Data partitioning strategies for simulating non-IID data distributions in the DDM-PS-Eval evaluation platform

Mikołaj Markiewicz, Jakub Koperwas

ICSOFT 2022
Lisbon, 11-13 July, 2022

Outline

Introduction

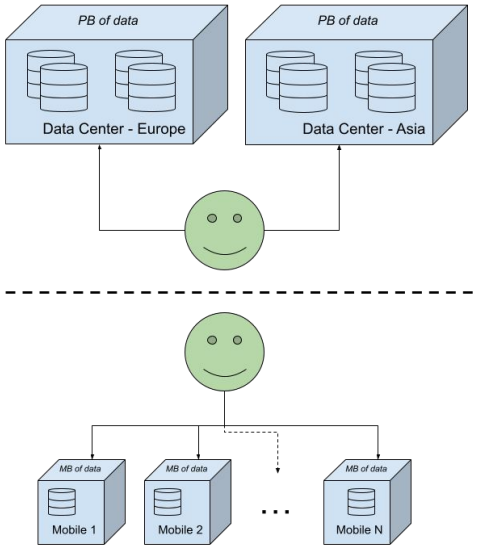
Problem statement – algorithms and distributed data

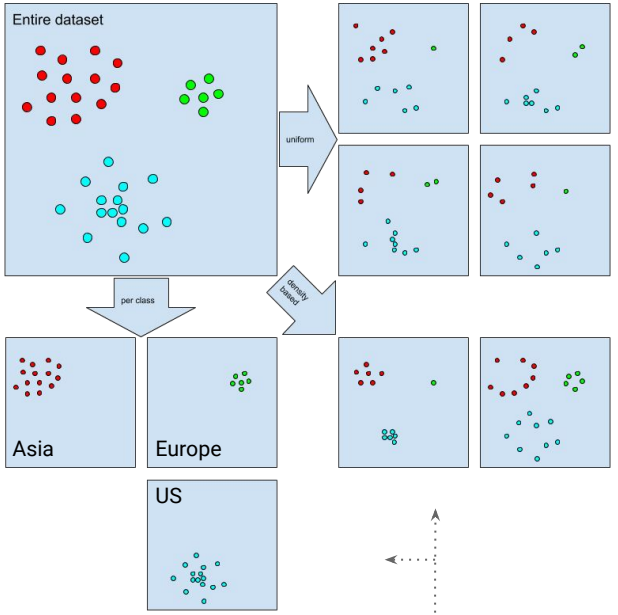
Target:

- Rapidly obtained and globally correct results

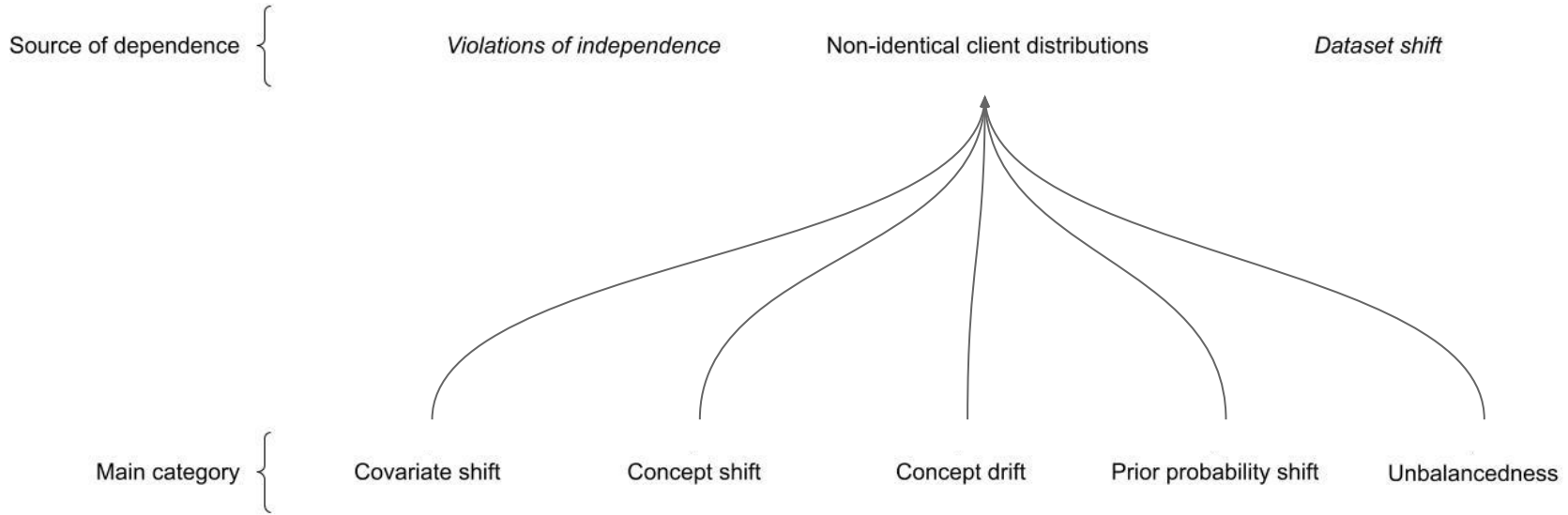
Issues:

- Large combined dataset size
- Data privacy issues
- Unknown data location
- Limiting transfer
- Unknown data distribution



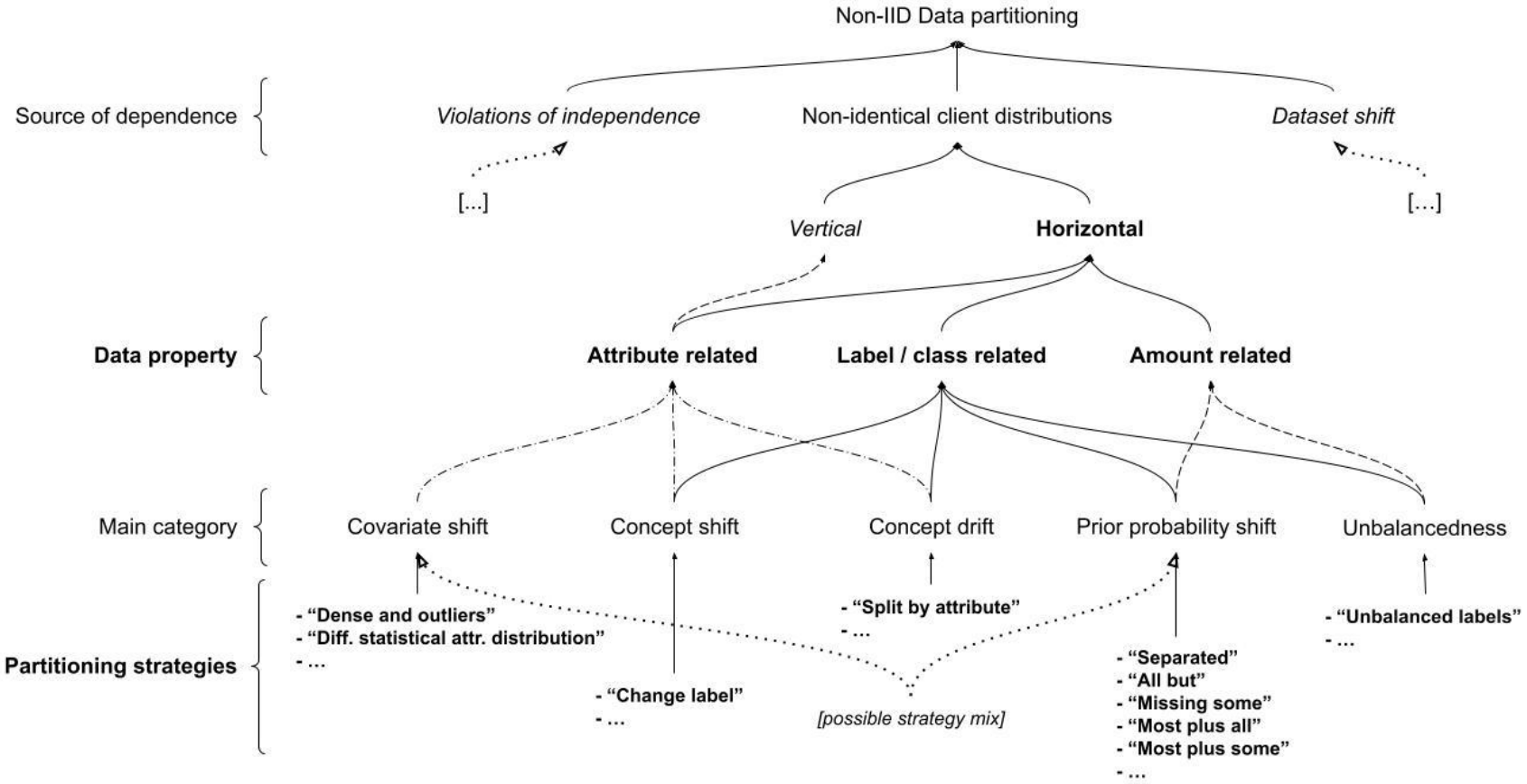


Non-IID data partitioning



Illustrated taxonomy of non-IID data regimes

(based on the work: Kairouz, P. et al., 2019. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977)

