7th European Business Intelligence Summer School (eBISS 2017) – Brussels, Belgium
Doctoral Colloquium

Information Profiling in the Data Lake – *Using Data Mining techniques*

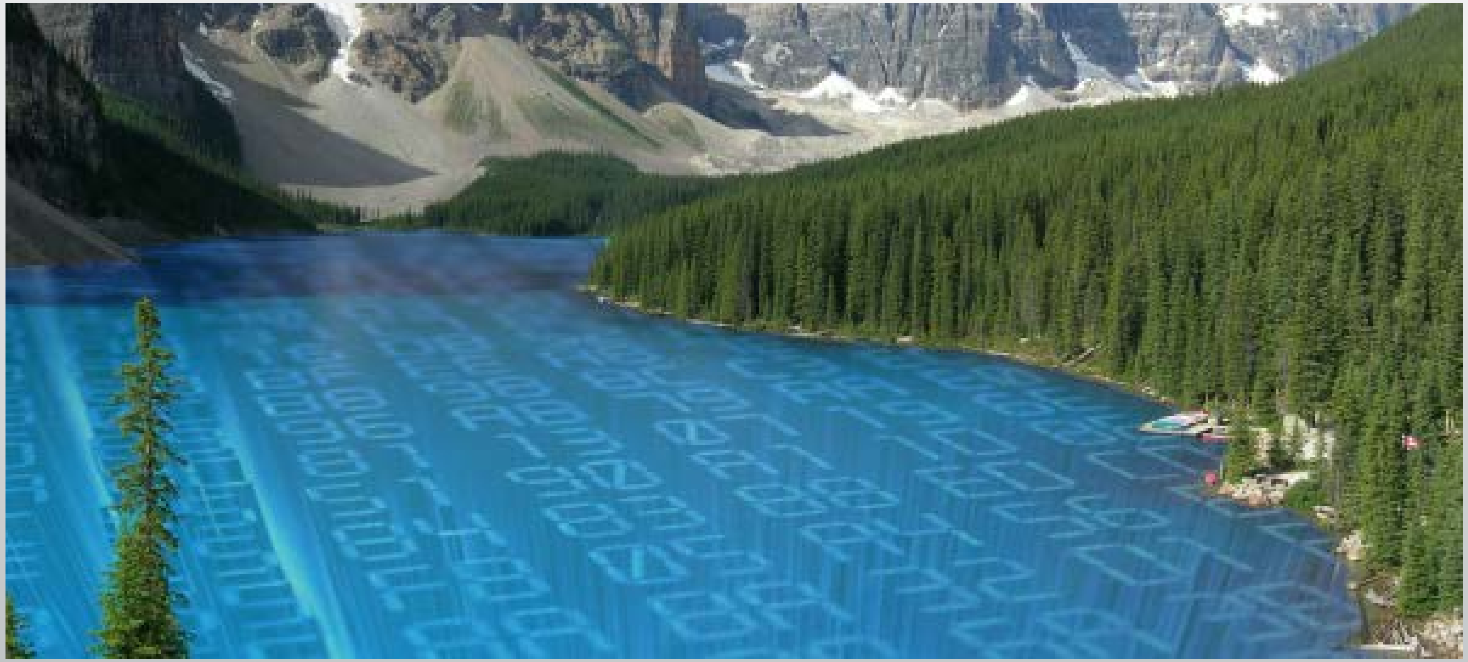
Presented by: Ayman El Serafi

Alberto Abelló , Oscar Romero (UPC)

Toon Calders (ULB, UAntwerpen)

Big Data Challenges

- **Data Lakes (DL)**: data repositories that store huge amounts of heterogeneous data in their original *raw format* :
  - Structured data (Relational DBs)
  - Semi-structured data (XML, JSON , RDF, …)
  - Unstructured text (free-text documents, e-mails, …)



Data governance: The need for metadata management

- Metadata is simply defined as: data which describes data
- To collect metadata in the DL we use **data profiling**
  - Data distributions
  - Statistics about instances and attributes
  - Unique data values

