A Multi-View Panorama of Data-Centric AI

Techniques, Tools, and Applications

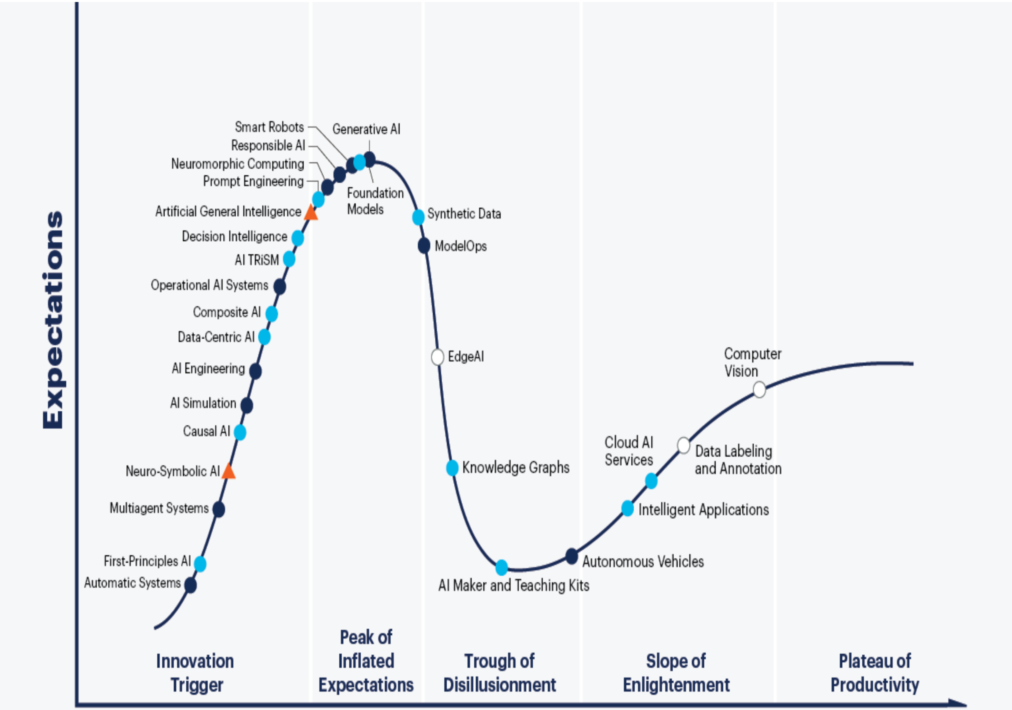
Alberto Fernández, University of Granada, **alfh@ugr.es**

Data-Centric AI

An innovation trigger for Machine Learning Research

Hype Cycle for Artificial Intelligence

Innovation & Impact for Business and Academia



Data-Centric AI

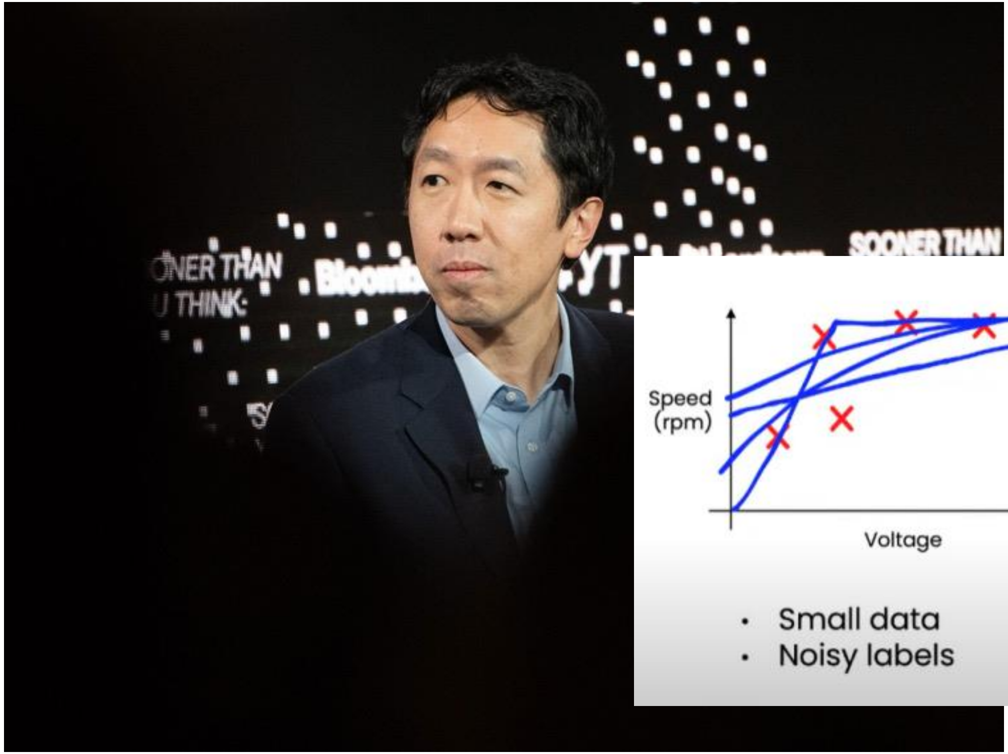
AI TRiSM

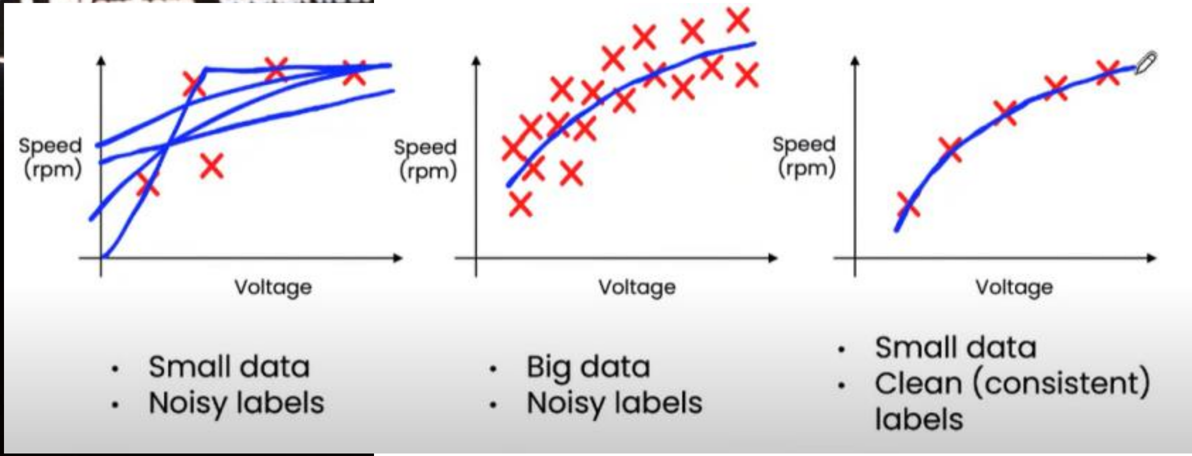
Responsible AI

Synthetic Data

[Gartner, What's New in Artificial Intelligence from the 2023 Gartner Hype Cycle]

Imperfect Data versus Smart Data



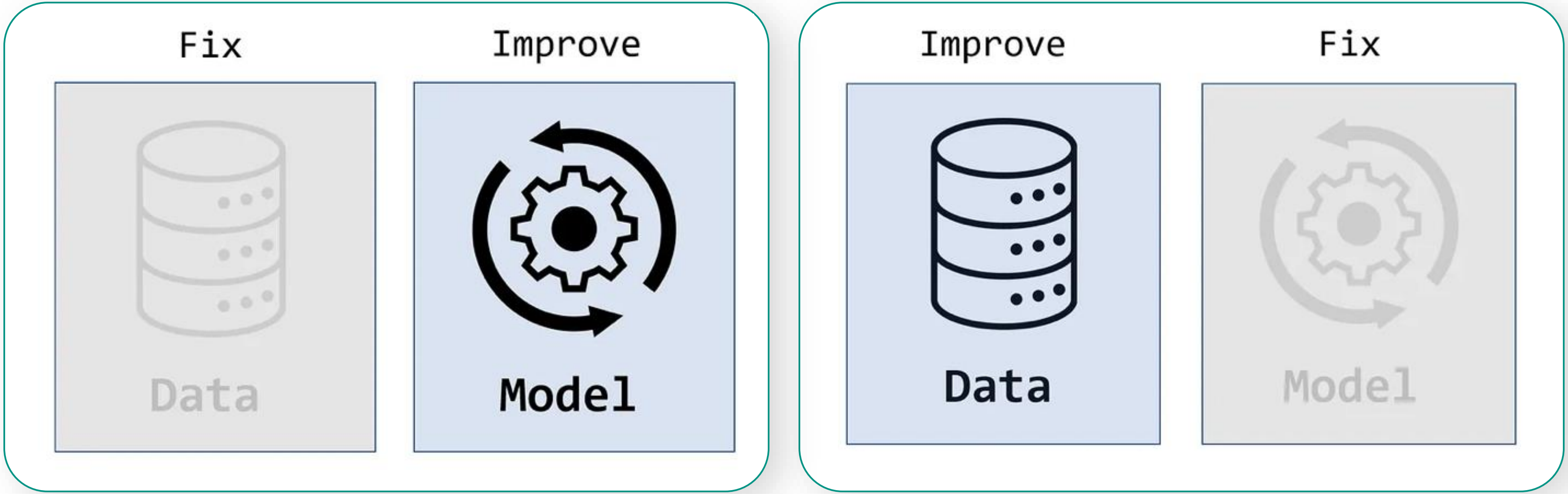


Andrew Ng: Unbiggen AI, IEEE Spectrum

Data-Centric AI Artificial Intelligence

- Model-Centric AI has reached a **point of saturation**. In terms of improvement potential, there is now more gain in shifting our attention towards **improving data.**

Model-Centric AI **Data-Centric AI**



Hands-on Tutorial

Data Centric AI: Tuning Model vs Improving Data

https://colab.research.google.com/drive/1UtyW47jVdfS9pxLf5MLyGmSdgmI0icz8?usp=sharing

Data flaws are not restricted to structure and format

Some data characteristics need to be considered

Imbalanced Data

Disproportion between concepts of interest. Worsens with concept rarity.

