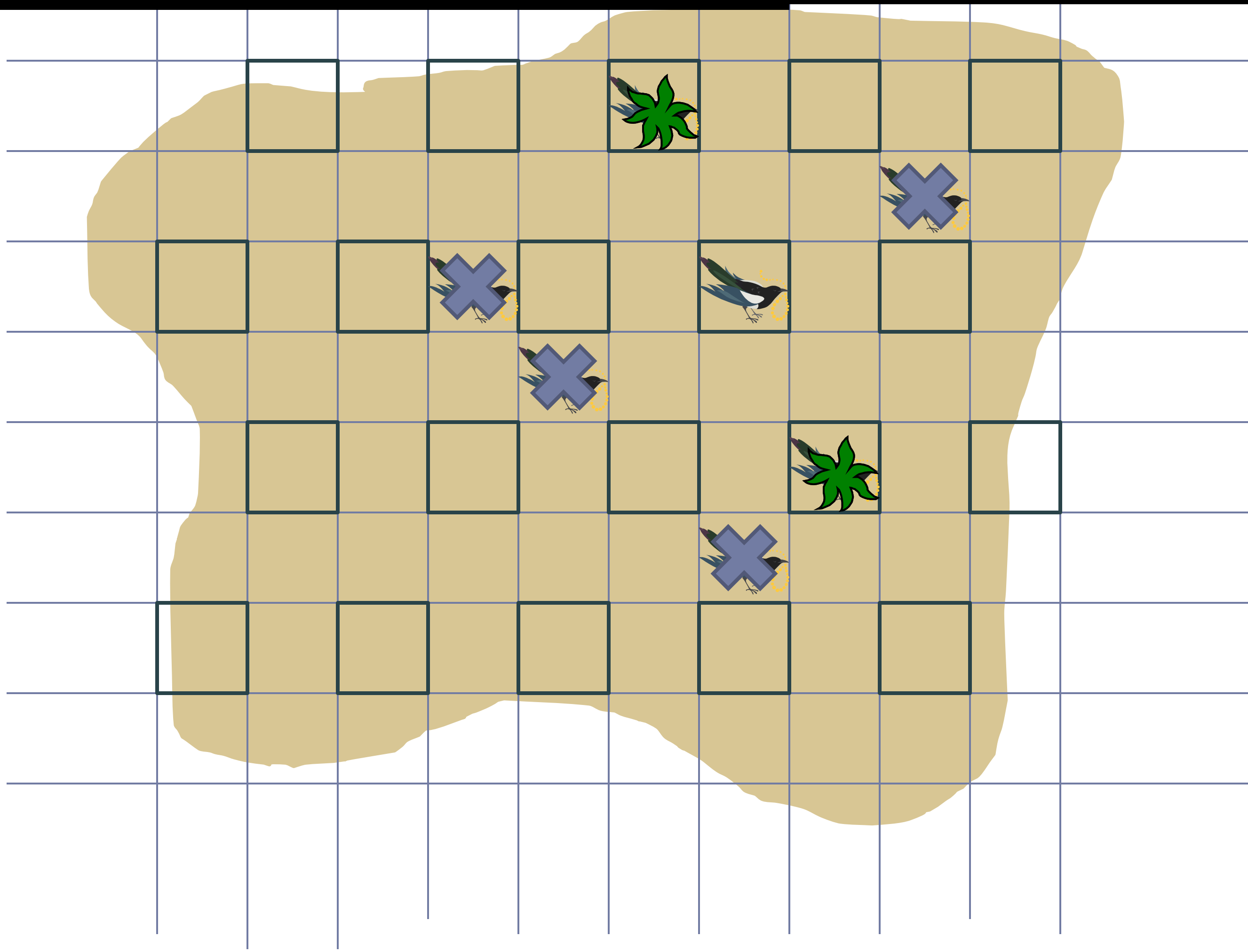
Disregarding costs to fishing industry

Full consideration of costs to fishing industry

Leathwick et al, 2008

Wildlife Surveys with Imperfect Detection

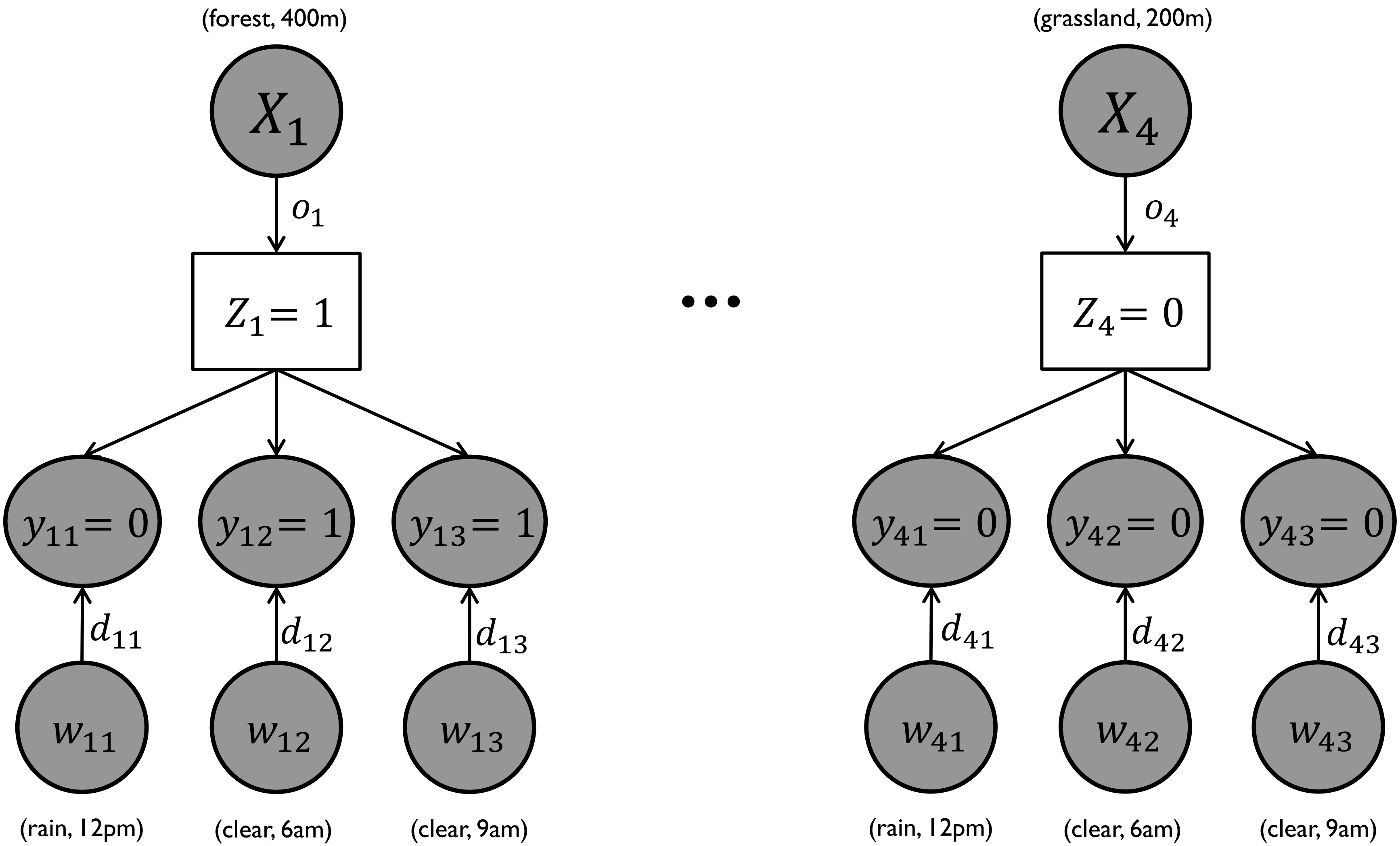
Partial Solution: Multiple visits: Different birds hide on different visits



