SAND2012-6949P

Estimating Extrapolation Risk in Supervised Machine Learning

Should I trust *this* prediction?

Art Munson, Philip Kegelmeyer

Sandia National Laboratories

August 23, 2012

National Nuclear Security Administration

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.

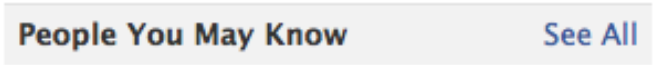
Sandia National Laboratories

Too Much Traffic to Monitor Manually

Maybe Machine Learning Can Help...

Web Search

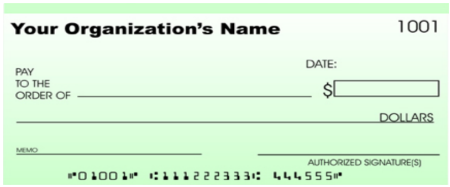
Friend Recommendations



Pose Recognition in Kinect

Winning Jeopardy

Reading Bank Checks



The IID Assumption in Machine Learning

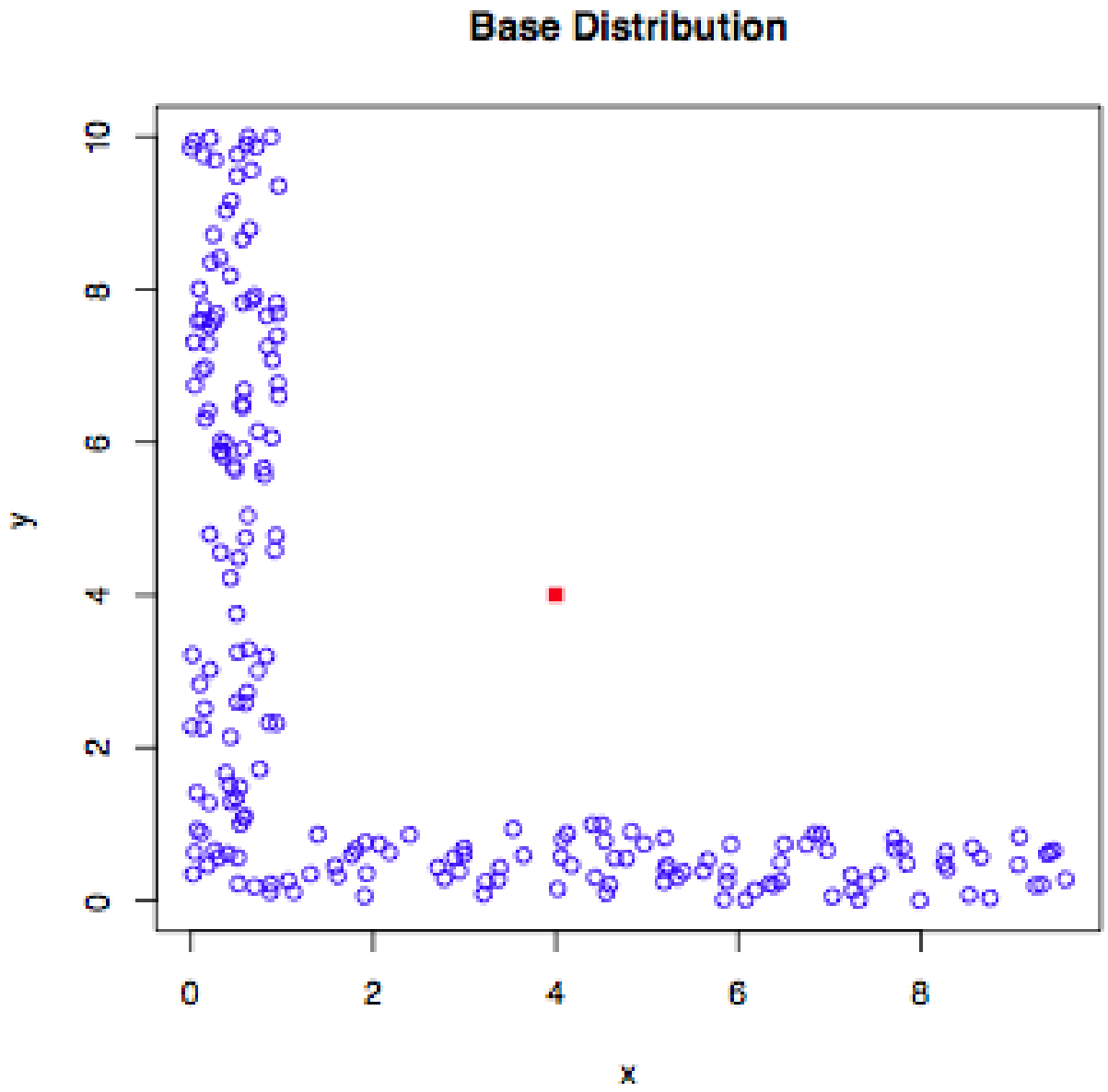
IID = **I**ndependent and **I**dentically **D**istributed
Assumes future data looks like past data.