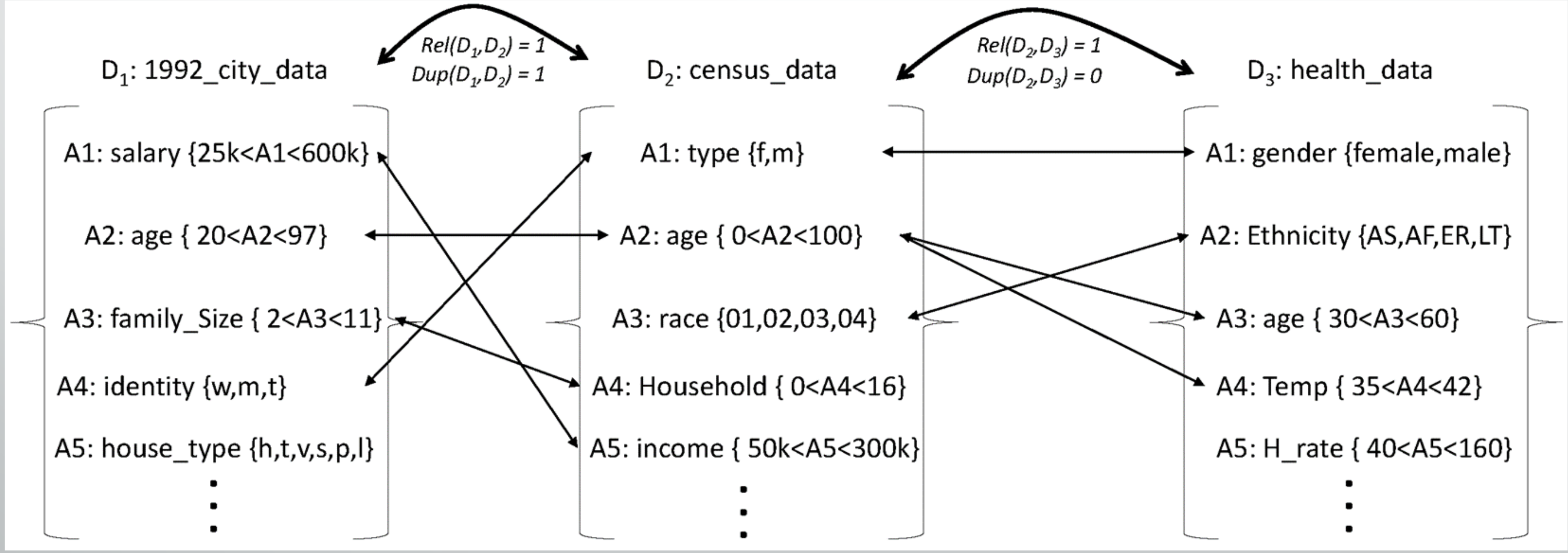
- **Schema matching + metadata** (from data profiling) = **Information profiling**
  - "Finding common information between datasets"
  - Information: the meaning of data, interpreted raw data
- There is a need for **automatic detection of patterns and relationships** in the DL for *<u>information retrieval and data analytics</u>*.

Our Goal: *automatic techniques* to generate the *mapping of data content ingested* in the data lake using metadata and data content matching

Naumann, Felix. "**Data profiling revisited.**" *ACM SIGMOD Record* 42.4 (2014): 40-49.

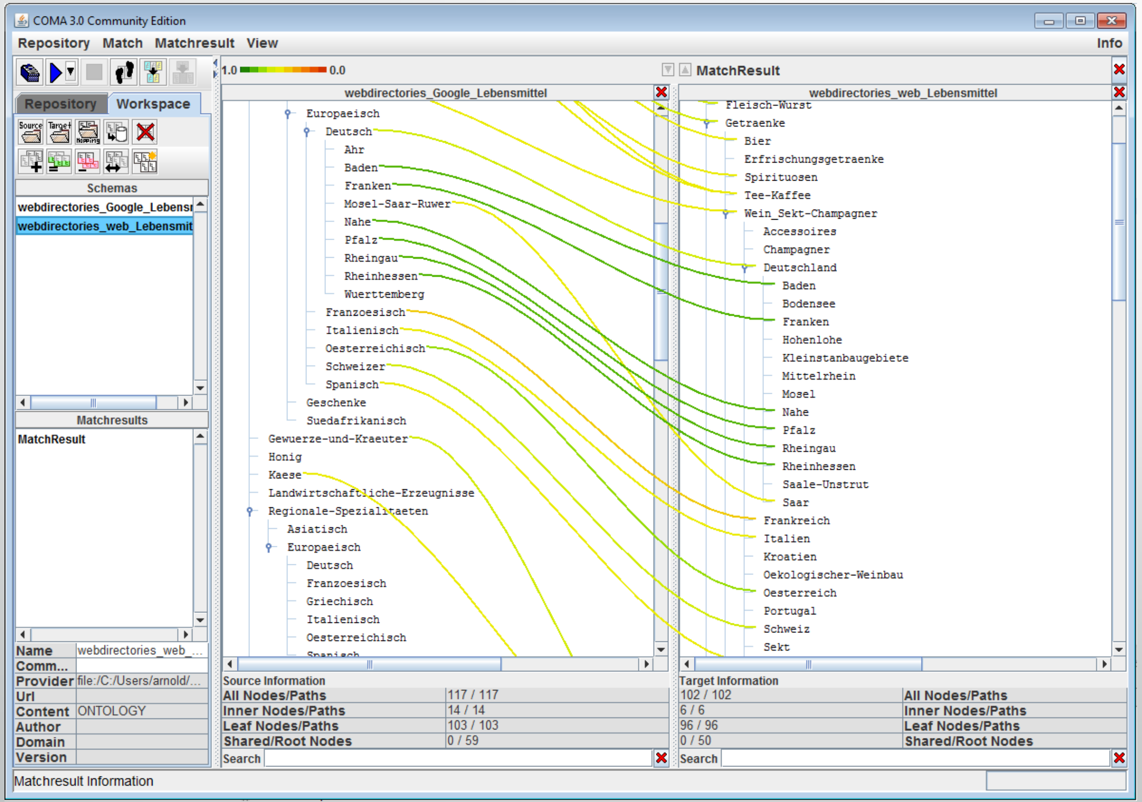
Goals

- Our goal is an organised DL where we are capable of:
  - Detecting similar schemas and relationships among data items from datasets
  - Detecting duplicate datasets
  - Detection "joinable" or "crossable" datasets
  - ...



