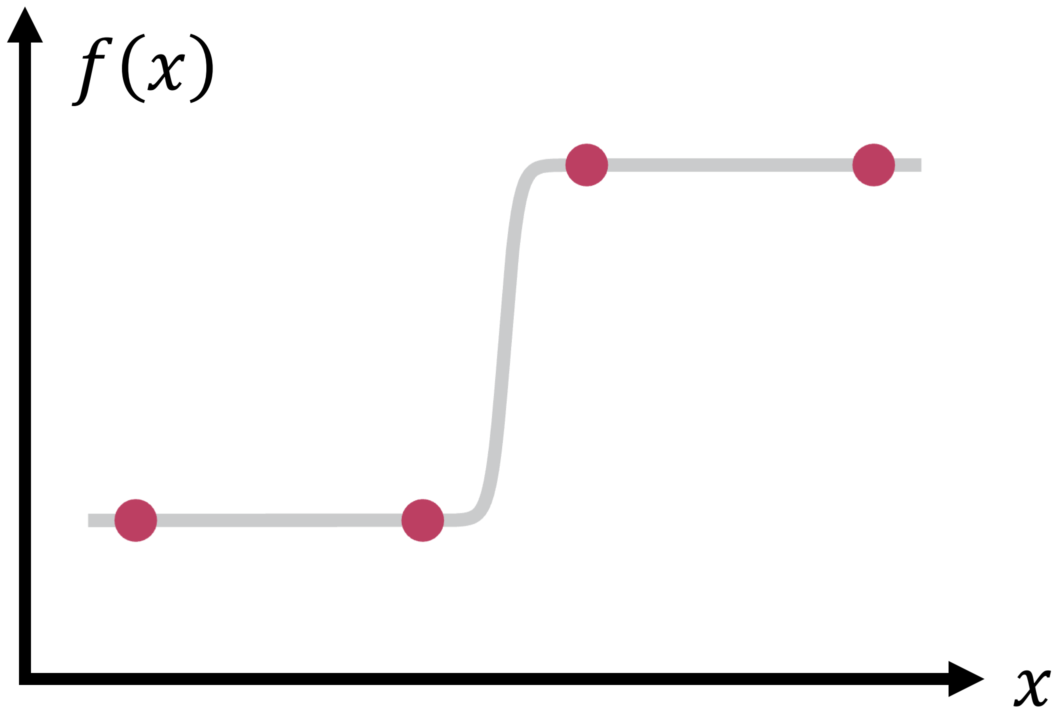
MAD Overview: Mixup for Augmenting Data in Myriad Scenarios

Madeline Navarro and Santiago Segarra
Department of Electrical and Computer Engineering, **Rice University**

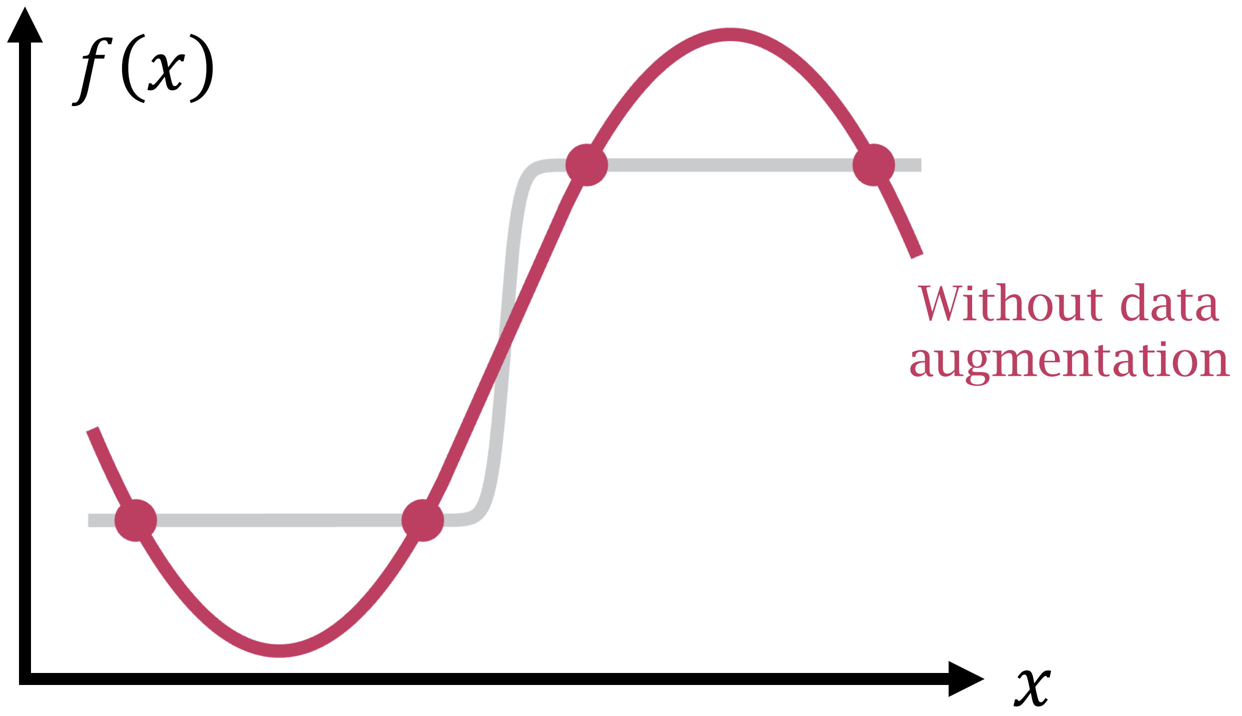
11 Jul 2024

Contact:
Email: nav@rice.edu

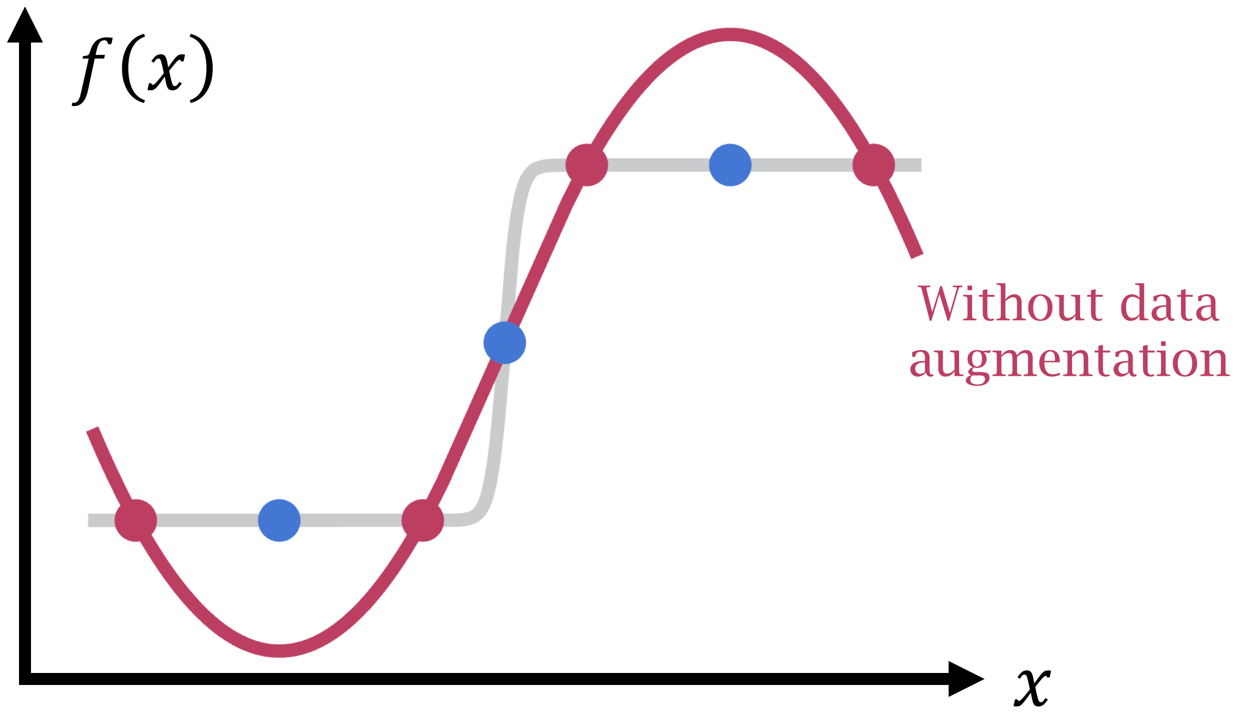
Data augmentation as implicit regularization



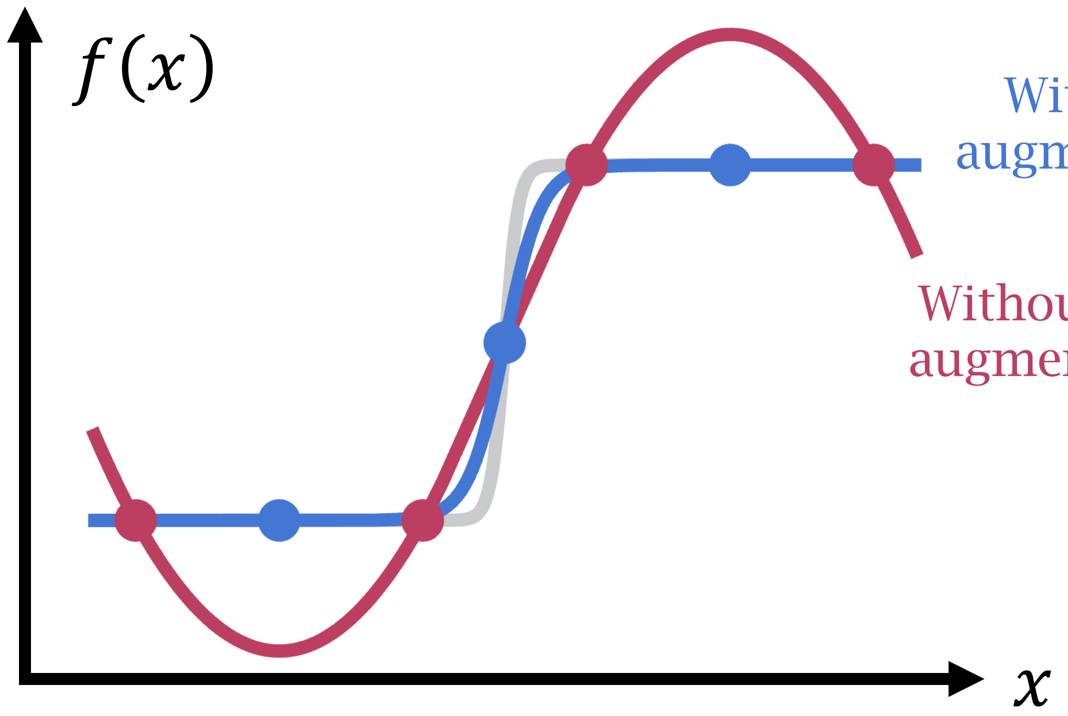
Data augmentation as implicit regularization



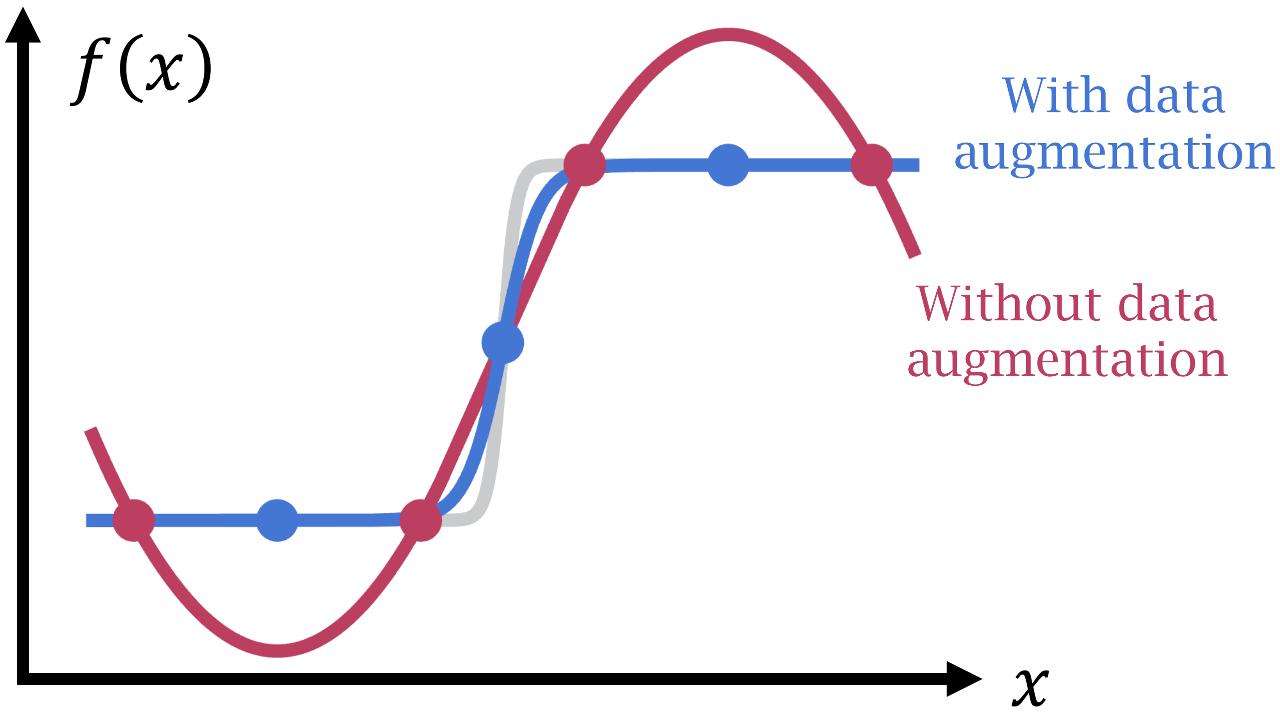
Data augmentation as implicit regularization



Data augmentation as implicit regularization



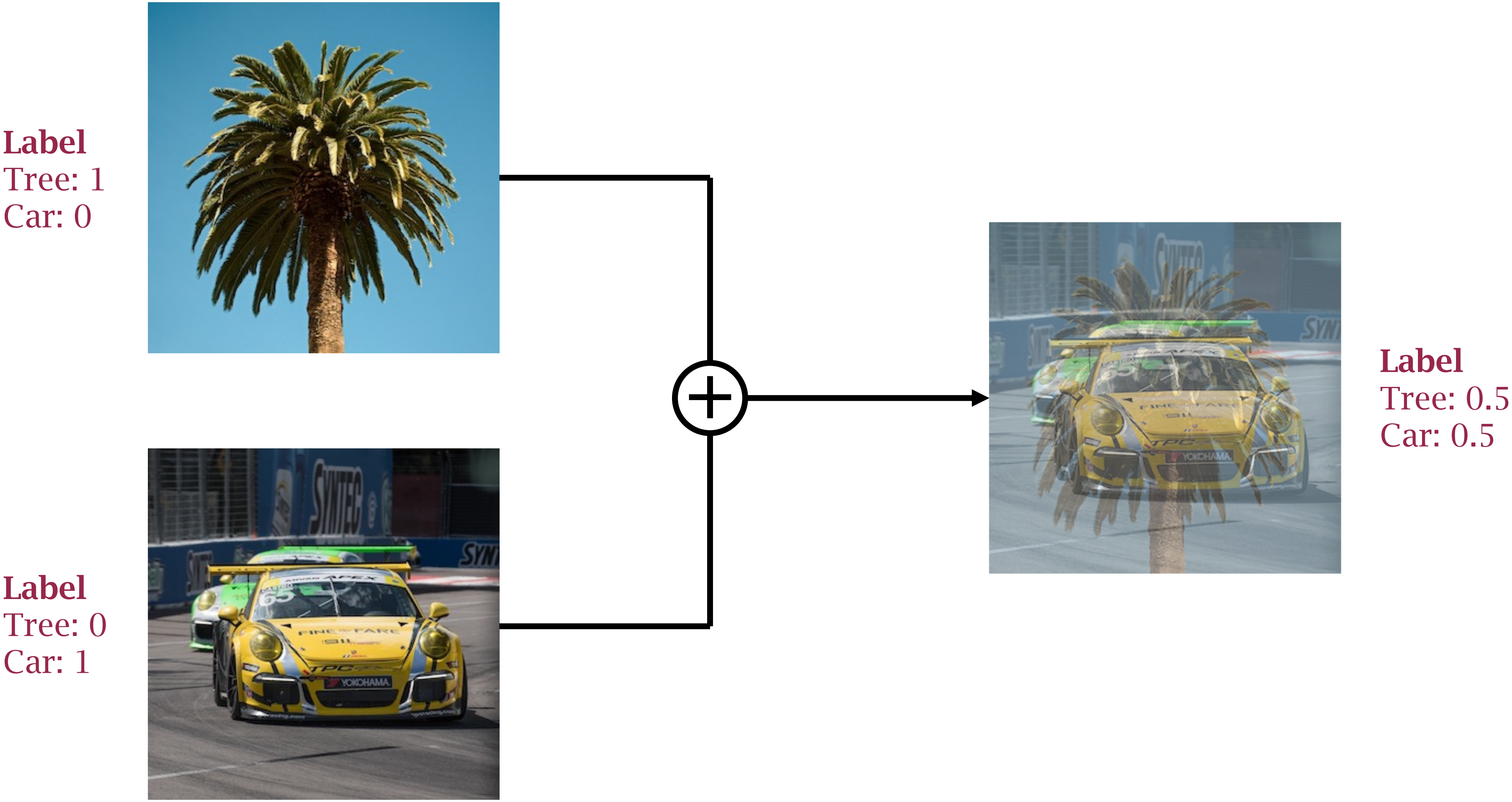
Data augmentation as implicit regularization



More training data

Avoid overfitting with intelligently generated data

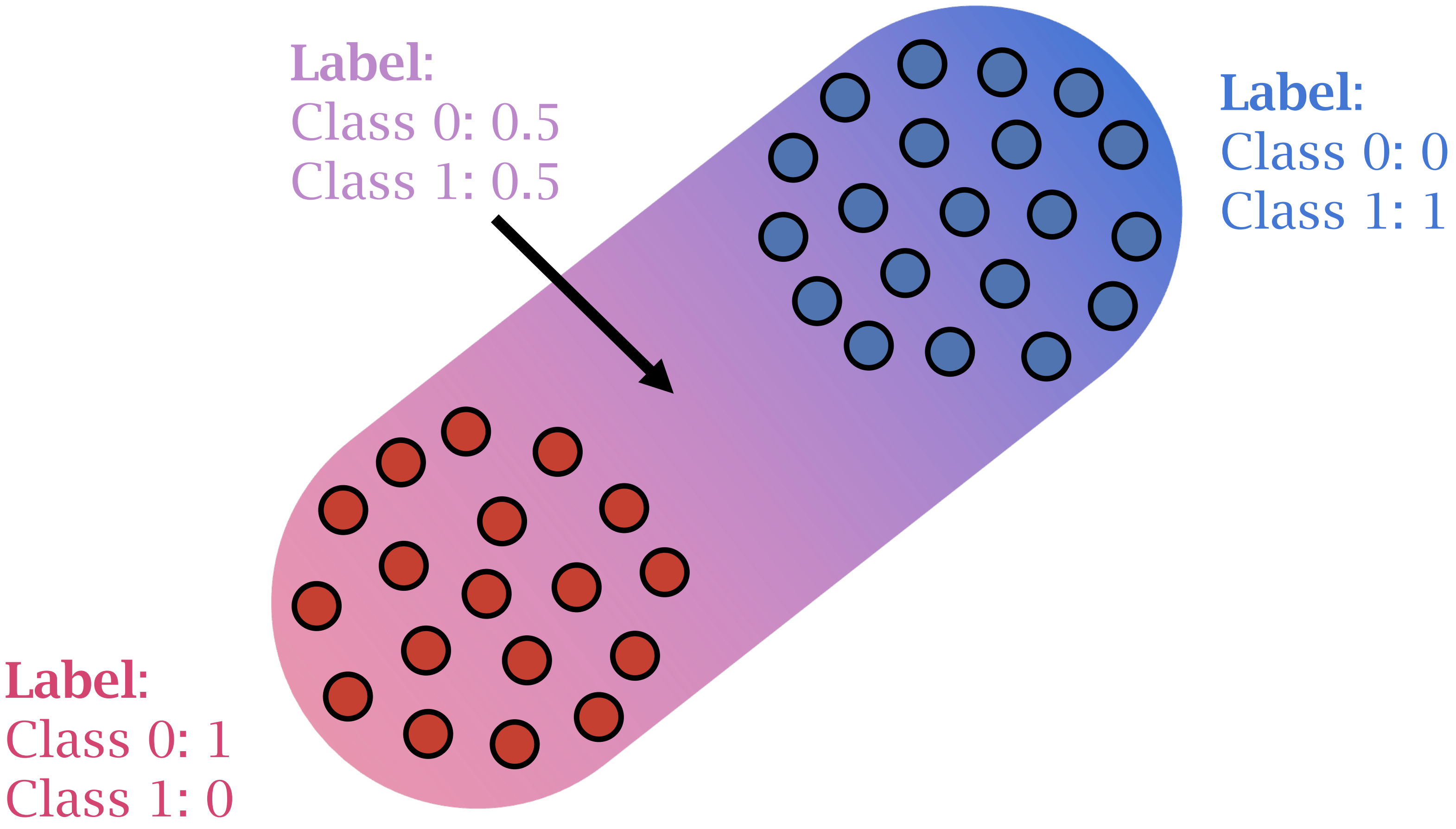
Mixup for data augmentation via linear combinations of data pairs



MAD directions

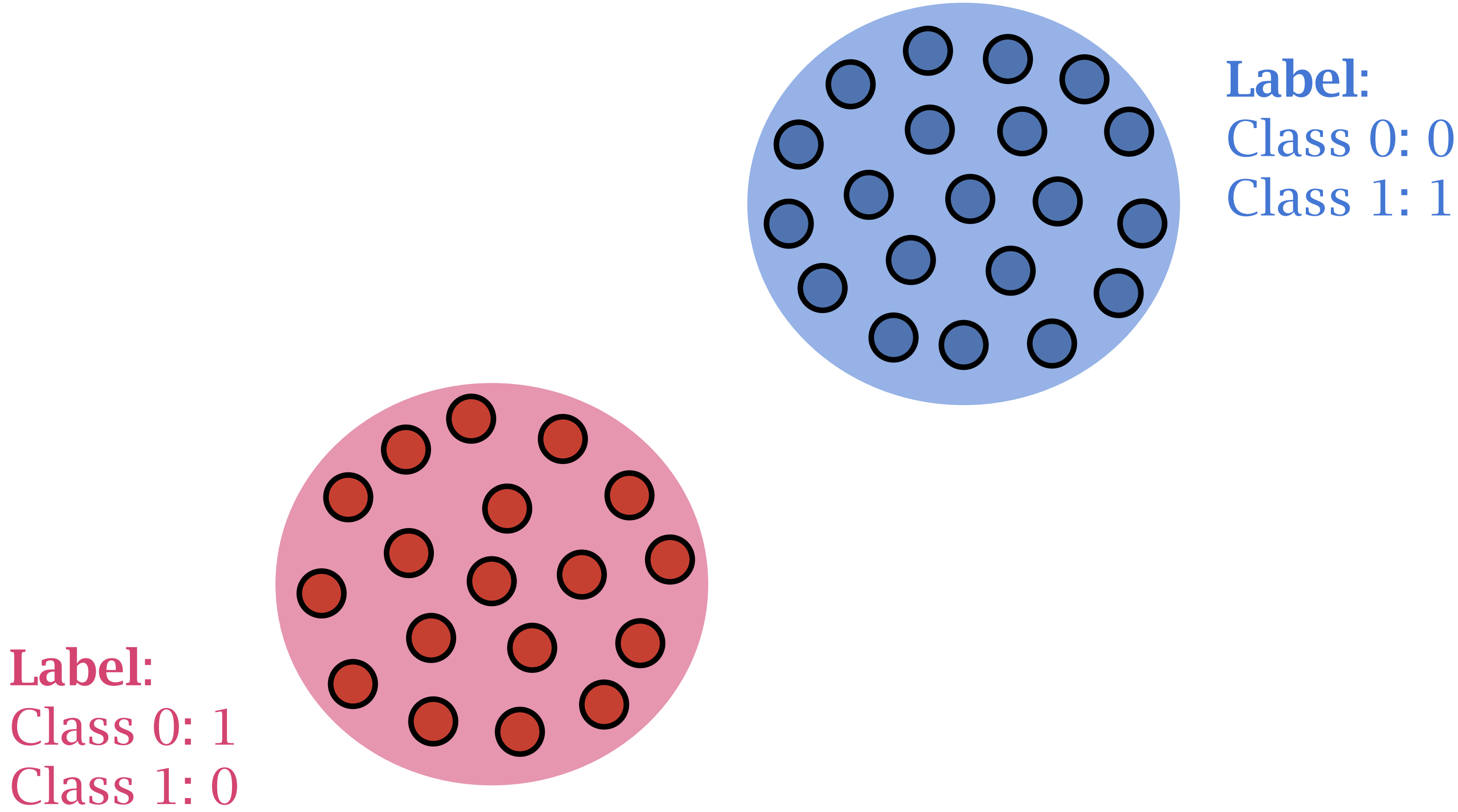
- Mixup method $\Longrightarrow$ Beyond pairwise linear mixup
- Mixup domain $\Longrightarrow$ Beyond Euclidean domains
- Mixup application $\Longrightarrow$ Beyond improving accuracy

When does pairwise linear mixup fail?



