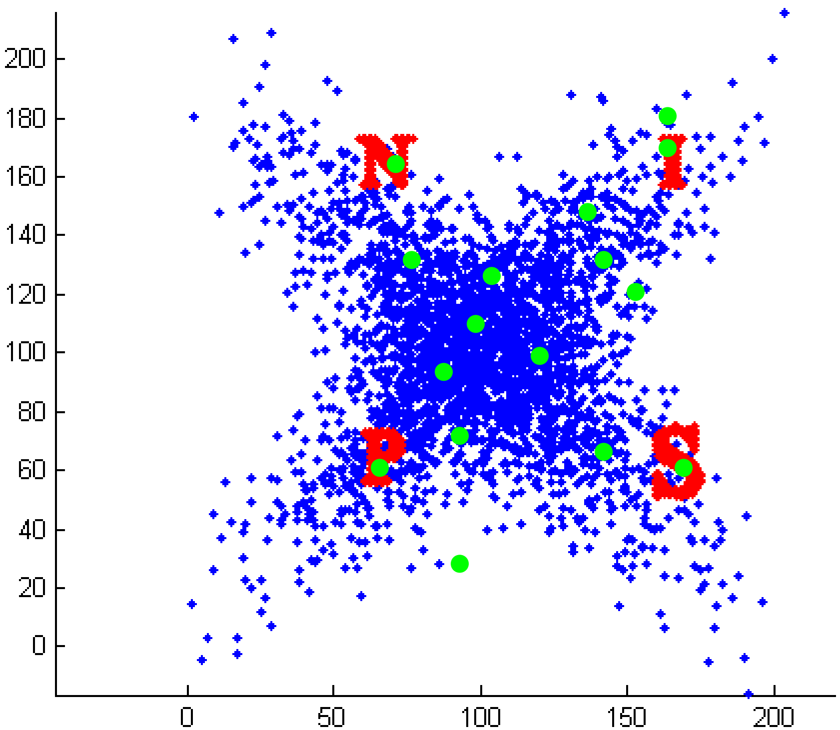
Rare Category Detection

Jingrui He
Machine Learning Department
Carnegie Mellon University

Joint work with Jaime Carbonell

What's Rare Category Detection

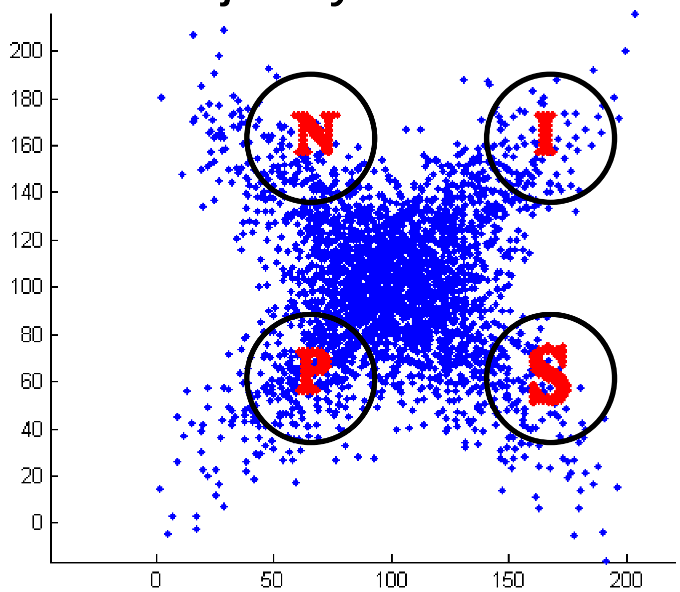


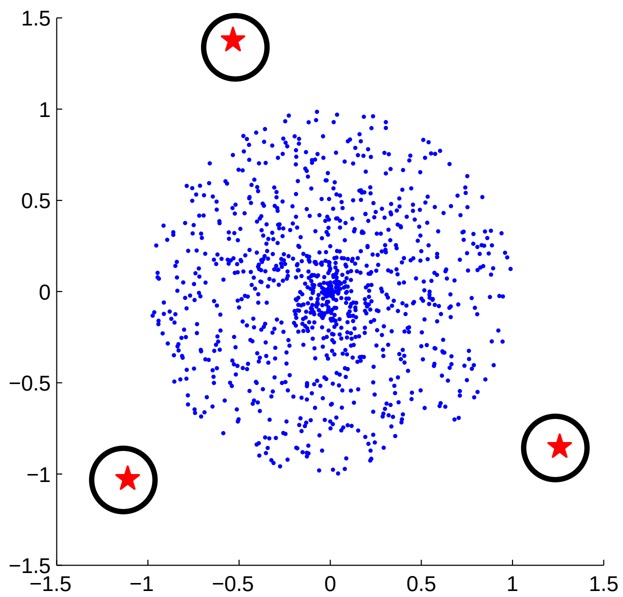
- Start de-novo
- Very skewed classes
  - Majority classes
  - *Minority classes*
- Labeling oracle
- Goal
  - Discover minority classes with *a few* label requests

Comparison with Outlier Detection

- Rare classes
  - A group of points
  - Clustered
  - Non-separable from the majority classes

- Outliers
  - A single point
  - Scattered
  - Separable





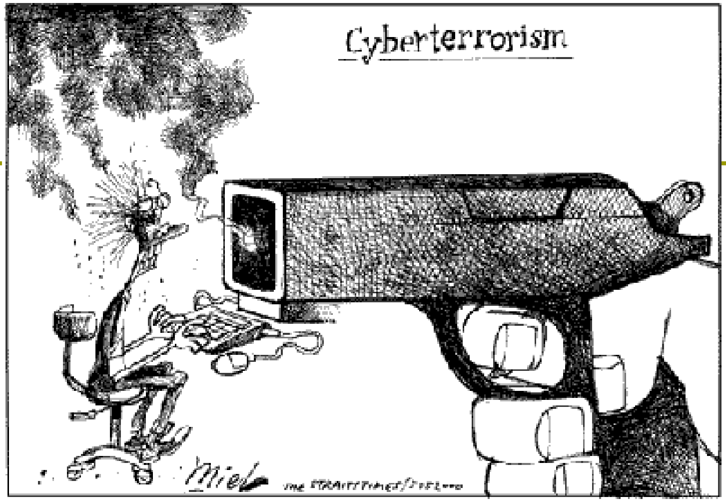
Comparison with Active Learning

- Rare category detection
  - Initial condition: *NO* labeled examples
  - Goal: *discover* the minority classes with the least label requests
- Active learning
  - Initial condition: labeled examples from *each* class
  - Goal: *improve* the performance of the current classifier with the least label requests

