Statistical Learning for business decision making: from big data to informative data

Thesis presented for the degree of
Doctor of Philosophy

XXXV cycle
Economics and Policies of Markets and Enterprises

University of Salerno

Coordinator: Professor Alessandra Amendola

Amendola

Supervisor: Professor Michele La Rocca

Candidate: Ivan Colosimo

Ivan Colosimo

.

Contents

Acknowledge 5

Introduction 7

1. Statistics and statistical learning 13

2. Data science: from data to information 14

1.1 Statistical decision support system (DDS) 17

1.2 Decision Making & Decision Support System 19

1.3 Information society 21

1.4 What we intend for Big Data Analytics 23

1.5 Big data analytics and the data-driven approach: tools to serve business 24

1.6 The steps in a Big Data Analytics project 25

3. The importance of data avaiability 28

3.1 How Big Data is collected 28

3.2 The data Catalog 31

3.3 Is data enough to extract information? 33

3.4 Data Catalog: why is important? 35

3.5 Data Governance: What we mean by it? 38

3.6 The problem of the lack of Data Governance in enterprise environment 39

3.7 Data Quality in the Data Governance Process 41

3.8 Data Quality Management 42

4. Data strategy to maximize the value of data. 43

4.1 How to implement a Data Strategy 43

4.2 Big Data: the role of data strategy 45

4.3 Impact and benefits on marketing, sales, and customer experience 46

4.4 The stages of a Data Strategy: from Assessment to the creation of a Roadmap 46

4.5 Implementation Strategies 47

4.6 Data-driven organization maturity model 48

5. A statistical decision support system for tourism data analytics 50

5.1 Big Social Data Analytics and travel experience in the integrated mobility business.52

5.2 Value configuration 52

5.3 Value creation and value capture mechanisms 53

5.4 Methodology 54

5.4.1 Recommender System: a predictive tool to create knowledge 54

5.4.2 Findings 57

5.4.3 Discussion and implications 60

5.4.4 Conclusions, limitations, and future research 61

6. Online data: Sentiment analysis with Python 63

6.1 Natural language processing and Sentiment Analysis 64

6.2 BERT-based Twitter Sentiment Analysis with Python 67

6.3 Dataset and Data Preparation 68

6.3.1 Data Collection & Data Labeling 68

6.3.2 Extraction of Text from Images 69

6.4 Results, limitations and extensions 72

7. Segmenting data with Python. A quantitative analysis on the household savings 75

7.1 Undestanding the determinants of the household savings in Italy 77

7.1.1 Literature review on household savings 77

7.1.2 The Italian context 79

7.2 Methodology 80

7.2.1 Data Collection 82

7.2.2 Data Preparation 82

7.2.3 Cluster Analysis – Data Analytics 83

7.2.4 Data Visualization 83

7.2.5 Pattern recognition 84

7.3 Findings and outcomes 84

7.4 Discussion and main implications 90

7.4.1 Theoretical implications 92

7.4.2 Practical implications 93

7.4.3 Social implications 93

7.4.4 Research limitations and future research 93

Appendices 95

References 98

Acknowledge

First, I would like to express my gratitude to my principal supervisor, Professor Michele La Rocca, for his continuous support and invaluable guidance throughout my executive doctoral study. It has been an immense privilege to be able to embark on this doctoral journey several years after completing my master's thesis. I finished my undergraduate studies in 2008 and never thought I would venture on this path 10 years later, moreover with a responsible job that occupies entire days (sometimes even weekends). If I managed to make it all coincide, I owe it to Prof. Michele La Rocca.

Special thanks to Professor Maria Teresa Cuomo, who was also invaluable and indispensable for the publications we were able to produce, for her support on the topics more related to marketing, and for providing the working group with her network of knowledge that enabled me to produce the scientific publications.

Thanks to Lorenzo Ricciardi Celsi, a colleague and friend who perhaps unintentionally was an incentive to emulate him.

I would like to thank my wife for the stimulus to throw myself into this adventure, for spending many moments together at home, and for not making me burdened by the fact that we could have gone out instead of staying home to study. For standing by me in life.

A final thanks to Roberto Sorrenti (my Manager) over the years of working at the ELIS Center. Without him, a PhD path did not exist in my mind.

Introduction

The advent of the data economy - defined as the value and the impact derived from the collection, storage, analysis, and utilization of data - has made indispensable in business:

- the knowledge of Statistics;
- the presence of the data scientist figure in the company to manage data and help make decisions based on empirical data;
- the need for easy access to available data;
- the need to implement a proper data strategy to maximize data value.

In this thesis, we will pitch out these four aspects that are fundamental in modern businesses. The first aspect is about the growing importance of Statistics in business, which has two purposes: synthesizing and generalizing.

Synthesizing means preparing the data collected in form (tables, graphs, numerical summaries) that enables a better understanding of the phenomena for which the survey was carried out. Synthesis meets the need for simplification, which stems from the human mind's limited capacity to handle articulated, complex, or multidimensional information. Descriptive Statistics is a specific field of Statistics that describe how to use techniques that allow a comprehensive study of a large amount of quantitative and qualitative information to highlight its characteristics, links, differences or associations among the variables surveyed.

Generalizing means extending the result of the analysis performed on a limited group of statistical units (sample) data to the entire community to which it belongs (universe, population). This operation of generalizing is carried out according to methods of induction, which represent the content of Inferential Statistics.

With the Inferential Statistics and Decision Theory closely related to the calculus of probability, it is possible to understand and interpret uncertain or random events and suggest rational rules of behavior and decision-making.

The contribution of Statistics is not limited only to the data analysis phase. Indeed, its real added value is expressed in the formulation of hypotheses of research, in the argumentation of theses, in the adoption of appropriate solutions and appropriate methodologies, in the choice of survey methods, in the formulation of the sample, in the procedures for extending the results to the reference universe.

Keeping these steps under control means producing reliable and cost-effective results and, in line with the definition of Business Statistics, master both techniques of Descriptive Statistics and Data and Inferential Analysis.

Statistical tools and methods thus enable us to deal rationally and effectively with various business problems.

By limiting the exposition to general purposes and thus ignoring the technical and specialist aspects of the applications of Statistics to the business environment, it must be said that this discipline is used successfully both in the analysis of production and management processes and results, and in assessment of the conditions of the market in which the company operates; in addition, the applications of Statistics for forecasting purposes, on market and sales trends, are beneficial, especially in the current economic situation; finally, the contribution of Statistics in the planning of marketing strategies is particularly significant.

Predict in the short or medium-long term how much the company will be able to sell in the future and what the market demand will be, understood as the primary demand for the asset in which the company is interested, is crucial for the entrepreneur and business manager because combining the two forecasts makes it possible to determine how much the company's market share may vary.

These kinds of forecasts involve "internal" assessments of the company's production capacity and "external" assessments of general market trends, absorption capacity, new entries or segments, or the need to strengthen the company's presence.

The role of Statistics can be fundamental to help a company to develop in several areas, for example:

- Evaluate if and to what extent customers are satisfied with the product that the company produces or with the service it offers. This is the so-called "customer satisfaction", a crucial aspect to which the company must increasingly refer because it allows it to acquire or consolidate competitive advantages on the market over time. Having a precise picture of the degree of satisfaction of one's customers means having solid tools to face the competition; knowledge of consumer preferences determines the success or failure of planned marketing strategies. The greater the customer loyalty, the lower the competitive pressure, the more solid the market position of the company and the more valuable is the stock of intangible resources ("fiduciary capital") available to the company to consolidate the position and plan new marketing strategies.

- Analyze the characteristics of one’s customers or potential customers to adopt differentiated commercial strategies or evaluate the entry of new products or services. Since for each type of product or service there is a certain heterogeneity of consumers, it is necessary to apply market segmentation techniques aimed at identifying the different market demands with respect to which to adapt suitable products and sales policies.

In recent years, thanks to the increased volume of data to manage, analyze and interpret, Statistics has been joined by Information Technology, because traditional approaches to information processing are not sufficient to manage large amounts of data (Big Data). Actually, the amount of data is not the real issue, but rather the way in which data are used: through the use of algorithms capable of processing many variables quickly and with few computational resources.

The executive research doctorate aims to train employees who, while maintaining their professional role in the company, can easly deepen the issues faced in the professional context. This work results from a close collaboration between the College of Professors of the University of Salerno and the ELIS Consortium started in the academic year 2019/2020. In particular this *executive doctoral thesis* is based on research activities carried out through the development of 3 industrial projects carried out for some large Italian companies (Ferrovie dello Stato Italiane, RAI, Cassa Depositi e Prestiti). The first goal of each project, therefore, is not the production of a scientific paper but rather the creation of a tool that could solve a scientific problem.

Among the most common problems companies face is the collection and interpretation of data.

Companies today collect huge amounts of data on customers, operations, social networks, etc. In this flourishing of data, the problem of enhancing the data itself emerges, that is, moving from data to information. Statistical Learning is the technique used to extract information from data and make predictions about possible phenomena that may occur.

Statistical learning therefore helps to make more informed decisions in an environment of uncertainty. And we know how uncertain today's world is.

Therefore, this thesis aims to provide an approach to data-aware use in business. Data, in fact, increasingly represent a corporate asset that must be enhanced in order to derive value from it: moving from data to information (Chapter 2). To be properly used, however, data, must possess certain characteristics (Chapter 3). The valorization of them comes through the implementation of a data strategy (chapter 4).

At this point, we will look at 3 practical applications.

1. Enhancing Traveller Experience in Integrated mobility services via big social data analytics

I had the opportunity to work on a project commissioned by Nugo (startup of the Ferrovie dello Stato group). Nugo offers door-to-door mobility services and my project was aimed at providing a scientific contribution showing how a data-driven approach can improve and enhance the tourist experience in integrated door-to-door mobility services. In particular, the data driven approach, thanks to the design of a recommendation system based on a big data analysis engine, makes it possible to: i) classify tourists' preferences for the most attractive Italian destinations in Google; ii) classify the main attractions (leisure, entertainment, culture, etc.) associated with individual tourist destinations, obtained from the analysis of relevant thematic websites such as: Tripadvisor, Minube and Travel365.

The deliverable of this activity was the creation of a software application called Rankingtrends, created using the following technologies: Python3, MySQL, Pandas, Docker. Furthermore, the work led to the drafting of a paper which was presented at the ACIEK Conference and than published in the journal *Technological Forecasting and Social Change*:

- *Enhancing Traveler Experience In Integrated Mobility Services Via Big Social Data Analytics*, Maria Teresa Cuomo, **Ivan Colosimo**, Lorenzo Ricciardi Celsi, Roberto Ferulano, Giuseppe Festa, Michele La Rocca. Conference ACIEK (Academy of Innovation, Entrepreneurship, and Knowledge) 2021: 14TH edition - Innovation, Management, and Governance for Sustainable Growth. 14-16 June, Paris Sorbonne (Virtual conference). ***Best paper award.***
- Maria Teresa Cuomo, **Ivan Colosimo**, Lorenzo Ricciardi Celsi, Roberto Ferulano, Giuseppe Festa, Michele La Rocca. Enhancing Traveller Experience in Integrated Mobility Services Via Big Social Data Analytics. *Technological Forecasting and Social Change*, Volume 176, 2022. (IF 10.884). https://doi.org/10.1016/j.techfore.2021.121460

2. Online data analysis: Sentiment analysis with Python

The second activity resulted from a project carried out for RAI. For some time, many comments, judgments and discussions on television broadcasts have been taking place on social networks. For this reason, RAI is interested in acquiring textual data referring to specific broadcasts to evaluate their approval. For this reason, natural language processing approaches have been proposed that capture users' sentiment towards broadcast.

- Software Application Instarai. Technologies used: Python3, MySQL, Pandas, Docker, Power BI.
- Online data analysis: Sentiment analysis with Python. Maria Teresa Cuomo, Lorenzo Baiocco, **Ivan Colosimo**, Egon Ferri, Michele La Rocca, Lorenzo Ricciardi Celsi. In Research Methods in Marketing, Business and Management: Theoretical and Practical Perspective (Emerald, Pantea Foroudi Editor).

3. Segmenting with big data analytics and Python: a quantitative analysis on the household savings

During the last academic year, we worked on a research activity that was inspired by a project commissioned by Cassa Depositi e Prestiti (CDP). CDP is a true "state bank," a financial institution that works through its services and group companies to support the country's development. CDP is the issuer of State-guaranteed postal savings bonds and passbook savings accounts. These products are distributed by Poste Italiane through its more than 12,000 branches located throughout the country. During the last academic year this study underlines as, under advanced analytics tools, household saving behaviours information and big data analytics may support data-driven decision approaches addressed in the managing of complex relationships in the financial arena. More punctually, using an exploratory and predictive analysis based on big data analytics and machine learning, this study aims to provide extensive customer profiling in the Italian household saving sector, supporting a data-driven decision-making approach. In this direction, the machine learning techniques have been aimed at behavioural segmentation, taking into consideration: i) homogeneous customers profile, ii) purchase/repayment paths and iii) churn.

- Software Application Household Monitoring. Technologies used: Python3, MySQL, Pandas, Docker, Tableau, Cloudera.
- Segmenting with big data analytics and Python. a quantitative analysis on the household savings. Maria Teresa Cuomo, Michele La Rocca, **Ivan Colosimo**, Debora Tortora, Lorenzo Ricciardi Celsi, Rosario Portera, Giuseppe Festa. Conference ACIEK (Academy of

Innovation, Entrepreneurship, and Knowledge) 2022: 16TH edition – Greening, Digitizing and Redefining Aim in an Uncertain and Finite World. 28-30 June, Seville.

- Segmenting with big data analytics and Python. a quantitative analysis on the household savings. Maria Teresa Cuomo, Michele La Rocca, **Ivan Colosimo**, Debora Tortora, Lorenzo Ricciardi Celsi, Rosario Portera, Giuseppe Festa. Published to *Technological Forecasting and Social Change.*

The dissemination of the three papers was done in a nonstatistical environment.

1. Statistics and statistical learning

The technology mentioned earlier has supported the development of increasingly sophisticated and fascinating data analysis methodologies. As is often the case, a solution was born from the combination of a need and an opportunity: machine learning.

Machine learning is undoubtedly one of the currently hottest trends, and as much as we talk about it, one might be led to think that traditional Statistics is now a distant memory. This section will try to better understand the distinction between Statistics, machine learning and statistical learning.

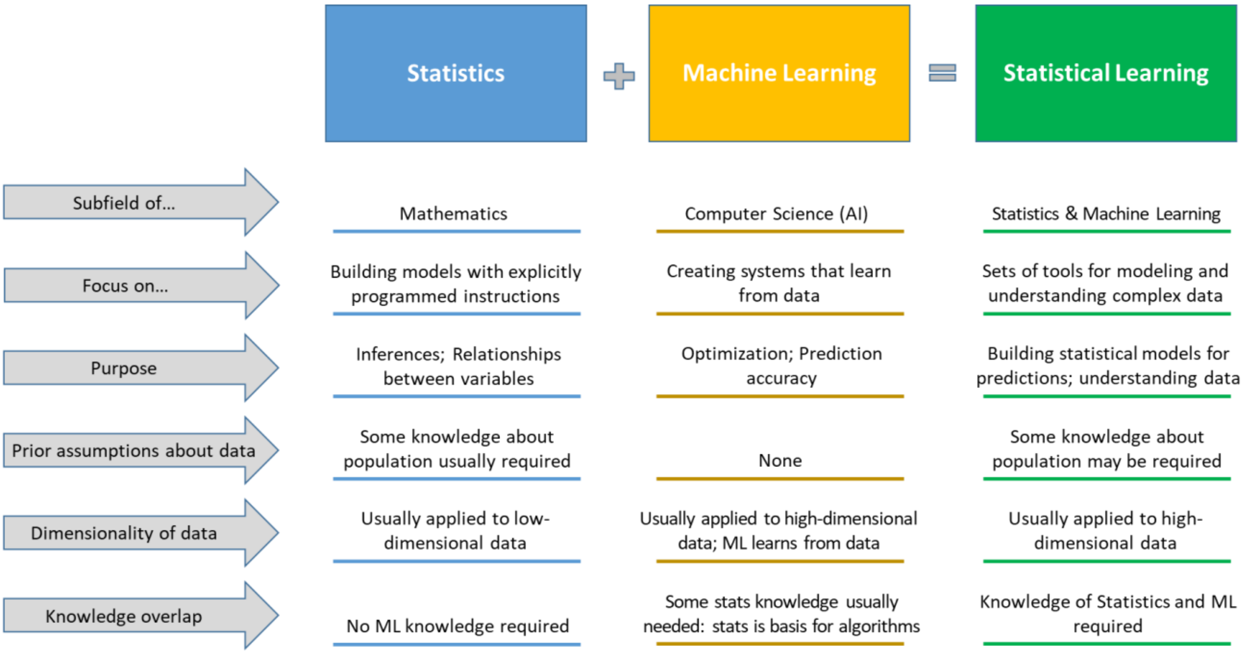
Machine Learning is the branch of computer science that utilizes past experience to learn from and use its knowledge to make future decisions. Machine Learning is at the intersection of computer science, engineering and Statistics. Machine learning aims to generalize a discernible pattern or create an unknown rule from given examples.

The table below illustrates the difference between statistical modelling and machine learning.

Statistical Modeling	Machine Learning
Formalization of relationship between variables in the form of mathematical equation	Algorithm that can learn from the data without relying on rule based programming.
Required to assume shape of the model curve prior to perform model fitting on the data	Does not need to assume underlying shape, as machine learning can learn complex patterns automatically based on the provided data.
Statistical models predict the output with accuracy of 85% and having 90% confidence about it.	Machine Learning just predict the output with accuracy of 85%.
In Statistical modeling, various diagnostic of parameters is performed, like p-value, and so on.	Machine Learning models do not perform any statistical diagnostic significance tests
Data will be split in 70%/30% to create training and testing data. Models developed on training and tested on testing data	Data will be split in 50%-25%-25% to create training, validation and testing data. Models developed on training and hyparameters are turned on validation data and finallyget evaluated against test data.
Statistical models can be developed on a single dataset called training data, as diagnostic are performed at both overall accuracy and individual variable level	Due to lack of diagnostic on variables, machine learning algorithms need to be trained on two dataset, called training and validation data, to ensure two point validation.
Statistical modeling is mostly used for research purposes	Machine Learning is very apt for implementation in a product environment
From the school of statistic and mathematics	From the school of computer science.

Table 1Difference between Statistical Modeling and Machine Learning. Source: (Dangeti, Pratap Statistics for Machine Learning)

Statistical learning refers to a vast set of tools for understanding data. These tools can be classified as supervised or unsupervised. Broadly speaking, supervised statistical learning involves building a statistical model for predicting or estimating an output based on one or more inputs. Problems of this nature occur in fields as diverse as business, medicine, astrophysics, and public policy. With unsupervised statistical learning, there are inputs but no supervising output; nevertheless, we can learn relationships and structure from such data. To illustrate some applications of statistical learning, we briefly discuss three real-world data sets considered in this book.



	Statistics	+	Machine Learning	=	Statistical Learning
Subfield of...	Mathematics		Computer Science (AI)		Statistics & Machine Learning
Focus on...	Building models with explicitly programmed instructions		Creating systems that learn from data		Sets of tools for modeling and understanding complex data
Purpose	Inferences; Relationships between variables		Optimization; Prediction accuracy		Building statistical models for predictions; understanding data
Prior assumptions about data	Some knowledge about population usually required		None		Some knowledge about population may be required
Dimensionality of data	Usually applied to low-dimensional data		Usually applied to high-dimensional data; ML learns from data		Usually applied to high-dimensional data
Knowledge overlap	No ML knowledge required		Some stats knowledge usually needed: stats is basis for algorithms		Knowledge of Statistics and ML required

Figure 1: Difference between Statistics, Machine Learning and Statistical Learning. Musio Image Akawipic

2. Data science: from data to information

Starting with the decision-making process in the company, we can see how it plays a central role in business management. The great importance given to this issue is justified by the fact that without a proper and well-structured decision-making process, managers of functional areas (but more generally, all levels of the company) could not validly determine their goals and how to achieve them.

Keeping the focus on the business context, we can define the decision-making process as detecting the problem, evaluating the alternatives, selecting the best solution among different options, and implementing the necessary actions to achieve the set goal. However, being an organization, an association composed of several subjects, in the decision-making process, one should also take into account "...the reactions of other subjects [also endowed with decision-making autonomy] with whom the decision maker is in a situation of interdependence"[1] (Costa, Gubitta and Pittino 2014 p. 37).

To best understand the internal process within the company, one must also consider the decision-making dynamics within an individual's mind. In fact, at any given moment, the individual is faced with a plurality of actions from which, through a specific process (which may or may not be conscious), he or she arrives at a choice and thus at the decision of one among several alternatives.

We can generally think of this process as a rational procedure on the part of the individual, but actually, based on Simon's studies, entirely rational choice is a utopian situation. In fact, if every individual were endowed with absolute rationality, this would imply that he or she is endowed with a set of stable preferences and is aware of all possible alternatives and the valuable information to evaluate them. All this would lead the subject to choose the best possible alternative to maximize his utility function. As can be clearly seen, however, this situation is not consistent with reality and is replaced with the concept of bounded rationality. In this view, it is believed that the human mind is presented with only some of the possible alternatives and that it is unable to calculate all the possible consequences, since the ultimate goal to be achieved is located at some future time after the choice is made. From this it follows that, because of the difficulties encountered during the decision-making process, the individual does not tend toward the optimal choice but rather toward the "satisfactory" choice. In summary, there has been a shift from a conception of an entirely rational man to one who is only intentionally rational.

In addition, the importance of the emotional part of decision-making depending on the system activated in the individual's mind, by Kahneman (2011) called System 1 and System 2 (see Hobfeld 2017 and Kannengiesser 2019). In particular, the former operates automatically and quickly, almost impulsively, without any effort or sense of voluntary control. On the other hand, the second one devotes attention to demanding mental activities that require a certain amount of effort and control, thus giving weight to the more rational aspect of the human mind. In general, the latter is based on System 1, and only if necessary, will the second be activated. Ultimately, therefore, it can be guessed from this theory that decision-making is different from purely rational decision-making due to the importance given to the part of impulsive and emotional thinking inherent in human nature.

Keeping in mind what has been described and the fact that organizations reflect the behaviors of the individuals that make up the organization, let us now shift our attention to the corporate level and try to understand the different forms of decision-making that can be encountered within them. The literature, in general, distinguishes between two types of organizational decisions: planned and unplanned decisions. The former refers to problems perceived as recurrent and routine and for the resolution of which a standard procedure is devised. Consequently, the resulting solutions are mainly based on well-defined procedures and rules usually shared by the actors involved. These decisions are implemented by the lower-middle management level and are based on an extensive repertoire of available information.

On the other hand, unscheduled decisions are generally related to circumstances that are difficult to relate to a standard procedure because the theme of uniqueness and significant impact on the company distinguishes them. These decisions, therefore, require an articulated decision-making process generally carried out by top management, in which various alternatives for action must be considered and analyzed with critical thinking. All this is accompanied by the collection of information not directly available in the company, which may nevertheless be incomplete or ambiguous. It follows, therefore, that more time will have to be devoted to the entire decision-making process than to routine decisions. Moreover, it may generally be subject to a higher probability of error.

The spread of artificial intelligence applications, i.e., algorithms that enable machines to emulate the cognitive abilities of human beings, has dramatically impacted the above decision-making processes. Artificial intelligence, in contrast to other types of automation (explicitly coded to act in a predefined way), is the ability to understand from data analysis what kind of action needs to be taken to complete a given task.

The three essential and general ingredients for artificial intelligence are Big Data, that is, the ability to use and benefit from large amounts of data; Algorithms suitable for using this data; and Machine Learning. The latter, in accordance with studies and research carried out by Accenture (2018), is the fundamental concept on which artificial intelligence is based today and consists of the set of possible learning methods that allow the machine and software to perform a task or activity without being programmed in advance and thus to perform it autonomously.

