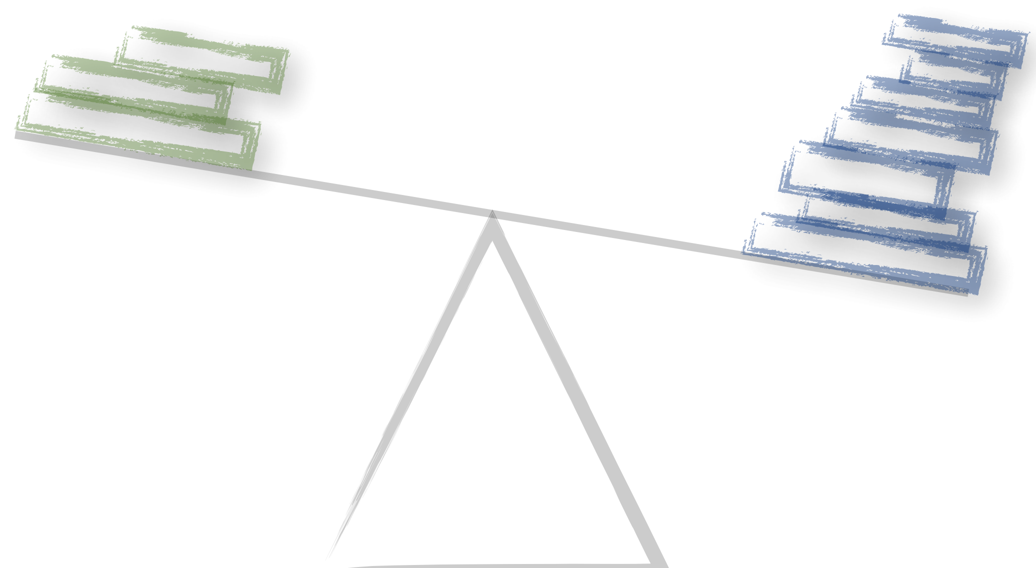
Sampling a longer life

Binary versus one-class classification revisited



Colin Bellinger, Shiven Sharma, Osmar R. Zaiane
and Nathalie Japkowicz

Motivation

- Class imbalance recognized as an important problem for two decades
- Generated many conference papers, workshops and special issues
    - Workshops CIPPP @ ICMLA 2012, LIDS I @ AAAI 2001, LIDS II @ ICML 2003, etc.
        - Workshop on Learning in the Presence of Class Imbalance and Concept Drift @ IJCAI 2017
        - Learning with Imbalanced Domains: Theory and Applications @ ECML 2017
    - Special Issues: SIGKDD Explorations Special Issue on Learning from Imbalanced Data Sets 2014
- We have advanced our understanding and developed many great methods

Motivation

- Class imbalance recognized as in important problem for two decades
  - Workshops CIPPP @ ICMLA 2012,
  - Special Issues

But which methods should we apply and when should we apply them?

- We have advanced our understanding and developed many great methods

Research Questions

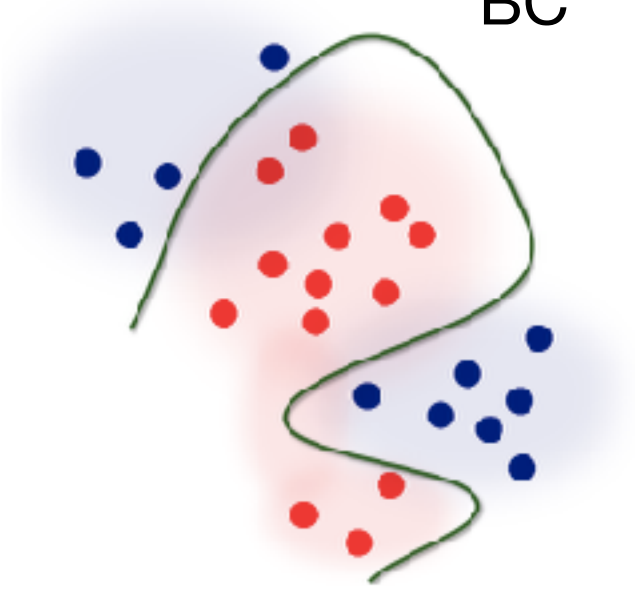
R1: *Given a dataset D, with complexity C, which paradigm should be applied?*

R2: *How are the paradigms affected by aspects of complexity?*

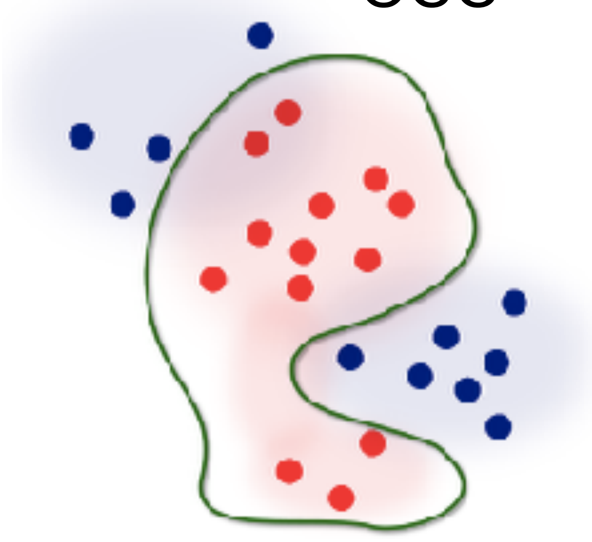
Classification Overview

- Classification Paradigms:
  - Binary classification (BC): learn from both classes
    - Make predictions about two
  - One-class classification (OCC): learn from a single class
    - Make predictions about two
    - Required for extreme imbalance
- Binary classifiers are perceived to be more powerful than OCC
  - Motivated much research to extend their usefulness
    - We focus on sampling

BC

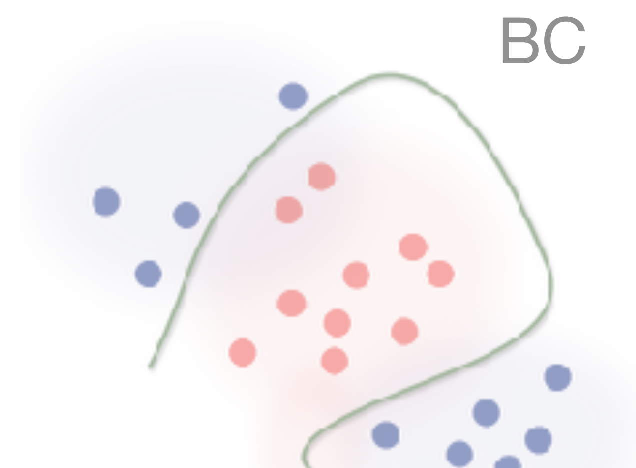


OCC

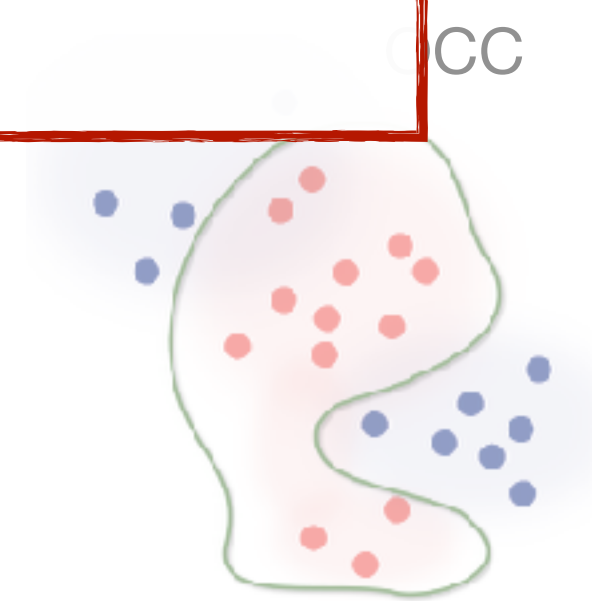


Classification Overview

- Classification Paradigms:
  - Binary classification (BC): learn from both classes
    - Make predictions about two
  - One-class classification (OCC): learn from a single class
    - Make predictions about two
    - Required for extreme imbalance
- Binary classifiers are perceived to be more powerful than OCC
  - Motivated much research to extend their usefulness

