Grant Agreement Number: 825225

Safe-DEED

www.safe-deed.eu

Report on Requirements and Design

Deliverable number	*D4.1*
Dissemination level	*PU = Public*
Delivery data	*due 31.05.2019*
Status	*Final*
Authors	*Marius Paraschiv, Ludovico Borrato, Yonas Kassa, Mihnea Tufiş*

Horizon 2020
European Union Funding
for Research & Innovation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825225.

Abstract

Safe-DEED is a project that takes a highly interdisciplinary approach, with professionals from software development, cryptography, data science, business innovation and the legal domain working together to enhance economic growth and improve trust in the modern data-driven economy.

It has been found that, even among large companies, many do not have a data valuation procedure in place. One component of Safe-DEED is concerned with Data Valuation. The purpose of this component is to provide value estimates for datasets, both by taking advantage of qualitative information supplied by the owner, and by exploiting the structure of the data itself.

In order to ensure the safety and anonymity of the datasets, a significant effort is also focused on scalable cryptographic protocols. The overall goal of Safe-DEED is to create value for companies and clients alike.

This document provides an initial assessment of the requirements and design of the Data Valuation Component (or DVC), taking into account data ingestion, analysis, scoring, valuation and suggestions for a presentation layer. An approach to context free valuation is also provided.

Table of Contents

List of Figures

Abbreviations

DVC	Data Valuation Component
ADAS	Automatic Data Analysis and Scoring
S2VM	Score-to-Value Mapping
WP	Work Package
CPL	Communication and Presentation Layer
QDSC	Qualitative Data Scoring Component

1 Executive Summary

This document represents Deliverable 4.1, a first requirements analysis of the Data Valuation Component (DVC) and its proposed implementation.

The purpose of the DVC is to provide users with the ability to estimate the knowledge and monetary value of a certain corpus of structured data based on a sample of the data and contextual information obtained via a questionnaire. A typical use case would have a customer provide access to a data source and answer a series of questions regarding both the data use and its characteristics. The series of answers would be provided in questionnaire-style, and they will establish the required *use context*. Due to an understanding of the situation in most companies, the person (or persons) answering the relevant questionnaire may not have the required aptitudes or necessary knowledge to provide a complete use context, hence a sub-component of the DVC attempts to automate the analysis based on the provided sample, and further increase the amount of information the platform requires in order to valuate the dataset. Furthermore, the valuation questionnaires are split into three categories to better suit the availability of individuals with relevant skills within various companies.

- The first part can be answered by business-focused employees,
- The second part will be answered by business-intelligence workers and,
- Finally, the technical questionnaire should be addressed to data scientists and data engineers.

When relevant scores are obtained from the questionnaires and the Automatic Data Analysis and Scoring (ADAS) sub-component, the score can also be mapped to an actual monetary value. Due to the context specific nature of data, economic scoring of data will be subject to availability of existing data-to-economic valuations in the given context. Since this part of economically valuating a given data corpus is not a mature discipline [7, 8] and it requires further research on possible data-to-economic mappings, novel algorithms and economic models, it has not yet been treated at the current stage of the project. However, since we have a dedicated sub task (task 4.3.2) addressing a context-aware economic model, economic scoring will be covered there in detail. A high level description of the pursued approach is given, and this relies on considering datasets from different domains and estimating a mapping function between their questionnaire score, automatic data quality score, and real market value. This task is performed by the context based economic model sub-component, and it is one of the main research topics of this work package (WP4).

As a final result is obtained, the information needs to be presented to the user in a clear and detailed manner. We will present a number of aspects that should be considered, when designing the Communication and Presentation Layer (CPL).

Finally, the situation when contextual information is not available may arise, and a method for context free valuation is proposed. This approach requires a database of sample - price pairs, and, as such, is not a viable option at the beginning of the platform's initial implementation.

2 Introduction

Reliably estimating the market value of a dataset is one of the most important challenges of our modern information economy. As part of Safe-DEED, WP4 is tasked with establishing a framework for valuation that can adapt to changes and to the requirements of a wide domain of data and market segments.

