DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

D5.1 Initial Evaluation of DataBench Metrics

Abstract

This deliverable is about setting-up a data and analytic infrastructure to perform technical evaluation of DataBench metrics collected in WP1-WP2. The result of this deliverable is to prepare the plan on how the technical validation will be structured, measured and reported.

Deliverable D.5.1	Initial Evaluation of DataBench Metrics
Work package	WP5
Task	5.1
Due date	31/12/2018
Submission date	04/03/2019
Deliverable lead	JSI
Version	1.0
Authors	JSI – Marko Grobelnik
Reviewers	Todor Ivanov Tomas Pariente Lobo

KEYWORDS

Big Data Validation, Knowledge Graphs, Ontology, Meta Learning

DISCLAIMER

COPYRIGHT NOTICE

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

Table of Contents

Table of Figures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

Executive Summary

This deliverable is about setting-up a data and analytic infrastructure to perform technical evaluation of DataBench metrics collected in WP1-WP2. The result of this deliverable is to prepare the plan on how the technical validation will be structured, measured and reported.

The key challenge in Big Data benchmarking is dealing with diversity of decisions that a stakeholder in a data processing related process (like data engineers, data analyst, decision makers) can make at the different stages in the analytics pipeline process. The key stages include data storage, data preparation, model construction, data and model representation, analytics, and model interpretation.

In our approach, the plan is (a) to structure the domain of Big Data process along many indicators in a form of a knowledge graph and with an upper ontological structure, (b) to make benchmarking data comparable across diverse data modalities, algorithms, tools and representations, and (c) to use the collected knowledge usable for daily tasks of data scientists end users in a form of recommendation service or interpretable aggregated knowledge.

WP5 will closely collaborate and provide technical input to the DataBench Toolbox (WP3) and provide a baseline for the evaluation exercise in WP4.

1. Introduction

Technology benchmarking, as a general matter, is about comparing different technological solutions and building blocks along different dimensions, measured via empirically observed indicators. Ideally, the measured environment is controllable to perform pivoting of different system parameters and to compare the outcomes. In the case of Big Data, the benchmarking approaches follow the same philosophy – however, an important issue is the complexity of an average Big Data project with many tuneable parameters along the stages of the pipeline. Many of these parameters can significantly change the overall performance of the executed benchmark.

Since the problem is not easy [1] and cannot be entirely solved in a clear rigorous scientific manner, we will approach the problem of an overall evaluation and validation of the collected metrics in a practical way to produce a result satisfying end-user needs.

In the next sections we will first present the overall approach in the WP5 along the three tasks, and in the continuation we will focus on the technical approach of the Task 5.1.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

2. WP5 task structure and objectives

WP5 is expected to validate and assess the correspondence of the technical KPIs and metrics and the resulting benchmarks collected and refined in WP1-WP2 and integrated in the Toolbox developed in WP3, to make sure they effectively correspond to the intentions of the original tools and needs of Industrial and research communities. The actual evaluation in WP4 will be served as a set of analytical insights into the data on different levels.

The workpackage consists of three tasks connected into a pipeline. Figures 1 and 2 depict the role of each of the following three tasks:

- **T5.1** – Systematization/ontologizing, storage, evaluation and validation of metrics based on data measurement including business data, sensor data/time series, media data, natural language data (incl. web/social media) and geospatial/spatiotemporal data
    - *Objective: To structure description of benchmarking experiments through indicators. To create a knowledge graph and an ontology of indicators [Figure 3].*
- **T5.2** – Assessment of technical usability, scalability, complexity and relevance of corresponding metrics and data, considering problems to be solved
    - *Objective: To generalize data from benchmarking experiments into a model via machine learning and other analytical techniques on the collected data from WP1 & WP2*
- **T5.3** – Assessment of the sustainability of the tool, finalisation of the methodology
    - *Objective: Integration of the tool into the DataBench Toolbox and positioning the DataBench methodology to be used in the data science industry*