Underrepresented Data

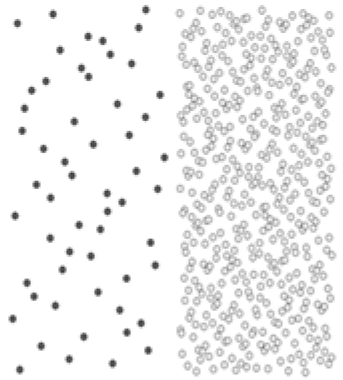
Concept subgroups with the same outcome, despite having different characteristics.

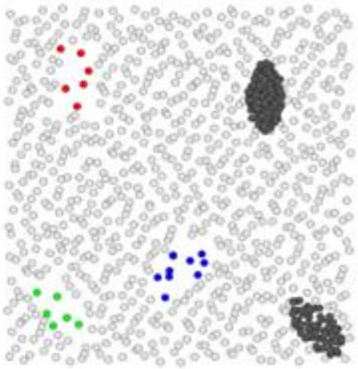
Overlapped Data

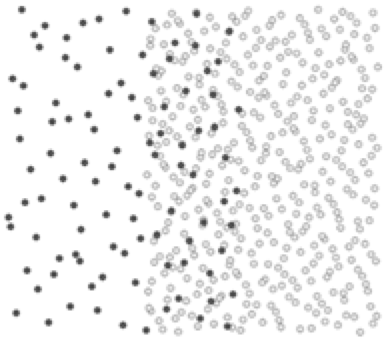
Concepts with similar characteristics but distinct outcomes.

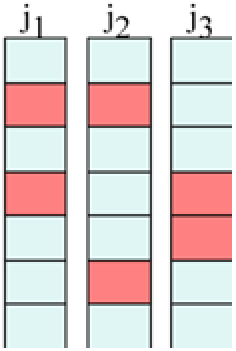
Missing Data

Missing information due to several reasons, e.g., non-disclosure and transmission/collection errors.









Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F. (2018). Data Intrinsic Characteristics. In: Learning from Imbalanced Data Sets. Springer, Cham.

Interplay between Data Intrinsic Characteristics

- In real-world domains, data characteristics **arise simultaneously**. However, we still lack a profound **understanding** of their interplay and methods to fully **define** and **quantify** them.
- There are current **open challenges in the intersection** between: *imbalance and overlap*, *imbalance and missing data*, *imbalance and privacy*, *privacy and fairness*, *imbalance and fairness*, …

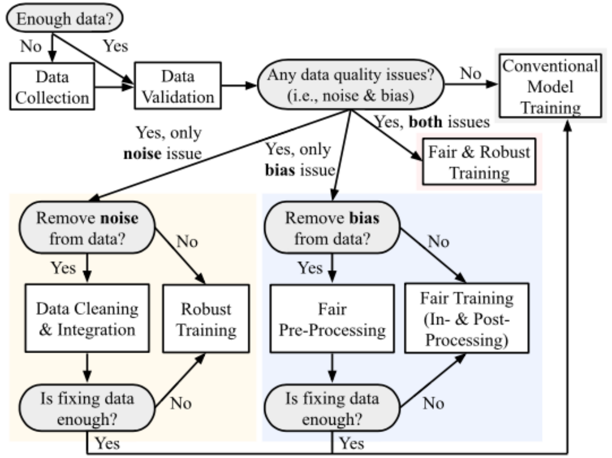
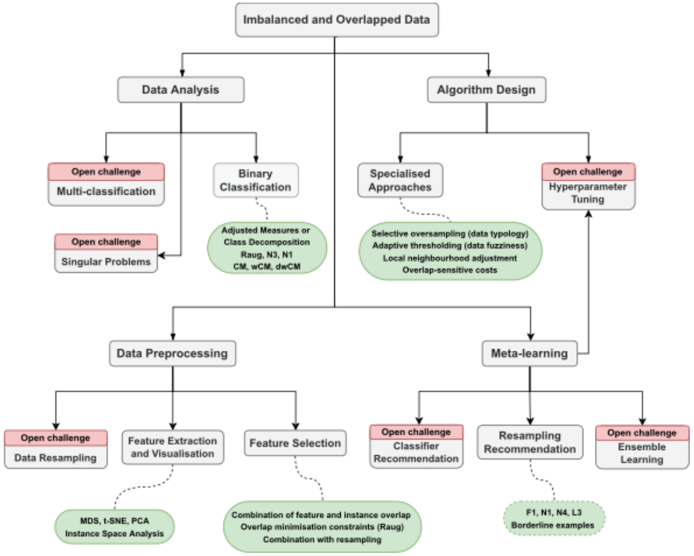
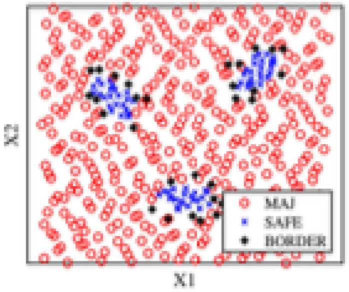
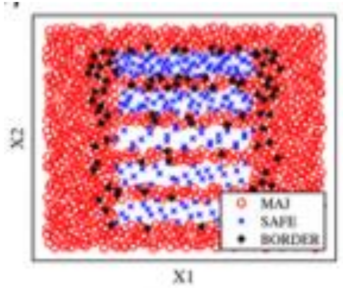


Fig. 2 Decision tree on how data-centric AI techniques connect with each other in one workflow







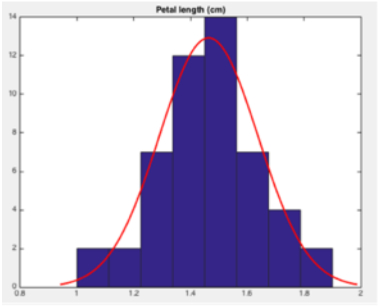
Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A., & Santos, J. (2023). A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research. Information Fusion, 89, 228-253.

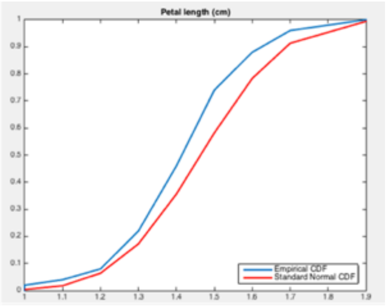
Data Profiling: Validating and Understanding Data

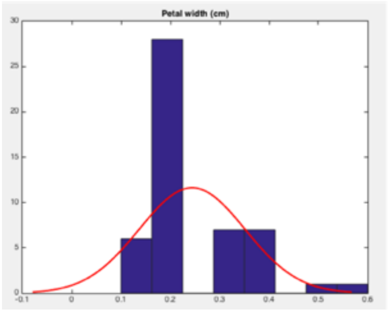
- Data Profiling involves ***iteratively*** examining the **structure**, **characteristics**, and **quality** of a dataset. This comprehends:
    - **Metadata Analysis:** Structure of data, including types, formats, constraints. Data should match the expected formats.
    - **Statistical Properties:** Basic statistical descriptors of data and feature distribution.
    - **Data Quality Assessment:** Checking for anomalies (e.g., inconsistencies, duplicates) or complicating factors (e.g., missing data, noisy data).
    - **Relationship and Interaction Analysis:** Identifying relationships in data and deriving possible insights to investigate further (e.g., dependencies, contraints).

