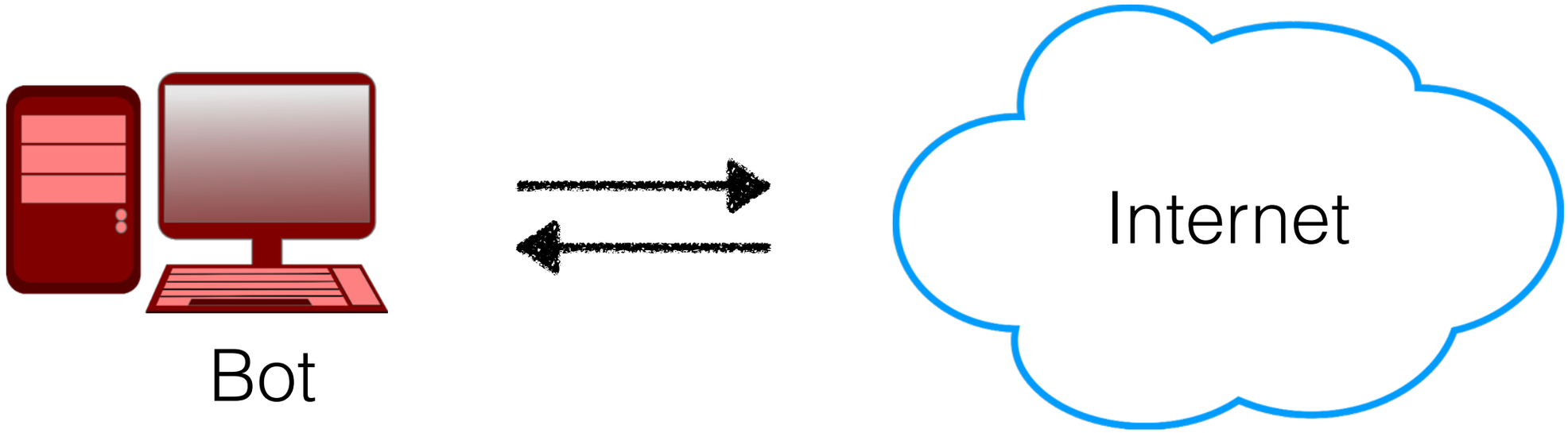
Conformal Clustering and its Application to Botnet Traffic

Giovanni Cherubin, Ilia Nouretdinov, Alexander Gammerman
Roberto Jordaney, Zhi Wang, Davide Papini, Lorenzo Cavallaro

Netflow, network traces

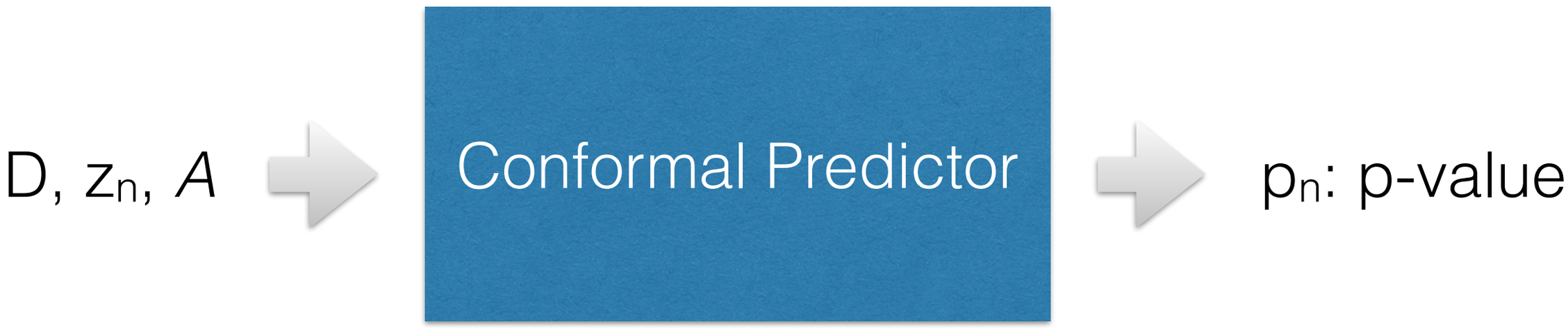


netflow	Date	Duration	IP_src	Port_src	IP_dst	Port_dst	TCP/ UDP
	Sent Packets	Recv Packets	Sent Bytes	Recv Bytes	Tot Packets	Tot Bytes	Flags…

