

**Grant Agreement Number: 825225**

**Safe-DEED**

**[www.safe-deed.eu](http://www.safe-deed.eu)**

## **D4.5 Big Data Valuation Component v.2**

<b>Deliverable number</b>	<i>D4.5</i>
<b>Dissemination level</b>	<i>Consortium</i>
<b>Delivery date</b>	<i>30 November 2021</i>
<b>Status</b>	<i>Final</i>
<b>Author(s)</b>	<i>Mihnea Tufiş</i>



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

## Changes Summary

Date	Author	Summary	Version
10.11.2021	Mihnea Tufiş	First draft	0.1
17.11.2021	Mihnea Tufiş	Ready for consortium review	1.1
26.11.2021	Ioannis Markopoulos	NOVA comments	1.2
27.11.2021	Lukas Helminger	KNOW comments	1.3
28.11.2021	Mihnea Tufiş	Integrate KNOW review	1.4
28.11.2021	Mihnea Tufiş	Integrate NOVA review	1.5
30.11.2021	Mihnea Tufiş	Final updates.	2.0

## **Executive summary**

This document complements deliverable D4.5 – Release of the Big Data Valuation Component version 2. This is the third release of the Data Valuation Component (DVC) following that in deliverables D4.2 and D4.4. It is supported by an updated architecture from the one presented in D4.1 and D4.4 and includes an extension of the data quality assessment, a new data utility assessment module, and integrations of results from WP5, including the chance estimators, deanonymisation risk analysis and an in-browser application of private set aggregation protocols for the calculation of the contextual scores. A version of the software was included in deliverable D6.3 – Personal data trials report final version – conducted by NOVA (ex-FNET) [17].

## Table of Contents

<b>1</b>	<b>Introduction .....</b>	<b>7</b>
<b>2</b>	<b>Implementation of the DVC .....</b>	<b>7</b>
2.1	Updates from the previous version .....	8
2.2	Architecture.....	8
2.2.1	Data Ingestion Layer.....	8
2.2.2	Qualitative Information Extraction and Data Scoring Component (QDSC).....	9
2.2.2.1	Systems & Economics .....	9
2.2.2.2	Legal & Obligations .....	9
2.2.2.3	Data Science .....	9
2.2.2.4	Data Properties.....	9
2.2.2.5	Business .....	10
2.2.2.6	Legal & Obligations .....	10
2.2.2.7	Data Science .....	10
2.2.2.8	Data Properties.....	10
2.2.2.9	Business .....	10
2.2.3	Automatic Data Analysis and Scoring (ADAS) .....	15
2.2.3.1	Data quality assessment.....	15
2.2.3.2	Data utility assessment .....	18
2.2.3.3	Deanonymisation risk analysis .....	19
2.2.4	Score-to-Value Mapping (S2VM) .....	20
2.2.5	Communication and Presentation Layer (CPL).....	20
2.3	Data flow.....	21
<b>3</b>	<b>Deliverable details .....</b>	<b>22</b>
3.1	Package structure .....	22
3.2	Library requirements.....	22
3.3	Running the DVC .....	23
<b>4</b>	<b>Conclusions .....</b>	<b>23</b>
4.1	Safe-DEED Personal data demonstrators.....	23
4.2	Exploitation plan .....	23
4.3	Next steps .....	24
<b>5</b>	<b>References .....</b>	<b>25</b>
<b>6</b>	<b>Annex.....</b>	<b>27</b>

## List of Figures

Figure 1: DVC - Load the data file.....	9
Figure 2: Context establishing process – Step 1 – Systems & Economics.....	10
Figure 3: Context establishing process – Step 2 – Legal & Obligations.....	11
Figure 4: Context establishing process – Step 3 – Data Science.....	12
Figure 5: Context establishing process – Step 4 – Data Properties .....	13
Figure 6: Context establishing process – Step 5 – Business.....	14
Figure 7: Snippet of the context-value mappings for Layer 2 - Legal, Terms, Obligations.....	15
Figure 8: Domain validity assessment. Enumerate the possible values that can appear in each column where a domain rule applies. ....	16
Figure 9: Format validity assessment. Enumerate both the rule type and the rule itself for each column for which a format check will be performed. ....	16
Figure 10: Uniqueness assessment. Enumerate the subset of columns over which the uniqueness of a data point will be checked. ....	17
Figure 11: Timeliness assessment. For each column, define both a time reference column and a decay factor. ....	17
Figure 12: Data utility assessment - Classification. Define the configuration for training a classification algorithm with the input dataset. ....	18
Figure 13: Deanonymisation risk analysis. Setup involves: the choice of k (for applying k-anonymisation), the choice of QIs and a range of values for the acceptable level of risk. ....	19
Figure 14: Deanonymisation risk analysis. Select an anonymisation configuration, based on the desired subset of QIs and their corresponding risk. ....	20
Figure 15: The structure of the aggregation of various metrics and scores in the DVC. ....	20
Figure 16: DVC - Output scores and colour codes.....	21
Figure 17: Data flow diagram (DFD) for the Data Valuation Component (DVC) .....	22
Figure 18: Structure of the repository (left) and of the Flask templates for the GUI (right).....	22
Figure 19: Complete architecture of the Data Valuation Process. It includes the Data Valuation Component, the Deanonymisation service and the PSA service.....	27

---

## Abbreviations

ADAS	Automatic Data Analysis and Scoring Component
CPL	Communication and Presentation Layer
CSV	Comma Separated Values
DFD	Data Flow Diagram
DIL	Data Ingestion Layer
DoA	Description of Action
DQA	Data Quality Assessment
DVC	Data Valuation Component
EUT	Fundacio Eurecat
FNET	Forthnet S.A. (now Nova)
GUI	Graphical User Interface
IFX	Infineon Technologies AG
JSON	JavaScript Object Notation
KUL	KU Leuven
KNOW	KNOW Research Center
ML	Machine Learning
NOVA	Nova S.A. (ex-Forthnet S.A.)
PSA	Private Set Aggregation
RSA	Research Studios Austria
QDSC	Qualitative Information Extraction and Data Scoring
QI	Quasi-Identifier
S2VM	Score-to-Value Mapping
WP	Work Package
XLS	Microsoft Excel Spreadsheet
XML	Extensible Markup Language

## 1 Introduction

This document accompanies the software release of version 2 of the big data valuation component in deliverable D4.5. The code is available in the same archive as this document, is stored on the Safe-DEED git repository, the Safe-DEED Next Cloud and is available on request.

Our solution considers that the value of data is generated from two main areas: data quality and data usability, which are assessed through the lens of the context in which the data will be used. The context is set by the user, during a context gathering procedure, based on which the relevant components of data quality and data usability are established.

The tool is trying to maximise the automation degree of all these processes, thus allowing for more in-depth analyses to support the value of data and a reduction of the time dedicated to the data valuation process.

Since this is a complex problem, the presentation of the results avoids the generation of a single aggregate value. Instead, the platform generates a set of scores (for different perspectives and at different levels of detail), thus informing the user on the strengths and weaknesses of the dataset they are assessing.