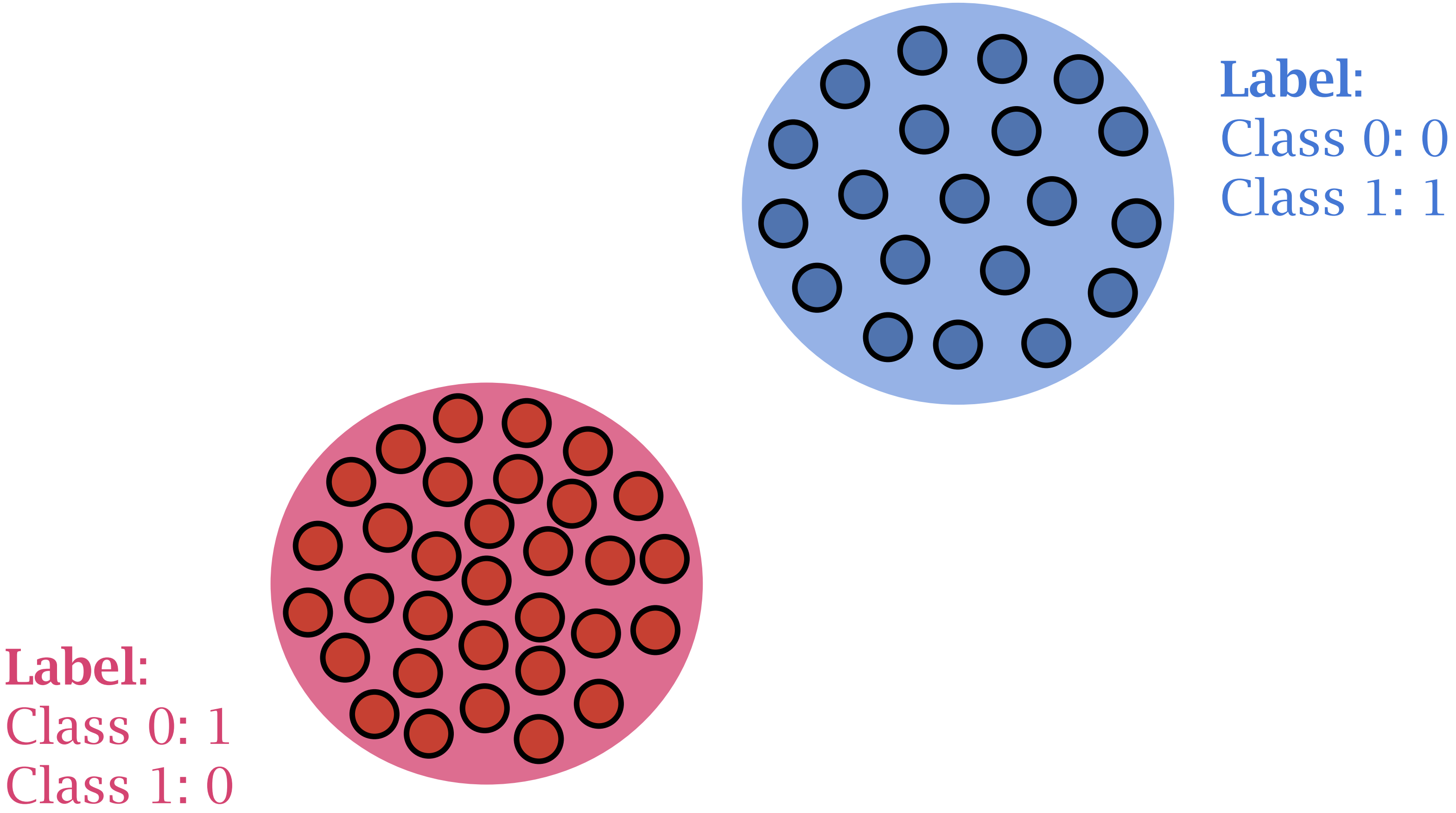
Linear mixup may add uncertainty in ways that are unhelpful

When does pairwise linear mixup fail?



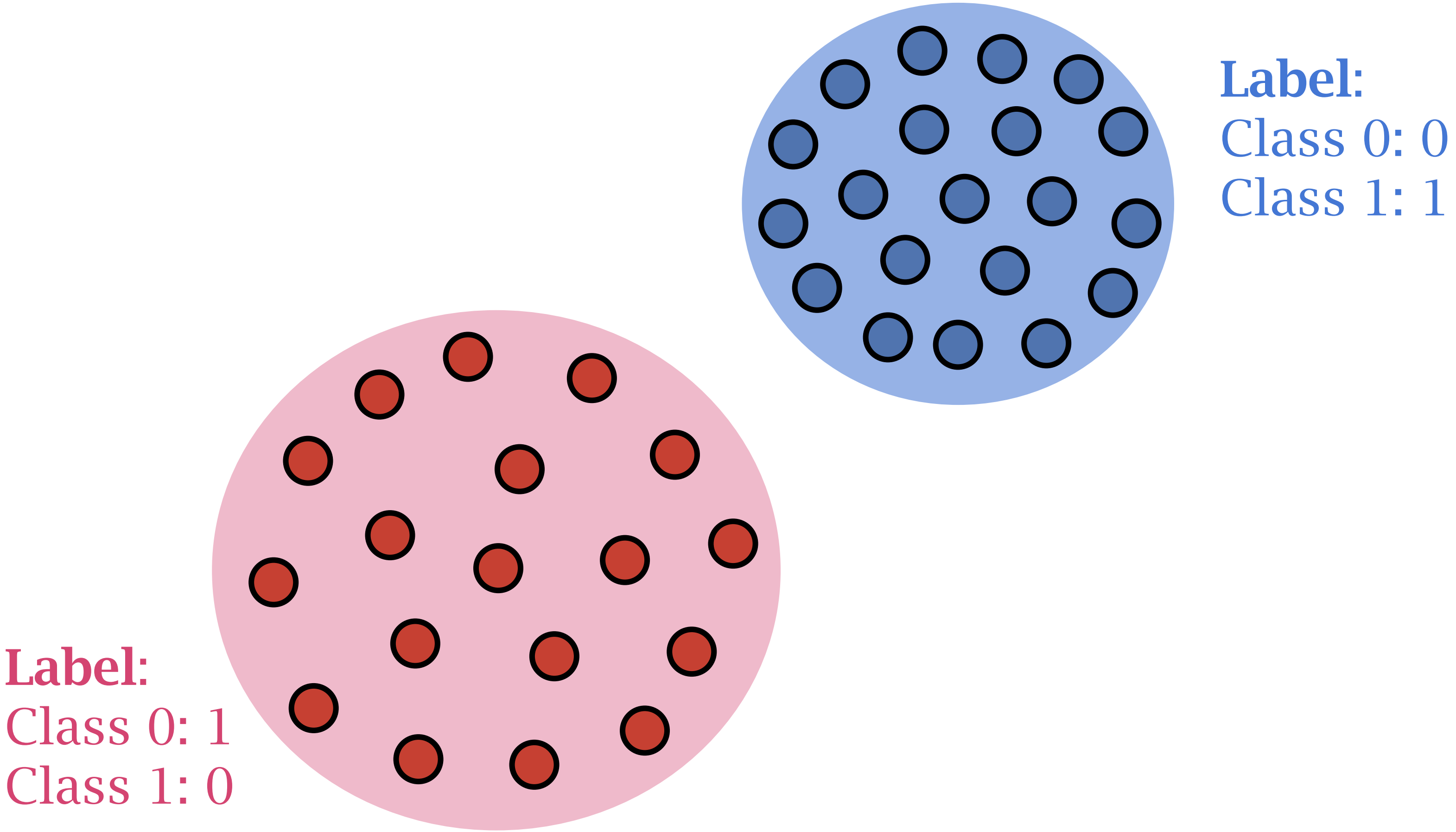
Pairwise mixup ignores most of the dataset when mixing two samples

When does pairwise linear mixup fail?



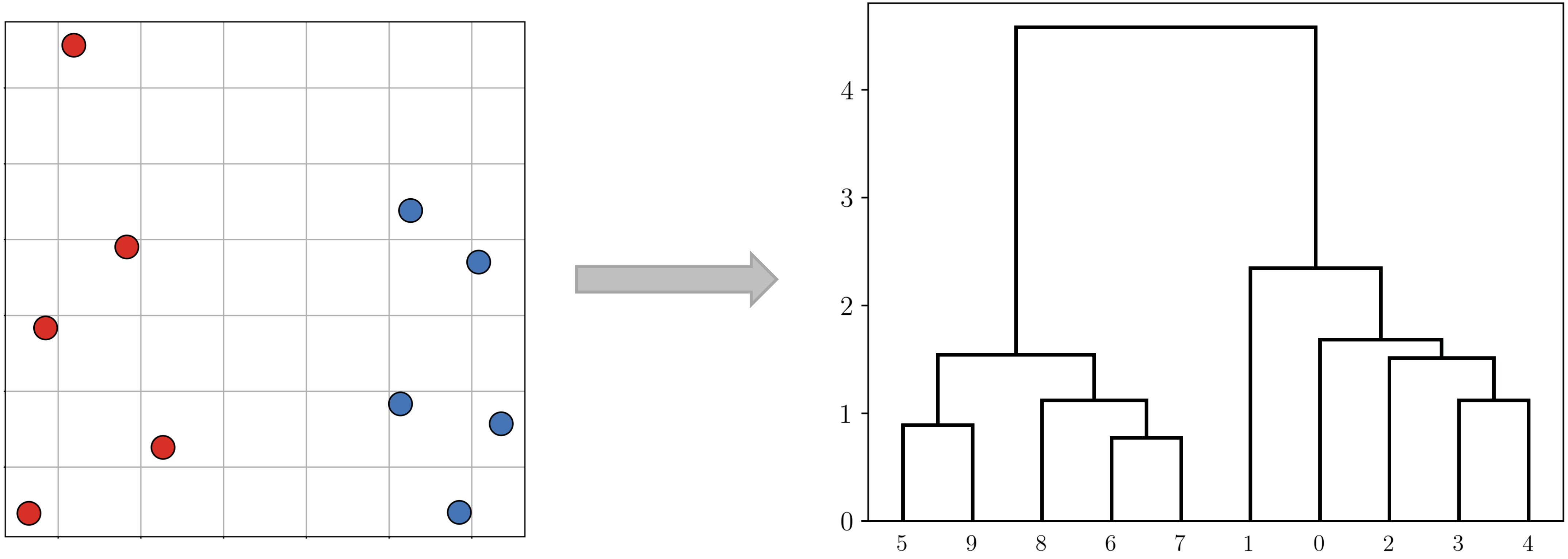
Pairwise mixup ignores most of the dataset when mixing two samples

When does pairwise linear mixup fail?



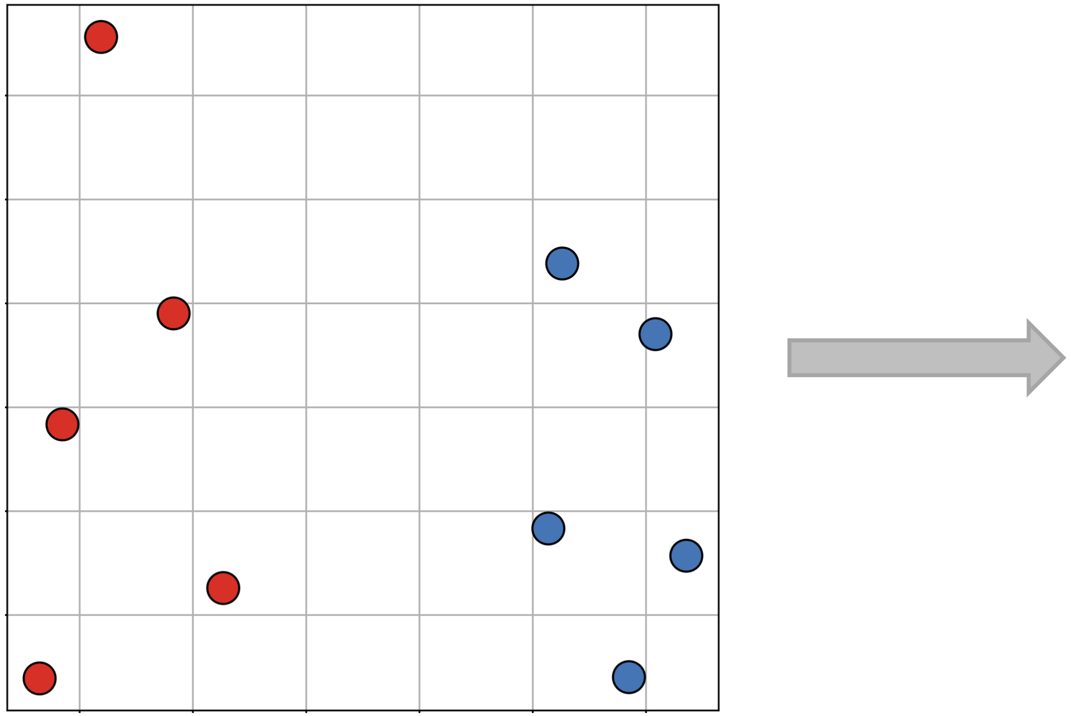
Pairwise mixup ignores most of the dataset when mixing two samples

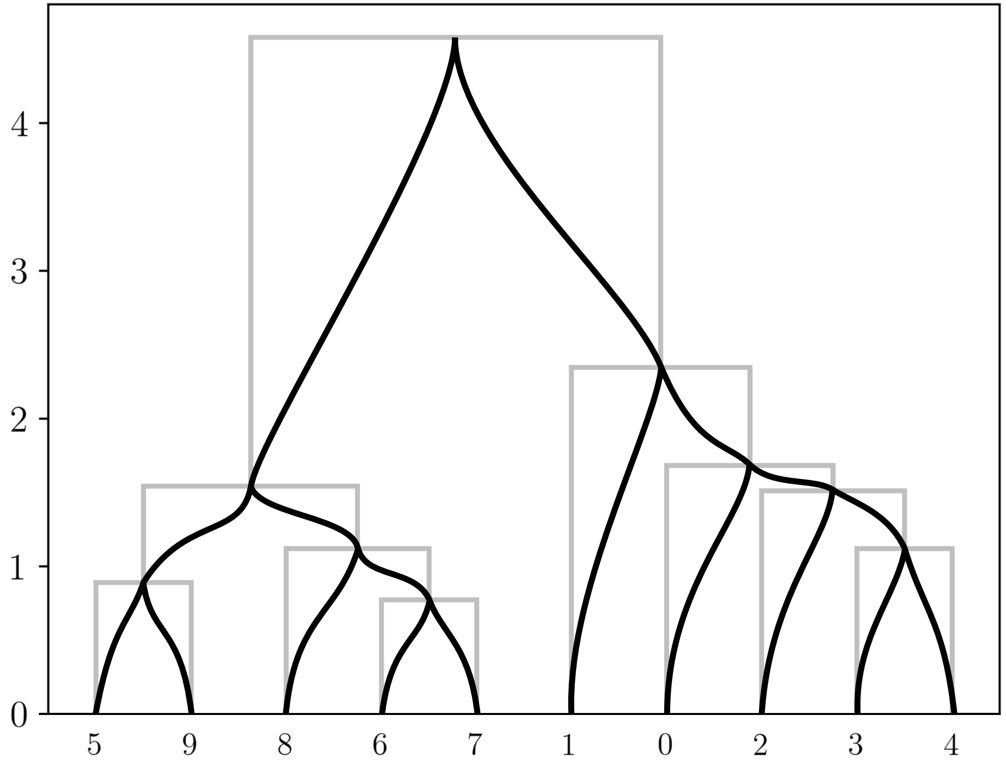
Clustering uses sample similarity to globally characterize datasets by their groups



Clustering methods such as hierarchical clustering use relationships among data to assign data to groups

Clustering uses sample similarity to globally characterize datasets by their groups





Clustering methods such as hierarchical clustering use relationships among data to assign data to groups

Convex clustering as a characterization of sample similarity

$$\{\hat{\mathbf{u}}_j(\lambda)\}_{j=1}^{T} = \underset{\mathbf{u}}{\operatorname{argmin}} \overbrace{\sum_{j=1}^{T} \|\mathbf{u}_j - \mathbf{x}_j\|_2^2}^{\text{Fidelity}} + \frac{\lambda}{1-\lambda} \overbrace{\sum_{i<j} w_{ij} \|\mathbf{u}_i - \mathbf{u}_j\|_1}^{\text{Fusion}}$$

Convex clustering tradeoff between fusing clusters and fitting to samples

K. Pelckmans, J. De Brabanter, J. A. Suykens, and B. De Moor, PASCAL 2005

Convex clustering as a characterization of sample similarity

$$\{\hat{\mathbf{u}}_j(\lambda)\}_{j=1}^{T} = \underset{\mathbf{u}}{\operatorname{argmin}} \overbrace{\sum_{j=1}^{T} \|\mathbf{u}_j - \mathbf{x}_j\|_2^2}^{\text{Fidelity}} + \frac{\lambda}{1-\lambda} \overbrace{\sum_{i<j} w_{ij} \|\mathbf{u}_i - \mathbf{u}_j\|_1}^{\text{Fusion}}$$

- $\mathbf{x}_j$: Each sample
- $\hat{\mathbf{u}}_j(\lambda)$: Cluster centroid for each sample at $\lambda \in [0,1]$
- λ: Fusion parameter

Convex clustering tradeoff between fusing clusters and fitting to samples

K. Pelckmans, J. De Brabanter, J. A. Suykens, and B. De Moor, PASCAL 2005

Convex clustering as a characterization of sample similarity

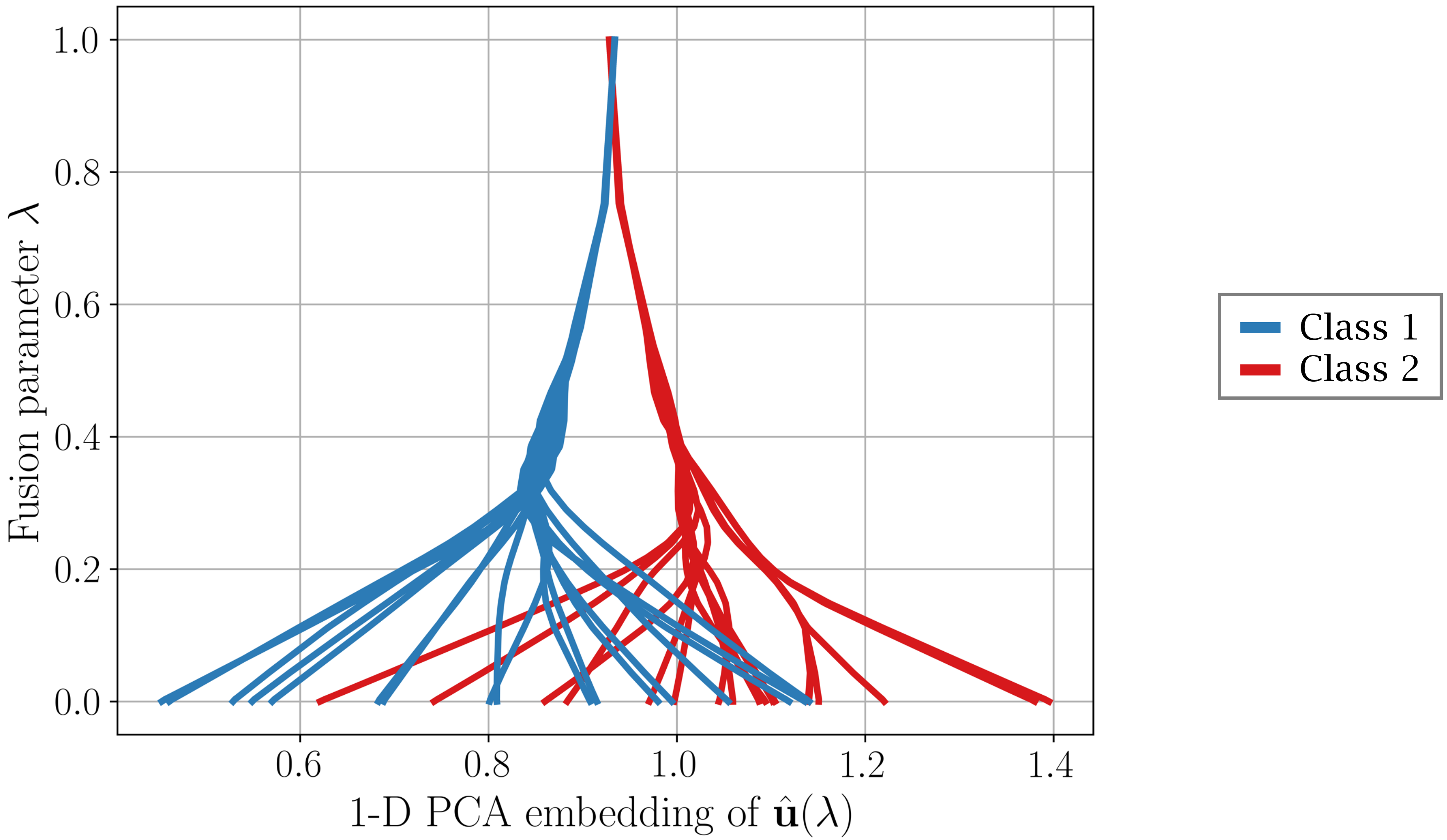
$$\{\hat{\mathbf{u}}_j(\lambda)\}_{j=1}^{T} = \underset{\mathbf{u}}{\operatorname{argmin}} \underbrace{\sum_{j=1}^{T} \|\mathbf{u}_j - \mathbf{x}_j\|_2^2}_{\text{Fidelity}} + \frac{\lambda}{1-\lambda} \underbrace{\sum_{i<j} w_{ij} \|\mathbf{u}_i - \mathbf{u}_j\|_1}_{\text{Fusion}}$$

