D2.7: Annual Report on Opportunities

Workpackage	WP2 – IMPACT: Framing the European Data Economy and maximize Impact
Editor(s):	Sonja Zillner (Siemens AG)
Responsible Partner:	Siemens AG
Contributors	Sonja Zillner, Ingra Marangoni (Siemens AG)
Internal Reviewer	Nuria de Lama, ATOS Andreas Metzger, UDE
Status-Version:	V1.0 for submission
Due to	31/3/2020
Submission Date:	31/3/2020
EC Distribution:	Public
Abstract:	To stimulate and foster the growth of the European data economy, the continuous identification of data-driven opportunities is a critical task. In this deliverable we introduce the design and outcome of the Data-driven Innovation (DDI) Framework including empirical data and findings derived from a complementary research study. We present statistical findings as well as findings related to the success patterns of data-driven innovation opportunities. Those findings are enhanced by the detailed description of successful start-ups of each success cluster as well as by an analysis of the progression of data-driven start-ups in the past years. Finally, the evaluation of the DDI Framework in classroom and workshop settings is summarized. This is the third version of a series of

This document is issued within the frame and for the purpose of the BDVE project. This project has received funding from the European Union's Horizon 2020 Programme (H2020-ICT-2016-2017) under Grant Agreement No. 732630

documents. In this report we have been able to present findings derived from the percentage-frequency analysis, the methodology and findings of the success pattern analysis, a more detailed representation of success data-driven start-ups, a reflections of how our sample changed over time as well as the ongoing evaluation of our workshops.

Table of Contents

List of Figures

List of Tables

Definitions, Acronyms and Abbreviations

Acronym	Title
BDV	Big Data Value
BDVe	Big Data Value ecosystem
BDVA	Big Data Value Association
DDI	Data-driven Innovation

Table 1 Definitions, Acronyms and Abbreviations

Executive Summary

The **Data-driven Innovation (DDI) Framework** addresses the challenges of identifying and exploring data-driven innovation in an efficient manner. It guides entrepreneurs in scoping promising data-driven business opportunities by reflecting the dynamics of supply and demand by investigating the co-evolution and interactions between the scope of the offering (supply) and the context of the market (demand) in systematic manner. The DDI Framework is consists of eight dimensions that are divided into supply side (Value Proposition, Data, Technology and Partner) and demand side aspects. (Ecosystem, Network Strategy, Revenue Strategy, Type-of-Business).

The DDI framework was developed and tested in the context BDVe project and is backed by empirical data and scientific research encompassing a quantitative and representative study of more than 90 data-driven business opportunities. This research study encompasses two parts:

First, to derive meaningful insights into trends, frequencies and distributions, we relied on *classical statistical data analysis*. Based on a percentage-frequency analysis, we could derive findings along the main dimensions of the DDI Framework:

Target Customer: Majority of data-driven start-ups (78%) are addressing B2B markets. Only two out of 90 start-ups of our sample focused on end-customer markets solely. Start-ups addressing end-user needs prefer already established channels to deliver their offering to the users. They tend to rely on partnerships established business partners to bring their offering to users. A second quite frequent strategy used by 19% of start-ups is position data-driven solutions as multi-sided market combining complementary offerings to align private and business needs.

75% of our start-up sample have developed a clear *sector focus*. Companies with clear *sector-focus* have a concrete customer segment(s) in mind for whom a concrete value proposition is delivered. In comparison, we also find start-ups that focus on technology with cross-domain impact. In general, their solution will be used by other intra- or entrepreneurs to build data-driven solutions for end-user.

Value Proposition and Data Value: Two in every three start-ups rely on data analytics in general for generating insights. Among the start-ups using data analytics, 83% rely on descriptive analytics in their offering (i.e. every second start-up). Compared to descriptive and predictive analytics, we can observe that diagnostics and prescriptive analytics are used in less frequency. Only every fifth data-driven start-up is offering solution for automating manual tasks or activities and match-making is observed only in 16% of the cases.

Data: There exist a wide range of different types of data sources that are relevant for developing data-driven innovation. Although only 19% of start-ups were addressing B2C markets, personal data was still the most frequently (67%) used in the analysed data-driven offerings. A second very popular type of data source are time-series and temporal data. 56% of start-ups in our sample rely on this type of data to generate value. The high frequency might be due to the popularity of using behavioral data that is tracked within each user interaction on the web and mobile devices and thus very likely to cover time-series data. Another very frequently used data source are geospatial data with 46% and the usage of Internet of Things (IoT) data is seen in 30% of our sample. Industrial data sources have only been used half as often in compared to personal data, and open data only half as often compared to industrial data.

Technology: Among the five technology areas listed in the Strategic Research and Innovation Agenda (SRIA), *Data Analytics* is used most frequently used. 82% of our start-up sample relied on some type of data analytics to implement data-driven value proposition. The usage of technologies of the *Data Management* area is seen in 41% of the cases and is very much in line with offerings addressing the challenges of processing unstructured data sources. Solutions for *Data Protection* are the least frequently addressed research challenge with 13%. When looking to which extend BDV SRIA Technologies are used in combination, we observed that more than half of the start-ups, precisely 59%, combine more or equal than two technologies, but only 22% combine more or equal than three SRIA technologies.

Network Strategy: For digital and data-driven innovations, network effects are important phenomena to be reflected. In our study, 57% of start-ups rely on network effects. We further distinguished different levels of network effects: First, when data-driven businesses can improve their offering the more data they have available, they are relying on network effects on data level. In our sample this was in 49% of start-ups the case. Second, when businesses are providing a technical foundation for others to build upon, network effects on infrastructure level can be observed. In our sample these have been 12% of the start-ups. Third, in cases where the number of marketplace participants is the key source of value, data-driven offerings can harness network effects on marketplace level. The low number of network effects on marketplaces in our study of 10% indicates the high challenges of building them.

Revenue Model: Often information about the type of revenue models used was difficult to find. Especially in cases when start-ups have been focusing on emerging technical advances, such as drones or autonomous driving, information about revenue models was not available. The most frequently used revenue model in our study was the subscription model. We observed in this context a strong correlation with the spread and high adoption of Software as a service (SaaS) approach, which brings a lot of flexibility when using for deploying data-driven innovation. The second most frequent revenue model is the selling of services in which the person’s time is payed for. Those revenue models are very often used for open software offerings as well as when offerings are not standardized or off-shelf. Advertisement as a revenue model is rarely observed. In our sample, only 2% of the start-ups are applying it and although might seem surprising, it reflects the high percentage of B2B models.

Type of Business: When positioning data-driven offerings on the market, one requires to reflect the associated business or innovation ecosystem. Data-driven service businesses that focus on establishing a partnership with one new business partner have been the most frequently used within our study; 78% used this strategy position their data-driven offering on the market. Compared to data-driven services is the development of data-driven marketplaces significantly more complex as a new marketplace / ecosystem need to be built up. Only 16% of companies in our sample relied on this approach. Another strategy is to identify an existing healthy ecosystem that is already in place which gives the opportunity to position the own offering as niche application. In our sample this strategy was observed in 12% of the cases.

Emerging technology businesses focus on technologies in a very early stage and anticipate a future ecosystem or not yet settled market. In our study it was seen in 9% of the case.

Second, in order to derive *success patterns* of data-driven innovations and understand similarities, we relied on *unsupervised cluster analysis*. Based on the cluster analysis, we could identify six clusters of data driven innovations

- **Cluster A "*Data Pre-Processing*":** The central focus of companies in the Cluster A "Data Pre-Processing" is on delivering solutions for the pre-processing of heterogeneous data sources, such as images or videos. Due to the addressed high technical complexity, companies tend to be very focused and do not provide additional analytics or automation capabilities. In addition, as companies of this cluster seem to focus on generic technical challenges without concrete customer value in mind, they are developing twice as often sector agnostic solutions compared to the overall sample-
- **Cluster B *"Internet-of-Things applications":*** Start-ups of the Cluster B rely on Internet of Things (IoT) Technology as part of their offering. As the IoT technology is integrated with multiple different types of technologies as well different types of data analytics, start-ups of this cluster are confronted with high data pre-processing, technical as well as data integration challenges. Although companies of this cluster rely on industrial data twice as often as others as well as tend to use different kinds of data sources, they are less likely to rely on unstructured data.
- **Cluster C: *"Industrial services":*** Companies of Cluster C are characterised by the usage of industrial data sources, and on the contrary the usage of unstructured data is less frequent in comparison to the overall sample. In terms of value proposition, they tend to cover the whole range of data analytics as well as provide high value through process automation. Companies of this cluster seem to be prepared to make us of available services for processing IoT data but do not include IoT technology as component of their overall offering.
- **Cluster D: *"Descriptive value"***: Companies in Cluster D are very focused on descriptive analytical services for non-industrial data sources. The usage of other analytical services is very significantly lower compared to the average, same is true for match-making functionalities or process automation capabilities. In addition, Cluster D companies are more likely to rely on semi-structured, media and time series data. All Cluster D offerings are positioned on the market as data-driven service and are generating income mainly by subscription (94% of the offerings) as well as by selling of services in half of the cases.
- **Cluster E *"Predictive value":*** All Cluster E companies focus on predictive analytics, often being combined with other analytical values, such as descriptive, diagnostic or prescriptive. They are 50% more likely to rely on unstructured data sources as well as 33% more likely to rely on personal data compared to the average and do not use industrial data. By relying on a smaller

number of different types technologies, they are confronted with less integration efforts, interfaces and partners. For generation revenue, Cluster E start-ups rely 50% more often on asset sale and selling of services and less frequently on subscription when compared to the average.

- **Cluster F *"Connecting peers"*:** The main value proposition of Cluster F start-ups is the match-making functionality allowing to connect supply and demand from business to consumer side with three quarters of consumers being end-customers and the remaining being business customer. Start-ups in Cluster F are very likely to rely on commission fee (60% compared to 10% in average), harness network effects on marketplace level and establish multi-sided markets / data-driven marketplaces. The high usage of personal data (87%) indicates that also in B2B marketplaces personal data is used for implementing match-making algorithm. Match-making is the central functionality provided, its implementation relies beside personal data to a high percentage on semi-structured and semantic data.

The results of the research study guided the fine-tuning and updating of the DDI framework as well as helped to identify **success patterns** of data-driven business opportunities.

By relying on the DDI Framework, we have now a method and comprehensive content in place that we can share with members of the wider data ecosystem for exploring data-driven business opportunities in the context of webinars, workshops as well as online coaching formats.

1 Introduction

In this series of reports, we are investigating *emerging business opportunities in the European Big Data Landscape* with the goal to promote their uptake in Europe. Based on the Data-driven Innovation (DDI) Framework developed in the first three years of the project (and documented in the first report of this series (Zillner, Timan and Kotterink, 2018) as well as in (Zillner, 2019)), we conducted a quantitative and qualitative study of a well-selected set of data-driven start-ups to identify promising patterns to guide future investment decisions. In addition, we evaluated the usage of the DDI Framework, Canvas and Guiding Questions in practical settings within a university course, in a series of Business Model Canvas[1] style workshops and webinars. Both, the results of the research study as well as the experiences in using the framework in industrial and academic trainings, provide valuable guidance for further fine-tuning and updating the DDI Framework.

By relying on the DDI Framework, we have now a method in place that we can share with members of the BDV ecosystem for exploring data-driven business opportunities. The DDI Framework complemented with content slides, comprehensive set of methods and guiding questions is used for industrial trainings and university lectures. The derived characteristics and patterns of successful start-ups help entrepreneurs, innovators, and managers to scope their data-driven business opportunities in a way that industrial investment decisions are becoming more likely.

With the DDI Framework approach we plan to engage in the future with entre- and intrapreneurs, SMEs and start-ups to help them scope promising business opportunities. The main impact of the work described in this series of reports is to provide a continuous assessment of emerging, data-driven business opportunities by designing a tool that can be offered to EU and national programmes addressing start-ups and SMEs that would be interested to validate or further explore the data component of their value propositions. The data collected through the DDI research study as well as applications in different contexts provide important insights to guide entrepreneurs in scoping and driving data-driven business opportunities by following success patterns and related findings.

The first part of the document (Section 0) provides an overview of the DDI Framework and explains how this deliverable relates to its two predecessors. Section 0 gives a summary of how the DDI Framework was developed and describes the methodology of our research study to evaluate the DDI Framework. The two subsequent sections are summarizing the findings of the study: While Section 4 summarizes all statistical-driven insights related to the eight dimensions of the DDI Framework, Section 0 covers the findings of the pattern analysis detailing the particularities of each data-driven success cluster. To make the above-mentioned findings more explicit we are describing in Section 6 for each cluster two examples in more detail along the dimensions of the DDI Framework. Due to the length of our study, we also had the

[1] The Business Model Canvas developed by Osterwalder (Osterwalder and Pigneur, 2010) has been the basis for many entrepreneurs, business developer, experts, etc. to structure brainstorming as well as the detailed exploration of business models by relying on a canvas completed with example categories as well as guiding questions.

chance to observe the performance of our sample-set over time, the findings are covered in Section 7. We conclude the deliverable with a short evaluation of some of the workshops that we have accomplished in the last year in Section 8 and an outlook to our future work in Section 0. Finally, in Section 0 the interested reader will find an updated[2] description of the DDI Ontology including all updates derived from the research study and practical elaborations.

[2] in comparison to the version initially reported in (Zillner, 2019)

2 Data-Driven Innovation (DDI) Framework

2.1 Overview of the DDI Framework

The **Data-driven Innovation (DDI) Framework** addresses the challenges of identifying and exploring data-driven innovation in an efficient manner. It guides entrepreneurs in scoping promising data-driven business opportunities by reflecting the dynamics of supply and demand by investigating the co-evolution and interactions between the scope of the offering (supply) and the context of the market (demand) in systematic manner.

The DDI Framework is based on a conceptual model in form of an ontology with a set of categories and concepts describing all relevant aspects of data-driven business opportunities. Its categories are divided into supply side and demand side aspects. On the *supply side* the focus is on the development of new offerings. For a clearly defined value proposition, this includes the identification of and access to required *data* sources, as well as the analysis of underlying *technologies*. On the *demand side* the focus is on the dynamics of the addressed markets and associated ecosystems. The analysis includes the development of a *revenue strategy*, a way forward of how to harness *network effects* as well as an understanding of the *type of business*. As data-driven innovations are never done in isolation, the identification and analysis of potential development *partners* as well as partners in the *ecosystem* help to align / balance the supply and demand aspects in a way that its competitive nature will stand out. An overview of the eight dimensions represented in the DDI Canvas is displayed in Figure 1.

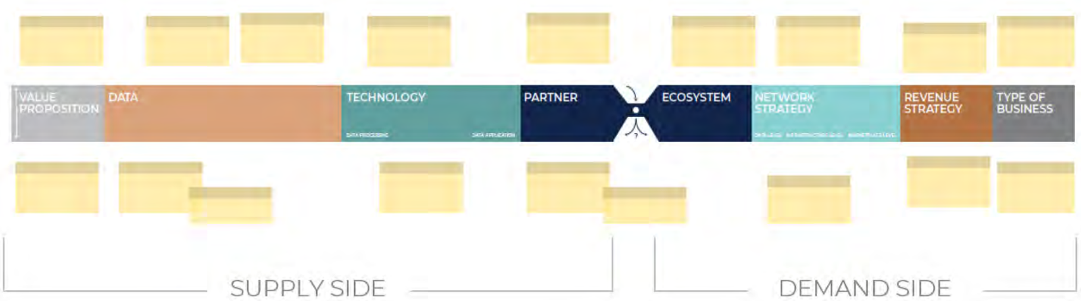


Figure 1 DDI Canvas with eight dimensions guiding the exploration of the relevant aspects of DDI

The DDI framework was developed and tested in the context of the Horizon 2020 BDVe project[3] and is backed by empirical data and scientific research encompassing a quantitative and representative study of more than 90 data-driven business opportunities. The results of the research study guided the fine-tuning and updating

[3] For more details see Zillner. S. D 2.6 Annual Report on Opportunities (BDVe Deliverable), April 2019 and Zillner. S. et al.: D 2.5 Annual Report on Opportunities (BDVe Deliverable), March 2018

the DDI framework as well as helped to identify **success patterns** of successful data-driven innovation.

Currently the DDI framework is used to run workshops with BDV PPP projects, data-driven start-ups, SMEs and with large enterprises. The DDI framework consists of the following two main elements:

- **DDI Canvas** guiding user in the exploring all relevant dimensions on the supply and demand side of a data-driven innovation in systematic manner as well.
- **DDI Guiding Questions** and methods that support user in exploring each dimension in the required level of detail by investigation the aspects mentioned by guiding questions as well as by applying complementary methods.

In the context of the BDVe project, a series data-driven innovation workshops in business settings as one-day or half-day trainings for professionals were designed and conducted[4]. In addition, the framework was evaluated with 35-50 students per year participating in the university course "Data-driven Innovation"[5] at the Technical University Munich in the summer semester of 2018 and 2019. In 2020 the DDI lecture will be offered online due to the Covid-19 crisis. By relying on the DDI Framework, we have a method and comprehensive content in place that we can share with members of the wider data ecosystem for exploring data-driven business opportunities in the context of webinars, workshops as well as online coaching formats.

2.2 Relation to Deliverable 2.5 and 2.6

We developed the DDI Framework over the last three years, and intermediary results were reported in BDVe Deliverable 2.5 and 2.6. While developing the DDI Framework we continuously integrated the feedback and learning we received from the research study as well as by applying it in academic and industrial settings. Below we summarize the major changes during the DDI development phase:

- **Name of the Framework**: Although the initial name "DemoX Model" (**D**ata-driven **e**cosystem **mo**delling) was well perceived and easy to remember, we decided to rename our approach into "Data-driven Innovation (DDI) Framework" as people did pay little attention to the underlying semantics of the term DemoX, i.e. an approach for the data-driven modeling of ecosystem, probably because ecosystem modelling is a very abstract concept. Due to that, the term DemoX was like an art name and we felt that we are losing the opportunity to convey with the name of the approach also aspects of its value proposition.

[4] DDI Workshop with PPP projects (Riga, 27.6.2019), with Corporates (Nürnberg 11.7.2018 / Erlangen 18.2018, with SMEs (Murcia, 1.10.2019); with EDI-Start-Ups (Bilbao, 6.11.2019), Webinar with Data-Pitch Start-ups (online 19.9.2019), DDI Presentation at Data-driven Business Model Workshop (Riga, 26.6.2019)

[5] https://www.ldv.ei.tum.de/en/lehre/data-driven-innovation/

- **Dimensions of the DDI Framework**: The dimensions on the supply and demand side of DDI Framework changed over the last three years. Some of the dimensions, i.e. framework conditions we initially assumed would be of high relevance, were removed in later versions, as we did neither find a) explicit information about this in our empirical investigations nor b) valuable methods to explore those aspects in the given context.
- **Layout of the DDI Framework**: The initial layout of a hexagon with supply and demand side dimensions was changed two times. First into an infinity symbol to indicate that the process of exploring DDI is a continuous and iterative one and that DDIs require to be focused on the supply side to able to pass the eye of the needle (in the middle of the infinity symbol) to reach the needed focus to explore options on the demand side without being overwhelmed with opportunities. As the infinity symbol was not appropriate to be used as canvas in the group brainstorming session along the process, we consolidated a linear design of the DDI Framework in the form of a banderole that can be easily be used and reused on any available whiteboard.
- **Start-up Analysis**. For the empirical evaluation of the DDI Framework, we decided to accomplish a research study about data-driven business opportunities. Due to the fact that in established ventures, such as SMEs or corporates, it is very difficult and less transparent to collect data about any type of emerging business opportunity / innovation, we decided in our study to concentrate on a representative set of data-driven start-ups. The design of the research study and first initial results were reported in D2.6. In this deliverable we will report on the continuation of the research study, i.e. the comprehensive statistical data analysis to derive important messages as well as the cluster / pattern analysis to identify promising success patterns and best practices of data-driven businesses
- **Start-up Examples**. In deliverable D2.6 we described some exemplary start-ups to motivate the DDI Framework. In this deliverable, we are using examples of start-ups to provide the reader illustrative examples while we discuss messages and patterns.
- **Evaluation** of our approach: As in D2.6 we will have a dedicated section in this deliverable that is summarizing the evaluation and assessment of the main accomplished workshops.

3 DDI Research Study

We conducted a research study to build and test the Data driven Innovation Framework. As already detailed in Deliverable D2.6, the research study is divided into four phases as shown in Figure 2.

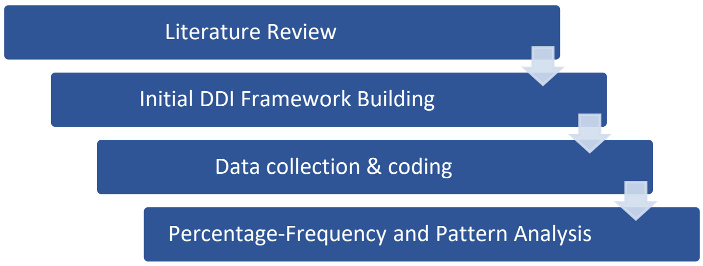


Figure 2 The four phases of the DDI Research Study

As Step 1 to 3 are reported in detail in Deliverable D2.5 and D2.6, we will not cover them in this deliverable. In D2.6 first initial outcomes of the Percentage Frequency Analysis were described. To ensure readability of documents and comprehensiveness of analysis, those results are seamlessly integrated into the next sections.

3.1 Methodology

To derive meaningful insights into trends, frequencies and distributions, we relied on classical statistical data analysis. In order to derive success patterns of data-driven innovations and understand similarities, we relied on unsupervised cluster analysis. In the following we will shortly discuss the methodological foundation of both approaches before we subsequently discuss in full detail and length the outcomes and findings of our study in Sections 4 and 5.

3.1.1 Percentage-Frequency Analysis

The first method we employed to assess which variables could shape data-driven business opportunities was a percentage-frequency analysis. The goal in using this method was to understand how frequent a variable was observed in our data.

From our dataset[6] already explained in the Deliverable 2.6, we know that we coincidently had 90 variables and 90 observations (start-ups) that were marked either

[6] Although the implementation of data-driven business opportunities is not restricted to a certain type of business or organisation, we decided in this study to focus on start-ups only due to two reasons: First, due to the lack of available public information as larger corporates and SMEs barely share information about their business or innovation designs and decisions; and second, due to high interdependence with existing *operations*. Innovation activities in large corporates and SMEs are often constrained by existing infrastructures, legacy systems or prior systems. For identifying successful data-driven start-ups, we needed to define a *measurement for success*. We decided to choose start-ups with funding between US$2M and US$10M to obtain the ones a) that had already convinced some ventures to invest into them b) that would already have their product validated and c) but that would still be a young start-up. (For more details see Deliverable 2.6)

by the presence of the variable (1) or the absence of it (0). For example, one of the variables described whether a start-up was doing business in the B2B domain. For the start-ups for which this was true, we marked a one, for the start-ups that did not target B2B we marked a zero. In our percentage-frequency analysis we then counted how many start-ups were marked with one and divided by the total number of observations for that variable. Using the same example, we have observed that 88 start-ups out of 90 were marked with one, which means 98% of companies target B2B customers in our model.

3.1.2 Pattern Analysis

For deriving promising success patterns of data-driven innovation, we relied on unsupervised cluster analysis. Cluster analysis is a descriptive modelling method. The aim is generating insights and comprehension about the data, enabling the discovery of unsuspected structures (Hand, Mannila and Smyth, 2001). It also allows the discovery of distribution patterns and correlations among the data features (Hartmann *et al.*, 2014). In our case, similar data-driven businesses are grouped into one cluster, and subsequently the characteristics of each cluster is explored and analysed in further detail.

(Hand, Mannila and Smyth, 2001) distinguishes between two different objectives for the cluster analysis. One is segmentation and the other is identifying natural subclasses. The authors also emphasize that the methods for clustering will always surpass the notion of distance.

The four steps of the cluster analysis

According to (KETCHEN Jr. and SHOOK, 1996) and (Mooi and Sarstedt, 2011) cited by (Hartmann *et al.*, 2014), the process consists of four steps:

- Choosing the cluster algorithm,
- Choosing the number of clusters,
- Selection of cluster variables,
- Validate and interpret the clustering results.

(Hand, Mannila and Smyth, 2001) highlights that the chosen cluster method requires to match the cluster objectives and needs to consider what specific insights shall be derived from the clusters, i.e. "*Data mining, after all, is about discovering the unexpected, so we must not be too determined in imposing our preconceptions on the analysis"* (Hand, Mannila and Smyth, 2001).

In the context of our DDI research study, we have been interested in identifying the core aspects that determine the nature of a successful data-driven innovation / business opportunity.

3.1.2.1 Choosing the Cluster Algorithm

Due to study similarities, we have decided to select the criteria established by Hartmann et al (2014), which is the partitioning method. This method is used to classify observations into different groups based on how similar they are, and there are different algorithms to choose from. In our case we selected the k-medoid algorithm instead of the k-means algorithm, also following the recommendation of Hartmann et al (2014) as the problem is similar[7].

3.1.2.2 Choosing the Number of Clusters

The logic of the partition-based cluster algorithm is to split the data set into *k* disjoint sets of points (clusters). The points within the set are as similar as possible, and as distant as possible to the neighbour clusters (Hand, Mannila and Smyth, 2001).

As with any machine learning algorithm, you need to specify parameters before you can run the model. In our case, we need to specify the number of clusters *k*. This is done by exploring the data. Here is common practice to use a software that allows you to play around with ease. In general, there is no best solution and rerun the analysis several times is the suitable option.

The decision about the optimal number of clustering variables was influenced by the findings of (Hartmann *et al.*, 2014). The authors highlight that there is no right or wrong answer as this is an unsupervised learning technique and the number of possible clustering variables depends on the sample size. As it is generally recommended to have a sample size of 2^m, being *m* the number of clustering variables, the ideal number of clustering variables would be between 6 (2^6=64) and 7 (2^7=128) as our sample set has 90 start-ups being analyzed. We finally used 6 variables to identify the clusters.

3.1.2.3 Selection of Cluster Variables

According to (Hartmann *et al.*, 2014) the essential and primary question to answer in the first step of the cluster analysis process is which variables from the model could potentially define a cluster. This step depends solely on the researcher's knowledge and (domain) expertise and raising different hypothesis is part of the process that will be validated on a later step. (Hand, Mannila and Smyth, 2001) highlights that in the clustering process, a scientific elaboration and insight can be judged as useful. We accomplish the selection of cluster variables in two steps.

Step 1: Identifying the Discriminative Power of the Variables

As first step, we need to select the variables that will determine our cluster analysis. In other words, we need to identify those variables that have discriminative power for describing data-driven innovation opportunities by relying percentage-frequency analysis

For doing so we first assessed the relevance of each variable by imposing a threshold determined by a range of discriminatory power. In our case, the latter meant excluding from the analysis the variables that would not determine a cluster, but instead would only help us understand how frequent a characteristic is for the cluster. For example, this means that as the variable B2B that was observed in 98% of the time, this high percentage would not be relevant to identify a data-driven business opportunity. Based on an extensive exploration of the data set, we decided to exclude variables with less than 20% or higher than 80% as potential candidates for determining a

[7] For more details regarding the specification of the algorithm, check (Hartmann *et al.*, 2014) on pg. 13-15

cluster. Due to their high or respectively low frequency, their interpretation did not bring new insights, i.e. they lacked discriminative power. At the end of this assessment, we found out that 60 variables were relevant for the next step, the pattern analysis.

Step 2: Selection of Variables

As next step, we needed to select six to seven variables out of those 60 variables. For determining the relevance of each variable for the cluster analysis, we continuously assessed its relevance for identifying the *core aspects that determine the nature of data-driving innovation.* For this selection step, it was important to understand which variables would be able to describe the critical aspects of the data-driven start-up. In this context, several different assumptions have been explored and tested:

- Shall we rely on a mix of variables ranging from different dimensions or shall we rely on variables of one dimension only?
- Shall we select variables from the supply side only, or from the demand side only? Or would it be variables from both sides?
- Would a set of specific variables be sufficient expressive in terms of cluster building, for instance, would a set of variables describing the value proposition or a set of variables describing the revenue model be sufficient to determine meaningful cluster of data-driven innovations.

The selection of variables was a highly iterative process that relied on a detailed analysis of each outcome for each optional cluster outcome by incorporating a mix of visual interpretation using a plot, number of overlaps, the percentages given by the Principle Components Analysis (PCA), and a qualitative question on whether it would answer the main question.

By using the ‘Cluster’ and ‘Factoextra’ libraries from RStudio and applying an iterative process using different combinations of variables, we present below two plots to illustrate our outcomes when testing different combinations of variables. From Figure 3 we see the PCA values in percentage on the axes, each colour block in the plot area is a candidate for a cluster, and each data point in the same colour is the observation.

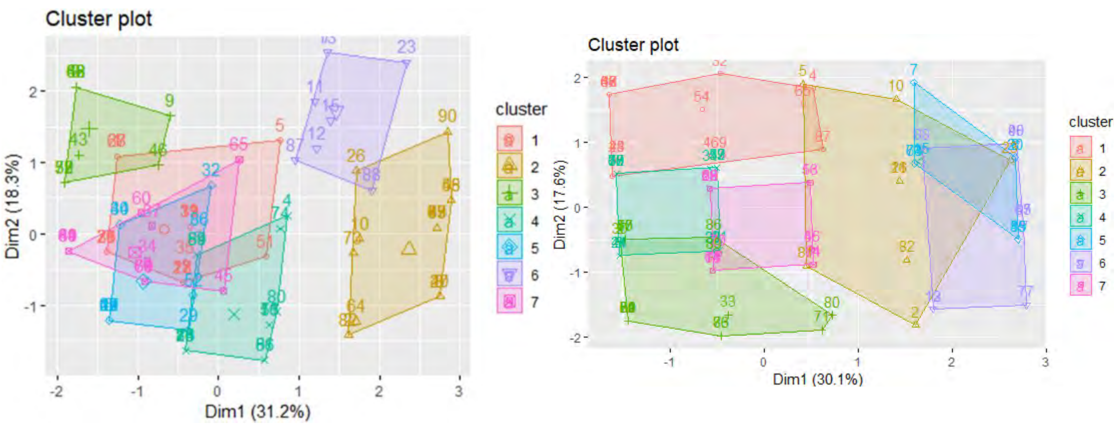


Figure 3 Two examples Cluster Plot from experimentation

After extensively challenging different outputs as the examples above, we identified three groups of variables that were different from each other but could result in good combinations for determining a cluster. The next step was assessing each group in more detail, which besides of all steps mentioned above included understanding the distribution of observations per potential cluster, and mainly which group could best answer our question. This analysis showed us that one of the three groups had a clear distribution for each cluster, they were unique and would answer our question.

The variables that identified the clusters were:

- **Match-making** (Dimension: Value proposition) enables the automated orchestration of value generation. Match-making algorithms are mapping the demand side requirements with the supply side resource capabilities. Typical examples that make use of matching algorithm are dating platforms.
- **Descriptive Analytics** (Dimension: Value proposition) is the most frequently used analytics. Its main objective is to explain what had happened in the past by providing the analyst, businessperson or expert a view of key metrics that measure the area of interest. The traditional business intelligence and data mining applications fall into this category. They provide a very important basis for developing a deeper understanding of the underlying data sources.
- **Predictive Analytics** (Dimension: Value proposition) is about forecasting. Its main objective is to predict what will happen in the future, for instance the estimated point in time of a machine outage or forecasting a quantifiable amount of customer, etc. A predictive model relies on a variety of variable data that have a relationship which the event the model aims to predict
- **Industrial Data** (Dimension: Data) refers to any data assets produced and used in industrial settings of all areas. Often those are data produced from productions lines, energy systems, infrastructures, etc. In general, industry data is "closed data" meaning that it is "owned" (in terms of access rights) by the entity operating the product or machine or thing producing the data and that the data is likely to cover confidential information, e.g. a log file from a medical MR might contain valuable information indicating IPRs of the MR itself. This also includes operational data in general produced in the context of operating any type of IT system
- The **Internet of Things (IoT)** (Dimension: Technology) is a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction[8].
- **Data-driven Marketplace** (Dimension: Type of Business) is a platform that connects supply and demand and greatly benefit from network effects. This

[8] https://www.i-scoop.eu/internet-of-things/

means the more provider & consumer (balanced) the more attractive the marketplace becomes

In our study we have been interested in the supply and demand side dynamics. And to which extend the supply side determines the demand side or vice versa. The above listed variable determining the six cluster are mainly describing dimension from the supply side. Why mainly? Although Data-driven Marketplace is describing a demand-side dimension, this aspect has high correlation with the match-making variable. **With the determining variables mainly describing aspects from the supply side, the findings from the cluster analysis related to the demand side is of high value for identifying future data-driven opportunities.**

3.1.2.4 Validation and Interpretation of the Cluster

After following Step 1 and 2 described above that allowed us to identify six clusters, we validated the selected clusters by following some additional steps. The first step was running a correlation analysis among the variables that identified the clusters, followed by the interpretation of the results and applicability through case studies.

The correlation analysis is a statistical method that quantifies the relationship between two variables. We used Pearson correlation coefficient as this does not assume normal distribution of the variable. It ranges from -1 to +1 with negative one means perfect negative correlation, zero means no correlation and positive one means perfect positive correlation. Our expectations were that the outcomes would be near zero, which would mean a weak correlation. From the figure below we observe that this happened to the correlation of almost all variables. The only case where we observed a stronger correlation between two variables was the *Data-driven Marketplace* and *Match-making*. This is coherent to our cluster interpretation as these two variables originated the Connecting Peers cluster. We used RStudio to run the Pearson correlation analysis and the outcome is Figure 4.

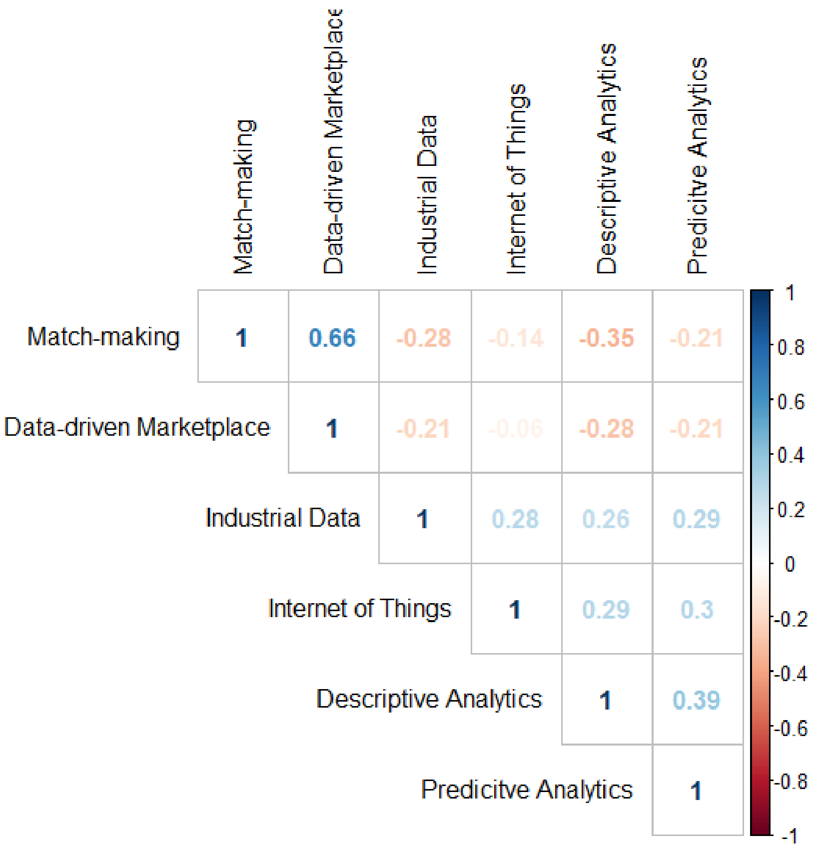


Figure 4 Pearson Correlation among variables

In order to assess the statistical validity of the Pearson Correlation values, we did a significance test using the probability value, also called p-value. With a significance level of 5%, meaning the threshold of the p-value ≤ 0.05, we assessed if the correlation outcome was significant. This is observed by the * in the table below and only the correlation between Match-making-Internet of Things and Internet of Things-Data-driven Marketplace were not significant. We will further validate this result by analyzing the outcomes from each cluster.

	Row	Column	Correlation	p-value	
1	Match-making	Descriptive Analytics	-0.35	0.001	*
2	Match-making	Predictive Analytics	-0.21	0.049	*
3	Descriptive Analytics	Predictive Analytics	0.39	0.000	*
4	Match-making	Industrial Data	-0.28	0.007	*
5	Descriptive Analytics	Industrial Data	0.26	0.014	*
6	Predictive Analytics	Industrial Data	0.29	0.006	*
7	Match-making	Internet of Things	-0.14	0.195	
8	Descriptive Analytics	Internet of Things	0.29	0.006	*
9	Predictive Analytics	Internet of Things	0.30	0.004	*
10	Industrial Data	Internet of Things	0.28	0.008	*

11	Match-making	Data-driven Marketplace	0.66	0.000	*
12	Descriptive Analytics	Data-driven Marketplace	-0.28	0.007	*
13	Predictive Analytics	Data-driven Marketplace	-0.21	0.049	*
14	Industrial Data	Data-driven Marketplace	-0.21	0.043	*
15	Internet of Things	Data-driven Marketplace	-0.06	0.566	

The next step of our validation process was to interpret our results. Rather than a static method, this approach was an iterative process that allowed us to derive the messages presented in the Section 4 and Section 0 while validating it by cross-checking the different results and applying to case studies, presented in Section 6.

3.1.2.5 DDI Cluster Results

The clustering process described above resulted in six different clusters, as shown in Table 2. The table describes on the left side the six variables (including their related dimension) that determined the process of cluster identification. On the right side, the identified clusters are displayed by describing to which extent each variable is present in each cluster.

An outstanding result is the fact that Cluster A is encompassing a set of start-ups that are characterised by the absence/non-fulfilment of the six discriminating variables. Therefore, for this cluster it was much more difficult to define a differentiating characteristic beyond the label “the other”. On the contrary, Cluster F is the only cluster that was determined by the two variables, e.g. *Match-making* and *Data-driven Marketplace.* Due to the rather high correlation of both variables, see Figure 4, this result is not very surprising. All other clusters indicated one key determining variable. It is important to mention that due to the relatively small sample size, all presented quantitative data related to the percentage distribution of the different variables of data driven innovation is indicative but not statistically significant.

Data driven Innovation Cluster Result							
		Cluster					
Variable	**Dimension**	**A**	**B**	**C**	**D**	**E**	**F**
Match-making	Value Proposition	0	0	0	0	1	13
Descriptive Analytics	Value Proposition	0	12	11	16	9	1
Predictive Analytics	Value Proposition	0	12	8	0	13	1
Industrial Data	Data	5	10	12	0	0	0
Internet of Things	Technology	0	14	0	3	0	1
Data-driven Marketplace	Type of Business	0	2	0	0	0	12
Count start-ups		20	14	12	16	13	15

Table 2 DDI Cluster Result

As nearly all determining variables are describing dimensions of the supply side, the identified clusters establish the basis to extract insights related success patterns on the demand side. Table 2 presents some general statistics about the different dimensions (value proposition, data, technology, revenue strategy, network strategy and type of business) of data driven innovation / business opportunities we could derive from our analysis of 90 start-ups. By comparing the six data driven innovation cluster by the type of value proposition they rely on, three different clusters could be identified. For instance, Cluster F relies on match-making capabilities to generate value for their different types of users. DDI Cluster D relies mainly on "descriptive analytics" to derive insights and value for customers whereas the Cluster E make primarily use of "predictive analytics". Cluster C is mainly determined by the usage of industrial data whereas Cluster B is characterised by Internet of Things applications.

3.1.3 Description of Clusters

Based on the above analysis, we could derive six clusters of data driven innovations (see Figure 5)

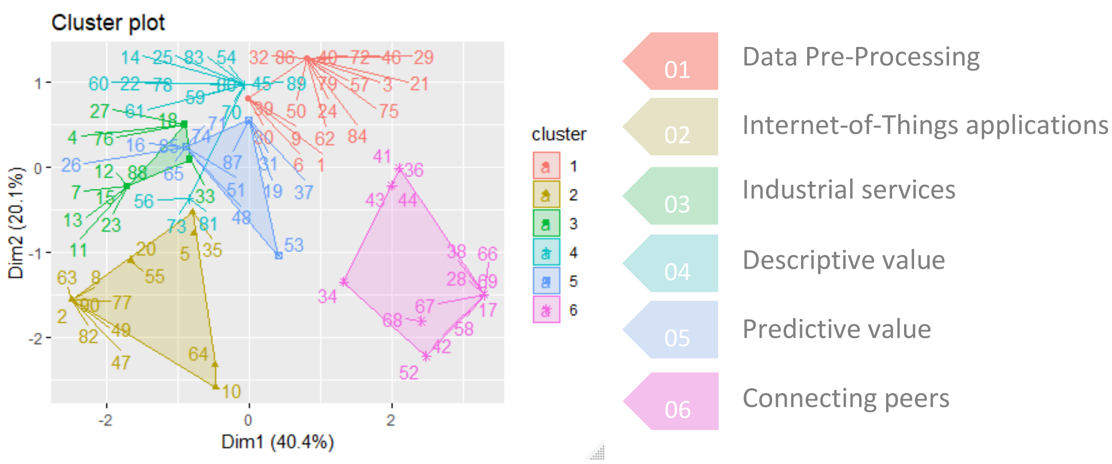


Figure 5 Overview of six main clusters

Cluster A "*Data Pre-Processing*": The central focus of companies in the Cluster A "Data Pre-Processing" is on delivering solutions for the pre-processing of heterogeneous data sources, such as images or videos. Due to the addressed high technical complexity, companies tend to be very focused and do not provide additional analytics or automation capabilities. In addition, as companies of this cluster seem to focus on generic technical challenges without concrete customer value in mind, they are developing twice as often sector agnostic solutions compared to the overall sample-

Cluster B *"Internet-of-Things applications":* Start-ups of the Cluster B rely on Internet of Things (IoT) Technology as part of their offering. As the IoT technology is integrated with multiple different types of technologies as well different types of data analytics, start-ups of this cluster are confronted with high data pre-processing, technical as well as data integration challenges. Although companies of this cluster rely on industrial data twice as often as others as well as tend to use different kinds of data sources, they are less likely to rely on unstructured data.

Cluster C: *"Industrial services":* Companies of Cluster C are characterised by the usage of industrial data sources, and on the contrary the usage of unstructured data is less frequent in comparison to the overall sample. In terms of value proposition, they tend to cover the whole range of data analytics as well as provide high value through process automation. Companies of this cluster seem to be prepared to make us of available services for processing IoT data but do not include IoT technology as component of their overall offering.

Cluster D: *"Descriptive value"*: Companies in Cluster D are very focused on descriptive analytical services for non-industrial data sources. The usage of other analytical services is very significantly lower compared to the average, same is true for match-making functionalities or process automation capabilities. In addition, Cluster D companies are more likely to rely on semi-structured, media and time series data. All Cluster D offerings are positioned on the market as data-driven service and are generating income mainly by subscription (94% of the offerings) as well as by selling of services in half of the cases.

Cluster E *"Predictive value":* All Cluster E companies focus on predictive analytics, often being combined with other analytical values, such as descriptive, diagnostic or prescriptive. They are 50% more likely to rely on unstructured data sources as well as 33% more likely to rely on personal data compared to the average and do not use industrial data. By relying on a smaller number of different types technologies, they are confronted with less integration efforts, interfaces and partners. For generation revenue, Cluster E start-ups rely 50% more often on asset sale and selling of services and less frequently on subscription when compared to the average.

Cluster F *"Connecting peers"*: The main value proposition of Cluster F start-ups is the match-making functionality allowing to connect supply and demand from business to consumer side with three quarters of consumers being end-customers and the remaining being business customer. Start-ups in Cluster F are very likely to rely on commission fee (60% compared to 10% in average), harness network effects on marketplace level and establish multi-sided markets / data-driven marketplaces. The high usage of personal data (87%) indicates that also in B2B marketplaces personal data is used for implementing match-making algorithm. Match-making is the central functionality provided, its implementation relies beside personal data to a high percentage on semi-structured and semantic data.

4 Findings from Percentage-Frequency Analysis

In the following subsections, we will summarize all finding derived from the percentage-frequency analysis. We will represent the findings along the eight dimension the DDI Framework.

4.1 Target Customer

4.1.1 Finding: Majority of start-ups address B2B markets

A first observation is that most start-ups address B2B markets. For instance, the company *Artomatix*[9] addresses the digital game business and provides support for 3D artist (professionals) in generating realistic 3D art of textures and texturing. Done manually, the generation of 3D art of textures and texturing is quite time consuming and tiring. By providing automatic support in the seamless generation of non-repetitive high-quality textures to build 3D worlds, characters and items, helps 3D artists to accomplish this task 10 times faster. The business customers of *Artomatix* are major studies or individual developers being addressed with different revenue models.

In our sample only two out of 90 start-ups focused on end-customer market solely. For example, the start-up *UnaliWear*[10] who is building and selling smart watches for supporting elderly people in their daily activities, for instance in finding the right way or by reminding them about medications.

This very low number of pure B2C start-ups might be surprising at first glance as in general start-ups are attributed to be very close to their customers as well as being highly skilled in exploring unknown user needs guided by an emphatic mindset. Is this a rumor? On a closer look, it becomes clear that start-ups accomplish the exercise of identifying user needs with many different target user groups ranging from end customers, such as individuals or households, to users in business settings, such as professionals with clear roles and tasks, to stakeholders in the sector. However, for delivering the offering to the users, start-ups seem to prefer already established channels. For instance, by integrating data-driven solutions as plug-in in established software applications the access to large customer bases can be realized. This specific strategy was for instance selected by *Artomatix* by giving 3D game developers access to their offering via Photoshop plug-in.

