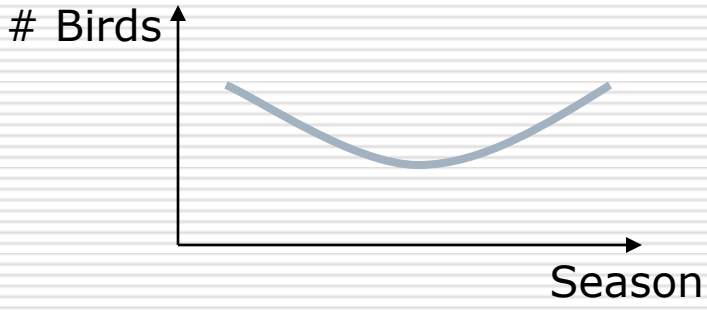
Detecting Statistical Interactions with Additive Groves of Trees

Daria Sorokina, Rich Caruana,
Mirek Riedewald, Daniel Fink

Domain Knowledge Questions

- Which features are important?
- What effects do they have on the response variable?
  - Effect visualization techniques

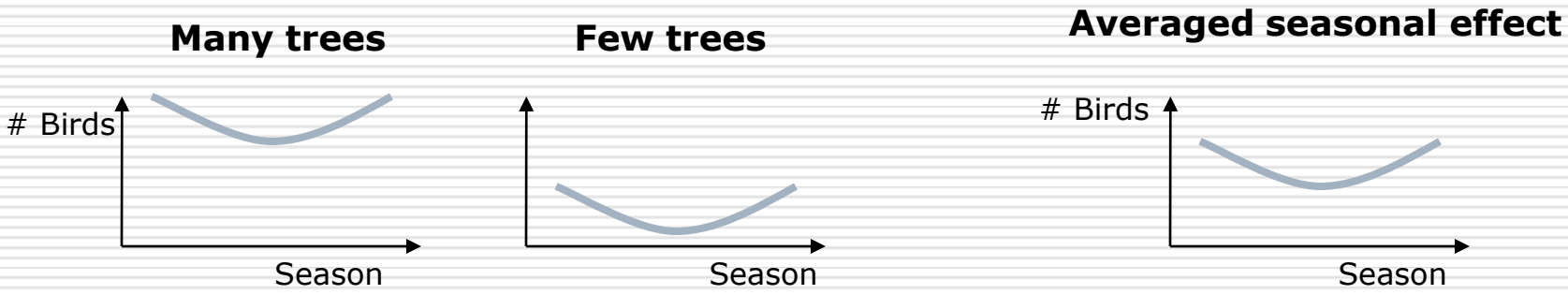
Toy example: seasonal effect on bird abundance



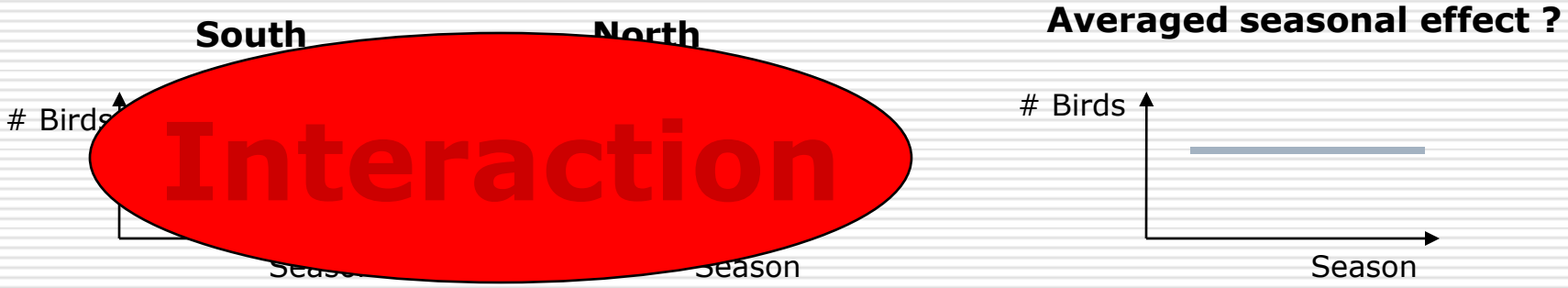
  - Is it always possible to visualize an effect of a single variable?

Visualizing effects of features

- Toy example 1: # Birds = F(season, #trees)



- Toy example 2: # Birds = F(season, latitude)



!

Statistical interactions are **NOT** correlations

!

Statistical Interactions

- *Statistical interactions* ≡ 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

$$F(x_1,x_2,\ldots x_n) =$$
$$G(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) + H(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n),$$

i.e., G does not depend on x_i, H does not depend on x_j

- Example:

$F(x_1,x_2,x_3) = \sin(x_1+x_2) + x_2 \cdot x_3$

  - x_1, x_2 interact
  - x_2, x_3 interact
  - x_1, x_3 do not interact

Interaction Detection Approach

How to test for an interaction:

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 the restricted model is as good as the unrestricted – there is no interaction.
   - If it fails to represent the data with the same quality – there is interaction.

Learning Method Requirements

1. Non-linearity
   - If unrestricted model does not capture interactions, there is no chance to detect them
2. Restriction capability (additive structure)
   - The performance should not decrease after restriction when there are no interactions

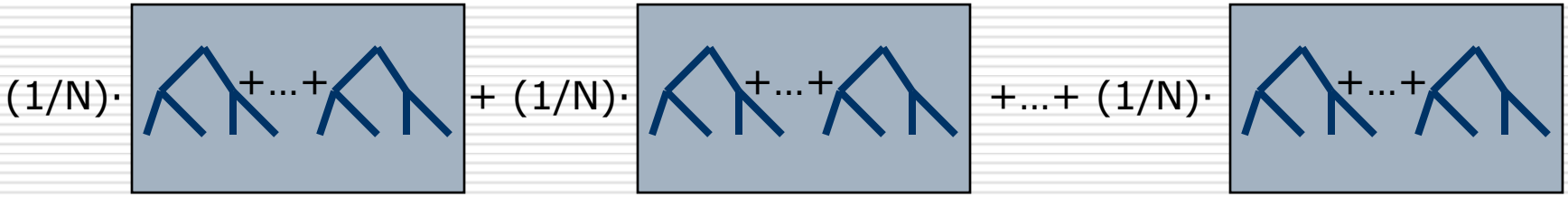
- Most existing prediction models do not fit both requirements at the same time
  - We had to invent our own algorithm that does