Netflow, network traces

	Date	Duration	TCP/ UDP	Sent Bytes	Port_dst	…
netflow_1	1248089563	2939	TCP	503	445	
netflow_2	1248089702	51	TCP	354	139	
…						

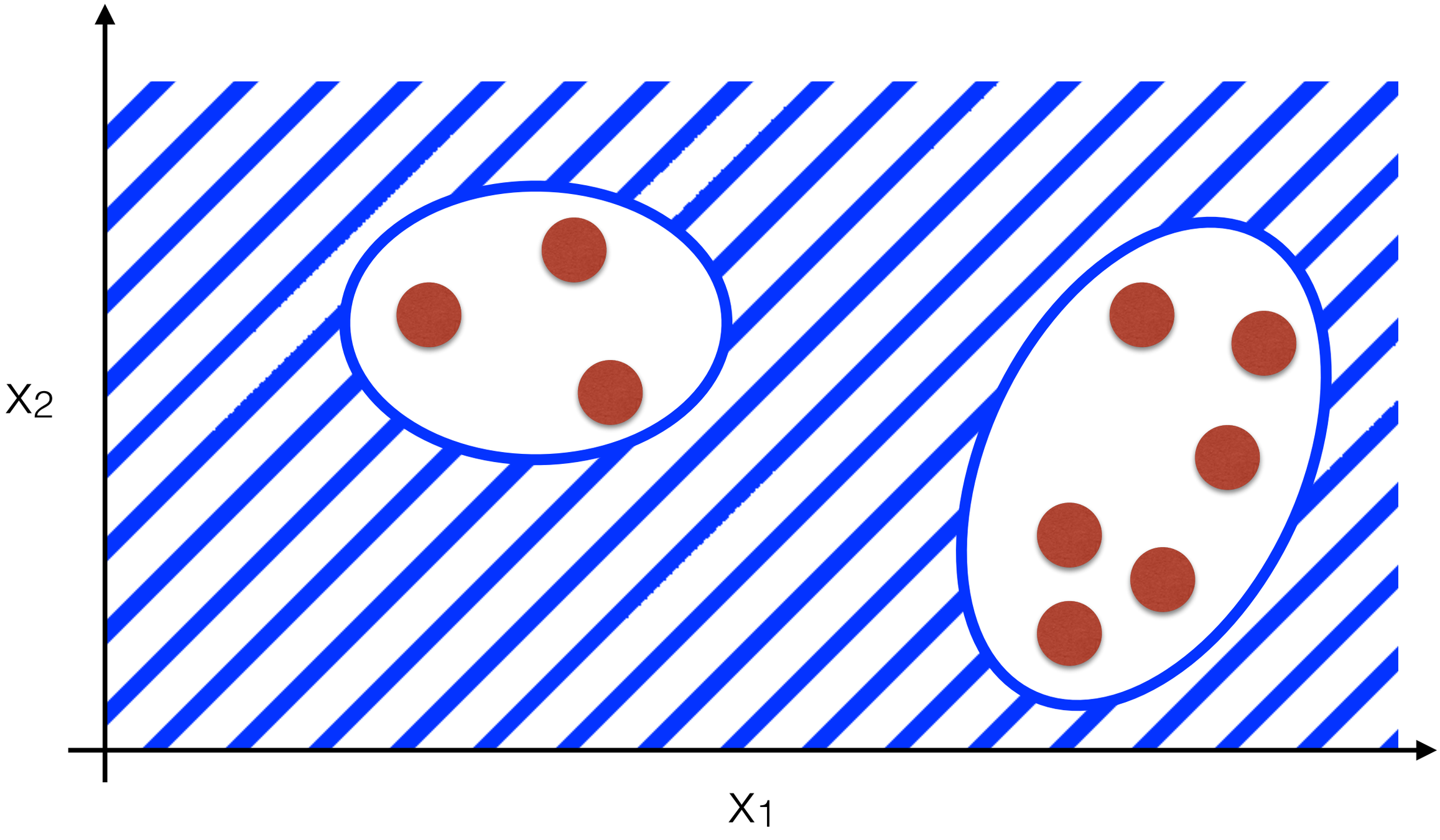
Conformal Predictor



Does z_n conform D for 1-ε confidence?