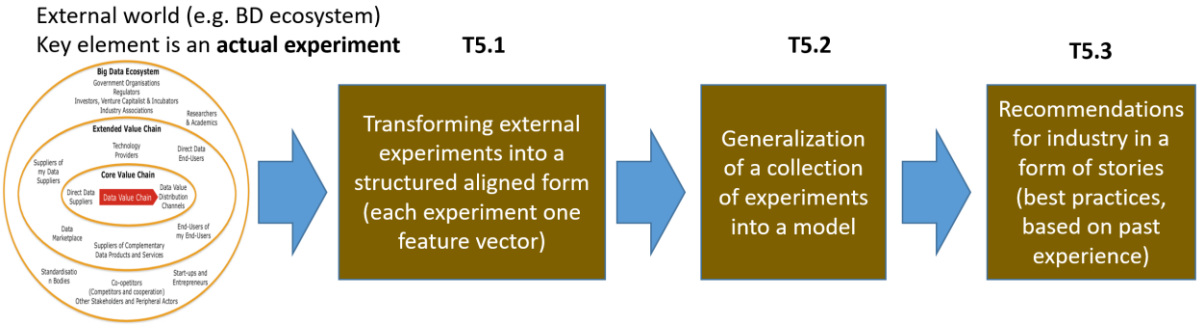

Figure 1: The pipeline of three tasks in WP5

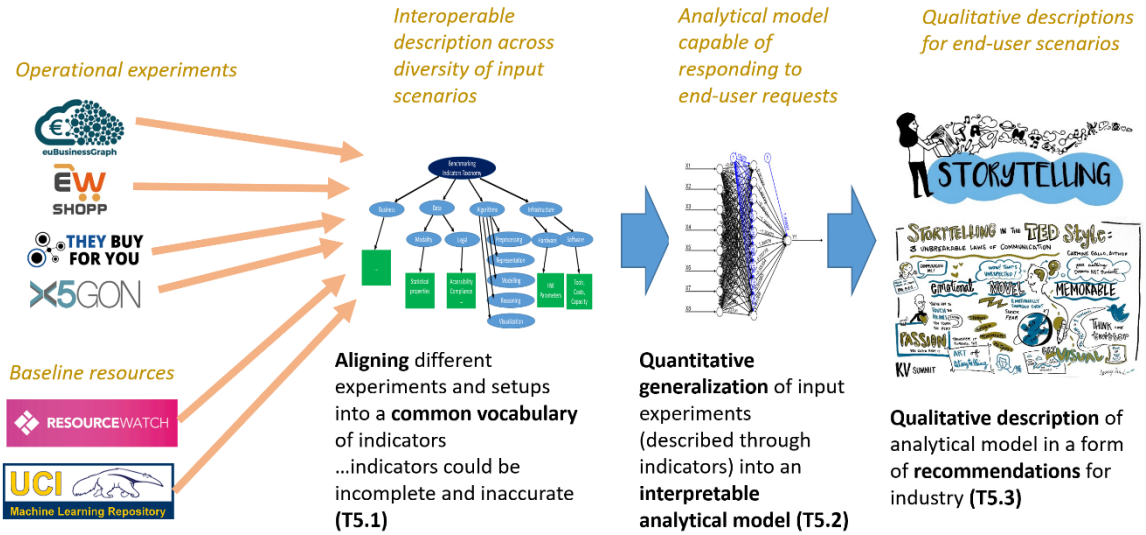

Figure 2: Detailed structure of WP5 along three tasks

In other words, WP5 will make sure to:

- collect the data and corresponding data schemas from WP1 and WP2;
- collect requirements from WP3 (DataBench Toolbox) for the solution to get integrated on the software engineering level (most probably the solution will be REST API interface);
- collect requirements from WP4 on the required analytic (quantitative and qualitative) input;
- organize the data in a knowledge graph;
- derive an appropriate ontology on the top of the knowledge graph;
- prepare analytic environment to allow extracting insights from the collected data from WP1 and WP2;
- prepare an analytical solution using collected data;
- integrate analytical solution into the DataBench toolbox (WP3);
- support evaluation process of the WP4;
- finally, the end task of WP5 is to ensure sustainable deployment of the analytical solution to allow reuse of the environment beyond the end of the project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

3. Knowledge graph representation of DataBench data (DataBenchKG)

Around 2010 Knowledge Graphs became a predominant solution to store and retrieve structured data in environments which require flexible and ever changing schema. Historically, the area of Knowledge Graphs stems from the area of Semantic Web, where EU in its FP5, FP6, FP7, and H2020 programs invested large amount of resources, and as a consequence EU has visible role in the corresponding communities and delivers some of the core resources in the area.