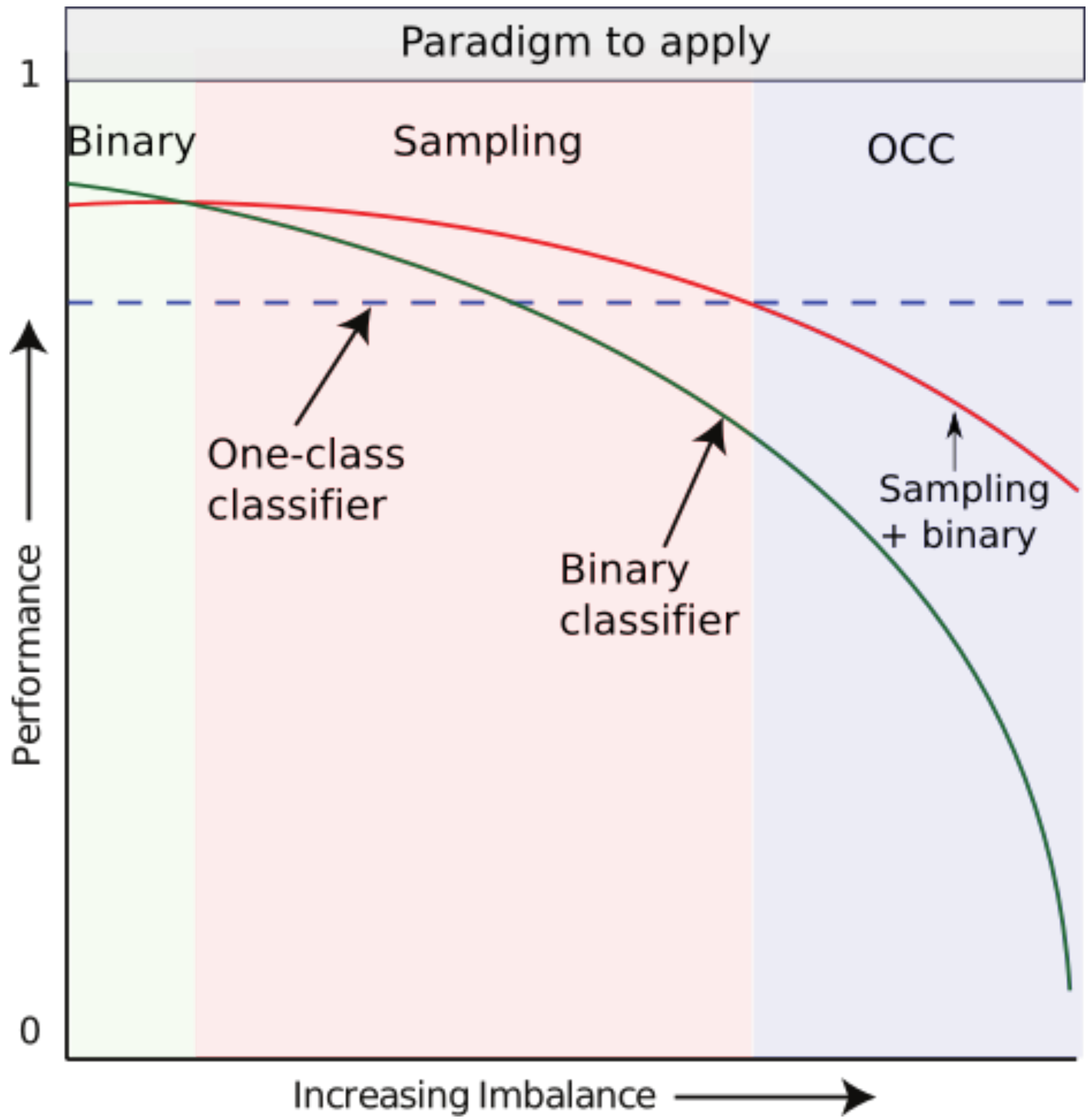


At which level of imbalance to switch to OCC?



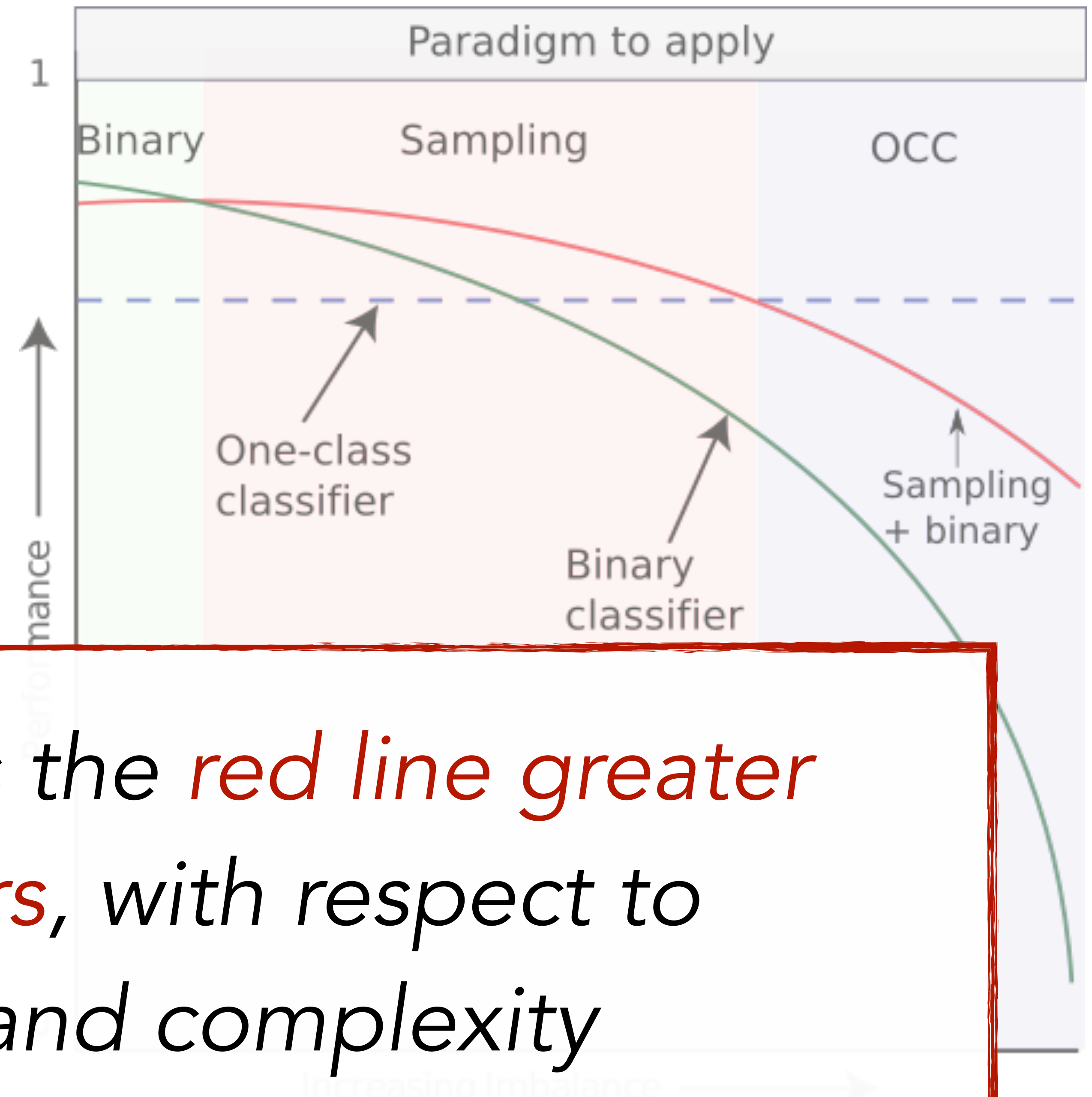
Perfomance Assessment

- Objective: intuitively assess the resilience of sampling to imbalance
  - Start with standard performance curves
  - x-axis = increasing class imbalance for a dataset
  - y-axis = classifier performance on the dataset



Perfomance Assessment

- Objective: intuitively assess the resilience of sampling to imbalance
    - Start with standard performance curves
    - x-axis = increasing class imbalance for a dataset
    - y-axis = … performance on the dataset



To what extent is the red line greater than the others, with respect to imbalance and complexity

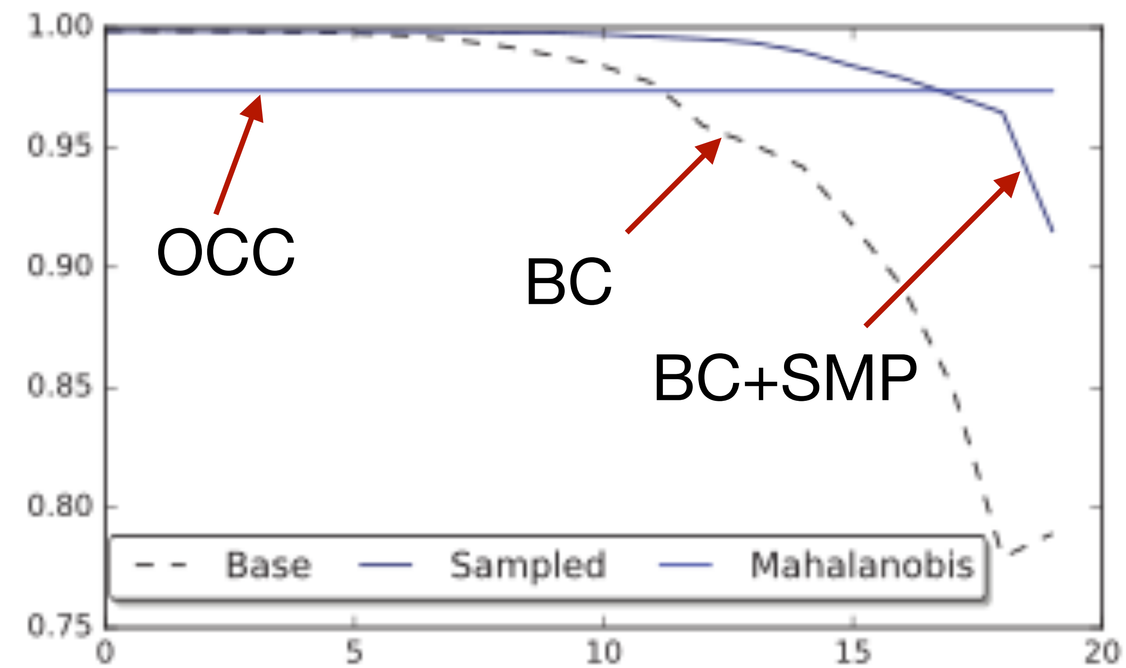
Experimental Framework

- 4 artificial datasets (various combinations of modality and overlap)
- 14 benchmark datasets
- Binary classifiers: NB, MLP, kNN, DT, SVM
- One-class classifiers: AE, one-class SVM, Mahalanobis distance
- Sampling methods: ROS, RUS, SMOTE, Borderline SMOTE, SMOTE with OSS, ADASYN
- Evaluation: g-mean over increasing imbalance
    - We report the best method from each category only
- To understand the impact of imbalance
    - Minority class under-sampled exponentially from the original size to 4