The spread of artificial intelligence applications, i.e., algorithms that enable machines to emulate the cognitive abilities of human beings, has impacted the above decision-making processes significantly. AI, in fact, unlike other types of automation (explicitly coded to act in a predefined way), from the analysis of data allows us to understand what kind of action needs to be taken to complete a given task through three ingredients:

- the Big Data, i.e., the ability to use and benefit from large amounts of data,
- the Algorithms suitable for using this data,
- Statistical learning.

The latter, in accordance with studies and research carried out by Accenture (2020), is the fundamental concept on which artificial intelligence is based today and consists of the set of possible learning methods that allow the machine and software to perform a task or activity without being programmed in advance and thus to perform it autonomously.

All of the more advanced technologies (such as Expert Systems, Predictive Systems, or Decision Support Systems) that can be applied within companies to support decision-making are based on Machine/Statistical Learning. In this regard, a strength of this topic is that there can be different types of machine learning algorithms: (1) supervised learning, where input and output examples are administered to make the artificial intelligence understand how to behave to achieve specific goals, (2) unsupervised learning, which is based on the analysis of results, i.e., in this case, the system learns autonomously and exclusively from its mistakes, (3) reinforcement learning, which uses a trial-and-error system in highly variable environments, i.e., a feedback loop of "rewards" when the AI achieves certain goals or results, or otherwise "punishments" so that it understands the correct actions to take and the wrong ones.

All of these concepts, as we shall see, have become increasingly prevalent in the scientific literature and society due to the increase in the amount of data generated by human beings.

1.1 Statistical decision support system (DDS)

As reported in the previous pages, the rapid development of technology and information technology has pushed business activities in every sector toward a "smart," data-driven approach. This has shifted the focus to the customer and his needs and requirements. In this context, information becomes a strategic resource to support decision-making and is considered a tangible business asset.

In this context, Decision Support Systems (DSS), software to support decision-makers that can provide valuable information for decision-making processes quickly and versatilely, thus helping them to be in complete control of their business environment and to make informed, data-driven decisions accordingly, are located.

DSSs, Decision Support Systems, are, therefore, software systems that make available to the user, the decision maker, a set of data analysis capabilities through the application of mathematical and statistical Machine Learning models in a rapid, interactive and straightforward manner to increase the efficiency and effectiveness of the decision-making process. From this definition, some aspects of DSS systems emerge that characterize them from the point of view of the service they provide to the decision-maker:

- Ease of use, intuitiveness and flexibility of the interface.
- The interactivity of the analytical environment and interface.
- Effectiveness and usefulness of the analytical models and data of interest.

From an application point of view, developing a decision support system basically goes through 4 main phases:

- *"Smart" phase*. It is the phase in which analysts, data scientists and domain experts work together to define the list of valuable data and information from inside and outside the company. This phase also identifies the problem to be addressed; data are selected according to it.
- *"Design" phase.* It is when analytical experimentation is carried out and ends with constructing the analytical model (often not just one) from which various possible solutions are generated.
- *"Choice" phase.* The different solutions developed are evaluated. This phase aims to choose the optimal solution (one or more than one) concerning the business context and the problem being addressed. Field testing activities are included in the choice process.
- *"Implementation" phase.* The DSS is realized by implementing the chosen solution.
- Finally, a fifth phase could be named that relates to end-user feedback, the moment when decision-makers begin to use the DSS, put in place in the company, in their daily lives, not only in a theoretical way but to make real decisions. Their evaluation is crucial to validate the previous steps, approve the DSS, or correct where necessary.

From the implementation point of view, one faces some technical issues during the above stages. Although they may seem unrelated to the ultimate goal, these choices should be noticed and addressed. Still, instead, it is essential to evaluate from the outset to avoid finding oneself having implemented DSS systems with shortcomings that are difficult to fix in retrospect:

- managing large amounts of data: this is a common challenge nowadays where the multiplicity of data sources and Big Data raises the question from a technological and architectural point of view of making choices that are appropriate for the expected amount of data and also for the type of analysis that you will want to apply.
- accessing different data sources on different platforms: this issue is closely related to the first one. In fact, we are increasingly faced with situations where data are found on different sources and platforms for various reasons. It is essential to have a complete overview of them to integrate them as best as possible and identify the beneficial information. Critical issues can be solved using Data Virtualization tools to create a single interface for accessing data by abstracting from their physical location and access modes.
- providing access to multiple users with different permissions because it is true that a DSS system must also be democratic, but that does not mean allowing everyone to see everything. It is necessary to provide a profiling policy and manage permissions to access and analyses different information. Aspects related to access and thus control and management of data can be quickly resolved through Data Governance, the discipline that deals with bringing order to data to allow complete control;
- managing historical versions of data: analytical models need historical datasets to be trained properly. On the other hand, these data are often not queried directly precisely because they relate to past periods and are no longer attractive. Managing historical data is, therefore, essential in terms of analytical feasibility and coping with query speed requirements.

1.2 Decision Making & Decision Support System

The most important concept to remember is that when we talk about artificial intelligence and, more specifically, Statistical Learning, we are talking about algorithms that essentially perform predictive processes without any skill or degree of judgment. When we talk about human intelligence, on the other hand, we look primarily at the inherent ability of humans to articulate judgment based on their intuition, experience, and creativity and untether, in part or whole, from a purely rational thought process. Numerous authors identify this characteristic as humans' primary competitive advantage over these increasingly sophisticated and intelligent machines. When the company faces uncertain and ambiguous situations for which no past data or evidence is available, the decision maker's intuition, imagination and creativity are the most effective means of identifying the best solution. In contrast, however, concerning AI, we can define prediction as "the process of filling in the information missing. Prediction takes the information one has, called data, and uses it to generate the information one does not have" (Agrawal, Gans, Goldfarb 2018). Thanks to this distinctive feature, in situations considered complex due to a large number of variables and data to be analyzed, among which to find valuable correlations, algorithms can identify in a short time new and creative opportunities for solving the considered problem. Moreover, thanks to new developments and advances in Deep Learning, these algorithms are now able to learn directly from raw data (Jarrahi 2018) and autonomously generate new knowledge, thus evaluating the benefits and costs of each possible scenario.

The result of this difference is a different approach of the two bits of intelligence toward the decision-making process that can take place in the company. A greater emphasis on judgment or prediction leads the decision-maker to develop different characteristics and require heterogeneous information from each other. Specifically, artificial intelligence is better able to support analytical decision-making, while conversely, human intelligence is more useful in an intuitive process (Jarrahi 2018). In particular, as reported by Shrestha, Ben-Menahem and Von Krogh 2019, to identify the best alternative, current algorithms and AI, in general, need well-delineated boundaries that define a specific area in which to process the available data and possibly create new data. In contrast, as mentioned earlier, the decision-making process followed by a person can be entirely unrelated to rational thinking and based on intuition and personal creativity. Linked to this aspect, in some situations, the decision-making process of a given algorithm may be complicated for humans to understand because it is based on purely rational functions, which, paradoxically, may lead to a solution perceived as not logical. This does not mean that the alternative indicated by the AI is not rational in itself, but that the logical process used to arrive at that point is so complex and difficult to understand that it may even seem unrelated to the initial problem, thus making its explanation even more arduous. Conversely, on the other hand, if a person makes the decision, that person can explain the logical thread and reasoning that led him or her to select a particular alternative over the others.

At the same time, however, human intelligence is limited by its inability to find new correlations among large amounts of data. It is exposed to a higher percentage of influence on the output. This influence, which can be curbed to some extent with AI, may be due to discordant interests among the parties involved in the decision-making process or the reduced time one has to make the decision. In general, however, it should be kept in mind that the decision-making processes faced by the enterprise are marked by a mix of all these characteristics peculiar to each intelligence. Consequently, the most advantageous strategy is to use both complementary (Jarrahi 2018).

1.3 Information society

As we said, the amount of data available has grown with the advent of the information society. The consequences are there for all to see: a smartphone in every pocket, a computer in every backpack, and an estimated 75 billion connected objects by 2025.

There are several definitions of "Big Data" in the literature today. The initial idea was that the volume of information had grown so large that it was no longer compatible with the memory used by computers for processing, so engineers needed to update their analysis tools. This is the origin of new technologies such as Google's Map-Reduce and its open-source equivalent, Hadoop, created by Yahoo. Over the years, several technologies have emerged to process big data: Apache Spark, Google Big Query, and many others. These technologies allow much more significant amounts of data to be processed than before, and most importantly, the data need not be arranged in ordered files or classic database tables.

In the meantime, data analysis technologies have been developed that make it possible to dispense with the rigid hierarchies and homogeneity of data, a necessary condition in the past. Likewise, large technology companies have developed new data processing techniques to provide insights and predict phenomena, becoming central to everyday life.

According to IDC (the world's premier company specializing in market research, consulting services, and event organization in the ICT and digital innovation sectors), data is the new basis for competitive advantage, whether structured or unstructured, human- or machine-generated, stored in data centers or the cloud. By leveraging large amounts and diversity of data to uncover patterns and pursue breakthrough ideas, an enterprise can win the war in the growing competitive enterprise landscape. Storage is an integral part of an organization’s data strategy as it actively contributes to storing and analyzing information. The challenge for IDC is to build storage systems that can handle large volumes of data but keep costs low without compromising performance. Moreover, in the age of data deluge and data proliferation, companies that still need to get up-to-date and evolve skill sets are falling behind. IDC believes that more enterprise infrastructures will use artificial intelligence and machine learning algorithms to support business agility needs and manage the skills gap. As the world's leading organizations seek to create and deliver digital experiences for IDC, there is no doubt that data, and the ability to extract meaningful information and actions from that data, will be at the center of these efforts.

The volume and variety of data continue to pervade organizations at all levels at an ever-increasing rate, convincing managers to derive value from it and determine its impact on the business. All our daily activities produce data: user posts on Facebook, a Google search, using an app, making a purchase on an online site, and in physical stores with credit cards and loyalty cards. And more: a photo, a video, a voice message, a tweet, itineraries, comments, and likes. Not to mention the data created by objects interconnected to the network: smart infrastructure in cities, sensors mounted on buildings and transportation, and smart home appliances. Each of us is constantly contributing to the production of data. To give an idea of what the term "we" means, we need to consider the number of people who comprise it: it includes Internet users, who number almost four billion worldwide, those who are active on social media, almost three billion, and smartphone owners, according to various estimates more than four billion, numbers of which are constantly being updated because they are constantly growing. It is clear, then, that with these numbers, the amount of information collected in real-time is expanding exponentially.

IDC predicts that by 2025 the global data sphere will increase to 163 billion zettabytes (one ZB equals one trillion gigabytes), which means ten times more than the 16.1 ZB of data that existed only two years ago. More than a quarter of this data will be real-time; of that, real-time IoT data will be more than 95 per cent. Numbers unthinkable in 1986, when (according to Gartner) the volume of data in circulation amounted to only 281 petabytes (a petabyte is one-millionth of a zettabyte), but also in 1993, when it had become 471 petabytes (+68%). Over the next seven years, however, the growth was much higher: by 2000, the volume had almost quintupled, reaching 2.2 exabytes (one exabyte is one-thousandth of a petabyte and one-thousandth of a zettabyte). Seven years later, in 2007, it was 65 exabytes, a thirty-fold increase. From then to 2016, however, the growth was more than 247 times. According to some estimates, the volume of data is increasing so much year on year that 90 per cent of what exists today was created in the last two years. According to IDC, by 2025, Third Platform technologies and services (cloud, social, mobile and big data) will drive about 75% of IT spending, growing at twice the rate of the total IT market. Over the next three years, digital transformation will reshape the entire macroeconomy as most global enterprises' revenue centers on digital products and services. [IDC website]

As can be seen, therefore, data are becoming increasingly important in the organization of production and exchange activities, so much so that they are considered an economic resource to all intents and purposes, indeed by far the most important resource in many sectors.

Indeed, thanks to advances in information and communication technology (ICT), organizations tend to collect data of all kinds, process them in real-time to improve decision-making processes and store them permanently so that they can be reused in the future or extract new knowledge.

1.4 What we intend for Big Data Analytics

As we have seen, the digital revolution has introduced new opportunities for analysis and action through the use of Big Data: not only are the masses of data growing exponentially compared to the past, but today new types of data are available, in all sectors: Digital Marketing, from purchase transactions, Internet of Things, machines in industry, from Social Media, Telco, apps and much more. Going beyond the traditional classification of Big Data, which refers to the 3Vs (Volume, Velocity, Variety), today by Big Data Analytics we mean all the techniques and technologies that allow to collect and manage Big Data, integrate them with traditional data, process them and transform them into "Small Data": understandable information useful to improve a decision-making process. To this end, the key aspects that must be present in a winning Big Data Analytics solution are:

- Ingestion & Storage
- Compute & Querying
- Sharing & Governance

While much attention has historically been paid to Ingestion and Storage processes, many Big Data Analytics platforms still need to meet expectations because they lack computation and querying performance. Instead, the selection of the best technologies based on use cases is often more important than the storage itself: only through processing can Big Data be transformed into information, drawing value from it. No less, proper management of Data Catalog, Lineage, and Governance are critical to the usability of data.

1.5 Big data analytics and the data-driven approach: tools to serve business

At the technological level, the last decade has seen the emergence of many new technologies explicitly designed to better manage Big Data. This technological revolution began with the advent of Hadoop, which represented a fundamental paradigm shift: in fact, all the most modern technologies that to date fall into the "Big Data" category share the basic principles of the Hadoop framework, namely:

- Horizontal scalability: the ability to dynamically scale a service by increasing or decreasing the number of nodes. This enabled overcoming the hardware limitations typical of vertical scalability (increasing server resources).
- Fault tolerance by design.
- Distributed storage (from HDFS to the most modern Object Storage).
- High degree of parallelism.

In recent years, amplifying the benefits-even economic benefits-of horizontal scalability, cloud computing has taken over disruptively.

While cloud providers initially offered mainly IaaS (Infrastructure as a Service, i.e., the ability to rent virtual machines hosted in their data centers) services, the offering has increasingly evolved toward PaaS (Platform as a Service) solutions. These provide a range of services-from Ingestion to Storage to Processing to Queries to Reporting-managed with unprecedented dynamism in scaling: in seconds each service can be scaled horizontally as needed, so that resource utilization is made efficient by paying only for what is used. Particularly when it comes to computing resources, "smart" use of PaaS cloud services can enable costs to be reduced by orders of magnitude.

As an alternative or alongside cloud providers' PaaS services, all major software vendor solutions specializing in data are evolving towards Container, or application virtualization. Thanks to this type of mechanism, applications can be ported to different environments, facilitating their management and guaranteeing service levels typical of distributed systems (fault tolerance, horizontal scalability, etc.).

It should be noted that the multiplicity and heterogeneity of technologies normally involved in a Big Data solution has created the need for new skills within teams, particularly among developers a "DevOps" figure is increasingly common, who gives awareness to the development team about the infrastructural aspects to be considered when designing a solution, and how best to manage them.

1.6 The steps in a Big Data Analytics project

A modern Big Data platform must be designed from a Data Fabric perspective: it is no longer just a decision-support system to enable users to do analytics, but it must also be able to interact with downstream and upstream systems and applications (whether web apps, mobile apps, external systems) typically via APIs, message queues, ESBs. Therefore, providing components within the architecture that deal with Application Integration becomes increasingly important.

While not all components are always necessary, here are which ones should be provided within a Data Platform:

- Data Storage. Depending on the nature of the data to be considered but especially depending on the processing and query requirements determined by the use-cases, the Data Storage area may contain a Data Lake Storage and/or a Data Warehouse layer. Over the years, we have witnessed the debate regarding whether the Data Warehouse should be replaced or flanked by a Data Lake. One of the most important choices to be made when designing a Data Platform relates to one of the following two methodological solutions:

- Data Lake + Data Warehouse: two dedicated tools are adopted, each with unique characteristics (e.g., the Data Lake enables very large historical depths at the highest level of detail at low cost, while a Data Warehouse layer enables greater dynamism and lower latency in queries). It is the recommended solution when dealing with a traditional Data Lake Storage.
- Data Lakehouse: leveraging the features of the most innovative Data Lake Engines, it is possible to combine the benefits of the Data Lake and those of the Data Warehouse in a single solution, keeping the data on open format storage (Data Lake storage) and using dedicated query acceleration and processing tools. Some examples of these solutions are Databricks, Snowflake, and Starburst. Cloud providers are proposing increasingly integrated solutions in this direction, such as Azure Synapse, and Amazon Redshift.
- Regardless of choice, the area of Data Storage will still be crowded since, from a Data Fabric perspective, in addition to having to contain legacy systems, it may be necessary to provide dedicated systems to support the more operational component

of integration with downstream systems, or linked to specific needs such as Graph Database or Time Series Database.

- Data Access. Given the multiplicity of source systems, the Data Access area is tasked with decoupling physical storage from the end user, be it a user, algorithm, or application. One of the most widely used ways to accomplish this layer is to use a Data Virtualization solution such as Denodo, Tibco DV, or Dremio.
- Data Management. Having to satisfy a multiplicity of use cases, no longer just analytical but increasingly related to Machine Learning/AI models, often in real-time, the Data Management area includes one or more technologies capable of implementing different data delivery styles. Delivery styles include the more traditional ETL/ELT tools, often joined by real-time streaming ingestion and processing systems, technologies that operate by replicating data in Change Data Capture, and sources accessed via Data Virtualization when it is preferable to avoid or limit data movement.

The concept of "upload flow" related to the Data Warehouse world has evolved and we now speak of "data pipelines," which can be batch or real-time, orchestrated through graphical tools or, depending on the skills of the work team, managed through code.

- *Business Information Consumption***.** The set of front-end tools, including static reporting, self-service reporting, dashboarding, interactive systems with natural language recognition and AI-based suggestions, etc.
- *Metadata Management*. As mentioned earlier, one of the aspects that should not be underestimated when designing a Big Data Analytics solution is Metadata Governance. This is where Data Catalog, Business Glossary, rule-based access management, Data Lineage, etc. tools are available.
- *Application Integration*. Technologies that enable integration with upstream and downstream applications. This category includes API management systems, Web App, IoT platform, No code/Low code development platform, etc.
- *Data Science Lab*. With a view to optimizing Time-to-Market in the development of use cases that rely on a Big Data Platform, a primary role is played by the Data Science Lab. This contains all the useful technologies to enable of Data Scientist to perform rapid prototyping, experimentation, and development of AI algorithms and models.

Finally, it is important to clearly define the tools to be used in the different case studies, as well as to share guidelines that, on the one hand, allow Data Scientists to work independently on models and their future evolutions and on the other hand create the conditions to manage effective industrialization. In this regard, well-designed model code versioning management methodologies and tools and CI/CD pipelines can make the difference between the successful industrialization of a model and the failure of a Data Science initiative due to technical difficulties or production release timelines that are not compatible with business needs.

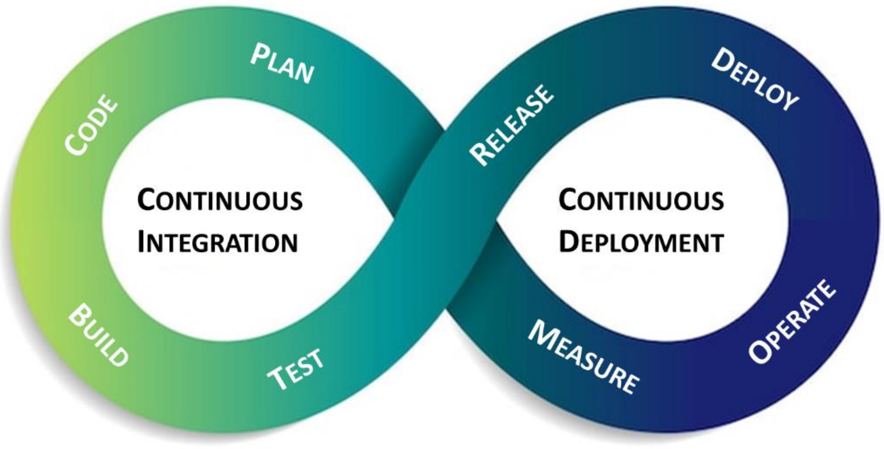


Figure 2: Continuous integration and continuous deployment CI/CD, or Continuous Integration and Continuous Deployment, refers to a methodology for the continuous deployment of products and/or services using automated processes and based on the concepts of integration, deployment, and distribution. Approaching the CI/CD methodology allows developers to overcome the normal code integration problems that typically occur in scenarios where development proceeds by parallel paths, and allows the result to be deployed automatically.

3. The importance of data avaiability

3.1 How Big Data is collected

"Big Data Analytics" refers, in a first approximation, to the collection, analysis and accumulation of massive amounts of data, which may include data of a personal nature, hypothetically even from different sources. The massive nature of processing operations brings with it the need to subject such sets of information (both stored and streaming) to automated processing, using algorithms and other advanced techniques, to identify correlations of a (mostly) probabilistic nature, trends and/or patterns.

Operationally, in the ICT sector, Big Data, as we said, is defined as a collection of data that cannot be acquired, managed and processed by "traditional" computer tools, software and hardware in a tolerable amount of time. However, there is no predefined size threshold for a data set to fall into the Big Data category.

From a descriptive point of view, it is common to find in the relevant literature, heavily influenced by the American experience, reference to some recurring characteristics with respect to the phenomenon under consideration. They are summarized in the 4 "Vs": the volume, regarding the enormous size of the data generated and collected; the variety, regarding the many types of data available, among which, in addition to traditional structured data, there are also semi-structured and unstructured data such as audio, video, web pages and text; the speed of processing operations; and the value that data take on when processed and analyzed, to allow the extraction of information that can contribute to the efficiency and quality of "traditional" production processes or intrinsically qualify the supply of goods and/or services, especially in terms of innovation and personalization.

Multiple other Vs suitable for characterizing Big Data have then been identified: among the most noteworthy are veracity, i.e., the quality and significance of the data collected or processed; valence, i.e., the degree to which the data are connected with other data; and visualization, i.e., the need to synthesize the most relevant data and the knowledge extracted from them in a visual and easily interpretable way.

The Big Data collection phase begins with the generation that takes place in the context of activities performed by users in a computerized that is, in the context of the so-called Internet of things. In the current context, in which virtually all media content is made available in digital format, and much of economic and social activity has migrated to the Internet, user activities, both online and offline, generate precisely large amounts of data.

First, online services, often populated by users' own content, are an excellent source for Big Data: think, for example, of e-mail, satellite navigation, and social networks, where users upload their own content (photos, videos, texts), sharing it publicly on digital platforms, apps, and websites. Added to this is the collection of data generated by the functionality of users' personal devices (such as smartphones, tablets, and personal computers.

Activities performed by users, even in the absence of direct interaction with an electronic device, generate (so-called offline) data and can provide relevant information about the behaviours and preferences of individuals. Consider, for example, the geolocation data of individuals provided by smartphones (where this feature appears to be enabled), even though the user does not actively interact with the device. Similarly, surveillance cameras, in capturing the presence and movements of individuals in a given area, capture data that can be processed to infer information about people's flows. Electronic payment instruments also capture information about the purchasing behaviour and preferences of the users who use them. In this regard, it is worth mentioning the initiative of Google, which has signed commercial agreements with some operators of payment card circuits to acquire information on the purchases made by consumers, which helps verify the effectiveness of personalized advertising campaigns, as well as for further profiling its users.

Another essential source of Big Data power is the Internet of Things (IoT), which sees applications both in the industrial sphere (e.g., in so-called predictive maintenance) and in the lives of individuals, from home automation to devices, often wearable devices (wearable devices) that record data about each individual (e.g., those related to sports activities and/or biological parameters).

The basic idea of IoT is to connect various real-world objects, such as sensors, actuators, RFID (Radio-Frequency Identification), barcode readers, cell phones, etc. - and make them cooperate with each other to accomplish a common task through the use of microprocessors in the objects.

It enables the development of applications in several key areas, think for example, in a Smart City, where citizens, through an application on their smartphones, have real-time access to data on traffic, available parking spaces, air quality, waiting times for public transportation, pharmacies open on duty, and the number of patients in emergency rooms. This is all thanks to interconnected sensors, which transmit their readings to a central server that processes and makes the information available to users.

