Data Pitch

H2020-ICT-2016-1

Project number: 732506

D2.3 [Updated DaaS platform]

Coordinator: Stefano Modafferi (University of Southampton IT Innovation Centre)

Quality reviewer: [name]

Deliverable nature:	Other
Dissemination level: (Confidentiality)	Public
Nature	
Document URL	
Work package	Wp2
Contractual delivery date:	30 Jun 2018
Actual delivery date:	
Version:	V1.0
Keywords:	Infrastructure; data analysis

Table of Contents

ABSTRACT

As discussed in D3.1, Data Pitch does not dictate the use of a particular infrastructure and 6 different cases are equally valid. Data can actually be hosted by the Data Providers, by the SME, in a commercial cloud or in one of the three different infrastructures directly controlled by Data Pitch.

In the first round of Data Pitch (2017-18) the infrastructure directly controlled by Data Pitch has been not directly requested by the different SMEs and Data Providers. In fact, among the possible choices the top pick was the commercial cloud. This is greatly related to the previous existence of commercial agreement either by the Data Provider or the SME for data sharing with well known worldwide platforms.

The support offered to the enablement of the data as a service paradigm has more been from a technical side on how to generally deal with data. To this extent, WP2 has created a technical survey and the results of the survey are presented and commented in this deliverable.

In the second round of Data Pitch (2018-19) more involvement is expected and a Data Provider has already requested the use of the Data Pitch data hosting infrastructure

EXECUTIVE SUMMARY

This short document is part of the D2.3 deliverable that is classified as “other”. It means that the report complements the installation of the infrastructures that are now ready to host data and experiments.

The document describes the different options available for hosting within Data Pitch and discusses how the data are ingested into the system.

Within the document the acceleration phase is also called the experimentation phase and the challenge winners are equivalently referred to as experimenters.

Also the document presents an analysis of the technical challenges faced by the SMEs in the first round of experimentation.

1 INTRODUCTION

This deliverable reports on the experience gained during the first cohort of acceleration and it provides an update on how the data are managed in Data Pitch from an infrastructure point of view. In fact, many options are possible regarding where to store and analyse the data: in the data provider infrastructure, in the experimenter (challenge winner) infrastructure, in the commercial cloud or in the Data Pitch infrastructure. The document briefly presents all of them. Data Pitch does not directly control all of the infrastructures. All the Data Pitch infrastructures offer a cloud-like approach where the end-user has control on the assigned resources. They also are able to support the choice and offer the best software stack for the execution of the experiment. There is not a standard proposal and the solution is personalised to the Data Provider and the challenge winner when the latter is identified.

2 UPDATE ON DATA AS A SERVICE (DAAS) INFRASTRUCTURE

One of the Data Pitch goals is to provide secure, transparent access to a wide range of virtualised data (Data-as-a-Service, or DaaS) and to facilitate a similarly wide range of data analytics. Following the Everything-as-a-Service (XaaS) paradigm, DaaS is based on the idea that data can be provided on-demand to the user regardless of geographic or organizational separation of provider and consumer. With the right DaaS solution, a company can combine data from disparate sources, including their own, and use the resulting knowledge to improve their business.

The infrastructure offer provided by Data Pitch has not undergone significant changes and it is described in D2.1.

In the second round a Data Provider has already requested Data Pitch to host its data and the deal is currently being finalised.

3 ANALYSIS OF DATA AS A SERVICE EXPERIENCE IN THE FIRST ROUND.

3.1 METHODOLOGY

A survey has been created and all the SMEs currently involved in the Data Pitch programme have been required to fill it in. The structure is available in the Annex to this deliverable. 14 out of 18 SMEs have provided their answers.

All the questions are related to technical aspects of the solution currently under implementation with few references to future expansion and to the relationship with the Data Providers. Questions and checkbox answers are purposely created to draw a high level picture with qualitative rather than quantitative information. This is also to preserve the sensitivity of the information shared by the SMEs.

Survey results are presented in an aggregated fashion and no filters have been applied. There is no cross check of the validity of what the SMEs state in the survey. Moreover, the business oriented environment nurtured by Data Pitch has possibly created an inclination of the SMEs towards the selling of their product rather than a rigorous assessment of the technical aspects of the solution.

This should not be considered a problem for the analysis that is meant to suggest the most important and promising area that technical challenges should consider to be really supportive for the SMEs and the Data Providers.

While we do not claim to draw general conclusions with statistical evidence basing on the narrow base of survey answers, we think that some interesting indications should be further considered when planning to operate in the business of data science. Results will be updated next year when the second round of SMEs will be participating in the acceleration period.

3.2 DATA HOSTING AND CONTROL

As explained in the previous section no SME or Data Provider has directly requested the data hosting offered by Data Pitch.

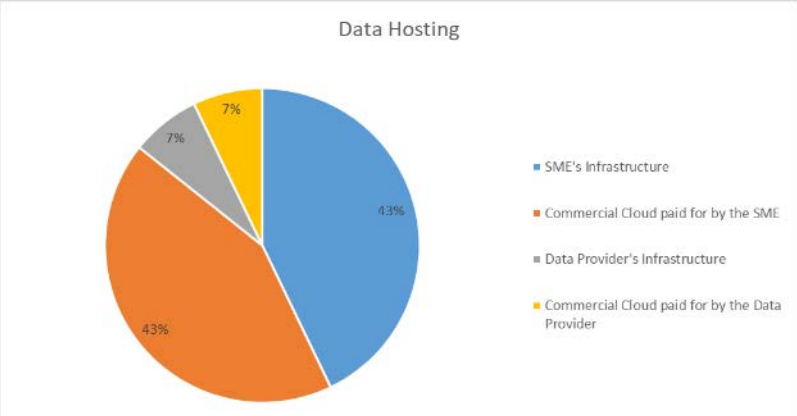


FIGURE 1: DATA HOSTING BREAKDOWN

As shown in the figure, during the experimentation phase the SMEs need to have access to the (a sample) of the data and this is mostly achieved using either the SME infrastructure of a commercial infrastructure paid by the SME.