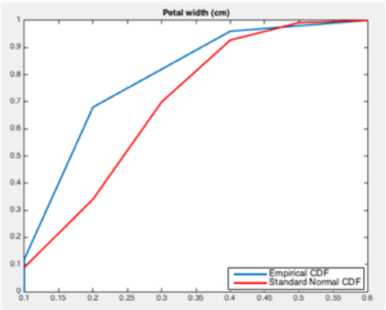
Data Profiling: Data Visualization

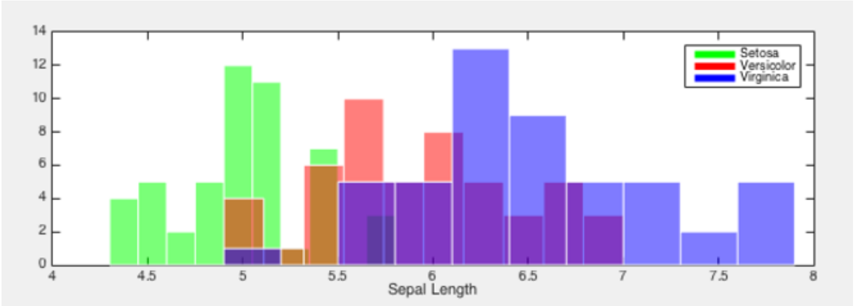
- Visualization goes hand in hand with data profiling, since it is crucial for feature assessment (feature distribution, outliers, symmetry, discriminative power…)

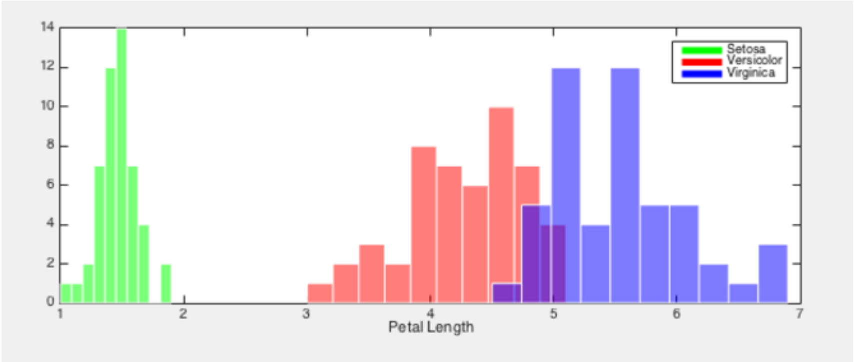




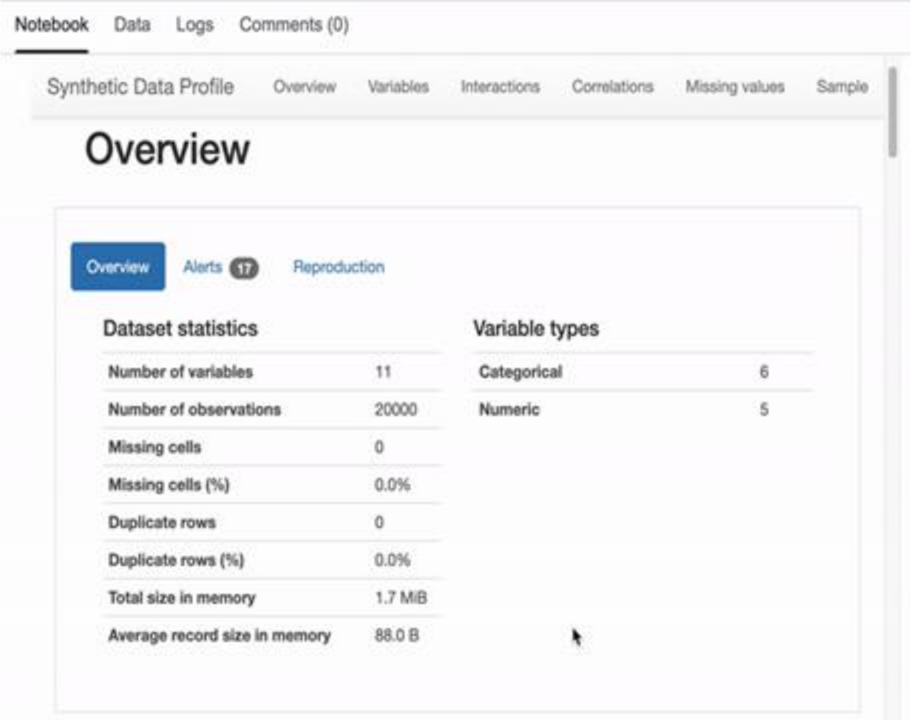








Data Profiling OSS: **YData-Profiling** (previously Pandas-Profiling)



- Automatic Generation of Data Quality Alerts
- Supports Tabular and Time-Series Data
- Comparison Report

Unique Missing Uniform Zeros
High cardinality High correlation Constant

https://docs.profiling.ydata.ai

12.2K stars

1.6K forks

```
pip install ydata-profiling
```

Clemente et al. (2023). ydata-profiling: Accelerating data-centric AI with high-quality data. Neurocomputing

Hands-on Tutorial

Data Centric AI: Data Profiling

https://colab.research.google.com/drive/1qAHxPa6lB0Cmc_V6sMr_pdQnbhXaAYS8?usp=sharing

Data Complexity

Characterizing data complexity and classification behaviour

Data Complexity

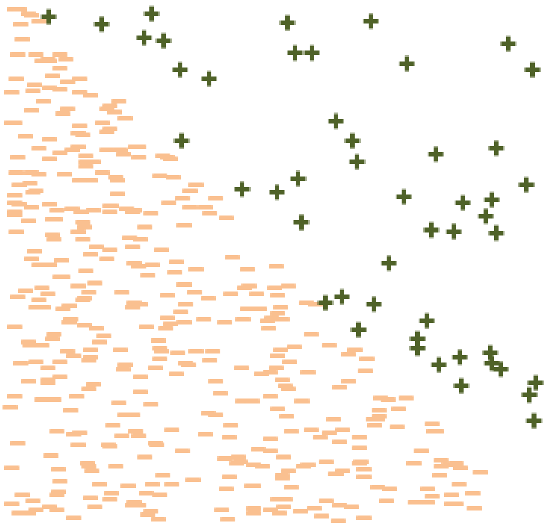
- Given data from a new problem, can we determine whether there exists a clean decision boundary between the classes?
- Are the classes intrinsically distinguishable?
- To what extent can this boundary be inferred by the automatic algorithms?
- Which classifiers can do the best job?

These questions are about the intrinsic complexity of a classification problem, and the match of a classifier's capability to a problem's intrinsic complexity.

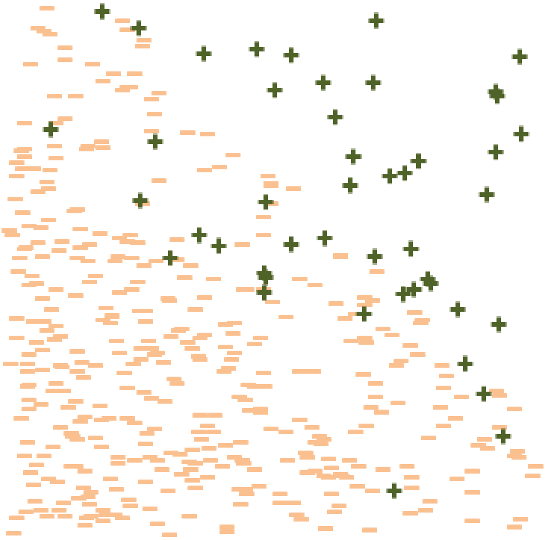
- Factors affecting performance can be:
  - The shape of the classes and thus the shape of the decision boundary
  - The amount of overlap bewteen the classes
  - The proximity of two classes
  - The number of informative samples available for training
  - (...)