Multiple Visit Data

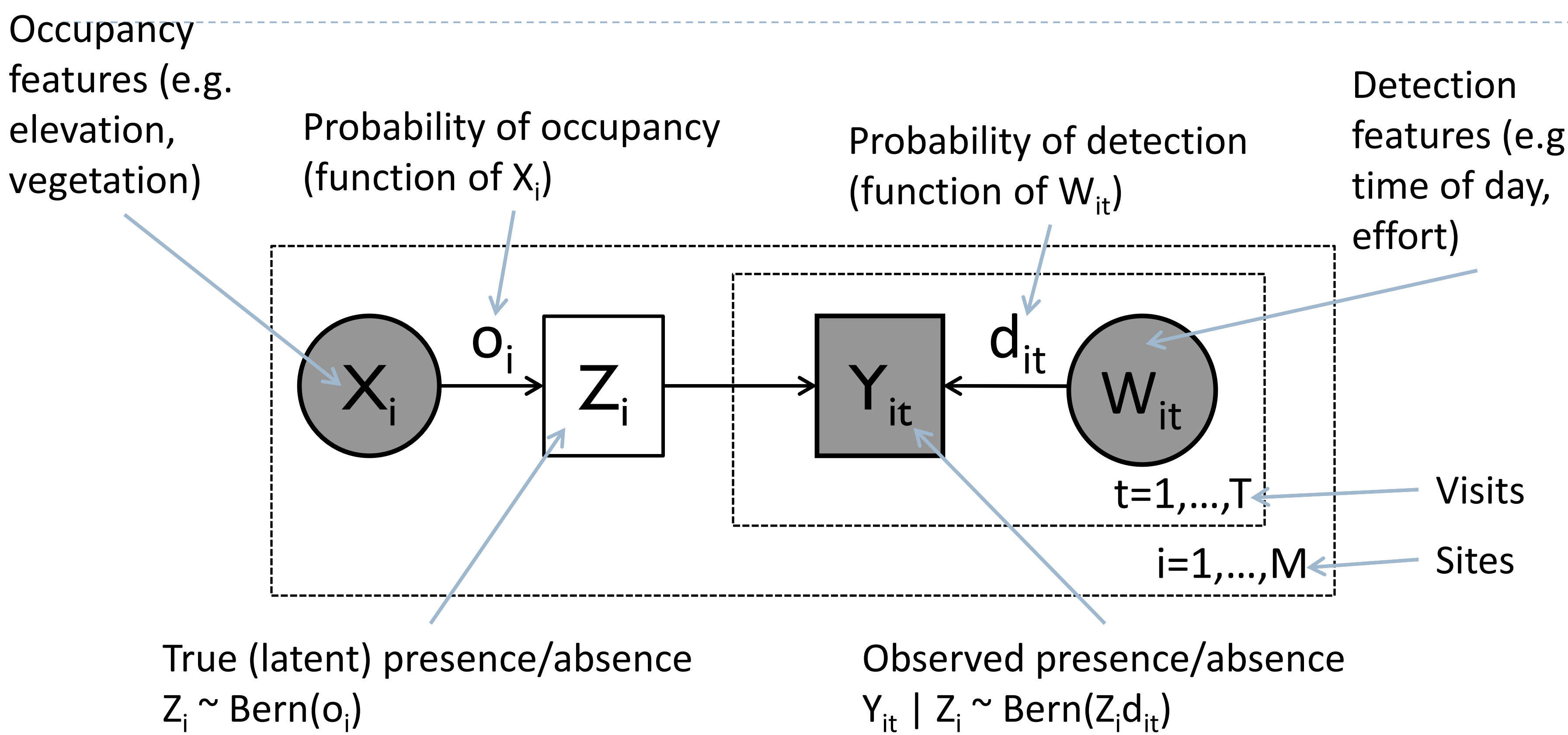
		Detection History		
Site	*True occupancy (latent)*	Visit 1 (rainy day, 12pm)	Visit 2 (clear day, 6am)	Visit 3 (clear day, 9am)
A (forest, elev=400m)	*1*	0	1	1
B (forest, elev=500m)	*1*	0	1	0
C (forest, elev=300m)	*1*	0	0	0
D (grassland, elev=200m)	*0*	0	0	0

Probabilistic Model with Latent Variable Z



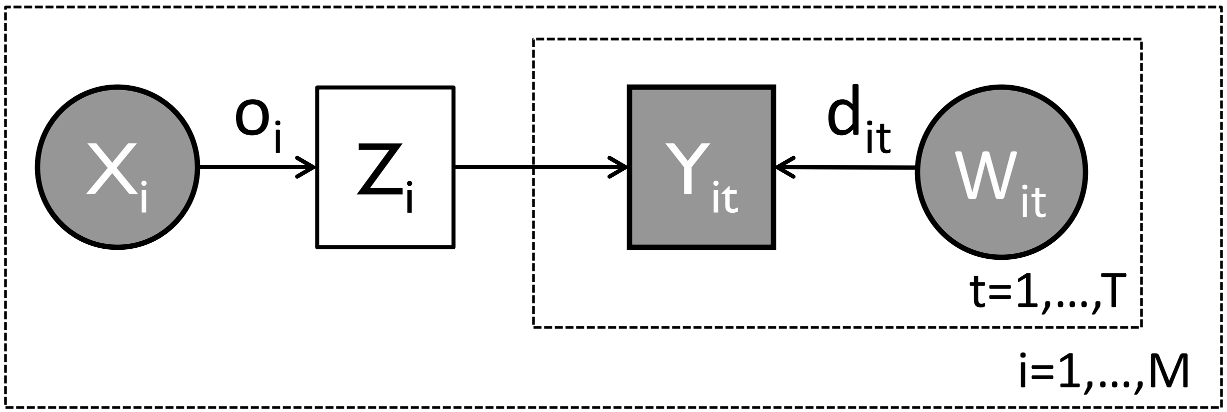
MacKenzie, et al, 2006

Occupancy-Detection Model



MacKenzie, et al, 2006

Parameterizing the model



$z_i \sim P(z_i|x_i)$: 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

- $\log \frac{F(X)}{1-F(X)} = \beta_0 + \beta_1 X^1 + \cdots + \beta_J X^J$
- $\log \frac{G(W)}{1-G(W)} = \alpha_0 + \alpha_1 W^1 + \cdots + \alpha_K W^K$
- Train via EM
- People tend to use very simple models: $J = 4, K = 2$