Additive Groves of Regression Trees
(Sorokina, Caruana, Riedewald; ECML'07)

- New regression algorithm
  - Ensemble of regression trees
- Based on
  - Bagging
  - Additive models
  - Combination of large trees and additive structure
- Useful properties
  - High predictive performance
  - Captures interactions
  - Easy to restrict specific interactions

Additive Groves

- **Additive models** fit additive components of the response function
- A **Grove** is an additive model where every single model is a tree
- **Additive Groves** applies bagging on top of single Groves



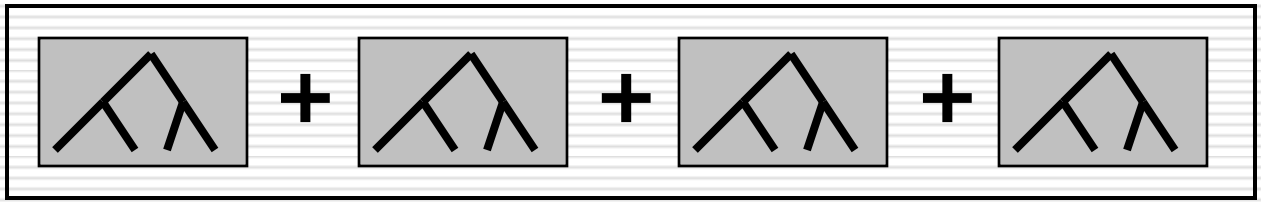
Interaction Detection Approach

How to test for an interaction:

1. Build a model from the data (no restrictions).
2. Build a restricted model – do not allow the interaction of interest.
3. Compare their predictive performance.
   - If the restricted model is as good as the unrestricted – there is no interaction.
   - If it fails to represent the data with the same quality – there is interaction.

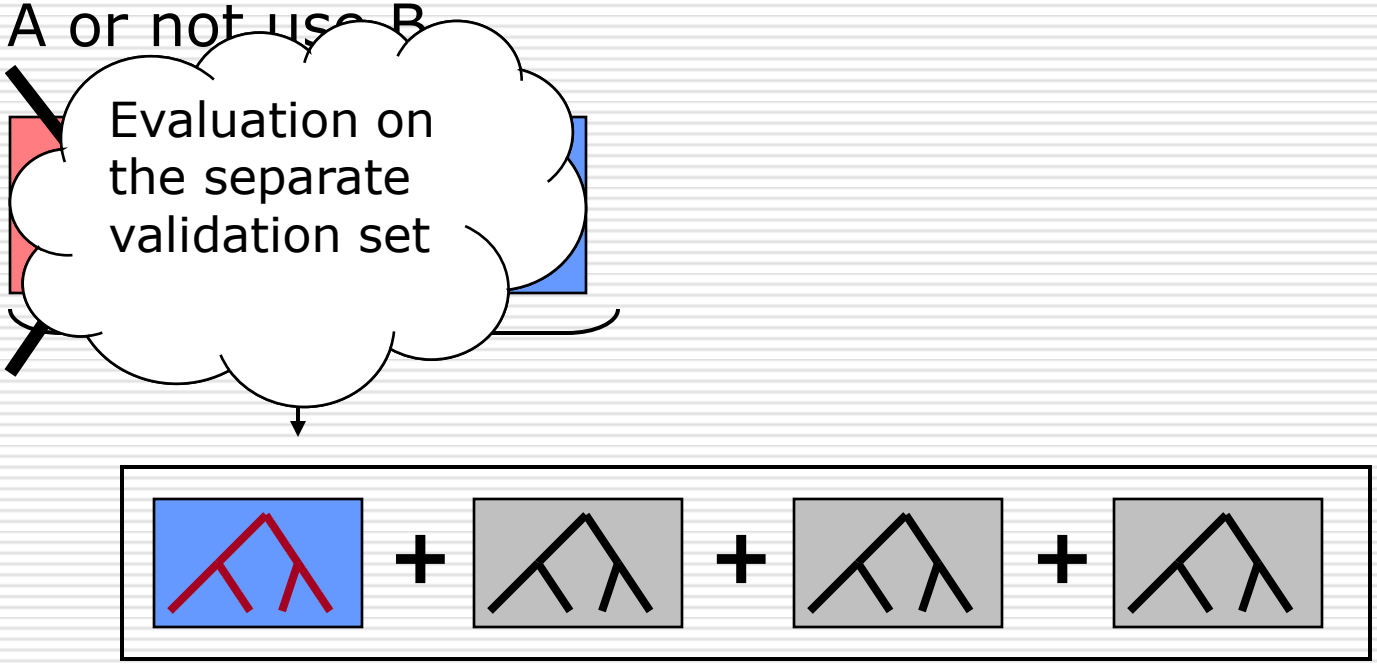
Training Restricted Grove of Trees

- The model is not allowed to have interactions between features A and B
- Every single tree in the model should either not use A or not use B