The valuation process is performed in two stages. Initially, the potential dataset is scored by a series of questionnaires that also help describe the nature of data; this helps the data ingestion layer to properly load the data and helps feed it to the automatic data analysis and scoring module (ADAS). The questionnaires gather information from the Data Owner related to data properties (e.g., type, features, usefulness for various algorithmic prediction methods), as well as use-cases and relevant costs for processing, transforming, collecting and storing the data, among others. The qualitative score enables the first stage of the DVC to assign a data value to a newly-received data corps. Once the value of data is calculated using the ADAS and the qualitative scores, the next step will be decided based on the availability of the use context. If a context is provided and the requirements of the context based economic model are satisfied, a price tag is assigned to the dataset. If the requirements of the context based economic model are not satisfied or context based data valuation is not required, the data will receive a context free value ranking, based on the overall information provided. Finally, the CPL (Communication and Presentation Layer) will present the assigned price / rank, as well as other details related to the data, various recommendations and possible applications, in a user friendly manner.

3 Data Valuation Component – Requirements

This task involves interviews with stakeholders, in order to collect input on the type of potential data customers, the economic contexts in which these customers would like to perform data valuation, the type of questions that they would ask, and the type of output that they would expect from the Big Data Valuation component. The collected input will be converted into specific functional requirements, while also taking into account technological requirements (scalability, fault tolerance, quality of experience, security, privacy, etc.). Finally, the task involves a high-level architectural design of the component, including the final choice of a technological stack.

3.1 Interviewing the stakeholders

The following results are based on an initial survey[1] designed to help us elicit a first set of requirements for the design of the Data Valuation Component.

The survey was sent to professionals from organizations whose business involves the consumption and/or production of data. The majority of the companies are based in Catalonia, but most of them have a global reach; some of these organizations are from the rest of Europe. A total of 19 participants – representing professionals using data from different perspectives (decision makers, managers, research, development) – responded to our questions. In the coming

[1] The survey is available at https://forms.gle/AX8BYun59zxuY9V19

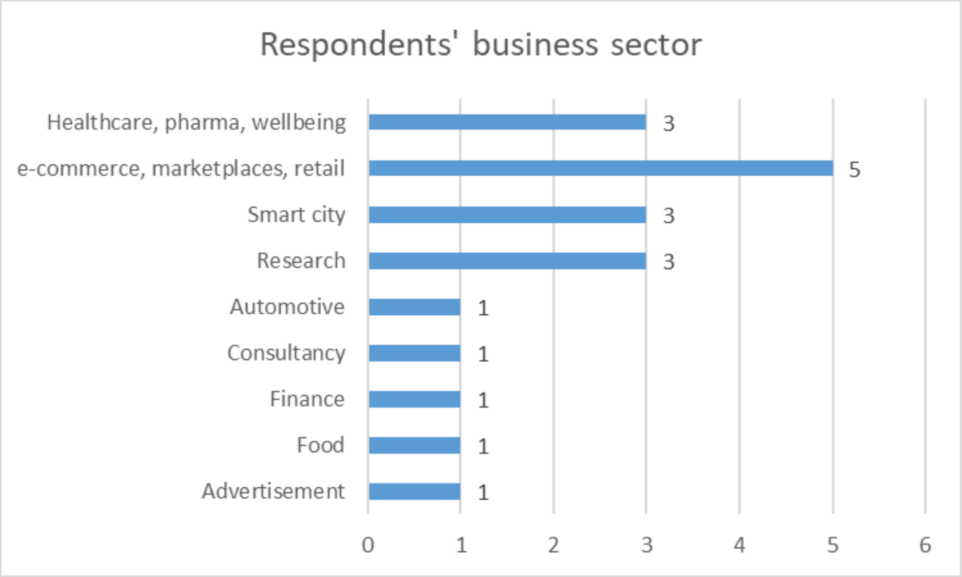

Figure 1: Business sector of respondent organizations

months, we plan to launch a large scale survey to get more diverse and representative insights from stakeholders in the data market; we plan to update the architecture subsequently.