Regression Tree Parameterization

- $\log \frac{F(x)}{1-F(x)} = f^0(x) + \rho_1 f^1(x) + \cdots + \rho_L f^L(x)$
- $\log \frac{G(w)}{1-G(w)} = g^0(w) + \eta_1 g^1(w) + \cdots + \eta_L g^L(w)$
- Perform functional gradient descent on F and G
- Could also use EM

Functional Gradient Descent with Latent Variables

- Loss function $L(F, G, y)$
- $F^0 = G^0 = f^0 = g^0 = 0$
- For $\ell = 1, \dots, L$
  - For each site i compute
    $$\tilde{z}_i = \partial L(F^{\ell-1}(x_i), G^{\ell-1}, y_i) / \partial F^{\ell-1}(x_i)$$
  - Fit regression tree f^ℓ to $\{\langle x_i, \tilde{z}_i \rangle\}_{i=1}^{M}$
  - For each visit t to site i, compute
    $$\tilde{y}_{it} = \partial L\left(F^{\ell-1}(x_i), G^{\ell-1}(w_{it}), y_{it}\right) / \partial G^{\ell-1}(w_{it})$$
  - Fit regression tree g^ℓ to $\{\langle w_{it}, \tilde{y}_{it} \rangle\}_{i=1,t=1}^{M,T_i}$
  - Let $F^\ell = F^{\ell-1} + \rho_\ell f^\ell$
  - Let $G^\ell = G^{\ell-1} + \eta_\ell g^\ell$

Experiment

- Algorithms:
  - Supervised methods:
    - S-LR: logistic regression from $(x_i, w_{it}) \rightarrow y_{it}$
    - S-BRT: boosted regression trees $(x_i, w_{it}) \rightarrow y_{it}$
  - Occupancy-Detection methods:
    - OD-LR: F and G logistic regressions
    - OD-BRT: F and G boosted regression trees
- Data:
  - 12 bird species
  - 3 synthetic species
  - 3124 observations from New York State, May-July 2006-2008
  - All features rescaled to zero mean, unit variance

Features

$X^{(1)}$	Human population per sq. mile
$X^{(2)}$	Number of housing units per sq. mile
$X^{(3)}$	Percentage of housing units vacant
$X^{(4)}$	Elevation
$X^{(5)} \ldots X^{(19)}$	Percent of surrounding 22,500 hectares in each of 15 habitat classes from the National Land Cover Dataset
$W^{(1)}$	Time of day
$W^{(2)}$	Observation duration
$W^{(3)}$	Distance traveled during observation
$W^{(4)}$	Day of year

Simulation Study using Synthetic Species

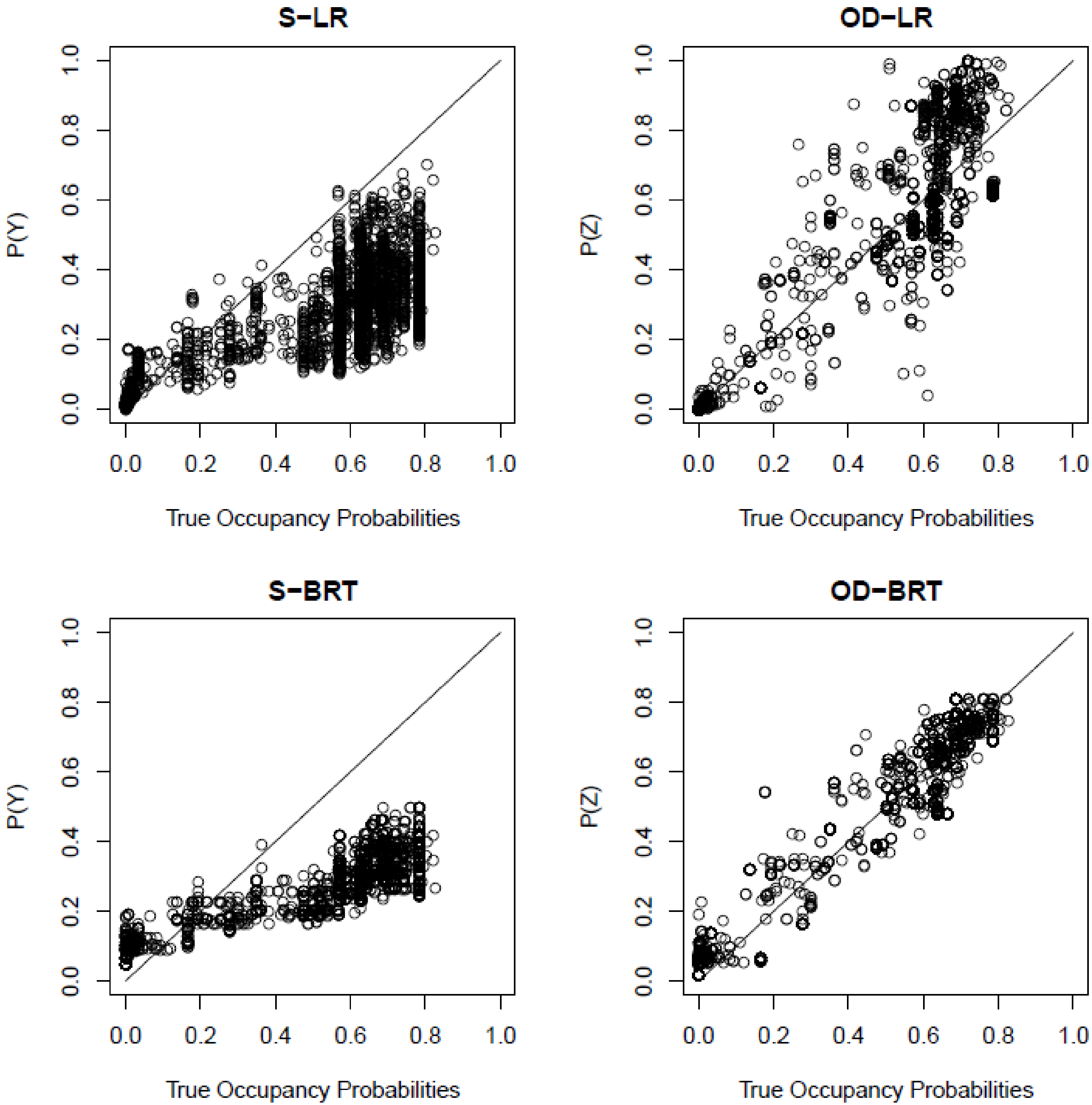
- Synthetic Species 2: F and G nonlinear

$$\log\frac{o_i}{1-o_i} = -2\left[x_i^{(1)}\right]^2 + 3\left[x_i^{(2)}\right]^2 - 2x_i^{(3)}$$

$$\log\frac{d_{it}}{1-d_{it}} = \exp(-0.5w_{it}^{(4)}) + \sin(1.25w_{it}^{(1)} + 5)$$