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.

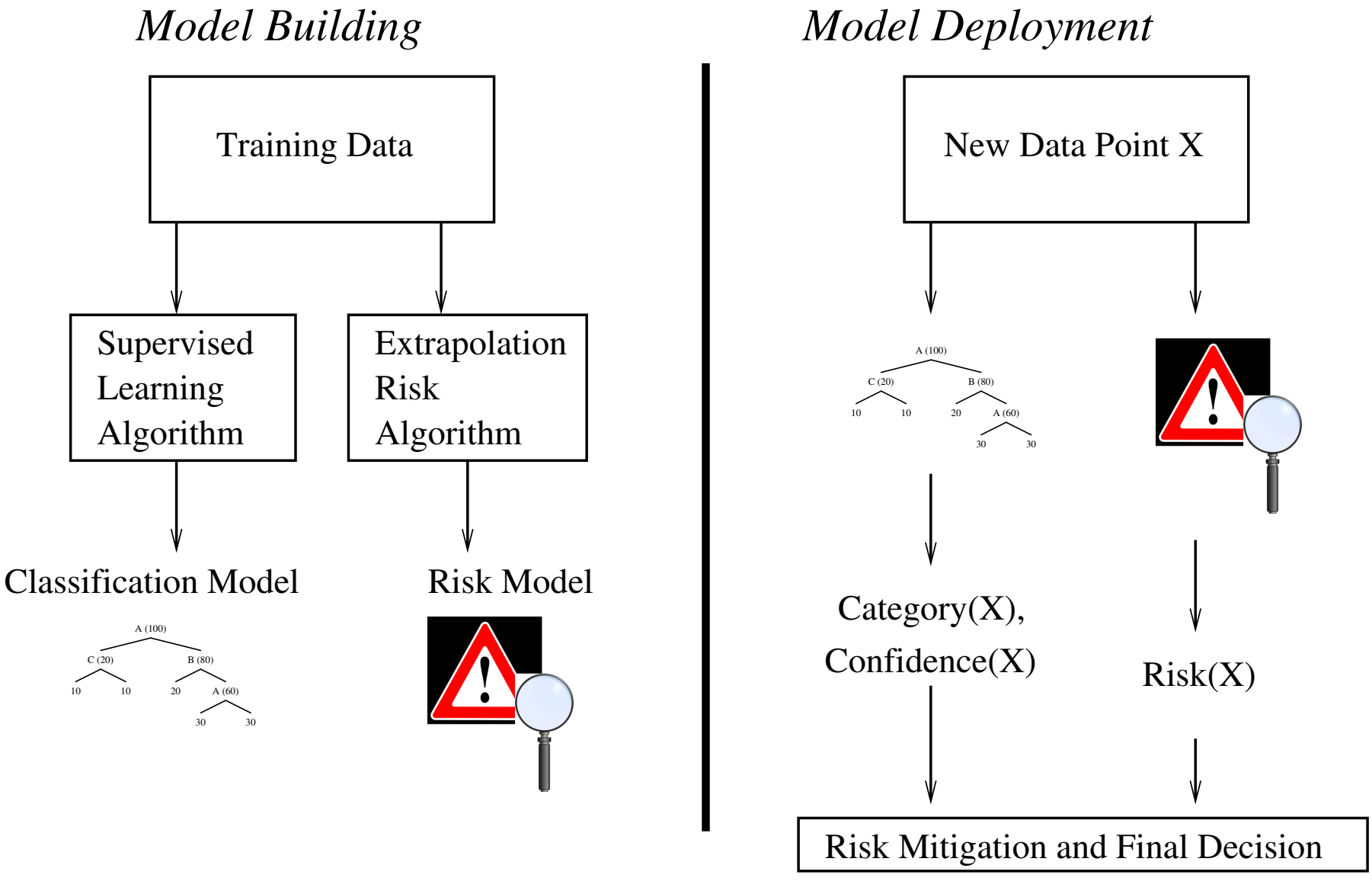
A Toy Example



Source: Hooker (2004).