- λ tunes between original dataset and total fusion (dataset mean)
  - $\lambda = 0$: T singleton clusters
  - $\lambda \in (0,1)$: Data samples begin to fuse into clusters
  - $\lambda = 1$: All samples in one cluster

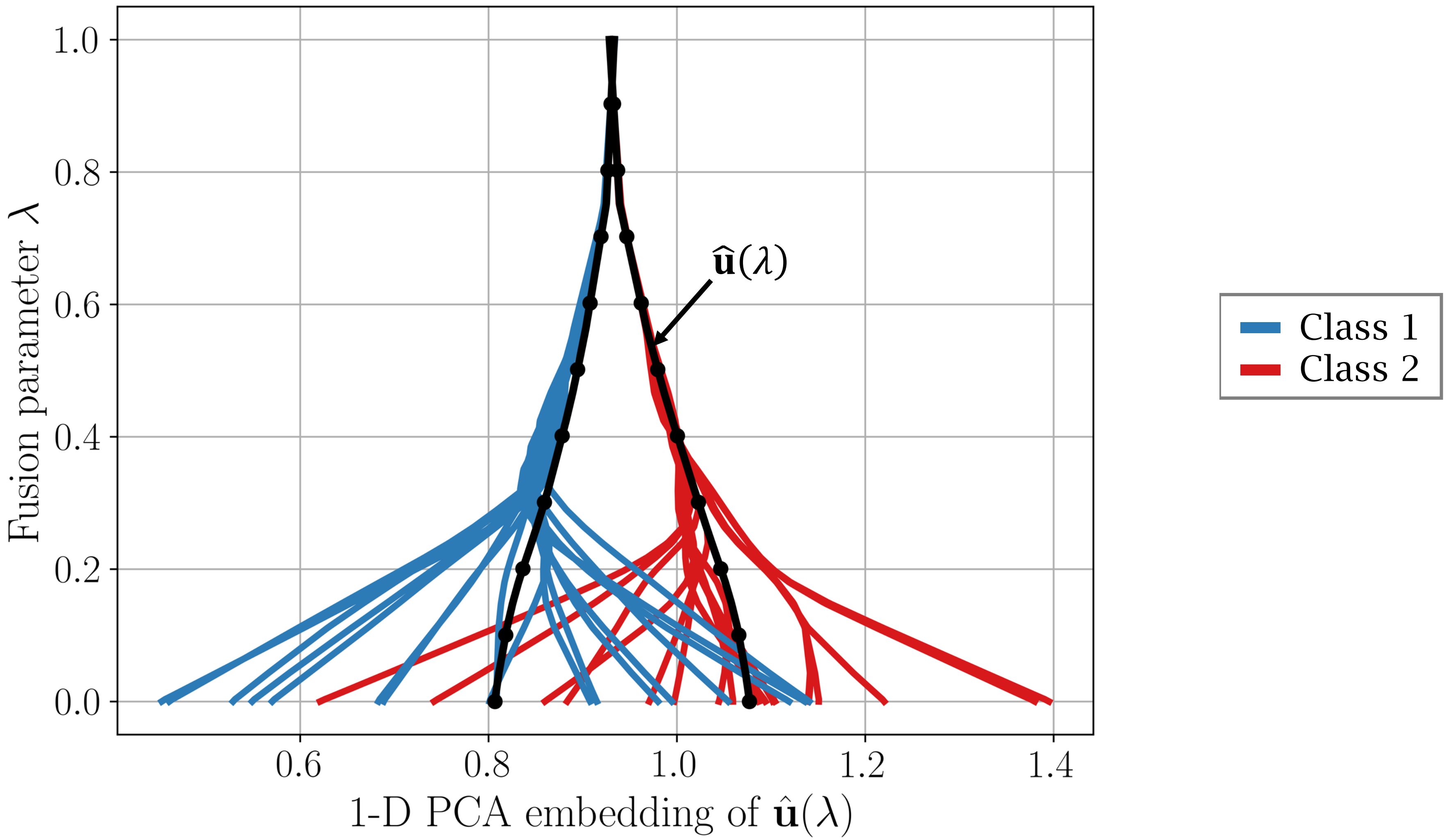
Convex clustering tradeoff between fusing clusters and fitting to samples

K. Pelckmans, J. De Brabanter, J. A. Suykens, and B. De Moor, PASCAL 2005

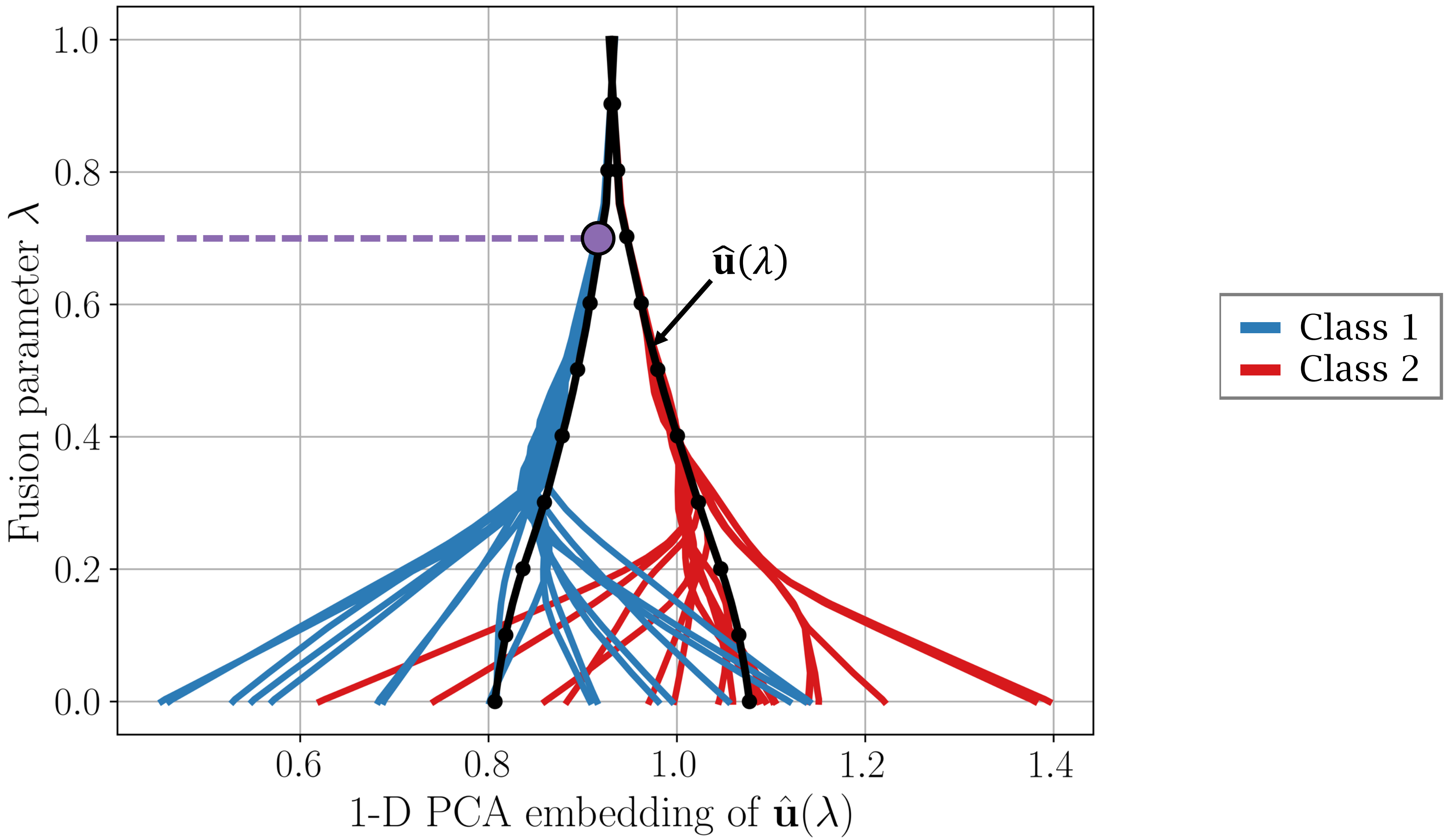
Convex clustering as a characterization of sample similarity



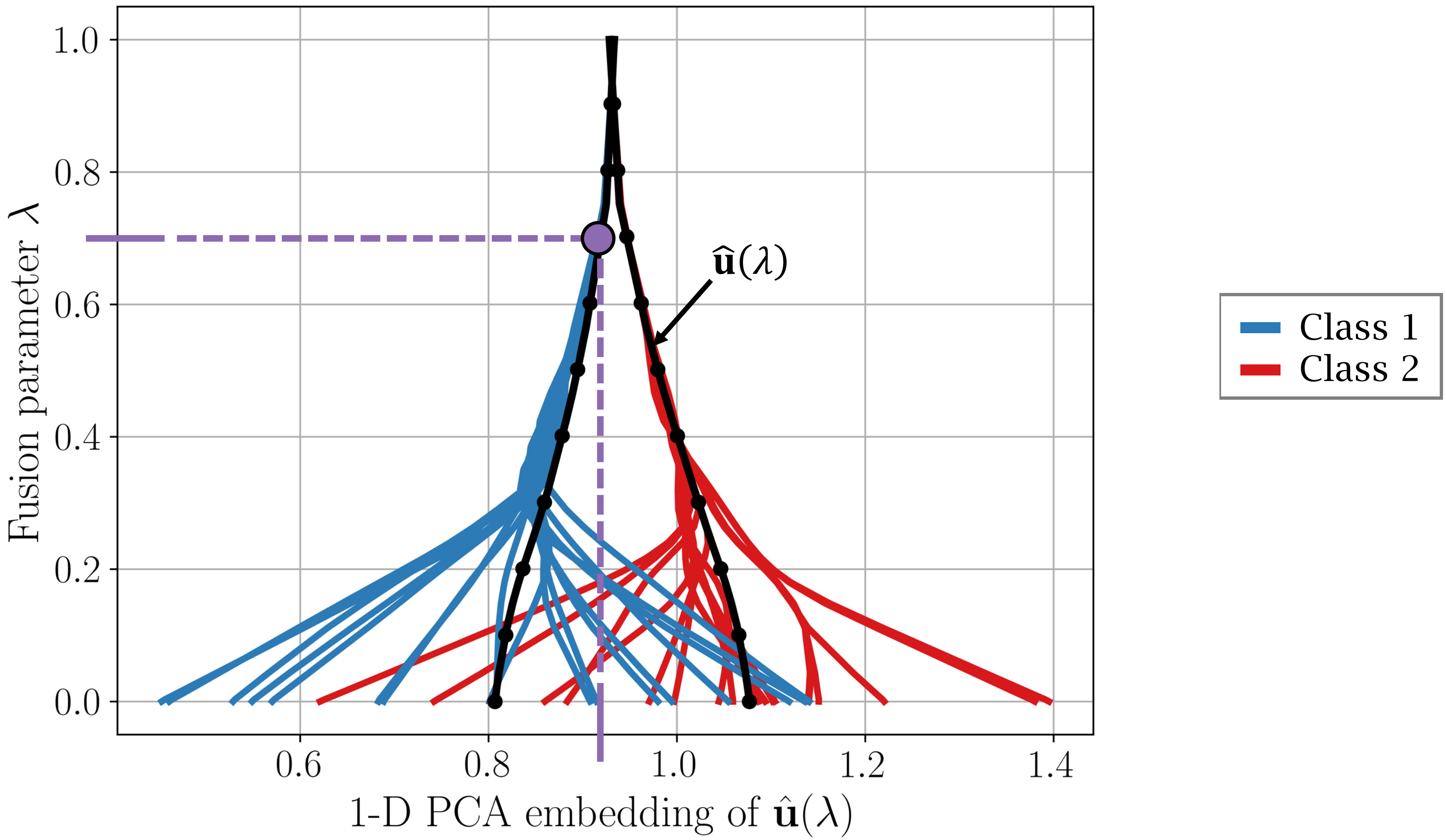
Convex clustering as a characterization of sample similarity



Convex clustering as a characterization of sample similarity

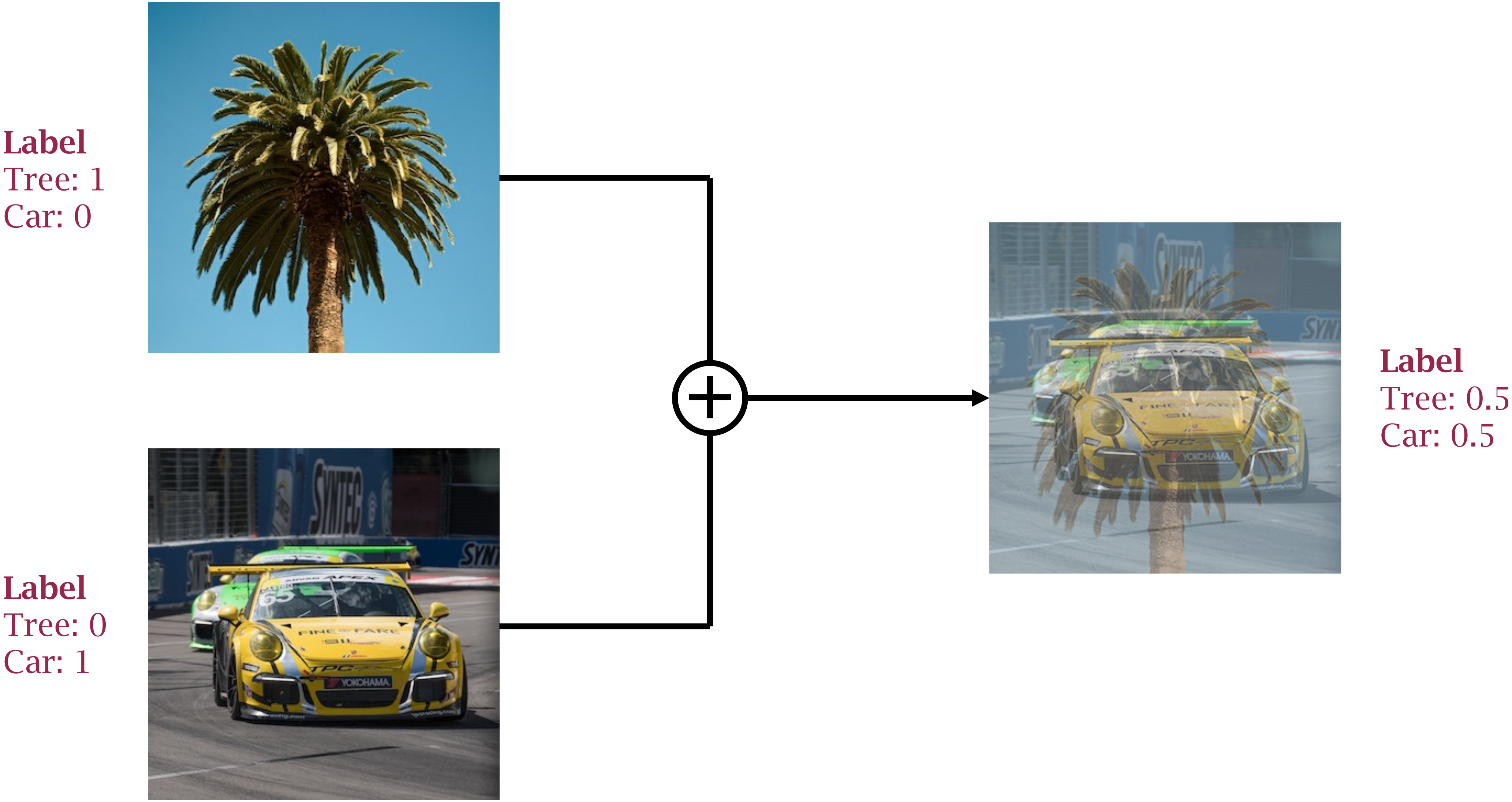


Convex clustering as a characterization of sample similarity



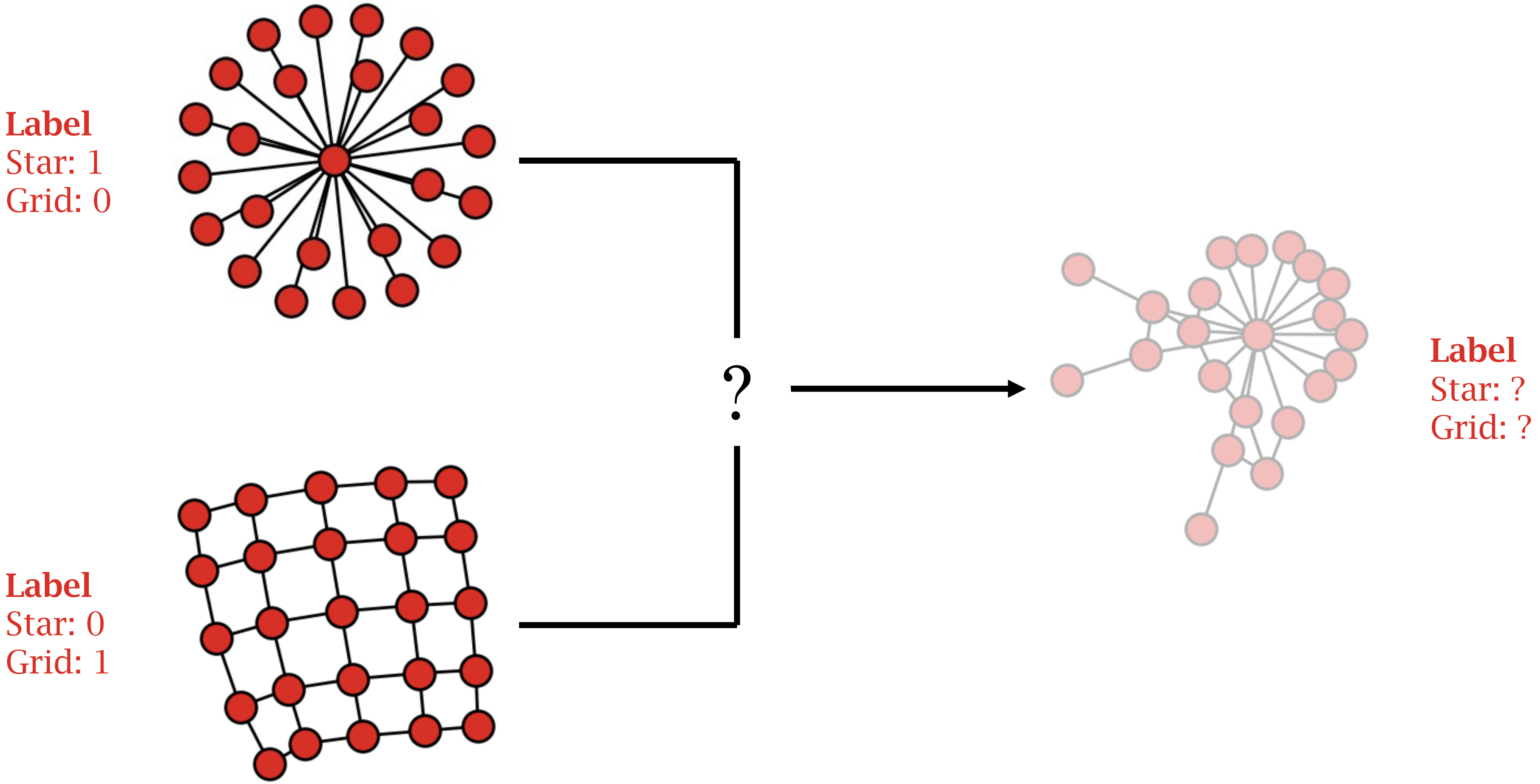
MAD directions

- Mixup method $\Longrightarrow$ Beyond pairwise linear mixup
- Mixup domain $\Longrightarrow$ Beyond Euclidean domains
- Mixup application $\Longrightarrow$ Beyond improving accuracy

