OML02 Re-validation Results

To get some insights about the validity of the OML02 sample for training and testing the dataset level models, we presented the same datasets to 5 independent judges who went through the datasets, did the same exercise of examining the textual description of each dataset and finally selected one of the 79 subject-areas we detected or assigned another subject-area which they find more appropriate. We present the instructions sent to those judges in Appendix 1 attached to this letter. The description of the demographics of those judges, the amount of time it took them to complete the annotation task and the percentage of agreement between the annotations by the judges and the ground-truth is given in Table 1 below. Overall, there is an average of 73.5% matching annotations between the reviewers and the ground-truth, considering each person individually.

Table 1: Description of the OML02 ground-truth reviewers and percentage of agreement

No.	**Gender**	**Degree**	**Specialisation**	**Duration**	**% Agreement**
1	M	B.Sc.	Pharmacy	5h 40min	87.6
2	M	PhD	Telecom. Engineering	2h 20 min	67.8
3	F	B.Sc.	Pharmacy	7h	66.8
4	F	PhD	Pharmacy	3h 30min	68.8
5	M	B.Sc.	Pharmacy	4h 45min	76.7

If we consider the majority of annotations by the 5 independent judges and the ground-truth annotations we already have in the paper , we can calculate the number of annotations which had the majority subject-area annotation by all of them (the 5 judges and the ground-truth, i.e., 6 votes in total) matching or not with those annotations in OML02 used in the experiment. This is given in Table 2 below. The first column indicates the highest number of votes achieved by a single annotation for the dataset, the second column indicates if there was a single annotation with this number of votes or if there were ties, the third column indicates whether the majority annotation led to the same one in OML02, or if there were ties which include the correct one and incorrect ones (labelled as “yes and no”) or whether none of the top majority annotations (“no”) were the same with OML02. The fourth column shows the number of datasets matching the properties described in the row, and the fifth column shows this as a percentage from the total number of datasets.

Table 2: Number of agreement between majority annotation from the independent judges and the ground-truth and the OML02 ground-truth annotations

Number of votes	Single absolute majority?	Ground-truth matching?	Number of datasets	% of datasets
6	Yes	Yes	70	34.5
5	Yes	Yes	51	25.1
4	Yes	Yes	40	19.7
3	Yes	Yes	24	11.8
3	No	Yes and no	5	2.5
2	Yes	Yes	4	2.0
2	No	Yes and no	7	3.4
2	No	No	1	0.5
1	No	Yes and no	1	0.5

As could be summarised from Table 2, a total of 93.1% of datasets had a single absolute majority annotation matching the one in the OML02 ground-truth (where “ground truth matching?” = “Yes”). Therefore, the vast majority of the datasets had a majority of reviewers in agreement with the annotations in the ground-truth, which indicates the overall validity of the annotations in OML02 without substantial disagreement. Those datasets had a clear subject-area based on their textual descriptions. We present examples of cases of datasets having mixed “yes and no” or “no” agreement in Table 3, which we have revised to check the reason for them. Their annotation in the OML02 ground-truth is given in the “OML02 annotation” column. The reason is listed in the last column, where “fuzzy” indicates that it is difficult to decide a single subject-area for the dataset (i.e., it is more subjective and can be different from the perspective of different persons) and “error” means that the 5 judges were not able to correctly indicate the correct subject due to flawed analysis of the textual description in our opinion (based on the revision of the textual descriptions found at the “link” column). The found majority annotations are also given, where “GT” indicates that the ground-truth annotation was found among the majority proposed annotations, and “other” indicates other annotations found. We note here that the reliability of the annotations by the 5 judges (volunteers) can not be fully guaranteed as they spent much less time (see Table 1) compared to the 2 annotators who labelled the datasets of OML02 (who spent more than 15 hours discussing each dataset and its description case-by-case), and they were probably not paying their utmost effort like the main annotators of OML02 who did a more dedicated effort as expressed by some errors in the cases in Table 3. We note that those cases only represent less than 7% of the number of datasets in OML02.