Through sensors placed in mobile devices, means of transportation, public infrastructure (airports, ports, train stations), and household appliances, IoT enables the digital encoding, transmission, and storage of information related to the operation of interconnected equipment and devices, both in the business/work environment and the activities of individuals. In this way, various types of data (environmental, geographic, and logistical), which, in general, exhibit multiple characteristics typical of Big Data, including heterogeneity, variety, lack of structure, strong space/time relationship, and rapid growth.

Data generated by users or "things" in the IoT context are then acquired through the electronic devices involved in the act of generation-such as smartphones, sensors (motion, temperature, humidity), cameras, input devices (keyboard, mouse), scanners, RFID, wearables and other IoT-specific devices, etc. - thus resulting in the availability of the entities that develop and operate these systems (e.g., in the case of smartphones, operating system providers), which, however, to acquire the availability of personal data must necessarily seek prior consent from the user who generated it.

Smartphones, in particular, play a central role in the acquisition of user-generated data, as they have numerous input devices (such as motion sensors, light sensors, location, keyboard and touch screen) integrated into a single Internet-connected tool that accompanies the user in all of his or her daily activities. We point out that in a smartphone, the vehicles for data acquisition are represented, on the one hand, by the operating system and, on the other, by the applications preinstalled or subsequently purchased and installed by the user.

In the second case, data are acquired by the respective developers.

More generally, all users' online activities (such as sending and receiving e-mails, satellite navigation, or using social networking services) -regardless of the device used- generate a huge amount of data, typically personal data, that feed copious acquisition activity. In this regard, as pointed out in some of the hearings held as part of this thesys, consider the way online services operate contributes to the multiplication of the possibilities of acquisition of every piece of data generated through the use of personal electronic devices. It is also emphasized that, on the part of users, "consent" to the processing of personal data that their online activities generate personal data, which their online activities generate, is intended to allow, where validly given, the free use of the same.

Technically, the acquisition of user-generated data presupposes the use of systems dedicated to their tracking, which, with specific reference to web browsing, consist of so-called cookies; the latter are text files that collect preferences (e.g., language, interface, the place from which access takes place, etc.) and consumer information (e.g., pages visited, texts transmitted, etc.) active in a website, allowing precise profiling, which, moreover, is updated with each subsequent access to the same site.

In this regard, it should be noted that among application and/or website developers, the practice of outsourcing tracking systems developed by major ICT players (such as Apple, Google, and Facebook), with the consequence that data acquired by the former are also within the availability of the latter, which, moreover, as developers of extremely popular operating systems and/or apps, are already in a privileged position to directly acquire user data from smartphones and/or related apps.

In the IoT context, environmental, geographic or logistical data are acquired from devices installed in homes or industrial production sites, business premises and the environment.

The management of IoT devices also often relies on standard solutions prepared by large ICT players (such as, for example, Amazon or Google), which are usually integrated with capabilities for processing the acquired data, guaranteeing them the availability of the data. So-called data brokers can also intervene in the data acquisition process, i.e., entities that aggregate data from different sources (mainly Internet sites) and organize them to make them available to third parties. Such brokers, operating simultaneously on multiple sites, realize essential economies of scale and scope (due to the variety of data collected on different sites) and enable them to increase the breadth and depth of data collection. Data brokers fuel an opaque market, especially for end users, who need to be put in a position to know the path of data captured by the websites and/or online platforms they access.

Finally, some data can be acquired without having to interface with users or otherwise with the entities that generate them. These are the so-called open data, generally produced by public entities and, by definition, freely accessible to all.

There needs to be more than the definition of Big Data to offer a complete picture of the phenomenon. Talking about Big Data is more than just talking about large data sets. the transformation taking place is more profound than that. The data collection and management process is changing, the technologies supporting the data lifecycle are evolving, and new skills for data exploitation are developing (the figure of the Data Scientist takes a central role in this context). In the following sections, we will try to answer the following questions.

3.2 The data Catalog

The infinite amount of data sources requires today's organizations to equip themselves with tools to gather and provide contextual information about the data in their various business systems, making it available with a simple search. This tool goes by "Data Catalog" or a "search engine" of business data. It proves to be the trump card for analysis of strategic Business areas and an ally for IT in data management challenges.

Let's start with a decidedly current business need, especially for B2C companies: to have an ongoing and expanded understanding of the customer and how he or she interfaces with the company. Having a broad view of the customer, the so-called Customer 360, requires the active participation of many functions, including Marketing, Sales, Customer Service and Finance.

Marketing, for example, needs data on customer-company interaction paths, considering all channels, customers' purchase history, past promotions offered, services they have used, and information on their habits and preferences. Only with such a view is it possible to propose personalized campaigns, avoid making communication mistakes and provide the necessary support to build more lasting relationships.

All this data is often produced and processed by different people and functions, and no one has a complete view of the end-to-end customer journeys and related data. When this data is not integrated into corporate data lakes or data warehouses because it is not deemed useful for analytical purposes and reports used by management, so-called data silos are generated: operational teams create and maintain their own databases and reports that are useful for their daily activities, without IT's knowledge.

When data are distributed in different systems and only partially available to some business areas, the solution to consider is precisely the adoption of a *Data Catalog*, supporting both Data Management activities and Business users.

It is a tool that collects and provides information related to data in different business systems, making it available with a simple search as if it were a "Google" of business data. It could be the inventory of data that the organization has, which contains valuable information to provide context to the data and allows IT, data scientists, and business analysts to find and better understand the data.

The Data Catalogue essentially helps collect, manage, and make available "data about data," technically called "metadata."

Modern Data Catalogs are "active" metadata management tools: they use Machine Learning algorithms to discover and collect metadata in an automated manner, enabling them to overcome the critical issues associated with not having complete knowledge of all the data and information systems an organization has.

The metadata collected in the catalogue can be of various kinds: technical, operational, business:

Technical metadata: catalogue and describe information systems, schemas, data types, some technical configurations of systems, physical data models, systems used for analytical purposes, applications and tools with which data are shared and/or analyzed.

Operational metadata: they describe the so-called data lineage, the path data takes between systems. They also include information about the performance of connections to systems and the transformations that data undergoes when it is moved or copied from one system to another.

Business metadata: refers mainly to the Glossary, ontologies, and semantic meaning of data. Also included in this metadata category is the assignment of business ownership and stewardship, i.e., the people responsible for the reliability of the data, who certify its quality and ensure that it is available on time for the business.

IT and business users identified as owners and stewards are left with the fundamental task of enriching and validating metadata to make the catalogue a valid tool available to all data consumers. Metadata enrichment can be done by describing and providing examples of use cases and defining service levels, criteria, and metrics to monitor data quality.

With the Data Catalog, all data consumers can search easily, quickly find out what data they have, understand what it looks like and what format it has, what it means to the business, where it resides, whether and how they can access it, whether it is suitable for doing reliable analysis (e.g., because it has been certified), and much more.

3.3 Is data enough to extract information?

When considered in isolation, data have little value, but they gain value when organised. For this reason, the processing phase, which involves organising unstructured raw data into actionable information for economic purposes, plays a central role in the entire Big Data chain. The analysis activity allows rapidly extracting knowledge from large masses of unstructured data to obtain information in a compact and easily interpretable format.

After an initial extraction phase-during which data are retrieved from the various available sources, selected and loaded into the memory of the processing system and subsequent integration of all the information that relates to the same elements or application domains, the actual analysis of the data intervenes, which takes place through analysis techniques and tools capable of bringing out of the raw unstructured data information amenable to interpretation and practical use.

Generally, analysis techniques mainly consist of algorithms, among which we distinguish between querying and learning algorithms. In comparison, the former aim to answer precise user queries posed in the form of queries, and the latter aim to extract new knowledge and theses and use advanced artificial intelligence techniques.

Implementing algorithms requires computational computer models involving hardware and software resources distributed in remote data centres. The intelligence of analytical techniques, together with the voluminousness and variety of data, is bringing out an essential innovation in the knowledge extraction process.

In the so-called data-driven analytical paradigm, data contribute not only to test theoretical hypotheses with statistical techniques but also to exploring new scenarios and deriving new theories, as well as, more generally, to discovering new knowledge through artificial intelligence algorithms. It is a wholly methodologically innovative approach to information acquisition and knowledge generation. It recognizes data as the driver and algorithms as the task of finding patterns that traditional methodology would perhaps only need help to identify (except for subjecting them to subsequent verification).

Within this new analytical paradigm, data are of central importance. Artificial intelligence programs learn through the availability of a vast number of examples. Thus, the data, as a source of information about the phenomenon to be studied, represents the very origin of the evolution of algorithms, so it is the availability of new data sources that enables the improvement of the algorithms employed and/or the development of new algorithms.

On the other hand, even when algorithms do not change over time, knowledge progress depends on data. For example, in several areas (think of online translations), improvements in recent years can be attributed not so much to algorithms, which essentially have not changed from the past, but to the availability of immense amounts of data, as well as to the somewhat more powerful computational capacity.

Finally, it should be considered that some activities typical of the digital environment make economic sense only if they are based on a large amount and variety of data (e.g., recommendation systems on the so-called long tail of online sales platforms).

The value of data is inversely proportional to its generality. The advantage in developing intelligent solutions to a particular user's problems derives precisely from the analysis of internally produced data: the most relevant datasets for an enterprise are those that the enterprise creates for itself because it knows the context in which they were created and the purposes for which they were created; the main innovations may come precisely from datasets built by an enterprise for internal use, not intended from the outset for third parties or the market.

Furthermore, the accuracy of algorithms increases with the diversity of data sources so that a data source weakly related to a phenomenon may have a more significant impact in improving the algorithm than a more precise and refined source closely related to the same phenomenon. It, therefore, becomes relevant for organizations to acquire and analyses so-called "*informative data*".

The use of data has become crucial in today's businesses. Data are used in businesses for a variety of purposes:

- Assess product suitability
- Measuring the efficiency/effectiveness of a process and being able to improve it
- Telling a story to communicate corporate values
- Identifying business models
- Identify new business opportunities
- Gather customer feedback
- Have an overview

The digital transformation of the past few years has increasingly noticeable effects on companies worldwide and thus, on consumers, profoundly changing the way we do business. This transformation is reflected not only in the evolution of the devices we use but especially in the application of data in our daily lives. The data we produce never stops, which is why companies must keep up with learning how to translate this information into profit.

But the word data does not mean information. Information results from data processing. Therefore, data is a known element, raw or elementary information and is usually made up of symbols that need to be processed and contextualised. Information, on the other hand, is an element resulting from processing more data that allows one to come to know something. Just having data is not enough.

3.4 Data Catalog: why is important?

Gartner reports that "each year, poor data quality costs organizations an average of $12.9 million." [Annual Gartner Report 2021]. This striking figure clearly shows what it means not to use quality data in business decisions. Further confirmation comes again from Gartner: "Aside from the immediate impact on revenue, poor quality data increases the complexity of data ecosystems and leads to poor decisions. In contrast, quality data provide better leads and improve customer understanding and relationships. Quality data is a competitive advantage." [Gartner Data & Analytics roadmap 2022]

But what does data quality mean?

It is a concept that goes beyond data quality per se and concerns the entire set of processes designed to monitor and increase, through analysis and improvement of all enterprise data, the characteristics that affect the ability of the data to meet stated or implied needs for the business. As Gartner further argues, "D&A (Data & Analytics) leaders must take pragmatic and targeted actions to improve the quality of enterprise data if they are to accelerate the digital transformation of their organizations." Scattered, redundant, or incorrect data creates clutter and inefficiencies within companies because it does not allow them to make the right decisions or act in a suitable timeframe. And this reduces opportunities for possible revenue. The term Data Quality is generically used to describe a process of analysis to which subjecting data, with the aim of analyzing and increasing its quality. Depending on the nature of the data and the purpose for which it is being analyzed, the term "quality" is often declined into the concepts of accuracy, consistency, completeness, and correctness (i.e., dimensions of quality) that define its general properties.

Analysis, and the subsequent process aimed at data quality, have, over time, become preparatory activities for decision support. The information content that data can express is functional to their ability to describe the environment from which they were drawn or observed. It is well known that using low-quality data can cause incorrect or inefficient decisions, damaging the economy of the company or government and all parties that base their decision-making processes. The problem of data quality plays a decisive role in many contexts: scientific, social, economic, political, etc. In this regard, we recall a striking case that occurred in space exploration: the explosion of the Space Shuttle Challenger was blamed on ten different categories of data quality problems, as described apodictically in the work of Fisher and Kingma (2001).

The relevant scientific communities have developed various techniques to analyze and improve data quality, declining well in the various dimensions that define it.

Currently, a challenge in data quality is to develop methodologies and tools that can analyses and improve data quality in a way that maximizes its informativeness. The development of methodologies is critical. scalable, i.e., capable of efficiently handling large amounts of data (e.g., Big Data), often found in many business and PA contexts.

Data Quality generically identifies activities and processes to analyses (and possibly improve) data quality in a database. However, data quality can be observed by emphasizing certain aspects that might be more relevant for the domain expert than others. To this end, data quality dimensions are proposed as a (qualitative) tool for evaluating data quality. It is important to note that these dimensions, some of which will be introduced below, can be defined at the schema, process or data level.

In the first case, the logical structure used to represent the data is analyzed to verify that it is adequate and suitable for obtaining data with the required qualitative characteristics.

Data accuracy is defined as the distance between a value v and a value v' considered the correct representation of the actual phenomenon that v is intended to express.

One can evaluate the accuracy of v by focusing on the following:

- Syntactic accuracy. It is checked that the value of the attribute v is present in the set of values in the domain.

- Semantic accuracy. In this case, one evaluates the accuracy of the v value by comparing it with its real counterpart v'.

Completeness is "the level of breadth, depth and appropriateness of a piece of data according to its purpose." (Wang and Strong 1996).

Consistency is defined in the literature as the "violation of one or more semantic rules defined on a data set" (Batini and Scannapieco 2006). Again, several levels of consistency can be identified:

- *Key consistency:* this is the simplest of the forms of consistency and requires that, within a relation schema (a table), no two tuples have the same value of an attribute used as a key.
- *Inclusion consistency:* a classic example of this is the "foreign key" of a relationship. It requires that a set of columns from one relation schema be contained in another set of columns from the same relation schema or another instance of a relation schema
- *Functional dependencies:* these are the best-known and most widely used. In general, given a relation R with attributes X and Y, Y is said to be functionally dependent on X (X->Y) if and only if each value of X is associated with a value defined in Y.

3.5 Data Governance: What we mean by it?

Data Governance is a set of practices and processes that help ensure the effective management of data resources within an organization. In the now vast set of concepts involving the world of data, it is possible to find a variety of definitions and declinations of Data Governance. Therefore, rather than reporting another formal one, it is preferable to approach the topic starting with the benefits it brings and the fundamental pillars that constitute it.

An organization’s data has become an asset only when it can be used across the board, going beyond the limited and partial use by individual areas, functions, and business units. A primary goal of data governance is to "break down the silos" of organizational data, which commonly form when individual business areas implement and use systems and applications that do not communicate with a data architecture designed to ensure integration, centralized management, and cross-distribution. Data Governance aims to harmonize the data lifecycle through defined and shared processes involving the participation of stakeholders from the various business units.

Another key objective of Data Governance is to ensure that data are handled and used correctly to enable decisions based on reliable data and to prevent misuse of personal and sensitive data for the organization. This goal is achieved if data use policies are defined and mechanisms are implemented to monitor data throughout its lifecycle, viz:

- when they are generated or acquired.
- in the internal transitions that the data undergo.
- when it is shared outside the organization.
- at times when the data is archived or deleted.

Analysts, data scientists, and managers who make decisions should be as autonomous as possible in using data for their work: they must be able to easily find what data is available in the company, know how to access it, understand its meaning and content, understand where the data comes from, and identify its level of reliability. Data users can also take on the " producer " role of enriched and reprocessed data sets, which could become part of the organization’s assets to be reused or shared with colleagues, even from different areas or functions. This scenario can become a reality when a modern data platform is complemented by processes and technologies that allow data assets (db tables, views, files, reports, etc.) to be catalogued, enriching them with metadata and key information for end users.

As can be easily guessed, to make better decisions or make processes more efficient, it is necessary to have data that can be trusted and that allow interpreting the current reality, data that are of high quality and up-to-date: this is precisely why data quality activities find an extremely relevant space.

It is clear from this initial overview that Data Governance is not only about the adoption of technological tools, but is a broad ecosystem that combines the review of processes, the definition of procedures and standards (formats, naming conventions, etc.), the identification of methodologies and best practices, training activities and facilitation of change and collaboration.

3.6 The problem of the lack of Data Governance in enterprise environment

It is well known that companies with good management, knowledge, and ability to use their data are much more likely to grow than companies that continue to operate "in silos," with non-integrated data and a lack of coordination between functions and people.

Data Governance fosters the mapping and management of data to ensure cross-domain visibility and widespread knowledge, enabling the sharing of insights that emerge from the data. It fosters the creation of a data-driven culture, adopting best practices in data handling and use, which over time leads to better decisions and benefits that become increasingly tangible.

In addition to the strategic benefits, Data Governance involves the implementation of control mechanisms to ensure compliance with business regulations and regulations imposed by legislative bodies at the national or supranational level, such as the General Data Protection Regulation (GDPR).

Data governance cannot be traced back to its technological and information technology dimensions, typical of the IT sector. To be genuinely representative, data governance must be understood as a series of multidisciplinary activities framed in a business context that cannot disregard aspects such as purpose, timing and, above all, budget to be devoted to the various activities envisaged by the organization.

Since this is a very complex discipline at the theoretical level, making it as simple as possible at the operational level is essential. To guide the proper definition and application of data governance, a framework capable of structuring and controlling it consistently must be prepared.

In several areas, such as finance, the data governance framework must be properly prepared by law, as prescribed by many provisions, starting with Law 231/2001. Thus, corporate data governance is more than just a simple though useful best practice, but an activity fully recognized by the national regulatory framework for data processing and management.

A data governance framework contemplates and makes synergistic contributions from organizational, technological and operational aspects. To cite one possible example, let us again resort to the provisions of the Data Management Association (DAMA), which envisages a veritable wheel with nine knowledge areas converging radially toward the center, consisting of data governance from the synergy of all competing aspects.

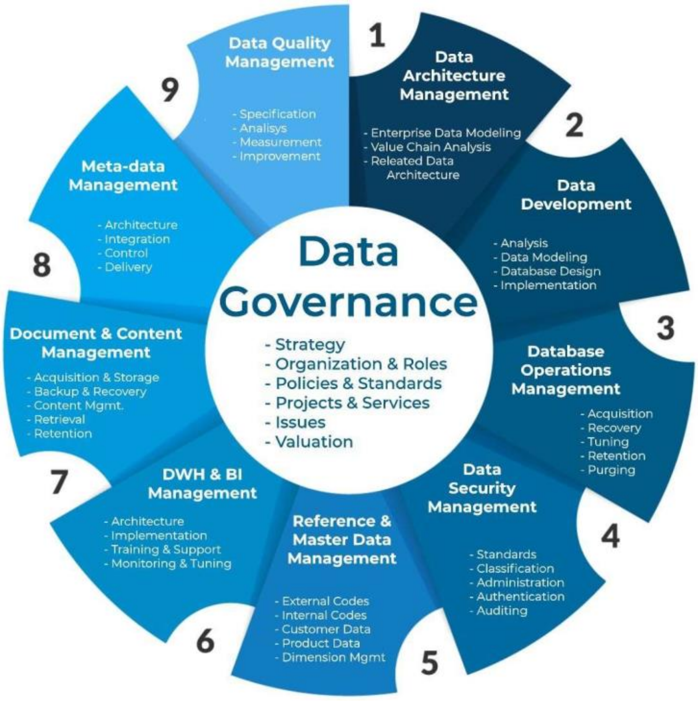


Figure 3 Source: DAMA (Data Management Association): Data Governance Framework

The nine areas that DAMA prepares for the establishment of a data governance framework are as follows:

1. *Data Architecture Management:* corresponding to the overall structure of the data and resources that constitute the enterprise architecture.
2. *Data Development (data modelling and design):* relating to all aspects of analysis, design, construction, testing and maintenance of operational phases.
3. *Database Operations Management (data storage and operations):* regarding the distribution of structured data asset storage, such as in various database systems.

4. *Data Security Management (data security):* with respect to expected compliance to ensure privacy and authentication according to current regulations.
5. *Reference & Master Data Management:* encompasses all functional operations to acquire, extract, transform, move, deliver, replicate, federate and virtualise structured data.
6. *Datawarehouse & Business Intelligence Management:* to ensure the management of descriptive data analysis for reporting activities useful to support decision-making processes.
7. *Document & Content Management:* includes all functional operations to store, protect, index and access unstructured data to ensure integration and interoperability with structured data.
8. *Metadata Management:* use of standards to reduce redundancy and ensure better data quality and use of Data Catalogue for the collection, categorisation, maintenance, integration, control, management and distribution of metadata.
9. *Data Quality Management:* consists of the definition, monitoring and maintenance of data integrity, as well as the progressive improvement of data quality.

3.7 Data Quality in the Data Governance Process

Data Quality is a dimension of Data Management frequently seen as a priority by companies and organizations that want to invest in governance for data improvement.

Data Governance drives formalizing data quality rules, quality measurement metrics, minimum control thresholds, and anomaly resolution processes. Effective Data Governance must harmonize data from different sources, eliminate inconsistencies, and bring out errors or noncompliance situations as early as possible to prevent them from negatively impacting the accuracy of analysis, all the more so when data are used to produce outputs that are shared with stakeholders outside the organization or to demonstrate compliance with regulations.

Data Quality Management activities that are "grafted" into a Data Governance framework can be grouped into four macro-activities that constitute a continuous cycle of analysis and refinement.

- *Discover:* Data Discovery and Data Profiling to examine the structure of datasets, quickly understand their content, and identify possible outliers.
- *Define:* definition of both technical and business requirements and rules valid for subsequent monitoring and identification of problems. These can be data format,

content, and availability requirements. Typically, an attempt is made to define requirements and rules that allow monitoring data quality along a few main dimensions: completeness, accuracy, timeliness, consistency, uniqueness, integrity, and compliance.

- *DQ rules application:* application of data quality rules to all relevant datasets across the enterprise, implementing defined rules and applying them systematically to datasets via computer systems.
- *Monitor:* analysis of data quality activities and results obtained to track data quality improvement over time. The purpose of monitoring metrics in a Data Quality system is to measure the quality of the data and, consequently the processes that use it.

3.8 Data Quality Management

It is appropriate to operate a Data Quality measurement to be effective by constructing a consistent set of rules, thresholds, and synthetic indicators functional for the qualitative enhancement of results.

It is, therefore, fundamental to define for each data domain which dimensions to measure data quality and with which rules. Typically, three classes of indicators can be identified:

- Availability
- Format
- Semantics

These indicator classes can group various Key Quality Indicators (KQIs) that succinctly express the level of data quality for a certain domain based on the results obtained from the implemented Data Quality rules. In addition to the more "technical" measurements belonging to the availability and format classes, modern Data Quality frameworks are increasingly pushing compliance with semantic conditions, i.e., the ability to verify whether a data item correctly represents a business phenomenon based on the interpretation of the data with respect to a specific context.

In this area, for example, profiling algorithms are being used to identify and classify the level of "sensitivity" of stored data for the suggestion of threshold values to be taken into account to define "good quality" data or processes by employing AI systems for a recommendation.

4. Data strategy to maximize the value of data.

4.1 How to implement a Data Strategy

The ability of companies to manage and extract value from data is, as we have seen, a key determinant of their success.

However, to date, the level of maturity on Data & Analytics issues is quite varied, and many companies still need to equip and structure themselves adequately. In these organizations, less than half of corporate data is used for decision-making; the majority of employees have access to data they should not see; analysts use a tiny fraction of their time on analytics; security breaches are increasingly common; and datasets proliferate in silos, making it more challenging to integrate them. Obviously, in this respect, larger companies, with consequently larger budgets, have a decidedly more significant advantage.

