AEGO

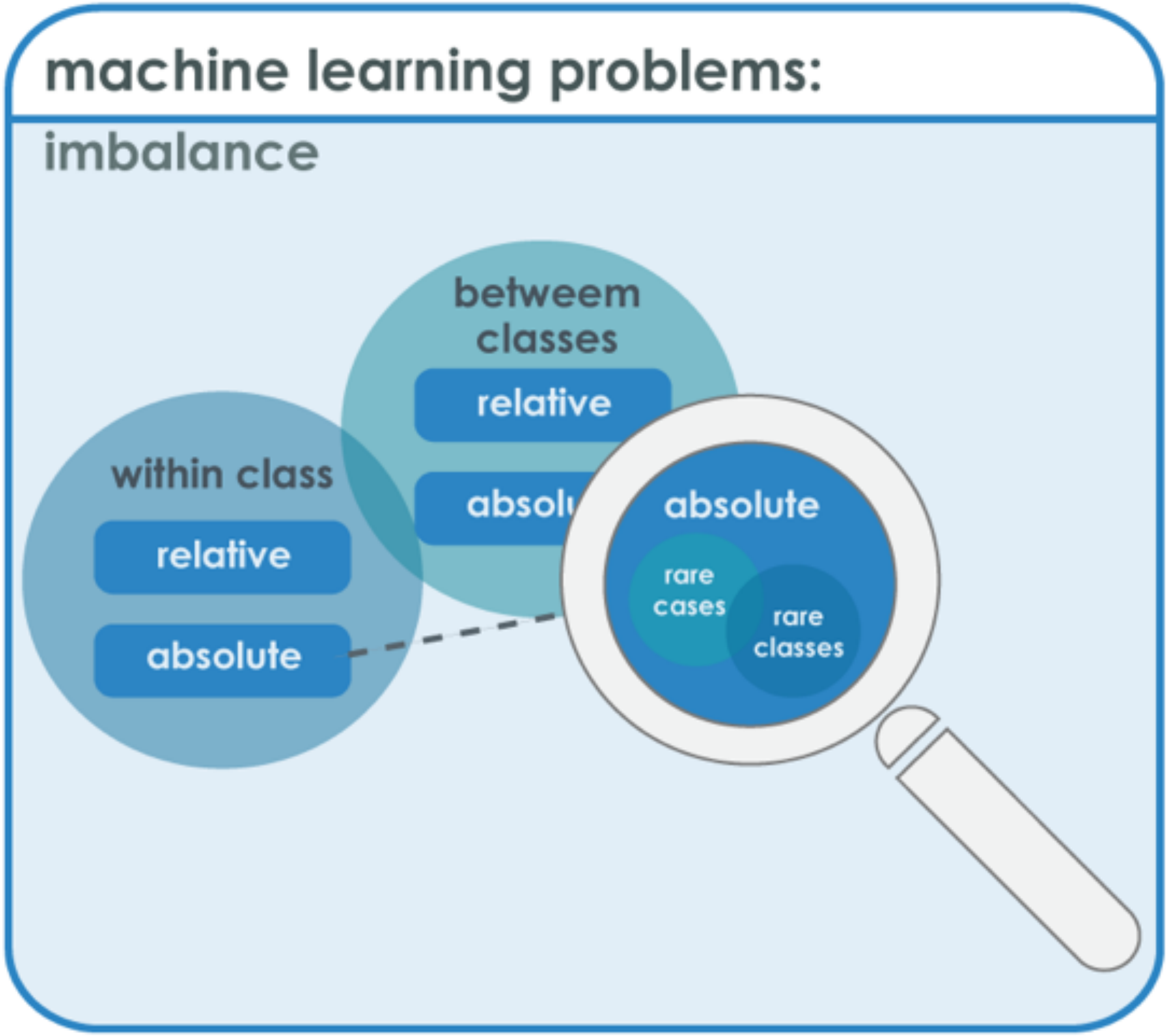
Denoising Autoencoder-Based Generative Minority Oversampling

Colin Bellinger
Electrical Engineering and Computer Science
University of Ottawa
cbell052@uottawa.ca

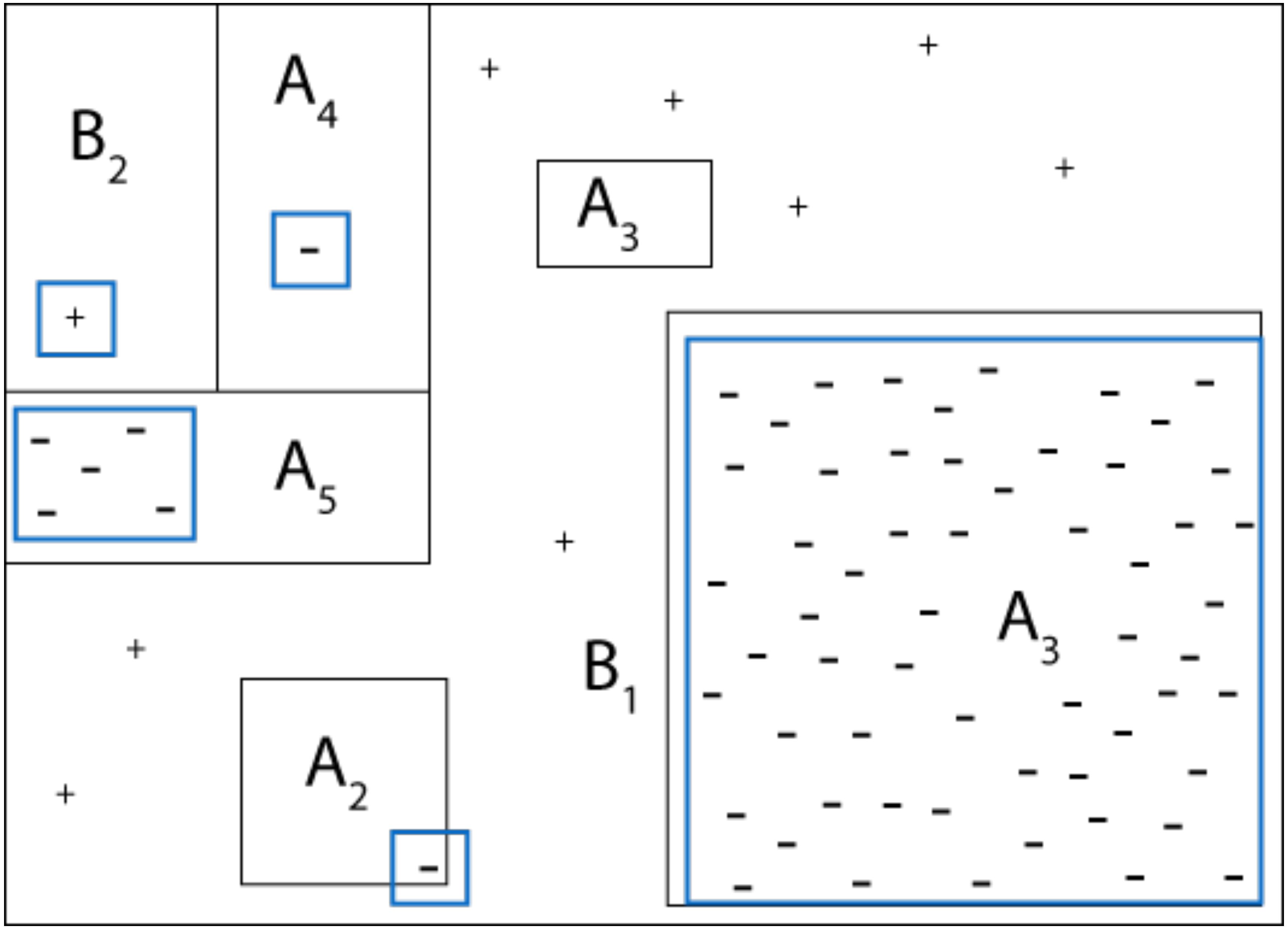
Outline

- Class imbalance
- Coping with Imbalance
    - Sampling
    - Cost adjustments
    - SMOTE
- AEGO: denoising AutoEncoder-based Generative Oversampling
- Experiments
    - Artificial
    - UCI
    - Gamma-ray spectral data
- Conclusions