The rest of the document is structured as follows: Section 2 describes the implementation details of the Data Valuation Component (DVC), including a description of the sub-components, the class diagram of the solution, the data flow between sub-components. Section 3 describes the structure, the dependencies and how to run the demonstrator package. Section 4 concludes the document and discusses the next steps in the development of the DVC.

This deliverable represents an updated version of the platform, and it follows D4.2 – Baseline prototype for data valuation released in November 2019 [20] and D4.4 – Big Data Valuation Component version 1, released in November 2020 [24]. The new features (see Section 2.1) are based on the results of the extensive state of the art review of methods for data valuation – deliverable D4.3 [23], extend the implementation of data quality assessment, and include a brand-new data utility assessment module. Additionally, it includes contributions from WP5: chance estimators, deanonymisation risk analysis and private set aggregation (PSA).

## 2 Implementation of the DVC

The goal of the DVC is to perform the valuation of a dataset over three aspects: data quality, data exploitability and economic value. The current version of the DVC receives a structured dataset, together with a context and rules for evaluating data quality, data utility, and the risk of deanonymisation (in the case of personal data). It returns three scores which describe the value of the data:

1. a score based on the contextual information provided
2. a score based on:
  - a. quality assessment
  - b. utility assessment
  - c. deanonymisation risk analysis
3. the aggregate value of the dataset, currently computed as a mean of the 2 previous scores.

In addition, the exchange of contextual information and the calculation of the contextual score (point 1 above) are achieved in a secure manner by implementing PSA protocols.

## 2.1 Updates from the previous version

The following is a list of the improvements to the Data Valuation Component (DVC) compared to version 1 of the platform in deliverable D4.4 [24].

1. Improvement of the architecture for the sub-components and the data flow, as well as subsequent code refactoring.
2. Implementation of new data quality metrics, as described in deliverable D4.3 [23]:
  - a. Timeliness. It is based on the age of each data point and a decay factor specific to each column.
  - b. Credibility.
  - c. Uniqueness.
3. Implementation of the new data utility module, which assesses the applicability of the dataset to train selected Machine Learning models: clustering, classification, regression.
4. Integration of the chance estimator developed by RSA as part of the data utility module.
5. Integration of the deanonymisation risk analysis developed by RSA as part of the privacy assessment module.
6. Integration of PSA protocols provided by KNOW, to achieve secure in-browser contextual valuation.
7. Adaptation of the GUI to each of the new functionalities described before.

## 2.2 Architecture

The DVC comprises the following modules, as illustrated in the data flow diagram in Figure 17.

1. Data Ingestion Layer (DIL);
2. Qualitative information extraction and Data Scoring Sub-Component (QDSC);
3. Automatic Data Analysis and Scoring (ADAS);
4. Score-to-Value Mapping (S2VM);
5. Communication and Presentation Layer (CPL).

The DVC needs to call two additional services:

- Anonymisation & deanonymisation risk analysis service.
- PSA service for securely computing contextual scores.

Figure 19 in Annex explains how all these components (DVC, deanonymisation service, PSA service) interact with each other to produce the complete Data Valuation Process.

### 2.2.1 Data Ingestion Layer

DIL represents the entry point of the data to the platform. The data is currently ingested via a GUI, where the path to the dataset is specified. The layer detects the data format and performs the suitable operations for uploading it. Currently, the only supported formats are CSV and XLS(X), with the capacity of choosing a specific datasheet for the latter.



**Figure 1: DVC - Load the data file**

## 2.2.2 Qualitative Information Extraction and Data Scoring Component (QDSC)

The first stage of the data valuation process directly involves the user, who is required to provide information about the context in which to value an input dataset. The context valuation is achieved through a multi-step questionnaire focusing on the following aspects:

1. Systems & Economics: Availability & access, Purpose
2. Legal & Obligations: Data protection, Legal-Terms-Obligations
3. Data Science: Tools, Format
4. Data Properties: Data velocity, Data transformations, Data quality, Data age
5. Business: Frequency of use, Benefits

### 2.2.2.1 Systems & Economics

The first section of the QDSC context establishing process collects information related to two main aspects surrounding the system and economics of the input dataset: availability and access of the dataset, and a declaration of the purpose of its usage (see Figure 2).

### 2.2.2.2 Legal & Obligations

The content of this section of the QDSC context establishing process is the result of the cooperation with our colleagues from WP3. Its objective is to collect the main information with respect to the data processing and legal practices concerning the input dataset (see Figure 3).

### 2.2.2.3 Data Science

This section is used to collect information about i) the tools necessary to perform the desired data science tasks, and ii) the format of the original data (see Figure 4).

### 2.2.2.4 Data Properties

This section collects information about relevant properties of the dataset, such as velocity, transformations, quality, age (see Figure 5).

## Systems &amp; Economics

Dashboard / Systems &amp; Economics

This section collects information about the following aspects related to the data set:

1. Availability & Access
2. Purpose

## Availability &amp; Access

Is this data easily accessible by all? ☐ Yes ☐ NoHow was the data generated? How many sources compose the data? ☐ Single source ☐ Aggregated / Multiple sourcesIs the data **enterprise** generated? ☐ Yes ☐ NoIs the data **machine** generated? ☐ Yes ☒ NoIs the data publicly available? ☐ Yes ☐ NoAre there known alternatives to the data set? ☐ Yes ☐ No

## Purpose

What do you want to achieve with the data?

- ☒ Spreadsheet-style calculations
- ☐ Charts and other visualizations
- ☒ ML - Classification
- ☐ ML - Regression
- ☐ ML - Clustering

Has it already been used for any of these?

- ☐ Spreadsheet-style calculations
- ☒ Charts and other visualizations
- ☐ ML - Classification
- ☐ ML - Regression
- ☐ ML - Clustering

Is the data representing time series? ☐ Yes ☒ No

Submit

Submit Layer 1 context and continue to Layer 2

Figure 2: Context establishing process – Step 1 – Systems &amp; Economics

## 2.2.2.5 Business

This is the final QDSC section and establishes the business context for the dataset. It requires information about the frequency of use of the input data and the expected benefits for various business areas.

## 2.2.2.6 Legal &amp; Obligations

The content of this section of the QDSC context establishing process is the result of the cooperation with our colleagues from WP3. Its objective is to collect the main information with respect to the data processing and legal practices concerning the input dataset (see Figure 3).

## 2.2.2.7 Data Science

This section is used to collect information about i) the tools necessary to perform the desired data science tasks, and ii) the format of the original data (see Figure 4).

## 2.2.2.8 Data Properties

This section collects information about relevant properties of the dataset, such as velocity, transformations, quality, age (see Figure 5).

## 2.2.2.9 Business

This is the final QDSC section and establishes the business context for the dataset. It requires information about the frequency of use of the input data and the expected benefits for various business areas.

## Data protection

Is the purpose of the processing activity clearly defined? ☒ Yes ☐ No

Does the dataset contain personal data? ☒ Yes ☐ No [i](#)

Does the dataset contain any of the following?