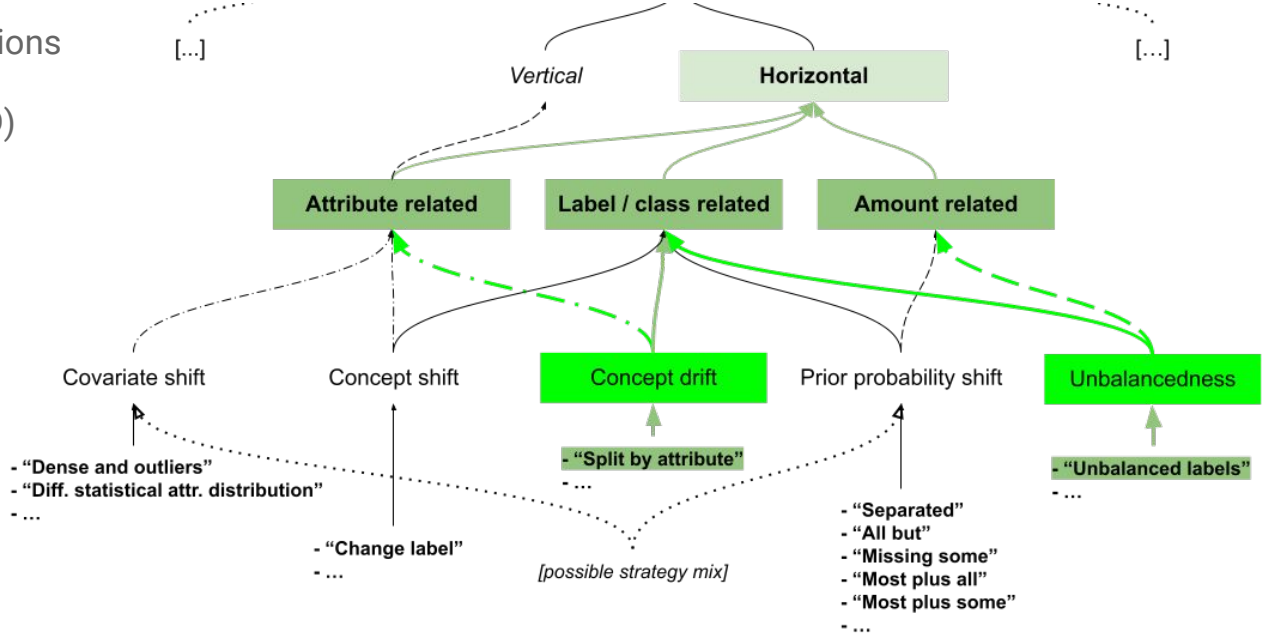


Non-IID data extended partitioning taxonomy

Minimal test suite – correct coverage

Evaluate with non-IID distributions

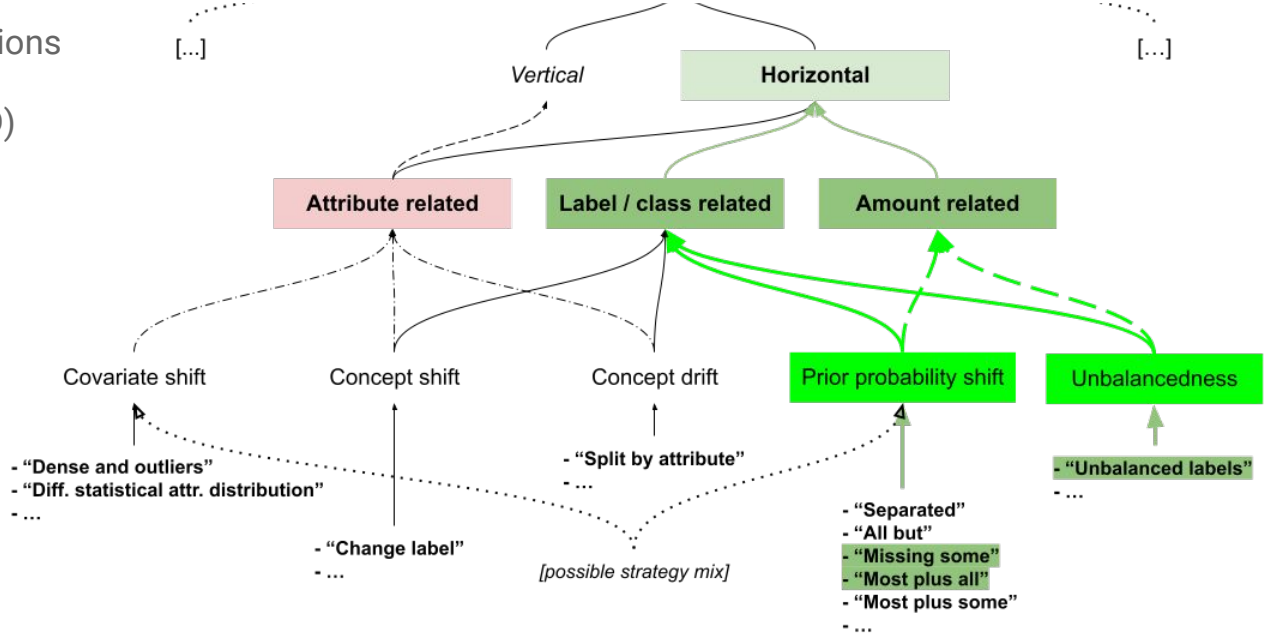
+ Uniform distribution (IID)



Minimal test suite – incomplete coverage

Evaluate with non-IID distributions

+ Uniform distribution (IID)



