

# Situational Awareness at Internet Scale: Detection of Extremely Rare Crisis Periods



Philip Kegelmeyer, [wpk@sandia.gov](mailto:wpk@sandia.gov), [csmr.ca.sandia.gov/~wpk](http://csmr.ca.sandia.gov/~wpk)

David Cieslak, [dcieslak@cse.nd.edu](mailto:dcieslak@cse.nd.edu), [www.nd.edu/~dcieslak/](http://www.nd.edu/~dcieslak/)

Larry Hall, Kevin Bowyer, and the USF AVATAR project, [morden.csee.usf.edu/avatar](http://morden.csee.usf.edu/avatar)

(USF CSE Symposium, presented by Philip Kegelmeyer, September 24, 2008)

## The Conclusion

When building **ensembles** of **decision trees** to classify extremely **skew** data, such as **Border Gateway Protocol (BGP)** data, use the **Hellinger Distance** metric, rather than the traditional **Infogain** metric, for increased accuracy.

(And be sure to use a **sensible metric** for “accuracy”.)

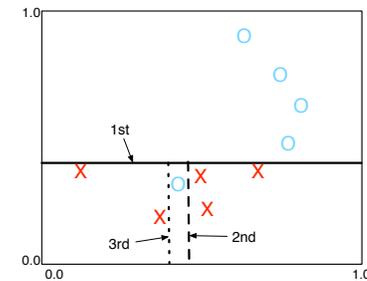
## Invented Training Data, for Search Relevance

Queries	Relevant? Truth	PageRank $a_1$	Fresh? $a_2$	Unique? $a_3$	...	Distinct? $a_K$
$q_1$	Yes	12	1003	0.97	...	0.12
$q_2$	Yes	99	2	0.33	...	0.03
$q_3$	No	3	27	0.12	...	0.13
$q_4$	Yes	16	183	0.08	...	0.58
$q_5$	No	17	665	0.36	...	0.64
$q_6$	No	44	1212	0.29	...	0.42
$q_7$	No	42	24	0.33	...	0.88
$q_8$	Yes	78	42	0.44	...	0.52
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$q_N$	No	12	3141	0.92	...	0.17

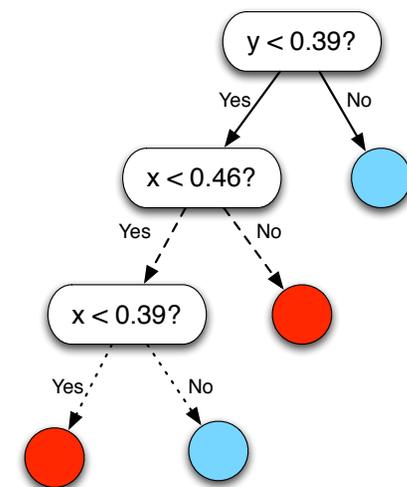
# Supervised Machine Learning Overview

Also known as: pattern recognition, statistical inference, data mining.

- Input: “ground truth” data.
  - Samples, with attributes, and *labels*.
  - Example: search result data
    - \* Samples: a query string
    - \* Attributes: features of the search
    - \* Labels: “relevant”, “irrelevant”
- Apply suitable method:  
decision trees, neural nets, SVMs.
- Output:  
rules for labeling new, *unlabeled* data.  
Equivalently:  
a partitioning of attribute space.



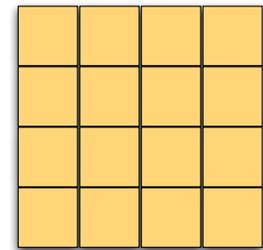
Attribute space partitioned.



Decision tree representation.

# Machine Learning, Before Ensembles

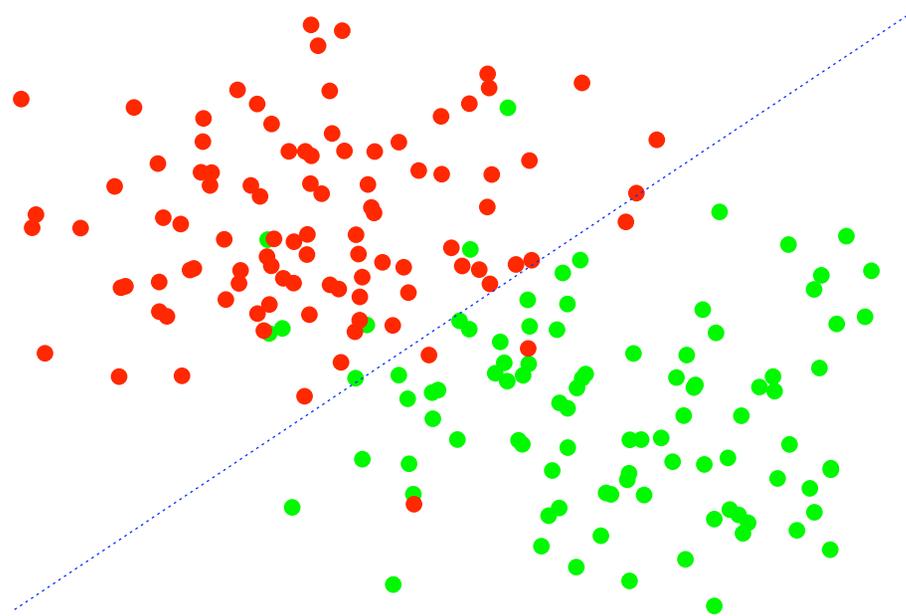
**Traditional:** Use 100% of training data to build a sage.



Sage sees all the data.

## Note: Even Sage is Not Perfectly Accurate

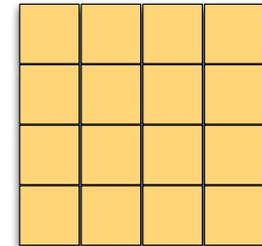
Class distributions can overlap inextricably.



“Bayes error” is the best any classifier can do.

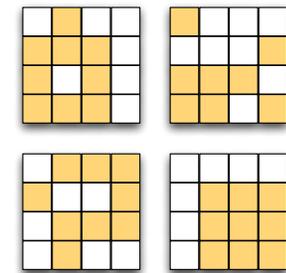
# Machine Learning, With Ensembles

**Traditional:** Use 100% of training data to build a sage.



Sage sees all the data.

**Ensembles:** Use randomized 100% of training data to build an expert. Repeat to build many experts. Vote them.



Each expert sees 2/3rds of the data.

The experts beat the sage[1]!