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

Approach: Intrinsic vs. Extrinsic Risk Estimation



Digression: Ensemble Learning

Ensemble machine learning: wisdom of crowds / committee of experts

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

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.

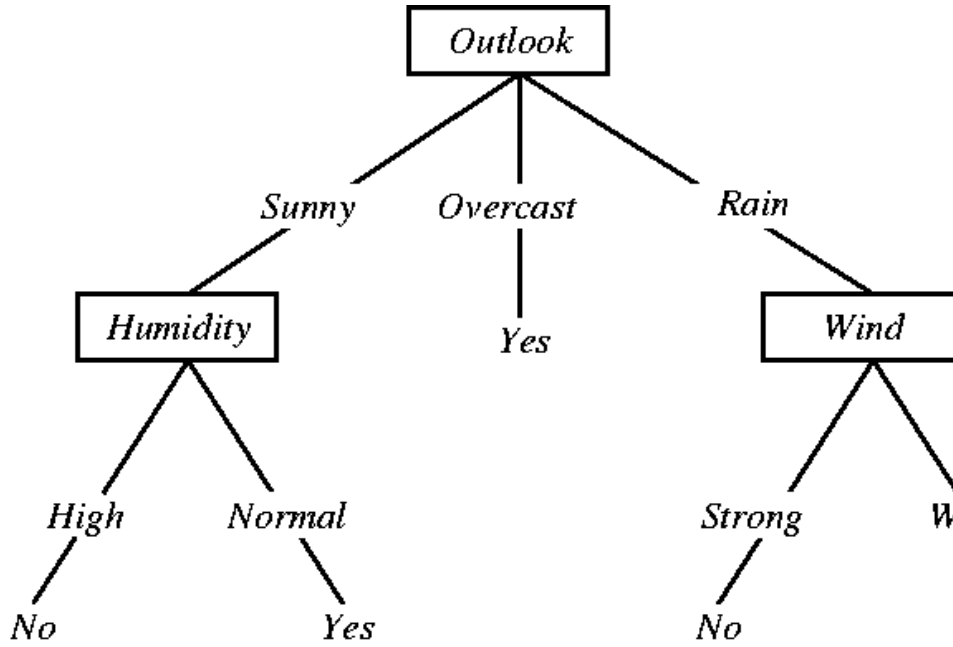
Remoteness: Intrinsic Risk Score for Tree Ensembles

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

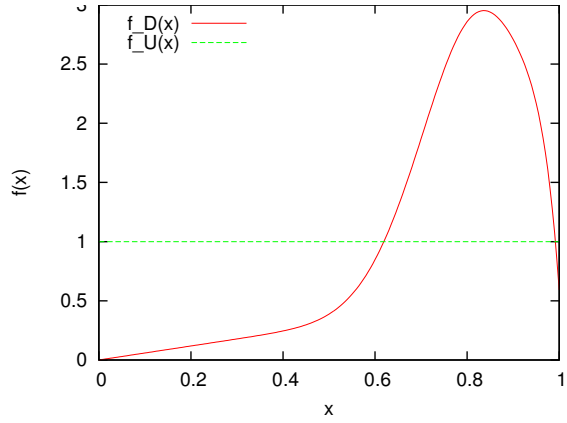
Extrapolation Risk Score

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.