In particular, knowledge graphs became the main solution to ensure interoperability in mid-sized to large enterprises, where the key role is to connect legacy data resources across the enterprises' IT infrastructure. Typically, knowledge graphs provide a thin infrastructure layer on the top of exising databases, connecting diverse data schemas and enabling data retrieval for flexible application scenarios.

As one of the key global resources in the area of Knowledge Graphs is "Linked Open Data Cloud" [3], which connects 1,234 datasets and schemas (as of June 2018). The LOD Cloud is maintained at Insight Centre for Data Analytics in Ireland with many international contributions. Figure 3 depics the structure of the LOD cloud.

How the Knowledge Graphs relate to the DataBench project? The information collected at various stages of the project (in particular WP1 and WP2) will be organised in a structured form to be easily accessible, structured along appropriate schemas, and interoperable with other related external semantic knowledge resources trying to standardize the domain of data management.

For that purpose, we will refer to the specific Knowledge Graph built in the DataBench project with the working name 'DataBenchKG'. In the next paragraphs we will describe the key ingredients of DataBenchKG, the envisioned implementation and the required characteristics.

Based on the preliminary analysis, we envision the information coming from the project to include the following sources (but not limited to, in the case of necessity to expand):

- **Questionnaires** – question-answer pairs, where the question part will be textual, while the answer part will be either structured in the multiple-choice lists or in minority cases textual descriptions.
- **Interviews** – the data will include pairs of (semi)structured questions and answers as unstructured textual descriptions.
- **Data science algorithms descriptions** – algorithms will be described in a form of structured descriptions as used in data science; whenever possible, the descriptions will be aligned with an ontology of machine learning and broader data science related algorithms; as much as possible we will use the existing efforts as part of the W3C Machine Learning Schema[1].

[1] https://www.w3.org/community/ml-schema/

- **Data science tools descriptions** – the tools (typically software systems) will be described in a form of structured descriptions; since we are not aware of any ontology to describe the tools and software packages, we plan to develop within the project a 'minimal viable ontology' satisfying the project needs.
- **Dataset descriptions** – to describe data, datasets and other types of data resources, we plan to use the existing body of knowledge from the area of Meta Learning (as a subfield of Machine Learning) [4][5]. The aim is to derive structured descriptions of characteristics of datasets which are commonly used in data science, statistics and broader in the area of data analytics. The process to extract data characteristics will be automated and will address general information about the datasets (modality, size) as well as shallow statistical properties (such as statistical distributions, correlation among the variables etc). During the course of the project we plan to compile a viable approach to define a schema satisfying the needs of the project. The major objective will be to automate the process of extracting such characteristics from diverse datasets.
- **Benchmarking tools description** – as part of the project we will address several benchmarking approaches and tools, and the goal will be to describe them in a structured way to perform comparison with particular focus on the DataBench platform. In particular, the aim is to connect in the DataBench framework related initiatives, such as other H2020 projects and other experimental setups.
- **Benchmarking experiments** – each benchmarking experiment performed either within DataBench framework or outside will measure and collect diverse KPIs (including memory consumption, time complexity, various metrics to estimate quality of analytic results, business aspects) – such collected data will be stored in the DataBenchKG in a structured way.
- **Benchmarking with machine learning models and datasets** – as part of the project we will perform extensive analytical tests combining a selection of ML and BigData algorithms with a selection of publicly available reference data sets. The aim is to create a recommendation tool to suggest what kind of algorithms should be used in particular data scenarios. The purpose of this task is to derive a generalized understanding on how different machine learning and analytical algorithms and tools perform when analysing different datasets and under broad range of parametrizations. The models will be represented in a structured form where a possible common method to represent machine learning models in an interoperable way is 'Predictive Model Markup Language' / PMML[2].

The above types of information will be systemized and stored as a knowledge graph (DataBenchKG), where individual knowledge and data and fragments will be aligned with external ontologies/schemas or ontologies/schemas constructed within the project (some of the semantic schemas required to represent the data are not developed yet). As a starting point we will use the well established semantic resources

[2] https://en.wikipedia.org/wiki/Predictive_Model_Markup_Language

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

including WikiData and LinkedOpenData. For technical and business concepts, where pre-existing semantic resources exists, we will align with the corresponding semantic schemas, like W3C Machine Learning Schema and Predictive Model Markup Language/PMML.

The collected data will be stored conceptually as a Knowledge Graph, whereas for the implementation of the actual storage will plan to use one of the well established and proven scalable graph databases such as Neo4J (https://neo4j.com/), ArangoDB (https://www.arangodb.com/) or similar. The final decision, which graph database to be used for DataBenchKG, will be taken at the beginning of the implementation phase.