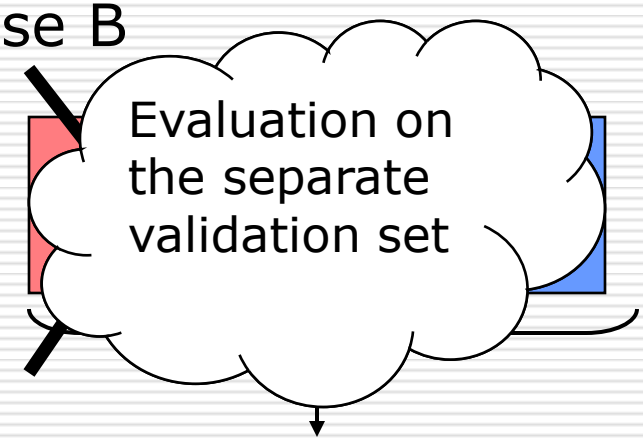
Training Restricted Grove of Trees

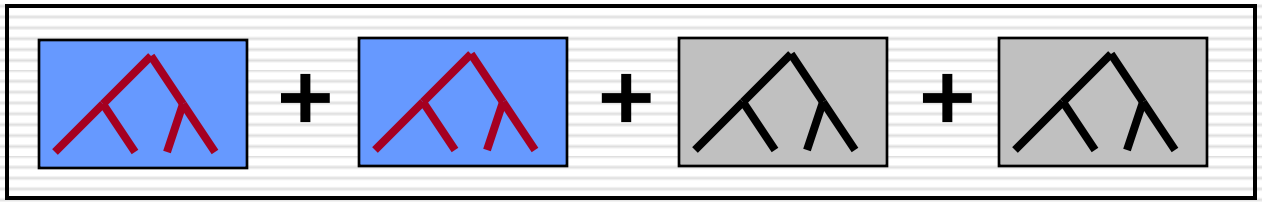
- The model is not allowed to have interactions between attributes A and B
- Every single tree in the model should either not use A or not use B



Training Restricted Grove of Trees

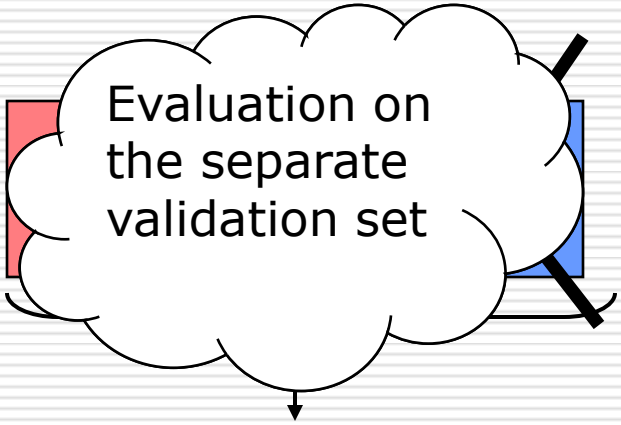
- The model is not allowed to have interactions between attributes A and B
- Every single tree in the model should either not use A or not use B

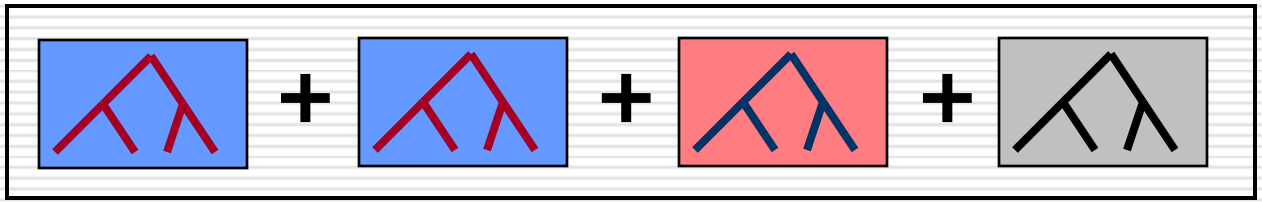




Training Restricted Grove of Trees

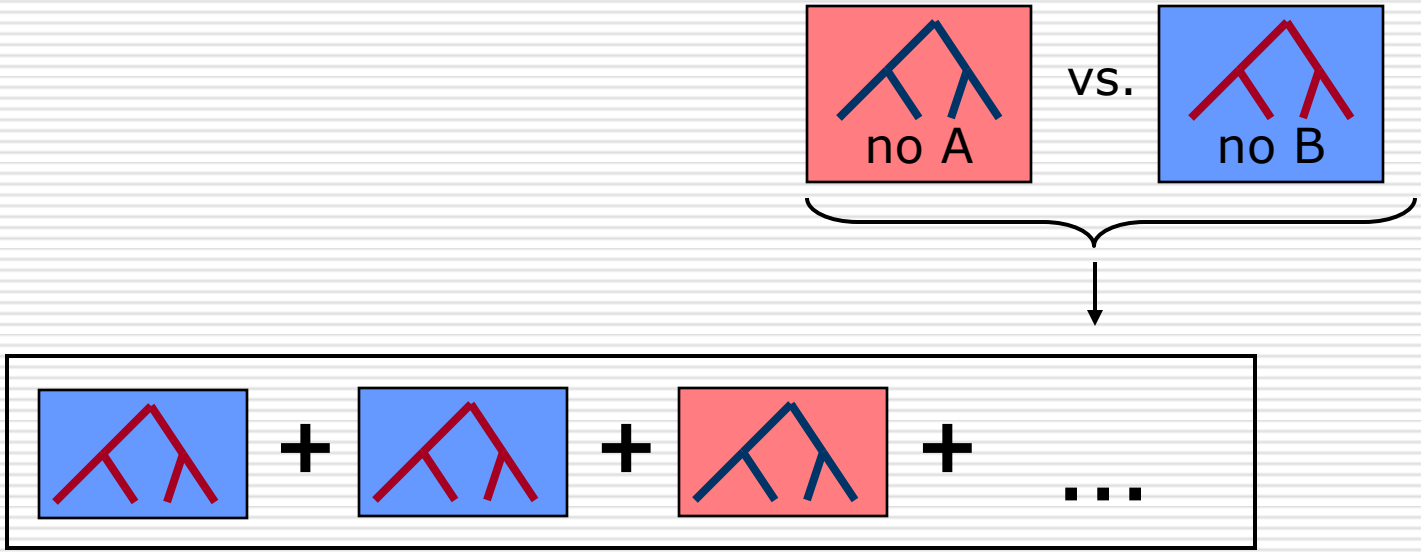
- The model is not allowed to have interactions between attributes A and B
- Every single tree in the model should either not use A or not use B