Overview of the results

Next, we discuss the results of the survey and the insights that we have gained in terms of what the data needs at different levels in an organization are, what kind of datasets are being transacted, how organizations currently evaluate datasets and what would they expect from a Data Valuation Component.

Figure 1 and Figure 2 below illustrate the business sector and the size of the companies and organizations taking part in the survey.

Similarly, Figure 3 illustrates the roles each of the respondents take in their organizations. This allows us to get a better understanding regarding data needs, data interactions and data exploitation at different levels in an organization.

Two different trends are shaping out in terms of transacting with data sets. The majority of the respondents (16) reported that they bought data from external sources; however only a minority (4) declared that they sold data that their business generated. Thus, there seems to be a clear need for companies to acquire data, doubled by a reluctance to share or sell whatever data they are generating.

Buying data

When acquiring data, respondents tend to look for the following types of data:

1. traffic, transportation, geo-location;
2. weather, satellite;
3. retail, customer-related (demographics, behaviour);

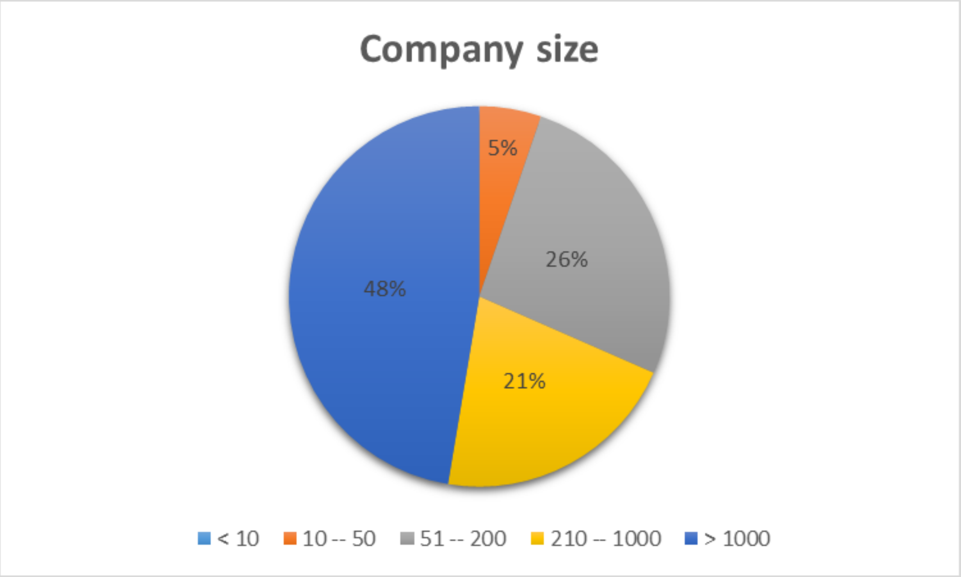

Figure 2: Business sector of respondent organizations

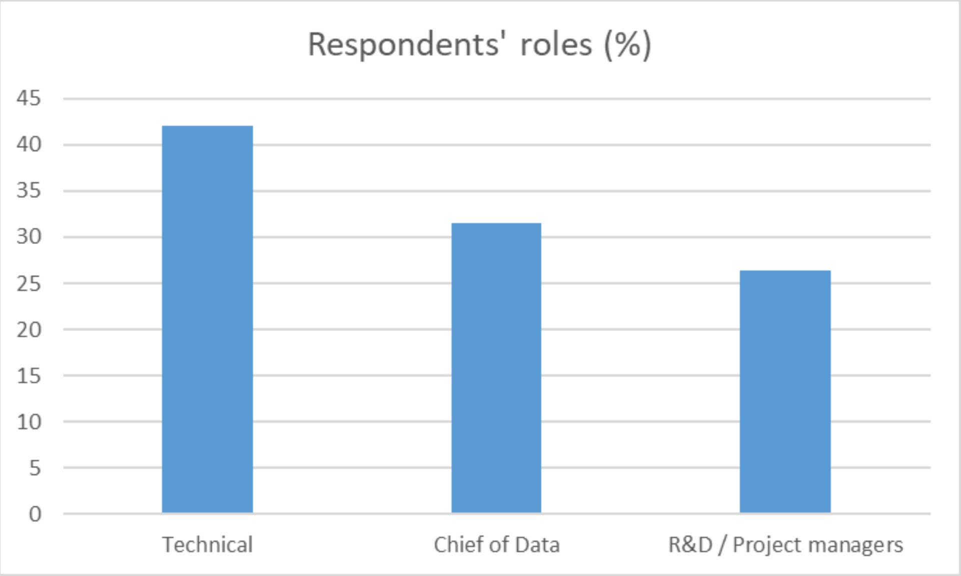

Figure 3: The roles of respondents within their organizations

4. social media;

5. company data, financial data.