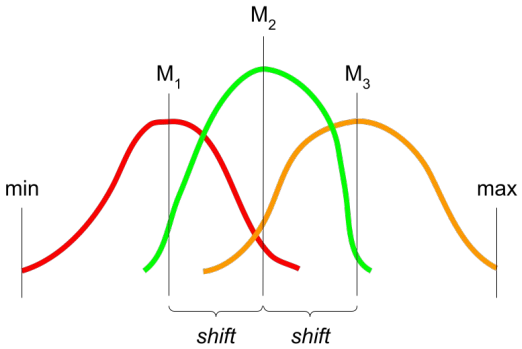
Partitioning strategies

Covariate Shift (Feature Distribution Skew)

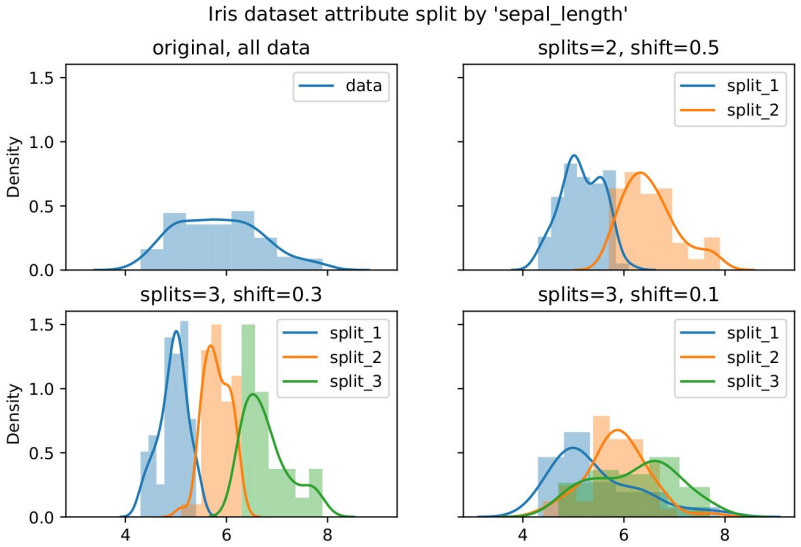
Controlling *"shift"* (skew) size:

$$\begin{cases} \sum_{n=2}^{splits} \frac{M_n - M_{n-1}}{max - min} = (splits - 1) \cdot shift \\ \forall n \quad \frac{M_n - M_{n-1}}{max - min} = shift \end{cases} \tag{4}$$

$$\wedge \quad splits > 1; \quad shift \in (0, \frac{1}{splits}]$$



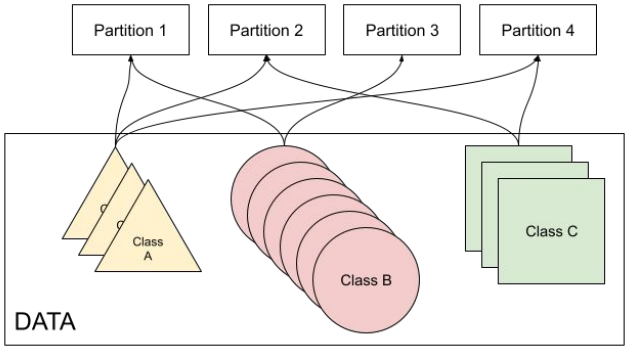
Different parameterisation results:



Prior Probability Shift (Label Distribution Skew)

Real-life examples:

- Single kangaroos in zoos around the world live wild in Australia
- A group of people from a single nation in a community emigrated from another country
- Mostly English texts in US or UK, and this language used all around the world



Different parameterisation possibilities and the 1st stage results:

```
L - labels >= D - partitions

 EmptyAdd=0
```

```
Legend:
 1st [] - majority labels
 2nd [] - additional labels
 3rd [] - empty partition labels
```

```
L - labels < D - partitions
```

```
Separated:      d1=[0, 3] + [1, 2]       + []
 D=3            d2=[1, 4] + [3, 0]       + []
 L=5            d3=[2]    + [1, 3]       + []
 Add=0

All but:        d1=[0, 3] + [1, 2]       + []
 D=3            d2=[1, 4] + [3, 0]       + []
 L=5            d3=[2]    + [1, 3, 4]    + []
 Add=-1

Most plus some: d1=[0, 3] + [1, 2]       + []
 D=3            d2=[1, 4] + [3, 0]       + []
 L=5            d3=[2]    + [1, 3]       + []
 Add=2

Most plus all:  d1=[0, 3] + [1, 2, 4]    + []
 D=3            d2=[1, 4] + [0, 2, 3]    + []
 L=5            d3=[2]    + [4, 1, 3, 0] + []
 Add=all
```