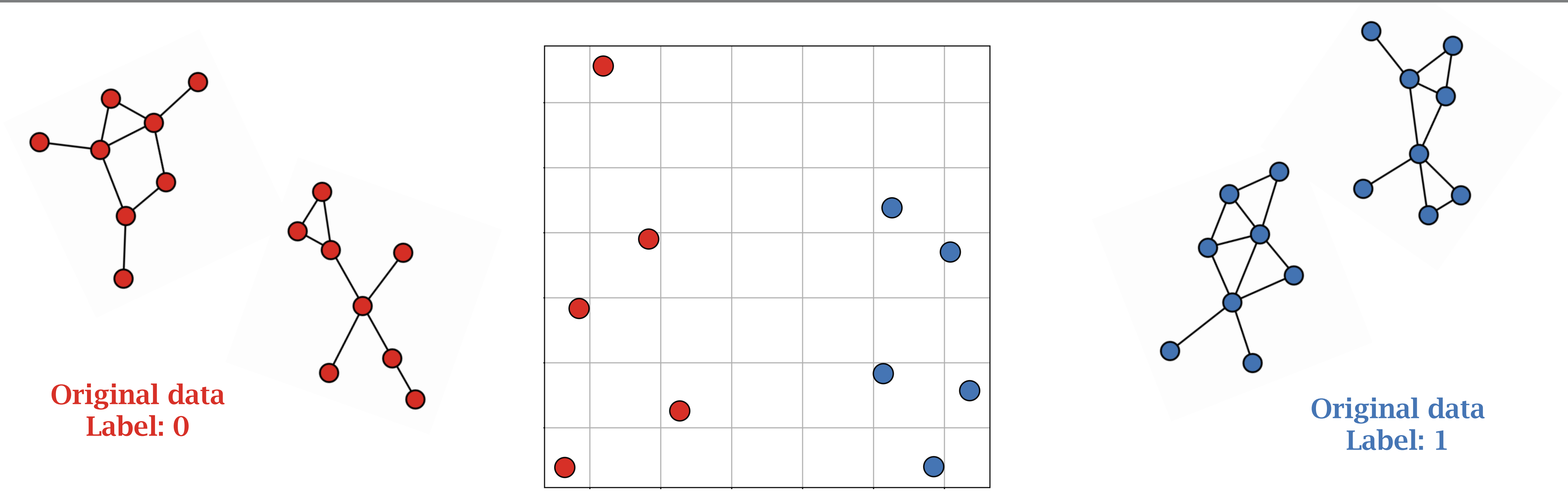


H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, ICLR 2017

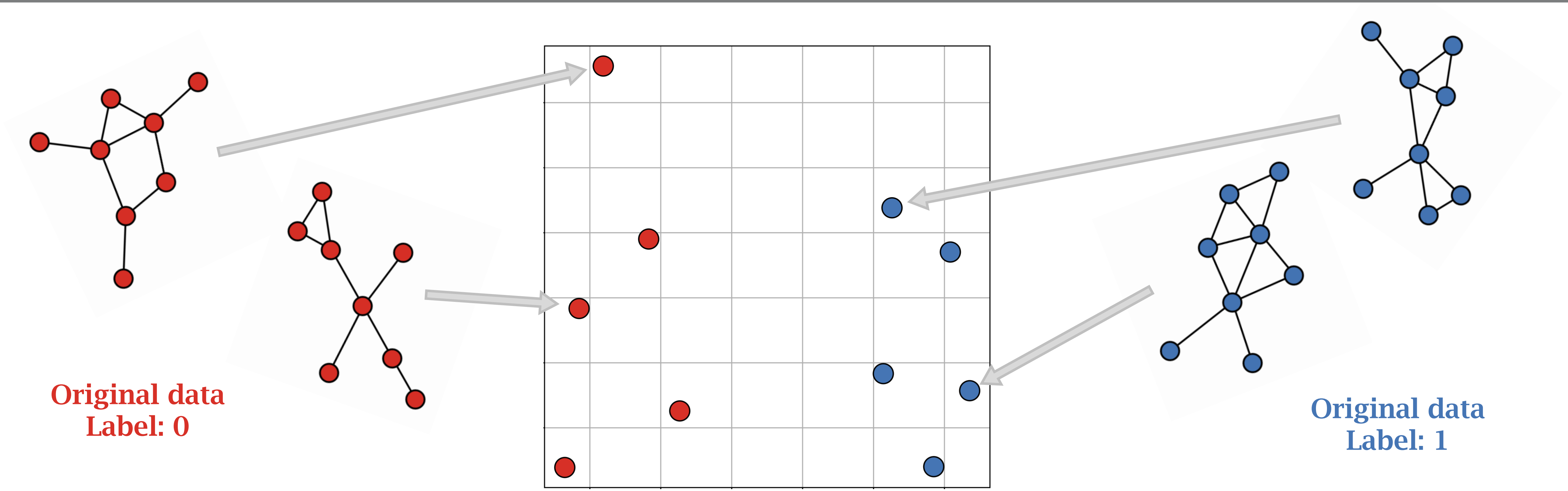
Non-Euclidean graph data is difficult to mixup



Graph Mixup for Augmenting Data (GraphMAD)

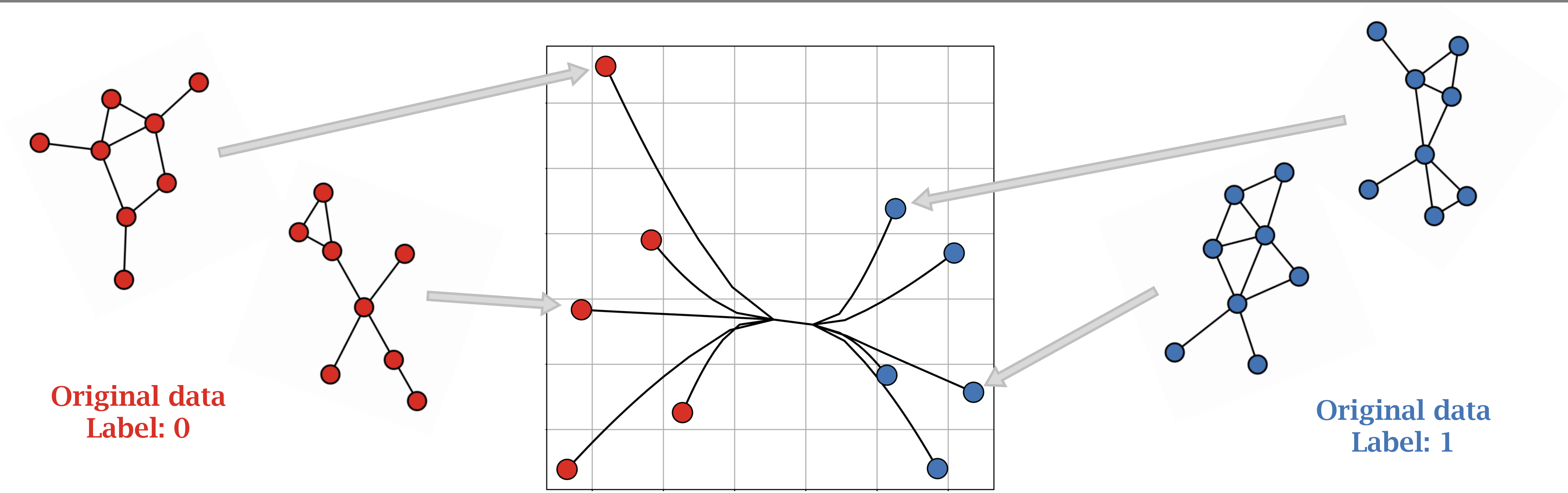


Graph Mixup for Augmenting Data (GraphMAD)



Step 1: Embed graphs

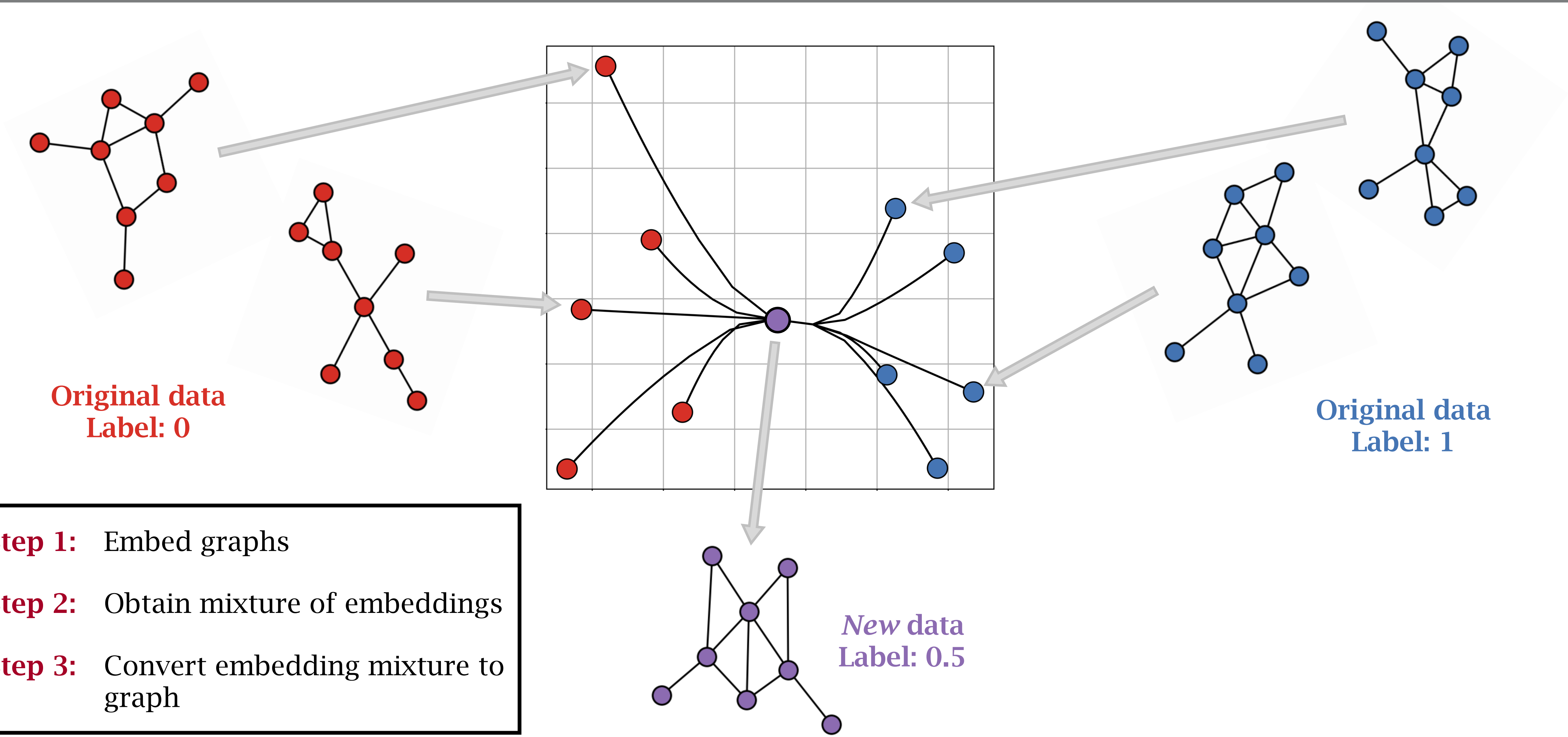
Graph Mixup for Augmenting Data (GraphMAD)



Step 1: Embed graphs

Step 2: Obtain mixture of embeddings

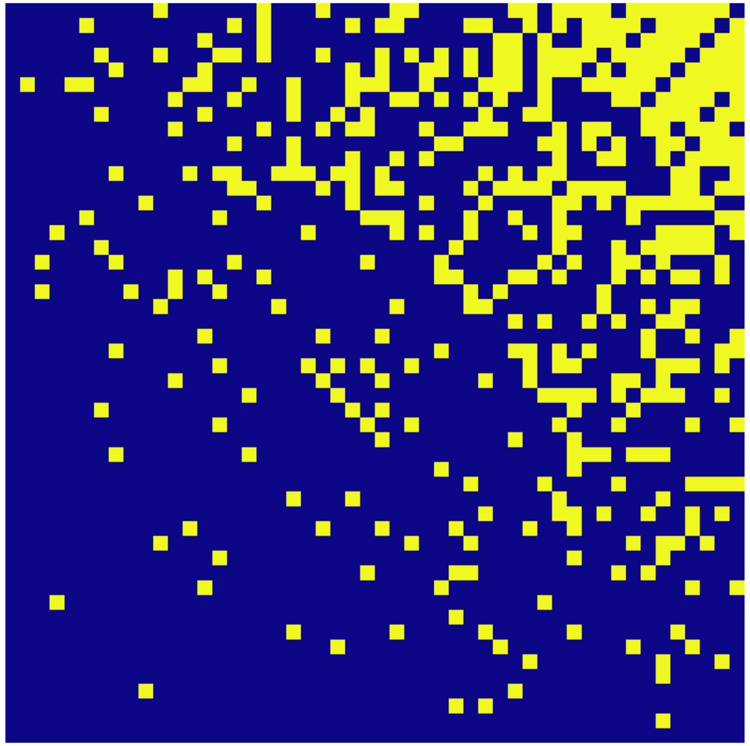
Graph Mixup for Augmenting Data (GraphMAD)



Step 1: Embed graphs

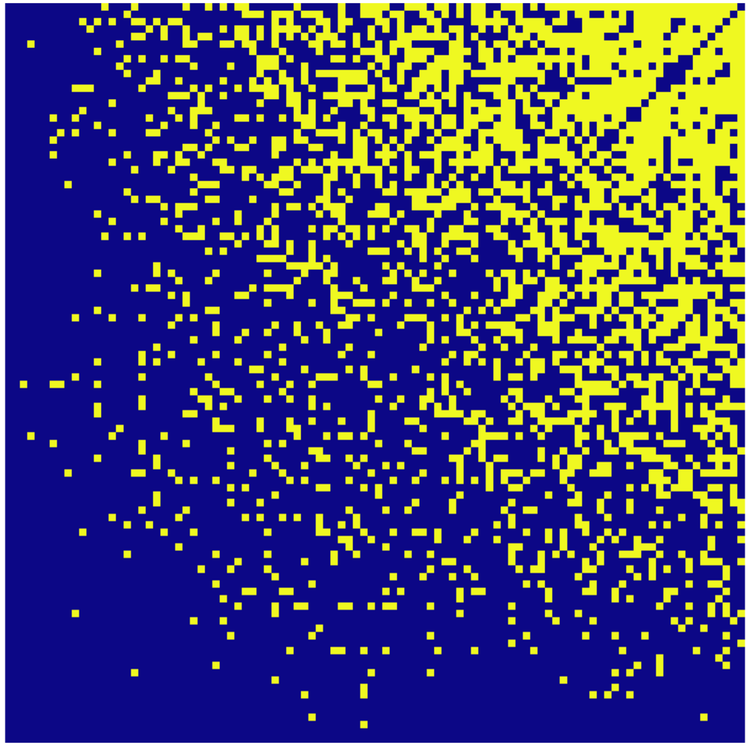
Step 2: Obtain mixture of embeddings

Step 3: Convert embedding mixture to graph



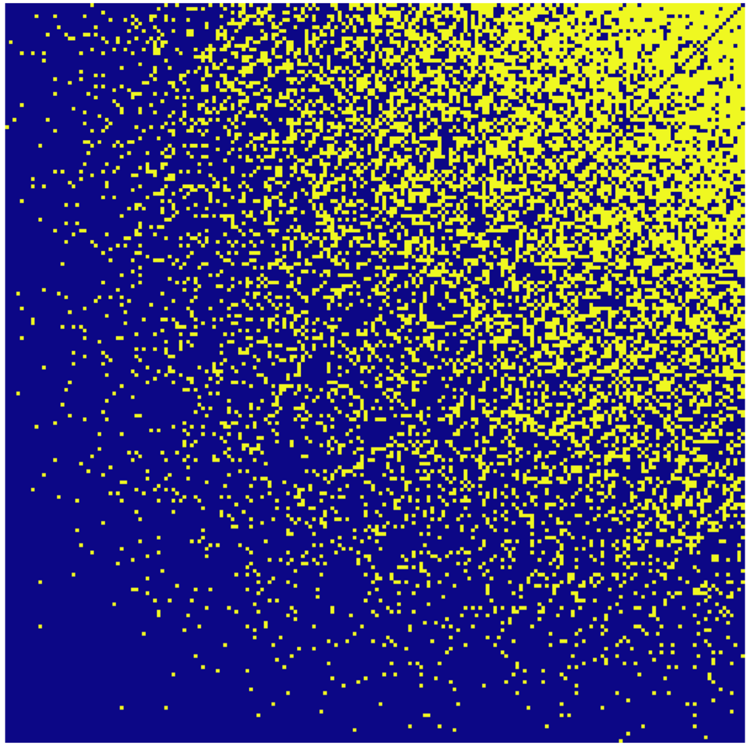
50 nodes

Limit objects as embedding space

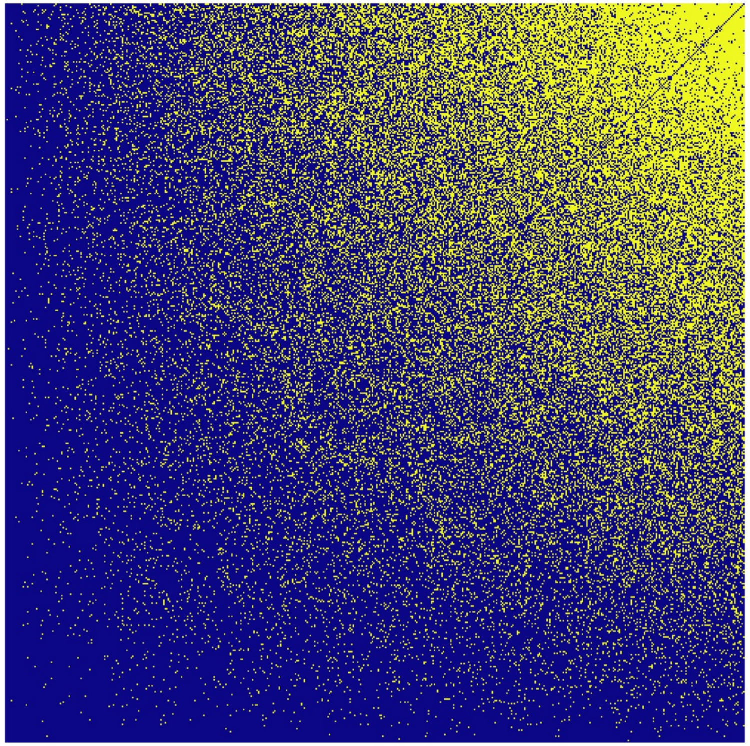


100 nodes

Limit objects as embedding space

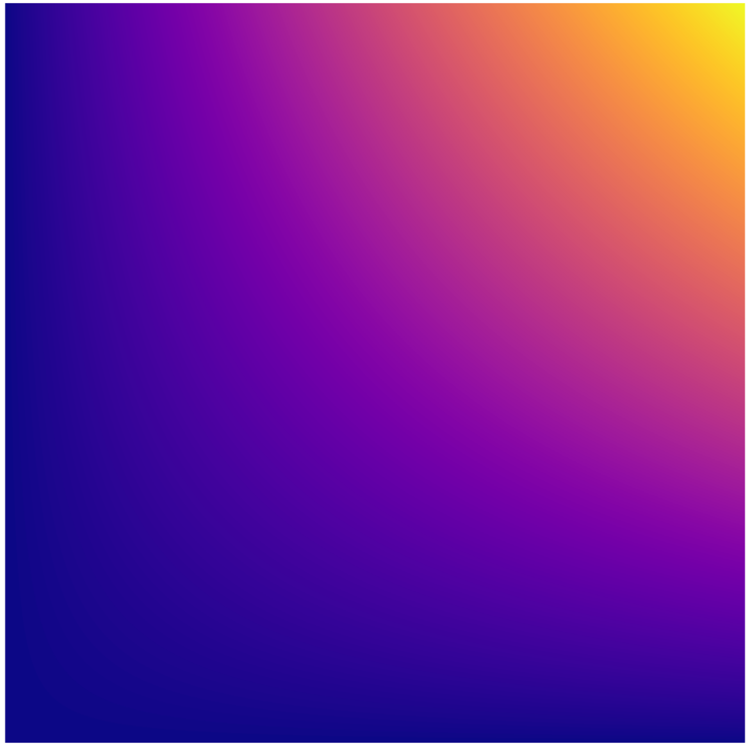


200 nodes



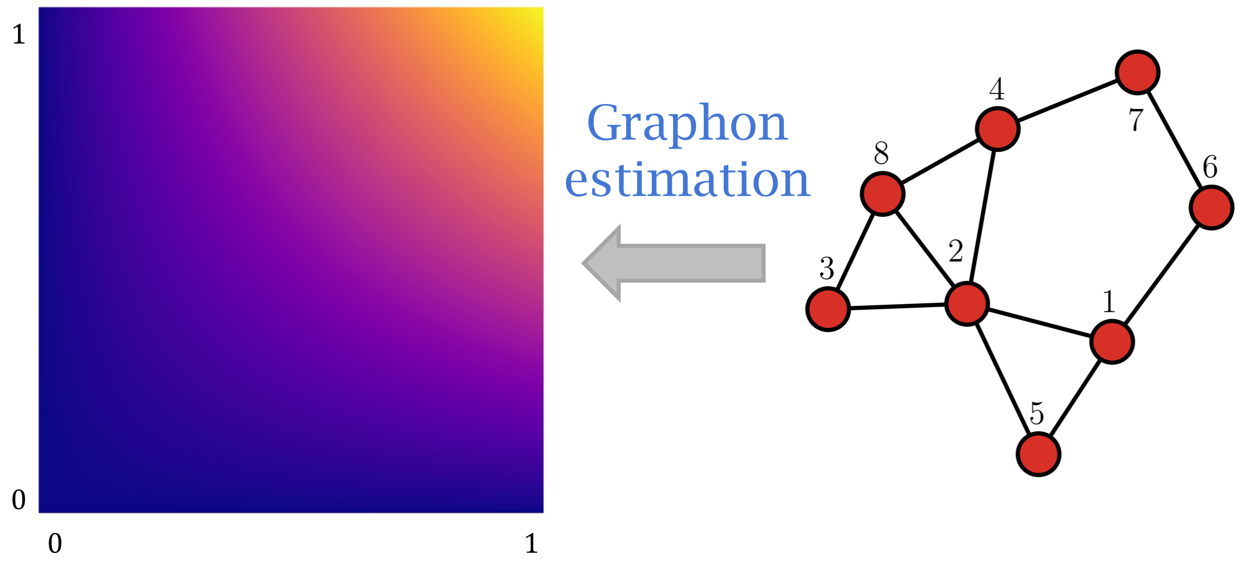
500 nodes

Limit objects as embedding space



Convergence to graphon
in cut distance

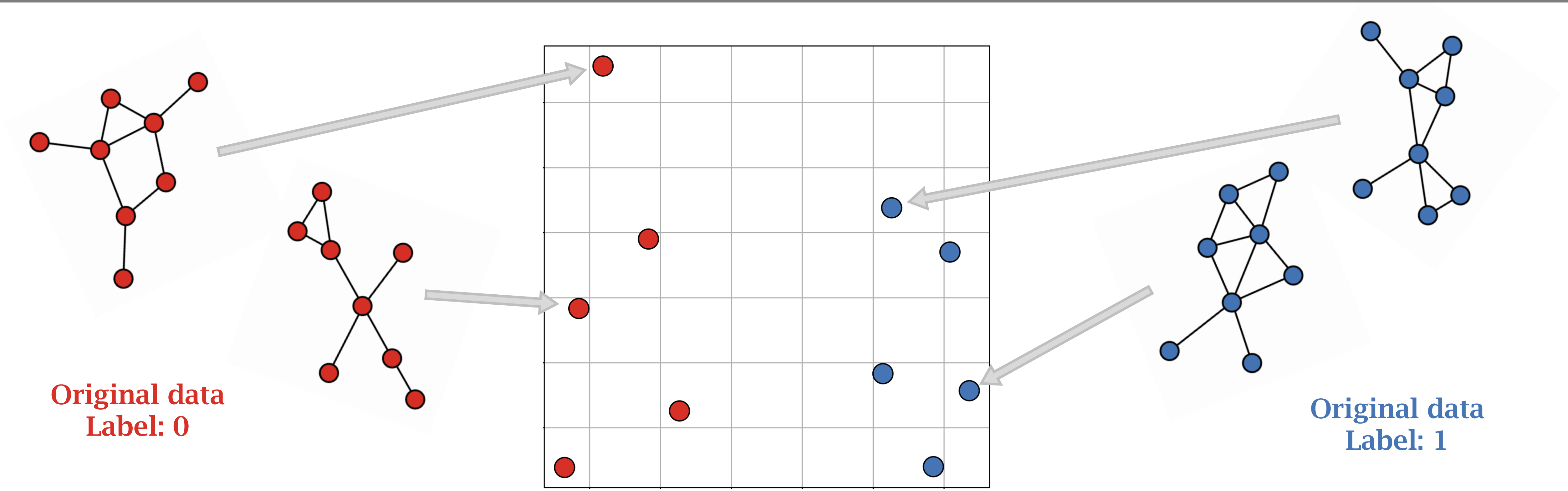
Limit objects as embedding space



Estimated graphon
$\widehat{W}: [0,1]^2 \to [0,1]$

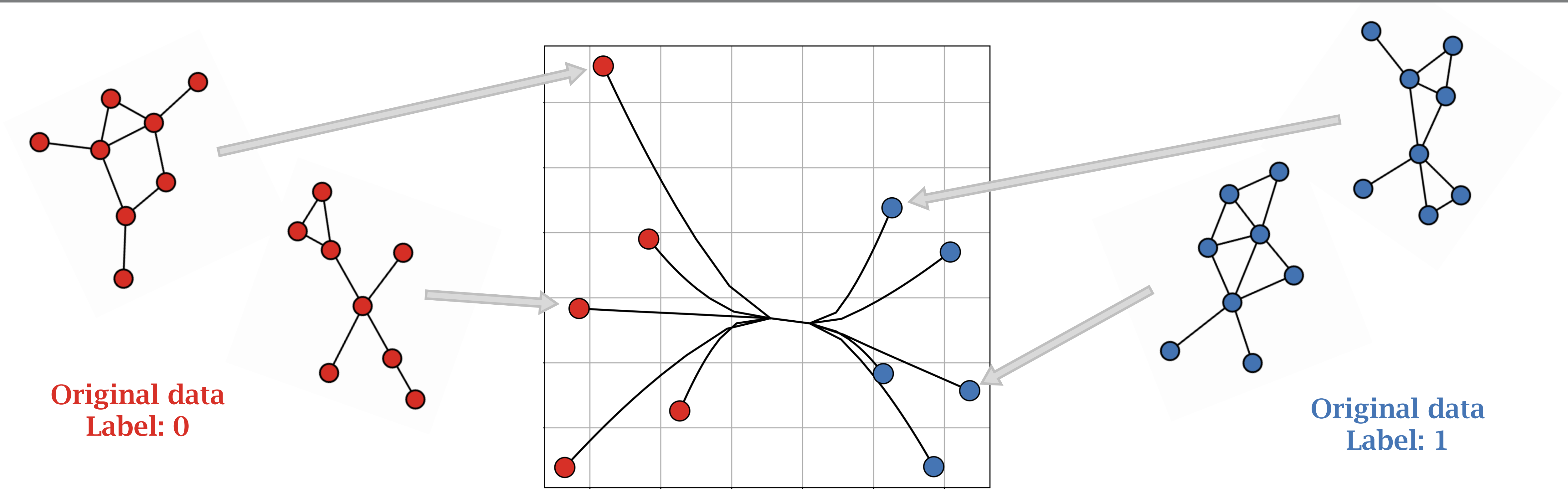
Graph
G

Graph Mixup for Augmenting Data (GraphMAD)



