SAND2012-9408P

Estimating Extrapolation Risk in Supervised Machine Learning

Should I trust *this* prediction?

Art Munson, Philip Kegelmeyer

Sandia National Laboratories

November 1, 2012

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

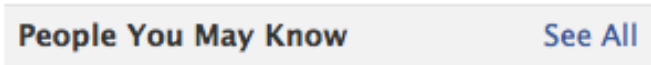
Too Much Traffic to Monitor Manually

Maybe Machine Learning Can Help...

Web Search

bing

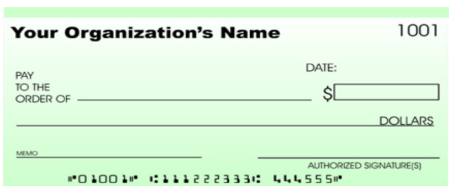
Friend Recommendations



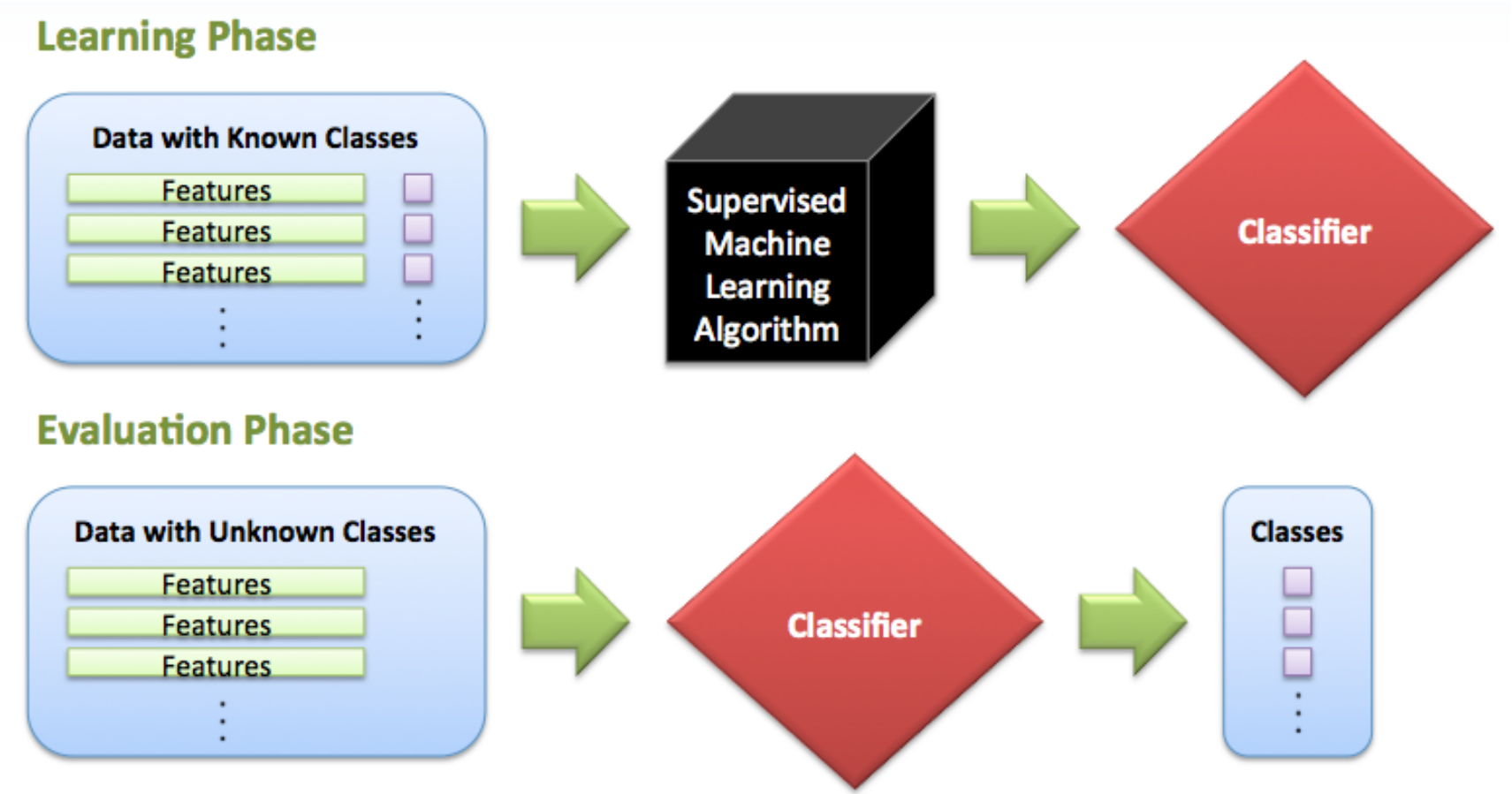
Pose Recognition in Kinect

Winning Jeopardy

Reading Bank Checks



Supervised Machine Learning from 10K Feet



Successful Applications:

- Bing (Microsoft)
- Kinect (Microsoft)
- Friend Recommendations (Facebook)

But is ML Suitable for Important Decisions?

But is ML Suitable for Important Decisions?

. . . like suggestions for who to date?

George is your best match.

MATCH	PROB OF FUN
George	0.75
Tom	0.7
Mike	0.68

I think you'll like Tom.

- Are you sure?
- How many matches have you made like mine?
- Why do you think I'll like Tom?

The IID Assumption in Machine Learning

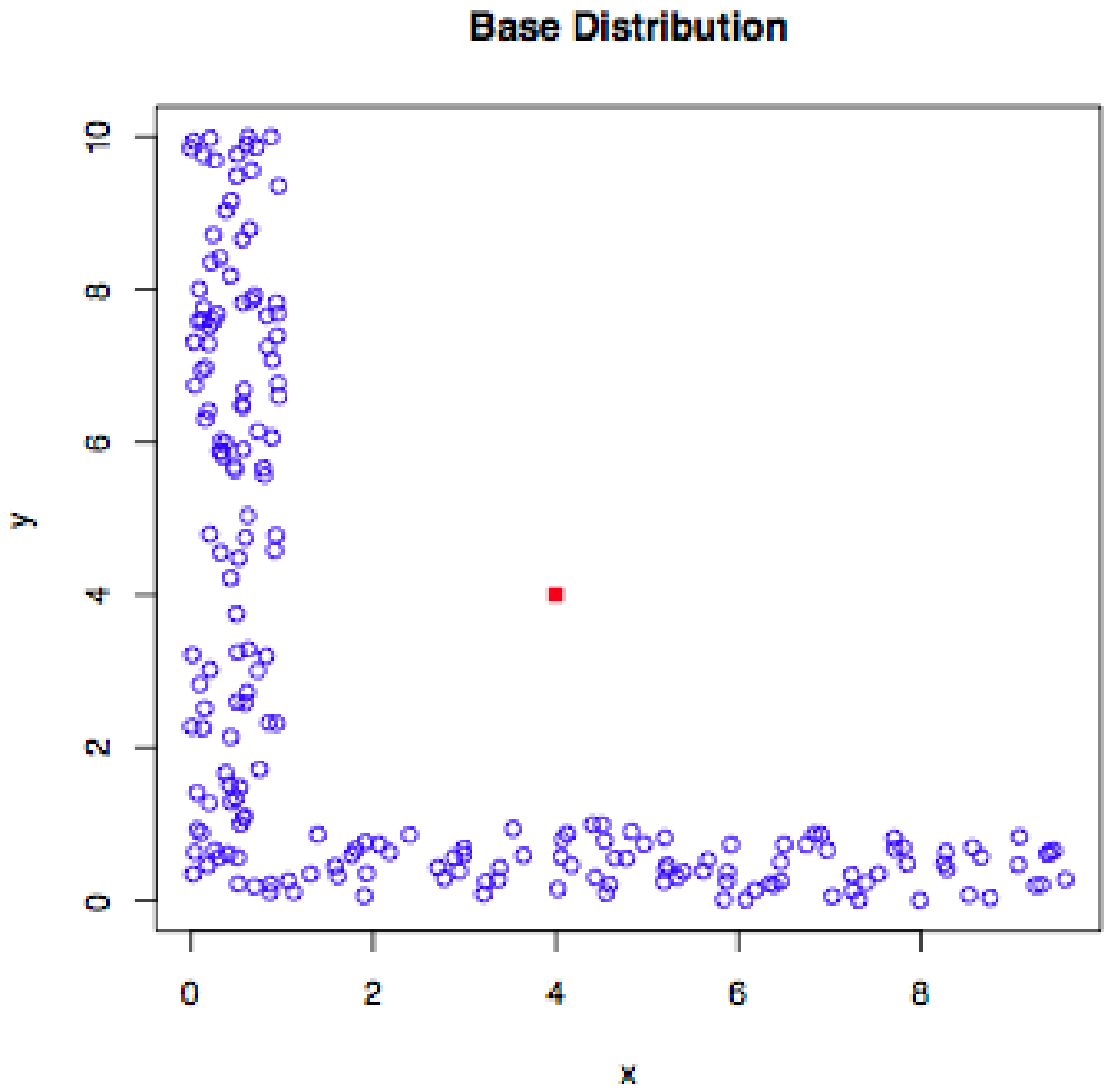
IID = **I**ndependent and **I**dentically **D**istributed
Assumes future data looks like past data.

