Topology and data reduction

Javier Perera-Lago

University of Seville

17th January, 2025

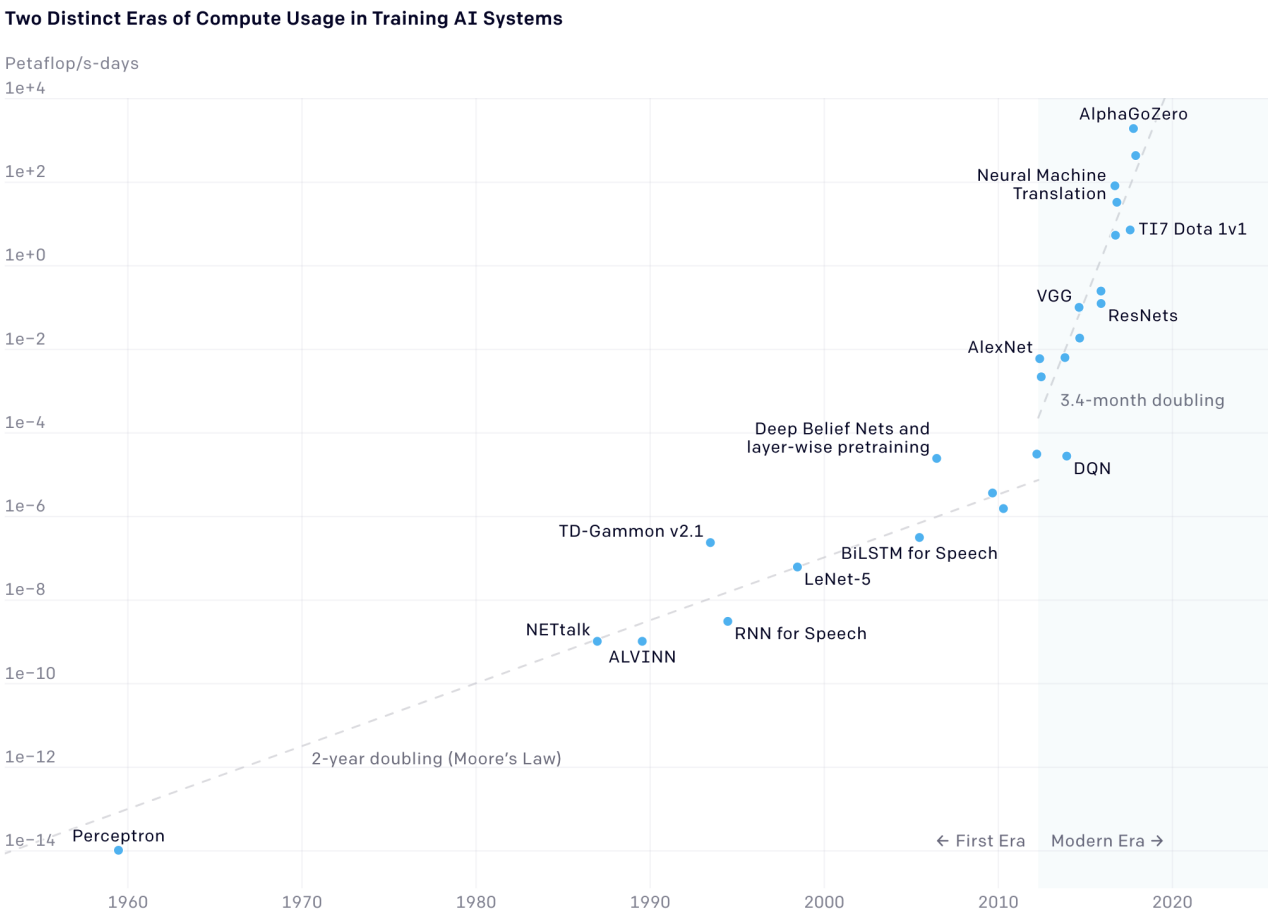
Artificial Intelligence: the training problem

Artificial Intelligence usually relies on Machine Learning models.

These models depend on a set of parameters that need to be adjusted. The setting or *learning* of the parameters requires a lot of real-world data.

Nowadays, we have more and more sophisticated models and more massive data sets. Because of this, the costs derived from developing new AI are growing continually.

Increasing computations in AI



Increasing computations in AI

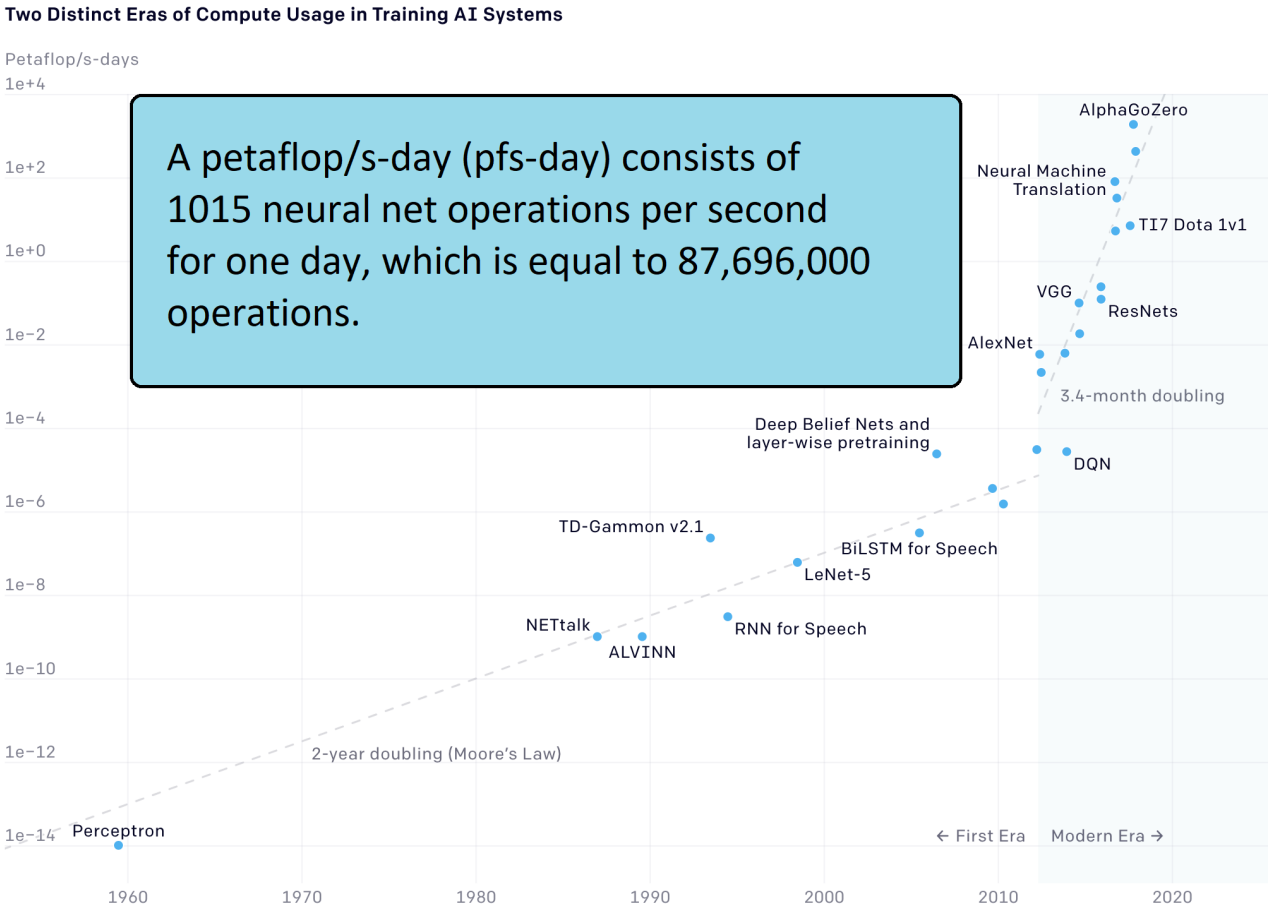


*Chart taken from the OpenAI blog: AI and compute

Increasing computations in Language Processing

Increasement of datasets and models for Natural Language Processing models:

Model	Year	Dataset Size	Number of parameters
BERT-Large	2018	13 GB	350 M
GPT-2-XL	2019	40 GB	1.5 B
ROBERTA	2019	160 GB	125 M
XLNet-Large	2020	158 GB	340 M
T5-11B	2020	750 GB	11 B

Data from: *Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green ai. Communications of the ACM, 63(12), 54-63.*

Red AI vs Green AI

- Red AI: AI research that seeks to improve the performance of models through the use of massive computational power without taking costs into account.

- Green AI: AI research that, in addition to seeking good results, seeks to reduce the consumption of resources.

Green AI: 4 approaches

According to the literature, there are four main ways to reduce the costs in Machine Learning:

- Compact Architecture Design
- Energy-efficient Training Strategies
- Energy-efficient Inference
- Efficient Data Usage

Green AI: 4 approaches

According to the literature, there are four main ways to reduce the costs in Machine Learning:

- Compact Architecture Design
- Energy-efficient Training Strategies
- Energy-efficient Inference
- Efficient Data Usage $\longleftarrow$ We will focus on this approach

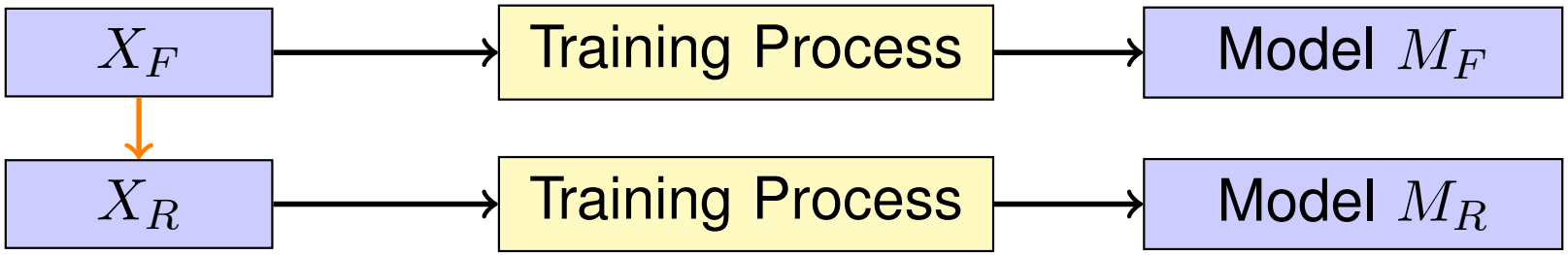
Efficient data usage: Data Reduction

We want to reduce the size of the dataset, trying that the reduced dataset gives us a good representation of the full dataset.

$$\text{Properties}(X_F) \approx \text{Properties}(X_R)$$

Efficient data usage: Data Reduction

The idea is to use the reduced dataset for model training instead of the full dataset, making the process less expensive and giving similar results.



Properties(X_F) $\approx$ Properties(X_R) $\Rightarrow$ Model $M_F \approx$ ModelM_R

Ways to reduce a dataset

There are two main ways of reducing the size of a dataset:

- **Reducing feature size**: eliminating irrelevant or redundant features diminishes the dataset size and mitigates the risk of overfitting.

$$X_{N\times D} \longrightarrow Y_{N\times d}\ (d << D)$$

- **Reducing sample size**: discarding redundant or noisy examples and alleviating imbalances between classes can improve the training process.

$$X_{N\times D} \longrightarrow Z_{n\times D}\ (n << N)$$

Dimensionality Reduction

There are many different methods to reduce the dimensionality of a dataset using topological information:

- Dimensionality reduction via PH optimization
- Topological Autoencoders
- UMAP (Uniform Manifold Approximation and Projection)
- FibeRed (Fiberwise dimensionality reduction)
- Topologically controlled lossy compression

Dimensionality Reduction

There are many different methods to reduce the dimensionality of a dataset using topological information:

- Dimensionality reduction via PH optimization
- Topological Autoencoders
- UMAP (Uniform Manifold Approximation and Projection)
- FibeRed (Fiberwise dimensionality reduction)
- Topologically controlled lossy compression

But we will focus on size reduction methods today.

Size Reduction: PH Landmarks

This is an instance selection method that ranks all the items in the dataset using a topology-based score.

It computes for every $x \in X$ the similarity between $PH_n(X)$ and $PH_n(X \setminus x)$ to see how informative it is for the whole dataset.

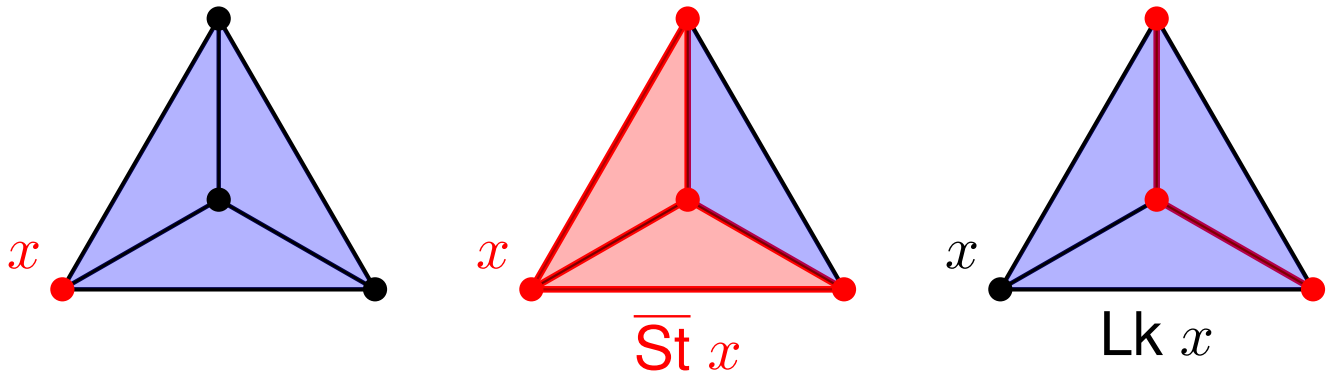
If $PH_n(X) \approx PH_n(X \setminus x)$, the point x can be discarded from the dataset without affecting the topology of the whole dataset.

Stolz, B. J. (2023).

Outlier-robust subsampling techniques for persistent homology. Journal of Machine Learning Research, 24(90), 1-35.