Types of Imbalance



Imbalance Realized

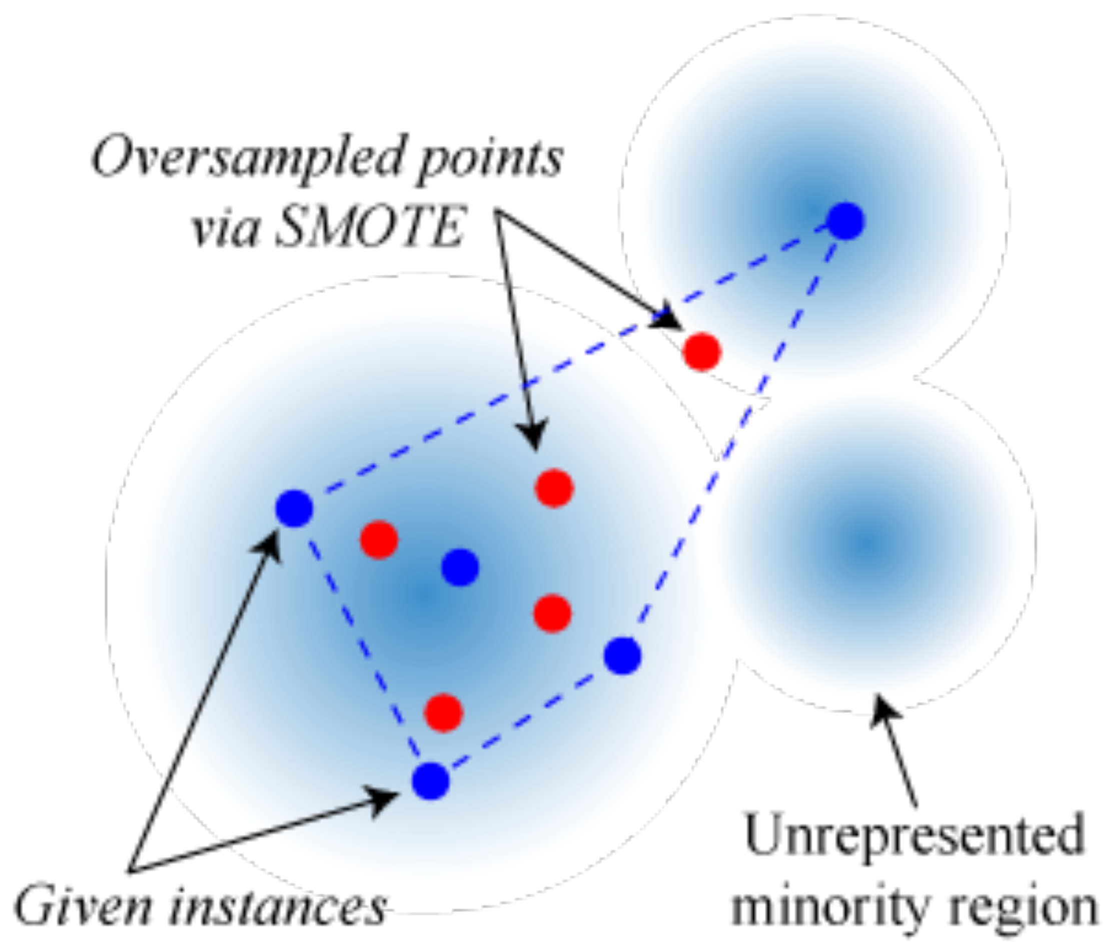


Coping with Imbalance

- Sampling
  - RUS
    - randomly undersampling majority class
  - ROS
    - randomly oversample minority class
  - Informative sampling
    - remove border/overlap
- Cost Adjustment
  - “encourage” correct classification of minority class

SMOTE

- Synthetically oversample minority class
- RUS majority



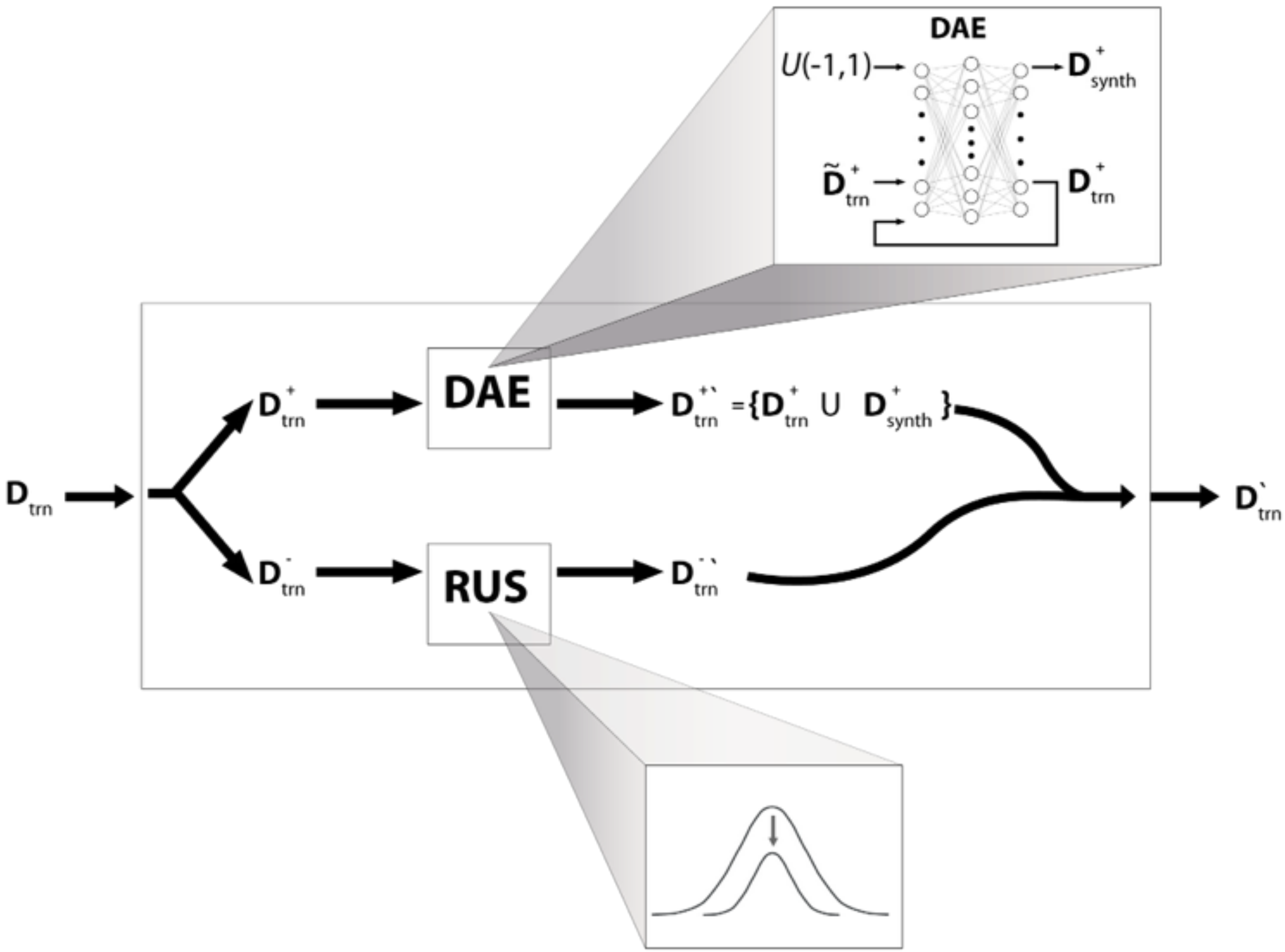
Pros and Cons

	Cost	ROS	RUS	SMOTE
Strengths	- Reflect cost - Force min clsf	- Simple	- Simple - Decrease bias of maj class	- Synth minority inst - RUS
Weaknesses	- Mod algo - Select costs - Overfit - Requires overlap - No new min info	- Overfit - Requires overlap - No new min info	- Lost info - Variable - No new min info	- Convex hull - *k*NN - Requires overlap

AEGO

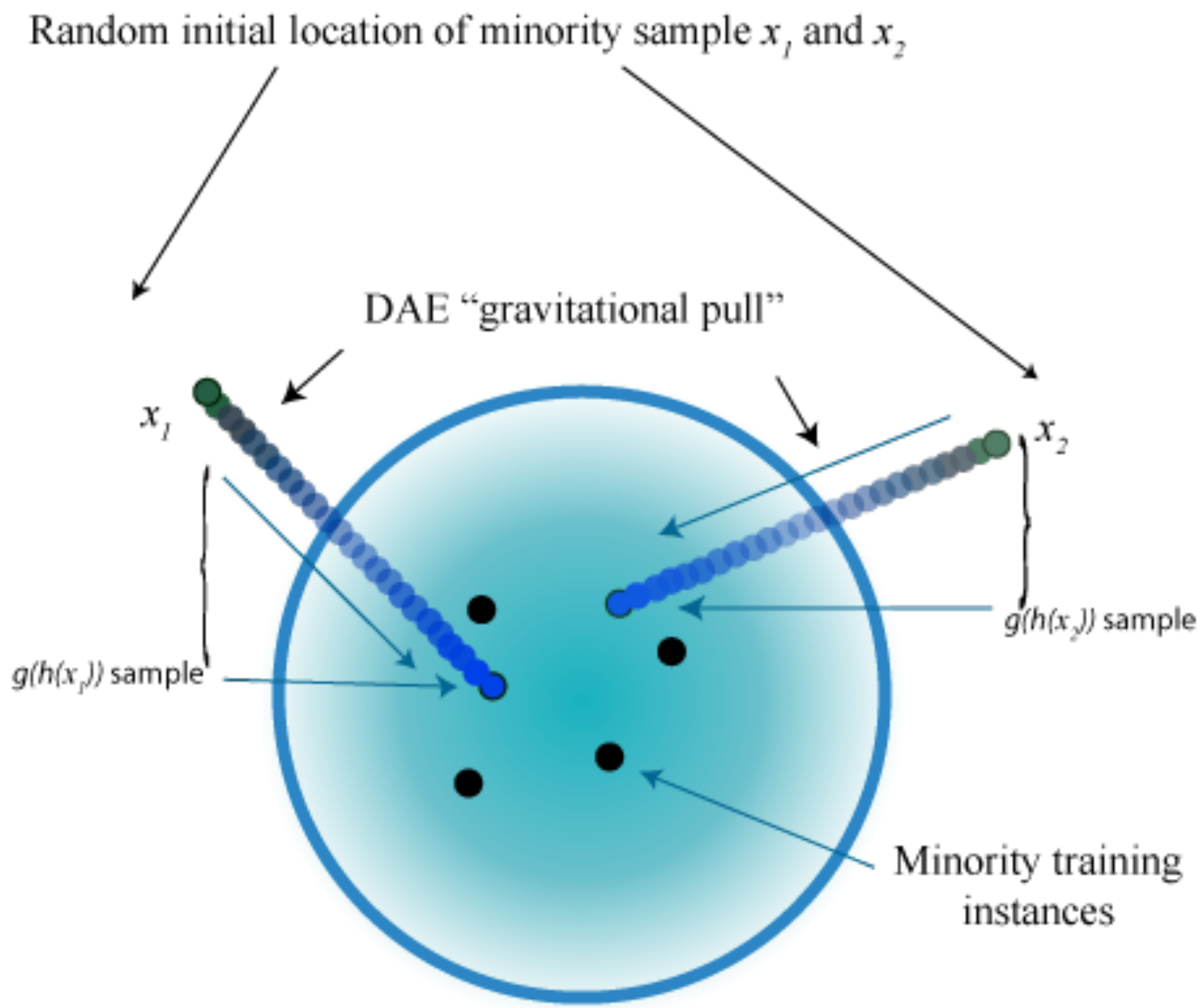
- Desire a system that:
  - takes advantage of RUS
  - generate diverse synthetic minority instance
    - influenced by minority training instance
      - perceived shape and density
      - expand convex-hull