App

Fraud detection

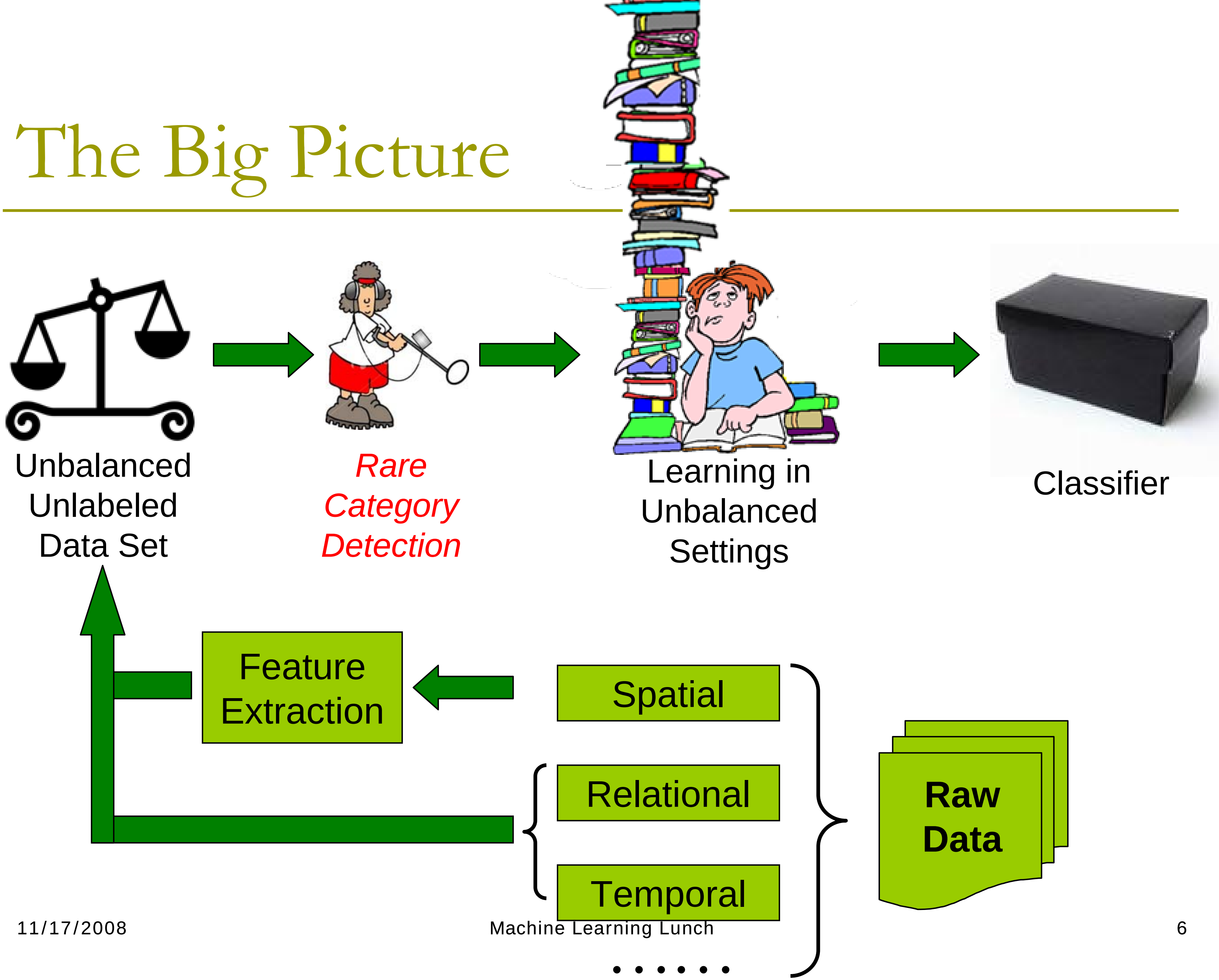
Network intrusion detection



Astronomy

Spam image detection

The Big Picture



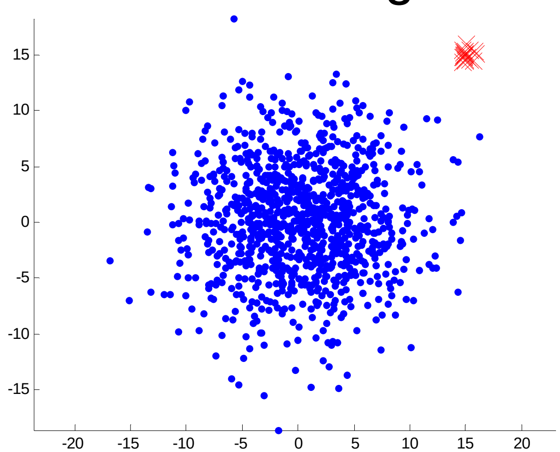
Outline

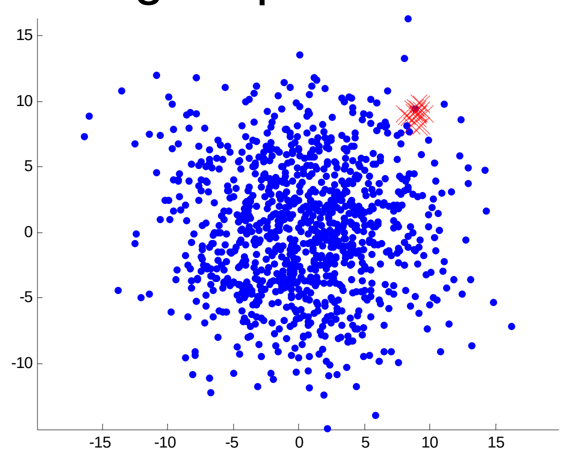
- Problem definition
- Related work
- Rare category detection for spatial data
  - Prior-dependent rare category detection
  - Prior-free rare category detection
- Conclusion

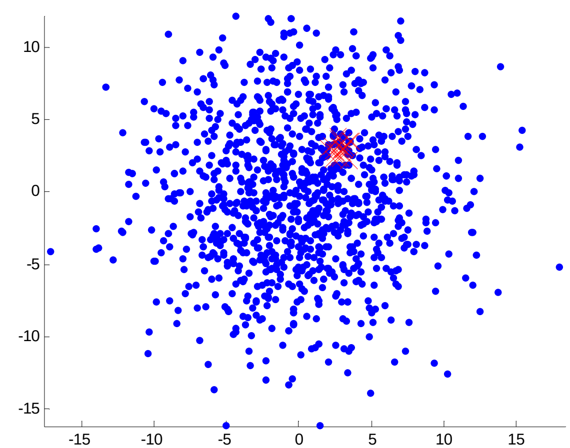
Related Work

- Pelleg & Moore 2004
  - Mixture model
  - Different selection criteria
- Fine & Mansour 2006
  - Generic consistency algorithm
  - Upper bounds and lower bounds
- Papadimitriou et al 2003
  - LOCI algorithm for groups of outliers

Separable or Near-separable







Outline

- Problem definition
- Related work
- Rare category detection for spatial data
  - Prior-dependent rare category detection
  - Prior-free rare category detection
- Conclusion

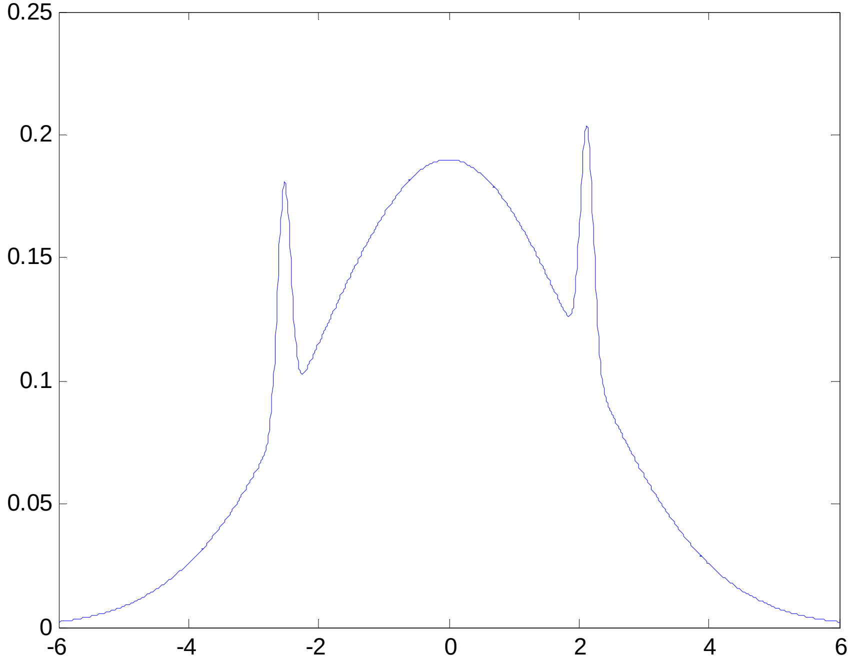
Notations

- Unlabeled examples: $S = \{x_1, \ldots, x_n\}$, $x_i \in \Re^d$
- *m* Classes: $y_i \in \{1, \ldots, m\}$
- *m-1* rare classes: $p^2, \ldots, p^m$
- One majority class: $p^1 >> p^c, 2 \le c \le m$

- Goal: find at least *ONE* example from *each* rare class by requesting *a few* labels