## Reminder: The Unaltered Training Data

Queries	Relevant? Truth	PageRank $a_1$	Fresh? $a_2$	Unique? $a_3$	...	Distinct? $a_K$
$q_1$	Yes	12	1003	0.97	...	0.12
$q_2$	Yes	99	2	0.33	...	0.03
$q_3$	No	3	27	0.12	...	0.13
$q_4$	Yes	16	183	0.08	...	0.58
$q_5$	No	17	665	0.36	...	0.64
$q_6$	No	44	1212	0.29	...	0.42
$q_7$	No	42	24	0.33	...	0.88
$q_8$	Yes	78	42	0.44	...	0.52
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$q_N$	No	12	3141	0.92	...	0.17

## First Expert Sees A Sampling With Replacement

Queries	Relevant? Truth	PageRank $a_1$	Fresh? $a_2$	Unique? $a_3$	...	Distinct? $a_K$
$q_1$	Yes	12	1003	0.97	...	0.12
$q_2$	Yes	99	2	0.33	...	0.03
$q_2$	Yes	99	2	0.33	...	0.03
$q_4$	Yes	16	183	0.08	...	0.58
$q_4$	Yes	16	183	0.08	...	0.58
$q_5$	No	17	665	0.36	...	0.64
$q_8$	Yes	78	42	0.44	...	0.52
$q_9$	No	59	7012	0.37	...	0.23
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$q_{N-1}$	Yes	36	1812	0.47	...	0.17

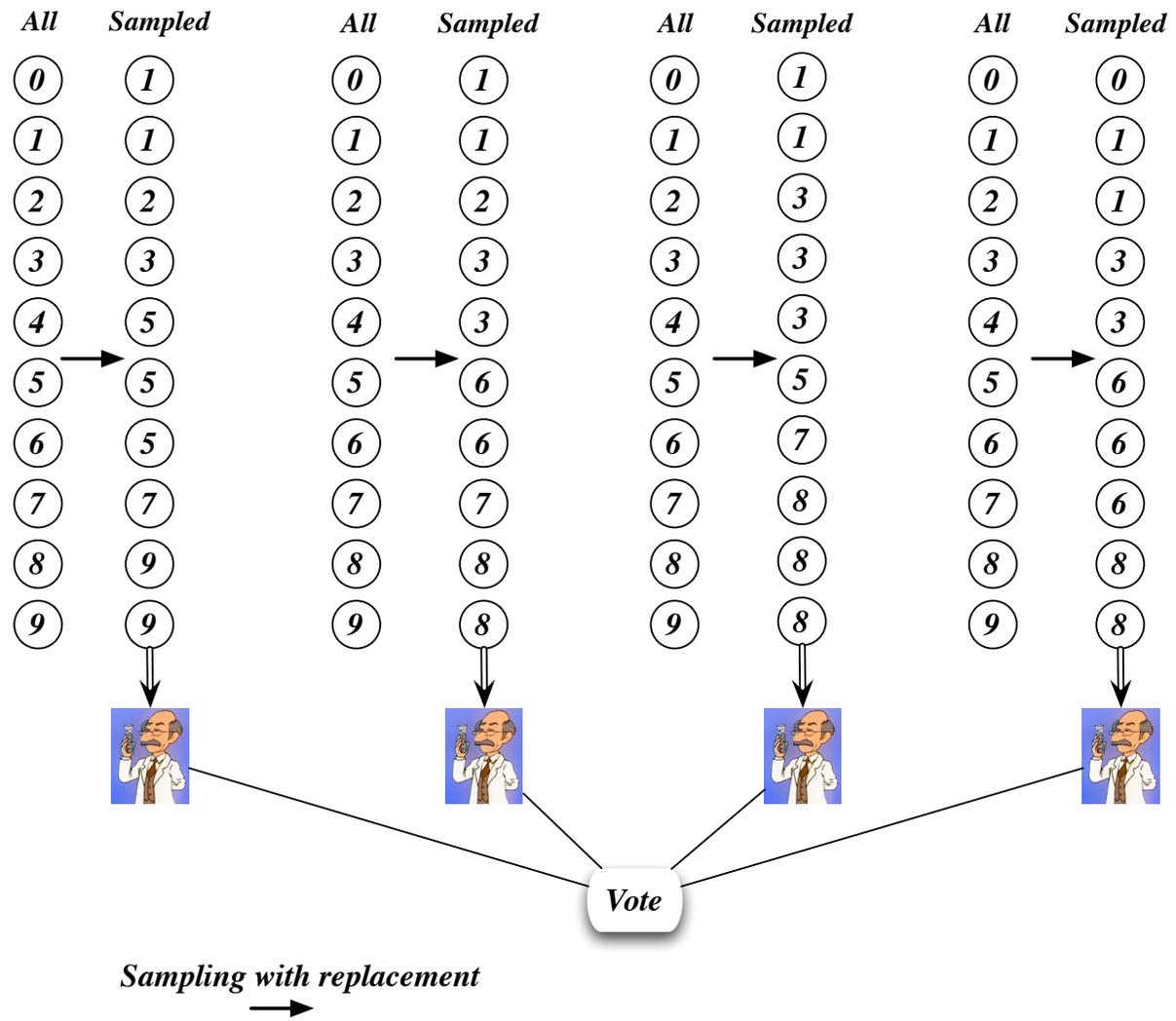
$q_2$  and  $q_4$  are repeated;  $q_3$  and others are missing.

## Second Expert Sees A Different Sampling

Queries	Relevant? Truth	PageRank	Fresh?	Unique?	...	Distinct?
		$a_1$	$a_2$	$a_3$	...	$a_K$
$q_1$	Yes	12	1003	0.97	...	0.12
$q_1$	Yes	12	1003	0.97	...	0.12
$q_2$	Yes	99	2	0.33	...	0.03
$q_3$	No	3	27	0.12	...	0.13
$q_3$	No	3	27	0.12	...	0.13
$q_3$	No	3	27	0.12	...	0.13
$q_6$	No	44	1212	0.29	...	0.42
$q_8$	Yes	78	42	0.44	...	0.52
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$q_N$	No	12	3141	0.92	...	0.17

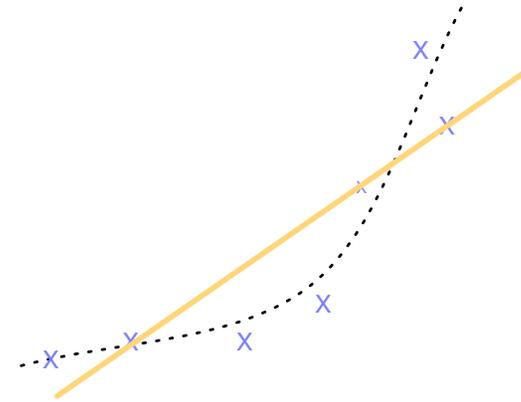
$q_3$  is repeated;  $q_4$  and others are missing.