What happens if:

- a new category appears?
- future data is noisier?
- a category evolves (e.g., malware)?

Answer: user gets a prediction, business as usual.

A Toy Example



Source: Hooker (2004).

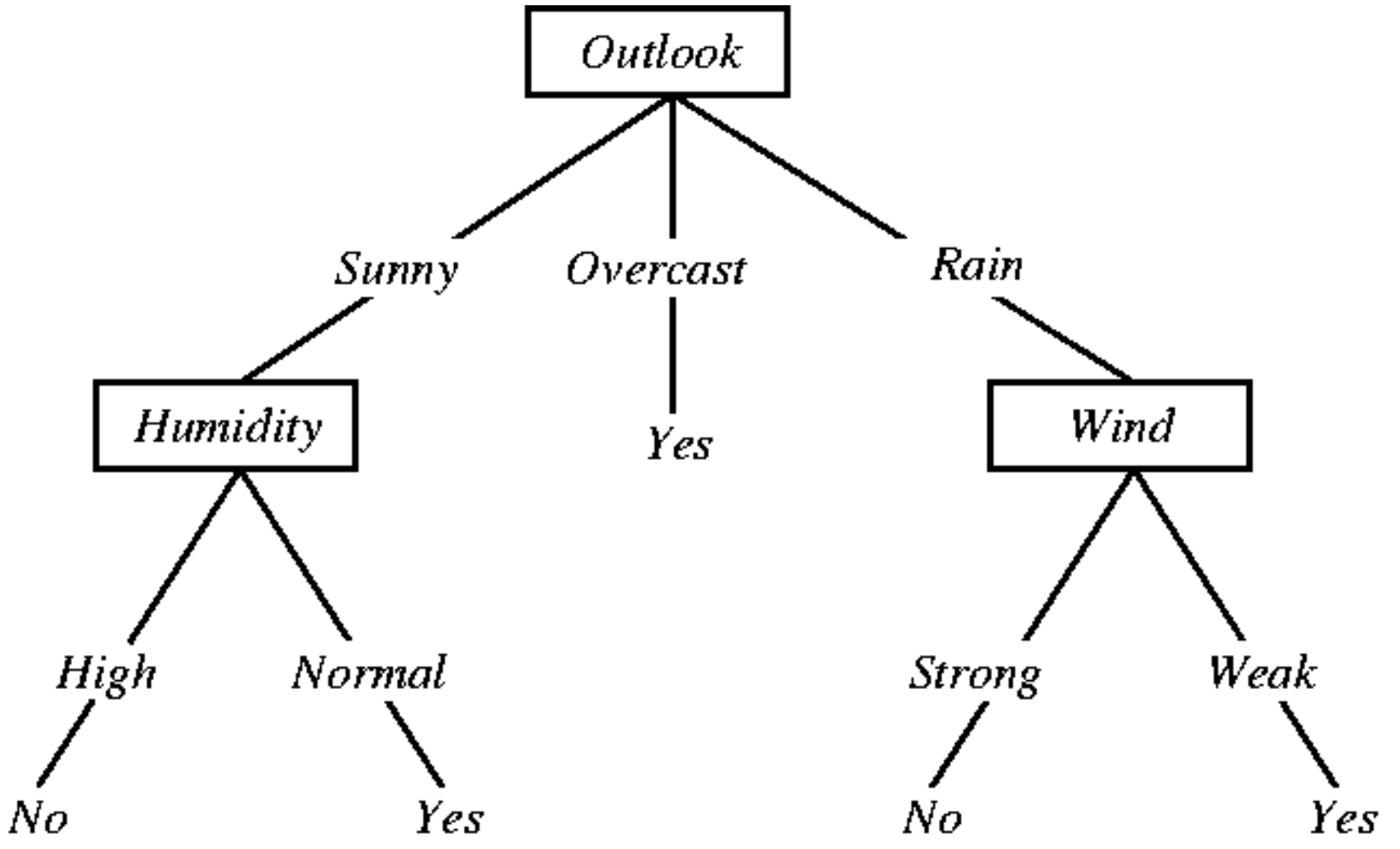
Can we detect when machine learning is extrapolating on new data?

Can we detect when machine learning is extrapolating on new data?