[9] https://artomatix.com/

[10] https://www.unaliwear.com/

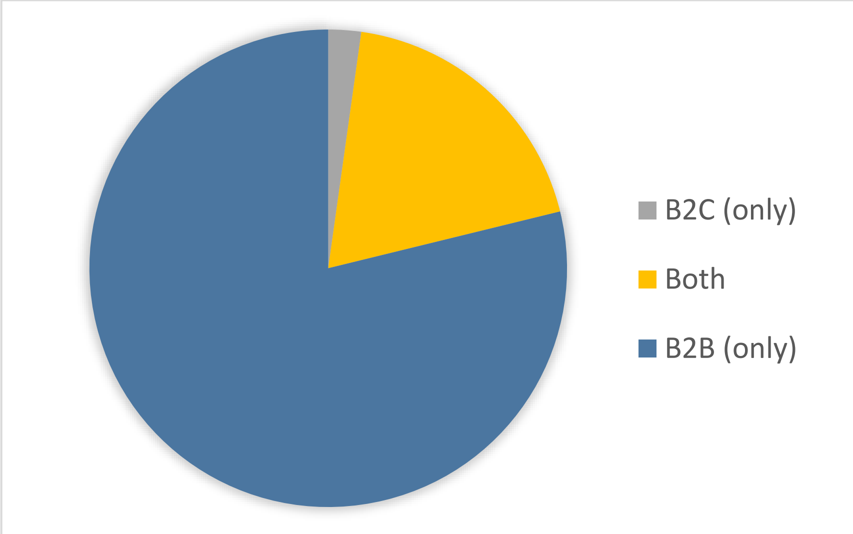


Figure 6 Majority of Data-Driven Start-ups address B2B Markets

A second frequent option to deploy data-driven solutions is via multi-sided markets combining with freemium offerings for private users with complementary business needs. Approximately 19% of start-ups address in such a way both B2B and B2C markets by building upon multi-sided business models or digital marketplaces. An example for multi-sided business models is the data analytics platform *Verv* of the company *Greenrunning* which allows collating and analyzing real-time electricity data to generate a range of data services souring energy, appliances and customer analysis. End-customers can use this service to analyse the 'energy signature' of their own electrical appliances which helps them to explore the costs of each appliance on a real-time basis. By giving usage permission of one owns data for analytical purposes, the service is free for end-customer and can be accessed via a mobile app. This allows *Greenrunning* to build a unique data set as basis for offering appliance usage data and personalised insights of interest to a wide range of industries including utilities, retail, home care and insurance. Those monitoring services are offered under licence agreements to business customers.

To summarize, we can state that start-ups are very good in exploring wide range of user needs. However, we also observe that in their value capture strategy they tend to rely on established business partners to bring their offering to users. This is caused on the one side by the limited willingness of end-customer to pay (money) for an offering. With end-customers being willing to pay only very small amount of money, companies need to build large scale customer bases (long tail phenomena) which again implies a lot of effort and investments for connecting and gaining customers. A popular shortcut for such situations is to partner or collaborate with large business players that have access to the needed customer base. The second way forward is the design of multi-sided markets that surpass the limited willingness of end-customers to pay money by introducing new currencies, such as the sharing of data.

4.1.2 Finding: Majority of start-ups have clear sector focus

75% of our start-up sample have a clear sector focus. Companies with clear *sector-focus* have a concrete customer segment(s) in mind for whom a concrete value proposition is delivered. For example, *CloudMedx*[11] Inc. designs artificial intelligence driven software for medical analytics. Clinical partners at all levels can derive meaningful and real-time insights from their data and intervene at critical junctures of patient care. Its underlying Clinical AI Computing platform uses healthcare specific NLP and Machine learning to generate real-time clinical insights at all points of care to improve patient outcomes. By relying on evidence-based algorithms and deep learning a wide variety of structured and unstructured data being stored in clinical workflows can be understood and used for decision making.

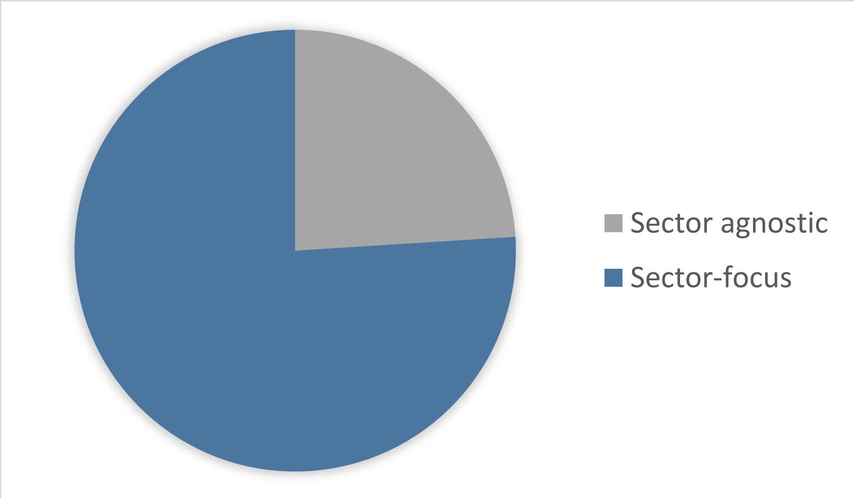


Figure 7 Three of four start-ups have a clear Sector Focus

Often, due to the particularity of the data the sector focus is obvious. For instance, data intelligence solution in the medical sector are very customized solution addressing the challenge of preprocessing and annotating the wide range of heterogeneous data sources in clinical and medical settings.

In comparison, we also find start-ups that focus on technology with cross-domain impact. In general, their solution will be used by other intra- or entrepreneurs to build data-driven solutions for end-user. For instance, the start-up *DGraph Labs*[12] is offering an open source distributed graph data base. The company is planning to release an enterprise version that is closed source as well as a hosted version (as it is easier to run hosted services for customers than trying to help them debug every issue on their own). Customers are using the service to build their own sector specific applications.

Summarizing, sector-specific data-driven offering are much more frequent than technology-driven solutions. This is caused by the very different pre-processing challenges of the data sources in the various sectors as well as the higher possibilities to identify target groups in concrete sector settings. Most sector-agnostic offerings are intermediate functionalities addressing developers to build customized solutions.

[11] http://www.cloudmedxhealth.com/

[12] https://dgraph.io/

4.2 Value Proposition / Data Value

Data value refers to the insights that can be generated out of data and how this can be used in a particular user or business context. In accordance to its value and complexity, we distinguish four different types of analytics that are used for generating different types of insights (Zillner, 2019):

- **Descriptive Analytics** is the most frequently used analytics. Its main objective is to explain what had happened in the past by providing the analyst, business person or expert a view of key metrics that measure the area of interest. The traditional business intelligence and data mining applications fall into this category. They provide a very important basis for developing a deeper understanding of the underlying data sources.
- **Diagnostic Analytics** aim to explain the root-case of a problem. Its main objective is to explain “why something happens”. Those applications are often based on rule-based or semantic model capturing important background knowledge as well as flexible dashboards empowering the expert / user to explore or filter relevant features.
- **Predictive Analytics** is about forecasting. Its main objective is to predict what will happen in the future, for instance the estimated point in time of a machine outage or forecasting a quantifiable amount of customer, etc. A predictive model relies on a variety of variable data that have a relationship which the event the model aims to predict.
- **Prescriptive Analytics** is the basis for decision support and decision automation. Its main objective is to inform the machine or the user about the best course of action or strategy. Prescriptive analytics requires a deep understanding of the underlying engineering, business, mental or other processes in order to transfer analytics results into recommended actions.

In addition, we observe that data is used to generate a lot of value for the automated orchestration of any type of consumers and producer, supply and demand, etc. So-called **Match-making** algorithms enable the automated orchestration of value generation by mapping the demand side requirements with the supply side resource capabilities. Typical examples that make use of matching algorithm are dating platforms. Moreover, data analytics is increasingly used for automating existing processes, e.g. tasks originally accomplished by humans are replaced by smart algorithms. **Process automation** includes all applications that help to replace manual tasks or activities by machines or algorithms. This can range from applications that automate one very particular human tasks in comprehensive manner, such as supporting lawyers finding relevant cases, to applications enabling the automated orchestration of existing processes and workflow, such as AI-based planning and scheduling algorithms that develop strategies or actions sequences for execution by intelligent agents, autonomous robots and unmanned vehicles. [13]

[13] All definition related to data value have been originally been defined in Deliverable D2.5 and are repeated in this document to increase readability.

In our study we are asking which type of approach data-driven start-ups are using too generate value.

Figure 8 gives an overview of how data value is generated across the sample set. The details we will discuss in the subsequent subsections.

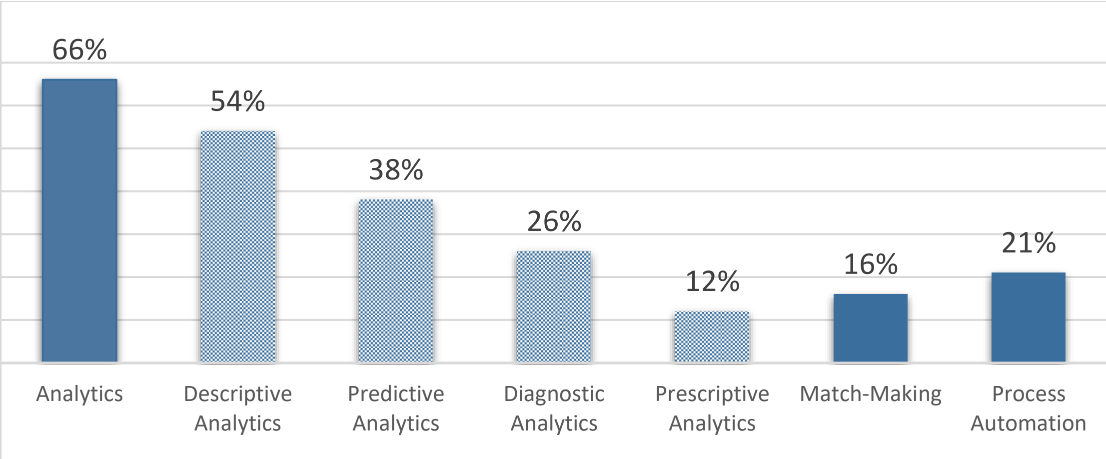


Figure 8 Overview of how data value is generated

4.2.1 Finding: 66% of start-ups use data analytics to generate value

Two in every three start-ups rely on data analytics in general for generating insights. Among the start-ups using data analytics, 83% rely on descriptive analytics in their offering (i.e. every second start-up).

> For instance, the start-up *Apptopia*[14] is using descriptive analytics to provide an app analytics, data mining and business intelligence services. They collect, measure, analyse and provide user engagement statistics for mobile apps and visualize the aggregated data in classical dashboards. The unique selling point of their offering is the high number of data points they are able to integrate and visualize, i.e. they state that they rely on "more different data points than nearly any other app data provider in the world". The insights, that can be generated by descriptive data on this large data sets, is of interest for the worldwide mobile app developer community as it allows them to compare their own app performance with competing or related apps. Whenever app developers are engaging with the *Apptopia* platform to benchmark their own apps, additional valuable data sets can be generated. By offering free-of charge descriptive analytics-based dashboards, *Apptopia* is able to attract large number of developers to use their platform which again allows them produce high valuable data sets that can be sold to business customers.

4 out of 10 start-ups of our sample set, relied on *predictive analytics* to generate value for its users.

[14] https://apptopia.com/

> For instance, the start-up *Visiblee*[15] collects IP addresses and cookies of all web-site visitors and uses this to predict in real-time the identity of unknown visitors. By relying on those real-time prediction, the company is able to increase the leads[16] by three times.

Compared to descriptive and predictive analytics, we can observe that diagnostics and prescriptive analytics are used in less frequency. In concrete numbers, only every fourth start-ups provided insights why something happened (diagnostic analytics) and only 12% of start-ups provided concrete recommendations about what to do next (prescriptive analytics). This might be caused by the fact, that both, diagnostic and prescriptive analytics require deep domain knowhow to be incorporated into the analytics, for instance knowledge about the underlying processes.

> For instance, the start-up *CheckRecipient*[17] relies on diagnostic analytics to prevent highly sensitive information being sent to the wrong people over email. By combining machine learning and natural language processing technologies, it is able to identify anomalies and mistakes in outgoing emails. By scanning historical email data, their algorithm can derive conventional usage patterns and behaviours in companies' email systems. Based on such patterns their platform can detect abnormal behaviour and mistakes in outgoing emails and stop them. Detailed warning explaining the anomaly classification are provide to the user. The development of such a solution requires access to large amount of historical emails, which is due to their high privacy constraints in general very difficult to access.
>
> The start-up *Iris Automation*[18] building collision-avoidance systems for industrial drones is offering prescriptive analytics. By relying on computer vision technology, their software solution is able to detect moving aircrafts and obstacles and to determine the safest course of action to avoid collisions.

The development of both offering require the access to very specific data sets and/or access to underlying domain knowledge. In the first example the access to large amount of historical emails is needed, which is due to high privacy constraints in general only in rare cases possible. In the second example, very specific domain knowledge needs to be coded into automated algorithms. This requires not only access to domain experts but often is also a time-consuming task which requires additional resources and often has clear limitation in terms of scalability of algorithms to other application scenarios (as underlying processes are in general very specific to one application).

4.2.2 Finding: 40% of start-ups are using combinations of data analytics

For implementing data-driven offerings, in general, several algorithm and approaches are combined. This is also true for the four different types of data analytics discussed before (see Section 4.2). In our sample, four out of ten start-ups use more than two

---

[15] https://www.visiblee.io/en/home/

[16] In sales context leads refer to contacts to potential customers.

[17] https://www.tessian.com/ (as links to the original website of CheckRecipient are automatically redirected to this website, we assume that the start-up has renamed themselves.

[18] https://www.irisonboard.com/

different types of data analytics, and 19% of start-ups rely even on three and more types of analytics to generate value.

For instance, *Eliq*[19] provides a comprehensive platform for the intelligent energy monitoring for utilities. The AI-powered app offers a wide range of insights

- by relying on *descriptive analytics* to show periodic energy consumption patterns that can be drilled down into different time frames, i.e. yearly, monthly, …, hourly)
- by relying on *diagnostics analytics Eliq* helps users to identify potential "energy leaks" or potential sources of energy theft.
- by integrating external data sources, such as extreme whether change forecast, *Eliq* can inform users that their energy consumption is likely to change significantly (*predictive analytics*). Utilities benefit from such information as they can customize marketing communication accordingly
- by relying on *prescriptive analytics,* the *Eliq* platform can not only inform users about increased energy consumption but also recommend strategies to overcome such high consumption scenarios, e.g. by upgrading or replacing devices with higher efficiencies. This allows utilities to establish a personalized and targeted user engagement.

Eliq is an example of a start-up that establishes a unique value proposition and competitive edge by offering a wide range of analytical services. We want to highlight that this is not a common pattern. The majority of start-ups (62%) is focusing on only one analytical offering.

4.2.3 Finding: Process automation and Match-making are less frequently used

Every fifth data-driven start-up is offering solution for a*utomating manual tasks or activities*.

For instance, *Warwick Analytics*[20] is offering a machine learning-based tool to automatically label unstructured text sources that serve the basis to automate processes involving customer interactions, for instance in contact centres or when processing insurance claims.

In a wide range of publications and studies related to the future of work, we find similar statements related to what will happen. For instance, (Frey and Osborne, 2013) was one of the first publications that made related estimates, such as *"According to our estimate, **47 percent of total US employment is in the high risk category**, meaning that associated occupations are potentially automatable over some unspecified number of years, **perhaps a decade or two**."* Other publications have similar messages, only the speed of the change might vary from study to study. For data-driven innovation we can infer that the potential for offerings addressing the aspect of automation has a much larger range of applications scenarios than today already addressed. However, again as the automation of business processes also requires deep domain expertise as well as related data sources, not everybody has the needed

[19] https://eliq.io/

[20] https://warwickanalytics.com/

resources to implement those automation solutions. In addition, we one need to keep in mind that not all types of business processes can benefit data analytic solutions. For instance, (Davenport, Jeanne and Robert, 2010) already list in a very early book which types of business processes are well-suited to be analysed for developing data driven solutions:

- *Data rich*, e.g. weather forecast
- *Information intensive*, e.g. disease management
- *Asset-intensive*, e.g. predictive maintenance for enabling effective usage of expensive resources
- *Labor-intensive*, e.g. crime prevention
- Dependent on *speed and timing*, e.g. real-time decision, scheduling for logistics
- Dependent on *consistency and control*, e.g. drug interaction or preliminary diagnosis
- Dependent on *distributed decision making*, e.g. supply chain optimization, load balancing
- *Cross-functional or cross-business* in scope, e.g. value-based healthcare, mobility as service
- *Low average success rate*, e.g. diagnose of rare disease.

Match-making is observed in only 16% of the cases, and the value proposition is uniquely diverse. It ranges from boat rentals such as Zizoo[21], to connecting readers to brands such as Selectionist[22], and connecting technical know-how in a crowd services platform such as Mila[23].

> For instance, on *Mila*, users can find, book, and rate tech-savvy people or offer their technical know-how. For instance, customers facing problems when configuring the WLAN router, can use the platform to look up who is able to help with the issue. Through a rating system, the specialists are evaluated by the customers and their abilities and prices are described.

4.3 Data

In general, we observe that the used data sources very much influence the efforts in data pre-processing as well as scope of offering. In case, a data-driven innovation is based on image data, we can conclude that some image segmentation algorithm needs to be in place. In accordance to the how specific or domain specific the underlying image data set is, new pre-processing image algorithm need to be

[21] www.zizoo.com

[22] www.business.selectionnist.com/

[23] https://www.mila.com/

developed. Or in case of personal data, GDPR compliant services and in case of industrial or operational data, data privacy methods need to be in place.

For that reason, we recommend to explore the data assets very early when scoping your data-driven innovation. The data exploration will help you to understand

a) whether the envisioned value proposition can be realized. Very often, we face the situation that the data quality is not good enough for generating the needed insights.
b) how much efforts are needed to create data in high quality. Often the raw data has not yet the data quality needed. The good message is that there exist many approaches to increase the quality of data for this scoped purpose. However, the expected return always needs to be aligned with the efforts needed. Other projects in the BDV PPP have made similar experiences (Metzger, Franke and Jansen, 2020)

In the following, we will give an overview which data types and sources are used how frequently in data-driven innovations.

There exist a wide range of different types of data sources that are relevant for developing data-driven innovation. An overview of different types of data sources was given in Deliverable 2.5. and Deliverable 2.6. How often the different types of data sources are used for is depicted in Figure 9.

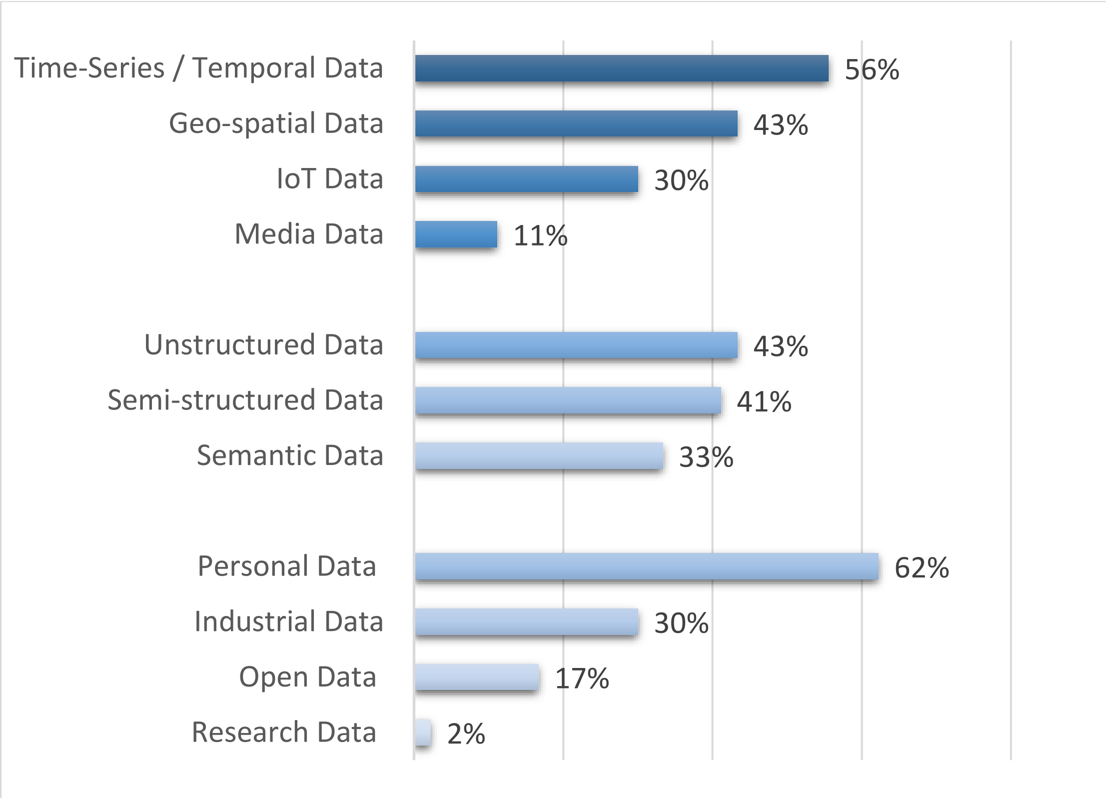


Figure 9 Overview of how frequent different data types have been used

4.3.1 Finding: Personal Data is most frequently used data source

Outstanding is the finding that *personal data* is most frequently used in data-driven offerings. This is a very impressive number given the fact that only a very low number of companies in our sample (19%) were addressing Business to Consumer markets. In consequence this also implies that a high percentage of start-ups addressing business customers in Europe[24] need to handle the constraints of the General Data Protection Regulation (GDPR).

> For example, *Oncora Medical*[25] is using personal data to fight cancer. The US-based company collects data about cancer patients including information related to treatments and clinical outcomes through an intuitive software used by doctors. Their objective is delivering predictions that can help design better radiation treatments for patients, as well as enabling precision medicine in radiation oncology. The data collected is personal data and thus is sensitive and has higher standards of protection.

In Europe, data concerning health receives special attention under the GDPR, and can only be used if the data subject has given explicit consent, if its processing is necessary for a series of diagnosis and prevention measures, as well as for public interests in the area of public health[26].

Industrial data, i.e. any data assets that are produced or used in industrial areas of all areas, is second type of data type which has high data protection requirements. In comparison to personal data, industrial data is used only half as often. Organisations seem to be reluctant (in particular if they do not see immediate value) to share their industrial and operational data with third parties, such as start-ups, because they are afraid to reveal relevant business secrets.

> For instance, the start-up *Locus*[27] is using collecting and processing real time fleet tracking data to optimize the supply chain performance. By analysing its customer's operational data, the company can predict customer´s behaviour, such as quantity and targeted locations as well as frequency of deliveries.
>
> Another example is *PlutoShift*[28] with its platform helping industrial customers to improve its operational efficiency by identifying inefficient patterns of energy usage by analysing customers' data stored in the cloud and operational sensor data. With energy being high cost driver, *PlutoShift* can help industrial customers to reduce resource consumption and operating cost.

In the ongoing public discussion about data sharing and protection, the usage of personal data versus industrial data are often handled /assumed as non-overlapping scenarios. We could not confirm this assumption in our sample data. In concrete, (15 startups) every second start-ups that is using industrial data is also making use of personal data. (or every fourth start-ups using personal data is also making use of

[24] As noted earlier our sample set is not restricted to European start-ups only, as we wanted to make sure that our analysis covers worldwide excellence. As we do not have precise numbers for European data companies, the sentence is formulated with some unambiguity.

[25] https://oncoramedical.com/

[26] See https://www.pega.com/insights/articles/gdpr-and-healthcare-understanding-health-data-and-consent for more details

[27] locus.sh

[28] plutoshift.com; previously called Pluto AI

industrial data). We are highlighting this observation, as the handling of so-called mixed data sources, i.e. data sources that are of type industrial and personal data, is not regulated in a thought through manner: each individual has the right of data portability. It is unclear how the right on data portability can be applied on mixed data sets given the fact that the access and usage of data of industrial data lies within the organization producing the data. In consequence, several players have access and usage rights in the context of mixed data source.

> For instance, *RuleX*[29], a company providing explainable AI in the context of business processes is relying on both, operational and personal data. Their value proposition is to provide augmented, automated and autonomous operational decision making. *RuleX* is providing solutions for a wide range of sectors, including banking, insurance, supply chain, CRM, manufacturing, energy or healthcare. When the customers' operational performance is influenced by end-users' interaction, any operational decision support needs to combine the processing of personal / user interaction data and operational data derived from the business processes. In cases personal data is involved, *RuleX* is providing GDPR-compliant data processing.

A second very popular type of data source are *time-series and temporal data*. 56% of start-ups in our sample rely on this type of data to generate value. The high frequency is likely because behavioral data is tracked within each user interaction on the web and mobile devices are very likely to cover time-series data. In addition, whenever changing events, such as weather, movement, trends, conditions, are monitored, temporal data is captured. Both aspects make this type of data source so popular.

> For instance, *Visiblee*[30], a US-based company is helping its customer by increasing their website lead generation by factor 3. *Visiblee* technology gathers personal data from website visitors, such as IP-addresses, cookies as well as metadata captured via a script installed on the webpage across time. By combining the customer data with additional external data sources covering relevant business information, *Visiblee* can enrich the information provided to its customer in real-time.

But the usage of time-related data does not automatically imply the usage of personal data.

> For instance, *Taranis*[31] is a company from Israel that supports farms to produce good crops by identifying, analyzing and treating early signs of crop threats. By combining observation data produced by different devices, such as drones, airplanes and satellites, insights and prediction about emergencies, nutrient deficiency, weeds, insects, disease and count tassels can be derived. By detecting areas that require specific treatments farmers get valuable guidance in optimizing their crop. All the data they are relying on is non-personal data.

[29] www.rulex.ai

[30] www.visiblee.io

[31] https://taranis.ag/

Another very frequently used data source are *geo-spatial data* with 46%, in other words 46% of data-driven offering a geographical component as part of their value generation.

> For instance, *Onfleet*[32] addresses the challenge of last-mile delivery for thousands of companies including the food and beverage industry. By relying on geo-spatial data to plan, track and monitor deliveries, they can offer an end-to-end route planning, dispatch, communication and analytics platform. By considering time, location, capacity and traffic, they can produce efficient routing solutions. This is completed by real time updates being send to drivers and by constantly monitoring and optimization of overall performance.

The usage of *Internet of Things (IoT) data* is seen in 30% of our sample. In IoT scenarios, large number of objects, such as physical devices, vehicles, home appliances, and other items embedded with electronics, software, sensors or actuators are connected and able to exchange data. We expect that in the near future the number and complexity of IoT scenarios, and thus the usage of IoT data, will increase significantly with the release of 5G. For now, it means that the start-ups are using the available technologies to connect in real time and generate value through it.

> For instance, *Senseye*[33] is offering a scalable predictive maintenance software that is used on the shop-floor by the maintenance and operations people. Their software takes machine condition and operations data from the factory history, IoT middleware or database solutions, and focuses on automatically delivering advanced Predictive Maintenance insights in an easily understandable manner. *Senseye* collects vibration data from accelerometers or high precision condition monitoring equipment, pressure data from changes in pressure and flow, torque from modern drive units, and current from Programmable Logic Controller (PLC) or via retrofit sensors that the customer might attach. With this data and additional ones that might be of value, *Senseye* is able to perform a health calculation of the machines and they deliver this insights and alerts through a well-designed web-based platform.

Open data is only used in 17% of our sample. The usage of open data[34] is compared to other data sources with 17% is rather low. In Table 3 we list some examples including a description of how the open data sources are used.

[32] https://onfleet.com/

[33] https://www.senseye.io

[34] In general, start-ups are not documenting details about the origin of their data sources on their website or public communication. For that reason, the below findings are based on the subjective assumptions of the research team's members involved in the coding process.

Start-up	Cluster	How open data is used?
TRX Systems	A	This location system delivering tracking services inside GPS-disabled buildings needs maps to run their application. The solution relies on **open maps**.
Taranis	B	Identify, analyze and treat early signs of crop threats to make informed decisions, lower costs and maximize yield. The solution needs **open maps** and **weather data**.
Eliq	B	Eliq provides software to monitor electricity usage and demand patterns. Their service relies on open **weather data**.
Emagin	C	Operational data from SCADA systems, enhanced by **GIS data** (open data)
Vivacity Labs	E	Vivacity Labs makes intelligent cameras to gather transport data, using the latest machine learning & computer vision techniques. Their services rely on **open maps**.

Table 3 Examples how start-ups benefit from open data

Although the success patterns analysis will only be discussed in Section 5, we will already now share some insights related to the usage of open data across the six identified cluster.

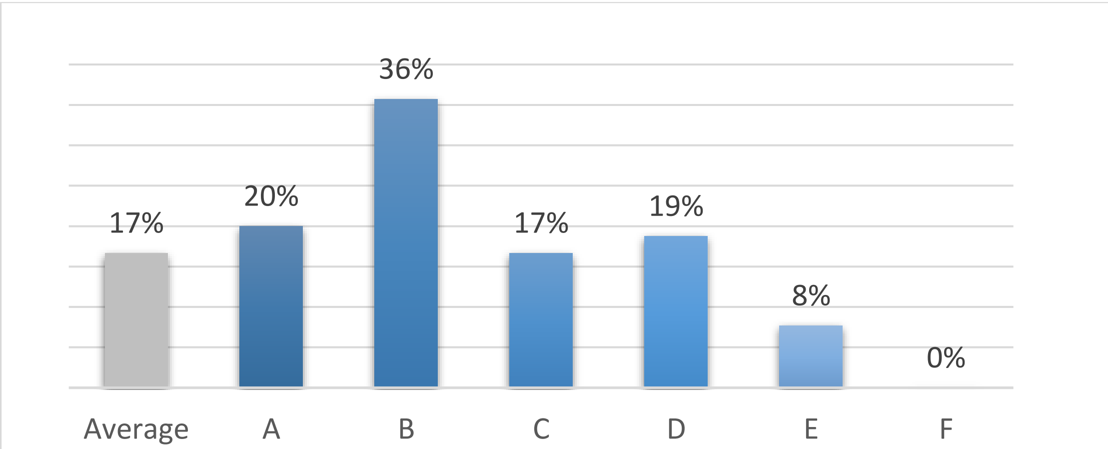


Figure 10 Open Data Usage across the six clusters

First finding is that we did not have any examples of start-ups in the *Connecting Peers Cluster* (F) (see Section 0) that are making use of open data. The main data-driven functionality or value provided by Cluster F start-ups is match-making. This is in most cases realized by means of personal profiles or extracted metadata. It seems that open data does not yet bring value to them in improving or coordinating the match-making functionality.

Second, the highest number of start-ups using open data were observed in the cluster *Internet-Of-Things- Applications* (B) that are characterised by fact that they are relying on Internet of Things (IoT) Technology as part of their offering. For instance, open map data as well as weather related open data brings value to their offerings.

4.3.2 Finding: 43% of start-ups using unstructured data in their offering

Another way of differentiating data sources is the extent to which the meaning of the data is represented explicitly in a structured formats. To make meaning of data explicit any type of unstructured data source, such as image, video, text, or audio, relies on very specific pre-processing algorithms and methods to extract initial information bits out of the raw data. In our study 43% percent of start-ups are pre-processing at least one type of unstructured data source. Note that in the cluster analysis (see Section 6), there is one group of start-ups (Cluster A) that rely on their capability of data pre-processing of a specific type unstructured data sources as their core capability.

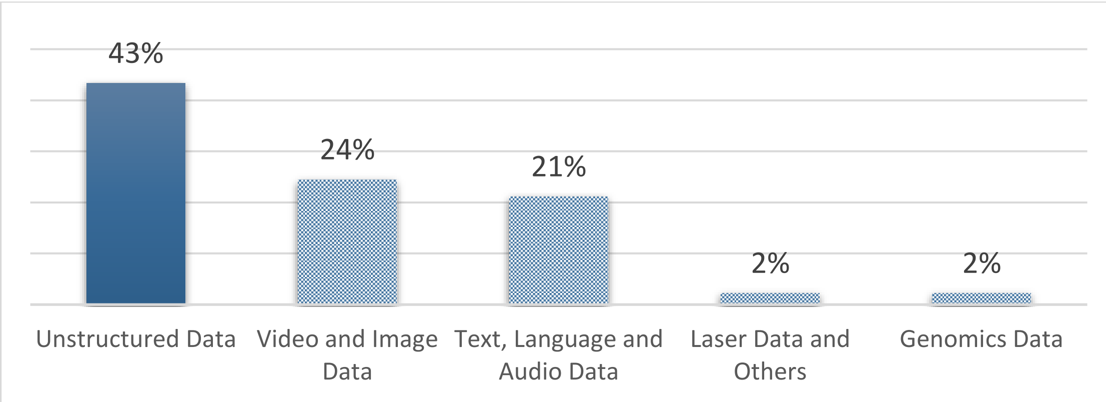


Figure 11 Unstructured Data Sources Usage

There is a wide range of unstructured data types, however image and video as well as text, language and audio data are the most popular ones. One in every fourth company using image or video data in its data-driven offering.

> For instance, the company *OncoraMedical*[35] pre-processes also radiology images to generate a personalized radiation treatment plan.

Similar frequent is the usage of text, language and audio data for implementation data-driven innovation. We observe this in 21% of the start-ups sample.

> For instance, *Warwick Analytics*[36] is offering an AI platform that is able to pre-process unstructured text data in close interaction with an human expert. By iteratively incorporating human's feedback, the provided text mining algorithm can be adapted very efficiently to the requested usage scenario and for maximizing accuracy.

[35] https://oncoramedical.com/

[36] https://warwickanalytics.com/

Laser and others as well as *data genomics* were rarely seen in our data, only two percent each.

> An example of laser usage comes from the start-up *Propeller Aero*[37]. They create tools and software for construction companies, mines, quarries, and landfills to collect, process, and visualize accurate survey data. One of their products is called Lidar and it means “light detection and ranging”. The product is able to send laser pulses at a feature (terrain, trees, or cliff faces) and measures the reflected pulses with a sensor, providing the distance between the features.
>
> In the data *genomics* field the company *Deskgen* is using genetic data to generate value to scientists. They help them to discover, understand and treat the root genetic causes of human diseases. Since 2012, the company has supported thousands of labs around the world with best-in-class CRISPR genome editing design and analysis technology, across basic research, drug discovery, translational research and IND-enabling studies.

4.3.3 Finding: Every 2nd company using unstructured data also rely on semantic technology

We were interested in the question, to which extent methods for the preprocessing of unstructured data are combined with semantic technologies. By using standardized semantics (vocabulary and representation formats) the information entities extracted from the unstructured data sources can be very easily be reused in different contexts. As the pre-processing of unstructured data sources often goes in hand with high efforts, we were interested whether additional efforts fostering the reuse of data was likely to be invested.

As starting point, we know that 43% of start-ups are relying on unstructured data and 33% of startups of our sample on semantic data (see also Figure 9).

In our sample set, every second company relying on unstructured data sets, also used semantic technologies (see Figure 11).

> For example, *iris.ai*[38] support researchers in finding relevant research papers or patents. Their solution is relying on natural language processing technologies to extract key data from massive collection of research data. In order to overcome the ambiguous labelling of research topics methods for clustering semantically similar documents are needed.

This example shows that in case valuable background knowledge is available, its incorporating into the data-driven solution seems to be very reasonable.

[37] https://www.propelleraero.com/

[38] https://iris.ai/

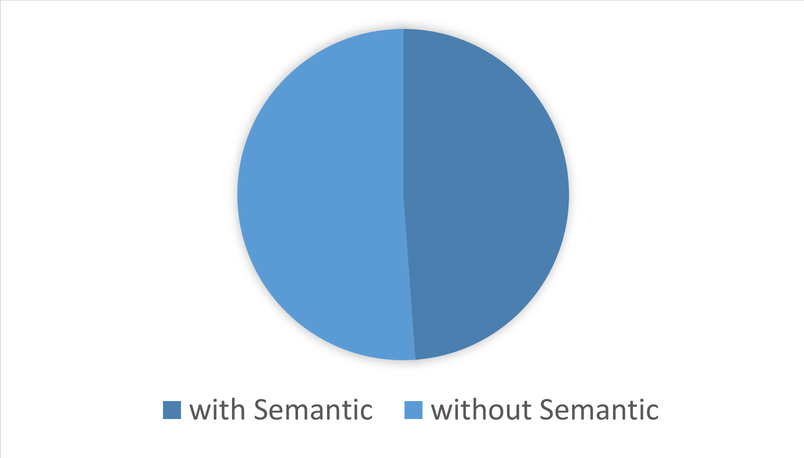


Figure 12 Unstructured Data combined with …

On the other side, similar number of start-ups (22%) using unstructured data did not incorporate explicit semantics (see Figure 12)

> For instance, *Artomatix*[39] is supporting artists and developers of the video gaming industry in generating solution in the faster generating of realistic 3D art textures. Its technology is based on computer graphics, Deep Learning and computer vision and uses generative neuronal networks to "imagine" new details of a texture in a way a human would do, i.e. it recognized objects in a video, can add texture and features automatically by relying on the "learned" knowledge what should be there. As those pre-processing steps do not rely on external background knowledge, the usage of semantic technologies does not improve the performance of the AI-solution.

Startups that are relying on semantics have a tendency to also use unstructured data. In numbers, 57% of start-ups that are using semantics are also relying on unstructured data (see Figure 13).

> For instance, *cortisol.io*[40] is offering a platform for natural language understanding which seamlessly integrates semantic capabilities, such as semantic search, semantic annotation or semantic supercomputing with core NLP modules converting text into semantic fingerprints. In addition, domain specific vocabulary required for semantic and NLP applications can be trained via unsupervised machine learning methods.

[39] https://artomatix.com/

[40] https://www.cortical.io

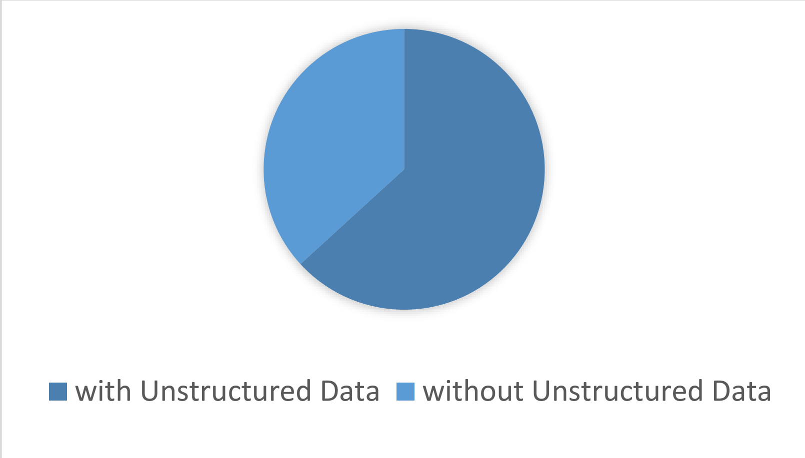


Figure 13 Semantic Data combined with….

This examples clearly demonstrates the high value of combining the two different technology disciplines in one application, like NLP, semantics and machine learning. However, there also exist applications in which semantics data can be applied in isolation.

> For instance, *Dgraph*[41] who is offering an open source, low latency, high throughput, native and distributed graph database are solely addressing the storage of data sets in graph structure, such as semantic knowledge graphs. The graph database is an intermediate asset for building knowledge-based applications.

In order to provide a unique selling provision, intermediate offering tend to have a clear focus on precise technical capabilities. Any integration efforts are out of scope and are likely to be addressed by end-user solution provider.

[41] https://dgraph.io/

4.4 Technology

The BDV Strategic Research and Innovation Agenda (Zillner *et al.*, 2017) is describing five technical priorities identified by the BDVA ecosystem and experts as strategic technical objectives. In our study, we were interested in which of these technical areas were most frequently covered when realizing data driven innovation.

In addition, to boost the adoption of AI and data technologies, it is of importance to ensure that technologies from related horizontal communities, such as High-Performance Computing, Internet-of-Things, Block-chain or Robotics can easily be aligned. As the different communities are putting efforts in aligning high-level reference architectures and interfaces, we have been interested in the question to which extend multiple technologies are used in combination.

4.4.1 Finding: Data Analytics is the most frequent

Figure 14 gives an overview how often each technical priority has been part of the data-driven innovations.

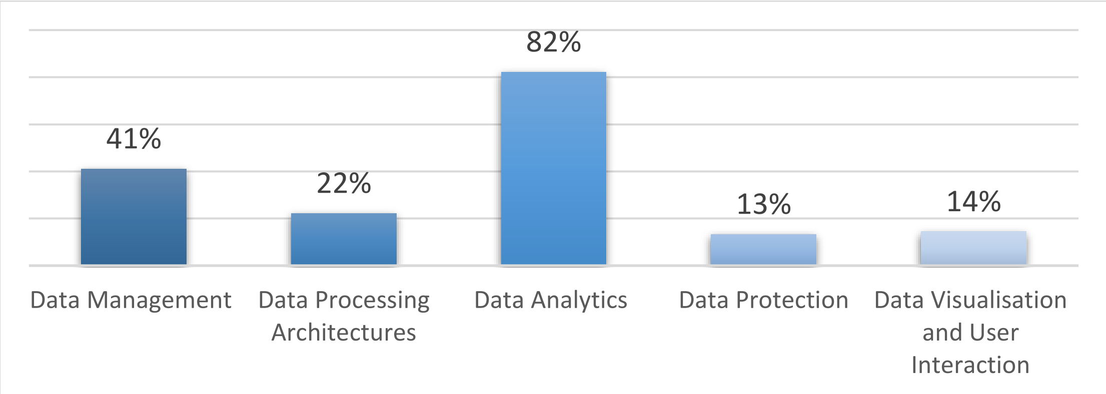


Figure 14 Coverage of the BDV SRIA Technical Priorities

Among the five technologies listed in the Strategic Research and Innovation Agenda (SRIA), *Data Analytics* is used most frequently. 82% of our start-up sample relied on data analytics to implement data-driven value proposition. Among the 18% of companies that do not use data analytics, 64% have a clear sector focus and 63% of them have received more than 3 funding rounds. On the other hand, they do not use more or equal to three technologies or complementary technologies, do not have more than three revenue models, but instead 64% use subscription as a revenue model.

The usage of *Data Management* is seen in 41% of the cases and is very much inline with offerings addressing the challenges of processing unstructured data sources (as discussed in Section 4.3.2). The usage of *Data Processing Architectures* reflects the numbers for IoT technology usage that will be discussed later. *Data Visualisation and User interaction* at 14% is strongly correlated with the usage of media data (11%) discussed before.

> One example is *Wegowise*[42] offering solutions that empower the owners and managers of over 40,000 buildings to understand, track, and improve building efficiency. The software provides timely, insightful, and actionable information to help our customers manage energy and water usage, saving not only resources, but time and money, too.

Own solutions for *Data Protection* are the least frequently addressed research challenge with 13%.

4.4.2 Finding: 59% of startups combine two or more SRIA Technologies

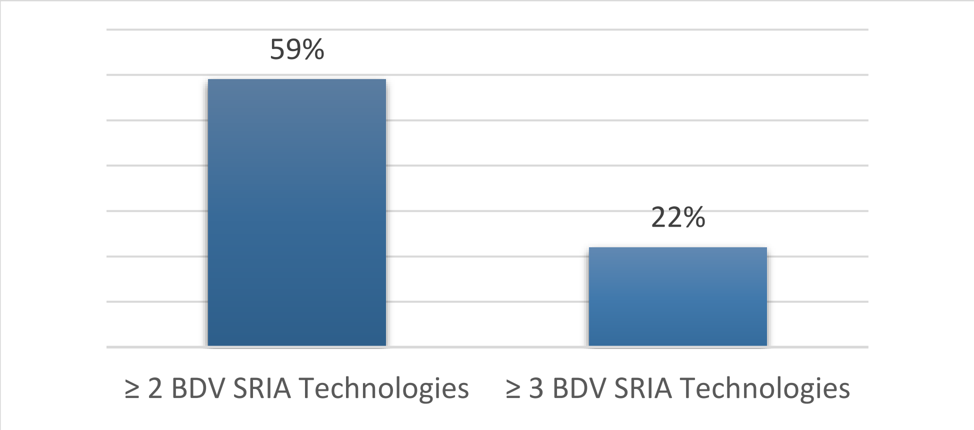


Figure 15 Usage of Multiple BDV SRIA Technologies

When looking at the usage of more or equal to two BDV SRIA Technologies we observed that more than half of the start-ups, precisely 59%, combine more than two technologies (see also Figure 15)

> *Uplevel Security*[43] is one example. They redefine security automation by using graph theory for real-time alert correlation. Their product creates a dynamic security graph (data management) for an organization based on incoming alerts, prior incident investigations and current threat intelligence (data protection). *Uplevel Security* then transforms the ingested data into subgraphs that continuously inform the main security graph. By automatically surfacing relationships, investigations no longer occur in isolation but begin with context

Less frequently observed, 22% of the companies combine more than 3 technologies.

> One example is the medical company *CloudMedx*[44], which started with the aim to make healthcare affordable, accessible and standardized for all patients and doctors. The company uses NLP and proprietary clinical contextual ontologies (data management) and deep learning (data analytics) to extract key clinical concepts from the electronic health records and serve them as insights to physicians and care teams with the goal to improve clinical operations, documentation, and patient care. In addition, *CloudMedx* is presenting the

[42] https://www.wegowise.com/

[43] https://www.uplevelsecurity.com/

[44] https://www.cloudmedxhealth.com/

results to dedicated teams through a user-friendly platform that allows for interactive predictive and prescriptive analytics to assess current metrics and build a path forward with informed decisions.

4.4.3 Finding: Only every fourth start-up uses two or more complementary technologies

When looking at the usage of Complementary Technologies in our study (see also Table 4) , meaning the usage of Blockchain, IoT, Robotics and High-Performance Computing, we generally observed a low presence of these types of technologies. Given that these complementary technologies are emerging in the market, the low frequency might indicate a promising usage for them in the future, in particular as soon the challenges related to the technical integration is available in off-the-shelf solutions.

The aggregated numbers for the usage of complementary technologies are as follows:

26% of the start-ups on average use *more than one complementary technology*. However, this number is strongly biased by the Internet of Things Applications cluster, where the usage of IoT is equal to 100%. Start-ups in the Industrial Services Cluster do not observe at all complementary technologies, and the other Clusters observe it 14% of the time on average.

4% of the start-ups on average observe *at least two types of complementary technologies*. For instance, only two clusters apply at least two types of complementary technologies, and they are the Data Pre-processing cluster and the Internet of Things Applications cluster. This might be a good indicative that these two clusters are the most challenging ones in terms of technological complexity as we will discuss in detail in Section 0.

Only a very small number of start-ups apply *at least three complementary technologies*. On average it is observed in only 1% of the time, and it comes from the Internet of Things Applications Cluster. From our sample, this means that only one company presents this characteristic and this company is *Carfit*[45].

> *Carfit* is a Palo Alto (USA) start-up that created a self-diagnostic and predictive maintenance platform in the connected car space, providing dealers and service providers with customized lead generation. Their solution applies IoT, Robotics and High-Performance Computing as complementary technologies by connecting noise, vibration and harshness knowledge to a sophisticated data analysis process that leads to predictive maintenance.