Table 3: analysis of the non-agreement cases by all the judges and the annotators with the OML02 ground-truth

Dataset	Link	OML02 annotation	Number of votes	Single absolute majority?	Found majority annotations	Reason
haberman	https://www.openml.org/d/43	Disease	3	Yes and no	GT: Disease Other: Hospital statistics	error
space_ga	https://www.openml.org/d/507	voting demographics	3	Yes and no	GT: voting demographics Other: Geographical measurements	error
vertebra-column	https://www.openml.org/d/1523	Disease	3	Yes and no	GT: Disease Other: Human Bones Measurements	fuzzy
covertype	https://www.openml.org/d/1596	Plant measurements	3	Yes and no	GT: Plant measurements Other: Geographical Measurements	fuzzy
ozone-level-8hr	https://www.openml.org/d/1487	Geographical Measurements	2	Yes and no	GT: Geographical Measurements Other: Pollution Measures, Gas sensing statistics	fuzzy
colic	https://www.openml.org/d/25	Health Measurements	2	Yes and no	GT: Health Measurements Other: Animal Profile, Hospital Statistics	fuzzy
SMSA	https://www.openml.org/d/1091	City Census Data	2	Yes and no	GT: City Census Data Other: Pollution Measures	fuzzy
ldpa	https://www.openml.org/d/1483	Motion patterns	2	Yes and no	GT: Motion patterns Other: Health	error

Dataset	Link	OML02 annotation	Number of votes	Single absolute majority?	Found majority annotations	Reason
					Measurements	
Physical_Activity_Recognition_Dataset_Using_Smartphone_Sensors	https://www.openml.org/d/4675	Motion patterns	2	No	Other: steel, signal measurements	error fuzzy

Thus, we believe that the amount of noise that might be introduced in OML02 by the fuzzy cases is minimal and within acceptable proportions, and should not therefore interfere with the performance of the models in the experiments. Yet, our results shows the robustness of our proposed approach and the models to achieve high performance exceeding 90% recall even with the existence of some noise (fuzzy cases) in the ground-truth (this is better than best performance in terms of agreement by the human judges from Table 1). To further investigate on that point, we reran the top performing models with the optimum similarity thresholds and using only the 93% of datasets having majority votes by judges and annotators and found that the evaluations metrics only vary slightly within +/-2%, which is insignificant.

In essence, the annotation task is not simple and is even difficult for human beings to complete, resulting from some fuzzy cases and as seen by some errors made by the judges. This explains the reason why we decided that both annotators had to do this annotation exercise together, so that they can take their time in discussing the details of each dataset and to reach consensus regarding a single subject-area they think best matches the dataset, yet without studying the underlying data stored in their attributes to prevent any bias towards our proposed algorithms. Both annotators did not have any previous knowledge about the datasets.