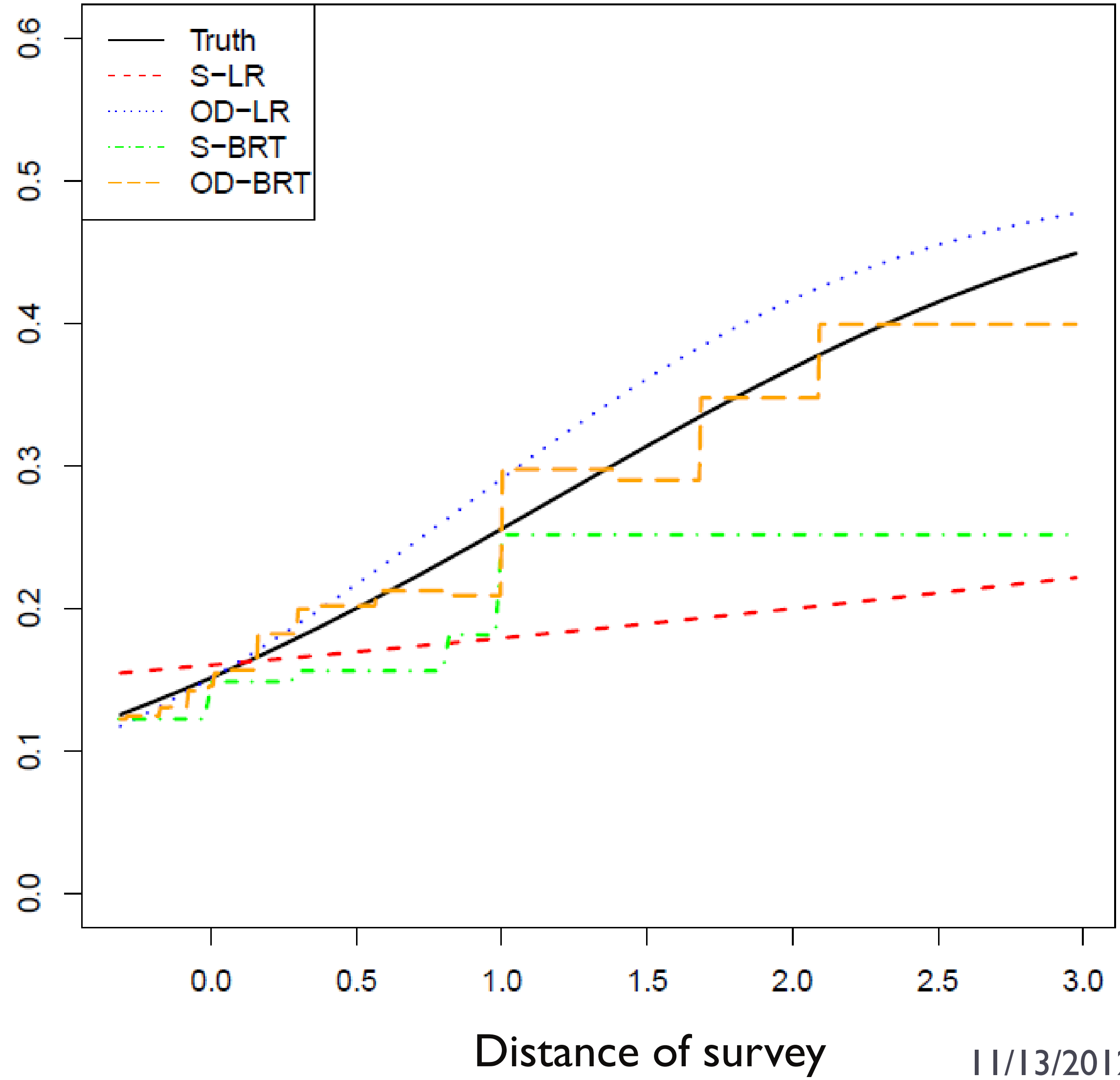
Predicting Occupancy

Synthetic Species 2



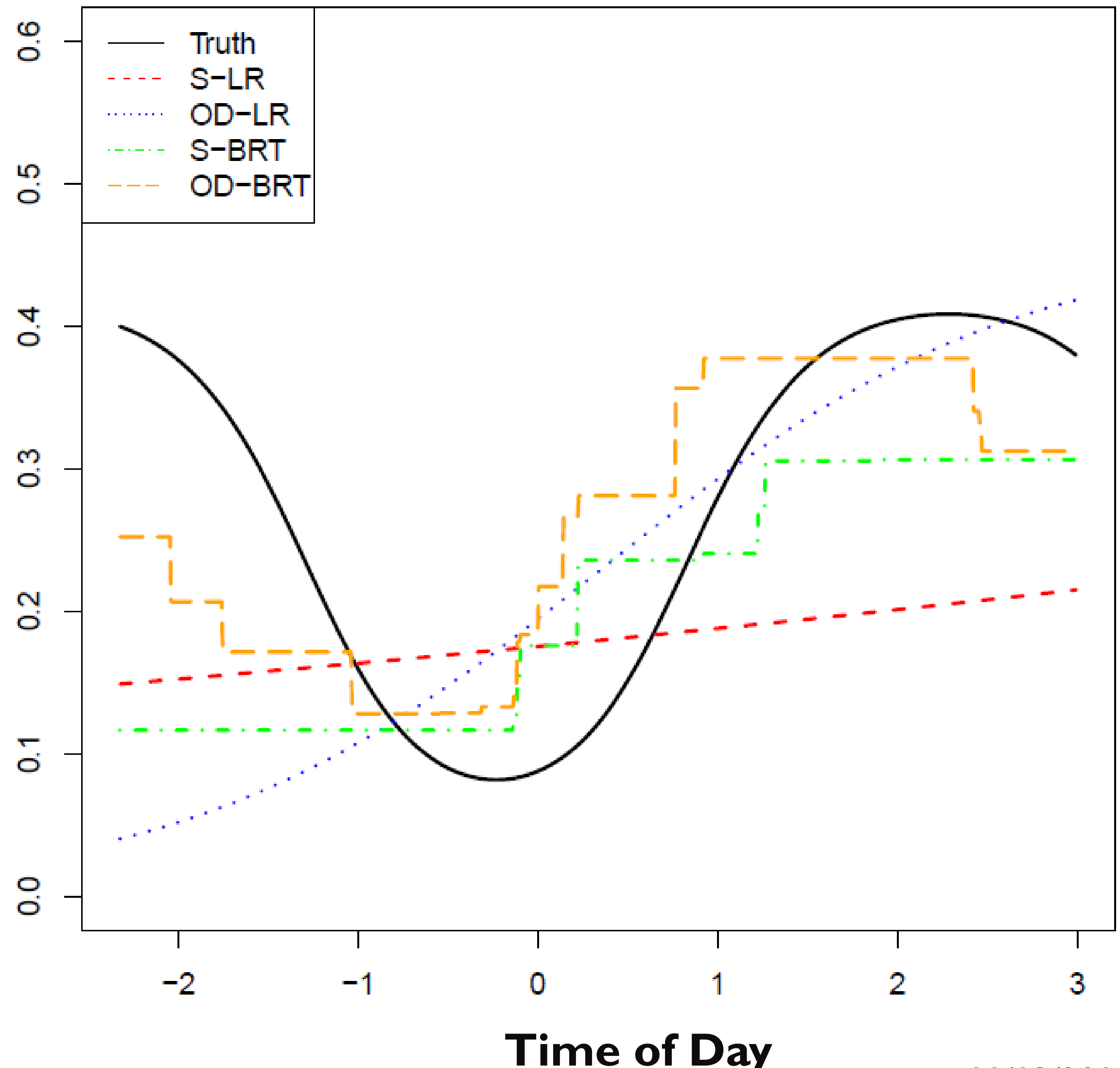
Partial Dependence Plot
Synthetic Species 1

