Data Pitch

H2020-ICT-2016-1

Project number: 732506

D2.5 Usage analysis, lessons learned, and recommendations

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

Quality reviewer: [Ryan Goodman (ODI)]

Deliverable nature:	Other
Dissemination level: (Confidentiality)	Public
Nature	
Document URL	
Work package	Wp2
Contractual delivery date:	31 December 2019
Actual delivery date:	31 December 2019
Version:	V0.1
Keywords:	Interactions, data analysis, impact assessment

Table of Contents

ABSTRACT

This report summarises the findings of the second round of Data Pitch (2018-19) about the use of technology within the cohort of accelerated SMEs. The evidence was collected through a survey[1]. 37 SMEs answered the survey during the period September-October 2019:

The results are presented according to the following categories:

a) Characteristics of data used in the solution
b) Creation of solution
c) Post-acceleration period

While the numbers of answers are not enough to draw conclusions that are significant from a statistical point of view, the document presents several interesting insights on how the technology has been used, showing the main technological barrier the SMEs met and how they have been overcome.

[1]. The survey also served the purpose of D7.2.

EXECUTIVE SUMMARY

This report presents the findings and insights of the work performed by the second cohort of SMEs accelerated in Data Pitch.

The report provides a detailed analysis of the questionnaire responses received from the second cohort of companies. This questionnaire was sent to all 29 companies and contained 32 questions in total. The same questions served the Deliverable D7.2. D2.5 focusses on the technological analysis, while D7.2 is about understanding the business implications and the overall impact of the acceleration program.

The second Data Pitch cohort has addressed a variety of problems using many different technologies. The solutions cover, as expected, a wide range of maturity levels and further technical development is required in several cases.

The technical support offered by Data Pitch has been appreciated both in the phase of the work plan definition and in the milestone verification. The SMEs have not required infrastructural support, preferring to use their own existing solution (or the one offered by the data providers). This confirms the findings that emerged in the analysis of the first cohort. The data storage infrastructure and the computational power is nowadays considered a commodity and SMEs prefer to pay for that, rather than investing time and effort for using a new one, even though this is offered without the cost of renting one.

1 INTRODUCTION

One of the Data Pitch objectives is to provide a Data-as-a-Service (DaaS) solution, e.g. secure, transparent access to a wide range of virtualised data and to facilitate the use of Machine Learning (ML) algorithms and advanced data analytics. DaaS is based on the concept that data can be provided on-demand to the user regardless of the geographic or organisational separation between the provider and consumer. The DaaS paradigm enables companies to combine data from different sources, including their own, and use the results for improving their business.

The analysis performed in this deliverable follows the structure of the analysis presented in Deliverable D2.3 this time 37[2] SMEs participated in the analysis.

1.1 SECTIONS

The main sections of this report are organised around three themes:

a) Characteristics of data used in the solution

b) Creation of solution

c) Post-acceleration

The report presents graphs only for a selection of questions considered most relevant for the analysis.

[2] The survey has been resubmitted to SMEs not answering to the one for the first cohort. Of this set, 8 SMEs provided an answer bringing the total to 37

2 METHODOLOGY

The survey was conducted by London Economics[3] who were commissioned by the Data Pitch consortium to carry out an independent assessment of the Data Pitch programme. 37 SMEs participated in the survey by filling in the questionnaire.

The questions are related to the characteristics of the data and solution, interactions between the users and data providers, the benefits of Data Pitch and partner exploitation plans.

Most of the questions contain checkboxes and ratings that allow drawing a qualitative impact assessment for the second round of experiments. Survey results are presented in an aggregated fashion, and no filters were used. The information provided by the respondents was not cross-checked, and it was assumed to be truthful.

The

term "paired data provider" identifies the data provider either identified by Data Pitch or the main one self-sourced by the SME when applying for the programme.

The presented results drawn from a sample of 37 respondents has allowed the partners to identify some trends that are relevant for the engagement of SMEs in a future Data Pitch style project.

2.1 DATA QUALITY

Simple basic data quality checks have been manually applied, and the general quality of the input was low. Where possible and appropriate, corrections have been applied (e.g. evident typo or clear inconsistencies).

3 INFORMATION ABOUT THE DATA USED IN THE SOLUTION

3.1 INTERACTION WITH DATA PROVIDERS

This information was broken down into two questions. #17 was about the interaction with the paired data providers (either self-identified or proposed by Data Pitch), #2 was about the interaction with other data providers included in the solutions.

How frequently (closely) we had to interact with the data provider for the development of the solution. (1 signifies 'No interaction' and 5 signifies 'Very close interaction')

[3] https://londoneconomics.co.uk

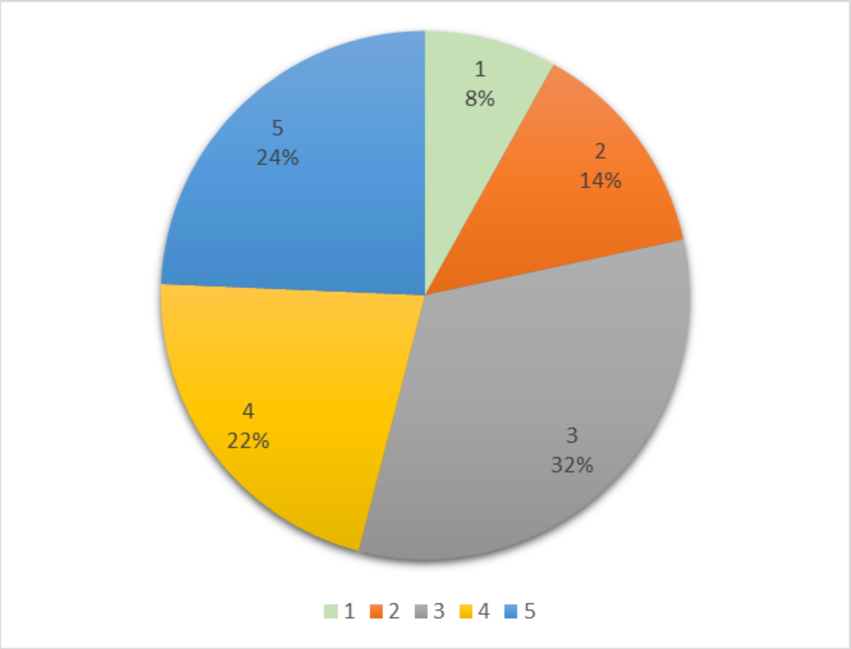

FIGURE 1: INTERACTION WITH PAIRED DATA PROVIDER DURING THE SOLUTION

How closely did you interact with Data Providers other than your partnered Data Provider? (1 signifies 'No interaction' and 5 signifies 'Very close interaction')