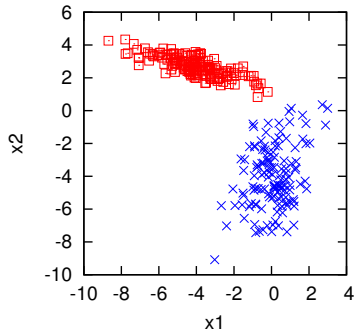


Confidence and Extrapolation Representation Trees (CERT)

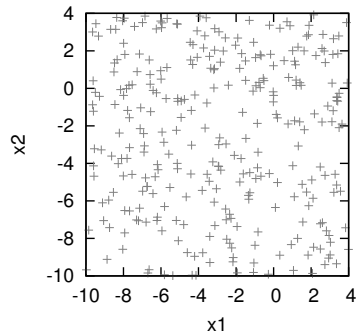
Hooker (2004) proposed CERT models for estimating extrap. risk.

- Idea: frame as classification problem.

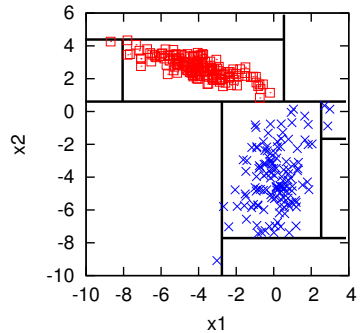
Class A (all train data)



Class B (background)