Step 1: Embed graphs

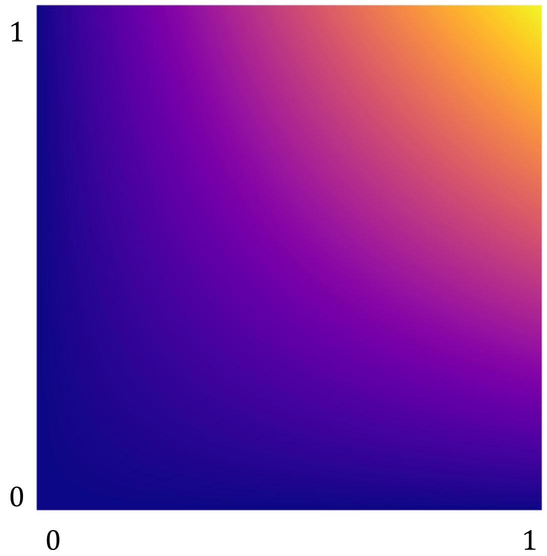
Graph Mixup for Augmenting Data (GraphMAD)



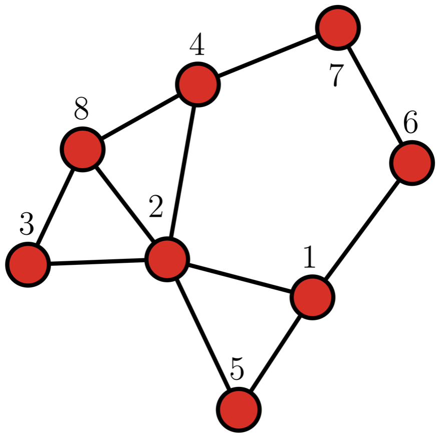
Step 1: Embed graphs

Step 2: Obtain mixture of embeddings

Stochastic inversion of graphon embedding allows multiple views of new data



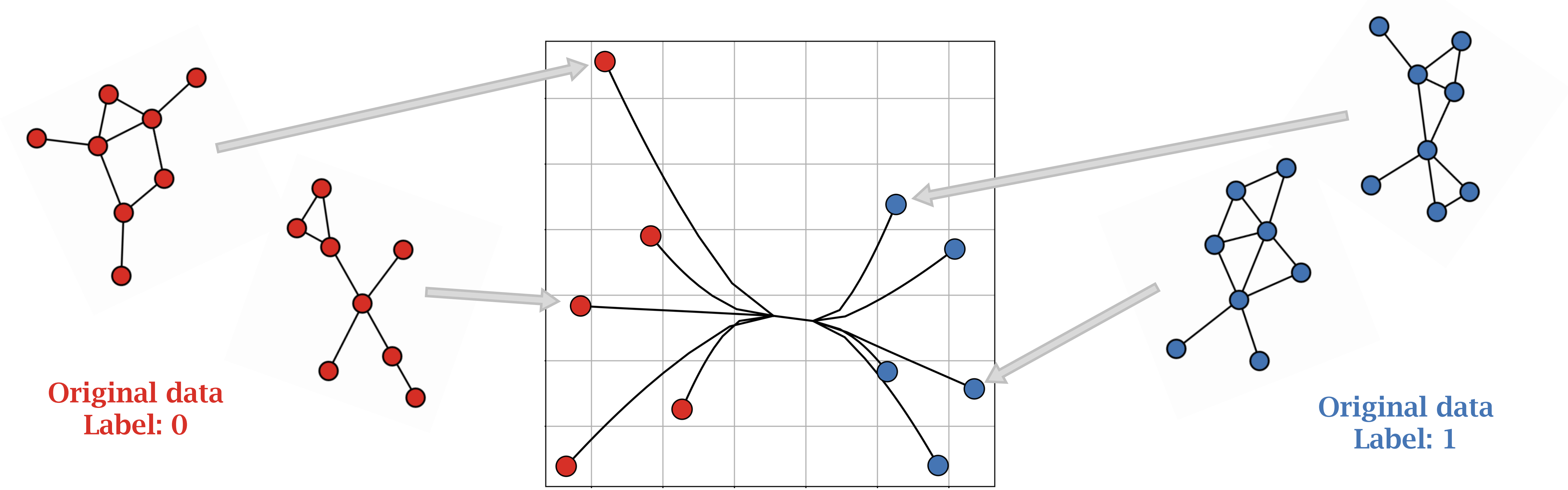
Graphon
$\mathcal{W}: [0,1]^2 \to [0,1]$



Sampled graph
$G \sim \mathcal{W}$

L. Lovász, AMS 2012

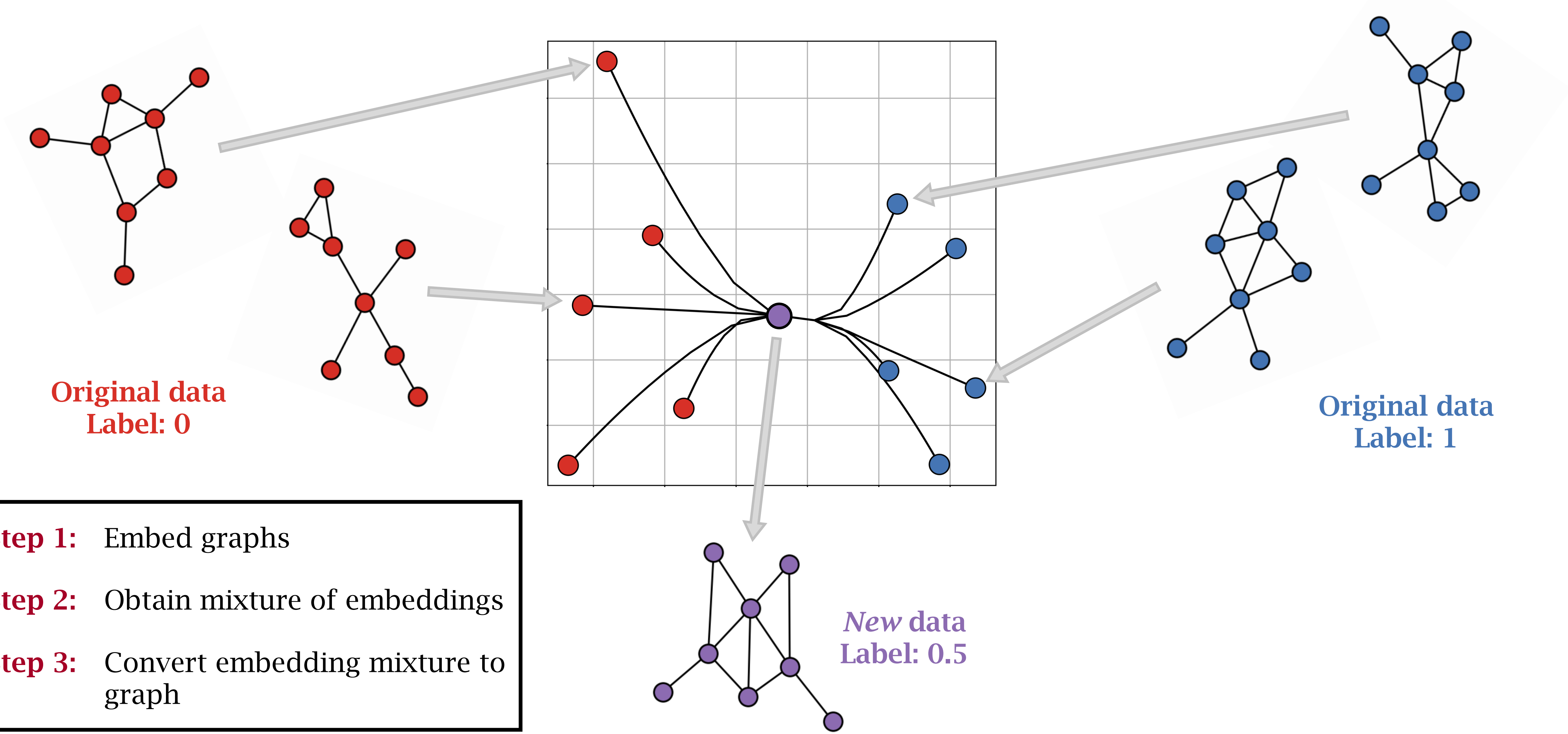
Graph Mixup for Augmenting Data (GraphMAD)



Step 1: Embed graphs

Step 2: Obtain mixture of embeddings

Graph Mixup for Augmenting Data (GraphMAD)



Step 1:	Embed graphs
Step 2:	Obtain mixture of embeddings
Step 3:	Convert embedding mixture to graph

GraphMAD improves performance and outperforms linear mixup on all datasets

Graph classification accuracy on molecule and bioinformatics datasets

Method		DD	PROTEINS	ENZYMES	AIDS	MUTAG	NCI109
Data mixup	**Label mixup**	**2 classes**	**2 classes**	**6 classes**	**2 classes**	**2 classes**	**2 classes**
None	None	68.77 ± 2.35	**69.51 ± 1.20**	26.43 ± 2.55	96.18 ± 2.57	84.59 ± 5.53	68.23 ± 2.13
Linear	Linear	67.01 ± 1.72	65.15 ± 2.53	24.88 ± 3.38	96.82 ± 1.39	85.71 ± 7.15	68.16 ± 2.72
	Sigmoid	64.89 ± 1.49	68.42 ± 3.94	24.76 ± 4.10	96.07 ± 1.42	85.71 ± 4.63	65.96 ± 2.34
	Logit	66.22 ± 3.82	69.25 ± 2.94	25.95 ± 5.48	96.07 ± 1.27	80.08 ± 5.60	66.81 ± 4.07
	Cvx. Clust.	68.22 ± 3.71	69.38 ± 2.04	24.64 ± 2.39	95.86 ± 1.88	**87.22 ± 4.96**	65.01 ± 3.07
Cvx. Clust.	Linear	67.11 ± 1.56	67.51 ± 2.62	**26.67 ± 6.49**	**97.15 ± 1.00**	**87.24 ± 4.21**	**68.61 ± 1.41**
	Sigmoid	68.23 ± 3.61	64.60 ± 5.07	**32.62 ± 6.35**	97.07 ± 1.35	85.20 ± 3.53	67.50 ± 2.06
	Logit	**70.07 ± 2.51**	67.26 ± 2.84	25.71 ± 4.26	95.87 ± 1.47	80.10 ± 14.77	65.33 ± 3.35
	Cvx. Clust.	**70.44 ± 3.79**	**71.18 ± 3.98**	24.52 ± 3.30	**97.22 ± 0.54**	85.71 ± 5.40	**68.54 ± 3.16**

Data augmentation with GraphMAD consistently outperforms linear mixup, and different label mixup functions can improve accuracy

MAD directions

- Mixup method ⟹ Beyond pairwise linear mixup
- Mixup domain ⟹ Beyond Euclidean domains
- Mixup application ⟹ Beyond improving accuracy

Machine learning models may act harmfully in the presence of sensitive information

Reuters

World

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

Dastin, Reuters 2018

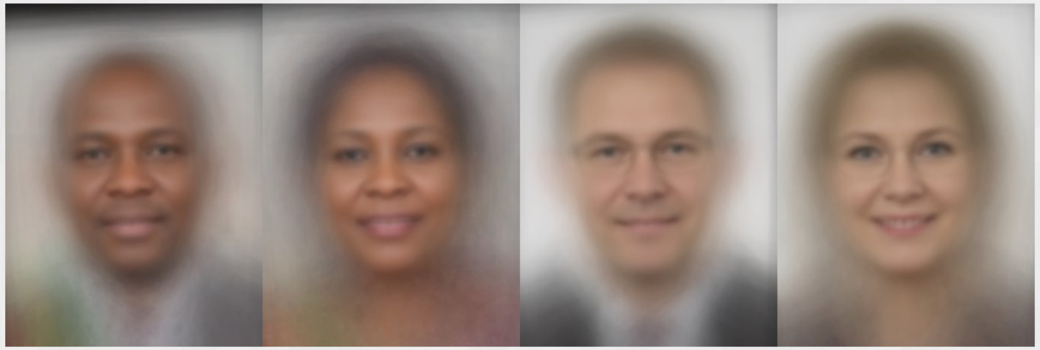
Online images amplify gender bias

Douglas Guilbeault, Solène Delecourt, Tasker Hull, Bhargav Srinivasa Desikan, Mark Chu & Ethan Nadler

Nature **626**, 1049–1055 (2024) | Cite this article

Guilbeault, Nature 2024

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Buolamwini and Gebru, FAT 2018

Group fairness encourages similar treatment for each group separately

Group fairness: Treatment invariant to different values of sensitive attribute

Group fairness encourages similar treatment for each group separately

Group fairness: Treatment invariant to different values of sensitive attribute

Demographic parity: Predictions $\hat{Y} = f(X)$ independent of sensitive attribute $Z \in \{0,1\}$

$$\mathbb{P}\left[\hat{Y} = y | Z = 0\right] = \mathbb{P}\left[\hat{Y} = y | Z = 1\right]$$

Group fairness encourages similar treatment for each group separately

Group fairness: Treatment invariant to different values of sensitive attribute

Demographic parity: Predictions $\hat{Y} = f(X)$ independent of sensitive attribute $Z \in \{0,1\}$

$$\mathbb{P}\left[\hat{Y} = y | Z = 0\right] = \mathbb{P}\left[\hat{Y} = y | Z = 1\right]$$

In practice: $\Delta\text{DP} = \widehat{\mathbb{E}}\left[\hat{Y} = y | Z = 0\right] - \widehat{\mathbb{E}}\left[\hat{Y} = y | Z = 1\right]$

Group fairness encourages similar treatment for each group separately

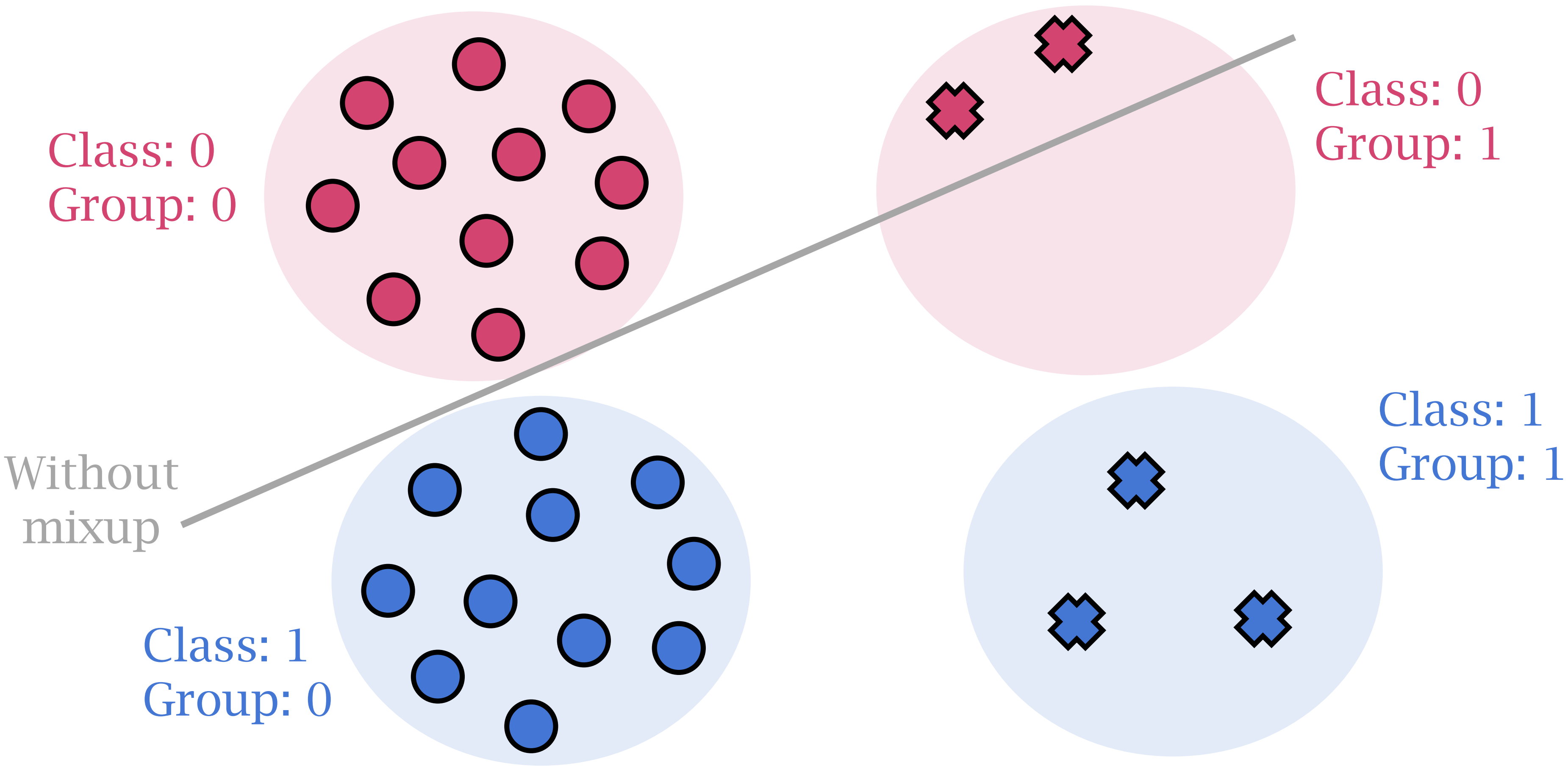
Group fairness: Treatment invariant to different values of sensitive attribute

Demographic parity: Predictions $\hat{Y} = f(X)$ independent of sensitive attribute $Z \in \{0,1\}$

$$\mathbb{P}[\hat{Y} = y | Z = 0] = \mathbb{P}[\hat{Y} = y | Z = 1]$$

In practice: $\Delta DP = \widehat{\mathbb{E}}[\hat{Y} = y | Z = 0] - \widehat{\mathbb{E}}[\hat{Y} = y | Z = 1] = 0$ DP achieved!