# “Bagging” is the Formal Name for This Method

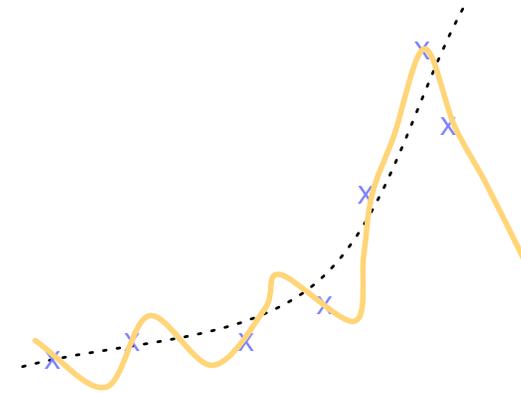


# Why Do Ensembles Work? (A)

- A statistical model is a *noisy* model of reality.
- Bias error:  
Model too simple, underfits.
- Variance error:  
Model too complex, overfits.
- Bias/variance is a trade-off.
- Ensembles:
  - Use methods with low bias...  
but high variance ...  
and average to reduce variance!
- Result:  
low bias error *and* low variance error.  
No hand tuning needed.



Too simple a model underfits the data.



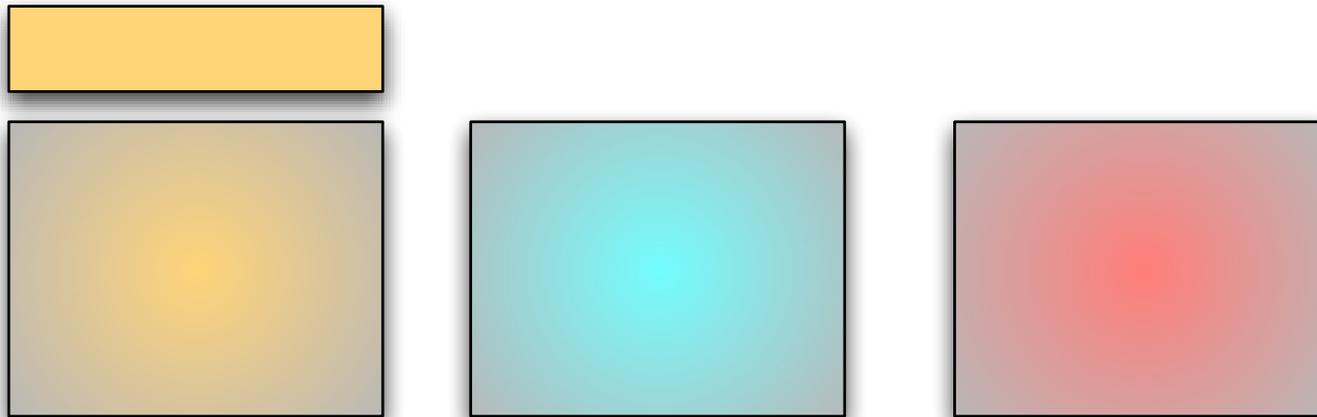
Too complex a model overfits the data.

## Why Do Ensembles Work? (B)

One key is *diversity* [5].

Imagine: three classes, each expert only 10% accurate, and when wrong, chooses at random among the three classes.

Then the crowd of experts is perfectly, 100% accurate!



One group of unconfused experts amid the foggy error.

Note: diverse, *random* error is difficult to achieve[2].

## The Conclusion

When building **ensembles** of **decision trees** to classify extremely **skew** data, such as **Border Gateway Protocol (BGP)** data, use the **Hellinger Distance** metric, rather than the traditional **Infogain** metric, for increased accuracy.

(And be sure to use a **sensible metric** for “accuracy”.)

# Skew vs. Supervised Machine Learning

When data is skew,  
the safe choice is  
“accurate” —  
and useless.



Minority class overwhelmed.

# Accuracy ( $A$ ) vs Class-Averaged Accuracy ( $A_c$ )

Classification confusion matrices.

		Truth	
		Q	W
Q		80	10
W		20	90

.....

$$\begin{aligned} A &= \frac{80 + 90}{100 + 100} \\ &= \frac{170}{200} = \mathbf{0.85} \end{aligned}$$

$$\begin{aligned} A_c &= \frac{1}{2} \left( \frac{80}{100} + \frac{90}{100} \right) \\ &= \mathbf{0.85} \end{aligned}$$

Balanced Data

# Accuracy ( $A$ ) vs Class-Averaged Accuracy ( $A_c$ )

Classification confusion matrices.

		Truth	
		Q	W
Q	80	10	
W	20	90	

---


$$\begin{aligned}
 A &= \frac{80 + 90}{100 + 100} \\
 &= \frac{170}{200} = 0.85
 \end{aligned}$$

$$\begin{aligned}
 A_c &= \frac{1}{2} \left( \frac{80}{100} + \frac{90}{100} \right) \\
 &= 0.85
 \end{aligned}$$

Balanced Data

		Truth	
		Q	W
Q	800	10	
W	200	90	

---


$$\begin{aligned}
 A &= \frac{800 + 90}{1000 + 100} \\
 &= \frac{890}{1100} = \mathbf{0.81}
 \end{aligned}$$

$$\begin{aligned}
 A_c &= \frac{1}{2} \left( \frac{800}{1000} + \frac{90}{100} \right) \\
 &= \mathbf{0.85}
 \end{aligned}$$

Skew Data

# Accuracy ( $A$ ) vs Class-Averaged Accuracy ( $A_c$ )

Classification confusion matrices.

	Truth			Truth			Truth	
	Q	W		Q	W		Q	W
Q	80	10	Q	800	10	Q	800	100
W	20	90	W	200	90	W	200	0
.....								
$A$	$= \frac{80 + 90}{100 + 100}$		$A$	$= \frac{800 + 90}{1000 + 100}$		$A$	$= \frac{800 + 0}{1000 + 100}$	
	$= \frac{170}{200} = 0.85$			$= \frac{890}{1100} = 0.81$			$= \frac{800}{1100} = \mathbf{0.72}$	
$A_c$	$= \frac{1}{2} \left( \frac{80}{100} + \frac{90}{100} \right)$		$A_c$	$= \frac{1}{2} \left( \frac{800}{1000} + \frac{90}{100} \right)$		$A_c$	$= \frac{1}{2} \left( \frac{800}{1000} + \frac{0}{100} \right)$	
	$= 0.85$			$= 0.85$			$= \mathbf{0.40}$	
	Balanced Data			Skew Data			Minority Class Overwhelmed	