Business confidential information:

☐ Revenue or financial data

☐ Data covered by copyright

☐ Trade secrets

Personal data:

☐ Sensitive data (e.g., medical)

☒ Digital traces (e.g., cookies, IP address, MAC address, search or browsing history etc.)

☒ Any other personal data

Was the processing of personal data carried out respecting privacy and data protection principles?

☒ Yes ☐ No ☐ Not applicable (no personal data) [i](#)

Is the controller able to demonstrate compliance with the EU Data Protection Principles

☒ Yes ☐ No ☐ Not applicable (no personal data) [i](#)

Have data subjects been given all the necessary information about processing their data in order to exercise their rights?

☒ Yes ☐ No ☐ Not applicable [i](#)

Have appropriate technical and organisational measures been implemented to ensure a level of security appropriate to the risk?

☒ Yes ☐ No [i](#)

Is the data encrypted? ☐ Yes ☒ No

Is the data anonymized? ☒ Yes ☐ No

Figure 3: Context establishing process – Step 2 – Legal &amp; Obligations

## Data Science

[Dashboard](#) / Data Science

This section collects information about the following aspects related to the data set:

1. Tools
2. Format

### Tools

Are there tools to clean and process this data? ☒ Yes ☐ No

Are there tools to profile and analyse the data from its current format? ☒ Yes ☐ No

Is there support concerning the use of these tools? Large ▼

### Format

What is the format of the data set file? Tabular file (CSV, TSV) ▼

Does it have a schema? ☒ Yes ☐ No

Does it adhere to a standard? Yes - Open ▼

Does it result from an export or a query from a relational database? ☒ Yes ☐ No

Is it in normalized form? ☐ Yes ☒ No ☐ Not available (Not in relational form)

**Figure 4: Context establishing process – Step 3 – Data Science**

## Data Properties

[Dashboard](#) / Data Properties

This section collects information about the following aspects related to the data set:

1. Data velocity
2. Data transformations
3. Data quality
4. Data age

### Data velocity

Is it streaming data? ☐ Yes ☒ No

How fast is the data generated?

### Data transformations

Any known transformations already done?

#### 1. Structure

- ☐ Add columns / rows
- ☐ Remove columns / rows
- ☐ Rename columns / index

#### 2. Missing values

- ☒ Drop
- ☐ Imputation

#### 3. Derived columns

- ☐ Relative values
- ☐ Aggregates
- ☐ Interpolations
- ☐ Combinations of existing columns
- ☐ Combinations of newly generated columns

Any known analyses already done and available?

- ☒ Univariate
- ☐ Bivariate
- ☐ Multivariate
- ☐ Features selection

### Data quality

Are all relevant fields complete? ☐ Yes ☒ No

Are all relevant fields error-free? ☒ Yes ☐ No

Are there known missing records? ☒ Yes ☐ No [?](#)

Are there known missing values in relevant fields? ☒ Yes ☐ No [?](#)

Does it have duplicates in relevant fields? ☐ Yes ☒ No

Does it complement or supplement an existing data set? ☐ Yes ☒ No [?](#)

Is the data noisy? ☐ Yes ☒ No

### Data age

How recent is the data set?

Is there a known later version of the data? ☐ Yes ☒ No

**Figure 5: Context establishing process – Step 4 – Data Properties**

## Frequency of use

Prior to the current use, when was the data used last time? Year

After the current use, when will the data be used next time? Year

## Benefits

Is the data already making money? ☒ Yes ☐ No

Will it improve the efficiency of an existing application or business process? ☒ Yes ☐ No

Does it complement an existing application? ☐ Yes ☒ No

Does it introduce a new channel to reach new customers? ☐ Yes ☒ No

Does it improve customer reach? ☒ Yes ☐ No

Which part of the business process does it contribute to?

- ☒ Sales
- ☒ Marketing
- ☐ Accounting
- ☒ Payroll
- ☒ Technical (includes R&D)
- ☐ Others

Which part of the organization will directly use this data?

- ☒ Executive
- ☒ Middle management
- ☐ Others

**Figure 6: Context establishing process – Step 5 – Business**

In the interest of reproducibility of the valuation process, the context is encoded as a JSON file and available for download and reuse. Finally, the component computes and returns a context-based score (QDSC-score).

The method for computing this score is inspired by current research on mapping data properties to data value [12] or datasheets for datasets [16], both of which are discussed in detail in deliverable D4.3 [23]. Essentially, there is a mapping between each potential answer to each question in the five steps before and a score between 0 and 1. At the end of the form, all the answers are summed up and the percentage score with respect to the maximum possible is computed.

```

"layer_2": {
  "protection": {
    "has_purpose": {
      "yes":1,
      "no":0
    },
    "is_personal": {
      "yes":0,
      "no":1
    },
    "business": {
      "revenue":0.34,
      "copyright":0.33,
      "trade":0.33
    },
    "personal": {
      "sensitive":0.34,
      "traces":0.33,
      "other":0.33
    },
    "is_principles_compliant": {
      "yes":1,
      "no":0,
      "NA":1
    },
  },

```

Figure 7: Snippet of the context-value mappings for Layer 2 - Legal, Terms, Obligations

## 2.2.3 Automatic Data Analysis and Scoring (ADAS)

It starts by preparing the loaded data and performs a set of analytic operations to extract the intrinsic properties of the dataset:

- Data shape (number of lines and columns) and size.
- Inference of the most plausible data type for each column (string, integer, float, datetime, etc.).
- The profile of each field:
  - missing values (percentage of missing values per column, using default markers – i.e., *NULL*, *NA*, *nan*, etc.).
  - distribution of the data from each field (histogram of the distribution of the possible values in each column).

### 2.2.3.1 Data quality assessment

Based on the information extracted about the structure of the dataset, the ADAS then initiates a multi-step process for collecting any data quality rules applicable to each of the columns of the dataset. The current version of the component performs data quality assessment using the following data quality measures:

- **Completeness.** Defined as the percentage of missing values or equivalent. The user has the option to define what values should be considered as missing values. By default, values of *NaN*, *nan*, *None*, *NA* or empty strings qualify as such.

- **Domain validity.** Defined as the percentage of values in a user-defined range / domain. If no such range is specified, all values are interpreted as valid.

Safe-DEED DVC

PRE-VALUATION ASSESSMENT

- Systems & Economics
- Legal & Obligations
- Data Science
- Data Properties
- Business

Logged in as:  
Start Bootstrap

## Data Quality Assessment :: Domain validity

Dashboard / Data Quality Assessment / Validity / Domain

Define the domain (range) for each field of the data set.  
Leave empty if no specific domain applies for the corresponding field.