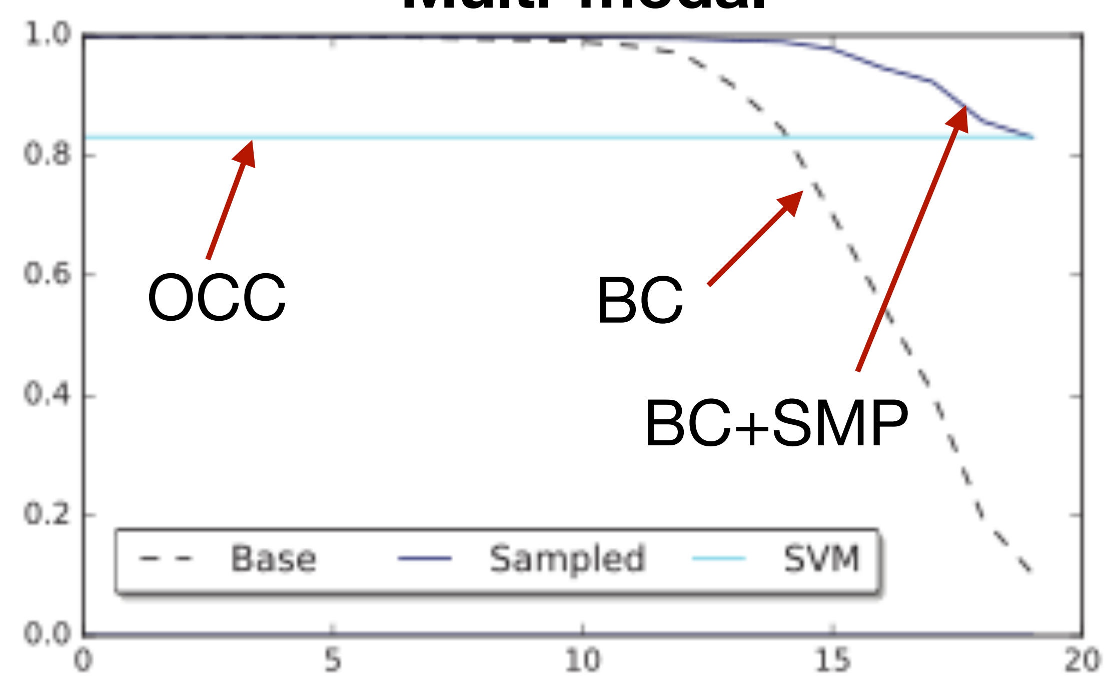
Results - Artificial Data

- Trends over unimodal and multimodal data with no overlap
- G-mean on the y-axis
- Increasing imbalance from left to right on the x-axis

Uni-modal



Multi-modal

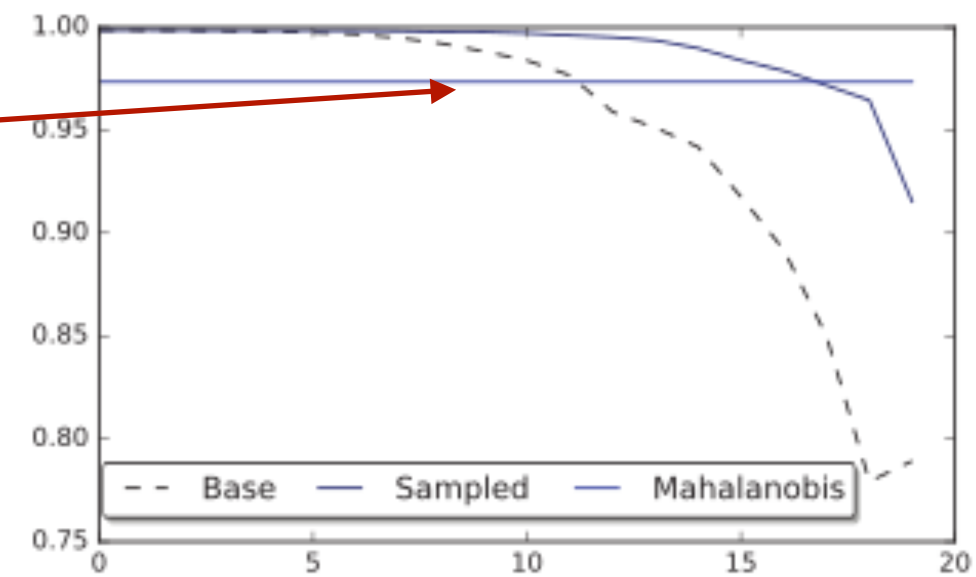


Results - Artificial Data

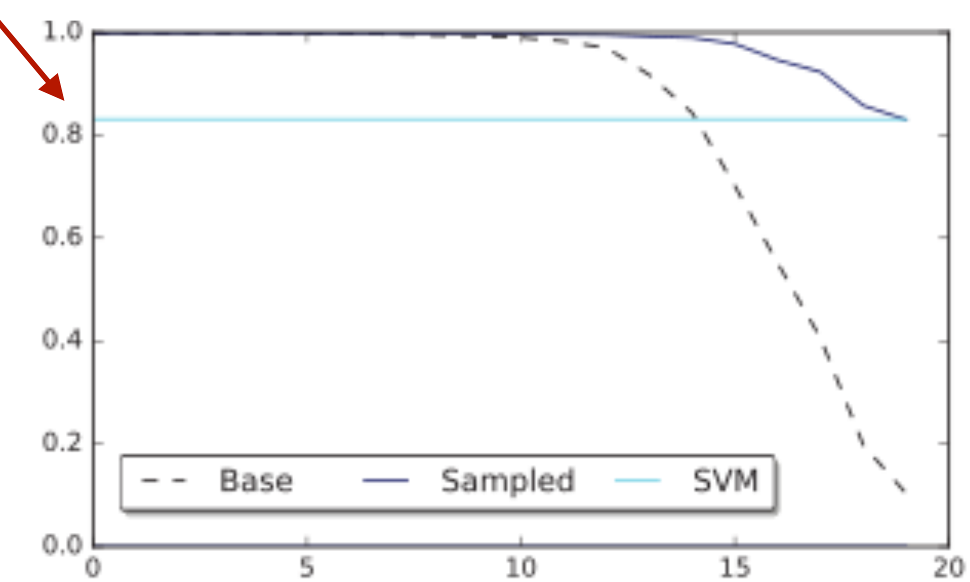
Trends over unimodal and multimodal data with no overlap

OCC: weaker on multi-modal data. 0.97 versus 0.81

Uni-modal



Multi-modal



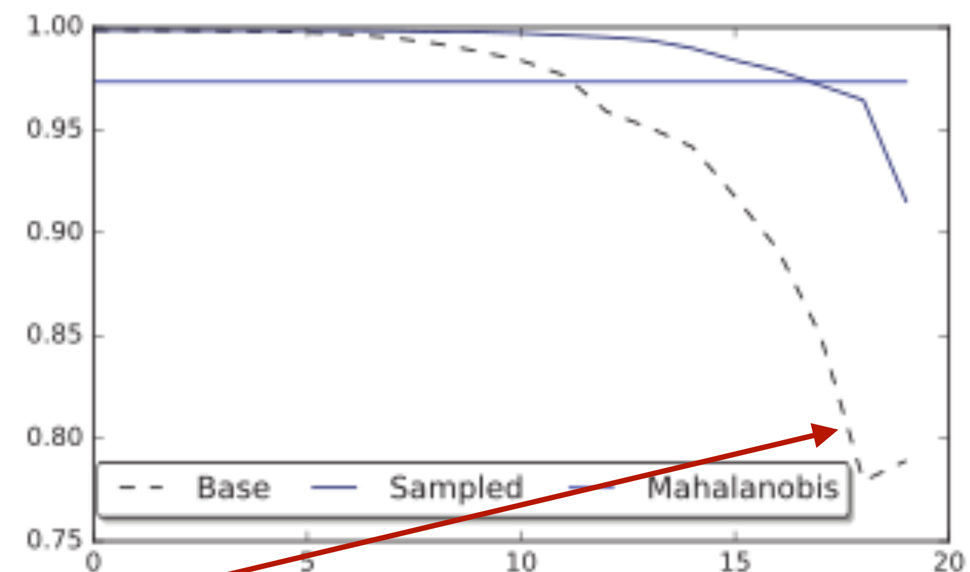
Results - Artificial Data

Trends over unimodal and multimodal data with no overlap

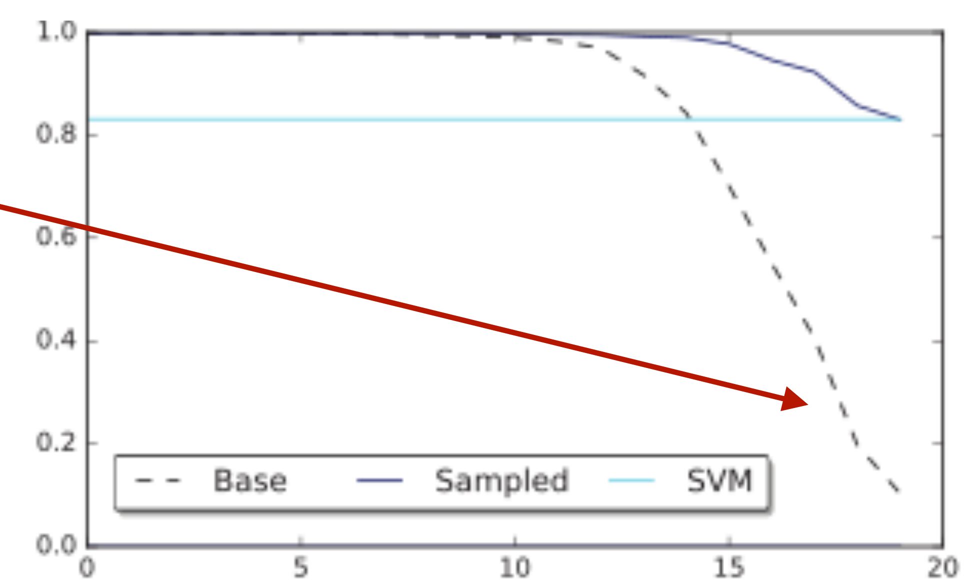
OCC: weaker on multi-modal data. 0.97 versus 0.81