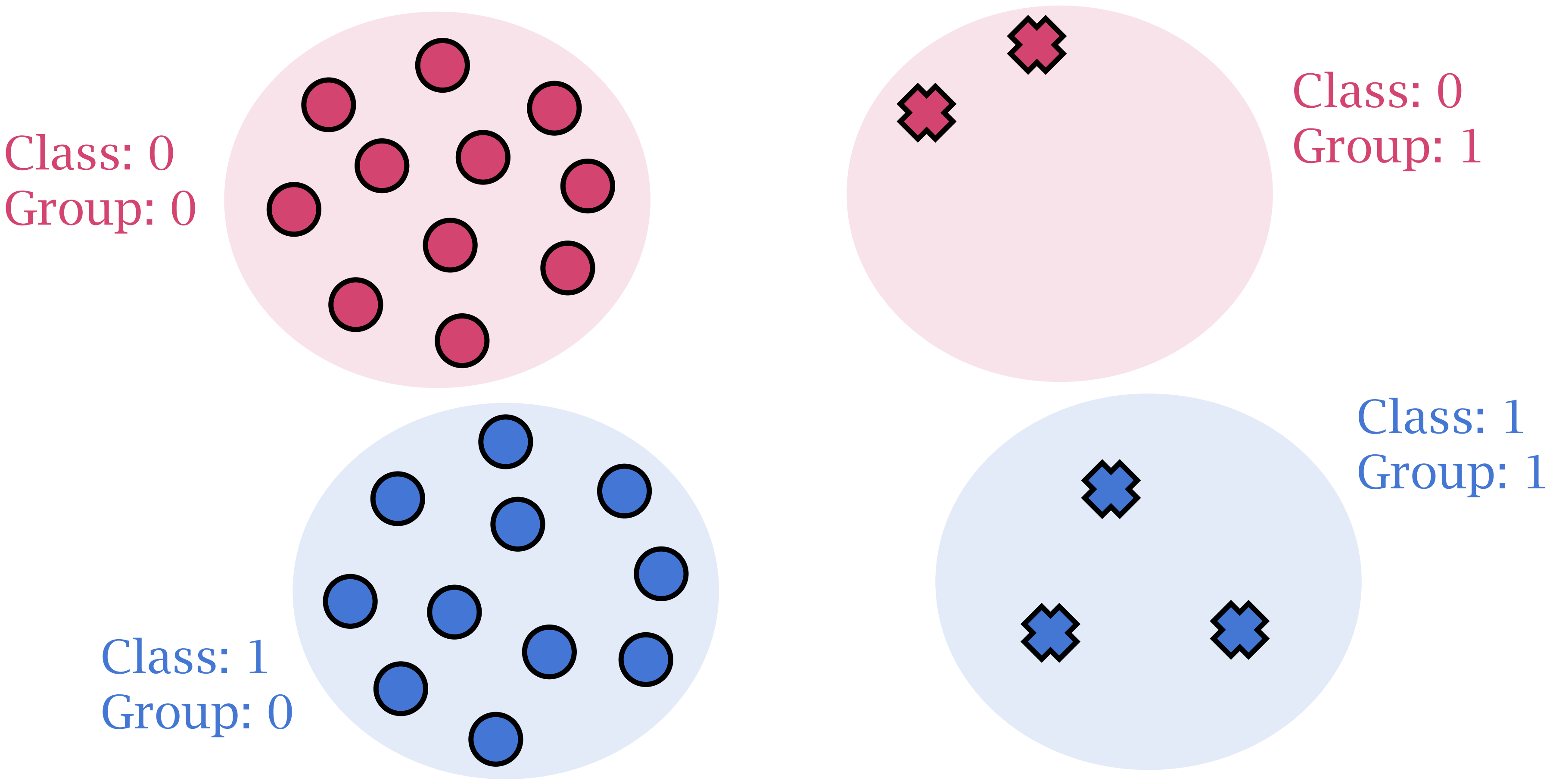
Fair SubGroup Mixup (FSGM) mixes samples across subgroups to mitigate bias



Bias due to underrepresented groups or shifts in distribution across groups

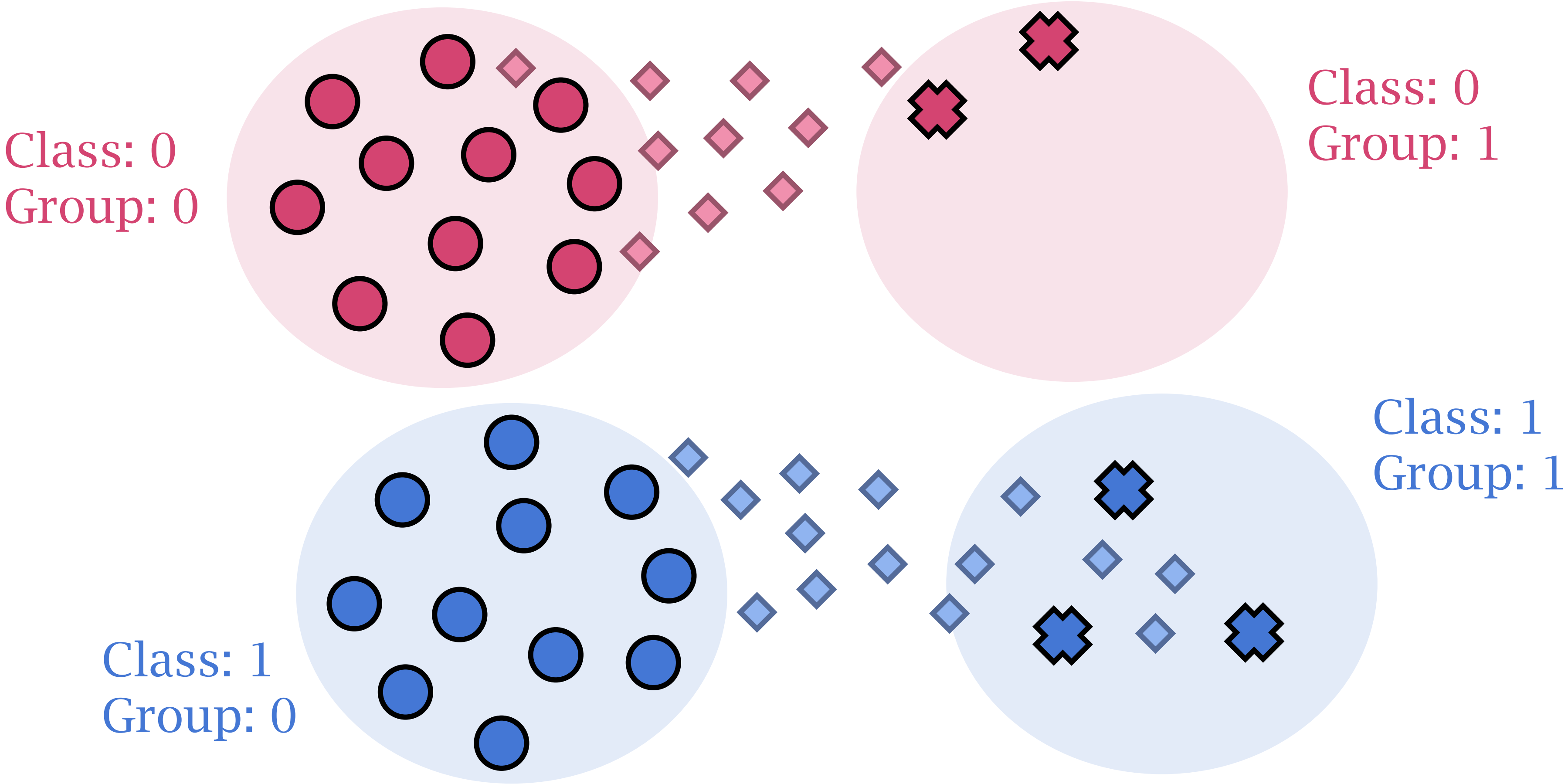
C.-Y. Chuang and Y. Mroueh, ICLR 2021

Fair SubGroup Mixup (FSGM) mixes samples across subgroups to mitigate bias



FSGM: Pairwise mixup between source subgroup and target subgroup

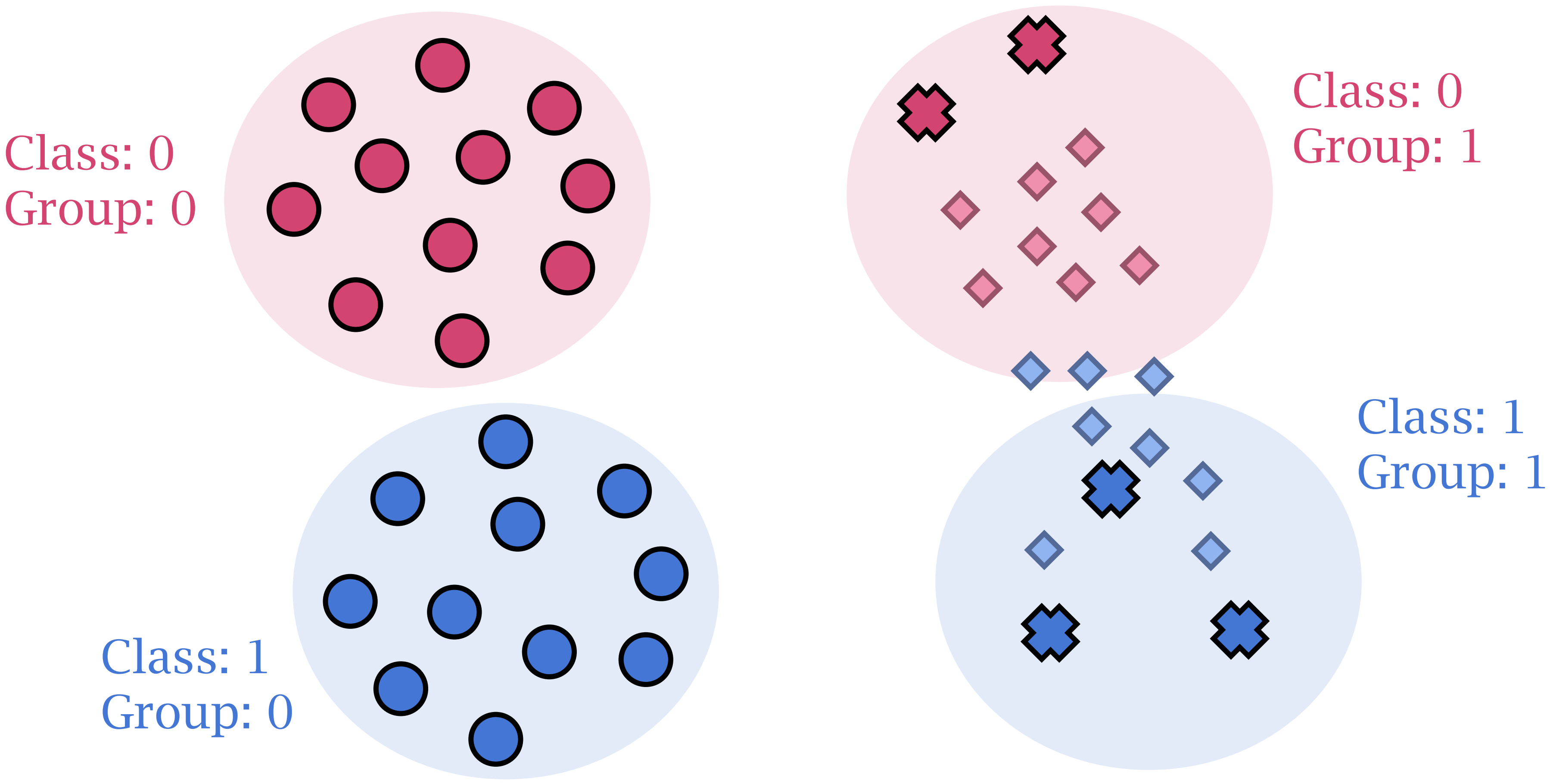
Fair SubGroup Mixup (FSGM) mixes samples across subgroups to mitigate bias



FSGM: Pairwise mixup between source subgroup and target subgroup

Mixup across groups promotes invariance between groups

Fair SubGroup Mixup (FSGM) mixes samples across subgroups to mitigate bias

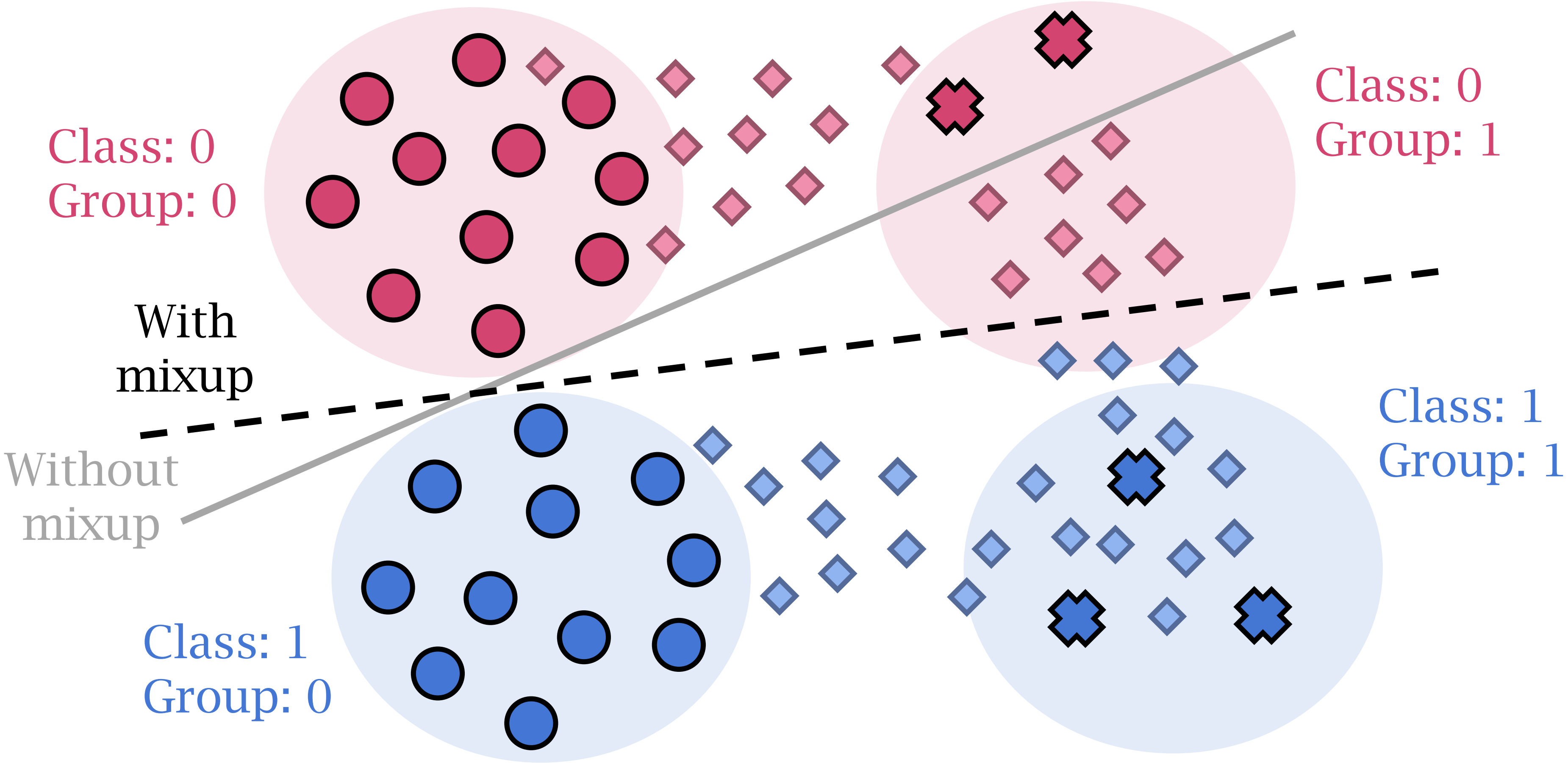


FSGM: Pairwise mixup between source subgroup and target subgroup

Mixup across classes promotes learning separately per group

nav@rice.edu

Fair SubGroup Mixup (FSGM) mixes samples across subgroups to mitigate bias

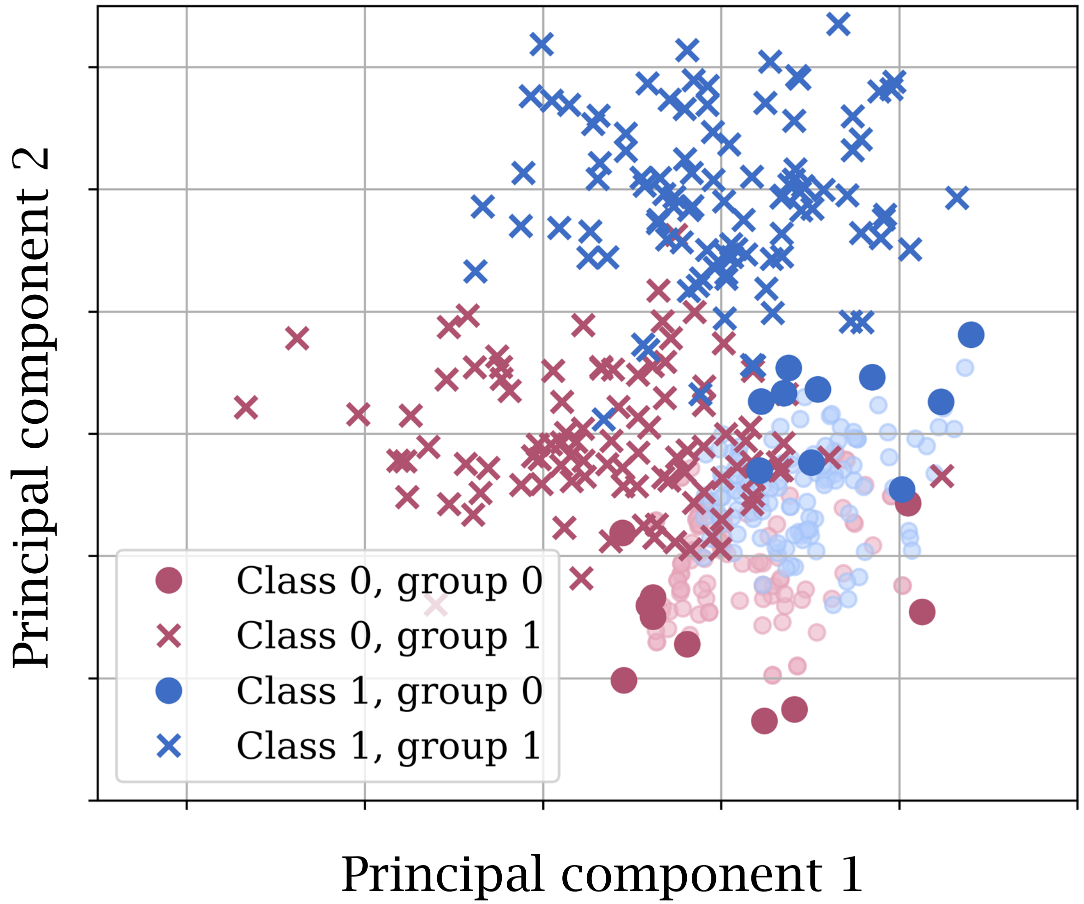


FSGM: Pairwise mixup between source subgroup and target subgroup

FSGM addresses two types of bias in data

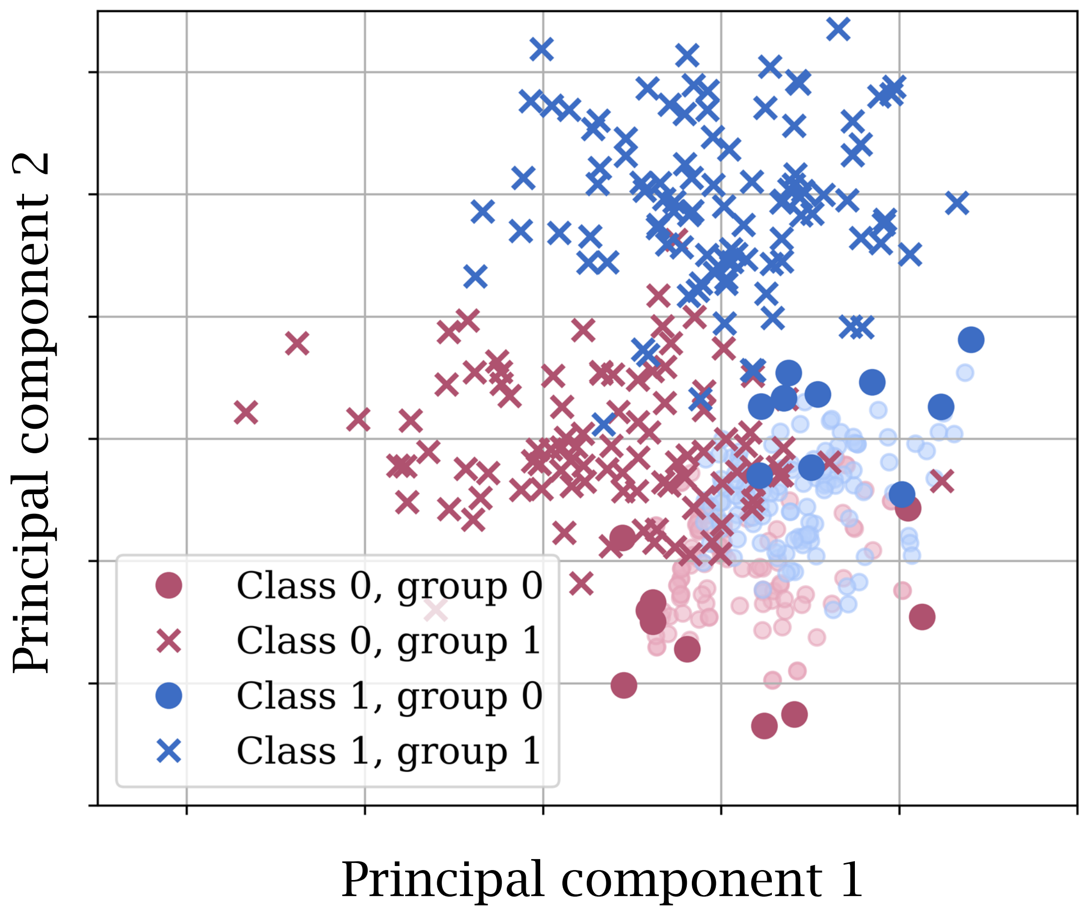
nav@rice.edu

Unbalanced groups result in unfair treatment of underrepresented group



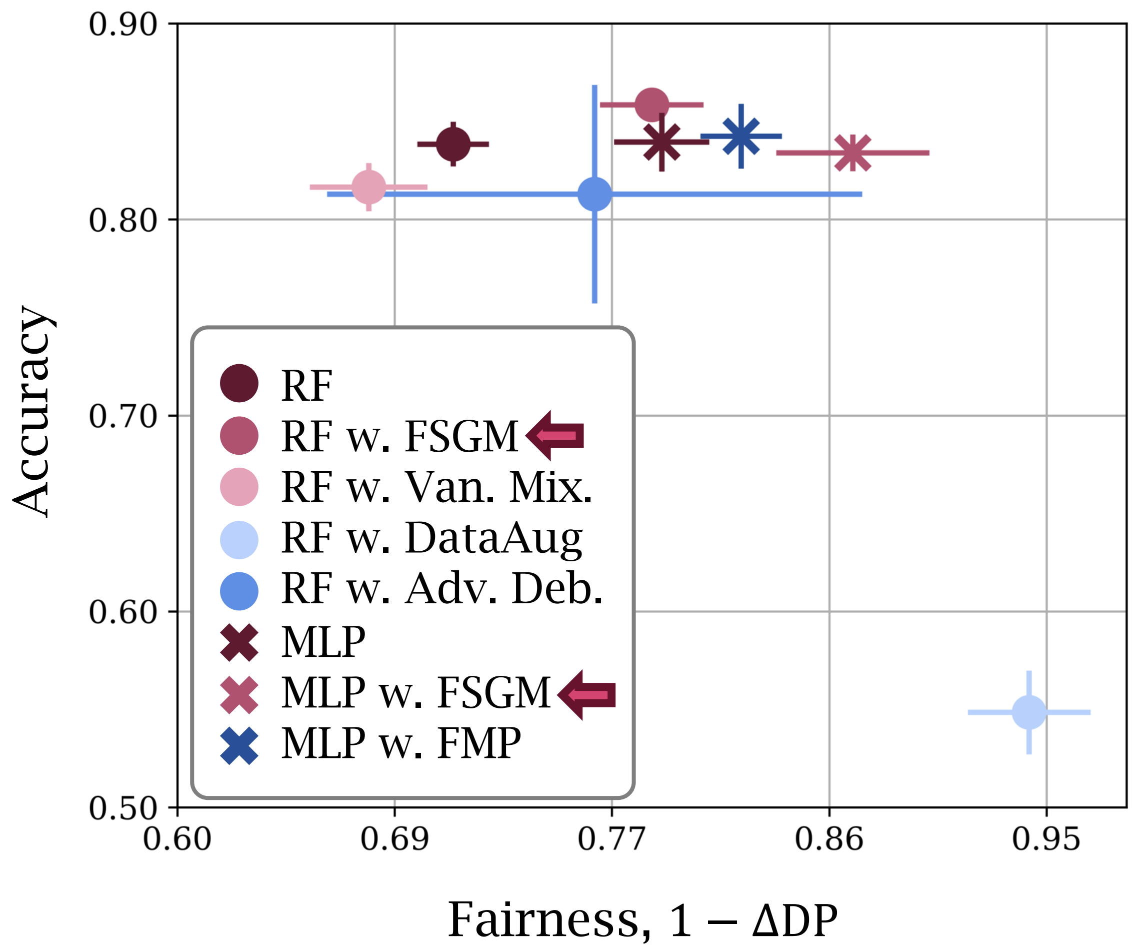
Unbalanced groups: Model treatment heavily influenced by overrepresented group