CP for anomaly detection

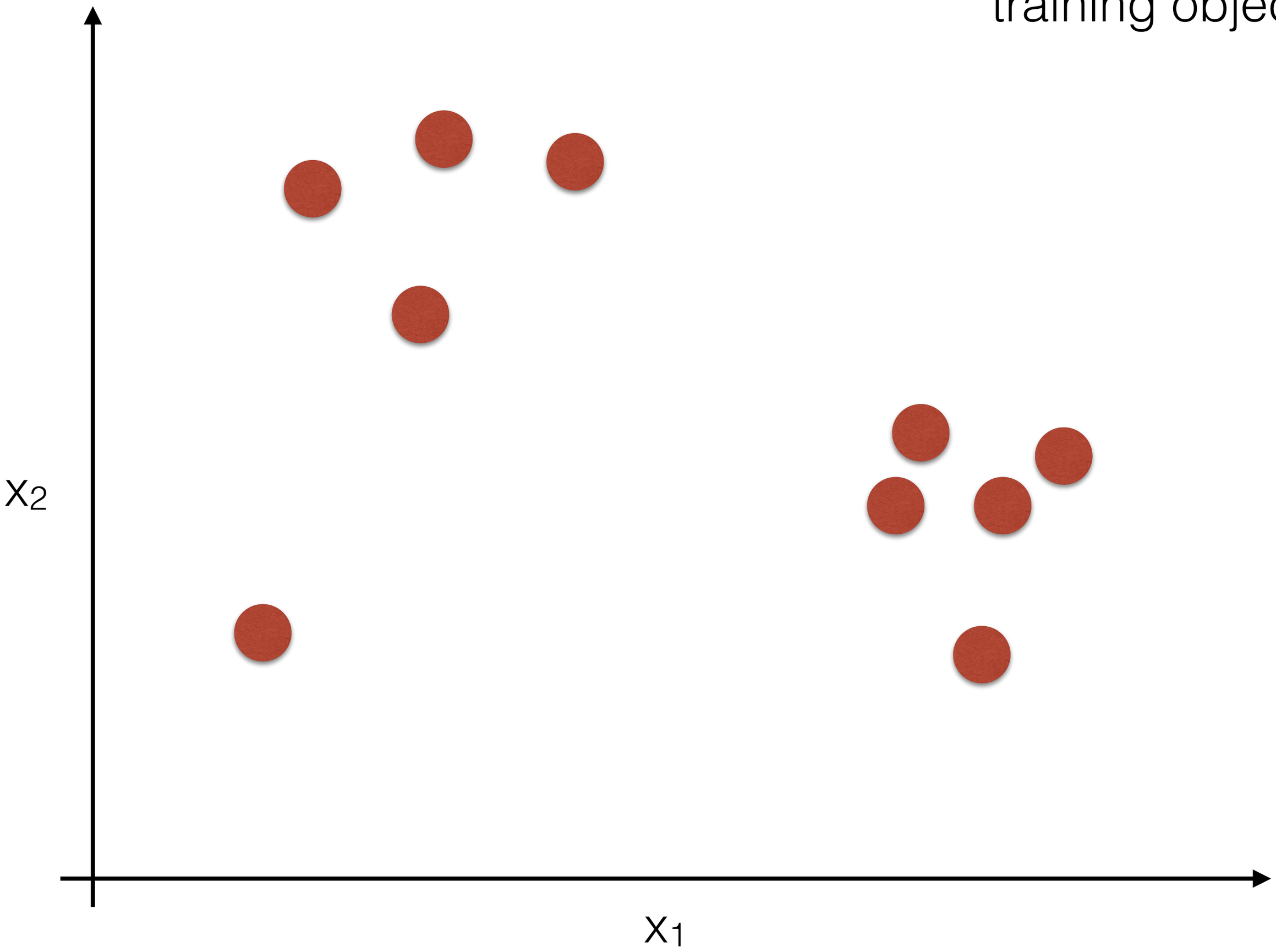
[Laxhammar11, Smith14]