Assumptions

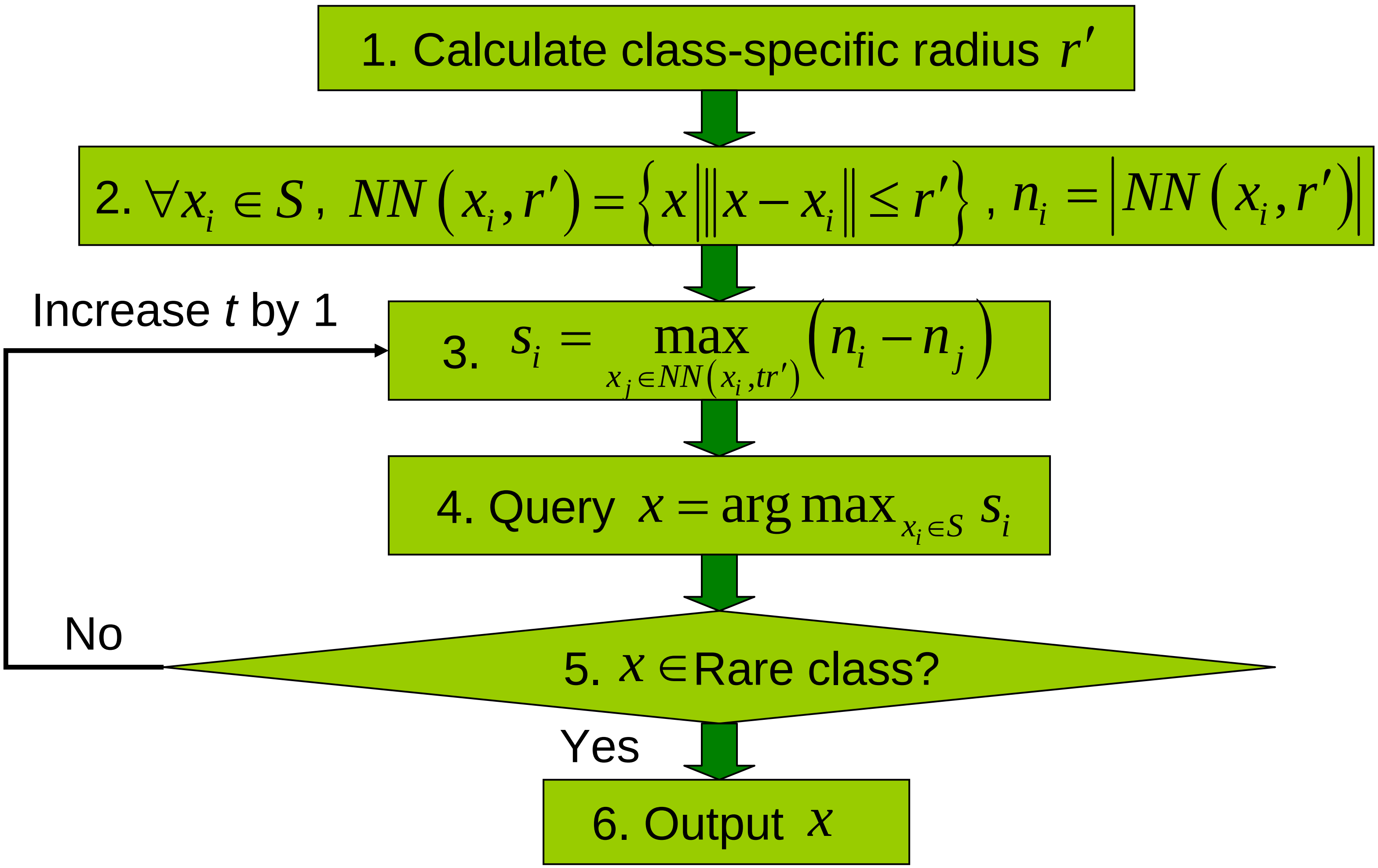
- The distribution of the majority class is sufficiently smooth
- Examples from the minority classes form compact clusters in the feature space



Overview of the Algorithms

- Nearest-neighbor-based methods
  - Methodology: local density differential sampling
  - Intuition: select examples according to the *change in local density*

Two Classes: NNDB

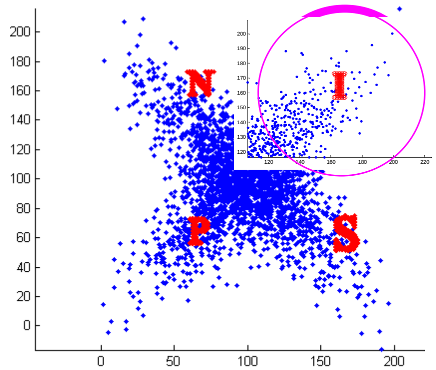


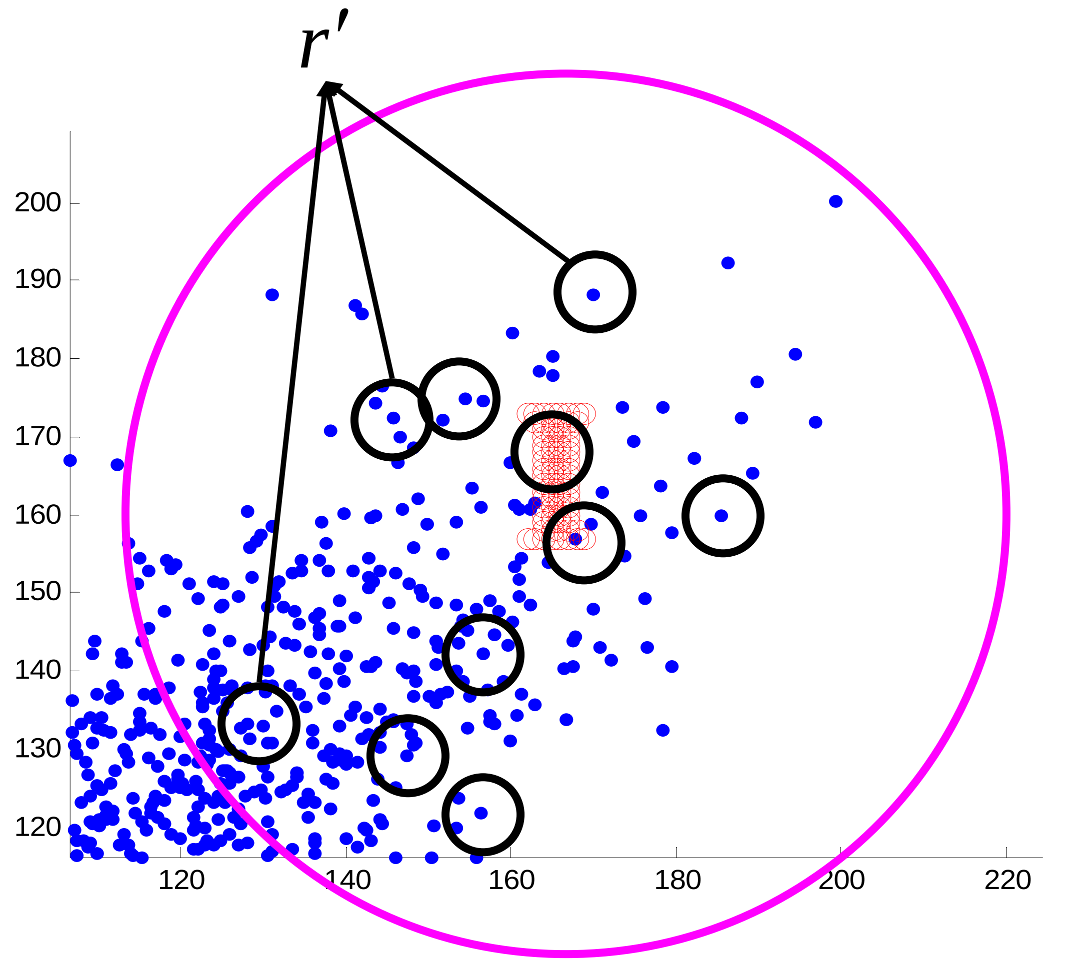
NNDB: Calculate Class-Specific Radius

- Number of examples from the minority class: $p^2 \rightarrow K = np^2$
- $\forall x_i \in S$, calculate the distance r_i^K between x_i and its K^{th} nearest neighbor
- The class-specific radius:

$$r' = \min_{i=1}^{n} r_i^K$$

NNDB: Calculate Nearest Neighbors





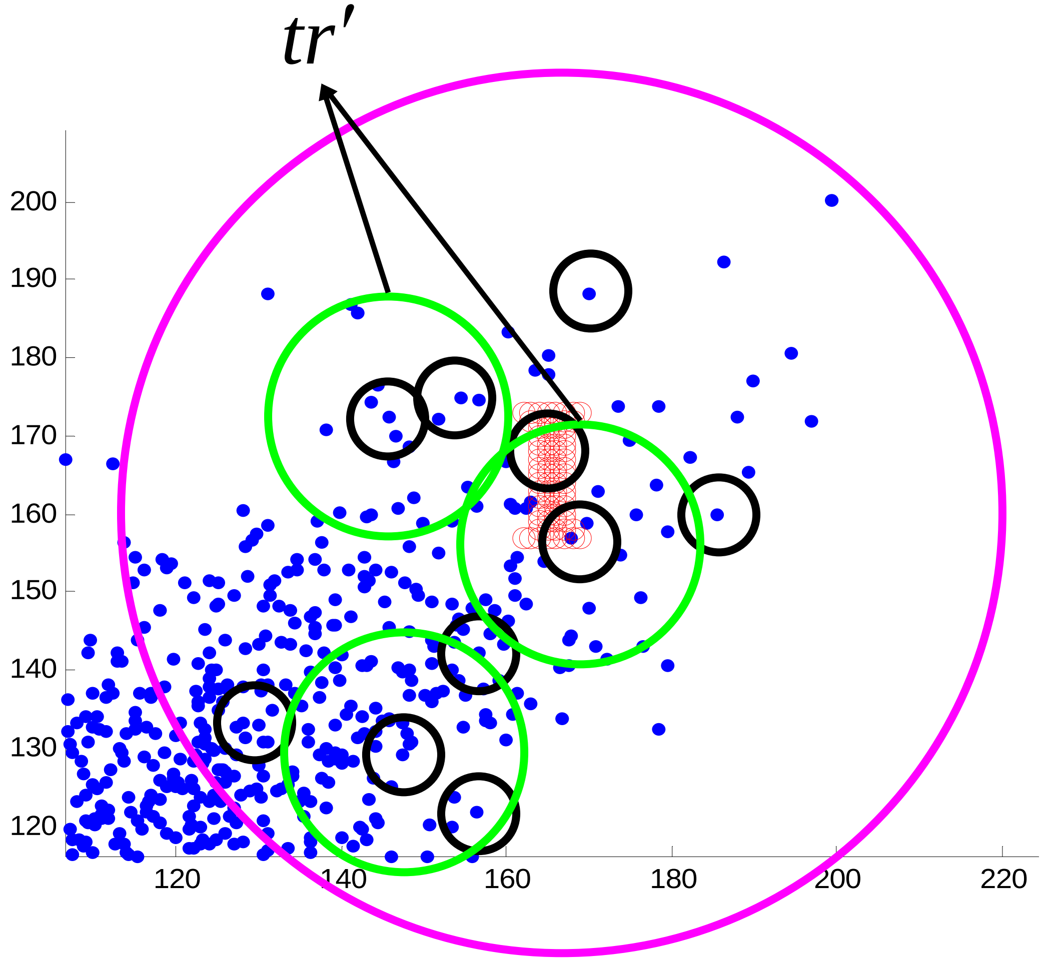
$$NN\left(x_i, r'\right) = \left\{x \,\middle|\, \left\|x - x_i\right\| \le r'\right\}$$

$$n_i = \left|NN\left(x_i, r'\right)\right|$$

NNDB: Calculate the Scores

$$s_i = \max_{x_j \in NN(x_i, tr')} \left(n_i - n_j \right)$$

Query $x = \arg\max_{x_i \in S} s_i$

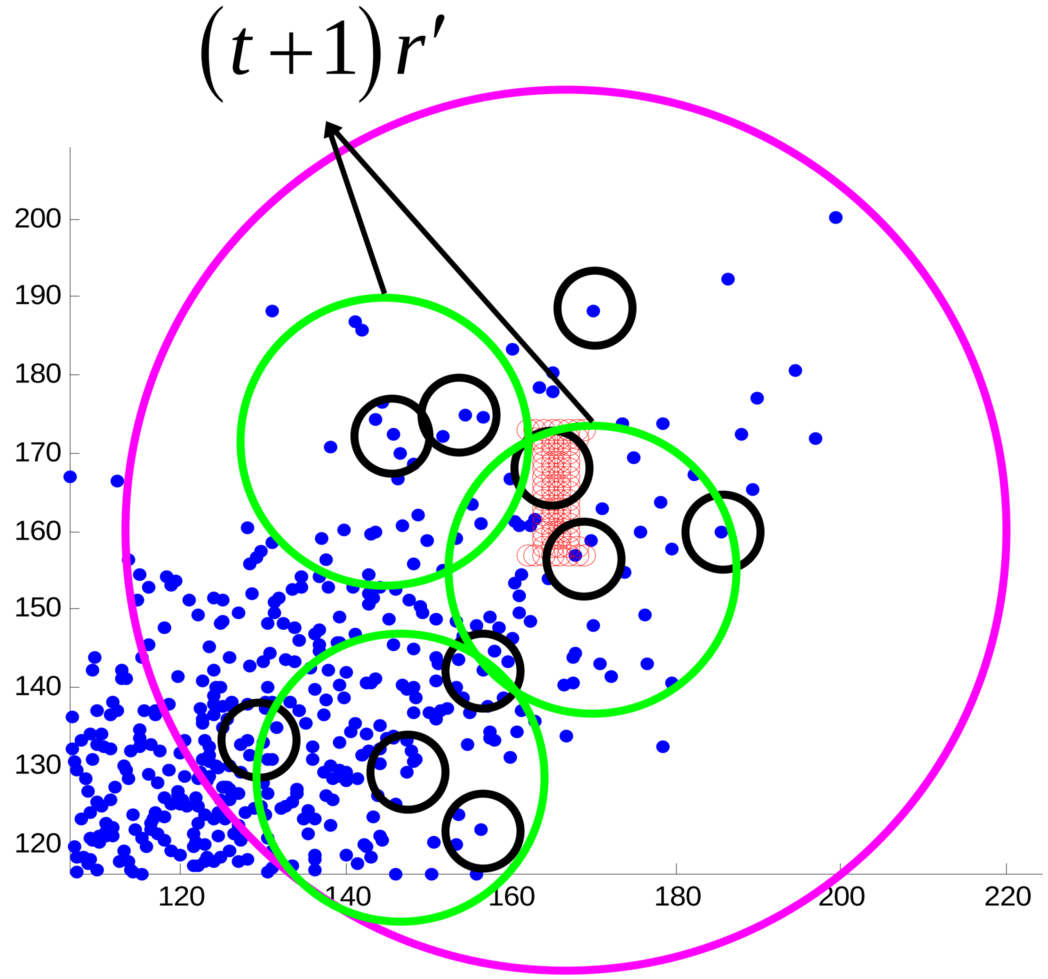


NNDB: Pick the Next Candidate

Increase t by 1

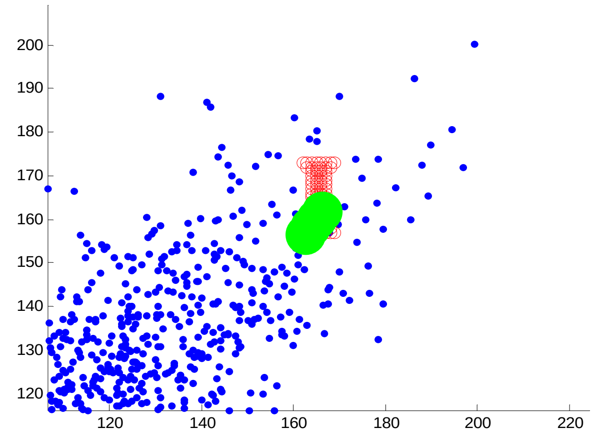
$$s_i = \max_{x_j \in NN(x_i,(t+1)r')} (n_i - n_j)$$