Training Restricted Grove of Trees

- The model is not allowed to have interactions between attributes A and B
- Every single tree in the model should either not use A or not use B



Higher-Order Interactions

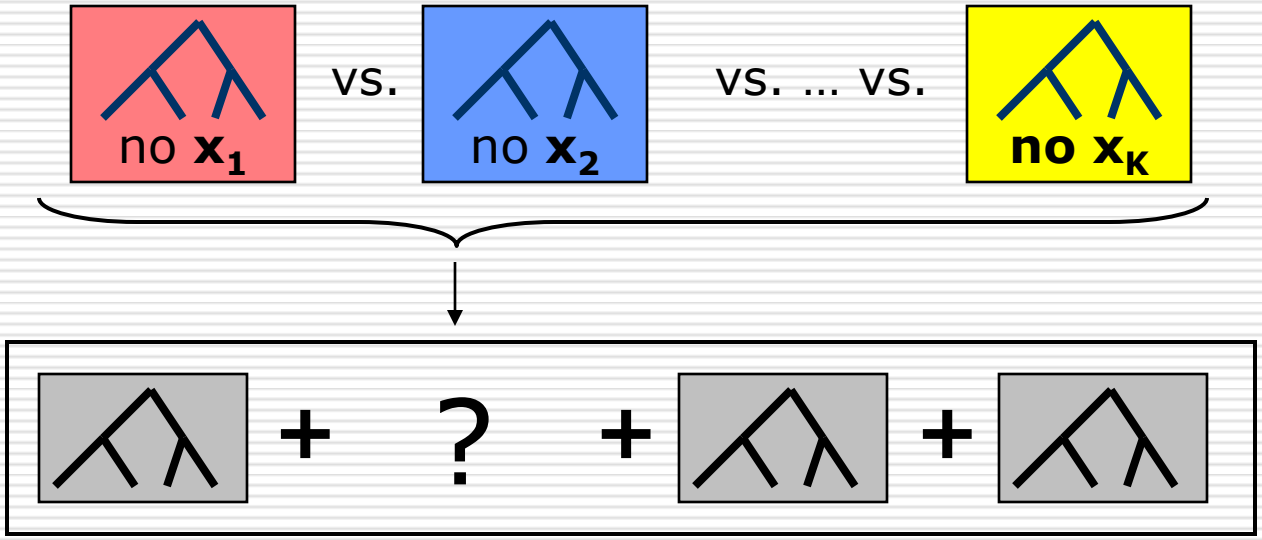
- $F(\mathbf{x})$ shows no K-way interaction between $x_1, x_2, \ldots, x_K$ when
$$F(\mathbf{x}) = F_1(\mathbf{x}_{\setminus 1}) + F_2(\mathbf{x}_{\setminus 2}) + \ldots + F_K(\mathbf{x}_{\setminus K}),$$
where each F_i does not depend on x_i

  - $(x_1+x_2+x_3)^{-1}$ – has a 3-way interaction
  - $x_1+x_2+x_3$ – has no interactions (neither 2 nor 3-way)
  - $x_1x_2 + x_2x_3 + x_1x_3$ – has all 2-way interactions, but no 3-way interaction

Higher-Order Interactions

- F(**x**) shows no K-way interaction between x_1, x_2, …, x_K when
  $$F(\mathbf{x}) = F_1(\mathbf{x}_{\setminus 1}) + F_2(\mathbf{x}_{\setminus 2}) + \ldots + F_K(\mathbf{x}_{\setminus K}),$$
  where each F_i does not depend on x_i

- K-way restricted Grove: K candidates for each tree



Quantifying Interaction Strength

- Performance measure: standardized root mean squared error

$$stRMSE = \sqrt{\frac{1}{N}\sum\left(F(x)-y\right)^2} \Big/ StD(y)$$

- Interaction strength: difference in performances of restricted and unrestricted models

$$I_{i,j} = stRMSE(R_{i,j}(x)) - stRMSE(U(x))$$

- Significance threshold: 3 standard deviations of unrestricted performance

$$I_{i,j} > 3 \cdot StD\left(stRMSE\left(U(x)\right)\right)$$

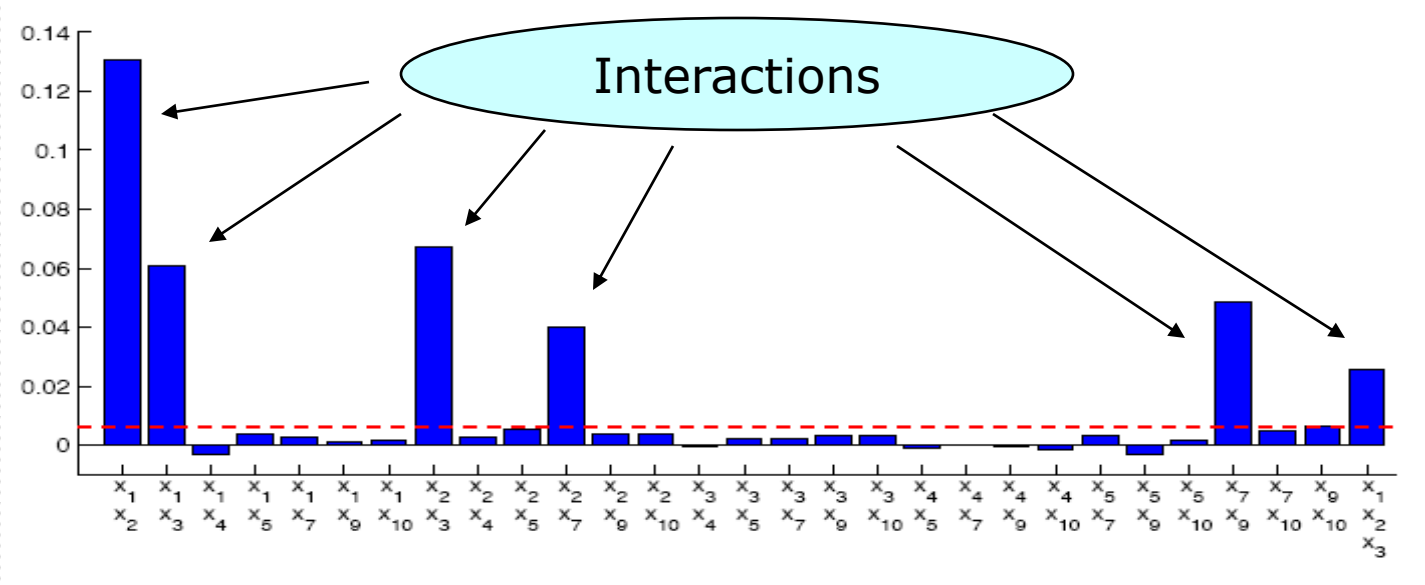
  - Randomization comes from different data samples (folds, bootstraps…)