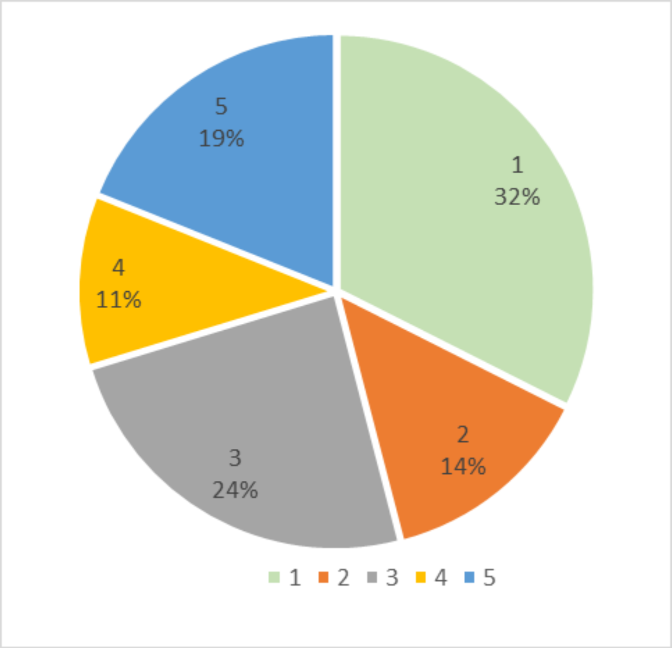

FIGURE 2: INTENSITY OF INTERACTION WITH ALL THE INVOLVED DATA PROVIDERS

The data shows that ~80% of the cohort (sum of classes 3,4,5) have had a good interaction with the data providers, while ~20% experienced problems. Despite being present in a limited number of cases, access to data remains the main barrier to develop innovative solutions. In future, better practices need to be put in place to reduce this value further. When coming to other data providers (see Figure 2) the level of interaction has dropped because this group of data providers were not the main benefit of the solution and this loosely coupling (if any at all) might have played a role in the level of interaction with the SME. Moreover, the question did not discriminate among closed and open data. Using open data usually does not require an interaction with the data owner beyond the description of the data itself for the benefit of the general public.

3.2 NUMBER OF DATASETS USED IN THE SOLUTION

A question (#3) tried to identify how many datasets are used in the solution.

How many datasets do you use in your solution? Please provide a total number of open, closed and self- generated datasets. (Dataset refers to sets of data that share the same features/characteristics and which your business either receives from a data provider or collects itself).

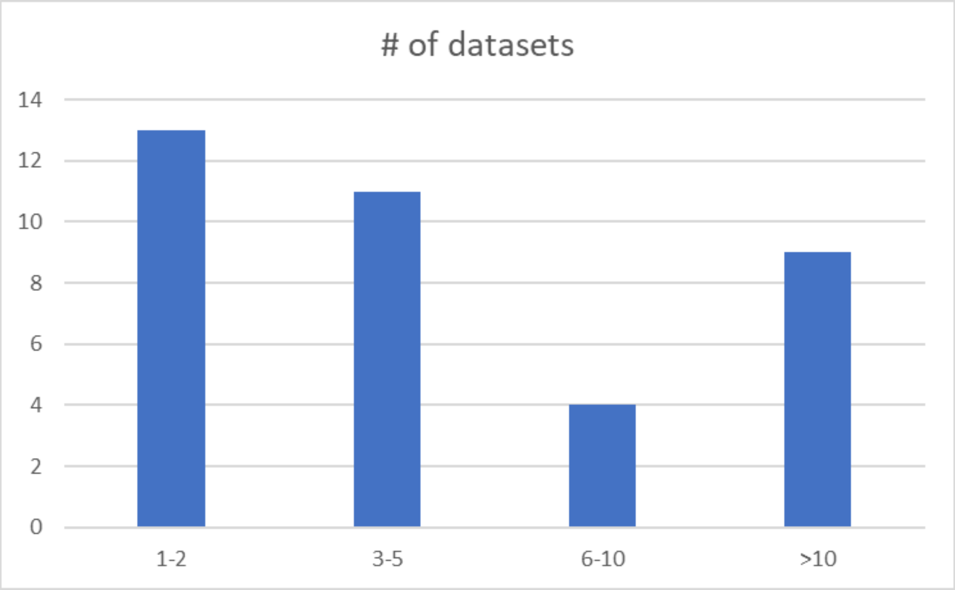

FIGURE 3: DATASETS USED IN THE SOLUTION

According to the responses to this question, the majority of solutions use less than 5 datasets (only two use just 1 dataset). The result shows that the number of datasets is relatively relevant for creating an innovative solution. Most of the cases have less than 5 datasets. On the other hand, by expanding the number of sources, solutions are likely more based on combining information. The breaking of data silos, constantly supported by the European Commission, is something that needs to be encouraged and supported as much as possible.

3.3 DATA CHARACTERISTICS

A question (#6) tried to group the solutions by the most important features present in the data.

Which characteristics of the data used in Data Pitch are the most important for your solution?

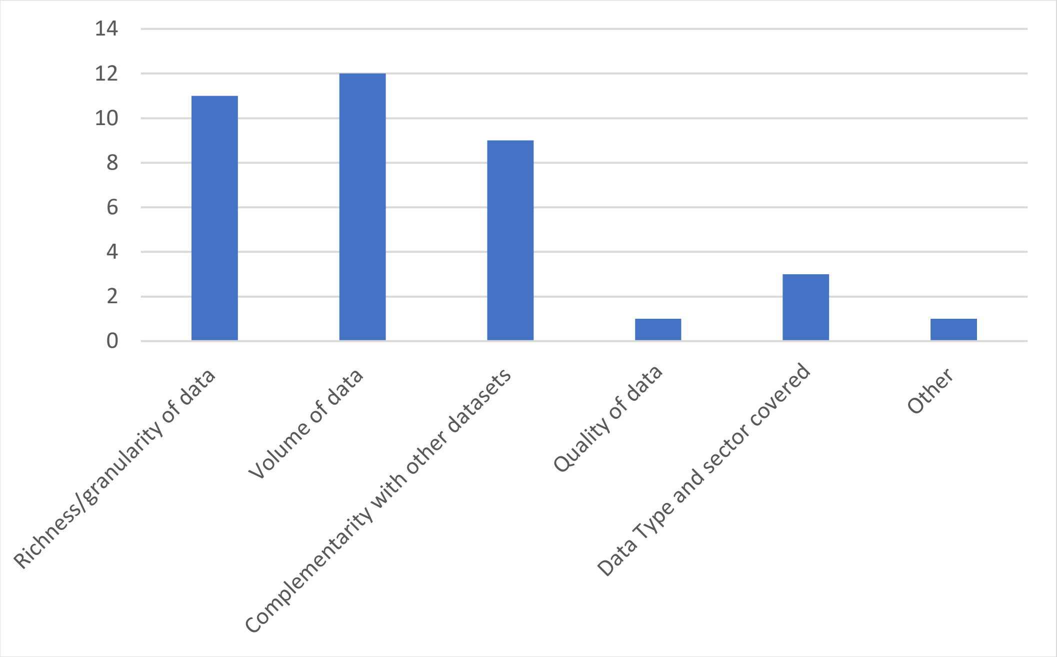

FIGURE 4: THE MOST IMPORTANT DATA CHARACTERISTICS

The survey has identified three main data characteristics were relevant in the context of the Data Pitch experiments, such as: a) richness/granularity of data, b) volume of data and c) complementarity with other datasets. These three characteristics are almost equally important (Figure 4) showing that not a single dimension of the big data landscape is more important than others and everything is driven by the applications (business) built on top of them.

3.4 SIZE OF OBTAINED DATASET

Few questions addressed the sizes in terms of records and space occupation of the datasets. These aspects were considered both regarding the paired datasets and within the complete solution (i.e. all the used datasets).