Schema Matching

- The goal of schema matching is to align two different schemata from two data sources to find their similar items (***instances*** or ***attributes***)
- It usually seeks to find correspondences between instances in the ***2 schemata*** (also commonly called mappings)
- A special type of schema matching is Ontology alignment:



Aligning values between two ontologies (source: COMA Community Tool)

Bernstein, P. a, Madhavan, J., & Rahm, E. (2011). **Generic Schema Matching , Ten Years Later.** Proceedings of the VLDB Endowment, 4(11), 695–701. http://doi.org/10.1007/s007780100057

Holistic Schema Matching

Rahm, E. (2016). **The Case for Holistic Data Integration**. In *ADBIS* (pp. 11-27).

- Holistic schema matching seeks to match ***multiple schemata*** together (which focuses on attribute mappings), instead of pairwise 1-to-1 matching (which focuses on instance-based matching)
- Goals of applying holistic schema matching in DLs:
  - Relationships: Detecting similar attributes from datasets stored in the DL
  - Duplication: Detecting duplicate schemata which can be grouped together
  - Singleton schema: detecting those schemata which have no strong linkages with any other dataset
- Holistic schema matching leads to the following **challenges:**
  - Heterogeneous data sources with different formats & representations requiring new schema matching techniques.
  - Large amounts of schemata, where schema matching techniques need to scale up with more efficient approaches.

Our Contribution: How Can Data Mining Help?

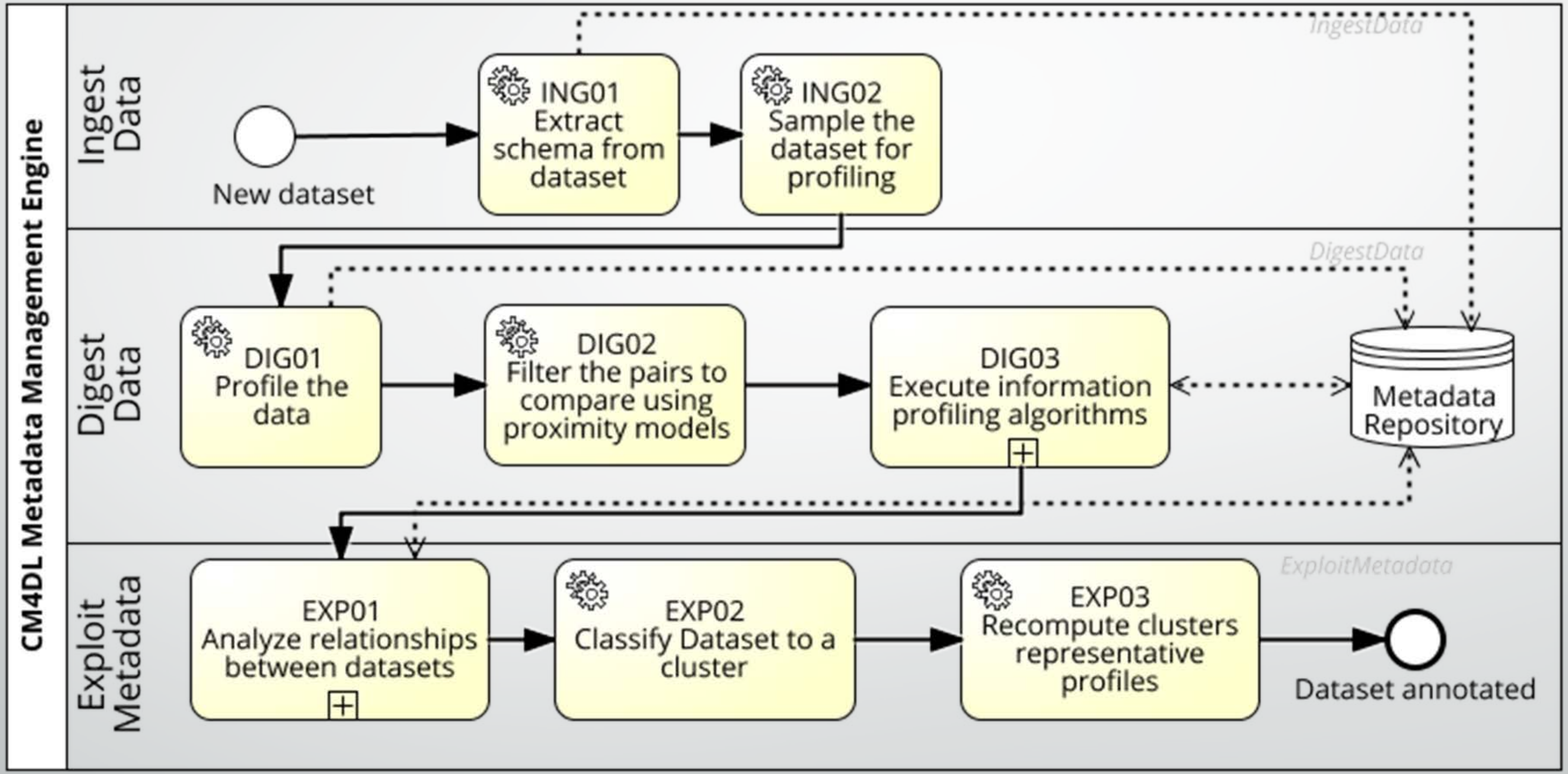
1. Finding attribute-level relationships and datasets similarity using schema matching & ontology alignment techniques