This solution is likely meaningful for the 6 months of the acceleration period while it is less realistic that a Data Provider would accept to have its own live data with high business value hosted by a commercial partner (i.e. the SME after the acceleration period).

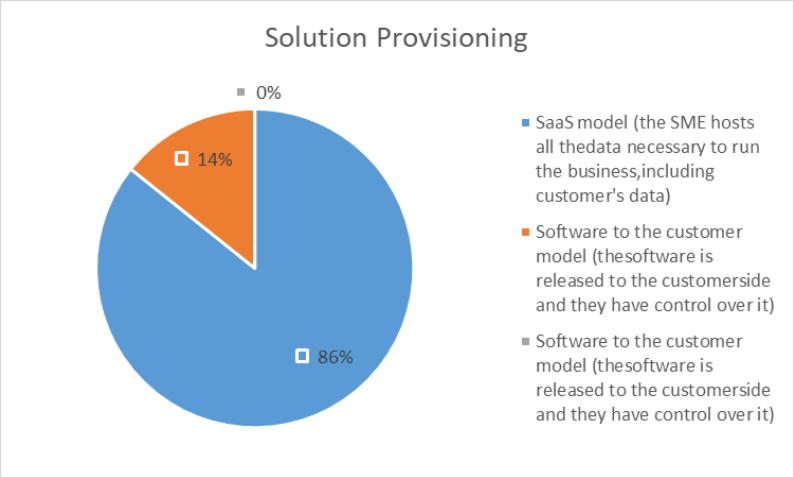


FIGURE 2: SERVICE PROVISIONING MODEL

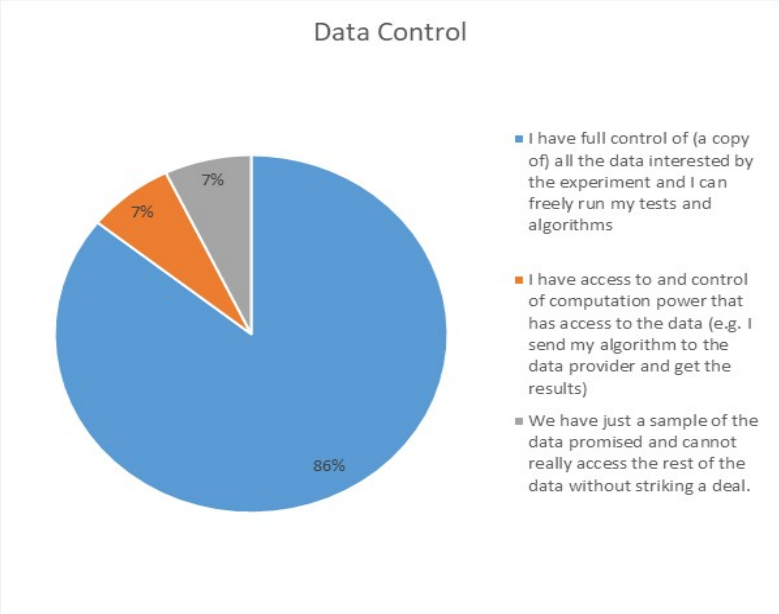


FIGURE 3: DATA CONTROL

The SMEs seem not really yet considering this problem and in fact the current vision for the service provisioning in the long term is shown in the figure below with the large majority trying to have a SaaS approach. The question in the survey was: *"What sentence below best describes how you will technically deliver your solution to your customers after the acceleration period?"*

Finally it is worth checking the type of control the SMEs have on the data (see Figure 3). In most of the case the Data Providers have agreed to give the dataset to the SME that can control it. This is the most straightforward model, but it is also the more risky for the Data Providers. Data Pitch has been available to investigate different models, but neither the SMEs nor the Data Providers were really interested to go beyond this model of control.

3.2.1 LESSON LEARNT

The solution identified for data hosting during the experimentation is not necessarily the same as will be going forward. It is important to increase the awareness in the SMEs of the business implication of the technical solution identified for the service delivery as this can be a very high barrier for the adoption of new and innovative tool for the extraction of value from data.

Making Data Providers aware that different models of data sharing can be explored could increase the providers willing to share their data. In fact the risk of giving the SME their data has been a deal breaking factors in some preliminary explorative discussions with potential Data Providers.

3.3 DATA CHARACHTERISTICS

This paragraph presents the general characteristics of the datasets. More than half of the SMEs are using 5 or more datasets (see Figure 4). This gives an indication of the complexity and variety of the data they deal with. Figure 4 shows the size of the closed (private) dataset that is the base of the solution. Here around 1/3 is rather large, while almost half of the datasets have an average size. The value of a dataset is not necessarily related to its sizes and often this private datasets are complemented with large public datasets.