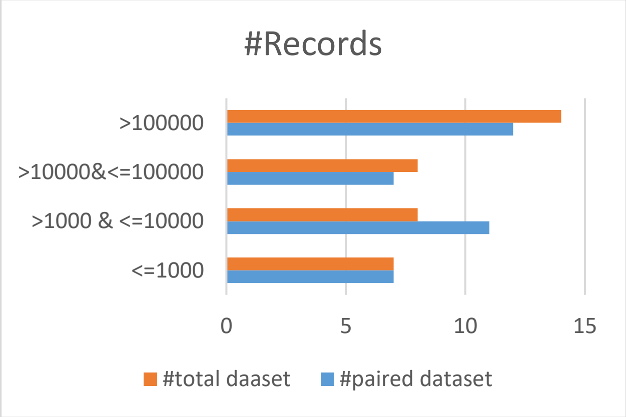

FIGURE 5: #RECORDS

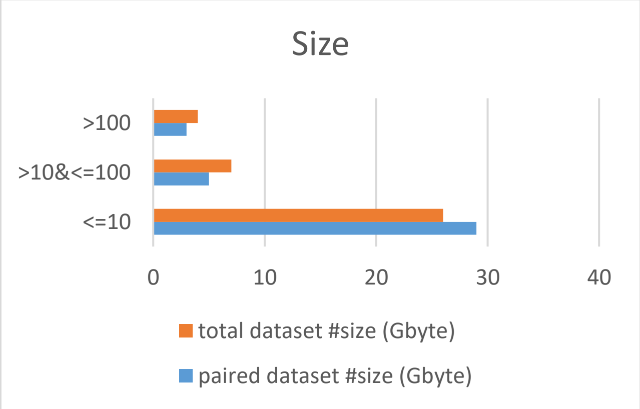

FIGURE 6: DATASET SIZE

Adding datasets have obviously produced an increment of size (both #records and space) in the solutions. However, as shown in Figure 7 (the figures are the same for both #record and size), most of the solutions (62%) have used only their paired dataset (100% section of the pie). The global picture this figures are balanced by the 30% of cases where the increment in size has been greater than 2/3 (66% section of the pie).

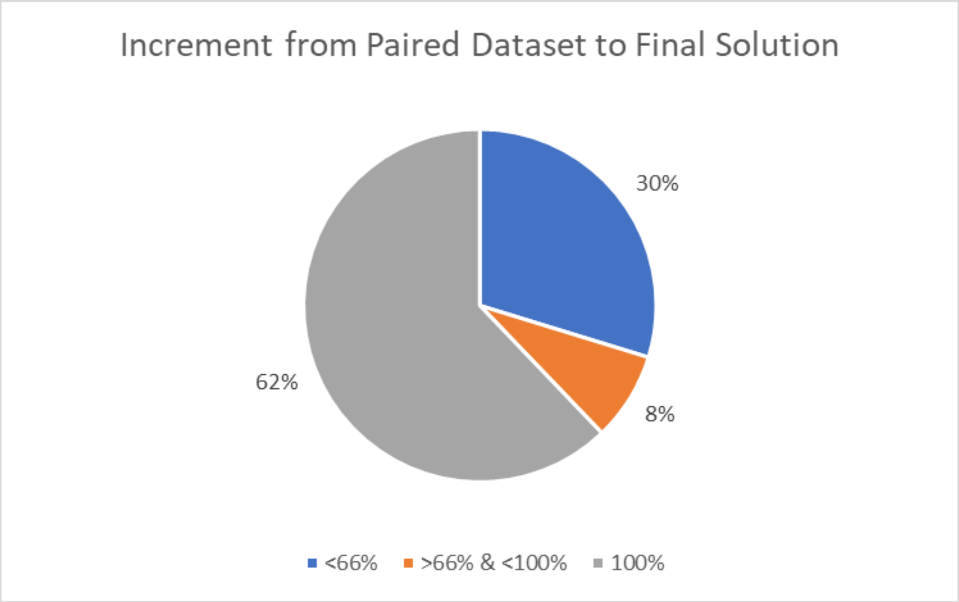

FIGURE 7: PERCENTAGE OF PAIRED DATA SIZE VS TOTAL DATA SIZE

3.5 TYPE OF DATA USED IN THE SOLUTION

Question #11 aimed at classifying the data according to the types.

Please indicate the main type of data used in your solution, select all that apply. a) Numeric, b) Text, c) Video, d) Audio, e) Geospatial, g) Image/graphics, h) Other.

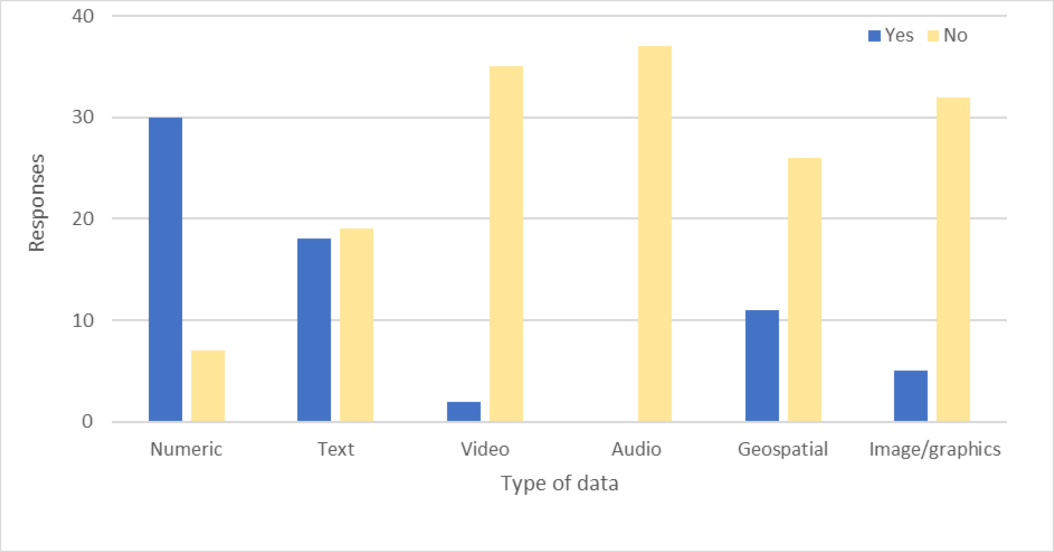

FIGURE 8: TYPE OF INPUT DATA USED IN THE SOLUTION

The results show a clear dominance of numerical data, about 81% (Figure 8). This is partly related to the type of challenges addressed in this round of acceleration. On the other hand, the numbers of answers do not allow to draw any statistically valid conclusion. In fact, the analysis of the first cohort in DataPitch and the landscape of data innovation as known to the author shows a good number of Situations in contrast with these results, thus confirming they are inconclusive.

3.6 STORAGE FORMAT OF DATA

Question #12 aimed at clustering the solutions according to the used storage support.