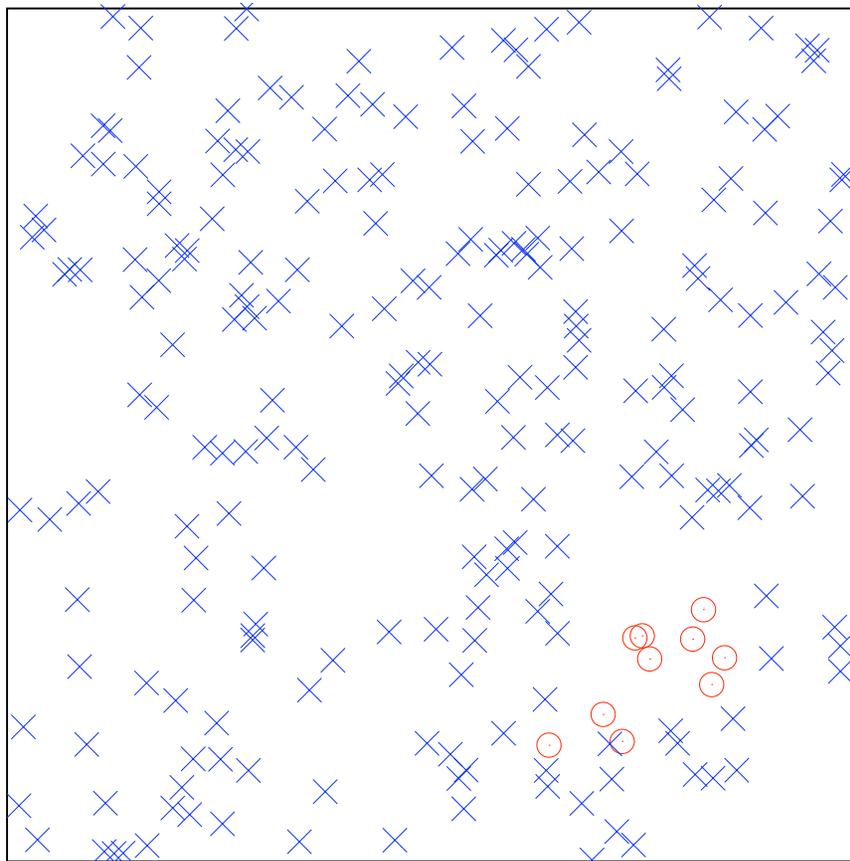
## The Conclusion

When building **ensembles** of **decision trees** to classify extremely **skew** data, such as **Border Gateway Protocol (BGP)** data, use the **Hellinger Distance** metric, rather than the traditional **Infogain** metric, for increased accuracy.

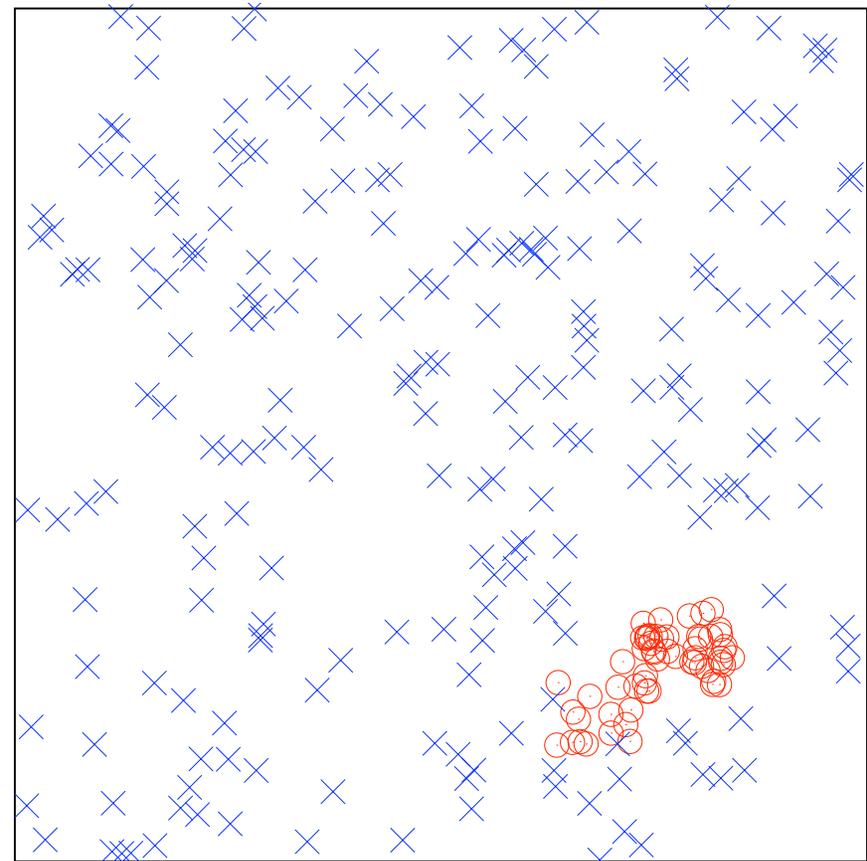
(And be sure to use a **sensible metric** for “accuracy”.)

# Previous Best Solution: SMOTE for Skew Data

Not enough minority data? Invent some!