BC: weaker on multi-modal data. 0.77 versus 0.10

Uni-modal



Multi-modal



Results - Artificial Data

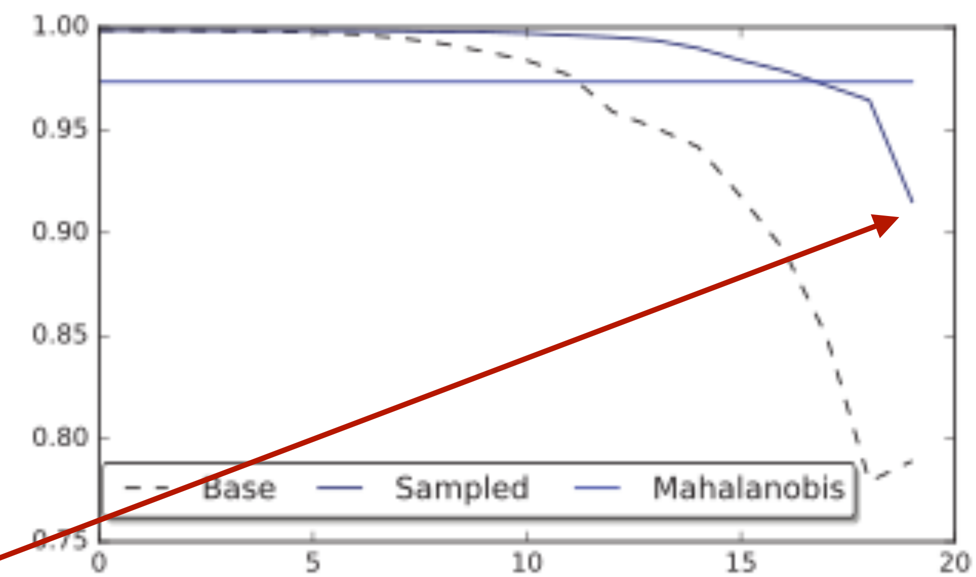
Trends over unimodal and multimodal data with no overlap

OCC: weaker on multi-modal data. 0.97 versus 0.81

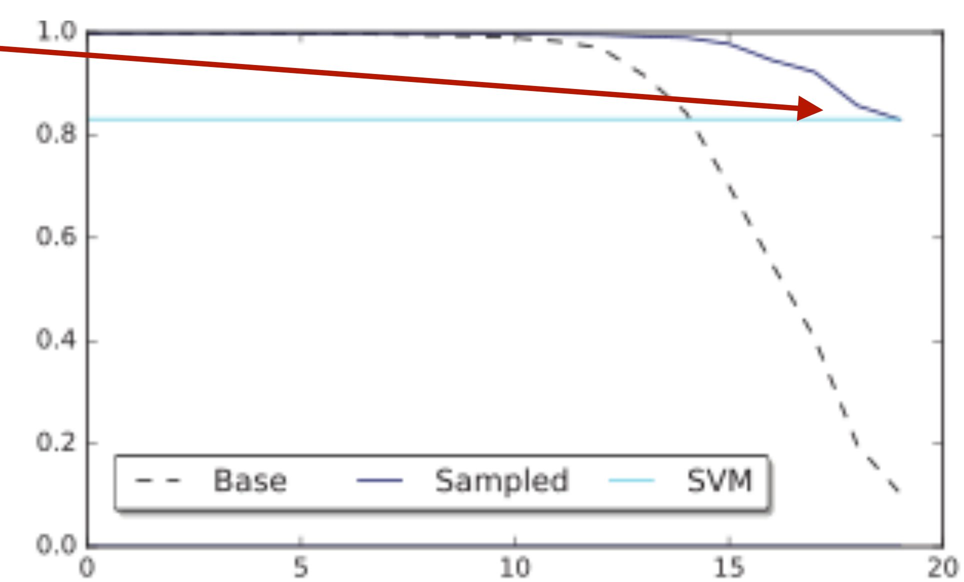
BC + SMP: weaker on multi-modal data. 0.92 versus 0.81

BC: weaker on multi-modal data. 0.77 versus 0.10

Uni-modal



Multi-modal

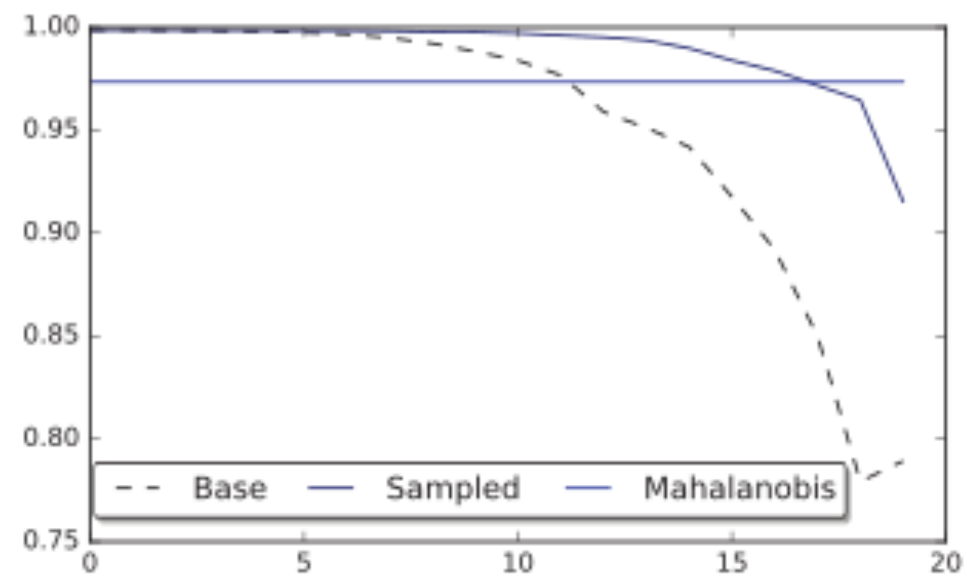


Results - Artificial Data

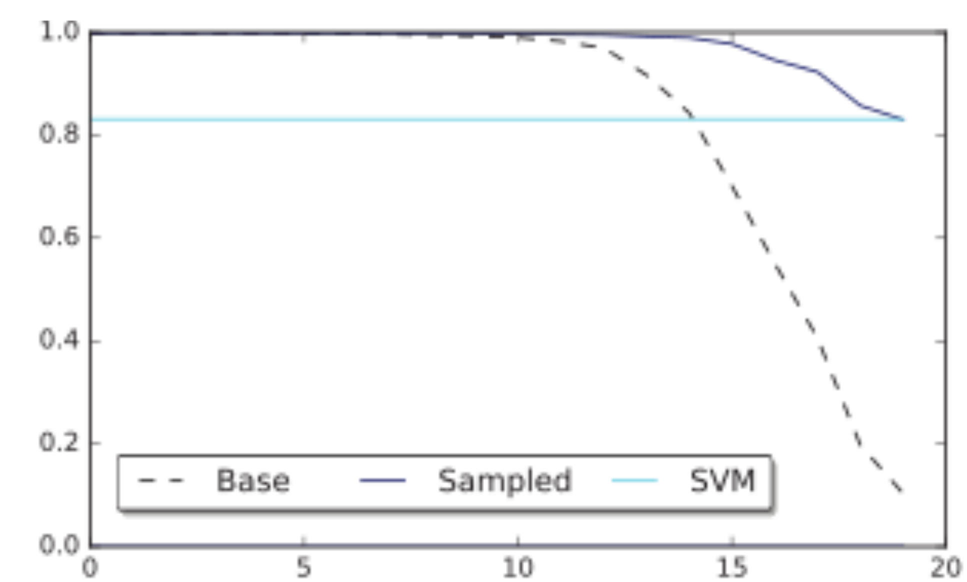
Trends over unimodal and multimodal data with no overlap

- OCC impacted by modality
- BC and BC+SMP are more affected by imbalance on multi-modal domains
- BC+SMP is always more robust than OCC on multi-modal