### Data Fields

CUSTOMER_ID	
BILLING_ACCOUNT_ID	
ASSET_ID	
ACTIVATION_DATE	
DEACTIVATION_DATE	
CITY	
STREET	
ZIP_CODE	
PERFECTURE_DESC	
ASSET_STATUS_ID	
INITIATION_CHANNEL	απόστημα, Call Center External
INITIATION_DEALER_ID	
PORTABILITY	Yes, No
LOOP_TYPE	
INTEGRATION_ID	
ASSET_STATUS_REASON	για την υπηρεσία, Άλλο, Ελλιπές
ASSET_STATUS_REASON_DESCR	είο (mail), Από φόρμα διακοπής
PROVIDER_DEST	
PROVIDER_SOURCE	

**Figure 8: Domain validity assessment. Enumerate the possible values that can appear in each column where a domain rule applies.**

- **Format validity.** Defined as the % of values that respect specific user-defined formatting rules. These rules follow a specific grammar: datetimes or regular expressions.

Safe-DEED DVC

PRE-VALUATION ASSESSMENT

- Systems & Economics
- Legal & Obligations
- Data Science
- Data Properties
- Business

Logged in as:  
Start Bootstrap

## Data Quality Assessment :: Format validity

Dashboard / Data Quality Assessment / Validity / Format

Define the format type and format rule expected for each column.  
Leave a pair empty if no format rule applies for the corresponding field.

### Data Fields

CUSTOMER_ID	str:regex	^\d-[A-Z0-9]{5}\$
BILLING_ACCOUNT_ID	str:regex	^\d-[A-Z0-9]{5}\$
ASSET_ID	str:regex	^\d-[A-Z0-9]{5}\$
ACTIVATION_DATE	datetime	%d/%m/%Y %H:%M:%S
DEACTIVATION_DATE	datetime	%d/%m/%Y %H:%M:%S
CITY		
STREET		
ZIP_CODE		
PERFECTURE_DESC		
ASSET_STATUS_ID		
INITIATION_CHANNEL		
INITIATION_DEALER_ID		
PORTABILITY		
LOOP_TYPE		
INTEGRATION_ID	str:regex	^\d-[A-Z0-9]{5}\$
ASSET_STATUS_REASON		
ASSET_STATUS_REASON_DESCR		
PROVIDER_DEST		
PROVIDER_SOURCE		

**Figure 9: Format validity assessment. Enumerate both the rule type and the rule itself for each column for which a format check will be performed.**



- **Uniqueness.** Defined as the % of non-duplicate values, with respect to a subset of columns.

**Figure 10: Uniqueness assessment. Enumerate the subset of columns over which the uniqueness of a data point will be checked.**

- **Credibility.** Defined as the percentage of values different from user-defined default values. If no such values are defined, all values will be considered credible.
- **Timeliness.** Defined by the following formula [11][23]:

$$Timeliness = e^{-(current\ timestamp - arrival\ timestamp) \times decay\ factor}$$

Records are compared to a reference column representing the arrival timestamp of the record. This is moderated by a decay factor, representing the rate at which the data in each column is aging. Both the reference column and the decay factor for each column are user-defined.

Data Fields	Reference Column	Decay Factor
CUSTOMER_ID	ACTIVATION_DATE	0.2
BILLING_ACCOUNT_ID	ACTIVATION_DATE	0.2
ASSET_ID	ACTIVATION_DATE	0.2
ACTIVATION_DATE	ACTIVATION_DATE	0.2
DEACTIVATION_DATE	DEACTIVATION_DATE	0.2
CITY	ACTIVATION_DATE	0.2
STREET	ACTIVATION_DATE	0.2
ZIP_CODE	ACTIVATION_DATE	0.2
PERFECTURE_DESC	ACTIVATION_DATE	0.2
ASSET_STATUS_ID	ACTIVATION_DATE	0.2
INITIATION_CHANNEL	ACTIVATION_DATE	0.2
INITIATION_DEALER_ID	ACTIVATION_DATE	0.2
PORTABILITY	ACTIVATION_DATE	0.2
LOOP_TYPE	ACTIVATION_DATE	0.2
INTEGRATION_ID	ACTIVATION_DATE	0.2
ASSET_STATUS_REASON	ACTIVATION_DATE	0.2
ASSET_STATUS_REASON_DESCR	ACTIVATION_DATE	0.2
PROVIDER_DEST	DEACTIVATION_DATE	0.2
PROVIDER_SOURCE	ACTIVATION_DATE	0.2

**Figure 11: Timeliness assessment. For each column, define both a time reference column and a decay factor.**

Just like in the case of contexts, in the interest of reproducibility, the data quality rules are encoded as a JSON file and available for download and reuse.

Next, the ADAS uses the loaded data and performs the data quality assessment against the provided quality rules. Quality scores are subsequently computed for each data quality dimension, which are finally aggregated (as a mean) into an ADAS-score.

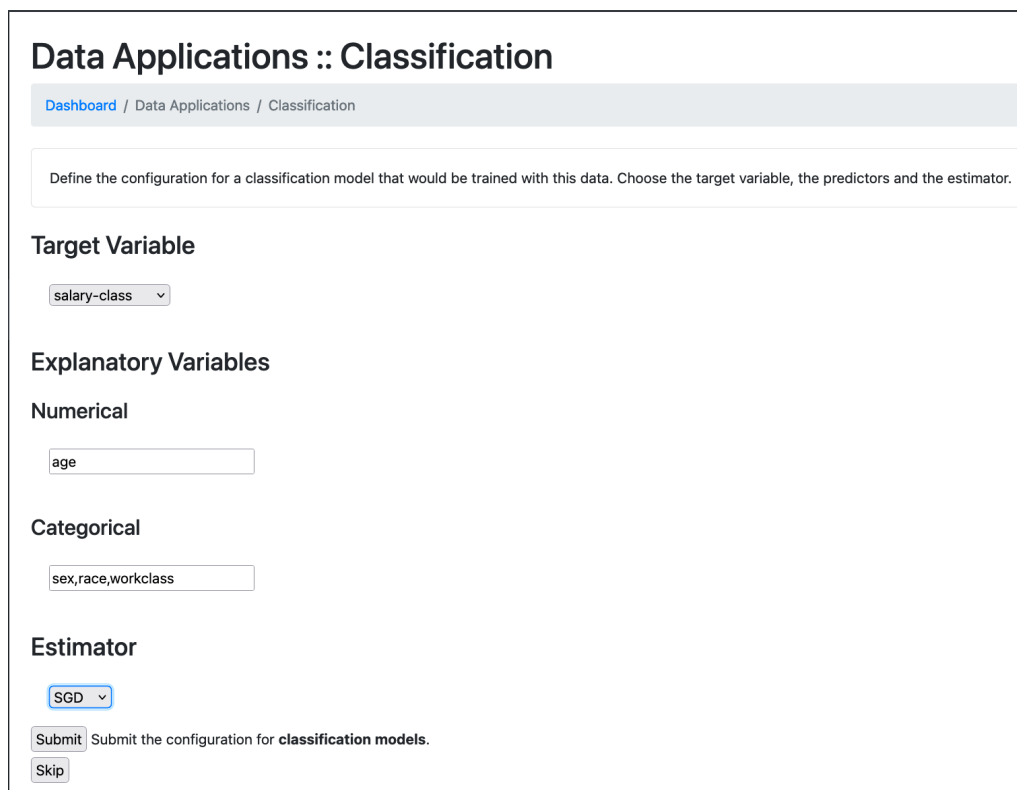
This employs a classic DQA framework, where each selected quality dimension is supported by one or several quality metrics [5][8]. For more details, please refer to the extended review of Data Quality Assessment methods in Section 4 of deliverable D4.3 [23].