How is this data stored? a) Semantic databases (i.e. RDF triples), b) Document oriented databases, c) Relational databases, d) Files

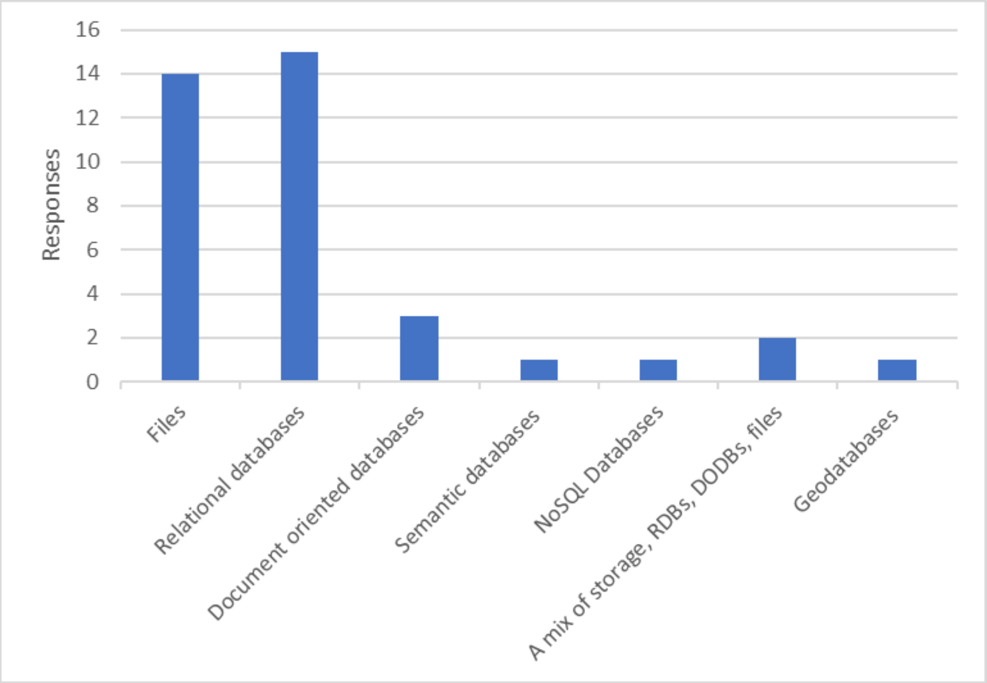

FIGURE 9: DATA STORAGE FORMAT

The results indicate that files and databases are the most relevant forms of data storage (Figure 9). The answers hint to a wide choice of general-purpose solutions that can handle at the same type a variety of data types. What is left unclear is, considering above all scalability and efficiency, how much fit for purpose are general-purpose solutions vs dedicated solutions able to focus on specific types of data.

3.7 LOCATION OF DATA STORAGE

One of the support offered by Data Pitch was computational capability. No SME have required access to this type of support, and question #13 investigated the type of solution adopted.

Where is the data stored? a) Data Provider's infrastructure, b) Commercial cloud paid for by the data provider, c) Your own infrastructure, d) Cloud paid for by your business

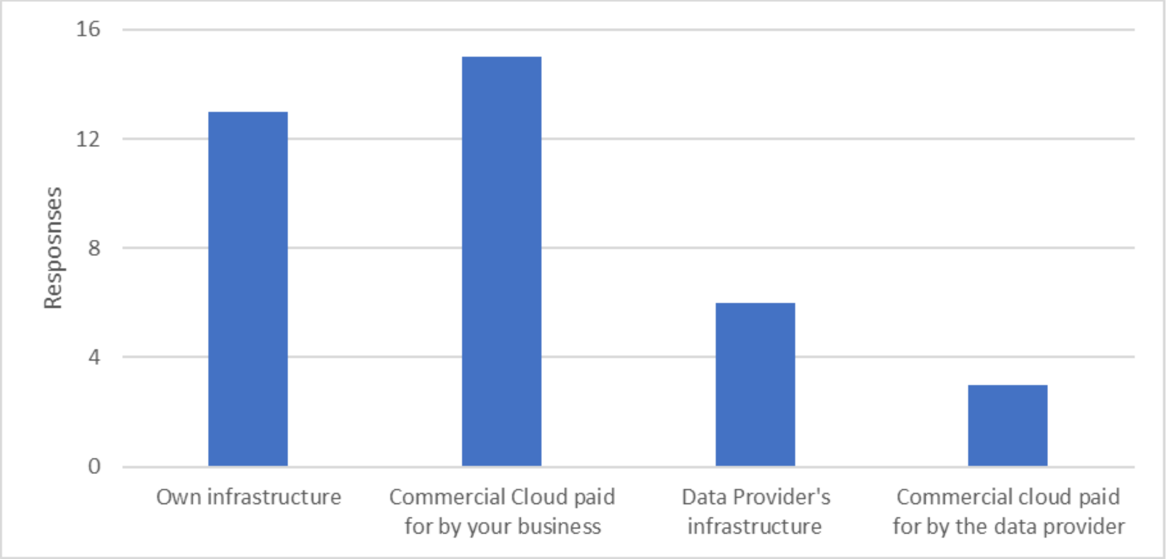

FIGURE 10: LOCATION OF DATA STORAGE

The results show that data storage is widely considered a commodity. A speculative analysis for justifying the missing opportunities of using infrastructure support is that setting up (or even just configure) a new environment is a cost that the SMEs are reluctant to pay, preferring to continue with their established and well-known solution.

3.8 DATA CONTROL

Question #14 was related to understanding how much control on the data the SME had. This is relevant to understand how innovation can be supported by building a trusted relationship among data providers and consumers that preserves the business sensitivity associated with data.

How much control do you have over the data when building your solution?

a) Full control (e.g. full copy of data freely available to me)
b) Partial control (e.g. data called through API when required)
c) No control (e.g. I send my algorithms to the Data provider and get the results)

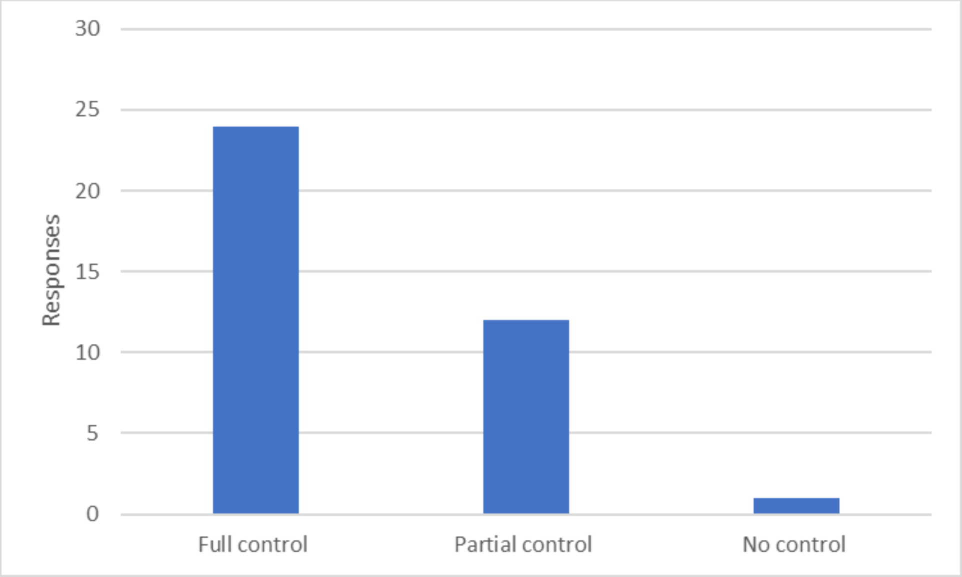

FIGURE 11: TYPES OF DATA CONTROL

The issue of data control can be a tricky one, especially if we consider shared ownership of data between the data provider and the user. The most desirable outcome for the users is to have full control of data, but partial control is also acceptable to provide that the conditions are advantageous for both parties (Figure 11).

3.9 DATA UPDATE PERIODICITY

One of the characteristic “V”s of Big Data sets is velocity. Question #15 classified the solutions according to their update frequency.