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

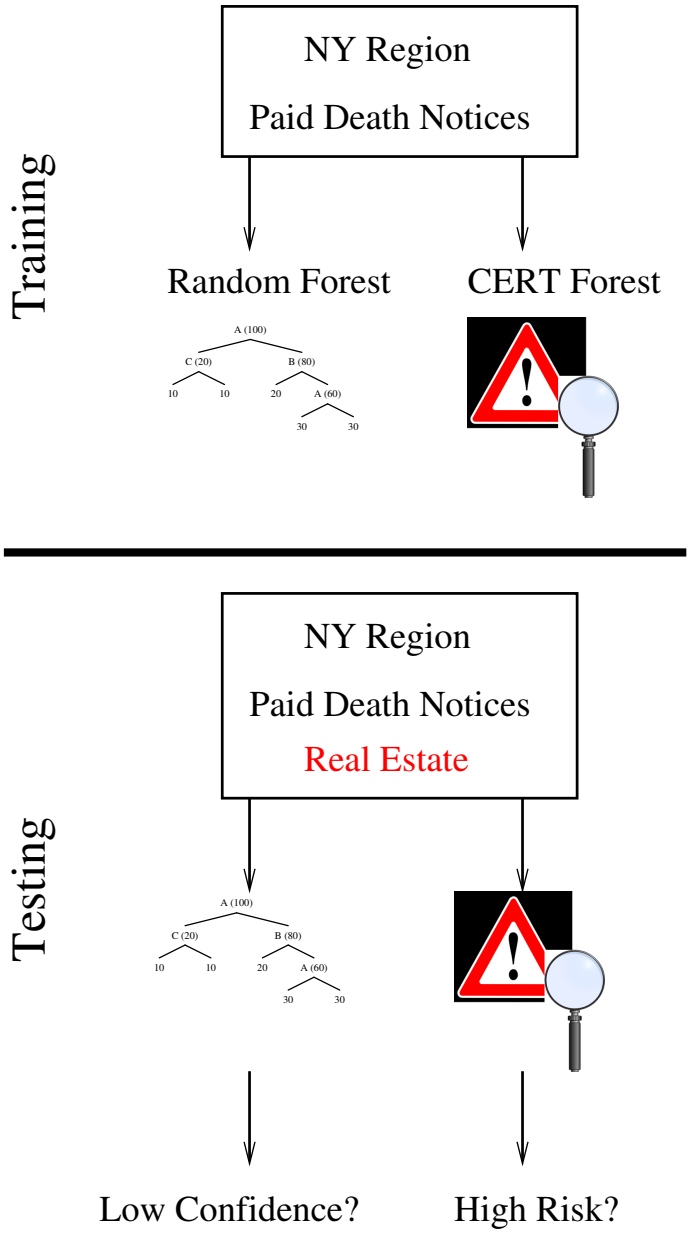


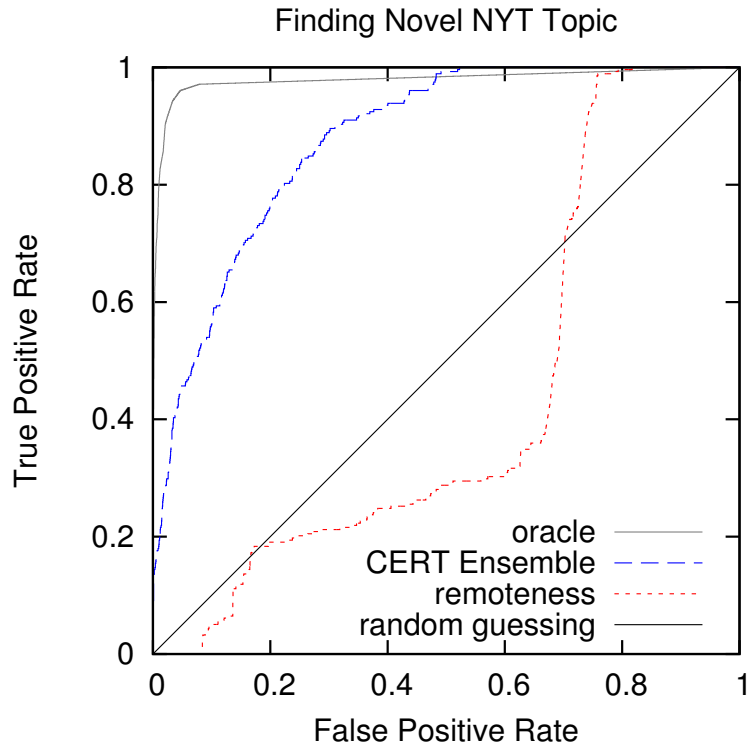
- Analytically compute expected # background points in a region.

Research Questions

1. Benefits from ensemble of CERT models?
   - A: Ensemble consistently improves risk estimation.
   - A: Pruning really is needed.
2. Remoteness vs. CERT Forests?
   (Intrinsic vs. Extrinsic)
3. Limitations?

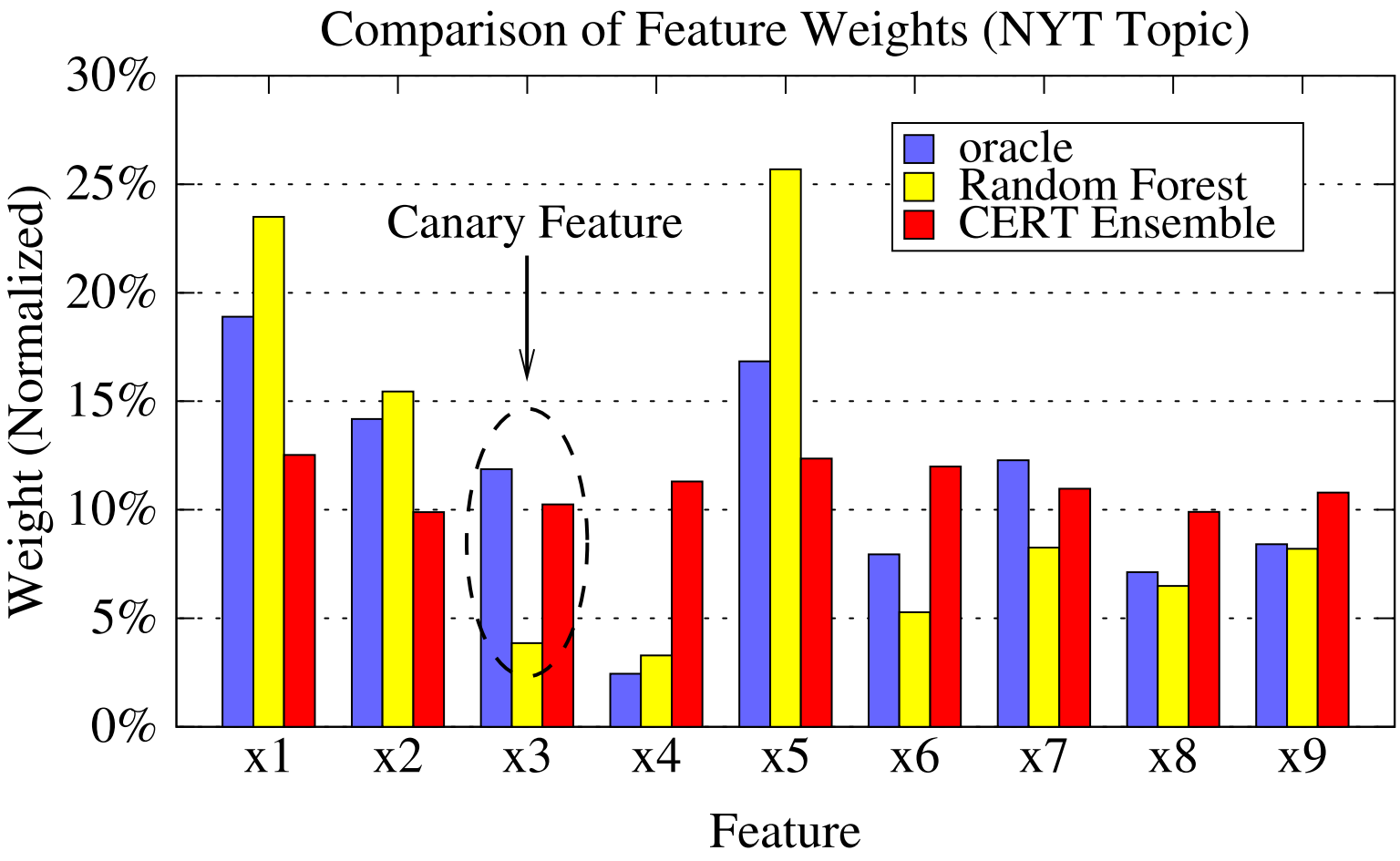
Take Away #1: Extrinsic Risk Model Needed