Unbalanced groups result in unfair treatment of underrepresented group



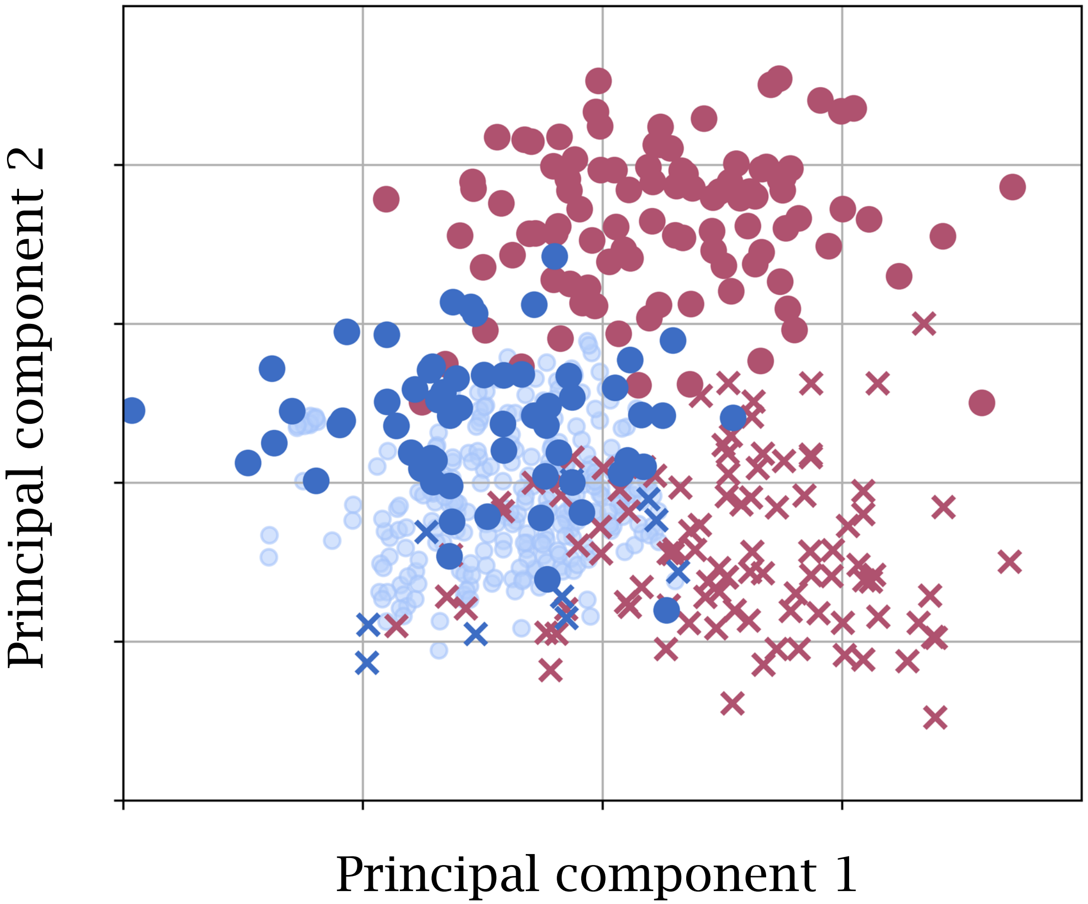
Mixup between classes of underrepresented group encourages more certain decision boundary

Mixup between classes of underrepresented group encourages confident decision boundary



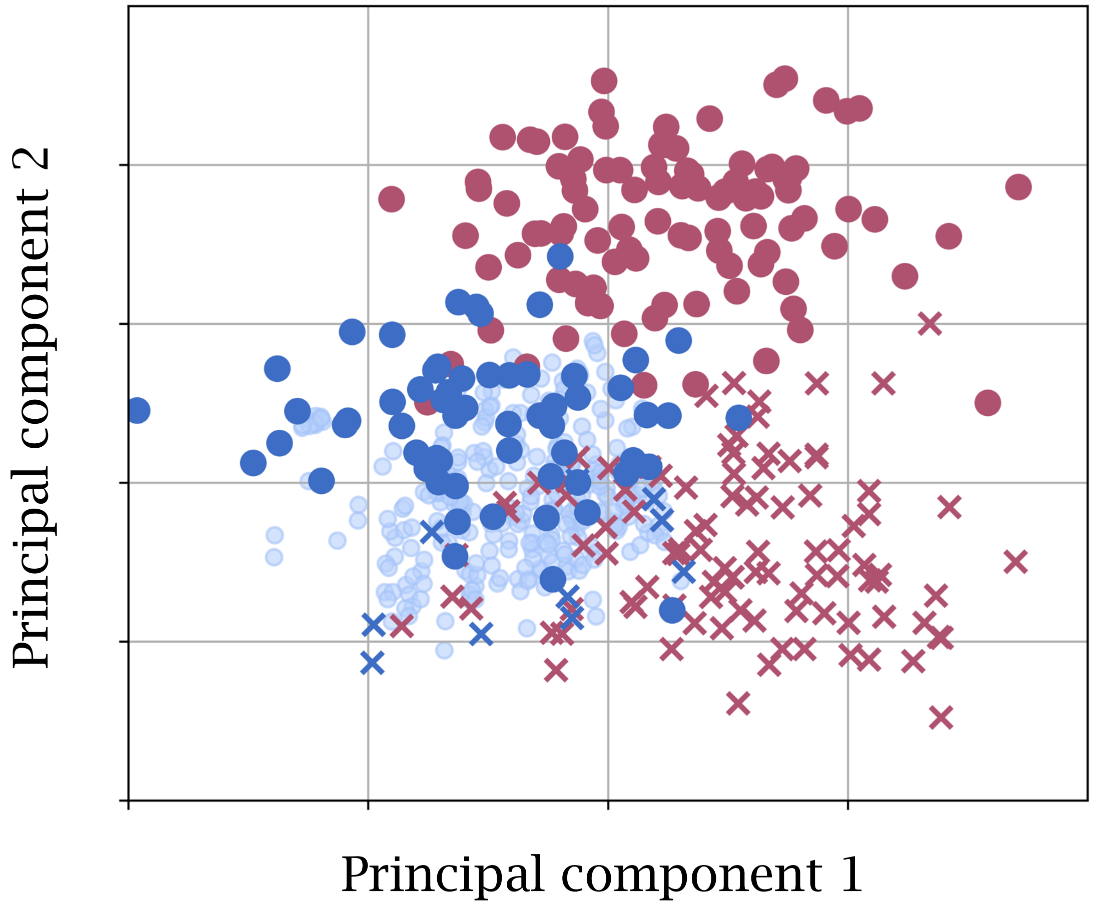
Fair SubGroup Mixup (FSGM) improves both accuracy and fairness above existing fairness and data augmentation methods

Unbalanced classes can violate demographic parity



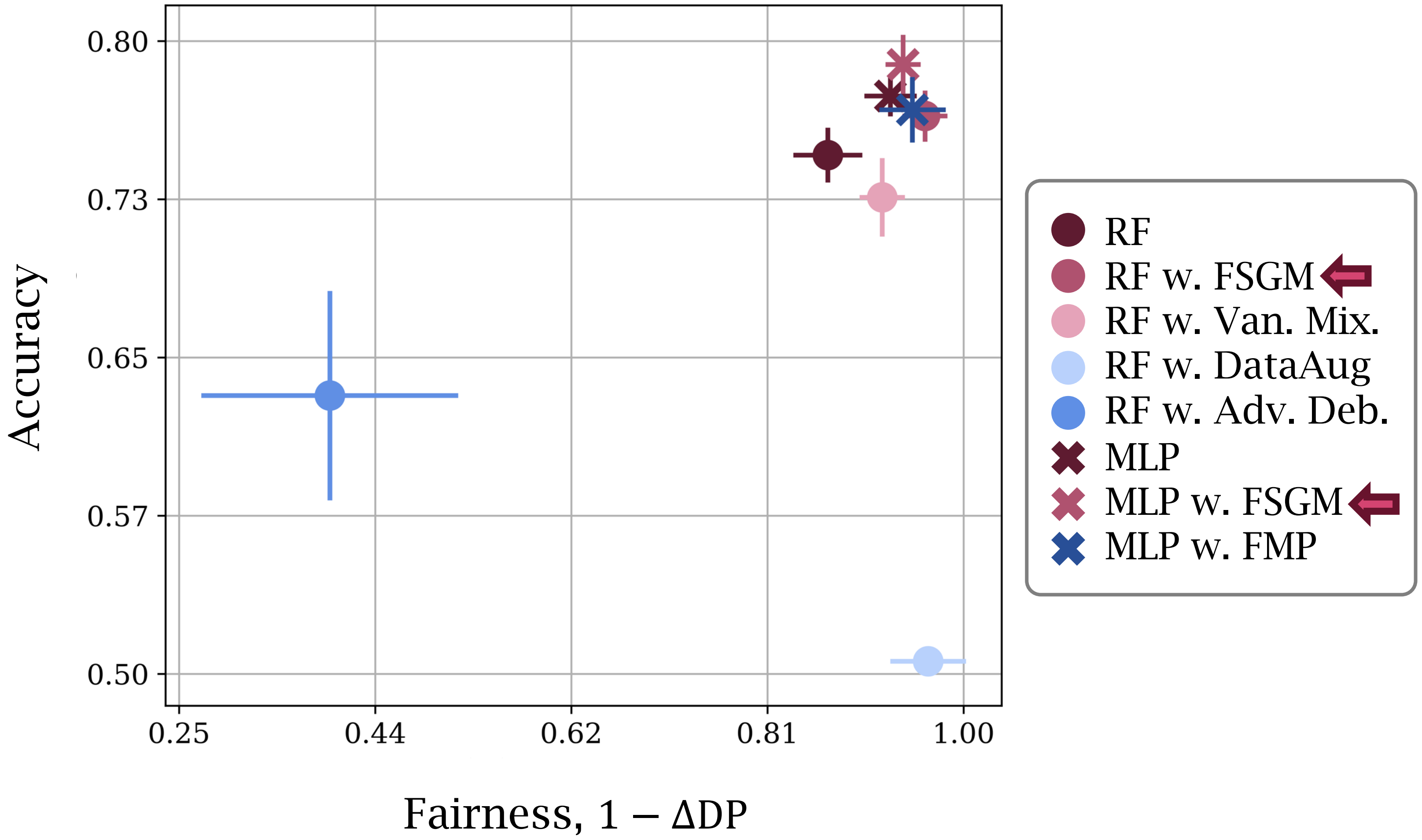
Unbalanced classes: Gaps between groups in minority class may result in demographic parity gap

Unbalanced classes can violate demographic parity



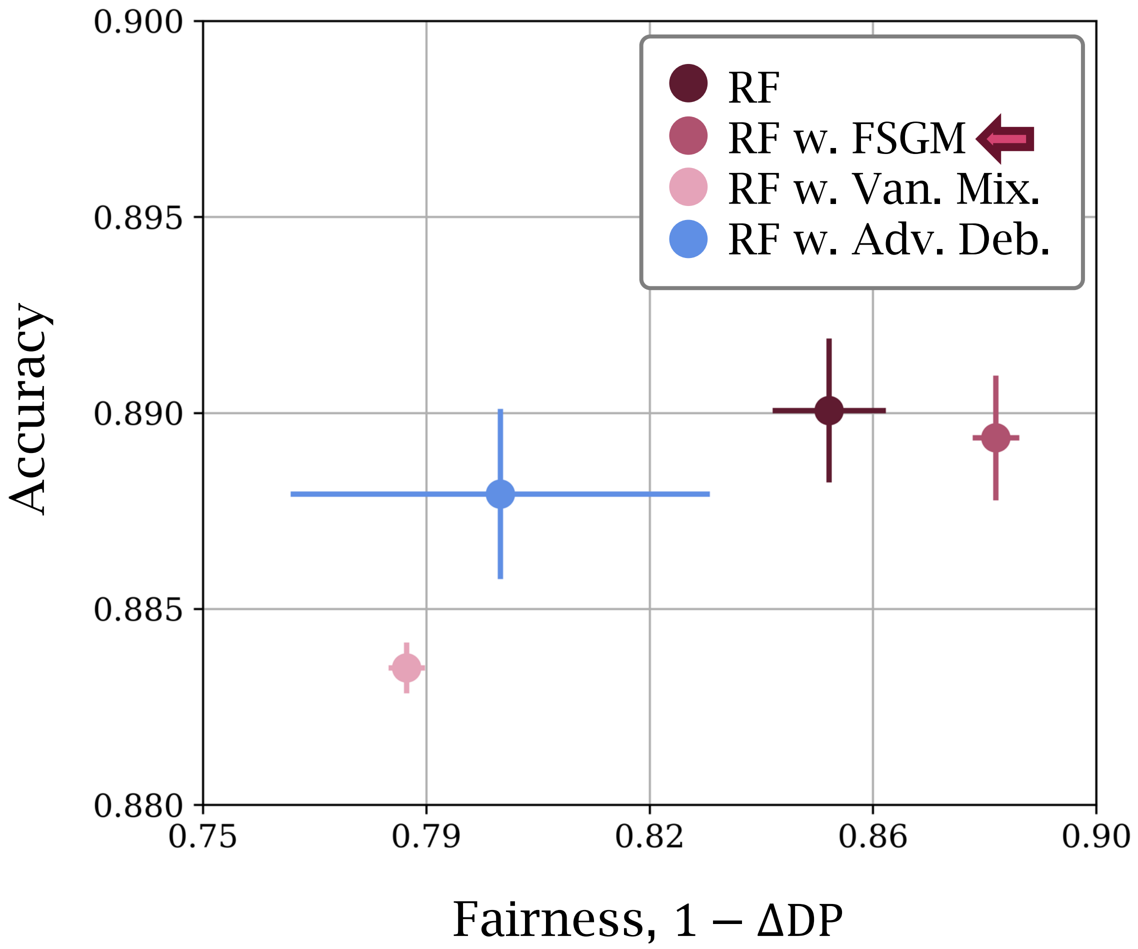
Mixup between groups of minority class encourages
similar group treatment, demographic parity

Mixup within underrepresented class encourages demographic parity



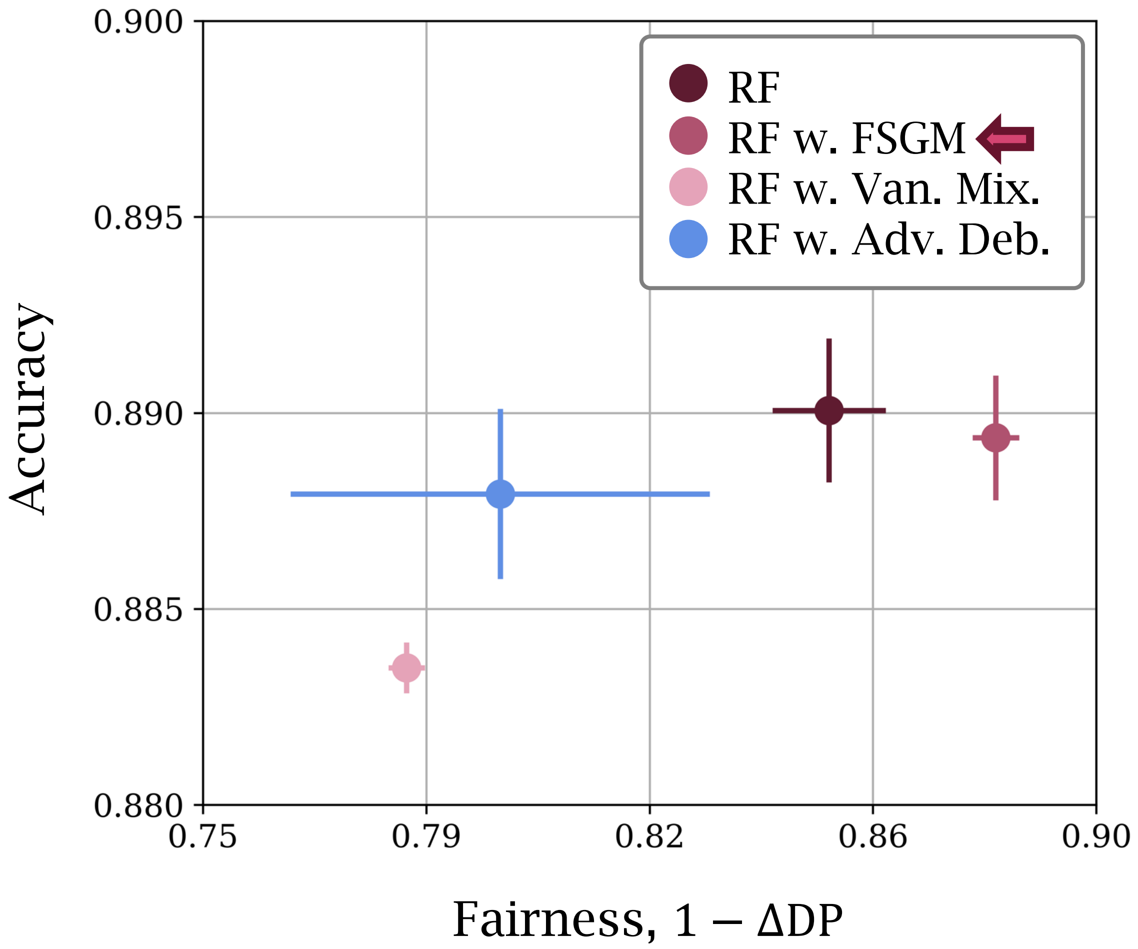
Fair SubGroup Mixup (FSGM) improves accuracy and achieves fairness rivaling the fairest method

Law school admission bar passage with race as protected attribute



Class: Bar passage (yes or no)
Group: Race (white or non-white)

Law school admission bar passage with race as protected attribute



On real-world benchmark dataset, FSGM improves fairness with robust accuracy compared to baselines

MAD overview

- Mixup method $\Longrightarrow$ Mixup using informative convex clustering
- Mixup domain $\Longrightarrow$ Mixtures of non-Euclidean graphs
- Mixup application $\Longrightarrow$ Applying mixup for improving model fairness

MAD overview

- Mixup method ⟹ Mixup using informative convex clustering

  Next steps – Theoretical and empirical evaluation of convex clustering for different applications and domains

- Mixup domain ⟹ Mixtures of non-Euclidean graphs

- Mixup application ⟹ Applying mixup for improving model fairness

MAD overview

- Mixup method ⟹ Mixup using informative convex clustering

  Next steps – Theoretical and empirical evaluation of convex clustering for different applications and domains

- Mixup domain ⟹ Mixtures of non-Euclidean graphs

  Next steps – Effects of mixtures of graphs for data augmentation via graphon theory

- Mixup application ⟹ Applying mixup for improving model fairness

MAD overview

- Mixup method $\implies$ Mixup using informative convex clustering

  Next steps – Theoretical and empirical evaluation of convex clustering for different applications and domains

- Mixup domain $\implies$ Mixtures of non-Euclidean graphs

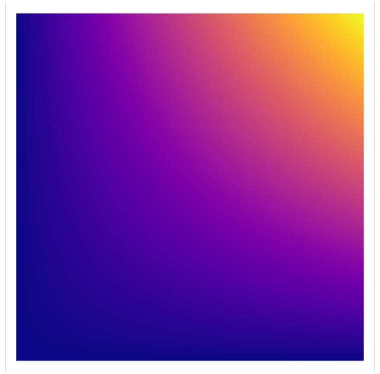
  Next steps – Effects of mixtures of graphs for data augmentation via graphon theory

- Mixup application $\implies$ Applying mixup for improving model fairness

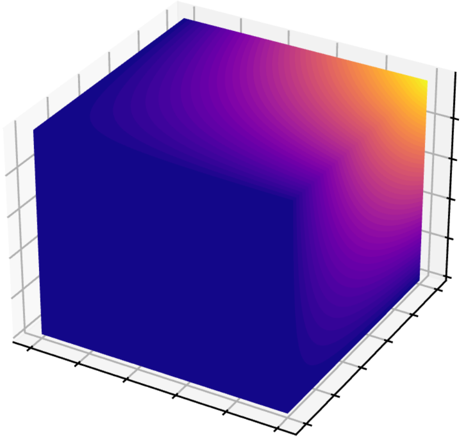
  Next steps – Convex clustering mixup for group fairness, individual fairness, or problems involving intersectionality