Size Reduction: PH Landmarks

Consider a simplicial complex X and a vertex $x \in V(X)$. We define the closed star and link of x as:



Note that:

1. $\overline{\text{St}}\ x$ is contractible.
2. $X = (X \setminus x) \cup \overline{\text{St}}\ x$
3. $\text{Lk}\ x = (X \setminus x) \cap \overline{\text{St}}\ x$

Size Reduction: PH Landmarks

Given a simplicial complex X and two subcomplexes $A, B \subset X$ such that $A \cup B = X$, we have the Mayer-Vietoris sequence:

$$\to H_n(A \cap B) \to H_n(A) \oplus H_n(B) \to H_n(X) \to H_{n-1}(A \cap B) \to$$

Considering $A = (X \setminus x)$ and $B = \overline{\mathsf{St}}\, x$, this translates into:

$$\to H_n(\mathsf{Lk}\, x) \to H_n(X \setminus x) \oplus H_n(\overline{\mathsf{St}}\, x) \to H_n(X) \to H_{n-1}(\mathsf{Lk}\, x) \to$$

Size Reduction: PH Landmarks

Since $\overline{\text{St}}\ x$ is contractible, $H_n(\overline{\text{St}}\ x) = 0\ \forall n$, and then:

$$\to H_n(\text{Lk}\ x) \to H_n(X \setminus x) \to H_n(X) \to H_{n-1}(\text{Lk}\ x) \to$$

If we assume that $H_n(\text{Lk}\ x) = H_{n-1}(\text{Lk}\ x) = 0$, then we have for each $n > 0$ the short exact sequence:

$$0 \to H_n(X \setminus x) \to H_n(X) \to 0$$

implying that $H_n(X \setminus x) \cong H_n(X)$.

Size Reduction: PH Landmarks

In practice, we do not work with homology but with persistent homology.

To make the computation easier, we restrict ourselves to a δ-neighborhood of x, getting $\mathrm{Lk}^{\delta}\ x$ and $\overline{\mathrm{St}}^{\delta}\ x$ instead of $\mathrm{Lk}\ x$ and $\overline{\mathrm{St}}\ x$.

Following the previous reasoning, we can argue that

$$PH_n(\mathrm{Lk}^{\delta}\ x) = 0 \Rightarrow PH_n(X) \cong PH_n(X \setminus x)$$

We will measure the similarity between $PH_n(X)$ and $PH_n(X \setminus x)$ by measuring how similar is $PH_n(\mathrm{Lk}^{\delta}\ x)$ to 0.

Size Reduction: PH Landmarks

If the barcode of $PH_n(\mathsf{Lk}^\delta\ x)$ is $B_n(\mathsf{Lk}\ x) = \{[b_i, d_i)\}_i$, we can define:

$$|B_n(\mathsf{Lk}^\delta\ x)| = \max_i(d_i - b_i)$$

We define the *PH outlierness* of x as:

$$\mathsf{out}_{PH}(x) = \max\{|B_0(\mathsf{Lk}^\delta\ x)|, |B_1(\mathsf{Lk}^\delta\ x)|, |B_2(\mathsf{Lk}^\delta\ x)|\}$$

This is the score that we use to select the most representative samples in X.

ε-representativeness

We ask ourselves:

How can we measure if a reduced dataset gives a good representation of the full dataset?

ε-representativeness

We ask ourselves:

How can we measure if a reduced dataset gives a good representation of the full dataset?

We will use the concept of ε**-representativeness**, which uses pairwise distances to measure the similarity between the full dataset and a reduced version of it.

Gonzalez-Diaz, R., Gutiérrez-Naranjo, M. A., & Paluzo-Hidalgo, E. (2022). Topology-based representative datasets to reduce neural network training resources. Neural Computing and Applications, 34(17), 14397-14413.

ε-representativeness

Let's assume we are trying to solve a classification task, and our dataset $\mathcal{D}$ is defined:

$$\mathcal{D} = \{(x, c_x) | x \in X \subset \mathbb{R}^n, c_x \in [[0, k]]\}$$

where $[[0, k]] = \{0, 1, 2, \cdots, k\}$. For each point $x \in X$, there is a label c_x that tells us its class. Each point belongs to one and only one class.

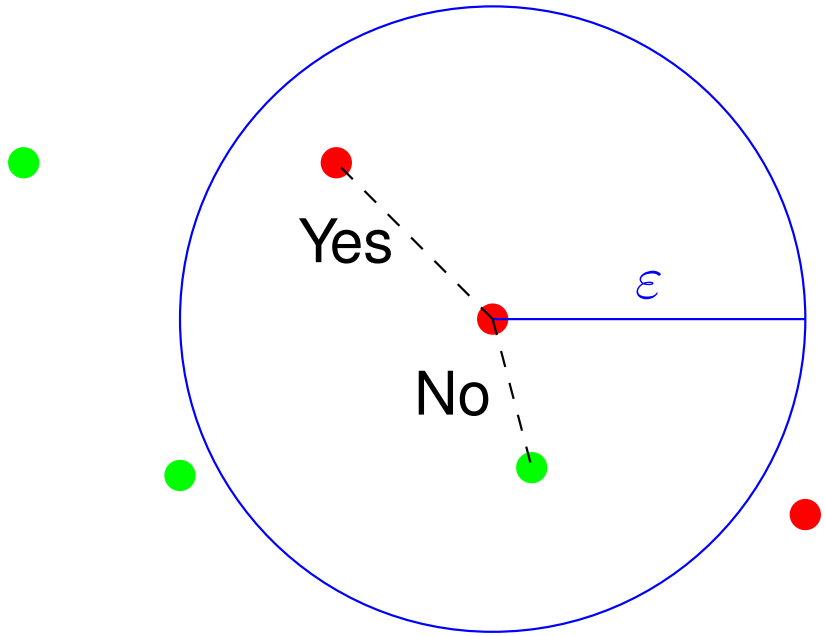
ε-representativeness

Definition: ε-representative point

Given a real number $\varepsilon > 0$ which we call the representation error, a labelled point (x, c_x) is ε-representative of $(\tilde{x}, c_{\tilde{x}})$ if $c_x = c_{\tilde{x}}$ and $||x - \tilde{x}|| \leq \varepsilon$. We denote $x \approx_\varepsilon \tilde{x}$.

ε-representativeness

Example of ε-representative points.



ε-representativeness

We extend ε-representativeness between pair of points to define the ε-representativeness between datasets:

Definition: ε-representative dataset

A dataset $\tilde{\mathcal{D}} = \{(\tilde{x}, c_{\tilde{x}}) | \tilde{x} \in \tilde{X} \subset \mathbb{R}^n, c_{\tilde{x}} \in [[0, k]]\}$ is ε-representative of $\mathcal{D} = \{(x, c_x) | x \in X \subset \mathbb{R}^n, c_x \in [[0, k]]\}$ if there exists an isometric transformation $f : \tilde{X} \to \mathbb{R}^n$, such that for any $(x, c_x) \in \mathcal{D}$ there exists $(\tilde{x}, c_{\tilde{x}}) \in \tilde{\mathcal{D}}$ satisfying that $f(\tilde{x}) \approx_\varepsilon x$.