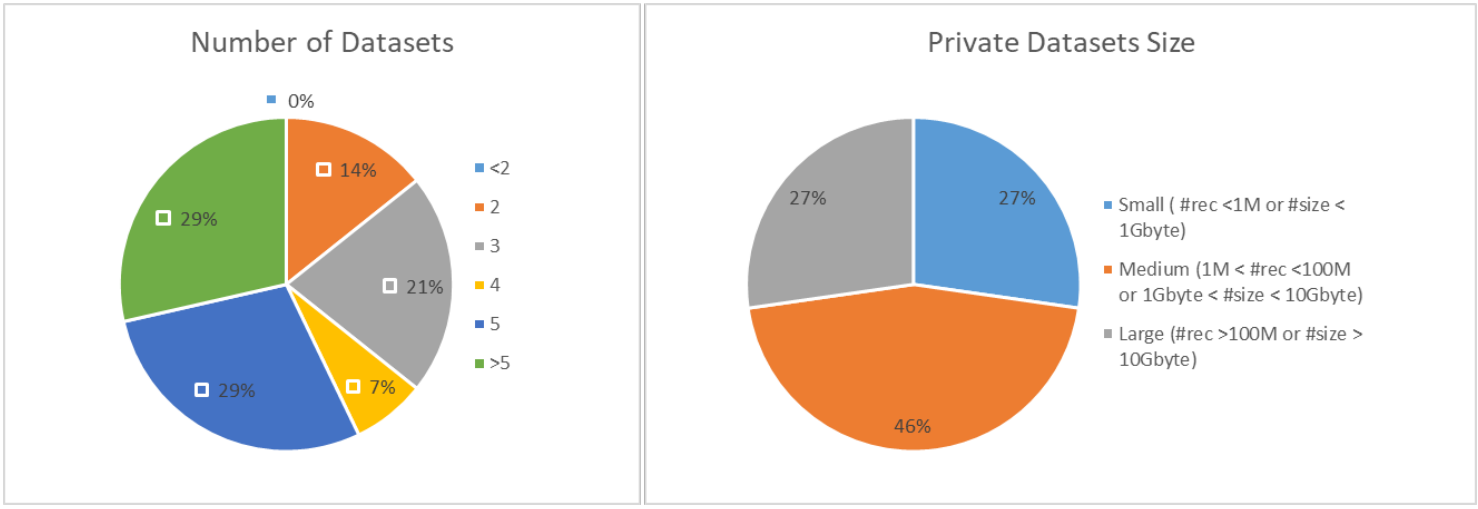


FIGURE 4: NUMBER OF DATASETS AND THEIR SIZE

The analysis of the data format is also interesting. Most of the data are in CSV (see Figure 5). This is an indication of an easy way for exporting data, but also on the need for having a relatively simple format that can be easily ingested and operated by many different systems.

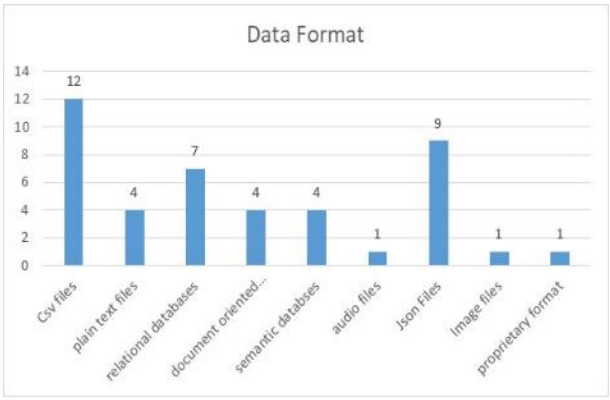


FIGURE 5: DATA FORMAT

Finally some useful information can be inferred by looking at the frequency of data update (see Figure 6). It is clear that the solution should be identified on stable data, but in few cases an update is important because data evolution is likely important in the solution.

When looking at the final solution the "static" approach almost not requiring an update (i.e. "historical data" and "occasional update" value) is still existing. This gives an indication that the proposed solution is likely one-off and therefore the SME needs to constantly identify new customers to be sustainable. Nevertheless the majority of the solutions are designed to deal with constantly evolving (at different paces) datasets.

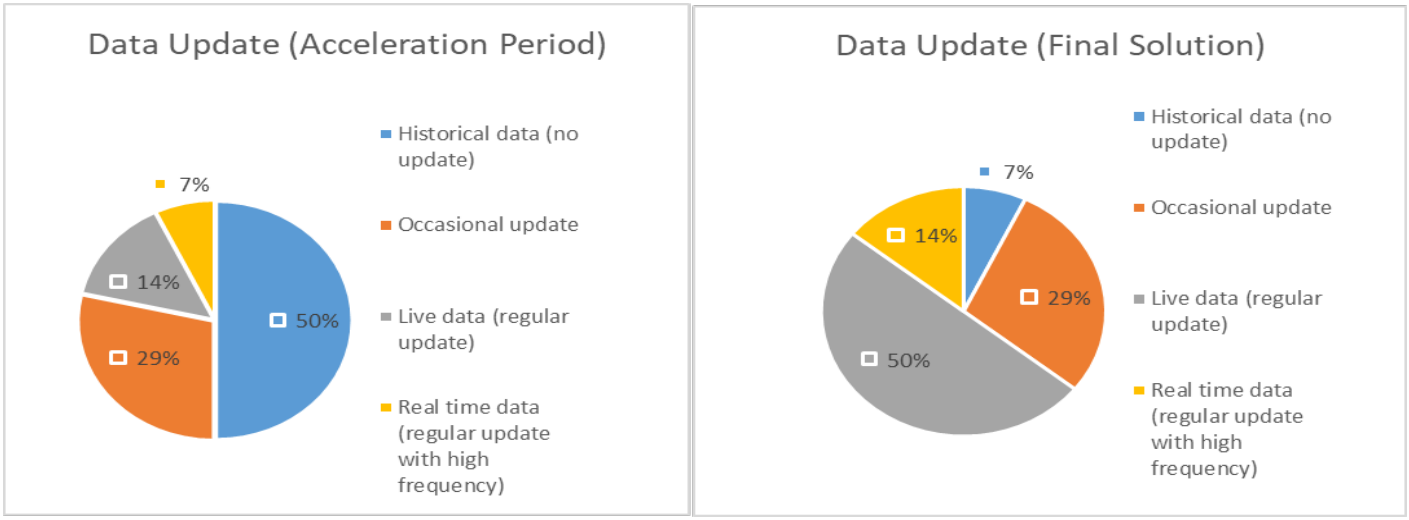


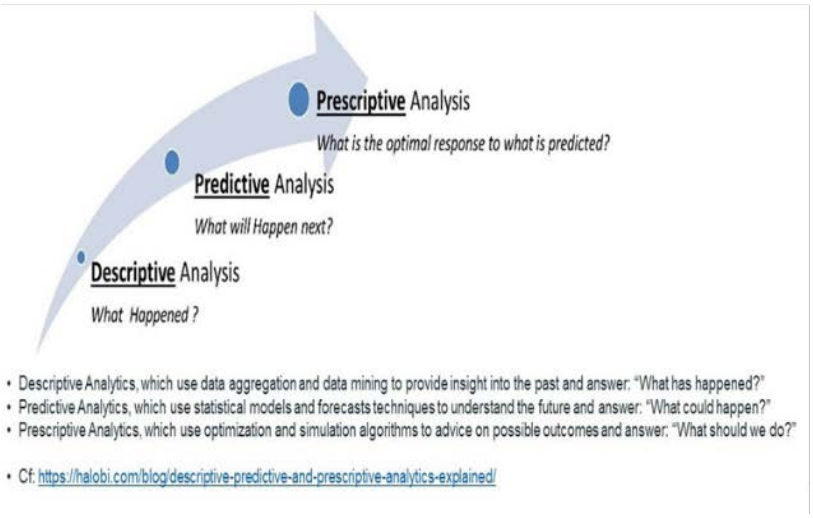
FIGURE 6: DATA UPDATE FREQUENCY

3.3.1 LESSON LEARNT

While big data pose specific challenges, it seems that the landscape of the business of extracting value from data is yet to be explored regardless of the data size. Lots of interesting and challenging problems are not yet completely solved and they have an important business value. Strengthening the instruments for dealing with data science is necessary and should be enabled at all level, not only for very large quantities of data.