Focus: decision tree ensembles

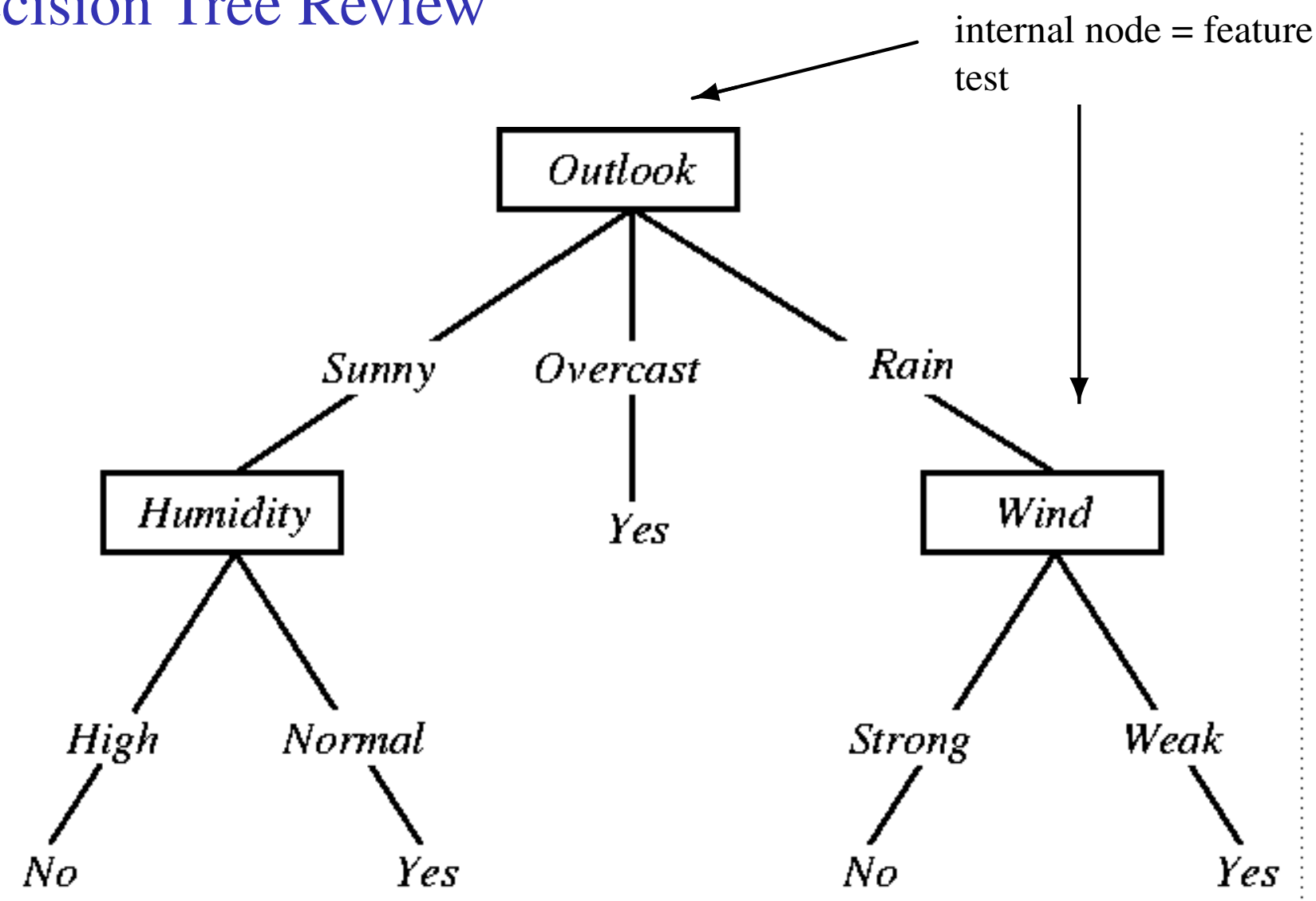
Outline

Decision Tree Review



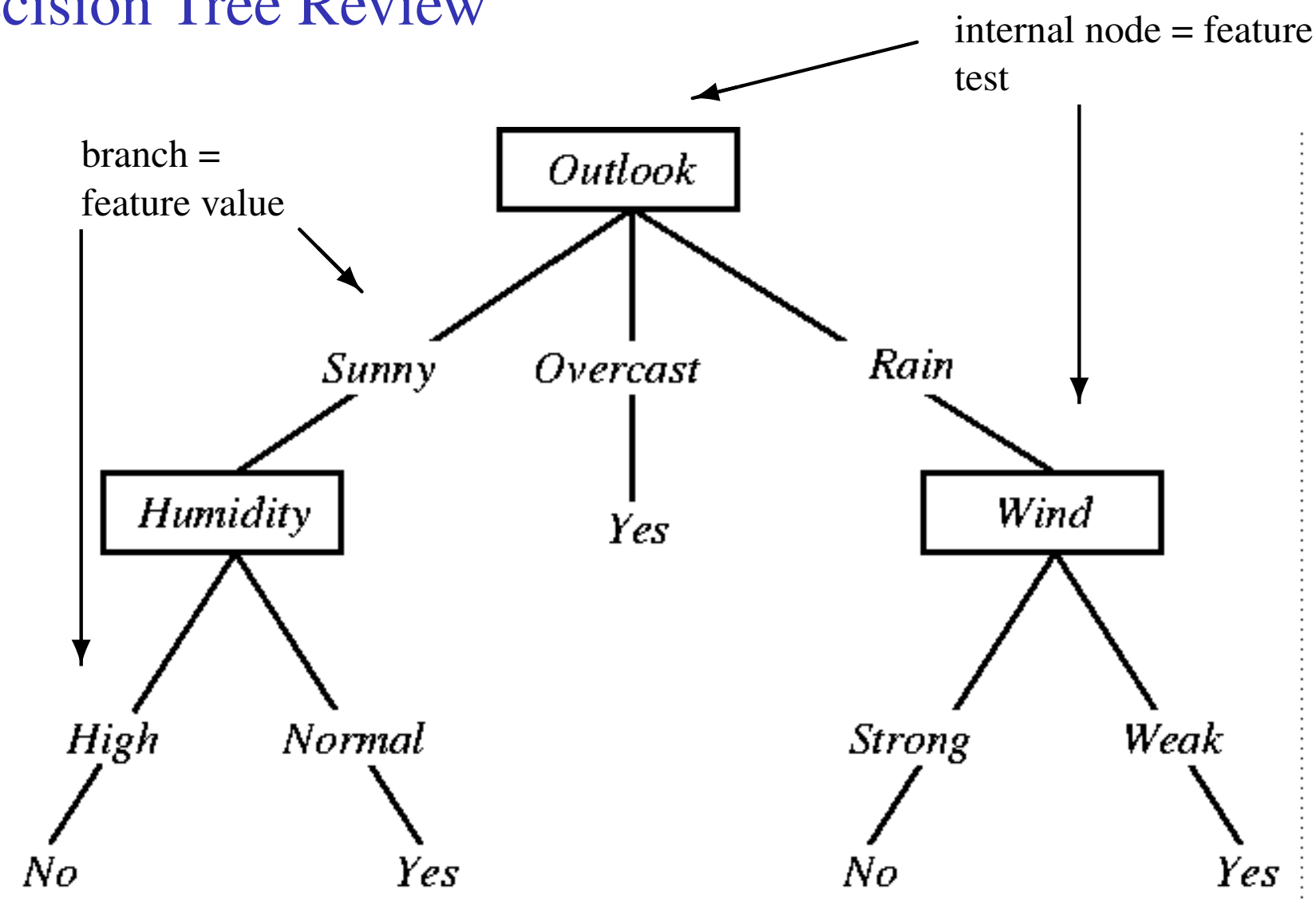
©Tom Mitchell, McGraw Hill, 1997

Decision Tree Review

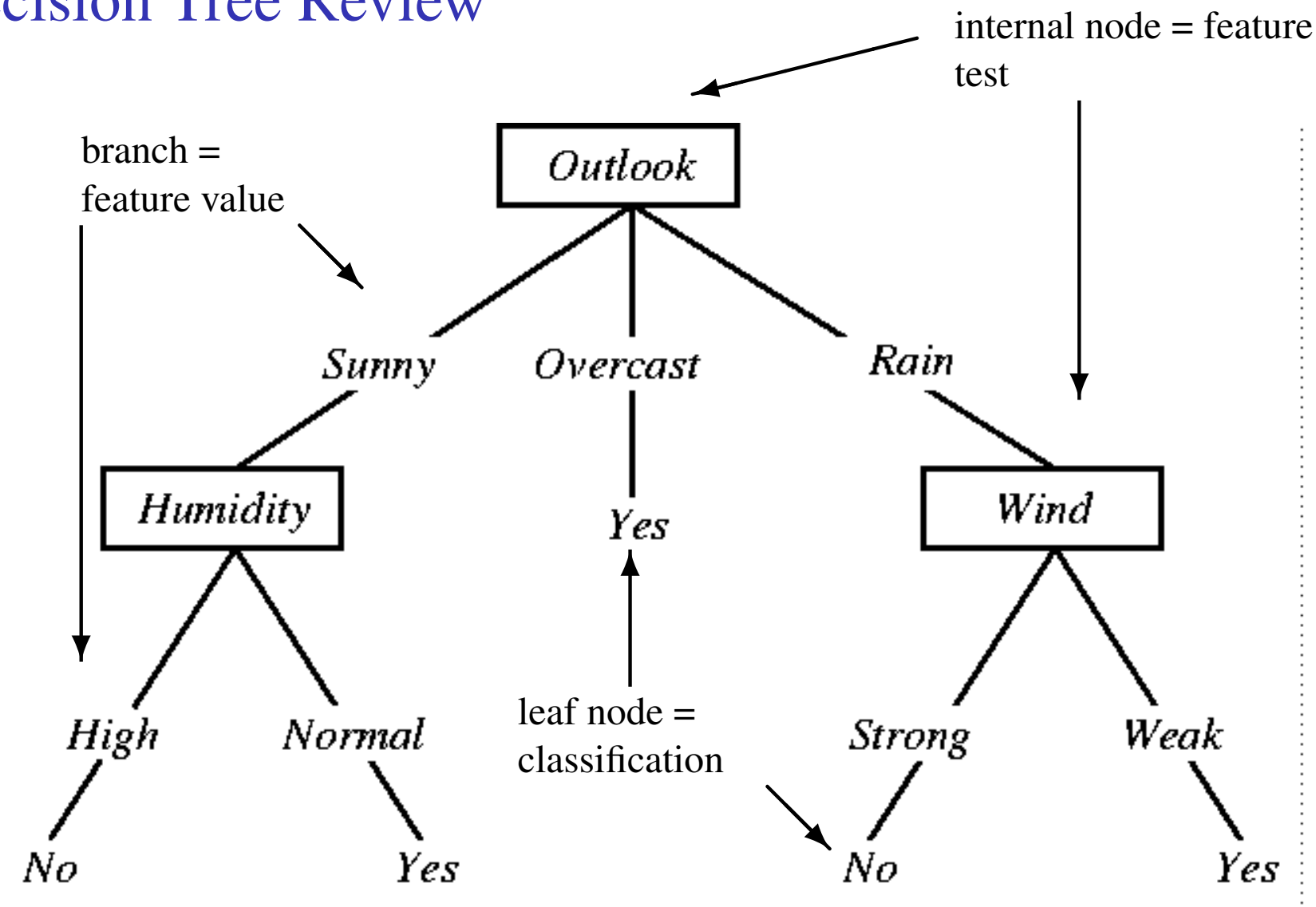


©Tom Mitchell, McGraw Hill, 1997

Decision Tree Review



Decision Tree Review



Decision Tree Strengths & Weaknesses

Strengths:

- Handle numeric and categorical features.
- Missing values are okay.
- Invariant to monotonic feature scaling.
- Robust to noisy training labels.
- Fast.

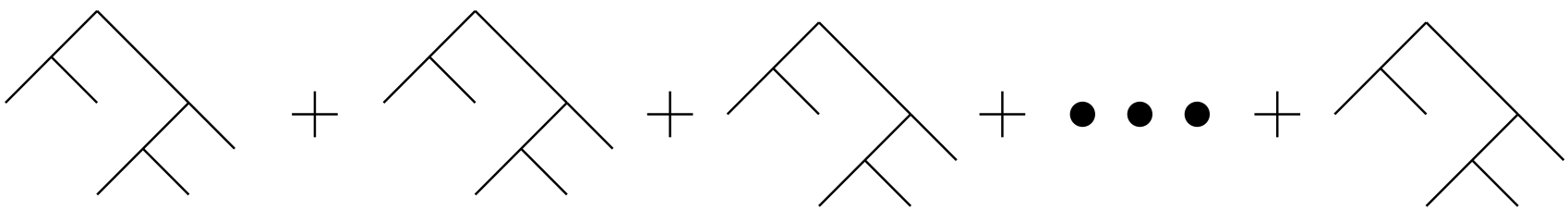
Weaknesses:

- High variance.