Uni-modal



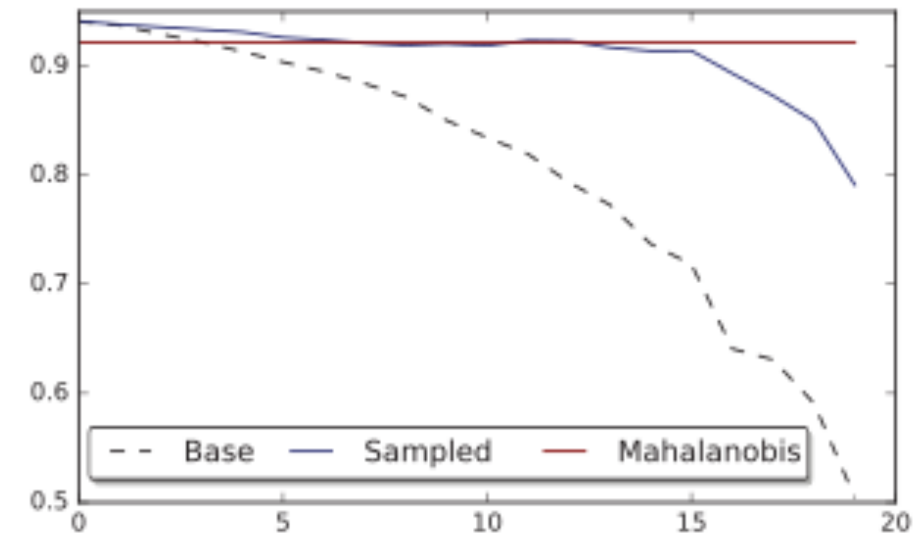
Multi-modal



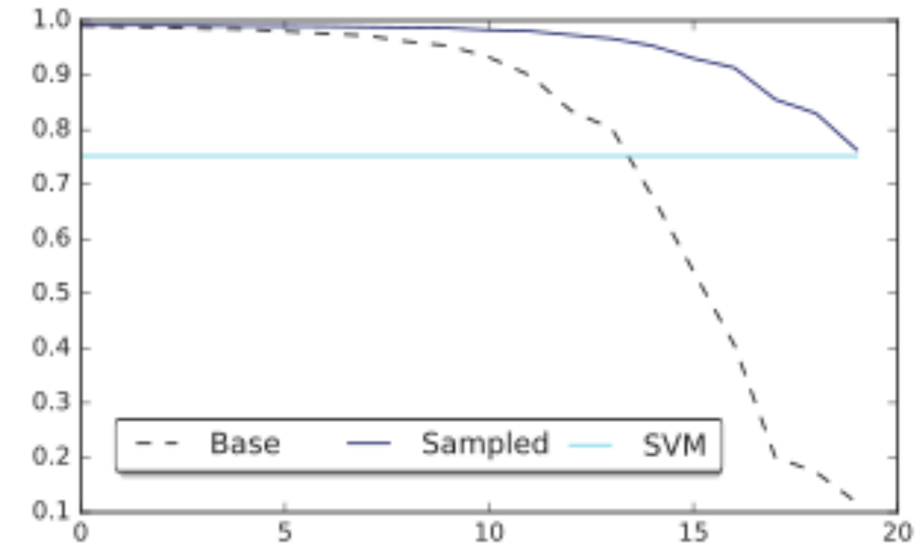
Results - Artificial Data

Trends over unimodal and multimodal data with overlap

Uni-modal



Multi-modal

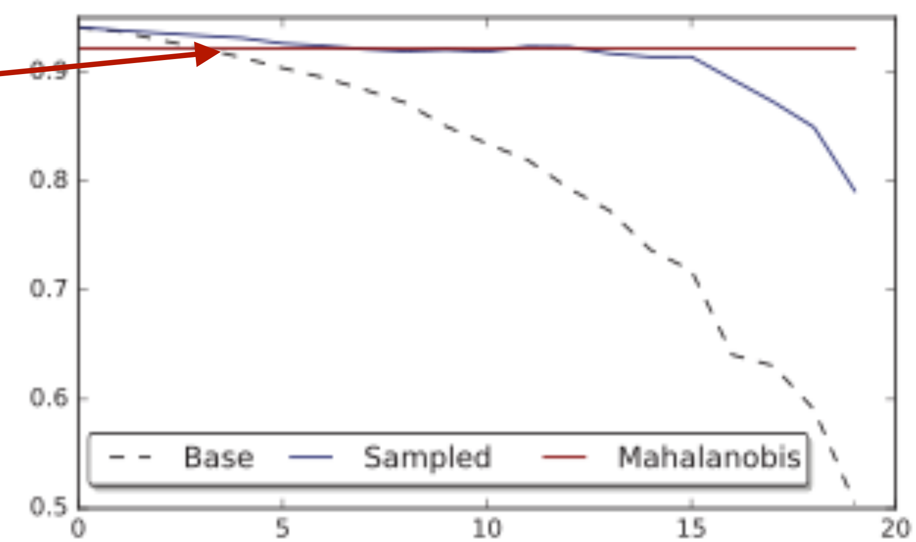


Results - Artificial Data

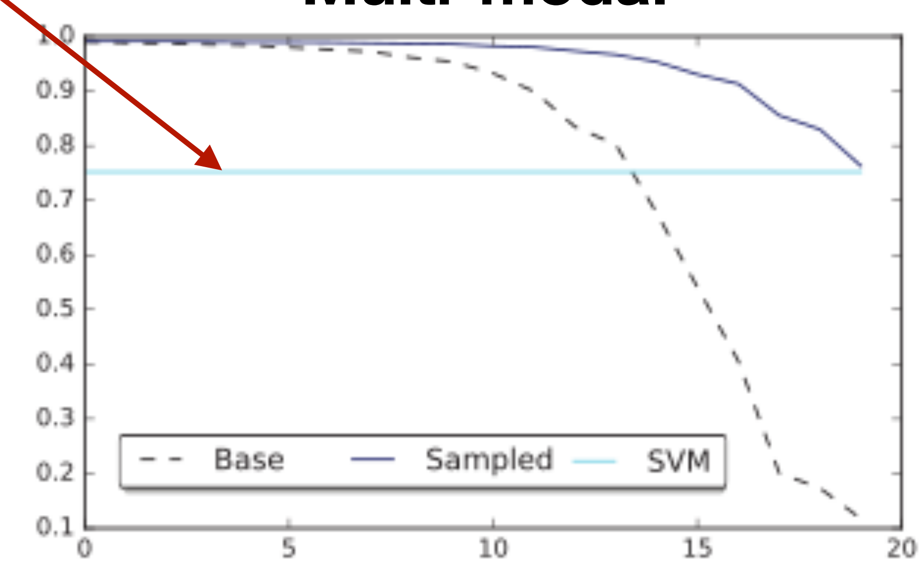
Trends over unimodal and multimodal data with overlap

OCC: weaker on multi-modal data.
0.94 versus 0.74

Uni-modal



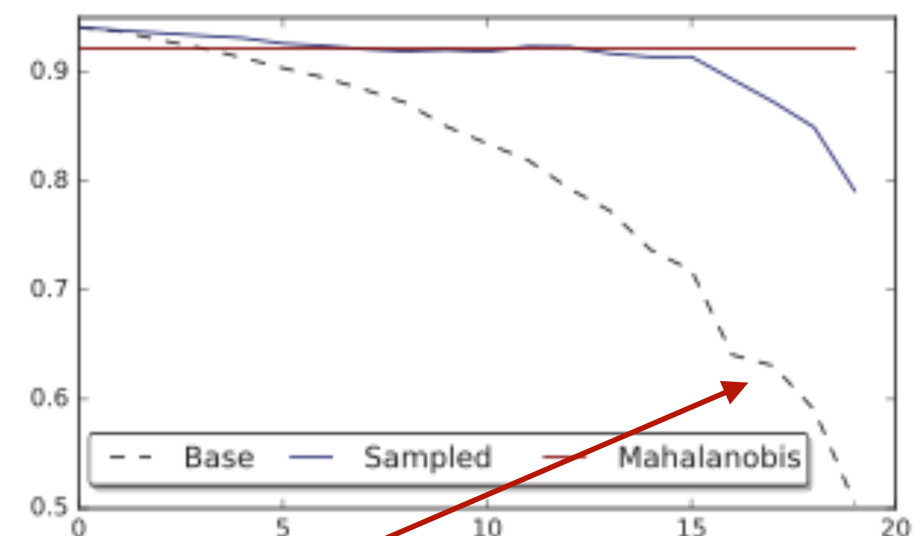
Multi-modal



Results - Artificial Data

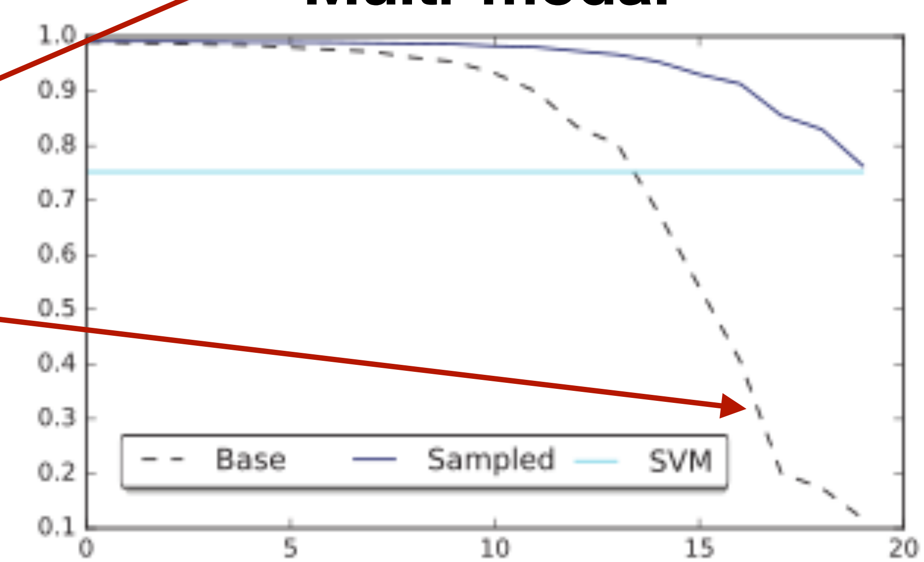
Trends over unimodal and multimodal data with overlap

Uni-modal



OCC: weaker on multi-modal data. 0.97 versus 0.81

Multi-modal



BC: weaker on multi-modal data. 0.50 versus 0.10

Results - Artificial Data

Trends over unimodal and multimodal data with overlap

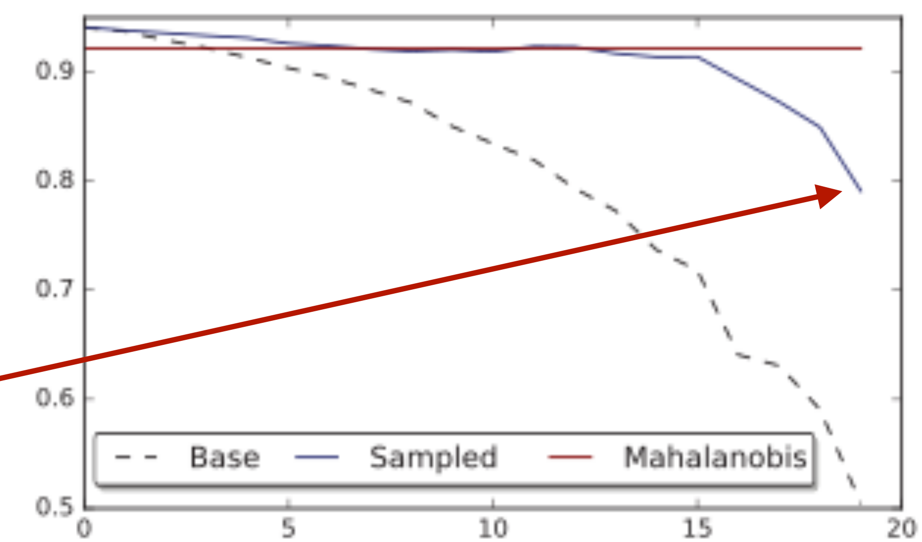
OCC: weaker on multi-modal data. 0.97 versus 0.81

BC + SMP: weaker on multi-modal data. 0.79 versus 0.74.