Canary Features

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



Anecdotal Success on Sandia Data

Task:

- Binary classification with $O(100)$ features.
- Existing SVM classifier with good accuracy, but trouble with rare anomalies.

Re-filtered SVM output using risk model:

- Fit CERT Forest using large unlabeled corpus.
- Checked predictions with high extrap. risk where SVM had high confidence.
- 70 of 75 points checked were outliers requiring human analyst.

Impact: group now uses separate models for classification and outlier detection, and both feed into analyst decision support tool.

Take Away #2: Intrinsic Risk Needed, Also

Task: predict if Windows binary file is malware or not.[1]

Training Data: 2010

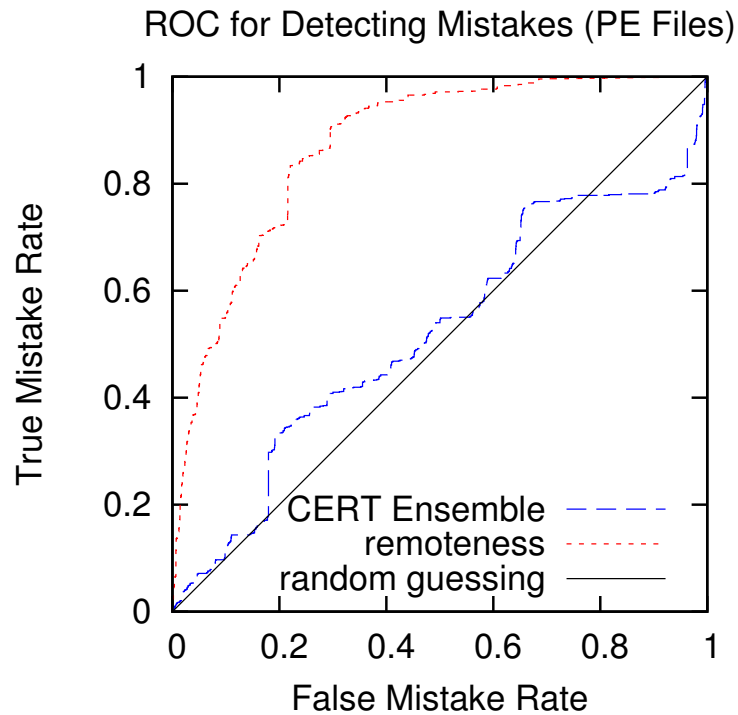
- 18,588 examples
- 44.8% malware

Testing Data: 2011

- 16,432 examples
- 79.3% malware

Setup:

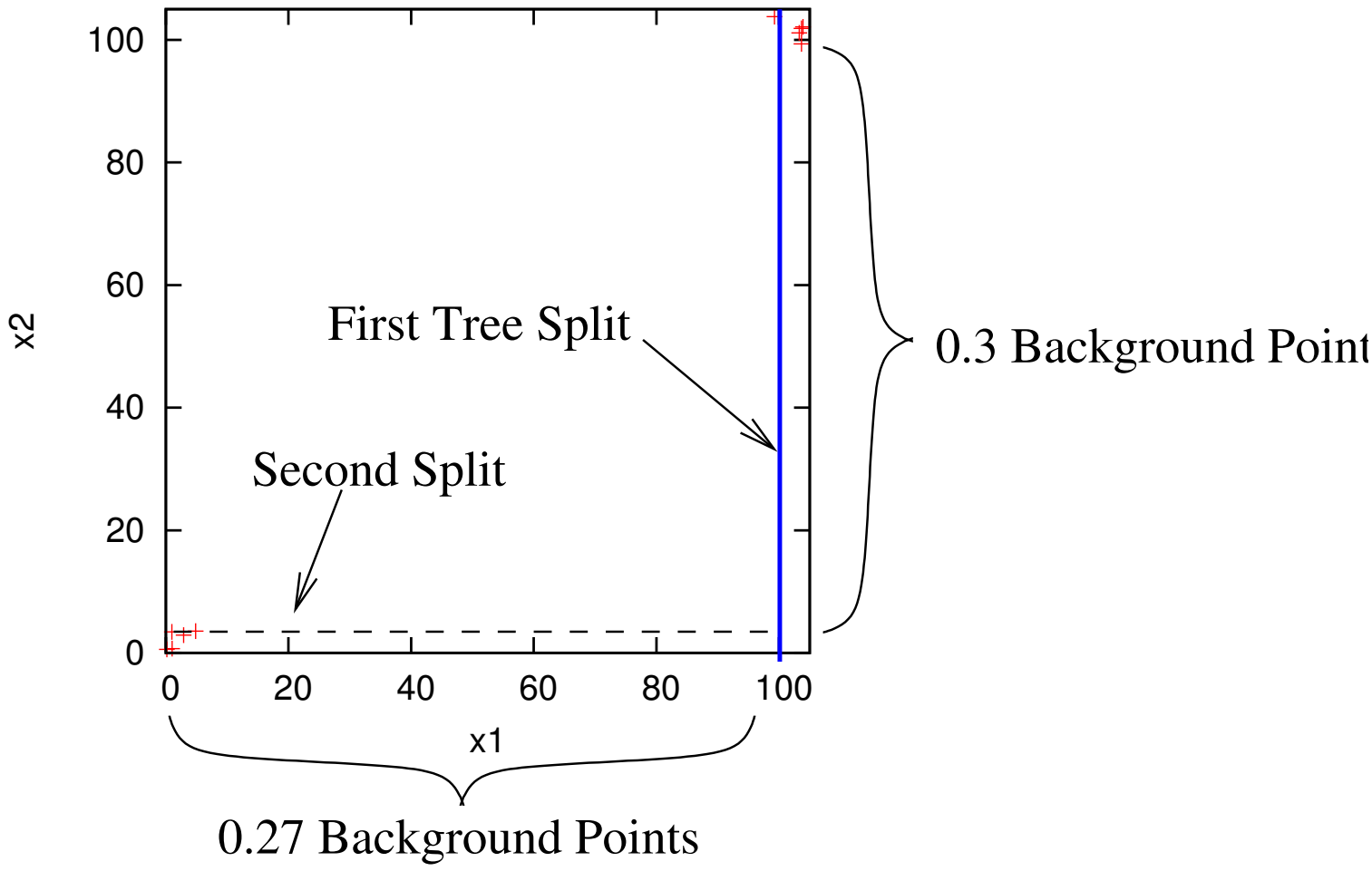
- Train classifier: goodware vs malware
- Estimate risk (test)
- Risk correlates with classifier mistakes?



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

Error Analysis for PE Task

CERT can prematurely declare points low-risk.



Conclusions & Next Steps

- Intrinsic and extrinsic risk metrics are complementary.
- Ensembles improve CERT's risk assessments.
- 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


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.