2. Using supervised machine learning (Classification models) for deciding “relatedness” between pairs of datasets for early-pruning tasks of pair-wise comparisons

3. Using clustering techniques to segment the datasets into similar groups of data that are related to each other.

4. Using frequent pattern mining to summarise different structures of data inside datasets from the DL.

Our Proposed Approach



OpenML: an online DL repository

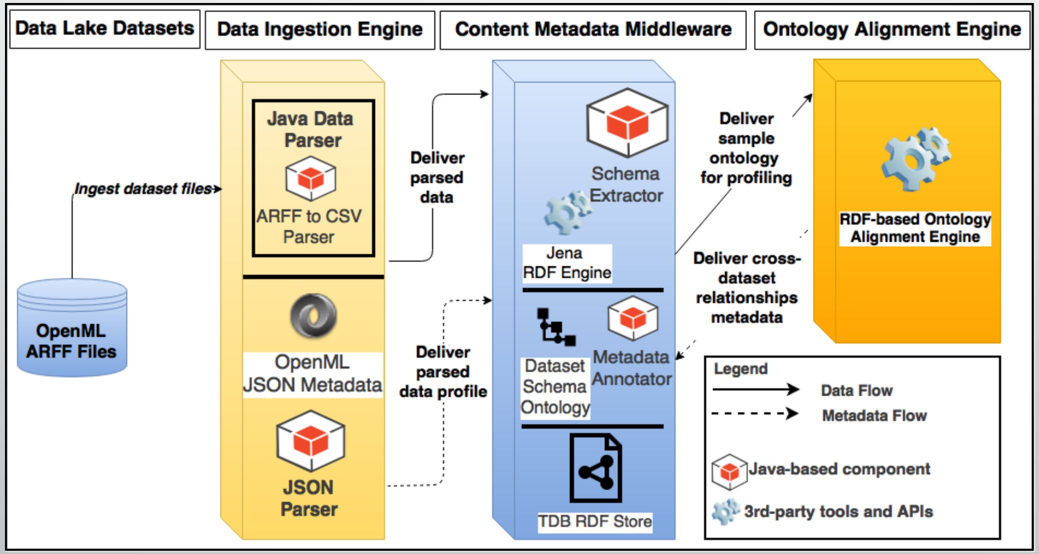
- Open Machine Learning datasets and collaborative data mining
- Datasets represented in flat tabular format
- Has more than 19500 datasets
- Data represent a huge domain of knowledge

Domain	Datasets IDs	Datasets Names
Plants	24,42,61	mushroom,soybean,Iris
Vehicles	9,21,207	autos,car,autoPrice
Business	4,29,223	labor,credit-a,stock
Sports	214,495,966	baskball,baseball-pitcher,analcatdatahalloffame
Health	35,37,51	Dermatology,diabetes,heart-h
Others	1,44,50	anneal, spambase, tic-tac-toe

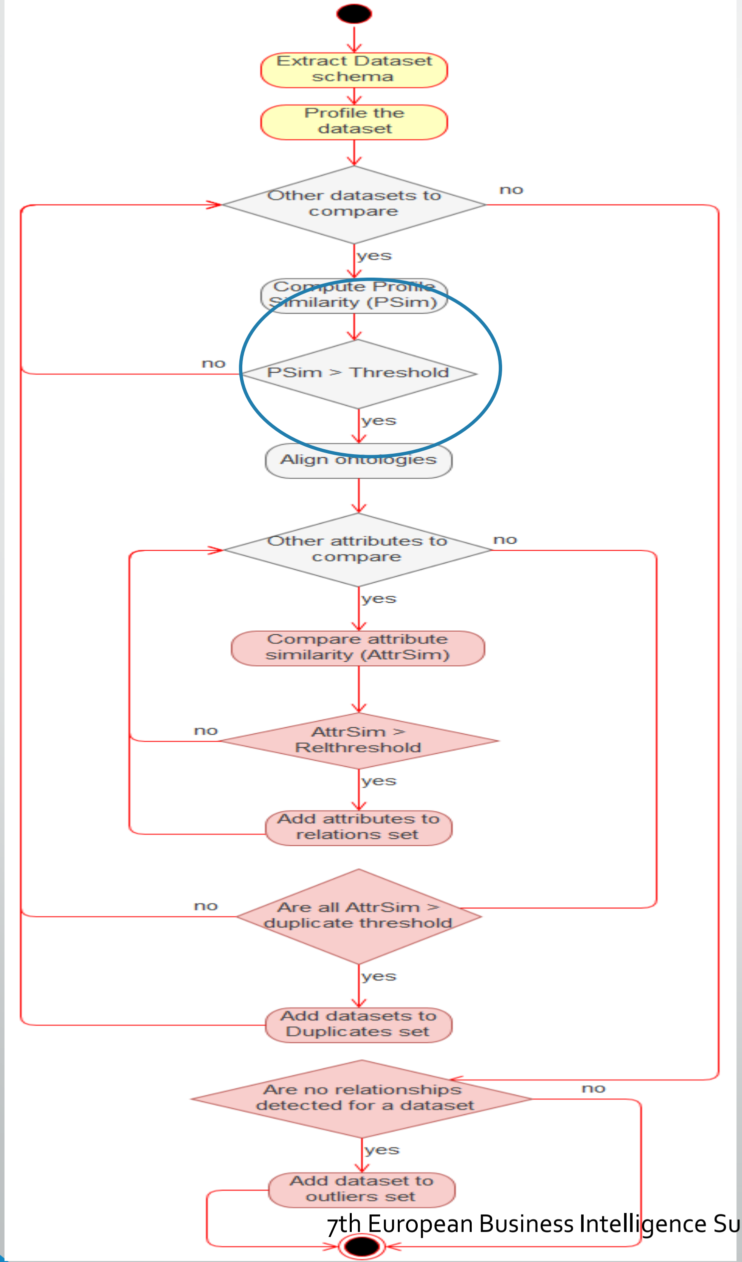
Research Paper 1: Towards Information Profiling: Data Lake Content Metadata Management