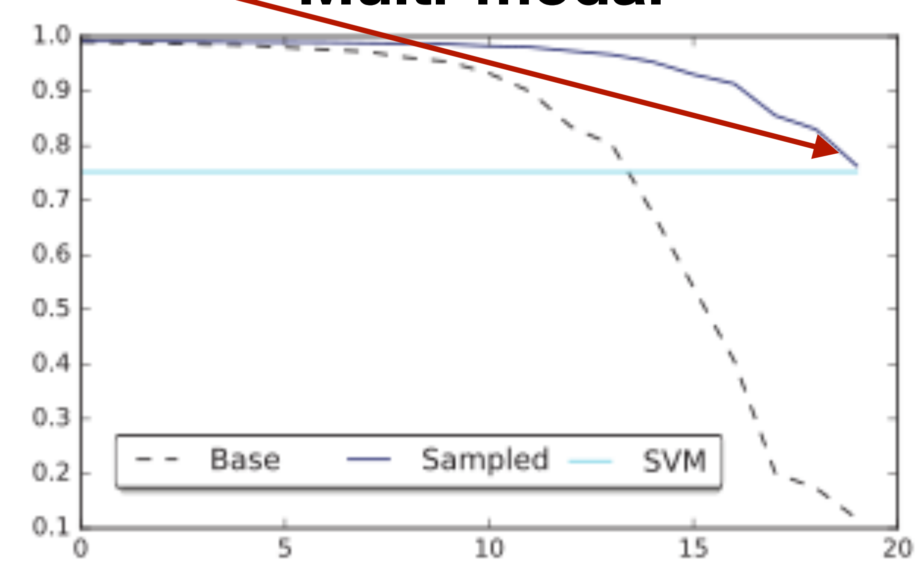
The degradation is small relative to others

BC: weaker on multi-modal data. 0.50 versus 0.10

Uni-modal



Multi-modal



Results - Artificial Data

Trends over unimodal and multimodal data with overlap

OCC: weaker on multi-modal data. 0.97 versus 0.81

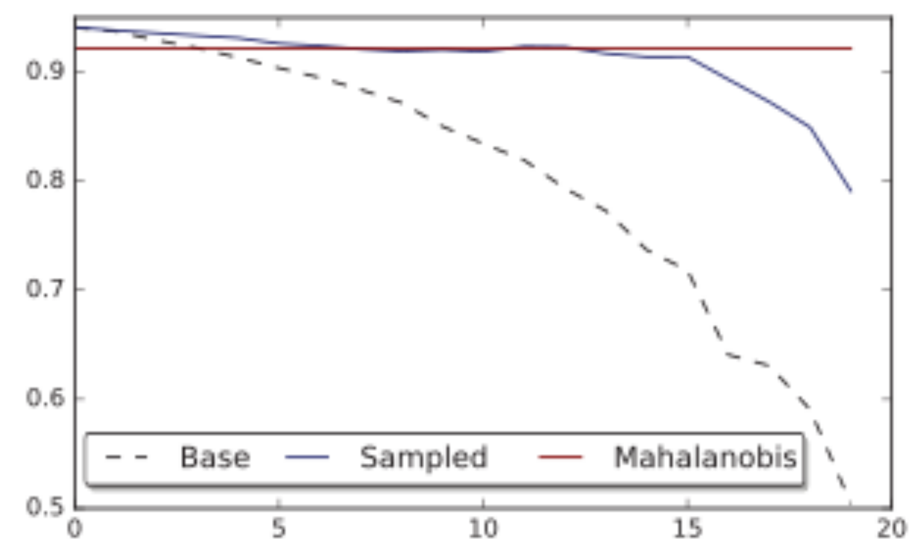
- OCC impacted by modality
- Again, imbalance is made worse by complexity
  - Particularly modality

BC+SMP: weaker on multi-modal data. 0.79 versus 0.74

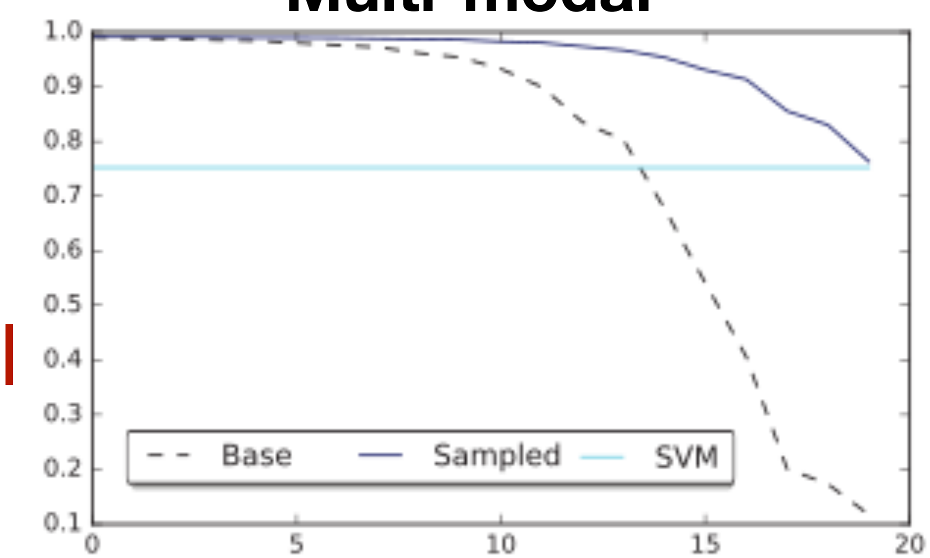
BC: weaker on multi-modal data. 0.50 versus 0.10

- BC+SMP is always more robust then OCC on multi-modal

Uni-modal



Multi-modal

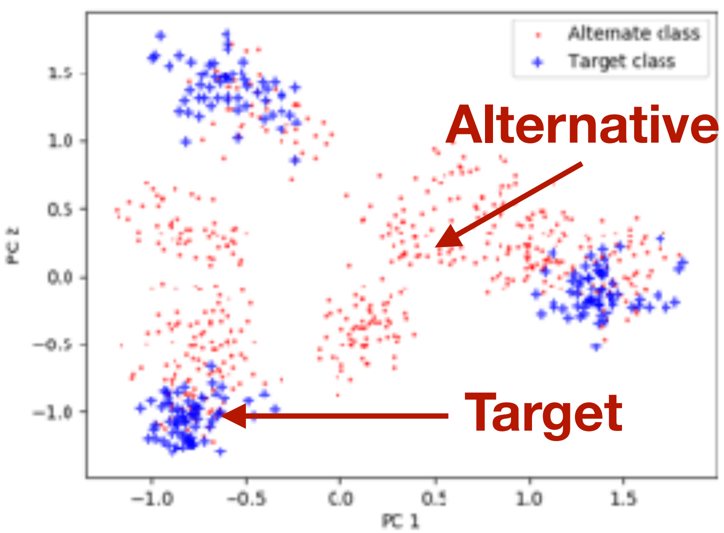


Results - Benchmark

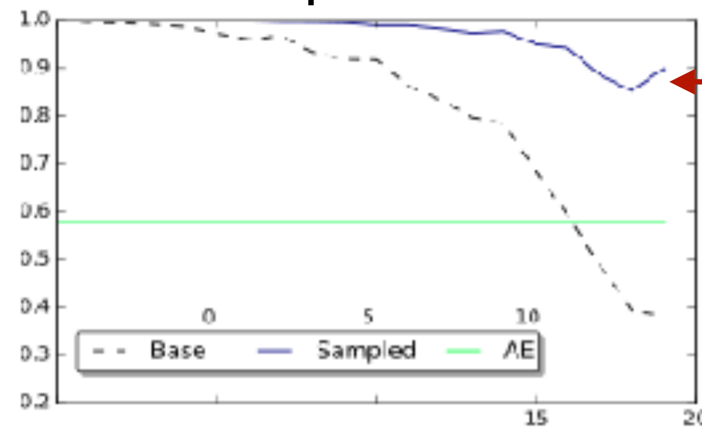
Hypothesis demonstrated on three benchmark datasets