ε-representativeness

ε-representative datasets preserve persistent homology:

ε-representativeness

ε-representative datasets preserve persistent homology:

Theorem 1 [1]

If the dataset $\tilde{\mathcal{D}}$ is ε-representative of $\mathcal{D}$, then

$$d_B(\mathrm{Dgm}_q(X), \mathrm{Dgm}_q(\tilde{X})) \leq 2\varepsilon$$

where $q \leq n$, $\mathrm{Dgm}_q(X)$ and $\mathrm{Dgm}_q(\tilde{X})$ are the persistence diagrams of the Vietoris-Rips filtrations computed from X and $\tilde{X}$, and d_B denotes the bottleneck distance between their persistence diagrams.

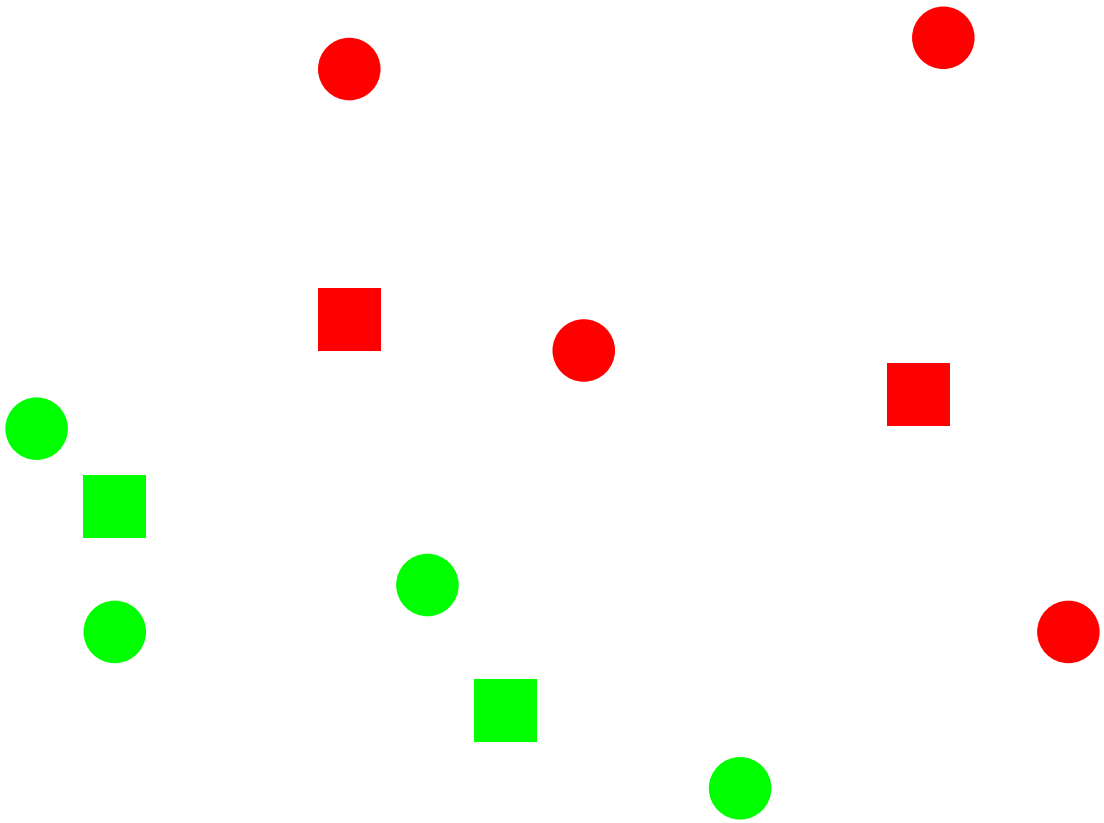
ε-representativeness

Given a dataset $\mathcal{D}$, a reduction $\mathcal{D}_R$ and an isometry $i : \mathcal{D}_R \to \mathbb{R}^d$, the minimum ε such that $\mathcal{D}_R$ is ε-representative dataset of $\mathcal{D}$ is:

$$\varepsilon^* = \max_{k=1,\ldots,c} \max_{x:c_x=k} \min_{x':c_{x'}=k} ||x - i(x')||$$

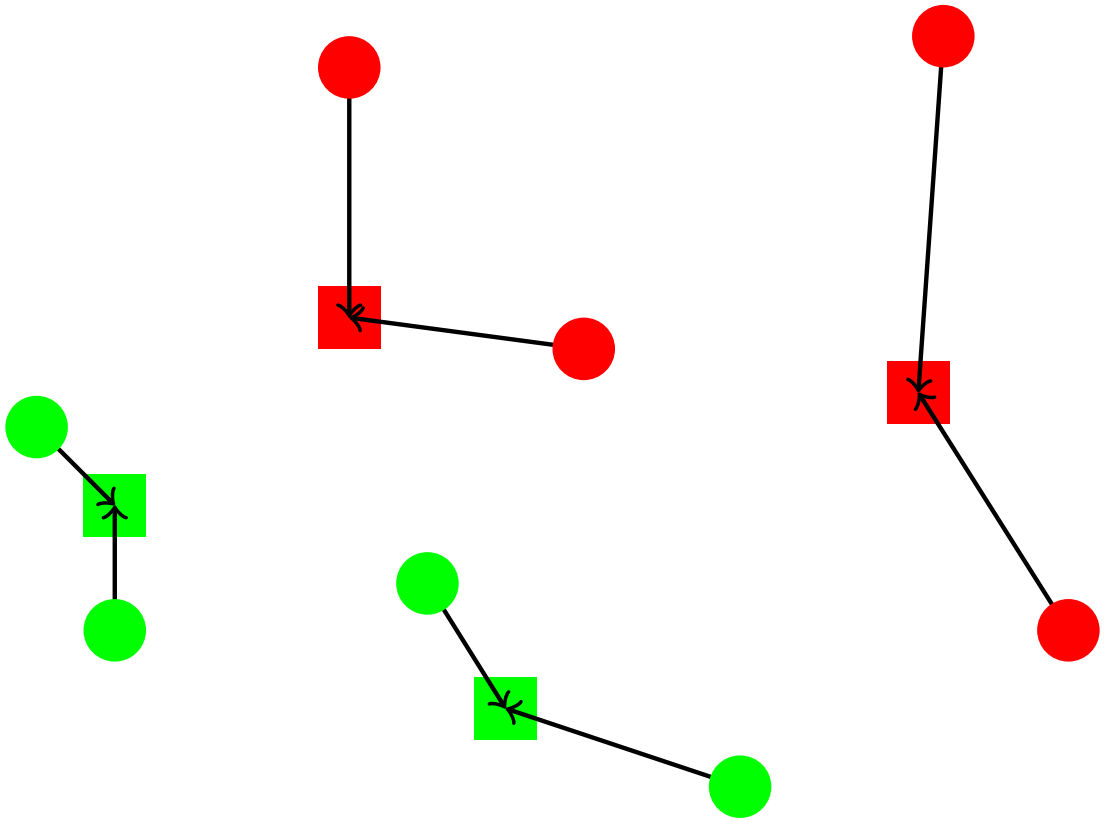
ε-representativeness

Let's consider this dataset with 2 classes. The square points form the reduced dataset. Which is its ε-representativeness?