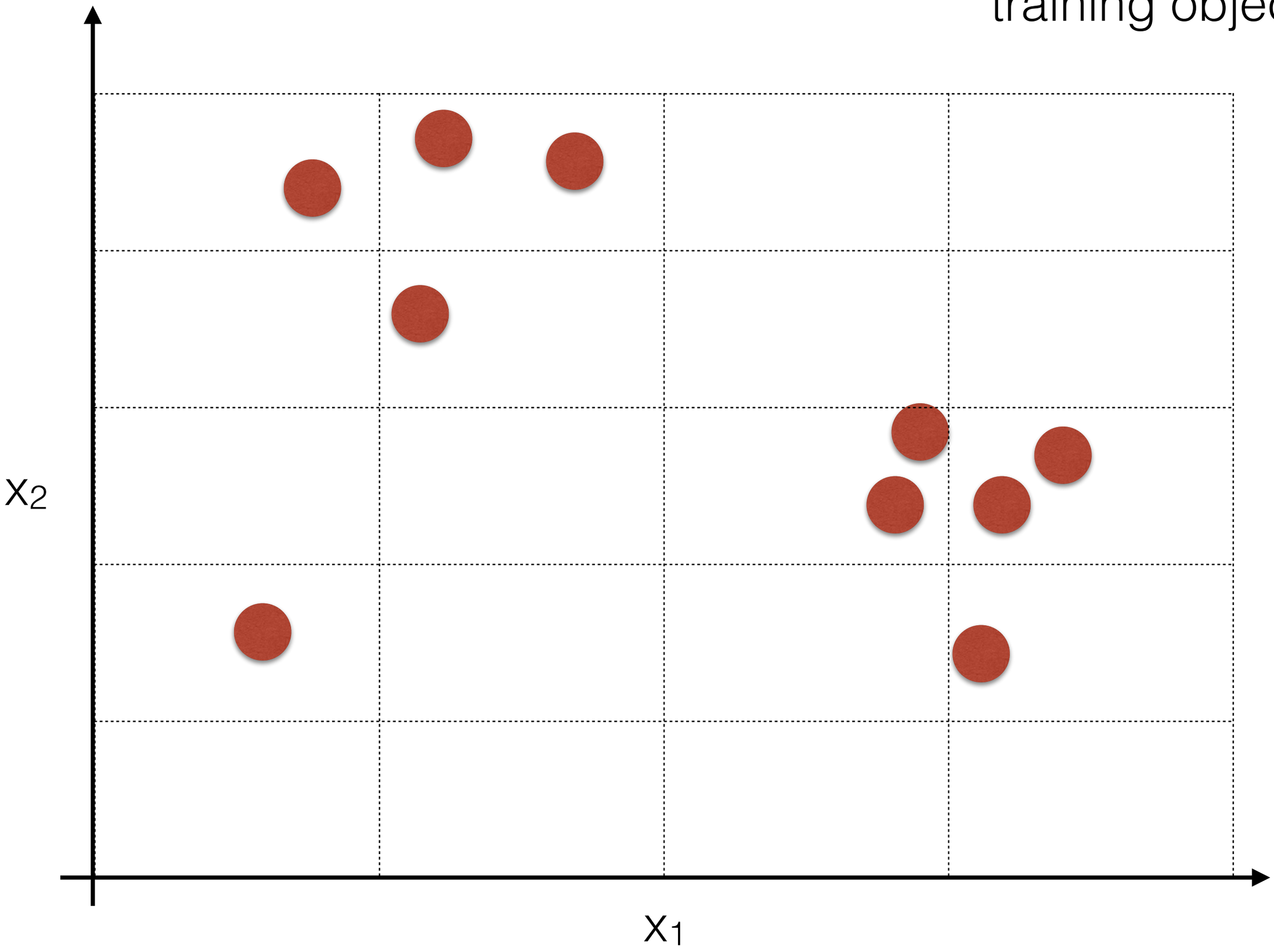
Conformal Clustering

- Conformal Predictors in unsupervised setting.
- Controls the objects left outside the clusters.
- Regulates the “depth” of clusters.