More in the paper

Results - Benchmark

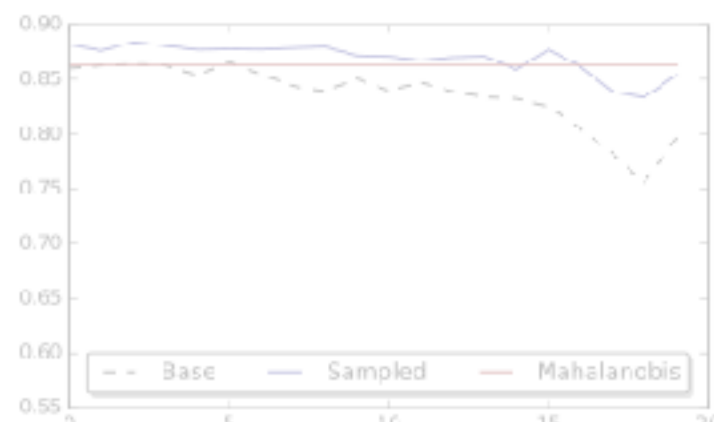


Delft Pump - Multi-modal + overlap

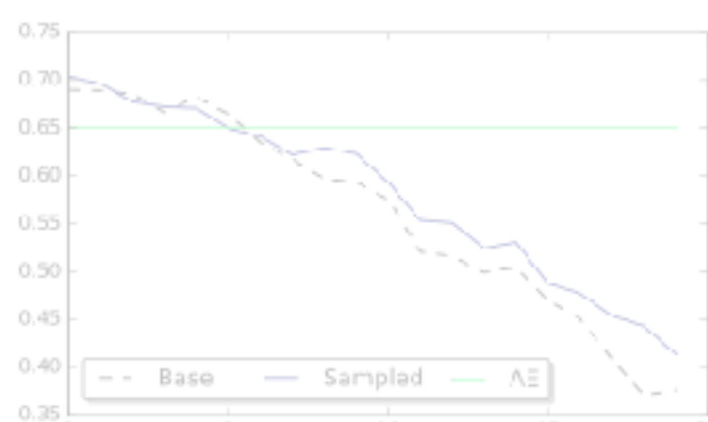


BC+SMP always stronger than OCC

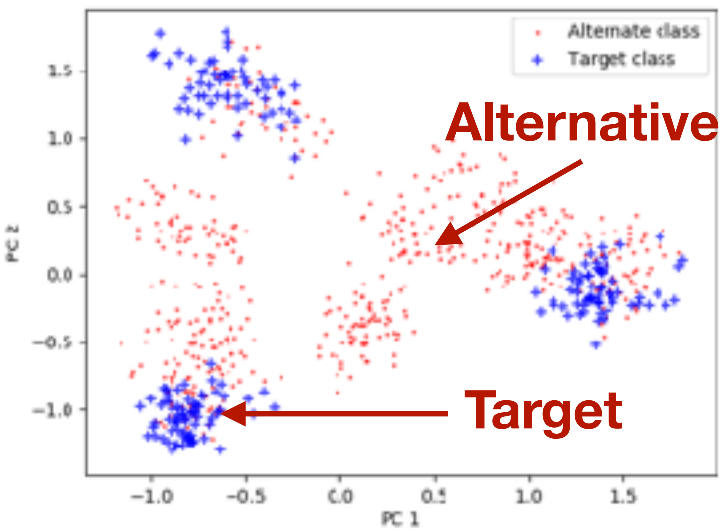
Biomed - uni-modal + moderate overlap



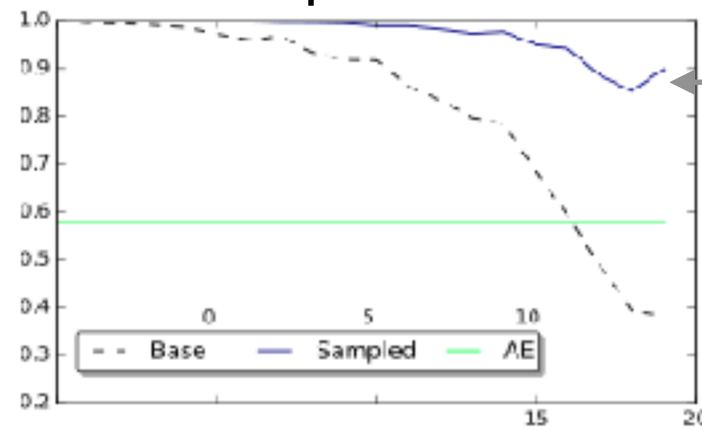
Arrhythmia - uni-modal + significant overlap



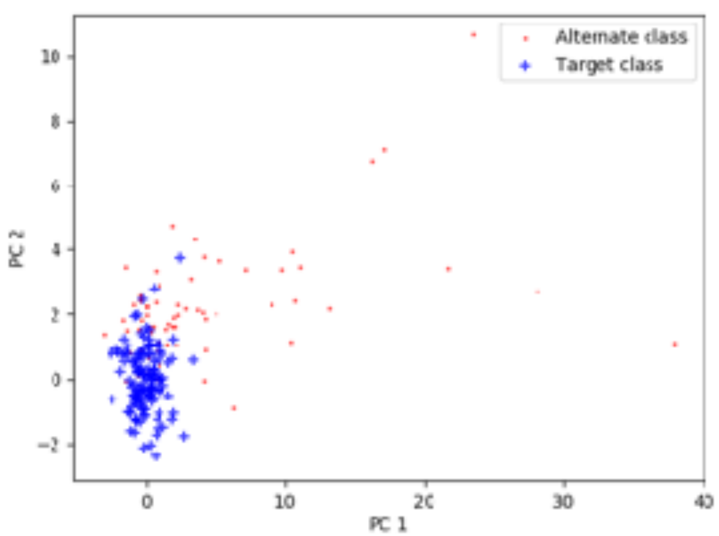
Results - Benchmark



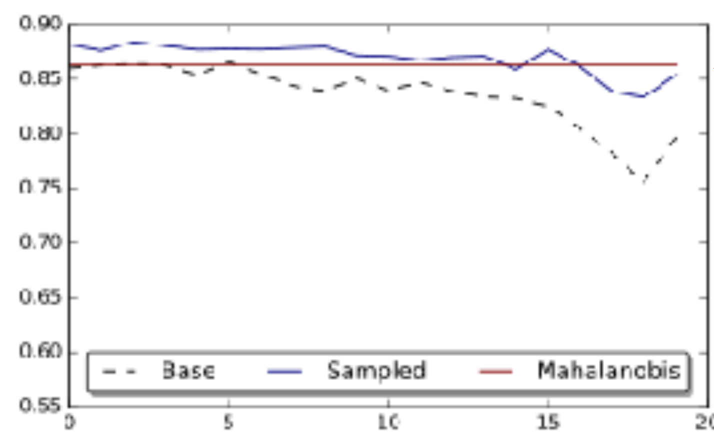
Delft Pump - Multi-modal + overlap



BC+SMP stronger than OCC

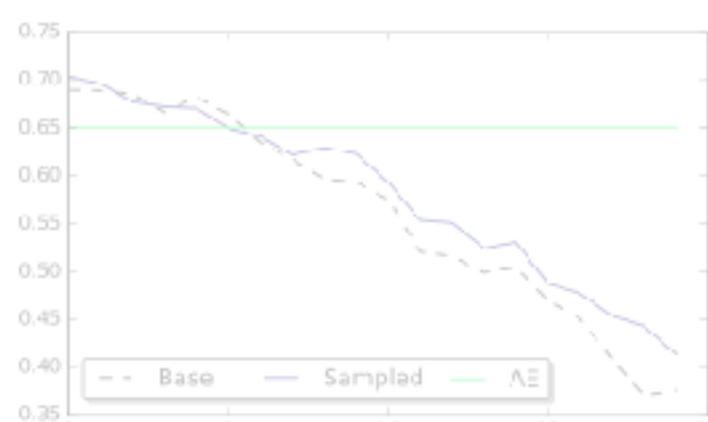


Biomed - uni-modal + moderate overlap

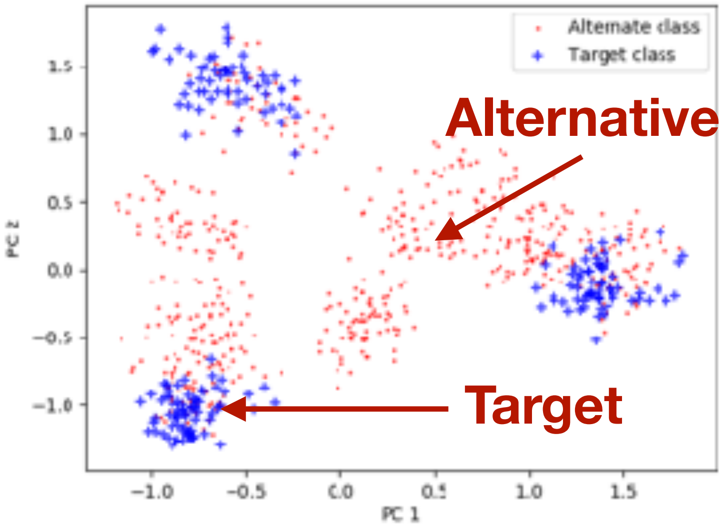


OCC better than BC+SMP for significant imbalance

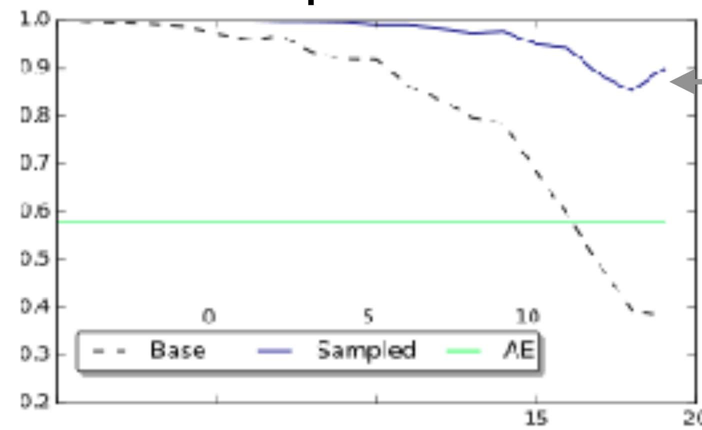
Arrhythmia - uni-modal + significant overlap