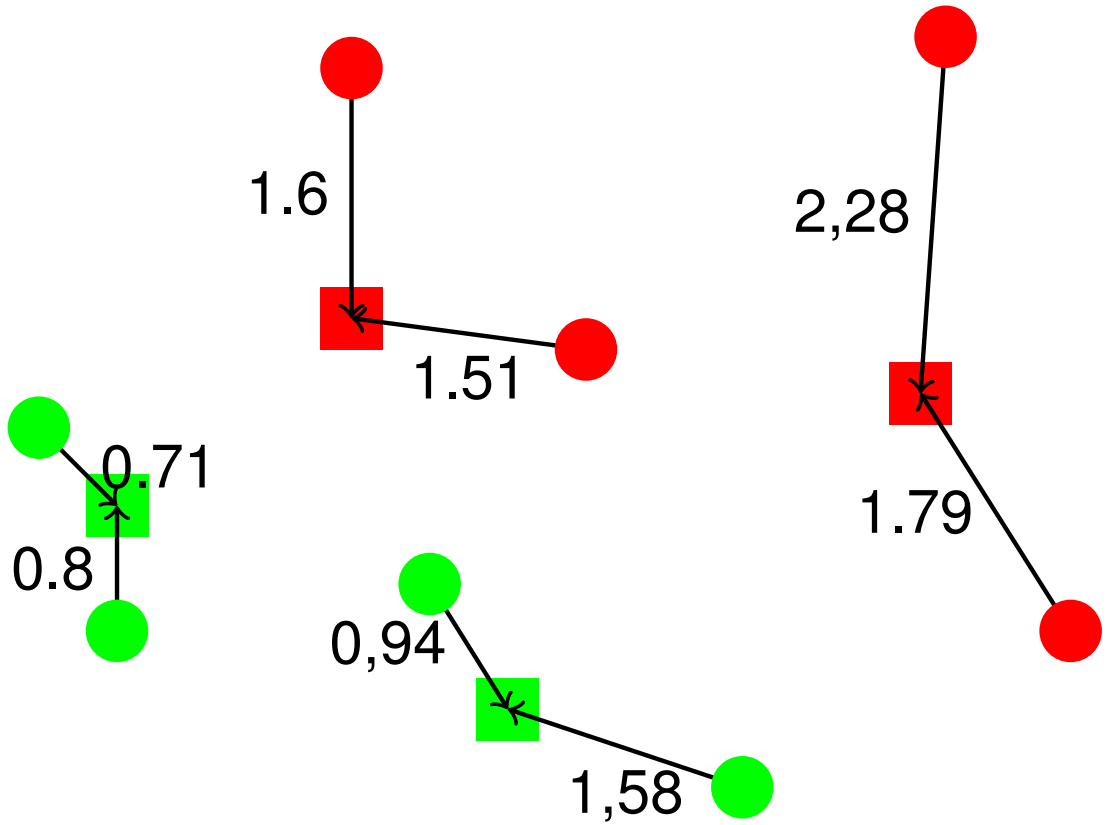
ε-representativeness

We take each point from $\mathcal{D}$ and we look for its closest point in $\mathcal{D}_R$ with the same class.



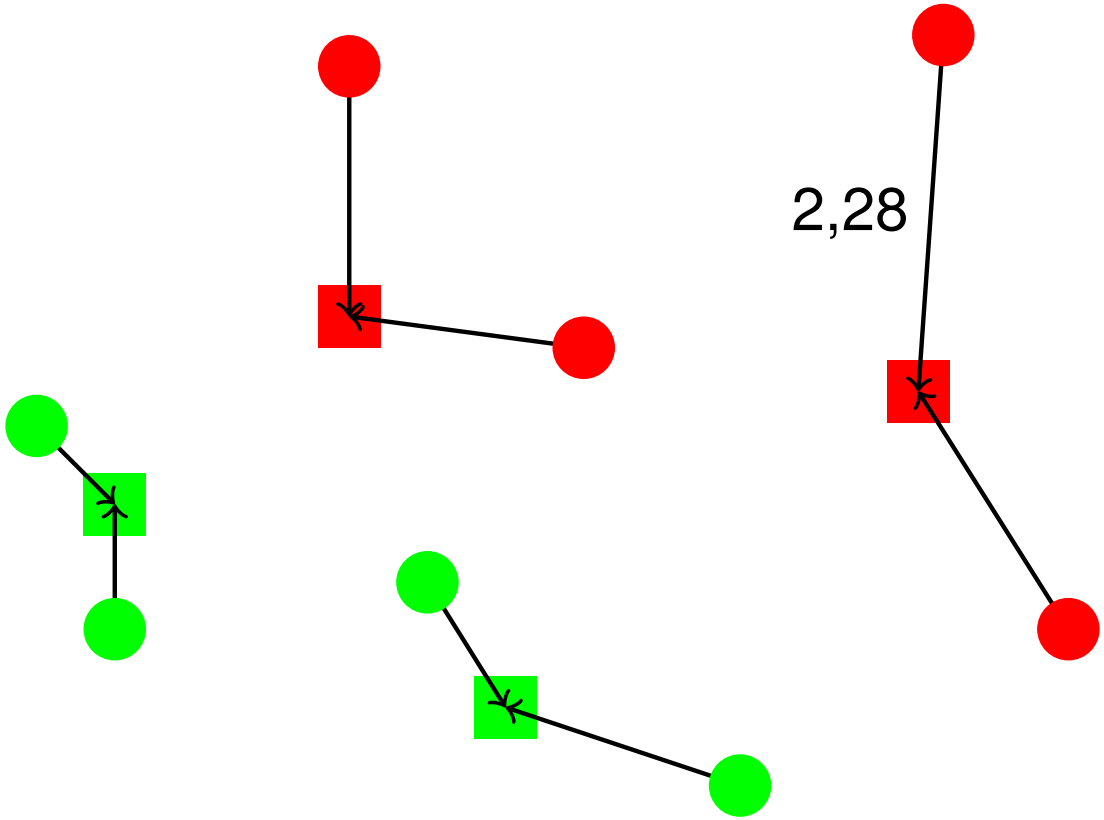
ε-representativeness

We compute the lengths of these arrows.



ε-representativeness

And we choose the maximum of these lengths.



ε-representativeness

Theorem 2 [1]

Let be $\mathcal{D}$ a binary dataset (with only two classes), and let be $\tilde{\mathcal{D}}$ a λ-balanced ε-representative dataset of $\mathcal{D}$. Let be $\mathcal{N}_w$ a perceptron with weights $w \in \mathbb{R}^{n+1}$. Then,

$$\varepsilon \leq \min \left\{ \frac{||wx||}{||w||} : (x, c_x) \in \mathcal{D} \right\} \Rightarrow \mathbb{A}(\mathcal{D}, \mathcal{N}_w) = \mathbb{A}(\tilde{\mathcal{D}}, \mathcal{N}_w)$$

where $\mathbb{A}$ denotes the accuracy of the classifier.

ε-representativeness

We now know that if the ε-representativeness is low:

1. The persistent diagrams of $\mathcal{D}$ and $\mathcal{D}_R$ built with Vietoris-Rips are similar.
2. $\mathcal{D}$ and $\mathcal{D}_R$ can have similar performance metrics for a perceptron.

So we ask ourselves...

ε-representativeness

We now know that if the ε-representativeness is low:

1. The persistent diagrams of $\mathcal{D}$ and $\mathcal{D}_R$ built with Vietoris-Rips are similar.
2. $\mathcal{D}$ and $\mathcal{D}_R$ can have similar performance metrics for a perceptron.