[45] https://car.fit/

4.5 Network strategy

4.5.1 Finding: 57% of start-ups rely on network effects

For digital and data-driven innovations, network effects are important phenomena to be reflected. A network effect occurs when a product or a service becomes more valuable to its users as more people use it (Shapiro and Varian, 1999). Network effects are also known as demand-side economics of scale and predominately exists in areas where networks are of importance, such as online social networks or online dating sites. A social network or dating site is more appealing to its user, the more users of interest it has. In consequence, harnessing network effects require developing a wider network of users in order to differentiate from competitors. For that reason, the critical mass of user and timing are key success factors in a network economy.

Due to the high impact of the network, competitors starting from 'ground zero' with no users in their network will face difficulties to enter the market successfully. In this discussion we are using the expression "network effect" to highlight the positive feedback (positive network externality[46]) i.e. the phenomena that already existing strength or weaknesses are reinforced, might lead to extreme outcomes. In the most extreme form, positive feedback can lead to a winner-takes-it-call market (e.g. Google).

As network effects impact the underlying economics and operation of data-driven innovation. Instead of producing products that are early on the market and differentiate from other offerings, the focus is now on scaling and scoping the demand perspective. Understanding network effects and its underlying market dynamics is crucial to position data-driven products, services and businesses well in the market. For doing so, data-driven-innovation can harness network effects on three different levels.

On data level: Data-driven businesses that can improve their offering the more data is available are relying on network effects on data level. Typical examples are navigation systems or recommendation engines which both become better the more – in this case behavioural data from users – can be collected and fed into the algorithm to produce more accurate traffic information or to better relate similar products with each other. In case data-driven offerings are based on network effects on data level, one needs to develop strategies of how to engage with customers, users or stakeholder to get access to the required data assets. In general, we can distinguish two strategies:

- **Increasing scale of data:** In cases when the outcome of the underlying algorithm can be improved by accumulating more data assets, the strategy is to attract more users in a way that more data can be collected. For

[46] For completeness we want also to mention the phenomena of negative network externalities which occurs when more users make a product less valuable (e.g. traffic congestion). Negative network effects are also referred as "congestion".

understanding how much data is needed, a detailed look at the algorithm is required. Some algorithm cannot be further improved after reaching a sufficiently large amount of data while other algorithms keep on improving continuously the more data is available without limit. In case of the former, the focus is to establish a critical amount of data traffic to ensure that the most accurate analytical model can be built. In the second case the objective of increasing the user base for building larger data assets will happen continuously. In both cases, the starting phase is the trickiest one, as one needs to attract users with a non-mature offering. In this phase a clear understanding of users' needs and interests will help building a promising strategy to overcome the classical chicken-egg problem. With increasing numbers of users, the efforts in attracting users will decrease and in some cases disappear.

- **Increasing scope of data:** The outcome of some algorithm can be improved the more diverging aspects / perspectives the incoming data set is representing. For instance, linking data from different sources establishes the basis to generate "super additive" insights (big brother phenomena) as more aspects of a subject of interest are reflected. This is also known as process of data contextualization allowing to significantly increase the quality of input data leading to better and more transparent outcomes. In such cases, the underlying question of how to increase the range of collected data from the already existing user bases needs to be investigated. In times when data regulations have not been so strict as today in Europe, this was a very impactful strategy of the GAFA (Google, Apple, Facebook, Amazon) organizations as they based their customer analytical insights on data from different types of services.

Example

The platform Insights (see Figure 16) of the company *Apptopia*[47] uses Big data technology to collect, measure, analyze and provide user engagement statistics for mobile apps.

[47] https://apptopia.com/

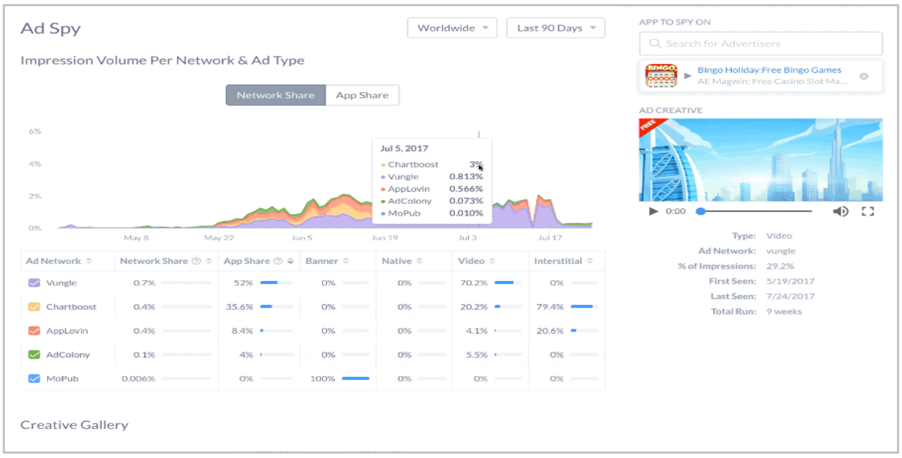


Figure 16 Screenshot of Insight Platform

The more app providers are producing data being connected to the platform, the more valuable the service gets. In order to get more real-time data, they attract app developers to connect to their platform by providing free data analytics products. With this free of charge value proposition, developers benefit in registering their mobile apps to the platform while giving the platform the permission to analyze user engagement data of the mobile app. High-priced subscription fee model for business customer, including Google, Pinterest, Facebook, NBC Universal, Deloitte, and others, benefiting from real-time engagement insights of mobile apps complements the revenue strategy of this offering.

In this context, multi-sided business models are the usual way forward. Per definition, a multi-sided business model brings together two or more distinct but interdependent groups of customers. Value is only created if all groups are attracted and addressed simultaneously. The intermediary, in our example the company *Apptopia*, generates value by facilitating interactions between the different customer groups, whereas the value increases when more users are attracted. The more app developer register on the platform the more accurate the statistics become and with an increasing number of business customer, *Apptopia* has the required resources to invest into advanced functionalities for app developers.

In our study every second start-up relied on network effects on data-level. This highlights the high importance of access to data sources in the context of data-driven innovations. For that reason, we see a sustainable strategy to get access to data as important success lever for data-driven innovations.

On infrastructure level: Data-driven businesses that harness network effects on the infrastructure level provide a technical foundation for others, i.e. third-party companies, to build upon. Based on a layer of common components, third-party players are invited to develop and produce an increasing number of data-driven offerings. This set-up is also known as product platforms (Hagel *et al.*, 2015), a prominent example is the android platforms that provides the technical foundation for others to build apps. This includes any types of tools and services that enable the plug-and-play building of a data-driven offerings, e.g. (open) standards, de-facto standards, APIs, standardized data models, etc. The more functionalities are available that help others to build and position innovative offerings better, faster, etc., the more attractive the offering itself becomes. The infrastructure layer itself has little value per se unless other users and partners create value on top of it.

For instance, the agricultural-robotics technology company *Skyx*[48] is neither offering hardware nor agriculture end-customer applications, but a software that enables a modular swarm of autonomous drones for spraying. By providing a technology to plan and control the mission of drones in real-time as well as to auto-pilot the entire fleet /swarm, it addresses the need of agri-spraying application developer applicators in building their solutions in higher quality and less cost by relying on an standardized approach. In addition, as the software is compatible with any commercially available hardware, cost in connecting the wide range of drones can significantly be reduced. Thus, Skyx provides tools and connectors for agri-spraying application developer to build their own solutions. The more drone hardware can be connected, and the more spraying functionalities can be provided, the more attractive is the overall offering for applicators.

Other examples that fall under this category are anonymization services, data quality services or data acquisition services. In general, network effects on infrastructure are more likely to happen in markets ((Hagel *et al.*, 2015):

- that benefit from tightly integrated and standardized products technology component that can be easily be connected and combined (e.g. due high technical complexity, high dynamics of development);
- that benefit from third parties to contribute important assets, such as hardware, data, etc.
- where third parties / customer benefit from the opportunity to implement their own sector-specific value proposition.

A critical success criterion for data-driven innovations that harness network effects on infrastructure level, is fair and balanced value creation and capturing. Third parties connecting the offering as technology providers or as application developers need to be able to generate significant value for themselves while innovators themselves need to yield strong returns at the same time. Thus, innovations with network effects on infrastructure level always impact the underlying economics and operations. As highlighted before, the scale and scope of the data-driven innovation is central to stimulate its demand side. In this context, open source software, de-facto standards, open APIs, and standardized data models are important levers (Choudary, 2015).

In our study, only 4% of data-driven start-ups focused on network effects on infrastructure level. This low percentage might be caused by the deep understanding and expertise required to exploit network effects on infrastructure level. As the focus is to identify common denominator/core components required for building a larger set of offerings, deep knowledge in the similarities and difference of the various offerings is needed.

[48] https://www.skyx.solutions/

On marketplace level: In cases where the number of marketplace participants is the key source of value, data-driven innovation offering network effects on marketplace level are central. Offerings that connect its participants in their specific roles, such as buyer and seller, consumer and producer, etc. allow that two participants can easily interact with each other. Two aspects increase the attractiveness of marketplaces:

- The number, quality and type of participants connected to the marketplace. In this context, balanced growth as well as critical mass of both participant types is required to ensure that promising counterparts can be found.
- The platform benefit from intelligent forms of matchmaking ensuring that the mapping of partners ensures nice fit of interest.

The purpose of connecting different participants via a marketplace is to exchange "something", i.e. goods and /or services. The asset being exchanged is also called core value unit and determines the design of the platform. In this context, three different types of core value units can be distinguished:

- Goods, in those cases the marketplace enables the exchange of physical products that can be described along the product categories and price expectations. Prominent examples are ebay or etsy. The matchmaking algorithm is based on facetted search based on a set of well-defined product categories.
- Standardized services are promoted as "off-the-shelf" offerings without means for customization. Typical examples are rides on Uber or boat rentals by Zizzo[49]. By standardizing the access to the service of interest, provider and consumer can be mapped automatically. The description of the service is used as input for the matching algorithm.
- Marketplaces offering non-standardized services, such as dating platforms, rely on the description of the service provider (also known as user profile). This description is then core value unit determining the matchmaking algorithm.

In general, in accordance to the core value unit different types of matchmaking algorithm are build. For instance, the start-up Selectionnist[50] is based on an advanced match-making service that connects consumers and brands by means of image recognition technology. Thus, consumer that locate a product of interest in a magazine can use a mobile app to snap a picture of the product which automatically forwards the customer to the product's online shop.

Network effects are built into the definition of a marketplace, i.e. all successful marketplaces benefit from network effects. This was also seen in our study.

The low number of marketplaces in our study, indicates the high challenges of building them. The challenges are less on the technical level but more on the level of building critical size and balanced user communities. Several strategies to attract users from the different communities have been implemented by the start-ups:

- **Complementary/additional value propositions offered to engage provider**: For instance *Zizzo*, the marketplace for renting boats, addressed the shortage of boat charter companies by providing them a powerful inventory boat

[49] https://www.zizoo.com/

[50] http://www.selectionnist.com/

management tool encouraging them to digitalize their processes which again was a necessary precondition for being able to join the marketplace.

- **Building of participants communities:** the founder of *Insurify*[51], the first online car insurance shopping platform *Insurify*[52] (TripAdvisor for car insurances), invested several years in building relationships with car insurance companies.
- **Provisioning of incentives to attract consumer**: The Hooch[53] hospitality app reward subscribed members with a voucher for one free drink every day, which helps cocktail bars to gain new customers.

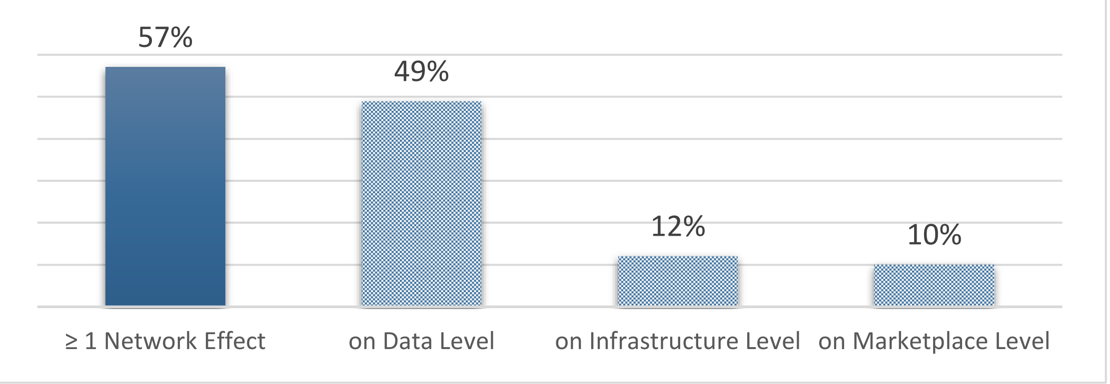


Figure 17 57% of start-ups harness network effects

To sum up the discussion about of network effects, we can state that it occurs most frequently on data level. This highlights the importance and impact of getting access to high quality data. In addition, we could observe that for businesses being based on network effects multi-sided business models are very likely to be present.

4.6 Revenue Model

4.6.1 Finding: Data-driven innovation relies on multiple revenue models

We have been interested in the question of how data-driven businesses are making money. Is this different from traditional businesses? And can we identify some dominant revenue models?

Our first finding is that often information about the type of revenue models used was difficult to find. Especially in cases when start-ups have been focusing on emerging technical advances, such as drones or autonomous driving, information about revenue models was not available. As emerging technology businesses are often seen as investment or bet on the future in a market not yet established, the absence of revenue-related information is not surprising. For 10 % of the companies analysed this was the case, i.e. no information about the revenue model could be found or inferred.

[51] https://insurify.com/

[52] https://insurify.com/

[53] https://hooch.co/

Our study confirmed the finding of (Attenberger, 2016) that revenue models have not changed through the usage of data technologies. The major difference to traditional businesses is that data-driven innovations rely on different types and combination of revenue streams that are continuously changing over time in order to address the specific user needs of each customer segment. On the one hand, we observe very *new forms of value propositions*, ranging from service offerings, the bundling and unbundling of offerings, intermediate offerings, product differentiations through versioning, etc. that allow to address the very specific user needs. On the other hand, the majority of data-driven innovations has – in comparison to traditional businesses – a different cost structure. With data and data offerings being cheap to reproduce and deliver, the typical cost structure of data-driven innovations relies on fixed costs for the development of the offerings but low variable cost. This kind of cost structure leads to substantial economics of scale as with more offerings sold, the lower the average costs of development become. In addition, as the reproduction and distribution costs are often marginal, the danger of price dumping and surplus of offerings in the competitive market is a frequent phenomenon. For instance, (Aitken and Gauntlett, 2013) counted more than 40,000 health apps in the app store being offered for free or for a very small price.

With this new cost structure for most of data-driven innovation, organisation have a new flexibility to *adjust the equation between value proposition and price* in accordance to the user needs of the various customer segments. In this context, companies elaborate the specific price level the targeted user group is willing to pay. The main objective for aligning the product version with pricing version for each customer segment is to attract more users and interactions as well as grow the community.

Figure 18 Revenue model of Bird.i (freemium model not explicitly described)

On average, data-driven companies in our study have 2.6 revenue models. The highest number of revenue models we found with the start-ups in our sample was 5. For instance the company bird.i[54] is using five different revenue models (see Figure 18) to attract different customer segments and to ensure market growth. The company offers satellite imagery available for everybody, i.e. to all different customer segments ranging from individual to business. Through their partnerships with a wide range of satellite image providers, they offer more data in comparison to single-provider data sources. In addition, they facilitate access to the images by enhancing the original data with valuable geo-referenced information. The first revenue model is a combination of freemium and advertisement (not explicitly listed in Figure 18 as for this revenue model no contract is needed). The other revenue models range from "pay-as-you-go" to three different types of subscription models. The first one is addressing individuals and restricts usage of images for private and public applications whereas the second and third subscription models is addressing commercial usage with moderate and extensive usage. This examples also shows how the value offerings differ in in each subscription category.

In addition, due to the high market dynamics and impact of network effects, revenue models are likely to change over time due to the market dynamics as well as their own strategies to actively harness network effects.

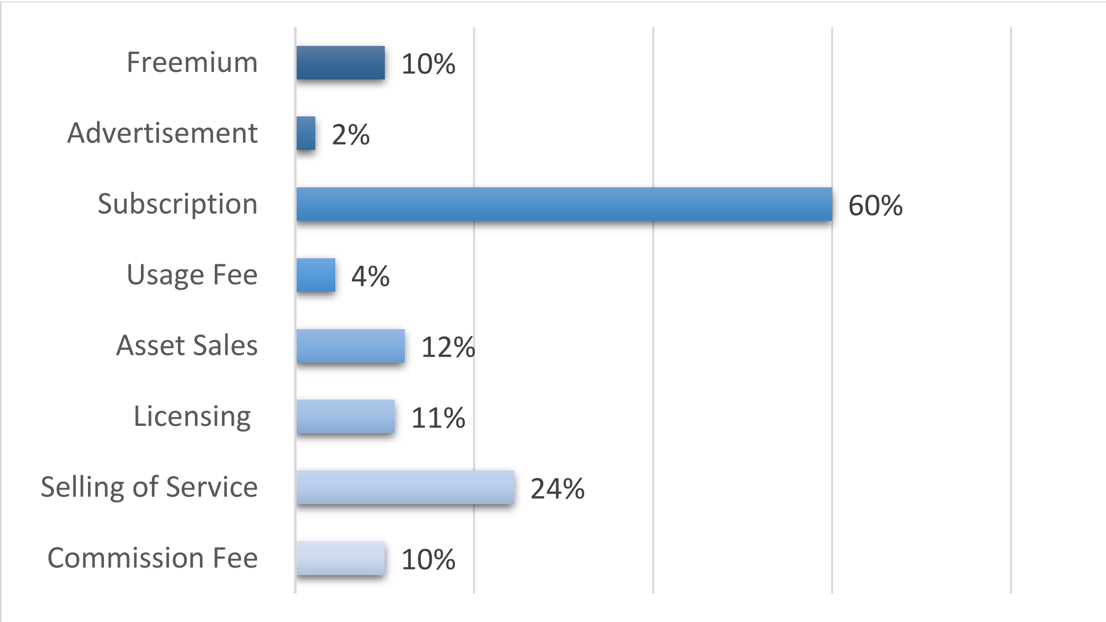


Figure 19 Overview of revenue models used for data-driven businesses

The most frequently used subscription model in our study was the subscription model. Due to big spread and high adoption of Software as a service (SaaS) approach which describes a software licensing and delivery model in which software is licensed on a subscription basis and is centrally hosted[55]. The deployment of data-driven innovation

[54] https://hibirdi.com/

[55] see https://en.wikipedia.org/wiki/Software_as_a_service

as SaaS-based offering brings a lot of flexibility. For instance, end-user do not require to install the software on their own system, they are relieved from the burden of operating the application themselves and they can access the software on-demand.

The second most frequent revenue model is the selling of services in which the person's time is payed for. Those revenue models are very often used for open software offerings as well as when offerings are not standardized or off-shelf.

Advertisement as a revenue model is rarely observed. In our sample, only 2% of the start-ups are applying it and although might seem surprising, it reflects the high percentage of B2B models. The two cases seen in our sample are both in the data-driven platform business, matching supply and demand.

> The first company to offer advertisement as a revenue model is *Influenster*. "A product discovery and reviews platform that enables consumers to find new products and get advice to make informed purchases" says the website. As a "search engine" website there is space for advertisement as the product itself might be a hard one to price. We can think about it as the search page of google and their advertisement strategy, it is a very similar logic.
>
> The second company to offer advertisement as a revenue model is *Bird.i* (which was already discussed before)

4.7 Type of Business

Readers Note: this dimension changed throughout the analysis of start-ups of our sample sets. In the Deliverable D2.6 we still used the "Value network strategy" which was updated in accordance to our findings. For that reason, the below explanations are more detailed than the explanation of the predeceasing dimensions.

Data-driven innovations can disrupt existing value chain. However, at the same time, we observe a large amount of "low hanging fruits", i.e. business opportunities in the scope of established processes (intern) or value chains (cross-organisational). In order to classify, the data-driven business opportunities we introduce some categories to distinguish their impact on the market and the value chain.
In order to come up with a sophisticated classification of data-driven business opportunities, we analysed existing approaches already available for the classification of traditional business opportunities. One important work in this context is (Ardichvili et al. 2003) who are classifying business opportunities into two dimensions *value creation capability* and *value sought.* Although both dimensions have on the first sight a good mapping to the DDI supply and demand side, the did not reflect the changing nature of underlying business ecosystems. As already discussed in the beginning of this deliverable are data-driven innovations are very rarely developed alone but rely on the collaboration between many partners in the value chain.

In other words, when positioning data-driven offerings on the market, one requires to reflect the associated business or innovation ecosystem. In order to address this important characteristic of data-driven innovation, we are classifying the type of business of data-driven innovation as follows. It aims to complement existing classification of business opportunities by addressing also the question to which degree one requires to reflect and incorporate the logic of the underlying business and innovation ecosystem. We have identified five different high-level categories of Type-of-Businesses, meaning how offerings are positioned on the market.

Improved Product with the strategy **to improve existing Product**. This strategy is only applicable for established companies that have a portfolio of products/offerings/services on the market. For that reason it does not appear in the statistics of Figure 20 . Existing offerings are enhanced by an additional data-driven functionality. For instance, gas turbine are complemented by an might predictive mainenance service. Such product improvements have implications how the product is postioned on the market. In consequence, business models, sales channels and organisation processes and structures need to be changed. For instance, need to be adapted require new revenue models(instead of selling the asset (e.g. an turbine, a predictive maintenance services is positioned on the market). This again implies changes in the business model revenue strategy as well as changes in the sales channel. For instance, the sales of an engine is a one-time sale while the maintenance services of an engine is likely to be offered on a subscribtion level.

4.7.1 Finding: Data-Driven Services are the most frequent Type-of Business

Figure 20 gives an overview of the frequency all all type-of businesses that have been of relevance for start-ups.

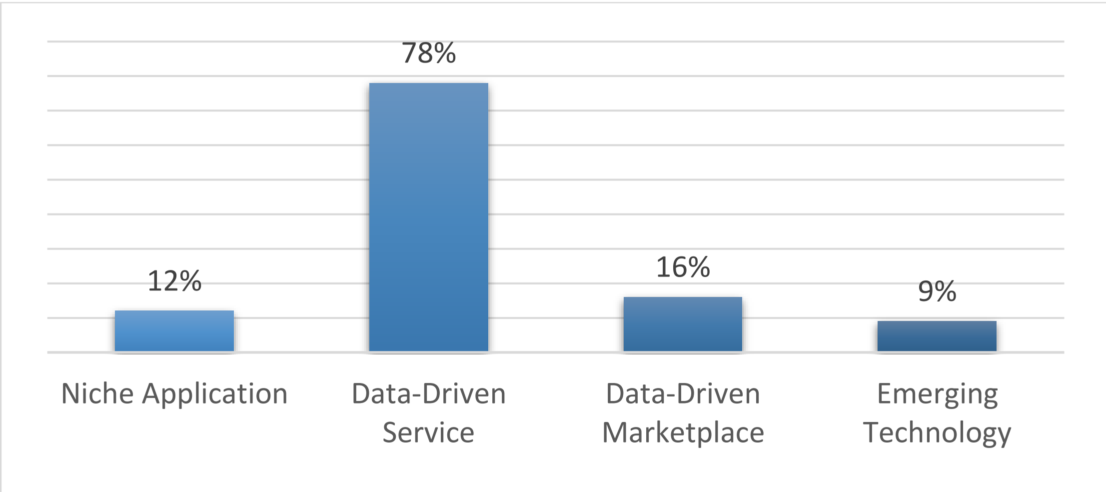


Figure 20 Type of Business

Data-driven Service with the strategy of *"Finding a new business partner"***.** This strategy here is to focus on one single customer and his or her business processes. Bases on a detailed understanding of his / her business processes including the pain points, happiness points and unaddressed user needs, new services for specific user needs are build. As the service is very focused on this one specific partner, the overall market and business ecosystem is only observed in an indirect manner.

In our study, the data-driven service business was with 78% the most frequently observed approach to position offerings on the market.

> For instance, *Arable* provides is agricultures BI solution based on in field-measurements as Software-as-a-Service (SaaS)-based service offering. In order to enable grower, advisors and businesses to play an proactive role in ensuring high quality and longevity of their agricultural operations, they can derive real-time, actionable monitoring and prediction related to weather risk and crop health by means of an tiered SaaS-offering with different level of services combined with IoT businesses. (The tier I service includes reporting, integrating and visualisation whereas the tier II services include predictions and advanced analytics).

When positioning data-driven innovation as service business on the market, only one target user group need to be considered when scoping the functionality of offering. We assume that due to the limited complexity, market opportunities are easier and faster to be addressed. This is also reflected in the high frequency of its usage in our sample set.

4.7.2 Finding: Data-driven Marketplaces are complex to build and less frequent

In comparison to the data-driven service business, is the development of **data-driven marketplaces** significantly more complex. Only 16% of companies in our sample relied on this approach. To make the new offering successful *a new market place / ecosystem need to be built.* Market participants on the supply as well as on the demand side, need to be attracted. In addition, it needs to be ensured that a critical number of participants are providing their assets and at the same time a critical number of participants is requesting it. In addition, the growth of the marketplace needs to be balanced on both, the supply and demand, side, in order to keep it attractiveness. Organisations have developed very different strategies to attract the different participants group, e.g. by providing needed IT services, analytics services, offering services for free, etc.

> For instance, the *zizoo*[56] established a global boat rental platform. They are building a global digital booking platform and website connecting suppliers (charter companies) to travelers worldwide, similar to "Booking.com for Boats". When building up this marketplace, the founder of the company where entering a market (the boat rental market) which was 10 years behind any other travel sectors. As the majority of boat charter companies, have not yet been digitalized, they needed to put a lot of effort in attracting the supply side to join their emerging marketplace. For instance, they are offering charter companies with a powerful inventory management tool and business intelligence for free. As they are making boat holidays affordable and accessible to everyone (bookings start from €20 a day), they were also able to attract the demand side.

At the centre of data-driven marketplaces is the match-making functionalities between supply and demand. This implies reaching out to a critical number of

[56] https://www.zizoo.com/

participants on both supply and demand side as well as finding an intelligent way to find matches.

Providing a match-making service does not automatically imply, that companies are building a marketplace. In particular, in situation when other players are having a better access to one or both target groups it is more reasonable to position the match-making offering as service for the companies with access to the target groups.

> For instance, the company Keywee[57] (see also Section 6.6) is providing match-making service that allow publishers and brand to find the audience for their articles. The technical basis is a natural language processing pipeline that allows extract key-terms per article to pull in readers in accordance to their profile's information. Keywee is offering this as a service to publisher who are already running online content sites attracting large number of visitors. Keywee receive commission fee, whenever new readers could be pulled in. Thus, they support publishers in the match-making between potentially interested readers and the content offering but do not establish a content marketplace themselves.

Another strategy is to identify an existing healthy ecosystem in place which gives the opportunity to position the own offering as niche application. The so-called **Nice Player** *leverage an existing ecosystem* by scoping a niche offering in accordance to the defined constraints of the dominant or key stone player of the ecosystem. Typical examples for such strategies are the thousands of apps offered in the I-phone or android ecosystem for mobiles. In our sample it was observed in 12% of the cases.

> One example is *AIMS Innovation*[58]. They develop AI and machine learning technology to give the world's largest companies deep insight and control into their most business-critical processes – like safely distributing electricity, shipping thousands of daily orders to ecommerce customers and delivering the results of medical tests to doctors quickly and reliably. They are positioning their offering in the Microsoft ecosystem. According to their website, they are the only Artificial Intelligence in IT Operations solution covering all core Microsoft enterprise technologies and gives deep business insight and control over critical business processes.

The last type of business categories is the so-called **Emerging technology** business that anticipates a future ecosystem or market. In our study it was seen in 9% of the case. As the market is not yet settled and the technology often in very early stages, they are scoped as investment into the future. Thus, revenue strategies cannot be implemented. The main focus of emerging technology businesses is building capabilities /assets ensuring a future competitive advantage.

For instance, the company *Carfit*[59] are working on creating the most comprehensive library on car's vibrations. Data related to noise, vibration or harshness is collected and generated and enhanced data analytics algorithm by incorporating automotive domain expertise. They are aiming car vibrations tracking device that can help lower

[57] https://keywee.co/

[58] https://www.aims.ai/

[59] https://car.fit/

car maintenance costs, increase efficiency and transparency about the car's operation. The self-diagnostic and predictive maintenance platform only brings real value to the end users when the vehicles are moving autonomously. They are aiming to address this future market, as today's drivers are in general very good in detecting anormal noises in the car. However, when cars are moving autonomously the need for remote monitoring will become very critical.

5 Findings from Success Pattern Analysis

This section discusses the detailed finding of our pattern analysis including the clusters derived. Table 4 provides an overview of all percentages per cluster and DDI dimensions / variables. In the following we will describe the indicative insights per cluster with concrete examples from the sample set.

DDI Dimensions					Avg.	A	B	C	D	E	F
					DDI Cluster Types Percentages						
General	≥ 3 funding rounds				59%	60%	57%	33%	81%	62%	53%
Value Proposition	Sector-agnostic				24%	45%	14%	33%	31%	15%	0%
Value Proposition	Target Customer	B2B			98%	95%	100%	100%	100%	92%	100%
Value Proposition	Target Customer	B2C			21%	20%	14%	0%	6%	8%	73%
Value Proposition	Target Customer	B2B & B2C			19%	15%	14%	0%	6%	0%	73%
Value Proposition	Offering	Data Value	Insight Generation	Analytics	66%	15%	93%	100%	100%	100%	13%
Value Proposition	Offering	Data Value	Insight Generation	Descriptive	54%	0%	86%	92%	100%	69%	7%
Value Proposition	Offering	Data Value	Insight Generation	Diagnostic	26%	10%	64%	58%	13%	23%	0%
Value Proposition	Offering	Data Value	Insight Generation	Predictive	38%	0%	86%	67%	0%	100%	7%
Value Proposition	Offering	Data Value	Insight Generation	Prescriptive	12%	5%	29%	25%	6%	15%	0%
Value Proposition	Offering	Data Value	Orchestration	Match-Making	16%	0%	0%	0%	0%	8%	87%
Value Proposition	Offering	Data Value	Automation	Process automation	21%	15%	29%	50%	13%	23%	7%
Value Proposition	Offering	Hardware			19%	30%	29%	8%	19%	23%	0%
Data	Time and Space Dimension	Time-Series / Temporal Data			56%	25%	100%	50%	88%	54%	27%
Data	Time and Space Dimension	Geo-spatial Data			43%	25%	100%	42%	44%	38%	20%
Data	Complex Data	IoT Data			30%	15%	100%	17%	25%	23%	7%
Data	Complex Data	Media Data			11%	5%	0%	0%	38%	8%	13%
Data	Data Type	Unstructured	Unstructured Data*		43%	60%	21%	25%	44%	62%	40%
Data	Data Type	Unstructured	Video and Image		24%	45%	21%	8%	19%	23%	20%
Data	Data Type	Unstructured	Text, Language and Audio		21%	25%	0%	25%	31%	23%	20%
Data	Data Type	Unstructured	Genomics		2%	0%	0%	0%	0%	15%	0%
Data	Data Type	Unstructured	Laser and others		2%	5%	0%	0%	6%	0%	0%
Data	Data Type	Semi-Structured Data			41%	30%	7%	17%	75%	38%	73%
Data	Data Type	Semantic Data			33%	40%	0%	50%	31%	15%	60%
Data	Origin or Data	Personal Data			62%	40%	57%	50%	63%	85%	87%
Data	Origin or Data	Industrial Data			30%	25%	71%	100%	0%	0%	0%
Data	Origin or Data	Open Data			17%	20%	36%	17%	19%	8%	0%
Data	Origin or Data	Research Data			2%	5%	0%	0%	0%	8%	0%
Tec hn	BD V SRI A	Data Management			41%	75%	43%	25%	31%	46%	33%

		Data Processing Architectures	22%	15%	86%	8%	19%	8%	0%
		Data Analytics	82%	70%	100%	75%	81%	100%	73%
		Data Protection	13%	5%	36%	25%	6%	8%	7%
		Data Visualization and User Interaction	14%	20%	21%	33%	13%	0%	0%
		≥ 1 SRIA Technology*	100%	100%	100%	100%	100%	100%	100%
		≥ 2 SRIA Technologies*	59%	65%	93%	58%	50%	54%	33%
		≥ 3 SRIA Technologies*	22%	20%	79%	17%	13%	8%	0%
	BDV Complementary Technologies	Blockchain	1%	0%	7%	0%	0%	0%	0%
		Internet of Things	20%	0%	100%	0%	19%	0%	7%
		Robotics	6%	5%	14%	0%	0%	15%	0%
		High Performance Computing	3%	10%	7%	0%	0%	0%	0%
		≥ 1 Complementary Technology*	26%	15%	100%	0%	19%	15%	7%
		≥ 2 Complementary Technologies*	4%	5%	21%	0%	0%	0%	0%
		≥ 3 Complementary Technologies*	1%	0%	7%	0%	0%	0%	0%
	≥ 3 SRIA & Complementary Technologies*		29%	20%	93%	17%	31%	15%	0%
Network Strategy	At least one Network Effect*		57%	35%	64%	50%	56%	46%	93%
	on Data Level		49%	30%	64%	42%	44%	31%	87%
	on Infrastructure Level		12%	10%	7%	17%	19%	15%	7%
	on Marketplace Level		10%	0%	7%	0%	0%	0%	60%
Revenue Strategy	Freemium		10%	5%	36%	8%	0%	0%	13%
	Advertisement		2%	0%	0%	0%	0%	0%	13%
	Subscription		60%	50%	71%	50%	94%	46%	53%
	Usage Fee		4%	10%	7%	8%	0%	0%	0%
	Asset Sales		12%	20%	14%	0%	6%	23%	7%
	Licensing		11%	25%	7%	0%	0%	23%	7%
	Selling of Service		24%	20%	29%	25%	50%	23%	7%
	Commission Fee		10%	0%	0%	0%	0%	0%	60%
	No information provided		10%	15%	7%	33%	0%	8%	0%
	More or equal to three different Revenue Models*		22%	25%	36%	25%	13%	8%	27%
Type of Business	Data-driven Service		78%	75%	93%	92%	100%	92%	20%
	Data-driven Marketplace		16%	0%	14%	0%	0%	0%	80%
	Niche Player		12%	20%	21%	8%	6%	8%	7%
	Emerging Technology		9%	15%	14%	0%	6%	8%	7%
*Variables derived by aggregating related ones									

Table 4 Description of Cluster

5.1 Start-ups in Cluster A (Data Pre-Processing) …

Start-ups in the Data Pre-processing cluster are sector agnostic in nearly half percent of the cases, the average for our sample was 24%. This means the companies have a generic solution and probably support others to develop their solution.

> *Cortical.io*[60] is an example of start-up offering Natural Language Understanding solutions based on Semantic Folding. Their two core modules are the *Retina Engine*, a basic NLU module that converts text into semantic fingerprints and performs operations like text segmentation and keyword extraction and the *Retina Database* encompassing domain-specific vocabulary that can be trained via unsupervised machine learning for any business domain. The two application modules *Semantic search* and *Semantic annotation* are augmenting the two core modules. Based on this comprehensive text processing architecture, *Cortical.io* has delivered solutions to a wide range of sectors, ranging from marketing, mobility and logistics, to insurance & financial services.

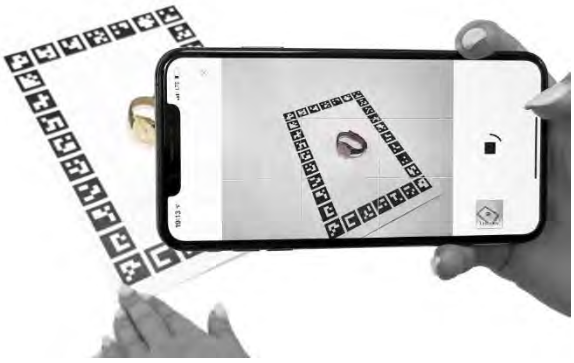
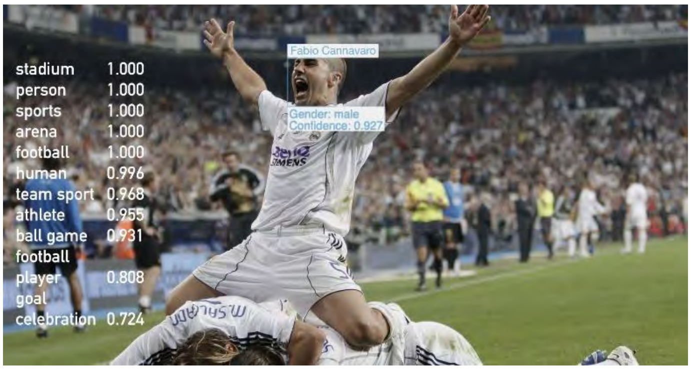
5.1.1 … address technical challenging pre-processing tasks

The pre-processing of unstructured data sources requires to solve wide range of technical challenges. The particularity and characteristics of data source determine the underlying complexity. As there are different types of unstructured data sources, with each type bringing forward very specific pre-processing challenges, offerings focusing on its data pre-processing are often unique and competitive by design.

We observe that offerings in this cluster abstain from adding more features (wider focus) but focus on solving very specific challenges (deeper focus) which again helps to bring forward a unique value proposition. In Table 5 we list some examples of data-driven start-ups that bring forward very specific and unique value propositions.

[60] https://www.cortical.io/

Startup	Value Proposition
Artomatix *www.artomatix.com*	Artificial Intelligence applied to art creation. Through different visual sources, such as photos and scans, the software can generate or improve realistic 3D art. The ability to colour match and then intelligently apply to a target set of textures is illustrated in the image below.  *Image Source: https://artomatix.com/3d-artist-showcase/alex-patel/#&gid=1&pid=1*
TRX System *www.trxsystems.com*	Indoor location software that uses sensor data, Bluetooth, Wi-Fi and inferred map in an Android device to track and map personnel operating indoors. The solution eliminates manual check-in at point tests, eliminates data recording error and provides a whole set of geo-location data in 3D format. The image below illustrates how all the captured information is visualized on an Android phone.  *Image Source: https://www.trxsystems.com/signal-mapper.html*

Cappasity *www.cappasity.com*	3D product imaging solution for websites, mobile apps, VR and AR applications. It provides an interactive in-store browsing experience, improving conversion as the customers can examine the product like they would in real life. The image below shows how the process of 3D image generation starts using the Cappasity app.  *Image source: https://cappasity.com/cappasity-app*
Valossa *www.valossa.com*	Video artificial intelligence that among many possibilities can detect faces, inappropriate content, driver & passenger recognition, and identify soccer player numbers to produce highlights based on metadata. The image below is a screenshot of how the software works: the numbers on the left are the degree of confidence for a specific feature. For example, the software identified with 100% confidence that the image is at a stadium, that there is a person and is related to sports.  *Image source: https://valossa.com/solutions/*

Iris Automation *www.irisonboard.com/*	Detect-and-avoid solution that enables Beyond Visual Line of Sight (BVLOS) operations for autonomous vehicles and Unmanned Aircraft Systems (UAS), truly understanding the aviation environment around it as if a pilot were on board. It can be used in cases such as collision avoidance, such as in the image below, pipeline and railway inspection.  *Image source: https://www.irisonboard.com/*
SchoolMint *www.schoolmint.com*	School enrolment and student behaviour management that help both schools and families to find, apply and re-register to the most suitable school. By turning all school data into one website, it is possible to get useful data such as how family search for schools, it also improves participation by educating families about the neighbourhood school and grow community engagement. Below is a screenshot on how it looks like to find schools based on the grade or location in the SchoolMint´s solution.  *Available at: www.schoolmint.com*

Table 5 Examples of unique value proposition of start-ups in the Pre-Processing Cluster

5.1.2 … are likely to rely on unstructured data

The above observations are based on the finding that 60% of the start-ups from the pre-processing cluster are using unstructured data as input for building their solutions. In average, data-driven solutions of the pre-processing cluster rely nearly 50% more often on unstructured data sources. Nearly half of them focus on video and image data, 25% use text, language and audio, 5% use laser and others.

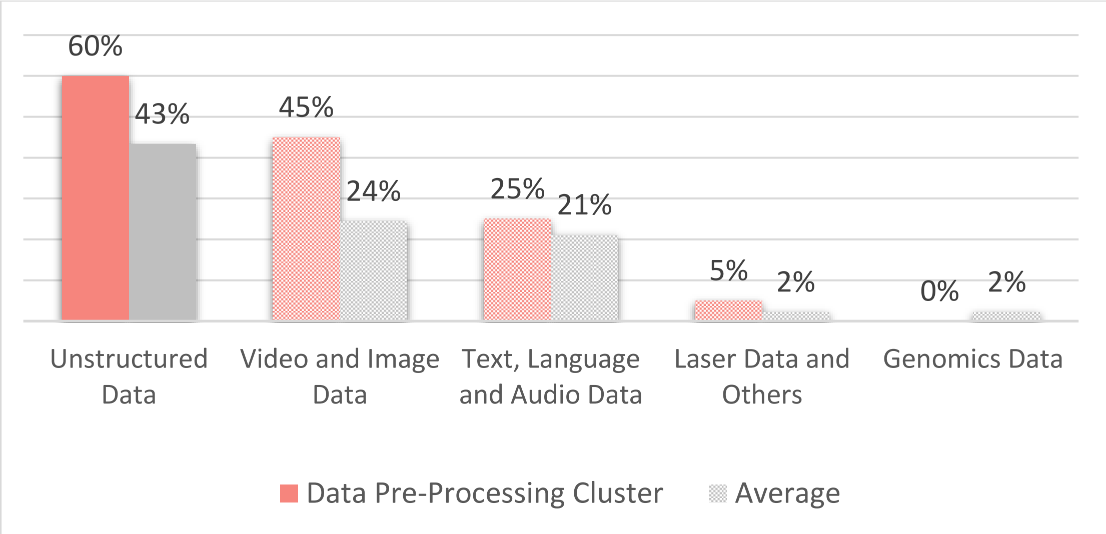


Figure 21 Usage of Unstructured Data in the Data Pre-processing Cluster

Nearly half of them focus on **video and image data**.

Some examples out of our sample are *Valossa*[61], a Finish start-up using video recognition and content intelligence software platform for businesses working with video. The *Valossa* offering combines multimodal video analysis and recognition with high level semantic inferencing to make sense of video content data. The video recognition software is available as a SaaS service and on-premise software solutions.

Another instance is the company *Iris Automation*[62] offering data-driven solutions for autonomous control by developing collision avoidance technology based on video data, allowing drones to see the world like a pilot does. By unlocking safe beyond visual line of sight flights, Iris Automation technology increases safety and efficiency of autonomous vehicles.

Text, language and audio data are also a relevant data source for 25% companies in the Pre-Processing Cluster.

Another example from our sample is *Altilia*[63], an Italian company developing big data solutions that enable organizations to analyze data and make more informed decisions. Through intelligent automation, their AI solution understand documents and natural language, giving context to any kind of

[61] https://valossa.com/

[62] https://www.irisonboard.com/

[63] http://altilia.ai/

typical and alternative data. For instance, their main solution called MANTRA solves the problems of collecting, extracting, harmonizing, indexing, querying, exploring, analyzing, understanding, and making sense of all available multi-structured data.

5.1.3 … heavily rely on technologies for data management

Moreover, technologies of the category BDV SRIA Data Management[64] are seen in 75% of the Pre-Processing offerings, while the average is 41%. The BDV SRIA Data Management priorities address in particular the challenges of processing the large amounts of data being made available in a variety of formats – ranging from unstructured to semi-structured to structured formats – such as reports, Web 2.0 data, images, sensor data, mobile data, geospatial data and multimedia data. For instance, important data types include numeric types, arrays and matrices, geospatial data, multimedia data and text. This includes methods for semantic annotation of data, the *handling of unstructured data*, such as videos, images or text in a natural language (including multilingual text), or specific domain data, such as Earth Observation data, have to be pre-processed and enhanced with semantic annotation. In addition, methods for ensuring semantic interoperability and high data quality need to be in place.

5.1.4 … include own hardware components in their offerings more often

Another characteristic of this cluster is the fact that hardware components are more frequently included in the offering compared to the average in the sample. While we observe one in every third data-driven solution encompassing hardware, we observe one in every fifth company doing the same on average.

For example, *Iris Automation* is a company offering detect and avoid technology. By using industrial cameras onboard of a device such as a drone, it possible to observe and interpret the world for full situational awareness of the operational environment. Their proprietary algorithm allows to make intelligent decisions in real-time by using location, heading, and speed, which leads to the determination of the safest course of action to avoid collisions. The hardware provided by Iris Automation can be integrated into any design, being a plug and play solution. The images below illustrate how the hardware looks like and is composed by a camera and a module.

[64] http://www.bdva.eu/SRIA

Figure 22 Iris Automation hardware is composed of a camera and a module. It can be integrated into any design, such as a drone[65]

For example, *Lynq*[66] is an American start-up that is changing how location data is communicated by allowing data to be transmitted for miles without networks or infrastructure. The hardware is a small device that works for groups of two to 12 people that are on a radio of 5km, and is a valuable solution for snow sports, travel, camping & hiking, and music festivals.

In addition, it was also part of our data collection to identify if the value proposition relied solely in an internal solution or if it also relied on external solutions. We find out that when a hardware was offered, an external solution was part of the process in 30% of the cases.

5.1.5 … are not likely to benefit from network effects

The pre-processing cluster is the one with the least network effects. This is related to the intrinsic definition of data pre-processing, where there is a low need for an input from other partners or applications to transform the solution into a more valuable business offer. From our data analysis, it was seen network effects on data level in 30% of the cases, 10% for infrastructure level and no network effects on marketplace level. This is in all three categories much lower than average.

For instance, *Playbasis*[67]offering customized gamification applications is not relying on network effects. The company provides a cloud-based off-the shelf mobile platform which supports business customers (e.g. fin-tech companies) to transform their channel-based customer engagement strategy into an omni-channel solution enabling new gamification-based opportunities for engagement, retention and conversion. They have developed a dedicate partnership program aiming to attract business customer to become solution partners whit whom they co-develop the new omni-channel solution. As their technology is based on a proprietary rule engine, their solution does not improve by the availability of more data (not network effects on data). In

[65] Image available at www.irisonboard.com/casia/

[66] https://lynqme.com/

[67] https://www.playbasis.com/

addition, although they rely on a dedicated partnering strategy for turning potential customer into solution partners, their performance of their offering does not depend on the number of partners. They simple need customer /partners to use the service and generate revenue. By relying on a software-as-a service-based subscription model, their business customers only need to pay for the service they use. Thus, the investment costs for setting-up the solution for business customers is solely captured by *Playbasis*.