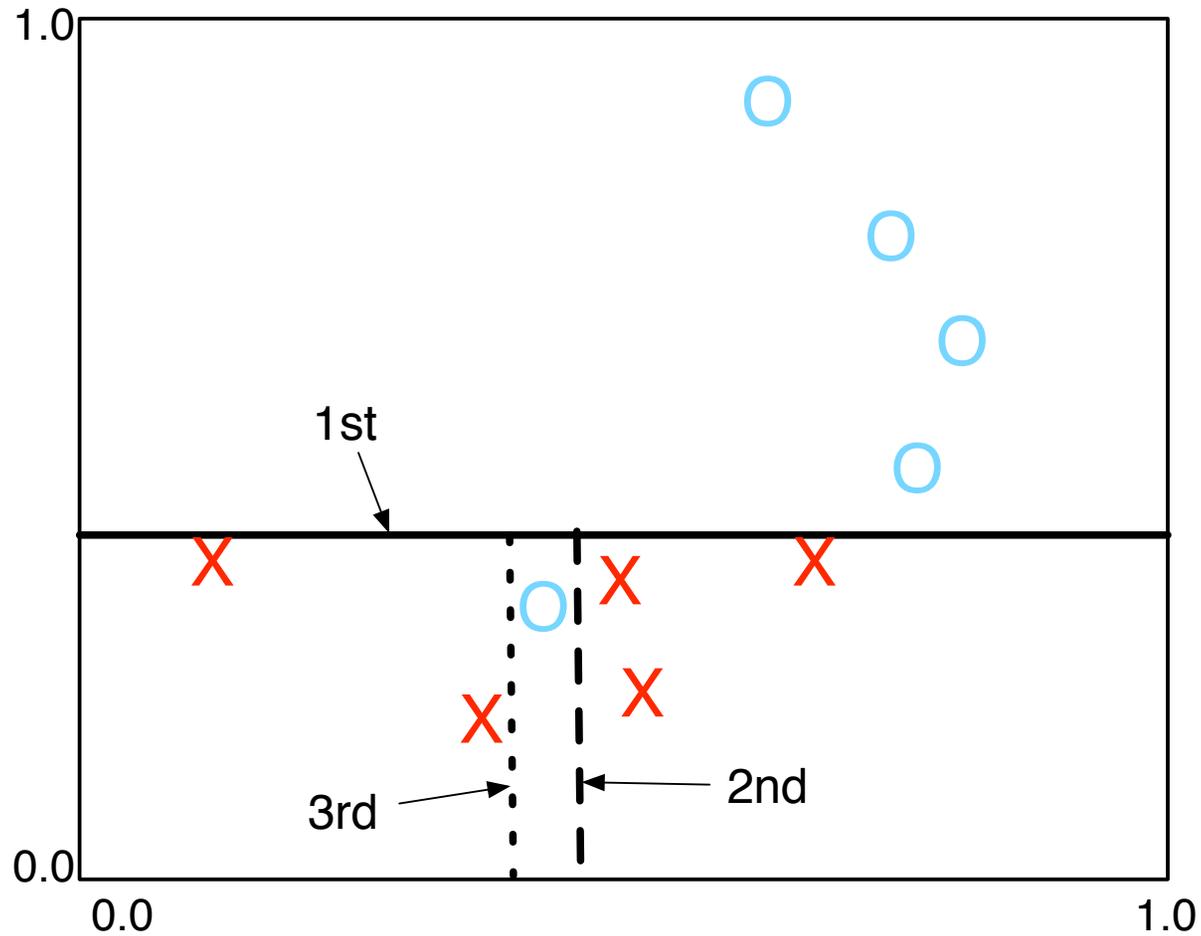
Minority class overwhelmed



Minority class filled out by SMOTE[3]

# Decision Trees and Purity Metrics

Partition attribute space to maximize purity of children. Recurse.



## Infogain is the Traditional Purity Metric

$W, Q$  are the classes of interest

$N$  = the total number of samples

$N_i$  = number of samples in class  $i$

$N^s$  = total number of samples in the  $L/R$  split

$N_i^s$  = number of samples in class  $i$  in  $L/R$  split

$$E = \sum_{i \in (W, Q)} -\frac{N_i^L}{N^L} \log_2 \frac{N_i^L}{N^L} + \sum_{i \in (W, Q)} -\frac{N_i^R}{N^R} \log_2 \frac{N_i^R}{N^R}$$

## Hellinger Pays Attention to Class Statistics

$W, Q$  are the classes of interest

$N$  = the total number of samples

$N_i$  = number of samples in class  $i$

$N^s$  = total number of samples in the  $L/R$  split

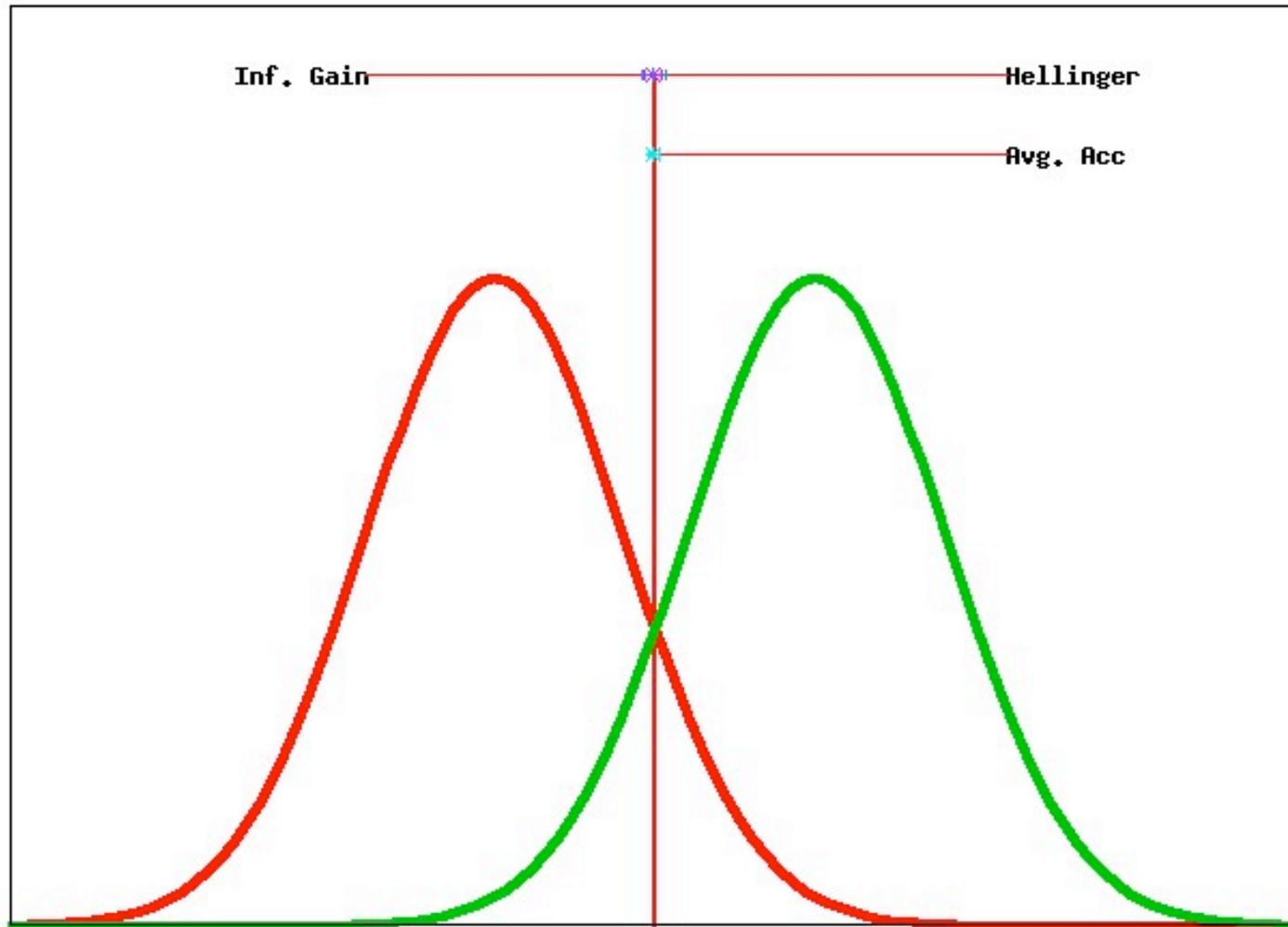
$N_i^s$  = number of samples in class  $i$  in  $L/R$  split

$$E = \sum_{i \in (W, Q)} -\frac{N_i^L}{N^L} \log_2 \frac{N_i^L}{N^L} + \sum_{i \in (W, Q)} -\frac{N_i^R}{N^R} \log_2 \frac{N_i^R}{N^R}$$

$$H = \sqrt{\left\{ \sqrt{\frac{N_Q^L}{N_Q}} - \sqrt{\frac{N_W^L}{N_W}} \right\}^2 + \left\{ \sqrt{\frac{N_Q^R}{N_Q}} - \sqrt{\frac{N_W^R}{N_W}} \right\}^2}$$