To reverse course, it is not enough to have specific figures for data management and data analytics, we need to adopt a strategic approach that aims to treat information as real assets. If a company wants to plan a compelling data journey and become data-driven, its first need is to determine what its level of maturity concerns Data & Analytics issues.

Data can be regarded as a "new commodity," with characteristics that are as peculiar as they are advantageous (they are an almost inexhaustible, reusable commodity with very low storage and transportation costs), which needs new rules for their treatment and use.

Therefore Infonomics (or information economics) was born: the set of theories and tools helpful in assigning economic meaning to data. It provides companies with a conceptual framework for valuing information and managing it as tangible assets.

Douglas 2017 declines this conceptual framework into three basic aspects: *monetization, management, and measurement*.

- *Monetization:* Concerns the possibility of generating, from data, measurable economic benefits. This can occur through direct monetization processes (such as the sale of raw data) or indirectly (e.g., when, as a result of data analysis, the efficiency of business processes is increased). For another example, recently, Inwit, the leading Italian tower company that operates about 23,000 towers throughout the country, has equipped all sites with sensors capable of collecting weather and air quality information. The tower-company goal is to collect and monetize data by selling it to ARPAs (environmental protection agencies). This is an opportunity that can only be realized if, first, a data strategy has been defined in the company, an inventory of all information assets has been taken, the feasibility of monetization projects has been verified, and the target market has been identified.

- *Management***:** Since there are yet to be commonly accepted principles and practices for managing data as tangible assets, it is possible to borrow the standard rules applied in managing other business assets. Or put into practice some models that have attempted to define principles for information asset management. For example, the models proposed by John Ladley [(Ladley 2012). *Data Governance. How to design, deploy, and sustain an effective data governance program*. MK] stipulate that data should be managed and accounted for like any other tangible or financial asset; that it should be a fundamental element of all processes in place; that the risks (known and potential) associated with its use should be assessed, both in terms of cost and responsibility (responsibility assumed by specific figures within the company); and that standards should be set for data quality, checked periodically.
- *Measuring:* As with management, models already in place for other assets can be applied to measure the actual and potential value generated by data-which are not yet part of recognized balance sheet activities, necessarily taking into account that the value of data can have both financial and "functional" connotations. Valuation models capable of capturing both facets are therefore needed so that questions about both the market value of data and its net contribution to profits and its ability to support decisions or influence key business drivers can be answered.

Data Strategy is the fundamental tool for successfully embarking on a Digital Transformation journey. Indeed, every transformational journey introduces new goals, technologies, and behaviors that the entire organization must know and share internally.

That is precisely what Data Strategy is for: clearly defining and sharing how the organization can leverage data to concretely implement the business strategy. By raising awareness of priorities, potential benefits, technologies, and methodologies needed, it enables you to:

- Identify and close any technology gaps.
- Identify and resolve critical organizational issues.
- Have a clear view of the maturity status of business processes.
- Assess business impacts and expected benefits.

To achieve these goals, it is essential to build a "roadmap" that, linked to the company's business strategy, communicates the strategies to be pursued and their impacts to the entire company. The Data Strategy Roadmap allows for simulations regarding costs, timing, goal achievement, and numerous other variables. In this way, it provides a comprehensive view of the activities to be introduced in the company and identifies the best strategies to respond resiliently to market shocks.

4.2 Big Data: the role of data strategy

To build an effective Data Strategy, surveying and assessing the necessary information is a must: only a complete understanding of the company's information assets makes it possible to build Big Data Analytics and future plans.

Counting on a huge amount of data is not enough; one must have a thorough, complete and reliable view of it. This is why the topic of Data Governance, defined by Gartner's glossary as: "mapping decision rights and the subsequent creation of an accountability framework to ensure the adoption of appropriate behaviours in the evaluation, production, consumption and control of data and related analytics practices, is of central importance." [Garter Data & Analytcs Roadmap Report 2022]

In this sense, Data Governance is to be understood in an extended and not merely technical sense.

As the volumes, variety, speed of generation and veracity of data increase, it is critical to know how to manage the entire flow and to understand how infrastructures should be built and maintained to manage them and extract their value.

Only through a path of Data Governance can this be done, ensuring that information is properly managed while simultaneously being able to define its methods, technologies, and behaviors inside and outside the organization.

Not mere data management, then, but the introduction of valid communication standards and clear guidelines. Both consider the full range of organic and extra-organic dependencies to create cross-functional mechanisms for integrating systems, people and processes within the Data Strategy journey. Data is the fil-rouge of these pillars: it not only permeates them but also enables them to evolve and generate value.

4.3 Impact and benefits on marketing, sales, and customer experience

The sudden change in buying and consumption habits that has occurred in recent years has prompted many companies to implement major strategic revisions to their marketing and sales plans. Consumers are more aware and demanding, and their opportunities for interaction and relationship, physical and digital, with companies have multiplied.

Therefore, Customer Experience has never been more strategic and differentiating, as well as crucial to the success of any business strategy.

Creating a direct link between Business Strategy and the roadmap of initiatives is a critical success factor in ensuring that the rest of the company takes up the levers operated by CEOs and Chief Marketing Officers: initiating, on the one hand, the redefinition of projects towards new goals, and on the other hand offering the ability to analyze their response times and values generated. Both are critical elements in Customer Centricity strategies, which must rely on a deep understanding of their customers, their needs and the tools to make them "tangible" and achievable.

While traditional Marketing has previously focused at length on measuring the performance of campaigns and activities, today, thanks to a more strategic view of data and technology, it can build them from the ground up with the assurance of a more significant business impact.

However, the data must be quality, accurate and updated with the necessary latency, up to real-time where necessary. Only in this way can they be transformed into meaningful insights to redefine campaigns and activities and develop personalized content across all touchpoints and communication channels.

4.4 The stages of a Data Strategy: from Assessment to the creation of a Roadmap

Awareness of the importance of capitalizing on one's current and future data assets is a prerequisite for successfully initiating and embarking on an evolutionary but sustainable strategic path. Indeed, along with business priorities, the true maturity of the business ecosystem must always be considered.

There are three main steps in implementing a Data Strategy and they are strongly interconnected.

The first, which is particularly effective in leveraging the "as-is" condition, is based on a technology-architectural assessment and is used to identify and assess state of the art in data and analytics. This step is dedicated to the study and analysis of communication channels, transformation processes and technologies used and serves to get a clear picture of the starting situation and identify gaps and points of attention with respect to business strategies. The second step, depending on the strategic business objectives, focuses on the company's evolving needs. Through analysis and interviews conducted in all key business areas, the desired use cases and any inefficiencies or pain points are identified, assigning to them complexity ratings of implementing their resolution and expected benefit ratings.

The last step involves the definition of a robust and modern technology architecture that, through the use of methodological best practices, is able to: enable the necessary new functionalities, respond to the needs gathered and, at the same time, preserve the value of the investments made through the implementation of a Roadmap.

As noted earlier, the Data Strategy roadmap is critical for any corporate "plan-ahead team" to redefine strategies and prepare for opportunities. Such an approach allows executives and managers to present the status of various initiatives and goals clearly and comprehensively to other directors, contemplating simulations and alternative scenarios. Unity of purpose and clarity of vision are critical factors in the success of any business strategy.

4.5 Implementation Strategies

Carving out a path based on the exact and specific needs of each company, entity, or organization is the foundation of any Data Strategy path. To the needs tracked, three key milestones must then be integrated:

- Alignment between business plan and business initiatives.
- An integrated strategic plan that contemplates project prioritization, mapping of information needs, evidence of economies of scale, roadmap management and other possible variables.
- Branches on possible architectural, methodological and organizational evolutions.

In today's complex socioeconomic environment, it is imperative to homogenize and simplify architectures. Carefully observing new trends with the goal of reducing costs and complexity: a critical success factor for both business leadership and in the strictly IT sphere.

Starting a Data Strategy journey also means involving the entire business ecosystem. For this, change management activities such as:

- Training on the use of new methodologies and technologies.
- Self-assessment systems.
- Skills mapping.

The goal of these and any additional Change Management actions is to direct pervasive and conscious change for the benefit of the entire organization. Supporting people in change and accompanying them at every step in organizational and procedural transformations, is a key factor in consolidating corporate welfare and, consequently, generating business growth.

Data Strategy represents a decisive opportunity for every industry to create value, react more quickly to events, support a broader set of decisions, or automate critical processes. Numerous studies identify 10% to 20% higher profits with basic analytics and over 30% with advanced analytics. Not to mention reputational benefits such as the ability to raise capital more efficiently and positive jumps in the stock market for publicly traded companies: companies that do Data Strategy do digital transformation, and that is what investors today expect when they look at a modern company. [Visualcapitalist report 2022].

In the next few pages, we will look at three practical examples of well-implemented data strategies that have produced value for the organisations that implemented them.

4.6 Data-driven organization maturity model

Becoming data-driven means embracing an evidence-based culture in which data can be trusted, and analysis is highly relevant, informative, and used to determine subsequent business actions. A data-driven maturity model can be based on how an organization leverages analytical tools to get evidence from data. In the chart below, we have tried to combine aspects that indicate what stage of maturity a company is in relation to its use of data.

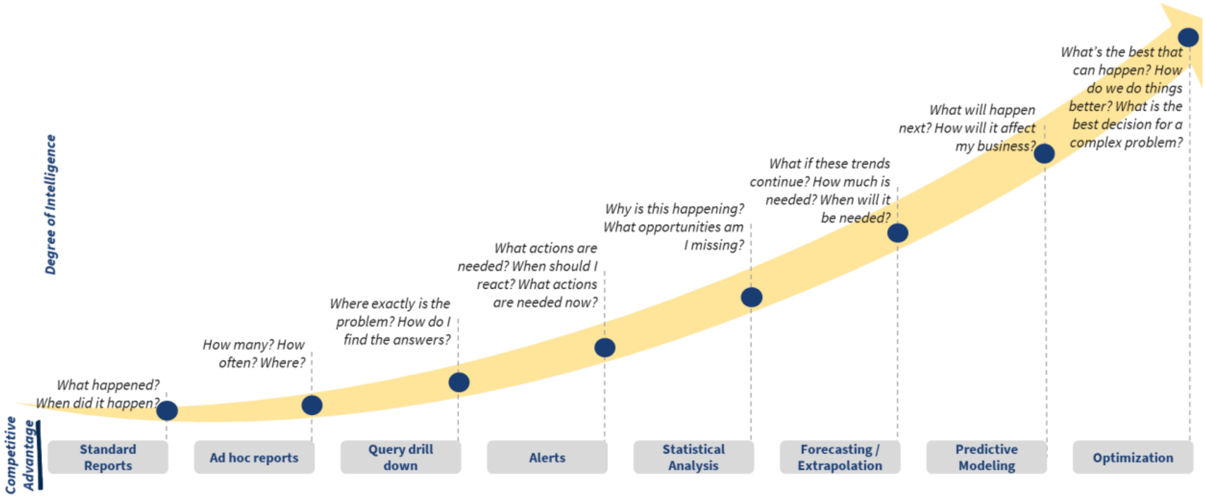

Degree of Intelligence
Competitive Advantage
What happened? When did it happen?
How many? How often? Where?
Where exactly is the problem? How do I find the answers?
What actions are needed? When should I react? What actions are needed now?
Why is this happening? What opportunities am I missing?
What if these trends continue? How much is needed? When will it be needed?
What will happen next? How will it affect my business?
What's the best that can happen? How do we do things better? What is the best decision for a complex problem?
Standard Reports
Ad hoc reports
Query drill down
Alerts
Statistical Analysis
Forecasting / Extrapolation
Predictive Modeling
Optimization

Figure 4: Data driven maturity model. Designed by author,

5. A statistical decision support system for tourism data analytics

This chapter is based on a paper presented at the Aciek 2021 conference. It discusses how the use of big social data – defined as: the vast amount of data generated by users on social media platforms, online communities, forums, blogs, and other social networking sites - can improve the user experience in planning a trip and provides to public administrations and operators in the turism sector the possibility to monitor tourist flows.

The COVID-19 pandemic has negatively impacted every sector worldwide. Still, the mobility industry seems the most disrupted (USD 31.7 billion in 2020 vs USD 55.3 billion in 2019 – EY Analysis Report), and mobility service providers will probably redefine their strategies (Renaud, 2020). Overall, industry players must demonstrate an unprecedented capacity to meet travellers' changing needs (Hall et al., 2020) and simultaneously initiate alternative and emergent methods (Nineties and Shutina, 2020; Wachyuni and Kusumaningrum, 2020). The new strategy to remain competitive will include their ability to reorganise and reactivate services combined with an effective interpretation of the market demand (Sigala, 2020). Therefore, the following question arises: can the mobility industry take advantage of this moment and, if so, how? Mobility companies are changing their core businesses and enriching market proposals with more integrated services. They are offering the possibility of purchasing integrated travel solutions with a single ticket, choosing between shared modes of transport: trains, subways, buses, ferries, car and bike sharing, taxis, and the opportunity to engage in traveller activities with multiple tourist services (de Oliveira and Cortimiglia, 2017) A data-driven approach could provide a valid solution for improving and enhancing traveller experience in integrated mobility services. In addition, many studies, ranging from value creation (e.g. Svensson and Gronroos, 2008; Vargo and Lusch, 2008; Zhang et al., 2020) and value configuration (Del Vecchio et al., 2018 *inter alia*) to value capture mechanisms (Bowman and Ambrosini, 2000 *inter alia*), have attempted to acknowledge and define how businesses' value could be co-created (Cabiddu et al., 2013; Merz et al., 2018). However, co-creation in mobility platforms is likely to remain contradictory for marketing and management scholars without better comprehending when and how to generate value and how the value can be estimated (Schulz et al., 2020). Accordingly, this study identifies the following aspects to understand how traveller experience can be improved: i) most appreciated travel destinations, and ii) main providers of entertainment and ancillary services that could enrich and upgrade the mobility business. Concerning these topics, this research intends to extract knowledge from data sources available on the Internet, i.e. Big Social Data (Cuomo et al., 2021) in compliance with the General Data Protection Regulation (GDPR), limiting the analysis to Italy. Thus, our research question is: how to improve and enrich the traveller experience using Big Social Data Analytics?

Before going on, it is necessary to introduce the useful concept of service-dominant (S-D) logic theory. The Service-Dominant (S-D) logic theory is a marketing and economic framework that shifts the focus from the traditional goods-dominant (G-D) perspective to a perspective centered around services. It emphasizes the fundamental role of service provision in creating value for customers and highlights the importance of co-creation and collaboration between customers and service providers.

In the S-D logic theory, value is not seen as something embedded in a product or a good, but rather as the result of the application of resources and competencies to address the needs and preferences of customers. The theory suggests that value is co-created through the integration of resources from both the provider and the customer, and it evolves through interactions and experiences.

This study takes a decision-making approach and applies a service-dominant (S-D) logic theory to integrate mobility services, and it analyzes how Big Social Data Analytics can be used to organize travellers' value propositions and engage them in co-creation (Del Vecchio et al., 2018; He et al., 2017; Wang and Alasuutari, 2017) of valuable mobility services (using travelers content support on the Web). Subsequently, by employing Recommender Systems, we surveyed travellers' requirements concerning Italian destinations to identify the two aspects mentioned above. Our proposal complements the literature on traveller experience by extending the notion of co-creation as the most productive and committed aspect that managers seek to develop with visitors in mobility markets. Furthermore, based on the theory of S-D logic (e.g. Svensson and Gronroos, 2008; Vargo and Lusch, 2008), our study clarifies how Big Social Data can be used as a key resource for the value configuration (Ardito et al., 2019; Wang and Alasuutari, 2017; Zhang, 2018) of travel experiences using the intelligent tool of recommendations (Garcia et al., 2011). Furthermore, given the relevance of Big Social Data Analytics (Bello-Orgaz et al., 2016; Nguyen and Jung, 2017), it is essential to understand how they can strengthen digital collaboration to co-value traveler experiences, with a final impact on creating value through a more appropriate value proposition.

On the one hand, the study findings will provide mobility managers with a better perspective on the advantages of a data-driven approach to improve the value generated for travellers. On the other hand, it would also help economic operators to increase revenues, reduce costs, and optimise business resources in value capture mechanisms. Therefore, to underline these considerations, the present study is organized as follows. We build on an existing study on Big Social Data Analytics and their impact on the co-creation of travel experiences in integrated mobility services. The proposed conceptual framework then describes the research design used for the empirical analysis, research hypotheses and data collection. The study concludes with a discussion of the theoretical, practical and social implications and limits of the proposed analysis, which led to the development of a software platform that is used by the marketing management of the Ferrovie dello Stato group (Nugo) which allows visibility of the movements of people flows and predict how they will evolve.

5.1 Big Social Data Analytics and travel experience in the integrated mobility business

While making travel plans, vacationers search for content that involves them by anticipating the travel experience, developing a strong sense of community linked to geo-referenced data, Big Data Analytics, the Internet of Things (IoT), end-user services, and cloud computing, combined with mobile technology and artificial intelligence, representing fundamental factors to deliver more personalized travel experiences and accomplish the transition towards more innovative and more competitive tourist systems (Susilo and Cats, 2014; Wang et al., 2016).

5.2 Value configuration

Among these factors, Big Data (i.e. complex data sets derived from smartphones, satellite images, social media platforms, and public monitoring devices) has undoubtedly brought benefits for tourism research, closely related to emerging technologies, new abilities in data processing, and innovative applications. Big data analytics, especially in integrated mobility services (Del Vecchio et al., 2019; Volo, 2019), represent a significant carter for value creation, contributing towards defining experiences in innovative technology. It has strengthened both traveler and business involvement and increased personalized offerings.

In this context, Big Data derived from user-generated content through popular social online services, namely Big Social Data, refers to the ‘social’ aspect of Big Data in response to various purposes of data analysis. Nonetheless, Big Social Data is used to extract information based on digital social interactions of people for either descriptive or predictive purposes to influence human decision-making and orientate business strategies for marketing (e.g. Wamba et al., 2015; Erevelles et al., 2016; Ducange et al., 2018). In particular, travelers’ voluntary characteristics represent a distinct trait, transforming them into members of an online community (Brandt et al., 2017) and may reveal what they cherish. These unstructured/semi-structured data are semantically richer than Big Data and represent a complete source for recommending the critical elements required for generating contextualized offers or co-creating personalized products/services with travelers, thereby promoting networking, collaboration, and innovation (Del Vecchio et al., 2018) and providing higher value even in real-time (Olshannikova et al., 2017).

5.3 Value creation and value capture mechanisms

From the perspective of travelers and S-D logic theory, active participation through online reviews and interaction with friends or peers on social media platforms may activate the co-creation process of the travel experience. According to Vargo and Lusch (2008), an S-D logic approach recommends that customers (businesses) are co-creators of value and are involved in the process of value co-design, where the roles of travelers and businesses are not specified (Svensson and Gronroos, 2008). The S-D logic “is firm-centric and managerially oriented” (Vargo and Lusch, 2008, p. 2) and originates in the foundational propositions that create value among businesses and tourists “in every aspect of the value chain and that it is the beneficiary who always uniquely and phenomenologically determines this value through value-in-use perceptions” (Merz et al., 2018, p. 79). Thus, a Big Social Data approach to co-create travel experiences increases the overall value for both mobility service providers and travelers. However, it is essential to help businesses increase revenues, reduce costs, and optimize resources to identify proper value capture mechanisms. In this regard, the amount of profit made cannot be determined solely on the basis of the analysis of business processes. According to Bowman and Ambrosini (Value creation versus value caption, 2000), although the source of differences in products manufactured (including production costs) across firms is associated with the use of specific resources peculiar to a firm, it also depends on the amount of profit realized by exchanging those products, in terms of capture mechanisms by the firm from customers and by resource suppliers from the firm. Therefore, comparisons are made with other suppliers and buyers in determining value capture. Profits are determined through exchanges the firm makes with resource sellers (including labour suppliers) and customers (Bowman and Ambrosini, 2000). Thus, value creation and value capture are distinct processes because the source that creates a value increment may or may not be able to capture or retain the value in the long run (Lepak et al. 2007). Based on these premises, we propose the conceptual research model and our research hypothesis as follows.

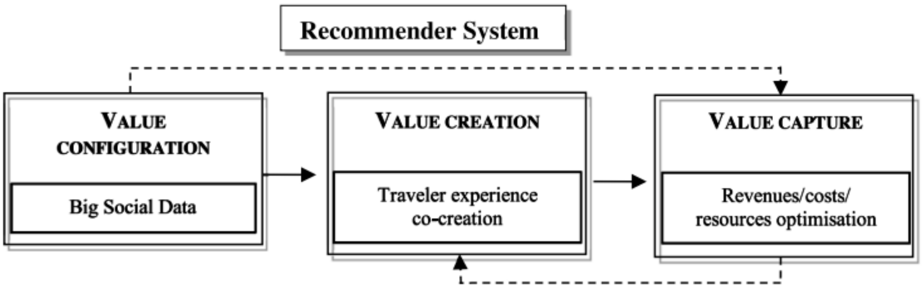


Figure 5: Recommender system schema

■ *A Big Social Data Analytics approach to experience co-creation affects the overall value for both travelers and decision-makers in integrated mobility services*

5.4 Methodology

5.4.1 Recommender System: a predictive tool to create knowledge

Recently, the number of services offered, as well as the number of customers using such services increased significantly, especially owing to the diffusion of cloud computing services. The popularity of social networks, mobile devices, and connected services has resulted in generating a large volume of data. Such user-generated data have accelerated the need for recommender systems "…to filter a set of items based on consumers' preferences, and thus predict a possible preference" (Pantano et al. 2019, p. 324). 'Recommender Systems' (RSs) have an essential role in helping users (i.e. travelers in this study) overcome data or service overloading situations, automatically indicating suitable data/services. The effectiveness of RSs is the ability to learn consumer preferences using past behaviour, predicting future trends and attitudes. Many studies show a significant use of recommended systems from e-commerce (Deng et al., 2014; Maniktala et al., 2016) and social commerce (Esmaeili et al., 2020) to tourism (Khalid et al., 2013; Yin et al., 2014). In other words, RSs support users in finding attractive and suitable solutions for a wide range of options. The main goal of these systems is to predict user's preferences available in the form of a recommendation list (Esamaieli et al., 2020, p. 1). As a result, the RS processes a large volume of data from social networks to reduce cold-start (Camacho et al., 2019) and make more adequate suggestions, specifically to satisfy the emerging needs of travelers (Logesh et al., 2019).

Materials and methods

This study deals with the support of Big Social Data Analytics for travel experience co-creation and focuses on integrated mobility services. Concerning the various transformations in travel experience, digital tourism is a growing trend in Italy, estimated at 14.2 billion euros, a quarter of the overall national tourism value that covers about 13% of the Italian GDP (Polytechnic of Milan, 2019). In 2019, Italy attracted 94 billion visitors, occupying fifth position worldwide (Italian Ministry of Tourism, 2020); in fact, the most accredited Italian estimates had foreseen a total turnover of almost 30 billion euros and an equally significant qualification of incoming tourism of 260 million visitors (-43,4% in 2020 compared to 2019, Cst, 2020), a contraction in the connected travel expenditure of around 4.5 billion (Demoskopika, 2020), and a potential recovery in 2022. Going ahead, the proposed RS requires a novel and efficient criterion for quickly ranking Italian tourist destinations based on Google search queries made available by Google Trends. Specifically, such a ranking criterion relies on the computation of a destination's popularity index, which is normalized in terms of the number of search queries made within the same geographical area (i.e. limited to the Italian territory) over a week. Later, given the collection of a weekly feed on the status of Google search queries concerning a selected set of Italian tourist destinations, each week, the proposed algorithm underlying the RS computes a ranking value for each destination. This value is calculated in terms of the deviation between the current popularity index and the average value of the time series of the weekly popularity indices obtained by the same destination over the previous four years. As a result, the final ranking of the tourism destination is obtained, and concerning each destination, the RS web scrapes the related current top attractions from three selected social networks: Tripadvisor, Minube, and Travel365. The result of this study was the creation of a software application, currently in operation, which is used by the marketing management of the Ferrovie dello Stato group (Nugo) to carry out analyzes on the movements of tourist flows. The use of this software application has also been proposed to

Public Administrations interested in the movements of tourist flows to better manage the services related to them.