Looking at the companies that don't acquire data, we see that they are in general medium or large companies (over 1000 employees), from the retail or health-care sectors. A reason for this could be that they are self-reliant on generating all the data that they need to achieve their business goals.

Next, we wanted to understand how organizations are evaluating the importance of several features of the data they acquire, in the context of their projects. The survey proposed a 5-levels scale (not important - 0, slightly important - 1, important - 2, somewhat important - 3, very important - 4):

- **Accuracy** (how close to reality is the data). Based on our results, this feature is the unanimously appreciated by data practitioners; all of our respondents rated it as being at least "important" and 14 of them (74%) appreciated it as being of utmost importance. Its average score was of 3.68;

- **Timeliness** (how up to date is the data). The minimum evaluation that this feature received by all but one of the respondents was "important" (i.e., in almost all of the cases it received a score of 2 or more). Its average score was of 3.12;

- **Completeness** (proportion of missing values). This feature had the exact distribution of responses as timeliness. A majority of respondents (15 - 79%) considered this feature to be at least "somewhat important", with the average importance of this feature being at 3.12;

- **Accessibility** (how easy it is to find the data). In some sense, this feature refers to the availability of a given dataset. Almost all respondents ranked it as being at least "important", but opinions were more uniformly split between the higher degrees of importance, with an average rating of 2.84;

- **Size**. This feature seems to have been the most divisive, with answers being very uniformly split between the different degrees of importance, resulting in an average score of 2.21;

- **Uniqueness** (proportion of unique records). The assessed importance seemed to be more uniform for this feature, with a majority of respondents (11) judging it "slightly important" or "important" at best. The average importance of this feature is 2.00.

This analysis allowed us to understand which features are of clear importance to all data practitioners and which of them are more divisive. Once the large scale survey will be conducted, we hope to better understand the expectations that different stakeholders have from the data they wish to acquire.

Finally, we compiled a list of other important attributes, as suggested to us by our respondents:

- Structure - whether the data is structured or not;

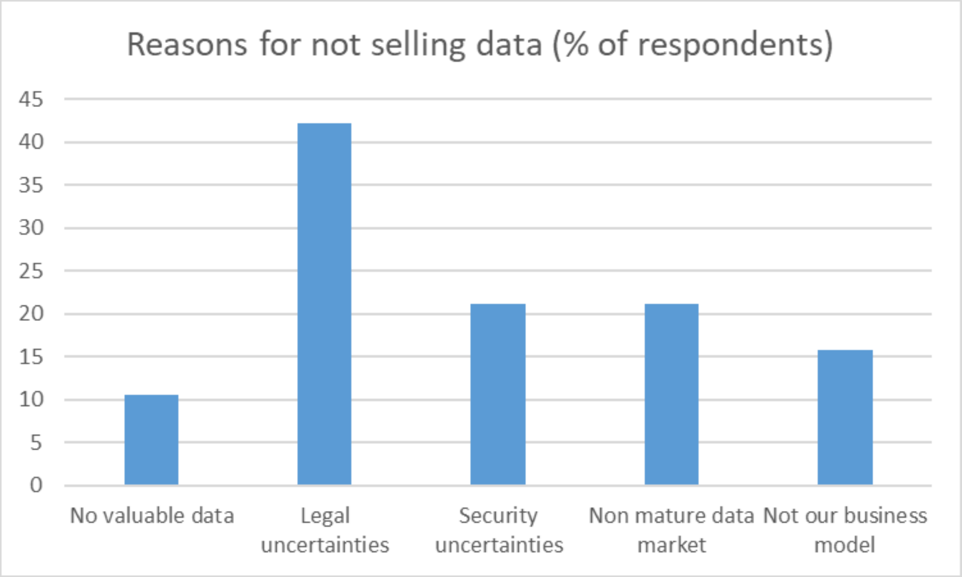

Figure 4: Reasons for not selling data

- Provenance - the origin of the data;
- Ethics - whether the data was obtained by respecting regulations for data collection, use and exchange;
- Standardization - whether the data respects standard domain formats;
- Metadata.