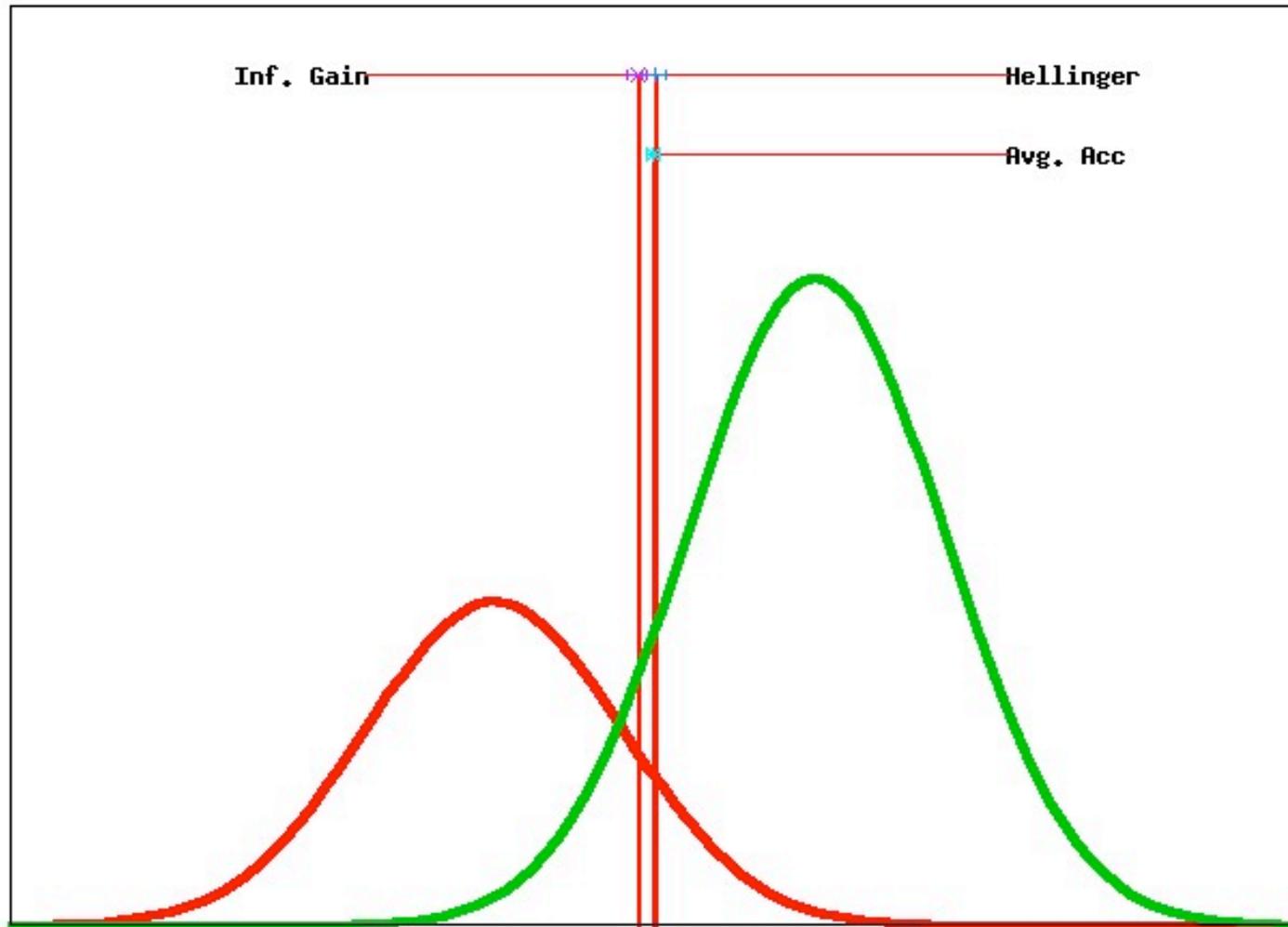
# Infogain vs Hellinger on Gaussian Data

(10000:10000)



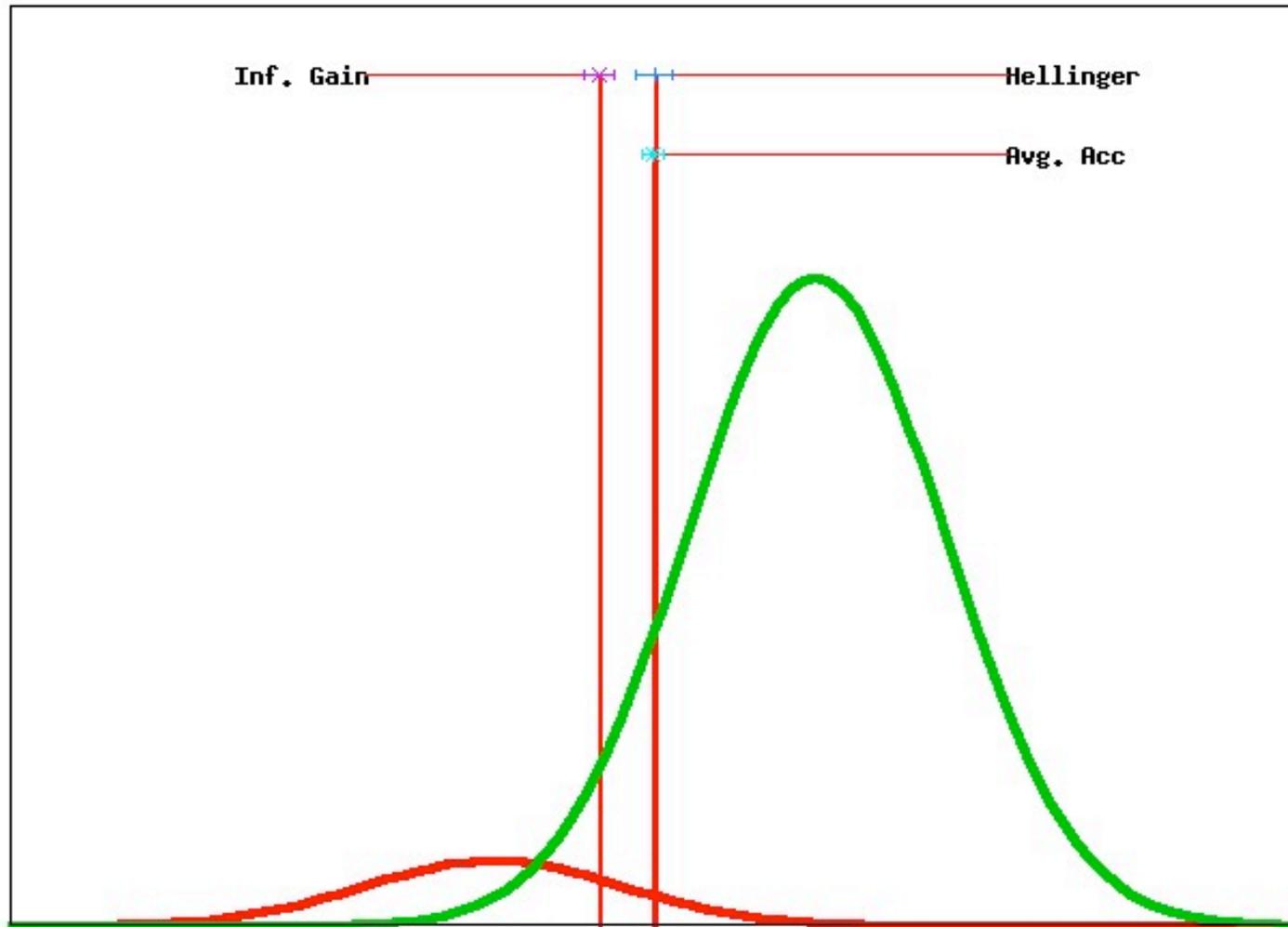
# Infogain vs Hellinger on Gaussian Data