3.4 TECHNICAL GOALS AND SOLUTION CLASSIFICATION

The business problems addressed in the Data Pitch challenges are quite widespread and similarly the technical solution identified are quite diverse. With reference to Figure 8, the survey asked the SMEs where their solution sits. The answers to this question describe a scenario where the initial step (i.e. descriptive analysis) is still not addressed and represents a relatively simple solution to real business problem. It needs to be said that the categorization chosen is not standard and subject to interpretation and this could have influenced the answers. In fact, prescriptive analysis is still quite technically complex and not the priority for the current Data Providers. The answers pointing at this category are likely more a long term vision than a factual positioning of the current solutions.



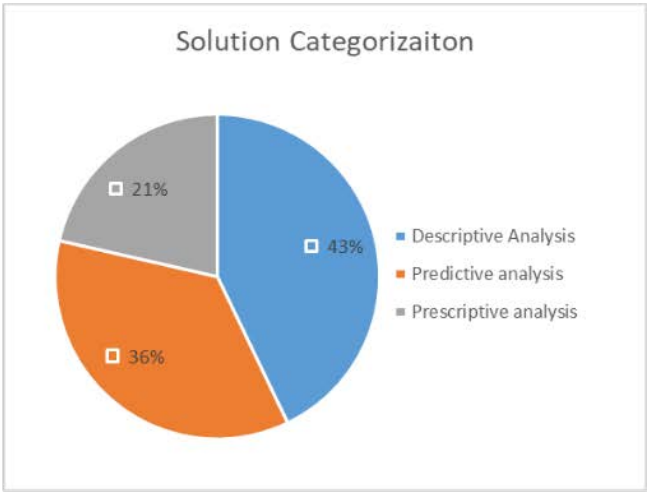


FIGURE 7: SOLUTION CATEGORIZATION

When moving to the actual technical solutions identified for addressing the problem (cf. Figure 9) the survey asked the SMEs to focus on a relatively small set of general possibilities and to pick only the most relevant aspect.

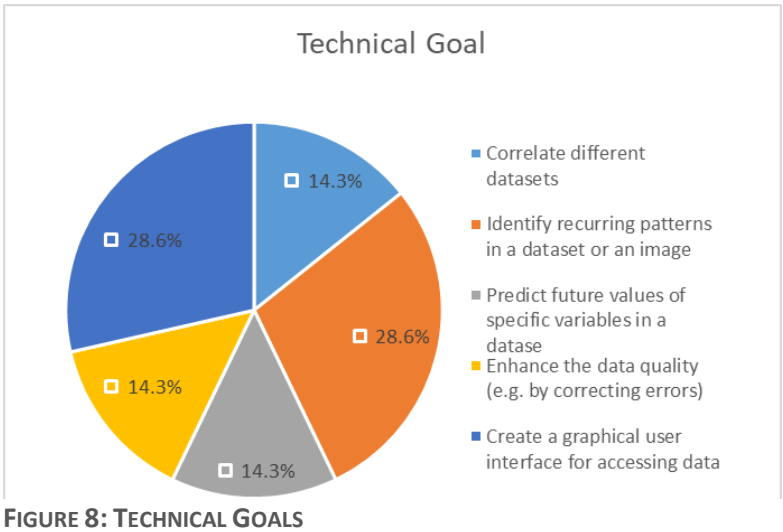


FIGURE 8: TECHNICAL GOALS

The results shows that the prediction of value is not necessarily the solution for actual business problems and that most of the solutions focus either on the identification of patterns or on the construction of visual tools for data access. The indication given by this information (and to some extent the slight inconsistency with Figure 8) indicates that seeking cutting-edge solutions are likely important, but also that a better alignment with solving actual business problems should be sought when providing incentives to the data science economy.

More in line with the expectations are Figure 10 and Figure 11. They show the technical solution adopted and a mix of standard and advanced techniques are currently being used.