training objects



training objects



p-values grid

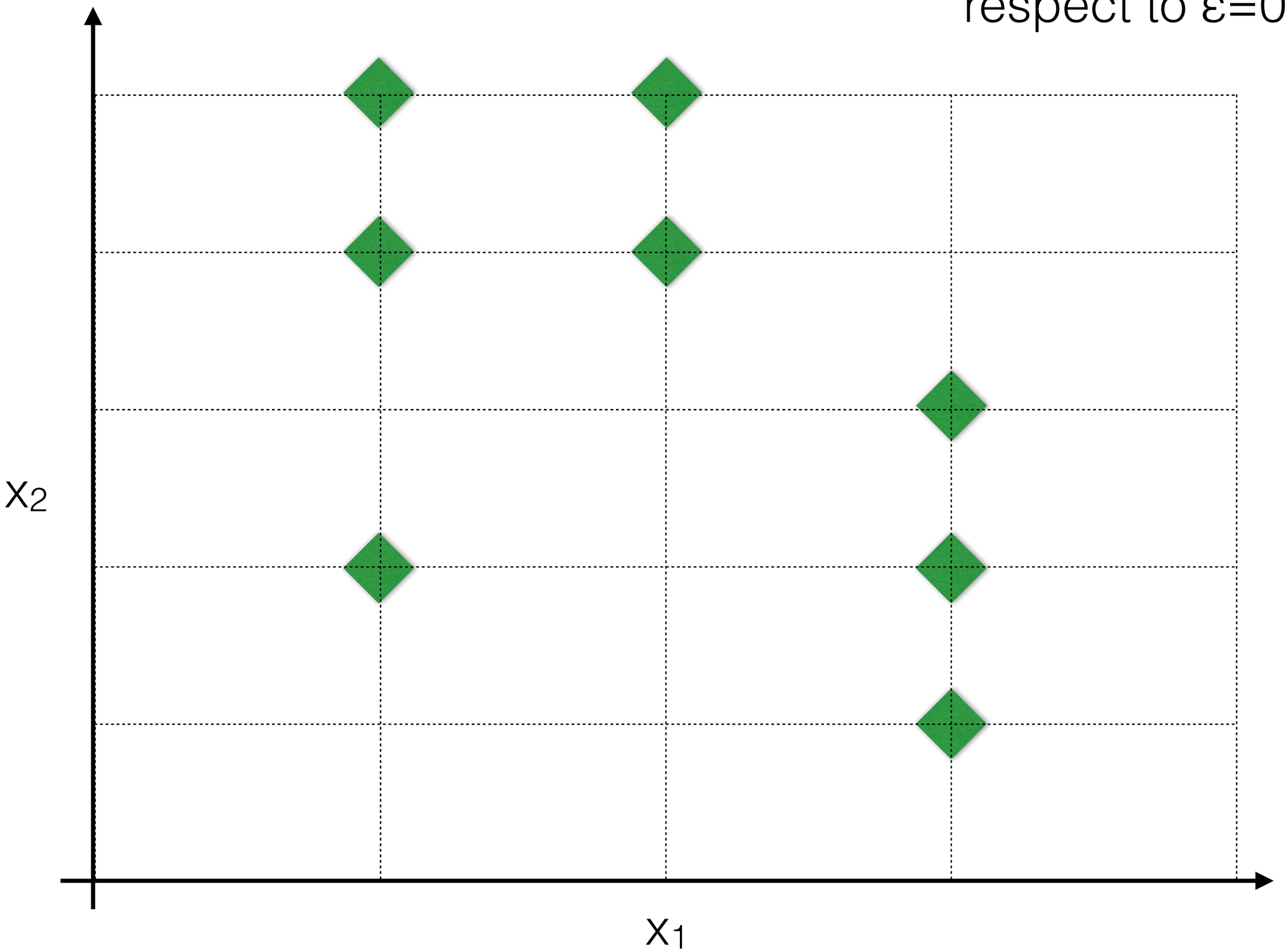
0.1 0.3 0.2 0.1 0.0

0.1 0.3 …

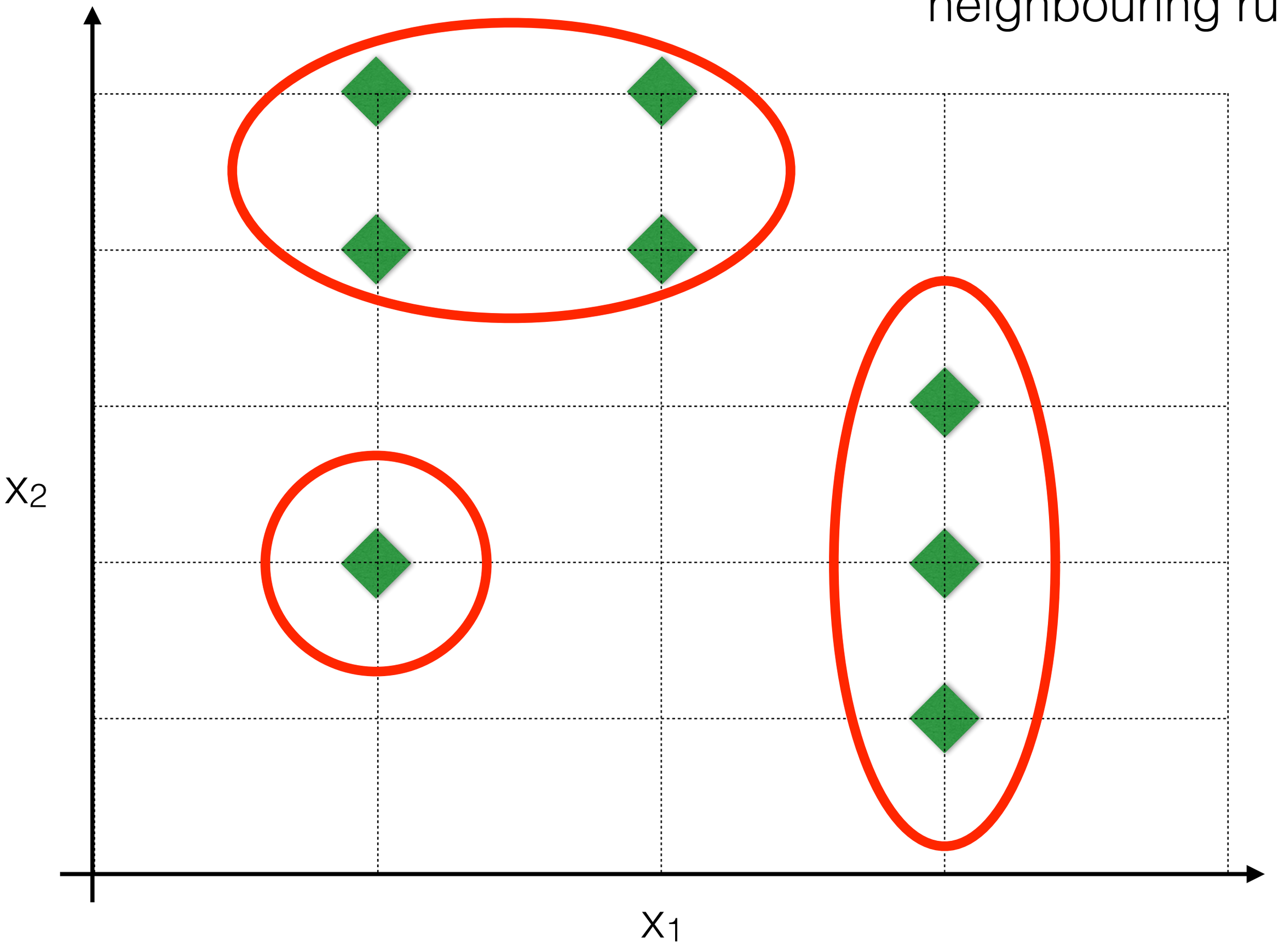
X_2

X_1