Query $x = \arg\max_{x_i \in S} s_i$



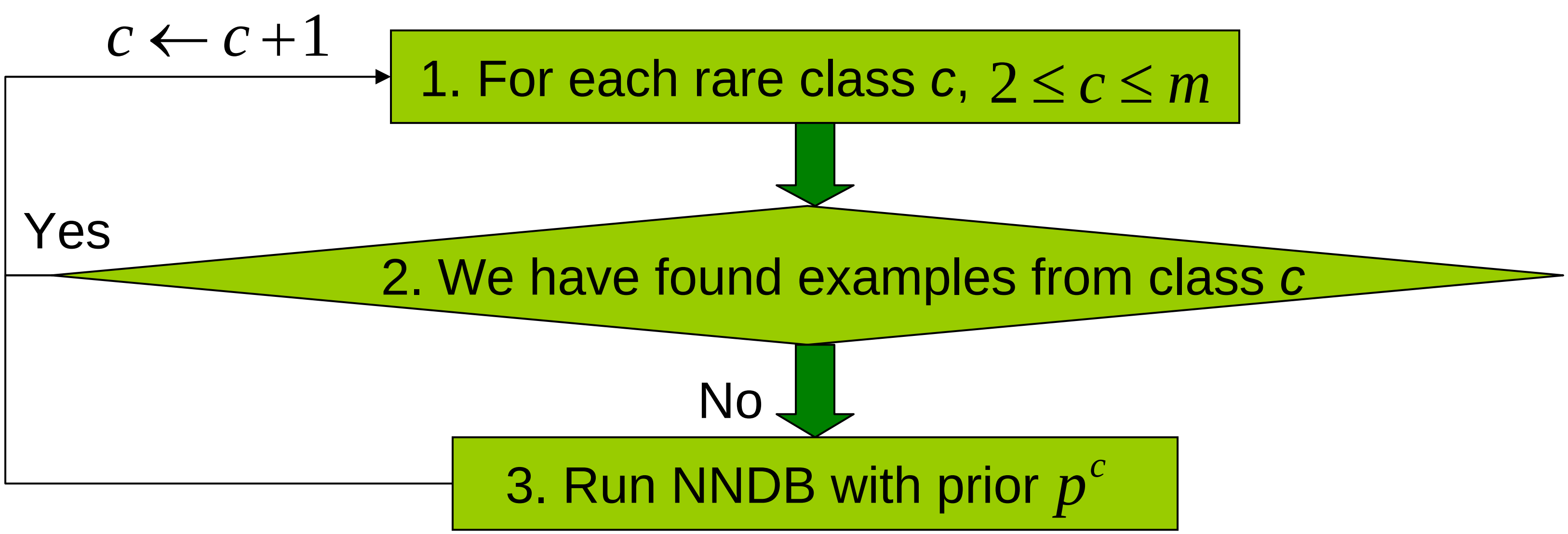
Why NNDB Works

- Theoretically
  - ***Theorem 1*** [He & Carbonell 2007]: under certain conditions, with high probability, after ***a few*** iteration steps, NNDB queries ***at least one*** example whose probability of coming from the minority class is ***at least 1/3***
- Intuitively
  - The score s_i measures the change in local density



Multiple Classes: ALICE

- *m-1* rare classes: $p^2, \ldots, p^m$
- One majority class: $p^1 \gg p^c, 2 \leq c \leq m$



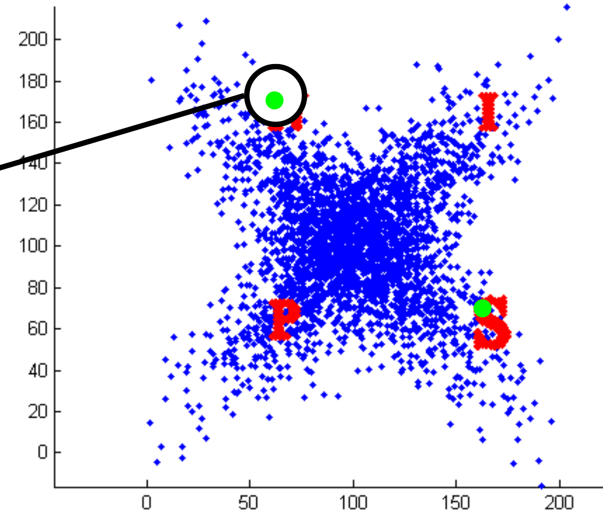
Why ALICE Works

- Theoretically
  - ***Theorem 2*** [He & Carbonell 2008]: under certain conditions, with high probability, in ***each outer loop*** of ALICE, after ***a few*** iteration steps in NNDB, ALICE queries ***at least one*** example whose probability of coming from ***one*** minority class is ***at least 1/3***

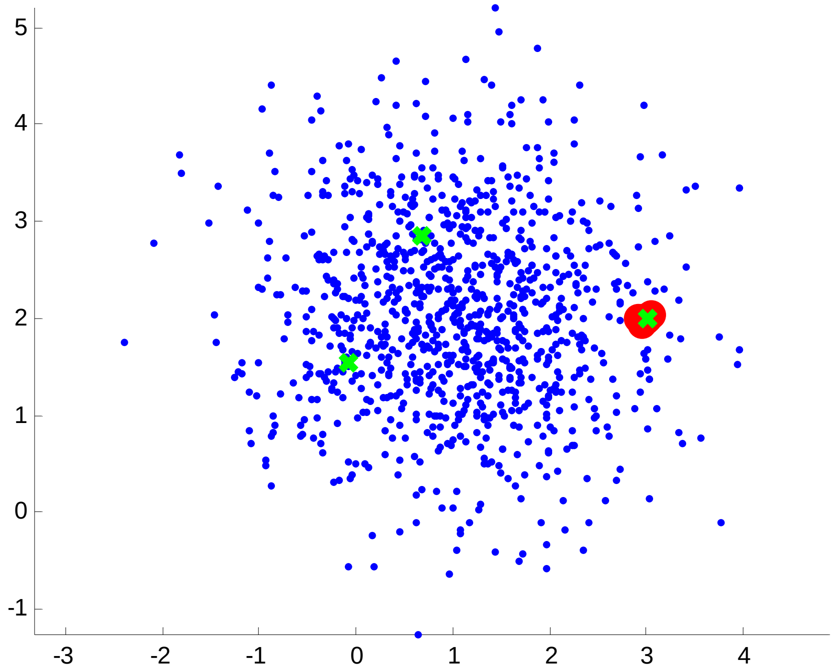
Implementation Issues

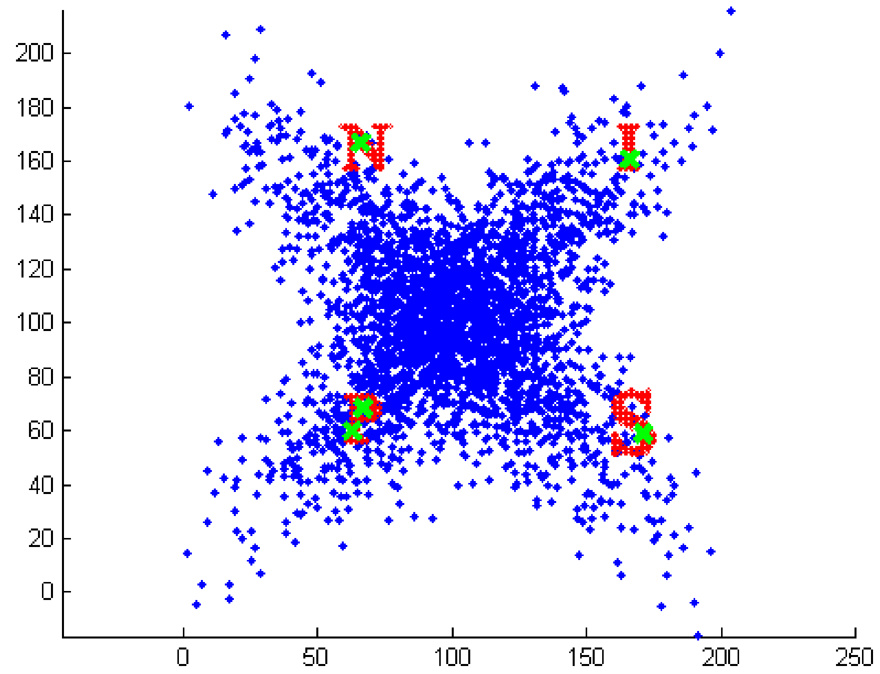
- ALICE
  - Problem: repeatedly sampling from the same rare class
- MALICE
  - Solution: relevance feedback