Data Complexity

Easy Dataset

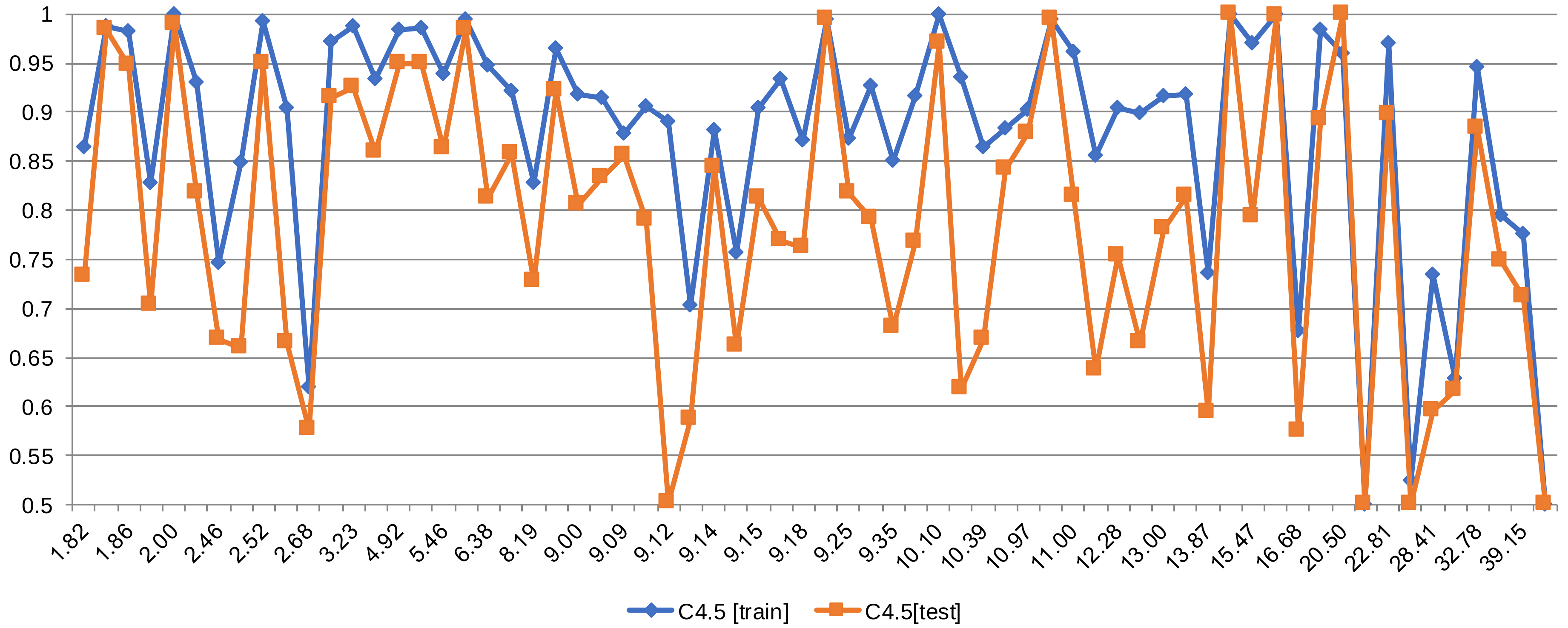


Hard Dataset

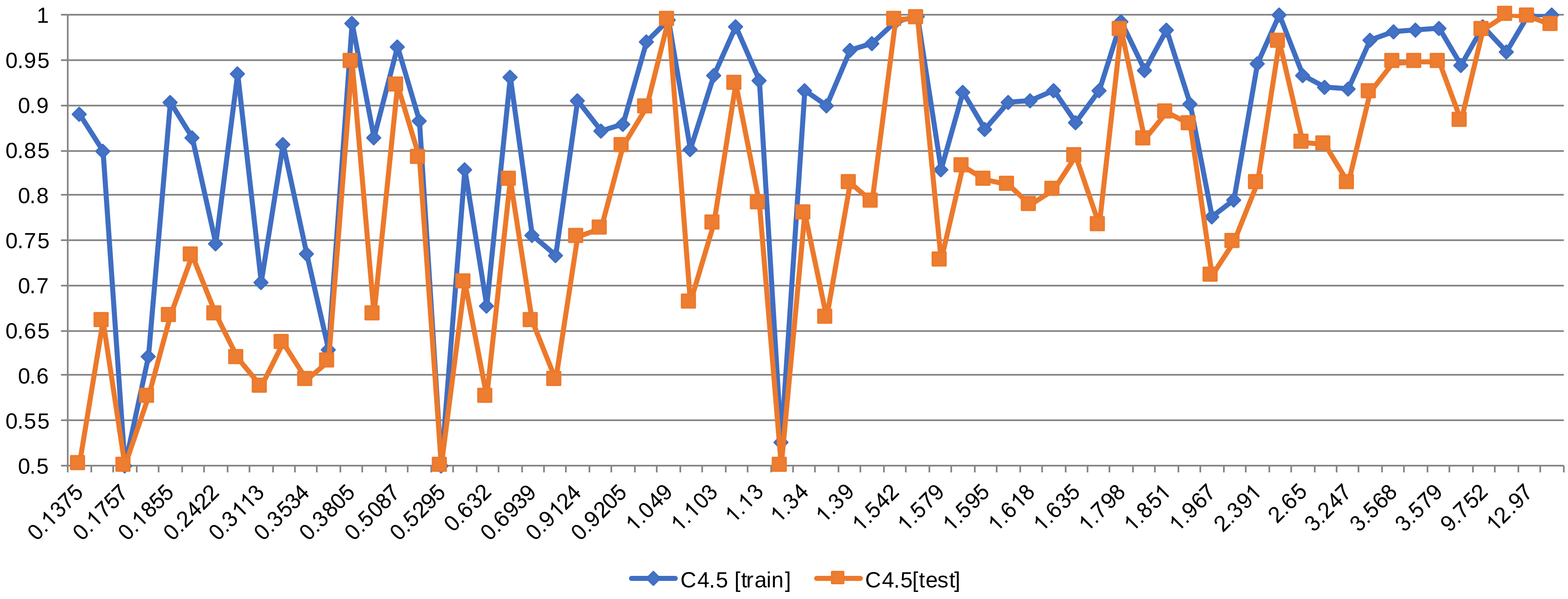


[Lopez et al. (2013). An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. Information Sciences]

Data Characterization (Imbalance Ratio metric)



Data Characterization (Overlap metric, F1)



Data Complexity: Learning Paradigms and Classifier Footprints

- Several real-world applications suffer from distinct (and often combined) data irregularities.
- Classifiers respond to different complexity factors **in their unique ways**:

Table 1
Examples of data irregularities in real-world applications.

Scenario	Type of data irregularity
Credit card fraud detection	Class imbalances, class skew [18]
Breast cancer diagnosis	Class imbalance, class skew, small disjuncts [19,20]
Market segmentation	Class imbalance, class skew [21]
Facial and emotion recognition	Small disjuncts [22]
Survey data	Unstructured missingness [23]
Phylogeny problem	Unstructured missingness [24]
Gene expression data	Unstructured missingness [25]
Visual object recognition	Structural missingness or absent features [17]
Software effort prediction	Unstructured and structural missingness [26]

- Max-margin Classifiers – sensitive to class imbalance, small disjuncts, class distribution skew, absent features, missing features.
- Neural Networks – sensitive to class imbalance, small disjuncts, absent features, missing features.
- *k*-Nearest Neighbours (*k*-NN) – sensitive to class imbalance, small disjuncts, absent features, missing features; immune to class distribution skew as it does not make any assumptions regarding the class-conditional distributions.
- Bayesian Inference – sensitive to class imbalance, small disjuncts, class distribution skew, absent features, missing features.
- Decision Trees – sensitive to class imbalance, small disjuncts, class distribution skew; inherently immune to feature missingness as branching is based only on the observed features.

Data Complexity Measures

- In practical applications, often a problem becomes difficult because of a **mixture of boundary complexity and sample sparsity effects.**
- **Data Complexity Measures started being organized into groups or categories:**

Ho and Basu *(2002)*

(1) Overlap of Individual Feature Values
(2) Separability of Classes
(3) Geometry, Topology, Density of Manifolds

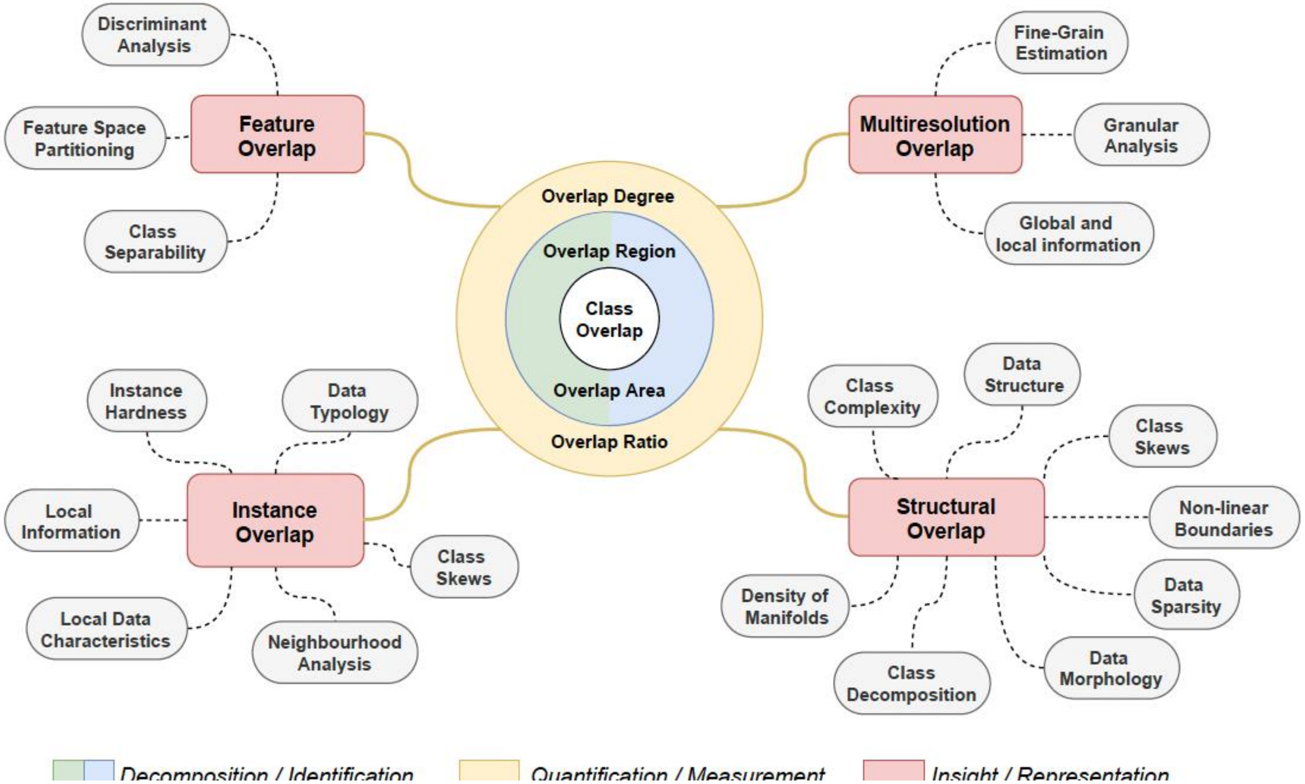
Sotoca *et al.* *(2005)*

(1) Overlap
(2) Class Separability
(3) Geometry and Density

Lorena *et al.* *(2019)*

(1) Feature-Based Measures
(2) Linearity Measures
(3) Neighbourhood Measures
(4) Netework Measures
(5) Dimensionality Measures
(6) Class Imbalance Measures

Characterisation of the class overlap problem



Data Complexity Measures: Feature-based measures

- **Feature-based measures:** Characterise how informative the available features are to separate the classes.