Correlations and Feature Selection

- Correlations between the variables hurt interaction detection
- Solution: feature selection.
  - Correlated features will be removed
- Also, feature selection will leave few variable pairs to check for interactions
  - As opposed to N^2

Experiments: Synthetic Data

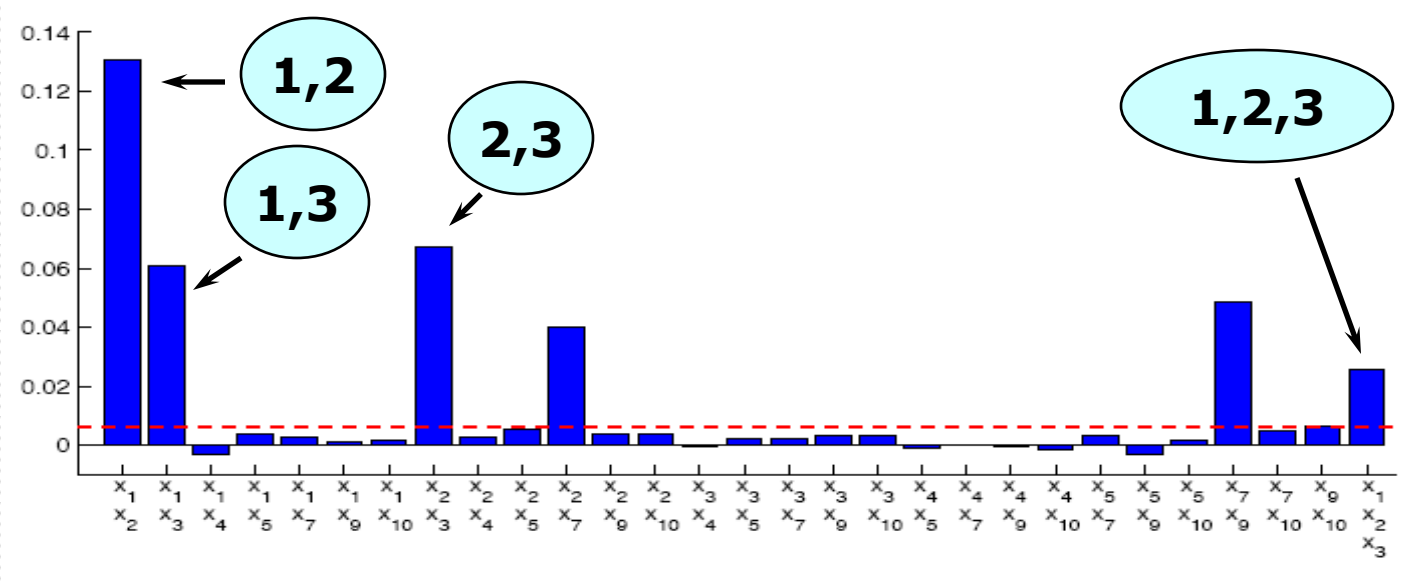
$$Y = \pi^{x_1 x_2} \sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}} \sqrt{\frac{x_7}{x_8}} - x_2 x_7$$



Daria Sorokina, Rich Caruana, Mirek Riedewald, Daniel Fink
Detecting Statistical Interactions with Groves of Trees.

Experiments: Synthetic Data

$$Y = \pi^{x_1 x_2}\sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}}\sqrt{\frac{x_7}{x_8}} - x_2 x_7$$