- OD-BRT has the least bias

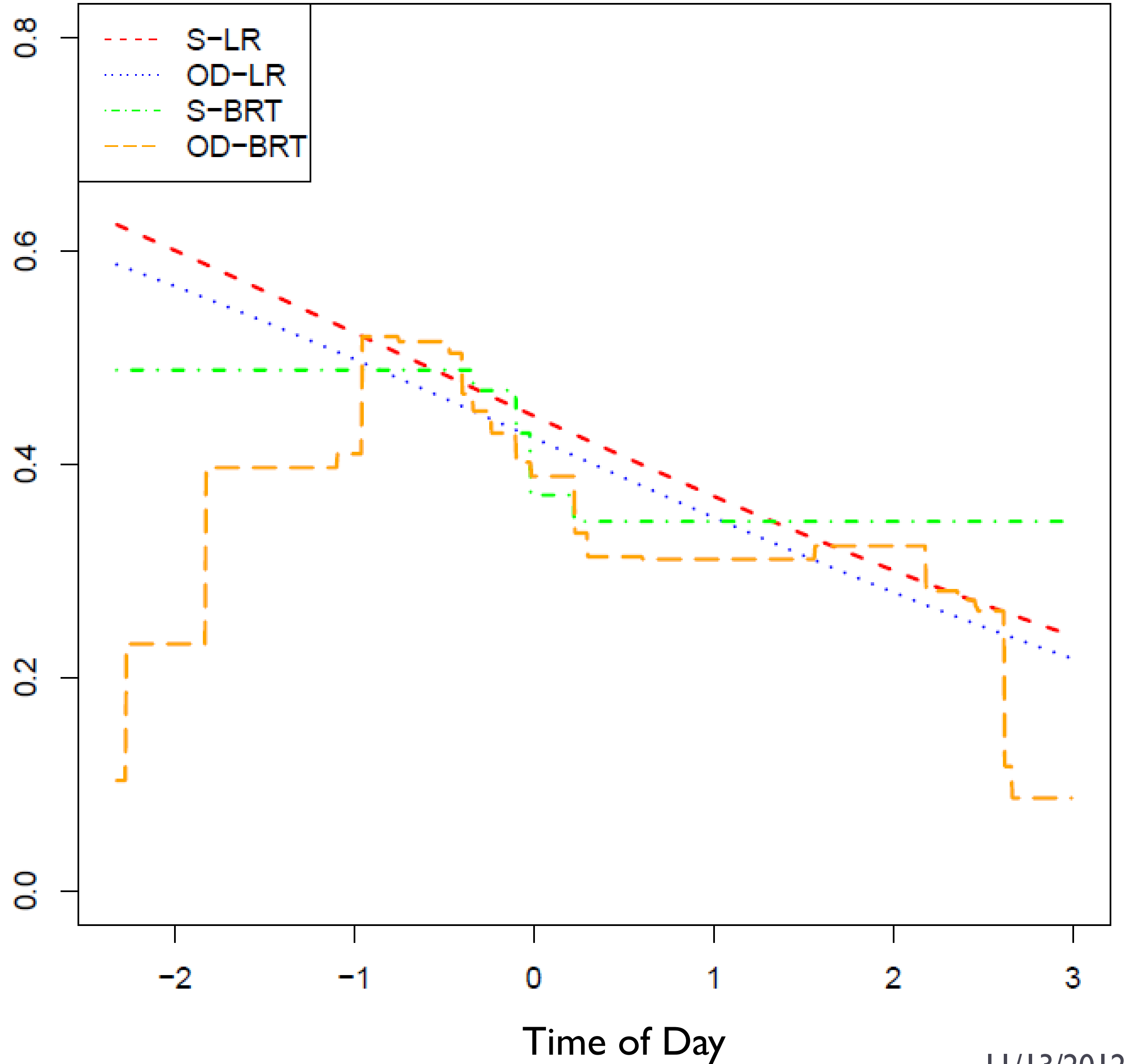


Partial Dependence Plot
Synthetic Species 3

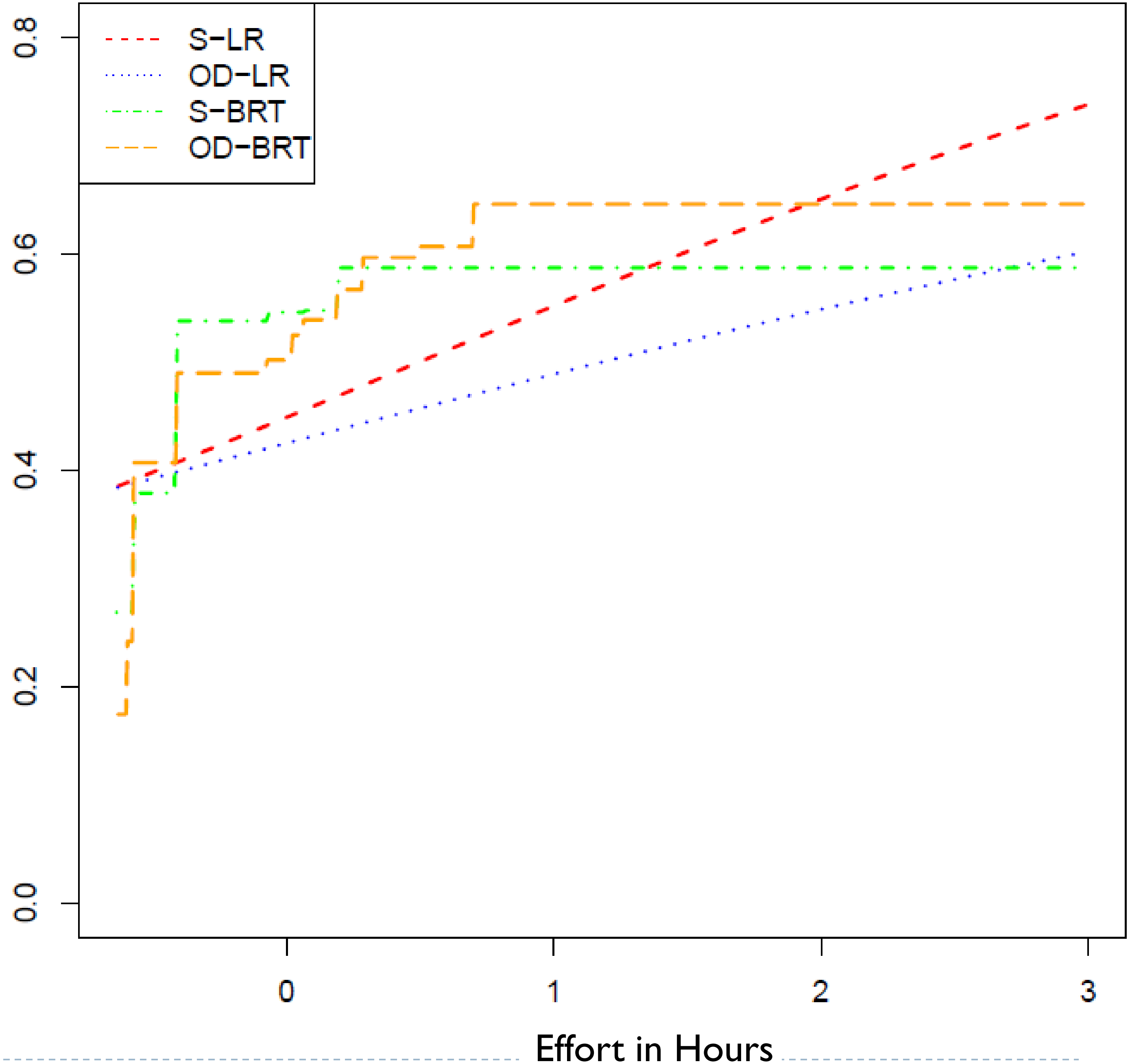
- OD-BRT has the least bias and correctly captures the bi-modal detection probability