Class-specific radius



Results on Synthetic Data Sets

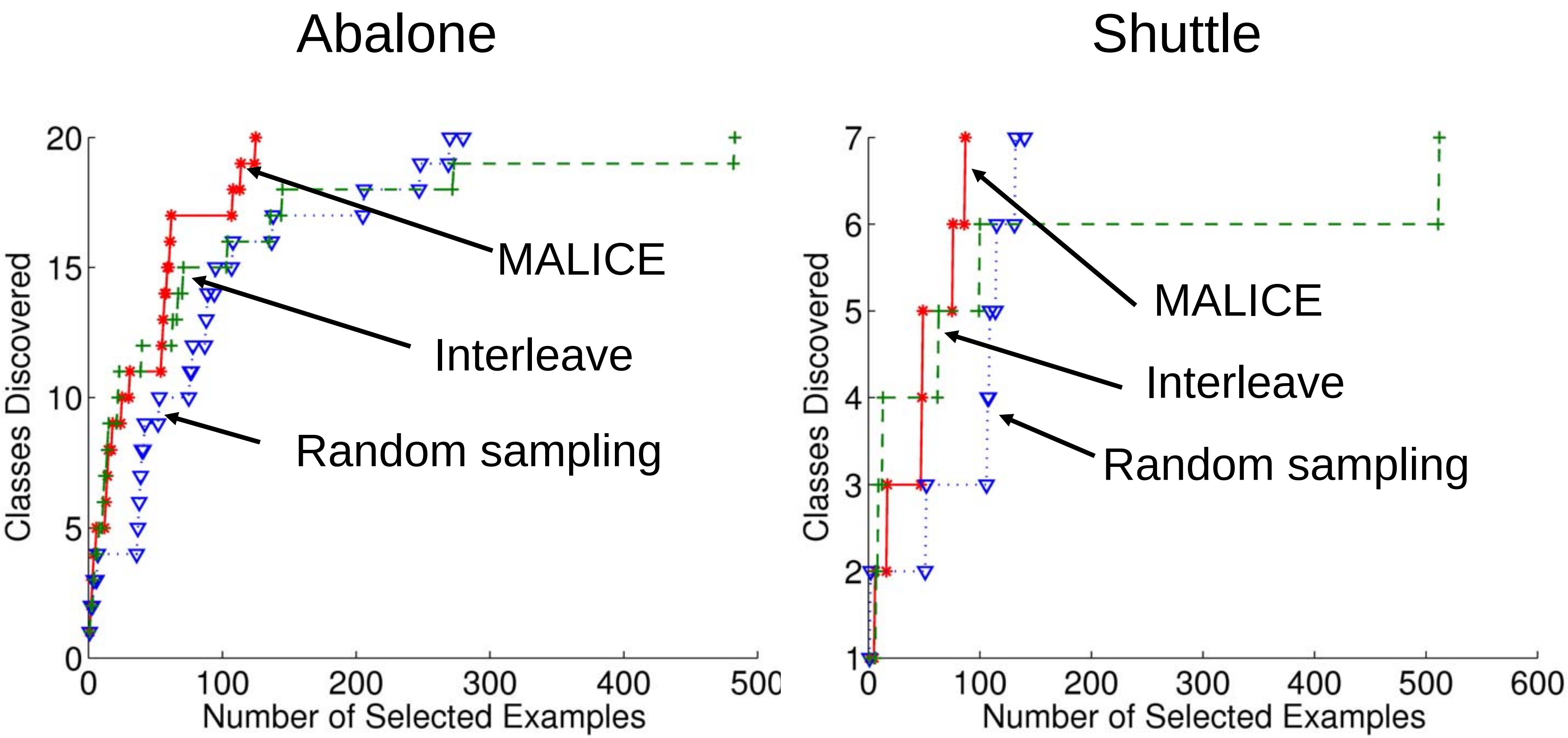




Summary of Real Data Sets

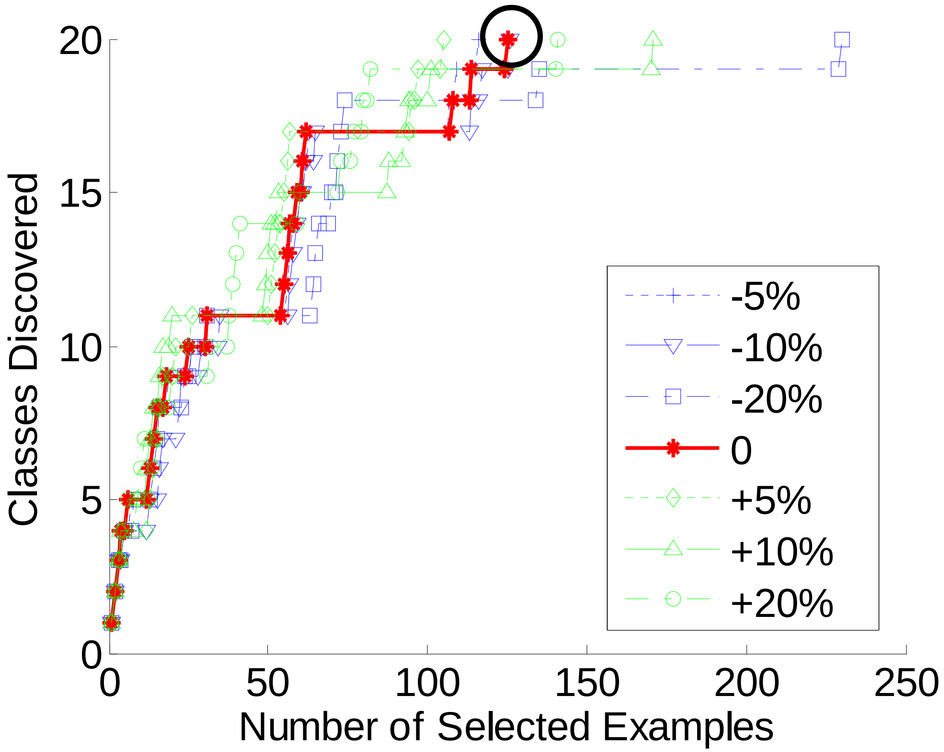
- Abalone
  - 4177 examples
  - 7-dimensional features
  - 20 classes
  - Largest class: 16.50%
  - Smallest class: 0.34%
- Shuttle
  - 4515 examples
  - 9-dimensional features
  - 7 classes
  - Largest class: 75.53%
  - Smallest class: 0.13%

Results on Real Data Sets

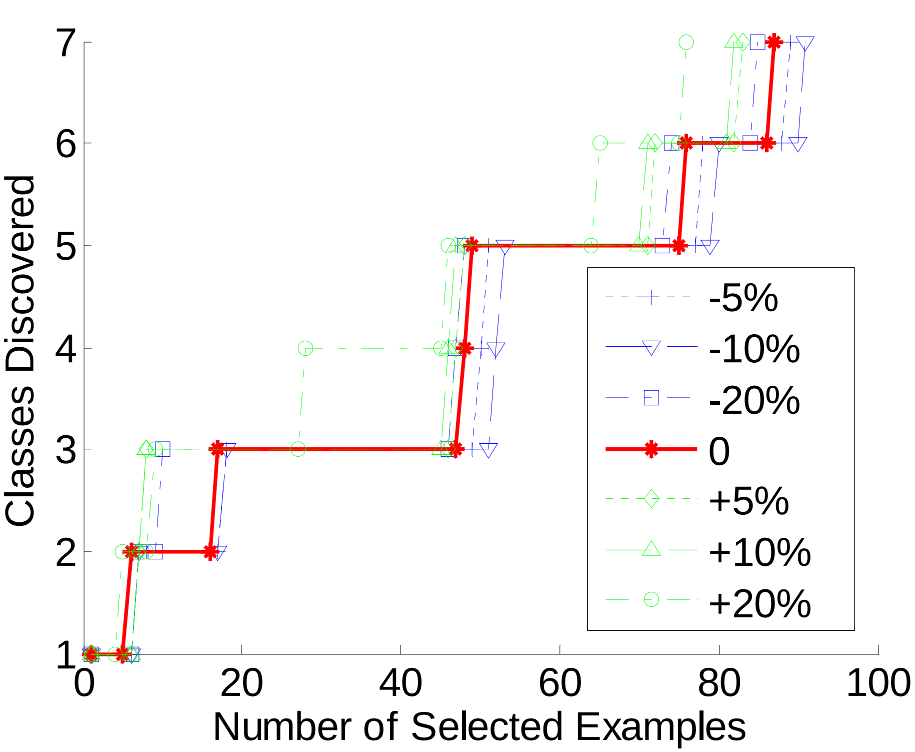


Imprecise priors

Abalone



Shuttle



Outline

- Problem definition
- Related work
- Rare category detection for spatial data
  - Prior-dependent rare category detection
  - Prior-free rare category detection
- Conclusion

Overview of the Algorithm

- Density-based method
  - Methodology: specially designed exponential families
  - Intuition: select examples according to the change in local density
  - Difference from NNDB (ALICE): *NO* prior information needed

Specially Designed Exponential Families [Efron & Tibshirani 1996]

- Favorable compromise between parametric and nonparametric density estimation
- Estimated density

Carrier density

$p \times 1$ parameter vector

$$g_\beta(x) = \boxed{g_0(x)} \exp\left(\boxed{\beta_0} + \boxed{\beta_1^T}\,\boxed{t(x)}\right)$$

Normalizing parameter

$p \times 1$ vector of sufficient statistics

SEDER Algorithm

- Carrier density: kernel density estimator
- $t(x) = \left[\left(x^1\right)^2, \ldots, \left(x^d\right)^2\right]^T$
- To decouple the estimation of different parameters
  - Decompose $\beta_0 = \sum_{j=1}^{d} \beta_0^j$
  - Relax the constraint such that

$$\int_{x^j} \frac{1}{\sqrt{2\pi}\sigma^j} \exp\left(-\frac{\left(x^j - x_i^j\right)^2}{2\left(\sigma^j\right)^2}\right) \exp\left(\beta_{0i}^j + \beta_1^j \left(x^j\right)^2\right) dx^j = 1$$