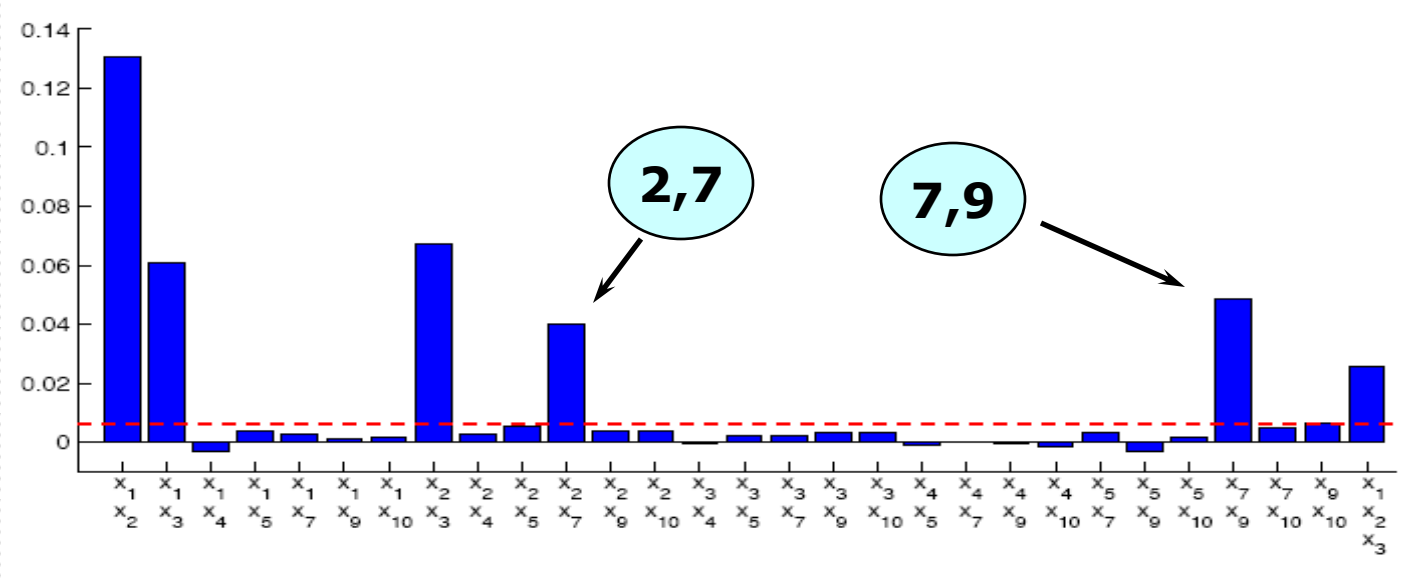


Daria Sorokina, Rich Caruana, Mirek Riedewald, Daniel Fink
Detecting Statistical Interactions with Groves of Trees.

Cornell University

Experiments: Synthetic Data

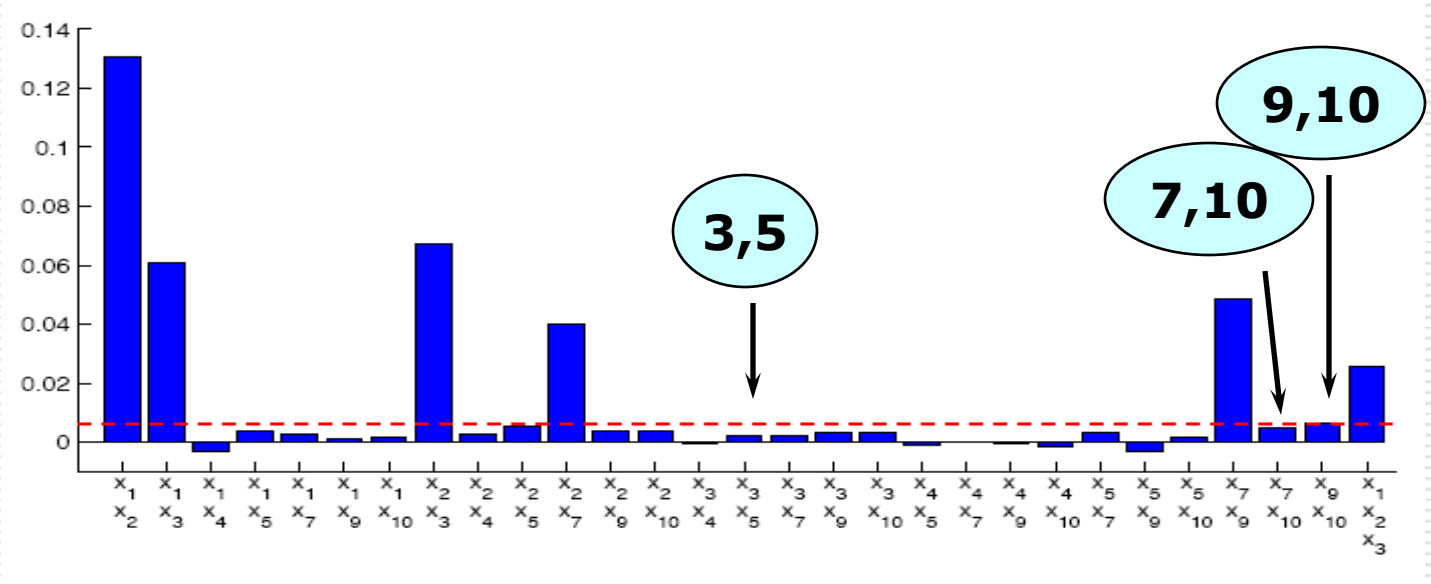
$$Y = \pi^{x_1 x_2} \sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}} \sqrt{\frac{x_7}{x_8}} - x_2 x_7$$



Experiments: Synthetic Data

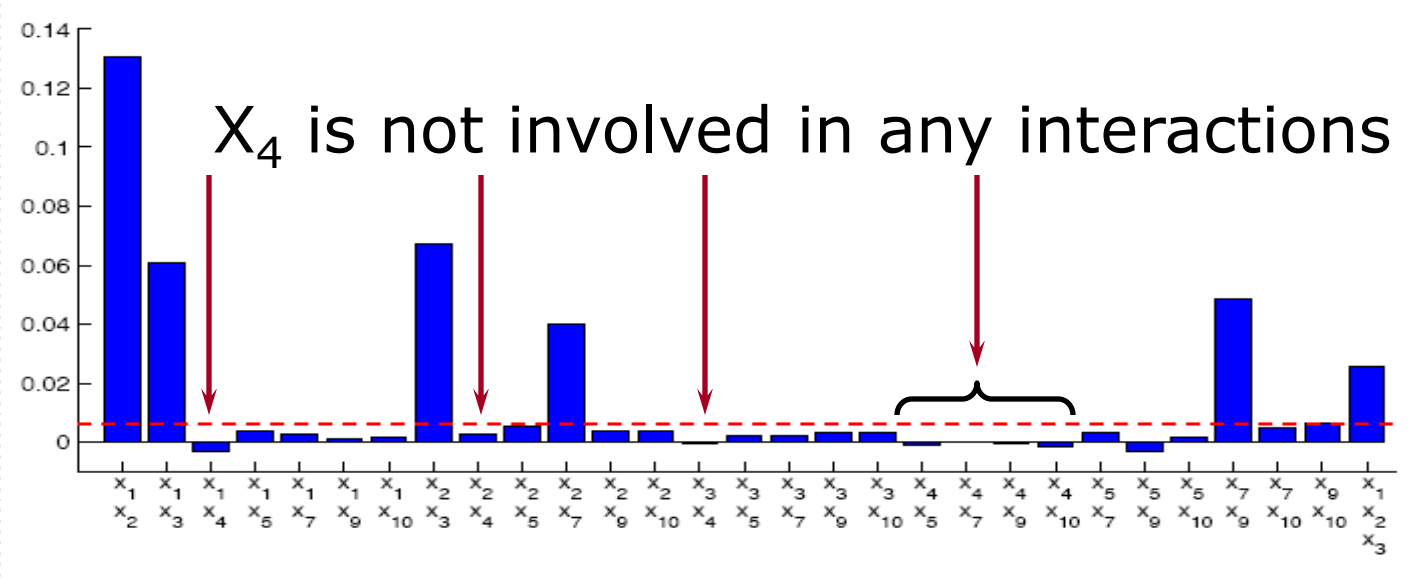
$$Y = \pi^{x_1 x_2}\sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}}\sqrt{\frac{x_7}{x_8}} - x_2 x_7$$