AEGO System



DAE
U(-1,1)
D+synth
D̃+trn
D+trn
Dtrn
D+trn
DAE
D+`trn ={D+trn U D+synth}
D`trn
D-trn
RUS
D-`trn

AEGO Sampling



Pros and Cons

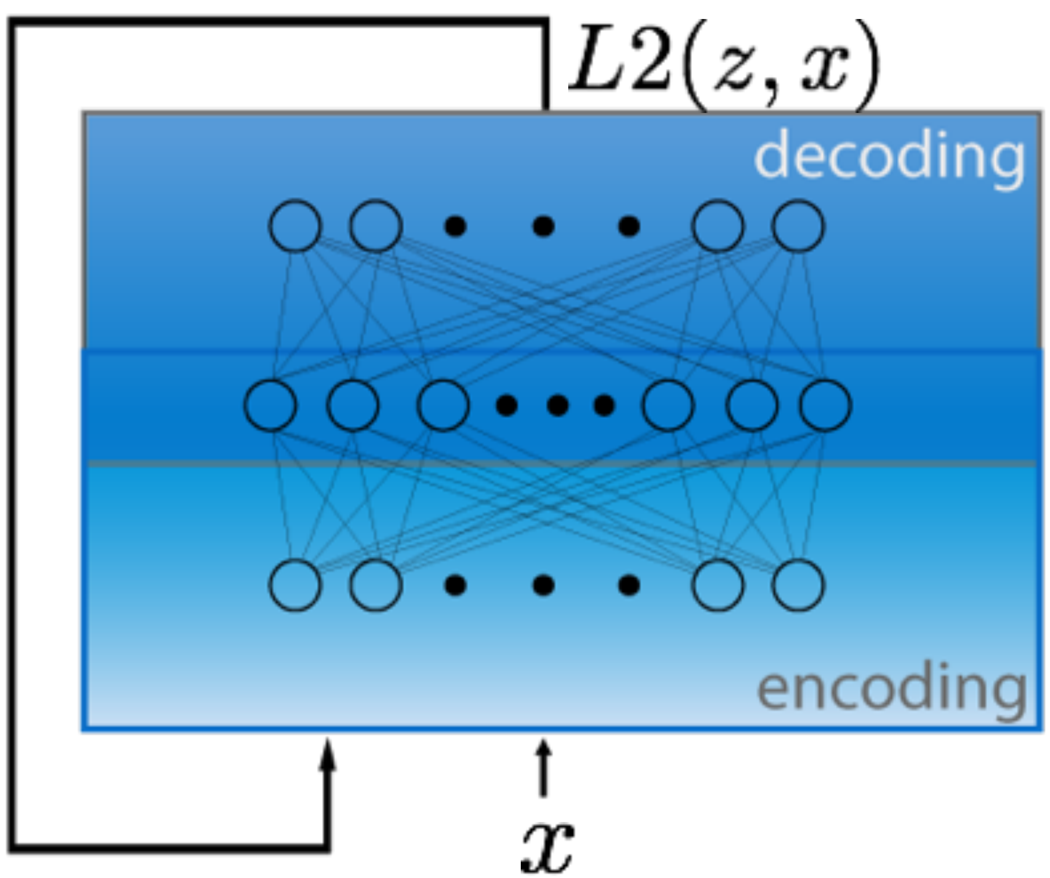
	Cost	ROS	RUS	SMOTE	AEGO
Strengths	- Reflect cost - Force min clsf	- Simple	- Simple - Dec bias of maj class	- Synth min - RUS	- Synth min - RUS - Outside convex hull - Density est.
Weaknesses	- Mod algo - Select costs - Overfit - Requires overlap - No new min info - Dep on min inst	- No new - Overfit - Requires overlap - No new min info - Dep on min inst	- Lost info - Variable - No new min info - Dep on min inst	- Convex hull - *k*NN - Requires overlap - Dep on min inst	- Parameters - Less obv synth set - Depends on min inst

Autoencoder

- Traditionally seen in OCC
  - Learns to “recognize” target class
  - Reject test instance with high RE
- Recent interest and advancements in DL community
  - Stacked AE

AE Components

- Traditional
  - compression to ensure learning
- Recent
  - over-complete with sparsity and denoising

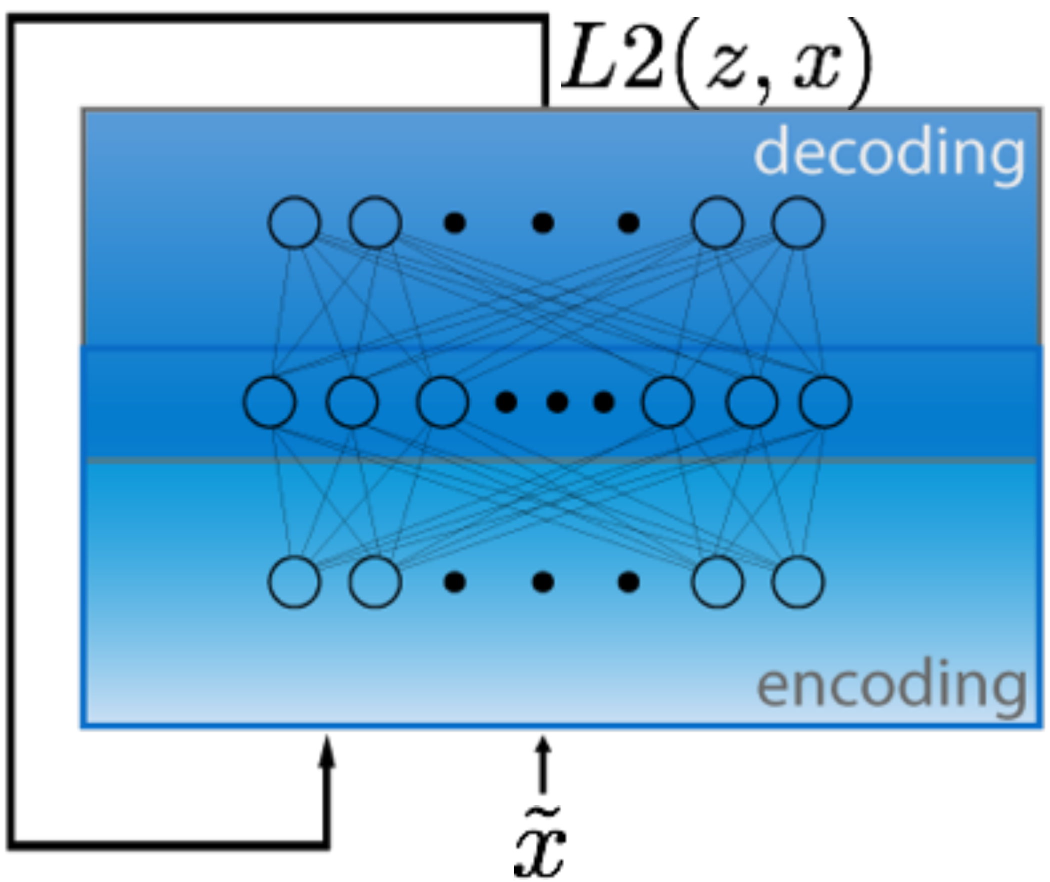


$$z = g_{\theta'}(y) = s'(\mathbf{W}'y + b')$$

$$h = f_{\theta}(x) = s(\mathbf{W}y + b)$$

Denoising Autoencoder

- Much discussion regarding learning
  - parametric vs. nonparametric?
- DAE shown to reproduce latent distribution