Review of Simple Ensemble Learning

Bagging: simple ensemble learning algorithm [1]:

- draw random sample of training data
- train a model using sample (e.g. decision tree)
- repeat N times (e.g. 25 times)
- bagged predictions: average predictions of N models



Ensemble Learning Intuition

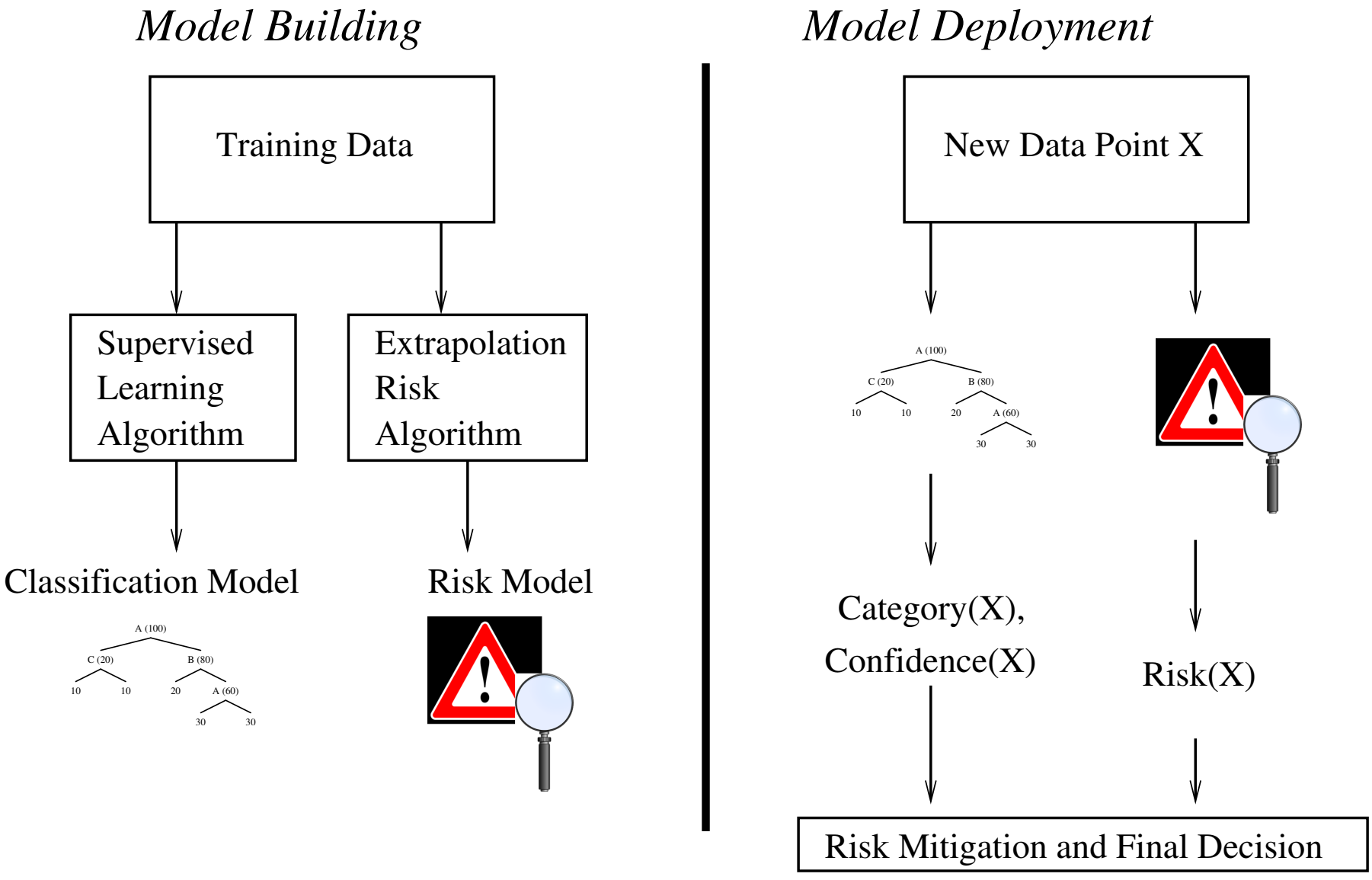
Ensemble machine learning: wisdom of crowds

Truth	**1**	**0**	**1**	**1**	**0**	**Accuracy**
Model 1	1	0	0	1	1	60%
Model 2	0	1	1	1	0	60%
Model 3	0	0	1	0	0	60%
Model 4	1	1	1	1	1	60%
Model 5	1	0	0	0	0	60%
Vote 1–5	1	0	1	1	0	100%

- No one model has to get it all right
- Performance of ensemble outperforms individuals
- Usually more reliable / robust
- Reduces variance

Outline

Approach: Intrinsic vs. Extrinsic Risk Estimation



Remoteness: Intrinsic Risk Score for Tree Ensembles

Data point z is *remote* with respect to class A if its average forest proximity to examples from A is low.

Remoteness(z) based on the closest class.

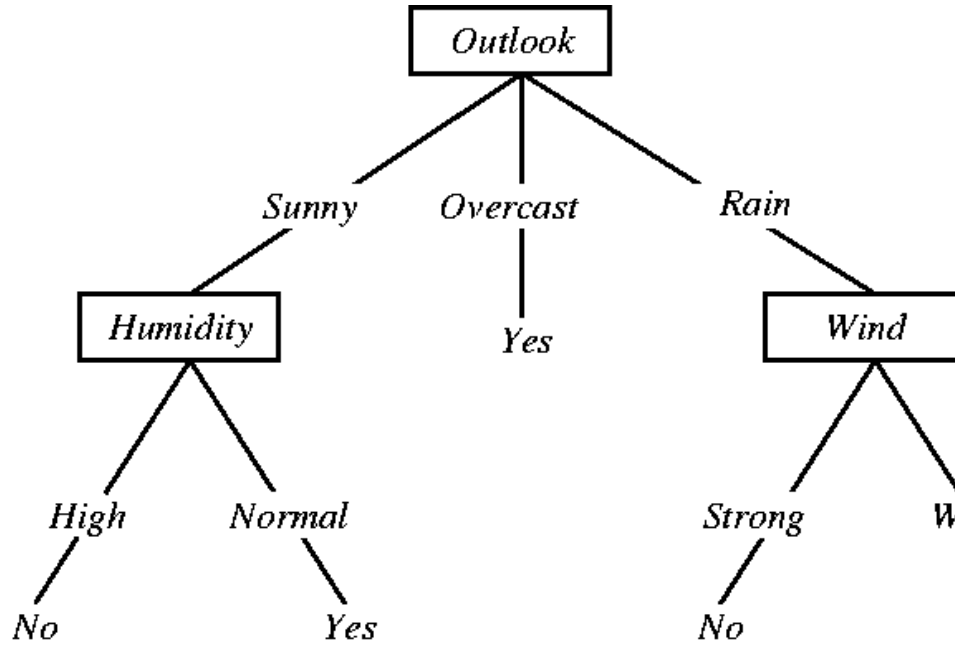
Remoteness: Intrinsic Risk Score for Tree Ensembles

Data point z is *remote* with respect to class A if its average forest proximity to examples from A is low.

Remoteness(z) based on the closest class.

Breiman's *forest proximity*:

- Points x and y are close to each other if they tend to land in the same leaves.
- Note:
  - non-Euclidean; invariant to monotonic scaling
  - categorical and numeric features
  - no triangle inequality



©Tom Mitchell, McGraw Hill, 1997

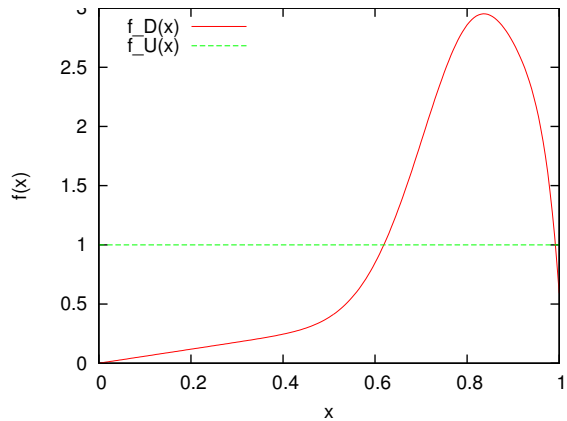
Extrapolation Risk Score

Following Hooker (2004), define extrapolation risk for data point x as

$$\text{Extrap}(x) = \frac{f_U(x)}{f_U(x) + f_D(x)}$$

- $f_U(x)$: data density at x assuming a uniform distribution
- $f_D(x)$: data density at x assuming the same distribution that generated the observed data D.

$\text{Extrap}(x) = 1$ for max. risk, and 0 for min. risk.

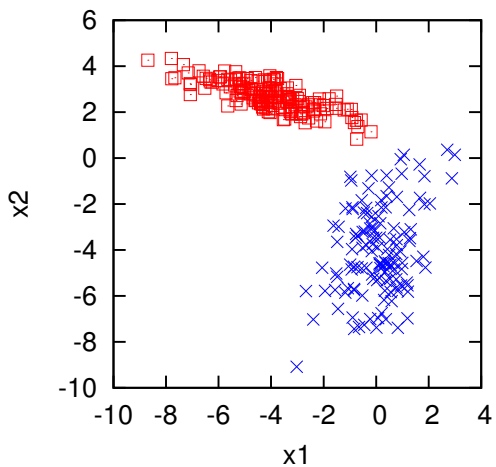


Confidence and Extrapolation Representation Trees (CERT)

Hooker (2004) proposed CERT models for estimating extrap. risk.