Parameter Estimation

- ***Theorem 3*** [To appear]: the maximum likelihood estimate $\hat{\beta}_1^j$ and $\hat{\beta}_{0i}^j$ of β_1^j and β_{0i}^j satisfy the following conditions: $\forall j \in \{1,\ldots,d\}$

$$\sum_{k=1}^{n}\left(x_k^j\right)^2 = \sum_{k=1}^{n}\frac{\sum_{i=1}^{n}\exp\left(\hat{\beta}_{0i}^j - \frac{\left(x_k^j - x_i^j\right)^2}{2\left(\sigma^j\right)^2}\right)E_i^j\left(\left(x^j\right)^2\right)}{\sum_{i=1}^{n}\exp\left(\hat{\beta}_{0i}^j - \frac{\left(x_k^j - x_i^j\right)^2}{2\left(\sigma^j\right)^2}\right)}$$

where

$$E_i^j\left(\left(x^j\right)^2\right) = \int_{x^j}\left(x^j\right)^2\frac{1}{\sqrt{2\pi}\sigma^j}\exp\left(-\frac{\left(x^j - x_i^j\right)^2}{2\left(\sigma^j\right)^2}\right)\exp\left(\hat{\beta}_{0i}^j + \hat{\beta}_1^j\left(x^j\right)^2\right)dx^j$$

Parameter Estimation cont.

- Let $\beta_1^j = \left(1 - \frac{1}{\boxed{b^j}}\right) \frac{1}{2\left(\sigma^j\right)^2}$

b^j: positive parameter

- $\forall j \in \{1, \ldots, d\} : \hat{b}^j \approx \frac{-B + \sqrt{B^2 + 4AC}}{2A}$

$\hat{b}^j \leq 1$

in most cases

where $B = \left(\sigma^j\right)^2, \; C = \frac{1}{n} \sum_{k=1}^{n} \left(x_k^j\right)^2$

$$A = \frac{1}{n} \sum_{k=1}^{n} \frac{\sum_{i=1}^{n} \exp\left(-\frac{\left(x_k^j - x_i^j\right)^2}{2\left(\sigma^j\right)^2}\right)\left(x_i^j\right)^2}{\sum_{i=1}^{n} \exp\left(-\frac{\left(x_k^j - x_i^j\right)^2}{2\left(\sigma^j\right)^2}\right)}$$

Scoring Function

- The estimated density

$$\tilde{g}_b(x) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\frac{1}{\sqrt{2\pi b^j}\,\sigma^j}\exp\left(-\frac{\left(x^j - b^j x_i^j\right)^2}{2\left(\sigma^j\right)^2 b^j}\right)$$

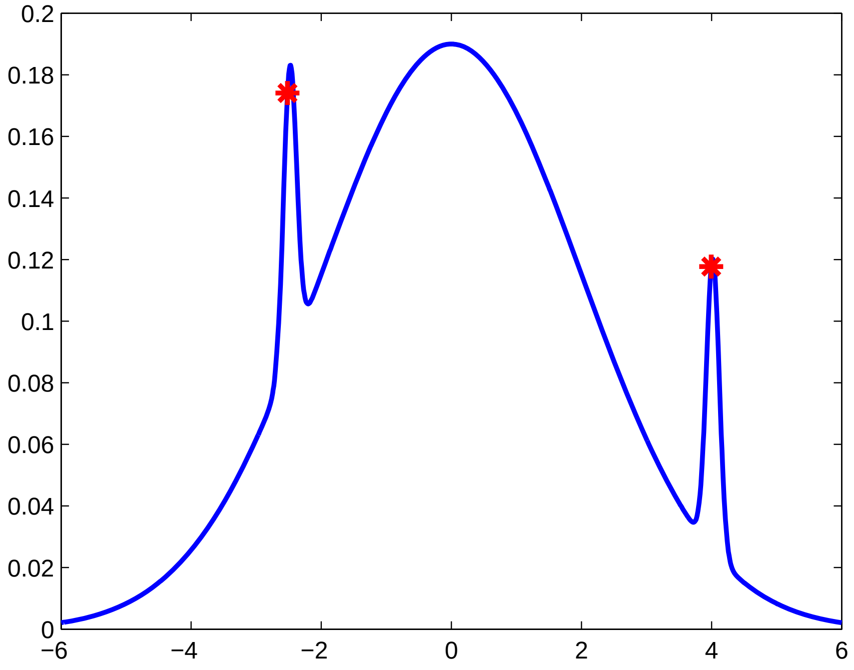
- Scoring function: norm of the gradient

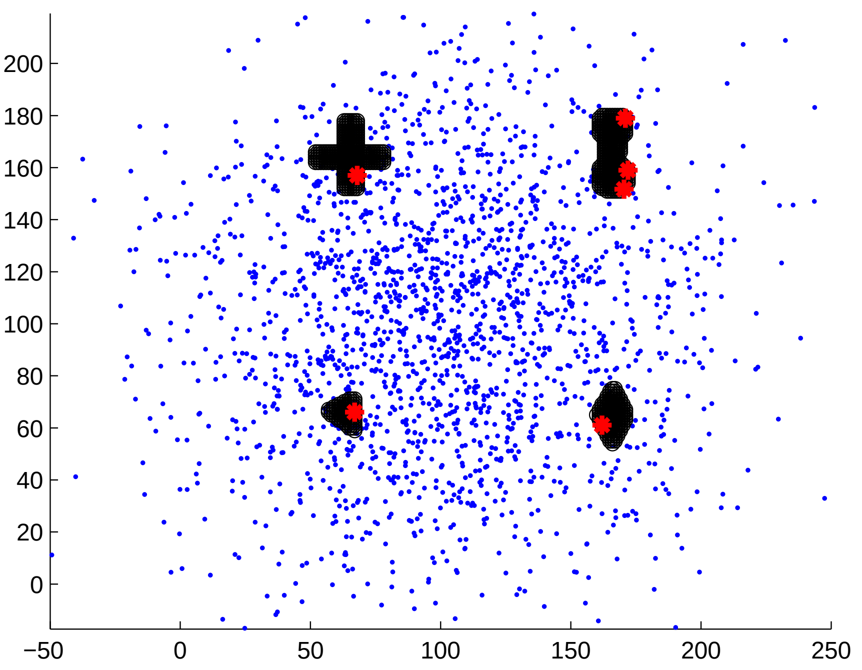
$$s_k = \sqrt{\sum_{l=1}^{d}\frac{\left(\sum_{i=1}^{n} D_i(x_k)\left(x_k^l - b^l x_i^l\right)\right)^2}{\left(\left(\sigma^l\right)^2 b^l\right)^2}}$$

where

$$D_i(x) = \frac{1}{n}\prod_{j=1}^{d}\frac{1}{\sqrt{2\pi b^j}\,\sigma^j}\exp\left(-\frac{\left(x^j - b^j x_i^j\right)^2}{2\left(\sigma^j\right)^2 b^j}\right)$$