(5000:10000)

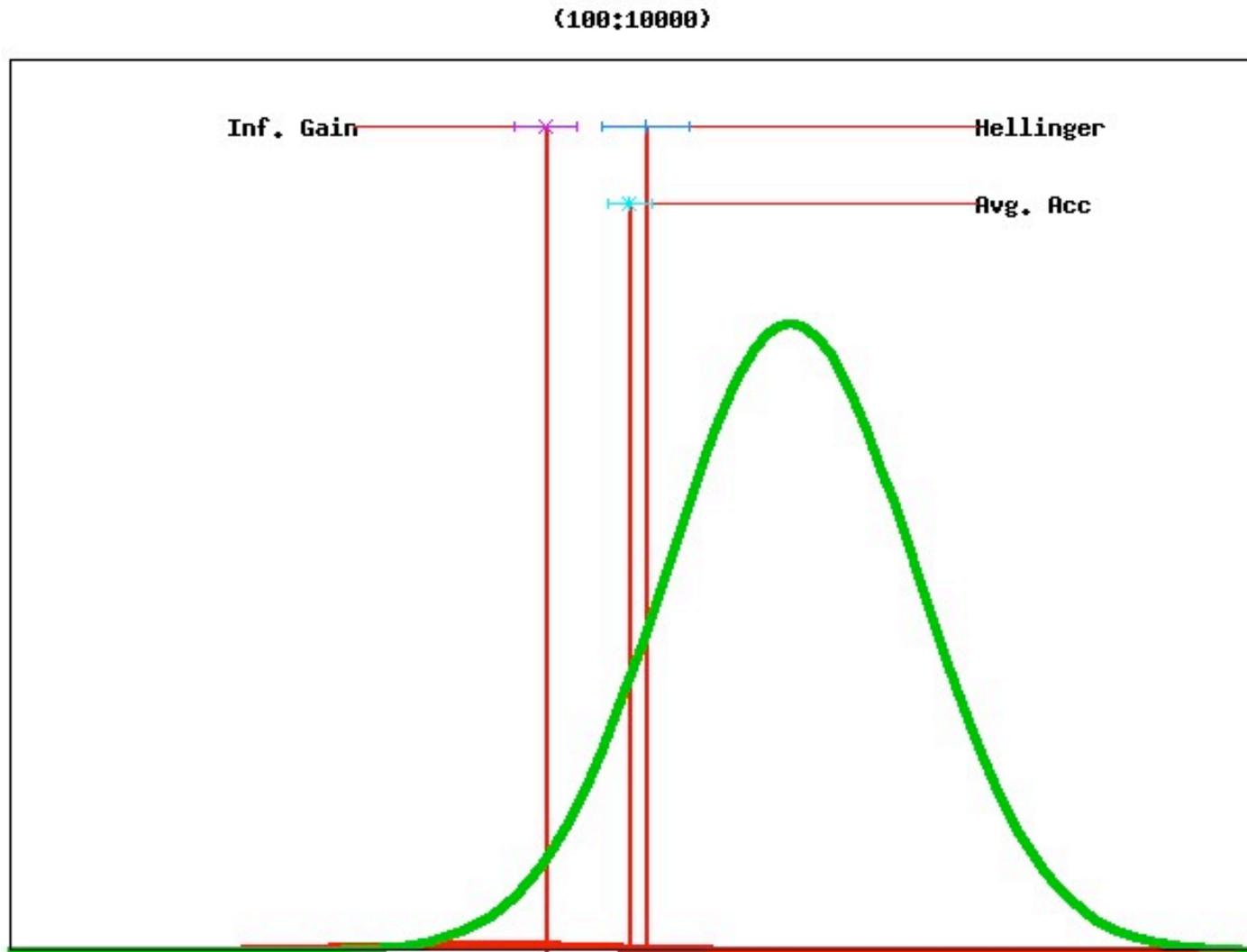


# Infogain vs Hellinger on Gaussian Data

(1000:10000)