x_5, x_8, x_{10} have small ranges by construction and do not influence response much. Interactions of all other variables are detected.



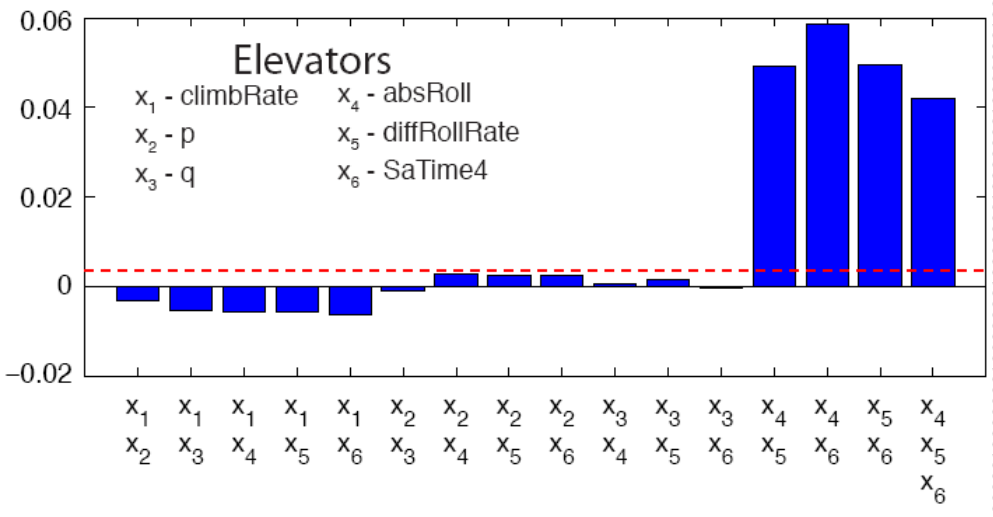
Experiments: Synthetic Data

$$Y = \pi^{x_1 x_2} \sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}} \sqrt{\frac{x_7}{x_8}} - x_2 x_7$$



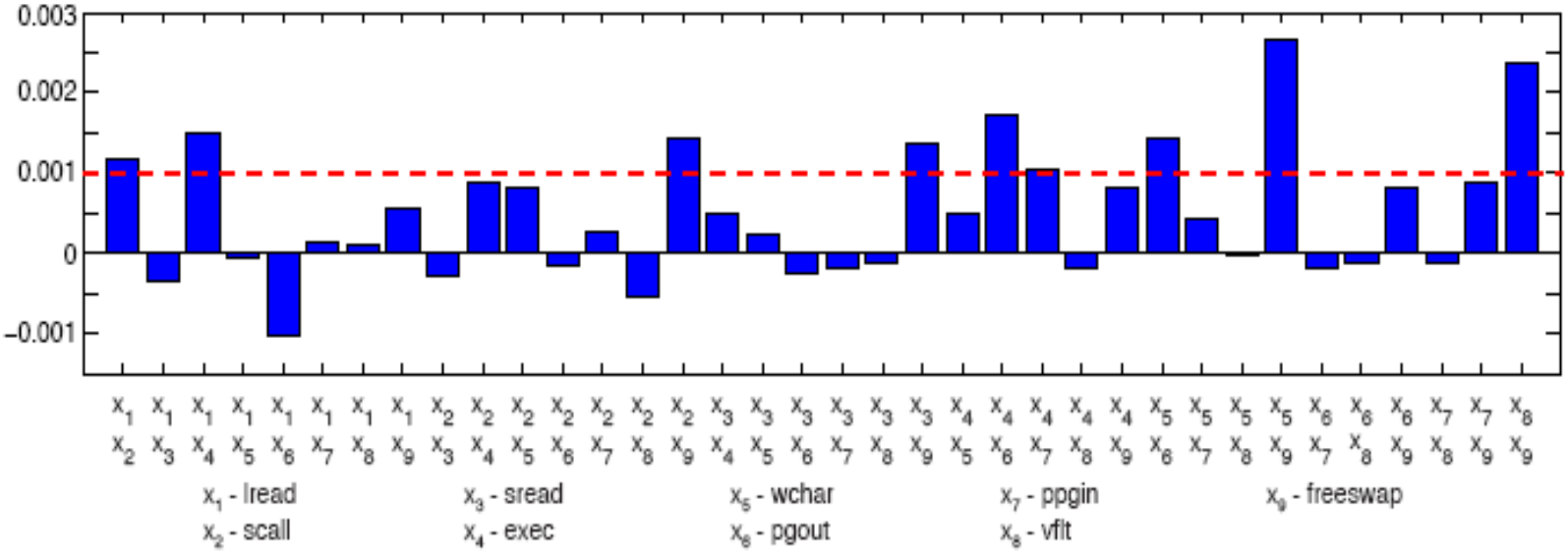
Experiments: Elevators

- Airplane control data set: predict required position of elevators
- 1 strong 3-way interaction
  - absRoll – absolute value of the roll angle
  - diffRollRate – roll angular acceleration
  - SaTime4 – position of ailerons 4 time steps ago