Data collection and Recommender System

The data necessary to carry out the study and to implement the software application were collected using three travel social networks, Tripadvisor, Minube, and Travel365, with each directed towards capturing focused information more efficiently than others (i.e. Skyscanner and Booking were removed for technical remarks). Globally, Tripadvisor is considered the most prominent travel review site (Yoo et al., 2016) and the largest travel platform (Jumpshot for Tripadvisor, worldwide data, April 2019, www.tripadvisor.com), and it is often used in managerial literature as a data source for sentiment analysis tasks (Valdivia et al., 2017). Minube, on the other hand, is a travel social networking platform that combines online comments, photo-sharing, and video-sharing, providing travelers with an instant way of capturing and sharing their travel experiences with their friends and the virtual community (Bastidas Manzano et al., 2018). Travel365 is an online travel guide that merges travelers' opinions with information and sites worldwide. Such data are vital for marketers of mobility services companies to offer targeted collective, integrated mobility services to prospective travelers and regular customers, as tourism is typically focused on distinct destinations and seasonality demand. After a preliminary phase of data extraction, transformation, and loading, a time series analysis is conducted to provide suitable input to the proposed RS (see Appendix for platform architecture and microservices). The methodological proposal follows the flow depicted in Figure 6. Thus, our RS can 1) collect attraction approval ratings, 2) compare attractions from different sources, and 3) dynamically filter destinations and attractions in terms of service providers. Furthermore, insight 1) can be obtained through Python libraries that allow access to the Google Trends tool, while insights 2) and 3) can be obtained through a dedicated automatic web scraping activity.

Remarks (Data and Technical analysis)

Data collection covers the period from March to December 2020 concerning the restrictions of European legislation regarding the General Data Protection Regulation (GDPR, 2012). Due to technological advancement over the past 25 years, legislation was no longer fit for purpose, and regulations managing users' data required to review, as demonstrated by the series of headline-grabbing security scandals jeopardising personal data (Reding, 2010; Krystlik, 2017). Hence, considering the need for safeguarding citizens' privacy due to the growing usage of computers and digital devices, the evolutionary process is focused on restricting personal data processing for reasons such as errors in the data processed or illegal processing operations. In this picture and according to the GDPR, even though it is possible to extract customer data, there are limitations to performing machine learning analysis. In addition, we investigated data in terms of i) homogeneity, wherein data structures vary by site and cannot be often integrated with each other, and ii) meaning and richness, as several websites present very old, insignificant, and poor information. In conclusion, we only used real-time data available on Google to filter out travelers' top-searched destinations, along with travel sites, to find information on popular tourist destination attractions.

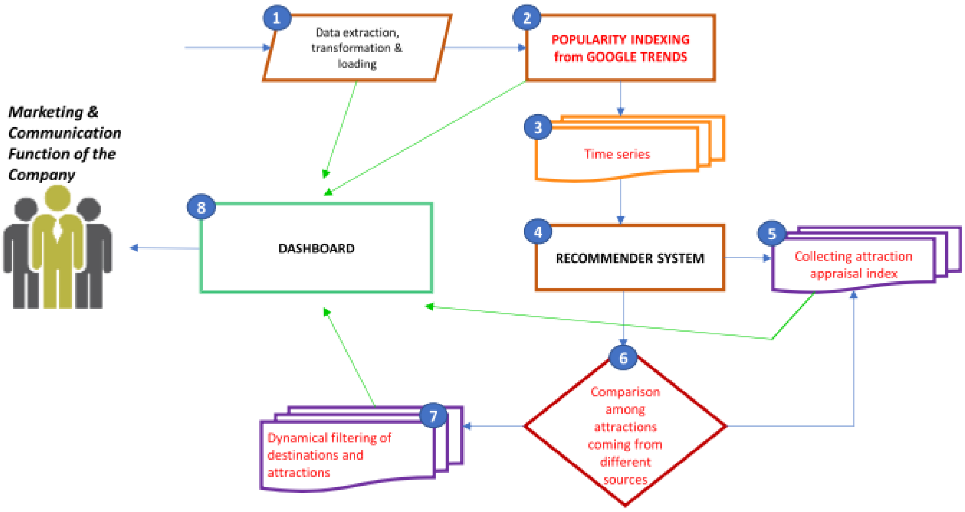

1
Data extraction, transformation & loading
2
POPULARITY INDEXING from GOOGLE TRENDS
3
Time series
Marketing & Communication Function of the Company
8
DASHBOARD
4
RECOMMENDER SYSTEM
5
Collecting attraction appraisal index
6
Comparison among attractions coming from different sources
7
Dynamical filtering of destinations and attractions

Figure 6: Decision support system archicture. Designed by authors

Some remarks based on technical analysis have been made. In this direction, some websites that provide the mechanism for blocking data downloads are not available, as they do not have a constant data structure that allows them to generate an automatic download code.

5.4.2 Findings

The proposed data-driven approach using the design of a RS (Figure 6) based on a Big Data Analytics engine makes it possible to: i) rank tourist preferences for the most attractive

Italian tourist destinations in Google, and ii) rank the main attractions (leisure, entertainment, culture) associated with single tourist destinations obtained from the analysis of relevant thematic websites, such as TripAdvisor, Minube, and Travel365. As discussed earlier, this empirical study examined online data that report the most visited and appreciated destinations, the most successful activities (cinemas, museums, theatres, parks), and services (including restaurants) relative to such destinations. We monitored 471 tourist destinations and classified them into three groups: seaside areas or beaches, mountains, and cities of art and culture. These destinations were selected from the most relevant Italian sources using the RS that allows to web-scrape the current top attractions from three social networks: Tripadvisor, Minube, and Travel365. By using the pytrends API (https://pypi.org/project/pytrends/) that allow to access data made publicly available by Google Trends, we define the so-called 'popularity index' P_k traveldest[k] for a specific tourist destination represented by the keyword traveldest (e.g. traveldest=′ Sorrento′) relative to the k-th week of the current year according to the following formula on a scale of non-negative real numbers from 0 to 100:

$$P_k = \frac{searches^{td}{}_k}{total\ searches_k} \quad \underset{k \in [1,..,52]}{Max} \quad \frac{total\ searches_k}{searches^{td}{}_k} * 100 \quad k = 1, \ldots, 52$$

where

searches^{td} denotes the number of searches made on Google during the k-th week of year and extracted via Google Trends relative to total searches, which accounts for the name of the selected tourist destination within the geographic area limited to Italy's borders (i.e. IT superscript); and

total searches represents the total number of searches on Google relative to all monitored tourist destinations within Italy's borders during the same week k in year. It means that year can be divided into 52 weeks.

Furthermore, the second factor of the multiplication in (1) is a normalization factor, represented by the maximum ratio, throughout the year, between the number of searches for the same destination and the total number of searches in Italy. This is just an adjustment we resorted to while designing the popularity index, forcing the score from (1) to always fall within the range [0,100] throughout the year. However, we should rank tourist destinations based on the popularity index. In that case, the eventual ranking will be considerably influenced by biases due to seasonality, since it will favour seaside destinations over mountain destinations during summertime and vice versa during the winter. Therefore, we resorted to the ranking value introduced below to obtain an unbiased popularity index.

$$Rv_{kt} = \frac{P_{k-1}+P_{k-2}+P_{k-3}+P_{k-t}}{t}\text{, k=1, …, 52 , t=1, …h}$$

Where P_{k-t}, for t = 1, …, h, denotes the popularity index of the traveldest evaluated in week k of year before. In the proposed results, we chose to rely on a horizon h = 4. Comparing tourism destinations in terms of Rv_{kt} rather than in terms of P_k total searches[k] allows us to make the RS independent of seasonality, characterizing the popularity of each group of monitored tourism destinations during the whole year. In simple terms, the popularity of the group accounting for mountain destinations, the group accounting for seaside destinations, and the group accounting for cities fluctuate around the same average value, when evaluated in terms of RV^Y traveldest[k], whatever the season of the year may be, in contrast to the evaluation made in terms of P^Y traveldest[k]. This way, given a determined week k of year Y, using RV^Y traveldest[k] we performed an unbiased comparison of all 471 destinations identified, showing the relative increase in the tourism destination's popularity represented by traveldest over the others, thus obtaining the overall ranking for the most sought after Italian tourism destinations during a specific week of the year.

In conclusion, it is evident that our findings, namely the Popularity Index and Ranking value in terms of the selection of tourism destinations proposed, validate the hypothesis, indicating that a Big Social Data Analytics approach to mobility experience co-creation affects the overall value for both travelers and integrated mobility services decision makers. Furthermore, given the collection of a weekly feed on the status of the Google search queries with respect to the selected list of 471 Italian tourist destinations, using the Recommender System, our algorithm calculates a ranking value reporting a popularity index for the same destination over the past four years. As a result, the final ranking of the tourism destination is obtained, and with respect to each destination, the RS allows to web-scrape the related current top attractions from the three social media websites previously identified: Tripadvisor, Minube, and Travel365.

5.4.3 Discussion and implications

Big Social Data is a valid approach to balance the planned and perceived value of the experience, with both the expected and the communicated value detaching the internal/external boundaries using the role of authentic travel-generated content (Ardito et al., 2019). Based on the S-D logic theory, the evidence from the framework completely supports the research hypothesis. Therefore, as a Big Social Data approach to the travel experience, co-creation positively influences the overall value for travelers and decision-makers in the mobility business, thus generating new forms of knowledge/information, innovation, and business (Narangajavana et al., 2019). In fact, the proposed Recommender System relies on innovative, efficient, and scalable criteria for ranking the most attractive Italian tourist destinations based on Google Trends search queries to provide relevant and helpful insights for the travel experience. The proposed approach appears to be well integrated with the existing tools and is used in the marketing and communication function of transport companies. Thus, on the one hand, this allows the creation of personalized marketing campaigns based on the knowledge of greater seasonality (greater preference flows) and, on the other, to select of the most sought-after attractions for the most preferred destinations. Achieving such a result is possible by adopting a data-driven strategy. In this respect, the success achieved by applying the Big Social Data approach to co-create travel experiences and service-dominant logic, particularly via recommender systems (Pantano et al., 2019), has been remarkable. However, this research shows the growing importance of Big Social Data for co-designing travel experiences based on creating a dashboard from a Recommender System (Figure 6). Thus, the proposed Big Social Data approach is valid. The study findings have several implications.

Theoretical implications

The Big Social Data approach and S-D logic provide information and knowledge from a system of recommendations (Esmaeili et al., 2020; Logesh et al., 2018) that improve the traveler experience based on the user-generated content and the use of the selected social networks. The application of an anthropological approach to Big Social Data constitutes a powerful method that deserves further development in future studies to enable both public and private sector organizations operating in the travel industry to improve their performance in terms of higher market share and profitability (Lalicic and Dickinger, 2019).

Practical implications

These insights are of great value to transport companies, as they allow segmentation and personalization of appropriate strategies for marketing proposals aimed at promoting the most attractive tourist destinations and directing seasonal tourism flows. The proposed approach has been implemented in a web application, namely the Recommender System, to make the results of this analysis available to corporate decision makers.

Social implications – impact on society and/or policy

Mobility operators must satisfy emerging needs, or renew old ones, on the basis of Maslow’s Pyramid (Maslow, A. H. (1943). A theory of human motivation. Psychological Review, 50(4), 370-396), linked to safety, which influences effective accessibility and pleasantness of the destinations, thereby affecting the actual demand for tourism and hospitality services. In this scenario, public agents must convert this weakness into an opportunity (Sigala, 2020) by investing in the phenomenon of under tourism and proximity tourism, as they are strictly associated with local development (Diaz-Soria, 2017) and as a valid solution to the dramatic freeze on the global hospitality offerings due to the COVID-19 pandemic.

5.4.4 Conclusions, limitations, and future research

Big Social Data and Recommender System seem to be the key sources of well-timed and rich knowledge (Bello-Orgaz et al., 2016; Garcia et al., 2011; Nguyen and Jung, 2017), helping in data-driven decision approaches that are oriented towards the management of complex relationships. Thus, applying a social approach to Big Data represents a new and alternative way to permit decision makers operating in the mobility industry and travelers to create valuable experiences. In this context, a data-driven approach can provide a virtuous mechanism of information/knowledge both on the supply side (business, community, institutions) and demand side (travelers) to transform and renew data flow among travelers, thus improving market share, revenue, profitability, and resource optimization (Lalicic and Dickinger, 2019) in terms of value capture and value for travelers. Despite the motivating outcomes, the main study limitation is the geographic area of the research process, as the country under analysis is Italy. Thus, extending the test to other countries would be useful, as this extension can be easily performed owing to the high scalability of the proposed procedure. Future research may consider comparing different clusters of national tourist destinations to evaluate the contrasting attributes of preferences for destinations. Reflecting on all these elements, the proposed research attempts to apply this decision-making approach to the mobility market in Italy, demonstrating how Big Social Data can be transformed into relevant value propositions in terms of tourism co-creation, along with the capability to enhance and enrich traveller experience. Faced with this new emphasis on data, there is a need to conduct more specific studies on the topic, moreover, if they can co-create and improve the travel experience in integrated mobility services. Therefore, the present research can be considered a first attempt to bridge this gap.

6. Online data: Sentiment analysis with Python

Premise

The second activity resulted from a project carried out for RAI. During the last years, many comments, judgments and discussions on television broadcasts have been taking place on social networks. For this reason, RAI is interested in acquiring textual data referring to specific broadcasts to evaluate their approval. For this reason, natural language processing approaches have been proposed that capture users' sentiment towards broadcast.

- Software Application Instarai. Technologies used: Python3, MySQL, Pandas, Docker, Power BI (used for the first time in Sanremo Festival 2021).
- Online data analysis: Sentiment analysis with Python. Maria Teresa Cuomo, Lorenzo Baiocco, **Ivan Colosimo**, Egon Ferri, Michele La Rocca, Lorenzo Ricciardi Celsi. In Research Methods in Marketing, Business and Management: Theoretical and Practical Perspective (Emerald, Pantea Foroudi Editor).

Companies are increasingly interested in knowing the purchasing behavior and opinions of their real and potential customers. All that complies with the new "post-modern" trend of marketing to develop a relationship with the customer that is different from the past, no longer convincing him/her to buy a product through the action of "pushing" towards the object of advertising (i.e., push strategy) but proposing a product different from all the others with unique and specific characteristics for the needs of customers. In other words, it would be desirable to “pull”, attract, (pull) customers towards their offers because these correspond exactly to what consumers want and need. To get ever closer to the predilections of end-users, companies already employ many diversified tools, such as: market surveys, focus groups, questionnaires and surveys, after-sales evaluations, collaborations with opinion leaders and with community of practice. Nevertheless, within these methods, companies work with experts of the market to obtain opinions, judgments, evaluations and feelings that can be useful for their activities. Obviously, all this information will be subject to the fact that interviewees know that their replies will be analyzed and will influence the results, or that the research field is limited to the questions or topics covered. In addition, to carry out the aforementioned investigations, it could be useful for companies to obtain "spontaneous" information, not obtained through a direct relationship with those who own them, in such a way to understand what are the "real" feelings and opinions associated with their product, service, brand or promotional activity. In particular, the study of opinions, feelings, judgments and emotions are the goal of the

Sentiment Analysis (SA), a statistical technique carried out on documents containing text, such as online comments made possible by the new potential of computers and by the immense availability of data that web users spontaneously generate every day. Thus, it is possible to identify the terms spontaneously associated with one's own product or those that are instead connected to the product of competitors. All that without "altering" the results and operating on infinitely larger volumes of data than those that it is possible to obtain from a questionnaire or, even less, from a focus group.

This chapter aims to present Sentiment Analysis, as a useful tool in the business field, in particular for management and scholars, starting from the description of the context in which it was born, observing the statistical theory and the logic on which it is based, finally examining a case study where he was has been applied for market analysis.

In actual fact, people often express their reactions, aspirations and desires through social media using a textual fragment of epigrammatic nature rather than writing long text. Further, unlike broadcast media, the content generated in online social media is immediate, spontaneous and unedited. Hence, many established industry players and scholars have started to analyze this ‘wisdom of crowd’.

6.1 Natural language processing and Sentiment Analysis

Sentiment analysis (SA) is part of the more general context of natural language processing (NLP), computational linguistics and textual analysis. It deals with automatically identifying and extracting opinions, feelings and emotions. contained in any textual document, an article, a review, a comment or a post on social networks, microblogs, blogs, forums.

Therefore, the goal of sentiment analysis is to identify and classify the opinions expressed on a specific subject, object, or topic or, more generally, in the entire document.

Several researches show that sentiment analysis turns out to be more difficult than a traditional data mining problem, such as topic-based classification, despite the generically lower number of classes, usually only two or three (Pang and Lee, 2008). One of the difficulties lies in the very little distinction that often exists between positive and negative sentiment (a distinction that, as mentioned earlier, is difficult even for a human being) but the main reason why sentiment analysis is more difficult than any topic detection problem is that this last can be solved with only the use of keywords which, unfortunately, do not work as well as for sentiment analysis (Turney, 2005). Further problems are due to syntactic ambiguities, the difficulty of determining the subjectivity/objectivity of sentences and texts, and the difficulty of deducing the domain dealt with, depending on the context.

Grimmer and Stuart (2013) – studying textual analysis– have elaborated four “principles" that every researcher must always keep in mind while examining a text:

1. “*All models are wrong, but some are useful*” (G.E.P. Box, 1976): the complexity of the language is so wide that any totally automatic analysis method cannot help but fail. The desire to create an automatic method with more and more rules to catch all the exceptions and nuances of a language is a frequent mistake that could create incorrect classifications. However, these methods can be useful for highlighting occurrences and frequencies of particular terms or associations among terms.
2. Quantitative methods help human beings, but they do not replace them: once again, due to the excessive complexity to which texts are subject, automatic text analysis techniques only make it possible to speed up the treatment and interpretation of texts, but they remain a support tool and not a substitute for human capabilities.
3. There is no perfect text analysis technique: in fact, every technique is designed for a specific purpose and is based on assumptions defined a priori, even more so when it comes to textual analysis. Think of the further limit that this type of analysis has in analyzing the differences between languages, based on the topics of discussion, the historical period, the age and gender of the writer.
4. Validate, validate, validate: Any new method or model must be able to be validated by the data itself. In particular, for supervised methods, validation is easier, they can be validated by checking the semantic attribution generated by the method and the objective semantic belonging of the text through post-classification reading. For unsupervised methods, validation is a more burdensome activity, since it requires the construction of controlled experiments, such as the insertion of texts whose semantic content is known but whose classification the algorithm does not know, and it is necessary to verify that the method assigns the text to the group which is assumed to be correct.

The main literature has focused particularly on the following classification problems related to sentiment analysis:

- Subjectivity classification, that deals with classifying a document into two classes: "objective" and "subjective" based on the "facts" and "opinions" contained therein (Tang et al, 2002).

- Sentiment polarity classification, that consists of identifying the orientation (positive, negative, neutral, etc.) of the opinion of the subjective sentences that make up a document in order to obtain a global classification.
- Opinion holder extraction, that deals with recognizing the author and the direct and indirect sources of the subjective sentences of the document.
- Object/feature extraction, especially in platforms – such as social networks and microblogs – where, basically, there is not only one specific and determined topic of discussion but, indeed, one is inclined to comment on several topics, it is extremely important to determine the entity subject to opinion.

The classifications above mentioned can be applied on different levels of granularity (whole document, single sentence, or single word). The classifications on different levels are not independent of each other but, on the contrary, each of them depends on the analysis of lower granularity: document-level classifications strongly depend on those in the sentence which, in turn, depend on those at the word level.

Document level sentiment analysis applies the classification process to an entire document, if it contains only the opinions of a single author on the same entity. Several researches (Zhao et al., 2014; Dhande, Patnaik, 2014; Sharma et al., 2014; Kumar et al., 2012) have been carried out on this subject both in the context of the classification of polarity and in that of subjectivity.

Sentiment level sentiment analysis deals with the recognition and classification of single sentences or short text messages. Various studies have been carried out in this regard, both in the context of the classification of polarity and in that of subjectivity. Among these, we can mention Yu, Hatzivassi-loglou,(2003), which used a supervised learning algorithm for identification and a method similar to that used in Turney (2005) for classifying sentences. subjective and Liu et al. (2005), who, on the other hand, formulated a simple method of aggregating the polarities of single words present in sentences.

Sentence-level sentiment analysis uses the a priori polarity of the individual words contained in the sentence itself. The polarity value is usually obtained by accessing an opinion word dictionary (called lexicon), built manually or automatically and suitably labeled. The manual creation of a lexicon involves the selection and classification of adjectives, nouns, verbs and adverbs starting from a traditional dictionary. For an automatic (or semi-automatic) creation, the techniques to be used are essentially two: dictionary-based approach (Fellbaum, 1998; Kim and Hovy, 2004) or corpus-based approach (Hatzivassiloglou and McKeown, 2004; Turney, 2005).

6.2 BERT-based Twitter Sentiment Analysis with Python

BERT stands for Bidirectional Encoder Representations from Transformers. It is a powerful natural language processing (NLP) model developed by Google. BERT is based on the Transformer architecture, which is a neural network architecture designed to process sequential data such as text. We propose a BERT-based sentiment analysis concerning tweets tagging Italian TV programs within a research project supported by a major Italian broadcasting company (ELIS Innovation Hub, 2021). The methodological approach followed for addressing this task is inspired by the six-step framework for managing data analytics projects introduced by Andreozzi et al. (2021), especially in terms of the structure of the data preparation activity.

Ultimately, we developed a model that processes tweets and returns the sentiment associated with them, thus offering the possibility to extract the maximum informative value from tweets and any photo attached to the tweets themselves.

So far, such a problem has been targeted mainly using Microsoft Azure APIs, yielding still unsatisfactory performance (average precision, recall and F1-score only very slightly above 50%), so the situation calls for a state-of-the-art neural network-based classifier exploiting an open-source development framework (i.e., Python and libraries such as TensorFlow, Keras, Hugging Face, etc.).

To train the model, from all the publicly available tweets tagging the TV programs of a major Italian broadcasting company, we extracted and labelled a suitable dataset. This task is application-specific since, as it very often happens in applied contexts (Andreozzi et al., 2021), an already labelled public dataset exhibiting the desired characteristics – in this respect, consisting of tweets tagging Italian TV programs – does not exist. Therefore, the extracted training set was labelled according to the following polarity classes: positive, negative, and neutral.

After dataset labelling, we carried out another important feature engineering step, namely we arranged a dedicated Python script for Optical Character Recognition (OCR) to extract text from any images attached to the tweets.

6.3 Dataset and Data Preparation

The considered dataset comprises 6,027 tweets containing an image or video attachments, with a training set of 4,746 tweets and a test set of 1,281 tweets.

Namely, as anticipated, the training set results from implementing a data preparation pipeline: data collection via scraping, data labelling, data balancing, and extraction of text from any images attached to the tweets.

6.3.1 Data Collection & Data Labeling

We downloaded the tweets by relying on a customized Python script. In this respect, Python, given a predefined hashtag, offers the opportunity to explore the latest published tweets with that hashtag (up to a predefined maximum bound, e.g., 5,000), downloading and saving text, date, and attached image (or the first frame of any attached GIFs and videos).

In order to label the data, we chose to exploit the potential of Python by relying on Pigeon (Germanidis, 2021), a simple widget that allows annotating a dataset of unlabeled examples directly from the development environment of a Jupyter notebook.

Data labeling is not always an easy task. Indeed, labeling a picture of a puppy as a dog or cat is a clear goal, while labeling a tweet as positive, negative or neutral, unfortunately, is not. Hence, to mitigate the subjectivity of the labelling action, we decided to enlarge the spectrum of training examples labelled as positive by including any slight appreciation. Similarly, we have broadened the spectrum of examples labeled negative to include those containing a slight criticism. This allowed us to obtain an adequately sized training set for the next aim to successfully train a neural network-based classifier. However, we must consider that sometimes the dividing line between the two classes above is not clearly drawn.