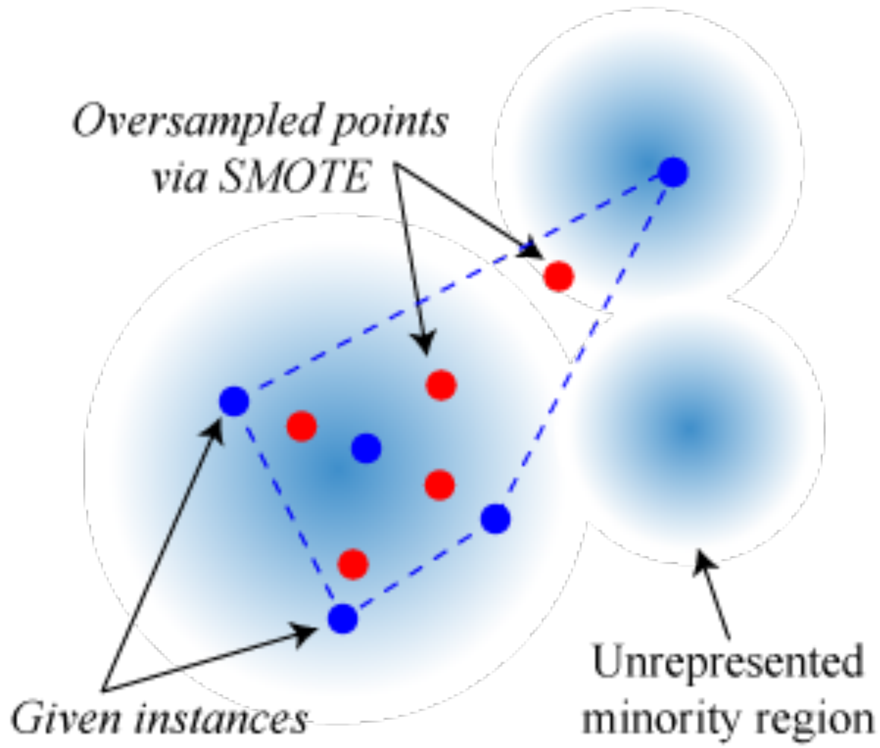


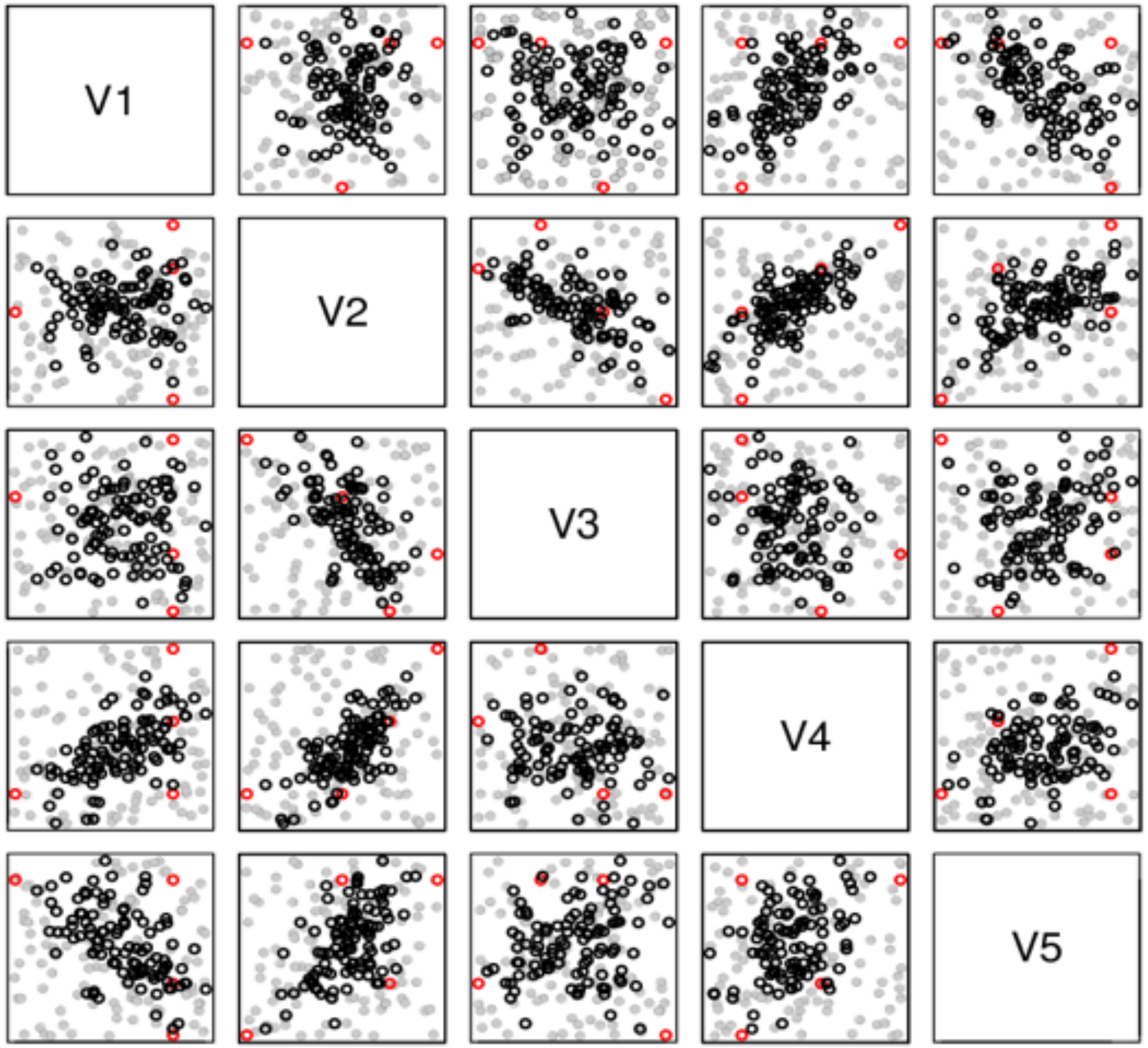
$$z = g_{\theta'}(y) = s'(\mathbf{W}'y + b')$$