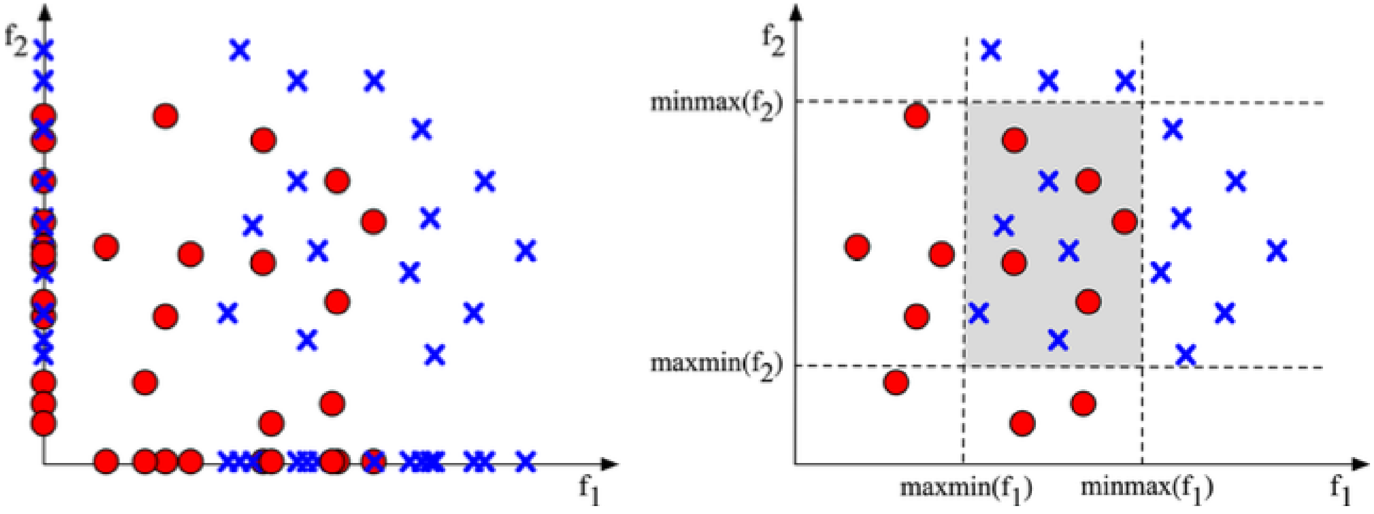
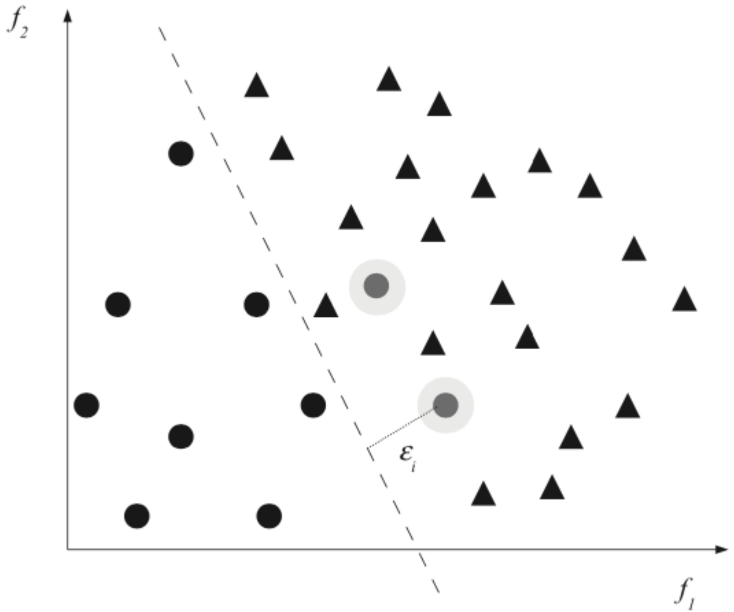


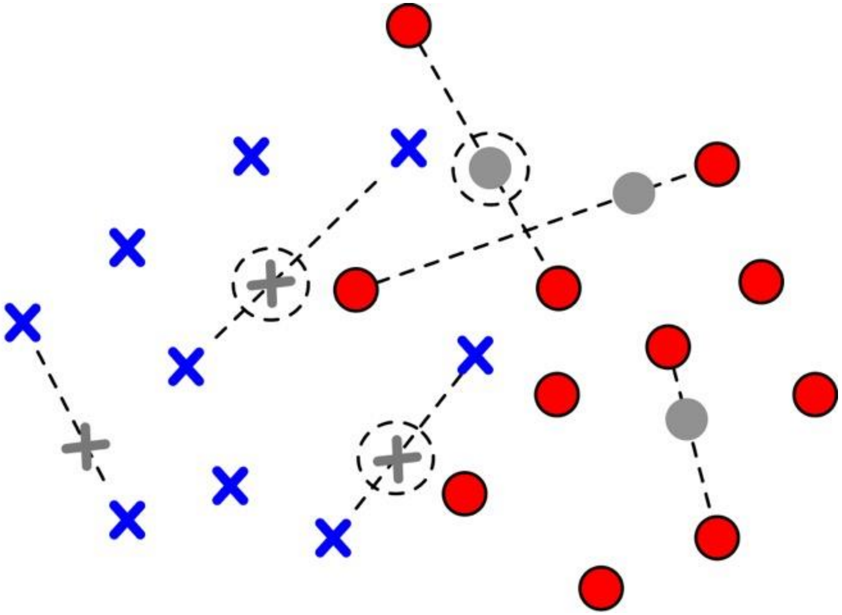
Fig. 5 Representations of F1 (leftside) and F2 (rightside) measures for the same dataset. Note how F1 projects data onto the axis to establish the amount of overlap, where f_1 is the feature with highest discriminative power, i.e., lowest overlap. In turn, F2 considers both features to define a region where classes coexist

Data Complexity Measures: Linearity measures

- **Linearity measures:** Quantify whether classes can be linearly separated.



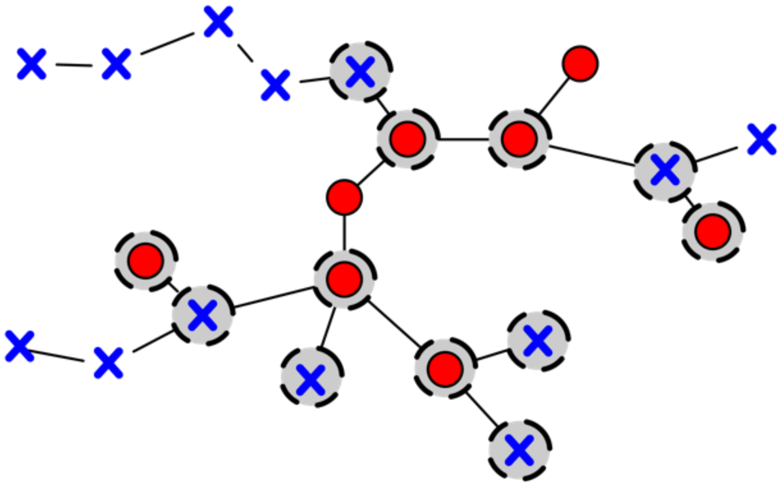
L1



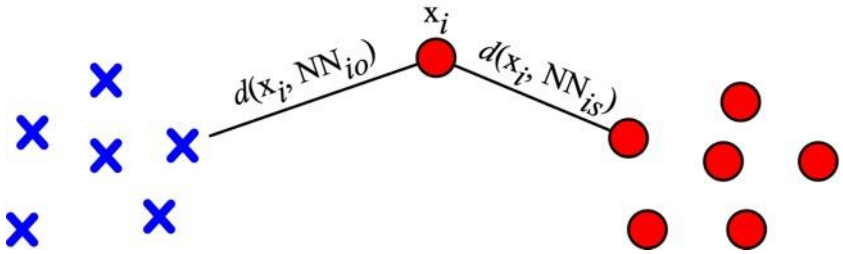
L3

Data Complexity Measures: Neighborhood measures

- **Neighborhood measures:** Characterize the presence and density of same or different classes in local neighborhoods.



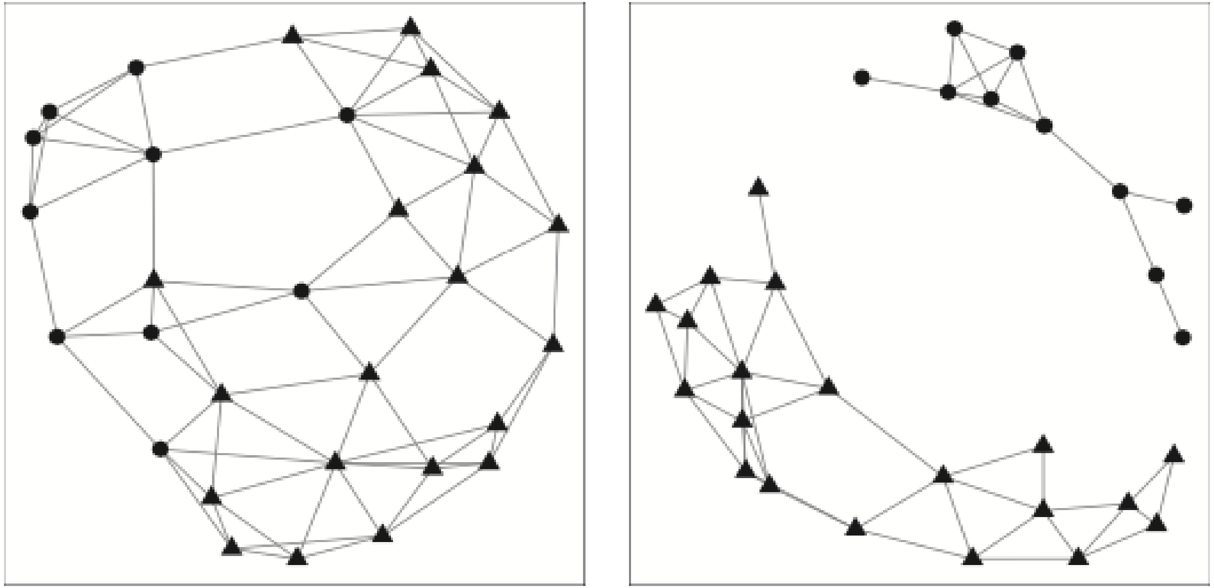
N1



N2

Data Complexity Measures: Network measures

- **Network measures:** Extract structural information from the dataset by modeling it as a graph.