The aim of constructing DataBenchKG, is the aggregation and analytics capabilities on top of the collected data. Most of the data types and sources (as listed above) ,to be stored in the DataBenchKG, are of a moderate scale and consequently we don't expect major implementation issues. For these data resources the out-of-box graph database engine will support basic aggregation operations such as search and baseline statistics. As we plan, the data source with the most data input will be coming from the 'Benchmarking with machine learning models and datasets' (generated by the tools from WP5), where we expect tens of thousands (or more) experiments to be performed and stored in the graph data engine. For the purpose to be scalable and having ability to perform modelling and aggregation, we could use alternative data storage engines, like the NoSQL database MongoDB or relational database PostgreSQL.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

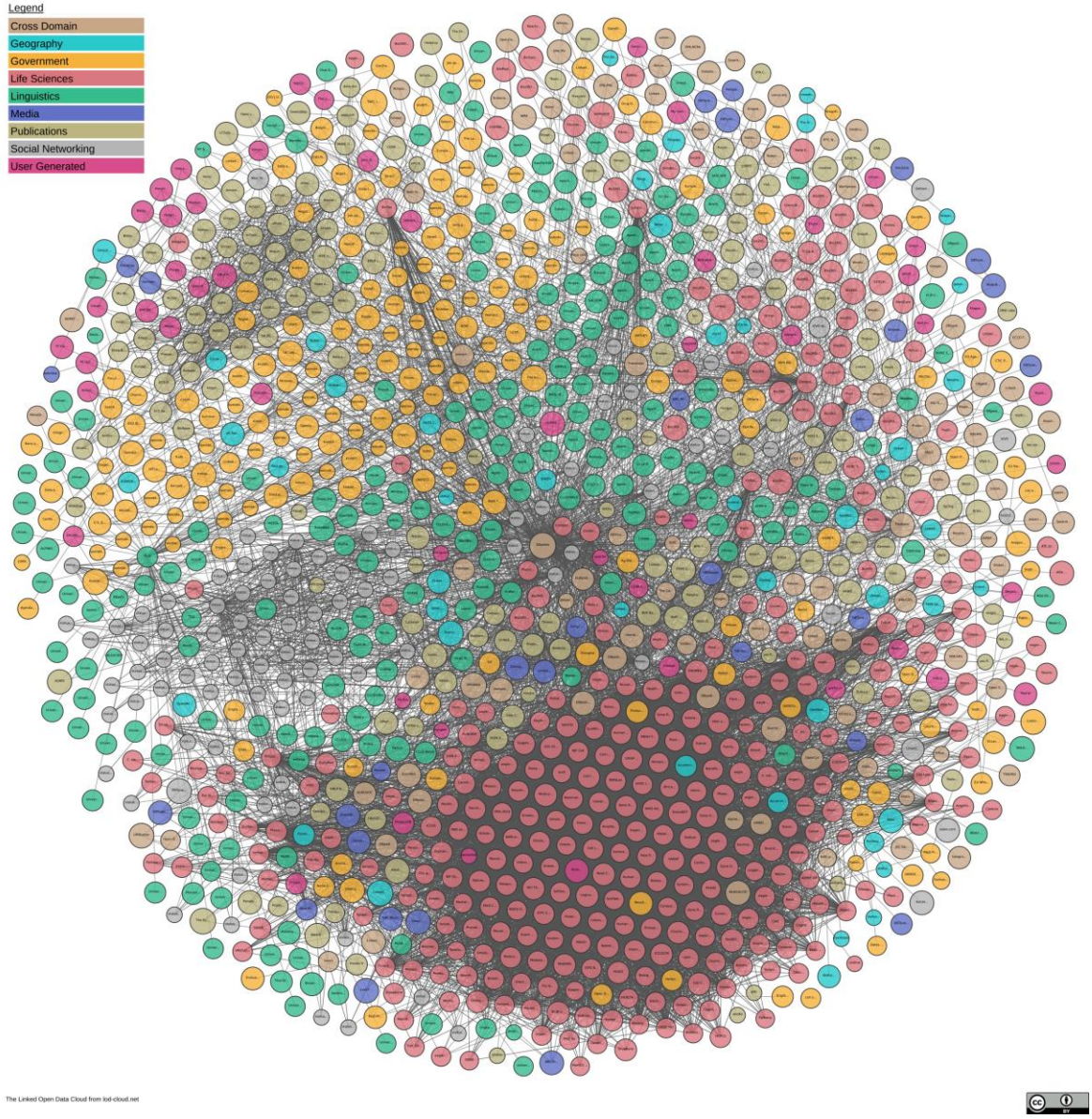

Figure 3: Depiction of Linked Open Data Cloud [3]. The structure of the graph of interconnected knowledge-graphs is color-coded addressing eight domains and cross-domain knowledge graphs.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