Also, for the sake of data cleaning, since the considered task is specifically focused on analysing the sentiment associated with the tweets tagging Italian TV programs, we discarded all tweets expressed in languages different from the Italian one, as well as all tweets matching the tag but whose contents were not explicitly referred to any Italian TV program.

Eventually, to evaluate the trained neural network-based classifier properly, we arranged the test set – consisting of 1,281 tweets – so that the number of test examples belonging to the negative polarity class – which in general are much less frequent – are balanced with the number of test examples belonging to the other two polarity classes (negative and neutral).

6.3.2 Extraction of Text from Images

To extract text from images, the best free open-source tool is Tesseract, a tool for OCR created by Hewlett Packard and currently sponsored by Google (Hewlett-Packard, 2021). Yet, Tesseract is trained to read text from black and white images and is not able to process automatically complicated images where the background is noisy and unstable. To solve this problem, we had to pre-process all training images with four transformations: (i) bilateral filtering, (ii) grayscaling, (iii) binarization, and (iv) inversion.

Then, we had to properly set Tesseract, taking only alphabetical characters and setting it to exploit Italian vocabulary. The proposed pipeline has been arranged to maximize performance over images formatted like Internet memes because most reaction pics on Twitter present themselves in the form of memes.

The first step we carried out on each image is bilateral filtering. In a nutshell, this filter helps remove the noise, but, by contrast with other filters, it preserves image edges instead of blurring them. This operation is performed by making sure that, when a point is blurred, the neighbours of that point that do not present similar intensities do not get blurred as well. The second operation we carried out on each image is projecting the RGB images into grayscale images. The last transformation we carried out is binarization to finally prepare the image for text extraction. A threshold value of 240 is defined so that, for every pixel, if the pixel value is smaller than the threshold, then the pixel is set to 0; otherwise, the pixel is set to 255. Since we have white text, we want to remove from the image everything that is not 255-white, that is, the colour of the text. Since Tesseract is trained to recognize black text, we must eventually invert the text colour from white to black.

Functional Architecture

The basic architecture (after the preprocessing stage described above) is quite simple and proposed in Fig. 7: the upper stream accounts for text processing while the lower stream accounts for image processing. Then, a fusion layer merges the outputs of the two streams, ultimately converging into a stream of classification layers.

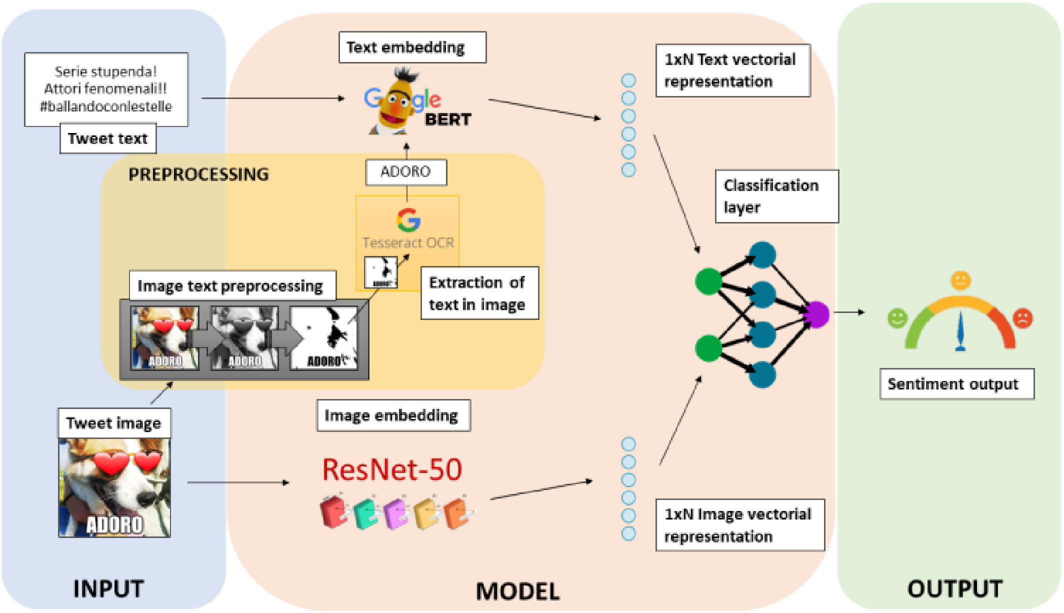


Figure 7: Functional architecture

For the text processing purpose (upper stream in Fig. 1), we chose to embed words into vectors using BERT, since this algorithm ensures state-of-the-art performance and it is also very easy to import in a TensorFlow environment, by means of the Hugging Face Transformers library (Hugging Face, 2021a).

Although many different implementations of BERT exist where fine-tuning was carried out with respect to different datasets and tasks in English, finding pre-trained architectures on Italian datasets is obviously more challenging.

In this respect, we made a comparison among four different backbones: (a) PoliBERT_SA, politic BERT based sentiment analysis (Gupta et al., 2020); (b) Neuraly, Italian BERT Sentiment model (Hugging Face, 2021b); BERTino, an Italian DistilBERT model (Muffo and Bertino, 2021); DistilBERT base multilingual model (Hugging Face, 2021c).

Out of these four backbones representing the state of art of BERT-based NLP techniques focused on Italian datasets, only the first two have been specifically fine-tuned for sentiment classification. Namely, PoliBERT_SA proved to be the best performing one, as it was fine-tuned for sentiment analysis on tweets (even if political ones). Thus it was chosen as the backbone of the upper stream of our functional architecture.

Relative to text processing, we first removed any URLs appearing in the tweets, as they are not informative; then, we removed all hashtags and transformed all emoji into text with similar meaning since the tokeniser of the proposed architecture is not trained to recognise them and they often carry strong sentiment information.

Afterwards, we merged the tweet text with the text extracted from the image thanks to Tesseract OCR (as shown in Fig. 1), using the Gated Multimodal Unit (GMU) approach proposed by John Arevalo et al. (2017).

Let us now consider the lower stream shown in Fig. 1, which, instead, is dedicated to image feature extraction. In this respect, ResNet50V2 (Keras, 2021) proved to be the most effective deep-learning-based object detection architecture for extracting any relevant sentiment-related information in the image that is not already contained in the extracted text. Yet, it must be noted that in a lot of tweets, the attached image contains no sentiment-related information at all, serving only as a picture of the object of the comment.

Subsequently, a fusion layer between the result of the TensorFlow PoliBERT sentiment analysis, on the one hand, and the result of the ResNet50V2 based object detection, on the other hand, is devised via simple concatenation.

Eventually, a classification layer is proposed in order to return the desired sentiment output. In this respect, classification consists of a layer of batch normalization, followed by three dense layers (of size 512, 128, 3, respectively) interspersed with 0.2 dropout; moreover, we chose a batch size of 4.

The text processing stream (the upper one in Fig. 1) was run on an nVIDIA GeForce MX150 graphics card equipped with 4GB GDDR5 RAM, whereas the image processing stream (the lower one in Fig. 1), due to its bigger dimensions, was trained on a dual-core Intel Core i7-7500U 2.70GHz (up to 3.50GHz) CPU.

Each experiment was carried out for ten epochs, although most of the information was usually extracted throughout the first three epochs. As an optimizer, we used RMSprop with an exponential decaying learning rate starting at 0.0001. The loss function we chose is classical cross-entropy.

Below we report in detail the layers of the overall neural network implementing the functional architecture proposed in Fig. 8.

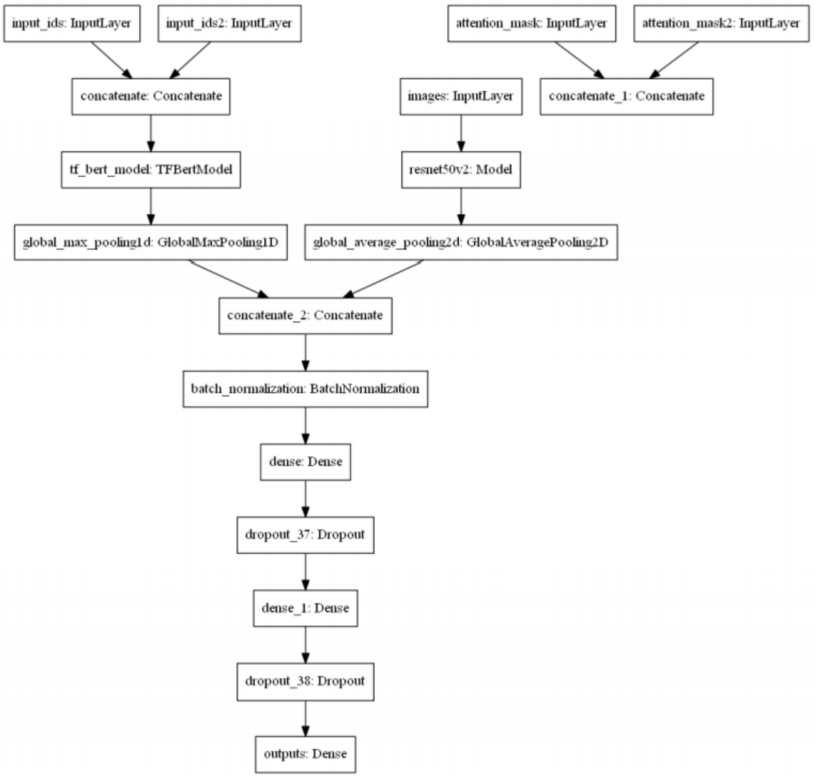

input_ids: InputLayer
input_ids2: InputLayer
attention_mask: InputLayer
attention_mask2: InputLayer
concatenate: Concatenate
images: InputLayer
concatenate_1: Concatenate
tf_bert_model: TFBertModel
resnet50v2: Model
global_max_pooling1d: GlobalMaxPooling1D
global_average_pooling2d: GlobalAveragePooling2D
concatenate_2: Concatenate
batch_normalization: BatchNormalization
dense: Dense
dropout_37: Dropout
dense_1: Dense
dropout_38: Dropout
outputs: Dense

Figure 8: Neural network architecture for BERT-based Twitter sentiment analysis

6.4 Results, limitations and extensions

Considering the very low baseline defined by the model currently in use by Italian broadcasting companies and based on Azure APIs (namely, on Azure Cognitive Services – Text Analytics (2021)), we obtained an upward bounce of performance of more than 10 percentage points on average. Namely, the performance in terms of precision increased up to 70% and in terms of F1-score up to 60% (see Table 2 below).

Remember that precision is the ratio of correctly predicted observations to the total number of predicted observations. Recall is the ratio of correctly predicted observations to all the observations in the considered class (either positive, negative or neutral sentiment). F1-score is the weighted average of precision and recall. Average precision in Table 2 is evaluated as the arithmetic mean among the precision results returned by the neural network with respect to each of the selected polarity classes of tweets received in input. Average recall and macro F1-score through the arithmetic mean of the corresponding per-class values, analogously to average precision.

Model	**Average Precision**	**Average Recall**	**Macro F1-Score**
Azure-based model	53%	44%	50%
BERT-based model	70%	55%	61%

Table 2: Performance comparison between the baseline model relying on Azure APIs and the proposed BERT-based model

The main responsible for the above-mentioned performance improvement is certainly the different approach to analyzing text. Since BERT came out, it has revolutionized the NLP field helping to exploit the potential lying in embedding the semantic value of words into numbers, and the considered use case confirmed its strength, by contrast with Azure APIs for text analytics.

Also, having the possibility to resort to a sentiment analysis model that has been pre-trained twice – the first time according to the classical BERT approach on a very big corpus of Italian training samples and the second time specifically with respect to Twitter sentiment analysis – has ensured a higher degree of generalization despite the relatively small dimensions of the training set.

Two other reasons for the increase in performance by comparison with the baseline model are the extraction of the text contained in the images attached to the tweets, as well as the valorization of the information content of the emoji. These features demonstrated to carry indeed a lot of information in terms of sentiment.

Anyway, as Table 1 itself suggests, there is still considerable room for improvement.

Indeed, as mentioned before, in sentiment analysis, there frequently occur some cases where the separation line between the positive and negative class is not clearly drawable at the training stage, and this reflects when it comes to using the model for inference purposes, into errors and inaccuracies that are unfortunately intrinsic to the methodological approach followed for training the model. The only way to address this is to purposedly rethink the data preparation stage, specifically focusing on identifying and correctly labelling a wide variety of training examples exhibiting this behaviour.

Another limitation is relative to the ineffectiveness of treating input tweets using ironic words: the negative sentiment that is usually associated with such tweets, especially with sarcastic ones, is still quite difficult to detect automatically and requires a dedicated effort in terms of deep learning (e.g., see Potamias et al. (2020) for an innovative approach recently proposed in the literature).

Finally, with respect to the image processing stream, there is still room for improvement as the convolutional neural network (currently implemented according to the ResNet50V2 architecture) could be trained to better learn the difference between side pics and reaction pics, focusing more on the latter pics rather than the former as carriers of sentiment information. This could be achieved by dedicated pre-training of the convolutional neural network and/or by restricting the search on the image through regional approaches.

On the basis of this work, a software application was created which is currently used in RAI to monitor the satisfaction of the broadcasts.

7. Segmenting data with Python. A quantitative analysis on the household savings

This chapter is based on a paper that was presented at the Aciek 2022 conference and analyzes household savings..

The research activity was inspired by a project commissioned by Cassa Depositi e Prestiti (CDP). CDP is a true "state bank," a financial institution that works through its services and group companies to support the country's development. CDP is the issuer of State-guaranteed postal savings bonds and passbook savings accounts. These products are distributed by Poste Italiane through its more than 12,000 branches located throughout the country. During the last academic year this study underlines as, under advanced analytics tools, household saving behaviors information and big data analytics may support data-driven decision approaches addressed in the managing of complex relationships in the financial arena. More punctually, using an exploratory and predictive analysis based on big data analytics and machine learning, this study aims to provide extensive customer profiling in the Italian household saving sector, supporting a data-driven decision-making approach. In this direction, the machine learning techniques have been aimed at behavioral segmentation, taking into consideration: i) homogeneous customers profile, ii) purchase/repayment paths and iii) churn.

- Software Application Household Monitoring. Technologies used: Python3, MySQL, Pandas, Docker, Tableau, Cloudera.

- Segmenting with big data analytics and Python. A quantitative analysis on the household savings. Maria Teresa Cuomo, Michele La Rocca, **Ivan Colosimo**, Debora Tortora, Lorenzo Ricciardi Celsi, Rosario Portera, Giuseppe Festa. Conference ACIEK (Academy of Innovation, Entrepreneurship, and Knowledge) 2022: 16TH edition – Greening, Digitizing and Redefining Aim in an Uncertain and Finite World. 28-30 June, Seville.

- Segmenting with big data analytics and Python. A quantitative analysis on the household savings. Maria Teresa Cuomo, Michele La Rocca, **Ivan Colosimo**, Debora Tortora, Lorenzo Ricciardi Celsi, Rosario Portera, Giuseppe Festa. Submitted to *Technological Forecasting and Social Change.*

In the advanced economic world, recent demographic changes (i.e., multi-income families, increased longevity, decreased fertility, and so on) have had a profound effect on household savings and their dynamics over time (Romei, 2021). In addition, the Covid-19 pandemic has caused a further and new leap in household savings (Mehta et al., 2020; Shehzad et al., 2021), evolving into higher savings rates (Statista, 2021), especially in relation to more defensive behaviours connected to growing socio-economic instability (Venieris & Gupta, 1986).

This trend has not spared Italy where, despite the recent decline in net wealth per capita, private households remain among the richest (Buklemishev, 2020; Ercolani et al., 2021) and least indebted in Europe (Acciari and Morelli, 2021). The propensity to Italian personal precautionary saving represents – also – a cultural trait of individual financial behaviour at the national level (Schunk, 2009; Fuchs-Schündeln et al., 2020). Therefore, research on this topic may prove to be crucial for predicting future behaviours regarding financial savings and finding a new, highly relevant fellow in the support provided by digital transformation (Reis et al., 2018).

On this stream, this study underlines as, under advanced analytics tools, household saving behaviours information and big data analytics may support data-driven decision approaches addressed in the managing of complex relationships in the financial arena. More punctually, using an exploratory and predictive analysis based on big data analytics and machine learning, this study aims to provide extensive customer profiling in the Italian household saving sector, supporting a data-driven decision-making approach. In this direction, the machine learning techniques have been aimed at behavioural segmentation, taking into consideration: i) homogeneous customers profile, ii) purchase/repayment paths and iii) churn.

Thus, our research question is as follows: in the actual context, are economic variables sufficient to explain and comprehend the household saving dynamics?

To answer to the abovementioned question a conceptual model based on a multidimensional approach to household savings has developed.

The outcomes obtained – by involving about 20millions of savers – showed several relevant clusters in the private savings sector in Italy, making the importance of a data-driven managerial decision-making approach. The research takes into consideration a great variety of savers, combining socio-demographic, psychological and economic aspects with significant policy implications. In addition, the study contributes to better understanding of the effect of digitalization in the context of economic research, that is a very actual topic for the digital transformation of the sector. Finally, the originality of the study consists in the high scalability of the proposed procedure that makes it able to be extended to many fields, also favouring cross-national and cross-cultural comparisons.

Therefore, to better underpin and enhance these considerations, the chapter is structured as follows. We draw on existing study in household saving and its impacts on customers financial behaviour to build our conceptual framework, that offers interesting research hypotheses. Succeeding, the chapter describes the research design used for the empirical analysis and data collection. Then, it concludes with a discussion of the managerial/practical and theoretical implications and limitations.

7.1 Undestanding the determinants of the household savings in Italy

7.1.1 Literature review on household savings

Furthermore, we can go through the main economic theories of saving to better understand the various drivers of household savings. Starting from the discounted utility model (Fisher, 1930), according to which the amount received in the future is less valued now than it will be later, and the Keynes positions (1936), describing savings decisions as stable over long periods of time and depending on the propensity to consume (linked to income increases) and the liquidity preference, two fundamental theories regarding saving challenged this focus on current income (Muradoglu and Taskın, 1996).

According to the life cycle theory (Ando and Modigliani, 1963), youth and the elder have different financial behaviours compared to the mature. Still, both of them diminish the savings rate for different reasons. Young people need to be supported by their parents while they do not reach employment age, while elders consume from the savings accumulated during their active life. Contextually, the increase in the average life span imposes the increase of the saving rate during the active life to maintain the level of consumption. The canonical permanent income (Friedman,

1957) considers that households respond to changes in permanent income, as for instance increasing the caution savings of the active population in order to compensate for a possible relative decrease of their income after retirement. Instead, they are not perceivable to changes in transitory income (Niculescu-Aron and Mihăescu, 2012).

Starting from these main theories, many studies bloomed, involving the psychological, behavioural and economic point of view as shown in Table 3.

Author	Field	Concept
Katona (1951, 1975)	Psychology	Individuals are influenced by their ability to save and their willingness to save, that is influenced by expectations and sentiment
Duesenberry (1949)	Psychology	The importance of peers on the likelihood of saving
Katona (1975) and Furnham (1985)	Psychology	Past savings experiences influence the likelihood of saving
Thaler and Shefrin (1981)	Behaviour	The behavioral life-cycle hypothesis: individuals have instincts to be both a planner who is concerned with lifetime utility and a doer who is focused on the present and practice mental accounting, applying different propensities to save in different categories of accounts
Campbell and Mankiw, 1990; Deaton, 1991	Economy	The role played in household saving by borrowing constraints
Kimball, 1990; Lusardi, 1998	Economy	The role played in household saving by precaution in the face of uncertainty
Xiao and Noring (1994)	Economy/psychology	In contrast with dominant economic models, they study several motives at once by

		arranging the motives in a hierarchy based on Maslow's hierarchy of needs
Boeree 1998, 2006	Economy/psychology	Savings motives are organized in a hierarchy, and that individuals move up the hierarchy as lower-level motives are satisfied. The savings motives in the hierarchy are (from low to high) physiological (basic), safety, security, love/societal, esteem/luxuries, and self-actualization

Table 3: Studies evolution on the household savings Source: author elaboration from: DeVaney et al., 2007; Bebczuk et al., 2015

7.1.2 The Italian context

Italy is one of the countries with the highest wealth-to-income ratio in the world. For instance, in 2016 the stock of net wealth owned by households was equivalent to six years of national income and almost eight years of household disposable income (Acciari and Morelli, 2021). While in 1966, the average personal net wealth per capita – that refers to the current value of all assets, tangible and intangible, that are under the control of the household sector, providing economic benefits to the holders, and over which property rights can be exercised – amounted to Euros 21,000 (2016 prices), at the end of 2006, just before the onset of the great financial crisis, the personal net wealth amounted to Euros 167,000, while it drop to Euros 141,000 in 2016. In 2018, household financial wealth decreased in Italy more than in the Eurozone. Despite this, the ratio between household net wealth and gross disposable income remains high in Italy, where the gross saving rate has headed towards 10%, albeit lower than the Eurozone value (Linciano et al., 2019). Instead, the incidence of household debt on GDP confirmed the distance between Italy and the Eurozone (in 2018, equal to 40% and 60%, respectively).

Afterwards, the outbreak of the Covid-19 pandemic early in 2020 had severe health, social and economic consequences at a global scale, which are expected to have worldwide consequences for a long time. Among the short time economic consequences, both in the US and in the euro area, there was a dramatic contraction of private consumption and a substantial jump in the saving rate due to the deterioration in economic conditions (decay in the labour market) together with other pandemic-related factors, such as the fear of infection, government-mandated restrictions and increased uncertainty due to health and economic perspectives (figure 9).

Italy also has recorded a sharper drop in household consumption (about 10% in 2020) and a stronger increase in private savings (above 20% in the last spring and above pre-pandemic figures at the end of 2020). More in detail, the long-lasting trend of an increasing preference for cash and deposits to the expense of bonds, equities and fund shares has been strengthened, confirming a household behaviour in terms of per capita financial investments significantly lower than in other main Eurozone countries (Linciano et al., 2020).

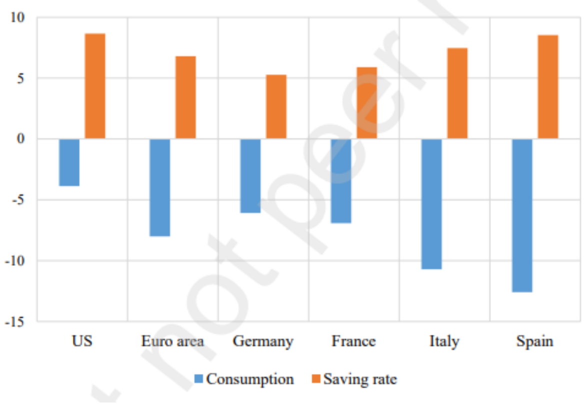


Figure 9: Changes in household consumption and the saving rate between 2019 and 2020. Source: Guglielmetti and Rondinelli, 2021

Because of its policy and economic implications, it is very relevant to understand if the recent Italian household behaviour was temporary and linked to lockdown measures and fears of infection; in other words, it would be useful to understand if most of the additional savings were involuntary or not. Simultaneously, it is necessary to investigate the precautionary reasons and financial investments of Italian households over time.

7.2 Methodology

By means of the information provided by the Big Data analysis, a profiling of household saving has been defined. To proceed in this direction, in the first phase of the study the hardware and software requirements necessary to perform the data processing have been considered. Data collection has been performed according to the so-called ETL (Extract, Transform, Load) process. The first stage in Data Collection consisted in extracting data from the identified data sources through suitably arranged queries (from traditional SQL to queries on data lakes via Impala or Hive). Remember that a data lake is a centralized repository that stores large volumes of structured, semi-structured, and unstructured data in its raw and original format. It is designed to store vast amounts of data from diverse sources, such as databases, logs, IoT devices, social media, and more, without the need for upfront data transformation or schema definition.