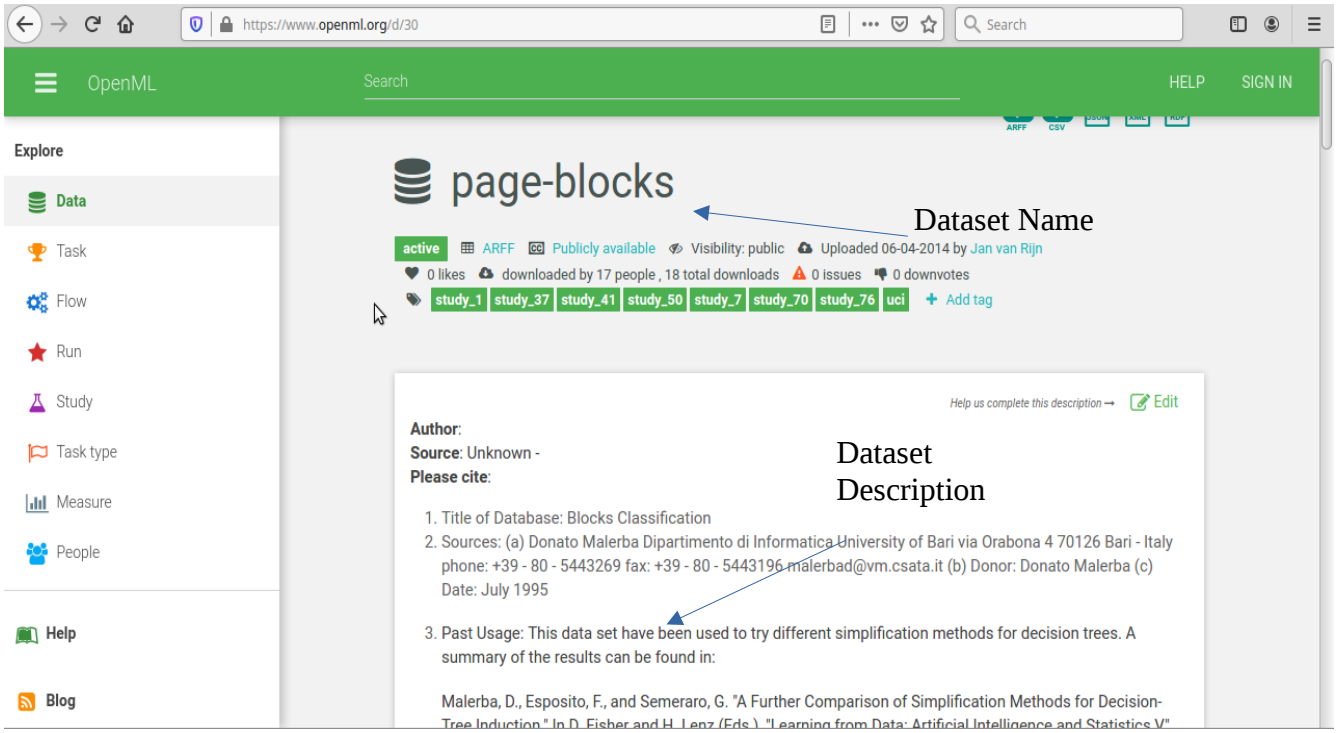
Appendix 1: OML02 Dataset Entity Description Analysis Instructions

Task

We have a group of datasets from an online repository called OpenML. It stores different datasets from multiple sources having information about different things (what we call "*entities*"). **Entities** are real-world concepts which could be described or measured like "car fuel prices", "plant measurements", "buildings", "shoes", "student grades", etc. We would like to identify the ***specific*** entity described or measured in the datasets using their textual description (see Fig. 1). In the attached Excel sheet, we present to you specific datasets that we would like to analyse. The task required is as follows:

1. Open the attached Excel sheet. Each row consists of a dataset from OpenML. We provide its ID and name in columns A and B. You only need to fill column D (or E) and optionally column F at the end of the sheet. ***Please time the total duration it takes to fill the Excel sheet***.

Fig. 1: Dataset page – Name and description



2. For each dataset row, open the online description page (column C: "dataset link"), read carefully its short description which describes what the dataset stores (and sometimes how the data was collected too) and decide what is the the most specific entity it talks about. Choose the best matching one from the given pre-defined entities in the drop-box in column D (this is a fixed given list of the specific entities you can choose from, which are the same for all datasets. The list can also be found in the sheet "Entities list" too, from the tabs at the bottom). If none of the options match your preference, then write your desired entity in (column E: "other_entity"). If you cannot understand the description or can't identify the entity from the description, then enter "unknown" in column E. You can optionally (as needed) add any extra comments or feedback to us using column F. We advise you to first read (go through) the whole list of options for the entities before completing this task so you can make sure you select the ***most relevant and specific option*** for each dataset.