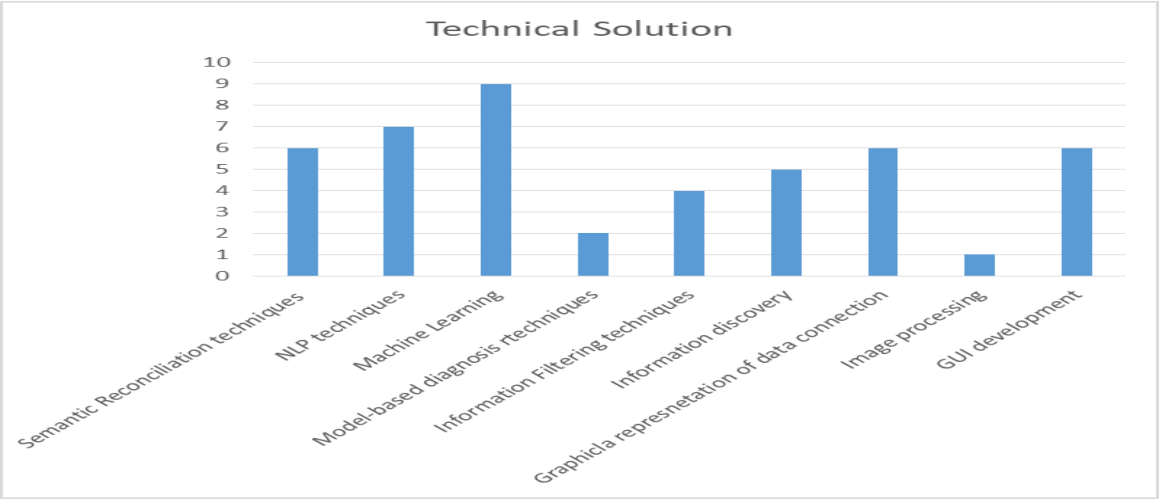


FIGURE 9: IDENTIFIED TECHNICAL SOLUTION

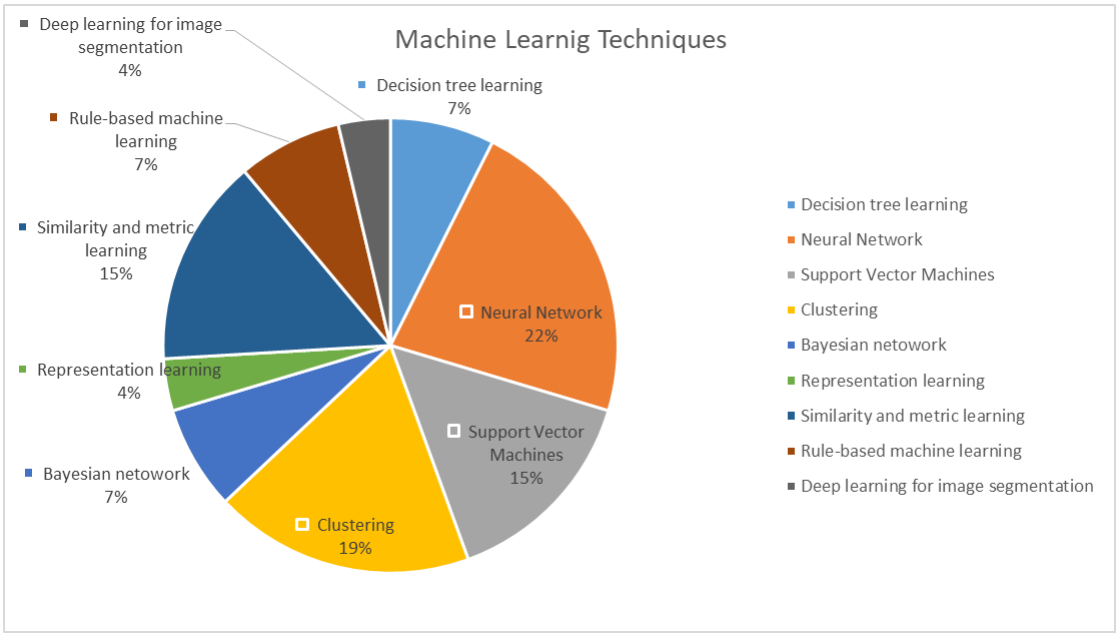


FIGURE 10: MACHINE LEARNING BASED SOLUTION BREAKDOWN

3.4.1 LESSON LEARNT

Existing business problems can often be solved with the smart application of relatively well-known techniques. It is the problem that drives the solution and not the other way around. It is likely possible to identify effective solution avoiding the complexity of cutting edge solutions. The general audience of business investors and operators are well aware of the problem they want to be solved and the level of innovation is only a second rank problem. On the other hand the business scenario is definitively evolving (possibly at a slower pace than expected) and the SMEs need to play a triple games: satisfying their customer now, be ready to evolve and to anticipate market needs. A deeper involvement of the SMEs within research-led academic problems might give them competitive advantage in the longer term over SMEs focusing on quotidian issues.

3.5 TECHNICAL CHALLENGES AND THEIR IMPACT ON THE SOLUTION

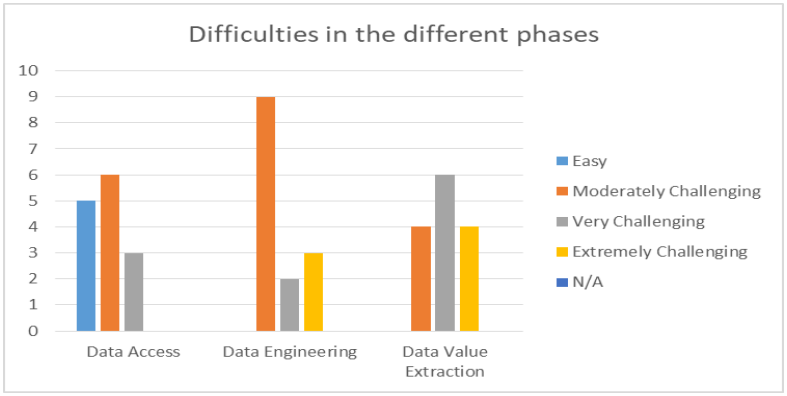


FIGURE 11: DIFFICULTIES ENCOUNTERED IN THE DIFFERENT PHASES

The surveys proposed to the SMEs a simplified breakdown of the value chain composed by "data access", "data engineering" and "data value extraction" (see Figure 10) . It asked the SMEs to identify how difficult and challenging each phase has been. The results are biased by the uneven distribution of the problems with some challenges requiring a more intensive and time consuming data engineering phase. The lack of easy mark in the data engineering and value extraction phases are proof that the challenges are not trivial and are an indication of the innovative aspects included in the proposed solutions.