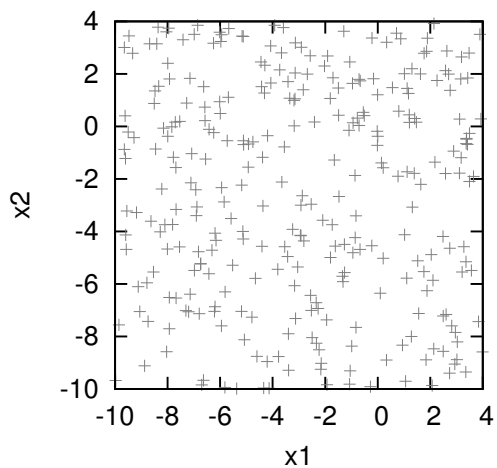
- Idea: frame as classification problem.

Class A
(all train data)



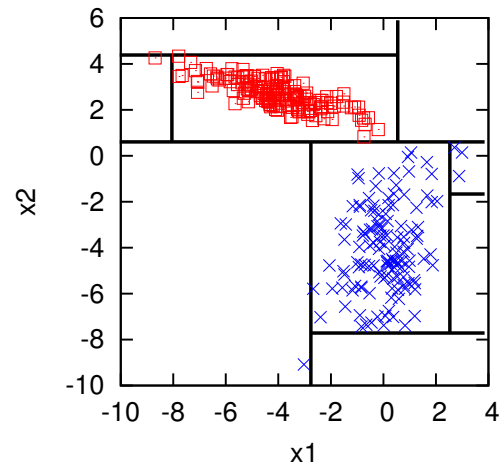
Vs.

Class B
(background)



$\Longrightarrow$

Result:



- Classification model predicts $\Pr(x \in \text{Class B}) \approx \mathit{Extrap}(x)$
- Decision tree learns bounding boxes.

CERT Insight: Avoid Uniform Sample

Problem:

$$\begin{aligned} \text{High dimensions} &\implies \text{sparsely sampled background} \\ &\implies \text{high variance} \end{aligned}$$

Solution: don't sample!

- Decision tree learning minimizes entropy of sub-regions R:

$$\text{Entropy}(R) = -p(A \mid R) \log_2 p(A \mid R) - p(B \mid R) \log_2 p(B \mid R)$$

with

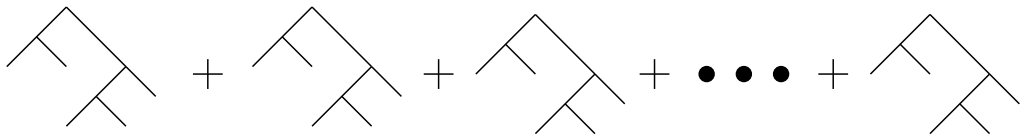
$$p(c \mid R) = \frac{N_c(R)}{N_A(R) + N_B(R)}$$

- Compute $N_B(R)$ analytically, using expected number of background points in R.

Outline

Research Questions

1. Benefits from ensemble of CERT models?



   - Better risk estimation?
   - Do you need to prune trees?

2. Remoteness vs. CERT Forests?
   (Intrinsic vs. Extrinsic)

3. Where do they break?

Experiment 1: Synthetic Data

Data

- Sample training points from mixture of 2D Gaussians.
- 250 points per mixture component.
- Try 1, 2, 3, and 4 components, with 10 random mixtures.

Model Fitting

- Train CERT model (baseline).
- Train bagged CERT model (100 trees).

Validation

- Compute true Extrap(x) across a uniform grid.
- Measure root mean squared error for model predictions at grid points.

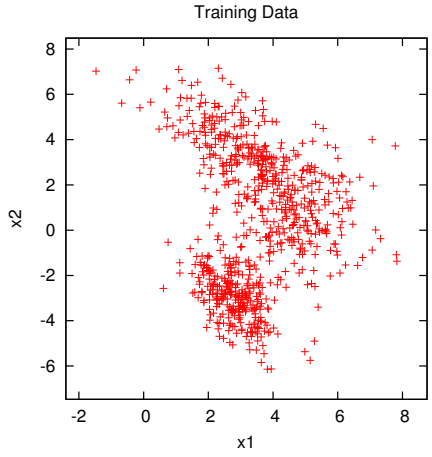
CERT Forest Beats CERT Tree

1 Gaussian		2 Gaussian		3 Gaussian		4 Gaussian	
tree	forest	tree	forest	tree	forest	tree	forest
0.170	0.083	0.161	0.108	0.190	0.107	0.164	0.108
0.166	0.089	0.152	0.100	0.172	0.091	0.134	0.075
0.148	0.093	0.198	0.099	0.181	0.109	0.189	0.134
0.181	0.097	0.154	0.091	0.155	0.095	0.127	0.077
0.142	0.087	0.180	0.112	0.148	0.094	0.133	0.089
0.184	0.113	0.165	0.089	0.179	0.090	0.170	0.098
0.185	0.107	0.179	0.104	0.138	0.083	0.139	0.082
0.173	0.085	0.244	0.206	0.164	0.097	0.172	0.099
0.191	0.085	0.266	0.213	0.205	0.149	0.158	0.135
0.203	0.114	0.201	0.091	0.188	0.111	0.130	0.084

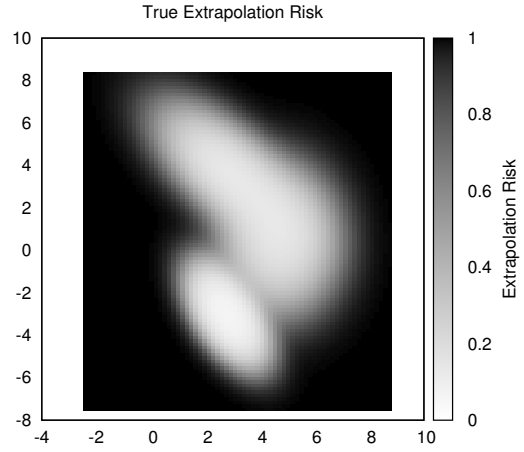
Error measure: root mean squared error. (Smaller is better.)

Pruning Needed to Prevent Overfitting

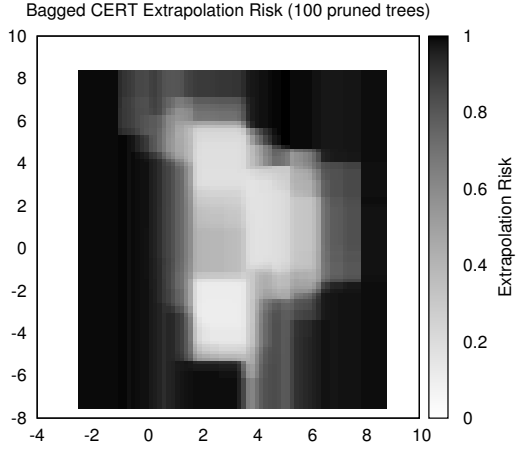
(a) Training Data (750 pts)