Partial Dependence Plot Blue Jay vs. Time of Day



Partial Dependence Plot Blue Jay vs. Duration of Observation



Open Problems

- Sometimes the OD model finds trivial solutions
  - Detection probability = 0 at many sites, which allows the Occupancy model complete freedom at those sites
  - Occupancy probability constant (0.2)
- Log likelihood for latent variable models suffers from local minima
  - Proper initialization?
  - Proper regularization?
  - Posterior regularization?
- How much data do we need to fit this model?
  - Can we detect when the model has failed?

Outline

- Two Cultures of Machine Learning
  - Probabilistic Graphical Models
  - Non-Parametric Discriminative Models
  - Advantages and Disadvantages of Each
- Representing conditional probability distributions using non-parametric machine learning methods
  - Logistic regression (Friedman)
  - Conditional random fields (Dietterich, et al.)
  - Latent variable models (Hutchinson, et al.)
- Ongoing Work
- Conclusions

Next Steps

- Modeling Expertise in Citizen Science
- From Occupancy (0/1) to Abundance (n)
- From Static to Dynamic Models

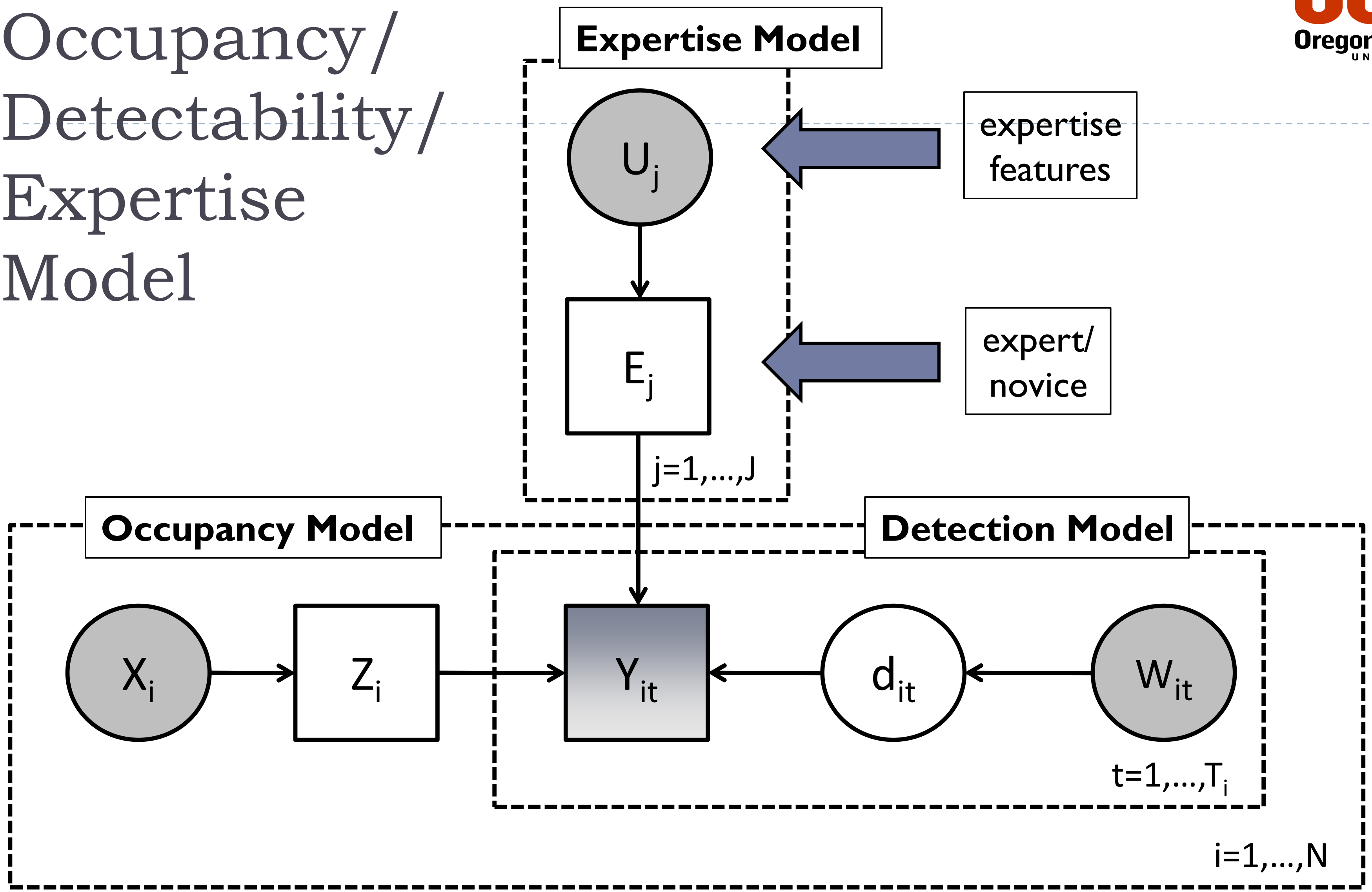
Modeling Expertise in Citizen Science

- Project eBird
  - Bird watchers upload checklists to ebird.org
  - 8,000-12,000 checklists per day uploaded
  - World-wide coverage 24x365
  - 38,599 observers; 336,088 locations
  - 2.4M checklists; 41.7M observations
  - All bird species (~3,000)

  - [Please volunteer! We need more observers in S.America]