respect to ε=0.1



neighbouring rule

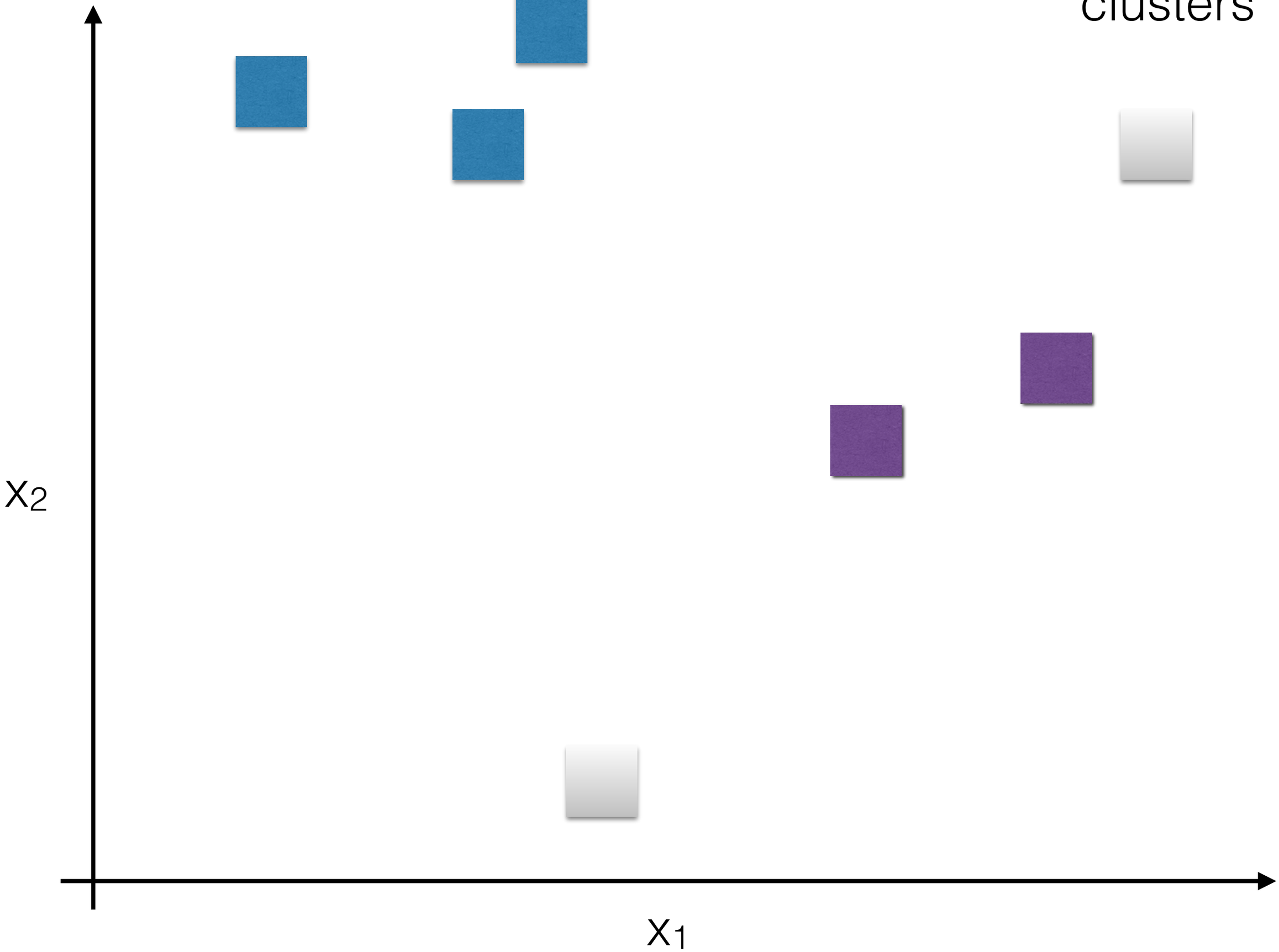


test set

x_2

x_1

clusters



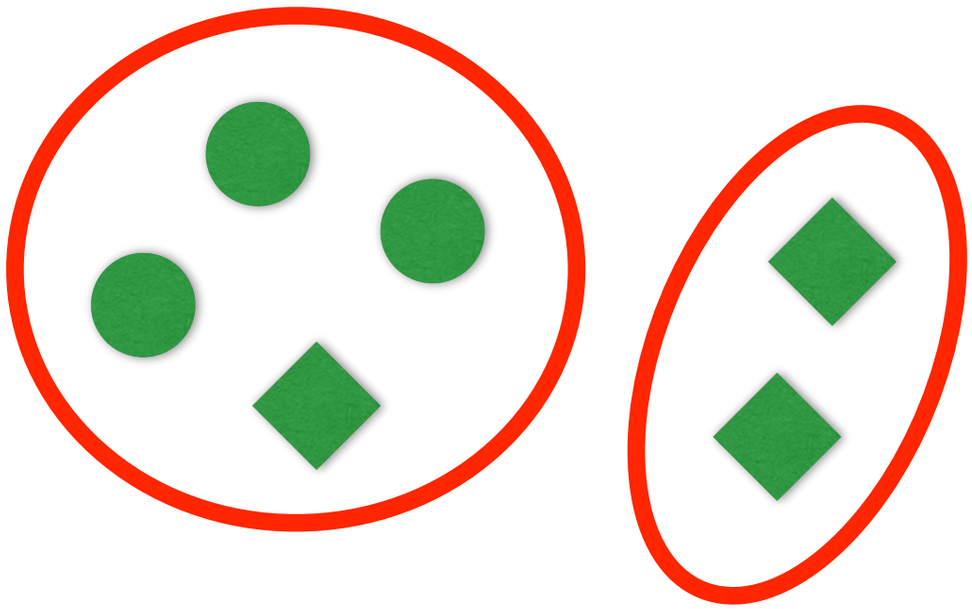
Our Approach

- Each network trace produces a feature vector.
- Normalisation.
- Dimensionality reduction (*t-SNE*).
- Non-conformity measures: *k-NN*, *KDE*.
- Performance measures: *Purity*, *Average P-Value.*