Once the user has parametrised the rules for the data quality metrics (DQD) corresponding to each column, our method checks the compliance of each column against the rule that applies to it. The result is a score between 0 and 1 for each column and each DQM. The scores are first aggregated over all columns [11][13] to obtain a data set level score for each of the six DQMs; these scores are then averaged (with equal weights) to obtain the Data Quality Score.

### 2.2.3.2 Data utility assessment

This module allows the user to assess the suitability of the dataset for training machine learning models for classification, regression, and clustering. In this version, the module allows for the instantiation of out of the box algorithms implemented in the scikit-learn library<sup>1</sup>:

- Regression: Ridge Regression.
- Classification: SGD Classifier.
- Clustering: MeanShift, KMeans.



**Figure 12: Data utility assessment - Classification. Define the configuration for training a classification algorithm with the input dataset.**

The utility score is produced by evaluating the performance of the model on the training sample using the following measures:

- Regression and Classification: R2 score.
- Clustering: Calinsky-Harabasz, Davies-Bouldin and Silhouette scores.

Additionally, the classification assessment includes the implementation of RSA's chance estimator [20], which gives a measure of how likely it is for the result of a trained classifier to be attributed to

<sup>1</sup> <https://scikit-learn.org/stable/index.html>

chance. This result is particularly interesting since it is calculated irrespective of the type of classification algorithm and can potentially be extended to regression [21].

The Data Utility Score is calculated as an average of the evaluation scores, depending on the number of ML algorithms included in the assessment.

### 2.2.3.3 Deanonimisation risk analysis

Deanonimisation risk analysis represents an initial attempt to in privacy as a dimension of data value. The tool developed by our WP5 partners at RSA is the first to be integrated within the privacy assessment module of the DVC, allowing for the quantification of the risk associated to the deanonimisation of sensitive columns. Thus, users are asked to choose:

- A value for  $k$  for the application of the  $k$ -anonymisation algorithm. This will be between 2 and the number of columns in the dataset. Leaving any side of the interval empty will force the use of the default min/max values for  $k$ .
- A subset of columns from the dataset, which will be considered as *quasi-identifiers* (QIs), when applying  $k$ -anonymisation.
- An acceptable range of values (0 – 100) for the *risk* of the data to be de-anonymised. Leaving any sides of the interval empty will force the use of the min/max value for the risk.

The screenshot shows the 'Safe-DEED DVC' interface. On the left is a dark sidebar with a menu: 'PRE-VALUATION ASSESSMENT', 'Systems & Economics', 'Legal & Obligations', 'Data Science', 'Data Properties', and 'Business'. Below the menu, it says 'Logged in as: Start Bootstrap'. The main content area has a title 'Privacy :: De-anonymisation Risk Analysis' and a breadcrumb 'Dashboard / Privacy / De-anonymisation Risk Analysis'. Below the title, there's a text box: 'Define the configuration for the de-anonymisation risk analysis. Please specify any of:'. It lists three bullet points: 'a range of values for k - integer between 2 and 9', 'a subset of columns (separated by commas) that you would want to include as QIs', and 'a range of values for the acceptable risk - float between 0 and 100'. Below this, there are three input fields: 'K from: 2 to 5' (with dropdown arrows), 'Choose the QI cols: race,workclass,marital-status', and 'Risk between: 0 Risk between: 45'. Each field has a note: '\* Leave empty if you want to use an open range on either side.' At the bottom, there are two buttons: 'Submit' and 'Skip'. The 'Submit' button is followed by the text 'Submit the configuration for de-anonymisation risk analysis.'

**Figure 13: Deanonimisation risk analysis. Setup involves: the choice of  $k$  (for applying  $k$ -anonymisation), the choice of QIs and a range of values for the acceptable level of risk.**

The result is the set of all possible combinations of QIs and their corresponding risk within the user-defined range. The user is then asked to choose the combination of QIs most suitable to the context. This choice can be made with respect to both the desired subset of QIs to use in the  $k$ -anonymisation and the (accepted) risk associated to these; the latter becomes the privacy assessment score.

Safe-DEED DVC

Search for...

PRE-VALUATION ASSESSMENT

- Systems & Economics
- Legal & Obligations
- Data Science
- Data Properties
- Business

Logged in as: Start Bootstrap

## Privacy :: De-anonymisation Risk Analysis :: Configuration

Dashboard / Privacy / De-anonymisation Risk Analysis

Select one of the following configurations, based on:

- the desired columns that you would want to include as QIs
- the desired k
- the acceptable value of de-anonymisation risk

QI set	k	Risk level	
['race', 'marital-status']	2	0.0	<input type="radio"/>
['race', 'workclass']	2	0.01	<input type="radio"/>
['workclass', 'marital-status']	2	0.01	<input type="radio"/>
['race', 'workclass', 'marital-status']	3	0.08	<input checked="" type="radio"/>

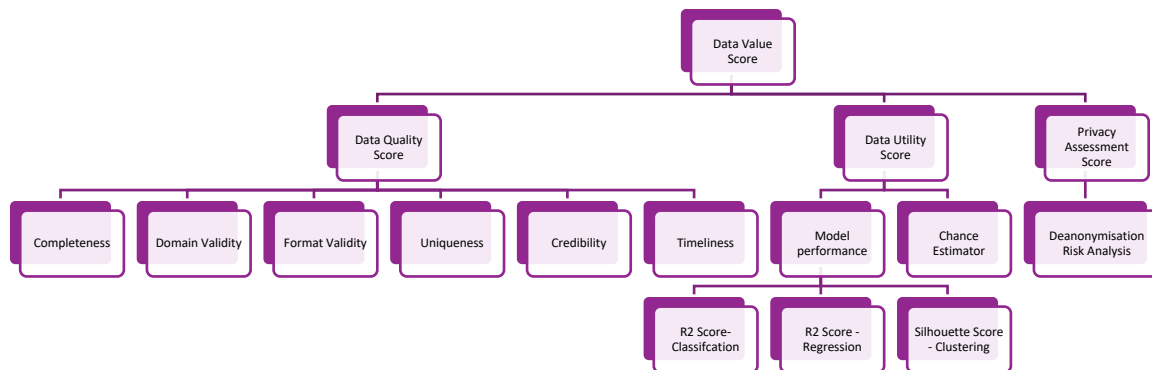
Submit Submit the configuration for de-anonymisation risk analysis.

**Figure 14: Deanonimisation risk analysis. Select an anonymisation configuration, based on the desired subset of QIs and their corresponding risk.**

Finally, the 3 previous scores (aggregated Data Quality Score, aggregated Data Utility Score, privacy assessment score) are aggregated in the ADAS-score, as an equally weighted average.