What is the update frequency (periodicity) of the data used in your solution?

a) Static data (not updated)
b) Occasional data (irregularly updated)
c) Live data (regularly updated)
d) Real-time data (regularly updated with high frequency)

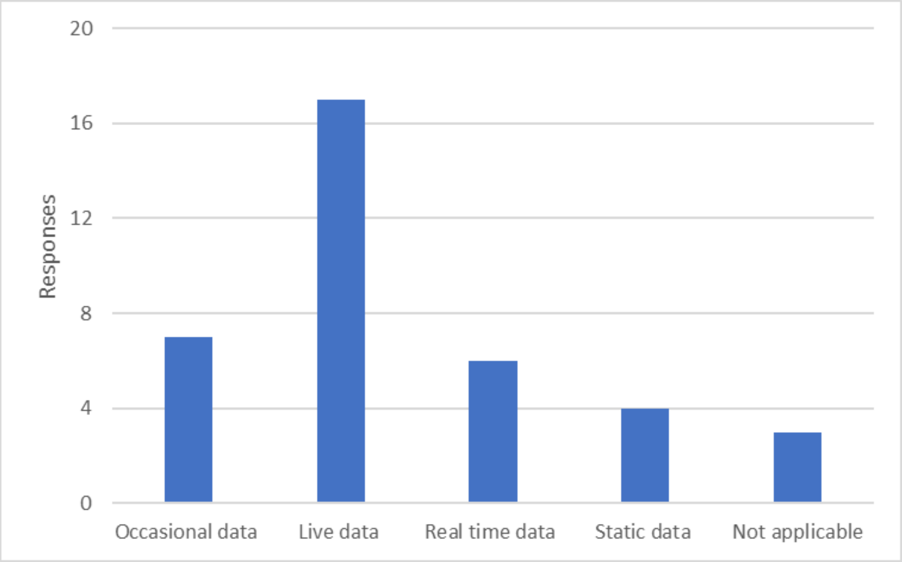

FIGURE 12: FREQUENCY OF DATA UPDATE IN THE SOLUTION

The frequency of data update can be linked to the intrinsic characteristic of the solution. For example, in the case of static data or occasional data the update frequency is low, indicating that the proposed solution is likely a one-off and therefore the SME needs to identify new customers to be sustainable constantly. However, the majority of SMEs surveyed work with live data and few of them with real-time data which generally require ad-hoc data analysis architecture.

4 CREATION OF SOLUTION

This section describes the outcome of the experiments and how the solution was affected by the specific features of data.

4.1 OBJECTIVE OF THE SOLUTION

Question #19 clustered the solutions by looking at the technical objectives. The options available were few in number and rather broad in subjects on purpose to encourage the SME to think to their solution with the right level of abstraction.

What is the primary technical objective of your solution? a) Combine and correlate different datasets, b) Identify patterns, c) Make predictions, d) Enhance data quality, e) Filter information, f) Visualise information, g) Create a user interface for data access

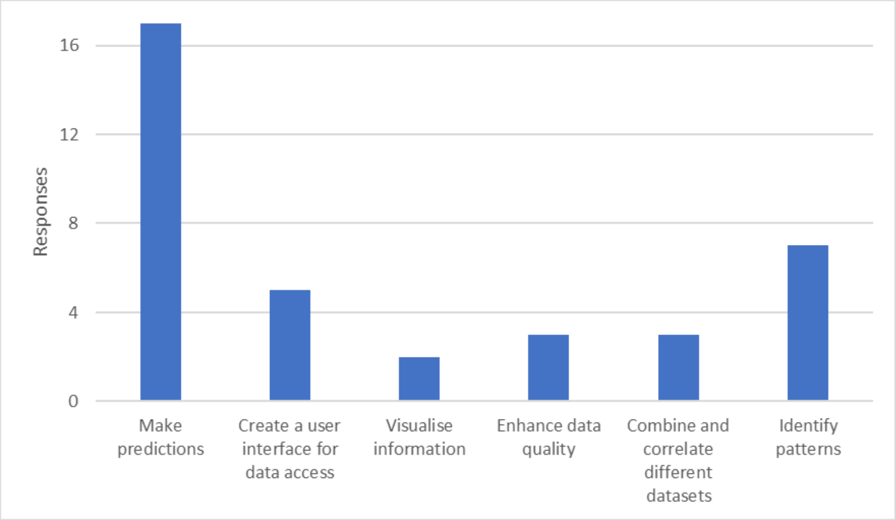

FIGURE 13: PRIMARY OBJECTIVE OF THE SOLUTION

According to the survey, the majority of applications use the available data for making a prediction and identifying patterns.

The reason for this result is partly because of the current market tendency. The most required solutions are about predictions based on existing data, and this tendency was reflected in the applications Data Pitch has received.

4.2 USE OF MACHINE LEARNING

Modern solutions largely use Machine Learning approach. This was not a requirement in the Data Pitch calls, but question #20 was proposed to understand the percentage of SMEs proposing an ML-based solution and 70% of solutions are indeed using ML-based solutions. The next question was about the different type of ML algorithms used.

4.3 TYPE OF MACHINE LEARNING USED IN THE SOLUTION

Many and quite different types of algorithms belong to the Machine Learning class. Question #21 aimed at a finer-grained classification of the solutions present in Data Pitch.

Which methods are used in your solution? Please select all that apply.

a) Regression algorithms (e.g. linear regression, logistic regression)
b) Instance-based algorithms (e.g. k-NN, SVM)
c) Decision tree algorithms (e.g. CART)
d) Bayesian algorithms (e.g. naïve Bayes, Bayesian Network)
e) Clustering algorithms (e.g. k-means, hierarchical clustering)
f) Association rule learning algorithms (Apriori, ECLAT)
g) Artificial neural network algorithms (e.g. MLP)
h) Deep learning algorithms (e.g. CNN, RNN, DBN)
i) Reinforcement learning algorithms (e.g. Q-Learning)
j) Ensemble algorithms (e.g. Random Forest, GBM)