- **Topic**: information profiling based on instance-based ontology alignment techniques
- **Research question**: How to extract RDF data summaries from textual datasets like CSV, in order to use ontology alignment and matching techniques to find relationships between the datasets?
- **Hypothesis**: Ontology alignment techniques can effectively extract
- Relationships and similarity between related/duplicate datasets.
- **Approach**: proposed an approach and an algorithm for efficiently and
- effectively handling the information profiling process in the DL.
- **Experiments**: test on an annotated gold-standard from OpenML. Measure computational performance based on execution time and effectiveness based on precision, recall, and F1 scores.

Alserafi, A., Abelló, A., Romero, O., & Calders, T. (2016, December). **Towards Information Profiling: Data Lake Content Metadata Management**. In IEEE International Conference on Data Mining Workshops (ICDMW), 2016 (pp. 178-185). IEEE.



The Relationships Discovery Algorithm



1. New Datasets ingested
2. Convert dataset to RDF
3. Profile Dataset
4. Filter pairs of datasets to compare
5. Compute similarity score using ontology alignment
6. Check score: related or duplicate or singleton

Worst Case Complexity:

[d(d-1) / 2] * m² where d is number of datasets in the data lake, and m is the number of instances in each dataset*

OpenML: Applying instance-based approaches

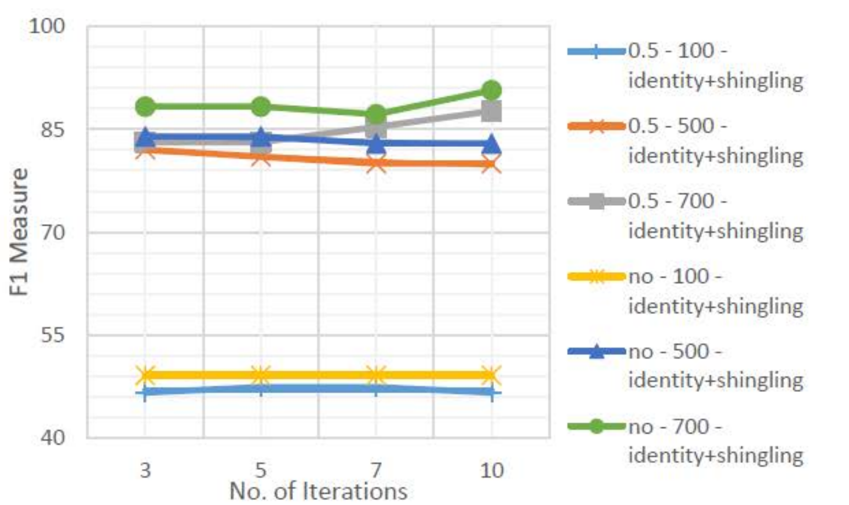
Dataset 1	Dataset 2	Similarity	isDuplicate	isRelated
9 Autos	975 Autos	0.897958	Yes	Yes width and width = 0.933814 body-style and body-style = 0.933814 bore and bore = 0.933814 city-mpg and city-mpg = 0.933814
38 sick	1000 hypothyroid	0.55089	Yes	Yes TBG_measured and "%27TBG%20measured%27 " = 0.87481 TBG and TBG = 0.87481 hypopituitary and hypopituitary = 0.87481 etc.
51 heart-h	171 primary-tumor	0.514138	No	Yes bone-marrow (yes-no) and exang (yes-no) = 0.581642 gender (m-f) and gender (m-f) = 0.408021

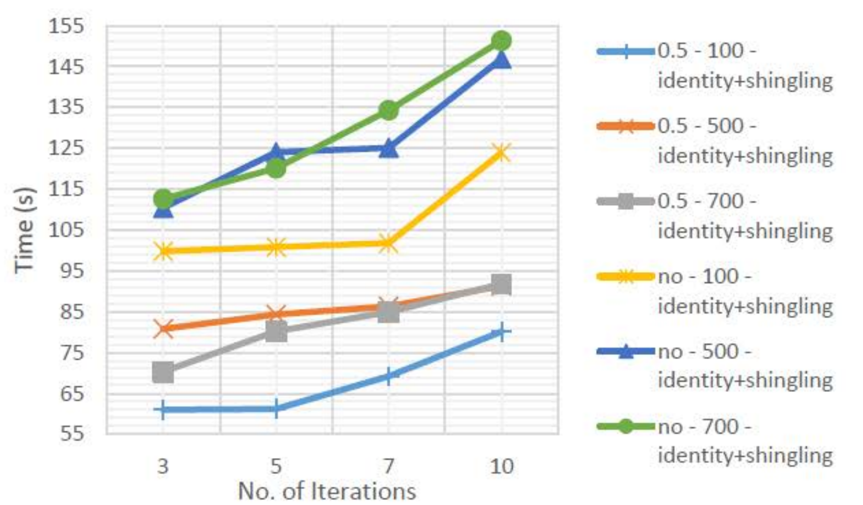
Singleton	Maximum Paris Similarity
48 tae	0.279586
50 tic-tac-toe	0
56 vote	0
40 sonar	0.328826