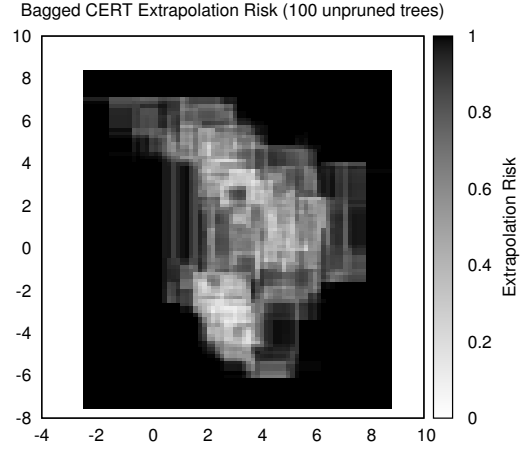
(b) Oracle



(c) Pruned. Avg Error = 0.095



(d) Unpruned. Avg Error = 0.207



Case Study I: Detect Novel NYT Topic

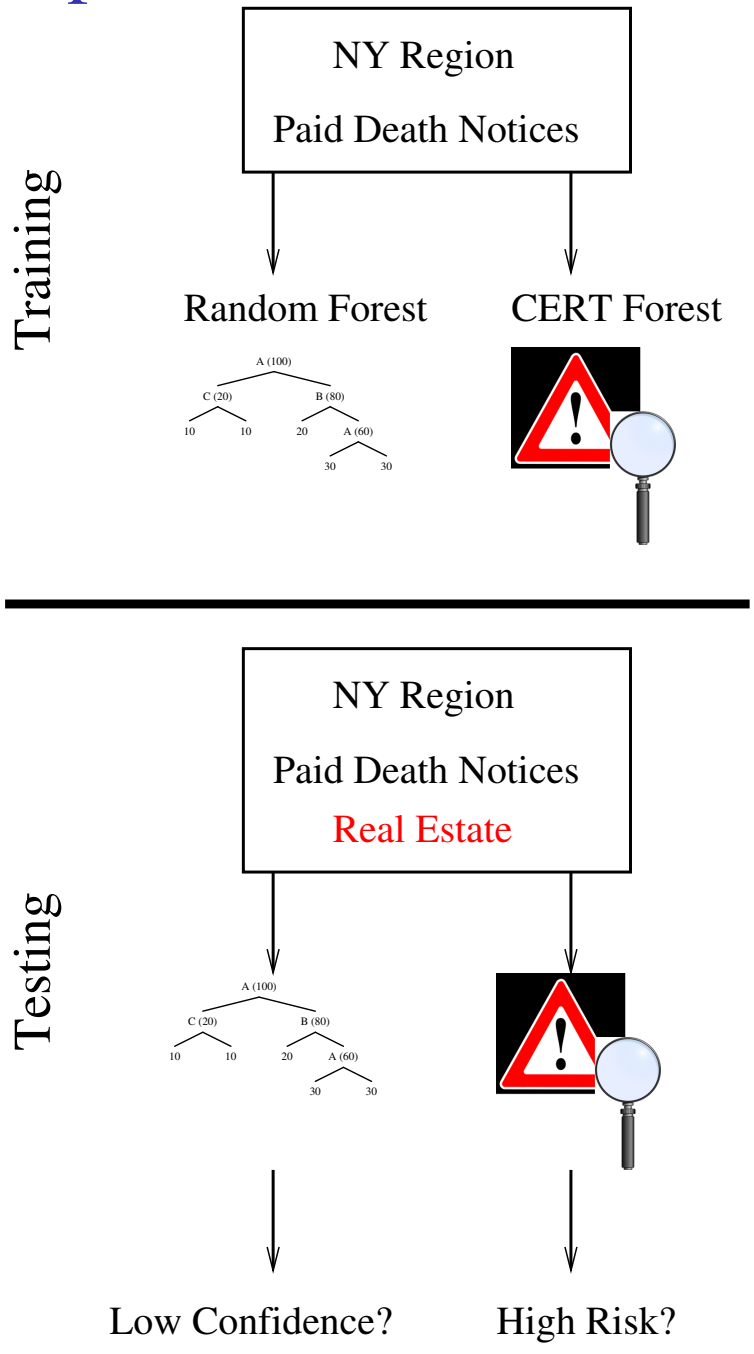
Experiment: Detect Novel NYT Topic

Data:

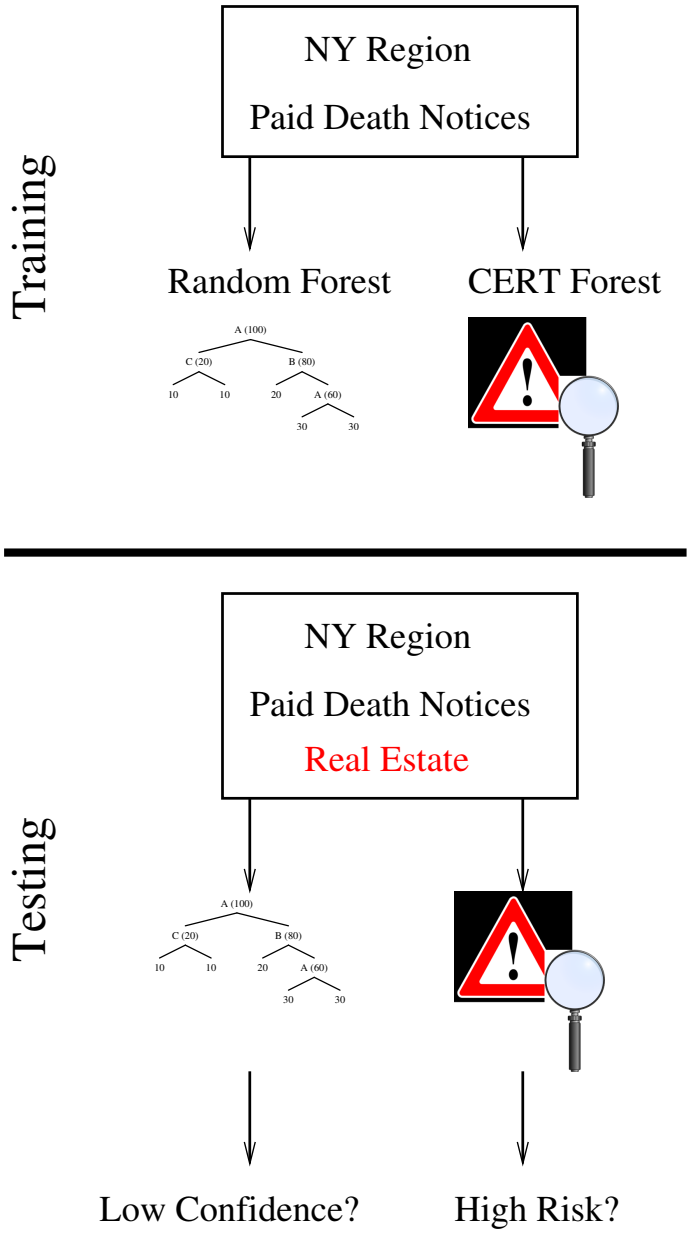
- 22,926 NYT articles
  - 48.9% NY Region
  - 48.6% Paid Death Notices
  - 2.4% Real Estate
- 9 numeric features (LSA)
- 1/2 train, 1/2 test

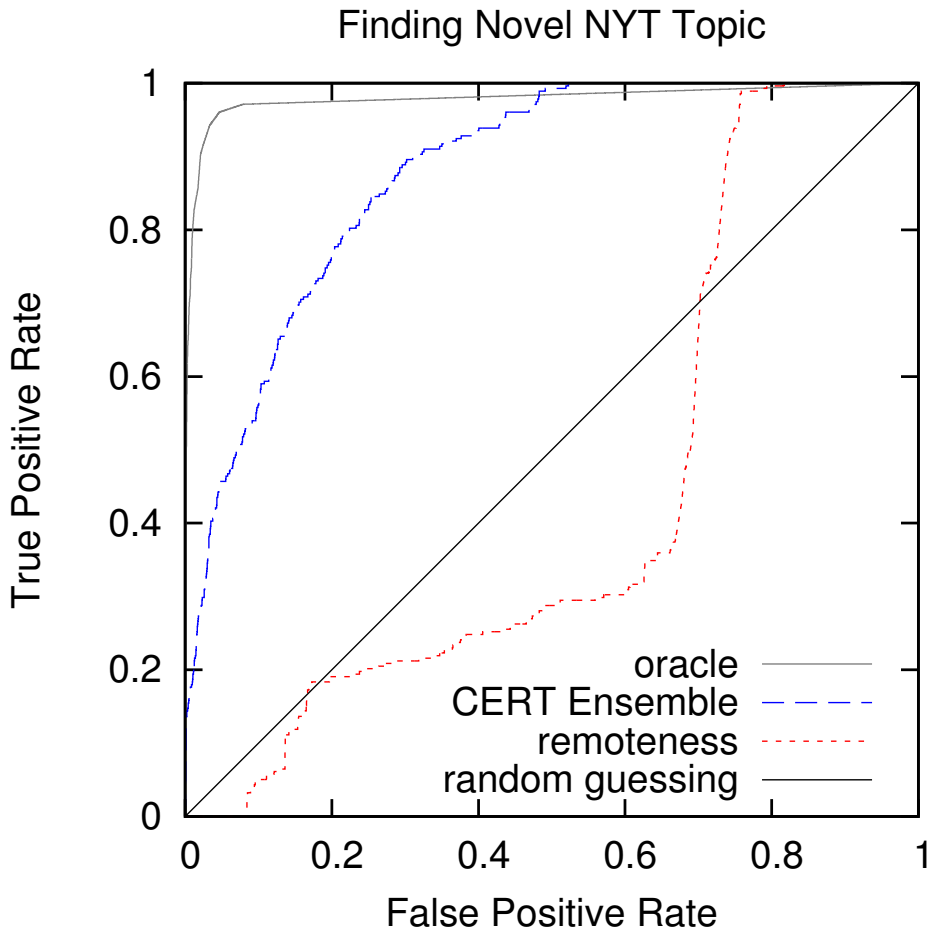
Experiment Design:

- Real estate topic omitted from training.
- Find real estate in testing?
  - Intrinsic: classifier confidence.
  - Extrinsic: risk model.