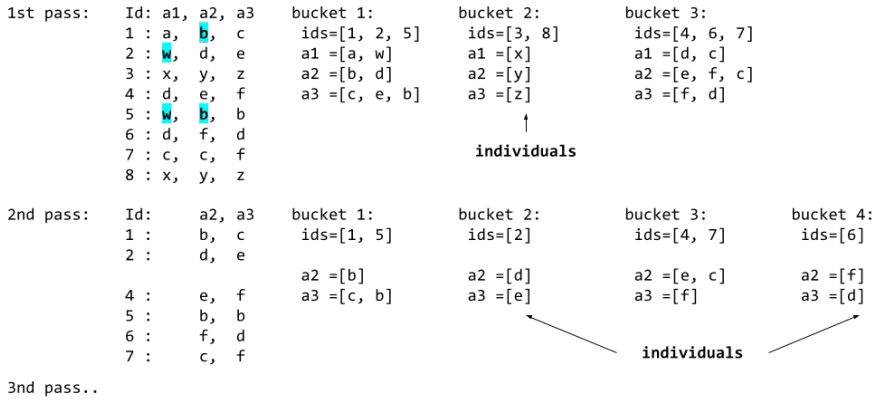
```
Separated:      d1=[0] + [] + []
 D=5            d2=[1] + [] + []
 L=3            d3=[2] + [] + []
 Add=0          d4=[]  + [] + [0, 1, 2]
 EmptyAdd=all   d5=[]  + [] + [0, 1, 2]

Separated:      d1=[0] + [] + []
 D=5            d2=[1] + [] + []
 L=3            d3=[2] + [] + []
 Add=0          d4=[]  + [] + [0, 1]
 EmptyAdd=2     d5=[]  + [] + [2, 0]

All but:        d1=[0] + [1] + []
 D=5            d2=[1] + [2] + []
 L=3            d3=[2] + [0] + []
 Add=-1         d4=[]  + [2] + [0, 1]
 EmptyAdd=2     d5=[]  + [1] + [2, 0]
```

Concept Drift (Same Label, Different Features)

Simplified algorithm:

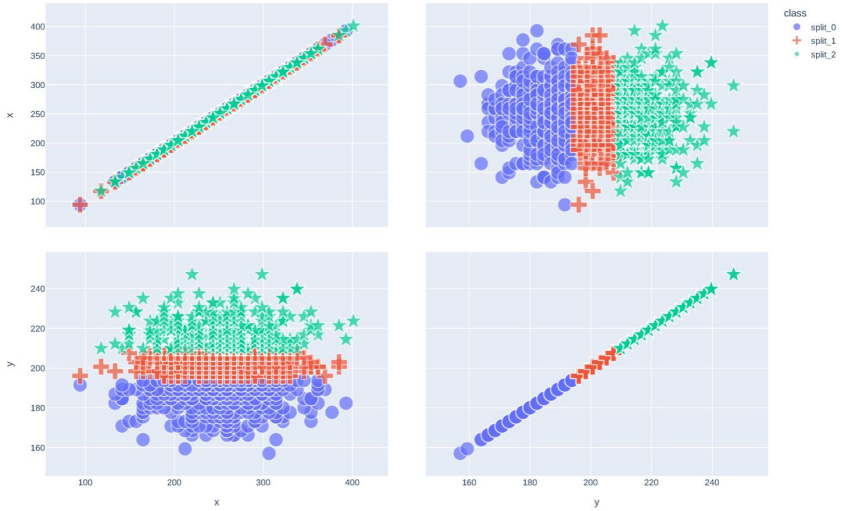
1. Extract data with chosen class,
2. Discretize numeric attributes,
3. Group data into *buckets* with the same attributes,
4. Find *“individuals”* in all-nominal data,
5. Check if the *“drift”* condition is fulfilled,
   - 5.1. Perform partitioning (go to step 6),
   - 5.2. Otherwise, remove *“individuals”* and attribute with the lowest entropy and repeat step 3.
6. Divide buckets across partitions,
7. Scatter *“individuals”*.

Nominal attributes example:

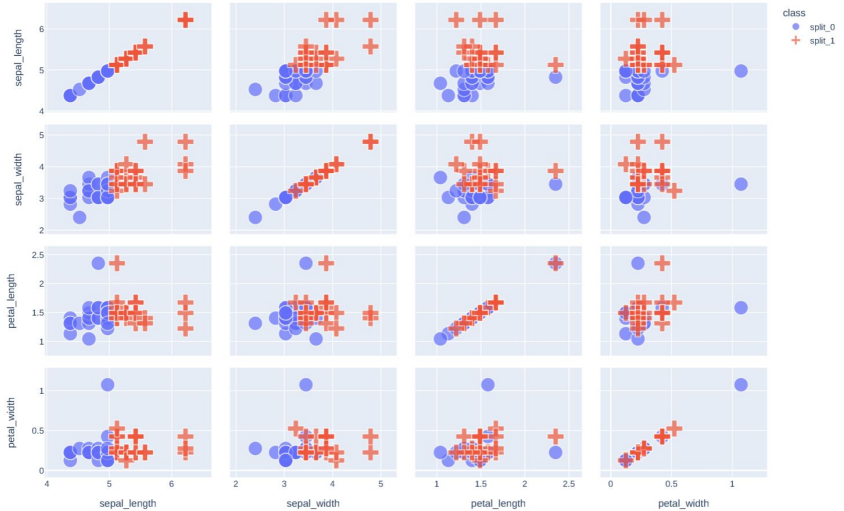


Concept Drift – sample results

3 *“drifts”* for a 2-dimensional Gaussian cluster:



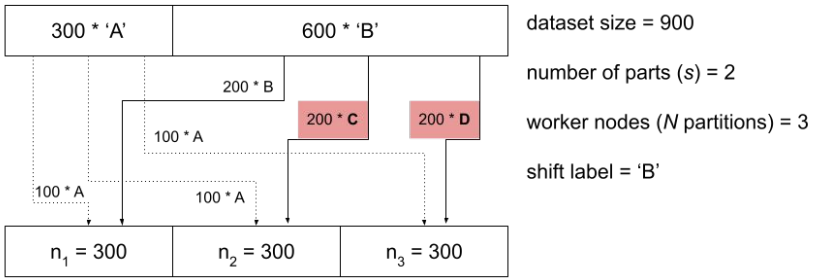
2 *“drifts”* for Iris dataset on *‘Iris-setosa’* class:



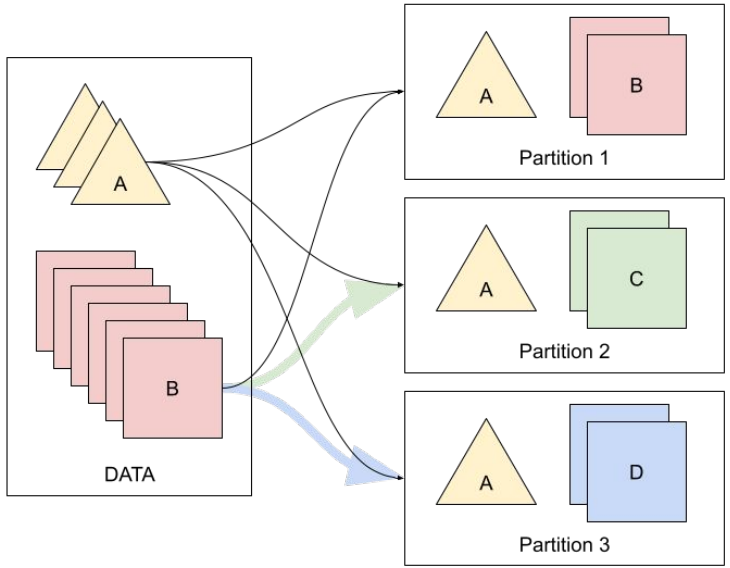
Concept Shift (Same Features, Different Label)

Creating artificial classes → *"mutate"* the label:

- The problem for supervised learning, e.g. non-fuzzy classifiers
- Transparent for unsupervised learning, e.g. clustering

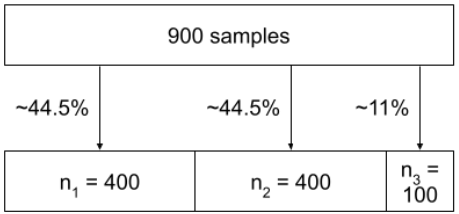


Example:



Unbalancedness (Quantity Skew)

Trivial approach example:



dataset size = 900

worker nodes (*N* partitions) = 3

large partitions quantity = 2

target ratio = ¼ = 100 / 400

```
C = 900
W = 3
T = 2

U = 0.25
```

Controlling unbalancedness (skew) ratio:

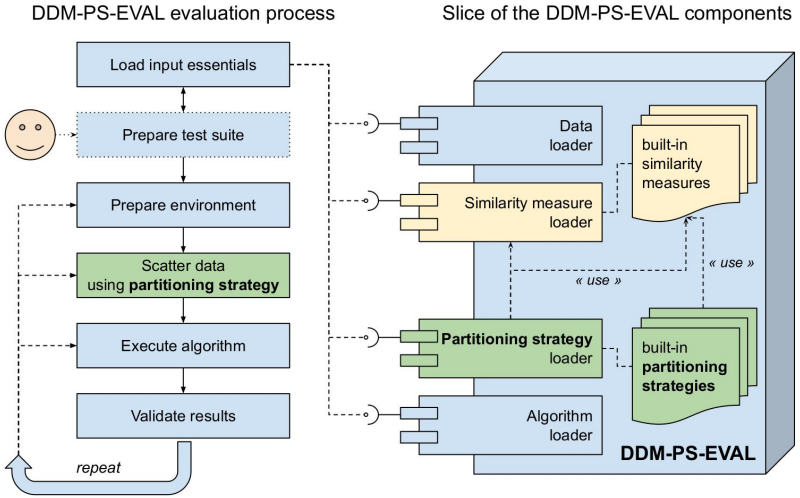
$$\begin{cases} T \cdot L + (W - T) \cdot S = C \\ \\ \frac{S}{L} = U \end{cases} \Longrightarrow \begin{cases} L = \frac{C}{U \cdot (W - T) + T} \\ \\ S = \frac{C - T \cdot L}{W - T} \end{cases} \quad (5)$$