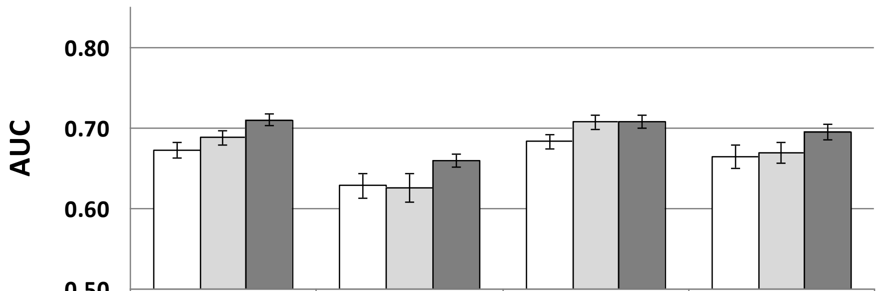
- Wide variation in "birder" expertise

Occupancy/ Detectability/ Expertise Model



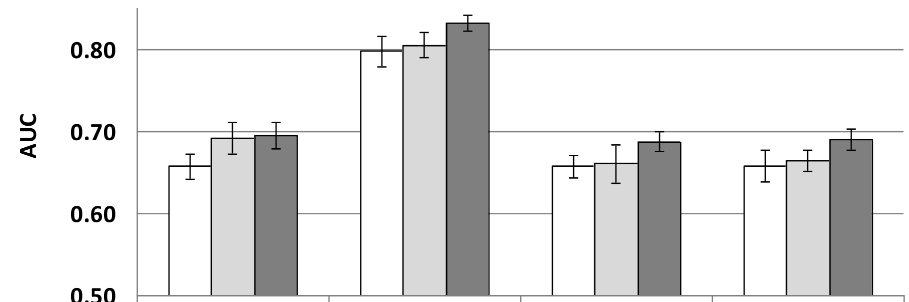
First Results

Average AUC on four common bird species



	Blue Jay	White-breasted Nuthatch	Northern Cardinal	Great Blue Heron
LR	0.6726	0.6283	0.6831	0.6641
OD	0.6881	0.6262	0.7073	0.6691
ODE	0.7104	0.6600	0.7085	0.6959

Average AUC on four hard-to-detect bird species



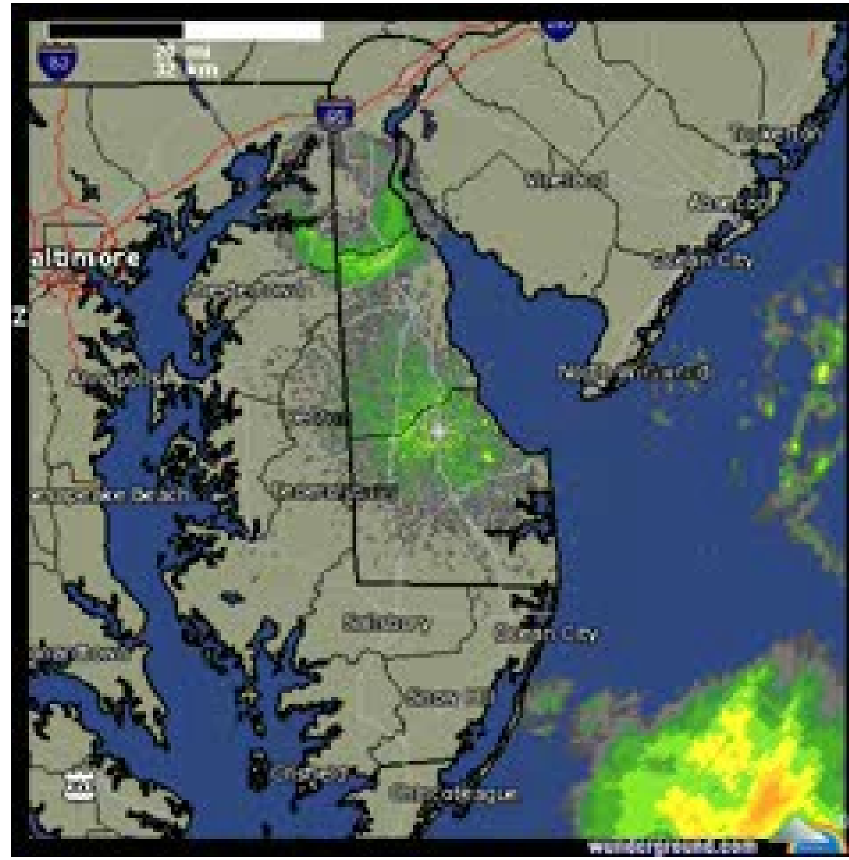
	Brown Thrasher	Blue-headed Vireo	Northern Rough-winged Swallow	Wood Thrush
LR	0.6576	0.7976	0.6575	0.6579
OD	0.6920	0.8055	0.6609	0.6643
ODE	0.6954	0.8325	0.6872	0.6903

- eBird data for May and June (peak detectability period) for NYState
- Expertise component trained via supervised learning

Jun Yu, Weng-Keen Wong, Rebecca Hutchinson (2010). *Modeling Experts and Novices in Citizen Science Data*. ICDM 2010.