$$h = f_\theta(\tilde{x}) = s(\mathbf{W}y + b)$$

$$\tilde{x}|x \sim \mathcal{N}(x, \sigma^2 \mathcal{I})$$

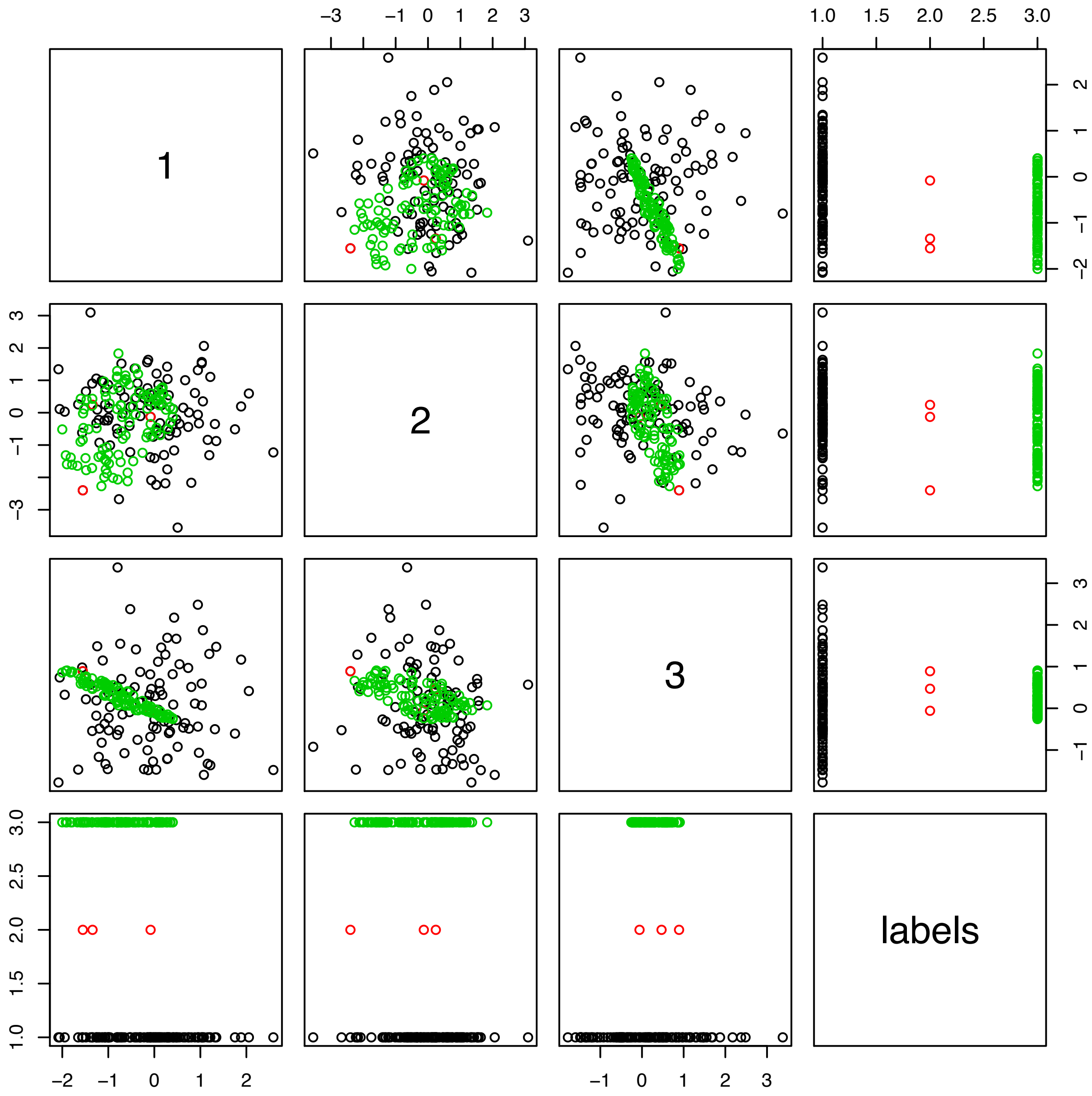
SMOTE Vs AEGO



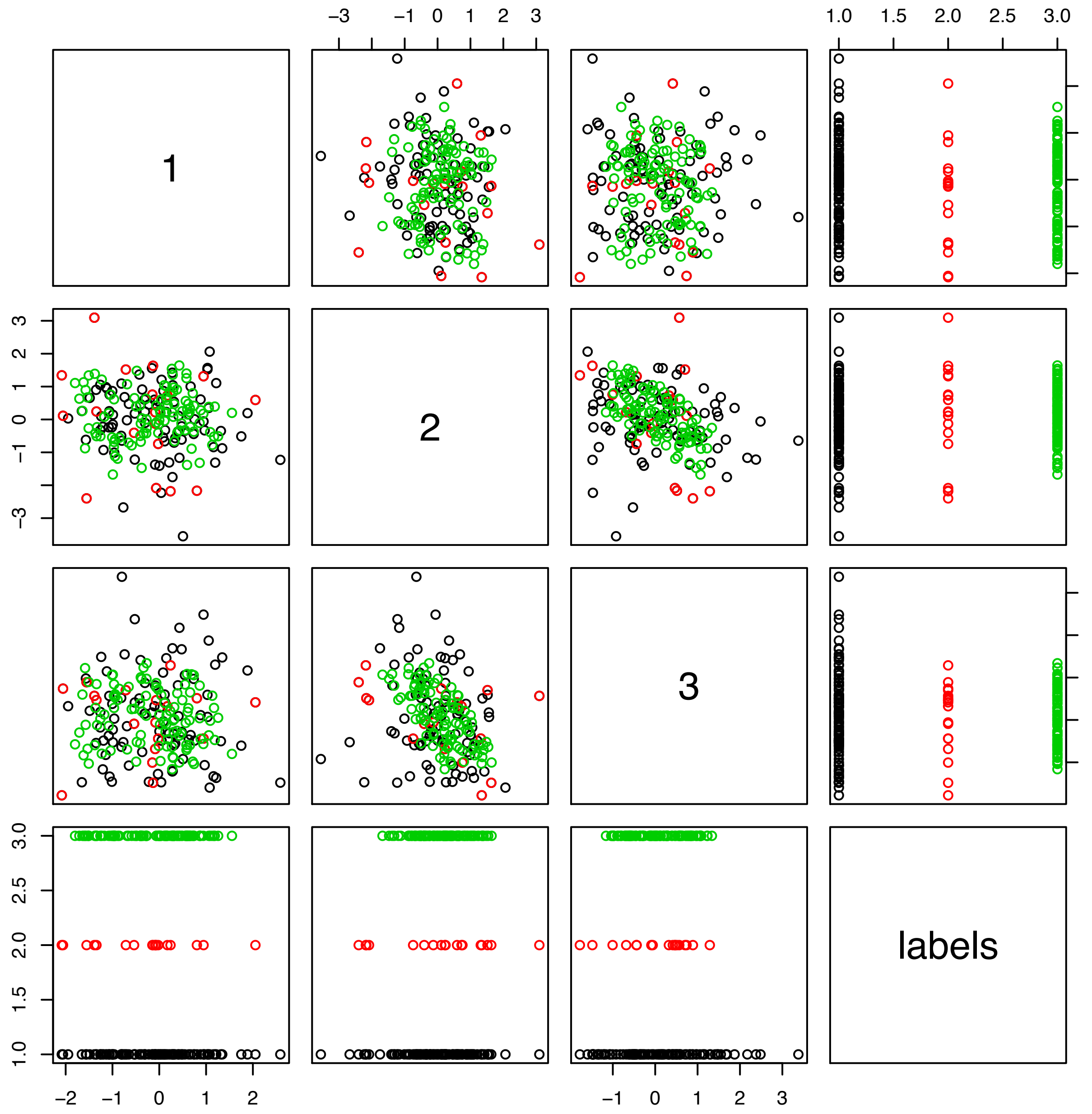


AEGO: Sample Size

Min trn = 3

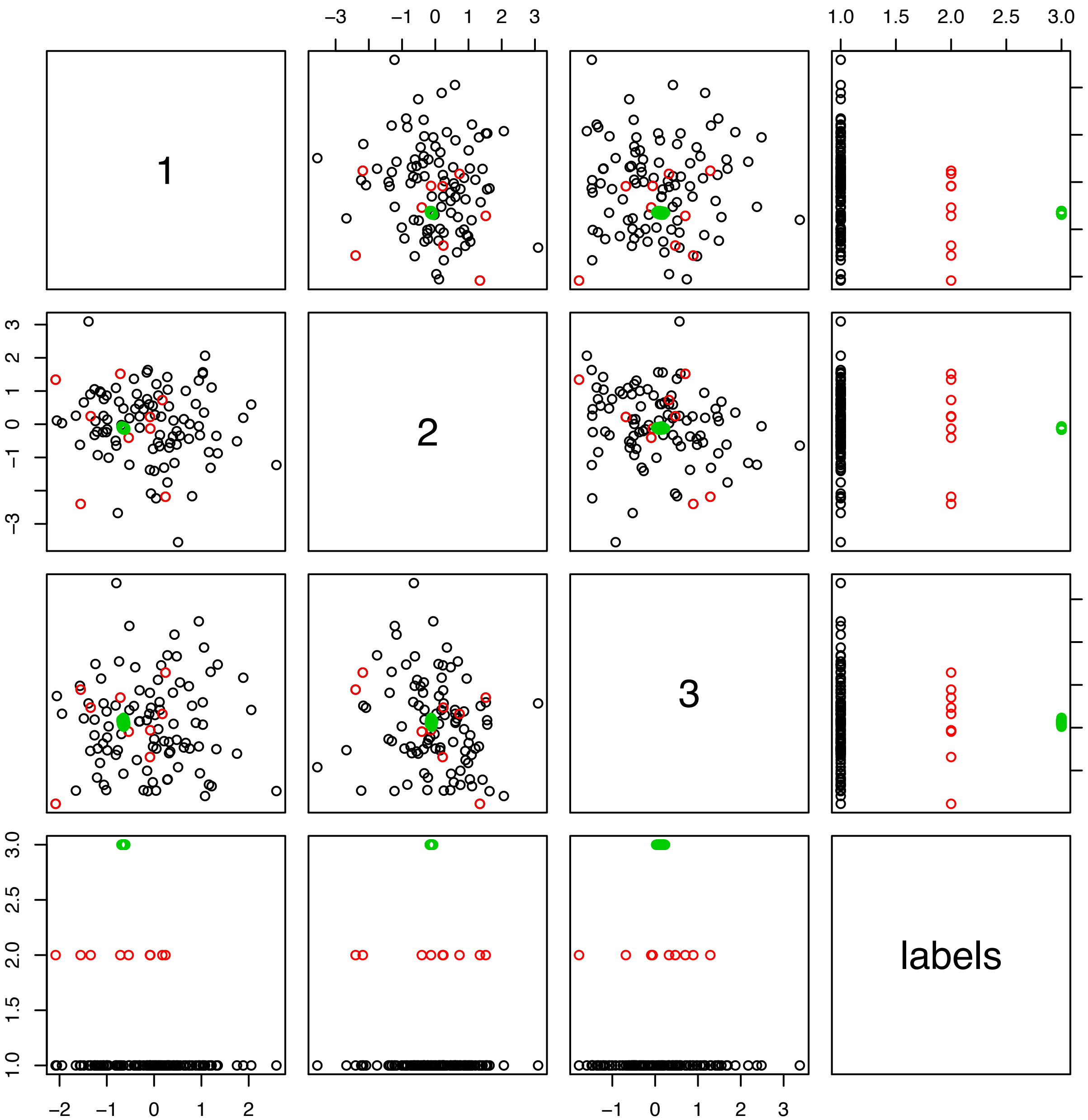


Min trn = 18

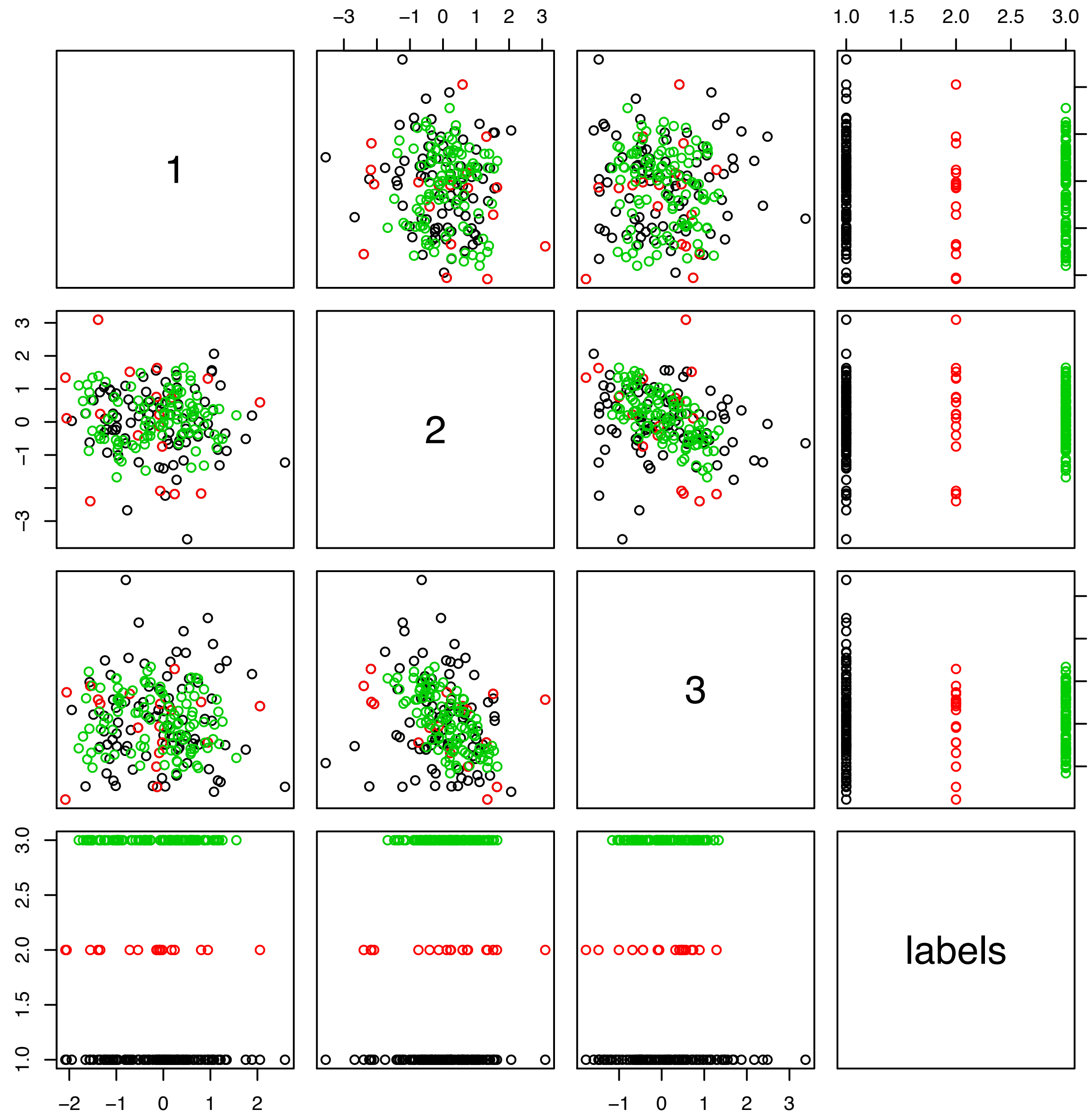


AEGO: DAE Vs AE