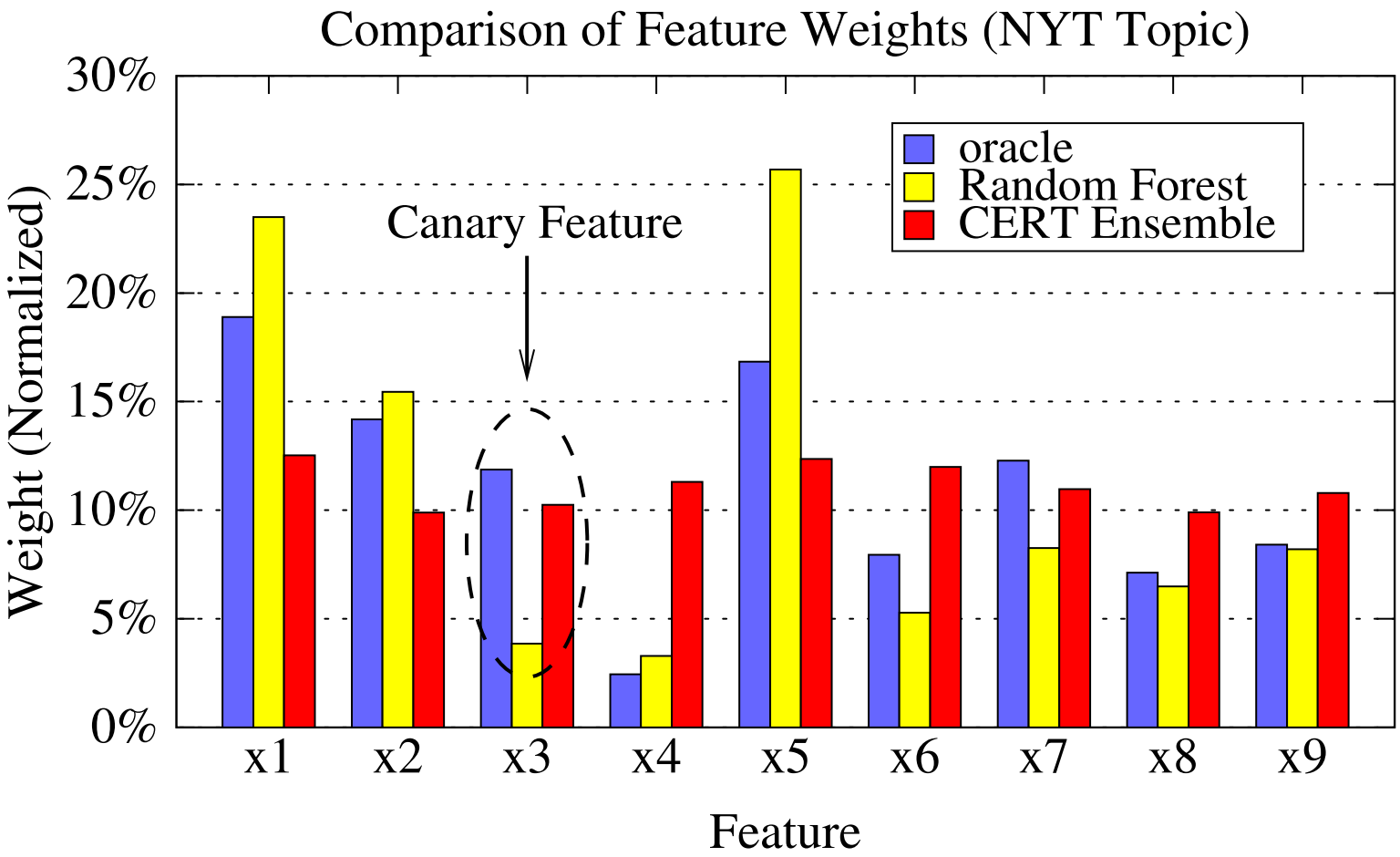
Take Away #1: Extrinsic Risk Model Needed





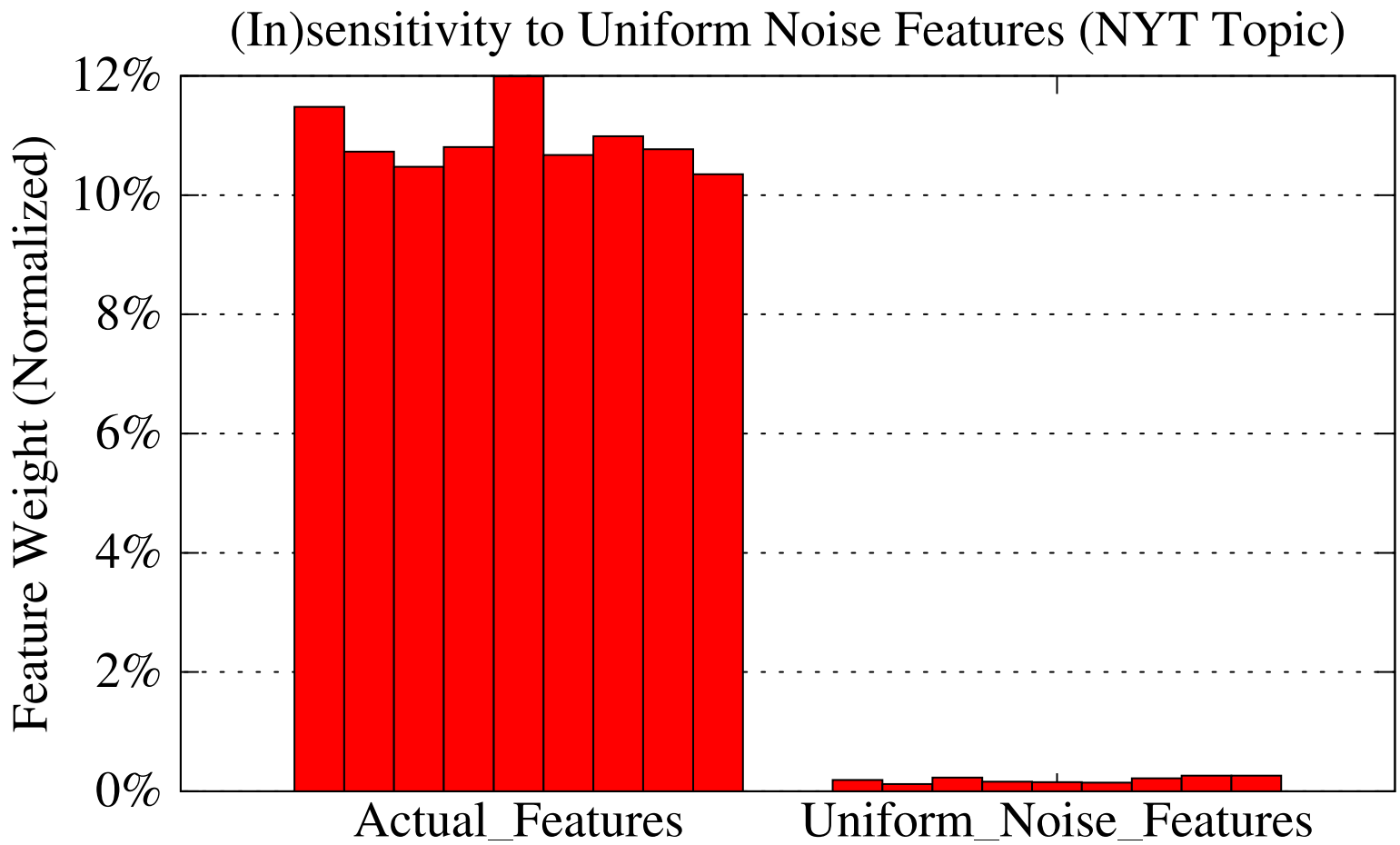
Canary Features

Classification model ignores feature x3
— which is important for finding the novel class.



CERT Forests Ignore Uniform Noise

Added 9 uniform noise features to NYT Novel Topic task.



Case Study II: Predict if EXE is Malware

Use Case: Predicting Reliability of Malware Classifier

Scenario: predict when model classifying EXE files might be wrong.[1]

Data:

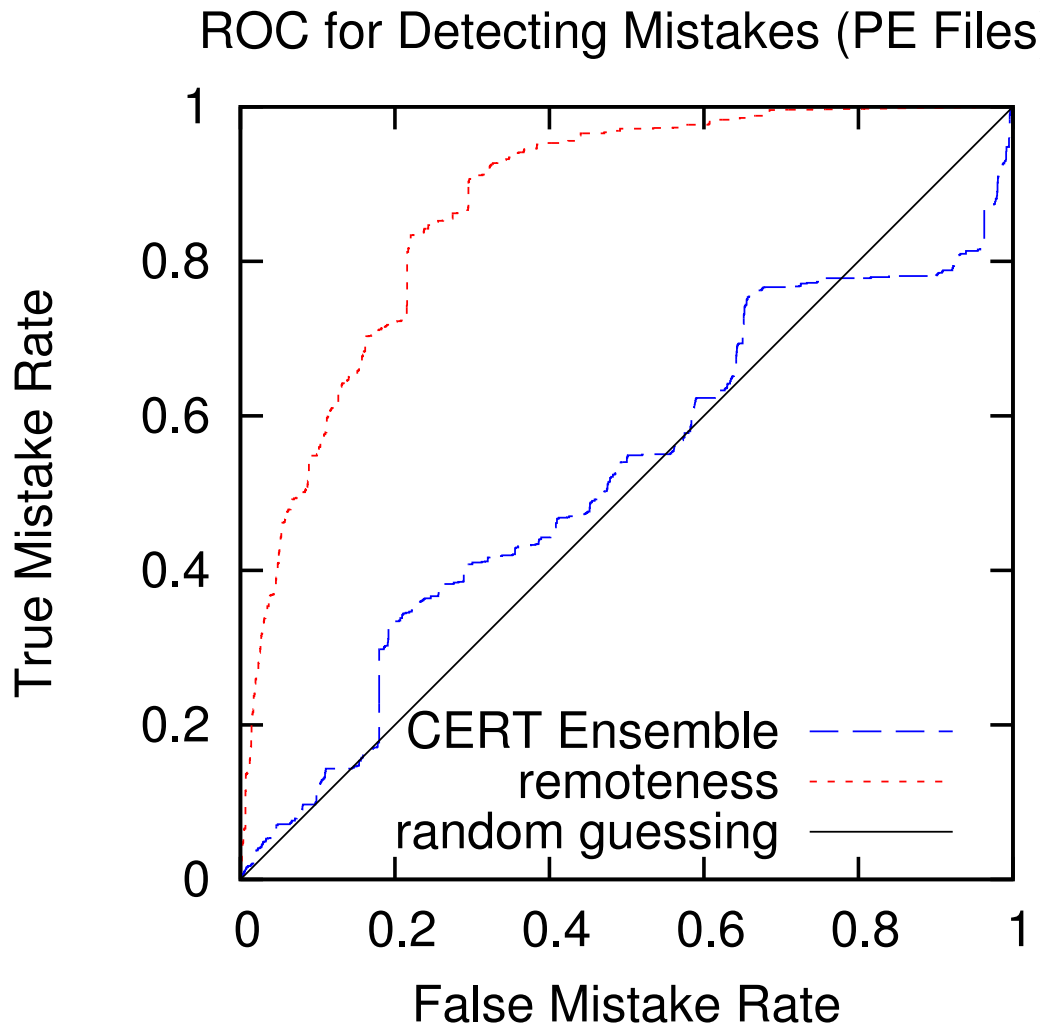
- Training Data: 2010
  - 18,588 examples
  - 44.8% malware
- Testing Data: 2011
  - 16,432 examples
  - 79.3% malware
- Extracted Features
  - 57 categorical features
  - 63 numeric features

Setup:

- Train classifier to predict goodware or malware.
- Train extrapolation model.
- Does classifier make mistakes on high risk test points?