Data Complexity Measures: Dimensionality Measures

- **Dimensionality Measures:** Evaluate data sparsity based on the number of samples relative to the data dimensionality.

  - **Average Number of Features per Dimension (T2):**

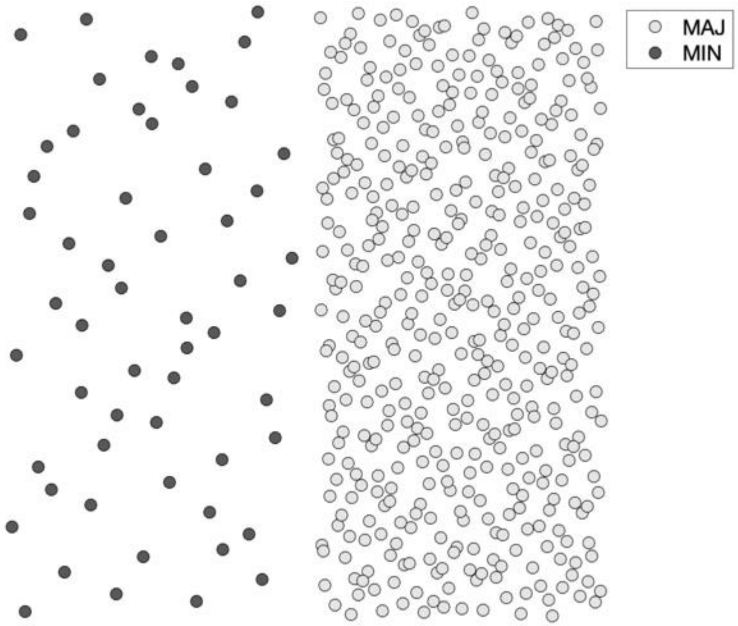
$$T2 = \frac{m}{n}$$

  - **Average Number of PCA Dimensions per Points (T3):**

$$T3 = \frac{m'}{n}$$

Data Complexity Measures: Class Imbalance Measures

- **Class Imbalance Measures:** Consider the ratio of the number of examples between classes.



- **Entropy of Class Proportions (C1):**

$$C1 = -\frac{1}{\log(n_c)} \sum_{i=1}^{n_c} p_{c_i} \log(p_{c_i})$$

- **Imbalance Ratio (C2):**

$$C2 = 1 - \frac{1}{IR}$$

$$IR = \frac{n_c - 1}{n_c} \sum_{i=1}^{n_c} \frac{n_{c_i}}{n - n_{c_i}}$$

Data Complexity Measures: Class Overlap Measures

- **Class Overlap Measures:** Consider class overlap as a concept comprising multiple sources of complexity (feature-level, instance-level overlap, structural overlap, multiresolution overlap).

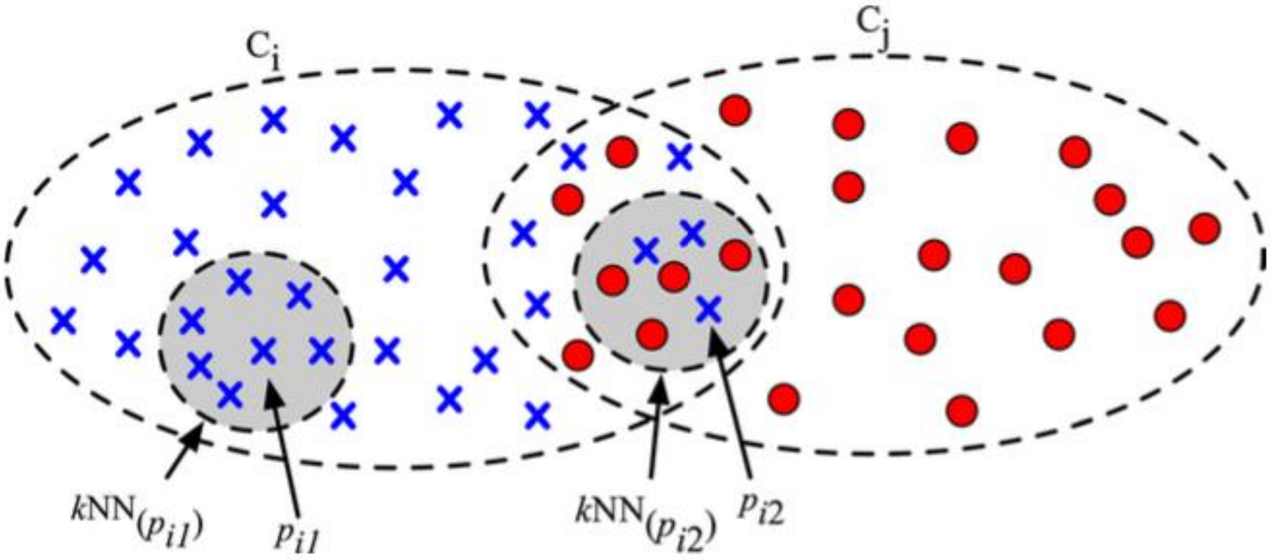
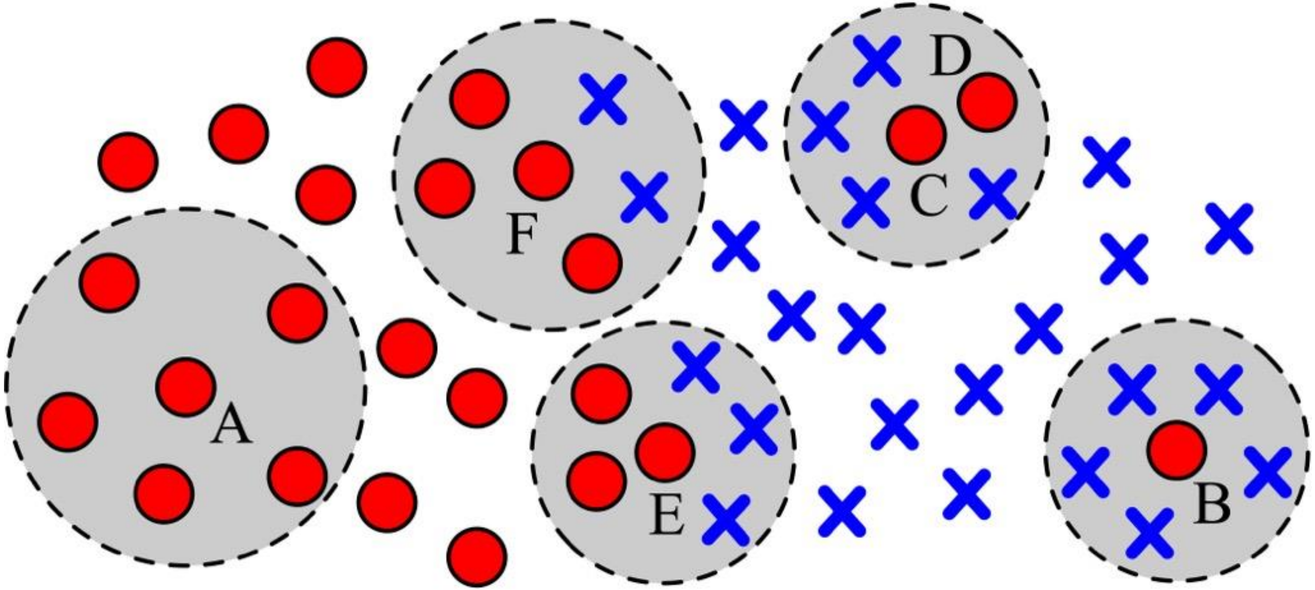


Fig. 18 Basic concepts for R-value computation. Note how $|\mathrm{kNN}(p_{i1}, C_j)| = 0$ and $|\mathrm{kNN}(p_{i2}, C_j)| = 4$, for $k = 6$. Adapted from Oh (2011)

Data Complexity Measures: Data Typology

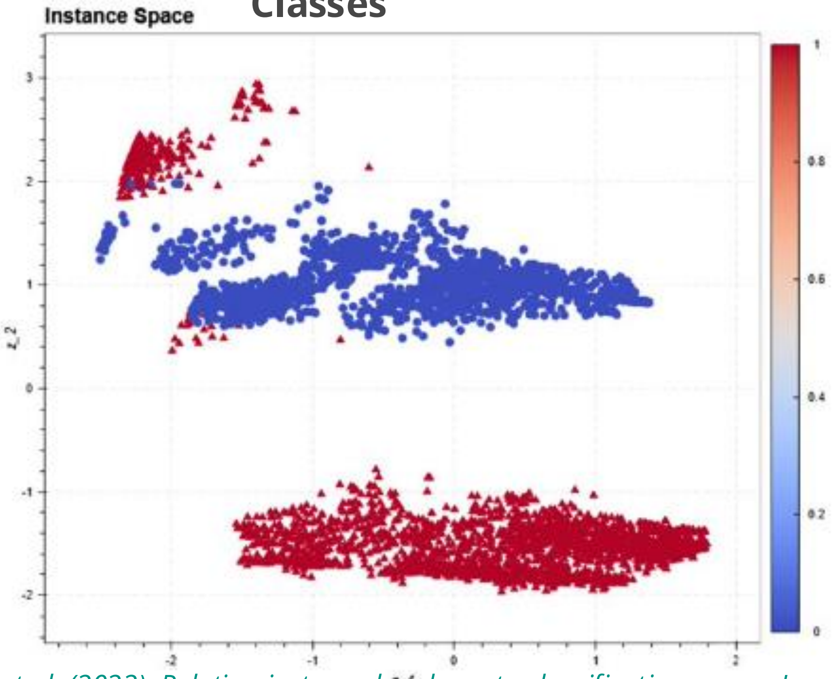
- **Data Typology:** Data complexity is mapped according to the types of examples in data – ***Safe*** (A), ***Borderline*** (E, F), ***Rare*** (C, D), and ***Outlier*** (B).