4. Construction of the Ontology of indicators/KPIs

The previous section described major data sources which are planned to be collected and used in the DataBench project. The data will be stored as a knowledge graph, which is conceptually a lower level data structure, operating with schemas and allowing baseline operations such as retrieval, counting and simple statistics. The upper part of the structure, systemizing the domain of data science and consequently its benchmaring will be defined as an ontology with all the relevant concepts connected with a relationship structure.

As a short introduction to ontologies, we could say they consists of four major elements:

- Concepts which describe the anchors and fundamental part of the structure – typically they are represented as nodes in a graph/network structure.
- Relations which connect concepts into a structure and establish interplay between individual meaningful building blocks. Relations could be either hierarchical (which allows multi-level aggregate representation of the world), as well as horizontal (to represent data instances of the world to be described by an ontology). Relations are typically edges in the graph/network structure.
- Attributes are an additional element which further describes either concepts or in more rare cases relationships. Attributes are typically just fields with values to further elaborate the context of given ontological element.
- Data or Knowledge Sets are just data which need to be systemized into an ontology. These are typically data instances as measured and observed from the environment. In the case of DataBench, the knowledge graph will have this role, where all the data will be stored and inserted into the ontology with the purpose to generate complex queries and aggregation.

Figure 4 illustrates the major elements of an ontology.

The most popular formalism, standardized at W3C, is Web Ontology Language (OWL) [7]. It allows description of an ontology with all the above mentioned elements. Several tools allow input and manipulation with OWL ontologies – the most popular is Protégé [8] which will be used also in the DataBench project to define and manipulate the ontological structure.

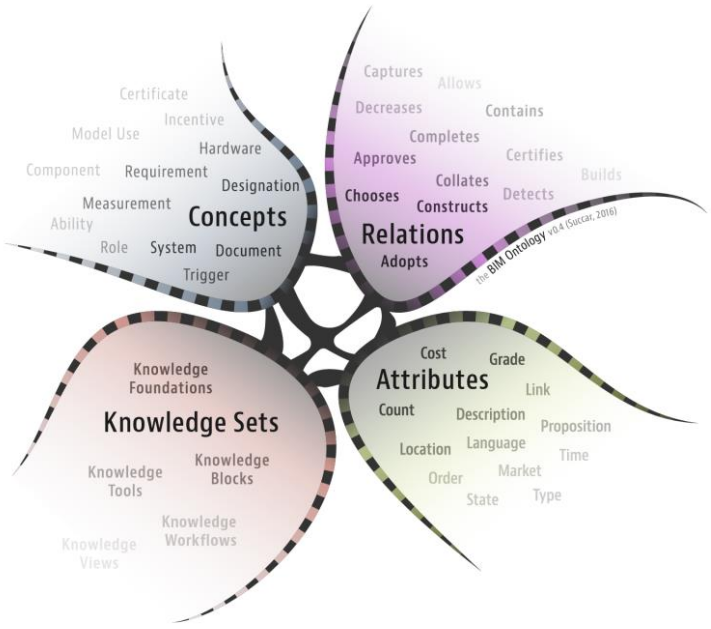

Figure 4: Depiction of major elements of an ontology [6].

The DataBench ontology will use the standard approach of modelling concepts and relationships. The input will come from the data sources, as described in section 3. The major building blocks of the ontology are depicted on Figure 5 visualizing all the major data inputs within the project.