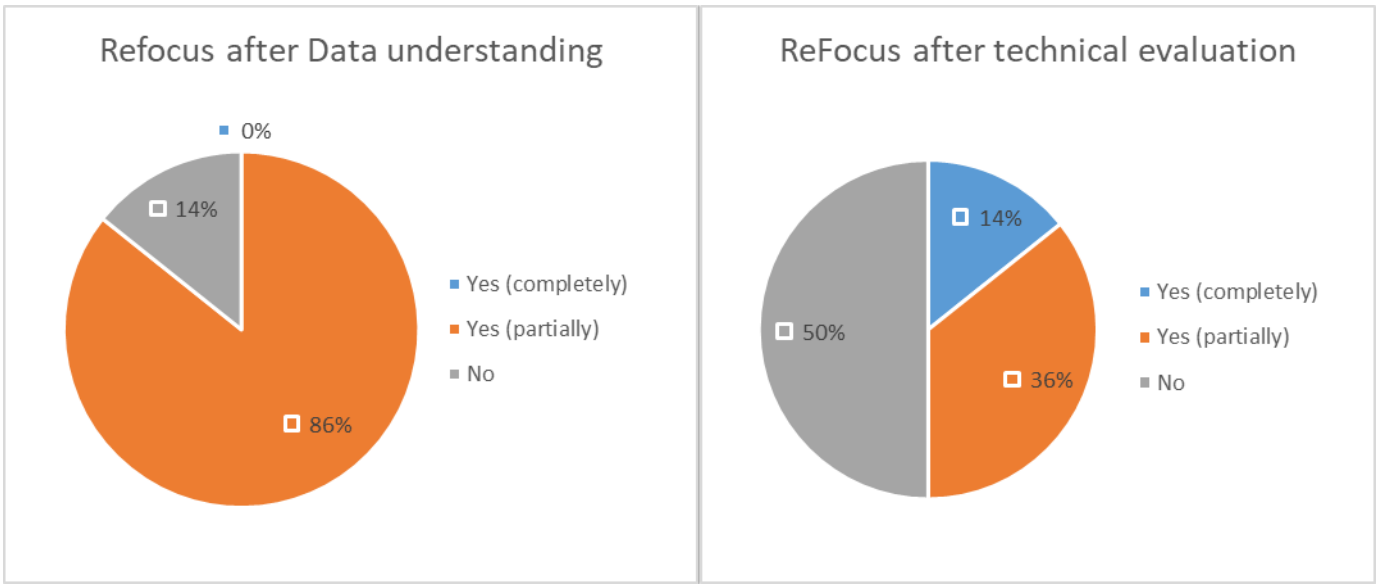


FIGURE 12: SOLUTION REFOCUSSING

Figure 12 shows that a gap existed in the perception of the problem by the Data Providers and the SMEs and that they had to refocus the original envisaged solution both as a results of a better understanding of the data and as a result of an accurate evaluation of the technical challenges and the related effort. This information shows that the process of identifying the right solution for a data related problem is a live and iterative process that requires all the actors to be involved. This need is backed up by the information shown in Figure 13. Over 60% of the SMEs have a constant interaction with the Data Providers to discuss and identify the best solutions and to solve problems related to the data understanding.

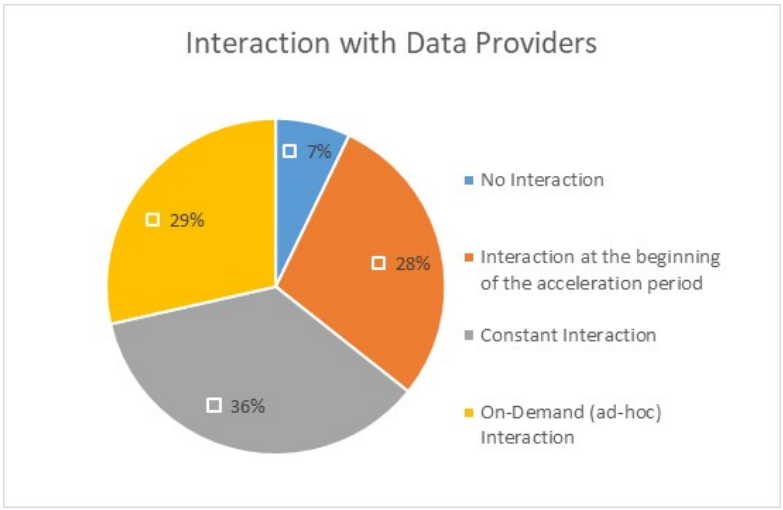


FIGURE 13: INTERACTION WITH DATA PROVIDERS

We have also asked the SMEs to produce free text around the "most difficult", "most satisfactory", "most

annoying" and "most enjoyable" technical challenge they have met. Table 1 shows an aggregated view of the answers provided. By analysing the asnwers, it is clear that low quality of data has been a common factor identified as an "annoying" issue. Data Pitch has offered to provide preliminary data cleaning services, but no SME requested the service. This is probably related to the fact that the necessary data cleaning was very specific and required a degree of domain knowledge, so the SME preferred to be in control of the process and the results.

TABLE 1: OVERVIEW OF THE COMMENTS ON THE TECHNICAL CHALLENGES FACED BY THE SMES

Most difficult	Most satisfactory	Most annoying	Most enjoyable
- Data understanding - Input preparation	- See the prototype working - Early validation of the approach	- Low Data quality	- Fine tuning the solution - Perform successful experiments under different conditions. - See the solution working in real condition

3.5.1 LESSON LEARNT

It is only when SME and Data Providers are in close dialogue with each other that many details are clarified and the right solution identified. Without one of the two partners involved the scenario is never completely clear and important details, but also opportunities, could be overlooked.

Another message to consider is the importance of the data engineering (including data quality improvement) phase. This phase is often overlooked and it deeply affect the effort needed and the quality of the results. Specific indications to the Data Providers should be provided so to manage their expectations.

3.6 SCALABILITY

A specific set of survey questions were addressing the scalability of the solutions (see Figure 15). The answers show a general awareness by the SMEs of the need to be ready to scale to maintain their business. In fact, the answers suggest an awareness of the technical complexity associated with the penetration in different verticals. This is a good sign that, by design, scalability is present in the SME plans. This is confirmed by the targeted economic verticals shown in Figure 16. In fact the average number of targeted verticals for each SME is just above 3.