### 2.2.4 Score-to-Value Mapping (S2VM)

Using the two previous scores (QDSC-score and ADAS-score), this component aggregates them into a final score (as an average) – the data value [11]. While this method is still simplistic, it validates the full data flow through the component and leaves the door open for the addition of economic models for data valuation, which will be integrated for the final version of the DVC.

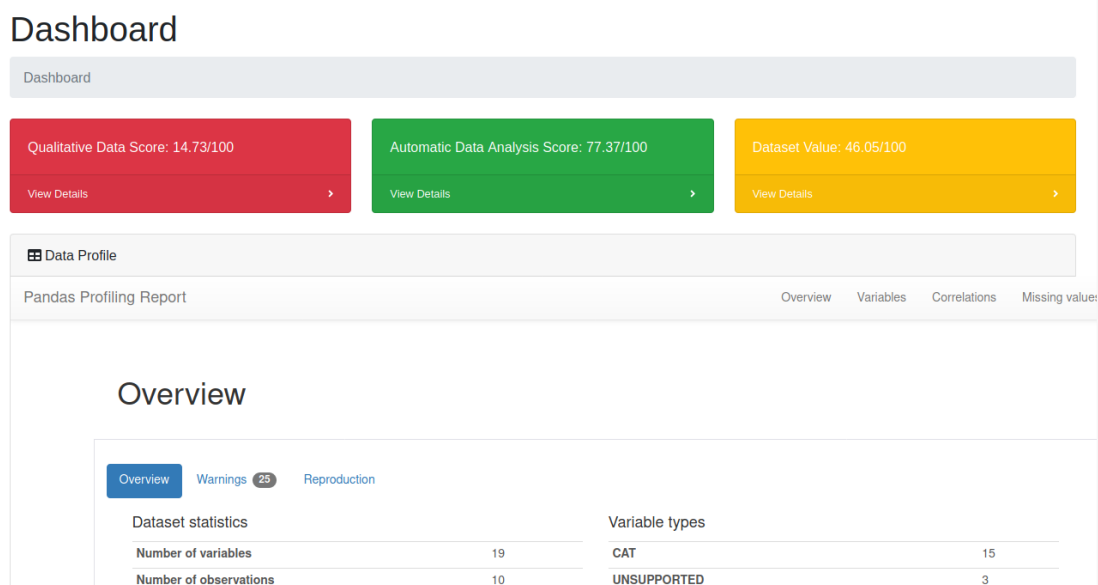


**Figure 15: The structure of the aggregation of various metrics and scores in the DVC.**

### 2.2.5 Communication and Presentation Layer (CPL)

This layer acts as an interface to report the results of the data valuation process. In an integrated GUI, it displays the following (see Figure 16):

- The QDSC and ADAS scores, together with a 3-colors code.
- The aggregated value of the dataset, together with a 3-colors code.
- A report of the dataset profile, available both as part of the GUI (HTML), as well as in PDF format.



**Figure 16: DVC - Output scores and colour codes**

## 2.3 Data flow

The DVC input/output structure is given below, together with the data flow diagram (Figure 17) for the entire component.

Input:

- A dataset (or a snapshot thereof) (only CSV and XLS(X) currently supported). Future versions will include support for semi-structured datasets (XML, JSON).
- User provided context.
- User defined data quality rules.
- User provided parametrisations of ML algorithms to be trained with the dataset

Output:

- A set of scores, which evaluate the input dataset from three perspectives: contextual data valuation, data quality assessment, aggregated data value.
- A set of reports based on the analysis of the intrinsic properties of the dataset (format, shape, data types, missing values, duplicates).
- Additionally, both user-provided context and user-defined data quality rules are stored and exchanged as JSON files.

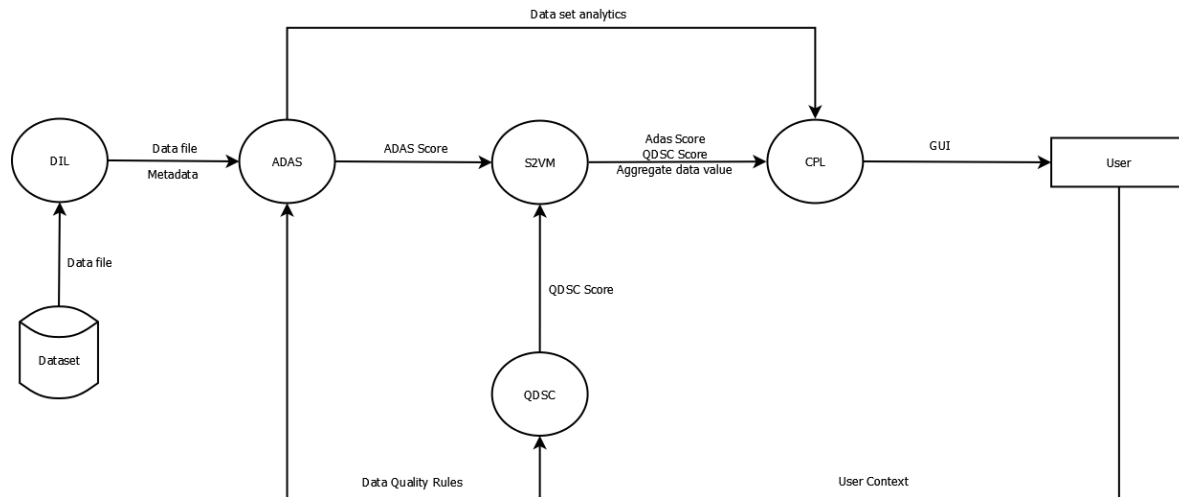


Figure 17: Data flow diagram (DFD) for the Data Valuation Component (DVC)

## 3 Deliverable details

### 3.1 Package structure

The structure of the package that forms deliverable D4.5 is depicted in Figure 18.

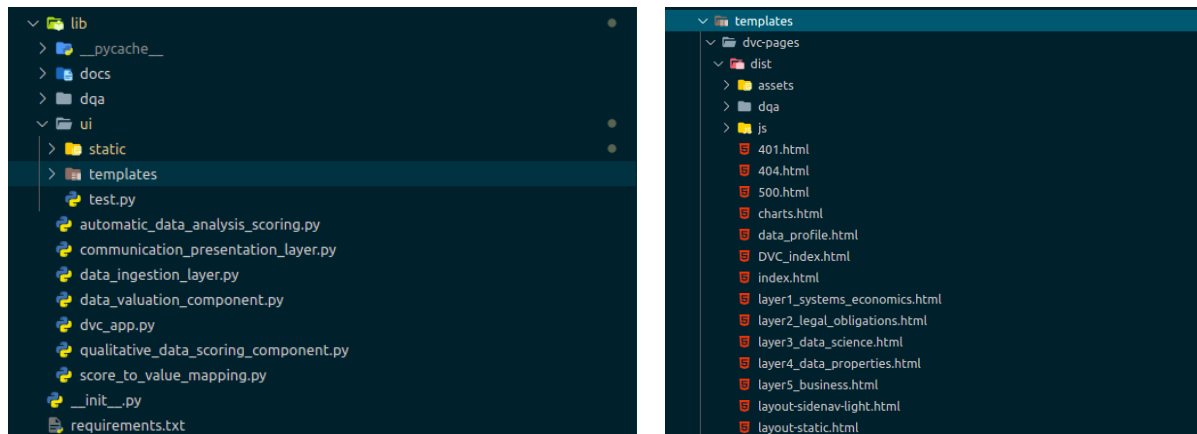


Figure 18: Structure of the repository (left) and of the Flask templates for the GUI (right)