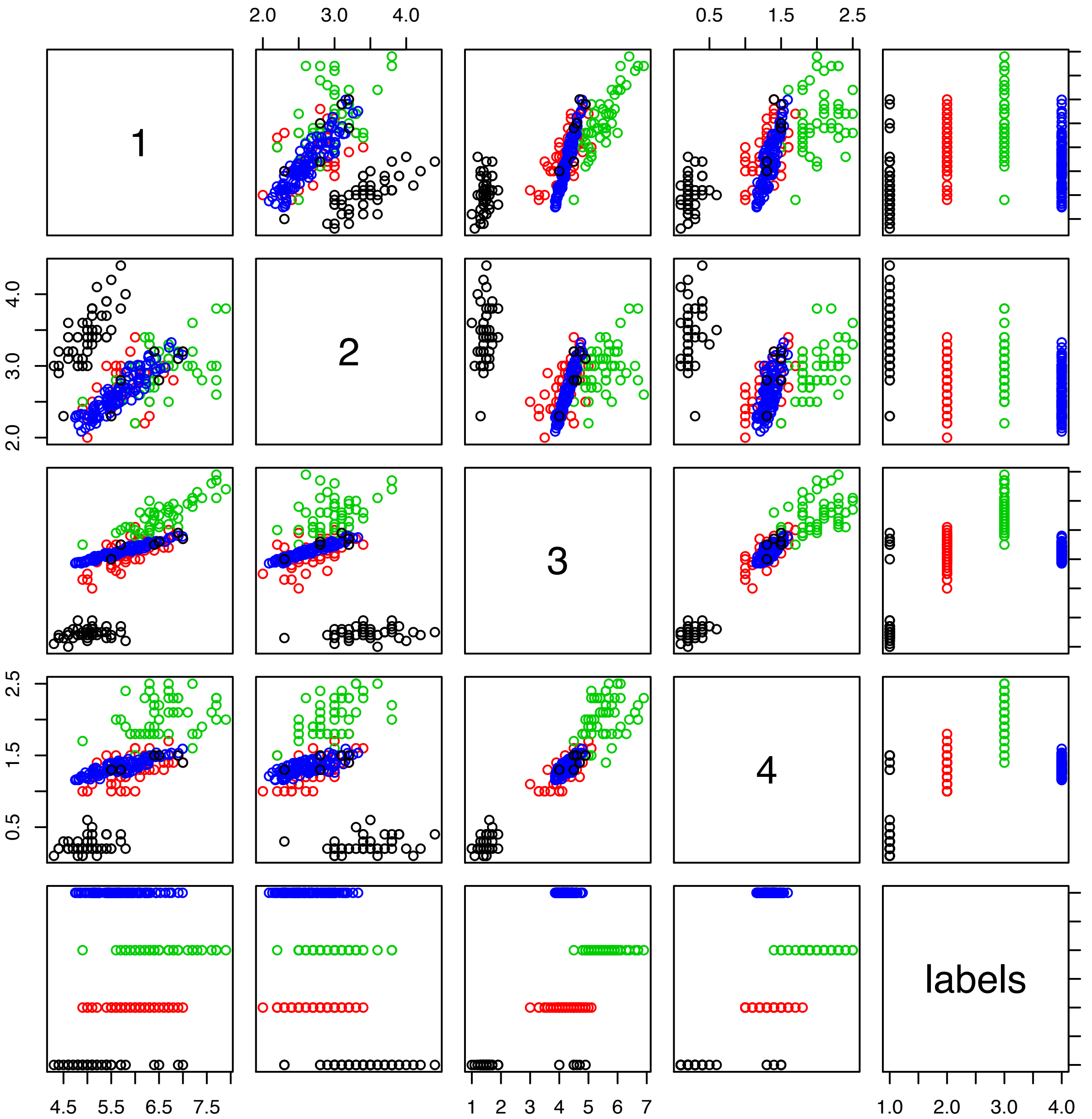
AE

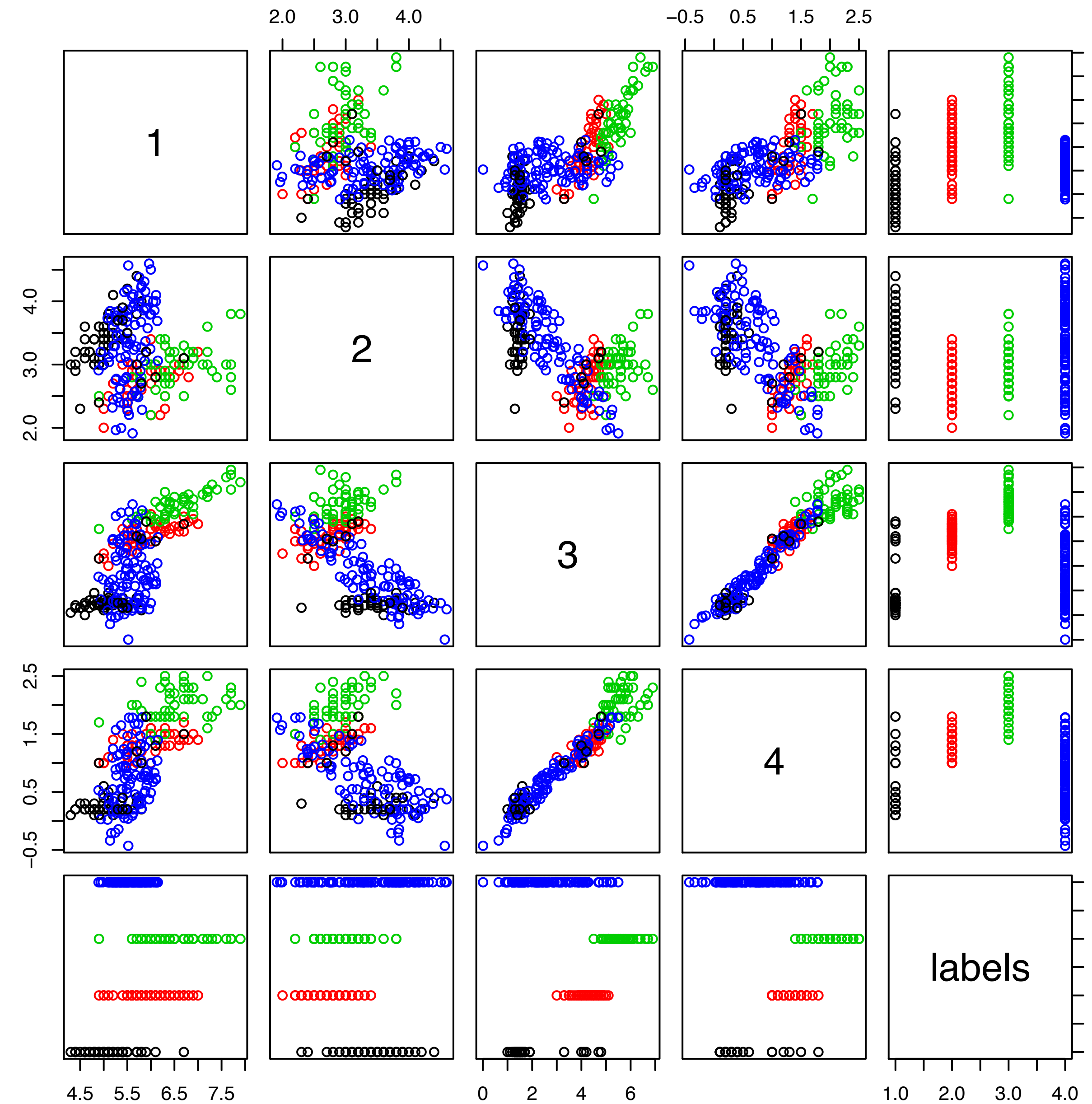


DAE



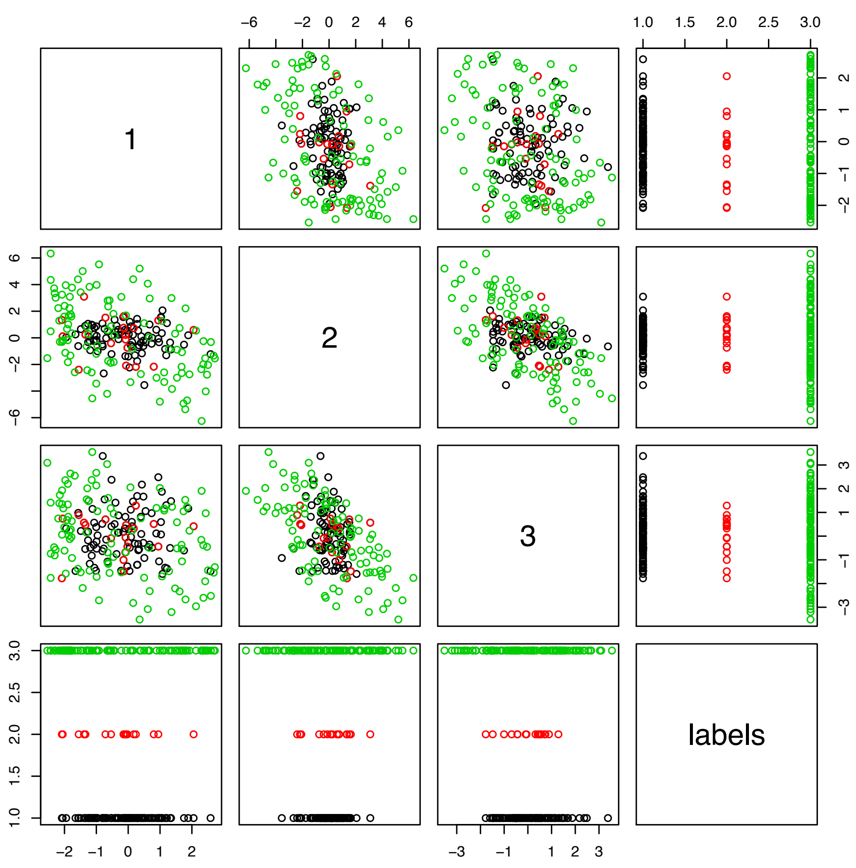
Iris Data

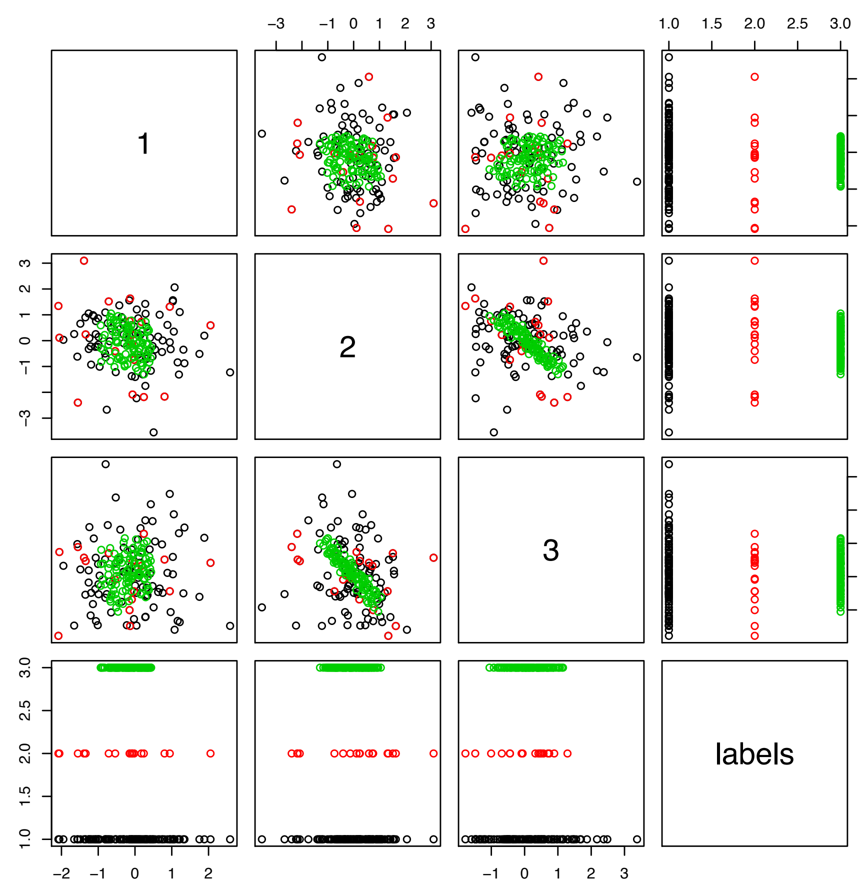


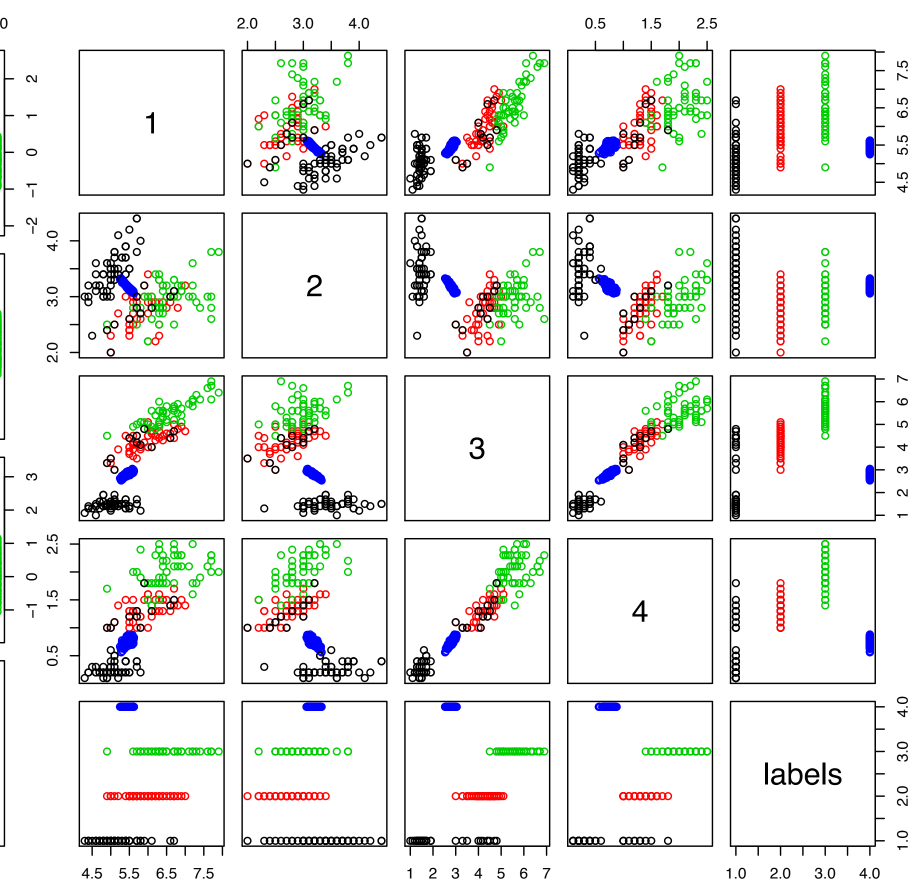


AEGO: Variability

Changing parameters







Discussion

- Generally good coverage
- Coverage depends on:
  - samples
  - parameters
- Nonetheless, robust to a wide range of parameters
- Ongoing research
  - training error and error propagation
  - stopping criteria

Experimental Method

- Modified UCI
  - *m* vs *n* classes with *m* underrepresented
  - 3 to 25 minority training instances
  - repeated 10 times
- Gamma-ray spectral data
  - +20,000 majority vs 49 minority
    - 5x2CV
- AUC evaluation
  - {BRUS, SMOTE, AEGO} + MLP