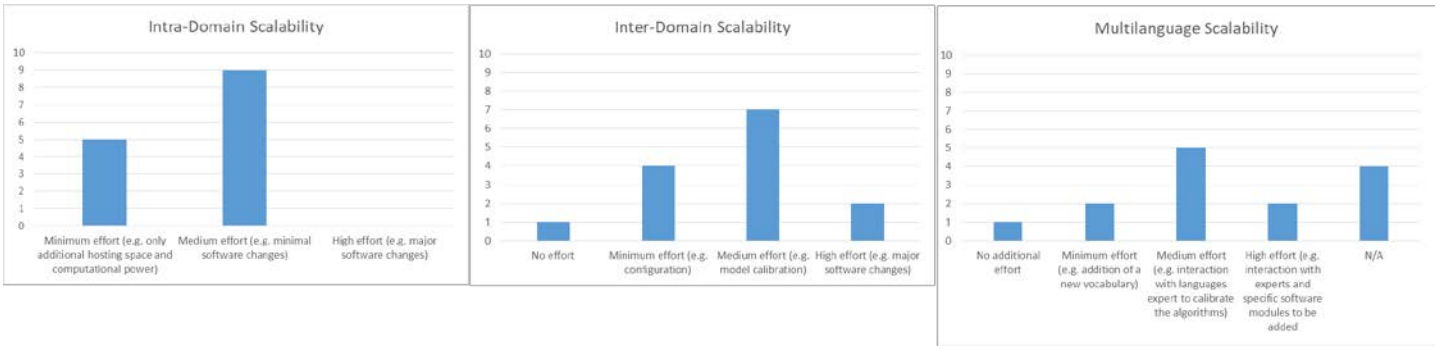


FIGURE 14: SCALABILITY ANALYSIS

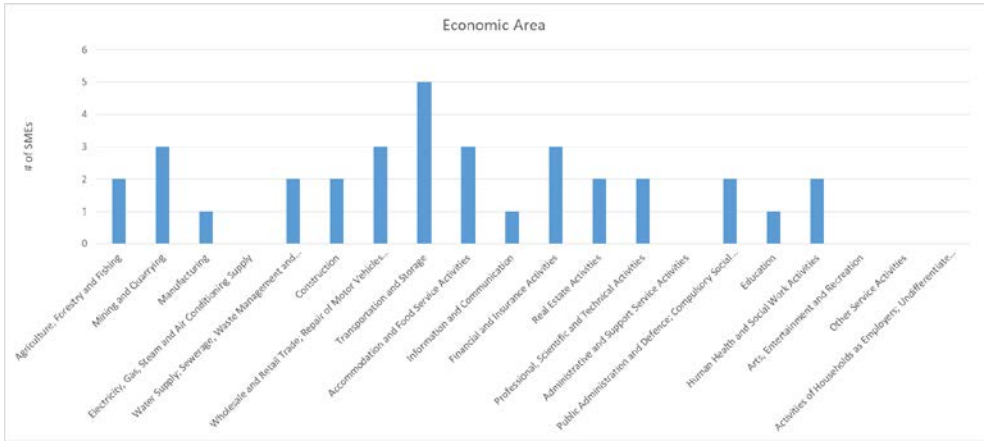


FIGURE 15: TARGETED VERTICALS

3.6.1 LESSON LEARNT

It is possible to generalise solutions across different domains to support scalability. This possibility needs to be taken into account since the initial design of the solution as the portability is never completely free. The “by design” paradigm is nowadays quite common and it should be further pushed to increase the solution long term sustainability.

4 CONCLUSION

This document reprised the different solutions available in Data Pitch to host data and experiments presented in D2.1. Further to that, the report provides an initial analysis about the technical challenges met by the SMEs during the first round of the acceleration period. This information is collected through a survey that is analysed.

The analysis is not based on a big enough critical mass to claim statistical significance, but, still, offers a few indications of areas that likely need a tuning to improve the effectiveness and long-term sustainability of measures aiming at stimulating the data science economy. This indications are reported in the different sections as lesson learnt and will be considered when supporting the SMEs participating to the second round of acceleration.

APPENDIX – SURVEY STRUCTURE

* Required

1. SME Name *
2. Data Provider Name(s) *

Please include the names of all the Data Providers sharing a closed dataset with you.

3. Consent to share the survey answer with the Data Pitch advisor *

Mark only one oval.

- o yes
- o no

Technical characteristics of the solution

Data related questions

4. Where is the data hosted *

Mark only one oval.

- o Data Provider's Infrastructure
- o Commercial Cloud paid for by the Data Provider SME's Infrastructure
- o Commercial Cloud paid for by the SME Other:

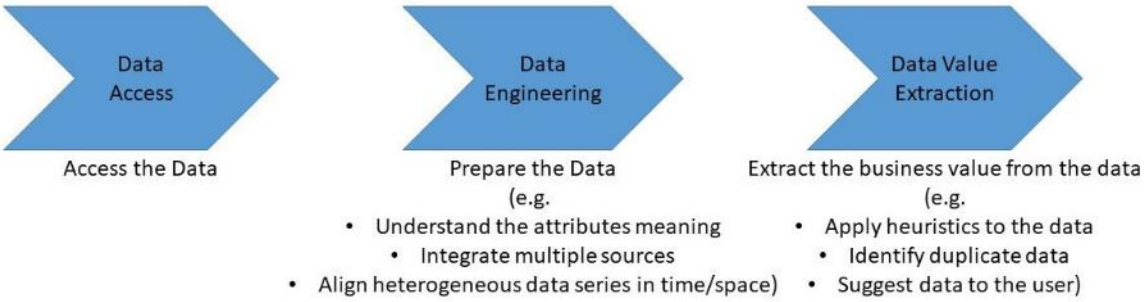
5. How do you work on the data? *

Mark only one oval.

- o I have full control of (a copy of) all the data interested by the experiment and I can freely run my tests and algorithms
- o I have access to and control of computation power that has access to the data (e.g. I send my algorithm to the Data Provider and get the results)
- o Other:

6. How many datasets does your solution use? *
7. What is the approximate size of the shared dataset you are working with as part of your solution (please specify either #entries or #physical dimension for each dataset referred to in the previous answer) *
8. What is the approximate total size of data entailed by your solution? *
9. Please indicate the format of all the data used in your solution *
10. Tell us the level of interaction on technical topics you have with your Data Providers (closed dataset) *
    - o No interaction
    - o Interaction at the beginning of the acceleration period
    - o Constant interaction
    - o On-Demand (ad-hoc) interaction
11. What is the frequency update of data during the acceleration period? *
    - o Historical data (no update)
    - o Occasional update
    - o Live data (regular update)
    - o Real time data (regular update with high frequency)
12. What is the frequency update of data for the final solution? *
    - o Historical data (no update)
    - o Occasional update
    - o Live data (regular update)
    - o Real time data (regular update with high frequency)

Questions related to the creation of your solution



13. With reference to the figure above, please rate the different phases of the value extraction chain in terms of how challenging they have been? *
    - o Easy
    - o Moderately challenging
    - o Very challenging
    - o Extremely challenging
    - o N/A (phase not present or not yet reached)
14. Which of the following best describes the technical objective of your solution? *

The sentences below purposefully represent broad categories, and it is likely that very different solutions might belong to the same category. E.g. a 'recommender system' and an 'entity reconciliation system' would fall in the "correlate different datasets" category. Conversely, if a chatbot type of solution stresses the identification of the (voice) patterns, it would fall into the "identify recurring patterns" while if it stresses the suggestion part it would be more a recommendation system, so falling into the "correlate different datasets". If more than one of these sentences is applicable to your solution, please select only the most relevant.