The key data types of the DataBench ontology will include raw data and collected (either calculated or estimated) KPIs. The major segments of the ontology will include:

- Business data and KPIs
- Datasets including modality, characteristics and legal framework of the data accessibility.
- Algorithms including all the major subgroups like pre-processing, model representation, modelling/analytics process with associated parameters, model interpretation/reasoning and visualization.
- Infrastructure describing benchmarking setups like hardware and software with associated tools using the infrastructure.
- Information on the process pipeline like data acquisition (including Edge, IoT…), storage, pre-processing, curation, visualization, usage.
- Architectural approaches where benchmarks are executed.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

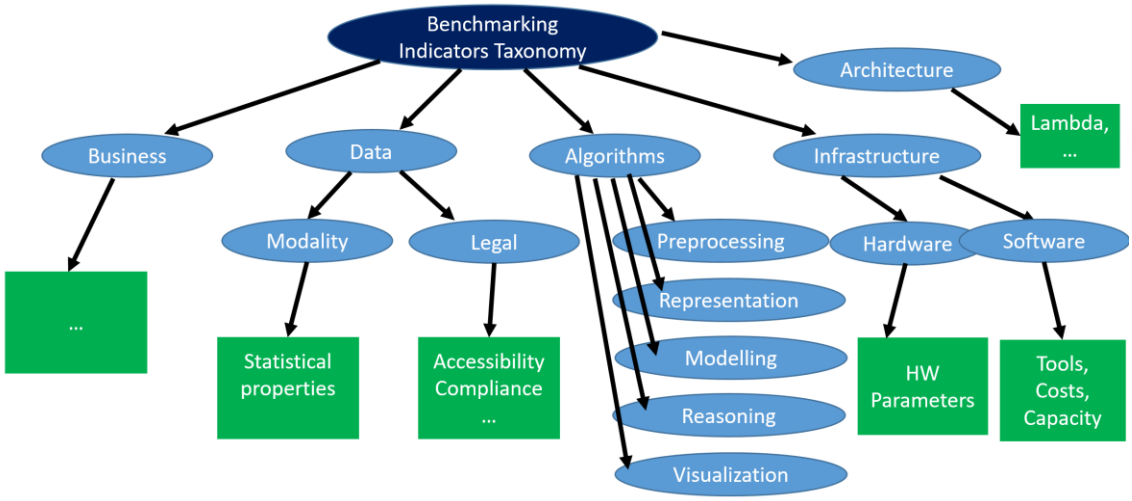

Figure 5: Depiction of an ontology of indicators.

The ontology structure (concepts relations, attributes) will be specified in collaboration with WP1, WP2 and WP3 which will contribute the data input to be organized. The operators to insert, aggregate and retrieve information from the ontology (either through lookup or some form of reasoning) will be defined in collaboration with WP3 (Toolbox) and WP4 (evaluation interpretation).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

5. Automatic extraction of indicators from data

The central object of machine learning, Big Data and more broadly Data Science is a data set. Typical approach, when analyzing data, is to use the experience and 'feeling' of the data scientist – some types of data modalities have usual characteristics (such as text and images) while others have unknown distributions and require some prior investigation, which usually doesn't happen. A typical data scientist uses his/hers favorite data science tools and through an empirical test and error paradigm converges to better or worse result.

In WP5 we want to overcome such a bad practice and provide a tool which will take a target data as input and create a meta description of such dataset for the purpose of deeper understanding and improved decision in the follow-up steps. Such meta descriptions would include not just generic indicators, but extract shallow (i.e. not as deep as a machine learning algorithm would reach) characteristics of the data. This includes describing probability distributions of the corresponding variables and insightful relations (through various statistical relational measures like correlation and beyond) between the variables. Those are typically never observed ahead of time by a data scientist and are only spotted (and not reported) by the analytic algorithm.

The major goal is not just to report characteristics of a data set, but rather link the extracted information with the properties of individual relevant analytic algorithms. As a result, the outcome would be a recommendation service where, given a data set, the system would propose which analytic algorithms would perform best and under what parametrization. As a general goal, we would like to create a 'landscape' of what kind of data performs well under what conditions and for what kind of tasks.

The area of data and algorithm characterization used to be known in Machine Learning as 'Meta Learning' which got lots of attention in the early years of Machine Learning and is not fully re-developed in the recent years in the time of BigData and Deep Learning.

The actual work in WP5 will include approaches which were used in the past and propose new approaches relevant for the todays' setups and techniques. In particular, we expect contributions on the side of one pass algorithms of large datasets to extract useful characteristics from diverse datasets.