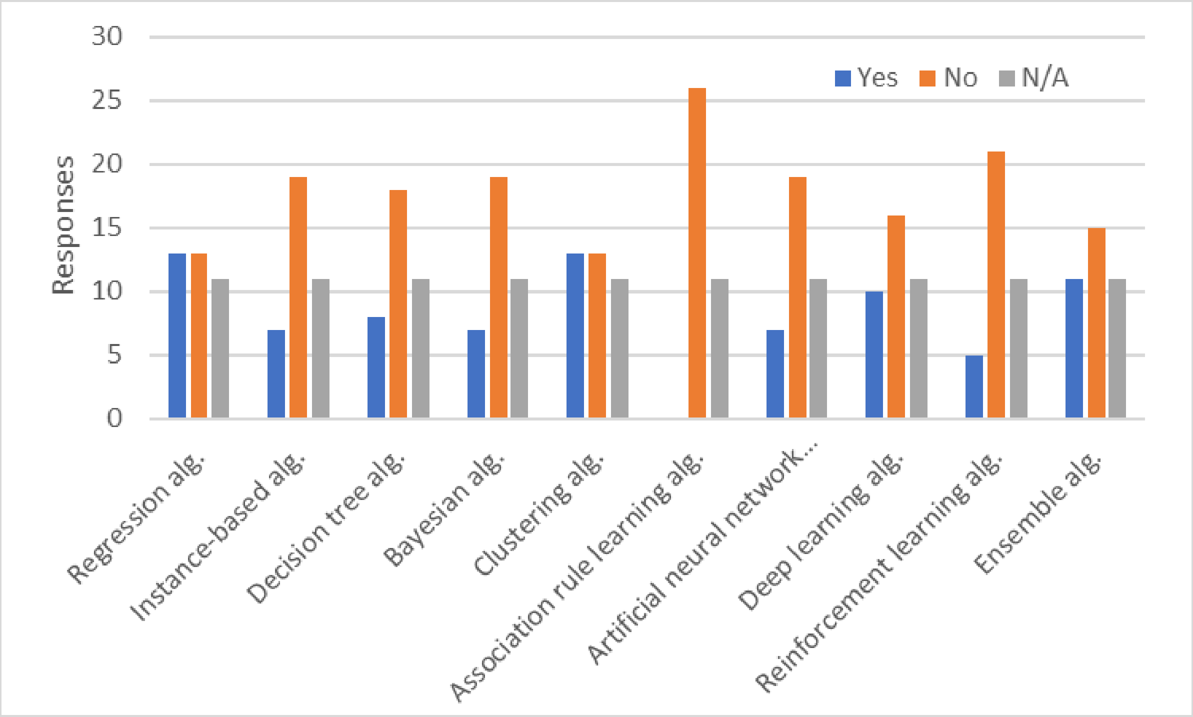

FIGURE 14: TYPE OF MACHINE LEARNING USED FOR THE SOLUTION

The most popular ML algorithms, according to the survey, are regression, clustering and ensemble algorithms, on average, they were used in about 1/3 of the projects. The picture shows a large variety of implemented technical solutions, and a need to tailor the ML algorithm to the business problem. Supporting the identification of the best algorithm to address a business problem is a relatively new open field worth exploring to further increase the support for open innovation.

4.4 CATEGORISATION OF SOLUTION

Question #22 required an abstraction from the technical aspects towards the end goal of the solution. Three classes have been proposed to the SME:

- Descriptive Analytics, which use data aggregation and data mining to provide insight into the past and future: "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 optimisation and simulation algorithms to advise on possible outcomes and answer: "What should we do?"

How would you categorise your solution? a) Descriptive, b) Predictive, c) Prescriptive.

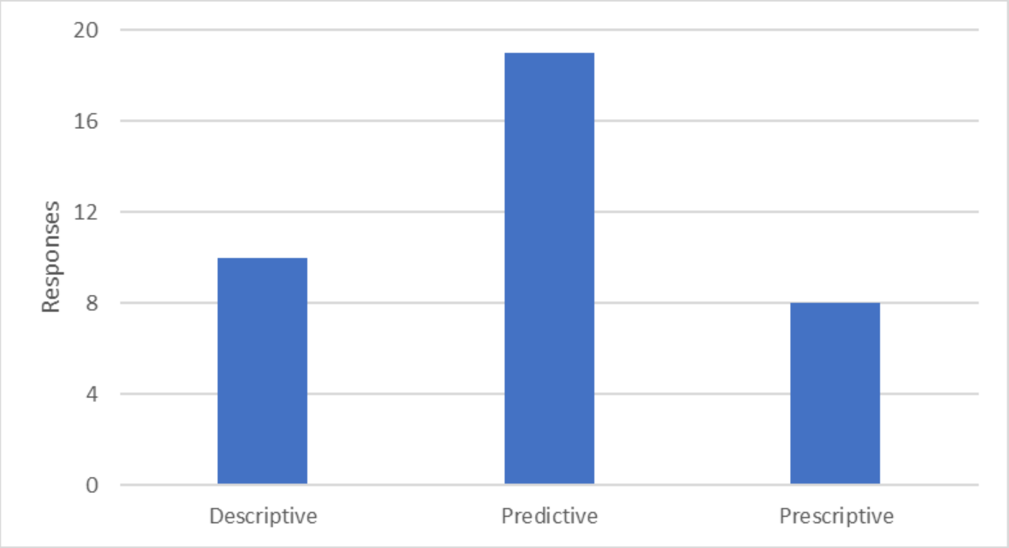

FIGURE 15: SOLUTION CATEGORIES

According to the survey, most of the projects fall, as expected, in the category of predictive applications. This is the field covering a large share of the market. However, a good proportion of the business is still within the descriptive category, showing that understanding the data also offers a good market opportunity. Few applications move in the prescriptive analytics where the goal is to understand how to manipulate the predictions. This field is still very challenging from a technological point of view, and the market is not demanding this type of application yet.

4.5 TECHNICAL CHALLENGES DURING THE ACCELERATION PERIOD

Question #18 was designed to investigate which step in the data value chain has been more challenging, and that possibly requires better support. It considers the typical value chain of:

- Data Access
- Data Engineering
- Building the software solution

For each of the below, please rate the difficulty you had during the acceleration period:(1 signifying easiest, 5 signifying hardest)

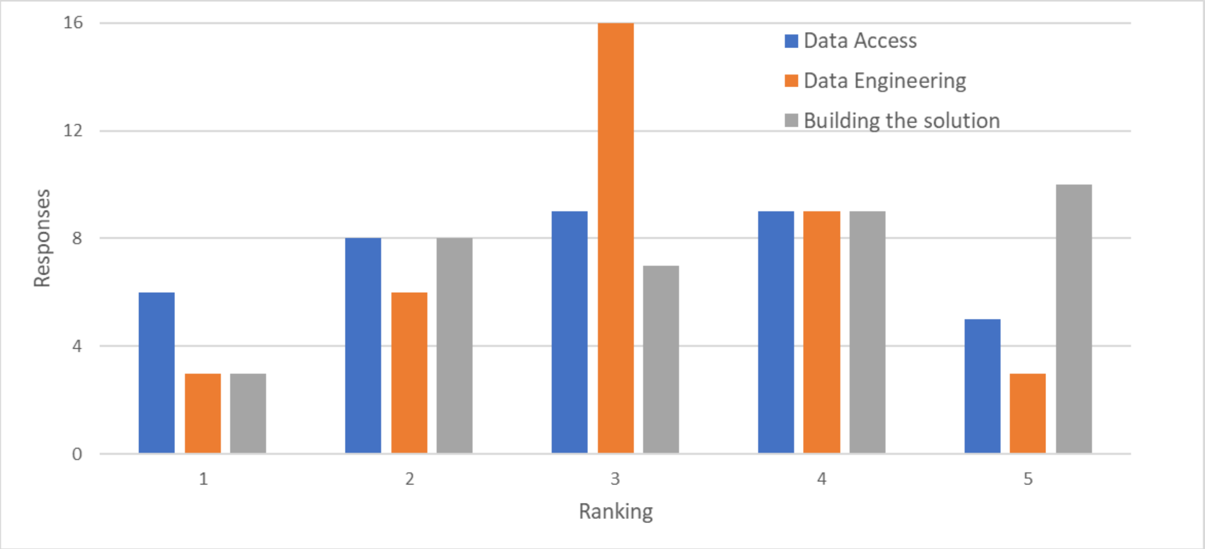

FIGURE 16: CHALLENGES IN CREATING THE SOLUTION

Most of the solutions focus on applying existing algorithms to solve specific business problems. In this approach, it is not a surprise to see that over 3/4 of SMEs found medium to high challenging data engineering. This phase of the data chain extraction (understanding and preparing your data) is intrinsically specific to each domain, even to each business problem. It is therefore required a considerable bespoke effort that is difficult to support with general purposes instruments. The phase of data access is more standard with respect to the engineering one. The 60%+ of SMEs showing difficulties in this phase indicates that more effort should be put in further easing the data access.

5 POST-ACCELERATION

5.1 DISTRIBUTION MODEL

The distribution model has an impact on the software architecture of the solution. Question #26 aimed at classifying the solutions according to it.