### 3.2 Library requirements

The current version of the prototype was developed using Python 3.7, runs as a Flask web app and requires the libraries enumerated in the file **requirements.txt**, included in the deliverable. At this stage, there are no particular requirements with respect to the operating system. Below we give a summary of the most important libraries necessary to run the DVC.

flask=1.1.2	numpy=1.19.1	scikit-learn=0.24.2
flask-cors=3.0.10	pandas=1.1.1	
matplotlib=3.3.1	pandas-profiling=2.9.0	

Additionally, the DVC interacts with the following versions of the libraries provided by WP5:

PSA=2.1.3 (NodeJS)	prioprivacy=0.1.8 (Springboot)
--------------------	--------------------------------

### 3.3 Running the DVC

This section describes how to run the current DVC prototype. This is more straightforward and requires less technical expertise than the previous version in D4.2 [20].

0. Make sure that all prerequisite libraries in *requirements.txt* are properly installed.
1. Unzip the file in the same archive or download the DVC package from the Safe-DEED Git repository.
2. Open a terminal window
3. Navigate to the *DVC* project folder and run the following command:  

```
> python3 lib/dvc_app.py
```

## 4 Conclusions

Deliverable D4.5 represents version 2 of the Data Valuation Component, a major improvement from the previous version in D4.4 [24]. The implementation is based on our research on context formalisation (see Section **Error! Reference source not found.**), methods for data valuation [23] and includes a module for assessing legal, ethical and privacy issues (see Section 2.2.2.5). The data valuation process incorporates modules for data quality assessment, ML utility assessment, deanonymisation risk analysis and relies on PSA protocols for the secure exchange of contextual information.

### 4.1 Safe-DEED Personal data demonstrators

A strategic decision was taken in the project to implement two versions of the demonstrator:

1. The first version is fully functional aiming at being used within the professional community building actions. This demonstrator version is characterised as confidential according to the project DoA.
2. The second version is used to address the wide audience through the project web site. This demonstrator version is public and is implemented beyond the DoA requirements. It exposes the complete functionality with all the confidential demonstrator scenarios and datasets to the wide community, but users are not able to upload their own data due to security reasons.

The DVC was incorporated in both demonstrators, documented in deliverables D6.2 and D6.3 [17]. Several dissemination activities conducted by NOVA were successful in gathering important user feedback on the different Safe-DEED tools, including the DVC. Among them, we mention

- Internal NOVA workshop based on the content from the demonstrators, involving professionals from various business areas (marketing, product development, analysis departments).
- Interviews with external professionals, working in various industries (consultancy, pharmaceutical, IT).
- Participation in the Horizon 2020 ReachOut project<sup>2</sup>, in which external beta testers interacted with the Safe-DEED tools, following definition of testing scenarios.

### 4.2 Exploitation plan

EUT plans to continue developing the DVC in the form of a data valuation platform. This platform will be implemented as a SaaS, with 2 deployment options currently contemplated:

1. A remote data valuation platform, hosted by EUT or
2. A downloadable solution, that an interested company can install and use within their premises.

---

<sup>2</sup> <https://www.reachout-project.eu/>

The platform was already included in the Horizon Results Booster<sup>3</sup> program, during which it received valuable guidance about the exploitation possibilities of the Data Valuation Platform, improvement of the maturity of the solution and started the development of a business plan.

Following these sessions, we were able to outline an exploitation roadmap for the near and mid-term future:

1. Implementation of the new UI/UX. This will make the project more marketable and encourage user adoption.
2. Implementation of the backend in a production-like environment. This will need to balance hosting costs and scaling necessities, which we assume will dramatically increase with user adoption.
3. Deciding on the licensing model, specific to each of the deployment scenarios.
4. Finding the initial user base (early adopters) for in-depth usage of the platform.

### 4.3 Next steps

Expanding on the points presented in the exploitation plan, here is a list of the next planned developments for the near future.

1. Improve the user interface and user experience of the platform. We have already started the working on a new and appealing UI/UX. These are meant to build a visual identity of the platform, improve the context gathering experience, speed up the user interaction for the data quality assessment process and present the results in a more intuitive, actionable way.
2. Extend the capabilities of the current data quality dimensions. For example, in the case of format validity, we plan to extend its use to allow for additional types of rules beyond the ones currently supported (string regex and datetime).
3. ML utility is currently evaluated against a small number of representative algorithms. A natural extension is to either
  - a. expand the list of algorithms that a user can choose to include in the assessment or
  - b. to offer a user the possibility to upload custom ML models and corresponding metrics.
4. Aggregate measures. Improve the currently basic aggregate measures by using economic models for data value, which make use of the declared context.

Finally, the success of the Data Valuation Platform is dependent on a continuous research effort. We have outlined the main research directions and their impact on the improvement of the DVC:

1. Integrating economic methods for mapping data properties to an economic estimate of the value of data. These could also lead to an improvement of the current aggregate measures, by assigning weights according to contextual information.
2. Developing recommender systems for contexts and data sets. In this scenario – which would be very suited for data markets – we rely on the contextual information, as well as the data quality and usability parameters to:
  - a. Recommend a dataset to a user that provides a “valuable context”
  - b. Recommend a context and a potential “buyer” for any dataset, based exclusively on its data quality and data usability assessments.
3. Develop an improved data usability assessment, by creating underlying ontologies of data uses and data value [2] [3].
4. Include measures of algorithmic fairness and trust.
5. Expand the capabilities of the privacy assessment module, operationalising the results of theoretical and empirical research on the economics of privacy [1].