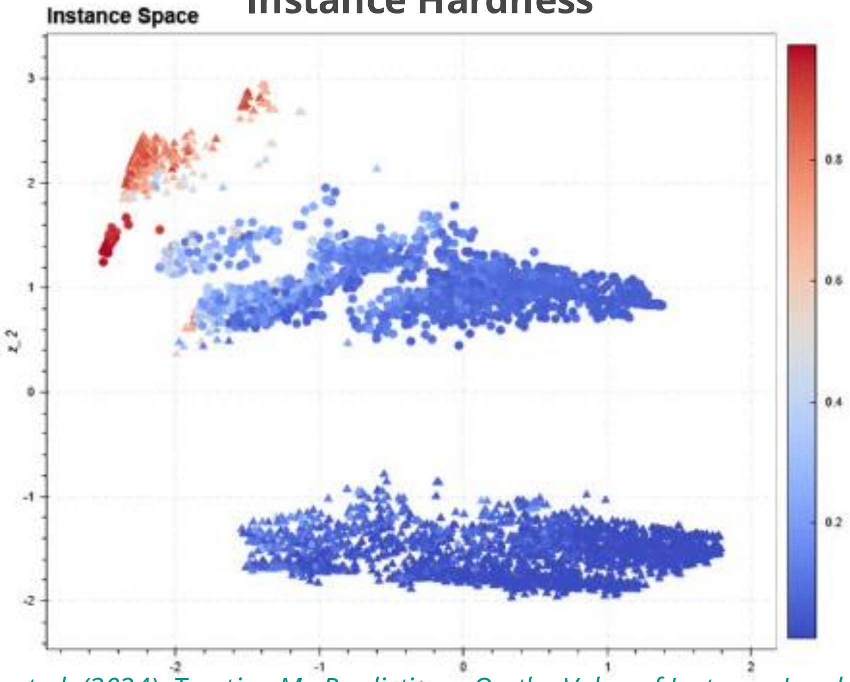
Data Complexity Measures: Instance Hardness

- **Instance Hardness or Instance-Level Complexity**

“Classes”



“Instance Hardness”

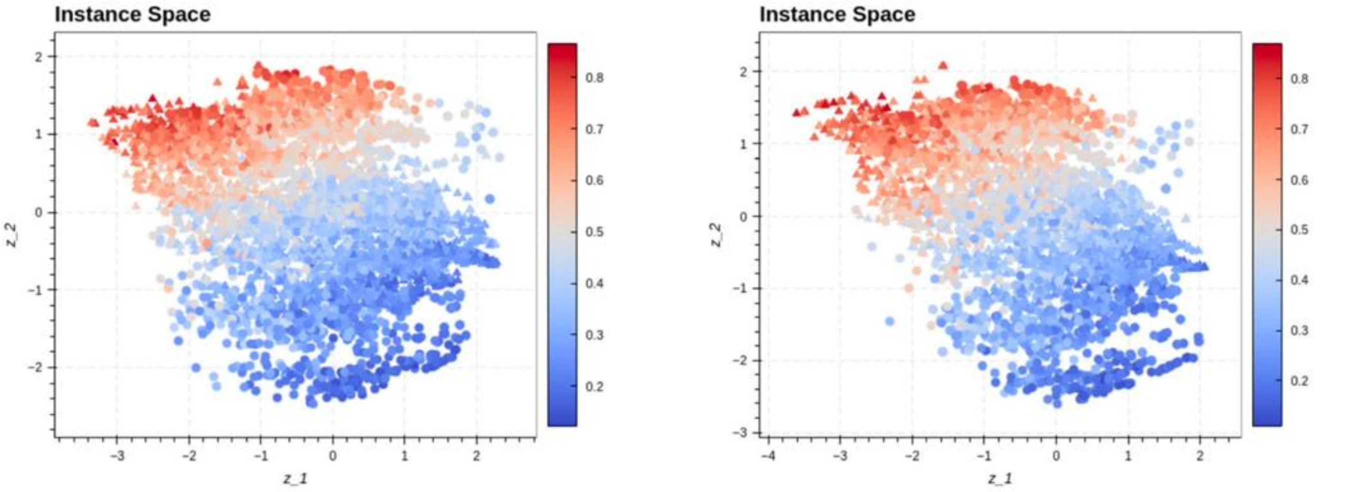


[Paiva et al. (2022). Relating instance hardness to classification performance in a dataset: a visual approach.] *Machine Learning*

[Lorena et al. (2024). Trusting My Predictions: On the Value of Instance-Level Analysis.] *ACM Computing Surveys*

Data Complexity: Applications

- **For a particular application:**



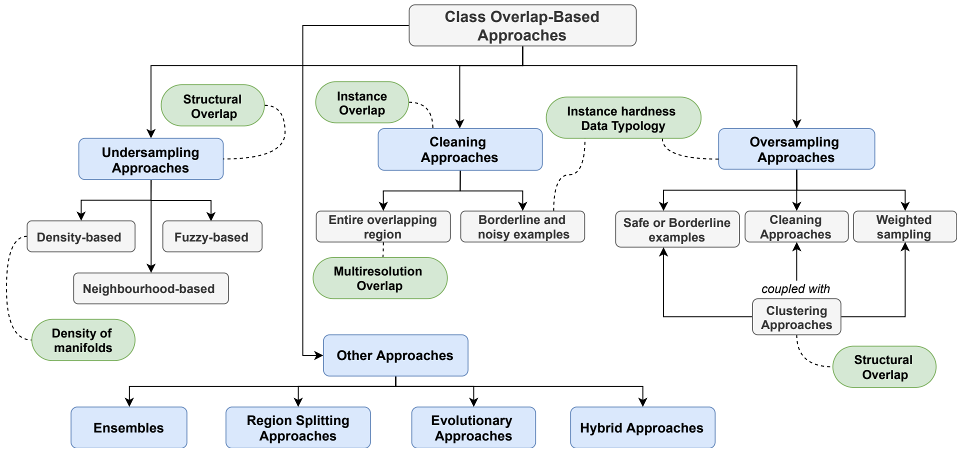
(a) COMPAS dataset ISA projections with `race` as an input attribute.

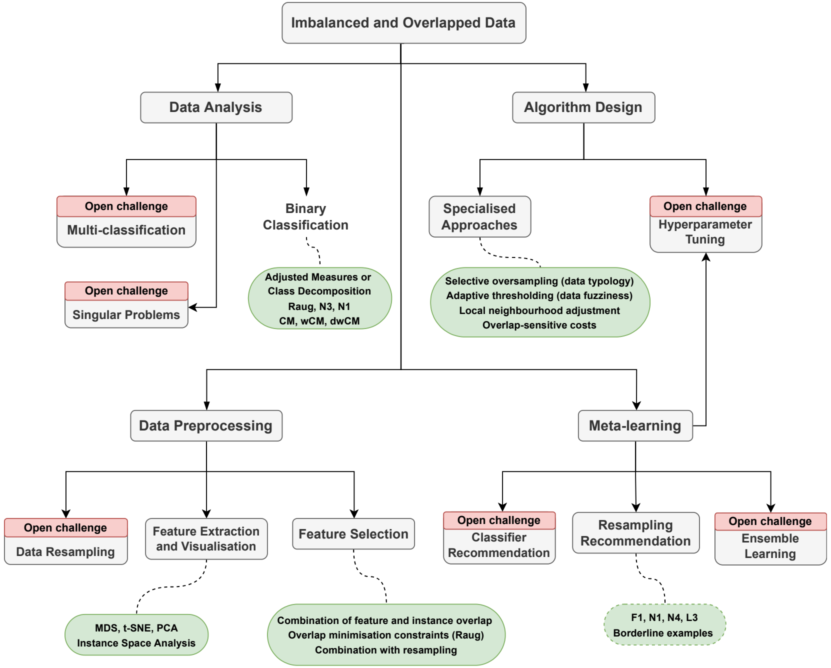
(b) COMPAS dataset ISA projections without `race` as an input attribute.