What distribution model do you envisage for your solution?

a) SaaS (Software as a Service)
b) On-premise deployment
c) Ad-hoc model (e.g. consultancy with engagement on a case-by-case basis)
d) Would prefer not to say

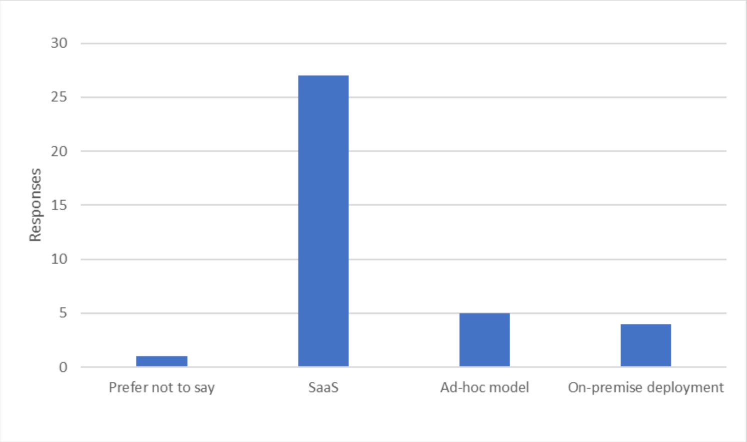

FIGURE 17: DISTRIBUTION MODEL

Most SMEs consider SaaS as a long-term distribution model. This model is widely adopted to provide ICT services. It is also the model that gives the SME full control of the data value chain. However, in a data analysis service, it assumes that the given SME hosts all the data necessary to run the analytic service. This type of solution may be not suitable with data-providers unwilling to move their data out of their premises. The good news is that most of the SMEs have shown a good degree of flexibility during the acceleration period, so they are likely able to adjust their distribution model should it be required by new data providers as a requirement to do business with them. Flexibility by design is an important element of successful SMEs.

5.2 SCALABILITY EFFORT

Innovative solutions need to survive in a competitive market and to reach a stability plateau after the start-up phase. One of the main issues to reach this state is the need to support scalability. Question #28 aimed at providing a qualitative and subjective measure of the envisaged effort to support scalability.

How much effort do you foresee this scalability requiring?

a) *Minimum effort (e.g. only additional hosting space and computational power)*
b) *Medium effort (e.g. minimal software changes)*
c) *High effort (e.g. major software changes*

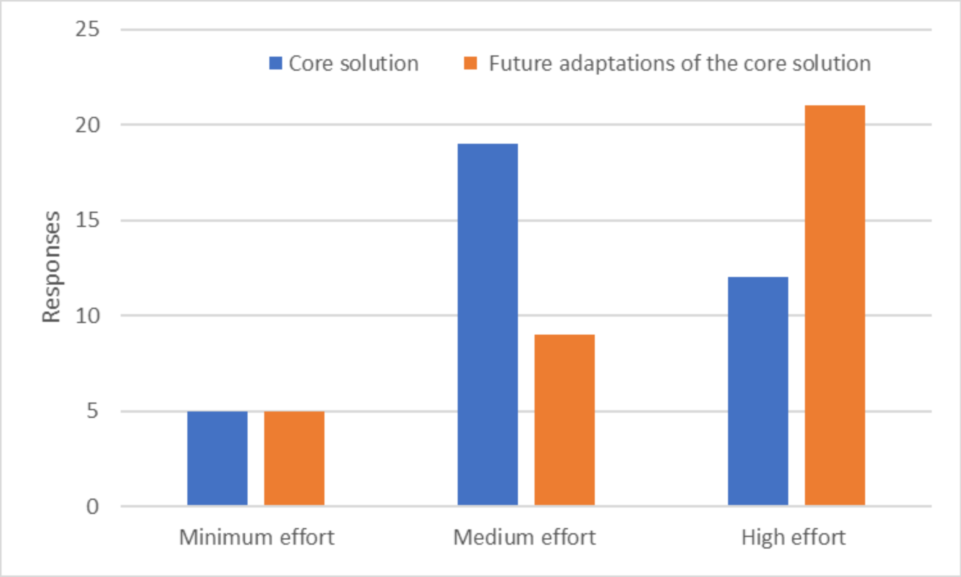

FIGURE 18: EFFORT REQUIRED FOR ACHIEVING SCALABILITY

Most of the respondents indicate that for achieving scalability of the core solution, a medium effort is required. The future adaptations of the core solution according to 55% of the respondents need high effort. To draw quantitative conclusions is out of the scope of the question. A possible conclusion is that an area of support for the SME can be to improve as much as possible the awareness of the scalability needs together with providing guidelines and best practices to support the "by design" approach with stress for the need of flexibility. A well-designed technical solution needs less effort to scale.

6 LESSONS LEARNED

The lesson learned with the second cohort about data access is in line with the results of the survey from the last year. The communication barrier with the data providers still exist, but the technology for accessing data is now mature, and the players in the field generally know how to deal with the technology stack.

"Existing business problems can be often solved with the smart applications of relatively well-known techniques" (c.f. analysis of the first cohort – D2.3). This finding is still true for the second cohort of accelerated SMEs.

The interaction with the SMEs as advisors and the analysis of their solutions have shown that, on average, a strong demand for totally new algorithms is not present. The business needs can largely be addressed with a clever application of known instruments, paired with a deep knowledge of the domains. This is perfectly fine from the open innovation perspective. The innovation can be described as a business-driven knowledge transfer from the scientific world (i.e. academia) to the Industry.

Addressing a business problem is often a matter of applying an existing algorithm to a specific problem. An adventure on exploring disruptive solutions is not necessarily what Corporates and Data Holders, in general, are willing to embark on. The approach to solve specific problems and then move to a bigger one is likely the best strategy to adopt.

From a SMEs perspective, this step by step approach requires a good degree of flexibility, and the best way for achieving it is to apply the "by design" paradigm. By considering possible evolutions and by taking an agile approach, the SME minimises the effort required to pivot or expand towards new problems and new domain.

7 SUMMARY

This report provided a summary of the survey which covered the second round of the Data Pitch acceleration. There were 37 SMEs (8 in the first and 29 in the second cohort) participating in the survey. These SMEs were operating in a wide range of domains, for example: finance, health, environment, broadcasting, etc. Although the sample with 37 respondents is relatively small, nevertheless some general trends can be identified: i) the main motivation of Data Pitch was from the perfective of SMEs to find out how to extract business value from large data. ii) A variety of technical solutions have been applied to solve many business problems.

Despite the need to overcome technical issues in the different stages of the data value chain, all the SMEs have arrived at the conclusion of the programme and with new technical solutions of different maturity levels. All of them have explored the potentiality of data innovation and are now ready to improve their technological offer, either scaling up in their domain or moving to new ones.

From an infrastructural point of view, after the experience of two cohorts of SMEs, it is now clear that computational power is considered a commodity and SMEs are ready to pay for it. Specific infrastructures and support are still required for addressing problems that present critical aspects of performances or security, but these types of scenarios have not been present in Data Pitch, possibly for the intention to include more generic situations.

APPENDIX – QUESTIONNAIRE

The survey text is provided below.

Data Pitch Impact Assessment

Introduction

London Economics [https://londoneconomics.co.uk] has been commissioned by the Data Pitch consortium to carry out an independent assessment of the Data Pitch programme. This survey collects information on your interaction with Data Pitch and the characteristics and features of your solution. No personal information is being collected in this survey. The survey will take approximately 15 minutes to complete.