---

<sup>3</sup> <https://www.horizonresultsbooster.eu>

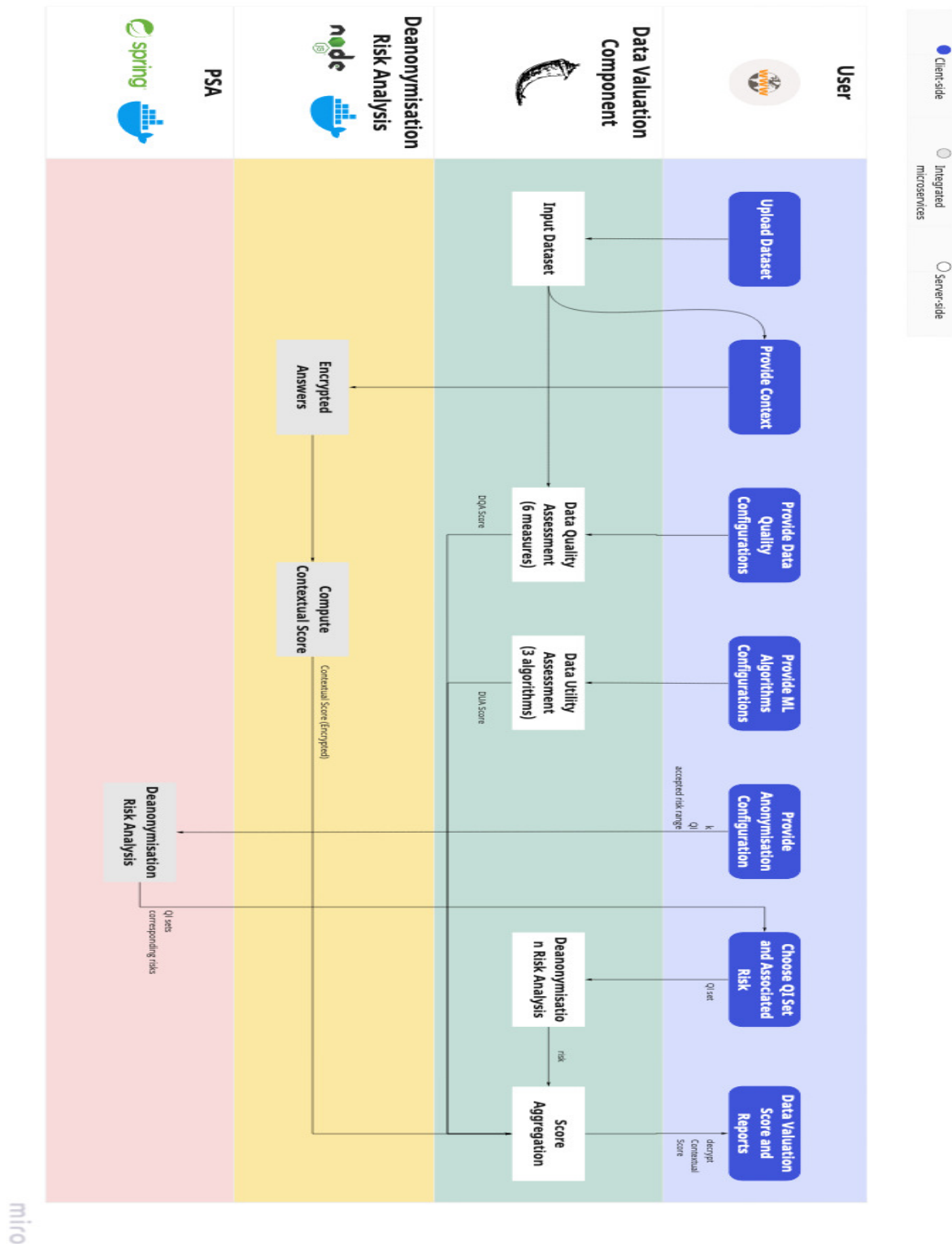


## 5 References

- [1] Acquisti, A., Taylor, C., Wagman, L. (2016). The Economics of Privacy. In: *Journal of Economics Literature*. pp. 442-492. 54-2.
- [2] Attard, J. and Brennan R. (2019). DaVe: A Semantic Data Value Vocabulary to Enable Data Value Characterisation. In: *Enterprise Information Systems*. pp. 239-261.
- [3] Attard, J., Orlandi, F., Auer, S. (2017). Exploiting the Value of Data through Data Value Networks. In: *Proceedings of the 10<sup>th</sup> International Conference on Theory and Practice of Electronic Governance*. pp. 475-484. New Delhi, India.
- [4] Bampoulidis, A. (2020). D5.10 – Report on the reidentification techniques on use-case data v2. H2020 Safe-DEED deliverable.
- [5] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, 41(3).
- [6] Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*. pp. 4349–4357.
- [7] Buolamwini, J., Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Friedler, S.A., Wilson, C. (eds). New York, NY, USA: PMLR; 2018. pp. 77–91.
- [8] Cai, L., and Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14(2), 1–10.
- [9] Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W., Choi, Y., Liang P., and Zettlemoyer, L. (2018). QuAC: Question Answering in Context. *CoRR*.
- [10] Erkut Erdem. 2018. Datasheet for RecipeQA.
- [11] Even, A., and Shankaranarayanan, G. (2006, November 10). Value-Driven Data Quality Assessment. *Proceedings of the 2005 International Conference on Information Quality*. MIT IQ Conference, MIT, Cambridge, MA, USA
- [12] Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. (2020). Datasheets for Datasets. <http://arxiv.org/abs/1803.09010>
- [13] Görz, Q., and Kaiser, M. (2012). An Indicator Function for Insufficient Data Quality – A Contribution to Data Accuracy. In H. Rahman, A. Mesquita, I. Ramos, and B. Pernici (Eds.), *Knowledge and Technologies in Innovative Information Systems* (Vol. 129, pp. 169–184). Springer Berlin Heidelberg.
- [14] Grass, A., Helminger, L., Schmid, F. (2021). D5.11 – Protocols for Privacy-Preserving Data Analytics and Secure Lead-Time Based Pricing v2/2. H2020 Safe-DEED deliverable.
- [15] Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2020). The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards. In D. Hallinan, R. Leenes, S. Gutwirth, and P. De Hert (Eds.), *Data Protection and Privacy: Data Protection and Democracy* (pp. 1–26). Oxford: Hart Publishing.
- [16] Kannan, K., Ananthanarayanan, R., and Mehta, S. (2018). What is my data worth? From data properties to data value. <http://arxiv.org/abs/1811.04665>
- [17] Markopoulos, I. (2020). D6.3 – Personal data trials report final version. H2020 Safe-DEED deliverable.
- [18] Paraschiv, M., Boratto, L., Kassa, Y., Tufiş M. (2019). D4.1 – Report on requirements and design. H2020 Safe-DEED deliverable.

- [19] Seck, I., Dahmane, K., Duthon, P., and Loosli, G. (2018). Baselines and a datasheet for the Cerema AWP dataset. CoRR. <http://arxiv.org/abs/1806.04016>
- [20] Taha, A.A., Bampoulidis, A., Lupu, M. (2019). Chance influence in datasets with a large number of features. In Haber P., Lampoltshammer T., Mayr M. (eds.), Data Science – Analytics and Applications (pp. 21-26). Springer Vieweg, Wiesbaden.
- [21] Taha A.A., Papariello, L., Bampoulidis, A., Knoth, P., Lupu, M. (2021). Formal Analysis and Estimation of Chance in Datasets Based on Their Properties. In IEEE Transactions on Knowledge and Data Engineering.
- [22] Tufiş, M., Boratto, L. (2019). D4.2 – Baseline prototypes for data valuation. H2020 Safe-DEED deliverable.
- [23] Tufiş, M., Boratto, L. (2020). D4.3 – Report on the context-aware and context-unaware valuation. H2020 Safe-DEED deliverable.
- [24] Tufiş, M. (2020). D4.4 – Big Data Valuation Component v.1. H2020 Safe-DEED deliverable.

## 6 Annex



**Figure 19: Complete architecture of the Data Valuation Process. It includes the Data Valuation Component, the Deanonimisation service and the PSA service.**