After the Data Collection from different databases, the phase of Data Preparation and reduction of dimensionality has been conducted, in order to have a consistent dataset, defined as the so-called Analytical Base Table (ABT).

Afterwards, advanced analytics algorithm using machine learning based Python libraries have been launched onto the ABT in order to perform analytics tasks such as user profiling, predicting churn phenomena, content recommendation to the user, price prediction, pattern recognition, and so on.

The results of such analytics tasks have been later collected and visualized through proper visualization tools (e.g., an executive dashboard). The related Data Visualization phase has been performed by means of several business intelligence tools such as PowerBI, Tableau and Qlik Sense, which are fully integrated with Python software functions.

The methodological phases followed were, as shown in Figure 10:

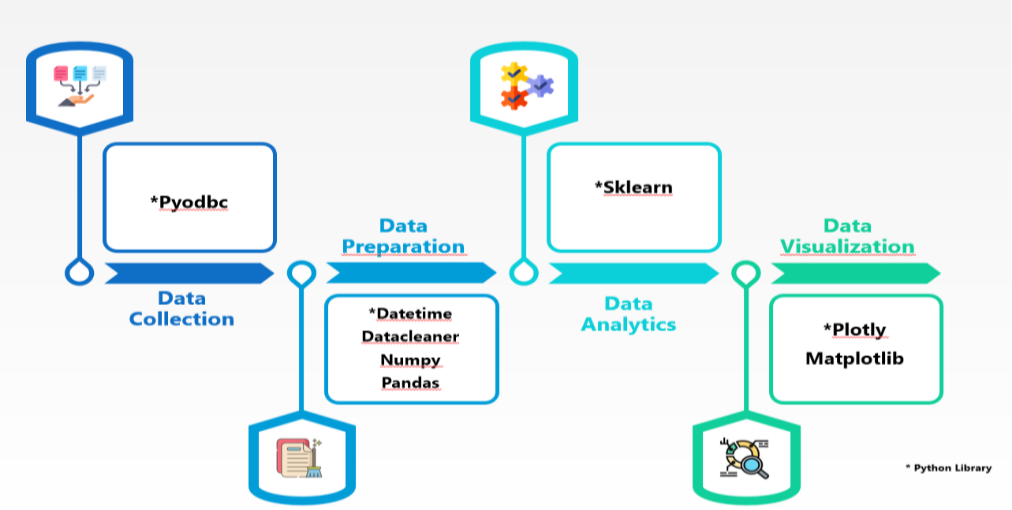


Figure 10: Methodology phases. Source: authors elaboration

- Data Collection;
- Data Preparation
- Cluster Analysis – Data Analytics;
- Data Visualization;
- Pattern recognition.

7.2.1 Data Collection

During the data collection phase, the data of savers of interest for the analysis were identified first. We can divide these data into three macro-categories: customer identification data, patrimonial data and data relating to movements.

The first category includes personal data, contactability, use of online services and customer seniority. The patrimonial data are, on the other hand, the number and amount of net deposits of products owned. In contrast, the data relating to movements are the number and type of operations carried out over time.

These data were loaded from the data lakes in the Python environment using the Pyodbc library, an open-source Python module that makes accessing ODBC databases simple. In this way, through queries, the data of interest were inserted into Pandas data frames.

7.2.2 Data Preparation

During the data preparation phase, raw data have been manipulated and transformed by changing the shape or structure of the data making them more suitable for learning algorithms.

The Data Preparation phase can be divided into two different phases.

In the first phase, using the Pandas and Numpy libraries, the data coming from the different databases have been re-elaborated and aggregated in order to build a final dataset containing in each row a unique saver and in the columns the features that characterize it. Since, as already mentioned, some data are time-dependent, variables have been created that take into consideration time intervals and their variation.

In this way a final dataset with more than 20 million observations has been realized.

In the second phase, the data were standardized. Since clustering techniques are based on the calculation of distances between observations, standardization is a highly recommended procedure in all those cases where a comparison is made between variables that have different units and orders of magnitude.

Indeed, it allows to scale data into suitable intervals that can be read by an unsupervised clustering algorithm. To perform this procedure, we used the Standard Scaled tool in the sklearn Python library based on the mean value and standard deviation of each variable.

7.2.3 Cluster Analysis – Data Analytics

Clustering is one of the most common tasks in data mining (Sun et al., 2017). The goal is to divide data items into groups according to pre-defined similarity or distance measure. More specifically, clusters should maximize the intra-cluster similarity and minimize the inter-cluster similarity. The K-means algorithm is one of the simplest and most widely used unsupervised learning algorithms to classify a given data set through a certain number of clusters fixed a priori (Hernández et al. 2012).

The K-means algorithm proceeds by combining adjacent data in a specific area and dividing them into several groups. It searches for a minimum degree of dispersion by clustering the data entry (every user) in a group using a distance-driven criterion iteratively, i.e. minimising the distance to the K-centroid of the belonging group compared to the one to the other groups (Kim et al. 2011).

In the K-means algorithm, the number of clusters depends on the initial setting of a K-value, here derived following the customary elbow method with the sum of squared errors (SSE) used as a performance indicator (Yuan et al. 2019).

7.2.4 Data Visualization

The most complex part of clustering is identifying the different groups identified by the algorithm.

The average results of each feature were compared for each of them to identify the different groups. We also used data visualization techniques such as the boxplot function of the seaborn library.

Using this function, it was possible to create boxplots of the variables of interest of each group, which are helpful graphic representations to verify the dispersion of certain numerical variables.

It also investigated the distribution of savers based on the group belonging to the Italian territory through interactive maps made through the library Plotly.

7.2.5 Pattern recognition

Once the different groups of savers were identified, the purchase and churn paths were analyzed. To such scope, we have served the diagrams of Sankey.

The diagrams of Sankey are a specific kind of diagram of flow, characterized by the presence of nodes connected by direct arrows that represent the flows of the process in the examination. The amplitude of the arrows represents the amount of operation. The direction of the arrows indicates the flow from a node to the successive one in the diagram.

The observation, for every cluster, goes through the flow of operations carried out in a determined temporal arc.

After identifying the savers belonging to every group, another type of data frame has been built, for every group, containing in every line a unique client, while in the columns been inserted the typology of movement executed by the saver in temporal order, thus identifying the typical behaviour.

Basing itself on the structure of the dataset realized to create the Sankey, it has been realized an algorithm that, choosing the number of last operations executed to analyze, allows for every customer to estimate the probability of the next operation in base to the successive operations executed by the savers that in the past have shown the same operational path.

Based on the same principle, the probability of abandonment – churn – of a saver was estimated by comparing the last operations carried out by a saver with those carried out by savers who have abandoned previously.

7.3 Findings and outcomes

Clustering on the data of almost 20 million clients in Italy has been performed (OECD 2022), Household net worth (indicator). Several relevant clusters emerged. Descriptive Statistics drilling down the properties of said clusters, including their delocalization, have been provided, together with the purchasing behaviour. Most of the outcomes suggest very useful insights to design a more valuable customer experience.

Through the k-means clustering algorithm applied to the dataset – suitably prepared so that a unique customer ID is associated with every row and all the related characteristics are reported in the columns - the following different profiles of customers have been identified.

Cluster no. 1. The largest group, consisting of 31% of the total number of clients, is made up of clients who are not active at all ***(inactive or dormant***), characterized each by the possession of an ordinary bank account but with a very low level of funds and few active movements.

*Cluster no. 2***.** The second group, counting 23% of the sample, is composed of *senior* customers with the highest average age (68 years), who are holders of ordinary bank account with pension credit, do not usually purchase other products and do not carry out a large number of cash transactions.

Cluster no. 3. The third group, amounting to 16% of the sample, can be associated with *loyal customers (savers)* with a high share of savings products and product purchases: they have average age and savings compared to the total number of customers analyzed (52 years and 35 thousand euros, respectively).

Cluster no. 4. The fourth group, amount to 10% of the sample, consists of *new customers*, that is, those who have the lowest customer seniority, who frequently use online services and are multi-banked.

The remaining 20% of customers are distributed among smaller groups with niche characteristics.

Cluster no. 5. Following this distinction, the fifth cluster was found to consist of very wealthy clients (*affluent investors*) with the largest assets.

Cluster no. 6. This is a group of *potential investors* with high average assets (although declining), high customer seniority, significant ownership of financial products, but who did not purchase other products in recent years and are not very active in their movements.

*Cluster no. 7***.** The seventh cluster was found to consist of clients who are very similar to the retiree cluster (senior plus), also including clients with a high average age although lower than the retiree cluster (62 years): they are owners of bank accounts with pension credit but compared to the larger cluster of retirees these are distinguished by a high level of physical purchases.

Cluster no. 8. The eighth cluster was found to consist of customers with the lowest average age among the groups identified (*young people online*), characterized by high use of online services, purchases made prevalently online, a greater presence of multi-bankers and higher amounts transferred, as well as high transactionality with online transfers.

Cluster no. 9. This is the cluster of digital potentials, amounting to 1% of the entire sample, with an average age of 44 years old, a declining amount of assets with average value of 10k€, almost inactive online accounts, and no purchases.

Cluster no. 10. This cluster can be identified with very few clients (0.1% of the total) who were potentially close to dropping out (*disappointed*), with very low and declining assets and a high level of early repayments.

Clusters 11 and 12. Finally, two particular groups have been identified, the first consisting of clients possessing products for people younger than 18 years old (*minors)* and the second gathering all clients who have abandoned savings services over the years (*former clients*).

From the territorial analysis of the clusters, a fairly homogeneous distribution of the various groups emerged, throughout Italy, with a higher number of clients in the North and South than in the Center, with a proportion of around 40%, 20%, 40%, respectively, (Figure 3).

The only exceptions are senior plus investors with a greater distribution in the South (over 50%, Figure 3) and affluent investors with more clients in Northern Italy (45%).

From the analysis of the purchase paths of clients, it emerged at a global level that clients who purchase a specific product tend to repurchase the same product with a higher percentage.

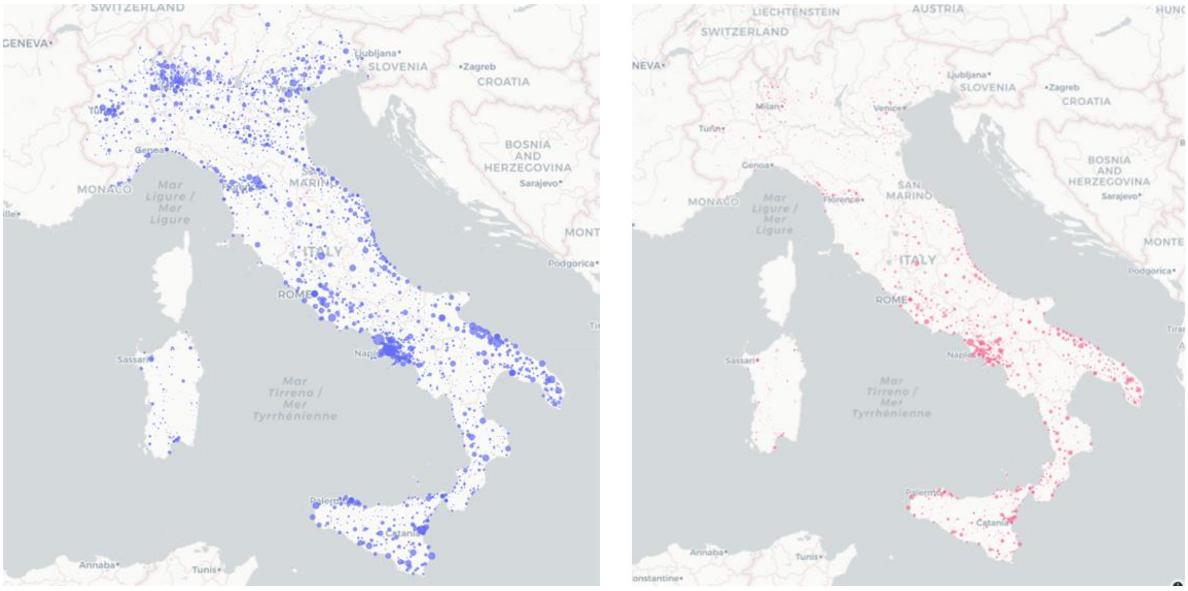


Figure 11: Purchasing paths and locations. Source: authors elaboration

For a non-disclosure agreement, the names of the specific financial products cannot be disclosed, therefore, we have divided the products into 5 broad categories:

- *Financial Product A (saving account sum tied for 6 month)*
- *Financial Product B (saving bond)*
- *Financial Product C (insurance policies)*
- *Financial Product D (saving account sum tied for 1 year)*
- *Financial Product E (saving account)*

More in detail, the purchase paths of the different clusters have been analyzed, and we report below some examples obtained on the most numerous or relevant clusters. For customers new to the devices, the most recurring purchase paths analyzed were the purchase of financial products, *Financial Product A* and *Financial Product B*. For customers belonging to the first path (purchase of smart booklet and subsequent issue of voucher) it was analyzed that 58% of customers with the same behavior tend to subsequently issue a subsequent *Financial Product B*, while 16% issue a *Financial Product C*. For clients belonging to the second path (purchase of smart booklet and subsequent issue of policy) it was analyzed that 61% of clients with the same path subsequently purchase a *Financial Product C*, while 26% purchase a *Financial Product B*.

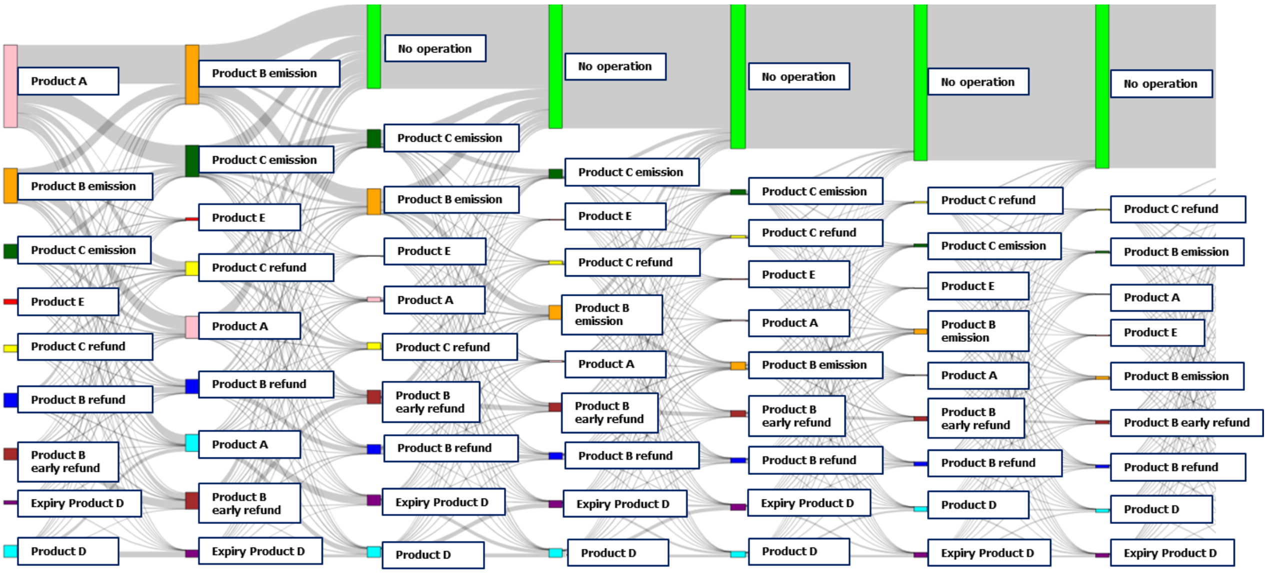


Figure 72: Sankey diagram cluster – New devices. Source: authors elaboration.

For affluent clients, the most recurring purchase path identified is the sequential issue of 3 *Financial Product B.* Analyzing the clients in the same cluster with the same purchase path who carried out further transactions, it appears that 82% of them issued a further *Financial Product B.*

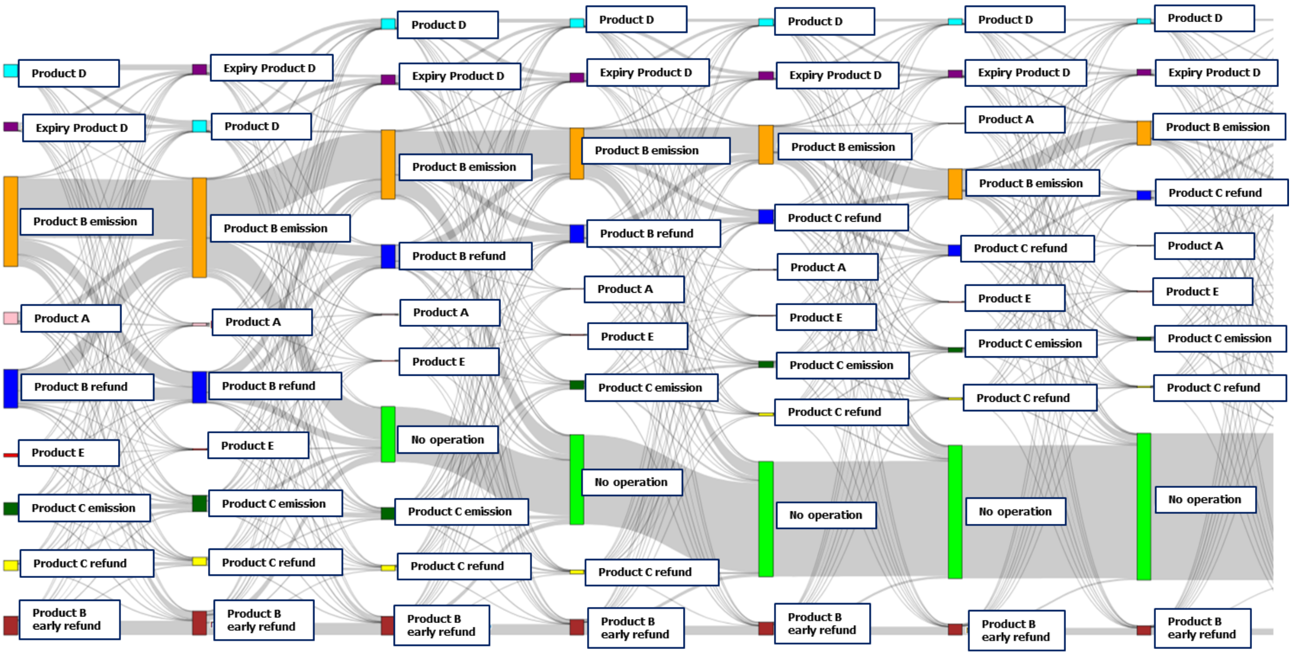


Figure 8: Sankey diagram cluster – Affluent

As regards saver clients, the most recurrent paths were: the issue of two consecutive *Financial Product B* or the expiration of a *Financial Product B* (expiry) and the subsequent issue of a new one. For clients belonging to the first path, a probability of 77% of issuing a new *Financial Product B* and 11% of early redemption of the *Financial Product B* was identified, while for clients with the second path, an almost equal probability of issuing a *Financial Product B* or redeeming it was identified (40 and 37% respectively).

For senior-plus clients, the most recurrent paths were: the issue of two consecutive *Financial Product B* or the redemption of two consecutive *Financial Product B*. For customers belonging to the first path, there was a 63% probability of issuing a new *Financial Product B* and a 23% probability of redeeming the *Financial Product B* in advance). In contrast, for customers with the second path, there was an 87% probability of redeeming another *Financial Product B* and only a 7% probability of issuing a new one.

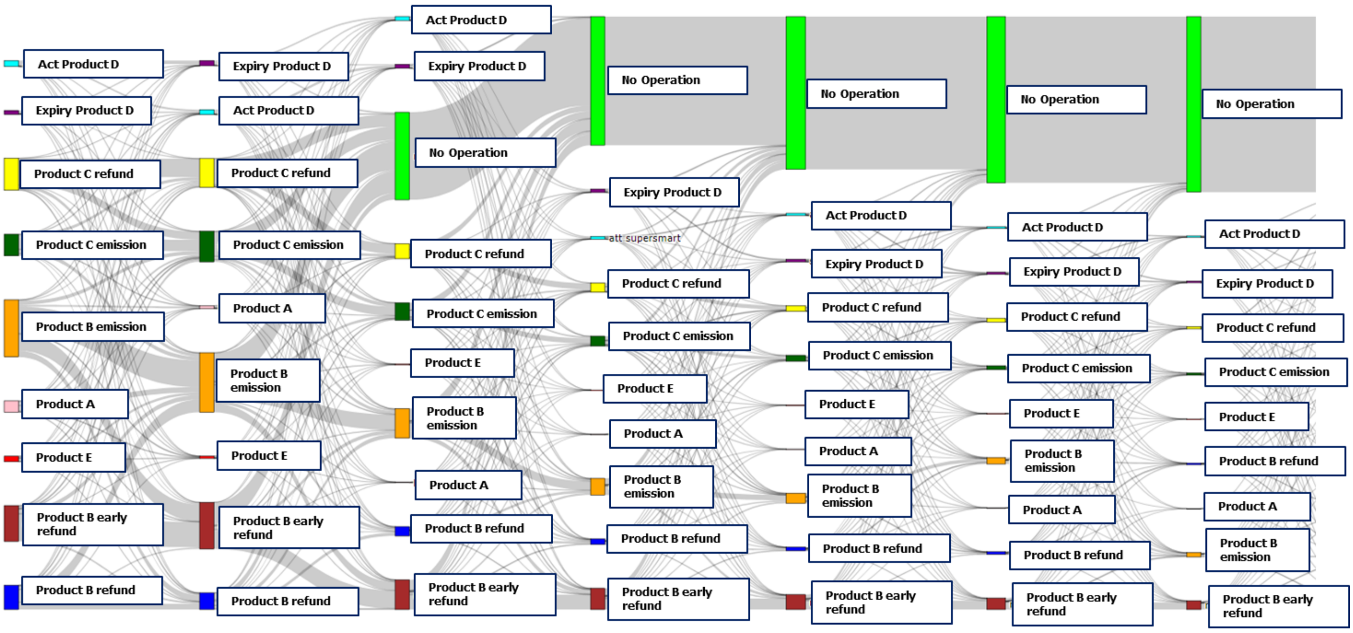


Figure 9: Sankey diagram cluster – Senior Plus

For young digital clients, the most recurrent paths were found to be: the issue of three consecutive *Financial Product B* or the issue of three consecutive *Financial Product D*. For customers belonging to the first path, there was a 79% probability of issuing a new *Financial Product B*, a 13% probability of redeeming a *Financial Product B* in advance and only a 3% probability of deactivating a *Financial Product D*), while for customers with the second path there was a 63% probability of

activating a further *Financial Product D*, a 28% probability of expiring a *Financial Product D* and only a 3% probability of issuing a *Financial Product B.*
From the analysis of the churn paths, it was found that clients who abandon (former clients) tended during the period analyzed to have a decrease in assets.