5.1.6 … are not characterized by a dominant revenue model

The pre-processing cluster is a good example of how diverse a revenue model could be. The most prevailing models are subscription being employed in half of the cases when the average is 61%, licensing in 25% in contrast to 11% on average, 10% usage fee when the average is 4%, 20% asset sales when the average is 12% and 20% selling of service when the average is 26%.

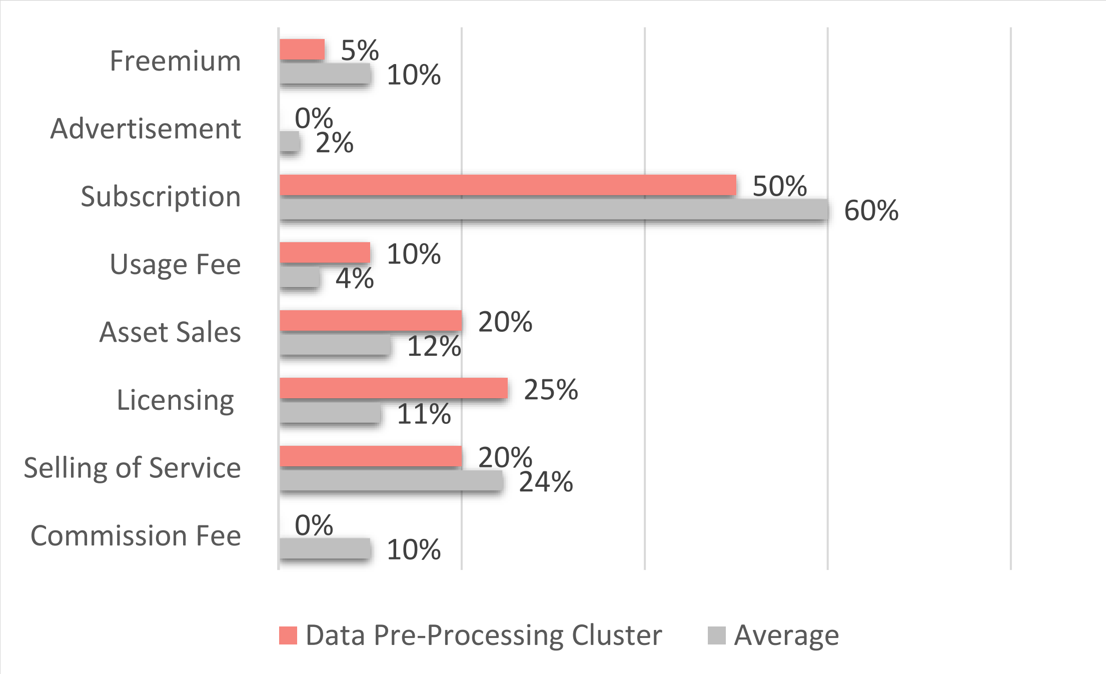


Figure 23 Revenue Model - Data Pre-Processing Cluster

We also wanted to understand if being sector focused or sector agnostic would change the revenue model pattern. From the figure below we can extract that sector focused companies diverse their revenue model slightly more than sector agnostic companies. Freemium and usage fee are not applied to sector agnostic solutions. Also, subscription being applied twice as much in the sector focused than in the sector agnostic ones.

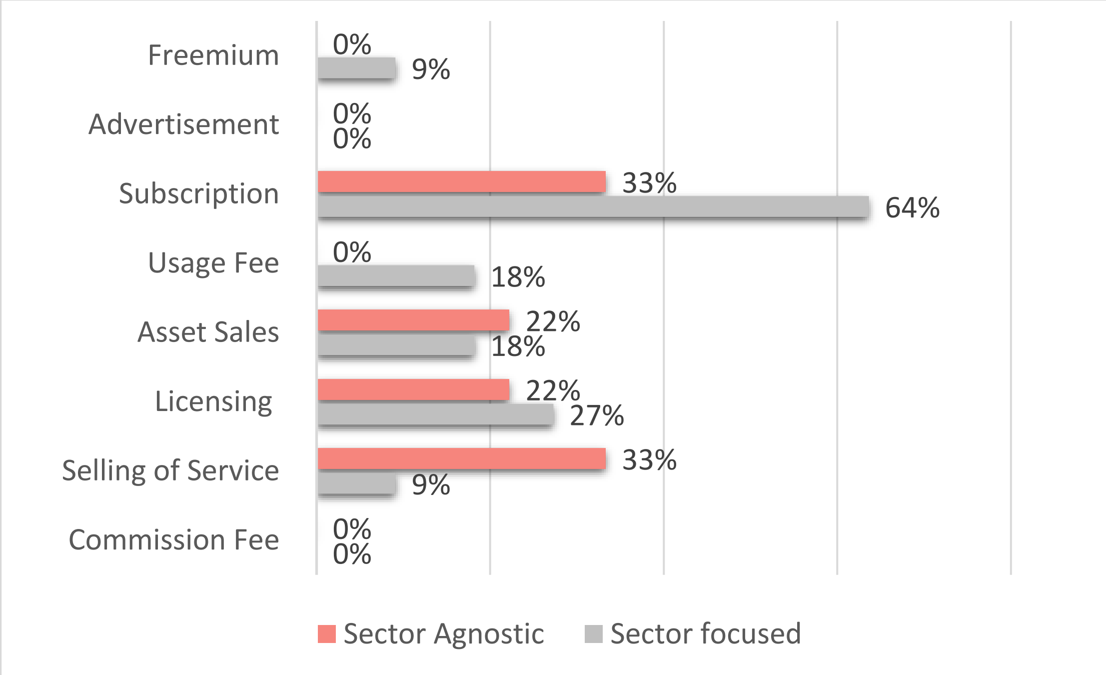


Figure 24 Revenue Model according to Sector Classification – Data Pre-Processing Cluster

For instance, *Valossa*[68] is a sector focused company using Subscription. The start-up provides video recognition and content intelligence software platform for businesses working with video. *Valossa* AI technology combines multimodal video analysis and recognition with high level semantic inferencing to make sense of video content data. *Valossa* builds computer software that sees, hears and profiles video content based on its semantic information. Beside their API and platform, they offer sector specific offerings in various sectors, such as Content Moderation or Sports Event recognition in Media & Entertainment, Human Analytics for Automotive or Retail, Interactive Cognitive Applications for Smart spaces, and Industrial Ai Application monitoring safety using live camera feed analysis. They offer *subscription revenue model* on a monthly basis with a fee of 199€. With this option, it is possible to use the video insight tools that include features such as face recognition training and download speech-to-text transcripts. The subscription is complemented with a usage fee as well as tailored pricing model for enterprises

In Figure 25 it is shown see how those options are offered in detail and which benefits the customer extracts from it.

[68] https://valossa.com/

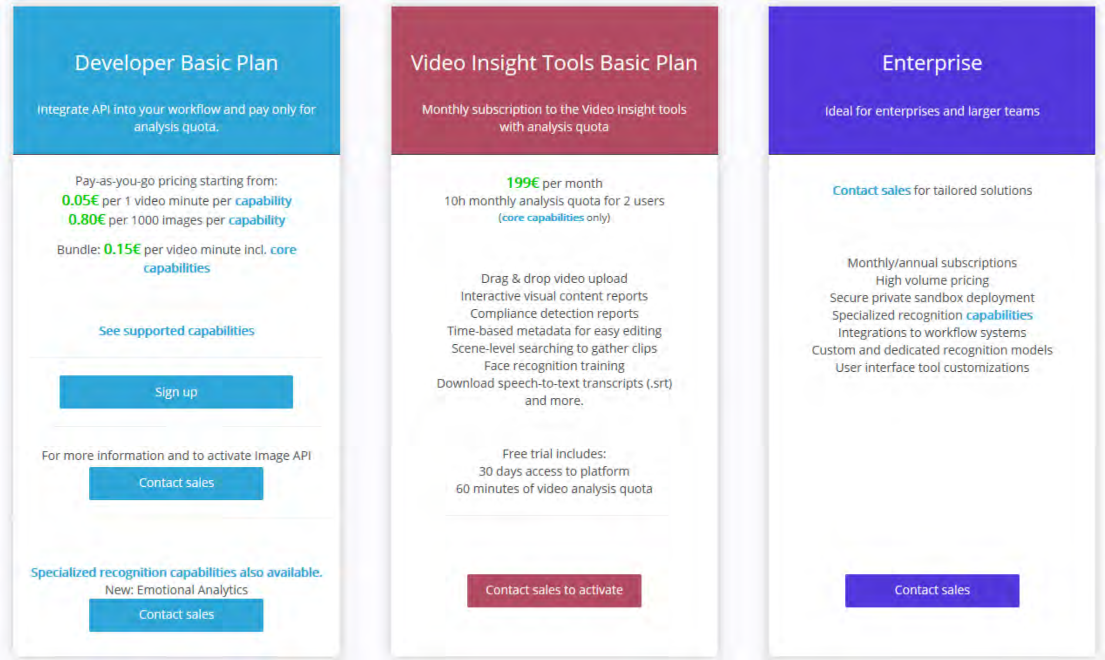


Figure 25 Subscription is one of the revenue model option from Valossa, a sector focused video recognition and content intelligence software platform[69]

Asset sales is marginally more used in sector agnostic solutions and *Licensing* is distributed more equally between sector agnostic and sector focused. The latter is 27% while the former is 22%.

Outstanding is the much higher percentage of 33% for selling of services in sector agnostic pre-processing companies compared the sector focused ones with 9%.

For instance, *ParallelDots*[70] offers a sector agnostic solution while relying on selling of services for generating revenue. The company develops artificial intelligence solutions for developers, start-ups, and enterprises by offering a complete stack of intelligent text APIs to enterprises. Those APIs are covering different functionalities, ranging from sentiment analysis to sarcasm detection. The founders are stating that they still tweaking their business model and revenue strategy to find out what is working out better and what isn't. The chosen revenue model for their solution is based on a differentiated subscription model with four different versions addressing starter, standard, business and advanced users with customized pricing model[71]. To complement the subscription strategy, they also develop with their own staff customized solutions for enterprises (selling of services).[72]

[69] Image source: https://valossa.com/

[70] https://www.paralleldots.com/

[71] For more details see https://www.paralleldots.com/pricing

[72] https://inc42.com/startups/parallel-dots-growth-story/

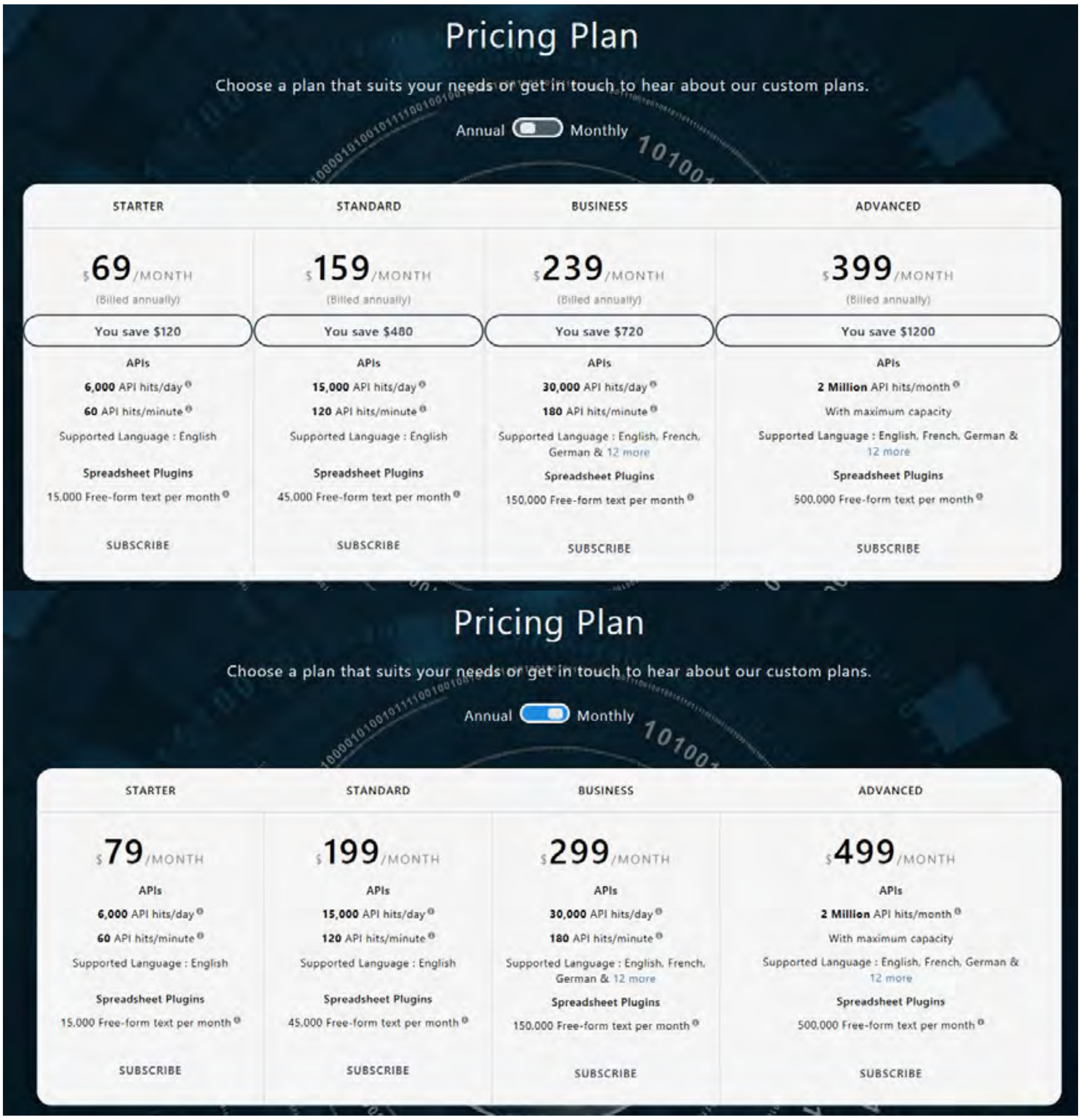


Figure 26 Differentiated Subscription Pricing Strategy from Paralelldots on annual and monthly basis

5.2 Start-ups in Cluster B: Internet of Things (IoT) application …

Start-ups of the Cluster B *offer Internet of Things (IoT) applications*, which include the IoT Technology – a complementary technology to the five BDV SRIA technological priorities - as part of their offering. As the deployment of IoT technologies often requires the integration of multiple different types of technologies as well as of multiple types of data sources, start-ups of this cluster are relying on data as well as technology integration functionalities as important aspect of their overall value proposition. By making the IoT technology ready to be used in end-customer applications they bring competitive value to their mainly business customer. A complementary aspect of their value proposition is the usage of various data analytics functionalities that allows them to bring wide range of insights to their customers. Although companies of this cluster rely on industrial data twice as often as others as well as tend to use many different types of data sources, they are less likely to rely on unstructured data.

5.2.1 … focus on sector specific solutions

Start-ups in the IoT application cluster offer sector specific solutions in 86% of the time. With their focus being on the deployment of IoT-based offerings, they need to put high efforts in data and technology integration, which in general is done in concrete sector applications.

> For instance, *Eliq*[73] offers AI-powered a SaaS-based solution for monitoring and optimising electricity use to gain a better understanding of your customers' energy usage patterns. The IoT-based platform is integrating data from smart meters, connected devices and weather data to produce real-time insights related to energy consumption patterns.

In cases offerings of this cluster did not have a clear sector focus, they aim to improve support activities or tasks that are requested in many different sectors.

> For instance, *Geopal*[74] provide a mobile workforce management solution that integrates seamlessly with the customers' existing office systems, such as ERP, CRM, GIS, BIM or Asset Management system. By seamlessly connecting field workers with the traditional back office systems and by providing real-time visibility of field operations, business operations in Power, Water, Telecoms, Oil & Gas, etc. can be made more efficient and safer.

5.2.2 … rarely use unstructured data

Unstructured data is less seen in the IoT application cluster when compared to the other clusters. While the average is at 43%, IoT start-ups make use of it in 21% of the cases. If seen, they are related to video and image. This might be due to the case that the data and technology integration is already challenging enough to ensure that companies can bring a competitive offering onto the market.

---

[73] https://eliq.io/

[74] https://www.geopal.com/

For example, the start-up *Arable*[75] does not use unstructured data in its solution. Its offering for agriculture companies is a solution to manage weather risk and crop health, delivering real-time, actionable insights from the field. Their exclusive hardware continually captures structured data such as precipitation, evapotranspiration, solar radiation and temperature. Their complementary hardware solution is an all-in-one irrigation management tool, weather station, and crop monitor, synthesizing both climate and plant data to produce actionable insights for all growing conditions.

And exceptions prove the rule. This is here the case when the analysis of unstructured data source can bring complementary but needed insights into the overall analysis. In such cases, it is very reasonable that the pre-processing of unstructured data is included in an IoT-offering.

For instance, *Taranis*[76] is a leading precision agriculture intelligence platform that provides a complete digital agronomy solution to identify, analyse and treat early signs of crop threats to allow farmers to make informed decisions leading to lower costs an maximize yield returns. To be able to provide precise insights the input needs to cover all relevant aspects. *Taranis* highlights that they have established the most comprehensive crop threat knowledge base relying on a wide range of data sources (features of the machine learning algorithm). This includes data from wide range of sensors providing real-time observational data from the field as well as surveillance imagery data captured by drones and satellites.

The image below illustrates the range of features that *Taranis* offers to its customers. Starting from the agricultural imagery solutions, *Taranis* has two proprietary technologies called UHR and AI^2. UHR captures multispectral imagery over fields which is then used by their scouting platform AI^2 that classifies and analyzes what is wrong in each acre. This platform is able to report plant population, detect when weed emerges, keep track of the field health, identify and categorize the top relevant diseases and detect growth irregularities. This is all analyzed and turned into real actions by their full analytics platform UHR.

[75] https://www.arable.com/

[76] https://taranis.ag/

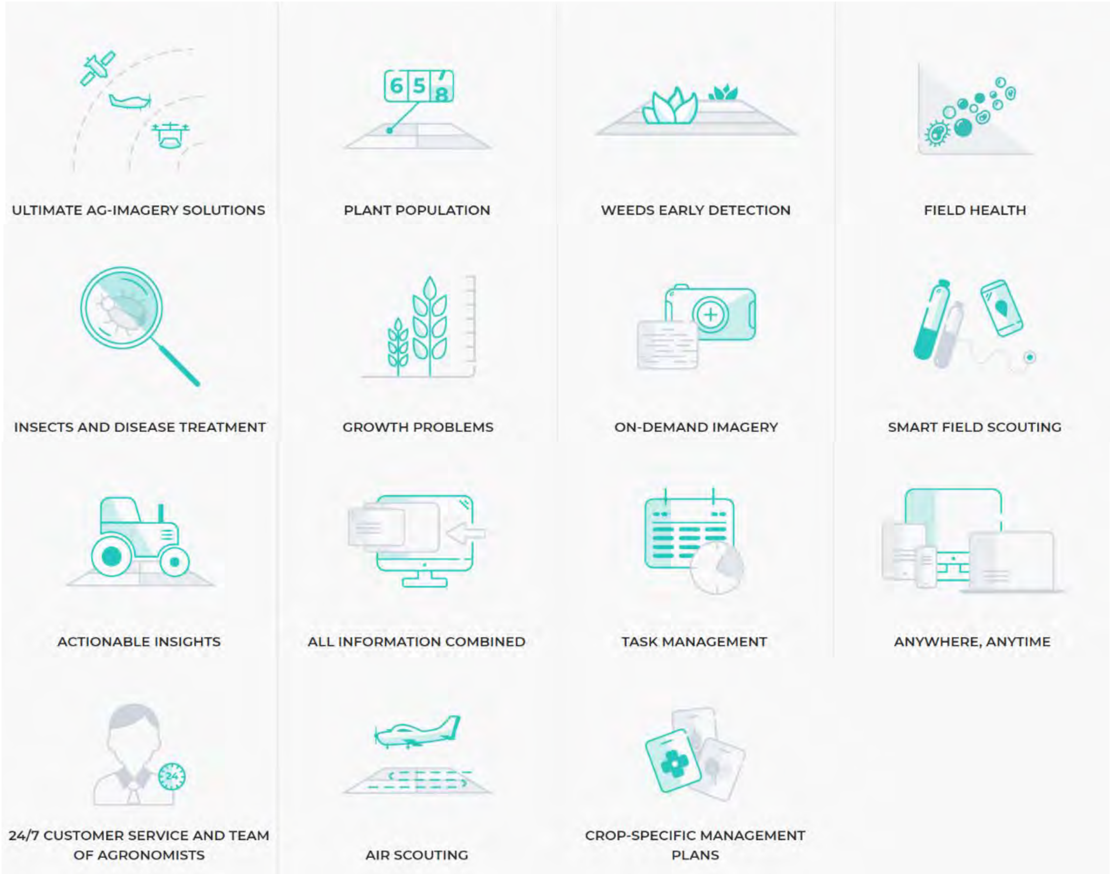


Figure 27 Comprehensive range of features of *Taranis*' agriculture intelligence platform

The difference between both start-ups, for instance the above mentioned *Arable* and *Taranis,* is their value proposition as well as data source used. *Taranis* is using imagery to monitor the fields and protect the crops against nutrient deficiency and insects. On the other hand, *Arable* uses a hardware that is fixed in a strategic location to capture different types of data and generate insights for plantation management, but no image is captured and analyzed.

5.2.3 … but frequently use industrial data

Start-ups in the *IoT application cluster* use industrial data more frequently compared to the average. This finding is not very astonishing, as industrial data[77] refers to any data assets produced and used in industrial settings of all areas, such as productions lines, energy systems, infrastructures, etc. that combine many distributed components into one functioning systems. Thus, the operational data of industrial settings is in general produced and stored in distributed locations and requires to be seamlessly connected in real-time.

[77] Industrial data how we defined it in our study

> For instance, Senseye[78] is offering cloud-based software for predictive maintenance taking information from industrial IoT sensors and platforms to automatically diagnose failures.

But we also have examples of data-driven offerings that rely on IoT technologies without any type of industrial data. For instance *Anagog*[79] sector specific solutions (e.g. In retail, banking, telecom, etc.) are based on a smartphone artificial intelligence engine that collects and used data from multiple smartphone sensors, correlates the data to visualize the smartphone user's journey and activities, extract his/her user behavior and provide predictions about what users are doing next in real-time. This edge AI engine established the basis to build customer-engagement solutions in various sectors that rely solely on end-customer behavioral data captured by smartphones.

5.2.4 … offer hardware solutions more frequently

Comparable to the data pre-processing cluster, start-ups in the *IoT application cluster* offer hardware more frequently than the average. In concrete numbers, 29% use hardware as distinguishing feature of their offering, compared to 19% in the overall sample.

> For instance, the agricultural business intelligence solutions based on in-field measurements to reduce crop health risks from *Arable*[80] relies on field-level weather and crop monitoring devices. As those hardware devices have not been available on the market before but are crucial for getting access to the required data, those hardware devices are a crucial part of their solution. The hardware can combine weather and plant measurements, sending it to the cloud. Its sensor is able to track crucial data that improves the activities performed in the fields, such as: precipitation, evapotranspiration, radiation, plant health, weather and harvest / event timing.

[78] https://www.senseye.io/de/

[79] https://www.anagog.com/

[80] https://www.arable.com/

Figure 28 Arable's device that combines weather and plant measurements in real-time[81]

5.2.5 … rely on the seamless combination of many different technologies

Start-ups in the *IoT application cluster* develop competences in data and technology integration as competitive assets. For instance, 93% offerings in this cluster use and integrate at least three different types of BDV SRIA or complementary technologies and 80% of offerings combine at least three different types of BDV SRIA technologies. The above numbers are very different to the average of all start-ups, e.g. only 22% of all start-ups integrate more than three SRIA BDV technologies and only 29% more than three BDV SRIA or complementary technologies. From those numbers we can derive there is a *high need of harmonization and interoperability of data and technology in the IoT application cluster*.

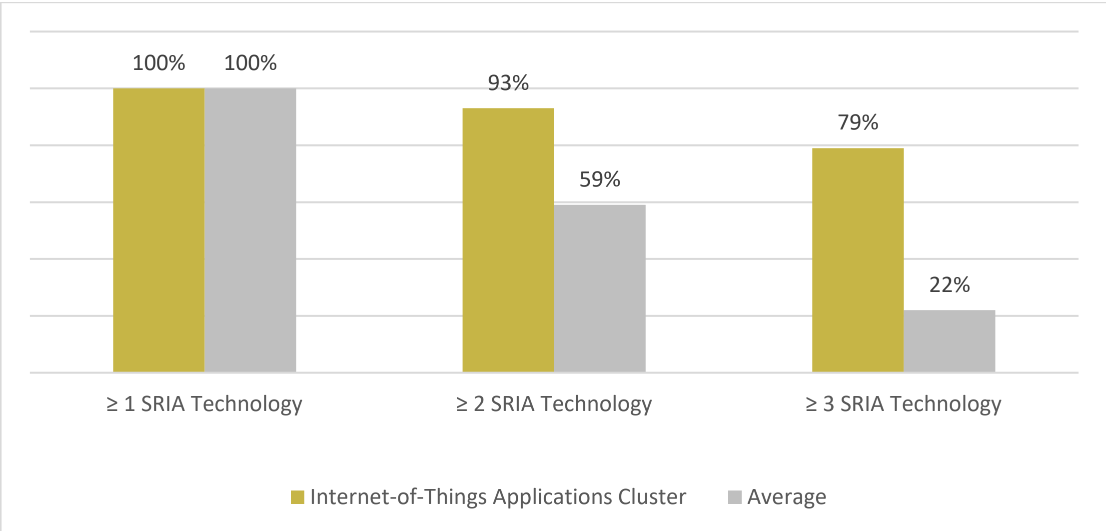


Figure 29 Usage of SRIA Technologies - Internet-of-Things Applications Cluster

CloudMedx[82] is an example of start-up applying at least three SRIA

[81] Image sources: https://www.arable.com/

[82] https://www.cloudmedxhealth.com/

technologies. The company designs artificial intelligence driven software for medical analytics. This means that medical notes, clinical history, demographics and medical procedures are merged and generate a clinical insight. For instance, the identified SRIA technologies used to generate such insights are SRIA Data Management, SRIA Data Analytics and SRIA Data Visualization and User Interaction. The image below illustrates how this process is organized. From the first step we can derive that SRIA Data Management is observed as there is a need for semantic annotation of unstructured and semi-structured data and semantic interoperability. From the second and third step we can observe the usage of SRIA Data Analytics as there is a semantic and knowledge-based analysis that continuously improve and provide real time interpretation of data. Finally, the SRIA Data Visualization and User Interaction is observed on how *CloudMedx* is presenting the results to dedicated teams, and this is done through a user-friendly platform that allows for predictive and prescriptive analytics to assess current metrics and build a path forward with informed decisions.

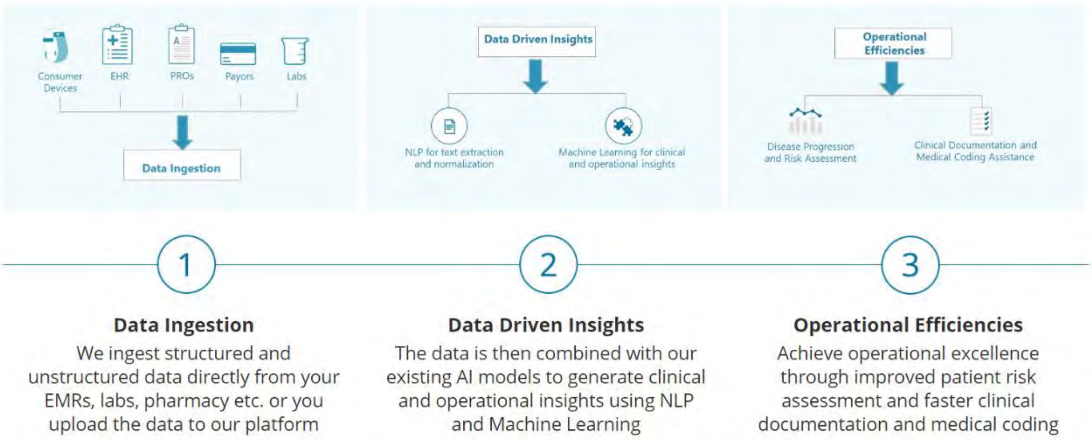


Figure 30 CloudMedx applies SRIA Data Management, Data Analytics and Data Visualization and User Interaction technologies through a three-step process approach

The efficient development and deployment of technologies priorised by BDV SRIA and IoT technologies is highly dependent on the *availability of wide range of interfaces and standards*. This will support the convergence of AI & Data and IoT technologies and solutions establishing the basis to build intelligent networks and systems that are becoming increasingly more capable to bring wide range of value to different industries. Both technologies are benefiting from bringing them together: AI and Data technologies bring value to IoT world through data analytics, improved decision making and means for data protection, whereas IoT brings value to the AI & Data world through the access to new data sources by providing means for connectivity, signaling and data exchange[83].

[83] See also https://www.businesswire.com/news/home/20191127005413/en/Artificial-Intelligence-AI-Big-Data-Data-Service

All start-ups in the IoT application cluster *rely on some type of data analytics as part of their offering*[84].

> For instance, the company *GeoPal*[85] offers for their business customer real-time visibility of their field operations through executive dashboards and KPI reports delivering increased operational efficiency and health & safety compliance.[86]

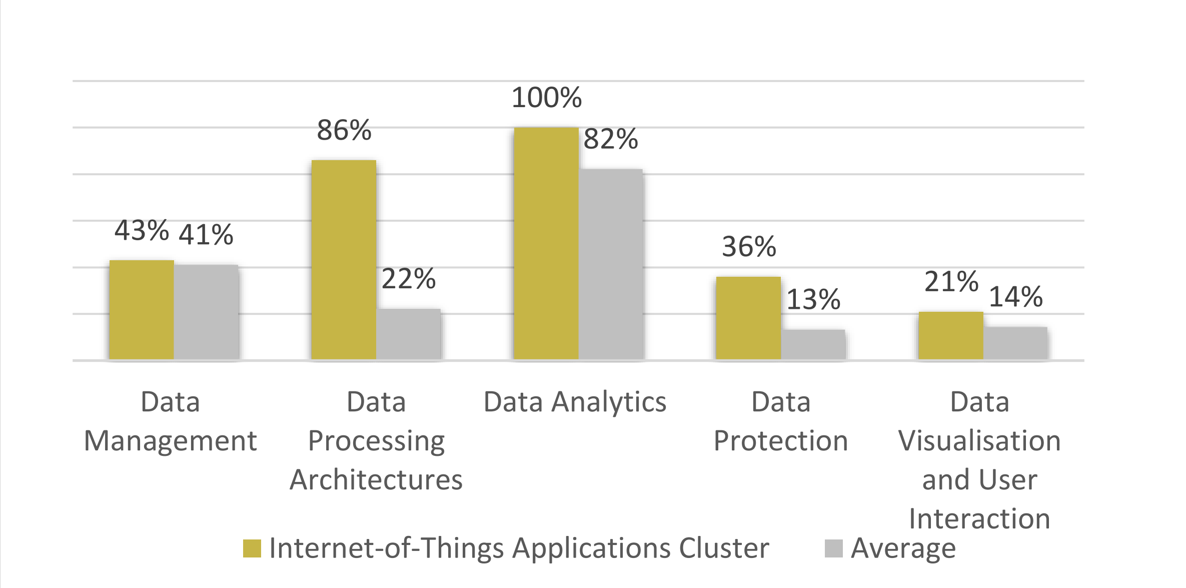


Figure 31 Technology - Internet-of-Things Applications Cluster

The frequent coverage of *data processing architectures* is not surprising as with the decentralized set-up of IoT application data-driven means for the efficient handling of data in rest and data in motion are of high importance.

Moreover, *data protection technologies seem to be more critical* for offerings in the IoT application cluster when compared to the other ones. In our sample, 36% of the IoT application offerings include data protection technologies into their solution, whereas in average this is happens only in 13% of cases. As in IoT setting high amount of data are produced and transmitted, and – in addition – the data exchanged is very often personal data or industrial data which require high privacy requirements to be fulfilled this finding is not surprising. Some companies transfer the need for high data protection standards into competitive features of their overall solution.

> For instance, the already discussed offering of *Anagog* (see Section 5.2.3) strongly makes use of personal data. They have established a privacy by design

---

[84] Note that the numbers in our study related to using analytics for generating value versus using BDV SRIA technologies are not always in the same, as the definition of BDV SRIA technologies is broader than the four dimensions of value generation.

[85] https://www.geopal.com/

[86] Note, that when we first analyzed the companies in 2018 GeoPal did not provide detailed information how data value was generated. While we are writing this report in 2019 very detailed information related to descriptive diagnostics offering can be found on the website.

approach that defines a transparent process to ensure the privacy of data. Their solution enables users to define and control their own personally identifiable information (PII), meaning that the private data will remain on the smartphone while still being used for providing personal and contextual services.[87]

5.2.6 ... use the freemium revenue model more frequently

IoT start-ups are more likely than others to use freemium as a part of their revenue strategy. The statistics is that one in every third company offer it.

For instance, the data analytics platform *Verv* of the company *Greenrunning* allows collating and analyzing real-time electricity data to generate a range of data services souring energy, appliances and customer analysis. End-customers can use this service to analyse the 'energy signature' of their own electrical appliances which helps them to explore the costs of each appliance on a real-time basis. By giving usage permission of one owns data for analytical purposes, the service is free for end-customer and can be accessed via a mobile app.

In addition, solutions of the IoT application cluster offer a wide range of revenue models. Subscription, freemium, selling of service, asset sales and usage fee are all applied in a higher frequency than the average. Licensing is the only one below average, and advertisement and commission fee are not used at all.

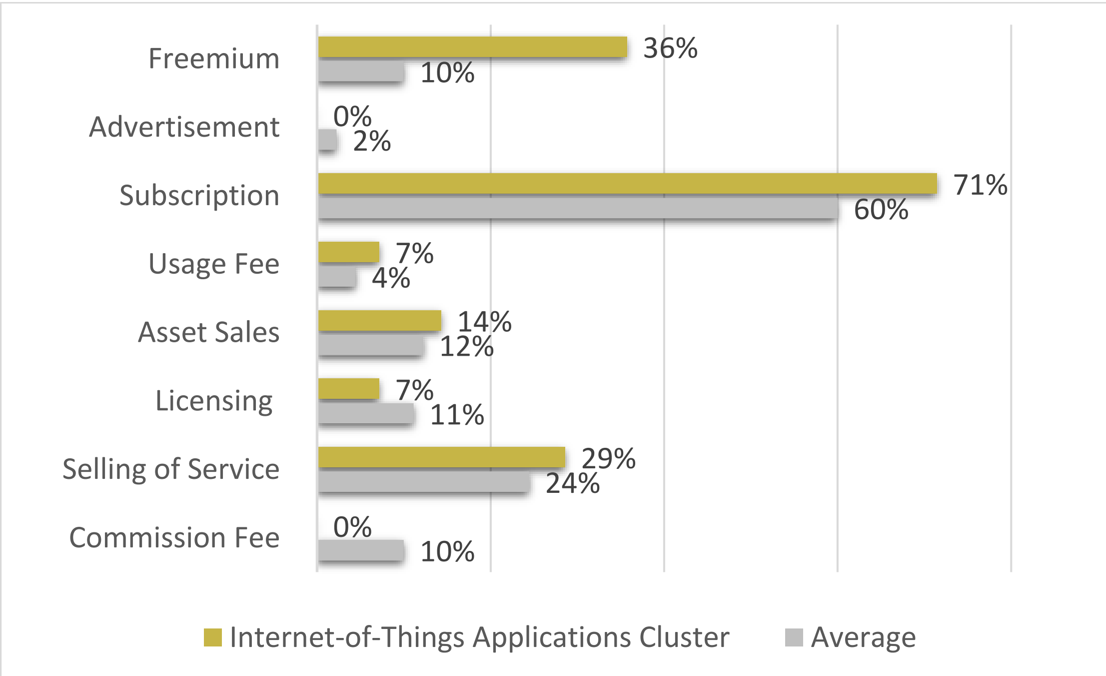


Figure 32 Revenue Model - Internet-of-Things Applications Cluster

[87] https://www.anagog.com/highly-sophisticated-patented-technology-designed-for-optimal-privacy/

5.2.7 … rely on network effects on data level

Network Effects on data level are seen in 64% of the cases in the *IoT Application Cluster*. Compared to the average sample, the percentage for IoT applications is one third higher. Due to the high amount of data produced by IoT devices, the size and amount of data is in general not seen as limiting factor but as opportunity.

For instance, the before mentioned *edge AI engine* developed by *Analog* enables mobile apps to understand user activities, behaviors and locations he or she visits in the "offline" world, and even predict what they'll do next, all on the phone itself. The more data the company gathers, the more efficient the app gets.

Another example is the American start-up *Swiftly*[88] that can increase accuracy by capturing more data. The Swiftly platform helping transit agencies and cities to improve urban mobility to improve transit system performance, service reliability, and real-time passenger information. The underlying AI algorithm streamline and analyze the transit data in real-time. With more data available the algorithms can continuously increase the accuracy and reliability of results.

5.2.8 … are twice as often niche player in a functioning ecosystem

Start-ups in the *Internet-of-Things application cluster* are twice as often a niche player compared to the overall sample. As the deployment of IoT technologies involves the interplay of many different technology components, it is more likely that businesses in this area decide to rely on strategic partnerships by joining an ecosystem that helps them and their partners to benefit from synergies by co-developing the complementary technology components.

For instance, *Carfit*[89] is revolutionizing the car maintenance field and it is not doing it alone. This global technology company for IoT/data analytics, NVH technology and Machine learning created a self-diagnostic and predictive maintenance platform for cars. Together with *NVIDIA*[90] and its innovation ecosystem, the firm has access to the technical knowledge and hardware that is necessary for accelerating the development of their cutting-edge products via an accelerator program. For the time being that the autonomous driving market is well established and mature, this partnership will create the ground for autonomous driving in general with autonomous maintenance being a specific but important niche application.

[88] https://www.goswift.ly/

[89] https://car.fit/

[90] the world's leader in Artificial Intelligence Computing hardware and the largest manufacturer of GPU's

5.3 Start-ups in Cluster C: Industrial Services ….

Companies the *Cluster Industrial Services* are characterised by the usage of industrial data sources, close customer relationships and offerings covering the all flavours of data value generation and process automation. Companies of this cluster seem to be prepared to make us of available services for processing IoT data but do not include IoT technology as component of their overall offering. More details about findings related to this cluster will be reported in the following subsections.

5.3.1 … do only address business customer

Exclusively focused in the B2B market, start-ups in the *Industrial Service Cluster* have no offerings for the end customer. The companies in this cluster act in the following areas and categories.

START-UP	AREA	CATEGORIES
PROTENUS	Healthcare and Health	Analytics, Big Data, Cyber Security, Health Care, Information Technology
CLOUDMEDX INC	Healthcare and Health	Artificial Intelligence, Big Data, Health Care, Information Technology, Machine Learning, Natural Language Processing, Neuroscience, Personal Health, Predictive Analytics
FRAUGSTER	Sector Agnostic	Artificial Intelligence, Big Data, FinTech
PLUTO AI	Environment (water, food, disasters management, etc.)	Analytics, Artificial Intelligence, Water
UPLEVEL SECURITY	Security	Cloud Security, Cyber Security, Data Visualization
ONCORA MEDICAL	Healthcare and Health	Health Care, Information Technology, Predictive Analytics
WEGOWISE	Energy	Analytics, Big Data, Energy Efficiency
RULEX	Sector Agnostic	Artificial Intelligence, Machine Learning, Predictive Analytics
QUINVI	Insurance & Financial Services	Accounting, Business Intelligence, SaaS, Software
AIMS INNOVATION	ICT	Analytics, Enterprise Software
TRADEGECKO	Retail & Consumer Goods	Business Intelligence, Enterprise Software, Payments, SaaS, Supply Chain Management
EMAGIN	Environment (water, food, disasters management, etc.)	Artificial Intelligence, Predictive Analytics, Water

Figure 33 Overview of start-ups from Industrial Services and their business focus.

Knowing that all start-ups in this cluster do make usage of industrial data, the exclusive focus on the B2B business is not surprising.

5.3.2 … tend to focus on insight generation and process automation

Startups in the *Industrial Service Cluster* are more likely to offer process automation - the average is 21%, while the industrial cluster is 50%. Also the degree to which Analytics is used for generating insights is high above average, for instance descriptive analytics with 92% compared to 54% in average, Diagnostic Analytics with 58% compared to 26% in average, Predictive Analytics 67% compared to 38% and Prescriptive Analytics 25% compared to 12%.

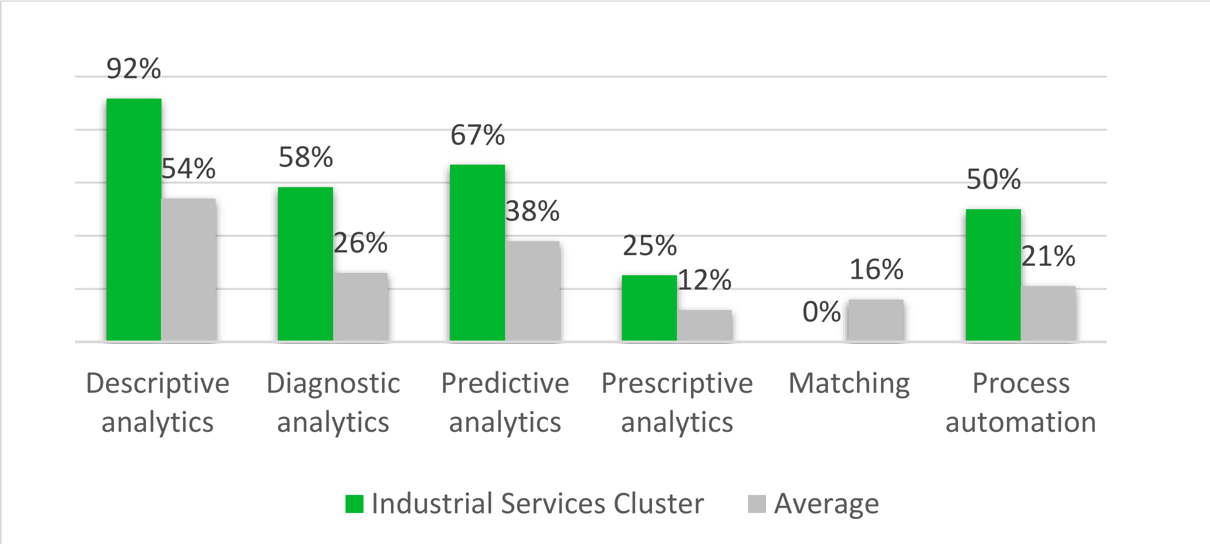


Figure 34 Generating Data Value - Industrial Services Cluster

For instances, the American start-up Qvinci[91] is offering a financial reporting and business intelligence solution for multiple industries. By processing the financial and business related data if their customers, they are able to provide solution for the automation of specific tasks, such as Automated Reporting or Workflow Management, as well as provide different types of analytical solutions, such as improved visibility and business intelligence (descriptive value) as well as error detection (diagnostic analytics)

5.3.3 … rely on industrial data and personal data

In this cluster all start-ups rely on industrial data. As this is compared to the average of 30% a quite outstanding result, the name of the cluster was selected accordingly to reflect this particularity.

In general, *industrial data* is "closed data" meaning that it is "owned" (in terms of access rights) by the entity operating the product, machine or thing producing the

[91] https://www.qvinci.com/

data and that the data is likely to cover confidential information, e.g. a log file from a medical MR might contain valuable information indicating IPRs of the MR itself. This also includes operational data in general produced in the context of operating any type of IT system (excel, definition sheet).

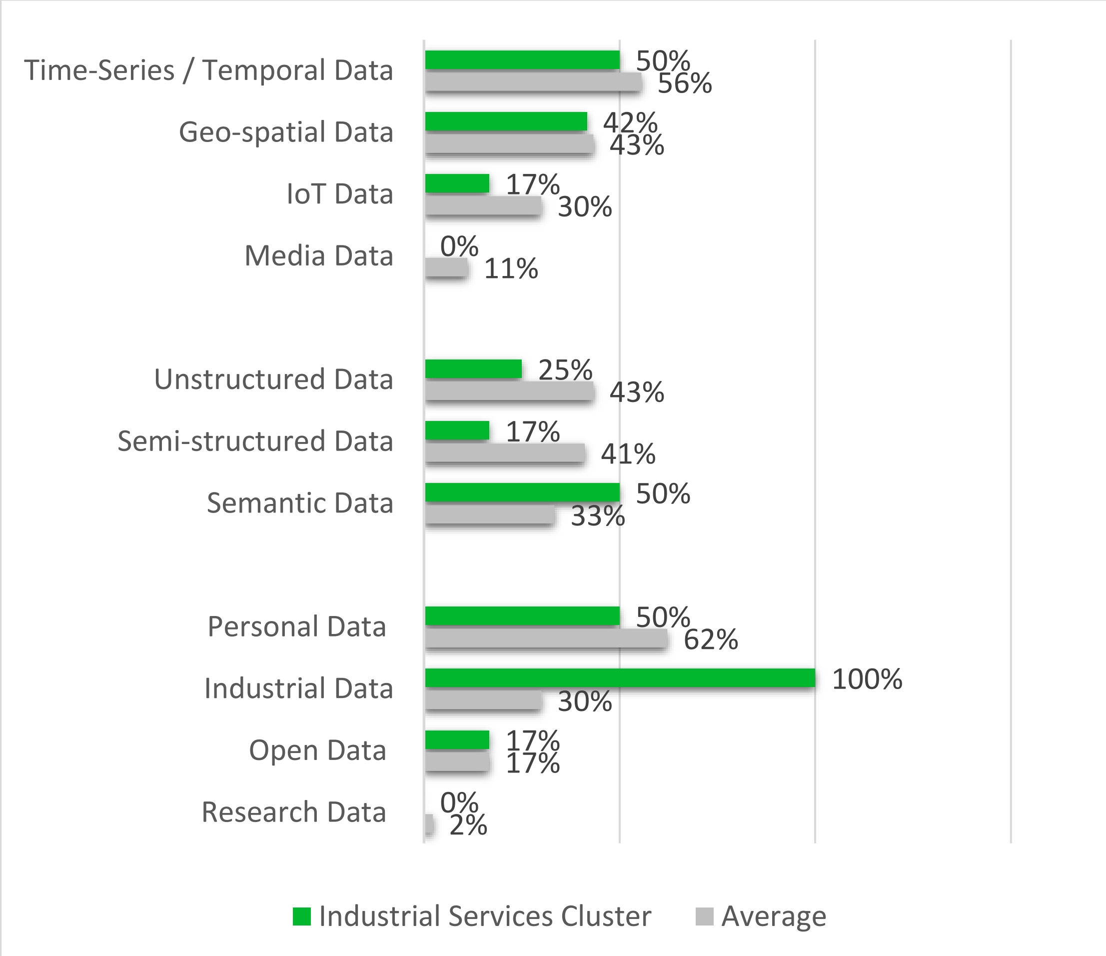


Figure 35 Type of Data – Industrial Services Cluster

Personal data is observed in 50% of the start-ups from the industrial cluster while the average is 62%. Although the companies in this cluster process a high amount of data, it was not expected to be the first place in personal data as the core idea is capturing value from industrial data. It is a similar logic to the pre-processing cluster, the one that uses personal data the least. But even if the cluster present a lower number compared to the average, it is still a high one. Every second startup using industrial data rely on personal data (mixed data). Therefore, there is a higher need for data protection (twice as much as average).