Selling data

As previously mentioned, only a very small number of companies (4 - 21%) actually sold data that they generated. Yet again, these are medium or large companies (e-commerce, consultancy), which are able to generate volumes of data useful to other businesses. An interesting segment is represented by organizations that generate data of public interest and are publishing them in an open and free manner. This is the case of research centers, which produce data as part of public funded projects and are required to publicly share this data in return. While this doesn't necessarily count as selling data for a profit, these datasets are available for use and should be valuated nevertheless. As for the types of data usually sold, they tend to correspond to those already reported when acquiring data (e.g., demographic, customer, retail, transportation, geo-location, weather, social media, financial).

Looking back at the majority of the participants in the survey that reported that their organizations **never sold** data, Figure 4 resumes the main reasons behind this.

The main reasons for which companies hesitate to sell their data relate to uncertainties that organizations seem to have about how they can further use the data, may they be legal (perhaps a poor understanding of the General Data Protection Regulation, GDPR), security (probably partially related to the legal aspects, as well) or commercial. Some respondents also reported that they do not own valuable data, an evaluation which will be interesting to follow in our large scale survey.

3.1.1 Platform output

Finally, in order to understand what different users are expecting from a Data Valuation Platform, we proposed several possible types of outputs and asked the participants to rate the importance of each of them:

- Short summary of the dataset. All respondents deemed this output to be at least "important" and 53% of them considered it "highly important";
- Quality assessment of the dataset. All respondents deemed this output to be at least "important" and 68% of them considered it "highly important";
- Economical value of the dataset. There was less agreement on the importance of this type of output; however, 79% of respondents did rate it as being at least "important" (26% "highly important");
- Exploitability. The Platform might suggest to the user a set of applications relevant to the given dataset in different domains. Most users (90%) also appreciate this information as being at least ”important", but there seems to be less consensus regarding its degree of importance.

As before, we allowed the users to suggest other types of outputs that they would expect from a Data Valuation Platform:

- Technical and business metadata: such data could be included in the summary;
- Data samples;
- Exclusivity - how many times that data has been sold or used, based on the premise that "rare" data is more valuable.

3.2 Functional Requirements

Based on the answer to the survey, the following functional requirements have been derived.

1. **The User must securely connect to the Data Valuation Component**.
2. **The User must fill a form to provide to the Component a use-context for their data**. The form will be divided into 3 sections and gather information about: data impact on the business, data overview, technical applications using the data.
   (a) **The business impact section**. The User gives information about: how the data was previously used by the company, the added value of the dataset for the business objectives, costs of collecting, storing and processing the data.

   (b) **The data-overview section**. The User gives information about: how the data was generated, brief description of the data (domain), description of the features, type of data (original, transformed), ethical issues about the data.

   (c) **The technical section**. The User gives information about: the types of analyses, statistical tasks and algorithms (including Machine Learning models) that were developed with the given data.

3. **The System must compute a Data Score**. Based on the information collected through the form, the System establishes an initial Data Score. This score, together with the format of the data are used as input parameters when loading the data.

4. **The User must securely upload a sample of the data set**. This is handled taking into account the format in which the data set is stored. The size of the sample depends on the User, who may choose to upload the entire dataset or a sample that they choose as being representative.

5. **The System infers the structure of the data**. The System will count the number of features, records, analyze indices and infer the most suitable data type for each feature.

6. **The System performs automatic data analysis of the sample dataset**. The System will attempt to establish the following about the dataset: provenance, timeliness, duplicate entries, features distribution, outliers, legal compliance. If the system is able to extract provenance and timeliness, these should be double checked against the answers provided by the User according to *Requirement 2*. Privacy compliance can be checked using techniques described in *WP5*.