Complexon as simplicial complex limit object

Dimension 1
Edge likelihoods



Dimension 2
Triangle likelihoods

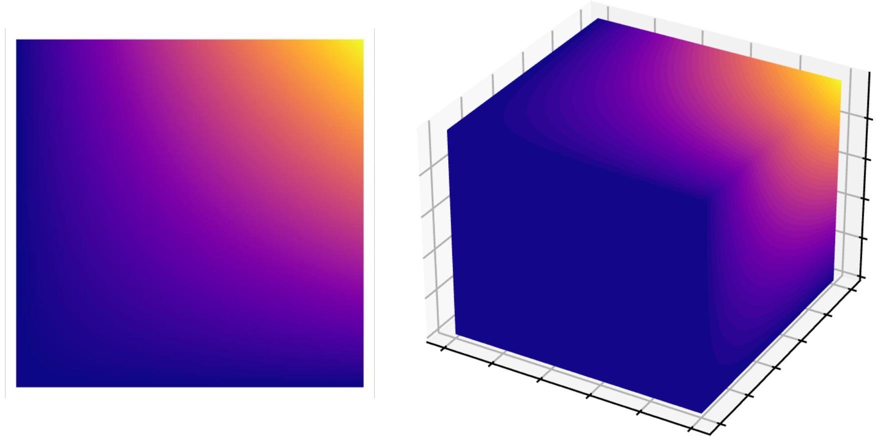


Complexon

$$\mathcal{W}\colon \bigsqcup_{d\geq 1} [0,1]^{d+1} \to [0,1]$$

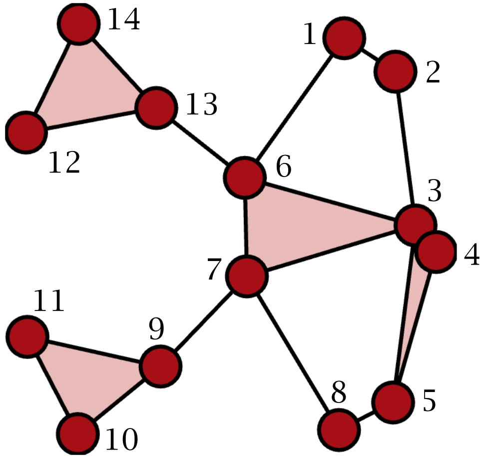
Dimension 1
Edge likelihoods

Dimension 2
Triangle likelihoods



Complexon

$$\mathcal{W}: \bigsqcup_{d \geq 1} [0,1]^{d+1} \rightarrow [0,1]$$

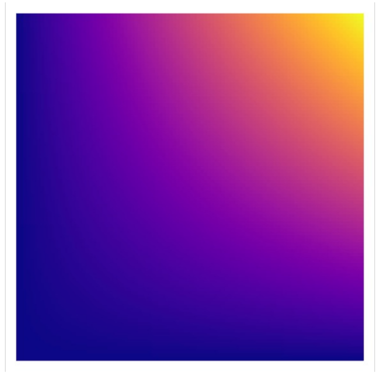


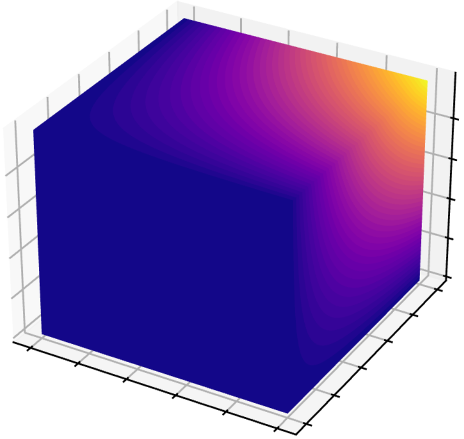
Sampled simplicial complex $K \sim \mathcal{W}$

Complexon as simplicial complex limit object

Dimension 1
Edge likelihoods

Dimension 2
Triangle likelihoods

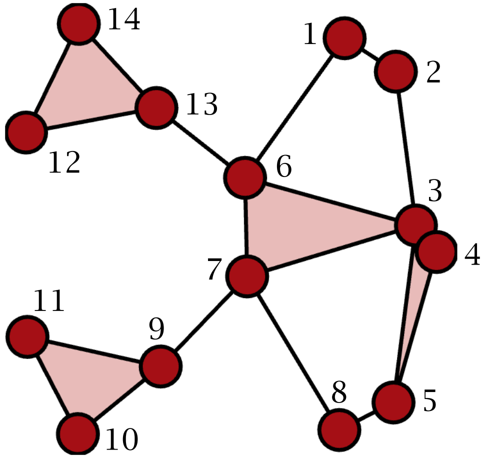




Complexon estimation

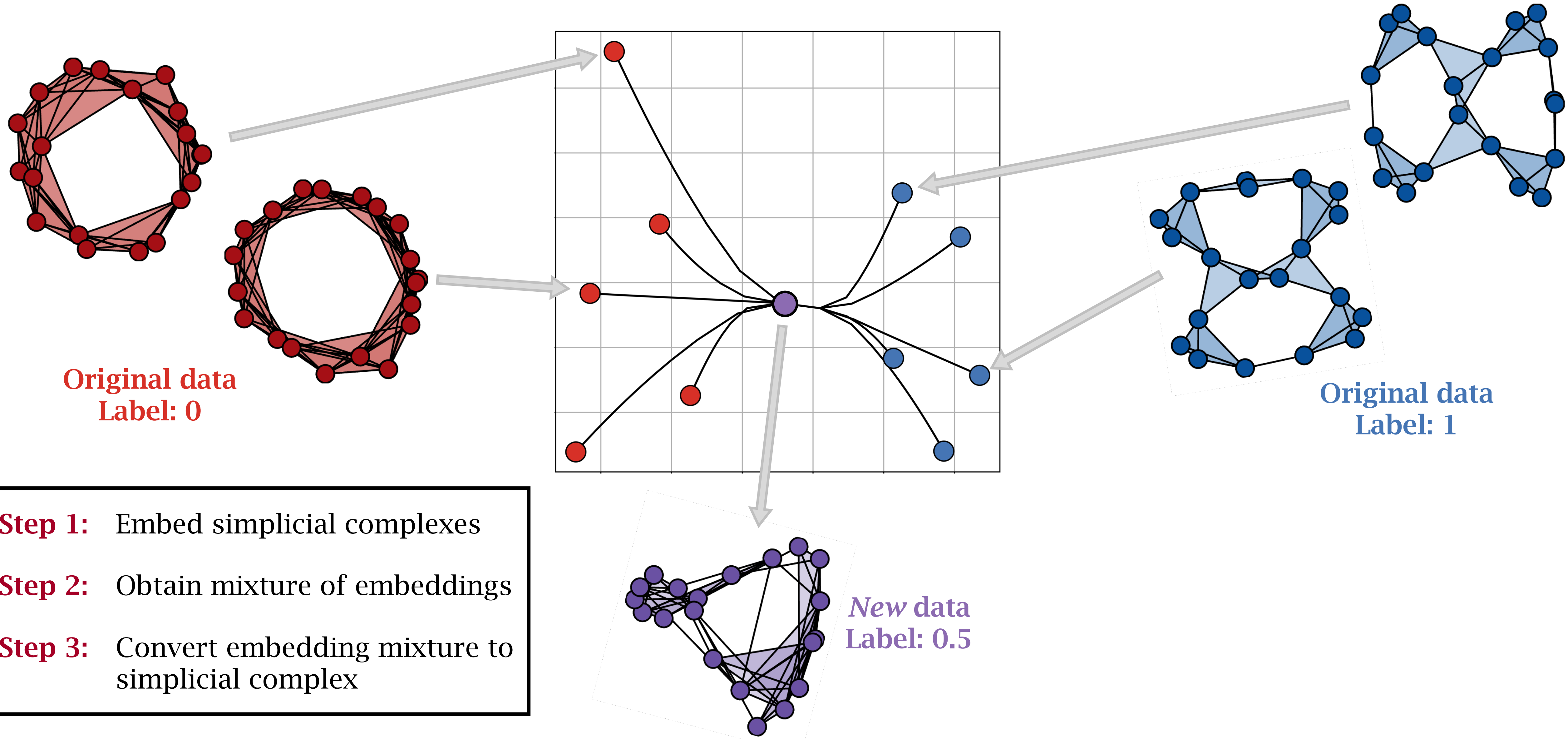
Estimated complexon

$$\widehat{\mathcal{W}}: \bigsqcup_{d\geq 1} [0,1]^{d+1} \to [0,1]$$



Simplicial complex
K

Simplicial Complex Mixup for Augmenting Data (SC-MAD)



Step 1: Embed simplicial complexes

Step 2: Obtain mixture of embeddings

Step 3: Convert embedding mixture to simplicial complex

GraphMAD improves performance and outperforms linear mixup on all datasets

Graph classification accuracy on social datasets

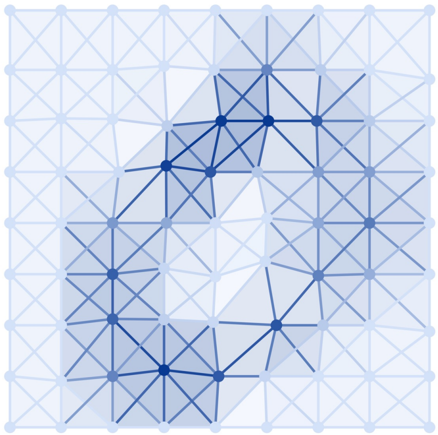
Method		COLLAB	IMDB-B	IMDB-M
Data mixup	**Label mixup**	**3 classes**	**2 classes**	**3 classes**
None	None	**80.00 ± 0.96**	73.14 ± 3.15	47.71 ± 4.25
Linear	Linear	77.60 ± 1.53	72.07 ± 2.06	47.24 ± 4.21
	Sigmoid	78.21 ± 1.16	**74.00 ± 2.14**	**49.67 ± 2.15**
	Logit	78.19 ± 1.61	72.64 ± 1.73	47.43 ± 2.45
	Cvx. Clust.	78.41 ± 0.99	71.43 ± 3.25	47.29 ± 5.21
Cvx. Clust.	Linear	78.93 ± 2.63	70.57 ± 4.89	45.52 ± 4.09
	Sigmoid	77.89 ± 1.30	**75.00 ± 5.13**	44.48 ± 2.78
	Logit	**80.39 ± 1.20**	73.43 ± 4.75	48.76 ± 2.43
	Cvx. Clust.	79.55 ± 2.29	71.43 ± 4.72	**49.71 ± 4.33**

Data augmentation with GraphMAD consistently outperforms linear mixup, and different label mixup functions can improve accuracy

SC-MAD for mixing complexons demonstrates consistent classification improvement

Simplicial complex classification accuracy on synthetic and real datasets

Method		Vietoris-Rips	MNIST
Data mixup	**Label mixup**	**2 classes**	**3 classes**
None	None	63.1 ± 1.67	78.2 ± 0.51
Linear	Linear	70.9 ± 0.51	80.2 ± 1.11
	Sigmoid	**71.9 ± 0.84**	68.7 ± 0.88
	Logit	59.4 ± 1.46	70.5 ± 0.33
	Cvx. Clust.	66.9 ± 1.93	80.5 ± 0.57
Cvx. Clust.	Linear	68.8 ± 1.96	80.4 ± 1.10
	Sigmoid	68.8 ± 1.56	**81.9 ± 0.072**
	Logit	70.9 ± 0.64	81.7 ± 0.49
	Cvx. Clust.	**73.8 ± 0.57**	**85.6 ± 0.52**



MNIST image 0

Both efficient linear mixup and informative convex clustering mixup improve classification performance

Class-dependent structure in complexon mixtures

Theorem For a set of simplicial complexes $\{(K_i, y_i)\}_{i=1}^{T}$ and their estimated complexons $\{\widehat{W}_i\}_{i=1}^{T}$, let $W_{\text{new}} = \sum_{i=1}^{T} \gamma_i \widehat{W}_i$ for $\sum_{i=1}^{T} \gamma_i = 1$ denote a complexon mixture.

Theorem For a set of simplicial complexes $\{(K_i, y_i)\}_{i=1}^{T}$ and their estimated complexons $\{\widehat{W}_i\}_{i=1}^{T}$, let $W_{\text{new}} = \sum_{i=1}^{T} \gamma_i \widehat{W}_i$ for $\sum_{i=1}^{T} \gamma_i = 1$ denote a complexon mixture.

Then, for the j-th estimate $\widehat{W}_j$, as $\gamma_j \to 1$ or $\widehat{W}_j \to \sum_{i \neq j} \frac{\gamma_i}{1-\gamma_j} \widehat{W}_i$,

$$\left| t(F, W_{\text{new}}) - t\left(F, \widehat{W}_j\right) \right| \to 0,$$

where F is any finite simplicial complex and $t(F, W)$ is the homomorphism density of F in W.

Theorem For a set of simplicial complexes $\{(K_i, y_i)\}_{i=1}^T$ and their estimated complexons $\{\widehat{W}_i\}_{i=1}^T$, let $W_{\text{new}} = \sum_{i=1}^T \gamma_i \widehat{W}_i$ for $\sum_{i=1}^T \gamma_i = 1$ denote a complexon mixture.

Then, for the j-th estimate $\widehat{W}_j$, as $\gamma_j \to 1$ or $\widehat{W}_j \to \sum_{i \neq j} \frac{\gamma_i}{1-\gamma_j} \widehat{W}_i$,

$$\left| t(F, W_{\text{new}}) - t\big(F, \widehat{W}_j\big) \right| \to 0,$$

where F is any finite simplicial complex and $t(F, W)$ is the homomorphism density of F in W.

Class-discriminative structure is present in complexon mixtures

Theorem For a set of simplicial complexes $\{(K_i, y_i)\}_{i=1}^T$ and their estimated complexons $\{\widehat{W}_i\}_{i=1}^T$, let $W_{\text{new}} = \sum_{i=1}^T \gamma_i \widehat{W}_i$ for $\sum_{i=1}^T \gamma_i = 1$ denote a complexon mixture.

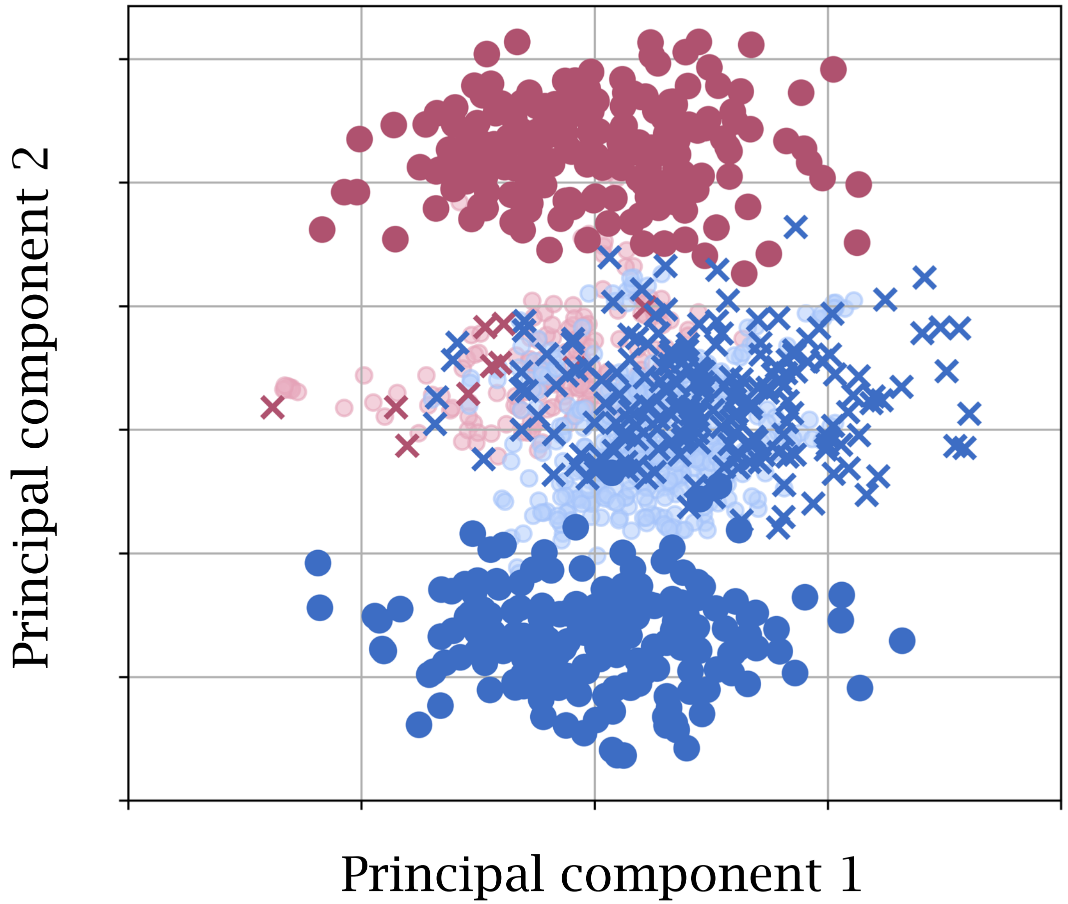
Then, for the j-th estimate $\widehat{W}_j$, as $\gamma_j \to 1$ or $\widehat{W}_j \to \sum_{i \neq j} \frac{\gamma_i}{1-\gamma_j} \widehat{W}_i$,

$$\left| t(F, W_{\text{new}}) - t\left(F, \widehat{W}_j\right) \right| \to 0,$$

where F is any finite simplicial complex and $t(F, W)$ is the homomorphism density of F in W.

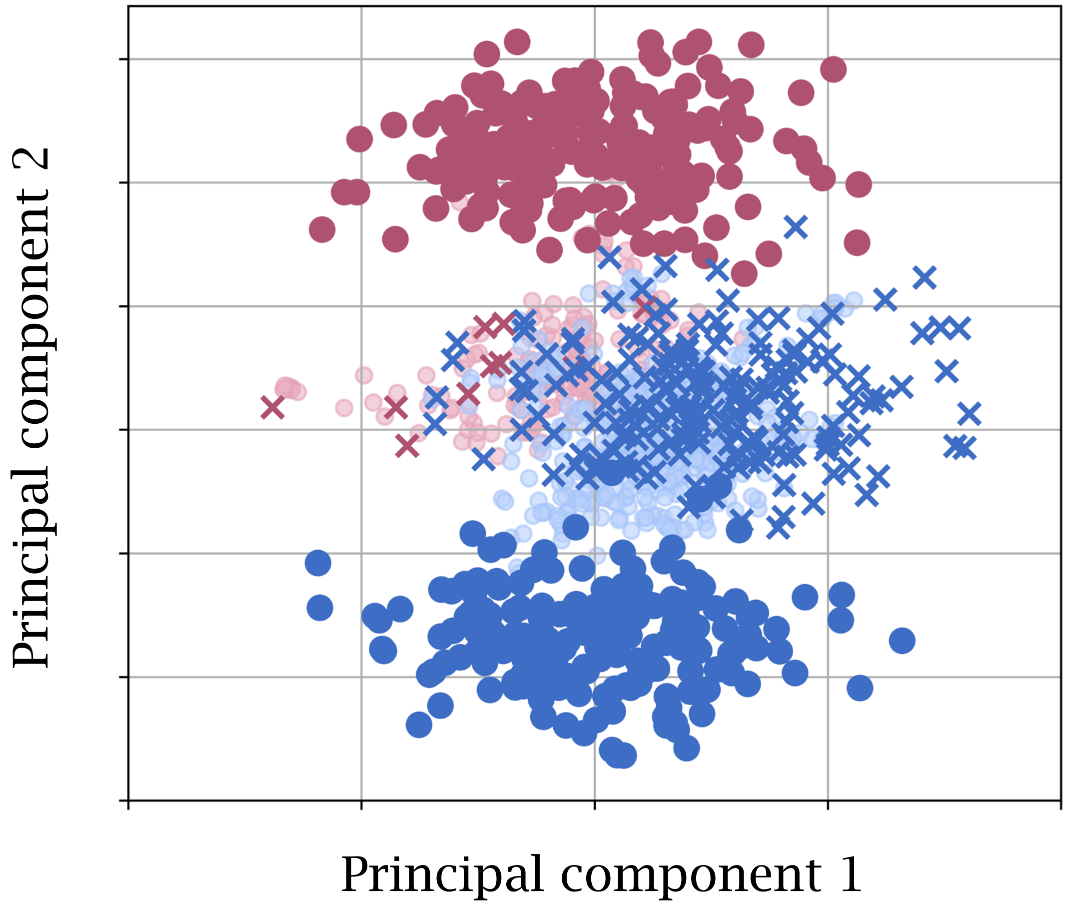
Class-discriminative structure is present in complexon mixtures

Mixup preserves class information when interpolating between classes



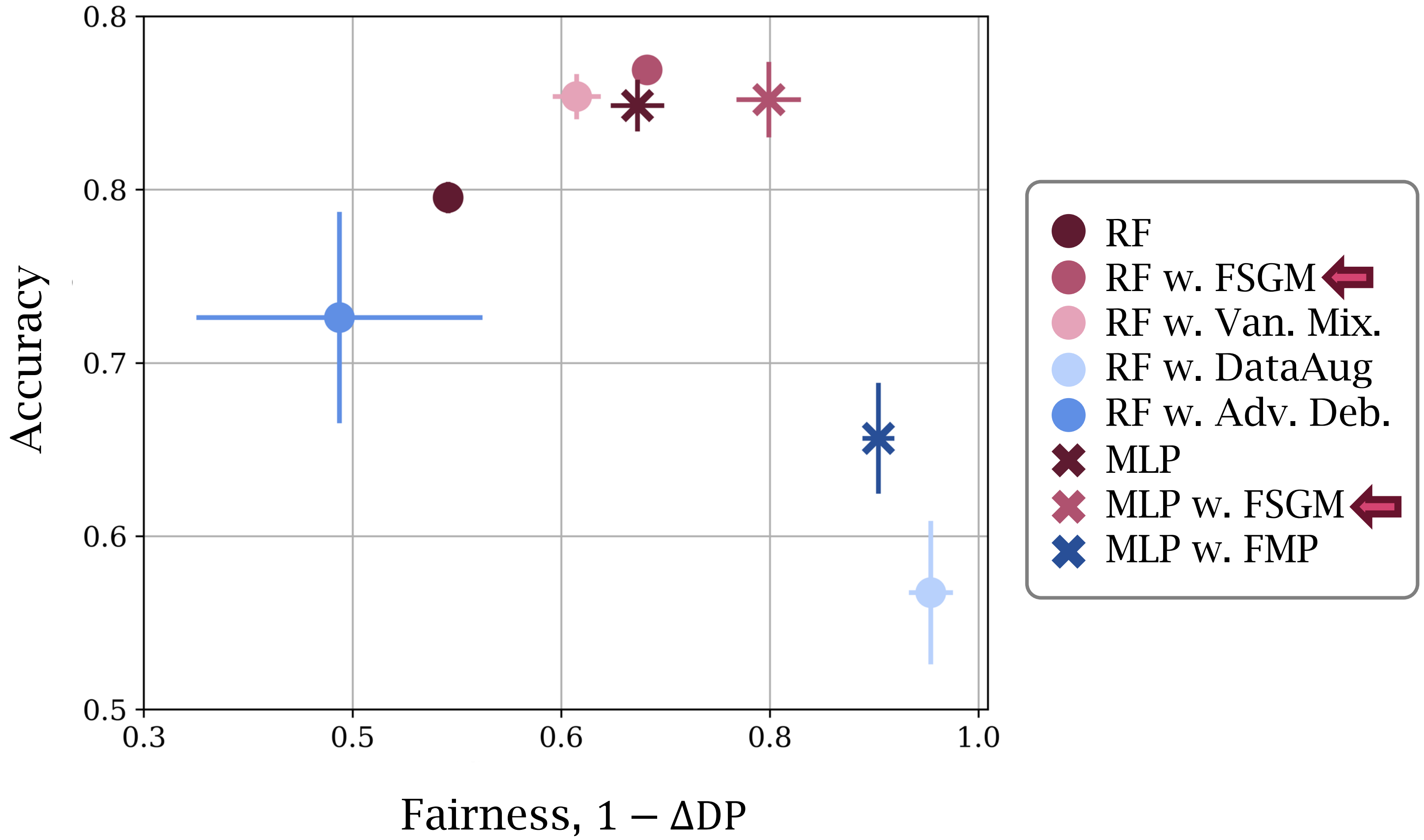
Underrepresented subgroup: Minority subgroup sensitive to unfair distribution shifts

Heavily underrepresented subgroup as underrepresented class and group



Unbalanced groups and classes
with distribution shift that contributes bias

Heavily underrepresented subgroup as underrepresented class and group



Fair SubGroup Mixup (FSGM) improves fairness while maintaining or improving accuracy