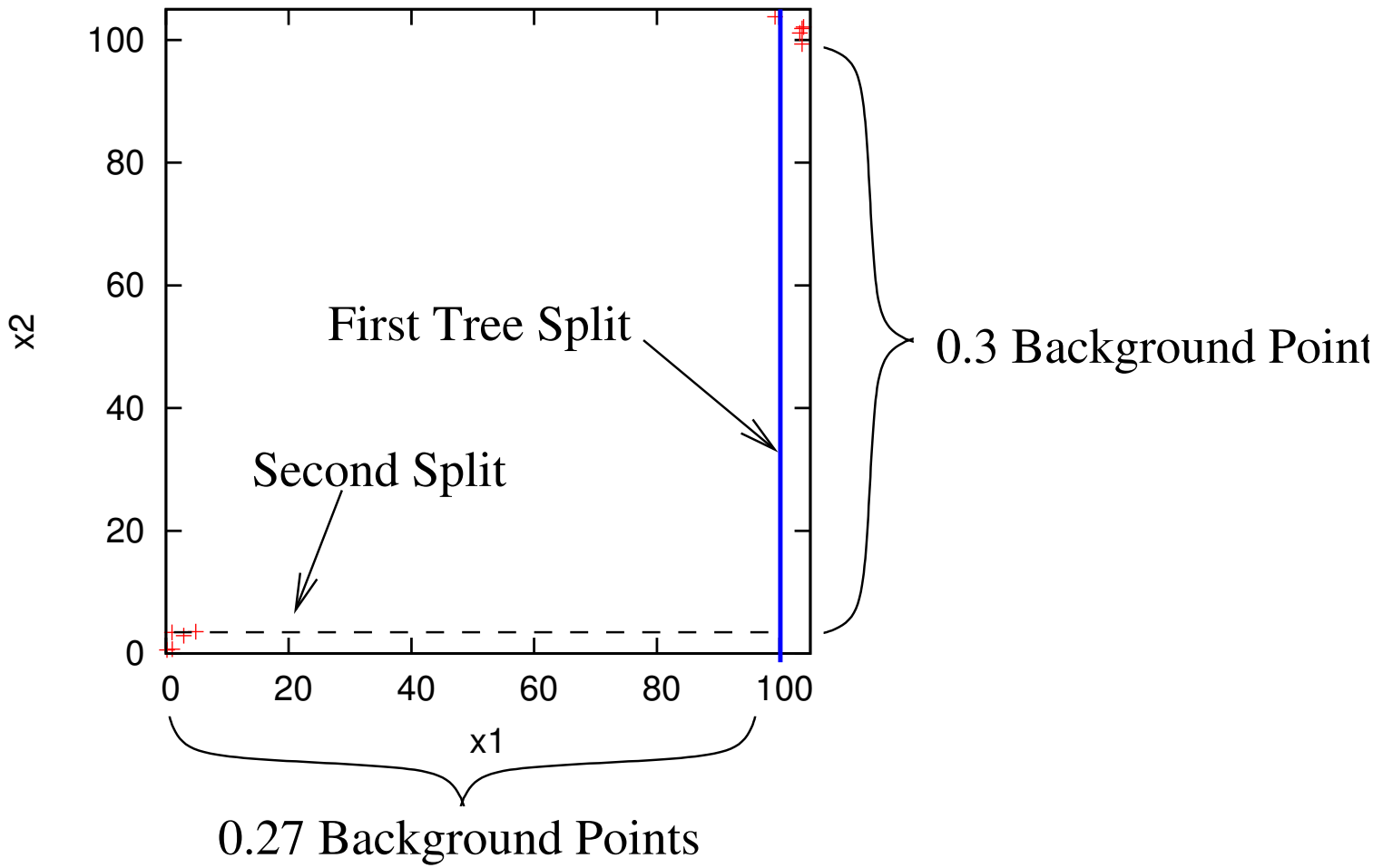
[1] Data from Ken Chiang, Michael Karres, and Levi Lloyd.

Take Away #2: Intrinsic Risk Needed, Also



Error Analysis for PE Task

CERT can prematurely declare points low-risk.



Intrinsic + Extrinsic = Better?

Algorithm 1: Simple Risk Combination Baseline

IR =intrinsic risk of x;

ER =extrinsic risk of x;

if *IR is high* **then**
    declare prediction risky;
else
    if *ER is high* **then**
        declare prediction risky
    else
        declare prediction safe

Conclusions & Next Steps

- Intrinsic and extrinsic risk metrics are complementary.
- Ensembles improve CERT's risk assessments.
- Yes, you should prune CERT.
- Characterized failure modes for CERT and remoteness score.

- Characterize types of problems each works well on?
- Benefit from combining?
- Exploring possible fixes for premature stopping in CERT.

Questions?

mamunso@sandia.gov

Bibliography I


Leo Breiman.
Bagging predictors.
Machine Learning, 24(2):123–140, 1996.


Giles Hooker.
Diagnosing extrapolation: Tree-based density estimation.
In Won Kim, Ronny Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 569–574, New York, NY, USA, 2004. ACM.


Parikshit Ram and Alexander G. Gray.
Density estimation trees.
In Chid Apté, Joydeep Ghosh, and Padhraic Smyth, editors, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 627–635, New York, NY, USA, 2011. ACM.

Backup Slides

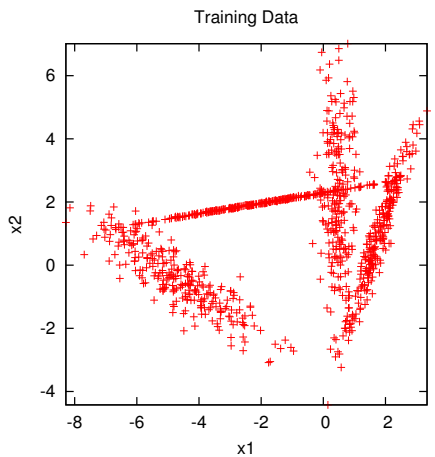
Aside

Cool. But wouldn't it be better to do density estimation from first principles?

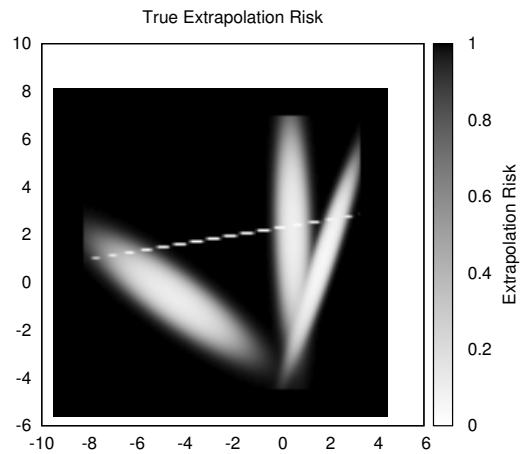
CERT vs DET

Compare density estimation trees [3] to CERT. Default params.

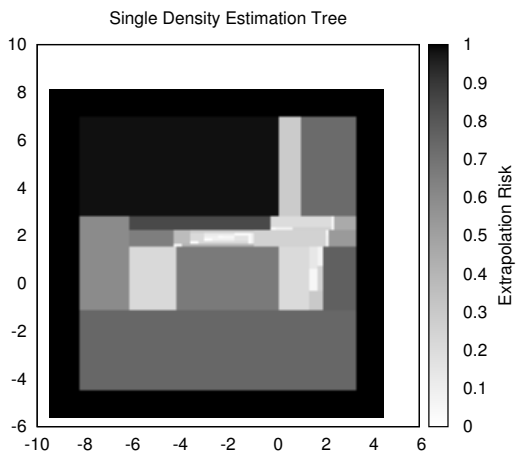
(e) Training Data (1000 pts)



(f) Oracle



(g) DET



(h) CERT