OpenML: Applying instance-based approaches

RESULTS	Humans (average)	Prototype
Time Duration	2 hours	60s – 151s
Precision	57.5%	76.2 – 100%
Recall	61.1%	78.9 - 89.5%
F1-score	55.6%	82 - 91%

For a sample of 15 annotated datasets, and 5 human annotators





OpenML: Applying instance-based approaches

- Conclusions
  - Automation of metadata collection and schema matching can lead to more efficient, easier, and accurate information profiling for the DL.
  - Instance based techniques are not effective in detecting similar attributes with transformed values .
  - Need to reduce processing time of ingested datasets to scale-up to hundreds, thousands, and millions of datasets.
  - Naïve approach of filtration can incorrectly eliminate positive cases that should be matched

Research Paper 2: DS-Prox: Dataset Proximity Mining for Governing the Data Lake

- **Topic**: early-pruning techniques for our information profiling approach at the dataset level using meta-features collected by profiling techniques

- **Research question**: How to use discovered content metadata for mining approximate proximity between pairs of datasets in the DL?

- **Hypothesis**: supervised machine learning techniques can effectively predict
- Relationships and similarity between related/duplicate datasets.

Alserafi, A., Calders, T, Abelló, A. & Romero, O.. (2017). **DS-Prox: Dataset Proximity Mining for Governing the Data Lake**. In Similarity Search and Applications (submitted). Springer.

- **Approach**: proposed an approach based on supervised machine learning techniques in order to mine similarity between datasets for efficiently and effectively handling the first early-pruning step of information profiling process in the DL.

- **Experiments**: test on an annotated gold-standard from OpenML. Measure computational performance based on efficiency gain and effectiveness based on precision and recall.

OpenML: Applying metadata-based approach

- New goal: reducing the computational complexity by replacing pairwise instance matching with metadata-based matching
- We focus on a novel approach for early-pruning within holistic schema matching for DLs matching using ***schema meta-features similarity mining***
- We use a **supervised machine learning** approach to compute the proximity between datasets in the DL, then we classify them into pairs of datasets which are possibly: related or duplicate datasets.
  - Experiment with multiple algorithms, like ensemble learners: AdaBoost, RandomForest, Bagging with Regression Trees, etc.
- Advantages:
  - Reduces similarity search processing, by early cheaper computations for filtration
  - Intelligently reduce the search-space without losing true-matching pairs
- Disadvantages:
  - Requires manual annotations of some training examples for the machine learning algorithm