Fig. 8 COMPAS dataset ISA projections with and without `race` as an input attribute, colored according to the IH value for each instance

Data Complexity: Applications

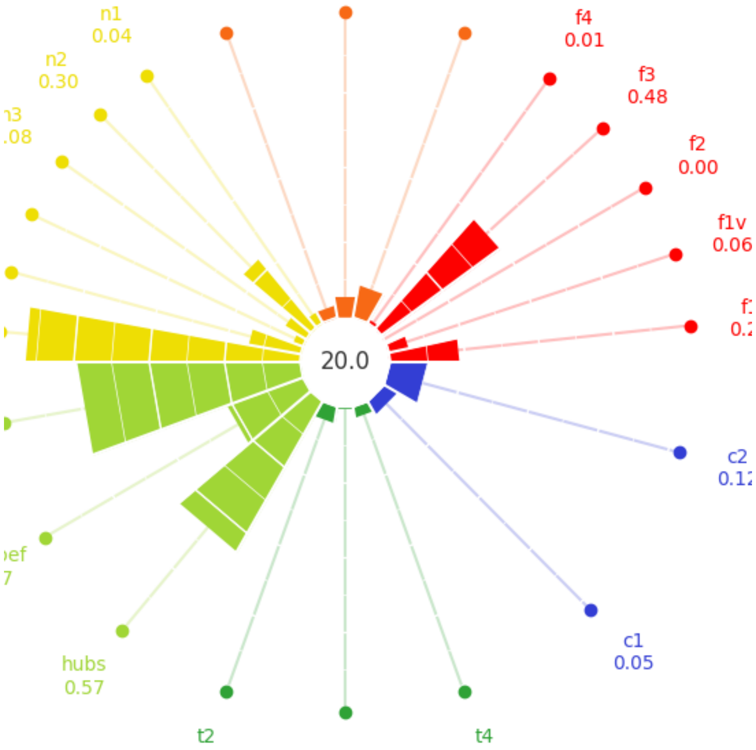
- **For research in classification methods:**





Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A., & Santos, J. (2023). A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research. Information Fusion, 89, 228-253.

Problexity: Problem Complexity Assessment



- Implements the data complexity measures as described in *Lorena et al. (2019).*

```
# Loading benchmark dataset from scikit-learn
from sklearn.datasets import load_breast_cancer
X, y = load_breast_cancer(return_X_y=True)

# Initialize CoplexityCalculator with default parametrization
cc = px.ComplexityCalculator()

# Fit model with data
cc.fit(X,y)
```

https://problexity.readthedocs.io/en/latest/

pip install problexity

[Komorniczak and Ksieniewicz (2023). Problexity – an open-source Python library for binary classification problem complexity assessment. Neurocomputing.]()

PyMFE: Python Meta-Feature Extractor

```
# Load a dataset
from sklearn.datasets import load_iris
from pymfe.mfe import MFE

data = load_iris()
y = data.target
X = data.data

# Extract default measures
mfe = MFE()
mfe.fit(X, y)
ft = mfe.extract()
print(ft)

# Extract general, statistical and information-theoretic measures
mfe = MFE(groups=["general", "statistical", "info-theory"])
mfe.fit(X, y)
ft = mfe.extract()
print(ft)
```

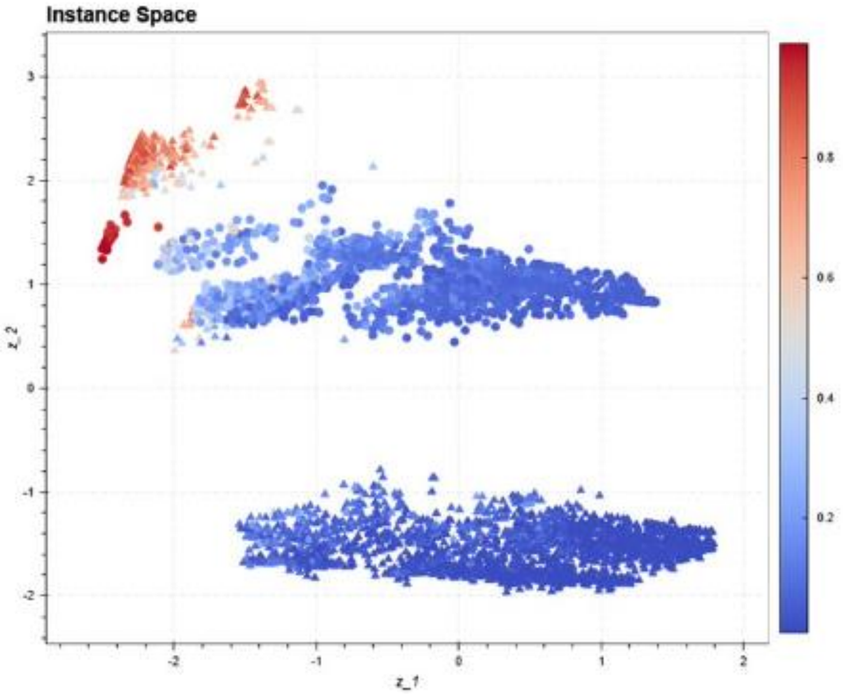
- Comprehensive suite of meta-features
- Different families of meta-features
- Several summarization functions

https://problexity.readthedocs.io/en/latest/

```
pip install pymfe
```

Rivolli et al. (2018). Towards reproducible empirical research in meta-learning. arXiv

PyHard: Instance Hardness Analysis in Machine Learning



- Uses Instance Space Analysis (ISA) to produce a hardness embedding of a dataset relating the performance of ML models to estimated instance hardness meta-features.

https://ita-ml.gitlab.io/pyhard/

```
pip install pyhard
```

[Paiva et al. (2021), PyHard: a novel tool for generating hardness embeddings to support data-centric analysis]()

References and Further Reading

- Das, S. Datta, B. Chaudhuri, Handling data irregularities in classification: Foundations, trends, and future challenges (2018), Pattern Recognition 81, 674–693.
- A. Fernández, S. García, M. Galar, M., R. Prati, B. Krawczyk, F. Herrera, Data Intrinsic Characteristics (2018), Springer International Publishing. pp. 253–277.
- I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, F. Herrera, Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data (2019), Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9, e1289.
- Fernández, A., del Río, S., López, V., Bawakid, A., del Jesus, M. J., Benítez, J. M., & Herrera, F. (2014). Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(5), 380-409.
- Seedat, N., Imrie, F., & van der Schaar, M. (2022). Dc-check: A data-centric ai checklist to guide the development of reliable machine learning systems. *arXiv preprint arXiv:2211.05764*.
- Jakubik, J., Vössing, M., Kühl, N., Walk, J., & Satzger, G. (2024). Data-centric artificial intelligence. *Business & Information Systems Engineering*, 1-9.
- Whang, S. E., Roh, Y., Song, H., & Lee, J. G. (2023). Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, *32*(4), 791-813.
- Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2023). Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*.

References and Further Reading

- Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. IEEE transactions on pattern analysis and machine intelligence, 24(3), 289-300.
- Lorena, A. C., Garcia, L. P., Lehmann, J., Souto, M. C., & Ho, T. K. (2019). How complex is your classification problem? a survey on measuring classification complexity. ACM Computing Surveys (CSUR), 52(5), 1-34.
- Komorniczak, J., & Ksieniewicz, P. (2023). problexity: An open-source Python library for supervised learning problem complexity assessment. Neurocomputing, 521, 126-136.
- Alcobaça, E., Siqueira, F., Rivolli, A., Garcia, L. P., Oliva, J. T., & De Carvalho, A. C. (2020). MFE: Towards reproducible meta-feature extraction. Journal of Machine Learning Research, 21(111), 1-5.
- Rivolli, A., Garcia, L. P., Soares, C., Vanschoren, J., & de Carvalho, A. C. (2018). Towards reproducible empirical research in meta-learning. arXiv preprint arXiv:1808.10406, 32-52.
- Paiva, P. Y. A., Smith-Miles, K., Valeriano, M. G., & Lorena, A. C. (2021). PyHard: a novel tool for generating hardness embeddings to support data-centric analysis. arXiv preprint arXiv:2109.14430.
- Paiva, P. Y. A., Moreno, C. C., Smith-Miles, K., Valeriano, M. G., & Lorena, A. C. (2022). Relating instance hardness to classification performance in a dataset: a visual approach. Machine Learning, 111(8), 3085-3123.
- Pascual-Triana, J. D., Fernández, A., Novais, P., & Herrera, F. (2024). Fair Overlap Number of Balls (Fair-ONB): A Data-Morphology-based Undersampling Method for Bias Reduction. arXiv preprint arXiv:2407.14210.
- Lorena, A. C., Paiva, P. Y., & Prudêncio, R. B. (2024). Trusting my predictions: on the value of Instance-Level analysis. ACM Computing Surveys, 56(7), 1-28.
- Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A., & Santos, J. (2023). A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research. Information Fusion, 89, 228-253.

Hands-on Tutorial

Data Centric AI: Data Complexity

https://colab.research.google.com/drive/1YLO5rHFDMfIUAMxaIDe3UYE5fabPAjwu?usp=sharing

ACKNOWLEDGEMENT

- This content was partially adapted from the tutorial “*A Multi-View Panorama of Data-Centric AI Techniques, Tools, and Applications*”, as part of the 27th European Conference on Artificial Intelligence.

- Thanks to my colleagues:
- **Miriam Seoane Santos,** University of Porto, **miriam.santos@fc.up.pt**
- **Pedro Henriques Abreu,** University of Coimbra, **pha@dei.uc.pt**

A Multi-View Panorama of Data-Centric AI

Techniques, Tools, and Applications

Alberto Fernández, University of Granada, **alfh@ugr.es**