So we ask ourselves...

Which data reduction method gives us the best ε-representativeness?

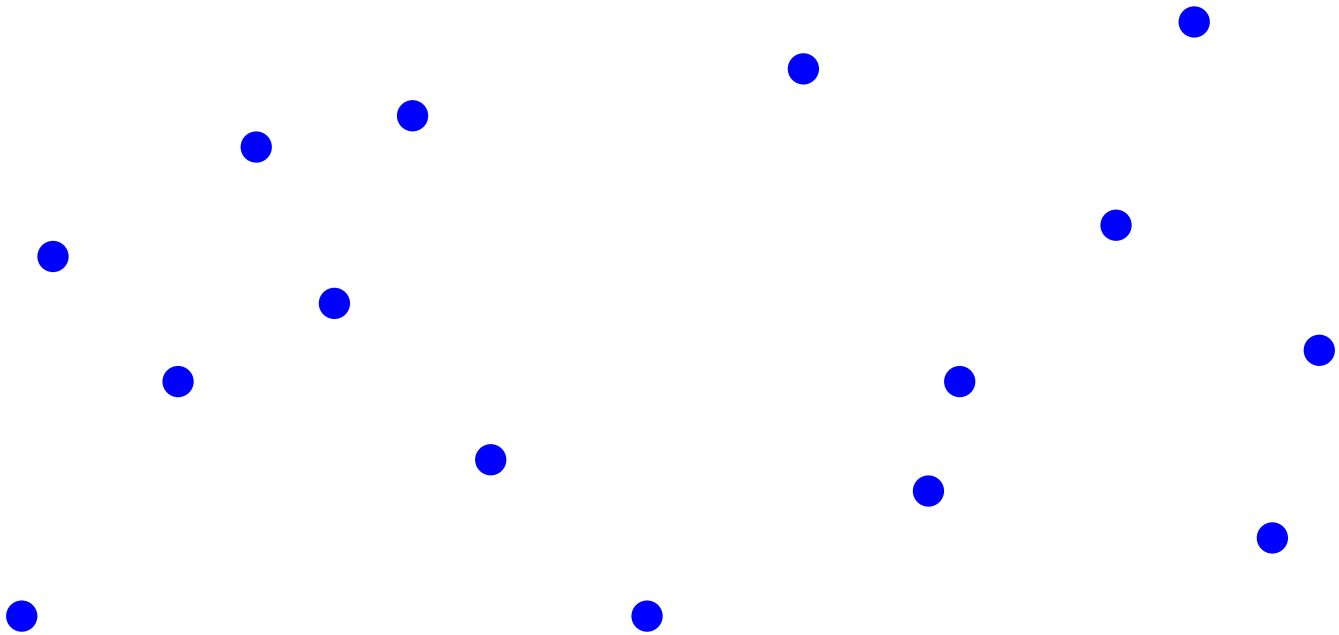
MaxMin Selection

The MaxMin Selection is a sequential algorithm that is applied class by class with these steps:

1. Pick an item $x \in de$ at random and include it in $\mathcal{D}_R$.
2. Pick the item $x \in \mathcal{D} \setminus \mathcal{D}_R$ that maximizes $\min_{x' \in \mathcal{D}_R} ||x - x'||$ and include it in $\mathcal{D}_R$.
3. Repeat the Step 2 until you get the desired number of items.

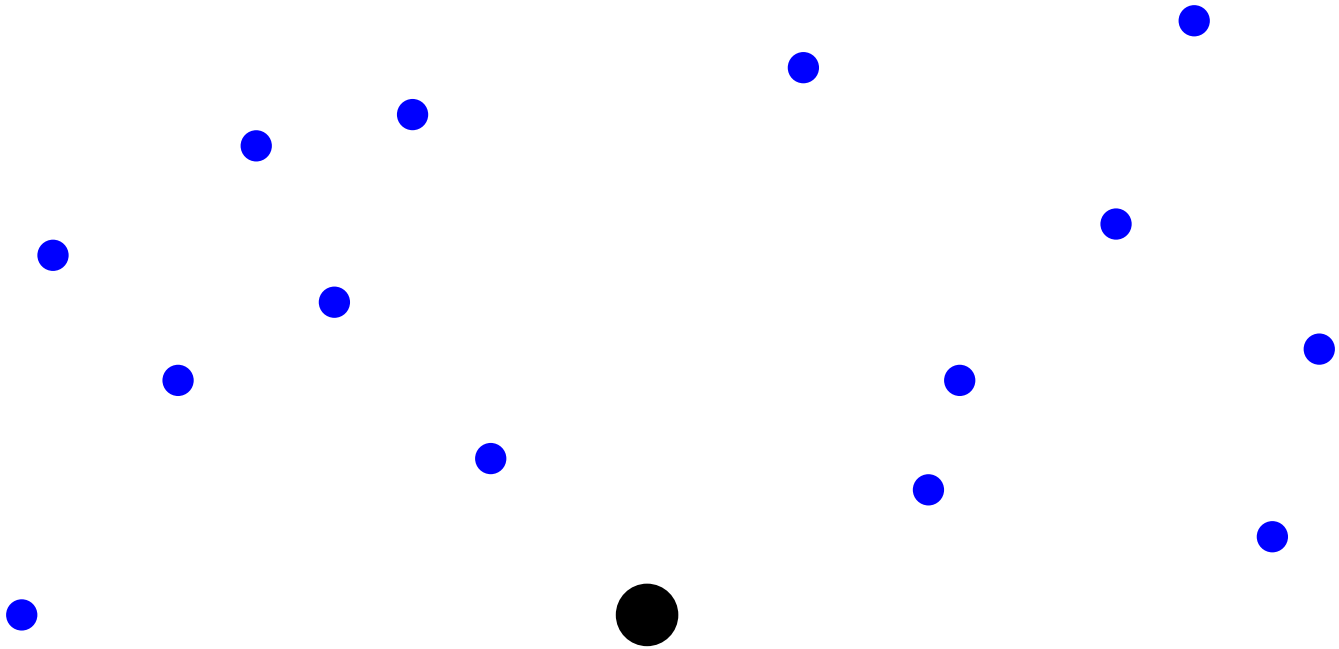
MaxMin Selection

Consider this dataset with just one class and 15 items.