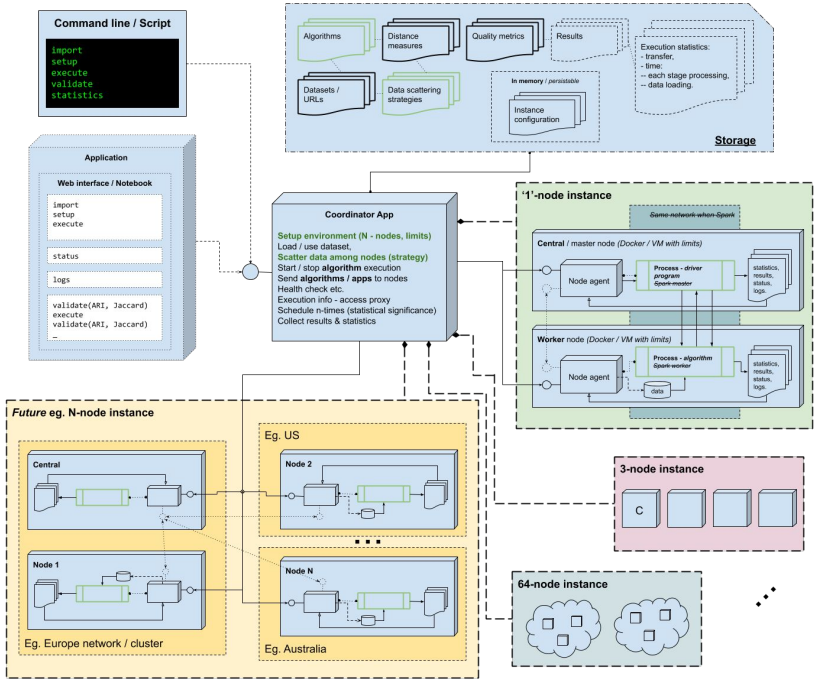
```
L = 400
S = 100
```

DDM-PS-Eval platform

Platform architecture – partitioning component

A simplified version with the most interesting components highlighted:





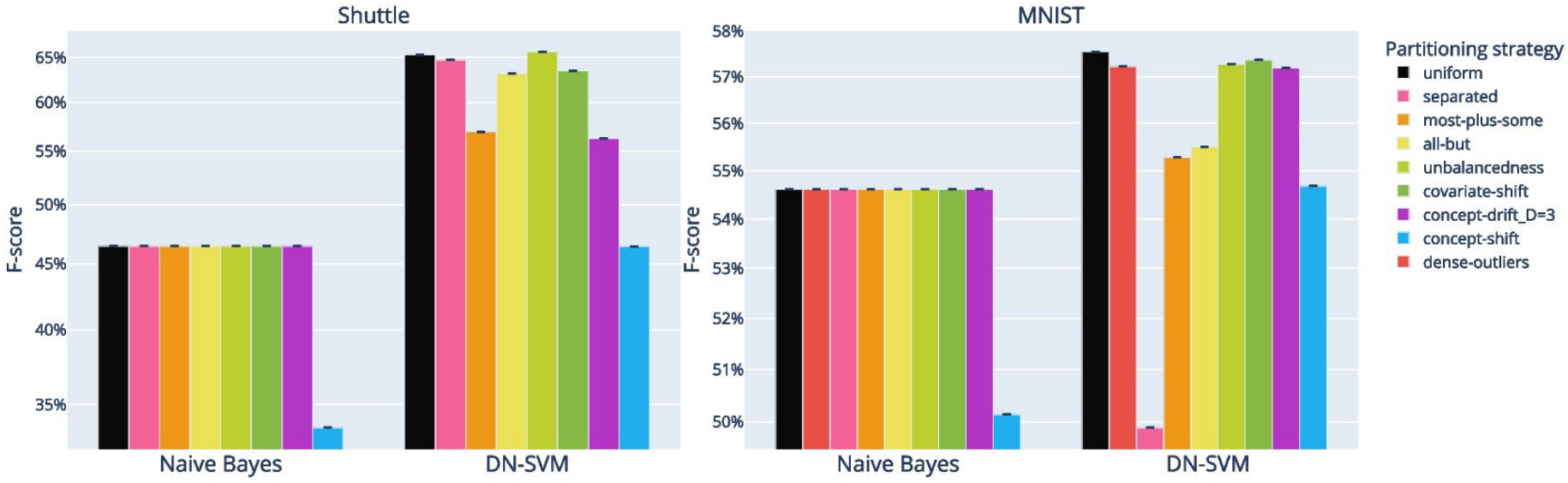
(Markiewicz, M. and Koperwas, J. (2022). Evaluation platform for ddm algorithms with the usage of non-uniform data distribution strategies. International Journal of Information Technologies and Systems Approach (IJITSA), 15(1):1–23.) The high-level design of the overall system architecture.