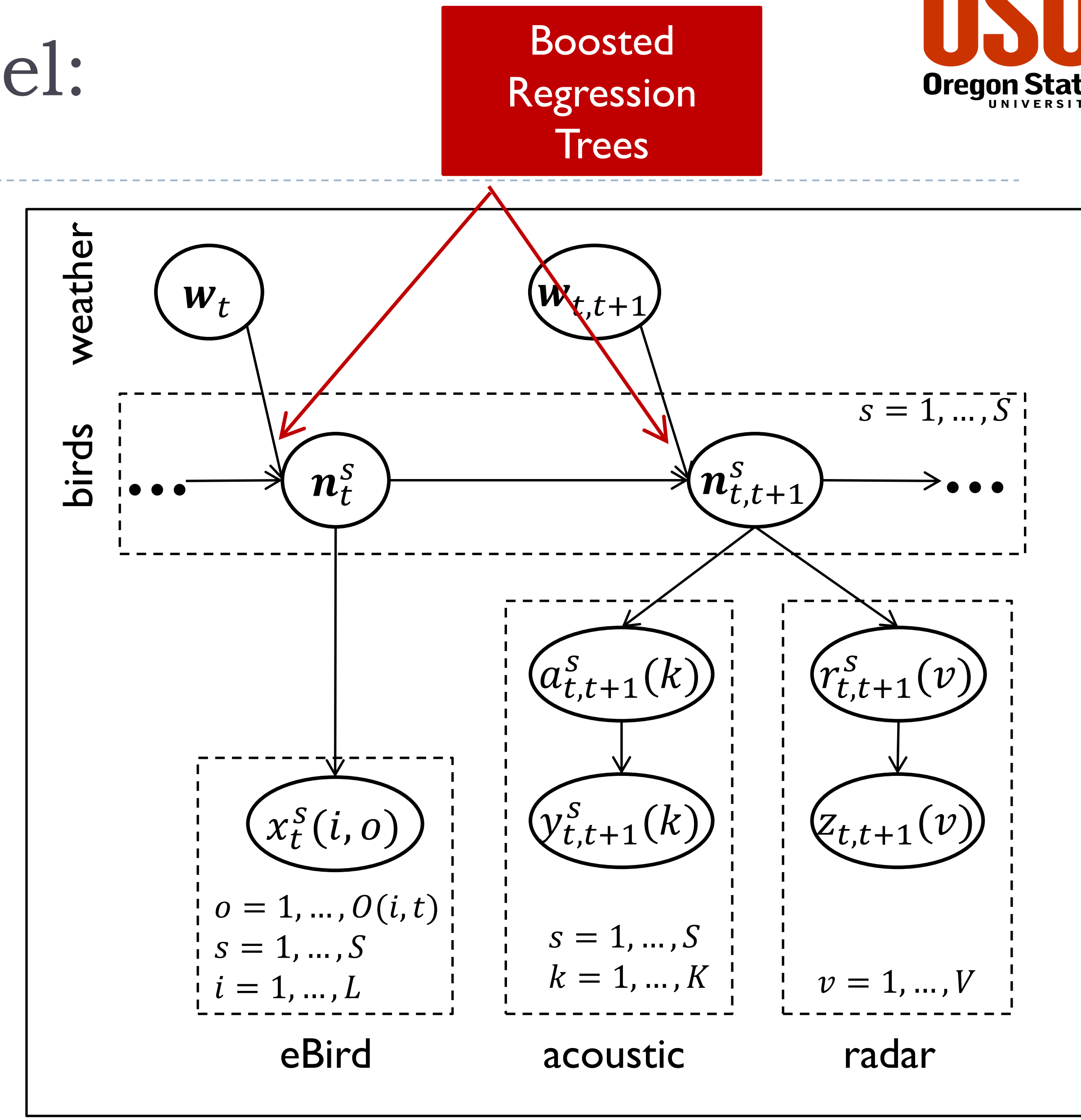
New Project: BirdCast

- Goal: Continent-wide bird migration forecasting
- Additional data sources:
  - Doppler weather radar
  - Night flight calls
  - Wind observations (assimilated to wind forecast model)



BirdCast Model:

- $n_t^s(c)$ = # of birds of species s at cell c and time t.
- $\boldsymbol{w}_t$ = weather variables (wind, temperature, precipitation)
- $x_t^s(i,o)$ = eBird count for visit o at site i species s and time t
- $y_{t,t+1}^s(k)$ = # of flight calls for species s at site k on the night $(t, t+1)$
- $z_{t,t+1}$ = # of birds (all species) observed at radar v on night $(t, t+1)$
- Occupancy changes each night



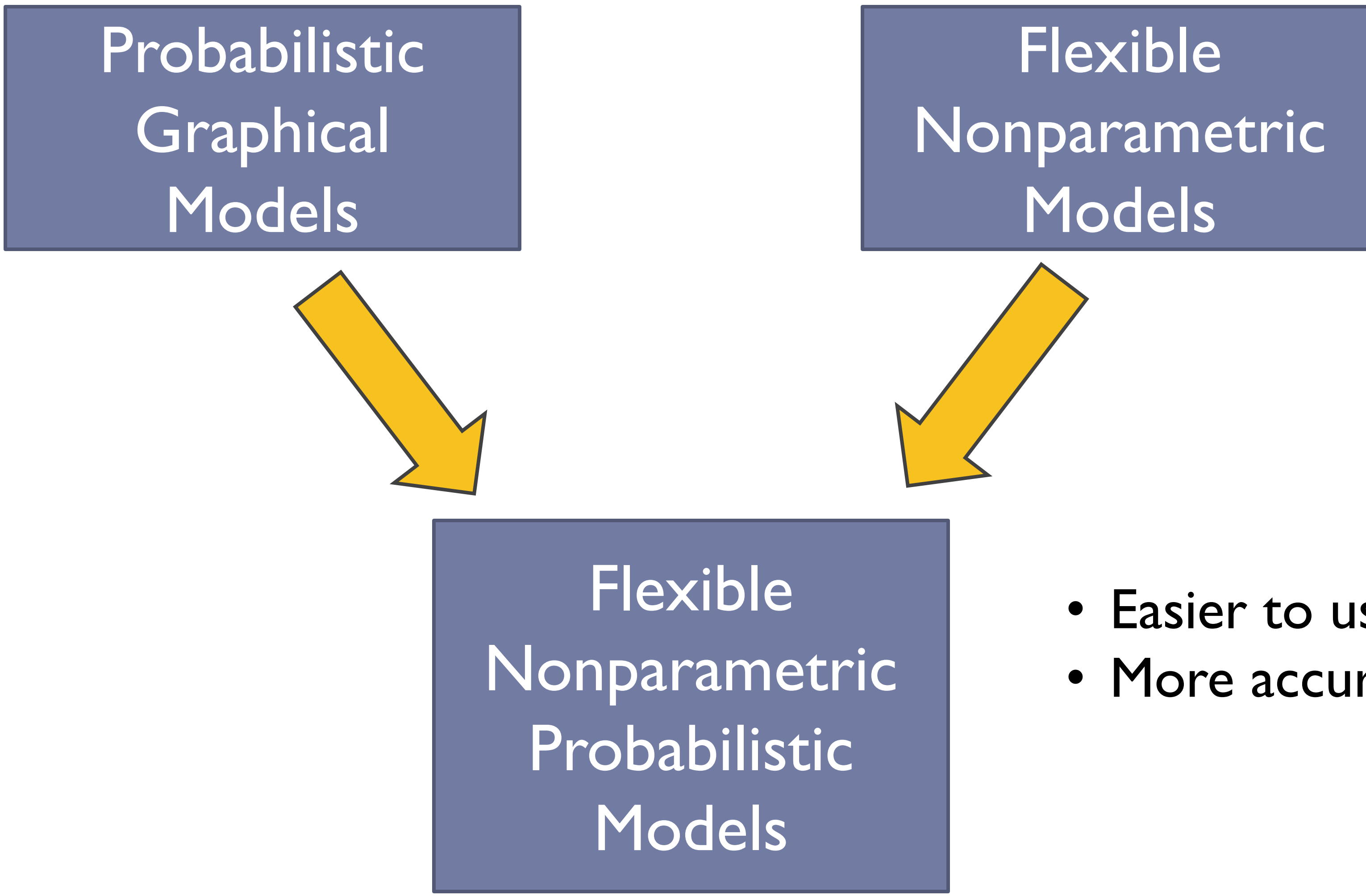
Outline

- Two Cultures of Machine Learning
  - Probabilistic Graphical Models
  - Non-Parametric Discriminative Models
  - Advantages and Disadvantages of Each
- Representing conditional probability distributions using non-parametric machine learning methods
  - Logistic regression (Friedman)
  - Conditional random fields (Dietterich, et al.)
  - Latent variable models (Hutchinson, et al.)
- Ongoing Work
- Conclusions

Concluding Remarks

- Gradient Tree Boosting can be integrated into probabilistic graphical models
  - Fully-observed directed models
  - Conditional random fields
  - Latent variable models
- When to do this?
  - When you want to condition on a large number of features
  - When you have a lot of data

Combining Two Approaches to Machine Learning



- Easier to use
- More accurate

Thank-you

- Adam Ashenfelter, Guo-Hua Hao: TreeBoosting for CRFs
- Rebecca Hutchinson, Liping Liu: Boosted Regression Trees in OD models
- Weng-Keen Wong, Jun Yu: ODE model
- Dan Sheldon: Models for Bird Migration
- Steve Kelling and colleagues at the Cornell Lab of Ornithology

- National Science Foundation Grants 0083292, 0307592, 0832804, and 0905885