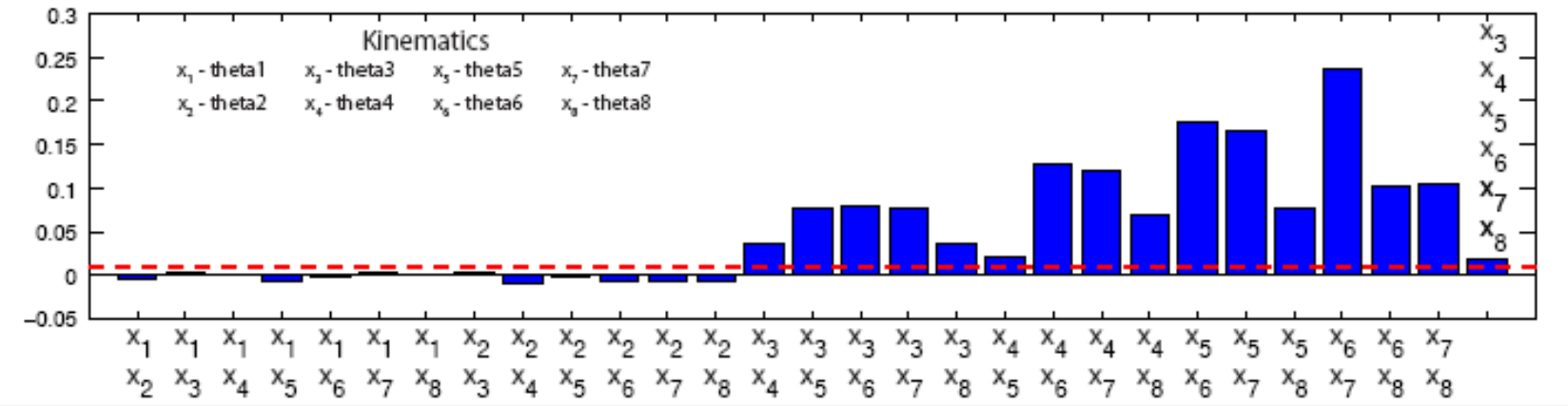
Experiments: CompAct

- Predict CPU activity from other computer system parameters
- A very additive, almost linear data set
- All detected interactions were fairly small and non-stable



Experiments: Kinematics

- Simulation of an 8-link robotics arm movements
- Predict distance between the end and the origin from values of joints angles
- Highly non-linear data set, contains a 6-way interaction



Experiments: House Finch Abundance Data

- Interaction (year, latitude)
  - corresponds to an eye-disease that affected house finches during the decade covered by the dataset

Summary

- Statistical interaction detection shows which features should be analyzed in groups
- We presented a novel technique, based on comparing restricted and unrestricted models
- Additive Groves is an appropriate learning method for this framework

Acknowledgements

- Our collaborators in Computer Science department and Cornell Lab of Ornithology:
  - Wes Hochachka
  - Steve Kelling
  - Art Munson

Appendix

- Related work
  - Statistical methods
  - (Friedman & Popescu, 2005)
  - (Hooker, 2007)
- Regression trees
- Trying to restrict bagged trees

Regression trees used in Groves

- Each split optimizes RMSE
- Parameter α controls the size of the tree
  - Node becomes a leaf if it contains ≤ α·|trainset| cases
  - 0 ≤ α ≤ 1, the smaller α, the larger the tree

(Any other type of regression tree could be used.)

Related work: early statistical methods

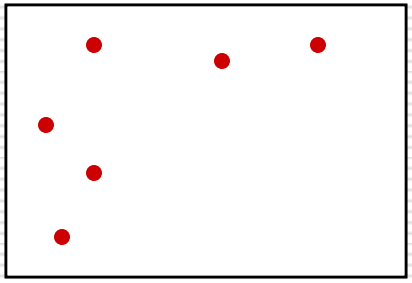
(Neter et. al., 1996) (Ott & Longnecker, 2001)

- Build a linear model with an interaction term
  - $\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + \beta x_1 \cdot x_2$
- Test whether β is significantly different from 0
- Problem: limited types of interaction

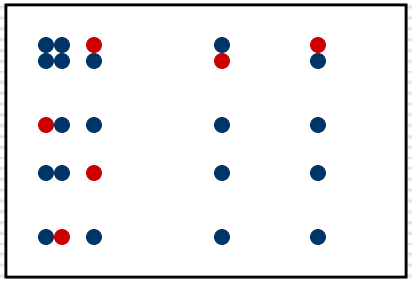
- Collect data for all combination of parameter values
- Find value of interaction term for each combination
- Test whether interaction is significant
- Problem: not useful for high-dimensional data sets

Related Work: Partial Dependence Functions (Friedman & Popescu, 2005)

- No interaction $\equiv E_{\setminus x}(F()) + E_{\setminus z}(F()) = E_{\setminus\{x,z\}}(F()) + E(F())$
  - But only if x and z are distributed independently
- Create "fake" data points in the data set
- Check for interactions in the resulting data
- Problem: fake interactions in the fake data
  - (Hooker, Generalized Functional ANOVA diagnostics, 2007)



Real data



Real and fake data

Related Work: Generalized Functional ANOVA Diagnostics (Hooker, 2007)

- Improvement on partial dependence functions algorithm
- Estimates joint distribution and penalizes the areas with small density
- Produces results based on real data
- High complexity
  - Dense grid
  - External density estimation

Would other ensembles work?

- Let's try to restrict bagging in the same way.
- Assume
  - A and B are both important
  - A is more important than B
  - There is no interaction between A and B
- First tree:
  - A is more important, the tree without B performs better. Choose the tree without B
- Second tree:
  - A is more important, the tree without B performs better. Choose the tree without B
- N-th tree:
  - …
- Now the whole ensemble consists of trees without B!
- B is important, so the performance dropped
  - But there was no interaction…