Experiments & results

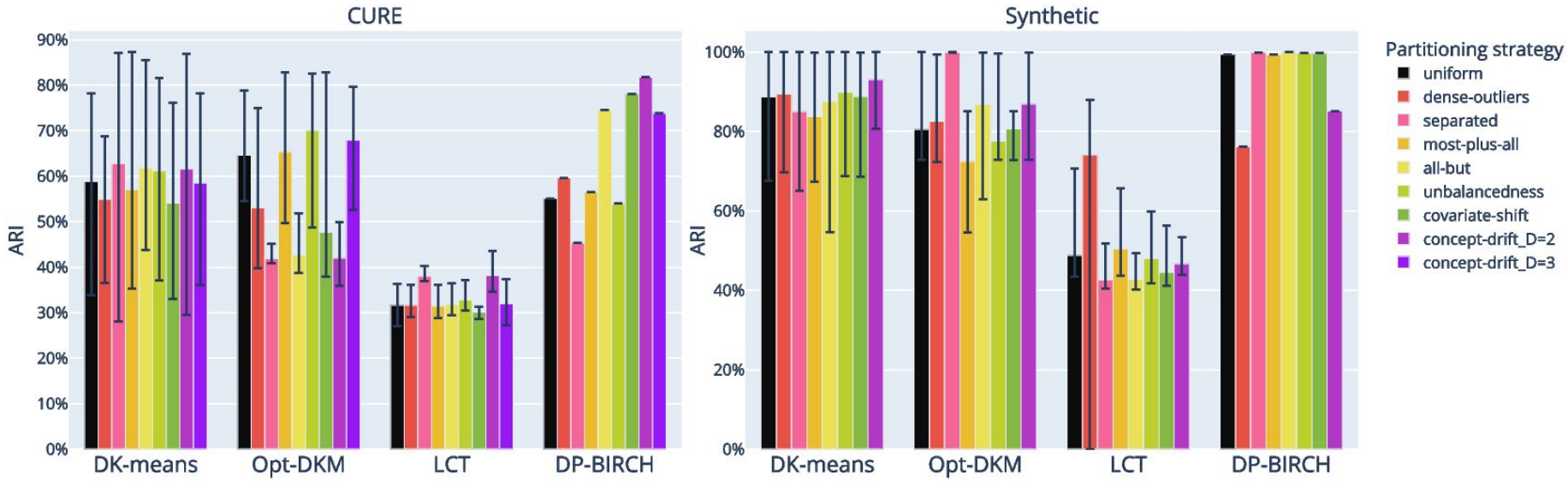
Experimental configurations

Data (classes/groups, number of samples)	Shuttle (7, 58.000), MNIST (10, 70.000), CURE (5, 2.000), Synthetic-Gaussians (7, 100.000)
Algorithms	Naive Bayes, DN-SVM, DK-means, Opt-DKM, LCT, DP-BIRCH
Worker nodes / partitions	Three workers for the first three literature datasets Four workers for the synthetic dataset
Partitioning strategies (parameterised separately for each dataset)	uniform, dense-outliers, covariate-shift, separated, most-plus-some, most-plus-all, all-but, concept-drift, concept-shift, unbalancedness

Quality results for classification



Quality results for clustering



(Negative) Impact on the results quality

Main non-IID category of data partitioning strategy	Naive Bayes	DN-SVM	DK-means	Opt-DKM	LCT	DP-BIRCH
Covariate shift	0	0 - L	0 - M	**0 - H**	**0 - H**	**0 - H**
Concept shift	**M - H**	**H**	0	0 - L	**0 - H**	0 - L
Concept drift	0	0 - M	0 - L	**0 - H**	**0 - H**	**L - H**
Prior probability shift	0	**0 - H**	0 - L	**0 - H**	**0 - H**	0 - M
Unbalancedness	0	0	0 - L	0 - L	**0 - H**	**L - H**

L, M, H indicate low (<5%), medium (<10%), high (≥10%) impact, respectively, and 0 indicates no noticeable impact or results that are better than those obtained for uniform data distribution

Sample time processing results (ms)

MAX - MNIST			
Impact	Algorithm	Strategy	Chart DN-SVM
0		uniform	107,069
H		dense-outliers	149,673
0		separated	20,245
0		most-of-one-plus-some	45,745
0		most-plus-all	67,137
0		all-but	54,878
H		unbalancedness=1	151,959
H		unbalancedness=0	149,228
0		covariate=2	98,239
M		covariate=3	113,235
0		concept-drift=2	107,318
0		concept-drift=3	104,365
H		concept-shift	119,338

MAX - Synthetic			
Impact	Algorithm	Strategy	Chart Opt-DKM
0		uniform	1,984
H		dense-outliers	2,347
H		separated	2,286
H		most-of-one-plus-some	2,298
L		most-plus-all	2,044
H		all-but	2,388
H		unbalancedness=1	2,535
H		unbalancedness=0	2,468
M		covariate=2	2,136
M		covariate=3	2,103
H		concept-drift=2	2,419
M		concept-drift=3	2,175
H		concept-shift	2,324

Conclusions

Conclusions and Future Work

- ★ Introduced extended non-IID data partitioning taxonomy
- ★ Presented new data partitioning methods simulating non-IID data dispersion
- ★ Illustrated partitioning impact on the distributed algorithms processing results
- ★ Extended DDM-PS-Eval platform with new partitioning strategies implementations

- ❑ Create data partitioning schemes for different dataset types, such for example textual datasets
- ❑ Prepare more realistic distributions of non-IID data by mixing multiple partitioning strategies
- ❑ Develop more sophisticated scattering of data with few labels for a large number of partitions
- ❑ ...

Q&A