> *Fraugster*[92] is a great representative of business that highly use personal and industrial data in every analysis they perform. As an anti-fraud start-up, *Fraugster* algorithm was trained using millions of transactions from every industry and region, being able to decide if a transaction is fraudulent or not

[92] https://fraugster.com/

within 15 milliseconds. For example, when a shopper makes a purchase on an e-commerce website, transaction data points are sent to *Fraugster*, and this includes personal data such as location and shopping behavior. Also, valuable industrial data, such as operational data about the running e-commerce business, is used. Next, the data intelligence service from Fraugster enriches every transaction with up to 2,500 additional data and their AI technology engine translates these data points into clear behavioral patterns and determines if the transaction is fraudulent or not.

Unstructured data is not highly employed in the industrial cluster. And in cases unstructured data is used, then it is mainly audio, text or language data.

For instance, *CloudMedx*[93] is highly relying on text as input for generating value to its customers. It provides artificial intelligence-based software solution for medical analytics that uses structured and unstructured data as input ranging from patient (personal data) to clinical operational data (industrial data). The text data used as input origins from Electronic Health Records (EMR) and Patient-Reported Outcome (PRO). By automatically extracting based on natural language processing (NLP) clinical and operational insights can be generated.

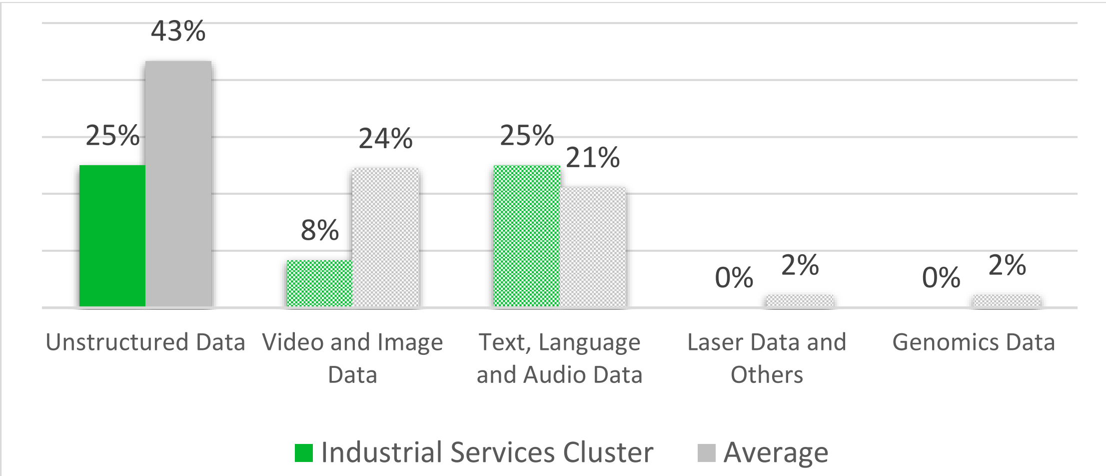


Figure 36 Unstructured Data: Industrial Services Cluster

5.3.4 … process IoT data but do not include IoT technology in their offering

17% of start-ups in the Industrial service cluster rely on IoT data, which is only half of the average. Outstanding here is, that although some start-ups are relying on IoT data, they do not include technical IoT offering as part of their solutions

The above mentioned anti-fraud company *Fraugster* is an example for this. Although the company relies on many open and external data points to enrich their customer's transaction information for improved fraud identification and

93 https://www.cloudmedxhealth.com/

predication, there is no indication that IoT technology would be relevant part of their offerings but that they are using off-shelf IoT technology for processing and incorporating data at different locations.

5.3.5 … provide little public information related to their revenue model

For the start-ups in the industrial service cluster it was more difficult to find public information about their revenue model. 33% do not share any information about their revenue model publicly, compared to 10% in average. This could be caused by the fact that start-ups in this cluster are establishing very close relationships with their customer to ensure that they get access to industrial data. And the more unique a partnership is the less public information will be available. For instance, the companies *Protenus*, CloudMedx, *Fraugster* and *Emagin* only share details related to their offering via their sales teams.

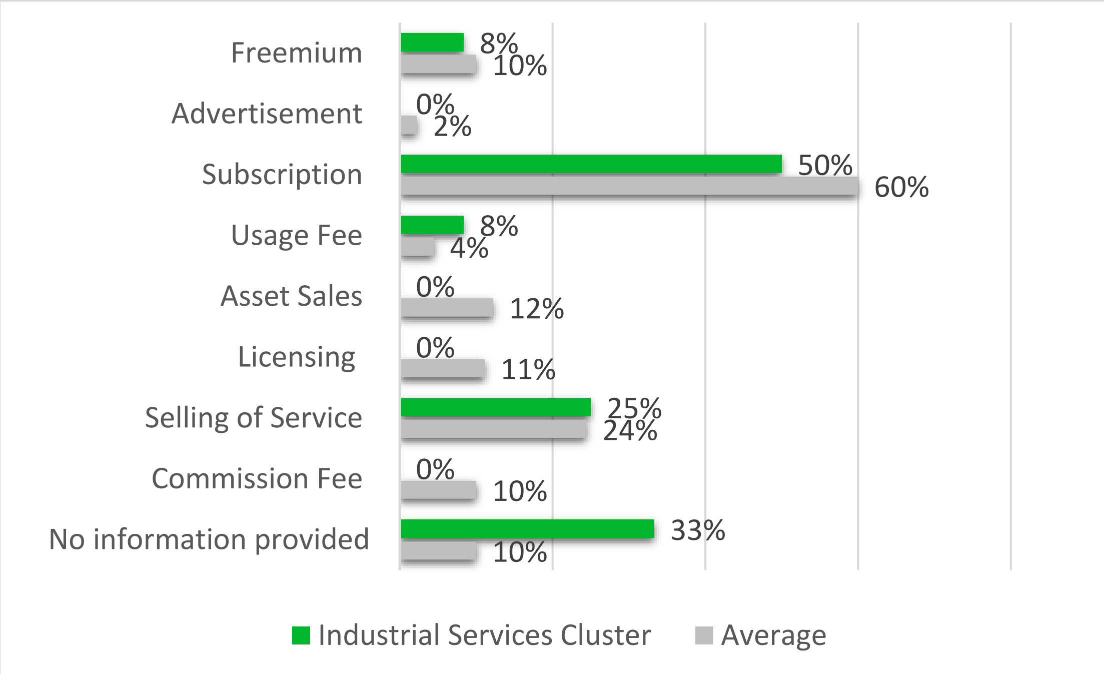


Figure 37 Revenue Strategy - Industrial Services Cluster

The most used revenue model in the industrial cluster is subscription, which is with 50% below the average. Similar to most of the companies offering subscription, they offer very tailored offering for different target groups and user needs.

For instance, *TradeGecko*[94] is an inventory and order management platform for modern wholesale merchants, offering subscription as a revenue model. Some of its services includes keeping track of the products as the company sells them and restock across multiple locations and channels; managing all the orders in one place, arrange fulfillments and getting paid 3x faster; organizing all the relationships in one place to speed up invoicing and cash flow; and plan

[94] https://www.tradegecko.com/

for growth with detailed custom reports on inventory and sales trends. As of end 2019, the company has included freemium as an alternative and has increased the subscription options as observed in the screenshot below in Figure 38. In addition, the company provide premium and pro offering for larger businesses with higher volumes including enterprise set-up and support services. Payment conditions for enterprises are negotiated via the sales channel.

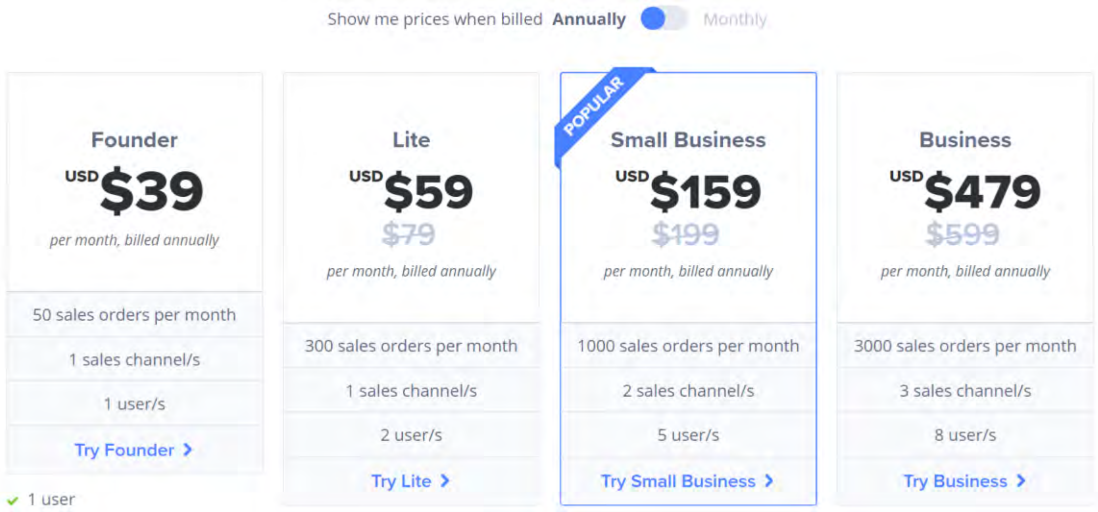


Figure 38 Pricing strategy of TradeGecko

Selling of Service is seen in 25% of the time, the same frequency as the average. Advertisement, asset sales, licensing and commission fee are not applied in this cluster.

5.3.6 … use network effects on infrastructure level above average

In average, start-ups in the industrial service cluster make less use of network effects compared to the average (50% in comparison to 57%). The slightly lower percentage of network effects on data level (42% to 49%) of might be caused by the fact that the access to industrial / operational data is more difficult and thus it is less likely that data-driven offerings rely on algorithm that improve with growth of data. Start-ups in the industrial service cluster do not harness network effects on marketplace level which is very much in line with the fact that no company in this cluster is positioning its offering as data-driven marketplace. Start-ups relying on network effects on infrastructural level help others to build their products and services, seems to be more likely in case industrial data is involved, i.e. we observe in this cluster a higher value then average, 17% in comparison to 10%.

For instance, *CloudMedx* is an example of start-up in the Industrial Services cluster that observe network effects on infrastructure level. *CloudMedx*'s goal is to improve the process of clinical care by using a proprietary clinical contextual ontology to improve patient journeys. The value of start-up's the more different clinical data sources of the different health provider can be integrated. With its offerings the company is addressing the challenge of integrating wide range of distributed clinical data source in various formats in efficient manner which again can helps hospitals and health service provider to implement improved processes and services for their customers / stakeholders. This means that the more connections to different software they are able to make, the faster hospitals and health service provider can implement data-driven services for their users. For example, different hospitals use different means to file Electronic Health Records (HER). For *CloudMedx*, this means that the challenge is being able to connect the different hospital IT system, healthcare data formats, without losing any information being of clinical relevance.

5.3.7 … rely mainly on data-driven services

Majority of start-ups of this cluster offer data-driven service, which is also with 92% above the average of 78%. The scoping of data-driven services benefits from close relationships with potential target customer. Due to need to get access to industrial / operational data, start-ups in the industrial service cluster seem to be in a good position to rescope or rethink established business processes, such as client interactions or service delivery systems and bring forward new service concepts based on the usage of data or data analytics that often allow to automate manual steps or to accomplish tasks in much better quality.

For instance, this is the case of *Protenus*[95], an American start-up that provides healthcare compliance analytics to help healthcare service providers to reduce risks. As patient privacy protection is an increasing concern due to the upsurge record of personal data, this start-up is helping healthcare institutions to better monitor possible data violations. And they are doing that by automating manual steps, auditing every access and creating alerts for vulnerable patients, an activity that is poorly performed without becoming a data-driven service.

None of the start-up in the digital service cluster is positioned as data-driven marketplace or emerging technology offering. However, niche offering can be found in this sector, also this is with 8% while the average is 12%.

A niche offering example is brought by AIMS Innovation (see also Section **¡Error! No se encuentra el origen de la referencia.**), the Norwegian start-up that delivers the first intelligent monitoring solution to dynamically determine monitoring thresholds without human intervention and then, proactively send "early warnings" before problems arise. As Microsoft Enterprise partner, AIMS Innovation is one of the many niche players in the comprehensive Microsoft ecosystem. Their offering can be accessed via the Microsoft App Store[96].

[95] https://www.protenus.com/

[96] https://appsource.microsoft.com/en-us/product/web-apps/aimsinnovation.aims

5.3.8 … are less likely to receive funding

Start-ups in the industrial cluster are less likely to receive more than three funding rounds. The companies in this cluster have observed it in only 33% of the cases, the average is equal to 59%, while the highest percentage goes to the descriptive cluster with 81%.

Some of our hypothesis for this fact is that they might be addressing smaller market and that their offerings might face limitation in terms of scalability. Another reason might be that due the closed nature of industrial data, offerings are only applicable to very specific set-ups. Also, as it is a business directly connected to industries, funding might come in a more isolated manner, such as private investors or serial entrepreneurs already consolidated in the business.

From our sample, we have three examples that receive below US$3M of funding and did not get any additional funding from the beginning of 2018 to the beginning of 2020. These companies are *Rulex* that received US$2.5M, AIMS Innovation also with US$2.5M and *Emagin*, with US$2.2M. By highlighting these three companies we do not want to pass the image that these businesses are not as successful as the other ones, as this would not be the case as all three business are fully running. What we want to convey instead is that start-ups from these cluster might present a particularity when it comes to funding needs.

	Funding Amount US$M	**Business**
Rulex	2.5	Rulex is an AI software to embed automated real time predictive intelligence in applications, infrastructure, and IoT edge apps
AIMS Innovation	2.5	AIMS Innovation is about application performance monitoring
Emagin	2.2	Emagin provides water and wastewater facilities with an operational intelligence platform that supports real-time decision making

Table 6 Companies from the Industrial Services Cluster might present a different need with regards to funding

An industrial company that has received more than three funding rounds is *Protenus.* According to Crunchbase, as of end of 2019 they have raised series C and had capture in total nearly 36 million dollars. Trusted by major healthcare organizations, they use artificial intelligence to understand how electronic medical records are accessed in the course of care and detect anomalies to ensure that every action taken inside a health system meets the high standards that patients expect. They serve an array of healthcare organizations across the U.S., including top-ranked hospitals, payers and health information exchanges.

5.4 Start-ups in Cluster D: Descriptive Value …

Companies in the *Descriptive Value Cluster* are primarily focused on descriptive analytical services for a wide range of non-industrial data sources. Their offerings are all positioned on the market as data-driven service and are generating income mainly by subscription. Start-ups in this cluster are more likely to receive funding. All further detailed findings related to this cluster along with illustrating examples, will follow in the next subsections

5.4.1 … focus on descriptive analytics for non-industrial data

The main focus of offerings from start-ups from the Descriptive Value Cluster is on generating descriptive insights for non-industrial data sources. The majority tend to be very focused on descriptive analytics, other type of data analytics (diagnostic, predictive or prescriptive analysis), process automation or match-making is only in rare cases seen as part of their offering

> *HYP3R* is a location-based marketing cloud company on a mission to make marketing efficient for businesses and delightful for consumers. The goal is to identify unique audiences at the desired location and target them with timely, relevant campaigns. In their offering, you can find a visual dashboard that compiles the information needed to make decisions. One of its customers is Marriot, the hotel chain. Every post shared publicly is filtered and made available to Marriot, as well as celebrity posts about the company and people commenting about travel.

5.4.2 … rely on extensive use of different data types

As Figure 39 indicates, start-ups on the Descriptive Value Cluster are making use of a wide range of data sources. The only type of data sources that are no trusted by them are industrial data and research data source. By nearly all other types of data, their usage is close, often even significantly above average.

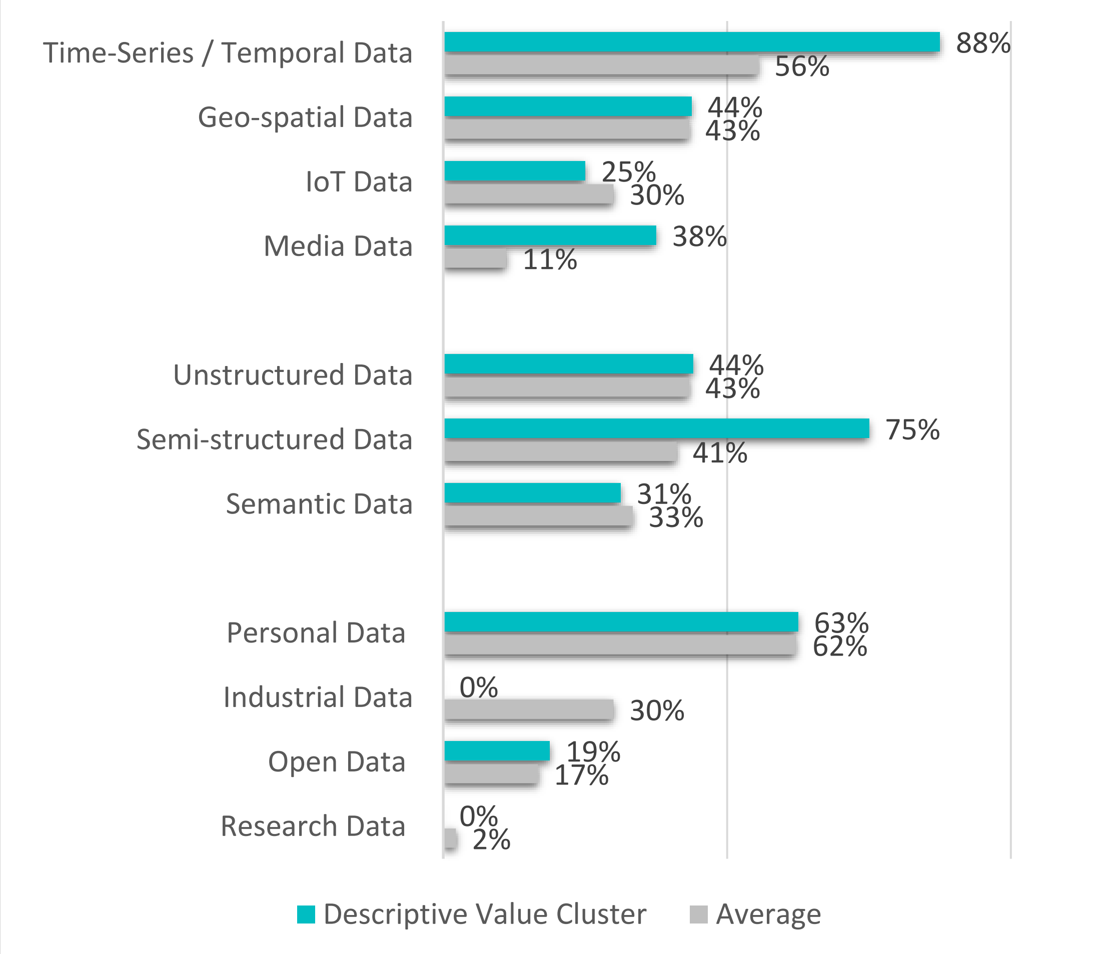


Figure 39 Type of Data - Descriptive Value Cluster

For instance, the usage of *Time series* and *Temporal data*, *Media Data* and *Semi-structured Data*, which is often social media, are significantly above average in this cluster and media data are more broadly explored in the descriptive cluster when compared to the average, while industrial data is not seen

Start-ups in the *Descriptive Value Cluster* rely on time series data in 88% of the time.

> For instance, the use of longitudinal data is central for *Edited*[97], an English start-up that is helping the world's leading brands to positioning their offerings at optimal price and time. Their real-time analytics solution for professionals in merchandising, buying, trading and strategy covers insights related to pricing, assortment as well as product metrics. *Edited* is keeping records of the historical sales performance (including launch data, discount, sales and more) for each offering in a longitudinal database covering more than seven years of records. By analysing all the related past time-series data, *Edited* can predict optimal roll-out times for new offering.
>
> *TVision Insights* is an audience measurement start-up using laser to generate value to large broadcasters and advertisers. The company's core technology uses data science and motion capture technology to passively assess "eyes on

[97] https://edited.com/

screen", the single most accurate way to measure person-level engagement with video content. The technology can identify who is watching TV and who is in the room but not watching. In other words, the captured data is forming a time series data and valuable interpretation can arise from it.

Semi-structured data is used in 75% of the cases, compared to 41% in the average sample. We observe that start-ups in this cluster rely to large extend on the processing of web-content as well as social media content, both data sources that are stored in semi-structured manner.

For instance, *Keywee* is extracting from online articles from the New York Times by using NLP technologies relevant keywords describing its content. While the company is extending the meta-data describing each single online article, they are also relying on any additional information conveyed by the semi-structured representation of the online article to feed their NLP pipeline.

Media data usage in this cluster is also with 38% significantly above average which is 11%. Media data includes all data sources used for media communications, covering a plural of mediums used for communications, such as television, mobile communication or social web.

Onclusive[98] (previously AirPR) is a data science company addressing the needs of professionals in communications and PR. They are offering a proprietary platform that is based on a comprehensive collection of communication content, i.e. their new scrawler has indexed some 5.5 billion articles which comes to about 100 terabytes of storage[99]. Based on this large collection of media content, the company is offering classical monitoring insights including social engagement metrics, e.g. by calculating metrics related to social shares in relation to sentiments or terms of relevance.

We do not only observe a higher usage of time series data, media data and semi-structured data compared to the average, we also see that they are likely to be used in combination.

For instance, *Apptopia*[100] uses different types of data to generate real-time engagement insights of mobile apps for their customers, including Google, Pinterest, Facebook, NBC Universal, Deloitte, and others. *Apptopia*'s analytics dashboards using engagement and usage data of more than 125000 mobile apps is able to provide reliable engagement predications for more than 7 million apps. For example, if an enterprise is interested in getting insights about the use of social media for a specific app, they can use *Apptopia*´s platform to search for similar apps that offer this service, and explore their advertising strategies in relation to their impact achieved, e.g. downloads by customers. For offering this service, *Apptopia* is relying on time series, media data and social media data.

[98] https://onclusive.com/

[99] May 2019 in accordance to https://www.swordandthescript.com/2019/05/onclusive-pr-attribution/

[100] https://apptopia.com/

5.4.3 … use mainly the subscription model as source of income

The *Descriptive Value Cluster* is the one with the highest frequency of subscription as a revenue model. Also, it was easier to identify a revenue model in this cluster. The correlation in this case is due to a more consolidated business, with a clear solution to the market and that is exploring the current available data at its best. While the subscription modality was found in 61% of the companies, on average, in the descriptive cluster this happened in 94% of the time. Figure 40 gives an overview of revenue models used in the Descriptive Value Cluster.

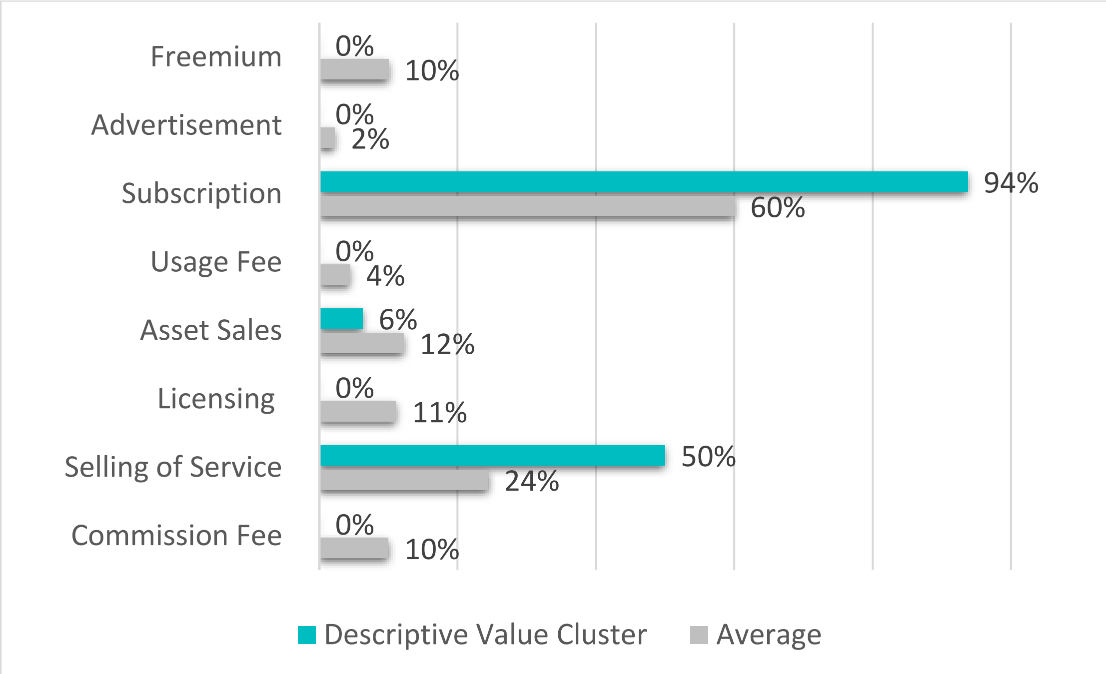


Figure 40 Revenue Model - Descriptive Value Cluster

For instance, *Rollbar*[101] is supporting software developers when deploying software with a service that automatically identifies, prioritises, and resolves coding errors. For their SaaS-service, they offer different types of subscription models depending on the stage and size of the company, also a free version for a specific amount of time. Figure 41 illustrates the pricing plans offered by *Rollbar*.

[101] https://rollbar.com/

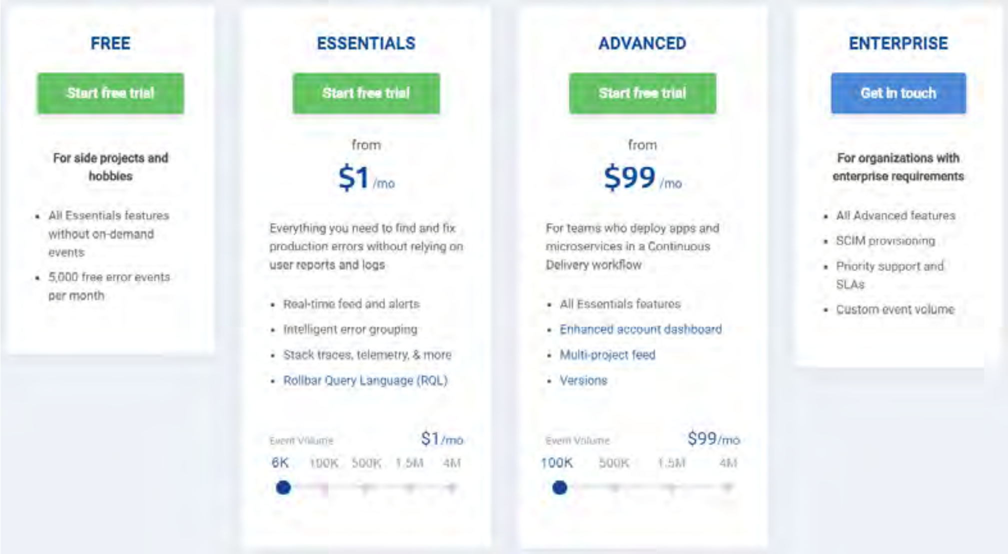


Figure 41 Overview of Rollbar's subscription models (source: rollbar.com)

Start-ups of the *Descriptive Value Cluster* are relying twice as often on selling of services as a revenue stream (average is 26% compared to 50% in the descriptive cluster). The Selling of Services revenue model is in general a very suitable income stream in case the data-driven offerings are not yet standardized, not off-shelf available or require additional education or process changes. For data-driven enterprise offerings often additional consulting services are provided to ensure that required support from experts to set-up and configure the data-driven offerings are in place.

For example, the start-up *Brand Embassy* was acquired by Nice Systems in May 2019 and operates under the new brand *NICE inContact*[102]. This cloud-based customer service platform is omnichannel and provides integrated social media, messaging, live chat and AI-powered chatbot service.

One of their revenue models is selling of service and, for example, if a customer has hired the call center solution from *NICE inContact*, the customer is also able to hire a partner from NICE inContact to help them on every stage of the business – from the pre-sales to the implementation.

5.4.4 … have networks effect on data and infrastructure level m, but no on marketplace level

In terms of network effects is the *Descriptive Value Cluster* quite close to the average findings. More than half of them making use of network effects on data level, while no network effects on marketplace level is observed. This is no astonishing result as with the high amount of different, open public available data sources used, network effects on data are a frequent phenomenon. On the other side as none of the start-ups positions its offering as marketplace, network effects on marketplace level are unlikely to find.

[102] https://www.niceincontact.com/call-center-software/nice-incontact-cxsuccess

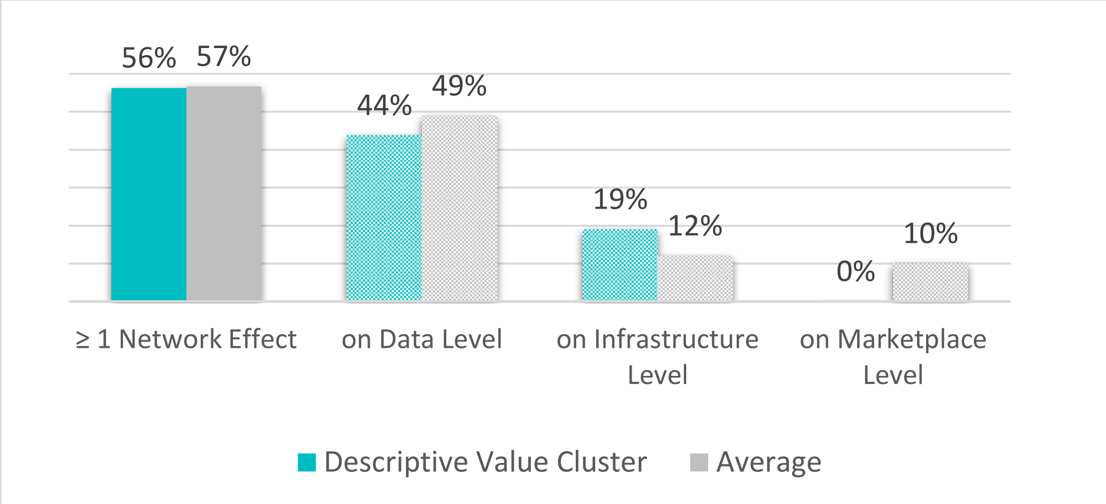


Figure 42 Network Effects - Descriptive Value Cluster

Network effects on infrastructure level are 50% higher than the average. Whenever companies are building own strength in connecting or integration others, network effects on infrastructure level becomes very likely.

> *PinMeTo*[103] is a Swedish online Software as a Service (SaaS) platform that makes sure multi-location brands and organisations can be found on multiple search, map, and social media platforms, including Apple, Facebook, Google, Instagram, Twitter, Foursquare and thousands of other apps and services. This consequently means that the more platforms are connected the more valuable the service gets, and that is exactly what network effects on infrastructure level is all about.

5.4.5 … are likely to receive funding

The amount of funding received by a start-up is a good KPI to measure its maturity, competitiveness and market potential. Is there significant larger amount of funding available for a specific group of start-ups, we can assume that the addressed market niche has promising potential.

As described in detail in Deliverable 2.6, for selecting our sample, we used the amount of funding as indicator for success, i.e. we wanted start-ups that already had their product validated, convinced venture capitalists about their business model, but were still on an early stage. Therefore, all our samples have received funding between US$2M and US$10M.

[103] https://www.pinmeto.com/

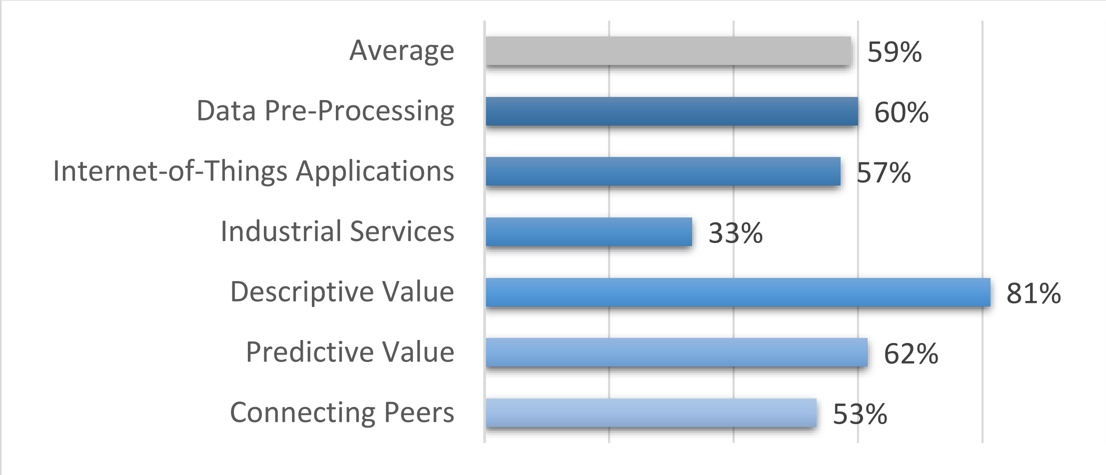


Figure 43 ≥ 3 Funding Rounds across Clusters

In terms of number of funding rounds, 59% of start-ups in average received at least three funding rounds. We have been also interested in whether the different clusters have been less or more attractive for investors. The answer is that the Descriptive Value Cluster was the most attractive cluster in terms of funding rounds, 81% of its start-ups participated in three or more funding rounds. Figure 43 gives an overview of the percentage of start-ups in each cluster that received more than three funding rounds.

So how to interpret those findings? The interpretation can be based on two aspects. First descriptive analytics is now available in a mature level so that it can be easily deployed in many different sectors (High technology readiness level). Secondly, the investors and market experts could identify high value for many different user groups leading to high market potential.

5.5 Start-ups in Cluster E: Predictive Value ….

All start-ups of the *Predictive Value Cluster* focus on predictive analytics, often being combined with other analytical values, such as descriptive, diagnostic or prescriptive. They are more likely to rely on personal data compared to the average, 50% more likely to use unstructured data, but tend to include only a smaller number of different types technologies into their offerings. For generation revenue, start-ups in this cluster rely 50% more often on asset sale and selling of services and less frequently on subscription when compared to the average. More details we will discuss in the following subsections.

5.5.1 … offer all predictive analytics

All start-ups to the *Predictive Value Cluster* of predictive insight for its user target group.

> For instance, *Desktop Genetics* is focusing on predictive analytics on genomics data and *DataRobot* is covering predictive machine models in their comprehensive machine learning API.

The predictive analytics is 23% of cases complemented with functionalities aiming to automate human tasks

> For instance, *Warwick Analytics*[104] is offering a machine learning-based tool to automatically label unstructured text sources that serve the basis to automate processes including predictions involving customer interactions, for instance in contact centres or when processing insurance claims.

5.5.2 … rely mainly on personal and unstructured but no industrial data

Start-ups of the *Predictive Value Cluster* tend to rely on most of the data sources a little bit below average, except for personal data and unstructured data sources.

An important characteristic of the predictive cluster is the *high usage of personal data*, which is 85% when compared to the average number of 62%. Therefore, we assume that companies in this cluster have a strategy to guarantee the safety of personal data, as requested by the GDPR.

> From our sample, *Visiblee* is helping B2B websites to become more effective by identifying anonymous visitors and converting them into clients. The SaaS solution collects IP addresses and cookies, which already mean personal data according to the GDPR. Using big data, machine learning and predictive analysis, they can identify in real time the unknown visitors and turn them into real clients.

[104] https://warwickanalytics.com/

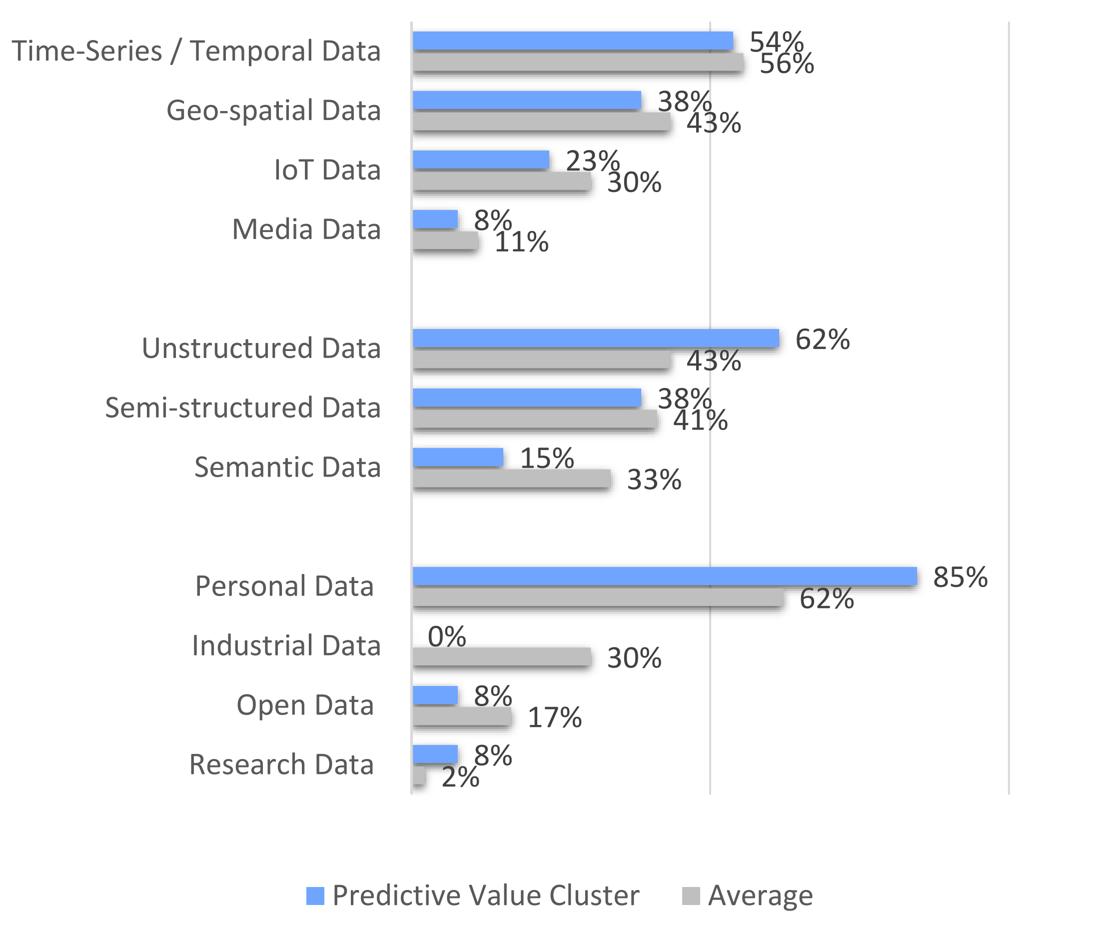


Figure 44 Type of Data - Predictive Value Cluster

A second characteristic of the *Predictive Value Cluster* is its high usage of unstructured data. While on average it is used 43% of the time, in the Predictive Value cluster it is used 62% of the time. When we analyze how this is distributed (see Figure 45), we see that is mostly video and image (23%) and text, language and audio (23%). In addition, this is the only cluster that is using genomics data (15%).

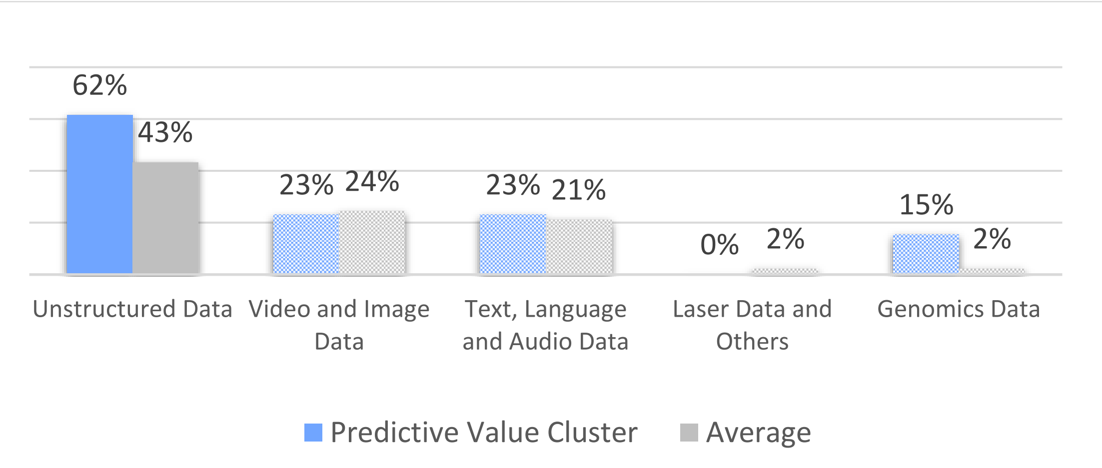


Figure 45 Unstructured Data: Predictive Value Cluster

Vivacity Labs[105] makes intelligent cameras to gather transport data, using machine learning and computer vision techniques. Their proprietary sensor hardware provides continuous and anonymous streams of vision data on all forms of urban mobility – from pedestrian to cyclist, car to HGV, and everything in-between. Using this data, businesses all over the UK are quantifying their operations, understanding travel patterns and justifying further investments.

The usage of *genetics data* is also outstanding for this cluster as it is the only cluster with occurrences. Genomics data refers to the genome and DNA data of an organism. They are used in bioinformatics for collecting, storing and processing the genomes of living things. Genomic data generally require a large amount of storage and purpose-built software to analyze.[106] In our sample, only two companies out of 90 used this type of data. In the predictive cluster only, this is equivalent to 15%.

The two companies using this type of data are *Desktop Genetics* and *Cyclica*. The former is building an AI application to re-engineer the human genome, helping researchers discover and treat the root genetic causes of human disease. The latter screens small-molecule drugs against repositories of structurally-characterized proteins or 'proteomes' to determine polypharmacological profiles. This means the goal is enhancing drug discovery by harnessing big data and predictive analytics, assessing the safety and efficacy of drugs.

And finally, this cluster is characterized by the absence of *industrial data*. Similar to the Descriptive Value Cluster (Section **¡Error! No se encuentra el origen de la referencia.**) and the Connecting Peers Cluster (Section **¡Error! No se encuentra el origen de la referencia.**) industrial data is used in none of offerings of the start-ups.

5.5.3 … rely on small number of different types of technologies

Start-ups in the *Predictive Value Cluster* are mainly relying on one type of technologies (see Figure 46). By relying on a smaller number of different types technologies, they are confronted with less effort in integration efforts, interfaces and partners.

---

[105] https://vivacitylabs.com/

[106] https://www.techopedia.com/definition/31247/genomic-data

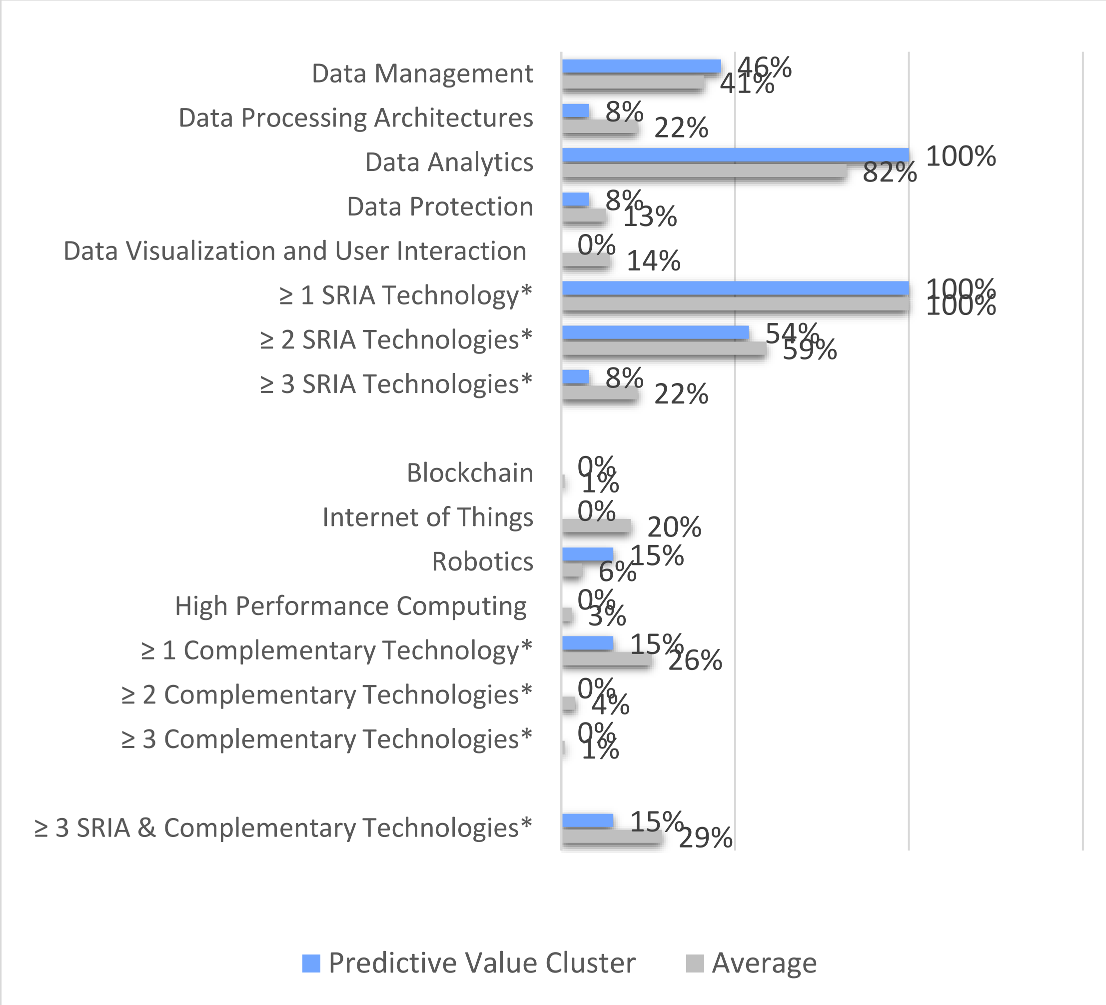


Figure 46 Type of Technology - Predictive Value Cluster

All startups in this cluster make use of data analytics and nearly every second provide innovative solutions in the area of data management by providing means for handling unstructured data sources.