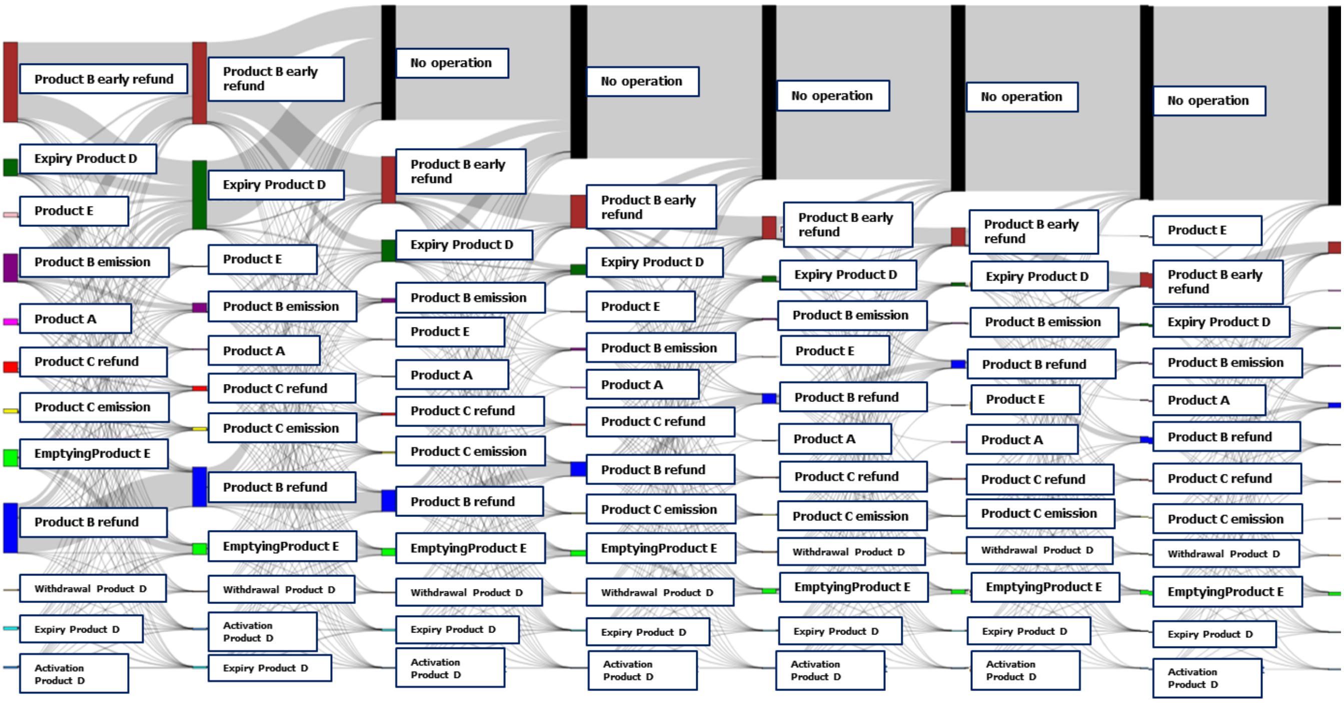


Figure 10: Sankey diagram cluster – Churn Cluster

The final and most frequent paths of abandonment (around 70% of cases) observed for this type of client are early redemption of *Financial Product B* and subsequent closure of the *Financial Product E* (21% of cases), early consecutive redemption of *Financial Product B* (maturities) (21%), consecutive redemption of *Financial Product B* (maturities) (14%), issue of a *Financial Product B* and subsequent early redemption of the *Financial Product B* (8%), emptying of the *Financial Product E* and consecutive closure of the *Financial Product E* (7%).

7.4 Discussion and main implications

To apply the exploratory analysis, a certain field of household saving has been chosen, i.e. financial savings, which is very popular in Italy.

The contribution of this work lies in the results obtained in terms of Data Analytics over a dataset accounting for the purchasing behaviour of almost 20 million financial savers. Our target in terms of Data Analytics is to perform unsupervised learning, namely clustering, intending to partition the data into a certain number of classes, such that each user (corresponding to a single data entry) is assigned to the cluster whose centroid is the closest to him/her.

We need first to go through the Data Preparation phase in order to succeed in this direction, and where we performed:

1. Data standardization, to scale data into suitable ranges that can be read by an unsupervised algorithm for clustering - in this respect, we used the Standard Scaled tool in the sklearn Python library based on the mean value and standard deviation of each variable.
2. Clustering, able to perform the dimensionality reduction of the dataset aimed at identifying the most suitable number of relevant clusters of users in the dataset.

This preparation activity paved the way for the proper and subsequent Clustering Analysis, performed employing the k-means algorithm. K-means clustering is relatively simple to implement, guarantees convergence, and is very flexible since it can be easily adapted to new examples. The clustering algorithm is efficient and scales well to large data sets (Fahad et al. 2014). K-means clustering can be implemented within the well-known MapReduce computational framework (Shirkhorshidi et al., 2014). Therefore, the overall procedure proposed here can be easily extended to Big Data using parallel computing and MapReduce software, such as Hadoop or Spark.

When we need to do clustering, it is essential to identify the most suitable ML algorithms for the analysis. In this paper, we choose to use k-means clustering.

It has several advantages concerning other competing techniques when considering the specific domain problem addressed in the paper and the data set dimension. Firstly, k-means clustering is relatively simple to implement, guarantees convergence, and is very flexible since it can be easily adapted to new examples. Secondly, in the case of large/big data applications, the clustering algorithm is very efficient and scales well to large data sets (Fahad et al., 2014). From the point of view of time complexity, k-means segmentation is linear in the number of statistical units, a crucial and desirable property when dealing with large data sets. Indeed, the computational cost is of order O(k*n*d), where k denotes the number of groups in the number of observations and d is the number of features. Finally, k-means clustering can be implemented within the well-known MapReduce computational framework (Shirkhorshidi et al., 2014). Therefore, the overall procedure proposed here can be easily extended to big data using parallel computing and software implementing MapReduce, such as Hadoop or Spark.

There are well-known drawbacks when using k-mean clustering. Firstly, the number of groups k must be chosen in advance. However, this can be considered as an additional flexibility of the overall procedure as it allows for the intervention of domain-specific knowledge and a more interpretable segmentation of the customers. Secondly, outliers can considerably impact the final grouping structure since outliers might drag centroids. However, the exact definition of outlying observation changes when dealing with big data. A small percentage of contamination results in a few outliers that can impact the final result in small data sets. However, when dealing with big data, the same small percentage of observations with quite different behaviour regarding the bulk of the data might be considered a sub-population that deserves specific attention. Finally, the well-known curse-of-dimensionality that affects k-means clustering small data sets is less severe here. We have millions of observations and only tens of features used for the clustering.

7.4.1 Theoretical implications

The study starts a big data analysis to profile customers in household savings (Attanasio & Banks, 2001) in Italy in order to support a data-driven managerial decision-making approach. The data-driven approach and multidimensional theory (economic, psychological and behavioural variables) provide insight and knowledge for more efficient clustering procedures (Punj & Stewart, 1986; Sun et al., 2017) that improve the saving experience. The application of such a composite approach to big data constitutes a powerful method that deserves further development in future studies to enable financial organizations operating to increase competitiveness and performance in terms of higher market share and profitability (Moro et al., 2014). Furthermore, at a theoretical level, this study contributes to a better understanding of the effect of digitalization in the context of economic research. The big data analysis of household savings can represent a valuable source of information to carry out new strands of research on personal financial behaviours.

7.4.2 Practical implications

These insights are of great value to financial companies, as they allow segmentation and personalization of proper strategies for marketing proposals aimed at promoting more homogenous profiles, more efficacious purchase/repayment paths, and clearer churn behaviours. Thus, the proposed approach provides for customized value propositions in terms of appropriate returns with respect to risks. Hence, the work focused on customer behaviour analysis of traditional banks/post offices. The data-driven management approach enabled by exploratory analysis, enriching and optimizing the organization of information assets, supports the creation of new systems, more in line with customer profiling.

7.4.3 Social implications

Despite the huge increase in digitalization and the spreadable use of digital devices to manage personal savings, the emerging profiles suggest some actions. Firstly, it would be useful to develop financial, and educational policies addressed to young and potential digital users, to build their beliefs on personal savings and help them to consciously manage them. Secondly, some policies to reduce the digital divide for elderly customers, ex-customers and wealthy senior customers are needed, so to exploit their value at a corporate level and to improve their personal benefits. Finally, the study underlines the social impact the household savings that need to be protected and encouraged using government programmes (Lusardi, 2008), in the actual context to understand whether the change in saving behaviour is permanent or transient (Mehta, 2020).

7.4.4 Research limitations and future research

Despite the interesting research, two main limitations seem to emerge. The first limitation of the study regards the geographic area of the dataset, since data referred to household saving in Italy. Thus, to overcome this limit, it would be useful to extend the test to other countries as well. Nonetheless, this extension can be easily performed due to the high scalability of the proposed procedure, that is the strength of the study. Future research, indeed, might compare different clusters of account holders across different countries to evaluate contrasting attributes of preferences about the financial choices in terms of savings. The second limitation of the research concerns the dataset as well as the results of the analysis, that are specifically relevant to the field of household savings. Thus, they cannot be extended to provide insight also in the field of bank savings, which is qualitatively very different. To fill this gap, it would be useful to extend the test to savings in Italy and other countries as well. To compare and contrast the results on financial savings at both a national and cross-national level would provide useful insights for professionals to manage personal savings.

Appendices

Appendix Chapter 5

The software made in correspondence with the development of this paper has an Architecture divided into 4 microsoftware:

- Software for downloading data from Google Trends;
- Google Trends data classification software;
- Back-End that exposes this data;
- Software Scraping (integrated in the Back end) that downloads data from the main sites: TripAdvisor, Minube and Travel 365;

The Front-End that communicates with the backend and shows the processed data.

Software microservice Googletrend

Description: it is the software that allows you to download the data relating to the cities that will be subsequently analyzed by the ranking software. It was written entirely in Python and through the pytrends module they are downloaded and saved on a MySQL database. It was developed to start automatically every Saturday at 01:00 at night (to avoid creating network traffic during the working day). For ease of publishing (distribution) support for Docker has also been included so that it can be quickly published on any server.

Technologies used: Python3, MySQL, Pandas, Docker

Software rankingtrends

Description: is the software that allows you to rank the N most popular cities with data extracted from Google Trends. Also written entirely in Python, it executes a series of algorithms that extract the cities that have a growing trend in the last 5 years, and then save these data on a MySQL DB. It was developed to start automatically every Saturday at 03:00, after the nugo / googletrends software has finished its extraction. For ease of publishing, Docker support has also been included in this software.

Technologies used: Python3, MySQL, Pandas, Docker

Microservice1. Backend

Description: Software for data exposure. Through a web server written in Flask, the backend provides a series of APIs for reading the aforementioned data. In addition, every Saturday at 05:00, it periodically automatically downloads the experiences related to the list of cities made available by the nugo / rankintrends software, taking information from the following sites: TripAdvisor, Minube and Travel 365. For ease of publication it has been added also Docker support for this software.

Technologies used: Python3, Flask, MySQL, Pandas, Docker

Microservice 2: Frontend

Description: The software provides the user with a graphical interface. The data is taken from the Nugo / backend software API and displayed in tabular form. The site presents a first page with the ranking data of the cities and related information (map, experiences by site, type), and a second, more specific page, with all the attractions taken from the previously mentioned sites. The software was developed with Docker support to facilitate publication.

Technologies used: Angular 10, HTML, SCSS, Typescript, Javascript, Docker

Appendix Chapter 6

The software made in correspondence with the development of this paper has an Architecture divided into 3 part:

- Software Scraping tool (integrated in the Back end) that downloads data from the main sites (Facebook, Twitter, Instagram.
- Database collecting data.
- Dashboard visualization.

Except for the webscraping tool, only proprietary market software (Microsoft) was used. This is because the request timeframe of the commissioning company was very tight, in particular, the activity started in September 2019 and had to be ready for the Sanremo festival Italian song (February 2020) to enable RAI's marketing people to intercept the broadcast's popularity through this tool.

Thus, in just a few months, fully functional software capable of processing information with the most advanced machine learning techniques was released.

Appendix Chapter 7

The software made in correspondence with the development of this paper has an Architecture divided into 3 part:

- Database collecting data;
- Data Analytics platform
- Dashboard visualization

Only market software (installed on premise) was used because of strict policies on data security issues.

References

Acciari, P., & Morelli, S. (2021). Wealth Transfers and Net Wealth at Death: Evidence from the Italian Inheritance Tax Records 1995–2016 (No. w27899). *National Bureau of Economic Research*, w27899, 1-46.

An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time 2021 (Santillana, Vespignani)

Analisi EY, 2021. https://assets.ey.com/content/dam/ey-sites/ey-com/it_it/topics/ey-mobility/mtt-report2020_211101.pdf.

Ando, A., & Modigliani, F., (1963). The Life-Cycle Hypothesis of Saving: Aggregate Implications and Tests. *American Economic Review*, 53(1), 55-84.

Ardito, L., Cerchione, R., Del Vecchio, P., Raguseo, E., 2019. Big data in smart tourism: challenges, issues and opportunities. *Current Issues in Tourism* 22 (15), 1805–1809.

Attanasio, O. R., & Banks, J. (2001). The assessment: household saving-issues in theory and policy. *Oxford review of economic policy*, *17*(1), 1-19.

Bastidas Manzano, A.B., Casado Aranda, L.A., Sanchez-Fernandez, J., 2018. La influencia de la Web en la reputacion online: el caso de Tripadvisor y Minube. *Revista Internacional de Turismo y Empresa*. RITUREM 2 (2), 3–27.

Batini, C., & Scannapieco, M. (2006). Data Quality: Concepts, Methodologies and Techniques. *Springer*

Bebczuk, R. N., Gasparini, L., Garbero, M. N., & Amendolaggine, J. (2015). Understanding the determinants of household saving: micro evidence for Latin America. *Documentos de Trabajo del CEDLAS.*

Bello-Orgaz, G., Jung, J.J., Camacho, D., 2016. Social big data: Recent achievements and new challenges. *Information Fusion* 28 (March), 45–59.

Buklemishev, O. V. (2020). Coronavirus crisis and its effects on the economy. *Population and Economics*, *4*, 13.

Bowman, C., Ambrosini, V., 2000. Value creation versus value capture: towards a coherent definition of value in strategy. *British journal of management* 11 (1), 1–15.

Brandt, T., Bendler, J., Neumann, D., 2017. Social media analytics and value creation in urban smart tourism ecosystems. *Information & Management* 54 (6), 703–713.

Cabiddu, F., Lui, T.W., Piccoli, G., 2013. Managing value co-creation in the tourism industry. *Annals of Tourism Research* 42, 86–107.

Camacho, L.A.G., Faria, J.H.K., Alves-Souza, S.N., Filgueiras, L.V.L., 2019. In *KDIR* (pp. 363-371). Social Tracks: Recommender System for Multiple Individuals using Social Influence.

Cuomo, M.T., Tortora, D., Foroudi, P., Giordano, A., Festa, G., Metallo, G., 2021. Digital transformation and tourist experience co-design: Big social data for planning cultural tourism. *Technological Forecasting and Social Change* 162, 120345.

Del Vecchio, P., Secundo, G., Maruccia, Y., Passiante, G., 2019. A system dynamic approach for the smart mobility of people: Implications in the age of big data. *Technological Forecasting and Social Change* 149, 119771.

Demoskopika, 2020, http://www.statistica.beniculturali.it/Visitatori_e_introiti_musei. htm.

Del Vecchio, P., Mele, G., Ndou, V., Secundo, G., 2018. Creating value from social big data: Implications for smart tourism destinations. *Information Processing & Management* 54 (5), 847–860.

Deng, S., Huang, L., Xu, G., 2014. Social network-based service recommendation with trust enhancement. *Expert Systems with Applications* 41 (18), 8075–8084.

de Oliveira, D.T., Cortimiglia, M.N., 2017. Value co-creation in web-based multisided platforms: A conceptual framework and implications for business model design. *Business Horizons* 60 (6), 747–758.

Diaz-Soria, I., 2017. Being a tourist as a chosen experience in a proximity destination. *Tourism Geographies* 19 (1), 96–117.

Ducange, P., Pecori, R., Mezzina, P., 2018. A glimpse on big data analytics in the framework of marketing strategies. Soft Computing 22 (1), 325–342.

DeVaney, S. A., Anong, S. T., & Whirl, S. E. (2007). Household savings motives. *Journal of Consumer Affairs*, 41(1), 174-186.

Dubey, R., Gunasekaran, A., Childe, S. J., Luo, Z., Wamba, S. F., Roubaud, D., & Foropon, C. (2018). Examining the role of big data and predictive analytics on collaborative performance in context to sustainable consumption and production behaviour. *Journal of Cleaner Production*, 196, 1508-1521.

Ercolani, V., Guglielminetti, E., & Rondinelli, C. (2021). Fears for the future: savig dynamics after the Covi-19 outbreak. Covid-19 note, *Bank of Italy*, forthcoming.

Erevelles, S, Fukawa, N, Swayne, L, 2016. Big data consumer analytics and the transformation of marketing. *J Bus Res* 69 (2), 897–904.

Esmaeili, L., Mardani, S., Golpayegani, S.A.H., Madar, Z.Z., 2020. A novel tourism recommender system in the context of social commerce. *Expert Systems with Applications* 149, 113301.

Fisher, I. (1930).*The Theory of Interest*. London: Macmillan.

Friedman, M. (1957). *A Theory of the Consumption Function*. Princeton: Princeton University Press.

Fuchs-Schündeln, N., Masella, P., & Paule-Paludkiewicz, H. (2020). Cultural determinants of household saving behavior. *Journal of Money, Credit and Banking*, 52(5), 1035-1070.

Galbraith, J.R. (2014). Organizational design challenges resulting from big data. *Journal of Organization Design*,3(1), 2-13.

Garcia, I., Sebastia, L., Onaindia, E., 2011. On the design of individual and group recommender systems for tourism. EXpert systems with applications 38 (6), 7683–7692.

Grable, J. E., & Lyons, A. C. (2018). An Introduction to Big Data. *Journal of financial service professionals*, 72(5).

Guglielminetti, E. & Rondinelli, C. (2021). Consumption and Saving Patterns in Italy during Covid-19 (June 22, 2021). Bank of Italy Occasional Paper No. 620, *Bank of Italy*. Available at https://ssrn.com/abstract=3891608.

Hans-Rüdiger Pfister, G. Böhm (2008) The multiplicity of emotions: A framework of emotional functions in decision making. Judgment and Decision Making Vol. 3, No. 1, January 2008

Hall, C.M., Scott, D., Go¨ssling, S., 2020. Pandemics, transformations and tourism: be careful what you wish for. Tourism Geographies 1–22.

He, W., Wang, F.K., Akula, V., 2017. Managing extracted knowledge from big social media data for business decision making. Journal of Knowledge Management 21 (2), 275–294.

Hernández, L., Baladrón, C., Aguiar, J., Carro, B., Sánchez-Esguevillas,A. (2012). Classification and Clustering of Electricity Demand Patterns in Industrial Parks. *Energies*. 5, 5215–5228.

Kannengiesser U. (2019) Empirical Evidence for Kahneman's System 1 and System 2 Thinking in Design. Conference: Human Behaviour in Design

Keynes, J. M. (1936). The supply of gold. *The Economic Journal*, 46(183), 412-418.

Khalid, O., Khan, M.U.S., Khan, S.U., Zomaya, A.Y., 2013. OmniSuggest: A ubiquitous cloud-based context-aware recommendation system for mobile social networks. IEEE *Transactions on Services Computing* 7 (3), 401–414.

Kim, Y., Ko, J.-M., Choi, S.-H. (2011). Methods for generating TLPs (typical load profiles) for smart grid-based energy programs. In: 2011 IEEE *Symposium on Computational Intelligence Applications In Smart Grid (CIASG)*. pp. 1–6. IEEE, Paris, French Guiana.

Krystlik, J., 2017. With GDPR, preparation is everything. Computer Fraud & Security (June), 5–8.

Ladley, J. (2012). *MK.* Data Governance. How to design, deploy, and sustain an effective data governance program.

Lalicic, L., Dickinger, A., 2019. An assessment of user-driven innovativeness in a mobile computing travel platform. *Technological Forecasting and Social Change* 144 (Jul), 233–241.

Laney, D. (2017). *Infonomics*: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage. Gartner.

Lepak, D.P., Smith, K.G., Taylor, M.S., 2007. Value creation and value capture: A multilevel perspective. Academy of management review 32 (1), 180–194.

Linciano, N., Caivano, V., Gentile, M., & Soccorso, P. (2020). Report on Financial Investments of Italian Households. Behavioural Attitudes and Approaches-2020 Survey. *Consob Statistics and analyses*, Rome.

Linciano, N., Costa, D., Gentile, M., & Soccorso, P. (2019). Report on financial investments of Italian households. Behavioural Attitudes and Approaches 2019 Survey, *Consob Statistics and analyses*, Rome.

Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., 2018. A personalised travel recommender system utilising social network profile and accurate GPS data. Electronic Government. an International Journal 14 (1), 90–113.

Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Li, X., 2019. Efficient user profiling based intelligent travel recommender system for individual and group of users.

Lusardi, A. (2008). *Household saving behavior: The role of financial literacy, information, and financial education programs* (No. w13824). National Bureau of Economic Research.

Mehta, S., Saxena, T., & Purohit, N. (2020). The new consumer behaviour paradigm amid COVID-19: Permanent or transient?. *Journal of Health Management*, *22*(2), 291-301.

Mezzanzanica, M., Boselli, R., Cesarini, M., & Mercorio, F. (2011). Data Quality through Model Checking Techniques. Proceedings of Intelligent Data Analysis (IDA), Lecture Notes in Computer Science vol. 7014 (p. 270-281). *Springer.*

Mezzanzanica, M., Boselli, R., Cesarini, M., & Mercorio, F. (2012). Data Quality Sensitivity Analysis on Aggregate Indicators. DATA 2012 - Proceedings of the International Conference on Data Technologies and Applications (p. 97-108). *SciTePress*.

Mezzanzanica, M., Boselli, R., Cesarini, M., & Mercorio, F. (2012). Toward the use of Model Checking for Performing Data Consistency Evaluation and Cleansing. *The 17th International Conference on Information Quality (ICIQ)*

Merz, M.A., Zarantonello, L., Grappi, S., 2018. How valuable are your customers in the brand value co-creation process? The development of a Customer Co-Creation Value (CCCV) scale. *Journal of Business Research* 82 (Jan), 79–89.

Mobile Networks and Applications 24 (3), 1018–1033.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22-31.

Muradoglu, G., & Taskın, F. (1996). Differences in household savings behavior: Evidence from industrial and developing countries. *The Developing Economies*, 34(2), 138-153.

Narangajavana Kaosiri, Y., Callarisa Fiol, L.J., Moliner Tena, M.A., Rodriguez Artola, R. M., Sanchez Garcia, J., 2019. User-generated content sources in social media: A new approach to explore tourist satisfaction. Journal of Travel Research 58 (2), 253–265.

Niculescu-Aron, I., & Mihăescu, C. (2012). Determinants of Household Savings in EU:What Policies for Increasing Savings? *Procedia - Social and Behavioral Sciences*, 483-492.

Nientied, P., Shutina, D., 2020. Tourism in transition, the post-Covid 19 aftermath in the Western Balkans.

Nguyen, D.T., Jung, J.E, 2017. Real-time event detection for online behavioral analysis of big social data. Future Generation Computer Systems 66 (Jan), 137–145.

Olshannikova, E., Olsson, T., Huhtam¨aki, J., K¨arkk¨ainen, H., 2017. Conceptualizing big social data. *Journal of Big Data* 4 (3), 1–19.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research*, 20, 134-148.

Reis, J., Amorim, M., Melão, N., & Matos, P. (2018, March). Digital transformation: a literature review and guidelines for future research. In World conference on information systems and technologies (pp. 411-421). *Springer, Cham.*

Romei V. (2021), "Global savers' $5.4tn stockpile offers hope for post-Covid spending", *Financial Times*, 18 April.

Schunk, D. (2009). What Determines Household Saving Behavior. *Jahrbücher für Nationalökonomie und Statistik*, *229*(4), 467-491.

Shehzad, K., Xiaoxing, L., Bilgili, F., & Koçak, E. (2021). COVID-19 and Spillover Effect of Global Economic Crisis on the United States' Financial Stability. *Frontiers in Psychology*, *12*, 104.

Statista (2021). Saving rate of households in Italy 2016-2022. Avalilable at https://www.statista.com/Statistics/1115319/saving-rate-of-households-in-italy/

Sun, L., Chen, G., Xiong, H., & Guo, C. (2017). Cluster Analysis in Data-Driven Management and Decisions. *Journal of Management Science and Engineering*, *2*(4), 227-251.

Venieris, Y. P., & Gupta, D. K. (1986). Income distribution and sociopolitical instability as determinants of savings: a cross-sectional model. *Journal of Political Economy*, *94*(4), 873-883.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165, 234-246.

Yuan, C., Yang, H., (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. J. 2, 226–235.