7. **The System must compute a Quality and Exploitability Score**. This score will be derived from the results of the automatic data analysis of the sample and from an evaluation of the applicability of different algorithms on the dataset.

8. **The System assigns a monetary value for the dataset**. This value will be based on the two different scores computed according to *Requirement 3* and *Requirement 7*, as well as the use-context provided in *Requirement 2*.

9. **The User must explore the results, displayed as a set of reports**.

   (a) **Summary about the data**. The User must see the following information: technical and business metadata, size, structure, description, cost of obtaining, legal compliance.

   (b) **Price assigned to the data set**. The User must see the final price computed according to *Requirement 8*. The User must also see how frequently was this dataset used by other Users.

   (c) **Quality assessment**. The User must explore the insights of the advanced analysis, produced according to *Requirement 7*.

   (d) **Exploitability report**. The User must see an exploitability report for the dataset: applicable analyses, applicable machine learning algorithms, most relevant application domains etc.

4 Data Valuation Component – Platform Architecture

The architecture of the DVC, represented schematically in Figure 5, is divided into a series of layers. Next, we provide an overview of the layers, and we'll describe them in greater detail in the following paragraphs:

- **QDSC (Qualitative information extraction and Data Scoring Sub-Component)**: the first stage of the data valuation process involves extracting information from the data owners themselves. This information extraction process will be developed to make it as less technical as possible, to ensure that it is accessible to a wider audience with different backgrounds from different domains. For this reason, it can be provided as a series of answers to a questionnaire with simplified forms. This helps place the valuation process into a market context, and this stage is also crucial in initializing the data ingestion layer with appropriate configurations.

- **Data Ingestion Layer**: this represents the entry-point of the data into the platform. The layer is tasked with detecting the nature of data and selecting the operations suitable for it. This layer may also provide the sampling and meta-data extraction, in line with the Safe-DEED safety and privacy requirements. This layer will also integrate with existing marketplaces, such as Data Market Austria.

- **ADAS (Automatic Data Analysis and Scoring)**: after the Data Ingestion Layer loads the data and prepares it for analytic operations, an initial series of tests is performed on the data sample. This serves to extract information about instance types, missing values, possible inaccuracies, and how good the data is for various types of machine learning algorithms (hence for various application tasks). Based on these results, an automatic data scoring will be generated, considering data quality assessment criteria such as data completeness, uniqueness, timeliness, validity, accuracy, consistency, and data relevance [9]. This assessment represents an automatic scoring of the sample, which is explicitly based on its structure and content.

- **S2VM (Score-to-Value Mapping)**: once the complete score is available, the data is considered to be ready for valuation. If the market segment to be addressed is provided and requirements for economic modeling are satisfied, the S2VM will attempt to map the score value to a relevant price value. This layer takes into account market dynamics and a history of previous, similar datasets, among other external factors.

- **CPL (Communication and Presentation Layer)**: finally, once the analysis results are available, they are presented to the end-user in a clear manner, together with some of the results from the ADAS layer, such that, besides the value estimate, the end-user would also extract data analytic results and possible application domains relevant for future uses of the dataset.