The following is a structured description of the data set characterization approach. Note that a similar procedures will be designed for other segments related to data processing (including data storage, acquisition, pre-processing, interpretation visualization, architecture):

- **Input**: a *data-set* to be analysed
    - *Example data-sets*: sensor data (streaming, partial, noisy), text (traditional documents, Web, social media, news, multilingual), multi-media (audio/images/video), geospatial/spatiotemporal, and traditional structured data from relational databases
- **Processing**:

    - *An algorithm* having either one pass through the data or sampling the data source (in the case of large size or a stream of data)
    - The algorithm collects properties of the data-set including statistical properties of the variables and shallow relations between variables
- **Output**: *actionable characterization* in a form of a vector of meta indicators and corresponding tuneable similarity metrics:
    - The **indicators will structure the metrics characteristics along several key dimensions** to make
        - (a) different datasets/experiments comparable, and
        - (b) to provide an interpretable metrics landscape based on how data is being used.
    - Some **key dimensions** include: the aspects of storage, access, streaming, data-modality, integration, semantics, pre/post-processing, modelling, reasoning and visualization, anonymization/privacy, and legal/copyright.
    - We expect to be able to measure the characteristics of **few hundreds** (up-to thousand) different datasets.
    - Among others, the metrics should allow visualization and exploration.

The broader context, where the above data characterization will be used is to perform modelling to establish relationship between data characteristics and algorithmic performance and business indicators. The result will be an interpretable model allowing to interpret how data, algorithms/tools and business KPIs are connected and where are the better or worse scenarios to be used in practice. The approach to link different classes of KPIs will be based on collecting sufficient data for each of the classes and to create statistical soft mappings among them. There is a risk in the cases where not enough data will be collected which we plan to mitigate with human interventions in a form of background knowledge (such as interpretable rules). The overall aim is to gain an interpretable correspondence between KPIs collected in diverse situations.

The sketch of the procedure is the following:

- **Input**: Dataset annotated with technical and business indicators.
    - Datasets, characterized and landscaped in the Task 5.1.
    - Dimensions impacting business decisions (Task 4.2) such as (a) scalability, (b) analytic task complexity, (c) technical usability, and (d) relevance.
- **Output**: Model/analytic mapping from dataset characteristics and methodology being used in the observed projects.
    - Datasets will be associated with tasks and possibly systems used.
    - The created model will statistically estimate mapping from a dataset and its associated tasks to challenges of how the data is being used in terms of selection of project infrastructure, methods, tools, user interfacing and legal challenges.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

6. Conclusions

WP5 has the role in the project to use the data technology for the purpose to evaluate and validate BigData benchmarking scenarios. The present document described three scenarios how WP5 will approach data benchmarking validation:

- Construction of the Knowledge Graph (DataBenchKG) to store all the data collected by the project into flexible schema graph database.
- Construction of the DataBench ontology to structure and systemize all the terms related to the BigData benchmarking and to allow a level of reasoning over the collected data.
- Using analytics techniques to characterize data resources and relate them to algorithmic, tools and business KPIs for the purpose of recommending what should be used in particular Big Data scenario.

Bibliography

[1] Chen, Yanpei, “We Don’t Know Enough to Make a Big Data Benchmark Suite“ An Academia-Industry View, Unpublished paper presented at the Workshop on Big Data Benchmarking. May 2012, San Jose, CA. https://amplab.cs.berkeley.edu/publication/we-dont-know-enough-to-make-a-big-data-benchmark-suite-an-academia-industry-view/

[2] Dan McCreary, “2018: The Year of Enterprise Knowledge Graphs” https://medium.com/@dmccreary/2018-the-year-of-enterprise-knowledge-graphs-66e868762b49 , Jan 2018

[3] Linked Open Data Cloud, https://lod-cloud.net/

[4] Sharan Vaswani, Meta Learning, http://www.cs.ubc.ca/labs/beta/Courses/CPSC532H-13/Slides/content-session-4-slides.pdf

[5] Vilalta, Ricardo, and Youssef Drissi. "A perspective view and survey of meta-learning." Artificial Intelligence Review 18.2 (2002): 77-95.

[6] Favio Vázquez, “Ontology and Data Science” https://towardsdatascience.com/ontology-and-data-science-45e916288cc5

[7] OWL – Web Ontology Language - https://www.w3.org/OWL/

[8] Protégé “A free, open-source ontology editor and framework for building intelligent systems” – https://protege.stanford.edu/

[9] Christiane Lemke, Marcin Budka, and Bogdan Gabrys: “Metalearning: a survey of trends and technologies”, Artificial Intelligence Review 2015; 44(1): 117–130.