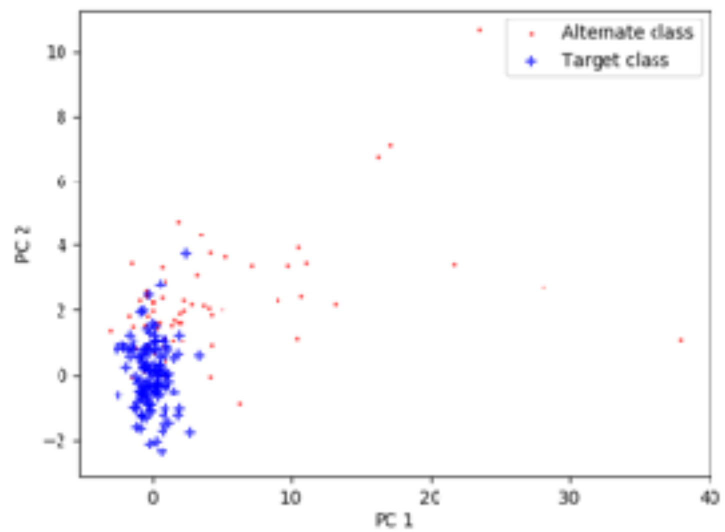
Results - Benchmark



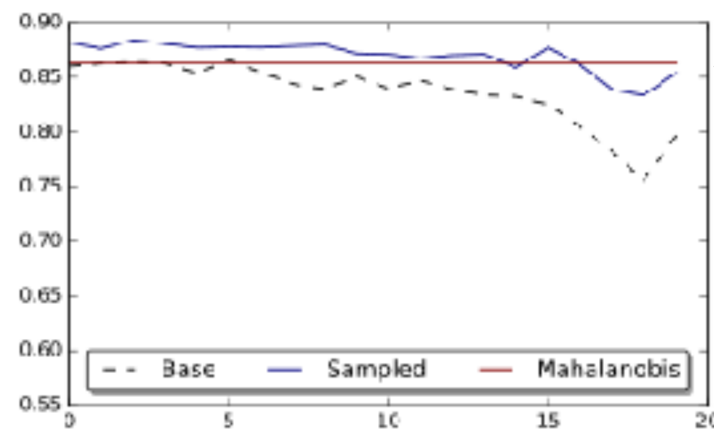
Delft Pump - Multi-modal + overlap



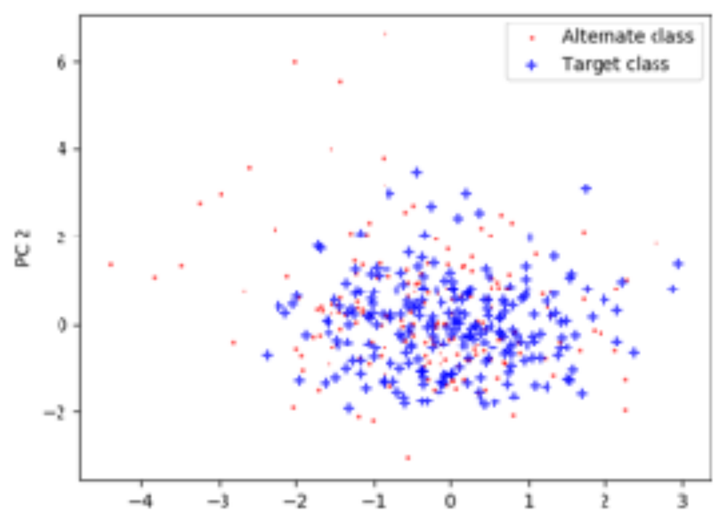
BC+SMP stronger than OCC



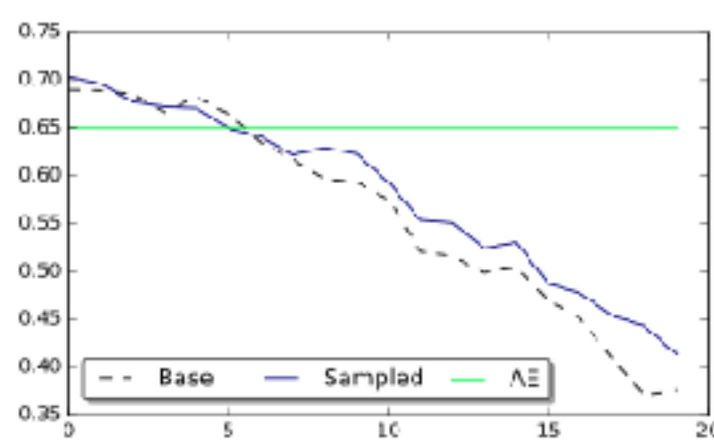
Biomed - uni-modal + moderate overlap



OCC better than BC+SMP for significant imbalance



Arrhythmia - uni-modal + significant overlap



OCC better than BC+SMP for moderate to significant imbalance

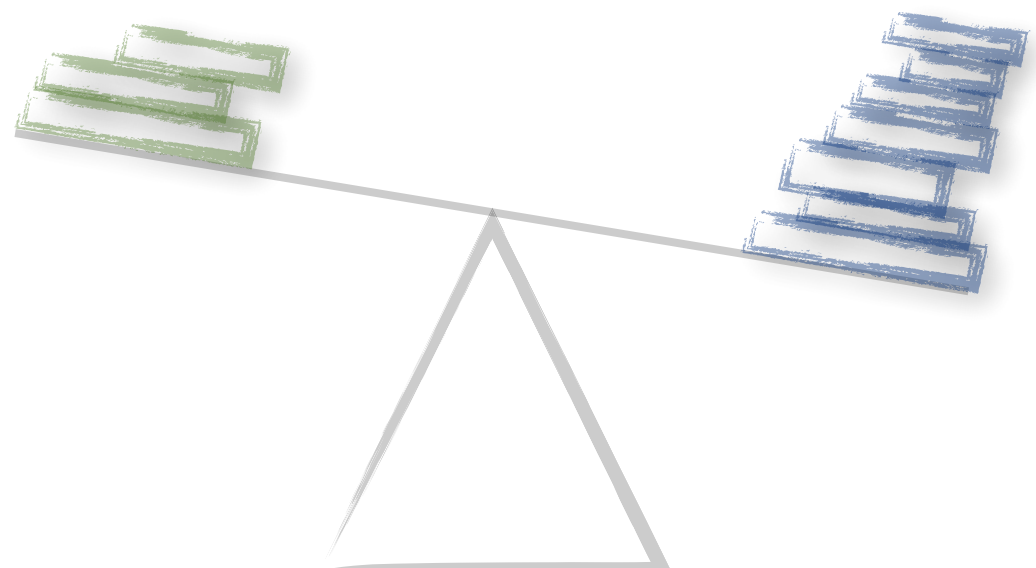
Conclusion

Application prescription

Modality \ Overlap	Low	High
Low	* BC+SMP/OCC	OCC
High	BC+SMP	BC+SMP

- Future work:
  - Other domain properties such as dimensionality, variance and feature correlations, etc.
  - Meta-learning on data properties and choice of classification method
  - Analyze the trends at an algorithmic level

Thank You!



UNIVERSITY OF ALBERTA

AMERICAN UNIVERSITY WASHINGTON, DC

FLUENT

Dataset	$\lvert M \rvert$	d	*Modality*	*Overlap*
Diabetes	66	8	unimodal	significant
Sonar	24	60	unimodal,spread	significant
Delft pump AR app.†	94	160	multi-modal	significant
Alphabets	749	15	multi-modal	significant
ForestC1	624	54	unimodal	significant
Biomed healthy†	33	5	unimodal,spread	moderate
Waveform 0†	149	21	unimodal	moderate
Heart	30	13	multi-modal	moderate
Cancer wpbc non-ret†	23	33	unimodal	significant
Spambase spam†	906	57	unimodal,spread	moderate
ForestC2C5	600	54	unimodal	moderate
Forest	742	54	bimodal	moderate
Ionosphere	31	34	unimodal	moderate
Arrhythmia normal†	91	278	unimodal	significant

Table 2: Classifier statistics for the benchmark datasets.

Dataset	B_{nat}	B_{imb}	S_{nat}	S_{imb}	OCC	ℓ	$size$
Delft pump	0.997	0.384	0.999	0.852	0.578	∞	$-/5/94$
Diabetes	0.657	0.226	0.743	0.634	0.615	∞	$-/42/66$
Sonar	0.721	0.418	0.773	0.656	0.543	∞	$-/9/24$
Alphabets	0.972	0.081	0.990	0.524	0.707	16.23	9/146/769
ForestC1	0.803	0.623	0.823	0.674	0.724	11.00	15/165/624
Waveform	0.868	0.269	0.880	0.719	0.777	5.20	5/26/149
Biomed	0.866	0.755	0.884	0.833	0.863	2.60	6/16/33
Cancer wpbc	0.566	0.330	0.611	0.490	0.496	2.30	4/9/23
Heart	0.821	0.575	0.830	0.731	0.740	2.20	4/9/30
Ionosphere	0.846	0.659	0.868	0.687	0.846	1.93	16/31/31
Spambase	0.895	0.458	0.909	0.604	0.801	1.37	16/22/906
ForestC2C5	0.827	0.474	0.851	0.604	0.754	1.28	25/32/600
Forest	0.900	0.546	0.911	0.647	0.789	1.24	12/15/742
Arrhythmia	0.680	0.379	0.703	0.413	0.650	1.20	39/47/91