For instance, *Cloudcherry* is applying two categories of BDV SRIA Technologies, they are data management and data analytics. Also, the start-up is disrupting the way organizations listen to the voice of customer offering out-of-the-box integrations with existing business applications, delivering predictive, prescriptive insights that enable business leaders to better understand customer needs.

The usage of more than three different types of SRIA Technologies is below average and observe only in 8% of the cases

For instance, *Carmera*[107] has built a real time change detection system to provide real-time 3D maps and navigation data for autonomous vehicles. Using mostly video as its data source, *Carmera*'s mission is to automate cities by kipping HD maps updated for leading OEMs and mobility-as-a-service operators, 100x faster and cheaper than ever possible before. This requires

---

[107] https://www.carmera.com/about/

robust data processing architectures, data management and data analytics, as data needs to be semantically annotated in digital formats, while creating interoperability standards and efficient technologies for the storage and exchange of semantic data and tools.

In relation to the usage of complementary technologies, we observe two opposing trends. On the one side, the usage of Robotics technologies is in this sector with 15% above average.

For example, *AuroRobotics*[108] is the Self-Driving Shuttle for travel within university campuses, corporate parks, and residential communities applying robotics. The solution is built on the basis of mechanical engineering and computer science, dealing with the design, construction, operation and use of robots as per their self-created autonomous driving shuttle.

However, as robotics is the only complementary technology start-ups in this cluster are integrating into their offering, the overall usage of complementary technologies is below average. In addition, the combined usage of more than 3 SRIA and complementary technologies is with 15% only half of the average, being the above-mentioned start-ups *Carmera* and *AuroRobotics* the examples.

5.5.4 … rely 50% more often on asset sale and selling of services

Start-ups in the *Predictive Value Cluster* rely on four different revenue models, subscription, asset sales, licensing and selling of service (see Figure 47). Until the moment the data was collected in 2018, freemium, advertisement, usage fee and commission fee were not observed.

[108] http://auro.ai/

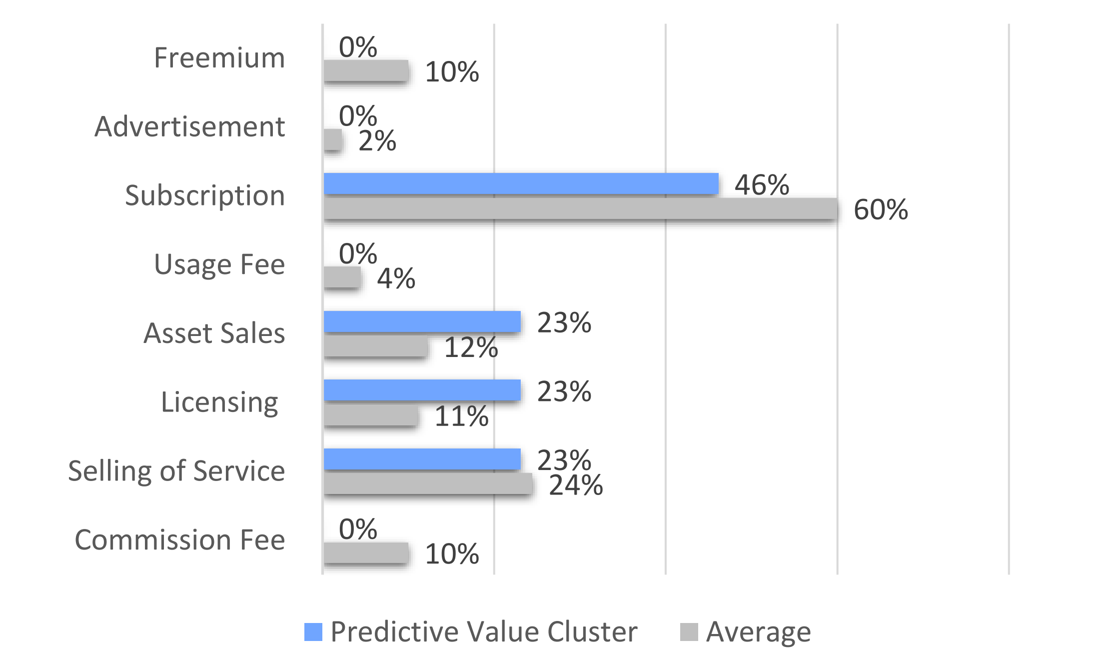


Figure 47 Revenue Model - Predictive Value Cluster

Two revenue models are the highlight of the predictive cluster, both above the average. *Licensing* and *asset sales* are observed in 23% of the cases, compared to the average of 11% and 12%, respectively.

> Another example applying licensing and asset sale is the start-up *Vivacity Labs*[109]. Their goal is to improve the process of cross vehicle traffic management in smart cities, and for that they built an intelligent sensor able to monitor up-to-the-minute city-wide transport status. They sell and install their V-shape proprietary hardware around a location and license their software that will fully classify road users, build a path analysis, predict traffic levels and in the future control traffic signals.

Asset sales are in the cluster observed twice as often than the average, with 23% compared to 12%. For developing accurate prediction models, the underlying data sources need to be comprehensive and complete. In many cases it is captured using an existing infrastructure, via mobile devices or with existing installed sensors. However, there are cases when relevant observations in the context of the subject of interest are not available. In order to ensure the access to high quality input data for such situations, specific hardware assets, such as sensors, cameras, etc., are installed to capture the missing data points. Whenever the required hardware is not available as off-the-shelf offering, they need to be newly designed and become then part of the overall offering.

In the *Predictive Value Cluster*, this happens in one every fifth company and might be an indicative of a very innovative solution built from scratch and tailored to a very

[109] https://vivacitylabs.com/

specific need. Below are some images of the hardware used by start-ups in the predictive cluster.

Startup	Hardware
UnaliWear *https://www.unaliwear.com*	Wearable technology that provides assistance for falls, medication reminders, and GPS location for its clients. Their mission is extending independence with dignity for millions of vulnerable people, and for that, they have built a watch with a patented battery system in the band that never needs to be removed for charging. The image below shows the device on the arm of a lady.  *Available at: https://www.unaliwear.com/about-unaliwear/*
VivacityLabs *https://vivacitylabs.com/*	Helping cities improve their transport network by providing insights through video analytics. VivacityLabs has built a proprietary solution composed by classification sensors and parking sensors that classify from pedestrian to cyclists in real time, allowing the prediction of a better transport system. The image below shows Vivacity's Lab hardware that is installed on lampposts.  *Available at: https://www.thesun.co.uk/motors/3572311/smart-traffic-lights-which-count-the-number-of-vehicles-on-the-road-to-launch-in-the-uk-and-they-will-give-priority-to-cyclists/*

Table 7 Two examples showing how Predictive Values start-ups are relying on hardware

The initial expectations in terms of revenue model for the predictive cluster was a high usage of the *subscription model*. However, the empirical data show that only 46% of the companies applied it, while the average was 61%.

> *Visiblee* is one of the start-ups that offer a subscription model, tailored to each customers' need. In Figure 48 the different types of subscription in accordance to the different need are displayed. They differ on number of user access, limitation of web traffic, leads nurturing, multi-source tracking possibilities, nominative identification, email retargeting, LinkedIn retargeting, API connection to CRM and dedicated account manager.

Figure 48 Pricing strategy of Visiblee[110]

5.5.5 …observe lower network effects when compared to others

For start-ups in the Predictive Value Cluster we observe a low percentage of network effects in general (see Figure 49). Networks effects on data level are 40% lower than average and network effects on marketplace level are not observed at all.

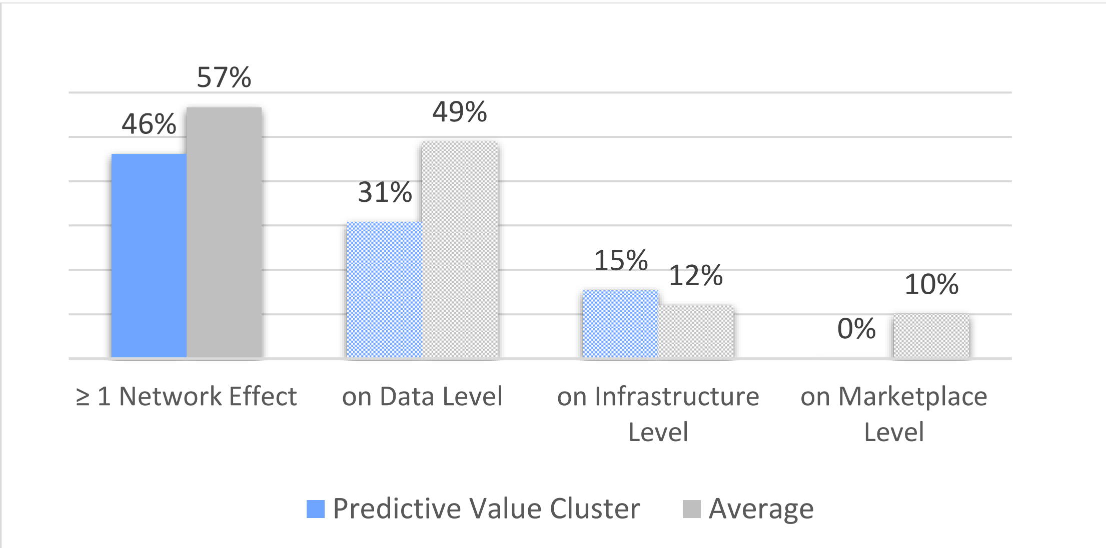


Figure 49 Network Effects - Predictive Value Cluster

[110] https://www.visiblee.io/en/pricing/

However, we observe that compared to average higher network effects on infrastructure level. Similar to Cluster C and D, start-ups in this cluster are also relying on connecting and integrating other applications, data sources, systems to build a competitive advantage.

> *CloudCherry* [111] a company of this cluster that was acquired by Cisco in 2019 is an example of a data-driven offering harnessing network effects on infrastructure level. The Experience Management platform offers an integration with existing business applications, delivering predictive and prescriptive insights that enable business leaders to better understand customer needs, and derive tangible business outcomes that can help drive true ROI. The more existing business application can be seamlessly be integrated without imposing extra efforts, the higher the accuracy of insights and predictions, the higher the value for business leaders.

5.5.6 … mainly focus on data-driven services

Data-driven services is for start-ups in the *Predictive Value Cluster* the most frequent option to position their offering on the market. They are using it with 92% above average (78%). Comparable pattern can be observed with start-ups from Cluster B and C.

> For example, the *Warwick Analytics* text-mining offering we discussed in the beginning of this section can seamlessly be integrated as additional functionality / service into other platforms, such as Intercom, Zendesk, Freshdesk or Salesforce.

[111] See https://www.crunchbase.com/organization/cloudcherry and https://www.cisco.com/c/en/us/products/contact-center/webex-experience-management/index.html?ccid=cc001647

5.6 Start-ups in Cluster F: Connecting Peers …

The core value proposition of start-ups of the *Connecting Peers Cluster* is the match-making functionality allowing to connect supply and demand from business to consumer side with three quarters of consumers being end-customers and the remaining being business customer. Start-ups in this sector are very likely to rely on commission fee, harness network effects on marketplace level and establish multi-sided markets / data-driven marketplaces. With 87% indicated the high percentage of personal data that also in B2B marketplaces personal data is used for implementing match-making algorithm. Match-making is the central functionality provided, its implementation relies beside personal data to a high percentage on semi-structured and semantic data. In the following subsections, we discuss those findings in further detail.

5.6.1 … are relying on match-making as central functionality

The match-making functionality is at the core of the Connecting Peers Cluster. It allows to map supply and demand, provider and consumer, etc. in accordance to predefined preferences, profile or meta-data. In Table 8 we are listing all start-ups in the cluster that are including match-making functionality as part of their offering.

Start-up	Sector	Match-making
Zizoo	Tourism & Sports	charter companies with end-user for Boat rental
Selectionist	Retail & Consumer Goods	print readers with online brands for buying
Shared2You	Retail & Consumer Goods	mobile movement data
CrossTarget	Retail & Consumer Goods	Mobile advertisements
Mila	Maintenance	Household with technical assistance
CleverTap	Retail & Consumer Goods	Engagement and analytics
Influenster	Retail & Consumer Goods	Informed purchasing based on reviews
Hooch	Media & Entertainment	Venue owner with end-user for Subscription cocktail app
Insurify	Insurance & Financial Services	Insurance provider with end-user for Car insurance contracts
Paysa	Human Resources	Salary information
LoanTab	Insurance & Financial Services	Loan money
BridgeU	Education	Education provider with students for personalizing curriculas
Bird.i	Earth Observation & Geospatial & Space	Satellite imagery

Table 8 Overview of Connecting Peers' start-up relying match-making functionality

100% of the start-ups in the connecting peers cluster have a clear *sector focus*, confirming our initial expectations that matching supply and demand would require a very specific topic to be explored. From connecting boat rental to customers, to a drinking membership, and to dating apps, the start-ups in these cluster are very diverse on their offerings but always with a clear customer target groups in mind.

As discussed before majority match-making algorithm rely on different types of data analytics, machine learning, and data-preprocessing. We have been interested to which extend start-ups in the connecting peers Cluster also provide some sort of data analytics as independent functionality.

> For instance, the US-based company *Hooch*[112] offered when we first analysed it in 2018 a subscription-only cocktail app and discovery platform, where customers are offered a free drink each day. The platform aligns the needs and interests of two different user groups in an efficient manner. End-user would get a classpass for drinks allowing them to get one free drink each day at over 400+ top bars and restaurants. Registered venue owners would use Hooch as free marketing tool that attracts people to their venues. The Hooch app was based on a simple search functionality allowing user to find interesting bars or venues. Complementary to this match-making functionality, Hooch was collecting large scale real-time data on the millennials[113] behaviour, drink preferences and spirits consumptions to identify spirits trends and other related patterns. The generated insights are offered to alcohol industry as service[114].

From our sample, we have observed that only 13% of the Connecting Peers companies offer some type of data analytics as complementary part of their match-making functionality.

5.6.2 … wide range of data sets are used for match-making

A wide range of different data types is observed in the connecting peers cluster, with exception to industrial and open data.

[112] Note that the founder of Hooch continuously reinvented, explored and relaunched their initial idea. Today their company is offering a "decentralized model for consumer rewards". See Hooch is https://techcrunch.com/2019/03/07/hooch-rewards and https://hoochrewards.com/

[113] Millennial is a person reaching young adulthood in the early 21st century.

[114] It is not explicitly mentioned how the revenue stream is generated for this scenario

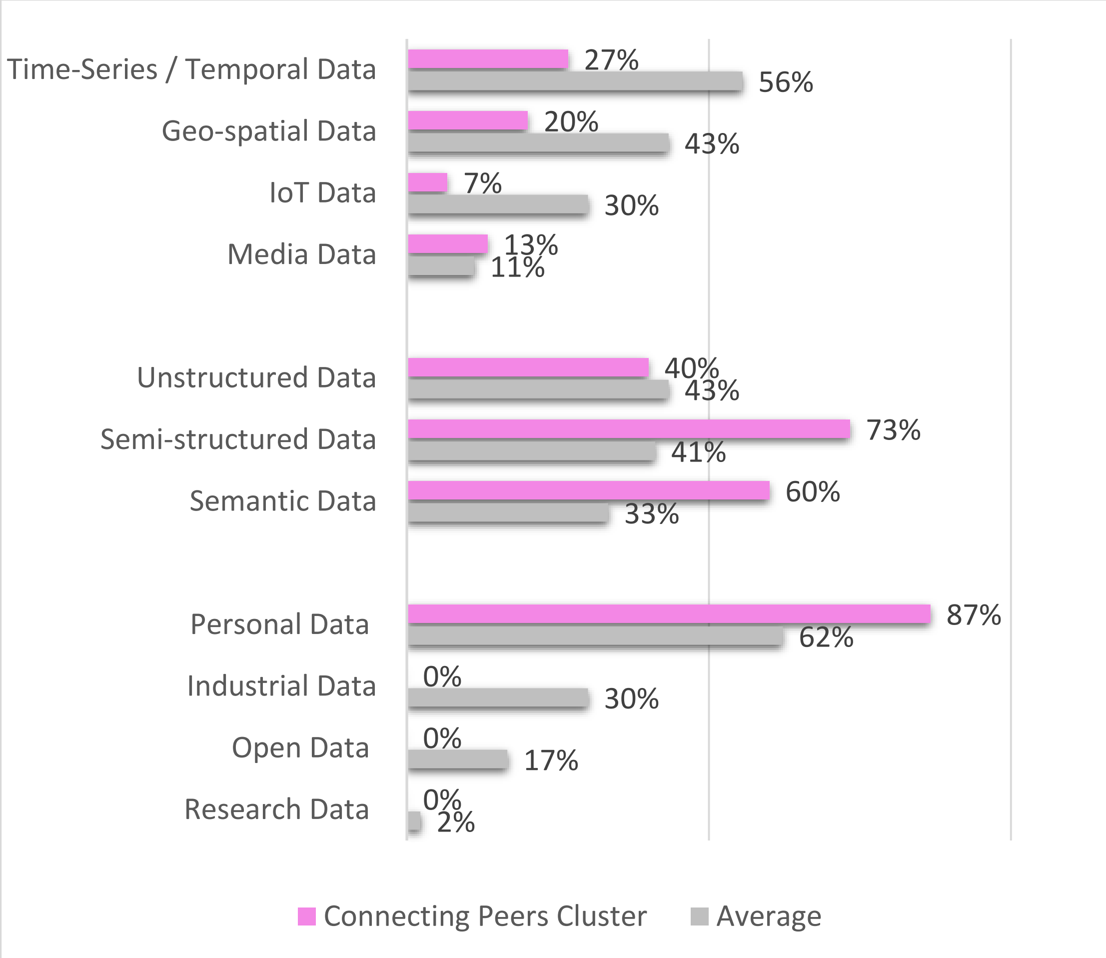


Figure 50 Type of Data - Connecting Peers Cluster

In the context of match-making the usage of personal data is of high relevance. Majority of match-making mechanism are based on personal profiles to find good matches

> For instance, *BridgeU*[115] offers to connect students with education providers such as school for enabling personalized curricula planning and organization. Students are requested to build a personal profile, incorporating data on subject interests, grades, social, professional and learning preferences. The *BridgeU* match-making algorithm relies on historical admission data and produces personalized university and course recommendations in accordance to the students' preferences and strengths. [116]

For that reason, Connecting Peers is the cluster with the highest frequency of using Personal Data, observed 87% of the time against the average of 62%.

To complement personal data, a wide range of other data sources are used as input for advancing match-making algorithms. In particular, the usage of semi-structured data with 73% and semantic with 60% is clearly above average. Also, 40% of the

[115] https://bridge-u.com/

[116] https://techcrunch.com/2017/10/17/bridgeu-raises-5-3m-to-close-the-gap-between-education-and-industry-needs/

companies use unstructured data as input data. On the other hand, industrial and open data are not observed.

> For instance, *Selectionnist*[117] is relying on image recognition technology to bridge online and offline world. By making images of well-known brand products in print magazines with the *Selectionnist* mobile app, print readers are directed to the brands' online shop offering all product details.

5.6.3 … are very often relying on commission fee

The *Connecting Peers Cluster* is the only one relying on commission fee, observed in 60% of the time against an average of 10%. This is therefore a characteristic feature of this cluster as no other cluster makes use of this revenue model.

> For instance, *LoanTap*[118] is an Indian-based online platform committed to deliver customized loan products to millennials. They offer tailor made loans for different solutions. In exchange, a commission fee is applied proportionally to the offered loan

From Figure 51 we can see that other types of revenue models are also applied in the Connecting Peers Cluster. subscription in 53%.

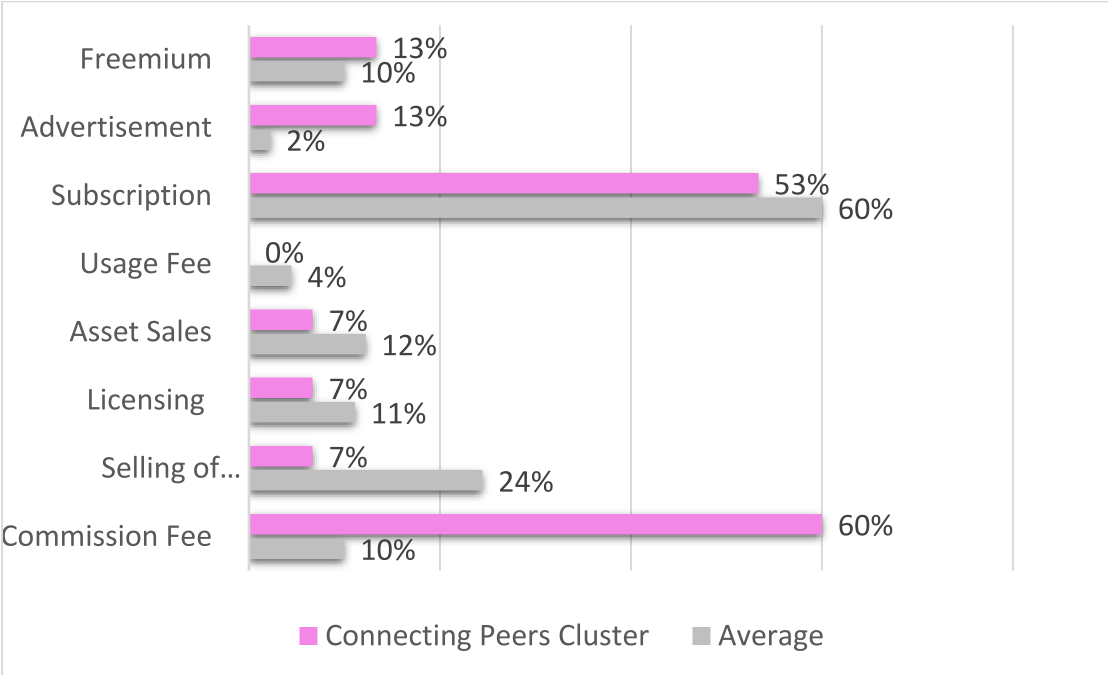


Figure 51 Revenue Model - Connecting Peers Cluster

Freemium is observed in 13% of the time and advertisement only occurs in this cluster with a frequency of 2%.

[117] https://www.crunchbase.com/organization/selectionnist

[118] https://loantap.in/

For instance, *Bird.i*[119] used to included freemium as part of their revenue strategy. *Bird.i*'s objective is to simplify the access to satellite imagery data for business user that aim to derive insights out of those images. *Bird.i* makes the satellite imagery accessible and affordable through a unique monthly subscription to their Portal or API. To allow new user to get familiar with their services, in 2018 they were still offering a freemium option. In 2020, this option was no longer available.

Subscription is offered in 53% of the cases, but below an average of 61%.

For, instance *Bird.i* is one of the start-ups offering monthly subscription for their Image Service.

Another example is the *Hooch* platform which offers a a subscription-only cocktail app and discovery platform, where customers are offered a free drink each day. For a monthly subscription fee ($9.99), people can get one free drink each day at over 400+ top bars and restaurants.

5.6.4 … have strong network effects on data and marketplace level

For digital and data-driven innovations, network effects are important phenomena to be reflected. A network effect occurs when a product or a service becomes more valuable to its users as more people use it (Shapiro and Varian, 1999). Network effects are also known as demand-side economics of scale and predominately exists in areas where networks are of importance, such as online social networks or online dating sites. A social network or dating site is more appealing to its user, the more users of interest it has. In consequence, harnessing network effects require developing a wider network of users in order to differentiate from competitors. For that reason, the critical mass of user and timing are key success factors in a network economy.

Network effects are highly observed in the *Connecting Peers Cluster* when compared to others. 93% of start-ups observe network effects, compared to 57% in average.

On Marketplace level is seen in 60% of the time, on data level in 87% of the time and on infrastructure level in 7% of the time

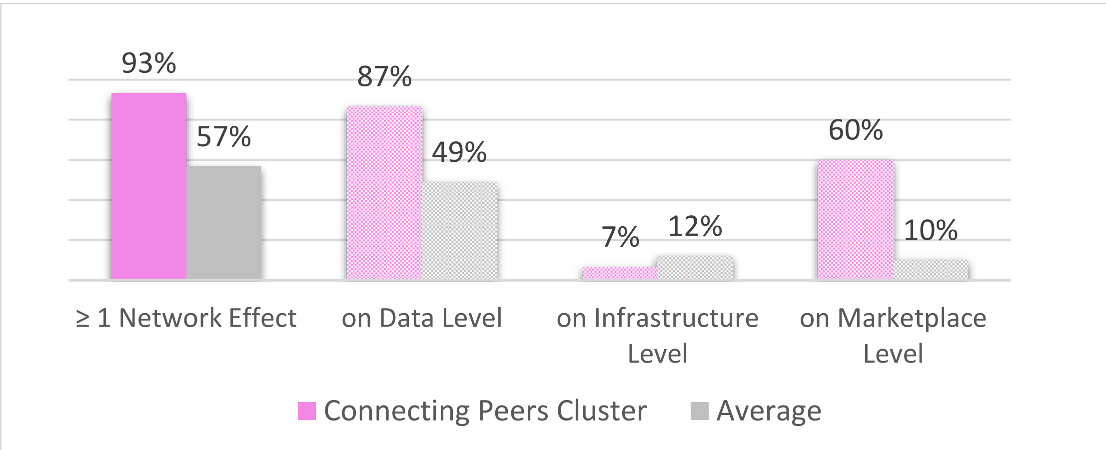


Figure 52 Network Effect - Connecting Peers Cluster

[119] https://hibirdi.com/

Network effects on marketplace level means that the more suppliers and users are on the market, the more interesting the value proposition becomes.

> *Hooch*[120] is a nice example for this case. When we started our research, their business model was offering a subscription-only cocktail app and discovery platform, where customers are offered a free drink each day. The app is based on a simple search functionality allowing user to find interesting bars or venues. In addition, Hooch is collecting large scale real-time data on the millennial's behavior, drink preferences and spirits consumptions to identify spirits trends and other related patterns. The generated insights are offered to alcohol industry as service. Nowadays they have changed their offering to a rewards business model. By connecting a credit card to the HOOCH app, every purchase a user made is transformed into TAP Dollars, their reward coin that rewards instantly. The user is then able to use it in a network of more than 250.000 partners. This means the company is still a marketplace but have changed its focused. It might also be an indicative that marketplaces can switch their businesses using the network they have built on a later stage.

On *data level* we observe that 87% of the start-ups are improving their value proposition by an ever increase usage of data.

> For instance, the French company *Selectionist* ia offering image recognition technology with the goal to connect readers of print journals with the world's largest brands through an application or a chatbot. They aim to bridge the gap between offline content and online experience by offering an advanced match-making service to connect consumer and brands. Their match-making algorithm is based on image recognition technology that continuously improves the more images from brands' product are in their data bases (more brands) as well as the more user request they receive. Thus, their offering is based on network effects on data level. The service is conceptualized as marketplace based on commission fee and with network effects on marketplace-level.

Network effect on infrastructure level has very low representation in the Connecting Peers sector. Only one start-up was classified accordingly. While there is an high overlap between companies focusing on network effects on data level and network effects on marketplace level, it seems that network effects on infrastructure and marketplace level are rather disjunct or opposing strategies, or simply two strategies that do not bring any synergies. While network-effects on marketplace level aims to connect people, network-effects on infrastructure level is about identifying new interaction that orchestrate users and resources in the ecosystem. By bringing forward methods and tools that allow other to connect their resources or access available resources to generate new value.

[120] https://www.crunchbase.com/organization/hooch

From our sample, *Zizoo*[121] is the only start-up in the Connecting Peers Cluster with network effects on infrastructure level. The company is building a data-driven marketplace for boat rental. The concept follows the vision to develop a *booking.com but for boats*. The challenges the founder *Zizoo* were facing when they started their business was the fact that the boat rental industry – although being with expected EUR 33 bn a very large market – was not yet digitalized. In consequence, charter companies were not able to connect with the platform about available boats as in the majority cases this information was only available in analog manner or self-made excel formats. In order to empower the charter companies to participate in the *Zizoo* marketplace they required support in moving to the digitalized world. The company well understand their needs and how to attract the provider of their marketplace by offering them a powerful inventory management tool and business intelligence for free. The more charter companies are connecting to their platform, the more time and efforts they can spend in extending and improving the set of tools for charter companies to manage their fleet of boats, the more attractive the marketplace becomes for them.

5.6.5 … are mainly as data-driven marketplace

80% of the startups in the connecting peers cluster are data-driven marketplace, against an average of 16%. This feature really defines the cluster and have been discussed in detail already in Section

To conclude this section, let’s have a look at *Mila*[122] which is a further start-up representing this cluster and highlights that there are no limitations a) in terms of who will be connected with whom as well as b) interaction with the physical world

Mila is a crowd services platform where users can find, book, and rate tech-savvy people or offer their technical know-how. For example, if a person doesn’t know how to install a Smart TV, he or she can find a technical specialist on *Mila*’s platform and book an appointment.

[121] https://www.zizoo.com/

[122] https://www.mila.com/en-de/

6 Example of Cluster Representatives

In the following section, we provide some examples of success stories of data-driven start-ups along the six clusters to give the reader an impression how clear and focused their supply and demand side can be pitched along the eight dimensions of the DDI Framework. Just ot recap. The dimension of the DDI Framework are divided into the supply and demand side of the offering.

On the *supply side* the focus is on the development of new offerings. For a clearly defined value proposition, this includes the identification of and access to required *data* sources, as well as the analysis of underlying *technologies.* On the *demand side* the focus is on the dynamics of the addressed markets and associated ecosystems. The analysis includes the development of a *revenue strategy*, a way forward of how to harness *network effects* as well as an understanding of the *type of business*. As data-driven innovations are never done in isolation, the identification and analysis of potential development *partners* as well as partners in the *ecosystem* help to align / balance the supply and demand aspects in a way that its competitive nature will stand out.

We want to highlight here that due to graphical limitations, the presentation of the data-driven offerings along the DDI Framework is a linear manner. However, the way of exploring the various options for each dimension as well as the interrelations between dimension is a very iterative and explorative one.

6.1 Cluster A: Pre-Processing Technologies

TRACTABLE

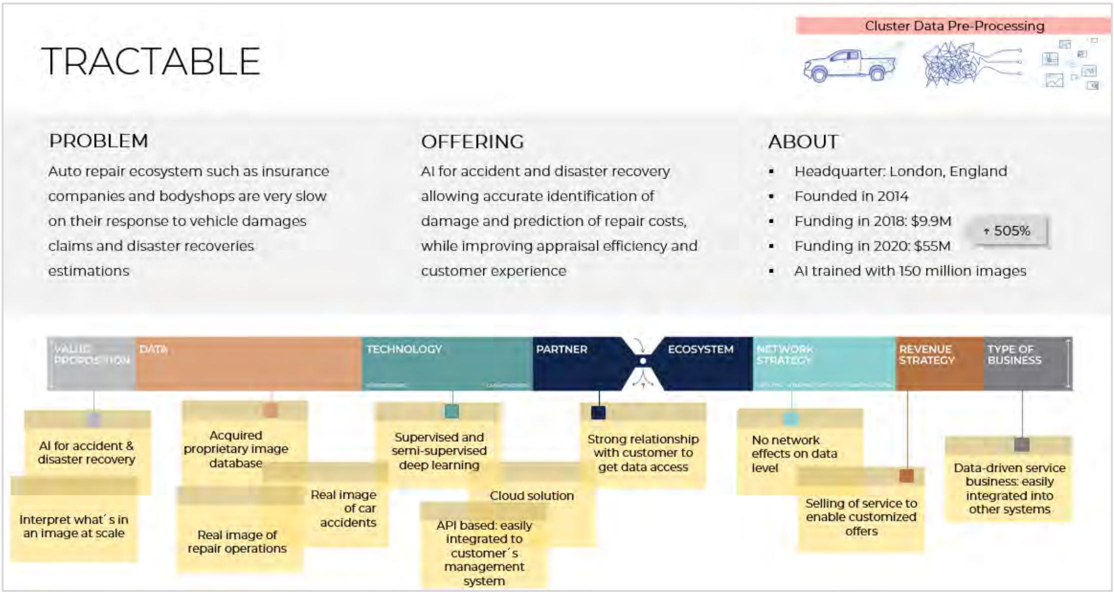


Figure 53 DDI Summary for Tractable

Tractable is a London-Based start-up founded in 2014 that uses Artificial Intelligence to look at damages and predict repair costs. Through deep learning tools, a visual inspection was initially performed on cars and more recently on medical imagery. The goal is to perform faster, cheaper and more accurately than humans, while offering a near zero marginal cost.

As a representative of the pre-processing cluster, *Tractable* has a high focus on the pre-processing of heterogeneous data sources. They rely heavily on the pre-processing of unstructured data and, in this case, image is the most used data source. The solution is cloud and API-based, meaning that can be easily integrated to the customer´s management system. In addition, the company does not focus on analytics, match-making or process automation and, therefore, no descriptive or prescriptive analytics is observed.

Problem	*Auto repair ecosystem such as insurance companies and bodyshops are very slow on their response to vehicle damages claims and disaster recoveries estimations*
Where	London, England
Founded	2014
Funding 2018	$9.9M
Funding 2020	$55M
Funding Δ	↑ 505%
Acquired?	No
Outstanding info	AI trained with 150 million images[123]
Cluster Rational	*Cluster Pre-Processing Technologies* *Tractable has a high focus on the pre-processing of heterogeneous data sources. They rely heavily on the pre-processing of unstructured data and, in this case, image is the most used data source*
Value Proposition	Insurance companies and bodyshops are able to interpret what is in an image at scale using AI for accident and disaster recovery Image understanding is their main data value. They do not focus on analytics, match-making or process automation and, therefore, no descriptive or prescriptive analytics is observed.
Data	Visual data such as real image of car accidents and repair operations, Tractable initially acquired a proprietary data asset[124]
Technology	Supervised and semi-supervised deep learning that capture the delicacy of real damaged vehicles and properties. The solution is cloud and API based, meaning that can be easily integrated to the customer´s management system
Partner	Tractable builds strong relationship with all their customers to get access customer's image data
Ecosystem	No information available

[123] Tractable has initially acquired a proprietary data assets and in the following set up a strong network with their customers to get access to their image data (see https://www.credit-suisse.com/microsites/pic/en/4th-private-innovation-circle/tractable.html)

[124] see https://www.credit-suisse.com/microsites/pic/en/4th-private-innovation-circle/tractable.html)

Network Strategy	No network effects on data level, they have a lot of methods in place to increase data quality
Revenue Strategy	Selling of service model; customer must get in touch with sales team to have a personalized offer
Type-of Business	Data-driven Service business as their offering can easily integrated into other systems

ARTOMATIX

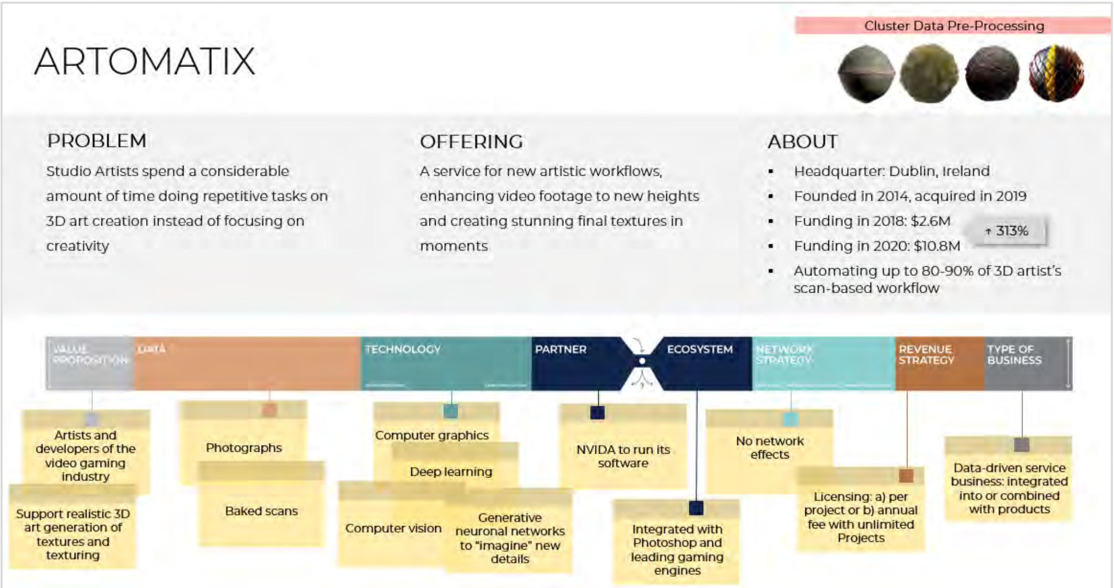


Figure 54 DDI Summary for Artomatix

Problem	*Studio Artists spend a considerable amount of time doing repetitive tasks on 3D art creation instead of focusing on creativity*
Where	Dublin, Ireland
Founded	2014
Funding 2018	$2.6M
Funding 2020	$10.8M
Funding Δ	↑ 313%
Acquired?	Yes, by Unity Technologies
Outstanding info	Automating up to 80-90% of the time-consuming and expensive parts of 3D artist's scan-based workflow. [125]
Cluster Rational	*Cluster Pre-Processing Technologies* *Artomatix's central focus is on delivering solutions for pre-processing images*
Value Proposition	*Artomatix*'s users are **artists** and **developers** of the video gaming industry that can benefit from a service that supports the realistic 3D art generation of textures and texturing. Before this

[125] https://artomatix.com/press/

	tedious task was done manually. With the suite of tools provided by *Artomatix*, artists can now do the same task 10 times faster.
Data	The data used for training and developing the algorithm are images such as photographs or baked scans.
Technology	The technology is based on computer graphics, Deep Learning and computer vision. It uses generative neuronal networks to "imagine" new details of a texture in a way a human would do, i.e. it recognized objects in a video, can add texture and features automatically by relying on the "learned" knowledge what should be there.
Partner	Artomatix partners with NVIDA to run its software on NVIDIAs graphic cards
Ecosystem	The software can be integrated with Photoshop and leading gaming engines like Unity and Unreal. In addition, Artomatix collaborates with several companies that adopt the solution *ArtEngine* as part of their offering, e.g. Unity to automate seam removal and general scan cleaning processes. In addition, Artomatix partners with other companies to co-develop new solutions, for instance with Dell new technology solution to enable studios everywhere to be more effective. [126]
Network Strategy	There are no network effects that need to be reflected
Revenue Strategy	The revenue model has changed over the past two years. *In 2018:* The company used three different Subscription models (Indie (Revenue < $100k/year), Professional (Revenue < $1M/year) and Enterprise (Revenue > $1M/year). Enterprises can license Artomatix's technology and build them into their existing process for an annual fee *In 2020:* The company licenses Artomatix's technology by a single project of any duration or by an annual fee with unlimited projects
Type-of-Business	The technology is offered as data-driven service which can be integrated into or combined with products

[126] More examples at: https://artomatix.com/partners/

6.2 Cluster B: Internet of Things Applications

SENSEYE

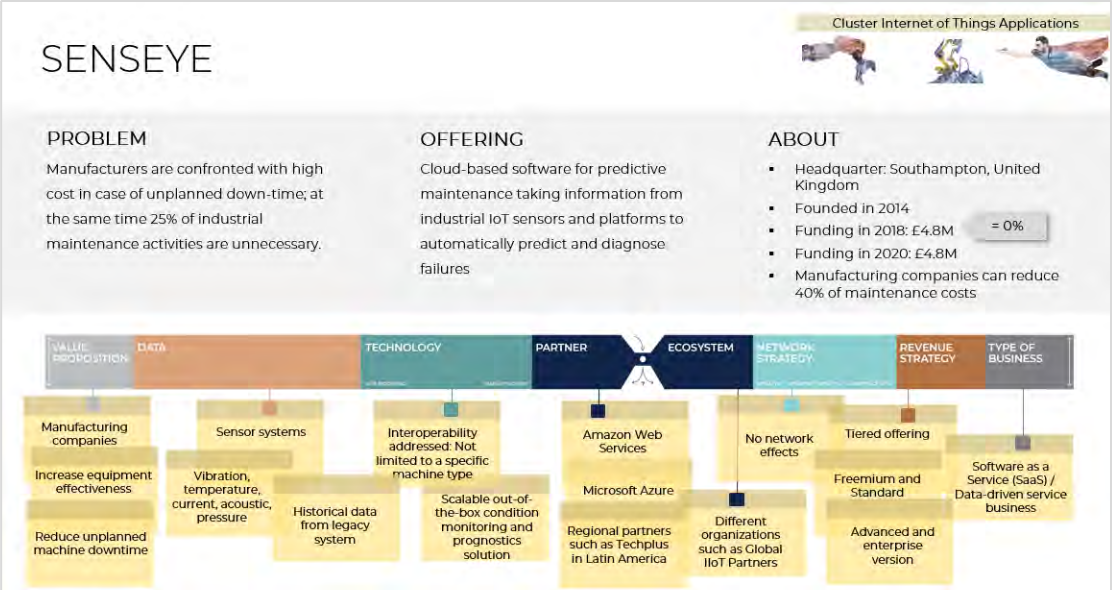


Figure 55 DDI Summary for Senseye

Senseye is a cloud-based software for predictive maintenance taking information from industrial IoT sensors and platforms to automatically diagnose failures. It helps manufacturers avoid downtime and save money by automatically forecasting machine failure without the need for expert manual analysis. While this affects many industries, in manufacturing over 25% of maintenance activities are unnecessary and produce additional failure risks. Further, 24% of the total manufacturing costs are attributed to downtime.

Problem	*Manufacturers are confronted with high cost in case of unplanned down-time; at the same time 25% of industrial maintenance activities are unnecessary.*
Where	Southampton, United Kingdom
Founded	2014
Funding 2018	£4.8M
Funding 2020	£4.8M
Funding Δ	= 0%
Acquired?	No
Outstanding info	Manufacturing companies can reduce 40% of maintenance costs[127]. *Senseye* make complex aerospace inspired technologies available to the broader market.
Cluster Rational	*Cluster Internet of Things Applications* *Senseye integrates IoT technology with multiple different types of technologies as well different types of data analytics. In addition,*

[127] https://www.senseye.io/

	they offer mean for data and technology integration as central aspect of their solution
Value Proposition	*Senseye* is a cloud-based software for predictive maintenance used by **manufacturing companies** across the globe. The solution takes information from industrial IoT sensors and platforms to automatically diagnose failures. Their data analytics solution reduces unplanned machine downtime (predictive analytics) and increases overall equipment effectiveness (optimization, prescriptive analytics)
Data	*Senseye* rely on different types of data sources covering condition indicators ranging from vibration, temperature, current, acoustic, pressure and other sensor systems, either captured by the machine PLC or retrofitted. In addition, historical data can be integrated into their platform
Technology	*Senseye* rely on different types of data analytics and on data & technology integration. Their solution is not limited to a specific machine type or require a team of engineers or data scientists to interpret the data. It is a scalable out-of-the-box condition monitoring and prognostics solution that works across all machine's types. It requires minimal deployment or configuration and provide an immediate 'time to first insight'
Partner	*Senseye* has an extensive list of Industrial and Technology partners. Just to mention a few, their cloud partners are Amazon Web Services and Microsoft Azure. In different parts of the world they count on regional partners, such as Techplus in Latin America and Rhino Assembly in the USA.[128]
Ecosystem	*Senseye* is part of an ecosystem formed by different organizations in areas such as Computerized Maintenance Management System, Maintenance Management System, Cloud Partners, Global IIoT Partners, IIoT Infrastructure Partners, Regional Partners, Trade Organization Partners and Educational Partners.[129]
Network Strategy	No network effects
Revenue Strategy	Available as a tiered offering: freemium, standard, advanced and enterprise version
Type-of-Business	Offered as Software as a Service (SaaS) / Data-driven service business

[128] More information on: https://www.senseye.io/our-partners/#1572623401971-cadaee85-ecac

[129] More information on: https://www.senseye.io/our-partners/#1572623401971-cadaee85-ecac

ARABLE

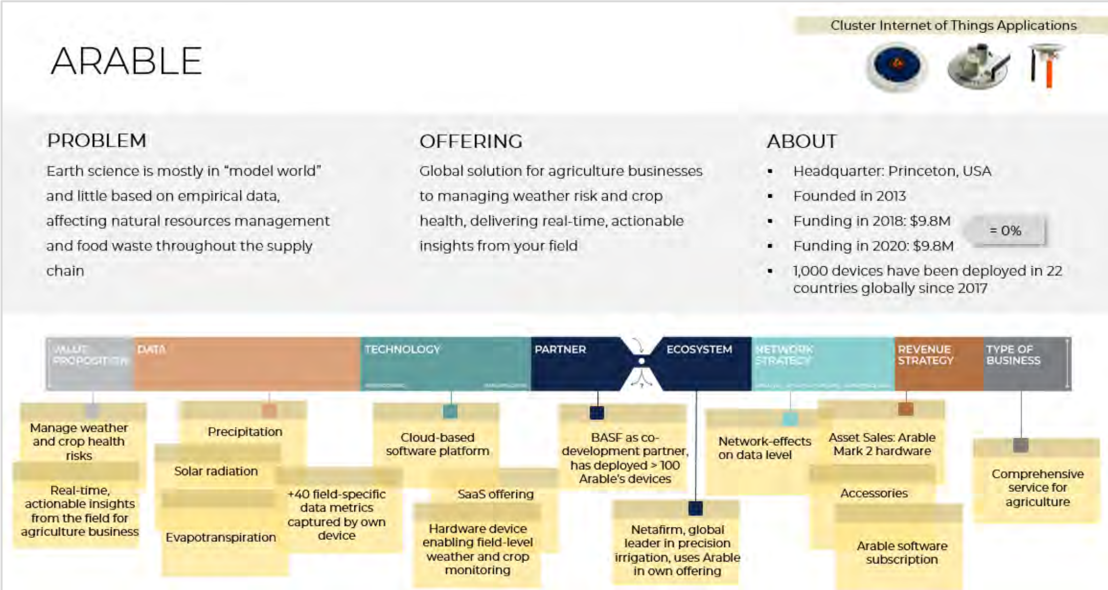


Figure 56 DDI Summary for Arable

Arable is an US-based company founded in 2013 offering agriculture businesses a global solution for managing weather and crop health risks, delivering real-time, actionable insights from the field.