Performance Measures

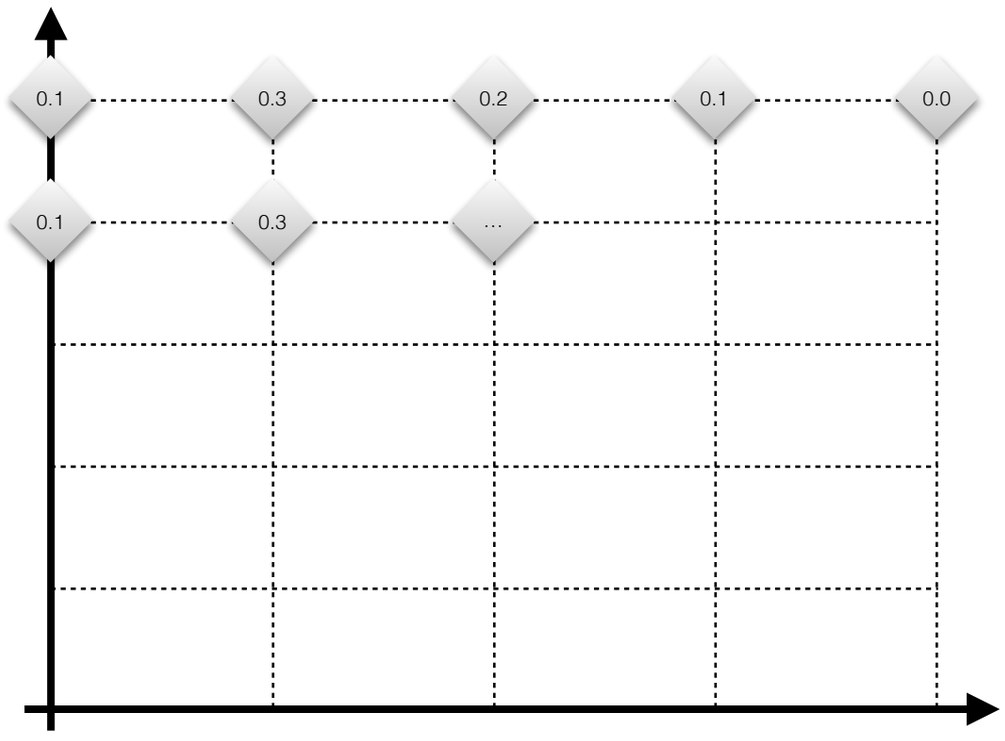
Purity

- How “pure” are the clusters.
- For the same ε the number of clusters is not influenced.



Average P-Value

- *Efficiency* criterion.
- Size of the prediction set.
- The smaller the prediction set the better.



Results (ε=0.2)

k-NN non-conformity measure

k	1	2	3	4	5	…	10
APV	**0.129**	0.139	0.141	0.147	0.160		0.193
Purity	**0.99**	0.97	0.97	0.96	0.96		0.92

KDE (Gaussian kernel) non-conformity measure

h	0.001	0.005	0.01	0.05	0.1	…	1.0
APV	0.404	0.332	0.299	0.165	**0.130**		0.221
Purity	**1.00**	0.98	**1.00**	0.99	0.99		0.92

Future work

- Avoid dimensionality reduction, reduce complexity.
- New criteria of accuracy.
- New non-conformity measures based on previous work in botnets detection (e.g.: BotFinder).
- Detection: “malicious” and “benign” data.

Bibliography

- [Vovk05] V. Vovk et al., *Algorithmic learning in a random world.* Springer, 2005.

- [Maaten08] L. van der Maaten et al., *Visualizing data using t-SNE.* Journal of Machine Learning Research, 2008.

- [Laxhammar11] R. Laxhammar et al., *Sequential conformal anomaly detection in trajectories based on hausdorff distance,* 2011.

- [Lei13] J. Lei et al., *A conformal prediction approach to explore functional data,* 2013.

- [Smith14] J. Smith et al., *Anomaly Detection of Trajectories with Kernel Density Estimation by Conformal Prediction.* Artificial Intelligence Applications and Innovations, Springer, 2014.

Thanks

Conformal Clustering and its Application to Botnet Traffic

Giovanni Cherubin, Ilia Nouretdinov, Alexander Gammerman
Roberto Jordaney, Zhi Wang, Davide Papini, Lorenzo Cavallaro