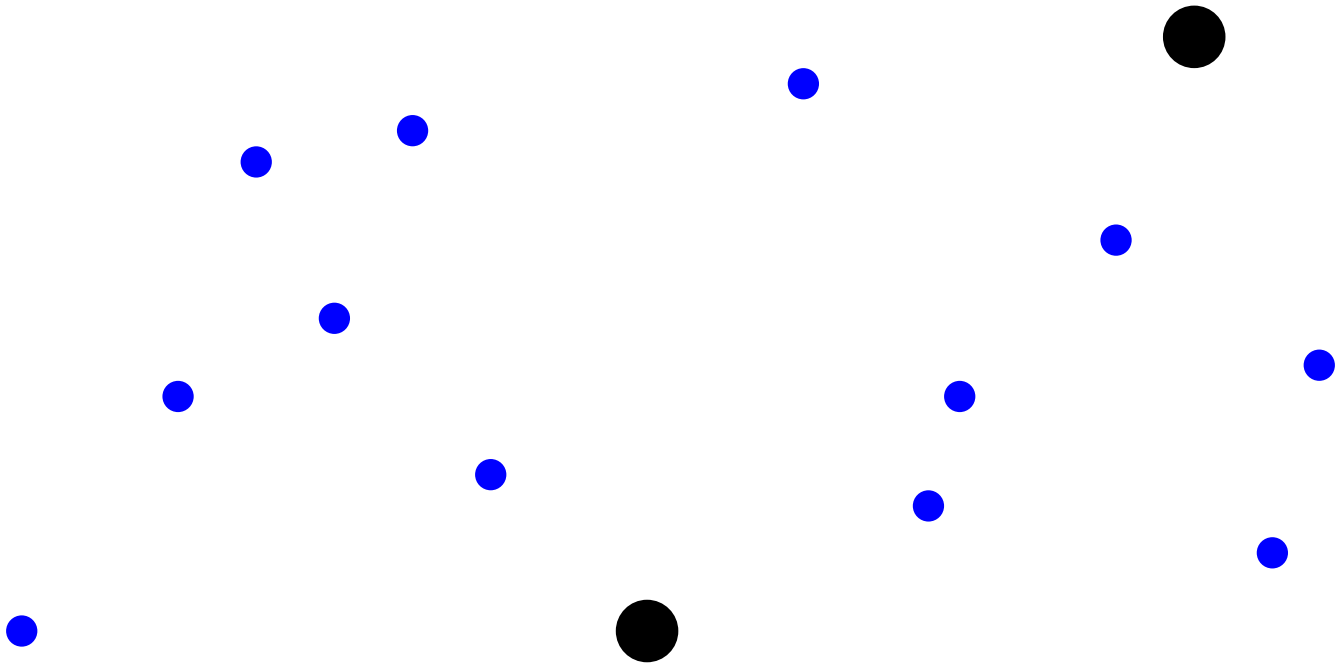
MaxMin Selection

We start with one item chosen at random.



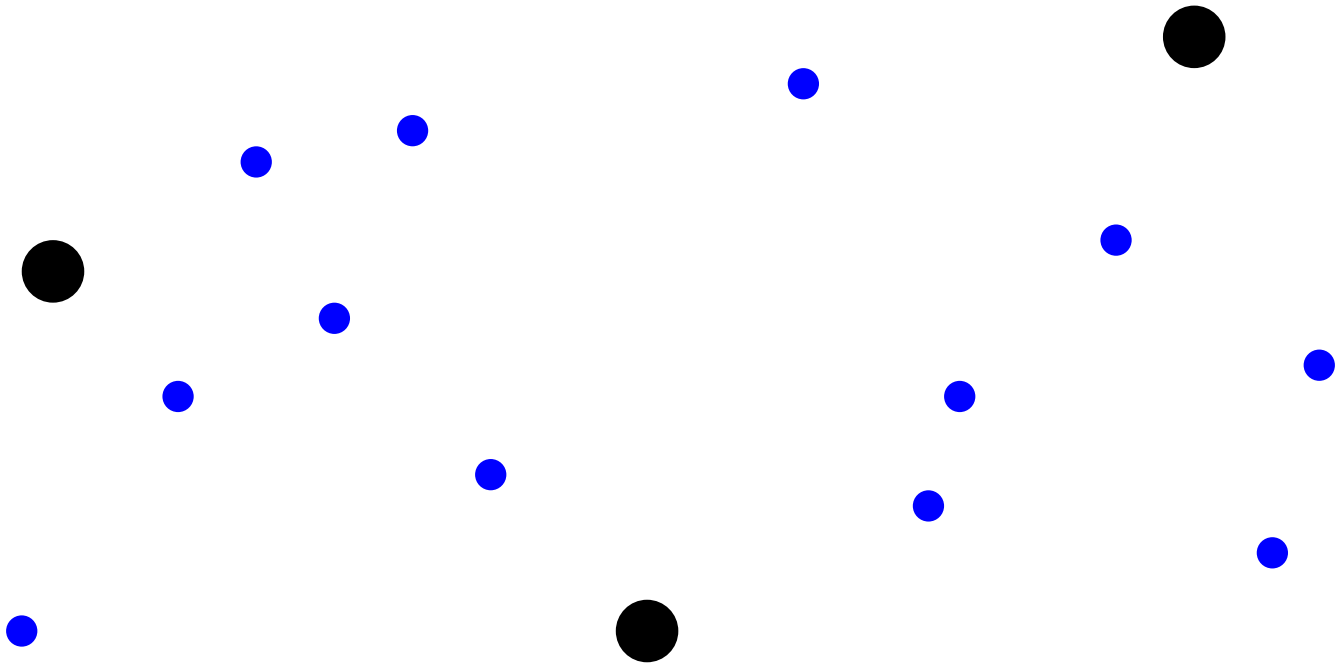
MaxMin Selection

Then, we choose the farthest item.



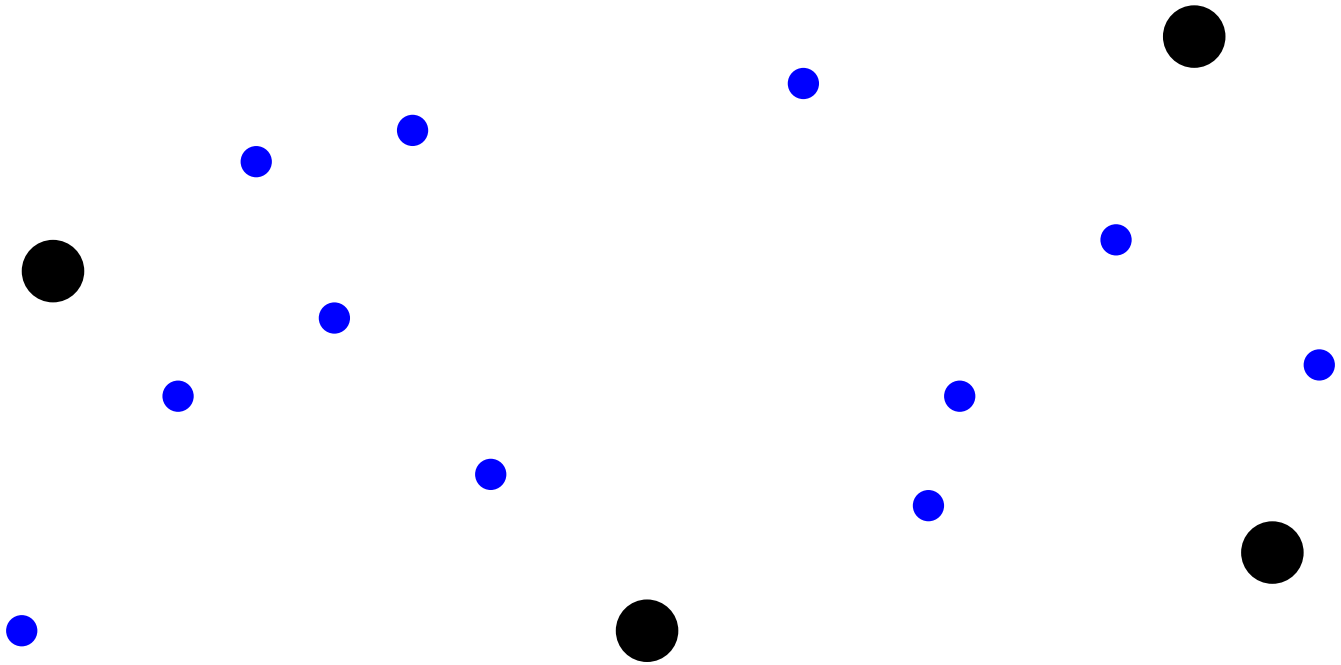
MaxMin Selection

And then, the item that is farthest from both selected items.