UCI Results

Compare by dataset

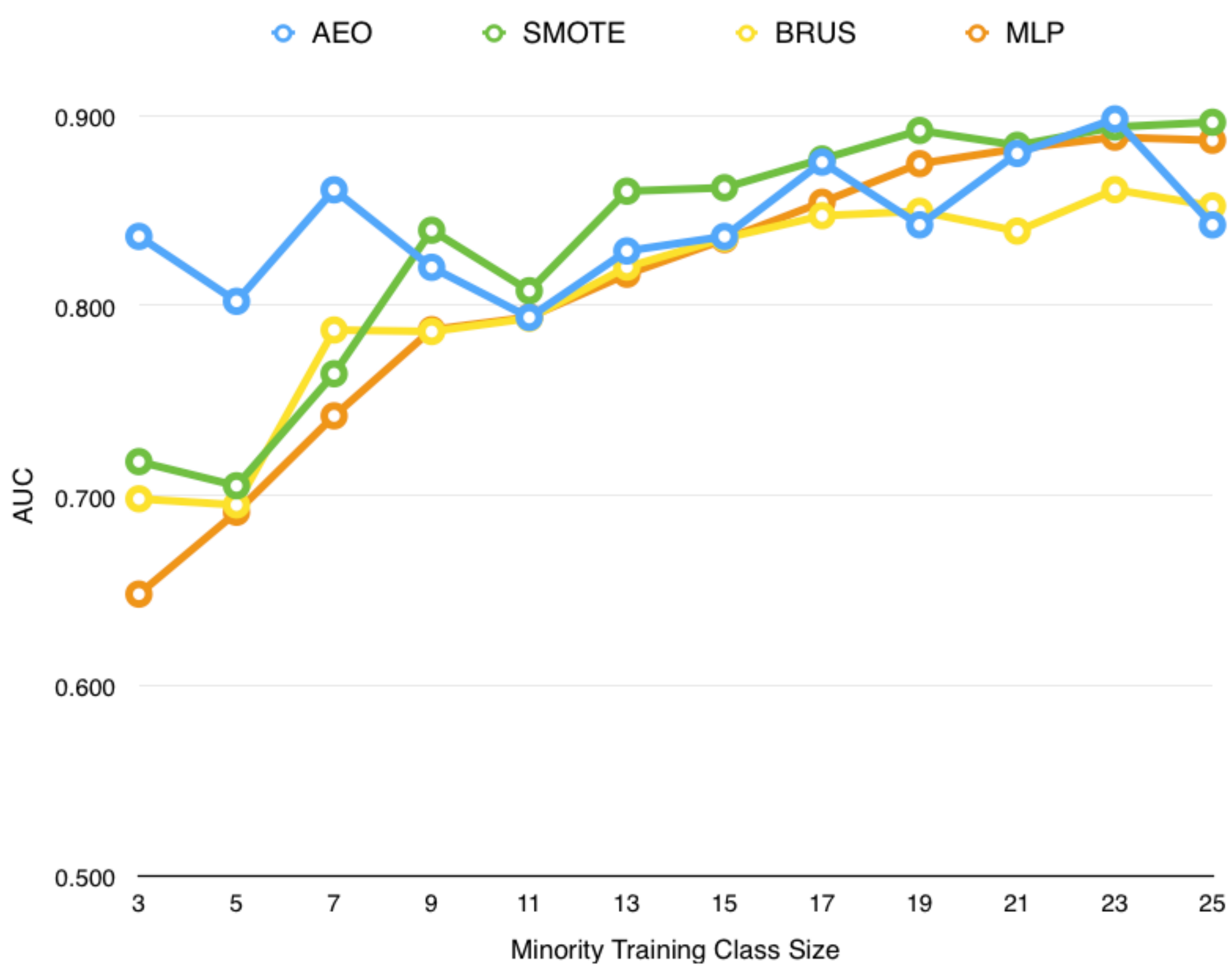
Dataset	AEGO	SMOTE	BRUS
Pen	3	*9*	0
Veh	*12*	0	0
Seg	*6*	5	1
Stat	*7*	4	0
Pima	4	*8*	0
Opt	3	*9*	0
Letter	*8*	4	0
Hab	6	*8*	0
Germ	3	*9*	0
Ecoli	*7*	5	0
Contra	*6*	4	2
Yeast	1	*12*	0
Wins	**6**	**6**	**0**

Compare by min trn size

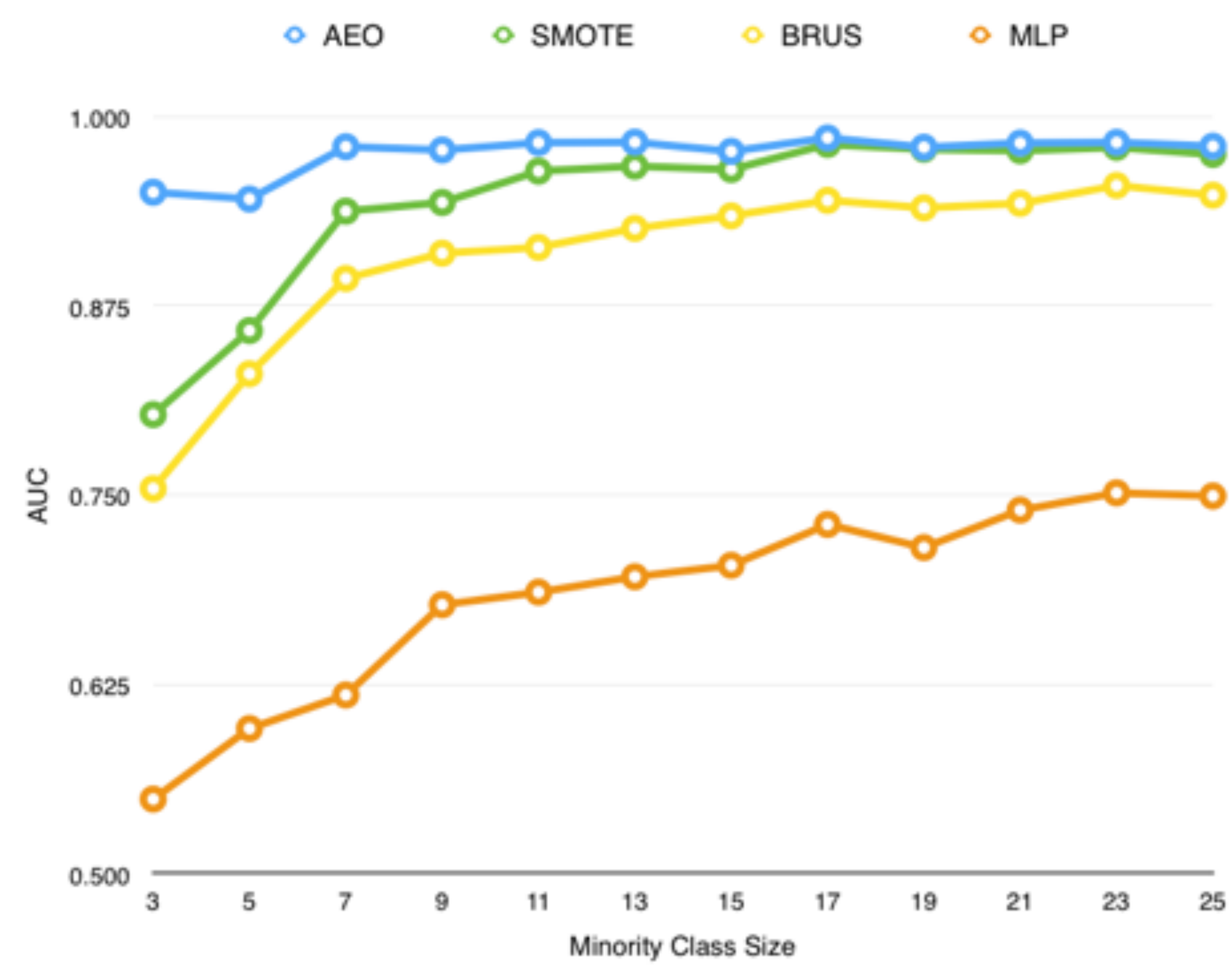
	AEGO	SMOTE	BRUS
3	*7*	4	0
5	*7*	4	0
7	*6*	5	0
9	*6*	5	0
10	*6*	4	0
13	4	*5*	1
15	*6*	5	0
17	4	*7*	0
19	*5*	4	1
21	5	*5*	0
23	*5*	4	0
25	3	*4*	0
Wins	**8**	**3**	**0**

UCI Results

Vehicle (van)



Pen digits (1,5,7)



Saanich Results

5x2CV

	MLP	BRUS	SMOTE	AEGO
1	0.893	0.766	0.757	0.869
2	0.592	0.679	0.745	0.827
3	0.676	0.720	0.719	0.881
4	0.803	0.728	0.844	0.888
5	0.910	0.729	0.861	0.909
6	0.842	0.798	0.945	0.957
7	0.626	0.798	0.856	0.927
8	0.937	0.707	0.755	0.874
9	0.599	0.746	0.746	0.833
10	0.980	0.834	0.952	0.975
Mean	0.786	0.742	0.818	0.894

Conclusion

- Novel form of Synthetic oversampling
  - denoising AutoEncoder-based Generative Oversampling (AEGO)
  - model minority class with DAE and “sample” the model
  - represents shape and density
  - expands beyond the convex-hull
- Regularly better on 124 UCI benchmark DS
  - notably strong with small minority training size
- Statistical better on Saanich domain

Thank You!

Colin Bellinger
Electrical Engineering and Computer Science
University of Ottawa
cbell052@uottawa.ca