- o Correlate different datasets
- o Identify recurring patterns in a dataset or an image
- o Predict future values of specific variables in a dataset
- o Enhance the data quality (e.g. by correcting errors)
- o Create a graphical user interface for accessing data

15. Which of the following technical characteristics describe your solution? *
    - o Semantic Reconciliation techniques (e.g. identification of same concepts across datasets) Natural Language Processing techniques (e.g. keyword identification,
    - o subject/predicate/object identification)) Machine Learning
    - o Model-based diagnosis techniques
    - o Information filtering techniques (e.g. to identify the best answer across a large set of answers)
    - o Information discovery according to the end-user context (e.g. context-based recommender systems)
    - o Graphical representation of data connection Image processing
    - o Voice processing
    - o Graphical user interface development
16. If your approach uses Machine Learning techniques, please select below what apply. *
    - o Machine Learning - Decision tree learning
    - o Machine Leaning - Neural Networks techniques
    - o Machine Learning - Association rule learning
    - o Machine Learning - Inductive logic programming
    - o Machine Learning - Support vector machines
    - o Machine Learning - Clustering
    - o Machine Learning - Bayesian networks
    - o Machine Learning - Reinforcement learning
    - o Machine Learning - Representation learning
    - o Machine Learning - Similarity and metric learning
    - o Machine Learning - Sparse dictionary learning
    - o Machine Learning - Bio-inspired approaches (e.g. genetic algorithms)
    - o Machine Learning - Rule-based machine learning

- o N/A
- o Other:

17. Please provide a description of the technical challenges you have faced *

Please list only one challenge for each of the following four categories: - Most difficult; Most satisfactory; Most annoying; Most enjoyable

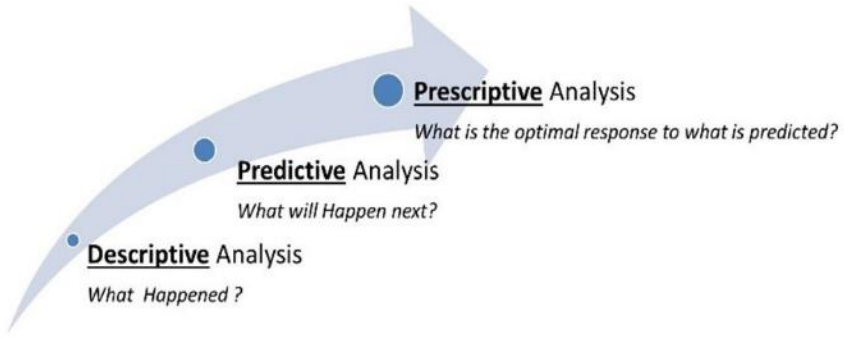
18. Have you had to change/re-focus your original technical idea because of a better understanding of the data as result of your initial data analysis?

- o Yes (completely)
- o Yes (partially)
- o No
- o Other:

19. Has a change/re-focus in your solution become necessary as a result of a technical challenge that is too time-consuming or difficult?

- o Yes (completely)
- o Yes (partially)
- o No
- o Other:

Questions related to the classification, scalability and portability of your solution



- Descriptive Analytics, which use data aggregation and data mining to provide insight into the past and answer: "What has happened?"
- Predictive Analytics, which use statistical models and forecasts techniques to understand the future and answer: "What could happen?"
- Prescriptive Analytics, which use optimization and simulation algorithms to advice on possible outcomes and answer: "What should we do?"

- Cf: https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/

20. With reference to the figure above, how would you categorise your solution? *

- o Descriptive Analysis
- o Predictive Analysis
- o Prescriptive Analysis
- o N/A

21. How much effort is necessary from a technical perspective to scale up your solution 10x within the same domain (same business, bigger volume) *

- o No effort
- o Minimum effort (e.g. only additional hosting space and computational power)
- o Medium effort (e.g. minimal software changes)
- o High effort (e.g. major software changes)

22. List the verticals (business domains) your solution is targeting (order the list from the most to least relevant) *

If possible refer to the NACE level1 codes listed here: https://en.wikipedia.org/wiki/Statistical_Classification_of_Economic_Activities_in_the_European_Community

23. How much effort is necessary from a technical perspective to port your solution to a different domain (inter-domains scalability) *

- o No effort
- o Minimum effort (e.g. configuration)

- o Medium effort (e.g. model calibration)
- o High effort (e.g. major software changes)

24. If your solution deals with natural languages, how much effort is necessary to add a new language? *
    - o No effort
    - o Minimum effort (e.g. configuration)
    - o Medium effort (e.g. model calibration)
    - o High effort (e.g. major software changes)
25. What sentence below best describes how you will technically deliver your solution to your customers after the acceleration period? (Please provide a brief explanation if you select 'other'.) *
    - o SaaS model (the SME hosts all the data necessary to run the business, including customer's data)
    - o Software to the customer model (the software is released to the customer side and they have control over it)
    - o Ad-hoc model (e.g. in a consultancy relationship where the interaction is defined case by case)
    - o Other:
26. Are you considering the possibility of releasing (part of) your solution as Open Source and exploring businesses models related to it? *
    - o Yes
    - o No
27. Please share with us any information about the technical development of your solution that you think might be beneficial for the programme, in particular for the next acceleration period.