Results on Synthetic Data Sets

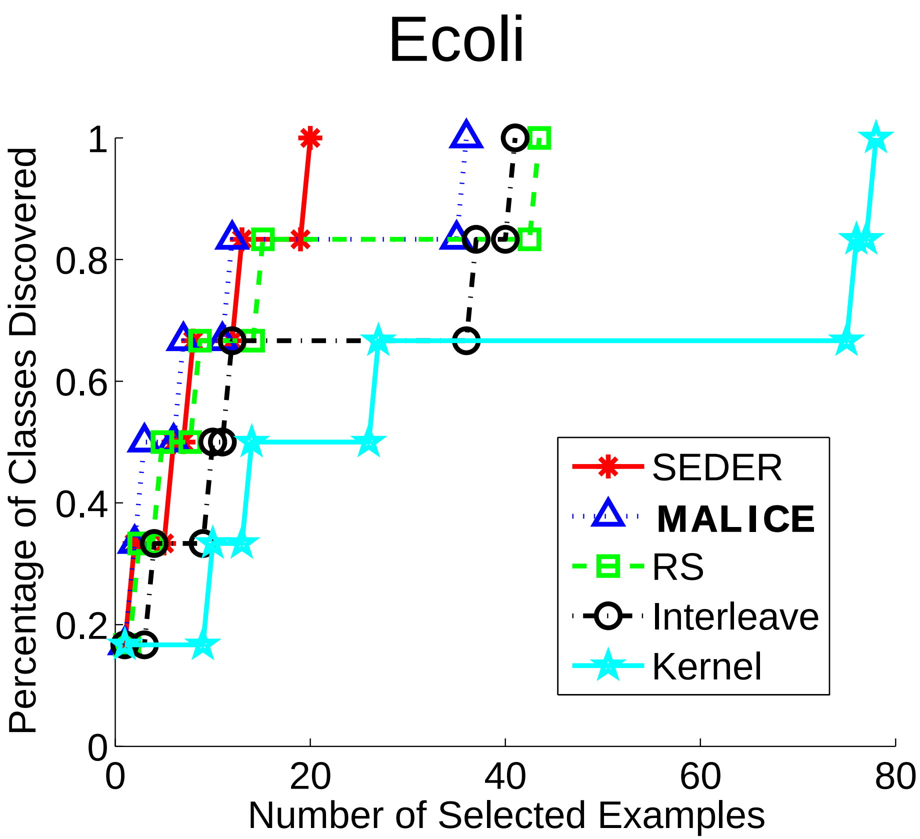


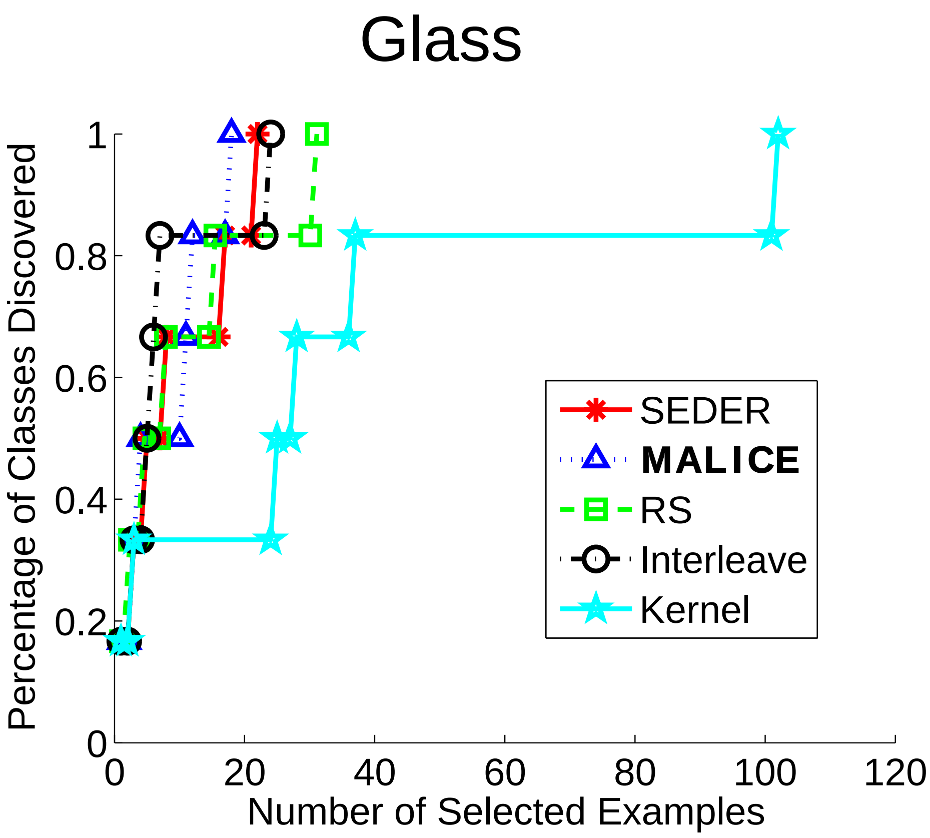


Summary of Real Data Sets

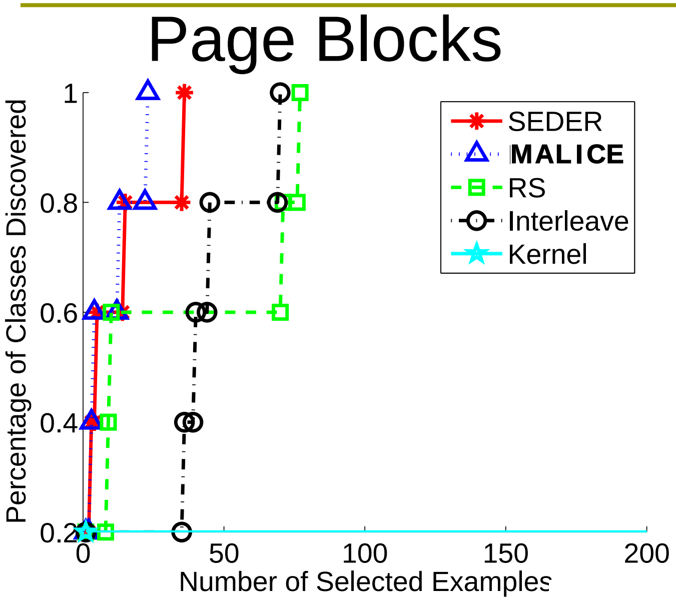
Data Set	*n*	*d*	*m*	Largest Class	Smallest Class
Moderately Skewed					
Ecoli	336	7	6	42.56%	2.68%
Glass	214	9	6	35.51%	4.21%
Extremely Skewed					
Page Blocks	5473	10	5	89.77%	0.51%
Abalone	4177	7	20	16.50%	0.34%
Shuttle	4515	9	7	75.53%	0.13%

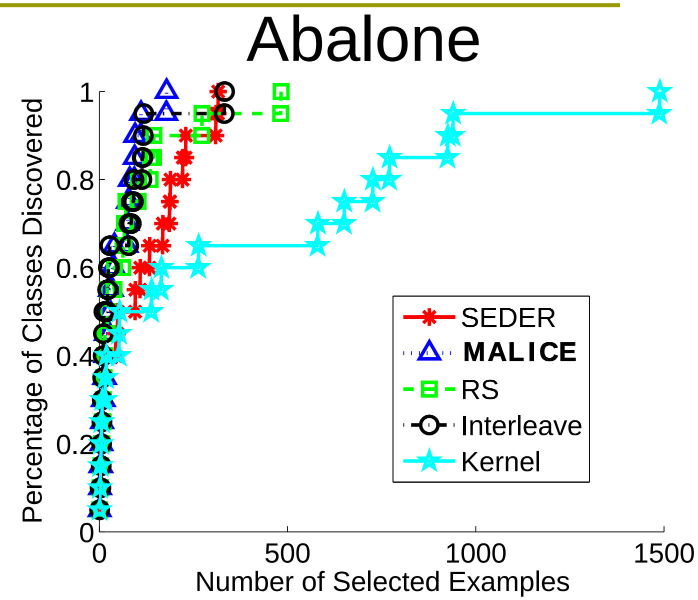
Moderately Skewed Data Sets

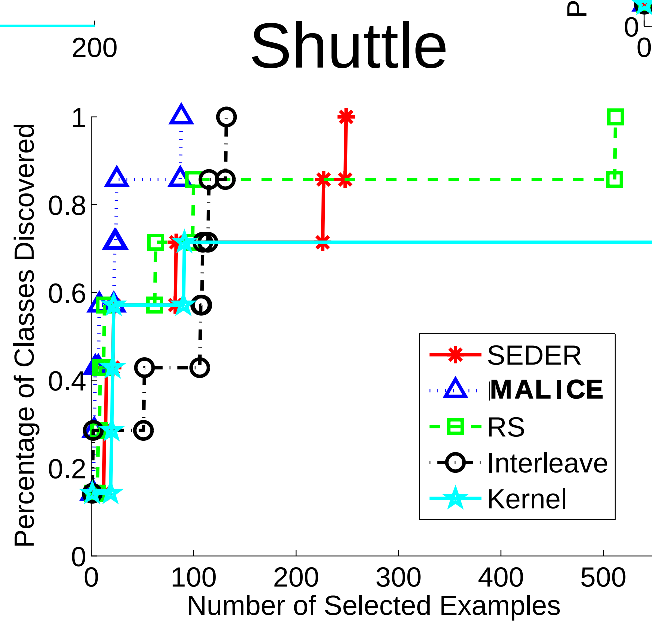




Extremely Skewed Data Sets







Conclusion

- Rare category detection
  - Open challenge
  - Lack of effective methods
- Nearest-neighbor-based methods
  - Prior-dependent
  - Local density differential sampling
- Density-based method
  - Prior-free
  - Specially designed exponential families

Thank You!