If you have any questions, please contact Moritz Godel, T +44 (0)20 3701 7708, mgodel@londoneconomics.co.uk

General Information

1. Name of your business: *

Information about the data used in your solution

2. How closely did you interact with Data Providers other than your partnered Data Provider?(1 signifies 'No interaction' and 5 signifies 'Very close interaction')

□

Information about the data used in your solution

3. How many datasets do you use in your solution? Please provide a total number of open, closed and self-generated datasets.(Dataset refers to sets of data that share the same features/characteristics and which your business either receives from a data provider or collects itself). *

Information about the data used in your solution

4. Without Data Pitch, would you have been able to access the same data (or equivalent data that would enable you to implement the same solution)? *

□ Yes

□ No

Information about the data used in your solution

5. What would prevent you from accessing the same data (or equivalent data that would enable you to implement the same solution)? *

□ I didn’t know this data existed

□ I couldn’t locate a provider of this data

□ Too expensive

□ Technical barriers

□ Legal/regulatory barriers

□ Other (please specify):

Information about the data used in your solution

6. Which characteristics of the data used in Data Pitch are the most important for your solution? *

□ Volume of data (enabling more precise predictions, greater coverage, etc.)

□ Richness/granularity of data (enabling higher quality solution, more relevant recommendations, etc.)

□ Complementarity with other datasets (‘missing piece of the puzzle’)

□ Other (please specify):

Information about the data used in your solution

7. How large is the dataset that you obtained from your main (partnered) data provider for use in Data Pitch? *

Number of entries/observations/records *

Size in Gigabytes *

8. What is(are) the primary unit(s) of observation (e.g. customers, card transactions, patent filings, images etc.)? *

Information about the data used in your solution

9. How large is the dataset that your solution uses in total? (I.e. any dataset(s) provided for Data Pitch + any dataset(s) you collected yourself or obtained from other data providers outside Data Pitch) *

Number of entries/observations/records *

Size in Gigabytes *

10. What is (are) the primary unit(s) of observation (e.g. customers, card transactions, patent filings, images etc.)? *

Information about the data used in your solution

11. Please indicate the main type of data used in your solution, select all that apply. *

□ Numeric

□ Text

□ Video

□ Audio

□ Geospatial

□ Image/graphics

□ Other (please specify):

12. How is this data stored? *

□ Semantic databases (i.e. RDF triples)

□ Document oriented databases

□ Relational databases

□ Files

□ Other (please specify):

Information about the data used in your solution

13. Where is the data stored? *

□ Data Provider's infrastructure

□ Commercial cloud paid for by the data provider

□ Your own infrastructure

□ Commercial Cloud paid for by your business

□ Other (please specify):

14. How much control do you have over the data when building your solution? *

□ Full control (e.g. full copy of data freely available to me)

□ Partial control (e.g. data called through API when required)

□ No control (e.g. I send my algorithms to the Data provider and get the results)

□ Other (please specify):

15. What is the update frequency (periodicity) of the data used in your solution? *

	Static data (not updated)	Occasional data (irregularly updated)	Live data (regularly updated)	Real time data (regularly updated with high frequency)	Not applicable
During the acceleration period	□	□	□	□	□
Mature commercial solution	□	□	□	□	□

Questions related to the creation of your solution

16. What resources/capabilities did Data Pitch funding enable you to acquire (rank in order of importance with 1 being most important ; answer N/A if Data Pitch funding was not used for a particular category)? *

	Ranking
Subject matter/domain knowledge	□
Business management skills	□
Software development skills	□
Data science/machine learning skills	□
Other IT skills	□
ICT infrastructure/hardware	□
Marketing/sales skills	□

Please specify any other important resources/capabilities and their relative ranks

Questions related to the creation of your solution

17. How closely did you interact with your partnered Data Provider? (1 signifies 'No interaction' and 5 signifies 'Very close interaction')

□

Questions related to the creation of your solution

18. For each of the below, please rate the difficulty you had during the acceleration period:(1 signifying easiest, 5 signifying hardest)

Data Access □

Data Engineering □

Building the solution □

Questions related to the creation of your solution

19. What is the primary technical objective of your solution? *

- □ Combine and correlate different datasets
- □ Identify patterns
- □ Make predictions
- □ Enhance data quality
- □ Filter information
- □ Visualise information
- □ Create a user interface for data access

Questions related to the creation of your solution

20. Does your solution use Machine Learning?

□ Yes

□ No

Questions related to the creation of your solution

21. Which methods are used in your solution? Please select all that apply. *

□ Regression algorithms (e.g. linear regression, logistic regression)

□ Instance-based algorithms (e.g. k-NN, SVM)

□ Decision tree algorithms (e.g. CART)

□ Bayesian algorithms (e.g. naïve Bayes, Bayesian Network)

□ Clustering algorithms (e.g. k-means, hierarchical clustering)

□ Association rule learning algorithms (Apriori, ECLAT)

□ Artificial neural network algorithms (e.g. MLP)

□ Deep learning algorithms (e.g. CNN, RNN, DBN)

□ Reinforcement learning algorithms (e.g. Q-Learning)

□ Ensemble algorithms (e.g. Random Forest, GBM)

□ Other (please specify):

Questions related to the creation of your solution

22. How would you categorise your solution?Descriptive Analytics, which use data aggregation and data mining to provide insight into the past and future: "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 optimisation and simulation algorithms to advise on possible outcomes and answer: "What should we do?" *

□ Descriptive

□ Predictive

□ Prescriptive

□ Other (please specify):

□

23. How different is your solution now from the idea you had at the start of your involvement in Data Pitch? (An answer of 1 signifies your solution matches the initial proposal exactly; an answer of 10 signifies that the solution is completely different from the proposal).

□

Questions related to the creation of your solution

24. Why is your solution different from the idea you had at the start of your involvement in Data Pitch? *

□ Technical reasons (original solution was too difficult to implement during the acceleration period)

□ Business reasons (changes to the solution resulting in a better/more marketable product)

□ Would prefer not to say

□ Other (please specify):

Information about the post-acceleration period

25. Who do you envisage as the primary customers for your solution? *

26. What distribution model do you envisage for your solution? *

□ SaaS (Software as a Service)

□ On-premise deployment

□ Ad-hoc model (e.g. consultancy with engagement on a case-by-case basis)

□ Would prefer not to say

□ Other (please specify):

27. Are you considering releasing your solution as Open Source? *

□ Yes, fully

□ Yes, parts of it

□ No

□ Would prefer not to say

Information about the post-acceleration period

28. How do you see the scalability of your solution over the next three years?(An answer of 1 signifies 'limited growth', an answer of 5 signifies 'significant growth').

Core solution (markets & customers as currently identified)	□
Future adaptations of the core solution (new markets/application areas/customer groups)	□

29. How much effort do you foresee this scalability requiring?

	Minimum effort (e.g. only additional hosting space and computational power)	Medium effort (e.g. minimal software changes)	High effort (e.g. major software changes)
Core solution	□	□	□
Future adaptations of the core solution	□	□	□

30. On a scale to 1-10, how unique is the product that your solution provides to your customers? Are there similar types of products out there, or is this a one-of-a-kind?(An answer of 1 signifies the product is 'not at all unique' and an answer of 10 signifies the product is 'completely unique').

.

□

31. On a scale of 1-5, how innovative is your solution?(An answer of 1 signifies low innovation, an answer of 5 signifies high innovation).

□

Concluding questions

32. Any other comments or remarks?