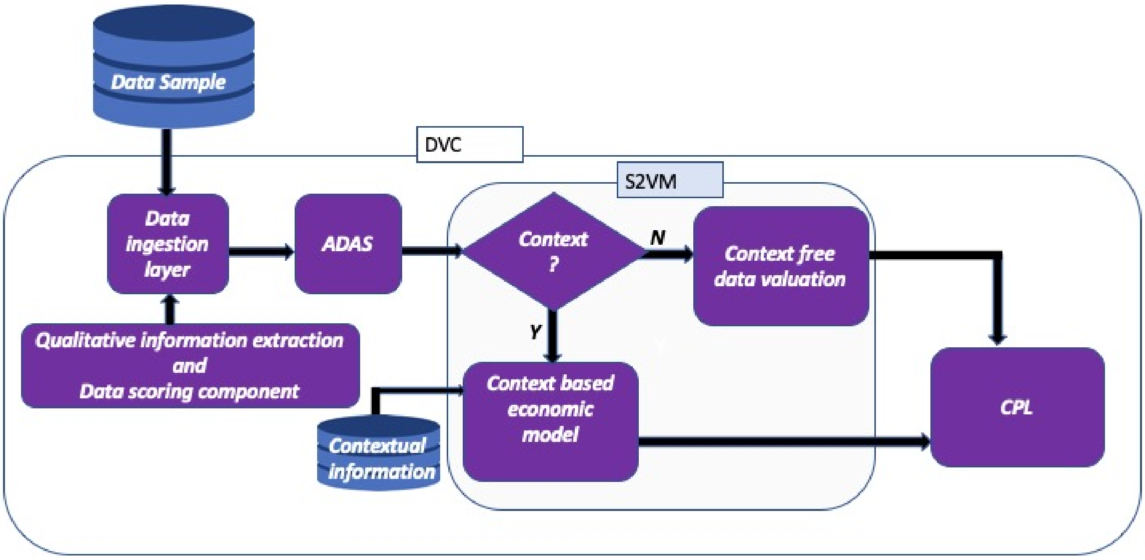
Data Sample
DVC
S2VM
Data ingestion layer
ADAS
Context ?
N
Context free data valuation
Y
Qualitative information extraction and Data scoring component
Context based economic model
Contextual information
CPL

Figure 5: DVC block structure.

4.1 Qualitative Information Extraction and Data Scoring Sub-Component (QDSC)

The first stage in the data valuation process consists in an initial interaction with the users of the Data Valuation Platform. This will be realized by the QDSC sub-module, an interface through which the user will provide information about the data, its usage, generation, and potential impact in specific domains. To acquire this information the user will answer a set of concise questions, as suggested in a recent paper by Gebru et al. [10]. Such information is also useful to place the data in a market context. These responses are also crucial in determining configuration settings for the data ingestion module (e.g., database settings, potential file settings, portions of the data to be analyzed, etc.). The interface of QDSC will have a user-friendly design in the form of a three parts questionnaire, depending on the qualification of the respondent:

- **Business-Intelligence** oriented questionnaire: containing questions regarding the behavior of data as an enabler in future projects and its potential to improve company workflow, products or services, questions describing the potential increase in value when merging one dataset with another, the cost-effectiveness of the data collection process or the costs of storing and processing the data;

- **Domain specific** questionnaire: containing questions regarding the generation of data, brief description of the data as a whole and the structure of the dataset. This questionnaire is useful to answer questions such as how and why the data was created, if it is original (raw) data or aggregate data, what information its fields contain, and whether it might raise any ethical or legal concerns;

- **Data Science** oriented questionnaire: containing all technical questions related to the data itself and its past usage and potential applications, this questionnaire will help us include qualitative information regarding the machine learning and statistical tasks it could be used for.

The information collected at this stage, together with the location of the data will be fed into the Data Ingestion Layer.

4.2 Data Ingestion Layer

The actual entry-point of the data into the platform is the Data Ingestion Layer. A series of problems need to be addressed before any type of analysis is performed on the sample. First, the data format needs to be established. This is because the analyses that will be performed at later stages are dependant on both the information gathered in the first stage (see Section 4.1) and the format of the data. In the initial implementation, due to its abundance in industry, the focus will be on tabular data (`CSV`, `TSV`, `XLS`, and relational databases). Next, the Data Ingestion Layer will load the data taking into account the constraints stated in the QDSC. Finally, the Data Ingestion Layer is tasked with identifying the necessary operations to load the data, selecting specified portions of the data, identifying the number of columns (features), the number of records (rows), and inferring the most suitable data type for each feature. Once these operations are completed the loaded data will be fed into ADAS. This layer will also integrate with existing marketplaces, such as Data Market Austria.