Complexity: *[d*(d-1) / 2]*
where d is number of datasets in the data lake

OpenML: Applying metadata-based approach

1
New dataset ingested

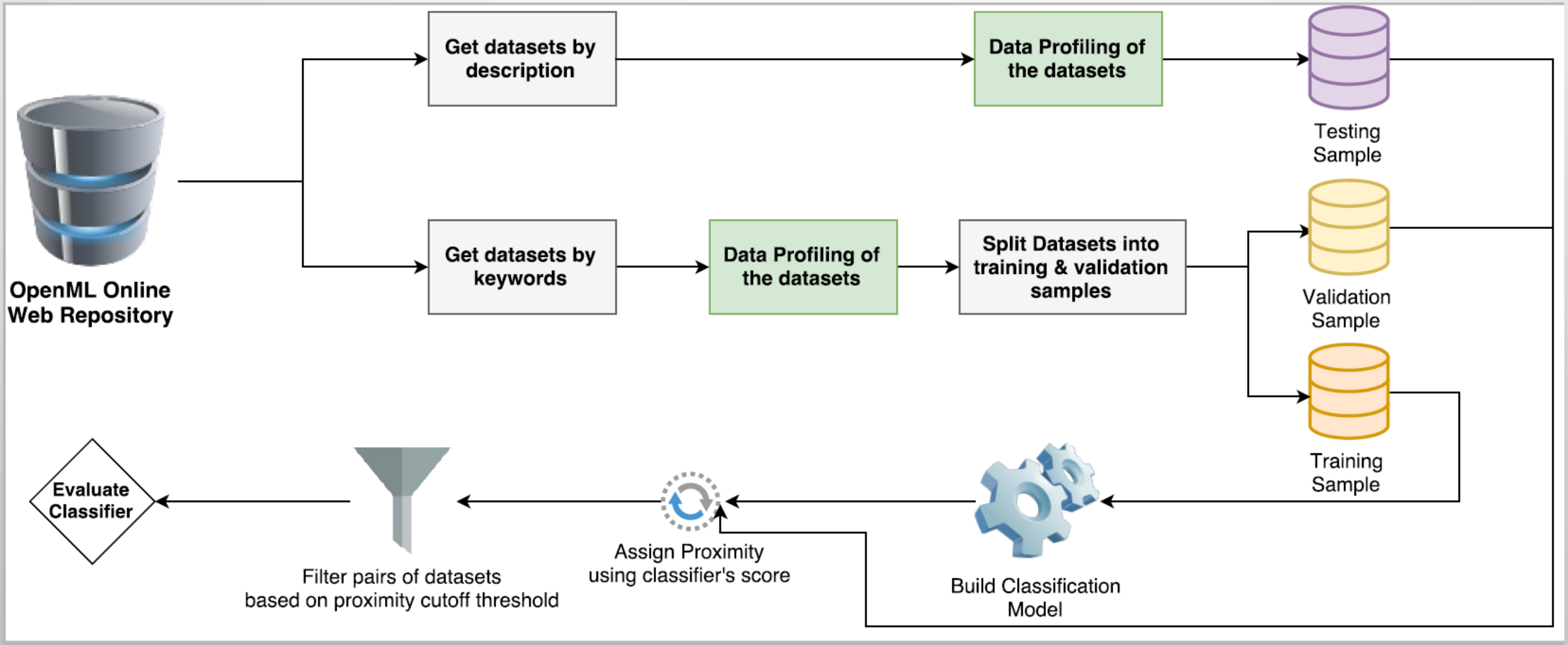
2
Profile dataset

3
Compute meta-features distances between all pairs of datasets

3
Apply classifier on all pairs of datasets to compute score

4
Filter pairs of datasets to compare based on score

5
Apply Schema matching



Experiments

- We test our approach with 540 annotated datasets from OpenML which were annotated manually based on their descriptions as:
  - Related datasets: datasets describing the same subject-area like (Diseases, Population Census, Cars, Sports, Digital Handwriting Recognition, Robot Motion Sensing, etc.)
  - Duplicate datasets: datasets describing the same subject-area , **and** where all attributes store similar information
    - This means most attributes describe the same real-world object, e.g. weight.
    - To help the in manual annotation, we use TF-IDF cosine-similarity between descriptions of datasets to find duplicates
- Note: attributes can have standardized, transformed, or differently coded data stored in them, but they still describe the same real-world object.
  - E.g. '*weight*' having a standardized value between 0 and 1 in one dataset and real values in kilograms [0,1000] in another.

- We train a classifier for each of the above 2 goals using an independent training set of 130 datasets.

Experiments

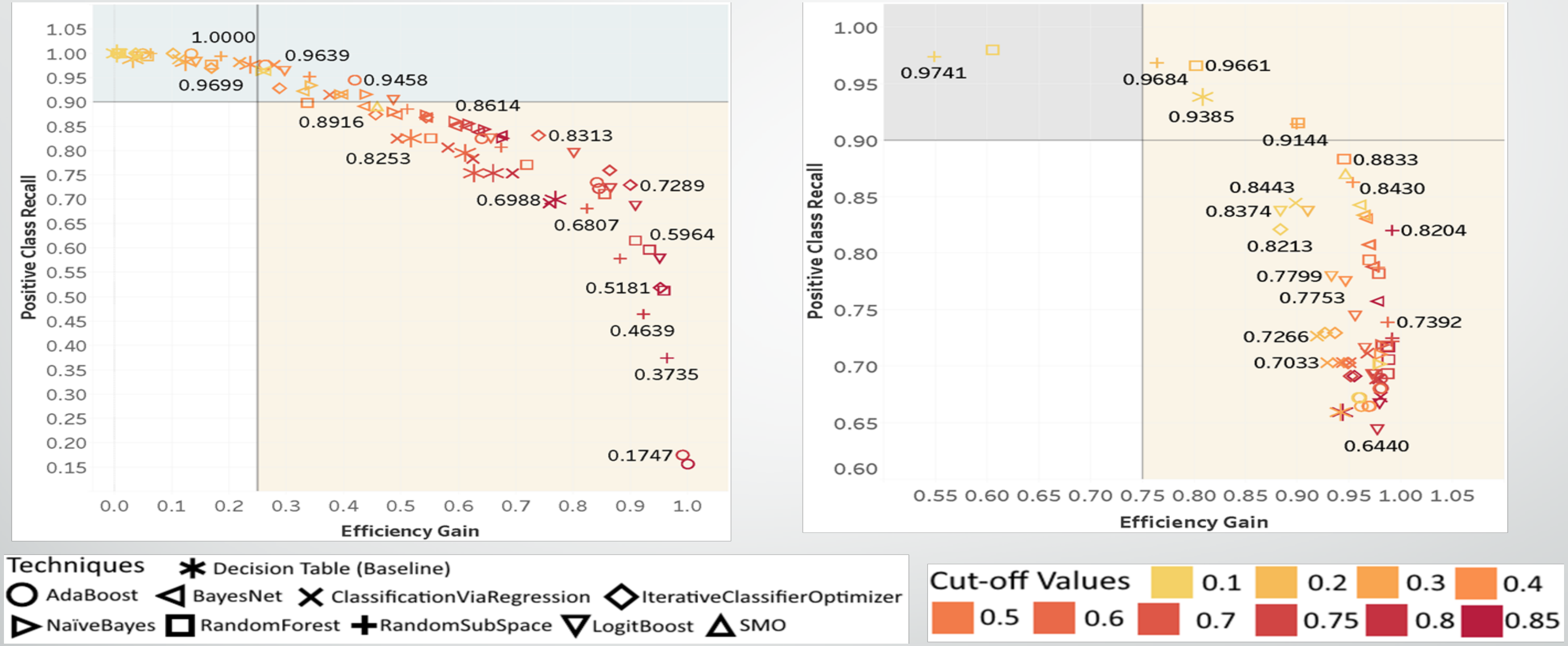
- When we apply the classifier on new pairs of datasets, we compute if they are related or not according to a ***minimum cut-off threshold*** for the score:

$$Rel(d_1, d_2) = \begin{cases} 1, & Sim(d_1, d_2) > c \\ 0, & \text{otherwise} \end{cases}$$