Problem	*Earth science is mostly in "model world" and little based on empirical data, affecting natural resources management and food waste throughout the supply chain*
Where	Princeton, USA
Founded	2013
Funding 2018	$9.8M
Funding 2020	$9.8M[130]
Funding Δ	= 0%
Acquired?	No
Outstanding info	1,000 devices have been deployed in 22 countries globally since 2017[131]
Cluster Rational	*Cluster Internet of Things Application* *Arable integrates IoT technology with multiple types of data analytics. In addition, data and technology integration is a central aspect of their solution.*
Value Proposition	*Arable*'s target users are growers, advisors, and businesses who aim to play a proactive role in the quality and longevity of their operations. The company offers a global solution to agriculture

[130] Information about funding amount was not consistent: Forbes indicates that Another source says that by April 2019 Arable has raised $13.2 million including from grants. (see https://thriveagrifood.com/tag/forbes/)

[131] https://thriveagrifood.com/tag/forbes/

	businesses for managing weather and crop health risks, delivering real-time, actionable insights from the field Their data analytics solution allows the continuous monitoring and predictive analytics to improve agriculture operation.
Data	The data used is captured by their own field-level weather and crop monitoring devices that collect over 40 field-specific data metrics, such as precipitation, solar radiation and evapotranspiration
Technology	To enable the access to data from anywhere in real time, a cloud-based software platform based on a tiered SaaS offering (different level of services) is combined with IoT hardware. Their solution also encompasses a hardware device enabling field-level weather and crop monitoring.
Partner	BASF[132] is also a co-innovation partner of Arable on their mission of understanding the nuanced effects that certain chemical combinations and weather events can have on crop outcomes, as well as on the farmer's bottom line. BASF deployed hundreds of Arable Mark devices across a variety of representative microclimates to collect field-level measurements in the last three years
Ecosystem	*Netafirm*, the global leader in precision irrigation, is a partner of Arable and integrates *Arable*'s data and services into Netafim's automated irrigation and fertigation cloud-based platform.[133]
Network strategy	As the prediction service improves with more data available, the solution of Arable is based on network-effects on data level.
Revenue strategy	In 2018 *Arable* is selling licenses for enterprise software to agribusinesses In 2020[134]: The current revenue model of *Arable* is based on three main revenue streams: asset sales via the Arable Mark 2 hardware, as well as a vast range of accessories such as Arable Solar and *Arable* ground anchor. In addition, subscription is offered via the *Arable* software subscription modality and Arable API 2.0
Type-of-Business	Data-driven service Business

[132] https://www.arable.com/2019/12/04/arable-and-basf-team-up-in-xarvio-field-manager-to-reduce-pesticides/

[133] https://www.arable.com/2020/01/31/netafim-arable-how-high-quality-data-enables-precision-irrigation-easily-affordably-and-globally/

[134] https://shop.arable.com/

6.3 Cluster C: Industrial Services

PLUTOSHIFT

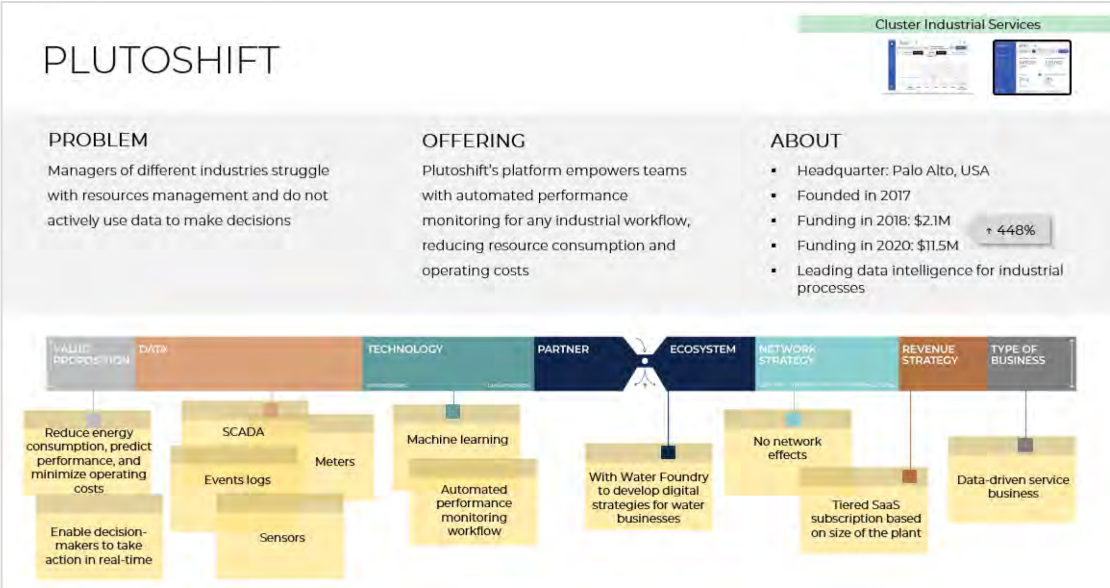


Figure 57 DDI Summary for Plutoshift

Problem	*Managers of different industries struggle with resources management and do not actively use data to make decisions*
Where	Palo Alto, USA
Founded	2017
Funding 2018	$2.1M
Funding 2020	$11.5M
Funding Δ	↑ 448%
Acquired?	No
Outstanding info	It is the leader in data intelligence for industrial processes.[135]
Cluster Rational	*Cluster Industrial Services* *Plutoshift covers the whole range of data analytics as well as provide high value through process automation. Their data sources are mainly industrial data*
Value Proposition	Intelligence platform that enables production and operating facilities to reduce energy consumption, predict performance, and minimize operating costs in wide range of sectors, such as Food & Beverage, Oil & Gas, Power & Renewables, Chemical and Manufacturing.[136]

[135] https://www.businesswire.com/news/home/20200114005292/en/Plutoshift-Continues-Momentum-Key-Additions-Leadership-Team

[136] The value proposition has changed with time. Now they are a full resource management platform for specific industries

	With the collected data, the solution enables decision-makers to take action in real-time to proactively solve the targeted problems. For instance, the *Plutoshift* analytics platform presents managers with a dashboard that quantifies the status of all assets at a given water treatment plant. These ratings, ranging from 0 to 100, take into account temperature and pressure readings in addition to other data from pumps and chlorinators to identify cause and effect relationships
Data	Data gathered from various internal and external sources, such as SCADA, meters, and event logs, and sensors. By utilizing historical data, Pluto can directly recommend steps to improve the functioning of plant infrastructure. This information is prescribed with assigned priority levels so that a human can step in and take the appropriate action.
Technology	*Plutoshift*'s analytics platform is backboned by machine learning, enabling the ingestion of large quantities of unstructured data, and therefore allowing an automated performance monitoring workflow. It is connected to the already existing data sources and with no need to touch the backend or go through an implementation cycle.
Partner	No information
Ecosystem	*Plutoshift* has established a partnership with Water Foundry, a global advisor in solving water-related challenges, to develop and implement digital strategies for businesses that rely on water[137]
Network Strategy	No network effects
Revenue Strategy	Tiered SaaS subscription model based on the size of the plant
Type of Business	Data-driven service business

[137] https://www.innovationews.com/Plutoshift-and-Water-Foundry-partner-to-bring-digital-transformation-to-businesses-that-rely-on-water/

FRAUGSTER

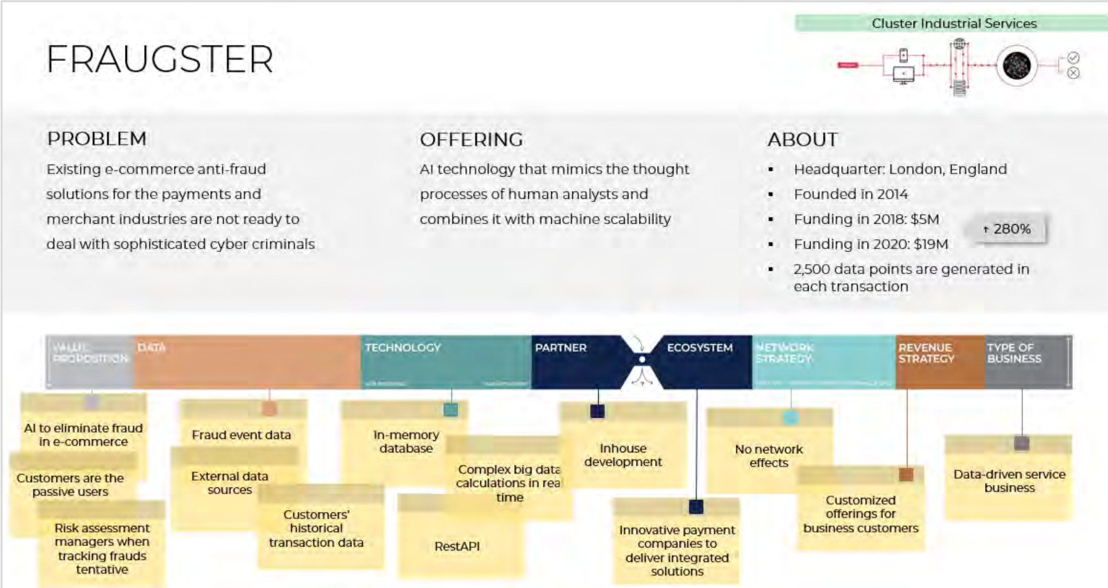


Figure 58 DDI Summary for Fraugster

Fraugster is a German-Israeli anti-fraud company that uses artificial intelligence to eliminate fraud and increase customers' profits on their e-commerce businesses. *Fraugster*'s proprietary technology allows the team to use state-of-the-art big-data analytics to pinpoint threats and isolate them from the rest of the user population.

Problem	*Existing e-commerce anti-fraud solutions for the payments and merchant industries are not ready to deal with sophisticated cyber criminals*
Where	London, England
Founded	2014
Funding 2018	$5M
Funding 2020	$19M
Funding Δ	↑ 280%
Acquired?	No
Outstanding info	They generate around 2,500 data points to each transaction, a considerable increase of information compared to the normal 60 data points[138]
Cluster Rational	*Cluster Industrial Services* *Fraugster covers the whole range of data analytics as well as provide high value through process automation. Their data sources are mainly industrial data, e.g. operational data*
Value Proposition	Anti-fraud company that uses artificial intelligence to eliminate fraud and increase customers' profits on their e-commerce businesses. **Customers** are the passive users when performing a

[138] https://ki-berlin.de/en/blog/article/fraugster-anti-fraud-the-story-makes-the-difference/

	payment in the e-commerce platform. Second user group are **risk assessment managers** when tracking frauds tentative
Data	Fraud event data is complemented with external data sources. Customers' historical transaction data
Technology	Their core proprietary technology is a massive parallel in-memory database, designed for extremely complex big data calculations in real time. Via a simple restAPI, an incoming event is received by Fraugster and complemented with external data sources. The process continues by looking for statistical twins within their records, both from the client's historical data as well as our overall data-pool. The event is compared then only to the relevant cluster of historical data (which grants far better and more accurate results than from comparing to the entire dataset) and returns the final fraud probability in a score format.
Partner	No information provided
Ecosystem	*Fraugster* works with the most innovative payment companies, such as Ingenico, Natixis, Worldline and RatePay, to deliver integrated solutions
Network strategy	No Network Effects
Revenue Strategy	No information provided
Type-of Business	Data-driven Service Business

6.4 Cluster D: Descriptive Value

KEYWEE[139]

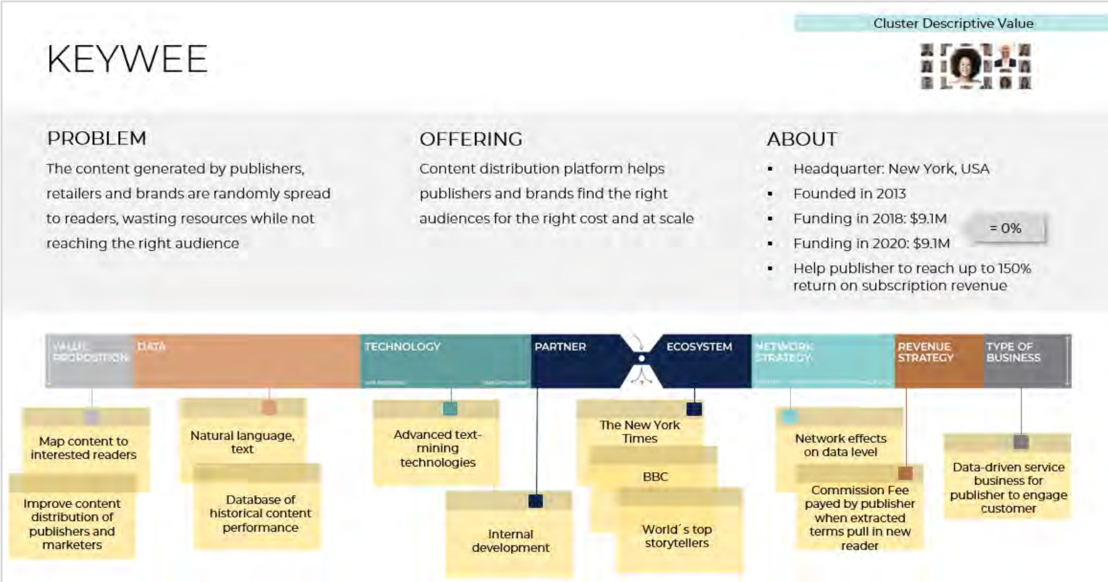


Figure 59 DDI Summary for Keywee

Problem	*The content generated by publishers, retailers and brands are randomly spread to readers, wasting resources while not reaching the right audience*
Where	New York, USA
Founded	2013
Funding 2018	$9.1M
Funding 2020	$9.1M
Funding Δ	= 0%
Acquired?	No
Outstanding info	Help publisher to reach up to 150% return on subscription revenue[140]
Cluster Rational	*Cluster Descriptive Value* *Keywee provides a match-making services based on natural language processing.* *It relies on semi-structured, media and time series data. The usage of industrial data is not observed.*
Value Proposition	Offering **publishers** and **marketers** a new approach to content distribution and performance measurement by scaling the content distribution campaigns quickly & efficiently. Also, it helps to connect with the right audience, and to discover what content resonates most with users.

[139] https://keywee.co/

[140] https://digiday.com/media/new-york-times-finds-new-subscribers-facebook/

Data	The company uses natural language processing to scan the content and understand the data. By employing its vast database of historical performance, the aim is then finding the right audience and distribute the content. While this process occurs, the company measures the performance and uses the data to enhance new content ideation and creation.
Technology	Advanced text-mining technology and a vast database of historical content performance to help publishers, retailers, and brands find the audiences most likely to meet specific business goals for their content: unlocking audience insights and infusing them into every step of the storytelling process
Partner Ecosystem	No information provided *Keywee* works with more than 400 of the world´s top storytellers, such as The New York Times, BBC and National Geographic
Network strategy	Networks effects on data level as the services become more precise /accurate the more data is available
Revenue Model	In 2018[141]: Subscription Selling of service In 2020[142]: Commission Fee: is payed whenever the match-making based on *Keywee* technology happens; i.e. a publisher sends a number of articles per month to *Keywee.* *Keywee* scans each article for relevant terms to pull in readers. Then the publisher buys traffic (for now, just on Facebook) against those terms. *Keywee* makes its money by taking a cut of the publisher's media spend.
Type of Business	Data-driven service business

[141] In 2018 no information related to revenue model could be found, in accordance to what they are doing it was very likely to rely on selling of services and/or subscription.

[142] https://digiday.com/media/new-york-times-finds-new-subscribers-facebook/

AGOLO[143]

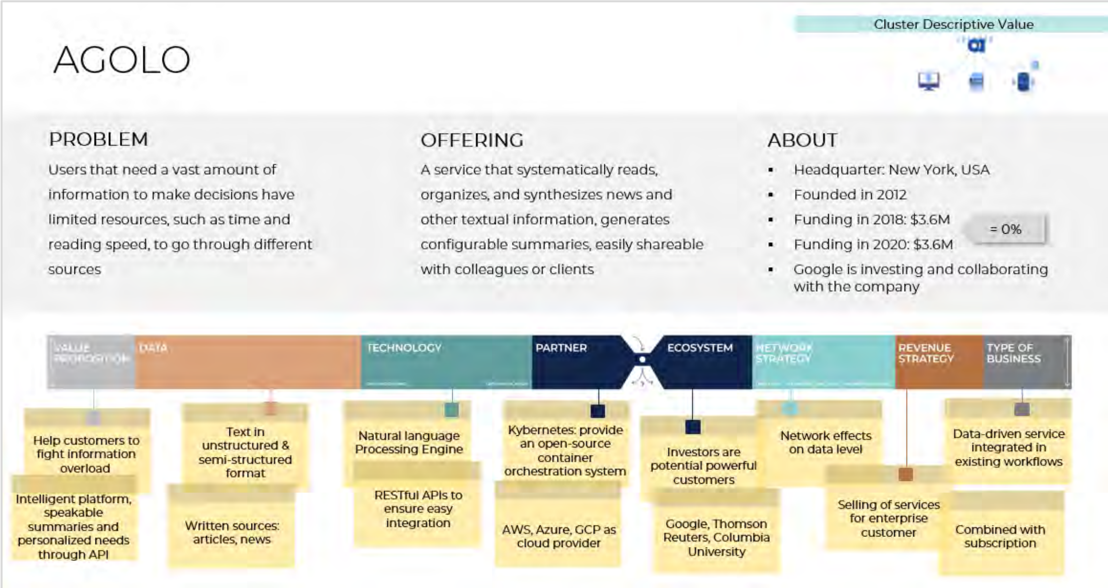


Figure 60 DDI Summary for Agolo

Agolo summarizes information faster and with broader coverage than any human. According to the company, they have the world's most advanced summarization software, helping its customers to fight information overload with summarization. They believe that the most interesting work is done when a variety of viewpoints come together.

As an example, they analyzed one of the Tesla´s earnings report. They scanned 58 different documents, analysts' reports, news articles and so forth. *Agolo* organized it into 7 different buckets, Tesla Model 3, Tesla Model X and so forth. Then the software wrote a summary around each bucket and finally a master summary saying, "These are the most important takeaways from the Tesla earnings report."

Problem	*Users that need a vast amount of information to make decisions have limited resources, such as time and reading speed, to go through different sources*
Where	New York, USA
Founded	2012
Funding 2018	$3.6M
Funding 2020	$3.6M
Funding Δ	= 0%
Acquired?	No
Outstanding info	Google is investing and collaborating with the company[144]
Cluster Rational	*Cluster Descriptive Value*

[143] https://www.agolo.com/

[144] https://www.techcty.com/agolo-attracts-microsoft-and-google-funding-with-ai-powered-summarization-tools/

	Agolo provides efficient means to summarize, synthesize and visualize news and information.
Value Proposition	Help customers to fight information overload with summarization through an intelligent platform, speakable summaries and personalized needs through API
Data	Text is the main source of data extracted from articles, news and different written sources in unstructured and semi-structured format
Technology	Natural language Processing Engine that builds upon a public or private cloud based on AWS, Azure or GCP and that is built with RESTful APIs enables seamless integration into current workflows.
Partner	*Kybernetes* providing an open-source container orchestration system for automating application deployment, scaling and management. AWS, Azure or GCP as cloud provider
Ecosystem	Most of *Agolo*'s investors[145], e.g. Google, Franklin Templeton Investments, Point72 Ventures, Thomson Reuters, M12, CRV and Columbia University are likely candidates for power customers with dedicated service agreements or co-developments
Network Strategy	Network effects on data level, as the algorithm can be improved with more data available
Revenue Strategy	Subscription in combination with Selling of services for enterprise customer
Type-of-Business	Data-driven Service Business

[145] https://www.agolo.com/company

6.5 Cluster E: Predictive Value

DESKTOP GENETICS [146]

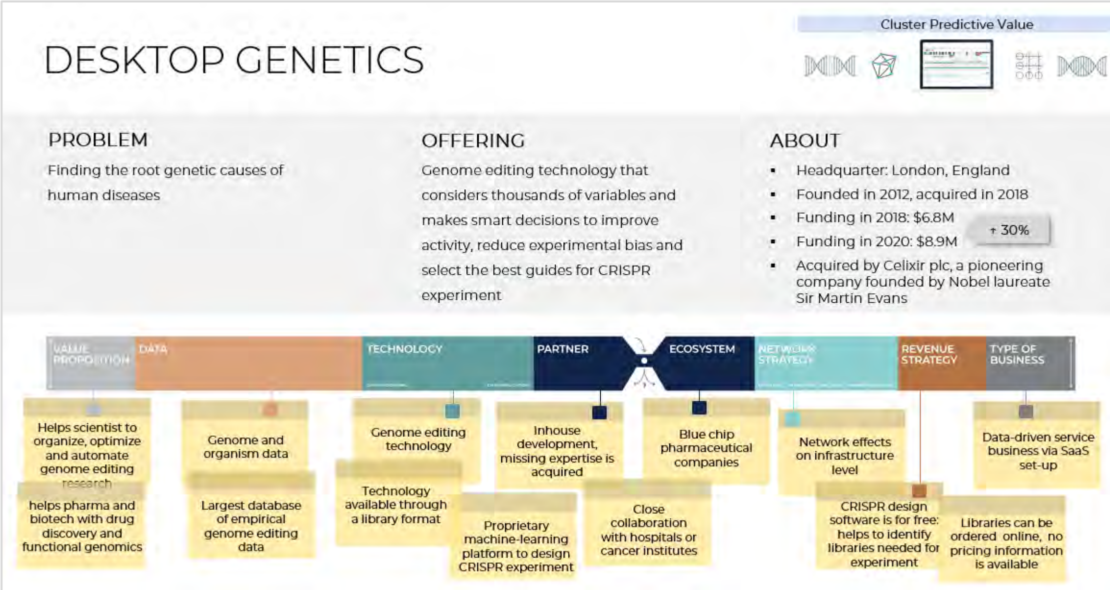


Figure 61 DDI Summary for Desktop Genetics

Desktop Genetics was established in 2012 and is a genome editing technology that considers thousands of variables and makes smart decisions to improve activity, reduce experimental bias and select the best guides for CRISPR experiment. Their aim is to help researchers discover and treat the root genetic causes of human disease.

Today, *Desktop Genetics* is a recognized leader in genome editing technology, staffed by dedicated team of genome editing experts, bioinformaticians and data scientists, driven by the real-world impact of CRISPR technology. *Desktop Genetics*'s tools and technologies are used by over 1800 organizations all over the world, and its projects contribute directly to several key partnerships to bring CRISPR into the clinic.

As a representative of the predictive cluster, the company relies on predictive analytics and sometimes complemented with means for process automation and other analytical values. Following the pattern of the cluster, it has a high focus on personal data and a higher usage of unstructured data. Also, no IoT and industrial data is observed.

Problem	*Finding the root genetic causes of human diseases*
Where	London, UK
Founded	2012
Funding 2018	$6.8M
Funding 2020	$8.9M
Funding Δ	↑ 30%
Acquired?	Yes, by Celixir

[146] https://www.deskgen.com/

Outstanding info	Acquired by *Celixir* plc, a pioneering company founded by Nobel laureate Sir Martin Evans[147]
Cluster Rational	*Cluster Predictive Value* *The product of Desktop Genetics - DESKGEN AI - is focusing on predictive analytics, in combination with descriptive analytics on genomics data (i.e. unstructured and personal data). The usage of IoT and industrial data is not observed.*
Value Proposition	Provide **genome editing experts, bioinformaticians and data scientists** with high quality, ready-to-use libraries that are AI-designed and tailored to their unique experiment.[148]
Data	Genome and organisms (Homo Sapiens, Mus musculus, Drosophila melanogaster, Escherichia coli, Arabidopsis thaliana, custom genome)[149] They claim is to be have the largest databases of empirical genome editing data of the world.
Technology	Their technology DESKGEN AI is a Genome editing technology that considers thousands of variables and makes smart decisions to improve activity, reduce experimental bias and select the best guides for CRISPR experiment. The technology is available through a library format, through genotyping analysis and mainly through their proprietary machine-learning platform[150]
Partner Ecosystem	Inhouse development *Desktop Genetics* partners include 4 of the top 5 blue chip pharmaceutical companies and world-renowned institutions such as the University College London Great Ormond Street Institute of Child Health, the Netherlands Cancer Institute, and the Copenhagen Center for Glycomics to support them in the development of advanced genome-editing applications
Network Strategy	On infrastructure level as the provide means for others to do the experiments. No network effect on data level as they only have very limited data with high pre-processing times
Revenue Strategy	In 2018: Subscription based on Software as a Service (SaaS) model In 2020: CRISPR design software is for free as this helps to identify all libraries needed for envisioned experiments

[147] https://www.deskgen.com/landing/#/company

[148] https://prismic-io.s3.amazonaws.com/desktopgenetics%2F324e0f9d-4eba-4f9b-8403-36005d070610_deskgen+library+catalog+v4.1.pdf

[149] See “order” bottom of their web-site

[150] https://www.deskgen.com/landing/#/company

	It is possible to order libraries online, no pricing information is giving, subscription model is likely as they mention that they continuously improve the library[151]
Type of Business	Data-driven service business

NEXOSIS / DATAROBOT

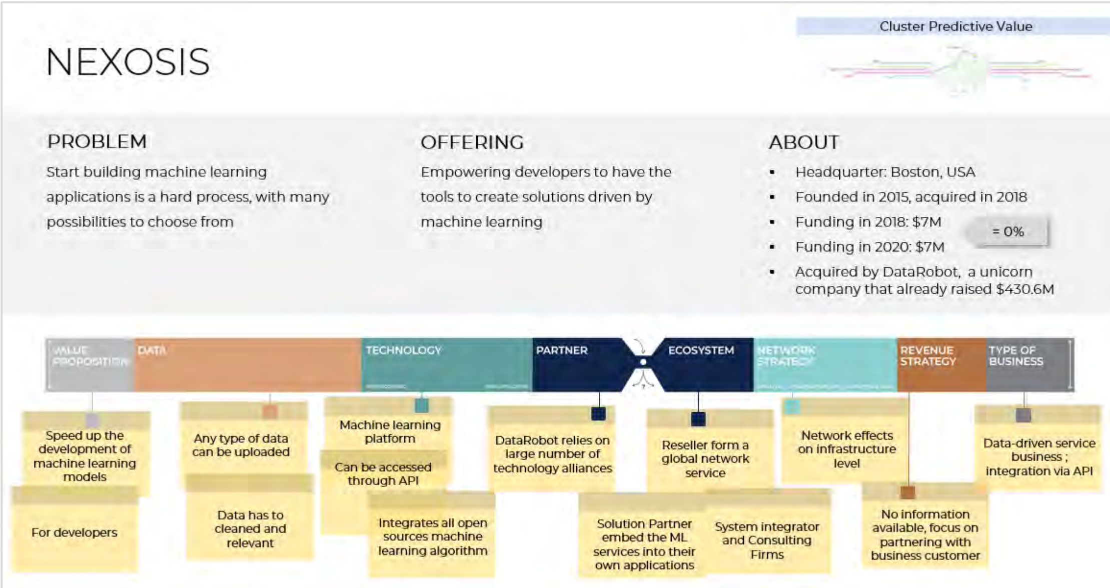


Figure 62 DDI Summary for Nexosis

Nexosis/DataRobot provides a machine learning API for developers. It is a developer platform for machine learning and their API allows for developers to quickly and easily build applications that leverage machine learning. They have automated key parts of machine learning so that building models is scalable and simplified.

The solution works on the following manner: first the user needs to upload the data, even if it comes from different sources but it has to be tidy and useful. Next, the user must choose the machine learning type is needed to solve the problem, such as time series. From this point on, different tasks can be kicked-off, including data categorization, aggregation and imputation. Finally, the platform packages the model up and deploys it to an API endpoint just for the customer, including all the previous steps so that when the model is called it quickly generates results tuned from and for the dataset.

[151] See how one example of their libraries is presented in https://www.deskgen.com/landing/#/resources/custom-designed-crispr-libraries

Problem	*Start building machine learning applications is a hard process, with many possibilities to choose from*
Where	Boston, USA
Founded	2015
Funding 2018	$7M
Funding 2020	$7M
Funding Δ	= 0%
Acquired?	Yes, by *DataRobot*
Outstanding info	*Nexosis* was acquired by *DataRobot*, the company founded in 2012 , already raised $430.6M and is being classified as unicorn
Cluster Rational	*Cluster Predictive Value* *Nexosis/DataRobot is offering a machine learning platform offering wide capabilities ranging from descriptive to predictive analytics. Any type of data being uploaded by the user upload can be processed*
Value Proposition	Provide a machine learning API for developers, making it easy to start building machine learning applications
Data	Any type of user data (even from different data sources) can be uploaded, however data has to cleaned and of relevance for the application / learning task in mind.
Technology	Machine learning platform including wide range of open sources machine learning algorithm and automated means for AI model generation and benchmarking The Machine learning platform can be accessed through an API or deployed on public or private cloud
Partner	DataRobot relies on large number of technology alliances providing platforms and applications integrating their product (ranging from AWS, Cloudera, tableau, etc.)[152]
Ecosystem	*DataRobot* has a growing partnership ecosystem on the demand side, e.g. - Value added reseller form a global network service provider that resell the ML platform - Solution Partner embed the ML platform into their own applications - System integrator and Consulting Firms use *DataRobot* to build horizontal ML solutions.
Network strategy	2018: No network effects 2020: Network effects on infrastructure level aiming to connect ML-related technology provider with AI solution developer
Revenue strategy	In 2018 the revenue model was a freemium with the limit of 10.000 predictions per month, and a customized offered if an enterprise needed dedicated support.

[152] https://www.datarobot.com/partners/

	In 2020 they put a strong focus on partnerships, therefore no information related to the revenue strategy was publicly made available on the website
Type-of-Business	Data-driven service business

6.6 Cluster F: Connecting Peers

INSURIFY

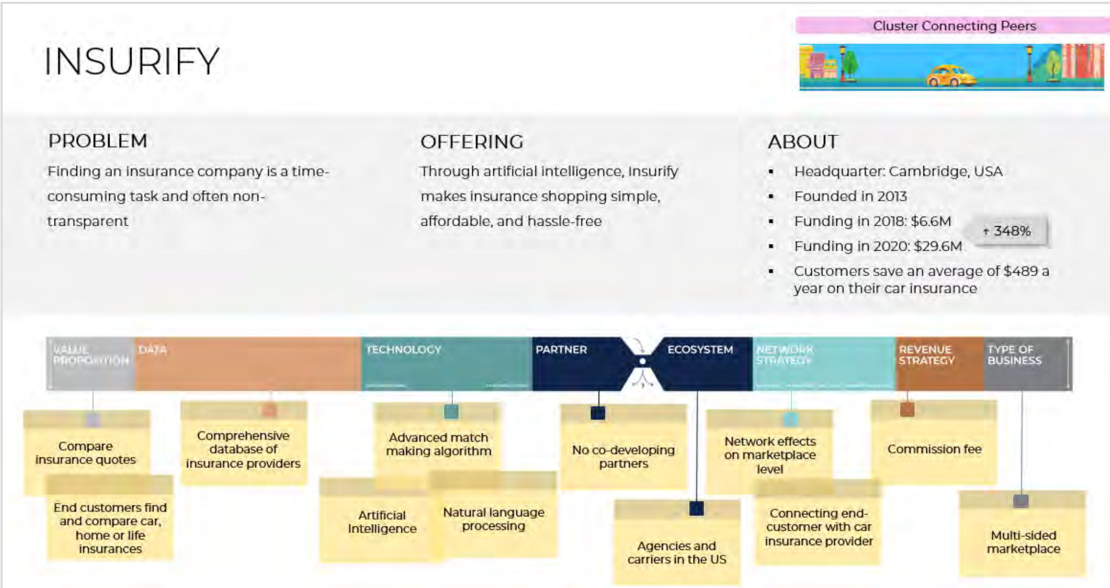


Figure 63 DDI Summary for Insurify

Problem	*Finding an insurance company is a time-consuming task and often non-transparent*
Where	Cambridge, USA
Founded	2013
Funding 2018	$6.6M
Funding 2020	$29.6M
Funding Δ	↑ 348%
Acquired?	No
Outstanding info	*Insurify* customers save an average of $489 a year on their car insurance when using the platform[153]
Cluster Rational	*Cluster Connecting Peers* *Insurify main value proposition is the match-making functionality between insurers (supply side) and owners (demand side). They rely on a commission fee revenue model, harness network effects on*

[153] https://insurify.com/company/about/

	marketplace level and establish multi-sided markets / data-driven marketplaces. In addition, the use personal data for implementing match-making algorithm
Value Proposition	Compare insurance quotes all in one place, matching the customer´s profile. In 2018 it was offered for comparing car quotes, as of 2020 life insurance and home insurance are also offered in their platform. Based on the customer profile, a personalized recommendation helping customer in find the appropriate insurance is provided.
Data	The company analyze millions of records, identify patterns and build models to match the customer´s profile. Comprehensive database of insurance providers including their offerings
Technology	Advanced match making algorithm based on Artificial Intelligence a Natural language processing
Partner Ecosystem	No co-developing partners are mentioned *Insurify* has partnered with the largest agencies and carriers in the US to quote 102 carriers in real-time, more than any other online platform for car insurance shopping. The founder of the company spent three years to build needed relationships within the industry.
Network Strategy	On marketplace level connecting end-customer with car insurance provider
Revenue Strategy	Commission fee as every time a user acquires an insurance through their website, a percentage of the deal is kept with Insurify[154]
Type-of-Business	Data-driven marketplace

[154] http://www.betaboston.com/news/2016/01/28/insurify-is-a-ai-powered-shopping-assistant-for-car-insurance

SELECTIONNIST[155]

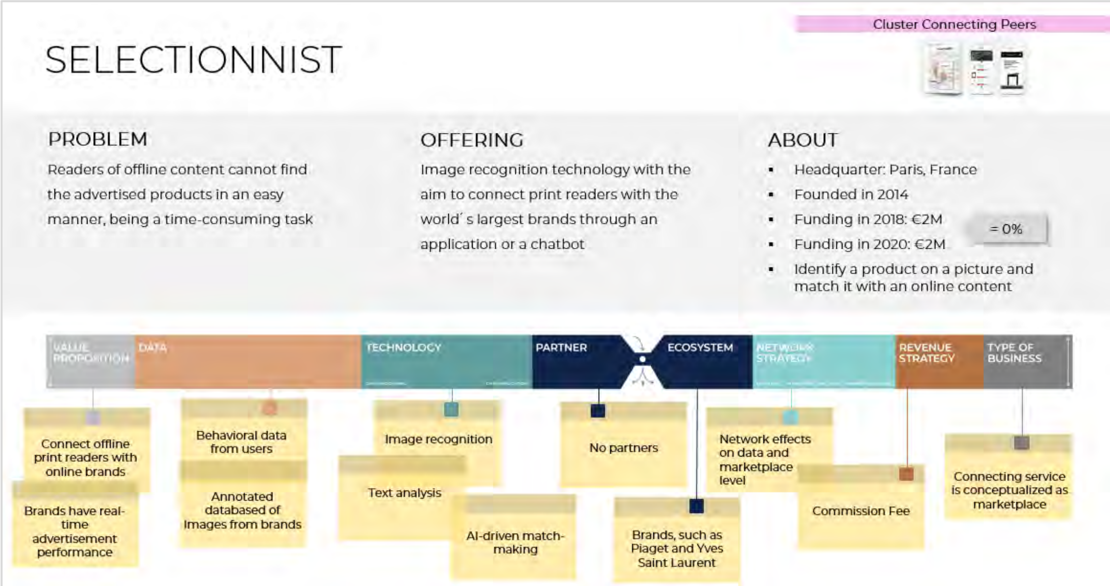


Figure 64 DDI Summary for Selectionnist

Selectionist is a France-based company found in 2014 offering image recognition technology with the goal to connect readers of print journals with the world's largest brands through an application or a chatbot. They aim to bridge the gap between offline content and online experience by offering an advanced match-making service to connect consumer and brands.

Problem	*Readers of offline content cannot find the advertised products in an easy manner, being a time-consuming task*
Where	Paris, France
Founded	2014
Funding 2018	€2M
Funding 2020	€2M
Funding Δ	= 0%
Acquired?	No
Outstanding info	Unique image recognition technology enables consumers to identify a product on a picture and match it with an online content
Cluster Rational	*Cluster Connecting Peers* *Selectionnist allows to connect brands (supply side) and consumers (demand side). Selectionnist's revenue model relies on commission fee, they harness network effects on marketplace level and establish a multi-sided market / data-driven marketplace*
Value Proposition	Match-making service to connect **offline print readers** with online brands.

[155] http://business.selectionnist.com/

	They address two different customer groups with different value proposition: • Value proposition for consumer: they locate and potentially purchase a product they spot in a magazine just by snapping a picture of it. • Value proposition for brands: brands can see in real-time how readers interact with their editorial & advertising in print magazines.
Data	Annotated databased of Image covering the product portfolio of brands being regularly updated by the brands, behavioral data from users
Technology	Image recognition technology, text analysis, AI-driven match-making
Partner Ecosystem	No partners are mentioned
	They managed to establish good partnerships with the well-known brands, such as Piaget, Yves Saint Laurent as well as top 30 French fashion magazines
Network strategy	Network effects on data level: the match-making algorithm is based on image recognition technology that continuously improves the more images from brands' product are in their data bases (more brands) as well as the more user request they receive. Network effects on marketplace level: The service is conceptualized as marketplace based on commission fee and requires to balanced growth of supply and demand
Revenue Model	The service is conceptualized as marketplace based on commission fee
Type-of-Business	Data-driven Marketplace

7 How our sample set performed over time

Two years passed since we first collected and coded a large set of successful data-driven start-ups. More precisely, the dates are from the first quarter of 2018 to the first quarter of 2020. Since the beginning, the objective of our research study was to systematically analyse and compare successfully implemented data-driven business opportunities. For instance, we have established our own criteria to measure success and they were:

- Funding between US$2M and US$10M
- Technology focus on Data Analytics and Artificial Intelligence

For more details regarding the why's and how's we performed our data collection and coding, please refer to Deliverable 2.6.

When consolidating all the content in this deliverable, we observed that the public presentation of many start-ups in our sample did no longer fit to our initial coding data. Many of them have evolved and adapted to new opportunities, some of them have been bought by other companies. Unfortunately, we do not have the resources to accomplish a detailed analysis of how data-driven start-ups adapt to the demand side over time. To answer such questions, a further research study would need to be scoped. Some companies from our sample set have considerably changed their value proposition since we first started our analysis. This shows the dynamic aspect of a data-driven business and how fast they can completely change their value proposition using their existing expertise and network. Founder seem to continuously investigate how to find more promising market or business opportunity by continuous opportunistic adaptations.

> For instance, HOOCH[156] is a great example coming from the USA. The first value proposition was a subscription-only cocktail app and discovery platform, where customers are offered a free drink each day and in contrast, bars and venues use it as a marketing tool while generating interesting consumer behavior. The platform united end-users to venue owners and consequently to the beverage industry. In the last two years their goal has switched from a cocktail-app only to a build community platform around shared experiences, all the while getting rewarded through a rewards program. Now customers can obtain rewards when they buy from partners of HOOCH such as for traveling, shopping or dining. They already have over 250.000 partners around the globe.

In order to get some indication of how (in terms of quantitative numbers) the sample set performed over time, we decided to do a comprehensive analysis of the change in funding received. This analysis will be covered in the following subsections.

In order to be able to interpret the numbers related to the funding amounts in the right manner, one need to keep in mind that the funding data provided is often not complete. This is due to several reasons, for instance not all funding data is public information. And although a vast majority of start-ups publicly expose how much they have raised in a funding round as this is seen as important success factor, there are

[156] https://www.crunchbase.com/organization/hooch

cases where the money is raised privately or the companies' revenues is already sufficient to reinvest in their own business.

When analyzing funding data to generate valuable insights and secure reliable analysis, we always have to keep the above mentioned aspects in mind. Simply due to lack of access to private data, we will rely in the following our analysis on two simplified assumptions:

- We assume that the funding amount published on Crunchbase is the most accurate one
- We assume that there is no other type of investment other than published on Crunchbase

In concrete, the section to follow will analyse four different aspects a) the founding year b) companies that outperformed in terms of funding and c) the funding variation between 2018 and 2020 and d) provide a qualitative analysis of the acquired companies to better understand their success leverage.

7.1 Foundation year

The foundation date is an insightful data item that can show us if data-driven businesses follow a pattern in accordance their age as well as the maturity of technology. When selecting the success criteria to our study, the foundation date was not of relevance as we did lack understanding how this could affect our sample set. The initial expectation was that most companies would be very young due to the data-driven aspect we were looking for.

The assumption that data-driven companies are rather young can be confirmed when looking at the table below. There is a similar behavior across all clusters and that is most of the companies in our data set were founded between 2012 and 2015. If we rank them, 2014 is the year where most firms were founded, followed equally by 2013 and 2015. Surprisingly, there are some businesses founded between 2004 and 2009, while from 2010 on at least one company was founded per cluster with exception of cluster D and E. To which extent the significant decline in founding rates was a consequence of the financial crisis in 2008 is subject of speculation. In addition, some general insights are presented below:

- 73% of the start-ups were founded between 2012 and 2015. We assume that with the emergence of the Big Data and later AI hype more funding did become available.
- 27% of the start-ups were founded in 2014
- Cluster A: 75% of the start-ups were founded between 2012-2015. 5% were founded in 2004, being the cluster with the oldest firm which is TRX Systems.
- Cluster B: foundation is spread across time, between 2008 and 2016. 57% were founded between 2012-2015.
- Cluster C: 42% of the start-ups were founded in 2014, the highest amount for this year among all clusters. Also, 17% were founded in 2017, which shows that very young company, for instance Plutoshift and Emagin, could validated their business model in a very fast pace. Also, 58% of the companies were founded between 2012 and 2015

- Cluster D: 75% of the start-ups were founded between 2012 and 2015. 31% of them were founded in 2012
- Cluster E: 85% of the companies were founded between 2012 and 2015. 38% of them were founded in 2015
- Cluster F: 87% of the companies were founded between 2012 and 2015. It is the cluster with the youngest companies in general

Founded Year	Avg.	Per Cluster					
		A	B	C	D	E	F
2004	1%	5%	0%	0%	0%	0%	0%
2007	1%	0%	0%	8%	0%	0%	0%
2008	1%	0%	7%	0%	0%	0%	0%
2009	2%	0%	7%	0%	6%	0%	0%
2010	4%	5%	7%	8%	0%	0%	7%
2011	9%	5%	7%	8%	19%	15%	0%
2012	13%	5%	21%	8%	31%	8%	7%
2013	17%	25%	7%	0%	19%	15%	27%
2014	27%	30%	21%	42%	19%	23%	27%
2015	17%	15%	7%	8%	6%	38%	27%
2016	6%	10%	14%	0%	0%	0%	7%
2017	2%	0%	0%	17%	0%	0%	0%

Table 9 Foundation year of start-ups from our sample set

7.2 Companies raised more than initial upper boundary over time

An interesting fact to look at is how many firms in the first quarter of 2020 have raised more than US$10M, as this was our upper boundary and success criteria when selecting the firms that would be part of our sample set. From the table below and considering that the firms that were acquired are included, we can conclude that:

- 37% of the companies continued raising money and are in 2020 above US$10M. Among different possibilities, this could mean:
  - The start-ups are growing their businesses and need more money to invest on human capital, regional or global expansion, R&D, sales force, etc.
  - The business needs a large investment before it can make considerable revenues. This is the case, for example, of companies that have their revenue model based on commission fee.
- Cluster A: if raising more than US$10M is the criteria for success, cluster A is the most successful one among all clusters with 45% of the companies raising more than US$10M
- Cluster B: 43% of the companies raised more than US$10M
- Cluster C: 42% of the companies raised more than US$10M
- Cluster D: 44% of the companies raised more than US$10M

- Cluster E: it is the cluster that received the least additional funding, with 15% of the companies raising more than US$10M
- Cluster F: 27% of the companies raised more than US$10M

Funding range 2020	Avg.	Per Cluster					
		A	B	C	D	E	F
2M-10M	63%	55%	57%	58%	56%	85%	73%
10M-20M	20%	25%	29%	33%	19%	8%	7%
20M-30M	11%	5%	14%	0%	25%	8%	13%
30M-40M	2%	5%	0%	8%	0%	0%	0%
40M-50M	0%	0%	0%	0%	0%	0%	0%
50M-60M	2%	10%	0%	0%	0%	0%	0%
60M-70M	0%	0%	0%	0%	0%	0%	0%
70M-80M	1%	0%	0%	0%	0%	0%	7%

Table 10 Funding amount that start-ups from our sample set have received until mid-March 2020

7.3 Funding variation between 2018 and 2020, incl. acquired companies

Since the beginning of our study we assume that successfully implemented data-driven business opportunities had received funding between US$2M and US$10M. Based on this assumption, it is of interest to analyze the variation in the funding amount over time. This is shown in the table below and further explored in the next sub-sections.

The difference between the two tables is that the first one includes the companies that were acquired, while the second one does not include. We wanted to check if there would be any interference in our interpretation if the acquired start-ups would be left out. By comparing both tables we come to the conclusion that there is no significant change, and therefore we will keep our analysis considering all the initial 90 start-ups.