4.3 Automatic Data Analysis and Scoring Sub-Component

After the representative sample is loaded by the Data Ingestion Layer, the ADAS sub-component provides with automatic analysis and scoring. This score will be combined with the score resulting from the QDSC and will both be fed into the S2VM sub-module. Below, we provide a high level description of a series of tests and operations that can be performed in order to characterize the dataset and indicate its potential application for a series of machine learning algorithms:

- Checking for anomalies and possible errors;
- Computing the correlation between independent features;
- Some algorithms, such as regression, time-series, and classification algorithms have difficulties in dealing with sparse data;
- Neural networks have problems with small datasets and sometimes with generalizing outside their training domain;
- Projections such as PCA, ICA, t-SNE, would better describe the separation boundary between classes. If this is nonlinear, linear models might need to be excluded;
- Checking for class imbalance in classification contexts.

4.3.1 Score-to-Value Mapping Sub-Component

Once the scores from the two previous layers are obtained, the aggregate score could be mapped to a relevant monetary value. The relationship between score and price is complex, and a topic of ongoing research within WP4. After establishing a hierarchy of datasets for a narrow market segment, and a quantitative distance-measure (the score) to differentiate them, one approach for establishing the price for a given score would be to consider the highest and lowest prices in the hierarchy as a reference point.

4.3.2 Communication and Presentation Layer

After the analysis is done and the valuation process is complete, the results are presented to the user in a clear and concise manner. The CPL will show the user a set of results including the value of the data, a summary of the data, possible machine learning models that can be leveraged based on the data, and possible application domains for the data.

5 Proposed Technology Stack

As discussed in the above sections, the data visualization platform involves four phases: i) data and questionnaire loading, ii) automatic data analysis and scoring, iii) score to value mapping, and iv) report generation. Since the data valuation platform will involve both statistical analysis and user interface modules, and in line with Safe-DEED's commitment to reduce dependence on proprietary software, we have initially selected the following open-source technology stack developed around the Python programming language and packages that can be integrated with it. More specifically:

- Python has built-in support for scientific computing and is easily extensible to perform statistical and machine learning tasks;
- Industry standard free and open-source packages including the SciPy ecosystem (open source), NumPy, and the Scikit-learn machine learning toolkit are readily available;
- The whole pipeline, including the web service and APIs, could be homogeneously developed based on the Python stack. As a result, the developed technology can easily be transferred and adopted.

The proposed technology stack includes, but is not limited to the following:

- Openstack [2];
- Python Flask API [5];
- Popular Python machine learning and data analysis packages, such as SciPy, Scikit-learn and Pandas [4, 3];
- UI libraries such as React and Vue [1];
- Data visualization tools, such as D3.js [6].

6 Conclusion

This deliverable described the initial requirements specification and proposed architecture of the data valuation platform. The proposed architecture splits the platform into a set of functional modules that are responsible for interacting with the user (e.g., the data owner), loading the data, analyzing and scoring, based on both the interaction with the user as well as the automatic analysis of the data. It assigns a value and a price tag for the analyzed data. To realize the platform, a set of open source tools are proposed. As the project evolves, these tools and frameworks may be updated.

7 References

[1] React A JavaScript library for building user interfaces

[2] Sefraoui O, Aissaoui M, Eleuldj M. OpenStack: toward an open-source solution for cloud computing. International Journal of Computer Applications. 2012 Oct;55(3):38-42.

[3] McKinney W. pandas: a foundational Python library for data analysis and statistics. Python for High Performance and Scientific Computing. 2011 Nov 18;14.

[4] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825-30.

[5] Flask (A Python Microframework), flask.pocoo.org/.

[6] D3 a JavaScript library for visualizing data with HTML, SVG, and CSS.

[7] Agarwal A, Dahleh M, Sarkar T. A marketplace for data: an algorithmic solution. arXiv preprint arXiv:1805.08125. 2018 May 21.

[8] Kannan K, Ananthanarayanan R, Mehta S. What is my data worth? From data properties to data value. arXiv preprint arXiv:1811.04665. 2018 Nov 12.

[9] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data quality assessment. Commun. ACM 45, 4 (April 2002), 211-218. DOI=http://dx.doi.org/10.1145/505248.506010

[10] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daume III H, Crawford K. Datasheets for datasets. arXiv preprint arXiv:1803.09010. 2018 Mar 23.