- Assessment measures for our early-pruning approach
  - Precision = $\frac{TP}{TP + FP}$
  - Recall = $\frac{TP}{TP + FN}$
  - Efficiency gain = $\frac{TP + FN}{N}$

Results

Rel(d1,d2) **Dup(d1,d2)**



- The approach can achieve a good optimization for the recall-efficiency trade-off
- Conclusion: our approach can only work for early-pruning to filter unnecessary comparisons, to be succeeded by more detailed matching techniques.

Research Paper 3: Keeping the Data Lake in Form: A Governance Framework using Information Profiling

- **Topic**: attribute-based schema matching, efficient similarity search algorithms and effective DL clustering techniques

- **Research questions**:

  1. How to use DM and schema matching techniques to effectively and efficiently profile the attributes in datasets to detect their relationships?
  1. What are the overall performance effects of using the framework's techniques in the DL?

- **Hypothesis**: information profiling techniques can support efficient holistic schema matching in the DL

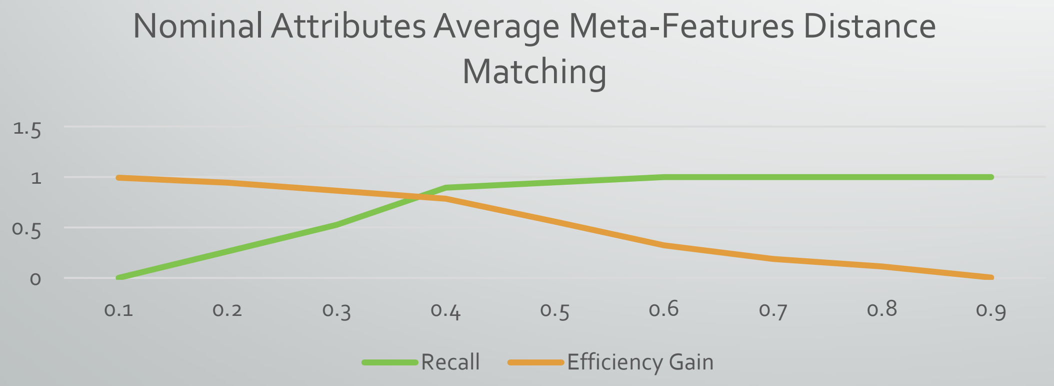
- **Approach**: applying the information profiling approach and the complete set of BPMN activities to flat tabular datasets

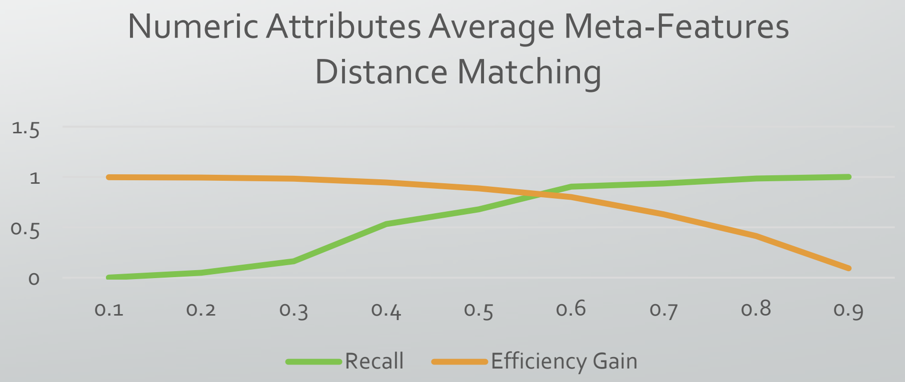
- **Experiments**: test on an annotated gold-standard from OpenML for effectiveness. Measure computational performance based on execution time and effectiveness based on precision & recall scores using huge amounts of datasets.

Preliminary Experimental Results

- Test on the annotated attributes from OpenML
  - Has 15 datasets
  - Has 62 related numeric attributes and 19 related nominal attributes

- Measure the recall and efficiency gains at different thresholds for the distance measures





Attribute-level meta-features matching

- Planned work
  - Improving numeric distance measure
  - Integrating the distance measure of all attributes for global similarity of datasets.
  - Efficiency improvement for large scale settings
    - Sampling for numeric values statistics computations
    - Indexing, ngram hashing, etc. for value comparisons
  - Possibly, construct learning model for weighting numeric and nominal distance measures

Research Status

Y= completed
P = partially completed
X = still to implement

Research Topic	**Ingestion**	**Digestion**	**Exploitation**
(01) Schema extraction from semi-structured data	Y		
(02) Data Profiling		Y	
(03) Proximity Mining		Y	
(04) Schema Matching		P	
(05) Indexing & Hashing		P	
(06) Dataset similarity metrics			P
(07) Clustering			X
(08) Metadata Visualization & Querying			P

Information Profiling in the Data Lake –

Using Data Mining techniques

THANK YOU VERY MUCH

☺

Discussion ...