MaxMin Selection

And we can go on until we want.



Applying data reduction

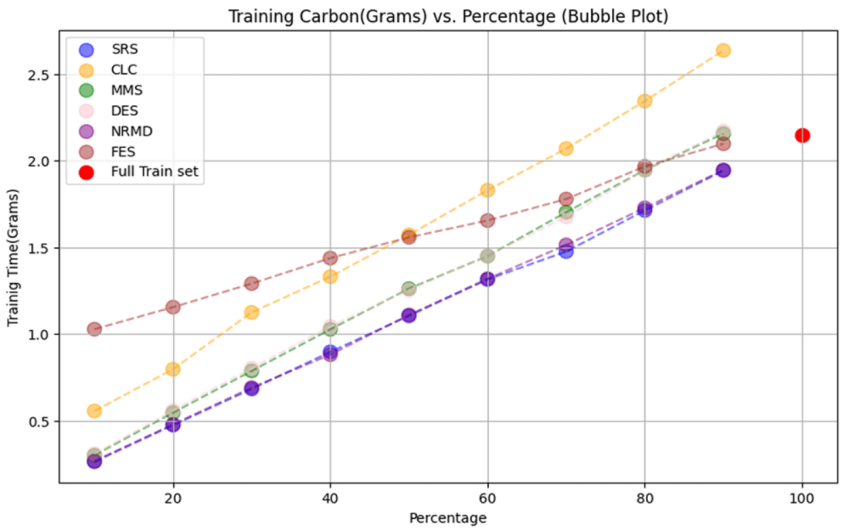
We applied some experiments about data reduction on the **Collision Dataset**.

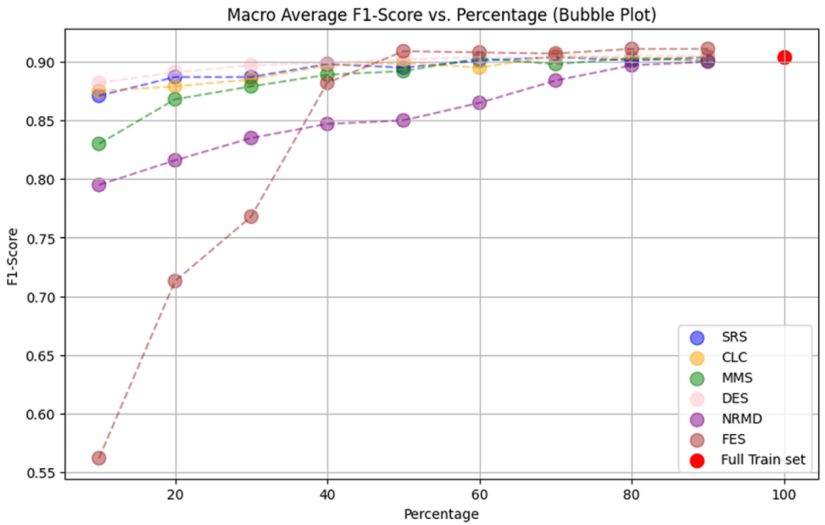
It consists on a set of simulations where a platoon of vehicles navigates an environment. The classification task consists in deciding whether the platoon will collide based on features such as the number of cars and their speed.

Mongelli, M., Ferrari, E., Muselli, M., & Fermi, A. (2019). Performance validation of vehicle platooning through intelligible analytics. IET Cyber-Physical Systems: Theory & Applications, 4(2), 120-127.

Applying data reduction

We trained a fixed Multi-Layer Perceptron with the full dataset and with many reduced dataset given by six different methods and we got the following results:





Applying data reduction

There is a significant correlation between ε-representativeness of the subset and the F1-score of the trained network.

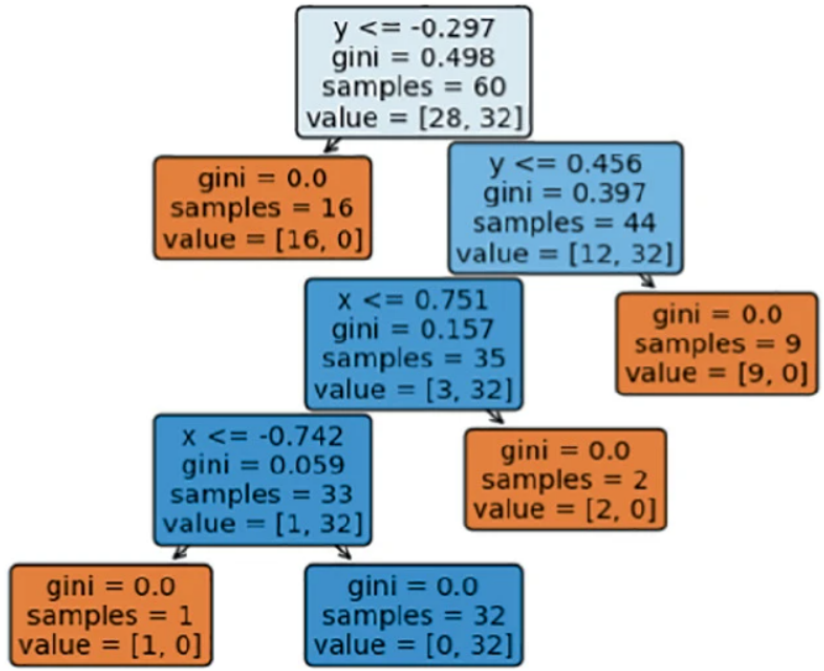
Perera-Lago, J., Toscano-Duran, V., Paluzo-Hidalgo, E., Gonzalez-Diaz, R., Gutiérrez-Naranjo, M. A., & Rucco, M. (2024). An in-depth analysis of data reduction methods for sustainable deep learning. Open Research Europe, 4(101), 101.

	Spearman's ρ	p-value
10%	-0.38	0.0
20%	-0.43	0.0
30%	-0.42	0.0
40%	-0.39	0.0
50%	-0.22	0.1
60%	-0.15	0.24
70%	-0.19	0.14
80%	-0.07	0.58
90%	-0.14	0.3

Applying data reduction

We also performed some experiments reducing the Collision Dataset in another family of models more interpretable by construction: Decision Trees.

Perera-Lago, J., Toscano-Durán, V., Paluzo-Hidalgo, E., Narteni, S., & Rucco, M. (2024, July). Application of the representative measure approach to assess the reliability of decision trees in dealing with unseen vehicle collision data. In World Conference on Explainable Artificial Intelligence (pp. 384-395). Cham: Springer Nature Switzerland.



Applying data reduction

In this case, we also found that:

- Subsets with better ε-representativeness train decision trees with higher accuracy
- Subsets with better ε-representativeness train decision trees more similar to the tree train with the full dataset in terms of feature importance

Thanks for your attention.