Including Acquired Companies							
Funding Δ 2018-2020	Avg.	Per Cluster					
		A	B	C	D	E	F
0%	33%	35%	36%	25%	25%	38%	40%
0%-50%	20%	20%	7%	17%	19%	31%	27%
51%-100%	8%	5%	0%	8%	19%	15%	0%
101%-200%	13%	10%	7%	25%	25%	8%	7%
201%-300%	11%	0%	29%	8%	13%	8%	13%
301%-400%	8%	15%	14%	8%	0%	0%	7%
401%-500%	1%	0%	0%	8%	0%	0%	0%
501%-1000%	3%	5%	7%	0%	0%	0%	7%
1000%-1500%	2%	10%	0%	0%	0%	0%	0%

Table 11 Funding variation of our sample set from 2018 to 2020 including the companies that were acquired

Excluding Acquired Companies							
Funding Δ 2018-2020	Avg.	Per Cluster					
		A	B	C	D	E	F
0%	30%	29%	42%	27%	20%	40%	25%
0%-50%	21%	24%	0%	18%	20%	30%	33%
51%-100%	9%	6%	0%	9%	20%	20%	0%
101%-200%	13%	12%	8%	18%	27%	0%	8%
201%-300%	12%	0%	25%	9%	13%	10%	17%
301%-400%	8%	12%	17%	9%	0%	0%	8%
401%-500%	1%	0%	0%	9%	0%	0%	0%
501%-1000%	4%	6%	8%	0%	0%	0%	8%
1000%-1500%	3%	12%	0%	0%	0%	0%	0%

Table 12 Funding variation of our sample set from 2018 to 2020 excluding the companies that were acquired

7.3.1 Companies with no increase in further investments since 2018

In general, one out of three firms did not receive any further investments from the first quarter of 2018 to the first quarter of 2020. Some of the main explanations might be:

- The start-up was **acquired by another firm**: In the analyzed period, 16% of the companies were acquired by other firms. Considering that among this 16% some received extra funding and some did not, we find out that 57% of them did not receive any further funding.
- **Private funding**: Companies that had no variation might have received private funding, for example of known partners, family or banks, and decided to keep this information private as it might be a strategic one.
- **Revenue growth**: Some types of businesses can generate revenues on early stages and therefore do not need further funding. They can reinvest a part of the revenues on the business without announcing it publicly, which also might be a strategic decision.
- **Out of the market**: The company could also be facing different types of problems that would run them out of the market, and it could take a while to notice it. These problems might surpass product validation, team building, technical qualifications, and economic & political aspects, among others.

7.3.2 Companies with extraordinary further investments

For this study, we are considering an extraordinary performance if the start-up has received at least five times the initial amount of funding we collected in the first quarter of 2018.

7.3.2.1 Between 501% and 1000%

The companies that have received an increase higher than 500% are:

- **Tractable (Cluster A):** Tractable is applying artificial intelligence to speed up accident and disaster recovery by using computer vision to perform visual damage appraisal instead of getting humans to do the job. The latest capital contribution of US$25M at the end of February 2020 will go on expanding the market footprint, which is already present in nine markets and will also use for continued product development by enhancing its AI, according to TechCrunch[157].

- **Locus.sh (Cluster B):** According to TechCrunch[158], Locus, an Indian startup that uses AI to help businesses map out their logistics, has raised $22 million in Series B funding in May 2019 to expand its operations in international markets
- **CleverTap (Cluster F):** According to TechCrunch[159], CleverTap, an India-based startup that lets companies track and improve engagement with users across the web, has pulled in $26 million in new funding thanks to a round led by Sequoia India.

7.3.2.2 Between 1001% and 1500%

- **CheckRecipient (Cluster A):** According to TechEU[160], the US-based VC firm Sequoia Capital has led a $42 million funding round in Tessian, a cybersecurity startup headquartered in London. This happened in February 2019 and the company plans to use the fresh capital to grow its team and expand its product line later in 2019.
- **Lynq (Cluster A):** According to Yahoo Finance[161], Lynq Technologies, Inc., the deep-tech startup that connects people and devices across miles without infrastructure, has announced a $6 million seed round in January 2020, amounting a total of US$34.8M in funding since its beginning. According to Dave Shor, the founder and CEO of Lynq, the plan ahead is increasing the size of the team and further developing scalable opportunities around the core technology.

7.4 Acquired Companies

7.4.1 Who are they?

Between 2018 and 2020, 16% of start-ups were acquired by other companies. Some interesting facts about these companies:

- Cluster A: 21% of acquired companies belong to Cluster A, they are:
  - SchoolMint acquired by Hero K12
  - Artomatix acquired by Unity Technologies
  - NVMdurance acquired by RPX Corporation

[157] https://techcrunch.com/2020/02/27/tractable-claims-25m-to-sell-damage-assessing-ais-to-more-insurance-giants/

[158] https://techcrunch.com/2019/05/13/indias-locus-raises-22-million-to-expand-its-logistics-management-business/

[159] https://techcrunch.com/2019/04/10/clevertap-raises-26-million/

[160] https://tech.eu/brief/british-email-security-startup-tessian-raises-42-million-from-sequoia/

[161] https://finance.yahoo.com/news/lynq-technologies-inc-closes-6-130000753.html

- Cluster B: 14% of acquired companies belong to cluster B, they are:
  - Placemeter (acquired by Netgear) acquired by Netgear
  - Mnubo acquired by Aspen Technologies

- Cluster C: 7% of acquired companies belong to cluster C, they are:
  - Uplevel Security acquired by McAfee

- Cluster D: 7% of acquired companies belong to cluster D, they are:
  - Brand Embassy acquired by NICE Systems

- Cluster E: 29% of acquired companies belong to cluster E, they are:
  - Cloudcherry acquired by Cisco
  - Nexosis acquired by DataRobot
  - AutoRobotics acquired by Ridecell
  - Desktop Genetics acquired by Celixir plc

- Cluster F: 21% of acquired companies belong to cluster F, they are:
  - Mila acquired by Swisscom
  - Influenster acquired by Bazaarvoice
  - Paysa acquired by Ceridian

7.4.2 Correlation between Type of Business and Acquisition Rate

While 64% of the acquired companies were a data-driven service, none of them were classified as emerging technologies.

Type of Business	%
Data-driven Service Business	64%
Data-driven Marketplace	29%
Emerging Technologies	0%

Table 13 Correlation between Type-of-Business and Acquisition Rate

7.4.3 Further findings

- 57% of the start-ups did not receive any further funding before the acquisition since we captured the data in the first quarter of 2018.
- The companies who were acquired without further funding had received US5.9M on average before the acquisition
- 38% received further funding before being acquired:
  - Artomatix from Cluster A was the company with the highest value of additional funding before the acquisition, quadrupling its initial amount
  - Mnubo from Cluster B comes next, more than tripling the initial funding amount

- Cloudcherry from Cluster E is the third one, more than duplicating the initial funding amount
- Uplevel Security from Cluster C is the fourth one, also more than duplicating the initial funding amount
- Desktop Genetics from Cluster E is the fifth one, with an increase of 30% in the funding amount before being acquired

- 21% of the companies had more than US$10M of funding before the acquisition

Startup Name / Cluster	Founded Date	Δ Funding Amount	Funding Amount US$M		Acquired by	Description
			2018	2020		
SchoolMint (A)	2013	0%	7.7	7.7	Hero K12	SchoolMint helps schools and families manage the admissions process, simply and cost effectively.
Artomatix (A)	2014	313%	2.6	10.8	Unity Technologies	Artificial Imagination: Artificial Intelligence applied to art creation
NVMdurance (A)	2013	0%	3.6	3.6	RPX Corporation	NVMdurance is accelerating the adoption of 3D NAND.
Placemeter (B)	2012	0%	9.4	9.4	Netgear	Placemeter is an urban intelligence platform. We quantify modern cities worldwide. The answers you need are all around you.
Mnubo (B)	2012	221%	5.0	16.1	Aspen Technologies	Big Data Analytics for The Internet of Things
Uplevel Security (C)	2014	112%	2.5	5.3	McAfee	Uplevel Security is the first adaptive system of intelligence that uses graph theory and machine learning to modernize security operations.
Brand Embassy (D)	2011	0%	4.0	4.0	NICE Systems	Brand Embassy helps companies to Acquire, Retain and Grow customers through digital customer service, at scale.
Desktop Genetics (E)	2012	30%	6.8	8.9	Celixir	Building an AI to re-engineer the human genome
Cloudcherry (E)	2014	129%	7.0	16.0	Cisco	Real-time Customer Sentiment Mapping and Experience Analytics
Nexosis (E)	2015	0%	7.0	7.0	DataRobot	Nexosis provides a machine learning API for developers

Auro Robotics €	2015	5%	2.1	2.2	Ridecell	Auro Robotics is the Self-Driving Shuttle for travel within university campuses, corporate parks, and residential communities.
Mila (F)	2013	0%	3.2	3.2	Swisscom	Mila is a crowd services platform. On Mila.com, users can find, book, and rate tech-savvy people or offer their technical know-how.
Influenster (F)	2010	0%	8.0	8.0	Bazaarvoice	A product discovery and reviews platform that enables consumers to find new products and get advice to make informed purchases.
Paysa (F)	2015	0%	4.0	4.0	Ceridian	Paysa is the world's first platform to empower individuals to maximize their salary across the span of their career.

Table 14 Overview of start-ups of our sample that have been acquired

8 Evaluation of DDI Framework in Practice

The usage of the DDI Framework has been introduced in different spheres during the last years. The goal was always to disseminate the available knowledge from our research, all the while a constant evaluation of the methodology was performed. Throughout different formats, such as webinars, workshops and an university lecture, we had the opportunity to evaluate the eight dimensions of the DDI framework were of relevance and the ones that would need some type of review.

As already described in deliverable D2.6 we have designed and conducted data-driven innovation workshops in business settings as one-day trainings for professionals. In addition to the one-day training we also conceptualized an half-day training as well as an webinar. In sum, the following workshops and webinars have been conducted:

- 5-day DDI Lecture at technical university in Munich in April 2019 with regular meetings throughout the summer semester
- DDI Presentation at Data-driven Business Model Workshop at the BDV Meetup in Riga, 26.6.2019
- Half-day DDI Workshop with PPP projects at the BDV Meetup in Riga, 27.6.2019
- 1-day DDI Workshop with Corporates (Nürnberg 11.7.2018 / Erlangen 18.2018,
- DDI Webinar with Data Pitch Start-ups (online 19.9.2019)
- DDI Workshop with SMEs (Murcia, 1.10.2019)
- DDI Panel at South Summit (Madrid, 8.10.2019)
- Half-day DDI Workshop with EDI-Start-Ups (Bilbao, 6.11.2019)

In D2.6 we have already evaluated the DDI Lecture with students as well as the DDI workshop in enterprise settings. To complement those evaluation we are focusing in the third deliverable series on the one-day workshop we conducted with SMEs. For us a new target group. Below we describe the evaluation of this events.

The workshop "*Success Patterns of Data-driven Business Opportunities*" was attended by 11 participants in October 2019. The workshop happened in Murcia, Spain, at the Ceeim - Centro Europeo de Empresas e Innovación de Murcia, a space dedicated to innovation in the region.

The region of Murcia is known for its export and entrepreneurial vocation. The Region of Murcia has set important assets for growing that allows progress towards convergence alongside most prosperous and advanced regions of Spain and Europe. It has an important number of specialized and competitive international areas that are contributing to overall development, generating new employment and, which is more important, they are setting a precedent to inspire and drive other companies. (http://www.ris3mur.es/2018/08/23/murcia-y-su-potencial-innovador/?lang=en)

The participants expected to develop the following four aspects:
• Develop and implement data-driven innovations
• Investigate and discover user / customer needs
• Familiarize with data processing steps
• Scope data-driven business opportunities as well as to position data-driven offerings

Based on the results of the quantitative and representative study of data-driven business opportunities reported in the previous deliverables, the workshop introduced patterns and success criteria of successful data-driven start-ups. The workshop had the duration of one day and followed the structure of the Data-Driven Innovation Framework. It covered the value proposition, the data aspect, the technology aspects encompassing the whole data value chain, the partner and ecosystem as well as the value network strategy, revenue model and type of business. Each participant received a printed material with a full-guidance package: the printed version of the Canvas where they were able to stick to the wall and brainstorm the ideas. In addition, a booklet with visual orientation, guiding questions and a how to use section. The quality of the material and the content was of high relevance for delivering the workshop.

During the day, the participants actively participated of the workshop. Every part of the canvas was explained with theoretical and practical examples. This dynamic also happened with the participants, with a hands-on moment at each of the 8 dimensions of the canvas. When they finished the supply side, the participants pitched their ideas in one minute and received feedback from Prof. Dr. Sonja Zillner.

To evaluate the workshop, we accomplished a quantitative and qualitative analysis. It was based on a questionnaire the participants filled out at the end of the workshop. The questionnaire covered questions to rate the relevance of content, methods and set-up. The results are divided into five parts: relevance, most valuable aspect, should be kept, improve, and dimension analysis.

RELEVANCE: The relevant learnings for the participants were based in three main aspects:

1. The methodology
2. The relevance of a specific part of the canvas
3. The Data-Driven focus

Nearly half of the participants have commented that it is of great value to have a road to organize a data-driven idea or established a new or running project. In some cases, the companies were in the process of managing innovation and needed a methodology. Therefore, a method that has a complete view of all necessary steps was of great value. For 55% of the participants, a specific part of the canvas was more appealing, being the value proposition the most mentioned one.

MOST VALUABLE ASPECT: It was a consensus among the group that having a clear methodology was the most valuable aspect of the workshop. Having the chance to practice following a clear method was a highlight, interpreted as a training for future appliance in other projects. Another valuable aspect often mentioned are the examples. The workshop has many of them and the participants really enjoy it, being a positive aspect for 36% of the participants. They have also appreciated the data-driven focus.

SHOULD BE KEPT: When asked about which components should be kept, 64% of the participants mentioned going through the whole methodology. It seems to be of value to have a detailed explanation of each dimension, and that it was easy to follow, interactive and practical. Again, the examples were mentioned by 18% of the participants – they believe it should be kept.

IMPROVE: For 40% of the participants there was not enough time to process all the information. It was suggested a two days workshop, where they could go home and process the supply side. The next day would be focused on the demand side, and as they could think better about the supply side, maybe new ideas would come. The participants also think that a specific dimension should be the further exposed. Here we understand that every person has its own needs in terms of project or specific knowledge. What we extract is that if it is possible to assess the level of knowledge from the participants beforehand, it would be easier to bring a more focused workshop. As this is often hard, a more general approach keeps being the ideal one. Also, while the workshop is happening, Sonja Zillner feels what is the need and then spends more or less time in a specific topic. The following aspect to be enhanced is giving a better description of the workshop before it happens, also with further instructions for the homework (where they have to pick an idea). Another aspect is updating or adapting some examples, mentioned by 2 participants. One says in terms of bringing more recent examples, and the other one says bringing examples of SME´s.

DIMENSION ANALYSIS. The following table shows the dimension analysis, where five means the highest rate and 1 the lowest rate. The participants could mark all of them with the same value, or assign different ones. First, we analyzed the dimensions with rate five. In decreasing order we have value proposition as the most relevant dimension (64%), followed by the Data aspect and Partner and Ecosystem (55% each), Technology and Type of Business (36%), Revenue strategy (26%) and Network Strategy (9%).

It is worth to highlight that the positive results for the supply side might be biased by the fact that participants were more attentive in the morning. As we could notice from above, it was often mentioned that one-day workshop might be too short and therefore they had less energy to focus on the content given in the second part of the workshop.

The dimension with rate four is led by Revenue Strategy (45%), Network Strategy (36%) and Value Proposition, Data Aspect and Type of Business (27%). As rate 4 is still a high number in our grade scale, we believe that the revenue strategy and network strategy are also of high relevance and interest of the participants.

The analysis of the dimensions with grade three, two or one tells us that the participants agree that all dimensions contribute to the overall knowledge. Given that the percentages are very low for these rates, most people gave at least grade 3 to the dimensions and we are confident that we should not exclude any dimension.

Grade	Value Proposition	Data	Technology	Partner & Ecosystem	Network Strategy	Revenue Strategy	Type of Business
5	64%	55%	36%	55%	9%	27%	36%
4	27%	27%	18%	9%	36%	45%	27%
3	0%	0%	36%	18%	45%	18%	36%
2	0%	9%	0%	18%	0%	9%	0%
1	9%	9%	9%	0%	9%	0%	0%

Table 15 Dimension Analysis: which dimensions are of most relevance for the workshop participants

9 Conclusion and Next Steps

This report provides a third update of the BDVe task aimed at documenting and reviewing emerging business opportunities. We finalized a quantitative study to identify success approaches and patterns in data-driven business opportunities that will provide guidance for investment decisions in the future.

In this study we could identify six success patterns of data-driven innovation opportunities. In addition, we were able to consolidate a large number of data-driven usage scenarios. Both results are very valuable source of content for future engagement format as the “inspirations from the real world” are in general very important source of inspiration for entrepreneurs.

By relying on the Data-driven Innovation Framework, we have now a proven method in place that we can share with all members of the BDV ecosystem to provide guidance in exploring and scoping data-driven business opportunities. The comprehensive content can be used for industrial workshops and educational set-ups. In the future, we plan to engage with the stakeholders of the BDV ecosystem, with focus on SMEs and start-ups, to help them scope promising business opportunities. Due to the covid-19 crisis, it becomes clear that engagement formats also need to move online. With the DDI Framework, we decided to accomplish the first online lecture at Technical University in Munich already in April 2020. Based on those experiences, we plan to design further interaction and teachings with our community in the upcoming month.

In this way, we aim to foster industrial investments in data-driven innovation in the BDV PPP ecosystem.

10 References

Aitken, M. and Gauntlett, C. (2013) *Patient Apps for Improved Healthcare: From Novelty to Mainstream*, *IMS Institute for Healthcare Informatics*.

Attenberger, Y. (2016) *Business Models and Big Data*. Master Thesis at the Munich School of Management, LMU Munich.

Choudary, S. P. (2015) *Platform Scale - How a new breed of startups is building large empires with minimal investment*. Platform Thinking Labs Pte Ltx.

Davenport, T., Jeanne, H. and Robert, M. (2010) *Analytics at Work: Smarter Decisions, Better ResultsNo Title*. Harvard Business Press.

Frey, C. B. and Osborne, M. (2013) *The Future of Employment*, *The Future of Employment*.

Hagel, J. *et al.* (2015) 'Patterns of disruption - Anticipating disruptive strategies in a world of unicorns, black swans and exponentials', *Deloitte University Press*, p. 28`.

Hand, D., Mannila, H. and Smyth, P. (2001) *Principles of Data Mining Cambridge*, *MIT Press*. doi: 10.1007/978-1-4471-4884-5.

Hartmann, P. M. *et al.* (2014) 'Big Data for Big Business? A Taxonomy of Data-driven BusinessModels used by Start-up Firms', *Cambridge Service Alliance Whitepaper*, pp. 1–29. doi: 10.1016/j.im.2014.08.008.

KETCHEN Jr., D. J. and SHOOK, C. L. (1996) 'THE APPLICATION OF CLUSTER ANALYSIS IN STRATEGIC MANAGEMENT RESEARCH: AN ANALYSIS AND CRITIQUE', *Strategic Management Journal*. doi: 10.1002/(sici)1097-0266(199606)17:6<441::aid-smj819>3.0.co;2-g.

Metzger, A., Franke, J. and Jansen, T. (2020) 'Ensemble deep learning for proactive terminal process management at duisport', in Brockke, Mendling, and Rosemann (eds) *Business Process Management Cases – Volume 2*. Springer.

Mooi, E. and Sarstedt, M. (2011) *A Concise Guide to Market Research*, *A Concise Guide to Market Research*. doi: 10.1007/978-3-642-12541-6.

Osterwalder, A. and Pigneur, Y. (2010) *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*. Toronto: Self Published.

Shapiro, C. and Varian, H. R. (1999) *Information rules: A Stragegic Guide to The Network Economy*, *Harvard Business School Press*. doi: 10.1145/776985.776997.

Zillner, S. *et al.* (eds) (2017) *Big Data Value Strategic Research and Innovation Agenda*.

Zillner, S. (2019) *D2.6. Annual Report in Opportunities, Deliverable 2.6*.

Zillner, S., Timan, T. and Kotterink, B. (2018) *D2.5 Annual Report in Opportunities*.

Appendix 1: DDI Ontology

Generic Information				
At least three funding rounds				Funding round is a modality of raising money from investors that start-ups often use. It can be a measurement of success as the higher the number of funding rounds, the more likely the business model is validated. At least three funding rounds is an extra indicative of the business model validation.
Value Proposition				
Sector-agnostic				
Target Customer	B2B			In **Business to Business (B2B)** the addressed customer is a business customer. The Value has a stronger emphasis on financial values, such as reducing costs and/or growth. In case emotional drivers are of relevance, they follow economic value of offering.
Target Customer	B2C			In **Business to Customer (B2C)** the addressed customer is an individual or household. The VP is typically a combination of functional and emotional benefits.
Target Customer	B2B & B2C			In multi-sided market, B2C and B2B market logics are combined to design new market set-ups.
Offering	Data Value	Insight Generation	Analytics	Data is acquired, acted upon, and sold within the data market. Corporate entities trade data internally in their organizations and between organizations.(IDC & OpenEvidece 2017)
Offering	Data Value	Insight Generation	Descriptive Analytics	Descriptive Analytics is the most frequently used analytics. Its main objective is to explain what had happened in the past by providing the analyst, business person or expert a view of key metrics that measure the area of interest. The traditional business intelligence and data mining applications fall into this category. They provide a very important basis for developing a deeper understanding of the underlying data sources.
Offering	Data Value	Insight Generation	Diagnostic Analytics	Diagnostic Analytics aim to explain the root-case of a problem. Its main objective is to explain "why something happens". Those applications are often based on rule-based or semantic model capturing important background knowledge as well as flexible dashboards empowering the expert / user to explore or filter relevant features.
Offering	Data Value	Insight Generation	Predictive Analytics	Predictive Analytics is about forecasting. Its main objective is to predict what will happen in the future, for instance the estimated point in time of a machine outage or forecasting a quantifiable amount of customer, etc. A predictive model relies on a variety of variable data that have a relationship which the event the model aims to predict.

Offering	Data Value	Insight Generation	Prescriptive Analytics	Prescriptive Analytics is the basis for decision support and decision automation. Its main objective is to inform the machine or the user about the best course of action or strategy. Prescriptive analytics requires a deep understanding of the underlying engineering, business, mental or other processes in order to transfer analytics results into recommended actions.
		Orchestration	Matching	Matching enables the automated orchestration of value generation. Match-making algorithms are mapping the demand side requirements with the supply side resource capabilities. Typical examples that make use of matching algorithm are dating platforms.
		Automation	Process automation	Process automation includes all applications that help to replace manual tasks or activities by machines or algorithms. This can range from applications that automate one very particular human tasks in comprehensive manner, such as supporting lawyers finding relevant cases, to applications enabling the automated orchestration of existing processes and workflow, such as AI-based planning and scheduling algorithms that develop strategies or actions sequences for execution by intelligent agents, autonomous robots and unmanned vehicles.
	Hardware			Process automation includes all applications that help to replace manual tasks or activities by machines or algorithms. This can range from applications that automate one very particular human tasks in comprehensive manner, such as supporting lawyers finding relevant cases, to applications enabling the automated orchestration of existing processes and workflow, such as AI-based planning and scheduling algorithms that develop strategies or actions sequences for execution by intelligent agents, autonomous robots and unmanned vehicles.
Data				
Time and Space Dimension	Time-Series / Temporal			Time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average. Another prominent example of time series data is the behavioral data tracked with any user interaction on web or mobile devices.
	Geo-spatial data			Geospatial data is data that has a geographic component to it. This means that the records in a dataset have location-related information tied to them such as geographic data in the form of coordinates, address, city, or ZIP code. GIS data is a form of geospatial data.

Complex Data	IoT Data		The Internet of things (IoT) is the network of physical devices, vehicles, home appliances, and other items embedded with electronics, software, sensors, actuators, and network connectivity which enable these objects to connect and exchange data. IoT data refers to the exchanged data in IoT networks. When analyzing IoT data several challenges need to be tackled: a) data structures as most sensors send out data with time stamps. Here the challenge is significant signals might happen only once in a while but require then attention, e.g. through static alerts based on thresholds (Same challenges as time-series data) b) IoT data is often characterized by a strong correlation of time-series / sensor data and other sources of unstructured data, such as log files c) many machine learning algorithm require to aggregate sequential data into cross-sectional representation. This translation might cause loss of critical information.
	Media Data		Media data encompasses all data sources used for media communications. This includes any type of communication channels through which news, entertainment, education, data or promotional messages are disseminated. This includes broadcasting and narrowcasting mediums such as newspapers and magazines TV, radio, billboards, direct mail, telephone, fax and internet. In this way, media covers a plural of mediums and thus very heterogeneous data sources. In the context of media data analysis there is a strong focus on video (e.g. YouTube) and social media data (e.g. Twitter, Facebook), etc.
Data Type	Unstructured Data	At least one type of Unstructured Data*	Unstructured data is free from data that is much more challenging to process as dedicated algorithm for extracting intermediate results from raw data are required. Each type of unstructured data relies on different and specific pre-processing algorithms and mechanism to extract metadata describing the content /semantics of the data. Often semantic models (meta-data and graph data are used to formally represent the extracted information from unstructured data).
		Video and Image Data	Image and Video data (Image data includes all mean for storing and transmitting photo graphic images as a digital file, e.g. tiff, jpeg, etc.
		Text, Language and Audio Data	Text or language data and audio data refer to any type of unstructured text-based data, such as a radiology report, a case description or emails. Basically, the text files encompass a long list of strings. Natural language techniques (NLP) are used to extract semantics from text documents. In accordance to the underlying language different NLP techniques have to be used.

		Genomics Data	Genomics data refers to the genome and DNA data of an organism. They are used in bioinformatics for collecting, storing and processing the genomes of living things. Genomic data generally require a large amount of storage and purpose-built software to analyze.
Data Type	Unstructured Data	Laser Data and others	Other formats, such as laser data
	Semi-Structured Data		Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. A typical example of semi-structured data is Web data. Web data refers to the wealth of data stored in the www. Most prominent data source is Wikipedia. When this data gets extracted /crawled from the web, the data is in semi-structured format. Another typical example of semi-structured data is the wealth of social media data such Facebook or Twitter.
	Semantic Data		Semantic data makes the content and semantic of data explicit and machine-processable. This includes *Graph data refers to all formally represented knowledge. Most prominent examples are medical ontologies or the Google knowledge graph. *Meta-data includes any type of data used for labelling or classifying other data sets. Metadata is defined as the data providing information about one or more aspects of the data; it is used to summarize basic information about data which can make tracking and working with specific data easier.
Origin or Data	Personal Data		Personal data is information relating to you only, which makes you identifiable – your name, photo, phone number, birth date, e mail address, car number plate, etc.
	Industrial Data		Industrial data refers to any data assets produced and used in industrial settings of all areas. Often those are data produced from productions lines, energy systems, infrastructures, etc. In general, industry data is "closed data" meaning that it is "owned" (in terms of access rights) by the entity operating the product or machine or thing producing the data and that the data is likely to cover confidential information, e.g. a log file from a medical MR might contain valuable information indicating IPRs of the MR itself. This also includes Operational data in general produced in the context of operating any type of IT system
	Open Data		Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. The goals of the open data movement are similar to those of other "open" movements such as open source, open hardware, open content, open government and open access

	Research Data	Research data is defined as recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created
Technology		
BDV SRIA Technologies	Data Management	The technical priority Data Management is motivated by the fact that more and more data is becoming available. This data explosion, often called "data tsunami" or "data lake", is triggered by the increasing amount of sensor data and social data, born in Cyber Physical Systems (CPS) and Internet of Things (IoT) applications. For more details we refer to (Zillner et al. 2017).
	Data Processing Architectures	The technical priority Data Processing Architectures is motivated by fast development and adoption of Internet of Things (IoT) technologies which is one of the key drivers of the Big Data phenomenon. Initially this phenomenon started by applying the existing architectures and technologies of Big Data that we categorize as data-at-rest, which is data stored in persistent storage. In the meantime the need for processing immense amounts of sensor data streams has increased. For more details we refer to (Zillner et al. 2017).
	Data Analytics	The technical priority Data Analytics aims to progress data analytics technologies for Big Data in order to develop capabilities to turn Big Data into value, but also to make those approaches accessible to the wider public. Data analytics will have a positive influence on all parts of the data value-chain to increase business opportunities through business intelligence and analytics while bringing benefits to both society and citizen. For more details we refer to (Zillner et al. 2017).
	Data Protection	The technical priority Data Protection addresses the need for advanced data protection and anonymization technologies in the areas of Big Data and data analytics. With more than 90% of today's data being produced in the last two years, a huge amount of person-specific and sensitive information coming from disparate data sources such as social networking sites, mobile phone applications, electronic medical record systems, etc., is being increasingly collected. For more details we refer to (Zillner et al. 2017).

	Data Visualization and User Interaction	The technical priority Data Visualization and User Interaction is addressing the need for advanced means for visualization and user interaction capable to handle the continuously increasing complexity and size of data to support the user in exploring and understanding effectively Big Data. Visual analytics is the science of analytical reasoning assisted by interactive user interfaces. Data generated from data analytics processes need to be presented to end users via (traditional or innovative) multi-device reports and dashboards which contain varying forms of media for the end-user, ranging from text, charts, to dynamic, 3D, and possibly augmented reality visualizations. For more details we refer to (Zillner et al. 2017)
BDV SRIA Technologies	At least one SRIA Technologies*	
	More or equal to three SRIA Technologies*	
BDV Complementary Technologies	Blockchain	A blockchain is a growing list of records, called blocks, that are linked using cryptography. Each block contains a cryptographic hash of the previous block, a timestamp, and transaction data (generally represented as a Merkle tree). By design, a blockchain is resistant to modification of the data. It is "an open, distributed ledger that can record transactions between two parties efficiently and in a verifiable and permanent way". For use as a distributed ledger, a blockchain is typically managed by a peer-to-peer network collectively adhering to a protocol for inter-node communication and validating new blocks. Source: https://en.wikipedia.org/wiki/Blockchain
	Internet of Things	The Internet of things (IoT) is a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers (UIDs) and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction. The definition of the Internet of things has evolved due to the convergence of multiple technologies, real-time analytics, machine learning, commodity sensors, and embedded systems.[1] Traditional fields of embedded systems, wireless sensor networks, control systems, automation (including home and building automation), and others all contribute to enabling the Internet of things. Source: https://en.wikipedia.org/wiki/Internet_of_things
	Robotics	Robotics is an interdisciplinary branch of engineering and science that includes mechanical engineering, electronic engineering, information engineering, computer science, and others. Robotics deals with the design, construction, operation, and use of robots, as well as computer systems for their control, sensory feedback, and information processing. Source: https://en.wikipedia.org/wiki/Robotics

	High Performance Computing	High-performance computing (HPC) is the use of super computers and parallel processing techniques for solving complex computational problems. HPC technology focuses on developing parallel processing algorithms and systems by incorporating both administration and parallel computational techniques. High-performance computing is typically used for solving advanced problems and performing research activities through computer modeling, simulation and analysis. HPC systems have the ability to deliver sustained performance through the concurrent use of computing resources. The terms high-performance computing and supercomputing are sometimes used interchangeably. Source: https://www.techopedia.com/definition/4595/high-performance-computing-hpc
BDV Complementary Technologies	At least one Complementary Technology*	
	More or equal to two Complementary Technologies*	
	More or equal to three Complementary Technologies*	
	More or equal to three SRIA & Complementary Technologies*	
Network Strategy		
At least one network effect*		
on Data Level		With higher amount as well as more different types of data sources, the envisioned value proposition can be improved. I.e. with more data, the underlying algorithm can be improved as they become more accurate. Typical Examples here fore are TomTom or Google applications. In addition, data is used to improve the matchmaking between demand and supply. (OECD 2015) (Choudary 2015)
on Infrastructure Level		On the infrastructure level the availability of tools, services and rules that enable the seamless integration of technology components as well as automated orchestration of processes value creation. This plug- and play nature of the tools and services establish the basis for others, external companies and producers, to build value and services on top of this infrastructure. (Choudary 2015)
on Marketplace Level		Network effects on marketplace level describe a situation in which the increased and balanced number of producer and consumer on a marketplace increases the value of the overall marketplace. It is based on the participants on the platform and their relationships. (Choudary 2015)
Revenue Strategy		

Freemium	The term "Freemium" is a combination of "free" and "premium". According to Hartmann et al., 2014), Freemium is the most dominant revenue model in the data ecosystem in particular for start-ups in the B2C area. The general idea is to provide data-driven offerings free of charge but charge money for additional features of the offering (premium). In addition, Freemium model is often combined with the advertising-based revenue model.
Advertisement	Here the revenue stream results from fees for advertisement of a particular product, service or brand. This means, people can consume data-driven offerings free of charge or with a discount in exchange for viewing paid-for advertisements (OECD, 2014d). Increasingly, advertisement is provided based on the profile and/or location of the consumers. Advertisement-based revenue models are also very common in multisided markets with a minimum of two distinct user groups that synergically support / complement each other.
Subscription	Revenue is generated by selling continuous access to a service. Examples of subscription-based models include regular (daily, monthly or annual) payments for access to the Internet, as well as access to digital content including data, news, music, video streaming, etc. The category also includes regular payments for software services and maintenance, hosting and storage, and customer "help" services.
Usage Fee	Revenue is generated by the use of a particular service; the more service is used, the more the customer pays. Usage fees are typically charged to customers for use of a particular (online) service – including most offers that are provided "as-a-Service" (XaaS), such as cloud computing based services. These services are offered through a pay-as-you go model, where usage fees are charged for the actual use of the service.
Asset Sales	Selling of goods (including digital content) – Asset sale is still used in the data ecosystem, mainly by IT infrastructure providers. But it is also used by service platform providers that sell sensor-equipped smart devices (including smartphones, smart meters and smart cars) as a source for generating data and delivering value-added services. Furthermore, it includes pay-per-download revenue models where users pay per item of download. These could include, for instance, data sets or other digital content such as e-books, videos, apps, games and music.
Licensing	This revenue model is often used to generate revenues from intangible assets that are protected through intellectual property rights (IPRs), such as patents and copyrights. Licensing may thus be used to monetize software and software components including algorithms, libraries and APIs. It may also be used for databases.

Selling of Service	This revenue model includes the provision of traditional B2B services such as IT consultancy services, software development and maintenance and helpdesk support. It also includes a wide range of long-term B2B services provided by Internet intermediaries such as web hosting, domain registration, and payment processing. It thus overlaps with the revenue models that are based on subscriptions and usage fees often used for IT service contracts.
Commission Fee	This is mainly used in B2C markets by intermediaries that use data analytics to better match supply and demand. Payment often will be calculated on the basis of a percentage of the price of products supplied, and it will only be obtained when successfully matching supply and demand – that is, when successfully providing businesses with customers.
No information provided	On the internet no information relate to the revenue strategy / model could be found.
More or equal to three different Revenue Models*	
Type of Business	
Data-driven Service	A data-driven service is a new or considerably changed service concept, client interaction channel, service delivery system or technological concept that individually, but most likely in combination, that is based on the usage of data or data analytics. (Footnote 1) The main focus of data-driven services is to transform traditional businesses and services into data-driven services (often also labelled as smart data).(footnote 2) Data-driven service business investigates the usage of data to improve traditional business outcomes. This ranges from industrial applications, data-driven marketing to process automation. Source / Footnote: 1 = adapted from B.van Ark et al.,(2003)"Services Innovation,Performance and Policy: A Review" June, 2003, 2= https://www.siemens.com/global/en/home/products/mobility/rail-solutions/services/digital-services/smart-data.html
Data-driven Marketplace	Data-driven marketplaces are establishing plug and play mechanisms allowing to connect users and things to plug in and orchestrates them towards efficient interactions. They enable efficient social and business interactions that are mediated by software. Source: Derived from (Choudary 2015)
Niche Player	A large number of organizations follow a niche strategy. Those companies emphasize differentiation by focusing on unique capabilities and leveraging key assets provided by others. (Iansiti & Levien 2004)

Emerging Technology	in emerging and potential disruptive market set-ups. Here we refer to relevant disruptive patterns described in (Hagel et al. 2015), such as the strategy of creating markets by connecting demand and supply in an innovative manner or the strategy to position offerings as platform that establish a foundation for others to build upon.

Table 16 DDI Ontology

11 Appendix 2: Sample set of the DDI research study

Startup	Cluster	Website	Description
Altilia	A	http://altilia.ai/	Altilia develops big data solutions that enable organizations to analyze data and make more informed decisions
Artomatix	A	https://artomatix.com/	Artomatix is a software company that uses Artificial Intelligence to create realistic 3D art creation
Avaamo	A	https://avaamo.ai/	Avaamo is a deep learning Virtual Assistant Platform for the enterprise
Cappasity	A	https://www.cappasity.com/	Cappasity is a cloud-based platform that lets online stores easily create and deliver 3D, VR and AR shopping experiences
Civil Maps	A	https://civilmaps.com/	Civil Maps provides a sensor-agnostic cognition platform and pipeline that enables city scale HD 3D Mapping, Edge Mapping and Localization
Cortical.io	A	https://www.cortical.io/	Cortical.io's offers AI-based natural language understanding solutions built on technology inspired by Neuroscience
DGraph Labs	A	https://dgraph.io/	Dgraph is a fast, transactional, native and distributed graph database
Iris Automation	A	http://www.irisonboard.com/	Iris Automation is an artificial intelligence and safety avionics company building collision-avoidance systems for autonomous vehicles
Iris.ai	A	https://iris.ai/	Iris is an AI Science Assistant, helping R&D double productivity when seeking out new opportunities in published research
Lynq	A	https://lynqme.com/	Lynq changes how location data is communicated, allowing data to be transmitted for miles without networks or infrastructure
Markable	A	https://markable.ai/	Markable - See Differently - powered by visual recognition AI technology
NVMdurance	A	http://nvmdurance.com/	NVMdurance is accelerating the adoption of 3D NAND
ParallelDots	A	https://www.paralleldots.com/	ParallelDots is a technology company that develops artificial intelligence solutions for developers, startups, and enterprises

Playbasis	A	http://www.playbasis.com/	Playbasis is about social engagement & gamification
SchoolMint	A	https://www.schoolmint.com/	SchoolMint helps schools and families manage the admissions process, simply and cost effectively
Tessian (Previously CheckRecipient)	A	https://www.tessian.com/	Tessian is building the first Human Layer Security platform to protect people using email
Tractable	A	https://tractable.ai/	Tractable is a software company that develops artificial intelligence for accident and disaster recovery
TRX Systems	A	http://www.trxsystems.com/	TRX Systems is the developer of NEON®, a location system delivering tracking services inside GPS-disabled buildings
Valossa	A	https://valossa.com/	Valossa provides video recognition and content intelligence software platform for businesses working with video
Ziko	A	https://zikto.com/	Zikto provides blockchain-based claim management SaaS to insurance companies
Anagog	B	https://www.anagog.com/	Anagog is revolutionizing the way you know your customers to craft the most personal customer experience
Arable	B	http://www.arable.com/	Arable is an agricultural business intelligence solution founded on in-field measurements
Carfit	B	https://car.fit/?lang=en	Carfit leverages car's vibrations with data science to anticipate maintenance needs
Eliq	B	http://eliq.io/	Eliq is a Swedish tech company that specializes in the analysis and visualization of energy data
GeoPal Solutions	B	https://www.geopal.com/	GeoPal is a web and mobile application. It is a Solutions Software to create world-class field workforces
Locus.sh	B	https://locus.sh/	Locus is an intelligent decision-making and automation platform for logistics
Mnubo	B	http://www.mnubo.com/	Mnubo is an IoT company that provides data analytics solutions
Onfleet	B	https://onfleet.com/	Onfleet develops software for last mile delivery logistics

Placemeter	B	http://www.placemeter.com/	Placemeter is an urban intelligence platform. We quantify modern cities worldwide. The answers you need are all around you
Senseye	B	https://www.senseye.io/	Senseye is a cloud-based software for predictive maintenance
Swiftly	B	https://www.goswift.ly/	Swiftly helps cities move by changing transportation and mobility through real-time information and better data
Taranis	B	https://taranis.ag/	Taranis offers a precision intelligence platform for agriculture
Verv (Previously Green Running)	B	https://verv.energy/	Verv is an intelligent home hub, which gives the power to take control of energy usage to ultimately reduce energy bills
Waycare	B	https://waycaretech.com/	WayCare optimizes traffic management systems leveraging predictive analytics, and enables two way vehicle to city communication
AIMS Innovation	C	https://www.aims.ai/	AIMS Innovation is about pplication performance monitoring
CloudMedx Inc	C	https://www.cloudmedxhealth.com/	CloudMedx Inc. designs artificial intelligence driven software for medical analytics
Emagin	C	https://www.emagin.ai/	Emagin provides water and wastewater facilities with an operational intelligence platform that supports real-time decision making
Fraugster	C	https://fraugster.com/	Fraugster is a German-Israeli anti-fraud company that uses artificial intelligence to eliminate fraud and increase customers' profits
Oncora Medical	C	https://oncoramedical.com/	Oncora Medical is a digital health company integrating big data and machine learning into radiation oncology
Plutoshift (Previously Pluto Ai)	C	https://plutoshift.com/	Plutoshift is an automated performance monitoring for industrial workflows
Protenus	C	https://www.protenus.com/	Protenus provides healthcare compliance analytics to help health systems reduce risks
Quinvi	C	https://www.qvinci.com/	Quinvi is about financial consolidation, reporting & benchmarking software

Rulex	C	http://www.rulex.ai/	Rulex is an AI software to embed automated real time predictive intelligence in applications, infrastructure, and IoT edge apps
TradeGecko	C	https://www.tradegecko.com/	TradeGecko is an inventory and order management platform for modern wholesale merchants
Uplevel Security	C	https://www.uplevelsecurity.com/	Uplevel Security offers adaptive system of intelligence that uses graph theory and machine learning to modernize security operations
WegoWise	C	https://www.wegowise.com/	WegoWise is a web-based utility analytics platform that integrates big data with energy efficiency for the building performance industry
Agolo	D	https://www.agolo.com/	Agolo summarizes information faster and with broader coverage than any human
Apptopia	D	https://apptopia.com/	Apptopia provides services in app analytics, data mining, and business intelligence for the mobile industry
Cordial	D	https://cordial.com/	Cordial is the foundation for tailored customer experiences
Edited	D	https://edited.com/	Edited is a retail technology company that helps the world's leading brands have the right product at the right price, at the right time
HYP3R	D	https://www.crunchbase.com/organization/hyp3r	HYP3R is a location-based marketing platform that surfaces powerful geosocial data to analyze, engage and acquire high-value customers
Keywee	D	https://keywee.co/	Keywee offers publishers and marketers an approach to content distribution and performance measurement
Nice in Contact (Previously Brand Embassy)	D	https://www.niceincontact.com/	Digital Transformation: NOW Omnichannel. Agile. Scalable customer service
Onclusive (Previously AirPR)	D	https://onclusive.com/	Onclusive is the data science company for communications and PR
PinMeTo	D	https://www.pinmeto.com/	PinMeTo helps chains and franchises reach all customers on all relevant online platforms

PlayVox	D	https://www.playvox.com/	Playvox is a quality assurance software for customer service
Propeller Aero	D	https://www.propelleraero.com/	Propeller Aero is a global leader in drone-mapping and analytics solutions that give worksites the power to measure and manage themselves
Rollbar	D	https://rollbar.com/	Rollbar is a software company that helps developers deploy software with solutions that identify, prioritize, and resolve coding errors
Scanalytics	D	https://www.scanalyticsinc.com/	Scanalytics is a sensor-based analytics and engagement platform for physical spaces
Sfara	D	https://www.sfara.com/	Sfara is leading a mobile revolution in safety and control
Shareablee	D	http://shareablee.com/	Shareablee is an online platform offering audience intelligence, competitive benchmarking, and insights for social media
TVision Insights	D	https://www.tvisioninsights.com/	TVision Insights is using cutting edge technology to measure what was once unmeasurable - how people really watch TV
AreaMetrics	E	https://areametrics.com/	AreaMetrics personalizes the world around you by leveraging sensors and data
Auro Robotics	E	http://www.auro.ai/	Auro Robotics is the Self-Driving Shuttle for travel within university campuses, corporate parks, and residential communities
Carmera	E	https://www.carmera.com/	Carmera provides real-time HD maps and navigation-critical data to autonomous vehicles
Cloudcherry	E	https://cloudcherry.com/	CloudCherry Experience Management platform that is disrupting the way organizations listen to the Voice of Customer
Cyclia	E	https://cyclicarx.com/	Cyclia is driving drug discovery, by doing more with AI
Desktop Genetics	E	https://www.deskgen.com/landing/#/	Desktop Genetics is an international company to help researchers discover and treat the root genetic causes of human disease

Metadata	E	https://www.metadata.io/	Metadata automates demand generation for B2B companies using data enrichment and targeted advertising
Nexosis	E	https://www.datarobot.com/nexosis/	Nexosis provides a machine learning API for developers
Taplytics	E	https://taplytics.com/	Taplytics Customer Experience Automation transforms data into universally actionable campaigns, without the risk of privacy breach
UnaliWear	E	https://www.unaliwear.com/	UnaliWear is a Wearable 'OnStar for People' + 'NEST for People'
Visiblee	E	https://www.visiblee.io/en/home/	Visiblee is a generation automation solution to boost B2B website conversion
Vivacity Labs	E	https://vivacitylabs.com/	Vivacity Labs makes intelligent cameras to gather transport data, using the latest machine learning & computer vision techniques
Warwick Analytics	E	http://www.warwickanalytics.com/	Warwick Analytics offers advanced predictive analytics and process automation
Bird.i	F	https://www.hibirdi.com/	Bird.i provides access to the latest, up to date satellite imagery, helping you obtain valuable insights & make better business decisions
BridgeU	F	https://bridge-u.com/	BridgeU helps young people around the world realise their potential, through empowering schools to provide smart, modern university
CleverTap	F	https://clevertap.com/	CleverTap is a customer lifecycle management and engagement platform that drives long-term retention and growth
CrossTarget	F	http://www.crosstarget.co.kr/	CrossTarget by OnnuriDMC is a data-driven mobile advertising tech company to ensure advertisers to have higher ROI than any others
Hooch	F	https://hoochrewards.com/	Hooch is a subscription-only cocktail app and discovery platform
Influenster	F	https://www.influenster.com/	Influenster is a digital destination where millions of consumers research and review products and where brands can connect with shoppers

Insurify	F	https://insurify.com/	Insurify is the most comprehensive insurance quotes comparison platform in America
LoanTab	F	https://loantap.in/	LoanTap is an online platform committed to deliver flexible loan products to salaried professionals
Mila	F	https://www.mila.com/en-de/	Mila is a crowd services platform
niki.ai	F	https://niki.ai/	Niki.ai is building internet commerce in the style that's most natural to Bharat
Paysa	F	https://www.paysa.com/	Paysa is helping employees and employers understand and identify opportunities in the labor market
Playment	F	https://playment.io/	Playment is a fully managed data labeling platform generating training data for computer vision models at scale
Selectionnist	F	http://business.selectionnist.com/	Selectionnist is an image recognition technology with the aim to connect print readers with the world's largest brands through an application or a chatbot
Shared2You	F	http://www.shared2you.com/	Shared2you is a company capturing data about the real-time movement and growth of the global mobile app economy
Zizoo	F	https://www.zizoo.com/	Zizoo is a leading boat holiday platform

Table 17 Sample set of the DDI research study