# Infogain vs Hellinger on Gaussian Data



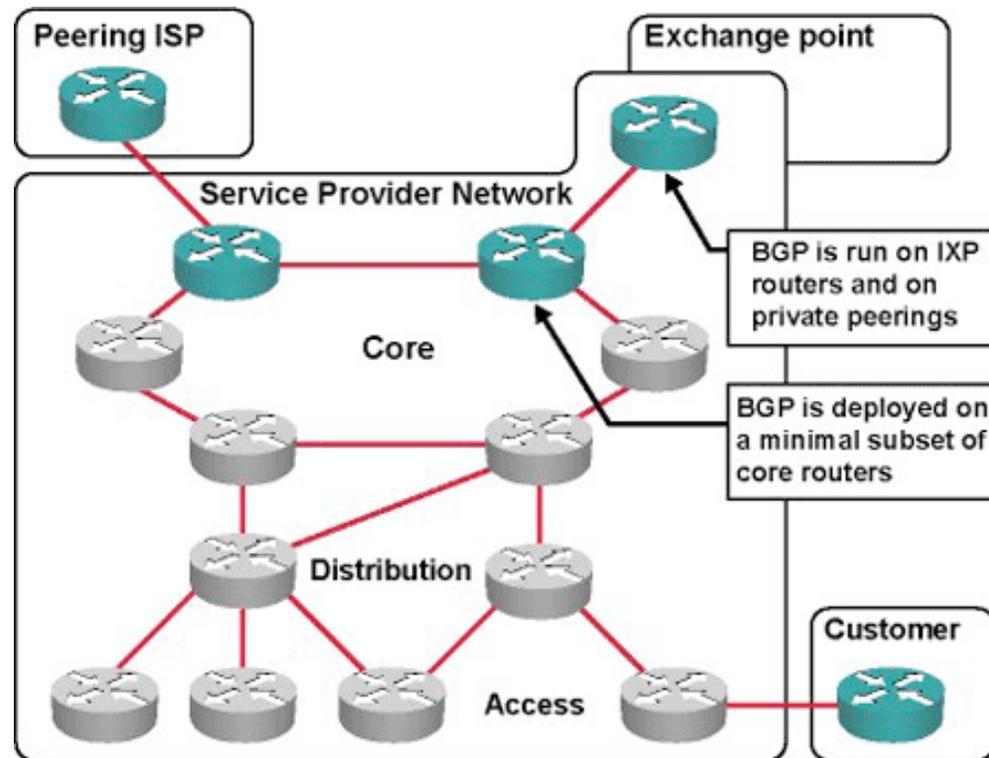
## The Conclusion

When building **ensembles** of **decision trees** to classify extremely **skew** data, such as **Border Gateway Protocol (BGP)** data, use the **Hellinger Distance** metric, rather than the traditional **Infogain** metric, for increased accuracy.

(And be sure to use a **sensible metric** for “accuracy”.)

# Border Gateway Protocol (BGP) Data

Count “withdrawals” and “announcements” over 30 second intervals.



(Our thanks to Max Planck, New Mexico Institute of Mining and Technology, for data acquisition and feature extraction.)

## Anecdote: Event 34, Internet Worm

- Start: 5:28 AM, 01/25/2003.
- End: Noon, 01/26/2003.
- $\Rightarrow$  2224 worm event samples.
- What happens as skew increases?



Proportion of worm samples	0.114	0.011	0.002	0.001
Infogain Bagging, $A_c$	0.897	0.833	0.773	0.757
Hellinger Bagging, $A_c$	0.899	0.838	0.783	0.769
Advantage of Hellinger	0.002	0.005	0.010	0.014

## New, Improved Event 34, With Extra SMOTE

- Start: 5:28 AM, 01/25/2003.
- End: Noon, 01/26/2003.
- $\Rightarrow$  2224 worm event samples.
- What happens as skew increases?



Proportion of worm samples	0.114	0.011	0.002	0.001
Infogain Bagging, $A_c$	0.897	0.833	0.773	0.757
SMOTE + Infogain Bagging, $A_c$	0.933	0.929	—	—
Hellinger Bagging, $A_c$	0.899	0.838	0.783	0.769
SMOTE + Hellinger Bagging, $A_c$	0.931	0.934	—	—

## Bagging, Hellinger vs. Infogain, Quantitatively

- *Skew* data: 19 datasets, Wilcoxon rank test  $\Rightarrow$  Hellinger statistically significantly more accurate than Infogain, at 95% confidence.
- *Balanced* data: 8 data sets, Wilcoxon rank test  $\Rightarrow$  no difference between Hellinger and Infogain, at 99% confidence.



Conclusion (so far): Always use Hellinger! It never hurts, and it often helps.

## The Conclusion

When building **ensembles** of **decision trees** to classify extremely **skew** data, such as **Border Gateway Protocol (BGP)** data, use the **Hellinger Distance** metric, rather than the traditional **Infogain** metric, for increased accuracy.

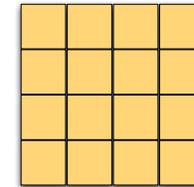
(And be sure to use a **sensible metric** for “accuracy”.)

## References

- [1] BANFIELD, R. E., HALL, L. O., BOWYER, K. W., BHADORIA, D., KEGELMEYER, W. P., AND ESCHRICH, S. A comparison of ensemble creation techniques. In *Proceedings of the Fifth International Conference on Multiple Classifier Systems, MCS2004* (2004), J. K. F. Roli and T. Windeatt, Eds., vol. 3077 of *Lecture Notes in Computer Science*, Springer-Verlag.
- [2] BANFIELD, R. E., HALL, L. O., BOWYER, K. W., AND KEGELMEYER, W. P. Ensemble diversity measures and their application to thinning. *Information Fusion Journal* 6, 1 (March 2005), 49–62.
- [3] CHAWLA, N. V., BOWYER, K. W., HALL, L. O., AND KEGELMEYER, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.
- [4] CHAWLA, N. V., HALL, L. O., BOWYER, K. W., AND KEGELMEYER, W. P. Learning ensembles from bites: A scalable and accurate approach. *Journal of Machine Learning Research* 5 (2004), 421–451.
- [5] CONDORCET, N. Essai sur l’application de l’analyse à la probabilité des décisions rendues à la pluralite des voix. Correspondence, 1785. Paris.
- [6] DOU, D., LI, J., QIN, H., KIM, S., AND ZHONG, S. Understanding and utilizing the hierarchy of abnormal BGP events. In *Proceedings of the 2007 SIAM International Conference on Data Mining* (2007), pp. 467–472.
- [7] RIPE NCC. RIPE routing information service raw data. [www.ripe.net](http://www.ripe.net).

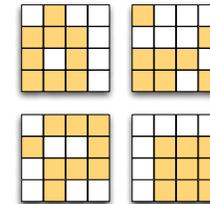
# Ensembles: Efficient, Robust, Optimal Accuracy

**Traditional:** Use 100% of training data to build a sage.



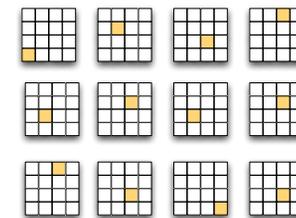
Sage sees all the data.

**Ensembles:** Use randomized 100% of training data to build an expert. Repeat to build many experts. Vote them.



Each expert sees 2/3rds of the data.

**Sandia:** Use a semi-random 1% of the training data to build a “bozo”. Repeat to build very many bozos. Vote them.



Each bozo sees a tiny fraction.

The experts beat the sage[1].

The bozos beat the experts[4].